

**Computational analysis of cancer transcriptomes:
drug response prediction in colorectal cancer and
gene regulatory networks and long non-coding genes in
medulloblastoma**

Dissertation

zur Erlangung des Grades eines Doktors der
Naturwissenschaften (Dr. rer. nat.) vorgelegt von

Thomas Risch

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Berlin, 2021

Betreuer*innen: Dr. Marie-Laure Yaspo
Prof. Dr. Annalisa Marsico

Erstgutachter*innen: Prof. Dr. Annalisa Marsico
Dr. Marie-Laure Yaspo

Zweitgutachter*in: Prof. Dr. Stephan Beck

Tag der Disputation: 08.02.2022

Abstract

During cancer development, malignant tumours accumulate genetic and epigenetic alterations that cause dysregulation of gene expression and cellular processes. Since the regulation of gene expression controls many cellular processes, understanding the transcriptome of malignant tumours provides insights into the biology of cancer. Key technology for the molecular analysis of whole cancer transcriptomes is next-generation sequencing (NGS) of RNA (RNA-seq) from bulk tumours. However, to derive information about cancer transcriptomes from RNA-seq data, a variety of computational tools and analyses are needed. The following work presents two cancer transcriptome studies addressing the computational analysis of RNA-seq data from colorectal carcinomas (CRC) and medulloblastomas (MB) by applying statistical and machine learning (ML) methods.

CRC is a clinically challenging disease because only a fraction of tumours responds to available chemo- and targeted therapies. Functional loss of the tumour suppressor *APC* has been suggested to represent the initial mutation, activating Wnt signalling. Additional events include mutually exclusive mutations in the *RAS/RAF* proto-oncogenes as well as in the TGF β , PI3K, and TP53 pathways. Routinely used biomarkers of resistance to the EGFR inhibitor cetuximab are *RAS/RAF* mutations that activate signalling downstream of EGFR. Still, a fraction of wild-type CRCs is resistant to cetuximab treatment. Addressing the need for a better molecular understanding of CRC in precision oncology, the OncoTrack consortium (Innovative Medicine Initiative) designed a multi-omics strategy integrating the establishment of a pre-clinical platform for CRC organoid and animal models. In the study presented below, we focused on the integrative analysis of gene expression and drug response data obtained from patient-derived xenografts (PDXs) treated with cetuximab. Applying statistical methods, we identified a signature of 241 genes associated with response to cetuximab, which allowed us to dissect the expression profiles of responding and non-responding CRC. We used a support vector machine (SVM), a supervised ML algorithm, to obtain a gene-expression-based classifier predictive of response to cetuximab. Here, we selected 16 highly predictive genes using multiple SVM recursive feature elimination. The built classifier outperformed *RAS/RAF* mutations as a predictor of cetuximab response and performed well in *RAS/RAF*-wild-type CRC that currently lacks biomarkers of cetuximab treatment outcome in clinical practice.

The second study addressed the molecular analysis of MB. MB, a tumour of the cerebellum, is the most common malignant brain tumour in children. Transcriptome profiling of MB using microarrays had revealed four tumour subgroups, namely WNT, SHH, Group 3 and Group 4, each related to distinct genetic alterations, molecular profiles, and clinical features. Recurrent mutations mainly cause pathway activation in WNT and SHH MB, respectively, whereas in Group 3 and Group 4 MB, gross chromosomal alterations are more prevalent and tumours express a specific cell-type- rather than a pathway-related gene signature. Additional molecular complexity has been identified within these four main subgroups, which could be dissected further into subtypes. However, the gene regulatory networks that contribute to the molecular heterogeneity in MB are only partially known, and the role of long non-coding (lnc) genes remains poorly addressed in this disease. To gain further insights into the molecular biology of MB, the PedBrain project was founded within the ICGC framework. As a contribution to this project, we sequenced and computationally analysed 164 MB RNA-seq samples. Addressing the heterogeneity of MB, we identified and validated molecular subclusters within the four main subgroups. Subgroup- and subcluster-specific gene expression profiles were analysed by functional enrichments and gene regulatory networks (GRNs) inferred from gene expression data. These GRNs revealed communalities and differences in gene regulation among subclusters and subgroups. By estimating the impact of TFs, we could unravel master regulators of subcluster-specific gene expression in a systematic fashion for the first time and highlight unknown regulators of Group 4 MB. Furthermore, we characterised lnc genes that were differentially expressed in MB. Among these genes, we identified 20 lnc genes that show brain-development-associated expression patterns, which is of interest due to the embryonic origin of MB. We identified a co-expression cluster that accumulates known cancer-related lnc genes and associates these genes with cancer-promoting protein biogenesis. Survival analyses revealed the lnc gene *MEG3* as a prognostic marker in SHH and Group 4 subcluster, potentially acting as a tumour suppressor that negatively regulates cell cycle and TGF β receptor expression.

Preface

Acknowledgments

I would like to express my sincere gratitude to Dr. Marie-Laure Yaspo, who was my supervisor, moreover, a mentor. I want to thank her for the advice, support, trust, and numerous constructive discussions, which promoted my work and research.

I also would like to thank my second supervisor Prof. Dr. Annalisa Marsico for her advice on bioinformatics questions and for the invitation to the scientific meetings of her research group, where I gained insight into diverse aspects of bioinformatics.

I also thank the whole OncoTrack and ICGC consortium for the great collaboration. I had the pleasure of working with Prof. Dr. Stephan Beck during the OncoTrack project and would like to thank him for agreeing to be a reviewer of my thesis. I also would like to thank Prof. Dr. Gregor Eichele for adding the gene *MEG3* to the GenePaint database.

I am also grateful to all current and former colleagues of the Research Group Gene Regulation & System Biology of Cancer. I very much appreciate the teamwork that helped us to complete many extensive research projects. A special thanks to Vyacheslav Amstislavskiy for organising and processing the huge amount of data as well as to Daniela Balzereit, Alexander Kovacsovics, Matthias Linser, Sabine Thamm, and Simon Dökel for their excellent work in the lab and for always providing high-quality NGS libraries. A special thanks also to the IT department that always provided great support.

I also would like to thank Anne and Tim for proofreading and commenting on the thesis manuscript.

In addition, I want to thank my parents and my whole family for their continuous and caring support during my journey that started long before the work on my PhD. I also wish to thank my friends for always being there for me and the countless happy moments. I would particularly like to thank my partner Tim for being my clown, intimate, and solid rock.

Publication and contributions

This thesis is built on research that was conducted within the framework of two research consortia: OncoTrack and ICGC PedBrain. My work contributed to these consortia as a member of the research group “Gene Regulation & System Biology of Cancer” at the Max Planck Institute for Molecular Genetics. The later presented work that was done within the OncoTrack consortium was published in M. Schütte, T. Risch, N. Abdavi-Azar, K. Boehnke, and D. Schumacher *et al.*, “Molecular dissection of colorectal cancer in pre-clinical models identifies biomarkers predicting sensitivity to EGFR inhibitors”, *Nature Communications*, 2017 (listed authors are co-first authors). The whole list of authors is shown in Figure A.1.

The later shown results include computational analyses that were done by myself. Since the studies presented below were done within research consortia, I will use the pronoun “we” to indicate scientific collaborators and myself when the studies are presented. Contributions of consortium partners and other research group members are indicated in the Method parts of the two conducted studies. These contributions relate to work that was done before the computational data analyses that are presented below, including, among others, the collection of tumour tissue and clinical data, all experiments (NGS, pre-clinical models, methylation arrays), NGS data processing (read mapping, quality control), and the calling of mutations and copy numbers.

Contents

List of Figures	IX
List of Tables	XII
List of abbreviations	XIII
1 Introduction	1
1.1 Thesis outline	1
2 Cancer and gene expression	3
2.1 Cancer - malignant neoplasm	3
2.2 The genome and gene expression regulation	4
2.3 Genetic and epigenetic alterations in tumours	7
2.4 Hallmarks of cancer	8
2.5 Long non-coding genes: Functions and roles in cancer	10
2.6 Omics technologies in cancer research	13
2.6.1 Next-generation RNA sequencing	14
3 Computational analysis of cancer transcriptomes	17
3.1 Preliminaries	17
3.1.1 Notations	17
3.1.2 Machine learning and supervised learning	18
3.1.2.1 Resampling methods	19
3.1.2.2 Stratified resampling for classes imbalance	20
3.1.2.3 Performance metrics	20
3.2 Processing and normalisation of RNA sequencing data	21
3.3 Frequently applied data analyses of cancer transcriptomes	22
3.3.1 Unsupervised clustering for molecular subtype identification	22
3.3.2 Differential gene expression analysis	23
3.3.3 Functional overrepresentation analysis	24

3.3.4	Survival data analysis	24
3.4	Gene expression-based tumour sample classification	25
3.4.1	Binary classification and support vector machines	26
3.4.2	Feature selection and support vector machines	30
3.4.2.1	Multiple support vector machine recursive feature elimination	31
3.5	Inference of gene regulatory networks from gene expression data	31
3.5.1	Gene Network Inference with Ensemble of trees - GENIE3	32
3.5.2	Downstream analysis of gene regulatory networks	34
3.6	Computational characterisation of lnc genes	35
3.6.1	Characterisation lnc genes based on their genomic organisation with coding genes	35
3.6.2	Inference of putative lnc gene functions from gene expression profiles	36
3.7	Summary and outlook	37
4	Colorectal carcinoma study	39
4.1	Biological and theoretical background on colorectal carcinoma and anti-EGFR therapy	39
4.2	Study: Anti-EGFR therapy outcome prediction in colorectal carcinoma	41
4.2.1	Motivation	41
4.2.2	Overview and research scope	41
4.3	Results	42
4.3.1	Xenograft cohort	42
4.3.2	Gene signature related to cetuximab response profiles	43
4.3.3	Classifier predictive of cetuximab response	45
4.4	Material and methods	47
4.4.1	Tissue collection	47
4.4.2	Next-generation sequencing	47
4.4.3	NGS data processing	48
4.4.4	CNV and SNV calling	48
4.4.5	Gene set overrepresentation analysis	48
4.4.6	Establishment and characterisation of PDX models	48
4.4.7	Drug testing scheme in xenografts	49
4.4.8	Identification of drug-response-associated genes	49
4.4.9	Construction of SVM classifier of cetuximab response	50
4.4.10	Validation of the classifier of cetuximab response	51
4.5	Discussion	52

5	Medulloblastoma study	55
5.1	Biological and theoretical background of medulloblastoma and lnc gene <i>MEG3</i>	55
5.1.1	Medulloblastoma: Tumour of the cerebellum	55
5.1.2	Development and structure of the cerebellum	55
5.1.3	Subgroups in medulloblastoma	57
5.1.4	Enhancer-mediated gene regulation in medulloblastoma	60
5.1.5	Kinase activity profiles in medulloblastoma	60
5.1.6	Subtypes within the four main MB subgroups	61
5.2	Study: Gene regulatory networks and characterisation of lnc genes in medulloblastoma .	63
5.2.1	Motivation	63
5.2.2	Overview and research scope	64
5.3	Results	65
5.3.1	Subgroup classification and subcluster identification	65
5.3.1.1	Medulloblastoma cohort overview and main subgroup classification	65
5.3.1.2	Evaluation of the identified intra-subgroup subcluster	66
5.3.2	Gene expression profiles and gene regulatory networks in medulloblastoma	70
5.3.2.1	Summary of differential gene expression in subgroups and subclusters	70
5.3.2.2	Inference and evaluation of the gene regulatory networks	72
5.3.2.3	Medulloblastoma subgroups	74
5.3.2.4	Subclusters within SHH medulloblastoma	78
5.3.2.5	Subclusters within Group 3 medulloblastoma	80
5.3.2.6	Subclusters within Group 4 medulloblastoma	83
5.3.2.7	Impact of copy number variations on transcription factor expression	86
5.3.2.8	Overlay of gene regulatory networks in medulloblastoma subgroups and subclusters	88
5.3.3	Lnc genes in medulloblastoma	92
5.3.3.1	Potential relevance of lnc genes in medulloblastoma	92
5.3.3.2	Characterisation of lnc genes: Divergent, antisense and intergenic	93
5.3.3.3	Impact of copy-number variations on lnc gene expression	96
5.3.3.4	Lnc genes with brain-development-related expression patterns	98
5.3.3.5	Lnc genes described in the literature and their context in medulloblastoma	102
5.3.3.6	Non-coding tumour suppressor <i>MEG3</i> as a prognostic biomarker in medulloblastoma	109
5.3.3.7	Identifying genes correlated with <i>MEG3</i> expression in medulloblastoma	113
5.3.3.8	<i>MEG3</i> as a regulator of proliferation and the TGF β pathway in medulloblastoma . .	114
5.3.3.9	<i>MEG3</i> expression in normal brain and cerebellum	120
5.4	Material and Methods	123
5.4.1	Preliminaries: Correlation coefficient evaluation	123
5.4.2	Discovery cohort	123
5.4.2.1	Tissue collection, clinical data, RNA-sequencing, and RNA-seq read processing . . .	123
5.4.2.2	DNA Methylation data and subgroup classification	123
5.4.2.3	CNV data	124
5.4.3	External resources	124
5.4.3.1	External medulloblastoma cohort	124
5.4.3.2	Active enhancers and ChIP-seq in medulloblastoma	124

5.4.3.3	BrainSpan: External RNA sequencing data from human brain tissue	124
5.4.3.4	FANTOM CAT	125
5.4.4	Transcriptome profiling	125
5.4.4.1	Clustering into subgroups	125
5.4.4.2	Molecular subcluster identification within subgroups	125
5.4.4.3	Differential gene expression analysis	126
5.4.4.4	Gene set overrepresentation analysis	127
5.4.5	Gene regulatory networks inference	127
5.4.5.1	Annotation of transcription factors	127
5.4.5.2	Implementation of GENIE3	127
5.4.5.3	GRN fitting score	128
5.4.5.4	Module detection in GRN	129
5.4.5.5	Network influence score of transcriptions factors	129
5.4.5.6	Network visualisation	130
5.4.5.7	Evaluation of the GRN validity	130
5.4.6	Characterisation of lnc coding genes	130
5.4.6.1	Considered lnc genes	130
5.4.6.2	Annotation of lnc gene types and coding partners	130
5.4.6.3	Expression correlation with coding partners	131
5.4.6.4	Annotation of tissue-/cell-type-specific gene expression	131
5.4.6.5	Annotation of lnc gene associated publications	132
5.4.6.6	Co-expression clustering	132
5.4.6.7	Gene expression-based survival analysis	132
5.4.6.8	<i>MEG3</i> binding motif and binding site prediction	133
5.4.6.9	<i>MEG3</i> -centred expression correlation analysis	133
5.4.6.10	Gene set enrichment in <i>MEG3</i> -correlated coding genes	134
5.4.6.11	DNA methylation analysis at the <i>MEG3</i> locus	134
5.5	Discussion	134
5.5.1	Molecular subgroup and subcluster identification using RNA-seq	135
5.5.2	Inference of gene regulatory networks in MB subgroups and subclusters	136
5.5.3	Expression profiles and gene regulatory networks in subgroup and subcluster	138
5.5.4	Characterisation of differentially expressed lnc genes in MB	143
6	Implications and conclusions	151
6.1	Machine learning-based classification and treatment outcome prediction in colorectal carcinoma	151
6.2	GRN inference and expression profiles and regulators of the molecular heterogeneity in medulloblastoma	151
6.3	Computational characterisation of lnc genes and their involvement in medulloblastoma	153
A	Appendix	157
A.1	Authors of the OncoTrack publication	157

Contents

A.2	Performance of anti-EGFR therapy outcome prediction	158
A.3	Expression pattern of TFs in MB	159
A.4	Expression pattern of lnc genes and related coding genes	180
A.5	Gene symbols and their description	195
	Bibliography	199
	Zusammenfassung	229
	Selbstständigkeitserklärung	230

List of Figures

2.1	Triplex formation between RNA and DNA.	12
3.1	SVM is a maximum margin classifier.	27
4.1	EGFR and RAS/RAF/MAPK signalling.	40
4.2	Analysed CRC xenograft cohort.	42
4.3	Comparison of gene sets related to cetuximab response profiles.	43
4.4	Gene signature associated with responsiveness to cetuximab treatment.	44
4.5	Validation of the predictive cetuximab response classifier.	46
5.1	Anatomy, histology, and development of the cerebellum.	56
5.2	ICGC PedBrain MB RNA-seq cohort summary.	65
5.3	Comparison of RNA-seq-based unsupervised clustering and DNA methylation-based classification of the PedBrain MB cohort.	66
5.4	Identification of subclusters within subgroups in the PedBrain MB cohort.	67
5.5	Identified subclusters in SHH, Group 3, Group 4 MBs.	68
5.6	Comparison of identified intra-subgroup subclusters to previously published subtype definition.	69
5.7	CNVs in subclusters.	70
5.8	Identified differentially expressed genes in MB subgroups and subclusters.	71
5.9	Cutoff evaluation for interaction weights of the GRN of MB subgroups.	72
5.10	Cutoff evaluation for interaction weights of the GRN of MB subclusters.	73
5.11	Functional enrichments and GRNs in MB subgroups.	76
5.12	Overlap of putative targets between the photoreceptor signature defining TFs CRX, NRL and RAX2.	76
5.13	GRN of SHH subclusters.	80
5.14	Functional enrichments in SHH subclusters.	80
5.15	GRN of Group 3 subclusters.	82
5.16	Functional enrichments in Group 3 subclusters.	82
5.17	Functional enrichments in Group 4 subclusters.	83
5.18	GRN of Group 4 subclusters.	86
5.19	Copy number and expression of TFs downregulated in WNT MB.	87
5.20	Copy number and expression of <i>NEUROD2</i>	88
5.21	Comparison of subgroup and subcluster GRNs.	89
5.22	Aggregated GRN in MB.	90
5.23	Functional enrichments among targets of selected TF.	91
5.24	<i>De novo</i> identification of MB subgroups based on the expression profiles of the 1643 most variable lnc genes.	92
5.25	Comparison of hierarchical clustering based on specifically upregulated coding or lnc genes.	93
5.26	lnc gene type annotation and coding neighbourhood.	94
5.27	Correlation between pairs of lnc genes and coding genes partners.	95

List of Figures

5.28	Expression levels of lnc genes and coding partners.	96
5.29	<i>RP1-234P15.4</i> expression is influenced by copy number variation.	97
5.30	Expression profile of <i>RP1-234P15.4</i> in MB.	98
5.31	Annotation of stem-cell-related as well as prenatal- and postnatal-brain-related expression of 20 lnc genes.	99
5.32	Expression profiles of <i>GLYCTK-AS1</i> and <i>GLYCTK</i> in MB and CB controls	100
5.33	Scatter plots of <i>GLYCTK-AS1</i> and <i>HES5</i> expression.	101
5.34	Annotation of literature knowledge.	102
5.35	Co-expression cluster containing <i>ZFAS1</i> , <i>GAS5</i> , <i>DANCR</i> , <i>PVT1</i> and <i>SNHG16</i>	105
5.36	Expression profile of <i>MYC</i> , <i>MYCN</i> , and <i>MYCL</i> in MB subgroups.	106
5.37	Comparison of Cc1 mean-pattern to <i>MYC</i> , <i>MYCN</i> , and <i>MYCL</i> expression in PedBrain MB samples.	107
5.38	Comparison of Cc1 mean-pattern to the summed normalised expression of <i>MYC</i> , <i>MYCN</i> , and <i>MYCL</i> in PedBrain MB samples.	108
5.39	<i>MEG3</i> expression in MB.	109
5.40	<i>MEG3</i> expression is prognostic of survival in MB.	110
5.41	Association between <i>MEG3</i> expression overall survival in MB subgroups.	111
5.42	<i>MEG3</i> expression is prognostic of survival in SHH subclusters MBs.	112
5.43	<i>MEG3</i> expression is prognostic of survival in Group 4 subclusters.	113
5.44	Evaluation of filtered spurious correlations in gene expression correlation analysis.	114
5.45	Genes and pathways negatively correlated with <i>MEG3</i> expression in MB.	116
5.46	<i>MEG3</i> is negatively correlated with <i>CDK1</i> , <i>CCNB1</i> and <i>MYC</i>	117
5.47	<i>MEG3</i> expression is a potential negative regulator of <i>TGFBR1</i>	119
5.48	Methylation of <i>MEG3</i> -DMR in MB subgroups.	120
5.49	<i>MEG3/Meg3</i> expression in human and mouse brain.	121
5.50	<i>MEG3</i> and <i>TGFBR1</i> correlation in human brain.	122
A.1	List of authors of the OncoTrack publication.	157
A.2	NIS of TFs downregulated in a subgroup or subcluster.	159
A.3	Expression profile of <i>HSF2</i> , <i>DEK</i> , and <i>HDAC2</i> in MB.	159
A.4	Expression profile of <i>MYC</i> and <i>HLX</i> in MB.	160
A.5	Expression profile of <i>NRL</i> and <i>CRX</i> in MB.	160
A.6	Expression profile of <i>OTX2</i> , <i>RREB1</i> , <i>NEUROG1</i> and <i>TBR1</i> in MB.	161
A.7	ISH of <i>RREB1</i> in P56 mice cerebellum.	162
A.8	Expression profile of <i>AGAP2-AS1</i>	162
A.9	Expression profile of <i>NEUROD2</i> , <i>ZBTB18</i> , and <i>LMX1A</i> in MB.	163
A.10	Expression profile of <i>CHD5</i> , <i>THRA</i> in MB.	164
A.11	Expression pattern of TFs with highest NIS in subgroups and subclusters.	165
A.12	Expression profiles of <i>ETS1</i> , <i>FLT1</i> , <i>KDR</i> , and <i>TEK</i> in MB.	166
A.13	Expression profile of <i>ZNF540</i> in MB.	167
A.14	Expression profile of <i>ATOH1</i> , <i>SOX2</i> , <i>GLI1</i> , and <i>NFATC1</i> in MB.	168
A.15	Expression profile of <i>SOX9</i> in MB.	169
A.16	Expression profile of <i>NEUROD6</i> in MB.	169
A.17	Expression profile of <i>MYT1</i> in MB.	169
A.18	Expression profile of <i>NEUROD1</i> in MB.	170
A.19	Expression profile of <i>MYC</i> and <i>HLX</i> in MB.	170
A.20	Scatter plots of <i>MYC</i> and <i>HLX</i> expression in ICGC MB samples.	171
A.21	Expression profile of <i>NRL</i> and <i>CRX</i> in MB.	171
A.22	Expression profile of <i>CRX</i> in MB.	171
A.23	Expression profile of <i>RAX2</i> in MB.	172
A.24	Expression profile of <i>EBF1</i> in MB.	172

A.25	Expression profile of <i>TWIST1</i> in MB.	173
A.26	Expression profile of <i>SOX11</i> in MB.	173
A.27	Expression profile of <i>NES</i> and <i>SOX9</i> in MB.	174
A.28	Expression profile of <i>AKT1</i> and <i>PIK3CA</i> in MB.	174
A.29	Expression profile of <i>LHX4</i> in MB.	175
A.30	Aggregated GRN in MB and differentially expressed TFs among subgroups.	176
A.31	Aggregated GRN in MB and differentially expressed TFs among SHH subclusters.	177
A.32	Aggregated GRN in MB and differentially expressed TFs among Group 3 subclusters.	178
A.33	Aggregated GRN in MB and differentially expressed TFs among Group 4 subclusters.	179
A.34	Expression profile of <i>MYC</i> in MB.	180
A.35	Expression profiles of <i>VPS9D1-AS1</i> in MB.	180
A.36	Expression profile of <i>LOXL1-AS1</i> in MB.	181
A.37	Expression profile of <i>DANCR</i> in MB.	181
A.38	Scatter plot of <i>MYC</i> and <i>DANCR</i> expression in ICGC MB samples.	181
A.39	Expression profile of <i>LINC-ROR</i> in MB.	182
A.40	Expression profile of <i>RMST</i> in MB.	182
A.41	Correlation of <i>HOTAIRM1</i> with <i>HOXA</i> genes.	183
A.42	Scatter plot of <i>FEZF1-AS1</i> and <i>FEZF1</i> expression in ICGC MB samples.	184
A.43	Scatter plot of <i>FEZF1-AS1</i> and <i>CDKN1A</i> (P21) expression in ICGC MB samples.	184
A.44	Expression profile of <i>GAS5</i> and <i>ZFAS1</i> in MB.	185
A.45	Scatter plot of <i>ZFAS1</i> and <i>ZEB1</i> expression in ICGC MB samples.	185
A.46	Expression profile of <i>ZEB1</i> in MB.	186
A.47	Scatter plot of <i>ZFAS1</i> and <i>NKD2</i> expression in ICGC MB samples.	186
A.48	Scatter plot of <i>ZFAS1</i> and <i>NKD2</i> expression in ICGC non-WNT MB samples.	186
A.49	Expression profile of <i>ZEB1</i> in MB.	187
A.50	Scatter plot of <i>GAS5</i> and <i>PDCD4</i> expression in ICGC MB samples.	187
A.51	Scatter plot of <i>GAS5</i> and <i>PTEN</i> expression in ICGC MB samples.	187
A.52	Scatter plot of <i>GAS5</i> and <i>CDKN1A</i> expression in ICGC MB samples.	188
A.53	Expression profile of <i>YBX1</i> in MB.	188
A.54	Pairwise correlation and scatter plots of <i>GAS5</i> , <i>ZFAS1</i> , <i>SNHG16</i> , <i>PVT1</i> , and <i>DANCR</i> expression in ICGC MB samples.	189
A.55	Expression profile of <i>MEG3</i> in Cavalli <i>et al.</i> MB cohort.	190
A.56	OS association with <i>MEG3</i> expression in SHH subtypes defined by Cavalli <i>et al.</i>	191
A.57	Expression profile of <i>CDK1</i> in MB.	192
A.58	Expression profile of <i>CCNB1</i> in MB.	192
A.59	Expression profile of <i>MYC</i> in MB.	192
A.60	Scatter plot of <i>MEG3</i> and <i>TGFBR1</i> expression in Cavalli <i>et al.</i> cohort.	193
A.61	Scatter plot of <i>MYC</i> and <i>TGFBR1</i> expression in ICGC Group 3 MB samples.	194
A.62	Scatter plot of <i>MYC</i> and <i>TGFBR1</i> expression in ICGC SHH and Group 4 MB samples.	194

List of Tables

3.1	Confusion matrix.	20
3.2	Performance metrics for binary classification.	21
3.3	Types of lnc genes based on the genomic position relative to coding genes.	36
4.1	Confusion matrix treatment outcome classification.	50
5.1	Clinical features of medulloblastoma subgroups.	58
5.2	Molecular features of medulloblastoma subgroups.	59
5.3	Summary of molecular subtype in MB defined by Cavalli et la.	62
5.4	Functional gene signatures positively associated with <i>MEG3</i> expression.	115
5.5	Cutoffs for the evaluation of measured correlation coefficients.	123
A.1	Performance of the SVM-based OT mini-classifier in individual external cohorts.	158
A.2	Performance of the SVM-based OT mini-classifier in merged external cohorts.	158

List of abbreviations

5-FU 5-fluorouracil	FANTOM CAT FANTOM CAGE Associated Transcriptome
A adenine	FC fold change
BH Benjamini and Hochberg	FDR false discovery rate
bp base pairs	FN false negative
BS binding site	FP false positive
C cytosine	FOLFOX FOLinic acid, Fluorouracil, Oxaliplatin
CAGE cap analysis of gene expression	FPKM fragments per kilobase of exon per million mapped reads
CB cerebellum	G guanine
ceRNA competing endogenous RNA	GABA gamma-butyric acid
cDNA complementary DNA	GENIE3 Gene Network Inference with Ensemble of trees
CI confidence interval	GLM generalised linear model
CLICK CLuster Identification via Connectivity Kernels	GO gene ontology
ChIP-seq chromatin immunoprecipitation sequencing	GRN gene regulatory network
ChOP chromatin oligo affinity precipitation	GRSBC Gene Regulation & System Biology of Cancer
CNV copy number variations	H&E hematoxylin and eosin
CRC colorectal carcinoma	HR hazard ratio
DGE differential gene expression	ICGC International Cancer Genome Consortium
DGEA differential gene expression analysis	indel insertion/deletion
DMR differentially methylated region	IG-DMR intergenic differentially methylated region
DNA deoxyribonucleic acid	IGL inner granular layer
duplex double helix	ISH <i>in situ</i> hybridisation
EGL extra granular layer	KEGG Kyoto Encyclopedia of Genes and Genomes
EMT epithelial-to-mesenchymal transition	KF Khambata-Ford
enh-GRN enhancer-mediated GRN	LCA large cell/anaplastic
expr-GRN expression-inferred GRN	lnc long non-coding
FANTOM Functional ANnotation Of the Mammalian genome	

List of Tables

lncRNA long non-coding RNA	RNA-seq RNA sequencing
LR likelihood ratio	RPKM reads per kilobase of exon per million mapped reads
LRT likelihood ratio test	rRNA ribosomal RNA
MAPK mitogen-activated protein kinase	SD stable disease
MB medulloblastoma	Shh Sonic Hedgehog
MBEN medulloblastoma with extensive nodularity	SMOTE Synthetic Minority Oversampling Technique
ML machine learning	SNV single nucleotide variation
miRNA microRNA	SVM support vector machine
mRNA messenger RNA	SVM-RFE SVM recursive feature elimination
MSVM-RFE multiple SVM recursive feature elimination	T thymine
NB negative binomial distribution	TAD topologically associating domains
NGS next-generation sequencing	T/C treated/control
NIS network influence score	TF transcription factory
NI-score network influence score	TFBS TF binding sites
NMF Non-negative matrix factorisation	TFO triplex-forming oligonucleotides
NV Novartis	TMM trimmed mean of M values
ORF open reading frame	TN true negative
OS overall survival	TP true positive
OT OncoTrack	triplex triple helix
PAD percentile absolute deviation	TSS transcription start site
PCR polymerase chain reaction	TTS triplex target site
PDO patient-derived organoids	U uracil
PDX patient-derived xenografts	UTR untranslated regions
PID Pathway Interaction Database	VZ ventricular zone
postn postnatal	WES whole exome sequencing
pren. prenatal	WGS whole genome sequencing
PRC2 polycomb repressive complex 2	Wnt Wntless and Int-1
PWM position weight matrices	WT wild-type
RNA ribonucleic acid	
RNApol RNA polymerase	

1 Introduction

Scientific research and medical description of cancer stretch back many thousands of years [1]. Written in Egypt on papyrus, the first medical record of cancer dates back to 3000 BC. Back then, the writer of this ancient text already recognised the examined tumour, a breast cancer, as a severe disease [2]. The first path-breaking scientific description of cancer was published by Johannes Müller in 1838. In a treatise about microscopic studies of tumour tissues, he was the first to describe that a tumour consists of a group of abnormal cells. This milestone paved the way for understanding cancer as a disease of cells. Müller also associated cancer with ageing, which is well approved today [1, 3]. With a more and more extended life expectancy, the risk of developing cancer is increasing dramatically [1]. Thus, cancer has become a leading cause of death together with cardiovascular diseases in the Western world [4].

The central aspect of current cancer research is the molecular analysis of tumours. Molecular studies are essential to understand this disease since all cellular processes are implemented in networks of molecule interactions involving DNA, RNA, and proteins. Gene expression, for instance, is regulated by such networks, while the regulation of gene expression controls most cellular processes [5]. Cancer develops due to dysregulation of gene expression and cellular programs caused by various mechanisms altering homeostasis in an initially normal cell [4].

High-throughput omics technologies and the computational analysis of omics data have been key to molecular studies of cancer [6]. Omics technologies allow the generation of molecular data on a large scale, such as the whole transcriptome of tumours. Next-generation sequencing is an important platform to investigate the whole transcriptome of cells by sequencing RNA (RNA-seq), which enables the measurement of gene expression levels. Therefore, RNA-seq is an important tool to study which cellular processes are dysregulated in cancer cells due to abnormal gene expression. However, large scale omics data require computational analyses to derive meaningful information [6]. These computational analyses mainly involve the application of statistical and machine learning methods. Especially machine learning can solve complex data analysis tasks and, therefore, has become an important tool. Even though the application of omics technologies and computational analyses has provided many new insights for cancer research, their introduction has led to even more molecular biological and clinical questions in cancer research.

1.1 Thesis outline

The primary scope of this thesis is the computational analysis of cancer transcriptomes, investigating two cancer types, colorectal cancer (CRC) and medulloblastoma (MB).

The first part of this thesis (Chapter 2) gives an overview of the molecular biology of cancer and surveys omics tools widely used in mole cancer research. An introduction of the hallmarks of cancer summarises the capabilities that are acquired by a tumour during cancer development. Additionally, long non-coding (lnc) genes and their lncRNA products are introduced as important regulators of cellular and cancer-associated processes.

The first part of this thesis (Chapter 2) gives an overview of the molecular biology of cancer and omics tools widely used in mole cancer research. The process of cancer development will be explained by introducing (1) cancer as a malignant disease, (2) gene expression and its regulation in the cell, and (3) mechanisms that lead to the cancer-causing dysregulation of gene expression in tumours. An introduction of the hallmarks of cancer summarises the capabilities that are acquired by a tumour during cancer development. Additionally, long non-coding (lnc) genes and their lncRNA products are introduced as important regulators of cellular and cancer-associated processes. The first part

1 Introduction

closes with an overview of omics technologies focusing on RNA-seq and advances that were gained in cancer research by applying these technologies. These advances include the discovery of the molecular heterogeneity within a single cancer type and the identification of prognostic and predictive biomarkers.

The second part of this thesis (Chapter 3) provides an introduction to the computational analysis methods used in this thesis including an outline of machine learning and supervised learning. Besides standard analyses in cancer research, three analysis tasks and their solutions via machine learning and statistical methods are introduced. The first task is the construction of a gene-expression-based classifier that predicts the response to therapy. Here, support vector machines (SVMs), a supervised machine learning method, offer a solution for this task. The second task comprises the inference of gene regulatory networks (GRNs) from gene expression data. GRNs depict the interactions between transcription factors (TFs), which function as regulators in the network, and other genes. Such a GRN can be obtained by applying the algorithm GENIE3 that is based on a random forest of regression trees (an ensemble machine learning method). The last task is the computational characterisation of lnc genes using two approaches. The first approach is the classification of lnc genes based on their position relative to protein-coding genes in the genome. The second approach comprises the analysis of expression profiles of lnc genes to infer their putative function.

The third part of this thesis (Chapter 4) presents a transcriptome study of CRC. Since only a fraction of CRCs respond to available chemo- and targeted therapies, this study aims to construct a classifier that predicts therapy outcome. In clinical practice, *RAS/RAF* mutations are used as biomarkers to predict the resistance of CRC to EGFR-targeting therapy. However, non-mutant tumours still show a wide range of treatment outcomes for EGFR-targeting therapy, highlighting the need for novel biomarkers. Addressing this question within the OncoTrack project, an SVM was used to build a gene-expression-based classifier that predicts response to anti-EGFR therapy.

The fourth part (Chapter 5) presents a cancer transcriptome study of MB. This study focuses on a deeper analysis of the molecular heterogeneity in MB by interfering GRNs underlying the heterogeneity of MB. Additionally, a computation characterisation of lnc genes that are related to the molecular heterogeneity of MB is performed since implications of lnc genes in MB are mostly unknown.

The last part of this thesis (Chapter 6) summarises the results and gained insights of the carried out studies in the light of current literature.

2 Cancer and gene expression

The following chapter will introduce cancer as a malignant disease, the genome and gene expression, cancer development via the dysregulation of gene expression, and omics technologies for the molecular analysis of tumours with an emphasis on RNA-seq.

2.1 Cancer - malignant neoplasm

The human body is a highly complex system of cells that develops from one single fertilised egg cell called a zygote [7]. The development from a single zygote to a complete human organism (as in other vertebrates) is determined by two essential processes. The first process is cell proliferation, which describes the cyclically repeated process (called cell cycle) of cell growth and cell division into two daughter cells leading to a multiplication of the cell number. The second process is cell differentiation, which is the sequential development from the undifferentiated zygote towards fully differentiated (functionally specialised) cell types that form the tissues and organs of the body. Even though both processes partially happen quite rapidly during development, they are always tightly regulated because the number of cells and cell types is comparable between individuals at a certain developmental stage [7, 8]. Proceeding cell differentiation is accompanied by a reduction in proliferation rate. Therefore, cells of the embryo usually show a higher proliferation rate than cells of the adult human body [8]. Overall, the number of cells is kept constant in the adult body — a state of homeostasis — via a careful balance between cell proliferation and loss of cells due to programmed cell death (apoptosis) or injuries [8, 9], illustrating cell proliferation control of as an essential aspect to sustain a healthy human body.

Cancer is a malignant neoplastic disease [4, 10] that evades the essential control of cell proliferation. Neoplasia describes an abnormal growth of cells resulting in a cell aggregation called a neoplasm or tumour [11, 12]. Based on the aggressiveness of growth, tumours are classified into benign or malignant. Benign tumours grow locally and do not invade surrounding normal tissue [12]. The more aggressive malignant tumours invade adjacent tissue, spread out, and form colonies of tumour cells at secondary sites distant and physically not connected to the primary tumour. These secondary sites are called metastases [12, 13]. This aggressive growth leads to damage of the invaded tissue by primary tumours or metastases [14]. Cancer always relates to malignant tumours, and the formation of metastases causes approximately 90% of cancer-related deaths [12].

The adult body consists of over 200 cell types [15], and cancer can develop from many different (pre- and postnatal) cell types. Thus, the term cancer comprises a vast collection of malignant neoplastic diseases. Most malignant tumours can be classified into four broad cancer categories based on the cell type of origin. (1) Carcinomas are tumours originating from epithelial cells and account for the majority of tumours, including breast and colon cancer. (2) Tumours that derive from cells types forming connective tissue are called sarcoma. (3) Leukaemia and lymphomas develop from haematopoietic (blood) cell types; leukaemia cells move freely in the bloodstream and do not form solid tumours. (4) Tumours originating from different parts of the nervous system belong to the category of neuroectodermal tumours, e.g. glioma and medulloblastoma. There are also tumour types which are not fitting into broad categories like melanomas. These derive from melanocytes, pigment cells of the skin and eye with a developmental origin close to neuroectodermal cells [12].

The central question is, which mechanisms can cause the transformation of normal cells into a malignant tumour in such a vast amount of cell types.

2.2 The genome and gene expression regulation

All biological processes that simultaneously happen in an organism and during its development are implemented in molecular biological networks defined by a system of molecule interactions. These networks describe the flow of matter and energy (metabolism) or information (signalling, regulation) [5]. The flow of genetic information is a central aspect in regulating the cells biological processes [5, 16].

Each cell carries deoxyribonucleic acid (DNA) that comprises the heritable genetic information. The DNA carries the genetic information and passes this information to the next generation due to its unique structure. The DNA is a macromolecule consisting of linear connected nucleotides (polymer). Each nucleotide is composed of a sugar (deoxyribose), a phosphate group, and a base. The phosphate group and sugar form the backbone of the DNA, where the phosphate group links the sugars of two nucleotides. Here, the phosphate group is linked to the 5' carbon atom of the first sugar and the 3' carbon atom of the second sugar. This linkage defines the direction of the DNA from 5' to 3'. The base, the third component, is linked to the sugar. There are four different bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Among others, these bases are important for the structure of the DNA. The DNA is present as a double-stranded molecule that forms a double helix (duplex) in the cell. These two strands are held together by hydrogen bonds formed between the base pairs A-T and G-C (Watson-Crick base pairing), where each base pair is split between strands. The two strands are reverse (opposite direction of strands) and complementary. This double-stranded complementary structure allows the generation of two identical copies of the DNA during DNA replication, where one strand is used as a template to synthesis the reverse strand. The DNA is replicated during cell proliferation, and each of the daughter cells receives one copy of the DNA to pass the genetic information to the next cell generation. Therefore, all successors of a zygote carry the same DNA/genetic information. The human DNA comprises approximately 3 billion base pairs organised in 23 pairs of macromolecules called chromosomes that comprise 22 autosomal and two sex chromosomes. The double (diploid) chromosome set arises from the inheritance of genetic information from both parents. The collection of all chromosomes defines the human genome that is located within the cell nucleus [16].

The bases facilitate not only the double-helix structure of the DNA. Unique linear combinations of the four bases along the DNA strands form sequence information that store the genetic information, each strand carrying differing information. The sequence information on the DNA is transcribed into ribonucleic acid (RNA) sequences, and the RNA is translated into protein sequences. The sequence information on the DNA carries templates for the functional molecules, proteins that fulfil almost all biological processes within a cell. The process of sequence transcription and translation represents the flow of genetic/sequence information in a cell. The two key steps in this process are called transcription and translation due to different building blocks of DNA, RNA, and protein sequences. RNA is similar to DNA despite few differences: RNA has ribose sugar, the base T is substituted by the base uracil (U), and RNA is single-stranded. Due to the similarity between DNA and RNA, the transfer of sequence information from DNA to RNA is a simple sequence transcription (rewriting). An RNA molecule is also called a transcript [16]. Proteins are macromolecules consisting of an amino acid chain that folds into a three-dimensional structure. The folding is essential for the molecular function of a protein. There are 20 different kinds of amino acids, and unique amino acid sequences relate to distinct proteins. Due to the higher number of amino acids compared to nucleotides (20 vs. 4), each amino acid is encoded by a nucleotide triplet called a codon. The codon facilitates the translation of nucleotide-based information (RNA) to amino acid-based information. A sequence of codons that encodes a protein is termed an open reading frame (ORF). RNA that carries protein information is referred to as messenger RNA (mRNA). Ribosomes (organelles of the cell) synthesise proteins by translating mRNA to proteins. These organelles are located in the cell cytoplasm, the space between the nucleus and cell membrane. [17].

However, only a fraction of the DNA relates to sequences that are transcribed and protein-coding. These protein-coding and transcribed DNA sequences are organised in single units (loci) scattered

along the genome. These loci relate to protein-coding genes. Protein-coding genes have a certain structure defined by alternating nucleotide blocks of exons and introns, where exons carry the protein-coding sequence. Both exons and introns are transcribed into RNA, but only exons remain in the mature mRNA because introns are removed from the RNA molecule by splicing. A mature mRNA consists of the (1) ORF, (2) untranslated regions (UTR) at the 5' and 3' end, (3) a 5' cap that protects the mRNA from degradation, and (4) a tail of adenosine monophosphate nucleotides (poly-A tail) that is important for mRNA stability and translation [16, 17]. Besides protein-coding genes, there are also non-coding genes: transcribed DNA loci that contain only RNA sequence information as an end product. Genes that transcribe into long non-coding (lnc) RNAs — a species of non-coding RNAs defined by a length of > 200 nucleotides — have a gene structure of exons/introns and are spliced like protein-coding genes. Non-coding RNAs also have various functions in biological processes, but especially lncRNAs and their genes are less well studied than proteins because the role of lncRNA has just been revealed recently [18].

A general definition of a gene is a DNA locus that is transcribed and carries sequence information for a functional product (RNA molecule or protein) [19]. The human genome contains over 20,000 protein-coding genes and an estimated number of ~60,000 lnc genes [20]. There are many known genes that carry sequence information for functional lncRNAs. However, the extent of how many genes that transcribe lncRNAs also produce a functional lncRNA is unknown because it is assumed that the transcription process itself and not the lncRNA can have a functional aspect [20]. Even though the mentioned definition of a gene includes that a gene carries sequence information for a functional product, all transcribed DNA loci that produce lncRNA are considered as a "lnc gene" in this thesis. Since human cells are diploid, a cell carries two copies of a gene — termed as two alleles — while each allele is paternally or maternally inherited.

The flow of genetic information from a gene to a functional molecule is called gene expression [21]. The regulation of gene expression is a central mechanism coordinating the biological processes of the cell [5, 16]. Here, the number of RNA or protein molecules that derived from a gene within a cell reflects the expression level of the gene. Changes in gene expression directly impact the transcriptome (the collection of transcripts existing within a cell) and proteome (the collection of proteins existing within a cell). Therefore, the regulation of gene expression controls which functional molecules are present within a cell and how many of them.

The most important determinant of gene expression is gene transcription, which is controlled by different mechanisms. One mechanism relates to the interaction between regulatory DNA sequences and transcription factors (TFs), a specialised type of proteins binding to these DNA sequences. TFs recognise and bind to regulatory sequences and form complexes with co-factor proteins or other TFs in order to control the initiation and activity of transcription that is performed by the RNA polymerase (RNAPol) enzyme. Here, TFs initiate transcription by guiding RNAPol to the transcription start site (TSS) of a gene. However, besides transcriptional activation, TFs can also function as transcriptional repressors. There are different families of TFs, and each family binds to a specific sequence motif, which allows the regulation of individual genes by different TFs [8]. Since TFs usually regulate the expression of genes located on a different chromosome than the gene encoding the TF, transcriptional control by TFs is referred to as regulation in *trans*. There are two main types of DNA regions that contain regulatory sequences bound by TFs: promoters and enhancers. These DNA regions control RNAPol-conducted transcription in *cis* (on the same chromosome). Promoters are essential for transcription initiation and are generally located in DNA regions upstream and at the TSS of a gene, where a promoter can extend up to several kilo-bases upstream of the TSS [8, 22]. Each gene has at least one promoter and one TSS. In contrast to promoters, enhancers are distant from a TSS. Due to the formation of DNA loops, TFs bound to an enhancer can form contact and interact with TFs/co-factors bound to a promoter to increase (enhance) transcription. Enhancers are frequently involved in cell-type-specific gene expression and can regulate the expression of more than one gene because of a dynamic formation of DNA loops and [8].

2 Cancer and gene expression

An additional mechanism of transcriptional regulation relates to epigenetics, involving the biochemical modification of DNA [8, 23]. DNA methylation represents the modification of the DNA by adding a methyl group to position five of a cytosine at CG dinucleotides (also CpG). DNA methylation at promoters and enhancers is generally associated with repressed transcription. However, more complex processes participate in epigenetic regulation. For example, DNA methylation at TF binding sites (TFBS) can block the binding for a subset of TFs, whereas other TFs are still capable of binding to methylated DNA. Additionally, DNA methylation is not necessarily the cause of the transcriptional repression and could just be associated with the repressive state [23]. The complex relationship between DNA methylation and transcription goes beyond the scope of this thesis and, therefore, is not further discussed (for further reading, please see [23]).

The spatial organisation of the DNA within the nucleus also strongly influences gene transcription. DNA does not exist as a long outstretch molecule inside the nucleus but in a spatially condensed form called chromatin, which is a complex of proteins and DNA. The smallest structural unit of chromatin is the nucleosome. A nucleosome consists of an octamer of the four core histones H2A, H2B, H3, and H4 as well as DNA wrapped around the histone octamer. In a next condensation step, several nucleosomes form a chromatin fibre. Further packing of the fibres results in highly condensed chromatin [8]. The compact organisation of DNA as chromatin restricts the access of TFs to the DNA. Therefore, regulatory DNA sequences that are accessible for TFs are located in nucleosome-free regions. The dynamic change of the DNA packing and accessibility modulates transcriptional regulation [24]. Additionally, specific transcriptional states are associated with specific biochemical modifications of histone proteins. For example, active transcription at promoter or enhancer regions is marked by H3K27ac (H3 - histone 3, K - amino acid lysine, 27 - position of K in the amino acid chain of H3, ac - acetylation of K) [25]. Biochemical modifications of the DNA (methylation) and histones that do not change the genome sequence and can be inherited define the epigenome. The constitution of the epigenome is different between cell types, unlike the genome sequence which is identical [26]. Proteins that can write, erase, or read epigenetic modifications are important transcription regulators besides TFs [25]. The organisation of chromatin into separated domains influences transcription as well. These topologically associating domains (TADs) are chromatin loops that are wider and not dynamic in contrast to chromatin loops between enhancers and promoters. TADs are one determinant of transcription because enhancer-promoter interactions generally happen only within a TAD and not across TAD boundaries [27].

Taken together, one central mechanism for the regulation of gene expression — the flow of genetic information — is the control of transcription via the binding of TFs to regulatory sequences in promoter and enhancer regions. The regulation of gene expression is a major aspect of controlling biological processes. Gene expression regulation also explains [28, 29] how the successor cells of a single zygote can differentiate and manifest certain traits (phenotypes) in terms of cell types (cell phenotypes) even though all cells carry the same genome (genotype) (a phenotype can also relate to the physical properties of an organism). Therefore, individual cell types relate to a specific gene expression profile/signature controlled by an underlying gene regulatory network (GRN) [29]. GRNs describe the interaction networks between TFs and their target genes via the binding to the promoters/enhancers of the target [5]. In humans, the number of different GRNs is at least the number of the over 200 cell types [15] highlighting the complexity of GRNs.

Protein-protein interactions in the nucleus, in the cytosol, or within the membrane define a large fraction of cellular activities. These protein-protein interactions relate to the flow of matter and energy, and information in terms of signals, which comprise intra- and extra-cellular signals. [5]. One aspect of the flow of signals is the transduction of extracellular signals. Here, extracellular signal molecules are recognised by transmembrane receptor proteins that propagate the signal into the cell by activating transducer or effector proteins. In the case of transducer activation, the transducer activates the effector. The effector protein leads to the final response to the signal. This cascade of signal propagation via protein-protein interaction is termed a signalling pathway. For example, the activation of a pathway

can result in the expression of genes that are positive regulators of cell proliferation, which is especially important during development. These pathways play an essential role in cell-to-cell communication and allow the coordination of processes between cells [17].

2.3 Genetic and epigenetic alterations in tumours

Regulation of gene expression and signalling operates within a pre-defined genetic program that ensures normal development and homeostasis of the adult body. Cancer is characterised by the dysregulation of this program — breaking the normal cell differentiation and proliferation — due to genetic and epigenetic alterations (changes of the genome and epigenome that can be inherited). Direct or cascade effects of these alterations dysregulate gene expression or other cellular processes. Cancer development is considered a multistep process where a tumour accumulates alterations. It is generally assumed that at least two alterations (hits) are necessary to develop cancer. In the classical view, a tumour arises from a single cell hit by a genetic or epigenetic tumour-initiating alteration. This initiating alteration provides the first trigger and a context for a cell to become a tumour. Cancer development involves clonal expansion where the initial alteration is passed to the following (tumour) cell generation [4]. During tumour promotion/progression, the tumour cell population expands and acquires additional alterations. Furthermore, tumour promotion/progression involves an evolutionary selection of tumour cells based on acquired tumour-driving alterations that have a selective advantage defined by alteration effects increasing the capability of proliferation, survival, invasion, or metastasis [4, 8]. The process of clonal evolution can lead to a heterogeneous collection of clonal populations within a tumour that is referred to as intratumour heterogeneity.

Cancer-causing alterations directly change the expression or the activity/functionality of a gene product. The potential effects of genetic alteration have been mostly described for protein-coding genes, as summarised in the following paragraph. Genetic alterations may consist of (1) point mutations (exchange of a single base - single nucleotide variants (SNVs), (2) insertions or deletions of few nucleotides (indel), (3) copy number variations (CNVs) (loss/deletion or gain/amplification of gene copy numbers), and (4) chromosomal aberrations/rearrangement (translocation, insertion, duplication, deletion, inversion of chromosome segments; includes CNVs) [4]. Here, the term mutation is mostly used for SNVs and indels. There are three possible consequences on the translated amino acid sequence when an SNV hits the ORF of a protein-coding gene: (1) missense (the encoded amino acid of a codon is changed), (2) nonsense (a stop codon is introduced resulting in a truncated protein), and (3) synonymous (the base substitution does not change the amino acid sequence)[12]. The consequence of indels is more complex. If the number of nucleotides of an indel is not evenly divisible by three (the codon length), the indel will introduce a frame-shift of the original ORF and change all amino acids following the indel [30]. Indels can also preserve the frame but introduce instead a small deletion or insertion which alter the protein structure and activity. Missense SNVs, nonsense SNVs, and indels can cause a change of protein activity/functionally. These changes can result either in a loss-of-function (loss of activity) or gain-of-function (change or enhancement of activity) of the corresponding gene protein product, where downstream effects dysregulate gene expression and cellular programs [4]. Outside of coding regions, mutations can also affect the activity of promoters and enhancers by modifying TFBSs for instance [12]. CNVs have a direct influence on gene expression. Copy number gain/amplification can lead to an elevated gene expression, whereas deletion of one allele (hemizygous deletion) can cause a reduction of gene expression which might lead to a haploinsufficiency (gene products of one allele are not sufficient to fulfil the original function of a gene). Deletion of both alleles of a gene (homozygous deletion) leads to the complete loss of a gene. Epigenetic alteration comprises, among others, the hypermethylation of promoters resulting in the inhibition (silencing) of transcription [4].

Tumour-initiating genetic alterations mainly relate to SNVs, indels, rearrangements, deletions, or amplifications that hit a gene. Broader chromosomal alterations occur more frequently later in

tumour development due to an increasing genetic/chromosomal instability during tumour progression. Cancer initiating alterations are generally somatic, meaning that they have occurred spontaneously in somatic cells (any cell excluding germ cells ,egg and sperm)). More rarely, cancer initiating mutation can be inherited (germline mutation). Germline mutations represent a predisposing factor for specific cancer types. Somatic mutations can be caused either by exposure to exogenous carcinogens including environmental agents damaging the genome/epigenome like UV light, cigarette smoke, radiations and industrial pollution or endogenous factors from inside an organism/cell. Endogenous factors include oxygen radicals that are produced by normal metabolism and errors in DNA replication. Since these endogenous factors are part of daily cell life, mechanisms like DNA repair are normally in place to protect the cell from these factors so that failure to repair DNA insults might lead to cancer [4].

However, the probability that a tumour-initiating mutation has hit a cell increases with lifetime due to the constant metabolism of the organism and cell renewal in many tissues involving DNA replication. Additionally, since at least two and most of the time several genetic/chromosomal mutations are necessary to develop cancer, the time between the first event and the fully developed cancer can span years up to several decades. Therefore, cancer is often an age-related disease. The average age of patients diagnosed with cancer is 65 years [4]. Nevertheless, cancer can also occur in infants and children (paediatric cancers). There is evidence that several paediatric cancers arise from embryonic or early postnatal cells. In general, tumour cells display some features of embryonic cells, including a less differentiated phenotype, high proliferation rate, and the capacity to invade other tissues. Therefore, a shorter development time is associated with tumours of embryonic/early postnatal origin because the cell-of-origin is already closer to a cancerous phenotype than most highly differentiated cells of the adult body [31]. A more extended time-lapse in the development of adult cancer is reflected by a higher number of genetic mutations compared to paediatric cancers [32].

Tumour-initiating/-driving genetic alterations generally hit two types of genes. These types include (1) proto-oncogenes, which become an oncogene and promote tumorigenesis after activation by an alteration, or (2) tumour suppressor genes, which are deactivated [4]. The positive regulation of proliferation or survival is a frequent function of proto-oncogenes in a normal cell. The change of proto-oncogenes to oncogenes by an activating alteration causes only an enhancement of the normal protein function. Therefore, oncogene proteins frequently promote the uncontrolled proliferation of tumour cells but can also cause abnormal differentiation or prevent apoptosis. Most oncogene proteins are part of a signalling pathway and different pathway components: growth factors (extracellular signal proteins), growth factor receptors, intracellular proteins of the signalling cascade (e.g. transducer, effectors), and TFs. The response to oncogenic pathway activation is the dysregulation of gene expression and cellular processes. Activation of oncogene proteins might be triggered by an increased expression of the gene (e.g. CNV amplification) or part of the gene containing the activated domain (e.g. case of gene fusions), or a change of the amino acid sequence (e.g. SNV, indel). In contrast to proto-oncogenes, tumour suppressor genes are negative regulators of proliferation and survival. Tumour suppressors function as inhibitors of pathways involving proto-oncogenes or as a regulator of cell cycle progression and apoptosis [8]. Inactivation of tumour suppressors can be caused, for example, by homozygous and hemizygous deletions, SNVs, indels, and promoter hypermethylation [8, 12].

2.4 Hallmarks of cancer

The model of tumour initiation and promotion/progression can be extended by the capabilities acquired by a malignant tumour during cancer development, as proposed by Hanahan and Weinberg [10]. These capabilities include eight "hallmarks of cancer":

- sustaining proliferation signalling,
- evading growth suppressors,
- avoiding immune-induced destruction (immune cells can recognise and destroy tumour cells),

- enabling replicative immortality (the number of replications is limited in somatic, differentiated cells),
- invasion and metastasis formation,
- inducing angiogenesis (inducing growth of blood vessels to supply the tumour with nutrients and oxygen),
- resisting cell death, and
- reprogramming energy metabolism (uncontrolled proliferation demands an adjustment of energy metabolism).

These hallmarks illustrate that cancer development is more than a tumour-cell-intrinsic process. It also involves interactions with the microenvironment defined by the system of tumour-surrounding, non-cancerous cells (such as immune cells) and tissue. Some of these hallmarks are outlined below.

The hallmark "sustaining proliferation signalling" involves the alterations of tumour suppressor genes and proto-oncogenes. Here, tumour suppressor genes and proto-oncogenes are part of different pathways that positively regulate proliferation or other cell behaviour [10]. Among these tumour suppressor genes are *adenomatous polyposis coli* (*APC*) and *patched 1* (*PTCH1*). *APC* is a negative regulator of the Wnt (Wingless and Int-1) signalling pathway by promoting the degradation of Catenin Beta 1 (*CTNNB1*), the downstream target of the Wnt pathway. *CTNNB1* forms a complex with transcription factors to regulate gene expression and promote proliferation. *PTCH1* is a receptor protein and negative regulator of Shh (Sonic Hedgehog) signalling by binding the receptor protein Smoothed (SMO). SMO propagates Shh signalling into the cell leading to cell proliferation via transcriptional regulation [12]. Mutated proto-oncogenes that sustain proliferation comprise, among others, *RAS* and *RAF* genes (like *KRAS*, *NRAS*, and *BRAF*) that propagate intracellular signalling downstream of receptor tyrosine kinases, a family of transmembrane receptors [12].

The hallmark "enabling replicative immortality" refers to the removal of DNA replication limits in tumour cells. These limits are defined by a limited number of replications in a normal cell. Here, the number of replications is controlled via telomeres that protect the ends of chromosomes. Telomeres are DNA sequences composed of specific tandem repeats and shorten over DNA replications. The erosion of telomeres triggers apoptosis or cellular senescence (a state where a cell stops proliferating but remains viable). However, the telomeric DNA can be elongated and maintained via the telomerase complex. Expression of the telomerase subunit protein Telomerase Reverse Transcriptase (*TERT*) leads to maintenance of telomeric DNA in cancer, creating immortalised tumour cells. Under normal conditions, *TERT* is expressed in embryonic cells and somatic cells of renewing tissues [10, 33].

Malignant tumour cells acquire the capability of "invasion and metastasis formation" by undergoing Epithelial-to-Mesenchymal Transition (EMT). Here, epithelial cells undergo a phenotypic change and acquire characteristics of mesenchymal cells. Epithelial cells form adherent junctions and are restricted in motility, whereas mesenchymal cells show high motility. EMT usually takes place during embryogenesis and wound healing. The phenotypic change of EMT is regulated by TFs, such as *TWIST1*, that regulate the migration of cells during embryogenesis. The dysregulation of these TFs promotes EMT in cancer, resulting in tumour cell invasion and metastasis formation [10].

Alterations of the tumour suppressor gene *tumor protein p53* (*TP53*) relates to the hallmarks "resisting cell death" and "avoiding growth suppressors". *TP53* responds to certain signals within a cell. These signals include DNA damage or suboptimal conditions of nucleotide pool levels, growth-promoting signals, glucose, and oxygen concentration. In response to these signals, *TP53* induces a cell cycle arrest, which allows the cell to repair DNA damage or normalise suboptimal conditions, or activates apoptosis. A central role of *TP53* as a tumour suppressor gene is underlined by frequent alteration of the gene across many different cancer types [10, 12].

In summary, the hallmarks of cancer represent properties acquired by the tumour, which collectively drive the evolutionary development of cancer. Additional aspects of the hallmarks of cancer discussed in the original publication of Hanahan and Weinberg are beyond the scope of the presented thesis.

2.5 Long non-coding genes: Functions and roles in cancer

Studying the effects of genetic and chromosomal alterations as well as gene expression dysregulation on protein-coding genes in cancer has been the primary source for understanding this disease. However, lnc genes have emerged as a widespread regulator of biological processes, including processes involved in cancer development in the past couple of years [34]. Lnc genes and their RNA products (lncRNA) are less well understood and less studied than coding genes and proteins. There are several reasons for this, as the lncRNA research is relatively young and a lower number of tools is available for studying molecular functions and mechanisms of lnc genes/-RNA [35]. Nevertheless, many aspects of lnc genes and lncRNA have been revealed by now.

The lower sequence conservation of lnc genes across species is a characteristic difference between lnc and protein-coding genes. However, the low sequence conservation of lnc genes does not automatically imply lacking functions. Lnc genes are conserved on the level of syntenic regions (conserved co-localisation of genes within a genomic region), short sequences, or secondary structure of the lncRNA product. Especially the conservation of the secondary structure of lncRNAs appears to be critical. Here, the structure is less dependent on the sequence conservation when compared to protein-coding genes (due to the missing step of translation in lncRNA biogenesis). Since protein sequences are encoded via codons, the ORF of a protein-coding gene needs to be conserved during evolution to maintain a protein's sequence, expression, and function [35].

Like proteins, lncRNAs fulfil their functions via interacting with other molecules including protein, RNA, and DNA/chromatin. Here, lncRNAs regulate processes on transcriptional, post-transcriptional, translational, or signalling levels [34, 38]. For example, lncRNAs act as competing endogenous RNA (ceRNA) by functioning as a "sponge" for a type of small RNA called a microRNA (miRNA) [34]. miRNAs negatively regulate mRNA expression on the post-transcriptional level via binding to the 3' UTR of mRNAs and inducing mRNA degradation [39]. By sponging miRNAs, ceRNAs act as an antagonist of miRNA-induced negative regulation of mRNA expression. Apart from the indirect regulation of mRNA expression, lncRNA can stabilise mRNA on a post-transcriptional level or regulate translation efficiency via direct lncRNA-mRNA interactions [34]. Through lncRNA-protein interactions, lncRNA can also stabilise protein expression on a post-translational level [40].

lncRNAs can directly interact with double-stranded DNA by forming an RNA:DNA:DNA triple helix (triplex) (Figure 2.1.a) [36, 37]. The triplex formation is facilitated by forming Hoogsteen hydrogen bonds between bases of the third nucleotide strand (RNA/DNA) and the DNA duplex. Here, only one strand of the duplex interacts with the third strand. The Hoogsteen hydrogen bonds are determined by steric constraints and available hydrogen donor and acceptor groups. These determinants limit triplex formation to certain base pairs (Figure 2.1.b) and three sequence motifs on the third nucleotide strand (Figure 2.1.c). Both base pairing and motifs depend on the orientation of the third strand relative to the duplex. Parallel and antiparallel orientations refer to forward and reverse Hoogsteen base pairing, respectively. While purine (G and A) and pyrimidine bases (U and C for RNA; T and C for DNA) of the third nucleotide strand can form hydrogen bonds with the duplex, only purine bases of the DNA duplex are involved in the base pairing. In the case of RNA-DNA:DNA triplexes (as shown in Figure 2.1.b), forward Hoogsteen base pairing includes the nucleotide triads C⁺-GC, U-AT, and G-GC, and reverse Hoogsteen base pairing includes the triads A-AT, U-AT and, G-GC. The triad C⁺-GC is only formed when the cytosine of the RNA strand is protonated, which requires an acid condition. Due to this special condition, it is unclear whether the C⁺-GC triad exists under physiological conditions. The Hoogsteen base pairs build triplexes via three RNA sequence motifs:

- pyrimidine/UC motif (rich of U and C, only forward Hoogsteen base pairing),
- purine/GA motif (rich of G and A, only reverse Hoogsteen base pairing), and
- purine-pyrimidine/GU motif (rich of G and U, forward or reverse configuration).

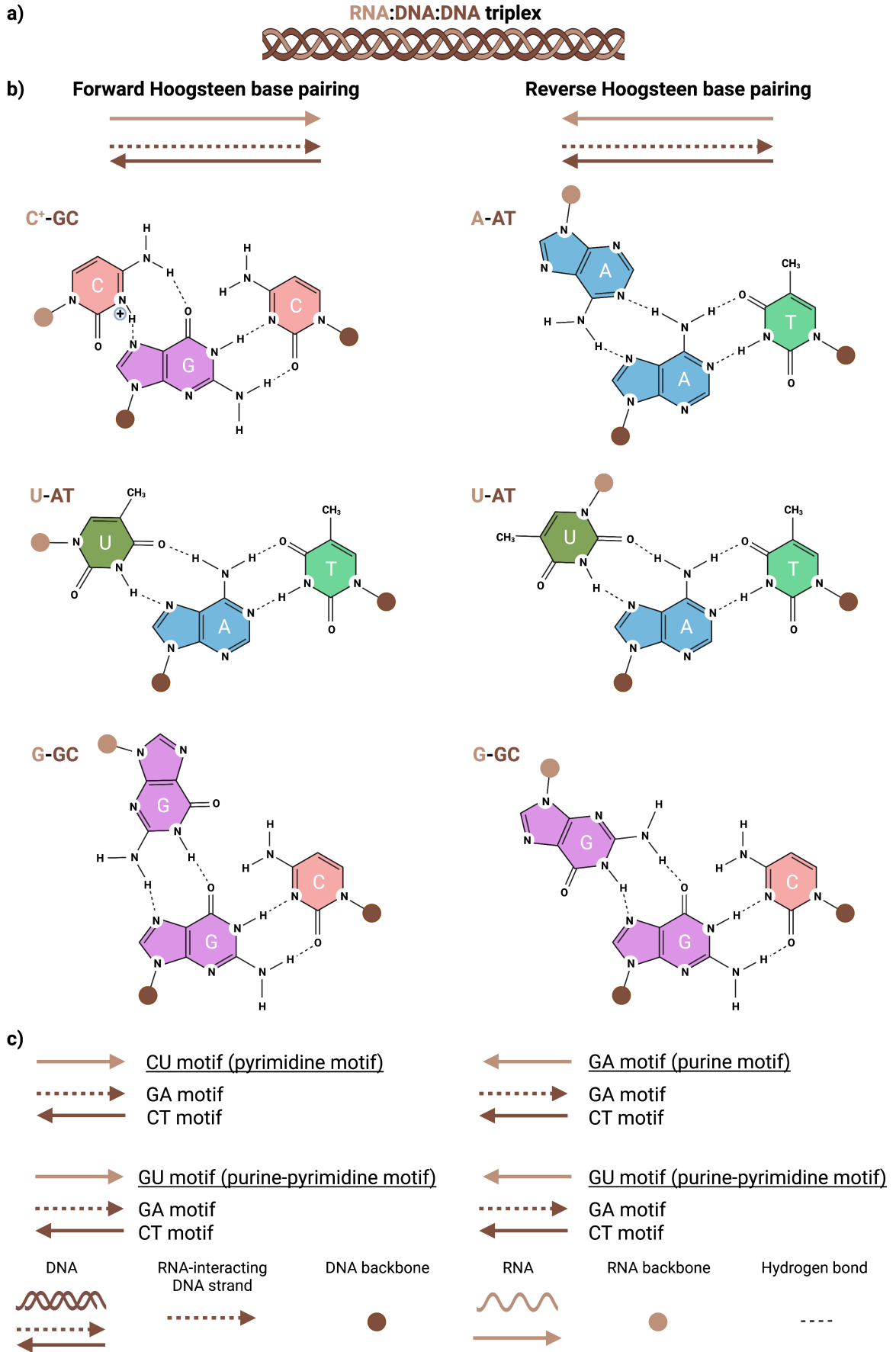


Figure 2.1 (preceding page): Triplex formation between RNA and DNA. **a)** Schematic illustration of an RNA:DNA:DNA triple helix (triplex). **b)** Forward and reverse Hoogsteen base pairing between single-stranded RNA and DNA duplex. Hydrogen bonds between nucleotide triads of RNA and DNA duplex are shown. Colours indicate the different DNA and RNA backbone as shown in the caption. The plus sign indicates protonation of a cytosine of the RNA strand. Arrows indicate 5' to 3' direction. **c)** RNA strand-related (third strand-related) sequence motifs that form a triplex with the duplex DNA. Left) forward Hoogsteen base pairing. Right) reverse Hoogsteen base pairing. [36, 37]. Created with BioRender.com.

RNA (third strand) nucleotides involved in a triplex are termed as triplex-forming oligonucleotides (TFO). A triplex target site (TTS) refers to polypurine DNA (duplex) regions that can form Hoogsteen hydrogen bonds with TFOs [36, 37].

These TTSs are enriched in promoters and enhancers [37]. Thus, the sequence-specific binding to DNA allows lncRNA to regulate transcription via guiding and recruiting TFs or chromatin-modifying proteins to cis-regulatory elements. Triplex-forming lnc genes frequently interact with the polycomb repressive complex 2 (PRC2), a chromatin-modifying protein complex that is involved in repressing gene expression, to regulate chromatin modification and transcription. Here, lncRNAs can regulate transcriptional in *cis* by regulating neighbouring genes or in *trans*. Another class of lncRNA is directly transcribed from enhancers. These enhancer-associated lncRNAs act mainly in *cis* by promoting enhancer-promoter loops or histone modifications at the promoter site when the lncRNA is brought into the vicinity via chromatin loops [40].

During the last years, an increasing number of oncogenic and tumour suppressive lnc genes that are dysregulated in cancer has been discovered [34, 41]. The functions of these lnc genes and their products relate to several cancer hallmarks, including proliferation signalling, growth suppressor regulation, replicative immortality, invasion and metastasis, and angiogenesis [34]. The reason for the dysregulation of lnc gene expression in cancer includes, among others, DNA methylation and copy number changes [41]. An interesting feature of lnc genes is the stronger cell/tissue type-specific expression compared to protein-coding genes [34]. This feature is also valid for cancer. Yan *et al.* [41] compared the expression between tumour cells and tissue-related normal cells for seven different cancer types. Among the lnc genes showing altered expression in tumours, ~60% of these lnc genes were specifically dysregulated in one cancer type.

Even though lnc genes are rarely studied, there are several lnc genes that have been well investigated in the context of cancer. The lnc gene Maternally Expressed 3 (*MEG3*; alias: *GTL2*) is one of the most studied non-coding tumour suppressors [42]. The *MEG3* RNA is located in the cellular nuclear compartment [43–45]. *MEG3* is expressed in numerous normal human tissues, such as different brain regions, including the cerebellum [46, 47]. Single-cell RNA-seq and ISH experiments in mice demonstrated that *Meg3* is a neuronal cell marker [44, 45, 48]. *MEG3* was found downregulated in various cancer types. Here, low expression of *MEG3* is associated with poor prognosis. Induced overexpression of *MEG3* in tumour cells impairs proliferation due to forced growth arrest that is caused by downregulation of cell cycle regulating factors like CDK1 and CCNB1 [49, 50].

MEG3 acts as a non-coding tumour suppressor via different mechanisms, including miRNA sponging, protein interactions, and DNA binding [51, 52]. *MEG3* functions as a tumour suppressor in breast cancer through RNA:DNA:DNA triplex formation and interaction with PRC2 resulting in negative regulation of the TGF β pathway via targeting genes such as *TGFB2*, *TGFB1*, and *TGFB2* [43]. Here, Mondal *et al.* have shown that *MEG3* regulates *TGFB1* through binding to an enhancer region that is located 131 kb upstream of *TGFB1* [43]. The authors sequenced DNA fragments bound by *MEG3*

via performing chromatin oligo affinity precipitation (ChOP). In ChOP, DNA fragments bound by the target RNA are pulled down using biotin-labelled probes of DNA oligonucleotides that are a reverse complement to sequences of the targeted RNA [43].

MEG3 is also involved in the regulation of the p53 pathway. High *MEG3* expression leads to an accumulation of p53 protein and enhanced p53-mediated transcription in several cancer types [51–53]. Zhou *et al.* reported that induced *MEG3* expression suppresses MDM2 expression in adenocarcinoma and osteosarcoma cell lines. This suppression of MDM2 by *MEG3* is relevant for tumour suppression because *MDM2* mediates p53 protein degradation [53]. Additionally, previous studies demonstrated that *MEG3* acts as a co-regulator and co-activator of p53-mediated transcription [53, 54]. Via binding to promoters, *MEG3* can also regulate gene expression of p53 targets such as *GDF15* (a growth factor that inhibits proliferation) leading to upregulation of *GDF15* [53]. Furthermore, p53 directly binds to *MEG3* RNA via its DNA binding domain [54]. In this context, it was proposed that *MEG3* promotes the tetramerisation of p53, which is essential for its transcription factor activity. It was further suggested that the p53-*MEG3* complex is recruited to target genes where *MEG3* dissociates, and p53 regulates the expression of the target [54].

MEG3 is located on chromosome 14q32.2 and part of the parentally imprinted *DLK1-MEG3* locus (alias: *DLK1-GTL2*, *DLK1-DIO3* locus) [55, 56]. This locus is conserved among several mammalian species including mice [55, 57]. Additionally to the imprinted locus, also the first ~900 nucleotides and certain secondary structures of *MEG3* are conserved among several mammalian species [47]. Due to the imprinting, the genes at the *DLK1-MEG3* locus are allele-specifically expressed depending on different epigenetic modifications (e.g. allele-specific DNA methylation on the maternal and paternal allele). *MEG3* belongs to the genes that are maternally expressed. In this locus, an intergenic differentially methylated region (IG-DMR) was identified that is located ~14kb upstream of *MEG3*. This IG-DMR serves as an imprinting control region and regulates allele-specific expression at this locus. The IG-DMR is unmethylated and methylated on the maternal and paternal allele, respectively. A second DMR, the *MEG3*-DMR, overlaps with the *MEG3* promoter and its first exon and regulates the allele-specific expression of *MEG3*. The *MEG3*-DMR is methylated on the paternal allele like the IG-DMR silencing *MEG3* expression on this allele [58]. A link between hypermethylation of the IG- and *MEG3*-DMR and associated with *MEG3* downregulation has been described in several cancer types (e.g. neuroblastoma, pheochromocytoma) [59–61]. The hypermethylation of the IG- and *MEG3*-DMR is thought to be one mechanism that causes downregulation of tumour suppressor *MEG3* in cancer. The methylation status of the IG- and *MEG3*-DMR is associated with the grade in meningiomas and overall survival in acute myeloid leukaemia [59]. However, in meningioma, a direct correlation between DMR methylation level and *MEG3* expression could not be found, suggesting that additional mechanisms probably contribute to the regulation of *MEG3* expression [62].

The assumed widespread role of lncRNAs as regulators and the specific expression patterns of lnc genes on the one hand, and a majority of unstudied lnc genes on the other hand, make lnc genes an interesting and growing research field, including for cancer research.

2.6 Omics technologies in cancer research

Omics technologies have been providing the basis for the breakthroughs in cancer research of the last decades and have become standard tools in cancer research. These tools describe high-throughput technologies that allow molecular analysis across the whole proteome (proteomics), transcriptome (transcriptomics), genome (genomics), and DNA methylome (DNA methylomics) or other classes of molecules or molecule modifications such as epigenetic modifications [63].

The array technology provided the first omics tools and facilitated, among others, genome-wide analysis of DNA methylation, single nucleotide polymorphisms (SNP, natural position-related base variations that occur in minimum %1 of a population), and gene expression on transcription level. The mentioned type of arrays are DNA chips. Such a chip is a glass slide that carries a two-dimensional

array of single-stranded oligonucleotide (DNA) probes on the surface. These probes are spotted (printed) on the slide at pre-defined positions and designed to detect a collection of desired nucleotide sequences. After extracting DNA (DNA is specially treated for analysis of DNA methylation) or RNA that is reverse transcribed into complementary DNA (cDNA), the extracted nucleotide sequences hybridise to complementary probes. Fluorescence signals are used to measure the qualitative or quantitative presents of nucleotide sequence (e.g. fragments of transcripts). Hence, measured expression levels relate to signal intensities in the case of arrays. Due to the pre-defined positions of the probes, each fluorescence signal can be related to different probes. Arrays that measure gene expression are called microarrays (even though all arrays are microarrays) or expression arrays. However, due to the need for pre-designing probes, microarrays are restricted to the pre-defined set of genes/sequences for analyses [64].

Next-generation sequencing (NGS) technologies overcome several limitations of arrays in the analysis of nucleotide sequences because NGS comprises the direct sequencing (reading) of nucleotide sequences at single-base resolution. In cancer research, frequent applications of NGS include whole genome sequencing (WGS), whole exome sequencing (WES, only DNA sequences at exon positions of protein-coding genes are sequenced), and RNA sequencing (RNA-seq, transcriptome sequencing). WGS and WES facilitate the detection of genetic and chromosomal alterations (SNVs, indels, CNVs, and chromosomal rearrangements). RNA-seq enables the quantification of gene expression on transcription level as well as the expression level quantification of transcribed genetic mutations and gene fusion due to chromosomal rearrangements [63].

Independently of the array or NGS platforms, the application of omics technologies has revealed that a single cancer type most often comprises a heterogeneous collection of molecular subgroups (also termed as subtypes). These subgroups are defined by a subgroup-specific molecular makeup comprising gene expression profiles and alterations. Differences between these subgroups are not limited to the molecular makeup and include clinical features such as overall survival (OS, time of survival of patients after diagnosis) or responsiveness to treatment [6, 63]. Understanding associations between certain molecular profiles and clinical features accelerated the discovery of prognostic and predictive molecular biomarkers. For example, these biomarkers can be mutations or expression levels of a single/collection of gene(s). Prognostic biomarkers allow assessing the future course of a disease in a patient and adjusting the extent of treatment to the risk of relapse or progression after treatment. Predictive markers indicate the chance of response or resistance to a certain therapy and support decision making regarding therapy options. The process of treatment adjustment based on molecular data is also called precision medicine or precision oncology in the case of cancer treatment. Precision medicine and oncology are substantially enabled by omics technologies [6].

Generally speaking, the molecular analysis of the cancer transcriptome by high-throughput technologies has been important to understand the characteristics and the biology of individual tumours, cancer subgroups, and cancer types. Here, RNA-seq has been a key technology that outperforms microarrays due to the independence of probe set and higher dynamic range of RNA-seq [65] for the molecular analysis of gene expression in tumours.

2.6.1 Next-generation RNA sequencing

NGS platforms of the second generation, including RNA-seq, are based on the sequencing of short sequences called reads. Among these platforms, the technology — initially developed by the company Solexa and is owned by Illumina now — has become the standard platform for NGS of short reads [66] and will be summarised for the application of RNA-seq by the following paragraphs.

The preparation of an RNA library is prior to the RNA sequencing and involves several steps starting with the extraction of RNA from a tissue sample. There are two extraction principles for RNA, disregarding small RNAs such as miRNA. The first principle is the purification for RNAs that carry a poly-A tail. Here, the aim is to extract mRNA that is generally polyadenylated. A considerable portion of

lncRNAs is also polyadenylated. The second principle was designed to extract the total RNA (independent of polyadenylation). This principle includes the depletion of ribosomal RNA (rRNA) that are the major subunits of ribosomes and account for more than 90% of the RNA within a cell. Besides RNA extraction, library preparation includes RNA fragmentation, reverse transcription into cDNA, ligation of pre-defined sequences (5' and 3' adapters and barcodes) being necessary for sequencing and processing step, size selection of fragments (generally a few hundred nucleotides), and amplification of cDNA fragments using polymerase chain reaction (PCR). The PCR step is designed to maintain strand-specific information, essential due to the occasional positional overlaps between genes on opposite strands [67].

After library preparation, a glass slide called flow cell is loaded with the single-stranded cDNA library. Adapter oligonucleotide sequences that are complementary to adapter sequences of the cDNA are attached to the surface of the flow cell. The adapter sequences of the cDNA hybridise with the slide-attached oligonucleotide adapters. Next, a second strand synthesis by polymerase starts from the slide-attached adapters. As a result, the second strand of a cDNA is attached to the flow cell via the oligonucleotide adapter. The double-stranded cDNA is denatured, and only the attached reverse strand of the cDNA remains. The next step is called bridge amplification. Here, the free end of the attached, single-stranded cDNA hybridises with the second type of slide-attached oligonucleotides; via this process, the cDNA folds into a bridge-like shape. The second strand synthesis and denaturation of the double-stranded cDNA follows; however, this time, both strands are attached. This bridge amplification is being repeated several times, leading to a local cluster of cDNAs that originally belonged to one cDNA template. Ultimately, a flow cell comprises millions of these clusters, each belonging to a different cDNA template of the library [66].

The sequencing process itself is referred to as "sequencing by synthesis" [66]. Fluorophore-labelled nucleotides that are blocked at the 3' end are used for this process. Due to the 3' blocking, these nucleotides can form a bond with a second nucleotide only at the 5' end. Depending on the base, a nucleotide is labelled with one of four different fluorophores, each emitting light in a different colour. The labelled nucleotides hybridise to the complementary base of a cDNA strand. The flow cell is imaged with different laser channels. Via the fluorophores' emitted light colour, the corresponding base at the sequenced position of a cDNA cluster can be determined. Next, the fluorophores are cleaved from the hybridised nucleotides and washed away and the 3' end is unblocked. The sequencing by synthesis is repeated over several cycles in which each cycle represents one additional position that is sequenced for an individual cDNA cluster. Here, only the end of the cDNAs is sequenced, and obtained sequences represent reads. The read length depends on the number of cycles and ranges between 50 and 150 nucleotides for RNA-seq.

The described sequencing procedure also allows generating single-end and paired-end reads. In the case of single-end reads, cDNA is sequenced only from one strand. Paired-end sequencing is the sequencing of the forward and the reverse strand of the cDNA. Since cDNAs represent fragments of the original RNA and reads are short, paired-end reads have the advantage of containing more sequence information of original RNA fragments [66]. At least 100 million sequenced paired-end reads, defining the sequencing depth, per RNA samples are recommended for a comprehensive expression level quantification of genes, including low expressed ones [68].

The sequencing of nucleotide sequences (DNA or RNA) extracted from a tissue sample or a collection of cells (e.g. cultured cells) is referred to as bulk sequencing, the opposition of single-cell sequencing. A bulk tissue sample of a solid tumour comprises tumour cells and non-tumour cells of the tumour microenvironment. The fraction of tumour cells within a tissue sample defines the tumour purity that can strongly differ between tumour samples. The tumour purity depends on the biopsy itself as a non-biological factor, the degree of tumour cell invasion, or the fraction of tumour-infiltrating immune cells. Tumour samples are generally not 100% pure [69] and, therefore, RNA-seq of bulk tumour tissue comprises a mixture of gene expression signals from tumour cells and cells of the microenvironment. This aspect needs to be considered in the analysis and interpretation of bulk RNA-seq data.

2 Cancer and gene expression

Overall, RNA-seq provides a powerful platform for the molecular analysis of cancer transcriptomes. However, to obtain relevant information from RNA-seq data, the application of computational processing and analyses using suited mathematical methods is necessary.

3 Computational analysis of cancer transcriptomes

The computational analysis of omics data includes data processing and subsequent data analyses in terms of data science [70], while this thesis concentrates on data analyses. Data science aims for the extraction of hidden and meaningful patterns from large data sets [71] and takes advantage of many different but interconnected fields, including statistics, probability theory, machine learning, data visualisation, databases, and computer science/high-throughput computing [71]. In this line, the following chapter will provide an overview of the analysis of cancer transcriptomes using RNA-seq data and suited computational and mathematical methods. After the Preliminaries (Section 3.1), a brief outline of RNA-seq data processing (Section 3.2) and frequently applied analyses in cancer research is given (Section 3.3). A more detailed description is provided for analyses and methods that will be the main focus of this thesis comprising

- the construction of gene-expression-based predictive classifiers (Section 3.4),
- the inference of GRNs (Section 3.5), and
- computational characterisation of lnc genes (Section 3.6),

while the first two points relate to machine learning tasks.

3.1 Preliminaries

The preliminaries include basic notations used within this thesis (Section 3.1.1) and a general introduction to machine learning (Section 3.1.2), focusing on supervised learning, resampling methods, and performance evaluation metrics used in supervised machine learning.

3.1.1 Notations

Data is the plural of datum. A datum is an abstract description of a real-world object (an entity, also termed as an instance in the field of machine learning). This abstract description comprises a number of collected or measured features/variables — both terms are interchangeable — for each entity. There are three different feature/variable types: numeric, nominal (categorical), and ordinal (categorical, but categories have a rank order). Entities and their features are usually organised within a matrix of N rows reflecting features and M columns reflecting entities. A data set can comprise one matrix or several matrices that are in direct or indirect relation [71].

The just-introduced terms will be used as follows. An entity is a patient/case or a single tumour tissue sample analysed via an omics platform. However, terms like sample size, sub-/resampling, and bootstrap sample relate to the statistical meaning of sample defined by a collection of entities from a bigger population. Features/variables of tumour samples comprise measured molecular data such as expression values or copy numbers of a gene. Features/variables of patients include clinical data such as OS, age, and responsiveness to treatments.

If not differently stated, an entity is a vector containing values of N features: $\mathbf{x} = x_1, \dots, x_N$ or $\mathbf{x} \in \mathbb{R}^N$, where each of the N feature values is a real number (\mathbb{R}). The transpose of the vector \mathbf{x} is given by \mathbf{x}^T . A single element within a matrix of N features by M entities is denoted by $x_{j,i}$ relating to the j th row (feature/variable) and i th column (entity/sample).

The copy number of a gene is indicated by 1N, 2N, 3N, and so on, where the number indicates the number of detected gene copies.

3.1.2 Machine learning and supervised learning

An important aspect in data analysis of big and high-dimensional data sets like omics data is the application of machine learning to solve certain problems and tasks. These problems are solved by applying an algorithm that learns a model from a data set (the input), while the learned model provides an output related to the original task. Here, the learning is driven by gaining experience to improve the knowledge or performance of the learned model. Models are learned by learning model parameters, while the number of parameters depends on the applied algorithm and the individual learning task. These model parameters should not be confused with hyperparameters, parameters of the model that cannot be learned and, therefore, need to be optimised in an extra tuning step. An additional aspect of the learning task is the selection of informative features that facilitate solving the task. The process of obtaining a learned model is called model training or fitting. Hence, a data set used for fitting a model is termed a training set. An independent data set used to evaluate the model's performance is a test or validation set [72].

The described concept of machine learning is used, among others, to solve unsupervised and supervised learning problems. Later presented data analyses within this thesis concentrate on supervised learning. Therefore, unsupervised learning is only shortly explained. Unsupervised learning is used for identifying hidden, meaningful structures (patterns) in the data. Clustering is one type of unsupervised learning. Clustering of entities is the identification of groups (clusters) of entities within a data set. (The same can be done for features.) Clustering methods find these clusters based on similarity/dissimilarity between entities. Therefore, entities assigned to the same cluster are more similar, and entities assigned to different clusters are more dissimilar to each other. Initially, these clusters are non-descriptive labels. In the case of the clustering of entities, the identification of features discriminating the found clusters can be used to gain information and knowledge about the found clusters [72].

In contrast to unsupervised learning, supervised learning deals with labelled data, where a label assigns an entity to a certain class. Here, a class relates to a collection of entities commonly sharing a characteristic of interest. These labels allow the supervised learning task classification. The classification between two classes (binary classification) is a common task and one focus of this thesis. Classification comprises the learning of a model — a classifier — that can map an entity based on its features to a class [72]. Here, a model is learned via induction, which is the inference of a classification model by the generalisation of seen training entities into general classification rules. Hence, the model is inferred by generalisation. Due to the learning of a generalised model that ideally has a general validity, a classifier allows the mapping of new, unseen entities to a class. Therefore, classifiers can make predictions for unseen data. The prediction of a class for unseen data is an important property of classifiers. In many real-world scenarios, the labelling of data is expensive and, therefore, not always possible, while the data themselves are cheap. Using one training set with labelled data for classifier training allows the identification of entities with a characteristic of interest in unseen data by class prediction [72, 73].

For classification tasks, the overall aim is to train a classifier with the best possible performance. Here, a minimal classification error defines the best performance; this error is the difference between the actual and predicted class. The classifier's performance is influenced by two major aspects that are termed as over- or underfitting of a learned model. In the case of overfitting, a model is tightly fitted to the training data, lacks generalisation, and depends on the training sets. Due to this dependency, repeating model fitting using slightly changed training data sets would cause big variations across learned models. Thus, overfitting is also termed as a high variance of a model. Overfitting is an issue since every data set contains some degree of noise that can be incorporated into the learned model due to overfitting. Underfitting relates to a model generalisation that misses an accurate representation of the training set. For example, underfitting can be caused by choosing a too simple model for complex classification problems. Underfitting is also termed as a high bias of a model. Model variance and bias are inversely related; high-bias models have low variance and *vice versa*. Therefore, finding a

trade-off between overfitting and underfitting, which corresponds to minimising the classification error/optimising the performance, is described by the bias–variance dilemma. A typical example of this dilemma is choosing the number of features. An increasing number of features makes it easier to fit a classifier because more features are available to infer rules that assign an entity to a class. However, an increasing number of features is attended by a higher chance of overfitting. A general rule is to avoid a number of features higher than the number of training entities. By contrast, reducing the number of features avoids overfitting, but underestimating the number of features leads to underfitting. A trade-off between model overfitting/variance and underfitting/bias can be achieved only by choosing an optimal number of features. The described example shows that features selection includes selecting informative features and finding an optimal number of features [72].

The described bias-variance dilemma illustrates that classification tasks are not trivial because the training of a model does not necessarily result in a classifier with low classification error. Thus, performance evaluation is essential in supervised learning. A classical setting for performance evaluation is using a test data set for evaluating hyperparameters of a model and an (external) validation set for evaluating a learned model/classifier¹. Test and validation sets are hold-out/independent from the training set to facilitate an evaluation of under- and overfitting. Still, the test set is part of the training step since it is used to tune the hyperparameters that are part of the model, which makes it necessary to use the validation set to evaluate the learned model [72].

However, a general issue of machine learning is a lack of individual data sets for training, testing, and validation. A common strategy is to split an available data set into a training and test set for model training to overcome this issue. This split is done several times to make full use of the available data set and rule out random effects by the split that influence the training and testing results. Here, resampling methods are used to obtain several data set splits, as summarised in the following section [72].

3.1.2.1 Resampling methods

Two resampling methods in supervised learning are commonly used to obtain data set splits into training and test sets: k -fold cross-validation and bootstrapping [72, 74].

In K -fold cross-validation, a data set is split into K partitions of equal or nearly equal size. Over $k = 1, \dots, K$ iterations, the entities of the k th partition are used as a test set and the remaining entities of the data form the training set. A classifier is learned on the training set. Afterwards, the trained classifier is used to predict the class of the entities in the test set. The performance is evaluated by averaging the classification error across the test folds or by pooling the classifier predictions of all test folds to calculate the classification error. The pooling version of K -fold cross-validation will be used in this dissertation due to small data sets. A common value of K is 10. Thus, 90 % and 10% of the data are used as a training and test set at each iteration, respectively. An advantage of K -fold cross-validation is that each entity of the data set is used once for testing, whereas other resampling techniques are more random and include entities for testing at different rates [72, 74].

Bootstrapping is a random and uniform resampling with replacement. Considering a data set with N entities, a bootstrap sample has a sample size of N . Due to the uniform sampling with replacement, each entity has a probability of $p = (1 - 1/N)^N$ for not being part of the bootstrap sample. Thus, a bootstrap sample contains ~63.2% of the original entities on average and ~36.8% are hold-out, while the hold-out (out-of-bag) entities form the test set. For accessing the performance of a classifier, bootstrapping is repeated several times (e.g. 50-100) and the classification error is calculated by using averaging or pooling across bootstrapping iterations. Bootstrapping has a lower variance of testing results across resamplings than K -fold cross-validation, especially for small data sets [74, 75].

¹Depending on the research field, the definition of a test and validation set can be interchanged.

3.1.2.2 Stratified resampling for classes imbalance

A common issue in classification tasks is an unbalanced number of entities between classes in a data set, termed class imbalance. The more and less frequent class is referred to as majority and minority class, respectively. Class imbalance of a training set can negatively affect the performance of a classifier when the imbalance was not considered during the training and evaluation [76]. Class imbalance is also an issue for resampling since sampled data do not necessarily include the same proportion of classes as the original data set. Stratified resampling addresses the issue of resampling unbalanced data. Here, each class is independently resampled. In the case of a split into a training and test set, the hold-out entities of each class are combined into the test set and the remaining entities form the training set. This stratification ensures that the proportion of classes is constant in the training and test set. The stratification can be applied for K -fold cross-validation and bootstrapping [74].

3.1.2.3 Performance metrics

After summarising supervised learning and testing strategies, this section introduced metrics for measuring the performance of trained classifiers, concentrating on metrics appropriate for imbalanced data.

In binary classification, the two classes are termed positive (+1) and negative (-1) class. Four different counts evaluate the possible cases of binary classification outcomes, as summarised by the confusion matrix in Table 3.1. True positive (TP) and true negative (TN) values summarise the number of correctly predicted entities for the positive and negative class by the classifier, respectively. False positive (FP) and false negative (FN) values summarise the number of incorrectly predicted entities. These four values are the basis for many performance metrics [77].

Table 3.1: Confusion matrix. TP - True Positive. FP - False Positive. TN - True Negative. FN - False Negative.

		Predicted class	
		Positive	Negative
Original class	Positive	TP	FN
	Negative	FP	TN

The introduced performance metrics are summarised in Table 3.2. The metrics sensitivity and specificity indicate the fraction of correctly predicted positive and negative class entities, respectively. The accuracy is the fraction of correctly predicted entities across both classes. However, in the case of class imbalance, the majority class has more weight in the calculation of the accuracy leading to an unbalanced performance measure. The balanced accuracy addresses this issue by equally considering sensitivity and specificity and can be used for class-unbalanced data sets. The F_1 -score can also be used for class-unbalanced data but puts an emphasis on the positive class. The F_1 -score can be useful when the sensitivity of a classifier is more relevant than its specificity [77].

Table 3.2: Performance metrics for binary classification [77].

Metric	Formula
sensitivity	$\frac{TP}{TP+FN}$
specificity	$\frac{TN}{TN+FP}$
accuracy	$\frac{TN+TP}{TP+FN+TN+FP}$
balanced accuracy	$\frac{\text{sensitivity}+\text{specificity}}{2}$
F ₁ -score	$\frac{2TP}{2TP+FP+FN}$

3.2 Processing and normalisation of RNA sequencing data

The sequencing step transforms RNA sequences that are expressed within a cell into digital read data. The sequenced RNA reads contain qualitative and quantitative information regarding which gene/transcript is expressed and the expression level for these genes/transcripts, respectively. In order to obtain this information, reads need to be mapped back to their original position in the genome. This is done during the processing step of read mapping/read alignment using computation in combination with suited read alignment methods. Since matured lncRNA and mRNA does not include introns, mostly RNA reads map to exonic genome regions. The exclusion of introns in the matured RNA introduces so-called split reads that stretch over neighbouring exons in the RNA but are split by an intron on the genome. A read's genomic position is matched to the coordinates of known genes and their exons to annotate a read to the originally transcribed gene. Generally, only reads with unique mapping positions are considered [78].

For gene expression level quantification, reads or fragments mapped to exon positions of a gene are summed up, resulting in a read or fragment count per gene. Fragments are counted when paired-end sequencing was performed to avoid that the same cDNA fragment is counted twice since a read pair (not a single read) is available per fragment. The obtained counts can be used for further analyses. However, in most cases, gene expression is quantified by Reads or Fragments Per Kilobase of exon model per Million mapped reads (RPKM or FPKM, respectively):

$$\text{RPKM}_g \text{ or } \text{FPKM}_g = \frac{r_g}{\frac{l_g}{10^3} \frac{N}{10^6}}, \quad (3.1)$$

where l_g is the exonic length and r_g is the read or fragments count of gene g , and N is the number of mapped reads or fragments, depending on the calculation of RPKM or FPKM. The quantification via RPKM/FPKM permits the comparison of expression levels between genes and samples because raw read/fragments counts are biased by the length of the exonic sequence of a gene and the sequencing depth. The length of the exonic sequence correlates with the length of the transcripts, and an increasing transcript length correlates with an increasing number of fragments that can be generated from one transcript. Therefore, genes with a longer exonic sequence accumulate more reads/fragments. The term "reads/fragments per kilobase of exon model" relates to the correction for the length bias and allows a comparison of expression values between genes. The sequencing depth of a sample can be individually chosen; however, the sequencing depth naturally varies even when the same depth is aimed for a collection of samples. The sequencing depth determines the pool of reads that is available for expression quantification. A smaller read pool results in a lower number of mapped reads;

although, when a gene expression level is constant across samples. The term "per million mapped reads" accounts for the sequencing depth bias and permits a comparison of expression values between samples [78, 79].

The calculation of RPKM/FPKM values does not correct all biases in RNA-seq-based gene expression quantification. An additional bias source is a different RNA/transcript composition between sequenced samples. Changes in the RNA composition refer to highly variable (such as unique expression of individual genes) or extreme high gene expression (accumulation of millions of reads by few genes). These changes introduce variations in the availability of reads from the original read pool. Such variations result in different read counts for an expressed gene, even when the gene is evenly expressed across samples. Robinson and Oshlack [80] proposed a correction for RNA composition effects by the weighted trimmed mean of M values (TMM). The TMM is based on the assumption that most genes have a similar expression level between samples. Here, the M-values relate to gene-wise log expression ratios between two samples. This assumption arises from processes that are active in every cell and carried out by so-called housekeeping genes. The TMM is used as a normalisation factor between a reference and a non-reference sample and can be interoperated into the calculation of RPKM/FPKM values. Detailed information on TMM calculation can be found in the publication of Robinson and Oshlack [80].

3.3 Frequently applied data analyses of cancer transcriptomes

After the processing, the obtained read counts and expression values per gene can be used for various computational analyses. This section provides a brief overview of four frequently performed analyses of cancer transcriptomes, while these analyses are connected and build on each other, starting with molecular subtype identification.

3.3.1 Unsupervised clustering for molecular subtype identification

The discovery of molecular subgroups has been an essential step in understanding cancer (see Section 2.6). The identification of molecular subgroups within cancer types by analysing gene expression data from tumours has become a standard task in cancer research [81–85]. These subgroups have distinct gene expression profiles defined by subgroup-specific up- or downregulation of genes. Several subgroup-specific factors can contribute to the distinct gene expression profiles of subgroups, including [85]

- genetic, chromosomal, epigenetic alterations,
- epigenetic changes during cancer development,
- heterogeneity of the tumour microenvironment (e.g. a degree of immune cell infiltration), and
- the cell-of-origin (an organ associated with one cancer type normally consists of many cell types).

The challenge of identifying molecular subgroups is that the subgroups are unknown at the beginning. The process of subgroup identification is defined by finding unknown, hidden, meaningful patterns within a gene expression data set of tumour samples. Therefore, finding subgroups is an unsupervised machine learning/clustering task (see Section 3.1.2).

Like for every other machine learning task, the selection of features mainly influences the analysis results. Additionally, since gene expression data are high-dimensional and contain many irrelevant or noise genes, feature selection is necessary for this data type. Since the patterns within the data

are unknown, relevant genes need to be selected in an unsupervised manner. The variance of gene expression across tumour samples is frequently used as a measure to select relevant genes [86] because subgroup-specific gene expression demands variation in gene expression.

Using the set of selected genes, a clustering algorithm is applied to identify clusters of tumours. These clusters represent the molecular subgroups. Many different algorithms can be applied for gene-expression-based tumour sample clustering [86, 87]. Commonly used algorithms are hierarchical clustering [87], consensus clustering [88], and negative matrix factorisation (NMF) [89] (please see related literature for information on the algorithms). After clustering, the found solution of clusters can be evaluated by different metrics such as the silhouette score that measures how well a sample fits its assigned cluster [86].

The obtained clusters of tumours are further analysed since the clustering does not necessarily provide the information on which genes contributed to the clusters of tumours in terms of subgroup-specific gene expression. Besides the molecular biological factors that contribute to the subgroups, it is expected that clusters of tumours show specific gene expression because clustering algorithms group tumours with similar expression profiles. The detection of subgroup-specific gene expression is done via differential gene expression analysis.

3.3.2 Differential gene expression analysis

Differential gene expression analysis (DGEA) is a statistical analysis to identify genes that show significant differences in expression between compared groups of samples, such as molecular subgroups. In the case of RNA-seq data, read counts per gene across samples are directly used to perform a parametric statistical analysis. Here, software tools such as "edgeR" [90] provide an implementation of statistical methods that can be used for RNA-seq data. Since the statistical analysis is parametric, it is necessary to know the underlying probability distribution of read counts. In edgeR, the read count Y_{gi} of gene g in sample i is modelled by a negative binomial distribution (NB):

$$Y_{gi} \sim \text{NB}(\mu_{gi}, \phi_g), \quad (3.2)$$

$$\mu_{gi} = N_i \pi_{gi}, \quad (3.3)$$

where μ_{gi} is the mean and ϕ_g the dispersion of the gene reads count, N_i is the total number of mapped reads, and π_{gi} is the fraction of reads that map to gene g in sample i . The testing for differential expression is done by fitting a generalised linear model (GLM) and applying a likelihood ratio test (LRT). The GLM is defined by

$$\log \mu_{gi} = \mathbf{x}_i^T \boldsymbol{\beta}_g + \log N_i, \quad (3.4)$$

where \mathbf{x}_i is a vector of covariates (predictors), and $\boldsymbol{\beta}_g$ is a vector of regression coefficients indicating the effects of a covariate on gene g , and $\log N_i$ is the intercept. Predictors of gene expression in the GLM are normally group assignments for each sample. (These group assignments are represented by dummy variables that have a value of 0 or 1.) Therefore, a fitted GLM is used for predicting gene expression across samples using group assignments as predictors. However, the goodness of fit of a GLM can differ since a fitted GLM only predicts gene expression well when the expression depends on the groups. Generally, not all expressed genes have an expression profile following the groups. An LRT is applied to test whether the goodness of fit of a GLM can be achieved just by chance. Here, the LRT compares the goodness of fit between the full GLM and a null model, which is the intercept of the GLM. The statistical value provided by the LRT can be used to calculate a p-value indicating whether the expression of a gene significantly depends on the tested groups. Significance means differential gene expression between groups. Since thousands of genes are tested in a DGEA, the p-values need to be correct for multiple testing using correcting methods such as the Benjamini and Hochberg (BH) procedure that is also termed as false discovery rate (FDR) [90, 91].

The FDR per gene (and fold-change between groups) can be used to select genes that are significantly differentially expressed between subgroups. Further analyses of a particular subgroup generally integrate the upregulated genes in this subgroup because these genes define the active biological processes. The active biological processes can be determined via a functional overrepresentation/enrichment analysis.

3.3.3 Functional overrepresentation analysis

An important element in functional overrepresentation analyses are databases that contain functional annotations of genes. These annotations include information on whether a gene is part of a certain pathway or assigned to a gene ontology (GO) term [92]. An ontology is a defined system of terms that have a defined relationship that can be used to describe an object. GO terms provide such a system to functionally describe genes and allow the functional characterisation of genes beyond pathways [93].

