From the Institut für Medizinische Informationsverarbeitung Biometrie und Epidemiologie (IBE) of the Ludwig-Maximilians-Universität (LMU) München
Director: Prof. Dr. Ulrich Mansmann

Development of an improved variant calling pipeline and the analysis of altered allelic transcription of recurrent mutations in Acute Myeloid Leukaemia

Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.) an der Medizinischen Fakultät der Ludwig-Maximilians-Universität München

submitted by
Aarif Mohamed Nazeer Batcha
from
Ramanathapuram, India
on
29.06.2018

Supervisor:	Prof. Dr. Ulrich Mansmann
Second evaluator:	Prof. Dr Konstantin Strauch
Co-supervisor:	PD Dr. med. Tobias Herold
Dean:	Prof. Dr. Reinhard Hickel
Date of oral defence:	17.12.2018

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Dean's Office
Medical Faculty

Affidavit

Nazeer Batcha, Aarif Mohamed

Surname, first name

Street

Zip code, town

Country

I hereby declare, that the submitted thesis entitled

Development of an improved variant calling pipeline and the analysis of altered allelic transcription of recurrent mutations in Acute Myeloid Leukaemia

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

Munich, 29.06.2018

Place, date

Aarif Mohamed Nazeer Batcha

Signature doctoral candidate

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Dean's Office
Medical Faculty

Confirmation of congruency between printed and electronic version of the doctoral thesis

Nazeer Batcha, Aarif Mohamed

Surname, first name

Street

Zip code, town

Country

I hereby declare that the electronic version of the submitted thesis, entitled

Development of an improved variant calling pipeline and the analysis of altered allelic transcription of recurrent mutations in Acute Myeloid Leukaemia

is congruent with the printed version both in content and format.

Munich, 29.06.2018

Place, date

Aarif Mohamed Nazeer Batcha

Signature doctoral candidate

Table of contents

List of abbreviations

AI	allelic imbalance
ABL	ABL proto-oncogene 1, non-receptor tyrosine kinase
AML	acute myeloid leukaemia
AMLCG	acute myeloid leukaemia co-operative group
ASE	allele-specific expression
ASXL1	ASXL transcriptional regulator 1
BCR	BCR, RhoGEF and GTPase activating protein
BCOR	BCL6 corepressor
bp	base pair
BWA	Burrows-Wheeler alignment tool
cDNA	complementary DNA
CDS	coding sequence of a gene
CEBPA	CCAAT enhancer binding protein alpha
CN-AML	cytogenetically normal acute myeloid leukaemia
CRE	cis-regulatory elements
CREST	clipping reveals structure
CRT	cyclic reversible termination
DIRAS3	DIRAS family GTPase 3
DKTK	deutsches konsortium für translationale krebsforschung
DNA	deoxyribonucleic acid
DNA-Seq	DNA sequencing
DNMT3A	DNA methyltransferase 3 alpha
ELN	European LeukemiaNet
EZH2	enhancer of zeste 2 polycomb repressive complex 2 subunit
FAB	French-American-British
FLT3	fms related tyrosine kinase 3
GATA2	GATA binding protein 2
GATK	genome analysis tool kit
$Genes^{MUT}$	gene with recurrent mutations
$Genes^{WT}$	genes without recurrent mutations
GTF	gene transfer format
HISAT	hierarchical indexing for spliced alignemtn of transcripts
IDH2	isocitrate dehydrogenase (NADP(+)) 2, mitochondrial
INDEL	insertion and deletion
inv	inversion
MAQ	mapping and assembly with qualities
MGE	mobile genetic element
MOM	maximum oligonucleotide mapping
NGS	next generation sequencing
NPM1	nucleophosmin 1
PHF6	PHD finger protein 6
PML	promyelocytic leukemia
PTPN11	protein tyrosine phosphatase, non-receptor type 11
RADAR	rigorously annotated database of A-to-I RNA editing
RARA	retinoic acid receptor alpha
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
RUNX1	runt related transcription factor 1
RUNX1T1	RUNX1 translocation partner 1
SAMtools	sequence alignment/map tools

SBL	seqeunce by ligation
SBS	sequence by synthesis
SF3B1	splicing factor 3b subunit 1
SMRT	single-molecule real-time seqeuncing
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SOAP	short oligonucleotide analysis package
SRSF2	serine and arginine rich splicing factor 2
STAG2	stromal antigen 2
STAR	spliced transcripts alignment to a reference
SV	structural variant
t	translocation
TCGA	the cancer genome atlas
TET2	tet methylcytosine dioxygenase 2
U2AF1	U2 small nuclear RNA auxiliary factor 1
v	version
VAF	variant allele frequency
WAI	weighted allelic imbalance
WES	whole exome sequencing
WGS	whole genome sequencing
WHO	world health organization
WT1	Wilms tumor 1
ZRSR2	zinc finger CCCH-type, RNA binding motif and serine/arginine rich 2

