Inferring the Clonal Identity of Single Cells from RNA-seq Data with Unique Molecular Identifiers

Ilse Ariadna Valtierra Gutiérrez

Munich 2020

Inferring the Clonal Identity of Single Cells from RNA-seq Data with Unique Molecular Identifiers

Ilse Ariadna Valtierra Gutiérrez

Dissertation
presented to the Faculty of Biology
of the Ludwig–Maximilian–University
Munich

Submitted by
Ilse Ariadna Valtierra Gutiérrez
from Mexico City, Mexico

Munich, 28.05.2020

1. Gutchter: Dr. Ines Hellmann
2. Gutachter: Prof. Dr. Dirk Metzler
Tag der Abgabe: 28.05.2020
Tag der mündlichen Prüfung: 16.10.2020

Eidestattliche Versicherung und Erklärung

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt ist.

München, den 28.05.2020

Ilse Ariadna Valtierra Gutierrez

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist und dass ich mich anderweitig einer Doktorprüfung ohne Erfolg **nicht** unterzogen habe.

München, den 28.05.2020

Ilse Ariadna Valtierra Gutierrez

Contents

Summary

Cancer is an evolutionary disease, in which heterogeneous populations of tumor cells can emerge, proliferate, and disappear depending on selective and neutral processes. This principle has been observed in many studies of acute myeloid leukemia (AML), which is the most common blood cancer in adults. Clonal heterogeneity and evolution have been proposed to play a role in the high relapse rate of this type of cancer. In order to understand this feature, it is crucial to have adequate clinical and experimental models that can provide enough data to elucidate the evolutionary history of a tumor, such as patient-derived xenografts (PDX). These models can be combined with high-resolution sequencing technologies, such as single-cell RNA-seq, to provide a detailed view of the heterogeneity and molecular features of the tumor. However, adequate analytical tools have to be applied and developed in order to fully exploit such datasets.

Here I present the analysis of the clonal heterogeneity of an AML patient and the corresponding PDX model, which was treated with multiple rounds of chemotherapy. This model allowed to study the response of the tumor populations to the pressure induced by the therapy, and the possible evolutionary forces behind it. Datasets for these AML samples were generated with multiple types of sequencing methods, one of which was single-cell RNA sequencing. To enable the analysis of somatic mutations and clonal populations in this kind of data, I developed a software package, *umivariants*, which is capable of extracting and proofreading variant sequences by making use of Unique Molecular Identifiers (UMIs), which are sequence barcodes that allow to distinguish reads that come from PCR amplification duplicates. The benefits of employing this proofreading approach for variant calling and for inferring the clonal identity of single cells were demonstrated. Finally, I applied *umivariants* to the analysis of the single-cell data of the AML PDX samples that were treated with chemotherapy, as well as other datasets with UMI-based sequencing.

Aims of the Thesis

The present work has been motivated by one of the crucial needs in biomedical science: to improve the understanding of cancer, a vastly complex and diverse set of diseases. In spite of increasingly sophisticated therapies, it is widely acknowledged that multiple aspects of its underlying biology still make it one of the most complex clinical challenges. This is particularly true of acute myeloid leukemia (AML), which has a high incidence of relapse despite a seemingly low number of causal genetic factors. One of the fundamental mechanisms behind this phenomenon is clonal heterogeneity, in which cell lineages acquire individual mutations and establish subpopulations within the tumor. Such subclones have been found to expand in AML after chemotherapy and become frequent in the tumor at relapse (Ding et al. [2012]). However, it is still unclear whether these rare subclones are the main drivers of relapse in AML, or how often, and how the clonal composition is shaped by all the genetic and environmental factors. Characterizing this heterogeneity requires not only sensitive and precise methods, but also an evolutionary approach.

The main objective of this dissertation was to establish a framework for the analysis of the clonal architecture in AML using the multiple omics datasets from clinical data at diagnosis and relapse, as well as patient-derived xenograft (PDX) samples. I will describe the strategy to call somatic cancer variants in the sequencing data, and how this information was used to infer subclones and their frequencies. The analysis of clonal architecture and reconfiguration of clonal frequencies during chemotherapy treatment will be shown.

The present study was made possible by the Munich Collaborative Research Center 1243. Within this framework, collaborators extracted the primary patient samples and established patient-derived xenografts, which allow to preserve clinical tumor samples in a mouse in-vivo system that is amenable to replication and experimental manipulation (Vick et al. [2015]). With one particular PDX sample, it was possible to perform an experiment in which samples were treated with multiple rounds of chemotherapy. These samples were then analyzed with diverse genomic, epigenomic, and transcriptomic methods.

Sequencing data of the PDX samples was prepared with diverse methods that employ sequence barcodes called unique molecular identifiers (UMIs). This was particularly important for the case of single-cell RNA sequencing (scRNA-seq). Therefore, another aim of this dissertation was to develop *umivariants*, a software package that can use UMIs to proofread and call sequence variants in datasets from multiple kinds of methods, particularly scRNA-seq. The package also uses the UMI-corrected data to estimate variant allele frequencies, genotypes, and clonal assignment or clonal frequencies in the samples. I will present the incorporation of *umivariants* to tools and pipelines for processing particular types of data, as well as demonstrations of *umivariants* for the

analysis of scRNA-seq and targeted sequencing datasets.

Abbreviations

AML	Acute Myeloid Leukemia
AML-LT	AML Long-Term Experiment
bp	Base-Pair(s)
BC	Sample Barcode
CNV	Copy Number Variant
CQS	Consensus Quality Score
DE	Differential Expression / Differentially Expressed
FDR	False Discovery Rate
FN, FNR	False Negative, False Negative Rate
FP, FPR	False Positive, False Positive Rate
GATK	Genome Analysis Toolkit
GoT	Genotyping of Transcriptomes
GT	Genotype
HSC	Hematopoietic Stem Cell
MaxQualSum	Maximum Quality Sum
PCR	Polymerase Chain Reaction
PDX	Patient Derived Xenograft
ROS	Reactive Oxygen Species
scRNA-seq	Single-Cell RNA Sequencing
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
TP, TPR	True Positive, True Positive Rate
TN	True Negative
UMI	Unique Molecular Identifier
VAF	Variant Allele Frequency
WT	Wild-Type

In loving memory of Jaskier, Harald, and Panda

Chapter 1

Introduction

1.1 Heterogeneity and Evolution in Cancer

1.1.1 The Biology of Cancer and Acute Myeloid Leukemia

Cancer can be defined as a set of diseases in which cells proliferate uncontrollably, which then invade regions outside their tissue of origin. In its many forms, it represents the second leading cause of deaths worldwide (World Health Organization [2020]). Cancer is a genomic disease, characterized by genetic and epigenetic factors that lead to a disruption of the molecular mechanisms that regulate the normal life cycle of healthy cells (Yates and Campbell [2012]). This is achieved, on one hand, by upregulating the signaling pathways that sustain proliferation and the mechanisms that lead to replicative immortality (e.g. telomere extension). On the other hand, cancers need to avoid programmed cell death, growth suppressors, and immune destruction (Hanahan and Weinberg [2000], Hanahan and Weinberg [2011]).

One particular family of cancers is that of leukemias. These are hematologic, i.e. blood cancers, which originate from cells of the hematopoietic stem cell lineage. These can be classified initially as acute or chronic depending on the rate at which the tumor cells divide; the other important element for their classification is the actual lineage of the tumor originating cell. Acute myeloid leukemia (AML) is the type that derives from abnormally or not differentiated hematopoietic stem and progenitor cells, which infiltrate the bone marrow and peripheral blood (Döhner et al. [2015]). AML is diagnosed when more than 20% of the bone marrow or peripheral blood cells can be morphologically classified as myeloblasts, or when the cells present certain recurrent genetic abnormalities (Arber et al. [2016]). AML entails a poor prognosis in adults over 60 years old, and can be successfully cured in only about 35-40% of adults below that age (Döhner et al. [2015]).

AML patients are particularly prone to high relapse rates. Across multiple studies, more than two thirds of the patients that enter complete remission have been observed to relapse, generally within the first 5 years after the initial diagnosis (Yanada et al. [2008], Verma et al. [2010], Oliva et al. [2018]). It has been shown that relapse AML samples can present additional or different genetic alterations compared to the original tumor (Ding et al. [2012], Klco et al. [2014], Shlush et al. [2017]). In order to obtain a deeper understanding of how these aggressive relapse tumors develop in AML, as well as other cancers, it is therefore not only important to analyze

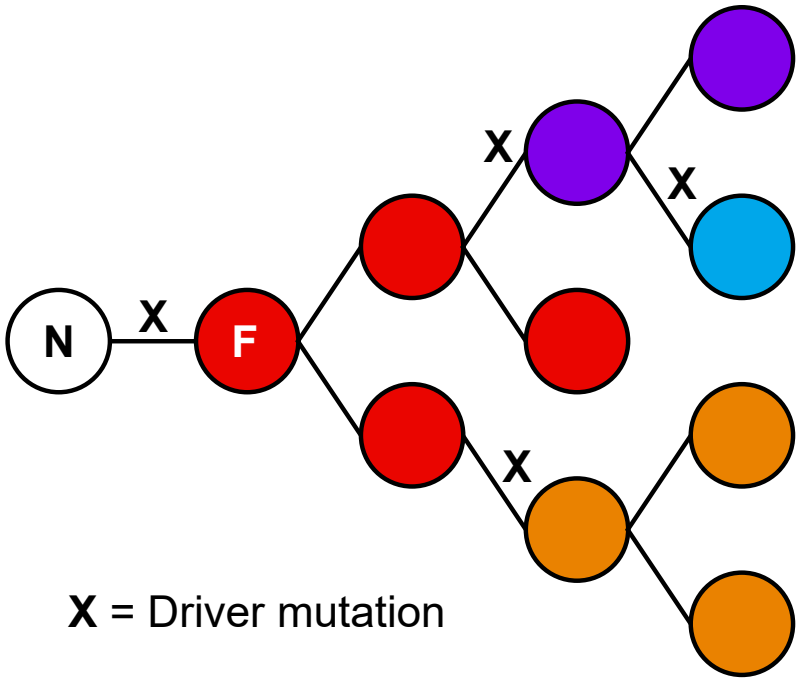


Figure 1.1: Representation of the clonal evolution process. Each node corresponds to a cell, where the first node (labeled 'N') is the normal cell from which the founder of the tumor ('F') originates. Each line to a new node represents a cell division, and each cross mark ('X') represents a driver mutation event that defines a new clone or subclone. Adapted from Nowell [1976].

the molecular features that characterize the relapse tumors, but also to study how such features emerge and expand in the cells. Evolutionary theory provides the exact framework to analyze how tumors diversify and adapt to the pressure of treatments.

1.1.2 Clonal Evolution and Heterogeneity

Given that cancers emerge from mutations and pathway deregulation that increase their proliferation - and thus their fitness - with respect to normal, healthy cells, it is natural to view this process as akin to the evolution of species: new traits emerge in a population, and are eventually fixed or lost by selection or drift. The basis for the paradigm of cancer as an evolutionary phenomenon was laid out by the seminal paper of Nowell [1976]. In this work, cancer evolution is described by a model of clonal evolution, in which the first neoplastic cell divides asexually, transfers its molecular lesions to the next generation, and the process is repeated as the tumor expands (figure 1.1). Thus, all cells from a neoplasm share a common set of mutations and/or chromosomal aberrations that can be traced to one ancestral cell. At the species level, genomic recombination (driven, for instance, by sexual reproduction) would lead to a mix and spread of those ancestral traits.

In the model of clonal evolution, tumor cells would not only retain a core of ancestral traits: new generations of these cells eventually and independently acquire novel sets of mutations, which are inherited to new offspring cells. This has the effect of increasing the diversity of the tumor cells, thus generating new populations called subclones. Subclonal diversity has been widely observed across tumors of all tissues, and has been frequently proposed as the origin of tumor cells that

resist chemotherapy, lead to relapse, initiate metastasis, and evade immunological control (Hu et al. [2010], Ding et al. [2012], Caswell and Swanton [2017]).

1.1.3 Mutations in cancer

Mutations in cancers are vastly diverse. Throughout the last decade, gigantic efforts by the international community have been undertaken to characterize a comprehensive catalog of cancer variants, giving birth to projects like The Cancer Genome Atlas (TCGA) or the Pan-Cancer Analysis of Whole Genomes (PCAWG) (Cancer Genome Atlas Research Network et al. [2013], ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium [2020]).

By analyzing variants from the big datasets of these consortia, a trend for their functional importance has been revealed and consolidated. First, there are mutations known as *drivers*, which directly increase the fitness of a tumor cell and are generally more frequently observed in multiple patients (Greenman et al. [2007]). Driver mutations can be categorized in two groups: gain-of-function in oncogenes, which are genes that stimulate cell growth, division, and survival; and loss-of-function in tumor suppressors, which regulate cellular growth, DNA repair, and cell cycle checkpoint activation (Lee and Muller [2010]). The appearance of new driver mutations has been associated with subclonal expansions (Bozic et al. [2010]). The counterpart to driver mutations are *passenger* mutations, which do not have a direct functional impact on cancer genes and fitness. Passenger mutations can increase their frequency as a result of hitchhiking with driver mutations during clonal expansion.

Another general category of cancer variants is by the kind of molecular lesion they entail. Among genetic variants, which are direct alterations of the DNA sequence, we find single-nucleotide variants (SNVs), also called point mutations, which entail substitutions among the different nucleotides. Next are insertions and deletions (jointly abbreviated indels), which are the direct addition or removal of short sequences with respect to the reference. SNVs and indels are also categorized in terms of the effect on a protein sequence: if they fall inside the protein coding sequence, they can be synonymous (no effect on protein sequence) or non-synonymous (altered protein sequence); otherwise they are catalogued as non-coding. Mutations affecting a large portion of DNA sequence, typically above 50 base-pairs (Liu et al. [2013]), fall in the category of structural or copy-number variants (SVs and CNVs, respectively). These include large insertions and deletions, segmental duplications, and chromosomal rearrangements. More aspects of these mutations will be covered in sections 1.2.2 and 1.2.3.

In the last decade, it has been postulated that mutations in tumors tend to present patterns that may reveal some of the underlying molecular and physiological mechanisms that led to their emergence. These patterns are known as mutational signatures (Nik-Zainal et al. [2012], Alexandrov et al. [2013]). Such signatures have been inferred from the study of the sequence context (i.e. adjacent sequence) of the catalogue of somatic mutations from samples of all cancer types in datasets like the TCGA. Some signatures have been associated with specific etiologies, such as C to T deaminations that were correlated with aging and would be originated endogenously. In contrast, other signatures have been attributed to mutagenic exposures such as ultraviolet radiation or smoking (Alexandrov et al. [2013]).

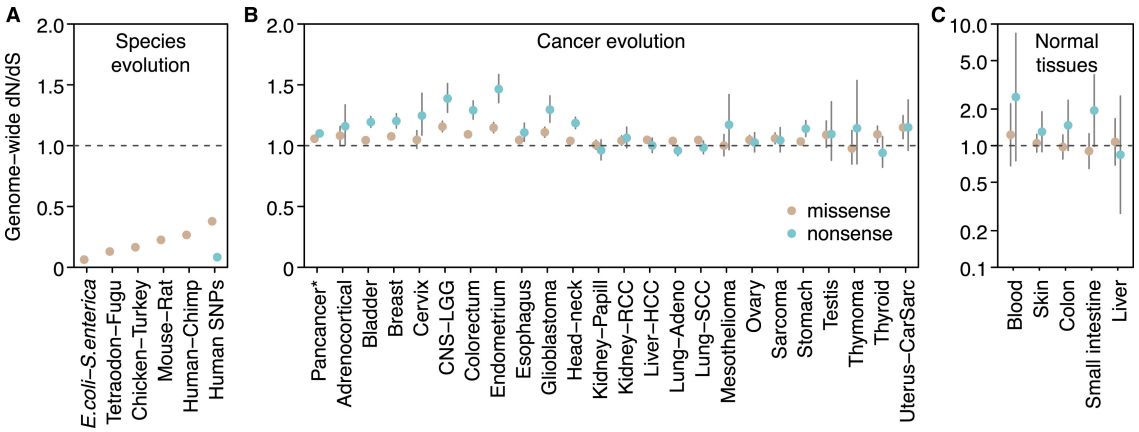


Figure 1.2: Estimation of genome-wide dN/dS values by Martincorena et al. [2017], across a) multiple species, b) tumor samples from the TCGA database, and c) normal tissues. Figure extracted directly from the original article by Martincorena et al. [2017], without modifications (CC BY 4.0 license, https://creativecommons.org/licenses/by/4.0/).

1.1.4 The Roles of Selection and Neutral Evolution

To complete the picture of cancer as an evolutionary process, the influence of natural selection and neutralism needs to be addressed. There has been an extensive debate over the last few years about whether Darwinian selection or genetic drift are the prominent forces that shape the survival and diversification of cancer in its environment, and analyses of some of the currently available big datasets of tumor genomes and mutations have led to contradictory interpretations. On one hand, the adaptation of a classical measure of selection by the rate of nonsynonymous versus synonymous mutations (dN/dS) yielded evidence for positive selection across most kinds of tumors from the TCGA dataset (figure 1.2; Martincorena et al. [2017]). However, different approaches reveal that certain cancers accumulate mutations at a rate that fits the expectations from a neutral model (Williams et al. [2016]).

With these contrasting views, and the accumulating evidence, one unifying framework has been to interpret the evolutionary forces in cancer as acting in a multi-step process (Wu et al. [2016]): selection can have a more decisive role in the early establishment of the tumor, when driver mutations confer a fitness advantage with respect to the normal tissue, even though positive and negative selection tend to overlap and cancel each other's effects. In later expansions, tumors accumulate mutations by drift. Both processes shape subclonal diversity.

1.1.5 Evolution in acute myeloid leukemia

Many of the characteristic driver mutations in AML fall in genes that regulate hematopoietic differentiation (NPM1, FLT3), DNA methylation (DNMT3A, TET2), or transcription factors (RUNX1, CEBPA), among others. Characteristic mutations in these genes include not only SNVs and indels, but also translocations and gene fusions (Saultz and Garzon [2016]). The majority of mutations that are detected in AML have been generally inferred to accumulate randomly throughout the life of the patient in healthy hematopoietic stem cells (Welch et al. [2012], Shlush et al. [2017]). The diversity of mutations and subclonal populations has found to be an important

feature of the AML samples at relapse (Ding et al. [2012], Klco et al. [2014], Shlush et al. [2017]).

In a targeted sequencing study of 1540 AML patients, Papaemmanuil et al. [2016] established that DNMT3A mutations were often acquired early in the founder clone (which could be attributed to clonal hematopoiesis), but eventually required additional mutations to achieve malignancy, particularly NPM1 (but also RUNX1). Other mutations that frequently appear in the founder clone of patients occur in TP53, IDH2, CEBPA, TET2, and NPM1 (Metzeler et al. [2016]). In contrast, RAS mutations are more often subclonal, with NRAS mutations being more frequent than those of KRAS (Papaemmanuil et al. [2016]). Another kind of subclonal mutations in AML, which are nevertheless frequently found in many patients, are FLT3 tandem duplications (Metzeler et al. [2016]).

In spite of the knowledge of such important driver mutations, it is not possible to use such information to predict the scenarios that would lead to potential relapses, especially if a rare subpopulation emerges with high fitness or resistance to therapy. In order to generate a better understanding of the evolutionary dynamics in AML, it is necessary to make use of methods that can provide a comprehensive profile of the genetic and molecular features that characterize the clonal and subclonal populations. Furthermore, it is necessary to employ experimental models where the effect of chemotherapies can be measured and interpreted in the context of tumor heterogeneity.

1.2 Analyzing Tumor Heterogeneity in Next-Generation Sequencing Data

1.2.1 The Landscape of Genomic and Transcriptomic Sequencing Methods

The last two decades have witnessed the surge and development of Next Generation Sequencing (NGS) methods, which have enabled the high-throughput analysis of hundreds of thousands of biological and clinical samples per year Goodwin et al. [2016]. The majority of studies and applications up to the present time have made use of the sequencing-by-synthesis approach of Illumina. This method relies on the generation of abundant short reads, typically within the range of 50 to 150 base pairs.

Sequencing analyses can be designed in ways that offer different compromises between the number of reads that cover a genomic region or the fraction of the genome that is covered by reads; i.e., the depth and breadth of coverage. Whole genome sequencing (WGS) offers the most comprehensive breadth of coverage, allowing the study of both coding and non-coding regions. However, sequencing sufficient samples at depths that would be ideal for variant calling and clonal inference, i.e. at least 200x reads (Griffith et al. [2015]), becomes expensive. Alternatively, it is possible to focus only on the protein coding regions: with whole exome sequencing (WES), exonic regions are specifically targeted, allowing deeper read coverage at a fraction of the cost of original WGS. Finally, when only a set of specific genes is of interest, it is possible to perform targeted amplicon sequencing (TAS) to obtain extremely deep coverage of that restricted set. In clinical applications and studies, WES and TAS are can be effectively used to accurately characterize patient samples where informative mutations are well known (Bewicke-Copley et al. [2019]). These methods usually require the PCR amplification of their target regions.

Sequencing has also been paramount to increase our capability to study gene expression. RNA sequencing has become a commonplace technique over the last decade, having displaced the use of microarrays due to its capability to characterize unexpected elements of the transcriptome. RNA-seq has been most frequently applied to the study of differentially expressed genes between samples. However, it has also provided huge opportunities for the study of alternative splicing, translation, RNA structure, and even spatial transcriptomics (Stark et al. [2019]). The same sequencing technology as for the genome can be used, but RNA first needs to be converted into a library of complementary DNA (cDNA) by means of reverse transcription (RT).

The sequencing techniques themselves are not the only ones that have enabled the analysis of ever more complex samples. Progress in sample and library preparation have been crucial to sequence not only full organisms, but also tissues and, particularly, single cells. The latter will be described in section 1.3

1.2.2 Somatic Single-Nucleotide Variant Calling

Sequencing technologies have made it possible to analyze millions of sequencing variants with high reliability. This feat has required the development of specialized computational and statistical methods. In general, such methods depend on comparing the observed sequence with a reference genome and/or control samples, as well as evaluating the amount and quality of evidence that come from the sequencing reads.

Some of the methods that have been designed to call somatic SNVs and indels are based on the direct analysis of the pileup of allelic coverage, such as VarScan (Koboldt et al. [2009], Koboldt et al. [2012]). VarScan calls germline and somatic variants by analyzing pileup files of tumor and matched-normal samples jointly, and applies a Fisher's exact test to determine significant differences between the VAFs of the tumor and normal samples. Another option is to apply Bayesian models on the coverage data, and two important somatic callers that use this approach are MuTect (Cibulskis et al. [2013], Benjamin et al. [2019]) and Strelka (Saunders et al. [2012], Kim et al. [2018]). Out of these algorithms, MuTect was the best performer in the ICGC-TCGA DREAM mutation calling challenge, in which many tools were benchmarked using in silico tumor genome data (Ewing et al. [2015]). Furthermore, MuTect2 has been shown to perform better at detecting low-frequency variants than Strelka2 (Chen et al. [2020]).

MuTect has been implemented as one of the tools from the Genome Analysis Toolkit (GATK). The GATK is not only designed as a software suite, but has also been conceived with a set of best practices (Van der Auwera et al. [2013]; figure 1.3). In the GATK pipeline, the first step is to align (i.e. map) the sequencing reads to a reference genome. For DNA sequencing (WGS, WES, TAS, etc.), this is done using BWA mem (Li [2013]); for RNA-seq, STAR is recommended (Dobin et al. [2013]). For the next steps, the reads are further pre-processed to reduce biases due to PCR amplification or sequencing technology. Using the Picard suite (Broad Institute [(Accessed: 2020/04/13; version 2.22.3]), reads coming from PCR duplicates of the same fragment are identified, and a 'Read Group' tag is added for sample identification. If the dataset is from RNA-seq, reads spanning splice junctions are divided into separate reads with one of the GATK modules. Afterwards, the GATK is used to recalibrate quality scores (which are provided with the sequencing reads) at locations with known SNVs and indels, which could be a hotspot for

systematic errors in sequencers. After the recalibration step, the reads are ready for actual variant calling.

The GATK tools and best practices were originally conceived to call single nucleotide polymorphisms (SNPs), i.e germline point mutations and small indels. This is done with the HaplotypeCaller (Poplin et al. [2017]. This software implements an assembly of haplotypes based on the reads around regions with candidate variants using de Bruijn-like graphs, an estimation of the likelihoods of the observed reads using a hidden Markov model (HMM), and a final estimation of the genotype likelihood.

In contrast to HaplotypeCaller, MuTect implements a Bayesian classifier, where the likelihood of the model with a called variant at a given location of the genome is compared to the likelihood of the variant being absent. These two likelihoods are estimated based on the number of reads supporting each of the alleles, and the sequence qualities (the full mathematical model is described in section 7.2.4). This is done to distinguish true variants from sequencing errors. The likelihood model is applied on both the tumor and, preferentially, a matched normal sample, which helps distinguish cancer somatic variants from germline or non-tumor variants. The likelihood values provide a measurement of confidence in the called variant, and allow the model to have no expectations of the variant allele frequency (VAF; equal to variant allele reads divided by total coverage). Therefore, low frequency variants, like the ones from tumor subclones, can still be detected and validated. In the latest version of the program, MuTect2, an assembly step like the one from HaplotypeCaller is also employed (Benjamin et al. [2019]).

1.2.3 Copy-Number Variant Calling

Copy-number variants (CNVs) are another kind of frequent genetic alteration in cancers. These comprise duplications and deletions of large chromosome segments of typically at least 50bp in length, but often spanning mega-bp regions or whole chromosome arms (Liu et al. [2013], Lauer and Gresham [2019]). In somatic cells, these generally come as a consequence of errors during homologous and non-homologous repair, particularly in genomic regions with high complexity or repeats (Hastings et al. [2009]). They can be produced by mechanisms of genomic instability in cancer and defective regulators of repair pathways or cell division.

Calling copy number variants from NGS data requires other approaches than simply comparing to a reference. One general principle is identifying positions where sequence coverage changes noticeably, called breakpoints, which are used to establish defined CNV segments (Liu et al. [2013]). This becomes complex to determine from short-read sequencing as the coverage signal is grainy and fuzzy throughout the genome, and even more so when it comes from WES sequencing only. Nevertheless, methods have emerged for both the WGS and WES scenarios. For instance, using the packages described in the MARATHON pipeline (Urrutia et al. [2018]), it is possible to call CNVs from either WGS or WES: CODEX/CODEX2 (Jiang et al. [2018]) are employed for coverage normalization and total copy number estimation, while FALCON or FALCON-X employ SNP information to assign CNVs to specific alleles (Chen et al. [2017]). This requires the usage of matched normal samples, with respect to which coverage estimates can be normalized and compared.

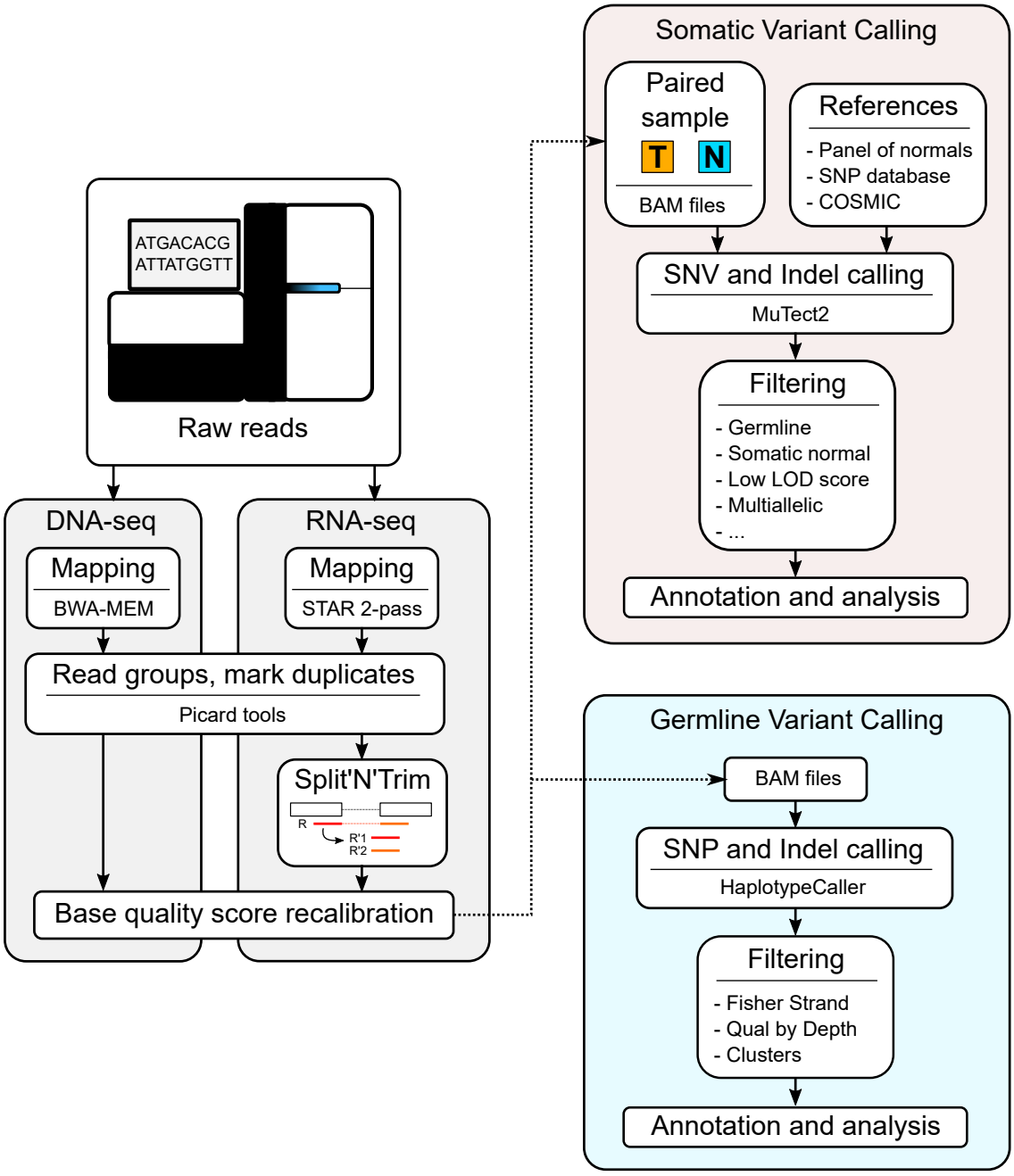


Figure 1.3: Pipelines for somatic and germline variant calling with the GATK best practices. Adapted from Van der Auwera et al. [2013], Cibulskis et al. [2013], Broad Institute [2014 (Accessed: 2020/05/10], and Broad Institute [2020 (Accessed: 2020/05/10].

1.2.4 Inference of Clonal Structure and Phylogeny

Sequencing data from cancer samples can provide information on the variants that characterize tumor subclones, as well as their relative fractions in the sample. In principle, this can be treated as a deconvolution problem, where the complete distribution of allele frequencies of all variants is actually a mixture of VAF distributions from all subclones. When the variants, subclones, and their frequencies are discerned from this mixture, it is possible to use this information to infer their evolutionary relations.

A wide diversity of algorithms have been designed to infer clonal compositions and phylogeny from DNA and also RNA sequencing data. An excellent overview is provided in (Schwartz and Schäffer [2017]). Some of these approaches tackle the inference of clonality and the phylogeny separately. One family of methods deals with the inference of subclones by VAF clustering approaches, and two popular examples are SciClone (Miller et al. [2014]) and PyClone (Roth et al. [2014]). SciClone applies a variational Bayesian mixture model that groups variants with closely-correlated frequencies in multiple samples. It has been used for a number of landmark studies on clonal heterogeneity in leukemias and other cancers (e.g. Klco et al. [2014]). PyClone applies a Markov Chain Monte Carlo (MCMC) Bayesian clustering algorithm, and is able to consider copy number information to convert VAFs into a cellular prevalence value (i.e. tumor fraction). However, PyClone can lead to overclustering of the variants when a large number of these is provided as input (Miller et al. [2014], Farahani et al. [2017]). Therefore, SciClone can be a more adequate tool for datasets with high numbers of variants, such as WGS. An alternative approach to these methods is Clomial (Zare et al. [2014]), a package that fit variant allele counts to a binomial model with a expectation-maximization algorithm. It runs one model for multiple numbers of clones, and chooses the best model (i.e. number of clones) based on maximum likelihood.

Clustering approaches do not themselves provide a framework for phylogenetic inference of the subclones, which requires the use of additional algorithms. ClonEvol (Dang et al. [2017]) is one such option: it parts from the assumption that the added cellular prevalence of daughter clones should not exceed that of their parent (sum rule), and clonal orderings should be the same across samples (cross rule). It tests these principles in all compatible phylogenies by bootstrapping the variants, discards trees that violate the rules, and infers the best consensus tree across samples.

Other methods provide a complete framework for deconvolution and phylogenetic inference. Canopy (Jiang et al. [2016]) is a framework that resolves clonal inference and phylogeny by sampling trees with an MCMC method. It can incorporate clonal and subclonal CNV events to the history. Like Clomial, it fits a model for a given number of clones at a time (although it would be for the likelihood of a tree), and can also be preceded by a binomial pre-clustering step.

When performing tumor clonal inference from bulk data, it should be considered that low sequencing depths and neutral evolution could lead to the inference of clusters with low-VAF variants which do not reflect the true phylogeny of the tumor. The recently developed framework MOBSTER (Caravagna et al. [2020]) attempts to fit such false-positive "neutral trail" clusters to a model with neutrality assumptions and remove them from the analysis. On the other hand, clonal inference precision can improve if sequencing is done at significantly higher depths (Griffith et al. [2015]). Another solution, however, is to adopt a single-cell based approach, which would provide exact estimates of mutational prevalence and co-ocurrence in the tumor cells. Such paradigm will

be discussed in the following section.

1.3 Single-Cell Sequencing for the Study of Tumor Evolution: Progress and Challenges

1.3.1 Single-cell RNA sequencing

One of the greatest breakthroughs in biology over the last decade has been the development of single-cell genomics and transcriptomics methods. This has been of particular interest to study cellular diversity in many contexts: during differentiation, development, and maintenance of tissues and organs; in health and disease; and in evolutionary comparisons of cellular composition and gene regulation. This progress has enabled the establishment of projects like the Human Cell Atlas, which has the ambitious goal of characterizing in detail all the cell types that can be found in the human body (Regev et al. [2017]).

In order to obtain this kind of knowledge, a huge focus has been placed on bringing gene expression to the single cell level, thus giving birth to single-cell RNA-seq (scRNA-seq). scRNA-seq techniques have grown exponentially in terms of their throughput and scalability: from tens of cells in a single experiment in 2009, to hundreds of thousands in the present (Svensson et al. [2018]). This has been made possible by the development of important methodological steps during library preparation and sequencing:

- Handling and separation of single cells. This has been accomplished by splitting and FACS-sorting the cells into single wells in plates, or by suspending them in an emulsion and sorting them in droplets or microfluidic devices.

- The incorporation of DNA barcodes to transcript sequences during first-strand synthesis, i.e. reverse transcriptions. Barcodes can be designed to label a specific cell in general (from here on labeled BCs), but one real advance has been the design of random barcodes that can tag individual transcripts. These are generally called Unique Molecular Identifiers (UMIs; Kivioja et al. [2011]). UMIs make it possible to collapse the reads that are products of PCR amplification, and instead count the original molecules (Islam et al. [2014]). This principle is illustrated in figure 1.4. Furthermore, the barcodes allow to pool single-cell libraries in early steps, as they can be demultiplexed during the analysis step based on the barcode sequence.

- PCR amplification of the cDNA library right after reverse transcription, which increases the starting material and reduces the molecules that are missed during fragmentation and sequencing.

