Combined modeling of bulk and single-cell genome sequencing data to dissect clonal heterogeneity in acute myeloid leukemia

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von
Raphael Hablesreiter

Berlin, 2024

Erstgutachter/in:	Prof. Dr.-Ing. Knut Reinert
Zweitgutachter/in:	Prof. Dr. med. Frederik Damm
Tag der Disputation:	22. November 2024

Declaration of authorship

Name: **Hablesreiter**

First name: **Raphael**

I declare to the Freie Universität Berlin that I have completed the submitted dissertation independently and without the use of sources and aids other than those indicated. The present thesis is free of plagiarism. I have marked as such all statements that are taken literally or in content from other writings. This dissertation has not been submitted in the same or similar form in any previous doctoral procedure.

I agree to have my thesis examined by a plagiarism examination software.

Date: June 18, 2024 Signature: _____

Abstract

Intra-tumor heterogeneity describes the coexistence of multiple genetically distinct subclones within the tumor of a patient resulting from somatic evolution, clonal diversification, and selection. It is a main causal driver of therapy resistance in the clinic by already containing subclones that are resistant to therapy or by subclones acquiring resistance to therapy. Therefore, the understanding of intra-tumor heterogeneity and tumor development may lead to new approaches and targets for treatment. In this thesis, I developed a method for the integrated analysis of bulk and single-cell DNA sequencing data of core-binding factor acute myeloid leukemia patients, which is defined by the presence of a *RUNX1-RUNX1T1* or *CBFB-MYH11* fusion gene. I generated a combined bulk and single-cell dataset of 9 core-binding factor acute myeloid leukemia patients with samples at diagnosis, complete remission and relapse. Using this method, I was able to reconstruct tumor development including somatic variants, somatic copy-number alterations and fusion genes from a single tumor sample and, if available, from merged diagnosis and relapse samples showing tumor evolution under the pressure of chemotherapy. I performed an in-depth analysis of small-scale and large-scale genomic alterations of leukemia patients and, moreover, demonstrate that my developed method can detect subclonal copy-number alterations with a higher resolution as compared to current methods.

Table of Contents

List of Figures

List of Tables

List of Inserts

Abbreviations

allo-SCT	allogeneic hematopoietic stem cell transplantation
AML	acute myeloid leukemia
AR	allelic ratio
Ara-C	cytarabine
ATRA	idarubicin-cytarabine-etoposide in combination with all-trans retinoic acid
BAF	B-allele frequency
BAM	Binary Alignment Map
BIH	Berlin Institute of Health
BM	bone marrow
CAD	coronary artery disease
CBF	core-binding factor
CCF	cancer cell fraction
CH	clonal hematopoiesis
CHIP	clonal hematopoiesis of indeterminate potential
chr/CHR	chromosome
CLL	chronic lymphocytic leukemia
CML	chronic myelogenous leukemia
CNA	copy-number alteration
CR	complete remission
CSV	comma-separated values
CVD	cardiovascular diseases
DA	daunorubicin and cytarabine
DGHO	Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e.V.
DNA	deoxyribonucleic acid
DNR	daunorubicin
DP	read depth
DTA	*DNMT3A*, *TET2* and *ASXL1*
ECOG	Eastern Cooperative Oncology Group
EFS	event-free survival
ELN	European LeukemiaNet
ENA	European Nucleotide Archive
FAB	French-American-British
FLA-IDA	fludarabine, cytarabine and idarubicin

GQ	genotype quality
GSP	gene specific primers
HAM	high-dose cytosine arabinoside and mitoxantrone
HD	high-dose
HET	heterozygous
HOM	homozygous
HSC	hematopoietic stem cell
ICD	International Statistical Classification of Diseases and Related Health Problems
ICE	idarubicin-cytarabine-etoposide
IDA	idarubicin
INDEL	insertion and deletion
ITD	internal tandem duplication
ITH	intra-tumor heterogeneity
IVA	idarubicin-etoposide-(intermediate-dose) cytarabine
JSON	Java Script Object Notation
LOH	loss of heterozygosity
MAF	minor allele frequency
MCMC	Markov chain Monte Carlo
MDS	myelodysplastic syndromes
MLL	mixed lineage leukemia
MTC	mitoxantrone, topotecan and cytarabine
NGT	numerical genotype
OS	overall survival
PB	peripheral blood
PCR	polymerase chain reaction
PDF	Portable Document Format
POS	position
PTD	partial tandem duplication
REF	reference
RO	number of reads with evidence of no mutation
ROI	region of interest
RTK	receptor tyrosine kinase
SAM	Sequence Alignment Map
SCNA	somatic copy-number alteration

SNP single-nucleotide polymorphisms

SNV single-nucleotide variant

TCGA the Cancer Genome Atlas

UMI unique molecular identifier

UPD uniparental disomy

UTR untranslated region

VAF variant allele frequency

WES whole-exome sequencing

WHO the World Health Organization

WT wild-type

Note: gene and protein symbols are not included in the list of abbreviations because of the internationally standardized nomenclature

I Background

In Germany almost 500,000 people are newly diagnosed with cancer every year and alone in 2020 approximately 13,560 people (*i.e.*, 5,640 women and 7,920 men) have been diagnosed with leukemia (ICD-10: C91–C95) with 4% of them being younger than 15 years [1,2]. The German Centre for Cancer Registry Data of the Robert Koch Institute reported that the incidence rate per 100,000 people for leukemias is with 12.9 in contrast to 8.0 higher in men than in women [1]. Figure 1 shows the age-specific incidence rates for women and men of 12,723 leukemia cases reported in 2019. Approximately 23% of them are diagnosed with acute myeloid leukemia (AML) which is the second most common type of leukemia after chronic lymphocytic leukemia (CLL) with 37% of newly diagnosed cases. The data shows that the incidence rate declines for children and minors, but then increases with age.

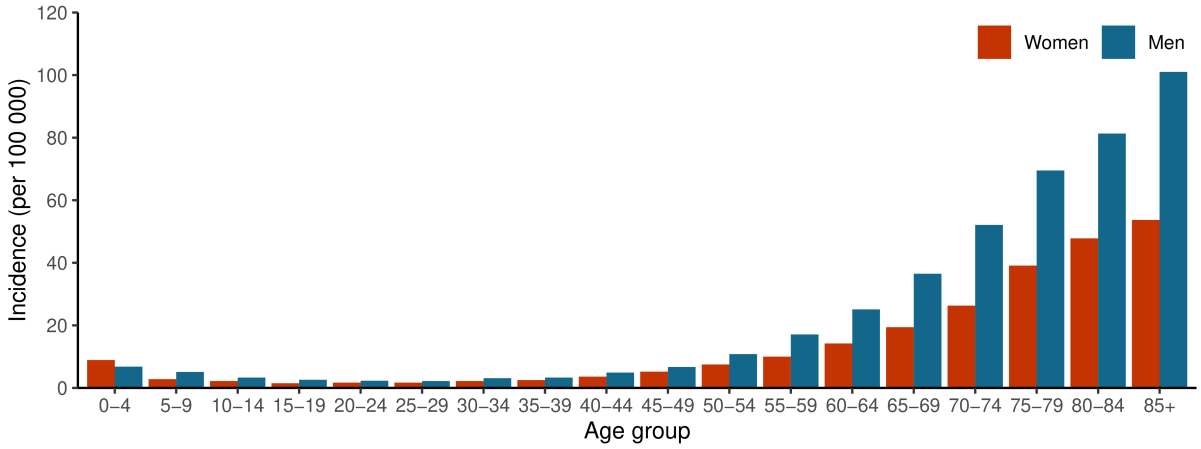

Figure 1: Age-specific incidence rates of leukemias (C91-C95) grouped by sex for Germany in 2019. Data derived from the German Centre for Cancer Registry Data of the Robert Koch Institute. [1]

1 Acute myeloid leukemia

AML is the most common acute leukemia and characterized by infiltration of the bone marrow by proliferative, clonal, abnormally differentiated and occasionally poor differentiated cells of the hematopoietic system [3]. During normal hematopoiesis mutational events (founder mutations) in hematopoietic stem cells (HSCs) primarily affecting genes involved in epigenetic regulation, such as DNA methylation (*e.g.*, *DNMT3A*, *IDH1/2* and *TET2*) or chromatin modification (*e.g.*, *ASXL1*), lead to a preleukemic state that by acquiring additional mutations (driver mutations) leads to leukemia [4,5]. This process is called leukemogenesis. AML is diagnosed if ≥20% myeloid blasts, including myeloblasts, monoblasts, and megakaryoblasts are detected in peripheral blood (PB) samples or bone marrow (BM) aspirates [6]. AML is classified according to the World Health Organization (WHO) by their 5th revised classification

of hematolymphoid tumors based on clinical parameters, phenotypic features and molecular genetic markers (*e.g.*, cytogenetics and mutation profiles) [7]. The circus plot, as shown in Figure 2, from Chen *et al.*, [8] visualizes common genetic events leading to pathogenesis of AML grouped by their functional categories. Some common genetic events, such as transcription factor fusions (*e.g.*, *MYH11-CBFB*), define distinct AML entities or subtypes (see Table 1) and are further used as guidance for risk stratification and treatment.

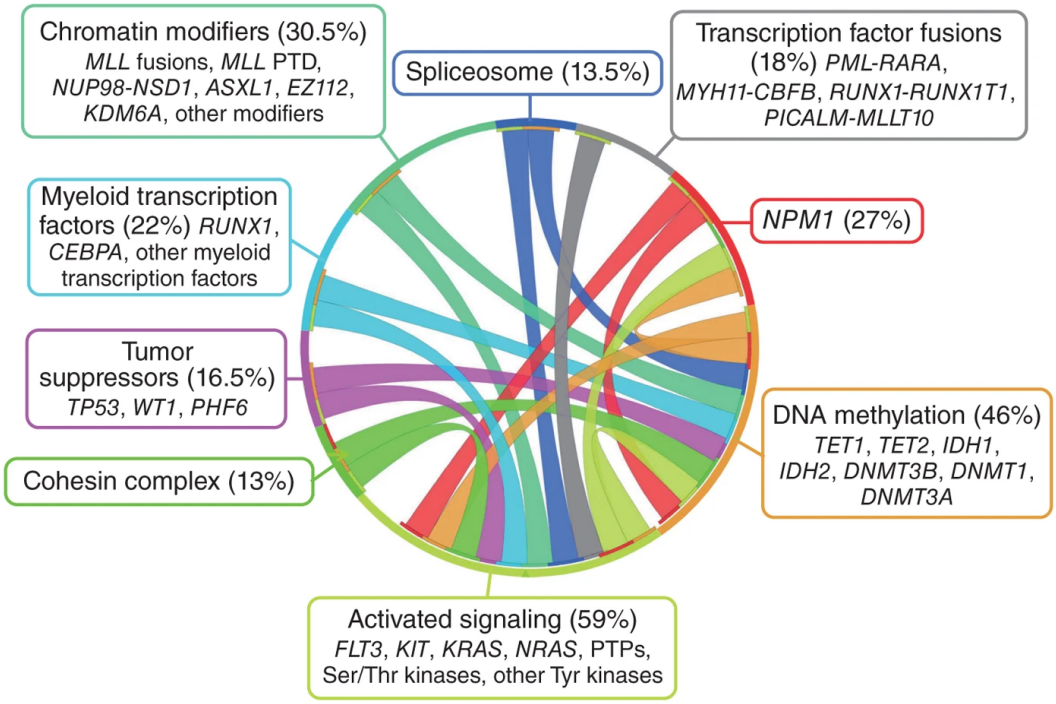

Figure 2: Most common genetic events leading to pathogenesis of acute myeloid leukemia (AML). Circos plot from Chen *et al.*, [8] visualizes by length of segment the proportion of gene alterations found in AML patients from one functional category. Bands connecting functional categories illustrate association between mutations in different pathways. Partial tandem duplication (PTD)

AML classification follows a hierarchy so that AML-defining recurrent genetic abnormalities (*e.g.*, t(8;21)(q22;q22.1)/*RUNX1-RUNX1T1*) outvote *TP53* mutations, followed by myelodysplasia-related gene mutations that supersede myelodysplasia-related cytogenetic abnormalities. Except for AML with recurrent genetic abnormalities, patients with 10-19% myeloid blasts are classified as myelodysplastic syndromes (MDS)/AML. Table 1 adapted from Döhner *et al.*, [6] and Juskevicius *et al.*, [9] lists common AML subtypes categorized by their 2022 European LeukemiaNet (ELN) risk classification (*i.e.*, favorable, intermediate and adverse) with genes that harbor frequent co-occurring mutations.

Decision-making for treating AML patients depends on patient fitness (*e.g.*, age, Eastern Cooperative Oncology Group (ECOG) performance status and pre-existing conditions) allowing them to undergo intensive chemotherapy or not and, thereafter, AML characteristics (*e.g.*, morphology and cytogenetics) for selection of induction therapy [10]. Furthermore, it is

necessary to use next-generation sequencing methods to get a comprehensive understanding of mutational status of AML specific genes (see Table 1 “Frequent co-occurring mutations”) so that, if applicable, patients can be treated using targeted therapy approaches such as tyrosine kinase inhibitors (*e.g.*, FLT3-inhibitiors in *FLT3* mutated AML patients [11]) [9].

Table 1: Risk classification of acute myeloid leukemias. 2022 European LeukemiaNet (ELN) risk classification of acute myeloid leukemia (AML) by genetic abnormalities at initial diagnosis with frequent co-occurring mutations. Adapted from Döhner *et al.*, [6] and Juskevicius *et al.*, [9].

Risk category	Genetic abnormality	Frequent co-occurring mutations
Favorable	t(8;21)(q22;q22.1)/*RUNX1*::*RUNX1T1*	*RUNX1*, *NRAS*, *Cohesin*[a], *ASXl2*, *ZBTB7A*, *ASXL1*, *EZH2*, *KDM6A*, *MGA*, *DHX15*
	inv(16)(p13.1q22) or t(16;16)(p13.1;q22)/*CBFB*::*MYH11*	*NRAS*, *KIT*, *FLT3*-ITD, *KRAS*
	Mutated *NPM1*, without *FLT3*-ITD	*DNMT3A*, *Cohesin*[a], *NRAS*, *IDH1*, *IDH2*^{R140}, *PTPN11*, *TET2*
	bZIP in-frame mutated *CEBPA*	*GATA2*, *NRAS*, *WT1*, *CSF3R*
Intermediate	Mutated *NPM1*, with *FLT3*-ITD	*DNMT3A*, *Cohesin*[a], *NRAS*, *IDH1*, *IDH2*^{R140}, *PTPN11*, *TET2*
	Wild-type *NPM1* with *FLT3*-ITD (without adverse-risk genetic lesions)	
	t(9;11)(p21.3;q23.3)/*MLLT3*::*KMT2A*	
	Cytogenetic and/or molecular abnormalities not classified as favorable or adverse	
Adverse	t(6;9)(p23.3;q34.1)/*DEK*::*NUP214*	*FLT3*-ITD, *KRAS*
	t(v;11q23.3)/*KMT2A*-rearranged	*KRAS*, *NRAS*
	t(9;22)(q34.1;q11.2)/*BCR*::*ABL1*	
	t(8;16)(p11.2;p13.3)/*KAT6A*::*CREBBP*	
	inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2) /*GATA2, MECOM(EVI1)*	*NRAS*, *PTPN11*, *SF3B1*, *KRAS*, *GATA2*, *ETV6*, *PHF6*, *RUNX1*, *BCOR*, *ASXL1*, *NF1*
	t(3q26.2;v)/*MECOM(EVI1)*-rearranged	
	-5 or del(5q); -7; -17/abn(17p)	
	Complex karyotype, monosomal karyotype	
	Mutated *ASXL1*, *BCOR*, *EZH2*, *RUNX1*, *SF3B1*, *SRSF2*, *STAG2*, *U2AF1*, and/or *ZRSR*	
	Mutated *TP53*	

1.1 Core-binding factor AML

The core-binding factors (CBF), a class of hematopoietic transcription factors consist of DNA-binding CBFα with three subunits (*i.e.*, Runt-related transcription factor 1-3, RUNX1-3) and non-DNA-binding, but binding affinity increasing, CBFβ encoded by *CBFB* [12]. CBF AML is cytogenetically defined by the presence of t(8;21)(q22;q22) or inv(16)(p13.1q22) resulting in *RUNX1-RUNX1T1* or *CBFB-MYH11* fusion gene, respectively (hereafter referred to as AML t(8;21) and AML inv(16)) [13]. For both CBF AML entities the protein fusions RUNX1-RUNX1T1 (see Figure 3a) and CBFβ-SMMHC (see Figure 3b) convert RUNX1 from an activator of transcription to a repressor of transcription by acting as a negative inhibitor for RUNX1 during development [12].

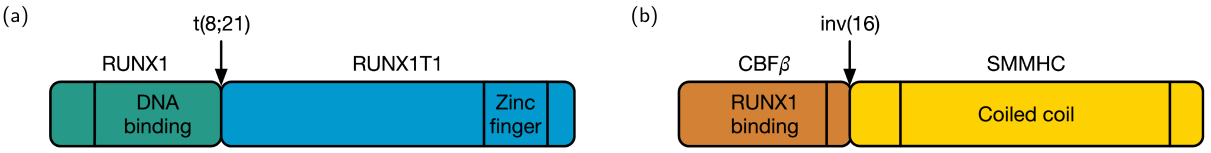

Figure 3: Structure of core-binding factor (CBF) fusion genes for AML with t(8;21) and inv(16). a) AML with t(8;21) is defined by the *RUNX1-RUNX1T1* fusion gene encoding for RUNX1-RUNX1T1 protein fusion and (b) AML with inv(16) is defined by the *CBFB-MYH11* fusion gene encoding for CBFβ-SMMHC protein fusion. Adapted from Speck *et al.*, [12] and Christen *et al.*, [14].

Both CBF AML entities have according to the 2022 European LeukemiaNet (ELN) risk classification a favorable prognosis as listed in Table 1 [6]. Jahn *et al.*, [5] and Opatz *et al.,* [15], have shown in 350 adult CBF AML patients that in both entities the most common mutated genes are genes involved in receptor tyrosine kinase (RTK)/RAS signaling, such as *NRAS*, *KIT* and *FLT3*. In this study, patients with t(8;21) show a more complex mutational landscape with somatic variants frequently found in genes involved in chromatin modification (*e.g.*, *ASXL1* and *ASXL2*) and DNA methylation (*e.g.*, *TET2* and *DNMT3A*), and in genes encoding for members of the cohesin complex (*e.g.*, *RAD21* and *SMC1A*). In comparison, they found that *WT1*, a transcription factor, and *BCORL1*, a transcriptional corepressor, were frequently mutated in inv(16) AML and mutations in genes involved in chromatin modification or belonging to the cohesin complex were very rare. *CEBPA* mutations improve overall survival (OS) for both CBF AML entities, whereas *KIT* mutations, which are frequent in both CBF AML entities, show a reduced OS only for patient with t(8;21) [16]. Recurrent secondary chromosomal abnormalities in CBF AML are trisomies of chromsomes 8, 21 and 22, with trisomy of chromosome 22 being the most common in inv(16). Deletions of chromosomes 9, X (for female patients) and Y (for male patients) are more frequent in t(8;21) [5,17].

1.2 Clonal hematopoiesis of indeterminate potential

Clonal hematopoiesis (CH) is defined by the presence of somatically acquired, cancer-associated mutations in hematopoietic cells without a history of a hematological malignancy and CH of indeterminate potential (CHIP) is defined by somatic mutations with a variant allele frequency (VAF) of at least 2%, resulting in at least 4% nucleated blood cells for heterozygous mutations.[18,19]. These somatic variants provide a fitness advantage to hematopoietic stem cells (HSCs) leading to accumulation of these cells [20]. Large sequencing studies have found that CHIP is associated with increased age and that at least 10% of persons older than 65 years [21] or older than 70 years [22] are carrying CHIP mutations. In sequencing data of blood derived samples from the Cancer Genome Atlas (TCGA), Xie *et al.*, [23] found that 5-6% of individuals over 70 years harbor somatic variants in genes involved in hematologic malignances.

The most commonly mutated genes in CHIP are epigenetic regulator genes, such as *DNMT3A*, *TET2* and *ASXL1*, also referred to as DTA mutations that are frequently mutated in leukemia [20]. It has also been shown in vitro that all DTA-mutations have effects on the self-renewal capacity of the HSC compartment and, especially, mutations in *TET2* show enhanced self-renewal capacities and age-related myeloid lineage predisposition [14,24,25]. CHIP can be further classified into myeloid (M-CHIP) and lymphoid CHIP (L-CHIP) depending if a mutation is located in a gene recurrently mutated in myeloid (*e.g.*, DTA mutations) or lymphoid malignancies (*e.g.*, *DUSP22*, *FAT1* and *KMT2D*) [26]. Here, mutations in *SRCAP* have shown a lymphoid bias and also an increase in DNA damage repair [27].

CHIP is associated with an increased risk for developing hematological malignancies and an increased risk for cardiovascular diseases (CVD) and coronary artery disease (CAD) ([20,26]. In AML patients, persistent CH-mutations with at least two mutations in more than 0.4% of the cells is strongly associated with lower leukemia-free survival and overall survival [28].

2 Cancer development and intra-tumor heterogeneity

In 1976 Peter C. Nowell [29] published “The Clonal Evolution of Tumor Cell Populations” proposing that most neoplasms arise from a single cell of origin that acquired a genetic variability leading to a proliferation advantage. Intra-tumor heterogeneity (ITH) describes the coexistence of multiple genetically distinct subclones within the tumor of a patient resulting from somatic evolution, clonal diversification, and selection [30]. Figure 4, adapted from Alessandro Lagana, 2022 [30], shows the different models of tumor development that results from cells acquiring somatic events. The clone harboring the somatic event that leads to tumor development is defined as founding clone. Linear evolution (Figure 4a) is the sequential acquisition of somatic events from one subclone to the next without a branching event. In the branching evolution (Figure 4b), the subclones acquire somatic events independently. Using targeted single-cell DNA sequencing (scDNA-Seq) it has been shown that linear and branching evolution exists in AML patients with branching evolution also showing convergent patterns defined by the independent acquisition of redundant somatic variants in subclones [31]. ITH is a main causal driver of therapy resistance in the clinic and the understanding of ITH and tumor development may lead to new targets for treatment [32]. Here, small subclones that are already present at a low cancer cell fraction prior to therapy or somatic variants acquired during therapy can drive resistance to therapy [33]. Figure 4c illustrates and example where two subclones are eradicated through therapy, but one subclone with acquired therapy resistance gave rise to new

clones after therapy. Furthermore, it has been shown that some genes have a specific clearance pattern and somatic variants in *DNMT3A*,*TET2, IDH1/2* and *KRAS* are persisting with a VAF >2.5% more than 30 days [34]. Early identification of subclones that potentially drive therapy resistance appears promising for treating patients, because if a patient harbors two somatic events that could be targets of therapy it is necessary to know which one has been acquired first to intervene in the hierarchical tumor development as early as possible [35].

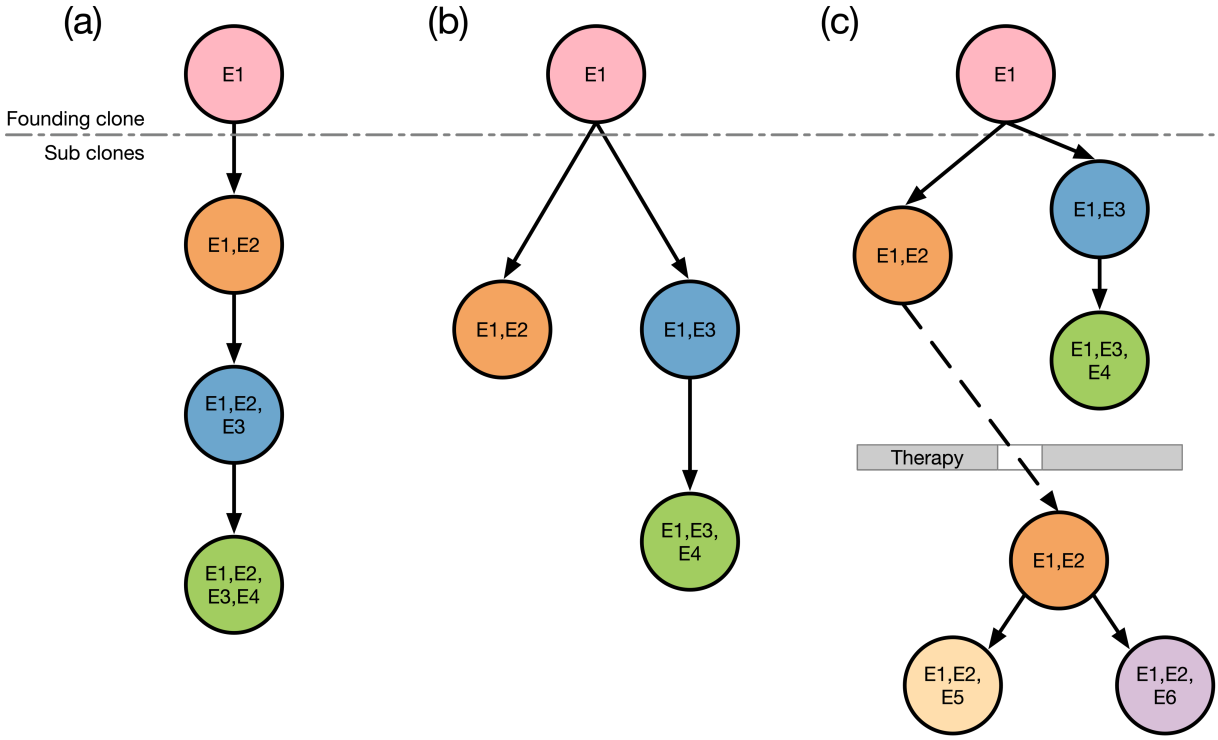

Figure 4: Models of tumor evolution. All models start with initiating somatic event E1 (*e.g.*, somatic variants) in the founding clone and subsequently, acquire additional somatic events E2-E6. (a) Linear evolution means that every subclone acquires somatic events E2-E4 subsequently. (b) Branching evolution means that during development subclones acquire somatic events independently. (c) Example of tumor evolution under the pressure of therapy with therapy resistant clone. Adapted from Alessandro Lagana, 2022 [30].

3 Reconstructing the history of somatic DNA alterations

Reconstructing the history of somatic DNA alteration allows for inferences on which mutations are in the same clone, to estimate the size of each clone and reconstruct the tumor phylogeny including common ancestors of clones or temporal order of clones [32]. In the following section information on methods and limitations of reconstructing ITH from bulk sequencing data (section 3.1) and single-cell DNA sequencing data (section 3.2) is presented.

3.1 Bulk sequencing

In general, bulk sequencing methods for reconstructing subclonal architecture of tumors try to estimate for somatic variants and/or copy-number alterations the fraction of cancer cells

harboring the variant referred to as cancer cell fraction (CCF) [36]. Copy-number alterations (CNAs) affect the VAF of mutations located in them (*e.g.*, a deletion increases the VAF of a mutation if the wild-type allele is lost) and, therefore, need to be considered when inferring CCF [37]. Here, some methods for CCFs are limited to SNVs only in copy-neutral regions (*e.g.*, sciClone [38]), whereas others can include SNVs in regions affected by copy-number alterations (*e.g.*, PyClone [39], PyClone-VI [40] and PhyloWGS [41]). Fu *et al.*, [42] developed a method that infers phylogeny from SNVs, structural variants (SVs) and CNAs using multi-regional tumor samples from one patient.

In Christen *et al.,* [14], we used targeted deep sequencing of somatic variants detected in patients with t(8;21) AML to reconstruct clonal evolution throughout therapy and multiple timepoints (*e.g.*, diagnosis, complete remission, relapse) using a pipeline of sciClone [38], ClonEvol [43] and Fishplot [44]. Here, only the use of targeted deep sequencing and multiple timepoints allowed us to infer the founding clone and the temporal order of subclones.

Bulk sequencing can hardly distinguish low level VAF mutations from sequencing library and sequencing error artefacts, so the information if somatic events are present in the same cell is lost [45]. It has been shown that relapse is often driven by clones which were only subclonally present at diagnosis and expanded thereafter, but bulk sequencing technologies failed to define its exact phylogeny [14,46]. Furthermore, reconstructing tumor phylogeny in bulk sequencing has its limitations with low VAF mutations where often multiple phylogenetic trees are possible precluding a generalizable conclusion. Here, the use of multiple samples – either collected at various timepoints or different anatomical sites - can improve the bioinformatical analysis and prediction strength [47,48].

3.2 Single-cell DNA sequencing

Lähnemann *et al.*, [49] specified eleven grand challenges in single-cell data science and, specifically, for single-cell genomics following challenges: (i) dealing with errors and missing data, (ii) scaling phylogenetic models to work with many cells and (iii) integrating multiple types of genomic variation (*e.g.*, SNVs and SCNAs) into phylogenetic models.

scDNA-Seq methods can be classified into whole-genome scDNA-Seq methods with shallow coverage throughout the genome suitable for detecting copy-number alterations and targeted scDNA-Seq methods with high coverage for specific cancer related regions suitable for accurate identification of SNVs [50]. Technical errors in single-cell data, as shown in Figure 5, are false positive errors with (a) erroneous bases that are introduced during amplification or

sequencing, and false negative errors, such as dropouts of (b) the whole locus or (c) one allele and (d) an imbalance amplification of one alle [51,52].

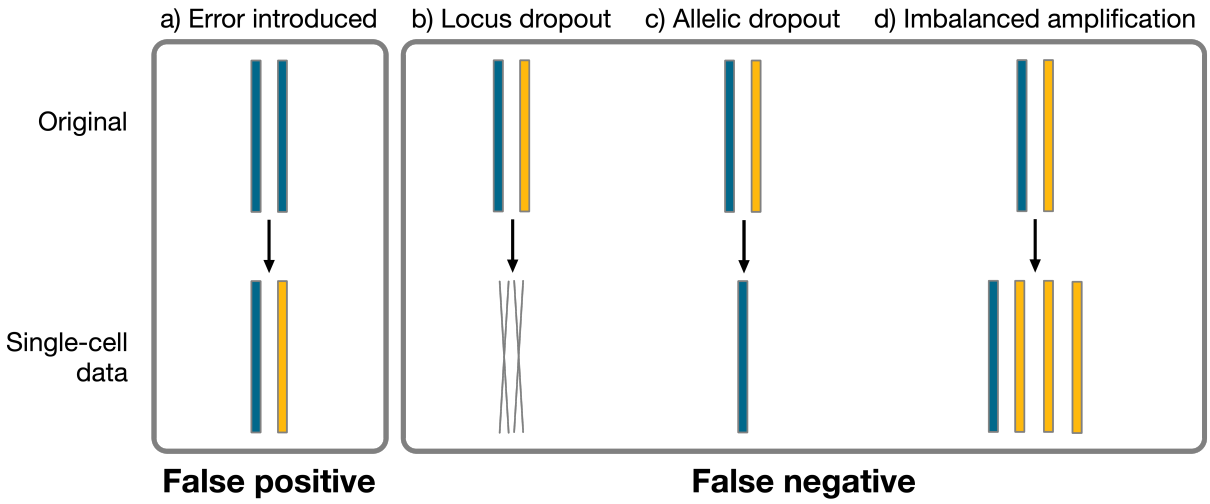

Figure 5: False positive and false negative calls in single-cell data. a) False positive call with amplification of sequencing error and introduction of erroneous base. False negative calls can be introduced by (b) a complete locus drop-out leading to no call at all, (c) allelic drop-out leading to wild-type or homozygous call depending on the allele that is lost and (c) imbalanced amplification leading also to misclassification of somatic variant. Adapted from Gawad *et al.*, [51] and Navin *et al.*, [52].

Recently developed methods for reconstructing phylogenies from scDNA-seq data are (i) the infinite-sites model, (ii) the k-Dollo model and (iii) the finite-sites model [53]. The infinite-sites model, which is used in SCITE [54], ∞SCITE [55], B-SCITE [56] OncoNEM [57] and Sciφ [58], and is the simplest phylogenetic model that allows variant to be gained once, but never to be lost. The (k-)Dollo model, which is used in SPhyR [59], SASC [60], PyDollo [61] and ConDoR [50], adds to the infinite-sites model the ability that a variant can be lost multiple or for the parameterized version k (=user defined integer) times. The finite-sites model, which is used in SiFit [62] and PhiSCS [63], lifts the limitation on losses and allows gains and losses multiple times. Additionally, COMPASS [64] uses a probabilistic model and SCARLET [53] uses a loss-supported phylogeny model to reconstruct tumor phylogenies.

Most of the tools were developed for inferring tumor phylogeny from approximately 100-1,000 cells, which is feasible for whole-genome scDNA-Seq, but not for targeted scDNA-Seq where cell numbers can go up to a few thousands cells per sample [64]. Only ∞SCITE [55], ConDoR [50] and COMPASS [64] are able to reconstruct tumor phylogeny in a reasonable amount of time for up to 10,000 cells, with ConDoR [50] and COMPASS [64] allowing for inferring tumor development using copy-number alterations and somatic variants. ConDoR [50] utilizes a constrained k-Dollo model with the assumption that SNVs and single-nucleotide polymorphisms (SNPs) can be lost only due to an overlapping copy-number alteration that happens only once in the phylogenetic tree. Copy-number clusters of cells used are needed in

advance to infer tumor phylogeny using ConDoR [50]. In comparison, COMPASS [64] does not need any a priori computation and performs a simulated annealing approach at first without SCNAs and in a subsequent step adds copy-number events to the phylogenetic tree.

4 MissionBio Tapestri Platform

The Tapestri platform (MissionBio) is a targeted scDNA-Seq platform that allows for the simultaneous detection of somatic variants (*i.e.*, SNVs and insertion or deletions (INDEL)), SCNAs and cell surface proteins. In this thesis, we used the targeted single-cell genomics workflow without protein detection [65]. The targeted regions are specified by the panel that is used for library preparation. Here, it is possible to use catalogue, published or, as in this thesis, custom panels.

Single-cell libraries are prepared using the Tapestri Instrument (MissionBio) that performs a two-step microfluidic approach as following: (i) single-cells are encapsulated in oil droplets with protease releasing DNA from histone and DNA binding proteins and (ii) the encapsulated cell lysate is again encapsulated with a reagent mix and barcoding beads. These barcoding beads consist of a read 1 sequence for Illumina indexes (Illumina), a 9-bp cell barcode (individual barcodes >3 Levenshtein distance [66] apart) and a common sequence that can bind to the amplified target region. Within each droplet gene specific primers (GSP) are used to amplify target regions and the common sequence attached to GSP bind to the barcoding beads resulting in DNA fragments with target region and cell barcode. Amplified products (single sample) can be pooled with Illumina sequencing chemistry and sequenced on an Illumina sequencer targeting approximately 5,000 cells and an average coverage per amplicon of 80x.

Sequencing reads in combination with panel information (*e.g.*, custom panel) are initially processed using the Tapestri pipeline (MissionBio). Here, adapter sequences are trimmed from raw reads using Cutadapt [67] and subsequently aligned to the reference genome using bwa-mem [68]. True barcodes are detected if they match a whitelist of known barcodes (exact match) and if they do not match exactly barcodes are selected using a Levenshtein distance [66] of 3. At first, cells are dropped if they have less than $10 * n\_amplicons$ total reads. With this subset a threshold is calculated using $0.2 * \sum_{i}^{cells} \sum_{j}^{amplicons} n_{ij}$ with n_{ij} as the number of reads of amplicon j in cell i. If this threshold is <10 then 10 is used as a threshold. The remaining cells are dropped if they have <80% of amplicons with reads above the calculated threshold or 10. The filtered cells are genotyped using the Genome Analysis Toolkit (GATK) [69] and stored as a Loom file [70], which is an efficient file format for large omics data. In case of MissionBio,

the Loom file consists of 5 layers (*i.e.*, numerical genotype (NGT), number of reads with evidence of mutation (AD), read depth (DP), genotype quality (GQ) and number of reads with evidence of no mutation (RO)). I used this file for downstream analysis as explained in section IV2.

II Overview

1 Aim of this thesis

The overall goal of this thesis is to investigate the subclonal architecture and tumor development of core-binding factor acute myeloid leukemia at a single-cell resolution. For this, I developed and applied a novel algorithm for the systematic integration of single-cell and bulk tumor sequencing data. Here the combined analysis of bulk sequencing data, with a high genome-wide resolution, and single-cell sequencing data, with a high clonal resolution should be used to unravel ITH for both large-scale (*e.g.*, SCNAs) and small-scale (*e.g.*, SNVs and INDELs) genetic alterations.

To achieve this goal, I first prepared a combined bulk and single-cell CBF-AML dataset. These two datasets were used to unravel clonal architecture of patients through an integrated analysis of bulk and single-cell genome sequencing data. Last, I validated the results of the phylogenetic trees using diagnostic information (*e.g.*, karyotype) and bulk sequencing data.

2 Patient cohort

The patient cohort for this thesis consists of 2 female and 7 male patients with CBF AML and samples at diagnosis (D), complete remission (CR) and relapse (Rel). The age of investigated patients at diagnosis ranged from 30 years to 67 years as shown in Table 2. According to the French-American-British (FAB) classification (M0-M7), the morphology of AML cells were classified as myelomonocytic leukemia (M4) (6/9), acute myeloblastic leukemia with maturation (M2) (2/9) and acute myelomonocytic/monocytic leukemia (M4/M5) (1/9) [71,72]. All patients reached CR and relapsed within 27 months (range 4 to 27 months). Overall survival (OS) is defined as the time between diagnosis and endpoint death (OS_{status}=1) or alive at last follow-up (OS_{status}=0) and ranges from 14 to 104 months. Relapse free survival (RFS), according to ELN 2017 [73], is defined for patients reaching CR as time between CR and Rel or OS event. In this cohort, 5 patients died (*i.e.*, patients 01, 02, 03, 07 and 09) and 4 patients (*i.e.*, patients 04, 05, 06 and 08) were censored at last follow-up.

Induction therapy for patients in this cohort consisted of daunorubicin (DNR) and cytarabine (Ara-C) (DA) (4/9), idarubicin-cytarabine-etoposide (ICE) (2/9), idarubicin-etoposide-(intermediate-dose) cytarabine (IVA) (2/9) and ICE in combination with all-trans retinoic acid (ATRA) (1/9). Consolidation treatment consisted of high-dose Ara-c (HD-Ara-c) and patient 09 received additionally DNR. In total, 7 patients received salvage treatment

consisting of mitoxantrone, topotecan and cytarabine (MTC) (3/7), high-dose cytosine arabinoside and mitoxantrone (HAM) (2/7), fludarabine, cytarabine and idarubicin (FLA-IDA) (1/7) and Ara-c combined with vosaroxin or placebo (1/7) followed by allogeneic hematopoietic stem cell transplantation (allo-SCT).

Table 2: Overview clinical data. Baseline characteristics are listed for patients with sex, age (in years), Eastern Cooperative Oncology Group (ECOG) performance status, French-American-British (FAB) classification, induction treatment, consolidation treatment, salvage treatment, relapse-free survival (RFS), overall survival (OS) and survival status (OS_{status}). RFS and OS are listed in months. If patient was alive at last follow-up OS_{status} is 0 and if patient died OS_{status} is 1. (Pat. = Patient, Tx = Treatment)

Pat.	Sex	Age	ECOG	FAB	Induction Tx	Consolidation Tx	Salvage Tx	RFS	OS	OS_{status}
01	male	44	90%	M2	ICE	high-dose Ara-c	MTC	24	35	1
02	male	48	100%	M4	ICE + ATRA	high-dose Ara-c	MTC	4	53	1
03	male	56	100%	M4	IVA	high-dose Ara-c	MTC I	14	33	1
04	male	30	90%	M4	DA	high-dose Ara-c	HAM	9	22	0
05	male	65	80%	M4/M5	ICE	high-dose Ara-c	FL-A-Ida	14	79	0
06	male	36	90%	M4	DA	high-dose Ara-c	HAM	11	104	0
07	male	67	100%	M4	DA	high-dose Ara-c	NA	9	14	1
08	female	52	100%	M4	DA	high-dose Ara-c	AraC + Vosaroxin/Placebo	14	80	0
09	female	56	80%	M2	IVA	MHD-ARAC/DNR	NA	27	36	1

Table 3 lists karyotypes, CBF type, *FLT3*- internal tandem duplication (ITD) status and *KIT* mutation status for patients in this cohort at diagnosis. Patients 01 and 09 were classified as a t(8;21) CBF AML with *RUNX1-RUNX1T1* gene fusion and the remaining 7 patients were classified as inv(16) CBF AML with *CBFB-MYH11* gene fusion. Patient 02 harbored a *FLT3*-ITD at diagnosis with an allelic ratio of 13%, which is associated with a poorer outcome in CBF AML patients in comparison to those without *FLT3*-ITD [74]. A c-KIT exon 8 frameshift mutation, which has a negative effect on relapse and RFS in CBF AML patients, has been detected in patient 03 at diagnosis [75]. Conventional G banding did not detect any secondary cytogenetic abnormalities in patients 02, 06 and 09. Secondary cytogenetic abnormalities, as defined by Han S. *et al.* [17], are those events that are present in addition to the CBF AML defining inv(16) or t(8;21) detected by conventional G-banding. Furthermore, they defined a complex karyotype with two or more secondary cytogenetic events and subclones as those that are only present in a fraction of metaphases (*e.g.*, +22 1/49) with the dominant clone as the one with the highest number of metaphases [17]. For patients 01 and 03 a trisomy of chromosome 8 (+8) and for patient 08 a trisomy of chromosome 22 (+22) has been detected by karyotype as shown in Table 3. It has been shown that trisomy of chromosomes 8, 21 and 22 are more common in patients with inv(16) [17]. Male patient 03 has a complex karyotype with a deletion of chromosome 11 and a trisomy of chromosome 8 in addition to inv(16). Cytogenetics in patient 04 identified a dominant clone harboring trisomy of chromosomes 13, 14 and 22 in

13/49 metaphases and a subclone harboring a trisomy of chromosome 22 in only 1/49 metaphases.