List of figures

List of tables

List of boxes

Abstract

Recurrent mutated genes in acute myeloid leukaemia are suspected to contribute to leukaemogenesis by different mechanisms but the ratios in which the recurrently mutated alleles are transcribed from DNA to RNA in the respective genes are widely unknown. A systematic comparison of variant allele frequencies of recurrent mutated genes was carried out using a large AML cohort (N=499). Around 95% of variants were detected to be transcribed from DNA to RNA by the application of a minimum read depth cut-off of 10x (90% transcribed among recurrent mutations). The analysis on 11 recurrently mutated genes in AML determined preferential mutant allele transcript abundance for *GATA2*^{MUT}, *RUNX1*^{MUT}, *TET2*^{MUT}, *SRSF2*^{MUT}, *IDH2*^{MUT} and *NPM1*^{MUT} and preferential wild-type transcript abundance for *PTPN11*^{MUT}, *CEBPA*^{MUT} and *WT1*^{MUT}, respectively. Presence of allelic imbalances among the common variants of *GATA2*^{WT}, *RUNX1*^{WT} and *IDH2*^{WT} were also demonstrated in patients without recurrent mutations in the respective genes. Further inquiry based on the differential expression of genes and transcript isoforms between patients with and without recurrent mutations in the respective genes showed no significant difference except for *SRSF2*, *CEBPA* and *WT1*. In summary, this study compared the variant allele frequencies of recurrently mutated genes and exhibits allele-specific transcript abundance of these genes in AML. The observed differences can be interpreted as a novel, currently underestimated mechanism how mutations contribute to leukaemogenesis and necessitate further analysis.

1 Introduction

Advancements in next generation sequencing (NGS) technologies have allowed us to study the heterogeneous and complex alterations in cancer genomes.[1,2] Although several optimized variant calling procedures and best practice guidelines are available for processing DNA sequencing (DNA-Seq), it still remains a challenge in the case of RNA sequencing (RNA-Seq).[3–5] The somatic sequence alterations obtained from variant calling procedures in the cancer genome are thought to disturb physiological protein function or gene expression, but the extent of such mutations transcribed from DNA to RNA are largely unknown.[6] Few studies have examined the variant allele frequency correlation between DNA and RNA and investigated the allelic imbalance among the somatic mutations in cancer genomes.[7–9] This study aims to improve the existing variant calling pipelines for targeted DNA-Seq and RNA-Seq and examine the differences in allelic proportions of recurrently mutated genes in acute myeloid leukaemia (AML).

1.1 DNA and RNA sequencing

Deoxyribonucleic acid (DNA) is the hereditary material in humans and most of the living organisms. The hereditary information is stored in the DNA sequences which consist of four nucleotide bases: adenine, guanine, thymine and cytosine, being chained by phosphate-deoxyribose backbones. Chromosomes are linear arrangement of condensed form of DNA molecules and histone proteins and the organisms containing two complete sets of homologous chromosomes in their cells are called diploid organisms (e.g.: Humans). Some of the DNA sequence fragments which code for protein molecules are termed as genes. A gene is the basic physical and functional unit of heredity. The variant form of any given gene is termed as allele. Diploid organisms with two identical alleles of the gene are known as homozygous at a gene locus, whereas those with two different alleles are termed as heterozygous. The process of determining the order of the arrangement of nucleotide bases in a DNA strand is called DNA sequencing. The sequence of DNA fragments contain the majority of genetic information necessary for life and thus any knowledge of it is helpful for the fundamental researches in biology. The Sanger sequencing technique is one of the initial sequencing methods for determining longer DNA molecules, proposed by Frederick Sanger (1977) using

chain termination inhibitors.[10] The improvement and implication of this technique revolutionized the field of biological research and led to the development of NGS technologies (also known as high-throughput or second generation sequencing) at the beginning of the millennium. In the last decade, rapid advancements in NGS machineries introduced sequencing techniques such as sequencing by synthesis (SBS), sequencing by ligation (SBL), cyclic reversible termination (CRT) and single-molecule real-time sequencing (SMRT).[11,12] Based on these techniques, different types of NGS technologies have been developed and were commercially offered in the form of various platforms: Illumna (Solexa) sequencing, Roche 454 sequencing, Ion torrent: Proton/PGM sequencing and ABI SOLiD sequencing. The initial process of any NGS workflow is the template preparation (amplification), followed by sequencing and imaging.[11] Widely used NGS methods in DNA-Seq are whole-genome sequencing (WGS), whole exome sequencing (WES), targeted sequencing and *de novo* sequencing. WGS consists of sequencing the entire genome of the species of interest (e.g. 3.2 billion bases in the human genome), whereas WES involves sequencing a selective capture of the known protein-coding regions of an entire genome (less than 2% in human genome). Thus WES is a cost-effective approach when compared to WGS with the drawback of losing non-coding but relevant regions, e.g. those with regulatory functions.[13] Targeted sequencing is another alternative for WGS and WES, when the research focus is highly restricted to specific regions of interest. Due to the short range of the areas of interest, typical targeted sequencing provides more than 10 times the confidence obtained by WGS (in terms of 'coverage'). This in turn provides greater opportunities for the researchers to identify and analyse variations in sequences with a much higher accuracy. *De novo* sequencing refers to sequencing a new genome without any reference sequences available for alignment. This involves sequencing reads with different fragment lengths ('insert sizes') and assembled into a set of overlapping segments which in turn represent a consensus region of DNA (sequence contigs).