It has been observed that UMI-based methods with early sample pooling, like SCRB or mcSCRB-seq, allow for a more sensitive quantification of gene expression at lower costs per cell, with increased power to detect differential expression compared to other methods (Ziegenhain et al. [2017], Bagnoli et al. [2018]). It must be acknowledged, however, that UMI-based methods require the incorporation of the barcodes via the primer for reverse transcription, and their detection is therefore restricted to the end of the sequence where they were placed. This is not an issue for full-length methods like Smart-seq (Ramsköld et al. [2012]) or Smart-seq2 (Picelli et al. [2013]),

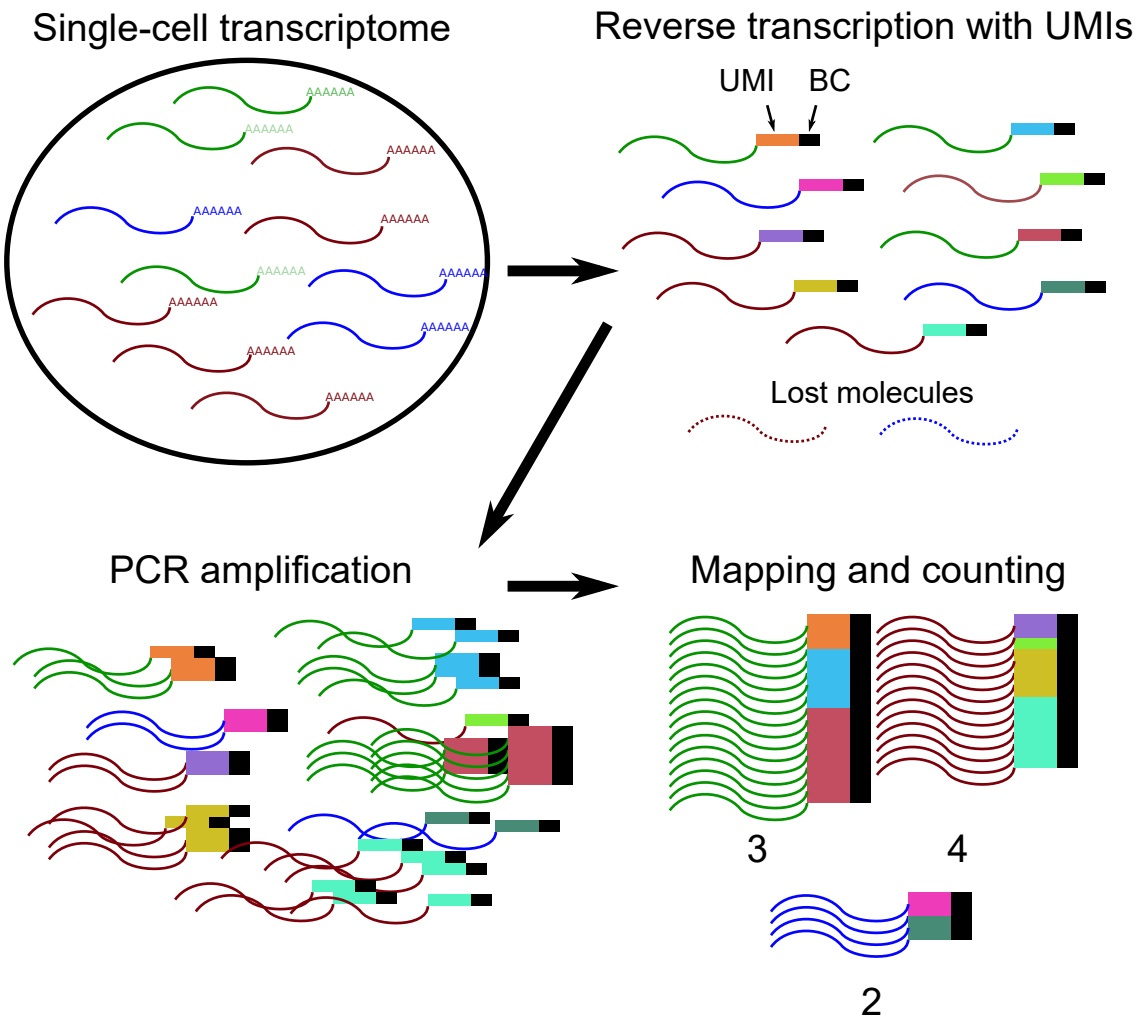


Figure 1.4: Using UMIs for transcript counting in scRNA-seq by collapsing PCR duplicates. Adapted from Islam et al. [2014].

which were shown to yield highly sensitive detection of gene expression in these benchmarks, albeit at the expense of sacrificing the use of UMIs and their methodological and quantification advantages. However, the recently developed Smart-seq3 protocol has been designed to produce full-length and UMI-tagged libraries simultaneously (Hagemann-Jensen et al. [2020]).

scRNA-seq methods have been widely adopted to characterize and analyze cancer cell populations since their inception. A few examples of their applications are the discovery of dormant, chemotherapy-resistant cells in acute lymphoblastic leukemia (Ebinger et al. [2016]), the inference of developmental hierarchies in oligodendroglioma (Tirosh et al. [2016]), and the association of clonal and driver mutations in AML (Petti et al. [2019]), or other hematological malignancies (Nam et al. [2019]), to their specific hematologic progenitor lineage.

1.3.2 Challenges in the analysis of somatic variants

Single cell data offer the unprecedented opportunity to analyze clonal evolution at the highest possible resolution. However, given the reduced amounts of starting material, single-cell methods are particularly sensitive to two noise from two opposite phenomena:

- If PCR amplification is employed, biases can be introduced and certain genes or regions can become overrepresented, and sequence errors can also be introduced early on the library preparation process.
- Certain genes or genomic regions will be missed due to insufficient coverage (especially in genes with low or no expression in scRNA-seq), a phenomenon also known as dropouts. Furthermore, amplification or sequencing errors could obscure variant detection.

To obtain precise variant calls in spite of those two error sources, several specialized methods have been developed. For instance, Monovar (Zafar et al. [2016]) employs sample pooling and dynamic programming to call variants and estimate genotype likelihoods across cells. $SCI\phi$ calls mutations in single cells by jointly inferring their phylogeny (Singer et al. [2018]). Alternatively, some pipelines have made use of HaplotypeCaller to call variants per cell, since these would appear as fixed SNPs on each cell (see Poirion et al. [2018]) and (Zhou et al. [2020]).

Bulk SNV calling methods have been evaluated for their direct use in scRNA-seq datasets (Liu et al. [2019]). On real and simulated SMART-seq2 data, Strelka2 and SAMtools-bcftools showed the best levels of sensitivity and specificity, and HaplotypeCaller also had good performance. On a 10x dataset, FreeBayes had the highest sensitivity, albeit with a high false discovery rate. MuTect2 was the least adequate algorithm in their comparison for direct application on scRNA-seq.

Currently, there are also approaches for CNV calling in single cells, particularly from scRNA-seq data. HoneyBADGER (Fan et al. [2018]) clusters the cells by smoothed minor allele frequencies of SNPs, divides them into two groups, calls CNV regions with a hidden Markov model, and repeats the procedure iteratively until no new CNVs are found. This also produces a CNV-based clonal phylogeny. CaSpER (Serin Harmanci et al. [2020]) applies a smoothing of expression signal to identify breakpoints, establishes segments with a hidden Markov model, and overlays them with allelic frequencies. The authors of these methods claimed that they could achieve CNV calling at 10 Mb or gene-length resolution, respectively. Finally, in a recent publication, Petti et al. [2019] employed a direct approach to call clonal variants that had been identified from enhanced WGS

datasets in scRNA-seq, and used this to analyze clinical leukemia samples. They developed two computational methods to perform variant pileup, UMI proofreading, and genotyping in 10x data: cb_sniffer and vartrix.

Aside from improving the computational frameworks that detect variants, experimental methods have been developed to increase coverage at variant sites. Genotyping of Transcriptomes (GoT, Nam et al. [2019]) implements a modified 10x platform in which a small fraction of transcripts from genes and loci of interest are deeply amplified with PCR during library preparation. UMI and cell barcode information are preserved, and it is possible to relate the amplicon variant calls to the single cell transcriptomes (and thus their inferred cell types) via the BC. They also developed the IronThrone pipeline to perform variant calling in GoT data. MutaSeq is a similar method designed for amplification during Smart-seq2 library preparation (Velten et al. [2018]). The method is highly sensitive, but has the disadvantage of lacking UMIs. These methods rely on the amplification of pre-specified loci during library preparation, but a method under development, Single-Cell Transcriptome and Genotype (scTAG-seq), will provide an opportunity to amplify any locus from an already available SCRB/mcSCRB-seq cDNA library (Johannes Bagnoli, Lucas Wange; personal communication).

As these methods mature, and new datasets emerge, it also becomes necessary to develop frameworks that can process the reads and barcodes, call the variants of interest, and use this information for the analysis of clonal heterogeneity.

1.3.3 Clonal inference or mapping at the single-cell level

Several methods have been designed to infer clonal phylogenies from single-cell variant estimates, while accounting for both false positive and negative calls due to amplification bias, sequencing errors, and sparse coverage. These have been based on clustering, like SCG (Roth et al. [2016]), or on a maximum likelihood approach with evolutionary assumptions: for the infinite sites model, there is OncoNEM (Ross and Markowetz [2016]), which implements neighbor search heuristics to explore the tree space, or SCITE (Jahn et al. [2016]), which implements an MCMC tree search algorithm. On the other hand, SiFit (Zafar et al. [2017]) and SciCloneFit (Zafar et al. [2019]) are based on the finite sites model. SCITE and SiFit produce single-cell lineage trees, while OncoNEM and SiCloneFit are able to cluster these into subclones. All of these methods, however, remain sensitive to very high dropout or false negative estimates, and may not be able to perform with very high numbers of SNVs (a problem with e.g. OncoNEM).

Due to the sparsity of single cell approaches, there has been an interest in the field to combine these datasets with the strengths from bulk sequencing. Some frameworks have been designed to infer clones jointly from bulk and single-cell data, such as bSCITE (Malikic et al. [2019]), and ddClone (Salehi et al. [2017]). On the other hand, other methods have been designed to use the variant information from a previously inferred clonal architecture to map single cells to their clones, and actually require scRNA-seq as input. Cardelino (McCarthy et al. [2020]) is a framework that can extract variant allele counts of SNVs in the single-cell data, model allelic imbalance, and estimate the likelihood of each cell to belong to each possible subclone while accounting for the phylogenetic structure. clonealign (Campbell et al. [2019]) is another package that can do such clonal assignment, but using CNV information.

scRNA-seq is particularly challenging for the direct inference of clonality. This is not only due to the inherent sequencing issues from single cell methods, but also because detection of a variant would depend on the expression of its gene. Furthermore, as gene expression has been demonstrated to occur in bursts which lead to stochastic monoallelic expression (Reinius et al. [2016], Larsson et al. [2019]), this directly affects allelic coverage in a given cell. Nonetheless, a few frameworks have been designed to handle this kind of data. SSrGE (Poirion et al. [2018]) determines expressed SNVs by fitting a LASSO regression between SNV detection and expression levels. Then it uses the regression to estimate weights for the SNVs, scores them for potential subpopulations, and clusters the cells. DENDRO (Zhou et al. [2020]) implements a beta-binomial framework to model variant detection given dropouts and bursting, determines the cell genotypes, and determines the subclones by clustering the cells.

While the progress in clonal inference at the single-cell level has advanced significantly with the development of the previous approaches, there is still a fundamental need to ensure that the signal of false positive variants is minimized in order to avoid the inference of spurious subpopulations, or the incorrect association of single cells with a subclone. Otherwise, the conclusions on other molecular and evolutionary features about such cells and populations could be misleading. It is therefore critical to establish pipelines that can make use of tools like UMIs and multiple types of omics data to proofread variant calling, and which can afterwards provide the corrected data to the clonal inference method of choice.

1.4 Optimizing the Analysis of Tumor Evolution from Bulk to Single-Cell Level

1.4.1 Unique Molecular Identifiers to proofread variant calling

The potential of UMIs for their proofreading capabilities has not only been used for scRNA-seq applications, but also for accurate and ultra-sensitive variant calling in deep targeted approaches. Protocols like single molecule molecular inversion probes (smMIPs) have employed UMIs to analyze low frequency variants in cancer samples, and throughout clonal hematopoiesis (Hiatt et al. [2013], Acuna-Hidalgo et al. [2017]). UMI-based approaches have also been used to obtain error-corrected immune profiles (Shugay et al. [2014], Turchaninova et al. [2016]). In order to perform such proofreading, these approaches have relied on calling a consensus sequence across the reads that are labeled with the same UMI (Salk et al. [2018]). The concept of the UMI consensus parts from the principle that one UMI should correspond to a single molecule (i.e. a transcript or a fragment of genomic DNA), and a unique final sequence is determined from all the reads that are barcoded with the UMI, which would correspond to PCR copies. By employing this technique, it is possible to reduce the number of amplification and sequencing errors that are considered when calling variants. An illustration of the concept is shown in figure 1.5.

Some computational methods have been created to proofread UMI-labeled sequences and call variants. MAGERI (Shugay et al. [2017]) is a framework that first assembles a consensus sequence from all the reads that are labeled with a UMI, keeping the most frequent nucleotide at each position. Then, it uses a beta-binomial distribution to model errors during PCR amplification and sequencing in order to exclude them from the true variant. A similar approach had been

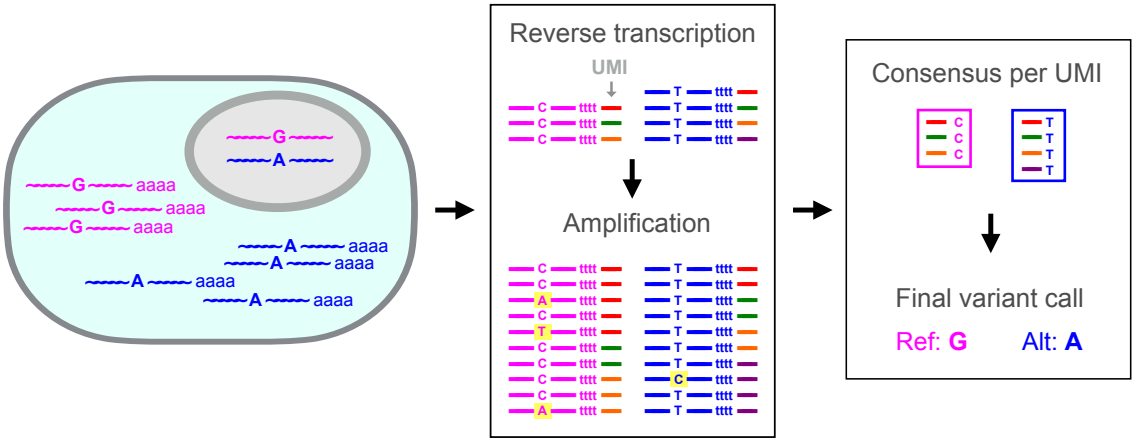


Figure 1.5: Calling the UMI consensus from multiple reads, and using the consensus sequences to call variants. This example is shown for cDNA reads of a single-cell transcriptome.

implemented in MIGEC, which calls consensus immunoglobulin sequences in RNA-seq data (Shugay et al. [2014]). Other approaches, as implemented by Hiatt et al. [2013] and in the fgbio suite (Fennel et al.), employ the sequence quality scores to calculate the likelihood of each possible nucleotide per position (i.e. given the error probability); this is used to estimate a posterior probability, and the highest value determines the consensus. Standard variant calling methods can be used afterwards. Finally, other methods perform variant calling after majority-vote UMI collapsing, such as cellSNP (Huang et al. [2019]) or UMI-VarCal (Sater et al. [2020]). However, these methods might not be suitable for the specific needs of clonal inference in scRNA-seq, and their UMI collapsing techniques need to be evaluated.

1.4.2 Analyzing clonal heterogeneity in leukemia from patient-derived xenografts

Clonal deconvolution within a single patient requires employing multiple samples, for instance from different timepoints or regions - either sections of the tumor or also from a metastasis (Schwartz and Schäffer [2017]). However, this poses a set of limitations for the comprehensive analysis of clonal heterogeneity and evolution, particularly for hematologic malignancies. First, this restricts the number of samples that can be realistically obtained and thus detection power, aside from the health burden and invasive methods required for their extraction. Second, it is not always possible to obtain sufficient biological material from the clinical sample to fully profile rare subclones, or even to detect them in the first place.

In order to produce additional material for the study of cancers, diverse experimental models have been established. In vitro systems, like cell lines and, more recently, 3D cultures and patient-derived organoids have enabled the characterization of many kinds of tumors (Kennedy and Barabé [2008], Gao and Chen [2015]). However, it has also been largely acknowledged that the physiological conditions in the environment of the tumor play a decisive role in its molecular composition, adaptation, and response to selective pressures, aside from the fact that successfully established cultures might present significant molecular differences to their original tumors (Gillet et al. [2013], Gao and Chen [2015]). Furthermore, AML has proven particularly difficult to preserve

in in-vitro cultures (Aigner et al. [2013]). Therefore, in vivo models that can provide a resemblance to the original host conditions represent a very valuable resource. Patient-derived xenografts (PDX) are a system in which primary tumor cells from the patient are transplanted and grown within an animal host, which is usually an immunodeficient mouse (Cho et al. [2016]).

PDX models have been a very useful tool to capture the heterogeneity of leukemia populations for nearly three decades (Cesano et al. [1992], Sawyers et al. [1992]). Cells from AML PDX samples have been shown to preserve many driver mutations that were present in the original sample patient at similar allele frequencies, and in some cases can even be engineered and serially transplanted (Vick et al. [2015], Wang et al. [2017]). Rare subclonal populations, and their corresponding variants, have also been demonstrated to develop successfully and acquire predominance in some PDX samples (Sandén et al. [2020]). Nevertheless, the versatility and throughput of PDX models make them attractive for the study of multiple biological properties by applying multi-omics methods, which can reveal valuable information about the biology of subclonal populations and the mechanisms that could enable them to resist therapies.

1.4.3 Developing software for the analysis of clonal variants

The aforementioned computational methods for UMI-based variant proofreading have enabled calling and correcting variants from the output of multiple protocols. However, each method has been designed for certain specific types of input, output, and usage. This increases the complexity and user workload during the integrative analysis of multiple UMI-based datasets. In order to facilitate the analysis of clonal variants and heterogeneity across multiple types of UMI-omics data, it is necessary to provide a set of tools and pipelines that provide homogeneous and reproducible results.

In this work, I will introduce *umivariants*, a package written in the R language to extract, proofread, and evaluate SNVs from a clonal architecture. The package also provides functions to assign single-cell or sample genotypes, estimate variant allele frequencies, and perform accurate cell-to-clone assignment. With this package, it is possible to obtain comparable results from different UMI-based protocols, regardless of whether they are DNA- or RNA-based, sequencing depth, barcode design, and so on. Furthermore, it makes it possible to evaluate the advantages of different UMI consensus methods in samples with complex clonal architectures.

Chapter 2

Clonal Heterogeneity in a PDX Model of Acute Myeloid Leukemia under Long-Term Chemotherapy

2.1 Description of the AML-PDX Model

One important question in AML evolution is how clonal heterogeneity is directly affected by the selective pressure of chemotherapy. Patient-derived xenograft (PDX) models have been used to maintain and study cancer heterogeneity, and therefore represent a very useful experimental tool to understand this phenomenon. A number of studies up to date have focused on analyzing the clonal structure in PDX models immediately after treating with chemotherapy once, or passaging multiple times without any selective pressure from chemotherapy (Sandén et al. [2020]). However, how such structure is affected after multiple applications of the treatment remains unexplored.

A collaborative project of the SFB 1243 was designed to provide some insights to that question. Samples from an established PDX model were engrafted into multiple mouse hosts, and these were separated into treatment groups that received up to three doses of chemotherapy. Tumor cells from these treated PDX were extracted before therapy, shortly after therapy, or after full-blown regrowth depending on the treatment group. I analyzed the clonal architecture of the treated and untreated PDX samples, as well as those from the original patient. Subclonal frequencies were estimated on each sample and compared along the treatment stages (i.e. the treatment groups with one, two, or three rounds of chemotherapy) in order to understand the evolutionary mechanisms that take place as the therapy progresses.

2.1.1 Clinical information and PDX sample generation

The clinical samples were extracted from a 52-year old patient at diagnosis that presented a case of acute myeloid leukemia with a t(2;11) translocation. Primary samples were taken from the bone marrow at three stages of the disease: diagnosis, first relapse, and second relapse. Three kinds of germline controls were extracted from the patient: minimal residual disease at full remission, after

Gene	AA	Chr	Position	Ref	Alt	V-D	V-R1	V-R2	V-491
NRAS	Q61K	chr1	115256530	G	T	0.003	0.0	0.037	0.082
DNMT3A	R882S	chr2	25457243	G	T	0.512	0.229	0.306	0.371
ETV6	P214L	chr12	12022535	C	T	0.487	0.214	0.249	0.42
KRAS	G12A	chr12	25398284	C	G	0.007	0.0	0.0	0.355
PTPN11	D61H	chr12	112888165	G	C	0.442	0.267	0.276	0.492
RUNX1	N136K	chr21	36252954	A	C	0.423	0.215	0.318	0.536
BCOR	P683fs	chrX	39932551	G	-	0.452	0.38	0.248	0.554

Table 2.1: SNVs of interest detected in the AML-LT HaloPlex analysis. The corresponding gene, amino acid change (AA), genomic coordinates (chromosome, position) in the hg19 reference, and alleles (reference and alternative) are shown. The last 4 columns indicate variant allele frequencies (V) at diagnosis (D), first relapse (R1), second relapse (R2), and AML491 (491). Data provided by Dr. Klaus Metzeler and Dr. Maja Rothenberg-Thurley (ELLF-LMU, SFB1243).

the first allogeneic hematopoietic stem-cell transplant, and after the second stem-cell transplant. A PDX line (AML491) was generated by collaborators from the Helmholtz Zentrum München (Dr. Binje Vick, Dr. Irmela Jeremias) from the first relapse sample, and transformed to express mCherry and luciferase (Vick et al. [2015]).

The patient and AML491 samples were analyzed by collaborators from the group of Dr. Klaus Metzeler for potential driver mutations at genes that are frequently mutated in AML. The Agilent HaloPlex system was used to amplify exonic regions of a panel of 67 AML-relevant genes (Metzeler et al., personal communication). 8 variants of interest were detected (table 2.1). The DNMT3A R882S mutation in particular was persistent in the remission sample.

A sample from the AML491 PDX was grown and serially passaged into new mouse hosts. The tumor content of one of such PDX samples (mouse 1021) was extracted, diluted, and transplanted into separate mouse hosts. These were treated with one, two, or three rounds of chemotherapy regime, and samples were taken at full-blown growth before the next treatment (stages I, III, V, and VII), or when the tumor population was depleted after the therapy (stages II, IV, and VI). In the particular case of stages V, VI, and VII, the tumors were not extracted from mouse 1021, but from PDX samples of stage IV. Due to the extended number of chemotherapy cycles that were used to treat the PDX samples, this was called the AML long-term experiment, or AML-LT (figure 2.1).

Tumor cells from the PDX samples were FACS-sorted and aliquoted for multiple kinds of genomic, transcriptomic, epigenomic, and functional analyses carried out by multiple groups of the SFB 1243. Two ancestral PDX samples of the LT-experiment (652 and 1021), i.e. that had been serially passaged from the original AML491, were also included as stage 0 (figure 2.2).

2.2 Clonal Inference in Whole-Genome Sequencing Data

2.2.1 Inferring the clonal phylogeny from somatic SNVs and Indels

3 primary patient tumor samples, as well as 5 LT PDX samples from stages I (1x), III (3x), and VII (1x) were used to call somatic SNVs and indels, using the complete-remission and bone-marrow donor samples as controls (sections 7.1.2 and 7.1.6). SNVs and indels were filtered based on the normal controls, depth > 50, and a maximum VAF of 0.75 (figure 2.3). SNVs that passed all

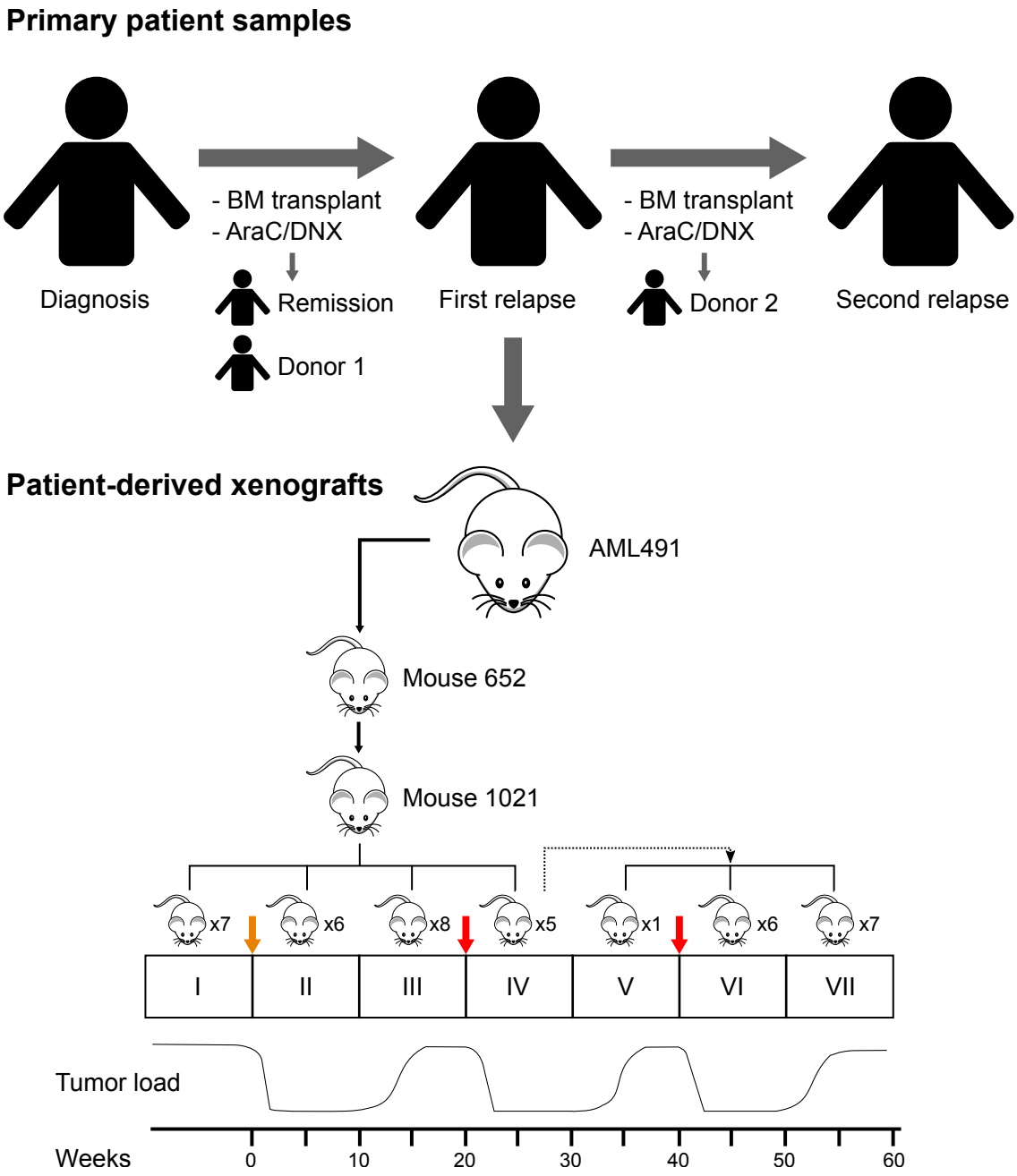


Figure 2.1: Scheme of the patient and PDX sample generation for the long-term experiment. In the PDX block, the orange arrow indicates chemotherapy regime 1 (2x AraC 50 + DNX; 1x AraC 100), and the red arrows indicate regime 2 (3x AraC 100).

Figure 2.2: PDX samples used for the multi-omics analyses of the AML-LT experiment, divided by chemotherapy stage.

filters were used to infer clone clusters using SciClone (Miller et al. [2014]). SciClone estimated 8 clusters in total (figure 2.4). Clusters 1 and 2 were fixed in the PDX samples. NRAS-Q61K was assigned to cluster 3, and KRAS-G12A to cluster 4. Cluster 7 was only detectable in the second-relapse sample, and contained the EZH2 SNV. Cluster 5 contained variants with low and inconsistent VAFs across all samples, thus being a likely artifact, which was in contrast with the consistent VAF values of variants in the other 7 clusters. Therefore, cluster 5 was excluded from subsequent analyses.

I used the remaining 7 clusters, comprising 6384 variants, to infer the clonal phylogeny and frequencies per sample with the ClonEvol package (Dang et al. [2017]). The inferred clonal phylogeny (figure 2.5) indicates linear evolution from diagnosis to relapse, followed by branching into the KRAS, NRAS, and second-relapse subclones. The KRAS and NRAS subclones produced an additional distinct subclone each, which were only detectable in PDX samples from later stages (III and VII).

Clonal fractions varied in the patient, and in the PDX samples throughout the long-term experiment (figure 2.6). In spite of the limited number of samples, this analysis provides a glimpse into the clonal dynamics that take place throughout chemotherapy stages. The KRAS and NRAS subclones were not detectable in the patient samples; furthermore, these had a considerable germline fraction. In the PDX samples, the stage I sample (28675) was predominated by KRAS, even though the NRAS subclone was still detected with a fraction of about 0.055. Two samples from stage III (26669 and 28648) had comparable fractions of both KRAS and NRAS, with sample 26669 also presenting the KRAS-2 subclone. The third sample (27077) was NRAS-predominant, with a KRAS fraction of 0.051. The stage VII sample (26617) consisted entirely of the NRAS and NRAS-2 subclones.

2.2.2 Mutational Profiles and Signatures

Given the well-delimited mutational profiles of each specific subclone, I was interested in analyzing whether there these showed any particular patterns. GC-loss mutations were the most abundant in

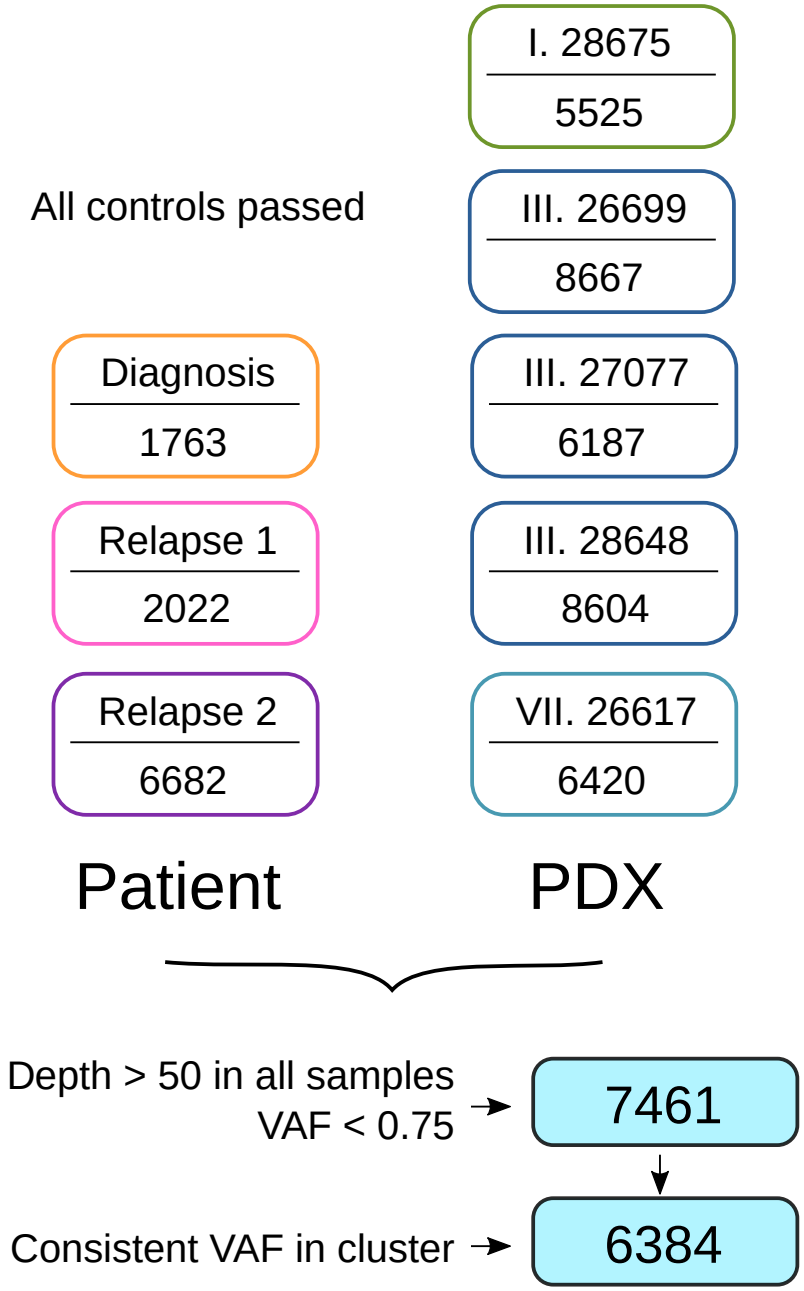


Figure 2.3: Number of SNVs and indels that passed the filters in WGS. In the upper part, the number of SNVs that passed the variant calling filters against all corresponding controls are shown. The lower part shows filters for depth, VAF, and cluster consistency across all samples. Complete remission was used as a control for all samples; Donor 1 was used as a control for first relapse, second relapse, and PDX; Donor 2 was used as a control for second relapse.

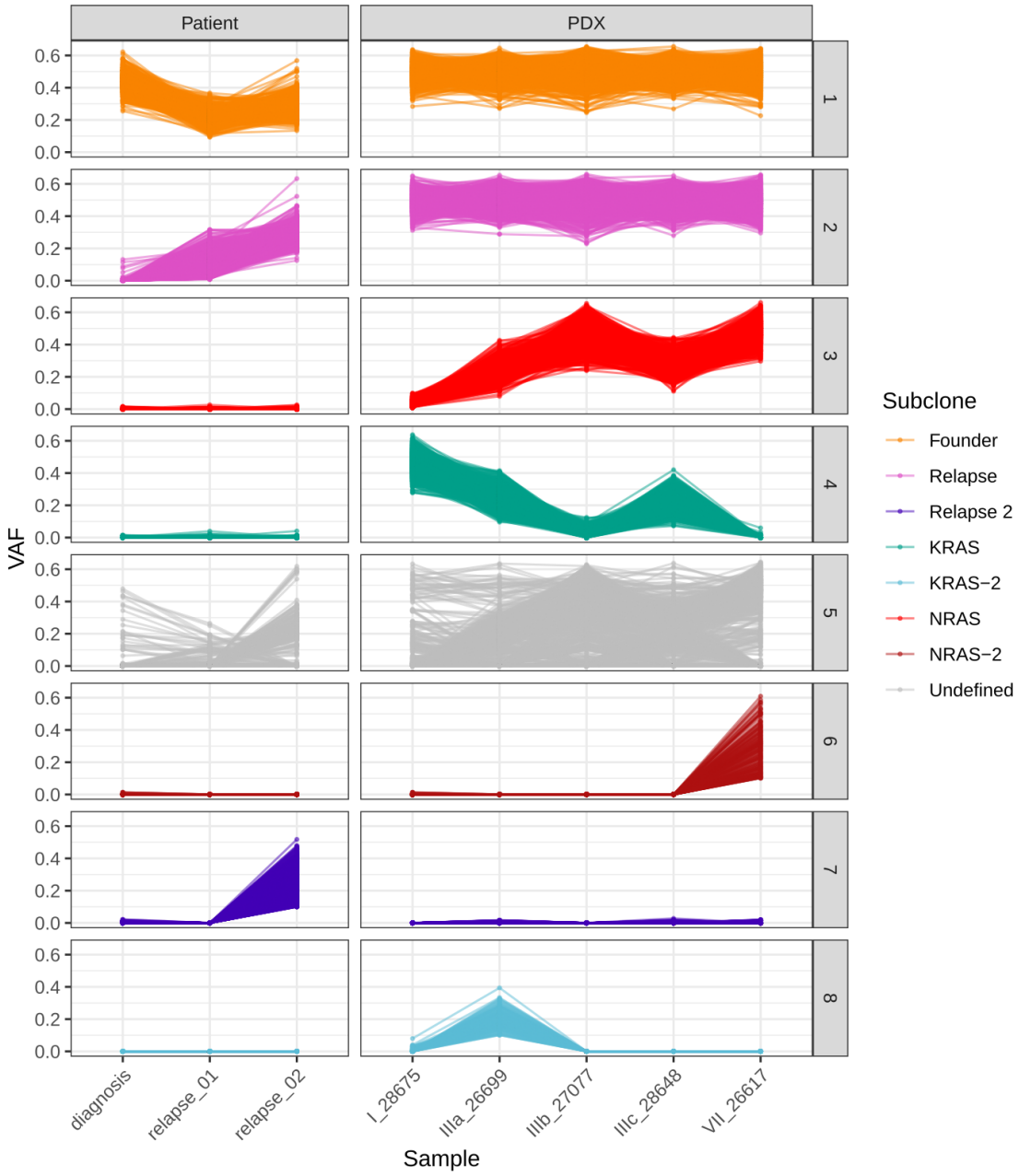


Figure 2.4: Variant allele frequencies of WGS SNVs and indels in the clusters inferred by SciClone.

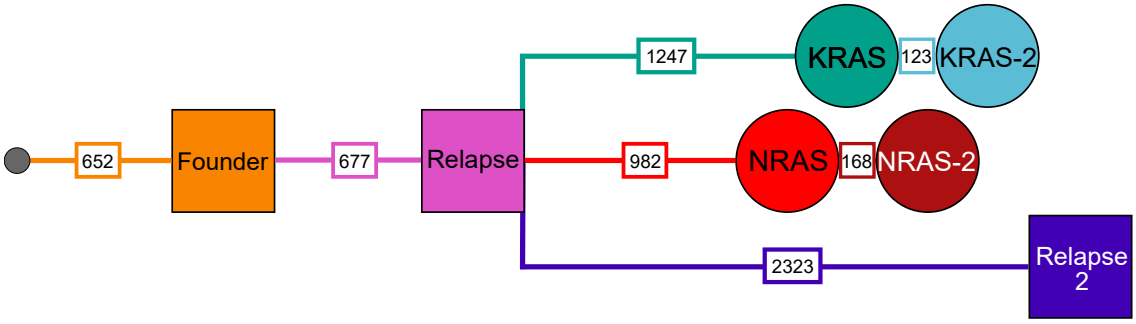


Figure 2.5: Clonal phylogeny of the AML-LT WGS dataset. Numbers in branches indicate the amount of SNVs that defined the clone or subclone. Squared nodes indicate clones that were present in the patient samples, and circular ones indicate subclones in the PDX samples.

the KRAS, NRAS, and second-relapse subclones (figure 2.7). An analysis of mutational signatures of SNVs that were unique to each subclone, performed by Christopher Alford, revealed that COSMIC signatures 18 and 24, with prevalent GC-loss mutations, were the most prevalent in these three subclones (figure 2.8; Christopher Alford, personal communication). Signature 18 is of particular interest, as its proposed etiology is the exposure to reactive oxygen species (ROS). This signature has been previously detected in several leukemia samples (Ma et al. [2018], Alexandrov et al. [2020]). In contrast, signature 1, which is associated with aging, was prevalent in the trunk clone, but was absent from KRAS, NRAS, and second-relapse.

2.2.3 Using the mutational profiles to estimate the timeline for subclone emergence

I estimated approximate age ranges per clone following a method that was introduced by Körber et al. [2019]. In the original approach, they used the proportion of changes in mutational signature 1 (which was taken as the molecular clock) between diagnosis and relapse tumors to approximate changes in somatic mutation rate and growth rate. This information was used to calculate the approximate age of the tumor. I adapted this approach to be able to estimate the age of each individual subclone, based on the prevalence of signature 1 on the sets of subclone-specific mutations (section 7.1.10). However, this had the limitation that it was not possible to estimate approximate times for the KRAS, NRAS, or second-relapse subclones, as their fractions of signature 1 were essentially 0.