Table 3: Overview cytogenetics. Core-binding factor (CBF) acute myeloid leukemia patients listed with either an inversion on chromosome 16 (inv 16) or a translocation between chromosome 8 and 21 t(8;21). The number of subclones harboring chromosomal abnormalities are indicated with a fraction of metaphases (*e.g.,* 1/49 meaning that one metaphase of 49 harbors the chromosomal abnormalities). Amplifications are indicated by "+" and deletions are indicated by "del". For patients 04 and 05 nuclear in situ hybridization (nuc ish) data is available. Clinical testing also identified a *FLT3* internal tandem duplication (*FLT3*-ITD) in patient 02 with an allelic ratio (AR) of 13%. (Pat. = Patient)

Pat.	Karyotype	CBF	*FLT3*-ITD	*KIT*
01	47, XY, +8, t(8;21) (q22;q22), inv (9) (p11 q12)	t(8;21)	WT	WT
02	46, XY, inv(16)	inv(16)	MUT (AR 13%)	WT
03	46, XY, del 11, inv 16, +8	inv(16)	WT	MUT
04	46, XY, inv(16) (p13q22) 7/47, idem, +22 1/49, idem, +13, +14, +22 13/49. nuc ish 16q22 (CBFBx2)(5CBFBsep3CBFBx1) 99/100, 11q23(MLLx2), 3q2(EVI1x2) 100	inv(16)	WT	WT
05	46, XY, inv(16) (p13q22) 7/46, idem, del(7)(q31q33) 13/46. nuc ish 3q26(EVI1x2), cen7(CEP7x2), 7q31(D7S486x2), 11q23(MLLx2) 100, 16q22(CBFBx2) 96/100	inv(16)	WT	WT
06	46, XY, inv(16) (p13q22)	inv(16)	WT	WT
07	46, XY, del(16)(p12), inv(16)(p13q22) 3/ 46, idem, +(3;12)(q12;p13), add(19)(q13) 5/46, idem, der(X)t(X;17)(p11;q11), del(5)(q23q34), +mar 4/46, idem, add(7)(q31)	inv(16)	WT	WT
08	47, XX, inv(16), +22	inv(16)	WT	WT
09	46, XX, t(8;21)	t(8;21)	WT	WT

Figure 6 visualizes overall survival ranging from 14 to 104 months with endpoints alive and dead for each patient. The time of sampling peripheral blood (PB) or bone marrow (BM) in reference to time of diagnosis is shown for samples at diagnosis, complete remission and relapse. For patients 07 and 08 no complete remission samples and for patient 06 no relapse sample were available. Written consent was obtained in accordance with the declaration of Helsinki and ethical approval was obtained from the local ethics committees of the cooperating institute.

DNA samples and cells at D, CR and Rel were derived from PB samples or BM aspirates. Percentage of blasts and if gene fusion has been detected (pos. = gene fusion has been detected and neg. = gene fusion has not been detected) are listed in Table 4. For patient 07 and 08 T-cells from diagnosis were used as a germline control for bulk analysis.

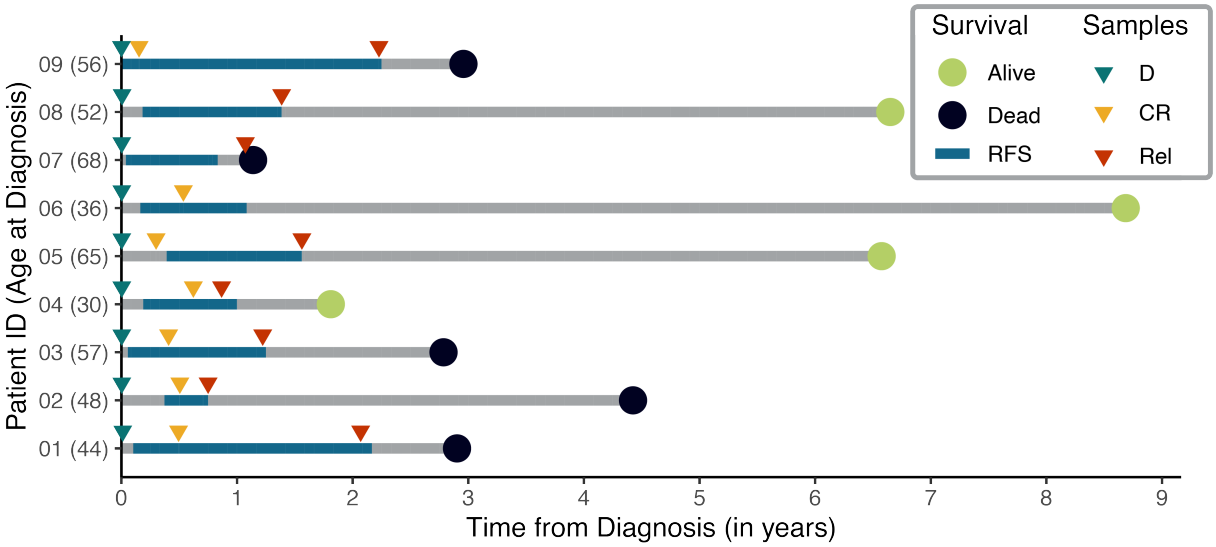

Figure 6: Timeline of disease progression and sample collection. This plot visualizes the survival time of each patient starting at diagnosis and ending with status alive or dead. Collection of samples at diagnosis (green), complete remission (yellow) and relapse (red) are visualized by triangles. Relapse-free survival (blue line) is defined as the time from complete remission to time of relapse.

Table 4: Overview patient samples. Samples derived from bone marrow (BM) or peripheral blood (PB) are listed with percentage of blasts and if gene fusion has been detected (pos. = detected and neg. = not detected). For patients without a complete remission sample (*i.e.*, patient 07 and 08) DNA from T-cells at diagnosis was used as a germline control. (N.A. = not available)

Patient	Sample	Type	DNA	Cells	Blast count [%]	Gene fusion
01	Diagnosis	BM	✓	✓	>50	pos.
	Complete remission	BM	✓	✓	<5	neg.
	Relapse	BM	✓	✓	30-40	pos.
02	Diagnosis	PB	✓	✓	84	pos.
	Complete remission	BM	✓	✓	<5	N.A.
	Relapse	BM	✓	✓	50	pos.
03	Diagnosis	PB	✓	✓	82	pos.
	Complete remission	BM	✓	✓	2-3	N.A.
	Relapse	BM	✓	✓	50	pos.
04	Diagnosis	PB	✓	✓	70	neg.
	Complete remission	BM	✓	✓	2	neg.
	Relapse	BM	✓	✓	17	pos.
05	Diagnosis	BM	✓	✓	60	pos.
	Complete remission	BM	✓	✓	<5	pos.
	Relapse	PB	✓	✓	59	pos.
06	Diagnosis	PB	✓	✓	95	pos.
	Complete remission	BM	✓	✓	<5	neg.
	Relapse	BM	✗	✗	15	pos.
07	Diagnosis	BM	✓	✓	90	pos.
	Complete remission	BM	∗	✗	0	pos.
	Relapse	BM	✓	✓	80	pos.
08	Diagnosis	BM	✓	✓	80	pos.
	Complete remission	NA	∗	✗	N.A.	N.A.
	Relapse	BM	✓	✗	20-25	pos.
09	Diagnosis	PB	✓	✓	40	pos.
	Complete remission	BM	✓	✓	<2	pos.
	Relapse	PB	✓	✗	50	pos.

✓ . . . Sample available
✗ . . . Sample missing
∗ . . . T-cells from diagnosis

3 Workflow

To uncover the full complexity of ITH with a targeted single-cell DNA sequencing approach, this thesis consists of (i) a bulk and (ii) a single-cell sequencing part as shown in Figure 7. For using the Tapestri (MissionBio) targeted single-cell DNA sequencing platform patient specific information on somatic variants, SCNAs and breakpoints of CBF specific fusion genes (*i.e.*, *CBFB-MYH11* and *RUNX1-RUNX1T1*) needs to be known prior.

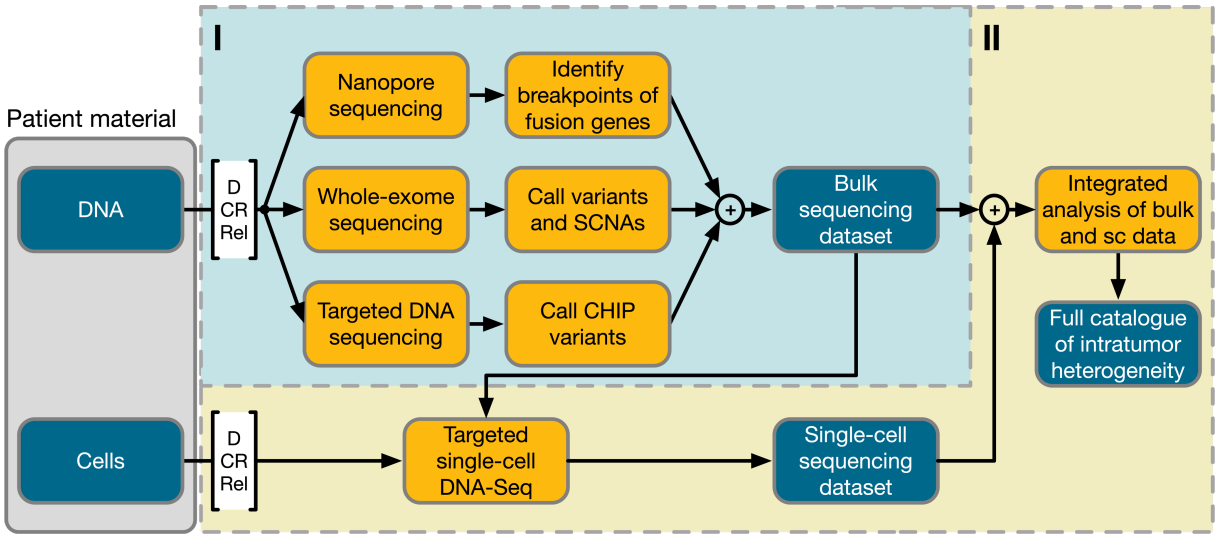

Figure 7: General workflow. The thesis consists of (I) a bulk-sequencing and (II) a single-cell sequencing part. In part I DNA samples are used for whole-exome sequencing, targeted DNA sequencing and Nanopore sequencing for calling somatic variants, somatic copy-number alterations and for identifying patient specific breakpoints of the core-binding factor fusion genes, respectively. In part II this information is used for targeted single-cell DNA sequencing and an integrated analysis using bulk and single-cell data to uncover the full catalogue of intratumor heterogeneity is performed. (CHIP = clonal hematopoiesis of indeterminate potential, D = diagnosis, CR = complete remission, Rel = relapse, SCNA = somatic copy number alterations, sc = single-cell)

In the bulk sequencing part, DNA samples at diagnosis, complete remission and relapse was used to generate whole-exome sequencing data. I used the whole exome sequencing data for calling somatic variants (*i.e.*, SNVs and INDELs) and SCNAs with the complete remission sample or the T-cells from diagnosis as a germline control. Patient specific breakpoints of the CBF AML gene fusions were identified using Nanopore (Oxford Nanopore Technologies) long-read sequencing data at diagnosis. For a better resolution on genes involved in clonal hematopoiesis, targeted DNA sequencing with a 25-gene CHIP panel using error-corrected reads, which has been well established in our group [19,76,77], was performed.

Subsequently, information on somatic variants, SCNAs and fusion gene breakpoints for each patient was combined to design custom panels (~200 amplicons per panel) for single-cell sequencing. The Tapestri platform (MissionBio) was used for generating single-cell libraries of all available samples and after sequencing the reads were initially processed using the Tapestri pipeline (MissionBio). Single-cell data was used in combination with prior knowledge from the

bulk sequencing part to reconstruct tumor phylogenies. I hypothesized that such integrated analysis of both datasets would enable conclusions to be drawn on intra tumor heterogeneity and changes in the clonal composition throughout the treatment.

III Bulk sequencing

In this chapter I present methods and results from bulk sequencing that were used as prerequisite for targeted single-cell DNA sequencing. I used whole-exome, targeted and Nanopore sequencing data to uncover somatic variants (*i.e.*, SNVs, INDELs and *FLT3*-ITDs), SCNAs and CBF AML specific fusion genes in 2 patients with t(8;21) AML and in 7 patients with inv(16) AML.