Such databases facilitate the functional characterisation of identified tumour subgroups via the set of subgroup-specifically expressed genes. At first, the databases are used to annotate the known functions of the subgroup-specifically expressed genes. Second, a statistical test is applied to test whether a certain gene function (e.g. DNA binding) is significantly overrepresented among the subgroup-specifically expressed genes. Such a statistical test is based on the hypergeometric probability distribution [92, 94]:

$$P(x = k) = \frac{\binom{l}{k} \binom{m-l}{n-k}}{\binom{m}{n}}, \quad (3.5)$$

where k is the number of subgroup-specifically expressed genes annotated for a certain function z , l is the number of subgroup-specifically expressed genes, m is the number of expressed genes in the data set, and n is the number of expressed genes that are annotated for function z . Significantly overrepresented functions among upregulated genes can help to determine the dysregulated biological processes and pathways that led to cancer formation within an identified molecular subgroup.

However, the characterisation of molecular subgroups within a cancer type is not limited to differentially expressed genes. Subgroups are compared by the frequency of mutations/alterations in individual tumour suppressors and proto-oncogenes and by the clinical outcome, such as the overall survival of patients [84].

3.3.4 Survival data analysis

Survival data represent a special data type. These data represent the time to a particular event such as death [95]. However, not all patients experience an event within a limited follow-up time. In such a case, the event time data are right-censored. Analyses of these data normally relate to the function $S(t)$ that is the probability of survival until time point t among the analysed group of patients. Several statistical methods can handle this kind of data [95].

The Kaplan-Meier estimator is used to plot the survival curve within a limited follow-up for a group of patients [95]. The Kaplan-Meier estimator can measure the fraction of patients that have survived until a given time point t_j [96]. The Kaplan-Meier curve obtained from the estimator is a right-continuous step function, where each step represents uncensored events in patients within a given period [95]. The Kaplan-Meier estimator of $S(t)$ is given by

$$\hat{S}(t) = \prod_{t_j < t} \frac{n_j - d_j}{n_j}, \quad (3.6)$$

where n_j is the number of patients that did not die before t_j , and d_j is the number of patients that died at t_j [95, 96]. By applying the Kaplan-Meier estimator to each subgroup, the survival curves can be plotted and compared. To determine differences between survival curves, the hazard ratio and the log-rank test are used.

3.4 Gene expression-based tumour sample classification

The log-rank test tests for differences between survival curves between two groups, denoted as group 1 and 2. Let t_j be the j th time point, $j = 1, \dots, J$. Let N_{1j} and N_{2j} the number of patients that survived until time point t_j in both groups, while $N_j = N_{1j} + N_{2j}$. Let O_{1j} and O_{2j} the number of patients that died at t_j in both groups, while $O_j = O_{1j} + O_{2j}$. The log-rank statistic is defined as

$$Z = \frac{\sum_{j=1}^J (O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^J V_j}}, \text{ where} \quad (3.7)$$

$$E_{1j} = O_j \frac{N_{1j}}{N_j} \text{ and} \quad (3.8)$$

$$V_j = \frac{O_j(N_{1j}/N_j)(1 - N_{1j}/N_j)(N_j - O_j)}{N_j - 1}. \quad (3.9)$$

The p-value of the log-rank statistic Z can be obtained via the chi-square distribution. This p-value indicates whether the survival curves of the two subgroups are significantly different [96].

The hazard function gives the probability that a patient dies within a given time period and is defined as $h(t) = \lim_{\Delta t \rightarrow 0} \frac{\text{probability a patient dies between } t \text{ and } t + \Delta t}{\Delta t}$ [96]. The hazard ratio (HR) is a ratio between two hazard functions values of two patient groups: $HR = h_1(t)/h_2(t)$. The HR indicates how much likelier it is that a patient dies in one or the other group [96]. For example, a HR= 2 would mean that group 1 has a two-times higher risk of dying than group 2 (in a comparison group 2 vs. group 1 the HR would be 0.5).

These three statistical methods are frequently used to analyse survival data in cancer research and to detect potential differences in clinical outcomes between cancer subgroups. However, these statistical methods can also be used to detect prognostic biomarkers. In the case of gene expression as a biomarker, a common strategy is to iterate over different cut-points of gene expression values, where a cut-point is used to split a set of analysed tumours into two groups. A log-rank test is used at each cut-point to find the cut-point with the lowest p-value, which is equivalent to the best separation of survival curves. By repeating this procedure for each expressed gene, a set of putative prognostic genes can be identified that need to be validated on an external cohort [97].

While the previous four sections outlined standard analysis methods in cancer research, these methods leave many different aspects of cancer transcriptomes analysis uncovered.

3.4 Gene expression-based tumour sample classification

The classification of tumour samples is a typical supervised machine learning task in cancer research. Learning tasks include the classification of tumours regarding (1) the likelihood of disease progression/relapse after treatment (prognostic) or (2) the chance of response or resistance to a therapy (predictive). This thesis focuses on predictive classifiers.

The identification of predictive biomarkers is essential to make treatment-related outcome predictions for tumours [98]. Such biomarkers are observable/measurable molecular features of tumours and include mutations/alterations (CNVs, SNV/indels), methylation levels of CpGs, and expression levels of genes. For example, the simplest biomarker would be a specific gene mutation that predicts the resistance of tumours to a certain therapy; as a consequence, only wild-type tumours would be treated with this therapy. Besides mutations, gene expression is a commonly used molecular data type for outcome prediction. Such predictions would be based on the expression level of a single gene or set of genes [98, 99]. Since this work concentrates on the analysis of cancer transcriptomes, gene expression data will be used for treatment outcome classification.

Several supervised machine learning algorithms can be used for treatment outcome classification. Among these algorithms is the SVM that was developed by Vapnik and colleagues [100, 101]. Two

publications by Statnikov *et al.* [102, 103] have shown that SVMs perform better than other supervised learning algorithms when applied to gene expression data. Here, the comparison included SVM, random forests (an ensemble of decision trees), K -nearest neighbours, and probabilistic neural networks. Due to the better performance on gene expression data, SVMs will be applied to solve the later-presented classification of treatment outcome.

3.4.1 Binary classification and support vector machines

This section introduces the four basic concepts behind the SVM using mathematical and non-mathematical descriptions. These concepts include linear classifiers, SVMs as a maximum-margin classifier, kernels, and soft-margin SVMs. Additionally, an adaptation of the soft-margin SVM will be described that can be used to learn from class-unbalanced data.

A binary classification problem can be mathematically formulated as follows, considering a training data set that contains two classes of entities. This training data set is defined by $S = \{ (\mathbf{x}_i, y_i) \mid 1 \leq i \leq M, \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y} \}$, where \mathbf{x}_i is a vector of N features describing an entity and y_i is the class label of entity \mathbf{x}_i . The input space that includes all entities of the training set is defined by \mathcal{X} . The binary label space of the positive and negative class is defined by $\mathcal{Y} = \{ -1, +1 \}$. A binary classification task can be solved by learning a function $f(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{Y}$, where the function $f(\mathbf{x})$ maps the entity \mathbf{x} from the input space \mathcal{X} to the label space \mathcal{Y} [72].

The first basic concept of an SVM is to find a linear decision boundary that separates two classes of entities within an N -dimensional space (assuming that the classes are linearly separable). Such a decision boundary can be defined by a linear discriminant function called a perceptron [104, 105]:

$$\text{sgn}(f(\mathbf{x})) = \text{sgn}(\mathbf{w}^T \mathbf{x} + b) = \begin{cases} +1 & \text{if } f(\mathbf{x}) > 0, \\ 0 & \text{if } f(\mathbf{x}) = 0, \\ -1 & \text{if } f(\mathbf{x}) < 0, \end{cases} \quad (3.10)$$

where $\mathbf{w} \in \mathbb{R}^N$ is a feature weight vector defining the orientation of the boundary, $b \in \mathbb{R}$ is a bias term moving the boundary parallel to its orientation, and b and \mathbf{w} control the function $f(\mathbf{x})$. The case $\mathbf{w}^T \mathbf{x} + b = 0$ defines a discriminant plane — a hyperplane — that separates both classes (Figure 3.1).

However, there are many solutions for a hyperplane to linearly separate two classes. It remains the question of the best separating hyperplane. An intuitive choice would be to place the hyperplane in the middle between both classes. This choice reflects a solution with a good generalisation of the classification task because this solution is robust to variations in unseen data. Such a generalising hyperplane can be found by maximising the margin between both classes of entities (Figure 3.1). Therefore, the maximisation of a class-separating margin is the second basic concept of an SVM [104, 105].

Combining both concepts, an SVM is a classifier that finds the maximal margin hyperplane. The width of the margin is given by $2/\|\mathbf{w}\|_2$, where $\|\mathbf{w}\|_2$ is the euclidean norm (or L_2 -norm) (Figure 3.1). Hence, the margin can be maximised by minimising $\|\mathbf{w}\|_2$. Entities that are the nearest to the generalising hyperplane are called support vectors. The function $f(\mathbf{x})$ of these support vectors define the margin boundaries (supporting planes) related to $\mathbf{w}^T \mathbf{x} + b = +1$ and $\mathbf{w}^T \mathbf{x} + b = -1$ (Figure 3.1) [104]. Therefore, an SVM tries to find \mathbf{w} and b such that the two inequalities [101]

$$\mathbf{w}^T \mathbf{x}_i + b \geq 1 \text{ if } y_i = +1 \text{ and} \quad (3.11)$$

$$\mathbf{w}^T \mathbf{x}_i + b \leq -1 \text{ if } y_i = -1 \quad (3.12)$$

are true for all entities of the training data set, while minimising $\|\mathbf{w}\|_2$. Both inequalities can be rewritten in the form

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1. \quad (3.13)$$

Considering this, the optimisation problem to find the maximal margin hyperplane consists of an objective function and a constraint:

$$\begin{aligned} & \text{minimise} && \|\mathbf{w}\|_2^2 && \text{by optimising } \mathbf{w} \text{ and } b, \\ & \text{subjected to} && y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, && \\ & && i = 1, \dots, M, && \end{aligned} \quad (3.14)$$

assuming both classes are linearly separable.

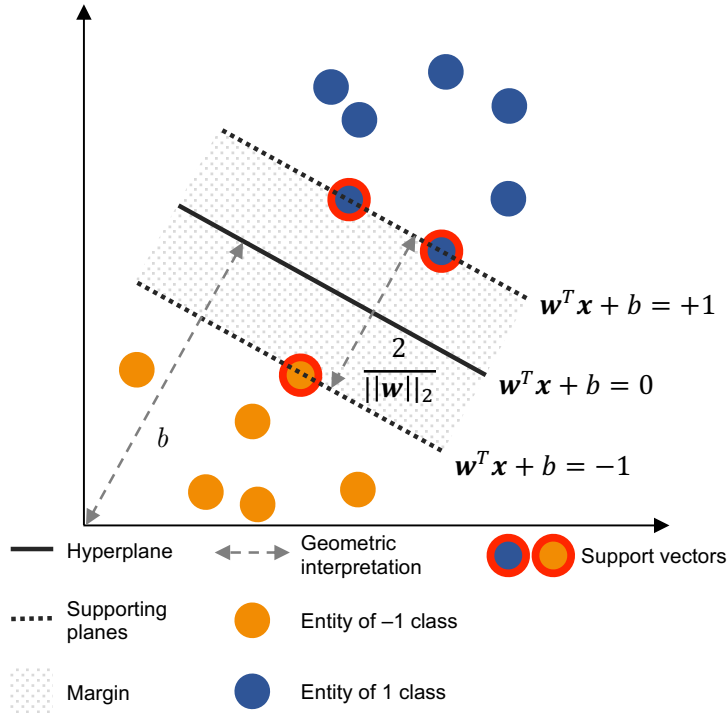


Figure 3.1: SVM is a maximum-margin classifier [104]. Grey, dashed arrows illustrate geometric interpretation of the values of b and $\frac{2}{\|\mathbf{w}\|_2}$ defined by the length of the arrows.

However, optimisation problems like this are difficult to solve due to their constraints. Therefore, the optimisation problem is reformulated from a primal to a dual optimisation problem that is easier to solve. The following paragraph will summarise the advantages of the dual representation (for a more detailed description and derivation of the dual representation see [105]).

Until now, it was assumed that the two classes of entities are linearly separable in input space. However, this is not the case for many real-world classification tasks. In order to solve tasks like that, the input space is mapped into a more complex or higher dimensional feature space \mathcal{F} using a mapping function $\phi(\mathbf{x})$ so $\mathcal{F} = \{\phi(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$. In this new feature space, the classes are linearly separable. A drawback is that mapping the input into the features space using a mapping function brings several problems. However, kernel functions or just kernels facilitate the mapping of data into the features space and the training of a classifier in this features space even without knowing the mapping function. The term kernel trick relates to this special characteristic of kernels. A kernel can be seen as a function that measures the similarity between entities, for example, by calculating the inner product between pairs of entities. Therefore, when a kernel is applied, a classification function $f(\mathbf{x})$ depends only on similarity measurements (dot products) between entities. Going back to the SVM, adding the unknown mapping function $\phi(\mathbf{x})$ to $f(\mathbf{x})$ and bringing $f(\mathbf{x})$ in the dual representation is a kernelisation of $f(\mathbf{x})$ (Equation 3.16) [105]. The dual representation involves that \mathbf{w} is formulated

3 Computational analysis of cancer transcriptomes

as a function:

$$\mathbf{w} = \sum_{i=1}^M y_i \alpha_i \phi(\mathbf{x}_i), \quad (3.15)$$

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^M y_i \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b \\ &= \sum_{i=1}^M y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b, \end{aligned} \quad (3.16)$$

where $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x})$ is the dot product, and $K(\mathbf{x}_i, \mathbf{x})$ is the kernel function. An additional aspect of the dual representation is α_i , which defines the influence of each \mathbf{x}_i (entity) for the final solution. Only entities that are support vectors have a non-zero value of α_i . This means only support vectors remain relevant for the solution of the dual representation. Intuitively, this appears to be correct considering the graphical representation of an SVM (Figure 3.1) [105]. Overall, the dual representation of an SVM provides a solvable optimisation problem, and the kernelisation provides a linear decision boundary in the feature space for classes of entities that are not linearly separable in the input space. Therefore, the application of kernels to solve initially non-linear decision boundaries is the third basic concept of an SVM. There are different types of kernel functions that are applied for SVMs. The simplest form of these kernels, a linear kernel, is the dot product between entities. Such a linear kernel will be used in the later presented study.

The so-far-presented description of an SVM relates to a so-called hard-margin SVM. The term derives from the fact that an entity can lie either on the border of the margin as a support vector or outside of the margin, which is a classification without errors. This means hard-margin SVMs presume non-noisy data, the absence of outliers, as well as non-overlapping classes at class boundaries (independent of input and feature space). However, this does not reflect most real-world data sets. Addressing this issue, the so-called soft-margin SVM was developed that relaxes these constraints. The primal representation of the optimisation problem of a soft-margin SVM includes two additional variables, the constant C and the a so-called slack variable ξ_i :

$$\begin{aligned} \text{minimise} \quad & \|\mathbf{w}\|_2^2 + C \sum_{i=1}^M \xi_i && \text{by optimising } \mathbf{w} \text{ and } b, \\ \text{subjected to} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, \dots, M, \\ & \xi_i \geq 0, i = 1, \dots, M. \end{aligned} \quad (3.17)$$

The slack variable ξ_i is an error due to misclassification. An entity is misclassified when it lies within the margin or crosses the hyperplane falling on the side of the opposite class. The hinge loss describes this classification error: $\max(0, 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b))$. The hinge loss is zero for correctly classified entities and increases with the degree of misclassification. As shown in Equation 3.17, the soft-margin optimisation problem minimises not only $\|\mathbf{w}\|_2^2$ to maximise the margin but also minimises the classification error. Additionally, the subject constraint is relaxed by subtracting ξ_i from the bound, which allows misclassification. The constant C is a factor of ξ_i and controls the influence of classification errors in the minimising problem. Hence, C is termed as a cost parameter since it controls the cost of classification errors. Additionally, this parameter can control the bias-variance trade-off of an SVM. A low value of C relates to a wide margin and a high bias (low variance) because classification errors have a low cost. A high C value corresponds to a narrow margin and high variance (low bias) because classification errors have a high cost. Thus, a high C increases the risk for overfitting. C itself is a hyperparameter because it is not predefined nor can be directly calculated. The tuning of C is done via a grid search where the performance of different C values is systematically evaluated by repeatedly splitting the data into training and test sets using resampling techniques (see Section 3.1.2.1) [105]. The optimisation problem of a soft-margin is also solved by using the dual representation of the problem.

However, this representation is not further described since the main aspects of the dual representation explained for the hard-margin SVM remain valid for the soft-margin SVM (please see [105] for more details). Taken together, the fourth basic concept of an SVM is the maximisation of a class-separating margin while minimising classification errors. Due to this concept, an SVM can generalise a learning task well in the presence of noisy data.

The soft-margin SVM is the generally-used version of an SVM. SVMs are relatively robust to overfitting also when the number of features exceeds the number of training entities. The reason for this property becomes obvious when looking at the objective function of the soft-margin SVM (Equation 3.17) and function $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$ [106]. A common source of overfitting by linear functions such as $f(\mathbf{x})$ are the optimised values (coefficients) of the feature weight vector \mathbf{w} . This optimised vector contains higher weights for features facilitating the solution of $f(\mathbf{x})$ compared to non-informative features. Here, there is a risk that the optimised feature weight vector highly depends on the training data leading to a lack of generalisation and overfitting. This problem is solved by adding additional constraints that keep weights small on average. This approach is termed shrinkage and is a regularisation technique. Regularisation is a concept that prevents overfitting in learning tasks [72]. The term $\|\mathbf{w}\|_2^2$ is a regulariser since the objective function 3.17 is minimised, which is leading to a shrinkage of the feature weights of \mathbf{w} [106]. Regularisers are used in the L_1 -norm or squared L_2 -norm (as for SVM) of the feature weight vector, where the norm of a vector is $\|\mathbf{w}\|_p := (\sum_j^M |w_j|^p)^{1/p}$. The L -norm of the regulariser affects the shrinkage of the feature weights. The squared L_2 -norm provides small, non-zero weights, and assigns similar weights among correlated features, called a grouping effect. Due to the small weights, L_2 -norm regularisation is capable of handling many features before overfitting. The L_1 -norm regularisation shrinks most weights towards zero resulting in a selection of features. However, correlated features are an issue for L_1 -norm regularisation. It will arbitrarily choose only one among several correlated features [107]. The squared L_2 -norm of \mathbf{w} ($\|\mathbf{w}\|_2^2$) is originally used for the SVM [106] and will be applied in a later-presented study. In this study, L_1 - and L_2 -norm regularisation for SVMs will be further discussed.

The last aspect of SVMs, which is addressed in this section, is the application of SVM to unbalanced data. SVM are affected by class imbalance, where the majority class has more influence in the SVM training. This is leading to a higher misclassification of the minority class. However, the objective function of the soft-margin SVM can be easily extended in a way that it can handle unbalanced data [108]:

$$\begin{aligned}
 &\text{minimise} && \|\mathbf{w}\|_2^2 + C^+ \left(\sum_{i:y_i=+1} \xi_i \right) + C^- \left(\sum_{i:y_i=-1} \xi_i \right) \quad \text{by optimising } \mathbf{w} \text{ and } b, \\
 &\text{subjected to} && y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, \dots, M, \\
 &&& \xi_i \geq 0, i = 1, \dots, M,
 \end{aligned} \tag{3.18}$$

where C^+ and C^- are cost parameters of the positive and negative class, respectively. By adding a cost parameter and an error term for each class (+ and -), classification errors per class can be independently weighted by tuning C^+ and C^- . By giving the minority class a higher C value than the majority class, an adjustment for class imbalance can be embedded in the SVM learning algorithm. Due to the different weighting of classes, this kind of SVM is also called (class) weighted SVM. Class-imbalance-aware learning is termed cost-sensitive learning. In this case, it is a cost-sensitive SVM. The ratio between the hyperparameters C^+ and C^- is not necessarily similar to the ratio between the number of entities per class [108, 109]. Therefore, value combinations of C^+ and C^- need to be tuned via grid search by using resampling and performance metrics that are suited for class-unbalanced data (see Section 3.1.2.2, 3.1.2.3).

As explained above in the context of machine learning, besides the learning algorithm, feature selection also has a major impact on the performance of the learned classifier. The next section introduces how SVMs can be used to select relevant features.

3.4.2 Feature selection and support vector machines

In classification, feature selection is the identification of features that discriminate between classes and service to solve the learning task. In the case of continuous numeric values such as gene expression values, features can discriminate between classes by differences in value range/distribution between the classes. Especially for high dimensional data such as omics data, feature selection is essential, considering the general rule that the number of features should be less than the number of training entities to avoid overfitting of a classifier. For example, gene expression data sets comprise around 20,000 measured coding genes, but 100 or fewer training entities are available. It leaves a big gap between the number of features and training entities [110].

Feature selection can be done via wrapper, filter, or embedded methods. The following paragraphs concentrate on the last two. Filter methods are applied as a pre-processing step prior to classifier training. Here, features are filtered using statistical methods such as correlation, t-test, and linear models. These methods test or measure the association between class labels and features. Coefficients and p-values derived from the statistical analysis function as filter criteria using ranking schemes or hard cutoffs. Filter methods work independently of the later applied learning algorithm, so the performance of the features for the actual classification task is unknown. This is a drawback of filter methods in terms of supervised learning. Nevertheless, filter methods facilitate the selection of relevant features that also help to understand the analysed data [110].

In contrast to filter methods, embedded methods incorporate feature selection in the classifier training and relate to the applied learning algorithm [110]. Guyon *et al.* [111] have published an SVM-embedded feature selection method that was tested for gene-expression-based tumour sample classification. The method is called SVM Recursive Feature Elimination (SVM-RFE). This method uses the weights (coefficients) of the feature vector w of a trained linear kernel SVM to rank and select features. Here, the squared feature weight $(w_j)^2$ is used as a ranking criterion of the j th feature. Important features have a high ranking criterion. The squaring is important to rank features by the magnitude of the weights independent of the sign of a weight, which indicates whether a feature is important for the positive or negative class. SVM-RFE works as follows:

- 1) Set \mathcal{U} of selected features contains all $j = 1, \dots, M$ features; set \mathcal{R} of ranked features is empty;
- 2) tune hyperparameters and train an SVM using features in \mathcal{U} ;
- 3) rank features by the criterion $(w_j)^2$;
- 4) eliminate feature j with the smallest ranking criterion from \mathcal{U} ;
- 5) add feature j to \mathcal{R} ;
- 6) repeat step 2)-5) until all features are ranked in \mathcal{R} .

Features that are added to \mathcal{R} at the end have the highest ranking/importance. Features are eliminated backwards (starting with all features), motivated by the idea that the feature with the smallest weight has the smallest relevance in the trained model. Eliminating this feature introduces the smallest change in the objective function. Due to the small changes, it is easier to find an optimal solution to the objective function [111].

After the feature ranking, the set \mathcal{R} of ranked features is used to find the minimal number of features with the best performance to reduce the features preventing overfitting. Hyperparameter C is tuned, and an SVM is trained starting with the feature of the highest rank in \mathcal{R} followed by a performance evaluation of the learned classifier. This procedure is step-wise repeated by using the $j = 1, \dots, M$ highest-ranked features. The j highest-ranked features with the best classification performance are chosen to train the final classifier [111].

Like other embedded feature selection methods, SVM-RFE has the advantage that it involves the evaluation of features in the context of the classification task that is to be solved. In combination with SVM as a classifier, SVM-RFE is among the best performing feature selection methods for cancer sample classification, as shown by Haury *et al.* [112]. However, Duan *et al.* [113] have published an extended version of SVM-RFE called Multiple SVM-RFE (MSVM-RFE) that reduces the chance of overfitting.

3.4.2.1 Multiple support vector machine recursive feature elimination

The difference between MSVM-RFE and SVM-RFE lies within the calculation of the feature ranking criterion. In SVM-RFE, the criterion $(w_j)^2$ is obtained by training the SVM on the whole data set one time after hyperparameter tuning. Duan *et al.* argued, even though SVMs integrate regularisation, there is a chance of overfitting for SVM-RFE because the whole data are presented to the SVM. Therefore, the authors decided to calculate the ranking criterion based on $r = 1, \dots, R$ ($R = 100$) SVMs trained on resampled data for each elimination step. Here, the ranking criterion c_j for feature j is defined by the signal-to-noise across the feature weights $(w_{jr})^2$ obtained from R trained SVMs:

$$w'_{jr} = \frac{w_{jr}}{\|\mathbf{w}_r\|}, \quad (3.19)$$

$$\mu_j = \frac{1}{R} \sum_{r=1}^R (w'_{jr})^2, \quad (3.20)$$

$$\sigma_j = \sqrt{\frac{\sum_{r=1}^R (w'_{jr} - \mu_j)^2}{R-1}}, \quad (3.21)$$

$$c_j = \frac{\mu_j}{\sigma_j}, \quad (3.22)$$

Equation 3.19 normalises weights to make them comparable across trained SVMs. Hyperparameter C is tuned per elimination step and used for all resamples [113].

The use of the signal-to-noise ratio of feature weights across trained SVMs derives from the idea to stabilise the estimation of the ranking criterion by being less dependent on the training data set. Applying MSVM-RFE for gene-expression-based tumour sample classification, the authors showed that MSVM-RFE could find solutions that have comparable or better solutions than SVM-RFE [113]. This justifies the usage of MSVM-RFE instead of SVM-RFE for the later-shown feature selection task.

Taken together, the classification of tumours (e.g. for predicting drug response) is a common machine learning task in cancer research and oncology. Gene expression data of tumours present a commonly used molecular data type for finding biomarkers and training predictive classifiers, while SVMs perform well on this data type. The introduced basic concepts of SVMs illustrate how this supervised learning algorithm solves classification tasks. Further, the application of SVMs is not limited to this learning task since SVMs can also be used as an embedded feature selection method (as described for SVM-RFE and MSVM-RFE). Additional to the embedded selection, filter methods based on statistical analyses are another option to select features. However, feature selection itself can be seen as a machine learning task. The following section elaborates on a specific application in which feature selection is the central learning task for obtaining GRNs.

3.5 Inference of gene regulatory networks from gene expression data

In biological terms, a GRN illustrates the interaction networks between TF and their target genes. This interaction is defined by the binding of TFs to the promoter or enhancer regions of a target gene. Via binding to these regions, TFs can regulate the gene expression of the target genes. Changes in a GRN lead to changes in gene expression levels. Therefore, GRNs depict the complex gene regulatory system that controls the cellular processes such as cell differentiation, molecule synthesis, energy transduction, or proliferation [114]. The fundamental role of GRNs in controlling the cellular process is also reflected by the fact that GRNs of cancer cells are different compared to normal cells. These cancer GRNs relate to numerous cancer cell types and disease-associated processes [15].

In cancer research, understanding the state of a GRN in a cancer cell provides direct insights into the gene regulatory system producing the gene expression profile leading to a malignant tumour. Here,

each state corresponds to a different GRN [29]. The studying of GRNs in cancer research refers to the approach of understanding cancer as a biological system defined by complex molecular interactions (the field of cancer systems biology). Here, the application of NGS technologies and the computational analyses of NGS data are essential to derive GRNs [15, 115].

Two NGS technologies are mainly used to obtain GRNs: chromatin immunoprecipitation sequencing (ChIP-seq) and RNA-seq. ChIP-seq comprises the pull-down of chromatin regions that are bound by TFs and DNA sequencing of the pull-down chromatin regions. However, a single ChIP-seq experiment is generally limited to one TF. This characteristic does not easily allow obtaining a genome-wide GRN [114, 116]. Because TF-DNA interactions are directly analysed via ChIP-seq, this technology is often used to validate GRNs that are inferred with other approaches [117]. Additionally, DNA sequences derived from ChIP-seq data can be used to identify the DNA motif bound by a certain TF. To obtain such a motif, a position weight matrix (PWM) is constructed containing the frequency of the four bases at each position of the motif. Once a PWM is constructed, it can be used to predict TF binding sites (TFBS) in DNA sequences without performing ChIP-seq. Therefore, databases containing experimentally obtained TF binding motifs (PWMs) are an important resource for GRN contraction or GRN validation [118].

Contrary to ChIP-seq, high-throughput technologies measuring gene expression such as RNA-seq allow obtaining genome-wide GRNs. Here, a GRN is inferred from gene expression data by analysing expression dependencies between genes. This approach is called reverse engineering of GRNs [114, 116]. Due to the genome-wide range, GRN inference from gene expression data is an attractive approach to construct GRNs. However, constructing a GRN from gene expression is a challenging task in the field of system biology [116]. Because of the complexity of this task, numerous network inference methods have been published. These methods were evaluated by the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project in so-called DREAM challenges [119]. In several DREAM challenges [119, 120], an algorithm developed by Huynh-Thu (called Gene Network Inference with Ensemble of trees – GENIE3) *et al.* [120] turned out to be the best or among the best performing methods. This algorithm is introduced in the following section.

3.5.1 Gene Network Inference with Ensemble of trees - GENIE3

In order to understand how network inference methods can reverse engineer GRNs, it is necessary to know the mathematical representation of a GRN. In mathematical terms, a GRN is a sparse, directed graph. In this graph, nodes represent genes, and edges indicate a regulatory relationship between two genes. Nodes in such a graph are only sparsely connected because of the biological nature of gene regulation, where a TF regulates a limited number of genes. A graph/GRN can be represented by a sparse, square matrix \mathbf{G} , where the number of rows and columns is equal to the number of genes in the network. Each matrix element G_{ji} is an edge between two genes indicating the effect of gene j on gene i . Therefore, effects between genes are directed, and gene j is the regulator (TF) of gene i (target). Besides being directed, a relationship between gene j and i can be signed (or unsigned) and weighted (or unweighted). The sign represents the nature of the regulatory relationship. This relationship can be repressive (negative sign) or activating (positive sign). Thus, a regulator gene j can be a repressor or an activator of gene expression. The weight (the absolute value of the weight for signed values) indicates the strength of a regulatory relationship. Due to the sparsity of matrix \mathbf{G} , most weights are zero indicating no relationship between gene j and i [114, 121]. Such a graph can be learned from a (learning) data set $L = \{\mathbf{x}_k | k = 1, \dots, M\}$, where $\mathbf{x}_k \in \mathbb{R}^N$ is a vector of expression values for N genes measured in the k th entity: $\mathbf{x}_k = (x_{k,1}, \dots, x_{k,N})^T$.

The authors of GENIE3 developed an algorithm that infers a GRN from gene expression data for a directed and unsigned graph with N nodes (genes). The general concept is as follows: the algorithm infers a GRN by exploiting directed expression dependencies between gene j and i to assign an

interaction weight to $G_{ji} \geq 0$, where $j, i = 1, \dots, N$ and large weight values relate to putatively true regulatory links. Besides the general concept, the algorithm is based on several ideas and assumptions to obtain the interaction weights of the graph [120].

The task of learning a graph including N genes is split into N subproblems. Each subproblem is an independent task of identifying the regulators of one given (target) gene by exploiting directed expression dependencies between the given target gene and the remaining input genes. A gene is selected as a regulator when its expression is predictive of (directly influencing) the target gene's expression. The just-described subproblem corresponds to a feature selection problem in supervised machine learning. In order to solve the feature selection problem, the authors assumed that the expression of gene i is given by a function f_i of the remaining input genes' expression, comparable to a multiple regression problem (more than one predictor). However, f_i itself is unknown [120].

To solve this unknown function, the authors decided to apply a random forest of regression trees, a supervised ensemble machine learning method [120]. Regression trees are applied for solving multiple regression problems [122]. A tree represents a learned model that predicts a target variable. Such a tree consists of several tree nodes (not to be confused with nodes of a graph) connected via branches in a tree-like structure. The tree starts with the root node and ends with terminal nodes, the so-called leaves; the remaining tree nodes are internal nodes. A tree node is a binary test that splits the analysed target variable based on a certain criterion. The main concept of a regression tree is to recursively split a target variable at each tree node based on a cut-point of a predictor variable in such a way that the overall variance of the target variable is minimised across the splits. Minimising the variance relates to the concept of finding the least-squares (as in linear regression models) [122].

In the case of GENIE3, a target variable is a vector of expression values of gene i measured across M samples, $\mathbf{x}_i = (x_{k,i} | k = 1, \dots, M)$, and a predictor variable is the expression vector of gene j , $\mathbf{x}_{j,j \neq i}$ [120]. The multiple regression problem for \mathbf{x}_i is solved as follows. Starting at the root node, vector \mathbf{x}_i is binary split into \mathbf{x}_{i_1} and \mathbf{x}_{i_2} by finding a predictor gene j and a cut-point s of \mathbf{x}_j , so

$$\mathbf{x}_{i_1} = (x_{k,i} | k : x_{k,j} < s) \text{ and} \quad (3.23)$$

$$\mathbf{x}_{i_2} = (x_{k,i} | k : x_{k,j} \geq s), \quad (3.24)$$

such that the overall variance is minimised across the two splits,

$$\frac{1}{|\mathbf{x}_i|} \left(\sum_{k: x_{k,j} < s} (x_{k,i} - \overline{\mathbf{x}_{i_1}})^2 + \sum_{k: x_{k,j} \geq s} (x_{k,i} - \overline{\mathbf{x}_{i_2}})^2 \right) = \frac{|\mathbf{x}_{i_1}|}{|\mathbf{x}_i|} \text{Var}(\mathbf{x}_{i_1}) + \frac{|\mathbf{x}_{i_2}|}{|\mathbf{x}_i|} \text{Var}(\mathbf{x}_{i_2}), \quad (3.25)$$

where $|\mathbf{x}_{i_1}|$ and $|\mathbf{x}_{i_2}|$ are the number of entities (cardinality) in each split, and $\overline{\mathbf{x}_{i_1}}$ and $\overline{\mathbf{x}_{i_2}}$ are the average in each split. Taken together, it is sought to find a value of j and s minimising Equation 3.25. The splits derived from the root node represent two new tree nodes. Each of the new nodes split \mathbf{x}_{i_z} again by finding the best j and s . This procedure is repeated until \mathbf{x}_{i_z} contains only one entity defining a leaf. At each tree node, j and s can be different [122].

However, for GENIE3, not only one but an ensemble of regression trees was trained to solve the regression problem. In ensemble learning, multiple, diverse predictive models are trained on the input data. The diversity of the models is obtained by resampling the input data (entities or features, or both). Each model on its own is weak in making predictions, but combining the predictions of all models provides a strong predictive ensemble model. Random forest is such an ensemble method [72]. In a random forest of regression trees, each tree is trained on a different bootstrap sample of the target variable \mathbf{x}_i (bootstrapping of entities) and at each tree node, R potential predictor variables are randomly selected and evaluated to find the best split. For GENIE3, the default parameters are $R = \sqrt{N}$ (N =all input genes) and 1000 trees per random forest, which are standard values of these two parameters for random forest. One random forest solves f_i corresponding to one subproblem i [120].

The authors of GENIE3 used regression trees because this method solves the function f_i without knowing the function's nature and can deal with non-linear relationships. Additionally, regression

3 Computational analysis of cancer transcriptomes

trees allow a ranking of genes used as a predictor. Here, genes are ranked by their importance I to predict the expression of the target gene i in the learned model. Assuming gene j was used for the split of \mathbf{x}_{i_z} into $\mathbf{x}_{i_{z+1}}$ and $\mathbf{x}_{i_{z+2}}$ at tree node \mathcal{N} , then the importance measure would be

$$I_j(\mathcal{N}) = |\mathbf{x}_{i_z}| \text{Var}(\mathbf{x}_{i_z}) - |\mathbf{x}_{i_{z+1}}| \text{Var}(\mathbf{x}_{i_{z+1}}) - |\mathbf{x}_{i_{z+2}}| \text{Var}(\mathbf{x}_{i_{z+2}}). \quad (3.26)$$

This equation means that the importance of a predictor gene is evaluated by the variance reduction due to the split at node \mathcal{N} . To calculate the importance of gene j for a single tree, I is summed across all tree nodes at which gene j was used for the split. The importance measure of gene j as a predictor of gene i can be extended to the ensemble of regression trees. Here, the importance measures per tree for gene j are averaged across all trees. The important measure that was obtained across the ensemble is equivalent to the interaction weight G_{ji} in the graph of the GRN. By calculating the importance measure for each gene j that was used to split a node in the ensemble of trees, a ranking of potential regulators of gene i can be achieved from all interaction weights $G_{ji} | j = 1, \dots, N$ [120].

Although GENIE3 provides a ranking of potential regulators for each gene based on interaction weights, the algorithm does not provide the final selection of regulators. To obtain such a final selection, a threshold value for the interaction weights is necessary. However, this threshold value needs to be chosen manually. This choice has to be done with caution because the threshold decides about the number of false positive (wrong edges) and false negative (missing edges) regulatory links within the final GRN. Additionally, even though all input genes are considered to be potential regulators when applying GENIE3, not all of these genes are TFs (a point that will be discussed in detail within a later presented study). A potential approach is to use databases of gene function annotations such as GO terms to identify TFs among the input genes and obtain the final GRN, as proposed by Cahan *et al.* [123].

Overall, GENIE3 is a useful algorithm to infer GRNs from gene expression data. Once a GRN and the related graph are obtained, several downstream analyses can be applied to gain more information out of the GRN.

3.5.2 Downstream analysis of gene regulatory networks

Biological networks such as GRNs generally have a modular structure defined by communities of genes. In a network, each community is a group of genes that are highly connected but less/rarely connect with genes outside of the group. These communities are also called modules or subnetworks. The community structure has a biological reason since gene products involved in the same process have many interactions and need to be expressed simultaneously. Otherwise, components of this process would be missing. Thus, genes associated with a common process are co-regulated by one or several potentially cooperating TFs. Due to the association of gene communities in GRNs with biological processes, the detection of such subnetworks can help to understand the biological functions of a GRN in the cell [124]. In terms of cancer biology, communities can relate to certain tumour subtypes or specific disease-related processes. Special algorithms detect such communities. One of these algorithms is a method called "map equation" published by Rosvall and Bergstrom [125]. This algorithm uses the flow of random walk in a network to identify communities. This method is not limited to biological networks and can generally be used to detect communities in networks.

Another important aspect of GRN analysis is determining the contribution of single regulatory genes (nodes) to the network. This contribution is frequently determined by the number of connections of a gene to other genes within the network. The number of connections is the degree of a node in a network. In directed graphs such as GRNs, the out-degree of a node is the number of target genes regulated by a TF node, and the in-degree of a node is the number of TFs regulating a gene. TFs with high out-degree are important regulators that influence the expression of many genes within the network [124, 126].

Cahan *et al.* proposed an approach called Network Influence Score (NIS) that extends the out-degree to determine the contribution of a TF to the network [123]. The authors used gene expression data comprising a wide collection of different cell/tissue types to infer GRNs. The NIS was applied to identify TFs with the strongest impact on cell-/tissue type-specific gene expression. The NIS integrates the out-degree of a TF and the level of expression change of the TF and target genes in a particular cell/tissue type. The level of expression change adds more information because the out-degree only indicates the number of targets but not in which range a TF can change the expression of a target gene. The authors estimated the level of expression change by scaling expression values gene-wise applying z-score normalisation across the sample collection of different cell/tissue types. Due to the z-score normalisation, the expression of a gene in any cell/tissue type is positioned to the expression distribution across the collected data set.

Taken together, GRNs depict the complex gene regulatory system that controls the cellular processes, while distinct GRNs define cancer cells. By applying GENIE3 — which is based on an ensemble machine learning algorithm — a weighted, unsigned, directed graph of a GRN can be inferred from expression data. The interaction weights provided by GENIE3 allow a ranking of the potential regulators of a gene. This facilitates the selection of putative regulators by manually choosing a weight threshold. To annotate TFs among the input genes, functional annotation databases can be used. Downstream analyses can extract information such as community structures of a GRN and the influence of individual TF within the network.

Introducing SVMs and GENIE3 illustrated how machine learning methods can solve specific tasks for the computational analysis of gene expression data. However, computational analyses of molecular data cannot always be solved by applying a single machine learning algorithm and require several methods and the integration of different data resources. An example of such an analysis is the computational characterisation of lnc genes.

3.6 Computational characterisation of lnc genes

Even though there is clear proof that lnc genes and their lncRNA products can have tumour promoting and suppressing functions by regulating cancer-relevant processes, the extent of lnc genes' involvement in cancer formation or suppression is still an open question. The main reason is that the function of most lnc genes is unknown due to a challenging functional characterisation and annotation of lnc genes. These challenges arise because lnc genes and lncRNA have specific features distinct from protein-coding genes and their product (see Section 2.5). One of these features is the low sequence conservation of lnc genes compared to protein-coding genes. This low conservation prevents a classification based on sequence-related functional domains as done for protein. Lnc genes are lower conserved because lncRNAs rather form structure-related (secondary and tertiary structures) functional domains than sequence-related domains [127, 128]. These two and additional attributes challenge the experimental characterisation of lnc genes [129]. Therefore, computational characterisation of lnc genes has been an important factor in analysing lnc genes. Additionally, computational high-throughput methods appear appropriate for this task because a higher number of lnc genes can be analysed in parallel [20, 130]. There are many approaches regarding how computational analyses can be used for lnc gene characterisation [130, 131]. However, applying all of them is not feasible. The following section introduces two approaches that will be applied within this thesis.

3.6.1 Characterisation lnc genes based on their genomic organisation with coding genes

One approach of lnc gene characterisation is the classification of lnc genes into certain types based on their genomic position relative to coding genes. These types include sense overlapping, sense intronic,

antisense, divergent, and intergenic (Table 3.6.1) [131, 132]. This thesis concentrates on the last three types (since sense overlapping and intronic lnc genes can be difficult to analyse due to the overlap with coding genes on the same strand). Additionally, via this characterisation, a coding gene in positional relationship to a lnc gene can be assigned as a coding partner to this lnc gene (e.g. the coding gene in divergent orientation to the lnc gene), defining the direct coding gene neighbourhood of lnc genes.

Table 3.3: Types of lnc genes based on the genomic position relative to coding genes [131, 132].

Type	Description
sense overlapping	a lnc and coding gene overlap on the same strand by at least one exon
sense intronic	exons of a lnc gene overlap with introns of a coding gene on the same strand
antisense	a lnc gene locus overlaps with a coding genes locus on the opposite strand
divergent	a lnc and coding gene are on opposite strands and the TSSs of both genes are in proximity; termed as diverged/bidirectional transcription
intergenic	a lnc falls between two coding genes independent of the strand

The positional classification of lnc genes arises from the knowledge that the organisation of the genome is not random [131]. Indeed, several publications have shown that divergent, antisense, and intergenic lnc genes have specific features [133–136]. For example, the direct coding gene neighbourhood of divergent and intergenic lnc genes is enriched for different functions than the neighbourhood of antisense lnc genes [135]. Additionally, divergent lnc genes have unique properties. The TSS of lnc genes is located in the same nucleosome-free region or promoter as the TSS of the coding gene partner [136].

A published work by Hon *et al.* focused on improving the positional classification of lnc genes by completing the annotation of 5' ends of lnc genes. The motivation behind this work was to obtain more complete gene models for lnc genes. Gene models for lnc genes are frequently incomplete because of lower expression and higher exosome sensitivity of lnc genes compared to coding genes [133]. The authors used cap analysis of gene expression (CAGE) data. CAGE relates to a protocol where only the 5' end of RNAs is sequenced, allowing expression quantification of individual transcription start sites (TSS) and genes by pooling start sides of one gene [133, 137]. The obtained lnc gene models and annotations of the lnc gene types divergent, antisense, and intergenic were published under the name FANTOM CAT (FANTOM CAGE-associated transcriptome), which was done within the framework of FANTOM (Functional ANnotation Of the Mammalian genome) [133].

The positional classification of lnc genes allows annotation of known, type-related properties to a lnc gene (like a shared nucleosome-free region for divergent genes). However, this classification is always limited to the genomic locus of a gene. Another approach takes advantage of gene expression data to infer putative lnc gene functions.

3.6.2 Inference of putative lnc gene functions from gene expression profiles

As introduced above, the control of gene expression is essential to orchestrate the biological processes within a cell. Therefore, gene expression is non-random, and the analysis of gene expression profiles can provide much information about genes. A fact that is frequently used for the computational characterisation of lnc genes [130].

The information in which tissue a lnc gene is expressed can imply potential gene functions since tissues fulfil different tasks in the human body. Gene expression data from different tissues can be used to identify tissue-specific expression of lnc genes by performing DGEA across the tissues [130]. Additionally, by comparing tumours against normal, matching tissue, it can be determined whether a lnc gene's expression is altered in cancer, which could indicate cancer-associated functions [41].

Co-regulation is often observed for genes that are involved in the same biological process. This co-regulation and the extensive functional annotation of coding genes can be used for the computational characterisation of lnc genes by performing a co-expression analysis integrating lnc and coding genes [130, 138]. There are two common ways to perform this analysis: co-expression clustering and gene-centred co-expression analysis [41, 130, 138, 139].

Co-expression clustering — an unsupervised machine learning task — is the identification of gene clusters. Each cluster is a group of strongly expression-correlated genes due to co-regulation, whereas the expression correlation across clusters is weak. For lnc characterisation, the co-expression clustering step includes coding and lnc genes. After obtaining the co-expression clusters, a functional overrepresentation analysis is performed per gene cluster. This analysis is possible because of the extensive functional annotations of coding genes in databases. Putative functional associations of a lnc gene can be inferred from the enriched biological processes of the co-expression cluster that is assigned to this lnc gene. This type of functional inference is called guilty-by-association since a putative function is determined by finding associations in terms of co-expression with functionally annotated genes [130, 138].

Co-expression clustering is practical for capturing positive expression correlations but not for negative correlations [138, 140]. However, a gene-centred co-expression analysis can capture positive and negative expression correlations between genes. Here, a single lnc gene is compared to all expressed coding genes to identify correlated genes using statistical measures like Pearson and Spearman correlation. The obtained correlation values can be used to perform a rank-based functional gene set enrichment analysis [139] or by selecting the significant or strongest negatively and positively correlated coding genes. In the second version, the overrepresentation analysis is separately applied to the selected negatively and positively correlated coding genes.

There is also an analysis method that is partially linked to co-expression. This characterisation method refers to the identification of the DNA binding site of lnc genes that regulate gene expression via triplex formation. There are two different ways to identify lnc gene binding sites. The first option is the conduction of ChIP-seq-like experiments [43]. The second option is a computational method that uses only sequence information and the Hoogsteen base-pairing model to identify binding sites of lnc genes [36]. However, the ChIP-seq-like experiments can also be used to obtain a PWM of the lnc gene binding motif. By predicting or experimentally obtaining lnc gene binding sites, putative targets of lnc genes can be identified and functional enrichments among targets indicate the processes regulated by the lnc gene.

The expression data of lnc genes can also be overlaid with clinical and other tumour-related molecular data. An overlay of gene expression and copy number data of lnc genes can be used to evaluate whether a gene is dysregulated due to genomic alterations in cancer. A copy-number-dependent expression of lnc genes can indicate cancer-associated functions. The detection of associations between lnc gene expression and clinical outcome, such as patient survival, can also point to lnc genes with putative cancer-associated functions [41].

Taken together, lnc genes have specific features that complicate their functional characterisation. Computational analyses provide several approaches for lnc gene characterisation including a classification based on the genomic position relative to coding genes, the analysis of gene expression profiles and co-expression, and the overlay of gene expression with tumour-related molecular and clinical data.

3.7 Summary and outlook

The last two chapters summarised molecular biological aspects of cancer development and the computational analysis of cancer transcriptome. Introduced molecular biological aspects of cancer included:

- the process of cancer development that is defined by the dysregulation of gene expression,
- hallmarks of cancers summarising capabilities obtained by malignant tumours,

3 Computational analysis of cancer transcriptomes

- the implication of lnc genes in cancer besides protein-coding genes, and
- the application of omics technologies for the molecular analysis of tumours with an emphasis on RNA-seq.

The introduction of the computational analysis of cancer comprised:

- processing of RNA-seq data to determine gene expression values,
- an outline of frequently applied cancer transcriptome analysis,
- the supervised machine learning algorithm SVM that can be used for classification and embedded feature selection via MSVM-RFE,
- the algorithm GENIE3 that infers GRNs from gene expression data, and
- the computational characterisation of lnc genes.

Each of the following two chapters presents a cancer transcriptome study that focuses on the application of the just-introduced computational analysis methods. The first study (Chapter 4) presents the application of SVM and MSVM-RFE for the construction of a gene-expression-based classifier that can predict the treatment outcomes of a targeted therapy in colorectal cancer patients. The second study (Chapter 5) presents the dissection of molecular heterogeneity in medulloblastoma, the analysis of GRNs related to this heterogeneity, and the computational characterisation of lnc genes that are differentially expressed between distinct molecular groups of medulloblastoma.

4 Colorectal carcinoma study

For copy right reasons, the pages 39 - 54 were removed from the online version.
Please see the original publication: <https://doi.org/10.1038/ncomms14262>.

5 Medulloblastoma study

Medulloblastoma (MB), a tumour of the cerebellum, is the most common malignant brain tumour among children [193]. The PedBrain Tumour Research Project was founded as part of the International Cancer Genome Consortium (ICGC) to gain further insights into the molecular biology of MB and other paediatric brain tumours. The ICGC PedBrain Project consists of a research consortium (coordinator: Prof. Dr. Peter Lichter) with several involved institutions (German Cancer Research Center, European Molecular Biology Laboratory, National Center for Tumor Diseases, Düsseldorf University, Heidelberg University, Heidelberg University Hospital, Max Planck Institute for Molecular Genetics, The Hospital for Sick Children in Toronto, and Arthur and Sonia Labatt Brain Tumour Research Center). As a consortium partner, we, the research group “Gene Regulation & System Biology of Cancer” located at the Max Planck Institute for Molecular Genetics, performed RNA sequencing of 164 MB samples and downstream analysis of the RNA-seq data.

The following chapter comprises the second study within this dissertation covering a comprehensive analysis of the MB transcriptome by using the MB RNA-seq cohort of ICGC PedBrain.

5.1 Biological and theoretical background of medulloblastoma and Inc gene *MEG3*

5.1.1 Medulloblastoma: Tumour of the cerebellum

Medulloblastomas represent malignant lesions of the cerebellum. Medulloblastoma has an embryonic origin and arises from progenitor cell populations [193]. MB accounts for most cases among malignant brain tumours during childhood but can occur into adulthood; however, MB is more likely to occur at an earlier age. Currently, the most common therapy options include safe resection, chemotherapy, and craniospinal radiation (radiation along the head-spine axis). Approximately 30% of the medulloblastoma patients die from this disease [83, 193]. Most survivors experience long-term side effects due to the treatment, including developmental, neurological, neuroendocrine, and psychosocial deficits [83]. In order to better understand the origins of MB, the following section will summarise the histology and development of the cerebellum.

5.1.2 Development and structure of the cerebellum

In humans, the cerebellum is the largest part of the hindbrain. It is located underneath the cerebrum lying dorsal (rear side) to the pons and medulla and is joint to the brainstem (Figure 5.1.a) [194]. The cerebellum is involved in sensory-motor processing and non-motor functions (e.g. emotional and cognitive processes) containing over half of the terminally differentiated neurons in the adult brain [195, 196]. The matured cerebellum is composed of the cerebellar nucleus, white matter, and three outer layers, each layer containing different cell types (Figure 5.1.b) [196]. The outermost layer is called the molecular layer and contains stellate and basket cells. Purkinje cells, candelabrum cells, and Bergmann glia are located in the Purkinje cell layer, which lies between the molecular and granular layer. The granular layer contains granule cells, Golgi cells, unipolar brush cells, and Lugaro cells [195, 197]. Based on their released neurotransmitters, the neuronal cell types can be grouped into either inhibitory gamma-butyric acid (GABAergic) or excitatory glutamatergic neurons. Purkinje, stellate,

5 Medulloblastoma study

basket, Lugaro, and Golgi cells are inhibitory GABAergic neurons. Granular and unipolar brush cells are excitatory glutamatergic neurons [195]. Bergmann glia are non-neural glial cells. The major cell types are Purkinje cells, granule cells, and Bergmann glia in the cerebellum [198].

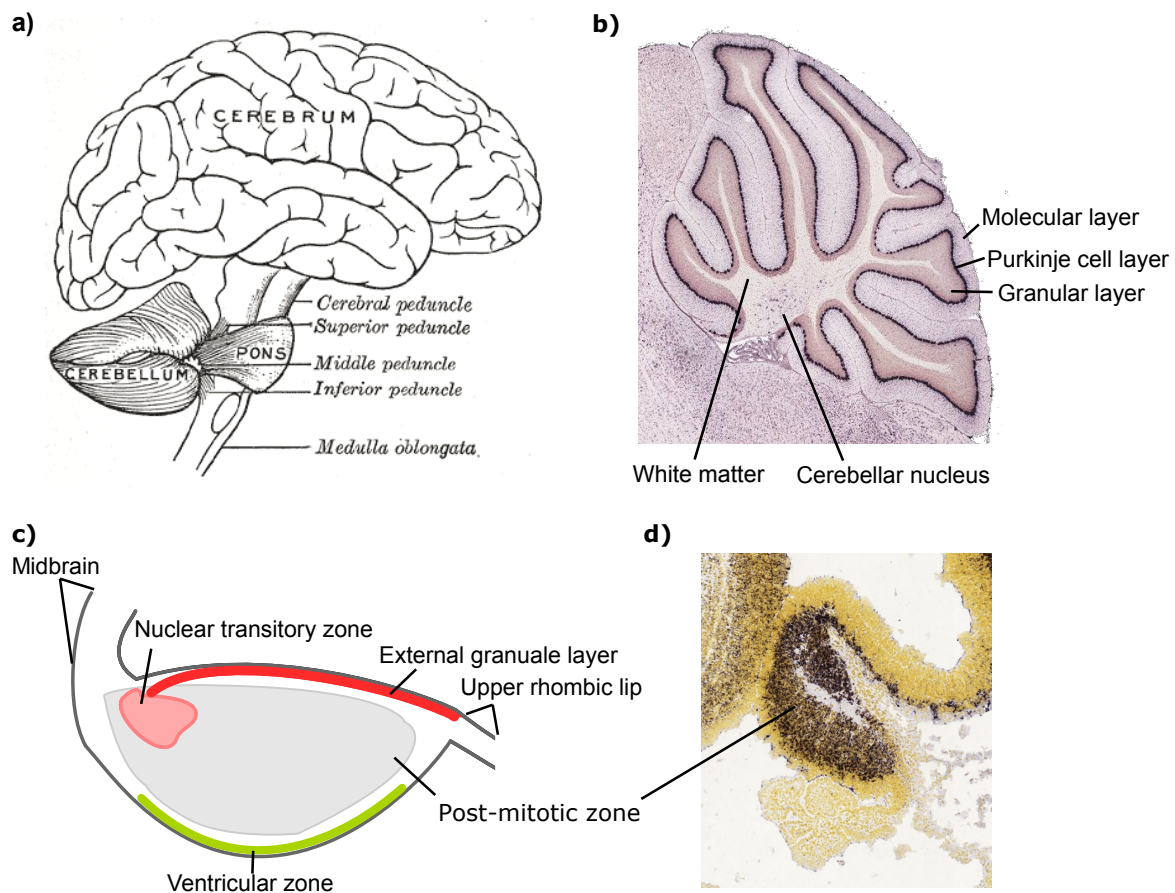


Figure 5.1: Anatomy, histology, and development of the cerebellum. **a)** Human adult brain. Reprint (public domain) from Henry Gray (1918) *Anatomy of the Human Body* [199]. Source: www.bartleby.com; Gray's Anatomy; FIG. 677. **b)** Mouse adult cerebellum sagittal (age: P56). ISH of the Purkinje cell marker *Calb1* [200]. Image credit for ISH: Allen Institute [201]. Labels were added to the image. **c)** Schematic illustration of the developing cerebellum in mice around E13.5 [196] [195]. **d)** Mouse cerebellum sagittal E13.5. ISH of *Tubb3* marking post-mitotic cells [202]. Image credit for ISH: Allen Institute [201]. Labels were added to the image.

The layered structure of the cerebellum arises from a complex cascade of cell migration during development. This developmental process will be explained using mice as a model organism. Around embryonic day (E)8, the cerebellar anlage (also called cerebellar primordium) emerges close to the midbrain-hindbrain boundary, called the isthmus [195, 196]. On a molecular level, the genes *Otx2*, *Gbx1*, *Fgf8*, *En1*, *En2*, *Wnt1*, and *Pax2* play an important role in the formation of the cerebellar anlage [195].

Atoh1-positive (alias *Math1*) precursors of glutamatergic neurons (granule and unipolar brush cells) arise from the upper rhombic lip. Around E13, these precursors start to migrate into the external granular layer (EGL) and the nuclear transitory zone (NTZ) (Figure 5.1.c), which will later form the cerebellar nucleus [195]. *Ptf1a*-positive precursors in the ventricular zone (VZ) give rise to GABAergic neurons of the cerebellum (Purkinje, stellate, basket, Lugaro, and Golgi cells) (Figure 5.1.c). Bergmann glia also arise from the VZ [195]. Purkinje cells emerge around E10–E13 [195]. After E13, post-mitotic

Purkinje cells start to migrate from the VZ into the direction of the EGL toward the middle of the cerebellar anlage forming a several-cells-thick Purkinje cell cluster in the post-mitotic zone (Figure 5.1.c-d) [195]. Around E18.5, Purkinje cells start to secrete Shh protein that serves as (neuro-)trophic factor by promoting the proliferation of granule precursors in the EGL (transit amplification). Here, the proliferation of granule precursors depends not only on Shh-signalling but also on Atoh1 expression [195, 196]. Atoh1 keeps granule precursors in a proliferative state responsive to Shh signalling by suppressing differentiation [196, 203]. Cell proliferation is initiated by the binding of the Shh protein to its receptor Patched 1 (Ptc1) that causes the release of the G-coupled receptor Smoothed (Smo) from Ptc1 suppression. Activated Smo inhibits Suppressor of Fused (SUFU) that binds and suppresses Gli-family transcription factors. Released Gli transcription factors translocate into the nucleus and transcriptionally activate proliferation-promoting target genes [204]. During transit amplification, the EGL can be separated into an outer EGL comprising proliferating granule precursors and an inner EGL containing post-mitotic differentiating granule cells [195]. The post-mitotic cells of the inner EGL migrate inward across the Purkinje cell cluster to form the inner granular layer (IGL) that later will mature into the granular layer. The proliferation of granule precursors, cell migration, and formation of the granular layer continues during the first weeks after birth — the EGL will disappear at the end of this time. The Purkinje cell cluster starts to form a mono-cell layer after birth due to Reelin (Reln) secretion by granule cells and precursors [195]. Stellate, basket and Golgi cells are derived postnatally, in which stellate and basket cells migrate from the cerebellar nucleus across the granule and Purkinje cell layer into the molecular layer [195, 205].

The cellular-heterogeneity of the cerebellum partially reflects the heterogeneity of MB because the cell of origin, including cell type and differentiation stage, is thought to be different or only partly overlapping between four MB consensus molecular subgroups [193]. However, numerous additional aspects contribute to the MB subgroups, which will be introduced in the following section.

5.1.3 Subgroups in medulloblastoma

Four different histological types are described for MB, based on histo-pathological patterns observed in H&E-stained tissue sections (hematoxylin and eosin): classic, desmoplastic (or nodular), large cell/anaplastic (LCA), and MB with extensive nodularity (MBEN) [206]. The classic histology type is the most common. Patients with LCA have a poor prognosis when this phenotype is not restricted to a small part of the tumour. Tumours with MBEN histology mainly occur in infants and have an excellent prognosis.

High-throughput cDNA microarray technologies have allowed the characterisation of medulloblastoma beyond histology. During these first attempts to define molecular subgroups in MB based on gene expression profiling using microarrays, researchers reported four, five or six molecular subgroups [82, 207–209]. Based on these studies, an international joint venture of researchers defined a consensus of four subgroups: WNT, SHH, Group 3, and Group 4 [210]. These four subgroups show distinct clinical and molecular features including CNVs, chromosomal rearrangements, mutations, DNA methylation, and gene expression (Table 5.1, 5.2) [83, 193].

WNT subgroup. 10% of medulloblastomas belong to the WNT subgroup, the rarest MB subgroup. WNT MBs have an excellent outcome with a five-year survival > 95% in paediatric cases. Progenitor cells in the lower rhombic lip of the developing brainstem are supposed to be the origin of WNT tumours. The β -catenin encoding proto-oncogene *CTNNB1* is mutated in 90% of WNT MBs, which causes activation of the Wnt pathway as reflected by the gene expression profile of this subgroup. The Wnt pathway transcriptionally activates downstream targets that promote the proliferation of the tumour cells. Additional recurrent mutations are far less frequent (*DDX3X* (36%), *SMARCA4* (19%), *TP53* (14%), *CSNK2B* (14%), *PIK3CA* (11%), and *EPHA7* (8%)). Chromosomal aberrations in WNT MBs are typically limited to monosomy 6 in 80%–85% of cases [83, 193].

Table 5.1: Clinical features of medulloblastoma subgroups.

Subgroups	WNT	SHH	Group 3	Group 4
% of cases	10	30	25	35
Age at diagnosis	Children, adults	Infants, children, adults	infants, children	infants, children
Histology	Classic, rarely LCA	Classic > desmoplastic > LCA > MBEN	Classic, LCA	Classic, rarely LCA
Metastasis at diagnosis (%)	5-10	15-20	40-45	35-40
Prognosis	Very good	Infants good, others intermediate	Poor	intermediate

Note: Infant: 0-3 years. Children: 4-17 years. Adults: > 17 years. Table adapted from Juraschka *et al.* [193]. Histology information added from [83].

SHH subgroup. The SHH subgroup accounts for approximately 30% of all MBs and occurs mostly in infants and adults. The outcome in this subgroup depends on several factors. One factor is the patient's age. Younger patients show a worse prognosis compared to adults. Additional factors of poor prognosis are metastasis at diagnosis, amplification of proto-oncogene *MYCN*, and mutations of the tumour suppressor *TP53*. Granule cell precursors of the EGL are thought to be the cell of origin for this subgroup [193]. SHH MBs typically carry germline mutations, somatic mutations, or copy number alterations that activate tumour-driving *Shh* signalling and, therefore, express a strong *Shh* signalling signature. The activated *Shh* signalling promotes the proliferation of SHH tumours [83, 193]. This *Shh* signature includes the upregulation of several TFs such as *Shh*-signalling-involved *GLI1/ GLI2* [82], the *Shh*-signalling-associated *ATOX1* [203], and the *Shh*-target *SOX2* [211]. Germline mutated genes include *PTCH1*, *SUFU*, and *SMO*. Two recurrently deleted or mutated negative regulators of the *Shh* pathway are *PTCH1* (43%) and *SUFU* (10%). Additionally, the *Shh* pathway and target genes are hit by activating mutations (*SMO* - 9%) or amplification (*GLI1* or *GLI2* - 9%; *MYCN* - 7%). Amplification of the *MYCN* paralog *MYCL* is rarely detected in SHH tumours. Alterations in TP53 and PI3K-AKT pathway occur in 9.4% and 10% of SHH MBs, respectively, which are also involved in tumorigenesis of this subgroup [193]. The activation of the PI3K-AKT is predominantly observed among adult SHH cases [212]. The promoter of telomerase reverse transcriptase (*TERT*) is mutated in 39% of all SHH MB cases. Loss of chromosome 9q (including *PTCH1*) and 10q (including *SUFU*) are the most recurrent cytogenetic events in the SHH subgroup [83, 193, 213]. Additionally, SHH tumours show differences that are related to age groups (infants, children, and adults) and associated with distinct gene expression and genomic profiles as well as clinical features [212, 214].

Group 3 subgroup. Group 3 tumours account for 25% of all MB cases. This subgroup is associated with the worst prognosis among all MB subgroups (five-year survival \leq 60%) and frequent metastasis at diagnosis (40%–45%). Group 3 tumours are thought to derive from neural stem cells. Group 3 and Group 4 are less defined by a specific activated pathway that drives the tumorigenesis such as in the subgroups WNT and SHH. However, *MYC* amplifications are recurrent in Group 3 MBs (17%) and are frequently associated with *PVT1-MYC* fusions [193]. The *PVT1-MYC* fusion causes an auto-activation of *MYC* through an *MYC* binding site in the *PVT1* promoter [213]. The subset of *MYC*-driven tumours shows the worst survival in MB and a distinct expression profile among Group 3 MB [208]. Besides an *MYC* signature, Group 3 MBs express a photoreceptor signature that is potentially induced via the upregulation of the TFs *NRL* and *CRX* [83, 207]; both TFs are involved in photoreceptor differentiation [215]. Recurrent SNVs can be found only in 5% of the Group 3 tumours comprising single or multiple mutations among *SMARCA4*, *KBTBD4*, *CTDNEP1*, and *KMT2D*. *MYCN* and *OTX2* amplification are

5.1 Biological and theoretical background of medulloblastoma and *Inc* gene *MEG3*

Table 5.2: Molecular features of medulloblastoma subgroups.

Subgroups	WNT	SHH	Group 3	Group 4
Proposed cell of origin	Progenitor cells in the lower rhombic lip	Granule precursors of the external granule layer	neural stem cells	Unknown
Recurrent gene amplification	-	<i>MYCN</i> , <i>GLI1</i> or <i>GLI2</i>	<i>MYC</i> , <i>MYCN</i> , <i>OTX2</i>	<i>SNCAIP</i> , <i>MYCN</i> , <i>OTX2</i> , <i>CDK6</i>
Recurrent SNVs	<i>CTNNB1</i> , <i>DDX3X</i> , <i>SMARCA4</i> , <i>TP53</i>	<i>PTCH1</i> , <i>TERT</i> , <i>SUFU</i> , <i>SMO</i> , <i>TP53</i>	<i>SMARCA4</i> , <i>KBTBD4</i> , <i>CTDNEP1</i> , <i>KMT2D</i>	<i>KDM6A</i> , <i>ZMYM3</i> , <i>KMT2C</i> , <i>KBTBD4</i>
Cytogenetic events (Gain)	-	3q, 9p	1q, 7, 18	7, 18q
Cytogenetic events (Loss)	6	9q, 10q, 17q	8, 10q, 11, 16q	8, 11p, X
Cytogenetic events (others)	-	-	i17q	i17q
Other recurrent events	-	-	<i>GFI1</i> and <i>GFI1B</i> enhancer hijacking	<i>PRDM6</i> <i>GFI1</i> , and <i>GFI1B</i> enhancer hijacking
Expression signature	Wnt-signalling	Shh-signalling	MYC and photoreceptor signature	Neuronal signature

Note: Table adapted from Juraschka *et al.* [193]. Expression signature information derived from [83].

identified in only 5% and 3% of cases, respectively. More frequent than SNVs is the upregulation of the oncogenes *GFI1* and *GFI1B* through enhancer hijacking in Group 3 tumours (15%–20%) and also a potential driver event. In Group 3 and Group 4 tumours, chromosomal rearrangements are far more frequent compared to WNT and SHH tumours. In Group 3, chromosomal changes include isochromosome 17q and loss of chromosomes 8, 10q, and 16q and gain of 1q, 7, and 18 [83, 193].

Group 4 subgroup. 35%-40% of medulloblastoma are classified as Group 4 and mainly diagnosed during childhood and adolescence. This subgroup shows an intermediate outcome with frequent metastasis at diagnosis (35%–40%) [193]. Overall, Group 4 MBs express a neuronal/neuronal-developmental profile. The cell of origin for Group 4 MB is not clear. Somatic mutations are rare events in Group 4, whereby *KDM6A*, *ZMYM3*, *KMT2C*, and *KBTBD4* are the most recurrent mutated genes (6%-9% of Group 4 cases). *GFI1* and *GFI1B* overexpression due to enhancer hijacking as well as *MYCN* and *OTX2* amplification are recurrent events that drive tumorigenesis in Group 4 and in Group 3 MB. Additionally, *CDK6* amplification (6%) and enhancer hijacking-induced *PRDM6* overexpression (17%) can frequently be found in this subgroup [193]. The gain of chromosomes 7 and 18q, isochromosome 17q, and loss of 8q, 8p, 11p, and X are common cytogenetic events in this subgroup [83, 193].

The four subgroups of medulloblastoma are accepted consensus clinical groups and recognised in the 2016 WHO classification of tumours of the central nervous system [193, 210]. Currently, gene expression and DNA methylation data are applied for subgroup classification in MB [193, 216, 217]. Besides the genetic, molecular, and clinical characteristics of the four subgroups, the implication of enhancers in the regulation of subgroup-specific gene expression was also studied in MB.

5.1.4 Enhancer-mediated gene regulation in medulloblastoma

Enhancers are *cis*-regulatory elements that are essential for tissue-/cell-specific expression. Enhancers regulate gene transcription via interaction with nearby promoters facilitated by DNA loops (see Section 2.2). In the framework of ICGC PedBrain, Lin *et al.* inferred enhancer-mediated GRNs to gain insights into MB subgroup-specific transcription [218].

Lin *et al.* analysed RNA-seq and H3K27ac ChIP-Seq data of 28 primary MB tumours. Here, we provided the RNA-seq data. H3K27ac histone marks were used to define active enhancer regions in MB, while excluding H3K27ac peaks close to transcription start sites (± 1 kb around the TSS). The H3K27ac signals in the enhancer regions were then used to define subgroup-specific active enhancers as well as common ones across subgroups. RNA-seq data were utilised to calculate gene expression values and to identify subgroup-specifically expressed genes. The sets of subgroup-specifically active enhancers and expressed genes were matched to infer putative enhancer targets. When gene and enhancer were located in the same TAD and when enhancer activity and gene expression were significantly positively correlated, a gene was assigned to an enhancer as a target, leading to infer a enhancer-mediated gene regulatory network. These inferred GRN contained subgroup-specifically expressed TFs as regulators and their putative target coding genes. Assignments between TFs and putative target genes were created using enhancers as mediators. When an enhancer showed enrichment of binding sites for a certain TF, the target genes of the enhancer were defined as targets of the TF. The regulatory impact of a TF was estimated based on the outdegree (the number of targets). These data were also used to identify TFs that showed binding site enrichment in WNT-, SHH-, Group 3-, Group 4-, WNT-SHH-, or Group3-Group4-specific enhancers [218]. Chromatin interactions were validated by 4C-seq for a subset of enhancer-gene assignments (a sequencing protocol that allows identification of chromatin interactions of a targeted genome region) [218, 219]. Herby, the authors confirmed the interaction of a Group 3-specific enhancer with the promoter of *TGFBR1* [218].

Furthermore, Lin *et al.* defined subgroup-specific super-enhancer. Super-enhancers are broad regions of spatially co-localised enhancer domains that have been shown to regulate genes involved in oncogenesis, maintenance of tumour cell identity, and cell-type-specific functions [218]. Overall, 92 TFs were associated with subgroup-specific super-enhancers. Among the super-enhancer associated TFs was also *LMX1A* that regulates the cell-fate of cells in the upper rhombic lip and is involved in the development of the cerebellum [218]. Additionally, *Lmx1a* is expressed in the NTZ of the cerebellum during mouse development. Lin *et al.* reported that *LMX1A* is specifically upregulated in Group 4 and, therefore, emphasised that *LMX1A* is a master regulator in Group 4 tumours and that cells of the upper rhombic lip and the NTZ could be the cell-of-origin of Group 4 tumours [218].

5.1.5 Kinase activity profiles in medulloblastoma

Most MB studies concentrated on the characterisation and profiling of mutations, cytogenetics, gene expression, and DNA methylation. However, the activity of kinases in MB was also studied. Zomerman *et al.* [220] profiled the kinase activity in MB using PamChips (peptide microarrays that allow profiling of tyrosine and serine/threonine kinases activity). The authors identified two major protein-signalling clusters in 50 MB samples including 13 SHH, 16 Group 3, and 19 Group 4 tumours (this paper did not include WNT MBs). Cluster-1 was associated with an MYC-like kinase activity profile similar to a kinase activity in hTERT immortalized retinal pigmented epithelial (RPE-1) cells when *MYC* or *MYCN* was overexpressed. Cluster 1 was active in all SHH samples without exception, in most Group 3, and in a small fraction of Group 4 tumours. Cluster-2 was associated with a neuronal differentiation expression signature and active in most Group 4 and in a minority of Group 3 MBs. Additional to the PamChips, the authors applied expression arrays for 48 MB samples of their cohort and identified two gene signatures that underlie protein-signalling cluster-1 and -2, respectively. Gene set enrichment analyses revealed that genes upregulated in cluster-1 vs. -2 samples were functionally associated with protein synthesis.

The so-far summarised aspects of MB concentrated on the four subgroups of MB. However, researchers started to investigate MB's molecular heterogeneity that goes beyond the four subgroups.

5.1.6 Subtypes within the four main MB subgroups

Three more recent publications by Cavalli *et al.*, Northcott *et al.*, and Schwalbe *et al.* addressed molecular subtypes of medulloblastoma beyond the four consensus subgroups by using different molecular data types and clustering approaches [221–223]. To avoid confusion between terms, from here on, the term subgroup always refers to the four main consensus subgroups (WNT, SHH, Group 3, and Group 4), whereas the term subtype addresses molecular subsets of MB within consensus subgroups [221–223].

Schwalbe *et al.* used the most variable 450K DNA methylation array probes across 428 MB samples and non-negative matrix factorisation (NMF is described in [89]) for sample clustering. The authors identified six subtypes that split each of the subgroups SHH, Group 3, and Group 4 into two subtypes, whereas the WNT subgroup remained integer as one group [223].

Since the heterogeneity of the SHH subgroup reflected by different age groups was already described, Northcott *et al.* concentrated on the identification of subtypes within Group 3 and Group 4 tumours (n=740) using 450K DNA methylation arrays. The clustering was done based on the most variable methylation array probes that were reduced to a 2D space using t-SNE. The algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise, [224]) was applied to the reduced feature space to cluster the samples. Eight subtypes were identified: four subtypes mostly represented Group 4 MBs, three subtypes related to Group 3 MBs, and one subtype included both Group 3 and Group 4 MBs with high proportions [222]. Thus, Northcott *et al.* identified a higher number of subtypes for Group 4 MBs than Schwalbe *et al.*.

Cavalli *et al.* used a different approach compared to Schwalbe *et al.* and Northcott *et al.* to define subtypes in a cohort of 763 MB samples [221]. The authors integrated both DNA methylation and gene expression data from microarrays, and each subgroup was independently clustered. The clustering was done by applying the Similar Network Fusion algorithm to calculate a similarity matrix between samples, followed by spectral clustering (see Wang *et al.* for further information about the Similar Network Fusion algorithm and subsequent spectral clustering [225]). Cavalli and colleagues identified two subtypes in WNT, four subtypes in SHH, three subtypes in Group 3, and three subtypes in Group 4 tumours. Interestingly, the authors reported that DNA methylation and gene expression provided complementary information for the identification of the subtypes. The analysis of the expression profiles of these subtypes was based on the most variable genes and not differential gene expression [221].

The subtypes of Cavalli *et al.* will be introduced in more detail below, since they provide the most comprehensive analysis. Schwalbe *et al.* did not provide data allowing a direct comparison of external data with their reported subtypes, and Northcott *et al.* did not include SHH MB in the subtype analysis [222, 223].

Cavalli *et al.* identified two WNT subtypes. The subtype WNT α that was characterized by monosomy 6 in 98% of the tumours was found more frequently in younger patients, whereas WNT β also occurred in adults (Table 5.3) [193, 221].

The SHH subgroup was split into four subtypes (Table 5.3). As mentioned above, SHH tumours show different molecular characteristics among the age groups infants, children, and adults. The four SHH subtypes reflect this age distribution. However, infant cases were split into the subtypes SHH β and SHH γ , both associated with a neural-development expression signature. In comparison to SHH γ , SHH β tumours are more metastatic, have a worse prognosis, and show frequent focal deletion of tumour suppressor *PTEN* (25%). The subtype SHH α occurs in children and shows recurrent *TP53* mutations as well as *MYCN* and *GLI2* amplifications. Subtype SHH δ represents adult SHH cases and is associated with *TERT* promoter mutations and Shh signalling [193, 221].

Table 5.3: Summary of molecular subtype in MB defined by Cavalli et al. [221].

Subgroup	Subtype	Age	Metastasis	5 years OS	CNV broad	CNV focal	Other events	Signatures
WNT	WNT α	infants, children (median 10 yo)	8.6%	97%	6 ⁻			
	WNT β	children, adults (median 20 yo)	21.4%	100%				
SHH	SHH α	children (median 8 yo)	20%	69.8%	9q ⁻ , 10q ⁻ , 17p ⁻	<i>MYCN</i> amp, <i>GLI2</i> amp, <i>YAP1</i> amp	<i>TP53</i> mut	DNA repair, cell cycle
	SHH β	infants (median 1.9 yo)	33%	67.3%		<i>PTEN</i> loss		neuronal development
	SHH γ	infants (median 1.3 yo)	8.9%	88%	Balanced genome			neuronal development
	SHH δ	adults (median 26 yo)	9.4%	88.5%		10q22 ⁻ , 11q23 ⁻	<i>TERT</i> promoter mut	SHH signalling
Group 3	Group 3 α	more infants, less children (median 4.82 yo)	43.3%	66.2%	7 ⁺ , 8 ⁻ , 10 ⁻ , 11 ⁻ , i17q			photo-receptor
	Group 3 β	mostly children (median 7.55 yo)	20%	55.8%		<i>OTX2</i> gain, <i>DDX31</i> loss	high <i>GFI1/1B</i> expression	neuronal differentiation, RNA processes
	Group 3 γ	infants, children (median 5 yo)	39.4%	41.9%	8 ⁺ , i17, 1q ⁺	<i>MYC</i> amp		RNA processes, Telomerase
Group 4	Group 4 α	infants, children (median 8.2 yo)	40%	66.8%	7q ⁻ , 8p ⁻ , i17q	<i>MYCN</i> amp, <i>CDK6</i> amp		neuronal development, cell migration
	Group 4 β	infants, children (median 10 yo)	40.7%	75.4%	i17q	<i>SNCAIP</i> dup		MAPK pathway, FGFR1 mutant receptor
	Group 4 γ	infants, children (median 7 yo)	38.7%	62.5%	7q ⁺ , 8p ⁻ , i17	<i>CDK6</i> amp		neuronal development, photoreceptor, PI3K pathway

Note: Table adapted from Cavalli *et al.* [221]. Infant: 0-3 years. Children: 4-17 years. Adults: > 17 years. ⁻ = loss. ⁺ = gain. amp = amplification. mut = mutation. yo = years old.

5.2 Study: Gene regulatory networks and characterisation of *lnc* genes in medulloblastoma

Subgroup Group 3 was subdivided into three subtypes (Table 5.3). Group 3 α tumours occur in infants and children. These tumours are associated with recurrent chromosome 8, 10, and 11 loss, chromosome 7 gain, and express a stronger photoreceptor signature compared to the remaining Group 3 subtypes. This subtype shows the highest frequency of metastasis (43.3%) but the best prognosis among Group 3 subtypes (66.2%). The Group 3 β subtype appears in children and is associated with *GFI1*/*GFI1B* overexpression, *DDX31* loss, *OTX2* gain, and a neuronal differentiation expression signature. Group 3 γ tumours occur in infants and children, are frequently metastatic, and show the worst 5 years survival (41.9%) among MB subtypes. Here, frequent *MYC* amplification is probably the reason for the bad outcome [193, 221].

Three Group 4 subtypes have been defined. They occur in infants and children and have a similar outcome (Table 5.3). Subtype Group 4 α shows *MYCN* and *CDK6* amplification and a neuronal development expression signature. Subtype Group 4 β associates with *SNCAIP* duplications and a MAPK and FGF signalling signature. Group 4 γ tumours show frequent *CDK6* amplification and expression signatures related to neuronal development, photoreceptors, and the PI3K pathway [193, 221].

In summary, different clustering approaches and data types have been used to identify molecular subtypes in MB [221–223]. Among these publications, Cavalli *et al.* identified distinct subtypes within each of the subgroups using arrays-based data associated with specific genomic aberrations and expression signatures.

5.2 Study: Gene regulatory networks and characterisation of *lnc* genes in medulloblastoma

5.2.1 Motivation

The consensus is that MB represents a heterogeneous collection of four distinct molecular tumour subgroups WNT, SHH, Group 3, and Group 4. The underlying genomic landscape of the subgroups is well studied. However, the current state of research of the MB transcriptome still leaves some aspects unstudied and, thus, has to be further refined. MB bulk transcriptome studies have been mainly based on the microarray technology and, therefore, are bound to the limitations of this technology. Also, the MB transcriptome was mostly studied based on differentially expressed coding genes since *lnc* genes are poorly covered on microarrays [82, 207, 208, 212, 214]. Transcription factors (TFs) that mainly contribute to subgroup-specific expression signatures are known for WNT and SHH tumours, due to associations of TFs with the activated pathways, and partially for Group 3 MBs. The TFs for Group 3 include *MYC* and potentially *CRX* and *NRL*. However, TFs like *CRX* and *NRL* are not functionally studied in MB leaving their regulatory impact unanswered. The regulation of Group 4-specific expression signature is not known. The enhancer-mediated GRN that was studied in MB covers only one aspect of gene regulation and disregards promoter-mediated gene regulation. Despite the consensus of the four main MB subgroups, the different number of reported potential subgroups before the consensus and the described intra-subgroup heterogeneity of SHH and Group 3 tumours indicate that the subgroups probably do not cover the whole complexity of MB. Molecular complexity that exceeds the subgroups has been studied by the identification of subtypes within the main subgroups. However, the analyses of these subtypes based mostly on microarrays. Additionally, the sets of subtype-specifically expressed genes and the GRNs underlying this subtype-specific expression remain to be investigated.

A transcriptome-centred study that uses deep RNA-seq data from a sufficiently large, informative MB cohort would shed light on the MB transcriptome.

5.2.2 Overview and research scope

We set out to perform a comprehensive analysis of the medulloblastoma transcriptome using RNA-seq data. The data foundation is an RNA-seq cohort comprising 164 MB and eight normal cerebellum samples that were generated and analysed within the framework of ICGC PedBrain in the laboratory of "Gene Regulation & System Biology of Cancer" located at the Max Planck Institute for Molecular Genetics.

In order to gain a deeper understanding of the medulloblastoma transcriptome, three aspects will be considered. The first aspect covers the dissection of the transcriptional heterogeneity beyond the known the four main MB subgroups since a subdivision of subgroups into molecular subclusters is expected to provide further insights into MB biology. (Subsets of MB within the four main subgroups identified by us will be called subclusters to distinguish our work from already published subtypes by using different terms.) This aspect is addressed via identifying molecular subclusters within each subgroup (intra-subgroup) by applying an unsupervised consensus clustering approach. Further, the subgroups and newly identified subclusters are characterised by differentially expressed coding as well as lnc genes, annotated for functional attributes (including biological processes and pathways) inferred by overrepresentation analysis.

The second aspect studied here focuses on the analysis of the transcriptional regulation networks that contribute to the subgroups and subclusters. Here, GRNs underlying the subgroups and subclusters are directly inferred from gene expression data. These GRNs are used to identify transcription factors that mainly contribute to subgroup- and subcluster-specific gene expression depicting the landscape of putative key regulatory genes in MB. The machine learning algorithm GENIE3 (see Section 3.5.1) is applied here to infer GRNs since GENIE3 learns regulatory links between TFs and putative targets directly from deep gene expression data, which provides an excellent basis for the inference of GRNs. The impact of individual TFs on gene expression is evaluated using the inferred GRNs and the previously proposed NIS (see Section 3.5.2).

The third aspect described in this thesis addresses the characterisation of lnc genes that are differentially expressed between MB subgroups and subclusters to better understand the implication of lnc genes in MB. Since the function of the majority of lnc genes is unknown and lnc genes have been rarely studied in MB (see Section 2.5 and 2.5), differentially expressed lnc genes are characterised in more details by integrating information streaming from several data resources, including the classification of lnc genes into the different lncRNAs categories, namely divergent, antisense, and intergenic as well as the annotation of tissue-specific expression patterns in brain and cerebellum. The tissue-related expression pattern may provide insights into potential developmental functions of lnc genes, which is of interest with regards to the embryonic origin of MB. The evaluation of expression correlation between divergent, antisense, and intergenic lnc genes and nearby coding genes informs on either the independent or co-regulation of lnc genes and their neighbouring coding genes. Following the guilty-by-association principle, this is used to infer potential functions of lnc genes from functional annotations of matched protein coding genes. Additionally, survival analyses will be performed to identify lnc genes whose expression is prognostic of OS in MB.

5.3 Results

The results of the conducted MB study are presented in three parts. The first part (Section 5.3.1) summarises the analysed cohort, the assignment of the MB samples to the four main subgroups, and the identification of the subclusters within the main subgroups. In the second part, the expression profiles of the main subgroups and identified subclusters as well as the inferred GRNs are described (Section 5.3.2). The characterisation of the differentially expressed lnc genes in MB is presented in the third part (Section 5.3.3).

5.3.1 Subgroup classification and subcluster identification

5.3.1.1 Medulloblastoma cohort overview and main subgroup classification

The RNA-seq (ICGC PedBrain) cohort comprised 164 MB samples and eight controls of the cerebellum (three postnatal and five prenatal samples; Figure 5.2.a). The cohort included 15 WNT, 47 SHH, 39 Group 3, and 63 Group 4 tumours. Here, the frequency of the four subgroups and the age distribution of the patients followed the expected clinical patterns (Figure 5.2.a-b, Section 5.1.3).

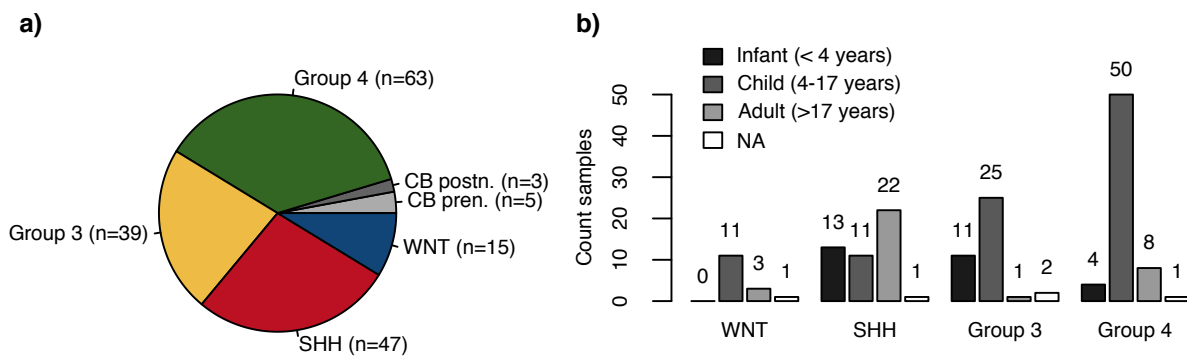


Figure 5.2: ICGC PedBrain MB RNA-seq cohort summary. **a)** Pie chart visualises the number of MB samples per subgroup and pre- (pren.) and postnatal (postn.) cerebellum (CB) controls. **b)** Bar plots show the age group distribution in subgroups.

The assignments of the four subgroups were derived from DNA methylation-based classification and provided by collaboration partners of the ICGC PedBrain project [222]. The DNA methylation-based classification was initially used to analyse the subgroups since this type of classification was chosen as the standard approach in the PedBrain project design. However, DNA methylation- and RNA-seq-based MB classification were compared to assess whether both data types were equally powerful for identifying the four MB subgroups. Using the RNA-seq data and assuming four clusters, we clustered the 164 MB samples (unsupervised) by considering the 6436 most variable coding genes and applying the algorithm NMF (Method Section 5.4.4.1). The consensus matrix of the NMF-based clustering indicated high stability of subgroup assignment for the clustered samples (Figure 5.3.a). The RNA-seq- and DNA methylation-based classification showed overall a high agreement (Figure 5.3.b). Few samples of Group 3 and Group 4 were interchanged in the RNA-seq-based clustering. However, some cases from these two subgroups are sometimes difficult to classify, as mentioned above (Section 5.1.6). Taken together, these results indicated that DNA methylation-based and RNA-seq-seq give similar classification results, which justifies using DNA methylation for subgroup classification.

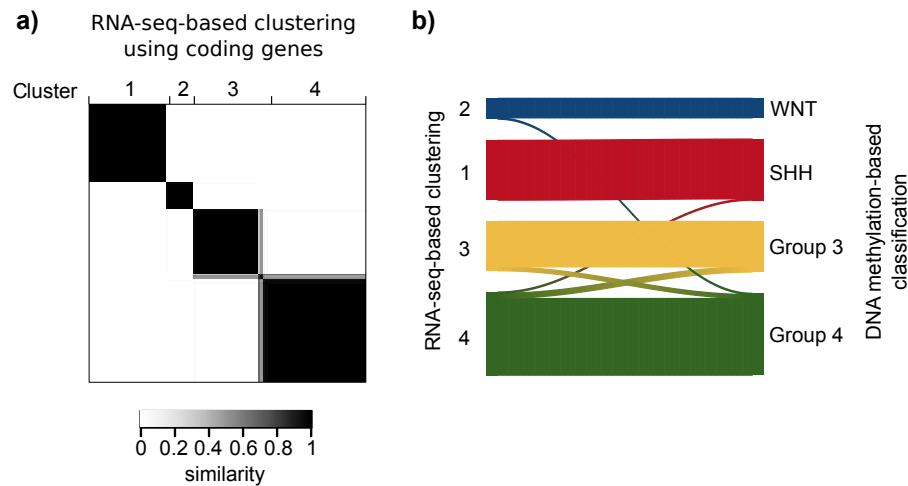


Figure 5.3: Comparison of RNA-seq-based unsupervised clustering and DNA methylation-based classification of the PedBrain MB cohort. **a)** Consensus matrix of the NMF clustering on RNA-seq-derived coding genes expression values. Heatmap shows the frequency of two MB samples falling into the same cluster over 60 NMF iterations. **b)** Sankey plot visualises the agreement between RNA-seq-based NMF clustering (panel a) and methylation-based classification.

5.3.1.2 Evaluation of the identified intra-subgroup subcluster

We identified subclusters within subgroups by analysing each subgroup independently. The clustering of the MB samples into subclusters included two steps (Method section 5.4.4.2). In the first step, we performed unsupervised consensus-based clustering using the most variable gene per subgroup. The consensus clustering was based on a consensus distance matrix derived from the mean pair-wise correlation between samples across subsamples of the most variable genes. Via this approach, we identified three subclusters in each of the main MB subgroups SHH, Group 3 and 4 (Figure 5.4.a-c). The small sample size of the WNT subgroup (15 cases) did not allow subcluster identification. Five samples of Group 4 could not be placed into a subcluster at this step, as indicated by the dendrogram (Figure 5.4.c). In the second step, we identified genes that showed differential expression between the subclusters of one subgroup by analysing each subgroup independently (results of the DEGA are summarised the following Section 5.3.2.1). These genes were used for hierarchical clustering of the samples to identify subclusters within a subgroup (semi-supervised clustering). Six tumour samples changed the cluster between the unsupervised and semi-supervised clustering, namely three SHH and three Group 4 samples (Figure 5.4.d-f). Measuring the goodness of the clustering for these samples by calculating the silhouette scores of the semi-supervised clustering showed that three samples had a negative silhouette score (MB179, MB236, and MB91) and two samples a low silhouette score <0.1 (MB265 and MB246) (a score of +1 would indicate a perfect assignment to a cluster). The negative and low silhouette scores indicated a difficult clustering of these samples. Therefore, these samples changed the cluster assignment between unsupervised and semi-supervised clustering. Additionally, five Group 4 samples without cluster assignment in the unsupervised clustering could be related to one cluster in the semi-supervised clustering showing a silhouette score between 0.07 and 0.29 (Figure 5.4.d-f). The expression profile of Group 4 sample MB177 did not match one of the three clusters and remained in the cluster that was assigned by the unsupervised clustering (Figure 5.4.f).

The three subclusters in each of the subgroups SHH, Group 3 and Group 4 were respectively called c1, c2 and c3 (Figure 5.5.a). SHH-c1 was a subset of SHH-c3 (Figure 5.5.a, 5.4.d). Here, genes upregulated in SHH-c3 compared to SHH-c2 were also upregulated in SHH-c1, but SHH-c1 MBs showed a distinct gene expression pattern compared to SHH-c2 and SHH-c3 (Figure 5.4.d). To put it briefly,

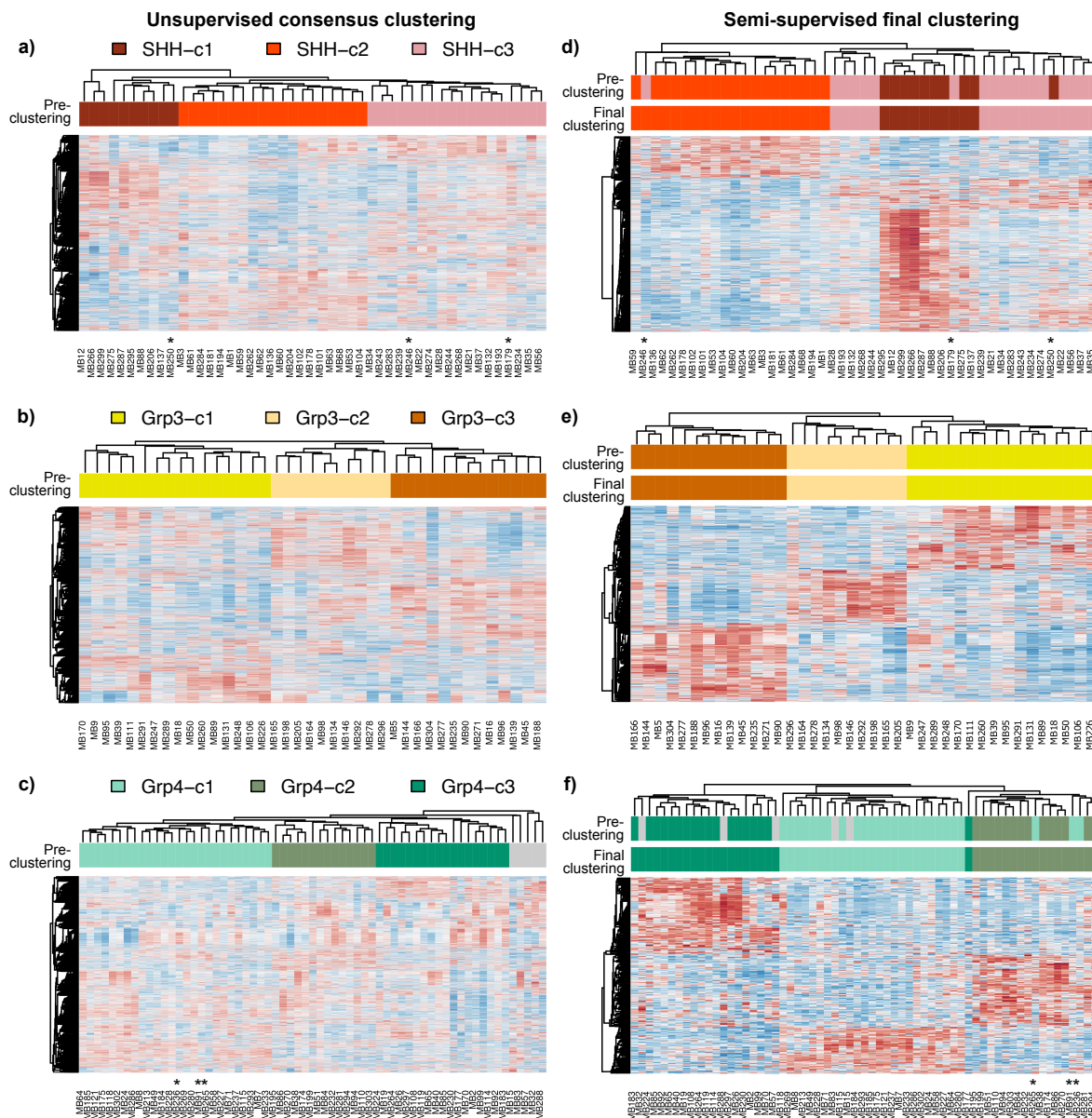


Figure 5.4: Identification of subclusters within subgroups in the PedBrain MB cohort. Consensus clustering was performed using coding genes. **a, b, c)** Heatmaps show that most variable genes that were used for consensus-based pre-clustering per subgroup: a) 3288 genes, b) 4193 genes, c) 2633 genes. Clustering results are indicated above as colour code. **d, e, f)** Heatmaps show subcluster-specifically upregulated genes within each subgroup: d) 1075 genes, e) 981 genes, f) 431 genes). Final clustering and pre-clustering results are indicated above the heatmap. **a, d)** SHH tumours. **a, d)** Group 3 tumours. **a, d)** Group 4 tumours. Grey-marked samples did not fit into the pre-clusters. Asterisk-marked MB samples that changed the subclusters between pre- and final clustering.

the expression profile of the subclusters SHH-c2 and SHH-c1 was different to subcluster SHH-c3, but the expression of subcluster SHH-c1 was different compared to subcluster SHH-c2 and SHH-c2. Here, SHH-c1 and SHH-c3 represented non-adult cases, whereas SHH-c2 represented adult SHH MB samples (Figure 5.5.b). Among the Group 3 subclusters, Grp3-c3 showed the highest fraction of infant cases; Grp3-c1 and Grp3-c2 were more frequent in children (Figure 5.5.c). The age groups were evenly distributed across Group 4 subclusters (Figure 5.5.d).

The RNA-seq-based MB subclusters that we identified were compared to intra-subgroup subtypes published by Northcott *et al.* and Cavalli *et al.* (Figure 5.6) [221, 222]. To compare to Northcott *et al.*, the ICGC PedBrain cohort was used. The comparison to Cavalli *et al.* was done using the microarray cohort from Cavalli *et al.*. This cohort was used in order to take advantage of the larger sample size and evaluate the stability of the identified subclusters on an external cohort.

We performed semi-supervised clustering to classify samples of the Cavalli *et al.* cohort into our RNA-seq-based subclusters, which was applied in a subgroup-wise way: subcluster-specifically expressed genes that we identified were mapped to the microarray expression probes, and NMF was performed to classify samples into three subclusters per main MB subgroup SHH, Group 3, and Group 4 (see Methods section 5.4.4.2).

Performing a Chi-squared test showed a significant overall agreement between our RNA-seq-based subclusters and the published subtypes ($p < 2e - 15$, Figure 5.6). Nevertheless, as shown in the Sankey diagrams in Figure 5.6, our RNA-seq-based subclusters showed a higher agreement with Cavalli *et*

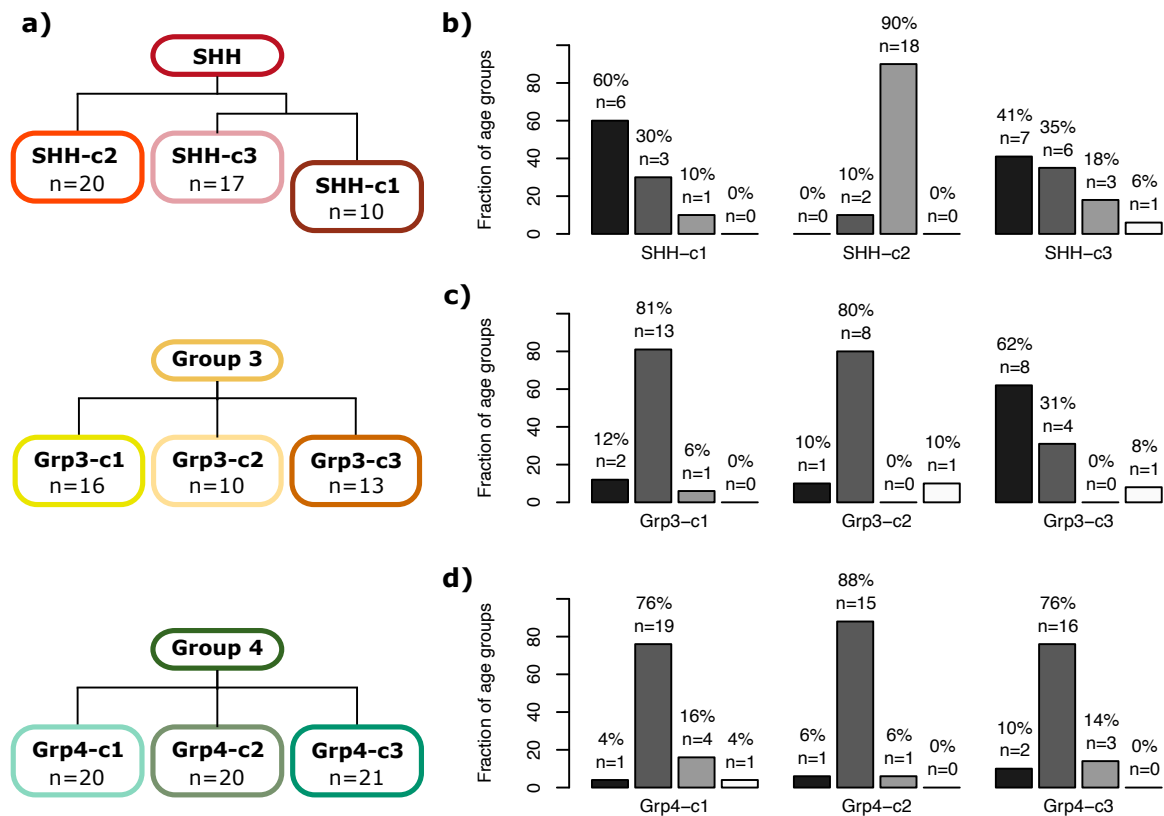


Figure 5.5: Identified subclusters in SHH, Group 3, Group 4 MBs. **a)** Trees show RNA-seq-based *de novo* identified subclusters on the PedBrain MB cohort. **b-d)** Age distribution in subclusters. Bars show the percentage of MB samples that belong to one of the three age groups per subcluster. **b)** SHH subclusters. **c)** Group 3 subclusters. **d)** Group 4 subclusters. Age groups relate to infant, child, and adult, as indicated by the caption. Numbers above the bars show the percentage (%) and absolute number (n) of samples.

al. than with the Northcott study (Figure 5.6). Compared to Northcott *et al.*, the subcluster Grp3-c2, identified by us, did not match with any of the reported subtypes. The subcluster Grp3-c3 split into the published subtypes III and IV, and the subcluster Grp4-c2 split into the published subtypes V and VI as defined by Northcott *et al.* (Figure 5.6.a). Northcott *et al.* have shown that their subtypes III and IV as well as V and VI are neighbouring subtypes indicating that these subtypes could be a subset of the Grp3-c3 and Grp4-c2 subcluster that we identified, respectively [222].

Subtypes published by Cavalli *et al.* within the main subgroups Group 3 and Group 4 showed overall a high concordance with our RNA-seq-based subclusters. The subcluster Grp3-c3 showed the highest frequency of infants among subclusters within the main subgroup Group 3. The subcluster Grp3-c3 matched with the published subtype Group 3 α that was also frequently found in infant cases [221]. Cavalli *et al.* reported frequent *MYC* amplifications in subtype Group 3 γ [221]. We found frequent *MYC* amplifications also in subcluster Grp3-c3 that matched with the published subtype Group 3 γ (Figure 5.6.b and 5.7.c). Frequent *CDK6* amplifications that were reported for the subtypes Group 4 β and Group 4 γ were also present in the matching subclusters Grp4-c1 and Grp4-c3, respectively. We could observe *MYCN* copy number gains that were reported for the published subtype Group 4 α also in our identified subcluster Grp4-c2 that matched this subtype (Figure 5.6.b and 5.7.d-e) [221].