The clonal ages for the founder and first-relapse clones are interpreted as time before obtaining the corresponding patient sample (in weeks). To translate them into actual times before and after initial diagnosis, I subtracted these values from the date at diagnosis. According to this estimate, the founder clone was established about a year before diagnosis (estimate with mean growth rate: 55.125 weeks before; range: 66.124 - 47.263 weeks; figure 2.9). The relapse clone was estimated to have been formed between 5.714 and 7.994 weeks before diagnosis, albeit at very low frequencies. The exact age of KRAS and NRAS clones could not be estimated due to the lack of signature 1 prevalence. However, I have tentatively placed them as formed by the time of diagnosis due to detection of the KRAS and NRAS variants at very low frequency in the HaloPlex dataset.

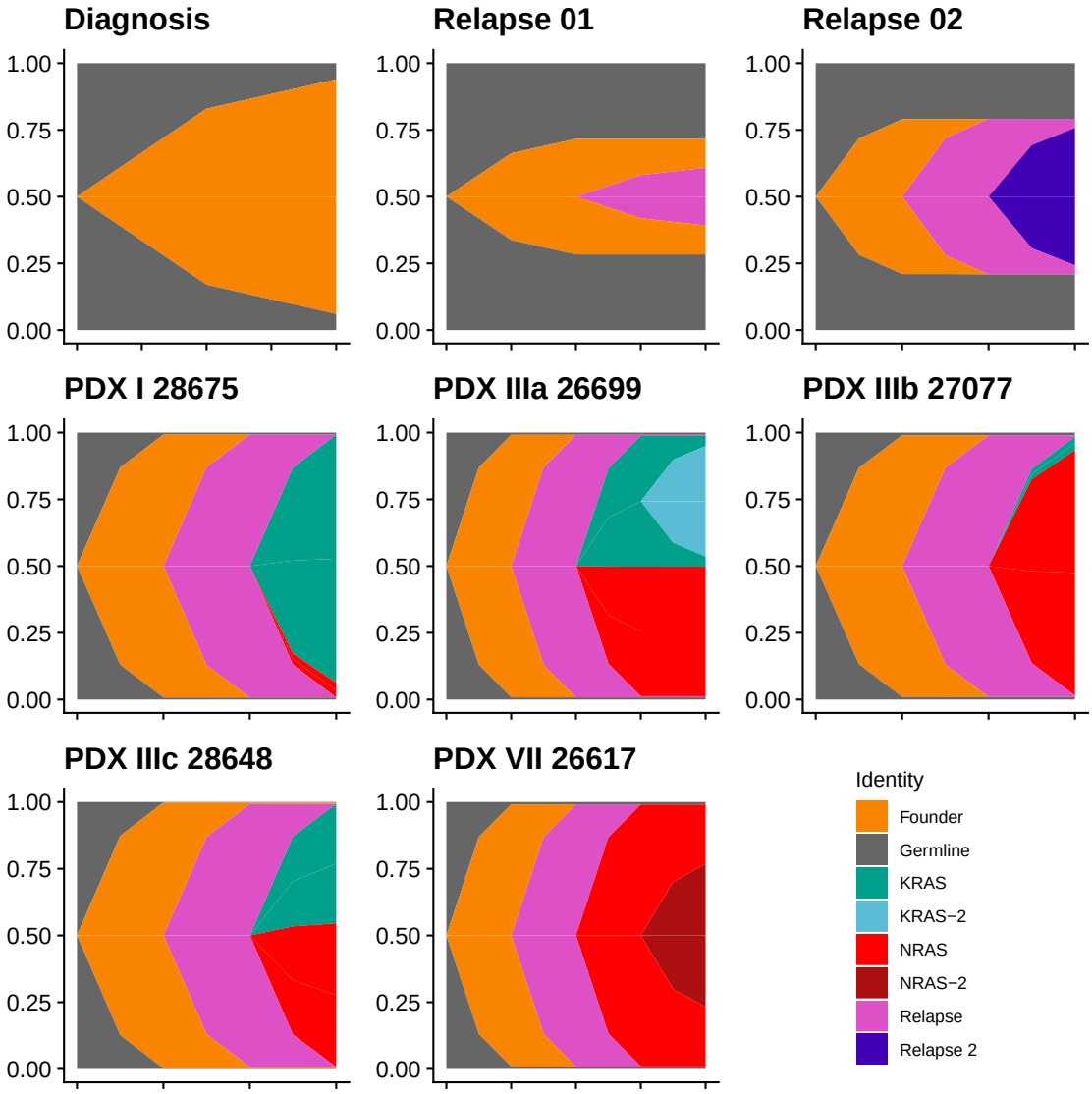


Figure 2.6: Clonal frequencies in the AML-LT WGS samples, represented as bell plots. The x axis indicates the emergence of each subclone, but does not represent real time. The frequency that was actually observed in the sample corresponds to the last point. Bell plots were generated using ggmuller (Noble [2019]).

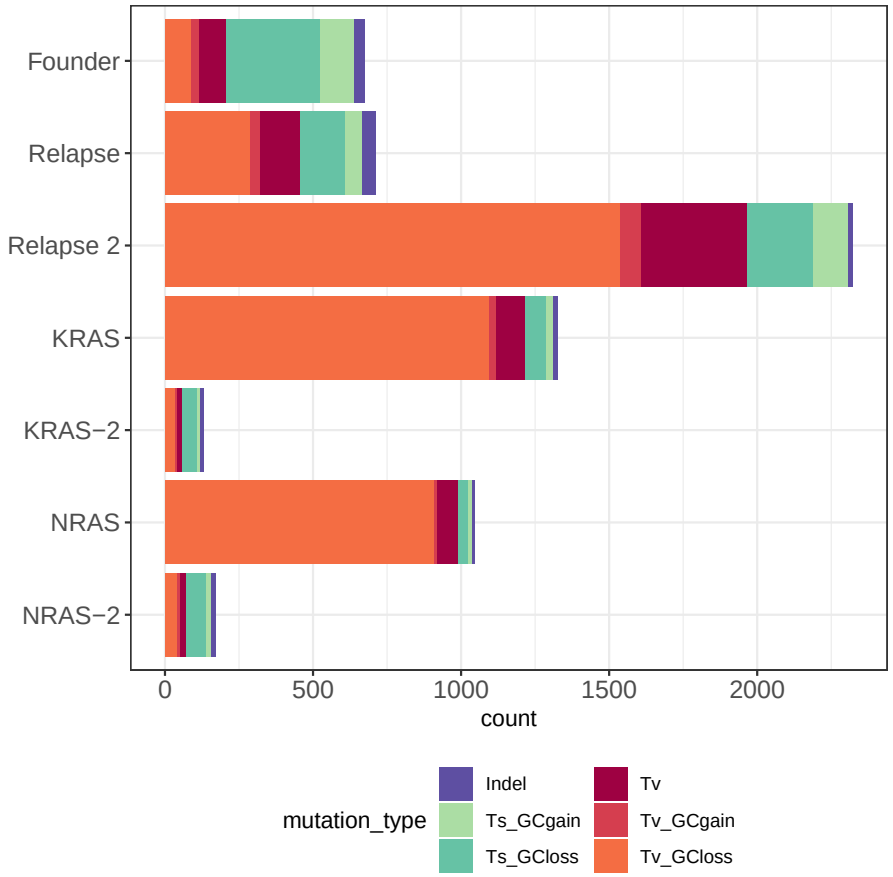


Figure 2.7: Number of variants per subclone that fall into the different types of substitution, or indels.

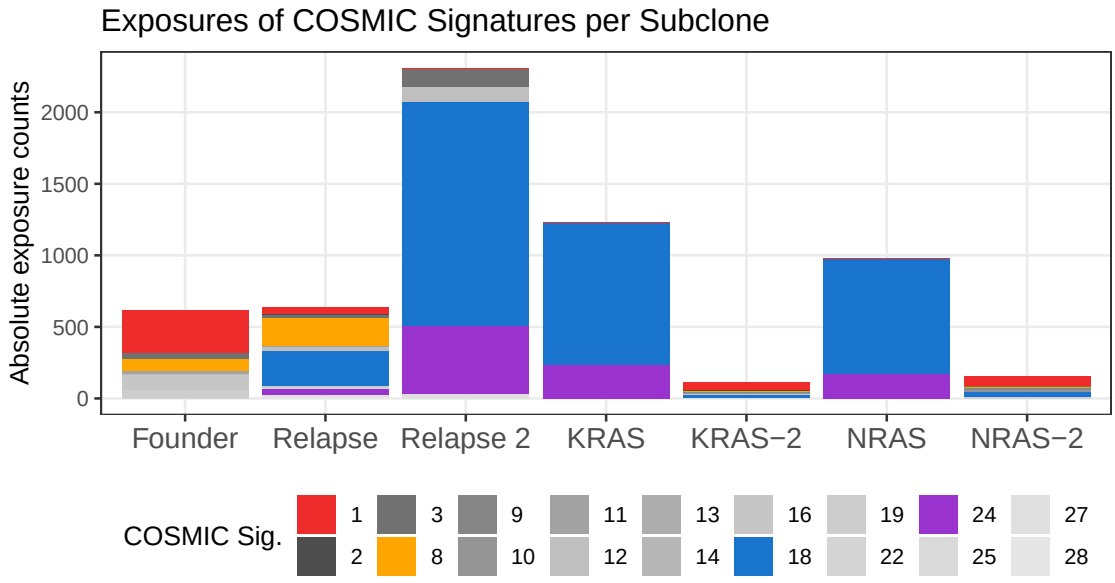


Figure 2.8: Mutational signature prevalence in the AML-LT WGS subclones. Courtesy of Christopher Alford.

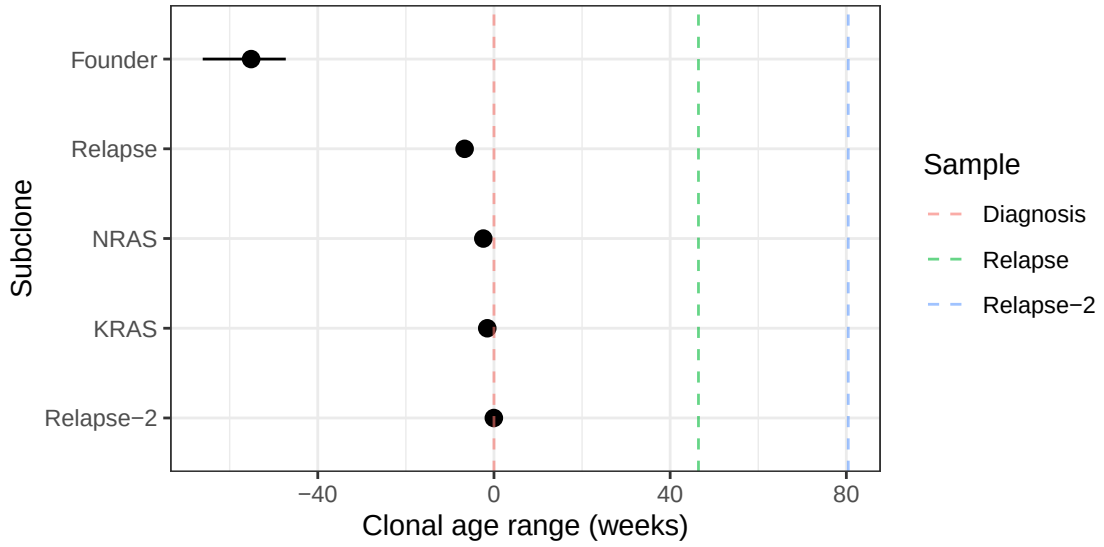


Figure 2.9: Estimated age per subclone with respect to the time of diagnosis. The x axis shows the time in weeks before or after diagnosis. Colored dashed lines indicate sampling times for diagnosis, first relapse, and second relapse.

2.3 Clonal Inference in Whole-Exome and Targeted Sequencing Data

2.3.1 Overview of the Dataset

The analysis of the WGS samples was important to establish the general clonal architecture that was present in the patient and that was preserved in the PDX samples. However, with only one sample from stages I and VII, it was not possible to determine if the frequencies of the KRAS and NRAS were necessarily the same in all samples from these stages. Furthermore, the three samples from stage III might not have been representative of how the KRAS and NRAS subclone frequencies changed after the first treatment with chemotherapy.

In order to obtain a more comprehensive picture of the frequencies of these subclones before and after each treatment, I used whole-exome and targeted sequencing data from additional samples (figure 2.2). WES data was available for 13 samples from stages I, III, and VII. Two kinds of TAS data were produced: HaloPlex (23 samples from stages 0, I, III, and VII), and SimSen-Seq (7 samples from stages 0, IV, and V).

2.3.2 Copy-Number Variants Were Stable in the PDX Samples

Given the diversity of samples that were available in the WES dataset, I considered that these would be useful to estimate a the whole diversity of CNV regions throughout all treatment stages. Allele-specific copy number variants were called on the exome data against normal controls using the MARATHON pipeline (Urrutia et al. [2018]). Genomic regions with 1 or more adjacent CNVs were filtered for a minimum length of 10 Mb. The full pipeline for CNV calling and filtering is described in chapter 7.1.8. 5 major CNVs could be detected in patient and PDX samples (figure

Chr	Start (Gb)	End (Gb)	Size (Mb)	Samples	Genes	Type
chr6	1.727	28.358	26.632	R2	812	Del; + Dup (R2)
chr6	28.358	36.713	8.355	R2, PDX	506	Del
chr7	90.004	158.851	68.848	R1, R2, PDX	1395	Del
chr16	31.771	33.962	2.191	D, PDX	130	Del
chr16	33.962	46.637	12.676	PDX	87	Del

Table 2.2: AML-LT CNV regions. Genomic coordinates (hg19), size, affected samples, and overlapping genes are shown. D = Diagnosis, R1 = Relapse 1, R2 = Relapse 2, PDX = PDX; Del = deletion; Dup = duplication.

2.10, table 2.2). Interestingly, the same unique set of CNVs was detected in all PDX samples, irrespective of the stage.

Two CNVs were present since the emergence of the founder or first-relapse clones. A 2.19-Mb deletion in chromosome 16 was observed in the diagnosis and PDX samples, and is adjacent to a 12.66-Mb deletion that is exclusive to the PDX samples and affected both chromosomes. A 68.85-Mb deletion in chromosome 7, covering most of the q-arm, was present in the first relapse, second relapse, and all PDX samples. One of the driver mutations of the second relapse, in the EZH2 gene, is in the major copy of this deletion. Finally, a complex set of CNVs can be observed on chromosome 6. An 8.35-Mb deletion was observed in the PDX samples, which could be affecting both major and minor copies similarly (or which could not be definitely associated to one chromosome). This same CNV was present in the second relapse sample, but was detected as a deletion in the minor copy and a duplication in the major copy. It is adjacent to a 26.63-Mb deletion/duplication in the second relapse.

In the pathway enrichment analysis of the CNV genes that were shared among all PDX samples (and thus both KRAS and NRAS subclones), the most significantly enriched pathways are related to immunological responses: interferon signaling, MHC antigen presentation, TCR signaling, and other related pathways. As seen on figure 2.12, this enrichment can be partly explained by the abundance of HLA genes that are located in chr6. Interestingly, there was also significant enrichment for oncogenic MAPK signaling, with genes such as BRAF, RASA4, and KDM7A. There was also enrichment for metabotropic glutamate/ pheromone receptors due to the abundance of TAS2R genes (in chr7), which encode taste receptors.

2.3.3 Inferring the clonal phylogeny with the Canopy package

Somatic SNVs and indels were also called in the WES patient and PDX data with MuTect2, following the GATK best practices. After filtering for enriched exonic regions, depth, VAF, and CNV overlap (see 7.1.6), 75 variants were kept for further analysis (figure 2.13).

The 75 SNVs and indels were used to re-infer the clonal architecture in the PDX samples using the Canopy package version 1.3.0 (Jiang et al. [2016]). The second-relapse sample was not included in the analysis, as variants from its corresponding subclone had not been called in the PDX samples. CNV events were not considered for clonal inference, as the identical set was called in all PDX samples and would therefore not be informative of the phylogeny.

The Canopy tree reflects the initially linear and then branching structure that was obtained from WGS, even though the NRAS-2 subclone was not detected (figure 2.14). In this tree, the

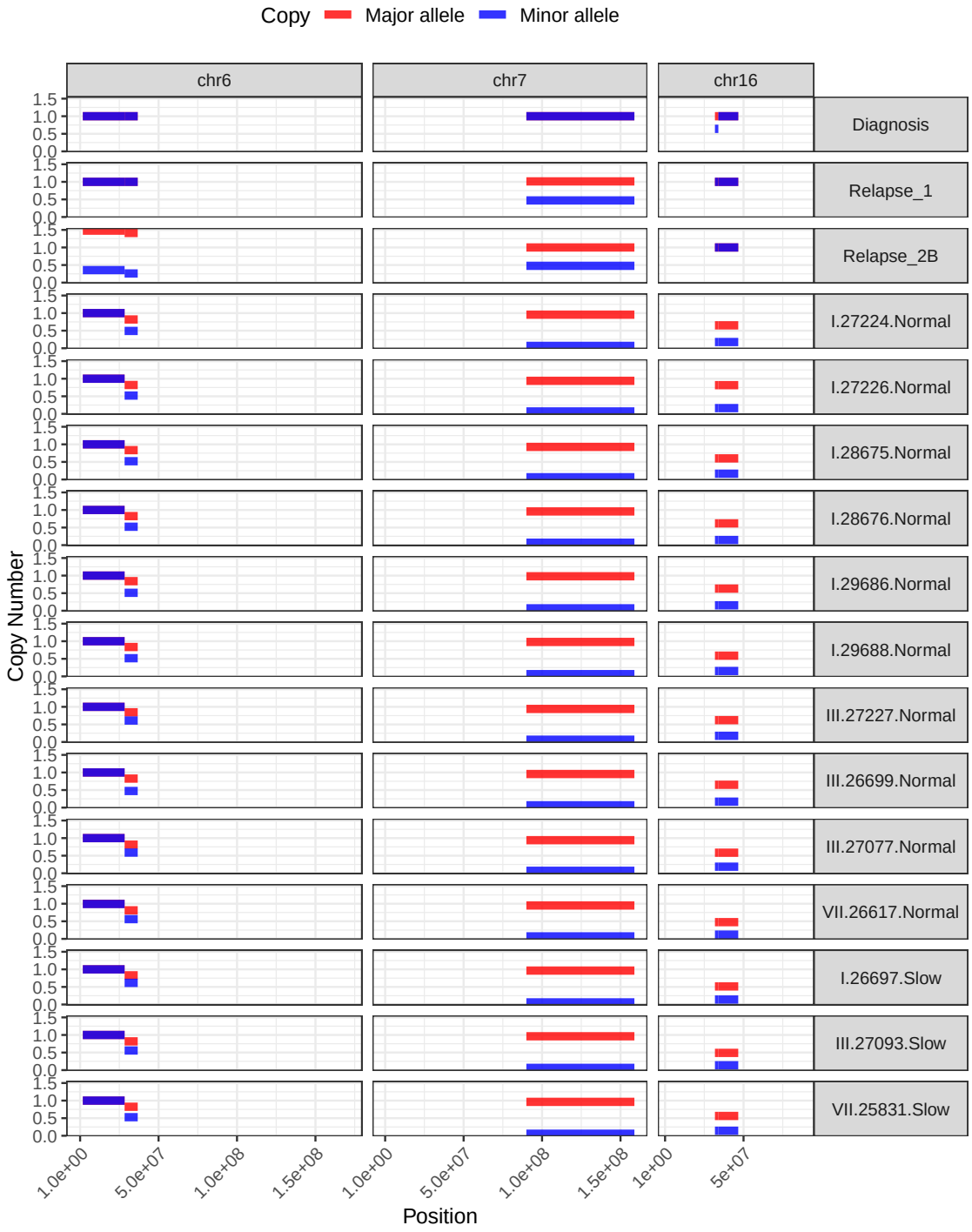


Figure 2.10: CNVs in the AML-LT exome dataset (patient and PDX samples). The x axis indicates chromosomal coordinates; full chromosome lengths are shown. The y axis indicates the copy number of the major or minor copy (which are indicated with colors).

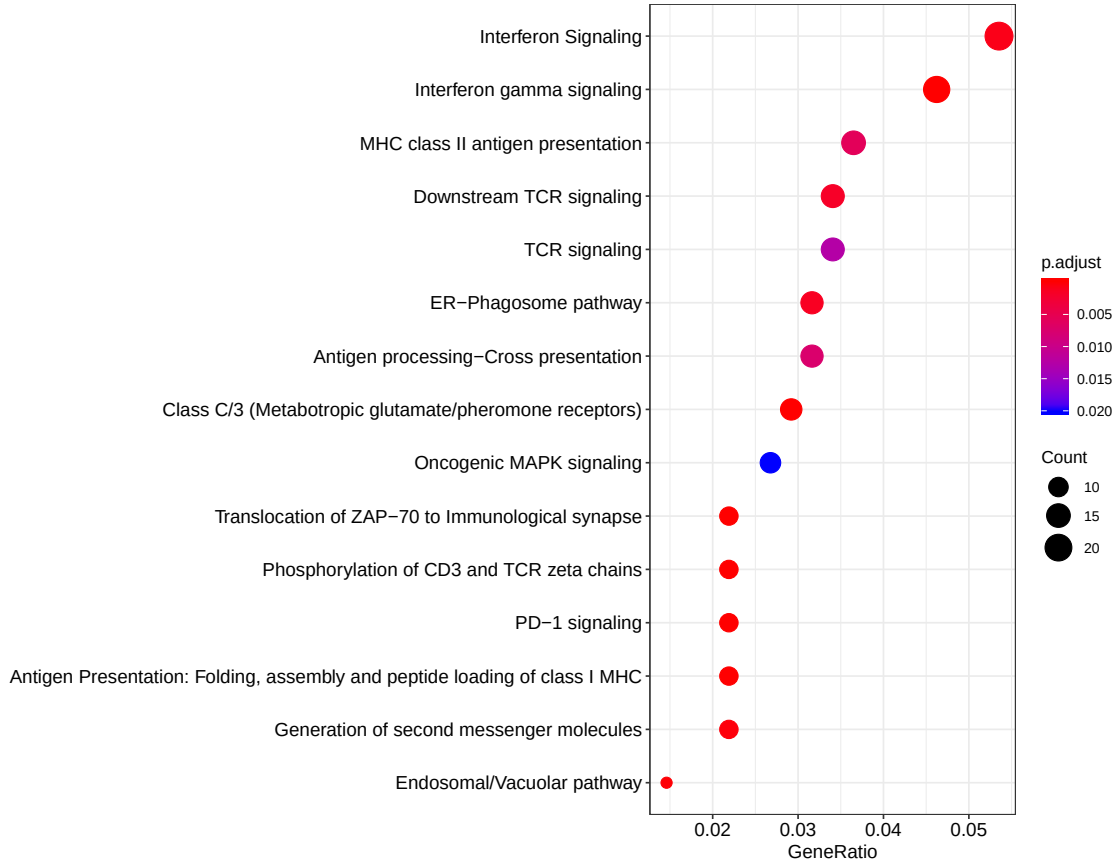


Figure 2.11: Reactome pathways that were enriched for genes in the CNV regions that were shared across all PDX samples.

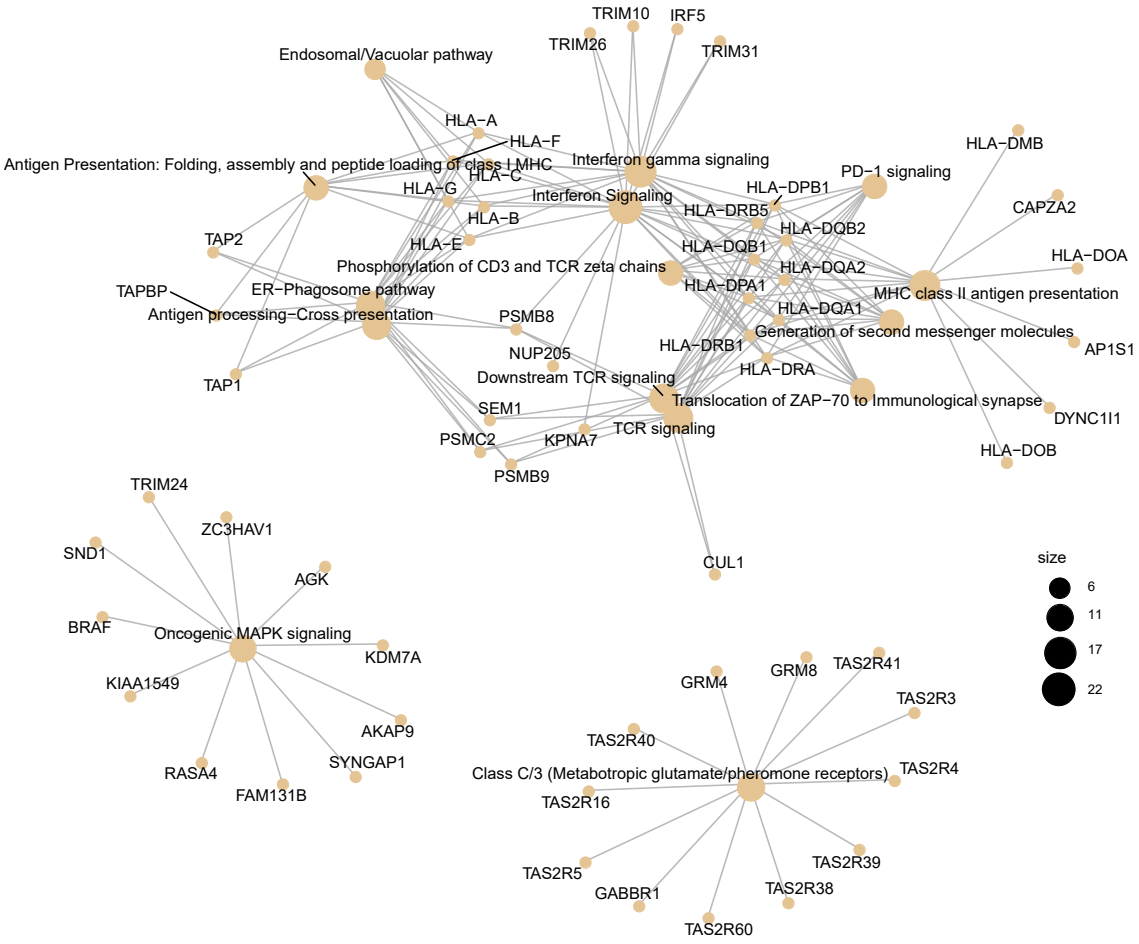


Figure 2.12: Gene networks of the reactome pathways that were enriched for genes in the CNV regions from PDX samples.

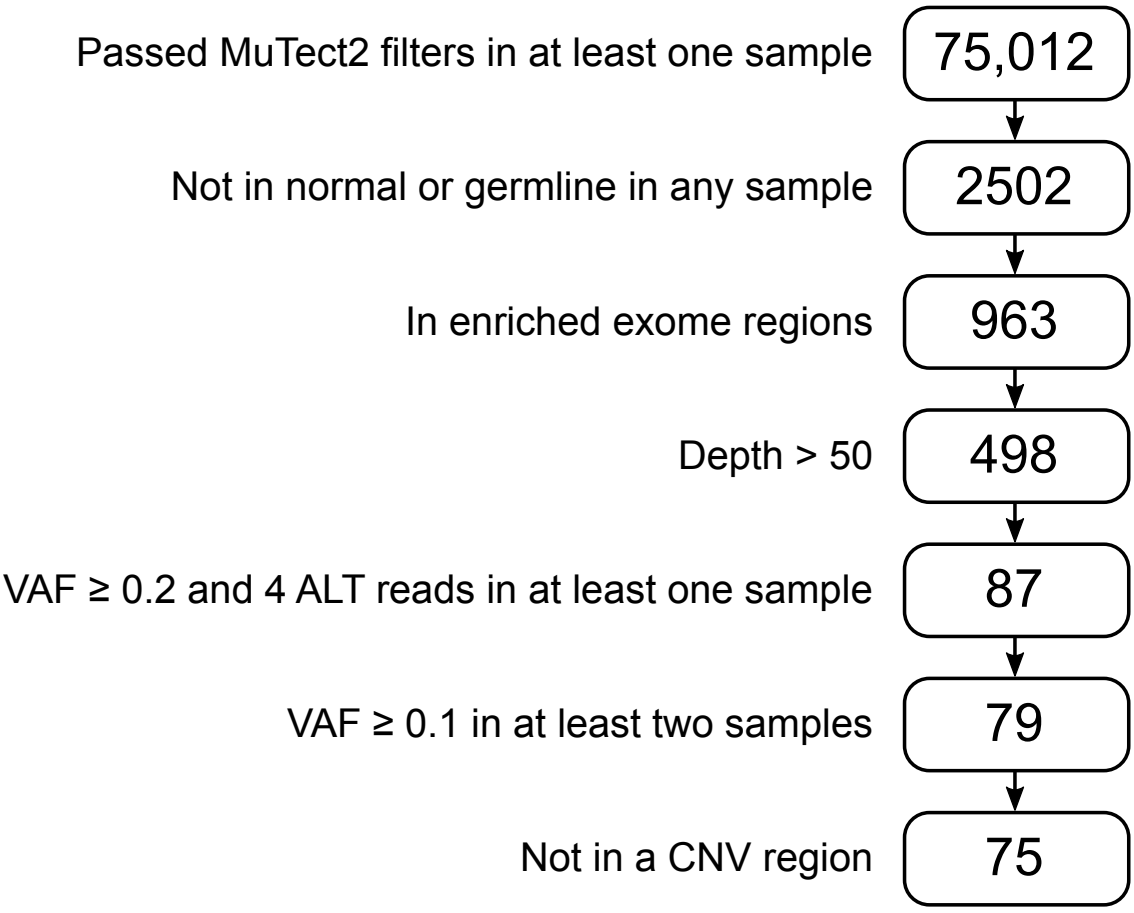


Figure 2.13: Exome SNV filters and remaining number of variants on each step

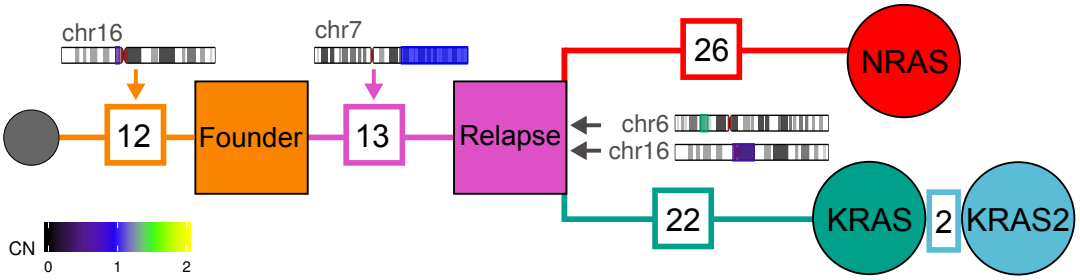


Figure 2.14: Clonal phylogeny inferred from the AML-LT exome dataset (without second relapse). Numbers in branches indicate the amount of SNVs that defined the clone or subclone. CNV events are indicated in the branches where they occurred; the added copy number (major + minor copy) is color coded.

NRAS subclone had more associated variants than KRAS + KRAS-2. The figure also shows the branches in which CNV events occurred, and these were all present in the final PDX samples (the last CNVs in chr6 and chr16 appeared after the relapse clone).

All 12 variants of the founder clone were completely fixed in the PDX samples, with VAFs generally close to 0.5 (figure 2.15). VAFs from the first-relapse clone were also mostly constant and in the range between 0.4 and 0.6, although some also had values between 0.2 and 0.4. VAFs from variants in the KRAS and NRAS clones were generally within close ranges in the samples where these variants appeared. One exception was one single SNV in the KRAS subclone, which had lower VAF values than the rest of the KRAS SNVs. This SNV was assigned to another KRAS branching subclone, but since this was not suggested by the WGS analysis and it was a single point of evidence, I assigned it to the initial KRAS subclone for now. The KRAS-2 subclone consisted of only 2 variants at this point, which were not simultaneously detected in a same sample. This could also be attributed to noise due to low frequencies.

2.3.4 Integration with ultrasensitive amplicon data

To obtain KRAS and NRAS clonal fractions in samples that were not subjected to WGS or WES, I employed the data from two deep targeted sequencing approaches. The first one was the aforementioned HaloPlex analysis (section 2.1.1). For the second approach, SNVs of the KRAS and NRAS subclones were sorted by coverage across samples and the correlation of their VAFs with respect to those of the KRAS and NRAS SNVs (Spearman's rho). The top sites were evaluated for multiplexed amplification and iteratively discarded until a suitable subset was found. These sites, which are shown in table 2.3, were used by Daniel Richter for genotyping additional PDX samples from stages 0, IV, and V using SiMSen-seq, which is a targeted, UMI-based ultra-sensitive sequencing protocol (Ståhlberg et al. [2017]). He called the targeted SNVs using MAGERI, version 1.1.1 (Shugay et al. [2017]).

To estimate clonal fractions from additional samples where only targeted amplicon re-sequencing data were available, KRAS and NRAS SNV VAFs were used directly to estimate their clonal frequencies. The mean VAF of SNVs in the panel that were assigned to the trunk of the phylogeny were used to estimate the total tumor fraction in the sample. The difference of the tumor fraction minus the KRAS + NRAS fractions was assigned to a combined founder/first relapse clone, as

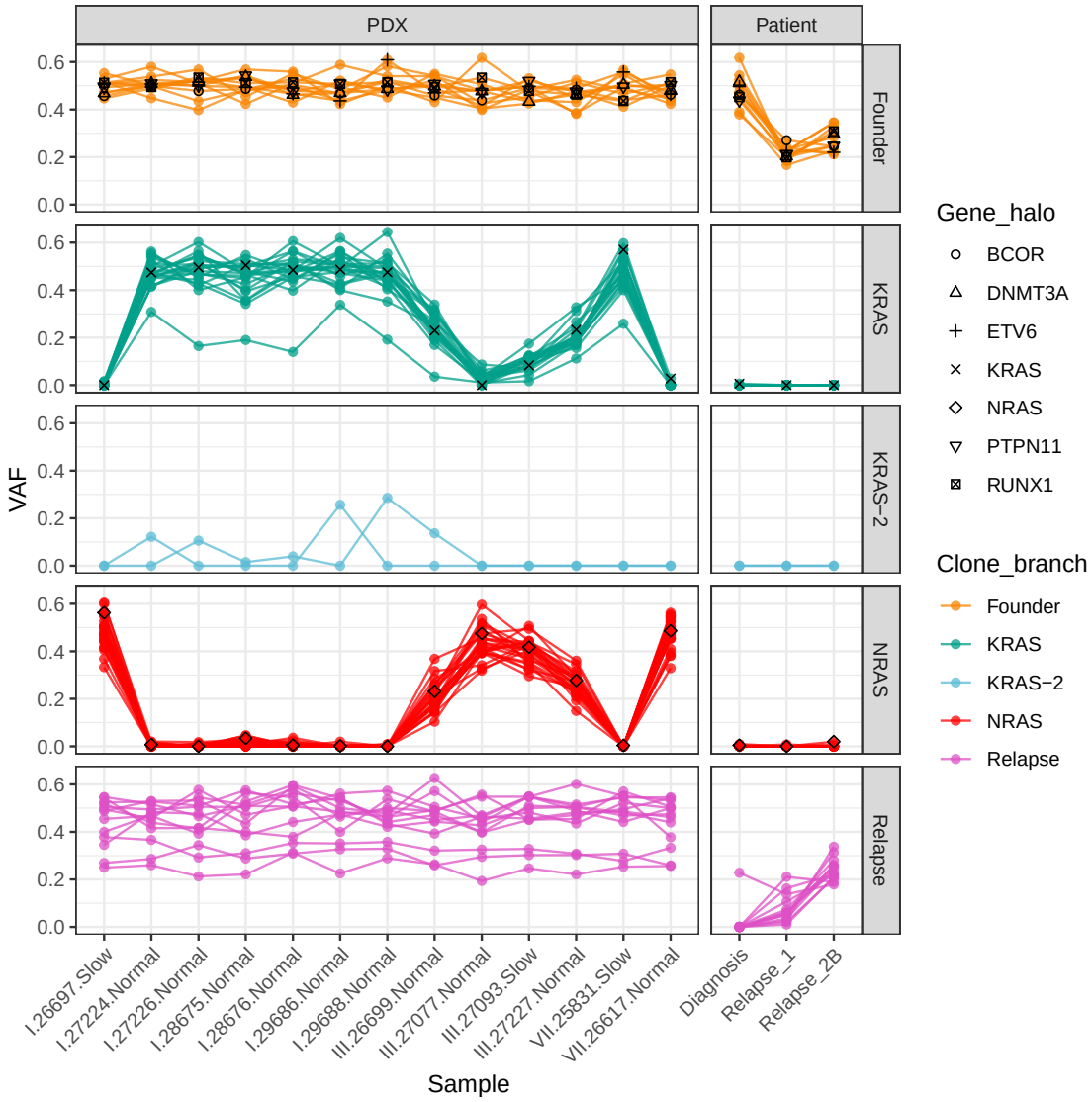


Figure 2.15: VAFs of the variants that define each clone in the AML-LT exome dataset. The plot is faceted vertically by sample type (PDX or patient), and horizontally by subclone. Shapes indicate SNVs in genes that were analyzed by HaloPlex, and their VAFs were obtained from HaloPlex read counts.