1 Sequencing file formats

Raw reads from an Illumina Sequencer are stored in the FASTQ format [78] consisting of four lines for each read as shown in Insert 1: (i) first line starting with "@" is a sequence identifier and can hold optional description, (ii) second line is the raw sequence, (iii) third line begins with "+" and can be optionally followed by the same sequence identifier as line 1 and (iv) the fourth line encodes the PHRED quality score of each base call for the sequence of line 2.

```
@A00643:342:HLHTJDRXY:1:2101:7383:1000 1:N:0:GTCTGTCA
NGTTAGCACATCATAGAGGAGCCAAAGTGATTTCAACAGGATGCAGCCTTGAAGATAAGCAGTGCCTTGAAAGATTCAGACCTCCCATAGGT
GGGTAATATTATGAGCACAGACTTTAAAACAGGAAATTTTGAAGGAAAATCACCTTTAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Insert 1: FASTQ example.

If reads are aligned to a reference sequence or reference genome than those aligned reads are stored in a Binary Alignment Map (BAM) file, which is a binarized/compressed version of the Sequence Alignment/Map format (SAM) [79]. The SAM format is a tab-delimited text format consisting of header lines starting with "@" followed by alignment lines with 11 mandatory fields (*e.g.*, mapping position) and a variable number of optional fields. These optional fields are each labeled with a tag and displayed as TAG:TYPE:VALUE. The type explains the format of the value, such as Z for string values.

2 Variant calling pipeline

I established a variant calling pipeline in Snakemake (v6.12.3) [80] for reads containing unique molecular identifier (UMI) that I used in Arends *et al.*, 2022 [19] and Arends *et al.*, 2023 [76]. The pipeline was also further developed and has been used by colleagues for Panagiota *et al.*, [77]. The pipeline consists of a preprocessing (section 2.1) and a variant calling part (section

2.2). The whole pipeline was used for the targeted sequencing data and only the variant calling part was used for the whole-exome sequencing data of this thesis.

Detailed parameters for each step of the pipeline can be found in the supplement.

2.1 Preprocessing

In the preprocessing part raw reads are aligned to a reference genome and processed to obtain consensus reads that can be used for variant calling.

At first, I used Picard's (v2.20.0) [81] ExtractIlluminaBarcodes and IlluminaBasecallsToSam to extract unmapped BAM (uBAM) files from Illumina basecalls. The uBAM stores the unaligned reads with the UMI sequence (*e.g.*, `RX:Z:TTATGATAT`) as a RX SAM tag. I aligned uBAMs to hg19 [82] reference genome using Picard's SamToFastq, bwa mem (v0.7.17) [68] and Picard's MergeBamAlignment subsequently. To group the reads based on UMIs and further create consensus reads with a minimum of 3 supporting UMIs, I executed fgbio's (v 0.6.1) [83] GroupReadsByUmi and CallMolecularConsensusReads. These error corrected reads were again mapped to hg19 [82] as before. For quality filtering, I executed fgbio's FilterConsensusReads with a minimum of 3 supporting UMIs, consensus bases with a quality >5 and default parameters.

2.2 Variant calling

In the variant calling part, BAM files are used for variant calling and the calls are subsequently annotated using different databases.

These preprocessed, error-corrected BAM files are used for variant calling with VarDictJava (v 1.8.2) [84] in single-sample mode with a minimum VAF of 0.1%, the reference genome used in preprocessing, the bed file provided by Twist BioScience and default parameters. In case of whole-exome sequencing data, aligned reads with the bed file of the library preparation kit were used as input for variant calling. I converted the raw variant calls with bcftools' (v1.11) [85] view, bcftools' index, bcftools' norm, and, subsequently, R (v4.0.5) [86] to the correct format for downstream tools. I used ANNOVAR (version 2020-06-07 23:56:37 -0400 (Sun, 7 Jun 2020)) to annotate unfiltered variant calls with the following databases: refGene [87], clinvar_2021050 [88], dbnsfp42c [89,90], gnomad_exome [91], avsnp150 (dbSNP140) [92], cosmic70 [93], revel [94], nci60, icgc28 [95], snp142 [92] and popfreq_all_20150413 (containing frequencies from 1000G, ESP6500, ExAC and CG46).

I further annotated these variant calls using R with three manually curated lists from our group to add information if genes are known AML drivers [96–98] or AML candidate genes

[99] and if variants are published CHIP variants [21,22,100–105] or CHIP hotspot mutations [106]. A variant was flagged as "important" if the gene is (i) a AML candidate or (ii) a AML driver gene, the variant (iii) is a known CHIP mutation or (iv) is associated with hematopoietic diseases according to the cosmic database [93]. Additionally, I calculated a measure for strand bias, annotated as FisherStrand, based on the Fisher's Exact Test as used in the Genome Analysis Toolkit (GATK) [107].

3 Detecting somatic variants and copy-number alterations

3.1 Whole-exome sequencing preparation

Libraries were generated from DNA samples (see Table 4) using the SureSelect Human All Exon v7 XT HS kit (Agilent) and sequenced on the NovaSeq 6000 platform (Illumina; 300 cycles, paired-end). Raw reads in FASTQ format were provided by the Genomics Platform of the Max Delbrück Center for Molecular Medicine (MDC) and Berlin Institute of Health at Charité (BIH).

3.2 Preprocessing

I used Trimmomatic (v0.36) [108] in paired-end mode with "`LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36`" and default parameters for quality trimming of raw reads and, subsequently, bwa mem and samtools sort (v1.11) [85] to align the reads to the hg19 [82] reference genome. To remove polymerase chain reaction (PCR) duplicates, I executed Picard's MarkDuplicates.

These aligned and deduplicated reads in BAM format I used for calling somatic variants (section 3.3) and somatic copy-number alterations (section 3.4).

3.3 Variant calling

I processed the whole-exome sequencing BAM files using the variant calling part (section 2.2) of the UMI variant calling pipeline I developed with a VAF threshold of 1%. I removed variants for quality criteria in each patient (for patient 06 filtering criteria were applied to diagnosis only) as following:

- VAF >1% at complete remission (*i.e.*, patients 01, 02, 03, 04, 05, 06 and 09) or in extracted T-cells from diagnosis (*i.e.*, patients 07 and 08)
- VAF <4% (VAF <6% for patient 03) at diagnosis and relapse
- read depth <50 at diagnosis and relapse

- variant read count <6 at diagnosis and relapse
- FisherScore ≥20 at diagnosis and relapse
- StrandBalance1 or StrandBalance2 is 0|1|NA at diagnosis and at relapse (NA = not available, which in this case are infinite values, such as divisions by zero)

Additionally, I removed variants if VAF at diagnosis is >1% and <4% and VAF at relapse is <5% or vice versa to reduce noise.

From the remaining set I removed variants that are annotated in SNP databases (except those with an important flag) with a minor allele frequency (MAF) >0.01% in the gnomad_exome [91], avsnp150 (dbSNP140) [92] or popfreq_all_20150413 (PopFreqMax) [109]. Additionally, synonymous and non-frameshift variants were dropped from the variant list. Remaining variants were manually checked and further filtered by visual inspection using Integrative Genomics Viewer (IGV) (v 2.11.6) [110].

Additionally, somatic variants were also called from raw reads by a collaboration partner as previously described in Yoshida *et al.*, [111] and Kataoka *et al.*, [112]. I used these variant calls to confirm manual filtering criteria.

For the identification of the *FLT3*-ITD in patient 02, I used ITDetect (v.1.4) [113] on all BAM files of this patient with the chromosomal location of *FLT3* (chr13:28608020-28608360) and default parameters.

3.4 Copy-number calling

I identified SCNAs using refphase (v0.1.1) [114,115] in combination with ASCAT (v3.1.0) [116] according to the ”Complete Example Workflow” in the repository. SCNAs are identified using the B-allele frequency (BAF) from heterozygous SNPs and the log read-depth ratio (LogR), which are calculated based on the read depth information at those positions [115].

I downloaded the dbSNP database (build: 151, reference: GRCh37.p13) [92] and used bcftools’ view to get position of overlapping SNPs with the bed file from Agilent. Then I used bcftools’ mpileup to pile up reads from the deduplicated BAMs at those positions and bcftools’ query to obtain the appropriate input format for further analysis. The format is a tab separated list with chromosome, chromosomal position, read count for the reference allele and the count for the alternative allele. These tables were processed using R and following libraries: ASCAT (v3.1.0) [116], refphase (v0.1.1) [114,115], dplyr (v1.1.1) [117], tidyr (v1.3.0) [118], glue (v1.6.2) [119], gtools (v3.9.4) [120] and readr (v2.1.4) [121]. At first, I filtered the germline sample (*i.e.*, complete remission or T-cells from diagnosis) for positions located only on autosomes, without additional calls on the same position and with a minimum read count of 50

reads for reference and alternative reads combined. Additionally, I used a sliding window filter allowing less than three positions in a range of 150bp. The positions for each tumor sample of a patient were filtered for positions in the normal sample and, additionally, a minimum read count of 50 reads for reference and alternative reads combined. Then LogR and BAFs are calculated for diagnosis and relapse samples with germline samples as reference. Each position is classified as following:

$$BAF_{normal} = \begin{cases} < 0.1\ germline\ homozygous \\ > 0.9\ germline\ homozygous \\ others, germline\ heterozygous \end{cases}$$

I used ASCAT's ascat.runAscat with `gamma=1` and default parameters, except for the diagnosis sample of patient 05 where I additionally set `rho_manual = .9` and `psi_manual = 2.1`. I executed refphase according to the example workflow and filtered results by visual inspection of LogR and BAF.

4 Detecting CHIP mutations

4.1 Targeted DNA sequencing data

Libraries were prepared using a commercially available library preparation kit (Twist BioScience) and a customized targeted sequencing panel (Twist BioScience) covering 45 genes recurrently mutated in CH (see Supplemental Table 1), which has been used in recent projects of our group [19,76]. These libraries are prepared using a 9 bp long unique molecular identifier (UMI) that is used for error correcting and therefore enables the calling of low VAF somatic variants. Libraries were sequenced on the NovaSeq 6000 platform (Illumina; 300 cycles paired-end) with read lengths of 148bp and a read length of 17bp for index i7 and 8bp for index i5.

4.2 Variant calling

I processed raw Illumina basecalls using my established variant calling pipeline as described in section 2. I filtered annotated variants from the UMI variant calling pipeline using the following quality criteria in each patient (for patient 06 filtering criteria were applied to diagnosis and for patient 08 to relapse only):

- VAF >1% at complete remission for patients 01, 02, 03, 04, 05, 06 and 09
- VAF <2% at diagnosis and relapse
- read depth <50 at diagnosis and relapse
- variant read count <4 at diagnosis and relapse

- FisherScore ≥20 at diagnosis and relapse
- StrandBalance1 or StrandBalance2 is 0|1|NA at diagnosis and at relapse (NA=not available, which in this case are infinite values, such as divisions by zero)

From the remaining set, I removed variants that are annotated in SNP databases (except those with an important flag) with a MAF >0.01% in the gnomad_exome [91], avsnp150 (dbSNP140) [92] or popfreq_all_20150413 (PopFreqMax) [109]. Additionally, synonymous variants, non-frameshift variants and variants already present in the final somatic variant list from whole-exome data were dropped from the variant list. Remaining variants were manually checked and further filtered by visual inspection using IGV.

5 Identifying breakpoints of fusion genes

I aligned Nanopore reads using Vulcan (v1.0.3) [122] to the humanG1Kv37 reference genome [123] with `--nanopore` and default parameters. This reference genome from the 1000 Genomes Project [123] was used because a reference genome without `chr` annotation is needed for downstream analysis. To identify the patient specific breakpoints of the fusion gene I performed NanoFG (v1.0) [124] on the mapped reads with `-s 'CBFB,MYH11'` in case of inv(16) (*i.e.,* patient 01 and 09) or `-s 'RUNX1,RUNX1T1'` in case of t(8;21) (*i.e.,* patient 02, 03, 04, 05, 06, 07 and 08), `-pdf 400` to get a FASTA sequence flanking ±400 bp of the breakpoint sequence and default parameters. In case of patient 03, I additionally used the do not filter option (`-df`) to get a result. I extracted quality metrices from reads and mapped reads using NanoStat (v1.6.0) [125].

6 Results

In this section, I present results from whole-exome, targeted and Nanopore sequencing. I called and identified somatic variants (*i.e.*, SNVs, INDELs and *FLT3*-ITDs), somatic copy-number alterations and CBF AML gene fusions (*i.e.*, *CBFB-MYH11* and *RUNX1-RUNX1T1*). These patient specific results have been used for establishing the custom panels for targeted single-cell DNA sequencing.

6.1 Somatic variants

I identified a total of 249 somatic variants using WES data of 9 CBF AML patients. Figure 8 shows the number of variants that are shared between diagnosis and relapse in grey, variants that are unique to diagnosis in blue and variants unique to relapse in red. If a variant has been

called at diagnosis or relapse according to the filtering criteria listed in section 3.1, the variant in the other sample has been added regardless of filtering criteria. In case of patient 06 all variants are unique to diagnosis, because there was no relapse sample available for this patient (see Table 4).

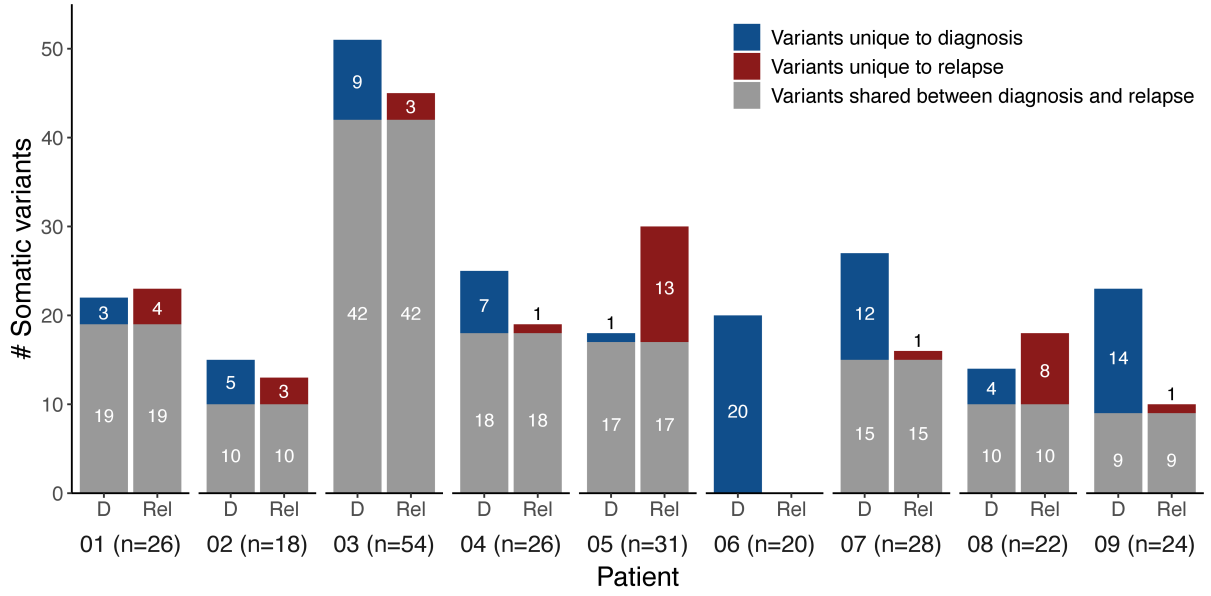

Figure 8: Overview of somatic variants called in whole-exome sequencing data. Bar plots show for each patient the number of detected mutations that are shared between diagnosis and relapse (grey) and those that are either unique to diagnosis (blue) or unique to relapse (red). Numbers in brackets are total number of somatic variants detected in this patient. For patient 06 all detected variants were unique to diagnosis, because there was no relapse sample available.

According to Christen *et al.,* [14] patient 09 is the only patient in this cohort that can be classified as genetically unstable with <40% of shared somatic variants between diagnosis and relapse, whereas the others can be classified as genetically stable with ≥40% of shared variants. For patients 01, 03 and 04 more than 60% of the identified somatic variants are shared between diagnosis and relapse. The mean number of somatic variants at diagnosis and relapse are comparable with 22.8 and (excluding patient 06) 21.8, respectively. Here, patient 02 with 18 unique variants has the lowest number and patient 03 with 54 unique variants has the highest number of detected somatic variants at both timepoints. It has to be noted that alignments of patient 03 did show a poor quality when inspected with IGV [110]. The mean VAF of gene mutations located on autosomes is with 25.0% at diagnosis higher than the mean VAF at relapse (excluding patient 06) with 13.6%. It is recommended by the onkopedia guidelines from the Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e.V. (DGHO) to examine patients every 1-3 months within the first two years and every 3-6 months for years 3-5 after reaching complete remission to detect a relapse at the earliest possible time [126]. Therefore, theVAF should be lower in relapse samples as observed in this cohort.

In Section 6.5, I present for each patient detected somatic variants and mutational dynamics in detail.

6.2 Somatic copy-number alterations

I was able to analyze 8 patients for copy-number alterations with the pipeline described in Section 3.4. The complete remission sample or T-cell at diagnosis (*i.e.*, patient 07 and 08) was used as a reference to call somatic copy-number alterations at diagnosis or relapse. For both samples of patient 03 and the relapse sample of patient 01 the quality was not sufficient for copy-number analysis. For patients 02, 04 and 06 no copy-number changes were detected using this method. In case of patient 02 and 06 this is consistent with clinical data from conventional G-banding as listed in Table 3, but in case of patient 04 the resolution of this method might not be sufficient enough to call copy-number alterations in only a small fraction (13 of 49 metaphases ~ 27% of cells) of a sample.

Figure 9 shows copy-number plots for all patients with at least one detected copy-number alteration (*i.e.*, patients 01, 05, 07, 08 and 09). Plots show the estimated purity, ploidy and copy-numbers for major allele in yellow and minor allele in blue within every autosomal chromosome for every analyzed sample. If there are no copy-number alterations in a region meaning that each allele has a copy-number of one, the section is highlighted in green. Results presented have been manually filtered.

At diagnosis of patient 01 (Figure 9a), I detected an amplification of chromosome 8 matching the diagnostic karyotype information for this patient listed in Table 3. Figure 9b shows both tumor samples of patient 05 with a deletion on chromosome 7 (chr7q34-q36.1) that was only subclonally present at diagnosis and becomes more dominant at relapse. Additionally, I detected a uniparental disomy (UPD) on the q-arm of chromosome 19 present in both tumor samples. A UPD is a copy-neutral loss of heterozygosity (LOH) meaning that one allele is absent, and the remaining allele has two copies. In patient 07 (Figure 9c), I identified a deletion on chr7 (7q34-q36.3) at diagnosis and relapse and an amplification on chr9 (9q34.3) and deletion on chromosome 19q13 at relapse. For patient 08 (Figure 9d) I detected an amplification of chromosome 22 only present at diagnosis. This is according to the G-banding results of patient 08 with 47 chromosomes in total and additional chromosome 22 (+22) as stated in Table 3. In the diagnosis sample of patient 09, I detected a UPD on the p-arm of chromosome 17 that was lost at relapse (see Figure 9e).

Detailed information on somatic copy-number alterations and their clinical relevance are presented in Section 6.5.

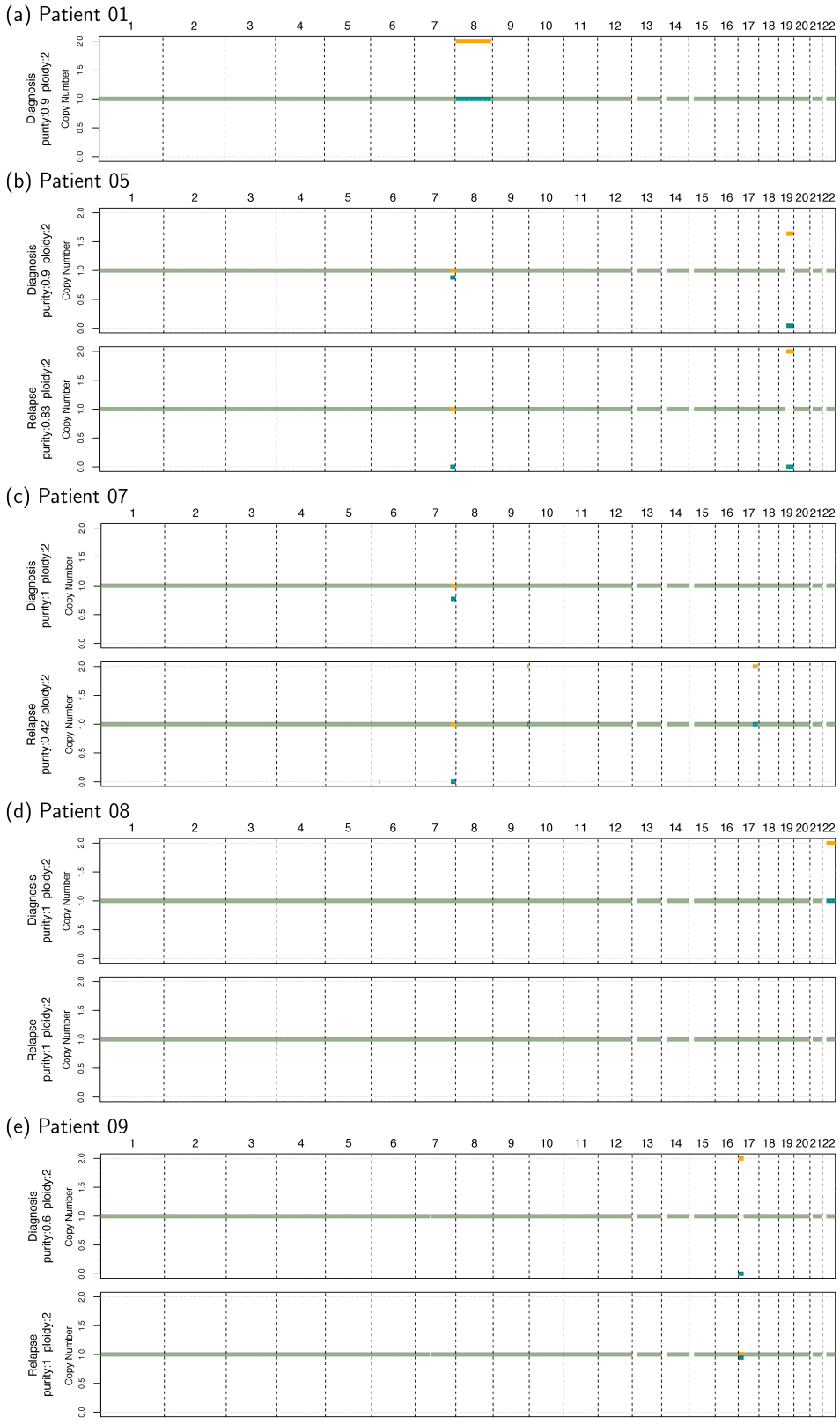

Figure 9: Overview somatic copy-number alterations. Genome plots of patients 01 (a), 05 (b), 07 (c), 08 (d) and 09 (e) for each analyzed sample purity, ploidy and copy-numbers of the major (yellow) and minor (blue) allele within every autosomal chromosome. If both alleles have a copy-number of 1 meaning that there is no copy-number alteration this section is highlighted in green. An amplification can be detected if the minor allele (blue) has a copy-number of 1 and the major allele (yellow) has a copy-number >1. A deletion is present if the major allele (yellow) has a copy-number of 1 and the minor allele (blue) a copy-number <1. A uniparental disomy is present if the major allele (yellow) has a copy-number of 2 and the minor allele (blue) a copy-number of 0.

6.3 CHIP variants

Using the 45 gene CHIP gene panel and deep sequencing, I detected 27 somatic variants in 6 of the 9 CBF AML patients that I did not discover previously using WES data. Table 5 lists all the variants that I manually filtered using IGV (v2.11.6) [110] and have been selected to be included in the single-cell panel. Due to the higher coverage and error-corrected reads, I was able to call variants with a VAF cut-off of 2% instead of 4% as in WES data. In patient 03 somatic variant *SF3B1* p.E903V with a VAF of 12.5% at diagnosis has the highest detected VAF. In patient 03, I detected most somatic variants (21/27) as for WES data (n=54) shown in Figure 8. For patients 01 and 05 two and for patients 04, 06 and 07 only one additional somatic variant was identified with the CHIP panel.

These variants are included in the detailed description of somatic variants in each patient in Section 6.5.

Table 5: Overview somatic variants called in targeted sequencing data. Detected variants are listed with gene and protein change, chromosome (Chr), start position, wild-type allele (Ref), variant allele (Alt) and variant allele frequency (VAF) in percent at diagnosis (D) and relapse (Rel). In total, I detected 28 somatic variants that were not identified using whole-exome sequencing data. Variants are grouped by patients and sorted alphabetically.

Patient	Variant	Chr	Start	Ref	Alt	VAF_D [%]	VAF_{Rel} [%]
01	*FLT3* p.D835Y	13	28592642	C	A	3.5	
01	*IDH2* p.R18P	15	90645570	C	G	9.4	
03	*CBL* p.S100T	11	119103260	T	A	5.0	
03	*CEBPA* p.P38R	19	33792851	G	C	4.4	4.9
03	*EZH2* p.A296V	7	148515190	G	A	7.3	
03	*EZH2* p.W543X	7	148511106	C	T	7.4	
03	*IDH1* p.N53H	2	209113350	T	G		4.2
03	*KIT* p.Y570S	4	55593643	A	C	4.8	3.8
03	*SETBP1* p.E903V	18	42532013	A	T	2.6	4.3
03	*SETBP1* p.P198S	18	42529897	C	T	5.0	2.7
03	*SF3B1* p.F632L	2	198267461	AAA	TAG	12.5	
03	*SF3B1* p.R630K	2	198267468	CTA	TTT	8.0	
03	*STAG2* p.A638E	X	123197789	C	A	5.4	
03	*STAG2* p.L574X	X	123196834	T	G	3.9	
03	*STAG2* p.L668P	X	123197879	T	C		7.4
03	*TET2* p.I19L	4	106155154	A	T	3.8	
03	*TET2* p.P1941A	4	106197488	C	G	3.9	
03	*TET2* p.P1956A	4	106197533	C	G		5.8
03	*TET2* p.Q1466K	4	106193934	C	A		5.0
03	*TET2* p.Q943X	4	106157926	C	T	5.6	2.0
03	*TET2* p.Y1255X	4	106164897	C	A	4.9	
03	*U2AF1* p.I10T	21	44524443	A	G	4.1	
03	*WT1* p.A154G	11	32456446	G	C		2.8
04	*BCORL1* p.A64T	X	129146938	G	A	2.5	
05	*CEBPA* p.A16T	19	33792918	C	T	3.4	
05	*RAD21* p.E157X	8	117870603	C	A		2.6
06	*BCORL1* p.R609X	X	129148573	C	T	6.5	
07	*TET2* p.N903Tfs*18	4	106157804	GA	G		3.5

6.4 Fusion gene breakpoints

I was able to detect CBF AML specific gene fusions (*i.e.*, *CBFB-MYH11* and *RUNX1-RUNX1T1*) using Nanopore sequencing data for every patient at diagnosis as described in

Section 5. Table 6 lists the identified gene fusions including information on the 5’ and 3’ location as well as the number of supporting reads. Additional to the location of breakpoints NanoFG [124] provides the FASTA sequence ±400 bp around each breakpoint which were used to establish the targeted single-cell panel. For patients 01, 05, 06 and 08 two breakpoints within less than 50 bp in the opposite direction have been identified.

Table 6: Overview breakpoints in core-binding factor genes. Detected gene fusions with fusion type, 5’ and 3’ breakpoints and positions are listed. Number of supporting reads are given as a fraction of total reads. In case of patient 02, 03 and 09 one breakpoint has been identified and for the other patients two breakpoints were identified. (SR = supporting reads)

Patient	Fusion type	5’ Gene	5’ Breakpoint	5’ Positon	3’ Gene	3’ Breakpoint	3’ Positon	SR
01	intron-intron	*RUNX1*	intron 5-6	21:36217048	*RUNX1T1*	intron 1-2	8:93081539	4/14
01	intron-intron	*RUNX1T1*	intron 1-2	8:93081575	*RUNX1*	intron 5-6	21:36217038	2/12
02	intron-exon	*CBFB*	intron 5-6	16:67123598	*MYH11*	exon 33-34	16:15815316	3/8
03	intron-intron	*CBFB*	intron 5-6	16:67130917	*MYH11*	intron 28-29	16:15826021	5/9
04	intron-exon	*CBFB*	intron 5-6	16:67129470	*MYH11*	exon 33-34	16:15815334	5/18
04	exon-intron	*MYH11*	exon 33-34	16:15815335	*CBFB*	intron 5-6	16:67129474	7/20
05	intron-exon	*CBFB*	intron 5-6	16:67129212	*MYH11*	exon 33-34	16:15815414	3/10
05	exon-intron	*MYH11*	exon 33-34	16:15815425	*CBFB*	intron 5-6	16:67129214	2/9
06	intron-exon	*CBFB*	intron 5-6	16:67125015	*MYH11*	exon 33-34	16:15815356	3/6
06	exon-intron	*MYH11*	exon 33-34	16:15815357	*CBFB*	intron 5-6	16:67125018	2/5
07	intron-exon	*CBFB*	intron 5-6	16:67123190	*MYH11*	exon 33-34	16:15815338	2/9
08	intron-intron	*MYH11*	intron 33-34	16:15815187	*CBFB*	intron 5-6	16:67116380	4/7
08	intron-intron	*CBFB*	intron 5-6	16:67116362	*MYH11*	intron 33-34	16:15815187	3/6
09	intron-intron	*RUNX1*	intron 5-6	21:36227306	*RUNX1T1*	intron 1-2	8:93053289	3/15

Figure 10 visualizes the intron-exon *CBFB-MYH11* gene fusion detected in patient 02 as example. As stated in Table 6, the 5’ breakpoint in *CBFB* is located intronic between exon 5 and 6 and the 3’ breakpoint in *MYH11* is located on exon 33.

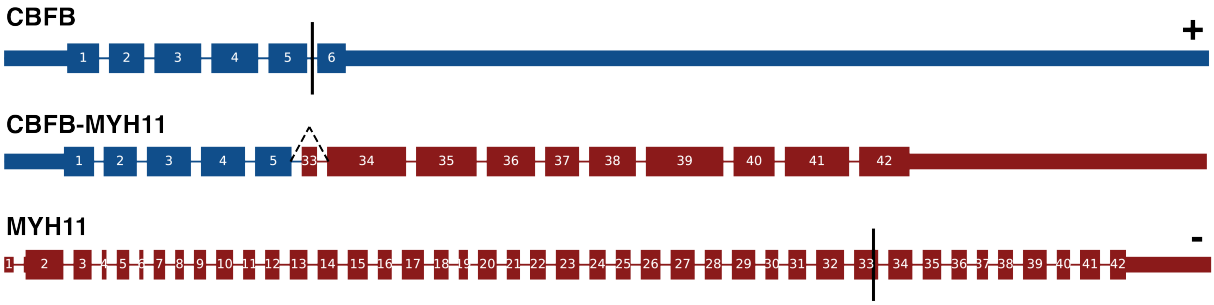

Figure 10: *CBFB-MYH11* gene fusion in patient 02. Gene fusion with *CBFB* (blue) on the positive strand harboring the intronic 5’ breakpoint and *MYH11* (red) on the negative strand harboring the exonic 3’ breakpoint at the bottom. Breakpoints are indicted by black lines. The resulting fusion gene in this patient with a length of 612 amino acids (aa) is visualized in the center.

6.5 Mutational dynamics and patients in detail

In this section, I present detailed information on somatic variants and copy-numbers for patients in this thesis. The information shown here was used for establishing the targeted single-cell panel.

6.5.1 Patient 01

I identified 28 somatic variants in patient 01 with t(8;21) CBF AML using WES and targeted sequencing data as listed in Table 7. *FLT3* p.D835Y and *IDH2* p.R18P have been detected using the targeted sequencing approach. Somatic variants in *FLT3*, which is part of the RTK family, can be found in approximately 30% of all AML patients and the most common SNV in *FLT3* is located on the activation loop D835 residue, as in this patient [127]. This patient harbors a mutation in epigenetic modifier *IDH2*, which has been shown to be associated with t(8;21) AML [5,15].

Table 7: Identified somatic variants in patient 01. In total I identified 28 somatic variants in patient 01 using whole-exome and targeted sequencing data. List is sorted in ascending order for gene names and contains all variants that were manually filtered and selected to be included in the single-cell panel. Variants in known AML driver genes are highlighted. Variant allele frequencies (VAFs) are shown in percent. † Somatic variants detected using targeted sequencing data.

Variant	Chr	Position	Ref	Alt	$VAF_{Diagnosis}$	$VAF_{Relapse}$
ABCC8 p.N500S	11	17464398	T	C	49.3	4.0
ADGB p.R1508Q	6	147109732	G	A	50.0	9.4
AZGP1 p.S111G	7	99569375	T	C	4.5	
CDH20 p.S794W	18	59221903	C	G	39.4	7.4
CDON p.V739I	11	125867249	C	T	46.4	1.8
CYP2A13 p.V80M	19	41594891	G	A	3.1	6.2
CYP2A13 p.D108N	19	41594975	G	A	4.8	5.2
FAT3 p.A4088V	11	92614032	C	T	44.9	4.3
***FLT3* p.D835Y**†	13	28592642	C	A	3.7	
FOXN4 p.A84V	12	109725967	G	A		5.4
GBP1 p.E99K	1	89525903	C	T		6.0
***IDH2* p.R18P**†	15	90645570	C	G	9.4	
MST1 p.T294S	3	49723881	G	C	8.1	3.5
OR2AG1 p.M309I	11	6807195	G	A	4.3	
OR5H1 p.T167S	3	97852040	A	T	5.3	2.4
PCDHB8 p.N464K	5	140559007	C	A	39.2	6.5
***PIK3C2A* p.R167K**	11	17167410	C	T	38.8	12.8
RBM10 p.R109X	X	47034471	C	T	97.0	4.3
RPAP3 p.A195V	12	48075532	G	A		4.2
SNAPC4 p.R166X	9	139289306	G	A	37.5	2.9
SYF2 p.A26V	1	25558650	G	A		4.1
TAF5L p.G500E	1	229730315	C	T	42.2	5.5
TNFRSF1A p.S56X	12	6443283	G	T	37.1	5.6
TTLL2 p.D410V	6	167754617	A	T	39.6	4.2
UGT2B7 p.D275E	4	69964361	T	A	4.1	3.3
ZBTB17 p.T236M	1	16271309	G	A	35.0	2.8
ZNF213 p.R147Hfs*28	16	3188459	G	AT	39.1	
ZSWIM4 p.A918T	19	13941646	G	A	36.5	3.0

Figure 11a visualizes VAF changes from diagnosis to relapse for all listed somatic variants in Table 7. Variants are highlighted if they are unique to diagnosis (blue), unique to relapse (red) or if the gene is a known AML driver (yellow). Most of the variants (19/28 ~ 68%) detected are shared between diagnosis and relapse. Figure 11b shows the copy-numbers, the log read-depth ratio (LogR) and the B-allele frequency (BAF) for chromosomes 1 to 22 at diagnosis. Sex chromosomes have been excluded for copy-number calling as described in section 3.4. I detected an amplification of chromosome 8 (trisomy 8) visible with the increase in LogR and the shift in BAF, which is matching the karyotype of this patient (see Table 3). I was not able to analyze the relapse sample due to low quality and the diagnosis sample is also of low quality as seen by the noisy LogR plot. Jahn *et al.*, [5] have shown that trisomy 8 and *FLT3* somatic variants have a negative impact on patient outcomes. The purity for this sample has been estimated with 90%, which is also reflected with some VAFs close to 50% when assuming heterozygosity and a VAF of 97% for *RBM10* p.109X located on chromosome X in this male patient at diagnosis. The relapse sample consisted of 30-40% blasts as listed in Table 4, which matches with the highest VAF of approximately 13% (26% of cells harboring that somatic variant) when assuming heterozygosity.

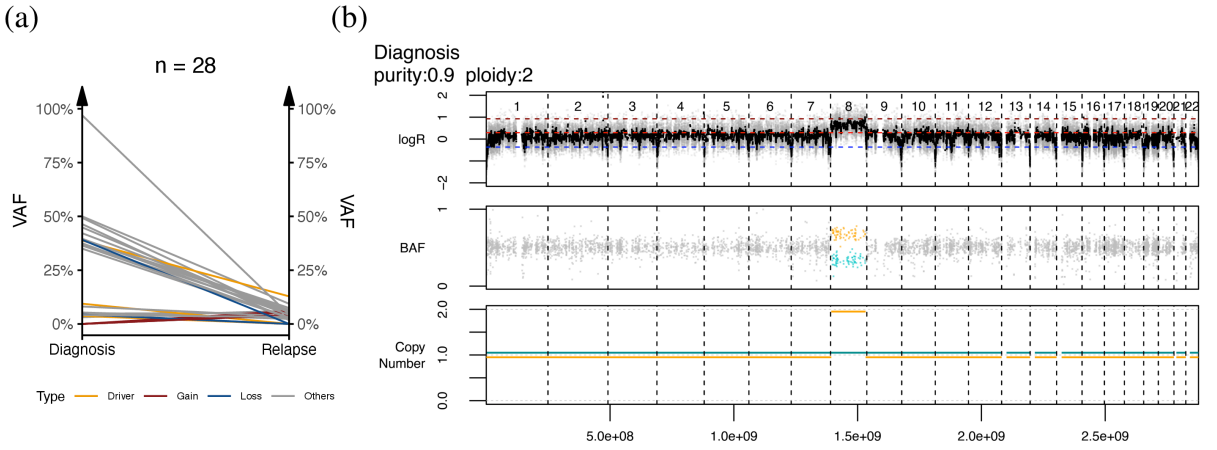

Figure 11: Mutational dynamics and copy-numbers of patient 01. (a) Changes in variant allele frequencies (VAFs) are shown from diagnosis on the left to relapse on the right. Variants are classified as gain if unique to diagnosis, as loss if unique to relapse, as driver if known AML driver gene and remaining as others. (b) Copy-numbers with log read-depth ratio (LogR) and B-allele frequency (BAF) for diagnosis of patient 01. Plot shows an amplification of chromosome 8. Relapse sample was not included due to poor quality.

6.5.2 Patient 02

I identified 18 somatic variants using whole-exome data in patient 02 as listed in Table 8. No additional mutations have been detected using the targeted sequencing approach. I detected in this male CBF AML patient with inv(16) somatic variants in *FLT3* and *KIT* which are genes involved in the RAS/RTK signaling pathway [14]. In detail, I detected *FLT3*-ITD, the most common genetic aberration in AML, with an AR of 15.5% and a VAF of 12.7% at diagnosis confirming clinical results (see Table 3) [128]. AR is defined as alternative read count divided by reference read count. It has been shown that somatic variants in *FLT3*, *KIT* and *RAS* of CBF AML patients with inv(16) lead to a survival/proliferation advantage of the cells and lead to poorer outcome for patients [129]. At relapse I detected a frame-shift mutation in *WT1* (*i.e.*, *WT1* p.R368Afs*5), which are known to be associated with *FLT3*-ITD in AML patients and show poor event-free survival (EFS) and OS in young patients (0-18 years) [130,131].

