Information Theoretical Prediction of Alternative Splicing with Application to Type-2 Diabetes Mellitus

Axel Rasche

2009

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Leiter [advisor]:
Dr. Ralf Herwig (Max-Planck-Institut für molekulare Genetik)

Betreuer [supervisor]:
Prof. Dr. Martin Vingron (Max-Planck-Institut für molekulare Genetik)

Einreichung [submission]: 18.08.2009
Gutachter [reviewer]:
Prof. Dr. Martin Vingron (Max-Planck-Institut für molekulare Genetik)
Prof. Dr. Winston Hide (Harvard School of Public Health)
Prof. Dr. Hans Lehrach (Max-Planck-Institut für molekulare Genetik)

Disputation [defence]: 11.01.2010
Promotionskommission [PhD comitee]:
Prof. Dr. Martin Vingron (Max-Planck-Institut für molekulare Genetik)
Prof. Dr. Hans Lehrach (Max-Planck-Institut für molekulare Genetik)
Prof. Dr. Knut Reinert (Freie Universität Berlin)
Dr. Roland Krause (Freie Universität Berlin)

Contents

1 Introduction

For biomedical research it is of major interest to identify the activity of genes in specific tissues of an organism. The gene's activity is determined by the amount of the gene's primary products, the transcripts. Transcript abundance is quantified with experimental technologies and noted as gene expression. However a gene does not always produce the same transcript but may encode several different variants by a particular pooling mechanism of the genetic sequence, called alternative splicing. Such a pooling mechanism is necessary to explain the comparatively low number of genes: ∼25 000 genes in humans vs. ∼20 000 in the nematode worm *caenorhabditis elegans* [63]. Alternative splicing controls condition dependent expression of specific variants. It is not surprising that even minor splicing disturbances can have pathological effects, i.e. may cause certain diseases [279].

Since organisms like human contain ∼25 000 active genes it is essential to use high-throughput data generation techniques for analysis of global gene expression. Considering alternative splicing, all these genes stand for ∼100 000 transcripts to be analysed [38]. Only recently the necessary amount of data can be generated by technologies like microarrays or RNA-Seq. Along with technological progress the large-scale data analysis methods have to advance to cope with new research subjects like alternative splicing.

In the course of my work I have developed a software pipeline for the analysis of alternative splicing and differential gene expression. It was developed and implemented within the statistical processing language R/BioConductor [238, 108] and comprises several steps such as quality control, preprocessing, statistical evaluation of expression changes and gene set evaluation. For the detection of alternative splicing a new method based on an information theoretic concept is introduced to the field of gene expression analysis. The method consists of a modification of Shannon's entropy to detect altered transcript abundance and is called ARH – Alternative splicing Robust prediction by Entropy.

The methods and their implementation have been applied to the disease domain of type-2 diabetes mellitus. First, a set of marker genes is identified by data integration and meta-analysis of diverse data resources using the differential expression pipeline. Second, alternative splicing is analysed with the alternative splicing pipeline with special focus on a set of marker genes and on functional sets of genes, i.e. pathways.

My thesis has a truly interdisciplinary character relating the fields of information theory and statistics with alternative splicing and type-2 diabetes mellitus. As a consequence it combines the mathematical goal of splicing detection with the implementation of a processing pipeline for standardised microarray analysis as well as the joined application to type-2 diabetes mellitus.

1.1 Alternative splicing

The genetic code is the archived storage form of biological construction plans for proteins, a hard disk in analogy to computer science. The hard disk is the DNA, instead of files the genes are written to the DNA. However the genetic information is not saved in one piece but is organised in units called exons. These exons are separated by introns not encoding a protein. The gene sequence is transcribed to RNA, similar to a read-out of a file to the short-term memory. In this short-term memory the code is modified by splicing out the introns. The splicing process is very precise by assembling the exons to one transcript. But it may happen that exons are rearranged by selecting alternative splice sites along with the introns. Through alternative splicing the read-out system comprises a multiplexing process for sharing different construction plans in one gene.

For example a gene may contain four exons (e_1, e_2, e_3, e_4) (see Figure 1.1). It produces two transcripts, transcript A with the exons (e_1, e_2, e_3, e_4) and transcript B with the exons (e_1, e_2, e_4). This effect is called exon skipping with an alternative splicing event in exon 3.

After splicing the transcript is translated to a protein. The protein is the final product of the original construction plan and a functional component of the biological cell. Alternative splicing allows to encode several different proteins from the same gene sequence.Thereby it also allows to change or to modulate the function of proteins suitable for the current biological condition. Furthermore, alternative splicing determines binding properties, intracellular localisation, enzymatic activity, protein stability and posttranslational modifications of a large number of proteins, for example reviewed in Stamm et al. [272]. Recent studies estimate alternative splicing to occur in about 92-98% of human multi-exon genes [306, 228]. Even taking into account non-effective splicing, a considerable number of transcripts contribute to the functional diversity of the proteome.

Since alternative splicing is a key mechanism of protein diversity the question follows what happens in the case of splicing aberration. Indeed alternative splicing is the cause of a variety of diseases like spinal muscular atrophy, myotonic dystrophy, premature ageing phenotypes or cystic fibrosis [2]. The number of gene alterations involved in the development of any type of cancer is so high and diverse that there are many opportunities for erroneous splicing events.

With experimental technologies it is possible to determine the abundance of transcripts, the gene expression. Similarly, the exon expression is the abundance of transcripts containing a certain specific exon segment. The relationship of gene and exon expression is accented by a suggested coupling of transcription and splicing regulation [196, 206, 111, 190]. Consequently, the analysis of the two aspects of expression should be performed in parallel, as implemented in the alternative splicing pipeline presented in Section 3.3.

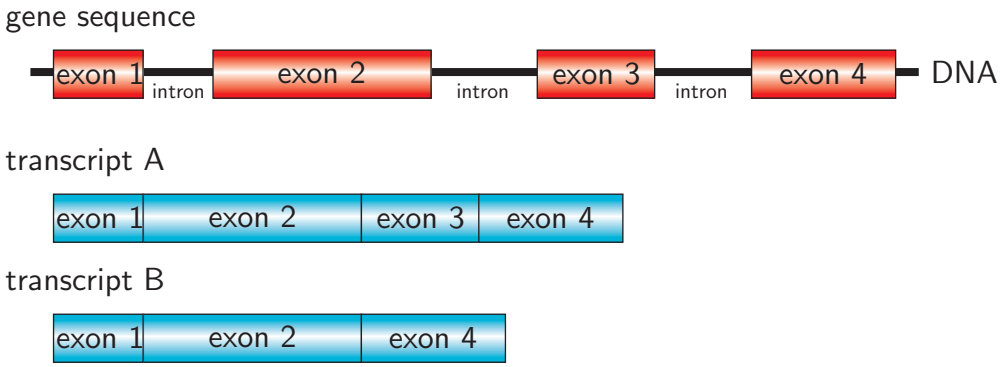


Figure 1.1: Example of an alternative splicing event. The gene is dispersed in exons over the DNA, separated by introns. The genetic sequence is transcribed and spliced to transcripts. In the process of splicing the introns are excised. Sometimes not only introns are excised but also one or more exons. The result are two transcripts encoded by the same genetic sequence.

1.2 Experimental techniques

Several experimental platforms for the analysis of alternative splicing exist such as microarrays (for example Affymetrix Exon Arrays), RNA-Seq (for example the Illumina sequencing system) and EST/mRNA libraries. The first two technologies, microarrays and RNA-Seq, are common in current laboratory facilities. Commercial providers offer complete systems with specific protocols, laboratory kits and necessary consumable articles.

Affymetrix GeneChip microarrays

Affymetrix microarrays consist of probes of 25 nucleotides length corresponding to transcript sequences that are synthesised on a slide with a photolithographic method. A dye labelled sample is injected on the slide. Labelled transcripts in the sample will hybridise to the corresponding probes. Afterwards the slide is scanned on the wavelength of the dye and the light intensity of the spots allows quantification of transcript abundance in the original sample, the gene expression.

The GeneChip© system comprises a number of technologies, design criteria and fixed protocols described in detail in Dalma-Weiszhausz et al. [78]. For this system several platforms are available depending on the intended use: Transcriptome mapping, gene expression profiling or genotyping accompanied by a custom array program. In the following the focus is on expression arrays.

Affymetrix arrays have been successfully used for analysis of gene expression for more than a decade. Rich experience is available for the interpretation of the data, see Clevert and Rasche [65] for an outline. Due to selection of probes and labelling protocol former microarray platforms have not been useful for detection of splice variants. As a consequence, existing alternative splicing in the samples under study imposes a severe and undetectable bias to those experiments. With the Affymetrix Exon Arrays this tool

is expanded from gene to exon level for quantitative studies of alternative splicing on a genome-wide scale. Although the alternative splicing detection bias is removed these new arrays pose additional hard challenges to the statistical analysis, namely the expression analysis on two information levels, gene and exon expression.

Illumina sequencing

Another technology facilitating alternative splicing analysis is RNA-Seq. The mRNA is sequenced by 2^{nd} generation sequencing technology, for example Illumina sequencing (i.e. Solexa technology) [142]. It is a variant of the shotgun sequencing approach providing a huge number of small sequences with 25 to 100 nucleotides.

Sequencing templates are immobilised on a suitable flow cell surface designed to present the DNA in a manner that facilitates access to enzymes while ensuring high stability of surface-bound template and low non-specific binding of fluorescently labelled nucleotides. Bridge amplification is employed to create up to 1000 identical copies of each single nucleotide sequence in close proximity. Illumina sequencing technology can achieve densities of up to ten million single nucleotide sequence clusters per square centimetre.

Illumina sequencing uses four fluorescently-labelled modified nucleotides to sequence the millions of clusters present on the flow cell surface. These nucleotides, specially designed to possess a reversible termination property, allow each cycle of the sequencing reaction to occur simultaneously in the presence of all four nucleotides. In each cycle, the polymerase is able to select the correct base to incorporate, with the natural competition between all four alternatives leading to higher accuracy than methods where only one nucleotide is present in the reaction mix at a time.

The Illumina sequencing approach is built on a very large number of short sequence reads. Deep sampling allows the use of statistical analysis, similar to conventional methods, to identify transcripts and to distinguish sequencing errors. Each raw read base has an assigned quality score so that the software can apply a weighting factor in calling differences and generating confidence scores.

1.3 Concepts of information theory as a measure of exon diversity

Information and entropy

A source of information sends symbols from a discrete alphabet $\mathcal{A} = \{a_1, ..., a_D\}$ with probabilities $p(a_1), ..., p(a_D)$. The probabilities suffice the equation $\sum_{d=1}^{D} p(a_d) = 1$. The information content of a single letter a_k of the alphabet is defined as $I(a_d) = -\log_2 p(a_d)$. This definition of the information content is motivated by the following characteristics:

- Non-negativity: The information content of a single letter is non-negative because $0 \leq p(a_d) \leq 1$ for $1 \leq d \leq D$.

- Monotony: Rare symbols shall have a high information content because they are unexpected, whereas frequent symbols shall have low information content. Thus information content has to increase reciprocally with its probability.
- Additivity: Information of D independent symbols $a_1, ..., a_D$ should be equal to the sum of information of the single symbols. This is implied in the functional characteristic of the logarithm

$$I(a_1, a_2) = -\log_2 p(a_1, a_2) = -\log_2 p(a_1) - \log_2 p(a_2) = I(a_1) + I(a_2)$$

The base of the logarithm, i.e. base 2, is motivated from information theory where information is stored in bits. The more bits are stored in a memory unit the higher is its information content. For example a unit of N bits can store one of 2^N binary coded digits which has an information content of $-\log_2 \frac{1}{2^N} = N$.

To characterise the information source completely the information content of the single letters is not sufficient. Therefore entropy is introduced which is defined by the expected information content or mean information content with respect to the probabilities of the alphabet $\mathcal{A}$, i.e.

$$H(a_1, ..., a_D) = -\sum_{d=1}^{D} p(a_d) \log_2 p(a_d). \tag{1.1}$$

This definition is motivated by the following characteristics:

- Continuity: The measure should be continuous, so that changing the values of the probabilities by a very small amount should only change the entropy by a small amount.
- Symmetry: The measure should be unchanged if the outcomes are re-ordered:

$$H(a_1, a_2) = H(a_2, a_1).$$

- Maximum: Uncertainty is highest when all possible events are equiprobable:

$$H(a_1, a_2) \leq H(a_1', a_2') \text{ with } p(a_1) \neq p(a_2) \text{ and } p(a_1') = p(a_2').$$

  For equiprobable events the entropy should increase with the number of outcomes, i.e. for $i = 1, ..., D, j = 1, ..., D + 1$:

$$H(a_1, ..., a_D) < H(b_1, ..., b_{D+1}) \text{ with } p(a_i) = \frac{1}{n} \text{ and } p(b_j) = \frac{1}{n+1}$$

It is important to note that the maximum of the entropy does not depend on the probabilities but only on the number of symbols:

$$\max(H) = H(a_1, ..., a_D) = \log_2 (D) \text{ with } p(a_i) = p(a_j) \text{ for } i, j \in \{1, ..., D\} \tag{1.2}$$

An example illustrating the definitions would be a fair dice with 6 possible outcomes $\{1, 2, 3, 4, 5, 6\}$. The probability for the 6 events are $\frac{1}{6}$ each. Information content of the

events is about 2.58 each. The entropy of the dice also is about 2.58. In contrast a non-fair dice with probabilities $\frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}$ has information contents 1, 3.32, 3.32, 3.32, 3.32, 3.32 and an entropy of 2.16.

Entropy is a measure of uncertainty in the information source. When letters are equiprobable the uncertainty is highest, as is entropy. If all but one letter have small probabilities the information source has low uncertainty and small entropy. The concept of information and entropy was introduced by Shannon [261] in the context of communication theory, see also Cover and Thomas [71].

Entropy was previously used in alternative splicing for quantification of global splicing disorders [247]. For the transcript fractions of a gene the entropy is calculated in normal and cancer tissue. Entropy of such transcript fractions is higher in cancer tissues than in normal tissues, indicating general splicing disruption in cancer. Here, entropy is applied to splicing prediction, analysing exon expression ratios.

Exon expression deviations as a source of information

A splicing event can be viewed as a disturbed exon expression between two biological samples. Disturbed exons contain information about the splicing which is measured in an information theoretical way. Exon expression ratios are contrasted by entropy taking noise into account. Extreme ratios deviating from the average ratio indicate splicing.

Looking at the example of an alternative splicing event in Section 1.1 there is a gene with four exons. In one sample transcript A is expressed comprising all four exons. In a different sample transcript B is expressed with exon 3 missing. Looking at the exon expression ratios the three constitutive exons, ratios will scatter around 1. For exon 3 the ratio deviates strongly from 1. With the gene as an information source the ratios are taken as letters. Scaling the ratios to a sum of 1, probabilities are derived for the four letters, i.e. exons. The ratio of exon 3 dominates the probabilities and the entropy of the information source is small. Thus splicing events correspond to small entropies. This basic idea is extended and adapted to practical needs in Chapter 4.

The need of prediction methods is two-fold. First, gene level predictions are needed to identify spliced genes and quantify the global amount of splicing between two biological conditions. Secondly, the exon level predictions are needed to identify the exact splicing event in the gene sequence. Although many isoforms and splicing events are known from EST/mRNA collections current databases do not cover all possible biological conditions. Therefore the presented method provides *de novo* predictions of splicing events not taking into account previous knowledge about transcript structures.

1.4 Type-2 diabetes mellitus

Type-2 diabetes mellitus is a rapidly increasing disease with more than 170 million persons affected worldwide, constituting more than 90% of all diabetic patients [283]. Type-2 diabetes mellitus poses a huge burden for the health care systems and is, thus, subject

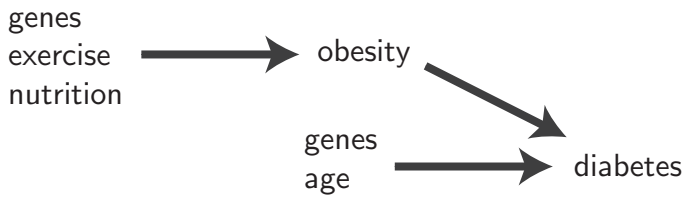


Figure 1.2: Main risk factors of type-2 diabetes mellitus. Missing physical exercise with over-nutrition in a disadvantageous genetic background leads to obesity. Obesity in a disadvantageous genetic background leads to diabetes at an early age.

to intensive biomedical research. Type-2 diabetes mellitus is a multigenic disease involving a high number of susceptibility genes and causes alteration of an entire network of genes. Several environmental and nutritional risk factors have been identified for type-2 diabetes mellitus, the most relevant being obesity where multiple molecular mechanisms have been proposed to link obesity to insulin resistance and β-cell failure (see Figure 1.2) [283]. Increased availability of food and reduced physical activity as consequences of modern lifestyle are the main drivers for an anticipated epidemic increase in type-2 diabetes mellitus patients in the years to come.

In the pathopysiology of type-2 diabetes mellitus, impaired insulin sensitivity and glucose intolerance are early phenomena, leading to hyperglycaemia, hyperlipidemia and, eventually, to a failure of pancreatic β-cells to produce and secrete a sufficient amount of insulin. However, most genes and their associated molecular network contributing to the onset and course of the disease are yet unknown. For example it is not clear which molecular effects lead to β-cell dysfunction [269], how enlarged fat mass causes insulin resistance [153] or what promotes the pathogenesis of the inflammation [79].

In the context of type-2 diabetes mellitus, a broad range of experimental techniques were applied for identification of disease genes reviewed in Rasche (2009) [239]. Genetic variation studies cover monogenic approaches with transgenic or knock-out mouse models and genome-wide approaches with association or linkage studies [64]. Such analyses have shown that several quantitative trait loci interact with each other and an effect of these variants on disease susceptibility is generally low. Multiple studies on the transcriptome level have been performed that emphasise the diversity of the disease and the complex pathophysiological interactions between different tissues, including fat, muscle, liver, pancreatic β-cells and brain [283, 285]. In several human studies, tissue biopsies from diabetic and normoglycaemic individuals have been profiled [117, 207]. In mouse studies differences in diet or mouse strains have been used to identify distinct expression profiles [35, 177, 211]. In the context of the onset of diabetes, several studies on the proteomic level have revealed differential expression of intracellular proteins as well as of secretory proteins in adipose tissue [58, 291]. A systematic evaluation of these large amounts of data, their common content as well as their specific differences, in particular in gene sets between human and rodent studies is performed in Rasche et al. [240].

Little is known about the role of alternative splicing in the onset and progression of type-2 diabetes mellitus. Functional effects of different isoforms Minn et al. [203] report

for the key player insulin. Similarly, isoform expression changes are found in signalling pathways and protein hubs [94, 95, 186]. Thus, splicing potential is given for type-2 diabetes mellitus markers. However, global analyses of splicing are still a gap in type-2 diabetes mellitus research and would complement available data sources.

The two mouse strains *NZO* and *NZL* are particularly interesting due to a polygenic background for type-2 diabetes mellitus similar to human [226]. The *NZL* mouse is a near relative to the *NZO* mouse with a higher prevalence of hyperglycaemia. The two strains are contrasted by the genetically different *SJL* mouse. The *SJL* strain stays lean and non-diabetic even on high-fat diet.

In this work the established alternative splicing analysis framework is applied to type-2 diabetes mellitus with a genome-wide assay. In a comprehensive study splicing of disease markers and global splicing states are assessed. Experiments are performed on the glycaemic and genetic background of two complementary mouse models *NZL* and *SJL* in the relevant tissues fat and liver.

To assess type-2 diabetes mellitus disease markers a robust list of such genes has to be established. Candidate gene lists are mainly focussed on special experimental approaches like sequencing [230, 298] or interactions between transcription and protein networks [191]. In Rasche et al. [240] the aim was to aggregate different heterogeneous resources and a list of 213 disease markers was established by conducting a statistical meta-analysis. For the purpose of the thesis the list was augmented for including the *NZO* mouse, with time series on disease stages and early stage expression experiments. Thus, the published approach has been extended and resulted in a list of 655 marker genes which is the starting point for the alternative splicing analysis.

1.5 Aims of the thesis

Merging information theory and statistics with molecular biology, the thesis tackles the challenge to achieve several aims. The interdisciplinary task balances applied mathematics and theoretical elegance vs. biomedical research to return practical, verifiable results. Here the detection of different isoforms between different biological conditions is the main mathematical objective. The thesis has three main purposes:

- Introduction of elements of information theory to the prediction to alternative splicing as a new statistical method.
- Implementation of a differential expression pipeline and alternative splicing pipeline in order to provide a standardised evaluation of Affymetrix 3' arrays and Exon Arrays.
- Assessment of genome-wide alternative splicing states in type-2 diabetes mellitus.

The differential expression pipeline enables marker identification and is the basis for the development of the alternative splicing pipeline. The information theoretical element in alternative splicing is elaborated for fast and robust splicing prediction. In the alternative splicing pipeline the splicing prediction joins the differential expression experience. These

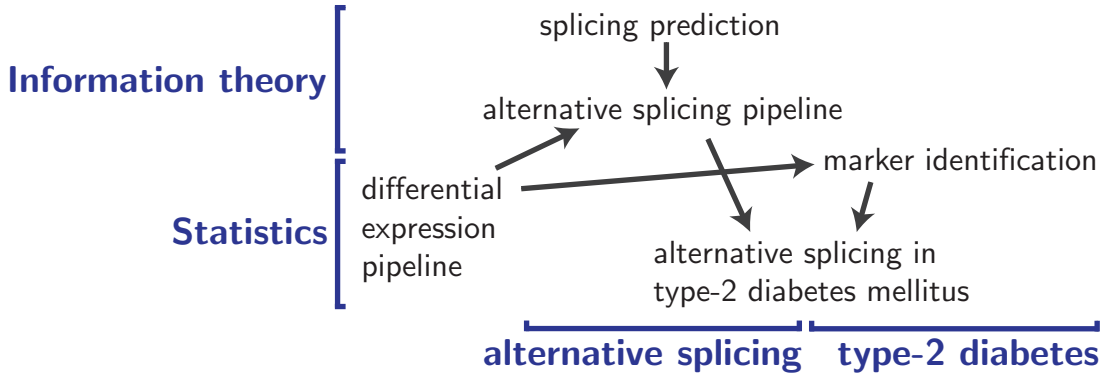


Figure 1.3: Objectives and their relations. The thesis relates information theory and statistics with alternative splicing and type-2 diabetes mellitus.

products are applied in type-2 diabetes mellitus for marker identification and analysis of global splicing states as depicted in Figure 1.3.

In **Chapter 2** basic aspects of the complex field of alternative splicing are introduced. Alternative splicing is a mechanism to increase the proteome from a smaller than expected code basis on the level of genes. A fraction of transcript variation is biologically relevant and related to functional changes. The precise splicing process is accompanied by an amount of splicing errors partially leading to a number of known splicing based diseases. Established therapies are rare but several approaches are available.

Global analysis of alternative splicing is performed with three technologies: EST or mRNA based alternative splicing databases, microarray experiments and RNA-Seq. All these technologies provide global estimates of alternative splicing and in the review the estimates are consolidated to estimate the prevalence of splicing. Specially designed splicing data sets are missing for exon arrays. In my work this is circumvented by contrasting a data set to a database. A manual selection and collection of literature based alternative splicing events in the AEdb is the necessary test set for the prediction evaluation in Section 4.4 [274].

In **Chapter 3** the implementation of the microarray analysis pipelines is elaborated. Comparison of results between different data sets and meta-analysis requires a standardised and controlled data analysis processing both for alternative splicing and differential expression [147]. Methods for differential expression analysis established for Affymetrix 3' gene expression arrays have been adapted to the specific design of the Exon Arrays. The alternative splicing pipeline consists of two branches, differential expression analysis and alternative splicing analysis. No method evaluation or accepted procedures are available for splicing evaluation. Thus, the alternative splicing analysis is developed in Chapter 4 and then implemented as a module in the pipeline.

Guidelines for the pipelines are implementation in R/BioConductor [238, 108], internal handling on Ensembl genes or exons with gene-wise analysis [38], modular design as well as division of complex experimental settings into different test cases. The application

of statistics depends on the experimental technology and the analysis objective [304]. Thus, the particular design of the Affymetrix 3' gene expression and Exon Array is introduced [97, 194]. The chip design allows to reinterpret the probe assignment to genes and transcripts in the light of the advancement of the genome sequence.

The processing of the arrays follows two fundamental principles: First, only biologically motivated corrections on the data are allowed with statistical models. Second, comparability between experiments is essential. The processing is highlighted with the steps quality control, test case determination, preprocessing and data evaluation [239]. Where research on 3' arrays is settled, the analysis of exon microarrays has posed new challenges to the computational analysis like data normalisation and presence tag calculation. Probe binding affinity is corrected by GC content of the probe sequences and intensity distributions are adjusted by quantile normalisation. Data evaluation establishes an array-wide gene analysis followed by the isolation of a set of alternatively spliced genes as well as a set of differentially expressed genes. It follows the gene set evaluation with over-representation analysis and group testing on a diverse set of functional resources like pathway databases, transcription factor targets, drug targets and tissue expression.

In **Chapter 4** the concept of entropy is introduced to the field of alternative splicing prediction. It develops a new method called ARH – Alternative splicing Robust prediction by Entropy [241]. The primary goal is to develop a method which is robust in the number of replicates and independent from the number of exons. For comparison, eight different methods proposed for splicing prediction on exon arrays are presented. In a broad evaluation the performance is assessed on several aspects like dependency on the numbers of exons, splicing prediction in the case of differential expression or no differential expression and robustness in the numbers of replicates. The evaluation runs on a tissue data set and in an artificial setting with a spike-in experiment resulting in a total of four different test settings: pairwise tissue comparison with database confirmed events, tissue specificity with database confirmed events, tissue specificity with RT-PCR validated events as well as the *in vitro* samples with generated events. The focus is on detection of exon skipping events. Design of the exon arrays is just adequate for this type of splicing events (see subsections 2.1.1 and 3.1.2).

In **Chapter 5** the power of the pipelines and the ARH prediction is shown with an application in the context of type-2 diabetes mellitus. An introduction to type-2 diabetes mellitus elucidates the interplay of different organs and factors to highlight two major organs for disease progression, adipose tissue and liver.

For marker identification a meta-analysis is performed on diverse qualitative and quantitative sources Rasche et al. [240]. The quantitative sources are microarray data sets processed with the differential expression pipeline. In every source disease relevance of genes is scored. Scores are summed up to a gene score rating the general relation of a gene to type-2 diabetes mellitus. Assessing consistency in the gene score another use arises of the entropy introduced in Section 4.1. High entropy identifies genes with consistent type-2 diabetes mellitus relevance over many sources.

For two mouse models of type-2 diabetes mellitus hybridisations on exon arrays have been performed at the German Institute of Human Nutrition (DIfE). Mice are all fed on

a high-fat diet and on this dietary background *NZL* animals develop obesity. Diabetic mice are separated by levels of blood glucose. In contrast the *SJL* animals do not develop obesity by genetic reasons. Samples of fat and liver tissue are prepared and with the alternative splicing pipeline spliced genes are identified and attributed to glycaemic or genetic causes.

2 Aspects of Alternative Splicing

In Chapter 2 basic aspects of the complex field of alternative splicing are introduced. Alternative splicing is a mechanism to increase the proteome from a smaller than expected code basis on the level of genes. A fraction of transcript variation is biologically relevant and related to functional changes. The precise splicing process is accompanied by an amount of splicing errors partially leading to a number of known splicing based diseases. Established therapies are rare but several approaches are available.

Global analysis of alternative splicing is performed with three technologies: EST or mRNA based alternative splicing databases, microarray experiments and RNA-Seq. All these technologies provide global estimates of alternative splicing and in the review the estimates are consolidated to estimate the prevalence of splicing. Specially designed splicing data sets are missing for exon arrays. In my work this is circumvented by contrasting a data set to a database. A manual selection and collection of literature based alternative splicing events in the AEdb is the necessary test set for the prediction evaluation in Section 4.4 [274].

2.1 Biological background

DNA is the basic storage form of genetic information. Due to its double-strand nucleotide structure it is possible to copy or inherit the information and provide repeated read-outs from the storage. The read-out process transcribes the DNA information to RNA. RNA is a short-term memory for biological information. On the one hand RNA is continuously processed and on the other hand RNA is permanently degraded. Newly transcribed RNA is called pre-mRNA.

Information on the DNA is structured in genes. Genes are segments on the DNA consisting of information blocks called exons. Exons are divided by introns, long DNA stretches which do not contribute to functional information. These introns are excised from the pre-mRNA by splicing. The RNA product after splicing is called mRNA. Finally,the mRNA is translated to a protein.

The vertebrate splicing machinery is not only capable of accurately recognising the small exons within the larger intron context, but is also able to recognise exons alternatively. In this process, an exon is incorporated into the mRNA, excised like an intron or exon bounds are varied. This process is called alternative splicing and is abundantly used in higher eukaryotes [295, 22]. For more information about the biochemical mechanisms of alternative splicing, readers may refer to some excellent reviews by Black [40], Nilsen [216], Matlin et al. [199], Hertel [125], Maniatis and Tasic [196]. The alternative transcripts or proteins encoded from the same gene are called isoforms.

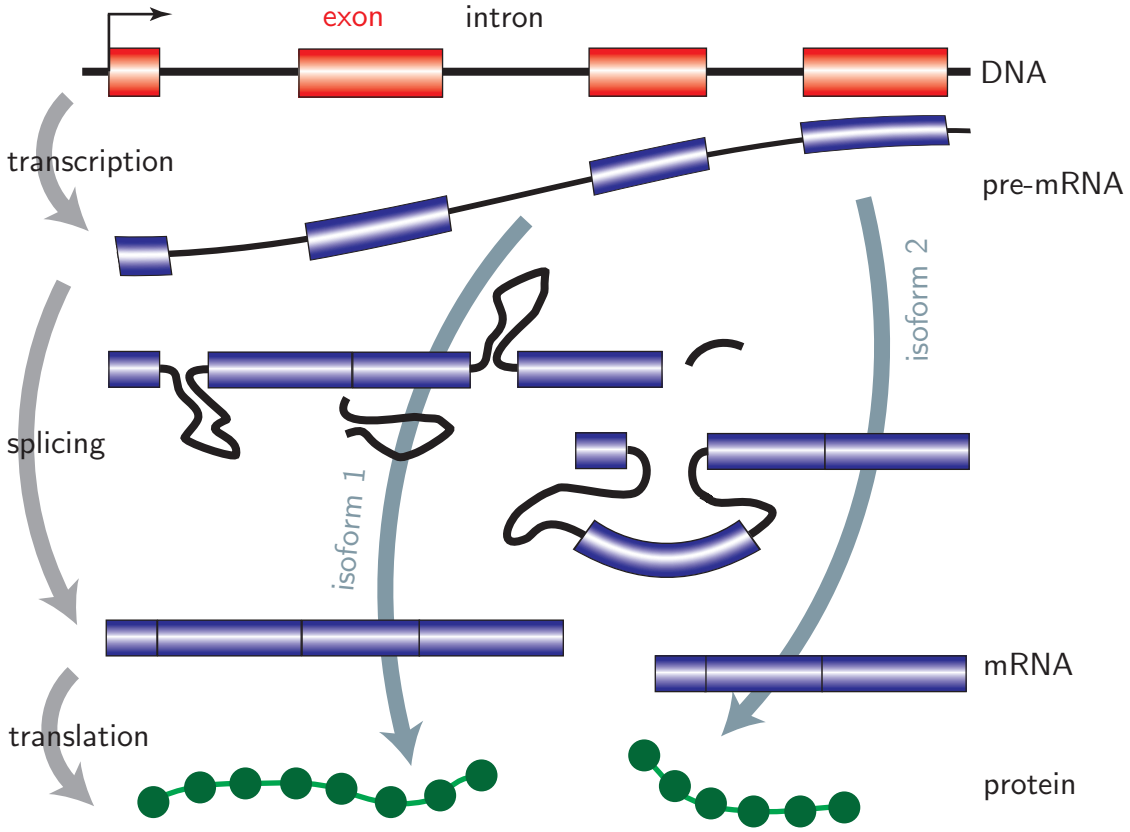


Figure 2.1: From DNA over RNA to protein. In the course of splicing from the same gene read-out different mRNA variants are processed for translation to different isoforms.

2.1.1 Alternative splicing patterns

Alternative splicing events can be subdivided into different classes (see Figure 2.2). Exons can be skipped at any position within the transcript. This is a simplifying view with microarrays in mind. Alternative transcription start (alternative first exon) or alternative transcription end (alternative last exon) are not alternative splicing in a proper sense but relate to the transcription process. The events are detectable by exon expression changes and thus are included in this work. Skipping one exon the event is called exon cassette mode. Database analysis from human curated sets revealed that cassette exons are the most common type, representing about half of alternative exons [295]. Additionally skipping adjacent exons it is a multiple exon skipping event. Several transcript variants can be present in a sample with skipping events in different exons constituting a mutually exclusive exon event. More splicing classes use different splice sites or ignore these sites resulting in alternative acceptor or donor sites or complete intron retention in the transcript.

Most alternative splicing patterns involve the choice of splice sites competing against

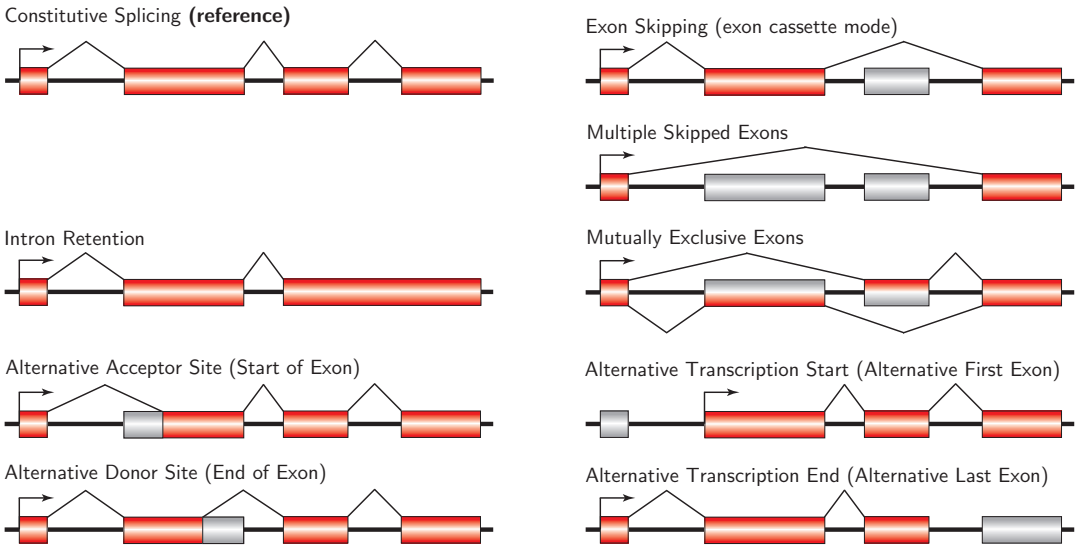


Figure 2.2: Splicing scope. From all splicing possibilities only the exon skipping events are detectable with exon arrays. They are visualised on the right part of the figure.

each other. One alternative splicing pattern in which this may not be the case is intron retention. Here, the choice is between splicing with intron excision and no splicing with the retention of an intron in the final mRNA [40]. The partially spliced RNA must then be exported to the cytoplasm. Thus for intron retention, the competition may be between splicing and mRNA transport rather than between two splicing patterns.

Not all splicing patterns are conserved between different species [40]. Even within a species, there appears to be remarkable variation in the use of particular splice variants [215]. A possible explanation for this is that alternative splicing provides an advantageous mechanism for testing new protein sequences during evolution. A transcript may be varied by a single point mutation extending an exon or creating a new exon. Encoding a new protein, it may comprise only a few percent of the original mRNA. Mutations altering the splicing allow the production of new proteins without crucial loss of the original protein.

2.1.2 Increase of the proteomic diversity

A major undertaking of the post-genomic era will be the description and functional characterisation of the proteome [115]. The number of proteins in the proteome is by no means equivalent to the number of genes, but can exceed it by orders of magnitude. Mechanisms that increase protein diversity in all metazoans include alternative polyadenylation, post-translational protein modifications as well as the use of patterns listed in Figure 2.2. Among these mechanisms, alternative pre-mRNA splicing is considered to be the most important source of protein diversity in vertebrates [196]. An overview about the number of consolidated transcripts per gene is depicted in Figure 2.3 for the

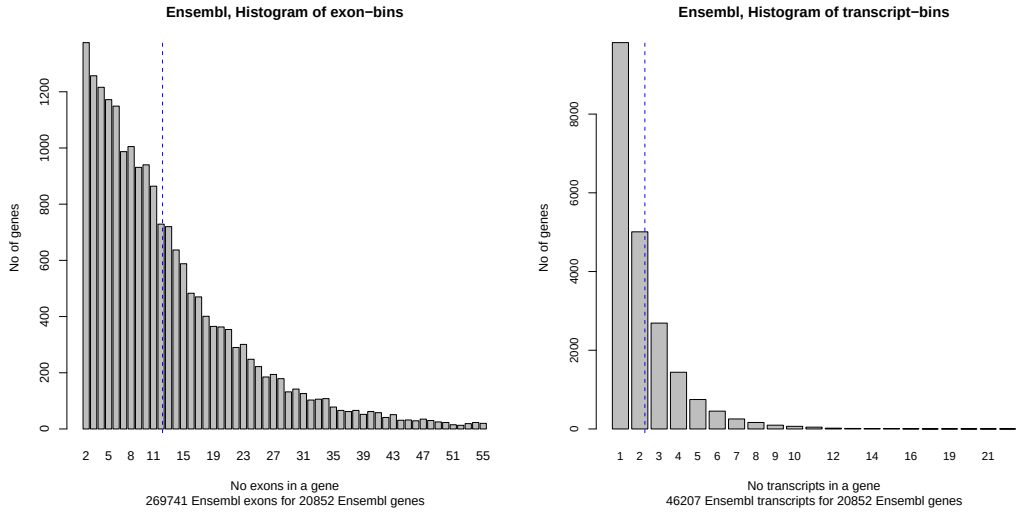


Figure 2.3: Ensembl 53 statistics. Ensembl gene identifier are filtered for known protein coding genes. The blue dashed line indicates the average. **Left:** Histogram of the number of exons per gene. **Right:** Histogram of the number of transcripts per gene.

Ensembl database [38]. 20 852 Ensembl genes have on average 13 exons (12.94) encoding 2 proteins (2.22).

How many alternatively spliced transcripts can be synthesised from a single gene? If each gene produces two isoforms, the proteome would be twice as big as the genome. The complexity of the proteome would be significantly higher if alternative splicing can create a much greater extent of transcript diversity [115]. It turns out that several genes do encode transcripts that are alternatively spliced to produce a vast number of different mRNAs. Mathematically the upper bound is given by the number of combinatorial combinations of the above listed alternative splicing patterns with the number of exons in the gene. However studies imply that in most of the splicing events only one or a few exons are affected [115, 319].

The regulation of alternative splicing producing this number of proteins under different biological conditions is still under intense investigation [295, 93]. The relative concentration of splicing-associated proteins can regulate alternative splice site selection. These proteins are often combined in complex ways into multiple layers of regulation. Many of these proteins can act either positively or negatively depending on their binding context.

2.1.3 Function and biological relevance

It is unquestionable that alternative splicing can generate mRNAs with important and distinct biological functions [29]. The more difficult question is what fraction of the extensive transcript variation generated by alternative splicing is truly biologically relevant and what fraction may be due to stochastic noise in the splicing process. Bioinformatic analyses indicate that 75% of alternative splicing events affect coding regions, with pre-

dicted effects ranging from subtle amino acid substitutions to removal of protein motifs or protein truncations [29].

By far the best-understood system of splicing regulation comes from the *somatic sex determination* pathway in Drosophila melanogaster [40]. A series of genetic studies identified the key regulators of the *somatic sex determination* pathway as RNA binding proteins that alter the splicing of particular transcripts.

Stamm et al. [272] review the function of alternative splicing. Three different main categories are listed:

- Introduction of stop codons
- Addition of new protein parts
- Influence on mRNA function

The function with most obvious biological consequences seems to be the introduction of stop codons. About 25-35% of alternative exons introduce frameshifts of stop codons into the pre-mRNA and 18-25% of transcripts will be switched off by stop codons caused by alternative splicing and nonsense-mediated decay [272].

A major part of splicing events leads to the addition of new protein parts. Changes in the protein primary structure can alter the binding properties of proteins, influence their intracellular localisation and modify their enzymatic activity or protein stability by diverse mechanisms [272]. One commonly found mechanism is the introduction of protein domains that are subject to posttranslational modification. The range of changes spans from a complete loss of function to very subtle modulations of function that are only detectable by specialised methods. Consistent with the idea that alternative splicing plays important roles in cellular function, bioinformatic and array data indicate that the process is more prominent in tissues with diverse cell types and among genes playing regulatory functions [29].

2.1.4 Splicing errors

A fundamental problem in pre-mRNA splicing is 'exon recognition', the process by which exons are distinguished from introns, and intron–exon boundaries are accurately defined. The average size of a human exon is 150 nucleotides, whereas introns average around 3500 nucleotides [196]. Thus, the splicing machinery must recognise small exon sequences located within vast stretches of intronic RNA. Moreover, splice sites are poorly conserved, and introns contain large numbers of cryptic splice sites. Cryptic splice sites are normally avoided by the splicing machinery, but can be selected for splicing when normal splice sites are altered by mutation. Mutations that alter splicing can allow production of new proteins without essential loss of the wild-type protein [40]. Although this might be advantageous for protein evolution, the high degree of variability in splicing makes it difficult to prove the significance of a splice variant that is not conserved across species.

Most of the known splicing errors are caused by mutations in the genomic DNA that destroy a normal splicing signal or create a new one [115, 141]. Other instances of

splicing errors are caused by mutations in splicing regulatory proteins or their binding sites. There are two possible fates for inappropriately spliced transcripts:

- translation or
- degradation.

In the first case these transcripts result in the production of aberrant proteins. Mostly these useless proteins will not affect the cell. In the second case erroneous spliced transcripts are just degraded.

Eukaryotes possess an mRNA surveillance system that scans newly synthesised mRNAs for the presence of premature stop codons and, if detected, degrades the defective mRNAs to prevent their translation [115]. A number of different splicing errors leads to the insertion of premature stop codons in a transcript. For example, improperly including exons or retaining introns that contain stop codons can direct a transcript to the *mRNA surveillance* pathway. In any case many alternatively spliced transcripts pass the surveillance system and are translated on a large scale [301].

Only for a handful of alternative splicing events a clear function and biological relevance is described today. However, experiments that address functional differences between protein isoforms encoded by alternatively spliced transcripts have not even been performed [227]. However, perhaps not all splicing events have functional consequence and may be considered to be splicing noise. Comparative genomics is a way to determine whether an alternative splicing event is real or represents noise. The nucleotide sequence of functionally neutral alternative exons will not be under selective pressure and should not be conserved in distantly related species, whereas functionally relevant alternative exons should be. However, alternative exons not evolutionarily conserved are not necessarily unimportant but might have evolved recently. This is not that far-fetched because single point mutations may create new splice sites [54, 141] and potentially introduce a new alternative exon. In fact, there are documented cases in which the splicing patterns of paralogous genes differ across species [115].

2.1.5 Alternative splicing in disease

A mounting body of evidence implicates splicing defects and altered splicing regulation as causes or modifiers of numerous pathologies [29, 218]. Both computational predictions and microarray experiments have identified hundreds of alternative splicing and aberrant splicing events associated with disease states, particularly various cancers [29]. Cartegni et al. [54] points to a misunderstanding of point mutations in the DNA possibly affecting splicing signals. Most splicing mutations affect the standard consensus splicing signals, and typically lead to skipping of the neighbouring exons. Less frequently, the mutations create an ectopic splice site or activate a cryptic splice site, thereby changing the overall splicing pattern of the mutant transcript [54]. Consequently Pagani et al. [227] link genomic changes in introns and exons to disease by several mechanisms.

These add to a growing list of alternative splicing changes and aberrant splicing events known to affect cellular features relevant for tumor growth, including cell transformation,

motility, invasiveness, and angiogenesis [29]. Consistent with the notion that changes in alternative splicing can be influential to cellular phenotypes French et al. [99] report correlations between transcript features and lymphoma grade, proper classification of histological distinct tumors, using Affymetrix Exon Arrays. As examples for diseases caused by alternative splicing the following have been described [2, 69, 91]: cystic fibrosis, spinal muscular atrophy, familial isolated growth hormone deficiency type II, Frasier syndrome, Frontotemporal dementia and parkinsonism linked to chromosome 17, retinitis pigmentosa and myotonic dystrophy.

Wang and Cooper [307] dissect the alternative splicing in cis-acting and trans-acting mutations as well as trans-dominant effects causing alternative splicing in disease [307, 91]. The mechanisms causing altered splicing involve disruption of either cis-acting elements within the affected gene or trans-acting factors that are required for normal splicing or splicing regulation. Effects in cis have a direct impact on the expression of only one gene whereas effects in trans have the potential to affect the expression of multiple genes.

Detailed knowledge about isoform expression provides the possibility to identify novel, more specific, and safer targets for drug design. In this regard, individual variation in splicing patterns related to population haplotypes may add yet another dimension to personalised medicine [29]. Recent analyses of HapMap cell lines document variations in alternative splicing among different individuals, an observation with significant basic and medical implications [29].

2.1.6 Therapy of diseases caused by alternative splicing

The proteome of a cell can rapidly change in response to extracellular stimuli through complex signal-transduction pathways [196]. Changes in protein composition can be regulated on many different levels, but have been studied primarily at the level of transcription and posttranslational protein modification. Stamm [271] describes several signals changing the selection of splice sites. Signalling pathways are activated for example by neuronal activity or stress. Yeo complementary reviews two high-throughput data sets for RNAi as well as compounds identifying targets and drugs directing alternative splicing events [318].

These studies call for the development of diagnostic and therapeutic means targeted at alternative splicing. Stoilov et al. [280] list six approaches as therapeutic strategies, namely antisense oligonucleotides, RNAi, ribozymes, SMaRT, low molecular weight drugs and expression of trans-acting factors.

More in detail Garcia-Blanco et al. [103, 102] describe the different approaches divided in conventional therapeutics, oligonucleotide-mediated therapies and RNA-based corrective therapy. The conventional therapeutics use small molecules to target specific isoforms of proteins or the gene expression. Antisense oligonucleotides were first used in the 1970s to inhibit a virus in tissue culture [277, 322]. A potential in human therapy has remained anticipated but unrealised. A group of methodologies has been developed to reprogram mRNAs to modify the outcome of alternative splicing decisions. E.g. splicing reactions can be redirected in the nascent primary transcript to prefer certain isoforms over others

Technology	Estimate	Cit.
alternative splicing database	40-60%	[183]
microarray	74-88%	[183, 149]
RNA-Seq	92-98%	[306, 228]

Table 2.1: Global splicing estimates. Current studies of different high-throughput technologies show different ranges of genome-wide alternative splicing prevalence.

(SMaRT).

2.2 Global analysis with high-throughput technologies

Currently three technologies facilitate genome-wide analysis of alternative splicing: EST or mRNA databases, splicing-sensitive microarrays as well as RNA-Seq experiments [29, 41]. EST/mRNA are collected in huge databases and already helped to determine the gene structure of the DNA. Sequence analysis shows that alternative exons often have unusual lengths, suboptimal splice sites and characteristic nucleotide patterns. Despite this progress alternative exons cannot be predicted *ab initio* from genomic data, which is due to the degenerate nature of splicing signals [115].

Often the interest is to analyse alternative splicing in a special sample under study. One would like to examine whether a particular exon is coregulated with others, and to test how a whole ensemble of splice variants is altered by a particular condition. Such condition specific analysis is possible in modern lab environments with microarrays and RNA-Seq.

Stating the existence of alternative splicing the question arises how widespread it is in the genome. Alternatively spliced exons can be found by sequence comparison of genomic, mRNA and EST sequences. Furthermore, a large number of alternative exons have been described in the literature [115]. The current state of databases lists 68% of all genes to be alternatively spliced with about 9 exons per gene [38, 273, 296]. Using a variety of experimental designs and biological samples, splicing-sensitive microarrays studies have produced consistent estimates of 74% to 88% of alternatively spliced genes in the human genome [149, 156]. Global splicing estimates are increasing in time and technology (see Table 2.1). Since the maximal bound is practically reached (with 98%) in RNA-Seq studies, estimates will probably converge in future.

2.2.1 Alternative splicing databases

The research on alternative splicing first was the domain of EST libraries. EST/mRNA databases are important to structure the genome sequences and identify genes. The goal of finding the exon-intron structure goes hand in hand with splice site detection. It was on this task when slowly the community realised that alternative splicing occurs, not rarely but for the majority of genes [205, 157, 149]. This opened the research field of

alternative splicing variant detection and was accelerated by the complete sequencing of the human genome [178].

The alternative splicing databases discussed in this Chapter are indispensable for designing arrays. The use of microarrays depends on accurate prediction of the gene and exon sites. Further the databases may help to support detected splice variants, as the most common variants are already listed in the databases. In this thesis the databases are also used for validation of splicing predictions (see Section 4.4).

The NAR database issue lists 24 databases for splicing [101]. Here the focus is on 17 databases for human or mouse using EST and mRNA sequences for splicing prediction. Excluded are special web services or databases which are not available at the time of writing or have not been updated for more than 6 years. An overview of the discussed databases is given in Table 2.2 with an associative visualisation in Figure 2.4.

Genome-wide analyses of alternative splicing are mainly based on publicly available sequence databases such as GenBank, UniGene, dbEST. Some databases also use protein sequences from RefSeq and Swiss-Prot/TrEMBL. EuSplice uses EST/mRNA data only for the validation of splicing events predicted by protein sequences. An exception is the H-DBAS database which relies solely on the H-Invitational project data [143], using manually annotated cDNA sequences. Most of the databases align the sequences to a genomic reference: NCBI, UCSC or Ensembl genome build.

Some early databases do not base on a genomic reference but align the EST and mRNA sequences with each other, e.g. use UniGene clusters as reference (STACK, PALS db, Xpro, EuSplice, EASED). Predictive methods based on EST and mRNA comparison have limited power, since they do not use information in the intronic part of the genome. Indeed, the splicing process is controlled by specific sequence motifs in the DNA flanking most of the intron sequences. These motifs, surrounded by a longer conserved consensus, provide valuable information for the location of splice sites through the alignment of EST and genomic sequences. Due to the above reasons, algorithms based on EST – genome pairwise comparison have provided more reliable tools for the detection of splice sites.

An in-depth discussion of the methods and associated problems is available in a review of Bonizzoni et al. [47]. Aligning the EST/mRNA sequences to the genome mostly algorithms like BLAST, BLAT or sim4 are used [23, 163, 96]. sim4 provides some quality indication for the alignment. Since these algorithms are very general several more specified methods have been developed and are often the basis for a database (ASPicDB, ECgene). The result of the alignment may now be used for two tasks: (1) deriving transcripts for genes taking alternative splicing into account as well as (2) exon and splice site definition. Some of the challenges described in the Subsection above are tackled with multiple alignment algorithms (ASPIC, ASAP) loosing some features like sensitivity scores or special sequences. The EST/mRNA sequences may also be assembled to full-length transcripts. The goal is to minimise the number of predicted transcripts that do not occur in nature. PALS db takes the longest mRNA sequence in each UniGene cluster as the reference sequence, which is aligned with ESTs and mRNA sequences in the same cluster to predict alternative splicing sites. Recently, several studies have suggested graphical methods to identify gene structure [217]. Splice graphs are constructed

Name	Organism	Source	Ref.	Alignment	Statistics (human)	Cit.
AEdb	various	literature	Ensembl	-	1609 RefSeq transcripts, 2940 AS events	[274]
Ensembl	human, mouse + diverse organisms	EST, mRNA, protein (dbEST, UniGene, InterPro)	NCBI	BLAST, Exonerate, EST_Genome, genscan	21 541 genes, 48 400 transcripts	[38, 138]
ASTD	human, mouse + 6 animals	EST/mRNA (GenBank); literature	Ensembl	AltSplice, manual	16 215 genes, 61 880 transcripts	[181, 273, 296]
ASAP	human, mouse + 13 animals	EST, mRNA (UniGene, GenBank, Entrez, RefSeq)	UCSC	BLAST	11 717 genes, 89 078 AS relations	[182, 165]
ASPicDB	human	EST, mRNA (RefSeq, UniGene)	NCBI	ASPic	18 442 genes, 229 123 transcripts	[56]
ECgene	human, mouse + 7 animals	EST, mRNA (dbEST, GenBank, RefSeq)	UCSC	ECgene (BLAT, sim4)	14 166 genes, 26 661 transcripts	[167, 166, 185]
SpliceNest	human, mouse	EST (UniGene)	Ensembl	sim4	33 270 EST cluster	[72, 121]
STACK	human	EST/mRNA (GenBank)	-	d2_cluster, PHRAP	270 515 cluster, 850 835 singletons	[60]
ProSplicer	human	proteins, EST, mRNA (Swiss-Prot, TrEMBL, UniGene, dbEST)	Ensembl	BLAST, sim4	21 786 genes	[135]
Hollywood	human, mouse	EST (GenBank, dbEST)	Ensembl	Genoa	∼22 200 genes, ∼79 000 cDNAs	[129]
H-DBAS	human	cDNA (H-Invitational)	UCSC	BLASTN, BLAT, EST2GENOME, manual	11 744 loci, 38 664 AS patterns	[293, 292]
TISA	human, mouse	EST, mRNA (dbEST, GenBank)	NCBI	BLAT, sim4	26 143 genes, 402 995 transcripts	[217]
PALS db	human, mouse	EST/mRNA (UniGene)	-	BLAST	33 111 UniGene cluster	[137]
Xpro	human + several eukaryotes	EST, mRNA (dbEST, GenBank)	-	BLAT	20 233 genes, 72 436 introns	[112]
FAST-DB	human	transcripts, EST/mRNA (UCSC, UniGene)	Ensembl	BLAST, sim4	12 538 genes, 201 245 exons	[113]
EuSplice	human, mouse + 21 organisms	DNA contig, mRNA, protein (RefSeq)	NCBI	none (by RefSeq)	23 678 genes, 95 822 AS events	[34]
EASED	human, mouse + 6 animals	EST/mRNA (GenBank)	Ensembl	BLAST	22 980 genes, 27 628 transcripts	[236]
MAASE	human, mouse	literature, EST, mRNA	UCSC	BLAT, sim4, manual annotation	1007 genes, 2217 AS events	[326]

Table 2.2: Alternative splicing databases. Overview about eukaryotic alternative splicing databases. Abbrv.: Ref., genomic reference; Cit., citation; AS, alternative splicing.