Ribonucleic acid (RNA) are single-stranded polymeric molecules, which are transcribed from DNA and are involved in several biological processes such as gene expression and regulation. RNA also comprises of four nucleotide bases: adenine, guanine, cytosine and uracil. Similar to DNA-Seq, RNA-Seq is the process of determining the arrangement of nucleotides within RNA. The library

preparation of RNA involves additional steps in which the RNA fragments are chemically labelled and are reverse-transcribed to complementary DNA (cDNA) fragments. This is followed by library amplification and sequencing, similar to that of the DNA-Seq. Different RNA-Seq methods in NGS include total RNA sequencing, mRNA sequencing, targeted RNA sequencing, small RNA and non-coding RNA sequencing. Researchers employ different sequencing methods based on the scientific question, sample type, read length, accuracy, time, cost, required read coverage and quality of the sequence data.[11–13]

1.2 Allelic imbalance

Transcription is a process in which the information encoded within the sequence of DNA (genes) are copied or transcribed into strands of RNA. The transcription is regulated by transcription factors, enhancers and other proteins through a variety of mechanisms.[14] Thus, the differences in gene regulation may be due to *cis*-regulatory or *trans*-regulatory changes. In a diploid cell, two allelic copies of each gene are present and thus the transcribed RNA strands can carry information from both alleles. However, the rate of expression of these two alleles must not necessarily be the same and can result in allelic imbalance (AI).[15] This phenomenon might be due to allele-specific expression, stability of the transcripts, copy number alterations or uniparental disomy.

Allele-specific expression (ASE) refers to the difference in the transcript abundance of the two allelic copies in a diploid organism. The ASE might be influenced by the variations in the enhancers or *cis*-regulatory elements (CRE), a non-coding DNA sequence with multiple activator and repressor binding sites for transcription regulation. The allele-specific differences in gene expression are also associated with the epigenetic phenomena of genomic imprinting and X-chromosome inactivation. Epigenetic phenomena refer to heritable changes in gene expression without any actual sequence alterations in the coding regions of DNA.[16] In the case of the X-chromosome, most of the genes on one copy of it are silenced in female mammals resulting in X-chromosome inactivation.[17] The whole inactivation of one copy of a gene lead to mono-allelic expression. Many other autosomal genes also show random choice of mono-allelic expression between maternal and paternal alleles due to genomic imprinting.[18,19] For example, the

DIRAS3 gene in humans is a paternally expressed gene which is located on chromosome 1 (maternally imprinted). This gene acts as a tumour suppressor gene and a loss of function of this gene is observed in ovarian and breast cancers.[20] In this case, the gene will not be expressed if an individual received both copies of chromosome 1 from the mother (uniparental disomy), which results in an increased risk of breast and ovarian cancer.[21] Other cause of AI might be due to the degradation of allele-specific transcript isoforms. RNA decay plays a major role in the process of gene expression and post-transcriptional regulation.[22,23] The presence of somatic mutations might affect the half-life of the mRNA transcripts and thus result in the premature decay of allele-specific transcript isoforms.

1.3 Variations in the sequences

Sequence variation on the DNA level refers to any genomic alterations in relation to the reference sequence (**Figure 1**).[24] The sequence variations are considered as either mutations (disease-causing change) or polymorphisms (non-disease-causing change). The most common form of sequence variations are single nucleotide variants (SNVs) and short insertions and deletions (INDELs).[25–27] The term SNV most often refers to a single nucleotide substitution of bases in a sequence when compared with the reference genome, irrespective of the frequency of its occurrence in a population. Any single nucleotide alteration which is observed in more than 1% of the population is considered as single nucleotide polymorphisms (SNPs).