Table 8: Identified somatic variants in patient 02. In total I identified 18 somatic variants in patient 02 using whole-exome and targeted sequencing data. List is sorted in ascending order for gene names and contains all variants that were manually filtered and selected to be included in the single-cell panel. Variants in known AML driver genes are highlighted. Variant allele frequencies (VAFs) are shown in percent.

Variant	Chr	Position	Ref	Alt	$VAF_{Diagnosis}$	$VAF_{Relapse}$
ASAP1 p.N369S	8	131149259	T	C		9.3
***FLT3*-ITD**	13	28608218	C	*	12.7	
HIST1H2AG p.V115Rfs*23	6	27101186	T	TCAGGC	13.6	
IL1RAPL1 p.P241S	X	29686564	C	T		4.7
KIAA0556 p.L1554F	16	27789039	C	T	21.8	32.5
***KIT* p.D816V**	4	55599321	A	T	16.4	
LRRC74A p.P36S	14	77294651	C	T	4.8	
NEFH p.A314V	22	29879421	C	T	42.9	19.0
NPTX1 p.P343S	17	78445582	G	A	43.6	22.7
PKHD1L1 p.P3049L	8	110491836	C	T	40.4	17.2
PTPN20 p.G28E	10	48755057	C	T	7.1	5.5
RAB2B p.T16M	14	21943027	G	A	38.8	35.5
RASGRP2 p.E408G	11	64503087	T	C	40.8	25.5
RINT1 p.S304T	7	105205770	T	A	6.2	10.8
RNPC3 p.M257I	1	104083975	G	A	37.0	38.6
TPR p.D1453V	1	186307169	T	A	47.1	24.3
***WT1* p.R368Afs*5**	11	32417914	G	GC		9.3
ZNRF4 p.R5H	19	5455516	G	A	6.6	

*CCAAACTCTAAATTTTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCT
GAAATCAACGTAGAAGTACTCATTATCTGAGGAGCCGGTCACCTCATTTGGAA

Figure 12a shows the VAF changes of detected somatic variants of Table 8 from diagnosis on the left to relapse on the right with more than 60% (11/18) of them found in both samples. No copy-number changes have been detected for patient 02 at diagnosis and relapse as shown in Figure 12b. The karyotype of this patient at diagnosis listed in Table 3 confirms this with no additional detected cytogenetical abnormalities to inv(16).

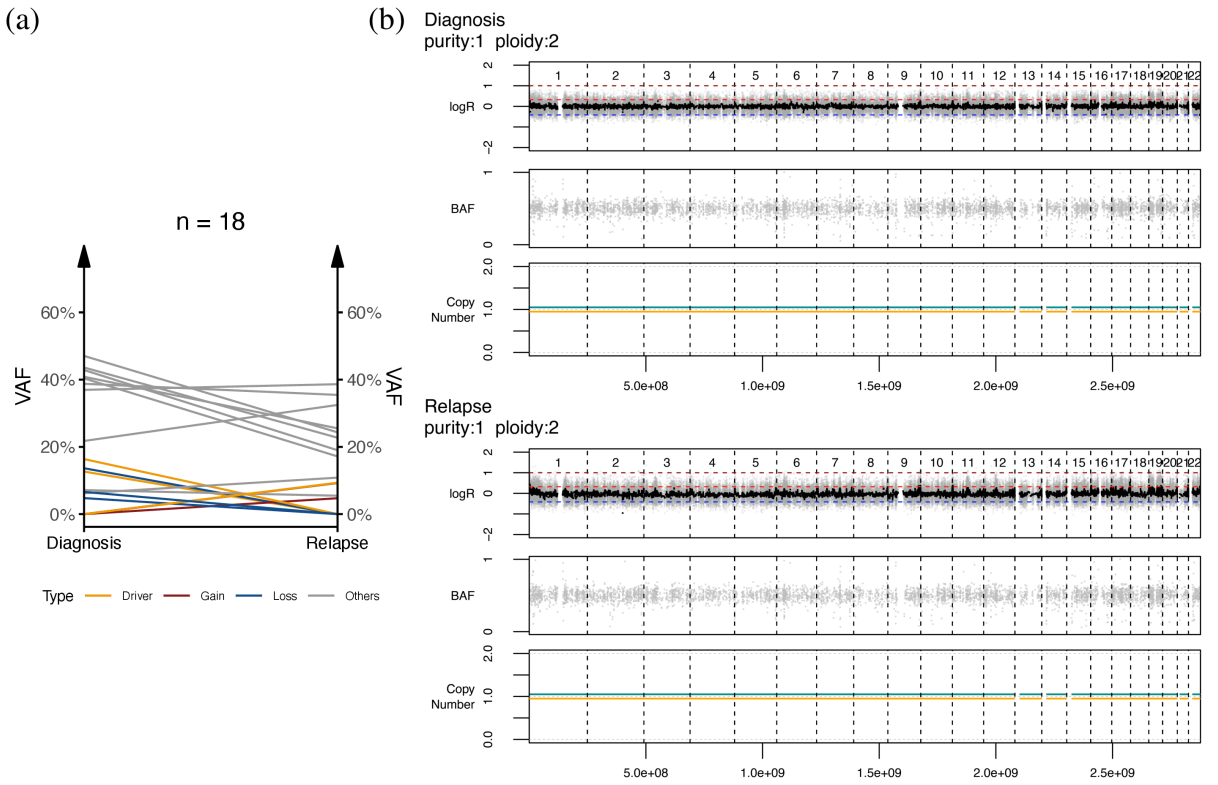

Figure 12: Mutational dynamics and copy-numbers of patient 02. (a) Changes in variant allele frequencies (VAFs) are shown from diagnosis on the left to relapse on the right. Variants are classified as gain if unique to diagnosis, as loss if unique to relapse, as driver if known AML driver gene and remaining as others. (b) Copy-numbers with log read-depth ratio (LogR) and B-allele frequency (BAF) for diagnosis and relapse of patient 02. Samples do not show any copy-number alterations.

6.5.3 Patient 03

I identified in total 75 somatic variants from whole-exome (n=54) and targeted sequencing data (n=21) in patient 05 as listed in Table 9. CBF AML specific somatic variants were found in genes involved in RTK/RAS signaling (*i.e.*, *CBL*, *KIT* and *KRAS*), chromatin modification (*i.e.*, *EZH2* and *SETD2*), methylation (*i.e.*, *IDH1* and *TET2*), cohesin complex (*i.e.*, *STAG2*), transcription (*i.e.*, *WT1*) and regulating splicing (*i.e.*, *SF3B1*) [5,132]. Furthermore somatic variants were found in tumor suppressors, such as *CHEK2*, *PTPRD* [133,134]. Multiple variants were detected in *KIT* (n=2), *SETBP1* (n=2), *SF3B1* (n=2), *STAG2* (n=3) and *TET2* (n=5).

Table 9: Identified somatic variants in patient 03. In total I identified 75 somatic variants at diagnosis (D) and relapse (Rel) in patient 03 using whole-exome and targeted sequencing data. List is sorted in ascending order for gene names and contains all variants that were manually filtered and selected to be included in the single-cell panel. Variants in known AML driver genes are highlighted. Variant allele frequencies (VAFs) are shown in percent. † Somatic variants detected using targeted sequencing data.

Variant	Chr	Position	Ref	Alt	VAF_D	VAF_{Rel}
ACSBG2 p.I250M	19	6177251	T	G	7.6	3.2
AHI1 p.T304K	6	135784283	G	T	13.9	2.9
***ANKRD26* p.D313Efs*5**	10	27366405	A	ATCGC GAACC	36.4	35.3
APELA p.I12F	4	165798517	A	T	7.2	1.8
***AR* p.R847K**	X	66942759	G	A	10.5	
ATP10D p.L1268F	4	47589086	A	T	4.7	0.8
AZGP1 p.N23T	7	99573576	T	G	5.6	4.6
AZGP1 p.P6S	7	99573628	G	A	11.4	4.5
BPGM p.R154L	7	134346720	G	T	10.0	2.6
***BRCA2* p.E2981K**	13	32953640	G	A	6.3	3.5
***CBL* p.S100T**†	11	119103260	T	A	5.0	
***CEBPA* p.P38R**†	19	33792851	G	C	4.4	4.9
***CHEK2* p.K306E**	22	29091840	TG	CA	11.1	
CLMN p.S689N	14	95669620	C	T	2.2	10.3
CXCR1 p.K197N	2	219029344	T	G		6.7
***EZH2* p.W543X**†	7	148511106	C	T	7.4	
***EZH2* p.A296V**†	7	148515190	G	A	7.3	
HIST1H4G p.K60fs	6	26247028	T	TAA	5.6	4.3
HSPA1L p.N283S	6	31778902	T	C	7.9	4.5
***IDH1* p.N53H**†	2	209113350	T	G		4.2
IFNA8 p.E60G	9	21409354	A	G	2.0	6.0
KCNQ3 p.D422N	8	133150208	C	T	51.7	41.0
KIAA0556 p.I1124S	16	27763064	T	G	47.4	37.7
***KIT* p.Y570S**†	4	55593643	A	C	4.8	3.8
***KIT* p.N822K**	4	55599340	T	A	44.6	39.5
***KRAS* p.G12C**	12	25398285	C	A	7.1	13.9
LAMB4 p.I1600K	7	107674672	A	T	8.6	4.2
LOC101059915 p.G58R	X	70887825	G	A	8.3	7.2
MAEL p.Y85F	1	166961944	A	T	9.6	5.4
MT1F splice-site	16	56692582	C	A	7.4	16.0
OR10G7 p.R233K	11	123909011	C	T	8.5	7.6
OR10G8 p.V28I	11	123900411	G	A	15.3	10.5
OR10G9 p.V298L	11	123894611	G	C	6.3	8.2
OR1M1 p.V40L	19	9204038	G	C		4.3
OR1M1 p.V69I	19	9204125	G	A	4.2	9.1
OR2A14 p.V203M	7	143826812	G	A	5.6	5.9
OR2A2 p.I67V	7	143806874	A	G	8.9	15.1
OR5T2 p.R161H	11	56000180	C	T	2.6	6.8
OR7E24 p.S69C	19	9361925	C	G	3.7	7.9
PBOV1 p.E57K	6	138539364	C	T	7.9	1.5
PKD1 p.A1368V	16	2161065	G	A	42.3	38.3
PPP1R12A p.X975K	12	80169711	A	T	5.2	13.0
PPP3CA p.S375R	4	101953512	A	T	8.8	
***PTPRD* p.I1298T**	9	8341102	A	G	40.0	61.3
***RNF216* p.R517_R518insL**	7	5760755	C	CTGA	30.2	26.9
SEPTIN7 p.L163H	7	35919516	T	A	10.0	
SERPINB13 p.V161F	18	61260187	G	T		11.5
***SETBP1* p.P198S**†	18	42529897	C	T	5.0	2.7
***SETBP1* p.E903V**†	18	42532013	A	T	2.6	4.3
***SETD2* p.S1817P**	3	47125689	A	G	3.8	
***SF3B1* p.F632L**	2	198267461	AAA	TAG	12.5	
***SF3B1* p.R630K**	2	198267468	CTA	TTT	8.0	
SLC24A5 p.L452F	15	48434399	C	T	7.7	
***STAG2* p.L574X**†	X	123196834	T	G	3.9	
***STAG2* p.A638E**†	X	123197789	C	A	5.4	
***STAG2* p.L668P**†	X	123197879	T	C		7.4
SYCP2 p.N969K	20	58455392	A	T	10.3	
SYNE1 p.E6971X	6	152551753	C	A	7.7	
TBC1D19 p.E73G	4	26640436	A	G	10.6	
***TET2* p.I19L**†	4	106155154	A	T	3.8	
***TET2* p.Q943X**†	4	106157926	C	T	5.6	2.0
***TET2* p.Y1255X**†	4	106164897	C	A	4.9	
***TET2* p.Q1466K**†	4	106193934	C	A		5.0
***TET2* p.P1941A**†	4	106197488	C	G	3.9	
***TET2* p.P1956A**†	4	106197533	C	G		5.8
THSD7B p.P554S	2	137928445	C	T	34.0	39.5
TMEM173 p.S236Y	5	138855922	G	T	34.9	40.0
TNFRSF10D p.T157K	8	23004486	G	T	2.1	5.4
TNFRSF14 p.S186R	1	2493118	C	G	11.1	7.6
U2AF1 p.I10T†	21	44524443	A	G	4.1	
USP19 p.V630M	3	49152252	C	T	42.1	34.8
***USP9X* p.V853Pfs*13**	X	41027388	T	TCGG CCCC	78.8	64.5
***WT1* p.A154G**†	11	32456446	G	C		2.8
ZBTB45 p.H508R	19	59025434	T	C	25.0	41.4
ZNF416 splice-site	19	58087302	C	T	3.2	6.6

Figure 13 shows the changes in VAFs from diagnosis to relapse for detected variants listed in Table 9. More than 60% (47/75) of the variants are shared between both tumor samples. INDEL *USP9X* p.V853Pfs*13 is the somatic variant with the highest VAF at diagnosis and relapse with 78.8% and 64.5%, respectively. For this patient I was not able to perform copy-number analysis due to the poor sample quality.

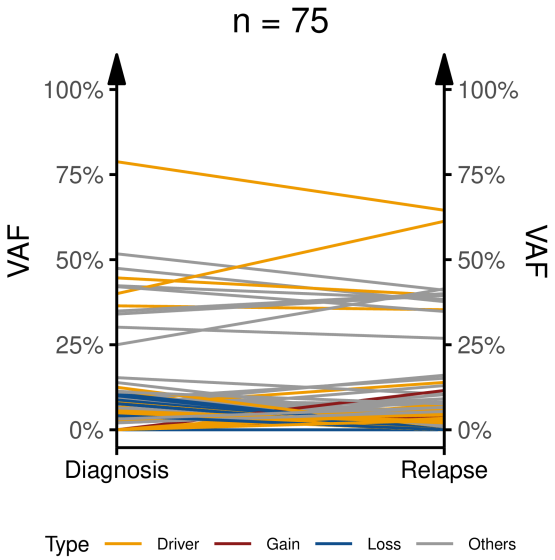

Figure 13: Mutational dynamics of patient 03. Changes in variant allele frequencies (VAFs) are shown from diagnosis on the left to relapse on the right. Variants are classified as gain if unique to diagnosis, as loss if unique to relapse, as driver if known AML driver gene and remaining as others.

6.5.4 Patient 04

I detected 27 somatic variants in patient 04 using whole-exome and targeted sequencing data as listed in Table 10. I identified a variant in *BCORL1* (*i.e., BCORL1 p.*A64T) using the error-corrected targeted sequencing data that is encoding for a transcriptional corepressor and has been associated with inv(16) CBF AML [5,135]. *SRCAP* is an epigenetic regulator and is involved in DNA damage repair [27]. High levels of *CHAF1B*, which is involved in CEBPA-mediated differentiation of leukemic cells, are associated with poor prognosis in leukemia patients [136]. I detected two variants in *NF1* (*i.e.*, *NF1* p.R1306X a stop-gain and *NF1* p.I679Dfs*21 a frame-shift variant), which is involved in RTK/RAS signaling, and it has been shown that loss of NF1 elevates RAS-MAPK signaling driving the development of AML .

Table 10: Identified somatic variants in patient 04. In total I identified 27 somatic variants in patient 04 using whole-exome and targeted sequencing data. List is sorted in ascending order for gene names and contains all variants that were manually filtered and selected to be included in the single-cell panel. Variants in known AML driver genes are highlighted. Variant allele frequencies (VAFs) are shown in percent. † Somatic variants detected using targeted sequencing data.

Variant	Chr	Position	Ref	Alt	$VAF_{Diagnosis}$	$VAF_{Relapse}$
ALDH5A1 p.C454Y	6	24533837	G	A	23.1	6.8
ALPL p.R59H	1	21889712	G	A	32.6	3.8
***BCORL1* p.A64T**†	X	129146938	G	A	2.5	
***CHAF1B* p.G161X**	21	37766948	G	T	48.8	7.2
CPSF1 p.Y1164D	8	145619936	A	C	49.1	7.1
CRB1 p.P309T	1	197316546	C	A	42.3	7.1
DENND2C p.G774R	1	115130514	C	T	46.8	8.7
EPHA5 p.Y961C	4	66197754	T	C	46.4	7.3
FUT6 p.P298S	19	5831687	G	A	27.5	3.2
GEMIN7 p.A100T	19	45593670	G	A	4.2	
MALRD1 p.R1588X	10	19678455	C	T	4.6	
***NF1* p.I679Dfs*21**	17	29553477	A	AC	19.5	
***NF1* p.R1306X**	17	29562981	C	T	7.4	
NR3C2 p.M778I	4	149035369	C	T	37.6	6.8
NRXN1 p.R1024Q	2	50699498	C	T	15.8	
PARP11 p.R84X	12	3931094	G	A		4.9
PRDM13 p.L156Q	6	100060978	T	A	0.4	4.3
RGPD8 p.D1388Y	2	113146360	C	A	3.9	4.4
RYR3 p.D2013N	15	33988595	G	A	7.3	
SCUBE3 p.C671S	6	35211476	G	C	41.3	4.2
***SRCAP* p.G2276Afs*7**	16	30747614	G	GATGC	28.3	7.8
THSD7B p.T307I	2	137814770	C	T	38.0	6.4
TMEM128 splice-site	4	4242207	C	T	55.7	4.7
WASHC2C p.D285H	10	46248044	G	C	2.4	6.6
WDR17 p.R640H	4	177070979	G	A	19.5	
WDR6 p.R1080C	3	49052671	C	T	18.1	4.2
ZNF587B p.F207I	19	58352661	T	A	4.3	4.4

Figure 14a shows the VAF changes from diagnosis on the left to relapse on the right for listed somatic variants in Table 10. The diagnosis sample contains 70% blasts and the relapse sample only 17% as listed in Table 4, which is reflected by the decreasing VAFs for the majority of mutations in the relapse sample (14/18 variants that were identified at both timepoints).

Despite the small fraction of blasts in the relapse sample, 70% of all somatic variants (18/27) are shared between diagnosis and relapse. No copy-number changes have been detected in this patient as shown in Figure 14b. In contrast, conventional G-banding show additional chromosomal abnormalities such as an amplification of chromosome 22 in 1 out of 49 metaphases and amplification of chromosomes 13,14 and 22 in 13 out of 49 metaphases. With approximately only 27% of all cells in the diagnosis sample harboring those chromosomal abnormalities they might be below the detection limit of this method.

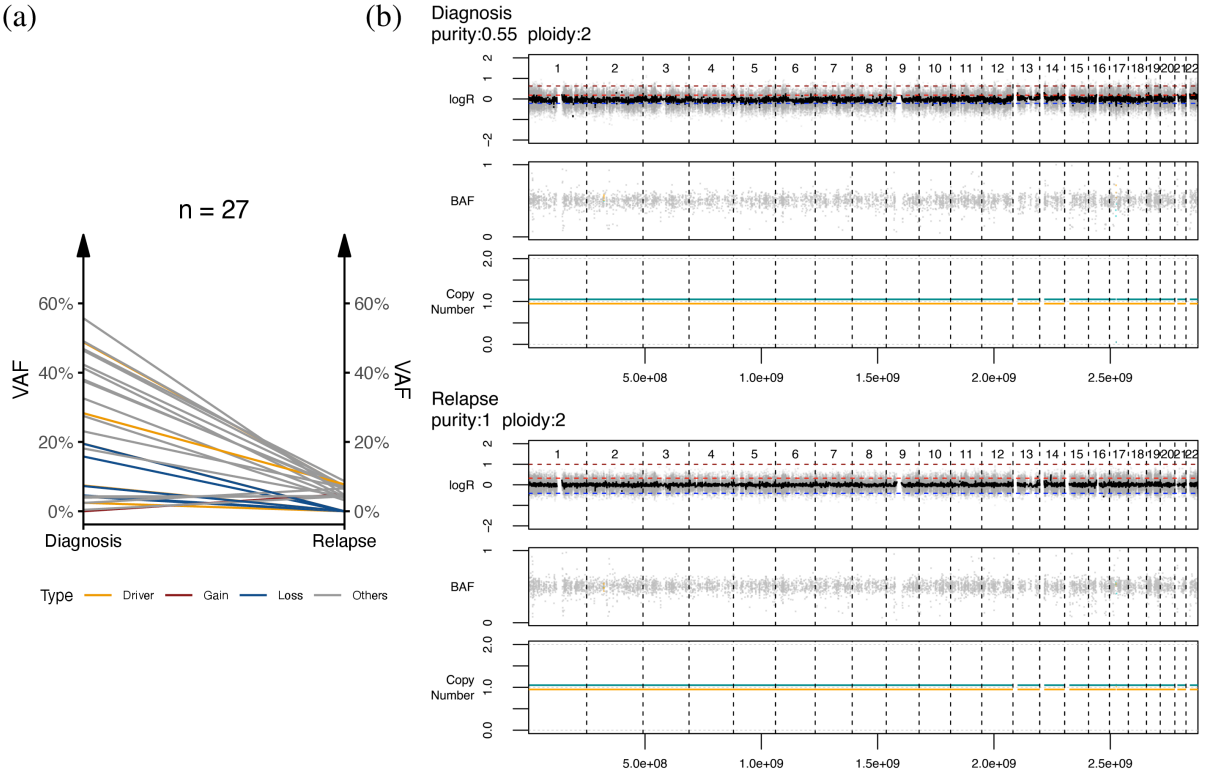

Figure 14: Mutational dynamics and copy-numbers of patient 04. (a) Changes in variant allele frequencies (VAFs) are shown from diagnosis on the left to relapse on the right. Variants are classified as gain if unique to diagnosis, as loss if unique to relapse, as driver if known AML driver gene and remaining as others. (b) Copy-numbers with log read-depth ratio (LogR) and B-allele frequency (BAF) for diagnosis and relapse of patient 04. Samples do not show any copy-number alterations.

6.5.5 Patient 05

I identified 33 somatic variants in patient 05 as listed in Table 11 using whole-exome and targeted sequencing data. *CEBPA* p.A16T and *RAD21* p.E157X with VAFs below 4% were detected using error-corrected targeted sequencing data enabling the detection of variants with a lower VAF. Nearly half of the identified somatic variants (16/33) in this patient are INDELS, specifically insertions with one exception.

Table 11: Identified somatic variants in patient 05. In total I identified 33 somatic variants in patient 05 using whole-exome and targeted sequencing data. List is sorted in ascending order for gene names and contains all variants that were manually filtered and selected to be included in the single-cell panel. Variants in known AML driver genes are highlighted. Variant allele frequencies (VAFs) are shown in percent. † Somatic variants detected using targeted sequencing data.

Variant	Chr	Position	Ref	Alt	$VAF_{Diagnosis}$	$VAF_{Relapse}$
AAK1 p.V639Sfs*3	2	69736454	C	CGTCACTGAGA		15.2
ADAMTSL1 splice-site	9	18661932	G	A	44.7	33.1
ADGRL4 p.C688F	1	79356849	C	A	35.6	32.0
ALDH8A1 p.F13Lfs*3	6	135271153	G	GTC		39.7
ALG1L p.S4L	3	125652507	G	A	5.4	
ASPSCR1 p.V275Pfs*11	17	79967033	G	CC		28.9
BTBD6 p.A432T	14	105716845	G	A	0.4	4.1
***CEBPA* p.A16T**†	19	33792918	C	T	3.4	
CORO1A p.R29_V30insGA	16	30196616	G	GGGGGGC		30.1
CRIM1 p.V312_S313insIV	2	36691736	G	GCATAGT	49.1	7.2
CRIM1 p.S313Ifs*66	2	36691742	T	TCATAGGGATGC		16.2
CX3CL1 p.T187M	16	57416565	C	T		4.3
DUSP3 p.H70_V71insGYD	17	41852220	A	ACATCATACC		30.4
FMN2 p.A374T	1	240256529	G	A		15.7
GAB4 p.A419T	22	17443634	C	T	38.2	21.7
ITPR1 p.V33Lfs*31	3	4562710	TG	T	39.5	33.9
KLHL8 p.N241H	4	88106447	T	G	47.3	43.8
MAP3K21 p.E50Q	1	233463922	G	C	41.2	13.8
MLLT6 p.P383S	17	36872730	C	T	44.6	34.0
***NBEAL2* p.V1424Rfs*46**	3	47042543	T	TGACGTGGCGGG		33.3
NFATC1 p.H288Rfs*11	18	77171128	C	CCGTCCCCG	13.4	23.5
***NFE2* p.P246_V247insKIVNLP**	12	54686542	C	CGGCAAGTTGACAATCTTG	42.9	36.2
NUTM2F p.R242Q	9	97084600	C	T	51.8	36.4
OR10G9 p.T13M	11	123893757	C	T	3.2	4.4
PAWR p.T267A	12	79990323	T	C	44.7	36.9
PTPRZ1 p.S165G	7	121616263	A	G	46.1	39.5
***RAD21* p.E157X**†	8	117870603	C	A		2.6
SGTA p.S26Afs*31	19	2768991	A	ACGAGAGGCCCCCGTGC	14.5	46.9
SIRT7 p.L347_R348insLPL	17	79870452	C	CGCAGGGGAA		16.9
SPATA13 p.V365Gfs*8	13	24798158	G	GGGGATCC		13.1
TMEM104 p.R101Lfs*28	17	72786382	T	TTCTCATCCTC		20.0
***WT1* p.S369Lfs*71**	11	32417910	G	GAGCGTACA		30.7
XKR4 p.V406I	8	56436049	G	A	50.2	41.5

Figure 15a shows the changes in VAF for somatic variants listed in Table 11 between diagnosis and relapse. For this patient only two variants were lost during disease and 14 variants were acquired at relapse, including somatic variants in AML drivers (*i.e.*, *NBEAL2* p.V1424Rfs*46, *RAD21* p.E157X and *WT1* p.S369Lfs*71). Figure 15b shows the copy-numbers including LogR and BAF for the diagnosis and relapse sample. Here, I detected a deletion on chromosome 7 (chr7q34-q36.1) confirming G-banding (see Table 3) visible by the drop in LogR and the shift in BAF that is already visible at diagnosis and is more dominant

at relapse. Additionally, this patient harbors a UPD on the q-arm of chromosome 19 visible by the shift in BAF and normal LogR.

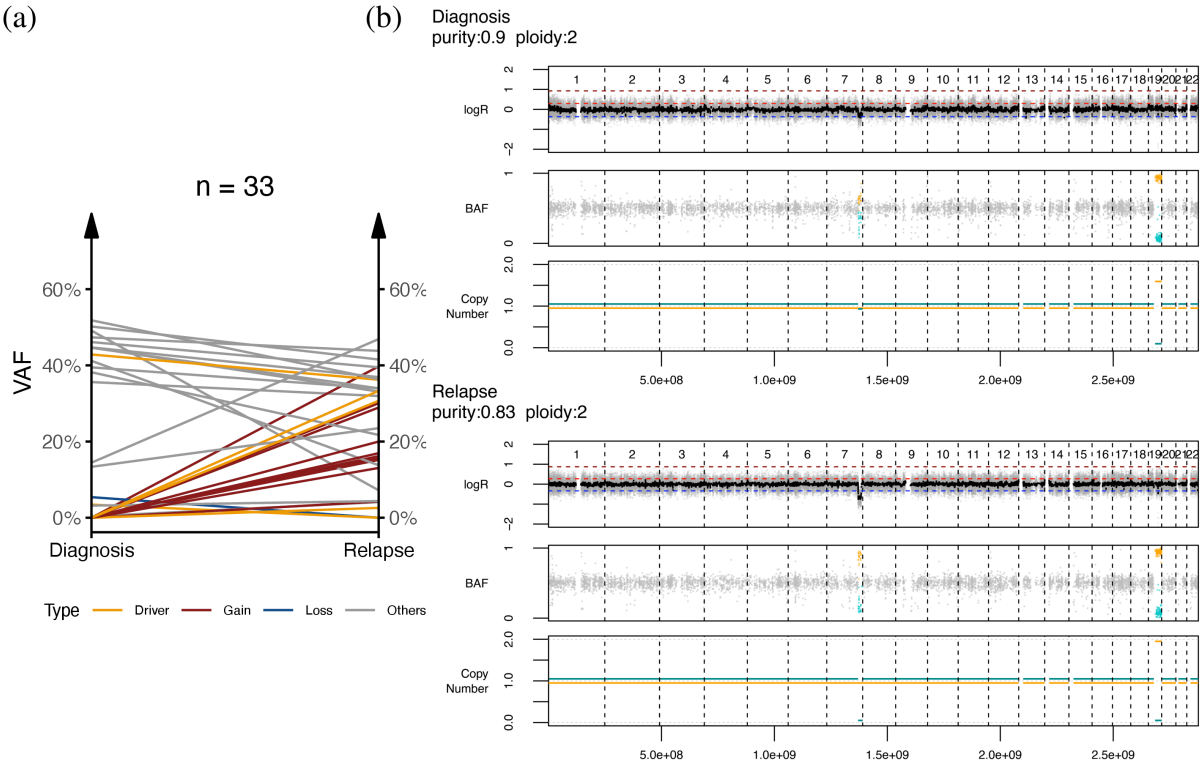

Figure 15: Mutational dynamics and copy-numbers of patient 05. (a) Changes in variant allele frequencies (VAFs) are shown from diagnosis on the left to relapse on the right. Variants are classified as gain if unique to diagnosis, as loss if unique to relapse, as driver if known AML driver gene and remaining as others. (b) Copy-numbers with log read-depth ratio (LogR) and B-allele frequency (BAF) for diagnosis and relapse of patient 05. Both samples show a deletion on chromosome 7, which is more dominant at relapse, and a uniparental disomy on the q-arm of chromosome 19.

6.5.6 Patient 06

I identified 21 somatic variants at diagnosis in inv(16) CBF AML patient 06 as listed in Table 12. For this patient no relapse sample was available. *BCORL1 p.*R609X was identified using the error-corrected targeted sequencing data. Mutations in BCORL1, which is a transcriptional corepressor, are more common in inv(16) patients [5,135]. Because from this patient only diagnosis and complete remission samples were available, I cannot show mutational dynamics.

Table 12: Identified somatic variants in patient 06. In total I identified 21 somatic variants at diagnosis for patient 06 using whole-exome and targeted sequencing data. List is sorted in ascending order for gene names and contains all variants that were manually filtered and selected to be included in the single-cell panel. Variants in known AML driver genes are highlighted. Variant allele frequencies (VAFs) are shown in percent. † Somatic variants detected using targeted sequencing data.

Variant	Chr	Position	Ref	Alt	$VAF_{Diagnosis}$
AZGP1 p.P6S	7	99573628	G	A	4.8
***BCORL1* p.R609X**†	X	129148573	C	T	3.8
C11orf80 p.G475C	11	66610493	G	T	45.8
FDXR p.A152T	17	72861876	C	T	50.9
LDHC p.G279R	11	18472510	G	A	43.6
LOC101059915 p.G89S	X	70887918	G	A	7.7
MBD3L2B p.R201Q	19	7051608	G	A	6.4
NCAPH2 p.R380W	22	50960444	C	T	42.0
NFE2L3 p.F260Lfs*35	7	26223343	CTTCT	C	40.8
***NRAS* p.Q61H**	1	115256528	T	A	44.7
OR10G8 p.V28I	11	123900411	G	A	11.4
OR1M1 p.L55F	19	9204083	C	T	4.6
OR1S1 p.I155T	11	57982680	T	C	5.2
OR1S2 p.K316E	11	57970708	T	C	4.8
OR8I2 p.T278M	11	55861616	C	T	8.7
PEG10 p.I171T	7	94293380	T	C	46.7
TERB2 p.G175D	15	45270687	G	A	6.6
TLE6 p.V236M	19	2989614	G	A	52.0
TREML4 p.L37F	6	41196497	C	T	6.2
USP48 p.E489X	1	22048240	C	A	41.0
ZNF236 p.N479S	18	74606987	A	G	32.0

Figure 16 shows copy-numbers including LogR and BAF for diagnosis of patient 06 with no detected copy-number alterations. The quality of this sample is poor as can be seen by the noisy LogR.

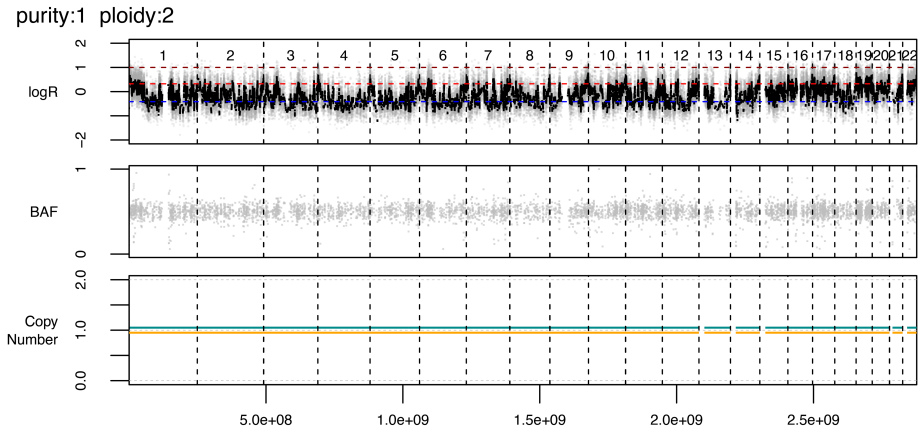

Figure 16: Copy-numbers of patient 06. Copy-numbers with log read-depth ratio (LogR) and B-allele frequency (BAF) for diagnosis of patient 06. This patient does not have copy-number abnormalities.

6.5.7 Patient 07

I detected in total 29 somatic variants at diagnosis and relapse in patient 07 using whole-exome sequencing and targeted sequencing data as listed in Table 13. *TET2* p.N903Tfs*18 with a VAF of 3.5% at relapse was identified using error-corrected targeted sequencing. This patient harbors multiple variants in genes involved in RTK/RAS signaling (i.e., *FLT3* p.D835Y, *KIT* p.D419del, *KIT* p.D816Y, *NF1* p.P2289Sfs*17 and *NRAS* p.G12A) and all of them have been detected exclusively at diagnosis [5].

Table 13: Identified somatic variants in patient 07. In total I identified 29 somatic variants in patient 07 using whole-exome and targeted sequencing data. List is sorted in ascending order for gene names and contains all variants that were manually filtered and selected to be included in the single-cell panel. Variants in known AML driver genes are highlighted. Variant allele frequencies (VAFs) are shown in percent. † Somatic variants detected using targeted sequencing data.

Variant	Chr	Position	Ref	Alt	$VAF_{Diagnosis}$	$VAF_{Relapse}$
ADSL p.E425X	22	40760965	G	T	12.1	7.8
ANK3 p.R610W	10	61827747	G	A	4.8	
CCDC74B p.R308H	2	130897150	C	T	6.7	
COL6A5 p.R1712L	3	130142716	G	T	1.5	8.3
CPZ p.S218R	4	8605893	C	A	40.8	27.6
DDI1 p.E322K	11	103908514	G	A	44.2	24.2
***DHX30* p.R864W**	3	47890128	C	T		11.1
DNAH11 p.P1626Lfs*24	7	21678611	CT	C	42.0	21.4
FAT1 p.V3388I	4	187530381	C	T	6.9	
FAT2 p.G1883R	5	150925041	C	T	2.0	10.6
FEM1A p.A401T	19	4793067	G	A	5.0	
***FLT3* p.D835Y**	13	28592642	C	A	4.1	
***KIT* p.D419del**	4	55589770	TACG	T	13.4	
***KIT* p.D816Y**	4	55599320	G	T	4.0	
MYO18B p.A408V	22	26165106	C	T	47.3	28.8
NBN p.T599Kfs*58	8	90965520	TG	T	31.2	15.4
NEXN p.R61X	1	78383884	C	T	42.2	14.0
***NF1* p.P2289Sfs*17**	17	29667528	G	TT	13.9	
***NRAS* p.G12A**	1	115258747	C	G	3.8	
***NSD1* p.N1969I**	5	176709479	A	T	40.4	26.5
PCDHGA2 p.P669S	5	140720543	C	T	10.8	2.8
PCLO p.P1151A	7	82595653	G	C	9.2	
PID1 p.H151Y	2	229890404	G	A	4.4	
POU4F2 p.T350M	4	147561779	C	T	40.7	17.4
SLC22A10 p.S476I	11	63072190	G	T	9.8	
SRRT p.R414Pfs*5	7	100482916	G	GC	34.4	23.3
***TET2* p.N903Tfs*18**†	4	106157804	GA	G		3.5
TEX15 p.D2855E	8	30695235	G	T	46.9	20.4
WDR81 p.R437C	17	1637249	C	T	1.4	15.2

Figure 17a shows the VAF changes of somatic variants listed in Table 13 between diagnosis and relapse of patient 07. For this patient, 12 variants are lost during disease development in this patient and only two variants are gained at relapse (*i.e.*, *DHX30* p.R864W and *TET2* p.N903Tfs*18). Despite the small difference in blasts for diagnosis and relapse (90% at diagnosis and 80% relapse, see Table 4) the mutational dynamics plot estimates a smaller amount of tumor cells at relapse. Figure 17b shows copy-numbers including LogR and BAF

for diagnosis and relapse. Here, I identified a deletion on chr7 (7q34-q36.3) visible at diagnosis and more dominant at relapse and an amplification on chromosome 17 present at relapse.

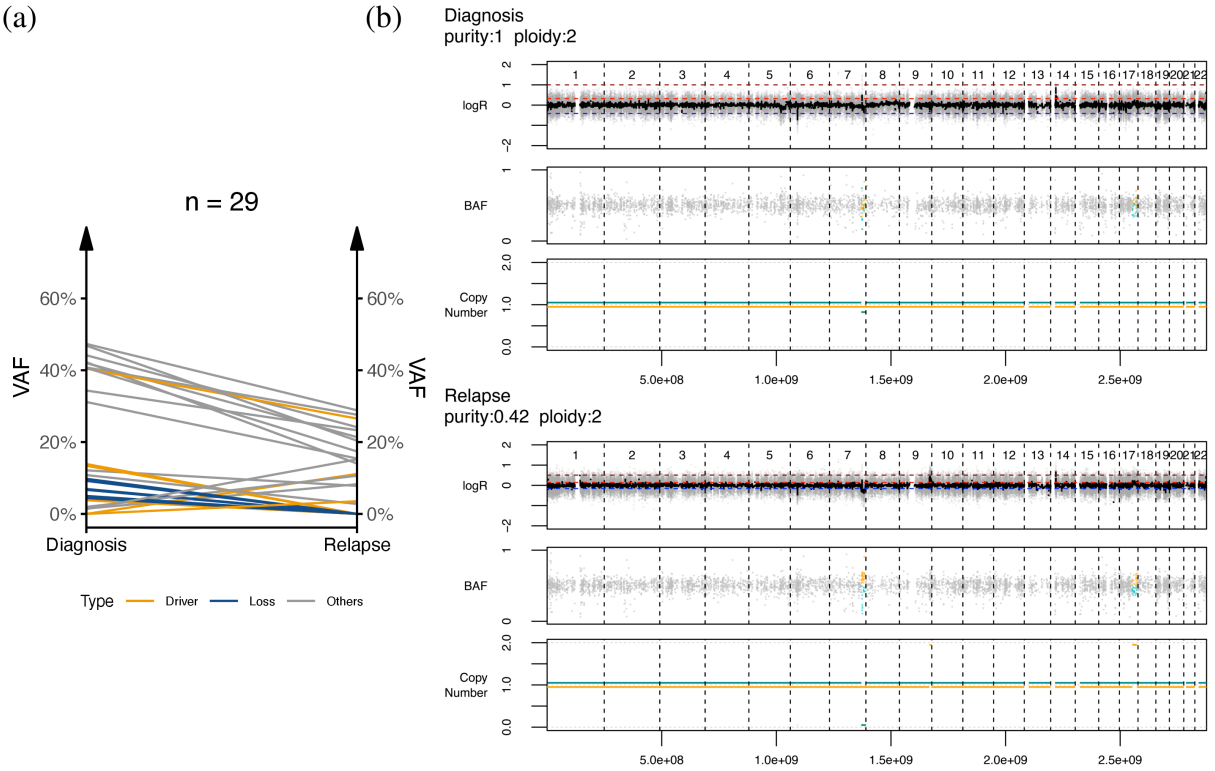

Figure 17: Mutational dynamics and copy-numbers of patient 07. (a) Changes in variant allele frequencies (VAFs) are shown from diagnosis on the left to relapse on the right. Variants are classified as gain if unique to diagnosis, as loss if unique to relapse, as driver if known AML driver gene and remaining as others. (b) Copy-numbers with log read-depth ratio (LogR) and B-allele frequency (BAF) for diagnosis and relapse of patient 07. This plot shows a deletion on chromosome 7 and amplification on chromosome 17 both dominant at relapse.

6.5.8 Patient 08

I identified 22 somatic variants in female inv(16) CBF AML patient 08 using whole-exome sequencing data (see Table 14). No additional variants have been detected using the targeted sequencing approach. Somatic variants were found in genes involved in RTK/RAS signaling (*i.e., FLT3* p.A680V), transcription (i.e., *WT1* p.D355Y) and the Fanconi pathway (*i.e.*, *FANCM* p.A166V) as well as in genes encoding for tumor suppressors (*i.e.*, *KMT2C* p.G908C) and epigenetic regulators (*i.e.*, *PHF6* p.G29X).

Table 14: Identified somatic variants in patient 08. In total I identified 22 somatic variants in patient 08 using whole-exome and targeted sequencing data. List is sorted in ascending order for gene names and contains all variants that were manually filtered and selected to be included in the single-cell panel. Variants in known AML driver genes are highlighted. Variant allele frequencies (VAFs) are shown in percent.

Variant	Chr	Position	Ref	Alt	$VAF_{Diagnosis}$	$VAF_{Relapse}$
ADAMTSL3 p.G611D	15	84581975	G	A	4.8	
ALKBH4 p.R179W	7	102098215	G	A		8.3
APOB p.R1815W	2	21234297	G	A		9.1
BOD1L2 p.L155P	18	54815007	T	C	5.5	6.2
COL5A1 p.E1571Rfs*53	9	137716447	G	GC	40.3	31.8
CPAMD8 p.R402G	19	17104288	G	C	5.1	1.5
***FANCM* p.A166V**	14	45605731	C	T		4.6
***FLT3* p.A680V**	13	28602329	G	A	28.1	
GATB p.E401K	4	152609912	C	T		7.3
HDX p.I369S	X	83695565	A	C		6.1
KLK9 p.R65H	19	51512445	C	T	42.6	35.2
***KMT2C* p.G908C**	7	151932949	C	A	11.5	10.0
LAMA2 p.G2629S	6	129807766	G	A	37.2	28.7
MAGT1 p.W169L	X	77112879	C	A	41.0	28.3
***PHF6* p.G29X**	X	133511732	G	T		36.8
POPDC2 p.R166H	3	119373455	C	T	32.3	
PTPN9 p.V292Wfs*37	15	75798109	AC	A	31.9	30.0
PTPRZ1 p.S204I	7	121616897	G	T	32.6	28.8
RGS9 p.A22V	17	63149547	C	T	39.0	28.0
SLC6A15 p.C90F	12	85277804	C	A		8.7
***WT1* p.D355Y**	11	32417953	C	A		8.7
XKRX p.R254C	X	100169917	G	A	8.2	

Figure 18a visualizes the VAF changes for somatic variants listed in Table 14 with only 45% (10/22) variants shared between diagnosis and relapse. Despite the difference in percentage of blasts for diagnosis (80%) and relapse (20-25%) as listed in Table 4, there is no real difference in VAFs between those two samples. Figure 18b shows copy-numbers with LogR and BAF for diagnosis and relapse of patient 08. I identified an amplification of chromosome 22 at diagnosis visible by the increase in LogR and shift in BAF. CBF AML patients with inv(16) and trisomy 22 show a very high RFS survival rate when compared to CBF AML patients without trisomy 22, in contrast this patient has a RFS of only 14 months (see Table 2 and Figure 6) [74]. At relapse the amplification of chromosome 22 is not detectable.

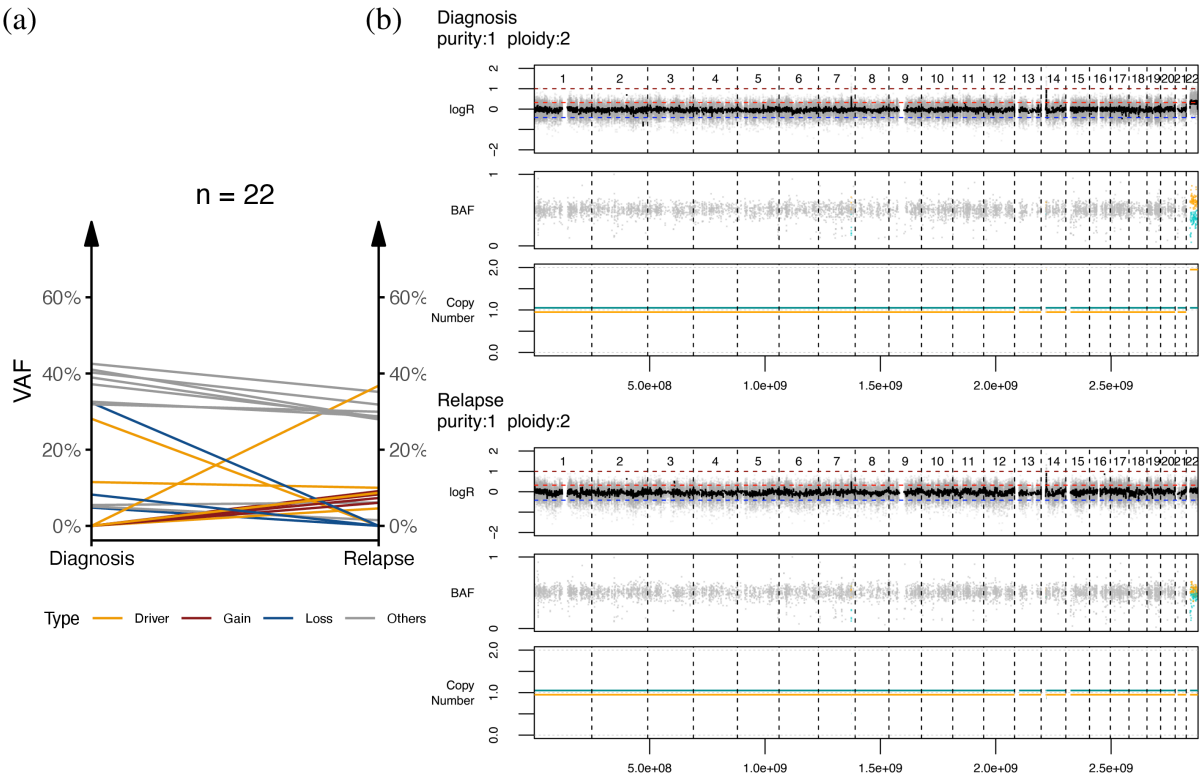

Figure 18: Mutational dynamics and copy-numbers of patient 08. (a) Changes in variant allele frequencies (VAFs) are shown from diagnosis on the left to relapse on the right. Variants are classified as gain if unique to diagnosis, as loss if unique to relapse, as driver if known AML driver gene and remaining as others. (b) Copy-numbers with log read-depth ratio (LogR) and B-allele frequency (BAF) for diagnosis and relapse of patient 08. Plot shows an amplification of chromosome 22 at diagnosis.

6.5.9 Patient 09

I identified 24 somatic variants at diagnosis and relapse by whole-exome sequencing data for female patient 09 as listed in Table 15. In detail, I detected somatic variants in genes involved in RTK/RAS signaling (*i.e.*, *KRAS* p.G13D) and in the cohesin complex (*i.e.*, *SMC3* p.L664Q) in this t(8;21) CBF AML patient . Mutations in genes encoding for members of the cohesin complex consisting of four core subunits (*i.e.*, SMC1A, SMC3, RAD21, and STAG protein) and are observed nearly exclusive for t(8;21) CBF AML .

Table 15: Identified somatic variants in patient 09. In total I identified 24 somatic variants in patient 09 using whole-exome and targeted sequencing data. List is sorted in ascending order for gene names and contains all variants that were manually filtered and selected to be included in the single-cell panel. Variants in known AML driver genes are highlighted. Variant allele frequencies (VAFs) are shown in percent.

Variant	Chr	Position	Ref	Alt	$VAF_{Diagnosis}$	$VAF_{Relapse}$
ANXA7 p.Y288C	10	75139895	T	C	26.4	0.9
ARV1 p.L192F	1	231131633	G	C	31.2	3.6
ATP10A p.E911del	15	25947088	ACCT	A	22.1	
CD207 p.S177N	2	71060812	C	T	32.5	1.4
COL6A3 p.P2252L	2	238245167	G	A	30.4	
CYP8B1 p.M204V	3	42916699	T	C	26.9	2.8
EFHD1 splice-site	2	233503196	T	-	25.5	
EIF2B4 p.R9S	2	27593157	G	T	20.7	
IFT46 p.Q294L	11	118415665	T	A	15.5	
***KRAS* p.G13D**	12	25398281	C	T	4.4	
KRT26 p.A296S	17	38926089	C	A	11.2	0.3
LAMB4 p.T571M	7	107720221	G	A	34.4	2.4
LRP1B p.G2237V	2	141457908	C	A	23.8	0.3
NWD1 p.W87X	19	16855293	G	A	22.3	
PAX7 p.A395V	1	19062154	C	T	25.0	
PHIP splice-site	6	79724935	T	C	31.4	4.0
SCN1B p.W183X	19	35524743	G	A	28.1	2.3
***SMC3* p.L664Q**	10	112356183	T	A	31.8	
SORL1 p.R1473X	11	121466379	C	T	30.5	
STAB1 p.A939V	3	52545694	C	T		4.2
STIM2 p.L512P	4	27019378	T	C	21.5	
TANC2 p.P1886L	17	61499000	C	T	7.3	
TTN p.E10394K	2	179536842	C	T	20.7	
USP22 p.A252_R253insGPS	17	20919146	T	TCGAAGGACC	54.3	

Figure 19a shows the VAF changes for patient 09 from diagnosis on the left to relapse on the right. For this patient only 9 out of 23 variants from diagnosis have been found at relapse and most of them with a VAF below 4%. Despite of 50% of blasts in the relapse sample as stated in Table 4 no variants with a VAF above 4.2% have been detected. Figure 19b shows the copy-numbers with LogR and BAF for diagnosis and relapse. At relapse I detected a UPD on chromosome 17 that is lost at relapse. *USP22* p.A252 R253insGPS is in the region of the UPD which can be seen by nearly double the VAF of the other variants in this patient, deriving from the LOH in this region.

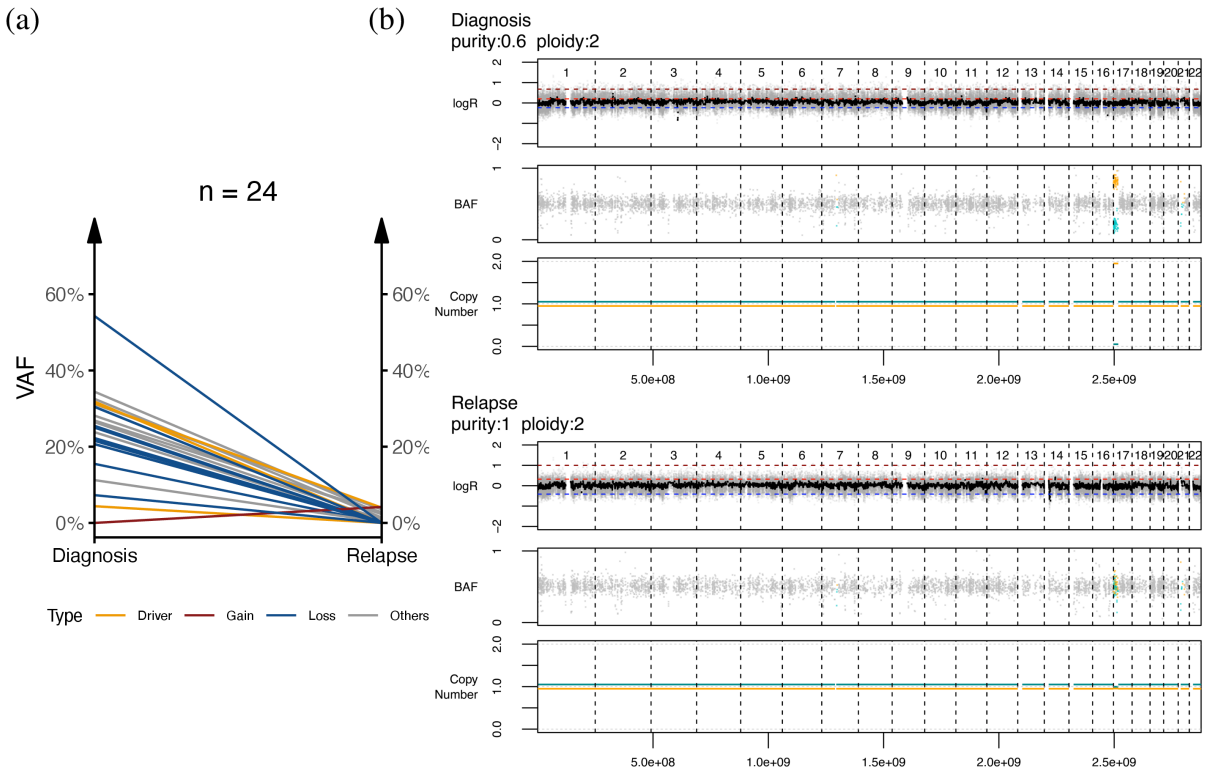

Figure 19: Mutational dynamics and copy-numbers of patient 09. (a) Changes in variant allele frequencies (VAFs) are shown from diagnosis on the left to relapse on the right. Variants are classified as gain if unique to diagnosis, as loss if unique to relapse, as driver if known AML driver gene and remaining as others. (b) Copy-numbers with log read-depth ratio (LogR) and B-allele frequency (BAF) for diagnosis and relapse of patient 09. Plot shows a uniparental disomy on chromosome 17.

IV Single-cell sequencing

In this chapter I present methods and results from the single-cell sequencing part of my thesis. Here, I show how I integrated bulk and single-cell data to reconstruct tumor phylogeny from CBF AML patients with somatic variants, fusion genes and copy-number alterations.

1 Custom targeted single-cell DNA-Seq panels

Three custom panels (see Table 16) were designed with a maximum size of 201 amplicons covering somatic variants, regions of copy-number alterations and fusion genes. Single-cell libraries were generated on the Mission Bio Tapestri platform using the Tapestri Single-Cell DNA Sequencing V2 kit (Mission Bio) and sequenced on the NovaSeq 6000 platform (Illumina; 300 cycles paired-end, 15% PhiX). Sequencing reads were processed using the Tapestri pipeline (Mission Bio, v2.0.2) with the respective panel and reference genome to obtain Loom files for each sample that are used for downstream analysis. For each panel a specific reference genome was created with the sequence of ±400 bp around of the patient specific breakpoints for each patient in that panel.

Table 16: Overview custom targeted single-cell DNA-Seq panels. The 9 patients were split on three custom panels ranging in size from 180 to 201 amplicons.

Panel	Patients	# Amplicons
CO-413	02, 04, 05, 08	180
CO-414	01, 03, 06	201
CO-415	07, 09	201

2 Identify variants and gene fusion in single-cells

I used an adapted version of the preprocessing script from COMPASS (v.1.1) [64] to obtain information on variants, the CBF gene fusion and the cell barcode for each selected cell in a sample (see section I4 for details on how cells are selected). Additionally, I extended the list of input parameters with `--use_whitelist_only` to obtain only variants provided by the whitelist and `--use_fusion` to obtain gene fusion information.

The first step in the preprocessing script is to identify variants either based on quality thresholds or the whitelist and then identify those cells that have at least 40% of those selected variants genotyped. For those filtered cells, I retrieved the barcodes and saved them as comma-separated values (CSV) with the suffix "-barcodes.csv" (see Insert 2).