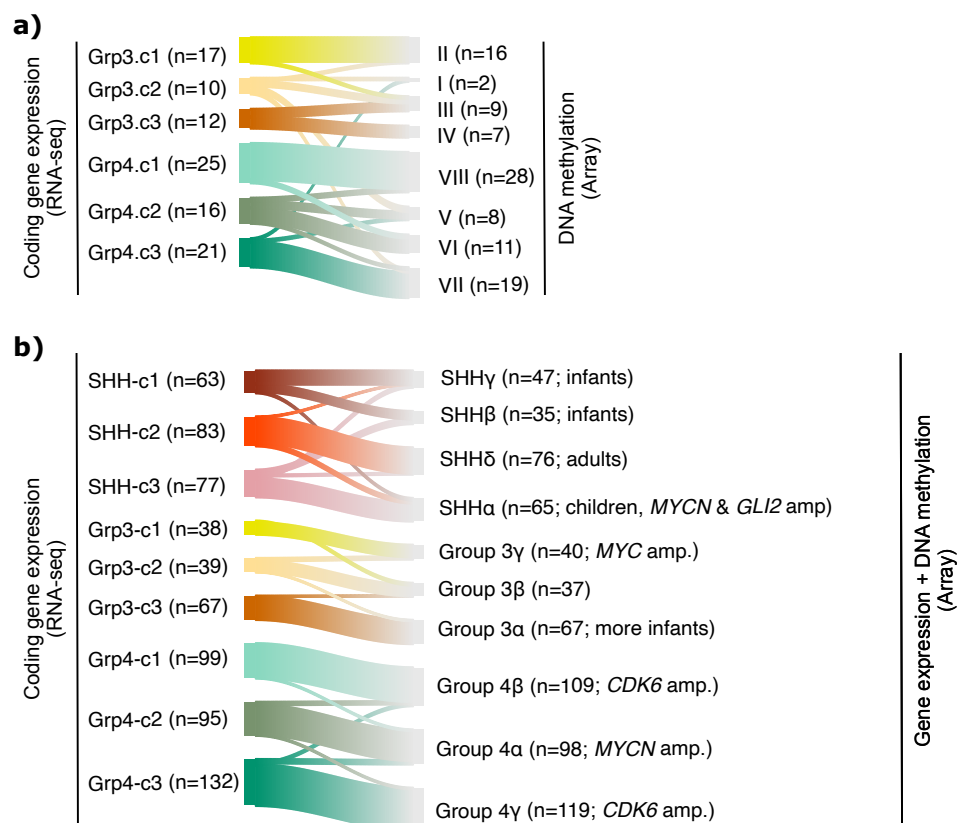


Figure 5.6: Comparison of identified intra-subgroup subclusters to previously published subtype definition. **a)** Sankey plot shows the comparison between RNA-seq-based *de novo* subclusters and published subtypes of Group 3 and 4 on the PedBrain MB cohort. Subtype assignments were taken from Northcott *et al.* [222]. **b)** Sankey plot shows the comparison between RNA-seq-based *de novo* subclusters and published subtypes of SHH, Group 3, and Group 4 on the Cavalli *et al.* cohort. Subtype assignments were taken from Cavalli *et al.* [221]. Subcluster assignments were achieved via semi-supervised clustering (see Methods section 5.4.4.2). The selection of subtype characteristics, as summarised above (Section 5.1.6), are shown in brackets. The number of samples is indicated in brackets.

Compared to the subclusters of the main subgroups Group 3 and Group 4, a higher discordance was observed between our subclusters and Cavalli's subtypes within the main subgroup SHH (Figure 5.6.b). The two published subtypes SHH α and SHH γ , which comprised infant SHH tumours, of the Cavalli study merged mainly into the subcluster SHH-c1 that we identified (Figure 5.6). Since subcluster SHH-c1 showed the highest fraction of infant cases compared to the remaining two SHH subclusters (Figure 5.5.b), the age distribution agreed between subcluster SHH-c1 and the matching published subtypes SHH α and SHH γ . The low number of ten SHH-c1 cases suggested that the sample size of our RNA-seq cohort was probably too small to identify more fine-grain molecular clusters among SHH-c1 cases. The adulthood-related subtype SHH δ and the childhood-related subtype SHH α by Cavalli *et al.* matched with our identified subcluster SSH-c2 and SSH-c3, respectively. Here, the age distributions of both subclusters agreed with their matching published subtypes (Figure 5.6.b and 5.5.b). Cavalli *et al.* reported that amplifications of *MYCN* and *GLI2* were more frequent in SHH α compared to the remaining subtypes within the main subgroup SHH [221]. We found a similar pattern for subcluster SHH-c3 that matched with the subtype SHH α (Figure 5.6.b and 5.7.a-b).

The high agreement between the RNA-seq-based subclusters that we identified and subtypes of Cavalli *et al.* showed that RNA-seq and the applied consensus clustering method could be used to identify subclusters within the main subgroups in MB. Additionally, the comparison suggests that the subclusters that we identified and the subtypes of the Cavalli study probably describe the same molecular subsets within the main MB subgroups, considering that Cavalli *et al.* identified two subtypes for infant SHH MBs. In contrast, we identified one subcluster for infant SHH MBs. The four main subgroups and the identified subclusters were the basis for further analyses of the MB transcriptome that we performed.

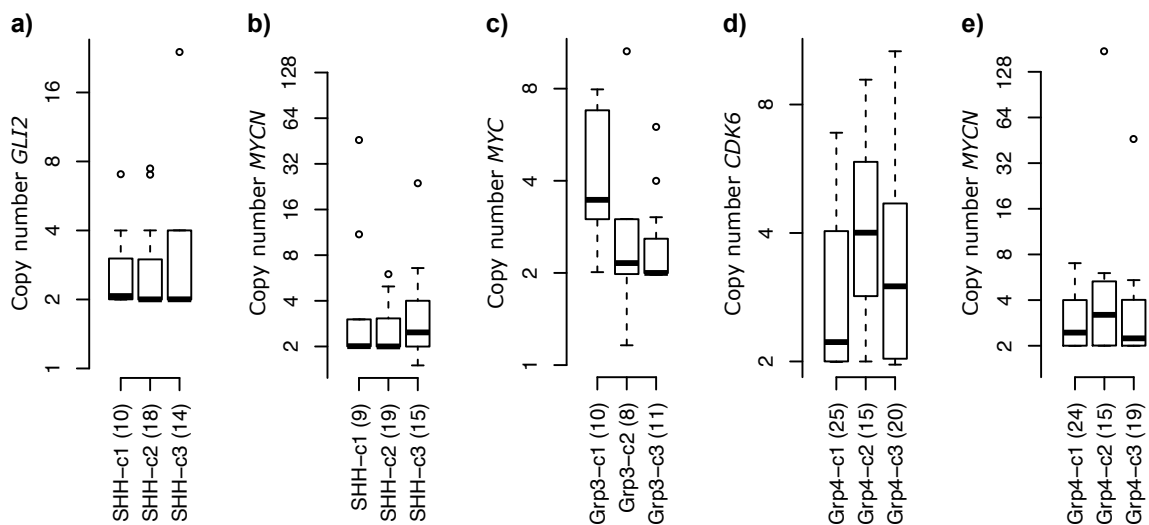


Figure 5.7: CNVs in subclusters. Box plots show distribution of gene copy numbers in subclusters. Axis is on log₂-scale.

5.3.2 Gene expression profiles and gene regulatory networks in medulloblastoma

5.3.2.1 Summary of differential gene expression in subgroups and subclusters

We performed a differential gene expression analysis based on fitting a GLM assuming an NB distribution and using a log-likelihood ratio test (Method Section 5.4.4.3, see Section 3.3.2). This procedure allows the detection of differential gene expression between several groups similar to an analysis of variance (ANOVA). Four independent DGEAs were performed. Three DGEAs related to the subgroup-

wise comparison between the subclusters within the main subgroups SHH, Group 3 and Group 4. One DEGA related to the comparison between the four MB subgroups. The DEGAs sought to identify genes that were specifically up- or downregulated in one subgroup or one subcluster among the subclusters of the same subgroup. The differential gene expression analysis between the four subgroups resulted in 1790 up- and 931 downregulated coding genes and 239 up- and 59 downregulated lnc genes (Figure 5.8.a; $FDR \leq 0.001$, $|\log_2(FC)| \geq 1$). Subclusters were defined by 1937 up and 775 downregulated coding genes, and 221 up- and 68 downregulated lnc genes (Figure 5.8.a; $FDR \leq 0.01$, $|\log_2(FC)| \geq 1$). Notably, 1192 coding and 163 lnc genes showed only among subclusters significant differential expression with the chosen cutoffs (Figure 5.8.b-c), indicating that the subclusters provide information about the transcriptional landscape in MB in addition to the subgroups.

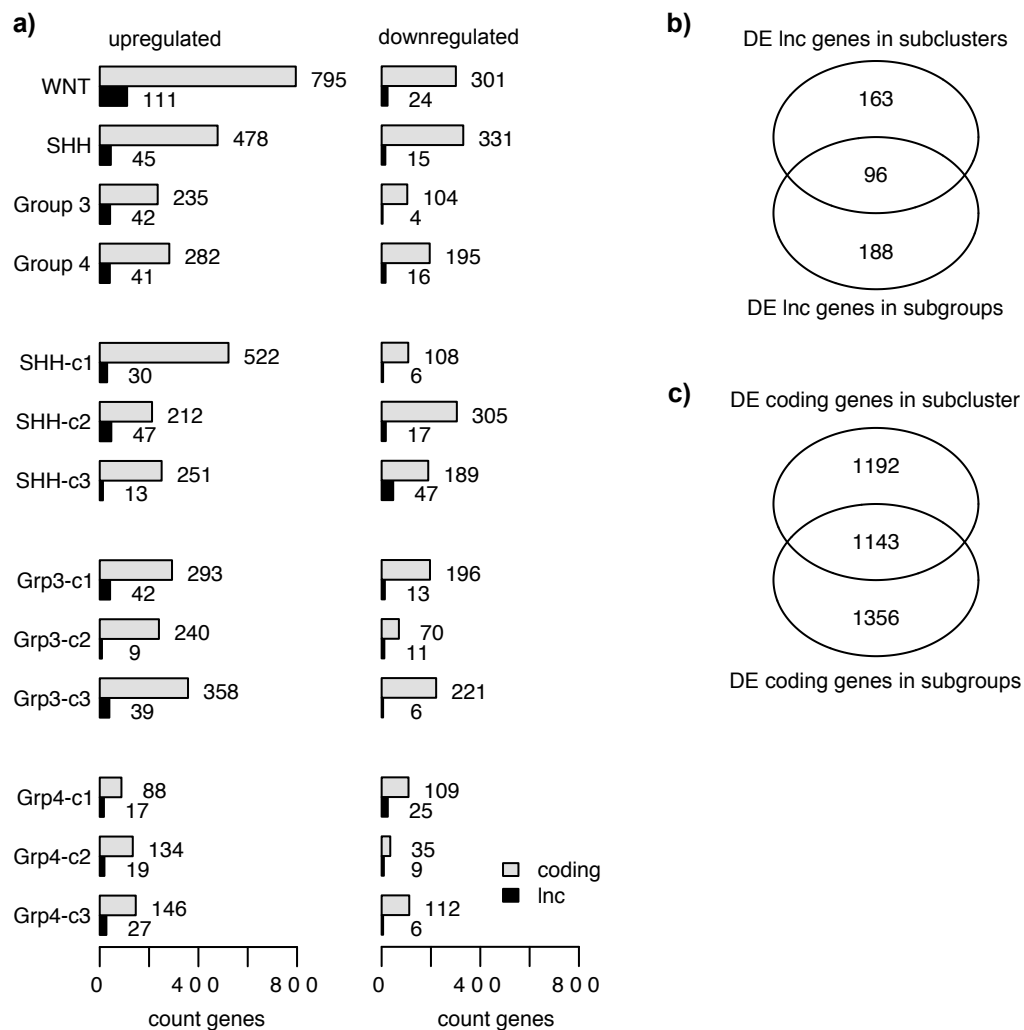


Figure 5.8: Identified differentially expressed genes in MB subgroups and subclusters. **a)** Bar plots show the number of differentially expressed genes between subgroups and subclusters of one subgroup. Left: subgroup and -cluster specifically upregulated genes. Right: subgroup and -cluster specifically downregulated genes. lnc and coding genes are separately shown. **b)** Venn diagram shows the set of lnc genes that were differentially expressed in subgroups and subclusters or exclusively. **c)** Venn diagram shows the set of coding genes that were differentially expressed in subgroups and subclusters or exclusively.

The just-presented sets of differentially expressed coding and lnc genes formed the basis for dissecting the transcriptional heterogeneity associated with subclusters and subgroups in MB. Thus, the

following presented analyses and results mainly relate to these differentially expressed genes. Due to the limited functional annotation of lnc genes, functional enrichments were based on coding genes. The identification of GRNs was among these analyses.

5.3.2.2 Inference and evaluation of the gene regulatory networks

We inferred four co-expression-based gene regulatory networks. These four GRNs represent TF-gene interactions among genes and TFs that were differentially expressed between the main subgroups or between subclusters within the subgroups SHH, Group 3, or Group 4. The introduced algorithm GENIE3 (Section 5.4.5) was applied to infer these GRNs from RNA-seq-based gene expression. Lnc and coding genes were considered as putative targets in the inferred GRNs of subgroups and subclusters. To select only strong and potentially meaningful associations between TFs and genes, we determined a strict cutoff for the GENIE3-provided weights of TF-gene interactions. The cutoff selection was based on a GRN fitting score, which was defined by the ratio between the enrichment of TFBS in promoters of predicted TF targets and the network density, which is the inverse of the sparsity (Method section 5.4.5). This score was chosen in order to obtain a GRN that shows a high TFBS enrichment over a high sparsity, as exemplarily shown for the GRN in the subgroup (Figure 5.9.a-c). The 0.997475, 0.984, 0.99, and 0.986 percentile of GENIE3 interaction weights were chosen as cutoff for the GRN of the MB main subgroups and of the subclusters within SHH, Group 3 and Group 4 MBs, respectively.

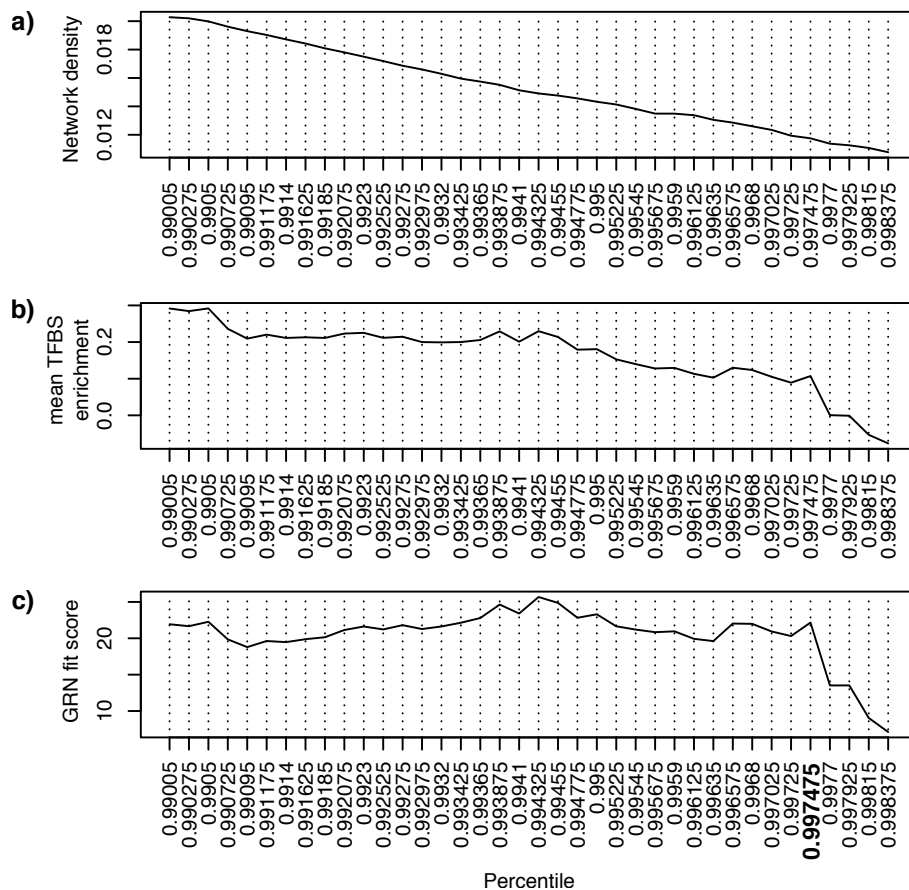


Figure 5.9: Cutoff evaluation for interaction weights of the GRN of MB subgroups. Line graphs show network evaluation values over the cutoff percentile of GENIE3 derived interaction weights. **a)** Network density. **b)** Average TFBS enrichment across TFs. **c)** Final GRN fitting score for cutoff evaluation (Method section 5.4.5). Chosen cutoff percentile is highlighted in bold.

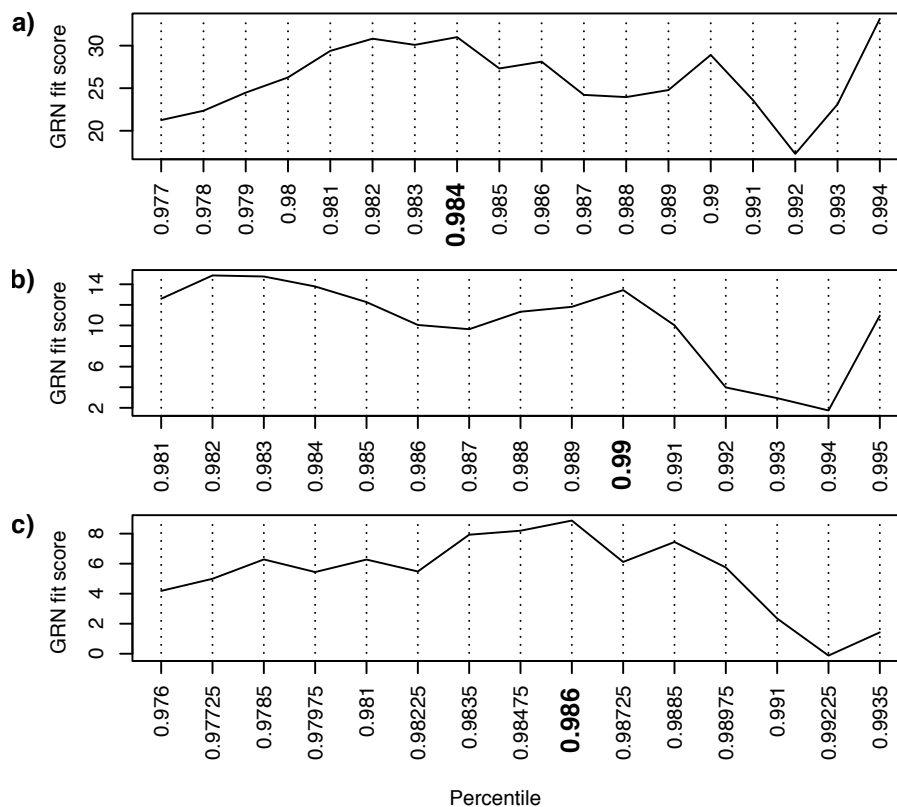


Figure 5.10: Cutoff evaluation for interaction weights of the GRN of MB subclusters. **a-c)** Line graphs show GRN fitting score for over the cutoff percentile of GENIE3 derived interaction weights (Method section 5.4.5). Chosen cutoff percentile is highlighted in bold.

Considering the four inferred GRNs collectively, we 7898 inferred putative interactions between 339 TFs as regulators and 3247 target genes comprising 2906 coding and 341 lnc genes. Here, putative interactions could be inferred for 78.7% (2906/3691 genes) and 75.7% (339/448 genes) of all differentially expressed coding and lnc genes, respectively. We evaluated the 7898 inferred TF-gene interactions integrating previously published ChIP-seq data of HLX, LHX2 and LMX1A in primary MB samples as well as ChIP-seq data of OTX2 and NEUROD1 in MB cell lines (Method Section 5.4.5.7) [218, 226]. Here, TF ChIP-seq peaks in promoter or enhancer regions of putative targets were used to validate TF-gene interactions. Assignments between enhancer and putative enhancer target genes in MB were taken from Lin et al. [218]. We performed a hypergeometric test to evaluate if the number of inferred TF targets that are associated with a TF ChIP-seq peak could be obtained by chance, considering the number of expressed genes as background. 67/84 predicted OTX2 targets ($p=3.7e-14$), 36/48 predicted NEUROD1 targets ($p=0.03$), 56/75 predicted HLX targets ($p=0.041$), 20/20 predicted LHX2 targets ($p=0.002$) and 41/43 predicted LMX1A targets ($p=6.8e-08$) overlap with ChIP-seq peaks in promoter or enhancer regions. Overall, 81% (220/270 interactions) of inferred TF-gene interactions were validated by ChIP-seq peaks for these five TFs. These results exemplify the robustness of the inferred GRNs and suggest that the later shown GRNs in MB subgroups and subclusters allow insights into regulatory mechanisms in MB.

In the following sections, the four inferred gene regulatory networks were visualised and utilised to determine the TFs that showed the strongest regulatory influence on gene expression in a particular subgroup or subcluster. Here, only positive regulation between TFs and targets was considered. The regulatory influence of TFs within the GRNs was measured by the NIS (see Section 3.5.2) and by the out-degree of TF nodes in the visualised GRN. The NIS integrates the number of putative TF target genes (out-degree) and the degree of dysregulation of the targets and the TF in a particular subgroup

or subcluster (Methods Section 5.4.5.5). Two versions of NIS plots will be presented. The first version shows TFs that were specifically upregulated in one subgroup or subcluster. The second version contains TFs specifically downregulated in one of the subgroups or subclusters. However, the NIS for these TFs was shown for the subgroups or subclusters that do not show downregulation since only positive regulation was considered, and TFs that were downregulated in one subgroup/subcluster could still influence the expression in the remaining subgroups/subclusters. The NIS is shown additionally to visualised GRN since the score allows a more straightforward interpretation of the impact of TFs on the regulatory network than the out-degree of TF nodes visualised by the node size. The visualisation of the GRNs allows the depiction of the network topology.

5.3.2.3 Medulloblastoma subgroups

Among the coding genes that were subgroup-specifically upregulated in WNT, SHH, Group 3 or Group 4 tumours, we detected functional enrichments that agreed with previously published data (see Section 5.1.3) (Figure 5.11.a) [83]. However, the WNT subgroup showed an enrichment for FGF signalling (Genomatix: $p = 3.02e-9$; Genomatix: $p = 3.02e-9$;) including the developmental gene *FGF8*, which was not described for this subgroup. *FGF8* is expressed during development in the isthmic organiser at the mid-/hindbrain boundary and is required to form this boundary and for cerebellum development [196].

The GRN that we inferred for the main subgroups contained 212 TFs as regulators, 1603 target genes (1391 coding genes and 183 lnc genes), and overall 3691 TF-gene interactions (Figure 5.11.b). A TF-gene interaction was predicted for 55.7% (1391/2499) and 64.4% (183/284) of subgroup-specifically expressed coding and lnc genes, respectively. The network showed six bigger subnetworks. Four subnetworks represented subgroup-specific upregulation of TFs and genes in one of the four main subgroups SHH, WNT, Group 3, or Group 4. Two subnetworks related to subgroup-specific downregulation in SHH or WNT MBs. TFs and genes downregulated in Group 3 or Group 4 did not form clear subnetworks. The TFs with the highest NI-scores in the subgroups WNT and SHH were associated with the activated Wnt and Shh pathway, respectively (Figure 5.11.c). In the WNT subgroup, these TFs included the Wnt-signalling targets *RUNX2*, *SP5* and *MSX2*, [234–236]. In SHH MBs, Shh-signalling associated TFs comprised the direct Shh targets *GLI1* and *GLI2*, and the Shh-signalling collaborating TF *ATOH1*, whose protein product is stabilised by Shh-signalling in MB and granule neuron progenitors, showed a high NIS [203, 237]. We found *SOX2* in the fourth position, a TF that is known to be Shh-dependent expressed and required for the tumour cell proliferation in SHH MB [211].

One of the subnetworks related to 117 genes/TFs that were downregulated in WNT MBs (Figure 5.11.b). This subnetwork contained the heat shock factor *HSF2* [238, 239] that can be a suppressor or promoter of cancer progression, the proto-oncogene *DEK* [240], and the histone deacetylase *HDAC2* [241], whose knockdown decrease tumour growth in SHH MB mouse models. We evaluated the NI-scores of these three TFs in the remaining three subgroups SHH, Group 3 and Group 4 because of the upregulation in these subgroups compared to WNT MB, and only positive-directed gene regulation was considered in our study. *HSF2* ranked in SHH, Group 3 and Group 4, *DEK* ranked in SHH, and *HDAC2* ranked in SHH and Group 3 MBs among the top TFs that were downregulated in another subgroup (Figure A.2). Interestingly, each of the three TFs was expressed at a similar level in SHH, Group 3, and Group 4 MB and showed upregulation compared to human cerebellum controls and WNT MB. *DEK* displayed the strongest upregulation (Figure A.3).

Via the performed TF ranking, we could show that the photoreceptor-differentiation-involved TFs *NRL*, *CRX*, and *RAX2* have the highest NI-scores in Group 3 MB [215, 242, 243], which agrees with the expressed photoreceptor signature in this subgroup (5.11) [83]. These three TFs cooperatively regulate gene expression in photoreceptors [242]. Our data suggested that this cooperation was apparently conserved in Group 3 MB, as indicated by the close position of these three TFs in the network due to a high number of shared targets (Figure 5.11.b, Figure 5.12). In our RNA-seq data, *NRL* and *CRX* were clearly expressed in all Group 3 tumours, whereas *RAX2* was virtually not expressed in a subset

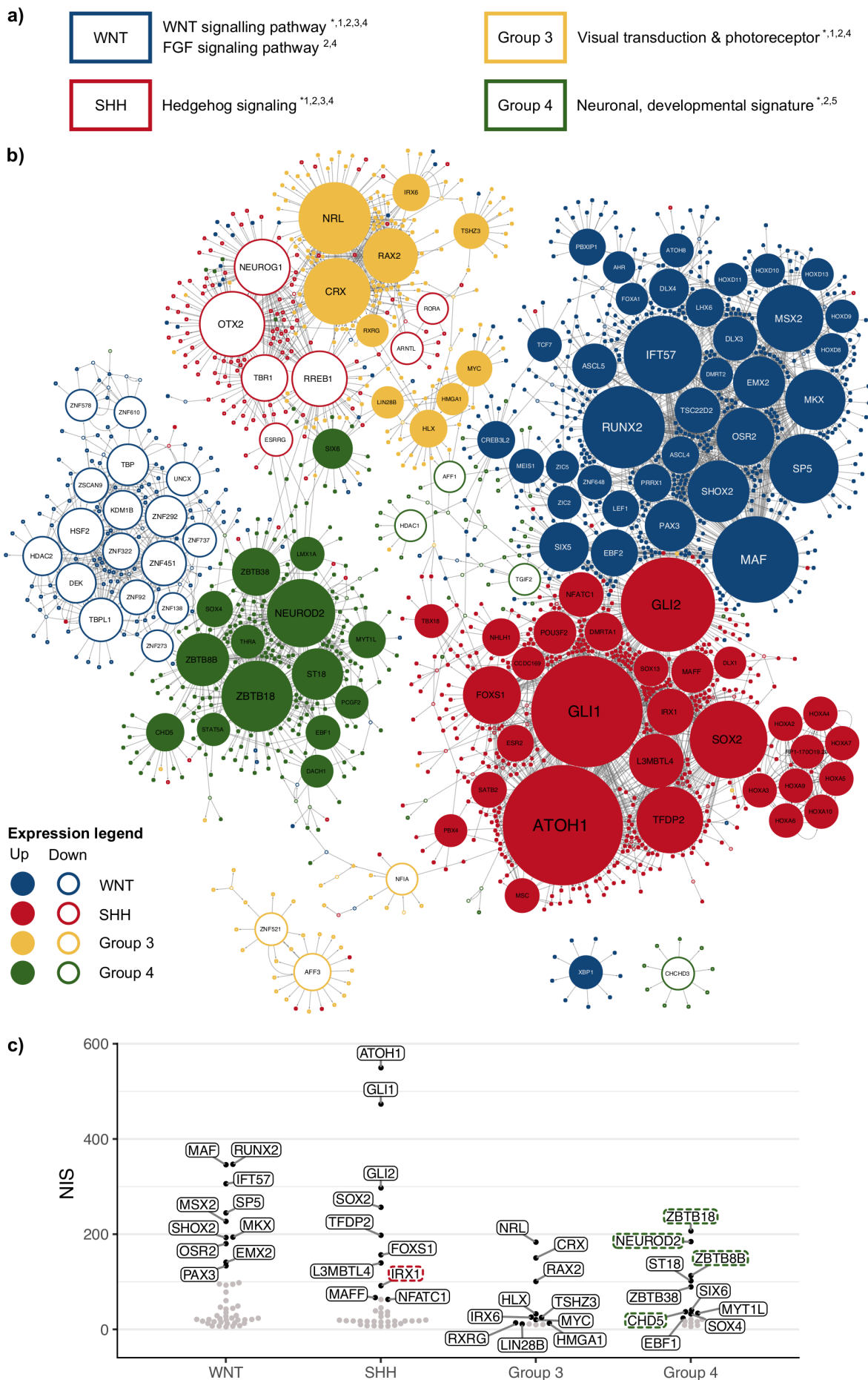


Figure 5.11 (preceding page): Functional enrichments and GRNs in MB subgroups. **a)** Enriched gene sets in subgroup-specific upregulated coding genes per subgroup. The asterisk indicates enrichments that are known from the literature [82, 83, 207, 208, 214]. Superscript 1-5 indicates the original source of the gene sets that show an enrichment:¹ KEGG, ² Reactome, ³ WikiPathways, ⁴ Genomatix, ⁵ GO Terms. FDR ≤ 0.05 . **b)** Inferred directed GRN in MB subgroups. Enlarged nodes represent TFs. The size of a TF node relates to the out-degree of a TF node. Target genes comprise coding and lnc genes and are shown as small nodes. Colours of the nodes indicate subgroup-specific expression. Filled nodes relate to subgroup-specific upregulation. Unfilled nodes relate to subgroup-specific downregulation. **c)** Bee swarm plot illustrates the ranking of TFs by their impact on subgroup-dependent gene regulation in GRN shown in panel a). The impact on gene regulation was measured by the NIS. Top-ten-ranked TFs per subgroup are shown. TFs that have not been mentioned as an important transcriptional regulator in the respective MB subgroup are highlighted by a dashed, coloured frame [82, 207–209, 218, 227–233]. **b-c)** Gene symbols are not written in italic to improve legibility.

of Group 3 samples, indicating that a contribution of *RAX2* is not essential for the regulation of the photoreceptor signature in Group 3 (Figure A.5, A.5). *NRL* and *CRX* have been previously mentioned to be higher expressed in Group 3 and potentially associated with the expressed photoreceptor signature in Group 3 MBs; *RAX2* was not mentioned in this context [207]. Lin *et al.* have reported a binding site enrichment for *RAX2* in active enhancers of Group 3 and 4 MB without further details [218]. The top-ranked TFs in Group 3 MB also included *MYC* and *HLX*. These TFs were higher expressed in Grp3-c1 vs. the remaining two subclusters within the main subgroup Group 3 (Figure A.4) and, therefore, will be discussed in the context of Group 3 subclusters in section 5.3.2.5.

Genes and TFs, including *OTX2*, *NEUROG1*, *TBR1* and *RREB1*, downregulated in SHH MBs formed a subnetwork closely located to the Group 3 subnetwork. Among all specifically downregulated TFs across the subgroups, *OTX2*, *NEUROG1*, *TBR1*, and *RREB1* showed the highest NIS in at least two non-SHH subgroups indicating a noteworthy influence of these TFs on gene expression in non-SHH MBs (Figure A.2.a). Our current results are supported by our previous work in Lin *et al.*, where a binding site enrichment for *OTX2* and *RREB1* in subgroup-specific enhancers of non-SHH MB was reported [218]. In our RNA-seq cohort, although all four TFs were downregulated in SHH tumours, they showed different expression patterns (Figure A.6). *OTX2* and *NEUROG1* did not show big differences in

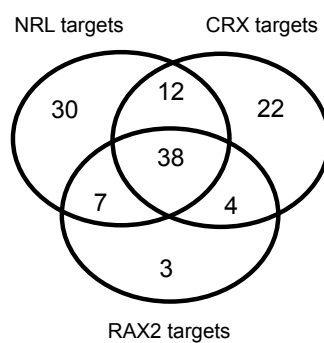


Figure 5.12: Overlap of putative targets between the photoreceptor signature defining TFs CRX, NRL and RAX2.

non-SHH subgroups, whereas *RREB1* and *TBR1* showed strong upregulation in Group 3 and Group 4 compared to WNT MB. The NIS ranking also reflected these expression patterns, where *RREB1* did not rank among the top TFs in the WNT subgroup. *RREB1* is expressed in the cerebellar nuclei and Purkinje cells layer of adult mice (Figure A.7) and was reported to be highly evolutionary conserved [244]. Overexpression of *RREB1* has been described, among other tumour types, in prostate cancer. In this tumour type, *RREB1* positively regulates expression of the lnc gene *AGAP2-AS1*, which promotes prostate cancer proliferation and migration via transcriptional repression of tumour- and metastasis-suppressing genes [245]. In our analysed RNA-seq MB cohort, *AGAP2-AS1* was differentially expressed between MB subgroups. However, we could not identify an obvious expression association between these two genes as indicated by the dissimilar expression patterns of *AGAP2-AS1* and *RREB1* in our data set (Figure A.8 and A.6). *OTX2* showed a high NIS in Group 3 (Figure A.2) and was closely positioned to *NRL*, *CRX*, and *RAX2* in the GRN (Figure 5.11.b) due to shared targets. *OTX2* has been previously shown to be a regulator of the photoreceptor expression signature and promoter of proliferation in MB [246]. *NEUROG1* is known to define non-SHH MBs and is expressed during brain development, particularly in Purkinje cell progenitors in the ventricular and intermediate zone of the developing cerebellum in the embryonic and postnatal stage [247–250]. *TBR1* is frequently mutated in Group 4 MB and known to be upregulated in Group 3 and Group 4 [251]. *TBR1* is expressed in the NTZ by cells that migrated from the rhombic lip into the NTZ [252]. Taken together, all four TFs are expressed in cell types of the cerebellum or associated with cerebellar development, while *RREB1* is not well known in MB.

Our performed DGEA and inferred GRN of the main subgroups could reveal TFs that were unknown to be specifically upregulated in Group 4 MBs or so far unknown to be an important regulator of gene expression in Group 4 MBs (Figure 5.11.b-c). The TFs *NEUROD2* and *ZBTB18*, which are involved in the development and differentiation of cerebellar neurons [253, 254], showed the highest NI-scores for Group 4 MBs in our TF ranking (Figure 5.11.c). Despite the upregulation in Group 4 samples, we observed that both TFs showed a slightly different expression profile across subgroups. *ZBTB18* was well expressed in all MB samples and subgroups, whereas *NEUROD2* was ubiquitously expressed only in WNT and Group 4 MBs (Figure A.9). Additionally, expression levels of *NEUROD2* were similar between Group 4 MBs and pre-/postnatal human cerebellum controls, whereas *ZBTB18* was upregulated around two-fold in Group 4 MB compared to the controls. However, considering the functions of *NEUROD2* and *ZBTB18* in cerebellar development and their potential high impact on gene expression in Group 4 MBs, *NEUROD2* and *ZBTB18* could play a decisive role in regulating the neuron-developmental signature in Group 4 MB. The third-ranked TF *ZBTB8B* has been recently mentioned to be upregulated in Group 4 MB without indicating a potential regulatory role [218]; however, this gene is poorly studied independently of a medulloblastoma context. Our data revealed that the TF *CHD5* was upregulated in Group 4 with a medium NIS in Group 4 (Figure 5.11.b-c). *CHD5* was reported to be a tumour-suppressor and chromatin-remodelling protein that is a transcriptional repressor or activator and necessary for neurogenesis [255]. The TF *LMX1A* did not rank among the top TFs in our performed analysis suggesting that this TF has a weaker impact on gene expression in Group 4, which is contrary to our previous work in Lin *et al.* [218]. In the currently analysed RNA-seq cohort, *NEUROD2* and *ZBTB18* showed a clear upregulation in all Group 4 samples, whereas *LMX1A* was low expressed in several Group 4 samples. This might explain why our estimation by the NIS suggests that *LMX1A* has a low impact on gene expression in Group 4 tumours compared to *NEUROD2* and *ZBTB18* (Figure A.9). The 8th-ranked TF *MYT1L* has been shown to play a regulatory role in Group 4 MB by Lastowska *et al.* [229].

Taken together, the functional enrichments and GRN in MB subgroups that we show above were supported in several aspects by the literature, which provides additional evidence to the ChIP-seq-based validation that our data and inferred GRN allow conclusions about regulatory mechanisms in MB. However, our presented results reveal activation of FGF signalling in WNT, *RAX2* as an additional regulator of the photoreceptor signature in Group 3, and *NEUROD2* and *ZBTB18* as a potential regulator of the neuronal-developmental signature in Group 4.

5.3.2.4 Subclusters within SHH medulloblastoma

The GRN that we inferred for the subclusters within the SHH subgroups split into two subnetworks relating to adult (SHH-c2) and non-adult (SHH-c1 and SHH-c3) SHH MBs (Figure 5.13a). The intermingled GRN of SHH-c1 and SHH-c3 probably arises from the hierarchical relationship between the two subclusters, where genes upregulated in subcluster SHH-c2 were also upregulated in subcluster SHH-c1, as summarised above (Figure 5.4).

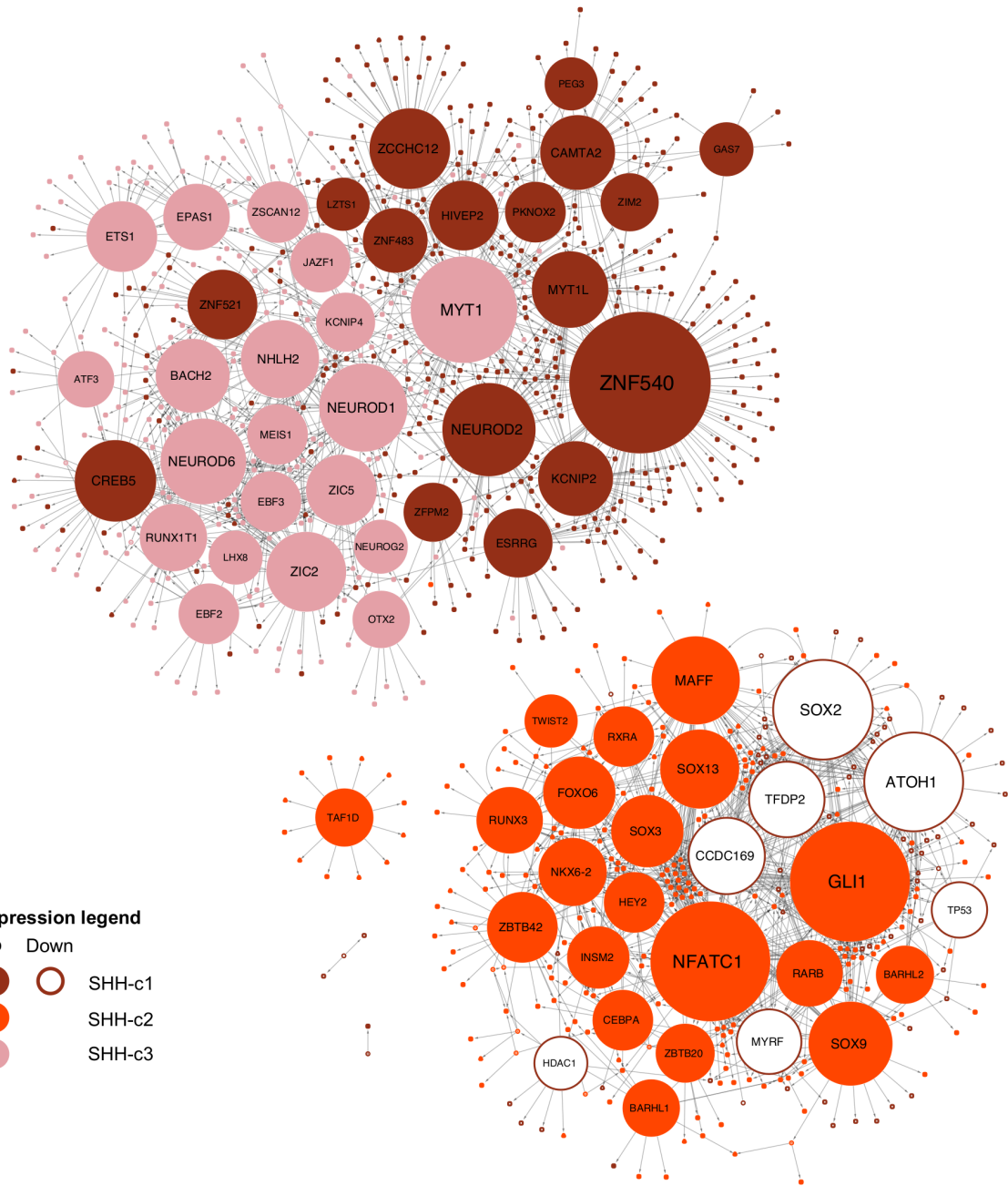
The TF *ZNF540* that was top-ranked for SHH-c1 MBs in our analysis is poorly studied (Figure 5.13) (Figure 5.13b). Our data showed that *ZNF540* was higher expressed in Group 3, Group 4, and SHH-c1 MBs compared to WNT and remaining SHH samples (Figure A.11, A.13). *NEUROD2* and *MYT1L*, which ranked among the top Group 4 TFs (Figure 5.11), showed the second and fourth highest NIS for SHH-c1 MBs, respectively. However, both TFs were lower expressed in SHH-c1 compared to Group 4 MB (Figure A.11). *Myt1l* was reported to be a neuron-specifically-expressed TF that ensures the cell fate of neurons by repressing non-neural programs and promotes neuronal differentiation in neural stem cells [257]. Taken together, our results show an overlap of TFs that were top-ranked in both SHH-c1 and Group 4 MBs (Figure A.11). These TFs might contribute to the enriched neuronal-developmental signature that we detected among the genes upregulated in subcluster SHH-c1 (5.14). This enrichment was previously described for the matching published subtypes SHH β and SHH γ of the Cavalli study [221].

The TFs *NFATC1* and *GLI1* showed the highest NIS in subcluster SHH-c2 (Figure 5.13b). Both genes ranked among the top ten TFs for the whole SHH subgroup (Figure 5.11). As introduced above, *GLI1* is an Shh-signalling target. *NFATC1* plays an important role in endothelial and osteoclast differentiation and is upregulated via Shh-signalling [258, 259]. The upregulation of *NFATC1* in adult SHH MBs was previously reported [256]. Our data showed that the TFs downregulated in SHH-c1 integrated into the subnetwork of SHH-c2 including the TFs *ATOH1* and *SOX2* that were top-scored for the SHH subgroup. In general, Shh-signalling-associated TFs in SHH-c2 were upregulated against non-adult subcluster SHH-c1 (infant) and SHH-c3 (children) or downregulated in SHH-c1 against SHH-c2 and SHH-c3. Here, the expression profiles followed the trend that Shh-signalling-associated genes were the lowest, intermediately, and the highest expressed in SHH-c1, SHH-c3, and SHH-c2, respectively (Figure A.11, A.14). At fourth position ranked *SOX9* (Figure 5.13b). *SOX9* showed an expression pattern among SHH subclusters that was similar to other Shh-signalling-associated TFs (Figure A.15, A.14). Like *SOX2* and *ATOH1*, *SOX9* was reported to be upregulated by Shh-signalling and is important for maintaining neuronal progenitors [237, 260, 261]. All three TFs could contribute to the functional enrichment of negative regulation of cell differentiation that we detected among genes upregulated in SHH-c2 MBs (Figure 5.14) since *SOX2* and *ATOH1* integrated into the subnetwork of SHH-c2 MBs (Figure 5.13a). Overall the GRN in SHH-c2 and its most influencing TFs supported the enrichment of active Shh-signalling that we observed among genes upregulated in SHH-c2 MBs (Figure 5.14). This enrichment was also reported for the matching subtype SHH δ by Cavalli *et al.* [221].

In subcluster SHH-c3, we detected an enriched TP53 signature (Figure 5.14); the matching published subtype SHH α is frequently *TP53*-mutated [221]. In subcluster SHH-c3, upregulated genes were enriched for angiogenic functions including *ETS1* and the angiogenesis-mediating receptor tyrosine kinases *FLT1* (*VEGFR1*), *KDR* (*VEGFR2*), and *TEK* (*TIE2*) (Figure A.12) [262, 263].

In subcluster SHH-c3, the three top-ranked TFs *NEUROD6*, *MYT1*, and *NEUROD1* are involved in neuronal differentiation [264–266]. *NEUROD6* has been reported to be upregulated in non-adult SHH [214]. As shown above, genes that showed upregulation in the subcluster SHH-c3 (children) compared to SHH-c2 (adults) were also upregulated in SHH-c1 MBs (infant) (Figure 5.4). The expression of the three TFs *NEUROD6*, *MYT1*, and *NEUROD1* followed a common pattern. Their expression was the highest, intermediate, and the lowest in SHH-c1, SHH-c3, and SHH-c2, respectively. This expression pattern was opposite to the expression of Shh-signalling-associated TFs in SHH-c2. *NEUROD1* showed the strongest upregulation in SHH-c1 and SHH-c3 compared to SHH-c2, followed by *NEUROD6* and *MYT1* (Figure A.11, A.16, A.17, A.18). *MYT1* showed a bimodal distribution in SHH-c2, where *MYT1*-

a)



b)

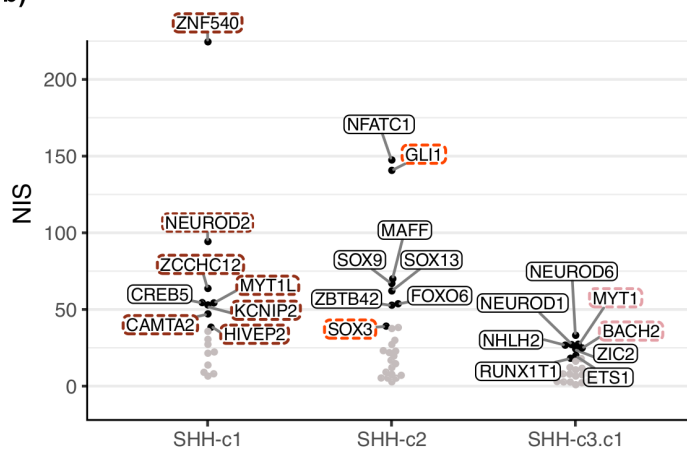


Figure 5.13 (preceding page): GRN of SHH subclusters. **a)** Inferred GRN of SHH subclusters. Enlarged nodes represent TFs. The size of a TF node relates to the out-degree of a TF node. Target genes comprise coding and lnc genes and are shown as small nodes. The colours of the nodes indicate subcluster-specific expression. Filled nodes relate to subcluster-specific upregulation. Unfilled nodes relate to subcluster-specific downregulation. **b)** Bee swarm plot illustrates the ranking of TFs by their impact on subcluster-dependent gene regulation in GRN shown in panel a). The impact on gene regulation was measured by the NIS. Top-eight-ranked TFs per subcluster are shown. TFs that have not been previously reported to be differentially expressed between SHH subclusters are highlighted by a dashed, coloured frame [207, 214, 221, 256]. **a-b)** Gene symbols are not written in italic to improve legibility.

high-expressing cases showed an expression level that was comparable to SHH-c1 and SHH-c3 cases (Figures A.17). Due to the function of these three TFs, SHH-c3 and SHH-c1 MBs (non-adult) could resemble a more differentiated cell type than SHH-c2 MBs (adult) with a potential higher differentiation in SHH-c1 vs. SHH-c3.

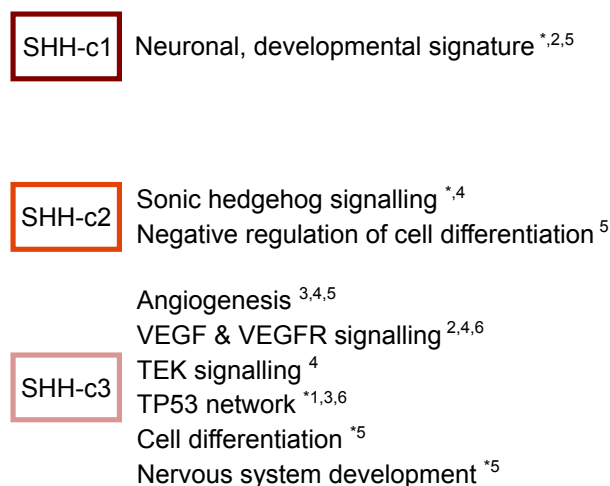


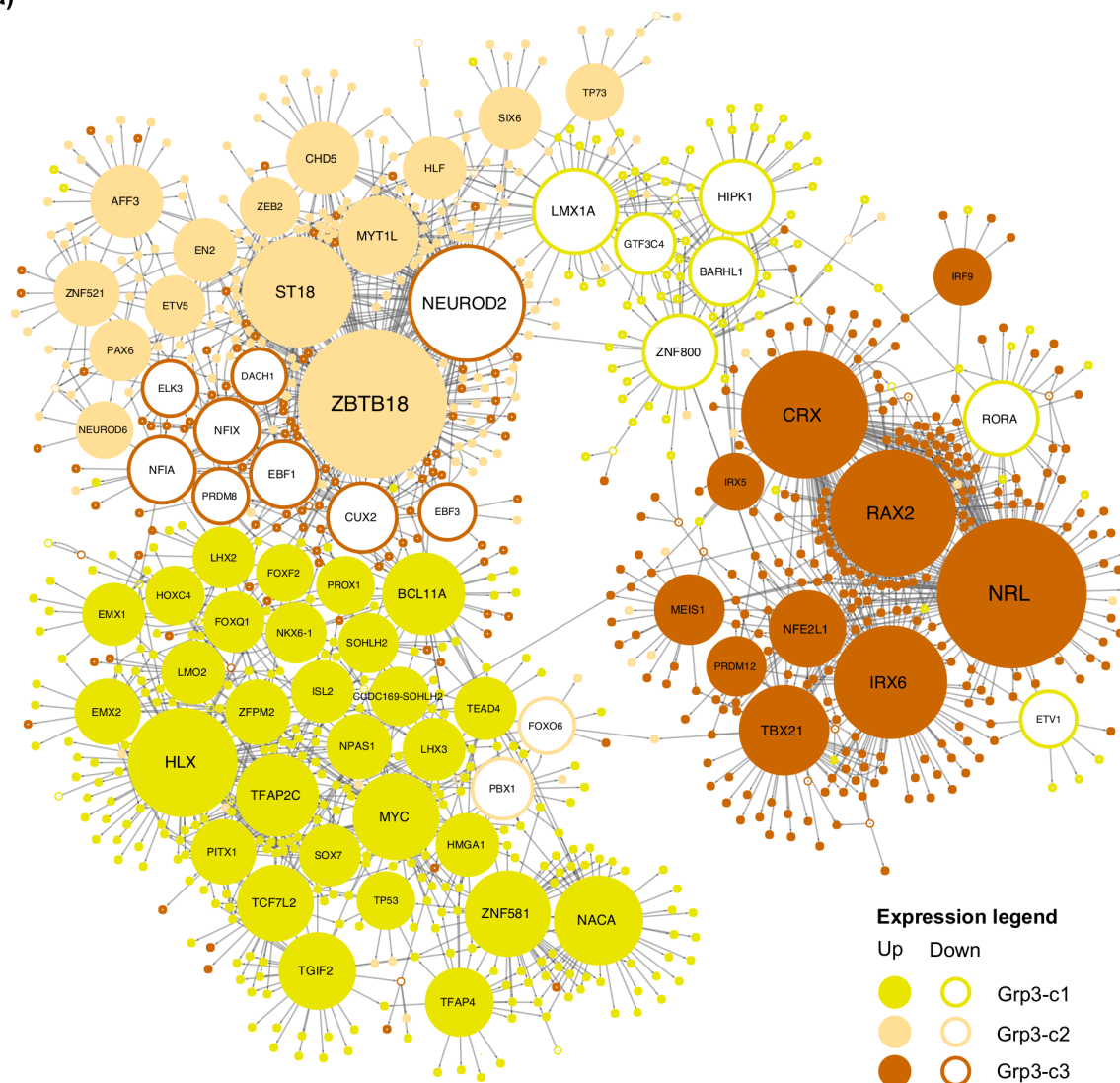
Figure 5.14: Functional enrichments in SHH subclusters. Enriched gene sets per SHH subcluster relate to subcluster-specifically upregulated coding genes. The asterisk indicates enrichments that are known from the literature [212, 221]. Superscript 1-5 indicates the original source of the gene sets that show an enrichment: ¹ KEGG, ² Reactome, ³ WikiPathways, ⁴ Genomatix, ⁵ GO Terms. FDR \leq 0.05.

5.3.2.5 Subclusters within Group 3 medulloblastoma

Genes and TFs that were specifically upregulated among the Group 3 subclusters formed three distinct subnetworks in the GRN (Figure 5.15a). Genes and TFs downregulated in Grp3-c1 formed a subnetwork but shared edges with the Grp3-c2 or Grp3-c3 subnetwork. Genes downregulated in Grp3-c3 MB integrated for the most part into the Grp3-c2 subnetwork.

In Grp3-c1 MB, the two top-ranked TFs were *MYC* and *HLX*, which are located on chromosome 8 and 1, respectively (Figure 5.15b). Grp3-c1 samples showed frequent and strong copy number gain of *MYC* resulting in a pronounced upregulation in Grp3-c1 tumours (Figure 5.7). The upregulation of *MYC*

a)



b)

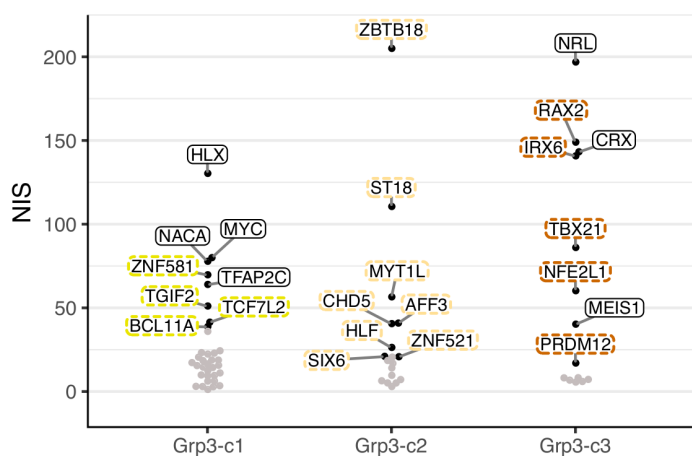


Figure 5.15 (preceding page): GRN of Group 3 subclusters. **a)** Inferred GRN of Group 3 subclusters. Enlarged nodes represent TFs. The size of a TF node relates to the out-degree of a TF node. Target genes comprise coding and lnc genes and are shown as small nodes. Colours of the nodes indicate subcluster-specific expression. Filled nodes relate to subcluster-specific upregulation. Unfilled nodes relate to subcluster-specific downregulation. **b)** Bee swarm plot illustrates the ranking of TFs by their impact on subcluster-dependent gene regulation in GRN shown in panel a). The impact on gene regulation was measured by the NIS. Top-eight-ranked TFs per subcluster are shown. TFs that have not been previously reported to be differentially expressed between Group 3 subclusters are highlighted by a dashed, coloured frame [207, 208, 221, 267, 268]. **a-b)** Gene symbols are not written in italic to improve legibility.

was also reflected by an enrichment for *MYC* targets that we detected among genes upregulated in subcluster Grp3-c1 (Figure 5.16), a feature that has been described for the matching published subtype Group3 γ [221]. However, in our RNA-seq data, the remaining two Group 3 subclusters displayed as well *MYC* upregulation compared to SHH and Group 4 MBs. In WNT MB, *MYC* was reported to be upregulated due to Wnt signalling [269]. *MYC* is known as an oncogene in MB, especially in a subset of Group 3 MBs where *MYC* is activated due to copy number gain [269]. Grp3-c1 tumours probably reflect this subset of Group 3 MB. *HLX*, which showed the highest rank, is involved in multiple developmental processes including CNS development and haematopoiesis [270–272]. In our MB cohort, *HLX* was well expressed in Grp3-c1 cases only, whereas *MYC* was clearly expressed in almost all MB samples (Figure A.19). An expression correlation between *HLX* and *MYC* was only present in Group 3 MB but not in non-Group 3 MB (Figure A.20.a-c). Lin *et al.* have reported a regulatory role of *HLX* in general for Group 3 MBs via binding to enhancer regions, but not in the context of Group 3 subclusters [218]. However, *HLX* upregulation in *MYC*-amplified Group 3 cases, equivalent to Grp3-c1 cases, has been previously reported [208]. Overall, the data that we presented suggest that *HLX* has a specific role in Grp3-c1 tumours associated with *MYC*.

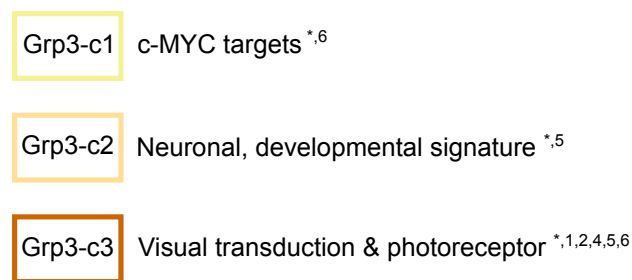


Figure 5.16: Functional enrichments in Group 3 subclusters. Enriched gene sets per Group 3 subcluster relate to subcluster-specifically upregulated coding genes. The asterisk indicates enrichments that are known from the literature [221]. Superscript 1-5 indicates the source of the gene sets that show an enrichment: ¹ KEGG, ² Reactome, ³ WikiPathways, ⁴ Genomatix, ⁵ GO Terms. FDR \leq 0.05.

Five of the top-ranked TFs in Grp3-c2 MB overlapped with TFs that were among the highest-ranked TFs in Group 4 MBs comprising *ZBTB18*, *ST18*, *SIX6*, *CHD5* and *MYT1L* (Figure 5.15b, A.11, 5.11).

NEUROD2, which was also highly ranked among the Group 4 TFs and downregulated in Grp3-c3, was incorporated into the Grp3-c2 GRN (Figure A.11, 5.11). These six Group 4-relevant TFs showed expression levels in Grp3-c2 MBs that were similar to Group 4 MBs (Figure A.11, A.9). Additionally, genes upregulated in subcluster Grp3-c2 (249 genes) showed significant overlap (96 genes, $p = 2.27 \times 10^{-117}$, hypergeometric test) with genes upregulated in subgroup Group 4 (323 genes). Genes upregulated in subcluster Grp3-c2 were enriched for a neuronal-developmental signature (Figure 5.16), as reported for the matching subtype Group 3 β by Cavalli *et al.* [221], like in subgroup Group 4. However, Group 3 relevant TFs were still upregulated in Grp3-c2 MBs compared to the remaining three subgroups (Figure A.11). The expression pattern of Group 3 and Group 4 relevant TFs indicated that Grp3-c2 represents a mixed phenotype of Group 3 and Group 4 MB.

NRL, *CRX*, and *RAX2* showed a strong influence on gene expression in the photoreceptor-signature-expressing subcluster Grp3-c3. As shown above, these three TFs were highly ranked for the whole subgroup Group 3 potentially regulating the photoreceptor signature of this subgroup (Section 5.3.2.3, Figure 5.11.c). All three TFs were upregulated in Group 3 MBs vs. the remaining subgroups but subcluster Grp3-c3 additionally showed significantly higher expression among Group 3 subclusters (Figure A.11, A.21, A.22, A.23). It explains why *NRL*, *CRX*, and *RAX2* can significantly impact expression in the whole subgroup Group 3 and additionally in subcluster Grp3-c3 that was also enriched for a photoreceptor signature (Figure 5.16). This enrichment was also reported for the matching subtype Group 3 α by Cavalli *et al.* [221]. In our analysis, *IRX6*, which is not well described in MB, ranked fourth with a slightly lower NI-score than *CRX*. In mice, it was reported that *Irx6* is required for terminal differentiation of specific retinal bipolar interneuron subtypes [273]. Bipolar interneurons transfer transduced signals from rods and cones into deeper cell layers of the retina [274]. In humans, it was reported that *IRX6* has a rod-enriched expression pattern [275]. *IRX6* indicates that potentially more TFs than *NRL*, *CRX*, and *RAX2* might contribute to the photoreceptor signature in MB.

5.3.2.6 Subclusters within Group 4 medulloblastoma

Subcluster-specifically upregulated genes were enriched for (1) a neuronal-developmental signature in all subclusters of Group 4 MBs, (2) FGF signalling in Grp4-c1, and (3) a photoreceptor signature and PI3K-AKT signalling in Grp4-c3 (Figure 5.17), as previously reported for related subtypes [221].

Grp4-c1	FGF signalling ^{*.4} Neuronal, developmental signature ^{*.5}
Grp4-c2	Neuronal, developmental signature ^{*.5} PI3K-AKT signalling ^{1,3}
Grp4-c3	Visual transduction & photoreceptor ^{*.1,2,4,5,6} PI3K-AKT signalling ^{*.1,3} Neuronal, developmental signature ^{*.5}

Figure 5.17: Functional enrichments in Group 4 subclusters. Enriched gene sets per Group 4 subcluster relate to subcluster-specifically upregulated coding genes. The asterisk indicates enrichments that are known from the literature [221]. Superscript 1-5 indicates the original source of the gene sets that show an enrichment: ¹ KEGG, ² Reactome, ³ WikiPathways, ⁴ Genomatix, ⁵ GO Terms. FDR ≤ 0.05 .

Genes and TFs upregulated in one of the three subclusters of Group 4 MBs formed three distinct subnetworks in GRN that we inferred (Figure 5.18a). The TF *EBF1*, which is expressed in pre- and postnatal mouse cerebellum [276], had the highest NIS in Grp4-c1 MB and ranked among the top

ten TFs of Group 4 MB (Figure 5.18b, 5.11). *EBF1* was well expressed in all MBs but clearly showed the highest expression in Grp4-c1 MB compared to the remaining subgroups and subclusters (Figure A.24); WNT and Grp3-c3 MBs showed the lowest expression; SHH, Grp4-c2, Grp4-c3, Grp3-c1 and Grp3-c2 MBs showed a similar medium expression level. These results indicate that the upregulation expression of *EBF1* is rather specific to Grp4-c1 tumours than a feature of the whole subgroup Group 4, highlighting a potential role of *EBF1* as TF in Grp4-c1 MBs. *EBF1* was reported known to be upregulated in Group 4, and Cavalli *et al.* has reported differential expression between Group 4 subtypes [221, 229]. A potential regulatory role of *EBF1* in Group 4 subclusters has not been described [218, 277]. TF *ZNF521* ranked at the eighth position, a gene that promotes proliferation in a SHH MB cell line [267]. In addition, *ZNF521* ranked among the top eight TFs of the Group 4-like subcluster Grp3-c2 and was upregulated in SHH-c1 among SHH subclusters (Figure 5.15).

Highly-ranked TFs in Grp4-c2 MB showed a distinct pattern because these TFs reached top NI-scores only in subcluster Grp4-c2. In the remaining Group 3 and Group 4 subclusters, at least one top-ranked TF overlapped with a top-ranked TFs of another subcluster or subgroup (Figure A.11). The second-ranked TF *TBR1* was already mentioned above among the TFs that were downregulated in the main subgroups SHH (Section 5.3.2.3). Interestingly, our data showed that several TFs/genes upregulated in subcluster Grp4-c2 have been linked with EMT, neuronal stem/progenitor cells, or stem-like cancer cells. Among these TFs was *TWIST1* that showed the highest NIS in Grp4-c2 tumours (Figure 5.18b). *TWIST1* has been reported in various cancer types, where *TWIST1* has different implications, including EMT and stemness [278]. Upregulation of *TWIST1* has been reported in WNT MB [279]. In our data, *TWIST1* was highly expressed in all Grp-c2 MBs on a level similar to *TWIST1*-high-expressing WNT MBs. Across all subgroups and subclusters, constant upregulation of *TWIST1* was only present in Grp4-c2 MBs (Figure A.25.a). Kahn *et al.* have shown that *TWIST1* is upregulated in metastases of Group 3 MB compared to primary tumours of this subgroup. The authors reported that *TWIST1* promotes metastasis in Group 3 tumours through transactivation of *BM11*, a mediator of *TWIST1*-induced EMT [280, 281]. We could detect a significant expression correlation between *TWIST1* and *BM11* only in Group 3 MB, but not in Group 4 MB (Figure A.25.c, A.25.b). The TF *SOX11* was in fourth position. Only Grp4-c2 MBs showed a specific upregulation of *SOX11* compared to the remaining subclusters and subgroups (Figure A.26). *SOX11* has been described to be involved in cerebellar development (pre- and postnatal stages) in mice and to be expressed in neuronal progenitors/immature neurons, mesenchymal stem cells, and cancer stem-like cells [282–284]. Our DGEA showed that Grp4-c2 tumours expressed *NES* and *SOX9* at a significantly higher level compared to the subcluster Grp4-c1 and Grp4-c3 (Figure A.27). Both genes have been reported to be associated with neural stem/progenitor cells, stem-like cancer cells, and EMT in cancer [285–289]. In our RNA-seq MB cohort, *SOX9* was higher expressed in SHH MB (where the gene is upregulated via SHH signalling [290]) compared to Grp4-c2 MBs. Across subgroups and subclusters, *NES* showed the highest expression in Grp4-c2 and WNT MBs (*NES* is a target of the WNT pathway [291]). Moreover, *NES* expression could be associated with the enrichment for PI3K-AKT signalling in Grp4-c2 MBs. Previous reports have shown that *Nes* expression is required for the activation of Pi3k-Akt signalling in cortical neural progenitor cells by phosphorylation of Akt [292]. We checked the expression of genes of the *AKT* and *PIK3* family in our data. These genes were well expressed in all MB samples (exemplarily shown for *AKT1* and *PIK3CA* in Figure A.28). Moreover, we found that the PI3K-AKT pathway was enriched among gene upregulated in subcluster Grp4-c2. The PI3K-AKT pathway was reported to play a role in the maintenance of pluripotency (stem-like features) and EMT [293, 294], which is consistent with the mentioned functions of *TWIST1*, *SOX11*, *SOX9*, and *NES*.

The photoreceptor signature defining TFs *NRL*, *CRX* and *RAX2* showed the highest NIS in Grp4-c3, which was in agreement with the observed gene set enrichment for this subcluster (Figure 5.18b, Figure 5.17). The TF *LHX4* ranked in the fourth position closely to *CRX*. *LHX4* was higher expressed in Group 3 and Group 4 MB than in WNT and SHH MB but Grp4-c3 tumours displayed the strongest and most specific upregulation, except for two outliers, supporting a regulatory role in Grp4-c3 MB (Figure A.29).

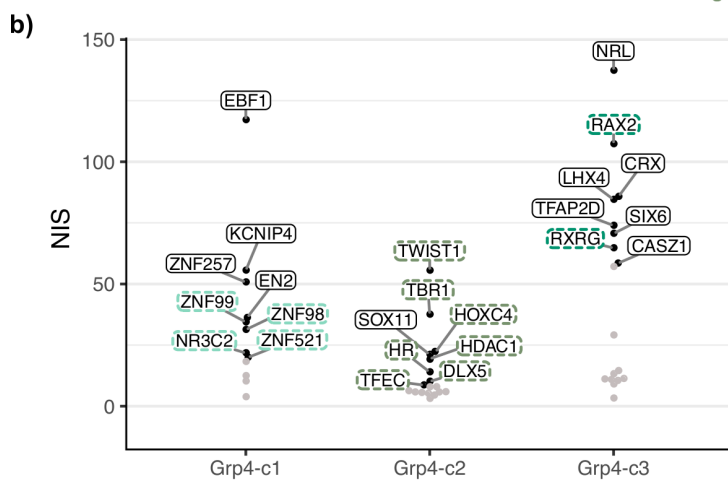
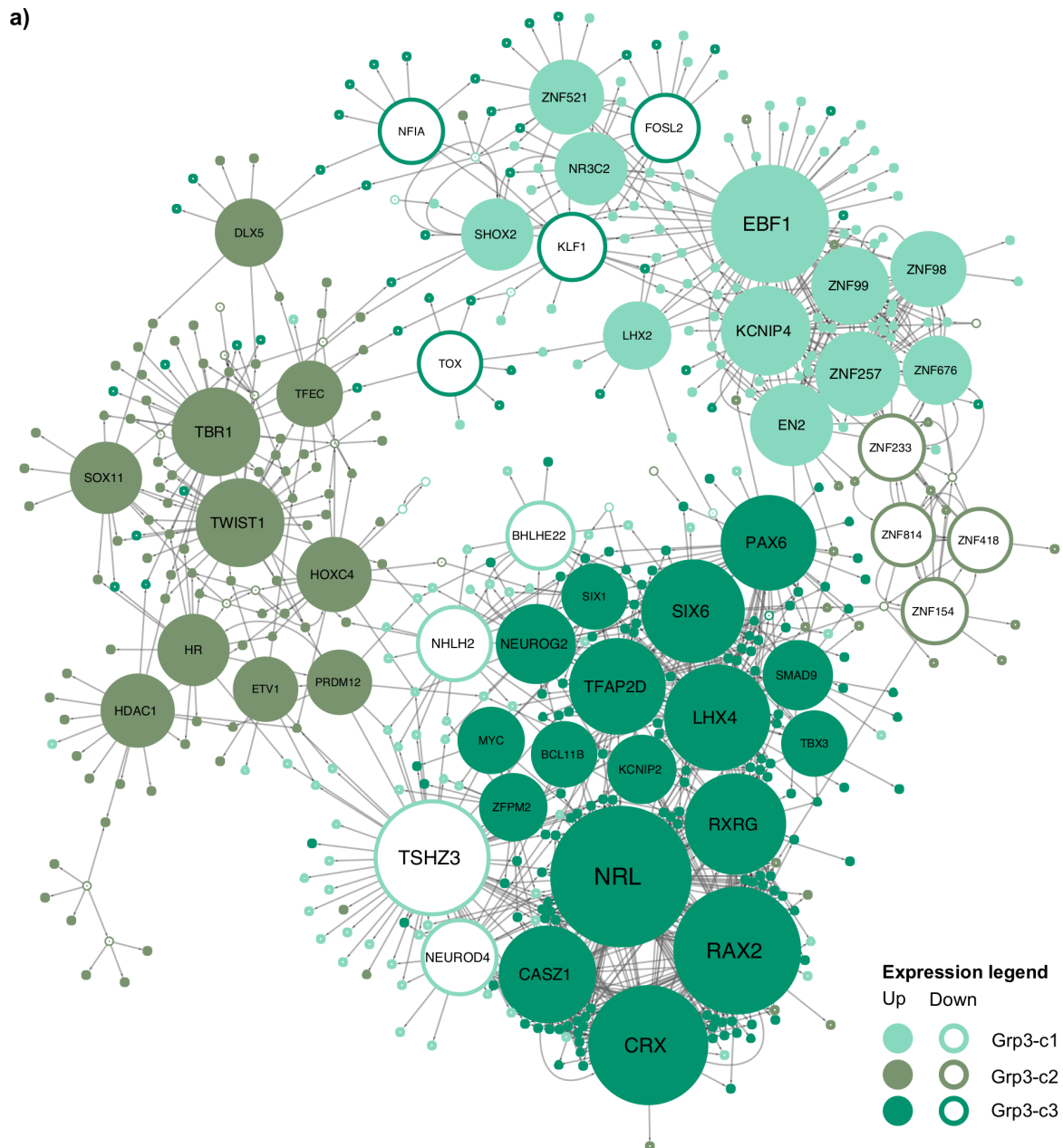


Figure 5.18 (preceding page): GRN of Group 4 subclusters. **a)** Inferred GRN of Group 4 subclusters. Enlarged nodes represent TFs. The size of a TF node relates to the out-degree of a TF node. Target genes comprise coding and lnc genes and are shown as small nodes. The colours of the nodes indicate subcluster-specific expression. Filled nodes relate to subcluster-specific upregulation. Unfilled nodes relate to subcluster-specific downregulation. **b)** Bee swarm plot illustrates the ranking of TFs by their impact on subcluster-dependent gene regulation in GRN shown in panel a). The impact on gene regulation was measured by the NIS. Top-eight-ranked TFs per subcluster are shown. TFs that have not been previously reported to be differentially expressed between Group 4 subclusters are highlighted by a dashed, coloured frame [208, 221, 233, 268]. **a-b)** Gene symbols are not written in italic to improve legibility.

LHX4 was reported to be involved in many developmental processes including in cone photoreceptors and retinal bipolar interneurons and, therefore, could contribute to the photoreceptor signature in subcluster Grp4-c3 [295–297].

5.3.2.7 Impact of copy number variations on transcription factor expression

Copy number variations are one mechanism that leads to dysregulation of gene expression in cancer (see Section 2.3). To evaluate the influence of CNVs on the above presented GRNs in MB, the following section will concentrate on the impact of CNVs on TF expression.

As presented above, 121 genes and TFs that were downregulated in WNT MB formed a network module (Section 5.3.2.3, 5.4.5.4). Here, 82/121 genes, which included 15/29 TFs, were located on chromosome 6 that is frequently monoallelically deleted in WNT MBs [251]. Thompson *et al.* already reported that a high fraction of genes downregulated in WNT MB are located on chromosome 6 and concluded a relation between chromosome 6 monosomy and gene expression [209]. For example in our data, the TFs *DEK* and *HSF2* showed a highly frequent copy number loss in WNT MB and copy number gains in non-WNT MBs (Figure 5.19a,d). 1N tumours showed a strong expression drop, whereas copy number gain did not influence gene expression, as previously reported for *DEK* in retinoblastoma (Figure 5.19b-c,e-f) [240]. The expression fold change comparing non-WNT vs. WNT MBs was 2.8 for *HSF2* and 2.6 for *DEK*, which is higher than the expected fold change of 2. However, 373 genes located on chromosome 6 and significantly downregulated in WNT tumours (FDR <0.001, subgroups pair-wise compared) displayed on average a fold change of 2. These data indicate that TFs and genes located on chromosome 6 are downregulated in WNT MB due to monosomy, but additional mechanisms also influence expression levels of these genes/TFs explaining the bigger fold change of *HSF2* and *DEK*.

Within the GRN of the subgroups, 184 genes that were specifically upregulated in Group 4 MB (including 20 TFs) formed a subnetwork, as shown above (Section 5.3.2.3, 5.4.5.4). Among these 184 Group 4-specific genes, 29/184 (15.7%) genes, including 6/20 (30%) TFs, were located on chromosome 17q that is known to be frequently gained in Group 3 and Group 4 [222]. Applying a hypergeometric test showed that it was significant ($p = 0.0208$) to find six TFs located on chromosome 17q among 20 TFs considering the 184 Group 4-specific genes as background. These six TFs located on chromosome 17q comprised *NEUROD2*, which showed the second-highest NIS (Figure A.30) and high expression in Group 4 (Figure 5.20.a). Lo *et al.* emphasised an association between copy number gain and upregulation of *NEUROD2* in MB [298]. This report of Lo *et al.* was evaluated using the analysed RNA-seq MB cohort. Comparing *NEUROD2* expression between 2N and >2N cases across the cohort showed an association between upregulation and copy number gain (Figure 5.20.c), where Group 3 and Group 4 tumours frequently showed a gain of *NEUROD2* (Figure 5.20.b). However, among

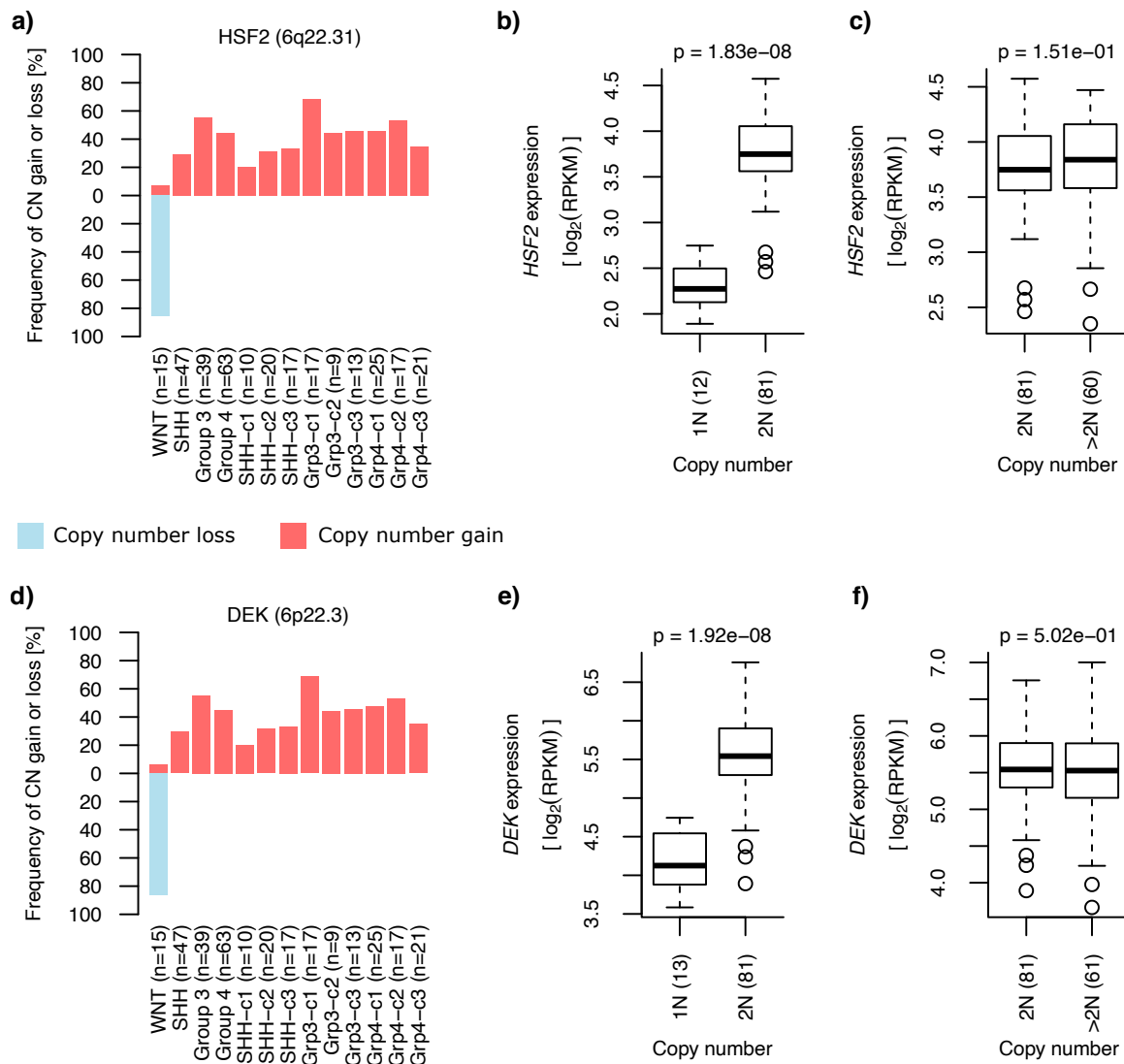


Figure 5.19: Copy number and expression of TFs downregulated in WNT MB. **a-c)** *HSF2*. **d-f)** *DEK*. **a, d)** Bar plots show the percentage of MB samples that showed a copy number loss or gain. **b-c, e-f)** Box plots compare the gene expression between samples with different copy numbers (1N, 2N, >2N). The numbers of samples are shown in brackets. Copy-number-based groups were tested for differential expression using Wilcoxon rank-sum test, and p-values are placed above the box plot.

Group 3 or Group 4 samples only, *NEUROD2* did not show a clear association between copy number gain and upregulation (Figure 5.20.d-e). Additionally, *NEUROD2* was lowly expressed in subcluster Grp3-c3 compared to the remaining Group 3 and Group 4 MB samples (Figure 5.20.a), even though the frequency of copy number gain was similar between Group 3 subclusters (Figure 5.20.b-c). Overall, the presented results emphasise that chromosome 17q copy number gain impacts the expression of TFs and the transcriptional regulation in Group 4 MBs. However, the expression profile of *NEUROD2* indicates that copy number and expression are not necessarily linear associated. Therefore, additional factors than copy number gain also influence the expression of *NEUROD2*.

5 Medulloblastoma study

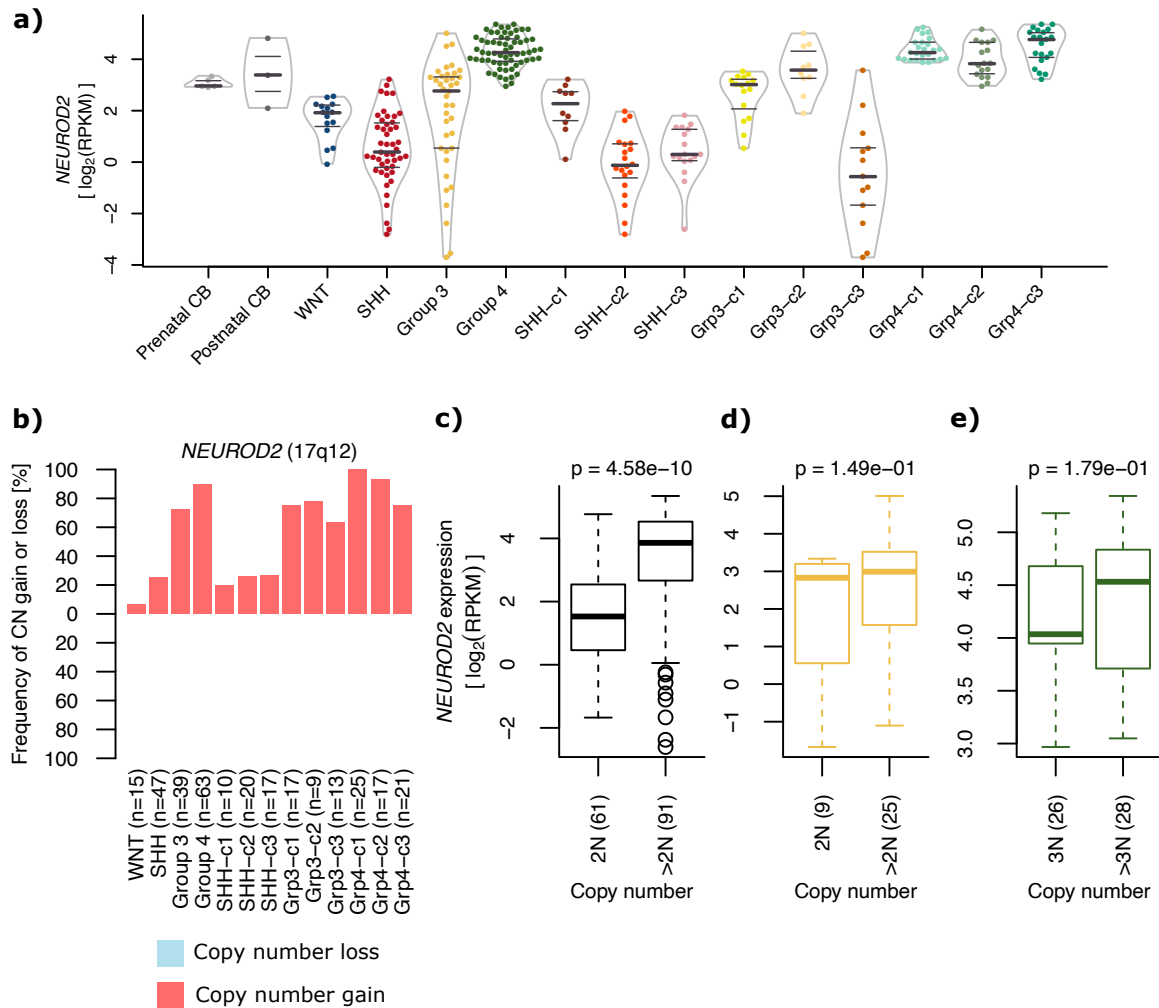


Figure 5.20: Copy number and expression of *NEUROD2*. **a)** Violin plot shows expression destitution of *NEUROD2*. 25%, 50% and 75% quantile are indicated by horizontal lines. Individual MB samples are shown as bee swarm plots. **b)** Bar plots show the percentage of samples that showed a copy number loss or gain. **c-e)** Box plots compare the gene expression between samples with different copy numbers (2N, >2N, 3N, >3N). The numbers of samples are shown in brackets. Copy-number-based groups were tested for differential expression using Wilcoxon rank-sum test, and p-values are placed above the box plot. **c)** Whole cohort. **d)** Group 3 samples. **e)** Group 4 samples.

5.3.2.8 Overlay of gene regulatory networks in medulloblastoma subgroups and subclusters

In the previous sections, we described the GRNs and their most influencing TFs in MB subgroups and subclusters (Figure A.11). In this section, we will focus on the comparison and rationalisation of the GRNs in subgroups and subclusters. The gene regulatory network of the subgroups was compared to the combined GRNs of the subclusters. Here, ~33% of the target genes, ~38% of the TFs, and 14% of the inferred TF-target interactions were shared between the GRNs of subgroups and subclusters (Figure 5.21.a-c). Considering that ~44% of the subgroup-specifically expressed genes were also significantly differentially expressed between subclusters, it would be expected that a maximum of 44% of targets and TFs in the subgroup network are also part of the subcluster networks. Overlaying the target genes

with published data showed that ~59% and ~52% of the target genes in the subgroup and subcluster GRNs, respectively, overlapped with genes whose expression had been reported to be correlated with the activity of nearby enhancers in MB (Figure 5.21.e-f) [218].

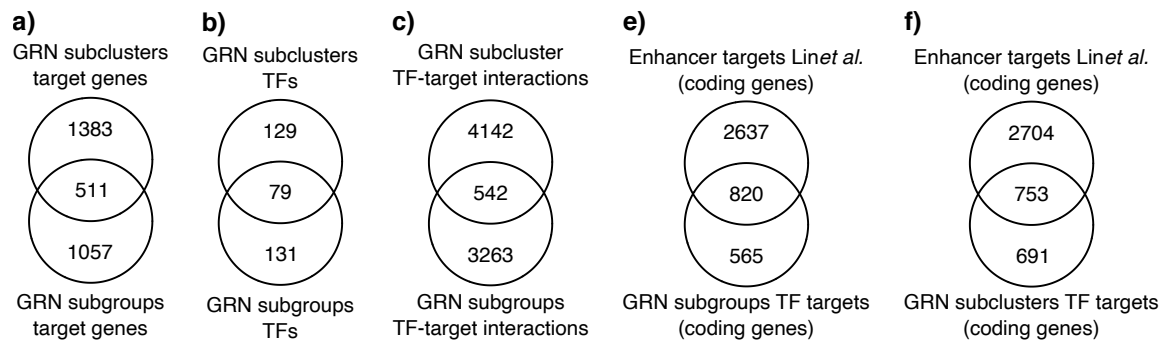


Figure 5.21: Comparison of subgroup and subcluster GRNs. **a-c)** Venn diagrams relate to a) target genes, b) TFs, and, c) inferred TF-target interactions. **d-e)** Venn diagrams compare the set of inferred TF targets and putative enhancer targets reported by Lin *et al.*. Both diagrams are restricted to coding genes because Lin *et al.* only reported coding genes [218]. TF targets in subgroup d) and subclusters e) are independently shown.

To further rationalise gene regulation in MB, the four GRNs of subgroups and subclusters were merged into one aggregated GRN. The aggregated GRN was used to represent the whole gene regulatory landscape in MB and to perform a TF-wise gene set overrepresentation analysis among TF targets.

We simplified the aggregated network to depict the whole gene regulatory landscape in MB and extract the most relevant information. Here, the aggregated GRN was reduced to the TFs, and two TFs were connected in the network when they shared a minimum of two targets. To visualise the relationship between subgroups and subclusters on the aggregated GRN, the aggregated GRN was plotted four times, and in each plot, the differentially expressed TFs of subgroups, SHH subclusters, Group 3 subclusters or Group 4 subclusters were individually highlighted, respectively (Figure 5.22; an enlarged version of this Figure is shown in Figure A.30, A.31, A.32, and A.33). TFs that were up- or downregulated in the same subgroup showed a strong tendency to be grouped together marking the subnetworks of WNT, SHH, Group 3 and Group 4 (Figure 5.22a). A dense subnetwork, including *RAX2*, *NRL* and *CRX*, comprised most TFs upregulated in Group 3 MB. TFs of the photoreceptor subclusters Grp3-c3 and Group4-c3 overlapped with Group 3 TFs. Most TFs of the subcluster Grp3-c1, including *HLX* and *MYC*, formed a subnetwork that was mostly separated from the Group 3 subnetworks. TFs upregulated in the adult SHH subcluster SHH-2 overlapped with TFs upregulated in the SHH subgroup. The upregulated TFs of non-adult SHH subcluster SHH-c1 overlapped partially with Group 4 TFs, whereas SHH-c3 TFs were placed between Group 3 and Group 4 TFs. Here, the GRN in subcluster SHH-c1 and SHH-c3 stretched beyond the GRN defined by Group 3 and Group 4 MB. TFs of the Group 4-like subcluster Grp3-c2 strongly overlapped with Group 4 TFs and partially with SHH-c3 TFs. Grp4-c2 TFs were placed mostly separately to the Group 3 and Group 4 GRN. TFs of subcluster Grp4-c1 partially overlapped with Group 4 TFs. In summary, the aggregated network depicted the commonalities and dissimilarities of GRNs among subclusters and between subgroups and subclusters.

Taking the aggregated GRN again, TF-wise gene set overrepresentation analysis among TF targets supported the aforementioned assumption that *ZBTB18* and *NEUROD2* are potential regulators of the neuronal-developmental signature in Group 4 and Grp3-c2 (Figure 5.23). The overrepresentation analysis also supported that *RAX2* is a regulator of the photoreceptor signature in cooperation with *NRL* and *CRX* in Group 3, Grp3-c3, and Grp4-c3 (Figure 5.23).

Overall, via the comparison and aggregation of the inferred GRNs, we could show that subgroup-associated regulatory networks partially contribute to subcluster-specific gene expression. However, the majority of the inferred TF-target interactions were specific for subgroups or subclusters. Here,

5 Medulloblastoma study

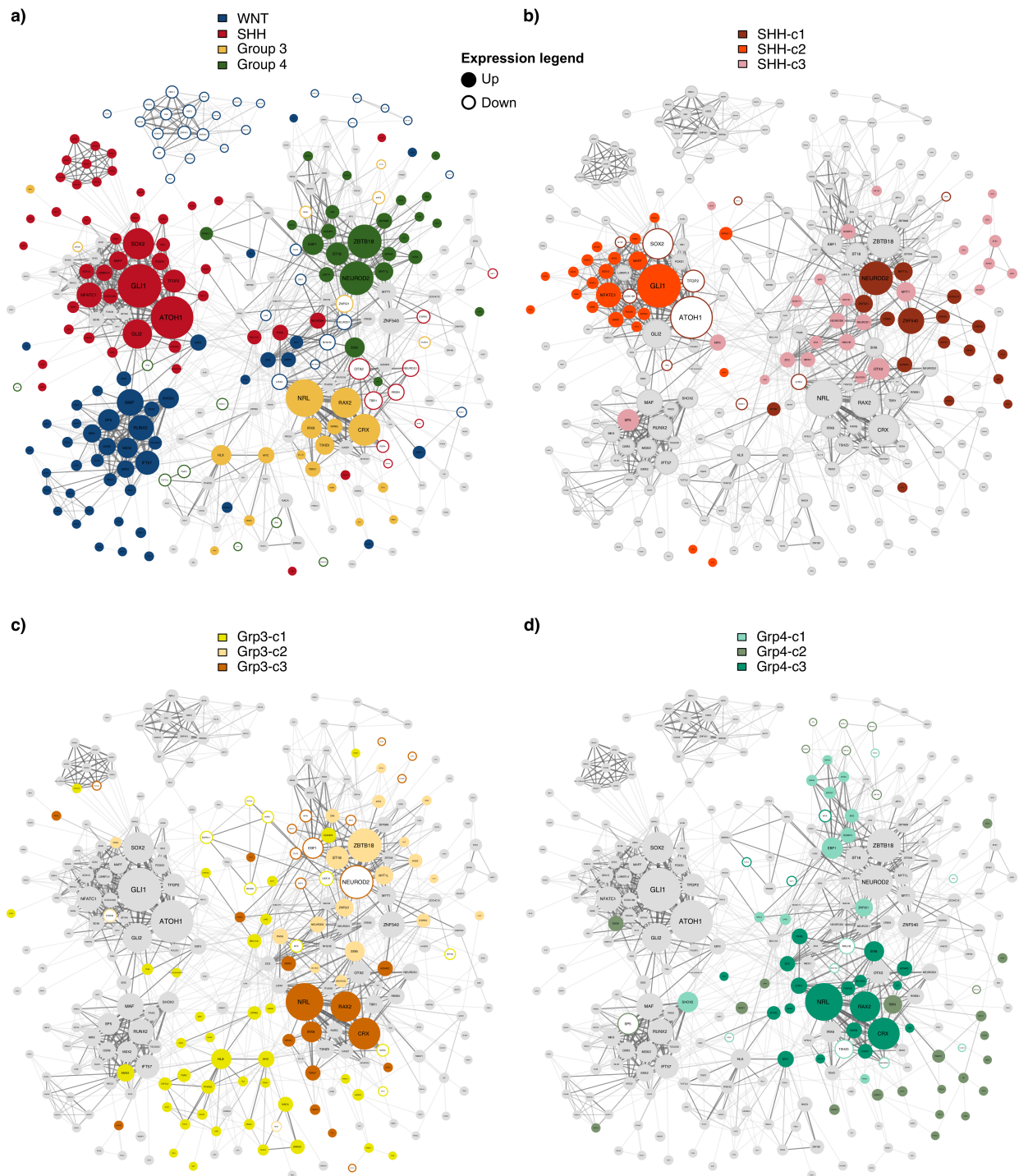


Figure 5.22: Aggregated GRN of MB. The aggregated GRN was reduced to TFs, and an edge between two TFs indicated at least two shared targets. The edge width indicates the number of shared targets. **a-d)** Colours highlight TFs that were differentially expressed between a) subgroups, b) SHH subclusters, c) Group 3 subclusters, d) Group 4 subclusters. Colour-filled circles indicate upregulation of TFs among subgroups or subclusters. White-filled circles indicate downregulation of TFs among subgroups or subclusters. Grey circles indicate no differential expression of TFs among subgroups or subclusters. Gene symbols are not written in italic to improve legibility.

subclusters expand and highlight new aspects of the gene regulatory networks in MB. Detected gene set enrichments among TF targets supported the identification TFs that contribute to the regulation of subgroup-specific gene expression signatures that were also present in subclusters.

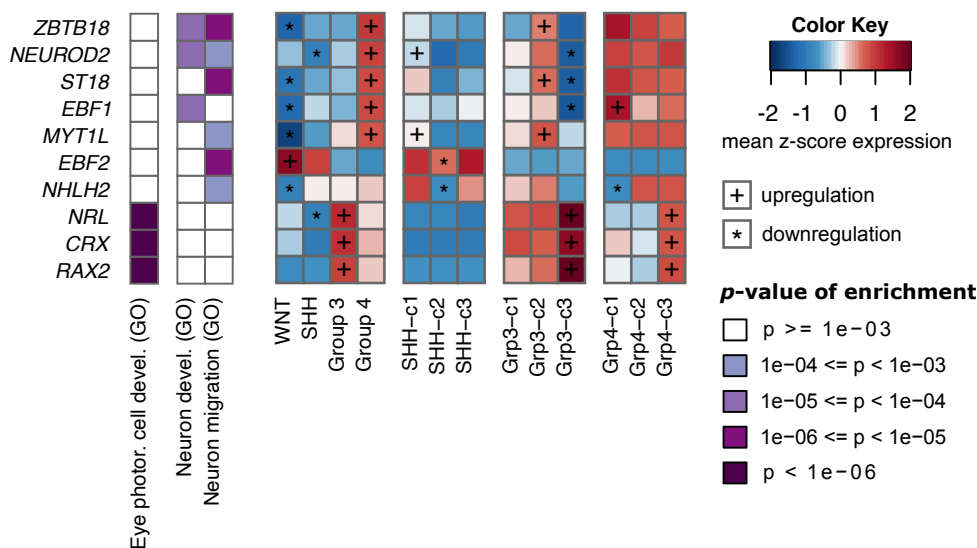


Figure 5.23: Functional enrichments among targets of selected TF. Per TF, functional enrichments and gene expression patterns in MB are shown. **Left heatmap)** P-values of functional enrichments are shown in categories as indicated by the caption. All functional enrichments showed an FDR < 0.05. The source of functional gene sets is indicated in brackets. **Right heatmap)** The heatmap summarises the expression patterns of the TFs in MB subgroups and subclusters. Significant up- or downregulation is indicated. Abbreviations: photor. - photoreceptor; devel. - development.

5.3.3 Lnc genes in medulloblastoma

Concentrating on lnc genes that are differentially expressed between subgroups or subclusters, in the next sections, we will present (1) a formal characterisation of lnc genes based on the genomic location relative to coding genes, (2) the annotation of brain-development-related expression patterns, (3) a summary of lnc genes that are described in the literature, and (4) the implication of the lnc tumour-suppressor *MEG3* in MB.

5.3.3.1 Potential relevance of lnc genes in medulloblastoma

As a simple test of the hypothesis that lnc genes have an implication in MB subgroups and subclusters, we evaluated whether lnc gene expression alone can recover MB subgroups and subclusters. Taking the 1643 most variable expressed lnc genes, MB samples were clustered by applying NMF and assuming four clusters (Figure 5.24). The lnc-gene-based clustering agreed for SHH and WNT to 100% with the methylation-based subgroup classification. 10/63 Group 4 and 3/39 Group 3 samples switched between lnc gene-based clusters. Discordances between Group 3 and Group 4 clustering can probably be explained by transcriptional similarities in a subsample of both subgroups [217], as reflected by the subclusters Grp3-c2 and Grp4-c3 (Section 5.3.2.5 and 5.3.2.6).

Applying hierarchical clustering, differentially expressed lnc genes mostly recovered the tumour subgroups and subclusters except for the subclusters SHH-c1 and SHH-c3 (Figure 5.25). A potential reason could be the low number of thirteen specifically upregulated lnc genes in SHH-c3.

Overall, the expression pattern of lnc genes resembled MB subgroups and subclusters, supporting the hypothesis that lnc genes have an implication in MB. The set of 448 identified differentially expressed lnc genes, as presented above (Section 5.3.2.1, Figure 5.8), was used to characterise the lnc gene landscape in MB.

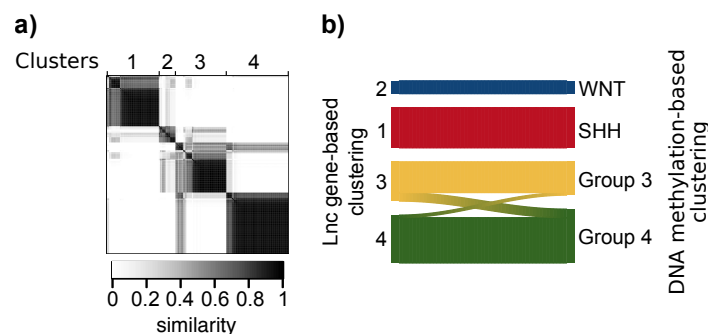


Figure 5.24: *De novo* identification of MB subgroups based on the expression profiles of the 1643 most variable lnc genes. **a)** Consensus matrix of the NMF clustering on RNA-seq-derived lnc genes expression values. Heatmap shows the frequency of two MB samples falling into the same cluster over 60 NMF iterations. **b)** Sankey plot visualises the agreement between RNA-seq-based NMF clustering (panel a)) and methylation-based classification.

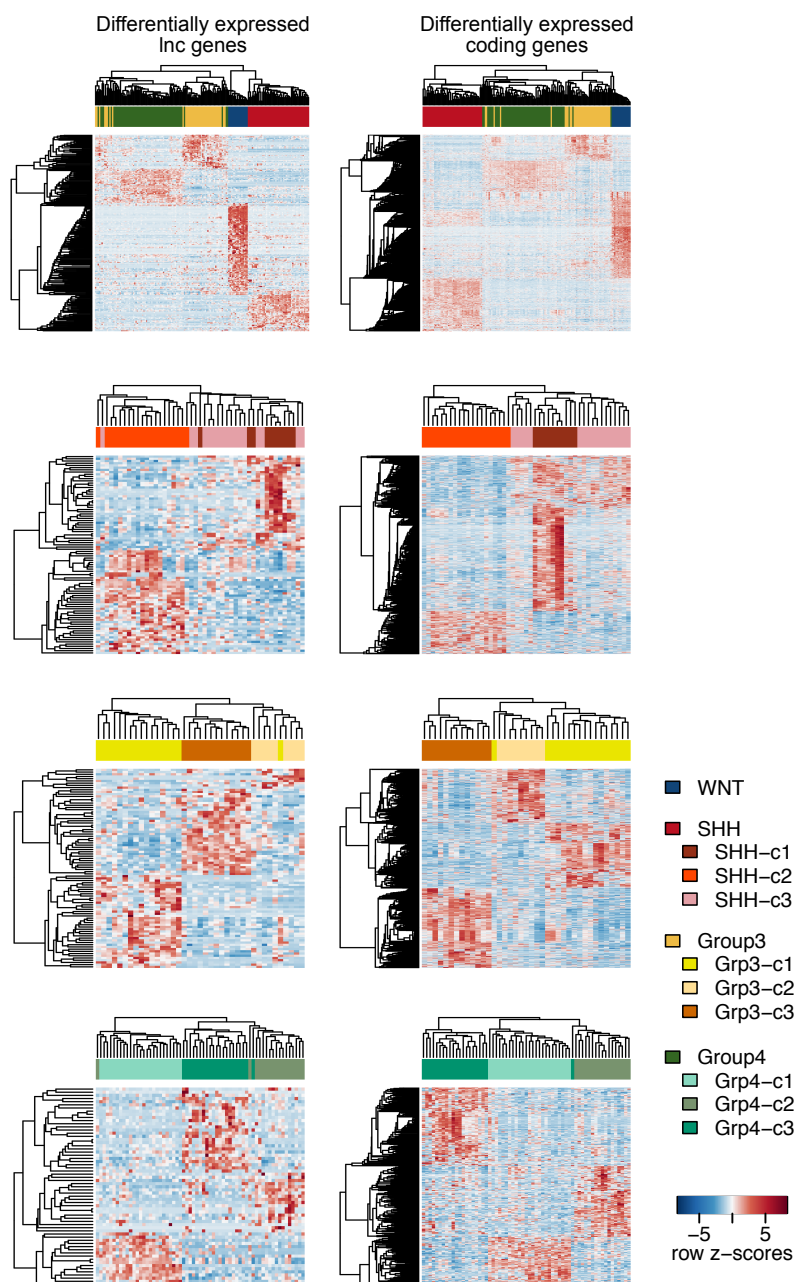


Figure 5.25: Comparison of hierarchical clustering based on specifically upregulated coding or lnc genes **Left column**) Heatmaps show lnc genes. **Right column**) Heatmaps show coding genes. The colours above the heatmaps indicate subgroup and subcluster assignments.