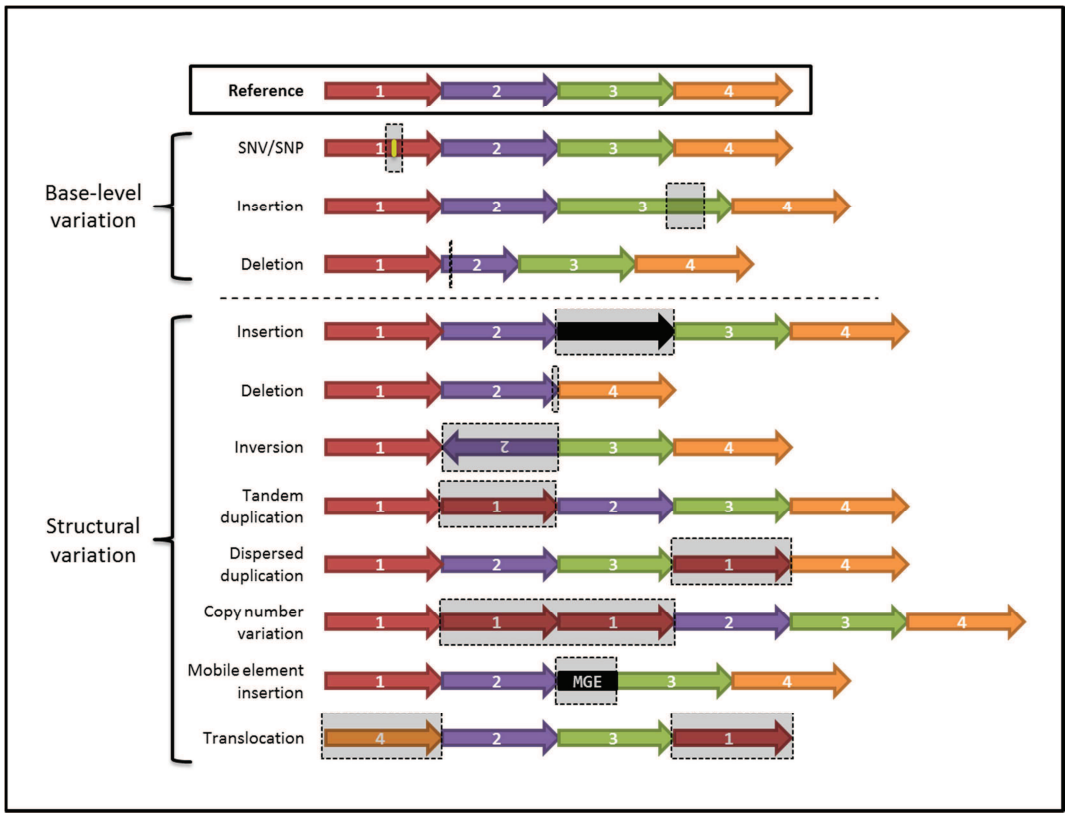

Figure 1: *Types of base-level and structural variations in DNA. The dashed line separates the base-level variants (occurs few bps) and structural variants (1Kbp or more).*[28] *The numbers indicate the order of DNA fragments. Yellow bar indicates single nucleotide substitution. Black bars represent insertion of sequence fragments. MGS represents mobile genetic elements like transposons and bacteriophage elements. Figure modified from Rahim et al.*[27]

Sequence variations comprise of insertions or deletion of one or more nucleotide bases are termed as INDELs. The relative frequency of such variants when compared with the reference genome is termed as variant allele frequency (VAFs). In the case of SNPs, the major and minor alleles refer to the first and second most common alleles occurring in a given population, respectively. These variants aid in understanding the pathogenesis of a disease and can also be used as a genetic markers.[29,30] They also help in understanding the structural and functional aspects of the protein biosynthesis.[31,32] Several annotation databases, such as dbSNP, COSMIC, ClinVar etc. have been developed. These databases include the information of SNP frequencies of existence in different population subsets as well as its prognostic information and clinical relevance.[33–35] Other sequence variations such as inversions, tandem duplications, dispersed duplications, copy number variations, insertions of mobile genetic elements (such as retro-transposons, bacteriophage elements (Mu) etc.), and translocations are termed as structural variants (**Figure 1**).[28,36–39] Effective identification of such genomic variants might

assist in investigating and inferring the associations between genotype and phenotype.

As mentioned before, the three major processes of variant detection are sequence alignment, variant calling and variant filtering. Thus, selecting and optimizing a sequence aligner and a variant caller followed by defining adequate filtering criteria are essential for detecting true variants and removing potential artefacts. The sequence variants are detected by aligning the digital sequence information generated by NGS machinery to the reference genome build. The reference genome to be used to align the read sequences is to be indexed using hash-based tables or Burrows-Weeler transform depending on the aligner algorithms along with the implementation of Ferragina – Manzini index.[40,41]

Some aligners are designed for short-read mapping (oligonucleotide fragments) such as SOAP, CUSHAW, MOM etc. whereas BWA-MEM, Bowtie2, CUSHAW3 etc. are optimized for longer read lengths (≥ 100bps).[42–47] Although these aligners perform relatively well in mapping DNA-Seq to the reference genome, they suffer from huge loss in alignment rates when mapping RNA-Seq, due to the innate complexity of the transcriptome sequencing.[48] RNA-Seq reads are not a continuous copy of genes which are transcribed from DNA. Since the non-contiguous exons are spliced together to form mature mRNA transcripts, the alignment of these sequence reads should account for such splice sites. This is achieved by the aligners by discovering splice junctions from the read coverage or by providing known exon-intron junctions, externally. Aligners such as TopHat, HISAT and STAR are optimized splice-aware aligners which are widely used for processing RNA-Seq.[49–51] Most of the sequence aligners are developed with multiple optimization parameters for addressing sensitivity, accuracy, memory, speed, alignment and mismatch rate. Selection of an optimal read aligner depends on the sequence read length, available hardware resources, alignment speed and time. An alternative of aligning sequence reads to the reference genome is to assemble short nucleotide sequences by using sequence overlaps (*de novo* assembly), but this is outside the focus of this study.[11,52]