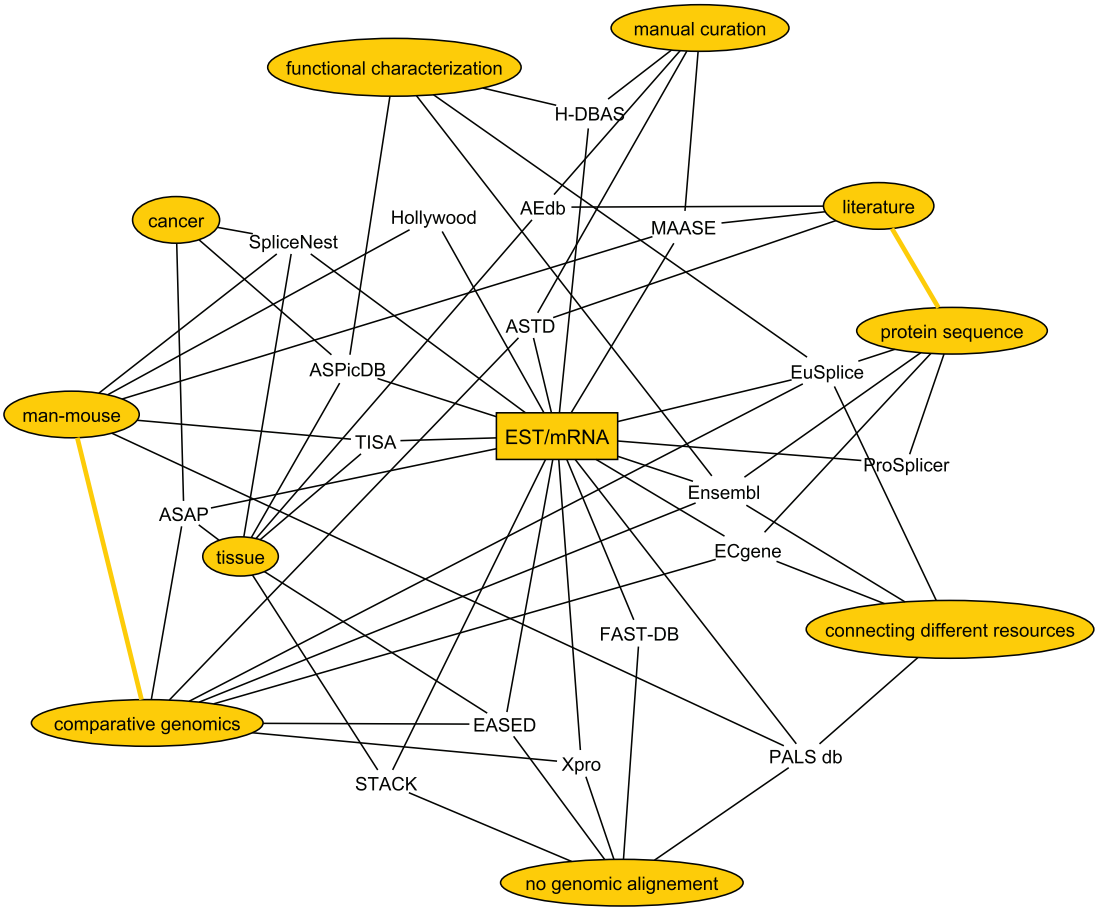


Figure 2.4: Relations of the alternative splicing databases. Alternative splicing databases assess the the same resources for different aims and use.

in which exons and introns constitute nodes and edges, respectively. However, generating all possible isoforms by graph traversal can produce many false transcripts. Many singular splicing examples are dispersed in the literature. Two databases try to gather a thimble of this knowledge by manual curation (AEdb, MAASE). Annotation of new results together with the publication will be crucial [133]. The era of high-throughput data has to be accompanied by in depth assays.

Only a few of the methods care about validation. For example EuSplice derives its prediction from protein sequences but uses EST/mRNA sequences for validation, similar for Xpro. SpliceNest is the only database, which performed PCR validation for their predictions [119, 120]. Comparing the databases with each other the results are often strongly discordant due to differences in the input, the genomic reference, the algorithms, alignment filtering and stringency as pointed out in [47]. A completely different approach is used in AEdb by manually collecting alternative splicing events from literature [274]. Therefore it is a consequently validation based database. Meanwhile it is integrated in

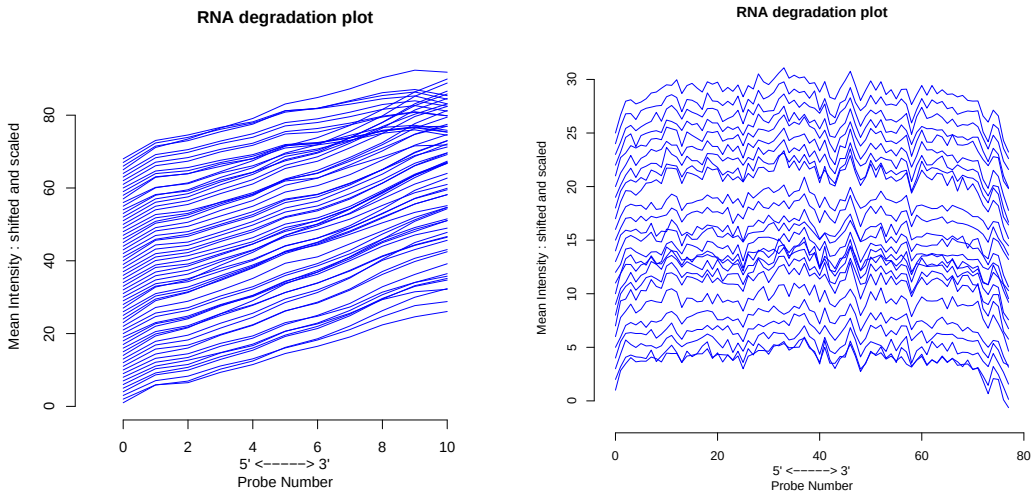


Figure 2.5: RNA degradation plots. RNA degradation plots are a quality control attribute for assessing microarrays. Due to the whole-transcript random primer protocol degradation now is rather constant. **Left:** RNA degradation in 3' gene expression arrays. Probes are chosen from the 3' end of the gene (see Section 3.2). **Right:** RNA degradation in exon arrays. Probes are dispersed over all exons (see Section 3.3).

ASTD [273].

Large differences are also in the presentation of the results. Four databases try to relate the genes and transcripts to different resources by connecting respective databases (Ensembl, ECgene, EuSplice, PALS db). Especially a functional characterisation of the genes or splice variants is provided by Ensembl, ASPicDB, H-DBAS and EuSplice. The input database may be stratified for sequences related to tissues or diseases. Taking this information into account some databases filter their results for tissue splicing or similar (SpliceNest, STACK, ASAP, ASPicDB, TISA, EASED). Results may also be compared between species. Five databases confine to human - mouse orthology (SpliceNest, Hollywood, TISA, PALS db, MAASE). More databases incorporate different genomes partially with the goal to study the evolution of splicing (Ensembl, ASTD, ASAP, ECgene, EUSplice, EASED, Xpro). Xpro is a comprehensive analysis of the splice sites themselves. MAASE focuses specifically on results for microarray design.

Focussing on genome-wide splicing prediction, a number of helpful studies is aside. The following selection are data sets which build on the above databases for splicing analyses with different research questions. T-STAG for example is a thorough tissue and cancer study relying on the SpliceNest database [118, 72]. HS3D is a selection of splice events for the training of machine learning algorithms [235]. SpliceInfo relates alternative splicing to RNA secondary structures [136]. ProSAS relates alternative splicing to changes in protein structures [39]. ASG provides a gallery of graph visualisations for alternative splicing [187].

In summary the databases cover a wide range of alternative splicing effects. However

Species	Topic	Cit.
human	transcript spike-in on HeLa cells, latin square design	[5]
human	colon cancer	[104]
human	11 tissues, muscle enrichment	[61, 80]

Table 2.3: Affymetrix Exon Array data sets. The three data sets are used in the validation of splicing prediction methods, see Section 4.4.

there is low consensus information over all resources. To proceed with a conservative selection of tissue splicing events, the AEdb is chosen as true positive set for the method evaluation in Section 4.4. Spliced exons in the AEdb are annotated to Ensembl, which is also the database used for annotation in the microarray pipelines in Chapter 3.

2.2.2 Microarrays

One recent development has been the transition of microarray studies of alternative splicing from the prototype stage to a tool for large-scale analysis of alternative splicing [184, 48] (see Table 2.4). Microarray techniques facilitate the detection of regulated splicing in large candidate pools and the identification of regulated splicing in biological contexts [183]. The design of most of the current microarrays has one basic flaw: the majority of the probes are not specific for different products from the same gene. The construction of splicing arrays requires sequence information uniquely associated with specific isoforms [48].

Spotted oligonucleotide microarrays employing probes designed to detect unprocessed and processed RNA have been used to monitor pre-mRNA splicing in yeast [62]. Conventional microarray-based approaches utilising oligonucleotides have been used for monitoring alternative splicing in mammalian cells [149, 53, 156, 134, 308, 180, 229]. The most extensive use of the latter approach was the application of „exon-junction" microarrays for the discovery of exon skipping events in human tissues and cell lines in Johnson et al. [149]. The authors used custom microarrays containing oligonucleotide probes complementary to mapped exon-exon junction sequences in RefSeq genes for the main purpose of discovering new alternative splicing events in human transcripts. Three data sets used Affymetrix custom arrays introducing the short oligonucleotide platform for alternative splicing [156, 308, 134].

The traditional labelling protocols necessitated the design of the probes toward the 3' end of the transcript in order to optimise the match of the labelled targets with the probes and thus were not all suited to detect alternative splicing events [48]. For the introduction of exon arrays monitoring the expression on the full length of a transcript Affymetrix had to develop and introduce the Whole Transcript Sense Target Labelling Assay. Using a random priming strategy, in combination with *in vitro* transcription-based

Species	Size	Probe placement	Manufacturer	Cit.
human, chimpanzee	2647 genes	exon body and splice junction	Agilent	[53]
human	316 genes	exon body and splice junction	Agilent	[180]
mouse	2647 genes	exon body and splice junction	Agilent	[229]
human	990 genes	tiling array	Affymetrix	[156]
human	10 000 genes	splice junction	Agilent	[149]
human	21 genes	exon body and splice junction	Affymetrix	[308]
yeast	933 genes	exon, intron and splice junction	cDNA spotted array	[62]
human	23 genes	exon body	fiber-optic microarray	[317]
rat	1600 genes	exon body	Affymetrix	[134]
human	364 genes	splice junction	fiber-optic microarray	[324]
human	86 genes	exon body and splice junction	Geniom OneR	[243]
human	17 939	exon body and splice junction	Agilent	[55]

Table 2.4: Alternative splicing microarray data sets. Low coverage microarrays have already been used for different splicing analyses.

Species	Coverage	Manufacturer	Cit.
human	10 tissues	Illumina	[306]
human	6 tissues	Illumina	[228]
human	2 cell lines	Illumina	[289]
mouse	1 tissue	454	[190]

Table 2.5: mRNA-Seq data sets. First data sets explore splicing by deep sequencing of the transcriptome.

linear amplification and a novel end-point fragmentation and labelling assay scheme, it provides a robust method for target labelling (see Figure 2.5). Exon Arrays meanwhile are used in different experimental settings, where data sets used in the thesis are listed in Table 2.3.

The current design of the Affymetrix Exon Arrays facilitates the analysis of exon skipping events. This type of splicing events covers more than 50% of all known splicing events [273]. On the other hand some splicing events are missed due to the selection of the probe position in the genome. Intron retention is not recognised at all. Changes in the splice boundaries are identifiable depending on the exact probe positions.

2.2.3 RNA-Seq

The introduction of 2^{nd} generation sequencing technologies opened new doors into the field of genomic sequencing. As understanding of these technologies becomes more widespread and new tools are being developed, so are new innovative ways of applying these technologies being created [197].

Given the low requirements of the new technology for a nucleotide sequence product, together with its deep coverage and base-scale resolution, its use has expanded to the field of transcriptomics [309]. However quantitative studies still need high number of replications [198].

Although a field recently opened three genome-wide sequencing studies are listed [208] in Table 2.5. Deep sequencing of cDNA from multiple human tissue types revealed thousands of new splicing junctions [306, 228]. Both studies conclude that approximately 92-98% of human multi-exon genes are subject to alternative splicing. That means at least 86% of all human genes. Wang et al. [306] identified over 22 000 tissue-specific alternative transcript events. These events cover alternative splicing, alternative polyadenylation and alternative promoter usage. Correlation between the tissue-specific patterns of alternative splicing and alternative polyadenylation events led the authors to the hypothesis, that these mechanisms might be coregulated.

3 Computational Analysis of Affymetrix Arrays

Comparison of results between different data sets and meta-analysis requires a standardised and controlled data analysis processing both for alternative splicing and differential expression [147]. Methods for differential expression analysis established for Affymetrix 3' gene expression arrays have been adapted to the specific design of the Exon Arrays. The alternative splicing pipeline consists of two branches, differential expression analysis and alternative splicing analysis. No method evaluation or accepted procedures are available for splicing evaluation. Thus, the alternative splicing analysis is developed in Chapter 4 and then implemented as a module in the pipeline.

Guidelines for the pipelines are implementation in R/BioConductor [238, 108], internal handling on Ensembl genes or exons with gene-wise analysis [38], modular design as well as division of complex experimental settings into different test cases. The application of statistics depends on the experimental technology and the analysis objective [304]. Thus, the particular design of the Affymetrix 3' gene expression and Exon Array is introduced [97, 194]. The chip design allows to reinterpret the probe assignment to genes and transcripts in the light of the advancement of the genome sequence.

The processing of the arrays follows two fundamental principles: First, only biologically motivated corrections on the data are allowed with statistical models. Second, comparability between experiments is essential. The processing is highlighted with the steps quality control, test case determination, preprocessing and data evaluation [65][1]. Where research on 3' arrays is settled, the analysis of exon microarrays has posed new challenges to the computational analysis like data normalisation and presence tag calculation. Probe binding affinity is corrected by GC content of the probe sequences and intensity distributions are adjusted by quantile normalisation. Data evaluation establishes an array-wide gene analysis followed by the isolation of a set of alternatively spliced genes as well as a set of differentially expressed genes. It follows the gene set evaluation with over-representation analysis and group testing on a diverse set of functional resources like pathway databases, transcription factor targets, drug targets and tissue expression.

[1] Parts of this Chapter appear in the Handbook of Research on Systems Biology Applications in Medicine edited by Dr. Andriani Daskalaki [65]; Copyright 2009, IGI Global, www.igi-global.com. Posted by permission of the publisher.

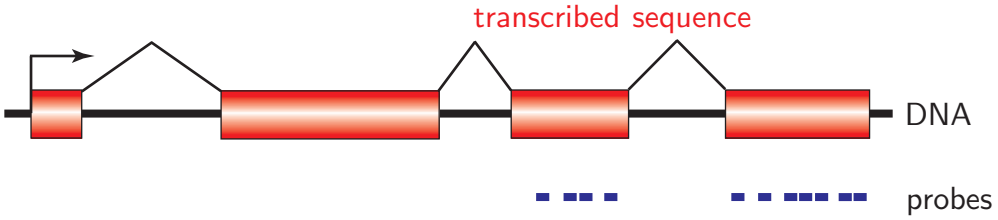


Figure 3.1: Probe placement 3' gene expression array. Probe sequences are selected from the transcribed regions of the coding sequence, preferably near the 3' end of the gene sequence.

3.1 Design of the GeneChip array

The basic element of the chip are probes of 25 basepair length spotted on quartz wafer slides with a photolitographic method [97]. In an experiment dye-labelled RNA from the sample under study is injected on the slide. The hybridisation depends non-linearly on the amount of transcripts in the sample [123, 213, 76]. The comparison between different samples is done by using several chips with multiple replicates per biological condition. The chip with the hybridised solution is scanned to the absorption spectrum of the dye. Analysis starts by exploiting the scanner image. From this image approximate hybridisation values are inferred for each probe. An exhaustive description of the technology is available in Dalma-Weiszhausz et al. [78] of by manufacturers manuals [10, 7, 194].

3.1.1 The 3' gene expression array

In the chip design, every probe is spotted with its perfect match probe (PM) and the so-called mismatch probe (MM). The PM have complete complementarity to their target sequence. In the MM sequence the 13th nucleotide is altered to its basepair complement. The idea is, that the MM measures the background expression. The PM signal than is composed by the background expression plus the gene specific expression.

The expression of a gene is measured by several probes with sequences unique to the respective transcript sequence. Reference transcript sequences are assembled from public sources like UniGene, GenBank, dbEST or RefSeq. A number of such probes collected in probe sets stands for independent measurements of the amount of transcripts for the gene. The number of probes in a probe set varies between chip platforms. For example in the popular mouse 430 2.0 array there are eleven probes in one probe set. Samples are prepared with the *in vitro* transcriptase protocol and expression is more independent from RNA degradation at the 3' end where the probe sequences are selected (see Figure 3.1).

3.1.2 The exon array

The exon arrays are an advancement of the 3' gene expression array but differ in two major points from the above described design:

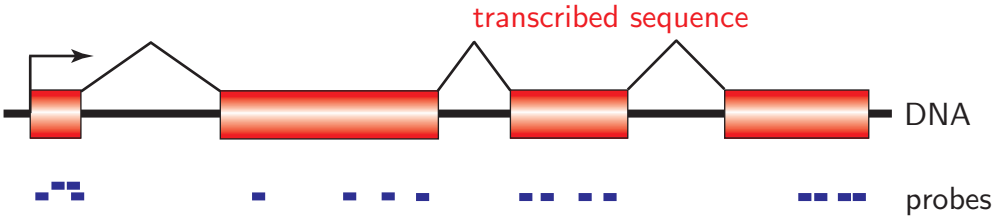


Figure 3.2: Probe placement exon array. Probe sequences are selected from the transcribed regions of the coding sequence, preferably four on each exon.

- Probes are distributed over all the exons;
- MM are replaced by a selected set of control sequences.

The first difference concerns the probe placement. With random primers in the whole-transcript protocol degradation is constant over the transcript and probes are not pressed to the 3' end. Instead of a probe set targeting one gene, now a probe set measures the expression of one exon of a gene. With the dispersion of the probe sets over all exons the hybridisation returns a more fine-grained picture of the gene expression (see Figure 3.2) [18, 13].

Due to a better sensitivity of the probes the number of probes is decreased in a probe set. Where the mouse 430 2.0 array has eleven probes in a probe set, the mouse exon array has four probes, sometimes less in a probe set. Still the exon arrays have a better coverage of probes per gene. In the mean a gene has about 13 exons, with one probe set per exon an average of 52 probes hit a gene.

The second difference are the drop of the MM. In the classic design, half of an array is reserved, one MM per PM. The amount of control probes is now reduced to a selected set of non-coding probe sequences. The control sequences are either chosen from non-coding regions of the genome (genomic controls) or randomly generated probe sequences not hitting any genomic sequence (antigenomic controls). The control sequences are selected for varying GC content with the goal of more than 1000 control sequences per possible GC number. On the mouse exon array are 20 744 genomic control and 16 943 antigenomic control probes.

Altogether the coverage of the genome by the arrays increases, continuing the whole-genome strategy. The mouse exon array measures 22 798 Ensembl genes, where the mouse 430A 2.0 measures 15 695 genes using the alternative assignments described in the next Subsection.

3.1.3 Alternative probe-gene assignments

Probes are assigned to probe sets. Probe sets are annotated to genes and transcripts in diverse databases and current annotation can be retrieved from the NetAffx homepage [4]. Since often several probe sets are mapped to one gene, the annotation introduces ambiguity into expression results. A probe that not completely hits its target gene

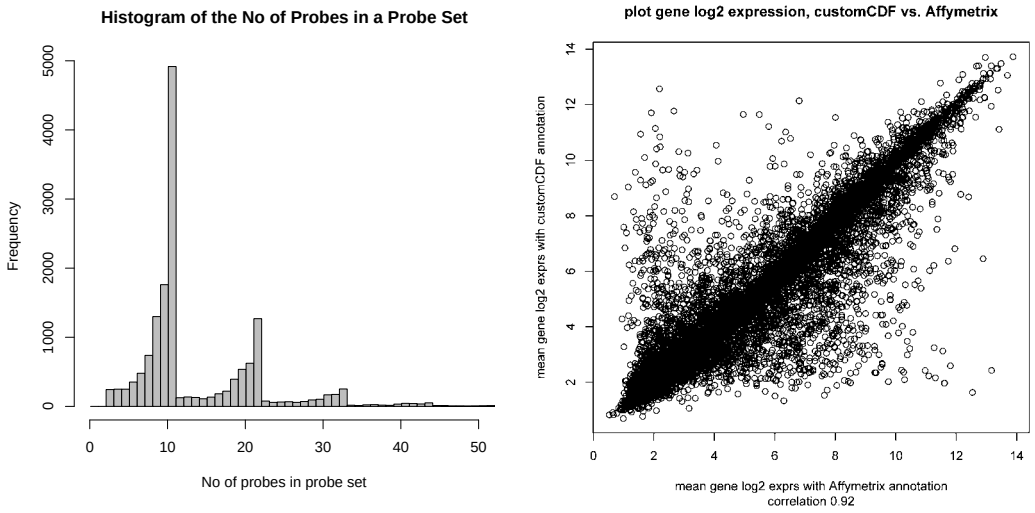


Figure 3.3: customCDF results vs. original annotation. Left: Histogram of the number of probe sets in the customCDF. **Right:** Probe set mean expression values plotted for customCDF assignment vs. Affymetrix original assignment.

sequence is supposed to introduce noise. The advancement of the sequence databases leads to altered gene sequence and extensive libraries for single nucleotide polymorphisms (SNP) are available. A probe can become obsolete because the probe sequence is not contained in the gene sequence anymore or its specificity is reduced by a SNP. It has come up, that probe to probe set annotation does not have to be fix but the probes can be reattached.

Dai et al. [77] – customCDF – presented new assignments and their purpose is three-fold:

- An injective mapping is possible from probes to genes of a specific sequence database;
- Probes not completely aligning to a gene are skipped;
- Probes aligning to a SNP position are skipped.

Affymetrix original annotation vs. customCDF

Because of differences between the original annotation and customCDF there are different results in the interpretation of the data. Since, in the customCDF, there is no fix number of probes in the probe set anymore statistical criteria may depend on the probe set size. In the following the differences between the assignments are elucidated on the Affymetrix mouse 430A 2.0 array.

The original probe sets are assigned to Ensembl genes in NetAffx resulting in 16 466 Ensembl genes (version 28 from 12.03.2009). The customCDF results in 15 768 genes (version 10 on Ensembl 46). From originally 496 468 probes on the array with the customCDF remain less than a half, 227 156 probes, in the new assignment. Probe set size

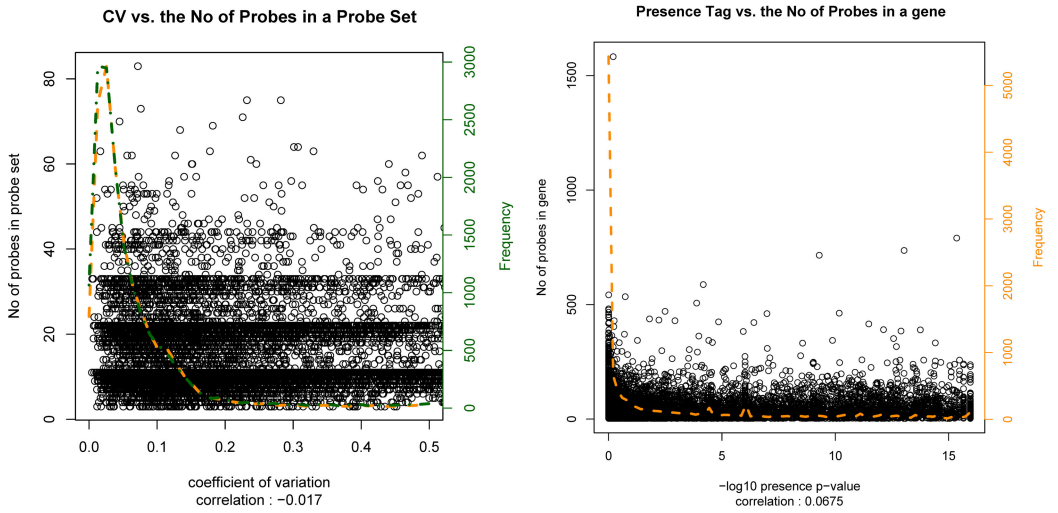


Figure 3.4: Coefficient of variations and presence tags. Left: In black is a scatterplot of the coefficient of variation vs. probe set size. A bigger probe set size leads to a smaller coefficient of variation. In dark orange is the histogram of coefficient of variations for the manufacturers annotation and in dark green the histogram of coefficient of variations for customCDF. **Right:** Scatterplot of -log10(presence p-values) vs. probe set size in customCDF annotation for exon arrays; In dashed orange is the histogram for the presence tags.

strongly vary around the original number of probes per probe set, in fact multiples of 11.

To compare the computations on the two assignments the intersection of 14 911 Ensembl genes is used. Calculation differences between the assignments are low. Expression log2-transformed values correlate with 0.92 (see Figure 3.3[2]). The fold-changes between treatment and control have the same distribution and a correlation coefficient of 0.95. The coefficient of variation is lower for customCDF with 0.269 compared to 0.278 (see Figure 3.4). The set of present genes has 8115 genes with NetAffx and 8393 with customCDF ending in an intersection of 7620 genes, a Jaccard index of 0.86 (the ratio $\frac{|A \cap B|}{|A \cup B|}$).

Since the probe set size is not constant anymore, calculations could depend on the number of probes. Due to a robust computation expression values and presence tags have correlation values near zero (see 3.4). Often values scatter around multiples of the original probe set size.

Differentially expressed genes are identified running the pipeline from the next Section 3.2 on both assignments. There are 996 differentially expressed genes on the original assignment and 964 differentially expressed genes on alternative assignment with an overlap of two-third of each differential expression set or a Jaccard index of 0.55.

In the exon arrays the number of probes depends on the number of exons in the gene and is thus highly variable. Both assignments have variable probe set size. Expression values

[2]If not differently noted for 3' gene expression array evaluations and images following versions are used: R 2.6.0, BioC 2.1.0, customCDF 10, Ensembl 46, Affymetrix mouse 430A 2.0 array. Original data is from a type-2 diabetes mellitus experiment introduced in Section 5.2.

and coefficient of variations show similar behaviour as above in the 3' gene expression comparison. A scatterplot of exon array presence tags vs. the higher and more variable number of probes per gene is provided in Figure 3.4 and has a correlation value of 0.068. Finally customCDF 11 covers 21 994 Ensembl genes vs. 27 006 genes in NetAffx 21. The advancement of the probe - gene assignments is continued by groups assessing the cross-hybridisation potential of the probes [160, 315].

3.2 Differential expression (DE) with 3' gene expression arrays

The pipeline presented is composed by standard tools and covers considerations elucidated in Clevert and Rasche [65]. The DE pipeline is applied in the meta-analysis for type-2 diabetes mellitus in Section 5.2 and is used as a standard operating procedure for various projects [86, 81, 240]. The analysis process is composed in R/BioC [238, 108]. As the different tools are available as BioC packages it is straightforward to keep the whole implementation of the pipeline in R. For a general discussion of differential expression analysis readers may refer to a number of excellent reviews [275, 173, 6, 126, 140, 145].

The pipeline workflow consists of the standard steps for processing microarrays and the steps are described in separate subsections (see Figure 3.5):

- quality control of raw data
- determine test cases
- preprocessing
- evaluation of the data and differential expression filter
- gene set evaluation: over-representation and group testing

The process is semiautomatic in the sense that only the test cases are manually specified. The combination of different tools and methods with several input and output formats requires a stringent handling of the identifier. The whole processing script depends exclusively on Ensembl genes leading to a massive reduction in the complexity of the processing and the script. Ensembl identifier are particularly simple to handle due to the BioMart interface and the biomaRt package [38, 162, 88]. Any resources, e.g. the gene sets, are mapped to Ensembl genes before evaluation.

In complex experiments it is not possible or advisable to preprocess all of the chips together. Thus the pipeline follows the guideline only to join for preprocessing what is later evaluated in conjunction. Thus experiments are divided in test cases and processing is looped for each test case.

For clarity reasons I follow some conventions. Hybridisation values are the raw probe signal values derived from the scanner image, i.e. the CEL file level. Intensity is the normalised probe signal and expression is the normalised and summarised value on the exon or gene level.

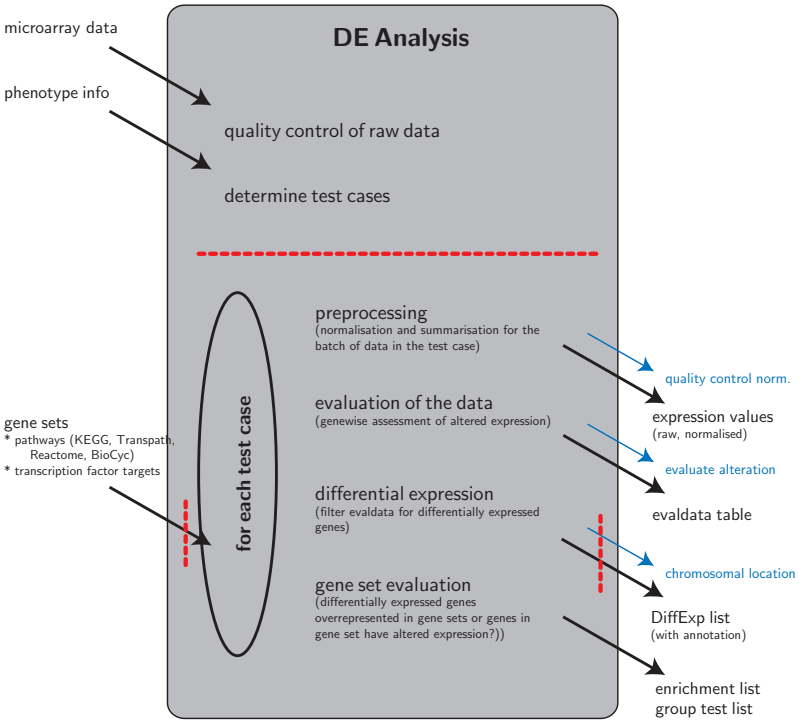


Figure 3.5: The 3' gene expression DE pipeline. A scheme for the analysis workflow. After manual determination of the test cases the data is processed automatically. Interfaces which need a mapping of Ensembl genes to probes or different gene identifier are marked with a red dotted line.

3.2.1 Experimental setup

Probe intensities and gene expressions show a variability which cannot always be attributed to specific biological or technical effects and thus is considered to be noise. Not only data preprocessing but also experimental planning has to account for this noise. A solid statistical analysis requires replicates. In technical replicates the same sample is used on several chips resulting in low noise levels. In biological replicates different samples are prepared from the same biological condition, e.g. different patients or animals are hybridised each on a single chip. The following empirical numbers of biological replicates are recommended by the manufacturer:

- Cell culture: 2-3 replicates;
- Animal system: 4-5 replicates;
- Human system: 5-6 replicates

Some problems and challenges accompany the current application of microarrays. A set of guidelines called Minimum Information about a Microarray Experiment - MIAME - is introduced by the International Microarray Gene Expression Data Society [50, 170, 224]. The MicroArray Quality Control (MAQC) project assesses inter- and intraplatform reproducibility of gene expression measurements [262]. The MAQC was initiated to address concerns about noise and preprocessing problems and more issues. The study is an important first step pushing microarrays toward clinical and regulatory settings. The experimental setup is described in Shi et al. [262]. Microarray products from different manufacturers are compared. Affymetrix products are presented with a very high reproducibility in- and across test sites with low variance in measurements.

Due to the noisy data microarray results always have to be verified. In the lab this is normally done with complementary RT-PCR experiments for selected genes. Without experimental validations statistical means for the consistency with other published data could serve, although this consistency is often low. The amount of verification is alleviated by the use of statistics in the analysis of the hybridisation results.

3.2.2 Quality control of raw data

A variety of errors from sample generation to scanning can lead to erroneous hybridisations. Most of these errors are detectable by checks within and between experimental arrays. All checks return images for a direct visual impression. A first test for the consistency is to correlate the hybridisations between the different chips. A corresponding heatmap facilitates the view, see Figure 3.6, for example to see if the different phenotypes cluster together [232].

Three images compare the chips by cumulative statistics. The histogram or boxplots for the hybridisation values and the RNA degradation assessed by controls on the chip. Often for the images it is helpful to use logarithmic data.

From the position of the probes on the chip and the hybridisation values visualisation of the chips hybridisation can be reconstructed [105]. This idea is continued by using a linear model for the expression of a probe set and highlight the positive and negative residuals

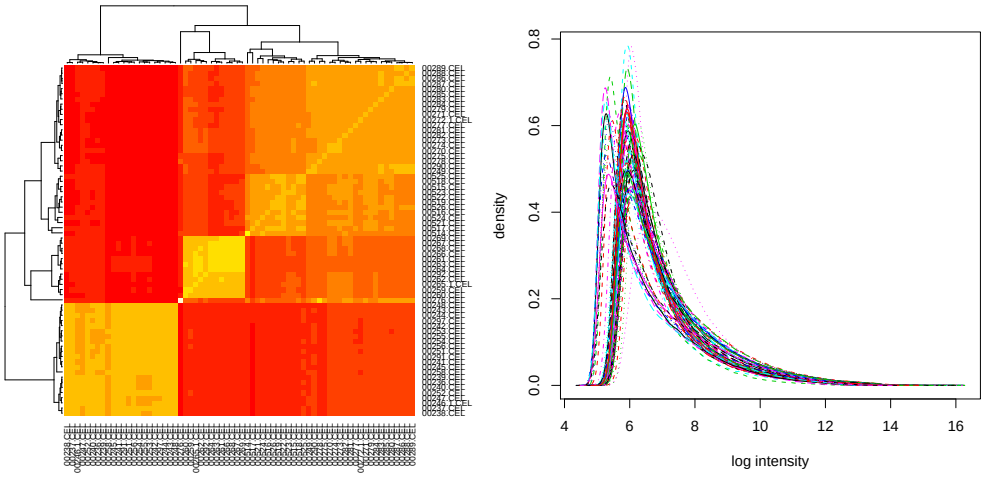


Figure 3.6: Intensity heatmap and histogram. Left: Heatmap of intensity correlations between chips. **Right:** Overlay of intensity histograms of the hybridisation intensities for different chips.

of the probes with red and blue [43, 42, 44]. Particular flaws during the hybridisation are identified. GeneChip arrays are very robust against such flaws, due to their design with several probes in a probe set dispersed over the whole chip.

The last images relate probes between chips. The hybridisation values of one chip are compared to the probe-wise median over all chips for (a) normal scatterplot, (b) quantile-quantile-plot and (c) MA-plot. The first plot is provided with linear and logarithmic values. The latter plots the difference to the mean of the logarithmic values.

3.2.3 Determine test cases

The DE pipeline is developed for case control studies where two groups of samples are compared with each other. Without loss of generality (W.l.o.g.) these groups are denoted as treatment and control. In terms of the wet lab experiment the treatment samples are any sort of modified or similar samples of interest, some special differentiation or cell types, e.g. disease samples. For the comparison one assumes some normal state of the tissue whereof the control samples are prepared.

Preprocessing for each test case may result in different expression values for the same gene and chip in different test cases. This is rarely the case but hinders the comparability of expression values. But the chips cannot be normalised in one batch, because an experiment may comprise several tissues. The processing would need another step to identify normalisation groups. This implementation follows the principle only to normalise what is compared. This has the advantage that other chip intensities do not disturb the expression results under study. The following steps of the pipeline are processed in a loop for every test case.

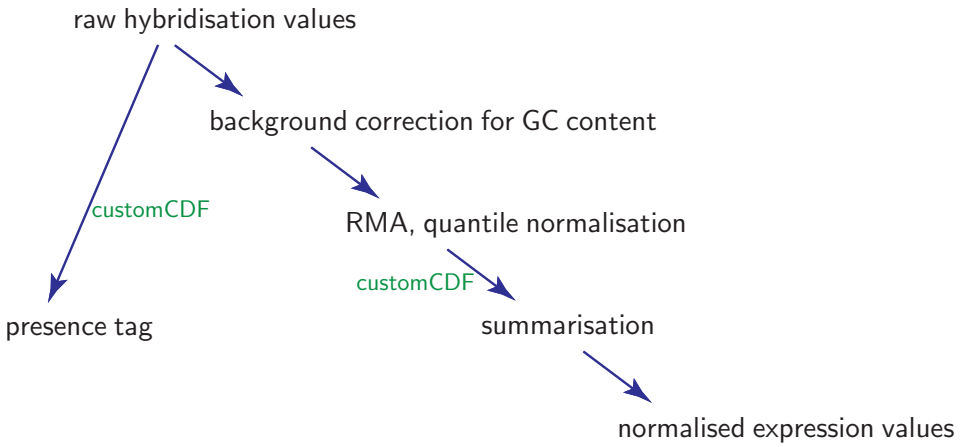


Figure 3.7: Steps in the preprocessing procedure. The right branch is the GC-RMA method using a three step procedure to calculate normalised expression values from raw hybridisation values. The left branch is the fourth step for calculating detection p-values for the genes. In both branches the customCDF probe-gene assignment is applied which is designated in the figure in green.

3.2.4 Preprocessing

The composition of the gene signal is not yet completely understood, although several groups develop mathematical models [189]. Such models are applied to correct the hybridisations. Preprocessing has to account for three major disturbing factors, the background signal, the probe binding affinity and the variance of the measurements resulting in noisy data. After several years of research for the optimal preprocessing a variety of methods are available for this task. The GC-RMA [312, 311, 313] accounts for the GC content of the probe sequences. The preprocessing is here elucidated as a four step procedure (see Figure 3.7):

1. Intrachip normalisation: The background correction removes unspecific intensities from the scanner images;
2. Interchip normalisation: Reduce non-biological differences between chips;
3. Summarisation: Probe intensities are combined into a single probe set expression value.
4. Presence tag: A detection p-value is computed from the probe hybridisations for every probe set.

Errors introduced in the preprocessing may corrupt further analysis. With the large number of genes on the array there may be low correlation between the samples when using few arrays. This has to be addressed in the experimental planning, see Subsection 3.2.1.

Intrachip normalisation

The hybridisations are affected by technical artefacts from protocol and image scanning, chemical background and optical background respectively. The optical background derives from the technical range of the scanning device eventually supplemented by the over-shining of the neighbouring spots not corrected by image analysis. The chemical background is explained by the probe hybridisation consisting of gene specific binding and unspecific binding. The unspecific binding, or cross hybridisation, comes from different RNA snippets, e.g. RNA from other genes. Unspecific binding has shorter binding times at the probes. By hybridisation over time unspecific binding faster reaches its equilibrium of hybridising and dissolving at the probe as gene specific binding does [76]. Thus the equilibrium is lower than for gene specific binding. MM were originally introduced to account for unspecific binding. However the MM intensities contain more gene specific binding than expected [325, 46, 146, 144].

The gene specific binding is target of all normalisation methods. Background signal and binding affinity differ from probe to probe and thus average estimates over all probes had little success. More success have models accounting for the nucleotide content of the probe sequence for estimating disturbing factors [123, 212, 263]. A higher GC content - higher number of G or C nucleotides in the probe sequence - is associated with a higher binding affinity due to three instead of two covalent bindings for a single nucleotide. The higher affinity leads to higher hybridisation values and increased variance. Position-specific probe affinity assigns every nucleotide an affinity depending on the position in the probe sequence. Dinucleotide models estimate affinity by neighbouring nucleotides [325]. However the dinucleotide models do not add much predictive power [313, 212].

The GC-RMA method assumes optical noise and unspecific binding to be independent and proposes the following statistical model [313, 312, 311, 310]:

$$\eta_p = \iota_p + O + N_p, \tag{3.1}$$

where

- η_p is the measured hybridisation for probe p,
- ι_p is a quantity proportional to RNA expression - the quantity of interest for probe p,
- O is optical noise and
- N_p is unspecific binding for probe p.

In the following the description is focussed on perfect match intensities ι, if necessary corresponding mismatches are denoted with index MM. Due to ignorable variance the optical background O is treated to be constant estimated with the minimal hybridisation observed over the whole array with subtraction of 1 for avoiding negatives: $\hat{O} = \min(\min(\eta_{\mathrm{PM},p}), \min(\eta_{\mathrm{MM},p})) - 1$.

It is assumed that the logged unspecific binding $\log(N_p)$ follows a bivariate-normal distribution with mean μ_p and variance $\sigma_p^2 = \mathrm{var}(\log(N_p))$. The mean μ_p depends on the binding affinity α_p with a smooth function h: $\mu_p = h(\alpha_p)$. Probe binding affinities

α_p are defined in the next paragraph. The function h is estimated as a loess curve on $\log(\eta_{\text{MM}} - \hat{O})$ vs. α_{MM}, the binding affinities of the MM, resulting in $\hat{h}$. Consequently μ_p is estimated with $\hat{\mu}_p = \hat{h}(\alpha_p)$. With more than 100 000 probes enough data is available for precise estimations of $\hat{h}$ and σ.

The binding affinity model incorporates a modified variant of the position-specific model from Naef and Magnasco [212]. It is a sum of base effects:

$$\alpha_p = \sum_{j=1}^{25} \sum_{k \in \{A,T,G,C\}} a_{j,k} \cdot 1_{b_j=k}, \tag{3.2}$$

where

- $j = 1, ..., 25$ designates the position along the probe p,
- k indicates the base letter,
- b_j represents the base at position j,
- $1_{b_j=k}$ is an indicator function, that is $\begin{cases} 1, & j-\text{th base is of type } k \\ 0, & \text{else} \end{cases}$ and
- $a_{j,k}$ represents the contribution to affinity of base k in position j.

Originally $a_{j,k}$ is estimated with a polynomial of degree 3, in GC-RMA it is estimated from the array data with a spline with 5 degrees of freedom [313]. In sum 100 affinity contributions $a_{j,k}$ are to estimate. The authors of Wu et al. [313] propose a maximum likelihood estimation or empirical bayes estimation; Explanations are skipped as standard models are used without details. Probe sequences are available in R/BioC via the package matchprobes [139].

Interchip normalisation

A chip specific bias is introduced in the experiment by RNA extraction, pipetting, temperature fluctuations, hybridisation efficiency and more. Possible sources are discussed in Hochreiter et al. [127] in more detail. Normalisation is the step to account for this bias and reduce the unwanted effects between the chips. Here the normalisation is done by quantile normalisation as proposed and implemented in Bolstad et al. [46], similar to RMA from the affy package [144, 105].

The hybridisation distribution of the arrays is corrected by an empirical distribution determined by ranking the intensities for each array. Then at each rank the probe intensities are set to the mean of the intensity values over the arrays. All arrays now have the same intensity distribution but for different genes and different positions within probe sets. Changes are visible in Figure 3.8.

At this step in different methods the PM are corrected by the MM values. Unfortunately this leads to a higher variance in the results especially for low intensities. RMA, like most of the current methods, ignores the MM information. Affymetrix indirectly supports this aspect by retiring the MM in the exon arrays.

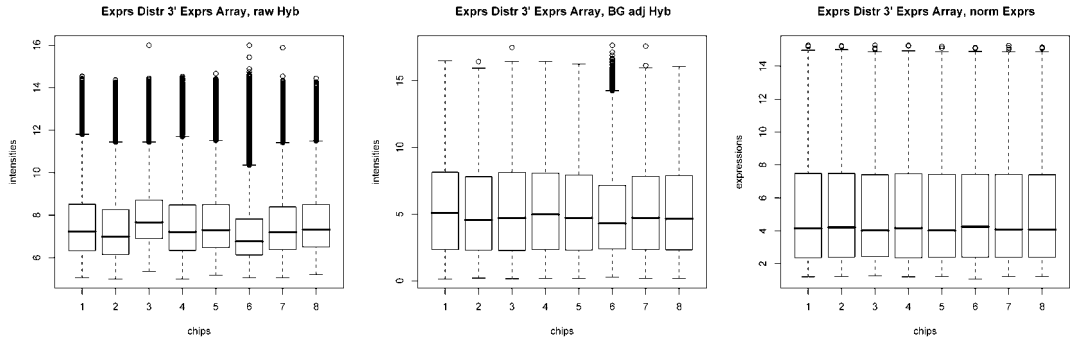


Figure 3.8: Boxplots for eight arrays during preprocessing. On the left hand side is the distribution of raw hybridisations for eight chips, in the centre distributions after background correction and on the right hand side expression distribution after preprocessing.

Summarisation

The summarisation is the last preprocessing step the corrected probe intensities are combined into a single probe set expression level. Integrated in RMA is the method medianpolish. Medianpolish is a multi-array method taking into account probe information across arrays. Examination of probe patterns show that the variability of probe intensities is lower across the arrays for a single probe than for probes in the same probe set. The medianpolish method proposes the following model [146]:

$$\log_2(\eta_{p,r}) = \mu + \alpha_p + \beta_r + \epsilon_{p,r}, \tag{3.3}$$

where

- $\eta_{p,r}$ are the hybridisations,
- μ is a baseline constant,
- α_p is a probe effect,
- β_r is an array effect and
- $\epsilon_{p,r}$ is a random error term.

The model is fitted robustly as a median decomposition with an algorithm from Tukey [302]. After model fitting the output is the gene expression $\Phi = \mu + \beta_r$.

Presence tag

In the analysis it is recommendable to filter for expressed, i.e. present, genes. Not expressed genes confuse the results because small changes in low intensities lead to high, unmotivated fold changes. The „detection p-value" is based on a comparison of raw PM hybridisations $\eta_{\mathrm{PM},p}$ to corresponding raw MM hybridisations $\eta_{\mathrm{MM},p}$ (see Figure 3.9) [16, 17, 7, 8, 9].

In a first step a Discrimination score is calculated for probe p:

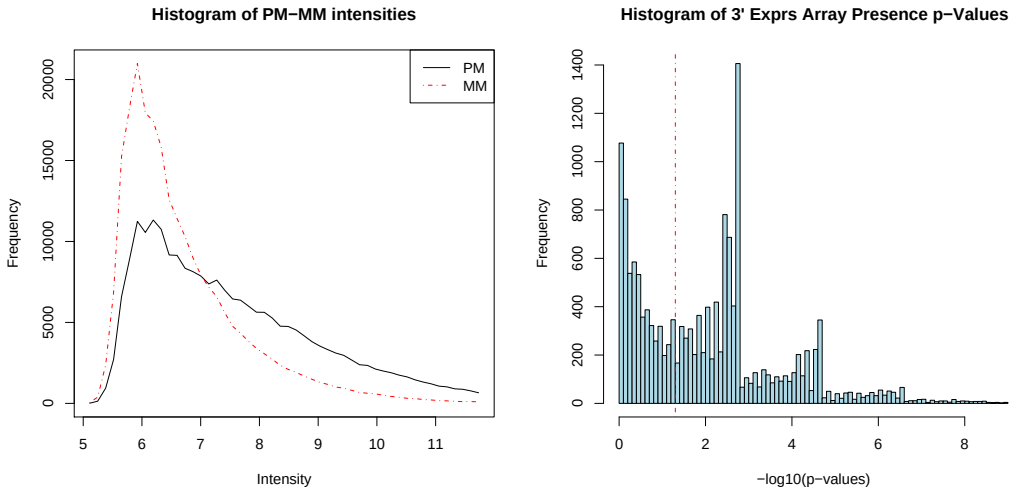


Figure 3.9: Distribution of probes and presence tags on 3' gene expression arrays. Left: Distribution of PM in black and MM in dotdashed red. **Right:** Distribution of presence tag p-values on -log10 scale. The dotdashed line represents the 0.05 threshold usually used for the detection.

$$R_p = \frac{\eta_{\mathrm{PM},p} - \eta_{\mathrm{MM},p}}{\eta_{\mathrm{PM},p} + \eta_{\mathrm{MM},p}} \tag{3.4}$$

The Discrimination score tends to 1 if the PM hybridisation $\eta_{\mathrm{PM},p}$ exceeds its MM hybridisation $\eta_{\mathrm{MM},p}$ and decreases if the intensities of the probe pair do not differ.

The second step is to compare the Discrimination scores of all probe pairs in the probe set to a predefined τ (the default is $\tau = 0.015$). The one-sided Wilcoxon signed rank test returns the p-value called the detection p-value. It is tested for the null hypothesis, that the discrimination scores are less or equal than τ:

$$H_0 : \mathrm{median}_{p\in\{\text{probe set}\}}(R_p) \leq \tau \tag{3.5}$$

The alternative hypothesis is the discrimination scores are greater than τ:
$H_A : \mathrm{median}_{p\in\{\text{probe set}\}}(R_p) > \tau$.

3.2.5 Evaluation of the data and differential expression filter

After the calculation of gene expression follows the gene-wise comparison of treatment expressions vs. control expression. The genes are assessed by the following criteria:

- presency
- variation
- alteration

The presency is evaluated with the detection p-value described in the preceding Subsection 3.2.4. Beside providing the mean of the treatment and control expressions the variation is analysed with the standard error of the mean and the coefficient of variation. The coefficient of variation is a mean-corrected standard deviation and facilitates a gene independent measure of variation. Finally the alteration is assessed with the fold change and statistical tests. The fold change, or ratio, indicates biological relevance of the observed expression change and is amended with the standard error of the fold change [169]. Statistical tests assess the significance of the change by combining the alteration and the variation. It may be a matter of taste to choose the t-test, Welch test or Wilcoxon test but the rank tests like Wilcoxon or permutation test require more replicates [128]. Ties are possible in the tests and exact calculations for rank tests are applied [132]. Specialised rank tests for the case of differential expression are still under investigation [51, 52, 130].

Due to the high number of statistical comparisons (∼8000-22 000 depending on the chip platform) often a multiple testing correction is applied for the statistical tests [275, 173, 126, 140]. Because most corrections analyse the distribution of p-values over the array results these corrections would be test case bounded. Evaluation results would not be comparable to other test cases and experiments. Since the pipeline is applied for evaluation with different pipeline outputs in Section 5.2 multiple testing is avoided in the above criteria in favour of a consequently gene-wise analysis. However for single experiments q-values are provided by Storey in the output [282, 281, 75].

Differential expression is filtered gene-wise by the three following criteria:

- The gene is expressed in at least one of two samples, treatment or control, with a p-value of at most 0.05;
- The ratio is minimally $4/3$ or maximally $3/4$ between treatment and control sample: $\frac{3}{4} \leq \frac{\Phi_t}{\Phi_c} \leq \frac{4}{3}$;
- If calculable, the Wilcoxon-test between the expressions for the two samples is significant with a p-value of at most 0.05.

The Wilcoxon-test is applied for settings with at least four replicates in both, treatment and control group. For at least three replicates the Welch test is applied. For less, the last criteria is skipped.

The filter for differential expression combines all the three evaluation criteria to decide about differential expression. Of course these criteria may be adjusted for the experimental setting. The list of differentially expressed genes may be very long, depending on the strength of the difference between the treatment and the control group. E.g. comparing tissues strong differences are expected.

3.2.6 Gene set evaluation: Over-representation and group testing

At this stage the analysis lead to a list of differentially expressed genes. Gene set evaluation is a way to ascend from the gene expression results to different biological levels. Gene sets are any sets of functionally related genes. Two approaches are implemented in this workflow:

Type	Resources	Cit.
Pathways	KEGG, Transpath, Reactome, BioCyc	[158, 249, 155, 174, 248]
transcription factor target sets	TransFac, ChIP-on-Chip	[200, 219, 220]
Genomic regions	QTL regions	[288]
GO categories	gene ontology database	[26]
Drug targets		[175]
Tissue expression		[286, 287]

Table 3.1: Gene set resources. The first column shows possibilities to define sets of functionally related genes. Implemented resources are listed in the last column.

- Over-representation
- Group testing

Both methods may address different biological questions. The gene sets can be defined by any functional genomics resource see Table 3.1.

Gene set analysis poses some intricate statistical challenges, especially when testing resources with tree structure like the gene ontology [26, 122]. Issues have been addressed by Goeman and Buhlmann [110].

Over-representation analysis

The over-representation analysis is based on the hypergeometric distribution. The overlap of the differentially expressed genes and the predefined gene set is assessed vs. the total number of genes on the array (see Figure 3.10) [169]. If there are n differentially expressed genes and W genes in the gene set the overlap of the differentially expressed genes and genes of the gene set is k. A the same time k is the number of successes in an urn model. The hypergeometric distribution describes the chance of k successes in n draws to hit W genes without replacement [252]. That is the probability

$$P(X=k) = \begin{cases} \dfrac{\binom{W}{k}\binom{N-W}{n-k}}{\binom{N}{n}}, & \max(0, W+n-N) \leq k \text{ and } k \leq \min(n, W) \\ 0, & \text{else} \end{cases}, \tag{3.6}$$

where

- N is the number of genes analysed on the array,
- W is the number of genes in the gene set,
- n the number of differentially expressed genes (the number of draws) and

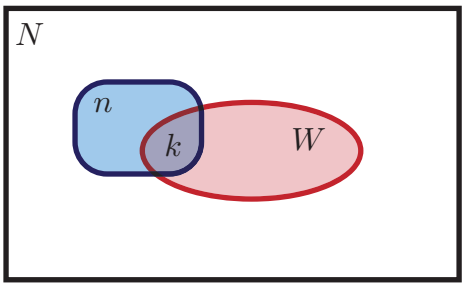


Figure 3.10: Sets compared by the hypergeometric distribution. The hypergeometric distribution quantifies the probability of an overlap k of n differentially expressed genes in a gene set with W genes within N analysed genes.

- k the number of differentially expressed genes in the gene set.

From the differential expression evaluation in Subsection 3.2.5 the number N of total genes under study is derived and n the number of differentially expressed genes. The size of the gene set W is available from the gene set resource. Now the overlap k is computed for the analysis. As output the p-value is provided, i.e. the probability $P(X \geq \ell)$ to hit at least an overlap of ℓ by chance:

$$P(X \geq \ell) = \sum_{k=\ell}^{\min(n,W)} P(X = k) \tag{3.7}$$

The number N constitutes a background where the gene sets are tested against. However different backgrounds come into question and lead to varying p-values in the test outcome. For example array generations cover different sets of genes and vary greatly in size. Thus, the same sample tested on different arrays results in differing over-representation results. Another possible choice for a background on an array is the set of present genes. But this introduces ambiguity as rarely the exactly same set of genes is expressed in the two conditions of a case study. This ambiguity necessitates to generate a consensus for the present genes. In fact the biggest set possible to argue about are all genes on the array. These are genes which have been evaluated and contain all present genes. Over-representation analysis has originally been introduced in Mootha et al. [207] as „gene set enrichment analysis" (GSEA).

Group testing

Group testing follows Makrantonaki et al. [195] and considers expression changes for all genes in the gene set by computing the gene-wise average over the replicates. The result are two vectors of expression means, i.e. μ_T for the treatment expressions and μ_C for the control group. The alteration is assessed for the complete gene set with a two-sided Wilcoxon signed rank test. The null hypothesis is

$$H_0 : \text{median}\,(\mu_T - \mu_C) = 0. \tag{3.8}$$

The alternative hypothesis is H_A : median $(\mu_T - \mu_C) \neq 0$. The p-value is computed if the gene set contains a minimal number of 3 expressed genes. Only expressed genes in the gene set are considered because, like in the case of fold changes, low intensities for not expressed genes lead to confusing results. The group testing is more sensitive for expression changes in the gene sets indicating biological aberrations below differential expression level.

3.3 Alternative splicing (AS) and differential expression with exon arrays

The exon arrays allow an analysis of the transcriptome on two different levels: the exon and the gene level. For the concept of the analysis pipeline this has two consequences:

1. The preprocessing has to be adapted due to the modified design (see Subsection 3.1.2) and the need to summarise and evaluate on exon and gene levels.
2. The evaluation cycle is extended by a completely new branch: alternative splicing analysis.

The concept of the expanded analysis pipeline is illustrated in Figure 3.11. The two chip generations, 3' gene expression and exon array, use the same technology and do well correlate in terms of expression [222, 5], thus it is reasonable that central conclusions are transferable. For example quality control is skipped, since the same procedures apply [15]. Optional approaches for a work flow are presented by several groups [20, 27, 161, 28, 223].