5.3.3.2 Characterisation of lnc genes: Divergent, antisense and intergenic

We integrated several resources to provide a systematic characterisation of the 448 differentially expressed lnc genes in MB: FANTOM CAT (FANTOM5)[133], BrainSpan (Allen Brain Atlas) [299], literature, and the here analysed MB RNA-seq data set. Using FANTOM CAT and Ensembl gene annotations, lnc genes were assigned to three different types indicating their position relative to coding and pseudogenes, as introduced in Section 3.6.1 (Method Section 5.4.6.2). We annotated 34%

5 Medulloblastoma study

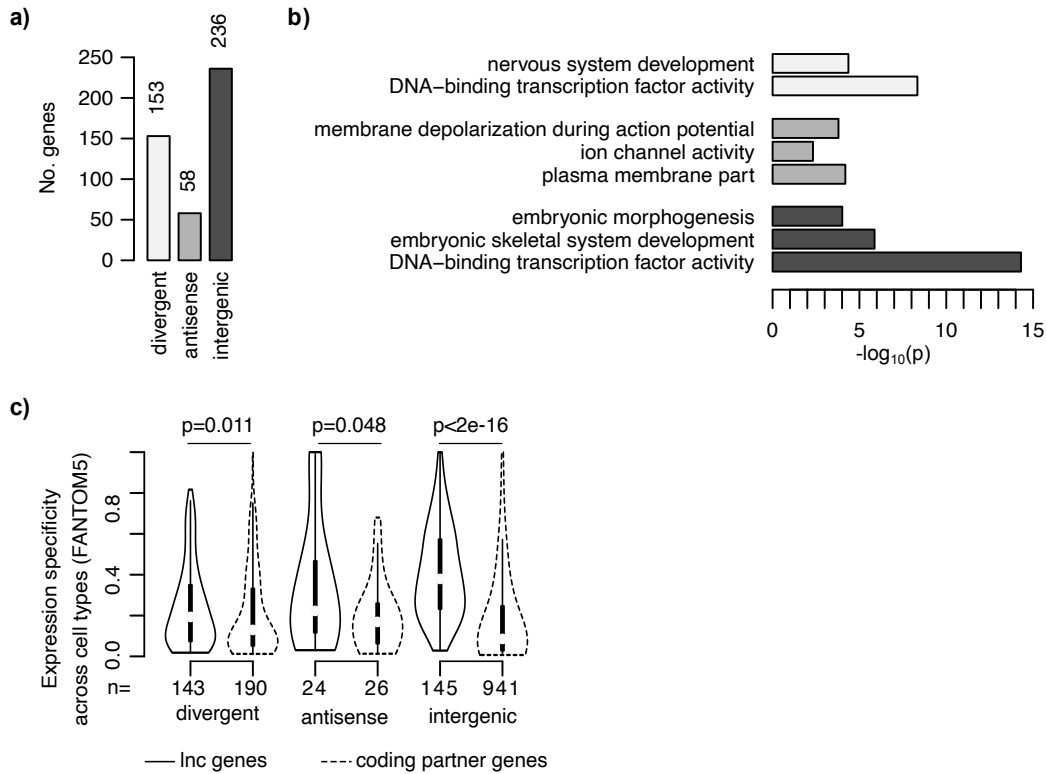


Figure 5.26: Lnc gene type annotation and coding neighbourhood. **a)** Bar plot shows the number of lnc genes that were annotated as divergent, antisense or intergenic. The number of genes is placed above each bar. **b)** Overrepresented gene ontology terms in 194, 55, 1397 expressed coding genes that were part of the coding neighbourhood of divergent, antisense and intergenic lnc genes, respectively (hypergeometric test). Colours of bars relate to lnc gene type as shown in panel a). **c)** Violin plots show the distribution of expression specificity scores across cell types for lnc and coding genes. Expression specificity score derived from FANTOM CAT (Method Section 5.4.3.4). P-values indicate that lnc genes of the three types show a significantly higher expression specificity score compared to their coding neighbourhood (one-sided Wilcoxon rank-sum test). The number of genes is placed below each violin plot.

(n=153), 13% (n=58) and 53% (n=236) of the differentially expressed lnc genes as divergent, antisense and intergenic, respectively (Figure 5.26.a). In comparison to the genome-wide-annotated lnc genes, divergent lnc genes were more frequent (MB: 34%; genome-wide: 22%), antisense lnc genes were less frequent (MB: 13%; genome-wide: 24%), and intergenic lnc genes showed a similar frequency (MB: 53%; genome-wide: 54%) among the detected differentially expressed lnc genes. Lnc and coding genes that were in divergent or antisense orientation to each other were considered as lnc-coding gene partners (Method Section 5.4.6.2). Four antisense lnc genes were in antisense orientation to two or three coding genes. For 41 divergent lnc genes, coding genes in antisense orientation were considered as a coding partner additional to the divergent coding partner. The majority of divergent lnc genes was partnered with a coding gene; five lnc genes were divergent to a pseudogene. For intergenic lnc genes, the definition of coding partners was less strict in terms of proximity. Coding and intergenic lnc genes were assigned as partners when they shared a common TAD, and the distance between the two genes was less than 500 kb (Method Section 5.4.6.2). Via this annotation, at least one coding gene partner could be assigned to 215/236 intergenic lnc genes. The remaining 21 intergenic lnc genes located within a TAD boundary or located in a TAD missing coding genes. We detected that

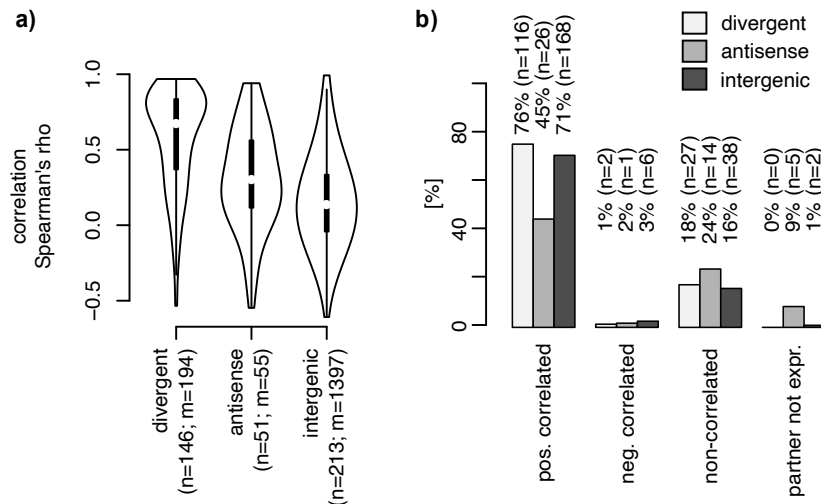


Figure 5.27: Correlation between pairs of lnc genes and coding genes partners. **a)** Violin plots show the distribution of the correlation coefficients between lnc-coding gene pairs for the three lnc gene types (Spearman correlation). 25% and 75% as well as 50% quantile are indicated by the lower and upper edge of the box and the point in the middle, respectively. The number of lnc and coding genes that define the genes pairs relate to n and m, respectively. **b)** Bar plots show the percentage of lnc genes that belong to one of the four correlation-based categories per lnc gene type. The percentage and number of lnc genes is placed above the bars.

194/209, 55/63, and 1397/1729 coding partners of divergent, antisense, and intergenic lnc genes were expressed in MB, respectively. The expressed coding partner neighbourhood of the three lnc gene types was characterised by performing a GO term overrepresentation analysis (5.26.b). The coding gene neighbourhood of intergenic and divergent lnc genes was enriched for DNA-binding transcription factor activity and for developmental processes, which is consistent with previous reports [134, 135]. However, in our analysis, the coding gene neighbourhood of these two lnc gene types enriched for different developmental processes. Due to the observed transcription factor activity enrichment, we checked how many lnc genes were partnered with TF. Here, 26/153 divergent and 73/236 intergenic lnc genes were partnered with a differentially expressed TF (e.g. *EBF1*, *NEUROD1*, *TWIST1*, *SOX9*). The coding gene neighbourhood of antisense lnc genes was enriched for cell membrane-associated processes (5.26.b).

To provide an additional evaluation of the potential biological relevance of the differentially expressed lnc genes in MB, we compared published expression specificity scores between lnc-coding partners per lnc gene type (Figure 5.26.c, Method Section 5.4.3.4). These expression specificity scores were based on CAGE data across 69 facet cell types and obtained from the FANTOM5 project [133]. In our comparison, all three lnc genes types showed a significantly higher expression specificity score compared to the coding partners. A general higher expression specificity for lnc genes compared to coding genes has been previously described [133]. Intergenic genes showed the highest expression specificity, which is in line with previous reports by Hon *et al.* [133].

To understand the transcriptional relationship between lnc-coding partners, the expression correlation between these partners was measured. This analysis included 145/153 divergent, 54/58 antisense and 173/236 intergenic lnc genes that were assigned to 150, 58, 476 expressed coding genes, respectively. We observed the general trend of positive correlation between lnc-coding partners and that divergent lnc coding genes showed the strongest positive correlation among the three lnc gene types (Figure 5.27.a). These results are in line with previous publications [135]. However, we also measured considerable negative correlation coefficients (< -0.3 , Method Section 5.4.1). To further subclassify the

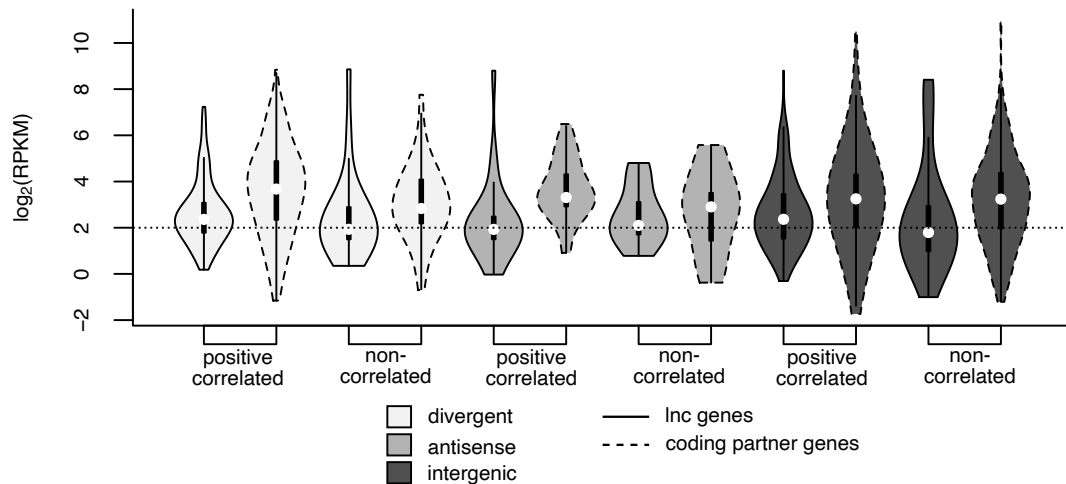


Figure 5.28: Expression levels of lnc genes and coding partners. Lnc gene and coding partners are split by lnc gene type (divergent, antisense, and intergenic) and correlation-based lnc categories. Violin plot silhouette shows expression density distribution. The 25%, 75%, and 50% quantile is indicated by the bottom and top edge of the black and the point in the middle, respectively.

three lnc gene types, the expression correlation analysis was utilised to assign differentially expressed lnc genes to four categories that indicate whether coding partners were positively correlated ($\rho \geq 0.3$), negatively correlated ($\rho \leq -0.3$), non-correlated ($-0.3 > \rho < 0.3$), or non-expressed (Figure 5.27.b). The category negative correlated and non-correlated implied that none of the coding partners was positively correlated. For the categories negative correlated, non-correlated and not expressed, we assured that the broader coding neighbourhood did not show positive expression correlation. Here, all coding genes within $\pm 100\text{kb}$ regardless of the strand orientation were additionally integrated to define these three categories.

We evaluated whether a positive correlation between lnc genes and coding partners is associated with similar expression levels between lnc and coding genes. As a basis for the comparison, we calculated the average expression per gene among the 15 samples with the highest expression levels because genes were mostly specifically upregulated in subsets of MB samples related to subgroups and subclusters. Taking the average of expression values across the whole cohort would not be an informative indicator of the actual expression level of a gene. Lnc genes of the most frequent categories positive correlated and non-correlated were compared to their coding partners per lnc gene type (Figure 5.28). Independent of the correlation category, lnc genes were lower expressed compared to coding partners for all lnc types, and lnc genes showed a similar expression level across types and categories. These data do not indicate associations between expression levels and positive correlation of lnc genes with their coding partners.

The minority of 95 lnc genes belonging to the categories negative correlated, non-correlated and non-expressed were of interest for further functional annotations. These lnc genes were transcribed independent of the coding partner and, therefore, are also potentially independent in their function or a potential negative regulator of the coding partner.

5.3.3.3 Impact of copy-number variations on lnc gene expression

The set of 95 lnc genes, which did not show a positive correlation with a coding partner or the coding gene neighbourhood, was overlaid with copy number data to identify lnc genes whose expression is influenced by copy number variations. Lnc genes that are influenced by CNVs could have relevant

functions. To identify lnc genes that show a significant drop of expression associated with a copy number loss or raise of expression associated with a copy number gain, expression values of tumour samples with 2N were compared against 1N or >2N for a given lnc gene, respectively.

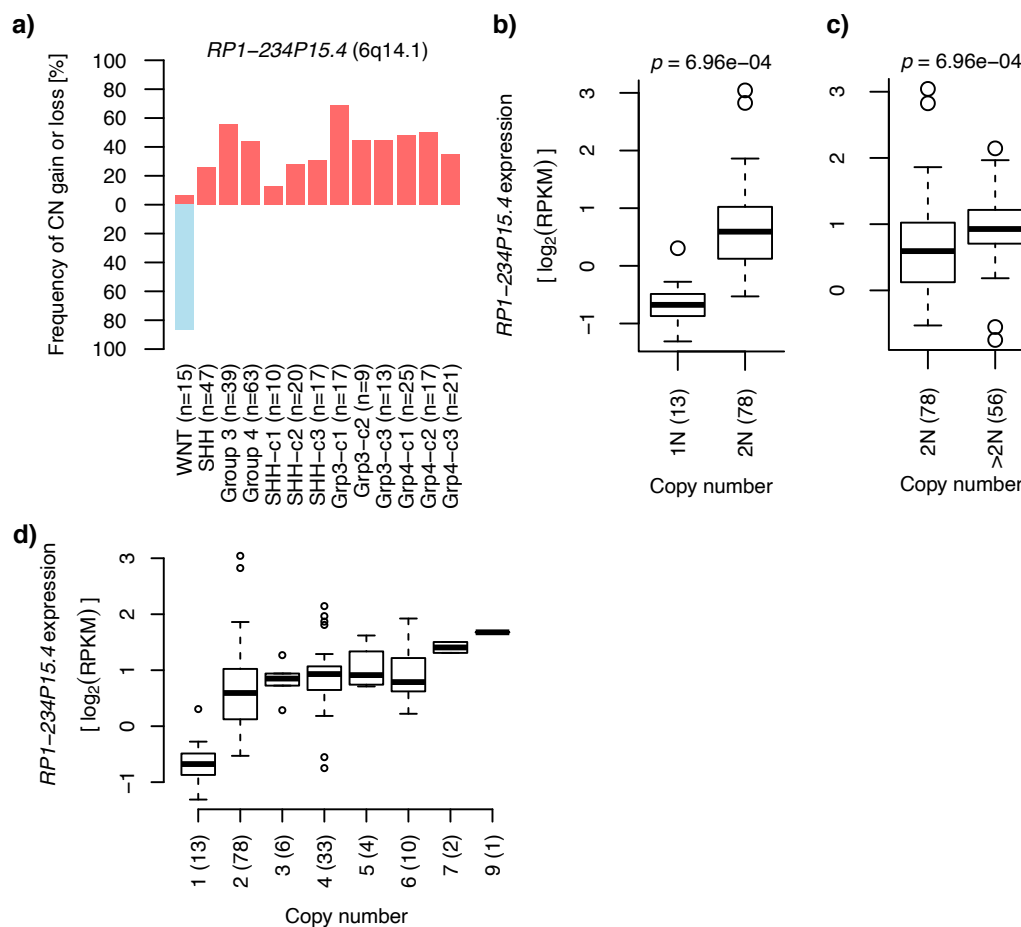


Figure 5.29: *RPI-234P15.4* expression is influenced by copy number variation. **a)** Bar plots show the percent of MB samples that showed a copy number loss or gain. **b-c)** Box plots compare the gene expression between samples with different copy numbers (1N, 2N, >2N). The numbers of samples are shown in brackets. Copy-number-based groups were tested for differential expression using the Wilcoxon rank-sum test. P-values are placed above the box plot. **d)** Box plots illustrate the expression of *RPI-234P15.4* over different copy numbers. The numbers of samples per box plot are shown in brackets.

Several lnc genes showed significant upregulation comparing MB samples with 2N and >2N copy numbers. However, for most lnc genes, overlaying gene expression values and discrete copy number levels did not support a direct CNV-expression association rather than spurious correlations due to subgroup-/subcluster-specific upregulation and copy number gain. We detected a convincing CNV-expression association only for the lnc gene *RPI-234P15.4* related to copy number loss (Figure 5.29.b). *RPI-234P15.4* is located on chromosome 6 and showed frequent copy number loss and downregulation in WNT MBs (Figure 5.30, 5.29.b). *RPI-234P15.4* followed CNV-expression patterns that were also observed for other coding genes located on chromosome 6, as described above (Section 5.3.2.7), and that related to frequent monoallelic deletion of chromosome 6 in WNT MBs. *RPI-234P15.4* also showed a significantly higher expression over copy number gain (Figure 5.29.d). However, the change of expression was only weak, and expression levels were comparable between copy number levels

3N-6N, indicating that copy number gain did not influence or only weakly influenced *RP1-234P15.4* expression (Figure 5.29.d-e). *RP1-234P15.4* expression in WNT MBs was similar to cerebellum controls but higher in the remaining subgroups compared to the control and WNT MBs (Figure 5.30).

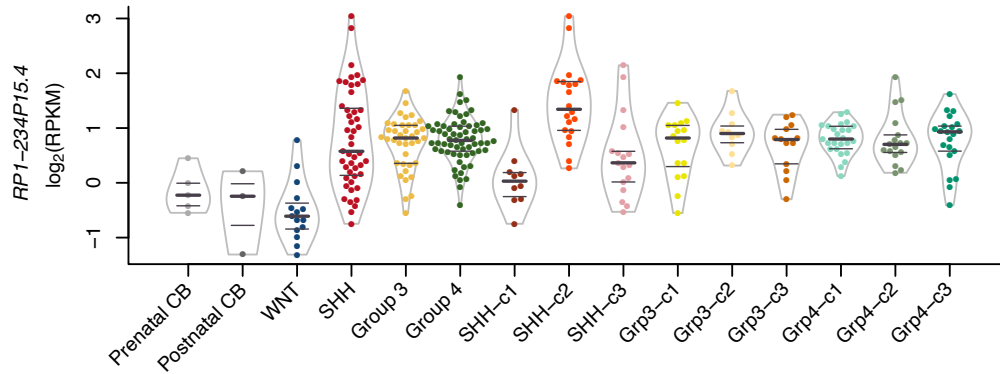


Figure 5.30: Expression profile of *RP1-234P15.4* in MB. Violin plots show expression density distribution. 25%, 50% and 75% quantiles are indicated by horizontal lines. Individual MB samples are shown as bee swarm plots.

5.3.3.4 Lnc genes with brain-development-related expression patterns

Since medulloblastoma is an embryonic tumour of cerebellar tissue, lnc genes with expression patterns related to the development of the cerebellum or brain could be of interest in the context of this disease. We utilised a gene expression data set of pre- and postnatal brain tissues from the BrainSpan project to identify lnc genes with the sought expression patterns. These patterns comprised a) up- or downregulation in the cerebellum in comparison to other brain tissues and b) differential expression between pre- and postnatal tissues of the cerebellum or whole brain. Additionally, FANTOM CAT data were integrated to annotate lnc genes that show enriched expression in embryonic or neural stem cells (as described in Method Section 5.4.3.4). Here, only lnc genes that were not positively correlated with the coding gene neighbourhood were considered for this analysis.

We detected 20 lnc genes that followed the mentioned expression patterns. Eight and two lnc genes were annotated for enriched expression in ESC and NSC, respectively (Figure 5.31). The lnc gene *GLYCTK-AS1* that was specifically upregulated in subgroup Group 4 and its subclusters showed upregulation in prenatal CB (data source: BrainSpan) and enriched expression in ESC (data source: FANTOM CAT; Figure 5.31). Due to the unstranded RNA-seq of the BrainSpan data set, *GLYCTK-AS1* and the coding antisense gene *GLYCTK* showed a strong correlation in expression ($\rho = 0.81$, $p = 1.39 \times 10^{-114}$, $n=487$). However, via our stranded RNA-seq samples of CB, we validated that *GLYCTK-AS1* was exclusively expressed in prenatal CB and that *GLYCTK-AS1* was independently expressed from *GLYCTK* (Figure 5.32.a-b). The integrated FANTOM5 CAGE data also supported the observation that *GLYCTK-AS1* and *GLYCTK* were independently expressed (Figure 5.33.a). Besides ESCs, *GLYCTK-AS1* was annotated for enriched expression in several cell/tissue ontologies in FANTOM CAT: 14 neuronal/brain-associated ontologies, sexual organs, and hindgut. Since *GLYCTK-AS1* showed development-associated expression patterns, we evaluated a potential expression correlation with the neural stem/progenitor cell marker *HES5* [300]. This analysis revealed a strong co-expression across prenatal human brain tissues (BrainSpan) ($\rho=0.67$, $p = 1.21 \times 10^{-31}$, $n=235$, Figure 5.33.b). We repeated this analysis using CAGE data (FANTOM5) showing that *GLYCTK-AS1* and *HES5* were either mutually exclusively expressed or strongly co-expressed in different subsets of samples (Figure 5.33.c). Here, in samples that express both genes, we could validate the strong expression correlation ($\rho = 0.66$, $p = 1.13 \times 10^{-28}$; $n=215$;

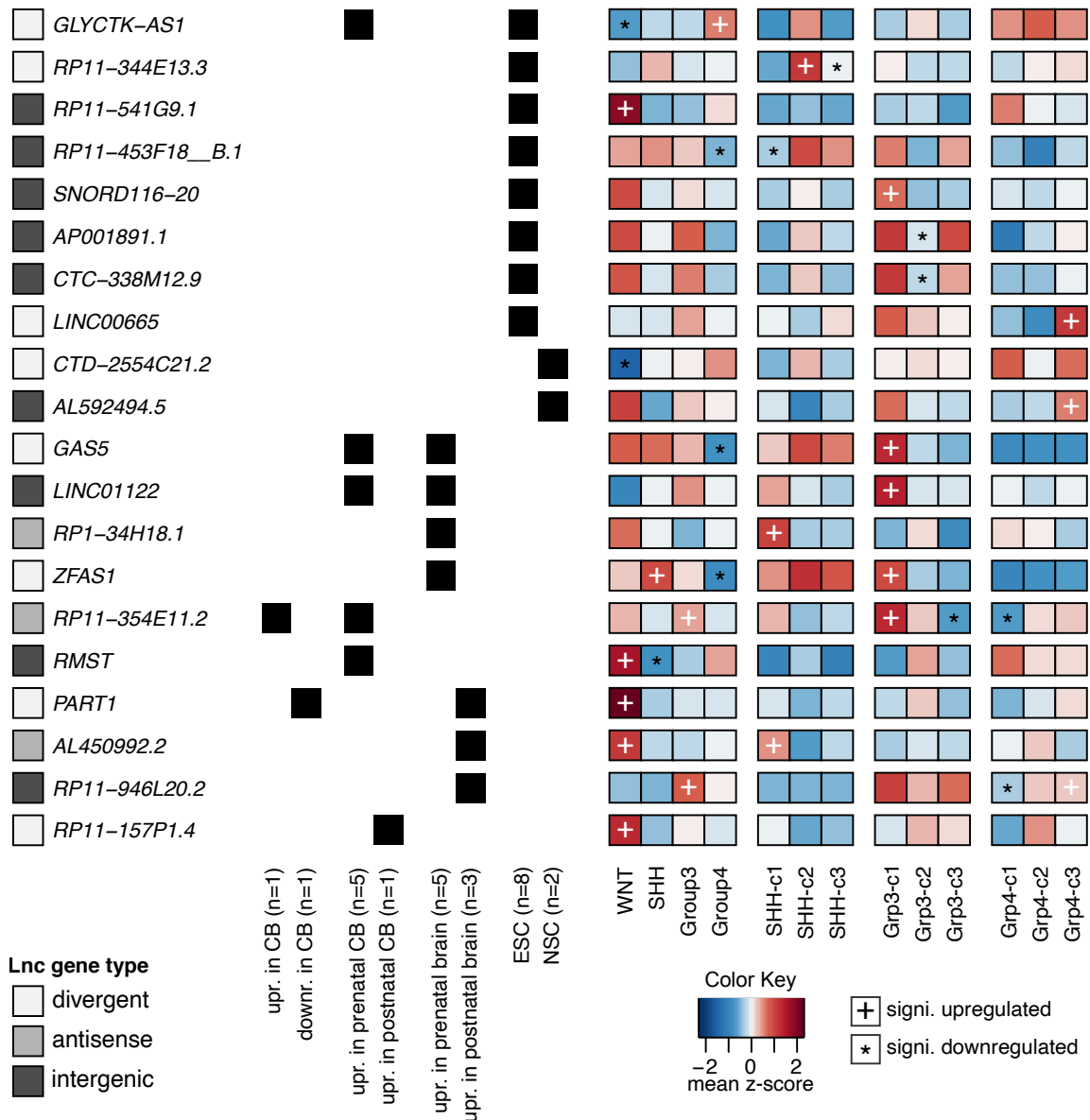


Figure 5.31: Annotation of stem-cell-related as well as prenatal- and postnatal-brain-related expression of 20 lnc genes. The gene *RP11-453F18_B.1* is also known as *FIRRE*. Genes are ordered by similar annotations. **Left)** Lnc gene symbols and their lnc gene type. **Middle)** Cerebellum-, brain-, prenatal-, postnatal-, and stem-cell-related expression patterns are split into eight categories and annotated per lnc gene, as indicated by black rectangles. Cerebellum- and brain-related expression patterns were identified using BrainSpan expression data (Method Section 5.4.6.4). The annotation for enriched expression in ESC and NSC derived from FANTOM CAT (Method Section 5.4.6.4). The number n of lnc genes that were annotated for one of the eight categories is indicated in brackets below. **Right)** The heatmap summarises the expression patterns of the lnc genes in MB subgroups and subclusters. Significant up- or downregulation is indicated. Abbreviations: ESC - Embryonic stem cells; NSC - neural stem cells; CB - cerebellum; upr. - upregulated; downr. - downregulated; signi. - significant.

Figure 5.33.d). In our MB cohort, *GLYCTK-AS1* and *HES5* showed significant but weak correlation in non-SHH MBs ($\rho = 0.31$, $p = 5.76e-4$, $n=117$; Figure 5.33.e). Among the non-Group 4 subclusters, only the subcluster SHH-c2, which represents adult SHH MBs, showed an expression of *GLYCTK-AS1* comparable to subgroup Group 4 MB. The remaining non-Group 4 subclusters showed a wide range of expression, and *GLYCTK-AS1* was highly expressed only in a fraction of samples (Figure 5.32.a). Taken together, these data suggest that *GLYCTK-AS1* is associated with developmental processes of neurons/cerebellum and could be expressed in neural progenitors.

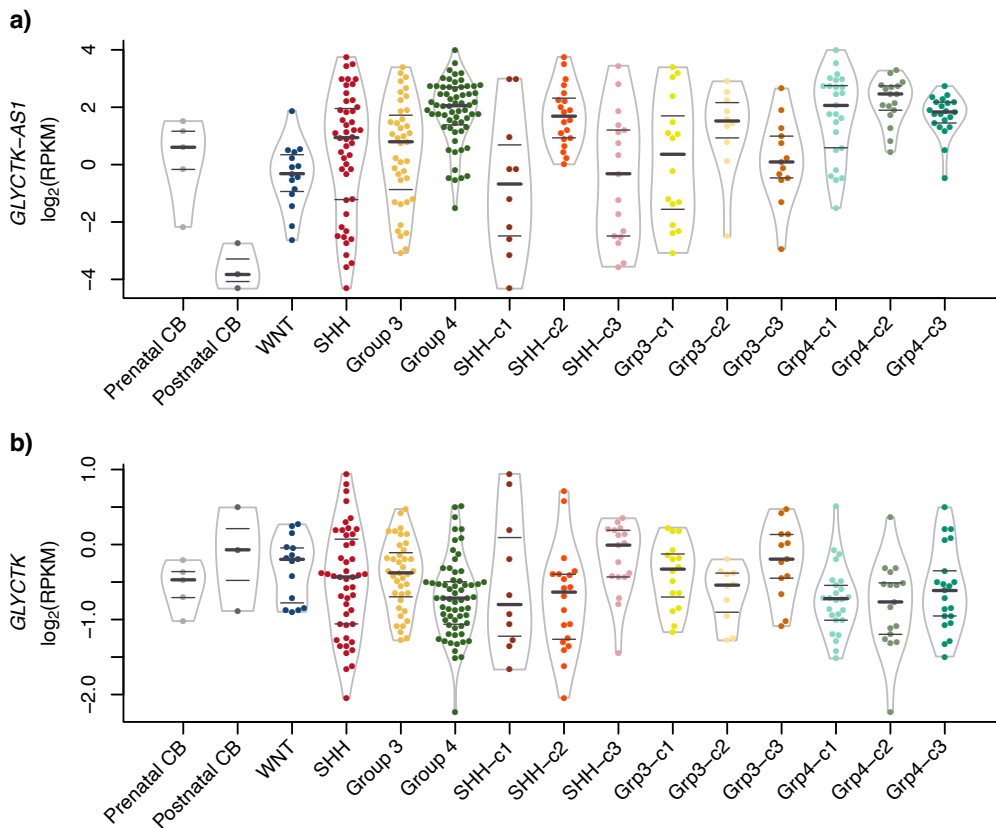


Figure 5.32: Expression profiles of *GLYCTK-AS1* and *GLYCTK* in MB and CB controls. **a)** *GLYCTK-AS1*. **b)** *GLYCTK*. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual MB samples are shown as bee swarm plots.

Three out of four lnc genes (*PART1*, *AL450992.2*, *RP11-946L20.2* and *RP11-157P1.4*) that showed upregulation in postnatal CB or whole-brain were strongly expressed in the WNT subgroup (Figure 5.31). This pattern could be explained by the reported lasting activation of Wnt signalling in the postnatal brain regulating synaptic functions [301].

Four (*GAS5*, *LINC01122*, *ZFAS1* and *RP11-354E11.2*) out of seven lnc genes that were higher expressed in prenatal CB or prenatal whole-brain showed significant upregulation in the *MYC* amplification-associated subcluster Grp3-c1 (Figure 5.31). The observed expression pattern of these four lnc genes is potentially similar to the expression pattern of *MYC* that shows higher expression in the developing brain and cerebellum, as previously reported [302, 303]. Among these four genes was *RP11-354E11.2*, which showed as the only gene seen in Figure 5.31 upregulation in CB in comparison to remaining brain regions.

Additionally to the BrainSpan data, we checked the expression of these 20 lnc genes in the cerebellum

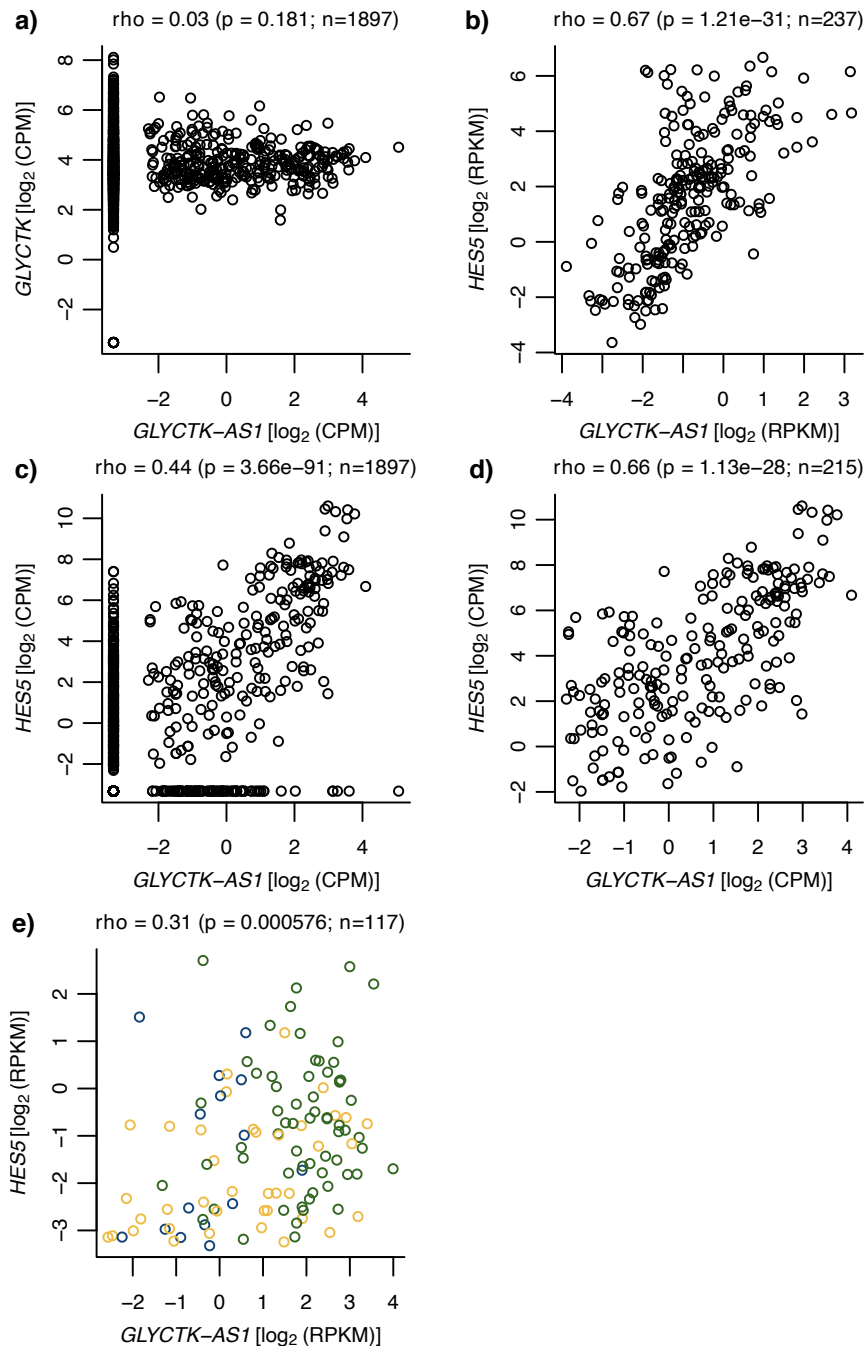


Figure 5.33: **a)** Scatter plot of *GLYCTK-AS1* and *GLYCTK* expression in FANTOM5 CAGE samples. **b-e)** Scatter plots of *GLYCTK-AS1* and *HES5* gene expression in **b)** prenatal human brain tissues from BrainSpan, **c)** all FANTOM5 CAGE samples, **d)** FANTOM5 CAGE samples that express both genes, **e)** ICGC non-SHH MB RNA-seq samples. Colours indicate MB subgroup: WNT=blue, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and the number of samples n are displayed above each plot.

controls of our analysed stranded RNA-seq cohort. Here, 19/20 lnc genes showed clear expression in stranded RNA-seq samples of pre- or postnatal CB or both. *AP001891.1* was only detected in BrainSpan data in prenatal and infant CB at a low expression level.

5.3.3.5 Lnc genes described in the literature and their context in medulloblastoma

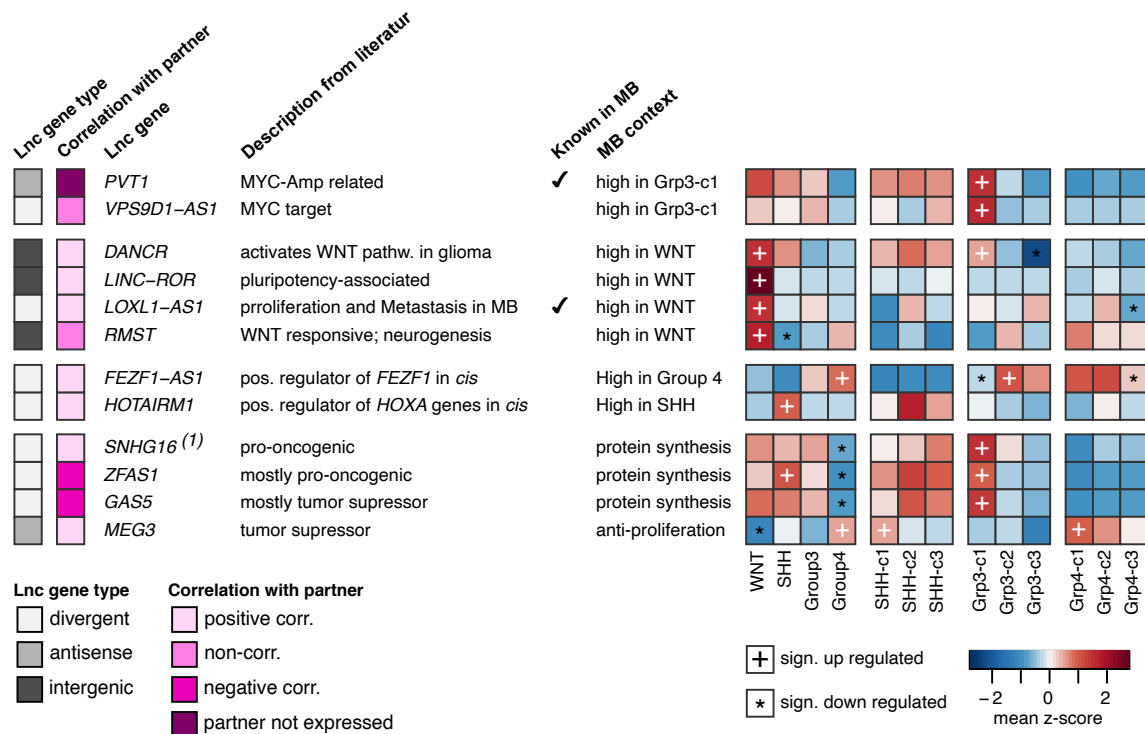


Figure 5.34: Annotation of literature knowledge. Related literature is mentioned in the main text. Right heatmap shows average expression in subgroups and subclusters. Significant up- and downregulation is indicated. ⁽¹⁾ SNHG16 has divergent and antisense coding partners and shows a weak positive correlation with one antisense partner.

The so-far shown characterisation of the detected differential expressed lnc genes was driven by data integration and external annotations. In the following part, we will focus on lnc genes that have been previously described in the literature and the reassessment of their reported functions in the context of MB.

In order to obtain an overview of how many of the 448 differentially expressed lnc genes have been mentioned in publications, Ensembl gene identifiers were matched to the PubMed database (Methods Section 5.4.6.5; database download February 20 2019). Ensembl (v70) identifiers of 194 lnc genes could be mapped to Entrez gene identifiers. Here, 174 lnc genes mapped to a single and 20 lnc genes corresponded to at least two Entrez identifiers. Among the 194 mapped lnc genes, 160 genes have been mentioned in at least one publication (range 1-189 publications). The majority of lnc genes has been mentioned in a few, up to three, publications (n=91, 57%). Among 48 lnc genes that were mentioned in at least seven publications, twelve genes were selected for further investigations in the context of MB using the RNA-seq cohort (Figure 5.34). These twelve lnc genes were selected because of their reported functions or implications in cancer.

The lnc genes *PVT1* and *VPS9D1-AS1* have been described to be associated with *MYC*. As summarised above (Section 5.1.5, 5.1.3), *PVT1* is known to be co-amplified with the upstream-located *MYC* in MB and upregulated in those cases, which is in line with the upregulation in Grp3-c1. However, in our data, *PVT1* was upregulated in both WNT and SHH vs Group 4 MBs, indicating a role of *PVT1* that is independent of Grp3-c1 tumours. Kawasaki *et al.* have shown that *VPS9D1-AS1* (alias *MYU*) is a direct target of *MYC* and stabilises the expression of the cell-cycle-promoting gene *CDK6* in cooperation with the RNA binding protein hnRNP-K (gene symbol *HNRNP-K*) in carcinoma [304]. However, we could

not find any evidence for an expression correlation between *VPS9D1-AS1* and *CDK6* in the presence of a ubiquitous expression of *HNRNPK* in MB (average subgroup expression between 39 and 50 RPKM) (Figure A.35.c). In the MB RNA-seq cohort, *MYC* was generally higher expressed in Group 3 tumours with additional elevated expression in subcluster Grp3-c1, but *VPS9D1-AS1* was upregulated only in Grp3-c1. Probably additional mechanisms prevent the expression of *VPS9D1-AS1* in Grp3-c2 and Grp3-c3 MBs (Figure A.34 and A.35.a-b).

Four lnc genes (*LOXLI-AS1*, *DANCR*, *LINC-ROR*, and *RMST*) that we selected based on literature were upregulated in WNT MBs (Figure 5.34). Among these four lnc genes was *LOXLI-AS1*. Gao *et al.* have reported *LOXLI-AS1* overexpression in MB tumours without evaluating subtypes or subclusters. The authors showed that *LOXLI-AS1* promotes cell proliferation and the formation of metastasis in MB via activation of the PI3K-AKT pathway [305]. Here, the authors knocked down *LOXLI-AS1* in MB cell lines D283 (Group 3/4-classified) and D341 (Group 3-classified) [305, 306]. Besides the upregulation that we observed in WNT MBs, *LOXLI-AS1* was consistently higher expressed in subcluster SHH-c2 and -c3, Grp3-c1 and -c3, and Grp4-c2 compared to normal cerebellum (Figure A.36).

The lnc gene *DANCR* was described as a bad prognosis factor in different cancer types [307–310]. *DANCR* is a direct target of *MYC* and a positive regulator of the Wnt/ β -catenin signalling in glioma [309, 310]. Also, *MYC* is known to be a target of Wnt/ β -catenin signalling [311]. The regulation of *DANCR* via Wnt-signalling and *MYC* might explain the high expression *DANCR* that we found in WNT and Grp3-c1 MB (Figure A.37 and 5.34). However, we observed a significant expression correlation between *DANCR* and *MYC* only in subcluster Grp3-c1 tumours. Additionally, expression patterns across subgroups indicated regulation of *DANCR* that might be independent of *MYC* (Figure A.37.a-b). Overall, our data suggested *MYC* as a potential regulator of *DANCR* in a subset of Group 3 MB excluding Grp3-c3 MB.

The lnc gene *LINC-ROR* functions as a microRNA sponge and, therefore, belongs to ceRNAs [312]. *LINC-ROR* is involved in the regulation of different pathways and processes including the TP53 pathway, EMT, pluripotency and hypoxia, and has been described to act pro-oncogenic in several cancer types. Only in glioma *LINC-ROR* has been described as a tumour suppressor [312]. In our MB cohort, *LINC-ROR* was almost exclusively expressed in WNT MBs and absent in pre- and postnatal cerebellum (Figure A.39 and 5.34). A link between upregulation of *LINC-ROR* and Wnt signalling is supported by reports in ovarian cancer [313]. However, since we did not see *LINC-ROR* expression in the normal cerebellum, the exclusive expression of *LINC-ROR* in WNT MB could be also linked to the cell-of-origin of this MB subgroup that is assumed to be outside of the cerebellum in the lower rhombic lip and developing brainstem, whereas the remaining MB subgroups most likely arise from cell types of the developing cerebellum [279, 314].

The lnc gene *RMST* was described as a regulator of neurogenesis and responsive to WNT signalling in developing forebrain, explaining the upregulation in WNT MBs [315, 316]. In our data, *RMST* was highly expressed in WNT MBs. The remaining subtypes showed a high variation in expression (Figure A.40).

The two divergent lnc genes *FEZF1-AS1* and *HOTAIRM1* did not show a common expression pattern among MB subgroups and subclusters, but both genes are described as positive regulators in *cis* of the nearby coding gene partner(s) (Figure 5.34). *HOTAIRM1* is part of the *HOXA* gene cluster and is located in a divergent and antisense position to *HOXA1* and *HOXA2*, respectively. In the analysed RNA-seq cohort, *HOTAIRM1* and several other *HOXA* genes were upregulated in SHH MBs (Figure 5.34, 5.11). Adult SHH subcluster SHH-c2 showed the highest expression of *HOTAIRM1*, which is in line with a previously reported upregulation of *HOXA* genes in adult SHH MB [214]. Li *et al.* have reported that *HOTAIRM1* positively regulates *HOXA1* in glioblastoma via inhibiting repressive histone marks and CpG modification in the *HOXA1* promoter by binding to DNMT1, DNMT3A, DNMT3A and EZH2. Here, *HOTAIRM1*-facilitated *HOXA1* upregulation promotes cell proliferation [317]. Wang *et al.* [318] reported several mechanisms of how *HOTAIRM1* regulates the expression of *HOXA* genes. In

the first mechanism, *HOTAIRM1* is involved in the dissociation of a sub-TAD that probably facilitates a silent state of *HOXA1-HOXA7*. Additionally, *HOTAIRM1* is a positive regulator of *HOXA1/HOXA2* and a repressive regulator for *HOXA4/HOXA5/HOXA6* via modulation of histone modification in the promoter regions. As emphasised by literature, we observed that *HOTAIRM1* showed the strongest correlation with *HOXA1* and *HOXA2* in comparison to the remaining HOXA genes in MB (Figure A.41). *HOTAIRM1* also showed a stronger correlation with *HOXA3*, but Wang *et al.* reported that *HOTAIRM1* has no impact on *HOXA3* expression [318]. Our data suggest that the positive regulation of *HOXA1* and *HOXA2* by *HOTAIRM1* is also valid in MB.

Chen *et al.* have reported *FEZF1-AS1* as a pro-oncogenic factor that is a positive regulator of the coding partner *FEZF1* in carcinoma without describing potential regulatory mechanisms [319]. The authors also reported that the knockdown of *FEZF1-AS1* or *FEZF* inhibits CRC cell proliferation and migration. Therefore, the authors speculated that the pro-oncogenic function of *FEZF1-AS1* is partially or fully carried out through *FEZF1* [319]. In our RNA-seq cohort, the lnc gene *FEZF1-AS1* was upregulated in Group 4, Grp3-c2 and Grp3-c3 MBs (Figure 5.34). *FEZF1-AS1* and the coding partner *FEZF* were strongly correlated ($\rho = 0.94$, $p = 1.39e-79$, $n=164$; Figure A.42). These results suggest that *FEZF1-AS1* regulates *FEZF1* in *cis* also in MB. Liu and colleagues published that *FEZF1-AS1* negatively controls the expression of *CDKN1A* (P21), a regulator of the cell cycle. However, we could not detect a relevant expression correlation between *CDKN1A* and *FEZF1-AS1* in MB (Figure A.43).

The following four lnc genes, *SNHG16*, *ZFAS1*, *GAS5* and *MEG3*, have been frequently described in the context of cancer (Figure 5.34). In the analysed MB cohort, *SNHG16*, *ZFAS1*, and *GAS5* were upregulated in WNT, SHH and Grp3-c1 tumours, whereas *MEG3* displayed an almost reverse expression profile associated with upregulation in Group 4 tumours.

Current literature suggests *SNHG16* as an oncogene in different cancer types [320, 321]. In bladder cancer, *SNHG16* negatively regulates *CDKN1A* (P21) via binding to PRC2 subunit EZH2 and PRC2-mediated epigenetically silencing of *CDKN1A*. In the analysed RNA-seq MB cohort, *SNHG16* and *CDKN1A* were not significantly correlated ($\rho = -0.12$, $p = 1.36e-01$; $n=164$). In non-small cell lung cancer, *SNHG16* acts as ceRNA targeting miR-146a, a microRNA that inhibits proliferation in a lung cancer cell line (A549) [321]. Moreover, *SNHG16* is a transcriptional target of *MYC*, explaining the high expression in WNT and Grp3-c1 tumours [322, 323].

We observed upregulation of *ZFAS1* in most SHH and Grp3-c1 tumours in comparison to prenatal cerebellum (Figure A.44). *ZFAS1* is well described in the context of cancer; oncogenic functions involve the EMT, NOTCH and p53 pathway [324]. *ZFAS1* is a ceRNA targeting miR-150 to increase the expression of *ZEB1*, a regulator of EMT [324]. It is also described that *ZEB1* is a target of Shh-signalling and upregulated in SHH MBs (Figure A.46) [325]. In the analysed MB cohort, *ZFAS1* and *ZEB1* showed a moderate positive correlation ($\rho = 0.46$, $p = 1.01e-09$, $n=164$; Figure A.45). Our data emphasised that *ZFAS1* might regulate *ZEB1* expression in addition to Shh-signalling in MB. In gastric cancer, it is reported that *ZFAS1* interacts with EZH2 and LSD1 (alias KDM1A) to downregulate the tumour suppressors *NKD2* and *KLF2* epigenetically [326]. In the analysed MB samples, *ZFAS1* and *NKD2* were moderately negatively correlated ($\rho = -0.43$, $p = 1.33e-08$, $n=164$; Figure A.47), but the negative correlation was more pronounced among non-WNT MBs ($\rho = -0.57$, $p < 2e-16$, $n=149$; Figure A.48). The upregulation of *NKD2* in WNT MBs has been reported before (Figure A.49) [218]. *NKD2* is induced via Wnt signalling and acts as an antagonist of this pathway [327]. Potential tumour suppressive roles of *NKD2* in MB have not been studied. Since *NKD2* antagonises WNT signalling via interactions upstream of CTNNB1 and the Wnt pathway is activated via mutation in CTNNB1 in most WNT MB cases, the antagonistic functions of *NKD2* could be eluded in WNT MBs [328]. However, *NKD2* might play tumour suppressive roles in the remaining subgroups. Here, our shown data suggest that *ZFAS1* could be a negative regulator of *NKD2* in non-WNT MBs.

We observed upregulation of *GAS5* in most WNT, SHH, and Grp3-c1 tumours in comparison to prenatal cerebellum (Figure A.44). The lnc gene *GAS5* is mostly described as a tumour suppressor,

but some studies also reported oncogenic functions [329–332]. *GAS5* features several functions in the context of tumour suppression. For example, *GAS5* can act as ceRNA targeting miR-21, a negative regulator of the tumour suppressor *PTEN* and *PDCD4* [329]. In the analysed MB cohort, *GAS5* showed a significant weak negative ($\rho = -0.33$, $p = 1.52e-05$, $n=164$) and moderate positive ($\rho = 0.45$, $p = 2.73e-09$, $n=164$) correlation with *PTEN* and *PDCD4*, respectively, indicating that *GAS5* expression might have a relevant influence on *PDCD4* expression mediated by miR-21 in MB but no influence on *PTEN* expression (Figure A.51, A.50). Previous reports also show that *GAS5* interacts with and stabilises YBX1 protein, a transactivator of *CDKN1A* (P21) [333]. However, we could find a relevant positive correlation between *GAS5* and *CDKN1A* while *YBX1* was ubiquitously expressed in MB ($\rho = 0.21$, $p = 6.38e-03$, $n=164$; Figure A.52 and A.53).

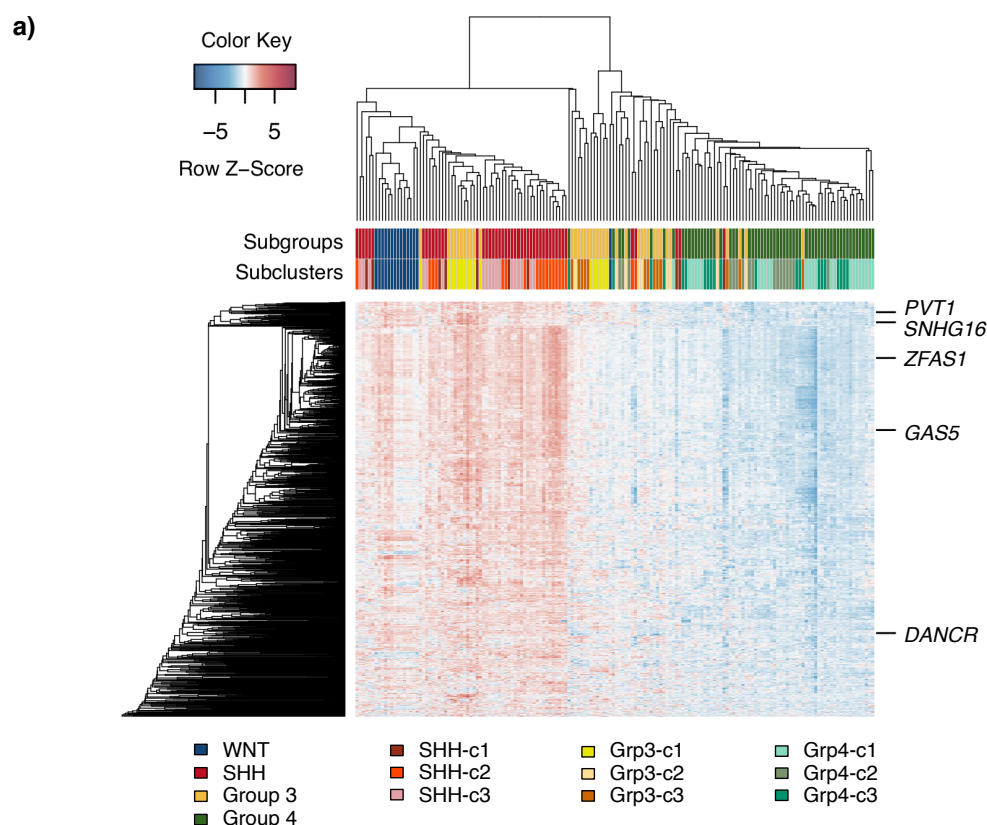


Figure 5.35: Co-expression cluster containing *ZFAS1*, *GAS5*, *DANCR*, *PVT1* and *SNHG16*. **a)** The heatmap depicts the co-expression cluster comprising 609 coding genes, 50 lnc genes, and 573 pseudo-genes across 164 MB samples. Colours above the heatmap indicate the subgroup and subcluster of a sample. **b)** Selection of processes that were enriched in the genes of the co-expression cluster (hypergeometric test, Method Section 5.4.4.4).

Among the twelve just-described lnc genes, the five lnc genes *PVT1*, *SNHG16*, *GAS5*, *ZFAS1*, and *DANCR* are probably the most studied ones in the context of cancer. In our MB cohort, these five lnc genes showed a similar gene expression pattern defined by upregulation in WNT, SHH, and Grp3-c1 MBs and were part of a common co-expression cluster (hereinafter called CLICK cluster

5 Medulloblastoma study

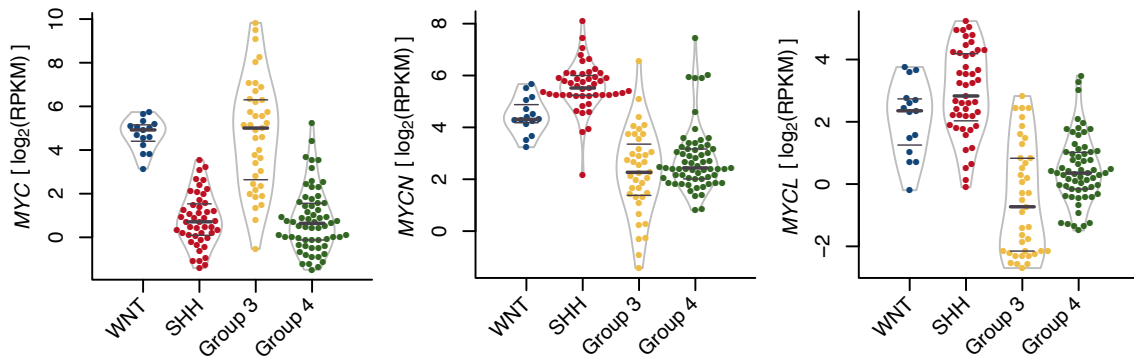


Figure 5.36: Expression profile of *MYC*, *MYCN*, and *MYCL* in MB subgroups. Violin plots show expression distribution. 25%, 50% and 75% quantiles are indicated by horizontal lines. Individual MB samples are shown as bee swarm plots.

1; acronym: Cc1), identified via the algorithm CLICK (Figure 5.34, 5.35.a; Method Section 5.4.6.6). *GAS5* and *ZFAS1* showed the strongest pairwise correlation among these five lnc genes ($\rho=0.87$, $p < 2e-16$, $n=164$; Figure A.54). Genes in Cc1 were functionally enriched for translation, metabolism of nucleotides, and transcriptional c-MYC targets; translation showed the strongest enrichment (Figure 5.35.b). The hallmark gene sets from MSigDB supported an enrichment for transcriptional c-MYC targets in Cc1 additionally (HALLMARK MYC TARGETS V1: $FDR=1.43e-16$; HALLMARK MYC TARGETS V2: $FDR=4.29e-16$) [334]. Kool *et al.* described a functional enrichment for protein synthesis in WNT-, SHH- and Group 3-matching MB subgroups, which agrees with the expression pattern of the co-expression cluster Cc1.

The enrichment for c-MYC targets in Cc1 provided a strong link between Cc1 and MYC-dependent gene regulation. As mentioned above, the lnc genes *SNHG16*, *PVT1*, and *DANCR* are MYC targets. Additionally, MYC is a well-known regulator of translational processes [335] that were enriched in Cc1. Surprisingly, *MYC* itself was not part of Cc1. However, it has been shown that members of the MYC gene family share a high fraction of transcriptional targets [336]. Therefore, we assumed that not only *MYC* but other family members could be additionally involved in the regulation of MYC targets in Cc1. All three family members, *MYC* (alias *c-MYC*), *MYCN* and *MYCL*, have been described to be involved in the formation of MB [269]. In our MB cohort, *MYC* was upregulated in WNT and Group 3 MBs and showed a different expression pattern than *MYCN* and *MYCL* that were upregulated in WNT and SHH MBs (Figure 5.36), which is in line with previous reports [269]. Overlaying gene expression levels of MYC family members with the mean-pattern of Cc1 indicate that the expression of all three TFs commonly contributed to the MYC target regulation in Cc1 (mean-pattern relate to the average relative expression of Cc1 in tumour samples; see Methods Section 5.4.6.6). Subgroup-specific upregulation of individual MYC family genes was correlated with upregulation of Cc1 in the respective subgroup. However, none of the three MYC family genes alone showed a strong correlation with the Cc1 mean-pattern across the whole cohort (Figure 5.37.a,d,g). Here, we observed a strong correlation with the Cc1 mean-pattern for *MYC* across non-SHH MBs and for *MYCN* and *MYCL* across non-Group 3 MBs, where *MYCN* showed a stronger correlation across non-Group 3 MBs compared to *MYCL* (Figure 5.37.b,f,i). Interestingly, the summed scaled expression of the three members of the MYC gene family displayed the strongest correlation with the Cc1 mean-pattern across the whole MB cohort (Figure 5.38.c). Here, two- and three-genes combinations of MYC family genes showed different strength of correlation with Cc1 mean-pattern in the following order: $MYC+MYCN+MYCL > MYC+MYCN > MYC+MYCL > MYCN+MYCL$ (Figure 5.38.a-d). The summed *MYCN* and *MYCL* expression showed a considerable high correlation with Cc1 mean-pattern across the whole cohort, but in fact, this correlation was limited to non-Group 3 MBs 5.38.d-e). Additionally, *MYCN* and *MYCL* alone showed a weaker correlation with Cc1 mean-pattern in non-Group 3 MBs compared to the

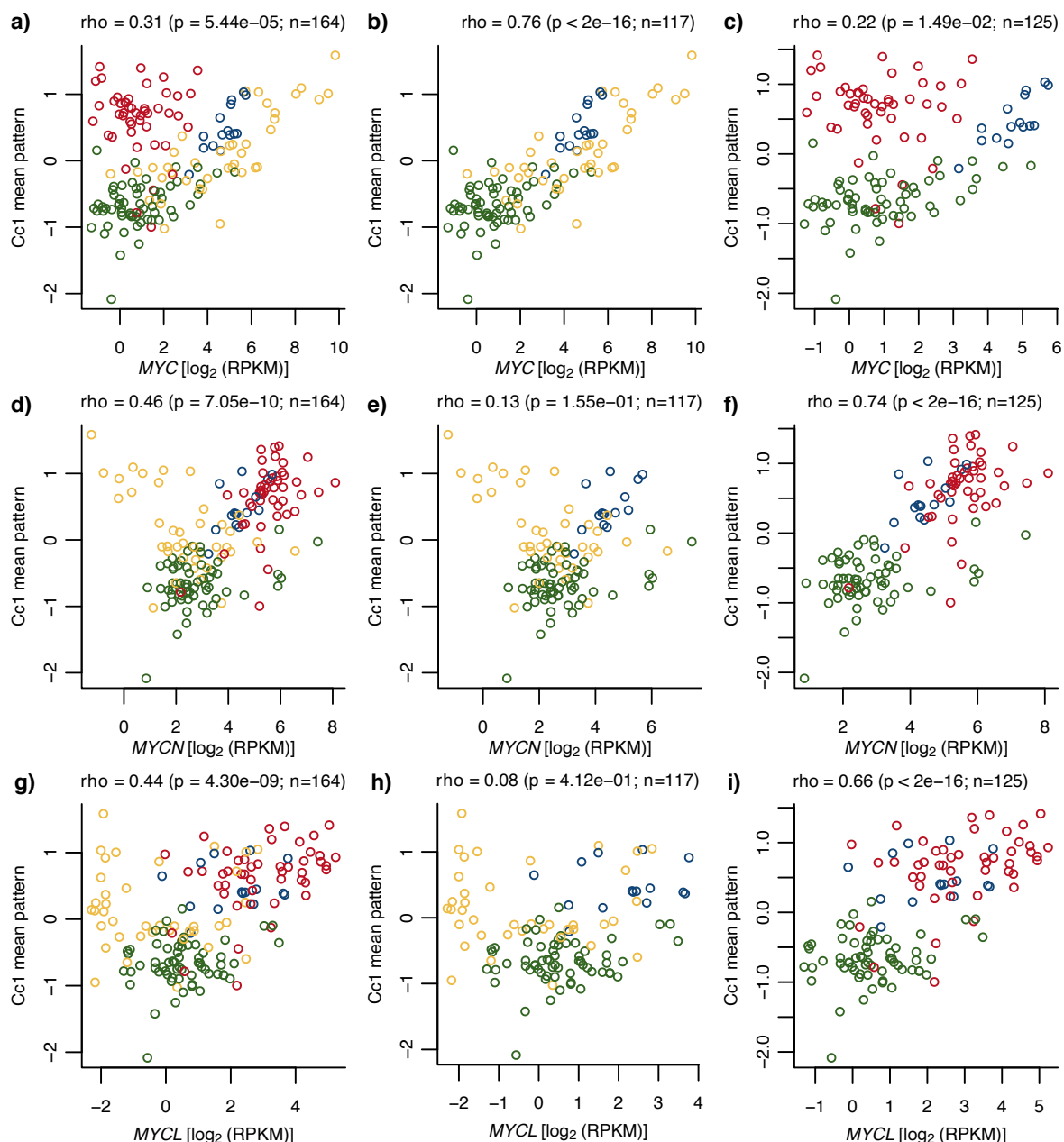


Figure 5.37: Comparison of Cc1 mean-pattern to *MYC*, *MYCN*, and *MYCL* expression in PedBrain MB samples. **a-c**) *MYC* expression vs. Cc1 mean-pattern. **d-f**) *MYCN* expression vs. Cc1 mean-pattern. **g-i**) *MYCL* expression vs. Cc1 mean-pattern. **a,d,g**) whole MB cohort. **b,e,h**) non-SHH MB samples. **c,f,i**) non-Group 3 MB samples. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and the number of samples *n* are displayed above each plot.

summed expression of both genes (Figure 5.37.f,i and 5.38.e).

Additionally to the common regulation of Cc1 by MYC family genes, Cc1 appeared to be linked to certain protein-signalling in MB. As summarised above in Section 5.1.3, Zomeran *et al.* reported a protein-signalling cluster (cluster-1) associated with an MYC-like kinase activity profile and protein synthesis that was active in SHH and a fraction of Group 3 MBs [220]; this paper did not include WNT MBs. The authors also provided an expression signature of 116 genes that were upregulated in MBs

5 Medulloblastoma study

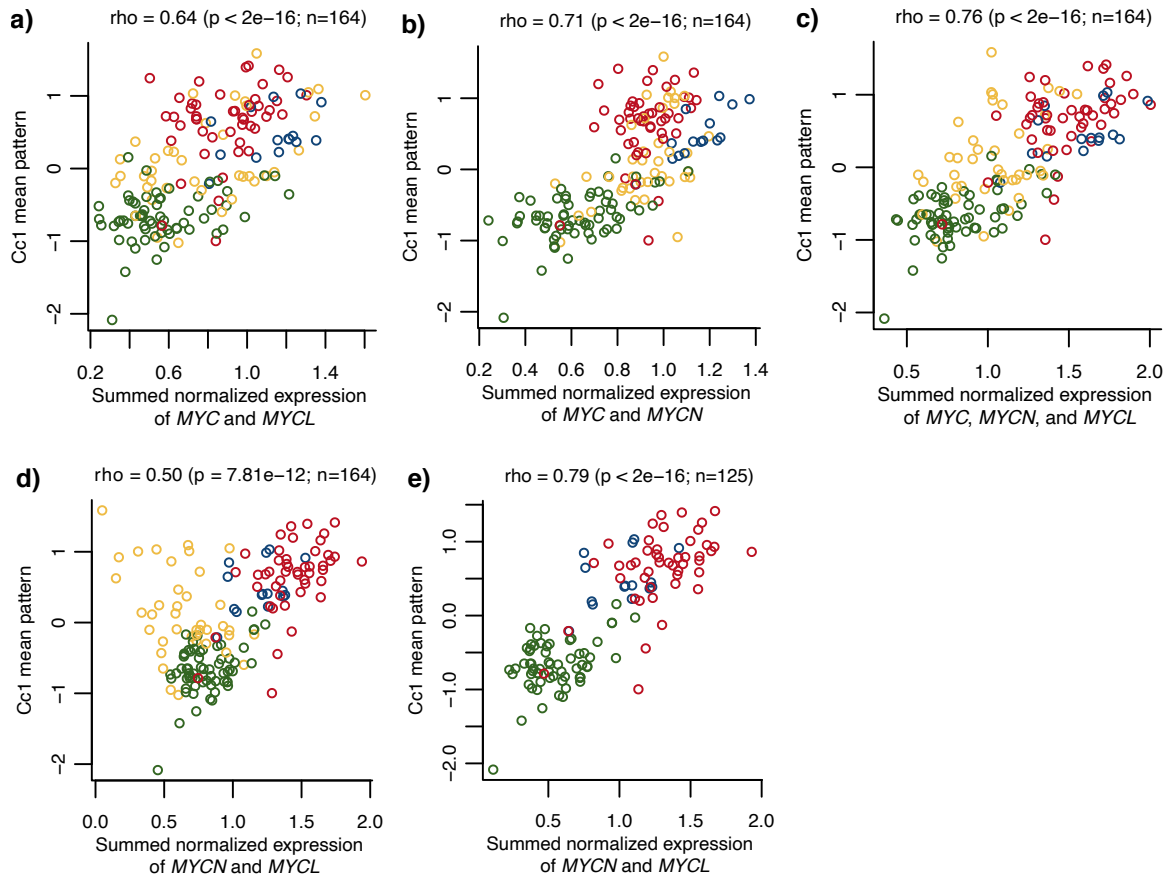


Figure 5.38: Comparison of Cc1 mean-pattern to the summed normalised expression of *MYC*, *MYCN*, and *MYCL* in PedBrain MB samples. Expression values were gene-wise log-transformed and scaled between 0 and 1. Normalised expression values for MYC family genes were summed per patient. Summed expression of **a)** *MYC* and *MYCL*, **b)** *MYC* and *MYCN*, **c)** *MYC*, *MYCN*, and *MYCL*, and **d-e)** *MYCN* and *MYCL*. **a-d)** whole PedBrain cohort. **e)** non-Group 3 samples. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and the number of samples n are displayed above each plot.

associated with protein-signalling cluster-1. We compared this signature to genes of Cc1. 92/116 genes mapped to the used Ensembl version, 78/92 genes were expressed in the ICGC MB cohort, 59/78 genes were part of the CLICK input (10132 genes), and 38/59 genes were part of Cc1. Applying a hypergeometric test and considering the CLICK input genes as background revealed that Cc1 was significantly enriched for genes upregulated in protein-signalling cluster-1 ($p = 2.90e-22$; Method Section 5.4.4.4). Overall, our analysis indicates that high expression of Cc1 resembles MBs with active cluster-1 protein-signalling as reported by Zomerman *et al.*

The shown data and related literature illustrated the various aspects of differential expressed lnc genes in MB. These aspects range from development associated expression pattern, pathways associations, and regulatory functions in *cis* and *trans*. Our co-expression analysis revealed the cluster Cc1 that included, among others, the lnc genes *PVT1*, *SNHG16*, *GAS5*, *ZFAS1*, and *DANCR*, all of them frequently described in cancer. Analyses showed that Cc1 is likely to be regulated by MYC family genes and is associated with protein translation processes and a distinct protein-signalling cluster.

Nevertheless, among the lnc genes that are well studied in the context of cancer and that were detected as differentially expressed in MB, was also the lnc tumour suppressor *MEG3* (Figure 5.34) [52]. The next three sections present the impact of *MEG3* in MB in detail.

5.3.3.6 Non-coding tumour suppressor *MEG3* as a prognostic biomarker in medulloblastoma

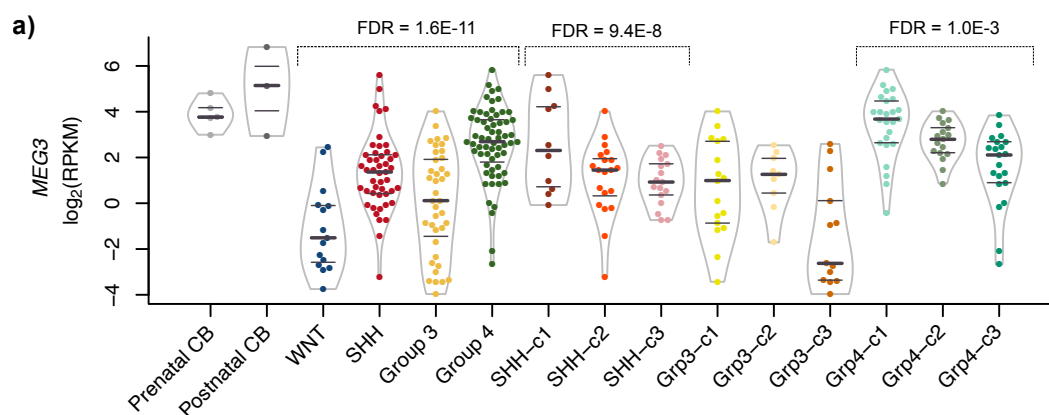


Figure 5.39: *MEG3* expression in MB. Violin plots show expression distribution. 25%, 50% and 75% quantiles are indicated by horizontal lines. Individual MB samples are shown as bee swarm plots. Shown FDR values relate to the differential expression among subgroups, subclusters of SHH MBs, and subclusters of Group 4 MBs.

Besides the biological role of lnc genes in MB, the evaluation of a correlation between lnc gene expression and clinical outcome was also of interest to understand potential clinical implications of lnc genes in MB. On the ICGC PedBrain discovery cohort, we performed systematic and gene-wise analyses for associations between gene expression and overall survival (OS). Here, tumour samples were split into low and high expression groups using an expression cutoff relating to an optimised separation of survival between the two groups ($p < 0.05$, Method Section 5.4.6.7). Here, a subsampling-based method was used to avoid overfitting (Method Section 5.4.6.7). We detected 82 lnc genes that showed a significant association with OS. However, only eight were measured on an external microarray MB cohort, published by Cavalli *et al.* [221], that we used for validation. (The Cavalli *et al.* cohort was used for survival analyses and validation due to a large cohort size and a long patient follow up.) The OS association could be validated on the external MB cohort for 2/8 lnc genes including the intergenic *MEG3* and the divergent *BAIAP2-AS1*. However, *BAIAP2-AS1* showed a weaker association with OS compared to *MEG3*. Additionally, *BAIAP2-AS1* was also strongly correlated with its divergent coding partner *BAIAP2* ($\rho = 0.75$), which would complicate an expression-based analysis without functional experiments¹. For this reason and the well-described tumour-suppressive function of *MEG3*, we investigated *MEG3* expression as a prognostic marker of survival in MB and potential biological mechanisms behind this association [52].

In our MB cohort, *MEG3* was significantly differentially expressed between subgroups and among subclusters of SHH and Group 4 tumours (Figure 5.39a). Group 4 and SHH MBs showed the highest and second highest expression of *MEG3* among main subgroups, respectively. Among subclusters within the subgroups Group 4 and SHH, *MEG3* was upregulated in SHH-c1 and Grp4-c1. *MEG3* showed the same expression pattern on the external validation cohort (Figure A.55). Additionally, *MEG3* was downregulated in MB compared to the normal cerebellum, as previously reported [337]. Only Grp4-c1 tumours showed *MEG3* expression comparable to normal cerebellum (Figure 5.39a).

¹*MEG3* was annotated as an antisense lnc gene that is positively correlated with its coding partner. However, newer Ensembl gene annotations (Ensembl v77 or higher) do not include this coding antisense gene that originally exactly overlapped with two exons of *MEG3* on the opposite strand, indicating this coding gene was annotated due to a technical artefact. Therefore, *MEG3* represents an intergenic lnc gene.

5 Medulloblastoma study

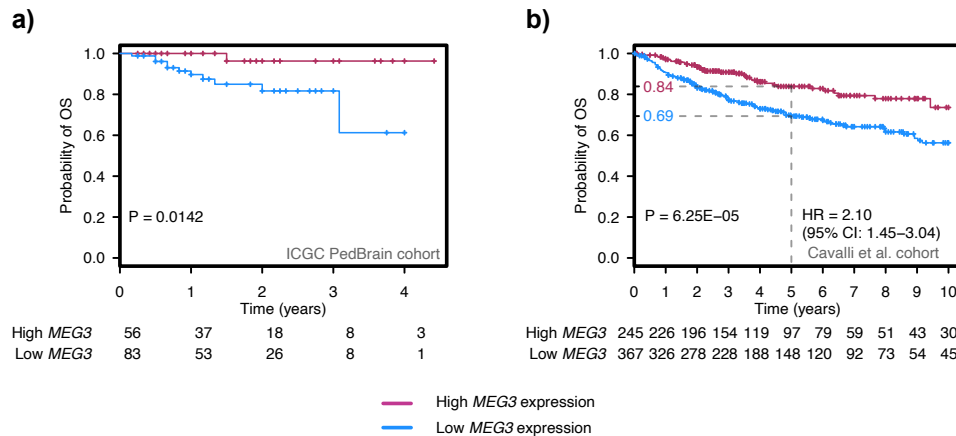


Figure 5.40: *MEG3* expression is prognostic of survival in MB. Kaplan-Meier curves show OS in *MEG3*-low- and -high-expressing MBs. MB samples that showed an *MEG3* expression \geq 60th percentile of the particular cohort were considered as high-expressing samples. **a)** ICGC PedBrain MB cohort. **b)** External validation cohort published by Cavalli *et al.* [221]. Shown p-value and hazard ratio relates to differences in survival between groups based on Cox regression (Methods section 5.4.6.7). Hazard ratio (HR) indicates risk comparing low-expressing vs. high-expressing samples. 95% confidence interval (CI) of HR is shown in brackets.

On the RNA-seq discovery and external validation cohort, *MEG3* expression was prognostic of OS in MB, where high *MEG3* expression was associated with a better outcome (Figure 5.40a-b). The robustly optimised expression cutoff, which we determined for the association between survival and *MEG3* expression using resampling, was around the 60th percentile of *MEG3* expression in our discovery and the external validation cohort. The top three cutoff solutions were the 60th, 59th and 58th percentile in the discovery and the 62nd, 61st and 60th in the validation cohort. We used the 60th percentile of *MEG3* expression as an optimised cutoff for both cohorts since it was the only solution that overlapped (equal to 4.48 RPKM in RNA-seq cohort). The external MB microarray cohort offered a more prolonged follow-up that allowed an estimation of the hazard ratio and the 5-years OS between *MEG3*-high- and low-expressing MBs (Figure 5.40b). The Kaplan-Meier curves showed an early separation, and the hazard ratio indicated a 2.1 times higher risk for *MEG3* low-expressing MBs to die of this disease (Figure 5.40a-b).

The larger external validation cohort was used to evaluate correlations between *MEG3* expression and overall survival in individual MB subgroups and subclusters taking the same *MEG3* expression value as cutoff as for the whole cohort (60th percentile). MB samples of the external cohort were classified into subclusters as described above (Section 5.3.1.2). The fractions of *MEG3*-high-expressing samples in the different subgroups of the external cohort followed the observed expression patterns of the discovery cohort (Figure 5.39). 91/172 samples in SHH, 130/164 samples in Group 4, 20/113 samples in Group 3, and 4/64 samples in WNT showed high *MEG3* expression. *MEG3* expression was able to predict OS in Group 4 and SHH but not in WNT and Group 3 MBs (Figure 5.41). However, the good outcome of WNT MBs with only 2/64 death cases did not provide data for meaningful survival analyses.

The observed significant association between OS and *MEG3* expression in the subgroups SHH and Group 4 varied in strength between or did not hold up in their subclusters. Among the SHH subclusters, only SHH-c1 showed a significant association. Here, *MEG3* expression stratified SHH-c1 cases into two groups displaying remarkably favourable in *MEG3*-high-expressing cases or miserable outcome with rapid death within few years in *MEG3*-low-expressing cases (Figure 5.42). Since our defined SHH-c1

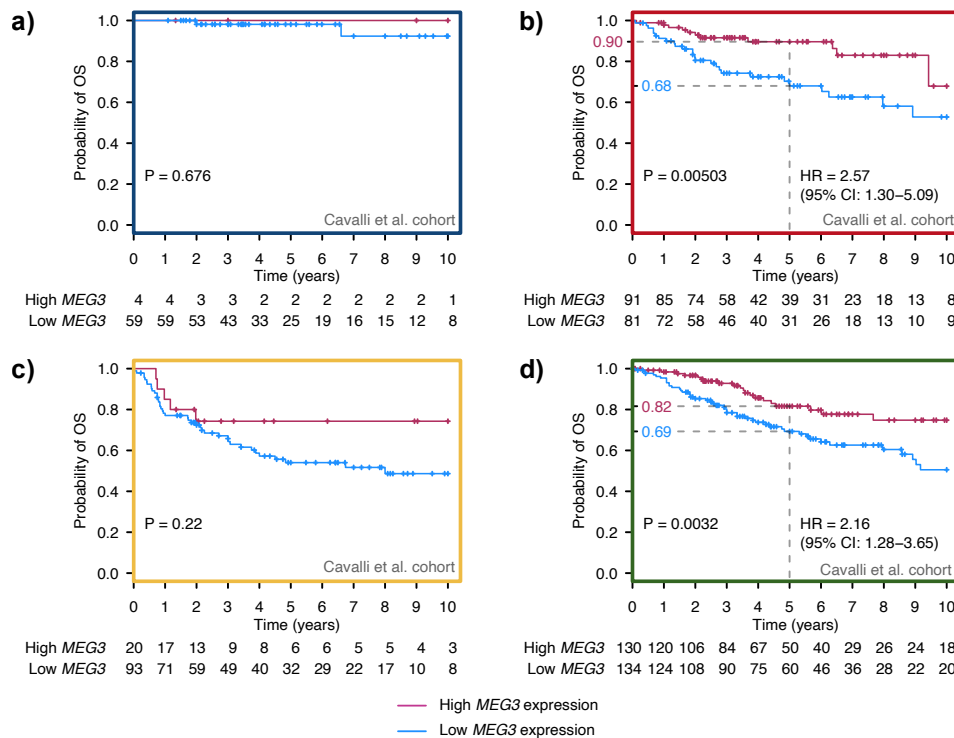


Figure 5.41: Association between *MEG3* expression overall survival in MB subgroups. Kaplan-Meier curves show OS in *MEG3*-low- and -high-expressing MBs. The external cohort published by *Cavalli et al.* is shown [221]. Shown p-value and hazard ratio relates to differences in survival between groups based on Cox regression (Methods section 5.4.6.7). Hazard ratio (HR) indicates risk comparing low-expressing vs. high-expressing samples. 95% confidence interval (CI) of HR is shown in brackets. **a)** WNT. **b)** SHH. **c)** Group 3. **d)** Group 4.

subcluster matched to two SHH subtypes of *Cavalli et al.* (SHH β and SHH γ), the sample stratification based on *MEG3* expression was repeated for the SHH subtypes of the *Cavalli* study. The Kaplan-Meier curves of the SHH subtypes looked similar to their related subclusters, and *MEG3* expression was able to stratify OS in SHH β and SHH γ MB independently and combined (Figure A.56). Among our subclusters of Group 4 MBs, *MEG3* expression was significantly associated with OS in Grp4-c2 and Grp4-c3 tumours (Figure 5.43). *MEG3*-high-expressing Grp4-c3 cases had a remarkably favourable outcome (Figure 5.43.c). *MEG3*-low-expressing Grp4-c2 cases showed the worst 5-years OS of 59% among *MEG3*-low-expressing and -high-expressing tumours of the Group 4 subclusters (Figure 5.43.b). Grp4-c2 *MEG3*-high-expressing samples had a 5-years OS that was similar to Grp4-c1 and Grp4-c3 *MEG3*-low-expressing samples at around 75%, which is the expected 5-years OS for Group 4 tumours [83].

Interestingly, the subclusters of SHH and Group 4 MB (SHH-c1, Grp4-c2, and Grp4-c3) that could be stratified via *MEG3* expression showed a certain expression distribution. Here, a larger number of the samples per subcluster was split between a *MEG3* expression level that was within/close to the normal cerebellum or below the expression range of the normal cerebellum (whole expression range: 7.66-113.37 RPKM; median: 15.69 RPKM), respectively (Figure 5.39a). In contrast, tumours of subgroups/subclusters that could not be stratified did show this clear split. Here, expression levels were either mostly below (WNT, SHH-c2, SHH-c3, and Group 3 subclusters) or within (Grp4-c1) the expression range of the normal cerebellum (Figure 5.39a). These results indicate that a wide range of *MEG3* expression that is split between normal expression and downregulation allows patient stratification of certain MB subclusters.

5 Medulloblastoma study

Taken together, our presented data showed that *MEG3* expression is a prognostic marker in MB, especially in SHH-c1 MBs, Grp4-c2, and Grp4-c3 MBs. *MEG3*-highly-expressing SHH-c1 and Grp4-c3 cases stood out due to a favourable prognosis. The shown results emphasise that *MEG3* acts as a non-coding tumour suppressor in MB. In order to understand the observed correlation between *MEG3* expression and OS in MB, we investigated the impact of *MEG3* expression on the tumour biology of MB.

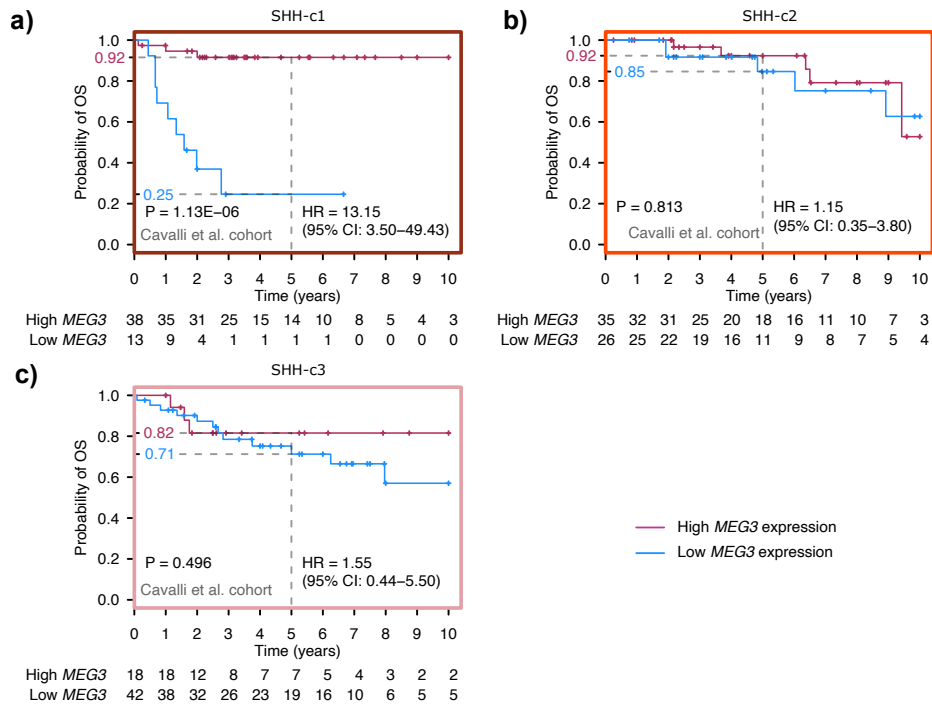


Figure 5.42: *MEG3* expression is prognostic of survival in SHH subclusters. Kaplan-Meier curves show OS in *MEG3*-low- and -high-expressing MBs. The shown cohort is published by *Cavalli et al.* [221]. Shown p-value and hazard ratio relates to differences in survival between groups based on Cox regression (Methods section 5.4.6.7). Hazard ratio (HR) indicates risk comparing low-expressing vs. high-expressing MB samples. 95% confidence interval (CI) of HR is shown in brackets. **a)** SHH-c1. **b)** SHH-c2. **c)** SHH-c3.

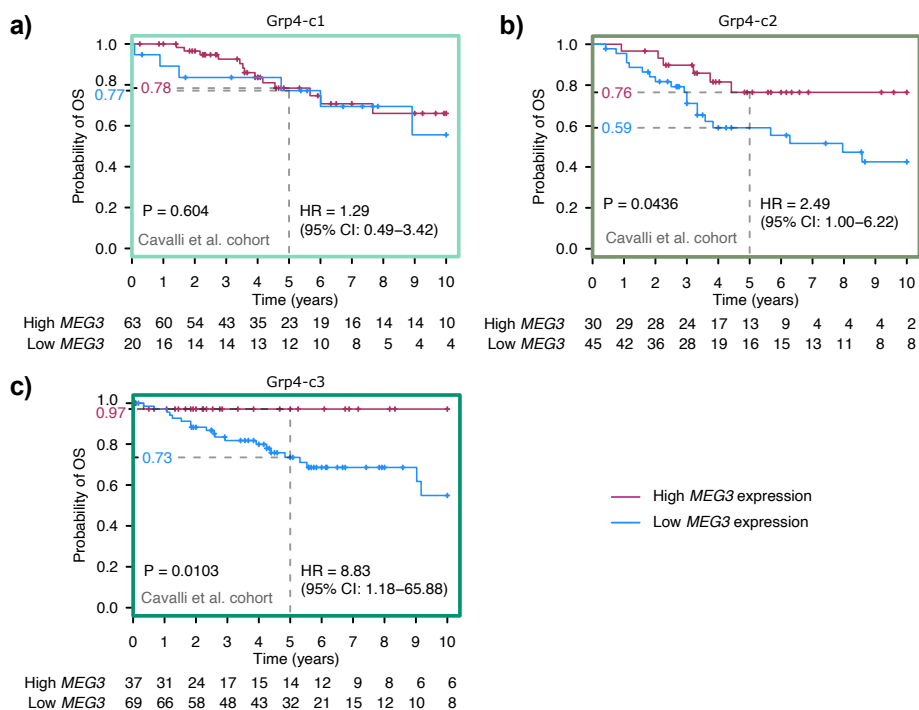


Figure 5.43: *MEG3* expression is prognostic of survival in Group 4 subclusters. Kaplan-Meier curves show OS in *MEG3*-low- and -high-expressing MBs. The shown cohort is published by *Cavalli et al.* [221]. Shown p-value and hazard ratio relates to differences in survival between groups based on Cox regression (Methods section 5.4.6.7). Hazard ratio (HR) indicates risk comparing low-expressing vs. high-expressing MB samples. 95% confidence interval (CI) of HR is shown in brackets. **a)** Grp4-c1. **b)** Grp4-c2. **c)** Grp4-c3.

5.3.3.7 Identifying genes correlated with *MEG3* expression in medulloblastoma

As previously described and summarised above (Section 2.5), *MEG3* acts as a tumour suppressor via multiple mechanisms. It can regulate gene expression by binding to DNA via triplex formation and cooperation with EZH2/PRC2 or as ceRNA (Section 2.5) [43, 51, 52, 338]. Therefore, we aimed for the identification of biological processes that are associated with *MEG3* expression in MB and potentially explain potential tumour-suppressive functions of *MEG3* in this disease. Here, we performed a gene expression correlation analysis on the PedBrain MB cohort using Spearman correlation to identify genes that were correlated with *MEG3* expression ($|\rho| \geq 0.3$, $FDR \leq 0.01$; Methods Section 5.4.6.9). This analysis was performed on the whole cohort as well as subgroup-wise on SHH, Group 3 and Group 4 MBs because of the subgroup-dependent association between *MEG3* expression and OS. The sample size of the WNT MBs ($n=15$) was too small to perform a robust expression correlation analysis.

Concentrating on results that were obtained from correlation analyses using the whole cohort, manual inspections of the results revealed spurious correlations. Here, we assumed the MB subgroups as a confounding factor, where spurious correlations were caused by subgroup-specific expression profiles that were shared between *MEG3* and a second gene. In order to filter out the spurious correlations, we applied a heuristic approach that was based on the assumption that non-spurious correlations are stable in at least one subgroup alone or after leaving out one of the four subgroups (Methods section 5.4.6.9). This heuristic approach was evaluated by comparing supposed spurious and non-spurious expression correlations on the basis of Spearman correlation coefficient ρ and the impact of the subgroups on gene expression. Here, likelihood ratios (LRs), which derived from the differential gene expression analysis between subgroups (Section 5.3.2.1; Methods Section 5.4.4.3), were taken as a gene-wise measure of subgroup-dependent expression. In the cohort-wide co-expression

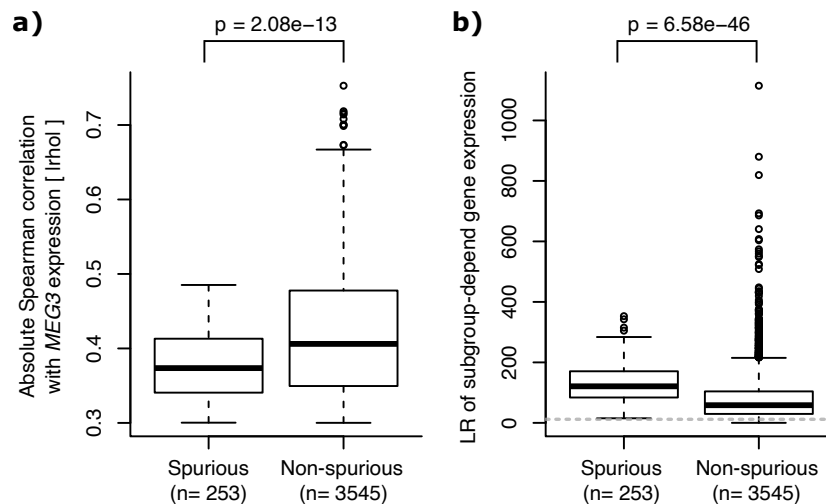


Figure 5.44: Evaluation of filtered spurious correlations in gene expression correlation analysis. **a)** Boxplots show the distribution of the absolute Spearman's rank correlation coefficient ρ of genes that are significantly correlated with *MEG3* expression. **b)** Boxplots show the distribution of gene-wise log-likelihood ratios (LR), indicating the strength of subgroup-dependent expression. The grey dashed line indicates the LR value where an FDR of 0.05 was reached. **a-b)** The two boxplots show genes related to spurious or non-spurious correlations. The shown p-values are based on a Wilcoxon rank-sum test.

analysis, 3798 coding genes were detected as significantly positively or negatively correlated with *MEG3* expression. A subset of 253 genes with supposed spurious correlations was identified showing significantly lower correlation coefficients as well as significantly higher likelihood ratios compared to genes with non-spurious correlations (Figure 5.44; Wilcoxon rank-sum test). The higher an LR value, the more gene expression depends on the subgroups. The opposing relationship between correlation coefficients and LRs suggested that the measured association of spurious correlations were mostly explained by subgroup-dependent expression as a confounding factor. These results indicate that the chosen heuristic approach identified potentially real spurious correlations that were caused by subgroup-dependent expression.

Results that derived from subgroup-wise correlation analyses using SHH, Group 3 or Group 4 MBs were not filtered for spurious correlations. Of course, also the subclusters could influence the correlation analysis as a confounding factor. However, due to the smaller sample size of the subclusters, we expected that spurious correlations were less frequent. Additionally, the chosen heuristic approach has limitations in terms of sample size, as described in the methods part (Method Section 5.4.6.9).

Overall, 3546, 2312, 533 and 2142 genes were significantly correlated with *MEG3* expression across the whole cohort, SHH, Group 3 or Group 4 MBs, respectively. Based on these four correlation gene sets, we identified biological processes and pathways that are associated with *MEG3* expression in MB.

5.3.3.8 *MEG3* as a regulator of proliferation and the TGF β pathway in medulloblastoma

We concentrated on genes, pathways and biological processes that were negatively correlated with *MEG3* expression to understand the potential tumour-suppressive function of *MEG3* in MB. We chose this focus since positively correlated genes showed a neuronal/developmental signature that was similar to Group 4-specific genes (Table 5.4) and negatively correlated genes pointed into directions that potentially explain tumour-suppressive functions of *MEG3* in MB, as detailed described below.

We identified the lowest number of negatively correlated genes (n=144) in Group 3 MBs in comparison to the whole cohort, and SHH and Group 4 MBs with 1938, 1459 and 1091 genes, respectively

Table 5.4: Functional gene signatures positively associated with *MEG3* expression. Genes positively correlated with *MEG3* expression (1607 genes) were tested for overrepresented gene signatures applying a hypergeometric test. Co-expression analyses were performed using the whole cohort (n=164). Methods section 5.4.6.9 and 5.4.6.10.

Source	Signature	p-value	FDR
Reactome	Neuronal System	7.81e-11	4.64e-09
Reactome	Developmental Biology	9.96e-06	1.71e-05

(Figure 5.45.a). In order to identify putative direct transcriptional targets among the negatively correlated genes, we integrated *in silico* predicted *MEG3* DNA binding sites (BS) into the analysis. The PWM of the binding motif was built on *MEG3* ChOP-seq data published by Mondal *et al.* (Methods Section 5.4.6.8). We obtained a PWM that comprised a GA-rich 25-nucleotide DNA binding motif, which is in line with previous publications [43, 339]. A gene was considered as a putative transcriptional target when its promoter or a regulating enhancer carried a *MEG3* binding site. Enhancer-gene pairs in MB were taken from Lin *et al.* [218]. Among the negatively correlated genes, a minority was associated with a *MEG3* binding site and ranged from 31-336 genes per gene set (Figure 5.45a).

We performed an overrepresentation analysis to identify biological processes and pathways that were enriched in the four sets of negatively correlated genes (whole cohort, SHH, Group 3, and Group 4) and their subsets of *MEG3*-BS-associated genes. The overrepresentation analysis was based on a hypergeometric test using gene signatures from KEGG, Reactome, WikiPathway and PID, which were downloaded from ConsensusPathDB (Methods Section 5.4.4.4) [340–344]. The reason for the integration of multiple pathway databases was to prove the reliability of the detected enrichments. We detected a strong enrichment for cell cycle processes among the genes that were negatively correlated across the cohort, SHH MBs, and Group 4 MBs. Here, the strongest enrichment was detected for the mitotic phase and the cell-cycle-related FOXM1 pathway [345, 346] (Figure 5.45.b). Cell-cycle-related processes were weakly or non-enrichment in the Group 3 gene set. Among *MEG3*-BS-associated genes, a similar enrichment pattern was seen for cell-cycle-related processes among the four gene sets (Figure 5.45.b). Enrichments for the TGF β pathway gained strength in *MEG3*-BS-associated genes and were present in the cohort, SHH, and Group 4 gene set with Group 4 showing the strongest enrichment.

We compared the overlap between the sets of genes that were negatively correlated with *MEG3* expression. This comparison showed that the fractions of genes exclusively detected within a set were similar and ranged between 34-36% per gene set (Figure 5.45c). The Group 3 gene set showed a similar frequency of overlapping genes compared to the remaining three gene sets. Twenty-three genes were in common between all four gene sets, but 13/23 genes belonged to the family of histone genes (Figure 5.45c). The expression of many histone genes was reported to be cell-cycle-dependent and, therefore, the genes that were detected in all four correlation gene sets probably relate to cell cycle processes [347]. Among others, the correlation gene sets contained the genes *CDK1*, *CCNB1*, *MYC*, and *TGFBR1*, which were of interest regarding the enriched pathways and processes (Figure 5.45c). These four genes also carried a *MEG3* BS in a promoter or associated enhancer.

MEG3 was moderately negatively correlated with *CDK1* and *CCNB1* across the MB cohort (Figure 5.46a-b). However, the negative correlation with *CDK1* and *CCNB1* was stronger collectively in SHH and Group 4 cases (Figure 5.46a-b). *CDK1* and *CCNB1* were expressed in all MB subgroups and subclusters, whereby *CCNB1* was higher expressed and showed more variation compared to *CDK1* (Figure A.57, A.58). Overall, the strong negative correlation with *CDK1* and *CCNB1* in SHH and Group 4 MBs and the predicted *MEG3* BS in promoter/enhancer regions indicate the *MEG3* negatively regulates the expression of *CDK1* and *CCNB1* in these two subgroups. The shown results substantiate the observed negative association between *MEG3* expression and mitotic cell cycle since *CDK1* and *CCNB1* genes are important regulators of the cell cycle during G2-M progression [348, 349].

5 Medulloblastoma study

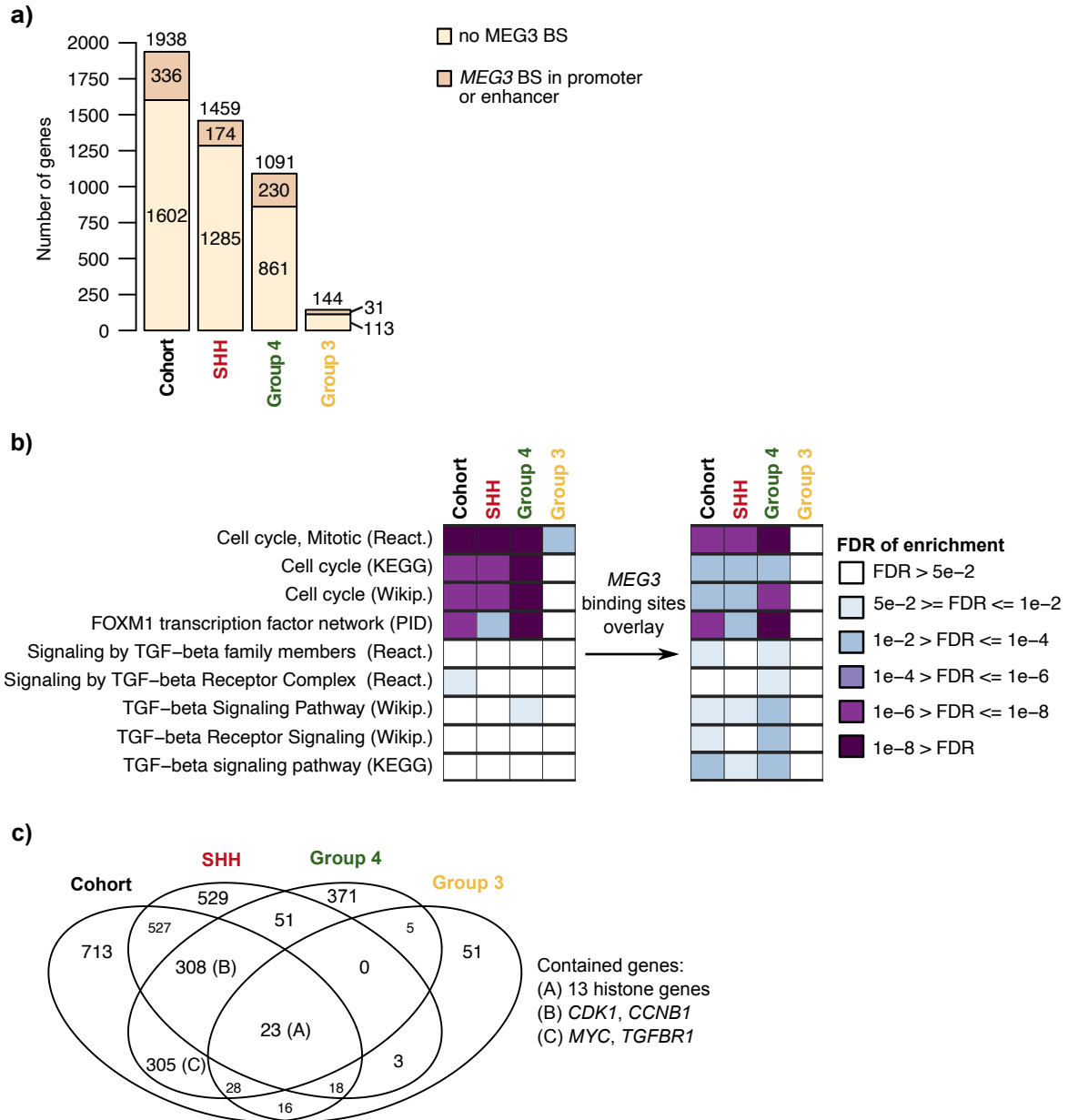


Figure 5.45: Genes and pathways negatively correlated with *MEG3* expression in MB. **a)** The stacked bar plots show the total number of genes that were negatively correlated with *MEG3* expression and the number of genes associated with and without *MEG3* BS, as indicated by the legend (Method Section 5.4.6.8, 5.4.6.9). Each bar relates to one gene set that was identified across the whole MB PedBrain cohort, SHH MBs, Group 3 MBs or Group 4 MBs. **b)** Heatmap visualises the enrichment of selected biological processes and pathways among genes that were negatively correlated with *MEG3* expression. The enrichment is given by the FDR of the applied hypergeometric test (Methods Section 5.4.6.10). Left heatmap: Genes that were negatively correlated with *MEG3* expression across the whole cohort, SHH MBs, Group 3 MBs, or Group 4 MBs, as shown in panel a). Right heatmap: Subset of genes that were negatively correlated with *MEG3* expression and were associated with a *MEG3* BS, as shown in panel a). **c)** The Venn diagram visualises the overlap of the four correlation gene sets that are shown in panel a). Positions of selected genes in the Venn diagram are given by the capital letters A-C in the brackets.