The aligned reads are then processed to call the variations in the sample sequence when compared to the reference genome. The detection of such variants depends

on accurate and precise identification of differences between reference genome and the aligned sequence reads and any form of bias might lead to erroneous results. The aligned reads tend to have systematic or random artefacts due to experimental or technology-specific errors and thus it is subjected to post-processing to mitigate these artefacts.[53–55] One of the main steps in the post-processing is the removal of duplicate reads in order to reduce the effect of PCR amplification bias, introduced during library preparation. The extent of the duplicate read removal depends on the depth and type of library sequenced and thus its accuracy mainly relies upon the error rate of the libraries prepared.[56] Several tool programs such as Picard, SAMtools, SEAL, FastUniq etc. process the aligned reads to remove or mark PCR duplicates.[54,55,57,58] Other post-processing steps include adding read group information, re-order aligned reads and sorting them in accordance with chromosomal position, name etc. and performing local re-alignment of INDELs in order to increase the detection rate and accuracy.[54,55,59,60] This is followed by the variant calling procedure. Many variant calling algorithms have been developed for a fast and accurate detection of the sequence variants. Most commonly used variant callers are VarScan, Genome Analysis Tool Kit (GATK) – Haplotype Caller, FreeBayes and SAMtools.[55,60–62] Other variant callers including PINDEL, BreakDancer and CREST are used for calling larger structural variants.[63–65] The focus of this study relies on SNVs and short INDELs

1.4 Acute myeloid leukaemia

Acute myeloid leukaemia (AML) is one of the most common acute blood cancer diseases in adults in which abnormal myeloblasts are found in blood and bone marrow. AML has an incidence rate of 3 to 4 cases per 100,000 individuals per year and is more frequent in older people.[66,67] It is a prognostic heterogeneous disease with characteristic set of cytogenetic abnormalities and somatic mutations.[68,69] The sequence of leukaemogenic events resulting in the transformation of normal haematopoietic stem cells into leukaemic blasts are reviewed by Horton *et al.*[70] The current treatment of AML mainly rely on chemotherapy and did not change substantially in the last 30 years.[71,72] The standard initial treatment of AML is the conventional 7+3 regimen, which consists of standard dose of cytarabine and anthracycline antibiotic (daunorubicin) for seven and three days, respectively.[72] Patients receiving such induction chemotherapy go to remission with no signs of

symptoms or disease, but this depends largely on the prognostic factors such as age and genetic abnormalities. With the current regimen, about 50% of patients can be cured but the majority of older patients have a very unfavourable outcome. They eventually relapse and become non-responsive to further treatment. The identification of distinct genetic alterations has already resulted in the development of more targeted treatment approaches (e.g. *FLT3*-inhibitors) that are currently entering routine treatment.[73,74] In recent years, other targeted therapies using immune checkpoint inhibitors, monoclonal or bi-specific T-cell engager antibodies, metabolic and pro-apoptotic agents are being heavily investigated.[74–76] These novel substances will hopefully help to improve the current unsatisfying results achieved with chemotherapy.

1.4.1 Classification of acute myeloid leukaemia

In 1976 and 1985, the French-American-British (FAB) co-operative group classified AML into eight subtypes (M0 – M7), whereas the identification of specific genetic aberrations, biological and clinical features led to a new classification scheme by the world health organization (WHO) in 2008.[77–79] Many cases were identified to have well characterized chromosomal aberrations, such as the chromosomal translocation t(8;21) and the chromosomal inversion inv(16), although many others were classified to be cytogenetically normal (CN-AML).[68,80,81] The chromosomal abnormalities lead to the formation of recurrent fusion genes such as *RUNX1-RUNX1T1*, *BCR-ABL1*, *PML-RARA* in AML.[9,82] Additionally, patients were observed to have point mutations in genes such as *CEBPA*, *RUNX1* etc.[83,84] In 2017, the European LeukemiaNet (ELN) classified the AML genetic abnormalities into three risk categories (favourable, Intermediate and Adverse) and suggested to report the frequencies, response rates and outcome measures in one of these categories (**Table 1**).