The parallel assessment of differential expression and alternative splicing has another advantage. The relation between alternative splicing and differential expression is an exciting field of research and the presented interpretation of the data enables a tight study of this subject.

3.3.1 Experimental setup and determination of test cases

Case studies are the most prevalent experiment design currently used, for example comparing disease - healthy tissues or altered transcript structure. Here the exon arrays can provide genome-wide splicing event search and are a straightforward expansion of normal expression experiments.

As in the DE pipeline, two groups of samples are compared with each other. W.l.o.g. the groups are denoted as treatment group and control group. The strength of alternative splicing analysis with exon arrays, is the individual preparation of the samples.

3.3.2 Preprocessing

The preprocessing challenges are similar to 3' gene expression arrays. Only the loss of the MM makes changes to the computation necessary. The preprocessing here focuses on four steps:

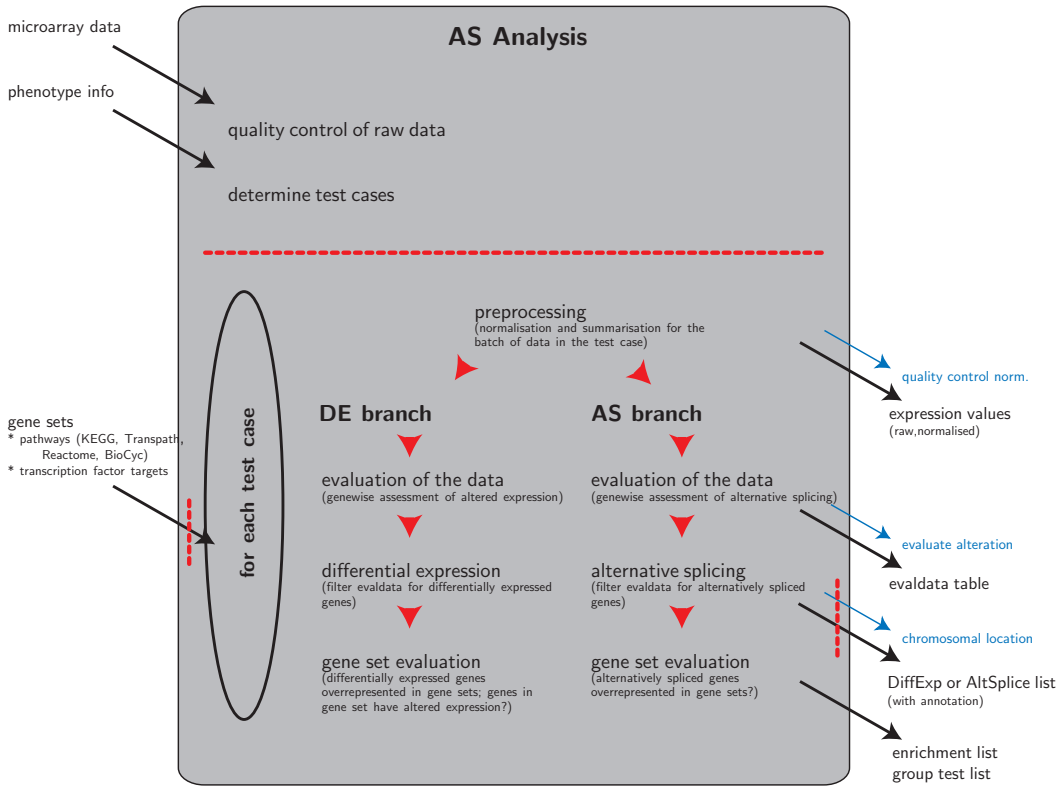


Figure 3.11: The exon array AS pipeline. A scheme for the analysis workflow. After manual determination of the test cases the data is processed automatically. Interfaces which need a mapping of Ensembl genes to probes or different gene identifier are marked with a red dotted line.

1. Intrachip normalisation that corrects for the GC content of the probes;
2. Interchip normalisation that reduces non-biological differences between chips;
3. Summarisation that combines the probe intensities into single gene or exon expression values;
4. Presence tag that computes a detection p-value from the probe hybridisations for every gene or exon.

Intrachip normalisation

Several studies highlight the GC content of the probe sequence as a major effect to correct [313, 212, 325]. Models for correction can be divided in three categories:

- content-dependent correction, the total number of G or C nucleotides in the probe sequence [12, 150];
- position-dependent correction, the binding affinity of a G or C nucleotide differs with the position within the probe sequence [311, 312, 313, 150];
- neighbour-dependent correction, the binding affinity of a G or C nucleotide in respect to the neighbouring nucleotides [325].

Affymetrix proposes a content-specific correction called PM-GCBG (Perfect Match minus GC Background Correction) [12]. The background probes are divided in so-called GC bins, where control probes are collected with the same GC content in the probe sequence. For a probe with a certain GC content the median of the corresponding GC bin is subtracted.

For the 3' gene expression arrays GC-RMA used a position dependent model to correct the hybridisation values for the GC content of the probe sequences [312, 311, 313, 310]. Since the model was developed for the 3' based *in vitro* transcriptase protocol the model is not necessarily appropriate for the new generation of whole-transcript protocols.

Genome tiling arrays use the same protocol and Johnson et al. [150] presented a model, the Model-based Analysis of Tiling-arrays (MAT) to tackle this task. The method combines content and position dependency of probe sequences in a linear model.

$$\log(\alpha_p) = t \cdot n_{p,k=T} + \sum_{j=1}^{25} \sum_{k \in \{A,C,G\}} \beta_{j,k} \cdot 1_{p,j,k} + \sum_{k \in \{A,C,G,T\}} \gamma_k \cdot n_{p,k}^2 + \epsilon_p \qquad (3.9)$$

where

- α_p is the affinity of probe p,
- $n_{p,k}$ is the nucleotide k count in probe p sequence, with k in $\{A, C, G, T\}$,
- t is the baseline value based on the number of T nucleotides in the probe sequence,
- $1_{p,j,k}$ is an indicator function with

$$1_{p,j,k} = \begin{cases} 1, & \text{if the nucleotide at position } j \text{ is } k \text{ in probe } p \text{ sequence} \\ 0 & \text{else} \end{cases},$$

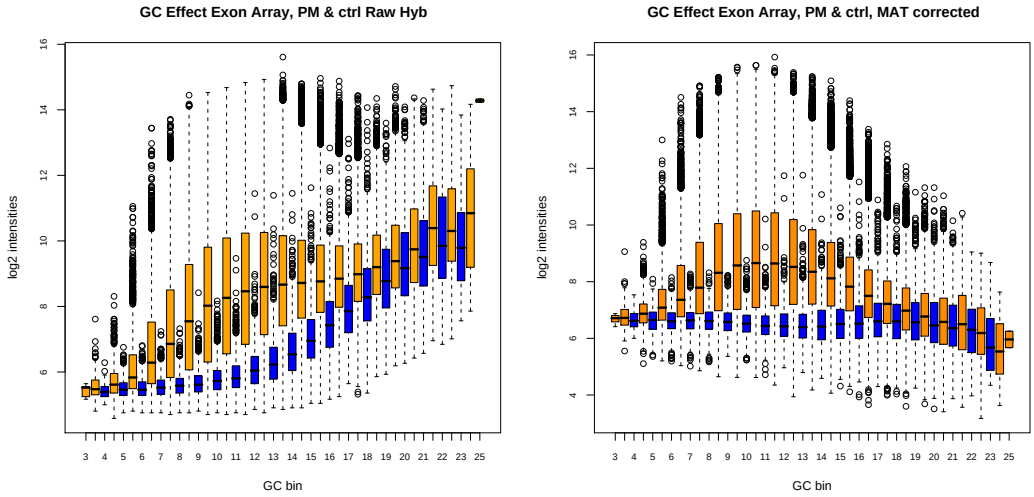


Figure 3.12: MAT GC-correction. On the left hand side are the distributions of the GC bins, orange the PM bins and blue the control bins. On the right hand side visualised the same bins with MAT affinity correction.

- $\beta_{j,k}$ is the effect of each nucleotide k in $\{A, C, G\}$ at each position j,
- γ_k is the effect of nucleotide k count squared and
- ϵ_p is the probe-specific error term following the assumption of a normal distribution.

The model comprises 80 parameters: 1 for t, 25×3 for β and 4 for γ. The parameters are estimated from the genomic background probes on the array. The performance of MAT can be assessed visually in Figure 3.12[3].

MAT is a probe affinity model. On linear intensity scale, the probe intensities are divided by an estimated probe affinity. It is not a model for background signal:

$$\iota_p = \frac{\eta_p}{\alpha_p}, \tag{3.10}$$

where

- η_p is the measured hybridisation of probe p,
- ι_p is the GC-corrected intensity of probe p and
- α_p is the binding affinity of probe p.

The parameters of the model are easy to estimate from the control probes and subsequently the probe affinities are easy to calculate for the PM. It is implemented for intrachip normalisation, similar to Kapur et al. [161], since it provides the most advanced GC correction for whole-transcript prepared samples.

[3]If not differently noted for exon array evaluations and images following versions are used: R 2.8.0, BioC 2.3.0, customCDF 11, Ensembl 49, Affymetrix Mouse Exon 1.0 ST Array. Original data is from the GGSC data set introduced in Subsection 5.3.1.

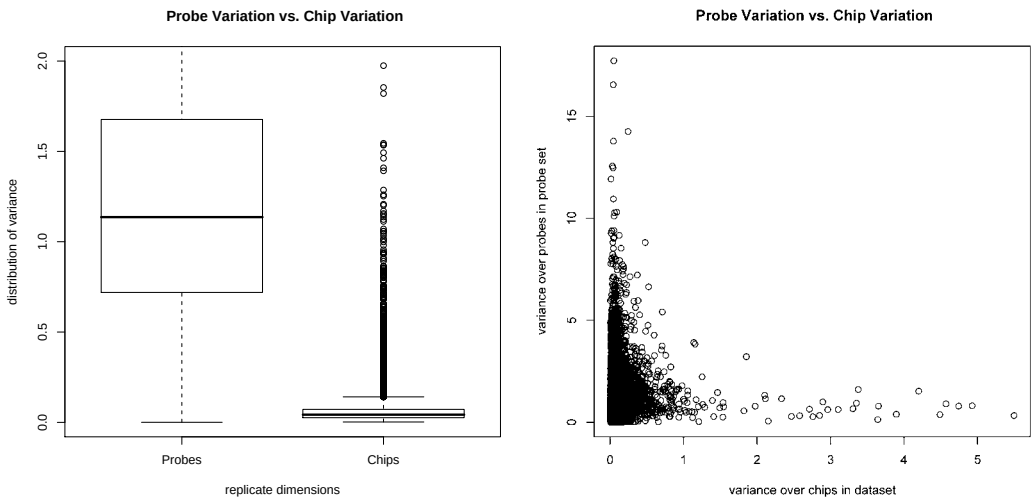


Figure 3.13: Expression variance in genes in two replicate dimensions. The predominant variance is the variance over the probes in a gene compared to probe variance over the replicates.

Interchip normalisation

To reduce unwanted effects between arrays the quantile normalisation has been successful with 3' gene expression arrays, like in RMA. Therefore normal quantile normalisation is applied from the limma package to adjust the intensity distributions over the arrays [268].

Summarisation

The evaluation on two levels (gene/exon), makes two summarisations necessary. Robust estimates with medianpolish as median decomposition were successful in 3' gene expression arrays. The median is computed now over the intensities in both replicate dimensions - arrays and probes. This provides robust summarised expression values also for low-replicate settings on gene and exon level.

Some figures like the coefficient of variation were computed by probe set replication over the arrays in the 3' gene expression arrays. Now the figures are calculated before summarisation on the probe level. To avoid summarisation means to use replication in two dimensions: 1) The different probe intensities within the probe set, 2) the arrays with the same biological condition. As the variation over the chips is smaller than over the probes it should be more easy to ignore the chips than the different probes (see Figure 3.13). This is straightforward with the need to compute the statistical numbers in low replicate settings [237]. Recent papers indicate improved accuracy for probe level analysis [188, 192, 193, 242].

In RMA, the summarisation uses a robust, median-based variant of ANOVA, called medianpolish [144]. The result is a chip-wise summarisation, where the subsequent eva-

luation depends on the replicates. A well developed summarisation method FARMS focuses on low-noise probes accompanied by a measure of variation within the probe set [127, 294]. It is useful for data sets with more than six replicates. The manufacturer recommends the model-based PLIER [19, 14]. Originally developed for 3' gene expression arrays it is also applicable for exon arrays.

Presence tag

For the exon arrays the detection call is calculated by a Wilcoxon signed rank test. Similar to the intrachip normalisation control probes in a chip are divided in GC bins following the number of G or C nucleotides in the probe sequence, e.g. a probe is in GC bin 5 because there are 3 G nucleotides and 2 C nucleotides in the 25 nucleotides probe sequence. Every probe intensity $\iota_{r,g,b}$ is directly compared to the 75%-quantile $Q_{0.75,r,g,b}$ of its corresponding GC bin b within the chip r. Thus there is a pairing along the gene g within the chip r. The pairings over replicates r are joined. The p-value of the one-sided Wilcoxon signed rank test is then calculated using the chip-wise pairing of probe intensities to control quantiles with the null hypothesis, that intensities are less or equal to bin quantiles (see Figure 3.14):

$$H_0 : \text{median}_{r,p}(\iota_{r,p,b} - Q_{0.75,r,p,b}) \leq 0 \tag{3.11}$$

The only threshold in this computation is the height of the quantile (75%) in the GC bin.

To calibrate the threshold various tissue experiments are compared (see Table 3.2). The presented values are a comparison to previous 3' gene expression presence tags from Subsection 3.2.4. For no tissue a definite set of present genes is available. Exon arrays cover more genes and although total numbers are higher, the ratios are often lower than compared to the corresponding 3' gene expression samples. In general about 90% of the genes called to be present on a HG-U133 Plus 2.0 are also present on the exon array. Thus, the 75% quantile of the control probe bins is an extrapolation of the 3' gene expression threshold to exon arrays.

As a replacement for the former detection calls from Subsection 3.2.4 Affymetrix introduced DABG (Detection Above BackGround) [12, 20]. Control probes are divided in GC bins following the number of G or C nucleotides in the probe sequence. For a single probe the probe intensity is compared to the intensity distribution of its corresponding GC bin. The quantile of the probe intensity within the GC bin is now treated like a p-value for the probe (For example if a probe has rank 60 in 1000 bin probes, it is assigned the p-value 0.06). The probe p-values are combined with the Fisher method to generate an exon-level probe set p-value. The manufacturer recommends not to use gene level detection calls due to the fact that not necessarily all exons are expressed. In fact a gene is considered to be expressed, if a certain number (half) of the exons is expressed.

Several arguments contradict this position of Affymetrix. First, the Fisher methods is very susceptible to the number of combined p-values. Since the number of probes in a probe set is not equal, due to the design or alternative assignments, p-values decrease

	exon array		HGU133plus2		Novartis 2002		Novartis 2004	
	total	ratio	total	ratio	total	ratio	total	ratio
breast	10 080	0.38	9703	0.56	–	–	–	–
cerebellum	13 624	0.51	9894	0.57	3700	0.45	3366	0.21
heart	10 336	0.39	7722	0.44	2502	0.31	2347	0.14
kidney	11 093	0.42	9124	0.52	2515	0.31	2965	0.18
liver	10 951	0.41	8024	0.46	1884	0.23	2525	0.15
muscle	10 021	0.38	8512	0.49	–	–	970	0.06
pancreas	10 060	0.38	7644	0.44	1778	0.22	2853	0.17
prostate	10 254	0.39	9732	0.56	–	–	4485	0.27
spleen	12 745	0.48	9552	0.55	3051	0.37	–	–
testis	14 261	0.54	11 105	0.64	4121	0.51	3933	0.24
thyroid	11 767	0.44	9830	0.56	4110	0.5	5583	0.34
total No genes	26 538		17 429		8148		16 437	

Table 3.2: Tissue presence numbers. For different tissue data sets the number of present genes and the corresponding ratio is computed. In the column 'total' is the total number of present genes and in the column 'ratio' is the quotient of the total number of present genes divided by the total number of genes on the array, the bottom row. Affymetrix used exactly the same samples for hybridisation on the exon arrays and HG-U133 Plus 2.0. The Novartis data sets represent older platforms [286, 287]: For Novartis 2002 the HG-U95A and for Novartis 2004 the HG-U133A as well as a custom array gnGNF1Ba. Abbrv.: HGU133plus2, HG-U133 Plus 2.0.

severely with increasing probe set size. Second the primary interest is to know which genes are expressed. If a gene is not expressed in any condition, evaluation may be skipped. If a gene is only expressed in one condition this leads to differential expression. Only if the gene is expressed in both conditions, it can be subject to alternative splicing. Third the DABG uses several thresholds in one algorithm. The combination of p-values, a cut-off of 0.05 for exon level p-values and the number of exons to be expressed in a gene.

3.3.3 Differential expression evaluation and filter

The same evaluation as in the DE pipeline applies for the exon array results. First a gene-wise assessment is computed by different statistical means for three criteria presency, variation and alteration. Where in the calculation of the criteria in 3' gene expression arrays only replication over arrays is used, now both replicate dimensions are available, over arrays and probes in gene/exon.

Differentially expressed genes are filtered combining the three criteria:

- The gene is expressed in at least one of two samples, treatment or control, with a p-value of at most 0.05.
- A ratio of minimal 4/3 or maximal 3/4 between treatment and control sample.

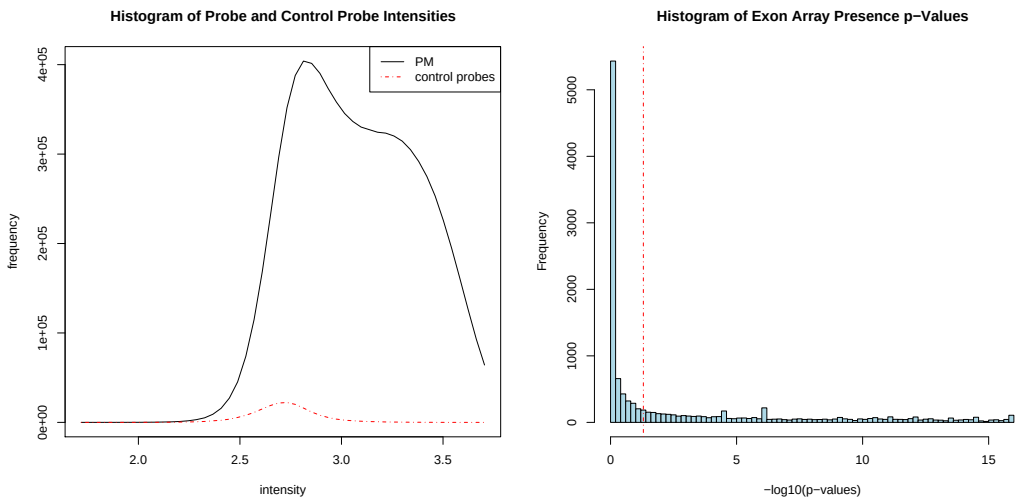


Figure 3.14: Distribution of probes and presence tags on exon arrays. Left: Distribution of PM in black and control values in dotdashed red. **Right:** Distribution of presence tag p-values on -log10 scale. The dotdashed line represents the 0.05 threshold usually used for the detection.

- If calculable, the Wilcoxon-test between the expressions for the two samples is significant with a p-value of at most 0.05.

Due to similar output as the DE pipeline the same gene set evaluation applies. That is over-representation by hypergeometric distribution and group testing by expression vectors. This shows a strength of the pipeline concept. Modules of the DE pipeline directly apply to the AS pipeline.

3.3.4 Alternative splicing evaluation and filter

Assessing alternative splicing poses the intricate statistical question to jointly evaluate gene expression and exon expression. Same criteria used for the genes may be applied to exons but do not evaluate splicing. Thus it is necessary to develop and evaluate new splicing criteria. This is performed in Chapter 4. Here the results are summarised for splicing identification as incorporated in the AS pipeline. Implemented criteria are Splicing Index/SI, MiDAS, PAC, ANOSVA and ARH. ARH provides gene-level predictions with the basic ARH, the ARH p-values and q-values. The q-values are deduced from the p-values similar as above by methods of Storey et al. [282]. ARH provides exon-level predictions with the splicing deviation.

In the pipeline again three criteria are used to filter for a set of alternatively spliced genes:

- Computation of ARH for exons present in at least one condition;
- ARH p-value is below 0.05;
- The gene is present in both conditions.

Algorithm 3.1 Definition of test cases in R.

```
# ------------------------------------------------------------------
TestCases.append = function(TestCases, Name, chip_treat, chip_ctrl) {
  NewCase = new("Case",
    chip_treat = chip_treat,
    chip_ctrl = chip_ctrl
  )
  TestCases.new = c(TestCases, list(NewCase))
  names(TestCases.new)[ length(TestCases.new) ] = Name
  return(TestCases.new)
}

phenoData = as.matrix(read.delim(phenoDataPath,
  header = TRUE, quote = "", row.names = 1)
)

Name = "TestCaseName"
chip_treat = rownames(phenoData)[ phenoData[ , "mouse" ] == "strain_t" ]
chip_ctrl = rownames(phenoData)[ phenoData[ , "mouse" ] == "strain_c" ]
TestCases = TestCases.append(
  TestCases = TestCases,
  Name = Name,
  chip_treat = chip_treat,
  chip_ctrl = chip_ctrl
)
```

ARH combines the variation and alteration criteria. Presency is filtered on both levels. Genes may only be spliced, if present in both conditions. Spurious results from non-expressed genes are avoided before assessing splicing by ARH. To filter for the spliced exons the following criteria is added:

- The splicing deviation is minimally 0.53 or the exon has maximal splicing deviation within the gene.

The recommendation of Affymetrix follows similar criteria by using (1) absolute $\log_2$ value of the splicing index below 0.05 and (2) t-test of normalised intensities less than 0.005 or 0.001 [61, 104, 20]. The normalised intensity is the exon expression divided by its gene expression. Finally for the set of alternatively spliced genes the over-representation analysis is used to identify biological processes affected by splicing.

3.4 Use of the pipelines for different research projects

The microarray pipelines are applied in various projects. Table 3.3 lists the data sets processed with one of the pipelines along the research projects. Altogether 74 data sets have been processed with 1019 test cases and 2990 arrays. Focus on the Affymetrix array platforms allowed automatic processing of 16 different chip types for the organisms human, mouse and rat. From all the data sets 47.3% were processed by Dr. Andriani Daskalaki, 13.5% by Reha Yildirimman and 8.1% by Dr. Mireia Vilardell.

As an example for the R implementation in algorithm 3.1 the test case definition is listed. A class *Case* was defined storing test case name and associated treatment as well as control files. In the listing, the function *TestCases.append* appends a new test case *TestCaseName* to the list *TestCases*. The phenotype information about two mouse strains is saved in a text file and read to R in the *phenoData* object. Thus files are selected by the mouse strain name.

The pipelines are implemented in R/BioC and run under any compilation of R provided that the necessary packages are installed, i.e. Linux/Unix/Windows NT. The running time of the DE pipeline is about half an hour per test case on a machine with AMD Opteron 852 CPU and 16 GB RAM. The considerably higher amount of data in the exon arrays leads to a running time of one hour per test case in the same technical setting.

Table 3.3: Applications of DE/AS pipeline. The pipelines are applied in a series of projects for various data sets. Abbrv.: TC, number of test cases; A, number of arrays; P, pipeline; Cit. citation of related output publication.

Project	Data sets	Platform	TC	A	Species	P	Cit.
AnEUploidy (EU)	AltugTeber	HG-U133 Plus 2	2	21	human	DE	
AnEUploidy (EU)	Amano	MG-U74Av2	2	25	mouse	DE	
AnEUploidy (EU)	Bahn	HG-U133A	1	15	human	DE	
AnEUploidy (EU)	Doherty	HG-U133A	4	3	human	DE	
AnEUploidy (EU)	Mao	HG-U133A	1	25	human	DE	
AnEUploidy (EU)	Mulligan	MG-U74Av2	1	8	mouse	DE	
carcinogenomics (EU)	ConnectivityMap	HG-U133A	307	546	human	DE	
carcinogenomics (EU)	E-MEXP-438	Mouse 430A 2.0	2	12	mouse	DE	
carcinogenomics (EU)	E-MEXP-82	MG-U74Av2	6	27	mouse	DE	
carcinogenomics (EU)	E-TABM-89	Mouse 430A 2.0	15	57	mouse	DE	
carcinogenomics (EU)	E-TOXM_11	RG-U34A	12	79	rat	DE	
carcinogenomics (EU)	E-TOXM_17	MG-U74Av2	8	30	mouse	DE	
carcinogenomics (EU)	E-TOXM_19	RG-U34A	6	58	rat	DE	
carcinogenomics (EU)	E-TOXM_21	RAE230A	3	15	rat	DE	
carcinogenomics (EU)	E-TOXM_28	RAE230A	6	27	rat	DE	
carcinogenomics (EU)	E-TOXM_34	Mouse 430A 2.0	8	104	mouse	DE	
carcinogenomics (EU)	E-TOXM_35	RAE230A	18	154	rat	DE	
cooperation	DifE HM	Mouse 430A 2.0	1	2	mouse	DE	
cooperation	DifE pancreas	Mouse 430A 2.0	1	2	mouse	DE	
cooperation	DifE pancreatic islets	Mouse 430A 2.0	12	12	mouse	DE	[86]
cooperation	RichardYaspo2009	HuEx 1.0 ST v2	1	4	human	AS	[245]
EMBRACE (EU)	Novartis Tissue 2002	HG-U95A	36	72	human	DE	
EMBRACE (EU)	Novartis Tissue 2002	MG-U74A	35	79	mouse	DE	
EMBRACE (EU)	Novartis Tissue 2004	HG-U133A	79	158	human	DE	
EMBRACE (EU)	Novartis Tissue 2004	gnGNF1Ba	79	158	human	DE	
EMBRACE (EU)	Novartis Tissue 2004	gnGNF1Musa	61	122	mouse	DE	
meta-analysis OI	E-GEOD-10334	HG-U133 Plus 2	1	247	human	DE	[81]
meta-analysis OI	E-GEOD-10526	HG-U133A	1	8	human	DE	[81]
meta-analysis OI	E-GEOD-2525	HG-U133A	3	9	human	DE	[81]
meta-analysis OI	E-GEOD-6751	HG-U133 Plus 2	2	59	human	DE	[81]
meta-analysis OI	E-GEOD-6927	HG-U133A	3	12	human	DE	[81]
meta-analysis OI	E-GEOD-7321	HG-U133A	1	2	human	DE	[81]
meta-analysis OI	E-GEOD-9723	HG-U133A	3	12	human	DE	[81]
meta-analysis T2DM	BiddingerKahn2005	MG-U74Av2	6	23	mouse	DE	[240]
meta-analysis T2DM	GuntonKahn2005	HG-U133A, HGU133B	1	25	human	DE	[240]
meta-analysis T2DM	LanAttie2003	MG-U74Av2	4	16	mouse	DE	[240]
meta-analysis T2DM	MoothaGroop2003	HG-U133A	2	42	human	DE	[240]
meta-analysis T2DM	NadlerAttie2000	Mu11KsubA, Mu11KsubB	1	10	mouse	DE	[240]

continued on next page

Table 3.3: Applications of DE/AS pipeline. The pipelines are applied in a series of projects for various data sets. Abbrv.: TC, number of test cases; A, number of arrays; P, pipeline; Cit. citation of related output publication.

Project	Data sets	Platform	TC	A	Species	P	Cit.
method comparison	AbduevaTriche2007	HuEx 1.0 ST v2	10	15	human	AS	[241]
method comparison	human tissue	HuEx 1.0 ST v2	39	33	human	AS	[241]
method development	GardinaTurpaz2006	HuEx 1.0 ST v2	1	20	human	AS	
method development	human tissue	HG-U133 Plus 2	11	33	human	DE	
PhysioSim (BMBF)	ESGEC	Mouse 430A 2.0	72	56	mouse	DE	
SysCo (EU)	BALBc	MoGene 1.0 ST v1	15	48	mouse	DE	
SysCo (EU)	C57BI	MoGene 1.0 ST v1	15	48	mouse	DE	
SysCo (EU)	E-GEOD-10532	Mouse 4302	1	6	mouse	DE	
SysCo (EU)	E-GEOD-10765	Mouse 4302	4	13	mouse	DE	
SysCo (EU)	E-GEOD-11199	HG-U133 Plus 2	1	24	human	DE	
SysCo (EU)	E-GEOD-11497	Mouse 4302	1	4	mouse	DE	
SysCo (EU)	E-GEOD-13147	Mouse 4302	1	4	mouse	DE	
SysCo (EU)	E-GEOD-14890	Mouse 4302	3	9	mouse	DE	
SysCo (EU)	E-GEOD-14891	Mouse 4302	3	8	mouse	DE	
SysCo (EU)	E-GEOD-2002	Mouse 4302	1	9	mouse	DE	
SysCo (EU)	E-GEOD-2973	Mouse 4302	11	37	mouse	DE	
SysCo (EU)	E-GEOD-360	HG-U95Av2	5	8	human	DE	
SysCo (EU)	E-GEOD-411	MG-U74Av2	6	17	mouse	DE	
SysCo (EU)	E-GEOD-4288	Mouse 4302	1	36	mouse	DE	
SysCo (EU)	E-GEOD-477	MG-U74Av2	1	5	mouse	DE	
SysCo (EU)	E-GEOD-5202	Mouse 4302	3	12	mouse	DE	
SysCo (EU)	E-GEOD-5555	Mouse 4302	1	42	mouse	DE	
SysCo (EU)	E-GEOD-5589	MG-U74Av2	10	34	mouse	DE	
SysCo (EU)	E-GEOD-6690	MG-U74Av2	1	4	mouse	DE	
SysCo (EU)	E-GEOD-7348	Mouse 4302	3	6	mouse	DE	
SysCo (EU)	E-GEOD-7649	MG-U74Av2	1	2	mouse	DE	
SysCo (EU)	E-GEOD-7769	Mouse 430A2	1	2	mouse	DE	
SysCo (EU)	E-GEOD-8621	Mouse 4302	1	12	mouse	DE	
SysCo (EU)	E-GEOD-9184	Mouse 4302	1	3	mouse	DE	
SysCo (EU)	E-GEOD-9509	Mouse 4302	1	18	mouse	DE	
SysCo (EU)	E-MEXP-1254	MG-U74Av2	2	12	mouse	DE	
SysCo (EU)	E-MEXP-1290	HG-U133 Plus 2	1	7	human	DE	
SysCo (EU)	E-TABM-102	Mouse 4302	1	30	mouse	DE	
SysCo (EU)	E-TABM-310	Mouse 4302	9	29	mouse	DE	
SysCo (EU)	IPP	HuGene 1.0 ST v1	40	39	human	DE	
SysProt (EU)	GGSC	MoEx 1.0 ST v1	6	25	mouse	AS	

4 Statistical Analysis of Alternative Splicing

In this Chapter the concept of entropy is introduced to the field of alternative splicing prediction. It develops a new method called ARH – Alternative splicing Robust prediction by Entropy [241]. The primary goal is to develop a method which is robust in the number of replicates and independent from the number of exons. For comparison, eight different methods proposed for splicing prediction on exon arrays are presented. In a broad evaluation the performance is assessed on several aspects like dependency on the numbers of exons, splicing prediction in the case of differential expression or no differential expression and robustness in the numbers of replicates. The evaluation runs on a tissue data set and in an artificial setting with a spike-in experiment resulting in a total of four different test settings: pairwise tissue comparison with database confirmed events, tissue specificity with database confirmed events, tissue specificity with RT-PCR validated events as well as the *in vitro* samples with generated events.

The focus is on detection of exon skipping events. Design of the exon arrays is just adequate for this type of splicing events (see subsections 2.1.1 and 3.1.2). For a gene with m exons this allows a combinatorial number of $\sum_{e=1}^{\lfloor \frac{m}{2} \rfloor} \binom{m}{e}$ events. Since the change of a majority of exons constitutes a gene expression change the limit of spliced exons is $\lfloor \frac{m}{2} \rfloor$.

4.1 Preliminaries

W.l.o.g. it follows the assumption of a two sample experiment, a case study design. This is the basic setting also used in differential expression experiments and is probably the most common experiment design. Methods for a 1-to-many design are proposed. But this design mainly occurs in tissue experiments. The 1-to-many design is reducible to the 1-to-1 design by taking all other conditions as controls.

On the array a set of probes is synthesised. The probes are mapped to exons. The mapping from probes to exons is not injectiv as an exon is measured by several probes. It is also not surjective as not all exons in a reference database have to be covered. On the other hand exons are mapped to genes with a surjective mapping iff probes cover all exons in the database. It follows a mapping from probes to genes. For all probes the mapping probe$\mapsto$exon$\mapsto$gene is called an assignment.

To describe alternative splicing and splicing prediction a gene g is chosen. Gene g consists of m exons indexed by e. Exon e is measured by n_e probes, so g has $n_g = \sum_{e=1,\ldots,m} n_e$

probes. W.l.o.g. there is a two sample study with 2 conditions $d = c, t$ for the control and treatment case respectively. The treatment case $d = t$ has q replicates and the control case $d = c$ has s replicates, both indexed with r.

For a probe p in the exon e in condition d and in replicate r there is the measured and preprocessed intensity $\iota_{p,e,d,r}$, i.e.

$$\iota_{p,e,d,r} \in \mathbb{R}^+ \tag{4.1}$$

Following an assignment exon or gene expressions can be computed from the set of assigned intensities.

The exon expression $\phi_{e,d}$ or gene expression Φ_d is a computed value notated as a function f of the probe intensities $\iota_{p,e,d,r}$ for a condition d.

$$\begin{aligned} \Phi_d &\in \mathbb{R}^+ \\ \phi_{e,d} &\in \mathbb{R}^+ \end{aligned} \tag{4.2}$$

where $\Phi_d = f(\iota_{p,e,d,r})$ with $f : \mathbb{R}^{n_g \cdot q+} \to \mathbb{R}^+$ and $\phi_{e,d} = f(\iota_{p,e,d,r})$ with $f : \mathbb{R}^{n_e \cdot q+} \to \mathbb{R}^+$. The expression corresponds to the summarised probe set value in 3' gene expression arrays.

If not differently noted the median is used for f with the median over the probes p and replicates r.

$$\begin{aligned} \Phi_d &= \text{median}_{p=1,...,n_g, r=1,...,q} \left(\iota_{p,e,d,r}\right) \\ \phi_{e,d} &= \text{median}_{p=1,...,n_e, r=1,...,q} \left(\iota_{p,e,d,r}\right) \end{aligned} \tag{4.3}$$

Certain methods may assume an expression computed only over the probes p namely $\phi_{e,d,r}$ and similarly for $\Phi_{d,r}$. Thus the replicate measurements are maintained and used in the method. RMA is frequently used where a chip estimate is added to the median. The median is a good choice for a robust expression value. In the following $f(\cdot)$ is avoided by using median$(\cdot)$.

The necessary notation is summarised in the following table:

gene	1 gene	g	
exon	m exons in gene g	$e = 1, ..., m$	
probe	n_e probes in exon e	$p = 1, ..., n_e$	
	n_g probes in gene g	$p = 1, ..., n_g$	$n_g = \sum_{e=1,...,m} n_e$
condition	2 conditions	$d = c, t$	c for control, t for treatment
replicate	q replicates in condition c	$r = 1, ..., q$	
	s replicates in condition t	$r = 1, ..., s$	
intensity	probe intensity	$\iota_{p,e,d,r} \in \mathbb{R}^+$	
expression	gene expression	$\Phi_{d,r} \in \mathbb{R}^+$	
	exon expression	$\phi_{e,d,r} \in \mathbb{R}^+$	

4.2 ARH

Looking at the expression ratios of the exons in a gene a concerted behaviour is expected following gene expression changes. Now exons deviating from the concerted behaviour attract attention. Using information theory the gene can be viewed as an information source and the information content of the exons is assessed with the entropy to rate the significance of a deviating exon. This is the core element of a splicing prediction method noted as ARH – Alternative splicing Robust prediction by Entropy. Different challenges in alternative splicing necessitate to embed the entropy into a five step splicing prediction procedure.

4.2.1 Algorithm

Splicing assessment is provided in two steps. In the first step analysis is performed on the gene level, i.e. ARH identifies spliced genes (see eq. (4.8)). In the second step analysis is performed on the exon level, i.e. splicing deviation (see eq. (4.4)) ranks the exons within a gene and identifies the skipped/included exons. For a gene g with m exons, two biological conditions and corresponding exon expressions $\phi_{e,t}$ and $\phi_{e,c}$, following quantities are computed:

1. The exon splicing deviation, ζ_e, measures the individual deviation of each exon from the median transcript change. Here, log ratios of exon fold changes are computed to account for symmetric measurement of up- or downsplicing. From these log ratios the median is subtracted to correct for global gene expression changes:
$$\zeta_e = \log_2\left(\frac{\phi_{e,t}}{\phi_{e,c}}\right) - \underset{e=1,\ldots,m}{\text{median}}\left(\log_2\left(\frac{\phi_{e,t}}{\phi_{e,c}}\right)\right). \tag{4.4}$$
2. The exon splicing probability is computed as a weighted absolute value of the splicing deviation ζ_e by
$$p_e = \frac{2^{|\zeta_e|}}{\sum_{e=1,\ldots,m} 2^{|\zeta_e|}}. \tag{4.5}$$
Note that for each gene $\sum_e p_e = 1$.
3. To measure whether the exon splicing probabilities are equally distributed or whether a single or a few exons dominate the probability distribution, the entropy is computed for each gene:
$$H_g\left(p_1, \ldots, p_m\right) = -\sum_{e=1}^{m} p_e \cdot \log_2\left(p_e\right). \tag{4.6}$$
4. Entropy defined in eq. (4.6) is dependent on the number of exons and can not be directly used for the comparison of different genes. Thus, in order to make the measure independent of the number of exons for a given gene, entropy is subtracted from its theoretical maximum:
$$\max(H_g) - H_g = \log_2(m) - H_g\left(p_1, \ldots, p_m\right). \tag{4.7}$$

5. Another necessary modification accounts for the strength of deviation within the gene. This is robustly estimated with the interquartile range of exon expression ratios, the 25%, $Q_{.25,e=1,\ldots,m}\left(\frac{\phi_{e,t}}{\phi_{e,c}}\right)$, and 75%, $Q_{.75,e=1,\ldots,m}\left(\frac{\phi_{e,t}}{\phi_{e,c}}\right)$, quantiles. A robust estimate for the amplitude is the interquartile ratio $\frac{Q_{.75,e}}{Q_{.25,e}}$. This ratio is close to 1 for low splicing probability and increases with deviations of a number of exons in the gene. The interquartile ratio is multiplied with the entropy index and constitutes the ARH splicing prediction:

$$\mathrm{ARH}_g = \frac{Q_{.75,e}}{Q_{.25,e}} \cdot (\max(H_g) - H_g)\,. \tag{4.8}$$

   Thus, ARH is suitable to compare the predictions across different genes. Large ARH values indicate splicing.

If a single exon deviates from the remaining exon expression ratios it dominates the splicing probability distribution (eq. (4.6)) resulting in a low entropy and a high ARH value. If a larger number of exons is spliced this measure is upweighted with an increased interquartile ratio greater than one. On the other hand if all exons have similar expression changes this leads to a high entropy with small interquartile ratio and consequently to a small ARH value.

The ARH prediction is implemented in R as `ARH` (see Algorithm 4.1) [238]. `ARH` returns gene level predictions, the outcome of equation (4.8). The function takes two input vectors (x and y in the implementation) for the exon (or probe set) expressions and one vector for the exon-gene grouping (f). To avoid division by zero the second vector is set to a minimum of 0.0001. Genes with only one exon or non-finite exon expressions are set to `NA`. The running time of the implementation is just a few minutes on a machine with AMD Opteron 852 CPU and 16 GB RAM.

4.2.2 Characteristics of ARH

ARH background distribution

For a given experiment the ARH values show a rapid decline from many near-zero values to few high ARH values. The ARH distribution shows little variation even between tissues (see Figure 4.1). To derive a biologically motivated background distribution, samples of the same biological conditions are compared. The human tissue data set from Clark et al. [61] entails data from 11 human tissues with 3 replicates each. In each tissue this allows three pairwise comparisons summing to 33 pairwise comparisons that were used for defining a background sample of ARH values. The distribution of these 33 comparisons provides thresholds for significant ARH values. The 95% quantile of the distribution is $Q_{\mathrm{ARH},.95} = 0.031$. The 95% quantiles of the 33 individual comparisons also cluster around that value (see Figure 4.2). For the 90%, 99%, and 99.9% quantiles of the background distribution the thresholds are $Q_{\mathrm{ARH},.9} = 0.023$, $Q_{\mathrm{ARH},.99} = 0.057$ and $Q_{\mathrm{ARH},.999} = 0.13$, respectively. The background distribution is also adequate to calculate p-values. The generalised extreme value distribution was found to fit best to the

Algorithm 4.1 Implementation of ARH in the R language.

```
ARH = function(x = "numeric", y = "numeric", f = "character", na.rm = FALSE) {
  if(any(x < 0, na.rm = TRUE) | any(y < 0, na.rm = TRUE)) {
    stop("What are negative expressions?")
  }
  y[ y < 0.0001 ] = 0.0001

  splicingDeviations = log2(x / y)
  splicingDeviationsMedian = split(splicingDeviations, f)
  splicingDeviationsMedian = sapply(X = splicingDeviationsMedian, FUN = median,
    na.rm = na.rm
  )
  splicingDeviationsMedian = splicingDeviationsMedian[
    match(f, names(splicingDeviationsMedian))
  ]
  splicingDeviations = 2^abs( splicingDeviations - splicingDeviationsMedian )
  rm(splicingDeviationsMedian)
  splicingProbabilitiesSum = split(splicingDeviations, f)
  splicingProbabilitiesSum = sapply(splicingProbabilitiesSum, sum, na.rm = na.rm)
  splicingProbabilitiesSum = splicingProbabilitiesSum[
    match(f, names(splicingProbabilitiesSum))
  ]
  splicingProbabilities = splicingDeviations / splicingProbabilitiesSum
  rm(splicingDeviations, splicingProbabilitiesSum)

  entropy = split(splicingProbabilities, f)
  entropy = entropy[ match(unique(f), names(entropy)) ]
  entropy = sapply(X = entropy, FUN = function(X)
    return( -sum(X * log2(X), na.rm = na.rm) )
  )
  iqrQuotient = x / y
  iqrQuotient = split(iqrQuotient, f)
  iqrQuotient = iqrQuotient[ match(unique(f), names(iqrQuotient)) ]
  iqrQuotient = sapply(X = iqrQuotient, FUN = quantile,
    probs = c(0.25, 0.75), na.rm = TRUE
  )
  iqrQuotient = iqrQuotient[ "75%" , ] / iqrQuotient[ "25%" , ]
  geneLength = table(f)
  geneLength = geneLength[ match(unique(f), names(geneLength)) ]

  arh = as.numeric(iqrQuotient * (log2(geneLength) - entropy))
  names(arh) = unique(f)
  good = split(is.finite(x) & is.finite(y) & is.finite(splicingProbabilities), f = f)
  good = sapply(good, function(X) return(sum(X) >= 2))
  good = good[ match(unique(f), names(good)) ]
  arh[ !good ] = NA

  return(arh)
}
```

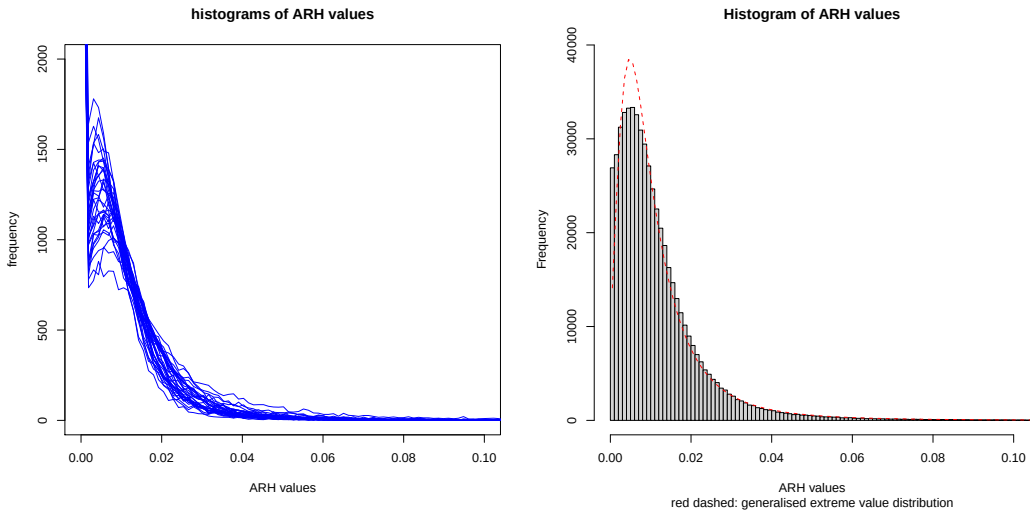


Figure 4.1: ARH histograms. **Left:** depicts the histograms for 33 two chip comparisons within one of the 11 tissues. This reflects a true biological background distribution only containing individual splicing variation. **Right:** Histogram of ARH background distribution derived from splicing predictions between the same biological conditions (ARH values equal to zero were skipped) and the fitted generalised extreme value distribution (red dashed line).

ARH background distribution due to a long heavy tail of large ARH values. Distribution parameters were fit with Matlab resulting in location = 0.006338, scale = 0.005507 and shape = 0.3329 (see Figure 4.1).

Exon-level analysis

In a spliced gene the splicing deviation ranks the exons in order to identify the most altered exons. With this ranking exons can be selected for example for wet-lab validation. Assessing the absolute splicing deviation as above, a global number of spliced exons is determined with the following thresholds: $Q_{\text{ARH\_sd},.9} = 0.43$, $Q_{\text{ARH\_sd},.95} = 0.53$, $Q_{\text{ARH\_sd},.99} = 0.75$ and $Q_{\text{ARH\_sd},.999} = 1.07$.

The exon-level splicing indication has to be symmetric in terms of up- or downsplicing. The swap of treatment and control samples changes an upspliced exon to a downspliced exon and vice versa. The absolute value of the $\log_2$ splicing deviation accounts for this symmetry. The dependence of the splicing probability on the fold changes was simulated for a gene with 13 exons, where $\log_2$ ratios are drawn from a normal distribution with $\mathcal{N}(0, 0.68)$ (see Figure 4.3).

Spliced exons are not necessarily adjacent. In the liver vs. pancreas tissue comparison the transcription factor *HNF4A* is an illustrative example depicted in Figure 4.9. Three exons were predicted to be spliced on positions 1, 4 and 5 with one confirmed event in position 4 in pancreas. The sum in the entropy formula is commutative and reflects the position independence of the exons.

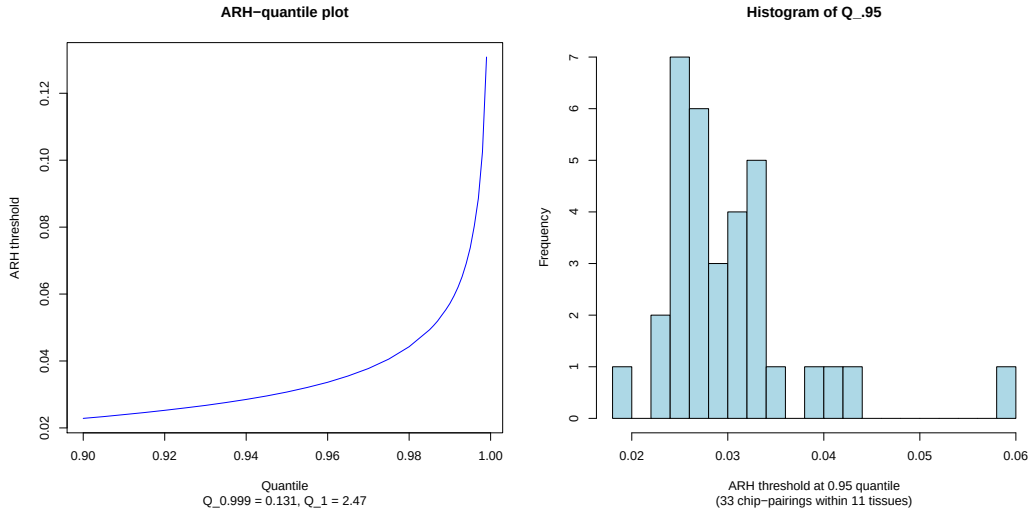


Figure 4.2: ARH quantiles. Left: The curve visualises the cut-offs of the quantiles for $Q_{.9}$ increasing to $Q_{.999}$. The maximal value constituting Q_1 is 2.67. **Right:** This is a histogram of the 33 threshold values for the 95% quantiles. The cut-off of the cumulative distribution is 0.03 and is the centre of the threshold distribution.

Gene-level analysis

A gene-level splicing prediction method requires to be sensitive in the deviation of a proportion of exons what is measured by ARH with the entropy and the interquartile ratio as weighting factor. A simulation is performed with varying number of spliced exons, where the linear ratio of the spliced exons is multiplied with a fold change of 3 (see Figure 4.3). ARH values reflected the number of spliced exons with a flat cap.

A strength of ARH is its low dependency on the total number of exons of a gene. In ARH the genes are sorted in bins by exon number and gene predictions are compared to the gene-bin maximal prediction. Comparing the entropy to the maximal entropy makes ARH independent of the number of exons (see Figure 4.3). See discussion Subsection 4.5.2 for details.

Performance with low number of experimental replicates

Since the costs of experimental replicates are often a limiting factor methods favourably require low number of replicates by computing robust predictions. Purdom et al. [237] were the first to address this aspect for FIRMA. ARH and other methods are compared using a single chip per condition (see Figure 4.4) highlighting the good performance of ARH. ARH predictions are only dependent on the robustness of exon expression calculation. Using the median over the probes but also over the replicates the method is robust in the number of replications.

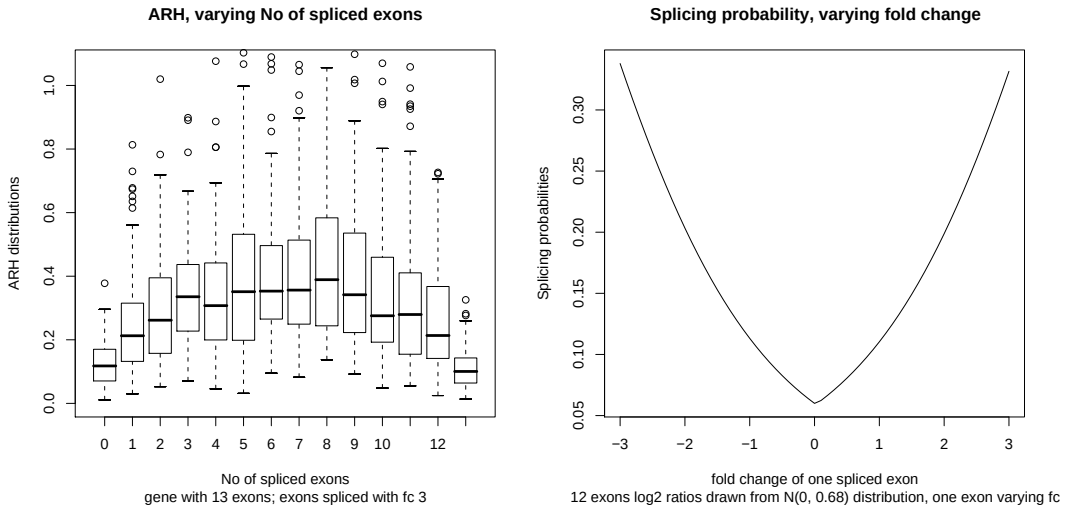


Figure 4.3: ARH simulations. Left: Simulation of ARH values (y-axis) with respect to number of spliced exons (x-axis). A gene with 13 exons is used for simulation corresponding to the average exon number in Ensembl for known protein coding genes. Log_2 ratios were drawn from a normal distribution with mean 0 and standard deviation 0.68 corresponding to the liver vs. pancreas comparison. Respective number of exons were upregulated with a factor of 3 indicating splicing. **Right:** Exons are drawn from the normal distribution and the 13th exon has varying fold change from $\log_2$ ratio -3 to 3.

4.3 Description of different methods

From a variety of prediction methods eight methods are gathered suitable for application on Affymetrix Exon Arrays. All presented methods make *de novo* predictions on the exon or gene level. Of course it is possible to check for known isoforms described in one of the splicing databases. But the most interesting question will be to identify previously unknown isoforms. Some methods are collected in the review of Cuperlovic-Culf et al. [74] in favour of known isoforms or are proposed in Affymetrix whitepapers [11]. All presented methods have been selected and eventually adapted to fit exon arrays. Other array designs facilitate their own specialised methods [260]. If analysis resides on known transcripts also isoform quantification is possible [308, 25].

4.3.1 Splicing index (SI)

The first presented method is an exon-wise prediction with a numerical index [270]. It is similar to the fold change and allows a quantitative analysis of splicing. For exons, normalised intensities (NI) are computed dividing exon expression by gene expression in each biological condition: $\frac{\phi_{e,d}}{\Phi_d}$. Then two biological conditions are compared by the ratio of the NI.

$$P_e^{\mathrm{SI}} = \log\left(\frac{\frac{\phi_{e,t}}{\Phi_t}}{\frac{\phi_{e,c}}{\Phi_c}}\right) = \log\left(\frac{\Phi_c}{\Phi_t}\cdot\frac{\phi_{e,t}}{\phi_{e,c}}\right) \tag{4.9}$$

The splicing index is the most common method propagated by Affymetrix in combination with MiDAS (s.b.) [61, 80, 104]. The method provides an appealing interpretation quantifying splicing effects similar to a fold change in gene expression differences.

4.3.2 SPLICE

The second method is also a numerical index with exon-wise prediction from Hu et al. [134]. A probe intensity $\iota_{e,p,d}$ is divided by the median of all probe intensities of the gene in the condition, $\mathrm{median}_{p\in\{1,\ldots,n_g\}}(\iota_{e,p,d})$. The probe-wise ratio of median corrected intensities is the predictor.

$$P_{e,p}^{\mathrm{SPLICE}} = \log\left(\frac{\frac{\iota_{e,p,t}}{\underset{p\in\{1,\ldots,n_g\}}{\mathrm{median}}(\iota_{e,p,t})}}{\frac{\iota_{e,p,c}}{\underset{p\in\{1,\ldots,n_g\}}{\mathrm{median}}(\iota_{e,p,c})}}\right) = \log\left(\frac{\iota_{e,p,t}\cdot\underset{p\in\{1,\ldots,n_g\}}{\mathrm{median}}(\iota_{e,p,c})}{\iota_{e,p,c}\cdot\underset{p\in\{1,\ldots,n_g\}}{\mathrm{median}}(\iota_{e,p,t})}\right) \tag{4.10}$$

Originally developed for a 1-to-many approach on 3' arrays, some adaptations are indispensable. The difference $PM - MM$ is replaced by the intensity ι, as there is no need to subtract anything after the background correction. The denominator averaged all tissues excluding the tissue under study. Here the average is replaced by the control values. The method does not take into account any replicates, thus the index r is omitted. The replicate probe values are condensed beforehand into a single probe value.