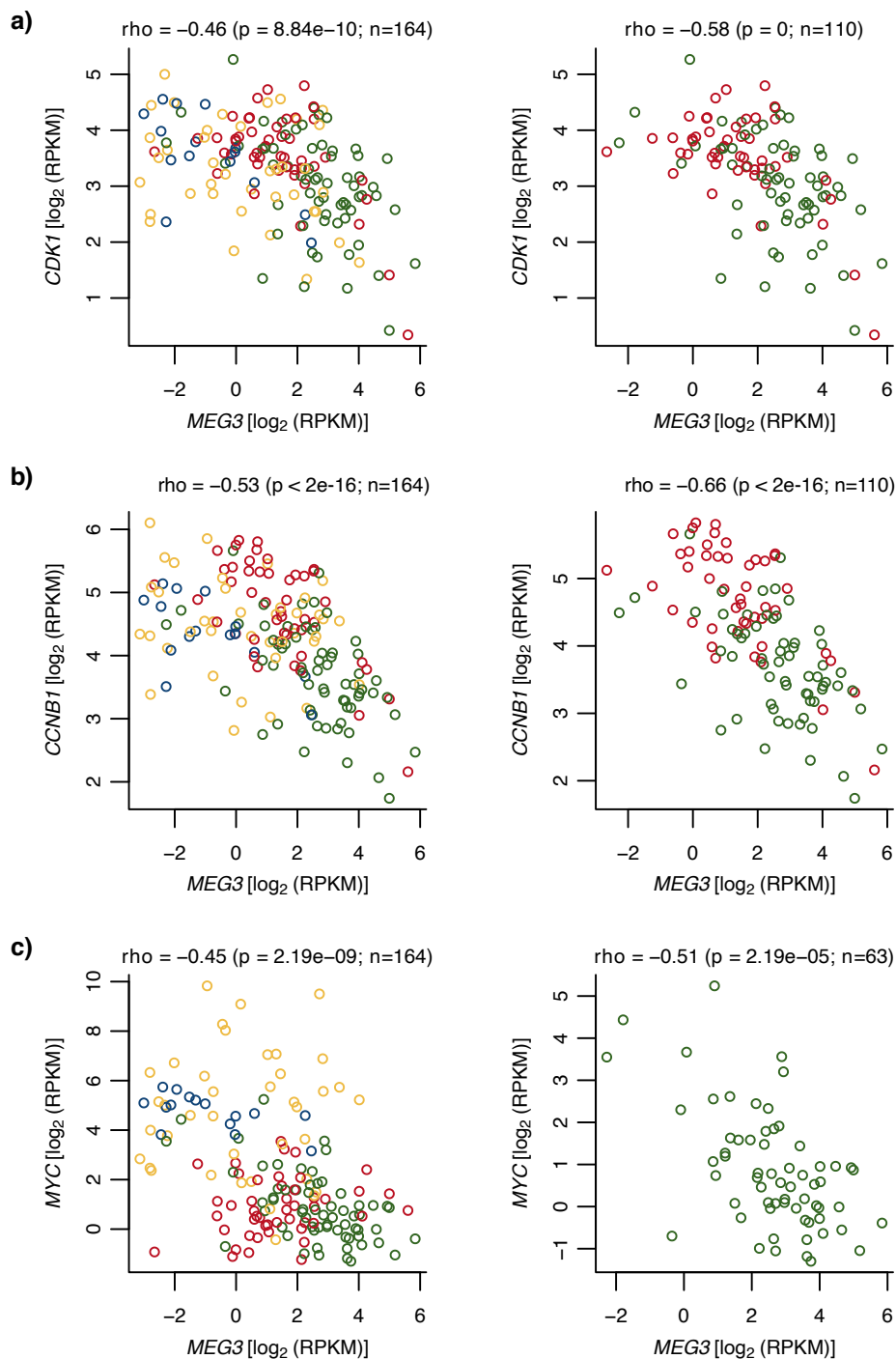


Figure 5.46: *MEG3* is negatively correlated with **a)** *CDK1*, **b)** *CCNB1*, and **c)** *MYC*. Scatter plots show **left)** whole cohort, **right** SHH and Group 4 samples or Group 4 samples only. Colours indicate subgroups: WNT=blue, SHH=red, Group 3=yellow, Group 4=green. Spearman correlation coefficient, related p-value, and the number of samples n are displayed above each plot.

The oncogene *MYC* showed a moderate correlation across the cohort and Group 4 MBs, respectively, but the correlation was stronger across Group 4 MBs (Figure 5.46.c). *MYC* was also differentially expressed among Group 4 subclusters (FDR = $9.51e-05$) with higher average expression and wider

expression range in the Grp4-c3 subcluster compared to Grp4-c1 and Grp4-c2 (Figure A.59). The observed negative correlation with *MYC*, the *MEG3* BS, and the wide expression range of *MYC* in Grp4-c3 cases indicate that *MEG3* could be predictive of survival in Grp4-c3 cases via regulating *MYC*.

We detected that the Transforming Growth Factor Beta Receptor 1 (*TGFBR1*), a receptor of the cancer-relevant TGF β pathway [350], was expressed in all MB subgroups with lower expression in Grp4-c1 among Group 4 subclusters. *TGFBR1* was negatively correlated with *MEG3* expression across the cohort, SHH MBs, and Group 4 MBs (Figure 5.47a-b). Here, the negative expression correlation between *MEG3* and *TGFBR1* was the most pronounced in Group 4. We validated that the significant negative correlation on the external MB cohort ($\rho = -0.43$, $p = 1.456e-35$, $n=763$), SHH MBs ($\rho = -0.17$, $p = 1.073e-02$, $n=223$), and Group 4 MBs ($\rho = -0.36$, $p = 3.205e-11$, $n=326$). Here, the bigger cohort size revealed a significant negative correlation also in WNT MBs ($\rho = -0.33$, $p = 5.543e-03$, $n=70$) (Figure A.60). Group 3 MBs showed no significant correlation in both cohorts (Figure 5.47b and A.60.d). The significant but weaker negative correlation values between *MEG3* and *TGFBR1* on the external microarray cohort could originate from the smaller dynamic range of expression microarrays compared to RNA-seq. Lin *et al.* have reported an intragenic enhancer of *TGFBR1* that interacts with the *TGFBR1* promoter and regulates *TGFBR1* expression in MB [218]. We found that this intragenic enhancer carried an *in silico* predicted *MEG3* BS region that spanned 39 bps (GA-rich, reverse strand) and contained six overlapping *MEG3* binding sites (Figure 5.47c). (The *MEG3* binding motif was 25 bps long and, therefore, fitted several times in the 39 bps long region.) To evaluate the regulatory strength of *MEG3* on *TGFBR1* expression, we utilised our inferred GRN of the main subgroups (Section 5.3.2.2 and 5.3.2). Based on the GENIE3-derived interaction weights, the six TFs that showed the highest interaction weights with *TGFBR1* were compared, considering *MEG3* as a TF. Here, *MEG3* showed clearly the strongest interaction weight among the top six TFs. *MYC* showed the second-highest interaction weight but was positively correlated with *TGFBR1* expression. *MYC* was significantly positively correlated with *TGFBR1* in Group 3 MB ($\rho = 0.39$, $p = 0.0154$, $n=39$), but not in SHH and Group 4 MBs ($\rho = 0.15$, $p = 0.12$; $n=110$; Figure A.61, A.62). The enhancer that carried a *MEG3* binding site in MB was different from the reported enhancer that carries a *MEG3* BS and regulates *TGFBR1* expression in breast cancer (as introduced above in Section 2.5). However, in MB, *MEG3* could regulate *TGFBR1* through a different enhancer since the results presented above provide a strong support. The negative regulation of *TGFBR1* by *MEG3* might provide a direct link for the negative association between *MEG3* expression and TGF β pathway enrichments that we observed in MB (Figure 5.45.b).

As summarised above in Section 2.5, *MEG3* can be regulated through the DNA methylation of the IG- and *MEG3*-DMR since *MEG3* is expressed from parentally imprinted region [58]. Using DNA methylation array data of the ICGC PedBrain cohort ($n=155$), we calculated average methylation level of the *MEG3*-DMR across the 33 covert CpGs of this DMR; the IG-DMR was not covered by the array (Methods Section 5.4.6.11). The average methylation level at the *MEG3*-DMR was significantly differentially methylated between the MB subgroups (Figure 5.48.a). An intermediated methylation level around 0.5 is expected for a parentally imprinted locus because only one of the two alleles is methylated [58, 351]. In MB, the median methylation level of the *MEG3*-DMR was 0.497 and 0.540 in Group 3 and 4 MB, respectively, and 0.611 and 0.670 in WNT and SHH MB, respectively (Figure 5.48.a). We observed the most robust preservation of the *MEG3*-DMR imprinting in Group 4 since the methylation levels in Group 4 were tightly distributed around the median of 0.540. Group 3 MBs showed a broader distribution and the most frequent hypomethylation in a subset of samples. WNT and SHH MB showed a trend for hypermethylation. However, *MEG3* expression and *MEG3*-DMR methylation level were not correlated in MB (Figure 5.48.b). Our observation is in line with a previous publication in the context of meningioma (a tumour of the meninges), where Zhang and colleagues concluded that mechanisms additional to DNA methylation contribute to the transcriptional regulation of *MEG3* [62].

Taken together, the pathways and biological processes that we found to be negatively correlated with *MEG3* expression might explain the discovered association between *MEG3* expression and OS in SHH and Group 4 MB, including the subgroup-dependent associations. A strong negative correlation with

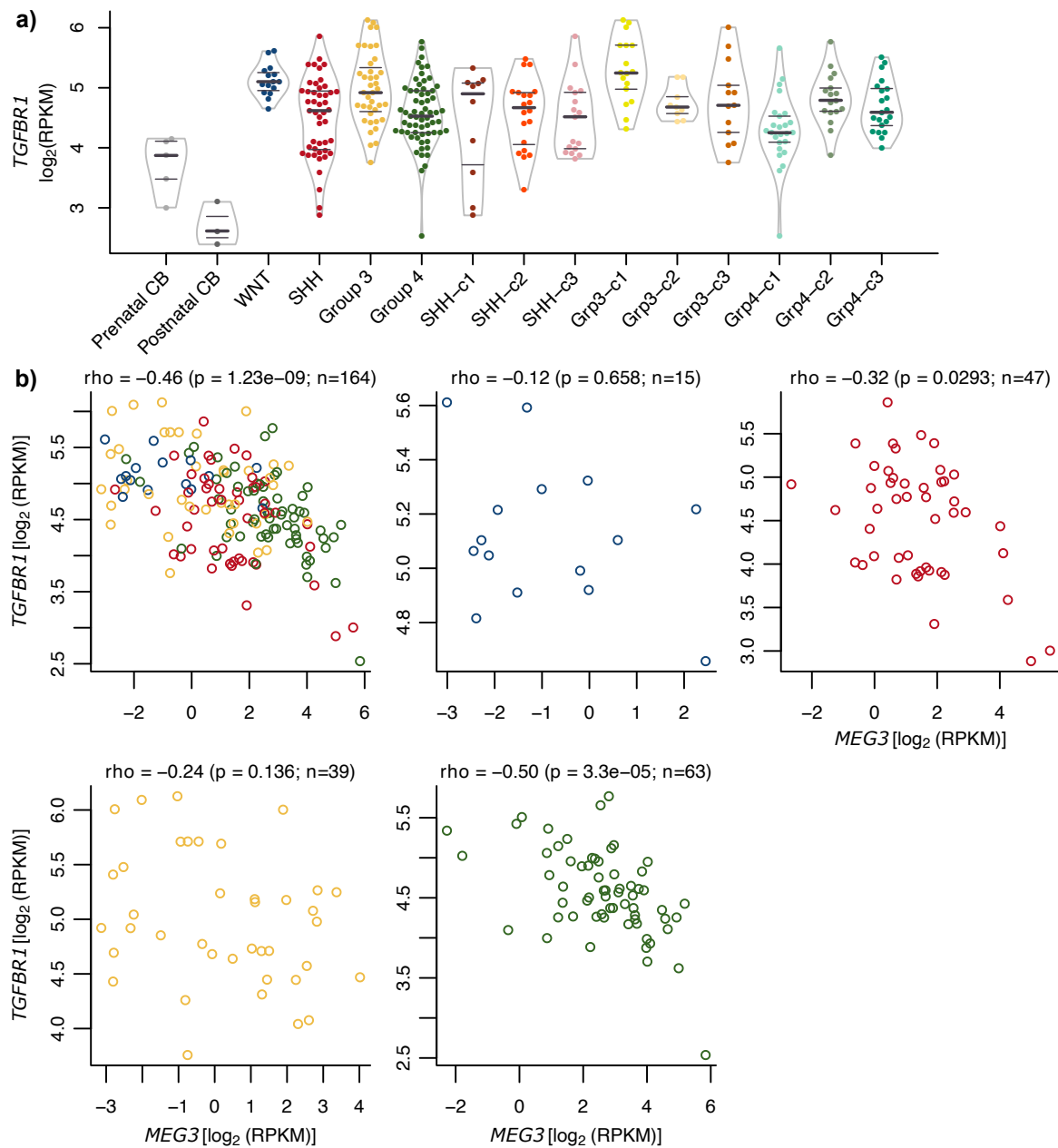


Figure 5.47: *MEG3* expression is a potential negative regulator of *TGFBR1*. **a)** Violin plots show expression profile of *TGFBR1* in ICGC PedBrain MBs. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual MB samples are shown as bee swarm plots. **b)** Scatter plots show expression of *MEG3* and *TGFBR1* across all PedBrain MB samples and across samples of each subgroup individually. Colours relate to panel a). Spearman correlation coefficient, related p-value, and number of samples *n* is displayed above each plot. **c)** Schematic of the *TGFBR1* locus. The location of the intronic enhancer (blue) and the enhancer-overlapping *MEG3* binding site (BS) region (purple) are highlighted, and coordinates are given in matching colours. Different transcripts are shown. Black boxes indicate exons. Arrows show the direction of transcription. **d)** GENIE3-inferred interaction weights between *TGFBR1* and the six top-ranked regulators.

mitotic cell cycle processes and the essential cell-cycle-controlling genes *CDK1* and *CCNB1* indicated that *MEG3* could inhibit proliferation also in MB. Negative regulation of *MYC* and the TGF β pathway could be at least in Group 4 an additional mechanism of how *MEG3* acts as a tumour suppressor in MB. The previous Sections showed in detail the expression pattern of *MEG3* in MB. However, understanding the expression pattern of *MEG3* in normal cerebellum during development and in adults could bring additional insights into the biological role of *MEG3* in MB.

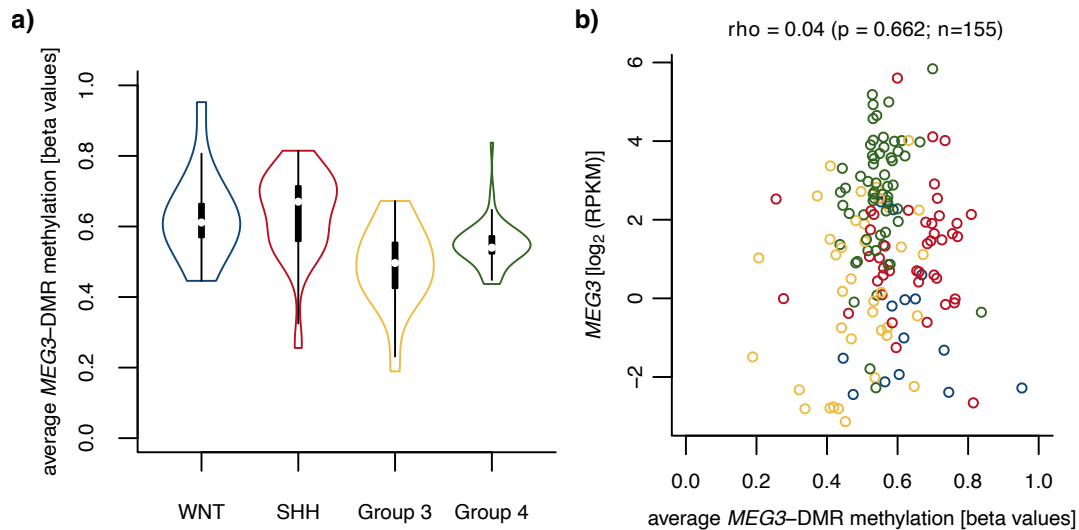


Figure 5.48: Methylation of *MEG3*-DMR in MB subgroups. 155 MB samples of the ICGC PedBrain cohort with matching DNA methylation and RNA-seq data are shown. **a)** Violin plots show average methylation levels across 33 CpG sites at the *MEG3*-DMR. The *MEG3*-DMR is significantly differentially methylated between MB subgroups ($p = 1.55e-09$, F-test). Upper and lower edge of black boxes indicate 25% and 75% quantile. White dot indicates the median. **b)** Scatter plot of average *MEG3*-DMR methylation levels and *MEG3* expression. Colours of data points indicate subgroups, as shown in panel a). See Method Section 5.4.6.11.

5.3.3.9 *MEG3* expression in normal brain and cerebellum

We depicted the expression of *MEG3* in the cerebellum of the human and developing mice. *MEG3* was lower expressed in prenatal vs. postnatal human cerebellum and non-cerebellum brain (Figure 5.49a). However, *MEG3* showed higher expression in pre- and postnatal CB compared to non-cerebellum brain tissues (Figure 5.49a). ISH in E14.5 mouse, obtained from the gene expression atlas GenePaint, showed clear expression of *Meg3* across the CNS with absent expression in the ventricular zone of the neocortex, midbrain and cerebellum (Figure 5.49.b)[352]. In E14.5 murine cerebellum, *Meg3* was expressed in the postmitotic differentiating and NT zone (Section 5.1.2). The ISH indicated that *Meg3* is expressed in postmitotic neurons but not in progenitor cells of the cerebellum, which is supported by previously published Single-cell RNA-seq of the mouse brain [48].

We detected that the strong negative expression correlation between *MEG3* and *TGFBR1* in MB was also present in human CB across pre- and postnatal stages (Figure 5.50.a). Integrating all available brain tissues in BrainSpan, the *MEG3*-*TGFBR1*-correlation was only present among prenatal samples (Figure 5.50.b-d). Our results indicate that regulation of *TGFBR1* by *MEG3* relates to a transcriptional program that is active in the prenatal brain, whereas in CB, this program is active in pre- and postnatal tissue.

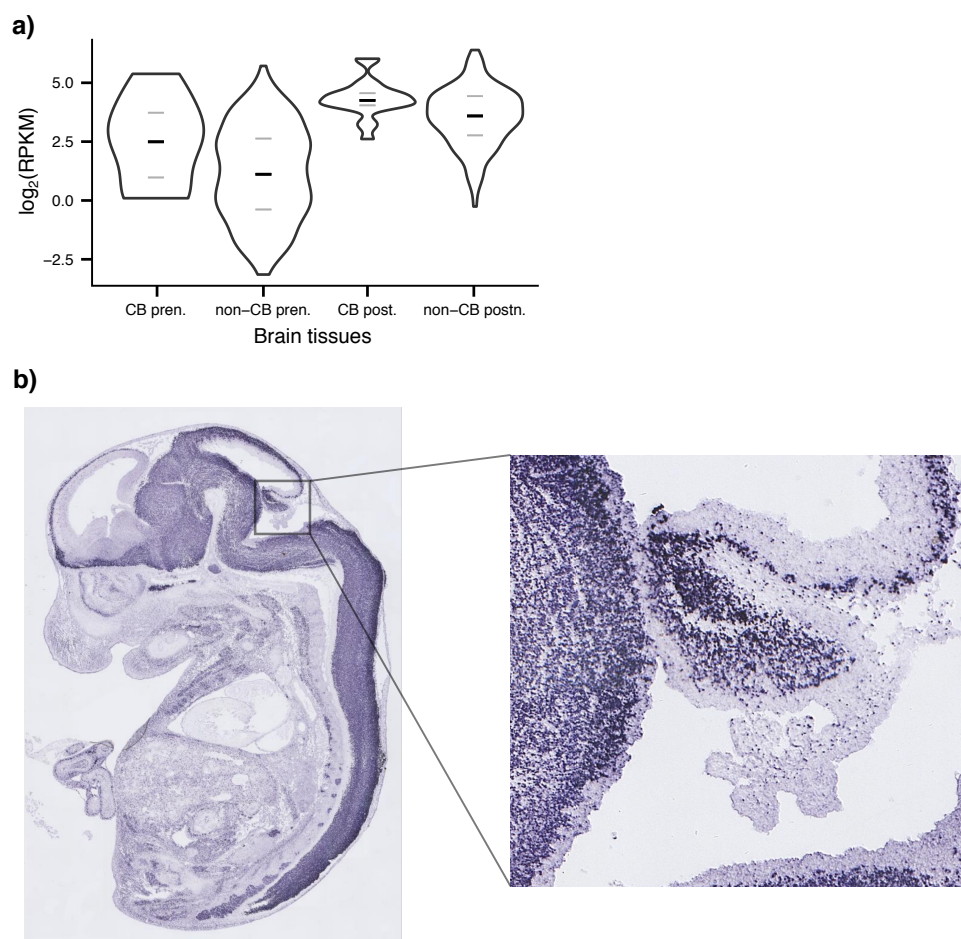


Figure 5.49: *MEG3/Meg3* expression in human and mouse brain. **a)** Violin plots show *MEG3* expression in human cerebellum (CB) and non-cerebellum brain tissues of BrainSpan[299]. Left two violin plots: prenatal tissue (pren.). Right two violin plots: postnatal tissue (postn.). **b)** ISH of *Meg3* in E14.5 mouse. Image credit for ISH: GenePaint expression atlas [352].

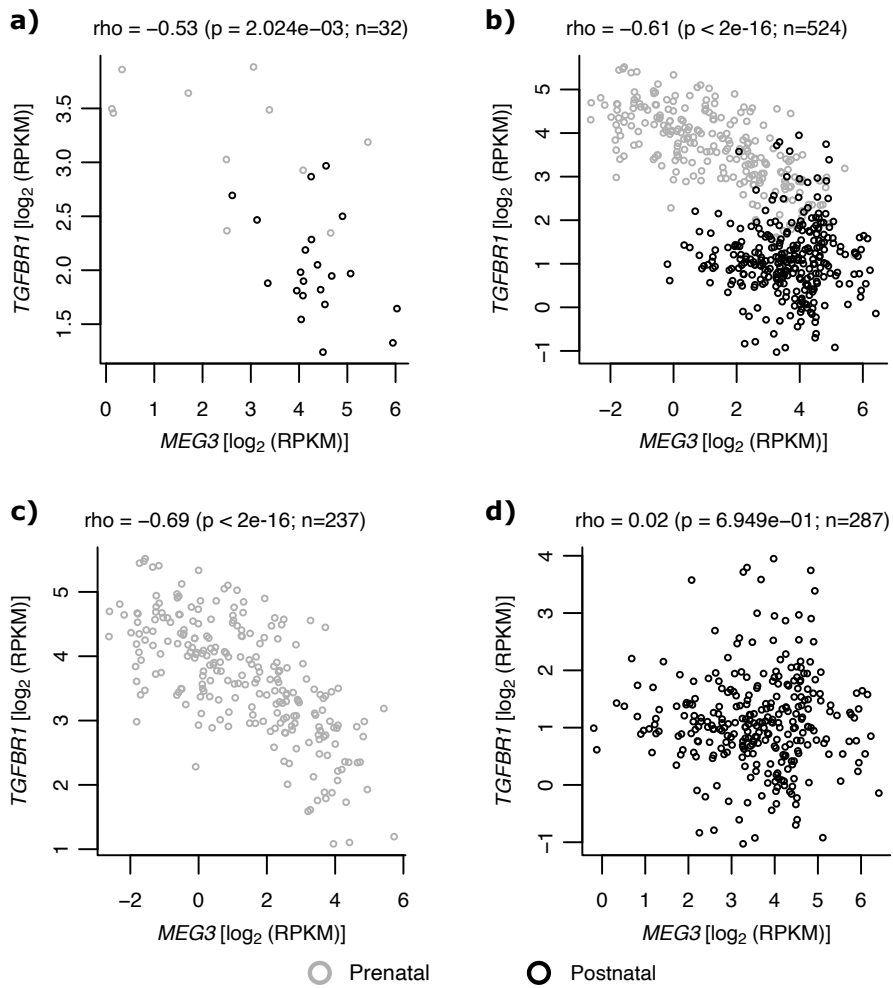


Figure 5.50: *MEG3* and *TGFBR1* correlation in the human brain. Scatter plots show *MEG3* and *TGFBR1* expression in **a)** pre- and postnatal cerebellum, **b)** pre- and postnatal brain tissues, **c)** prenatal brain tissues, and **d)** postnatal brain tissues. Spearman correlation coefficient, related p-value, and the number of samples n are displayed above each plot. RNA-seq data derived from BrainSpan.

5.4 Material and Methods

Contributions of third persons comprising colleagues and ICGC PedBrain consortium partners are indicated at the begin of each section.

5.4.1 Preliminaries: Correlation coefficient evaluation

Correlation between two expressed genes was measured using Spearman or Pearson correlation coefficients. For the evaluation of individual measured coefficients, several cutoffs were used [353]:

Table 5.5: Cutoffs for the evaluation of measured correlation coefficients.

Evaluation	Cutoff
non-correlated	$ \text{coefficient} < 0.3$
weak correlation	$0.3 \leq \text{coefficient} < 0.5$ and significant (p, FDR < 0.05)
moderate	$0.5 \leq \text{coefficient} < 0.6$ and significant (p, FDR < 0.05)
strong	$0.6 \leq \text{coefficient} $ and significant (p, FDR < 0.05)

5.4.2 Discovery cohort

5.4.2.1 Tissue collection, clinical data, RNA-sequencing, and RNA-seq read processing

The tissue collection of medulloblastoma (164 MB samples) and cerebellum controls (eight samples) was done by PedBrain consortium partners. “[A]ll patient material was collected after receiving informed consent according to ICGC guidelines and as approved by the institutional review board of contributing centres” [222]. Consortium partners provided clinical information, including the overall survival and age of the patients.

Tissue sample processing, RNA-sequencing, and RNA-seq read mapping was done by members of the Gene Regulation & System Biology of Cancer (GRSBC) Lab, as previously described: “RNA was extracted from fresh frozen tissue samples using the AllPrep DNA/RNA/Protein Mini kit (Qiagen) including DNase I treatment on a column. All samples were subjected to quality control on a Bioanalyzer instrument. RNA sequencing libraries were prepared from 10 μ g of total RNA. Strand-specific RNA sequencing was performed following a protocol described previously. Sequencing was carried out with 2 \times 51 cycles on a HiSeq 2000 instrument (Illumina). All reads were aligned to the human reference genome (1000 genomes version of human reference genome hg19/GRCh37) using BWA (v 0.5.9-r16). Aligned reads were converted to the SAM/BAM format using SAMtools.” [218]. Mapped reads were annotated to Ensembl v70 [354]. Read counts per gene were used to calculate RPKM values (reads per kilobase per million reads) [78]. RPKM values were normalised for transcriptome composition effects based on trimmed mean of M values (TMM) using edgeR (v3.10.2) [80].

5.4.2.2 DNA Methylation data and subgroup classification

The DNA methylation data of the discovery cohort were provided by PedBrain consortium partners. Here, Infinium HumanMethylation450 BeadChip arrays (450k array) were used to perform the DNA methylation profiling, as previously described [222]. For further information on microarray-based DNA methylation profiling, please see [355] and [356]. Subgroup classifications of the MB samples were based on 450k array data and provided by PedBrain consortium partners following the procedure of Hovestadt *et al.* [357].

5 Medulloblastoma study

For the DNA methylation analysis that was performed in this thesis, the provided IDAT-files were processed using the Bioconductor package `minfi` (v1.30.0; R version 3.6.0) [358]. Intensities were normalised using the `preprocessFunnorm` function with default parameters. The `getBeta` function was used to obtain beta values. Beta values were corrected for potential inter-chip batch effects using `ComBat` from the `sva` package (v3.28.0) [359]. Here, subgroup assignments were additionally provided to avoid the corrections of tumour-related features.

5.4.2.3 CNV data

Copy number estimations based on whole-genome sequencing for the discovery cohort were provided by PedBrain consortium partners [222]. The provided tumour-purity-corrected copy numbers were integrated for further analysis.

5.4.3 External resources

Several external resources were integrated into the analysis and for validation. These external resources comprised an MB cohort, RNA sequencing data of human brain tissues, CAGE data of human tissue and cell types, and a catalogue of lnc gene loci and loci-associated features.

5.4.3.1 External medulloblastoma cohort

The external medulloblastoma cohort relates to the introduced study of Cavalli *et al.* that included 763 primary MB samples compressing 70 WNT, 233 SHH, 144 Group 3, and 326 Group 4 tumours. The authors used Affymetrix Gene 1.1 ST arrays for the profiling of gene expression. Processed gene expression data, including gene mapping to Ensembl, were downloaded from Gene expression Omnibus (GEO; accession ID: GSE85217). Subgroup, subtype, and clinical information were taken from supplemental material published by Cavalli *et al.* [221]. This cohort was used for a comparison between reported subtypes and identified subclusters as well as survival-related analyses due to the larger cohort size and longer patient follow-up.

5.4.3.2 Active enhancers and ChIP-seq in medulloblastoma

Coordinates of active enhancers in MB as well as putative gene targets of these enhancers were taken from supplementary data published by Lin *et al.* [218]; this study was introduced above in Section 5.1.4. Since the work of Lin *et al.* was done in the framework of ICGC PedBrain, collaboration partners provided processed ChIP-seq data of HLX, LMX1A, and LHX2 from primary MB samples. ChIP-seq of HLX was done from a Group 3 MB, LHX2 from a Group 3 and 4 sample, and LMX1A from a Group 4 sample. Here, ChIP-seq peaks were called using MACS and peaks with a p-value < 1e-09 were considered; sample-matched input DNA was used as background [218, 360].

Published ChIP-seq of NEUROD1 and OTX2 that had been done from MB cell lines (D283 - Group 3/4, D341 - Group 3) was downloaded from Gene Expression Omnibus (GEO) (GEO accession ID: GSE92585; sample accession ID: GSM2432952, GSM2432944, GSM2432942, GSM2432950, GSM2432957, GSM2432949) [226, 306]. These data were processed by members of the GRSBC lab. ChIP-seq reads were mapped to the human reference genome (1000 genomes version of human reference genome hg19/GRCh37) using BWA (v 0.5.9-r16). ChIP-seq peaks were called using MACS applying a p-value < 1e-09 as cutoff for considered peaks.

5.4.3.3 BrainSpan: External RNA sequencing data from human brain tissue

The BrainSpan data set comprises 524 RNA-seq samples from 26 different human brain structures/regions and different pre- and postnatal stages. The data set includes also eleven prenatal and 21 post-

natal cerebellum tissue samples. RPKM values were downloaded from the BrainSpan website [299]: http://www.brainspan.org/api/v2/well_known_file_download/267666525. Here, the RNA had been sequenced using an unstranded protocol.

5.4.3.4 FANTOM CAT

The FANTOM CAT resource is part of the FANTOM5 project and provides an accurate atlas of lnc gene loci and a collection of annotated features that relate to these loci [133, 361]. In FANTOM CAT, the CAGE technology provided precise locations of 5' ends of lnc genes on the genome, which allowed the accurate annotation of the three types of lnc genes divergent, antisense, and intergenic [133]. The CAGE data set covers a comprehensive collection of different human tissue and cell types and has been used to calculate a score that evaluates the expression specificity of individual lnc genes across 69 different cell types facets; cell type-specific expression had been termed as “enriched expression” by the authors [133].

The resource FANTOM CAT was obtained by downloading supplementary information of the original publication as provided by Hon *et al.* [133]. Related CAGE-based gene expression data were downloaded from the FANTOM5 website (https://fantom.gsc.riken.jp/5/suppl/Hon_et_al_2016/data/expression/expression_atlas/FANTOM_CAT.expression_atlas.gene.lv3_robust.rle_cpm.tsv.gz).

5.4.4 Transcriptome profiling

5.4.4.1 Clustering into subgroups

The RNA-seq-based *de novo* clustering of MB samples into four subgroups was done by taking the most variable protein-coding genes, applying non-negative matrix factorisation with 60 iterations, and assuming four clusters (see Section 3.3.1). Most variable genes were selected using the 85th percentile of the absolute deviations to the median of the expression value of a gene termed as percentile absolute deviation (PAD) (with reference to the median absolute deviation - MAD [362]):

$$\tilde{x}_g = \text{median}(\mathbf{x}_g) \quad (5.1)$$

$$PAD_g = P_{85}(|\mathbf{x}_g - \tilde{x}_g|), \quad (5.2)$$

where \mathbf{x}_g is a vector of expression values of gene g . PAD was calculated on \log_2 -transformed RPKM values to obtain the relative deviation as well as on raw RPKM values to obtain the absolute deviation of gene expression. Selected genes showed a relative PAD > 0.8 and an absolute PAD > 1. The absolute PAD was used to avoid the selection of lowly expressed genes. The PAD was chosen over simple standard deviation since a more robust measurement of deviation was needed because the cohort was heterogeneous and an unbalanced number of tumours per subgroup between 15 and 63 MB samples. The 85-th percentile was chosen to also catch expression deviation related to the smallest subgroup in consideration of the two-sided properties of the PAD. Before applying NMF, RPKM values per gene were \log_2 -transformed RPKM values adding a pseudo-value of 0.5 and scaled (a mean = 0; standard deviation = 0), and scaled values < 0 were set to 0.

5.4.4.2 Molecular subcluster identification within subgroups

Molecular subclusters within the subgroups SHH, Group 3, and Group 4 were *de novo* identified using an unsupervised consensus clustering approach with subsequent semi-supervised clustering; the number of WNT MBs (n=15) was too small for subclusters identification. This consensus approach included the calculation of a consensus distance matrix of pair-wise distances between MB samples. This consensus distance matrix formed the basis for agglomerative hierarchical clustering of the samples. The resulting dendrogram indicated the final subcluster of each sample.

Subcluster identification was done subgroup-wise for each subgroup independently. Tumour samples were clustered based on the most variable genes selected by the PAD, as explained above (Section 5.4.4.1) taking the 75-th percentile; genes with a relative PAD > 0.7 and an absolute PAD > 1 were selected (SHH: 3288 genes; Group 3: 4193 genes; Group 4: 2633 genes). RPKM values were \log_2 -transformed adding a pseudo-value of 0.5. The consensus distance matrix was defined by the average distance between sample pairs across subsamples of the most variable genes (genes are equal to features and samples to entities). In 1000 iterations, only 90% of the originally selected genes were used for the calculation of pair-wise distances using Pearson correlation coefficients (PCC; pair-wise distance = $1 - \text{PCC}$). The consensus matrix was obtained by averaging across the 1000 iterations. Agglomerative hierarchical clustering with average linkage was applied to the consensus matrix and samples were assigned to a subcluster based on the obtained dendrogram. Subsequently followed the semi-supervised clustering of the samples. This clustering step was semi-supervised because genes for the clustering were selected based on differential expression between identified subclusters within a subgroup (differential gene expression analysis is explained in the next Section 5.4.4.3). Here, a common distance matrix based on PCC and agglomerative hierarchical clustering with average linkage was used for the clustering of the samples. The resulting dendrogram was used to assign samples to their final subclusters.

5.4.4.3 Differential gene expression analysis

Differential gene expression analyses were performed using the R package edgeR by fitting a GLM and applying an LRT (see Section 3.3.2). For the GLM, subgroup or subcluster assignments per tumour sample were used as categorical predictors coded as $n - 1$ dummy variables, where n is the number of subgroups/subclusters since one subgroup/subcluster is defined as the intercept. Differential expression between subgroups or subcluster within a subgroup was tested using a LRT by comparing the null model M_0 against the full model M_1 :

$$M_0 = \theta_0 \tag{5.3}$$

$$M_1 = \theta_1, \tag{5.4}$$

where $\theta_0 \in \Theta_0$, $\theta_1 \in \Theta_1$, and $\Theta_0 \subset \Theta_1$. The parameter set Θ_0 of the null model M_0 contains the intercept of the fitted GLM. The parameter set Θ_1 of the full model M_1 contains the intercept and $n - 1$ dummy variables of subgroup/subcluster assignments. Via the LRT is gene-wise tested if the full set of subgroups/subclusters explains significantly better the observed read count distribution across the data set than the intercept alone.

Coding and lnc genes showing ≥ 1 RPKM in at least six tumours among the analysed sample set were considered for differential gene expression analysis. For the detection of differential expression between subclusters within a subgroup, only samples belonging to the respective subgroup were used for the analysis. Obtained p-values were corrected for multiple testing using the Benjamin-Hochberg (BH) procedure. Genes were at first selected based on FDR: $\text{FDR} \leq 0.01$ for subcluster comparison and $\text{FDR} \leq 0.001$ for subgroup comparison. The FDR only indicates whether gene expression depends on subgroups/subclusters but does not indicate in which subgroup/subcluster a gene is up- or downregulated. Therefore, fold change values between compared sample groups were used to identify subgroups/subcluster-specific up- or downregulation. A gene was considered as subgroup-/subcluster-specific upregulation when a particular subgroup/subcluster showed a fold change ≥ 2 against the remaining subgroups/subclusters allowing the exception that the fold change against one of the remaining subgroups/subclusters was ≥ 1.75 . This exception was used to prevent missing relevant differential expression due to the high number of tested sample groups. Additionally, some subgroups share some similarities such as Group 3 and Group 4 (Section 5.1.6). The same scheme was

used for subgroup-/subcluster-specific downregulation using inverse fold change values. Additionally, subgroup-/subcluster-specifically expressed genes needed to show an absolute difference ≥ 1 RPKM between compared sample groups.

5.4.4.4 Gene set overrepresentation analysis

Gene set overrepresentation analyses were performed using the Genomatix Pathway System (Genomatix GmbH, Munich) and other gene function annotation databases in combination with a hypergeometric test (see Section 3.3.3). GO term annotations for genes were downloaded from Ensembl [354], and pathway annotations of the databases PID, KEGG, Reactome, and WikiPathways were downloaded from Consensus Path DB [344]. 15028 expressed coding genes were defined as background for the hypergeometric test. Obtained p-values were corrected for multiple testing using BH procedure. GO terms and pathways were considered as significantly enriched when they displayed an $FDR \leq 0.05$ and minimum overlap of three genes with the analysed gene set. Analysed gene sets comprised subgroup-/subcluster-specifically upregulated coding genes as well as coding genes that were significantly correlated with lnc genes expression, as later described in Section 5.4.6.10.

5.4.5 Gene regulatory networks inference

5.4.5.1 Annotation of transcription factors

TFs were annotated using a set of selected 34 GO terms, as previously described [123], and the Genomatix Genome Analyzer (Genomatix GmbH, Munich) that provides a database of transcription factors. The selected GO terms indicated nucleic acid or DNA binding, and transcriptional regulation. Genes were selected based on GO terms when a single annotated GO term indicated the functions nucleic acid/DNA binding and transcriptional regulation or when both functions were indicated by at least two annotated GO terms. After manual inspection, 15 genes were removed from the TF list because the literature did not provide strong evidence of TF functions.

5.4.5.2 Implementation of GENIE3

Gene regulatory networks were inferred from gene expression data applying the in Section 3.5.1 introduced algorithm GENIE3 that provides weights of regulatory links between regulators and putative targets [120]. An R implementation of GENIE3 (downloaded from <http://www.montefiore.ulg.ac.be/huynhthu/GENIE3.html>) was run using default parameters, \log_2 normalised RPKM values, and without pre-defining regulatory genes. In the resulting network, the set of defined TFs, as described in the previous section, was used as regulators. Since GENIE3 does not provide information about the nature of the regulatory relationship, Spearman's rho was calculated to assign negative and positive regulation between pairs of TFs and analysed genes. However, in the later analysed GRNs only links associated with positive regulation were considered.

In separate runs, four gene regulatory networks were inferred that corresponded to the subgroups and the subclusters in SHH, Group 3, and Group 4. Two different schemes were performed to construct the GRN for the subgroups and subclusters, respectively. The GRN representing subgroups was constructed running GENIE3 on the union between the differentially expressed genes and the most 8078 variable genes across the cohort (overall 8504 genes). The most variable genes were selected as described in Section 5.4.4.1 using PAD with a 85-th percentile; genes were expressed in minimum ten MB samples with ≥ 0.5 RPKM. The calculated weights of regulatory links were directly used for further processing, and only genes differentially expressed between subgroups were considered for the GRN.

To infer the regulatory networks representing the subclusters in SHH, Group 3 or Group 4, the following steps were separately applied to each subgroup. Weights of regulatory links were calculated by taking cohort-wide and intra-subgroup effects into account. Intra-subgroup weights were calculated applying GENIE3 to MB samples of the considered subgroup. Here, analysed genes comprised the most

variable genes within the considered subgroup and genes that were differentially expressed between subclusters of this subgroup. The most variable genes within a subgroup were selected taking the 75th percentile of the PAD; genes were expressed in minimum five samples with ≥ 0.5 RPKM (SHH: 3241 genes, Group 3: 4229 genes, Group 4: 2519 genes). Cohort-wide weights were calculated integrating all MB samples and the union between genes that were used for subgroup GRN and intra-subgroup weights inference. The final weights of regulatory links were achieved by taking the geometric mean between the cohort-wide and intra-subgroup weights of a certain link. Only differentially expressed genes between subclusters were considered in the final subcluster GRN per subgroup.

5.4.5.3 GRN fitting score

After the calculation of regulatory weights via GENIE3, a threshold for regulatory links with potentially relevant weights was evaluated because the weights allow a ranking of regulatory links. However, the value range of the weights is context-dependent and the weights do not have a statistical meaning in terms of significance [120]. The authors of GENIE3 did not provide a procedure to obtain a threshold for relevant weights. An own-developed procedure was applied for threshold determination using a GRN fitting score that was based on the enrichment predicted transcription factor binding site (TFBS) in promoter regions of putative transcription factors targets as well as the density of the network.

The enrichment of TFBSs was calculated as follows. A promoter region was defined by 2000 bp upstream and 50 bp downstream of a transcription start site that was annotated by Ensembl; overlapping alternative promoters of a gene were merged. TFBSs were predicted by applying the MatInspector tool (Genomatix GmbH, Munich) that uses position weight matrices and optimised thresholds [363]. For each TF with available position weight matrices, an enrichment of TFBSs in promoters of putative targets was evaluated by calculating a z-score for the observed TFBS frequency. Here, the observed frequency of TFBSs in target promoters was compared to the empirical distribution of expected TFBS frequencies in promoter or intergenic regions, similar to a previously published approach [364]; intergenic regions had a defined length of 2050 bps. The TFBS frequency b was given as TFBSs per Mb and defined by $b = s \cdot 1000000 / n$, where s is the number of TFBSs and n is the number of examined nucleotides. Let b_{OBS} and b_{EXP} denote the observed and expected frequency of TFBSs, respectively. The number of TFBSs is normalised by the number of examined nucleotides to address different promoter lengths introduced by merged overlapping promoters. The promoter background included 54078 promoters of expressed genes (RPKM ≥ 0.5 in minimum 5 samples), and the intergenic background comprised around 50000 regions. Taking a TF targeting k genes with m promoters and b_{OBS} observed TFBS frequency, the empirical distributions of TFBS frequencies b_{EXP} was achieved by drawing m regions from the promoter or intergenic background 2000 times. The mean and standard deviation of b_{EXP} distributions was used to calculate the z-score of b_{OBS} representing a measure of TFBS enrichment.

The z-score of TFBSs formed the basis for the threshold determination. Here, GRN fitting scores were calculated at different weight cutoffs, and the best GRN fitting score was chosen. For each GRN, 10-15 different cutoffs were evaluated taken at a chosen set of i -th percentiles of the weight matrix returned by GENIE3. This matrix included all tested genes and not only differentially expressed genes. For an evaluated weight cutoff, only TF-target links were considered displaying a weight higher than the cutoffs and positive regulation; TFs and genes needed to be differentially expressed between subgroups/subclusters of the analysed GRN. The lowest tested percentile was individually chosen per inferred GRN and represented a weight cutoff where the TF-target link list included between four and five TFs per target on average. The mean of TFBS frequency z-scores across TFs was calculated for the selected TF-target links. The GRN fitting score was defined by the ratio between the mean of TFBS frequency z-scores (z) and the network density (d) at the evaluated cutoff:

$$\text{GRN fitting score} = \frac{z}{d}. \quad (5.5)$$

The network density d was defined by the ratio between the number of selected links l and the maximum number of links r in a fully connected network:

$$d = \frac{l}{r} \quad (5.6)$$

$$r = N_{TF} \cdot N_{TAR}, \quad (5.7)$$

where N_{TF} is the number of TFs and N_{TAR} is the number of targets in the selected list of TF-targets links. The network density was used here as penalty of the TFBS enrichment because GRNs have the characteristics to be sparse [121] and, therefore, the GRN fitting score was defined in favour of a high sparsity (low density). The TFBS enrichment can be seen as the overall performance of the TF-gene interaction prediction. The 0.99725th, 0.985th, 0.989th and 0.9835th percentile were taken as the optimal threshold for the GRN of the main groups, and SHH, Group 3 and Group 4 sub-clusters, respectively. The final list of TF-targets links contained only links with a significant Spearman's rho (FDR ≥ 0.05).

5.4.5.4 Module detection in GRN

After the inference was the GRNs using GENIE3, the topology of the GRNs was analysed by detecting modules that provide biologically relevant information like co-expression/-regulation and functional associations, as summarised in Section 3.5.2. Modules were detected applying the map equation algorithm for each of the four GRNs; the map equation algorithm detects hierarchical modules in networks using random walks, as described in [125]. The map equation algorithm had been implemented as part of the Infomap software by the original authors and was downloaded from www.mapequation.org (v0.18.1) [365]. The infomap tool was run considering the networks as weighted undirected and using following parameters: -i 'pajek' -N 100 -clu -tree -bftree -map -to-nodes -flow-network -node-ranks -seed 634711. The GENIE3-based interaction weights between TFs and putative targets were also provided as link weights to the Infomap tool. The module detection was also necessary for the calculation of the network influence score.

5.4.5.5 Network influence score of transcriptions factors

In order to allow the identification of TFs that have potentially the most influence on subgroup-/subcluster-specific gene expression in MB, a previously described network influence score (NIS), since it allows to rank TFs by their influence on gene expression for individual cell types, summarised in Section 3.5.2 [123]. In the presented MB study, cell types are represented by subgroups and subclusters that also relate to specific gene expression. The NIS was calculated per TF and a certain subgroup/subcluster as follows:

$$NIS_{S_i}(r_i) = \sum_{j=1}^N z(t_j) + N \cdot z(r_i) \quad , \quad t \in O_k \wedge t \in L_{r_i} \wedge r \in O_k, \quad (5.8)$$

where r_i is the i -th TF (regulator), t_j is the j -th target of TF r_i , L_{r_i} is the target set of TF r_i , N is the number of targets of TF r_i , O_k is the set of genes belonging to the k -th detected module of the GRN, the z function is the average of expression z-scores among the tumour sample set S_i for a TF or target, S_i represents samples of a certain subgroup or subcluster. The limitation that both the TF and the target need to be part of the same module is also part of the original definition of the NIS [123]. This allows the integration of the network topology in the calculation of the NIS. In the original definition of the NIS, $z(t_i)$ and $z(r_i)$ were additionally weighted by the mean expression of the target/regulator since genes can display different functional expression ranges, as later discussed [123]. The calculation of the z-scores was also adapted to the analysed MB cohort [123]. For GRN related to

the subgroups, the whole MB cohort was considered for the z-score normalisation of gene expression values per gene. For each of the three GRNs that related to subclusters within a certain subgroup, only samples of the respective subgroup were considered z-score normalisation (e.g. for the GRN of the SHH subclusters only SHH tumours were considered). In order to adjust for differences in sample sizes between subgroups or subclusters, z-scores were calculated taking the weighted mean and weighted standard deviation where each subgroup or subcluster had equal weight.

5.4.5.6 Network visualisation

The inferred GRNs were visualised using Cytoscape (v2.8.3) using the AllegroLayout plugin app. Networks were visualised as a directed network and the node size related to the outdegree of genes; since only TFs represent nodes with outgoing links in the GRN, enlarged nodes always represented TFs in the visualised network.

5.4.5.7 Evaluation of the GRN validity

ChIP-seq data of HLX, LMX1A, LHX2, NEUROD1, and OTX2 — as described in Section 5.4.3.2 — were used for evaluation of the GRN validity. A Gene was assassinated as a putative target when a ChIP-seq peak overlapped a promoter region of the gene or an enhancer that regulates the gene. Promoters covered 2000 bps upstream and 50 bps downstream of transcription start sites. Enhancer-gene assignments were taken from Lin *et al.* [218]. The four GRNs were not individually evaluated but as a whole by taking the union of all inferred TF-target links. For the evaluation, the fraction of targets validated by ChIP-seq was considered and a hypergeometric test was applied to check if the observed number of validated targets could be obtained by chance compared to the number of expressed genes associated with ChIP-seq peaks; overall, 23222 genes were expressed in the MB cohort.

5.4.6 Characterisation of lnc coding genes

5.4.6.1 Considered lnc genes

The set of genome-wide lnc genes was selected based on Ensembl v70. lnc genes were annotated via the gene biotypes antisense, lincRNA, and processed transcript. Additionally, lnc genes that overlapped exon regions of coding genes on the sense strand were excluded. The basic set of lnc genes comprised 11207 genes. This basic set was tested for differential expression together with coding genes, as described in Section 5.4.4.3. The set of lnc genes that were differentially expressed between subgroups or subclusters was further characterised.

5.4.6.2 Annotation of lnc gene types and coding partners

The three lnc gene types divergent, antisense, and intergenic — as introduced in Section 2.5 — were annotated integrating FANTOM CAT (Section 5.4.3.4). In the case that a lnc gene was in divergent orientation to one coding gene/pseudogene and in antisense orientation to another coding gene/pseudogene, this lnc gene was generally annotated as divergent gene. The authors of FANTOM CAT classified lnc genes as divergent when its strongest transcription start site was within ± 2 kb of any start site of a lnc gene and a coding gene/pseudogene on the opposite strand [133]. For the annotation of antisense lnc genes, classifications of FANTOM CAT and Ensembl were combined since the FANTOM CAT definition of antisense gene was too strict. In FANTOM CAT, a lnc gene was classified as an antisense gene when at least 50% of its gene body overlapped with a gene body of a coding gene/pseudogene on the opposite strand [133]. In Ensembl: “Antisense RNAs: Locus that has at least one transcript that intersects any exon of a protein-coding locus on the opposite strand, or published evidence of antisense regulation of a coding gene.” [132]. When a lnc gene was annotated as antisense by Ensembl but as intergenic by

FANTOM CAT, the lnc gene was assigned to the antisense type. Annotations for divergent lnc genes were only based on FANTOM CAT; at this point, Ensembl did not provide this annotation [132]. The remaining lnc genes belonged to the type intergenic.

The annotation of lnc gene types also allowed to define coding gene partners. Coding gene partners of antisense and divergent lnc genes were in the respective orientation. An antisense gene could be in antisense orientation to more than one coding gene, where all of these coding genes were considered as coding partners. A divergent gene could be in divergent orientation to a coding gene but also in antisense orientation to an additional coding gene; both, divergent and antisense coding genes were considered as coding partners for divergent lnc genes. The coding partners of intergenic lnc genes included coding genes within ± 1 Mb and the same topological associating domain (TAD) that was shared between the lnc and coding gene. TAD boundaries were downloaded for human dorsolateral prefrontal cortex from the 3D Genome Browser since data for cerebellum were not available [366].

5.4.6.3 Expression correlation with coding partners

The correlation between lnc genes and their coding partners was measured using Spearman correlation in order to analyse the transcriptional dependencies. Additionally, lnc genes were classified based on their correlation with coding partners into the classes positively correlated, negatively correlated, non-correlated, and coding partner not expressed. Positive correlation was defined by Spearman correlation coefficients $\rho \geq 0.3$ and negative correlation by $\rho \leq -0.3$ (FDR < 0.05), in relation to potentially relevant correlation. None correlation related to $-0.3 > \rho > 0.3$. A lnc gene was classified as

- "positively correlated" when at least one coding partner was positively correlated and none of the coding partners negatively correlated,
- "negatively correlated" when at least one coding partner was negatively correlated,
- "non-correlated" when all coding partners did not display positive or negative correlations, and
- "coding partner not expressed" when all coding partners were not expressed.

For negatively correlated and non-correlated lnc genes, all coding genes within ± 100 kb were considered for the classification to ensure that these lnc genes did not show positive correlation with the direct coding gene neighbourhood because it was from interest to identify lnc genes that are not co-expressed with coding gene partners/neighbourhood.

5.4.6.4 Annotation of tissue-/cell-type-specific gene expression

For the annotation of tissue-/cell-type-specific gene expression, FANTOM CAT and the BrainSpan data set were used. Here, expression patterns related to development were of interest due to the embryonic origin of MB. The annotation of tissue-/cell-type-specific enriched expression by FANTOM CAT was integrated to identify lnc genes that are specifically/enriched expressed in embryonic and neural stem cells. The BrainSpan data set was used to identify lnc gene that showed upregulation in pre- or postnatal cerebellum/brain by comparing pre- and postnatal tissue samples (analyses was individually done for cerebellum and whole brain) as well as up- or downregulation in cerebellum vs. the remaining brain regions. Gene expression data of the BrainSpan data set are only provided as RPKM values and, therefore, read count-based analyses using edgeR, as in Section 5.4.4.3, could not be done. Significant differential expression was detected using the Wilcoxon rank-sum test. Genes expressed by minimum 0.2 RPKM in at least three analysed samples were tested. A gene was significantly differentially expressed between tested groups when it showed an absolute mean \log_2 fold change > 1 and an FDR < 0.05 (BH procedure); a pseudo-value of 0.5 was added to RPKM values for \log_2 transformation.

5.4.6.5 Annotation of lnc gene associated publications

In order to facilitate a comprehensive annotation of publications that relate to the analysed set of lnc genes, the database gene2pubmed was downloaded from NCBI (February 20 2019, <ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz>) [367]. This database uses Entrez gene IDs. The mapping of Ensembl and Entrez gene IDs was downloaded from the Ensembl website using the BioMart online tool [368]. One Ensembl ID could map to several Entrez IDs. Lnc genes with at least seven publications were screened for implications in MB or other cancer types as well as comprehensive functional characterisations of these lnc genes. Based on this screening, twelve potentially interesting lnc genes were selected for further characterisation.

5.4.6.6 Co-expression clustering

Lnc and coding genes were clustered into co-expression clusters applying the algorithm CLICK (CLuster Identification via Connectivity Kernels) that is part of the software tool EXPANDER (v7.11) [369]. The most variable genes among subgroups and subclusters (Section 5.4.4.2) as well as all differentially expressed genes (Section 5.4.4.3) were clustered (overall 10132). RPKM values were \log_2 transformed and z-score normalised. The CLICK parameter expected homogeneity was tested between the values 0.5 and 0.8 in 0.01 steps. Based on the average separation between and the average homogeneity within co-expression clusters, the value 0.59 was chosen as the best solution. Overall, 29 clusters were identified comprising 9544 genes. Identified co-expression clusters were further analysed when lnc genes of interest were part of a cluster.

5.4.6.7 Gene expression-based survival analysis

In order to identify lnc genes that displayed a significant association between gene expression and overall survival, tumour samples were split into two groups that represented high and low expression of an analysed lnc gene using a defined cutoff. A log-rank test (Section 3.3.4) was applied, testing for significant differences in survival between the groups. Simply using the cutoff with lowest p-value could lead to an overestimation of the effect of the prognostic factor [97]. In order to avoid overfitting, a subsampling approach was used to find an optimised robust expression cutoff. For each gene, expression cutoffs were evaluated between the 25% and 75% percentile in 1% steps. Each cutoff was evaluated based on the signal-to-noise ratio of test statistic values (chi-square) that were obtained from log-rank tests across 1000 cohort subsamples taking 50% of the available samples per subsample. The cutoff (percentile) showing the highest signal-to-noise ratio was selected as an optimised cutoff. The final evaluation of the expression of a lnc gene as a potential prognostic marker was based on the p-value of the log-rank test using the optimised cutoff on the whole cohort ($p < 0.05$). The application of the signal-to-noise that was obtained across subsamples of the data set refers to the MSVM-RFE approach for feature selection (see Section 3.4.2.1). The signal-to-noise ratio provides the information whether a cutoff relates to a meaningful split of patients (differences in survival) that is stable across many subsamples.

Lnc and coding genes were tested for association with survival using the ICGC MB cohort at first. Overall survival data were available for 155/164 MB samples. The average follow-up time of patients was ~2 years. 16 patients with a follow up > 5 years were excluded from the analysis since none of these patients experienced death and the general low follow-up time leaving 139 samples for the analysis. The union of the most variable genes across the cohort and differentially expressed genes between subgroups/subclusters, which were expressed in at least 25% of the sample by 0.5 RPKM, were tested (8182 genes). Here, 82 lnc genes showed a significant association with overall survival (p-values < 0.05). Here, only the p-value was used and not the FDR to define significance since a subsequent validation procedure was applied using an independent cohort in order to ensure the identification of potentially true prognostic markers [370].

The subsequent validation was done on the microarray cohort published by Cavalli *et al.* [221]. External RNA-seq MB cohorts were not available. Survival data were available for 612 patients overall. The average follow-up time was ~5 years. Survival data were right-censored at ten years because only 10% of the patients exceeded a follow-up time of 10 years. Among the 82 lnc genes that showed significant association with survival using the ICGC cohort, eight lnc genes that could be mapped to the microarray probes showed sufficient expression on the microarray. Lnc genes that showed an \log_2 expression intensity > 5.2 (comparable to 0.5 RPKM [65]) in at least 25% of the MB samples were analysed. Optimised expression cutoffs were evaluated as for the ICGC MB cohort.

Time-dependent survival of *MEG3* high- and low-expressing MB samples was visualised using Kaplan-Meier curves plotted with the `km.coxph.plot` function of the `survcomp` R package (v1.20.0) [371]. Differences in survival between sample groups were evaluated using Cox regression-based hazard ratio and p-value [96].

5.4.6.8 *MEG3* binding motif and binding site prediction

Data published by Mondal *et al.* were used to obtain a PWM of the *MEG3* DNA binding motif [43]. The authors have provided a list of 532 *MEG3* ChOP-seq peaks (Section 2.5) associated with genes that were deregulated after *MEG3* knockdown [43]. The sequences of these 532 peaks were used to obtain a PWM using the algorithm CoreSearch that is part of the Genomatic Genome Analyser (Genomatic GmbH, Munich) [372]. The following parameters were changed from the default: `-MOTIF=2`, `-MATSIM=0.85`, `-CORE=8`. Setting the parameter `MOTIF=2` allows that a found motif can match to several positions per sequence, which was critical since the *MEG3* binding motif is composed of GA-repeats [372]. The parameter `MATSIM` defines the minimum similarity of the found motif across sequences that were used to build the PWM. This parameter was set from 80% (default) to 85% because 85% provided a shorter motif with 25 bps (33 bps with 80%) that was closer to reports of Mondal *et al.*. CoreSearch initially identifies a core sequence of a motif that is used to obtain the final motif. The length of this core sequence is related to the parameter `-CORE` and was set from 7 to 8 due to the GA-repeats of the *MEG3* binding motif.

The obtained PWM was used to predict *MEG3* binding-sides in promoter and enhancer regions. *MEG3* binding-sides were assigned to genes via associated promoter and enhancer regions. Enhancer regions and enhancer-gene associations in MB were taken from Lin *et al.* [218].

5.4.6.9 *MEG3*-centred expression correlation analysis

Genes significantly correlated with *MEG3* expression were identified calculating Spearman correlation and using the MB RNA-seq cohort ($|\rho| > 0.3$; FDR < 0.05 , BH procedure). Here, expression correlation was measured across the cohort as well as across SHH, Group 3, or Group 4 samples resulting in four correlation gene sets; the WNT subgroup was too small for an individual analysis. Analysed genes were expressed with > 1 RPKM in at least five samples and were among the top 75% of expressed genes with the highest standard deviation. Negatively and positively correlated genes were independently considered in downstream analyses.

Visual inspection revealed spurious correlations within the correlation gene set that was obtained across the cohort; here, subgroup-dependent expression was a confounding factor. A heuristic approach was applied to identify spurious correlation. This approach was based on the assumption that non-spurious correlations should remain stable after removing one subgroup or at least within one subgroup. Through measuring correlation in different settings, the confounding effect of subgroup-dependent expression was systematically evaluated. The evaluation within each subgroup was essential to capture subgroup-dependent correlations that are potentially not detected after removing the subgroup associated with dependent correlation from the analysis. Subgroup-dependent correlations are expected due to the molecular heterogeneity of the subgroups. Among others, the correlation stability between gene pairs was evaluated by comparing the correlation coefficient measured across

the whole cohort to the coefficient measured in a different setting (e.g. removing one subgroup). The comparison of two correlation coefficients was based on Fisher's z-transformation (z-test) that can also be applied to Spearman and not only Pearson correlation coefficients [373]. A one-sided z-test was applied testing if the coefficient that was measured across the whole cohort was bigger than coefficient that was measured in a different setting using absolute coefficient. A correlation was defined as spurious when the correlation was not stable 1) after removing one subgroup and 2) within the removed subgroup. A non-stable correlation showed a significant drop of the coefficient in the tested setting (z-test, $p < 0.05$) and a p-value > 0.075 of the coefficient in the tested setting (related to non-relevant correlations). The second condition was used because a coefficient in the tested setting could be significantly smaller but still significant/relevant (for example, a drop from 0.8 to 0.6). A more tolerant significance level of 0.075 was used since the tested settings always comprised a smaller number of MB samples, compared to the whole cohort, resulting in a lower statistical power. Identified spurious correlations were excluded from the list of significantly correlated genes.

5.4.6.10 Gene set enrichment in *MEG3*-correlated coding genes

A gene set enrichment analysis was applied to identify significantly enriched functions among genes correlated with *MEG3* expression, which was similar to Section 5.4.4.4. Each of the four correlation gene sets, which related to the whole cohort, SHH, Group 3 or Group 4, were independently analysed. Here, positively and negatively correlated genes were separately analysed. The enrichment analysis was repeated for correlated genes that carried a *MEG3* binding site in a promoter or enhancer (Section 5.4.6.8).

5.4.6.11 DNA methylation analysis at the *MEG3* locus

Since DNA methylation changes in specific regions around the *MEG3* locus can regulate *MEG3* expression, as summarised in Section 2.5 (page 13), DNA methylation patterns for these regions were analysed in MB. DNA methylation data (Infinium HumanMethylation450 BeadChip array) for 155 MB samples that matched to the RNA-seq cohort were provided by collaboration partners of ICGC PedBrain [222]. Data were processed and normalised using the minfi R package (v1.14.0) [374]. Between-sample normalisation was done using the preprocessFunnorm function of minfi and potential batch effects were removed using ComBat of the sva R package (v3.18.0) [359, 375].

The IG-DMR was not covered by the array. The *MEG3*-DMR was covered by 33 probes that spanned chr14:101290556-101293856. Coordinates of *MEG3*-associated DMRs were obtained from [376]. Average DNA methylation levels (based on beta value) for the *MEG3*-DMR were calculated across the 33 covered probes for each MB sample. An F-Test was performed using the dmpFinder function of minfi to test for significantly different *MEG3*-DMR methylation levels between subgroups.

5.5 Discussion

We presented a MB transcriptome study that was based on a deeply sequenced RNA-seq cohort of 164 tumours, comprising the expected fractions of the MB subgroups WNT, SHH, Group 3, and Group 4 [83] (Section 5.3.1.1). Our MB transcriptome study concentrated on three aspects.

First, we explored the molecular heterogeneity beyond the four subgroups by identifying nine molecular subclusters within the subgroups SHH, Group 3, and Group 4. Here, each subgroup was split into three subclusters. These subclusters were validated by and related to recently published MB subtypes (Section 5.3.1.2).

Second, the consensus subgroups and the defined subclusters provided the basis for deeper analyses of the molecular heterogeneity in MB. We identified differentially expressed lnc and coding

genes among subgroups and among subclusters of one subgroup (Section 5.3.2.1). These specifically-expressed genes were used to understand the subgroups and subclusters in MB by the identification of functional enrichments and underlying gene regulatory networks (Section 5.3.2.2, 5.3.2). We inferred four GRNs from MB gene expression data. These GRNs depict the landscape of transcription factors that have a high impact on specific gene expression in MB subgroups and subclusters (Section 5.3.2). The inferred GRNs allowed to portray specific and common gene regulatory programs between subgroups and subclusters.

Third, we computationally characterised lnc genes that we detected as differentially expressed among MB subgroups and subclusters to gain insights into functional roles of lnc genes in the MB (Section 5.3.3). At first, we achieved a systematic characterisation of lnc genes by two forms of classification. The first classification was based on the genomic position relative to coding genes. The second classification utilised the expression correlation with neighbouring coding genes to derive different lnc gene categories. We performed DGEAs to identify lnc genes that show brain/cerebellum development-associated expression patterns, which highlighted potentially disease-relevant candidates. Twelve lnc genes that have been frequently described in cancer were re-evaluated in the context of MB. Here, several regulatory links that were reported in other contexts could be validated in MB. We identified the co-expression cluster Cc1 that contained several reported cancer-associated lnc genes. Cc1 was associated with protein biogenesis and protein signalling profiles regulated by MYC family genes. Additionally, survival analyses and co-expression-based functional inference highlighted *MEG3* as potential tumour suppressor in MB and prognostic marker in subclusters of SHH and Group 4.

5.5.1 Molecular subgroup and subcluster identification using RNA-seq

The dissection of the molecular complexity of MB has been a central part of understanding this disease. Here, in previous studies, mainly 450k methylation arrays and expression microarrays were used to identify the four main subgroups [193, 216, 217]. Therefore, we evaluated the capability of RNA-seq to identify the four subgroups. Thereby, we verified that RNA-seq is suitable for this task since the RNA-seq-based MB sample clustering and the 450k array-based sample classification showed a high agreement regarding the identification of the four subgroups. (Figure 5.3 in Section 5.3.1.1).

Seeking a deeper understanding of the molecular heterogeneity in MB beyond the four consensus subgroups, we identified and explored three molecular subclusters within each of the main subgroups SHH, Group 3, and Group 4. These subclusters were strongly supported by external subtypes that were published by Cavalli *et al.* parallel to the work on our study (Section 5.3.1.2) [221]. Additionally, the sets of subcluster-specifically expressed genes that we identified facilitated the classification of the external MB samples (Cavalli *et al.*) into the subclusters. In this way, we could prove that these subcluster-associated genes are stable across cohorts and technologies since Cavalli *et al.* had used microarrays [221]. Besides the validation of the subclusters, by comparing the subclusters to the subtypes of Cavalli *et al.*, we could relate the subclusters to detailed clinical features and somatic copy number aberrations. Our analysed RNA-seq cohort was too small to obtain such detailed information when considering the high number of molecular subclusters and the short follow-up time of patients in the RNA-seq cohort. Cavalli *et al.* had integrated 763 tumour samples with a sufficiently long follow-up into their study [221].

Our described subclusters were also compared to subtypes of Group 3 and Group 4 MBs published by Northcott *et al.* [222]. These published subtypes of Group 3 and Group 4 MBs showed a robust agreement with our subclusters, but subtypes published by Cavalli *et al.* showed a better agreement with the subclusters (Section 5.3.1.2). For example, our subcluster Grp3-c2 matched with subtype Group 3 β by Cavalli *et al.* but did not match with any subtype by Northcott *et al.*. Cavalli *et al.* have reported that the split of the subtypes Group 3 β and Group 3 γ has been mainly supported by expression data [221]. Here, it is important to note that Northcott *et al.* had used only array-based DNA methylation data for subtype identification, whereas Cavalli *et al.* had used array-based methylation

and gene expression data [221, 222]. Our RNA-seq-based subclusters support the reports of Cavalli *et al.* that certain MB subtypes can be only identified using gene expression because subcluster Grp3-c2 did not agree with any DNA methylation-based subtype of Northcott *et al.*

Additional differences between our subclusters and published subtypes include SHH-c1 vs. Cavalli *et al.* as well as Grp3-c3 and Grp4-c2 vs. Northcott *et al.* (Section 5.3.1.2). Here, one subcluster corresponded to two reported subtypes. The RNA-seq cohort of our study that was used to identify the MB subclusters has a considerably smaller size ($n=164$) than the two external cohorts used for subtype identification (over 700 MB samples). This might explain why we did not observe a further split of the subclusters SHH-c1, Grp3-c3, and Grp4-c2. However, considering that our RNA-seq cohort is ~4.5 times smaller than the external cohorts and the overall good agreement with subtypes of Cavalli *et al.*, RNA-seq provides a sensitive platform for the identification of molecular subtypes/subclusters within the four main subgroups.

We applied an unsupervised consensus clustering approach and a subsequent semi-supervised clustering to identify the presented subclusters subgroup-wisely. The second step was semi-supervised because it integrated genes that were differentially expressed between the subclusters within one subgroup. These differentially expressed genes represented additional information that supported the clustering. Therefore, the semi-supervised clustering was favoured over the unsupervised clustering for the final cluster assignments. Due to the integration of additional information, the semi-supervised clustering should allow a correct cluster assignment even for samples that are difficult to cluster. This assumption was proven by five Group 4 samples that could not be assigned to a cluster by the unsupervised clustering. The consensus clustering approach — based on a consensus distance matrix defined by the average pair-wise dissimilarity of samples across subsamples of genes — that we applied differs from the classical version of consensus clustering [377]. For classical consensus clustering, a consensus matrix is used that contains the pair-wise consensus rate between samples. The consensus rate is defined by the frequency of two samples being assigned to a common cluster across repeated clustering runs on random subsamples of analysed data [377]. A direct comparison of our applied consensus distance matrix with the classical consensus clustering approach would be of interest to evaluate the performance of the consensus distance matrix procedure (e.g. based on samples that were difficult to cluster by the unsupervised clustering). Even though the performance of the consensus distance matrix procedure was not directly evaluated, the subclusters that we found were strongly supported by external data. These results underline that our applied consensus distance matrix clustering approach is suitable for identifying meaningful molecular subclusters; however, a second semi-supervised clustering step is necessary to correct potential misclassification.

5.5.2 Inference of gene regulatory networks in MB subgroups and subclusters

One focus of the transcriptome study presented above was to infer gene regulatory networks that underlie the subgroup- and subcluster-specific expression. We inferred the shown GRNs from gene expression data in order to take advantage of the deeply sequenced RNA-seq MB cohort. The networks were inferred using the ensemble machine learning algorithm GENIE3 that is based on a random forest of regression trees (see Section 3.5.1) [120]. Many different algorithms are available for the inference of GRNs from gene expression data. However, among these algorithms, GENIE3 has shown a good or the best performance in the DREAM4 and DREAM5 challenges [119, 120]. Even though GENIE3 is a well-performing method for inferring GRNs, it provides only a weight (score, criterion) for ranking TF-gene interaction — like many other GRN inference methods [119] — leaving the choice of a threshold for the weight that controls the number of false positive and false negative predictions to the user [378]. A number of methods have been proposed for selecting features based on importance scores (like interaction weights) from tree-based ensemble methods, such as GENIE3 [379]. However, many of these methods are computationally extensive because they require multiple runs of the

learning algorithm on permuted data [379, 380]. (In the case of GENIE3, this procedure needs to be repeated for each gene of the input data set.) Furthermore, these methods select a maximum number of features (TFs in GRNs) that show at least some predictive power on the target variable. This design of feature selection might not be the best choice in a GRN inferred from gene expression, considering regulatory cascades, as pointed out by Huynh-Thu and Geurts [379]. Such cascades (e.g. $TF_1 \rightarrow TF_2 \rightarrow \text{target}$) have the effect that TFs upstream in the cascade (like TF_1 in the example) are still predictive of the expression of the target gene. Due to these indirect effects in a GRN, selecting a maximum number of informative features would introduce many (false positive) indirect links. Therefore, these feature selection approaches for tree-based methods are not necessarily appropriate for gene-expression-based GRN inference [379].

Addressing the problem of determining a threshold for relevant TF-gene interaction weights, we proposed a GRN fitting score above. Here, we did not evaluate the interaction weights directly as done by the approaches that were designed for tree-based methods. We examined different thresholds by evaluating the resulting GRN. Here, we used the proposed GRN fitting score for the evaluation. This fitting score was based on the ratio between the average enrichment of predicted TFBSs and the network density. The average TFBS enrichment represents a performance measure of the interaction prediction and can be used alone to reduce false positives. An increasing TFBS enrichment would correlate with a decreasing number of false positives and an increasing number of true positives among predicted targets. However, it can be expected that the change of average TFBS enrichment over an increasing threshold will flatten above a certain threshold since the true positive predictions accumulate among higher interaction weights [381]. GRNs are sparse [121] because a gene is regulated by only one or a few TFs, and, therefore, network sparsity can be understood as one objective in the GRN reconstruction [382]. Thus, we used the network density as an additional indication for the false positive and false negative levels. Here, we assumed that a decreasing network density (increasing sparsity) is correlated with a decreasing number of false positives at the cost of a higher false negative level. Overall, the proposed GRN fitting score combines two objectives by using the network density as a penalty of the average TFBS enrichment to evaluate a threshold of interaction weights based on the inferred GRN.

Finding a cutoff for the interaction weights only filters the GENIE3 output, whereas hyperparameters and options of GENIE3 influence the algorithm directly. GENIE3 hyperparameters comprise the number of randomly drawn genes per regression tree (R) and the number of trees in the random forest (see Section 3.5.1) [120]. Defaults hyperparameter values were used ($R = \sqrt{N}$, $N =$ number of input genes; 1000 trees) that were suggested for the application of GENIE3 before [119, 120]. These hyperparameter values are commonly used for random forest [383]. The tuning of the GENIE3 hyperparameters was not addressed in the study above since the tuning step remains challenging for GENIE3, as already pointed out by the authors of this method [120]. An option that could be tested in the future is to apply a grid search that includes the two hyperparameters and the cutoff for the interaction weights using the introduced GRN fitting score as an evaluation metric. Besides the hyperparameters, GENIE3 also provides the option to predefine a set of known TFs to evaluate regulatory interactions [120]. However, this option should be used with caution when considering the algorithm (Section 3.5.1). GENIE3 obtains a ranking of regulatory interactions by a random forest of regression trees. Each tree node minimises a loss function by selecting gene j (putative regulator) and a sample split that minimises the expression variation of gene i (target) (see Section 3.5.1). Here, the capability to be a regulator of gene i is compared between input genes. When the set of regulators is predefined but the true regulator of gene i is not part of the predefined set, the regulatory interactions with gene i and the predefined regulators will probably be overrated, leading to false positive results. For this reason, we run GENIE3 considering all input genes as potential regulators since probably not all human TFs are known; additionally, non-TF genes can also influence gene expression.

We obtained four GRNs by applying GENIE3 to our RNA-seq MB cohort to identify the regulatory networks that underlie the expression profiles of subgroups and subclusters within the main subgroups

SHH, Group 3, and Group 4. By integrating published ChIP-seq data performed in MB of the TFs *HLX*, *LHX2*, *LMX1A*, *OTX2*, and *NEUROD1*, we could show a good validation rate of 81% for the overall 270 putative targets of these five TFs. These results indicate a good overall validity for the inferred GRNs in MB. In comparison, a published community method — which integrates numerous inference methods to obtain a community ranking of regulatory TF-gene interactions and performed as good as or even better than GENIE3 — showed a validation rate of ~70% [119]. Additional to the validation based on ChIP-seq, known aspects about gene regulation in MB also supported the robustness of our inferred GRNs (Section 5.3.2.3). However, the cutoffs that we determined for the interaction weights were rather strict and chosen in favour of a high network sparsity when we selected the best GNR fitting score, which was a conservative choice. This assured the inference of reliable GRNs, as indicated by the validation rate of 81%, but could miss potentially true regulatory interactions as the later discussed interaction between *MEG3* as regulator of *TGFBR1*. Nevertheless, robust GRNs based on strict cutoffs are preferred for genome-wide analyses considering that we could directly validate inferred regulatory interactions for only five of 339 TFs due to the limited number of ChIP-seq experiments performed in MB.

We used the inferred GRNs to identify TFs that potentially have the strongest impact on subgroup- and subcluster-dependent gene expression (Section 5.3.2). TFs were ranked based on a previously described NIS, which has been applied to identify the TFs with the strongest impact on cell/tissue type-specific gene expression in regulatory networks inferred from gene expression [123]. The performed MB study showed similarities to the original NIS publication. We inferred GRNs also from expression data, and the analysed MB cohort also comprised a collection of cell types represented by subgroups and subclusters. Therefore, the NIS appeared to fit the task to identify TFs with strong influence on subgroup- and subcluster-dependent gene expression in MB. We used an adapted version of the NIS above (Method Section 5.4.5.5). Here, we did not weight genes (targets and TFs) by their expression level since functional relevant genes can be expressed on individual expression levels. This can be exemplified by the two TFs *ATOH1* and *CRX* that showed a high NIS in the subgroup SHH and Group 3, respectively (Section 5.3.2.3). *ATOH1* showed an average expression level of 40 RPKM in SHH samples, whereas *CRX* was expressed at 12 RPKM on average in Group 3 samples.

Overall, the GRNs that we inferred from gene expression data included the majority of the detected differentially expressed genes (78%) in MB. The high validation rate by ChIP-seq data and the literature support strongly suggest that our inferred GRNs provided a reliable basis to identify TFs with a high impact on gene expression in MB and reveal new aspects of gene regulation in MB.

5.5.3 Expression profiles and gene regulatory networks in subgroup and subcluster

We overlaid the inferred GRNs and NIS-based TF ranking with functional enrichments that we detected among subgroup- and subcluster-specific upregulated genes. Comparing these analyses provided links between gene regulation and expression profiles that are characteristic for MB subgroup and subcluster. For example, top-ranked TFs that we found in WNT and SHH MBs reflect the respectively activated pathways in these subgroups (Section 5.3.2.3) [83]. However, the performed analyses revealed more details for the subgroups Group 3 and 4, their subclusters, and the SHH subclusters.

We observed that the expression profile of Group 4 MBs is characterised by a neuronal-developmental signature, as previously described [83]. Our performed TF ranking provides evidence that *ZBTB18* and *NEUROD2* probably strongly contribute to this neuronal-developmental signature in accordance with the reported function of both TFs in the development and differentiation of cerebellar neurons [253, 254]. In our data set, these two TFs showed a strong regulatory influence and specific upregulation in all Group 4 MB and in subcluster Grp3-c2 that expressed a Group 4-like profile. Moreover, the potential regulatory role of *ZBTB18* and *NEUROD2* in MB was supported by a functional enrichment for neuron

development among putative target genes that we inferred. Upregulation of *NEUROD2* (alias *NDRF*) and *ZBTB18* (alias *RP58*, *ZNF238*) is known for Group 4 MBs [218, 230]. However, a regulatory role of these two TFs has not been reported for MB or subgroup Group 4, including the work of Lin *et al.* [218].

Although the inferred GRNs strongly suggest both TFs as important regulators, the question arises what kind of functional role *ZBTB18* and *NEUROD2* might have in Group 4 and Grp3-c2 MBs. The expression of *ZBTB18* and *NEUROD2* in pre- and postnatal human cerebellum shown above was also reported in mice where differentiating and matured cerebellar cell types express both TFs [253, 254]. Since *ZBTB18* and *NEUROD2* expression is part of normal pre- and postnatal cerebellum, and only *ZBTB18* showed upregulation in Group 4 MB vs controls in our analysis (~2.2-fold), the question remains whether or not the expression of *ZBTB18* and *NEUROD2* has a cancer-relevant implication for Group 4/Grp3-c2 MBs or simply reflects transcriptional programs of cerebellar cell types that are conserved in the tumours. By comparing the expression profile of human bulk MB samples to single-cell RNA-seq data of the developing mouse cerebellum, Vladoiu *et al.* have recently shown that the expression profile of Group 4 MB is similar to that of unipolar brush cells at mouse embryonic stage E16 or E18 [314]. In the mouse cerebellum single-cell data of Vladoiu *et al.*, *Neurod2* and *Zbtb18* have been among the top differentially expressed genes that were upregulated in pre- and postnatal populations of unipolar brush cells, in granule cells, and other interneuron types. The data of Vladoiu *et al.* support a general role of *Neurod2* and *Zbtb18* in the differentiation of GABAergic and glutamatergic interneurons in the cerebellum, as recently shown [253, 314, 384]. Considering the work of Vladoiu *et al.* and the expression of *ZBTB18* and *NEUROD2* in pre- and postnatal human cerebellum, as shown above, *ZBTB18* and *NEUROD2* expression could indeed simply reflect an expression program of differentiating or differentiated cerebellar cells (potentially unipolar brush cells at E16 or E18) that is conserved in Group 4 MB. Nevertheless, previous publications have reported functions of *NeuroD2* and *Zbtb18* that could be relevant for Group 4 tumours [253, 254, 384, 385]. In the mouse cerebellum, *NeuroD2* promotes the postnatal survival of basket and stellate cells (molecular layer) and granule cells during early postnatal time points by upregulating genes that are potentially implicated in the survival of cerebellar neurons [254, 384]. *Zbtb18* promotes neuron survival in mouse neocortex at late prenatal (E16.5, E18.5) and early postnatally (P2, latest tested) stages but not in earlier prenatal stages (E14.5, earliest tested). Therefore, *Zbtb18* might also promote postnatal survival cerebellar cells such as *NeuroD2* [385]. A study of *Zbtb18* in mouse cerebellum evaluated the effect of *Zbtb18* knockout on neuron survival at E14.5 without seeing differences between wild-type and knockout. However, considering the study of the neocortex, E14.5 is probably too early to see potential effects of *Zbtb18* knockout on neuron survival in the cerebellum [253, 385]. Transferring these reports to MB, *ZBTB18* and *NEUROD2* can potentially promote cell survival and tumour maintenance in Group 4 MB. The capacity of TFs — that are associated with rather a cell-type-specific signature than obvious oncogenic processes in MB such as *ZBTB18* and *NEUROD2* — to maintain MB tumours has been recently shown for *NRL* in Group 3 MB by Garancher *et al.* [268]. However, at this point, additional experiments in MB are necessary in order to answer the question of whether or not reported survival-promoting functions of *ZBTB18* and *NEUROD2* in cerebellar/neocortical neurons could also play a role in Group 4 and Grp3-c2 MB maintenance. If this assumption is correct, other alterations cause the initiation of tumourigenesis and *ZBTB18* and *NEUROD2* would promote tumour maintenance. Since both TFs are well expressed in the normal cerebellum, it is probably unlikely that *ZBTB18* and *NEUROD2* carry an oncogenic function in MB like the TF *OTX2* [246].

Tatard *et al.* have reported that exogenous overexpression of *ZBTB18* inhibits tumour growth in an SHH MB cell line (DAOY) [386]. However, overexpression of *ZBTB18* probably has subgroup-dependent effects that differ between SHH and Group 4 MBs. Here, it should be considered that SHH MB resembles a granule progenitor and Group 4 MB a differentiating/mature cell type [314]. In the normal cerebellum, only granule progenitors react to Shh-signalling and proliferate; initiation of differentiation goes along with a post-mitotic stage [386]. Tatard *et al.* concluded that *Zbtb18* might

inhibit the SHH MB cell line by introducing cell differentiation blocking response to Shh-signalling [386]. Since Group 4 MB already reflects a differentiating/mature cell type and does not depend on Shh-signalling, *ZBTB18* most likely has different effects in subgroup Group 4 compared to SHH MB.

Besides Group 4 MB, the GRNs revealed new aspects about gene regulation in Group 3, Grp3-c3, and Grp4-c3 MB. The characteristic photoreceptor expression signature that we detected for Group 3 tumours agrees with previous reports [387]. However, this photoreceptor signature was also expressed in subcluster Grp3-c3 and Grp4-c3. The performed TF ranking highlighted the cooperating of TFs *NRL*, *CRX*, and *RAX2* in the regulation of the photoreceptor signature in Group 3, Grp3-c3 and Grp4-c3 MB. This regulatory role of *NRL*, *CRX*, and *RAX2* was supported by functional enrichments for photoreceptor development among putative targets of the three TFs in MB, as shown above, and the known regulatory functions of these TFs in photoreceptor differentiation [215, 242, 243]. The role of *NRL* and *CRX* as master regulators of the photoreceptor signature in MB was confirmed by a more recent study by Garancher *et al.* — published in parallel to the work on the present study. The authors also showed an upregulation of both TFs in Group 3 MB and subtypes of Group 3 and Group 4 [268] where these subtypes match to our MB subclusters Grp3-c3 and Grp4-c3. The authors did not mention *RAX2* in this context. Lin *et al.* have reported a TFBS enrichment for *RAX2* in subgroup specific-enhancers of Group 3 and Group 4 [218]. However, in our present study, *RAX2* was virtually not expressed in several Group 3 and Grp4-c3 samples, whereas *NRL* and *CRX* were expressed in all Group 3 MB except for three samples of the Group 4-like subcluster Grp3-c2. Considering the expression levels that we observed for *RAX2* in MB and the reported function of *Rax2* as a simple transcriptional modulator by enhancing the transactivating function of *Nrl* and *Crx* in mice [242], the maintenance of the photoreceptor signature in MB depends on *NRL* and *CRX* expression, but *RAX2* expression is dispensable.

Our performed TF ranking in subcluster Grp3-c1 suggests that the top-ranked TFs *MYC* and *HLX* mainly influence gene expression in Grp3-c1 MBs. The oncogenic activation of *MYC* in a subset of Group 3 tumours due to copy number gain/amplification and *PVT1-MYC* fusion is well described [193, 269]. Subcluster Grp3-c1 and matching previously reported subtypes (Group3 γ , II) reflect this subset of *MYC*-driven Group 3 MB [221, 222]. In contrast to *MYC*, oncogenic functions have not been reported for *HLX* in MB. However, information scattered across several publications may help to understand the observed expression association between *MYC* and *HLX* in Grp3-c1 MB. The expression correlation between *HLX* and *MYC* in bulk Group 3 MB that we observed above was also reported on single-cell level of MB, as published by Hovestadt *et al.* [388], underlining a strong expression association between both TFs in Group 3 MB. Two mechanisms might explain the expression association between *HLX* and *MYC* in Group 3 and the upregulation in Grp3-c1 MB, respectively. *HLX* is located on chromosome 1q, and it was reported that copy number gain of chromosome 1q leads to upregulation of *HLX* in MB [298]. Cavalli *et al.* described that subtype Group3 γ , matching subcluster Grp3-c1, is not only defined by recurrent chr. 8 (*MYC*) but also recurrent chr. 1q gain. Therefore, co-amplification/-gain of *MYC* and *HLX* in subcluster Grp3-c1 might be a possible mechanism causing the upregulation of both TFs in these tumours. Lin *et al.* have reported that *HLX* binds to an *MYC*-regulating enhancer that is active in Group 3 and WNT MB, which provides a potential mechanism for the expression correlation between *HLX* and *MYC* in Group 3 MB [218]. Moreover, both mechanisms might work together to enhance *MYC* expression. A deeper understanding of the regulatory role of *HLX* in *MYC*-driven Group 3 MB via functional experiments is needed since this subset shows the worst survival among Group 3 tumours [221].

The remaining two subclusters Grp4-c1 and Grp-c2 within Group 4 were not defined by photoreceptor-associated TFs as in subcluster Grp4-c3. The function of the TF *EBF1* that we found to be top-ranked in subcluster Grp4-c1 is not well described in MB. In terms of *EBF1* gene expression patterns in MB, previous reports mentioned upregulation in Group 4 MB and differential expression among Group 4 subtypes [221, 229]. In terms of gene regulation, an enrichment of *EBF1* binding sites has been reported in somatic lowly methylated regions in SHH MB and enhancers active in Group 3 and Group 4 MB [218, 277]. However, as we presented above, genes upregulated in Grp4-c1 MB and putative targets

of *EBF1* were enriched for neuronal-developmental functions indicating that *EBF1* contributes to an expression signature that is associated with neuron differentiation on top of the Group 4-common signature regulated by *ZBTB18* and *NEUROD2*.

We found several interesting features for the subcluster Grp4-c2. In Grp4-c2 MB, we identified highly-ranked TFs, upregulated genes, and enriched pathways that were related to stemness or EMT/mesenchymal features or both, comprising PI3K-AKT pathway, the TFs *TWIST1*, *SOX11*, *SOX9*, and the gene *NES* [278, 283, 285, 288, 289, 293, 294]. A direct link between EMT, *SOX9*, and PI3K-AKT signalling has been described for SHH MB. Here, PI3K-AKT signalling stabilises *SOX9* protein by inhibiting GSK3/FBW7-dependent *SOX9* degradation. Elevated protein levels of the TF *SOX9* lead to EMT and higher cell motility in SHH tumours associated with metastasis and worse clinical outcome [289]. This reported mechanism could also promote metastases in Grp4-c2 MB since we found *SOX9* and genes of the PI3K-AKT pathway to be upregulated in this subcluster. In our analysis, the TF *TWIST1* showed the highest NIS for Grp4-c2 MB. *TWIST1* is a the well-described EMT-initiating TF [278] supporting the idea that Grp4-c2 MB undergoes EMT that might promote metastasis formation. *TWIST1* was not only the top-ranked TF in Grp4-c2 MB but also specifically upregulated when compared to non-WNT tumours. Therefore, Grp4-c2 could be marked by frequent metastasis. However, Cavalli *et al.* reported the same frequency of metastasis at diagnosis for Group 4 subtypes [221]. Since the published subtypes matched well to our subclusters Group 4 MB described above (see Section 5.1.3), the reported same frequency of metastasis between Group 4 subtypes does not support the idea of EMT-driven metastasis in Grp4-c2 MB at first glance. Nevertheless, it should be considered that Group 4 MB shows a generally high frequency of metastasis at diagnosis (40%) among MB subgroups which is twice as much as in SHH MB (see Section 5.1.3). Therefore, the mentioned factors could still promote EMT-associated metastasis in Grp4-c2 tumours, and alternative or partially overlapping mechanisms lead to metastasis in the remaining two Group 4 subclusters. For example, we detected an enrichment of the FGF and PI3K-AKT signalling pathway for the remaining two subclusters Grp4-c1 and Grp4-c3, respectively. Besides the PI3K-AKT signalling, several pathways can promote EMT in cancer including the FGF pathway [389]. Nevertheless, the results could be interpreted in another way, too. In cancer, EMT and stemness are disease-driving features that can occur independently but frequently co-occur [390]. EMT is linked with metastasis since it promotes motility and invasion of tumour cells, whereas expression of stemness-associated factors is related to tumour cell proliferation and maintenance [390, 391]. Therefore, these crossroads between stemness and EMT suggest that rather stemness-associated features than EMT-associated features might play a disease-relevant role in Grp4-c2 tumours [390, 391]. In this case, common or subcluster-independent mechanisms of forming metastasis would explain the similar frequency of metastasis between Group 4 subclusters (subtypes). Overall, the data that we presented highlight Grp4-c2 MB as an interesting subcluster since it upregulates EMT- and stemness-related factors and Group 4 MB is more known to resemble an immature/mature cell type. Further experiments are needed to understand the involvement of EMT and stemness factors in Grp4-c2 tumours.

The TF ranking for the SHH subgroup simply reflected the activation of the *Shh* pathways. However, new insights into the age-related molecular heterogeneity of SHH MB were provided by analysing the GRN and expression profiles of the SHH subclusters. Comparing the subclusters presented above to MB subtypes published by Cavalli *et al.* underlined that the identified SHH subclusters SHH-c1 SHH-c3, and SHH-c2 represent infant, childhood, and adult SHH MB cases, respectively. Among the identified subclusters, the expression profiles of the SHH subclusters showed the most complex relation. These data indicated that infant and adult SHH MBs have a distinct expression profile, and childhood SHH MBs share expression profiles with the remaining two age groups. This is in line with several publications that had concentrated on the dissection of the molecular heterogeneity within SHH MBs [212, 214, 221]. However, none of these publications has defined the set of genes

that childhood SHH MB shares with infant or adult cases. Here, the genes presented above that are differentially expressed between SHH subclusters provide new insights into the age-group-related expression pattern.

TFs that were highly ranked in the SHH subclusters also followed the age-group-related expression pattern indicating a regulatory link. Here, *ATOH1* and *NEUROD1* might make a main contribution to the age-group-related expression pattern. In mouse MB models, it has been shown that *NeuroD1* overexpression promotes neuronal differentiation and leads to downregulation of Shh targets such as *Gli1*, whereas *Atoh1* overexpression represses neuronal differentiation including downregulation of *NeuroD1* [237, 392]. The apparent inverse association between these two TFs and their regulation of cell differentiation-repressing or -promoting programmes has been recently validated in human MB (published during the work on the present study). Here, Hovestadt *et al.* demonstrated that *NEUROD1* and *ATOH1* expression is negatively correlated and higher in infant or adult SHH MB, respectively. Additionally, the authors reported that *NEUROD1* expression in infant SHH MB correlates with the expression signature of immature/mature granule neurons and unipolar brush cells, whereas *ATOH1* expression in adult SHH MB correlates with the signature of progenitors of these two cell types [388]. These reports fit well with the results shown above, including the expression patterns of *ATOH1* and *NEUROD1* and the observed functional enrichments related to repression or promotion of cell differentiation in adult and the non-adult SHH subclusters, respectively. Our presented expression profiles of *ATOH1* and *NEUROD1* in MB suggest that the children subcluster SHH-c3 might be influenced by both TFs, explaining the intermediate expression profile of SHH-c3 MBs.

Besides upregulated TFs, the visualised GRNs also depicted the contribution of subgroup- or subcluster-specifically downregulated TFs to gene regulation in MB. In the context of subgroups, TFs downregulated in WNT or SHH MB showed a high impact on gene expression in non-WNT and non-SHH MB, respectively. Here, previous work in Lin *et al.* study confirmed that *OTX2*, *RREB1*, and *TBR1* are potentially involved in gene regulation in non-SHH MB based on binding site enrichments detected in active enhancers and super-enhancers [218]. In our MB study presented above, *DEK* was among the TFs that were specifically downregulated in WNT MB. Here, we found *DEK* also to be upregulated in non-WNT MB vs. normal cerebellum. These results are supported by a reported upregulation of *DEK* in Group 4 MB vs. fetal neural brain [230]. However, *DEK* is not yet well studied in the context of MB. *DEK* has been reported to be an oncogene in many different cancer types [393]. Therefore, it might be of interest to study *DEK* in the context of non-WNT MB for future research [393].

Our performed overlay of copy number and gene expression of selected TFs emphasise a general contribution of copy number variation on expression levels of TFs indicating that copy number changes influence GRNs in MB. However, copy number variations alone did not always fully explain expression patterns and, therefore, additional mechanisms most likely influence expression levels of TFs as well.

The GRNs that we inferred from gene expression data (expr-GRNs) depicted regulatory links but did not resolve to which extent TF binding in promoter or enhancer elements played a role in gene regulation [394]. By integrating putative enhancer-targeted genes in MB, published by Lin *et al.*, we showed that over 50% of predicted TF target genes are potentially regulated via TF binding in enhancer elements [218]. The high fraction of enhancer-targeted genes underlines the reported idea that enhancers essentially contribute to transcriptional changes in cancer [395]. Furthermore, the fact that not all genes in our GRNs were assigned to be an enhancer target indicates a contribution of promoter-mediated gene regulation in the expr-GRNs. Promoter-mediated gene regulation is a type of regulation that is not covered by the previously published enhancer-mediated GRN (enh-GRN) of the Lin study. An additional advantage of the expr-GRN compared to the enh-GRN is the independence from available TF binding motifs for GRN construction. Here, the enh-GRN is biased by the research status of individual TFs. Additionally, the quality of available TF motifs can differ, and TFs of the same family recognise the same or highly similar motifs, which complicates the assignment of regulatory links to a particular TF when several TFs of the same family are expressed [396]. On the other hand, in expr-GRNs, the individual contributions of highly correlated TFs cannot always be

deconvolved, which results in the inference of false or indirect links in the case of a regulatory cascade ($TF_1 \rightarrow TF_2 \rightarrow \text{target}$). Here, expr-GRNs could be improved by the integration of TFBS predictions. Overall, our study presented above extends the previous work of Lin *et al.*

Lastowska *et al.* published a GRN for MB inferred from microarray gene expression data, too, but concentrating on a specific aspect. At first, the authors performed a mutagenesis experiment to identify 17 genes (including six TFs) that were associated with SHH MB formation in *Ptch*^{+/-} mice tumour models. Second, the authors inferred an expression-based GRN in MB. Several of the found SHH MB-associated genes were part of a subnetwork that was connected to gene regulation in Group 4 (highlighting *MYT1L*) and non-Group 4 MBs. However, due to the authors' focus, the GRN of Lastowska *et al.* shows only a fraction of the GRN in MB compared to the expr-GRNs that we presented above.

All in all, the individually inferred GRNs of subgroups and subclusters within subgroup SHH, Group 3, and Group 4 depicted the landscape of TFs that probably predominantly contribute to the transcriptional heterogeneity in MB linking TFs to certain transcriptional signatures. The dissection of gene regulation in MB highlighted TFs that might be of interest for future studies including so-far unknown master regulators of Group 4 expression profiles, stemness- and EMT-associated TFs in Grp4-c2 MB, *HLX* in Grp3-c1 MB, and *DEK* in non-WNT MBs.

5.5.4 Characterisation of differentially expressed lnc genes in MB

Besides the analysis of the expression profiles and GRNs in subgroups/subclusters, the study presented above also focused on the characterisation of differentially expressed lnc genes in MB.

Unsupervised clustering and differential gene expression showed that expression of lnc genes follow the profiles of MB subgroups and subclusters. This analysis provides the general idea of a potential role of lnc genes in MB. We selected the set lnc genes that we further analysed based on differential expression among subgroups or subclusters. Overall, the performed DGEA identified a smaller number of lnc than coding genes. As shown above and in previous reports, lnc genes have a generally lower expression than coding genes, which might be a potential explanation for the smaller number of detected differentially expressed lnc genes (Section 5.3.3.2) [133]. A lower gene expression has the consequence that fewer reads are available. The lower read number, in turn, leads to less statistical power to detect differential gene expression by the applied DGEA method edgeR. This method assumes a negative binomial distribution for the modelling of read counts [397, 398] (see Section 3.3.2). Additionally, we applied the same cutoffs to coding and lnc genes. The chosen average absolute difference of ≥ 1 RPKM between groups represents probably a strict cutoff for lnc genes. All these points indicate that the 448 lnc genes detected above comprise only the highest expressed lnc genes that are differentially expressed in MB. Therefore, these 448 lnc genes represent a robust set for detailed characterisations.

The computational characterisation that we performed for the 448 lnc genes in MB focused on several aspects:

- a systematic general characterisation by the position relative to coding genes and expression correlation with neighbouring coding genes,
- the identification of lnc genes interesting in the context of MB, and
- inferring potential implication of individual lnc genes in MB.

Here, (1) several external resources (FANTOM CAT, BrainSpan, and Ensembl), (2) literature, and (3) the analysed MB RNA-seq cohort were used for the characterisation [133, 299, 354].

Previous reports showed that the classification of lnc genes into different types, which is based on their position relative to coding and pseudo-genes, provides insights into the molecular genetics and biology of lnc genes and their transcribed lncRNAs, as previously shown [133, 135]. For this reason, we classified the 448 analysed lnc genes into the three types divergent, antisense, and intergenic. Here, the resource FANTOM CAT provided essential information for the classification of divergent lnc genes since FANTOM CAT was developed to improve lnc gene models by an accurate annotation of

transcription start sites [133] (see Section 3.6.1). To understand the three lnc gene types in the context of MB, we compared the analysed lnc genes based on the three types. This comparison included the positional coding gene partners/neighbourhood of the lnc genes. The functional enrichments detected above for transcription factor activity and developmental processes in coding gene neighbourhood of divergent and intergenic lnc genes is supported by previous genome-wide studies [134, 135]. Therefore, these functional associations are rather general than MB-related characteristics of coding partners of these two lnc gene types. However, the coding gene neighbourhood of divergent lnc genes was enriched for nervous system development. This enrichment is of interest because nervous system development-associated genes/TFs can have tumourigenic functions in MB or contribute to its molecular heterogeneity [193]. Considering this and a described regulation of coding partners by divergent lnc genes in *cis* [399], divergent lnc genes might regulate the expression of coding genes with implications in MB.

The higher expression specificity that we observed for lnc genes when compared to their coding neighbourhood is a generally known characteristic of lnc genes [133]. However, the strong positive expression correlation that we detected between divergent lnc genes and their coding partners in MB would not necessarily support a significantly higher expression specificity for divergent lnc genes. (The observed strong positive expression correlation is in line with previous reports [135].) This discrepancy might be explained by the general lower expression of lnc genes compared to coding ones, as shown above and previously reported [133]. The lower expression could have introduced a bias in the expression specificity score that was obtained from FANTOM CAT [133]. The lower expression might increase the probability that expressed lnc genes remain undetected due to limitations of every RNA quantification technology including the CAGE technology that has been used to calculate the expression specificity scores [133]. However, the published expression specificity score should still provide useful information. However, probably only pronounced differences in cell-specific expression as seen between intergenic and coding genes should be considered significant. Previous studies showed that intergenic lnc genes have a general higher expression specificity because their transcription initiation sites frequently overlap with enhancers [133]. The association between intergenic lnc genes and enhancers is of interest because enhancers are involved in cancer-associated transcriptional changes [395]. Additionally, oncogenic functions have been reported for lnc genes that are transcribed from super-enhancer loci [400]. Therefore, further characterisation of intergenic lnc genes by obtaining H3K4me1 histone marks to distinguish promoter- and enhancer-associated TSS in MB would be of interest to understand better the role of intergenic lnc genes in this disease.

In our analyses, the majority of lnc genes were significantly positively correlated. Only nine lnc genes showed a significant negative correlation with coding partners. This negative correlation might be interesting due to the fact that lnc RNAs can act as a negative regulator of transcription in *cis* [401]. However, for none of these nine lnc genes, such a function has yet been reported. The observed significant negative correlation might not imply *cis*-regulating functions for these lnc genes and could be caused by coincidence. The negative correlation between *ZFAS1* and the divergent coding partner *ZNFX1* observed above was also reported in Head and Neck Squamous Cell Carcinomas. However, across breast cancer cell lines, *ZFAS1* and *ZNFX1* no expression correlation was reported [402, 403]. Additionally, *DLGAP1-AS1*, which we detected to be negatively correlated with the antisense coding gene *DLGAP1*, is only one out of five lnc genes in antisense orientation to the large *DLGAP1* locus that spans nearly one Mb. The genomic location of both genes does not indicate a strong dependency between *DLGAP1-AS1* and *DLGAP1*.