Table 1: *2017 ELN risk stratification by genetics*[72]

Risk category*	Genetic abnormality
Favorable	t(8;21)(q22;q22.1); *RUNX1-RUNX1T1* inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11* Mutated *NPM1* without *FLT3*-ITD or with *FLT3*-ITD^{low}† Biallelic mutated *CEBPA*
Intermediate	Mutated *NPM1* and *FLT3*-ITD^{high}† Wild-type *NPM1* without *FLT3*-ITD or with *FLT3*-ITD^{low}† (without adverse-risk genetic lesions) t(9;11)(p21.3;q23.3); *MLLT3-KMT2A*‡ Cytogenetic abnormalities not classified as favorable or adverse
Adverse	t(6;9)(p23;q34.1); *DEK-NUP214* t(v;11q23.3); *KMT2A* rearranged t(9;22)(q34.1;q11.2); *BCR-ABL1* inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); *GATA2*,*MECOM*(*EVI1*) – 5 or del(5q); – 7; – 17/abn(17p) Complex karyotype,§ monosomal karyotype‖ Wild-type *NPM1* and *FLT3*-ITD^{high}† Mutated *RUNX1*¶ Mutated *ASXL1*¶ Mutated *TP53*#

The table shown above is taken from the 2017 ELN risk classification by Dönner *et al.*[72]
Frequencies, response rates, and outcome measures should be reported by risk category, and, if sufficient numbers are available, by specific genetic lesions indicated.
*Prognostic impact of a marker is treatment-dependent and may change with new therapies.
†Low, low allelic ratio (<0.5); high, high allelic ratio (≥0.5); semiquantitative assessment of *FLT3*-ITD allelic ratio (using DNA fragment analysis) is determined as ratio of the area under the curve "*FLT3*-ITD" divided by area under the curve "*FLT3*-wild type".
‡The presence of t(9;11)(p21.3;q23.3) takes precedence over rare, concurrent adverse-risk gene mutations.
§Three or more unrelated chromosome abnormalities in the absence of 1 of the WHO-designated recurring translocations or inversions, that is, t(8;21), inv(16) or t(16;16), t(9;11), t(v;11)(v;q23.3), t(6;9), inv(3) or t(3;3); AML with *BCR-ABL1*.[79]
‖Defined by the presence of 1 single monosomy (excluding loss of X or Y) in association with at least 1 additional monosomy or structural chromosome abnormality (excluding core-binding factor AML).
¶These markers should not be used as an adverse prognostic marker if they co-occur with favorable-risk AML subtypes.
#*TP53* mutations are significantly associated with AML with complex and monosomal karyotype.[72]

1.4.2 Recurrent mutations in AML

Somatic mutations in AML are thought to contribute to leukaemogenesis either by improving the ability of hematopoietic cells to proliferate or by preventing cells from maturing. The proliferation might be through the activation of intracellular signals that contribute to growth and survival (eg: *FLT3*, *KIT*), whereas the inhibition might be by blocking cell differentiation or enhancing self-renewal using altered transcription factors (*CEBPA*, *NPM1*, *RUNX1*, etc.).[85,86]