Originally, the protruding probes were clustered with an algorithm called NEIGHBORHOOD. This step can be avoided with the pre-knowledge of the assignment and directly deduce a prediction for the exon. An exon-level prediction is calculated from the probe predictors by considering the median:

$$P_e^{\mathrm{SPLICE}} = \underset{p\in\{1,\ldots,n_e\}}{\mathrm{median}}\left(P_{e,p}^{\mathrm{SPLICE}}\right) \tag{4.11}$$

4.3.3 Pattern-based correlation (PAC)

The pattern-based correlation (PAC) compares the treatment exon expression $\phi_{e,t}$ to a scaled treatment gene expression Φ_t [11]. The treatment gene expression Φ_t is scaled by the quotient of the general exon expression, $\mathrm{median}_{p,d,r}(\iota_{e,p,d,r})$, through the general gene expression, $\mathrm{median}_{e,p,d,r}(\iota_{e,p,d,r})$. The predictor P_e^{PAC} is the scaled gene expression subtracted from the treatment exon expression.

$$P_e^{\mathrm{PAC}} = \phi_{e,t} - \Phi_t\cdot\frac{\mathrm{median}_{p,d,r}(\iota_{e,p,d,r})}{\mathrm{median}_{e,p,d,r}(\iota_{e,p,d,r})} \tag{4.12}$$

PAC was also developed for 1-to-many experiments, where the minuend and the subtrahend can be correlated over all conditions. For the two sample setting, the correlation is not applicable and is replaced by the difference. The presented variant was applied in French et al. [99].

4.3.4 Analysis of splice variation (ANOSVA)

A two-way ANOVA is applied on $\log_2$ probe intensities $\iota_{e,p,d,r}$ with a factor for the exon α_e in the gene and a factor for the biological condition β_d. The interaction factor $\gamma_{e,d}$ between the exon factor and the condition factor indicates splicing [320, 66].

$$\log_2\left(\iota_{e,p,d,r}\right) = \mu + \alpha_e + \beta_d + \gamma_{e,d} + \varepsilon_{e,p,d,r} \tag{4.13}$$

The $\varepsilon_{e,p,d,r}$ is a gaussian error term in the probes p and replicates r. The null hypothesis assumes the effects can be explained by exon effects in α_e or overall gene effects in β_d:

$$H_0 : \gamma_{e,d} = \gamma_{i,j} \text{ for any } (e,d) \neq (i,j) \tag{4.14}$$

The alternative is $H_A : \gamma_{e,d} \neq \gamma_{i,j}$ for at least one pair of (e,d) and (i,j). The statistical model is fit on the available probe intensities and thus does not need a summarisation step. The resulting p-value is the predictor P_g^{ANOSVA}. Testing for the two main factors facilitates analysis of exonic variation and differential expression. Due to the amount of data available for the fit, tests for genes with many exons have higher power and are more likely to return a significant p-value than shorter genes (see Figure 4.7). Using the more robust Kruskal-Wallis rank test for the same null hypothesis increases sensitivity by a drastic decrease of specificity.

4.3.5 Microarray detection of alternative splicing (MiDAS)

Beside the splicing index Affymetrix proposes an exon-level t-test between the conditions [11]. Exon normalised intensities are the exon expression $\phi_{e,d}$ divided by gene expression Φ_d. Logged normalised intensities $\log\left(\frac{\phi_{e,d}}{\Phi_d}\right)$ are compared between samples. A t-test is applied over the replicates r between both conditions c, t:

$$H_0 : \mu_{t,r=1,\ldots,s}\left(\log\left(\frac{\phi_{e,t}}{\Phi_t}\right)\right) = \mu_{c,r=1,\ldots,q}\left(\log\left(\frac{\phi_{e,c}}{\Phi_c}\right)\right). \tag{4.15}$$

Thus the exons are tested for differential inclusion between conditions and the resulting p-values is the predictor P_e^{MiDAS}.

Originally, the method was developed for 1-to-many experiments. In that case the normalised intensities are compared with an ANOVA test over the different conditions. In a 1-to-1 experiment setting the test reduces to a t-test. For expression estimates the manufacturer proposes PLIER [19].

MiDAS is implemented in a command-line tool `apt` (Affymetrix power tools) [20]. After preprocessing, the probe intensities are passed to `apt` for MiDAS computation. In the description of `apt`, the inventors note that the output is not a p-value but similar to a p-value. Characteristics of the output suggest some tuning of the test, for example for multiple testing or exon number correction, not described in the above cited documentation.

4.3.6 Microarray analysis of differential splicing (MADS)

The MADS splicing prediction is a four step procedure, following the same idea as MiDAS [316]. For each probe, the ratio of probe intensity to the estimated gene expression index is calculated:

$$\hat{\iota}_{e,p,d,r} = \frac{\iota_{e,p,d,r}}{\Phi_d}. \tag{4.16}$$

Two separate one-sided t-tests assess whether the ratios of a probe are significantly higher or lower in one sample group over another:

$$H_0 : \mu_{e,p,t,r=1,\ldots,s}\left(\hat{\iota}_{e,p,t,r}\right) = \mu_{e,p,c,r=1,\ldots,q}\left(\hat{\iota}_{e,p,c,r}\right) \tag{4.17}$$

with alternative hypotheses

$$\begin{aligned} H_{A1} : \mu_{e,p,t,r=1,\ldots,s}\left(\hat{\iota}_{e,p,t,r}\right) &> \mu_{e,p,c,r=1,\ldots,q}\left(\hat{\iota}_{e,p,c,r}\right) \\ H_{A2} : \mu_{e,p,t,r=1,\ldots,s}\left(\hat{\iota}_{e,p,t,r}\right) &< \mu_{e,p,c,r=1,\ldots,q}\left(\hat{\iota}_{e,p,c,r}\right). \end{aligned}$$

The obtained probe $p_{e,p}$-values are summarised for each alternative hypothesis to exon p_e-values using the Fisher method as follows (Notation for the alternative hypotheses is temporarily skipped). The $p_{e,p}$-values for individual probes are transformed via the formula $x = -2 \cdot \ln(p_{e,p})$. Under the null hypothesis that the exon targets are not spliced, the $p_{e,p}$-values follow a uniform $[0, 1]$ distribution, and the transformed p-values follow χ^2_2 distribution. The sum of the transformed p-values follows $\chi^2_{2 \cdot n_e}$ distribution where n_e equals the number of probes. From the sum of the transformed p-values distribution the exon p_e-value is deduced.

For the two summarised p-values $p_{A1,e}, p_{A2,e}$ corresponding to the two alternative hypotheses the lower p-value is chosen to be predictive for splicing:

$$P_e^{\text{MADS}} = \min_{A1,A2}\left(p_{A1,e}, p_{A2,e}\right) \tag{4.18}$$

For gene expression index computation, the developers propose an iterative probe selection algorithm [316].

4.3.7 Finding isoforms using robust multichip analysis (FIRMA)

The probe residues of the RMA gene expression estimation are proposed as predictors for splicing in Purdom et al. [237]. RMA was mentioned in the preprocessing of 3' gene

expression arrays, see Subsection 3.2.4 and especially equation 3.3. In the final summarisation step a gene expression is estimated for each replicate by fitting the following additive model for each gene:

$$\log_2(\iota_{e,p,d,r}) = \mu + \alpha_p + \beta_r + \epsilon_{e,p,d,r} \tag{4.19}$$

where

- $\iota_{e,p,d,r}$ is an intensity matrix of a particular gene,
- μ is a base line constant,
- α_p is a probe effect,
- β_r is an array effect and
- $\epsilon_{e,p,d,r}$ is a random error term.

The equation is fitted for both conditions separately. For the fit of the equation (4.19) Purdom et al. [237] use an iteratively reweighted least squares method resulting in the estimations $\hat{\mu}, \hat{\alpha}_p, \hat{\beta}_r$. The residuals of the fit are the probe level splicing predictors:

$$R_{e,p,d,r} = \log_2(\iota_{e,p,d,r}) - \hat{\mu} - \hat{\alpha}_p - \hat{\beta}_r. \tag{4.20}$$

Still this is a probe prediction and an exon level prediction is computed by the median:

$$P_{e,d,r}^{\text{FIRMA}} = \underset{p=1,\ldots,n_e}{\text{median}} \left(\frac{R_{e,p,d,r}}{s} \right), \tag{4.21}$$

where s is an estimate of standard error.

The additional divisor s is estimated by the median absolute deviation of the residuals and helps to make the scores comparable between different genes [237].

FIRMA is implemented using CEL file level, since the splicing prediction is a direct result of the summarisation procedure. Computation is performed with the aroma package [31]. Thus it is the only method in the evaluation with its own preprocessing. The implementation of the authors also requires special annotation files. At the time of processing, only a combined Affymetrix/Ensembl probe-exon-gene assignment was available.

In order to include FIRMA in the method evaluation the maximal prediction was selected from the Affymetrix probe sets within an Ensembl gene, thus generating gene level estimates. Furthermore the method provides chip-wise predictions remaining with varying results across replicates. The average for the treatment replicates was found to perform best as a final predictor:

$$P^{\text{FIRMA}} = \underset{r=1,\ldots,s}{\text{mean}} \left(\max_{e=1,\ldots,m} \left(P_{e,d,r}^{\text{FIRMA}} \right) \right), \tag{4.22}$$

4.3.8 Correlation

The correlation between the exon expression vectors for the two conditions is the predictor [259]:

$$P_g^{\text{Correlation}} = \varrho_{e=1,\ldots,m}(\phi_{e,t}, \phi_{e,c}) \tag{4.23}$$

Algorithm 4.2 Implementation of the splicing index in the R language.

```
gene_xpr_treat = split(probe_int[ , chip_treat ], f = probe_map[ , "gene_id" ])
gene_xpr_treat = gene_xpr_treat[ match(x = unique(probe_map[ , "gene_id" ]),
  table = names(gene_xpr_treat))
]
gene_xpr_treat = sapply(gene_xpr_treat, median)
gene_xpr_ctrl = split(probe_int[ , chip_ctrl ], f = probe_map[ , "gene_id" ])
gene_xpr_ctrl = gene_xpr_ctrl[ match(x = unique(probe_map[ , "gene_id" ]),
  table = names(gene_xpr_ctrl))
]
gene_xpr_ctrl = sapply(gene_xpr_ctrl, median)
exon_xpr_treat = split(probe_int[ , chip_treat ], f = probe_map[ , "exon_id" ])
exon_xpr_treat = exon_xpr_treat[ match(x = unique(probe_map[ , "exon_id" ]),
  table = names(exon_xpr_treat))
]
exon_xpr_treat = sapply(exon_xpr_treat, median)
exon_xpr_ctrl = split(probe_int[ , chip_ctrl ], f = probe_map[ , "exon_id" ])
exon_xpr_ctrl = exon_xpr_ctrl[ match(x = unique(probe_map[ , "exon_id" ]),
  table = names(exon_xpr_ctrl))
]
exon_xpr_ctrl = sapply(exon_xpr_ctrl, median)

SplicingIndex = { log2(exon_xpr_treat) - log2(exon_xpr_ctrl)
  + stretch(log2(gene_xpr_ctrl)) - stretch(log2(gene_xpr_treat))
}
```

Deviation of the correlation from 1 indicates splicing. The developers hypothesise that in the absence of splicing the exon expression pattern between two conditions should be highly correlated, with a Pearson correlation coefficient close to 1 [259]. Differences in splicing and therefore differences in exon signal patterns between the conditions will result in a decrease in the gene's correlation. This decrease in correlation has again an exon number effect in itself. For constant exon splicing deviation and increasing exon number, the correlation will increase due to a decreasing effect of a single exon in the Pearson estimator.

4.3.9 Practical implementation of the methods

All methods are implemented in R except MiDAS and FIRMA. MiDAS values were calculated on the standard preprocessing with the Affymetrix Power Tools in version 1.8.0. FIRMA values were calculated with the package aroma.affymetrix using RMA preprocessing. The running time of all methods in sum is about three hours on a machine with AMD Opteron 852 CPU and 16 GB RAM.

Due to the character of the predictions the methods can be categorised into scores (Splicing Index, SPLICE, PAC, FIRMA, Correlation) or tests (ANOSVA, MiDAS, MADS). Also some methods provide exon level prediction (Splicing Index, SPLICE, PAC, MiDAS, MADS, FIRMA) or gene level prediction (ANOSVA, Correlation). All methods were implemented with the same preprocessing of data except FIRMA which requires

	tissue pairwise average	liver vs. pancreas	liver vs. pancreas, 1-to-1	liver vs. pancreas, no DE	liver vs. pancreas, DE	tissue specific average	muscle vs. rest, AEdb	muscle vs. rest, RT-PCR	spike-in
ARH	**0.83**	**0.86**	**0.84**	**0.80**	**0.96**	**0.86**	**0.86**	**0.97**	**0.99**
SI	0.70	0.74	0.73	0.60	0.86	0.75	0.71	0.95	0.96
SPLICE	0.69	0.78	0.73	0.70	0.86	0.75	0.62	0.88	0.96
PAC	0.63	0.75	0.74	0.59	0.90	0.72	0.64	0.96	0.96
ANOSVA	0.76	0.78	0.72	0.73	0.86	0.7	0.6	0.84	0.98
MiDAS	0.68	0.71	–	0.59	0.83	0.62	0.48	0.85	0.95
FIRMA	0.69	0.73	0.69	0.72	0.73	0.75	0.74	0.92	0.75
MADS	0.68	0.69	–	0.48	0.88	0.71	0.49	0.67	0.98
Cor.	0.74	0.69	0.65	0.76	0.58	0.78	0.75	0.73	0.75
exon No.	0.83	0.79	0.79	0.77	0.81	0.84	0.92	0.93	–

Table 4.1: AUC for different test settings and methods. Exon number predictor (last row) refers to the number of exons per gene. Abbrv.: Cor., correlation; SI, splicing index; exon No., exon number; DE, differential expression.

RMA. The preprocessing follows Subsection 3.3.2.

As an example the implementation of the splicing index is provided in Algorithm 4.2. The preprocessed intensities $\iota_{e,p,d,r}$ are in the object *probe_int* with the arrays in the column. The vectors *chip_treat* and *chip_ctrl* denominate the treatment and control arrays respectively. The matrix *probe_map* reflects the probe - exon - gene assignment. Exon and gene vectors have different length and the function *stretch* adapts gene numbers to corresponding exons. Output is a vector with splicing indices for the exons.

4.4 Evaluation of alternative splicing prediction methods

In general systematic evaluations were thriving steps in method development [70, 145, 59]. ARH is compared with eight existing splicing prediction methods listed in Section 4.3. All methods were compared with the same preprocessing of the data except FIRMA which requires RMA. MiDAS values were calculated on the standard preprocessing with the Affymetrix Power Tools in version 1.8.0.

Methods are evaluated in different test settings using true positive events from splicing database AEdb, transcript spike-in experiment as well as RT-PCR validations. The test settings thus span a broad range of challenges for splicing prediction. Ordering the predictions by decreasing splicing indication constitutes a classifier that allows visualising

the performance of the predictions with the receiver operating characteristic (ROC). ROC curves are visualised with the ROCR-package in R [264, 265, 92]. Using the ROC the performance was quantified with the area under the curve (AUC; see Table 4.1), likewise AUC was computed with ROCR.

4.4.1 Probe assignment and selection of splicing events from the AEdb

For the human tissue data set, probe-exon assignments are drawn from latest genome annotations of Dai et al. [77] in version 11 for Ensembl exons. Exon to gene assignment was retrieved via BioMart from Ensembl 49 [162, 38] and resulted in 232 376 exons that correspond to 26 538 genes.

AEdb contains confirmed splicing events extracted from the literature [274, 172]. The AEdb sequence flat file was downloaded (http://www.ebi.ac.uk/asd/aedb/) and the splicing events were filtered by splicing mechanism (cassette exon events), species (human, mouse, rat) as well as the availability of a sequence for the events. Eight tissues overlap in AEdb and the tissue benchmark data sets. Events attributed to these 8 tissues were selected and the corresponding sequences of the alternative exons were aligned to exon sequences from Ensembl 49 for exact matches. For each tissue this resulted in a list of Ensembl exon identifier: heart 13, kidney 28, liver 27, muscle 26, pancreas 2, spleen 15, testis 43 and thyroid 12. Events may attribute to more than one tissue. For tissue specificity such events are skipped. For pairwise comparison just the events specific to either of the two tissues are used as true positive set. As an example in Table 4.2 is the excerpt of confirmed events for the case of muscle specificity.

4.4.2 Test data set 1: Tissue data with literature confirmed events

As a true positive set for judging the methods performances, an independent data set is chosen from the manually curated AEdb [274]. Experimental data was available for 11 human tissues with 3 experimental replicates per tissue. Confirmed events in AEdb were available for 8 of these tissues what allows for 28 pairwise tissue comparisons (see Figure 4.4). Due to issues discussed in Subsection 4.5.1 it is possible to rank the performances of the methods but not to assess the overall performance.

The benchmark test set was analysed with respect to different aspects. The pairwise tissue comparisons correspond to highly diverse biological conditions leading to a lot of variation in the benchmarks. In the analysis the average performance is provided across the 28 pairwise comparisons (see Figure 4.4). For an in-depth discussion of alternative splicing attributes the liver vs. pancreas test case is chosen because it is representative for the average performance. For this test case AEdb returns 27 exon events in 18 genes. The methods not only differ by performance but also by the predicted splicing events. The commonality of the predictions is assessed by looking at the overlaps between methods. The top 250 predictions constitute about $\sim 1\%$ of all genes on the array. The commonality Table 4.3 reflects a limited overlap between the methods. For the 18 confirmed genes the ARH values and the corresponding quantiles in the ARH

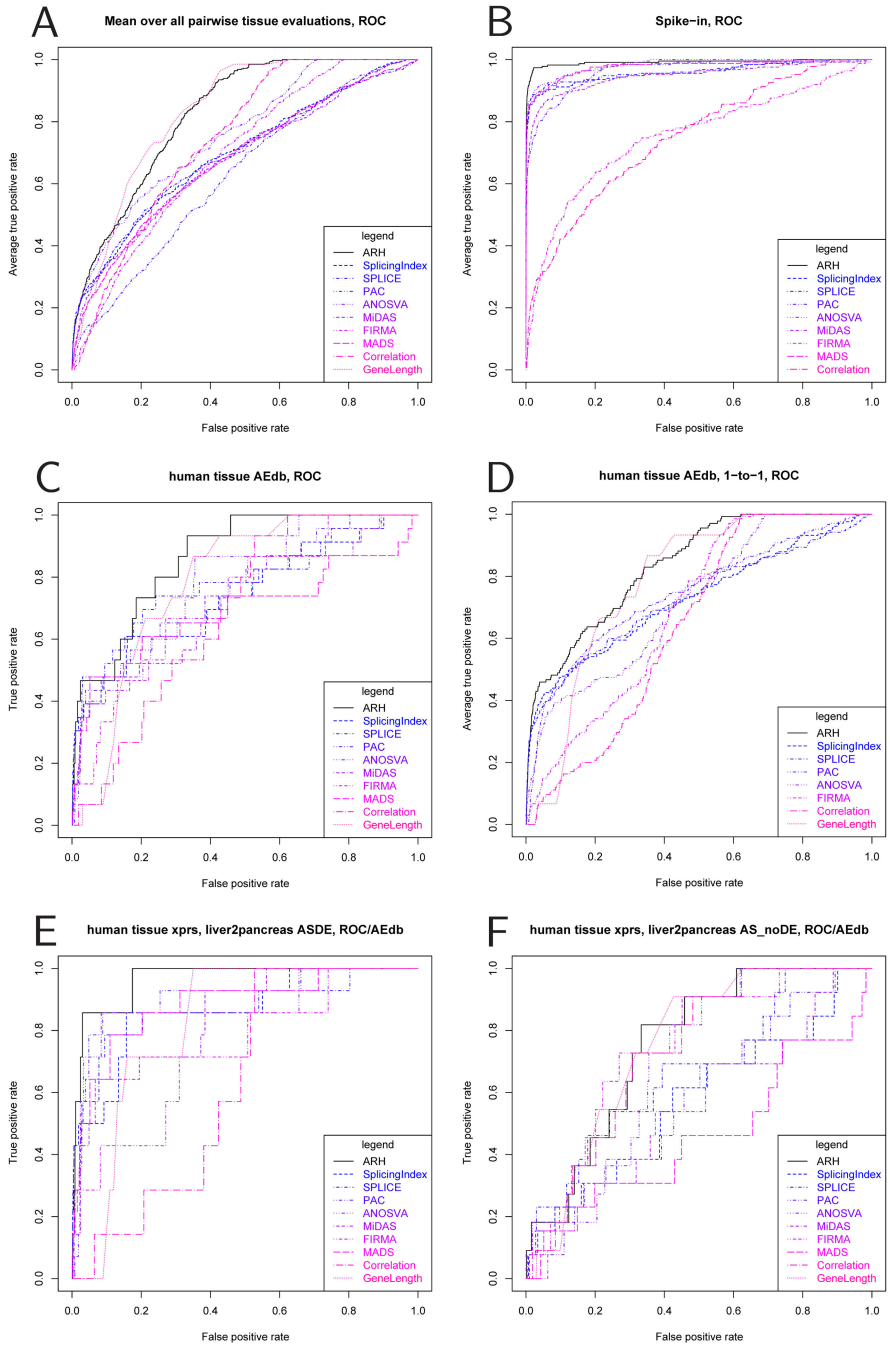


Figure 4.4: ROC curves for different aspects of methods performance. A: Overall performance across the 28 pairwise tissue comparisons with respect to AEdb confirmed splicing events (performances vertically averaged). **B:** HeLa cell line data with spiked transcripts as true positive set. **C:** Liver vs. pancreas individual splicing predictions. **D:** Performance of methods with only 1 of the 3 replicates. **E:** Performance for confirmed events in differentially expressed genes only. **F:** Performance for confirmed events in non-differentially expressed genes only.

Exon	Gene	HGNC	EN	Tissue S.	Tissue pairwise comparison
ENSE00001065029	ENSG00000159023	EPB41	29	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001427474	ENSG00000183091	NEB	212	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001072853	ENSG00000135636	DYSF	55	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00000980710	ENSG00000147573	TRIM55	11	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00000980706	ENSG00000147573	TRIM55	11	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00000980708	ENSG00000147573	TRIM55	11	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00000709651	ENSG00000035403	VCL	22	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001530953	ENSG00000149294	NCAM1	25	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001530952	ENSG00000149294	NCAM1	25	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001513560	ENSG00000017427	IGF1	11	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00000912834	ENSG00000198838	RYR3	108	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00000673711	ENSG00000198838	RYR3	108	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001540626	ENSG00000198838	RYR3	108	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00000912951	ENSG00000198838	RYR3	108	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00000673710	ENSG00000198838	RYR3	108	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001542491	ENSG00000182871	COL18A1	73	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001368345	ENSG00000182871	COL18A1	73	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001389173	ENSG00000183963	SMTN	25	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid
ENSE00001319641	ENSG00000183963	SMTN	25	muscle	brain2muscle, colon2muscle, heart2muscle, kidney2muscle, liver2muscle, muscle2pancreas, muscle2skeletalMuscle, muscle2spleen, muscle2testis, muscle2thyroid

Table 4.2: AEdb confirmed splicing events for muscle. A subset of the AEdb true positive set for muscle specificity. Abbrv.: HGNC, human genome nomenclature identifier; EN, exon number; Tissue S., tissue specificity.

	ARH	SI	SPLICE	PAC	ANOSVA	MiDAS	FIRMA	MADS	Cor.	exon No.
ARH	-	0.58	0.59	0.27	0.31	0.3	0.06	0.19	0.04	0
SI	183	-	0.77	0.28	0.33	0.27	0.07	0.18	0.03	0.01
SPLICE	185	217	-	0.28	0.35	0.3	0.07	0.2	0.03	0.01
PAC	106	108	109	-	0.13	0.16	0.06	0.13	0.03	0.002
ANOSVA	118	123	129	59	-	0.25	0.03	0.23	0.02	0.05
MiDAS	115	107	114	70	101	-	0.04	0.17	0.02	0.02
FIRMA	27	31	31	29	13	20	-	0.04	0	0.02
MADS	80	78	83	57	93	73	17	-	0.01	0.02
Cor.	18	14	13	14	8	9	0	4	-	0
exon No.	1	5	5	1	22	9	8	10	0	-

Table 4.3: Overlaps for the top 250 genes of each prediction method. In the test case liver vs. pancreas the top 250 predicted genes are compared between the methods. The lower left triangular matrix contains absolute numbers and the upper right triangular matrix contains the Jaccard index. Abbrv.: Cor., correlation; SI, splicing index.

background distribution, the p-values of the generalised extreme value fit and q-values for FDR correction following Dabney et al. [75] are listed in Table 4.4.

Furthermore, tissue specificity is analysed by comparing a selected tissue against the 10 remaining tissues (see Figure 4.5). This leads to considerable variance in the intensities for the control group. ARH is robust for such variance, as it works only with an overall control exon expression ignoring the variation. For example comparing muscle to non-muscle tissues this variance challenges the methods in their robustness for noise in the measurements and results in a strong spread of performances. For muscle, the AEdb contains 19 confirmed exon events in 10 genes (see Table 4.2). In Das et al. [80], the authors use the same human tissue data set to establish a list of muscle-enriched exons whereof 17 events have been validated with RT-PCR. Since the study was performed on an older genome build, the probe set region of the 17 events was updated with the UCSC Genome Annotations Lift Tool to the current genome build (Assembly Mar 2006) [164]. The original regions intersect with 13 Ensembl exons in 11 genes constituting the list of validated events used for analysis. Since the RT-PCR assays are generated specifically on the samples under study, the ROC are more specific than AEdb confirmed events (see Figure 4.5). It is a major advantage of ARH that it is robust to noise within the samples. The effect is exemplified in Figure 4.6 with two case studies of different prediction quality.

4.4.3 Test data set 2: Spike-in transcripts

In Abdueva et al. [5] another benchmark data set was presented with spike-in hybridisations of 24 transcripts. For genes not-expressed in HeLa cells, the mRNA is added at concentrations of 0, 2, 32, 128, 512 pM in a Latin square design by five groups (see Table 4.5). The data set has the advantage, that expression strength is exactly known in every sample. The samples have a very homogenous background such that noise can

	ARH	ARH background	ARH p-value	ARH q-value
ENSG00000005471	1.75	$3.23{\cdot}10^{-6}$	$8.12{\cdot}10^{-7}$	0.00039
ENSG00000131979	1.33	$3.23{\cdot}10^{-6}$	$1.86{\cdot}10^{-6}$	0.00065
ENSG00000135447	0.50	$2.91{\cdot}10^{-5}$	$3.31{\cdot}10^{-5}$	0.0038
ENSG00000101076	0.37	$5.65{\cdot}10^{-5}$	$8.01{\cdot}10^{-5}$	0.0065
ENSG00000148584	0.31	$8.07{\cdot}10^{-5}$	0.00014	0.0089
ENSG00000015475	0.22	0.00023	0.00038	0.017
ENSG00000171105	0.14	0.00090	0.0015	0.042
ENSG00000143257	0.11	0.0017	0.0026	0.061
ENSG00000105325	0.023	0.10	0.12	0.68
ENSG00000183337	0.020	0.13	0.15	0.75
ENSG00000142192	0.016	0.21	0.23	0.92
ENSG00000170632	0.015	0.23	0.25	0.96
ENSG00000082701	0.011	0.35	0.38	0.99
ENSG00000197965	0.0084	0.46	0.50	0.99
ENSG00000100429	0.0078	0.49	0.54	0.99
ENSG00000163606	0.0069	0.54	0.60	0.99
ENSG00000106633	0.0034	0.73	0.84	0.99
ENSG00000010932	0.00036	0.87	0.98	0.99

Table 4.4: Table of ARH values for the liver vs. pancreas confirmed events. The AEdb confirmed splicing events for the test case liver vs. pancreas are highlighted with their corresponding ARH values. In ARH are the normal linear ARH values. In 'ARH background' are the quantiles within the background distribution with no biological splicing. In ARH p-value are the p-values of the fit with the generalised extreme value distribution on the background distribution. In ARH_q are the q-values of the ARH q-value p-value distribution (q-values are fitted to all genes).

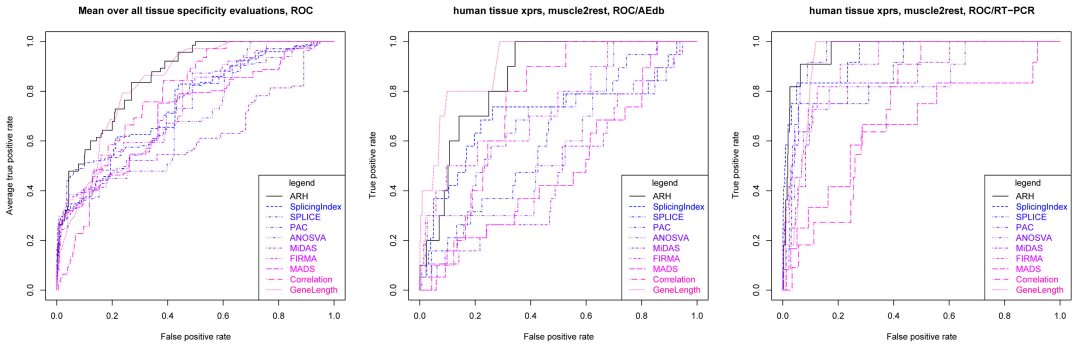


Figure 4.5: Tissue specificity. Left: Overall performance across the 8 tissue specificity comparisons with respect to AEdb confirmed splicing events (performances vertically averaged). **Centre:** Comparison of muscle vs. non-muscle tissue data invoking additional experimental noise with AEdb confirmed splicing events. **Right:** Muscle vs. non-muscle tissue data with RT-PCR validated true positive set.

be neglected. All true positives are known due to the closed collection of spiked genes. Following the original handling of the data the Affymetrix probe-probe set-transcript cluster assignment is used.

The 24 transcripts are not spliced by experiment. Generation of splicing events follows an idea of Beffa et al. [27] and reassigns exons from one spike group to a different group. The five experimental groups facilitate 10 pairwise comparisons. In each comparison exons with 0, 2, 32, 128, 512 pM are assigned to genes with 2, 0, 512, 32, 128 pM concentration respectively. The true positives in this data set are characterised by differentially expressed genes with generically spliced exons at extreme fold changes. The environment excluding the 24 transcripts has no expression change at low variability. The results show a general increase in methods performance compared to the tissue data with ARH being the best performing method (see Figure 4.4).

4.5 Discussion

Entropy was introduced to alternative splicing for quantification of global splicing disorders [247]. Fractions of transcript variants for a gene are assessed by entropy. Computing entropy ratios between conditions splicing disorders are quantified. Ritchie et al. [247] state that fraction entropies are generally higher in cancer tissues compared to normal tissues. This effect is observable for proper alternative splicing, not for alternative transcription start site and not for alternative polyadenylation. It is the first study with a quantitative estimate of splicing disruption in cancer.

Entropy in this study is computed on the fractions of transcript variants. Thus entropy distribution depends on the number of possible transcript variants and possibly results also may depend on the exon number. Unfortunately the authors of Ritchie et al. [247] do not argue about such a possible dependency. Thus it is not sure if the entropy ratios in the study not also show any tendency to high-number genes as discussed here in

HGNC	TC ID	Group	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5
AAK1	2558150	1	0	2	32	128	512
COPS4	2733928	1	0	2	32	128	512
MRPS5	2564599	1	0	2	32	128	512
RUNDC3A	3722872	1	0	2	32	128	512
TRIM55	3101514	1	0	2	32	128	512
EDNRB	3518766	2	512	0	2	32	128
GALK2	3593339	2	512	0	2	32	128
KCNH6	3730698	2	512	0	2	32	128
KRT7	3415320	2	512	0	2	32	128
SEC22B	2355615	2	512	0	2	32	128
ARL6IP2	2548776	3	128	512	0	2	32
C1orf187	2320392	3	128	512	0	2	32
NOSTRIN	2514216	3	128	512	0	2	32
POU2F2	3863435	3	128	512	0	2	32
SNTB2	3666601	3	128	512	0	2	32
GFRA1	3308241	4	32	128	512	0	2
GLYATL1	3331822	4	32	128	512	0	2
MRS2	2898452	4	32	128	512	0	2
SERGEF	3365136	4	32	128	512	0	2
SNX24	2826343	4	32	128	512	0	2
INHBA	3047581	5	2	32	128	512	0
ARD1B	2774900	5	2	32	128	512	0
PAX9	3532793	5	2	32	128	512	0
SLC39A14	3089360	5	2	32	128	512	0

Table 4.5: The spike-in transcripts latin square scheme. For not present genes in HeLa cells, the transcripts are spiked into the RNA sample. The numbers in the experiment columns refer to pM for spike-in clone concentrations. Abbrv.: HGNC, human genome nomenclature ID; TC ID, transcript cluster identifier; Exp., experiment.

subchapter 4.5.2.

In my work entropy is introduced to prediction of alternative splicing. Exon expression ratios are calculated directly between conditions and later assessed by entropy. Also ARH accounts for entropy distribution by subtraction from maximal entropy.

4.5.1 General performance of methods and study design

The prediction of alternative splicing remains a challenge. In general, performance of all methods is not very good, in particular with respect to the tissue data set. This is due to the fact that splicing prediction poses particular problems to the statistical analysis. A gene encodes several transcripts on the one hand and consists of different exons on the other hand. For each product or each exon a separate analysis is performed to test potential splicing using the same measurements in several tests. Approaches for the comparison of methods can be found in Purdom et al. [237] with a simulation model and in Beffa et al. [27] with the re-ordering of spike-in data. The advantage of the human tissue data set is the challenge to identify splicing events in a non-artificial, experimental setting.

The confirmed events used for this study are in any sense independent from the computations. This has effects on the performance of the methods with respect to two aspects. On the one hand the confirmed events are not generated from the tissue samples on the chips. Thus, some of the AEdb splicing effects may be weak or not appropriate. For example, if the splicing event is confirmed in one tissue but the isoform is not tested specifically in the second tissue, then this result would turn into a false positive in the light of the experimental data because the isoform may be present in both tissues. On the other hand, strong splicing differences can be expected between tissues. The number of confirmed splicing events is low concerning recent predictions of up to 95% of spliced human multi-exon genes [228]. With the AEdb there are only a few events of unknown strength. The methods may predict successfully many real, existing events before marking the confirmed events. These aspects may in part explain low sensitivity of the results. In particular the test settings only allow relative rating of the methods, not computation of overall performance. In particular the test settings only allow relative rating of the methods, not computation of overall performance.

A major advantage of ARH is the robustness concerning noise within the samples and exon expression variability along the gene. This is probably the explanation of the performance spread of the methods in the tissue specificity settings in Figure 4.5. To elucidate this in detail the focus is on the muscle vs. non-muscle case with two case studies in Figure 4.6. The prediction ranks for the case studies are in Table 4.6. Where *HNF4A* in Figure 4.9 is an illustrative example easy to predict the genes *CSDE1* and *DYSF* have different prediction quality in the methods. Another advantage of ARH concerns experimental design, not only prediction performance: ARH provides reliable predictions for few replicates. This aspect is presented in Subsection 4.2.2. All of the analysed methods are suitable for exon skipping prediction. This is by choice and adequate to the Affymetrix Exon Array platform. Of course all of the methods are successful to identify

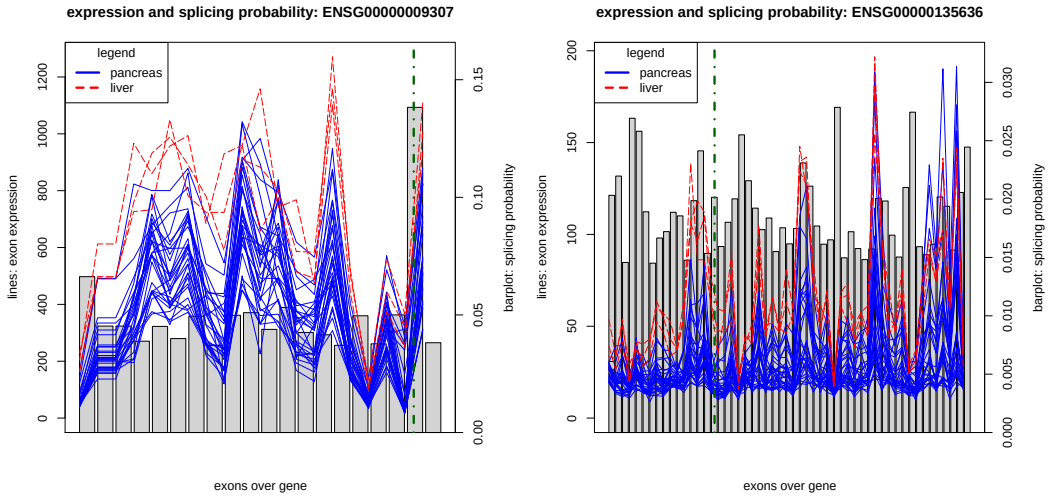


Figure 4.6: Expression and splicing probabilities of *CSDE1* and *DYSF*. The two examples are from the muscle vs. non-muscle test setting. In both examples is a very high noise level within the samples along the replicates. **Left:** The left example *CSDE1* is a true positive from the RT-PCR validated study. The ARH values is 0.47 for *CSDE1*. **Right:** The right example *DYSF* is an AEdb confirmed event and ARH value is 0.058 for *DYSF*. The true positive exon in *DYSF* (vertical green dotdashed line) has an ARH splicing deviation value of 0.474 and is thus significant on the 0.1 level (see subsection 4.2.2).

	HNF4A			*CSDE1*			*DYSF*		
	rank	total	ratio	rank	total	ratio	rank	total	ratio
ARH	231	26 538	0.0087	71	26 538	0.0027	653	26 538	0.025
SI	266	232 376	0.0011	121	232 376	0.00052	11 742	232 376	0.051
SPLICE	279	232 376	0.0012	7491	232 376	0.032	9861	232 376	0.042
PAC	7401	232 376	0.032	725	232 376	0.0031	14 488	232 376	0.062
ANOSVA	638	26 538	0.024	2774	26 538	0.1	199	26 538	0.0075
MiDAS	1343	232 376	0.0058	19 095	232 376	0.082	5183	232 376	0.022
FIRMA	2173	26 538	0.082	3	26 538	0.00011	716	26 538	0.027
MADS	1527	232 376	0.0066	56 528	232 376	0.24	21 421	232 376	0.092
Cor.	5511	26 538	0.21	6780	26 538	0.26	6519	26 538	0.25
Exon No.	3892	26 538	0.15	2227	26 538	0.084	236	26 538	0.0089

Table 4.6: Ranks and quantiles for different case studies. For every method three columns are introduced: In 'rank' is the minimal rank of the method sorted for decreasing splicing indication (low rank is better), 'total' the total number of ranks, i.e. exon level or gene level prediction as well as 'ratio' the quantile of the prediction, the rank divided by the total number of ranks. The third column is comparable between all predictions. The expression and indication for *HNF4A* is visualised in Figure 4.9 and *CSDE1* as well as *DYSF* in Figure 4.6. Abbrv.: SI, splicing index; Cor., correlation; Exon No., exon number.

splicing events. The study is novel in the field by relating the new proposed method ARH in competition to the existing methods.

4.5.2 Predictors vs. number of exons in the gene

In differential expression settings the number of probes are mostly constant across the genes on the array. This is not true anymore with the exon arrays. Predictions are calculated for genes with strongly differing number of exons. Ideally, a method is independent on the number of exons in a gene. The performance of the different methods is investigated with respect to this feature. The genes were partitioned in bins referring to exon numbers. Boxplots for the distribution of the predictions were calculated per bin and are shown in Figure 4.7. Here, genes with the same number of exons were assigned to the same bin. With increasing number of exons the probability of a false positive prediction increases. Focussing on the exon level does not avoid the problem. Sorting the predictions by decreasing splicing indication, genes with high number of exons are still preferred.

A majority of the methods shows a dependency on the number of exons. Especially statistical tests are susceptible to the increasing splicing indication with increasing exon number. Statistical tests become sensitive with increasing exon number and detect decreasing splicing differences. In the AEdb test setting this misleadingly improves performance. In order to make the ARH gene level prediction independent of the number of exons per gene the entropy values were compared to their possible maximum. This maximum is only dependent on the number of exons and thus constant over the exon bin. Thus, ARH corrects for the number of exons per gene.

Interestingly, Figure 4.4 and Table 4.1 demonstrate that the number of exons per gene is per se already a well-performing splicing prediction exceeding several of the computational methods. This may be a consequence of exon number bias in the AEdb compared to genome-wide data from Ensembl annotation. In the Ensembl database gene number distribution decreases with increasing exon number in the genes with a mean of 13 exons per gene. The AEdb, in contrast, shows a fairly differing distribution of number of exons with a peak between 7 and 18 exons per gene and a mean of 25. This observation reflects a selection bias of the manual curation. The AEdb genes show increasing coverage of Ensembl genes with increasing exon number as visualised with the ratio of AEdb exon number bins divided by the Ensembl exon number bins in Figure 4.8.

4.5.3 Alternative splicing and differential expression

Alternative splicing in the presence of gene expression changes demands the methods to account for differential expression. Affymetrix, and several other groups, conceptualise the ideal splicing events without differential expression [20]! Looking at the liver vs. pancreas setting differentially expressed genes are filtered as denoted in Subsection 3.2.5. From 23 confirmed exon events 12 events are in 6 differentially expressed genes and 11 events are in 9 genes without considerable expression changes. The DE pipeline identifies

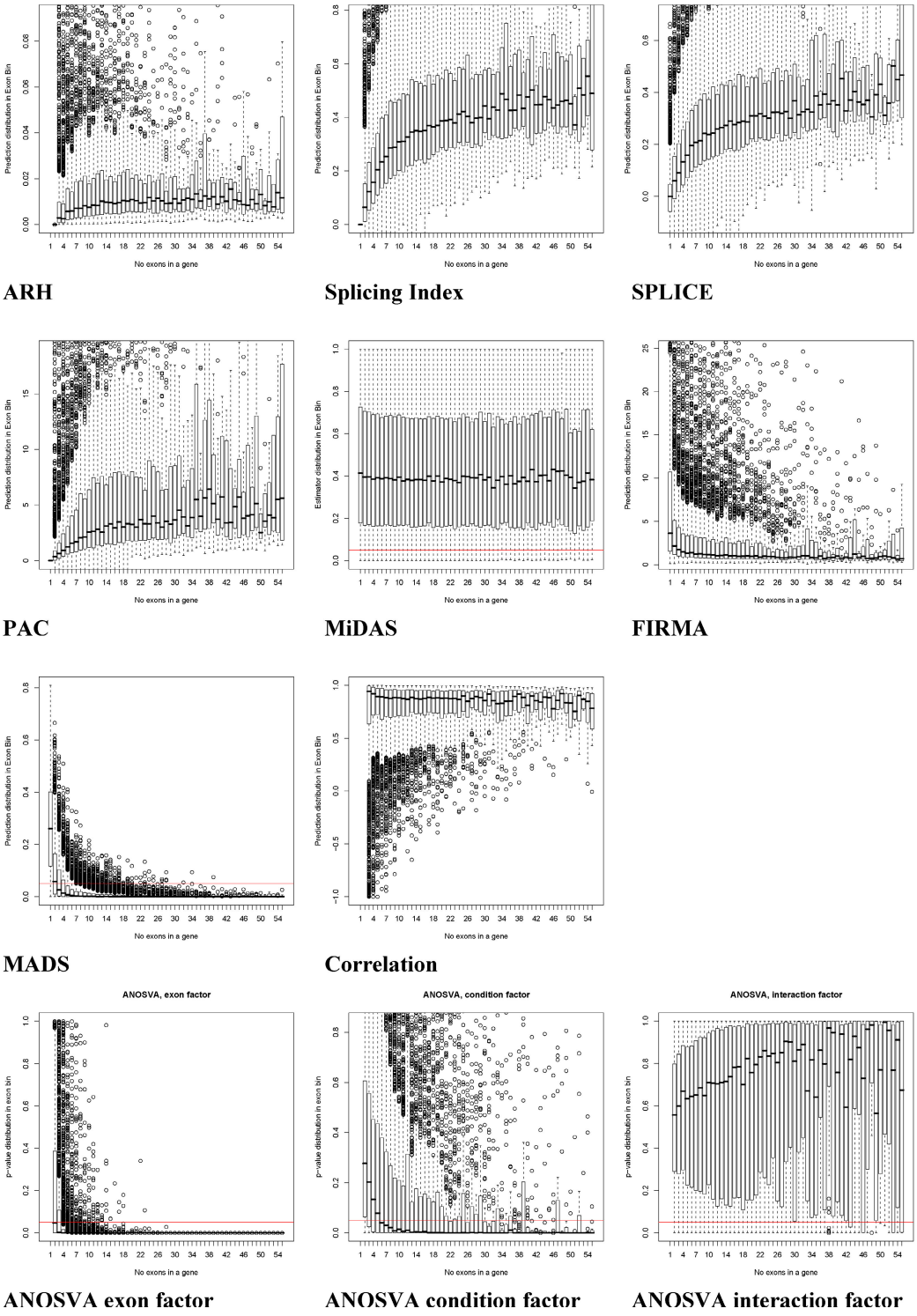


Figure 4.7: Dependency of the different methods on the exon number. Boxplots of prediction values (y-axis) with respect to genes with the same number of exons (x-axis). For the statistical tests the red horizontal line indicates an 0.05 p-value.

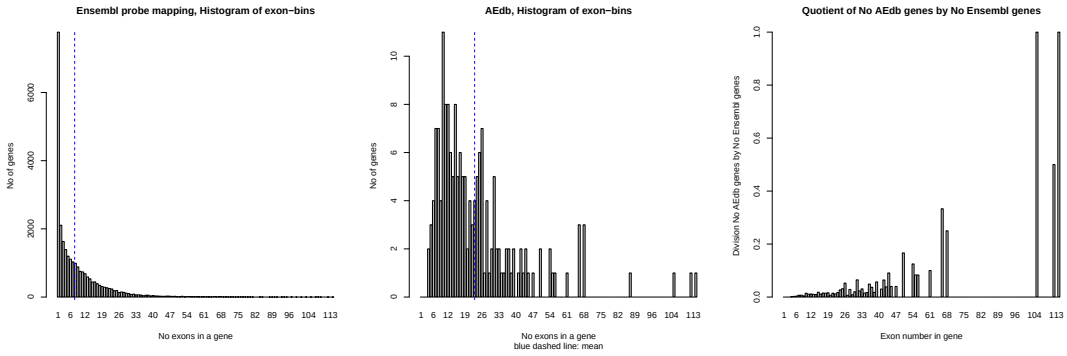


Figure 4.8: Exon number statistics in assignments. Genes are divided in bins by the number of assigned exons. The histogram represents the number of genes in a bin. The blue, dashed line indicates the mean of the exon number. **Left:** The Ensembl probe mapping comprises all genes and corresponding exons from Ensembl 49 with assigned probes following the mapping of Dai et al. [76] in version 11. **Centre:** From the AEdb all cassette exon events with known sequence are filtered. The event sequences are compared to the Ensembl exon sequences and filtered for perfect matches. For the resulting exons the corresponding Ensembl genes are drawn from Ensembl 49. **Right:** For every bin the number of genes in the AEdb histogram is divided by the number of Ensembl genes. This illustrates that genes in the AEdb tend to have more exons than genes have in general, i.e. a bias towards high exon number genes.

4615 differentially expressed genes and 963 alternatively spliced genes with an intersection of 719 genes. These numbers imply that alternative splicing cannot be modelled without differential expression. A coupling of the transcription and the splicing machinery is indicated in several studies [111, 206, 272]. However there is no genome-wide assessment of this intersection.

The true positive set is split into the 12 differential expression events and 11 non-differential expression events and subsequently the ROC curves are computed using only the two subsets of confirmed events. Surprisingly all methods perform better to identify splicing with differential expression (see Figure 4.4). In case of no differential expression several methods are challenged to deviate alternative splicing from random. How do methods take differential expression into account? Splicing Index calculates a normalised intensity dividing the exon expression by its gene expression. ANOSVA introduces a condition factor in its statistical test. In ARH the median of the $\log_2$ ratios is subtracted.

4.5.4 Predictions with two arrays

Since the costs for the arrays are always an issue in academic settings methods favourably require low number of replicates. The methods shall be robust in the number of arrays. Purdom et al. [237] are the first to address this aspect for FIRMA. Here it is addressed by computing the ROC curve for only one chip per condition (see Figure 4.4). In liver vs. pancreas are three chips in each condition allowing nine pairings of single chips. MiDAS and MADS require replicates and are excluded from this test. ARH predictions are only dependent on the robustness of exon expressions. Using the median over the probes but

also over the replicates the method is robust in the number of replications.

4.5.5 Exon expression variability

Exon expressions are variable across the gene. Figure 4.9 elucidates the complex nature of exon expression. In previous gene expression experiments this variability was taken into account to be noise in the probes. The exon arrays point to a deeper transcription pattern in terms of splicing. Similar expression variability is found in RNA-Seq data. This indicates that the driving element of expression variability is not just probe intensity variance.

The variation can be decomposed into several dimensions: For example variation in replicates, within exon/gene probe variation as well as exon expression variation within a gene. The replicate variation is low compared to the within exon or gene variation as already pointed out in the preprocessing in Figure 3.13. The probe variation within exons is lower than within genes (see Figure 4.10). These two aspects already show that intensities or exon expressions are probably not drawn from a single gene population.

For seven of the human tissues RNA-Seq is available from Wang et al. [306]. Reads from the public available Illumina sequencing lanes are aligned to Ensembl database[1]. Reads aligning within Ensembl exons denote an expression level called the exon read count. The third dimension of variation is compared on the gene variation over exon expressions using array expression and read counts (see Figure 4.10). The coefficient of variation in RNA-Seq data is even bigger than in exon arrays, in general for about 90% of the genes. Thus, the exon expression variability seems not to be a technological artefact.

In summary, the assumption that all exons in a gene have the same expression does not hold in general. Thus, a uniform distribution cannot be assumed. Similar observations led Shah and Pallas [259] to the identification of the correlation as an indicator for splicing. ARH has been shown to cope with variable exon expression. Taking the ratio of the exon expressions between the biological conditions levels out the expression changes. The logarithm to the base 2 of the ratios saliently reflects splicing peculiarities in the exon expressions. With the entropy ARH weights the expression ratios to each other, identifying genes with deviating ratios.

4.6 Approaches with negative results

The track of research on the prediction methods was bordered by several trials with less success. Major approaches are described to indicate possible pitfalls for follow-up studies.

Statistical tests have the nice attribute to increase power with increasing data. Increasing data is in return a consequence of increasing exon number. Statistical tests for a given

[1] Tissues were aligned against all ENSEMBL 53 human cDNA sequences using bowtie v0.9.9.3 allowing 2 mismatches within first 24 bp and maximum number of mismatch maq like qualities of 70 [179]. For each tissue aligned reads were then counted into their specific contig (in this case transcripts) afterward, not keeping track of the uniqueness. Reads are not normalised according to the length of the exon but a simple scaling in terms of using the overall reads of the two states was performed.

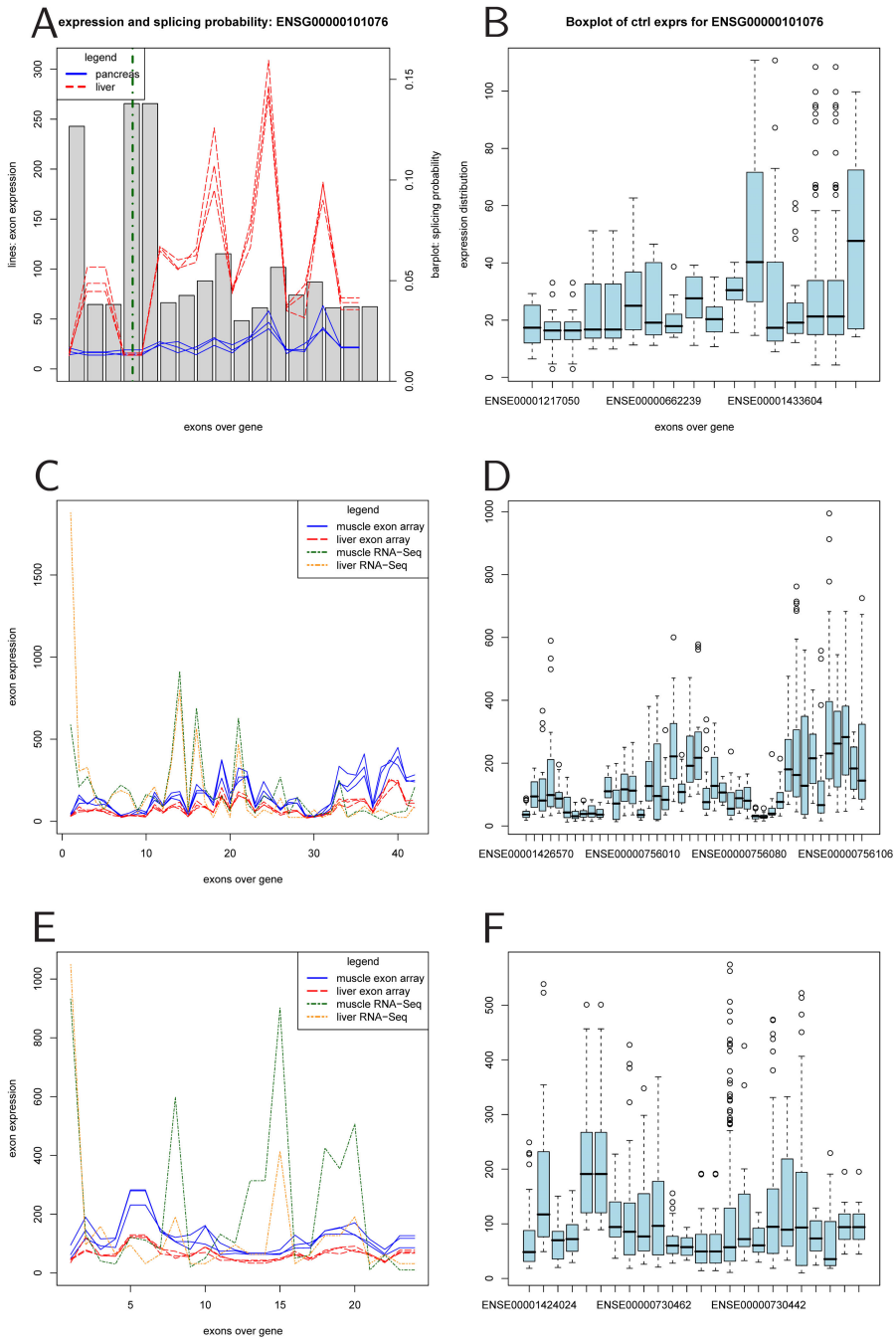


Figure 4.9: Singular examples for tissue exon expression. A: *HNF4A* is a gene with a confirmed splicing event between liver and pancreas (exon 4, green dot-dashed line). The lines (y-axis, left scale) show the exon expressions ordered by genomic position (x-axis). The bars (y-axis, right scale) correspond to the splicing probability values of the respective exons. The ARH value for *HNF4A* is 0.37, corresponding to a p-value of $8.01 \cdot 10^{-5}$. **C, E:** The lines in red and blue show the exon expressions ordered by genomic position and in green and orange show RNA-Seq exon counts. To compare the expression measures on the scale sequencing counts are divided by their median and multiplied with the array expression median. **B, D, F:** For one tissue the distribution of probe intensities is depicted as exon-wise boxplots.