Gene	Chrom	Position	Ref	Alt	Subclone
CREB3	chr9	35735991	G	T	KRAS
CSMD3	chr8	114111154	C	A	NRAS
KRAS	chr12	25398284	C	G	KRAS
NLRP3	chr1	247588686	G	T	KRAS
NRAS	chr1	115256530	G	T	NRAS

Table 2.3: AML-LT variants targeted with SiMSen-seq. Genomic coordinates and associated clone (from the exome analysis) are shown.

the latter could not be distinguished. The germline fraction was calculated as 1 - tumor fraction.

The addition of samples with respect to the WGS dataset helps to complete the picture of the clonal dynamics (figure 2.16). One interesting feature is the detection of a low fraction of the relapse clone in the diagnosis sample. Next, it can be observed that stages 0 and I show a predominance of KRAS (either pure or KRAS + KRAS-2) in most samples, which becomes an intermediate frequency in most samples of stage III. However, from stage IV onwards most samples tend to have one predominant subclone once again, with a majority of them showing a larger KRAS fraction.

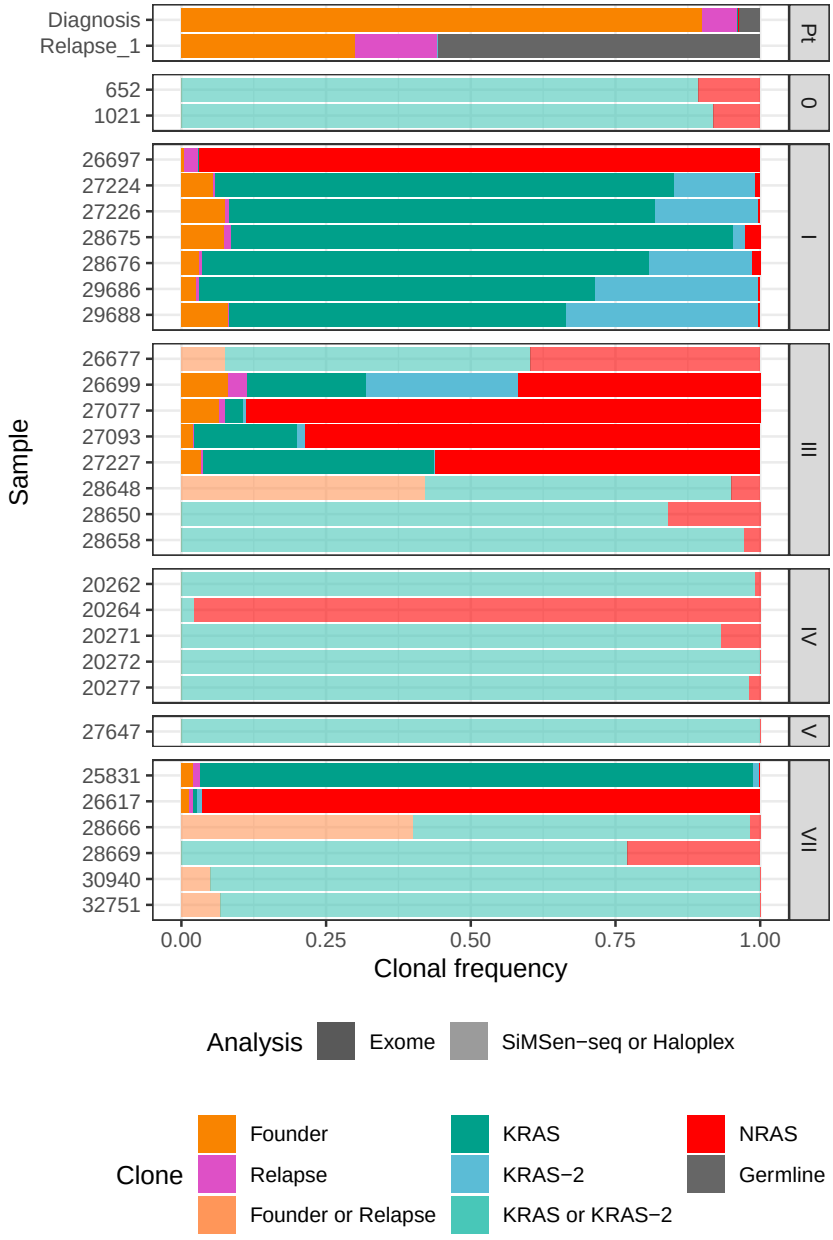


Figure 2.16: Clonal frequencies in the AML-LT exome and targeted datasets. The type of dataset is indicated by the transparency.

Chapter 3

umivariants: An R Package for Analyzing Clonal Variants in Sequencing Data with Unique Molecular Identifiers

3.1 Calling Variants from Single-Cell RNA-Seq with Unique Molecular Identifiers

3.1.1 An approach to proofread variants in scRNA-seq for the study of clonal heterogeneity

As demonstrated by the case of the AML-LT experiment, the detailed analysis of somatic variants is important to infer cancer heterogeneity and to track clonal frequencies. For this very purpose, it is critical to tap into the potential of single-cell techniques such as scRNA-seq methods. These can not only provide the highest possible resolution into the subclonal variant composition, but have become sufficiently accessible and affordable to enable their widespread use. However, as described in section 1.3, these methods rely heavily on PCR amplification to increase the signal from the starting biological material. The amplification step is a potential source of sequence errors that can lead to call false positive variants, or to mask the true ones. UMIs therefore represent an important tool to proofread such errors, as amplification duplicates can be collapsed into a single value that removes the signal of errors that come from individual reads.

My aim is to improve variant calling in scRNA-seq in order to enable a precise estimate of the subclonal composition in tumor datasets. To fulfill this goal, here I propose a general approach for the application of UMIs to proofread sequences and call variants in scRNA-seq (figure 3.1). In principle, it would not depend on direct variant discovery on the scRNA-seq dataset, but require a set of variants that have been called and associated with clones in data from bulk DNA-seq methods, such as WGS or WES (as recommended by McCarthy et al. [2020] or Petti et al. [2019]).

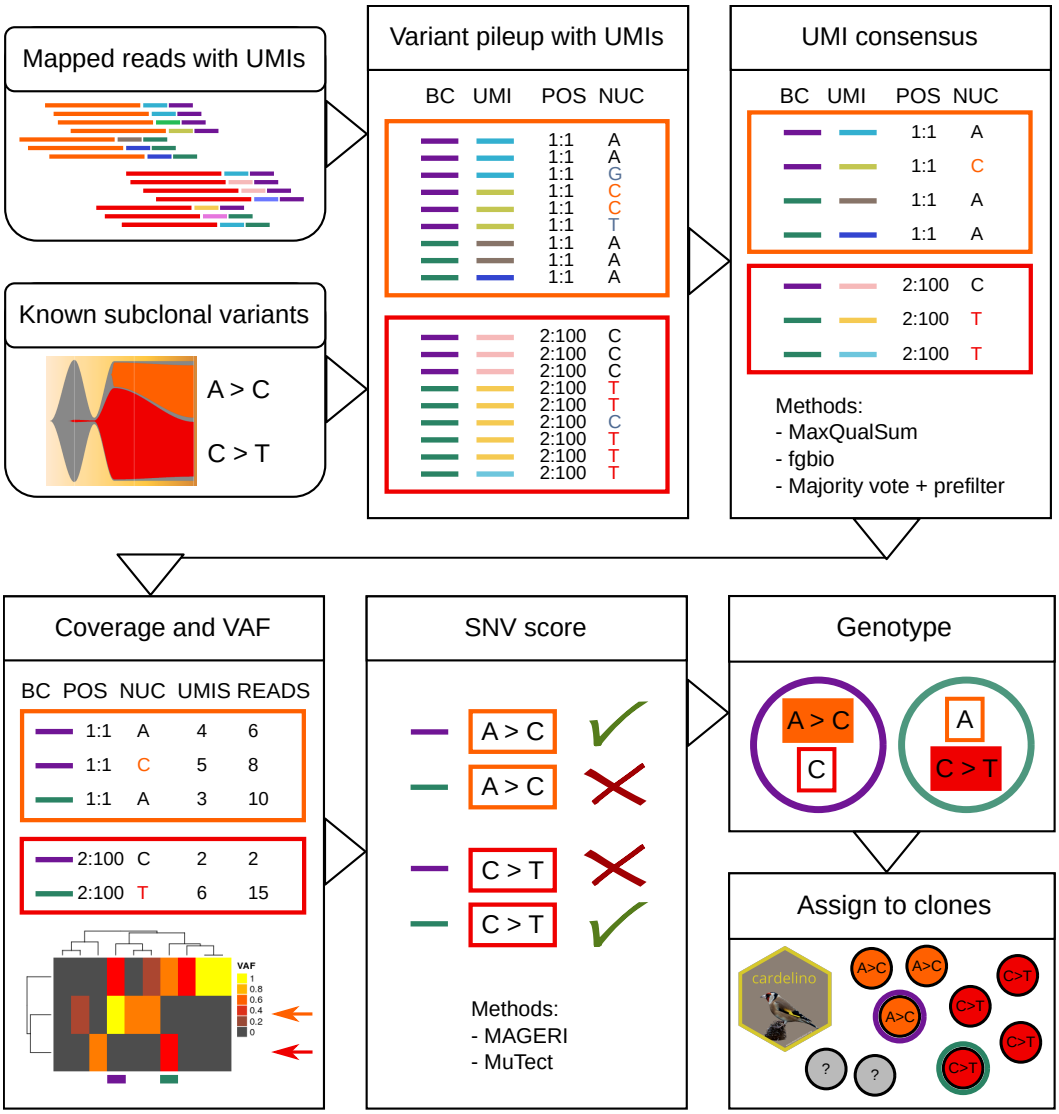


Figure 3.1: Flowchart with the approach for variant proofreading and calling in scRNA-seq using UMIs.

Sequences of the variant sites would be extracted from the scRNA-seq reads to generate a pileup table. The proofreading step would be performed on the pileup by generating a consensus across all values per UMI per position. After proofreading, the UMI-consensus sequences are used to calculate the coverage (total and per allele) of each variant per cell. These corrected coverage values are provided to a variant calling model, which gives a confidence score. This information is used to determine the genotype of each variant per cell. The mutational profile of each cell is finally used to assign it to one of the clones from the bulk DNA-seq annotation. The steps of this approach have been implemented in the *umivariants* R package.

3.1.2 Extracting variants and reads with UMIs

In order to extract the variants of interest and their sequences in the scRNA-seq dataset, input files with the corresponding information have to be provided to *umivariants*. Sequences at the variant positions are scanned in the file with the UMI-barcoded reads, and the nucleotide or indel value at the site is reported per read with its corresponding UMI and sample barcode (BC) sequences. In *umivariants*, this is achieved with the scan_UMI_bam() function.

To extract the variants of interest, the user needs to provide the following information in an input file: their chromosomal coordinates (chromosome name, start position, optionally end position and strand), the reference allele, and the variant allele. The input file with the variants can be provided in three different formats. The first one is the Variant Call Format (VCF), which was designed by the 1000 Genomes Project Analysis Group (Danecek et al. [2011]). This is the standard format of the output files produced by MuTect2 and HaplotypeCaller, as well as most other widely used variant callers. The second possibility is the BED format. This is a tabular format where the first three obligatory columns contain the coordinates of the variant. The actual alleles need to be provided in optional fields. The final possible input format is a custom text table, which needs to contain the required columns with the variant location and alleles.

scRNA-seq reads from which the variant sequences are to be extracted should be mapped with the appropriate software. I would recommend using zUMIs (Parekh et al. [2018]) with the STAR aligner (Dobin et al. [2013]), but CellRanger (Zheng et al. [2017]) is an appropriate pipeline for 10x datasets as well. Both zUMIs and CellRanger produce a BAM file (Li et al. [2009]) with tags (i.e. fields) that contain the UMI and sample BC sequences that were associated with each read. The exact name of the tags containing the UMIs and BCs needs to be provided to scan_UMI_bam(). If a pipeline that incorporates the UMI/BC sequences to the read identifier was employed, the sequences can be parsed using separators or regular expressions.

Variant sequences are extracted from the reads in the BAM file together with their corresponding UMIs and BCs. The function is capable of handling SNVs, indels, and multi-nucleotide replacements; it can also make use of multiple threads to process multiple variant sites simultaneously (see section 7.2.2). The output of this step is a pileup table that contains the information of the start position of the variant (chromosome and location), the UMI sequence, the sample barcode (BC) sequence, the sequence quality score (Phred-scaled and ASCII-converted), mapping quality score, and optionally the read query name. Such table contains the necessary information for subsequent steps of UMI-consensus based proofreading.

3.1.3 UMI proofreading and collapsing with consensus

The UMI consensus step consists in collapsing the sequences from all reads that are associated with the same UMI, position, and sample BC into a single value. The goal of this step is to harness the power of all available PCR duplicates to reduce the number of amplification or sequencing errors that might lead into false positive variant calls. The principle behind the UMI consensus is the assumption that, since every UMI should be bound to a single transcript during the first-strand synthesis, a unique sequence is expected for any transcript-UMI combination. That means that all the reads from each UMI should have the same sequence. Any deviations would be attributed to amplification or sequencing errors.

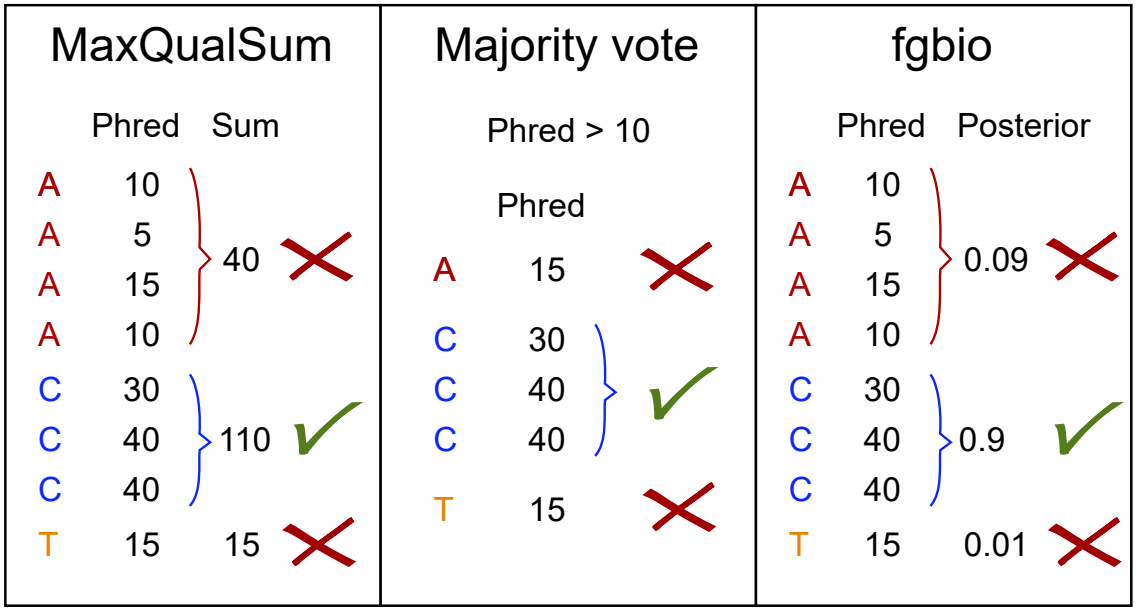


Figure 3.2: Graphical illustration of the UMI consensus methods implemented in *umivariants*

To compute the UMI consensus, parameters like the nucleotide frequency or sequence quality can be considered. Three different methods that make use of this information have been implemented in *umivariants* (figure 3.2; section 7.2.3). The first one is the **Majority vote**, in which the allele with the highest frequency is taken as the consensus. Sequences can be filtered to have a minimum quality value before computing this consensus. The second method is called **MaxQualSum** method, which stands for 'Maximum Quality Sum'. In this method, the consensus based on the sum of quality scores (Phred scores) of the reads of a UMI that have a certain allele (i.e. nucleotide or indel). This method was conceived as an alternative to the direct majority vote, in order to balance the evidence for each allele in cases in which multiple reads of low quality support one value, and fewer reads with high quality support another. The third consensus method available is **fgbio**, an implementation of the method from the fgbio software suite (Fennel et al.). This consists of a likelihood model of each possible nucleotide at the variant site, given their count and Phred scores. The consensus nucleotide is the one with the highest posterior probability, estimated from the likelihood of the nucleotides. The fgbio method provides a robust confidence estimate of the quality of the consensus sequence, but is only adequate for SNVs and not for indels.

The three UMI consensus methods estimate a Phred-scaled consensus quality score (CQS), which attempts to represent the sequence quality of the consensus reads. In the case of the majority vote method, the CQS is estimated with the method implemented in MAGERI (Shugay et al. [2017]), which is calculated from the frequency of the consensus allele in the UMI and scaled to a maximum value of 40 (see section 7.2.3). The CQS in MaxQualSum is equal to the mean Phred score of the sequences with the consensus allele. The CQS of the fgbio method is obtained from the posterior probability of the consensus nucleotide, after applying the likelihood model.

The output table of the UMI consensus step of *umivariants* contains the UMI, sample BC, position, consensus nucleotide, total reads per UMI, read count and fraction with the consensus nucleotide, and specific output of the consensus method. Based on this output, the proofread

sequence values and the UMI counts per sample can be used to compute the coverage of the variant site (total and per allele), as well as variant allele frequencies (VAF). These two metrics can be estimated and plotted by *umivariants* to give an overview of the support for variant calling, and are required for steps further downstream in the pipeline (variant scoring and genotyping).

3.1.4 Variant calling and genotyping per sample

After proofreading the input sequences with the UMI consensus, it is possible to perform variant calling on each of the single cells. To do this, the allelic coverage and CQS values need to be provided to a mathematical framework that can determine if the variant can be effectively detected in the sample. The variant caller can estimate a score that reflects the confidence on the detected variant.

In order to perform the variant calling and scoring, I have implemented two methods in *umivariants*. These are implementations of two published variant callers. The first one is **MuTect**. This is the Bayesian classifier of the somatic variant caller from the Genome Analysis Toolkit (Cibulskis et al. [2013]). The score of this method corresponds to a log odds (LOD) score of the likelihood of presenting the variant versus having only the reference sequence (section 7.2.4).

The second implemented SNV calling method is **MAGERI** (Shugay et al. [2017]). In contrast to MuTect, this model was designed for UMI sequencing data, where the consensus was computed. MAGERI is based on a beta-binomial model which was designed to account simultaneously for amplification and sequencing errors, based on empirically-determined polymerase substitution rates (section 7.2.4). One fundamental difference that needs to be considered in the *umivariants* implemetation of MAGERI with respect to the original software is that the latter performs an assembly of all the reads that belong to a UMI in order to generate a single consensus read, while the *umivariants* version only gets the consensus for a single site. The original MAGERI pipeline aligns the UMI consensus reads to the reference with a Smith-Waterman algorithm, while *umivariants* expects an input dataset that has already been mapped to the reference.

Each of these two scoring techniques has some limitations. The MuTect method requires extensive calculations of the variant calling likelihood, based on the Phred scores from all the consensus sequences. On the other hand, the MAGERI score is only defined for substitutions, not for indels, and to obtain fully accurate parameters, extensive analyses of the error rates of specific polymerases would need to be conducted.

The scores reported using either the MuTect or MAGERI methods can be employed to produce a final list that reflects the observed genotype of all variants of interest per sample. To this end, a threshold on the SNV score from the variant caller can be applied to determine if the variant was detected on each cell or not. Combined with the coverage and VAF values, the status can be further classified into one of the following genotypes: homozygous reference (0/0), heterozygous (0/1), homozygous variant (1/1), or undetermined (-/-). In scRNA-seq, it is important to consider that low expression and coverage values can lead to the detection of a single allele. Therefore, an additional heterozygosity score, based on the binomial distribution, was implemented in *umivariants* to label potential heterozygous variants when these are genotyped as homozygous variant due to low coverage (see section 7.2.5).

3.1.5 Assigning single cells to clones

After calling the variants and their corresponding genotypes on each of the cells, it is possible to proceed to the final step of the *umivariants* approach: inferring what is the clone or subclone from which each of the cells originated. If the expected clonal structure is available (i.e. from a WES or WGS analysis), the phylogeny can be used for clonal assignment based on the genotypes and coverage values.

Two methods were implemented for clonal assignment of single cells. The first one, **Simple Assignment**, labels the cell with the clone that has the highest number of detected variants (based on the genotype). If all the detected variants are shared between possible ancestral and daughter subclones, the method assigns the most ancestral one to the cell. The second clonal assignment method is **Cardelino** (McCarthy et al. [2020]), which takes the variant allele coverage, total coverage, and a the clonal phylogeny in matrix format as inputs. Cardelino assigns the clone to each cell based on a Bayesian mixture model that is applied to the coverage values, and the estimated posterior probability is provided in the output. Regardless of the approach, cells can be left unassigned if there is no coverage of the variants that can inform of the subclone.

3.2 Benchmarking the UMI Consensus and SNV Scoring Methods from *umivariants* for Variant Calling and Clonal Assignment

3.2.1 Benchmarking framework for UMI consensus and SNV score methods

It is expected that the usage of the UMI consensus as a proofreading step should increase the precision in variant calling and clonal assignments. However, this statement needs to be demonstrated with actual scRNA-seq data with UMIs. Furthermore, it is important to quantify the actual improvement in the variant and clonal assignment estimate that is obtained from the application of this procedure, if any. This lead me to benchmark the consensus and score functions of *umivariants* with an AML scRNA-seq dataset, using a set of ground-truth variants. The evaluation of the variant calls with and without the consensus methods was used to determine changes in false-positive variant detection through the false positive and false discovery rates, together with the impact on false negative variants (false negative rate). To evaluate how this affects subsequent clonal assignment, the benchmarking rates from the UMI consensus methods were used to simulate single-cell allelic coverage datasets from different clonal phylogeny configurations. These data were in turn used to calculate how many cells could be assigned to their true subclone of origin.

The data that I employed for this analysis was a scRNA-seq dataset that was generated for 28 PDX samples of the AML-LT experiment (figure 2.2; section 7.1.5). 2276 cells passed the initial quality filters, which included library size and a proportion of mouse reads below 30% (Johannes Bagnoli, personal communication). As a ground truth set of variants, I called the set of core heterozygous germline variants (i.e. with VAF = 0.5) from the AML long-term exome dataset on the scRNA-seq dataset (7.1.7). The rationale for this procedure was that heterozygous SNPs should be present in the genome sequence of all tumor cells of the PDX samples, and these should be detectable given enough coverage.

The effect of initial read coverage on the ability to detect the variants and reduce false postive

	In ground truth set	Not in ground truth set
Called in scRNA-seq	True positive (TP)	False positive (FP)
Not called in scRNA-seq	False negative (FN)	True negative (TN)

Table 3.1: Confusion matrix of the UMI-consensus method benchmark in AML-LT scRNA-seq data. The values were estimated for the variant call results of each combination of UMI-consensus method, SNV calling method, and SNP coverage threshold.

Rate	Formula
False Positive Rate (FPR)	$FP/(FP+TN)$
False Negative Rate (FNR)	$FN/(FN+TP)$
False Discovery Rate (FD)	$FP/(FP+TP)$

Table 3.2: Performance rates of the UMI-consensus method benchmark in AML-LT scRNA-seq data. The rates were estimated from the confusion matrix of each combination of UMI-consensus method, SNV calling method, and SNP coverage threshold.

calls was evaluated by subsetting the ground truth set of SNPs according to the minimum coverage of the site across all cells, ranging from 1 to 40. The three UMI-consensus methods (MaxQualSum, majority vote, fgbio) were used to proofread the sequences in the scRNA-seq dataset at SNP sites that passed each coverage threshold. The UMI consensus tables were used to perform variant calling with the MuTect or MAGERI methods, and the variants were classified as called in the scRNA-seq dataset if they passed the score thresholds (6.3 and 40, respectively). The variant calling results were compared to a modality where no UMI consensus method was employed. A confusion matrix was estimated from the variant calling results per UMI consensus method, SNV calling method, and coverage threshold to analyze how many of the called SNPs were true positives, false positives, true negatives, or false negatives (table 3.1).

The values of the confusion matrix of each combination of UMI consensus method, SNV caller, and SNP coverage were used to estimate a series of performance rates that reflect the proportion of true or false results in the scRNA-seq SNP calls. In particular, the usage of UMI consensus methods was expected to reduce the false positive (FPR) and false discovery rates (FDR), due to a better control of false positive calls. However, I also wanted to evaluate whether this step would also have a negative impact when calling true positive variants, and thus estimated the false negative rate (FNR) as well to evaluate how many true SNPs were missed.

3.2.2 The MaxQualSum Consensus and MAGERI Score Minimized the False Discovery Rate

In the benchmark analysis of the UMI consensus methods in the AML-LT scRNA-seq data, it could be observed that different combinations of UMI and SNV consensus methods were able to reduce the FPR, FDR, and FNR more efficiently depending on the different coverage values (3.3). Up to the 20-read coverage threshold, the combination of MaxQualSum + MuTect yields the lowest FPR. At higher coverage values, any UMI consensus method + MAGERI worked better. With respect to the FDR, the use of any UMI consensus method lead to a reduction of the FDR compared

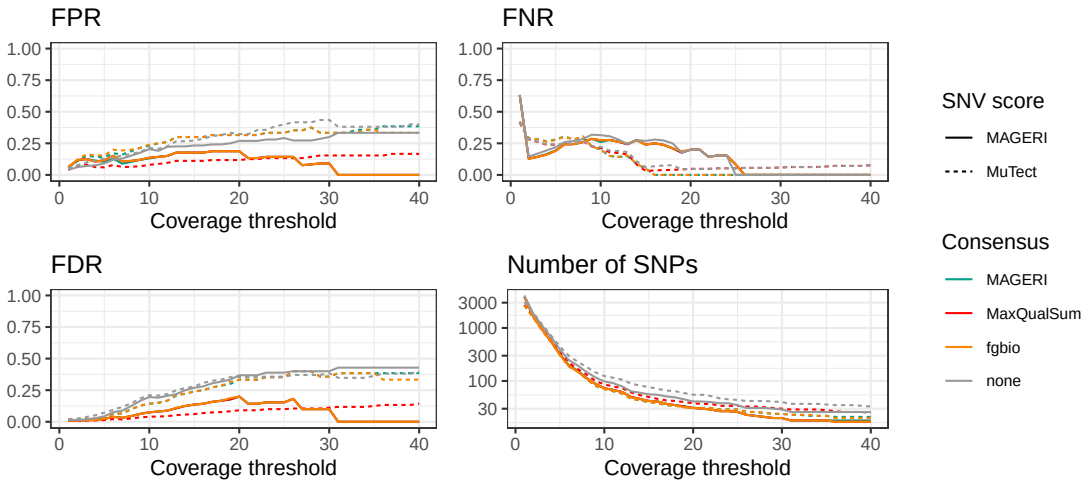


Figure 3.3: Classification rates of the UMI consensus and SNV score method benchmark with the AML-LT data. The x-axis represents the read coverage threshold. The shown rates are: false positive rate (FPR), false negative rate (FNR), and false discovery rate (FDR). The plot on the lower right corner shows the number of SNPs that passed the coverage threshold and were not filtered out during consensus and score estimation.

to the control without UMI consensus. It is particularly interesting to compare the maximum FDR value (0.45) of the non-consensus group + MAGERI at coverage 40 with the corresponding analyses where a UMI consensus was generated, the latter being nearly equal to zero. Similarly to the case of the FPR, the lowest values for the FDR were obtained with MaxQualSum + MuTect up to a coverage of 26 reads, and any UMI consensus method + MAGERI lead to the lowest FDR at higher coverage values.

Apart from the impact on the reduction of false positive variant calls, it was important to determine if the application of UMI consensus methods lead to a detectable increase in the amount of true SNPs that were excluded from the variant calls. However, this did not appear to be the case: for any given variant calling method, the FNR values produced when employing the UMI consensus methods were nearly equal to the control without a consensus (figure 3.3, upper right side). The lowest FNR results for up to a coverage of 8 reads were obtained by any UMI consensus method + MAGERI. At higher thresholds, a lower FNR was obtained with the fgbio or majority consensus methods + MuTect.

3.2.3 Evaluating the impact on clonal assignment

The previous benchmark was useful to establish the advantage of using UMI consensus methods to reduce false positive variant calls, without an impact on the detection of true variants. The following step was to evaluate whether such variant proofreading can contribute to a more precise assignment of single cells to their subclones. In order to approach this, I used the FPR and FNR obtained with each combination of UMI consensus and SNV calling methods to simulate variant counts from cells that originated of a specific clonal phylogeny configuration. With such variant counts, cells would be assigned to one of the possible subclones, and the proportion of cells that were assigned to their true subclones could be evaluated.

For this analysis, clonal phylogenies of 3, 5, or 7 subclones were designed with equal or mixed branch lengths (figure 3.4). In the case of trees with unequal branch lengths, the branch length value of the new subclonal generations was increased either once in a lineage, or twice for one of the 7-subclone trees ('7 - Double'). Two trees were designed to reflect two scenarios of more complex branch length inequalities: one where the number of variants was the highest for the founder clone, and decreased on each new generation of subclones ('Large trunk, small tips'); and another with the opposite scenario, in which few variants were at the trunk, and increased for the last generation of daughter cells ('Small trunk, large tips'). A total of 105 SNVs in-silico generated SNVs were distributed along the branches of each of the trees.

To evaluate the impact of variant calling benchmark rates on the single-cell assignments to these clonal phylogenies, I implemented a method based on the single-cell allele coverage matrix simulation function from OncoNEM (Ross and Markowetz [2016]; section 7.3.2). In brief, coverage per allele of each variant site per cell was simulated, depending on the true genotype of the cell. These allelic counts were modified to reflect variant calling errors based on the FPR and FNR that were obtained in the UMI consensus benchmark for each combination of methods at fixed coverage thresholds of 5, 20, or 40 reads. Cells were then assigned back to the subclones using Cardelino. The number of correct assignments, incorrect assignments, and unassigned cells was estimated.

3.2.4 Error rates derived from the UMI consensus method benchmarks improved clonal assignments

Allelic counts simulated with the FPR and FNR values that resulted from the application of UMI consensus methods led to higher proportions of correctly assigned cells across all different clonal phylogenies, but especially those with more subclones or more variation in the branch lengths (figure 3.5). In the simplest clonal phylogeny (3 clones, equal branch lengths), all consensus and score methods performed comparably well across all coverage values, and even the control without consensus yielded high correct assignment rates 1 in most cases (except for a coverage threshold of 40 + MuTect). Increasing the complexity of the clonal phylogeny (more clones, unequal branch lengths) made the assignment more sensitive to the FPR and FNR. Interestingly, increasing the number of clones seemed to have a stronger influence on the correct assignment rate than branch length inequalities: median correct assignment rates were lower on the 5- or 7-clone configurations (0.98 and 0.95, respectively), but were comparable between the versions of a given number of clones with equal or unequal branch lengths (0.97 and 0.93). The highest drop in the assignment rate was observed in the most complex clonal configurations: 7 clones with large trunk and small tips (median = 0.83), or small trunk and large tips (median = 0.85).

Different consensus and score method combinations yielded the optimal correct assignment rate per coverage threshold, largely reflecting the FPR and FNR values that were obtained on the SNP calling benchmark. For the coverage thresholds of 5 and 20 reads, the highest median correct assignment rate across phylogenies was obtained with the MaxQualSum + MuTect combination. In the 40-bp threshold, any consensus method + MAGERI score yielded comparably good correct assignment rates (median 1). It is also worth pointing out that the correct assignment rates were generally comparable for within a phylogeny and threshold in combination with any UMI

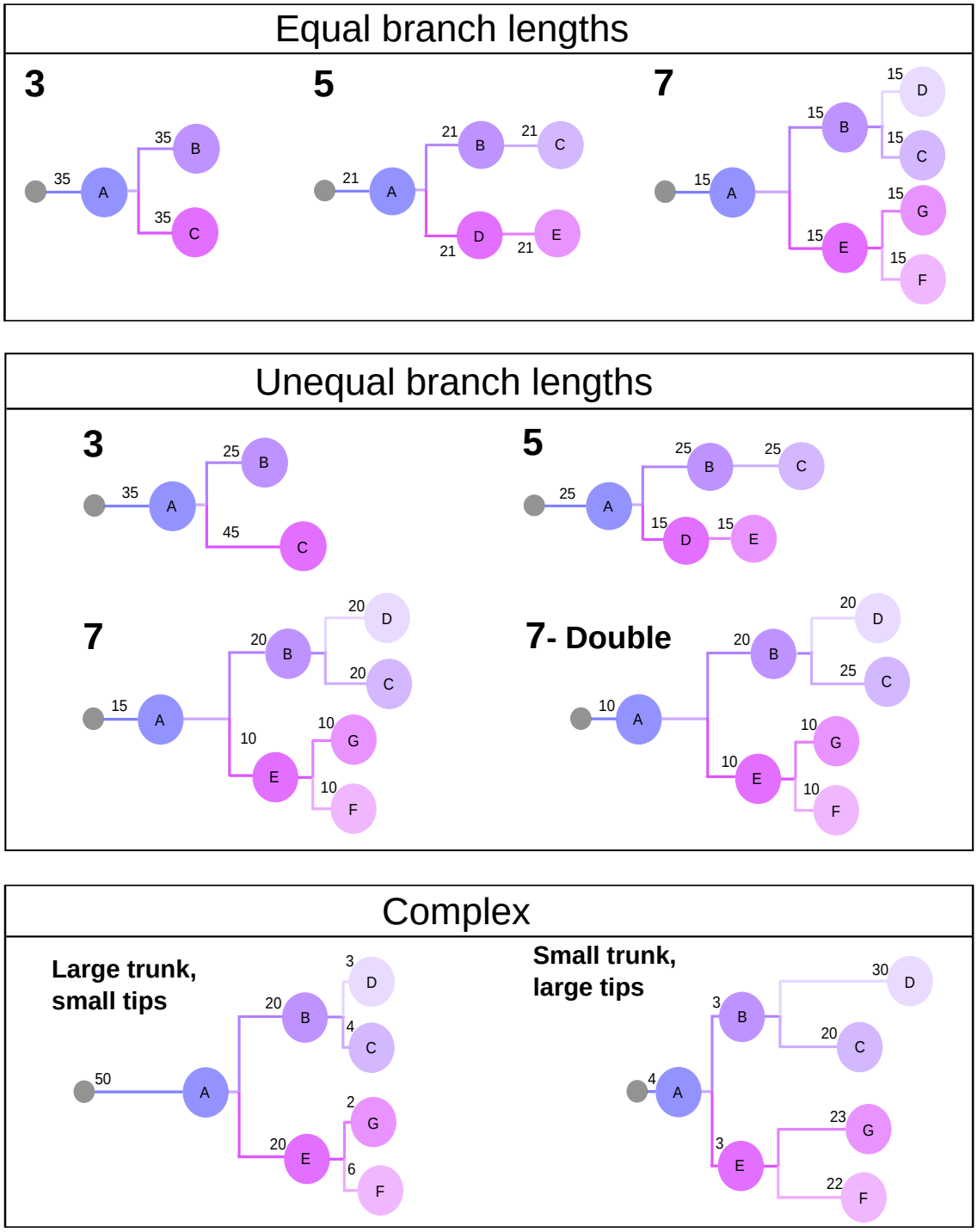


Figure 3.4: Simulated clonal phylogeny configurations for the evaluation of clonal assignment. Branches are labeled with the corresponding number of variants.

consensus method, and in the 5- and 40-read thresholds was able to improve the assignment rate without any consensus method.

3.3 Performance Efficiency

3.3.1 Runtime Measurement Setup

The *umivariants* approach is capable of producing accurate variant calls and clonal assignments in scRNA-seq, as demonstrated by the results of previous benchmarks. However, one practical consideration that has to be contemplated is the amount of computational resources and running time that have to be employed for each of the steps, especially for larger datasets with a large number of reads or variants to analyze. In order to provide some expectation for these requirements, I benchmarked the runtime efficiency and resource usage of the most computationally intensive functions of umivariants, namely the UMI pileup, the UMI consensus, and SNV score calculation. I also address what are the performance benefits of the multithreaded processing of such steps.

I employed subsampled datasets derived from the AML-LT scRNA-seq data. The original BAM file was subset with different numbers of reads ranging from 5×10^4 to 1×10^8. Six replicate BAM files were produced for each number of reads. On each BAM file, increasing numbers of variants were called (5 to 200), which correspond to the n variants with the highest coverage in the original BAM file. Finally, I ran the pileup, consensus, and SNV score steps on each combination of total reads and number of variants 3 times with the microbenchmark R package, which estimated runtime statistics on each case. The mean runtime values per triplicate test were subsequently averaged among the 6 replicates. Each umivariants step was run using 1 or 5 threads. These tests were run on a Scientific Linux 7.3 server with 88 threads and 500 Gb of RAM.

3.3.2 Runtime depends on the number of SNVs in the pileup step, and on the number of reads in the UMI consensus and SNV calling steps

In the sequence pileup step, increasing the number of reads or variants, or using a single core, generally result in longer runtimes. Employing 5 threads instead of one reduced the runtime by factors ranging from 1.74 to 4.83; this was nearly always more beneficial for 100 or 200 variant (figure 3.6). Interestingly, the actual increase in runtime does not scale consistently in its entirety when increasing the number of reads, and in some cases a drop can actually be observed for a higher library size. This could be a consequence of lacking direct coverage in some of the chosen variants, as the reads were randomly subset.

In a benchmark using 5 threads, the runtime of the UMI consensus methods scaled exponentially in the lol-log scale with the number of reads (figure 3.7). The fgbio and majority vote methods were the most computationally intensive. The runtime values increased moderately with higher numbers of SNVs, but such an increase was not scaled.