```
ds.ca["barcode"][filtered_cells].tofile(os.path.join(
  outdir,basename.replace(".cells", "")+"_barcodes.csv"), sep = ",")
```

Insert 2: Barcodes of selected cells are written to a comma-separated file.

The *FLT3*-ITD is detected in a cell when the alternative allele in amplicon `AMPL135278`, which is the amplicon designed specifically for *FLT3*-ITD of patient 02, has a length greater than 1. This also changes the variant name to *FLT3*-ITD. Fusion gene reads are added as variant with name "Fusion inv(16): CBFB-MYH11" or "Fusion t(8;21): RUNX1-RUNX1T1" in case of patients 01 and 09. If reads were found on the specific fusion gene chromosomes the number of reads are annotated as reference and alternative allele count and classified as heterozygous [137]. If there were two breakpoints present in a patient the reads on both amplicons were summed up.

The information is written to a CSV file containing the following columns: (i) the chromosome (CHR), (ii) the start position (POS), (iii) the reference allele (REF), (iv) the alternative allele (ALT), (v) the gene name (REGION), (vi) the variant annotation (NAME), (vii) the population frequency from the 1000 genomes project [123] (FREQ) and additional columns for each selected cell. A variant in each cell is annotated with reference counting reads, alternative counting reads and the genotype (0=wild-type, 1=heterozygous, 2=homozygous or 3=missing) delimited by colons. Additionally, barcodes (*e.g.,* 'AACAACTGGCCAGTCTCA-1') and read counts of fusion amplicons for each filtered cell are saved as CSV files for further downstream analysis.

3 Reconstruction of tumor phylogeny

Many of the available methods for reconstructing tumor phylogeny can only perform on approximately 100-1,000 cells in a reasonable amount of time, as described in section I3.2. COMPASS [64] and ConDoR [50] are two of the available methods able to infer tumor phylogeny with copy-number alterations. In addition to information on variants, ConDoR needs clustering information of copy-number profiles of single-cells, which was not possible with my samples. Moreover, clone sizes of inferred phylogenies did not match with whole-exome results. Figure 20 shows the inferred tumor phylogeny from COMPASS of patient 04 at diagnosis with copy-number alterations. COMPASS showed a loss on chromosome 17 for two tumor clones but was not able to retrieve known copy-number alterations from karyotype (*i.e.*, amplification on chromosome 13,14 and 22 as listed in Table 3).

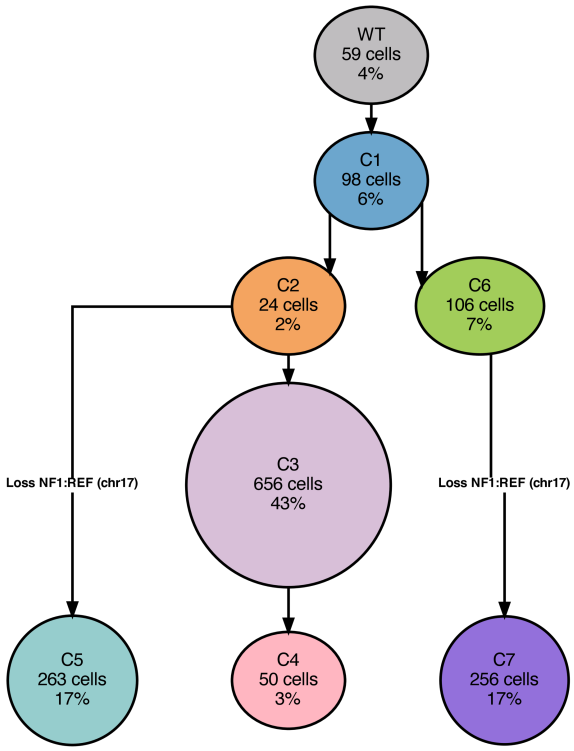

Figure 20: Simplified tumor phylogeny of patient 04 at diagnosis using COMPASS with copy-number alterations. Tree starts at the top with the wild-type (WT) cell fraction and 7 tumor clones (C1-C7). Somatic variants for each tumor clone are not shown. COMPASS [105] shows a deletion on chromosome 17 for tumor clones C5 and C7, but only amplifications of chromosome 13, 14 and 22 have been detected by karyotype.

Therefore, I developed a 2-step approach to infer the tumor phylogeny for each patient: (i) reconstruct tree with COMPASS using only somatic variants and gene fusion information and (ii) identify SCNAs in each node with wild-type cells as reference.

3.1 Step 1: Infer trees based on somatic variants

For inferring the tumor phylogeny, I used COMPASS (v1.1) [64] in mutation mode (`--CNA 0`), the corresponding sex (*i.e.*, for patient 08 and 09 `--sex female` and `--sex male` for others), with 10 Markov chain Monte Carlo (MCMC) chains in parallel and 20,000 iterations in each (`--nchains 10 --chainlength 20000`), and default parameters. COMPASS generates 5 output files (i-v, information from https://github.com/cbg-ethz/COMPASS) with an additional custom file (vi):

(i)&(ii) **[sample]_tree.{gv/json}:** the inferred tree in graphviz and JavaScript Object Notation (json) format

(iii) **[sample]_cellAssignments.tsv:** hard assignments of cells to nodes, and whether cell was inferred to be a doublet (in which case the node assignment is unreliable).

(iv) **[sample]_cellAssignmentsProbs.tsv:** posterior attachment probabilities of cells to nodes

(v) **[sample]_nodes_genotypes.tsv:** genotype of each SNV for each node (0: no mutation; 1: heterozygous mutation; 2: homozygous mutation)

(vi) **[sample]_data.csv:** custom file containing cell numbers and hex color codes for each node

Figure 21a shows the original visualization of phylogenetic trees form COMPASS. For better visualization, I modified the code to output trees with events on branches between nodes (see Figure 21b) in graphviz format which is converted to PDF with Graphviz's dot [138].

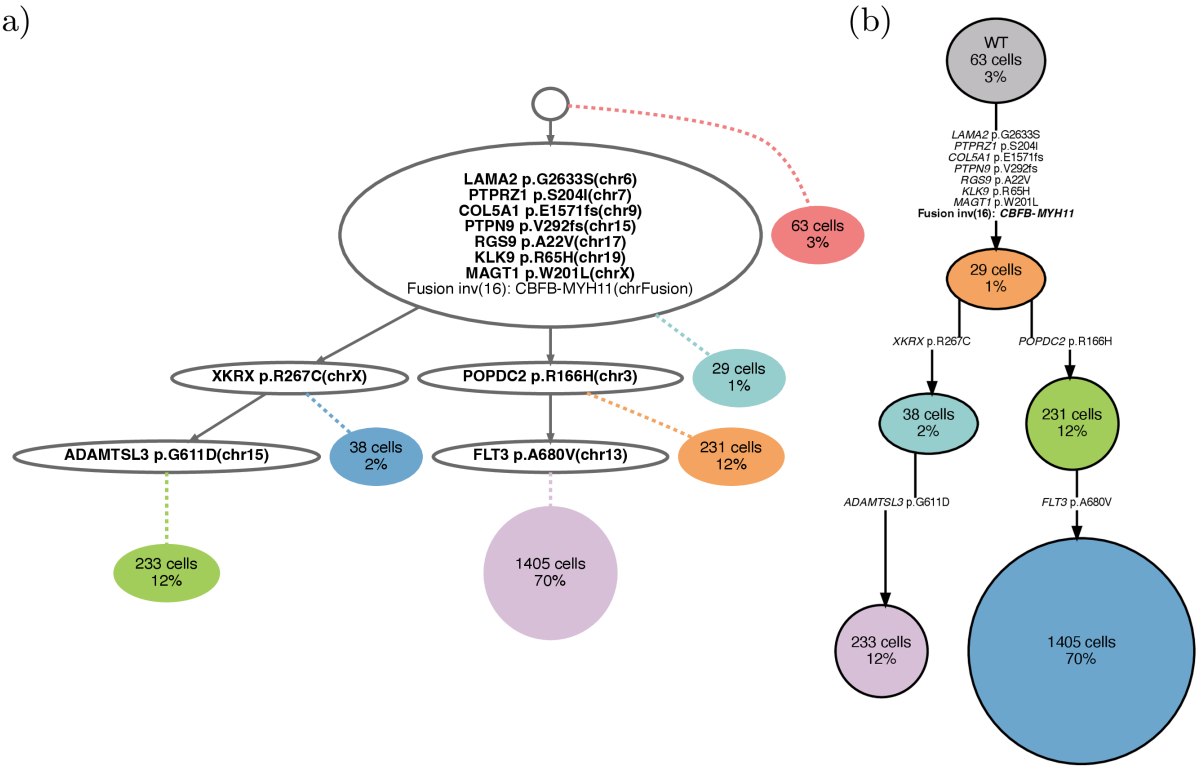

Figure 21: Visualization of inferred trees. (a) Original visualization of phylogenetic trees from COMPASS [105] and (b) modified version with events located on branches between nodes. Phylogenetic tree shown is from patient 08 at diagnosis with somatic variants and gene fusion only.

3.2 Step 2: Identify nodes with somatic copy-number alterations

For the analysis of SCNAs of each tumor clone in an inferred phylogenetic tree I used R with following packages: reticulate (v1.28) [139], ggplot2 (v3.4.4) [140], data.table (v1.14.8) [141], stringr (v1.5.0) [142], ggpubr (v0.6.0) [143], dplyr (v1.1.1) [117], jsonlite (v1.8.5) [144], readxl (v1.4.2) [145] and ggh4x (v0.2.6) [146]. Additionally, I used Python 3 (v3.7.9) [147] with mosaic (v2.2) (Mission Bio).

Each custom panel (*i.e.*, CO-413, CO-414 and CO-415) contains specifically designed amplicons for copy-number analysis that are initially grouped by chromosomes (see Figure 22,

Grouped amplicons). A region of interest (ROI) is defined as a segment of a known SCNA for a patient identified by copy-number calling or karyotype. Regions for copy-number analysis are classified into copy-number neutral regions (see Figure 22: R_1, R_2 and R_3) and ROIs (see Figure 22: ROI_1 and ROI_2). If ROI covers a chromosome only partially then amplicons on that chromosome are split into regions left and right of the ROI and the ROI itself (see Figure 22: Regions for analysis, R_2, ROI_2 and R_3). Amplicons are selected for analysis if more than 75% or 50% (for amplicons in ROIs) of cells have a minimum read depth of 30. Copy-number neutral regions are excluded from further analysis if they consist of less than 4 amplicons (see Figure 22: region R_3).

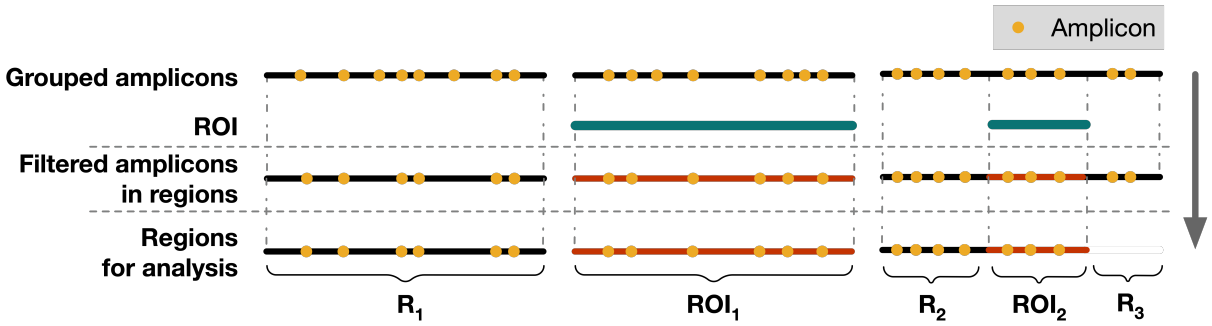

Figure 22: Grouping and filtering copy-number amplicons in regions. To obtain regions for the analysis, amplicons for copy-number analysis are initially grouped by chromosomes (CHR) and then masked with regions of interest (ROIs). ROIs are patient specific segments of known copy-number alterations.

I merged cell assignments from COMPASS with cell barcodes and, subsequently, removed cells that have been marked as doublet. This results in a list of cell barcodes for each tumor clone and the wild-type cell fraction. Then I calculated the ploidies p_{ij} for each cell j at amplicon i using `Cnv.compute_ploidy` from mosaic (v2.2) (Mission Bio) with cells from the wild-type fraction as diploid reference and default parameters. For cells j of each tumor clone n I performed following steps:

- calculate variance $\mathrm{Var}(p_{ij})$ for amplicon i of ploidies p_{ij}
- calculate Z-scores Z_{ij} for ploidies p_{ij}
- filter cell j of tumor clone n if $Z_{ij} < 2$
- calculate ploidies of copy-neutral regions (R_1, R_2, …) and ROIs (ROI_1, ROI_2, …) using a weighted mean, where weights are the normalized variance of amplicon i from 1 to 0.1 so that higher variance amplicons are adding less to the ploidy
- (optionally) center ploidies of ROIs by subtracting the mean of the ploidies of copy-neutral regions (R_1, R_2, …) – 2 (=ploidy of copy-neutral region)

I used the weighted mean for ploidies to allow a less strict filtering of high variance amplicons.

In patients 05 and 09 I detected UPDs using genotype information of SNPs for cells in each without missing data from each cell in regions of known UPDs. I used the fraction of wild-type, heterozygous and homozygous cells in each node for deciding if a UPD is present or absent.

4 Detecting clones in remission

For each complete remission sample, I used the preprocessing script as described in paragraph 2 to identify variants and the existence of the fusion gene. At first, I selected cells that had at least one variant annotated as heterozygous or homozygous or cells where the gene fusion is present using R. The infinite-sites model, which is the simplest phylogeny model, allows every variant in a phylogenetic tree to change only one time from wild-type to mutated (*i.e.,* heterozygous or homozygous) and is never lost [53]. Applying the infinite-sites model, I classified every selected cell with the clone that has matching somatic variants and is the furthest away from the wild-type fraction. Results were visualized using R with packages ggplot2 (v3.4.4) [140] and latex2exp (v0.9.6) [148].

5 Results

In this section I present results from reconstructing the history of DNA alterations using somatic variants, the existence of the CBF gene fusion and chromosomal alterations. Phylogenetic trees from patients that do not have additional chromosomal alterations are shown in section 5.2. Inferred tumor phylogeny from patients with SCNAs including amplifications, deletions and UPDs are presented in section 5.3.

In general, a phylogenetic tree consists of nodes (representing a cell clone) and branches connecting those nodes. The trees in this thesis visualize the tumor development starting with the wild-type cell fraction from which the tumor clones connected by branches emerge. Events listed along those branches (*e.g.*, *BCORL1* p.R609X) are acquired from one node to the next. This means that every clone in the tree contains all the somatic events that are listed from that node back to the wild-type cell fraction. Nodes on two different branches have a common ancestor but have acquired additional somatic events independently.

Cell numbers in this section can differ between results from the Tapestri pipeline (Mission Bio, v2.0.2), genotyping and phylogenetic trees due to different thresholds for each analysis.

5.1 Tapestri pipeline results

In total 21 samples from 9 patients were analyzed using the Tapestri pipeline (Mission Bio, v2.0.2), representing different timepoints of the disease stage (*i.e.*, diagnosis (D), complete remission (CR) or relapse (Rel)), as listed in Table 17. Samples with less than 1,000 detected cells (*i.e.*, diagnosis sample of patient 05 and relapse sample of patient 07) and samples with no coverage on patient specific fusion amplicons (*i.e.*, patient 03) were removed from further downstream analysis. For the remaining samples the number of detected cells ranges from 1,637 to 7,540 with a mean read/cell/amplicon depth of 35 and 203.

Table 17: Tapestri pipeline run metrices. In total 21 samples with three different panels with a size ranging from 180 to 201 amplicons were analyzed using the Tapestri pipeline (Mission Bio, v2.0.2). Samples are available at diagnosis (D), complete remission (CR) and relapse (Rel). Fusion is yes, if reads were found on the patient specific fusion amplicon and vice versa. Depth is mean reads/cell/amplicon. (Pat. = Patient)

Pat.	Panel (# Amplicons)	Sample	Reads [x10^6]	Fusion	# Cells	Depth	Analysis
01	CO414 (201)	D	191	yes	3337	57	yes
		CR	391	yes	2818	149	yes
		Rel	184	yes	4574	50	yes
02	CO413 (180)	D	251	yes	7540	35	yes
		CR	242	yes	4103	83	yes
		Rel	363	yes	4125	142	yes
03	CO414 (201)	D	213	no	5287	60	no
		CR	330	no	4227	103	no
		Rel	187	no	4468	54	no
04	CO413 (180)	D	346	yes	1637	203	yes
		CR	341	yes	5333	107	yes
05	CO413 (180)	D	341	yes	15	43035	no
		CR	289	yes	2351	201	yes
		Rel	250	yes	2665	142	yes
06	CO414 (201)	D	185	yes	2687	79	yes
		CR	293	yes	2459	130	yes
07	CO415 (201)	D	290	yes	5884	64	yes
		Rel	183	yes	711	384	no
08	CO413 (180)	D	370	yes	2094	267	yes
09	CO415 (201)	D	291	yes	4777	88	yes
		CR	237	yes	2144	140	yes

5.2 Tumor development without copy-number alterations

In the following section the tumor development visualized with phylogenetic trees of patients without additional copy-number alterations are shown. For those patients I solely used somatic variants and the existence of the fusion gene to infer tumor phylogeny.

5.2.1 Patient 06

For patient 06 I used 11 of 21 somatic variants and the *CBFB-MYH11* fusion for inferring the tumor development at diagnosis. Figure 23 shows the percentage of cells classified as wild-type (WT), heterozygous (HET), homozygous (HOM) or missing for each somatic variant detected in bulk sequencing and, additionally, the presence of the *CBFB-MYH11* gene fusion for (a) diagnosis and (b) remission. A cell is classified as HET if the gene fusion is present (see Section 2). I excluded variants if the amplicon is not covered (*i.e.*, *LOC101059915* p.G89S and *TREML4* p.L37F), there was no change between diagnosis and complete remission (*i.e.*, *OR10G8* p.V28I) or less than 5 cells were mutated at diagnosis (*e.g.*, *AZGP1* p.P6S). Bars of variants that were excluded from inferring tumor phylogeny are shown in faded colors.

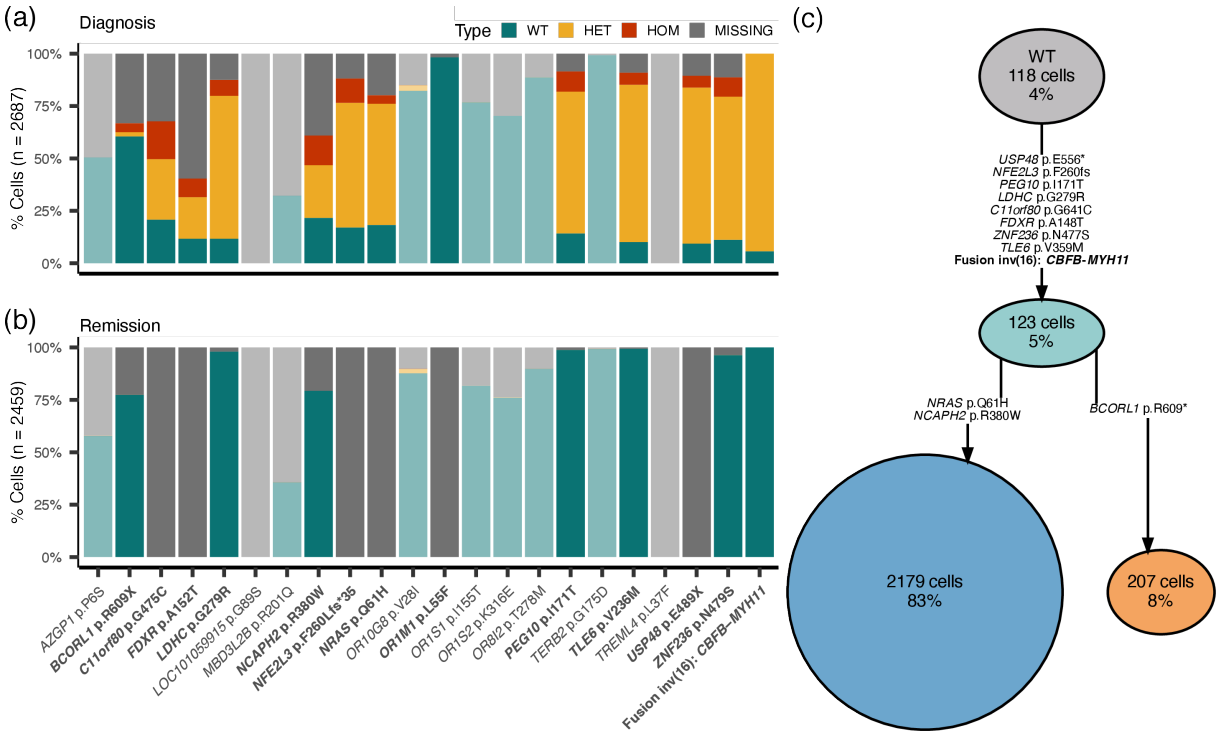

Figure 23: Patient 06 single-cell results. (a) Genotyping information of all 2,687 cells detected at diagnosis. Every variant in each cell is classified as wild-type (WT), heterozygous (HET), homozygous (HOM) or missing (MISSING). If the gene fusion is detected the cell is classified as HET. Variants that were excluded from further analysis are presented in shaded colors. (b) Genotyping information of all 2,459 cells detected at remission. (c) Inferred phylogenetic tree of patient 06 at diagnosis consisting of 3 tumor clones with branching event after *CBFB-MYH11* harboring founding clone.

Figure 23c shows the inferred tree for the diagnosis sample of patient 06 consisting of 3 tumor clones and the wild-type (WT) cell fraction. From the founding clone which contains the

CBFB-MYH11 gene fusion two subclones with a distinct set of variants develop. The *NRAS* p.Q61H is the dominant tumor clone in this sample with 83% of all cells, which is in line with the inferred VAF of 44% in bulk sequencing. *BCORL1* p.R609X has been detected in bulk sequencing with a VAF of 3.8% leading to roughly 7.6% of mutated cells, which matches with the size of the *BCORL1* clone in the single-cell data as shown in Figure 23c.

5.2.2 Patient 02

Figure 24 shows the genotyping results for 18 somatic variants and the CBFB-MYH11 fusion gene of patient 02 for samples at (a) diagnosis, (b) remission and (c) relapse. I excluded variants from further analysis if the amplicon was not covered (*i.e.*, *KMT2C* p.G908C and *PTPN20* p.G28E) or there were less than 5 mutated cells in both tumor samples (*i.e.*, *IL1RAPL1* p.P241S and *LRRC74A* p.P36S). *FLT3*-ITD, the most detected genetic aberration in AML, was detected in 29.7% of all cells at diagnosis (*i.e.*, 2,011 cells HET and 91 cells with HOM) and in 22.5% of all cells at relapse (*i.e.*, 837 cells heterozygous and 84 cells homozygous) [128]. The percentage of mutated cells at diagnosis with *FLT3*-ITD can be converted to an allelic ratio of approximately 15% when assuming heterozygosity. This matches the allelic ratio of 13% detected with clinical testing as listed in Table 3. In contrast to single cell sequencing, the *FLT3*-ITD was only identified in the diagnosis sample by whole exome sequencing (see Table 8), highlighting the difficulty in detecting this aberration in conventional bulk sequencing. I used 15 of 18 somatic variants and the information of the *CBFB-MYH11* gene fusion to reconstruct the tumor development of patient 02.

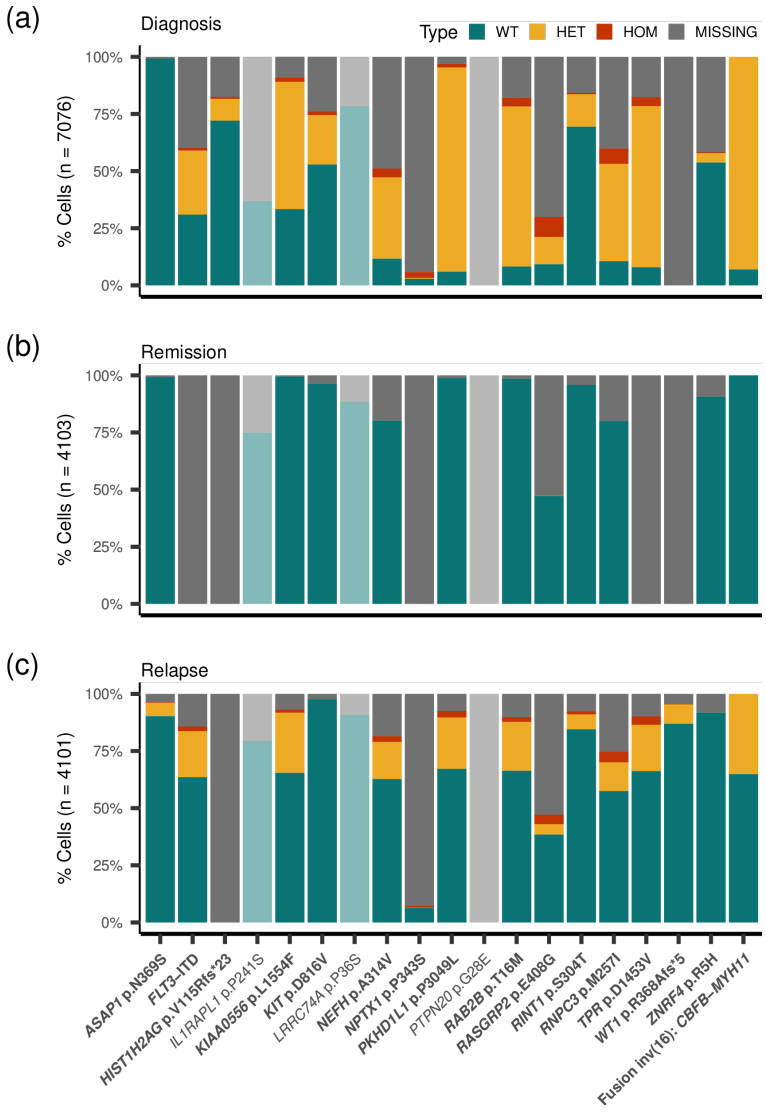

Figure 24: Single-cell genotyping information patient 02. Each variant in each cell is classified as wild-type (WT), heterozygous (HET), homozygous (HOM) and missing. Genotype information for all cells detected at (a) diagnosis (n = 7,076), (b) remission (n = 4,103) and (c) at relapse (n = 4,101). Variants excluded from further analysis are presented in shaded colors.

I merged cells from diagnosis and relapse to infer the tumor pyhlogeny as shown in Figure 25a. Numbers in each clone represent the number of cells from the diagnosis and relapse sample seperated by a slash. The fraction of wild-type cells in this patient is 4% at diagnosis and 64% at relapse. From the founding clone with the *CBFB-MYH11* gene fusion and a clone size of 26 cells (*i.e.*, 3 cells at diagnosis and 23 cells at relapse) additional somatic variants are acquired. The phylogenetic tree at diagnosis (Figure 25b) consists of a diagnosis specifc branch with *KIT* p.D816V, *HIST1H2AG* p.V115Rfs*23 and *ZNRF4* p.R5H subclones and a second branch with a dominant *FLT3*-ITD clone and a *RINT1* p.S304T clone. At relapse, the *KIT* branch is lost and the tumor progresses by acquiring two additional variants (*i.e.*, *WT1* p.R368Afs*5 and *ASAP1* p.N369S). The somatic variants in the diagnosis-specific *KIT* branch are also not detectable in bulk sequencing data of the relapse sample (see Table 8).

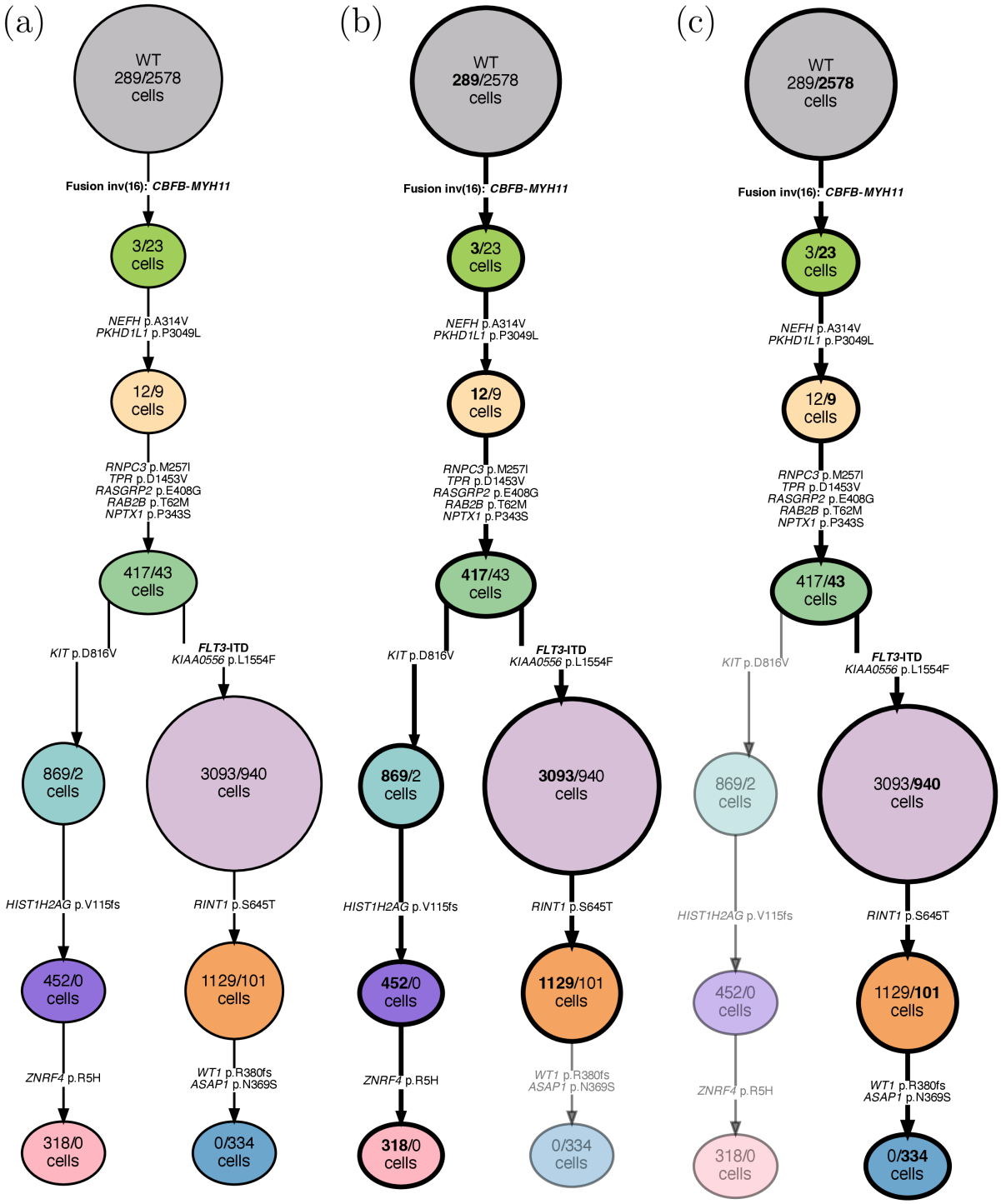

Figure 25: Inferred phylogenetic tree of patient 02. (a) Inferred tree from merged diagnosis and relapse samples with in total 10912 cells (*i.e.,* 6,882 cells from diagnosis and 4,030 cells from relapse), 15 somatic variants and the *CBFB-MYH11* gene fusion. Numbers in each node represent the number of cells from diagnosis and relapse separated by a slash. (b) Diagnosis tree with branching event resulting in a *FLT3*-ITD and a *KIT p.*D816V branch. (c) Relapse tree with a new *WT1* clone, but without a branching event.

5.3 Tumor development with copy-number alterations

In this section I present inferred phylogenetic trees from patients that have additional copy-number alterations detected by bulk sequencing or conventional G-banding. As described in section 3, I first reconstructed the tumor development using only somatic variants and the presence of the CBF gene fusion and subsequently use the cells of each clone to call copy-numbers. Ploidies for each clone are calculated with the cells in the wild-type fraction as reference.

5.3.1 Patient 08

For patient 08 I used 11 of 21 somatic variants and the *CBFB-MYH11* gene fusion as highlighted in Figure 26 for inferring the phylogenetic tree at diagnosis. I excluded variants that are shaded in Figure 26, because the amplicon was not covered (*i.e.*, *CPAMD8* p.R402G, *HDX* p.I369S, *KMT2C* p.G908C, *PHF6* p.G29X, *SLC6A15* p.C90F and *WT1* p.D355Y) or there were less than 5 mutated cells in the sample (*i.e.*, *ALKBH4* p.R179W. All the variants that have less than 5 mutated cells except for *BOD1L2* p.L155P are unique to relapse as listed in Table 14.

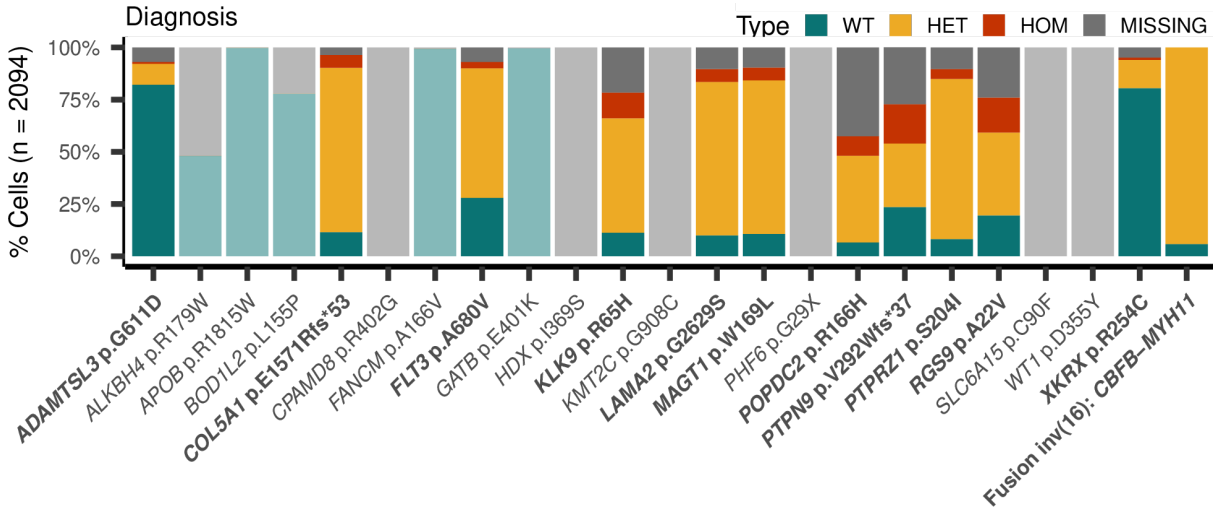

Figure 26: Single-cell genotyping information patient 08. In total 2,094 cells have been genotyped at diagnosis as wild-type (WT), heterozygous (HET), homozygous (HOM) and missing (MISSING) for each somatic variant. I excluded variants in faded colors from further downstream analysis because the amplicon did not work (*e.g.*, *WT1* p.D355Y) or there were less than 5 cells mutated (*e.g.*, *ALKBH4* p.R179W).

Figure 27a shows the inferred phylogenetic tree for patient 08 at diagnosis based on somatic variants and the gene fusion only. From the founding clone harboring the *CBFB-MYH11* gene fusion two clones each consisting of two subclones emerge. Here the dominant clone with 70% of all cells acquired *POPDC2* p.R166H and *FLT3* p.A680V additionally to the somatic variants in the founding clone. Only the founding clone was detectable in the relapse sample with bulk sequencing as listed in Table 14, whereas the dominant clone from diagnosis

was absent at the time of relapse sample analysis. No viable cells from relapse were available from this patient to confirm this observation via single-cell genotyping. Vice versa, the relapse-specific variants were below detection threshold in the single-cell analysis, hinting that they were not present at diagnosis and acquired later.

For each tumor clone, I calculated the ploidy as described in section 3.2 by using the wild-type cells (n = 63) as reference. Figure 27b shows the ploidy for chromosomes 7, 11, 13, 14 and 22 in each clone of the tree. Here individual dots represent the calculated ploidy of an amplicon that has passed quality filtering. The variance of each amplicon is visualized by the strength of the black color - a darker color means a smaller variance. Due to using the wild-type cells as a reference the ploidy of each amplicon in the wild-type clone is exactly 2. For smaller clones, the variance of the ploidies of each amplicon is larger than for clones with a higher number of cells. This can be seen when comparing the founding clone with 29 cells to the 233 cells comprising *ADAMTSL3* clone. In the *POPDC2* p.R166H clone an increase in ploidy on chromosome 22 is already visible, but the amplification on chromosome 22 is clearly detectable in the *FLT3* p.A680V clone (SCNAs are highlighted in bold).

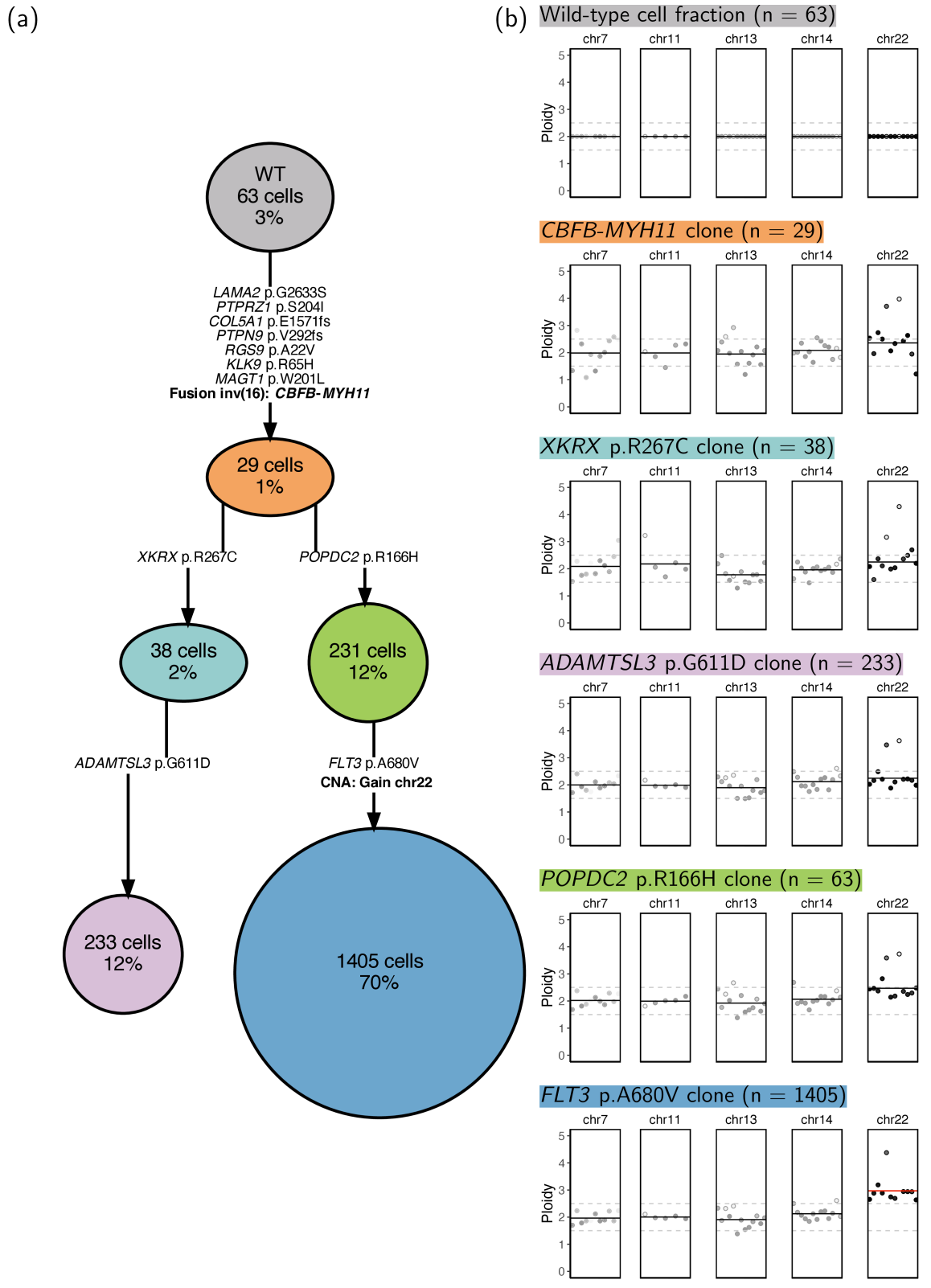

Figure 27: Inferred phylogenetic tree from patient 08 at diagnosis with ploidies for each clone. (a) Inferred tree at diagnosis from somatic variants and *CBFB-MYH11* gene fusion with the founding clone harboring the CBF gene fusion and the *FLT3* p.A680V clone harboring an amplification of chromosome 22. (b) Ploidies of amplicons and regions in each node using the wild-type cell fraction as reference. Dots show ploidy of an amplicon and, additionally, the variance by their shade.

5.3.2 Patient 04

I used 19 of 27 somatic variants and the *CBFB-MYH11* gene fusion for reconstructing tumor development in patient 04 as highlighted in bold in Figure 28. I excluded variants in shaded colors because there were less than 5 mutated cells in the tumor sample or the amplicon was not covered (*i.e.*, *RGPD8* p.D1388Y and *WASHC2C* p.D285H). I removed *ZNF587B* p.F207I because it was part of the wild-type cell fraction in the first inferred tree and, therefore, it would affect downstream analysis especially detecting remaining clones in complete remission.

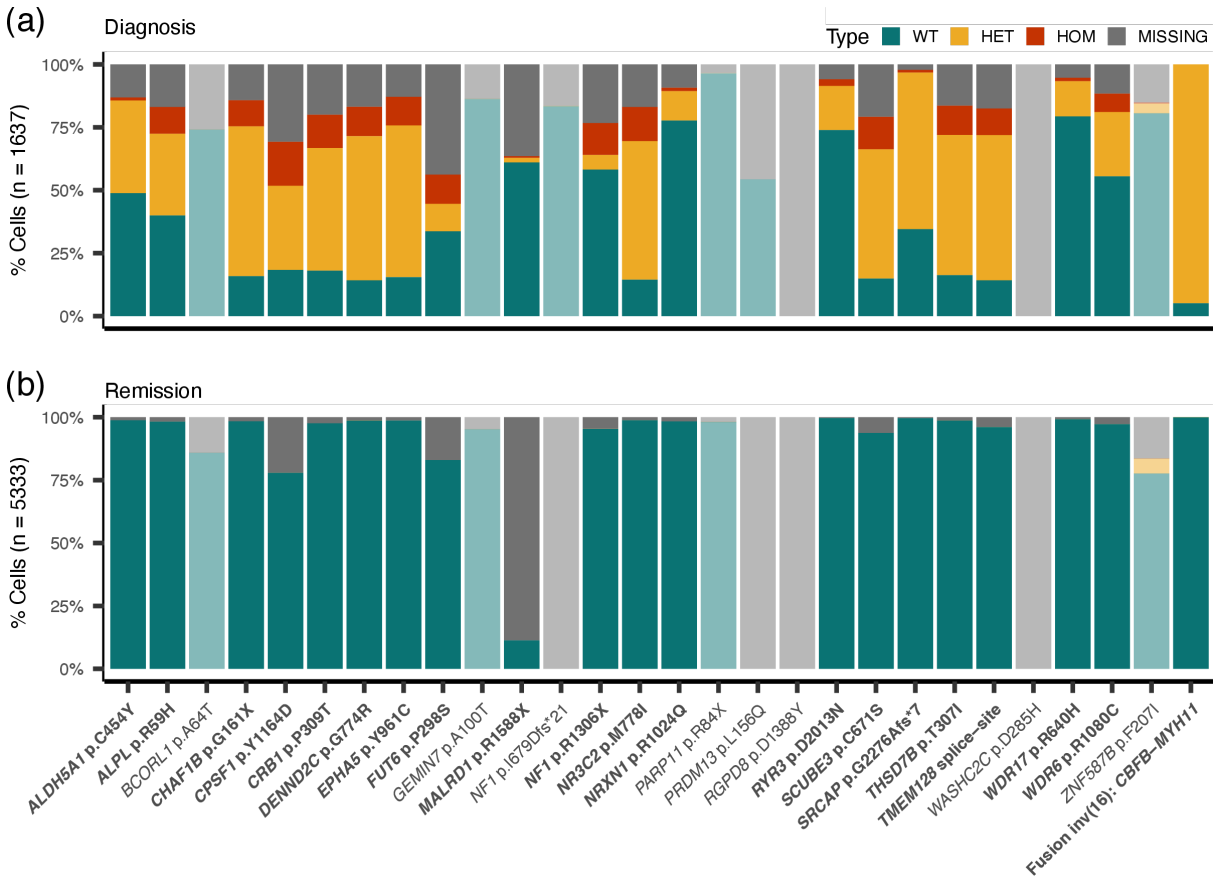

Figure 28: Single-cell genotyping information patient 04. In total 1,637 cell have been genotyped (a) at diagnosis and (b) 5,333 cells at relapse. Variants were excluded from further analysis if there were less than 5 mutated cells in the tumor sample or the amplicon did not work (*i.e.*, *RGPD8* p.D1388Y and *WASHC2C* p.D285H). Additionally, *ZNF587B* p.F207I has been removed from tree reconstruction because it is part of the wild-type clone. Excluded variants are visualized in shaded colors.

Figure 29a shows the inferred tree for patient 04 at diagnosis consisting of 6 tumor clones with the founding clone (n = 59) containing the *CBFB-MYH11* gene fusion. The SNV in *NF1* and the frame-shift variant in *SRCAP*, which are known AML driver genes, form two distinct clones. Copy-number calling in bulk sequencing data did not show any SCNAs for patient 08, but subclonal amplifications of chromosomes 13, 14 and 22 are present in the karyotype at diagnosis (see Table 3). Figure 29b visualizes the ploidy for chromosomes 7, 11 and the chromosomes of interest (*i.e.*, chr 13, 14 and 22) in each tumor clone in comparison to the 59

wild-type cells. The *NF1* clone with 24% of all cells (n = 362) shows amplifications on chromosomes 13, 14 and 22. This matches the number of metaphases (13/49 ~27%) with an additional chromosome detected by conventional G-banding (see Table 3). This clone seems to be lost at relapse, because both variants are unique to diagnosis as listed in Table 10.

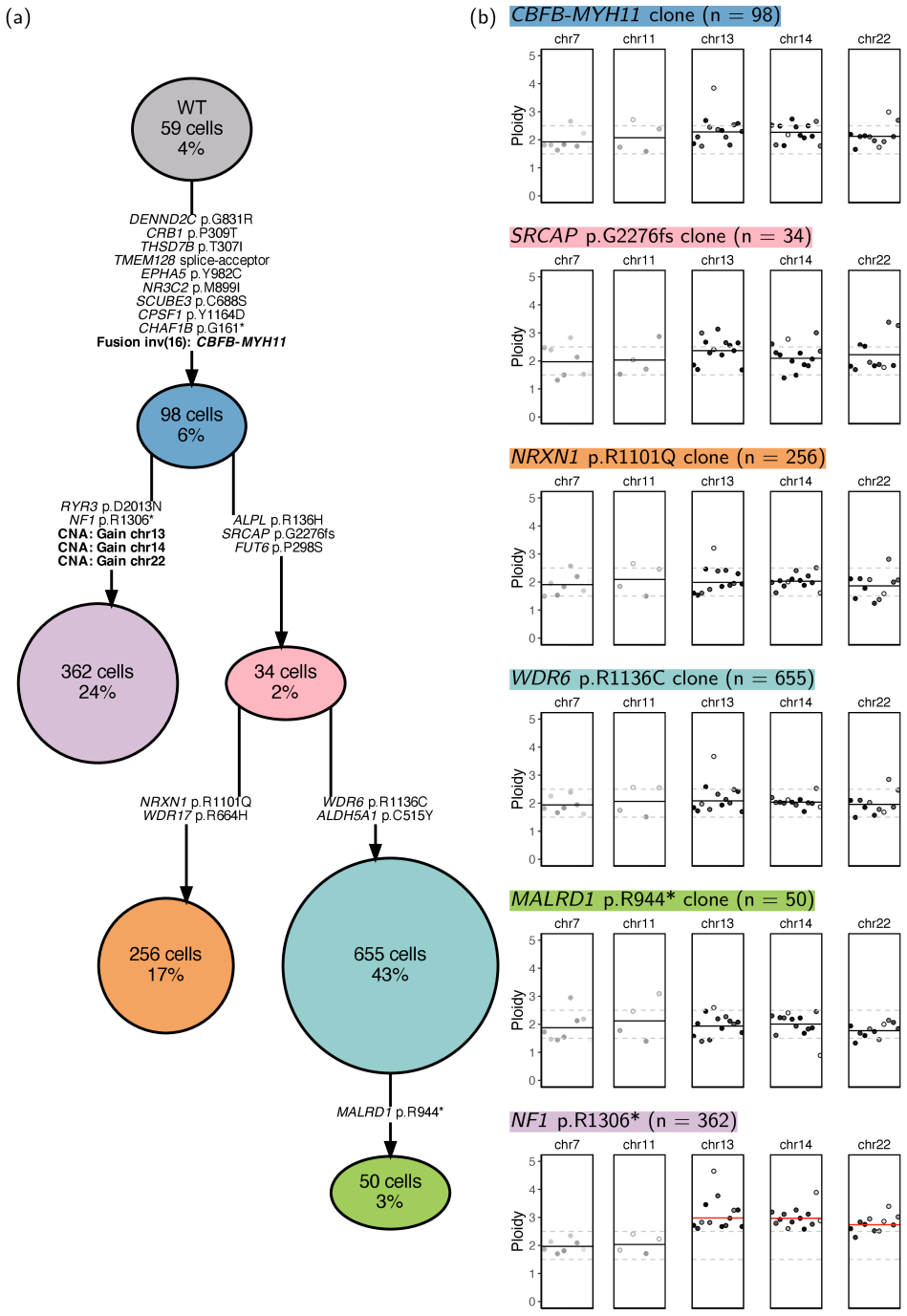

Figure 29: Inferred phylogenetic tree from patient 04 at diagnosis with ploidies of tumor clones. (a) Inferred phylogenetic tree with in total 6 tumor clones and the founding clone harboring the *CBFB-MYH11* gene fusion. (b) Ploidies of chromosomes 7, 11, 13, 14 and 22 for every tumor clone in comparison to the wild-type cells. The *NF1* p.R1306* clone has amplifications on chromosome 13, 14 and 22. For other tumor clones no copy-number changes are detected.

5.3.3 Patient 07

In case of patient 07 I used 26 of 29 somatic variants and the *CBF-MYH11* gene fusion for inferring the tumor phylogeny at diagnosis. Figure 30 shows the genotypes for detected cells in the (a) diagnosis and (b) relapse sample. I excluded *MYO18B* p.A408V and *NF1* p.P2289Sds*17 because the amplicon was not covered and *PCDHGA2* p.P669S, because there were less than 5 mutated cells in the diagnosis sample. For this patient I did not merge the tumor samples for inferring the phylogenetic tree, because only 711 cells have been detected by the Tapestri pipeline for the relapse sample. Only a fraction of the required cell number was available as input for Mission Bio single-cell DNA library preparation for the relapse sample of this patient, resulting in the low cell output as well as skewed variant allele fractions. However, the sample is shown here, to demonstrate which clones were found in the relapse sample, independent of cell fractions.

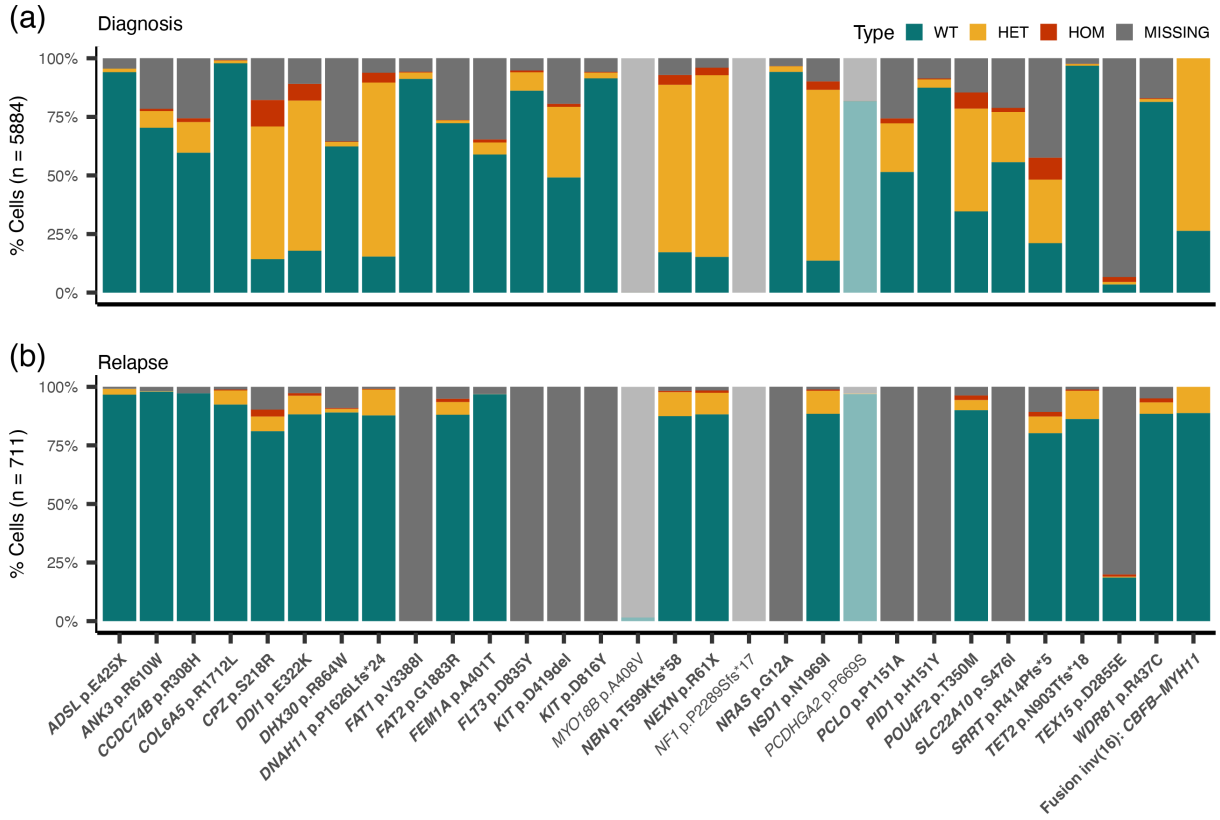

Figure 30: Single-cell genotyping information patient 07. In total 5,884 cells have been genotyped (a) at diagnosis and 711 cells (b) at relapse. I excluded variants in shaded colors because there were less than 5 mutated cells (*PCDHGA2* p.P669S) or the amplicon was not covered (*i.e.*, *MYO18B* p.A408V and *NF1* p.P2289Sds*17).

Figure 31a shows the simplified phylogenetic tree at diagnosis with a CHIP clone harboring *TET2* p.N924fs and an AML/tumor clone. I detected this *TET2* variant using error-corrected targeted sequencing as described in section III4. This separation between the CHIP clone and the AML clone persists throughout the course of the disease as shown in Figure 31b, which is consistent with the bulk data of an increasing CHIP clone at relapse. The complete

inferred tumor phylogeny of the diagnosis sample is shown in Figure 31c. The complex phylogenetic tree consists of 11 tumor clones with the founding clone harboring the *CBFB-MYH11* gene fusion. Interestingly, from the founding clone the two *KIT* variants (*i.e*, *KIT* p.D816Y and *KIT* p.D418del), *NRAS* p.G12A and *FLT3* p.D835Y evolve into distinct subclones, with some of the subclones acquiring additional mutations.

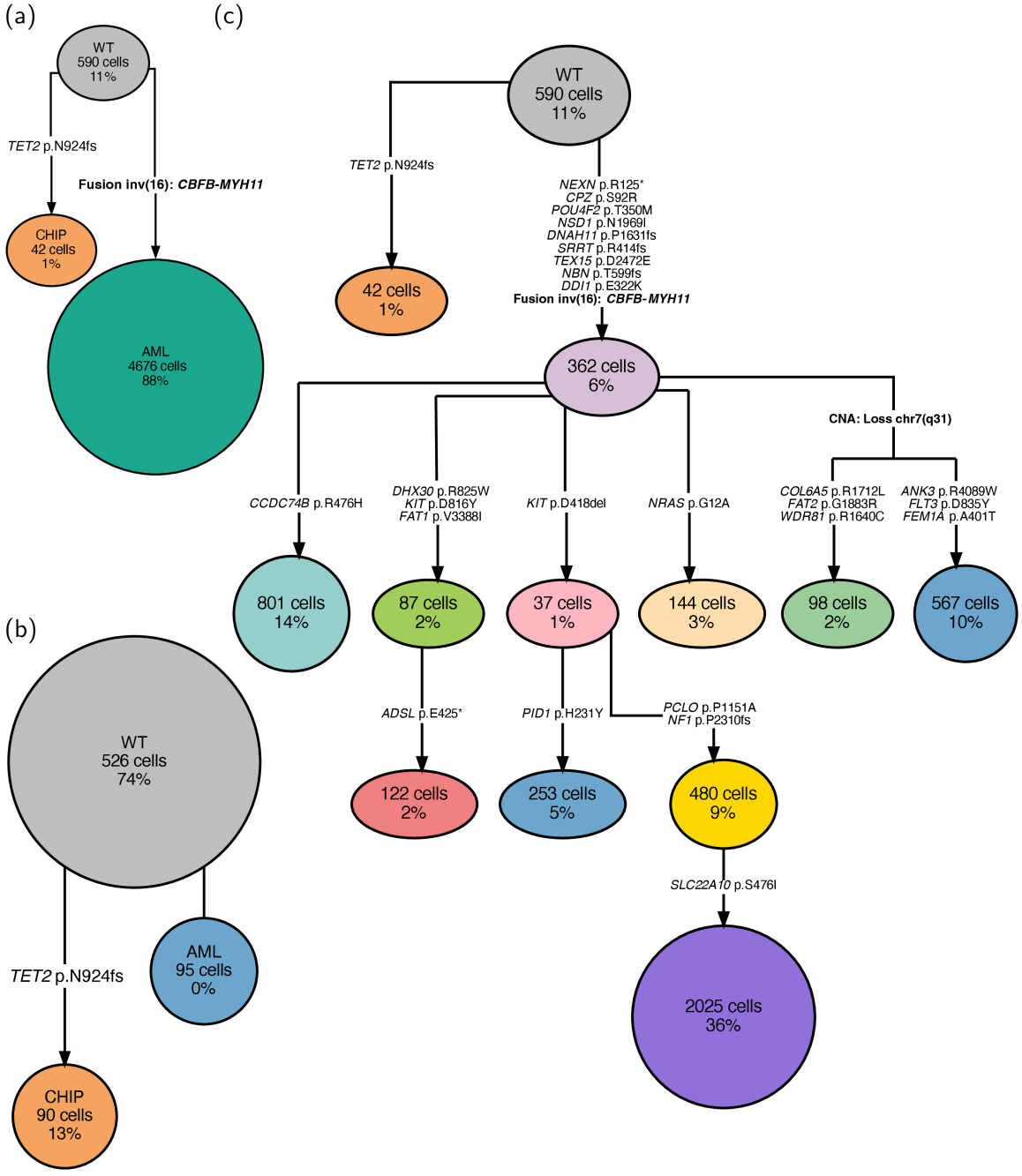

Figure 31: Inferred phylogenetic trees from patient 07. (a) Simplified phylogenetic tree with a dominant AML /tumor clone and a small 42 cells comprising clonal hematopoiesis of indeterminate potential (CHIP) clone at diagnosis. (b) Simplified tree at relapse with the distinct CHIP clone that persists throughout the course of the disease. (c) Complete phylogeny consisting of 11 tumor clones with the founding clone harboring the *CBFB-MYH11* gene fusion and a CHIP clone for patient 07 at diagnosis

For each tumor clone in the sample Figure 32 visualizes the ploidies of chromosome 11, 13 and region of interest 7q31. Here I detected in clones *COL6A5* p.R1712L and *FLT3* p.D835Y a deletion in the region of interest. The deletion in the *NRAS* p.G12A clone I did not consider, because only one amplicon was below a ploidy of 1.5. Also, I did not count the *NF1* p.P2310Sfs clone as deleted for chr7q31, because it shows a borderline deletion and more importantly neither the ancestral nor the descendant clone show a deletion.

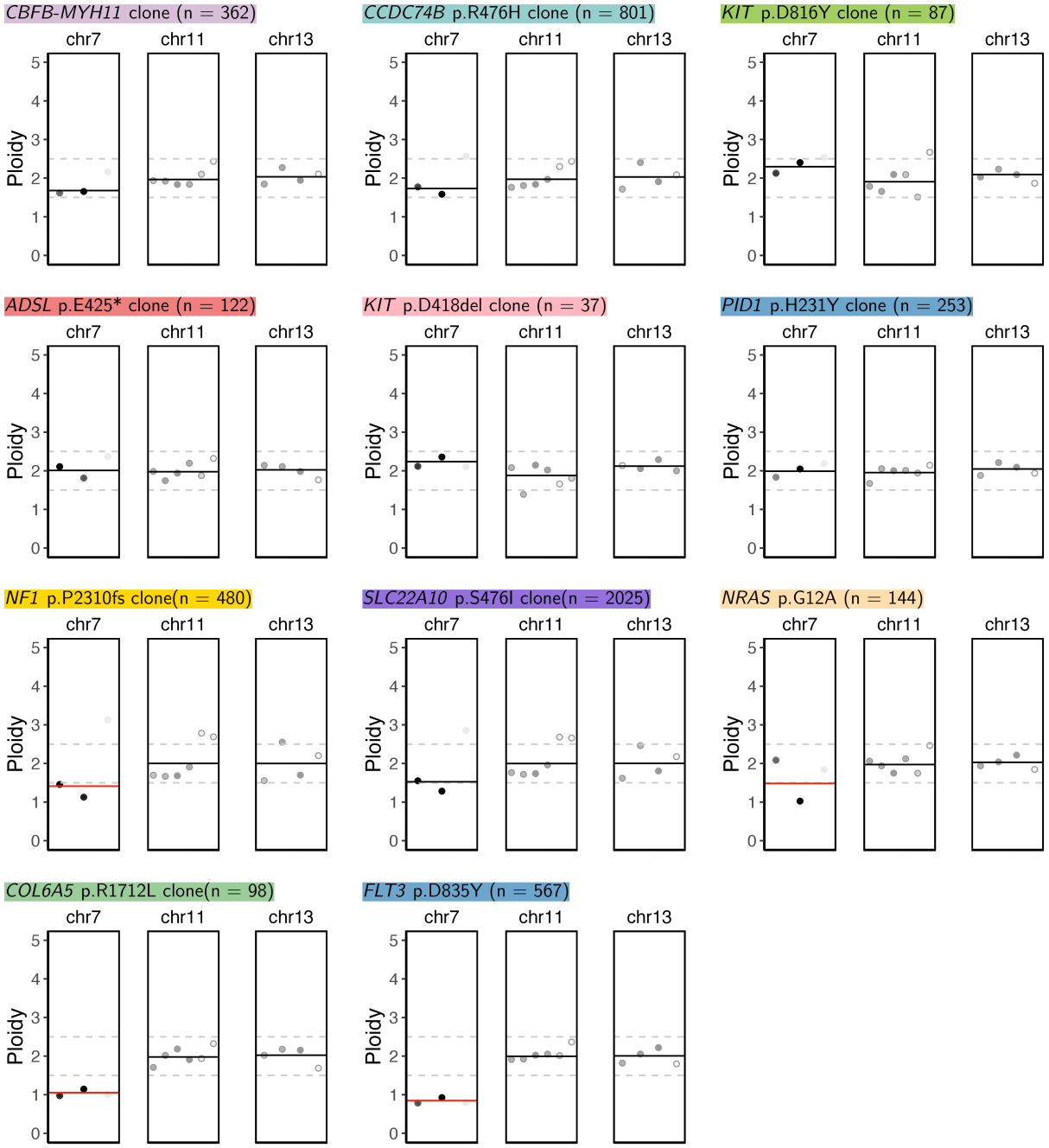

Figure 32: Ploidies of tumor clones of patient 07 at diagnosis. For each tumor clone ploidies for chromosomes 11, 13 and the region of interest on chromosome 7 in comparison to the wild-type cell fraction are shown. The deletion on chromosome 7q31 can be detected in the *COL6A5* p.R1712L and *FLT3* p.835Y clone.

5.3.4 Patient 01