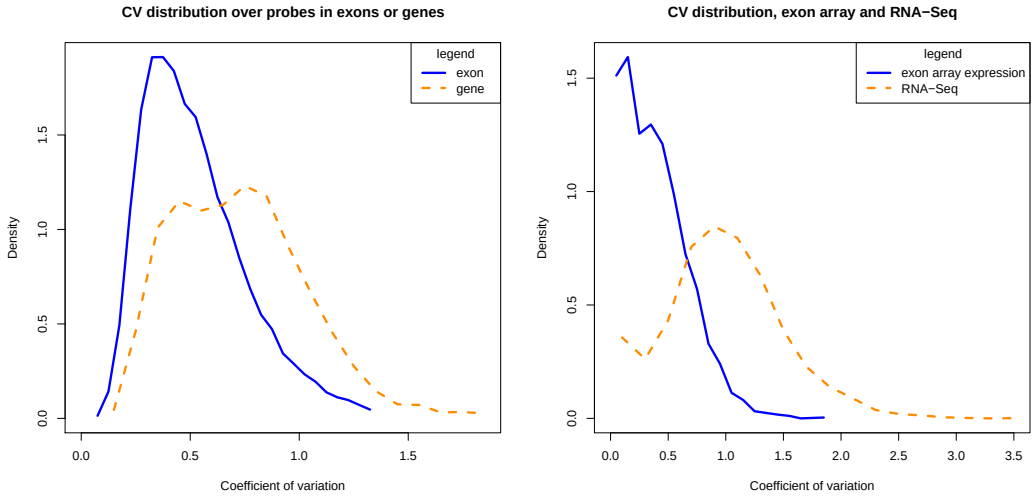


Figure 4.10: Different dimensions of variation. Variation is quantified with the coefficient of variation. **Left:** The probe variation within the exons is in blue and within the genes in dashed orange. **Right:** The exon expression variation within genes is compared between technologies. In blue is the gene variation in exon arrays and in dashed orange the variation in RNA-Seq.

p-value threshold tend to be significant for gene with more exons. Genes with same exon number can be sorted in bins. Only for the genes in one bin the tests are comparable. This problem is inherent to statistical tests and attempts to circumvent this problem did not work (correction of z-score, divide test estimate by its exon number bin average).

One branch of development was an ANOVA test on the differences of probe intensities between biological conditions. Despite of considerable performance of the test the development was stopped due to a severe exon number dependency. Performance of the test could not be separated from the performance of the exon number estimator.

Normal ANOVA works with the assumption of a gaussian distribution within the data. One attempt was to use more robust tests with rank methods like the Kruskal-Wallis test. The robustification indeed lead to better performance in the method evaluation. Unfortunately this improvement was on cost of specificity. The robust tests were far to sensitive with arbitrarily small p-values.

The idea behind such a test was to use probe level data instead of summarised exon and gene expressions. Probes correspond to the level of the measurements free from any summarisation effects. Therefore analysis could start one level lower before summarisation. This idea is supported by the observation of improved accuracy for probe level analyses [188, 192, 193, 242]. However, for the prediction methods there was no advantage using probe level data.

With a variety of prediction methods available in Section 4.3 it could be straightforward to combine available predictions, joining the strengths of different methods improving the performance. Predictions were ranked from best to worst splicing indication. Using the ranks commonality predictions are defined by applying geometrical means, e.g. the

best rank for every gene/exon. Several means like the sum of the ranks, the mean rank or the euclidean norm on the ranks were applied. These combination methods had average performance compared to the original methods.

Now ARH is applied on linear data. Logarithm is often recommended in gene expression analysis (see Section 3.2). ARH and other methods were also applied on logged data without improvement. Computation of the entropy directly on $\log_2$ ratios lead to severely increased sensitivity also for minor ratio changes. Thus the $\log_2$ is only used for the distinction of up- and downsplicing. After this distinction, ratios are linearised robustifying the proposed method.

ARH corrects for exon number by subtraction from the maximal entropy. The theoretical goal is to compare the entropy in terms of length of the random vector. Two types of normalisation are proposed in information theory: (1) the normalised entropy with division by maximal entropy and (2) the entropy rate with division by the random vector length. Both types of normalisation were not successful. This observation indicates a major importance for the prediction in the deviation from maximal entropy. Just the deviation seems to be invariant from exon number.

Finally, in the method evaluation the number of confirmed or validated events is quite low for all of the three test data sets. For the tissue data set one of the tissue EST databases was selected from Subsection 2.2.1, the T-STAG/SpliceNest database [118, 72]. The database predicts splicing events from the EST collections and comprises a high-throughput prediction in itself. Although the qualitative statements about the performance of the methods were similar, the ROC and AUC characteristics did not convince. Thanks to a suggestion of Dr. Stefan Haas the method evaluation was switched to the AEdb with manually curated splicing events [274].

5 Alternative Splicing in Type-2 Diabetes Mellitus

In Chapter 5 the power of the pipelines and the ARH prediction is shown with an application in the context of type-2 diabetes mellitus. An introduction to type-2 diabetes mellitus elucidates the interplay of different organs and factors to highlight two major organs for disease progression, adipose tissue and liver.

For marker identification a meta-analysis is performed on diverse qualitative and quantitative sources Rasche et al. [240][1]. The quantitative sources are microarray data sets processed with the differential expression pipeline. In every source disease relevance of genes is scored. Scores are summed up to a gene score rating the general relation of a gene to type-2 diabetes mellitus. Assessing consistency in the gene score another use arises of the entropy introduced in Section 4.1. High entropy identifies genes with consistent type-2 diabetes mellitus relevance over many sources.

For two mouse models of type-2 diabetes mellitus hybridisations on exon arrays have been performed at the German Institute of Human Nutrition (DIfE). Mice are all fed on a high-fat diet and on this dietary background *NZL* animals develop obesity. Diabetic mice are separated by levels of blood glucose. In contrast the *SJL* animals do not develop obesity by genetic reasons. Samples of fat and liver tissue are prepared and with the alternative splicing pipeline spliced genes are identified and attributed to glycaemic or genetic causes.

5.1 Biology and genetics of type-2 diabetes mellitus

Type-2 diabetes mellitus (T2DM, formerly called noninsulin-dependent diabetes mellitus (NIDDM), or adult-onset diabetes) is a disorder that is characterised by high blood glucose in the context of insulin resistance and relative insulin deficiency. While it is often initially managed by increasing exercise and dietary modification, medications are typically needed as the disease progresses.

The polygenic nature of type-2 diabetes mellitus is now well established and several mouse models including *NZO*, *BTBR* etc. have been studied to analyse diabetes susceptibility on a complex genetic background [64]. Linkage analyses have shown that several quantitative trait loci interact with each other and with the environment to elicit obesity

[1] Parts of this Chapter appear in the Handbook of Research on Systems Biology Applications in Medicine edited by Dr. Andriani Daskalaki [239]; Copyright 2009, IGI Global, www.igi-global.com. Posted by permission of the publisher.

(blood glucose mmol / l)	Fasting		2h after 75g oral glucose load
Diabetes mellitus	≥ 7.0	or	≥ 11.1
Impaired glucose tolerance	< 7.0	and	7.8 – 11
Impaired fasting glucose	6.1 – 6.9	and	< 7.8
Normoglycaemia	< 6.1	and	< 7.8

Table 5.1: Diagnostic classification of type-2 diabetes mellitus. Diagnostic criteria of diabetes mellitus and other categories of hyperglycaemia [283].

syndromes that are potentially diabetic. Several recent genome-wide association studies have identified novel candidate genes for type-2 diabetes mellitus but the effect of these variants on disease susceptibility is generally low, with odds ratios mostly around 1.5 [98, 114, 253, 256, 266, 276, 323]. Multiple studies on the transcriptome level have been performed that emphasise the diversity of the disease and the complex pathophysiological interaction between different tissues, including fat, muscle, liver, pancreatic β-cells and brain [283]. In several human studies, tissue biopsies from diabetic and normoglycaemic individuals have been profiled [117, 207]. In mouse studies differences in diet or mouse strains have been used to identify distinct expression profiles [35, 211, 177]. Complementary ChIP-on-Chip studies reveal the associated gene regulatory network of important transcription factors active in the relevant tissues [219, 220]. In the context of the onset of diabetes, several studies on the proteomic level have revealed differential expression of intracellular proteins as well as of secretory proteins in adipose tissue [58, 291]. Despite the availability of these large amounts of data, their common content as well as their specific differences, in particular in gene sets between human and rodent studies, has not been systematically evaluated until recently with Rasche et al. [240].

5.1.1 Diabetes mellitus

Abnormally high level of glucose in blood are the main characteristic of diabetes mellitus [83, 283]. Healthy people mediate blood glucose, whereas in diabetics glucose levels remain high. Insulin regulates the blood glucose level. In diabetes insulin is not produced at all, insufficiently or not as effectively as needed. Most common forms are type-1 diabetes (5% of the cases, an autoimmune disorder) and type-2 diabetes (95%, obesity associated). Some rare variants exist, e.g. by single gene mutations.

Type-2 diabetes mellitus generally occurs in obese adults. Many underlying factors contribute to high blood glucose levels. Resistance of the body to insulin is an important factor, ignoring its insulin secretions. Therefore type-2 diabetes mellitus is a combination of deficient secretion and deficient insulin action. The rise of obesity in the population is the driving force behind the increase of diabetes. Today it can be difficult to maintain healthy body weight in the presence of abundant food and a sedentary life.

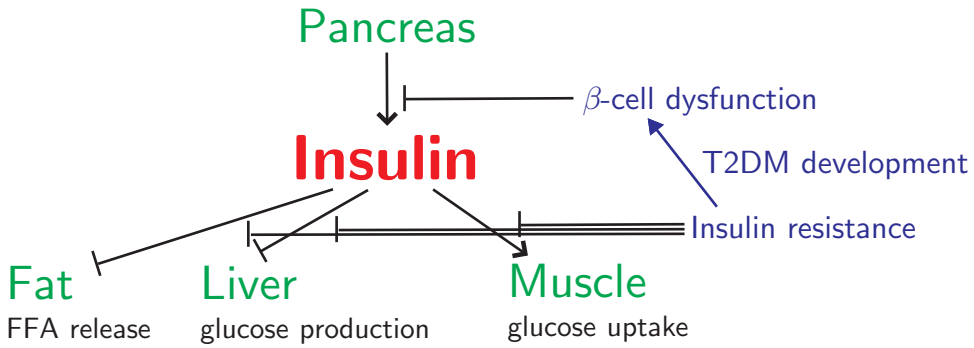


Figure 5.1: Insulin as main regulator of energy balance. The transition from insulin resistance to insulin production deficiency is a major aspect in type-2 diabetes mellitus progression. The progression is accelerated by factors like adipokines, inflammation, glucose overflow and free fatty acids. FFA denotes free fatty acids.

Diagnosis

Being overweight or obese is defined by looking at the Body Mass Index:

$$\text{BMI} = \frac{\text{weight in kg}}{(\text{height in m})^2}. \tag{5.1}$$

A Body Mass Index of 18-25 is healthy, 25-30 overweight and above that level obese.

Diabetes mellitus is diagnosed on the basis of WHO recommendations from 1999, including two criteria: fasting glucose and 2h after 75g oral glucose load (see Table 5.1) [283, 269]. Criteria are combined into a practicable diagnostic classification. Impaired fasting glucose and impaired glucose tolerance are conditions predisposing overt diabetes mellitus. A substantial part of people with these problems will progress to overt diabetes if not treated [283].

5.1.2 Physiology

Describing the physiology of type-2 diabetes mellitus means talking about the energy control in an organism. Three molecules are the substantial interactors regulating the energy supply in the body [83, 283]:

Glucose is an essential energy source for the body.

Insulin is the regulator of circulating glucose levels and energy balance (see Figure 5.1). Insulin increases uptake of glucose into fat and muscle tissue, and formation of glycogen in the liver and skeletal muscle.

Glucagon is the opponent of insulin and rises in scarcity. Glucagon activates glycogen breakdown. Glucagon also helps the body to use alternative resources such as fat and proteins.

Blood glucose levels are variable depending on the needs of metabolism, rising for three reasons: diet, breakdown of glycogen or hepatic synthesis of glucose. Glycogen is a short-term energy reserve generated from glucose. Glucose is regulated by insulin and some

other hormones. Glucose abundancy releases insulin from pancreatic islets β-cells and stimulates the following:

- liver to store glucose as glycogen;
- muscle to absorb glucose from the blood and store glucose as glycogen and
- cells to convert glucose in ATP.

Fasting results in reduced blood glucose level, leading to lower insulin and higher glucagon. Glucagon raises blood glucose by calling of glycogen from the liver short term reserve and glucose production by converting amino acids in the liver. Glucagon level is stimulated by several causes, like protein-rich food or stress. When fasting for some time, the liver is exhausted by glycogen but continues to make glucose from amino acids.

Tissues

In every tissue inside the cell, glycolysis uses some of the glucose. Glycolysis is a central pathway of carbohydrate metabolism which occurs in all body cells and releases energy and carbohydrate intermediates for use in metabolism. Four tissues are of major importance for glucose management:

Liver produces and consumes glucose and buffers glucose levels. It is one of the most important organs in this interplay. From digestion the liver receives glucose-rich blood and removes large amounts of glucose to mediate the blood glucose level.

Fat stores energy as fat. Fatty acids are assembled to triglycerides.

Skeletal muscle stores energy as glycogen.

Pancreatic islets Pancreatic α-cells produce glucagon and pancreatic β-cells produce insulin. Pancreatic β-cells detect the rise of blood glucose and respond with the release of insulin and at the same time α-cells lower the release of glucagon and thus the production of glucose from other sources.

5.1.3 Pathogenesis

Impaired insulin sensitivity and peripheral insulin resistance are central factors in the pathogenesis of type-2 diabetes mellitus [269, 283, 284]. In Insulin resistance a normal insulin concentration returns a subnormal biological response. In carbohydrate metabolism insulin insensitivity leads to insufficient glucose usage in muscle and fat as well as increased glucose production in the liver. In protein and fatty acid metabolism insulin insensitivity leads to decreased intracellular uptake of amino acids and increased lipid breakdown and thereof increased free fatty acids.

In insulin resistance hepatic glucose production is insufficiently suppressed. Insulin insensitivity is balanced by the β-cells with additional insulin production. Eventually the β-cells are not able to produce and secrete a sufficient amount of insulin probably due to genetic defects. Long term glucose overflow (hyperglycaemia) leads to decreased sensitivity of the β-cells and to its apoptosis. Free fatty acid overflow (hyperlipidemia) provokes insulin secretion and consequently the decline of insulin storages. Abnormal levels of free

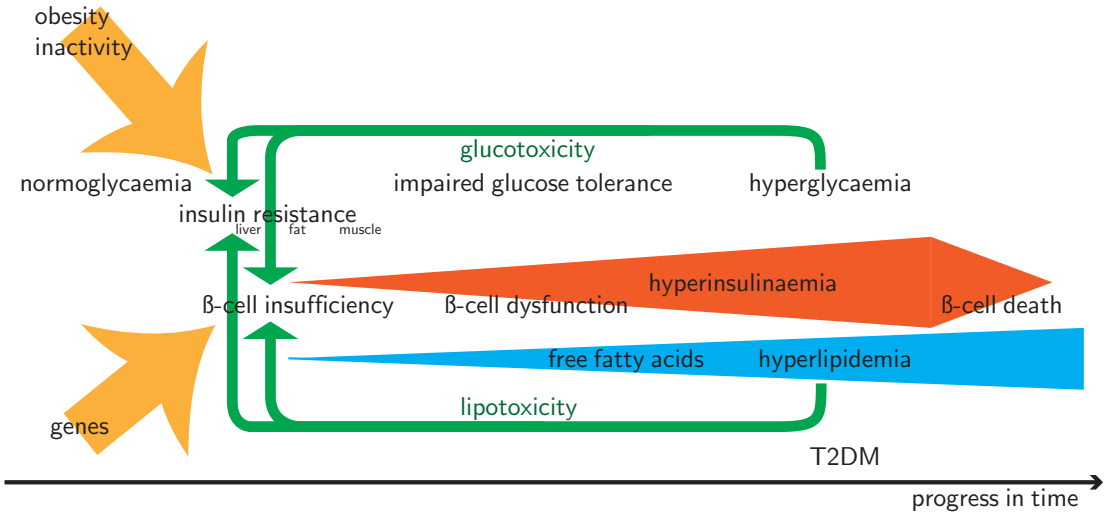


Figure 5.2: Pathogenesis of type-2 diabetes mellitus. The figure is sort of a visual linear approximation of the disease progress in type-2 diabetes mellitus. Peripheral insulin resistance and β-cell insufficiency lead to increased insulin and free fatty acid levels. Advanced insulin resistance leads to impaired glucose tolerance. Glucotoxicity and lipotoxicity constitute cycles worsening the basic insulin resistance and β-cell dysfunction. Abbreviations: IR, insulin resistance; hyperglycaemia, glucose overflow; glucotoxicity, toxic effects of glucose; hyperlipidemia, free fatty acid overflow; lipotoxicity, toxic effects of free fatty acids; hyperinsulinaemia, insulin overflow.

fatty acids are a consequence of impaired fatty acid metabolism. Due to massive increase of fat in type-2 diabetes mellitus patients the free fatty acid release is increased. Free fatty acids accumulate in muscle and disturb carbohydrate metabolism as well as glucose uptake aggravating insulin resistancy. Furthermore free fatty acids increase the glucose production in liver.

Open questions

However, most genes and their associated molecular network contributing to the onset and course of the disease are yet unknown. An understanding of the interplay between obesity and insulin resistance is crucial but not completely resolved [64, 21, 89, 210].

There is a strong correlation between obesity and diabetes. In fact, for every kilogram gained on a population level, diabetes rates increase linearly [305]. Many details are unknown how enlarged fat mass causes insulin resistance [153, 221]. Obesity is associated with an increase in adipocyte secretion of chemokines, which promote macrophage infiltration. In addition to increased macrophage infiltration, obesity is associated with increased macrophage activation. Activated macrophages produce cytokines that can negatively impact on insulin sensitivity. Does insulin resistance cause inflammation or vice versa [79]? One major hypothesis is, when maximal fat tissue expandability is reached the inflammation is started due to a stress reaction [116]. Inflammation in adipose tissue promotes insulin resistance in different organs, like liver, by several pathways [221, 251].

Muscle cells ingest major amounts of glucose via the *Slc2a4* transporter mediated by insulin. In type-2 diabetes mellitus translocation is impaired of *Slc2a4* to the cell membrane and accounts for the malfunctioning glucose usage. Exact molecular characterisation of the defect was not possible yet [269].

In addition it was not possible to identify the molecular defects leading to β-cell dysfunction despite intense research on the complex intramolecular network of the insulin secretion cascade. Pancreatic islets try to compensate for higher insulin requirements by increasing β-cell mass. Maximal cell mass seems to be genetically bounded and after reaching this maximum β-cell mass declines [154, 209]. This cell mass decline is caused by apoptosis of β-cells and followed by decreased insulin capacity. One hypothesis sees glucose overflow leading to excessive deposit of reactive oxygen species causing β-cell death [283].

In summary type-2 diabetes mellitus has a complex pathogenesis with insulin insensitivity in different organs and β-cell secretion aberration. This effects are amplified by disturbed fatty acid metabolism.

5.1.4 Genetics

The role of genetics in type-2 diabetes mellitus is indicated by the familial clustering of insulin sensitivity and insulin secretion, the higher concordance rate of type-2 diabetes mellitus in monozygotic vs. dizygotic twins and the high prevalence of type-2 diabetes mellitus in certain ethnic groups [84]. Concordance rates were 88% in monozygotic twins compared to dizygotic twins for impaired glucose tolerance. A positive family history is related to a 2.4 fold increased risk [283]. Two main strategies seek to identify genetic factors: the genome-wide scanning and the candidate gene approach. In genome-wide scanning for the same species genotypes are compared to each other to narrow down disease related regions. In the candidate gene approach gene sequences of physiologically important proteins are compared among population samples.

Mutations of a single gene can result in disease. This happens in rare forms of diabetes. Such mutations can be investigated with sequencing to find the responsible SNP in the DNA. Type-2 diabetes mellitus is assumed to be polygenic. Disease genes may show subtle but common differences in the gene sequence. It is difficult to link these common gene variations to an increased risk of developing type-2 diabetes mellitus. Therefore it is a remarkable result, that microarray study results converge on the same functional modules by deriving metabolic pathways from expression results [299].

Genome-wide scanning

Association studies are performed on patient cohorts raised over years. Genotyping microarrays isolate chromosomal regions or SNP [256, 266, 276, 323, 98, 114, 253]. Positive associations are found in one or more studies. However the following functional characterisation or positional cloning of causative genes has mostly been unsuccessful [84].

The major genome-wide association studies settled down on 15 genomic loci with 20 candidate genes reviewed for example in Doria et al. [84] or Stumvoll et al. [285].

In the linkage approach the genome of affected family members is compared using genetic markers. This locates genes by the rationale, that family members not only share the phenotype but also chromosomal regions surrounding the involved gene. Alterations are combined with the family genealogy over several generations and affected sibling pairs linking parts of the genome to the risk of developing diabetes.

The genome-wide scan is also used in mice. So-called quantitative trait loci are isolated through backcrossing between susceptible and unsusceptible strains. Such studies demand a much smaller number of individuals as in the human case due to genomic homogeneity of in-bred strains [231]. It is easier to follow or direct the family history in animal models. The genetic component may be linked to the expression level combining expression microarrays and genotyping arrays [176]. Thus, it integrates two different information levels and results in narrow genomic candidate regions.

Genetic linkage and association studies often have poor replicability. Because of the late onset of type-2 diabetes mellitus, susceptibility gene variants may exist in the control group and reduce the power of the studies. Beside some more factors are attributed to low replicability like ethnic stratification or gene-by-gene and gene-by-environment interactions.

Candidate gene approach

In the candidate gene approach specific genes are selected preferably due to their functional role in type-2 diabetes mellitus [83, 230, 24]. In unrelated individuals, the statistical association of an allele and the phenotype is tested. But also the candidate gene approach had minor success in identifying causative factors. Variants were extensively analysed in many candidate genes but the initial association could mostly not be replicated in follow-up studies [225].

The candidate gene approach is scientifically more simple focussing on disease status and alleles or haplotypes in insulin signalling or glucose metabolism. Dean and McEntyre [83] as well as Parikh and Groop [230] describe work and results performed on the most promising candidates. An exhaustive collection of genome regions and assured genetical factors is provided by OMIM under the identifier #125853 [225]. A consequent step after selecting candidate genes is to generate animal models by genetic manipulation. Such models are reviewed in Nandi et al. [214] or Clee and Attie [64].

5.1.5 Animal models

Type-2 diabetes mellitus affects the basic metabolic process and therefore is traceable in all organisms from human over mice and rats down to *caenorhabditis elegans*, where the most relevant pathways are found as ageing pathways. Collection of human tissue samples demands the cooperation of many medical institutions. In addition, nutrition and lifestyle are not under control in contrast with lab animal models. With animal models more

finegrained study designs are possible by controlled environments with regard to nutrition and lifestyle but also genetics. Time series for different disease states and backcrossing experiments for genetical insight are further features. In the following some strains are presented. Although rat models for type-2 diabetes mellitus exist like the *Zucker Diabetic Fatty rat* or *Zucker Fatty rat* the focus remains on mouse models with view on the generated data sets in Subsection 5.2.1 and Section 5.3.

A complete overview about the mouse models used in type-2 diabetes mellitus research is provided by [64] highlighting the history of the mouse strains and their susceptibility to impaired glucose tolerance or type-2 diabetes mellitus. For example the *C57BL/6* is the most important mouse model accounting for 14% of all experiments. It shows diabetes-susceptible and diabetes-resistant aspects [300]. With a so-called *ob* mutation in the *Leptin* gene, the same mouse strain becomes obese and develops hyperglycemia. These mice compensate insulin resistance by making more β-cells and insulin. Most other strains except *C57BL/6* do develop diabetes with *Leptin* defect. The *BTBR* strain shows strong diabetes-susceptibility. Crossings of *BTBR* and *C57BL/6* are more glucose intolerant than either parental strain, suggesting interactions between strain specific alleles. Insight into the metabolism and insulin resistance drawn from mouse models is described in Nandi et al. [214]. The authors break down the plurality of knock-out and transgenic mice by phenotypes and tissue to find unsuspected players, e.g. transcription factors, which emerge from the underlying studies.

The project in mind needs an animal model reflecting the human metabolic syndrome. Such a model is the *New Zealand obese* (*NZO*) mouse strain, an in-bred polygenic mouse model [37, 36]. This strain is characterised by several features:

- Morphologic changes in the pancreas with an altered number and size of the islets,
- early signs of insulin resistance in the first weeks in white and brown adipose tissue, muscle as well as liver,
- hepatic glucose overproduction and impaired first-phase insulin secretion [303] and
- similar characteristic of the human metabolic syndrome like hypertension elevated cholesterol, hyperglycemia and hyperinsulinemia [226, 73].

The *NZO* mouse separates from the *New Zealand black* mouse, its nearest relative, by reduced body temperature, increased nutrition intake in portion and frequency as well as lower activity [152].

The *NZL/Ltj* (*NZL*) mouse strain was developed in the Jackson Lab in 2004 with a genome consisting of 96.88% *NZO* and 3.12% *New Zealand black* genome. The new *NZL* strain develops obesity with severe hyperglycaemia, as known from the *NZO* strain. Prevalence for hyperglycaemia is higher with 79% of the *NZL* mice compared to 50% in *NZO* at week 20 and for blood glucose more than 250 mg/dl [202, 3]. Therefore the *NZL* mouse differs weakly from the *NZO* strain with main differences higher rate of hyperglycaemic animals and a higher reproduction rate. For backcrossings performed in the *NZO* mouse with a genetically different, completely lean and non-diabetic mouse the *Swiss Jackson laboratory* (*SJL*) strain is mostly used. With *NZO* and *SJL* crossings several loci have been localised for type-2 diabetes mellitus in the mouse genome [233, 57].

In summary, major findings can arise from a variety of organisms to understand human metabolism. A caveat is, that the disease in the animal models may have causes different from the human setting. So the results lack comparability and have to be reproduced in different models. All of the described mice models are available from the Jackson Labs including the focussed type-2 diabetes mellitus models [3].

5.2 Marker identification for type-2 diabetes mellitus by meta-analysis

The goal of this meta-analysis approach is to generate additional value by combining individual studies and by extracting consistent information [240]. Several meta-analysis studies have been previously applied within other disease domains, such as cancer [244] or Alzheimer [33] using different types of data [68]. With respect to type-2 diabetes mellitus some recent approaches have been published: In Tiffin et al. [298] several computational prediction methods have been combined in order to identify a common set of type-2 diabetes mellitus genes. The authors assessed the accordance of the prediction methods resulting in a candidate gene list of 99 different genes. For type-1 diabetes mellitus a web-resource tracks the expression behaviour of genes in several tissues [267]. Liu et al. [191] applied enrichment analysis to previously defined gene sets and protein-protein interactions using data from different species and tissues from the Diabetes Genome Anatomy Project [1] and identified a subnet of insulin signalling proteins and nuclear receptors. In contrast to Liu et al. [191] or Rhodes et al. [244] the presented approach is not limited to transcriptome studies. Data is accumulated from different levels of molecular interaction such as genetic information using knock-out mice and SNP, gene regulatory and gene expression information as well as information on protein signalling and protein interactions. In order to reduce technical bias of transcriptome measurements this data type is restricted to experiments that were performed on the Affymetrix GeneChip platform. Involving several parts of the body a common signature for type-2 diabetes mellitus cannot be found with a single tissue. Therefore, similar to Liu et al. [191], relevant tissues are combined such as liver, muscle, adipose tissue and pancreas. Although mouse models are available for aspects of the disease, it is unclear, whether these mice have diabetes for the same reason as humans do. With a span over human and several mouse models a more global view of the disease is generated.

Using a Bootstrap [49] scoring approach a core set of 655 genes is computed that shows significant disease relevance in the data sets under study. Here, the gene expression profiles are used along with qualitative data comprising literature, genetic and proteomic sources. Besides known genes this approach exhibits a large fraction (499) of yet barely characterised novel candidate genes. These genes have been further validated in the functional context of networks and exhibit high potential for understanding pathways and pathway crosstalk associated with type-2 diabetes mellitus. Gene set over-representation analyses infers the deranged parts of the physiology by type-2 diabetes mellitus using gene ontology terms [26], common pathway resources [151, 158, 249] and information on

week	liver	muscle	fat
4	2 HFD / 3 SD	2 HFD / 2 SD	2 HFD / 3 SD
5	1 HFD		1 HFD
7	1 HFD		1 HFD
8	2 HFD / 4 SD	2 HFD / 2 SD	2 HFD / 4 SD
9	1 HFD		1 HFD
10	1 HFD		1 HFD
11	1 HFD		1 HFD
12	2 HFD / 4 SD	2 HFD / 2 SD	2 HFD / 4 SD

Table 5.2: Study design of the ESGEC data set. The data set has three dimensions: time, tissue and diet. Abbrv.: HFD, high-fat diet; SD standard diet.

the associated gene regulatory network [219, 220, 201].

The Section follows the work presented in Rasche et al. [240] but comprises new data sets specific for the alternative splicing analysis in Section 5.3. The project data is introduced in the next Subsection 5.2.1, it is focussed on time series correlated to disease stages as well as early stage expression alterations on the *NZO* mouse.. Different sources are joined in Subsection 5.2.2 and the identification of the marker set follows in 5.2.3. Considering many years of research on diabetes candidate genes relate to diverse established resources. Here the marker set is related to networks in Subsection 5.2.4.

5.2.1 Early stage gene expression changes

DNA microarrays have been used to dissect various aspects of type-2 diabetes mellitus reviewed by Sun [290]. Inside physiologic and pathologic conditions transcriptomics permits a more comprehensive understanding of gene sets involved in the mechanisms of type-2 diabetes mellitus. In human studies are conducted for example in adipose tissue returning disperse results, i.e. linking genes to lipid and glucose metabolism, membrane transport and promotion of the cell cycle. In skeletal muscle probably the most important finding is the upregulation of the *oxidative phosphorylation* pathway in accord with rat results [207].

A variety of studies applies microarrays to animal models and cultured cells. In type-2 diabetes mellitus they returned a tremendous amount of information about the pathophysiology. Studies *in vivo* and *in vitro* profiling adipocytes from intra-abdominal and subcutaneous adipose tissue lead to coordinated depot-specific differences in expression of genes in embryonic development and pattern specification. Diet effects alter the expression of hundreds of genes primarily related to lipid metabolism and transcription factors in adipocyte differentiation. In rat skeletal muscle the activation of the *nutrient-sensing hexosamine biosynthesis pathway* decreased genes involved in oxidative phosphorylation confirming results from human studies. In mouse assays unfortunately all of the mice were at least 14 weeks of age and thus provide little insight into the early phase of pathogenesis.

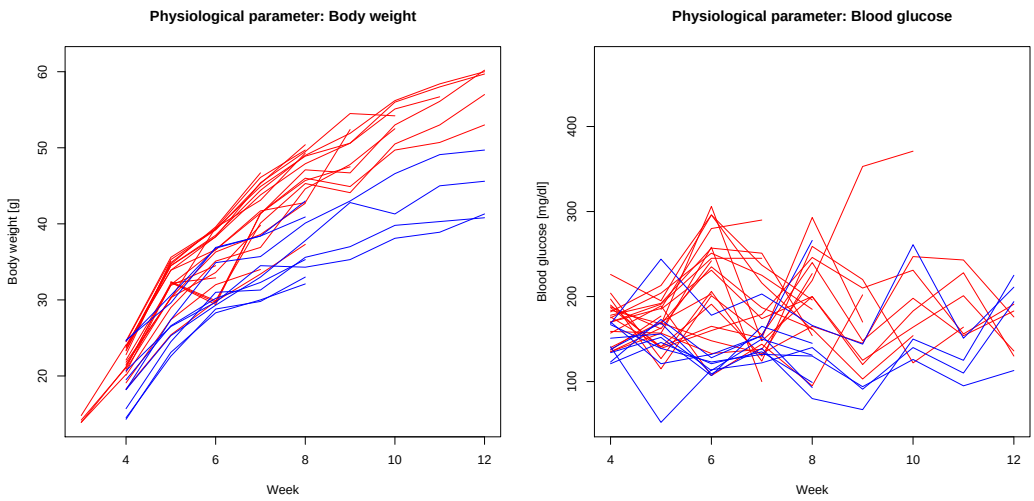


Figure 5.3: Physiological parameters for the ESGEC mice. Red curves visualise the development of high-fat diet animals and blue curves of standard diet animals. **Left:** On the x-axis is the time in weeks and on the y-axis the body weight in gram. **Right:** On the x-axis is the time in weeks and on the y-axis the blood glucose level in mg/dl.

Microarrays allow the categorisation of disease stages according to the changes in the mRNA expressed. This is used in parallel in several tissues at the same time or in the same tissue at several time points. The *NZO* mouse shows the main characteristics of the human metabolic syndrome as introduced in Subsection 5.1.5. Project data was generated on the *NZO* mouse with the aim to characterise early stage gene expression changes and the acronym ESGEC — Early Stage Gene Expression Changes — refers to this data set in the remain of the work. Study design follows the idea to choose time points before and during obesity development to identify candidate genes and metabolic processes contributing to adiposity.

The ESGEC data set comprises samples with variation in three dimensions: time, tissue and diet and is outlined in Table 5.2. Male *NZO* animals were separated from its mothers in week 3 and set on the respective diets, high-fat diet and standard diet, and characterised up to week 12. High-fat diet contains 15.3 MJ/kg by 32.5% sugar, 17.1% protein, and 14.6% fat and standard-diet contains 12.8 MJ/kg by 36.5% starch, 19% protein, 4.7% sugar and 3.3% fat among others. Characterisation includes body weight, body fat, blood glucose, blood plasma insulin and cytokines. Dr. Tanja Dreja characterised and prepared the samples in the group of Dr. Hadi Al-Hasani at the German Institute of Human Nutrition (DIfE) in Potsdam-Rehbrücke [85]. After preparation the samples were passed to a company for hybridisation on Affymetrix mouse 430 2.0 arrays. This data set complements the study on the exon arrays in Section 5.3 by identifying a marker gene set for the *NZO* and *NZL* model mice.

For comparison of diet changes the data listed in Table 5.2 facilitates six simple cases. Differential expression results at week 8 and 12 in each of the three tissues between

high-fat and standard diet. As high-fat samples are underrepresented at week 8 and 12 exactly the samples are joined from week 7 to 9 and from week 10 to 12 respectively. Additionally, a time series analysis is performed over all available time points in each tissue on the high-fat diet. In the time series expressions are correlated with the weeks of age representing disease stage for the mice.

5.2.2 Mapping, preprocessing and categorisation of data

Data sets were selected from heterogeneous sources that target different levels of cellular information. Data categories are either binary or quantitative. An overview about the type-2 diabetes mellitus specific data sets is given in Table 5.3.

Binary data was introduced by incorporating medical reviews, phenotype information (for example from knock-out genes), results from proteome analysis [283, 58, 83, 225, 204, 3, 214] as well as published candidate gene lists from previous studies or models [298, 1, 230, 168]. Binary information was assigned according to the fact whether the gene had been identified in the study or not.

Quantitative data was incorporated by evaluating data from differential gene expression and time series microarray studies [117, 207, 35, 177, 211]. In order to extract comparable information across the different studies data is restricted to the same technological platform (Affymetrix GeneChip studies). Furthermore, in order to conduct standardised data normalisation only studies were taken into account that published and provided the raw data (CEL file level). Each individual microarray study was normalised using the pipeline described in the chapters 3.2 with customCDF in version 8.

For transcriptome studies that are targeting differential expression three bits of information are stored – the fold-change indicating the alteration of the gene when comparing the diabetic state with the normal state, the standard error of the fold-change computed from the replicated experiments in that study and the expression p-value (presence-call) that indicates whether or not the gene is expressed in the target samples under study. In time series studies the correlation is stored between phenotypic characteristics, for example blood glucose, and the gene expression levels with the coefficient of variation and the expression p-value.

However, some issues are to consider applying and integrating microarrays. The total number of probes on a microarray and the selection of the probe sequences from the gene/transcript sequences differs between the chip manufacturers and therefore hinders comparability. Data analysis is complex due to the large amount of genes and possible study designs. Therefore, the focus is on case-control and time series *in vivo* studies on Affymetrix platforms in this project as well as in the public studies.

A central pre-requisite of any meta-analysis approach is the consolidation of the different identifier types, for example coming from different organisms and from different versions of chips. The Ensembl database was used as the backbone annotation for all studies (see Figure 5.4) [38]. Any identifier are mapped on their mouse Ensembl gene identifier (version 44). Mapping and merging of the information has been done within R and the BioConductor package collection [238, 108]. The total number of genes under study

Data set	Category	Species	Tissue	Study research question	Cit.
StumvollGoldstein2005	Qualitative	human		medical review about T2DM	[283]
DeanMcEntyre2004	Qualitative	human		medical review about selected candidate genes	[83]
OMIM	Qualitative	human		medical review about T2DM	[225]
PubMedGeneRIF	Qualitative	human/ mouse		text mining in the NCBI geneRIF	[204]
KOmiceJAX	Qualitative	mouse		mouse models with phenotype T2DM	[3]
NandiAccili2004	Qualitative	mouse		mouse models with phenotype Insulin Resistance	[214]
ChenHess2005	Qualitative	rat	fat	Secretory proteins in adipose tissue	[58]
SundsenBergsten2006	Qualitative	human	blood	differential protein expression	[291]
MoothaGroop2003	Quantitative	human	muscle	patients with T2DM/impaired glucose tolerance and controls	[207]
GuntonKahn2005	Quantitative	human	pancreas	patients with T2DM vs. controls	[117]
LanAttie2003	Quantitative	mouse	fat/muscle/liver/ pancreas	diabetic mice vs. controls	[177]
BiddingerKahn2005	Quantitative	mouse	fat/muscle/liver	diabetic mice vs. controls	[35]
NadlerAttie2000	Quantitative	mouse	fat	diabetic mice with different level of hyperglycaemia	[211]
ESGEC 2006	Quantitative	mouse	fat/muscle/liver	diabetic mice vs. controls week 8 and 12	(5.2.1)
ESGEC 2006	Quantitative	mouse	fat/muscle/liver	time series week 4 to 12	(5.2.1)

Table 5.3: Overview on the data sets used for the T2DM meta-analysis. The table lists the heterogeneous set of data sets and resources combined in the meta-analysis. Abbrv.: T2DM, type-2 diabetes mellitus.

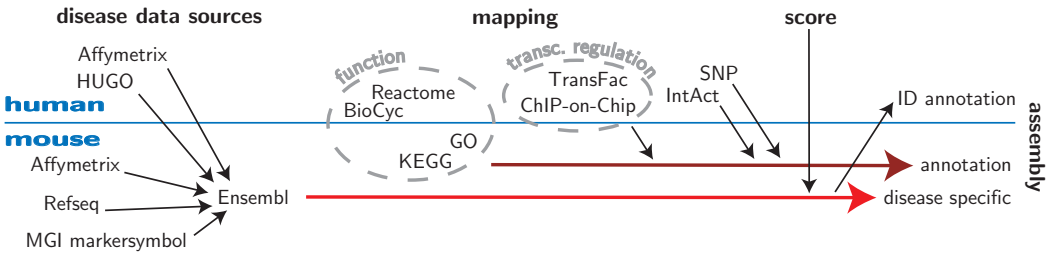


Figure 5.4: Identifier mapping scheme. For the different type-2 diabetes mellitus data sources the identifier are mapped to Ensembl mouse gene identifier. Then follows score calculation and for public use annotation of different identifier. With a similar procedure the functional annotation and gene set resources are mapped to Ensembl mouse for subsequent analysis. The assembly is the tabular output of the collected source data annotated with gene sets as well as score and entropy. Abbrv.: ID, identifier.

comprises 18 439 Ensembl mouse genes representing the union of the homologue genes from all data sources.

5.2.3 Identification of marker genes – generality vs. specificity

Numerical scores were computed for all genes in each individual study, the scores were summarised and the summarised scores were compared against a random sample at the 99.9 percentile. This procedure determines a cut-off score value of 4.3 and identifies a set of 655 genes with a score exceeding this cut-off. In the following this procedure is explained in four steps. (1) Generating gene scores, (2) isolate candidate genes by resampling, (3) relate to different candidate gene studies and (4) account for consistency in gene scores.

Scoring type-2 diabetes mellitus relevance of genes across studies

In order to score the different categories of information, i.e. binary counts and quantitative gene expression values, for each category the scores are summarised of the individual experiments. For binary information the counts were grouped in sub-categories, for example knock-out mice described in two reviews only get a single count.

For quantitative information, the score of the g-th gene in the u-th study, $s_{g,u}$, was computed as follows:

$$s_{g,u} = \begin{cases} |\log_2(f_{g,u})| \left(1 - \frac{\zeta_{g,u}}{f_{g,u}}\right) (1 - p_{g,u}), & p_{g,u} \leq 0.1 \text{ and } \frac{\zeta_{g,u}}{f_{g,u}} \leq 1 \\ 0, & \text{else} \end{cases} . \tag{5.2}$$

Here, $f_{g,u}$ is the fold change, $p_{g,u}$ is the average detection p-value and $\zeta_{g,u}$ is the standard error of the ratio derived from the experimental replicates of the study. Thus, the fold

change is weighted with its reproducibility across the experimental replicates and with the likelihood of the gene being expressed in the study's target samples. A similar formula applies for correlation studies:

$$s_{g,u} = \begin{cases} |c_{g,u}| \cdot v_{g,u} \cdot (1 - p_{g,u}), & p_{g,u} \leq 0.1 \\ 0, & \text{else} \end{cases} . \tag{5.3}$$

Here, $c_{g,u}$ is the correlation to a certain phenotypic parameter, $v_{g,u}$ the coefficient of variation of the gene's signal across experimental replicas. The formula is applied on the data of Nadler et al. [211] and ESGEC time series project data. In Nadler et al. mice from three different strains (*B6*, *BTBR* and F2 intercrosses) are separated in five classes with increasing hyperglycemia. The Kendall rank correlation between the classes and the gene expression is calculated.

The total score of the gene was computed as the sum across all individual study scores. A common approach in meta-analyses is to apply the same statistical test to congeneric studies and combine the resulting p-values by the Fisher method or z-Score. This is not practicable in the case at hand, since the experimental sources are too heterogenous and not for every study integrated a p-value can be calculated.

To contrast the different data sources a correlation heatmap is provided in figure 5.5. In order to measure the dependency of the scoring method on published data – particularly review papers – the correlation is computed between the scores derived from the qualitative and quantitative data. The correlation is 0.08 indicating that the transcriptome data is rather independent of the published review knowledge. In the 'qualitative' category of the study, comprising reviews/OMIM, knock-out models and PubMedGeneRIF [283, 58, 83, 225, 204, 3, 214, 230], 507 genes are related with the disease. Only a small proportion (108 corresponding to 16.5%) of those genes were also found in the type-2 diabetes mellitus candidate list, so that the computed scores do not replicate literature knowledge to a dominating extent.

Using a leave-one-out cross-validation with these studies the significance is measured of the overlap of each of these studies with the candidate list. The qualitative studies are the benchmark for the scoring approach. The scoring is calculated without the respective qualitative study. The hypergeometric distribution of the qualitative study gene set and the notional candidate set assigns a p-value. This p-value reflects the success of the score to identify the genes from the qualitative study. For all of the qualitative reference sets highly significant p-values are computed (for example Stumvoll et al. [283]: $1.1 \cdot 10^{-13}$, Dean and McEntyre [83]: 0.011, OMIM [225]: 0.0026, PubMedGeneRIF [204]: $1.83 \cdot 10^{-50}$).

Sampling for significance

Summarised scores are compared against a random sample at the 99.9 percentile. This procedure determines a cut-off score value of 4.3 and identifies a set of 655 genes with a score exceeding this cut-off. Randomly, one would expect 18 out of the 18 439 genes to exceed the threshold. Cutting at the 99 percentile results in 1808 genes (expecting 184 by

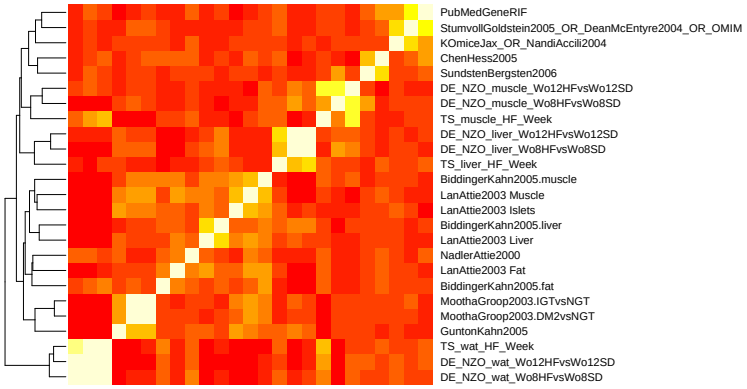


Figure 5.5: Correlation of data sources. The image visualises the scorepoint correlations between the sources as a heatmap with light yellow for maximal correlation. Both categories of sources are intermingled.

chance), cutting at the 98, 97, 96 and 95 percentiles would result in 2412, 2811, 3125 and 3403 selected genes (305, 553, 738 and 922 randomly expected genes). Thus, the ratio of detected vs. expected significant scores increases with percentile of the random sample from 3.7 to 36.4, indicating a necessary precondition for the validity of the selection procedure, see Figure 5.6.

In order to assess the significance of the overall gene scores generated gene scores are computed. For this bootstrap [49] random score are drawn from each study. The sum of the drawn study scores determines the score for a virtual gene. The distributions of the original scores (black line) and the random scores (blue dotdashed line) are shown in figure 5.6. Using the random distribution as background sample genes are assigned to be "significant" that are above the 99.9 percentile of the background distribution.

Overlap to previous predictions of type-2 diabetes mellitus genes

In the original publication Rasche et al. [240] a first marker list with 213 genes was established with the same method. That first study excludes nine microarray data sets derived from the ESGEC data and one proteomic study from Sundsten et al. [291]. However both studies overlap by 184 marker genes and therefore the additional data sets are a specific refinement of the marker set (see Figure 5.7). The score vectors over the genes correlate with a Pearson correlation of 0.78 and a Kendall rank correlation of 0.6.

From seventeen genes in the OMIM description of type-2 diabetes mellitus (Diabetes mellitus, noninsulin dependent, #125853, [225]) four genes have a significant score in this study: *Retn*, *Gpd2*, *Vegfa* and *Akt2* (see Table 5.4). *Retn* represents an adipocytokine which has been implied to play roles in obesity, diabetes, and insulin resistance [278, 321]. Interestingly, *Retn* is only deregulated in one of two studies involving adipose tissue. In contrast, differential expression for *Vegfa* was observed in pancreatic islets whereas *Gpd2*

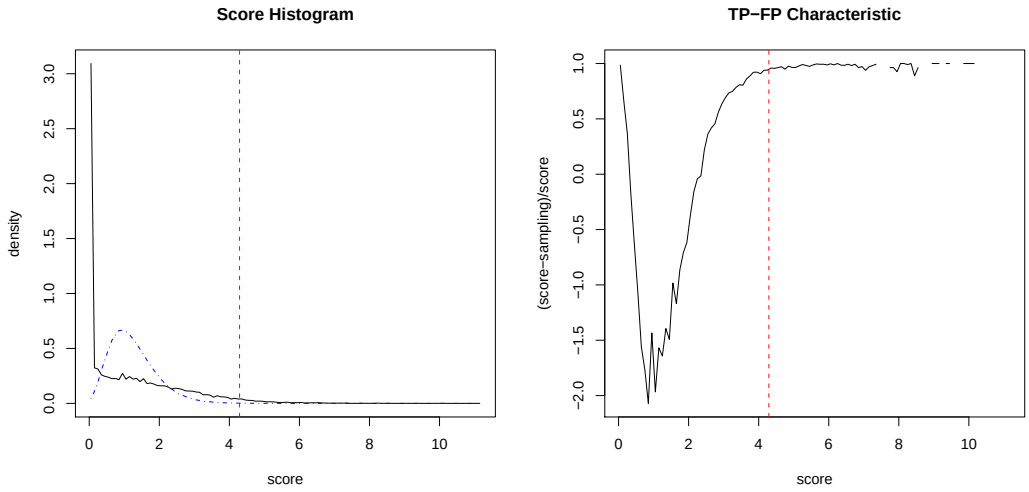


Figure 5.6: Score histogram & TP-FP characteristic of the score. Left: Histogram of gene scores (black line) and background distribution of scores derived from bootstrap sampling (blue dotdashed line) [49]. The vertical red-dashed line marks the cut-off for the type-2 diabetes mellitus candidate gene list. **Right:** Compare the sampled false positive rate to the assumed true positives. The image is derived from the left figure. It shows on the y-axis the formula $y = \frac{x-S}{S}$ with x the gene score and S the sampled background distribution at x.

did not show tissue-specific expression.

Several previous studies already published type-2 diabetes mellitus candidate lists allowing us to assess common content. The overlap to the list of the Diabetes Genome Anatomy Project [1], being also the source of some of the transcriptome data sets used for this meta-analysis [117, 207, 35], results in a p-value of 0.002. Using the same resource, with a less conservative selection of data sets, Liu et al. identified 82 genes related to insulin signalling with an overlap of 18 genes to the candidate list containing several strongly connected proteins [191]. More selective is a review of sequencing candidates by Parikh and Groop [230] leading to a p-value of $3.1 \cdot 10^{-9}$. In Tiffin et al. [298] 99 candidates were published as partial overlap of several electronic candidate prediction methods. This results in a p-value of $1.8 \cdot 10^{-8}$ comparing it with the marker list. In summary, the type-2 diabetes mellitus candidate gene list includes a small amount of candidate genes from previous studies and, further, leads to an additional set of 589 genes not identified in the other studies. Subtracting those genes for which disease information is available from the incorporated reviews the presented approach identifies 499 novel type-2 diabetes mellitus candidate genes.

Biological validity of the type-2 diabetes mellitus candidate set is assessed by comparison to existing studies and disease gene repositories such as OMIM and genome wide association studies. The union of the medical reviews [283, 83], genetic sources [225, 3, 214] and the PubMed hits [204] contains 507 genes with an overlap of 108 genes (16.5%) to the candidate genes. However, at present only a few genes from genome-wide association

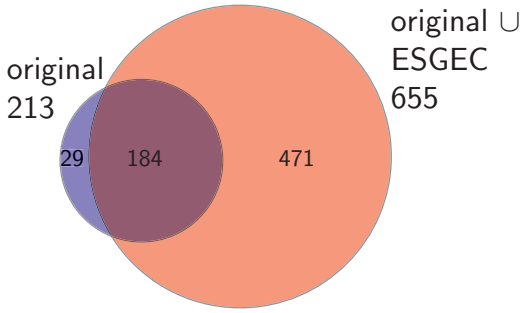


Figure 5.7: Overlap of original and expanded marker set study. The 213 original marker genes were published in Rasche et al. [240] and in the study under discussion original transcriptomic sources were expanded by nine additional project specific microarray data sets.

studies have been functionally characterised in humans [84].

Since the contribution of most of the known risk alleles to the development of type-2 diabetes is rather small, one might conclude that many additional genetic factors are still unknown. Therefore, and since there is no unambiguous set of candidates that defines truly positive disease genes in a polygenic context, this analysis may provide guidance for future systematic investigation of candidate genes and further validation studies.

Table 5.5 reflects a limited overlap of the type-2 diabetes mellitus genes predicted by this study with those predicted by other bioinformatics methods. This difference can be explained by the differences in the data domain used for the predictions (for example, sequence data, gene expression data, PPIs) and differences in the methods themselves. The lack of overlap is not unique to this study and seems to be a common problem with any two prediction studies. In particular, one study – Tiffin et al. [298] – compared seven different analysis methods and found that there was no gene common to all studies.

Accounting for experimental study bias

Since data from multiple tissues is analysed in human and mouse, it is likely that for some cases an individual experiment is dominating the score, for example, if the gene is active only in a single tissue. In order to identify those genes an entropy-based numerical criterion is computed. Entropy is used as an indicator for measuring generality and specificity of a candidate gene with respect to the different studies.

Let be $s_{g,u}$ the score of the g-th gene in the u-th study, then H_g is a measure for the uniformity of the score distribution over the individual experiments:

$$H_g = -\sum_u \frac{s_{g,u}}{\sum_k s_{g,k}} \log_2 \left(\frac{s_{g,u}}{\sum_k s_{g,k}} \right) \tag{5.4}$$

Entropy is low if a single study has a major contribution on the overall score. On the other hand, entropy is high if most of the studies account equally for the score.

Source Name	markersymbol	StumvollGoldstein2005	DeanMcEntyre2004	OMIM	PubMedGeneRIF	KOmiceJax	NandiAccili2004	ZegginiDIAGRAM2008	Frayling2007	Score	Entropy
ENSMUSG00000026827	Gpd2			*						8.92	3.83
ENSMUSG00000004056	Akt2			*	*	*	*			7.33	3.47
ENSMUSG00000012705	Retn			*	*					6.62	2.97
ENSMUSG00000023951	Vegfa			*				*		4.94	3.44
ENSMUSG00000038894	Irs2	*		*	*	*				3.91	2.11
ENSMUSG00000041798	Gck	*	*	*	*	*				3.86	2.22
ENSMUSG00000037370	Enpp1			*	*					3.59	2.91
ENSMUSG00000024985	Tcf7l2			*	*				*	3.5	2.86
ENSMUSG00000020679	Tcf2		*	*	*				*	3.18	1.61
ENSMUSG00000017950	Hnf4a	*	*	*	*					3.11	2.35
ENSMUSG00000055980	Irs1	*	*	*	*	*	*			3.06	1.71
ENSMUSG00000029556	Tcf1		*	*	*	*				3	1.59
ENSMUSG00000073134				*	*	*	*			3	1.59
ENSMUSG00000070561	Kcnj11	*	*	*	*				*	2.87	2.06
ENSMUSG00000034701	Neurod1		*	*	*					2.39	1.48
ENSMUSG00000029644	Pdx1			*		*				2	1
ENSMUSG00000027223	Mapk8ip1			*						1	0

Table 5.4: Results for type-2 diabetes mellitus OMIM genes. The OMIM database associates 17 genes with type-2 diabetes mellitus. Some are discussed in different reviews. Not all genes are detected in the meta-analysis, suggesting that expression differences are low in disease samples.

	this study	RascheHerwig2008	TiffinHide2006	DiabetesGenomeCG	ParikhGroop2004	OMIM	DoriaKahn2008	KitanoMuramatsu2004	LiuKasif2007_IS	JiangHancock2007
this study	**655**	0.27	0.02	$4.5{\cdot}10^{-3}$	0.01	0.01	$3.0{\cdot}10^{-3}$	0.03	0.03	0.01
RascheHerwig2008	184	**213**	0.03	0.01	0.02	0.03	0	0.04	0.02	0.01
TiffinHide2006	18	10	**102**	0	0.01	0.01	0	0.02	0.02	0.01
DiabetesGenomeCG	3	2	0	**8**	0.04	0	0	0.01	0	0.01
ParikhGroop2004	9	5	1	1	**18**	0.17	0.06	0.04	0.02	0
OMIM	4	6	1	0	5	**17**	0.06	0.03	0.04	0.02
DoriaKahn2008	2	0	0	0	2	2	**19**	0.01	0	0
KitanoMuramatsu2004	25	13	4	1	6	5	2	**148**	0.09	0.07
LiuKasif2007_IS	18	7	3	0	2	4	0	20	**92**	0
JiangHancock2007	6	2	2	1	0	2	0	14	0	**68**
Cit.		[240]	[298]	[1]	[230]	[225]	[84]	[168]	[191]	[148]

Table 5.5: Commonality of different studies and models. In the lower left triangular matrix are total numbers of overlap. In the triangular matrix part is the fraction of overlap to the union of candidates, the Jaccard index. Identifier from the original studies are mapped to Ensembl 44 and small differences in the number of candidates may apply.