Nevertheless, our investigation of expression correlation between lnc genes and coding gene partners facilitated the definition of correlation-based lnc gene categories. These categories allowed us to lnc genes that were not significantly positively correlated with coding partners in MB. This set of lnc genes qualified for further analyses because additional characterisations were based on expression data. Here, a strong positive correlation between lnc genes and coding partners would not allow inferring potential functions of individual lnc genes independent of the coding partner. However, it

is important to note that significance alone is probably not an appropriate indicator of whether lnc and coding partners show a potentially relevant co-expression. A significance level can be reached for non-meaningful/low correlation coefficients in larger data sets as in the analysed ICGC MB cohort due to high statistical power. Here, in our study, a Spearman correlation coefficient > 0.3 was used as a cutoff for potentially relevant co-expression between partners. A significance level (FDR < 0.05) was already reached at a coefficient around 0.16.

We identified 95 that were not significantly positively correlated with coding partners in MB among the 448 lnc genes that we computationally characterised. For 20 of these 95 lnc genes, integration and analyses of external data (BrainSpan, FANTOM CAT) revealed expression patterns that are associated with brain/cerebellum development, NSCs, and ESCs (Figure 5.31). These expression patterns are of particular interest due to the embryonic origin of MB and, therefore, may point to lnc genes that play a role in this disease (see Section 5.1.1). However, many of these 20 lnc genes are un- or rarely studied. Among these 20 lnc genes, development-related expression has been reported for *GAS5*, *RMST*, *RP11-453F18__B.1* (alias *FIRRE*) [315, 404, 405]. Additionally, our annotation of development-related expression patterns highlighted the antisense lnc gene *GLYCTK-AS1*. Only *GLYCTK-AS1* was expression-enriched in ESCs and upregulated in pre- vs. postnatal cerebellar tissue displaying an exclusive expression in prenatal cerebellum (Figure 5.31 and Figure 5.32.a). Additionally, we detected a strong expression correlation between *GLYCTK-AS1* the neural stem cell marker *HES5* [300] in the prenatal human brain (BrainSpan), which supports an implication of *GLYCTK-AS1* in brain/cerebellum development. The resource FANTOM CAT provides the information that the TSS of *GLYCTK-AS1* is located in an enhancer, whereas the antisense coding gene *GLYCTK* is transcribed from a promoter, which is underlining independent transcription of *GLYCTK* and *GLYCTK-AS1* [133]. The transcription of *GLYCTK-AS1* from an enhancer supports a cell-type-specific expression of this gene [406]. Here, we observed exclusive expression in the prenatal cerebellum. Additionally, FANTOM CAT annotations comprised enriched expression mainly in neuronal/brain-associated cell/tissue ontologies for *GLYCTK-AS1* [133]. As already pointed out, additional H3K4me1 marks are necessary to validate the enhancer-associated transcription of *GLYCTK-AS1* in MB. Interestingly, current gene annotations in mice contain an uncharacterised lnc gene (D030055H07Rik) that is in antisense orientation to *Glyctk* (GRCm38, Ensembl release 100) [407]. If *GLYCTK-AS1* is conserved in mice, ISH experiments might provide insights into the potential role of *GLYCTK-AS1* in brain/cerebellum development and cell-type-related expression. Overall, our results strongly suggest that *GLYCTK-AS1* is linked to cerebellar development and potentially stem/progenitor cells. The upregulation of *GLYCTK-AS1* in Group 4 MB, which we showed above, might be interesting because this subgroup has rather been associated with im-/mature cell types than progenitor/stem cells and, therefore, *GLYCTK-AS1* might reveal new aspects about Group 4 MB [314]. Additionally, we showed a constant upregulation of *GLYCTK-AS1* in subcluster SHH-c2 compared to the remaining SHH subclusters, which might point to a specific role of *GLYCTK-AS1* in adult vs. non-adult SHH tumours.

Twelve lnc genes, which we selected based on previous reports of these lnc genes, were reassessed in the context of MB using the analysed RNA-seq MB cohort. Here, we depicted subgroup- and subcluster-related expression patterns and previously reported regulatory links between lnc and coding genes were evaluated in MB. This analysis revealed some interesting aspects. We found the lnc gene *LINC-ROR* to be exclusively expressed in WNT MB and absent in the normal cerebellum. The reported cell proliferation- and metastasis-promoting function of *LOXL1-AS1* in MB that had been studied in Group 3/4-classified MB cell lines [305] might be especially relevant for subcluster Grp3-c1 and -c3, and Grp4-c2 because we observed consistent high expression in these subclusters of Group 3 and Group 4 compared to normal cerebellum. Based on the strong positive correlation of *FEZF1-AS1* and *HOTAIRM1* with their coding partners, we could verify that the reported *cis*-regulatory function of both lnc genes [317, 319] is preserved in MB. The regulatory roles of both lnc genes in *cis* might be of interest in MB since the coding partners *HOXA1/HOXA2* and *FEZF1* have been shown to promote tumour cell proliferation in glioblastoma and colorectal carcinoma, respectively [317, 319].

Overall, our evaluation of previously reported regulatory (direct or indirect) links between lnc and coding genes was limited to a small number of reported links because publications were manually evaluated. Here, the integration of databases that collect reported targets of lnc genes (such as "LncRNA2Target") might facilitate a faster and more comprehensive evaluation of literature-based knowledge in MB and should be considered for future studies [408].

We identified the co-expression cluster Cc1 in MB that showed high expression in WNT, SHH and Grp3-c1 MB. This co-expression cluster included five of the lnc genes that we selected via literature, *PVT1*, *SNHG16*, *GAS5*, *ZFAS1* and *DANCR*. The accumulation of these five lnc genes in the co-expression cluster Cc1 is noticeable since all of these lnc genes have been frequently reported to be involved in cancer [307–310, 320, 321, 324, 329, 331, 332, 409]. Further, among the genes of Cc1, we detected functional enrichments that suggest that these five lnc genes are associated with MYC-target expression and translational processes in MB, following the approach guilt-by-association (see Section 3.6.2). Previous publications of these lnc genes support these functional associations. *PVT1*, *SNHG16*, and *DANCR* are known MYC targets in humans providing a direct regulatory link between these lnc genes and the MYC-target enrichment in the co-expression cluster [213, 310, 322]. In prostate cancer, Chen *et al.* have identified a co-expression cluster that contained *GAS5* and *ZFAS1*. This cluster was enriched for RNA-processing and protein translation, similar to co-expression cluster Cc1, indicating that the strong correlation between these two lnc genes and the functional association with translational processes is not only present in MB [330]. Moreover, a direct association with ribosomes and ribosome biogenesis, reflecting translational processes, has been reported for *SNHG16*, *ZFAS1*, and *GAS5* [332, 402, 410].

Additionally, the role of *GAS5* needs to be evaluated in MB since it is mostly known as a tumour suppressor [329]. Nevertheless, two publications showed oncogenic functions of *GAS5* depend on the expression of specific isoforms. This is an aspect that should be considered for future studies of *GAS5* in MB [331, 332]. Among these five lnc genes, only *PVT1* is known in MB, but just in the context of co-amplification or fusion with *MYC* in MYC-driven Group 3 tumours [213]. The expression profiles of *PVT1* shown above and the reported oncogenic functions in other cancer types indicate that *PVT1* might play a role also in WNT and SHH MB as well [409].

Overall, additional experiments are needed in order to understand the exact functional implication of these five lnc genes in MB since these lnc genes fulfil their oncogenic functions by a variety of mechanisms [320, 324, 326, 409].

Interestingly, our evaluation of the expression correlation between mean pattern of Cc1 and members of the MYC gene family members strongly suggests that the co-expression cluster Cc1 might be collectively regulated by *MYC*, *MYCN*, and *MYCL* in MB in association with subgroup-specific dependency. Here, especially *MYC* and *MYCN* have been reported to drive MB formation (see Section 5.1.3) [269]. Therefore, the co-expression cluster Cc1 probably represents tumour-promoting processes that MYC family members collectively regulate in MB. Here, the enrichment in Cc1 for translational processes shown above provides a link between a tumour-promoting role of Cc1 and MYC family genes. Translational processes are known to be regulated by MYC, where increased translation/protein synthesis is considered to enhance cell-cycle progression in cancer [411].

Our performed analyses strongly suggest a correspondence between co-expression cluster Cc1 and a reported MYC-like kinase activity profile that has been described in MB by Zomerman *et al.* (see Section 5.1.5) [220]. The authors showed that *MYC* or *MYCN* (*MYCL* was not tested) can induce this MYC-like kinase activity profile. However, Zomerman *et al.* reported that several MB samples that were associated with a MYC-like kinase activity profile did not show high expression of *MYC*, *MYCN* or *MYCL*. Therefore, the authors concluded that *MYC* or *MYCN* abnormalities (overexpression and amplification) are dispensable for the activation of a MYC-like kinase activity profile in medulloblastoma. In this case, other aberrations might induce this kinase profile. However, this conclusion conflicts with our results shown above of the collective regulation of the co-expression cluster Cc1 by MYC family members considering that Cc1 is associated with the MYC-like kinase profile. The results presented

above highlight that the summed overall expression of *MYC*, *MYCN*, and *MYCL* needs to be considered to fully capture the impact of MYC family members in individual MB tumours, an aspect that was not considered by Zomerman *et al.* [220].

In the study that we presented above, the most detailed functional characterisation was done for *MEG3* using clinical data, a gene-centred co-expression analysis, and predicted DNA binding sites. As previously proposed [41], associations with clinical outcomes can point to lnc genes with cancer-relevant functions. To detect associations between gene expression and OS, it was essential to determine an expression cutoff that allowed an optimised split of cases into two groups associated with different survival rates. Here, we applied a signal-to-noise-/subsampling-based approach to avoid the overfitting of detected associations. Overfitting needs to be addressed when values of a continuous prognostic factor are categorised [97]. The top-three cutoff solutions for *MEG3* expression of the discovery and external cohort varied around an overlapping solution. This demonstrates that the applied approach truly allowed an estimation of robust cutoffs by avoiding overfitting, considering the different sample size (164 vs. ~700 samples) and technologies used for gene expression quantification of the two cohorts (RNA-seq vs. microarrays).

We chose a gene-centred co-expression analysis for *MEG3* because genes that are negatively correlated with *MEG3* expression were of interest. Therefore, co-expression clusters were not applicable. Co-expression cluster analyses can be performed in an unsigned or signed manner. Signed analyses result in clusters comprising only positive correlation, whereas unsigned analyses result in clusters comprising positive and negative correlation. Signed analyses are to be preferred over unsigned ones because signed analyses provide more interpretable and biological relevant clusters. The identification of co-expression clusters that only relate to negative correlation is not possible [138, 140]. Additionally, gene-centred co-expression/guilty-by-association analyses have been previously used in lnc characterisation. This allows for analysing positive and negative correlations independently [139]. We improved the inference of *MEG3* functions in MB by integrating predicted binding sites of this lnc genes. Hereby, the detection of particular pathways that are negatively associated with *MEG3* expression was enabled. The usefulness of predicted binding sites for the characterisation of the lnc gene is also highlighted by a recently published pan-cancer study by Chiu *et al.*, in which binding site predictions and co-expression analyses were combined [412]. The authors used the Triplexator tool to predict lnc gene binding sites because it works independently of ChOP data.

Independent of the binding site prediction, our inspection of detected correlations between *MEG3* and coding genes revealed potential spurious correlations since subgroup-dependent expression was a confounding factor. Here, subgroup-dependent gene expression profiles reflected different marginal distributions that might have caused the detection of spurious correlations [413]. We applied a heuristic approach to detect spurious correlations (Section 5.4.6.9, 5.3.3.7). Our evaluation suggests that this approach successfully selected spurious correlations, but the fraction of missed spurious correlations remains unknown. This heuristic approach has the advantage that the removal of effects of supposed confounding factors can be well controlled, for example, by choosing a cutoff for the strength of correlation that should remain after removing one subgroup or within a single subgroup. However, this approach has the disadvantage that confounding effects are not fully captured. The application of model-based correction of confounding effects in co-expression analyses might be an option to capture the whole effect of confounders [414, 415]. However, the controllability of effect removal in the performed analyses was important in our study since truly correlated gene pairs could still be under the effect of subgroup-dependent expression.

We showed that *MEG3* expression levels in MB tumours can stratify SHH and Group 4 cases into outcome-related groups (Section 5.3.3.6). Furthermore, we demonstrated that *MEG3*-based outcome stratification in these two subgroups was restricted to the subclusters SHH-c1, Grp4-c2, and Grp4-c3.

MEG3-high-expressing Grp4-c3 cases showed a strikingly favourable outcome, and *MEG3*-low-expressing Grp4-c2 tumours showed the worst five-year survival rate among Group 4 cases. Molecular subtypes of Group 4 tumours that are reported by Cavalli *et al.* or Schwalbe *et al.* do not provide a good

stratification of patients with Group 4 tumours in terms of overall survival [221, 223]. However, our results show that risk stratification of Group 4 cases by jointly using molecular subclusters/subtypes and *MEG3* expression could improve the identification of low risk (Grp4-c3 + *MEG3*-high), high risk (Grp4-c2 + *MEG3*-low), and medium risk (remaining Group 4 samples) Group 4 patients.

We demonstrated in SHH-c1 (infant SHH MB) cases, high expression of *MEG3* strikingly marked favourable outcomes (similar to Grp4-c3 MB), whereas low *MEG3* expression marked patients with rapid death and a miserable outcome. A higher tumour sample number allowed Cavalli *et al.* to define two subtypes of infant SHH MB jointly represented by subcluster SHH-c1. Here, in both subtypes SHH γ and SHH β , *MEG3*-expression-based stratification could still define patient groups with significantly different OS. Cavalli *et al.* showed that subtype SHH β represents infant SHH MB cases with a worse OS compared to subtype SHH γ . However, a *MEG3*-expression-based stratification might improve outcome prediction in infant SHH MB when comparing the 5 years overall survival between subtype SHH γ/β and *MEG3* expression-stratified patient groups (Figure 5.3, 5.42, A.56).

By identifying enriched processes among genes that are negatively correlated with *MEG3* expression, we could show that *MEG3* probably acts as a tumour suppressor in MB via negatively regulating mitotic cell cycle and TGF β pathway (see Section 5.3.3.8). Similar tumour suppressive functions of *MEG3* have been reported, for example, in breast cancer (see Section 2.5) [43, 416]. We observed this potential tumour suppressive function of *MEG3* in MB in a subgroup-dependent manner for SHH and Group 4 MB, providing a functional link and an explanation why *MEG3* expression is prognostic in SHH and Group 4 subclusters but not in Group 3 MBs.

We could highlight *CDK1*, *CCNB1*, *MYC*, and *TGFBR1* as putative *MEG3* targets by identifying predicted *MEG3* binding sites in promoter and enhancer regions of these genes. Here, *MEG3* probably acts as a negative regulator, taking the negative expression correlation between *MEG3* and these genes into account. *CDK1* and *CCNB1* (regulators of G2-to-M phase progression in the cell cycle [349]) provide a potentially direct regulatory link for the observed negative association between *MEG3* expression and mitotic cell cycle in SHH and Group 4 MB. These data indicate that *MEG3* can induce cell growth arrest in SHH and Group 4 MB, a tumour suppressive function of *MEG3* that has been described in several cancer types including downregulation of *CDK1* and *CCNB1* [50, 59].

We detected a negative correlation between *MYC* and *MEG3*. This negative correlation seemed to be more relevant for Group 4 MB because it was stronger in this subgroup. Grp4-c3 tumours that were stratified based on *MEG3* expression, in turn, showed upregulation of *MYC* among Group 4 subclusters. These data indicate that high *MEG3* expression in a subset of Grp4-c3 MBs might impair tumour progression via downregulation of the oncogene *MYC* [269, 417]. Huang *et al.* reported that *MEG3* functions as ceRNA in bladder cancer and indirectly, negatively controls *MYC* mRNA expression through a complex cascade involving miR-27a, *PHLPP2*, and JUN, which impairs tumour invasion [418]. However, the data shown above suggest that *MEG3* might regulate *MYC* in MB via binding a *MYC*-regulating enhancer.

Among genes that were negatively correlated with *MEG3*, we could show that a functional enrichment for the TGF β pathway increased after filtering for genes carrying a *MEG3* binding site. This increase of enrichment highlights *MEG3* as a negative regulator of the TGF β pathway in MB. As shown above, a predicted *MEG3* binding site was located in an enhancer of *TGFBR1* that has been shown to interact with the *TGFBR1* promoter in MB [218]. This provides a possible mechanism for how *MEG3* might regulate the TGF β pathway in MB. We showed that *TGFBR1* and *MEG3* expression are significantly negatively correlated in both SHH and Group 4 MB but stronger in Group 4 tumours. Additionally, we found the strongest functional enrichment for the TGF β pathway among genes that were negatively correlated with *MEG3* in Group 4 tumours. These data indicate that negative regulation of the TGF β pathway by *MEG3* might be especially relevant for Group 4 MB. Furthermore, we also observed a strong negative correlation between *MEG3* and *TGFBR1* in developing brain tissues and in pre- and postnatal cerebellum in humans. Here, our findings suggest that the negative regulation of *TGFBR1* by *MEG3* is a transcriptional program that is active during the development of the brain and cerebellum. TGF β

signalling has been previously linked to genes that are affected by genetic events in Group 3 MB and to metastasis formation and progression in SHH MB [222, 419, 420]. However, an implication of the TGF β pathway in Group 4 MB still needs to be evaluated since this pathway shows a tumour-suppressive or promoting function in cancer [421].

We observed a downregulation of *MEG3* in MB (excepting Grp4-c1 tumours) compared to normal cerebellum (excepting Grp4-c1 tumours), a finding that is supported by previous reports [337]. This downregulation provides additional evidence that *MEG3* might act as a tumour suppressor in MB. Our findings suggest that the observed capability of *MEG3* to stratify SHH-c1, Grp4-c2, and Grp4-c3 MB is associated with a certain expression distribution in these subclusters. Here, the fraction of MB samples that displayed either *MEG3* expression within or below the expression range of normal cerebellum was similar. These results indicate that *MEG3* has a critical expression level to act as a tumour suppressor in MB, which is frequently reached in SHH-c1, Grp4-c2, and Grp4-c3 MB but infrequently in WNT, SHH-c2, SHH-c3, and Group 3 tumours. In the case of subcluster Grp4-c1, it remains unclear whether the general high *MEG3* expression or unknown subcluster-dependent mechanisms — counteracting tumour suppressive functions of *MEG3* — prevented a stratification of Grp4-c1 cases using *MEG3* expression. The possibility that a marker can be prognostic only in one subcluster within a subgroup is supported by Cavalli *et al.*. The authors described a subtype-dependent implication of *TP53* mutations among SHH subtypes, where a *TP53* mutation was linked to a desperate OS in SHH α , but not in the renaming SHH cases [221].

Several lnc genes have recently been reported to have implications in MB comprising *CCAT1*, *CDKN2B-AS1*, *linc-NeD125*, *Nkx2-2as* [422], and *UCA1* [423]. None of these genes was among the detected differentially expressed lnc genes. This indicates the existence of MB-relevant lnc genes that do not follow subgroup-/subcluster-specific expression, despite the possibility that these lnc genes were expressed at a level that did not pass the strict cutoffs of the DGEA. Therefore, for future studies, it would be of interest to expand the performed characterisation to lnc genes that are highly variably expressed in MB or differentially expressed between controls and MB.

Joshi *et al.* [424] and Kesharwani *et al.* [425] recently published an analysis of lnc genes in MB, but with a much more limited scope compared to our study presented above. Both publications concentrated on expression profiles associated with subgroups and did not address subclusters. Even though Joshi *et al.* and Kesharwani *et al.* reported more lnc genes associated with survival in MB, in none of the studies, a potential mechanism between lnc gene expression and the observed survival association was investigated. Using the Cavalli cohort, Kesharwani *et al.* reported *MEG3* to be prognostic in SHH MB but did not evaluate subtypes of SHH MB nor provided further details [425]; here, Group 4 MB were also not mentioned. Kesharwani *et al.* used microarray data (covering only a fraction of lnc genes on the genome) and Ingenuity pathway analysis to identify potential regulators and enriched biological functions of differentially expressed lnc genes per subgroup. However, for the enrichment analysis, the authors did not take into account that most lnc genes are not studied and, therefore, lack annotations in gene functional databases [425]. Using the ICGC MB RNA-seq cohort, Joshi *et al.* identified nine co-expression clusters of lnc genes and linked four of them to subgroup-specific expression. However, they did not provide further functional annotations of the co-expression clusters [424]. Joshi *et al.* additionally concentrated on classifying MB samples into subgroups using lnc gene expression [424]. Considering the work of Joshi *et al.* and Kesharwani *et al.*, the study that we presented above provides a more detailed analysis and extensive characterisation of lnc genes in MB than the most recent genome-wide publications on this topic in MB.

6 Implications and conclusions

This chapter recapitulates the main findings and associated literature related to this thesis, focusing on computational analyses of CRC and MB cancer transcriptomes.

6.1 Machine learning-based classification and treatment outcome prediction in colorectal carcinoma

The process of treatment choices for cancer patients has only recently developed from a "one-size-fits-all" treatment per cancer type towards precision oncology that considers the tumours' genomic and molecular makeup [426]. The application of high-throughput omics technologies facilitated the step towards precision oncology by enabling the analysis of a tumour's whole molecular landscape. The genomic and molecular information can be used to identify biomarkers and construct models (like classifiers) predicting treatment outcome [426]. Omics data types that are commonly used for precision oncology include mutations, copy number variations, gene expression, DNA methylation, and proteomics [426]. However, gene expression, DNA methylation, and proteomics data show the most predictive potential [426, 427]. Statistical and machine learning approaches (such as (M)SVM-RFE) play an essential role in biomarker identification and model construction [426]. In order to successfully apply these approaches, common issues like overfitting and class imbalance for classification tasks need to be addressed [426].

The cetuximab response classifier that we presented above (OT mini-classifier) has a potential impact on treatment choice in CRC. The good performance of the OT mini-classifier underlines: using a cost-sensitive SVM and stratified resampling provides a solution — minimally interfering with the learning step and training data — for the application of (M)SVM-RFE to class-unbalanced data. The good performance also shows that MSVM-RFE can be applied for the selection of predictive expression signatures and the construction of predictive classifiers in the context of precision oncology. The reported higher predictive potential of gene expression compared to mutation data in precision medicine [426, 427] is supported by the gene-expression-based OT mini-classifier that outperformed the commonly used *KRAS/NRAS/BRAF* mutation status to predict the outcome of cetuximab treatment in CRC. Additionally, the gene-expression-based classifier serves the need for biomarkers/tools predicting cetuximab treatment outcome in *RAS/RAF* wild-type CRCs. Overall, the built OT mini-classifier suggests the use of gene-expression-based classifiers to complement (or instead of) mutation status whenever possible to predict cetuximab treatment outcome in CRC. This aspect should be considered for the future development of precision oncology guidelines for CRC patients, although cost-benefit as well as regulatory issues also come into play.

6.2 GRN inference and expression profiles and regulators of the molecular heterogeneity in medulloblastoma

The application of systems biology approaches has been an important element in cancer research to dissect the complex molecular interactions defining the biological system of cancer [115]. The identification of transcriptional gene regulatory networks, which play an essential role in the control of numerous biological and cellular processes [428], is a major challenge in systems biology [116]. Reverse engineering of GRNs using gene expression data is a relevant approach. This approach enables the genome-wide inference of regulatory TF-gene interactions in combination with high-throughput

6 Implications and conclusions

technologies like RNA-seq. In contrast, ChIP-seq techniques focus on individual TFs [116]. The large number of available algorithms that infer GRNs from expression data, such as GENIE3, allow an importance ranking of TF-gene interactions by providing interaction weights [119]. However, the choice of an interaction weight threshold, which controls the number of false positive and false negative predictions, is left to the user [120, 378]. A number of methods have been proposed for selecting features based on importance scores (such as interaction weights) from tree-based ensemble methods, such as GENIE3 [379]. However, many of these methods are computationally extensive because they require multiple runs of the learning algorithm on permuted data [379, 380]. Additionally, these methods cannot deal well with regulatory cascades (e.g. $TF_1 \rightarrow TF_2 \rightarrow \text{target}$) because TFs upstream in the cascade are still predictive of the expression of the downstream target gene. Therefore, these methods would select many (false positive) indirect links [379]. In contrast to these methods, we did not evaluate the interaction weights directly but examined different thresholds by evaluating the resulting GRNs. Thus, our proposed approach addresses the GRN inference task directly. For the evaluation we proposed a GRN fitting score that is based on the ratio between the average of predicted TFBS enrichments across TFs and the network density. Additional work is necessary to evaluate the reliability of the GRN fitting score but this score might represent a comprehensible measure that assists in finding a threshold for trustworthy TF-gene interaction weights for gene-expression-based GRN inference.

In MB, GRNs have mostly been studied under particular aspects including subgroup-specific gene expression [268], enhancer-mediated GRNs of subgroups [218], and small GRNs comprising SHH-MB-promoting genes identified via mutagenesis screening [229]. In contrast to these publications, the GRNs that we presented in this thesis were inferred from gene expression data without limiting the analysis to certain aspects of gene regulation in MB. For example, it can be assumed that our gene-expression-based GRNs comprise not only enhancer-mediated but also promoter-mediated gene regulation, a type of regulation that is not covered by the previously published enhancer-mediated GRN of the Lin study. The inference of GRNs allowed us to depict the regulatory maps and landscape of master regulators of the subgroup- and subcluster-specific gene expression in MB. The presented work extends the enhancer-mediated GRN published by Lin *et al.* and systematically illustrates GRNs in MB subclusters (subtypes) for the first time.

The subclusters that we identified and analysed relate to the recently emerging research of molecular subtypes within the four main MB subgroups in MB [221–223, 429]. These subtypes within the four main MB subgroups represent a further development revealing and dissecting molecular heterogeneity of MB [210]. Generally speaking, the gene expression profiles of MB subtypes are less well studied compared to the four MB subgroups. Work by Northcott *et al.* [222], Schwalbe *et al.* [223], and Sharma *et al.* [429] provided only limited information on the expression profile of MB subtypes. Cavalli *et al.* analysed the functional enrichments of subtype-related expression profiles but used microarrays and only identified the most variable expressed genes in each subgroup without reporting specific up- or downregulation of genes for a particular subtype [221].

The subclusters that we analysed in this thesis match the subtypes of Cavalli *et al.* [221]. Therefore, expression profiles and GRNs of subclusters that we presented can be directly linked to the subtypes of Cavalli *et al.*, extending further the work of these authors. Here, we identified concrete signatures of genes specifically expressed in MB subcluster; hence, providing a more defined characterisation of these subtypes. Here again, the use of RNA-seq data for the identification of tumour subclusters corroborates previous reports that particular subtypes can only be revealed via gene expression data analysis [221]. These results suggest that also gene expression and not only DNA methylation data should be considered to find a consensus of MB subtypes. This aspect was only partially addressed by Sharma *et al.* [429] who defined a consensus of MB subtypes based on DNA methylation data by comparing subtype identification approaches that were used by Northcott *et al.*, Schwalbe *et al.*, and Cavalli *et al.*. Thus, a complete evaluation of the importance of gene expression for defining MB

subtypes is still missing.

Overall, the study presented above on MB molecular heterogeneity and associated GRNs provide novel findings and hypotheses that are of interest for future MB studies:

- *ZBTB18* and *NEUROD2* as master regulators of the Group 4 expression signature with a hypothesised role associated with Group 4 tumour maintenance,
- a contribution of *RAX2* to the photoreceptor signature in MB,
- *HLX* as a putative important regulator in *MYC*-driven Group 3 MB (subcluster/subtype Grp3-c1/Group 3 γ), and
- high expression of EMT-/stemness-linked TFs (*TWIST1*, *SOX11*, and *SOX9*) and genes (*NES*) in Grp4-c2 (Group 4 α) MB, which is potentially relevant for this MB subset.

6.3 Computational characterisation of lnc genes and their involvement in medulloblastoma

Previous cancer studies have revealed lnc genes that regulate disease-relevant processes on different levels (transcriptional, post-transcriptional, translational, or signalling) by interacting with proteins, RNAs, DNA, and chromatin [34, 38]. Here, lnc genes function as oncogenes or tumour suppressors [34, 41]. However, the gene function is unknown for the majority of lnc genes [127], while the number of lnc genes exceeds the number of protein-coding genes [430]. The functional characterisation and classification of lnc genes remains challenging. One reason is the low evolutionary sequence conservation of lnc genes compared to protein-coding genes, preventing a classification based on sequence-related functional domains [127]. A second reason is the formation of rather structure-related (secondary and tertiary structures) than sequence-related functional domains [127, 128]. The application of high-throughput NGS technologies together with computational analyses has formed the basis for genome-wide studies of lnc genes, where RNA-seq and the analysis of expression data, e.g. by building co-expression networks and performing DGEAs, is commonly used for the functional prediction of lnc genes [127, 431]. Considering the current knowledge, there are different possibilities for annotating and classifying lnc genes but a "gold standard" that would provide a universal system for lnc gene characterisation is still missing [432]. For example, the classification of lnc genes according to their genomic location relative to protein-coding genes has revealed specific features for the classes of divergent, antisense, and intergenic lnc genes including different transcriptional relations between lnc genes and their coding gene neighbours [133–135]. However, only a minority of previously published pan-cancer studies of lnc genes proactively integrate this classification to obtain a more detailed lnc gene characterisation [41, 139, 412, 433–438]. Here, three out of nine studies integrated location-based lnc classification comprising the classes antisense and intergenic (and sense-overlapping lnc genes) [139, 434, 437], but only one of these three studies integrated divergent lnc genes as an additional class [437]. Two of the three studies evaluated the expression correlation with coding gene neighbours [139, 437]. Pan-cancer studies are probably a good indicator for the current design of computational analyses to characterise lnc genes in cancer genome-wide considering the large amount of data and integrative work undertaken.

The computational analysis, which is presented in this thesis, of lnc genes differentially expressed in MB underlines that a location-based lnc gene classification provides additional information enhancing the computational characterisation of lnc genes in cancer. A valuable piece of information resulting from this classification is the annotation of neighbouring protein-coding partners, which permits the evaluation of the expression correlation between lnc genes and coding partners. Since co-expression analyses are important to infer functions of lnc genes following the guilty-by-association principle [41, 431, 433, 439], a strong correlation between lnc and coding partners would not allow functional characterisation of the lnc gene independent of the coding partner. An aspect that is especially important for divergent lnc genes that show frequent strong expression correlation with the divergent

6 Implications and conclusions

coding partner, due to a shared nucleosome-free region or promoter [136], and a potential prevalent regulatory role of the coding partner in *cis* [440]. Additionally, the results that we presented highlight that the integration of binding sites for RNA:DNA:DNA triplex-forming lnc genes improve functional predictions using co-expression with protein-coding genes, a consideration that was also taken into account in a recently published pan-cancer study [412].

Two genome-wide studies of lnc genes in the context of MB have been recently published by Joshi *et al.* [424] and Keshewani *et al.* [425]. Both studies focused on subgroup-dependent expressed lnc genes and gave an overview of lnc genes that are associated with survival [424] (including *MEG3* for the SHH subgroup) [425], but did not provide potential functional mechanisms for MB. Additionally, Joshi *et al.* concentrated on lnc gene expression-based subgroup classification [424]. Keshewani *et al.* used microarray data, covering only a fraction of lnc genes on the genome, and applied functional enrichment analyses directly to lnc genes without considering that most lnc genes lack functional annotations [425].

The lnc gene characterisation that we presented above gives more explicit details on the involvement of lnc genes in MB compared to the previous work of Joshi *et al.* and Keshewani *et al.*. Within our study, computational characterisations included lnc genes that were subgroup- and subcluster-specifically expressed. Among the analysed lnc genes, we identified twenty lnc genes that showed expression patterns associated with developmental processes in the cerebellum or whole brain. These developmental expression patterns are of interest — because of the embryonic origin of MB [193]. Here, the unstudied lnc gene *GLYCTK-AS1* stands out since our results suggest that *GLYCTK-AS1* might be linked to cerebellar development and potentially to neural stem/progenitor cells. The upregulation of *GLYCTK-AS1* in Group 4 MB might be interesting because this subgroup has rather been associated with im-/mature cell types than progenitor/stem cells [314] and, therefore, *GLYCTK-AS1* might reveal new aspects about Group 4 MB. The co-expression cluster Cc1 that we identified, comprising lnc and coding genes, might point out to MB-relevant processes:

- The cluster contained five lnc genes (*PVT1*, *SNHG16*, *GAS5*, *ZFAS1*, and *DANCR*) that have been frequently described to be involved in cancer [307–310, 320, 321, 324, 329, 331, 332, 409]. However, besides *GAS5* and *ZFAS1* [330], these lnc genes have not been described in association to each other. Therefore, the presence of these five cancer-related lnc genes in the co-expression cluster might suggest a functional association between all five lnc genes in MB.
- We detected a functional enrichment for translational processes in Cc1, which enhance cell-cycle progression in cancer [411].
- Summed expression values of members of the *MYC* gene family strongly correlated with the expression profile of Cc1, which was enriched for *MYC* target genes. These results suggest additive effects of *MYC*, *MYCN*, and *MYCL* in regulating *MYC* target expression in MB. Therefore, a collective assessment of the *MYC* family genes in MB tumours should be considered to determine the cooperative influence of these genes, given that *MYC* and *MYCN* are involved in MB formation [269].

The accumulation of cancer-associated lnc genes in Cc1 paves the way for future MB studies to fully understand the functional role of the lnc genes and their association with *MYC* family genes as regulators.

Our study demonstrates an involvement of *MEG3* as a non-coding tumour suppressor in MB and prognostic marker in subtypes within SHH and Group 4 MB. Therefore, our results provide new findings for *MEG3* as a prognostic marker in MB; Keshewani *et al.* [425] reported *MEG3* as a prognostic marker only for the whole SHH subgroup. Overlaying expression correlation analyses and predictions of *MEG3* binding site suggests that *MEG3* might negatively regulate mitotic cell cycle (targetting *CDK1* and *CCNB1*), TGF β pathway (targetting *TGFBR1*), and *MYC* in MB via triplex formation in promoter and enhancer regions. Regulatory links between *MEG3* these pathways and target genes have been previously reported in other cancer types [43, 50, 59, 418]. Additional functional experiments of *MEG3* in MB are needed to validate our results and hypotheses. However, our work highlights *MEG3* as a

6.3 Computational characterisation of *lnc* genes and their involvement in medulloblastoma

promising prognostic marker in MB and adds MB to the list of cancer types where *MEG3* potentially acts as a tumour suppressor (see [52]).

A Appendix

A.1 Authors of the OncoTrack publication

Moritz Schütte^{1,*}, Thomas Risch^{2,*}, Nilofar Abdavi-Azar^{2,*}, Karsten Boehnke^{3,*}, Dirk Schumacher^{4,5,*}, Marlen Keil⁶, Reha Yildirimman¹, Christine Jandrasits², Tatiana Borodina¹, Vyacheslav Amstislavskiy², Catherine L. Worth², Caroline Schweiger⁷, Sandra Liebs⁸, Martin Lange⁹, Hans-Jörg Warnatz², Lee M. Butcher^{10,11}, James E. Barrett¹⁰, Marc Sultan², Christoph Wierling¹, Nicole Golob-Schwarzl^{7,12}, Sigurd Lax¹³, Stefan Uranitsch¹⁴, Michael Becker⁶, Yvonne Welte^{4,15}, Joseph Lewis Regan⁹, Maxine Silvestrov^{4,15}, Inge Kehler⁸, Alberto Fusi⁸, Thomas Kessler¹, Ralf Herwig¹⁶, Ulf Landegren¹⁷, Dirk Wienke¹⁸, Mats Nilsson^{17,19}, Juan A. Velasco³, Pilar Garin-Chesa²⁰, Christoph Reinhard²¹, Stephan Beck¹⁰, Reinhold Schäfer^{4,5}, Christian R.A. Regenbrecht^{4,15,**}, David Henderson²², Bodo Lange^{1,**}, Johannes Haybaeck^{7,12,**}, Ulrich Keilholz^{8,**}, Jens Hoffmann^{6,**}, Hans Lehrach^{1,2,23} & Marie-Laure Yaspo^{2,**}

¹Alacris Theranostics GmbH, Fabeckstr. 60-62, D-14195 Berlin, Germany. ²Max Planck Institute for Molecular Genetics, Department of Vertebrate Genomics/Otto Warburg Laboratory Gene Regulation and Systems Biology of cancer, Ihnestrasse 73, D-14195 Berlin, Germany. ³Eli Lilly and Company, Lilly Research Laboratories, Quantitative Biology, Avda. de la Industria 30, Alcobendas, 28108 Madrid, Spain. ⁴Charité-Universitätsmedizin Berlin, Institute of Pathology, Laboratory for Molecular Tumour Pathology, Charitéplatz 1, 10117 Berlin, Germany. ⁵German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69192 Heidelberg, Germany. ⁶Experimental Pharmacology and Oncology Berlin-Buch GmbH (EPO), Robert-Rössle-Str. 10, 13125 Berlin, Germany. ⁷Institute of Pathology, Medical University of Graz, Auenbruggerplatz 25, 8036 Graz, Austria. ⁸Charité-Universitätsmedizin, Charitéplatz 1, 10117 Berlin, Germany. ⁹Bayer Pharma AG, Müllerstraße 178, 13353 Berlin, Germany. ¹⁰UCL Cancer Institute, University College London, London WC1E 6BT, UK. ¹¹Department of Surgery and Cancer, Imperial College London, London W12 0NN, UK. ¹²Center for Biomarker Research in Medicine, Stiftingtalstrasse 5, 8010 Graz, Austria. ¹³Department of Pathology, Hospital Graz Süd-West, Göstinger Straße 22, 8020 Graz, Austria. ¹⁴Department of Surgery, Hospital Brothers of Charity Graz, Marschallgasse 12, 8020 Graz, Austria. ¹⁵CPO-Cellular Phenomics& Oncology, Berlin-Buch GmbH, Robert-Rössle-Str. 10, 13125 Berlin, Germany. ¹⁶Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, Ihnestrasse 73, D-14195 Berlin, Germany. ¹⁷Department of Immunology, Genetics and Pathology, SciLifeLab, Uppsala University, Box 815, SE-751 08 Uppsala, Sweden. ¹⁸Merck KGaA, Frankfurter Str. 250, 64293 Darmstadt, Germany. ¹⁹Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Tomtebodavägem 23A, Solna, Stockholm 17165, Sweden. ²⁰Boehringer Ingelheim RCV GmbH & Co KG, Dr. Boehringer-Gasse 5-11, A-1121 Wien, Austria. ²¹Eli Lilly and Company, Lilly Research Laboratories, Oncology Translational Research, Lilly Corporate Center, Indianapolis, Indiana 46285, USA. ²²Bayer Pharma AG, Global External Innovation & Alliances, Müllerstraße 178, 13353 Berlin, Germany. ²³Dahlem Centre for Genome Research and Medical Systems Biology, Fabeckstr. 60-62, 14195 Berlin, Germany. * These authors contributed equally to this work. ** These authors jointly supervised this work.

Figure A.1: List of authors of the OncoTrack publication.

A.2 Performance of anti-EGFR therapy outcome prediction

Table A.1: Performance of the SVM-based OT mini-classifier classifier in individual external cohorts.





Cohorts	Classifier	Samples analyzed	TP	FP	FN	TN	Sensitivity	Specificity	Balanced accuracy
 (n=48)	Mutation status in <i>KRAS/BRAF/NRAS</i>	whole cohort (n=48)	10	6	4	28	0.71	0.82	0.77
	Mutations status in <i>KRAS</i> codon 12/13	whole cohort (n=48)	11	19	3	15	0.79	0.44	0.61
	OT mini classifier cross validation	whole cohort (n=48)	14	2	0	32	1.00	0.94	0.97
 (n=60)	Mutation status in <i>KRAS/BRAF/NRAS</i>	whole cohort (n=60)	12	8	3	37	0.80	0.82	0.81
	Mutation status in <i>KRAS</i> codon 12/13	whole cohort (n=60)	12	20	3	25	0.80	0.56	0.68
	OT mini classifier	whole cohort (n=60)	14	5	2	39	0.88	0.89	0.88
	OT mini classifier	Wild-type for <i>KRAS</i> codon 12/13 (n=32)	11	3	1	17	0.92	0.85	0.88
	OT mini classifier	Wild-type for all <i>KRAS/BRAF/NRAS</i> (n=20)	11	3	1	5	0.92	0.63	0.77
 (n=36)	Mutation status in <i>KRAS/BRAF/NRAS</i>	whole cohort (n=36)	5	11	3	17	0.63	0.61	0.62
	Mutation status in <i>KRAS</i> codon 12/13	whole cohort (n=36)	5	20	3	8	0.63	0.29	0.46
	OT mini classifier	whole cohort (n=36)	7	3	1	25	0.88	0.89	0.88
	OT mini classifier	Wild-type for <i>KRAS</i> codon 12/13 (n=25)	5	1	0	19	1.00	0.95	0.98
	OT mini classifier	Wild-type for all <i>KRAS/BRAF/NRAS</i> (n=16)	5	1	0	10	1.00	0.91	0.95
	Mutation status in <i>KRAS/BRAF/NRAS</i>	whole cohort excluding SD cases (n=32)	3	11	1	17	0.75	0.61	0.68
	Mutation status in <i>KRAS</i> codon 12/13	whole cohort excluding SD cases (n=32)	3	20	1	8	0.75	0.29	0.52
	OT mini classifier	whole cohort excluding SD cases (n=32)	4	3	0	25	1.00	0.89	0.95
	OT mini classifier	Wild-type for <i>KRAS</i> codon 12/13 & excluding SD cases (n=23)	3	1	0	19	1.00	0.95	0.98
OT mini classifier	Wild-type for all <i>KRAS/BRAF/NRAS</i> & excluding SD cases (n=14)	3	1	0	10	1.00	0.91	0.95	
 (n=68)	Mutation status in <i>KRAS</i> codon 12/13	Cases with mutation information (n=59)	20	19	2	18	0.91	0.49	0.70
	OT mini classifier	whole cohort (n=68)	17	6	8	37	0.68	0.86	0.77
	OT mini classifier	Wild-type for <i>KRAS</i> codon 12/13 (n=39)	15	3	5	16	0.75	0.84	0.80
	Mutation status in <i>KRAS</i> codon 12/13	cases with mutation information & excluding SD cases (n=42)	5	19	0	18	1.00	0.49	0.74
	OT mini classifier	whole cohort excluding SD cases (n=49)	5	6	1	37	0.83	0.86	0.85
OT mini classifier	Wild-type for <i>KRAS</i> codon 12/13 & excluding SD cases (n=24)	4	3	1	16	0.80	0.84	0.82	

Table A.2: Performance of the SVM-based OT mini-classifier in merged external cohorts.

Merged cohorts	Classifier	Samples analyzed	TP	FP	FN	TN	Sensitivity	Specificity	Balanced accuracy
EPO+NV+KF (n=164)	Mutation status in <i>KRAS/BRAF/NRAS</i>	Cases with mutation information (n=96)	17	19	6	54	0.74	0.74	0.74
	Mutation status in <i>KRAS</i> codon 12/13	Cases with mutation information (n=155)	37	59	8	51	0.82	0.46	0.64
	OT mini classifier	whole cohort (n=164)	38	14	11	101	0.78	0.88	0.83
	OT mini classifier	Wild-type for <i>KRAS</i> codon 12/13 (n=96)	31	7	6	52	0.84	0.88	0.86
	OT mini classifier	Wild-type for all <i>KRAS/BRAF/NRAS</i> mutations (n=36)	16	4	1	15	0.94	0.79	0.87

A.3 Expression pattern of TFs in MB

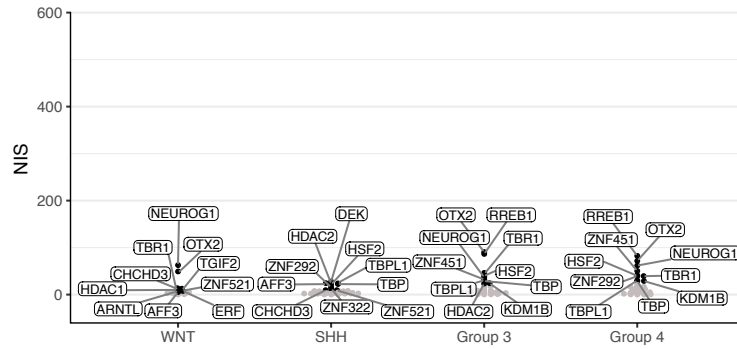


Figure A.2: NIS of TFs downregulated in one subgroup.

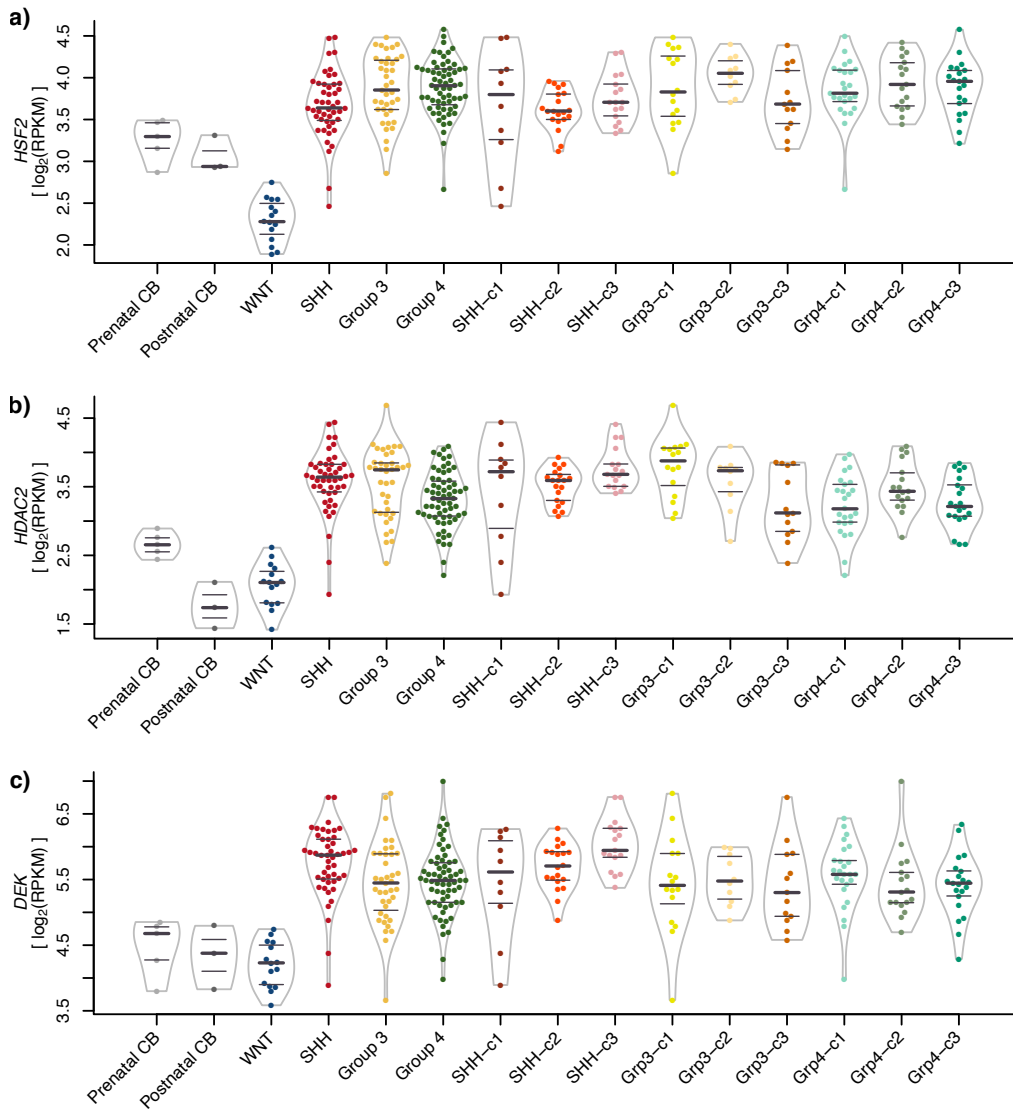


Figure A.3: Expression profile of *HSF2*, *DEK*, and *HDAC2* in MB. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A Appendix

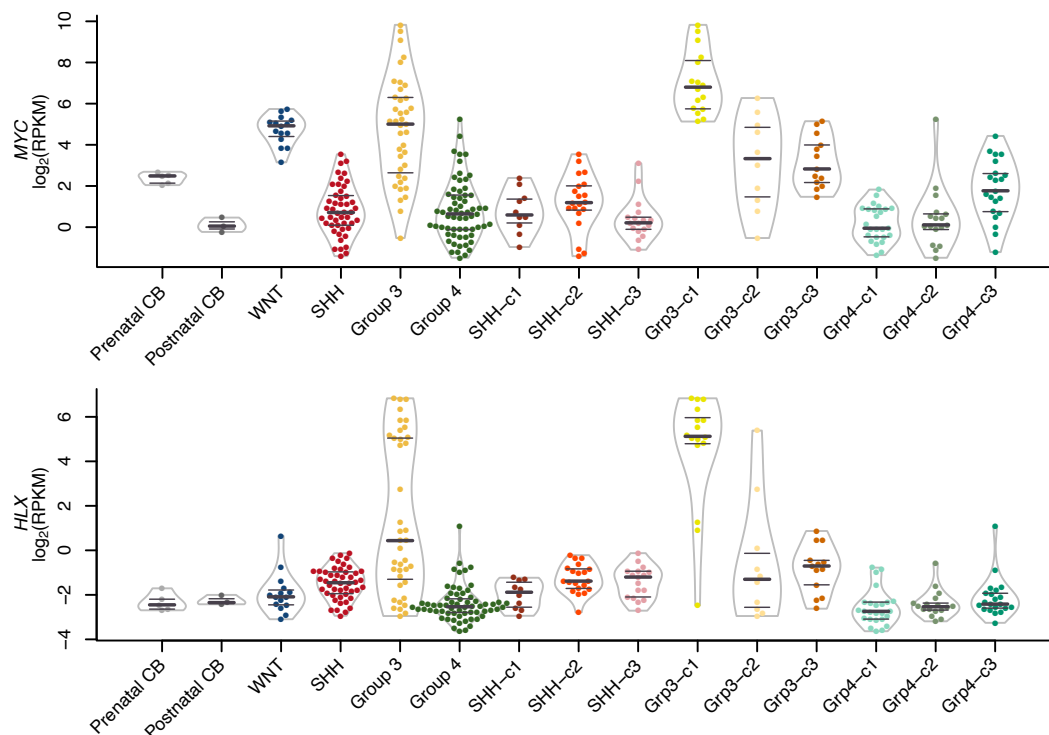


Figure A.4: Expression profile of *MYC* and *HLX* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

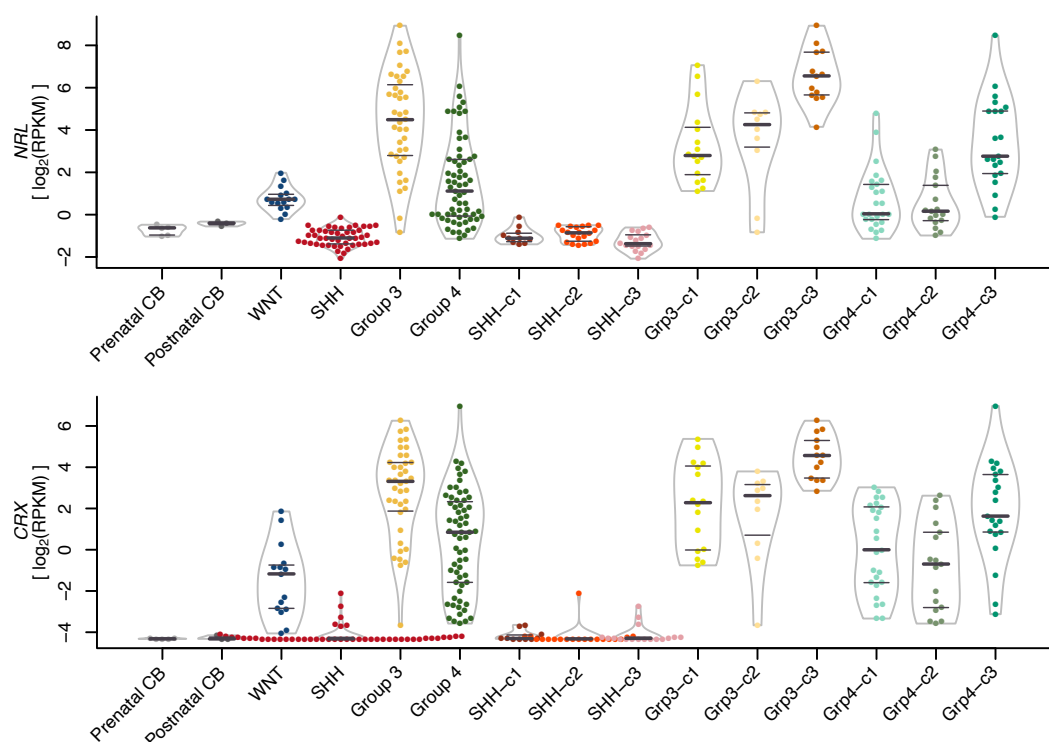


Figure A.5: Expression profile of *NRL* (top) and *CRX* (bottom) in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A.3 Expression pattern of TFs in MB

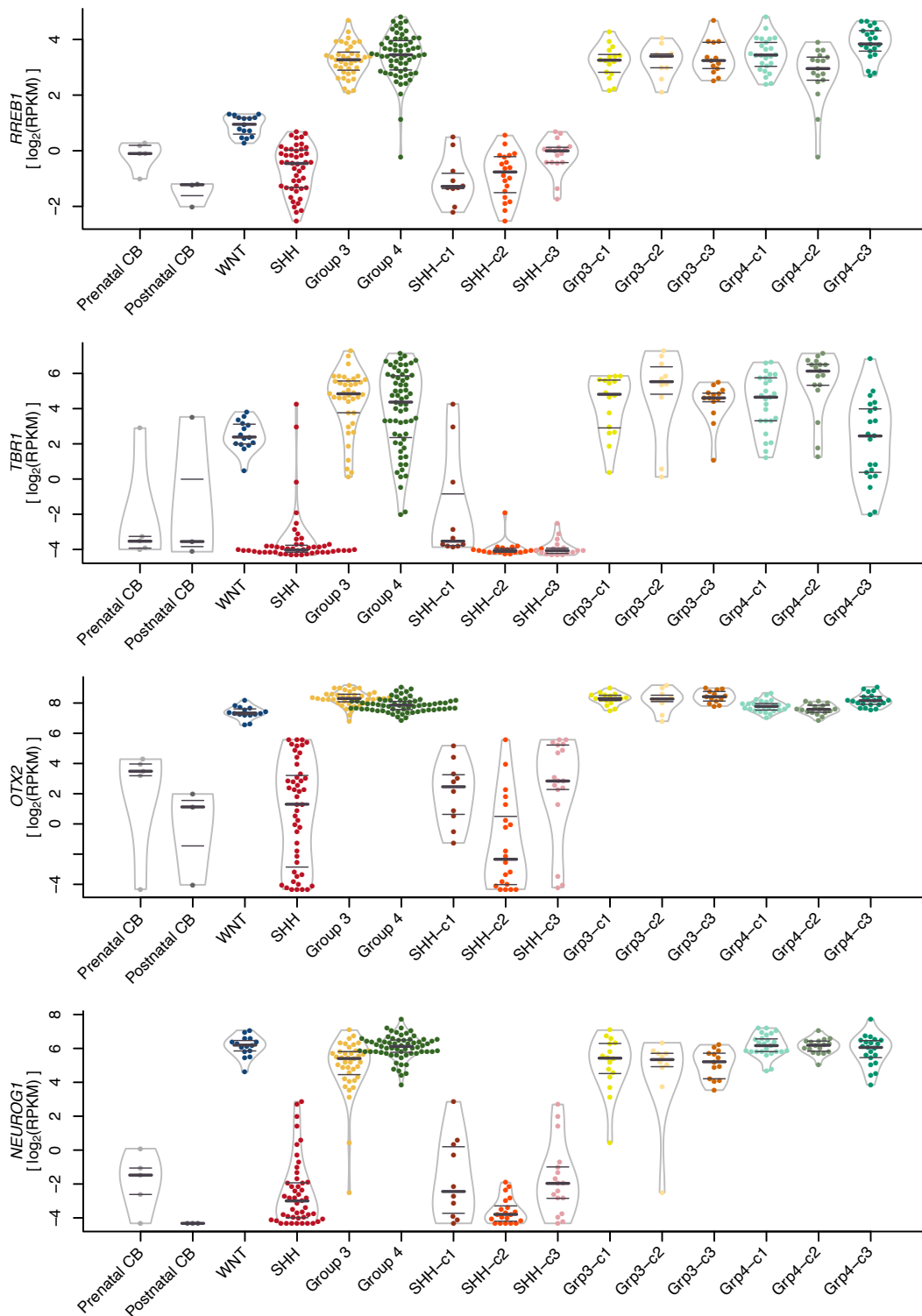


Figure A.6: Expression profile of *OTX2*, *RREB1*, *NEUROG1* and *TBR1* in MB. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

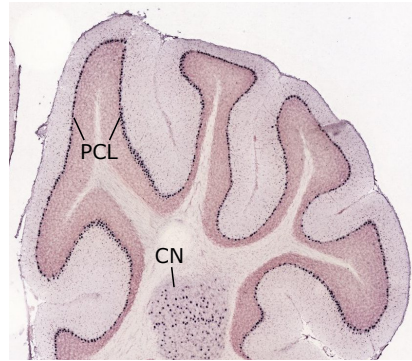


Figure A.7: ISH of *RREB1* in P56 mice cerebellum. CN - cerebellar nuclei. PCL - Purkinje cell layer. Image credit for ISH: Allen Institute [201]. Labels were added to image.

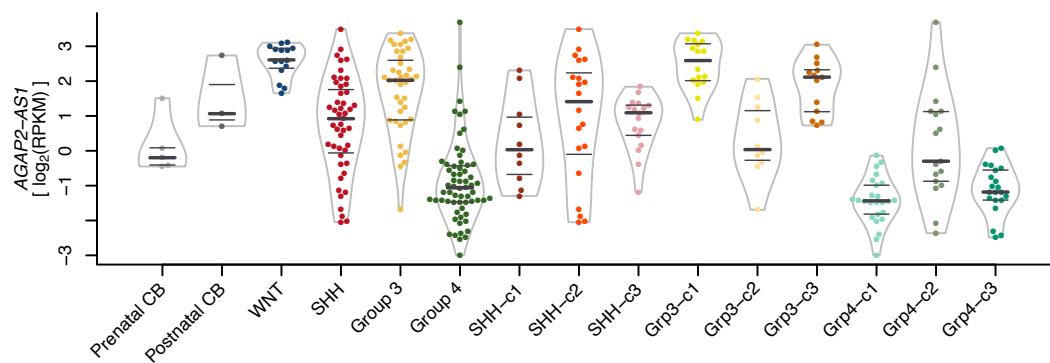


Figure A.8: Expression profile of *AGAP2-AS1* and in MB. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

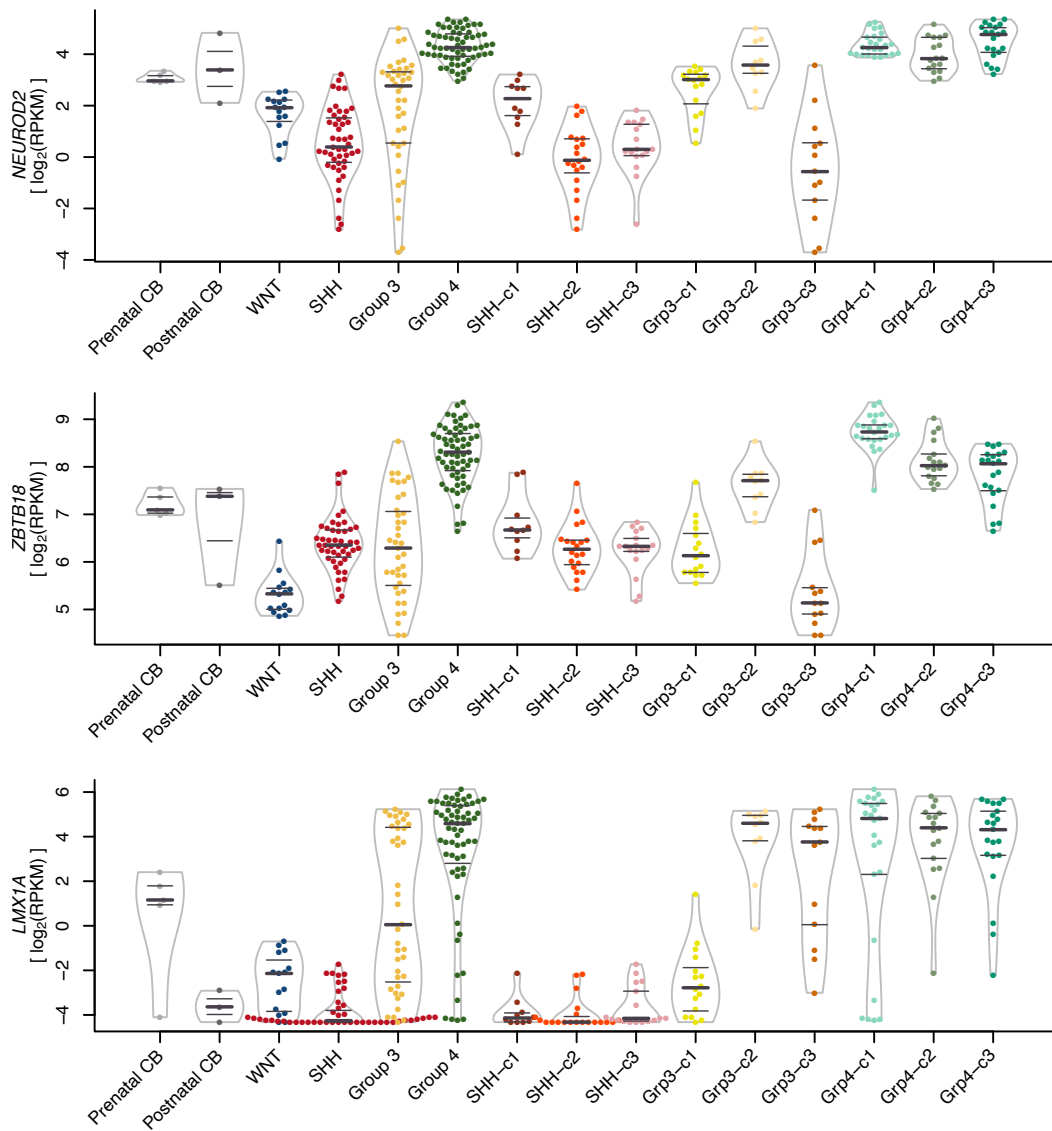


Figure A.9: xpression profiles of *NEUROD2*, *ZBTB18*, and *LMX1A* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

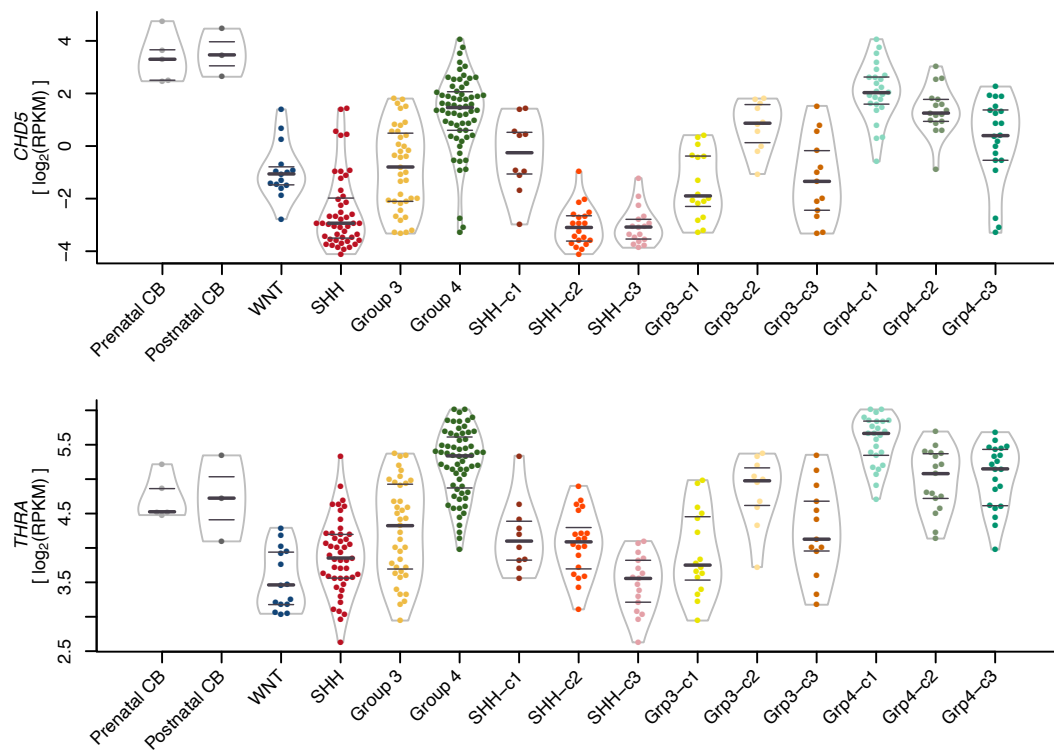


Figure A.10: Expression profile of *CHD5*, *THRA* in MB. Violin plots show expression distribution. 25%, 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A.3 Expression pattern of TFs in MB

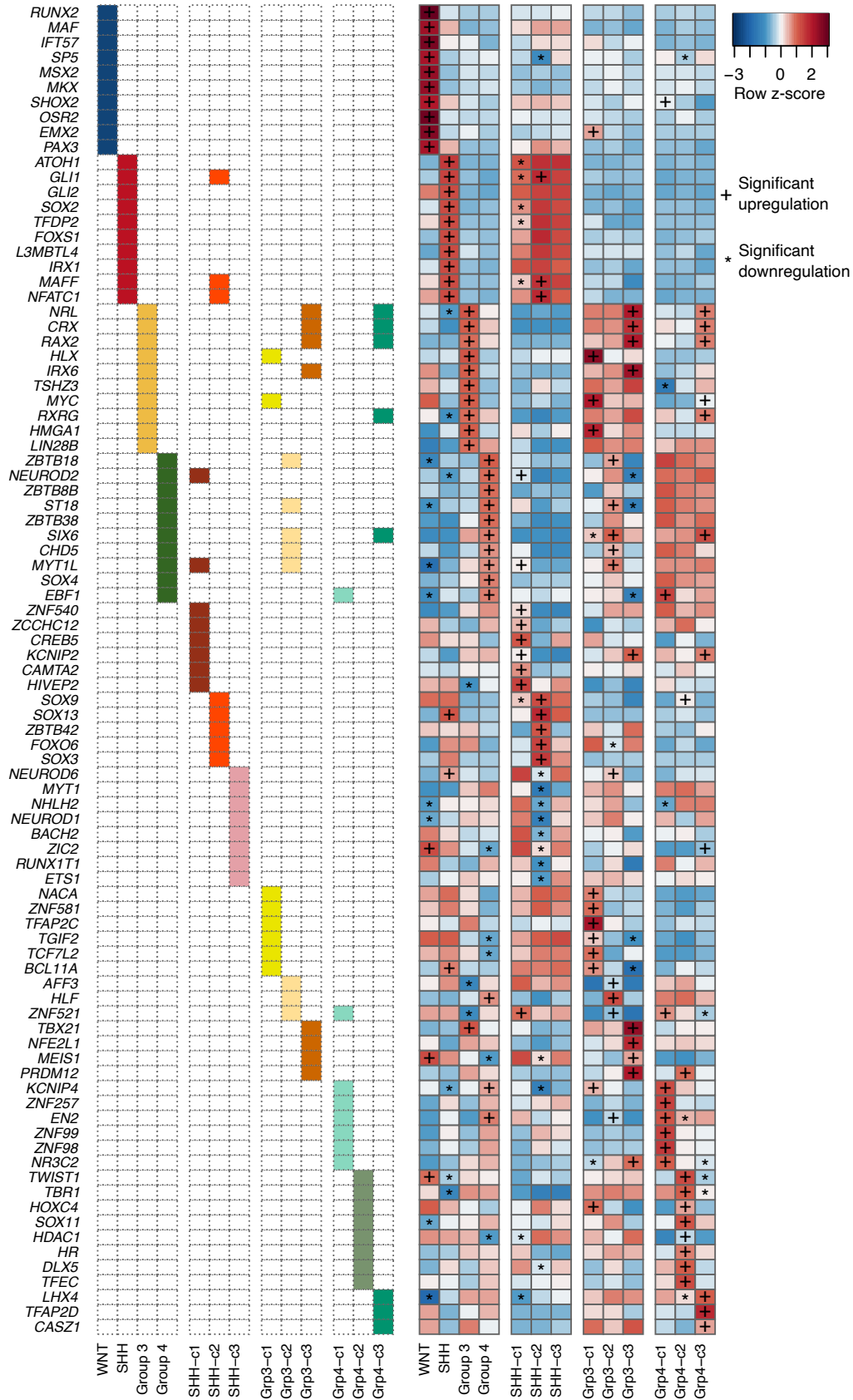


Figure A.11: Expression pattern of TFs with highest NIS in subgroups and subclusters. **Left)** Coloured rectangles annotate top-ranked TFs per subgroups and subclusters. Different colours relate to individual subgroups and subclusters, as indicated at the bottom. **Right)** Heatmap shows meant expression in subgroups and subclusters. Significant up- and downregulation among subgroups and subclusters within a subgroup are indicated.

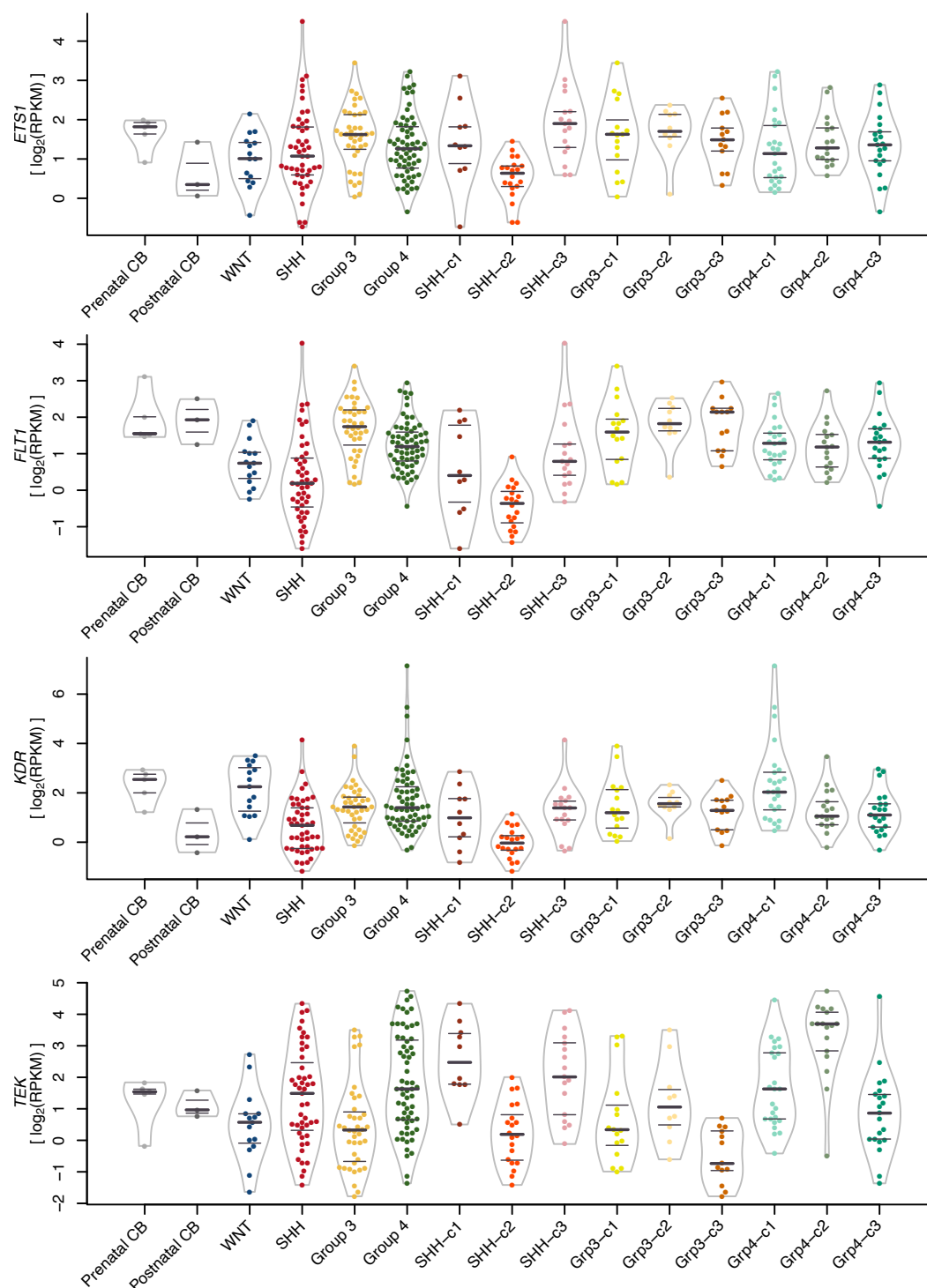


Figure A.12: Expression profiles of *ETS1*, *FLT1*, *KDR*, and *TEK* in MB. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

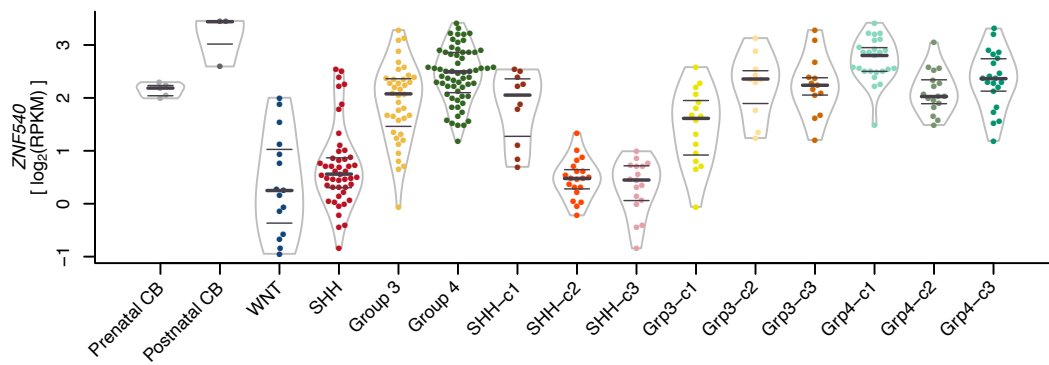


Figure A.13: Expression profile of *ZNF540* in MB. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A Appendix

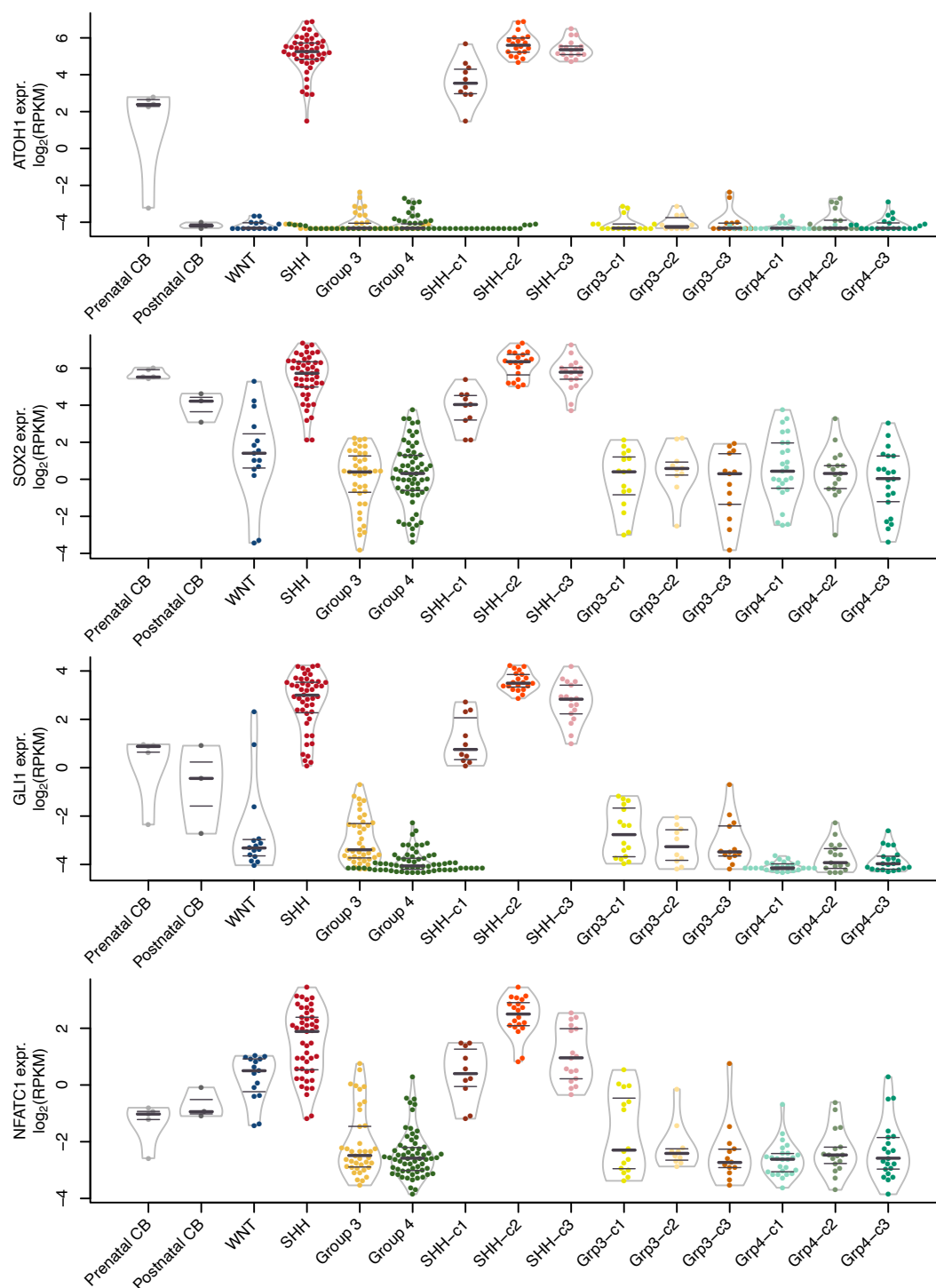


Figure A.14: Expression profile of *ATOH1*, *SOX2*, *GLI1*, and *NFATC1* in MB. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

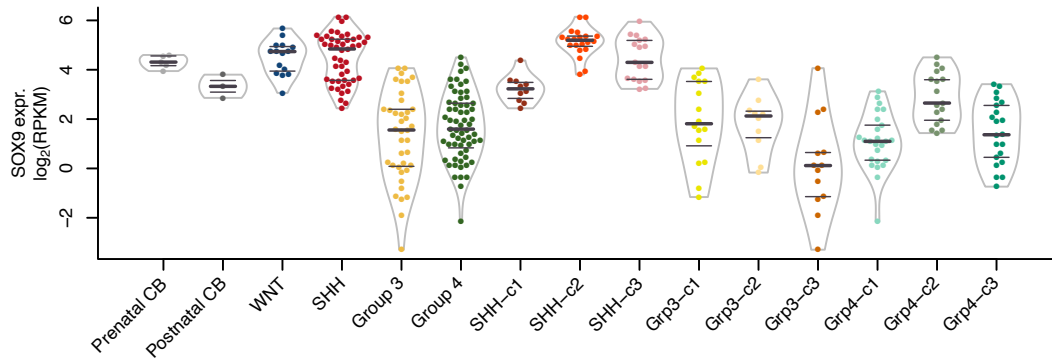


Figure A.15: Expression profile of *SOX9* in MB. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

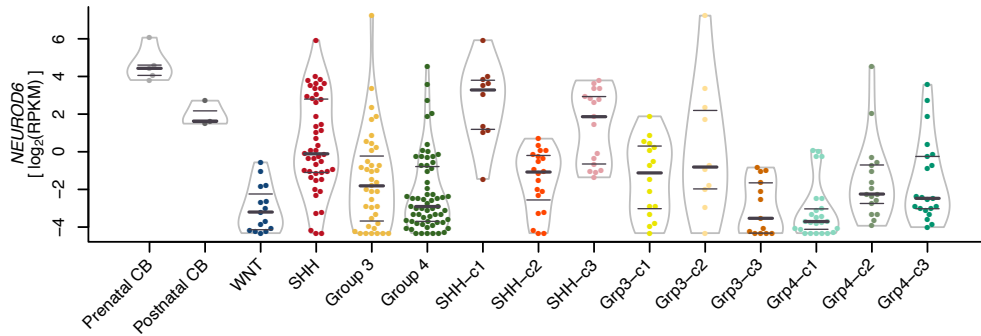


Figure A.16: Expression profile of *NEUROD6* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

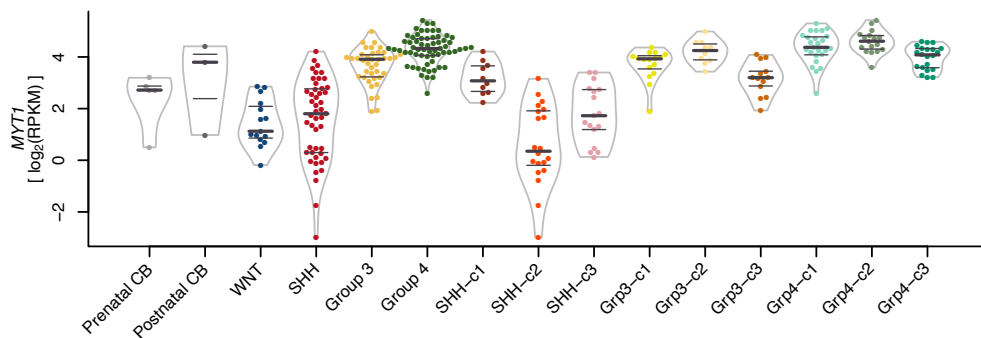


Figure A.17: Expression profile of *MYT1* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A Appendix

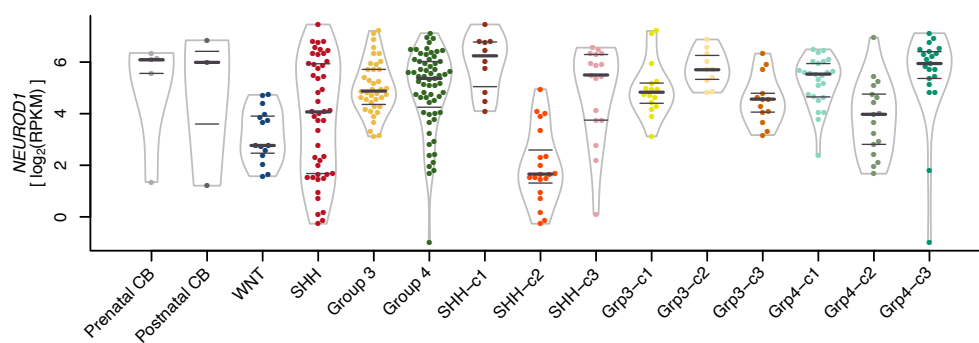


Figure A.18: Expression profile of *NEUROD1* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

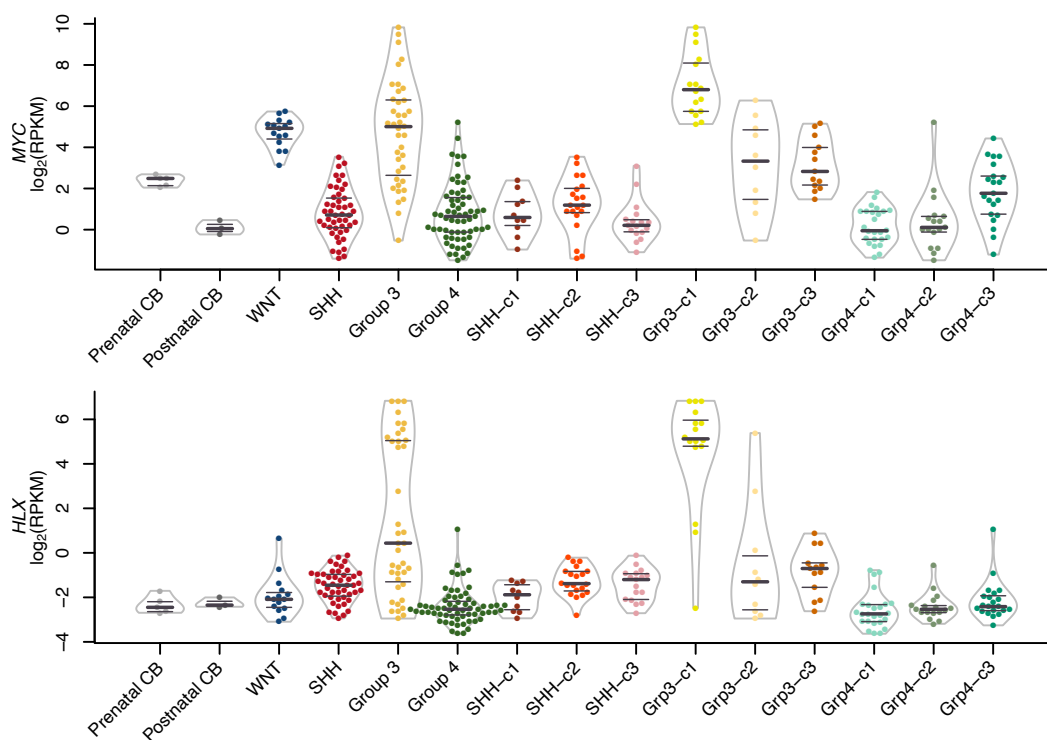


Figure A.19: Expression profile of *MYC* and *HLX* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

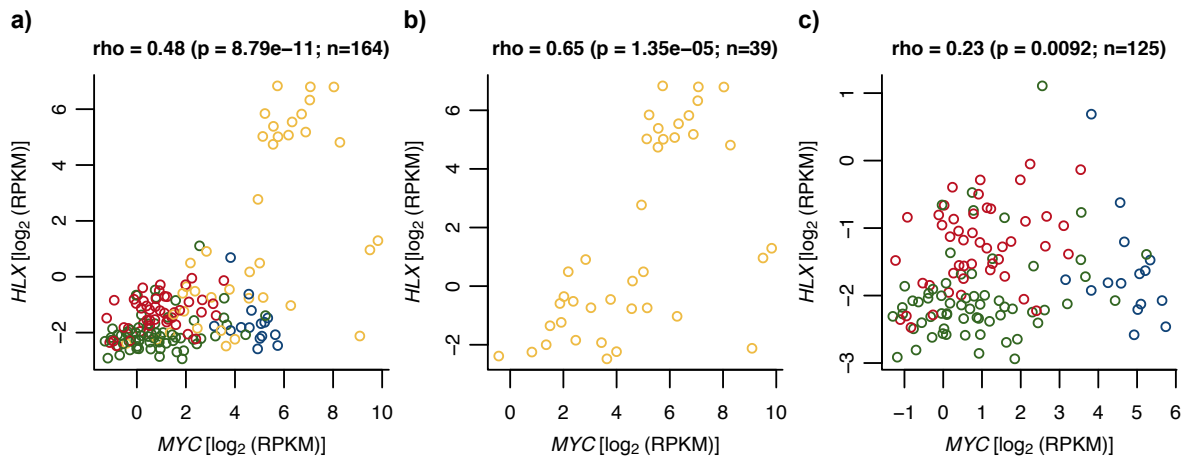


Figure A.20: Scatter plots of *MYC* and *HLX* expression in ICGC MB samples. **a)** Whole cohort. **b)** Group 3 MB. **c)** non-Group 3 MB. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green.

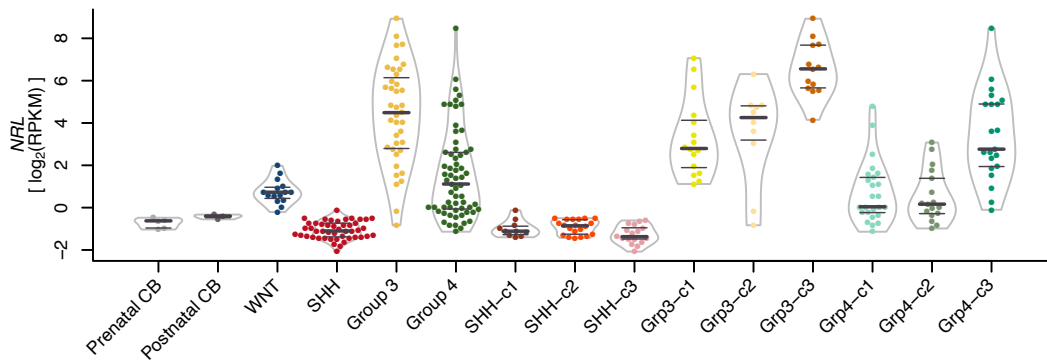


Figure A.21: Expression profile of *NRL* and *CRX* in MB. Violin plots show expression destitution. 25%, 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

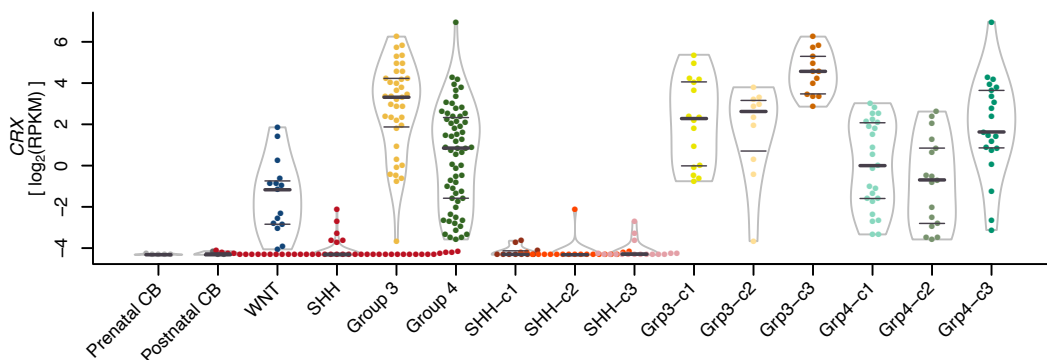


Figure A.22: Expression profile of *CRX* in MB. Violin plots show expression destitution. 25%, 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

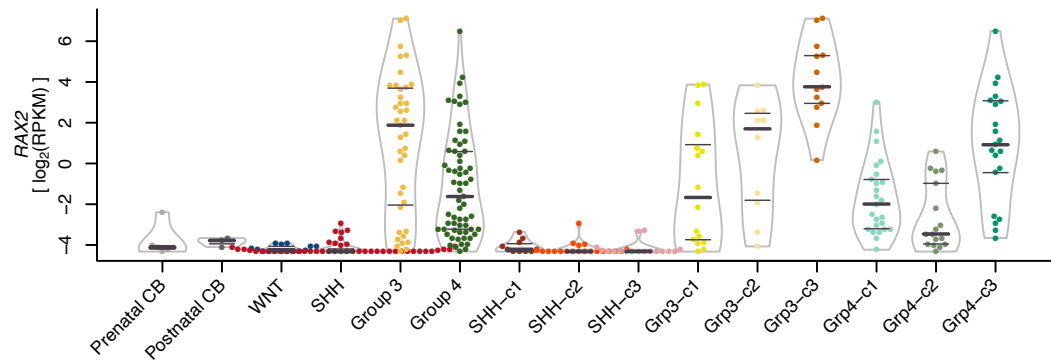


Figure A.23: Expression profile of *RAX2* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

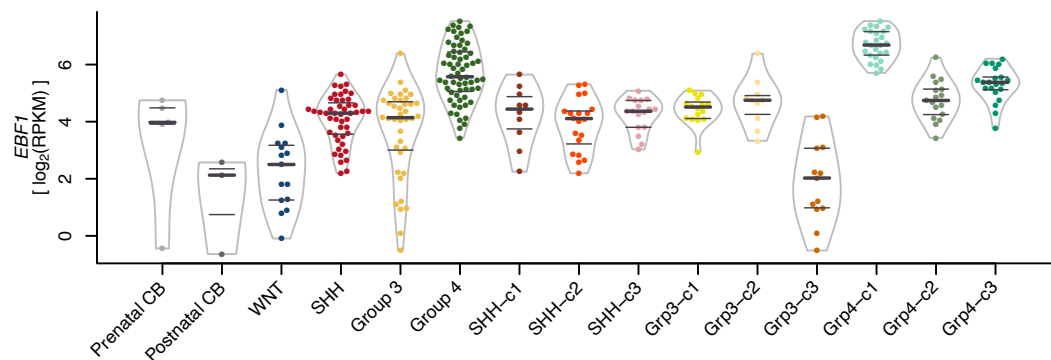


Figure A.24: Expression profile of *EBF1* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

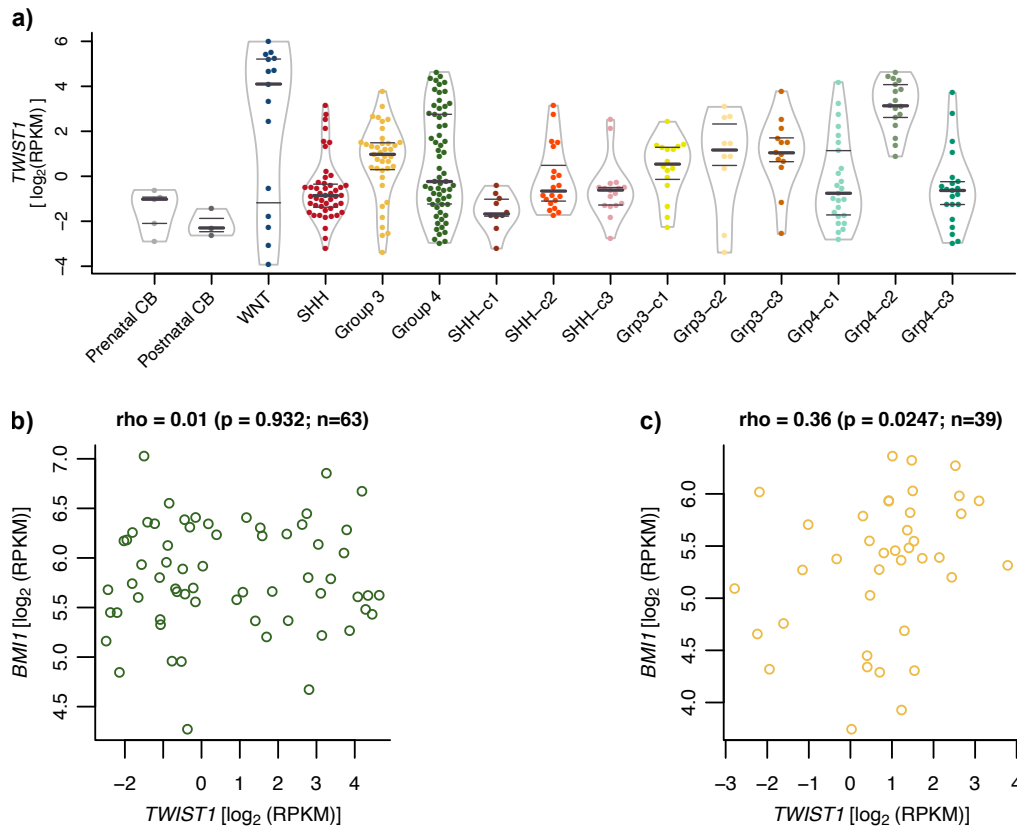


Figure A.25: **a)** Expression profile of *TWIST1* in MB. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots. **b+c)** Scatter plot of *TWIST1* and *BMI1* expression in b) Group 4 and c) Group 3 MB samples. Colours indicate MB subgroups as shown in panel a).

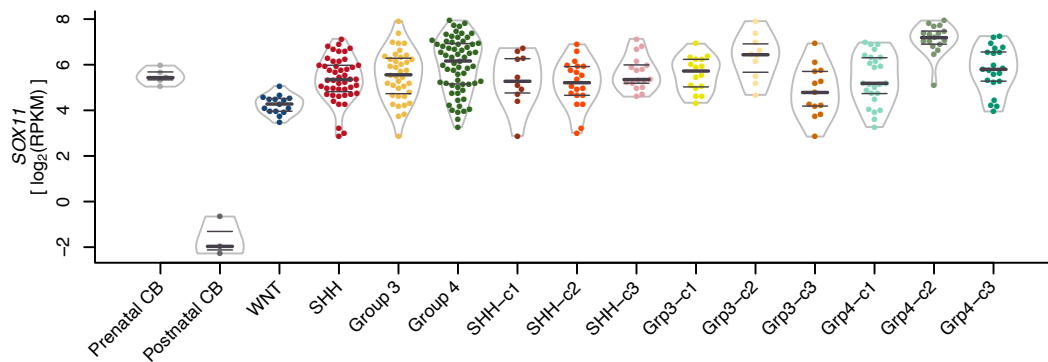


Figure A.26: Expression profile of *SOX11* in MB. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A Appendix

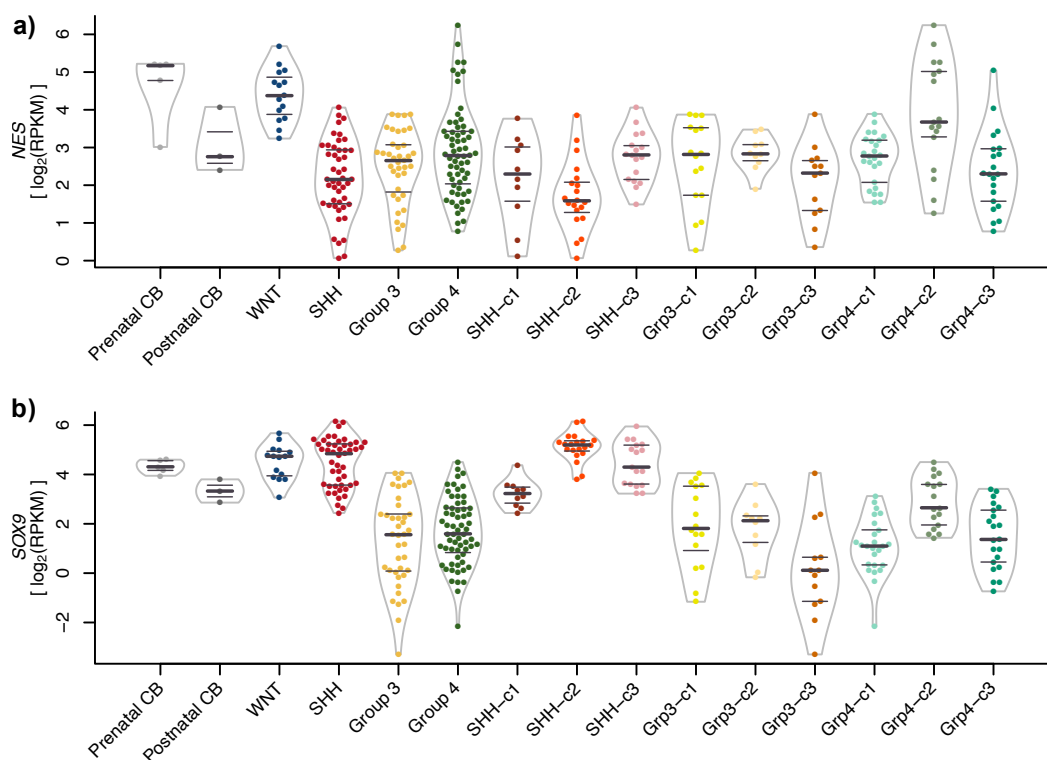


Figure A.27: Expression profile of **a) NES** and **b) SOX9** in MB. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

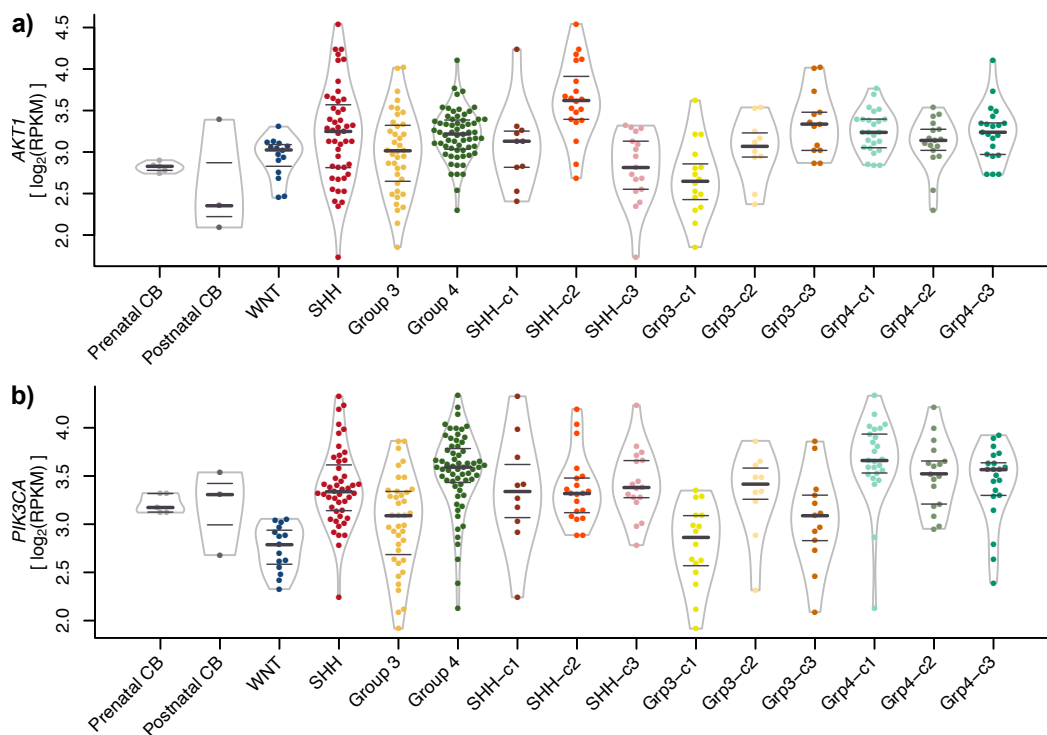


Figure A.28: Expression profile of **a) AKT1** and **b) PIK3CA** in MB. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

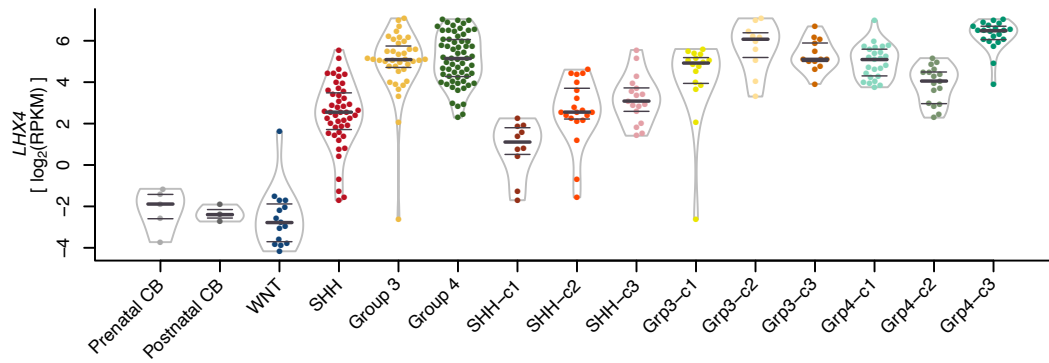


Figure A.29: Expression profile of *LHX4* in MB. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

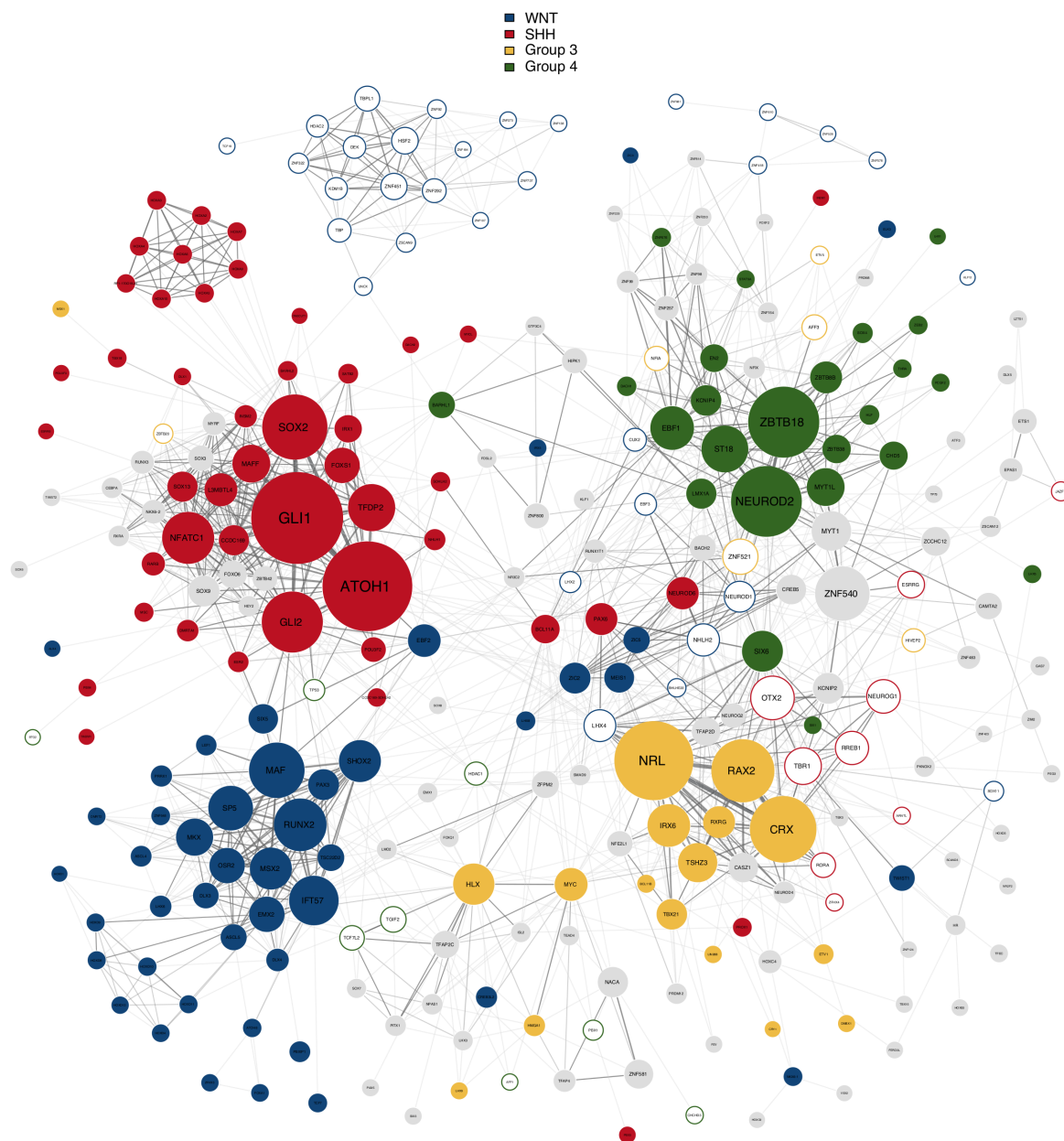


Figure A.30: Aggregated GRN in MB and differentially expressed TFs among subgroups. Caption as in Figure 5.46.