For patient 01 samples at all three timepoints (*i.e.,* diagnosis, remission and relapse) were available and I used 15 of 28 variants and the *RUNX1-RUNX1T1* gene fusion for reconstructing the tumor development. Figure 33 shows the fraction of genotyped cells at (a) diagnosis, (b) complete remission and (c) relapse and variants excluded for downstream analysis in faded colors. I removed variants if the amplicon was not covered (*i.e.*, *IDH2* p.R18P, *OR5H1* p.T167S and *UGT2B7* p.D275E), less than 5 mutated cells have been detected in the tumor samples or if the mutation (as for *PIK3C2A* p.R167K) is part of the wild-type cell fraction. I included *FLT3* p.D835V that I detected at diagnosis in 129 cells (*i.e.*, 112 heterozygous and 17 homozygous) accounting for 4% of all cells for downstream analysis, because *FLT3* variants can be found in approximately 30% and SNVs at the 835 residue of *FLT3* are the most common [149]. This variant was also detectable in WES (*i.e.*, alternative read count = 2, VAF = 1.2%) and targeted sequencing data (*i.e.*, alternative read count = 14, VAF = 1.3%), but was filtered out due to the small VAF. *ZNF213 p.*R147Hfs*28 (chr16:3188459 G>AT) has been called in the single cell data as *ZNF213* p.R147fs (chr16:3188458 C>CA) and *ZNF213* p.R147L (chr16:3188459 G>T), because they were in the same clone I merged them in the inferred phylogenetic tree (see Figure 34a).

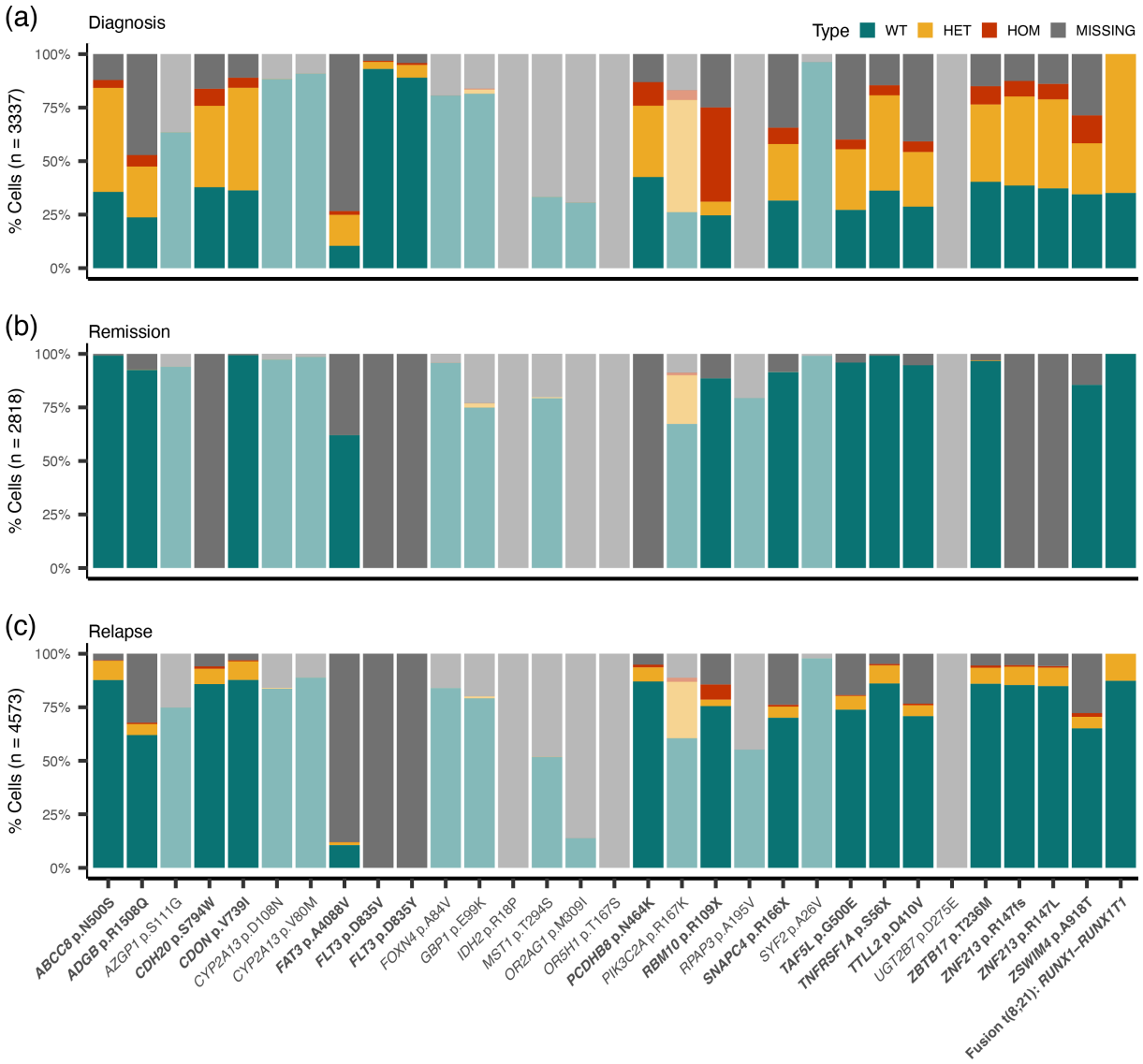

Figure 33: Single-cell genotyping information patient 01. Bar plots show amount of wild-type (WT), heterozygous (HET), homozygous (HOM) and missing classified cells for each variant in the (a) diagnosis, (b) remission and (c) relapse sample. If reads where found in case of the gene fusion *RUNX1-RUNX1T1* the cell was classified as heterozygous. Variants in bold were further used for downstream analysis.

Figure 34a shows the already updated inferred phylogenetic tree of the diagnosis and relapse sample for patient 01. In this patient the founding clone with only *ZBTB17* p.T318M does not harbor the *RUNX1-RUNX1T1* gene fusion. From the founding clone we can distinguish 3 subclones that have acquired additional variants and the CBF gene fusion. From the *ZNF213* clone two diagnosis specific subclones emerge that have acquired distinct *FLT3* variants each: *FLT3* p.D835Y and *FLT3* p.D835V. Both clones are not detectable at relapse. For each of the 6 tumor clones, Figure 34b visualizes the ploidy of chromosomes 4, 8 and 11. The cells of the founding *ZBTB17* clone did not have copy-number changes, but from the *RUNX1-RUNX1T1* clone onwards I detected an amplification of chromosome 8.

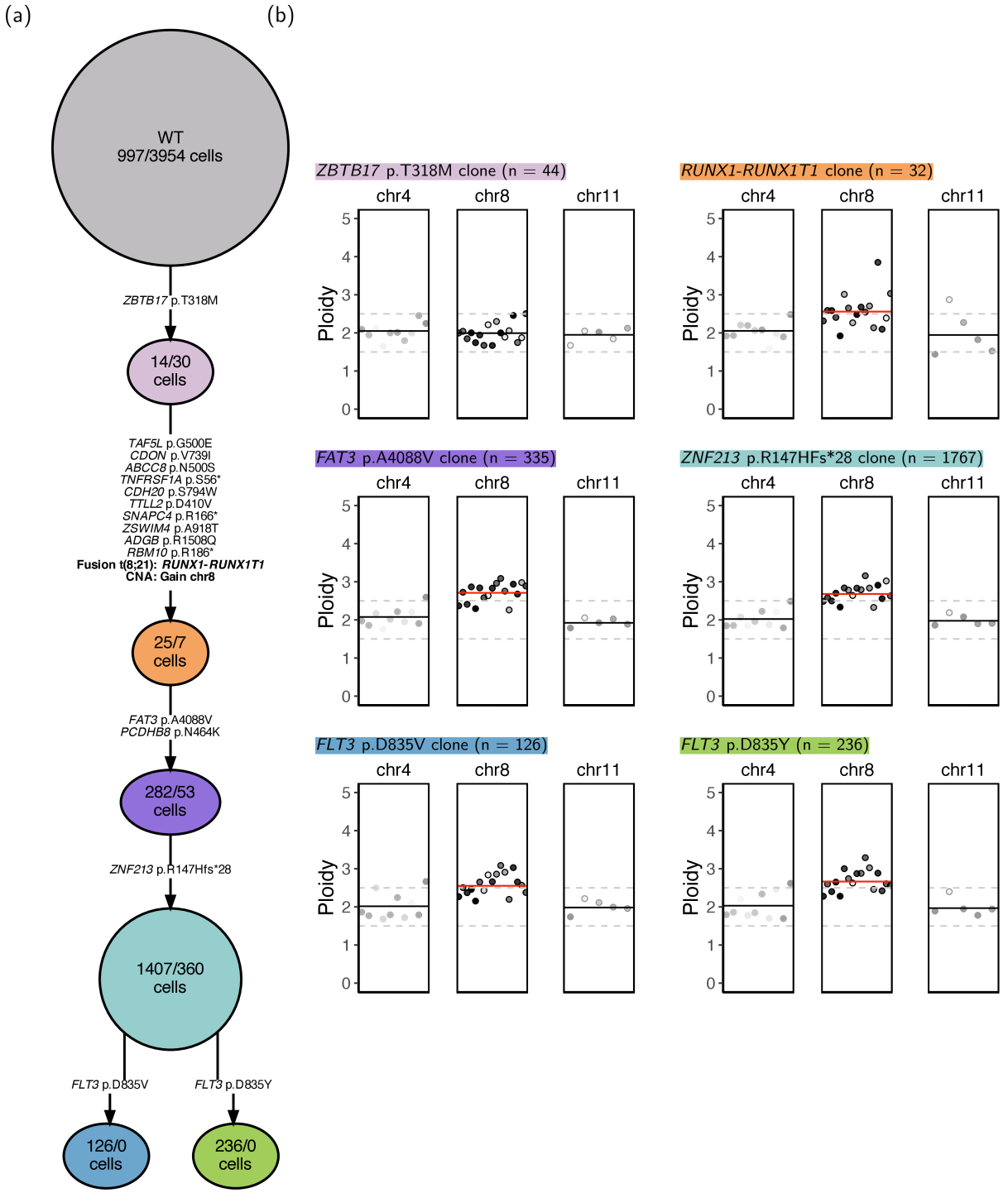

Figure 34: Inferred phylogenetic tree of patient 01 with ploidies for each tumor clone. (a) Inferred tree with in total 7,895 cells (*i.e.*, 3,330 cells from diagnosis sample and 4,565 cells from relapse), 15 variants and fusion gene information of patient 01. In each node the number of cells originating from diagnosis and relapse are separated by a slash. At diagnosis two distinct subclones with *FLT3* somatic variants at the same residue are present. (b) Ploidies for chromosomes 4, 8 and 11 for cells in each tumor clone in comparison to the wild-type cell fraction. An amplification of the whole chromosome 8 is detected from the *RUNX1-RUNX1T1* clone onwards.

5.3.5 Patient 09

I used 22 of 24 somatic variants and the *RUNX1-RUNX1T1* gene fusion as highlighted in Figure 35 for reconstructing the tumor phylogeny of patient 09 at diagnosis. I excluded *TANC2* p.P1886L because the amplicon was not covered and *STAB1* p.A939V because there were less than 5 mutated cells in the tumor sample.

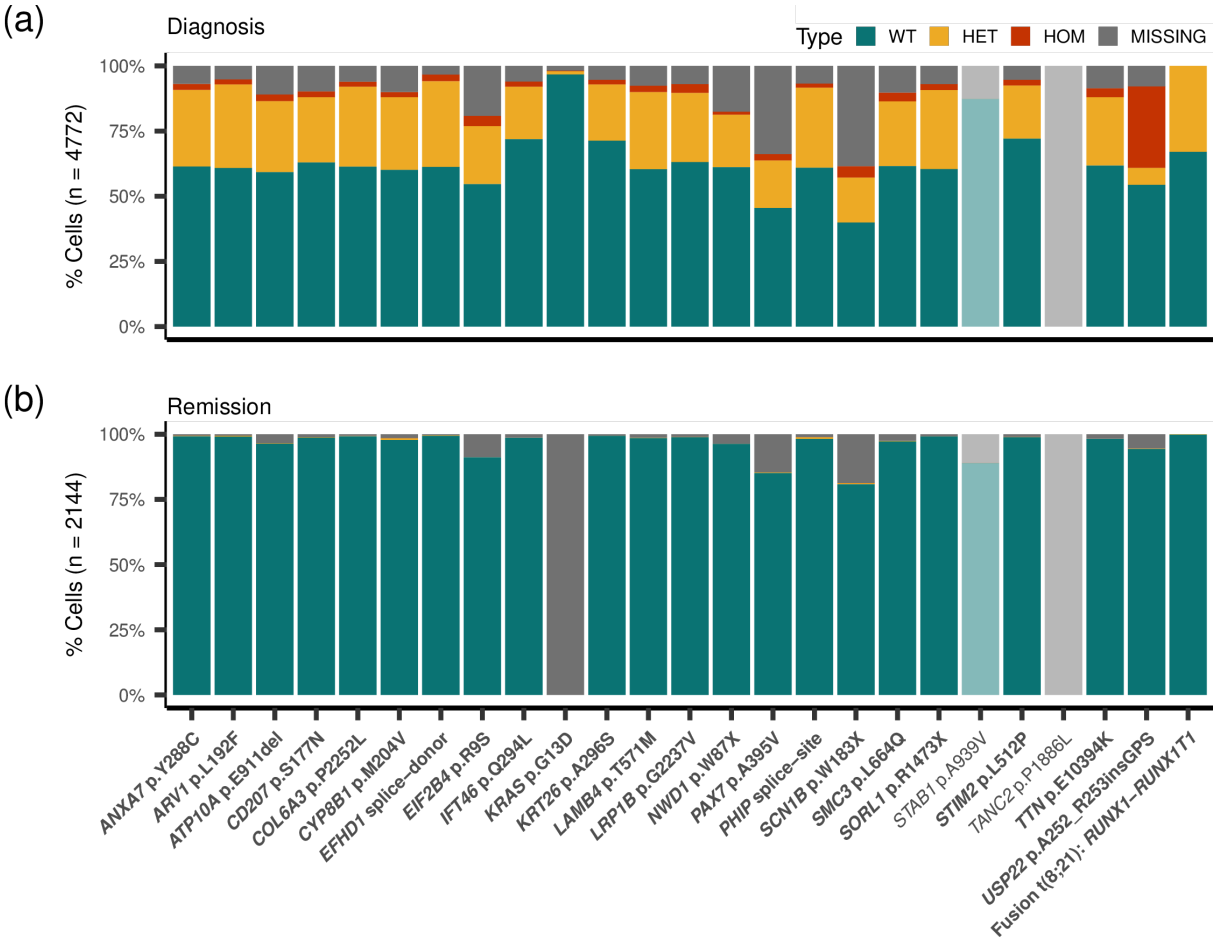

Figure 35: Single-cell genotyping information patient 09. Bar plots show amount of wild-type (WT), heterozygous (HET), homozygous (HOM) and as missing classified cells for each variant in the (a) diagnosis (4,772 cells) and (b) remission (2,144 cells) sample. I excluded *TANC2* p.P1886L because the amplicon did not work and *STAB1* p.A939V because there were less than 5 cells mutated in the tumor sample.

This female patient has a UPD on the p-arm of chromosome 17. This means that one allele is lost and the other one has two copies resulting in a ploidy of 2, which is the ploidy of a normal region. Therefore, I used the gentotype information of two SNPs in that region for measuring the fraction of hetorozygous and homozygous variants. Figure 36a shows the inferred phylogenetic tree consisting of 7 tumor clones for patient 09 at diagnosis. From the 18 cells containing founding clone with variants *ARV1* p.L192F and *SCN1B* p.W183*, 4 subclones emerge with additionally acquired variants. The *LRP1B* p.G2237V clone is ancestral to two subclones with distinct variants. The first subclone is the largest (30% of all cells) harboring *STIM2* p.L607P, *IFT46* p.Q345L and *KRT26* p.A296S mutations, whereas the other subclone

with *KRAS* p.G13D consists of only 64 cells. Barplots in Figure 36b show percentage of cells that are classified as wild-type (WT), heterozygous (HET) or homozygous (HOM) for *USP22* p.252insGPS located in the region of the UPD on chromosome 17 (see Figure 19). I had to infer the UPD solely on this somatic variant, because there were no additional somatic variants or SNPs covered with this panel.

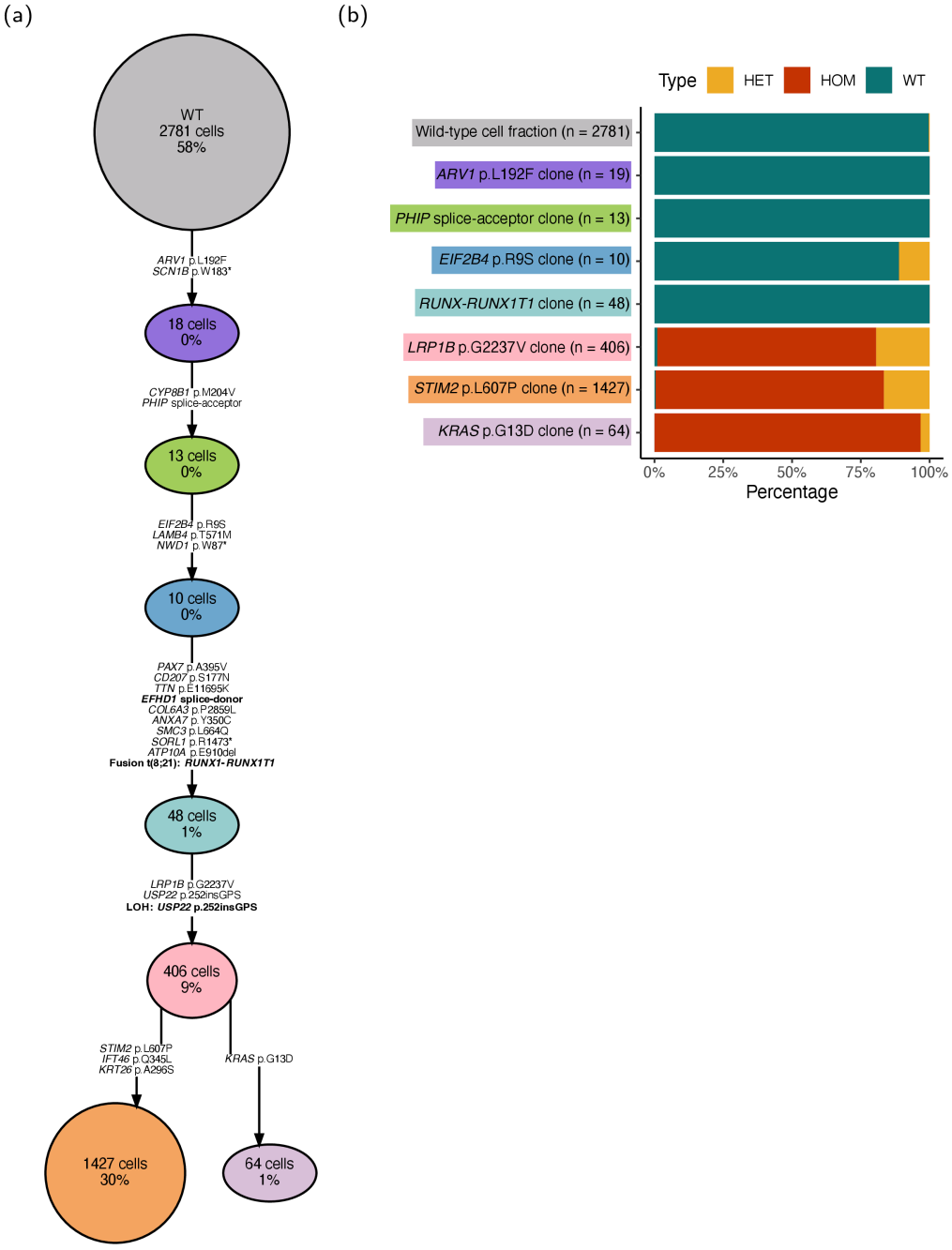

Figure 36: Inferred phylogenetic tree from patient 09 at diagnosis. (a) Inferred tumor phylogeny from 4,772 cells resulting in 7 tumor clones and branching event at the *LRP1B* p.G2237V clone. (b) Barplots show for each clone of the phylogenetic tree the percentage of as wild-type (WT), heterozygous (HET) and homozygous (HOM) classified cells for *USP22* p.252insGPS in each clone. *USP22* p.252insGPS is in the region of the uniparental disomy (UPD) on chromosome 17 detected using whole-exome sequencing data.

5.3.6 Patient 05

I selected 25 of 33 somatic variants and the *CBFB-MYH11* gene fusion to infer the tumor phylogeny for patient 05 at relapse as shown in Figure 37. I did not use the diagnosis sample for further analysis, because only 14 cells have been detected by the Tapestri pipeline as listed in Table 17. I excluded variants if less than 5 mutated cells have been genotyped or the amplicon was not covered. INDELs *CRIM1* p.V312_S313insIV (chr2:366917336 G>GCATAGT) and *CRIM1* p.S313Ifs*66 (chr2: 36691742 T> TCATAGGGATGC) listed in Table 11 have been called as *CRIM1* p.S313fs (chr2:36691738 A>ATAGTCATAGTATCCCC) in the single-cell data. *ADGRL4* p.C688F is annotated as *ELTD1* p.C688F in this analysis. Instead of *ASPSCR1* p.V275Pfs*11 as listed in Table 11 the Tapestri pipeline (Mission Bio, v2.0.2) called *ASPSCR1* p.V352fs and *ASPSCR1* p.V352L.

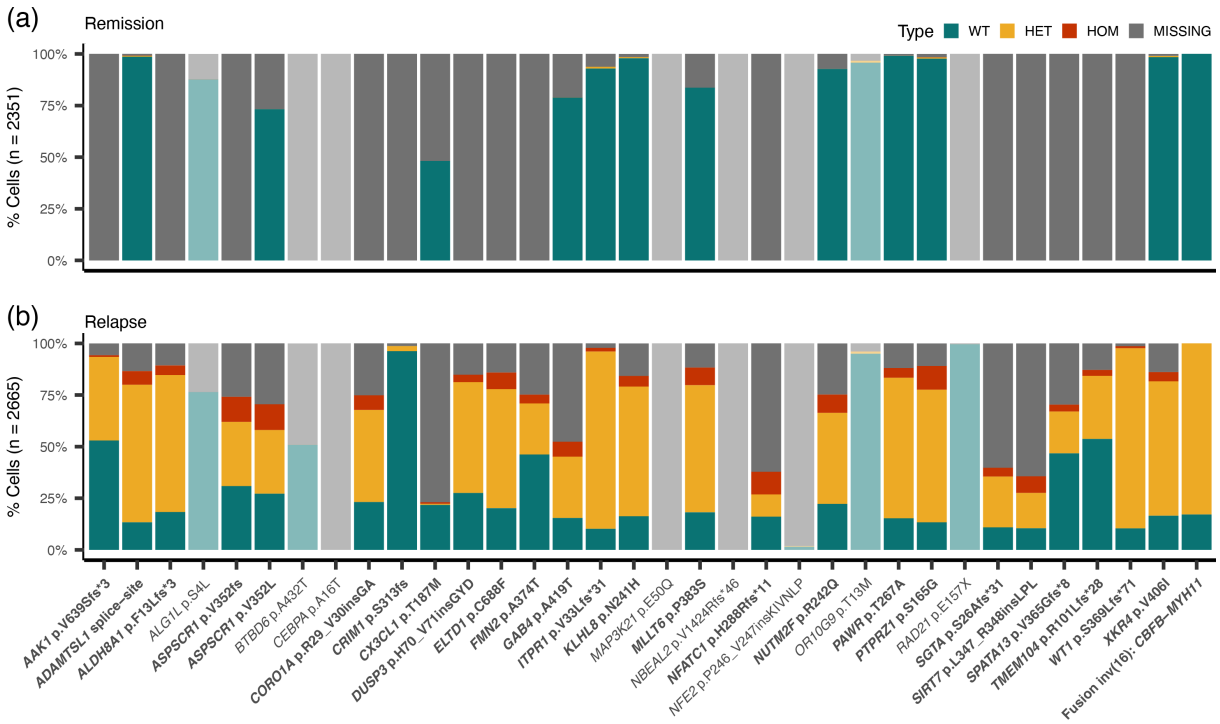

Figure 37: Single-cell genotyping information patient 05. Bar plots show amount of wild-type (WT), heterozygous (HET), homozygous (HOM) and missing classified cells for each variant in the (a) remission and (c) relapse sample. If reads where found in case of the gene fusion *CBFB-MYH11* the cell was classified as heterozygous. Variants in bold were further used for downstream analysis.

Figure 38a shows the inferred tree for patient 05 at diagnosis with 8 tumor clones. Here the founding clone harbors the *CBFB-MYH11* gene fusion. *ASPSCR1* p.V352fs and *ASPSCR1* p.V352L have been called in the same clone and, therefore, I merged them as *ASPSCR1* p.V275Pfs*11, as the variant was annotated in the whole exome sequencing data. The ploidy of each tumor clone and for chromosomes 7, 11, 13, 14 and 15 is visualized in Figure 38b. The deletion on chromosome 7 (q34q36) is already present in the founding clone. To identify the UPD on the q-arm of chromosome 19, I used the percentage of wild-type versus

mutated cells of two SNPs (*i.e.*, rs4254439 and rs8105710) in the region of interest to estimate at which node the UPDs starts. Figure 38c shows that already the founding clone harbors the UPD at chr19q.

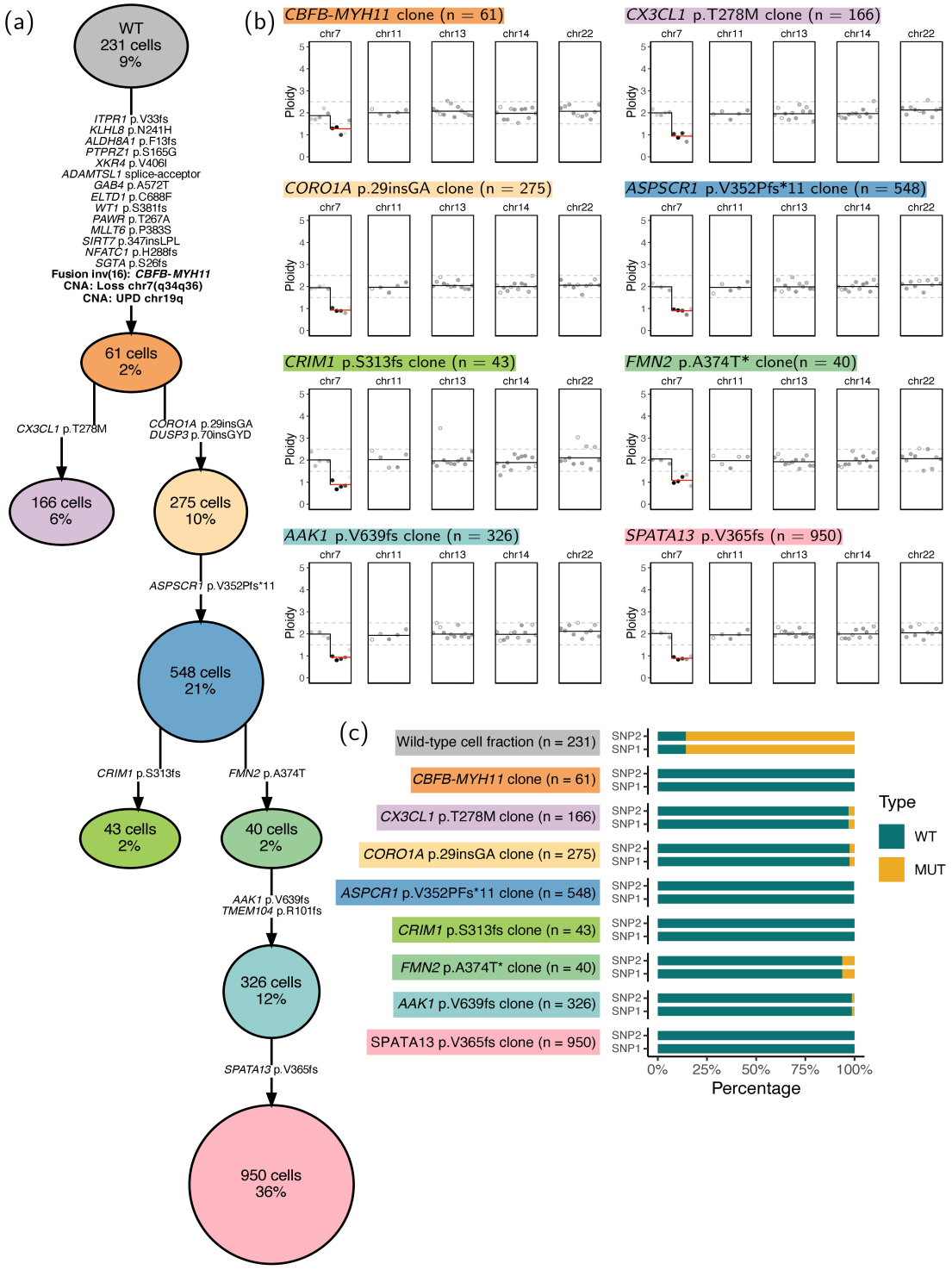

Figure 38: Inferred phylogenetic tree of patient 05 at relapse with ploidies for each tumor clone. (a) Inferred tumor phylogeny from 25 somatic variants and the *CBFB-MYH11* gene fusion. (b) Ploidies for chromosomes 7, 11, 13, 14 and 22 for cells in each tumor clone in comparison to the wild-type cell fraction. (c) Bar plots showing percentages of wild-type cells and mutated cells, which includes cells classified as heterozygous and homozygous, for two single nucleotide polymorphisms (SNPs) (*i.e.*, rs4254439 and rs8105710) in region of uniparental disomy on chromosome 19.

5.3.7 Patient 03

I selected only 12 of 75 somatic variants in patient 03 to reconstruct the tumor phylogeny as shown in Figure 39. For this patient the gene fusion amplicon had no reads and, therefore, no information if the gene fusion is present within a cell could be obtained. The bad quality of the samples continues in the single-cell data, because for this patient I had problems with obtaining high quality variant calls (see section III6.5.3), I was not able to perform copy-number anylsis (see section III6.2) and I had problems with detecting the *CBFB-MYH11* gene fusion breakpoint (see section III6.4).

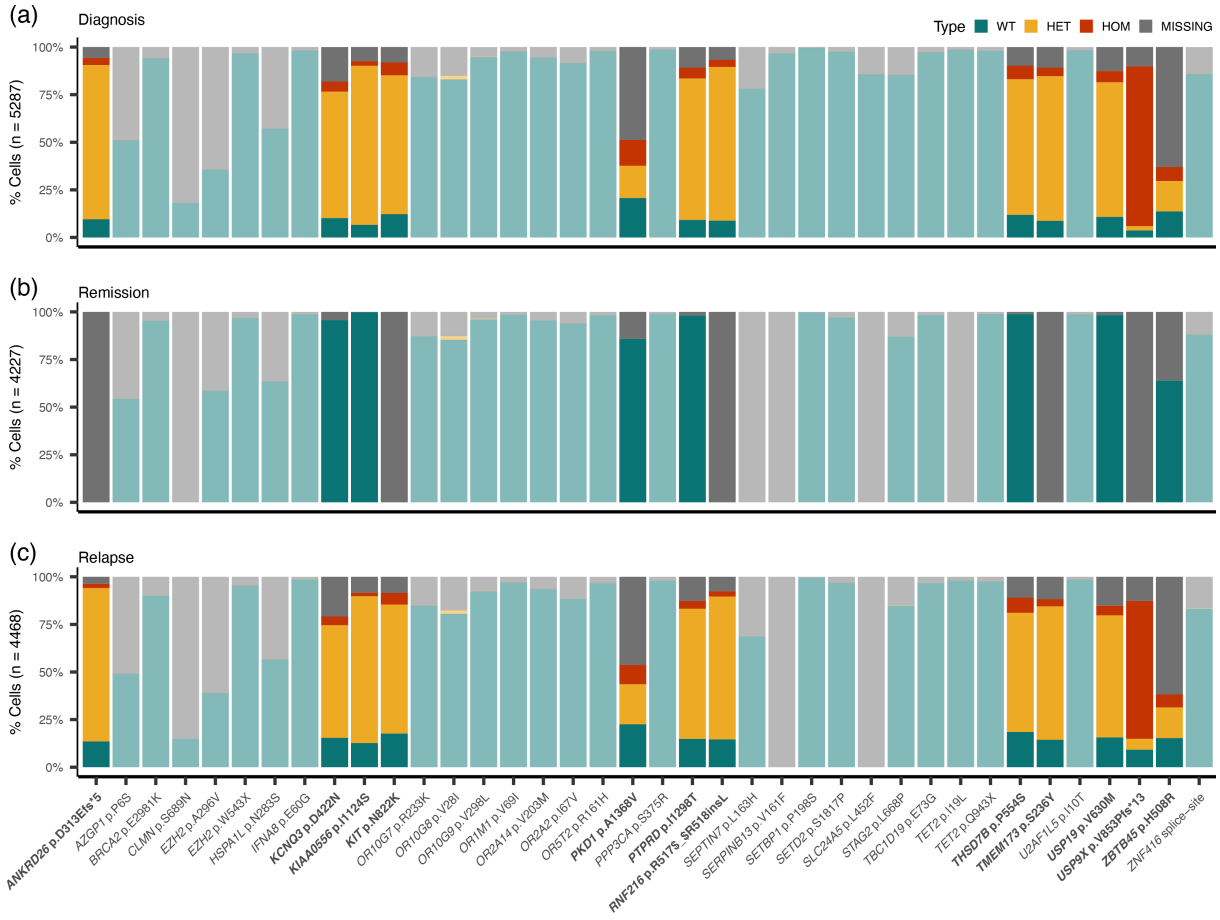

Figure 39: Single-cell genotyping information patient 03. Bar plots show amount of wild-type (WT), heterozygous (HET), homozygous (HOM) and missing classified cells for each variant in the (a) remission and (c) relapse sample. Variants in bold were further used for downstream analysis and variants with only as missing classified cells in all samples were removed.

Figure 40a shows the inferred phylogenetic tree for diagnosis and relapse of patient 03 with 4 tumor clones. Each tumor clone acquires additional somatic variants, but there is no branching event with distinct clones. Due to the fact that only a small fraction of variants were used to reconstruct the tumor phylogeny the real phylogenetic tree might differ. I detected as shown in Figure 40b, a deletion on chromosome 11 that is present from the founding clone onwards. The small amplification of chromosome 8 in the founding clone is not reliable due to

the small clone size (n=15). At first I was not able to call the deletion on chromsome 11, because as stated in the karyotype for this patient (see Table 4) I used the whole chromosome for ploidy calling. Some amplicons had a ploidy <2, but only in a specific region. Then I tried to use the Nanopore data from fusion gene detection to narrow down the region of interest. Here, I executed the "Human variation workflow" (v1.7.1,https://github.com/epi2me-labs) for copy-number analysis from EPI2ME (Oxford Nanopore Technologies) on the aligned reads from section III6.4 using the humanG1Kv37 reference genome [57], `--bam_min_coverage 3`, `--bin_size 500` and default paramters. With this I identified a deletion on chromsome 11 (*i.e.*, chr11:25045000-47305000), but no amplification of chromosome 8.

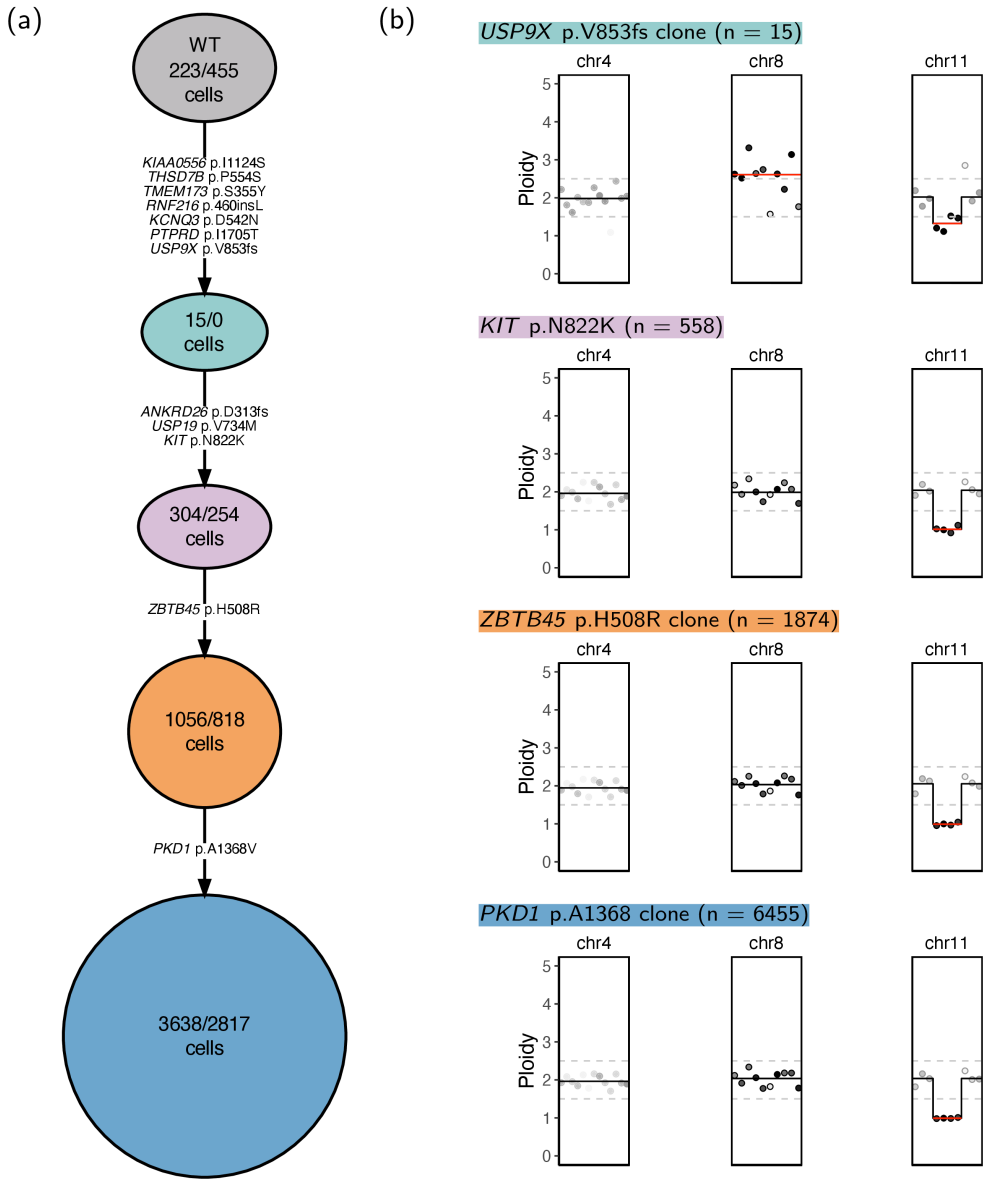

Figure 40: Inferred phylogenetic tree of patient 03 with ploidies for each tumor clone. (a) Inferred tumor phylogeny from 12 somatic variants show a linear tree with 4 tumor clones. (b) Ploidies for chromosomes 4, 8 and 11for cells in each tumor clone in comparison to the wild-type cell fraction.

5.4 Residual tumor clones in complete remission

I used the complete remission sample of each patient to detect remaining tumor cells. Figure 41 shows for each patient with available cells at complete remission (see Table 4) the number of wild-type cells (n_{WT}) and the number of cells that harbor at least one as heterozygous or homozygous classified variant (n_{MUT}). Due to the small number of mutated cells at complete remission it is not possible to reconstruct a clonal hierarchy. Therefore, I matched each mutated cells to a tumor clone in the reconstructed tumor phylogeny as shown in section 5.2 or 5.3 according to the infinite-sites assumption [50]. This means that mutation in a phylogenetic tree can only be gained once and not be lost. Here all missing genotypes were set to wild-type to

obtain a very conservative assumption of the remaining tumor clones. Furthermore, this skews the analysis towards the founding clone, because there are more possibilities (*e.g.*., more somatic variants in the founding clone) for the assignment than to the late tumor clones defined by only one additional acquired somatic variant. This can be seen for patients where the founding clone harbors ≥2 somatic events (*i.e.*, patients 03, 04, 05 and 06) and with more than 38% (ranging from 38.7% in patient 04 to 94.1% in patient 05) of all detected mutated cells assigned to the founding clone. The phylogenetic tree of patient 05 shows more somatic events that are acquired from the wild-type cell fraction to the founding clone than at later stages, therefore, most mutated cells were assigned to the founding clone.

The highest number of remaining cells, I detected in patient 05 with 34 mutated cells and the lowest number of remaining cells with only 4 mutated cells in patient 06. The remaining tumor clone size ranged from 0.16% in patient 06 to 1.54% in patient 09. Interestingly, patient 05 and 09 with both >1% of mutated cells in complete remission also tested positive for the gene fusion in clinical testing at complete remission as shown in Table 4. For patient 01 (Figure 41a) I detected 25 mutated cells that I was able to assign to tumor clones present at relapse. Here, I did not detect any cells harboring diagnosis specific *FLT3* mutations (see Figure 34a). Patient 02 has the shortest RFS of this cohort with only 4 months. For this patient I detected mutated cells that I was able to assign to diagnosis (1/20 cells) and relapse (2/20 cells) specific clone. In case of patient 03, I detected mutated cells that were assigned to all tumor clones (n=4) of the combined phylogenetic tree (see Figure 40). At complete remission of patient 06 I detected only 4 mutated cells.

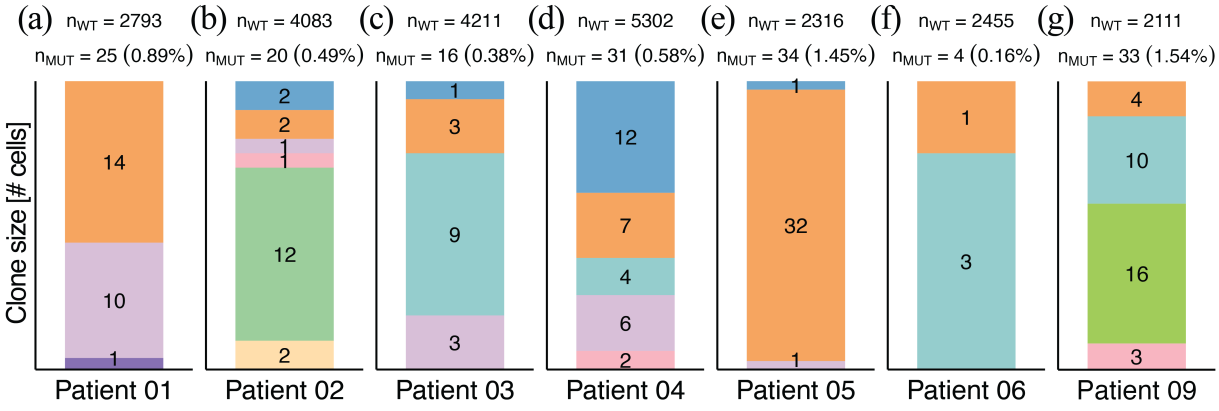

Figure 41: Detected tumor clones in complete remission. Bar plots show for each patient the number of mutated (n_{MUT}) cells and by color their corresponding clone in the phylogenetic tree. Mutated cells are cells that have at least one heterozygous or homozygous somatic variant or the specific gene fusion. The number of wild-type (n_{WT}) cells is stated at the top of each plot. Numbers in brackets are the percentage of mutated cells in the sample.

V Discussion

Patients with CBF AML, defined by the presence of a *RUNX*1-*RUNX1T1* or a *CBFB*-*MYH11* fusion gene, have a high relapse rate of 30-40% within 5 years (RFS <5 years) after standard induction therapy, despite being classified as AML with favorable risk [150]. ITH is a main causal driver of chemotherapy resistance and, furthermore, it has been shown that small subclones already present prior to therapy or acquired during therapy can drive chemotherapy resistance [32,33]. For a better understanding of ITH, it is necessary to identify and order temporally small- and large-scale somatic alterations, which can help to further improve our understanding of tumor development and acquired chemotherapy resistance in leukemia patients.

In this thesis, I performed an integrated analysis of bulk and targeted single-cell DNA sequencing data to uncover ITH and tumor development in a cohort of 9 relapsed CBF AML (*i.e.*, 7 patients with inv(16) AML and 2 patients with t(8;21) AML) patients. To obtain this, I initially used the bulk sequencing data to identify somatic variants (*i.e.*, SNVs and INDELs), SCNAs and CBF specific gene fusions of each patient. With this information I used to prepare targeted panels for generating the scDNA-Seq dataset. For the integrated analysis, I developed a novel method capable of reconstructing the clonal architecture of patients in this cohort using combined information on somatic variants and cytogenetic abnormalities (*i.e.*, SCNAs and fusion genes). To validate the inferred tumor phylogenies, I used information obtained from clinical testing, WES and conventional G-banding.

1 Aim I: Preparation of a combined bulk and single-cell CBF AML dataset

The first aim of this thesis was to generate a comprehensive bulk and single-cell dataset. I used three different bulk-sequencing methods (*i.e.*, whole-exome, targeted DNA and Nanopore sequencing) to obtain mutation and copy-number profiles as well as specific breakpoints of fusion genes for each patient. I developed a pipeline for variant calling that I used for whole-exome and targeted sequencing data. This pipeline was additionally used by me for two other projects of our group (Arends *et al.*, 2022 [19] and Arends *et al.*, 2023 [76]) and was further developed and used by colleagues (Panagiota *et al.*, [77] and Arends *et al.*, 2024 [151]). This pipeline is now an established workflow in our group for identifying CHIP mutations in error-corrected sequencing data of blood derived samples.

I identified 277 somatic variants in total with ranging from 18 to 75 mutations per patient from whole-exome and targeted data combined (listed in section III of this thesis). In general, the tumor content of samples at relapse is lower than at diagnosis and, therefore, the VAF of detected somatic variants (*i.e.*, the mean VAF at diagnosis is 25.0% and the mean VAF at relapse (excluding patient 06) is 13.6%), which might result from the tight check-up schedule recommended by the DGHO guidelines [126]. This also limits the detection of somatic variants and SCNAs in small subclones, due to the smaller fraction of reads carrying the genetic or chromosomal abnormality. Of those 9 patients, 4 have died within 5 years from diagnosis and the remaining were alive at last follow up, which is for patients 05, 06 and 08 more than 6 years after diagnosis.

Patients harbor at least two somatic variants in regions of known AML driver genes [96–98]. In 5 patients (*i.e.*, patients 01, 05, 07, 08 and 09) I identified SCNAs, such as amplification, deletions and UPDs. The mutational landscape of those 9 patients with mutations in *KIT*, *WT1* and *FLT3* as common AML driver mutations and the detected secondary chromosomal abnormalities (*e.g.*, trisomies of chromosome 8 and 22) are comparable with published data [5,14,17,74,96]. Although the patient numbers are too low to constitute a representative cohort with 2 t(8;21) and 7 inv(16) AML patients, *NRAS* mutations are underrepresented in this patient collection compared to other inv(16) cohorts [5,152]. In the cohort of Jahn *et al.*, [5] *NRAS*, which is part of the RTK/RAS signaling, is the most frequent mutated gene in *CBFB-MYH11* type CBF AML patients with the highest number of events in codon Q61. Here, only patient 06 carries a mutation in this region (*i.e.*, *NRAS* p.Q61H) [5]. Fröhling S. *et al.,* [153] have shown that in their AML cohort patients with a mutation in *CEBPA* have a favorable prognosis, which is also true for patient 05 *CEBPA* p.A16T with an OS of 6.6 years. Patient 08 harbors a trisomy of chromosome 22, which is associated with a higher RFS in CBF AML patients. This is not the case for this patient with a RFS of 14 months, however this patient has the second longest OS in this cohort of 6.7years.

The resolution for detecting subclonal copy-number alterations (*e.g.*, trisomies of chromosomes 13, 14 and 22 in patient 04) in whole-exome sequencing data was not sufficient and, therefore, I used additional information from cytogenetics for establishing the targeted single-cell sequencing panel. In total three custom panels (see Table 16) for the Tapestri platform (MissionBio) consisting of approximately 200 amplicons were designed. This was done to keep sequencing costs low, because doubling the amplicon doubles also the amount of necessary paired-end reads (*e.g.*, 200 amplicons = $80x10^6$ paired-end reads per sample, 400 amplicons = $160x10^6$ paired-end reads per sample).

2 Aim II: Integrated analysis of bulk and single-cell data

For the second aim, the integrated analysis of bulk and single-cell sequencing data, I developed a method that enables the detection of copy-number alterations within small subclones of tumor samples. Here, I used parts of COMPASS [64] and further developed it to include gene fusions and to detect SCNAs (*i.e.*, amplifications, deletions and UPDs) within a phylogenetic tree. For patients with single-cell data from diagnosis and relapse I inferred the tumor phylogeny from cells from both samples combined. Initially, I performed the analysis on the individual samples of those patients to be able to check if the combined results, resemble the sample specific trees. The developed method improves existing methods and more importantly enables the detection of SCNAs that were not detected using existing tools (*i.e.*, ConDoR [50] and COMPASS [64]). This I managed by using the fraction of wild-type cells of each phylogenetic tree to call ploidies of tumor clones.

Because the data is very sparse and noisy, I reduced the cells for estimating ploidies of a region using the Z-score as a measure to remove noisy amplicons of cells within a clone. Sashittal *et al.*, [50] cluster cells based on their copy-number profiles prior to inferring tumor phylogeny and Zhang *et al.* [154] place SNVs on a copy-number tree , which was not possible for patients in this cohort due to noise in the single-cell data. This might result from the different coverage for each amplicon and, moreover, the variability of these amplicons within the cells of a sample, as pointed out by Sollier *et al.*, [64]. The panels used in this thesis are custom panels and may not be as thoroughly tested and optimized as off-the shelf panels. Here, primers were included that might not work, but the possibility to have an amplicon that covers a region of a somatic event is more important than perfect amplicons. It has to be noted that a single-cell is not a complete single-cell, but rather a snapshot of parts of a single-cell and when combining similar cells the power to retrieve information increases [155]. This can be seen throughout ploidies of tumor clones with larger clones having less variance in their estimation (*e.g.*, ploidies of *CBFB-MYH11* clone versus *FLT3* p.A680V clone of patient 08 at diagnosis shown in Figure 27).

Morita *et al.*, [31] classified patients based on a linear and branching tumor evolution, in contrast, all patient in this cohort (except for patient 03) show a branching phylogenetic tree. This might result from using a targeted panel specifically designed for each patient instead of an off-the-shelf panel with a limited number of amplicons overlapping patient specific mutations. This allows to infer tumor phylogenies from a larger number of somatic events per patient and, therefore, lead to a more complex phylogenetic tree. Patient 07 has the most

complex phylogenetic tree in this cohort with several distinct clones harboring variants in genes involved in RAS/RTK and also the shortest OS.

For all patients, except for patient 03, I could identify gene fusions in single-cells and, further show that in 7 of 8 patients the *RUNX1-RUNX1T1* or *CBFB-MYH11* gene fusion is already present in the founding clone. Patient 03 did show poor quality throughout each bulk sequencing method, but due to limited number of patients I did not exclude this patient from single-cell analysis. In CBF AML the gene fusion inhibits affected cells to differentiate and secondary variants lead to a proliferative advantage and to an accelerated proliferation of blasts, confirming these results [150]. That the founding clone harbors the CBF gene fusion supports the effectiveness of clinical testing of minimal residual disease using real-time quantitative PCR with a sensitivity of 10^4-10^5 [126]. Moreover, it is important to identify additional somatic variants of the founding clone, because it had been shown that for AML patients with late relapse (>5 years) the founding clone persisted and gave rise to relapse [156]. Patients 01, 02 and 03 with samples at diagnosis and relapse also support this with the founding clone and a large portion of the tree persisting throughout the course of the disease. Schwede *et al.*, [35] pointed out that it is necessary to investigate the mutational order of chromosomal abnormalities in combination with somatic variants in AML patients. They further explained that if using a FLT3 inhibitor as therapy for a patient with a *FLT3* mutations and inv(16) it is important to know which is the initiating event. Furthermore, they pointed out that also for patients with *IDH1/2* and *FLT3* mutations it is important to know the clonal order, due to using a targeted therapy that targets the gene with the earlier mutations, so that no clones remain after treatment. This would be the case for patient 01 harboring *FLT3* and *IDH2* mutations (see Table 7), but unfortunately the amplicon for *IDH2* p.R18P in scDNA-Seq failed. Here, the comparison of VAFs between *IDH2* p.R18P and *FLT3* p.D835Y with 9.4% and 3.7%, respectively, suggests the *IDH2* variant was acquired earlier or in distinct clones. Due to the limitation of identifying ITH from low level VAF mutations in bulk sequencing data, it is not possible to decide for one possibility [45]. The advantage of scDNA-Seq can be shown in patient 07, where I detected a CHIP clone independently of the AML clone persisting throughout therapy detected at diagnosis and relapse. Hirsch *et al.*, [28] also found that in several patients of their cohort repeated chemotherapy had no impact on CHIP clones. In line with this finding is the relative increase of the CHIP clone in the relapse sample of patient 07 (see Figure 31), which must be considered with caution due to the limited number of cells sequenced at relapse.

The tumor phylogenies of patient 04 and 08 with inv(16) AML show relapse specific *WT1* mutations, which might disrupt the immunogenic potential of WT1 and drive immune escape

after allogenic stem cell transplantation [157]. That *WT1* mutations are acquired during the disease has also been shown in other sequencing studies [158,159]. In patients with multiple mutations in genes involved in RAS/RTK they are located in distinct clones, which has been shown previously in bulk data [14,17] and was recently confirmed in single-cell data [35]. Here, patient 01 harbors two *FLT3* variants (*i.e.*, FLT3 p.D835V and FLT3 p.D835Y) and patient 02 harbors one *FLT3*-ITD and one *KIT p.*D816V mutation in two distinct clones. Patient 07 harbors 5 mutations in genes involved in RAS/RTK signaling genes that are in 4 distinct clones (*i.e.*, *FLT3 p.*D835Y, *KIT* p.D419del, *NRAS* p.G12A and *KIT* p.D816Y with *NF1* p.P2289Sfs*17 in the same clone). Itzykson *et al.*, [160] have shown that CBF AML patients with clonal interference (clonal heterogeneity) of signaling genes (*i.e.*, *KIT*, *NRAS*, *KRAS*, *FLT3*, *JAK2* and *CBL*), which is the case in patients 01, 02 and 07, have a significant lower EFS than only a single clone. Here, patients 02 and 07 with a RFS of 4 and 9 months, respectively, support this finding. In contrast, patient 01 has with a RFS of 24 months the second longest RFS of this cohort.

For the investigation of remaining tumor cells at complete remission I used the inferred tumor phylogeny of each patient (inferred tree from diagnosis, relapse or combined) to assign each cell to a tumor clone using the infinite-sites assumption. Due to the small number of mutated cells (4-34 cells) and sparse information of single-cell data it was not possible to infer a phylogenetic tree at complete remission. Here, I can show that using somatic events additional to gene fusion elevates the number of detected mutated cells at complete remission. Patient 05 and 09 with the highest percentage of mutated cells at complete remission (*i.e.*, 1.45% and 1.54%) also tested positive in clinical testing for the gene fusion at complete remission as shown in Table 4.

3 Aim III: Validation

I validated the inferred phylogenetic trees including copy-number alterations using existing bulk and clinical information of each patient. Initially, I compared the number of mutated cells in the single-cell data and the estimated clone sizes of the inferred phylogenetic trees with VAFs from WES. This step was a reason for using COMPASS [64] instead of ConDoR [50], because the estimated wild-type cell fraction from ConDoR did not match bulk and clinical data. In detail, the estimated fraction of wild-type cells was smaller than what would be assumed from WES, which is negatively affecting the estimation of ploidies, because the wild-type fraction is used as a reference. Here, I used the somatic variant with the highest VAF of a sample or the percentage of blasts for each sample (see Table 4) to estimate the percentage