For example, the gene *Serpina1b* has an outstanding score (7.713, rank 27/18 439) in this study. This is due to a very high fold-change in a single experiment; consequently, entropy is low (1.14, rank 11 014/18 439). In contrast, other genes show more consistent alteration across many different studies, for example *Pdk4* (9.39, rank 8/18 439) indicated by higher entropy (3.8, 214/18 439). Differential expression of *Pdk4*, a major regulator of glucose metabolism, has been found in fat, pancreatic islets and skeletal muscle but not in liver. The thirty genes with highest scores are listed in Table 5.6.

A plot of the entropy vs. the score is given in Figure 5.8. The Pearson correlation between the score and the entropy is 0.79. Most of the type-2 diabetes mellitus marker genes have high entropy and, thus, contribute to expression changes in many of the experiments.

5.2.4 Beyond the marker set

Genes do not act as individual units, they collaborate in overlapping pathways, the deregulation of which is a hallmark for the disease under study. In order to integrate pathway information and to derive cellular network information on the selected genes, functional annotation is added from pathway databases such as KEGG, Reactome, BioCyc [151, 158, 249], GO [26], protein-protein interaction databases such as IntAct [124] and databases on transcription factors such as TransFac [201]. Genetic variation of a

Rank	MGI symbol	Score	Entropy	Rank	MGI symbol	Score	Entropy
1	Cyp2e1	11.06	2.96	16	Cstb	8.37	3.42
2	Elovl6	10.79	3.36	17	Dhrs7	8.28	3.79
3	Tst	10.18	3.42	18	Ctss	8.24	3.64
4	Thrsp	10.04	3.67	19	Ccl2	8.23	2.87
5	Fasn	9.93	3.55	20	Apobec1	8.22	3.03
6	Acly	9.88	3.76	21	Pik3r1	8.17	3.93
7	Atf3	9.45	3.37	22	Mod1	8.12	3.78
8	Pdk4	9.39	3.79	23	Cfd	8.09	3.35
9	Lgals3	9.11	3.39	24	Hsd11b1	7.92	3.27
10	XR_003396	9.1	3.13	25	Srebf1	7.87	3.66
11	Gpd2	8.92	3.83	26	Serpine2	7.8	3.32
12	Scd1	8.75	3.38	27	Serpina1a	7.71	1.14
13	Ccrn4l	8.56	3.46	28	Ddah1	7.71	2.95
14	Agt	8.51	3	29	Cd14	7.56	3.65
15	Lgmn	8.43	3.51	30	Tyrobp	7.36	3.24

Table 5.6: Top 30 type-2 diabetes mellitus candidate genes. Top thirty type-2 diabetes mellitus candidate genes out of the 655 markers

gene was described with the number of associated SNP. The number of SNP in the coding and surrounding region of the gene is noted for mouse and human [38]. A particular biomedical interest is on genes that can be used for drug development. This characteristic has been previously assigned to the gene's ability to provide binding sites for biochemical well-characterised (i.e. druggable) compounds [131, 250]. The selected candidates were evaluated with respect to this information. In this Subsection the computed type-2 diabetes mellitus gene set has been used to identify biological networks on different layers of cellular information such as signalling and metabolic pathways, a comprehensive gene regulatory network and protein-protein interactions.

Relation to monogenic mouse models for type-2 diabetes mellitus

A variety of genetic studies have been performed in the last decades. At least 22 genetically engineered mouse models with type-2 diabetes mellitus phenotype have been studied in detail [3, 214]. Of those, five genes show a significant score in the meta-analysis: *Akt2*, *Slc2a4*, *Ptpn1*, *Slc2a2* and *Ppp1r3c*. Consistent with previous reports, the insulin-regulated glucose transporter *Slc2a4* is down-regulated in the insulin resistant state in adipose tissue but not in skeletal muscle. Likewise, down-regulation of glucose transporter *Slc2a2* in pancreatic islets confirms previous reports and reflects deterioration of β-cell function in the course of insulin resistance and diabetes. On the other side *Slc2a2* is also changed in liver. *Ptpn1* is expressed in all tissues showing only small fold-changes. Several genes from OMIM or KO-mice do not change at all on the expression level. This indicates that only the complete loss of the associated protein alters the system whereas

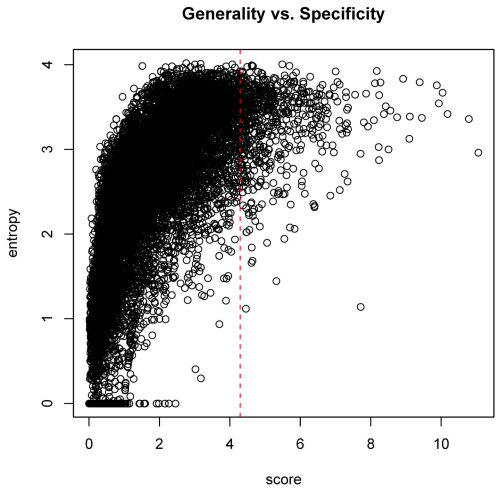


Figure 5.8: Dispersion of entropy. In the upper right corner are target genes related to type-2 diabetes mellitus by consistent and strong alteration in expression. In the lower right corner are genes with strong alteration specific e.g. to tissue or species. The red-dashed line indicates the cut off for the candidate genes.

the gene's expression is not altered in type-2 diabetes mellitus.

Relation to human and rodent association and linkage studies

Recently, a total of 20 candidate genes for type-2 diabetes mellitus have been identified and replicated in humans through multiple genome-wide association studies of common variants by using high-density SNP mapping approaches [84]: *Cdkal* (score 0), *Cdkn2a* (score 0)/*Cdkn2b* (4.0, 898/18 439), *Fto* (2.9, 2545/18 439), *Hhex* (4.3, 671/18 439), *Igf2bp2* (3.1, 2067/ 18 439), *Kcnj11* (2.9, 2520/18 439), *Pparg* (4.3, 651/18 439), *Slc30a8* (0.1, 12 895/18 439), *Notch2* (0.14, 12 526/18 439), *Thada* (0.20, 12 176/18 439), *Adamts9* (1.1, 8210/18 439), *Jazf1* (0.28, 11 656/18 439), *Cdc123* (3.5, 1500/18 439), *Camk1d* (2.0, 4622/18 439), *Ide* (5.0, 312/18 439), *Kif11* (0.75, 9554/18 439), *Tspan8* (2.1, 4352/18 439), *Lgr5* (0.53, 10 497/18 439) and *Tcf7l2* (3.5, 1435/18 439). Interestingly, only two of these genes show a high score in the meta-analysis, *Pparg* and *Ide*, although also *Cdkn2b* and *Tcf7l2* are significant on the less restrictive 0.01 level. On the other hand, from the data one could infer that *Fto* and *Hhex* act in pancreatic islets. *Cdkal1* and *Cdkn2a* are not expressed in the transcriptional studies. These genes show very low expression levels or might be active in tissues not included in the study. Since the meta-analysis approach takes into account several data sets from DNA microarrays, the candidate genes have a bias towards transcripts whose expression is changed in the context of type-2 diabetes mellitus. Moreover, the gene variants from association studies may not result in altered gene expression and, for most SNP found in association studies, there is a lack of functional information since the variation mostly occurs in non-coding regions of the genes. In

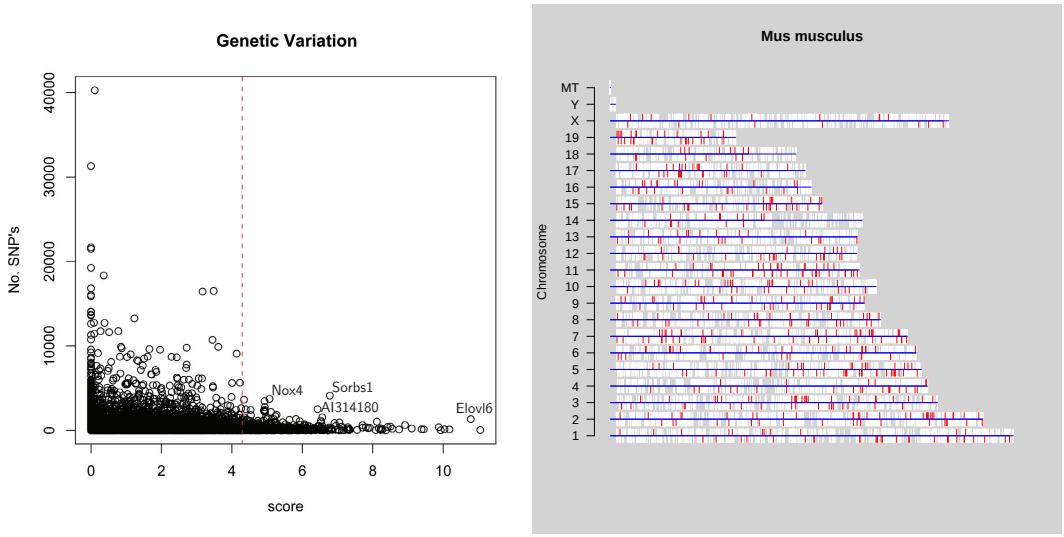


Figure 5.9: Genetic variation & Chromosomal localisation of type-2 diabetes mellitus genes. **Left:** On the x-axis is the score and on the y-axis is the number of identifier in the coding and surrounding region of the gene. A general tendency of disease genes to high genetic variation is not observable. This survey only relates the number of varied nucleotides. No conclusion is possible about single polymorphisms. The red-dashed line indicates the cut off for the candidate genes. **Right:** In a picture of the geneplotter package in R/BioC the set of significant genes has been marked with red [107, 106]. The significant genes are spread over the whole genome. It indicates that no genomic region can be accountable for the heritable prevalence.

order to correlate the type-2 diabetes mellitus genes with genetic variation the number of known SNP is plotted for the genes vs. the score (Figure 5.9). No general tendency to highly variable genes is observable. Two genes of the candidate list show high variation by high score, *Sorbs1* (4130) and *AI314180* (2523).

A further issue of the study was the chromosomal localisation of the type-2 diabetes mellitus genes. The distribution of the genes over the mouse genome is displayed in Figure 5.9. Using the hypergeometric distribution on local sliding windows across the chromosome significantly enriched chromosomal regions are identified. However, none of these regions convinced since they are sparsely occupied. Rather conversely, one observes that type-2 diabetes mellitus affects a wide range of physiological phenomena spanning loci in the entire genome.

Assessing functional annotation with over-representation analyses

Disease related networks were investigated with four different types of network information – biological pathways [151, 158, 249], protein-protein interaction networks [124], gene regulatory networks [219, 220, 201] and functional annotation using GO annotations [26] (see Table 5.7). These networks define – by annotation – sets of associated genes. The hypergeometric distribution compares the overlap between the superset and the gene group to the overlap of a random selection of two gene groups with the same size. Thus

Resource	Species	Resource content	Version	No. sets	Cit.
KEGG	mouse	pathway	11.03.08	201	[158]
ConsensusPathDB	human	pathway	19.04.07	1617	[155]
OdomYoung2004	human	study of selected TF in liver and pancreas	publication	6	[220]
OdomYoung2006	human	study of selected TF in liver	publication	6	[219]
TransFac	mouse	sequence motifs for TF	10.2	187	[201]
GO molecular function	mouse	ontology	Ensembl 44	1010	[26, 38]
GO cellular component	mouse	ontology	Ensembl 44	366	[26, 38]
GO biological process	mouse	ontology	Ensembl 44	2170	[26, 38]

Table 5.7: Overview about the network level. Overview about the network level, e.g. gene set, resources used in the meta-analysis approach. Abbrv.: TF, transcription factor.

it is possible to assign each annotation item (pathway, GO term etc.) a p-value that reflects enriched occurrence of candidate genes.

This analysis is facilitated by a product from the DE pipeline in Subsection 3.2.6 about gene set evaluation. In case of GO terms only genes are included with evidence codes IC, IMP, TAS or IDA to rely on the same confidence level as in the above mentioned resources. The p-values are corrected for multiple testing using q-values following Storey for the control of the false discovery rate [281, 282].

The over-representation analysis results provide confidence in the markers. Categories on the physiological level span two pathway resources (KEGG, ConsensusPathDB) [158, 155] and the GO tree [26]. Altogether, 5563 gene sets are analysed, whereof 799 (14.4 %) are significant with a p-value below 0.05. As greater parts of the metabolism are affected by type-2 diabetes mellitus, multiple pathways have a significant over-representation p-value. For example, in KEGG 48 out of 202 pathways have a p-value lower than 0.05 (23.8%). However results for different pathways are not independent. For example, the 130 genes annotated with *insulin signalling pathway* and the 45 genes annotated with *type II diabetes mellitus* share 31 genes.

Since several pathway resources are used in parallel, one can compare the findings for consistency, assuming the resources are independent. Some of the top pathways *PPAR signalling*, *adipocytokine signalling* and *insulin signalling pathway* are well related to type-2 diabetes mellitus. The study identifies KEGG pathways *Fatty acid metabolism* and *polyunsaturated fatty acid biosynthesis* in the enriched pathways what is complemented by the ConsensusPathDB pathways *fatty acid elongation – saturated*, *fatty acid β-oxidation I/II/IV* and by the GO categories *fatty acid metabolic process*, *fatty acid biosynthetic process*, *fatty acid beta-oxidation*, *fatty acid elongation*, *positive regulation of fatty acid biosynthesis* as well as *positive regulation of fatty acid metabolism*. The KEGG pathway *complement and coagulation cascades* is complemented by the ConsensusPathCB pathways *alternative complement activation*, *initial triggering of complement*, *complement*

SigSet	Set	Sig	All	p-value	q-value	Pathway description
21	75	655	18439	$9.22 \cdot 10^{-14}$	$1.81 \cdot 10^{-2}$	PPAR signalling pathway
17	71	655	18439	$3.22 \cdot 10^{-10}$	$1.81 \cdot 10^{-2}$	Adipocytokine signalling pathway
14	48	655	18439	$7.00 \cdot 10^{-10}$	$1.81 \cdot 10^{-2}$	Valine, leucine and isoleucine deg.
21	130	655	18439	$6.14 \cdot 10^{-9}$	0.02	Insulin signalling pathway
8	19	655	18439	$1.30 \cdot 10^{-7}$	0.02	Polyunsaturated fatty acid bios.
11	45	655	18439	$3.53 \cdot 10^{-7}$	0.02	Type II diabetes mellitus
11	47	655	18439	$5.67 \cdot 10^{-7}$	0.02	mTOR signalling pathway
13	69	655	18439	$7.88 \cdot 10^{-7}$	0.02	VEGF signalling pathway
22	186	655	18439	$7.99 \cdot 10^{-7}$	0.02	Focal adhesion
13	73	655	18439	$1.54 \cdot 10^{-6}$	0.02	Pancreatic cancer
15	110	655	18439	$7.95 \cdot 10^{-6}$	0.02	Leukocyte transendothelial mig.

Table 5.8: Gene set over-representation of the most significant KEGG pathways. All are the genes under consideration, **Sig** the number of candidate genes, **Set** is the number of genes in the pathway under study and **SigSet** the overlap of genes in the pathway and the candidate genes. p-values were computed as in Subsection 3.2.6. q-values are the multiple testing corrected p-values [281, 282].

cascade, *classical complement pathway*, *classical antibody-mediated complement activation*, *alternative complement pathway* and the GO categories *defense response*, *cellular defense response*, *complement activation, classical pathway* and *complement activation, alternative pathway.*

For 281 type-2 diabetes mellitus gene candidates there is information on the associated biochemical pathways (using the KEGG database). Whereas most genes (271) are associated with a single or a few (up to 11) pathways, some genes exhibit a higher interconnection such as *Mapk3* (32 pathways), *Mapk1* (32), *Pik3r1* (27), *Akt2* (25), *Akt1* (25), *Prkcb1* (19), *Ccnd1* (17), *Aldh9a1* (16), *Cdc42* (15) and *Mapk9* (13). The importance of *Mapk1*, *Pik3r1*, *Rasa1* and *Socs2* is also supported by Liu et al. [191] as members of an insulin signalling subnet derived from protein-protein-interactions.

Type-2 diabetes mellitus-related protein-protein interactions

Protein-protein interactions have been taken from the IntAct database denoting the number of interactions and interactors registered for the type-2 diabetes mellitus candidate genes. The ratio of interactors to interactions indicates whether the protein participates in big complexes or binds with single proteins. Figure 5.10 shows the number of interactions and the score for all genes under study. There is no trend for preferential selection of highly interacting genes in this type-2 diabetes mellitus candidate list. The high-scored genes comprehend a few genes with many interactions like *Mapk1*, *Pik3r1*, *Stat3*, *App*, and *Rela* in mouse with at least 20 listed interactions. The large number of interactions of *Mapk1* and *Pik3r1* is consistent with their participation in many of the signalling pathways. *Actb*, *Capza2*, *Myh11* and *Myh9* have more than 700 interactors, indicating big polymers. In human *Tsc22d1*, *Tnfrsf1b*, *Ndrg1*, *App* and *Nfkbia* have most

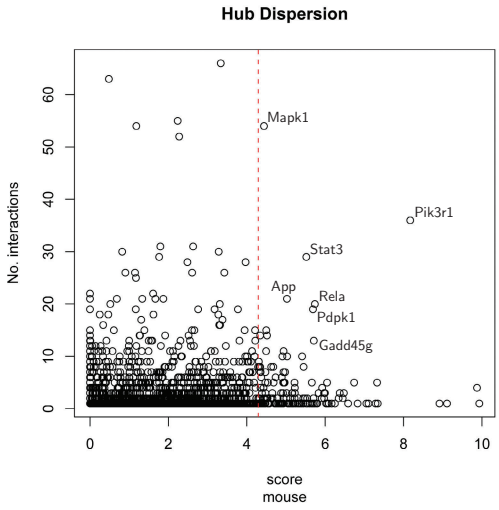


Figure 5.10: Number of protein-protein-interactions vs. score. Scatterplot of the number of mouse protein interactions in IntAct vs. the type-2 diabetes mellitus gene score. The vertical red-dashed line indicates the significance cut-off value of the score. *Mapk1* and *Pik3r1* are highlighted as genes with more than 30 interactions.

interactions. *Lmna* is the only gene with more than 300 interactors.

Mapping the IntAct interactions on Ensembl genes and coerce the human net and mouse net a graph is derived with 5975 nodes and 233 188 edges (data not shown). If one considers the edges between significant genes and their non-significant nearest neighbours there are still 2490 nodes and 32 727 edges. This shows that the disease genes strongly interact with main physiological triggers and deregulate essential parts of the metabolic network. Reducing the interactions on the 655 type-2 diabetes mellitus genes results in 173 nodes and 1293 edges visualised in Figure 5.11. Such a network constitutes the core of a topological model as proposed in Schlitt and Brazma [254].

Protein-protein interactions are still very sparse or derived from high-throughput experiments with low overlap and low reproducibility so that results have to be carefully cross-checked. For example, a protein complex arises from one experiment of Collins et al. [67] with vague relationship to type-2 diabetes mellitus in the network of the candidate genes.

Type-2 diabetes mellitus-related gene regulatory network

In order to study the information content of the set of selected disease genes on gene regulation, in the study are analysed a) the transcription factors present in the significant set and b) known target sets of transcription factors for over-representation. Analysis is often hampered because transcription factors are known to be expressed at a very low level and fold changes are commonly low. Moreover, many transcription factors are regulated by phosphorylation (e.g. *Foxa*'s) and/or ligand binding (e.g. *Ppar*'s). As a

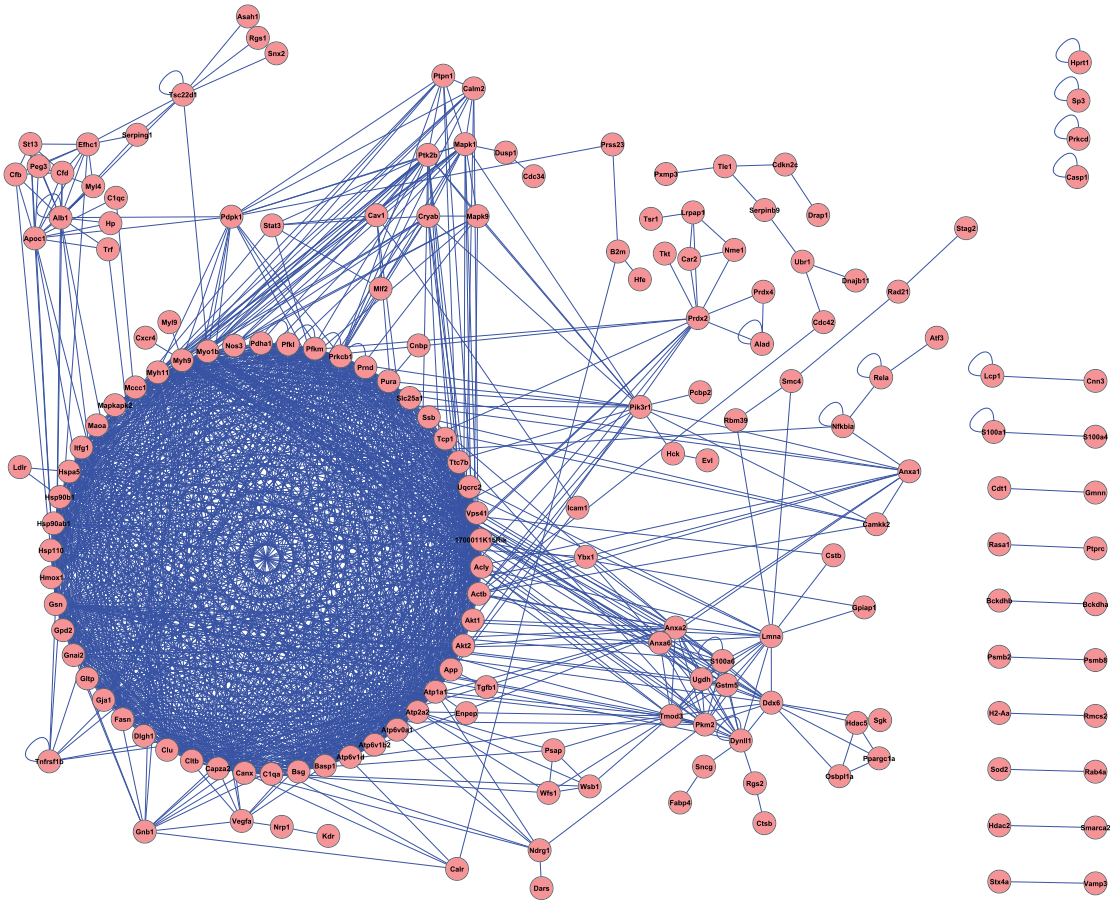


Figure 5.11: Protein-protein interactions of type-2 diabetes mellitus marker genes. Protein interaction network of the significant gene list. The arcs are the interactions in IntAct between significant genes. The interaction network in mouse and human have been united. In the upper part is *Mapk1* as an intermediator between *Pik3r1* and *Irs1*. In the lower part is a cluster stemming from a polymer binding in mouse synapses.

Ensembl	HGNC	SigSet	Set	Sig	All	p-value
ENSMUSG00000017950	HNF4A	224	4267	655	18439	$3.98 \cdot 10^{-11}$
ENSMUSG00000037025	FOXA2	58	857	655	18439	$1.90 \cdot 10^{-6}$
ENSMUSG00000029556	TCF1	56	964	655	18439	0.000204
ENSMUSG00000043013	ONECUT1	60	1247	655	18439	0.0101
ENSMUSG00000026641	USF1	67	1458	655	18439	0.0176
ENSMUSG00000025958	CREB1	88	2075	655	18439	0.0438

Table 5.9: Gene set over-representation of significant transcription factor target sets from Odom et al. [219] (2006). The target sets with a p-value below 0.05 are displayed. In the rows are the different transcription factors. Column identifier as in Table 5.8.

result, important core regulators including *Onecut1* (score 1.3, rank 7311/18 439), *Hnf4a* (3.1, 2033/18 439), *Tcf1* (3, 2238/18 439), and *Foxa2* (2.5, 3317/18 439) are not in the candidate list. Collecting transcription factors from Odom et al. [219, 220], TransFac [201] and the GO category GO:0003700 [26] in mouse and human with evidence codes IC, IMP, TAS or IDA identifies 502 transcription factors. Thereof 30 transcription factors received a high score in this type-2 diabetes mellitus set: *Pparg, Sfpi1, Stat3, Srebf1, Nr1d1, Fos, Epas1, Rela, Sp3, Ybx1, Bhlhb2, Cbfb, Cebpa, Irf8, Pura, Foxo1, Cebpb, Nr1d2, Cited2, Klf7, Sox18, Max, Klf10, Ccrn4l, Cnbp, Drap1, Klf9, Nfil3, Hdac2* and *Atf3*. *Srebf1* and *Ybx1* are expressed only in mouse but in every tissue. *Cebp*'s and *Srebp*'s are important regulators of lipid metabolism and adipogenesis and were found differentially expressed in the course of insulin resistance and type-2 diabetes mellitus. Consistent changes could be identified in the tissues under study (fat: all but *Nfil3*; liver: *Srebf1, Ccrn4l, Ybx1, Bhlhb2, Klf10*; muscle: *Atf3, Klf10, Nfil3*; pancreatic islets: *Ccrn4l, Atf3, Ybx1, Bhlhb2, Klf10, Nfil3*). *Pparg* is expressed predominantly in fat where its expression is altered.

In total, target sets of 199 transcription factors have been investigated as gene sets for over-representation analysis. Table 5.9 shows the transcription factors from Odom et al. [219]. For example, *Cebpa* is highly significant. It is expressed in adipose tissue and modulates the expression of *Leptin*. *Cebpa* shows some correlation with the level of hyperglycemia in [211]. Alteration is also observable in liver. A gene regulatory network comprising the regulatory interactions of the significant genes and the significant and enriched transcription factors is shown in Figure 5.12. Obvious are the five hubs, the core regulatory circuit derived from [219]. Target and regulator at the same time are *Foxa2, Tcf1, Hnf4a, Onecut1, Creb1, Sfpi1, Stat3, Srf, Srebf1* and *Cebpa*.

The gene regulatory network associated with the type-2 diabetes mellitus candidate set is generic in the sense that all interactions are displayed regardless whether the genes are expressed in a specific tissue or not. This network can be tuned towards tissue specificity by taking into account tissue-related gene expression and other characteristics. Using tissue expression data sets (Su et al. [286]) the representation is assessed of the different tissues in the type-2 diabetes mellitus candidate list. A total of 552 genes from the list are included in the tissue expression panel, where 353 (64%) are expressed in fat, 210

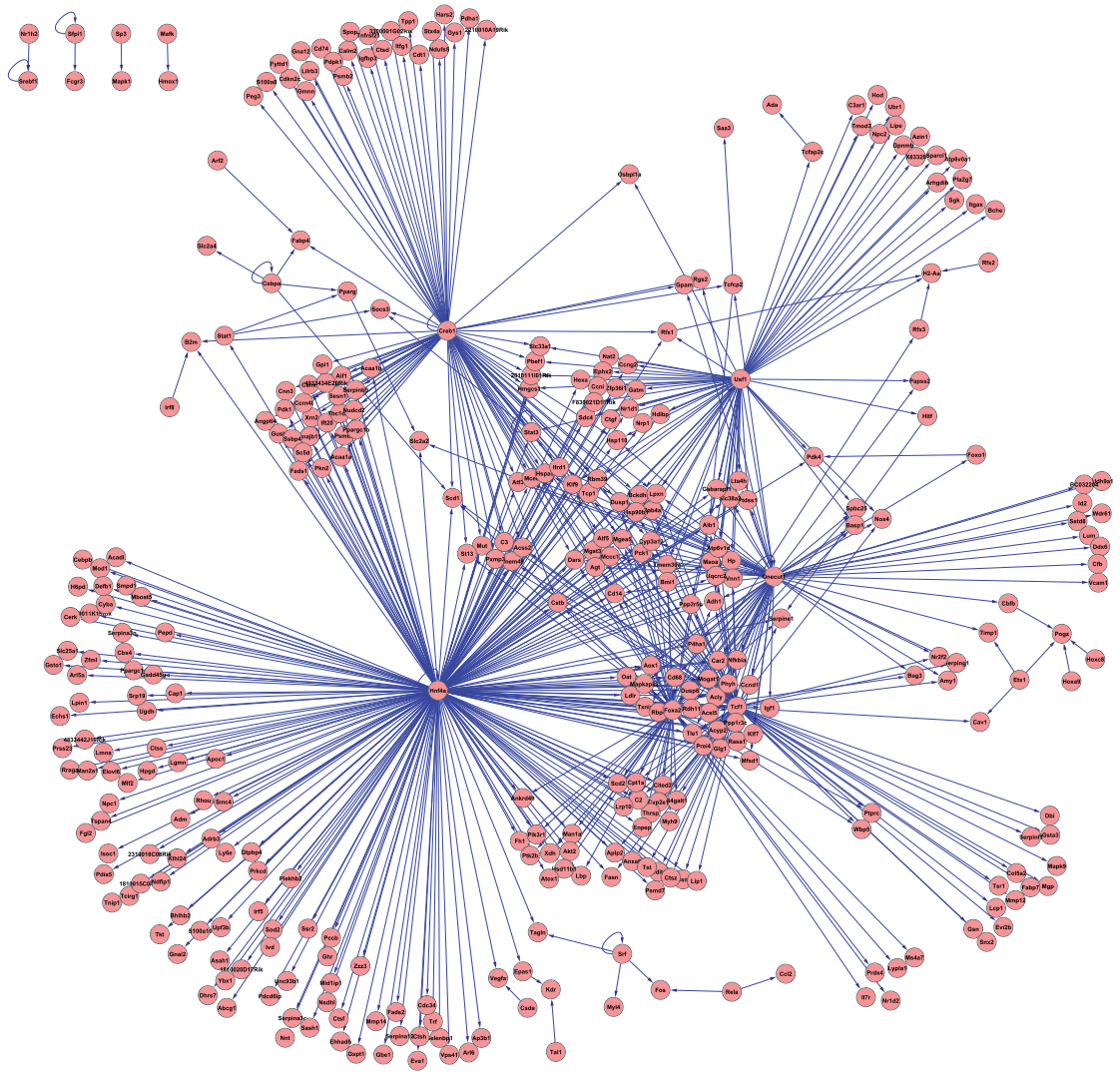


Figure 5.12: Gene regulations in the type-2 diabetes mellitus marker gene set. Gene regulatory network composed of the significant genes. Significant transcription factors and transcription factors with enriched target sets with respect to the type-2 diabetes mellitus candidate gene list. Thick ends of the arrows point to transcription factors, thin ends point to target genes.

strain/phenotype	liver	fat
NZL hyperglycaemic	2	4
NZL normoglycaemic	4	5
SJL normoglycaemic	5	5

Table 5.10: Study design of the GGSC data set. The table shows the number of replicates.

(38%) in muscle and 231 (42%) in liver. An intersection of 137 genes is expressed in all three tissues (data not shown).

There are further limitations in analysing gene regulatory networks. Information of transcription factor binding sites – besides computationally predicted sites – is sparse and the knowledge on target sets of transcription factors is limited. In Table 5.9 the p-values for six target sets of regulators are listed that have been derived from ChIP-on-Chip data. The ChIP-on-Chip data might also help in characterising the 499 unknown type-2 diabetes mellitus genes as being potential transcription factor targets. The overlap between this uncharacterised subset and the transcription factor target sets are: *Hnf4a* 224 genes, *Foxa2* 58 genes, *Usf1* 67 genes, *Tcf1* 56 genes, *Creb1* 88 genes and *Onecut1* 60 genes. However, this technique is still error-prone and generates a lot of false positive targets due to the different steps in the experiment. Commonly, one ends up with large targets sets containing thousands of genes [219, 220]. Here, new methods of computational analysis that combine ChIP-on-Chip predicted targets with sequence analysis of their promoter regions have to be developed.

5.3 Evaluation of alternative splicing with exon arrays

The meta-analysis provides a detailed picture of gene expression in type-2 diabetes mellitus. However in the marker list is for example *Ppargc1a*. In Monsalve et al. [206] the authors point out that *Ppargc1a* is not only a transcription factor but also a splicing factor. It is a coactivator of *Pparg* and the *Pparg* expression changes in fat and liver were never subject of genome-wide splicing assays. Transcriptional or genetical changes of factors may cause splicing changes. Therefore mouse experiments were performed with a diabetic and a genetic dimension to separate respective effects.

Alternative splicing in type-2 diabetes mellitus was aside in the last years in terms of high-throughput experimentation. Efforts focussed on genetics with association studies and gene expression both using microarrays. This is in contrast to the observation, that the relevance of splice variants is well known for several key players in type-2 diabetes mellitus. Ratios of splice variants are correlated to insulin levels in type-2 diabetes mellitus for genes like *Tnfrsf1b, Ptpn1* and *Insulin* [257, 94, 95, 203]. The insulin receptor (*Insr*) has two isoforms with different affinities to bind insulin. In human the low affinity variant is reported to be increased in type-2 diabetes mellitus [258].

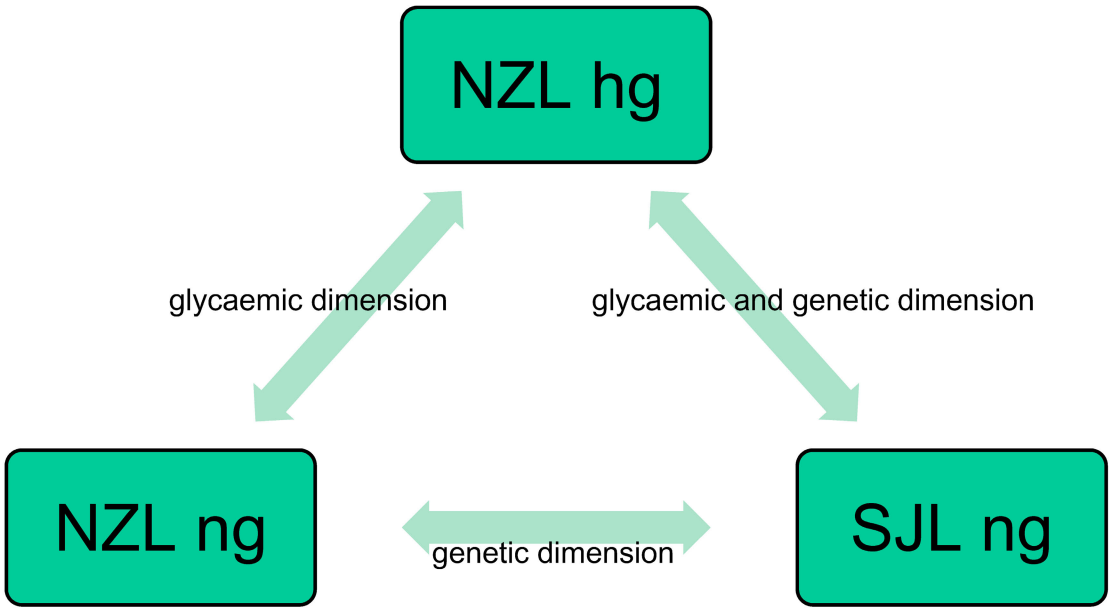


Figure 5.13: Glycaemic and genetic splicing changes. Obese *NZL* mice separate in a hyperglycaemic and a normoglycaemic group. Both groups differ from the always lean *SJL* strain. Abbrv.: hg, hyperglycaemic; ng, normoglycaemic.

5.3.1 Glycaemic and genetic splicing changes

The polygenic mouse model *NZO* has been successful to identify type-2 diabetes mellitus marker genes in the ESGEC data set. The closely related *NZL* mouse is applied to generate another data set, now for the identification of spliced genes. The second data set is called GGSC — glycaemic and genetic splicing changes.

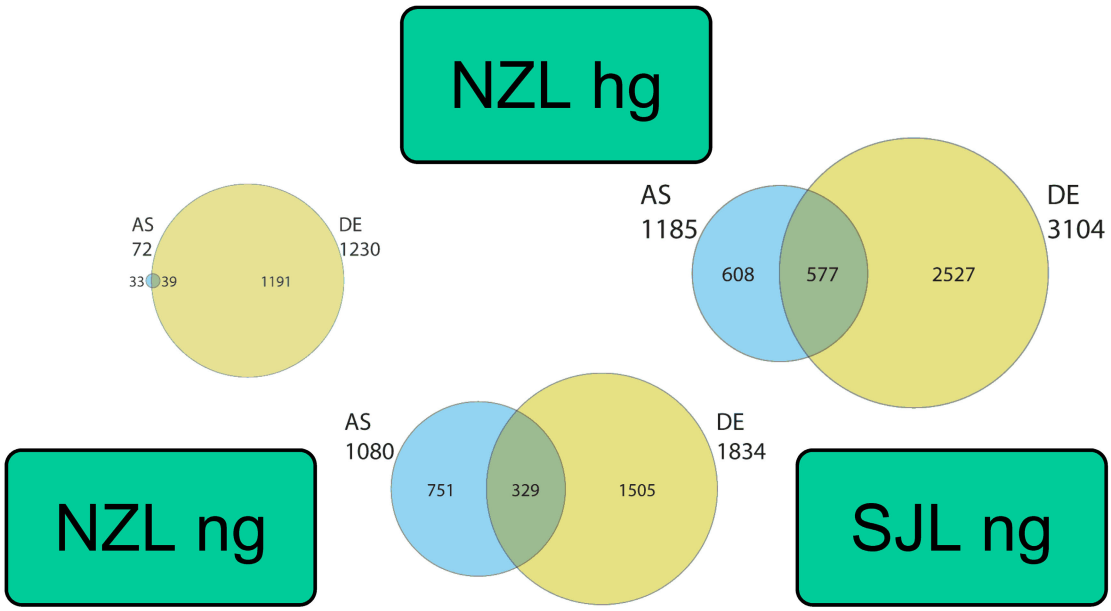
The *NZL* mice on a high-fat diet all develop severe obesity but not all develop diabetes. These two groups are separated by measurements on the blood glucose, i.e. in a hyperglycaemic and a normoglycaemic group. On the other hand the *SJL* mice on high-fat diet do not develop obesity nor diabetes. This implicates distinct genetic differences between *NZL* and *SJL* strains. Figure 5.13 visualises the experimental setting including test cases for the glycaemic, the genetic as well as the combined glycaemic and genetic dimension.

The two experimental dimensions are analysed in the tissues with considerable changes in the marker gene *Pparg*. *Pparg* regulates fatty acid storage and glucose metabolism. Many insulin sensitising drugs used in the treatment of diabetes target *Pparg* as a means to lower serum glucose without increasing pancreatic insulin secretion. The genes activated by *Pparg* stimulate lipid uptake and adipogenesis by fat cells. *Pparg* knockout mice fail to generate adipose tissue when fed a high-fat diet. Excess fat in obesity is known to be associated with an inflammatory state of adipose tissue. It is not definite how the inflammatory state causes type-2 diabetes mellitus or how some mice pertain normal glycaemia levels. On the other hand for regulating glucose levels the key organ is the liver. For the two tissues the numbers of arrays are listed in Table 5.10.

	Fat		Liver	
Rank	MGI	ARH p	MGI	ARH p
1	Tph2	$1.9{\cdot}10^{-6}$	1810049H19Rik	$8.2{\cdot}10^{-8}$
2	Itgad	0.0017	Ela2a	$2.0{\cdot}10^{-6}$
3	Matr3	0.0014	Ela1	$6.5{\cdot}10^{-5}$
4	Lamc1	0.017	Acot3	$8.7{\cdot}10^{-5}$
5	Cuzd1	0.0011	Pla2g1b	0.00017
6	Ctnna2	0.00063	Wdr47	0.00022
7	Reln	0.022	Upp2	0.00057
8	Fbln1	0.0085	Moxd1	0.00096
9	Fgf13	0.0055	Ugt2b37	0.0012
10	Kcnma1	0.047	Pnpla3	0.0013
11	Slc38a5	0.014	Prei4	0.0016
12	Actg2	0.0075	Mgst3	0.0017
13	Cilp	0.027	C730027P07Rik	0.0017
14	Ntrk3	0.011	Tsc22d3	0.0017
15	Pik3ap1	0.042	AC121985.3	0.0018
16	Zbtb16	0.0082	Aldh3a2	0.002
17	AC133081.4	0.0039	Egfr	0.0022
18	Zranb3	0.026	Txnrd1	0.0022
19	Rnf128	0.012	Il15ra	0.0025
20	Ankle1	0.0063	Tagln	0.0026
21	Cxcl10	0.0064	Cltb	0.0029
22	Nkain4	0.0099	Rcan1	0.0032
23	BC055004	0.01	Cyp7a1	0.0033
24	Ccl2	0.0027	Palld	0.0033
25	Capg	0.039	Dnajb1	0.0033
26	Clec12a	0.0077	B3galt1	0.0033
27	Mrpl37	0.018	Zbtb16	0.0037
28	Stab2	0.049	Gpr110	0.0039
29	Sgk1	0.0077	Mug2	0.004
30	Pctk3	0.048	Slc25a22	0.004
31	Mogat1	0.0093	1100001G20Rik	0.0042
32	Cldn10a	0.016	Morf4l2	0.0043
33	Ly9	0.037	Sds	0.0044
34	Mtap2	0.039	Acaa1a	0.0057
35	Fhl5	0.035	2310076L09Rik	0.0059
36	Rcan1	0.016	Ipo11	0.006
37	Il1b	0.022	Agpat5	0.0061
38	Nrg1	0.038	BC023882	0.0069
39	Sh3pxd2a	0.048	Snhg8	0.0076
40	Aif1l	0.034	Ide	0.0077

Table 5.11: Top 40 spliced genes in the glycaemic dimension. The spliced genes in the setting *NZL* hyperglycaemic vs. *NZL* normoglycaemic for both tissues. Abbrv.: MGI, mouse genome informatics identifier; ARH p, ARH p-value.

fat tissue:



liver tissue:

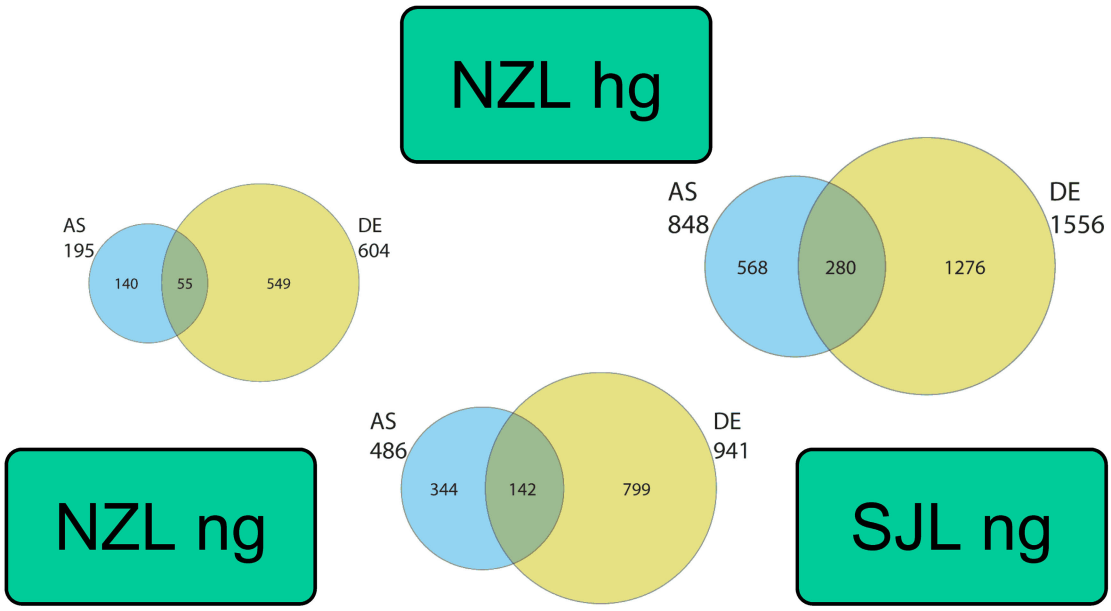


Figure 5.14: Splicing overview. Following the scheme of Figure 5.13 the size of the alternative splicing set of genes is visualised in blue. Alternatively spliced genes are related to the differentially expressed genes in yellow by Venn diagrams. The three Venn diagrams stand for the three dimensions in the data set. **Top:** Fat tissue. **Below:** Liver tissue. Abbrv.: hg, hyperglycaemic; ng, normoglycaemic.

After weaning at week 3, male *NZL* and *SJL* mice were fed a high-fat diet and characterised. high-fat diet contains 15.3 MJ/kg by 32.5% sugar, 17.1% protein and 14.6% fat among others. Characterisation includes body weight, body fat and blood glucose. The tissue samples were prepared at week eight. Dr. Tanja Dreja characterised and prepared the samples in the group of Dr. Hadi Al-Hasani at the German Institute of Human Nutrition (DIfE) in Potsdam-Rehbrücke [85]. After preparation the samples were passed to a company for hybridisation on Affymetrix Mouse Exon 1.0 ST arrays.

5.3.2 Splicing states in type-2 diabetes mellitus

On the GGSC exon array data set the tools developed and arranged in the previous chapters are applied with the AS pipeline. The ARH splicing prediction identifies sets of alternatively spliced genes, see Table 5.11 for a selection of top candidates. The AS pipeline with its parallel branches also identifies sets of differentially expressed genes for direct comparison of alternative splicing and differential expression on the same biological samples. The respective sets undergo functional evaluation with the gene set evaluation, in particular over-representation analysis.

Fat

For the GGSC samples splicing differences are visualised in Figure 5.14. The test cases for adipose tissue show little glycaemic splicing but more genetic splicing than in liver. The latter corresponds to the fact that the *SJL* mouse strain does not increase adipose tissue. In the *NZL* hyperglycaemic vs. *SJL* normoglycaemic test case the glycaemic and genetic splicing effects sum up. From the 1185 alternatively spliced genes it shares 769 genes with *NZL* normoglycaemic vs. *SJL* normoglycaemic and 42 genes with the 72 genes from the *NZL* hyperglycaemic vs. *NZL* normoglycaemic test case.

In the genetic dimensions the *Pik3r1* gene is spliced. It is one of the type-2 diabetes mellitus marker genes and is particularly interesting as it is a hub in the protein-protein-network as pointed out in Subsection 5.2.4. It participates in several signalling pathways. *Pik3r1* encodes three transcripts and the respective expression levels are studied in Lefai et al. [186] for muscle and fat tissue related to type-2 diabetes mellitus.

With the over-representation test the alternative splicing gene sets are linked to the functional level. For example looking at KEGG pathways one pathway is notably significant, the *ECM-receptor interaction*. the pathway has in the *NZL* hyperglycaemic vs. *NZL* normoglycaemic test case a p-value of 0.031 (see Table 5.13). Seven out of 204 pathways have a p-value below 0.05. The *ECM-receptor interaction* pathway is significant in all of the three dimensions together with *Leukocyte transendothelial migration*. The pathways are also significant in the differential expression gene set evaluation. The differential expression over-representation results closely corresponds to the ESGEC data set comparing high-fat diet vs. standard diet at week 8. However for differential expression 44 out of 204 pathways are significant with an emphasis to inflammation pathways. Inflammation or immune reactions are marginally affected by splicing.

Ensembl gene	MGI	ARH	ARH p	Chr.	Start	Tissue
ENSMUSG00000035385	Ccl2	0.11	0.0027	11	81849079	fat
ENSMUSG00000019970	Sgk1	0.073	0.0077	10	21648148	fat
ENSMUSG00000012187	Mogat1	0.068	0.0093	1	78507744	fat
ENSMUSG00000021775	Nr1d2	0.049	0.021	14	19071325	fat
ENSMUSG00000036086	Zranb3	0.045	0.026	1	129998352	fat
ENSMUSG00000024397	Aif1	0.040	0.036	17	35309230	fat
ENSMUSG00000056737	Capg	0.038	0.039	6	72499267	fat
ENSMUSG00000001131	Timp1	0.036	0.045	X	20450494	fat
ENSMUSG00000028459	Cd72	0.036	0.046	4	43460607	fat
ENSMUSG00000027346	Prei4	0.13	0.0016	2	132413357	liver
ENSMUSG00000026688	Mgst3	0.13	0.0017	1	169323882	liver
ENSMUSG00000032085	Tagln	0.11	0.0026	9	45737711	liver
ENSMUSG00000047547	Cltb	0.11	0.0029	13	54698387	liver
ENSMUSG00000051748	1100001G20Rik	0.092	0.0042	11	83565996	liver
ENSMUSG00000036138	Acaa1a	0.082	0.0057	9	119256848	liver
ENSMUSG00000056999	Ide	0.073	0.0077	19	37388470	liver
ENSMUSG00000037071	Scd1	0.065	0.010	19	44481876	liver
ENSMUSG00000023087	Ccrn4l	0.058	0.014	3	51028855	liver
ENSMUSG00000029322	Plac8	0.057	0.015	5	100985499	liver
ENSMUSG00000026356	Dars	0.045	0.026	1	130311931	liver
ENSMUSG00000026739	Bmi1	0.040	0.035	2	18604996	liver
ENSMUSG00000068874	Selenbp1	0.040	0.035	3	94748328	liver
ENSMUSG00000038776	Ephx1	0.040	0.036	1	182947558	liver
ENSMUSG00000066441	Rdh11	0.038	0.040	12	80289976	liver
ENSMUSG00000032786	Alas1	0.035	0.046	9	106149505	liver

Table 5.12: Spliced marker genes in the glycaemic dimension. The spliced marker genes in the setting *NZL* hyperglycaemic vs. *NZL* normoglycaemic. In fat are 9 spliced markers and in liver are 16. Abbrv.: MGI, mouse genome informatics identifier; ARH p, ARH p-value, Chr., chromosome; Start, chromosomal start position.

Consistent with these findings the gene *Kcnma1* is spliced in the glycaemic dimension with an ARH p-value of 0.047. It encodes a voltage gated ion channel and is known to be spliced in type-2 diabetes mellitus [82]. Another example is *Kcnd3*, it is integral to the cell membrane and is spliced in the *NZL* hyperglycaemic vs. *SJL* normoglycaemic test case with an ARH p-value of 0.041. *Kcnd3* is member of a gene family encoding potassium voltage-gated channels. Voltage-gated potassium channels represent the most complex class of voltage-gated ion channels from both functional and structural standpoints. Their diverse functions include regulating cell volume among others. More importantly it is known for two transcript variants differing by a protein kinase C site [234, 171, 314]. Such sites are subject to phosphorylation, i.e. a posttranslational modification. Thus *Kcnd3* is an example where alternative splicing directs such posttranslational modifications [272].

Liver

As already noted the marker gene set comprises the splicing factor coactivator peroxisome proliferator-activated receptor γ coactivator 1 α *(Ppargc1a)*. It is spliced in *NZL* hyperglycaemic vs. *SJL* normoglycaemic. As type-2 diabetes mellitus is a polygenic disease it cannot be expected that a splicing factor like *Ppargc1a* acts solely but interacts with different markers. In Subsection 5.2 a marker set was constituted. For example in the *NZL* hyperglycaemic vs. *NZL* normoglycaemic condition 16 out of 195 alternatively spliced genes are type-2 diabetes mellitus marker (see Table 5.12). In fact the proportion of alternatively spliced genes is generally higher in the marker set: $\frac{195}{21994} = 0.0089 < \frac{16}{655} = 0.024$.

In parallel another marker gene, the insulin degrading enzyme (*Ide*), is spliced between *NZL* normoglycaemic and the two other conditions with ARH p-value 0.0077 (both) visualised in Figure 5.15. Table 5.14 provides a detailed picture of exon expression and splicing indication for *NZL* hyperglycaemic vs. *NZL* normoglycaemic. For *Ide* Farris et al. [90] identify six distinct transcripts in human with most of the variance attributable to alternative polyadenylation sites. At least one of the variants is catalytically inefficient. Deletion of insulin-degrading enzyme (*Ide*) in mice causes hyperinsulinaemia and glucose intolerance. *Ide* is also a candidate from association studies in humans [84].

Genome-wide association studies have been performed in humans to identify genetic candidates for type-2 diabetes mellitus as described in Subsection 5.1.4 and reviewed in Doria et al. [84]. Two such candidate genes are spliced in liver, *Ide* and *Notch2*. *Notch2* is also spliced along the genetic dimension. Actually for most of these genes functional validation was not possible. The genetic changes identified in the human genome are mostly in non-coding parts of the DNA. Nonetheless it is particularly interesting to find these genes again in context of genetic changes. Many of the splicing signals and splice sites depend on nucleotide sequences in the introns as discussed in Subsection 2.1.5 [307].

Vegfa encodes a protein that is often found as a disulfide linked homodimer. This protein is a glycosylated mitogen that has various effects like promoting cell migration and inhibiting apoptosis among others. Alternatively spliced transcript variants, encoding either freely secreted or cell-associated isoforms, have been characterised. It is member of the OMIM candidate list for type-2 diabetes mellitus as discussed in Subsection 5.2.3.

Tissue	SigSet	Set	Sig	All	p-value	Pathway
fat	4	232	72	21 994	0.00709	Cytokine-cytokine receptor interaction
fat	2	82	72	21 994	0.0297	Hematopoietic cell lineage
fat	2	84	72	21 994	0.031	ECM-receptor interaction
fat	2	89	72	21 994	0.0345	ErbB signaling pathway
fat	2	94	72	21 994	0.0381	Toll-like receptor signaling pathway
fat	1	12	72	21 994	0.0386	Prion disease
fat	2	105	72	21 994	0.0465	Leukocyte transendothelial migration
liver	7	62	195	21 994	$1.26{\cdot}10^{-6}$	Metabolism of xenobiotics by cytochrome
liver	6	60	195	21 994	$1.51{\cdot}10^{-5}$	Retinol metabolism
liver	5	41	195	21 994	$3.01{\cdot}10^{-5}$	Linoleic acid metabolism
liver	6	69	195	21 994	$3.39{\cdot}10^{-5}$	Drug metabolism - cytochrome P450
liver	3	29	195	21 994	0.00212	Bile acid biosynthesis
liver	3	29	195	21 994	0.00212	Biosynthesis of unsaturated fatty acids
liver	4	72	195	21 994	0.00385	Arachidonic acid metabolism
liver	2	11	195	21 994	0.00408	Cysteine metabolism
liver	3	39	195	21 994	0.00496	Glycerolipid metabolism
liver	2	20	195	21 994	0.0134	Limonene and pinene degradation
liver	2	28	195	21 994	0.0254	Porphyrin and chlorophyll metabolism
liver	3	72	195	21 994	0.0262	PPAR signaling pathway
liver	2	32	195	21 994	0.0326	Alanine and aspartate metabolism
liver	2	33	195	21 994	0.0345	Lysine degradation
liver	1	4	195	21 994	0.035	Biotin metabolism
liver	3	84	195	21 994	0.0388	Pyrimidine metabolism
liver	3	90	195	21 994	0.046	Prostate cancer
liver	2	39	195	21 994	0.0468	Ribosome
liver	2	40	195	21 994	0.049	Valine, leucine and isoleucine degradation
liver	2	40	195	21 994	0.049	Drug metabolism - other enzymes

Table 5.13: Gene set over-representation of the most significant KEGG pathways. Listed are the pathways with a p-value below 0.05. **'All'** are the genes under consideration, **'Sig'** the number of candidate genes, **'Set'** is the number of genes in the pathway under study and **'SigSet'** the overlap of genes in the pathway and the candidate genes. p-values were computed as in Subsection 3.2.6.