The runtime of both SNV calling methods, MuTect and MAGERI, scale exponentially in the log-log scale with the number of reads. The runtime with MuTect increases considerably after 1 million reads. In contrast, the calculation of the MAGERI score is more efficient, finishing in less

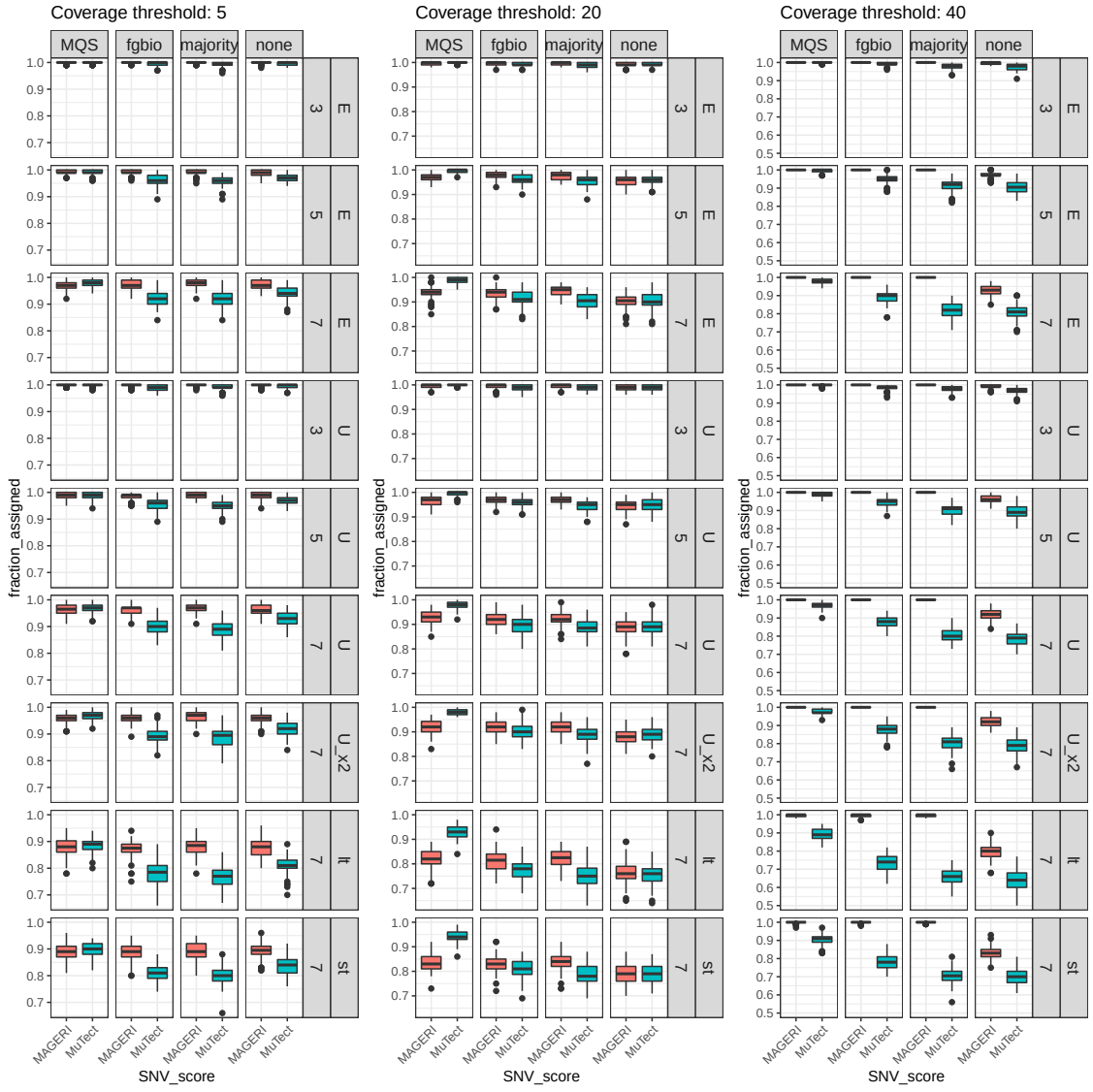


Figure 3.5: Fraction of cells assigned to their true clone in the simulation of clonal phylogenies and variant allele coverage according to the FPR and FNR from the UMI consensus benchmark. The plot is divided in three columns representing discrete coverage cut-offs at which the FPR and FNR were taken. On each of these columns, clonal phylogeny configurations are faceted horizontally, the UMI consensus methods of the corresponding error rates are faceted vertically, and the x axis contains the variant calling method. E = equal branch lengths; U = unequal branch lengths; U_x2 = unequal, double; lt = large trunk and small tips; st = small trunk and large tips.

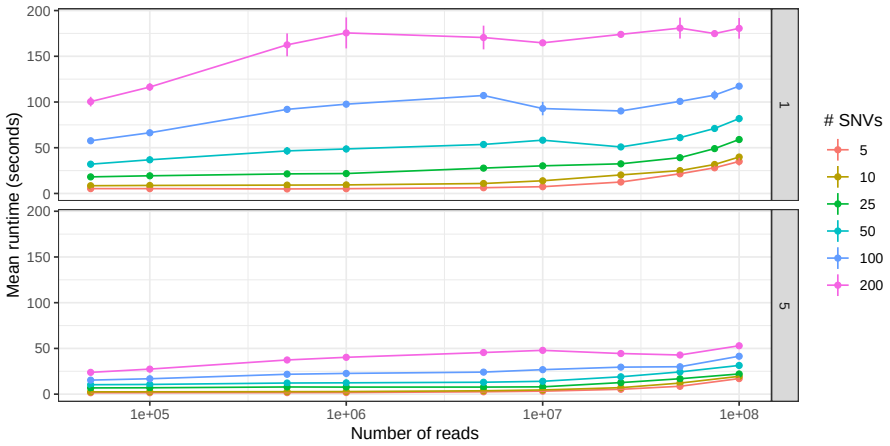


Figure 3.6: Results of the UMI pileup performance tests. The log_{10}-scaled number of subsampled reads is shown on the x-axis. The plotted values correspond to the mean of 6 replicates. The error bars represent the standard deviation. The line and point color represents the number of called variants, and the shape indicates the number of threads used for running the pileup.

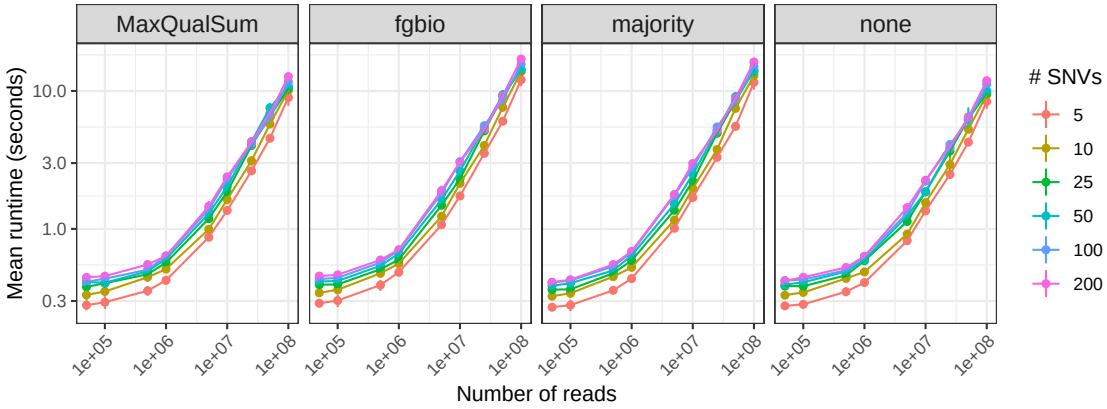


Figure 3.7: Results of the UMI consensus performance tests. The plot is $log_{10} - log_{10}$-scaled. The plotted values correspond to the mean of 6 replicates. The error bars represent the standard deviation. The line and point color represents the number of called variants. 5 threads were used. The plots are faceted by consensus method; a control without consensus was also computed.

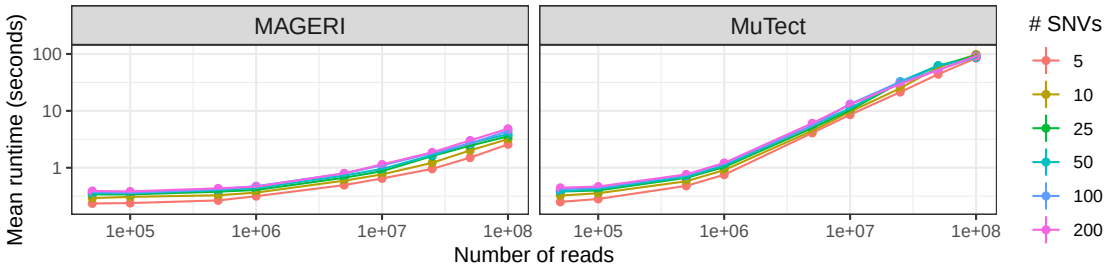


Figure 3.8: Results of the SNV calling performance tests. The log_{10}-scaled number of subsampled reads is shown on the x-axis. The plotted values correspond to the mean of 6 replicates. The error bars represent the standard deviation. The line and point color represents the number of called variants, and the shape indicates the number of threads used for running the pileup. 5 threads and the MaxQualSum consensus method were used. The plots are faceted by SNV calling method.

than a second for up to 10 million reads, and in 3 seconds for 100 million (figure 3.8). Similarly to the UMI consensus step, runtimes are similar across different numbers of SNVs, so these were not a critical factor in increasing the runtime.

3.4 Comparison of the Original MAGERI Method to the *umivariants* Implementation Using Ultra-Sensitive Genotyping Data

As mentioned in section 3.1.4, the original MAGERI software assembles the reads that are associated with a UMI prior to mapping using its own local aligner. Therefore, variant calling results produced with the original MAGERI pipeline could potentially present significant differences with the ones from its implementation in *umivariants*, even if the same UMI consensus and SNV scoring methods are applied. In order to confirm whether such differences would be prevalent, I compared the performance of both methods using the SiMSen-seq dataset of the AML-LT experiment (section 2.3.4), which incorporates UMIs for ultra-sensitive variant genotyping.

For the *umivariants* analysis, SiMSen-seq reads were mapped using BWA-MEM, and the output BAM files were UMI-tagged using fgbio. The UMI-consensus and SNV calling steps were performed using the methods and parameters that had the closest resemblance to the orignal MAGERI software defaults: a majority vote UMI consensus with a minimum of 5 reads, and a MAGERI SNV score threshold of 20. Even though this corresponds to a bulk targeted amplicon sequencing dataset, the input files and variant sequences can be processed just like the scRNA-seq data. Variant calling with the original MAGERI software (version 1.1.1) was performed by Daniel Richter. 7 samples with two replicates were analyzed.

I compared the UMI coverage, VAF, and SNV call status of the 5 targeted variants between both MAGERI implementations on each replicate of the samples. The UMI coverage ratio was estimated by dividing $uCov_{MAGERI}/uCov_{umivariants}$, where $uCov$ is the UMI coverage calculated with one of the methods. I also estimated the VAF difference by subtracting $VAF_{MAGERI} - VAF_{umivariants}$.

The VAF and UMI coverage values were largely similar across variants between both MAGERI

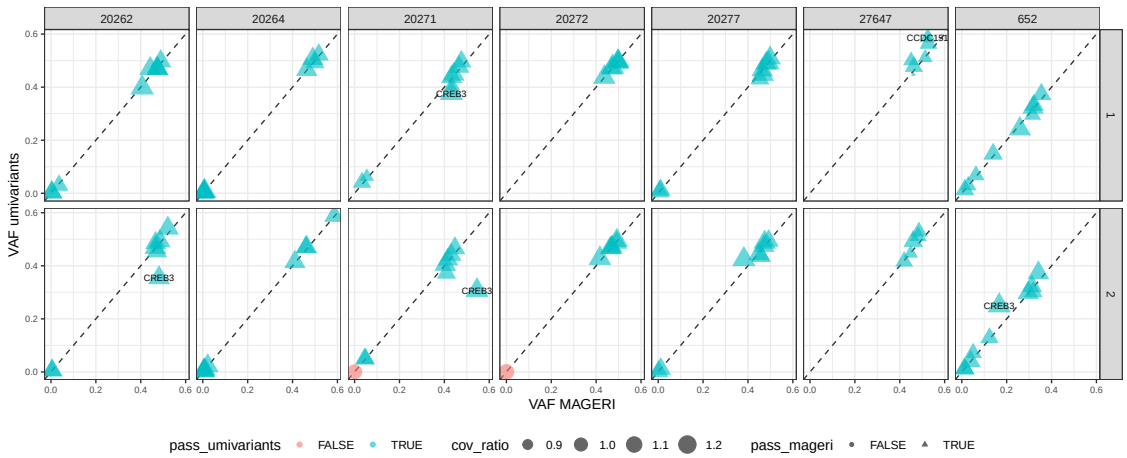


Figure 3.9: Comparison of VAF, UMI coverage, and variant acceptance (i.e. score threshold pass) when comparing variant processing and calling with MAGERI and its implementation in umivariants. The plot is faceted by sample (vertically) and replicate (horizontally).

implementations, with VAF differences below 0.01 and UMI coverage ratios within a range of 0.8 to 1.21 in a majority of cases (3.9). One exception was the CREB3 variant, which actually presented a VAF difference > 0.5 in four instances. Furthermore, it had the largest VAF difference overall with a value of 0.238 (mouse 20271, replicate 2). The KRAS and CDH2 variants also presented some instances of VAF differences higher than 0.25.

It is noteworthy that variant calling (i.e. evaluation of variants that passed the score threshold) agreed between both implementations; i.e. they both classified the same variants as being true or false on each sample and replicate. Furthermore, UMI coverage ratios stayed close to 1 in most cases, even with a total UMI coverage value in the tens of thousands. One conclusion from this is that it is possible to rely on the *umivariants* implementation of MAGERI for variant calling in most cases, even if the assembly step is not executed. Further adjustments could be performed on the upstream processing of the reads (such as mapping, trimming, etc.) or during quality filtering to minimize any differences in the UMI coverage and VAF values with respect to the results from the original MAGERI algorithm.

Chapter 4

Analysis of scRNA-seq Data with *umivariants*

4.1 Calling Variants in *Genotyping of Transcriptomes* Data

4.1.1 Description of the dataset

To demonstrate the potential of *umivariants* for accurate variant calling in scRNA-seq, I performed a re-analysis of the data from the publication by Nam et al. [2019]. In their work, they analyzed samples from patients with calreticulin (CALR)-mutated myeloproliferative neoplasms: 6 of the sequenced patients presented essential thrombocytonemia (ET), and 5 had myelofibrosis (MF). scRNA-seq data from these samples was generated with the 10x protocol. In the publication, the authors generated a method for targeted amplification of loci of interest during the library preparation step: Genotyping of Transcriptomes (GoT). They employed GoT to increase the coverage of variants in CALR and other genes that were only sparsely covered in the 10x data.

GoT data were processed with their own computational pipeline, IronThrone-GoT. In this method, variants are extracted from the reads based on fixed expected positions in the amplicons. The reads preserve their UMIs and sample barcodes, and a majority-vote approach is applied for the UMI consensus, with any ties being decided by the read with the highest Phred score. Reference and variant alleles are interpreted based on the user-provided configuration file, and the UMIs that present each allele on each cell are afterwards counted.

Both the 10x and GoT data from this publication have been made publicly available (see section 7.4). Furthermore, variant coverage counts obtained with IronThrone-GoT were also provided. Therefore, I was interested in using all these available data to determine if *umivariants* could effectively analyze the same variants in both kinds of datasets, and produce variant calls that were comparable to the ones from IronThrone-GoT.

Sample	Cells	Mutated genes / Variant ID
ET01	6811	CALR 1, JAK2?
ET02	1135	CALR 2, *SH2B3*
ET03	3587	CALR 1, XBP1, *NOTCH1*
MF01	965	CALR 2, SF3B1
MF05	8475	CALR 2, NFE2, SF3B1

Table 4.1: Genes with somatic variants in the 10 GoT samples. ET = essential thrombocytonemia; MF = myelofibrosis. Variants/genes in italics were not amplified by GoT. Taken from the extended figure 3 and supplementary table 1 of Nam et al. [2019].

ID	hg38 Pos	Variant	Transcript	cDNA Pos	AA
CALR 1	19:12943751	52-bp del	ENST00000316448.9	1099	L367Tfs*46
CALR 2	19:12943813	A>ATTGTC	ENST00000316448.9	1154	K385Nfs*47
JAK2	9:5073770	G>T	ENST00000381652.3	1849	V617F
NFE2	12:54292710	GCTCT>G	ENST00000540264.2	782	Q261fs
SF3B1	2:197402649	G>C	ENST00000335508.10	1984	H662D

Table 4.2: GoT variants with start positions in the hg38 genome and in the cDNA of selected transcripts, as well as amino acid (AA) change. Variants are identified by the gene name and, in the case of CALR, an additional number. Taken from the extended figure 3 and supplementary table 1 of Nam et al. [2019].

4.1.2 *umivariants* is able to recover the reported variants in the 10x and GoT-amplicon datasets

I started my re-analysis of the Nam et al. [2019] data with *umivariants* by calling five variants of interest (table 4.2) that the authors reported in the 10x datasets of five samples were made readily available in BAM format (table 4.1). The variants included two types of indels in CALR: one 52-bp deletion and one one 5-bp insertion. Using *umivariants*, it was possible to call the reported variants on the 10x samples, including the CALR deletion and the insertion (figure 4.1). Variants were only detected in a low number of cells per sample, which was in agreement with the original publication.

To demonstrate the performance of *umivariants* in the analysis of GoT data, I processed the reads from the ET02 sample and called the CALR insertion (CALR 2) that was reported as present in the sample. I applied the MaxQualSum, majority vote, and no UMI consensus methods to determine whether these led to differences in the reported coverage or genotype. 942 cells showed a coverage of at least one UMI with 2 reads in the CALR 2 site. The total UMI coverage values per cell were generally similar between the results from both *umivariants* and IronThrone-GoT, with Spearman correlation values > 0.95 between the two approaches (figure 4.2).

430 cells had at least one UMI with the CALR 2 insertion. As a proxy to the coverage of the variant allele per cell, and to determine the number of cells in which both alleles could be detected, I computed and compared the VAF values per cell reported by *umivariants* and IronThrone-GoT. The observed CALR 2 VAF per cell in mutant cells ranged from 0.053 to 1. The VAF estimate based on the counts from the MaxQualSum consensus showed the highest correlation ($\rho = 0.931$) with the VAFs based on the allelic counts from IronThrone-GoT, while the one from a control without consensus showed the lowest correlation ($\rho = 0.873$; figure 4.3).

The allelic coverage estimates of the CALR 2 variant from each pipeline were used to assign

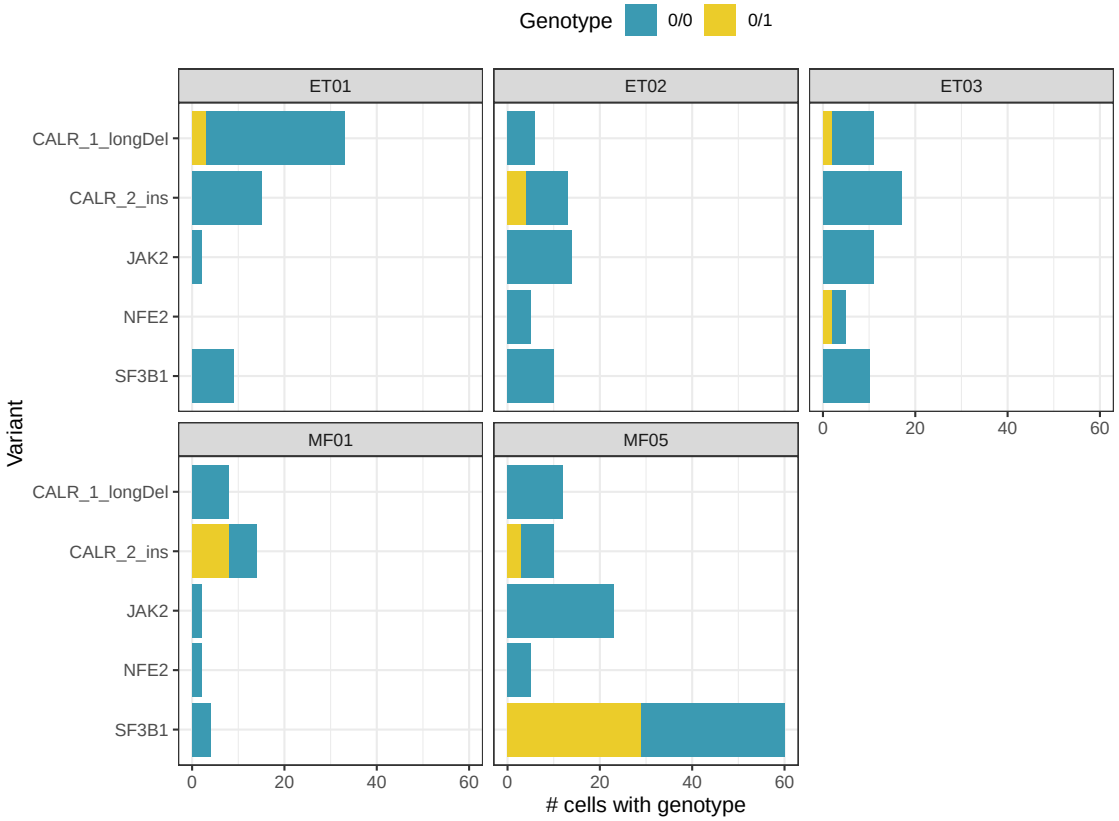


Figure 4.1: VAFs of the variants of interest in five 10x datasets of the GoT publication. Variants were called using *umivariants*. In the ET01 dataset, UMIs were not filtered by number of reads in order to show detection of the CALR deletion.

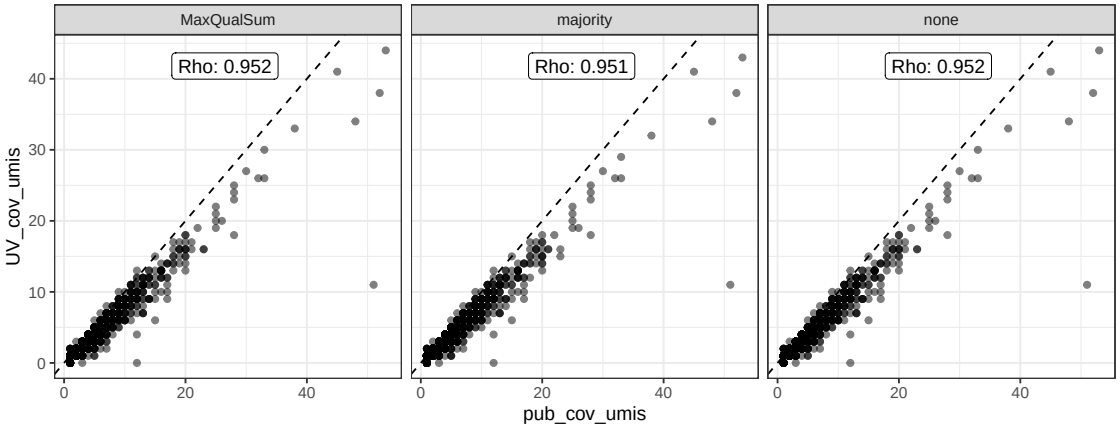


Figure 4.2: Comparison of CALR insertion UMI coverage values estimated by *umivariants* or IronThrone-GoT in the GoT ET02 sample. The plots are faceted by the consensus method used in *umivariants*. Spearman's ρ values between both analyses per consensus method are shown.

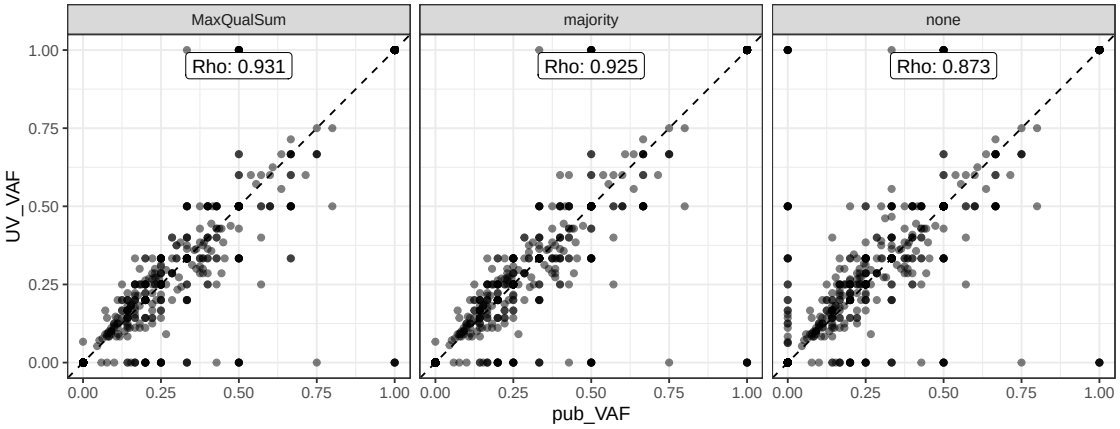


Figure 4.3: Comparison of CALR insertion VAF values estimated by *umivariants* or IronThrone-GoT in the GoT ET02 sample. The plots are faceted by the consensus method used in *umivariants*. Spearman's ρ values between both analyses per consensus method are shown.

the genotype of each cell. When comparing the genotype assignments of the *umivariants* results with the MaxQualSum consensus to the IronThrone-GoT values, 93.47% of all cells showed the exact same genotype in both approaches, with an agreement of 96.2% of cells classified as variant, i.e. genotypes of either 0/1 or 1/1 (figure 4.4). 2.44% of the cells had coverage dropouts in the *umivariants* analysis and could not be genotyped.

4.1.3 Consistent cluster labeling between *umivariants* and the original publication

In the original publication, the authors used the GoT data to analyze differences in gene expression between mutant and wild-type cells, particularly in the context of clusters of different hematopoietic progenitors. I analyzed if the genotyping results produced by *umivariants* led to any appreciable differences in the gene expression patterns of mutant cells across cell clusters. To that end, I used the ET02 sample with the CALR 2 insertion on which I compared the *umivariants* and IronThrone-GoT genotypes in the previous section.

I processed the publicly available ET02 scRNA-seq gene expression count matrices with Seurat (Stuart et al. [2019]), and performed dimensionality reduction using UMAP (McInnes et al. [2018]) with Louvain clustering. Each single cell was labeled with the genotypes that were estimated from the *umivariants* (MaxQualSum) and IronThrone-GoT analyses. From an initial visual inspection, cells with the genotypes of both approaches appeared to be equally distributed in the different clusters, as expected from the high agreement rate between the genotypes of both estimations (figure 4.5). Accordingly, the fraction of mutant cells (i.e. with genotype 0/1 or 1/1) on each cluster cluster was similar between the annotations of *umivariants* and IronThrone-GoT (figure 4.6). This fraction was generally lower in the *umivariants* annotation, but this was moderate and consistent in all clusters, indicating that the coverage and genotype dropouts were not biased by a specific cell group.

I performed a differential expression analysis of the CALR 2 mutant and WT cells, based on the genotypes of either *umivariants* or IronThrone-GoT. Using Seurat, I counted how many

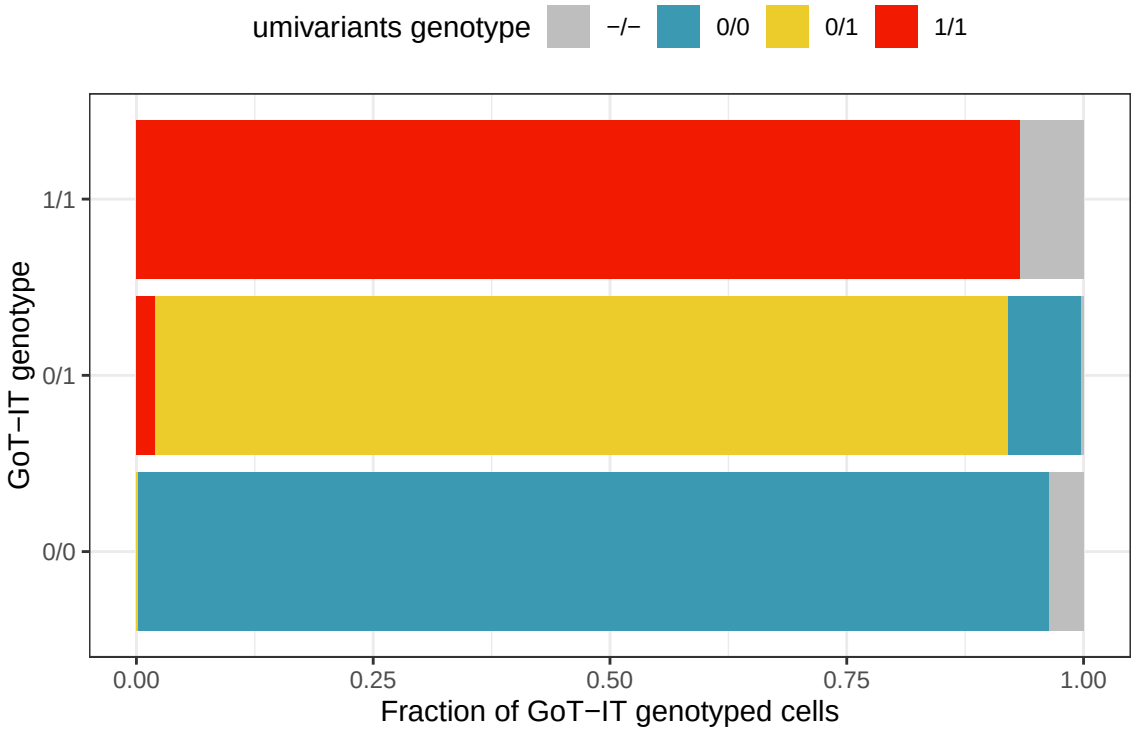
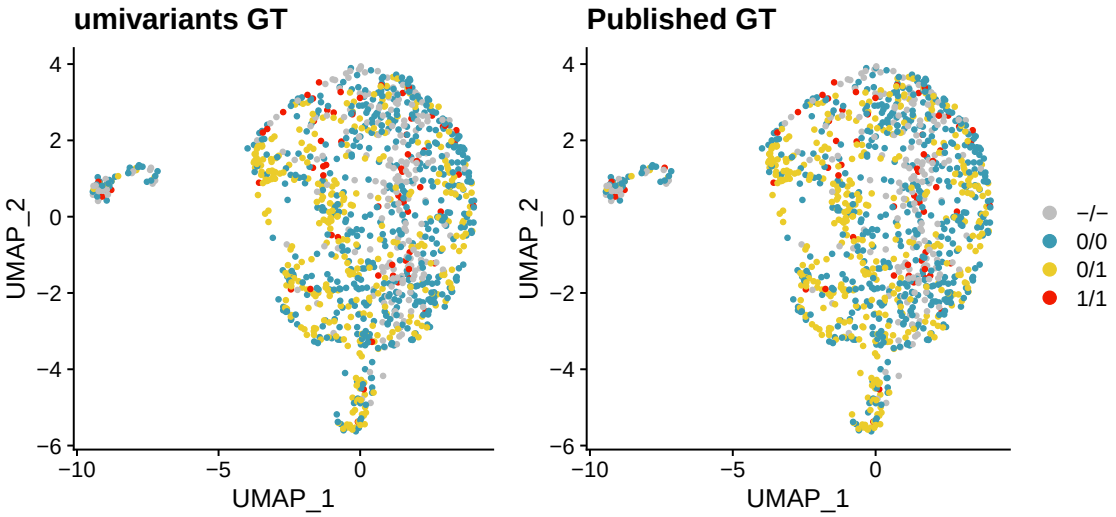
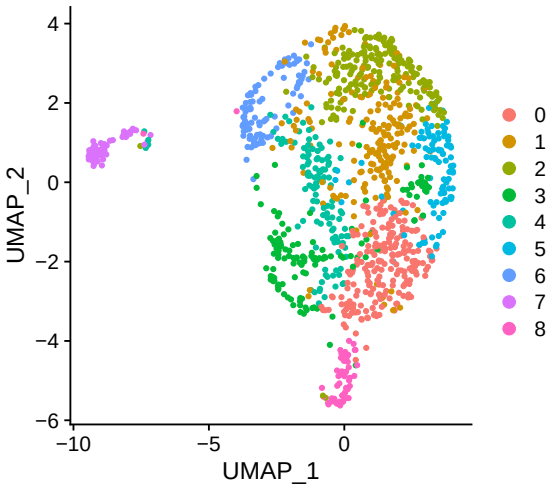


Figure 4.4: Comparison of CALR insertion genotype per cell estimated by *umivariants* with the MaxQualSum consensus or IronThrone-GoT in the GoT ET02 sample. The fraction of cells with a specific genotype in the IronThrone-GoT analysis that were assigned to each possible genotype in the *umivariants* analysis are shown with the color of the bars.



(a) UMAP projection with CALR variant 2 genotypes per cell estimated with *umivariants* (left) or with IronThrone-GoT (right).



(b) UMAP projection with estimated clusters.

Figure 4.5: UMAP projection and genotyping of the ET02 sample.

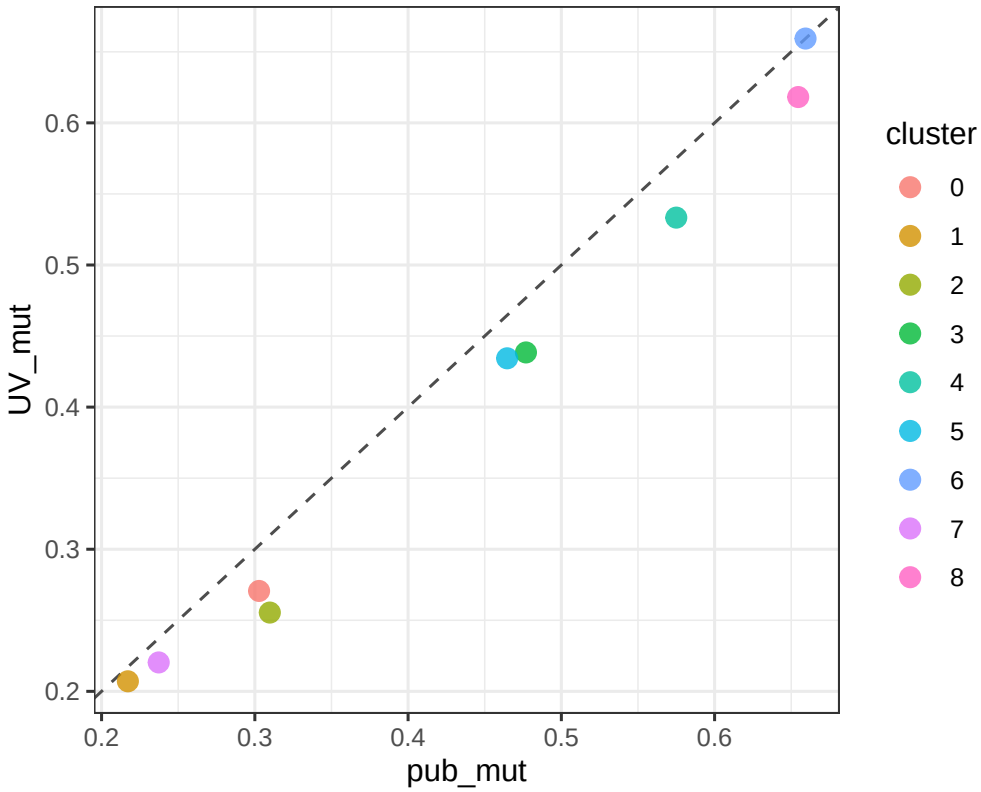


Figure 4.6: Fractions of CALR mutant cells per cluster in the ET02 sample determined from the IronThrone-GoT or *umivariants*.

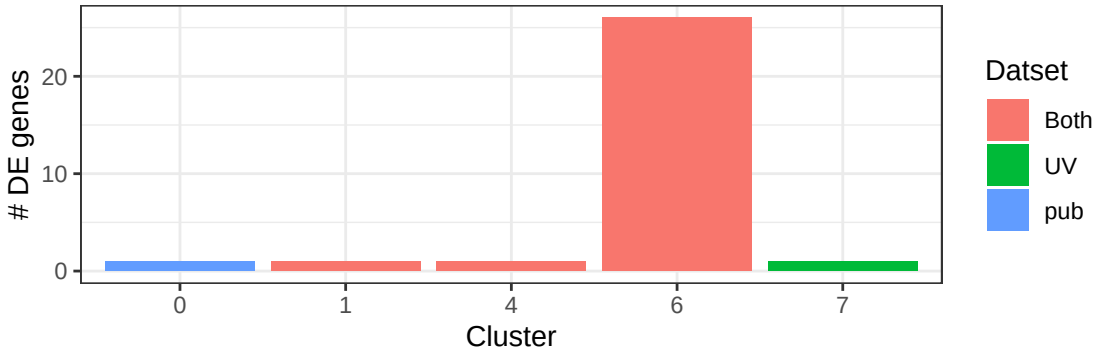


Figure 4.7: Number of DE genes in the ET02 sample that were found per cluster in CALR 2 mutant versus WT cells, depending on the genotype by *umivariants* or IronThrone-GoT. Counts are colored depending on whether a gene was classified as DE in the cells of that cluster when using the annotation of only one or both genotyping approaches.

differentially expressed (DE) genes were found per cluster. Then I counted how many of those DE genes were retrieved when using only one of the genotype annotations, or both. Only two genes (one in cluster 0 and one in cluster 7) were not found in the mutant classification with both methods, while any other DE gene instances were shared (figure 4.7). Therefore, genotype assignments from both *umivariants* and IronThrone-GoT can be used to profile gene expression values in mutant cells with comparable results.

4.2 Assigning Clones in the AML-LT scRNA-seq Data Using *umivariants*

After validating *umivariants* for the reliable genotyping of mutant cells in the GoT dataset, it was now possible to use this approach for the study of mutant cells and subclonal heterogeneity in AML. To that end, the scRNA-seq data that was generated for the AML-LT experiment was available, and I already demonstrated its utility for the benchmarking of UMI consensus methods (section 3.2). To increase the potential coverage of the variant sites of interest, the scRNA-seq dataset was sequenced under 3 different configurations (section 7.1.5). The reads from all three approaches were combined, and SNVs from the subclones defined in the exome dataset (section 2.3.3).