of the wild-type fraction. My developed method is not relying on COMPASS for inferring copy-number alterations in subclones of a phylogenetic, it is only necessary to have cell barcodes for all tumor clones and the wild-type fraction of a phylogenetic tree.

The *FLT3*-ITD at diagnosis of patient 02 was estimated by clinical testing with an AR of 13% (see Table 3), in WES data I identified it with a VAF of 12.7% (see Table 8) and the fraction of mutated cells versus total cells of single-cell genotyping is approximately 29.7% (see Figure 24). The single-cell estimation is higher than just doubling the VAF from bulk because cells are counted as heterozygous if the VAF within a cell is >20%. This shows that the single-cell data is comparable with bulk and clinical data. In case, of patient 04 I identified a tumor clone harboring a trisomy of chromosomes 13, 14 and 22 with a clone size of 24% matching the metaphases with this abnormality by karyotype (13/49 ~27%). This highlights the capability of my developed method to return reliable results and, as represented by these SCNAs in patient 04, the possibility to detect SCNAs that were not detected by existing methods [64].

4 Conclusion

To conclude, I developed a method that improves our understanding of ITH in AML patients by uncovering copy-number alterations in small subclones that current methods were not able to detect. Moreover, establishing and analyzing patients with targeted panels that cover the mutation and copy-number landscape including the possibility to track CBF gene fusions in single-cell for each patient in this cohort has never been done before. For patients with cells at diagnosis and relapse (*i.e.*, 01, 02 and 03) I performed a combined analysis showing the change in clonal composition under the pressure of intensive chemotherapy.

The developed method in this thesis uncovers tumor development and improves available methods in a reproducible manner. Additionally, it can be easily applied with any tool that infers tumor phylogeny by only using R and Python. Furthermore, this method can be applied on available datasets with a priori information on copy-numbers.

Additionally, a larger patient cohort would help to better understand inter-tumor heterogeneity of CBF AML and, further, help to understand acquired chemotherapy resistance. When using T-cells as a germline reference, as I did for patient 07 and 08, information could be drawn already from the diagnosis sample. Copy-number and fusion gene breakpoints could be retrieved also from Nanopore sequencing with higher depth. Here, maybe targeted panels from patient specific driver, leukemia related mutations and, if diagnosis and relapse samples are available, from somatic variants present at two timepoints can be used to establish a targeted

panel. This might limit the number of mutations per patient and make this a more labor- and cost-effective approach for future studies with larger patient cohorts. A limitation of this study is the missing information on cell types within a sample. Here, studies using surface protein markers or gene expression as an additional data layer would be helpful to gain further insights on tumor development. Especially, when investigating preleukemic mutations and therapy resistance the cell type information would be of high interest.

Bibliography

[1] German Centre for Cancer Registry Data, Robert Koch Institute: Database Query with estimates for cancer incidence, prevalence and survival in Germany, based on data of the population based cancer registries (DOI: 110.18444/5.03.01.0005.0017.0001 [Inzidenz, Prävalenz]; DOI: 10.18444/5.03.01.0005.0016.0001 [survival]). Mortality data provided by the Federal Statistical Office, (2022). www.krebsdaten.de/database (accessed February 25, 2024).

[2] F. Erdmann, C. Spix, A. Katalinic, M. Christ, J. Folkerts, J. Hansmann, K. Kranzhöfer, B. Kunz, K. Manegold, A. Penzkofer, K. Treml, G. Vollmer, S. Weg-Remers, B. Barnes, N. Buttman-Schweiger, S. Dahm, J. Fiebig, M. Franke, I. Gurung-Schönfeld, J. Haberland, M. Imhoff, K. Kraywinkel, A. Starker, P. von Berenberg-Gossler, A. Wienecke, Cancer in Germany 2017/2018, (2022) 170.

[3] H. Döhner, D.J. Weisdorf, C.D. Bloomfield, Acute Myeloid Leukemia, N. Engl. J. Med. 373 (2015) 1136–1152.

[4] E. Peroni, M.L. Randi, A. Rosato, S. Cagnin, Acute myeloid leukemia: from NGS, through scRNA-seq, to CAR-T. dissect cancer heterogeneity and tailor the treatment, J. Exp. Clin. Cancer Res. 42 (2023) 1–22.

[5] N. Jahn, T. Terzer, E. Sträng, A. Dolnik, S. Cocciardi, E. Panina, A. Corbacioglu, J. Herzig, D. Weber, A. Schrade, K. Götze, T. Schröder, M. Lübbert, D. Wellnitz, E. Koller, R.F. Schlenk, V.I. Gaidzik, P. Paschka, F.G. Rücker, M. Heuser, F. Thol, A. Ganser, A. Benner, H. Döhner, L. Bullinger, K. Döhner, Genomic heterogeneity in core-binding factor acute myeloid leukemia and its clinical implication, Blood Adv. 4 (2020) 6342–6352.

[6] H. Döhner, A.H. Wei, F.R. Appelbaum, C. Craddock, C.D. DiNardo, H. Dombret, B.L. Ebert, P. Fenaux, L.A. Godley, R.P. Hasserjian, R.A. Larson, R.L. Levine, Y. Miyazaki, D. Niederwieser, G. Ossenkoppele, C. Röllig, J. Sierra, E.M. Stein, M.S. Tallman, H.-F. Tien, J. Wang, A. Wierzbowska, B. Löwenberg, Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN, Blood 140 (2022) 1345–1377.

[7] J.D. Khoury, E. Solary, O. Abla, Y. Akkari, R. Alaggio, J.F. Apperley, R. Bejar, E. Berti, L. Busque, J.K.C. Chan, W. Chen, X. Chen, W.-J. Chng, J.K. Choi, I. Colmenero, S.E. Coupland, N.C.P. Cross, D. De Jong, M.T. Elghetany, E. Takahashi, J.-F. Emile, J. Ferry, L. Fogelstrand, M. Fontenay, U. Germing, S. Gujral, T. Haferlach, C. Harrison, J.C. Hodge, S. Hu, J.H. Jansen, R. Kanagal-Shamanna, H.M. Kantarjian, C.P. Kratz, X.-Q. Li, M.S. Lim, K. Loeb, S. Loghavi, A. Marcogliese, S. Meshinchi, P. Michaels, K.N. Naresh, Y. Natkunam, R. Nejati, G. Ott, E. Padron, K.P. Patel, N. Patkar, J. Picarsic, U. Platzbecker, I. Roberts, A. Schuh, W. Sewell, R. Siebert, P. Tembhare, J. Tyner, S. Verstovsek, W. Wang, B. Wood, W. Xiao, C. Yeung, A. Hochhaus, The 5th edition of the World Health Organization classification of Haematolymphoid Tumours: Myeloid and histiocytic/dendritic neoplasms, Leukemia 36 (2022) 1703–1719.

[8] S.-J. Chen, Y. Shen, Z. Chen, A panoramic view of acute myeloid leukemia, Nat. Genet. 45 (2013) 586–587.

[9] R. Juskevicius, M.A. Thompson, A. Shaver, D. Head, Clinical presentation, diagnosis, and classification of acute myeloid leukemia, in: Acute Leukemias, Springer International Publishing, Cham, 2021: pp. 11–55.

[10] C.D. DiNardo, H.P. Erba, S.D. Freeman, A.H. Wei, Acute myeloid leukaemia, Lancet 401 (2023) 2073–2086.

[11] C. Negotei, A. Colita, I. Mitu, A.R. Lupu, M.-E. Lapadat, C.E. Popovici, M. Crainicu, O. Stanca, N.M. Berbec, A review of FLT3 kinase inhibitors in AML, J. Clin. Med. 12 (2023). https://doi.org/10.3390/jcm12206429.

[12] N.A. Speck, D.G. Gilliland, Core-binding factors in haematopoiesis and leukaemia, Nat. Rev. Cancer 2 (2002) 502–513.

[13] R. Alaggio, C. Amador, I. Anagnostopoulos, A.D. Attygalle, I.B. de O. Araujo, E. Berti, G. Bhagat, A.M. Borges, D. Boyer, M. Calaminici, A. Chadburn, J.K.C. Chan, W. Cheuk, W.-J. Chng, J.K. Choi, S.-S. Chuang, S.E. Coupland, M. Czader, S.S. Dave, D. de Jong, M.-Q. Du, K.S. Elenitoba-Johnson, J. Ferry, J. Geyer, D. Gratzinger, J. Guitart, S. Gujral, M. Harris, C.J. Harrison, S. Hartmann, A. Hochhaus, P.M. Jansen, K. Karube, W. Kempf, J. Khoury, H. Kimura, W. Klapper, A.E. Kovach, S. Kumar, A.J. Lazar, S. Lazzi, L. Leoncini, N. Leung, V. Leventaki, X.-Q. Li, M.S. Lim, W.-P. Liu, A. Louissaint Jr, A. Marcogliese, L.J. Medeiros, M. Michal, R.N. Miranda, C. Mitteldorf, S. Montes-Moreno, W. Morice, V. Nardi, K.N. Naresh, Y. Natkunam, S.-B. Ng, I. Oschlies, G. Ott, M. Parrens, M. Pulitzer, S.V. Rajkumar, A.C. Rawstron, K. Rech, A. Rosenwald, J. Said, C. Sarkozy, S. Sayed, C. Saygin, A. Schuh, W. Sewell, R. Siebert, A.R. Sohani, R. Tooze, A. Traverse-Glehen, F. Vega, B. Vergier, A.D. Wechalekar, B. Wood, L. Xerri, W. Xiao, The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid neoplasms, Leukemia 36 (2022) 1720–1748.

[14] F. Christen, K. Hoyer, K. Yoshida, H.-A. Hou, N. Waldhueter, M. Heuser, R.K. Hills, W. Chan, R. Hablesreiter, O. Blau, Y. Ochi, P. Klement, W.-C. Chou, I.-W. Blau, J.-L. Tang, T. Zemojtel, Y. Shiraishi, Y. Shiozawa, F. Thol, A. Ganser, B. Löwenberg, D.C. Linch, L. Bullinger, P.J.M. Valk, H.-F. Tien, R.E. Gale, S. Ogawa, F. Damm, Genomic landscape and clonal evolution of acute myeloid leukemia with t(8;21): an international study on 331 patients, Blood 133 (2019) 1140–1151.

[15] S. Opatz, S.A. Bamopoulos, K.H. Metzeler, T. Herold, B. Ksienzyk, K. Bräundl, S. Tschuri, S. Vosberg, N.P. Konstandin, C. Wang, L. Hartmann, A. Graf, S. Krebs, H. Blum, S. Schneider, C. Thiede, J.M. Middeke, F. Stölzel, C. Röllig, J. Schetelig, G. Ehninger, A. Krämer, J. Braess, D. Görlich, M.C. Sauerland, W.E. Berdel, B.J. Wörmann, W. Hiddemann, K. Spiekermann, S.K. Bohlander, P.A. Greif, The clinical mutatome of core binding factor leukemia, Leukemia 34 (2020) 1553–1562.

[16] J.P. Patel, M. Gönen, M.E. Figueroa, H. Fernandez, Z. Sun, J. Racevskis, P. Van Vlierberghe, I. Dolgalev, S. Thomas, O. Aminova, K. Huberman, J. Cheng, A. Viale, N.D. Socci, A. Heguy, A. Cherry, G. Vance, R.R. Higgins, R.P. Ketterling, R.E. Gallagher, M. Litzow, M.R.M. van den Brink, H.M. Lazarus, J.M. Rowe, S. Luger, A. Ferrando, E. Paietta, M.S. Tallman, A. Melnick, O. Abdel-Wahab, R.L. Levine, Prognostic relevance of integrated genetic profiling in acute myeloid leukemia, N. Engl. J. Med. 366 (2012) 1079–1089.

[17] S.Y. Han, K. Mrózek, J. Voutsinas, Q. Wu, E.A. Morgan, H. Vestergaard, R. Ohgami, P.M. Kluin, T.K. Kristensen, S. Pullarkat, M.B. Møller, A.-I. Schiefer, L.B. Baughn, Y. Kim, D. Czuchlewski, J.R. Hilberink, H.-P. Horny, T.I. George, M. Dolan, N.K. Ku, C. Arana Yi, V. Pullarkat, J. Kohlschmidt, A. Salhotra, L. Soma, C.D. Bloomfield, D. Chen, W.R. Sperr, G. Marcucci, C. Cho, C. Akin, J. Gotlib, S. Broesby-Olsen, M. Larson, M.A. Linden, H.J. Deeg, G. Hoermann, M.-A. Perales, J.L. Hornick, M.R. Litzow, R. Nakamura, D. Weisdorf, G. Borthakur, G. Huls, P. Valent, C. Ustun, C.C.S. Yeung, Secondary cytogenetic abnormalities in core-binding factor AML harboring inv(16) vs t(8;21), Blood Adv. 5 (2021) 2481–2489.

[18] S. Jaiswal, B.L. Ebert, Clonal hematopoiesis in human aging and disease, Science 366 (2019) eaan4673.

[19] C.M. Arends, S. Dimitriou, A. Stahler, R. Hablesreiter, P.M. Strzelecka, C.M. Stein, M. Tilgner, R. Saiki, S. Ogawa, L. Bullinger, D.P. Modest, S. Stintzing, V. Heinemann, F. Damm, Clonal hematopoiesis is associated with improved survival in patients with metastatic colorectal cancer from the FIRE-3 trial, Blood 139 (2022) 1593–1597.

[20] H. Ahmad, N. Jahn, S. Jaiswal, Clonal hematopoiesis and its impact on human health, Annu. Rev. Med. 74 (2023) 249–260.
[21] G. Genovese, A.K. Kähler, R.E. Handsaker, J. Lindberg, S.A. Rose, S.F. Bakhoum, K. Chambert, E. Mick, B.M. Neale, M. Fromer, S.M. Purcell, O. Svantesson, M. Landén, M. Höglund, S. Lehmann, S.B. Gabriel, J.L. Moran, E.S. Lander, P.F. Sullivan, P. Sklar, H. Grönberg, C.M. Hultman, S.A. McCarroll, Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence, N. Engl. J. Med. 371 (2014) 2477–2487.
[22] S. Jaiswal, P. Fontanillas, J. Flannick, A. Manning, P.V. Grauman, B.G. Mar, R.C. Lindsley, C.H. Mermel, N. Burtt, A. Chavez, J.M. Higgins, V. Moltchanov, F.C. Kuo, M.J. Kluk, B. Henderson, L. Kinnunen, H.A. Koistinen, C. Ladenvall, G. Getz, A. Correa, B.F. Banahan, S. Gabriel, S. Kathiresan, H.M. Stringham, M.I. McCarthy, M. Boehnke, J. Tuomilehto, C. Haiman, L. Groop, G. Atzmon, J.G. Wilson, D. Neuberg, D. Altshuler, B.L. Ebert, Age-related clonal hematopoiesis associated with adverse outcomes, N. Engl. J. Med. 371 (2014) 2488–2498.
[23] M. Xie, C. Lu, J. Wang, M.D. McLellan, K.J. Johnson, M.C. Wendl, J.F. McMichael, H.K. Schmidt, V. Yellapantula, C.A. Miller, B.A. Ozenberger, J.S. Welch, D.C. Link, M.J. Walter, E.R. Mardis, J.F. Dipersio, F. Chen, R.K. Wilson, T.J. Ley, L. Ding, Age-related mutations associated with clonal hematopoietic expansion and malignancies, Nat. Med. 20 (2014) 1472–1478.
[24] L. Busque, J.P. Patel, M.E. Figueroa, A. Vasanthakumar, S. Provost, Z. Hamilou, L. Mollica, J. Li, A. Viale, A. Heguy, M. Hassimi, N. Socci, P.K. Bhatt, M. Gonen, C.E. Mason, A. Melnick, L.A. Godley, C.W. Brennan, O. Abdel-Wahab, R.L. Levine, Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis, Nat. Genet. 44 (2012) 1179–1181.
[25] F. Christen, R. Hablesreiter, K. Hoyer, C. Hennch, A. Maluck-Böttcher, A. Segler, A. Madadi, M. Frick, L. Bullinger, F. Briest, F. Damm, Modeling clonal hematopoiesis in umbilical cord blood cells by CRISPR/Cas9, Leukemia 36 (2022) 1102–1110.
[26] A. Niroula, A. Sekar, M.A. Murakami, M. Trinder, M. Agrawal, W.J. Wong, A.G. Bick, M.M. Uddin, C.J. Gibson, G.K. Griffin, M.C. Honigberg, S.M. Zekavat, K. Paruchuri, P. Natarajan, B.L. Ebert, Distinction of lymphoid and myeloid clonal hematopoiesis, Nat. Med. 27 (2021) 1921–1927.
[27] C.-W. Chen, L. Zhang, R. Dutta, A. Niroula, P.G. Miller, C.J. Gibson, A.G. Bick, J.M. Reyes, Y.-T. Lee, A. Tovy, T. Gu, S. Waldvogel, Y.-H. Chen, B.J. Venters, P.-O. Estève, S. Pradhan, M.-C. Keogh, P. Natarajan, K. Takahashi, A.S. Sperling, M.A. Goodell, SRCAP mutations drive clonal hematopoiesis through epigenetic and DNA repair dysregulation, Cell Stem Cell 30 (2023) 1503-1519.e8.
[28] P. Hirsch, R. Tang, N. Abermil, P. Flandrin, H. Moatti, F. Favale, L. Suner, F. Lorre, C. Marzac, F. Fava, A.-C. Mamez, S. Lapusan, F. Isnard, M. Mohty, O. Legrand, L. Douay, C. Bilhou-Nabera, F. Delhommeau, Precision and prognostic value of clone-specific minimal residual disease in acute myeloid leukemia, Haematologica 102 (2017) 1227–1237.
[29] P.C. Nowell, The clonal evolution of tumor cell populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression, Science 194 (1976) 23–28.
[30] A. Laganà, Computational approaches for the investigation of intra-tumor heterogeneity and clonal evolution from bulk sequencing data in precision oncology applications, Adv. Exp. Med. Biol. 1361 (2022) 101–118.
[31] K. Morita, F. Wang, K. Jahn, T. Hu, T. Tanaka, Y. Sasaki, J. Kuipers, S. Loghavi, S.A. Wang, Y. Yan, K. Furudate, J. Matthews, L. Little, C. Gumbs, J. Zhang, X. Song, E. Thompson, K.P. Patel, C.E. Bueso-Ramos, C.D. DiNardo, F. Ravandi, E. Jabbour, M. Andreeff, J. Cortes, K. Bhalla, G. Garcia-Manero, H. Kantarjian, M. Konopleva, D.

Nakada, N. Navin, N. Beerenwinkel, P.A. Futreal, K. Takahashi, Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics, Nat. Commun. 11 (2020) 5327.

[32] M. Tarabichi, A. Salcedo, A.G. Deshwar, M. Ni Leathlobhair, J. Wintersinger, D.C. Wedge, P. Van Loo, Q.D. Morris, P.C. Boutros, A practical guide to cancer subclonal reconstruction from DNA sequencing, Nat. Methods 18 (2021) 144–155.

[33] N. McGranahan, C. Swanton, Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future, Cell 168 (2017) 613–628.

[34] J.M. Klco, C.A. Miller, M. Griffith, A. Petti, D.H. Spencer, S. Ketkar-Kulkarni, L.D. Wartman, M. Christopher, T.L. Lamprecht, N.M. Helton, E.J. Duncavage, J.E. Payton, J. Baty, S.E. Heath, O.L. Griffith, D. Shen, J. Hundal, G.S. Chang, R. Fulton, M. O'Laughlin, C. Fronick, V. Magrini, R.T. Demeter, D.E. Larson, S. Kulkarni, B.A. Ozenberger, J.S. Welch, M.J. Walter, T.A. Graubert, P. Westervelt, J.P. Radich, D.C. Link, E.R. Mardis, J.F. DiPersio, R.K. Wilson, T.J. Ley, Association between mutation clearance after induction therapy and outcomes in acute myeloid leukemia, JAMA 314 (2015) 811.

[35] M. Schwede, K. Jahn, J. Kuipers, L.A. Miles, R.L. Bowman, T. Robinson, K. Furudate, H. Uryu, T. Tanaka, Y. Sasaki, A. Ediriwickrema, B. Benard, A.J. Gentles, R. Levine, N. Beerenwinkel, K. Takahashi, R. Majeti, Mutation order in acute myeloid leukemia identifies uncommon patterns of evolution and illuminates phenotypic heterogeneity, Leukemia (2024) 1–10.

[36] L.Y. Liu, V. Bhandari, A. Salcedo, S.M.G. Espiritu, Q.D. Morris, T. Kislinger, P.C. Boutros, Quantifying the influence of mutation detection on tumour subclonal reconstruction, Nat. Commun. 11 (2020) 6247.

[37] S.C. Dentro, D.C. Wedge, P. Van Loo, Principles of reconstructing the subclonal architecture of cancers, Cold Spring Harb. Perspect. Med. 7 (2017) a026625.

[38] C.A. Miller, B.S. White, N.D. Dees, M. Griffith, J.S. Welch, O.L. Griffith, R. Vij, M.H. Tomasson, T.A. Graubert, M.J. Walter, M.J. Ellis, W. Schierding, J.F. DiPersio, T.J. Ley, E.R. Mardis, R.K. Wilson, L. Ding, SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution, PLoS Comput. Biol. 10 (2014) e1003665.

[39] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, S.P. Shah, PyClone: statistical inference of clonal population structure in cancer, Nat. Methods 11 (2014) 396–398.

[40] S. Gillis, A. Roth, PyClone-VI: scalable inference of clonal population structures using whole genome data, BMC Bioinformatics 21 (2020) 571.

[41] A.G. Deshwar, S. Vembu, C.K. Yung, G.H. Jang, L. Stein, Q. Morris, PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors, Genome Biol. 16 (2015) 35.

[42] X. Fu, H. Lei, Y. Tao, R. Schwartz, Reconstructing tumor clonal lineage trees incorporating single-nucleotide variants, copy number alterations and structural variations, Bioinformatics 38 (2022) i125–i133.

[43] H.X. Dang, B.S. White, S.M. Foltz, C.A. Miller, J. Luo, R.C. Fields, C.A. Maher, ClonEvol: clonal ordering and visualization in cancer sequencing, Ann. Oncol. 28 (2017) 3076–3082.

[44] C.A. Miller, J. McMichael, H.X. Dang, C.A. Maher, L. Ding, T.J. Ley, E.R. Mardis, R.K. Wilson, Visualizing tumor evolution with the fishplot package for R, BMC Genomics 17 (2016) 880.

[45] G.D. Evrony, A.G. Hinch, C. Luo, Applications of single-cell DNA sequencing, Annu. Rev. Genomics Hum. Genet. 22 (2021) 171–197.

[46] S. Cocciardi, A. Dolnik, S. Kapp-Schwoerer, F.G. Rücker, S. Lux, T.J. Blätte, S. Skambraks, J. Krönke, F.H. Heidel, T.M. Schnöder, A. Corbacioglu, V.I. Gaidzik, P.

Paschka, V. Teleanu, G. Göhring, F. Thol, M. Heuser, A. Ganser, D. Weber, E. Sträng, H.A. Kestler, H. Döhner, L. Bullinger, K. Döhner, Clonal evolution patterns in acute myeloid leukemia with NPM1 mutation, Nat. Commun. 10 (2019) 2031.

[47] J. Kuipers, K. Jahn, N. Beerenwinkel, Advances in understanding tumour evolution through single-cell sequencing, Biochim. Biophys. Acta Rev. Cancer 1867 (2017) 127–138.

[48] T. Kader, M. Zethoven, K.L. Gorringe, Evaluating statistical approaches to define clonal origin of tumours using bulk DNA sequencing: context is everything, Genome Biol. 23 (2022) 43.

[49] D. Lähnemann, J. Köster, E. Szczurek, D.J. McCarthy, S.C. Hicks, M.D. Robinson, C.A. Vallejos, K.R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C.S.-O. Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. de Barbanson, A. Cappuccio, G. Corleone, B.E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T.J. Lobo, E.M. Keizer, I. Khatri, S.M. Kielbasa, J.O. Korbel, A.M. Kozlov, T.-H. Kuo, B.P.F. Lelieveldt, I.I. Mandoiu, J.C. Marioni, T. Marschall, F. Mölder, A. Niknejad, L. Raczkowski, M. Reinders, J. de Ridder, A.-E. Saliba, A. Somarakis, O. Stegle, F.J. Theis, H. Yang, A. Zelikovsky, A.C. McHardy, B.J. Raphael, S.P. Shah, A. Schönhuth, Eleven grand challenges in single-cell data science, Genome Biol. 21 (2020) 31.

[50] P. Sashittal, H. Zhang, C.A. Iacobuzio-Donahue, B.J. Raphael, ConDoR: tumor phylogeny inference with a copy-number constrained mutation loss model, Genome Biol. 24 (2023) 272.

[51] C. Gawad, W. Koh, S.R. Quake, Single-cell genome sequencing: current state of the science, Nat. Rev. Genet. 17 (2016) 175–188.

[52] N.E. Navin, Cancer genomics: one cell at a time, Genome Biol. 15 (2014) 452.

[53] G. Satas, S. Zaccaria, G. Mon, B.J. Raphael, SCARLET: Single-cell tumor phylogeny inference with copy-number constrained mutation losses, Cell Syst. 10 (2020) 323-332.e8.

[54] K. Jahn, J. Kuipers, N. Beerenwinkel, Tree inference for single-cell data, Genome Biol. 17 (2016) 86.

[55] J. Kuipers, K. Jahn, B.J. Raphael, N. Beerenwinkel, Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors, Genome Res. 27 (2017) 1885–1894.

[56] S. Malikic, K. Jahn, J. Kuipers, S.C. Sahinalp, N. Beerenwinkel, Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data, Nat. Commun. 10 (2019) 2750.

[57] E.M. Ross, F. Markowetz, OncoNEM: inferring tumor evolution from single-cell sequencing data, Genome Biol. 17 (2016) 69.

[58] J. Singer, J. Kuipers, K. Jahn, N. Beerenwinkel, Single-cell mutation identification via phylogenetic inference, Nat. Commun. 9 (2018) 5144.

[59] M. El-Kebir, SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error, Bioinformatics 34 (2018) i671–i679.

[60] S. Ciccolella, C. Ricketts, M. Soto Gomez, M. Patterson, D. Silverbush, P. Bonizzoni, I. Hajirasouliha, G. Della Vedova, Inferring cancer progression from Single-Cell Sequencing while allowing mutation losses, Bioinformatics 37 (2021) 326–333.

[61] A. McPherson, A. Roth, E. Laks, T. Masud, A. Bashashati, A.W. Zhang, G. Ha, J. Biele, D. Yap, A. Wan, L.M. Prentice, J. Khattra, M.A. Smith, C.B. Nielsen, S.C. Mullaly, S. Kalloger, A. Karnezis, K. Shumansky, C. Siu, J. Rosner, H.L. Chan, J. Ho, N. Melnyk, J. Senz, W. Yang, R. Moore, A.J. Mungall, M.A. Marra, A. Bouchard-Côté, C.B. Gilks, D.G. Huntsman, J.N. McAlpine, S. Aparicio, S.P. Shah, Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer, Nat. Genet. 48 (2016) 758–767.

[62] H. Zafar, A. Tzen, N. Navin, K. Chen, L. Nakhleh, SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models, Genome Biol. 18 (2017) 178.
[63] S. Malikic, F.R. Mehrabadi, S. Ciccolella, M.K. Rahman, C. Ricketts, E. Haghshenas, D. Seidman, F. Hach, I. Hajirasouliha, S.C. Sahinalp, PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data, Genome Res. 29 (2019) 1860–1877.
[64] E. Sollier, J. Kuipers, K. Takahashi, N. Beerenwinkel, K. Jahn, COMPASS: joint copy number and mutation phylogeny reconstruction from amplicon single-cell sequencing data, Nat. Commun. 14 (2023) 4921.
[65] Mission Bio, Inc., Performance of the Tapestri ® Platform for Single-Cell Targeted DNA Sequencing, (2019).
[66] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics. Doklady 10 (1965) 707–710.
[67] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet J. 17 (2011) 10.
[68] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, ArXiv [q-Bio.GN] (2013). http://arxiv.org/abs/1303.3997.
[69] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.
[70] Linnarsson Lab, Loom file format specs - Loom file format version 3.0.0, Linnarsson Lab (2019). https://linnarssonlab.org/loompy/format/index.html (accessed March 13, 2024).
[71] J.M. Bennett, D. Catovsky, M.T. Daniel, G. Flandrin, D.A. Galton, H.R. Gralnick, C. Sultan, Proposed revised criteria for the classification of acute myeloid leukemia. A report of the French-American-British Cooperative Group, Ann. Intern. Med. 103 (1985) 620–625.
[72] J.M. Bennett, D. Catovsky, M.T. Daniel, G. Flandrin, D.A. Galton, H.R. Gralnick, C. Sultan, Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group, Br. J. Haematol. 33 (1976) 451–458.
[73] H. Döhner, E. Estey, D. Grimwade, S. Amadori, F.R. Appelbaum, T. Büchner, H. Dombret, B.L. Ebert, P. Fenaux, R.A. Larson, R.L. Levine, F. Lo-Coco, T. Naoe, D. Niederwieser, G.J. Ossenkoppele, M. Sanz, J. Sierra, M.S. Tallman, H.-F. Tien, A.H. Wei, B. Löwenberg, C.D. Bloomfield, Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel, Blood 129 (2017) 424–447.
[74] S. Kayser, M. Kramer, D. Martínez-Cuadrón, J. Grenet, K.H. Metzeler, Z. Sustkova, M.R. Luskin, A.M. Brunner, M.A. Elliott, C. Gil, S.C. Marini, Z. Ráčil, P. Cetkovsky, J. Novak, A.E. Perl, U. Platzbecker, F. Stölzel, A.D. Ho, C. Thiede, R.M. Stone, C. Röllig, P. Montesinos, R.F. Schlenk, M.J. Levis, Characteristics and outcome of patients with core-binding factor acute myeloid leukemia and FLT3-ITD: results from an international collaborative study, Haematologica 107 (2022) 836–843.
[75] H. Ayatollahi, A. Shajiei, M.H. Sadeghian, M. Sheikhi, E. Yazdandoust, M. Ghazanfarpour, S.F. Shams, S. Shakeri, Prognostic importance of C-KIT mutations in core binding factor acute myeloid leukemia: A systematic review, Hematol. Oncol. Stem Cell Ther. 10 (2017) 1–7.
[76] C.M. Arends, T.G. Liman, P.M. Strzelecka, A. Kufner, P. Löwe, S. Huo, C.M. Stein, S.K. Piper, M. Tilgner, P.S. Sperber, S. Dimitriou, P.U. Heuschmann, R. Hablesreiter, C. Harms, L. Bullinger, J.E. Weber, M. Endres, F. Damm, Associations of clonal hematopoiesis with recurrent vascular events and death in patients with incident ischemic stroke, Blood 141 (2023) 787–799.

[77] V. Panagiota, J.F. Kerschbaum, O. Penack, C.M. Stein, C.M. Arends, C. Koenecke, P.M. Strzelecka, A. Kloos, L. Wiegand, A. Lasch, R. Altwasser, A. Halik, R. Gabdoulline, J. Thomson, K. Weibl, G.-N. Franke, C. Berger, J. Hasenkamp, F. Ayuk, I.-K. Na, G. Beutel, U. Keller, L. Bullinger, G.G. Wulf, N. Kröger, V. Vucinic, M. Heuser, F. Damm, Clinical implications and dynamics of clonal hematopoiesis in anti-CD19 CAR T-cell treated patients, HemaSphere 7 (2023) e957.

[78] P.J.A. Cock, C.J. Fields, N. Goto, M.L. Heuer, P.M. Rice, The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, Nucleic Acids Res. 38 (2010) 1767–1771.

[79] Global Alliance for Genomics & Health (GA4GH), Sequence Alignment/Map Format Specification, (2023). http://samtools.github.io/hts-specs/SAMv1.pdf.

[80] F. Mölder, K.P. Jablonski, B. Letcher, M.B. Hall, C.H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S.O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, J. Köster, Sustainable data analysis with Snakemake, F1000Res. 10 (2021) 33.

[81] Broad Institute, Picard Tools, Broad Institute, GitHub Repository, Accessed: 2020/08/20; Version 2.20.0 (n.d.). http://broadinstitute.github.io/picard/.

[82] D.M. Church, V.A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.-C. Chen, R. Agarwala, W.M. McLaren, G.R.S. Ritchie, D. Albracht, M. Kremitzki, S. Rock, H. Kotkiewicz, C. Kremitzki, A. Wollam, L. Trani, L. Fulton, R. Fulton, L. Matthews, S. Whitehead, W. Chow, J. Torrance, M. Dunn, G. Harden, G. Threadgold, J. Wood, J. Collins, P. Heath, G. Griffiths, S. Pelan, D. Grafham, E.E. Eichler, G. Weinstock, E.R. Mardis, R.K. Wilson, K. Howe, P. Flicek, T. Hubbard, Modernizing reference genome assemblies, PLoS Biol. 9 (2011) e1001091.

[83] T. Fennell, N. Homer, fgbio, (n.d.). https://github.com/fulcrumgenomics/fgbio (accessed June 3, 2021).

[84] Z. Lai, A. Markovets, M. Ahdesmaki, B. Chapman, O. Hofmann, R. McEwen, J. Johnson, B. Dougherty, J.C. Barrett, J.R. Dry, VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research, Nucleic Acids Res. 44 (2016) e108.

[85] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, H. Li, Twelve years of SAMtools and BCFtools, Gigascience 10 (2021). https://doi.org/10.1093/gigascience/giab008.

[86] R Core Team, R: A Language and Environment for Statistical Computing, (2021). https://www.R-project.org/.

[87] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. 35 (2007) D61-5.

[88] M.J. Landrum, J.M. Lee, M. Benson, G.R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J.B. Holmes, B.L. Kattman, D.R. Maglott, ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Res. 46 (2018) D1062–D1067.

[89] X. Liu, X. Jian, E. Boerwinkle, dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions, Hum. Mutat. 32 (2011) 894–899.

[90] X. Liu, C. Li, C. Mou, Y. Dong, Y. Tu, dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs, Genome Med. 12 (2020) 103.

[91] S. Chen, L.C. Francioli, J.K. Goodrich, R.L. Collins, M. Kanai, Q. Wang, J. Alföldi, N.A. Watts, C. Vittal, L.D. Gauthier, T. Poterba, M.W. Wilson, Y. Tarasova, W. Phu, R. Grant, M.T. Yohannes, Z. Koenig, Y. Farjoun, E. Banks, S. Donnelly, S. Gabriel, N. Gupta, S. Ferriera, C. Tolonen, S. Novod, L. Bergelson, D. Roazen, V. Ruano-Rubio, M. Covarrubias, C. Llanwarne, N. Petrillo, G. Wade, T. Jeandet, R. Munshi, K. Tibbetts,

Genome Aggregation Database Consortium, A. O'Donnell-Luria, M. Solomonson, C. Seed, A.R. Martin, M.E. Talkowski, H.L. Rehm, M.J. Daly, G. Tiao, B.M. Neale, D.G. MacArthur, K.J. Karczewski, A genomic mutational constraint map using variation in 76,156 human genomes, Nature 625 (2023) 92–100.

[92] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (2001) 308–311.