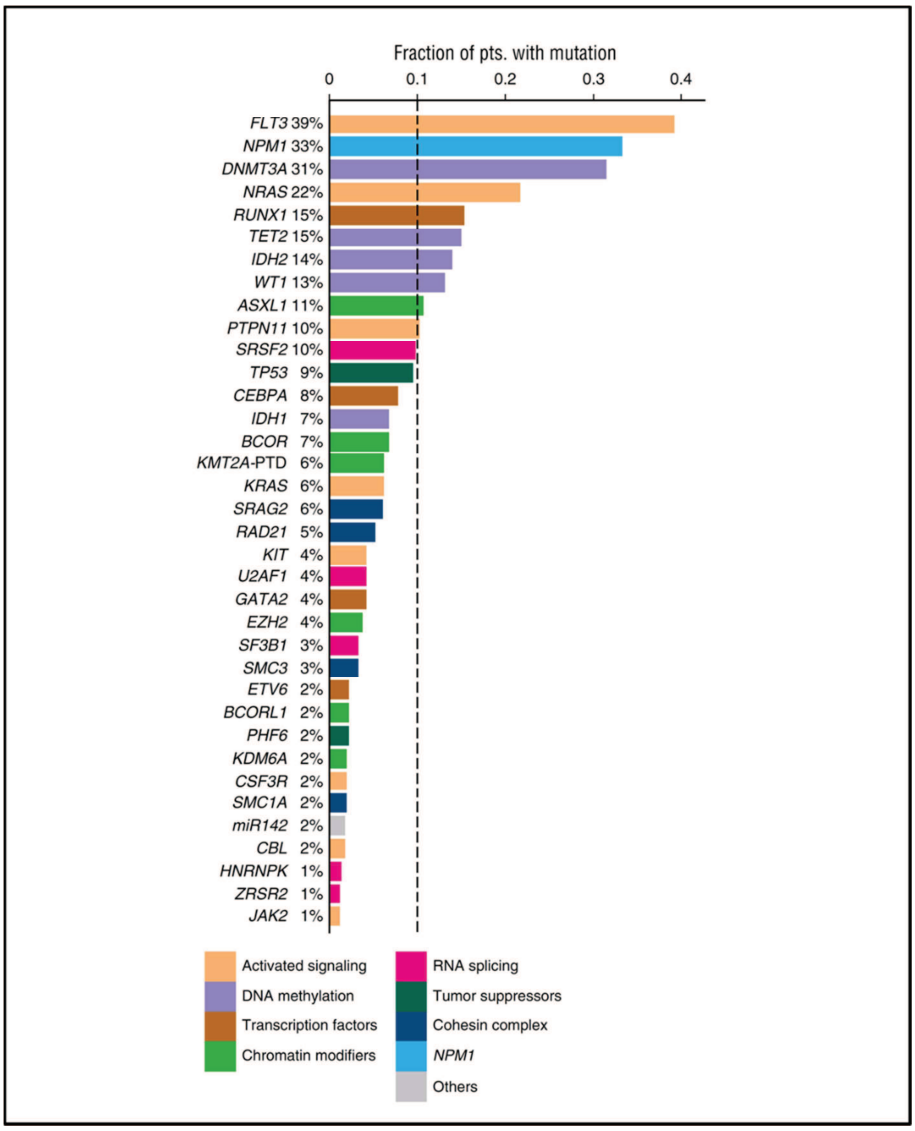

Figure 2: *Overview of driver gene mutations in AML. Histogram showing the frequency of driver gene mutations detected in >1% of patients (N=664). Bars are colored according to the functional category assigned to each driver gene. The figure is taken from the the overview of driver mutations in AML by Metzeler et al.*[87]

Extensive analyses on 200 AML patients by The Cancer Genome Atlas (TCGA) showed that 23 genes were recurrently mutated in AML.[9] A recent study by Lindsley *et al.* showed the presence of somatic mutations in genes such as *SRSF2*, *SF3B1*, *U2AF1*, *ZRSR2*, *ASXL1*, *EZH2*, *BCOR* or *STAG2* to be distinct genetic subtype for secondary AML diagnosis.[88] The proportion of the occurrence of these recurrent 'driver mutations' differ among the particular genes (**Figure 2**).[9,87] Mutations in *NPM1* and *CEBPA as well as FLT3*-internal tandem duplications (*FLT3*-ITDs) are widely used as prognostic and predictive markers as per the suggestions of the ELN.[72] In the case of CEBPA, bi-allelic mutations are considered as a distinctive entry because only those cases define a clinical and pathologic entity,

and are associated with favourable outcome.[89,90] However, other recurrent somatic mutations in *DNMT3A*, *IDH1*, *IDH2* etc. and their significance in disease prognosis are still under study. Germline mutations in at least 10 genes were associated with the inherited forms of myeloid neoplasm.[9,91,92] This thesis focuses on the 36 genes which were observed in more than 1% of the AML study population in a previous study trial (**Figure 2**) and the proportion of somatic mutations transcribed from DNA to RNA in these recurrently mutated genes in AML.[87]

1.5 Study background

The selection of a sequence aligner and variant caller highly influence the accuracy of variant detection.[93–95] One of the major difficulties in implementing a standard variant calling procedure for RNA-Seq is its inherent intricacies.[48] Recent advancement in computation algorithms enabled the researchers to develop and establish efficient splice-aware aligners to map the transcriptome sequence to the reference genome.[50,51] However, an accurate and reliable variant calling procedure in RNA-Seq is still a challenge.[96] Some of the best practice guidelines for variant calling are available for DNA-Seq, but no gold-standard pipelines have been established for both DNA-Seq and RNA-Seq.[3,97,98] The reliability of the detected variants in RNA are determined by comparing those variants in DNA.[96,99] An observation of mutation in DNA will have a 100% allele frequency in RNA in a haploid region, whereas around 50% or 100% of allele frequencies would be observed in the case of heterozygous or homozygous variants, respectively. However, the abundance of the heterozygous SNPs can be different in tumour cells when compared with the normal cells.[100] Some studies showed high correlation between the mutant allele frequencies of DNA and RNA.[7,101] Castle *et al.*'s study on mice tumour cell lines demonstrated a 99% concordance among the tumour mutations between DNA and RNA.[7] This study used the exome and transcriptome sequencing to determine the correlation between the mutation allele frequency in DNA and RNA followed by the measurement of AI (**Box 1**). They also suggested that the genes are equally transcribed in proportion to their DNA VAFs irrespective of mutated and wild-type allele.[7]

```
Imbalance = (RNA mutation allele frequency)
              minus (DNA mutation allele frequency)
```

Box 1: *Definition of allelic imbalance used by Castle et al.*[7]