With the combined coverage of all three sequencing configurations, I attempted to assign cells to one of the KRAS or NRAS subclones. However, one hurdle became evident at this stage: only 66 cells had any coverage of a variant allele. Furthermore, only 10 of the 75 clonal variants had coverage for the variant allele in the single cells; another 19 had coverage of the reference allele only. KRAS G12A was actually the best covered variant, found in 27 cells (4.8).

One potential reason behind the coverage of few of the variant sites could be the actual expression level of the genes that contain them. After normalization with scran (Lun et al. [2016]; performed by Beate Vieth), only 18 of the genes with associated SNVs were found to be effectively expressed in the dataset, generally with a median below 5 normalized counts (figure 4.9). Some of the most highly expressed genes in this set had SNVs from the founder clone, such as ETV6 or

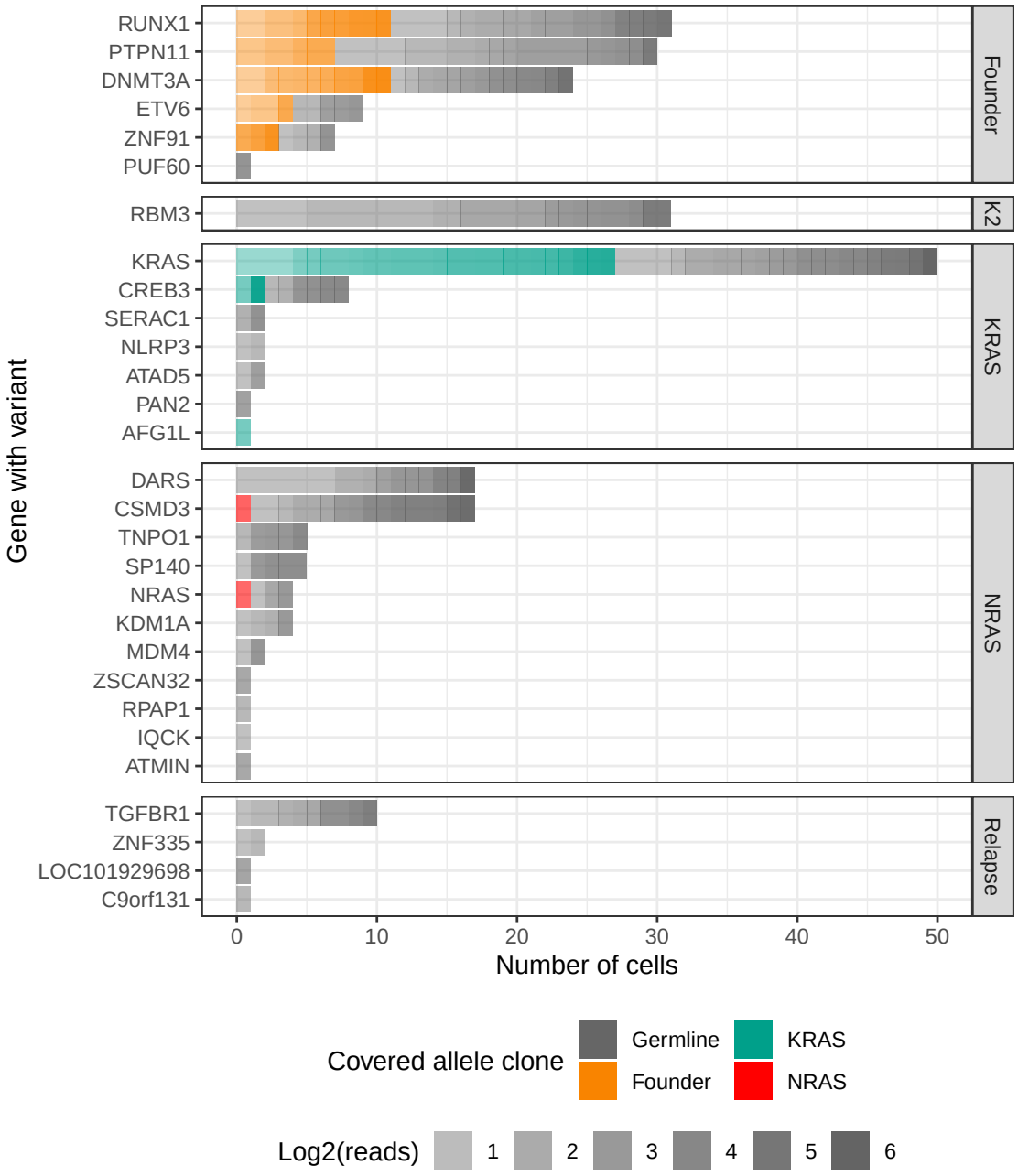


Figure 4.8: Allelic coverage at clonal SNV sites per cell in the AML-LT scRNA-seq dataset. The bars show the number of cells with a specific log_2 read coverage per gene on either reference or alternative allele (labeled 'Germline' or by the clone, respectively). log_2 read coverage is indicated by transparency. Genes are faceted by their corresponding subclone inferred with Canopy in the exome data. K2 = KRAS-2.

RUNX1.

Subclonal variants could be called in 66 cells. 30 of these were assigned to the KRAS clone, 2 to NRAS, and 34 to the founder. For the assignment, I used Cardelino through its *umivariants* wrapper, with a posterior probability cutoff of 0.35. Their exact distribution across PDX samples and therapy stages can be seen in figure 4.10. Given the low number of cells assigned to a clone per sample, these do not reflect the clonal frequencies inferred from bulk sequencing. On the other hand, all KRAS and NRAS cells had only one detectable variant from either of these subclones (figure 4.11).

When analyzing the transcriptional and clustering profiles of the cells that could be assigned to a subclone or the founder clone, it was observed that the KRAS and founder cells were homogeneously spread among different clusters (4.12; Philipp Janssen, personal communication). The two NRAS cells were placed into clusters associated with hematopoietic stem cells and common myeloid progenitors, but with such a low number of cells it cannot be determined if NRAS cells in general would be confined to such clusters.

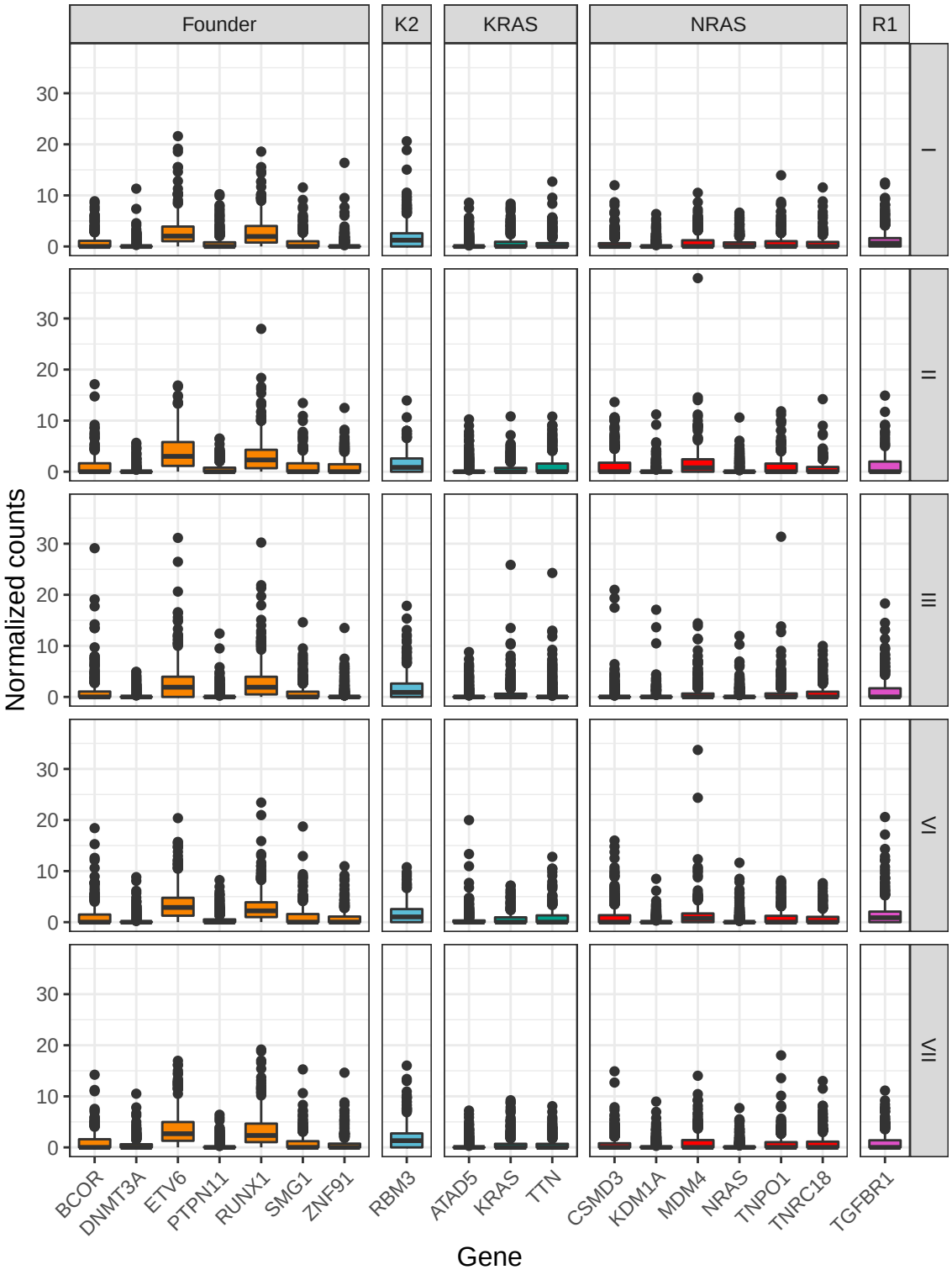


Figure 4.9: Normalized counts of genes with clonal SNVs that were expressed above background levels in the AML-LT scRNA-seq data

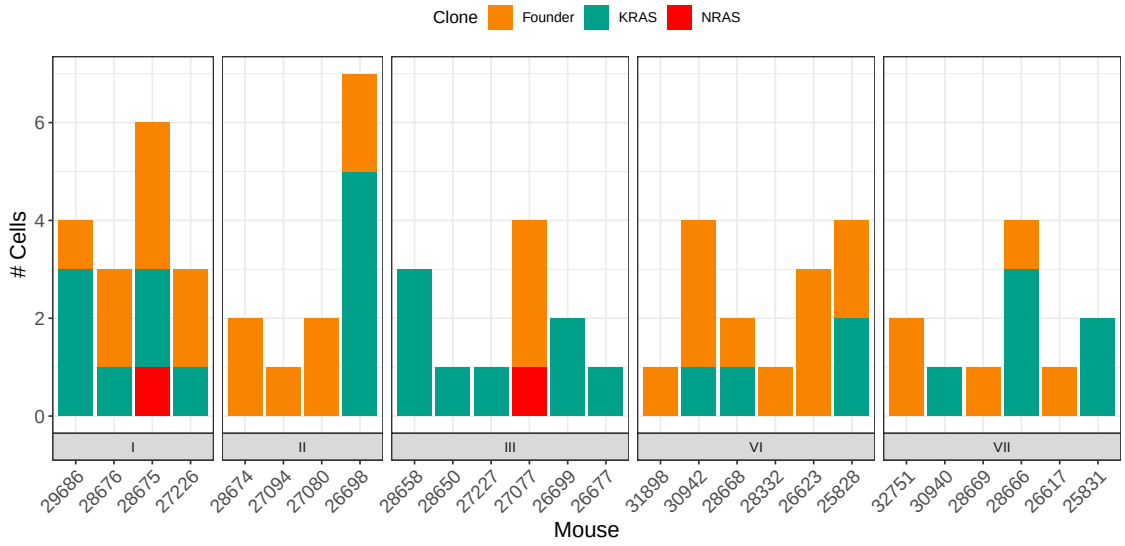


Figure 4.10: Number of cells assigned to each clone per sample and treatment stage in the AML-LT mcSCRB-seq dataset.

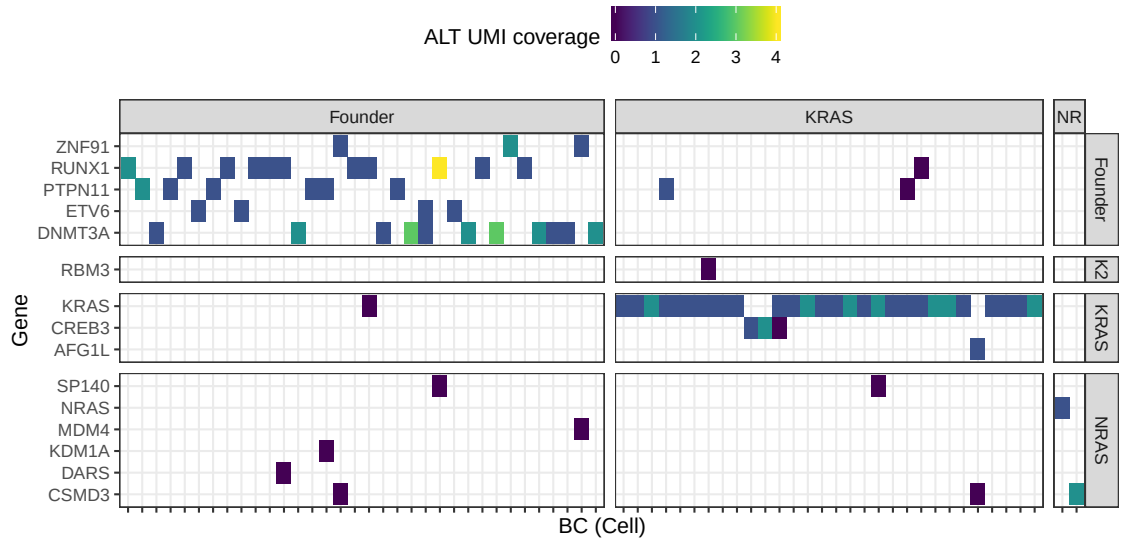


Figure 4.11: Total UMI coverage of the variant allele (ALT) in cells that were assigned to a clone or subclone. The plot is faceted by SNV clone (rows) and clone assigned to the cell (columns). NR = NRAS. K2 = KRAS-2.

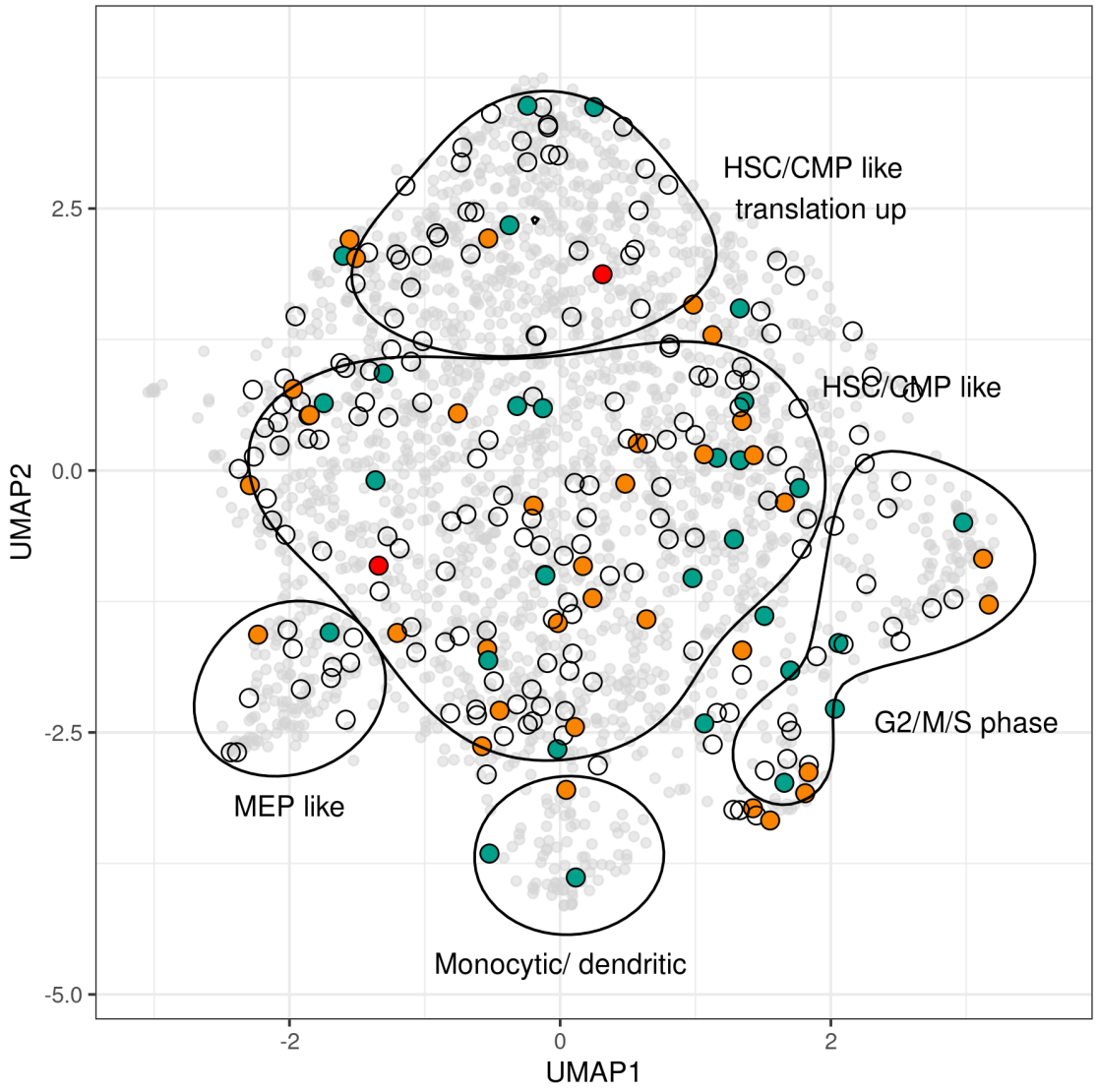


Figure 4.12: UMAP projection of the cells of the AML-LT scRNA-seq dataset, colored by clonal assignment. Courtesy of Philipp Janssen. HSC = hematopoietic stem cell, CMP = common myeloid progenitor, MEP = megakaryocyte-erythroid progenitor.

Chapter 5

Discussion

5.1 The Challenges of Analyzing Clonal Heterogeneity and Evolution in AML

Inferring clonal heterogeneity and evolution in AML, just like in the cancer field as a whole, represents a highly complex challenge that has required the development of multiple experimental and computational methods. In order to have sufficient power to call sufficient sequence variants that inform of such heterogeneity, it is critical to analyze sufficient samples with the highest possible breadth and depth of sequence coverage. It is also ideal to analyze a particular cancer dataset with multiple omics methods that can provide different strengths and support for the multiple molecular lesions; however, the different requirements for quality control, analysis, and integration remain non-trivial.

With the advent of scRNA-seq methods in particular, it has become possible to study clonal heterogeneity in high detail in AML, and to associate specific subclonal populations with molecular phenotypes and cell types (van Galen et al. [2019], Petti et al. [2019]). However, these studies have also revealed that the coverage obtained from the single-cell transcriptome alone is too sparse to allow the analysis of the complete profile of subclonal variants. This becomes increasingly difficult when the genes with variants of interest have low or no expression in the cells. Therefore, various techniques have been designed to amplify the signal of expected somatic variants (Nam et al. [2019], van Galen et al. [2019], Velten et al. [2018]). The use of deep amplification and sequencing, while enhancing the signal of the variants of interest, can also increase the probability of inducing false positive mutations due to technical errors. It is therefore important to make use of tools like UMIs to enable proofreading and filtering such errors, as has been done in methods like GoT (Nam et al. [2019]), or the currently developed scTAG-seq: a method that will enable amplification on readily available cDNA libraries with UMIs (Johannes Bagnoli, Lucas Wange, et al., personal communication).

The approach and software package proposed in this dissertation, *umivariants*, was designed to enable proofreading and analyzing sequence variants based on UMIs, with the end goal of using such information to study clonal heterogeneity in cancer samples. Other computational methods have been developed over the past years to use UMI information to call and/or proofread variants

Method	Data type	Functions
umivariants	Multi-omics with UMIs	Variant pileup, UMI consensus, SNV scoring, genotyping, clone mapping, chimeric UMI pre-testing
cellSNP + Vireo	scRNA-seq (UMI or not)	Variant calling, UMI collapsing, genotyping, demultiplexing
MAGERI	UMI-TAS	UMI consensus, SNV calling (with score)
UMI-VarCal	UMI-TAS	UMI collapsing, SNV calling
cb_sniffer + vartrix	10x	Variant pileup, UMI collapsing, genotyping

Table 5.1: Examples of software that handle variants in UMI / sc-RNA sequencing data, including *umivariants*. Sources: cellSNP + Vireo: Huang et al. [2019]; MAGERI: Shugay et al. [2017]; UMI-VarCal: Sater et al. [2020]; cb_sniffer + vartrix: Petti et al. [2019].

as well (table 5.1). These different methods are generally good solutions for specific tasks and input types: e.g. cellSNP and Vireo for calling haplotypes in and demultiplexing scRNA-seq datasets (Huang et al. [2019]), or MAGERI for variant calling in TAS datasets. However, these are often restricted to one format or data type. In contrast, *umivariants* provides a framework that can help transition seamlessly among different types of UMI data, and to generate outputs that are easy to compare and combine. Therefore, it represents a solution to the complex problem of sequence quality control and integration among multiple samples and methods.

5.2 *umivariants* Tools and Pipelines Allow Efficient and Precise Variant Calling in scRNA-seq and Other UMI-Based Methods

umivariants provides a universal framework for extracting and analyzing variants in multiple kinds of sequencing data that contain UMIs. With its modular and sequential functions, it is possible to design efficient and flexible pipelines that make use of UMI sequence proofreading by estimating the consensus. Its functionality was demonstrated with different datasets from the AML-LT experiment, as well as those from external publications.

As shown throughout the examples in this work, umivariants is capable of handling multiple types of UMI sequencing data (scRNA-seq, targeted amplicons, etc.) by providing multiple options to read the necessary input. Sample barcodes and UMIs can be directly extracted from the BAM file as tags, parsed from the read names, and BCs can even be omitted or provided manually. Variants of interest can be provided as a text table or VCF file, and indel positions are adequately adjusted. After the reads and barcodes are extracted, the user remains in control of downstream quality control, as well as the exact methods and parameters that should be applied for UMI consensus and SNV calling. The results from each step are stored in data frames that can be easily manipulated.

In order to evaluate if the UMI consensus procedure effectively contributed to a reduction in variant calling errors and more precise assignment of mutant cells to clones, I performed a benchmarking analysis using a scRNA-seq dataset with ground-truth variants (section 3.2). I evaluated the false positive, false discovery, and false negative rates that were obtained when calling variants after applying the UMI consensus; these were compared to a control analysis without this step. The results showed that the usage of UMI consensus methods led to a better control of

the false positive calls, without an increase in false negative ones. By using these benchmarking rates to generate simulated datasets based on clonal structures of different complexities, it was also revealed that such reduction in false positive calls was useful to improve the assignment of single cells to their true clones.

From the benchmarking tests, it could be observed that the MaxQualSum method yielded good performance in low coverage scenarios, which would generally be the case with scRNA-seq (as seen in section 3.2). It was also the best option for scRNA-seq targeted amplification methods, such as GoT (section 4.1.2). The methods that were implemented originally in MAGERI, i.e. the majority-vote consensus and its homonymous SNV calling method, would be adequate for the analysis of datasets sequenced at higher depths. For a test dataset of deeply genotyped variants, the umivariants implementation of MAGERI performed comparably to the original software in nearly all cases, in spite of foregoing the assembly step of the original approach (section 3.4). Verifying that appropriate variant calling and genotyping was performed with *umivariants* in the GoT and SiMSen-seq datasets was important to demonstrate that it is a reliable approach for the eventual analysis of other important features of the cells or samples that have to be interpreted in the context of the genetic makeup, such as the molecular phenotype (i.e. gene expression) of mutant cells.

The development of scRNA-seq library preparation and analysis is a fast-paced, highly dynamic field. Because of its modularity and flexibility with input files and parameters, *umivariants* has the potential to enable the analysis of data generated with future generations of methods, as well as to contribute to the development of new ones. With options in the UMI-consensus functions to retain and analyze non-consensus UMI reads, it will be possible to evaluate proportions of chimeric UMIs, i.e. reads from different samples that were mistakenly barcoded with the same UMI and BC sequences (Dixit [2016]). The information of the analysis would then be helpful to design strategies against this phenomenon. Furthermore, multi-threading options have been implemented in the functions for pileup, consensus, and SNV scoring (see section 3.3). These will be important to scale the analysis of ever-increasing numbers of variants, samples, and sequencing depths. The current implementation can already alleviate some bottlenecks in computationally intensive calculations (figure 3.6).

In its current version, *umivariants* is able to retrieve and call a consensus in substitutions and moderately complex indels. Appropriate detection and control of the latter is not a straightforward task, as it can be influenced by all steps in the variant calling pipeline, from mapping to the statistical model, and might often be lost in regions of high genomic complexity. In the future, it would also be relevant to adapt and apply *umivariants* for the analysis of further types of sequence variants. One possibility would be to provide allelic counts that could be used in a single-cell CNV calling framework that requires SNP allelic coverage, such as HoneyBADGER or CaSpER (see section 1.3.2). In this scenario, umivariants should be provided with the complete SNP profile of the sample.

5.3 The Subclonal Architecture of the AML Long-Term Experiment Pinpoints the Different Evolutionary Forces Acting on the Patient and the PDX

The AML long-term experiment represents a unique dataset for the comprehensive analysis of clinical and experimental samples of an AML tumor. The combination of primary patient samples with PDX models that underwent extensive therapy regimes provided a detailed view into the effects of selective pressure from chemotherapy on the clonal composition of the tumor. The availability of data from multiple genomic sequencing methods at different breadths and depths of coverage (WGS, WES, TAS) was critical to enable calling variant and clonal inference in the patient and PDX, resulting in an overarching clonal structure that connected the populations observed in all of the different available samples.

The subclonal architectures and phylogenies inferred from WGS (figure 2.5) and WES (figure 2.14) converged in a common evolutionary history, in which the subclone that was present on the first-relapse sample descended directly from the founder clone (which was predominant at diagnosis), and three different subclones emerged from the first-relapse one: KRAS, NRAS, and second-relapse. It is noteworthy that the observed clonal populations in the relapse and PDX samples descended from the initial one that was predominant at diagnosis, which contrasts with cases from other studies in which such relapse populations were actually derived from different lineages that could even be traced to pre-leukemic stem cells (Shlush et al. [2017]).

The founder clone was characterized by a fixed set of known AML driver mutations in genes like DNMT3A, RUNX1, BCOR, PTPN11, and ETV6. This profile would confer the tumor the necessary shift in fitness to expand its population, representing the initial selective event in this sample. The SNV profile of first-relapse subclone likely does not reflect a product of positive selection, as few non-synonymous mutations were actually damaging (2 in WGS and 2 in WES), and none of these have been recognized as AML drivers. However, this is the subclone where the deletion of the chromosome 7 q-arm appeared, a lesion that has been associated with poor prognosis (Papaemmanuil et al. [2016]). Finally, the second-relapse was likely defined by selection after chemotherapy, possibly driven by the EZH2 SNV and the overlapping deletion in chromosome 7. Furthermore, this subclone contains the largest number of mutations that were swept with the drivers in a single event. The second relapse sample also contains the NRAS Q61K mutation at a low frequency (0.037).

In contrast with the founder and both relapse subclones, the KRAS and NRAS subclones were not prevalent in the patient samples, and the SNVs of these two genes were only found at very low frequency in the diagnosis sample. However, these two subclones constituted the bulk of the tumor population in the PDX samples. The KRAS and NRAS variants might be the only relevant drivers that distinguish these subclones. Several non-synonymous mutations of these subclones are present in large genes that frequently present rare mutations in public exome datasets, such as TTN (KRAS) or USH2A (KRAS-2) (Shyr et al. [2014]). Other variants are present in genes with diverse functions regarding the cytoskeleton, extracellular matrix, translation, respiration, and translation; however, they appear not to have direct associations with leukemias, even if they have been associated to other cancers (such as ADAM12 in NRAS, or NLRP3 in KRAS).

The CNV profiles obtained from all WES samples revealed that, as expected for AML, the chromosomal structure remained largely stable after the appearance of specific events in the relapse, second-relapse, and the most recent common ancestor of the KRAS and NRAS clones. The deletion of the q-arm in chromosome 7 in the first-relapse subclone was the most extensive CNV in the dataset, and might have had an impact on pathways such as MAPK and BRAF signaling. CNV profiles were identical among all PDX samples, further indicating a lack of PDX-specific selection. This makes a contrast to the observations of a study that found pervasive mouse-specific selection and evolution in 1110 PDX samples from 24 cancer types by analyzing their CNV profiles (Ben-David et al. [2017]).

The analysis of mutational signatures of both the patient and PDX subclones showed that the signals from mutational signatures 18 and 24 increased in the relapse and PDX subclones, becoming predominant in the second relapse, KRAS, and NRAS (2.8). Mutational signature 18 was the most prevalent in these cases, and has been potentially associated with DNA damage by reactive oxygen species (ROS). High levels of ROS have been observed across many leukemias, including AML, and have been associated with the control of processes like proliferation and hematopoietic differentiation (Prieto-Bermejo et al. [2018]). The induction of ROS can be mediated by the activation of oncogenes such as RAS, which has been observed to shift metabolism towards anaerobic fermentation and to the increase in ROS production during growth in glucose-depleted medium (Chiaradonna et al. [2006]). This physiological background could contribute to the production of the somatic variants that would eventually constitute the background of the KRAS, NRAS, and second-relapse subclones. In contrast, mutational signature 1, which is associated with aging, had a high frequency in the founder clone; was lost in the KRAS, NRAS, and second-relapse subclones; and increased again in the KRAS-2 and NRAS-2 subclones. In the case of the founder clone, the presence of this signature agrees with the notion that most mutations that initiate AML originate from randomly accumulated mutations in hematopoietic stem cells through the lifetime of the patient (Welch et al. [2012]). The presence of signature 1 was useful to attempt to estimate the age of the founder clone at about 1 year before diagnosis, but the same procedure could not be applied to the other subclones (section 2.9).

5.4 Clonal Fractions in the AML-LT PDX Samples May Reflect Subclone-Specific Sensitivity to Therapies, without Ongoing Adaptation

Tumor populations of the AML-LT PDX samples were characterized by the expansion of the KRAS and NRAS subclones, in contrast to their rare frequencies in the patient. RAS-protein mutations have long been known to be sufficient to induce malignancy in mouse cells (Malumbres and Barbacid [2003]). It has been observed that oncogenic transformation by RAS in mouse has different requirements than in human: namely, stimulation of the RalGEF pathway instead of Raf or PI3K (Hamad et al. [2002]). These mechanisms could have been relevant for the initial engraftment of these populations in the PDX, while other properties would influence the frequencies of these two subclones.

In mouse models, it has been observed that KRAS G12D induces aggressive proliferation in

hematopoietic stem cells, eventually depleting this population, and the KRAS subclone population could be following this regime (Sabnis et al. [2009]). In contrast, cells from NRAS-predominant samples were found to be associated with more stem-cell features and lower expression of cell cycle genes (Philipp Janssen, personal communication), even though the NRAS Q61K variant has been typically associated with MAPK pathway activation (Li et al. [2012], Posch et al. [2016]).

One striking feature of the AML-LT samples that underwent one round of chemotherapy was the transition from a majority of KRAS-dominant samples to more samples with intermediate or high NRAS frequencies. This was largely reverted after additional rounds of treatment. In most samples from stage IV onwards, KRAS was again the predominant subclone. It is important to point out that the first round of chemotherapy had a treatment with cytarabine and daunorubicin, while only cytarabine was used for subsequent treatments. Changes in clonal frequency and decline rate that were observed in this experiment could reflect a specific sensitivity of the cells from the KRAS subclone to daunorubicin. After this agent was removed, differences in the replicative potential of the KRAS cells might have given them a fitness advantage once again. It could be especulated that, in those samples where NRAS was predominant and which had not been treated with daunorubicin, this was due to a stochastic depletion of the KRAS population during engraftment, which would further agree with the lack of selection that was observed in these subclones.

The ability to enable the analysis the mutational profile of the AML-LT single cells from the scRNA-seq data that was generated in the project was one motivation behind the development of *umivariants*. This became possible only for a very limited number of cells, due to low expression and coverage values of the genes and sites that contained the defining subclonal mutations (4.8). Nevertheless, some insights into the molecular profiles of subclonal cells could still be obtained. The analysis of the gene expression profiles of the cells that could be assigned to clones, as well as additional inference of cell populations based on the clonal frequencies from bulk methods, indicated that the transcriptional profiles of the KRAS and NRAS cells were generally similar among themselves and throughout the stages of chemotherapy (figure 4.12; Philipp Janssen, personal communication). This provides additional evidence of a largely neutral evolutionary process in the samples of this experiment, in which no adaptation to the chemotherapy was induced.

Chapter 6

Final Conclusions and Perspectives

UMI-based sequencing technologies will remain under active development and usage in the analysis of tumor and tissue heterogeneity in the foreseeable future. Many of these methods are being actively developed at the moment for bulk and especially single-cell applications. The *umivariants* package has a the potential to contribute to the analysis of the data generated by this diversity of methods, given the features, flexibility, and modularity it offers. It already offers robust and user-friendly solutions for variant pileup, proofreading through UMI consensus, variant calling, and genotyping, which can easily be adapted and extended for future format considerations and dataset magnitudes.

The new opportunities provided by the development of these sequencing and analysis tools will be critical to expand our understanding of AML heterogeneity, evolution, and mechanisms of relapse in both clinical and experimental settings. The use of this tools was essential to study the AML long-term experiment, which represents one example of AML tumor cells that evolve neutrally without adaptation to the pressure of chemotherapy. With the analysis of the genetic profile of these samples in bulk and single-cell datasets, it was possible to observe the response of its subclonal populations to additional treatments, and to contrast their situation with the tumor populations in the patient that were driven by selection.

Single-cell analyses in additional AML PDX models with future methods that yield a higher coverage of the clonal variants will help elucidate the molecular features of subclonal populations that drive their fitness and diversity. I expect *umivariants* to be a useful tool in the analysis of those datasets, and thus contribute to our knowledge on how AML is shaped by evolution.

Chapter 7

Materials and Methods

7.1 AML-LT Experiment

7.1.1 PDX Model Generation and Long-Term Treatment

AML PDX models were established from the primary patient samples by Dr. Binje Vick, from the group of Prof. Dr. Irmela Jeremias (Helmholtz Zentrum München, SFB1243). Tumor cells were transformed to express enhanced firefly luciferase and mCherry, and transplanted into the PDX model to generate the initial AML491 PDX line (Vick et al. [2015]). PDX cells were transplanted into a donor mice 652 -> 1021, and from the latter, cells were transplanted into the PDX mice that would constitute the LT experiment. Tumor burden was monitored by bioluminescence imaging.

The AML PDX samples were treated with up to three regimes of chemotherapy (figure 2.1). The first regime was a treatment over 32 days with 1 mg kg^{-1} of daunorubicin, one dose of 50 mg kg^{-1} of cytarabine, and one dose of 100 mg kg^{-1} of cytarabine. The second and third regimes were a treatment over 32 days with three doses of 100 mg kg^{-1} of cytarabine. Due to clinical signs of illness, tumor samples from mice of stage IV were re-transplanted and used to generate the samples from stages V, VI, and VII. Cells were extracted after full-blown growth (I, III, V, VII), or shortly after the application of chemotherapy when the cells were depleted (II, IV, VI).

7.1.2 Whole-Genome Sequencing

WGS data of the patient and PDX samples were generated by the group of Dr. Philipp Greif (ELLF-LMU, SFB1243). WGS reads from 5 PDX samples (paired-end, 150 bp) were aligned to a concatenated human and mouse genome (hs37d5-GRCm38) using bwa mem (version 0.7.15-r1140; Li [2013]). Reads from human chromosomes were coordinate-sorted and extracted with samtools version 1.8 (Li et al. [2009]) and used for downstream analysis.

7.1.3 Whole-Exome Sequencing

WES data of the patient and PDX samples were generated by Julia Niggemeyer (Metzeler Group, ELLF-LMU, SFB1243). Paired-end 100 bp whole exome sequencing (WES) reads were mapped using BWA-MEM. Reads from patient samples were mapped to the human reference genome,

Run ID	UMI length	BC length	Read length
Run 1	10	14	50
RFC150	10	14	150
RFC18_100	10	16	100

Table 7.1: AML-LT scRNA-seq dataset configurations.

version hg19. Reads from the PDX samples were mapped to a concatenated human-mouse reference genome (hg19 - GRCm38), and reads mapping to standard human chromosomes were extracted with samtools version 1.8.

7.1.4 Targeted Amplicon Sequencing and Ultra-Sensitive Genotyping

TAS data was generated with the HaloPlex AML panel of 67 genes by Dr. Maja Rothenberg-Thurley and Dr. Klaus Metzeler (ELLF-LMU, SFB1243).

PDX samples of stages IV and V were re-sequenced using SiMSen-seq (Ståhlberg et al. [2017]) by Daniel Richter (SFB1243). The 5 SNVs described in table 2.3 were amplified. Libraries were sequenced with 100-bp paired-end reads.

7.1.5 Single-Cell RNA-Sequencing

cDNA libraries of 2276 cells from 28 samples of the AML-LT experiment (stages I, II, III, IV, VI, and VII) were successfully prepared with the mcSCRB-seq protocol (Bagnoli et al. [2018]) by Johannes Bagnoli. These were generated under three different configurations of library preparation and/or sequencing, which are shown in table 7.1. In all cases, BCs were a composite of the SCRB BC sets and a TruSeq i7 anchor.