[93] J.G. Tate, S. Bamford, H.C. Jubb, Z. Sondka, D.M. Beare, N. Bindal, H. Boutselakis, C.G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S.C. Jupe, C.Y. Kok, K. Noble, L. Ponting, C.C. Ramshaw, C.E. Rye, H.E. Speedy, R. Stefancsik, S.L. Thompson, S. Wang, S. Ward, P.J. Campbell, S.A. Forbes, COSMIC: The Catalogue Of Somatic Mutations In Cancer, Nucleic Acids Res. 47 (2019) D941–D947.

[94] N.M. Ioannidis, J.H. Rothstein, V. Pejaver, S. Middha, S.K. McDonnell, S. Baheti, A. Musolf, Q. Li, E. Holzinger, D. Karyadi, L.A. Cannon-Albright, C.C. Teerlink, J.L. Stanford, W.B. Isaacs, J. Xu, K.A. Cooney, E.M. Lange, J. Schleutker, J.D. Carpten, I.J. Powell, O. Cussenot, G. Cancel-Tassin, G.G. Giles, R.J. MacInnis, C. Maier, C.-L. Hsieh, F. Wiklund, W.J. Catalona, W.D. Foulkes, D. Mandal, R.A. Eeles, Z. Kote-Jarai, C.D. Bustamante, D.J. Schaid, T. Hastie, E.A. Ostrander, J.E. Bailey-Wilson, P. Radivojac, S.N. Thibodeau, A.S. Whittemore, W. Sieh, REVEL: An ensemble method for predicting the pathogenicity of rare missense variants, Am. J. Hum. Genet. 99 (2016) 877–885.

[95] J. Zhang, R. Bajari, D. Andric, F. Gerthoffert, A. Lepsa, H. Nahal-Bose, L.D. Stein, V. Ferretti, The international cancer genome consortium data portal, Nat. Biotechnol. 37 (2019) 367–369.

[96] E. Papaemmanuil, M. Gerstung, L. Bullinger, V.I. Gaidzik, P. Paschka, N.D. Roberts, N.E. Potter, M. Heuser, F. Thol, N. Bolli, G. Gundem, P. Van Loo, I. Martincorena, P. Ganly, L. Mudie, S. McLaren, S. O'Meara, K. Raine, D.R. Jones, J.W. Teague, A.P. Butler, M.F. Greaves, A. Ganser, K. Döhner, R.F. Schlenk, H. Döhner, P.J. Campbell, Genomic Classification and Prognosis in Acute Myeloid Leukemia, N. Engl. J. Med. 374 (2016) 2209–2221.

[97] N. Potter, F. Miraki-Moud, L. Ermini, I. Titley, G. Vijayaraghavan, E. Papaemmanuil, P. Campbell, J. Gribben, D. Taussig, M. Greaves, Single cell analysis of clonal architecture in acute myeloid leukaemia, Leukemia 33 (2019) 1113–1123.

[98] F.C. Brown, P. Cifani, E. Drill, J. He, E. Still, S. Zhong, S. Balasubramanian, D. Pavlick, B. Yilmazel, K.M. Knapp, T.A. Alonzo, S. Meshinchi, R.M. Stone, S.M. Kornblau, G. Marcucci, A.S. Gamis, J.C. Byrd, M. Gonen, R.L. Levine, A. Kentsis, Genomics of primary chemoresistance and remission induction failure in paediatric and adult acute myeloid leukaemia, Br. J. Haematol. 176 (2017) 86–91.

[99] Z. Sondka, S. Bamford, C.G. Cole, S.A. Ward, I. Dunham, S.A. Forbes, The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers, Nat. Rev. Cancer 18 (2018) 696–705.

[100] P. Desai, N. Mencia-Trinchant, O. Savenkov, M.S. Simon, G. Cheang, S. Lee, M. Samuel, E.K. Ritchie, M.L. Guzman, K.V. Ballman, G.J. Roboz, D.C. Hassane, Somatic mutations precede acute myeloid leukemia years before diagnosis, Nat. Med. 24 (2018) 1015–1023.

[101] R. Acuna-Hidalgo, H. Sengul, M. Steehouwer, M. van de Vorst, S.H. Vermeulen, L.A.L.M. Kiemeney, J.A. Veltman, C. Gilissen, A. Hoischen, Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life, Am. J. Hum. Genet. 101 (2017) 50–64.

[102] S. Abelson, G. Collord, S.W.K. Ng, O. Weissbrod, N. Mendelson Cohen, E. Niemeyer, N. Barda, P.C. Zuzarte, L. Heisler, Y. Sundaravadanam, R. Luben, S. Hayat, T.T. Wang, Z. Zhao, I. Cirlan, T.J. Pugh, D. Soave, K. Ng, C. Latimer, C. Hardy, K. Raine, D. Jones, D. Hoult, A. Britten, J.D. McPherson, M. Johansson, F. Mbabaali, J. Eagles, J.K. Miller,

D. Pasternack, L. Timms, P. Krzyzanowski, P. Awadalla, R. Costa, E. Segal, S.V. Bratman, P. Beer, S. Behjati, I. Martincorena, J.C.Y. Wang, K.M. Bowles, J.R. Quirós, A. Karakatsani, C. La Vecchia, A. Trichopoulou, E. Salamanca-Fernández, J.M. Huerta, A. Barricarte, R.C. Travis, R. Tumino, G. Masala, H. Boeing, S. Panico, R. Kaaks, A. Krämer, S. Sieri, E. Riboli, P. Vineis, M. Foll, J. McKay, S. Polidoro, N. Sala, K.-T. Khaw, R. Vermeulen, P.J. Campbell, E. Papaemmanuil, M.D. Minden, A. Tanay, R.D. Balicer, N.J. Wareham, M. Gerstung, J.E. Dick, P. Brennan, G.S. Vassiliou, L.I. Shlush, Prediction of acute myeloid leukaemia risk in healthy individuals, Nature 559 (2018) 400–404.

[103] S. Jaiswal, P. Natarajan, A.J. Silver, C.J. Gibson, A.G. Bick, E. Shvartz, M. McConkey, N. Gupta, S. Gabriel, D. Ardissino, U. Baber, R. Mehran, V. Fuster, J. Danesh, P. Frossard, D. Saleheen, O. Melander, G.K. Sukhova, D. Neuberg, P. Libby, S. Kathiresan, B.L. Ebert, Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease, N. Engl. J. Med. 377 (2017) 111–121.

[104] M. Frick, W. Chan, C.M. Arends, R. Hablesreiter, A. Halik, M. Heuser, D. Michonneau, O. Blau, K. Hoyer, F. Christen, J. Galan-Sousa, D. Noerenberg, V. Wais, M. Stadler, K. Yoshida, J. Schetelig, E. Schuler, F. Thol, E. Clappier, M. Christopeit, F. Ayuk, M. Bornhäuser, I.W. Blau, S. Ogawa, T. Zemojtel, A. Gerbitz, E.M. Wagner, B.M. Spriewald, H. Schrezenmeier, F. Kuchenbauer, G. Kobbe, M. Wiesneth, M. Koldehoff, G. Socié, N. Kroeger, L. Bullinger, C. Thiede, F. Damm, Role of Donor Clonal Hematopoiesis in Allogeneic Hematopoietic Stem-Cell Transplantation, J. Clin. Oncol. 37 (2019) 375–385.

[105] G. Collord, The pre-clinical evolution of haematological malignancies, University of Cambridge, 2019. https://www.sanger.ac.uk/theses/gc8-thesis.pdf.

[106] J.E. Feusier, S. Arunachalam, T. Tashi, M.J. Baker, C. VanSant-Webb, A. Ferdig, B.E. Welm, J.L. Rodriguez-Flores, C. Ours, L.B. Jorde, J.T. Prchal, C.C. Mason, Large-scale identification of clonal hematopoiesis and mutations recurrent in blood cancers, Blood Cancer Discov. 2 (2021) 226–237.

[107] D. Caetano-Anolles, Fisher's Exact Test, GATK (n.d.). https://gatk.broadinstitute.org/hc/en-us/articles/360035532152-Fisher-s-Exact-Test (accessed November 27, 2023).

[108] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120.

[109] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (2010) e164.

[110] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nat. Biotechnol. 29 (2011) 24–26.

[111] K. Yoshida, T. Toki, Y. Okuno, R. Kanezaki, Y. Shiraishi, A. Sato-Otsubo, M. Sanada, M.-J. Park, K. Terui, H. Suzuki, A. Kon, Y. Nagata, Y. Sato, R. Wang, N. Shiba, K. Chiba, H. Tanaka, A. Hama, H. Muramatsu, D. Hasegawa, K. Nakamura, H. Kanegane, K. Tsukamoto, S. Adachi, K. Kawakami, K. Kato, R. Nishimura, S. Izraeli, Y. Hayashi, S. Miyano, S. Kojima, E. Ito, S. Ogawa, The landscape of somatic mutations in Down syndrome-related myeloid disorders, Nat. Genet. 45 (2013) 1293–1299.

[112] K. Kataoka, Y. Shiraishi, Y. Takeda, S. Sakata, M. Matsumoto, S. Nagano, T. Maeda, Y. Nagata, A. Kitanaka, S. Mizuno, H. Tanaka, K. Chiba, S. Ito, Y. Watatani, N. Kakiuchi, H. Suzuki, T. Yoshizato, K. Yoshida, M. Sanada, H. Itonaga, Y. Imaizumi, Y. Totoki, W. Munakata, H. Nakamura, N. Hama, K. Shide, Y. Kubuki, T. Hidaka, T. Kameda, K. Masuda, N. Minato, K. Kashiwase, K. Izutsu, A. Takaori-Kondo, Y. Miyazaki, S. Takahashi, T. Shibata, H. Kawamoto, Y. Akatsuka, K. Shimoda, K. Takeuchi, T. Seya, S. Miyano, S. Ogawa, Aberrant PD-L1 expression through 3'-UTR disruption in multiple cancers, Nature 534 (2016) 402–406.

[113] S. Lee, C.-H. Sun, H. Jang, D. Kim, S.-S. Yoon, Y. Koh, S.C. Na, S.I. Cho, M.J. Kim, M.-W. Seong, J.M. Byun, H. Yun, ITDetect: a method to detect internal tandem

duplication of FMS-like tyrosine kinase (FLT3) from next-generation sequencing data with high sensitivity and clinical application, BMC Bioinformatics 24 (2023) 62.

[114] T.B.K. Watkins, E.C. Colliver, M.R. Huska, T.L. Kaufmann, E.L. Lim, C.B. Duncan, K. Haase, P. Van Loo, C. Swanton, N. McGranahan, R.F. Schwarz, Refphase: Multi-sample phasing reveals haplotype-specific copy number heterogeneity, PLoS Comput. Biol. 19 (2023) e1011379.

[115] T.B.K. Watkins, E.L. Lim, M. Petkovic, S. Elizalde, N.J. Birkbak, G.A. Wilson, D.A. Moore, E. Grönroos, A. Rowan, S.M. Dewhurst, J. Demeulemeester, S.C. Dentro, S. Horswell, L. Au, K. Haase, M. Escudero, R. Rosenthal, M.A. Bakir, H. Xu, K. Litchfield, W.T. Lu, T.P. Mourikis, M. Dietzen, L. Spain, G.D. Cresswell, D. Biswas, P. Lamy, I. Nordentoft, K. Harbst, F. Castro-Giner, L.R. Yates, F. Caramia, F. Jaulin, C. Vicier, I.P.M. Tomlinson, P.K. Brastianos, R.J. Cho, B.C. Bastian, L. Dyrskjøt, G.B. Jönsson, P. Savas, S. Loi, P.J. Campbell, F. Andre, N.M. Luscombe, N. Steeghs, V.C.G. Tjan-Heijnen, Z. Szallasi, S. Turajlic, M. Jamal-Hanjani, P. Van Loo, S.F. Bakhoum, R.F. Schwarz, N. McGranahan, C. Swanton, Pervasive chromosomal instability and karyotype order in tumour evolution, Nature (2020). https://doi.org/10.1038/s41586-020-2698-6.

[116] P. Van Loo, S.H. Nordgard, O.C. Lingjærde, H.G. Russnes, I.H. Rye, W. Sun, V.J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C.M. Perou, A.-L. Børresen-Dale, V.N. Kristensen, Allele-specific copy number analysis of tumors, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 16910–16915.

[117] H. Wickham, R. François, L. Henry, K. Müller, D. Vaughan, dplyr: A Grammar of Data Manipulation, (2023).

[118] H. Wickham, D. Vaughan, M. Girlich, K. Usher, Tidyr: Tidy messy data, V1. 2.0, (2022).

[119] J. Hester, J. Bryan, glue: Interpreted String Literals, (2022).

[120] B. Bolker, G.R. Warnes, T. Lumley, gtools: Various R Programming Tools, (2022). https://CRAN.R-project.org/package=gtools.

[121] H. Wickham, J. Hester, J. Bryan, readr: Read Rectangular Text Data, (2023). https://CRAN.R-project.org/package=readr.

[122] Y. Fu, M. Mahmoud, V.V. Muraliraman, F.J. Sedlazeck, T.J. Treangen, Vulcan: Improved long-read mapping and structural variant calling via dual-mode alignment, Gigascience 10 (2021) giab063.

[123] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (2015) 68–74.

[124] C. Stangl, S. de Blank, I. Renkens, L. Westera, T. Verbeek, J.E. Valle-Inclan, R.C. González, A.G. Henssen, M.J. van Roosmalen, R.W. Stam, E.E. Voest, W.P. Kloosterman, G. van Haaften, G.R. Monroe, Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing, Nat. Commun. 11 (2020) 2861.

[125] W. De Coster, S. D'Hert, D.T. Schultz, M. Cruts, C. Van Broeckhoven, NanoPack: visualizing and processing long-read sequencing data, Bioinformatics 34 (2018) 2666–2669.

[126] C. Röllig, F.A. Ayuk, J. Braess, M. Heuser, M.G. Manz, J. Passweg, D. Reinhardt, R.F. Schlenk, A. Zebisch, Akute Myeloische Leukämie (AML), Onkopedia Leitlinen (2023). https://www.onkopedia.com/de/onkopedia/guidelines/akute-myeloische-leukaemie-aml (accessed March 5, 2024).

[127] C.C. Smith, K. Lin, A. Stecula, A. Sali, N.P. Shah, FLT3 D835 mutations confer differential resistance to type II FLT3 inhibitors, Leukemia 29 (2015) 2390–2392.

[128] J.-S. Ahn, H.-J. Kim, FLT3 mutations in acute myeloid leukemia: a review focusing on clinically applicable drugs, Blood Research 57 (2022) 32.

[129] P. Paschka, J. Du, R.F. Schlenk, V.I. Gaidzik, L. Bullinger, A. Corbacioglu, D. Späth, S. Kayser, B. Schlegelberger, J. Krauter, A. Ganser, C.-H. Köhne, G. Held, M. von Lilienfeld-Toal, H. Kirchen, M. Rummel, K. Götze, H.-A. Horst, M. Ringhoffer, M. Lübbert, M. Wattad, H.R. Salih, A. Kündgen, H. Döhner, K. Döhner, Secondary genetic lesions in acute myeloid leukemia with inv(16) or t(16;16): a study of the German-Austrian AML Study Group (AMLSG), Blood 121 (2013) 170–177.

[130] R. Rampal, M.E. Figueroa, Wilms tumor 1 mutations in the pathogenesis of acute myeloid leukemia, Haematologica 101 (2016) 672–679.

[131] N. Niktoreh, C. Walter, M. Zimmermann, C. von Neuhoff, N. von Neuhoff, M. Rasche, K. Waack, U. Creutzig, H. Hanenberg, D. Reinhardt, Mutated *WT1*, *FLT3-ITD,* and *NUP98-NSD1* fusion in various combinations define a poor prognostic group in pediatric acute myeloid leukemia, J. Oncol. 2019 (2019) 1609128.

[132] I. van der Werf, A. Wojtuszkiewicz, M. Meggendorfer, S. Hutter, C. Baer, M. Heymans, P.J.M. Valk, W. Kern, C. Haferlach, J.J.W.M. Janssen, G.J. Ossenkoppele, J. Cloos, T. Haferlach, Splicing factor gene mutations in acute myeloid leukemia offer additive value if incorporated in current risk classification, Blood Adv. 5 (2021) 3254–3265.

[133] H. Janiszewska, A. Bąk, K. Skonieczka, A. Jaśkowiec, M. Kiełbiński, A. Jachalska, M. Czyżewska, B. Jaźwiec, M. Kuliszkiewicz-Janus, J. Czyż, K. Kuliczkowski, O. Haus, Constitutional mutations of the CHEK2 gene are a risk factor for MDS, but not for de novo AML, Leuk. Res. 70 (2018) 74–78.

[134] R. Godfrey, D. Arora, R. Bauer, S. Stopp, J.P. Müller, T. Heinrich, S.-A. Böhmer, M. Dagnell, U. Schnetzke, S. Scholl, A. Östman, F.-D. Böhmer, Cell transformation by FLT3 ITD in acute myeloid leukemia involves oxidative inactivation of the tumor suppressor protein-tyrosine phosphatase DEP-1/ PTPRJ, Blood 119 (2012) 4499–4511.

[135] M. Li, R. Collins, Y. Jiao, P. Ouillette, D. Bixby, H. Erba, B. Vogelstein, K.W. Kinzler, N. Papadopoulos, S.N. Malek, Somatic mutations in the transcriptional corepressor gene BCORL1 in adult acute myelogenous leukemia, Blood 118 (2011) 5914–5917.

[136] V. Madan, H.P. Koeffler, Differentiation therapy of myeloid leukemia: four decades of development, Haematologica 106 (2021) 26–38.

[137] github.com/cbg-ethz/compass:issue7, (n.d.). https://github.com/cbg-ethz/COMPASS/issues/7 (accessed February 13, 2024).

[138] E.R. Gansner, E. Koutsofios, S.C. North, K.-P. Vo, A technique for drawing directed graphs, IEEE Trans. Softw. Eng. 19 (1993) 214–230.

[139] K. Ushey, J.J. Allaire, Y. Tang, reticulate: Interface to "Python," (2023).

[140] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, (2016). https://ggplot2.tidyverse.org.

[141] T. Barrett, M. Dowle, A. Srinivasan, J. Gorecki, M. Chirico, T. Hocking, data.table: Extension of `data.frame`, (2023). https://r-datatable.com.

[142] H. Wickham, stringr: Simple, Consistent Wrappers for Common String Operations, (2022).

[143] A. Kassambara, ggpubr: "ggplot2" Based Publication Ready Plots, (2023). https://rpkgs.datanovia.com/ggpubr/.

[144] J. Ooms, The jsonlite package: A practical and consistent mapping between JSON data and R objects, ArXiv [Stat.CO] (2014). http://arxiv.org/abs/1403.2805 (accessed January 12, 2024).

[145] H. Wickham, J. Bryan, readxl: Read Excel Files, (2023).

[146] T. van den Brand, ggh4x: Hacks for "ggplot2," (2023).

[147] G. Van Rossum, F.L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

[148] S. Meschiari, latex2exp: Use LaTeX Expressions in Plots, (2023).

[149] A.M. Al-Subaie, B. Kamaraj, The structural effect of FLT3 mutations at 835th position and their interaction with Acute Myeloid Leukemia inhibitors: In silico approach, Int. J. Mol. Sci. 22 (2021) 7602.

[150] C. Darwish, K. Farina, D. Tremblay, The core concepts of core binding factor acute myeloid leukemia: Current considerations for prognosis and treatment, Blood Rev. 62 (2023) 101117.

[151] C.M. Arends, K. Kopp, R. Hablesreiter, N. Estrada, F. Christen, U.M. Moll, R. Zellinger, W.D. Schmitt, J. Sehouli, M. Fleischmann, I. Ray Coquard, A. Zeimet, F. Raspagliesi, C. Zamagni, I. Vergote, D. Lorusso, N. Concin, L. Bullinger, E.-I. Braicu, F. Damm, Dynamics of clonal hematopoiesis under DNA-damaging treatment in patients with ovarian cancer, Leukemia (n.d.).

[152] Z.J. Faber, X. Chen, A.L. Gedman, K. Boggs, J. Cheng, J. Ma, I. Radtke, J.-R. Chao, M.P. Walsh, G. Song, A.K. Andersson, J. Dang, L. Dong, Y. Liu, R. Huether, Z. Cai, H. Mulder, G. Wu, M. Edmonson, M. Rusch, C. Qu, Y. Li, B. Vadodaria, J. Wang, E. Hedlund, X. Cao, D. Yergeau, J. Nakitandwe, S.B. Pounds, S. Shurtleff, R.S. Fulton, L.L. Fulton, J. Easton, E. Parganas, C.-H. Pui, J.E. Rubnitz, L. Ding, E.R. Mardis, R.K. Wilson, T.A. Gruber, C.G. Mullighan, R.F. Schlenk, P. Paschka, K. Döhner, H. Döhner, L. Bullinger, J. Zhang, J.M. Klco, J.R. Downing, The genomic landscape of core-binding factor acute myeloid leukemias, Nat. Genet. 48 (2016) 1551–1556.

[153] S. Fröhling, R.F. Schlenk, I. Stolze, J. Bihlmayr, A. Benner, S. Kreitmeier, K. Tobis, H. Döhner, K. Döhner, CEBPA mutations in younger adults with acute myeloid leukemia and normal cytogenetics: prognostic relevance and analysis of cooperating mutations, J. Clin. Oncol. 22 (2004) 624–633.

[154] L. Zhang, H.W. Bass, J. Irianto, X. Mallory, Integrating SNVs and CNAs on a phylogenetic tree from single-cell DNA sequencing data, Genome Res. 33 (2023) 2002–2017.

[155] L. Pachter, Caltech BI/BE/CSS 183: Introduction to Computational Biology and Bioinformatics, (2022). https://github.com/pachterlab/BI-BE-CS-183-2023.git.

[156] M. Yilmaz, F. Wang, S. Loghavi, C. Bueso-Ramos, C. Gumbs, L. Little, X. Song, J. Zhang, T. Kadia, G. Borthakur, E. Jabbour, N. Pemmaraju, N. Short, G. Garcia-Manero, Z. Estrov, H. Kantarjian, A. Futreal, K. Takahashi, F. Ravandi, Late relapse in acute myeloid leukemia (AML): clonal evolution or therapy-related leukemia?, Blood Cancer J. 9 (2019) 7.

[157] S. Vosberg, P.A. Greif, Clonal evolution of acute myeloid leukemia from diagnosis to relapse, Genes Chromosomes Cancer 58 (2019) 839–849.

[158] P.A. Greif, L. Hartmann, S. Vosberg, S.M. Stief, R. Mattes, I. Hellmann, K.H. Metzeler, T. Herold, S.A. Bamopoulos, P. Kerbs, V. Jurinovic, D. Schumacher, F. Pastore, K. Bräundl, E. Zellmeier, B. Ksienzyk, N.P. Konstandin, S. Schneider, A. Graf, S. Krebs, H. Blum, M. Neumann, C.D. Baldus, S.K. Bohlander, S. Wolf, D. Görlich, W.E. Berdel, B.J. Wörmann, W. Hiddemann, K. Spiekermann, Evolution of Cytogenetically Normal Acute Myeloid Leukemia During Therapy and Relapse: An Exome Sequencing Study of 50 Patients, Clin. Cancer Res. 24 (2018) 1716–1726.

[159] S. Vosberg, L. Hartmann, K.H. Metzeler, N.P. Konstandin, S. Schneider, A. Varadharajan, A. Hauser, S. Krebs, H. Blum, S.K. Bohlander, W. Hiddemann, J. Tischer, K. Spiekermann, P.A. Greif, Relapse of acute myeloid leukemia after allogeneic stem cell transplantation is associated with gain of WT1 alterations and high mutation load, Haematologica 103 (2018) e581–e584.

[160] R. Itzykson, N. Duployez, A. Fasan, G. Decool, A. Marceau-Renaut, M. Meggendorfer, E. Jourdan, A. Petit, H. Lapillonne, J.-B. Micol, P. Cornillet-Lefebvre, N. Ifrah, G. Leverger, H. Dombret, N. Boissel, T. Haferlach, C. Preudhomme, Clonal interference of

signaling mutations worsens prognosis in core-binding factor acute myeloid leukemia, Blood 132 (2018) 187–196.

Data availability

Supplementary data, tables and analysis scripts can be found at the following GitHub repository:

- https://github.com/RaphaelHablesreiter/PhD_Thesis_SupplementaryMaterial

The variant calling pipeline that was used for the analysis of bulk sequencing data and the modified version of COMPASS are available under following GitHub repositories:

- https://github.com/RaphaelHablesreiter/umi-processing
- https://github.com/RaphaelHablesreiter/COMPASS:develop

The raw sequencing data for this thesis have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB75961 (https://www.ebi.ac.uk/ena/browser/view/PRJEB75961).

Supplement

List of Supplemental Tables

Supplemental Table 1: List of genes (n=45) covered by the customized targeted sequencing panel.

Gene	Exons covered	Gene	Exons covered
ASXL1	full coding sequence	*KRAS*	full coding sequence
ATM	full coding sequence	*MPL*	Exon 10
BCOR	full coding sequence	*MYD88*	full coding sequence
BCORL1	full coding sequence	*NF1*	Exons 28-38
BRAF	Exon 15	*NOTCH1*	Exons 26,27,34
BRCC3	full coding sequence	*NPM1*	Exon 11
CALR	Exons 8-9	*NRAS*	full coding sequence
CBL	full coding sequence	*PHF6*	Exons 3-5, 7-8
CEBPA	full coding sequence	*PPM1D*	full coding sequence
CHEK2	full coding sequence	*PTPN11*	full coding sequence
CSF3R	Exons 14,17	*RAD21*	full coding sequence
DNMT3A	full coding sequence	*RUNX1*	full coding sequence
ETV6	full coding sequence	*SETBP1*	Exons 4-9
EZH2	full coding sequence	*SF3B1*	full coding sequence
FLT3	Exons 6, 14,15, 20	*SRSF2*	full coding sequence
GATA1	Exon 2	*STAG2*	full coding sequence
GATA2	full coding sequence	*STAT3*	full coding sequence
GNAS	full coding sequence	*TET2*	full coding sequence
GNB1	full coding sequence	*TP53*	full coding sequence
IDH1	full coding sequence	*U2AF1*	full coding sequence
IDH2	full coding sequence	*WT1*	full coding sequence
JAK2	full coding sequence	*XPO1*	Exon 14
KIT	Exons 8-11,17		

Supplemental Table 2: Whole-exome sequencing metrices.

Patient	Sample	Reads R1/R2 [x10^6]	Reads mapped (raw) [x10^6]	PCR Duplicates	Reads mapped (filtered) [x10^6]	Mean coverage
01	D	48.75/48.75	94.56	17.13 %	78.29	191
01	CR	46.02/46.02	89.17	11.93 %	78.45	183
01	Rel	51.82/51.82	100.32	13.89 %	86.25	202
02	D	50.71/50.71	98.7	19.77 %	79.12	203
02	CR	55.87/55.87	108.61	18.89 %	88.02	224
02	Rel	48.33/48.33	93.96	20.69 %	74.46	192
03	D	54.36/54.36	105.21	25.01 %	78.81	194
03	CR	64.42/64.42	125.14	22.21 %	97.26	255
03	Rel	56.53/56.53	108.82	23.4 %	83.23	216
04	D	43.34/43.34	84.33	17.8 %	69.26	174
04	CR	54.14/54.14	105.02	21.21 %	82.65	219
04	Rel	57.68/57.68	111.79	21.81 %	87.34	232
05	D	56.12/56.12	108.59	20.91 %	85.79	226
05	CR	49.43/49.43	95.81	20.21 %	76.38	198
05	Rel	54.55/54.55	106.05	19.84 %	84.94	218
06	D	58.82/58.82	113.21	16.31 %	94.65	232
06	CR	51.45/51.45	99.19	16.54 %	82.67	206
07	D	39.98/39.98	77.8	11.63 %	68.7	162
07	CR	39.59/39.59	76.78	12.03 %	67.46	158
07	Rel	37.52/37.52	72.68	11.31 %	64.35	151
08	D	50.09/50.09	97.49	14.87 %	82.93	201
08	CR	55.97/55.97	108.44	17.99 %	88.85	221
08	Rel	59.76/59.76	116.18	15.33 %	98.29	242
09	D	66.83/66.83	130.03	17.95 %	106.61	268
09	CR	67.33/67.33	130.98	14.44 %	111.98	278
09	Rel	43.99/43.99	85.18	14.25 %	72.95	175

Supplemental Table 3: Targeted sequencing data metrices.

Patient	Sample	Reads R1/R2 [x10^6]	Reads mapped (raw) [x10^6]	Reads mapped (consensus) [x10^6]	Reads mapped (filtered) [x10^6]	Mean target coverage
01	D	7.85/7.85	15.6	2.38	1.15	496
01	CR	10.21/10.21	20.29	3.19	1.56	778
01	Rel	13.97/13.97	27.81	4.23	1.95	944
02	D	11.26/11.26	22.4	3.52	1.64	808
02	CR	11.02/11.02	21.93	3.53	1.71	808
02	Rel	12.85/12.85	25.55	4.06	1.94	956
03	D	7.77/7.77	15.43	2.32	1.17	571
03	CR	7.81/7.81	15.54	2.45	1.14	575
03	Rel	7.9/7.9	15.68	2.08	1.16	541
04	D	5.6/5.6	11.13	1.27	0.69	343
04	CR	11.92/11.92	23.74	3.56	1.8	903
04	Rel	10.22/10.22	20.32	3.21	1.58	758
05	D	9.63/9.63	19.15	2.47	1.29	653
05	CR	8.11/8.11	16.11	2.42	1.12	558
05	Rel	10.7/10.7	21.27	3.28	1.67	767
06	D	7.8/7.8	15.53	1.61	0.84	437
06	CR	14.5/14.5	28.84	4.55	2.41	1233
07	D	10.91/10.91	21.67	3.37	1.66	735
07	Rel	8.86/8.86	17.62	2.44	1.09	522
08	Rel	14.33/14.33	28.49	4.49	2.09	1037
09	D	11.94/11.94	23.73	3.69	1.71	814
09	CR	10.41/10.41	20.68	3.18	1.42	722
09	Rel	12.03/12.03	23.92	3.74	1.79	859

Supplemental Table 4: Nanopore sequencing metrices.

Patient	Sample	Raw reads [x10^6]	Mean raw read length [bp]	Reads mapped [x10^6]	Read length N50 [bp]	Mean coverage
01	Diagnosis	2.13	9940.1	1.98	23659	6.19
02	Diagnosis	3.92	6243.2	3.8	11048	7.19
03	Diagnosis	1.62	11281.6	1.46	27968	5.34
04	Diagnosis	4.24	6875.4	3.82	24192	8.43
05	Diagnosis	4.26	4072.2	2.87	6871	3.48
06	Diagnosis	2.26	5176.9	2.58	11947	3.32
07	Diagnosis	4.02	6786.4	3.69	13453	7.9
08	Diagnosis	4.46	3207.3	3.76	5617	3.99
09	Diagnosis	5.21	6352.2	4.72	12829	9.57

Zusammenfassung

Intratumorale Heterogenität beschreibt die Koexistenz mehrerer genetisch unterschiedlicher Subklone innerhalb des Tumors eines Patienten, die durch somatische Evolution, klonale Diversifizierung und Selektion entstehen. Intratumorale Heterogenität ist eine der Hauptursachen für Therapieversagen und Therapieresistenz in der Behandlung. Das Verstehen der intratumoralen Heterogenität und der Tumorevolution kann zu neuen Therapieansätzen führen. In dieser Arbeit habe ich eine Methode zur integrierten Analyse von Bulk- und Einzelzell-DNA-Sequenzierungsdaten von Patienten mit Core-Binding-Factor akuter myeloischer Leukämie entwickelt. Diese definiert sich durch das Vorhandensein eines *RUNX1-RUNX1T1* oder *CBFB-MYH11*-Fusionsgens. Ich habe einen Datensatz aus Bulk- und Einzelzell-DNA-Sequenzierungsdaten mit Proben zum Zeitpunkt der Diagnose, der Remission und des Rezidivs von insgesamt 9 Patienten mit Core-Binding-Factor akuter myeloischer Leukämie generiert. Mit der von mir entwickelten Methode konnte ich die Tumorevolution einzelner Tumorproben oder, wenn vorhanden, von Proben der Diagnose und des Rezidivs unter dem Einfluss der Chemotherapie anhand somatischer Varianten, somatischer Kopienzahlveränderungen und von Fusionsgenen rekonstruieren. Mit dieser Methode konnte ich bei Leukämiepatienten die klonale Komposition analysieren und darüber hinaus habe ich gezeigt, dass die von mir entwickelte Methode subklonale Kopienzahlveränderungen mit einer größeren Genauigkeit als derzeitige Methoden erkennen kann.

List of publications

Arends CM, Kopp K, **Hablesreiter R**, […], Braicu EI, Damm F. Dynamics of clonal hematopoiesis under DNA-damaging treatment in patients with ovarian cancer. Leukemia. 2024 Apr 18. doi: 10.1038/s41375-024-02253-3. PMID: 38637689

Noerenberg D, Briest F, Hennch C, Yoshida K, **Hablesreiter R**, […], Damm F. Genetic Characterization of Primary Mediastinal B-Cell Lymphoma: Pathogenesis and Patient Outcomes. J Clin Oncol. 2024 Feb 1;42(4):452-466. doi: 10.1200/JCO.23.01053. PMID: 38055913

Briest F, Noerenberg D, Hennch C, Yoshida K, **Hablesreiter R**, […], Damm F. Frequent ZNF217 mutations lead to transcriptional deregulation of interferon signal transduction via altered chromatin accessibility in B cell lymphoma. Leukemia. 2023 Nov;37(11):2237-2249. doi: 10.1038/s41375-023-02013-9. PMID: 37648814

Arends CM, Liman TG, Strzelecka PM, Kufner A, Löwe P, Huo S, Stein CM, Piper SK, Tilgner M, Sperber PS, Dimitriou S, Heuschmann PU, **Hablesreiter R**, Harms C, Bullinger L, Weber JE, Endres M, Damm F. Associations of clonal hematopoiesis with recurrent vascular events and death in patients with incident ischemic stroke. Blood. 2023 Feb 16;141(7):787-799. doi: 10.1182/blood.2022017661. PMID: 36441964.

Arends CM, Dimitriou S, Stahler A, **Hablesreiter R**, […], Heinemann V, Damm F. Clonal hematopoiesis is associated with improved survival in patients with metastatic colorectal cancer from the FIRE-3 trial. Blood. 2022 Mar 10;139(10):1593-1597. doi: 10.1182/blood.2021014108. PMID: 34932794

Christen F, **Hablesreiter R**, […], Briest F, Damm F. Modeling clonal hematopoiesis in umbilical cord blood cells by CRISPR/Cas9. Leukemia. 2022 Apr;36(4):1102-1110. doi: 10.1038/s41375-021-01469-x. PMID: 34782715

Hoyer K, **Hablesreiter R**, […], Sinn M, Damm F. A genetically defined signature of responsiveness to erlotinib in early-stage pancreatic cancer patients: Results from the CONKO-005 trial. EBioMedicine. 2021 Apr;66:103327. doi: 10.1016/j.ebiom.2021.103327. PMID: 33862582

Christen F, Hoyer K, Yoshida K, Hou HA, Waldhueter N, Heuser M, Hills RK, Chan W, **Hablesreiter R**, […], Damm F. Genomic landscape and clonal evolution of acute myeloid leukemia with t(8;21): an international study on 331 patients. Blood. 2019 Mar 7;133(10):1140-1151. doi: 10.1182/blood-2018-05-852822. PMID: 30610028

Frick M, Chan W, Arends CM, **Hablesreiter R**, […], Damm F. Role of Donor Clonal Hematopoiesis in Allogeneic Hematopoietic Stem-Cell Transplantation. J Clin Oncol. 2019 Feb 10;37(5):375-385. doi: 10.1200/JCO.2018.79.2184. PMID: 30403573

Acknowledgements

Foremost, I would like to thank my supervisor Frederik Damm for giving me the opportunity to perform my graduate work in his lab and providing me this exciting topic for my research. During the last years in his lab, I have learned a lot and have grown as a scientist by being responsible for planning the experiments, analyzing and interpreting the results critically.

Furthermore, I would like to thank Prof. Dr. Knut Reinert for his willingness to supervise my thesis and function as a reviewer.

I would like to thank my colleagues in the lab of AG Damm, who constantly provided an enjoyable work atmosphere inside and outside of the lab. Here, I would like to thank especially Paulina Strzelecka and Natalia Barreas for their scientific input as well as their experimental support. A special thanks goes to my fellow PhD colleagues Catarina Stein, Laura Wiegand and Pelle Löwe who made long days in the lab and the office shorter and more fun.

My deepest gratitude goes to my colleagues and friends Friederike Christen, Laura Grunewald and Kaja Hoyer, who not only helped me with their various expertise but always had an open ear for problems and supported me during tougher periods of my PhD time. I am very lucky, that I could always count on your support.

Thanks to my parents and my sister Julia for their constant support during the time of my PhD work. Thank you for always believing in me. And last, I am grateful that Catharina always supported me during this years of long working days and me in front of the screen. You are always there for me, when I need someone to talk to or to get my mind of while dancing with you under strobe lights.