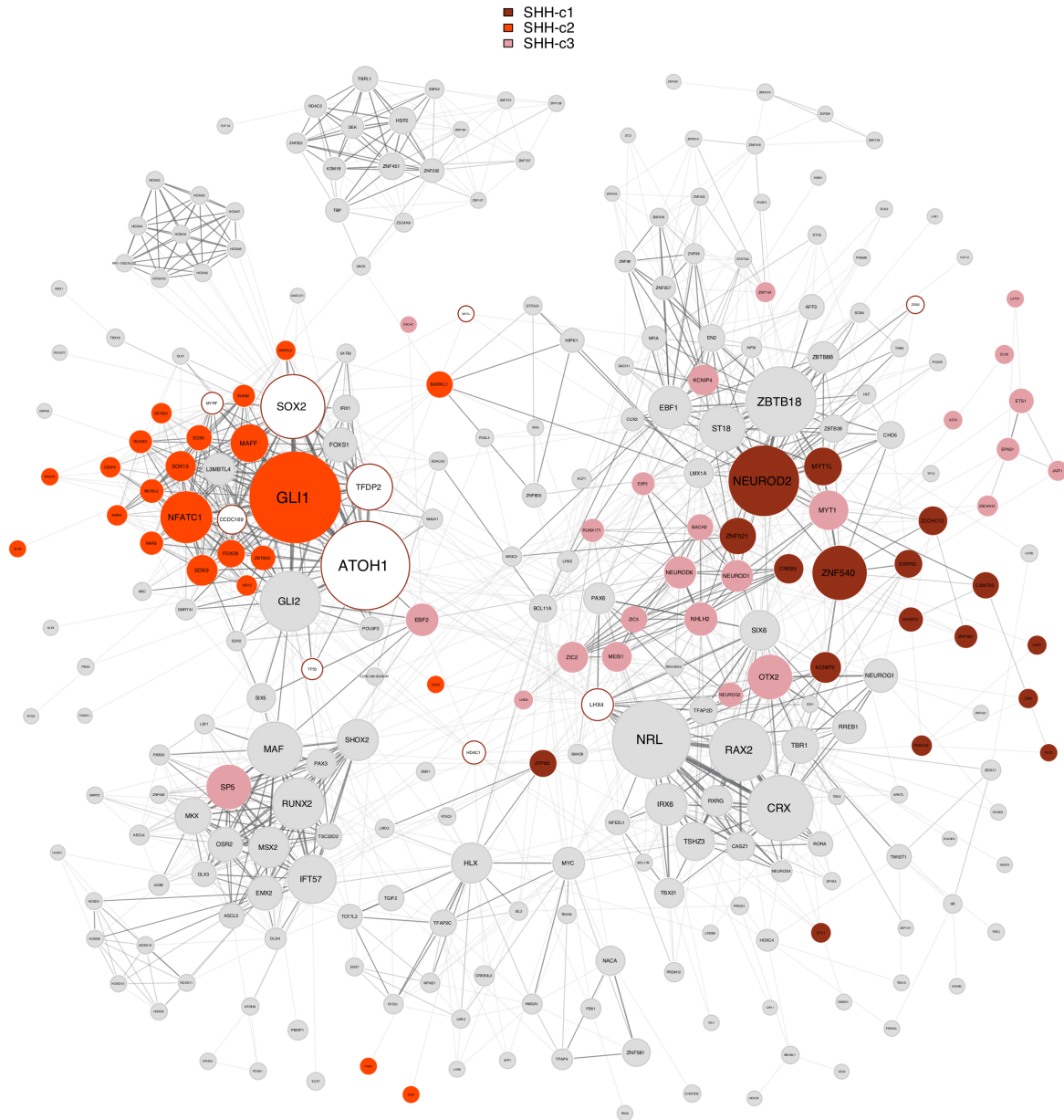


Figure A.31: Aggregated GRN in MB and differentially expressed TFs among SHH subclusters. Caption as in Figure 5.46.

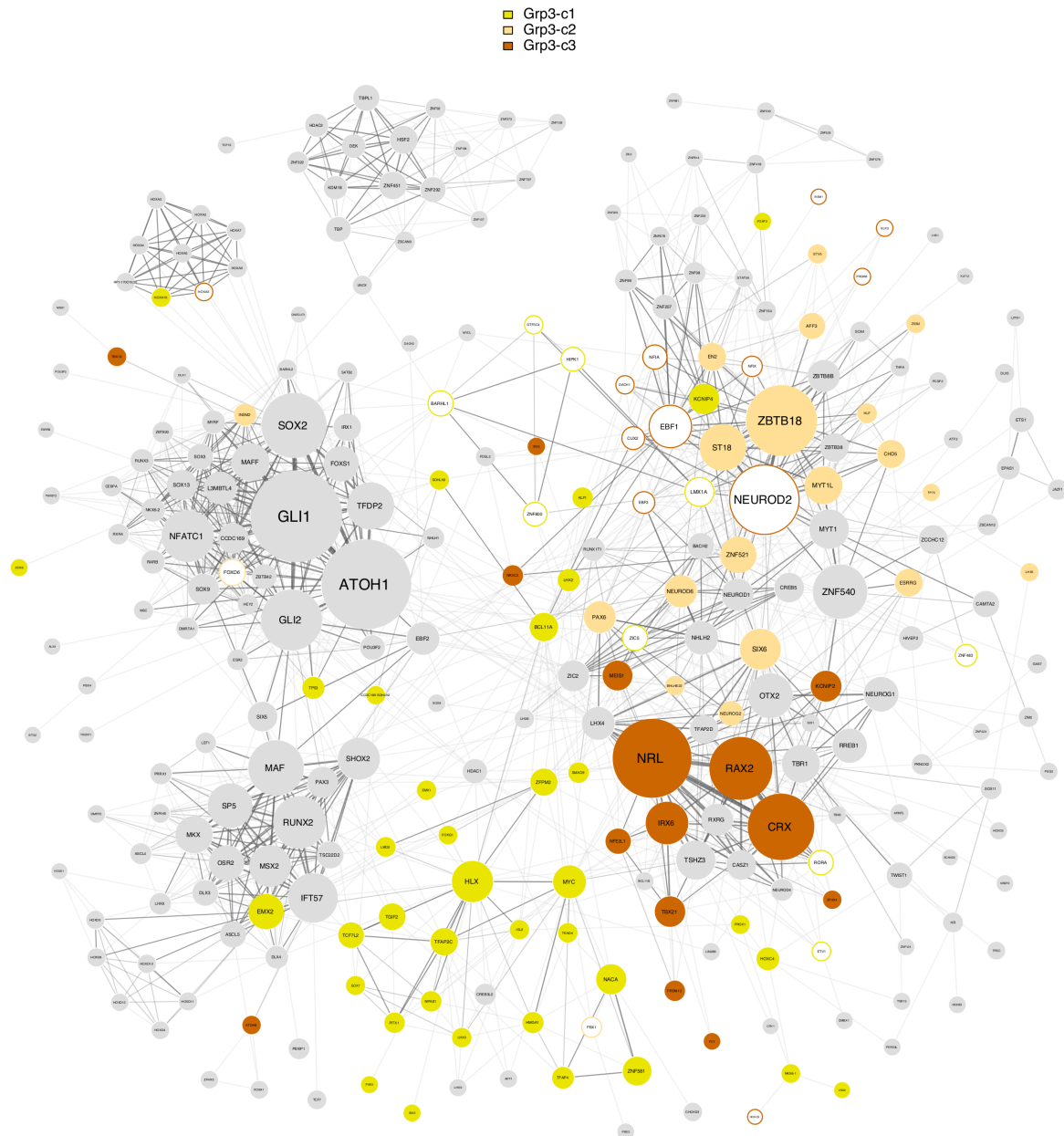


Figure A.32: Aggregated GRN in MB and differentially expressed TFs among Group 3 subclusters. Caption as in Figure 5.46.

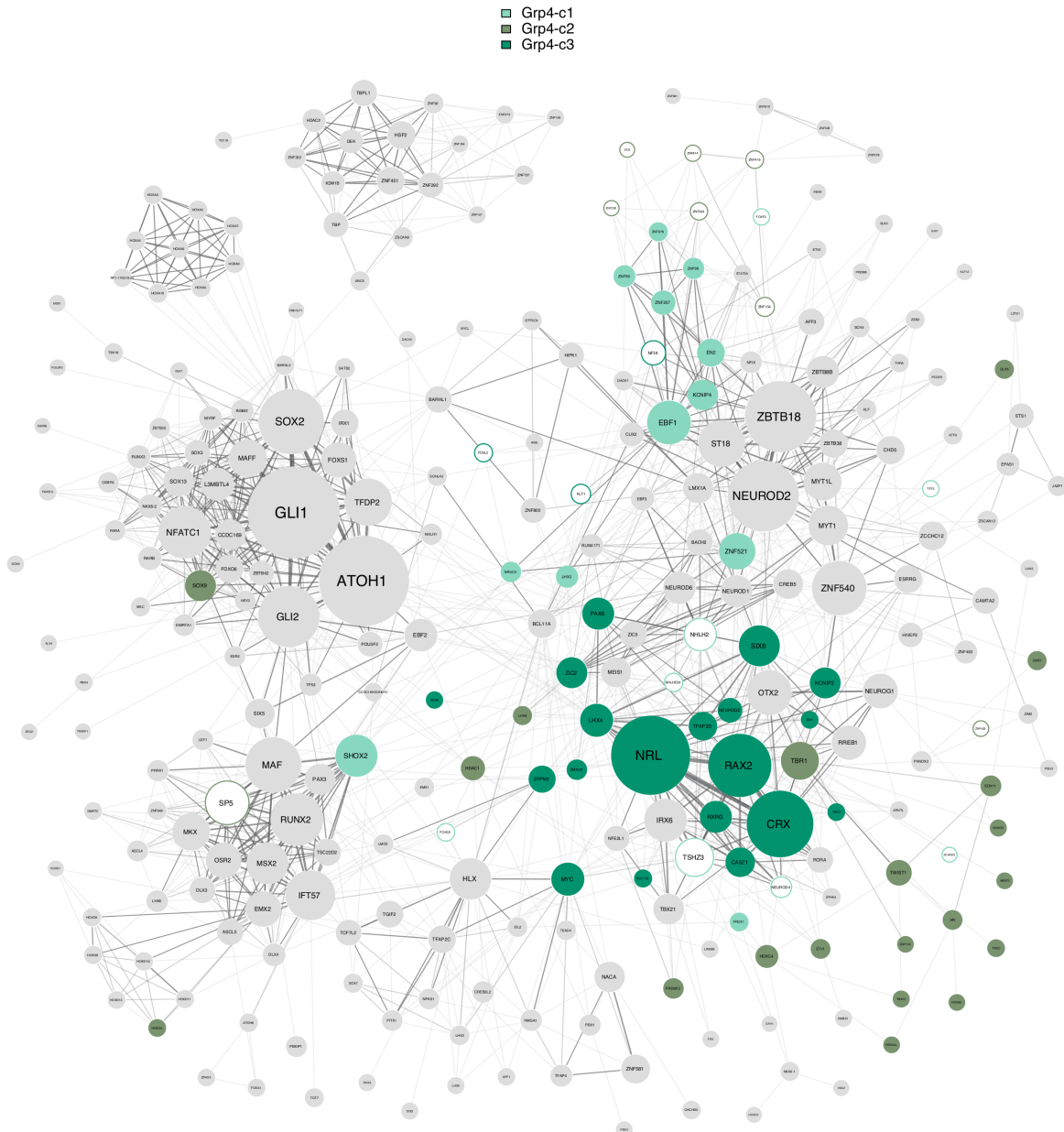


Figure A.33: Aggregated GRN in MB and differentially expressed TFs among Group 4 subclusters. Caption as in Figure 5.46.

A.4 Expression pattern of lnc genes and related coding genes

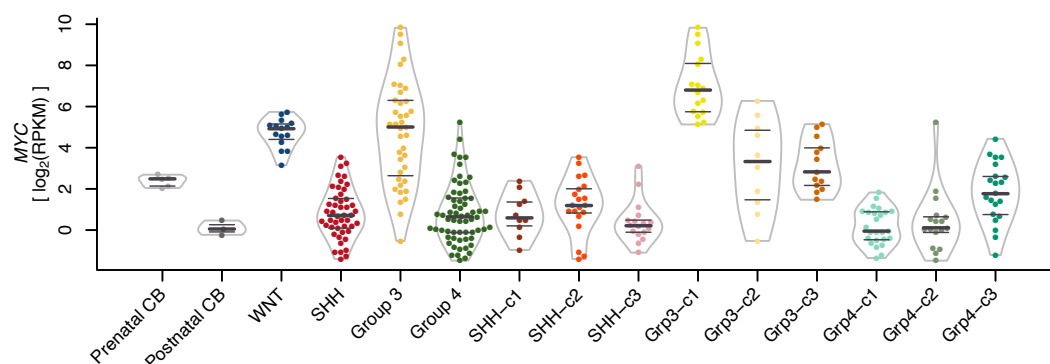


Figure A.34: Expression profile of MYC in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

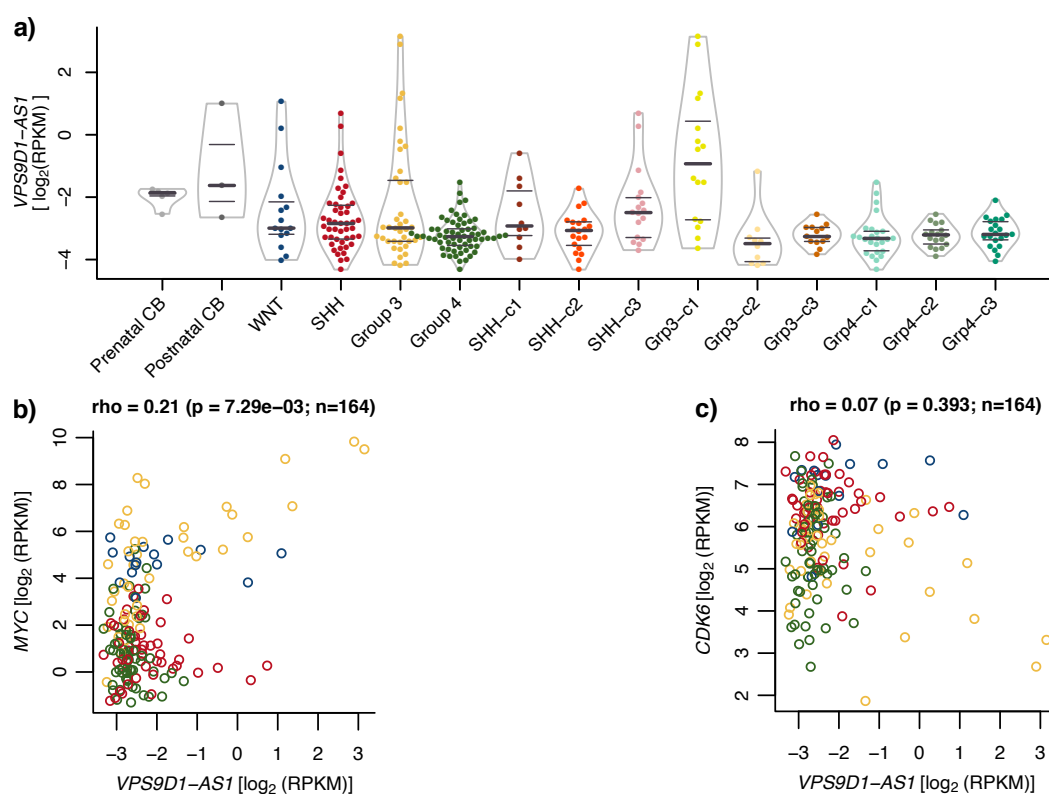


Figure A.35: Expression profiles of *VPS9D1-AS1* in MB. **a)** Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots. **b+c)** Scatter plot of b) *VPS9D1-AS1* and *MYC* and c) *VPS9D1-AS1* and *CDK6* expression in ICGC PedBrain MB samples. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green.

A.4 Expression pattern of lnc genes and related coding genes

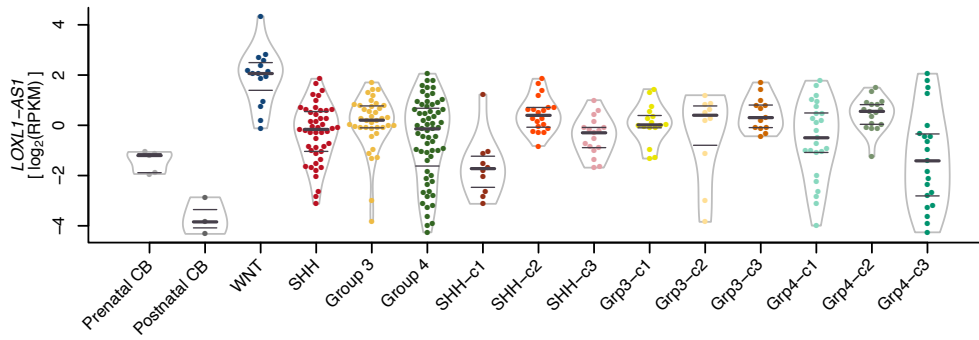


Figure A.36: Expression profile of *LOXL1-AS1* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

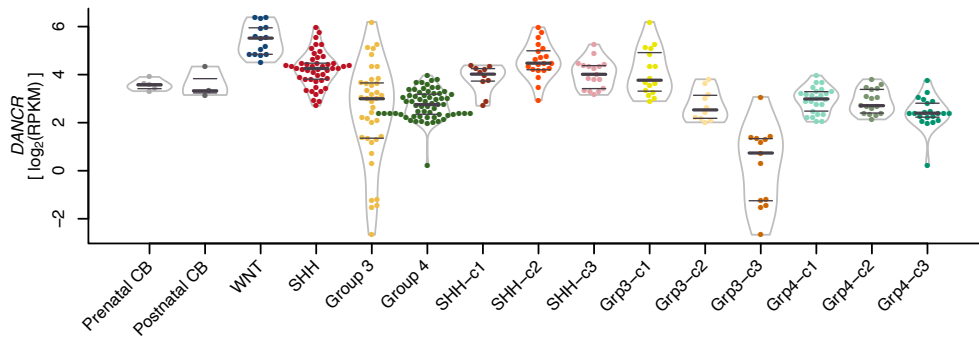


Figure A.37: Expression profile of *DANCR* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

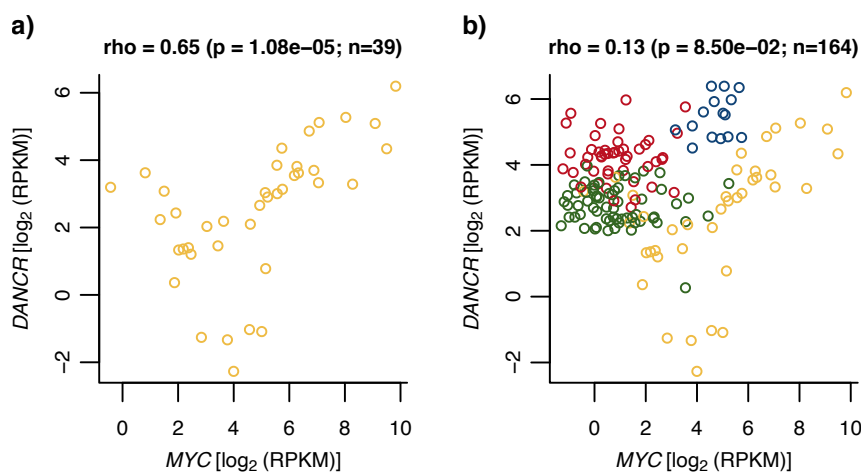


Figure A.38: Scatter plot of *MYC* and *DANCR* expression in ICGC MB samples. **a)** Group 3 tumours. **b)** All four subgroups. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green.

A Appendix

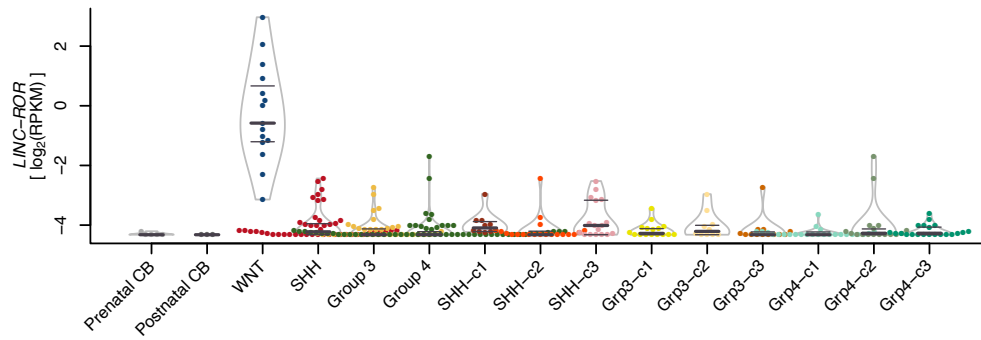


Figure A.39: Expression profile of *LINC-ROR* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

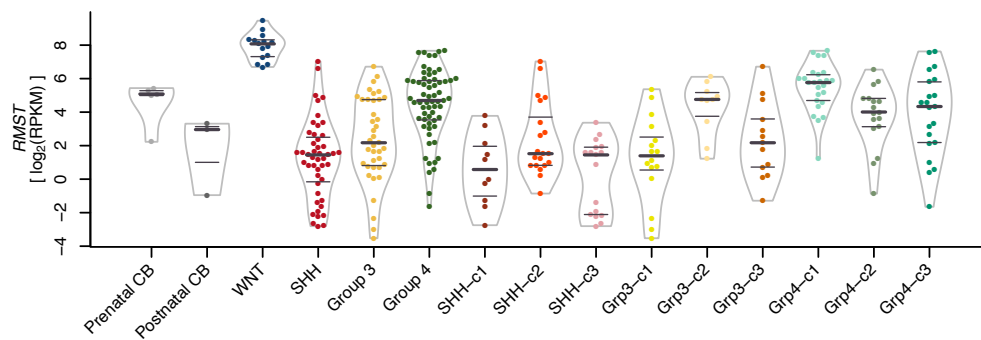


Figure A.40: Expression profile of *RMST* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A.4 Expression pattern of *lnc* genes and related coding genes

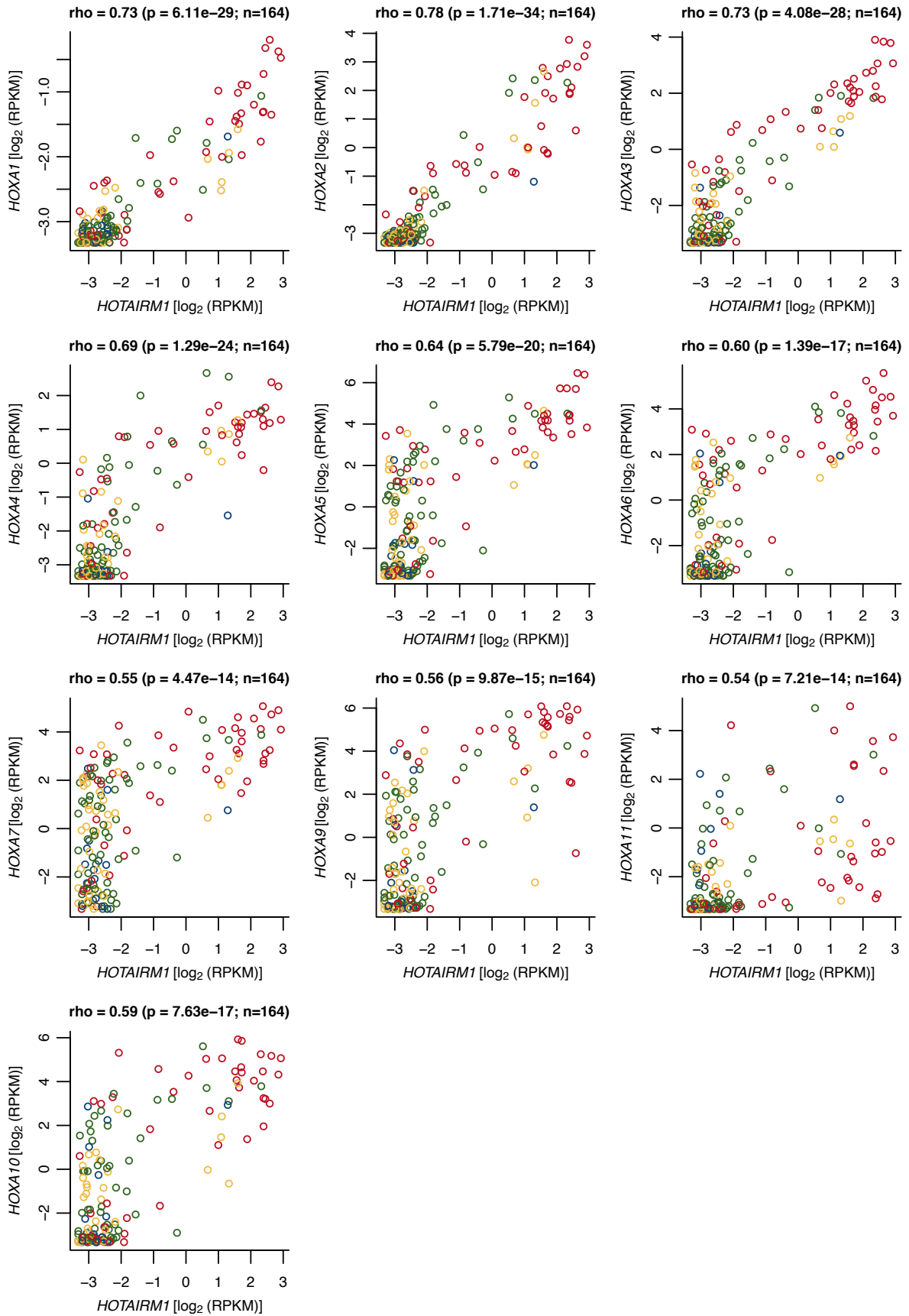


Figure A.41: Correlation of *HOTAIRM1* with *HOXA* genes. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green.

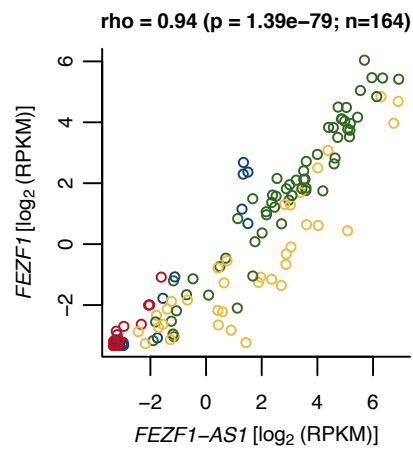


Figure A.42: Scatter plot of *FEZF1-AS1* and *FEZF1* expression in ICGC MB samples. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

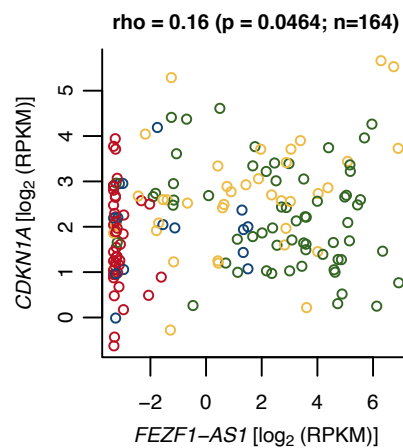


Figure A.43: Scatter plot of *FEZF1-AS1* and *CDKN1A* (P21) expression in ICGC MB samples. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

A.4 Expression pattern of lnc genes and related coding genes

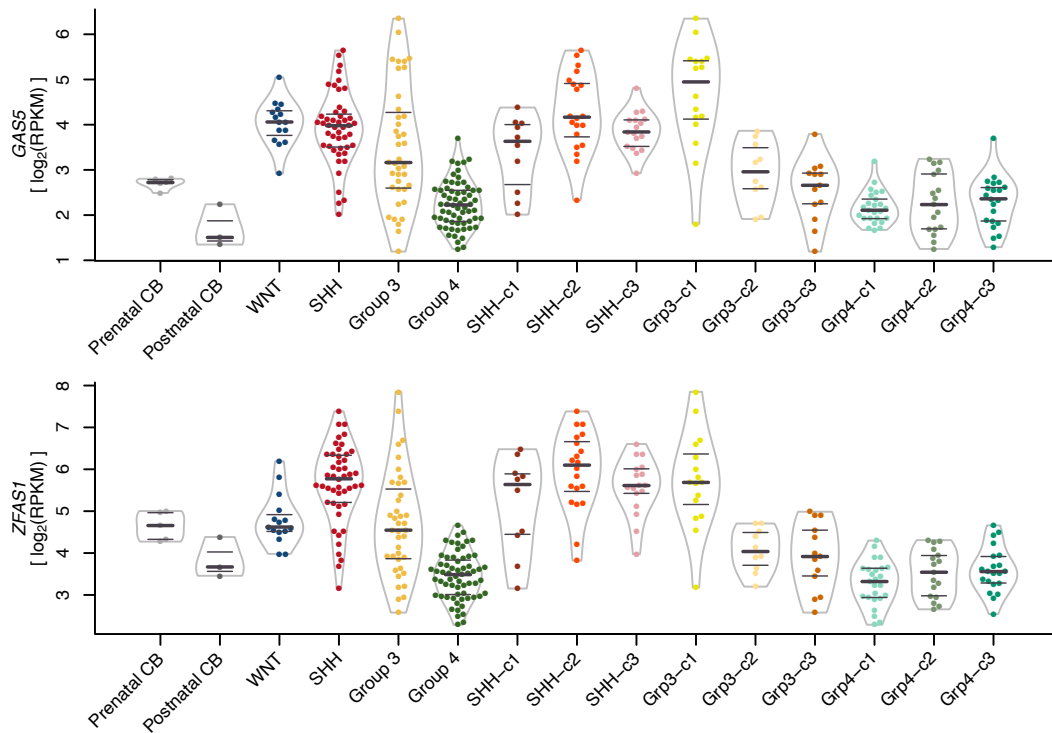


Figure A.44: Expression profile of *GAS5* and *ZFAS1* in MB. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

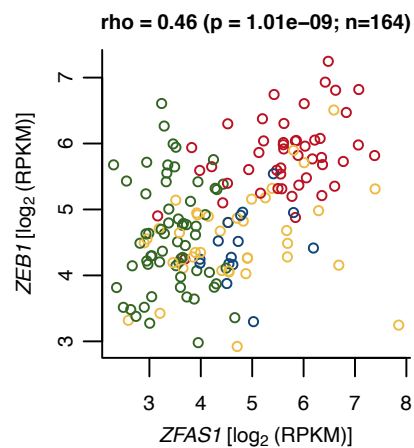


Figure A.45: Scatter plot of *ZFAS1* and *ZEB1* expression in ICGC MB samples. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

A Appendix

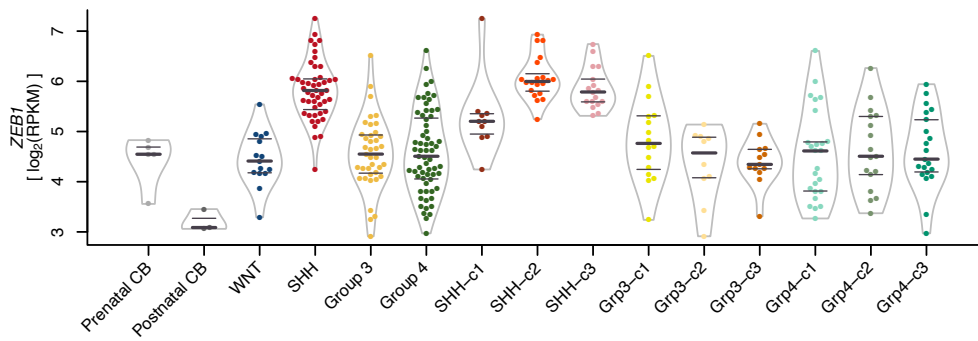


Figure A.46: Expression profile of *ZEB1* in MB. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

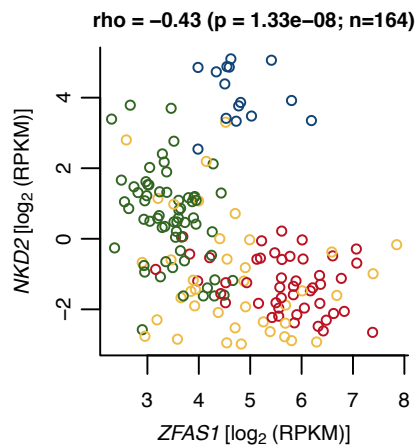


Figure A.47: Scatter plot of *ZFAS1* and *NKD2* expression in ICGC MB samples. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

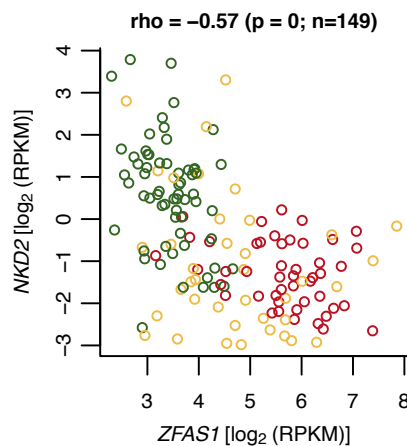


Figure A.48: Scatter plot of *ZFAS1* and *NKD2* expression in ICGC non-WNT MB samples. Colours indicate MB subgroups: SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

A.4 Expression pattern of *lnc* genes and related coding genes

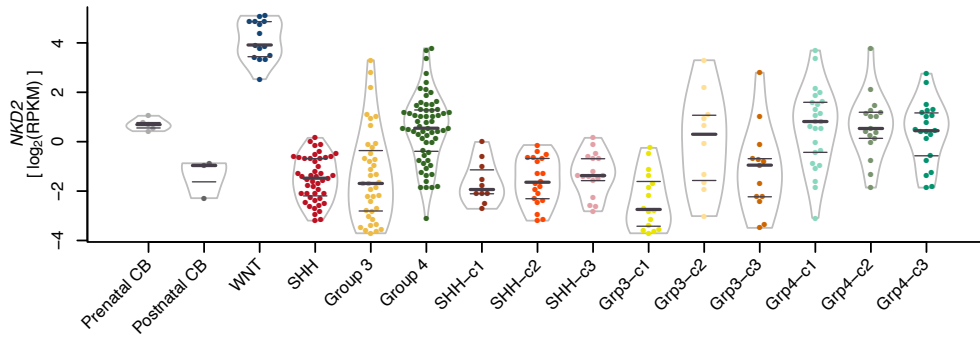


Figure A.49: Expression profile of *ZEB1* in MB. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

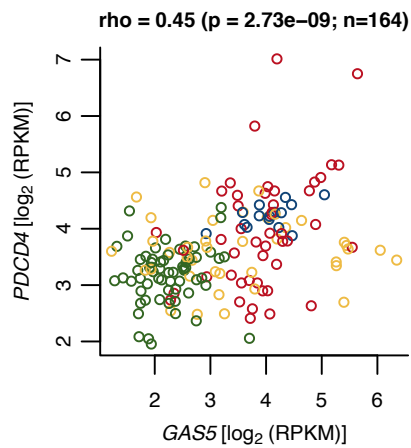


Figure A.50: Scatter plot of *GAS5* and *PDCD4* expression in ICGC MB samples. Colours indicate MB subgroups: SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

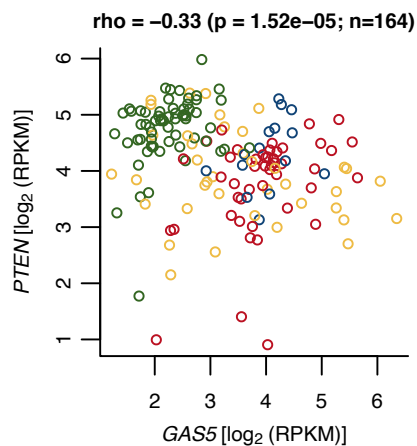


Figure A.51: Scatter plot of *GAS5* and *PTEN* expression in ICGC MB samples. Colours indicate MB subgroups: SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

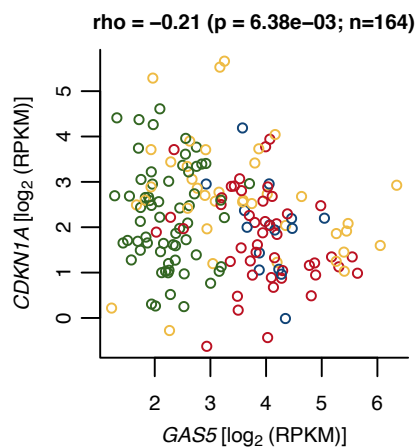


Figure A.52: Scatter plot of *GAS5* and *CDKN1A* expression in ICGC MB samples. Colours indicate MB subgroups: SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

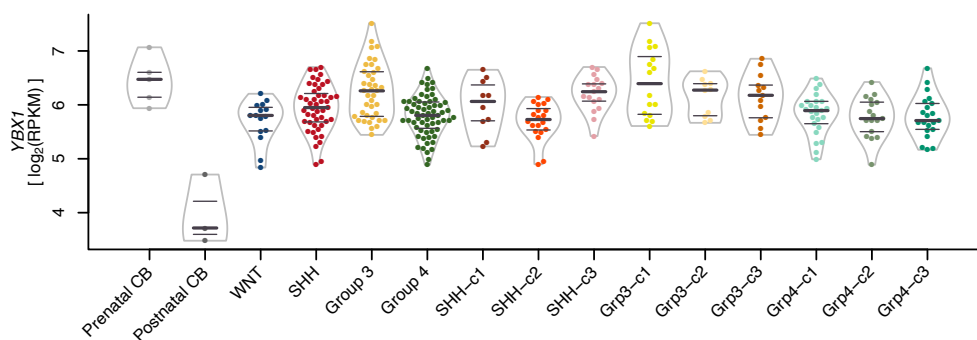


Figure A.53: Expression profile of *YBX1* in MB. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A.4 Expression pattern of lnc genes and related coding genes

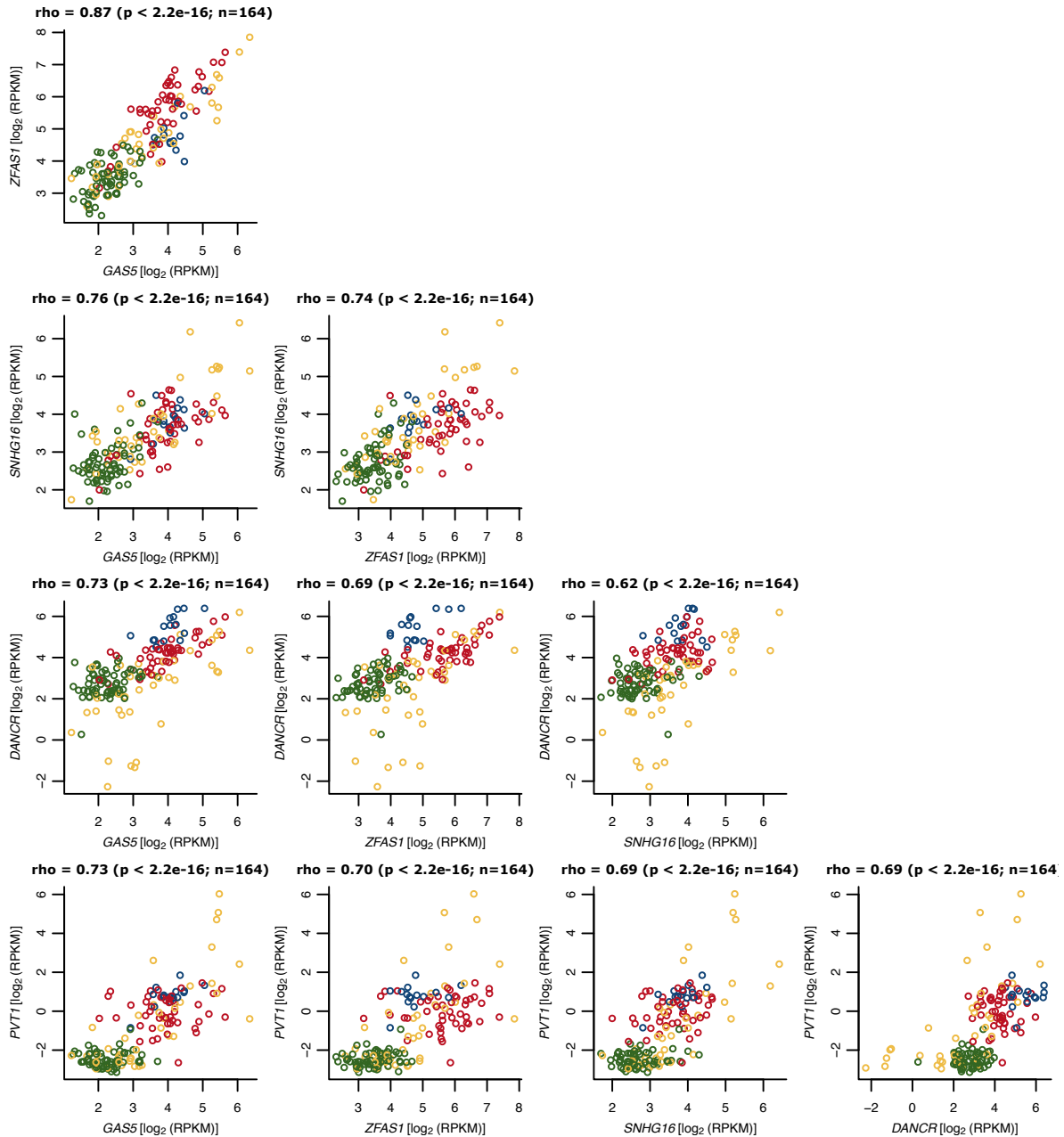


Figure A.54: Pairwise correlation and scatter plots of *GAS5*, *ZFAS1*, *SNHG16*, *PVT1*, and *DANCR* expression in ICGC MB samples. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above each plot.

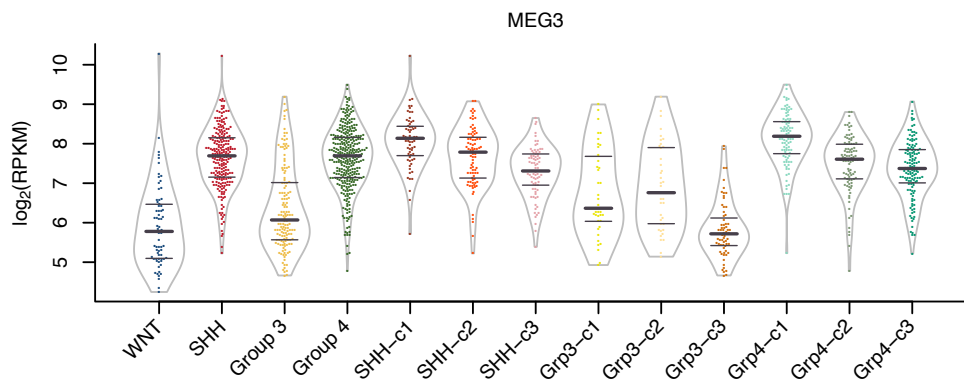


Figure A.55: Expression profile of *MEG3* in Cavalli *et al.* MB cohort [221]. Violin plots show expression distribution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A.4 Expression pattern of lnc genes and related coding genes

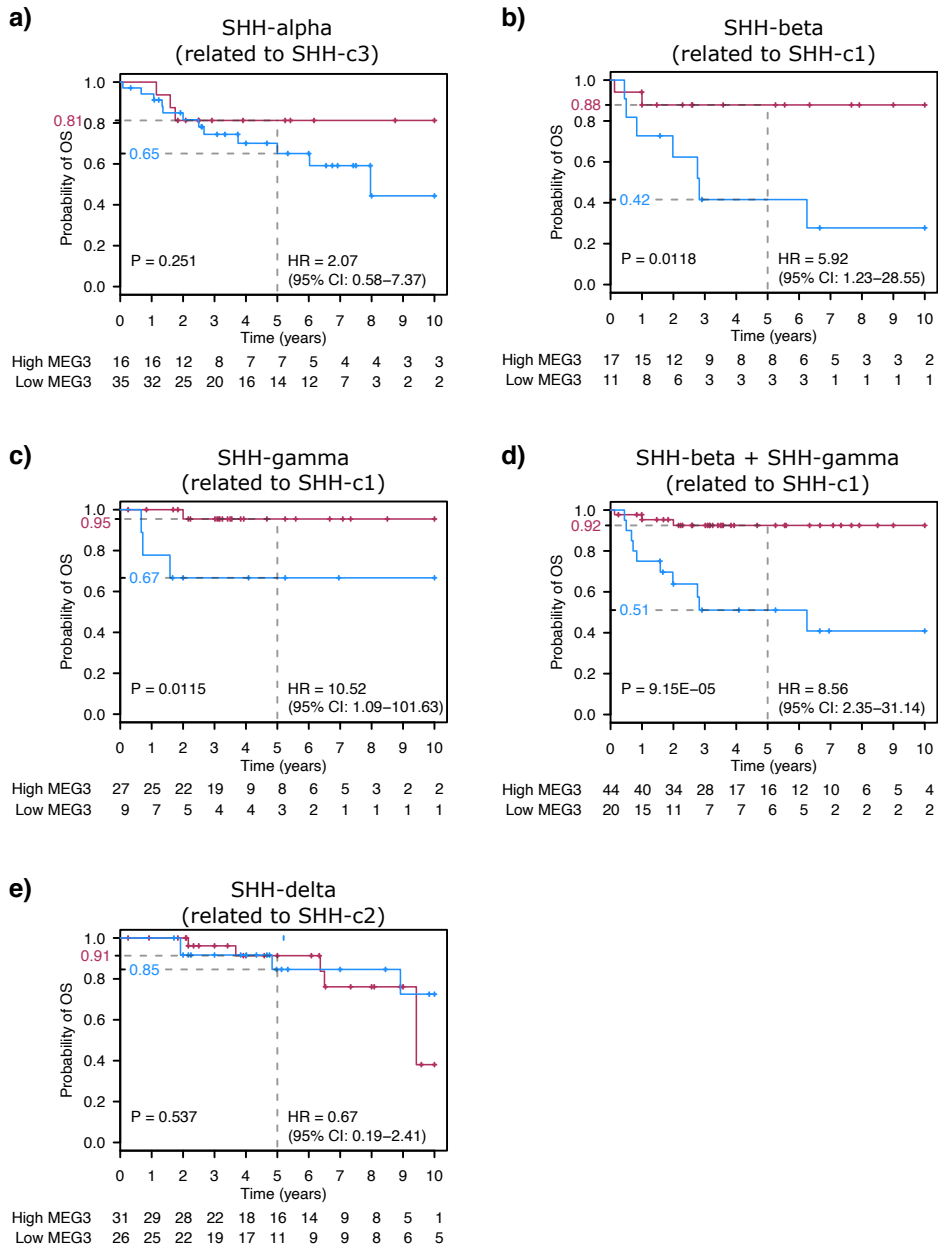


Figure A.56: OS association with *MEG3* expression in SHH subtypes defined by Cavalli *et al.* [221]. Kaplan-Meier curves show OS in *MEG3* low- and high-expressing MBs. **a)** SHH-alpha. **b)** SHH-beta. **c)** SHH-gamma. **d)** SHH-beta + SHH-gamma. **e)** SHH-delta. Shown p-value and hazard ratio relates to differences in survival between groups based on Cox regression (Methods section 5.4.6.7). Hazard ratio (HR) indicates risk comparing low-expressing vs. high-expressing samples. 95% confidence interval (CI) of HR is shown in brackets.

A Appendix

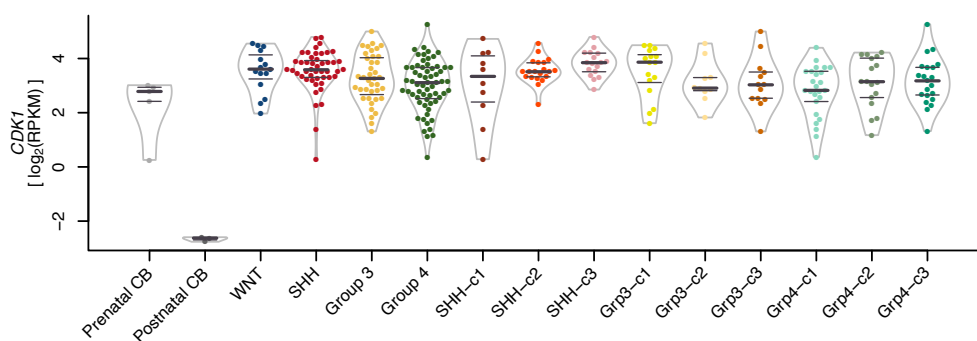


Figure A.57: Expression profile of *CDK1* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green.

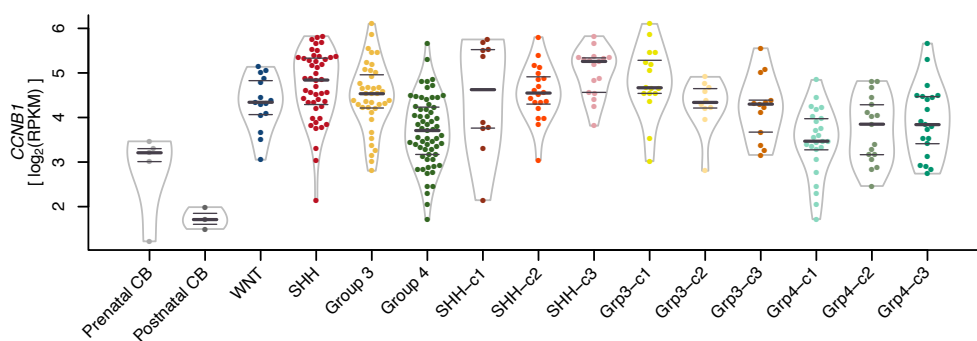


Figure A.58: Expression profile of *CCNB1* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green.

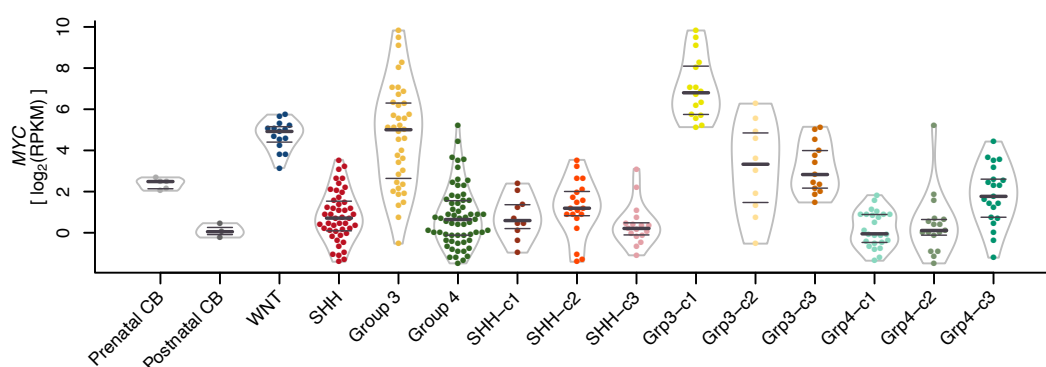


Figure A.59: Expression profile of *MYC* in MB. Violin plots show expression destitution. 25% , 50% and 75% quantiles are indicated by horizontal lines. Individual samples are shown as bee swarm plots.

A.4 Expression pattern of *lnc* genes and related coding genes

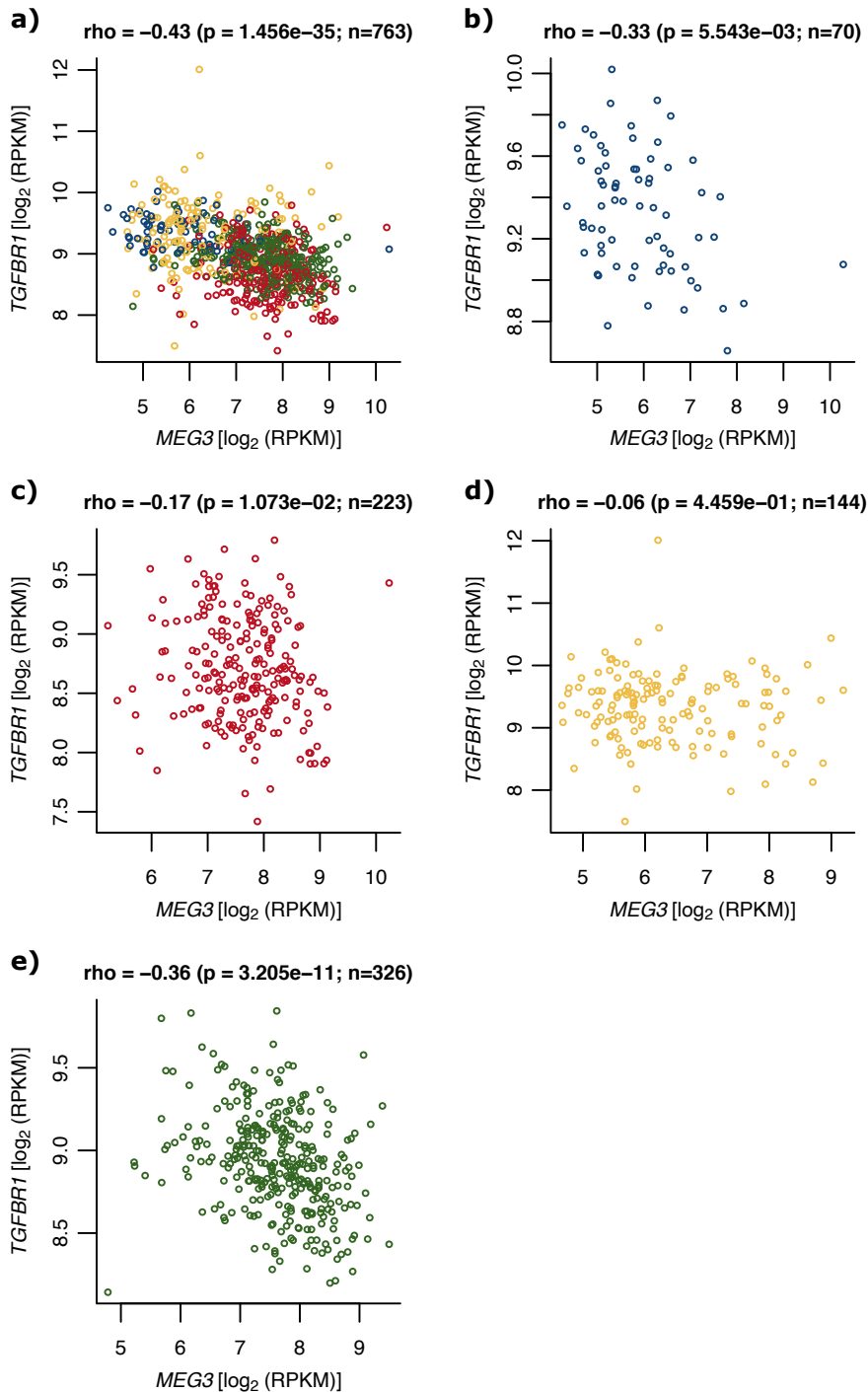


Figure A.60: Scatter plot of *MEG3* and *TGFBR1* expression in Cavalli *et al.* cohort. **a)** Across the whole cohort. **b)** WNT MB. **c)** SHH MB. **d)** Group 3 MB. **d)** Group 4 MB. Colours indicate MB subgroups: WNT=blue, SHH=red, Group 3 = yellow, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples *n* is displayed above each plot.

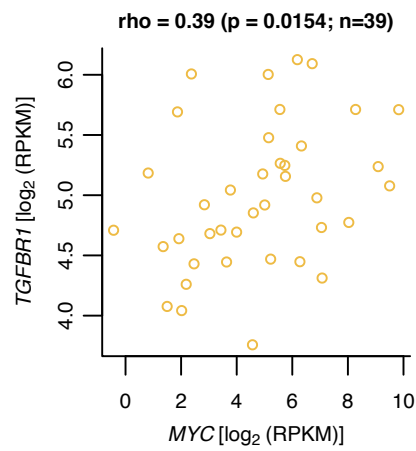


Figure A.61: Scatter plot of *MYC* and *TGFBR1* expression in ICGC Group 3 MB samples. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

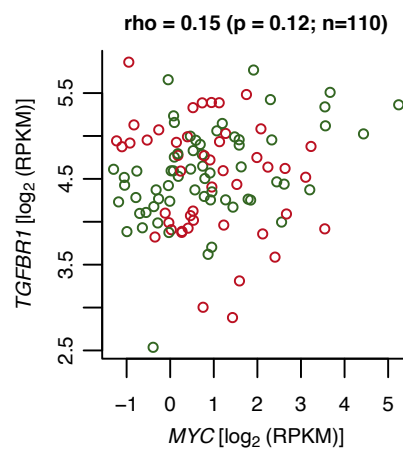


Figure A.62: Scatter plot of *MYC* and *TGFBR1* expression in ICGC SHH and Group 4 MB samples. Colours indicate MB subgroups: SHH=red, Group 4 = green. Spearman correlation coefficient, related p-value, and number of samples n is displayed above the plot.

A.5 Gene symbols and their description

Symbol and description of genes mentioned within the text.

Gene symbol	Description
<i>AGAP2-AS1</i>	AGAP2 antisense RNA 1
<i>AKT1</i>	v-akt murine thymoma viral oncogene homolog 1
<i>APC</i>	adenomatous polyposis coli
<i>AREG</i>	amphiregulin
<i>ATOH1</i>	atonal homolog 1 (Drosophila)
<i>AURKA</i>	aurora kinase A
<i>BAIAP2</i>	BAI1-associated protein 2
<i>BAIAP2-AS1</i>	BAIAP2 antisense RNA 1 (head to head)
<i>BMI1</i>	BMI1 polycomb ring finger oncogene
<i>BRAF</i>	v-raf murine sarcoma viral oncogene homolog B
<i>CCAT1</i>	colon cancer associated transcript 1 (non-protein coding)
<i>CCNB1</i>	cyclin B1
<i>CDK1</i>	cyclin-dependent kinase 1
<i>CDK6</i>	cyclin-dependent kinase 6
<i>CDKN1A</i>	cyclin-dependent kinase inhibitor 1A (p21, Cip1)
<i>CDKN2B-AS1</i>	CDKN2B antisense RNA 1
<i>CEACAM7</i>	carcinoembryonic antigen-related cell adhesion molecule 7
<i>CHD5</i>	chromodomain helicase DNA binding protein 5
<i>CRNDE</i>	colorectal neoplasia differentially expressed (non-protein coding)
<i>CRX</i>	cone-rod homeobox
<i>CSNK2B</i>	casein kinase 2, beta polypeptide
<i>CTDNBP1</i>	CTD nuclear envelope phosphatase 1
<i>CTNNB1</i>	catenin (cadherin-associated protein), beta 1, 88kDa
<i>DANCR</i>	differentiation antagonizing non-protein coding RNA
<i>DDX3X</i>	DEAD (Asp-Glu-Ala-Asp) box helicase 3, X-linked
<i>DDX31</i>	DEAD (Asp-Glu-Ala-Asp) box polypeptide 31
<i>DEK</i>	DEK oncogene
<i>DIO3</i>	deiodinase, iodothyronine, type III
<i>DLGAP1</i>	discs, large (Drosophila) homolog-associated protein 1
<i>DLGAP1-AS1</i>	DLGAP1 antisense RNA 1
<i>DLK1</i>	delta-like 1 homolog (Drosophila)
<i>E2F5</i>	E2F transcription factor 5, p130-binding
<i>EBF1</i>	early B-cell factor 1
<i>EGFR</i>	epidermal growth factor receptor
<i>EN1</i>	engrailed homeobox 1
<i>EN2</i>	engrailed homeobox 2
<i>EOMES</i>	eomesodermin
<i>EPHA7</i>	EPH receptor A7
<i>ERBB2</i>	v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2
<i>EREG</i>	epiregulin
<i>ETS1</i>	v-ets avian erythroblastosis virus E26 oncogene homolog 1
<i>FEZF1</i>	FEZ family zinc finger 1
<i>FEZF1-AS1</i>	FEZF1 antisense RNA 1
<i>FGF8</i>	fibroblast growth factor 8 (androgen-induced)
<i>FGFR1</i>	fibroblast growth factor receptor 1

A Appendix

<i>FIRRE</i>	firre intergenic repeating RNA element
<i>FLT1</i>	fms-related tyrosine kinase 1
<i>FYN</i>	FYN oncogene related to SRC, FGR, YES
<i>GAS5</i>	growth arrest-specific 5 (non-protein coding)
<i>GBX1</i>	gastrulation brain homeobox 1
<i>GDF15</i>	growth differentiation factor 15
<i>GFI1</i>	growth factor independent 1 transcription repressor
<i>GFI1B</i>	growth factor independent 1B transcription repressor
<i>GLI1</i>	GLI family zinc finger 1
<i>GLI2</i>	GLI family zinc finger 2
<i>GLYCTK</i>	glycerate kinase
<i>GLYCTK-AS1</i>	GLYCTK antisense RNA 1
<i>HDAC2</i>	histone deacetylase 2
<i>HES5</i>	hes family bHLH transcription factor 5
<i>HLX</i>	H2.0-like homeobox
<i>HNRNPK</i>	heterogeneous nuclear ribonucleoprotein K
<i>HOTAIRM1</i>	HOXA transcript antisense RNA, myeloid-specific 1
<i>HOXA1</i>	homeobox A1
<i>HOXA2</i>	homeobox A2
<i>HOXA3</i>	homeobox A3
<i>HOXA4</i>	homeobox A4
<i>HOXA5</i>	homeobox A5
<i>HOXA6</i>	homeobox A6
<i>HOXA7</i>	homeobox A7
<i>HSF2</i>	heat shock transcription factor 2
<i>IRX6</i>	iroquois homeobox 6
<i>KBTBD4</i>	kelch repeat and BTB (POZ) domain containing 4
<i>KDM6A</i>	lysine (K)-specific demethylase 6A
<i>KDR</i>	kinase insert domain receptor (a type III receptor tyrosine kinase)
<i>KLF2</i>	Kruppel-like factor 2
<i>KMT2C</i>	lysine (K)-specific methyltransferase 2C
<i>KMT2D</i>	lysine (K)-specific methyltransferase 2D
<i>KRAS</i>	Kirsten rat sarcoma viral oncogene homolog
<i>LHX2</i>	LIM homeobox 2
<i>LHX4</i>	LIM homeobox 4
<i>LINC-ROR</i>	long intergenic non-protein coding RNA, regulator of reprogramming
<i>LINC01122</i>	long intergenic non-protein coding RNA 1122
<i>LMX1A</i>	LIM homeobox transcription factor 1, alpha
<i>LOXL1-AS1</i>	LOXL1 antisense RNA 1
<i>MAP2K1</i>	mitogen-activated protein kinase kinase 1
<i>MEG3</i>	maternally expressed 3 (non-protein coding)
<i>MSX2</i>	msh homeobox 2
<i>MYC</i>	v-myc avian myelocytomatosis viral oncogene homolog
<i>MYCL</i>	v-myc avian myelocytomatosis viral oncogene lung carcinoma derived homolog
<i>MYCN</i>	v-myc avian myelocytomatosis viral oncogene neuroblastoma derived homolog
<i>MYT1</i>	myelin transcription factor 1
<i>MYT1L</i>	myelin transcription factor 1-like
<i>NES</i>	nestin
<i>NEUROD1</i>	neuronal differentiation 1
<i>NEUROD2</i>	neuronal differentiation 2

<i>NEUROD6</i>	neuronal differentiation 6
<i>NEUROG1</i>	neurogenin 1
<i>NFATC1</i>	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1
<i>NKD2</i>	naked cuticle homolog 2 (Drosophila)
<i>NKX2-2AS</i>	NKX2-2 antisense RNA 1
<i>NRAS</i>	neuroblastoma RAS viral (v-ras) oncogene homolog
<i>NRL</i>	neural retina leucine zipper
<i>OTX2</i>	orthodenticle homeobox 2
<i>PART1</i>	prostate androgen-regulated transcript 1 (non-protein coding)
<i>PAX2</i>	paired box 2
<i>PDCD4</i>	programmed cell death 4 (neoplastic transformation inhibitor)
<i>PDGFRA</i>	platelet-derived growth factor receptor, alpha polypeptide
<i>PHLPP2</i>	PH domain and leucine rich repeat protein phosphatase 2
<i>PIK3CA</i>	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
<i>PRDM6</i>	PR domain containing 6
<i>PTCH1</i>	patched 1
<i>PTEN</i>	phosphatase and tensin homolog
<i>PVT1</i>	Pvt1 oncogene (non-protein coding)
<i>RAX2</i>	retina and anterior neural fold homeobox 2
<i>REG4</i>	regenerating islet-derived family, member 4
<i>RMST</i>	rhabdomyosarcoma 2 associated transcript (non-protein coding)
<i>RREB1</i>	ras responsive element binding protein 1
<i>RUNX2</i>	runt-related transcription factor 2
<i>SIX6</i>	SIX homeobox 6
<i>SMARCA4</i>	SWI/SNF related, matrix asso., actin depen. regulator of chromatin, subf. a, mem. 4
<i>SMO</i>	smoothened, frizzled family receptor
<i>SNCAIP</i>	synuclein, alpha interacting protein
<i>SNHG16</i>	small nucleolar RNA host gene 16 (non-protein coding)
<i>SOX2</i>	SRY (sex determining region Y)-box 2
<i>SOX9</i>	SRY (sex determining region Y)-box 9
<i>SOX11</i>	SRY (sex determining region Y)-box 11
<i>SP5</i>	Sp5 transcription factor
<i>STI8</i>	suppression of tumorigenicity 18 (breast carcinoma) (zinc finger protein)
<i>SUFU</i>	suppressor of fused homolog (Drosophila)
<i>TBR1</i>	T-box, brain, 1
<i>TEK</i>	TEK tyrosine kinase, endothelial
<i>TERT</i>	telomerase reverse transcriptase
<i>TGFB2</i>	transforming growth factor, beta 2
<i>TGFBR1</i>	transforming growth factor, beta receptor 1
<i>TGFBR2</i>	transforming growth factor, beta receptor II (70/80kDa)
<i>THRA</i>	thyroid hormone receptor, alpha
<i>TP53</i>	tumor protein p53
<i>TWIST1</i>	twist family bHLH transcription factor 1
<i>UCA1</i>	urothelial cancer associated 1 (non-protein coding)
<i>VPS9D1-AS1</i>	VPS9D1 antisense RNA 1
<i>WNT1</i>	wingless-type MMTV integration site family, member 1
<i>YAP1</i>	Yes-associated protein 1
<i>YBX1</i>	Y box binding protein 1
<i>ZBTB8B</i>	zinc finger and BTB domain containing 8B
<i>ZBTB18</i>	zinc finger and BTB domain containing 18

A Appendix

<i>ZEB1</i>	zinc finger E-box binding homeobox 1
<i>ZFAS1</i>	ZNFX1 antisense RNA 1
<i>ZMYM3</i>	zinc finger, MYM-type 3
<i>ZNF521</i>	zinc finger protein 521
<i>ZNF540</i>	zinc finger protein 540
<i>ZNFX1</i>	zinc finger, NFX1-type containing 1

Bibliography

- [1] Margaret A. Knowles. *Introduction to the cellular and molecular biology of cancer*. eng. 4th ed. New York: Oxford University Press, 2005. ISBN: 1-280-75813-9.
- [2] Steven I. Hajdu. "A note from history: Landmarks in history of cancer, part 1". In: *Cancer* 117.5 (2011), pp. 1097–1102. ISSN: 0008543X. DOI: 10.1002/cncr.25553.
- [3] Steven I. Hajdu. "A note from history: Landmarks in history of cancer, part 3". In: *Cancer* 118.4 (2012), pp. 1155–1168. ISSN: 10970142. DOI: 10.1002/cncr.26320.
- [4] Raymond W Ruddon. *Cancer biology*. New York, 2007.
- [5] Gašper Tkačik and William Bialek. "Cell Biology: Networks, Regulation and Pathways". In: *Encyclopedia of Complexity and Systems Science*. New York, NY: Springer New York, 2009, pp. 719–741. DOI: 10.1007/978-0-387-30440-3_48.
- [6] Sameek Roychowdhury and Arul M. Chinnaiyan. "Translating cancer genomes and transcriptomes for precision oncology." In: *CA: a cancer journal for clinicians* 66.1 (2016), pp. 75–88. ISSN: 1542-4863. DOI: 10.3322/caac.21329.
- [7] Jamie A Davies. *Life unfolding : how the human body creates itself*. First edit. New York: Oxford University Press. ISBN: 0-19-967354-3.
- [8] Geoffrey M Cooper. *The cell : a molecular approach*. Washington, DC, 2000.
- [9] Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti et al. "An estimation of the number of cells in the human body". In: *Ann. Hum. Biol.* 40.6 (2013), pp. 463–471. ISSN: 03014460. DOI: 10.3109/03014460.2013.807878.
- [10] Douglas Hanahan and Robert A. Weinberg. "Hallmarks of cancer: The next generation". In: *Cell* 144.5 (2011), pp. 646–674. ISSN: 00928674. DOI: 10.1016/j.cell.2011.02.013.
- [11] "Neoplasia". In: *Encycl. Cancer*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 2474–2474. DOI: 10.1007/978-3-642-16483-5_4011.
- [12] Robert A. Weinberg. *The biology of cancer*. 2. ed. New York, NY [u.a: Garland Science, 2014. ISBN: 978-0-8153-4528-2.
- [13] Ruth J. Muschel. "Metastasis". In: *Encycl. Cancer*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 2260–2261. DOI: 10.1007/978-3-642-16483-5_3671.
- [14] "Malignant Tumor". In: *Encycl. Cancer*. Ed. by Manfred Schwab. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 2150–2150. ISBN: 978-3-642-16482-8. DOI: 10.1007/978-3-642-16483-5_3519.
- [15] Frank Emmert-Streib, Matthias Dehmer and Benjamin Haibe-Kains. "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks." In: *Front Cell Dev Biol* 2 (2014), p. 38. DOI: 10.3389/fcell.2014.00038.
- [16] Jeremy M. Berg, John L. Tymoczko and Lubert Stryer. *Biochemistry*. 5. ed., in. New York, NY: Freeman, 2002. ISBN: 0-7167-4684-0.
- [17] B Alberts, A Johnson, J Lewis, P Walter, M Raff and K Roberts. *Molecular biology of the cell*. New York, NY, 2002.
- [18] Kevin C. Wang and Howard Y. Chang. "Molecular mechanisms of long noncoding RNAs". In: *Mol Cell* 43.6 (2012), pp. 904–914. DOI: 10.1016/j.molcel.2011.08.018. Molecular.

Bibliography

- [19] Graziano Pesole. "What is a gene? An updated operational definition". In: *Gene* 417.1-2 (2008), pp. 1–4. DOI: 10.1016/j.gene.2008.03.010.
- [20] "The landscape of long noncoding RNAs in the human transcriptome". In: *Nat Genet* 47.3 (2015), pp. 199–208. ISSN: 1546-1718. DOI: 10.1038/ng.3192.
- [21] Yufei Huang. "Gene Expression". In: *Encyclopedia of Systems Biology*. New York, NY: Springer New York, 2013, pp. 791–791. DOI: 10.1007/978-1-4419-9863-7_819.
- [22] Tae Hoon Kim, Leah O. Barrera, Ming Zheng, Chunxu Qu, Michael A. Singer, Todd A. Richmond et al. "A high-resolution map of active promoters in the human genome." In: *Nature* 436.7052 (2005), pp. 876–80. ISSN: 1476-4687. DOI: 10.1038/nature03877.
- [23] Peter A. Jones. "Functions of DNA methylation: Islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13.7 (2012), pp. 484–492. ISSN: 14710056. DOI: 10.1038/nrg3230.
- [24] Ty C. Voss and Gordon L. Hager. "Dynamic regulation of transcriptional states by chromatin and transcription factors". In: *Nature Reviews Genetics* 15.2 (2014), pp. 69–81. ISSN: 14710056. DOI: 10.1038/nrg3623.
- [25] Tong Ihn Lee and Richard A. Young. "Transcriptional regulation and its misregulation in disease". In: *Cell* 152.6 (2013), pp. 1237–1251. ISSN: 10974172. DOI: 10.1016/j.cell.2013.02.014.
- [26] Vani Brahmachari and Shruti Jain. "Epigenetics". In: *Encyclopedia of Systems Biology*. New York, NY: Springer New York, 2013, pp. 665–669. DOI: 10.1007/978-1-4419-9863-7_567.
- [27] "Chromatin Domains: The Unit of Chromosome Organization". In: *Molecular Cell* 62.5 (2016), pp. 668–680. ISSN: 10974164. DOI: 10.1016/j.molcel.2016.05.018.
- [28] "Phenotype". In: *Encyclopedia of Cancer*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 2856–2856. DOI: 10.1007/978-3-642-16483-5_4514.
- [29] Sui Huang, Ingemar Ernberg and Stuart Kauffman. "Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective." In: *Seminars in cell & developmental biology* 20.7 (2009), pp. 869–76. ISSN: 1096-3634. DOI: 10.1016/j.semcdb.2009.07.003. eprint: NIHMS150003.
- [30] "Frameshift Mutation". In: *Encyclopedia of Cancer*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1454–1454. DOI: 10.1007/978-3-642-16483-5_2266.
- [31] "The prenatal origins of cancer." In: *Nature reviews. Cancer* 14.4 (2014), pp. 277–89. ISSN: 1474-1768. DOI: 10.1038/nrc3679.
- [32] Zachary R. Chalmers, Caitlin F. Connelly, David Fabrizio, Laurie Gay, Siraj M. Ali, Riley Ennis et al. "Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden". In: *Genome Medicine* 9.1 (2017), pp. 1–14. ISSN: 1756994X. DOI: 10.1186/s13073-017-0424-2.
- [33] Jerry W. Shay. "Telomerase". In: *Encyclopedia of Cancer*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 3631–3636. DOI: 10.1007/978-3-642-16483-5_5715.
- [34] Adam M. Schmitt and Howard Y. Chang. "Long Noncoding RNAs in Cancer Pathways." In: *Cancer cell* 29.4 (2016), pp. 452–463. ISSN: 1878-3686. DOI: 10.1016/j.ccell.2016.03.010.
- [35] "Unique features of long non-coding RNA biogenesis and function." In: *Nature reviews. Genetics* 17.1 (2016), pp. 47–62. ISSN: 1471-0064. DOI: 10.1038/nrg.2015.10.
- [36] Fabian A. Buske, Denis C. Bauer, John S. Mattick and Timothy L. Bailey. "Triplexator: Detecting nucleic acid triple helices in genomic and transcriptomic data". In: *Genome Research* 22.7 (2012), pp. 1372–1381. ISSN: 10889051. DOI: 10.1101/gr.130237.111.

- [37] Yue Li, Junetha Syed and Hiroshi Sugiyama. “RNA-DNA Triplex Formation by Long Noncoding RNAs.” In: *Cell chemical biology* 23.11 (2016), pp. 1325–1333. ISSN: 2451-9448. DOI: 10.1016/j.chembiol.2016.09.011.
- [38] Guodong Yang, Xiaozhao Lu and Lijun Yuan. “LncRNA: A link between RNA and cancer”. In: *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1839.11 (2014), pp. 1097–1109. ISSN: 18764320. DOI: 10.1016/j.bbagr.2014.08.012.
- [39] Sonja Hombach and Markus Kretz. “Non-coding RNAs: Classification, Biology and Functioning.” In: *Advances in experimental medicine and biology* 937 (2016), pp. 3–17. ISSN: 0065-2598. DOI: 10.1007/978-3-319-42059-2_1.
- [40] Sandra U. Schmitz, Phillip Grote and Bernhard G. Herrmann. “Mechanisms of long noncoding RNA function in development and disease”. In: *Cellular and Molecular Life Sciences* 73.13 (2016), pp. 2491–2509. ISSN: 14209071. DOI: 10.1007/s00018-016-2174-5.
- [41] Xiaohui Yan, Zhongyi Hu, Yi Feng, Xiaowen Hu, Jiao Yuan, Sihai D. Zhao et al. “Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers”. In: *Cancer Cell* 28.4 (2015), pp. 529–540. ISSN: 18783686. DOI: 10.1016/j.ccell.2015.09.006.
- [42] Frank J. Slack and Arul M. Chinnaiyan. “The Role of Non-coding RNAs in Oncology”. In: *Cell* 179.5 (2019), pp. 1033–1055. ISSN: 10974172. DOI: 10.1016/j.cell.2019.10.017.
- [43] Tanmoy Mondal, Santhilal Subhash, Roshan Vaid, Stefan Enroth, Sireesha Uday, Björn Reinius et al. “MEG3 long noncoding RNA regulates the TGF- β pathway genes through formation of RNA-DNA triplex structures.” In: *Nat Commun* 6 (July 2015), p. 7743. DOI: 10.1038/ncomms8743.
- [44] Naomi Habib, Yinqing Li, Matthias Heidenreich, Lukasz Swiech, Inbal Avraham-Davidi, John J Trombetta et al. “Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons.” In: *Science* 353.6302 (Aug. 2016), pp. 925–8. DOI: 10.1126/science.aad7038.
- [45] Adarsh S Reddy, David O’Brien, Nilambari Pisat, Claire T Weichselbaum, Kristina Sakers, Miriam Lisci et al. “A Comprehensive Analysis of Cell Type-Specific Nuclear RNA From Neurons and Glia of the Brain.” In: *Biol Psychiatry* 81.3 (Feb. 2017), pp. 252–264. DOI: 10.1016/j.biopsych.2016.02.021.
- [46] Xun Zhang, Yunli Zhou, Kshama R Mehta, Daniel C Danila, Staci Scolavino, Stacey R Johnson et al. “A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells.” In: *J Clin Endocrinol Metab* 88.11 (Nov. 2003), pp. 5119–26. DOI: 10.1210/jc.2003-030222.
- [47] Chringma Sherpa, Jason W Rausch and Stuart Fj Le Grice. “Structural characterization of maternally expressed gene 3 RNA reveals conserved motifs and potential sites of interaction with polycomb repressive complex 2.” In: *Nucleic Acids Res* 46.19 (Nov. 2018), pp. 10432–10447. DOI: 10.1093/nar/gky722.
- [48] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao et al. “Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding.” In: *Science* 360.6385 (Apr. 2018), pp. 176–182. DOI: 10.1126/science.aam8999.
- [49] Soudeh Ghafouri-Fard and Mohammad Taheri. “Maternally expressed gene 3 (MEG3): A tumor suppressor long non coding RNA.” In: *Biomed Pharmacother* 118 (Oct. 2019), p. 109129. DOI: 10.1016/j.biopha.2019.109129.
- [50] R Qin, Z Chen, Y Ding, J Hao, J Hu and F Guo. “Long non-coding RNA MEG3 inhibits the proliferation of cervical carcinoma cells through the induction of cell cycle arrest and apoptosis.” In: *Neoplasma* 60.5 (2013), pp. 486–92. DOI: 10.4149/neo_2013_063.
- [51] Maggie M Balas and Aaron M Johnson. “Exploring the mechanisms behind long noncoding RNAs and cancer.” In: *Noncoding RNA Res* 3.3 (Sept. 2018), pp. 108–117. DOI: 10.1016/j.ncrna.2018.03.001.

Bibliography

- [52] Arwa Al-Rugeebah, Mohammed Alanazi and Narasimha Reddy Parine. "MEG3: an Oncogenic Long Non-coding RNA in Different Cancers." In: *Pathol. Oncol. Res.* (2019). ISSN: 1532-2807. DOI: 10.1007/s12253-019-00614-3.
- [53] Yunli Zhou, Ying Zhong, Yingying Wang, Xun Zhang, Dalia L Batista, Roger Gejman et al. "Activation of p53 by MEG3 non-coding RNA." In: *J Biol Chem* 282.34 (Aug. 2007), pp. 24731–42. DOI: 10.1074/jbc.M702029200.
- [54] Juanjuan Zhu, Shanshan Liu, Fuqiang Ye, Yuan Shen, Yi Tie, Jie Zhu et al. "Long Noncoding RNA MEG3 Interacts with p53 Protein and Regulates Partial p53 Target Genes in Hepatoma Cells." In: *PLoS One* 10.10 (2015), e0139790. DOI: 10.1371/journal.pone.0139790.
- [55] Carol A Edwards, Andrew J Mungall, Lucy Matthews, Edward Ryder, Dionne J Gray, Andrew J Pask et al. "The evolution of the DLK1-DIO3 imprinted domain in mammals." In: *PLoS Biol* 6.6 (June 2008), e135. DOI: 10.1371/journal.pbio.0060135.
- [56] N Miyoshi, H Wagatsuma, S Wakana, T Shiroishi, M Nomura, K Aisaka et al. "Identification of an imprinted gene, Meg3/Gtl2 and its human homologue MEG3, first mapped on mouse distal chromosome 12 and human chromosome 14q." In: *Genes Cells* 5.3 (Mar. 2000), pp. 211–20. DOI: 10.1046/j.1365-2443.2000.00320.x.
- [57] M Paulsen, S Takada, N A Youngson, M Benchaib, C Charlier, K Segers et al. "Comparative sequence analysis of the imprinted Dlk1-Gtl2 locus in three mammalian species reveals highly conserved genomic elements and refines comparison with the Igf2-H19 region." In: *Genome Res* 11.12 (Dec. 2001), pp. 2085–94. DOI: 10.1101/gr.206901.
- [58] Shau-Ping Lin, Neil Youngson, Shuji Takada, Hervé Seitz, Wolf Reik, Martina Paulsen et al. "Asymmetric regulation of imprinting on the maternal and paternal chromosomes at the Dlk1-Gtl2 imprinted cluster on mouse chromosome 12." In: *Nat Genet* 35.1 (Sept. 2003), pp. 97–102. DOI: 10.1038/ng1233.
- [59] Yuqing He, Yanhong Luo, Biyu Liang, Lei Ye, Guangxing Lu and Weiming He. "Potential applications of MEG3 in cancer diagnosis and prognosis." In: *Oncotarget* 8.42 (Sept. 2017), pp. 73282–73295. DOI: 10.18632/oncotarget.19931.
- [60] D Astuti, F Latif, K Wagner, D Gentle, W N Cooper, D Catchpoole et al. "Epigenetic alteration at the DLK1-GTL2 imprinted domain in human neoplasia: analysis of neuroblastoma, pheochromocytoma and Wilms' tumour." In: *Br J Cancer* 92.8 (Apr. 2005), pp. 1574–80. DOI: 10.1038/sj.bjc.6602478.
- [61] Vladimir Balik, Josef Srovnal, Igor Sulla, Ondrej Kalita, Tatiana Foltanova, Miroslav Vaverka et al. "MEG3: a novel long noncoding potentially tumour-suppressing RNA in meningiomas." In: *J Neurooncol* 112.1 (Mar. 2013), pp. 1–8. DOI: 10.1007/s11060-012-1038-6.
- [62] Xun Zhang, Roger Gejman, Ali Mahta, Ying Zhong, Kimberley A. Rice, Yunli Zhou et al. "Maternally expressed gene 3, an imprinted noncoding RNA gene, is associated with meningioma pathogenesis and progression." In: *Cancer Res.* 70.6 (2010), pp. 2350–8. DOI: 10.1158/0008-5472.CAN-09-3885.
- [63] Aniruddha Bhati, H. Garg, A. Gupta, H. Chhabra, A. Kumari and T. Patel. "Omics of cancer." In: *Asian Pacific journal of cancer prevention : APJCP* 13.9 (2012), pp. 4229–33. ISSN: 2476-762X. DOI: 10.7314/apjcp.2012.13.9.4229.
- [64] Kenichi Inaoka, Yoshikuni Inokawa and Shuji Nomoto. "Genomic-Wide Analysis with Microarrays in Human Oncology." In: *Microarrays (Basel, Switzerland)* 4.4 (2015), pp. 454–73. ISSN: 2076-3905. DOI: 10.3390/microarrays4040454.
- [65] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo and Xuejun Liu. "Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells." In: *PLoS One* 9.1 (2014), e78644. DOI: 10.1371/journal.pone.0078644.

- [66] Sara Goodwin, John D. McPherson and W. Richard McCombie. “Coming of age: Ten years of next-generation sequencing technologies”. In: *Nature Reviews Genetics* 17.6 (2016), pp. 333–351. ISSN: 14710064. DOI: 10.1038/nrg.2016.49.
- [67] Brian T. Wilhelm and Josette Renée Landry. “RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing”. In: *Methods* 48.3 (2009), pp. 249–257. ISSN: 10462023. DOI: 10.1016/j.ymeth.2009.03.016.
- [68] “Sequencing depth and coverage: Key considerations in genomic analyses”. In: *Nature Reviews Genetics* 15.2 (2014), pp. 121–132. ISSN: 14710056. DOI: 10.1038/nrg3642.
- [69] Dvir Aran, Marina Sirota and Atul J. Butte. “Systematic pan-cancer analysis of tumour purity”. In: *Nature Communications* 6 (2015), pp. 1–12. ISSN: 20411723. DOI: 10.1038/ncomms9971.
- [70] Bonnie Berger, Jian Peng and Mona Singh. “Computational solutions for omics data”. In: *Nature Reviews Genetics* 14.5 (2013), pp. 333–346. ISSN: 14710056. DOI: 10.1038/nrg3433.
- [71] J D Kelleher and B Tierney. *Data science*. The MIT Press essential knowledge series. Cambridge, MA: The MIT Press, 2018. ISBN: 9780262347020.
- [72] Peter. Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge, MA, 2012.
- [73] Eyke Hüllermeier, Thomas Fober and Marco Mernberger. “Inductive Bias”. In: *Encyclopedia of Systems Biology*. New York, NY: Springer New York, 2013, pp. 1018–1018. DOI: 10.1007/978-1-4419-9863-7_927.
- [74] Brian J Parker, Simon Günter and Justin Bedo. “Stratification bias in low signal microarray studies.” In: *BMC Bioinformatics* 8 (Sept. 2007), p. 326. DOI: 10.1186/1471-2105-8-326.
- [75] Ji Hyun Kim. “Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap”. In: *Computational Statistics and Data Analysis* 53.11 (2009), pp. 3735–3745. ISSN: 01679473. DOI: 10.1016/j.csda.2009.04.009.
- [76] Charles X Ling and Victor S Sheng. “Class Imbalance Problem”. In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I Webb. Boston, MA: Springer US, 2010, p. 171. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_110.
- [77] Yasen Jiao and Pufeng Du. “Performance measures in evaluating machine learning based bioinformatics predictors for classifications”. In: *Quantitative Biology* 4.4 (2016), pp. 320–330. ISSN: 20954697. DOI: 10.1007/s40484-016-0081-2.
- [78] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer and Barbara Wold. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature Methods* 5.7 (2008), pp. 621–628. ISSN: 15487091. DOI: 10.1038/nmeth.1226.
- [79] Cole Trapnell, Brian a Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.” In: *Nature biotechnology* 28.5 (2010), pp. 511–5. ISSN: 1546-1696. DOI: 10.1038/nbt.1621.
- [80] Mark D Robinson and Alicia Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data.” In: *Genome biology* 11.3 (2010), R25. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-3-r25.
- [81] Charles M Perou, T Sørlie, Michael B Eisen, M van de Rijn, Stefanie S Jeffrey, Christian A Rees et al. “Molecular portraits of human breast tumours.” In: *Nature* 6797 (2000), pp. 747–52. ISSN: 0028-0836. DOI: 10.1038/35021093.
- [82] Paul A Northcott, Andrey Korshunov, Hendrik Witt, Thomas Hielscher, Charles G Eberhart, Stephen Mack et al. “Medulloblastoma comprises four distinct molecular variants.” In: *J Clin Oncol* 29.11 (Apr. 2011), pp. 1408–14. DOI: 10.1200/JCO.2009.27.4324.

Bibliography

- [83] Paul a Northcott, David T W Jones, Marcel Kool, Giles W Robinson, Richard J Gilbertson, Yoon-Jae Cho et al. "Medulloblastomics: the end of the beginning." In: *Nat. Rev. Cancer* 12.12 (2012), pp. 818–34. ISSN: 1474-1768. DOI: 10.1038/nrc3410.
- [84] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien de Reyniès, Andreas Schlicker, Charlotte Soneson et al. "The consensus molecular subtypes of colorectal cancer." In: *Nat Med* 21.11 (Nov. 2015), pp. 1350–6. DOI: 10.1038/nm.3967.
- [85] Felipe De Sousa E. Melo, Louis Vermeulen, Evelyn Fessler and Jan Paul Medema. "Cancer heterogeneity - A multifaceted view". In: *EMBO Reports* 14.8 (2013), pp. 686–695. ISSN: 1469221X. DOI: 10.1038/embor.2013.92.
- [86] Taosheng Xu, Thuc Duy Le, Lin Liu, Ning Su, Rujing Wang, Bingyu Sun et al. "CancerSubtypes: An R/Bioconductor package for molecular cancer subtype identification, validation and visualization". In: *Bioinformatics* 33.19 (2017), pp. 3131–3133. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx378.
- [87] Rui Xu and Donald C. Wunsch. "Clustering algorithms in biomedical research: A review". In: *IEEE Reviews in Biomedical Engineering* 3 (2010), pp. 120–154. ISSN: 19373333. DOI: 10.1109/RBME.2010.2083647.
- [88] Stefano Monti, Pablo Tamayo, Jill Mesirov and Todd Golub. "Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data". In: *Machine Learning* 52.1-2 (2003), pp. 91–118. ISSN: 08856125. DOI: 10.1023/A:1023949509487.
- [89] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub and Jill P Mesirov. "Metagenes and molecular pattern discovery using matrix factorization." In: *Proc Natl Acad Sci U S A* 101.12 (Mar. 2004), pp. 4164–9. DOI: 10.1073/pnas.0308531101.
- [90] Mark D Robinson, Davis J McCarthy and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–40. DOI: 10.1093/bioinformatics/btp616.
- [91] Davis J McCarthy, Yunshun Chen and Gordon K Smyth. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." In: *Nucleic acids research* 40.10 (2012), pp. 4288–97. ISSN: 1362-4962. DOI: 10.1093/nar/gks042.
- [92] Adi L Tarca, Gaurav Bhatti and Roberto Romero. "A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity." In: *PloS one* 8.11 (2013), e79217. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0079217.
- [93] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry et al. "Gene ontology: tool for the unification of biology." In: *Nature genetics* 1 (2000), pp. 25–9. ISSN: 1061-4036. DOI: 10.1038/75556.
- [94] Yong Wang. "Hypergeometric Distribution". In: *Encyclopedia of Systems Biology*. New York, NY: Springer New York, 2013, pp. 929–930. DOI: 10.1007/978-1-4419-9863-7_492.
- [95] Sandra S. Stinnett. "Survival analysis". In: *Evidence-Based Otolaryngology* 404 (2008), pp. 567–577. DOI: 10.1007/978-0-387-49979-6_25.
- [96] Shuangge Ma. "Survival Analysis, Fundamental Statistical Techniques". In: *Encycl. Syst. Biol.* New York, NY: Springer New York, 2013, pp. 2030–2032. DOI: 10.1007/978-1-4419-9863-7_249.
- [97] D G Altman, B Lausen, W Sauerbrei and M Schumacher. "Dangers of using "optimal" cutpoints in the evaluation of prognostic factors." In: *J Natl Cancer Inst* 86.11 (June 1994), pp. 829–35. DOI: 10.1093/jnci/86.11.829.

- [98] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis and Dimitrios I Fotiadis. "Machine learning applications in cancer prognosis and prediction." In: *Comput Struct Biotechnol J* 13 (2015), pp. 8–17. DOI: 10.1016/j.csbj.2014.11.005.
- [99] Zeenia Jagga and Dinesh Gupta. "Machine learning for biomarker identification in cancer research - developments toward its clinical application." In: *Per Med* 12.4 (Aug. 2015), pp. 371–387. DOI: 10.2217/pme.15.5.
- [100] Bernhard E. Boser, Vladimir N. Vapnik and Isabelle M. Guyon. "Training Algorithm Margin for Optimal Classifiers". In: *Perception* (1992), pp. 144–152.
- [101] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (1995), pp. 273–297. ISSN: 0885-6125. DOI: 10.1007/BF00994018.
- [102] Alexander Statnikov, Constantin F Aliferis, Ioannis Tsamardinis, Douglas Hardin and Shawn Levy. "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis." In: *Bioinformatics* 21.5 (Mar. 2005), pp. 631–43. DOI: 10.1093/bioinformatics/bti033.
- [103] Alexander Statnikov, Lily Wang and Constantin F Aliferis. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification." In: *BMC Bioinformatics* 9 (July 2008), p. 319. DOI: 10.1186/1471-2105-9-319.
- [104] S.R. Alty, H.K. Lam and J. Prada. "On the Applications of Heart Disease Risk Classification and Hand-written Character Recognition using Support Vector Machines". In: *Computational Intelligence and Its Applications*. IMPERIAL COLLEGE PRESS, 2012, pp. 213–253. DOI: 10.1142/9781848166929_0009.
- [105] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000. ISBN: 0521780195.
- [106] Jussi Kujala, Timo Aho and Tapio Elomaa. "A walk from 2-norm SVM to 1-norm SVM". In: *Proceedings - IEEE International Conference on Data Mining, ICDM* (2009), pp. 836–841. ISSN: 15504786. DOI: 10.1109/ICDM.2009.100.
- [107] Joseph O. Ogutu, Torben Schulz-Streeck and Hans-Peter Piepho. "Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions." In: *BMC proceedings* 6 Suppl 2.Suppl 2 (2012), S10. ISSN: 1753-6561. DOI: 10.1186/1753-6561-6-S2-S10.
- [108] Chih Chung Chang and Chih Jen Lin. "LIBSVM: A Library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2.3 (2011). ISSN: 21576904. DOI: 10.1145/1961189.1961199.
- [109] Edgar Osuna, Robert Freund and Federico Girosi. "Support Vector Machines : Training and Applications". In: *Massachusetts Institute of Technology* 9217041.1602 (1997).
- [110] Isabelle Guyon and André Elisseeff. "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182. ISSN: 1532-4435.
- [111] Isabelle Guyon, Jason Weston, Stephen Barnhill and Vladimir Vapnik. "Gene selection for cancer classification using support vector machines". In: *Machine Learning* 46.1-3 (2002), pp. 389–422. ISSN: 08856125. DOI: 10.1023/A:1012487302797.
- [112] Anne Claire Hauray, Pierre Gestraud and Jean Philippe Vert. "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures". In: *PLoS ONE* 6.12 (2011), pp. 1–12. ISSN: 19326203. DOI: 10.1371/journal.pone.0028210. arXiv: 1101.5008.

Bibliography

- [113] Kai Bo Duan, Jagath C. Rajapakse, Haiying Wang and Francisco Azuaje. “Multiple SVM-RFE for gene selection in cancer classification with expression data”. In: *IEEE Transactions on Nanobioscience* 4.3 (2005), pp. 228–233. ISSN: 15361241. DOI: 10.1109/TNB.2005.853657.
- [114] Yong Wang. “Gene Regulatory Networks”. In: *Encycl. Syst. Biol.* New York, NY: Springer New York, 2013, pp. 801–805. DOI: 10.1007/978-1-4419-9863-7_364.
- [115] W Du and O Elemento. “Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies.” In: *Oncogene* 34.25 (June 2015), pp. 3215–25. DOI: 10.1038/onc.2014.291.
- [116] Clément Carré, André Mas and Gabriel Krouk. “Reverse engineering highlights potential principles of large gene regulatory network design and learning.” In: *NPJ Syst Biol Appl* 3 (2017), p. 17. DOI: 10.1038/s41540-017-0019-y.
- [117] Albertha J.M. Walhout. “What does biologically meaningful mean? A perspective on gene regulatory network validation”. In: *Genome Biology* 12.4 (2011), pp. 1–7. ISSN: 14747596. DOI: 10.1186/gb-2011-12-4-109.
- [118] Andrei Lihu and Ștefan Holban. “A review of ensemble methods for de novo motif discovery in ChIP-Seq data”. In: *Briefings in Bioinformatics* 16.6 (2015), pp. 964–973. ISSN: 14774054. DOI: 10.1093/bib/bbv022.
- [119] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho et al. “Wisdom of crowds for robust gene network inference.” In: *Nat Methods* 9.8 (July 2012), pp. 796–804. DOI: 10.1038/nmeth.2016.
- [120] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel and Pierre Geurts. “Inferring regulatory networks from expression data using tree-based methods.” In: *PLoS One* 5.9 (Sept. 2010). DOI: 10.1371/journal.pone.0012776.
- [121] Vladimir Filkov. “Identifying Gene Regulatory Networks from Gene Expression Data”. In: *Handbook of Computational Molecular Biology*. Ed. by Srinivas Aluru. Oxford: Chapman and Hall/CRC, 2005. Chap. 27, pp. 27–1–27–29.
- [122] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York, 2013. ISBN: 978-1-4614-7137-0. DOI: 10.1007/978-1-4614-7138-7.
- [123] Patrick Cahan, Hu Li, Samantha A. Morris, Edroaldo Lummertz da Rocha, George Q. Daley and James J. Collins. “CellNet: Network Biology Applied to Stem Cell Engineering”. In: *Cell* 158.4 (2014), pp. 903–915. DOI: 10.1016/j.cell.2014.07.020.
- [124] Theodosia Charitou, Kenneth Bryan and David J Lynn. “Using biological networks to integrate, visualize and analyze genomics data.” In: *Genet Sel Evol* 48 (Mar. 2016), p. 27. DOI: 10.1186/s12711-016-0205-1.
- [125] Martin Rosvall and Carl T Bergstrom. “Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems.” In: *PLoS One* 6.4 (Apr. 2011), e18209. DOI: 10.1371/journal.pone.0018209.
- [126] Deepak Sharma and Avadhesh Surolia. “Degree Centrality”. In: *Encyclopedia of Systems Biology*. New York, NY: Springer New York, 2013, pp. 558–558. DOI: 10.1007/978-1-4419-9863-7_935.
- [127] Sonali Jathar, Vikram Kumar, Juhi Srivastava and Vidisha Tripathi. “Technological Developments in lncRNA Biology.” In: *Adv Exp Med Biol* 1008 (2017), pp. 283–323. DOI: 10.1007/978-981-10-5203-3_10.