Reads from all three configurations were pre-processed uniformly. Reads were mapped to a concatenated hg38-mm10 (human-mouse) genome using zUMIs Parekh et al. [2018], with STAR 2.6 (Dobin et al. [2013]). Cells were filtered for high quality by Johannes Bagnoli with a minimum of 13,000 reads, 2000 UMIs mapping to exons and 3000 UMIs mapping to introns or exons, 1000 detected genes, 50% of reads mapping to introns or exons, 25 UMIs mapping to ERCC sequences, and a maximum of 30% reads mapped to mm10 genes. GATK best practices were used to pre-process the mapped reads (Van der Auwera et al. [2013]). Pileup and consensus of exome clonal variants (described in section 2.3) were done with umivariants. In the case of RFC18_100 data, BCs were trimmed to 14 bases after pileup (an extra pair of TT at the 3' end was removed). Consensus, VAF, and genotype were estimated from the reads of all three datasets jointly.

7.1.6 Somatic SNV Calling

WGS data

SNVs and indels were called following the GATK best practices, version 3.6 (Van der Auwera et al. [2013]), with MuTect2 (Cibulskis et al. [2013]), using the remission and post-bone-marrow transplantation samples (labeled "Donor") as normal controls. The ExAC release 0.3.1 (Lek et al. [2016]) was used as a reference for common SNPs. SNVs that were annotated in ExAC with a VAF > 0.05 were excluded.

Variants from the primary patient WGS data were called by collaborators using the DKFZ OTP pipeline (Reisinger et al. [2017]; Sebastian Vosberg, personal communication).

SNVs and short indels from both patient and PDX samples were further filtered to have a minimum coverage of 50 reads in all samples, a maximum allele frequency of 0.75 in any sample, and a VAF = 0 in normal controls (except for DNMT3A R882S, which was persistent at remission). SNVs and indels were also removed if they had a $P - value > 0.01$ in a binomial test $B(VAF > 0.5|K, N)$ given K variant allele reads and N total coverage. After applying these filters, 7461 SNVs and indels remained.

WES data

SNVs and indels were called following the GATK best practices with MuTect2 (version 3.6), using the remission and post-bone-marrow transplantation samples as normal controls. ExAC 0.3.1 was used as a polymorphism reference. Variants were filtered to have a minimum depth of 50 reads and a maximum VAF of 0.75 across all patient and PDX samples, at least two variant allele reads and a VAF of 0.1 in at least one tumor sample, and a VAF of 0 in controls (except for DNMT3A R882S). Sites with more than one variant allele were excluded.

7.1.7 SNP Calling

Germline SNPs and indels were called following the GATK best practices with HaplotypeCaller, version 3.6. A set of core heterozygous SNPs was defined as being called in the patient diagnosis and complete remission samples (i.e. before any bone marrow transplantation), with a VAF = 0.5 on both samples, a depth of 20 reads, and a frequency ≤ 0.05 in ExAC 0.3.1.

7.1.8 CNV Calling and Overlapping Gene Analysis

Allele-specific copy number variants were called on the exome data using the MARATHON pipeline (Urrutia et al. [2018]), which makes use of the CODEX2 and FALCONX packages (Chen et al. [2017], Jiang et al. [2018]). The pipeline estimates the copy number based on the coverage per allele at heterozygous SNP sites. The core set of heterozygous SNPs from the diagnosis and remission exomes was given as input to CODEX2 to normalize the coverage on the patient and PDX tumor samples with respect to the three control samples (first full remission, and remission after first and second BM transplants). FALCONX was used to call chromosomal regions with allele-specific copy number differences between each tumor-control pair based on the coverage at heterozygous SNPs which were observed in both samples. Major and minor copy number estimates were homogenized along stretches of 1 Mbp, and segments were delimited based on local copy number differences of 0.3.

Raw CNV calls were further processed to yield a unique set of consensus CNV regions for all samples. First, I obtained the intersection of CNV regions per sample that were called against each control. Afterwards, I merged them across all samples using the reduce() function of the GenomicRanges R package (Lawrence et al. [2013]) (i.e. to combine segments with at least a 1-bp overlap), and filtered any regions below 10 Mbp in length. To establish segments within the merged regions where the CNV was not present in all samples, I used the GenomicRanges disjoin() function to obtain all possible non-overlapping CNV segments. Any adjacent segments

that were below 1 Mbp in length were merged. I also merged larger segments with the smaller ones if they were present in less than 3 samples, or more than 13 (out of 16). This strategy kept regional differences in copy number per sample, but vastly reduced the number of smaller, separate segments to be considered. The final copy number that was reported per segment corresponds to the mean between the copy number values per sample (i.e. from the FALCON-X output), weighted by the length of the CNV segment. Segments that were not called in a sample were assigned a copy number of 1 (for both major and minor).

In order to find which genes could be affected by the CNV events, I extracted them by their overlap with the CNV coordinates based on the GRCh37.75 annotation, importing the corresponding GTF file as a txdb object in R. ENSEMBL gene identifiers were converted to their gene names and Entrez Gene ID with the biomaRt package v2.42.0 (Durinck et al. [2009]). The number of genes that overlapped with each CNV is shown in table 2.2. To find the broader functional context of these CNV genes, I ran a pathway enrichment analysis with the ReactomePA package (Yu and He [2016]). Gene sets for pathway enrichment were defined by the CNVs that were present per sample or per chromosome.

7.1.9 Clonal Inference

The set of filtered SNVs and indels from the WGS analysis were used for cluster inference with the SciClone package, version 1.1.0 (Miller et al. [2014]). I used the remaining 7 clusters, comprising 6384 variants, to infer the clonal phylogeny and frequencies per sample with the ClonEvol package, version 0.99.11 (Dang et al. [2017]).

In the case of the data from the 13 WES samples, the Canopy package version 1.3.0 was employed (Jiang et al. [2016]). The second relapse sample and CNVs were excluded from the analysis. Canopy was set to perform inference from 3 to 10 clones with an initial binomial clustering step of 100 EM runs. MCMC sampling of trees was done on each number of clusters using Canopy's canopy.sample.cluster.nocna function, using 100 chains with a minimum of 20000 iterations and a maximum of 200000. The first 100 iterations were excluded as burn-in. The number and composition of clones was determined with the maximum Bayesian Information Criterion (BIC), and the best clonal phylogeny for the chosen architecture was determined by maximum likelihood. Clonal fractions per sample were directly extracted from the Canopy output.

7.1.10 Clonal Age Analysis

I estimated approximate age ranges per clone following a method that was introduced by Körber et al. [2019]. To calculate the time that would take for each clone to acquire its set of mutations, I used the following equation:

$$T(m) = \frac{m}{\mu * \lambda}$$

where m is the number of mutations on the WGS dataset that were acquired on that particular clone (i.e. excluding any ancestral ones), μ is the mutation rate, and λ is the growth rate. The growth rate was calculated from the bioluminescence proliferation assays of KRAS- or NRAS-predominant PDX samples. For the founder, first-relapse, and second-relapse clones I took the mean growth rate of these assays. The standard deviation of the growth rates was used to provide upper and

Package	Version
GenomicRanges	1.38.0
GenomicAlignments	1.22.1
GenomicFeatures	1.31.1
Rsamtools	2.2.1
VariantAnnotation	1.32.0
Cardelino	0.99.0
tidyverse	1.2.1
dplyr	0.8.3
tidyr	1.0.0
parallel	3.6.0
Biostrings	2.5.4.0
data.table	1.12.8
dtplyr	1.0.1
stringdist	0.9.5.5
ComplexHeatmap	2.2.0

Table 7.2: Versions of *umivariants* dependencies (R packages)

lower limits. Growth rates were provided as doublings per week. The mutation rate per subclone was calculated with $\mu = \frac{\mu_{base}}{cov_{50} * sig_1}$, where μ_{base} is the basal somatic mutation rate in humans, which was estimated as a median of $2.8 * 10^{-7} bp^{-1} generation^{-1}$ by Milholland et al. [2017]), cov_{50} is the number of sites covered at 50x across all WGS samples (due to the minimum coverage required for SNV calling), and sig_1 is the ratio of the mutational signature 1 in the subclonal SNVs to those present in the parent clone (or just the fraction of signature 1 for the founder clone).

7.2 *umivariants* Development and Testing

7.2.1 Software Versions

umivariants version 0.0.0.9000 was developed in R version 3.6.0. Essential dependencies are shown in table 7.2.

7.2.2 Variant pileup

To extract variants from UMI-barcoded reads, variants are imported from the input file and converted into a GenomicRanges object (Lawrence et al. [2013]). Sequences at each variant position are extracted from reads in the input BAM file that overlap the variant coordinate, using the stackStringsFromBam() function from the GenomicAlignments package (Lawrence et al. [2013]). This same function is used to extract the Phred-scaled sequence quality scores. The input BAM file needs to be coordinate-sorted, and a BAM index file (.bai extension) needs to be created in its same location.

Tags containing the UMI and BC, read query names, and mapping quality values of each analyzed read are extracted with the Rsamtools package (Morgan et al. [2019]), and merged with the stackStringsFromBam() output. If minimum sequence and/or mapping quality values are provided by the user, reads are filtered by these values. UMIs and BCs are parsed depending on the input BAM file configuration:

- If the UMI and barcode are contained in tags, their names can be provided directly as parameters. These are used as part of the scanning parameters in scanBamParam() from Rsamtools.

- If the UMI and BC are contained in the same string, these can be split by position.

- If the UMI is contained in the read query name (QNAME), it can be extracted with different methods, such as splitting the QNAME by a separator character, a regular expression, or a fixed position.

- If the sample contains no BC, a sample name can be provided as a parameter, or the UMI can be copied into the BC column.

Each variant position is handled individually, a process which can be divided among multiple threads using the 'cores' option.

The type of variant at each location is determined from the alleles presented in the input SNV table. Four possible types of variants can be handled by the scan_UMI_bam() function: SNVs (i.e. single-nucleotide substitutions), insertions, deletions, and multi-nucleotide substitutions (i.e. replacements, see Danecek et al. [2011]). While SNVs are extracted directly with the call to stackStringsFromBam(), the other three types of variants require additional processing to extract the full sequence of the alleles and to obtain quality values that can be employed for subsequent consensus estimation:

- **Multi-nucleotide substitutions:** if multiple substitutions occur contiguously, the sequences can be piled up directly with the call to stackStringsFromBam(). Quality values of all nucleotides are averaged for subsequent steps of UMI consensus and scoring.

- **Deletions:** two kinds of characters can be used to represent a deletion by default when using stackStringsFromBam(): '-', for deletions represented with a D in the CIGAR string, and '.', for large deletions that were assigned to an N. The N character can also be used to convey any case in which a large chromosomal region is skipped, such as in a splice junction. In order to confirm that the extracted string matches the expected deletion, an additional function verifies that the length of the '-' or '.' string matches the length of the annotated deletion.

- **Insertions:** the position of the read that contains the insertion is extracted from the CIGAR string. The cigar_breaker() function from *umivariants* parses the CIGAR string and provides a table where each row corresponds to the different operations in succession. If the CIGAR contains the character 'I', preceded by the length of the insertion, the corresponding row of the cigar_breaker() table is used to extract the position of the insertion in the read. Instead of extracting the sequence with stackStringsFromBam(), the complete reads are extracted with Rsamtools scanBam(), and the substring is extracted based on the cigar_breaker() table.

7.2.3 UMI-Consensus Models

MaxQualSum model

Quality scores are converted from ASCII format (i.e. character) to a numeric Phred-scaled value: $Q_{Phred} = A - O$, where A is the numeric value of the character in ASCII format, and O is an offset value. The latter is defined by the Illumina/CASAVA version, which is equivalent to 33 in versions 1.8 and above. The raw score of each allele (i.e. single nucleotide, multiple nucleotide, or indel) per UMI, position, and sample BC is determined by adding the Phred scores of all reads that present the allele. The consensus allele is determined by the highest sum of Phred scores. In the case of a tie between two or more alleles, the UMI is flagged for potential exclusion by the user; they can be discarded in the output table already if specified, which is the default. The mean Phred score of the nucleotides/indels with the consensus allele is taken as the consensus quality score.

The Phred quality score (Q) is equal to $q = -10log_{10}(p)$ (Ewing and Green [1998]), where p is the error probability defined by CASAVA. In the QUAL field, it is converted into the ASCII character with the equivalent rounded value + 33 (Illumina $\geq$ 1.8) or + 64 (Illumina $<$ 1.8).

Majority vote model

After the pileup step, read sequences are filtered with a user-provided minimum Phred score value. The number of reads that contain each observed allele per UMI, position, and sample BC are counted. The allele with the highest frequency is taken as the consensus. UMIs with ties between two or more alleles are also flagged in this case, and optionally discarded.

The consensus quality score is computed after Shugay et al. [2017] with the following equation:

$$CQS = \frac{40 * (4 * F_{cons} - 1)}{3}$$

where F_{cons} is the frequency of the consensus nucleotide in the UMI.

fgbio likelihood model

The complete fgbio model is described in the Wiki of the project's Github page (Fennel et al.; https://github.com/fulcrumgenomics/fgbio/wiki/Calling-Consensus-Reads). The model is based on the likelihood of each possible nucleotide substitution to originate from an amplification or sequencing error. The consensus is estimated using the following steps on each nucleotide from each read that overlaps the position of interest. The equations were taken from the fgbio GitHub Wiki page.

1. ASCII Phred scores are converted to their numeric value (Q), which is converted back to the error probability by $P_{err} = 10^{-Q/10}$. (P_{err} is written as $P_{Q'}$ in the original reference, where Q' is a re-scaled Phred score required in cases of systematic over-estimation, which is not applied here.)

2. The combined error probability for amplification and sequencing is estimated with this

formula:

$$P'_{err} = Err_{post} * (1 - P_{err}) + (1 - Err_{post}) * P_{err} + (Err_{post} * P_{err} * 2/3)$$

Err_{post} is an expected error after UMI incorporation; i.e. it would correspond to errors during amplification, and is a parameter provided by the user.

3. The next step is the estimation of the likelihood of each nucleotide (A, C, G, or T) to be the one in the real molecule. The likelihood of each base (B) is estimated with this formula:

$$L_{Call=B} = \prod_i \begin{Bmatrix} P'_{err,i}/3 & if B \neq B_i \\ (1 - P'_{err,i}) & if B = B_i \end{Bmatrix}$$

4. The posterior probability of each base for being the correct consensus is calculated by dividing the likelihood of the base by the sum of the likelihoods of all four bases, and the base with the highest posterior probability is taken as the consensus:

$$Post_{Call=B} = \frac{LL_{Call=B}}{\sum_{C \in \{A,C,G,T\}} LL_{Call=C}}$$

5. The posterior probability of the consensus is converted back to an error probability by $P_{err} = 1 - Post_{Call}$. The final error estimate is done by incorporating the probability of having an error before incorporating the UMIs (e.g. during reverse transcription), where Err_{pre} is a user-provided prior for such error:

$$P'_{err} = Err_{pre} * (1 - P_{err}) + (1 - Err_{pre}) * P_{err} + (Err_{pre} * P_{err} * 2/3))$$

As a final step, this is converted back to a Phred-scaled consensus quality score:

$$Q_{call} = -10 * log10(P'_{err})$$

No consensus model

This option was incorporated to generate controls for the evaluation of the three UMI consensus models. Counts and frequencies of each allele per UMI, position, and sample BC are estimated. The mean Phred score per allele is calculated, but consensus quality scores are not provided. Quality or frequency ties among alleles are not flagged or excluded.

7.2.4 SNV-Calling Models

MuTect model

The MuTect model for variant calling was created by Cibulskis et al. [2013], and the following equations were extracted from this source with permission under RightsLink license number 4836740058907.

The main principle consists in performing a likelihood ratio test between the two possible variant calling models: M_0, in which there is no variant at the site, and M_f^m, in which variant m

exists with frequency f. M_0 is eqivalent to M_f^m with f = 0, The likelihood of M_f^m is defined as:

$$L(M_f^m) = P(b_i|e_i, r, m, f) = \prod_{i=1}^{d} P(b_i|e_i, r, m, f)$$

where r is the reference allele $r \in A, C, G, T$, b_i is the called base of read i, and e_i is the sequence error. $P(b_i|e_i, r, m, f)$ is defined from e_i depending on whether b_i is equal to the r, m, or another sequence.

Variant detection in the tumor (T) is done by dividing the likelihood of the two models to obtain a LOD score:

$$LOD_T(m, f) = log_{10}\left(\frac{L(M_f^m)P(m, f)}{L(M_0)(1 - P(m, f))}\right) \leq log_{10}\delta_T$$

$\leq log_{10}\delta_T$ is a decision threshold value equal to 6.3 in the publication, and in the *umivariants* default parameter. If $LOD_T(m, f) \leq log_{10}\delta_T$, the variant is considered to be called.

MAGERI model

The MAGERI variant calling model was published by Shugay et al. [2017]. It is reproduced here as permitted by the Creative Commons BY 4.0 License.

MAGERI is based on an implementation of the Beta-Binomial model which covers the inference of errors generated during both PCR amplification and sequencing. It is defined for single based substitutions, defined into six categories (A>C / T>G; A>G / T>C; A>T / T>A; C>A / G>T; C>G / G>C; C>T / G>A).

Error frequencies per substitution type (xy) are fitted with a Beta distribution:

$$\epsilon_{xy}\ Beta(\alpha_{xy}, \beta_{xy}); x, y \in A, T, C, G$$

where the values for α_{xy}, β_{xy} were estimated by the authors based on empirical polymerase error rates published in Shagin et al. [2017]. Such parameters are stored internally in *umivariants.*

The total observed error counts n_{xyi} at position i given coverage N_i are modelled with the Beta-binomial distribution:

$$n_{xyi} \sim BetaBinom(N_i, \alpha_{xy}, \beta_{xy})$$

Quality scores of the variant call (Q score) are defined as:

$$Q = -10log_{10}PBetaBinom(n_xyi, N_i, \alpha_{xyi}, \beta_{xy})$$

and Q is capped to a maximum value of 100. A minimum value of Q can be used as a threshold to define a variant as called (defaults to 20).

7.2.5 Variant genotyping

Variants are genotyped per sample BC using the UMI-consensus and SNV score table by following the following steps:

- The initial classification is based exclusively on the VAF: 0 or 1 for the homozygous cases, any value in-between for heterozygotes, and '-/-' for variant sites without coverage.

- If the SNV score was calculated and is provided as input, any variants with VAF > 0 which did not pass the score threshold are re-classified as $0/0$.

- For variants with a VAF $= 1$ that is the result of very low coverage, a heterozygosity score is calculated to determine the probability of being an actual heterozygote.

The heterozygosity score is estimated from a binomial test per site and sample:

$$P_{het} = B(a|d, \theta)$$

where a is the variant allele UMI coverage, d is the total coverage, and θ is a probability prior. θ is calculated as

$$\theta = \frac{(1 - P_{miss})}{CN}$$

CN is the copy number of the site at the sample (default $= 2$). P_{miss} is the probability of missing one allele (i.e with 0 reads) based on the read coverage, and is calculated with

$$P_{miss} = B(0|d_r, P_{cov}/CN)$$

where d_r is the read depth at the site. P_{cov} is the coverage probability, and is empirically calculated on the whole dataset as the fraction of samples (BCs) with a coverage > 1 read for that site.

In variants with a VAF $= 1$, if the P-value of the heterozygosity score and $P_{missing}$ are above user-specified thresholds (0.05 and 0.8 as default, respectively), then the genotype is assigned as $0/1$.

7.3 Benchmark of Variant Calling and Clonal Assignment with the UMI Consensus Methods

7.3.1 Benchmarking true and false positive calls

The set of core exome SNPs that was described in section 7.1.7 was used as a ground truth dataset to evaluate variant calls in the AML-LT scRNA-seq dataset after calling the UMI consensus with the 3 different methods implemented in *umivariants*, as well as the control without consensus. The read coverage of all SNPs was calculated in the complete scRNA-seq datastet. SNP subsets were defined by filtering the sites according to each coverage value. The SNPs were taken as true positives and any other possible variants were taken as false positives. The confusion matrix in table 3.1 was estimated based on this classification of true and false positives; undetected SNPs were called as false negatives. True negatives were defined if the variant call from MAGERI or MuTect was equal to FALSE, and the variant had a different allele from the expected SNP.

7.3.2 Benchmarking clonal assignment with the variant calling rates

Clonal phylogenies were designed to vary in the total number of clones, and the branch length (i.e. number of variants) on each of the daughter clones . These trees were written in Newick format with the desired branch lengths, and imported into R with the ape package (Paradis and Schliep [2019]). Ancestral and offspring clones were integrated in each tree.

On each of these clonal phylogenies, sets of randomly selected SNVs were assigned to each of the corresponding clones. The number of SNVs per clone was taken from the branch lengths. Based on these SNV-clone assignments, sets of single cells were randomly assigned to the clones based on a simulated clonal frequency. SNV detection on each cell was simulated with a method based on the simulator from the OncoNEM package (Ross and Markowetz [2016]): a binary cell-SNV matrix was generated with the true SNVs per clone for each cell, and coverage values of the reference and variant alleles of each SNV were sampled from a binomial distribution to reflect SNV detection and site coverage. A random number of these values was sampled based on the FPR to assign false positives, on the FNR to assign false negatives, or on a dropout rate (0.1) to assign lack of coverage. With these simulated allele counts, cells were assigned to the set of subclones using Cardelino (McCarthy et al. [2020]). The number of correct assignments, incorrect assignments, and unassigned cells was estimated based on the ground truth annotation of each simulated cell.

7.4 GoT Data Analysis

The 10x and GoT sequencing data from the ET and MF samples were downloaded from the Gene Expression Omnibus using the accession number GSE117826. 10x Cell Ranger count matrices were available for all samples in Matrix Market Exchange Format (MEF; .mtf). BC whitelists of each sample were also downloaded from GEO. Some of the 10x datasets (ET01, ET02, ET03, MF01, and MF05) were directly available and downloaded in BAM format.

Reads for the ET02-GoT dataset were downloaded from the NCBI Sequence Read Archive (SRA) in FASTQ format from accesion numbers SRR7613783 and SRR7613784. The ET02 FASTQ files were mapped to the hg38 genome using zUMIs with STAR 2.7. The variant sites described in table 4.2 were retrieved from the BAM files and analyzed using *umivariants* as described in 4.1.

References

R. Acuna-Hidalgo, H. Sengul, M. Steehouwer, M. van de Vorst, S. H. Vermeulen, L. A. L. M. Kiemeney, J. A. Veltman, C. Gilissen, and A. Hoischen. Ultra-sensitive sequencing identifies high prevalence of clonal Hematopoiesis-Associated mutations throughout adult life. *Am. J. Hum. Genet.*, 101(1):50–64, July 2017. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2017.05.013. URL `http://dx.doi.org/10.1016/j.ajhg.2017.05.013`. [14]

M. Aigner, J. Feulner, S. Schaffer, R. Kischel, P. Kufer, K. Schneider, A. Henn, B. Rattel, M. Friedrich, P. A. Baeuerle, A. Mackensen, and S. W. Krause. T lymphocytes can be effectively recruited for ex vivo and in vivo lysis of AML blasts by a novel CD33/CD3-bispecific BiTE antibody construct. *Leukemia*, 27(5):1107–1115, Apr. 2013. ISSN 0887-6924, 1476-5551. doi: 10.1038/leu.2012.341. URL `http://dx.doi.org/10.1038/leu.2012.341`. [16]

L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, M. Imielinsk, N. Jäger, D. T. W. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, Aug. 2013. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12477. URL `http://dx.doi.org/10.1038/nature12477`. [3]

L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, PCAWG Mutational Signatures Working Group, G. Getz, S. G. Rozen, M. R. Stratton, and PCAWG Consortium. The repertoire of mutational signatures in human cancer. *Nature*, 578 (7793):94–101, Feb. 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-1943-3. URL `http://dx.doi.org/10.1038/s41586-020-1943-3`. [23]

D. A. Arber, A. Orazi, R. Hasserjian, J. Thiele, M. J. Borowitz, M. M. Le Beau, C. D. Bloomfield, M. Cazzola, and J. W. Vardiman. The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20):2391–2405, May 2016. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2016-03-643544. URL `http://dx.doi.org/10.1182/blood-2016-03-643544`. [1]

J. W. Bagnoli, C. Ziegenhain, A. Janjic, L. E. Wange, B. Vieth, S. Parekh, J. Geuder, I. Hellmann, and W. Enard. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat.*

Commun., 9(1):2937, July 2018. ISSN 2041-1723, 2041-1723. doi: 10.1038/s41467-018-05347-6. URL https://www.nature.com/articles/s41467-018-05347-6. [10, 76]

U. Ben-David, G. Ha, Y.-Y. Tseng, N. F. Greenwald, C. Oh, J. Shih, J. M. McFarland, B. Wong, J. S. Boehm, R. Beroukhim, and T. R. Golub. Patient-derived xenografts undergo mouse-specific tumor evolution. *Nat. Genet.*, Oct. 2017. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3967. URL http://dx.doi.org/10.1038/ng.3967. [71]

D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, and L. Lichtenstein. Calling somatic SNVs and indels with mutect2. Dec. 2019. URL https://www.biorxiv.org/content/10.1101/861054v1.supplementary-material. [6, 7]

F. Bewicke-Copley, E. Arjun Kumar, G. Palladino, K. Korfi, and J. Wang. Applications and analysis of targeted genomic sequencing in cancer studies. *Comput. Struct. Biotechnol. J.*, 17: 1348–1359, Nov. 2019. ISSN 2001-0370. doi: 10.1016/j.csbj.2019.10.004. URL http://dx.doi.org/10.1016/j.csbj.2019.10.004. [5]

I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. U. S. A.*, 107(43):18545–18550, Oct. 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1010978107. URL http://dx.doi.org/10.1073/pnas.1010978107. [3]

Broad Institute. Calling variants in rnaseq. https://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq, 2014 (Accessed: 2020/05/10). [8]

Broad Institute. Somatic short variant discovery (snvs + indels). https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels-, 2020 (Accessed: 2020/05/10). [8]

Broad Institute. Picard tools. http://broadinstitute.github.io/picard/, (Accessed: 2020/04/13; version 2.22.3). [6]

K. R. Campbell, A. Steif, E. Laks, H. Zahn, D. Lai, A. McPherson, H. Farahani, F. Kabeer, C. O'Flanagan, J. Biele, J. Brimhall, B. Wang, P. Walters, I. Consortium, A. Bouchard-Côté, S. Aparicio, and S. P. Shah. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol.*, 20(1):54, Mar. 2019. ISSN 1465-6906. doi: 10.1186/s13059-019-1645-z. URL http://dx.doi.org/10.1186/s13059-019-1645-z. [13]

Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120, Oct. 2013. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.2764. URL http://dx.doi.org/10.1038/ng.2764. [3]

G. Caravagna, T. Heide, M. J. Williams, L. Zapata, D. Nichol, K. Chkhaidze, W. Cross, G. D. Cresswell, B. Werner, A. Acar, L. Chesler, C. P. Barnes, G. Sanguinetti, T. A. Graham, and A. Sottoriva. Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat. Genet.*, 52(9):898–907, Sept. 2020. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-020-0675-5. URL http://dx.doi.org/10.1038/s41588-020-0675-5. [9]

D. R. Caswell and C. Swanton. The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. *BMC Med.*, 15(1):133, July 2017. ISSN 1741-7015. doi: 10.1186/s12916-017-0900-y. URL http://dx.doi.org/10.1186/s12916-017-0900-y. [3]

A. Cesano, J. A. Hoxie, B. Lange, P. C. Nowell, J. Bishop, and D. Santoli. The severe combined immunodeficient (SCID) mouse as a model for human myeloid leukemias. *Oncogene*, 7(5): 827–836, May 1992. ISSN 0950-9232. URL `https://www.ncbi.nlm.nih.gov/pubmed/1570153`. [16]

H. Chen, Y. Jiang, K. N. Maxwell, K. L. Nathanson, and N. Zhang. ALLELE-SPECIFIC COPY NUMBER ESTIMATION BY WHOLE EXOME SEQUENCING. *Ann. Appl. Stat.*, 11(2): 1169–1192, June 2017. ISSN 1932-6157. doi: 10.1214/17-AOAS1043. URL `http://dx.doi.org/10.1214/17-AOAS1043`. [7, 77]

Z. Chen, Y. Yuan, X. Chen, J. Chen, S. Lin, X. Li, and H. Du. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci. Rep.*, 10(1):3501, Feb. 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-60559-5. URL `http://dx.doi.org/10.1038/s41598-020-60559-5`. [6]

F. Chiaradonna, E. Sacco, R. Manzoni, M. Giorgio, M. Vanoni, and L. Alberghina. Ras-dependent carbon metabolism and transformation in mouse fibroblasts. *Oncogene*, 25(39):5391–5404, Aug. 2006. ISSN 0950-9232. doi: 10.1038/sj.onc.1209528. URL `http://dx.doi.org/10.1038/sj.onc.1209528`. [71]

S.-Y. Cho, W. Kang, J. Y. Han, S. Min, J. Kang, A. Lee, J. Y. Kwon, C. Lee, and H. Park. An integrative approach to precision cancer medicine using Patient-Derived xenografts. *Mol. Cells*, 39(2):77–86, Feb. 2016. ISSN 1016-8478, 0219-1032. doi: 10.14348/molcells.2016.2350. URL `http://dx.doi.org/10.14348/molcells.2016.2350`. [16]

K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31(3):213–219, Mar. 2013. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.2514. [6, 8, 41, 76, 82]

P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, Aug. 2011. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btr330. URL `http://dx.doi.org/10.1093/bioinformatics/btr330`. [39, 80]

H. X. Dang, B. S. White, S. M. Foltz, C. A. Miller, J. Luo, R. C. Fields, and C. A. Maher. ClonEvol: clonal ordering and visualization in cancer sequencing. *Ann. Oncol.*, 28(12):3076–3082, Dec. 2017. ISSN 0923-7534, 1569-8041. doi: 10.1093/annonc/mdx517. [9, 20, 78]

L. Ding, T. J. Ley, D. E. Larson, C. A. Miller, D. C. Koboldt, J. S. Welch, J. K. Ritchey, M. A. Young, T. Lamprecht, M. D. McLellan, J. F. McMichael, J. W. Wallis, C. Lu, D. Shen, C. C. Harris, D. J. Dooling, R. S. Fulton, L. L. Fulton, K. Chen, H. Schmidt, J. Kalicki-Veizer, V. J. Magrini, L. Cook, S. D. McGrath, T. L. Vickery, M. C. Wendl, S. Heath, M. A. Watson, D. C. Link, M. H. Tomasson, W. D. Shannon, J. E. Payton, S. Kulkarni, P. Westervelt, M. J. Walter, T. A. Graubert, E. R. Mardis, R. K. Wilson, and J. F. DiPersio. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, Jan. 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature10738. URL `http://dx.doi.org/10.1038/nature10738`. [xiii, 1, 3, 5]

A. Dixit. Correcting chimeric crosstalk in single cell RNA-seq experiments. *bioRxiv*, 2016. doi: 10.1101/093237. URL `https://github.com/asncd/schimera`. [69]

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan. 2013. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/bts635. URL `http://dx.doi.org/10.1093/bioinformatics/bts635`. [6, 39, 76]

H. Döhner, D. J. Weisdorf, and C. D. Bloomfield. Acute myeloid leukemia. *N. Engl. J. Med.*, 373 (12):1136–1152, Sept. 2015. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMra1406184. URL http://dx.doi.org/10.1056/NEJMra1406184. [1]

S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomart. *Nat. Protoc.*, 4(8):1184–1191, July 2009. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2009.97. URL http://dx.doi.org/10.1038/nprot.2009.97. [78]

S. Ebinger, E. Z. Özdemir, C. Ziegenhain, S. Tiedt, C. Castro Alves, M. Grunert, M. Dworzak, C. Lutz, V. A. Turati, T. Enver, H.-P. Horny, K. Sotlar, S. Parekh, K. Spiekermann, W. Hiddemann, A. Schepers, B. Polzer, S. Kirsch, M. Hoffmann, B. Knapp, J. Hasenauer, H. Pfeifer, R. Panzer-Grümayer, W. Enard, O. Gires, and I. Jeremias. Characterization of rare, dormant, and Therapy-Resistant cells in acute lymphoblastic leukemia. *Cancer Cell*, 30 (6):849–862, Dec. 2016. ISSN 1535-6108, 1878-3686. doi: 10.1016/j.ccell.2016.11.002. URL http://dx.doi.org/10.1016/j.ccell.2016.11.002. [12]

A. D. Ewing, K. E. Houlahan, Y. Hu, K. Ellrott, C. Caloian, T. N. Yamaguchi, J. C. Bare, C. P'ng, D. Waggott, V. Y. Sabelnykova, ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants, M. R. Kellen, T. C. Norman, D. Haussler, S. H. Friend, G. Stolovitzky, A. A. Margolin, J. M. Stuart, and P. C. Boutros. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods*, 12(7): 623–630, July 2015. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3407. URL http://dx.doi.org/10.1038/nmeth.3407. [6]

B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Res.*, 8(3):186–194, Mar. 1998. ISSN 1088-9051. doi: 10.1101/gr.8.3.186. URL https://www.ncbi.nlm.nih.gov/pubmed/9521922. [81]

J. Fan, H.-O. Lee, S. Lee, D.-E. Ryu, S. Lee, C. Xue, S. J. Kim, K. Kim, N. Barkas, P. J. Park, W.-Y. Park, and P. V. Kharchenko. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.*, June 2018. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.228080.117. URL http://dx.doi.org/10.1101/gr.228080.117. [12]

H. Farahani, C. P. E. de Souza, R. Billings, D. Yap, K. Shumansky, A. Wan, D. Lai, A.-M. Mes-Masson, S. Aparicio, and S. P Shah. Engineered in-vitro cell line mixtures and robust evaluation of computational methods for clonal decomposition and longitudinal dynamics in cancer. *Sci. Rep.*, 7(1):13467, Oct. 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-13338-8. URL http://dx.doi.org/10.1038/s41598-017-13338-8. [9]

T. Fennel, N. Homer, and Fulcrum Genomics. fgbio: Tools for working with genomic and high throughput sequencing data. URL https://github.com/fulcrumgenomics/fgbio. [15, 40, 81]

D. Gao and Y. Chen. Organoid development in cancer genome discovery. *Curr. Opin. Genet. Dev.*, 30:42–48, Feb. 2015. ISSN 0959-437X, 1879-0380. doi: 10.1016/j.gde.2015.02.007. URL http://dx.doi.org/10.1016/j.gde.2015.02.007. [15]

J.-P. Gillet, S. Varma, and M. M. Gottesman. The clinical relevance of cancer cell lines. *J. Natl. Cancer Inst.*, 105(7):452–458, Apr. 2013. ISSN 0027-8874, 1460-2105. doi: 10.1093/jnci/djt007. URL http://dx.doi.org/10.1093/jnci/djt007. [15]

S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):333–351, May 2016. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg.2016.49. URL http://dx.doi.org/10.1038/nrg.2016.49. [5]

C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, Mar. 2007. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature05610. URL `http://dx.doi.org/10.1038/nature05610`. [3]

M. Griffith, C. A. Miller, O. L. Griffith, K. Krysiak, Z. L. Skidmore, A. Ramu, J. R. Walker, H. X. Dang, L. Trani, D. E. Larson, R. T. Demeter, M. C. Wendl, J. F. McMichael, R. E. Austin, V. Magrini, S. D. McGrath, A. Ly, S. Kulkarni, M. G. Cordes, C. C. Fronick, R. S. Fulton, C. A. Maher, L. Ding, J. M. Klco, E. R. Mardis, T. J. Ley, and R. K. Wilson. Optimizing cancer genome sequencing and analysis. *Cell Syst*, 1(3):210–223, Sept. 2015. ISSN 2405-4712. doi: 10.1016/j.cels.2015.08.015. URL `http://dx.doi.org/10.1016/j.cels.2015.08.015`. [5, 9]

M. Hagemann-Jensen, C. Ziegenhain, P. Chen, D. Ramsköld, G.-J. Hendriks, A. J. M. Larsson, O. R. Faridani, and R. Sandberg. Single-cell RNA counting at allele and isoform resolution using smart-seq3. *Nat. Biotechnol.*, 38(6):708–714, June 2020. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-020-0497-0. URL `http://dx.doi.org/10.1038/s41587-020-0497-0`. [12]

N. M. Hamad, J. H. Elconin, A. E. Karnoub, W. Bai, J. N. Rich, R. T. Abraham, C. J. Der, and C. M. Counter. Distinct requirements for ras oncogenesis in human versus mouse cells. *Genes Dev.*, 16(16):2045–2057, Aug. 2002. ISSN 0890-9369. doi: 10.1101/gad.993902. URL `http://dx.doi.org/10.1101/gad.993902`. [71]

D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan. 2000. ISSN 0092-8674. doi: 10.1016/S0092-8674(00)81683-9. URL `https://www.ncbi.nlm.nih.gov/pubmed/10647931`. [1]

D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, Mar. 2011. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2011.02.013. URL `http://dx.doi.org/10.1016/j.cell.2011.02.013`. [1]

P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira. Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, 10(8):551–564, Aug. 2009. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2593. URL `http://dx.doi.org/10.1038/nrg2593`. [7]

J. B. Hiatt, C. C. Pritchard, S. J. Salipante, B. J. O'Roak, and J. Shendure. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.*, 23(5):843–854, May 2013. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.147686.112. URL `http://dx.doi.org/10.1101/gr.147686.112`. [14, 15]

L. Hu, C. McArthur, and R. B. Jaffe. Ovarian cancer stem-like side-population cells are tumourigenic and chemoresistant. *Br. J. Cancer*, 102(8):1276–1283, Apr. 2010. ISSN 0007-0920, 1532-1827. doi: 10.1038/sj.bjc.6605626. URL `http://dx.doi.org/10.1038/sj.bjc.6605626`. [3]