Ensembl exon	E median hg	E MAD hg	E median ng	E MAD ng	G median hg	G MAD hg	G median ng	G MAD ng	ARH SD	Chr.	Start	End
ENSMUSE00000145331	2200	2728	660	683	1100	838	1100	893	1.7	19	37388470	37388562
ENSMUSE00000544899	1600	1289	2200	1122	1100	838	1100	893	-0.51	19	37370373	37370455
ENSMUSE00000544872	1200	816	900	311	1100	838	1100	893	0.4	19	37345995	37346062
ENSMUSE00000145301	1300	788	1000	741	1100	838	1100	893	0.38	19	37389895	37390007
ENSMUSE00000544894	1500	1207	1700	906	1100	838	1100	893	-0.28	19	37362588	37362698
ENSMUSE00000544882	2000	1149	1600	945	1100	838	1100	893	0.27	19	37355075	37355242
ENSMUSE00000544886	1200	154	1400	289	1100	838	1100	893	-0.26	19	37358611	37358722
ENSMUSE00000145284	660	467	540	420	1100	838	1100	893	0.23	19	37399638	37399807
ENSMUSE00000544879	1400	894	1200	1020	1100	838	1100	893	0.21	19	37352231	37352503
ENSMUSE00000145261	1700	401	1900	546	1100	838	1100	893	-0.21	19	37389324	37389486
ENSMUSE00000544893	1200	465	1400	270	1100	838	1100	893	-0.19	19	37361603	37361723
ENSMUSE00000326867	820	164	900	320	1100	838	1100	893	-0.18	19	37403471	37403678
ENSMUSE00000544876	1300	354	1400	393	1100	838	1100	893	-0.17	19	37351005	37351066
ENSMUSE00000544871	620	612	550	540	1100	838	1100	893	0.13	19	37343231	37345327
ENSMUSE00000544890	1400	302	1400	578	1100	838	1100	893	-0.12	19	37359696	37359787
ENSMUSE00000145293	2100	1530	1900	1355	1100	838	1100	893	0.12	19	37386573	37386653
ENSMUSE00000145313	570	405	520	423	1100	838	1100	893	0.1	19	37392467	37392589
ENSMUSE00000441012	580	233	600	484	1100	838	1100	893	-0.095	19	37404918	37405110
ENSMUSE00000544900	980	807	1000	756	1100	838	1100	893	-0.095	19	37372542	37372664
ENSMUSE00000544896	1300	611	1400	583	1100	838	1100	893	-0.088	19	37365160	37365304
ENSMUSE00000145257	1600	260	1500	708	1100	838	1100	893	-0.019	19	37384278	37384381
ENSMUSE00000145264	2400	1000	2300	824	1100	838	1100	893	0.018	19	37388000	37388091
ENSMUSE00000502797	1400	321	1300	304	1100	838	1100	893	0.017	19	37378384	37378486
ENSMUSE00000544874	500	510	480	439	1100	838	1100	893	0.01	19	37346645	37346717
ENSMUSE00000544869	120	22	110	67	1100	838	1100	893	0	19	37376153	37378168

Table 5.14: Exon expression of *Ide*. The gene Ide is spliced in liver in the setting *NZL* hyperglycaemic vs. *NZL* normoglycaemic with an ARH p-value of 0.0077. ARH splicing deviation indicates splicing strength of the exons. Abbrv.: E, exon; G, gene; hg, hyperglycaemic; ng, normoglycaemic; MAD, median absolute deviation; ARH SD, ARH splicing deviation; Chr., chromosome; Start, chromosomal start position; End, chromosomal end position.

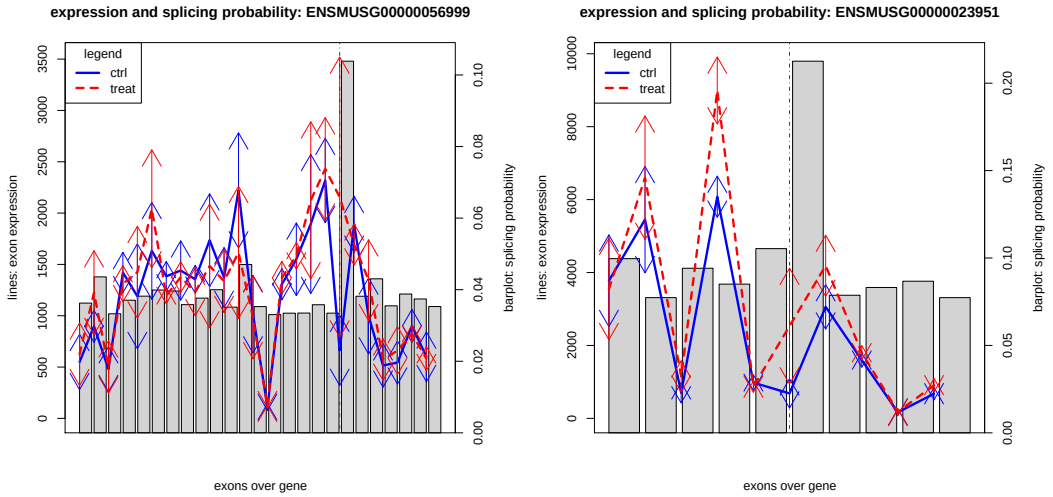


Figure 5.15: Spliced events along the experiment dimensions. The lines (y-axis, left scale) show the exon expressions ordered by genomic position (x-axis). The arrows visualise the median absolute deviation around the median expression values. The bars (y-axis, right scale) correspond to the splicing probability values of the respective exons. **Left:** *Ide* with a splicing event between *NZL* hyperglycaemic and *NZL* normoglycaemic (exon 19, green dot-dashed line). **Right:** *Vegfa* with a splicing event between *NZL* normoglycaemic and *SJL* normoglycaemic (exon 6, green dot-dashed line).

Consistent with the human genetic indication it is spliced in the GGSC mouse data set along the genetic dimensions with a p-value of 0.00099 and 0.0015 (see Figure 5.15).

In general at least one fourth of the spliced genes are also differentially expressed. This indicates a close link between alternative splicing and differential expression. Effects detected as gene expression changes are probably often the up- or downregulation of transcript variants. In particular *Ppargc1a* as transcription and splicing factor provides a possible explanation linking the two regulation levels.

6 Conclusion and Future Work

After developing and establishing several tools it was possible to perform genome-wide analysis of alternative splicing in type-2 diabetes mellitus, especially in marker genes. A combination of modules was necessary to achieve these results. As it was an interdisciplinary work, the presented tools can be divided to pursue three different aims. The methodological aim was to improve splicing analysis by information theory and evaluate competing methods. The bioinformatic aim was to implement a pipeline for alternative splicing analysis comprising the methodological advancement. The biological aim was to apply method and implementation on the example of type-2 diabetes mellitus.

6.1 Expanding the splicing analysis

With ARH the concept of entropy is introduced to the field of splicing prediction. ARH is based on a simple, robust model waging the exon expression ratio deviations in a gene. A deviation in an exon leads to a dominating effect on the entropy and finally to a significant ARH splicing indication. The exon expression ratios take into account probe effects and variable exon mRNA abundances. ARH is rated in the method comparison where it outperforms the existing methods.

This is the first study to comprehensively compare splicing prediction methods on the same platform. In a broad evaluation the performance is assessed of the existing methods on several aspects like robustness in the numbers of replicates, and the dependency on the numbers of exons. The evaluation runs in a biological setting with a tissue data set and in an artificial setting derived of a spike-in experiment.

The compared methods predict exon skipping events. This is in accordance with the Affymetrix Exon Array design. Although many transcripts are already available in major sequence databases in a specific sample under study a previously unknown transcript variant can be present. ARH and the methods under discussion allow for *de novo* predictions amending the transcript databases. However, with a new class of methods estimation of isoform ratios is possible [308, 25, 245]. For most genes the number of combinatorial possible isoforms is far too large for such an isoform quantification. Using the transcript sets from the databases united with ARH predictions the total number of isoforms should be small enough for robust isoform ratio estimations.

Many possibilities of alternative splicing analyses are constrained by the design of the exon arrays. Possibly the next chip generations allow new types of predictions like exon junction analysis, intron retention or different exon splice sites. Assessment of these splice events depends on adequate probe placement and constitution of probe sets. Thus

assessment is ascribed to expression analysis and it is straightforward to use information theory to expand the analysis.

The method evaluation rates pure prediction performance, ignoring the remaining processing steps. In differential expression analyses the methods for assessing expression changes are often evaluated in conjunction with the preprocessing methods. Also for splicing predictions it is possible that the performance of the methods depends on data preprocessing. This would mean to assess a combinatorial number of combinations between preprocessing and prediction methods. The evaluation environment presented in the thesis facilitates such combinations assessments replacing methods by preprocessing-prediction pairs.

The ARH splicing prediction method is developed for exon arrays. However, another high-throughput experimental technology facilitating expression analysis is RNA-Seq [246, 245]. A tissue data set is already available by Wang et al. [306]. Thus, the whole development and evaluation architecture also applies for RNA-Seq data. Actually not all of the discussed methods can be applied. The different technology-dependent characteristics of expression data make adaptations also for ARH indispensable. Efforts for such an adaptation are ongoing.

6.2 Refinement of microarray analysis

Three gene expression evaluation pipelines are presented:

- A differential expression pipeline for 3' gene expression arrays,
- a time series pipeline for 3' gene expression arrays as well as
- an alternative splicing and differential expression pipeline for exon arrays.

Complex data sets are divided into different test cases, for example case control studies. The pipelines have a modular structure allowing for improvements and adaptations for new tasks in microarray analysis. The pipelines are used as standard evaluation tools inside and outside of the Max Planck Institute for Molecular Genetics. Due to the standardisation of the data processing, results are comparable between experiments. Thus, it was possible to integrate different experiments and perform a meta-analysis.

In a test case of differential expression analysis two sets of hybridisations are compared. This setting directly expands for alternative splicing analyses where the same sets of hybridisations are evaluated for differential splicing between two biological conditions with gene-wise and exon-wise evaluation. These two branches, differential expression, and alternative splicing, are assembled into the alternative splicing pipeline for exon arrays.

Due to the modular structure it was possible to adapt the pipeline to time series analyses comparing one set of hybridisations to a phenotype. Thus, a different type of experiment design is exploited with same procedures. Similarly, the pipeline may be adapted for ANOVA. A statistical model using phenotypic parameters may be tested on a set of samples. Genes are identified where expression follows the model, i.e. the combination of phenotypic parameters.

For exon arrays, a preprocessing and statistical evaluation is developed on probe level allowing for a decreased number of replicates in experiments. Improvement of analyses by probe level evaluation is also suggested for 3' gene expression arrays [188, 192, 193, 242]. By reimplementation of the 3' gene expression preprocessing for probe level analysis replicate numbers could be lowered.

The preprocessing of the 3' gene expression as well as the exon arrays comprises four steps which partially have the goal to alleviate experimental effects by statistical means. Still intensities or expressions are quite abstract numbers well comparable for a probe or exon across samples but difficult to relate among genes. The overall goal should be to relate these expression levels to units of measurements, e.g. parts per million or mol per litre. To achieve this goal it will be necessary to improve quantification of measurement curves for optical and chemical estimation of hybridisation values. An example is the probe intensity dependence on the GC content. The GC-RMA correction uses a model for the probe background signal and the MAT correction models binding affinity by GC content. Still probably not all of such effects are quantified and background modelling is developing [109, 100, 255].

Current experimental designs of microarray experiments always relate different samples with each other due to the abstract nature of expression levels. A tight correspondence of expression levels to units of measurements would also advance the use of microarrays for diagnostic applications or generally for analysis of a single biological condition. In such an experiment one sample would provide a report of present genes and the activity of present pathways.

6.3 Type-2 diabetes mellitus with alternative splicing

The tools developed for microarrays and splicing prediction are applied in order to assess the role of alternative splicing in human disease pathology with respect to type-2 diabetes mellitus. Sets of spliced genes are identified in key tissues for type-2 diabetes mellitus. In the GGSC data set test cases relate splicing changes to glycaemic and genetic differences in the samples. For selected results it will be necessary to perform validations by different experimental techniques, e.g. RT-PCR.

Follow-up effects of splicing manifest on the level of proteins. Thus the sets of spliced marker genes have to be compared with the Pfam database for protein families. Spliced exons may contain functional or binding domains. Due to aspects discussed in Chapter 2, splicing changes are not necessarily functionally relevant. However, since distinct biological conditions are compared within the same tissues a connection to type-2 diabetes mellitus is probable.

A collection of genes connected to different biological and genetical context is presented. Each of these results is worth to be continued. In particular some of the genes may relate as targets to therapies presented in Chapter 2. Thus, new approaches for possible treatments may arise from the identification of spliced genes. The comparison of genetically different samples, also in human, could in part explain the genetic prevalence for type-2

diabetes mellitus.

A core set of 655 type-2 diabetes mellitus candidate genes is identified by a meta-analysis of existing data sources. The relation of these genes is explored for disease relevant information and by using over-representation analysis, biological networks are identified on different layers of cellular information such as signalling and metabolic pathways, gene regulatory networks and protein-protein interactions. Still the current meta-analysis strongly depends on transcriptome experiments and other biological levels could improve the identification of marker genes. Since major genome-wide association studies performed in human finished recently it is possible to include genetic marker indication for example by translating the LOD scores to gene score points as new summands for the gene score [84].

The approach at hand includes both, genes with low but consistent expression changes across the different studies as well as strongly differentially expressed genes with respect to a single study. Entropy is an indicator for measuring generality and specificity of a candidate gene with respect to the different studies. The correlation between the score and the entropy is 0.80. However, most of the type-2 diabetes mellitus genes have high entropy and, thus, contribute to expression changes in many of the experiments.

A simple over-representation analysis based on the hypergeometric distribution has been applied in order to characterise the type-2 diabetes mellitus set on the network level including pathways, regulatory networks and protein-protein interactions. In general, there is a high consistency of the results of the overrepresentation analysis when screening different databases. In return the different networks are the base for raising a topological model of type-2 diabetes mellitus, the union of interactions for the 655 marker genes. Such a topological model is a step on a systems biology path for merging the functional genomics data into a mathematical model for type-2 diabetes mellitus and Anja Thormann assembled the data from different databases. Subsequent analysis of the model revealed power law structure similar to the complete known human protein-protein interaction network and low average for shortest paths [297].

Modelling approaches in type-2 diabetes mellitus have to tackle enormous challenges like incorporating genetics, nutrition and mitochondria. For this challenges, no predecessors exist in other diseases. With the advent of high-throughput experiments and the methods of systems biology chances to resolve these issues arose. Partial models exist only for isolated aspects of type-2 diabetes mellitus [168, 32, 148, 159, 191].

References

[1] Diabetes genome anatomy project (2002). URL http://www.diabetesgenome.org/. www.diabetesgenome.org

[2] Alternative Splicing and Disease. Progress in Molecular and Subcellular Biology. Springer (2006)

[3] Human disease and mouse model detail for NIDDM (2006). URL http://www.jax.org/. Jackson Labs

[4] Affymetrix Inc. (2009). URL www.affymetrix.com

[5] Abdueva, D., Wing, M.R., Schaub, B., Triche, T.J.: Experimental comparison and evaluation of the affymetrix exon and u133plus2 genechip arrays. PLoS ONE **2**(9), e913 (2007). DOI 10.1371/journal.pone.0000913. URL http://dx.doi.org/10.1371/journal.pone.0000913

[6] Affymetrix: GeneChip® Expression Analysis. Affymetrix, Inc. (1999). URL www.affymetrix.com

[7] Affymetrix: Microarray Suite User's Guide. Affymetrix, Inc. (2000). URL www.affymetrix.com

[8] Affymetrix: GeneChip® Expression Analysis, Data Analysis Fundamentals. Affymetrix, Inc. (2001). URL www.affymetrix.com

[9] Affymetrix: algorithms description document. Affymetrix, Inc. (2002). URL www.affymetrix.com

[10] Affymetrix Data Sheet: Genechip® exon array system for human, mouse, and rat. Data sheet, Affymetrix, Inc. (2005). URL www.affymetrix.com

[11] Affymetrix GeneChip Exon Array Whitepaper Collection: Alternative transcript analysis methods for exon arrays. Tech. Rep. 1.0, Affymetrix, Inc. (2005). URL www.affymetrix.com

[12] Affymetrix GeneChip Exon Array Whitepaper Collection: Exon array background correction. Tech. Rep. 1.0, Affymetrix, Inc. (2005). URL www.affymetrix.com

[13] Affymetrix GeneChip Exon Array Whitepaper Collection: Exon probeset annotations and transcript cluster groupings. Tech. Rep. 1.0, Affymetrix, Inc. (2005). URL www.affymetrix.com

[14] Affymetrix GeneChip Exon Array Whitepaper Collection: Gene signal estimates from exon arrays. Tech. Rep. 1.0, Affymetrix, Inc. (2005). URL www.affymetrix.com

[15] Affymetrix GeneChip Exon Array Whitepaper Collection: Quality assessment

of exon and gene arrays. Tech. Rep. 1.0, Affymetrix, Inc. (2007). URL www.affymetrix.com

[16] Affymetrix Reference Guide: Statistical algorithms reference guide. Guide, Affymetrix, Inc. (2001). URL www.affymetrix.com

[17] Affymetrix Reference Guide: Statistical algorithms reference guide. Guide, Affymetrix, Inc. (2007). URL www.affymetrix.com

[18] Affymetrix Technical Note: Genechip® exon array design. Technical note, Affymetrix, Inc. (2005). URL www.affymetrix.com

[19] Affymetrix Technical Note: Guide to probe logarithmic intensity error (plier) estimation. Technical note, Affymetrix, Inc. (2005). URL www.affymetrix.com

[20] Affymetrix Technical Note: Identifying and validating alternative splicing events. Technical note, Affymetrix, Inc. (2006). URL www.affymetrix.com

[21] Al-Hasani, H., Joost, H.G.: Nutrition-/diet-induced changes in gene expression in white adipose tissue. Best Pract Res Clin Endocrinol Metab **19**(4), 589–603 (2005). DOI 10.1016/j.beem.2005.07.005. URL http://dx.doi.org/10.1016/j.beem.2005.07.005

[22] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: Molecular biology of the cell. 4th edn. Garland Science (2002)

[23] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J Mol Biol **215**(3), 403–410 (1990). DOI 10.1006/jmbi.1990.9999. URL http://dx.doi.org/10.1006/jmbi.1990.9999

[24] Andrulionyte, L.: Transcription factors as candidate genes for type 2 diabetes. Phd, University of Kuopio (2007)

[25] Anton, M.A., Gorostiaga, D., Guruceaga, E., Segura, V., Carmona-Saez, P., Pascual-Montano, A., Pio, R., Montuenga, L.M., Rubio, A.: SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays. Genome Biol **9**(2), R46 (2008). DOI 10.1186/gb-2008-9-2-r46. URL http://www.ncbi.nlm.nih.gov/pubmed/18312629

[26] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nature Genetics **25**, 25–29 (2000)

[27] Beffa, C.D., Cordero, F., Calogero, R.A.: Dissecting an alternative splicing analysis workflow for genechip exon 1.0 st affymetrix arrays. BMC Genomics **9**, 571 (2008). DOI 10.1186/1471-2164-9-571. URL http://dx.doi.org/10.1186/1471-2164-9-571

[28] Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D.J., Jensen, R.V., Majewski, J.: Gene expression and isoform variation analysis using affymetrix exon arrays. BMC Genomics **9**, 529 (2008). DOI 10.1186/1471-2164-9-529. URL http://dx.doi.org/10.1186/1471-2164-9-529

[29] Ben-Dov, C., Hartmann, B., Lundgren, J., Valcarcel, J.: Genome-wide analysis of alternative pre-mrna splicing. J. Biol. Chem. **283**(3), 1229 – 1233 (2008). DOI 10.1074/jbc.R700033200. URL http://www.ncbi.nlm.nih.gov/pubmed/18024428

[30] Bengtsson, H., Bullard, J., Hansen, K.D.: affxparser: Affymetrix File Parsing SDK (2008). R package version 1.14.0

[31] Bengtsson, H., Simpson, K., Bullard, J., Hansen, K.: aroma. affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Tech. rep., Department of Statistics, University of California, Berkeley (2008)

[32] Bergman, R.N.: Pathogenesis and prediction of diabetes mellitus: lessons from integrative physiology. Mt Sinai J Med **69**(5), 280–290 (2002)

[33] Bertram, L., McQueen, M.B., Mullin, K., Blacker, D., Tanzi, R.E.: Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. Nat Genet **39**(1), 17–23 (2007). 1061-4036 (Print) Journal Article Meta-Analysis Research Support, Non-U.S. Gov't

[34] Bhasi, A., Pandey, R.V., Utharasamy, S.P., Senapathy, P.: Eusplice: a unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes. Bioinformatics **23**(14), 1815–1823 (2007). DOI 10.1093/bioinformatics/btm084. URL http://dx.doi.org/10.1093/bioinformatics/btm084

[35] Biddinger, S.B., Almind, K., Miyazaki, M., Kokkotou, E., Ntambi, J.M., Kahn, C.R.: Effects of diet and genetic background on sterol regulatory element-binding protein-1c, stearoyl-coa desaturase 1, and the development of the metabolic syndrome. Diabetes **54**(5), 1314–23 (2005). 0012-1797 (Print) Journal Article

[36] Bielschowsky, M., Bielschowsky, F.: A new strain of mice with hereditary obesity. In: Proc Univ Otago Med Sch, vol. 31, pp. 29–31 (1953)

[37] Bielschowsky, M., Goodall, C.M.: Origin of inbred nz mouse strains. Cancer Res **30**(3), 834–836 (1970)

[38] Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X.M., Flicek, P., Graf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., Hubbard, T.J.P.: Ensembl 2006. Nucleic Acids Res **34**(Database issue), D556 – 61 (2006). DOI 10.1093/nar/gkj133. URL http://www.ncbi.nlm.nih.gov/pubmed/16381931

[39] Birzele, F., Küffner, R., Meier, F., Oefinger, F., Potthast, C., Zimmer, R.: Prosas: a database for analyzing alternative splicing in the context of protein structures. Nucleic Acids Res **36**(Database issue), D63–D68 (2008). DOI 10.1093/nar/gkm793. URL http://dx.doi.org/10.1093/nar/gkm793

[40] Black, D.L.: Mechanisms of alternative pre-messenger rna splicing. Annu. Rev. Biochem **72**, 291 – 336 (2003). DOI 10.1146/annurev.biochem.72.121801.161720. URL `http://www.ncbi.nlm.nih.gov/pubmed/12626338`

[41] Blencowe, B.J.: Alternative splicing: new insights from global analyses. Cell **126**(1), 37 – 47 (2006). DOI 10.1016/j.cell.2006.06.023. URL `http://www.ncbi.nlm.nih.gov/pubmed/16839875`

[42] Bolstad, B.: Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization. Ph.D. thesis, UNIVERSITY OF CALIFORNIA (2004)

[43] Bolstad, B.: affyPLM: Methods for fitting probe-level models (2007). URL `http://bmbolstad.com`. R package version 1.14.0

[44] Bolstad, B., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R., Speed, T.: Quality assessment of affymetrix genechip data. In: Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., Huber, W. (eds.) Bioinformatics and Computational Biology Solutions using R and Bioconductor, pp. 42–3. Springer, New York (2005)

[45] Bolstad, B.M.: affyio: Tools for parsing Affymetrix data files (2009). R package version 1.10.0

[46] Bolstad, B.M., Irizarry, R.A., Astrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19**(2), 185–93 (2003). Comparative Study Evaluation Studies Journal Article Validation Studies England

[47] Bonizzoni, P., Rizzi, R., Pesole, G.: Computational methods for alternative splicing prediction. Brief Funct Genomic Proteomic **5**(1), 46 – 51 (2006). DOI 10.1093/bfgp/ell011. URL `http://www.ncbi.nlm.nih.gov/pubmed/16769678`

[48] Bracco, L., Throo, E., Cochet, O., Einstein, R., Maurier, F.: Methods and platforms for the quantification of splice variants' expression. Prog Mol Subcell Biol **44**, 1–25 (2006)

[49] Bradley, E.: Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics **7**(1), 1–26 (1979)

[50] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M.: Minimum information about a microarray experiment (miame)-toward standards for microarray data. Nat Genet **29**(4), 365–371 (2001). DOI 10.1038/ng1201-365. URL `http://dx.doi.org/10.1038/ng1201-365`

[51] Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P.: Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett **573**(1-3), 83–92 (2004). DOI 10.1016/j.febslet.2004.07.055. URL `http://dx.doi.org/10.1016/j.febslet.2004.07.055`

[52] Breitling, R., Herzyk, P.: Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. J Bioinform Comput Biol **3**(5), 1171–1189 (2005)

[53] Calarco, J.A., Xing, Y., Caceres, M., Calarco, J.P., Xiao, X., Pan, Q., Lee, C., Preuss, T.M., Blencowe, B.J.: Global analysis of alternative splicing differences between humans and chimpanzees. Genes Dev **21**(22), 2963 – 2975 (2007). DOI 10.1101/gad.1606907. URL `http://www.ncbi.nlm.nih.gov/pubmed/17978102`

[54] Cartegni, L., Chew, S.L., Krainer, A.R.: Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet **3**(4), 285 – 298 (2002). DOI 10.1038/nrg775. URL `http://www.ncbi.nlm.nih.gov/pubmed/11967553`

[55] Castle, J.C., Zhang, C., Shah, J.K., Kulkarni, A.V., Kalsotra, A., Cooper, T.A., Johnson, J.M.: Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet **40**(12), 1416–1425 (2008). DOI 10.1038/ng.264. URL `http://dx.doi.org/10.1038/ng.264`

[56] Castrignano, T., D'Antonio, M., Anselmo, A., Carrabino, D., D'Onorio, D.A., D'Erchia, A.M., Licciulli, F., Mangiulli, M., Mignone, F., Pavesi, G., Picardi, E., Riva, A., Rizzi, R., Bonizzoni, P., Pesole, G.: ASPicDB: a database resource for alternative splicing analysis. Bioinformatics **24**(10), 1300 – 1304 (2008). DOI 10.1093/bioinformatics/btn113. URL `http://www.ncbi.nlm.nih.gov/pubmed/18388144`

[57] Chadt, A., Leicht, K., Deshmukh, A., Jiang, L., Scherneck, S., Bernhardt, U., Dreja, T., Vogel, H., Schmolz, K., Kluge, R., Zierath, J., Hultschig, C., Hoeben, R., Schürmann, A., Joost, H., Al-Hasani, H.: Tbc1d1 mutation in lean mouse strain confers leanness and protects from diet-induced obesity. Nature Genetics **40**(11), 1354–1359 (2008)

[58] Chen, X., Cushman, S., Pannell, L., Hess, S.: Quantitative proteomic analysis of the secretory proteins from rat adipose cells using a 2d liquid chromatography-ms/ms approach. J. Proteome Res. **4**(2), 570–577 (2005)

[59] Choe, S., Boutros, M., Michelson, A., Church, G., Halfon, M.: Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biol **6**(2), R16 (2005)

[60] Christoffels, A., van Gelder, A., Greyling, G., Miller, R., Hide, T., Hide, W.: Stack: Sequence tag alignment and consensus knowledgebase. Nucleic Acids Res **29**(1), 234–238 (2001)

[61] Clark, T., Schweitzer, A., Chen, T., Staples, M., Lu, G., Wang, H., Williams, A., Blume, J.: Discovery of tissue-specific exons using comprehensive human exon microarrays. Genome Biol **8**(4), R64 (2007)

[62] Clark, T.A., Sugnet, C.W., Ares, M.: Genomewide analysis of mrna processing in yeast using splicing-specific microarrays. Science **296**(5569), 907–910 (2002). DOI 10.1126/science.1069415. URL `http://dx.doi.org/10.1126/science.1069415`

[63] Claverie, J.M.: Gene number. what if there are only 30,000 human genes? Science **291**(5507), 1255–1257 (2001)

[64] Clee, S.M., Attie, A.D.: The genetic landscape of type 2 diabetes in mice. Endocr Rev **28**(1), 48 – 83 (2007). DOI 10.1210/er.2006-0035. URL `http://www.ncbi.nlm.nih.gov/pubmed/17018838`

[65] Clevert, D.A., Rasche, A.: Handbook of Research on Systems Biology Applications in Medicine, chap. The Affymetrix GeneChip Microarray Platform, pp. 248–258. IGI Global (2009)

[66] Cline, M.S., Blume, J., Cawley, S., Clark, T.A., Hu, J.S., Lu, G., Salomonis, N., Wang, H., Williams, A.: ANOSVA: a statistical method for detecting splice variation from expression data. Bioinformatics **21 Suppl 1**, i107 – 15 (2005). DOI 10.1093/bioinformatics/bti1010. URL `http://www.ncbi.nlm.nih.gov/pubmed/15961447`

[67] Collins, M.O., Husi, H., Yu, L., Brandon, J.M., Anderson, C.N., Blackstock, W.P., Choudhary, J.S., Grant, S.G.: Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. J Neurochem **97 Suppl 1**, 16–23 (2006). 0022-3042 (Print) Comparative Study Journal Article Research Support, Non-U.S. Gov't

[68] Conlon, E.M., Song, J.J., Liu, A.: Bayesian meta-analysis models for microarray data: a comparative study. BMC Bioinformatics **8**(1), 80 (2007). 1471-2105 (Electronic) Journal article

[69] Cooper, T.A., Wan, L., Dreyfuss, G.: Rna and disease. Cell **136**(4), 777–793 (2009). DOI 10.1016/j.cell.2009.02.011. URL `http://dx.doi.org/10.1016/j.cell.2009.02.011`

[70] Cope, L.M., Irizarry, R.A., Jaffee, H.A., Wu, Z., Speed, T.P.: A benchmark for Affymetrix GeneChip expression measures. Bioinformatics **20**(3), 323 – 331 (2004). DOI 10.1093/bioinformatics/btg410. URL `http://www.ncbi.nlm.nih.gov/pubmed/14960458`

[71] Cover, T.M., Thomas, J.A.: Elements of Information Theory. 2nd edn. Wiley-Interscience (2006)

[72] Coward, E., Haas, S., Vingron, M.: SpliceNest: visualizing gene structure and alternative splicing based on EST clusters. Trends in Genetics **18**(1), 53–55 (2002)

[73] Crofford, O.B., Davis, C.K.: Growth characteristics, glucose tolerance and insulin sensitivity of new zealand obese mice. Metabolism **14**, 271–280 (1965)

[74] Cuperlovic-Culf, M., Belacel, N., Culf, A.S., Ouellette, R.J.: Data analysis of alternative splicing microarrays. Drug Discov Today **11**(21-22), 983 – 990 (2006). DOI 10.1016/j.drudis.2006.09.011. URL `http://www.ncbi.nlm.nih.gov/pubmed/17055407`

[75] Dabney, A., Storey, J., Warnes, G.: q-value: Q-value estimation for false discovery rate control. R package version **1**, — (2006)

[76] Dai, H., Meyer, M., Stepaniants, S., Ziman, M., Stoughton, R.: Use of hybridiza-

tion kinetics for differentiating specific from non-specific binding to oligonucleotide microarrays. Nucleic Acids Res **30**(16), e86 (2002)

[77] Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., Watson, S.J., Meng, F.: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res **33**(20), e175 (2005). DOI 10.1093/nar/gni179. URL `http://www.ncbi.nlm.nih.gov/pubmed/16284200`

[78] Dalma-Weiszhausz, D., Warrington, J., Tanimoto, E., Miyada, C.: DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols, vol. 410, chap. The affymetrix GeneChip platform: an overview., pp. 3–28. Elsevier Inc. (2006). DOI DOI:10.1016/S0076-6879(06)10002-6

[79] Dandona, P., Aljada, A., Bandyopadhyay, A.: Inflammation: the link between insulin resistance, obesity and diabetes. Trends Immunol **25**(1), 4–7 (2004)

[80] Das, D., Clark, T.A., Schweitzer, A., Yamamoto, M., Marr, H., Arribere, J., Minovitsky, S., Poliakov, A., Dubchak, I., Blume, J.E., Conboy, J.G.: A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. Nucleic Acids Res **35**(14), 4845 – 4857 (2007). DOI 10.1093/nar/gkm485. URL `http://www.ncbi.nlm.nih.gov/pubmed/17626050`

[81] Daskalaki, A., Rasche, A.: Informatics in Oral Medicine: Advanced Techniques in Clinical and Diagnostic Technologies, chap. Meta-Analysis approach for the identification of molecular networks related to infections of the oral cavity. IGI Global (—). (in press)

[82] Davies, K.P., Zhao, W., Tar, M., Figueroa, J.C., Desai, P., Verselis, V.K., Kronengold, J., Wang, H.Z., Melman, A., Christ, G.J.: Diabetes-induced changes in the alternative splicing of the slo gene in corporal tissue. Eur Urol **52**(4), 1229–1237 (2007). DOI 10.1016/j.eururo.2006.11.028. URL `http://dx.doi.org/10.1016/j.eururo.2006.11.028`

[83] Dean, L., McEntyre, J.: The Genetic Landscape of Diabetes. NCBI (2004). Open Access Book

[84] Doria, A., Patti, M.E., Kahn, C.R.: The emerging genetic architecture of type 2 diabetes. Cell Metab **8**(3), 186–200 (2008). DOI 10.1016/j.cmet.2008.08.006. URL `http://dx.doi.org/10.1016/j.cmet.2008.08.006`

[85] Dreja, T.: Microarray-basierte expressionsanalysen des weißen fettgewebes der nzo-maus sowie der langerhansschen inseln der nzl-maus: Zwei modelle für das metabolische syndrom. Ph.D. thesis, Mathematisch-Naturwissenschaftliche Fakultät, Universität Potsdam (2009)

[86] Dreja, T., Jovanovic, Z., Rasche, A., Kluge, R., Herwig, R., Tung, Y.C.L., Joost, H.G., Yeo, G.S., Al-Hasani, H.: Diet-induced gene expression of isolated pancreatic islets from a polygenic mouse model for the metabolic syndrome. Diabetologia **53**(2), 309–320 (2009). (Open Access)

[87] Dudoit, S., Yang, Y., Callow, M., Speed, T.: Statistical methods for identifying dif-

ferentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica **12**(1), 111–140 (2002)

[88] Durinck, S., Huber, W., Davis, S.: biomaRt: Interface to BioMart databases (e.g. Ensembl, Wormbase, Gramene and Uniprot) (2009). R package version 1.13.2

[89] English, S.B., Butte, A.J.: Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes. Bioinformatics **23**(21), 2910 – 2917 (2007). DOI 10.1093/bioinformatics/btm483. URL `http://www.ncbi.nlm.nih.gov/pubmed/17921495`

[90] Farris, W., Leissring, M.A., Hemming, M.L., Chang, A.Y., Selkoe, D.J.: Alternative splicing of human insulin-degrading enzyme yields a novel isoform with a decreased ability to degrade insulin and amyloid beta-protein. Biochemistry **44**(17), 6513–25 (2005). AG12749/AG/United States NIA NS046324/NS/United States NINDS Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United States

[91] Faustino, N.A., Cooper, T.A.: Pre-mRNA splicing and human disease. Genes Dev **17**(4), 419 – 437 (2003). DOI 10.1101/gad.1048803. URL `http://www.ncbi.nlm.nih.gov/pubmed/12600935`

[92] Fawcett, T.: Roc graphs: Notes and practical considerations for researchers. Machine Learning **31**, 1–37 (2004)

[93] Fedor, M.J.: Alternative splicing minireview series: combinatorial control facilitates splicing regulation of gene expression and enhances genome diversity. J. Biol. Chem. **283**(3), 1209 – 1210 (2008). DOI 10.1074/jbc.R700046200. URL `http://www.ncbi.nlm.nih.gov/pubmed/18024424`

[94] Fernandez-Real, J.M., Botas-Cervero, P., Lainez, B., Ricart, W., Delgado, E.: An alternatively spliced soluble tnf-alpha receptor is associated with metabolic disorders: a replication study. Clin Immunol **121**(2), 236–41 (2006). Journal Article Research Support, Non-U.S. Gov't United States

[95] Fernández-Real, J.M., Straczkowski, M., Lainez, B., Chacón, M.R., Kowalska, I., López-Bermejo, A., García-España, A., Nikolajuk, A., Kinalska, I., Ricart, W.: An alternative spliced variant of circulating soluble tumor necrosis factor-alpha receptor-2 is paradoxically associated with insulin action. Eur J Endocrinol **154**(5), 723–730 (2006). DOI 10.1530/eje.1.02145. URL `http://dx.doi.org/10.1530/eje.1.02145`

[96] Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., Miller, W.: A computer program for aligning a cdna sequence with a genomic dna sequence. Genome Res **8**(9), 967–974 (1998)

[97] Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D.: Light-directed, spatially addressable parallel chemical synthesis. Science **251**(4995), 767–773 (1991)

[98] Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W., Shields, B., Harries, L.W., Barrett, J.C., Ellard, S., Groves, C.J., Knight, B., Patch, A.M., Ness,

A.R., Ebrahim, S., Lawlor, D.A., Ring, S.M., Ben-Shlomo, Y., Jarvelin, M.R., Sovio, U., Bennett, A.J., Melzer, D., Ferrucci, L., Loos, R.J., Barroso, I., Wareham, N.J., Karpe, F., Owen, K.R., Cardon, L.R., Walker, M., Hitman, G.A., Palmer, C.N., Doney, A.S., Morris, A.D., Smith, G.D., Hattersley, A.T., McCarthy, M.I.: A common variant in the fto gene is associated with body mass index and predisposes to childhood and adult obesity. Science **316**(5826), 889–94 (2007). 1095-9203 (Electronic) Journal Article Research Support, Non-U.S. Gov't

[99] French, P.J., Peeters, J., Horsman, S., Duijm, E., Siccama, I., van den Bent, M.J., Luider, T.M., Kros, J.M., van der Spek, P., Smitt, P.A.S.: Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. Cancer Res **67**(12), 5635–5642 (2007). DOI 10.1158/0008-5472.CAN-06-2869. URL `http://dx.doi.org/10.1158/0008-5472.CAN-06-2869`

[100] Furusawa, C., Ono, N., Suzuki, S., Agata, T., Shimizu, H., Yomo, T.: Model-based analysis of non-specific binding for background correction of high-density oligonucleotide microarrays. Bioinformatics **25**(1), 36–41 (2009). DOI 10.1093/bioinformatics/btn570. URL `http://dx.doi.org/10.1093/bioinformatics/btn570`

[101] Galperin, M.Y.: The molecular biology database collection: 2008 update. Nucleic Acids Res **36**(Database issue), D2–D4 (2008). DOI 10.1093/nar/gkm1037. URL `http://dx.doi.org/10.1093/nar/gkm1037`

[102] Garcia-Blanco, M.: Alternative Splicing and Disease, vol. 44, chap. Alternative splicing: therapeutic target and tool, pp. 47–64. Springer (2006)

[103] Garcia-Blanco, M., Baraniak, A., Lasda, E.: Alternative splicing in disease and therapy. Nature Biotechnology **22**(5), 535–546 (2004)

[104] Gardina, P.J., Clark, T.A., Shimada, B., Staples, M.K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., Davies, C., Williams, A., Turpaz, Y.: Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics **7**, 325 (2006). DOI 10.1186/1471-2164-7-325. URL `http://www.ncbi.nlm.nih.gov/pubmed/17192196`

[105] Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A.: affy—analysis of affymetrix genechip data at the probe level. Bioinformatics **20**(3), 307–315 (2004). DOI http://dx.doi.org/10.1093/bioinformatics/btg405

[106] Gentleman, R.: geneplotter: Graphics related functions for bioconductor. BioConductor package collection **-**, – (2006). R package version 1.10.0

[107] Gentleman, R., Biocore: geneplotter: Grapics related functions for Bioconductor (2009). R package version 1.20.0

[108] Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Li, F.L.C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H., Zhang, J.: Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology **5**, R80 (2004)

[109] Gharaibeh, R.Z., Fodor, A.A., Gibas, C.J.: Background correction using dinucleotide affinities improves the performance of gcrma. BMC Bioinformatics **9**, 452 (2008). DOI 10.1186/1471-2105-9-452. URL http://dx.doi.org/10.1186/1471-2105-9-452

[110] Goeman, J.J., Buhlmann, P.: Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics **23**(8), 980 – 987 (2007). DOI 10.1093/bioinformatics/btm051. URL http://www.ncbi.nlm.nih.gov/pubmed/17303618

[111] Goldstrohm, A.C., Greenleaf, A.L., Garcia-Blanco, M.A.: Co-transcriptional splicing of pre-messenger rnas: considerations for the mechanism of alternative splicing. Gene **277**(1-2), 31–47 (2001)

[112] Gopalan, V., Tan, T.W., Lee, B.T.K., Ranganathan, S.: Xpro: database of eukaryotic protein-encoding genes. Nucleic Acids Res **32**(Database issue), D59 – 63 (2004). DOI 10.1093/nar/gkh051. URL http://www.ncbi.nlm.nih.gov/pubmed/14681359

[113] de la Grange, P., Dutertre, M., Martin, N., Auboeuf, D.: FAST DB: a website resource for the study of the expression regulation of human gene products. Nucleic Acids Res **33**(13), 4276 – 4284 (2005). DOI 10.1093/nar/gki738. URL http://www.ncbi.nlm.nih.gov/pubmed/16052034

[114] Grant, S.F.A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., Styrkarsdottir, U., Magnusson, K.P., Walters, G.B., Palsdottir, E., Jonsdottir, T., Gudmundsdottir, T., Gylfason, A., Saemundsdottir, J., Wilensky, R.L., Reilly, M.P., Rader, D.J., Bagger, Y., Christiansen, C., Gudnason, V., Sigurdsson, G., Thorsteinsdottir, U., Gulcher, J.R., Kong, A., Stefansson, K.: Variant of transcription factor 7-like 2 (tcf7l2) gene confers risk of type 2 diabetes. Nat Genet **38**(3), 320–323 (2006). 1061-4036 10.1038/ng1732 10.1038/ng1732

[115] Graveley, B.: Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics **17**(2), 100–107 (2001)

[116] Guilherme, A., Virbasius, J.V., Puri, V., Czech, M.P.: Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes. Nat Rev Mol Cell Biol **9**(5), 367–377 (2008). DOI 10.1038/nrm2391. URL http://dx.doi.org/10.1038/nrm2391

[117] Gunton, J.E., Kulkarni, R.N., Yim, S., Okada, T., Hawthorne, W.J., Tseng, Y.H., Roberson, R.S., Ricordi, C., O'Connell, P.J., Gonzalez, F.J., Kahn, C.R.: Loss of arnt/hif1beta mediates altered gene expression and pancreatic-islet dysfunction in human type 2 diabetes. Cell **122**(3), 337–49 (2005). 0092-8674 (Print) Journal Article

[118] Gupta, S., Vingron, M., Haas, S.A.: T-stag: resource and web-interface for tissue-specific transcripts and genes. Nucleic Acids Res **33**(Web Server issue), W654–W658 (2005). DOI 10.1093/nar/gki350. URL http://dx.doi.org/10.1093/nar/gki350

[119] Gupta, S., Zink, D., Korn, B., Vingron, M., Haas, S.A.: Genome wide identification

and classification of alternative splicing based on EST data. Bioinformatics **20**(16), 2579 – 2585 (2004). DOI 10.1093/bioinformatics/bth288. URL http://www.ncbi.nlm.nih.gov/pubmed/15117759

[120] Gupta, S., Zink, D., Korn, B., Vingron, M., Haas, S.A.: Strengths and weaknesses of est-based prediction of tissue-specific alternative splicing. BMC Genomics **5**(1), 72 (2004). DOI 10.1186/1471-2164-5-72. URL http://dx.doi.org/10.1186/1471-2164-5-72

[121] Haas, S.A., Beissbarth, T., Rivals, E., Krause, A., Vingron, M.: Genenest: automated generation and visualization of gene indices. Trends Genet **16**(11), 521–523 (2000)

[122] Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., Consortium, G.O.: The gene ontology (go) database and informatics resource. Nucleic Acids Res **32**(Database issue), D258–D261 (2004)

[123] Hekstra, D., Taussig, A., Magnasco, M., Naef, F.: Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. Nucleic Acids Research **31**(7), 1962 (2003)

[124] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R.: Intact: an open source molecular interaction database. Nucleic Acids Res **32**(Database issue), D452–5 (2004). 1362-4962 (Electronic) Journal Article

[125] Hertel, K.J.: Combinatorial control of exon recognition. J. Biol. Chem. **283**(3), 1211 – 1215 (2008). DOI 10.1074/jbc.R700035200. URL http://www.ncbi.nlm.nih.gov/pubmed/18024426

[126] von Heydebreck, A., Huber, W., Gentleman, R.: Differential Expression with the Bioconductor Project. Bioconductor Project Working Papers **7**, 1–17 (2004)

[127] Hochreiter, S., Clevert, D.A., Obermayer, K.: A new summarization method for Affymetrix probe level data. Bioinformatics **22**(8), 943 – 949 (2006). DOI 10.1093/bioinformatics/btl033. URL http://www.ncbi.nlm.nih.gov/pubmed/16473874

[128] Hollander, M., Wolfe, D.: Nonparametric statistical methods. John Wiley New York, N. Y (1973)

[129] Holste, D., Huo, G., Tung, V., Burge, C.B.: HOLLYWOOD: a comparative relational database of alternative splicing. Nucleic Acids Res **34**(Database issue),

D56 – 62 (2006). DOI 10.1093/nar/gkj048. URL http://www.ncbi.nlm.nih.gov/pubmed/16381932

[130] Hong, F., Breitling, R., McEntee, C., Wittner, B., Nemhauser, J., Chory, J.: RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics **22**(22), 2825 (2006)

[131] Hopkins, A.L., Groom, C.R.: The druggable genome. Nat Rev Drug Discov **1**(9), 727–30 (2002). 1474-1776 (Print) Journal Article

[132] Hothorn, T., Hornik, K.: exactRankTests: Exact Distributions for Rank and Permutation Tests (2006). R package version 0.8-18

[133] Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., Pierre, S.S., Twigger, S., White, O., Rhee, S.Y.: Big data: The future of biocuration. Nature **455**(7209), 47–50 (2008). DOI 10.1038/455047a. URL http://dx.doi.org/10.1038/455047a

[134] Hu, G.K., Madore, S.J., Moldover, B., Jatkoe, T., Balaban, D., Thomas, J., Wang, Y.: Predicting splice variant from dna chip expression data. Genome Res **11**(7), 1237–1245 (2001). DOI 10.1101/gr.165501. URL http://dx.doi.org/10.1101/gr.165501,http://www.ncbi.nlm.nih.gov/pubmed/11435406

[135] Huang, H.D., Horng, J.T., Lee, C.C., Liu, B.J.: Prosplicer: a database of putative alternative splicing information derived from protein, mrna and expressed sequence tag sequence data. Genome Biol **4**(4), R29 (2003)

[136] Huang, H.D., Horng, J.T., Lin, F.M., Chang, Y.C., Huang, C.C.: Spliceinfo: an information repository for mrna alternative splicing in human genome. Nucleic Acids Res **33**(Database issue), D80–D85 (2005). DOI 10.1093/nar/gki129. URL http://dx.doi.org/10.1093/nar/gki129

[137] Huang, Y.H., Chen, Y.T., Lai, J.J., Yang, S.T., Yang, U.C.: Pals db: Putative alternative splicing database. Nucleic Acids Res **30**(1), 186–190 (2002)

[138] Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., Clamp, M.: The ensembl genome database project. Nucleic Acids Res **30**(1), 38–41 (2002)

[139] Huber, W., Gentleman, R.: matchprobes: a bioconductor package for the sequence-matching of microarray probe elements. Bioinformatics **20**(10), 1651–1652 (2004). DOI 10.1093/bioinformatics/bth133. URL http://dx.doi.org/10.1093/bioinformatics/bth133

[140] Huber, W., von Heydebreck, A., Vingron, M.: Analysis of microarray gene expression data. In: Balding, D.J., Bishop, M., Cannings, C. (eds.) Handbook of Statistical Genetics, 2nd edn., pp. 162–187. John Wiley & Sons, Ltd. (2003)

[141] Hull, J., Campino, S., Rowlands, K., Chan, M.S., Copley, R.R., Taylor, M.S., Rockett, K., Elvidge, G., Keating, B., Knight, J., Kwiatkowski, D.: Identification

of common genetic variation that modulates alternative splicing. PLoS Genet **3**(6), e99 (2007). DOI 10.1371/journal.pgen.0030099. URL http://dx.doi.org/10.1371/journal.pgen.0030099

[142] Illumina Systems and Software, Technology Spotlight: DNA Sequencing with Solexa Technology. Technology spotlight, Illumina, Inc. (2007). URL http://www.illumina.com/

[143] Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., Yura, K., Miyazaki, S., Ikeo, K., Homma, K., Kasprzyk, A., Nishikawa, T., Hirakawa, M., Thierry-Mieg, J., Thierry-Mieg, D., Ashurst, J., Jia, L., Nakao, M., Thomas, M.A., Mulder, N., Karavidopoulou, Y., Jin, L., Kim, S., Yasuda, T., Lenhard, B., Eveno, E., Suzuki, Y., Yamasaki, C., ichi Takeda, J., Gough, C., Hilton, P., Fujii, Y., Sakai, H., Tanaka, S., Amid, C., Bellgard, M., de Fatima Bonaldo, M., Bono, H., Bromberg, S.K., Brookes, A.J., Bruford, E., Carninci, P., Chelala, C., Couillault, C., de Souza, S.J., Debily, M.A., Devignes, M.D., Dubchak, I., Endo, T., Estreicher, A., Eyras, E., Fukami-Kobayashi, K., Gopinath, G.R., Graudens, E., Hahn, Y., Han, M., Han, Z.G., Hanada, K., Hanaoka, H., Harada, E., Hashimoto, K., Hinz, U., Hirai, M., Hishiki, T., Hopkinson, I., Imbeaud, S., Inoko, H., Kanapin, A., Kaneko, Y., Kasukawa, T., Kelso, J., Kersey, P., Kikuno, R., Kimura, K., Korn, B., Kuryshev, V., Makalowska, I., Makino, T., Mano, S., Mariage-Samson, R., Mashima, J., Matsuda, H., Mewes, H.W., Minoshima, S., Nagai, K., Nagasaki, H., Nagata, N., Nigam, R., Ogasawara, O., Ohara, O., Ohtsubo, M., Okada, N., Okido, T., Oota, S., Ota, M., Ota, T., Otsuki, T., Piatier-Tonneau, D., Poustka, A., Ren, S.X., Saitou, N., Sakai, K., Sakamoto, S., Sakate, R., Schupp, I., Servant, F., Sherry, S., Shiba, R., Shimizu, N., Shimoyama, M., Simpson, A.J., Soares, B., Steward, C., Suwa, M., Suzuki, M., Takahashi, A., Tamiya, G., Tanaka, H., Taylor, T., Terwilliger, J.D., Unneberg, P., Veeramachaneni, V., Watanabe, S., Wilming, L., Yasuda, N., Yoo, H.S., Stodolsky, M., Makalowski, W., Go, M., Nakai, K., Takagi, T., Kanehisa, M., Sakaki, Y., Quackenbush, J., Okazaki, Y., Hayashizaki, Y., Hide, W., Chakraborty, R., Nishikawa, K., Sugawara, H., Tateno, Y., Chen, Z., Oishi, M., Tonellato, P., Apweiler, R., Okubo, K., Wagner, L., Wiemann, S., Strausberg, R.L., Isogai, T., Auffray, C., Nomura, N., Gojobori, T., Sugano, S.: Integrative annotation of 21,037 human genes validated by full-length cdna clones. PLoS Biol **2**(6), e162 (2004). DOI 10.1371/journal.pbio.0020162. URL http://dx.doi.org/10.1371/journal.pbio.0020162

[144] Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., Speed, T.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**(2), 249 (2003)

[145] Irizarry, R., Wu, Z., Jaffee, H.: Comparison of Affymetrix GeneChip expression measures. Bioinformatics **22**(7), 789–794 (2006)

[146] Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P.: Summaries of affymetrix genechip probe level data. Nucleic Acids Research **31**(4), e15 (2003)

[147] Jensen, L.J., Steinmetz, L.M.: Re-analysis of data and its integration. FEBS Letters **579**(8), 1802–1807 (2005)

[148] Jiang, N., Cox, R.D., Hancock, J.M.: A kinetic core model of the glucose-stimulated insulin secretion network of pancreatic beta cells. Mamm Genome **18**(6-7), 508 – 520 (2007). DOI 10.1007/s00335-007-9011-y. URL `http://www.ncbi.nlm.nih.gov/pubmed/17514510`

[149] Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., Shoemaker, D.D.: Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science **302**(5653), 2141 – 2144 (2003). DOI 10.1126/science.1090100. URL `http://www.ncbi.nlm.nih.gov/pubmed/14684825`

[150] Johnson, W.E., Li, W., Meyer, C.A., Gottardo, R., Carroll, J.S., Brown, M., Liu, X.S.: Model-based analysis of tiling-arrays for ChIP-chip. Proc. Natl. Acad. Sci. USA **103**(33), 12,457 – 12,462 (2006). DOI 10.1073/pnas.0601180103. URL `http://www.ncbi.nlm.nih.gov/pubmed/16895995`

[151] Joshi-Tope, G., Gillespie, M., Vastrik, I., DEustachio, P., Schmidt, E., Bono, B.d., Jassal, B., Gopinath, G., Wu, G., Matthews, L., Lewis, S., Birney, E., Stein, L.: Reactome: a knowledgebase of biological pathways. Nucl. Acids Res. **33**(suppl 1), 428–432 (2005)

[152] Jurgens, H., Schurmann, A., Kluge, R., Ortmann, S., Klaus, S., Joost, H., Tschop, M.: Hyperphagia, lower body temperature, and reduced running wheel activity precede development of morbid obesity in New Zealand obese mice. Physiological Genomics **25**(2), 234 (2006)

[153] Kahn, B.B., Flier, J.S.: Obesity and insulin resistance. J Clin Invest **106**(4), 473–81 (2000). 0021-9738 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review

[154] Kahn, S.E., Hull, R.L., Utzschneider, K.M.: Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature **444**(7121), 840–846 (2006). DOI 10.1038/nature05482. URL `http://dx.doi.org/10.1038/nature05482`

[155] Kamburov, A., Wierling, C., Lehrach, H., Herwig, R.: Consensuspathdb–a database for integrating human functional interaction networks. Nucleic Acids Res **37**(Database issue), D623–D628 (2009). DOI 10.1093/nar/gkn698. URL `http://dx.doi.org/10.1093/nar/gkn698`

[156] Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., Tammana, H., Gingeras, T.R.: Novel rnas identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res **14**(3), 331–342 (2004). DOI 10.1101/gr.2094104. URL `http://dx.doi.org/10.1101/gr.2094104`

[157] Kan, Z., States, D., Gish, W.: Selecting for functional alternative splices in ests. Genome Res **12**(12), 1837–1845 (2002). DOI 10.1101/gr.764102. URL `http://dx.doi.org/10.1101/gr.764102`

[158] Kanehisa, M., Goto, S.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research **28**(1), 27–30 (2000)

[159] Kansal, A.R.: Modeling approaches to type 2 diabetes. Diabetes Technol Ther **6**(1), 39–47 (2004). DOI 10.1089/152091504322783396. URL http://dx.doi.org/10.1089/152091504322783396