- [128] Tim R. Mercer and John S. Mattick. “Structure and function of long noncoding RNAs in epigenetic regulation”. In: *Nature Structural and Molecular Biology* 20.3 (2013), pp. 300–307. ISSN: 15459993. DOI: 10.1038/nsmb.2480.
- [129] Loyal A Goff and John L Rinn. “Linking RNA biology to lncRNAs.” In: *Genome Res* 25.10 (Oct. 2015), pp. 1456–65. DOI: 10.1101/gr.191122.115.
- [130] Junichi Iwakiri, Michiaki Hamada and Kiyoshi Asai. “Bioinformatics tools for lncRNA research.” In: *Biochim Biophys Acta* 1859.1 (Jan. 2016), pp. 23–30. DOI: 10.1016/j.bbagr.2015.07.014.
- [131] Letizia Da Sacco, Antonella Baldassarre and Andrea Masotti. “Bioinformatics tools and novel challenges in long non-coding RNAs (lncRNAs) functional analysis.” In: *Int J Mol Sci* 13.1 (2012), pp. 97–114. DOI: 10.3390/ijms13010097.
- [132] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski et al. “GENCODE: the reference human genome annotation for The ENCODE Project.” In: *Genome Res* 22.9 (Sept. 2012), pp. 1760–74. DOI: 10.1101/gr.135350.111.
- [133] Chung-Chau Hon, Jordan A Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen J L Rackham, Julian Gough et al. “An atlas of human long non-coding RNAs with accurate 5’ ends.” In: *Nature* 543.7644 (Mar. 2017), pp. 199–204. DOI: 10.1038/nature21374.
- [134] Moran N Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev et al. “Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.” In: *Genes Dev* 25.18 (Sept. 2011), pp. 1915–27. DOI: 10.1101/gad.17446611.
- [135] Sai Luo, J Yuyang Lu, Lichao Liu, Yafei Yin, Chunyan Chen, Xue Han et al. “Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells.” In: *Cell Stem Cell* 18.5 (May 2016), pp. 637–52. DOI: 10.1016/j.stem.2016.01.024.
- [136] Benjamin S Scruggs, Daniel A Gilchrist, Sergei Nechaev, Ginger W Muse, Adam Burkholder, David C Fargo et al. “Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin.” In: *Mol Cell* 58.6 (June 2015), pp. 1101–12. DOI: 10.1016/j.molcel.2015.04.006.
- [137] Mutsumi Kanamori-Katayama, Masayoshi Itoh, Hideya Kawaji, Timo Lassmann, Shintaro Katayama, Miki Kojima et al. “Unamplified cap analysis of gene expression on a single-molecule sequencer.” In: *Genome Res* 21.7 (July 2011), pp. 1150–9. DOI: 10.1101/gr.115469.110.
- [138] Sipko van Dam, Urmo Vösa, Adriaan van der Graaf, Lude Franke and João Pedro de Magalhães. “Gene co-expression analysis for functional classification and gene-disease predictions.” In: *Brief Bioinform* 19.4 (July 2018), pp. 575–592. DOI: 10.1093/bib/bbw139.
- [139] Jiao Yuan, Haiyan Yue, Meiyang Zhang, Jianjun Luo, Lihui Liu, Wei Wu et al. “Transcriptional profiling analysis and functional prediction of long noncoding RNAs in cancer.” In: *Oncotarget* 7.7 (Feb. 2016), pp. 8131–42. DOI: 10.18632/oncotarget.6993.
- [140] Mike J Mason, Guoping Fan, Kathrin Plath, Qing Zhou and Steve Horvath. “Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells.” In: *BMC Genomics* 10 (July 2009), p. 327. DOI: 10.1186/1471-2164-10-327.
- [141] Moritz Schütte, Thomas Risch, Nilofar Abdavi-Azar, Karsten Boehnke, Dirk Schumacher, Marlen Keil et al. “Molecular dissection of colorectal cancer in pre-clinical models identifies biomarkers predicting sensitivity to EGFR inhibitors.” In: *Nat Commun* 8 (Feb. 2017), p. 14262. DOI: 10.1038/ncomms14262.
- [142] Toni M Brand, Mari Iida and Deric L Wheeler. “Molecular mechanisms of resistance to the EGFR monoclonal antibody cetuximab.” In: *Cancer Biol Ther* 11.9 (May 2011), pp. 777–92. DOI: 10.4161/cbt.11.9.15050.

Bibliography

- [143] Simonetta M Leto and Livio Trusolino. "Primary and acquired resistance to EGFR-targeted therapies in colorectal cancer: impact on future treatment strategies." In: *J Mol Med (Berl)* 92.7 (July 2014), pp. 709–22. DOI: 10.1007/s00109-014-1161-2.
- [144] D J Huels and O J Sansom. "Stem vs non-stem cell origin of colorectal cancer." In: *Br J Cancer* 113.1 (June 2015), pp. 1–5. DOI: 10.1038/bjc.2015.214.
- [145] Moisés Blanco-Calvo, Ángel Concha, Angélica Figueroa, Federico Garrido and Manuel Valladares-Ayerbes. "Colorectal Cancer Classification and Cell Heterogeneity: A Systems Oncology Approach." In: *Int J Mol Sci* 16.6 (June 2015), pp. 13610–32. DOI: 10.3390/ijms160613610.
- [146] Andrea Bertotti, Eniko Papp, Siân Jones, Vilmos Adleff, Valsamo Anagnostou, Barbara Lupo et al. "The genomic landscape of response to EGFR blockade in colorectal cancer." In: *Nature* 526.7572 (Oct. 2015), pp. 263–7. DOI: 10.1038/nature14969.
- [147] Markus Morkel, Pamela Riemer, Hendrik Bläker and Christine Sers. "Similar but different: distinct roles for KRAS and BRAF oncogenes in colorectal cancer development and therapy resistance." In: *Oncotarget* 6.25 (Aug. 2015), pp. 20785–800. DOI: 10.18632/oncotarget.4750.
- [148] Muhammad W Saif. "Colorectal cancer in review: the role of the EGFR pathway." In: *Expert Opin Investig Drugs* 19.3 (Mar. 2010), pp. 357–69. DOI: 10.1517/13543781003593962.
- [149] Felipe De Sousa E Melo, Xin Wang, Marnix Jansen, Evelyn Fessler, Anne Trinh, Laura P M H de Rooij et al. "Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions." In: *Nat Med* 19.5 (May 2013), pp. 614–8. DOI: 10.1038/nm.3174.
- [150] Laetitia Marisa, Aurélien de Reyniès, Alex Duval, Janick Selves, Marie Pierre Gaub, Laure Vescovo et al. "Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value." In: *PLoS Med* 10.5 (2013), e1001453. DOI: 10.1371/journal.pmed.1001453.
- [151] Anguraj Sadanandam, Costas A Lyssiotis, Krisztian Homicsko, Eric A Collisson, William J Gibb, Stephan Wullschleger et al. "A colorectal cancer classification system that associates cellular phenotype and responses to therapy." In: *Nat Med* 19.5 (May 2013), pp. 619–25. DOI: 10.1038/nm.3175.
- [152] Andreas Schlicker, Garry Beran, Christine M Chresta, Gael McWalter, Alison Pritchard, Susie Weston et al. "Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines." In: *BMC Med Genomics* 5 (Dec. 2012), p. 66. DOI: 10.1186/1755-8794-5-66.
- [153] Daniel Vallböhmer, Syma Iqbal, Dong Yun Yang, Katrin E Rhodes, Wu Zhang, Michael Gordon et al. "Molecular determinants of irinotecan efficacy." In: *Int J Cancer* 119.10 (Nov. 2006), pp. 2435–42. DOI: 10.1002/ijc.22129.
- [154] D G Haller. "Safety of oxaliplatin in the treatment of colorectal cancer." In: *Oncology (Williston Park)* 14.12 Suppl 11 (Dec. 2000), pp. 15–20.
- [155] Valerie M Nelson and Al B Benson. "Status of targeted therapies in the adjuvant treatment of colon cancer." In: *J Gastrointest Oncol* 4.3 (Sept. 2013), pp. 245–52. DOI: 10.3978/j.issn.2078-6891.2013.035.
- [156] Eric Van Cutsem, Claus-Henning Köhne, István Láng, Gunnar Folprecht, Marek P Nowacki, Stefano Cascinu et al. "Cetuximab plus irinotecan, fluorouracil, and leucovorin as first-line treatment for metastatic colorectal cancer: updated analysis of overall survival according to tumor KRAS and BRAF mutation status." In: *J Clin Oncol* 29.15 (May 2011), pp. 2011–9. DOI: 10.1200/JCO.2010.33.5091.

- [157] Carsten Bokemeyer, Eric Van Cutsem, Philippe Rougier, Fortunato Ciardiello, Steffen Heeger, Michael Schlichting et al. "Addition of cetuximab to chemotherapy as first-line treatment for KRAS wild-type metastatic colorectal cancer: pooled analysis of the CRYSTAL and OPUS randomised clinical trials." In: *Eur J Cancer* 48.10 (July 2012), pp. 1466–75. DOI: 10.1016/j.ejca.2012.02.057.
- [158] Wendy De Roock, Bart Claes, David Bernasconi, Jef De Schutter, Bart Biesmans, George Fountzilas et al. "Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis." In: *Lancet Oncol* 11.8 (Aug. 2010), pp. 753–62. DOI: 10.1016/S1470-2045(10)70130-3.
- [159] Andrea Sartore-Bianchi, Miriam Martini, Francesca Molinari, Silvio Veronese, Michele Nichelatti, Salvatore Artale et al. "PIK3CA mutations in colorectal cancer are associated with clinical resistance to EGFR-targeted monoclonal antibodies." In: *Cancer Res* 69.5 (Mar. 2009), pp. 1851–7. DOI: 10.1158/0008-5472.CAN-08-2466.
- [160] Shirin Khambata-Ford, Christopher R Garrett, Neal J Meropol, Mark Basik, Christopher T Harbison, Shujian Wu et al. "Expression of epiregulin and amphiregulin and K-ras mutation status predict disease control in metastatic colorectal cancer patients treated with cetuximab." In: *J Clin Oncol* 25.22 (Aug. 2007), pp. 3230–7. DOI: 10.1200/JCO.2006.10.5437.
- [161] Madeline A Lancaster and Juergen A Knoblich. "Organogenesis in a dish: modeling development and disease using organoid technologies." In: *Science* 345.6194 (July 2014), p. 1247125. DOI: 10.1126/science.1247125.
- [162] Danielle M Burgenske, David J Monsma, Dawna Dylewski, Stephanie B Scott, Aaron D Sayfie, Donald G Kim et al. "Establishment of genetically diverse patient-derived xenografts of colorectal cancer." In: *Am J Cancer Res* 4.6 (2014), pp. 824–37.
- [163] Hui Gao, Joshua M Korn, Stéphane Ferretti, John E Monahan, Youzhen Wang, Mallika Singh et al. "High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response." In: *Nat Med* 21.11 (Nov. 2015), pp. 1318–25. DOI: 10.1038/nm.3954.
- [164] Sylvia Julien, Ana Merino-Trigo, Ludovic Lacroix, Marc Pocard, Diane Goéré, Pascale Mariani et al. "Characterization of a large panel of patient-derived tumor xenografts representing the clinical heterogeneity of human colorectal cancer." In: *Clin Cancer Res* 18.19 (Oct. 2012), pp. 5314–28. DOI: 10.1158/1078-0432.CCR-12-0372.
- [165] Enzo Medico, Mariangela Russo, Gabriele Picco, Carlotta Cancelliere, Emanuele Valtorta, Giorgio Corti et al. "The molecular landscape of colorectal cancer cell lines unveils clinically actionable kinase targets." In: *Nat Commun* 6 (Apr. 2015), p. 7002. DOI: 10.1038/ncomms8002.
- [166] Manoel Nunes, Patricia Vrignaud, Sophie Vacher, Sophie Richon, Astrid Lièvre, Wulfran Cacheux et al. "Evaluating patient-derived colorectal cancer xenografts as preclinical models by comparison with patient clinical data." In: *Cancer Res* 75.8 (Apr. 2015), pp. 1560–6. DOI: 10.1158/0008-5472.CAN-14-1590.
- [167] Marc van de Wetering, Hayley E Francies, Joshua M Francis, Gergana Bounova, Francesco Iorio, Apollo Pronk et al. "Prospective derivation of a living organoid biobank of colorectal cancer patients." In: *Cell* 161.4 (May 2015), pp. 933–45. DOI: 10.1016/j.cell.2015.03.053.
- [168] Jean-Yves Douillard, Kelly S Oliner, Salvatore Siena, Josep Tabernero, Ronald Burkes, Mario Barugel et al. "Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer." In: *N Engl J Med* 369.11 (Sept. 2013), pp. 1023–34. DOI: 10.1056/NEJMoa1305275.
- [169] Louisa Rafa, Anne-Frédérique Dessein, Louise Devisme, David Buob, Stéphanie Truant, Nicole Porchet et al. "REG4 acts as a mitogenic, motility and pro-invasive factor for colon cancer cells." In: *Int J Oncol* 36.3 (Mar. 2010), pp. 689–98. DOI: 10.3892/ijo_00000544.

Bibliography

- [170] Kumar S Bishnupuri, Qizhi Luo, Nabendu Murmu, Courtney W Houchen, Shrikant Anant and Brian K Dieckgraefe. “Reg IV activates the epidermal growth factor receptor/Akt/AP-1 signaling pathway in colon adenocarcinomas.” In: *Gastroenterology* 130.1 (Jan. 2006), pp. 137–49. DOI: 10.1053/j.gastro.2005.10.001.
- [171] D J Jonker, C S Karapetis, C Harbison, C J O’Callaghan, D Tu, R J Simes et al. “Epiregulin gene expression as a biomarker of benefit from cetuximab in the treatment of advanced colorectal cancer.” In: *Br J Cancer* 110.3 (Feb. 2014), pp. 648–55. DOI: 10.1038/bjc.2013.753.
- [172] Y Takahashi, P Sheridan, A Niida, G Sawada, R Uchi, H Mizuno et al. “The AURKA/TPX2 axis drives colon tumorigenesis cooperatively with MYC.” In: *Ann Oncol* 26.5 (May 2015), pp. 935–942. DOI: 10.1093/annonc/mdv034.
- [173] Scott Kopetz. “Targeting SRC and epidermal growth factor receptor in colorectal cancer: rationale and progress into the clinic.” In: *Gastrointest Cancer Res* 1.4 Suppl 2 (2007), S37–41.
- [174] Paulina Pechańska, M. Becker, T. Mayr, B. Hinzmann, H.-P. Adams, I. Klamann et al. “Mutation Status of KRAS, BRAF, PIK3CA and Expression Level of AREG and EREG Identify Responders to Cetuximab in a Large Panel of Patient Derived Colorectal Carcinoma Xenografts of All Four UICC Stages”. In: *J. Cancer Ther.* 04.02 (2013), pp. 678–693. ISSN: 2151-1934. DOI: 10.4236/jct.2013.42083.
- [175] Dmitri Parkhomchuk, Tatiana Borodina, Vyacheslav Amstislavskiy, Maria Banaru, Linda Hallen, Sylvia Krobitsch et al. “Transcriptome analysis by strand-specific sequencing of complementary DNA.” In: *Nucleic Acids Res* 37.18 (Oct. 2009), e123. DOI: 10.1093/nar/gkp596.
- [176] Heng Li and Richard Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform.” In: *Bioinformatics* 25.14 (July 2009), pp. 1754–60. DOI: 10.1093/bioinformatics/btp324.
- [177] Ruibin Xi, Angela G Hadjipanayis, Lovelace J Luquette, Tae-Min Kim, Eunjung Lee, Jianhua Zhang et al. “Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion.” In: *Proc Natl Acad Sci U S A* 108.46 (Nov. 2011), E1128–36. DOI: 10.1073/pnas.1110574108.
- [178] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin et al. “VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.” In: *Genome Res* 22.3 (Mar. 2012), pp. 568–76. DOI: 10.1101/gr.129684.111.
- [179] Cornelis A Albers, Gerton Lunter, Daniel G MacArthur, Gilean McVean, Willem H Ouwehand and Richard Durbin. “Dindel: accurate indel calls from short-read data.” In: *Genome Res* 21.6 (June 2011), pp. 961–73. DOI: 10.1101/gr.112326.110.
- [180] I Fichtner, W Slisow, J Gill, M Becker, B Elbe, T Hillebrand et al. “Anticancer drug response and expression of molecular markers in early-passage xenotransplanted colon carcinomas.” In: *Eur J Cancer* 40.2 (Jan. 2004), pp. 298–307. DOI: 10.1016/j.ejca.2003.10.011.
- [181] In Sock Jang, Elias Chaibub Neto, Juistin Guinney, Stephen H Friend and Adam A Margolin. “Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data.” In: *Pac Symp Biocomput* (2014), pp. 63–74.
- [182] Sarah Vluymans. “Learning from unbalanced data”. In: *Stud. Comput. Intell.* 807.9 (2019), pp. 81–110. ISSN: 1860949X. DOI: 10.1007/978-3-030-04663-7_4.
- [183] Ramkiran K. Gouripeddi, V. N. Balasubramanian, S. Panchanathan, J. Harris, A. Bhaskaran and R. M. Siegel. “Ranking predictors of complications following a drug eluting stent procedure using support vector machines”. In: *Comput. Cardiol.* 36 (2009), pp. 345–348. ISSN: 02766574.
- [184] Sebastián Maldonado, Richard Weber and Fazel Famili. “Feature selection for high-dimensional class-unbalanced data sets using Support Vector Machines”. In: *Inf. Sci. (Ny)*. 286 (2014), pp. 228–246. ISSN: 00200255. DOI: 10.1016/j.ins.2014.07.015.

- [185] Rok Blagus and Lara Lusa. "SMOTE for high-dimensional class-imbalanced data." In: *BMC Bioinformatics* 14 (Mar. 2013), p. 106. DOI: 10.1186/1471-2105-14-106.
- [186] Jiapeng Yin, Jian Hou, Zhiyong She, Chengming Yang and Han Yu. "Improving the performance of SVM-RFE on classification of pancreatic cancer data". In: *Proceedings of the IEEE International Conference on Industrial Technology* 2016-May (2016), pp. 956–961. DOI: 10.1109/ICIT.2016.7474881.
- [187] Isabelle Guyon, Jason Weston, Stephen Barnhill and Vladimir Vapnik. "Gene selection for cancer classification using support vector machines". In: *Machine Learning* 46.1-3 (2002), pp. 389–422. ISSN: 08856125. DOI: 10.1023/A:1012487302797.
- [188] Li Wang, Ji Zhu and Hui Zou. "Hybrid huberized support vector machines for microarray classification and gene selection." In: *Bioinformatics* 24.3 (Feb. 2008), pp. 412–9. DOI: 10.1093/bioinformatics/btm579.
- [189] Anthony W Tolcher. "Stable disease is a valid end point in clinical trials." In: *Cancer J* 15.5 (), pp. 374–8. DOI: 10.1097/PP0.0b013e3181b0bb05.
- [190] Justin M Balko and Esther P Black. "A gene expression predictor of response to EGFR-targeted therapy stratifies progression-free survival to cetuximab in KRAS wild-type metastatic colorectal cancer." In: *BMC Cancer* 9 (May 2009), p. 145. DOI: 10.1186/1471-2407-9-145.
- [191] Jesús García-Foncillas, Yu Sunakawa, Dan Aderka, Zev Wainberg, Philippe Ronga, Pauline Witzler et al. "Distinguishing Features of Cetuximab and Panitumumab in Colorectal Cancer and Other Solid Tumors." In: *Front Oncol* 9 (2019), p. 849. DOI: 10.3389/fonc.2019.00849.
- [192] K Trumpi, I Ubink, A Trinh, M Djafarihamedani, J M Jongen, K M Govaert et al. "Neoadjuvant chemotherapy affects molecular classification of colorectal tumors." In: *Oncogenesis* 6.7 (July 2017), e357. DOI: 10.1038/oncsis.2017.48.
- [193] Kyle Juraschka and Michael D Taylor. "Medulloblastoma in the age of molecular subgroups: a review." In: *Journal of neurosurgery. Pediatrics* 24.4 (2019), pp. 353–363. ISSN: 1933-0715. DOI: 10.3171/2019.5.PEDS18381.
- [194] Susan Standring, ed. *Gray's Anatomy: The Anatomical Basis of Clinical Practice (41st ed.)* Elsevier, Sept. 2015.
- [195] Hassan Marzban, Marc R Del Bigio, Javad Alizadeh, Saeid Ghavami, Robby M Zachariah and Mojgan Rastegar. "Cellular commitment in the developing cerebellum." In: *Front. Cell. Neurosci.* 8.January (2014), p. 450. ISSN: 1662-5102. DOI: 10.3389/fncel.2014.00450.
- [196] Thomas Butts, Mary J Green and Richard J T Wingate. "Development of the cerebellum: simple steps to make a 'little brain'". In: *Development* 141.21 (2014), pp. 4031–4041. DOI: 10.1242/dev.106559.
- [197] Mitsuhiro Hashimoto and Masahiko Hibi. "Development and evolution of cerebellar neural circuits." In: *Dev Growth Differ* 54.3 (Apr. 2012), pp. 373–89. DOI: 10.1111/j.1440-169X.2012.01348.x.
- [198] Salvador Martinez, Abraham Andreu, Nora Mecklenburg and Diego Echevarria. "Cellular and molecular basis of cerebellar development." In: *Front Neuroanat* 7 (2013), p. 18. DOI: 10.3389/fnana.2013.00018.
- [199] Henry Gray. *Anatomy of the Human Body*. 20th ed. Philadelphia and New York: Lea and Febiger, 1918.
- [200] Joshua J White and Roy V Sillitoe. "Development of the cerebellum: from gene expression patterns to circuit maps." In: *Wiley Interdiscip Rev Dev Biol* 2.1 (), pp. 149–64. DOI: 10.1002/wdev.65.

Bibliography

- [201] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard et al. "Genome-wide atlas of gene expression in the adult mouse brain." In: *Nature* 445.7124 (Jan. 2007), pp. 168–76. DOI: 10.1038/nature05453.
- [202] Daniver Morales and Mary E Hatten. "Molecular markers of neuronal progenitors in the embryonic cerebellar anlage." In: *J Neurosci* 26.47 (Nov. 2006), pp. 12226–36. DOI: 10.1523/JNEUROSCI.3493-06.2006.
- [203] Antoine Forget, Laure Bihannic, Sara Maria Cigna, Coralie Lefevre, Marc Remke, Monia Barnat et al. "Shh signaling protects Atoh1 from degradation mediated by the E3 ubiquitin ligase Huwe1 in neural precursors." In: *Dev. Cell* 29.6 (2014), pp. 649–61. DOI: 10.1016/j.devcel.2014.05.014.
- [204] Martine F Roussel and Mary E Hatten. "Cerebellum development and medulloblastoma." In: *Curr Top Dev Biol* 94 (2011), pp. 235–82. DOI: 10.1016/B978-0-12-380916-2.00008-5.
- [205] J.K. Fahrion, Y. Komuro, N. Ohno, Y. Littner, C. Nelson, T. Kumada et al. "Cerebellar Patterning". In: *Patterning Cell Type Specif. Dev. CNS PNS*. Vol. 1. Elsevier, 2013, pp. 211–225. ISBN: 9780123972651. DOI: 10.1016/B978-0-12-397265-1.00042-3.
- [206] David N Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K Cavenee et al. "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary." In: *Acta Neuropathol* 131.6 (June 2016), pp. 803–20. DOI: 10.1007/s00401-016-1545-1.
- [207] Marcel Kool, Jan Koster, Jens Bunt, Nancy E Hasselt, Arjan Lakeman, Peter van Sluis et al. "Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features." In: *PLoS One* 3.8 (Aug. 2008), e3088. DOI: 10.1371/journal.pone.0003088.
- [208] Yoon-Jae Cho, Aviad Tsherniak, Pablo Tamayo, Sandro Santagata, Azra Ligon, Heidi Greulich et al. "Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome." In: *J Clin Oncol* 29.11 (Apr. 2011), pp. 1424–30. DOI: 10.1200/JCO.2010.28.5148.
- [209] Margaret C Thompson, Christine Fuller, Twala L Hogg, James Dalton, David Finkelstein, Ching C Lau et al. "Genomics identifies medulloblastoma subgroups that are enriched for specific genetic alterations." In: *J Clin Oncol* 24.12 (Apr. 2006), pp. 1924–31. DOI: 10.1200/JCO.2005.04.4974.
- [210] Marcel Kool, Andrey Korshunov, Marc Remke, David T W Jones, Maria Schlanstein, Paul A Northcott et al. "Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas." In: *Acta Neuropathol* 123.4 (Apr. 2012), pp. 473–84. DOI: 10.1007/s00401-012-0958-8.
- [211] Julia Ahlfeld, Rebecca Favaro, Pierfrancesco Pagella, Hans A Kretzschmar, Silvia Nicolis and Ulrich Schüller. "Sox2 requirement in sonic hedgehog-associated medulloblastoma." In: *Cancer Res* 73.12 (June 2013), pp. 3796–807. DOI: 10.1158/0008-5472.CAN-13-0238.
- [212] Marcel Kool, David T W Jones, Natalie Jäger, Paul A Northcott, Trevor J Pugh, Volker Hovestadt et al. "Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothed inhibition." In: *Cancer Cell* 25.3 (Mar. 2014), pp. 393–405. DOI: 10.1016/j.ccr.2014.02.004.
- [213] Paul A Northcott, David J H Shih, John Peacock, Livia Garzia, A Sorana Morrissy, Thomas Zichner et al. "Subgroup-specific structural variation across 1,000 medulloblastoma genomes." In: *Nature* 488.7409 (Aug. 2012), pp. 49–56. DOI: 10.1038/nature11327.

- [214] Paul A Northcott, Thomas Hielscher, Adrian Dubuc, Stephen Mack, David Shih, Marc Remke et al. “Pediatric and adult sonic hedgehog medulloblastomas are clinically and molecularly distinct.” In: *Acta Neuropathol* 122.2 (Aug. 2011), pp. 231–40. DOI: 10.1007/s00401-011-0846-7.
- [215] Joseph A. Brzezinski and Thomas A. Reh. “Photoreceptor cell fate specification in vertebrates”. In: *Development* 142.19 (2015), pp. 3263–3273. DOI: 10.1242/dev.127043.
- [216] David Capper, David T W Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm et al. “DNA methylation-based classification of central nervous system tumours.” In: *Nature* 555.7697 (Mar. 2018), pp. 469–474. DOI: 10.1038/nature26000.
- [217] Andrey Korshunov, Lukas Chavez, Paul A Northcott, Tanvi Sharma, Marina Ryzhova, David T W Jones et al. “DNA-methylation profiling discloses significant advantages over NanoString method for molecular classification of medulloblastoma.” In: *Acta Neuropathol* 134.6 (Dec. 2017), pp. 965–967. DOI: 10.1007/s00401-017-1776-9.
- [218] Charles Y Lin, Serap Erkek, Yiai Tong, Linlin Yin, Alexander J Federation, Marc Zapatka et al. “Active medulloblastoma enhancers reveal subgroup-specific cellular origins.” In: *Nature* 530.7588 (Feb. 2016), pp. 57–62. DOI: 10.1038/nature16546.
- [219] Peter H L Krijger, Geert Geeven, Valerio Bianchi, Catharina R E Hilvering and Wouter de Laat. “4C-seq from beginning to end: A detailed protocol for sample preparation and data analysis.” In: *Methods* 170 (Jan. 2020), pp. 17–32. DOI: 10.1016/j.ymeth.2019.07.014.
- [220] Walderik W Zomeran, Sabine L A Plasschaert, Siobhan Conroy, Frank J Scherpen, Tiny G J Meeuwssen de Boer, Harm J Lourens et al. “Identification of Two Protein-Signaling States Delineating Transcriptionally Heterogeneous Human Medulloblastoma.” In: *Cell Rep* 22.12 (Mar. 2018), pp. 3206–3216. DOI: 10.1016/j.celrep.2018.02.089.
- [221] Florence M.G. Cavalli, Marc Remke, Ladislav Rampasek, John Peacock, David J.H. Shih, Betty Luu et al. “Intertumoral Heterogeneity within Medulloblastoma Subgroups”. In: *Cancer Cell* 31.6 (2017), 737–754.e6. DOI: 10.1016/j.cccell.2017.05.005.
- [222] Paul A. Northcott, Ivo Buchhalter, A. Sorana Morrissy, Volker Hovestadt, Joachim Weischenfeldt, Tobias Ehrenberger et al. “The whole-genome landscape of medulloblastoma subtypes”. In: *Nature* 547.7663 (2017), pp. 311–317. ISSN: 0028-0836. DOI: 10.1038/nature22973.
- [223] Edward C Schwalbe, Janet C Lindsey, Sirintra Nakjang, Stephen Crosier, Amanda J Smith, Debbie Hicks et al. “Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study.” In: *Lancet. Oncol.* 18.7 (2017), pp. 958–971. DOI: 10.1016/S1470-2045(17)30243-7.
- [224] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: KDD’96. Portland, Oregon: AAAI Press, 1996, 226–231.
- [225] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno et al. “Similarity network fusion for aggregating data types on a genomic scale.” In: *Nat Methods* 11.3 (Mar. 2014), pp. 333–7. DOI: 10.1038/nmeth.2810.
- [226] Gaylor Boulay, Mary E. Awad, Nicolo Riggi, Tenley C. Archer, Sowmya Iyer, Wannaporn E. Boonseng et al. “OTX2 activity at distal regulatory elements shapes the chromatin landscape of group 3 medulloblastoma”. In: *Cancer Discov.* 7.3 (2017), pp. 288–301. DOI: 10.1158/2159-8290.CD-16-0844.
- [227] R C Rostomily, O Bermingham-McDonogh, M S Berger, S J Tapscott, T A Reh and J M Olson. “Expression of neurogenic basic helix-loop-helix genes in primitive neuroectodermal tumors.” In: *Cancer Res* 57.16 (Aug. 1997), pp. 3526–31.

Bibliography

- [228] Kin-Mang Lau, Queeny Kwan Yi Chan, Jesse C S Pang, Fanny Man-Ting Ma, Kay K W Li, Walter Wai Yeung et al. "Overexpression of HMGA1 deregulates tumor growth via cdc25A and alters migration/invasion through a cdc25A-independent pathway in medulloblastoma." In: *Acta Neuropathol* 123.4 (Apr. 2012), pp. 553–71. DOI: 10.1007/s00401-011-0934-8.
- [229] Maria Lastowska, Hani Al-Afghani, Haya H Al-Balool, Harsh Sheth, Emma Mercer, Jonathan M Coxhead et al. "Identification of a neuronal transcription factor network involved in medulloblastoma development." In: *Acta Neuropathol Commun* 1 (July 2013), p. 35. DOI: 10.1186/2051-5960-1-35.
- [230] Cornelia M Hooper, Susan M Hawes, Ursula R Kees, Nicholas G Gottardo and Peter B Dallas. "Gene expression analyses of the spatio-temporal relationships of human medulloblastoma subgroups during early human neurogenesis." In: *PLoS One* 9.11 (2014), e112909. DOI: 10.1371/journal.pone.0112909.
- [231] Kay Ka-Wai Li, Yan Qi, Tian Xia, Yu Yao, Liangfu Zhou, Kin-Mang Lau et al. "CRMP1 Inhibits Proliferation of Medulloblastoma and Is Regulated by HMGA1." In: *PLoS One* 10.5 (2015), e0127910. DOI: 10.1371/journal.pone.0127910.
- [232] Sivan Gershanov, Helen Toledano, Shalom Michowiz, Orit Barinfeld, Albert Pinhasov, Nitza Goldenberg-Cohen et al. "MicroRNA-mRNA expression profiles associated with medulloblastoma subgroup 4." In: *Cancer Manag Res* 10 (2018), pp. 339–352. DOI: 10.2147/CMAR.S156709.
- [233] Rosa Angélica Castillo-Rodríguez, Víctor Manuel Dávila-Borja and Sergio Juárez-Méndez. "Data mining of pediatric medulloblastoma microarray expression reveals a novel potential subdivision of the Group 4 molecular subgroup." In: *Oncol Lett* 15.5 (May 2018), pp. 6241–6250. DOI: 10.3892/ol.2018.8094.
- [234] Tripti Gaur, Christopher J. Lengner, Hayk Hovhannisyan, Ramesh A. Bhat, Peter V N Bodine, Barry S. Komm et al. "Canonical WNT signaling promotes osteogenesis by directly stimulating Runx2 gene expression." In: *J. Biol. Chem.* 280.39 (2005), pp. 33132–40. DOI: 10.1074/jbc.M500608200.
- [235] Ian J. Huggins, Tomas Bos, Olivia Gaylord, Christina Jessen, Brianna Lonquich, Angeline Puranen et al. "The WNT target SP5 negatively regulates WNT transcriptional programs in human pluripotent stem cells." In: *Nat. Commun.* 8.1 (2017), p. 1034. DOI: 10.1038/s41467-017-01203-1.
- [236] Y. Zhai, A. Iura, S. Yeasmin, A. B. Wiese, R. Wu, Y. Feng et al. "MSX2 is an oncogenic downstream target of activated WNT signaling in ovarian endometrioid adenocarcinoma." In: *Oncogene* 30.40 (2011), pp. 4152–62. DOI: 10.1038/onc.2011.123.
- [237] Olivier Ayrault, Haotian Zhao, Frederique Zindy, Chunxu Qu, Charles J. Sherr and Martine F. Roussel. "Atoh1 inhibits neuronal differentiation and collaborates with Gli1 to generate medulloblastoma-initiating cells." In: *Cancer Res.* 70.13 (2010), pp. 5618–27. DOI: 10.1158/0008-5472.CAN-09-3740.
- [238] Yun-Hua Zhong, Hong-Zhong Cheng, Hao Peng, Shi-Cong Tang and Ping Wang. "Heat shock factor 2 is associated with the occurrence of lung cancer by enhancing the expression of heat shock proteins." In: *Oncol Lett* 12.6 (Dec. 2016), pp. 5106–5112. DOI: 10.3892/ol.2016.5368.
- [239] J. K. Björk, M. Åkerfelt, J. Joutsen, M. C. Puustinen, F. Cheng, L. Sistonen et al. "Heat-shock factor 2 is a suppressor of prostate cancer invasion." In: *Oncogene* 35.14 (2016), pp. 1770–84. DOI: 10.1038/onc.2015.241.
- [240] Erica Riveiro-Falkenbach and María S Soengas. "Control of tumorigenesis and chemoresistance by the DEK oncogene." In: *Clin Cancer Res* 16.11 (June 2010), pp. 2932–8. DOI: 10.1158/1078-0432.CCR-09-2330.

- [241] Sonia Coni, Anna Barbara Mancuso, Laura Di Magno, Giulia Sdruscia, Simona Manni, Silvia Maria Serrao et al. “Selective targeting of HDAC1/2 elicits anticancer effects through Gli1 acetylation in preclinical models of SHH Medulloblastoma.” In: *Sci. Rep.* 7.February (2017), p. 44079. DOI: 10.1038/srep44079.
- [242] Qing-liang Wang, Shiming Chen, Noriko Esumi, Prabodh K Swain, Heidi S Haines, Guanghua Peng et al. “QRX, a novel homeobox gene, modulates photoreceptor gene expression.” In: *Hum. Mol. Genet.* 13.10 (2004), pp. 1025–40. DOI: 10.1093/hmg/ddh117.
- [243] C.-M. Amy Chen and Constance L Cepko. “The chicken RaxL gene plays a role in the initiation of photoreceptor differentiation.” In: *Development* 129.23 (2002), pp. 5363–75. DOI: 10.1242/dev.00114.
- [244] Liang Ming, Ronit Wilk, Bruce H Reed and Howard D Lipshitz. “Drosophila Hindsight and mammalian RREB-1 are evolutionarily conserved DNA-binding transcriptional attenuators.” In: *Differentiation* 86.4-5 (), pp. 159–70. DOI: 10.1016/j.diff.2013.12.001.
- [245] Bingqing Hui, Hao Ji, Yetao Xu, Juan Wang, Zhonghua Ma, Chongguo Zhang et al. “RREB1-induced upregulation of the lncRNA AGAP2-AS1 regulates the proliferation and migration of pancreatic cancer partly through suppressing ANKRD1 and ANGPTL4.” In: *Cell Death Dis.* 10.3 (2019), p. 207. DOI: 10.1038/s41419-019-1384-9.
- [246] Jens Bunt, Nancy E. Hasselt, Danny A. Zwijnenburg, Mohamed Hamdi, Jan Koster, Rogier Versteeg et al. “OTX2 directly activates cell cycle genes and inhibits differentiation in medulloblastoma cells.” In: *Int. J. cancer* 131.2 (2012), E21–32. DOI: 10.1002/ijc.26474.
- [247] Ettore Salsano, Laura Croci, Emanuela Maderna, Linda Lupo, Bianca Pollo, Maria Teresa Giordana et al. “Expression of the neurogenic basic helix-loop-helix transcription factor NEUROG1 identifies a subgroup of medulloblastomas not expressing ATOH1.” In: *Neuro. Oncol.* 9.3 (2007), pp. 298–307. DOI: 10.1215/15228517-2007-014.
- [248] Edwin A. Obana, Travis G. Lundell, Kevin J. Yi, Kryslaine L. Radomski, Qiong Zhou and Martin L. Doughty. “Neurog1 Genetic Inducible Fate Mapping (GIFM) Reveals the Existence of Complex Spatiotemporal Cyto-Architectures in the Developing Cerebellum”. In: *Cerebellum* 14.3 (2015), pp. 247–263. ISSN: 14734230. DOI: 10.1007/s12311-014-0641-9.
- [249] Eui-seok J. Kim, Kei Hori, Alex Wyckoff, Lauren K. Dickel, Edmund J. Koundakjian, Lisa V. Goodrich et al. “Spatiotemporal fate map of neurogenin1 (Neurog1) lineages in the mouse central nervous system”. In: *J. Comp. Neurol.* 519.7 (2011), pp. 1355–1370. DOI: 10.1002/cne.22574.
- [250] “Neurogenin1 expression in cell lineages of the cerebellar cortex in embryonic and postnatal mice”. In: *Dev. Dyn.* 238.12 (2009), pp. 3310–3325. DOI: 10.1002/dvdy.22165.
- [251] David T. W. Jones, Natalie Jäger, Marcel Kool, Thomas Zichner, Barbara Hutter, Marc Sultan et al. “Dissecting the genomic complexity underlying medulloblastoma”. In: *Nature* 488.7409 (2012), pp. 100–105. DOI: 10.1038/nature11284. eprint: NIHMS150003.
- [252] Andrew J Fink, Chris Englund, Ray A M Daza, Diane Pham, Charmaine Lau, Mary Nivison et al. “Development of the deep cerebellar nuclei: transcription factors and cell migration from the rhombic lip.” In: *J Neurosci* 26.11 (Mar. 2006), pp. 3066–76. DOI: 10.1523/JNEUROSCI.5203-05.2006.
- [253] Valérie Baubet, Chaomei Xiang, Aliah Molczan, Laura Roccograndi, Svetlana Melamed and Nadia Dahmane. “Rp58 is essential for the growth and patterning of the cerebellum and for glutamatergic and GABAergic neuron development.” In: *Development* 139.11 (June 2012), pp. 1903–9. DOI: 10.1242/dev.075606.

Bibliography

- [254] J M Olson, A Asakura, L Snider, R Hawkes, A Strand, J Stoeck et al. “NeuroD2 is necessary for development and survival of central nervous system neurons.” In: *Dev Biol* 234.1 (June 2001), pp. 174–87. DOI: 10.1006/dbio.2001.0245.
- [255] Chris M Egan, Ulrika Nyman, Julie Skotte, Gundula Streubel, Siobhán Turner, David J O’Connell et al. “CHD5 is required for neurogenesis and has a dual role in facilitating gene expression and polycomb gene repression.” In: *Dev Cell* 26.3 (Aug. 2013), pp. 223–36. DOI: 10.1016/j.devcel.2013.07.008.
- [256] Hani Al-Halabi, Andre Nantel, Almos Klekner, Marie-Christine Guiot, Steffen Albrecht, Peter Hauser et al. “Preponderance of sonic hedgehog pathway activation characterizes adult medulloblastoma.” In: *Acta Neuropathol* 121.2 (Feb. 2011), pp. 229–39. DOI: 10.1007/s00401-010-0780-0.
- [257] Moritz Mall, Michael S Kareta, Soham Chanda, Henrik Ahlenius, Nicholas Perotti, Bo Zhou et al. “Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates.” In: *Nature* 544.7649 (Apr. 2017), pp. 245–249. DOI: 10.1038/nature21722.
- [258] Kuan Shen Wong, Kira Rehn, Sharina Palencia-Desai, Vikram Kohli, Wynn Hunter, Juli D Uhl et al. “Hedgehog signaling is required for differentiation of endocardial progenitors in zebrafish.” In: *Dev Biol* 361.2 (Jan. 2012), pp. 377–91. DOI: 10.1016/j.ydbio.2011.11.004.
- [259] Masataka Asagiri and Hiroshi Takayanagi. “The molecular understanding of osteoclast differentiation”. In: *Bone* 40.2 (2007), pp. 251–264. ISSN: 87563282. DOI: 10.1016/j.bone.2006.09.023.
- [260] Abby Sarkar and Konrad Hochedlinger. “The sox family of transcription factors: versatile regulators of stem and progenitor cell fate.” In: *Cell Stem Cell* 12.1 (Jan. 2013), pp. 15–30. DOI: 10.1016/j.stem.2012.12.007.
- [261] Charlotte E Scott, Sarah L Wynn, Abdul Sesay, Catarina Cruz, Martin Cheung, Maria-Victoria Gomez Gaviro et al. “SOX9 induces and maintains neural stem cells.” In: *Nat Neurosci* 13.10 (Oct. 2010), pp. 1181–9. DOI: 10.1038/nn.2646.
- [262] Jürgen Dittmer. “The biology of the Ets1 proto-oncogene.” In: *Mol Cancer* 2 (Aug. 2003), p. 29. DOI: 10.1186/1476-4598-2-29.
- [263] Michael Jeltsch, Veli-Matti Leppänen, Pipsa Saharinen and Kari Alitalo. “Receptor tyrosine kinase-mediated angiogenesis.” In: *Cold Spring Harb Perspect Biol* 5.9 (Sept. 2013). DOI: 10.1101/cshperspect.a009183.
- [264] Martine Uittenbogaard, Kristin K Baxter and Anne Chiaramello. “NeuroD6 genomic signature bridging neuronal differentiation to survival via the molecular chaperone network.” In: *J Neurosci Res* 88.1 (Jan. 2010), pp. 33–54. DOI: 10.1002/jnr.22182.
- [265] Francisca F Vasconcelos, Alessandro Sessa, Cátia Laranjeira, Alexandre A S F Raposo, Vera Teixeira, Daniel W Hagey et al. “MyT1 Counteracts the Neural Progenitor Program to Promote Vertebrate Neurogenesis.” In: *Cell Rep* 17.2 (Oct. 2016), pp. 469–483. DOI: 10.1016/j.celrep.2016.09.024.
- [266] Abhijeet Pataskar, Johannes Jung, Pawel Smialowski, Florian Noack, Federico Calegari, Tobias Straub et al. “NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program.” In: *EMBO J* 35.1 (Jan. 2016), pp. 24–45. DOI: 10.15252/embj.201591206.
- [267] Raffaella Spina, Gessica Filocamo, Enrico Iaccino, Stefania Scicchitano, Michela Lupia, Emanuela Chiarella et al. “Critical role of zinc finger protein 521 in the control of growth, clonogenicity and tumorigenic potential of medulloblastoma cells.” In: *Oncotarget* 4.8 (Aug. 2013), pp. 1280–92. DOI: 10.18632/oncotarget.1176.

- [268] Alexandra Garancher, Charles Y. Lin, Morgane Morabito, Wilfrid Richer, Nathalie Rocques, Magalie Larcher et al. “NRL and CRX Define Photoreceptor Identity and Reveal Subgroup-Specific Dependencies in Medulloblastoma.” In: *Cancer Cell* 33.3 (2018), 435–449.e6. DOI: 10.1016/j.ccell.2018.02.006.
- [269] Martine F Roussel and Giles W Robinson. “Role of MYC in Medulloblastoma.” In: *Cold Spring Harb Perspect Med* 3.11 (Nov. 2013). DOI: 10.1101/cshperspect.a014308.
- [270] Masahiro Kawahara, Ashley Pandolfi, Boris Bartholdy, Laura Barreyro, Britta Will, Michael Roth et al. “H2.0-like homeobox regulates early hematopoiesis and promotes acute myeloid leukemia.” In: *Cancer Cell* 22.2 (Aug. 2012), pp. 194–208. DOI: 10.1016/j.ccr.2012.06.027.
- [271] T J Lints, L Hartley, L M Parsons and R P Harvey. “Mesoderm-specific expression of the divergent homeobox gene Hlx during murine embryogenesis.” In: *Dev Dyn* 205.4 (Apr. 1996), pp. 457–70. DOI: 10.1002/(SICI)1097-0177(199604)205:4<457::AID-AJA9>3.0.CO;2-H.
- [272] Michael D Bates, Dana T Dunagan, Lynn C Welch, Ajay Kaul and Richard P Harvey. “The Hlx homeobox transcription factor is required early in enteric nervous system development.” In: *BMC Dev Biol* 6 (July 2006), p. 33. DOI: 10.1186/1471-213X-6-33.
- [273] Erin N Star, Minyan Zhu, Zhiwei Shi, Haiquan Liu, Mohammad Pashmforoush, Yves Sauve et al. “Regulation of retinal interneuron subtype identity by the Iroquois homeobox gene Irx6.” In: *Development* 139.24 (Dec. 2012), pp. 4644–55. DOI: 10.1242/dev.081729.
- [274] Krishna K Ghosh, Sascha Bujan, Silke Haverkamp, Andreas Feigenspan and Heinz Wässle. “Types of bipolar cells in the mouse retina.” In: *J Comp Neurol* 469.1 (Jan. 2004), pp. 70–82. DOI: 10.1002/cne.10985.
- [275] Alyssa Kallman, Elizabeth E Capowski, Jie Wang, Aniruddha M Kaushik, Alex D Jansen, Kimberly L Edwards et al. “Investigating cone photoreceptor development using patient-derived NRL null retinal organoids.” In: *Commun Biol* 3.1 (Feb. 2020), p. 82. DOI: 10.1038/s42003-020-0808-5.
- [276] Laura Croci, Seung-Hyuk Chung, Giacomo Masserdotti, Sara Gianola, Antonella Bizzoca, Gianfranco Gennarini et al. “A key role for the HLH transcription factor EBF2COE2, O/E-3 in Purkinje neuron migration and cerebellar cortical topography.” In: *Development* 133.14 (July 2006), pp. 2719–29. DOI: 10.1242/dev.02437.
- [277] Volker Hovestadt, David T W Jones, Simone Picelli, Wei Wang, Marcel Kool, Paul A Northcott et al. “Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing.” In: *Nature* 510.7506 (June 2014), pp. 537–41. DOI: 10.1038/nature13268.
- [278] Zhixiang Zhao, Mohammad Aminur Rahman, Zhuo G Chen and Dong M Shin. “Multiple biological functions of Twist1 in various cancers.” In: *Oncotarget* 8.12 (Mar. 2017), pp. 20380–20393. DOI: 10.18632/oncotarget.14608.
- [279] Paul Gibson, Yiai Tong, Giles Robinson, Margaret C Thompson, D Spencer Currie, Christopher Eden et al. “Subtypes of medulloblastoma have distinct developmental origins.” In: *Nature* 468.7327 (Dec. 2010), pp. 1095–9. DOI: 10.1038/nature09587.
- [280] Suzana A Kahn, Xin Wang, Ryan T Nitta, Sharareh Gholamin, Johanna Theruvath, Gregor Hutter et al. “Notch1 regulates the initiation of metastasis and self-renewal of Group 3 medulloblastoma.” In: *Nat Commun* 9.1 (Oct. 2018), p. 4121. DOI: 10.1038/s41467-018-06564-9.
- [281] Muh-Hwa Yang, Dennis Shin-Shian Hsu, Hsei-Wei Wang, Hsiao-Jung Wang, Hsin-Yi Lan, Wen-Hao Yang et al. “Bmi1 is essential in Twist1-induced epithelial-mesenchymal transition.” In: *Nat Cell Biol* 12.10 (Oct. 2010), pp. 982–92. DOI: 10.1038/ncb2099.
- [282] M Rex, R Church, K Tointon, R M Ichihashi, S Mokhtar, D Uwanogho et al. “Granule cell development in the cerebellum is punctuated by changes in Sox gene expression.” In: *Brain Res Mol Brain Res* 55.1 (Mar. 1998), pp. 28–34. DOI: 10.1016/S0169-328X(97)00354-9.

Bibliography

- [283] Siu Man Tsang, Erik Oliemuller and Beatrice A Howard. "Regulatory roles for SOX11 in development, stem cells and cancer." In: *Semin Cancer Biol* 67.Pt 1 (Dec. 2020), pp. 3–11. DOI: 10.1016/j.semcancer.2020.06.015.
- [284] Yong Wang, Lu Lin, Helen Lai, Luis F Parada and Lei Lei. "Transcription factor Sox11 is essential for both embryonic and adult neurogenesis." In: *Dev Dyn* 242.6 (June 2013), pp. 638–53. DOI: 10.1002/dvdy.23962.
- [285] Jakub Neradil and Renata Veselska. "Nestin as a marker of cancer stem cells." In: *Cancer Sci* 106.7 (July 2015), pp. 803–11. DOI: 10.1111/cas.12691.
- [286] Aurora Bernal and Lorena Arranz. "Nestin-expressing progenitor cells: function, identity and therapeutic implications." In: *Cell Mol Life Sci* 75.12 (June 2018), pp. 2177–2195. DOI: 10.1007/s00018-018-2794-z.
- [287] Alice Jo, Sahitya Denduluri, Bosi Zhang, Zhongliang Wang, Liangjun Yin, Zhengjian Yan et al. "The versatile functions of Sox9 in development, stem cells, and human diseases." In: *Genes Dis* 1.2 (Dec. 2014), pp. 149–161. DOI: 10.1016/j.gendis.2014.09.004.
- [288] Maribel Aguilar-Medina, Mariana Avendaño-Félix, Erik Lizárraga-Verdugo, Mercedes Bermúdez, José Geovanni Romero-Quintana, Rosalío Ramos-Payan et al. "SOX9 Stem-Cell Factor: Clinical and Functional Relevance in Cancer." In: *J Oncol* 2019 (2019), p. 6754040. DOI: 10.1155/2019/6754040.
- [289] Aldwin Suryo Rahmanto, Vasil Savov, Andrä Brunner, Sara Bolin, Holger Weishaupt, Alena Malyukova et al. "FBW7 suppression leads to SOX9 stabilization and increased malignancy in medulloblastoma." In: *EMBO J* 35.20 (Oct. 2016), pp. 2192–2212. DOI: 10.15252/embj.201693889.
- [290] Fredrik J Swartling, Vasil Savov, Anders I Persson, Justin Chen, Christopher S Hackett, Paul A Northcott et al. "Distinct neural stem cell populations give rise to disparate brain tumors in response to N-MYC." In: *Cancer Cell* 21.5 (May 2012), pp. 601–613. DOI: 10.1016/j.ccr.2012.04.012.
- [291] Lu Cui, Yuan Guan, Zepeng Qu, Jingfa Zhang, Bing Liao, Bo Ma et al. "WNT signaling determines tumorigenicity and function of ESC-derived retinal progenitors." In: *J Clin Invest* 123.4 (Apr. 2013), pp. 1647–61. DOI: 10.1172/JCI65048.
- [292] Xiao-Jing Xue and Xiao-Bing Yuan. "Nestin is essential for mitogen-stimulated proliferation of neural progenitor cells." In: *Mol Cell Neurosci* 45.1 (Sept. 2010), pp. 26–36. DOI: 10.1016/j.mcn.2010.05.006.
- [293] Jason S L Yu and Wei Cui. "Proliferation, survival and metabolism: the role of PI3K/AKT/mTOR signalling in pluripotency and cell fate determination." In: *Development* 143.17 (Sept. 2016), pp. 3050–60. DOI: 10.1242/dev.137075.
- [294] Wenting Xu, Zhen Yang and Nonghua Lu. "A new role for the PI3K/Akt signaling pathway in the epithelial-mesenchymal transition." In: *Cell Adh Migr* 9.4 (2015), pp. 317–24. DOI: 10.1080/19336918.2015.1016686.
- [295] H Li, D P Witte, W W Branford, B J Aronow, M Weinstein, S Kaur et al. "Gsh-4 encodes a LIM-type homeodomain, is expressed in the developing central nervous system and is required for early postnatal survival." In: *EMBO J* 13.12 (June 1994), pp. 2876–85.
- [296] Revathi Balasubramanian, Andrew Bui, Qian Ding and Lin Gan. "Expression of LIM-homeodomain transcription factors in the developing and mature mouse retina." In: *Gene Expr Patterns* 14.1 (Jan. 2014), pp. 1–8. DOI: 10.1016/j.gexp.2013.12.001.
- [297] Diego F Buenaventura, Adrienne Corseri and Mark M Emerson. "Identification of Genes With Enriched Expression in Early Developing Mouse Cone Photoreceptors." In: *Invest Ophthalmol Vis Sci* 60.8 (July 2019), pp. 2787–2799. DOI: 10.1167/iovs.19-26951.

- [298] Ken C Lo, Michael R Rossi, Tania Burkhardt, Scott L Pomeroy and John K Cowell. “Overlay analysis of the oligonucleotide array gene expression profiles and copy number abnormalities as determined by array comparative genomic hybridization in medulloblastomas.” In: *Genes Chromosomes Cancer* 46.1 (Jan. 2007), pp. 53–66. DOI: 10.1002/gcc.20388.
- [299] Jeremy a Miller, Song-Lin Ding, Susan M Sunkin, Kimberly A Smith, Lydia Ng, Aaron Szafer et al. “Transcriptional landscape of the prenatal human brain.” In: *Nature* 508.7495 (2014), pp. 199–206. DOI: 10.1038/nature13185.
- [300] T Ohtsuka, M Sakamoto, F Guillemot and R Kageyama. “Roles of the basic helix-loop-helix genes Hes1 and Hes5 in expansion of neural stem cells of the developing brain.” In: *J Biol Chem* 276.32 (Aug. 2001), pp. 30467–74. DOI: 10.1074/jbc.M102420200.
- [301] Carolina A Oliva, Jessica Y Vargas and Nibaldo C Inestrosa. “Wnts in adult brain: from synaptic plasticity to cognitive deficiencies.” In: *Front Cell Neurosci* 7 (Dec. 2013), p. 224. DOI: 10.3389/fncel.2013.00224.
- [302] C Ruppert, D Goldowitz and W Wille. “Proto-oncogene c-myc is expressed in cerebellar neurons at different developmental stages.” In: *EMBO J* 5.8 (Aug. 1986), pp. 1897–901.
- [303] K A Zimmerman, G D Yancopoulos, R G Collum, R K Smith, N E Kohl, K A Denis et al. “Differential expression of myc family genes during murine development.” In: *Nature* 319.6056 (1986), pp. 780–3. DOI: 10.1038/319780a0.
- [304] Yoshihiro Kawasaki, Mimon Komiya, Kosuke Matsumura, Lumi Negishi, Sakiko Suda, Masumi Okuno et al. “MYU, a Target lncRNA for Wnt/c-Myc Signaling, Mediates Induction of CDK6 to Promote Cell Cycle Progression.” In: *Cell Rep* 16.10 (Sept. 2016), pp. 2554–2564. DOI: 10.1016/j.celrep.2016.08.015.
- [305] Ran Gao, Rui Zhang, Cuicui Zhang, Yingwu Liang and Weining Tang. “LncRNA LOXL1-AS1 Promotes the Proliferation and Metastasis of Medulloblastoma by Activating the PI3K/AKT Pathway.” In: *Anal Cell Pathol (Amst)* 2018 (2018), p. 9275685. DOI: 10.1155/2018/9275685.
- [306] Delyan P Ivanov, Beth Coyle, David A Walker and Anna M Grabowska. “In vitro models of medulloblastoma: Choosing the right tool for the job.” In: *J Biotechnol* 236 (Oct. 2016), pp. 10–25. DOI: 10.1016/j.jbiotec.2016.07.028.
- [307] Wei Liu, Qin-Peng Wang and Jia Guo. “Prognostic significance of long non-coding RNA DANCR expression in human cancers: A systematic review and meta-analysis.” In: *Biosci Rep* (Mar. 2019). DOI: 10.1042/BSR20181627.
- [308] Lei Pan, Wei Liang, Jianmei Gu, Xueyan Zang, Zhenhua Huang, Hui Shi et al. “Long noncoding RNA DANCR is activated by SALL4 and promotes the proliferation and invasion of gastric cancer cells.” In: *Oncotarget* 9.2 (Jan. 2018), pp. 1915–1930. DOI: 10.18632/oncotarget.23019.
- [309] Jun Li and Liang Zhou. “Overexpression of lncRNA DANCR positively affects progression of glioma via activating Wnt/ β -catenin signaling.” In: *Biomed Pharmacother* 102 (June 2018), pp. 602–607. DOI: 10.1016/j.biopha.2018.03.116.
- [310] Yunqi Lu, Zhongyi Hu, Lingegowda S Mangala, Zachary E Stine, Xiaowen Hu, Dahai Jiang et al. “MYC Targeted Long Noncoding RNA DANCR Promotes Cancer in Part by Reducing p21 Levels.” In: *Cancer Res* 78.1 (Jan. 2018), pp. 64–74. DOI: 10.1158/0008-5472.CAN-17-0815.
- [311] Babita Madan, Nathan Harmston, Gahyathiri Nallan, Alex Montoya, Peter Faull, Enrico Petretto et al. “Temporal dynamics of Wnt-dependent transcriptome reveal an oncogenic Wnt/MYC/ribosome axis.” In: *J Clin Invest* 128.12 (Dec. 2018), pp. 5620–5633. DOI: 10.1172/JCI122383.
- [312] Yan Pan, Chen Li, Jing Chen, Kai Zhang, Xiaoyuan Chu, Rui Wang et al. “The Emerging Roles of Long Noncoding RNA ROR (lincRNA-ROR) and its Possible Mechanisms in Human Cancers.” In: *Cell Physiol Biochem* 40.1-2 (2016), pp. 219–229. DOI: 10.1159/000452539.

Bibliography

- [313] Yanhui Lou, Huanhuan Jiang, Zhumei Cui, Lingzhi Wang, Xiangyu Wang and Tian Tian. “LincROR induces epithelial-to-mesenchymal transition in ovarian cancer by increasing Wnt/ β -catenin signaling.” In: *Oncotarget* 8.41 (Sept. 2017), pp. 69983–69994. DOI: 10.18632/oncotarget.19545.
- [314] Maria C Vladoiu, Ibrahim El-Hamamy, Laura K Donovan, Hamza Farooq, Borja L Holgado, Yogi Sundaravadanam et al. “Childhood cerebellar tumours mirror conserved fetal transcriptional programs.” In: *Nature* 572.7767 (Aug. 2019), pp. 67–73. DOI: 10.1038/s41586-019-1158-7.
- [315] Shi-Yan Ng, Gireesh K Bogu, Boon Seng Soh and Lawrence W Stanton. “The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis.” In: *Mol Cell* 51.3 (Aug. 2013), pp. 349–59. DOI: 10.1016/j.molcel.2013.07.017.
- [316] Giuliana Caronia-Brown, Angela Andereggi and Rajeshwar Awatramani. “Expression and functional analysis of the Wnt/beta-catenin induced mir-135a-2 locus in embryonic forebrain development.” In: *Neural Dev* 11 (Apr. 2016), p. 9. DOI: 10.1186/s13064-016-0065-y.
- [317] Qi Li, Chengya Dong, Jiayue Cui, Yubo Wang and Xinyu Hong. “Over-expressed lncRNA HOTAIRM1 promotes tumor growth and invasion through up-regulating HOXA1 and sequestering G9a/EZH2/Dnmts away from the HOXA1 gene in glioblastoma multiforme.” In: *J Exp Clin Cancer Res* 37.1 (Oct. 2018), p. 265. DOI: 10.1186/s13046-018-0941-x.
- [318] Xue Q D Wang and Josée Dostie. “Reciprocal regulation of chromatin state and architecture by HOTAIRM1 contributes to temporal collinear HOXA gene activation.” In: *Nucleic Acids Res* 45.3 (Feb. 2017), pp. 1091–1104. DOI: 10.1093/nar/gkw966.
- [319] Na Chen, Dan Guo, Qiong Xu, Minhui Yang, Dan Wang, Man Peng et al. “Long non-coding RNA FEZF1-AS1 facilitates cell proliferation and migration in colorectal carcinoma.” In: *Oncotarget* 7.10 (Mar. 2016), pp. 11271–83. DOI: 10.18632/oncotarget.7168.
- [320] Xianxiang Cao, Jing Xu and Dong Yue. “LncRNA-SNHG16 predicts poor prognosis and promotes tumor proliferation through epigenetically silencing p21 in bladder cancer.” In: *Cancer Gene Ther* 25.1-2 (Feb. 2018), pp. 10–17. DOI: 10.1038/s41417-017-0006-x.
- [321] Wei Han, Xuemei Du, Min Liu, Jing Wang, Lixin Sun and Yongchun Li. “Increased expression of long non-coding RNA SNHG16 correlates with tumor progression and poor prognosis in non-small cell lung cancer.” In: *Int J Biol Macromol* 121 (Jan. 2019), pp. 270–278. DOI: 10.1016/j.ijbiomac.2018.10.004.
- [322] Shangfeng Li, Shengkai Zhang and Jie Chen. “c-Myc induced upregulation of long non-coding RNA SNHG16 enhances progression and carcinogenesis in oral squamous cell carcinoma.” In: *Cancer Gene Ther* 26.11-12 (Nov. 2019), pp. 400–410. DOI: 10.1038/s41417-018-0072-8.
- [323] Jonathan R Hart, Thomas C Roberts, Marc S Weinberg, Kevin V Morris and Peter K Vogt. “MYC regulates the non-coding transcriptome.” In: *Oncotarget* 5.24 (Dec. 2014), pp. 12543–54. DOI: 10.18632/oncotarget.3033.
- [324] Anbang He, Shiming He, Xuesong Li and Liqun Zhou. “ZFAS1: A novel vital oncogenic lncRNA in multiple human cancers.” In: *Cell Prolif* 52.1 (Jan. 2019), e12513. DOI: 10.1111/cpr.12513.
- [325] Shalini Singh, Danielle Howell, Niraj Trivedi, Ketty Kessler, Taren Ong, Pedro Rosmaninho et al. “Zeb1 controls neuron differentiation and germinal zone exit by a mesenchymal-epithelial-like transition.” In: *Elife* 5 (May 2016). DOI: 10.7554/eLife.12717.
- [326] Fengqi Nie, Xiang Yu, Mingde Huang, Yunfei Wang, Min Xie, Hongwei Ma et al. “Long noncoding RNA ZFAS1 promotes gastric cancer cells proliferation by epigenetically repressing KLF2 and NKD2 expression.” In: *Oncotarget* 8.24 (June 2017), pp. 38227–38238. DOI: 10.18632/oncotarget.9611.

- [327] W Zeng, K A Wharton, J A Mack, K Wang, M Gadbaw, K Suyama et al. “naked cuticle encodes an inducible antagonist of Wnt signalling.” In: *Nature* 403.6771 (Feb. 2000), pp. 789–95. DOI: 10.1038/35001615.
- [328] Tianhui Hu and Cunxi Li. “Convergence between Wnt- β -catenin and EGFR signaling in cancer.” In: *Mol Cancer* 9 (Sept. 2010), p. 236. DOI: 10.1186/1476-4598-9-236.
- [329] Chenhui Ma, Xuefei Shi, Qingqing Zhu, Qian Li, Yafang Liu, Yanwen Yao et al. “The growth arrest-specific transcript 5 (GAS5): a pivotal tumor suppressor long noncoding RNA in human cancers.” In: *Tumour Biol* 37.2 (Feb. 2016), pp. 1437–44. DOI: 10.1007/s13277-015-4521-9.
- [330] Xin Chen, Chao Yang, Shengli Xie and Edwin Cheung. “Long non-coding RNA GAS5 and ZFAS1 are prognostic markers involved in translation targeted by miR-940 in prostate cancer.” In: *Oncotarget* 9.1 (Jan. 2018), pp. 1048–1062. DOI: 10.18632/oncotarget.23254.
- [331] Joseph Mazar, Amy Rosado, John Shelley, John Marchica and Tamarah J Westmoreland. “The long non-coding RNA GAS5 differentially regulates cell cycle arrest and apoptosis through activation of BRCA1 and p53 in human neuroblastoma.” In: *Oncotarget* 8.4 (Jan. 2017), pp. 6589–6607. DOI: 10.18632/oncotarget.14244.
- [332] Yingyi Zhang, Xinya Su, Zhe Kong, Fangqiu Fu, Pu Zhang, Dan Wang et al. “An androgen reduced transcript of LncRNA GAS5 promoted prostate cancer proliferation.” In: *PLoS One* 12.8 (2017), e0182305. DOI: 10.1371/journal.pone.0182305.
- [333] Yongchao Liu, Jing Zhao, Wenhong Zhang, Jun Gan, Chengen Hu, Guangjian Huang et al. “LncRNA GAS5 enhances G1 cell cycle arrest via binding to YBX1 to regulate p21 expression in stomach cancer.” In: *Scientific reports* 5.March (2015), p. 10159. ISSN: 2045-2322. DOI: 10.1038/srep10159.
- [334] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov and Pablo Tamayo. “The Molecular Signatures Database (MSigDB) hallmark gene set collection.” In: *Cell Syst* 1.6 (Dec. 2015), pp. 417–425. DOI: 10.1016/j.cels.2015.12.004.
- [335] Jan van Riggelen, Alper Yetil and Dean W Felsher. “MYC as a regulator of ribosome biogenesis and protein synthesis.” In: *Nat Rev Cancer* 10.4 (Apr. 2010), pp. 301–9. DOI: 10.1038/nrc2819.
- [336] Elizabeth A Raetz, Marianne K H Kim, Philip Moos, Marlee Carlson, Carol Bruggers, David K Hooper et al. “Identification of genes that are regulated transcriptionally by Myc in childhood tumors.” In: *Cancer* 98.4 (Aug. 2003), pp. 841–53. DOI: 10.1002/cncr.11584.
- [337] Danielle Ribeiro Lucon, Cristiane de Souza Rocha, Rogerio Bastos Craveiro, Dagmar Dilloo, Izilda A Cardinalli, Denise Pontes Cavalcanti et al. “Downregulation of 14q32 microRNAs in Primary Human Desmoplastic Medulloblastoma.” In: *Front Oncol* 3 (2013), p. 254. DOI: 10.3389/fonc.2013.00254.
- [338] Syuzo Kaneko, Roberto Bonasio, Ricardo Saldaña-Meyer, Takahaki Yoshida, Jinsook Son, Koichiro Nishino et al. “Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin.” In: *Mol Cell* 53.2 (Jan. 2014), pp. 290–300. DOI: 10.1016/j.molcel.2013.11.012.
- [339] Chao-Chung Kuo, Sonja Hänzelmann, Nevcin Sentürk Cetin, Stefan Frank, Barna Zajzon, Jens-Peter Derks et al. “Detection of RNA-DNA binding sites in long noncoding RNAs.” In: *Nucleic Acids Res* 47.6 (Apr. 2019), e32. DOI: 10.1093/nar/gkz037.
- [340] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi and Mao Tanabe. “Data, information, knowledge and principle: back to metabolism in KEGG.” In: *Nucleic Acids Res* 42.Database issue (Jan. 2014), pp. D199–205. DOI: 10.1093/nar/gkt1076.
- [341] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati et al. “The Reactome Pathway Knowledgebase.” In: *Nucleic Acids Res* 46.D1 (Jan. 2018), pp. D649–D655. DOI: 10.1093/nar/gkx1132.

Bibliography

- [342] Martina Kutmon, Anders Riutta, Nuno Nunes, Kristina Hanspers, Egon L Willighagen, Anwesha Bohler et al. “WikiPathways: capturing the full diversity of pathway knowledge.” In: *Nucleic Acids Res* 44.D1 (Jan. 2016), pp. D488–94. DOI: 10.1093/nar/gkv1024.
- [343] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay et al. “PID: the Pathway Interaction Database.” In: *Nucleic Acids Res* 37.Database issue (Jan. 2009), pp. D674–9. DOI: 10.1093/nar/gkn653.
- [344] Ralf Herwig, Christopher Hardt, Matthias Lienhard and Atanas Kamburov. “Analyzing and interpreting genome data at the network level with ConsensusPathDB.” In: *Nat Protoc* 11.10 (Oct. 2016), pp. 1889–907. DOI: 10.1038/nprot.2016.117.
- [345] Ulrich Schüller, Qing Zhao, Susana A Godinho, Vivi M Heine, René H Medema, David Pellman et al. “Forkhead transcription factor FoxM1 regulates mitotic entry and prevents spindle defects in cerebellar granule neuron precursors.” In: *Mol Cell Biol* 27.23 (Dec. 2007), pp. 8259–70. DOI: 10.1128/MCB.00707-07.
- [346] Xi Chen, Gerd A Müller, Marianne Quaas, Martin Fischer, Namshik Han, Benjamin Stutchbury et al. “The forkhead transcription factor FOXM1 controls cell cycle-dependent gene expression through an atypical chromatin binding mechanism.” In: *Mol Cell Biol* 33.2 (Jan. 2013), pp. 227–36. DOI: 10.1128/MCB.00881-12.
- [347] Qianyun Mei, Junhua Huang, Wanping Chen, Jie Tang, Chen Xu, Qi Yu et al. “Regulation of DNA replication-coupled histone gene expression.” In: *Oncotarget* 8.55 (Nov. 2017), pp. 95005–95022. DOI: 10.18632/oncotarget.21887.
- [348] Marcos Malumbres and Mariano Barbacid. “Cell cycle, CDKs and cancer: a changing paradigm.” In: *Nat Rev Cancer* 9.3 (Mar. 2009), pp. 153–66. DOI: 10.1038/nrc2602.
- [349] Olivier Gavet and Jonathon Pines. “Progressive activation of CyclinB1-Cdk1 coordinates entry to mitosis.” In: *Dev Cell* 18.4 (Apr. 2010), pp. 533–43. DOI: 10.1016/j.devcel.2010.02.013.
- [350] Joan Massagué. “TGF β in Cancer”. In: *Cell* 134.2 (2008), pp. 215–230. ISSN: 00928674. DOI: 10.1016/j.cell.2008.07.001.
- [351] Natalia Pervjakova, Silva Kasela, Andrew P. Morris, Mart Kals, Andres Metspalu, Cecilia M. Lindgren et al. “Imprinted genes and imprinting control regions show predominant intermediate methylation in adult somatic tissues.” In: *Epigenomics* 8.6 (2016), pp. 789–99. DOI: 10.2217/epi.16.8.
- [352] Axel Visel, Christina Thaller and Gregor Eichele. “GenePaint.org: an atlas of gene expression patterns in the mouse embryo.” In: *Nucleic Acids Res*. 32.Database issue (2004), pp. D552–6. DOI: 10.1093/nar/gkh029.
- [353] R Artacho, C Lujano, A B Sanchez-Vico, C Vargas Sánchez, J González Calvo, P R Bouzas et al. “Nutritional status in chronically-ill elderly patients. Is it related to quality of life?” In: *J Nutr Health Aging* 18.2 (2014), pp. 192–7. DOI: 10.1007/s12603-013-0385-0.
- [354] Fiona Cunningham, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent et al. “Ensembl 2015.” In: *Nucleic Acids Res* 43.Database issue (Jan. 2015), pp. D662–9. DOI: 10.1093/nar/gku1010.
- [355] Daniel Zilberman and Steven Henikoff. “Genome-wide analysis of DNA methylation patterns.” In: *Development* 134.22 (Nov. 2007), pp. 3959–65. DOI: 10.1242/dev.001131.
- [356] C S Wilhelm-Benartzi, D C Koestler, M R Karagas, J M Flanagan, B C Christensen, K T Kelsey et al. “Review of processing and analysis methods for DNA methylation array data.” In: *Br J Cancer* 109.6 (Sept. 2013), pp. 1394–402. DOI: 10.1038/bjc.2013.496.

- [357] Volker Hovestadt, Marc Remke, Marcel Kool, Torsten Pietsch, Paul A Northcott, Roger Fischer et al. “Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays.” In: *Acta Neuropathol* 125.6 (June 2013), pp. 913–6. DOI: 10.1007/s00401-013-1126-5.
- [358] Jean-Philippe Fortin, Timothy J Triche and Kasper D Hansen. “Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi.” In: *Bioinformatics* 33.4 (Feb. 2017), pp. 558–560. DOI: 10.1093/bioinformatics/btw691.
- [359] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe and John D Storey. “The sva package for removing batch effects and other unwanted variation in high-throughput experiments.” In: *Bioinformatics* 28.6 (Mar. 2012), pp. 882–3. DOI: 10.1093/bioinformatics/bts034.
- [360] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoutte, David S Johnson, Bradley E Bernstein et al. “Model-based analysis of CHIP-Seq (MACS).” In: *Genome Biol* 9.9 (2008), R137. DOI: 10.1186/gb-2008-9-9-r137.
- [361] Marina Lizio, Jayson Harshbarger, Hisashi Shimoji, Jessica Severin, Takeya Kasukawa, Serkan Sahin et al. “Gateways to the FANTOM5 promoter level mammalian expression atlas.” In: *Genome Biol* 16 (Jan. 2015), p. 22. DOI: 10.1186/s13059-014-0560-6.
- [362] Peter J. Rousseeuw and Christophe Croux. “Alternatives to the Median Absolute Deviation”. In: *Journal of the American Statistical Association* 88.424 (1993), pp. 1273–1283. ISSN: 0162-1459. DOI: 10.1080/01621459.1993.10476408.
- [363] K Cartharius, K Frech, K Grote, B Klocke, M Haltmeier, A Klingenhoff et al. “MatInspector and beyond: promoter analysis based on transcription factor binding sites.” In: *Bioinformatics* 21.13 (July 2005), pp. 2933–42. DOI: 10.1093/bioinformatics/bti473.
- [364] Shannan J Ho Sui, James R Mortimer, David J Arenillas, Jochen Brumm, Christopher J Walsh, Brian P Kennedy et al. “oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.” In: *Nucleic Acids Res* 33.10 (2005), pp. 3154–64. DOI: 10.1093/nar/gki624.
- [365] Martin Rosvall, Alcides V Esquivel, Andrea Lancichinetti, Jevin D West and Renaud Lambiotte. “Memory in network flows and its effects on spreading dynamics and community detection.” In: *Nat Commun* 5 (Aug. 2014), p. 4630. DOI: 10.1038/ncomms5630.
- [366] Yanli Wang, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang et al. “The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions.” In: *Genome Biol* 19.1 (Oct. 2018), p. 151. DOI: 10.1186/s13059-018-1519-9.
- [367] Donna Maglott, Jim Ostell, Kim D Pruitt and Tatiana Tatusova. “Entrez Gene: gene-centered information at NCBI.” In: *Nucleic Acids Res* 39.Database issue (Jan. 2011), pp. D52–7. DOI: 10.1093/nar/gkq1237.
- [368] Rhoda J Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich et al. “Ensembl BioMart: a hub for data retrieval across taxonomic space.” In: *Database (Oxford)* 2011 (2011), bar030. DOI: 10.1093/database/bar030.
- [369] Roded Sharan, Adi Maron-Katz and Ron Shamir. “CLICK and EXPANDER: a system for clustering and visualizing gene expression data.” In: *Bioinformatics* 19.14 (Sept. 2003), pp. 1787–99. DOI: 10.1093/bioinformatics/btg232.
- [370] Hongda Chen, Phillip Knebel and Hermann Brenner. “Empirical evaluation demonstrated importance of validating biomarkers for early detection of cancer in screening settings to limit the number of false-positive findings.” In: *J Clin Epidemiol* 75 (July 2016), pp. 108–14. DOI: 10.1016/j.jclinepi.2016.01.022.

Bibliography

- [371] Markus S Schröder, Aedin C Culhane, John Quackenbush and Benjamin Haibe-Kains. “survcomp: an R/Bioconductor package for performance assessment and comparison of survival models.” In: *Bioinformatics* 27.22 (Nov. 2011), pp. 3206–8. DOI: 10.1093/bioinformatics/btr511.
- [372] F Wolfertstetter, K Frech, G Herrmann and T Werner. “Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm.” In: *Comput Appl Biosci* 12.1 (Feb. 1996), pp. 71–80. DOI: 10.1093/bioinformatics/12.1.71.
- [373] Leann Myers and Maria J. Sirois. “Spearman Correlation Coefficients, Differences between”. In: *Encycl. Stat. Sci.* Hoboken, NJ, USA: John Wiley & Sons, Inc., 2004. DOI: 10.1002/0471667196.ess5050.
- [374] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen et al. “Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.” In: *Bioinformatics* 30.10 (May 2014), pp. 1363–9. DOI: 10.1093/bioinformatics/btu049.
- [375] Jean-Philippe Fortin, Aurélie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig et al. “Functional normalization of 450k methylation array data improves replication in large cancer studies.” In: *Genome Biol* 15.12 (Dec. 2014), p. 503. DOI: 10.1186/s13059-014-0503-2.
- [376] David Monk, Joannella Morales, Johan T den Dunnen, Silvia Russo, Franck Court, Dirk Prawitt et al. “Recommendations for a nomenclature system for reporting methylation aberrations in imprinted domains.” In: *Epigenetics* 13.2 (2018), pp. 117–121. DOI: 10.1080/15592294.2016.1264561.
- [377] Yasin Şenbabaoğlu, George Michailidis and Jun Z Li. “Critical limitations of consensus clustering in class discovery.” In: *Sci Rep* 4 (Aug. 2014), p. 6207. DOI: 10.1038/srep06207.
- [378] Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona and Jean-Philippe Vert. “TIGRESS: Trustful Inference of Gene REgulation using Stability Selection.” In: *BMC Syst Biol* 6 (Nov. 2012), p. 145. DOI: 10.1186/1752-0509-6-145.
- [379] Vân Anh Huynh-Thu and Pierre Geurts. “Unsupervised Gene Network Inference with Decision Trees and Random Forests”. In: *Gene Network Inference*. Vol. 1883. Berlin, Heidelberg: Springer Berlin Heidelberg, 2019, pp. 195–215. ISBN: 9781493988822. DOI: 10.1007/978-1-4939-8882-2_8.
- [380] Frauke Degenhardt, Stephan Seifert and Silke Szymczak. “Evaluation of variable selection methods for random forests and omics data sets”. In: *Briefings in Bioinformatics* 20.2 (2019), pp. 492–503. ISSN: 14774054. DOI: 10.1093/bib/bbx124.
- [381] Vân Anh Huynh-Thu and Guido Sanguinetti. “Gene Regulatory Network Inference: An Introductory Survey”. In: *Methods in Molecular Biology*. Vol. 1883. 2019, pp. 1–23. ISBN: 9781493988822. DOI: 10.1007/978-1-4939-8882-2_1. eprint: 1801.04087.
- [382] Spencer Angus Thomas and Yaochu Jin. “Reconstructing biological gene regulatory networks: Where optimization meets big data”. In: *Evolutionary Intelligence* 7.1 (2014), pp. 29–47. ISSN: 18645917. DOI: 10.1007/s12065-013-0098-7.
- [383] Philipp Probst, Marvin N. Wright and Anne Laure Boulesteix. “Hyperparameters and tuning strategies for random forest”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.3 (2019), pp. 1–15. ISSN: 19424795. DOI: 10.1002/widm.1301. arXiv: 1804.03515.
- [384] Alexander Pieper, Stephanie Rudolph, Georg L Wieser, Tilmann Götze, Hendrik Mießner, Tomoko Yonemasu et al. “NeuroD2 controls inhibitory circuit formation in the molecular layer of the cerebellum.” In: *Sci Rep* 9.1 (Feb. 2019), p. 1448. DOI: 10.1038/s41598-018-37850-7.

- [385] C Xiang, V Baubet, S Pal, L Holderbaum, V Tatard, P Jiang et al. “RP58/ZNF238 directly modulates proneurogenic gene levels and is required for neuronal differentiation and brain expansion.” In: *Cell Death Differ* 19.4 (Apr. 2012), pp. 692–702. DOI: 10.1038/cdd.2011.144.
- [386] Valérie M Tatard, Chaomei Xiang, Jaclyn A Biegel and Nadia Dahmane. “ZNF238 is expressed in postmitotic brain cells and inhibits brain tumor growth.” In: *Cancer Res* 70.3 (Feb. 2010), pp. 1236–46. DOI: 10.1158/0008-5472.CAN-09-2249.
- [387] Paul A Northcott, Andrey Korshunov, Stefan M Pfister and Michael D Taylor. “The clinical implications of medulloblastoma subgroups.” In: *Nat Rev Neurol* 8.6 (May 2012), pp. 340–51. DOI: 10.1038/nrneuro1.2012.78.
- [388] Volker Hovestadt, Kyle S Smith, Laure Bihannic, Mariella G Filbin, McKenzie L Shaw, Alicia Baumgartner et al. “Resolving medulloblastoma cellular architecture by single-cell genomics.” In: *Nature* 572.7767 (Aug. 2019), pp. 74–79. DOI: 10.1038/s41586-019-1434-6.
- [389] David M Gonzalez and Damian Medici. “Signaling mechanisms of the epithelial-mesenchymal transition.” In: *Sci Signal* 7.344 (Sept. 2014), re8. DOI: 10.1126/scisignal.2005189.
- [390] Isabel Fabregat, Andrea Malfettone and Jitka Soukupova. “New Insights into the Crossroads between EMT and Stemness in the Context of Cancer.” In: *J Clin Med* 5.3 (Mar. 2016). DOI: 10.3390/jcm5030037.
- [391] Pedro M Aponte and Andrés Caicedo. “Stemness in Cancer: Stem Cells, Cancer Stem Cells, and Their Microenvironment.” In: *Stem Cells Int* 2017 (2017), p. 5619472. DOI: 10.1155/2017/5619472.
- [392] Yan Cheng, Shengyou Liao, Gang Xu, Jian Hu, Duancheng Guo, Fang Du et al. “NeuroD1 Dictates Tumor Cell Differentiation in Medulloblastoma.” In: *Cell Rep* 31.12 (June 2020), p. 107782. DOI: 10.1016/j.celrep.2020.107782.
- [393] Yong Teng, Liwei Lang and Catherine E Jauregui. “The Complexity of DEK Signaling in Cancer Progression.” In: *Curr Cancer Drug Targets* 18.3 (2018), pp. 256–265. DOI: 10.2174/1568009617666170522094730.
- [394] Douglas Vernimmen and Wendy A Bickmore. “The Hierarchy of Transcriptional Activation: From Enhancer to Promoter.” In: *Trends Genet* 31.12 (Dec. 2015), pp. 696–708. DOI: 10.1016/j.tig.2015.10.004.
- [395] Ken J Kron, Swneke D Bailey and Mathieu Lupien. “Enhancer alterations in cancer: a source for a cell identity crisis.” In: *Genome Med* 6.9 (2014), p. 77. DOI: 10.1186/s13073-014-0077-3.
- [396] Giovanna Ambrosini, Ilya Vorontsov, Dmitry Penzar, Romain Groux, Oriol Fornes, Daria D Nikolaeva et al. “Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study.” In: *Genome Biol* 21.1 (May 2020), p. 114. DOI: 10.1186/s13059-020-01996-3.
- [397] Sophie Lamarre, Pierre Frasse, Mohamed Zouine, Delphine Labourdette, Elise Sainderichin, Guojian Hu et al. “Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size.” In: *Front Plant Sci* 9 (2018), p. 108. DOI: 10.3389/fpls.2018.00108.
- [398] Xiaobei Zhou, Helen Lindsay and Mark D Robinson. “Robustly detecting differential expression in RNA sequencing data using observation weights.” In: *Nucleic Acids Res* 42.11 (June 2014), e91. DOI: 10.1093/nar/gku310.
- [399] Lidia Chellini, Valentina Frezza and Maria Paola Paronetto. “Dissecting the transcriptional regulatory networks of promoter-associated noncoding RNAs in development and cancer.” In: *J Exp Clin Cancer Res* 39.1 (Mar. 2020), p. 51. DOI: 10.1186/s13046-020-01552-8.

Bibliography

- [400] Yuan Tan, Yuejin Li and Faqing Tang. “Oncogenic seRNA functional activation: a novel mechanism of tumorigenesis.” In: *Mol Cancer* 19.1 (Apr. 2020), p. 74. DOI: 10.1186/s12943-020-01195-5.
- [401] Alessandro Fatica and Irene Bozzoni. “Long non-coding RNAs: new players in cell differentiation and development.” In: *Nat Rev Genet* 15.1 (Jan. 2014), pp. 7–21. DOI: 10.1038/nrg3606.
- [402] Herah Hansji, Euphemia Y Leung, Bruce C Baguley, Graeme J Finlay, David Cameron-Smith, Vandre C Figueiredo et al. “ZFAS1: a long noncoding RNA associated with ribosomes in breast cancer cells.” In: *Biol Direct* 11.1 (Nov. 2016), p. 62. DOI: 10.1186/s13062-016-0165-y.
- [403] Tomasz Kolenda, Kacper Guglas, Magda Kopczyńska, Anna Teresiak, Renata Bliźniak, Andrzej Mackiewicz et al. “Oncogenic Role of ZFAS1 lncRNA in Head and Neck Squamous Cell Carcinomas.” In: *Cells* 8.4 (Apr. 2019). DOI: 10.3390/cells8040366.
- [404] Jiajie Tu, Geng Tian, Hoi-Hung Cheung, Wei Wei and Tin-Lap Lee. “Gas5 is an essential lncRNA regulator for self-renewal and pluripotency of mouse embryonic stem cells and induced pluripotent stem cells.” In: *Stem Cell Res Ther* 9.1 (Mar. 2018), p. 71. DOI: 10.1186/s13287-018-0813-5.
- [405] Ezgi Hacısuleyman, Loyal A Goff, Cole Trapnell, Adam Williams, Jorge Henao-Mejia, Lei Sun et al. “Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre.” In: *Nat Struct Mol Biol* 21.2 (Feb. 2014), pp. 198–206. DOI: 10.1038/nsmb.2764.
- [406] Sven Heinz, Casey E Romanoski, Christopher Benner and Christopher K Glass. “The selection and function of cell type-specific enhancers.” In: *Nat Rev Mol Cell Biol* 16.3 (Mar. 2015), pp. 144–54. DOI: 10.1038/nrm3949.
- [407] Andrew D Yates, Premanand Achuthan, Wasii Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta et al. “Ensembl 2020.” In: *Nucleic Acids Res* 48.D1 (Jan. 2020), pp. D682–D688. DOI: 10.1093/nar/gkz966.
- [408] Liang Cheng, Pingping Wang, Rui Tian, Song Wang, Qinghua Guo, Meng Luo et al. “LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse.” In: *Nucleic Acids Res* 47.D1 (Jan. 2019), pp. D140–D144. DOI: 10.1093/nar/gky1051.
- [409] Onayemi Titilayo Onagoruwa, Gargi Pal, Chika Ochu and Olorunseun O Ogunwobi. “Oncogenic Role of PVT1 and Therapeutic Implications.” In: *Front Oncol* 10 (2020), p. 17. DOI: 10.3389/fonc.2020.00017.
- [410] Lise Lotte Christensen, Kirsten True, Mark P Hamilton, Morten M Nielsen, Nkerorema D Damas, Christian K Damgaard et al. “SNHG16 is regulated by the Wnt pathway in colorectal cancer and affects genes involved in lipid metabolism.” In: *Mol Oncol* 10.8 (Oct. 2016), pp. 1266–82. DOI: 10.1016/j.molonc.2016.06.003.
- [411] Davide Ruggero. “Translational control in cancer etiology.” In: *Cold Spring Harb Perspect Biol* 5.2 (Feb. 2013). DOI: 10.1101/cshperspect.a012336.
- [412] Hua-Sheng Chiu, Sonal Somvanshi, Ektaben Patel, Ting-Wen Chen, Vivek P Singh, Barry Zorman et al. “Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context.” In: *Cell Rep* 23.1 (Apr. 2018), 297–312.e12. DOI: 10.1016/j.cellrep.2018.03.064.
- [413] Mona Kulkarni, Antonio Hernandez Conte, Aaron Huang, Lorraine Lubin, Takahiro Shiota and Saibal Kar. “Coronary artery disease, acute myocardial infarction, and a newly developing ventricular septal defect: surgical repair or percutaneous closure?” In: *J Cardiothorac Vasc Anesth* 25.6 (Dec. 2011), pp. 1213–6. DOI: 10.1053/j.jvca.2011.08.006.
- [414] Princy Parsana, Claire Ruberman, Andrew E Jaffe, Michael C Schatz, Alexis Battle and Jeffrey T Leek. “Addressing confounding artifacts in reconstruction of gene co-expression networks.” In: *Genome Biol* 20.1 (May 2019), p. 94. DOI: 10.1186/s13059-019-1700-9.

- [415] Nicholas A Furlotte, Hyun Min Kang, Chun Ye and Eleazar Eskin. “Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity.” In: *Bioinformatics* 27.13 (July 2011), pp. i288–94. DOI: 10.1093/bioinformatics/btr221.
- [416] Lin Sun, Yu Li and Bangxiang Yang. “Downregulated long non-coding RNA MEG3 in breast cancer regulates proliferation, migration and invasion by depending on p53’s transcriptional activity.” In: *Biochem Biophys Res Commun* 478.1 (Sept. 2016), pp. 323–329. DOI: 10.1016/j.bbrc.2016.05.031.
- [417] Meital Gabay, Yulin Li and Dean W Felsher. “MYC activation is a hallmark of cancer initiation and maintenance.” In: *Cold Spring Harb Perspect Med* 4.6 (June 2014). DOI: 10.1101/cshperspect.a014241.
- [418] Chao Huang, Xin Liao, Honglei Jin, Fei Xie, Fuxing Zheng, Jingxia Li et al. “MEG3, as a Competing Endogenous RNA, Binds with miR-27a to Promote PHLPP2 Protein Translation and Impairs Bladder Cancer Invasion.” In: *Mol Ther Nucleic Acids* 16 (June 2019), pp. 51–62. DOI: 10.1016/j.omtn.2019.01.014.
- [419] Donya Aref, Connor J Moffatt, Sameer Agnihotri, Vijay Ramaswamy, Adrian M Dubuc, Paul A Northcott et al. “Canonical TGF- β pathway activity is a predictor of SHH-driven medulloblastoma survival and delineates putative precursors in cerebellar development.” In: *Brain Pathol* 23.2 (Mar. 2013), pp. 178–91. DOI: 10.1111/j.1750-3639.2012.00631.x.
- [420] Karthiga Santhana Kumar, Anuja Neve, Ana S Guerreiro Stucklin, Claudia M Kuzan-Fischer, Elisabeth J Rushing, Michael D Taylor et al. “TGF- β Determines the Pro-migratory Potential of bFGF Signaling in Medulloblastoma.” In: *Cell Rep* 23.13 (June 2018), 3798–3812.e8. DOI: 10.1016/j.celrep.2018.05.083.
- [421] Hiroaki Ikushima and Kohei Miyazono. “TGFbeta signalling: a complex web in cancer progression.” In: *Nat Rev Cancer* 10.6 (June 2010), pp. 415–24. DOI: 10.1038/nrc2853.
- [422] Pietro Laneve, Jessica Rea and Elisa Caffarelli. “Long Noncoding RNAs: Emerging Players in Medulloblastoma.” In: *Front Pediatr* 7 (2019), p. 67. DOI: 10.3389/fped.2019.00067.
- [423] Xie Zhengyuan, Xiao Hu, Wang Qiang, Li Nanxiang, Cai Junbin and Zhang Wangming. “Silencing of Urothelial Carcinoma Associated 1 Inhibits the Proliferation and Migration of Medulloblastoma Cells.” In: *Med Sci Monit* 23 (Sept. 2017), pp. 4454–4461. DOI: 10.12659/msm.904675.
- [424] Piyush Joshi, George Jallo and Ranjan J Perera. “In silico analysis of long non-coding RNAs in medulloblastoma and its subgroups.” In: *Neurobiol Dis* 141 (July 2020), p. 104873. DOI: 10.1016/j.nbd.2020.104873.
- [425] Varun Keshwani, Mamta Shukla, Don W Coulter, J Graham Sharp, Shantaram S Joshi and Nagendra K Chaturvedi. “Long non-coding RNA profiling of pediatric Medulloblastoma.” In: *BMC Med Genomics* 13.1 (June 2020), p. 87. DOI: 10.1186/s12920-020-00744-7.
- [426] George Adam, Ladislav Rampásek, Zhaleh Safikhani, Petr Smirnov, Benjamin Haibe-Kains and Anna Goldenberg. “Machine learning approaches to drug response prediction: challenges and recent progress.” In: *NPJ Precis Oncol* 4 (2020), p. 19. DOI: 10.1038/s41698-020-0122-1.
- [427] Eoghan R Malone, Marc Oliva, Peter J B Sabatini, Tracy L Stockley and Lillian L Siu. “Molecular profiling for precision cancer therapies.” In: *Genome Med* 12.1 (Jan. 2020), p. 8. DOI: 10.1186/s13073-019-0703-1.
- [428] Guy Karlebach and Ron Shamir. “Modelling and analysis of gene regulatory networks.” In: *Nat Rev Mol Cell Biol* 9.10 (Oct. 2008), pp. 770–80. DOI: 10.1038/nrm2503.
- [429] Tanvi Sharma, Edward C Schwalbe, Daniel Williamson, Martin Sill, Volker Hovestadt, Martin Mynarek et al. “Second-generation molecular subgrouping of medulloblastoma: an international meta-analysis of Group 3 and Group 4 subtypes.” In: *Acta Neuropathol* 138.2 (Aug. 2019), pp. 309–326. DOI: 10.1007/s00401-019-02020-0.

Bibliography

- [430] Florian Kopp and Joshua T Mendell. “Functional Classification and Experimental Dissection of Long Noncoding RNAs.” In: *Cell* 172.3 (Jan. 2018), pp. 393–407. DOI: 10.1016/j.cell.2018.01.011.
- [431] Xing Chen, Ya-Zhou Sun, Na-Na Guan, Jia Qu, Zhi-An Huang, Ze-Xuan Zhu et al. “Computational models for lncRNA function prediction and functional similarity calculation.” In: *Brief Funct Genomics* 18.1 (Feb. 2019), pp. 58–82. DOI: 10.1093/bfpg/ely031.
- [432] Julien Jarroux, Antonin Morillon and Marina Pinskaya. “History, Discovery, and Classification of lncRNAs.” In: *Adv Exp Med Biol* 1008 (2017), pp. 1–46. DOI: 10.1007/978-981-10-5203-3_1.
- [433] Christopher R Cabanski, Nicole M White, Ha X Dang, Jessica M Silva-Fisher, Corinne E Rauck, Danielle Cicka et al. “Pan-cancer transcriptome analysis reveals long noncoding RNAs with conserved function.” In: *RNA Biol* 12.6 (2015), pp. 628–42. DOI: 10.1080/15476286.2015.1038012.
- [434] Bo Ye, Jianxin Shi, Huining Kang, Olufunmilola Oyebamiji, Deirdre Hill, Hui Yu et al. “Advancing Pan-cancer Gene Expression Survival Analysis by Inclusion of Non-coding RNA.” In: *RNA Biol* 17.11 (Nov. 2020), pp. 1666–1673. DOI: 10.1080/15476286.2019.1679585.
- [435] Yulan Deng, Shangyi Luo, Xinxin Zhang, Chaoxia Zou, Huating Yuan, Gaoming Liao et al. “A pan-cancer atlas of cancer hallmark-associated candidate driver lncRNAs.” In: *Mol Oncol* 12.11 (Nov. 2018), pp. 1980–2005. DOI: 10.1002/1878-0261.12381.
- [436] Yue Gao, Xin Li, Hui Zhi, Yunpeng Zhang, Peng Wang, Yanxia Wang et al. “Comprehensive Characterization of Somatic Mutations Impacting lncRNA Expression for Pan-Cancer.” In: *Mol Ther Nucleic Acids* 18 (Dec. 2019), pp. 66–79. DOI: 10.1016/j.omtn.2019.08.004.
- [437] Joana Carlevaro-Fita, Andrés Lanzós, Lars Feuerbach, Chen Hong, David Mas-Ponte, Jakob Skou Pedersen et al. “Cancer lncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis.” In: *Commun Biol* 3.1 (Feb. 2020), p. 56. DOI: 10.1038/s42003-019-0741-7.
- [438] Yongsheng Li, Tiantongfei Jiang, Weiwei Zhou, Junyi Li, Xinhui Li, Qi Wang et al. “Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers.” In: *Nat Commun* 11.1 (Feb. 2020), p. 1000. DOI: 10.1038/s41467-020-14802-2.
- [439] Maite Huarte. “The emerging role of lncRNAs in cancer”. In: *Nature Medicine* 21.11 (2015), pp. 1253–1261. ISSN: 1546170X. DOI: 10.1038/nm.3981.
- [440] Noa Gil and Igor Ulitsky. “Regulation of gene expression by cis-acting long non-coding RNAs.” In: *Nat Rev Genet* 21.2 (Feb. 2020), pp. 102–117. DOI: 10.1038/s41576-019-0184-5.

Zusammenfassung

Maligne Tumore akkumulieren während ihrer Entwicklung genetische und epigenetische Veränderungen, die eine Fehlregulation von Genexpression und zellulären Prozessen verursachen. Da die Regulation der Genexpression viele zelluläre Prozesse steuert, bietet die Erforschung des Transkriptom von malignen Tumoren Einblicke in die Biologie von Krebs. Schlüsseltechnologie für die molekulare Analyse von Krebstranskriptome ist Next-Generation-Sequencing (NGS) von RNA (RNA-seq) aus Tumorgewebe. Die folgende Arbeit präsentiert zwei Krebstranskriptomstudien, die sich mit der computergestützten Analyse von RNA-seq-Daten von kolorektalen Karzinomen (KRK) und Medulloblastomen (MB) unter Anwendung statistischer und maschineller Lern-Methoden (ML) befassen.

KRK ist eine klinisch herausfordernde Krankheit, da nur ein Bruchteil der Tumoren auf verfügbare Chemo- und zielgerichtete Therapien anspricht. Es wird angenommen, dass der funktionelle Verlust des Tumorsuppressors APC die initiale Mutation darstellt. Zusätzliche Ereignisse umfassen Mutationen in jeweils einem von drei RAS/RAF Proto-Onkogenen sowie Mutationen in den TGF-, PI3K- und TP53-Signalwegen. Routinemäßig verwendete Biomarker für eine Resistenz gegen den EGFR-Inhibitor Cetuximab sind RAS/RAF-Mutationen, die den Signalweg nach EGFR aktivieren. Dennoch sind einige der Wildtyp-KRKs gegen die Behandlung mit Cetuximab resistent. Um die Notwendigkeit eines besseren molekularen Verständnisses von KRK in der Präzisionsonkologie zu adressieren, hat das OncoTrack-Konsortium (Innovative Medicine Initiative) eine Multi-Omics-Strategy entwickelt, die die Einrichtung einer präklinischen Plattform für KRK-Organoid- und Xenotransplantat-Modelle integriert. In der unten vorgestellten Studie konzentrierten wir uns auf die integrative Analyse von Genexpressionsdaten und Daten zur Effektivität von Cetuximab-Behandlungen, die aus behandelten patientenbezogenen Xenotransplantaten (PBXs) gewonnen wurden. Mittels statistischer Methoden identifizierten wir eine Signatur von 241 Genen, die mit dem Ansprechen auf Cetuximab assoziiert sind. Wir verwendeten eine Support Vector Machine (SVM), ein ML-Algorithmus, um einen auf Genexpression basierenden Klassifikator zu erhalten, der das Ansprechen auf Cetuximab vorhersagt. Hier haben wir 16 prädiktive Gene mittels multiple SVM recursive feature elimination ausgewählt. Der entwickelte Klassifikator übertraf RAS/RAF-Mutationen als Prädiktor für das Ansprechen von Tumoren auf Cetuximab und schnitt gut bei RAS/RAF-Wildtyp KRK ab, für den es derzeit keine Biomarker für Cetuximab Behandlungsergebnis in der klinischen Praxis gibt.

Die zweite Studie befasste sich mit der molekularen Analyse von MB. MB, ein Tumor des Kleinhirns, ist der häufigste bösartige Hirntumor bei Kindern. Transkriptomanalysen von MB unter Verwendung von Microarrays hatten vier Haupt-Tumorgruppen ergeben, nämlich WNT, SHH, Gruppe 3 und Gruppe 4, die unterschiedliche genetische Veränderungen, molekulare Profile und klinische Merkmale zeigen. Hauptsächlich Mutationen verursachen eine Signalwegaktivierung in WNT bzw. SHH MB, während in Gruppe 3 und Gruppe 4 MB chromosomale Veränderungen häufiger vorkommen und Tumore exprimieren eher eine zelltypspezifische Gensignatur. Innerhalb dieser vier Hauptgruppen wurde eine zusätzliche molekulare Komplexität erkannt, die zur Identifizierung von Subtypen innerhalb der Hauptgruppen führte. Die Genregulationsnetzwerke, die zur molekularen Heterogenität bei MB beitragen, sind jedoch nur teilweise bekannt, und die Rolle von langen nicht-kodierenden (lnc) Genen wurde bei dieser Krankheit nur unzureichend untersucht. Um weitere Einblicke in die Molekularbiologie von MB zu gewinnen, wurde im Rahmen des ICGC das Projekt PedBrain gegründet. Als Beitrag zu diesem Projekt haben wir 164 MB RNA-seq-Proben sequenziert und analysiert. Um die Heterogenität von MB zu adressieren, identifizierten und validierten wir molekulare Subcluster innerhalb der vier Hauptgruppen. Subgruppen- und subcluster-spezifische Genexpressionsprofile wurden durch funktionelle Anreicherungen und Genregulationsnetzwerke (GRNs), die aus Genexpressionsdaten abgeleitet wurden, analysiert. Diese GRNs zeigten Gemeinsamkeiten und Unterschiede in der Genregulation zwischen Subclustern und Hauptgruppen. Durch Abschätzen des Einflusses von TFs konnten wir zum ersten Mal systematisch Hauptregulatoren der subclusterspezifischen Genexpression entschlüsseln und unbekannte Regulatoren der Gruppe 4 MB hervorheben. Darüber hinaus charakterisierten wir lnc-Gene, die in MB differentiell exprimiert waren. Unter diesen Genen haben wir 20 lnc-Gene identifiziert, die Expressionsmuster aufweisen, die mit der Gehirnentwicklung assoziiert sind, was aufgrund des embryonalen Ursprungs von MB von Interesse ist. Wir identifizierten einen Co-Expressionscluster, der bekannte krebsbezogene lnc-Gene akkumuliert und diese Gene mit der krebsfördernden Proteinbiogenese in Verbindung bringt. Überlebensanalysen zeigten das lnc-Gen MEG3 als prognostischen Marker in SHH- und Gruppe-4-Subclustern, das möglicherweise als Tumorsuppressor wirkt, der den Zellzyklus und die TGF-Rezeptorexpression negativ reguliert.

Selbstständigkeitserklärung

Name: Risch

Vorname: Thomas

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht. Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Thomas Risch

Berlin,