Y. Huang, D. J. McCarthy, and O. Stegle. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.*, 20(1):273, Dec. 2019. ISSN 1465-6906. doi: 10.1186/s13059-019-1865-2. URL `http://dx.doi.org/10.1186/s13059-019-1865-2`. [15, 68]

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, Feb. 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-1969-6. URL http://dx.doi.org/10.1038/s41586-020-1969-6. [3]

S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, 11(2): 163–166, Feb. 2014. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2772. URL http://dx.doi.org/10.1038/nmeth.2772. [10, 11]

K. Jahn, J. Kuipers, and N. Beerenwinkel. Tree inference for single-cell data. *Genome Biol.*, 17: 86, May 2016. ISSN 1465-6906. doi: 10.1186/s13059-016-0936-x. URL http://dx.doi.org/10.1186/s13059-016-0936-x. [13]

Y. Jiang, Y. Qiu, A. J. Minn, and N. R. Zhang. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 113(37):E5528–37, Sept. 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1522203113. URL http://dx.doi.org/10.1073/pnas.1522203113. [9, 27, 78]

Y. Jiang, R. Wang, E. Urrutia, I. N. Anastopoulos, K. L. Nathanson, and N. R. Zhang. CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol.*, 19(1):202, Nov. 2018. ISSN 1465-6906. doi: 10.1186/s13059-018-1578-y. URL http://dx.doi.org/10.1186/s13059-018-1578-y. [7, 77]

J. A. Kennedy and F. Barabé. Investigating human leukemogenesis: from cell lines to in vivo models of human leukemia. *Leukemia*, 22(11):2029–2040, Nov. 2008. ISSN 0887-6924, 1476-5551. doi: 10.1038/leu.2008.206. URL http://dx.doi.org/10.1038/leu.2008.206. [15]

S. Kim, K. Scheffler, A. L. Halpern, M. A. Bekritsky, E. Noh, M. Källberg, X. Chen, Y. Kim, D. Beyter, P. Krusche, and C. T. Saunders. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, 15(8):591–594, Aug. 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0051-x. URL http://dx.doi.org/10.1038/s41592-018-0051-x. [6]

T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, 9 (1):72–74, Nov. 2011. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.1778. URL http://dx.doi.org/10.1038/nmeth.1778. [10]

J. M. Klco, D. H. Spencer, C. A. Miller, M. Griffith, T. L. Lamprecht, M. O'Laughlin, C. Fronick, V. Magrini, R. T. Demeter, R. S. Fulton, W. C. Eades, D. C. Link, T. A. Graubert, M. J. Walter, E. R. Mardis, J. F. Dipersio, R. K. Wilson, and T. J. Ley. Functional heterogeneity of genetically defined subclones in acute myeloid leukemia. *Cancer Cell*, 25(3):379–392, Mar. 2014. ISSN 1535-6108, 1878-3686. doi: 10.1016/j.ccr.2014.01.031. URL http://dx.doi.org/10.1016/j.ccr.2014.01.031. [1, 5, 9]

D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, Sept. 2009. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btp373. URL http://dx.doi.org/10.1093/bioinformatics/btp373. [6]

D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22(3):568–576, Mar. 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.129684.111. URL http://dx.doi.org/10.1101/gr.129684.111. [6]

V. Körber, J. Yang, P. Barah, Y. Wu, D. Stichel, Z. Gu, M. N. C. Fletcher, D. Jones, B. Hentschel, K. Lamszus, J. C. Tonn, G. Schackert, M. Sabel, J. Felsberg, A. Zacher, K. Kaulich, D. Hübschmann, C. Herold-Mende, A. von Deimling, M. Weller, B. Radlwimmer, M. Schlesner, G. Reifenberger, T. Höfer, and P. Lichter. Evolutionary trajectories of IDHWT glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer Cell*, 35(4):692–704.e12, Apr. 2019. ISSN 1535-6108, 1878-3686. doi: 10.1016/j.ccell.2019.02.007. URL `http://dx.doi.org/10.1016/j.ccell.2019.02.007`. [23, 78]

A. J. M. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, Jan. 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-018-0836-1. URL `https://doi.org/10.1038/s41586-018-0836-1`. [14]

S. Lauer and D. Gresham. An evolving view of copy number variants. *Curr. Genet.*, May 2019. ISSN 0172-8083, 1432-0983. doi: 10.1007/s00294-019-00980-0. URL `http://dx.doi.org/10.1007/s00294-019-00980-0`. [7]

M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, 9 (8):e1003118, Aug. 2013. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1003118. [77, 79]

E. Y. H. P. Lee and W. J. Muller. Oncogenes and tumor suppressor genes. *Cold Spring Harb. Perspect. Biol.*, 2(10):a003236, Oct. 2010. ISSN 1943-0264. doi: 10.1101/cshperspect.a003236. URL `http://dx.doi.org/10.1101/cshperspect.a003236`. [3]

M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536 (7616):285–291, Aug. 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature19057. URL `http://dx.doi.org/10.1038/nature19057`. [76]

A. Li, Y. Ma, M. Jin, S. Mason, R. L. Mort, K. Blyth, L. Larue, O. J. Sansom, and L. M. Machesky. Activated mutant NRas(Q61K) drives aberrant melanocyte signaling, survival, and invasiveness via a rac1-dependent mechanism. *J. Invest. Dermatol.*, 132(11):2610–2621, Nov. 2012. ISSN 0022-202X, 1523-1747. doi: 10.1038/jid.2012.186. URL `http://dx.doi.org/10.1038/jid.2012.186`. [72]

H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 2013. URL `http://arxiv.org/abs/1303.3997`. [6, 75]

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug. 2009. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btp352. URL `http://dx.doi.org/10.1093/bioinformatics/btp352`. [39, 75]

B. Liu, C. D. Morrison, C. S. Johnson, D. L. Trump, M. Qin, J. C. Conroy, J. Wang, and S. Liu. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget*, 4(11):1868–1881, Nov. 2013. ISSN 1949-2553. doi: 10.18632/oncotarget.1537. URL `http://dx.doi.org/10.18632/oncotarget.1537`. [3, 7]

F. Liu, Y. Zhang, L. Zhang, Z. Li, Q. Fang, R. Gao, and Z. Zhang. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biol.*, 20(1):242, Nov. 2019. ISSN 1465-6906. doi: 10.1186/s13059-019-1863-4. URL `http://dx.doi.org/10.1186/s13059-019-1863-4`. [12]

A. T. L. Lun, K. Bach, and J. C. Marioni. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.*, 17:75, Apr. 2016. ISSN 1465-6906. doi: 10.1186/s13059-016-0947-7. URL `http://dx.doi.org/10.1186/s13059-016-0947-7`. [60]

X. Ma, Y. Liu, Y. Liu, L. B. Alexandrov, M. N. Edmonson, C. Gawad, X. Zhou, Y. Li, M. C. Rusch, J. Easton, R. Huether, V. Gonzalez-Pena, M. R. Wilkinson, L. C. Hermida, S. Davis, E. Sioson, S. Pounds, X. Cao, R. E. Ries, Z. Wang, X. Chen, L. Dong, S. J. Diskin, M. A. Smith, J. M. Guidry Auvil, P. S. Meltzer, C. C. Lau, E. J. Perlman, J. M. Maris, S. Meshinchi, S. P. Hunger, D. S. Gerhard, and J. Zhang. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature*, 555(7696):371–376, Mar. 2018. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature25795. URL `http://dx.doi.org/10.1038/nature25795`. [23]

S. Malikic, K. Jahn, J. Kuipers, S. C. Sahinalp, and N. Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat. Commun.*, 10(1): 2750, June 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10737-5. URL `http://dx.doi.org/10.1038/s41467-019-10737-5`. [13]

M. Malumbres and M. Barbacid. RAS oncogenes: the first 30 years. *Nat. Rev. Cancer*, 3(6): 459–465, June 2003. ISSN 1474-175X. doi: 10.1038/nrc1097. URL `http://dx.doi.org/10.1038/nrc1097`. [71]

I. Martincorena, K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell. Universal patterns of selection in cancer and somatic tissues. *Cell*, 0(0), Oct. 2017. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2017.09.042. URL `http://www.cell.com/article/S0092867417311364/abstract`. [4, 103]

D. J. McCarthy, R. Rostom, Y. Huang, D. J. Kunz, P. Danecek, M. J. Bonder, T. Hagai, R. Lyu, HipSci Consortium, W. Wang, D. J. Gaffney, B. D. Simons, O. Stegle, and S. A. Teichmann. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nat. Methods*, 17(4):414–421, Apr. 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-020-0766-3. URL `http://dx.doi.org/10.1038/s41592-020-0766-3`. [13, 37, 42, 85]

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. Feb. 2018. URL `http://arxiv.org/abs/1802.03426`. [56]

K. H. Metzeler, T. Herold, M. Rothenberg-Thurley, S. Amler, M. C. Sauerland, D. Görlich, S. Schneider, N. P. Konstandin, A. Dufour, K. Bräundl, B. Ksienzyk, E. Zellmeier, L. Hartmann, P. A. Greif, M. Fiegl, M. Subklewe, S. K. Bohlander, U. Krug, A. Faldum, W. E. Berdel, B. Wörmann, T. Büchner, W. Hiddemann, J. Braess, K. Spiekermann, and AMLCG Study Group. Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood*, 128(5):686–698, Aug. 2016. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2016-01-693879. URL `http://dx.doi.org/10.1182/blood-2016-01-693879`. [5]

B. Milholland, X. Dong, L. Zhang, X. Hao, Y. Suh, and J. Vijg. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.*, 8:15183, May 2017. ISSN 2041-1723. doi: 10.1038/ncomms15183. URL `http://dx.doi.org/10.1038/ncomms15183`. [79]

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, and L. Ding. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.*, 10(8):e1003665, Aug. 2014. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1003665. [9, 20, 78]

M. Morgan, H. Pagès, V. Obenchain, and N. Hayden. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import, 2019. URL `http://bioconductor.org/packages/Rsamtools`. [79]

A. S. Nam, K.-T. Kim, R. Chaligne, F. Izzo, C. Ang, J. Taylor, R. M. Myers, G. Abu-Zeinah, R. Brand, N. D. Omans, A. Alonso, C. Sheridan, M. Mariani, X. Dai, E. Harrington, A. Pastore, J. R. Cubillos-Ruiz, W. Tam, R. Hoffman, R. Rabadan, J. M. Scandura, O. Abdel-Wahab, P. Smibert, and D. A. Landau. Somatic mutations and cell identity linked by genotyping of transcriptomes. *Nature*, 571(7765):355–360, July 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1367-0. URL `http://dx.doi.org/10.1038/s41586-019-1367-0`. [12, 13, 53, 54, 67]

S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. Van Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, A. Menzies, S. Martin, K. Leung, L. Chen, C. Leroy, M. Ramakrishna, R. Rance, K. W. Lau, L. J. Mudie, I. Varela, D. J. McBride, G. R. Bignell, S. L. Cooke, A. Shlien, J. Gamble, I. Whitmore, M. Maddison, P. S. Tarpey, H. R. Davies, E. Papaemmanuil, P. J. Stephens, S. McLaren, A. P. Butler, J. W. Teague, G. Jönsson, J. E. Garber, D. Silver, P. Miron, A. Fatima, S. Boyault, A. Langerød, A. Tutt, J. W. M. Martens, S. A. J. R. Aparicio, Å. Borg, A. V. Salomon, G. Thomas, A.-L. Børresen-Dale, A. L. Richardson, M. S. Neuberger, P. A. Futreal, P. J. Campbell, M. R. Stratton, and Breast Cancer Working Group of the International Cancer Genome Consortium. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, May 2012. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2012.04.024. URL `http://dx.doi.org/10.1016/j.cell.2012.04.024`. [3]

R. Noble. *ggmuller: Create Muller Plots of Evolutionary Dynamics*, 2019. URL `https://CRAN.R-project.org/package=ggmuller`. R package version 0.5.3. [24]

P. C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, Oct. 1976. ISSN 0036-8075. doi: 10.1126/science.959840. URL `https://www.ncbi.nlm.nih.gov/pubmed/959840`. [2]

E. N. Oliva, J. Franek, D. Patel, O. Zaidi, S. A. Nehme, and A. M. Almeida. The real-world incidence of relapse in acute myeloid leukemia (aml): A systematic literature review (slr). *Blood*, 132(Supplement 1):5188–5188, Nov. 2018. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2018-99-111839. URL `https://ashpublications.org/blood/article/132/Supplement%201/5188/265727/The-RealWorld-Incidence-of-Relapse-in-Acute`. [1]

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, and P. J. Campbell. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.*, 374(23):2209–2221, June 2016. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1516192. URL `http://dx.doi.org/10.1056/NEJMoa1516192`. [5, 70]

E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528, Feb. 2019. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/bty633. URL `http://dx.doi.org/10.1093/bioinformatics/bty633`. [85]

S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, and I. Hellmann. zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience*, May 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy059. [39, 76]

A. A. Petti, S. R. Williams, C. A. Miller, I. T. Fiddes, S. N. Srivatsan, D. Y. Chen, C. C. Fronick, R. S. Fulton, D. M. Church, and T. J. Ley. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat. Commun.*, 10(1):3660, Aug. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11591-1. URL `http://dx.doi.org/10.1038/s41467-019-11591-1`. [12, 37, 67, 68]

S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, 10(11):1096–1098, Nov. 2013. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2639. URL `http://dx.doi.org/10.1038/nmeth.2639`. [10]

O. Poirion, X. Zhu, T. Ching, and L. X. Garmire. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. *Nat. Commun.*, 9(1): 4892, Nov. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07170-5. URL `https://doi.org/10.1038/s41467-018-07170-5`. [12, 14]

R. Poplin, V. Ruano-Rubio, M. A. DePristo, T. J. Fennell, M. O. Carneiro, G. A. Van der Auwera, D. E. Kling, L. D. Gauthier, A. Levy-Moonshine, D. Roazen, K. Shakir, J. Thibault, S. Chandran, C. Whelan, M. Lek, S. Gabriel, M. J. Daly, B. Neale, D. G. MacArthur, and E. Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. Nov. 2017. URL `https://www.biorxiv.org/content/early/2017/11/14/201178`. [7]

C. Posch, M. Sanlorenzo, I. Vujic, J. A. Oses-Prieto, B. D. Cholewa, S. T. Kim, J. Ma, K. Lai, M. Zekhtser, R. Esteve-Puig, G. Green, S. Chand, A. L. Burlingame, R. Panzer-Grümayer, K. Rappersberger, and S. Ortiz-Urda. Phosphoproteomic analyses of NRAS(G12) and NRAS(Q61) mutant melanocytes reveal increased CK2α kinase levels in NRAS(Q61) mutant cells. *J. Invest. Dermatol.*, 136(10):2041–2048, Oct. 2016. ISSN 0022-202X, 1523-1747. doi: 10.1016/j.jid.2016.05.098. URL `http://dx.doi.org/10.1016/j.jid.2016.05.098`. [72]

R. Prieto-Bermejo, M. Romo-González, A. Pérez-Fernández, C. Ijurko, and Á. Hernández-Hernández. Reactive oxygen species in haematopoiesis: leukaemic cells take a walk on the wild side. *J. Exp. Clin. Cancer Res.*, 37(1):125, June 2018. ISSN 0392-9078, 1756-9966. doi: 10.1186/s13046-018-0797-0. URL `http://dx.doi.org/10.1186/s13046-018-0797-0`. [71]

D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, and R. Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, 30(8):777–782, Aug. 2012. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.2282. URL `http://dx.doi.org/10.1038/nbt.2282`. [10]

A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, and Human Cell Atlas Meeting Participants. The human cell atlas. *Elife*, 6, Dec. 2017. ISSN 2050-084X. doi: 10.7554/eLife.27041. URL `http://dx.doi.org/10.7554/eLife.27041`. [10]

B. Reinius, J. E. Mold, D. Ramsköld, Q. Deng, P. Johnsson, J. Michaëlsson, J. Frisén, and R. Sandberg. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.*, 48(11):1430–1435, Nov. 2016. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3678. URL `http://dx.doi.org/10.1038/ng.3678`. [14]

E. Reisinger, L. Genthner, J. Kerssemakers, P. Kensche, S. Borufka, A. Jugold, A. Kling, M. Prinz, I. Scholz, G. Zipprich, R. Eils, C. Lawerenz, and J. Eils. OTP: An automatized system for managing and processing NGS data. *J. Biotechnol.*, 261:53–62, Nov. 2017. ISSN 0168-1656, 1873-4863. doi: 10.1016/j.jbiotec.2017.08.006. URL `http://dx.doi.org/10.1016/j.jbiotec.2017.08.006`. [77]

E. M. Ross and F. Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.*, 17(1):69, 2016. ISSN 1465-6906, 1474-760X. doi: 10.1186/s13059-016-0929-9. URL `http://dx.doi.org/10.1186/s13059-016-0929-9`. [13, 45, 85]

A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, 11(4):396–398, Mar. 2014. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2883. URL `http://dx.doi.org/10.1038/nmeth.2883`. [9]

A. Roth, A. McPherson, E. Laks, J. Biele, D. Yap, A. Wan, M. A. Smith, C. B. Nielsen, J. N. McAlpine, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nat. Methods*, 13(7):573–576, July 2016. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3867. URL `http://dx.doi.org/10.1038/nmeth.3867`. [13]

A. J. Sabnis, L. S. Cheung, M. Dail, H. C. Kang, M. Santaguida, M. L. Hermiston, E. Passegué, K. Shannon, and B. S. Braun. Oncogenic kras initiates leukemia in hematopoietic stem cells. *PLoS Biol.*, 7(3):e59, Mar. 2009. ISSN 1544-9173, 1545-7885. doi: 10.1371/journal.pbio.1000059. URL `http://dx.doi.org/10.1371/journal.pbio.1000059`. [72]

S. Salehi, A. Steif, A. Roth, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. ddclone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biol.*, 18(1):44, Mar. 2017. ISSN 1465-6906. doi: 10.1186/s13059-017-1169-3. URL `http://dx.doi.org/10.1186/s13059-017-1169-3`. [13]

J. J. Salk, M. W. Schmitt, and L. A. Loeb. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.*, Mar. 2018. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg.2017.117. URL `http://dx.doi.org/10.1038/nrg.2017.117`. [14]

C. Sandén, H. Lilljebjörn, C. Orsmark Pietras, R. Henningsson, K. H. Saba, N. Landberg, H. Thorsson, S. von Palffy, P. Peña-Martinez, C. Högberg, M. Rissler, D. Gisselsson, V. Lazarevic, G. Juliusson, H. Ågerstam, and T. Fioretos. Clonal competition within complex evolutionary hierarchies shapes AML over time. *Nat. Commun.*, 11(1):579, Feb. 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-14106-0. URL `http://dx.doi.org/10.1038/s41467-019-14106-0`. [16, 17]

V. Sater, P.-J. Viailly, T. Lecroq, É. Prieur-Gaston, É. Bohers, M. Viennot, P. Ruminy, H. Dauchel, P. Vera, and F. Jardin. UMI-VarCal: a new UMI-based variant caller that efficiently improves low-frequency variant detection in paired-end sequencing NGS libraries. *Bioinformatics*, 36(9):2718–2724, May 2020. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btaa053. URL `http://dx.doi.org/10.1093/bioinformatics/btaa053`. [15, 68]

J. N. Saultz and R. Garzon. Acute myeloid leukemia: A concise review. *J. Clin. Med. Res.*, 5(3), Mar. 2016. ISSN 1918-3003, 2077-0383. doi: 10.3390/jcm5030033. URL `http://dx.doi.org/10.3390/jcm5030033`. [4]

C. T. Saunders, W. S. W. Wong, S. Swamy, J. Becq, L. J. Murray, and R. K. Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, July 2012. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/bts271. URL http://dx.doi.org/10.1093/bioinformatics/bts271. [6]

C. L. Sawyers, M. L. Gishizky, S. Quan, D. W. Golde, and O. N. Witte. Propagation of human blastic myeloid leukemias in the SCID mouse. *Blood*, 79(8):2089–2098, Apr. 1992. ISSN 0006-4971. URL https://www.ncbi.nlm.nih.gov/pubmed/1562735. [16]

R. Schwartz and A. A. Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.*, 18(4):213–229, Apr. 2017. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg.2016.170. URL http://dx.doi.org/10.1038/nrg.2016.170. [9, 15]

A. Serin Harmanci, A. O. Harmanci, and X. Zhou. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat. Commun.*, 11(1):89, Jan. 2020. ISSN 2041-1723. doi: 10.1038/s41467-019-13779-x. URL http://dx.doi.org/10.1038/s41467-019-13779-x. [12]

D. A. Shagin, I. A. Shagina, A. R. Zaretsky, E. V. Barsova, I. V. Kelmanson, S. Lukyanov, D. M. Chudakov, and M. Shugay. A high-throughput assay for quantitative measurement of PCR errors. *Sci. Rep.*, 7(1):2718, June 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-02727-8. URL http://dx.doi.org/10.1038/s41598-017-02727-8. [83]

L. I. Shlush, A. Mitchell, L. Heisler, S. Abelson, S. W. K. Ng, A. Trotman-Grant, J. J. F. Medeiros, A. Rao-Bhatia, I. Jaciw-Zurakowsky, R. Marke, J. L. McLeod, M. Doedens, G. Bader, V. Voisin, C. Xu, J. D. McPherson, T. J. Hudson, J. C. Y. Wang, M. D. Minden, and J. E. Dick. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature*, 547(7661):104–108, July 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature22993. URL http://dx.doi.org/10.1038/nature22993. [1, 4, 5, 70]

M. Shugay, O. V. Britanova, E. M. Merzlyak, M. A. Turchaninova, I. Z. Mamedov, T. R. Tuganbaev, D. A. Bolotin, D. B. Staroverov, E. V. Putintseva, K. Plevova, C. Linnemann, D. Shagin, S. Pospisilova, S. Lukyanov, T. N. Schumacher, and D. M. Chudakov. Towards error-free profiling of immune repertoires. *Nat. Methods*, 11(6):653–655, June 2014. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.2960. URL http://dx.doi.org/10.1038/nmeth.2960. [14, 15]

M. Shugay, A. R. Zaretsky, D. A. Shagin, I. A. Shagina, I. A. Volchenkov, A. A. Shelenkov, M. Y. Lebedin, D. V. Bagaev, S. Lukyanov, and D. M. Chudakov. MAGERI: Computational pipeline for molecular-barcoded targeted resequencing. *PLoS Comput. Biol.*, 13(5):e1005480, May 2017. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1005480. [14, 32, 40, 41, 68, 81, 83]

C. Shyr, M. Tarailo-Graovac, M. Gottlieb, J. J. Y. Lee, C. van Karnebeek, and W. W. Wasserman. FLAGS, frequently mutated genes in public exomes. *BMC Med. Genomics*, 7:64, Dec. 2014. ISSN 1755-8794. doi: 10.1186/s12920-014-0064-y. URL http://dx.doi.org/10.1186/s12920-014-0064-y. [70]

J. Singer, J. Kuipers, K. Jahn, and N. Beerenwinkel. Single-cell mutation identification via phylogenetic inference. *Nat. Commun.*, 9(1):5144, Dec. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07627-7. URL http://dx.doi.org/10.1038/s41467-018-07627-7. [12]

A. Ståhlberg, P. M. Krzyzanowski, M. Egyud, S. Filges, L. Stein, and T. E. Godfrey. Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. *Nat. Protoc.*, 12(4):664–682, Apr. 2017. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2017.006. URL http://dx.doi.org/10.1038/nprot.2017.006. [32, 76]

R. Stark, M. Grzelak, and J. Hadfield. RNA sequencing: the teenage years. *Nat. Rev. Genet.*, 20 (11):631–656, Nov. 2019. ISSN 1471-0056, 1471-0064. doi: 10.1038/s41576-019-0150-2. URL http://dx.doi.org/10.1038/s41576-019-0150-2. [6]

T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, 3rd, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of Single-Cell data. *Cell*, 177(7):1888–1902.e21, June 2019. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2019.05.031. URL http://dx.doi.org/10.1016/j.cell.2019.05.031. [56]

V. Svensson, R. Vento-Tormo, and S. A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.*, 13(4):599–604, Apr. 2018. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2017.149. URL http://dx.doi.org/10.1038/nprot.2017.149. [10]

I. Tirosh, A. S. Venteicher, C. Hebert, L. E. Escalante, A. P. Patel, K. Yizhak, J. M. Fisher, C. Rodman, C. Mount, M. G. Filbin, C. Neftel, N. Desai, J. Nyman, B. Izar, C. C. Luo, J. M. Francis, A. A. Patel, M. L. Onozato, N. Riggi, K. J. Livak, D. Gennert, R. Satija, B. V. Nahed, W. T. Curry, R. L. Martuza, R. Mylvaganam, A. J. Iafrate, M. P. Frosch, T. R. Golub, M. N. Rivera, G. Getz, O. Rozenblatt-Rosen, D. P. Cahill, M. Monje, B. E. Bernstein, D. N. Louis, A. Regev, and M. L. Suvà. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, 539(7628):309–313, Nov. 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature20123. URL http://dx.doi.org/10.1038/nature20123. [12]

M. A. Turchaninova, A. Davydov, O. V. Britanova, M. Shugay, V. Bikos, E. S. Egorov, V. I. Kirgizova, E. M. Merzlyak, D. B. Staroverov, D. A. Bolotin, I. Z. Mamedov, M. Izraelson, M. D. Logacheva, O. Kladova, K. Plevova, S. Pospisilova, and D. M. Chudakov. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat. Protoc.*, 11(9): 1599–1616, Sept. 2016. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2016.093. URL http://dx.doi.org/10.1038/nprot.2016.093. [14]

E. Urrutia, H. Chen, Z. Zhou, N. R. Zhang, and Y. Jiang. Integrative pipeline for profiling DNA copy number and inferring tumor phylogeny. *Bioinformatics*, 34(12):2126–2128, June 2018. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/bty057. URL http://dx.doi.org/10.1093/bioinformatics/bty057. [7, 26, 77]

G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, 43:11.10.1–33, 2013. ISSN 1934-3396, 1934-340X. doi: 10.1002/0471250953.bi1110s43. URL http://dx.doi.org/10.1002/0471250953.bi1110s43. [6, 8, 76]

P. van Galen, V. Hovestadt, M. H. Wadsworth, II, T. K. Hughes, G. K. Griffin, S. Battaglia, J. A. Verga, J. Stephansky, T. J. Pastika, J. L. Story, G. S. Pinkus, O. Pozdnyakova, I. Galinsky, R. M. Stone, T. A. Graubert, A. K. Shalek, J. C. Aster, A. A. Lane, and B. E. Bernstein. Single-Cell RNA-Seq reveals AML hierarchies relevant to disease progression and immunity. *Cell*, 0(0), Feb. 2019. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2019.01.031. URL http://www.cell.com/article/S0092867419300947/abstract. [67]

L. Velten, B. A. Story, P. Hernandez-Malmierca, J. Milbank, M. Paulsen, C. Lutz, D. Nowak, J.-C. Jann, C. Pabst, T. Boch, W.-K. Hofmann, C. Mueller-Tidow, S. Raffel, A. Trumpp, S. Haas, and L. M. Steinmetz. MutaSeq reveals the transcriptomic consequences of clonal evolution in acute myeloid leukemia. Dec. 2018. URL https://www.biorxiv.org/content/early/2018/12/21/500108. [13, 67]

D. Verma, H. Kantarjian, S. Faderl, S. O'Brien, S. Pierce, K. Vu, E. Freireich, M. Keating, J. Cortes, and F. Ravandi. Late relapses in acute myeloid leukemia: analysis of characteristics and outcome. *Leuk. Lymphoma*, 51(5):778–782, May 2010. ISSN 1042-8194. doi: 10.3109/10428191003661852. URL http://dx.doi.org/10.3109/10428191003661852. [1]

B. Vick, M. Rothenberg, N. Sandhöfer, M. Carlet, C. Finkenzeller, C. Krupka, M. Grunert, A. Trumpp, S. Corbacioglu, M. Ebinger, M. C. André, W. Hiddemann, S. Schneider, M. Subklewe, K. H. Metzeler, K. Spiekermann, and I. Jeremias. An advanced preclinical mouse model for acute myeloid leukemia using patients' cells of various genetic subgroups and in vivo bioluminescence imaging. *PLoS One*, 10(3):e0120925, Mar. 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0120925. URL `http://dx.doi.org/10.1371/journal.pone.0120925`. [xiii, 16, 18, 75]

K. Wang, M. Sanchez-Martin, X. Wang, K. M. Knapp, R. Koche, L. Vu, M. K. Nahas, J. He, M. Hadler, E. M. Stein, M. S. Tallman, A. L. Donahue, G. M. Frampton, D. Lipson, S. Roels, P. J. Stephens, E. M. Sanford, T. Brennan, G. A. Otto, R. Yelensky, V. A. Miller, M. G. Kharas, R. L. Levine, A. Ferrando, S. A. Armstrong, and A. V. Krivtsov. Patient-derived xenotransplants can recapitulate the genetic driver landscape of acute leukemias. *Leukemia*, 31 (1):151–158, Jan. 2017. ISSN 0887-6924, 1476-5551. doi: 10.1038/leu.2016.166. URL `http://dx.doi.org/10.1038/leu.2016.166`. [16]

J. S. Welch, T. J. Ley, D. C. Link, C. A. Miller, D. E. Larson, D. C. Koboldt, L. D. Wartman, T. L. Lamprecht, F. Liu, J. Xia, C. Kandoth, R. S. Fulton, M. D. McLellan, D. J. Dooling, J. W. Wallis, K. Chen, C. C. Harris, H. K. Schmidt, J. M. Kalicki-Veizer, C. Lu, Q. Zhang, L. Lin, M. D. O'Laughlin, J. F. McMichael, K. D. Delehaunty, L. A. Fulton, V. J. Magrini, S. D. McGrath, R. T. Demeter, T. L. Vickery, J. Hundal, L. L. Cook, G. W. Swift, J. P. Reed, P. A. Alldredge, T. N. Wylie, J. R. Walker, M. A. Watson, S. E. Heath, W. D. Shannon, N. Varghese, R. Nagarajan, J. E. Payton, J. D. Baty, S. Kulkarni, J. M. Klco, M. H. Tomasson, P. Westervelt, M. J. Walter, T. A. Graubert, J. F. DiPersio, L. Ding, E. R. Mardis, and R. K. Wilson. The origin and evolution of mutations in acute myeloid leukemia. *Cell*, 150(2):264–278, July 2012. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2012.06.023. URL `http://dx.doi.org/10.1016/j.cell.2012.06.023`. [4, 71]

M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva. Identification of neutral tumor evolution across cancer types. *Nat. Genet.*, 48(3):238–244, Mar. 2016. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3489. URL `http://dx.doi.org/10.1038/ng.3489`. [4]

World Health Organization. Cancer. `https://www.who.int/en/news-room/fact-sheets/detail/cancer`, Apr. 2020. URL `https://www.who.int/en/news-room/fact-sheets/detail/cancer`. Accessed: 2020-4-27. [1]

C.-I. Wu, H.-Y. Wang, S. Ling, and X. Lu. The ecology and evolution of cancer: The Ultra-Microevolutionary process. *Annu. Rev. Genet.*, 50:347–369, Nov. 2016. ISSN 0066-4197, 1545-2948. doi: 10.1146/annurev-genet-112414-054842. URL `http://dx.doi.org/10.1146/annurev-genet-112414-054842`. [4]

M. Yanada, G. Garcia-Manero, G. Borthakur, F. Ravandi, H. Kantarjian, and E. Estey. Relapse and death during first remission in acute myeloid leukemia. *Haematologica*, 93(4):633–634, Apr. 2008. ISSN 0390-6078, 1592-8721. doi: 10.3324/haematol.12366. URL `http://dx.doi.org/10.3324/haematol.12366`. [1]

L. R. Yates and P. J. Campbell. Evolution of the cancer genome. *Nat. Rev. Genet.*, 13(11): 795–806, Nov. 2012. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3317. URL `http://dx.doi.org/10.1038/nrg3317`. [1]

G. Yu and Q.-Y. He. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.*, 12(2):477–479, Feb. 2016. ISSN 1742-206X, 1742-2051. doi: 10.1039/c5mb00663e. URL `http://dx.doi.org/10.1039/c5mb00663e`. [78]

H. Zafar, Y. Wang, L. Nakhleh, N. Navin, and K. Chen. Monovar: single-nucleotide variant detection in single cells. *Nat. Methods*, 13(6):505–507, June 2016. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3835. URL `http://dx.doi.org/10.1038/nmeth.3835`. [12]

H. Zafar, A. Tzen, N. Navin, K. Chen, and L. Nakhleh. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol.*, 18(1):178, Sept. 2017. ISSN 1465-6906. doi: 10.1186/s13059-017-1311-2. URL `http://dx.doi.org/10.1186/s13059-017-1311-2`. [13]

H. Zafar, N. Navin, K. Chen, and L. Nakhleh. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res.*, 29(11):1847–1859, Nov. 2019. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.243121.118. URL `http://dx.doi.org/10.1101/gr.243121.118`. [13]

H. Zare, J. Wang, A. Hu, K. Weber, J. Smith, D. Nickerson, C. Song, D. Witten, C. A. Blau, and W. S. Noble. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.*, 10(7):e1003703, July 2014. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1003703. URL `http://dx.doi.org/10.1371/journal.pcbi.1003703`. [9]

G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, 8:14049, Jan. 2017. ISSN 2041-1723. doi: 10.1038/ncomms14049. [39]

Z. Zhou, B. Xu, A. Minn, and N. R. Zhang. DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing. *Genome Biol.*, 21(1):10, Jan. 2020. ISSN 1465-6906. doi: 10.1186/s13059-019-1922-x. URL `http://dx.doi.org/10.1186/s13059-019-1922-x`. [12, 14]

C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard. Comparative analysis of Single-Cell RNA sequencing methods. *Mol. Cell*, 65(4):631–643.e4, Feb. 2017. ISSN 1097-2765, 1097-4164. doi: 10.1016/j.molcel.2017.01.023. URL `http://dx.doi.org/10.1016/j.molcel.2017.01.023`. [10]

List of Figures

List of Tables

Acknowledgements

This whole journey was only possible due to the wonderful people that gifted me with all their wisdom and support. First and foremost, I would like to express my deep gratitude to Ines Hellmann. You have not only been a great and incredibly patient mentor, but also the exact kind of person I've always pictured as the role model of a scientist. Thank you for giving me this incredible chance. Second, I would really like to thank Wolfgang Enard for always giving me a broad and creative perspective on every project, and for always sharing your enthusiasm. Both of you have managed to create the best working atmosphere ever, and it has been a privilege to see how you approach science.

I would like to thank the members of my thesis advisory committee who helped me shape many aspects of my PhD project, and even got me thinking about my career path as a whole: Dirk Metzler, Oliver Weigert, and Christiane Fuchs. I am also truly indebted to our collaborators within the SFB who made the realization of the AML long-term project possible: Binje Vick, Irmela Jeremias, Klaus Metzeler, Maja Rothenberg-Thurley, Karsten Spiekermann, Sebastian Vosberg, Philipp Greif. A special thank you to Alex Graf for taking great care of the server and helping me install all this crazy code!

All the people in the Enard lab, past and present, have made this whole experience even more joyful and exciting than Europapark. My first thank you goes to Maria, because we started this whole variant calling adventure together, and it's been marvelous to *not* exist with you ever since. Thanks for making me feel dahoam right away. Next big DANK goes to Johannes, Daniel, and Simon, for always producing so much amazing data, for trying out *umivariants* and making it grow, and for all the great tunes and lols. Big thank you to Beate for being our statistical savior and for the many geeky laughs; to Bria for being the best buddy in AML drama and heavy-metal fun; to Swati for all the guidance in bioinformatics and in life; to Christoph for your scRNA-seq knowledge and for all the rides I hitched; to Philipp for the fantastic analysis of the LT single cells, and the tips to slay at darts; and to Lu for all the fun we had with scTAG-seq, cellSNP, and being latino. Big thanks to Ines B for always cheering me up with your beautiful smile and all the kitty pictures. I am hugely grateful to Chris for his terrific work with the mutational signatures, and for being just so amazingly nice; to Charlotte for setting up the benchmarks of the UMI consensus, overcoming frustrations with MAGERI, and a lot of happy moments; and to Vroni for working on the simulation of scRNA-seq data. Thanks to Johanna, Zane, Jessy, Aleks, Mari, Zeynep, Erdem, for all the singing, dancing, cake indulging, and the many fascinating things in biology and beyond that I learned from you. Gracias a Karin for the language exchange and health advice. Thanks to the pretty ladies that keep the happiness levels of the lab high and steady: Elly, Daisy, Leyla,

Freya, and of course, Anita. To anyone I forgot: thanks for making this such a cool time!

I also want to thank the wonderful people from the SFB/IRTG 1243 that made it possible to have such a rich and smooth grad school life. Many thanks to Elizabeth Schroeder-Reiter for organizing all of the fantastic seminars, courses, and retreats; and for caring so much about our well-being. Big thanks as well to Elke Hammerbacher for saving my contract and helping me deal with bureaucracy so many times! Thanks to all the fellow PhDs who were great friends and collaborators: Christina, Julia, Will, Martina, Konrad...

I am of course always grateful to my loving family on both sides of the Atlantic: to my mom, Ku (we are the Hootsforce!), Laura, and all my other blood relatives who gave me everything to be able to embark in this odyssey. To the family that adopted us when we arrived in Germany for the first time: Sigrid, Harald, Werita. To my dearest Cosa, Mariel, Evil Clan, Sucias, Machaca Gang, Summer Breeze Crew, and all the mates with whom I shared school and/or home: I always keep you in my black heart.

Finally, nothing would be worthwhile without those who make a house a true and proper *madriguera*. Juan, my Lebencito, this world would be pointless without your existence. One piece of my soul is forever with you. Puchi, you are our peace and calm; the softest, warmest bundle of love that has ever existed. Gordi, your heart was the biggest and purest in this multiverse, and I miss you every second. This one's for you.