[160] Kapur, K., Jiang, H., Xing, Y., Wong, W.H.: Cross-hybridization modeling on affymetrix exon arrays. Bioinformatics **24**(24), 2887–2893 (2008). DOI 10.1093/bioinformatics/btn571. URL http://dx.doi.org/10.1093/bioinformatics/btn571

[161] Kapur, K., Xing, Y., Ouyang, Z., Wong, W.H.: Exon arrays provide accurate assessments of gene expression. Genome Biol **8**(5), R82 (2007). DOI 10.1186/gb-2007-8-5-r82. URL http://www.ncbi.nlm.nih.gov/pubmed/17504534

[162] Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., Birney, E.: Ensmart: a generic system for fast and flexible access to biological data. Genome Res **14**(1), 160–169 (2004). DOI 10.1101/gr.1645104. URL http://dx.doi.org/10.1101/gr.1645104

[163] Kent, W.J.: Blat–the blast-like alignment tool. Genome Res **12**(4), 656–664 (2002). DOI 10.1101/gr.229202.ArticlepublishedonlinebeforeMarch2002. URL http://dx.doi.org/10.1101/gr.229202.ArticlepublishedonlinebeforeMarch2002

[164] Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: The human genome browser at ucsc. Genome Res **12**(6), 996–1006 (2002). DOI 10.1101/gr.229102. ArticlepublishedonlinebeforeprintinMay2002. URL http://dx.doi.org/10.1101/gr.229102.ArticlepublishedonlinebeforeprintinMay2002

[165] Kim, N., Alekseyenko, A.V., Roy, M., Lee, C.: The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. Nucleic Acids Res **35**(Database issue), D93 – 8 (2007). DOI 10.1093/nar/gkl884. URL http://www.ncbi.nlm.nih.gov/pubmed/17108355

[166] Kim, N., Shin, S., Lee, S.: ECgene: genome-based EST clustering and gene modeling for alternative splicing. Genome Res **15**(4), 566 – 576 (2005). DOI 10.1101/gr.3030405. URL http://www.ncbi.nlm.nih.gov/pubmed/15805497

[167] Kim, P., Kim, N., Lee, Y., Kim, B., Shin, Y., Lee, S.: ECgene: genome annotation for alternative splicing. Nucleic Acids Res **33**(Database issue), D75 – 9 (2005). DOI 10.1093/nar/gki118. URL http://www.ncbi.nlm.nih.gov/pubmed/15608289

[168] Kitano, H., Oda, K., Kimura, T., Matsuoka, Y., Csete, M., Doyle, J., Muramatsu, M.: Metabolic syndrome and robustness tradeoffs. Diabetes **53 Suppl 3**, S6–S15 (2004). 0012-1797 (Print) Journal Article Review

[169] Klipp, E., Herwig, R., Kowald, A., Wierling, C., Lehrach, H.: Systems Biology in Practice. Wiley-VCH (2005)

[170] Knudsen, T.B., Daston, G.P., Society, T.: Miame guidelines. Reprod Toxicol **19**(3), 263 (2005)

[171] Kong, W., Po, S., Yamagishi, T., Ashen, M.D., Stetten, G., Tomaselli, G.F.: Isolation and characterization of the human gene encoding ito: further diversity by alternative mrna splicing. Am J Physiol **275**(6 Pt 2), H1963–H1970 (1998)

[172] Koscielny, G., Texier, V.L., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M., Harrington, E., Boué, S., Eyras, E., Plass, M., Lopez, F., Ritchie, W., Moucadel, V., Ara, T., Pospisil, H., Herrmann, A., Reich, J.G., Guigó, R., Bork, P., von Knebel Doeberitz, M., Vilo, J., Hide, W., Apweiler, R., Thanaraj, T.A., Gautheret, D.: Astd: The alternative splicing and transcript diversity database. Genomics **93**(3), 213–220 (2009). DOI 10.1016/j.ygeno.2008.11.003. URL `http://dx.doi.org/10.1016/j.ygeno.2008.11.003`

[173] Kreil, D., Russell, R.: Tutorial section: There is no silver bullet-a guide to low-level data transforms and normalisation methods for microarray data. Briefings in Bioinformatics **6**(1), 86–97 (2005)

[174] Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A., Wingender, E.: Transpath: an integrated database on signal transduction and a tool for array analysis. Nucleic Acids Res **31**(1), 97–100 (2003)

[175] Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S., Golub, T.R.: The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science **313**(5795), 1929–1935 (2006). DOI 10.1126/science.1132939. URL `http://dx.doi.org/10.1126/science.1132939`

[176] Lan, H., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Mui, E.T., Flowers, M.T., Schueler, K.L., Manly, K.F., Williams, R.W., Kendziorski, C., Attie, A.D.: Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genet **2**(1), e6 (2006). 5803701/phs 66369/phs Hl56593/hl/nhlbi Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't United States

[177] Lan, H., Rabaglia, M.E., Stoehr, J.P., Nadler, S.T., Schueler, K.L., Zou, F., Yandell, B.S., Attie, A.D.: Gene expression profiles of nondiabetic and diabetic obese mice suggest a role of hepatic lipogenic capacity in diabetes susceptibility. Diabetes **52**(3), 688–700 (2003)

[178] Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray,

A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., Consortium, I.H.G.S.: Initial sequencing and analysis of the human genome. Nature **409**(6822), 860–921 (2001)

[179] Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L.: Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol **10**(3), R25 (2009). DOI 10.1186/gb-2009-10-3-r25. URL `http://dx.doi.org/10.1186/gb-2009-10-3-r25`

[180] Le, K., Mitsouras, K., Roy, M., Wang, Q., Xu, Q., Nelson, S.F., Lee, C.: Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. Nucleic Acids Res **32**(22), e180 (2004). DOI 10.1093/nar/gnh173. URL `http://www.ncbi.nlm.nih.gov/pubmed/15598820`

[181] Le, T.V., Riethoven, J.J., Kumanduri, V., Gopalakrishnan, C., Lopez, F., Gautheret, D., Thanaraj, T.A.: AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. BMC Bioinformatics **7**, 169 (2006). DOI 10.1186/1471-2105-7-169. URL `http://www.ncbi.nlm.nih.gov/pubmed/16556303`

[182] Lee, C., Atanelov, L., Modrek, B., Xing, Y.: Asap: the alternative splicing annotation project. Nucleic Acids Res **31**(1), 101–105 (2003)

[183] Lee, C., Roy, M.: Analysis of alternative splicing with microarrays: successes and challenges. Genome Biol **5**(7), 231 (2004). DOI 10.1186/gb-2004-5-7-231. URL `http://dx.doi.org/10.1186/gb-2004-5-7-231`

[184] Lee, C., Wang, Q.: Bioinformatics analysis of alternative splicing. Brief Bioinform **6**(1), 23–33 (2005)

[185] Lee, Y., Lee, Y., Kim, B., Shin, Y., Nam, S., Kim, P., Kim, N., Chung, W.H., Kim, J., Lee, S.: ECgene: an alternative splicing database update. Nucleic Acids Res **35**(Database issue), D99 – 103 (2007). DOI 10.1093/nar/gkl992. URL `http://www.ncbi.nlm.nih.gov/pubmed/17132829`

[186] Lefai, E., Roques, M., Vega, N., Laville, M., Vidal, H.: Expression of the splice variants of the p85alpha regulatory subunit of phosphoinositide 3-kinase in muscle and adipose tissue of healthy subjects and type 2 diabetic patients. Biochem J **360**(Pt 1), 117–26 (2001). Journal Article Research Support, Non-U.S. Gov't England

[187] Leipzig, J., Pevzner, P., Heber, S.: The alternative splicing gallery (asg): bridging the gap between genome and transcriptome. Nucleic Acids Res **32**(13), 3977–3983 (2004). DOI 10.1093/nar/gkh731. URL `http://dx.doi.org/10.1093/nar/gkh731`

[188] Lemieux, S.: Probe-level linear model fitting and mixture modeling results in high accuracy detection of differential gene expression. BMC Bioinformatics **7**, 391 (2006). DOI 10.1186/1471-2105-7-391. URL `http://dx.doi.org/10.1186/1471-2105-7-391`

[189] Li, C., Wong, W.H.: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A **98**(1), 31–6 (2001). 1r01hg02341-01/hg/nhgri Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. United States

[190] Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., Darnell, R.B.: Hits-clip yields genome-wide insights into brain alternative rna processing. Nature **456**(7221), 464–469 (2008). DOI 10.1038/nature07488. URL `http://dx.doi.org/10.1038/nature07488`

[191] Liu, M., Liberzon, A., Kong, S.W., Lai, W.R., Park, P.J., Kohane, I.S., Kasif, S.: Network-based analysis of affected biological processes in type 2 diabetes models. PLoS Genet **3**(6), e96 (2007). DOI 10.1371/journal.pgen.0030096. URL `http://www.ncbi.nlm.nih.gov/pubmed/17571924`

[192] Liu, X., Lin, K.K., Andersen, B., Rattray, M.: Including probe-level uncertainty in model-based gene expression clustering. BMC Bioinformatics **8**, 98 (2007). DOI 10.1186/1471-2105-8-98. URL http://dx.doi.org/10.1186/1471-2105-8-98

[193] Liu, X., Milo, M., Lawrence, N.D., Rattray, M.: Probe-level measurement error improves accuracy in detecting differential gene expression. Bioinformatics **22**(17), 2107–2113 (2006). DOI 10.1093/bioinformatics/btl361. URL http://dx.doi.org/10.1093/bioinformatics/btl361

[194] Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., Brown, E.L.: Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol **14**(13), 1675–1680 (1996). DOI 10.1038/nbt1296-1675. URL http://dx.doi.org/10.1038/nbt1296-1675

[195] Makrantonaki, E., Adjaye, J., Herwig, R., Brink, T.C., Groth, D., Hultschig, C., Lehrach, H., Zouboulis, C.C.: Age-specific hormonal decline is accompanied by transcriptional changes in human sebocytes in vitro. Aging Cell **5**(4), 331–344 (2006). DOI 10.1111/j.1474-9726.2006.00223.x. URL http://dx.doi.org/10.1111/j.1474-9726.2006.00223.x

[196] Maniatis, T., Tasic, B.: Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature **418**, 236–243 (2002)

[197] Marguerat, S., Wilhelm, B.T., Bähler, J.: Next-generation sequencing: applications beyond genomes. Biochem Soc Trans **36**(Pt 5), 1091–1096 (2008). DOI 10.1042/BST0361091. URL http://dx.doi.org/10.1042/BST0361091

[198] Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y.: Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res **18**(9), 1509–1517 (2008). DOI 10.1101/gr.079558.108. URL http://dx.doi.org/10.1101/gr.079558.108

[199] Matlin, A.J., Clark, F., Smith, C.W.J.: Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol **6**(5), 386 – 398 (2005). DOI 10.1038/nrm1645. URL http://www.ncbi.nlm.nih.gov/pubmed/15956978

[200] Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., Wingender, E.: Transfac: transcriptional regulation, from patterns to profiles. Nucleic Acids Res **31**(1), 374–378 (2003)

[201] Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E.: Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res **34**(Database issue), D108–10 (2006). 1362-4962 (Electronic) Journal Article

[202] McInerney, M.F., Najjar, S.M., Brickley, D., Lutzke, M., Abou-Rjaily, G.A., Reifsnyder, P., Haskell, B.D., Flurkey, K., Zhang, Y.J., Pietropaolo, S.L., Pietropaolo, M., Byers, J.P., Leiter, E.H.: Anti-insulin receptor autoantibodies are not requi-

red for type 2 diabetes pathogenesis in nzl/lt mice, a new zealand obese (nzo)-derived mouse strain. Exp Diabesity Res **5**(3), 177–185 (2004). DOI 10.1080/15438600490478029. URL http://dx.doi.org/10.1080/15438600490478029

[203] Minn, A.H., Lan, H., Rabaglia, M.E., Harlan, D.M., Peculis, B.A., Attie, A.D., Shalev, A.: Increased insulin translation from an insulin splice-variant overexpressed in diabetes, obesity, and insulin resistance. Mol Endocrinol **19**(3), 794–803 (2005). 58037-01/United States PHS Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. United States

[204] Mitchell, J.A., Aronson, A.R., Mork, J.G., Folk, L.C., Humphrey, S.M., Ward, J.M.: Gene indexing: characterization and analysis of nlms generifs. AMIA Annu Symp Proc —, 460 – 464 (2003)

[205] Modrek, B., Lee, C.: A genomic view of alternative splicing. Nat Genet **30**(1), 13–19 (2002). DOI 10.1038/ng0102-13. URL http://dx.doi.org/10.1038/ng0102-13

[206] Monsalve, M., Wu, Z., Adelmant, G., Puigserver, P., Fan, M., Spiegelman, B.M.: Direct coupling of transcription and mrna processing through the thermogenic coactivator pgc-1. Mol Cell **6**(2), 307–316 (2000)

[207] Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., Groop, L.C.: Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet **34**(3), 267–73 (2003). 1061-4036 (Print) Journal Article

[208] Muers, M.: RNA: splicing, counting, coordinating and controlling the alternatives. Nature Reviews Genetics —, — (2008)

[209] Muoio, D.M., Newgard, C.B.: Mechanisms of disease: molecular and metabolic mechanisms of insulin resistance and beta-cell failure in type 2 diabetes. Nat Rev Mol Cell Biol **9**(3), 193–205 (2008). DOI 10.1038/nrm2327. URL http://dx.doi.org/10.1038/nrm2327

[210] Mutch, D.M., Clement, K.: Unraveling the genetics of human obesity. PLoS Genet **2**(12), e188 (2006). DOI 10.1371/journal.pgen.0020188. URL http://www.ncbi.nlm.nih.gov/pubmed/17196040

[211] Nadler, S.T., Stoehr, J.P., Schueler, K.L., Tanimoto, G., Yandell, B.S., Attie, A.D.: The expression of adipogenic genes is decreased in obesity and diabetes mellitus. PNAS **97**(21), 11,371–11,376 (2000)

[212] Naef, F., Magnasco, M.: Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. Physical Review E **68**(1), 11,906 (2003)

[213] Naef, F., Socci, N.D., Magnasco, M.: A study of accuracy and precision in oligonucleotide arrays: extracting more signal at large concentrations. Bioinformatics **19**(2), 178–184 (2003)

[214] Nandi, A., Kitamura, Y., Kahn, C.R., Accili, D.: Mouse models of insulin resistance. Physiol Rev **84**(2), 623 – 647 (2004). DOI 10.1152/physrev.00032.2003. URL http://www.ncbi.nlm.nih.gov/pubmed/15044684

[215] Nembaware, V., Lupindo, B., Schouest, K., Spillane, C., Scheffler, K., Seoighe, C.: Genome-wide survey of allele-specific splicing in humans. BMC Genomics **9**, 265 (2008). DOI 10.1186/1471-2164-9-265. URL http://dx.doi.org/10.1186/1471-2164-9-265

[216] Nilsen, T.W.: The spliceosome: the most complex macromolecular machine in the cell? Bioessays **25**(12), 1147 – 1149 (2003). DOI 10.1002/bies.10394. URL http://www.ncbi.nlm.nih.gov/pubmed/14635248

[217] Noh, S.J., Lee, K., Paik, H., Hur, C.G.: TISA: tissue-specific alternative splicing in human and mouse genes. DNA Res **13**(5), 229 – 243 (2006). DOI 10.1093/dnares/dsl011. URL http://www.ncbi.nlm.nih.gov/pubmed/17107969

[218] Novoyatleva, T., Tang, Y., Rafalska, I., Stamm, S.: Alternative Splicing and Disease, vol. 44, chap. Pre-mRNA missplicing as a cause of human disease, pp. 27–46. Springer (2006)

[219] Odom, D.T., Dowell, R.D., Jacobsen, E.S., Nekludova, L., Rolfe, P.A., Danford, T.W., Gifford, D.K., Fraenkel, E., Bell, G.I., Young, R.A.: Core transcriptional regulatory circuitry in human hepatocytes. Mol Syst Biol **2**, 2006 0017 (2006). 1744-4292 (Electronic) Journal Article

[220] Odom, D.T., Zizlsperger, N., Gordon, D.B., Bell, G.W., Rinaldi, N.J., Murray, H.L., Volkert, T.L., Schreiber, J., Rolfe, P.A., Gifford, D.K., Fraenkel, E., Bell, G.I., Young, R.A.: Control of pancreas and liver gene expression by hnf transcription factors. Science **303**(5662), 1378–81 (2004). 1095-9203 (Electronic) Journal Article

[221] Ogawa, W., Kasuga, M.: Cell signaling. fat stress and liver resistance. Science **322**(5907), 1483–1484 (2008). DOI 10.1126/science.1167571. URL http://dx.doi.org/10.1126/science.1167571

[222] Okoniewski, M., Hey, Y., Pepper, S., Miller, C.: High correspondence between Affymetrix exon and standard expression arrays. BIOTECHNIQUES **42**(2), 181 (2007)

[223] Okoniewski, M., Miller, C.: Comprehensive Analysis of Affymetrix Exon Arrays Using BioConductor. PLoS Computational Biology **4**(2), 1–6 (2008)

[224] Oliver, S.: On the miame standards and central repositories of microarray data. Comp Funct Genomics **4**(1), 1 (2003). DOI 10.1002/cfg.238. URL http://dx.doi.org/10.1002/cfg.238

[225] OMIM: Online mendelian inheritance in man, omim (tm) (2000). MIM Number: 125853: 04.10.2005

[226] Ortlepp, J.R., Kluge, R., Giesen, K., Plum, L., Radke, P., Hanrath, P., Joost, H.G.: A metabolic syndrome of hypertension, hyperinsulinaemia and hypercholesterolaemia in the new zealand obese mouse. Eur J Clin Invest **30**(3), 195–202 (2000)

[227] Pagani, F., Baralle, F., Genome, H., RNA, S., RNA, M.: Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet **5**(5), 389–96 (2004)

[228] Pan, Q., Shai, O., Lee, L.J., Frey, B.J., Blencowe, B.J.: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet **40**(12), 1413–1415 (2008). DOI 10.1038/ng.259. URL http://dx.doi.org/10.1038/ng.259

[229] Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A.L., Mohammad, N., Babak, T., Siu, H., Hughes, T.R., Morris, Q.D., Frey, B.J., Blencowe, B.J.: Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol Cell **16**(6), 929 – 941 (2004). DOI 10.1016/j.molcel.2004.12.004. URL http://www.ncbi.nlm.nih.gov/pubmed/15610736

[230] Parikh, H., Groop, L.: Candidate genes for type 2 diabetes. Rev Endocr Metab Disord **5**(2), 151–76 (2004). 1389-9155 (Print) Journal Article Review

[231] Permutt, M.A., Wasson, J., Cox, N.: Genetic epidemiology of diabetes. J Clin Invest **115**(6), 1431 – 1439 (2005). DOI 10.1172/JCI24758. URL http://www.ncbi.nlm.nih.gov/pubmed/15931378

[232] Ploner, A.: Heatplus: A heat map displaying covariates and coloring clusters (2009). R package version 1.12.0

[233] Plum, L., Kluge, R., Giesen, K., Altmüller, J., Ortlepp, J.R., Joost, H.G.: Type 2 diabetes-like hyperglycemia in a backcross model of nzo and sjl mice: characterization of a susceptibility locus on chromosome 4 and its relation with obesity. Diabetes **49**(9), 1590–1596 (2000)

[234] Po, S.S., Wu, R.C., Juang, G.J., Kong, W., Tomaselli, G.F.: Mechanism of alpha -adrenergic regulation of expressed hKv4.3 currents. Am J Physiol Heart Circ Physiol **281**(6), H2518–2527 (2001). URL http://ajpheart.physiology.org/cgi/content/abstract/281/6/H2518

[235] Pollastro, P., Rampone, S.: Hs3d - homo sapiens splice sites dataset. Nucleic Acids Research, 2003 Annual Database Issue. (2003). URL http://www.sci.unisannio.it/docenti/rampone/. Release 1.2

[236] Pospisil, H., Herrmann, A., Bortfeldt, R.H., Reich, J.G.: EASED: Extended Alternatively Spliced EST Database. Nucleic Acids Res **32**(Database issue), D70 – 4 (2004). DOI 10.1093/nar/gkh136. URL http://www.ncbi.nlm.nih.gov/pubmed/14681361

[237] Purdom, E., Simpson, K.M., Robinson, M.D., Conboy, J.G., Lapuk, A.V., Speed, T.P.: FIRMA: a method for detection of alternative splicing from exon array data. Bioinformatics **24**(15), 1707 – 1714 (2008). DOI 10.1093/bioinformatics/btn284. URL http://www.ncbi.nlm.nih.gov/pubmed/18573797

[238] R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2005)

[239] Rasche, A.: Handbook of Research on Systems Biology Applications in Medicine,

chap. Approaching Type 2 Diabetes Mellitus by Systems Biology, pp. 358–373. IGI Global (2009)

[240] Rasche, A., Al-Hasani, H., Herwig, R.: Meta-Analysis Approach identifies Candidate Genes and associated Molecular Networks for Type-2 Diabetes Mellitus. BMC Genomics **9**, 310–326 (2008). (Highly Accessed, Open Access)

[241] Rasche, A., Herwig, R.: ARH: Predicting Splice Variants from Genome-wide Data with Modified Entropy. Bioinformatics **26**, 84–90 (2009). (Open Access)

[242] Rattray, M., Liu, X., Sanguinetti, G., Milo, M., Lawrence, N.D.: Propagating uncertainty in microarray data analysis. Brief Bioinform **7**(1), 37–47 (2006)

[243] Relógio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R.B., Valcárcel, J.: Alternative splicing microarrays reveal functional expression of neuron-specific regulators in hodgkin lymphoma cells. J Biol Chem **280**(6), 4779–4784 (2005). DOI 10.1074/jbc.M411976200. URL `http://dx.doi.org/10.1074/jbc.M411976200`

[244] Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., Chinnaiyan, A.M.: Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A **101**(25), 9309–14 (2004). 0027-8424 (Print) Journal Article Meta-Analysis Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[245] Richard, H., Schulz, M., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., Haas, S.A., Yaspo, M.L.: Prediction of alternative isoforms from exon expression levels in rna-seq experiments. — —, — (—). (under review)

[246] Richard, H., Schulz, M.H., Sultan, M., Magen, A., Haas, S., Vingron, M., Yaspo, M.L.: Genome wide analysis of alternative splicing using second generation sequencing. In: Alternative Splicing Special Interest Group, ISMB 2008, Toronto (2008)

[247] Ritchie, W., Granjeaud, S., Puthier, D., Gautheret, D.: Entropy measures quantify global splicing disorders in cancer. PLoS Computational Biology **4**(3), e1000,011 (2008). DOI 10.1371/journal.pcbi.1000011

[248] Robertson, M.: Reactome: clear view of a starry sky. Drug Discov Today **9**(16), 684–685 (2004). DOI 10.1016/S1359-6446(04)03217-9. URL `http://dx.doi.org/10.1016/S1359-6446(04)03217-9`

[249] Romero, P., Wagg, J., Green, M., Kaiser, D., Krummenacker, M., Karp, P.: Computational prediction of human metabolic pathways from the complete human genome. Genome Biology **6**(1:R2), 17 (2004)

[250] Russ, A.P., Lampel, S.: The druggable genome: an update. Drug Discov Today **10**(23-24), 1607–10 (2005). 1359-6446 (Print) Journal Article

[251] Sabio, G., Das, M., Mora, A., Zhang, Z., Jun, J.Y., Ko, H.J., Barrett, T., Kim, J.K., Davis, R.J.: A stress signaling pathway in adipose tissue regulates hepatic insulin

resistance. Science **322**(5907), 1539–1543 (2008). DOI 10.1126/science.1160794. URL http://dx.doi.org/10.1126/science.1160794

[252] Sachs, L., Hedderich, J.: Angewandte Statistik. 12th edn. Springer (2006). ISBN 3-540-32160-8

[253] Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J., Hughes, T.E., Groop, L., Altshuler, D., Almgren, P., Florez, J.C., Meyer, J., Ardlie, K., Bengtsson, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H.N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Rastam, L., Speliotes, E.K., Taskinen, M.R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjogren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S.B., Chirn, G.W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., Purcell, S.: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science **316**(5829), 1331–1336 (2007). DOI 10.1126/science.1142358. URL http://dx.doi.org/10.1126/science.1142358. 1095-9203 (Electronic) Journal article

[254] Schlitt, T., Brazma, A.: Modelling gene networks at different organisational levels. FEBS Lett **579**(8), 1859–1866 (2005). DOI 10.1016/j.febslet.2005.01.073. URL http://dx.doi.org/10.1016/j.febslet.2005.01.073

[255] Schuster, E.F., Blanc, E., Partridge, L., Thornton, J.M.: Estimation and correction of non-specific binding in a large-scale spike-in experiment. Genome Biol **8**(6), R126 (2007). DOI 10.1186/gb-2007-8-6-r126. URL http://www.ncbi.nlm.nih.gov/pubmed/17594493

[256] Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., Prokunina-Olsson, L., Ding, C.J., Swift, A.J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.Y., Conneely, K.N., Riebow, N.L., Sprau, A.G., Tong, M., White, P.P., Hetrick, K.N., Barnhart, M.W., Bark, C.W., Goldstein, J.L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T.A., Watanabe, R.M., Valle, T.T., Kinnunen, L., Abecasis, G.R., Pugh, E.W., Doheny, K.F., Bergman, R.N., Tuomilehto, J., Collins, F.S., Boehnke, M.: A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science **316**(5829), 1341–1345 (2007). DOI 10.1126/science.1142382. URL http://dx.doi.org/10.1126/science.1142382

[257] Sell, S.M., Reese, D.: Insulin-inducible changes in the relative ratio of ptp1b splice variants. Mol Genet Metab **66**(3), 189–92 (1999). ISRSA0003/RS/United States DRS Journal Article Research Support, U.S. Gov't, P.H.S. United states

[258] Sesti, G., Federici, M., Lauro, D., Sbraccia, P., Lauro, R.: Molecular mechanism of insulin resistance in type 2 diabetes mellitus: role of the insulin receptor variant

forms. Diabetes Metab Res Rev **17**(5), 363–73 (2001). Journal Article Research Support, Non-U.S. Gov't Review England

[259] Shah, S.H., Pallas, J.A.: Identifying differential exon splicing using linear models and correlation coefficients. BMC Bioinformatics **10**, 26 (2009). DOI 10.1186/1471-2105-10-26. URL http://dx.doi.org/10.1186/1471-2105-10-26

[260] Shai, O., Morris, Q.D., Blencowe, B.J., Frey, B.J.: Inferring global levels of alternative splicing isoforms using a generative model of microarray data. Bioinformatics **22**(5), 606–613 (2006). DOI 10.1093/bioinformatics/btk028. URL http://dx.doi.org/10.1093/bioinformatics/btk028

[261] Shannon, C.E.: A mathematical theory of communication. Bell Systems Technology Journal **27**, 379–423 (1948)

[262] Shi, L., Reid, L.H., Jones, W.D., Shippy, R., Warrington, J.A., Baker, S.C., Collins, P.J., de Longueville, F., Kawasaki, E.S., Lee, K.Y., Luo, Y., Sun, Y.A., Willey, J.C., Setterquist, R.A., Fischer, G.M., Tong, W., Dragan, Y.P., Dix, D.J., Frueh, F.W., Goodsaid, F.M., Herman, D., Jensen, R.V., Johnson, C.D., Lobenhofer, E.K., Puri, R.K., Schrf, U., Thierry-Mieg, J., Wang, C., Wilson, M., Wolber, P.K., Zhang, L., Amur, S., Bao, W., Barbacioru, C.C., Lucas, A.B., Bertholet, V., Boysen, C., Bromley, B., Brown, D., Brunner, A., Canales, R., Cao, X.M., Cebula, T.A., Chen, J.J., Cheng, J., Chu, T.M., Chudin, E., Corson, J., Corton, J.C., Croner, L.J., Davies, C., Davison, T.S., Delenstarr, G., Deng, X., Dorris, D., Eklund, A.C., h. Fan, X., Fang, H., Fulmer-Smentek, S., Fuscoe, J.C., Gallagher, K., Ge, W., Guo, L., Guo, X., Hager, J., Haje, P.K., Han, J., Han, T., Harbottle, H.C., Harris, S.C., Hatchwell, E., Hauser, C.A., Hester, S., Hong, H., Hurban, P., Jackson, S.A., Ji, H., Knight, C.R., Kuo, W.P., LeClerc, J.E., Levy, S., Li, Q.Z., Liu, C., Liu, Y., Lombardi, M.J., Ma, Y., Magnuson, S.R., Maqsodi, B., McDaniel, T., Mei, N., Myklebost, O., Ning, B., Novoradovskaya, N., Orr, M.S., Osborn, T.W., Papallo, A., Patterson, T.A., Perkins, R.G., Peters, E.H., Peterson, R., Philips, K.L., Pine, P.S., Pusztai, L., Qian, F., Ren, H., Rosen, M., Rosenzweig, B.A., Samaha, R.R., Schena, M., Schroth, G.P., Shchegrova, S., Smith, D.D., Staedtler, F., Su, Z., Sun, H., Szallasi, Z., Tezak, Z., Thierry-Mieg, D., Thompson, K.L., Tikhonova, I., Turpaz, Y., Vallanat, B., Christophe, V., Walker, S.J., Wang, S.J., Wang, Y., Wolfinger, R., Wong, A., Wu, J., Xiao, C., Xie, Q., Xu, J., Yang, W., Zhang, L., Zhong, S., Zong, Y., Jr, W.S.: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. **24**(9), 1151 – 1161 (2006). DOI 10.1038/nbt1239. URL http://www.ncbi.nlm.nih.gov/pubmed/16964229

[263] Siddiqui, A.S., Delaney, A.D., Schnerch, A., Griffith, O.L., Jones, S.J.M., Marra, M.A.: Sequence biases in large scale gene expression profiling data. Nucleic Acids Res **34**(12), e83 (2006). DOI 10.1093/nar/gkl404. URL http://www.ncbi.nlm.nih.gov/pubmed/16840527

[264] Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T.: Rocr: visualizing classifier performance in r. Bioinformatics **21**(20), 3940–3941 (2005). DOI 10.1093/

bioinformatics/bti623. URL http://dx.doi.org/10.1093/bioinformatics/bti623

[265] Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T.: ROCR: Visualizing the performance of scoring classifiers. (2007). URL http://rocr.bioinf.mpi-sb.mpg.de/. R package version 1.0-2

[266] Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C., Froguel, P.: A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature **445**(7130), 881 – 885 (2007). DOI 10.1038/nature05616. URL http://www.ncbi.nlm.nih.gov/pubmed/17293876

[267] Smink, L.J., Helton, E.M., Healy, B.C., Cavnor, C.C., Lam, A.C., Flamez, D., Burren, O.S., Wang, Y., Dolman, G.E., Burdick, D.B., Everett, V.H., Glusman, G., Laneri, D., Rowen, L., Schuilenburg, H., Walker, N.M., Mychaleckyj, J., Wicker, L.S., Eizirik, D.L., Todd, J.A., Goodman, N.: T1dbase, a community web-based resource for type 1 diabetes research. Nucleic Acids Res **33**(Database issue), D544–9 (2005). 1362-4962 (Electronic) Journal Article Research Support, Non-U.S. Gov't

[268] Smyth, G.K.: Limma: linear models for microarray data. In: Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., Huber, W. (eds.) Bioinformatics and Computational Biology Solutions using R and Bioconductor, pp. 397–420. Springer, New York (2005)

[269] Spinas, G.A., Lehmann, R.: Der diabetes mellitus: Diagnose, klassifikation und pathogenese. Schweizerisches Medizin-Forum **20**, 519–525 (2001)

[270] Srinivasan, K., Shiue, L., Hayes, J.D., Centers, R., Fitzwater, S., Loewen, R., Edmondson, L.R., Bryant, J., Smith, M., Rommelfanger, C., Welch, V., Clark, T.A., Sugnet, C.W., Howe, K.J., Mandel-Gutfreund, Y., Jr, M.A.: Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods **37**(4), 345 – 359 (2005). DOI 10.1016/j.ymeth.2005.09.007. URL http://www.ncbi.nlm.nih.gov/pubmed/16314264

[271] Stamm, S.: Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. Human Molecular Genetics **11**(20), 2409–2416 (2002)

[272] Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T., Soreq, H.: Function of alternative splicing. Gene **344**, 1–20 (2005)

[273] Stamm, S., Riethoven, J.J., Le, T.V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N.L., Thanaraj, T.A.: ASD: a bioinformatics resource on alternative splicing. Nucleic Acids Res **34**(Database issue), D46 – 55 (2006). DOI 10.1093/nar/gkj031. URL http://www.ncbi.nlm.nih.gov/pubmed/16381912

[274] Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O., Zhang, M.Q.: An alternative-exon database and its statistical analysis. DNA Cell Biol **19**(12), 739–756

(2000). DOI 10.1089/104454900750058107. URL http://dx.doi.org/10.1089/104454900750058107

[275] Steinhoff, C., Vingron, M.: Normalization and quantification of differential expression in gene expression microarrays. Brief Bioinform **7**(2), 166–177 (2006). DOI 10.1093/bib/bbl002. URL http://dx.doi.org/10.1093/bib/bbl002

[276] Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S., Baker, A., Snorradottir, S., Bjarnason, H., Ng, M.C.Y., Hansen, T., Bagger, Y., Wilensky, R.L., Reilly, M.P., Adeyemo, A., Chen, Y., Zhou, J., Gudnason, V., Chen, G., Huang, H., Lashley, K., Doumatey, A., So, W.Y., Ma, R.C.Y., Andersen, G., Borch-Johnsen, K., Jorgensen, T., van Vliet-Ostaptchouk, J.V., Hofker, M.H., Wijmenga, C., Christiansen, C., Rader, D.J., Rotimi, C., Gurney, M., Chan, J.C.N., Pedersen, O., Sigurdsson, G., Gulcher, J.R., Thorsteinsdottir, U., Kong, A., Stefansson, K.: A variant in cdkal1 influences insulin response and risk of type 2 diabetes. Nat Genet **39**(6), 770–775 (2007). DOI 10.1038/ng2043. URL http://dx.doi.org/10.1038/ng2043

[277] Stephenson, M., Zamecnik, P.: Inhibition of Rous sarcoma viral RNA translation by a specific oligodeoxyribonucleotide. Proceedings of the National Academy of Sciences of the United States of America **75**(1), 285 (1978)

[278] Steppan, C.M., Bailey, S.T., Bhat, S., Brown, E.J., Banerjee, R.R., Wright, C.M., Patel, H.R., Ahima, R.S., Lazar, M.A.: The hormone resistin links obesity to diabetes. Nature **409**(6818), 307–12 (2001). 0028-0836 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[279] Stier, H., Wrede, P., Kleffe, J.: Handbook of Research on Systems Biology applications in Medicine, chap. Alternative Splicing and Disease, pp. 291 – 311. IGI Global (2009)

[280] Stoilov, P., Meshorer, E., Gencheva, M., Glick, D., Soreq, H., Stamm, S.: Defects in Pre-mRNA Processing as Causes of and Predisposition to Diseases. DNA and Cell Biology **21**(11), 803–818 (2002)

[281] Storey, J.: A direct approach to false discovery rates. Journal Of The Royal Statistical Society Series B **64**(3), 479–498 (2002)

[282] Storey, J., Taylor, J., Siegmund, D.: Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society Series B **66**(1), 187–205 (2004)

[283] Stumvoll, M., Goldstein, B.J., van Haeften, T.W.: Type 2 diabetes: principles of pathogenesis and therapy. Lancet **365**(9467), 1333–1346 (2005). DOI 10.1016/S0140-6736(05)61032-X. URL http://dx.doi.org/10.1016/S0140-6736(05)61032-X

[284] Stumvoll, M., Goldstein, B.J., van Haeften, T.W.: Pathogenesis of type 2 diabetes. Endocr Res **32**(1-2), 19–37 (2007)

[285] Stumvoll, M., Goldstein, B.J., van Haeften, T.W.: Type 2 diabetes: pathogenesis

and treatment. Lancet **371**(9631), 2153–2156 (2008). DOI 10.1016/S0140-6736(08)60932-0. URL http://dx.doi.org/10.1016/S0140-6736(08)60932-0

[286] Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., Patapoutian, A., Hampton, G.M., Schultz, P.G., Hogenesch, J.B.: Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A **99**(7), 4465–4470 (2002). DOI 10.1073/pnas.012025199. URL http://dx.doi.org/10.1073/pnas.012025199

[287] Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M.P., Walker, J.R., Hogenesch, J.B.: A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A **101**(16), 6062–6067 (2004). DOI 10.1073/pnas.0400782101. URL http://dx.doi.org/10.1073/pnas.0400782101

[288] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A **102**(43), 15,545–15,550 (2005). DOI 10.1073/pnas.0506580102. URL http://dx.doi.org/10.1073/pnas.0506580102

[289] Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., Yaspo, M.L.: A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science **321**(5891), 956–960 (2008). DOI 10.1126/science.1160342. URL http://dx.doi.org/10.1126/science.1160342

[290] Sun, G.: Application of DNA microarrays in the study of human obesity and type 2 diabetes. OMICS **11**(1), 25 – 40 (2007). DOI 10.1089/omi.2006.0003. URL http://www.ncbi.nlm.nih.gov/pubmed/17411394

[291] Sundsten, T., Eberhardson, M., Goransson, M., Bergsten, P.: The use of proteomics in identifying differentially expressed serum proteins in humans with type 2 diabetes. Proteome Sci **4**, 22 (2006). DOI 10.1186/1477-5956-4-22. URL http://www.ncbi.nlm.nih.gov/pubmed/17163994

[292] Takeda, J.I., Suzuki, Y., Nakao, M., Barrero, R.A., Koyanagi, K.O., Jin, L., Motono, C., Hata, H., Isogai, T., Nagai, K., Otsuki, T., Kuryshev, V., Shionyu, M., Yura, K., Go, M., Thierry-Mieg, J., Thierry-Mieg, D., Wiemann, S., Nomura, N., Sugano, S., Gojobori, T., Imanishi, T.: Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. Nucleic Acids Res **34**(14), 3917 – 3928 (2006). DOI 10.1093/nar/gkl507. URL http://www.ncbi.nlm.nih.gov/pubmed/16914452

[293] Takeda, J.I., Suzuki, Y., Nakao, M., Kuroda, T., Sugano, S., Gojobori, T., Imanishi, T.: H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. Nucleic

Acids Res **35**(Database issue), D104 – 9 (2007). DOI 10.1093/nar/gkl854. URL http://www.ncbi.nlm.nih.gov/pubmed/17130147

[294] Talloen, W., Clevert, D.A., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S., Gohlmann, H.W.H.: I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. Bioinformatics **23**(21), 2897 – 2902 (2007). DOI 10.1093/bioinformatics/btm478. URL http://www.ncbi.nlm.nih.gov/pubmed/17921172

[295] Thanaraj, T., Stamm, S.: Regulation of Alternative Splicing, vol. 31, chap. Prediction and statistical analysis of alternatively spliced exons, pp. 1–31. Springer (2003)

[296] Thanaraj, T.A., Stamm, S., Clark, F., Riethoven, J.J., Le, T.V., Muilu, J.: ASD: the Alternative Splicing Database. Nucleic Acids Res **32**(Database issue), D64 – 9 (2004). DOI 10.1093/nar/gkh030. URL http://www.ncbi.nlm.nih.gov/pubmed/14681360

[297] Thormann, A.: Topologische Analyse von Protein-Protein-Interaktionsnetzwerken. bachelor thesis, Freie Universität Berlin (2007)

[298] Tiffin, N., Adie, E., Turner, F., Brunner, H.G., van Driel, M.A., Oti, M., Lopez-Bigas, N., Ouzounis, C., Perez-Iratxeta, C., Andrade-Navarro, M.A., Adeyemo, A., Patti, M.E., Semple, C.A.M., Hide, W.: Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. Nucleic Acids Res **34**(10), 3067 – 3081 (2006). DOI 10.1093/nar/gkl381. URL http://www.ncbi.nlm.nih.gov/pubmed/16757574

[299] Toye, A., Gauguier, D.: Genetics and functional genomics of type 2 diabetes mellitus. Genome Biol **4**(12), 241 (2003). Journal Article Research Support, Non-U.S. Gov't Review England

[300] Toye, A.A., Lippiat, J.D., Proks, P., Shimomura, K., Bentley, L., Hugill, A., Mijat, V., Goldsworthy, M., Moir, L., Haynes, A., Quarterman, J., Freeman, H.C., Ashcroft, F.M., Cox, R.D.: A genetic and physiological study of impaired glucose homeostasis control in c57bl/6j mice. Diabetologia **48**(4), 675–86 (2005). 0012-186X (Print) Comparative Study Journal Article Research Support, Non-U.S. Gov't

[301] Tress, M.L., Bodenmiller, B., Aebersold, R., Valencia, A.: Proteomics studies confirm the presence of alternative protein isoforms on a large scale. Genome Biol **9**(11), R162 (2008). DOI 10.1186/gb-2008-9-11-r162. URL http://dx.doi.org/10.1186/gb-2008-9-11-r162

[302] Tukey, J.: Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts (1977)

[303] Veroni, M.C., Proietto, J., Larkins, R.G.: Evolution of insulin resistance in new zealand obese mice. Diabetes **40**(11), 1480–1487 (1991)

[304] Vingron, M.: Bioinformatics needs to adopt statistical thinking. Bioinformatics **17**(5), 389–390 (2001)

[305] Virtue, S., Vidal-Puig, A.: It's not how fat you are, it's what you do with it that

counts. PLoS Biol **6**(9), e237 (2008). DOI 10.1371/journal.pbio.0060237. URL http://dx.doi.org/10.1371/journal.pbio.0060237

[306] Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B.: Alternative isoform regulation in human tissue transcriptomes. Nature **456**(7221), 470–476 (2008). DOI 10.1038/nature07509. URL http://dx.doi.org/10.1038/nature07509

[307] Wang, G.S., Cooper, T.A.: Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet **8**(10), 749 – 761 (2007). DOI 10.1038/nrg2164. URL http://www.ncbi.nlm.nih.gov/pubmed/17726481

[308] Wang, H., Hubbell, E., shan Hu, J., Mei, G., Cline, M., Lu, G., Clark, T., Siani-Rose, M.A., Ares, M., Kulp, D.C., Haussler, D.: Gene structure-based splice variant deconvolution using a microarray platform. Bioinformatics **19 Suppl 1**, i315–i322 (2003)

[309] Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet **10**(1), 57–63 (2009). DOI 10.1038/nrg2484. URL http://dx.doi.org/10.1038/nrg2484

[310] Wu, J., with contributions from James MacDonald Jeff Gentry, R.I.: gcrma: Background Adjustment Using Sequence Information (2009). R package version 2.14.0

[311] Wu, Z., Irizarry, R.: Stochastic models inspired by hybridization theory for short oligonucleotide arrays. In: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology, pp. 98–106. ACM New York, NY, USA (2004)

[312] Wu, Z., Irizarry, R.: Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. Journal of Computational Biology **12**(6), 882–893 (2005)

[313] Wu, Z., Irizarry, R., Gentleman, R., Martinez-Murillo, F., Spencer, F.: A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association **99**(468), 909–918 (2004)

[314] Xie, C., Bondarenko, V., Morales, M., Strauss, H.: Closed-State Inactivation in Kv4. 3 Splice Forms is Differentially Modulated by Protein Kinase C. Biophysical Journal **96**(3S1), 655–655 (2009)

[315] Xing, Y., Kapur, K., Wong, W.H.: Probe selection and expression index computation of Affymetrix Exon Arrays. PLoS ONE **1**, e88 (2006). DOI 10.1371/journal.pone.0000088. URL http://www.ncbi.nlm.nih.gov/pubmed/17183719

[316] Xing, Y., Stoilov, P., Kapur, K., Han, A., Jiang, H., Shen, S., Black, D.L., Wong, W.H.: Mads: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays. RNA **14**(8), 1470–1479 (2008). DOI 10.1261/rna.1070208. URL http://dx.doi.org/10.1261/rna.1070208

[317] Yeakley, J.M., Fan, J.B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M.S., Fu, X.D.: Profiling alternative splicing on fiber-optic arrays. Nat Biotechnol **20**(4), 353–358 (2002). DOI 10.1038/nbt0402-353. URL http://dx.doi.org/10.1038/nbt0402-353

[318] Yeo, G.: Splicing regulators: targets and drugs. Genome Biology **6**(12), 240 (2005)

[319] Yeo, G., Holste, D., Kreiman, G., Burge, C.B.: Variation in alternative splicing across human tissues. Genome Biol **5**(10), R74 (2004). DOI 10.1186/gb-2004-5-10-r74. URL http://dx.doi.org/10.1186/gb-2004-5-10-r74

[320] Yoshida, R., Numata, K., Imoto, S., Nagasaki, M., Doi, A., Ueno, K., Miyano, S.: A Statistical Framework for Genome-Wide Discovery of Biomarker Splice Variations with GeneChip Human Exon 1.0 STArrays. In: International Conference on Genome Informatics, vol. 17, p. 88. UNIVERSAL ACADEMY PRESS INC. (2006)

[321] Youn, B.S., Yu, K.Y., Park, H.J., Lee, N.S., Min, S.S., Youn, M.Y., Cho, Y.M., Park, Y.J., Kim, S.Y., Lee, H.K., Park, K.S.: Plasma resistin concentrations measured by enzyme-linked immunosorbent assay using a newly developed monoclonal antibody are elevated in individuals with type 2 diabetes mellitus. J Clin Endocrinol Metab **89**(1), 150–6 (2004). 0021-972X (Print) Journal Article Research Support, Non-U.S. Gov't

[322] Zamecnik, P., Stephenson, M.: Inhibition of Rous sarcoma virus replication and cell transformation by a specific oligodeoxynucleotide. Proceedings of the National Academy of Sciences of the United States of America **75**(1), 280–284 (1978)

[323] Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R.B., Rayner, N.W., Freathy, R.M., Barrett, J.C., Shields, B., Morris, A.P., Ellard, S., Groves, C.J., Harries, L.W., Marchini, J.L., Owen, K.R., Knight, B., Cardon, L.R., Walker, M., Hitman, G.A., Morris, A.D., Doney, A.S.F., (WTCCC), W.T.C.C.C., McCarthy, M.I., Hattersley, A.T.: Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. Science **316**(5829), 1336–1341 (2007). DOI 10.1126/science.1142364. URL http://dx.doi.org/10.1126/science.1142364

[324] Zhang, C., Li, H.R., Fan, J.B., Wang-Rodriguez, J., Downs, T., Fu, X.D., Zhang, M.Q.: Profiling alternatively spliced mrna isoforms for prostate cancer classification. BMC Bioinformatics **7**, 202 (2006). DOI 10.1186/1471-2105-7-202. URL http://dx.doi.org/10.1186/1471-2105-7-202

[325] Zhang, L., Miles, M.F., Aldape, K.D.: A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol **21**(7), 818–821 (2003). DOI 10.1038/nbt836. URL http://dx.doi.org/10.1038/nbt836

[326] Zheng, C.L., Kwon, Y.S., Li, H.R., Zhang, K., Coutinho-Mansfield, G., Yang, C., Nair, T.M., Gribskov, M., Fu, X.D.: MAASE: an alternative splicing database designed for supporting splicing microarray applications. RNA **11**(12), 1767 – 1776 (2005). DOI 10.1261/rna.2650905. URL http://www.ncbi.nlm.nih.gov/pubmed/16251387

Notation and Abbreviations

AS	alternative splicing
c	biological condition index: $c = c$ for control, t for treatment
DE	differential expression
ESGEC	early stage gene expression changes, project data set
e	exon index
η	raw probe hybridisation value
GGSC	glycaemic and genetic splicing changes, project data set
g	gene index
ι	normalised probe intensity
MM	mismatch probe
NZL	NZO derived in-bred polygenic mouse model for T2DM
NZO	*New Zealand obese*, in-bred polygenic mouse model for T2DM
p	probe index
ϕ	exon expression
Φ	gene expression
PM	perfect match probe
r	replicate index
SJL	*Swiss Jackson laboratory*, in-bred lean non-diabetic mouse model
SNP	single nucleotide polymorphism
T2DM	type-2 diabetes mellitus
W.l.o.g.	Without loss of generality

In all cases not introduced in the thesis statistical notation follows Sachs and Hedderich [252] and information theoretical notation follows Cover and Thomas [71].

Typesetting and Layout This document was created using the LaTeX-Suite for layout and typesetting as provided by MiKTeX 2.7 and operated through LyX 1.6.3.

Publications

[1] Clevert, D.A., Rasche, A.: Handbook of Research on Systems Biology Applications in Medicine, chap. The Affymetrix GeneChip Microarray Platform, pp. 248–258. IGI Global (2009)

[2] Daskalaki, A., Rasche, A.: Informatics in Oral Medicine: Advanced Techniques in Clinical and Diagnostic Technologies, chap. Meta-Analysis approach for the identification of molecular networks related to infections of the oral cavity. IGI Global (—). (in press)

[3] Dreja, T., Jovanovic, Z., Rasche, A., Kluge, R., Herwig, R., Tung, Y.C.L., Joost, H.G., Yeo, G.S., Al-Hasani, H.: Diet-induced gene expression of isolated pancreatic islets from a polygenic mouse model for the metabolic syndrome. Diabetologia **53**(2), 309–320 (2009). (Open Access)

[4] Rasche, A.: Handbook of Research on Systems Biology Applications in Medicine, chap. Approaching Type 2 Diabetes Mellitus by Systems Biology, pp. 358–373. IGI Global (2009)

[5] Rasche, A., Al-Hasani, H., Herwig, R.: Meta-Analysis Approach identifies Candidate Genes and associated Molecular Networks for Type-2 Diabetes Mellitus. BMC Genomics **9**, 310–326 (2008). (Highly Accessed, Open Access)

[6] Rasche, A., Herwig, R.: ARH: Predicting Splice Variants from Genome-wide Data with Modified Entropy. Bioinformatics **26**, 84–90 (2009). (Open Access)

[7] Rasche, A., Yildirimman, R., Herwig, R.: Handbook of Systems Toxicology, chap. Integrative Analysis of microarray data – a path for systems toxicology. John Wiley & Sons Ltd. (—). (accepted)

[8] Richard, H., Schulz, M., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., Haas, S.A., Yaspo, M.L.: Prediction of alternative isoforms from exon expression levels in rna-seq experiments. — —, — (—). (under review)

Acknowledgements

I am very indebted to Dr. Ralf Herwig. Not only he offered me a motivating topic, but competently lead me into computational biology, a dynamic field of science at the interface between mathematics, biology and computer science. An excellent working environment is established by Prof. Dr. Hans Lehrach and by Prof. Dr. Martin Vingron at the Max Planck Institute for Molecular Genetics. It was an intellectual pleasure to discuss and learn from them.

This work would not have been possible without the support from various sides. Dr. Hadi Al-Hasani and Dr. Tanja Dreja from the German Institute of Human Nutrition (DIfE) performed the experiments on Affymetrix arrays, in particular the ESGEC and GGSC data sets, and introduced me to type-2 diabetes mellitus. Together with Dr. Arif Malik and Dr. Frank Kleinjung it was an enjoyable scientific collaboration. Dr. Stefan Haas provided insight and experience about alternative splicing and splicing prediction. Anja Thormann continues this work by implementation of the T2DM-GeneMiner web tool releasing the results of the meta-analysis study to the community and composing a huge topological model from the marker list. The following persons read parts of the manuscript: Dr. Hadi Al-Hasani, Simon Bungers, Lukas Chavez, Dr. Anita Daskalaki, Dr. Tanja Dreja, Dr. Stefan Haas, Hendrik Hache, Linda Hallen, Dr. Ralf Herwig, Dr. Martje Tönjes and Dr. Christoph Wierling.

Assistance on the way of finishing was provided by my office partner Lukas Chavez. Beside science the student association STA at the institute offered exchange and distraction, organised by Martje Tönjes, Linda Hallen and many others.

Financial support was appreciated by the Max Planck Society, the BMBF, the European Union and by my family. The ESGEC data set was generated in the Nutrigenomik project PhysioSim. The GGSC data set was supported by the European Union under its 6th Framework Programme with the grant SysProt [LSHG-CT-2006- 037457].

Parts of Chapter 3, Computational Analysis of Affymetrix Arrays, and Chapter 5, Alternative Splicing in Type-2 Diabetes Mellitus, appear in the Handbook of Research on Systems Biology Applications in Medicine edited by Dr. Andriani Daskalaki [65, 239]; Copyright 2009, IGI Global, www.igi-global.com. Posted by permission of the publisher.

It was great fun working on this thesis, an intensive time learning about a field I was not acquainted with before.

Zusammenfassung

Für die biomedizinische Grundlagenforschung ist es von besonderem Interesse, die Aktivität von Genen in verschiedenen Geweben eines Organismus zu bestimmen. Die Genaktivität wird hier bestimmt durch die Menge der direkten Produkte eines Gens, die Transkripte. Die Häufigkeit der Transkripte wird durch experimentelle Technologien quantifiziert und als Genexpression bezeichnet. Aber ein Gen produziert nicht immer nur ein Transkript, sondern kann mehrere Transkripte herstellen mittels der parallelen Kodierung, dem sogenannten alternativen Spleissen. Solch ein Mechanismus ist notwendig um die grosse Zahl an Proteinen und die verhältnismässig kleine Anzahl an Genen zu erklären: ∼25 000 Gene im Menschen gegenüber ∼20 000 im Fadenwurm *caenorhabditis elegans*. Alternatives Spleissen kontrolliert die Expression von verschiedenen Transkriptvarianten unter verschiedenen Bedingungen. Es ist nicht überraschend, dass auch kleine Fehler beim Spleissen pathologische Wirkung entfalten, d.h. Krankheiten auslösen können.

Da Organismen wie der des Menschen etwa 25 000 verschiedene Gene besitzen, war es notwendig, für die Analyse der globalen Genexpression Hochdurchsatzmethoden zur Datengenerierung zu entwickeln. Mit dem alternativen Spleissen stehen all diesen Genen mehrere Transkripte gegenüber. Erst seit kurzem kann die notwendige Menge an Daten generiert werden durch Technologien wie z.Bsp. Microarrays oder Sequenzierungstechnologie der neuesten Generation. Gleichzeitig mit dem technischen Fortschritt müssen die Datenanalyseverfahren mithalten, um neuen Forschungsfragen zu entsprechen.

Im Laufe dieser Arbeit wird eine Softwarepipeline vorgestellt für die Analyse von alternativem Spleissen sowie differentieller Genexpression. Sie wurde entwickelt und implementiert in der Programmiersprache und Statistik-Software R und BioConductor und umfasst die Schritte Qualitätskontrolle, Vorverarbeitung, statistische Auswertung der Expressionsveränderungen und Genmengenauswertung. Für die Erkennung von alternativem Spleissen wird die Informationstheorie in das Gebiet der Genexpression eingeführt. Die vorgestellte Lösung besteht aus einer Erweiterung der Shannon-Entropie auf die Erkennung veränderter Transkripthäufigkeiten und heisst ARH – Alternatives Spleissen Robuste Vorhersage mittels Entropie.

Der Nutzen der entwickelten Methoden und Implementierungen wird aufgezeigt am Beispiel von Daten zum Typ-2 Diabetes Mellitus. Mittels Datenintegration und Metaanalyse von unterschiedlichen Datenquellen werden Markergene bestimmt mit Fokus auf differentielle Expression. Danach wird alternatives Spleissen untersucht mit speziellem Fokus auf die Markergene und funktionelle Genmengen, d.h. Stoffwechselwegen.