In another study, O'Brien *et al.* showed only 14% overlap among the detected SNVs when comparing the WES and the RNA-Seq from 27 lung cancer pairs of tumour and matched normal samples.[102] Main reasons discussed by the authors were the presence of SNVs in the low coverage regions in either WES or RNA-Seq, allele-specific transcript abundance in RNA-Seq, location of SNVs outside the WES capture regions or RNA-editing.[102] However, another study on RNA-editing sites suggest very few DNA and RNA sequence differences and the occurrence of RNA-editing sites are quite rate.[103] Rhee et al. studied the AI of more than 100,000 somatic mutations from more than 2,000 cancer specimens across five human solid tumour types in TCGA cohort (AML not included).[8] They compared the WES and transcriptome sequencing and observed the AI among the nonsense SNVs and frameshift INDELs as well as among splice site mutations.[8] The splice site mutation-harbouring alleles were observed to be relatively over-expressed when compared with the wild-type alleles.[8] A Comprehensive analysis of AML patients by TCGA (151 paired DNA- and RNA-Seq samples) showed allelic biases in the expression of mutations among *DNMT3A*, *PHF6*, *RUNX1*, *TET2*, *TP53* and *WT1*.[9] TCGA compared either WGS or WES with the RNA-Seq and included only the SNVs which were detected at a minimum read depth cut-off of 10x in the RNA-Seq.[9] Some of the imbalances were discussed to be due to copy number events, hemizygous variants (all PHF6 mutations were from male patients and were located on X chromosome) or the loss of heterozygosity.[9] Celton and colleagues also observed the existence of ASE among the low expressed *GATA2*-mutated AML samples.[104]

The phenomenon of AI was observed in different cancer types and was associated with genomic imprinting, copy number alterations or epigenetic mechanisms. Although allele-specific transcript abundance was observed among the AML samples, there was no systematic understanding of such imbalances among the recurrently mutated genes in AML. Mutational screening of *NPM1*, *FLT3*, *RUNX1*

and *CEBPA* are recommended by ELN to be in routine practice for AML prognosis and the investigation of AI among these genes in a large cohort might provide additional knowledge in understanding the development of leukaemia.

1.6 Objectives of this study

The overall objective of this thesis was to compare the allelic proportions of recurrent mutations transcribed from DNA to RNA and to determine the existence of allelic imbalance among genes which are recurrently mutated in AML. The following are the objectives of this study.

1. To detect potential sequence variants (SNVs and small INDELs) from the targeted DNA- and RNA-Seq using optimized sequence aligners and variant callers.
2. To define, implement and optimize the imperative filtering criteria on raw variants in order to eliminate potential sequence artefacts and enrich the called variants.
3. To evaluate the proportions of recurrent mutations transcribed from DNA to RNA and to investigate the existence of AI among the genes^{MUT} which are recurrently mutated in AML.
4. To analyse the presence of AI among the genes^{WT} which did not harbour any recurrent mutations and to study the impact of recurrent mutations on their respective gene and transcript isoform expression levels.

2 Methods

The overall work in this thesis involves the improvement of the conventional variant calling pipeline for targeted DNA-Seq and RNA-Seq followed by the analysis and validation of AI among the recurrently mutated genes^{MUT} in AML. The processing of raw sequencing reads of both DNA and RNA to call the sequence variants and the comparison of sequence aligners and variant callers are discussed in section 2.2 and 2.3. The called raw variants from both sequences were filtered for potential sequence artefacts (section 2.4) and were used for downstream analyses. The comparison of VAFs of the genes^{MUT} and mutation types^{MUT} in patients harbouring recurrent mutations and genes^{WT} in patients with wild-type status were analysed to detect any AI among them (section 2.6). Finally, the gene-level and transcript-level differential expression analyses were carried out between the patients with mutant and wild-type status in the respective genes (section 2.7).

2.1 Study population

This study included 499 AML samples from four independent cohorts: the German AML co-operative group (AMLCG), the German cancer consortium (Deutsches Konsortium für Translationale Krebsforschung, DKTK), the Cancer Genome Atlas (TCGA) and the HELSINKI cohort. The primary cohort consisted of 246 samples from the AMLCG study population. The participants sequenced for both DNA and RNA in AMLCG-1999 (n=45) and AMLCG-2008 (n=201) trials were included in this study[87,105]. The diagnosis of AML was based on the criteria recommended by the WHO.[79] All the patients included in the primary cohort received cytarabine- and anthracycline- based induction treatment. Further details regarding the treatment protocols and patient selection were published in previous studies.[87,105] The study protocols were in accordance with the Declaration of Helsinki and approved by the institutional review boards of the participating centres. All patients provided written informed consent for inclusion on the clinical trial and genetic analyses. Additional exclusion criteria based on the coverage statistics of the variants on the genes of interest were described in the later sections (**Figure 3**). The validation cohort includes AML samples from three external cohorts: DKTK (n=40), TCGA (n=116) and HELSINKI (n=97) cohorts[9,106–108]. The targeted DNA-Seq covering the

genes of interest was available in the case of DKTK samples, whereas only WES data sets were available for TCGA and the HELSINKI cohorts. Nevertheless, the downstream processing and analyses were carried out in all three validation cohorts by considering the differences among them as well.

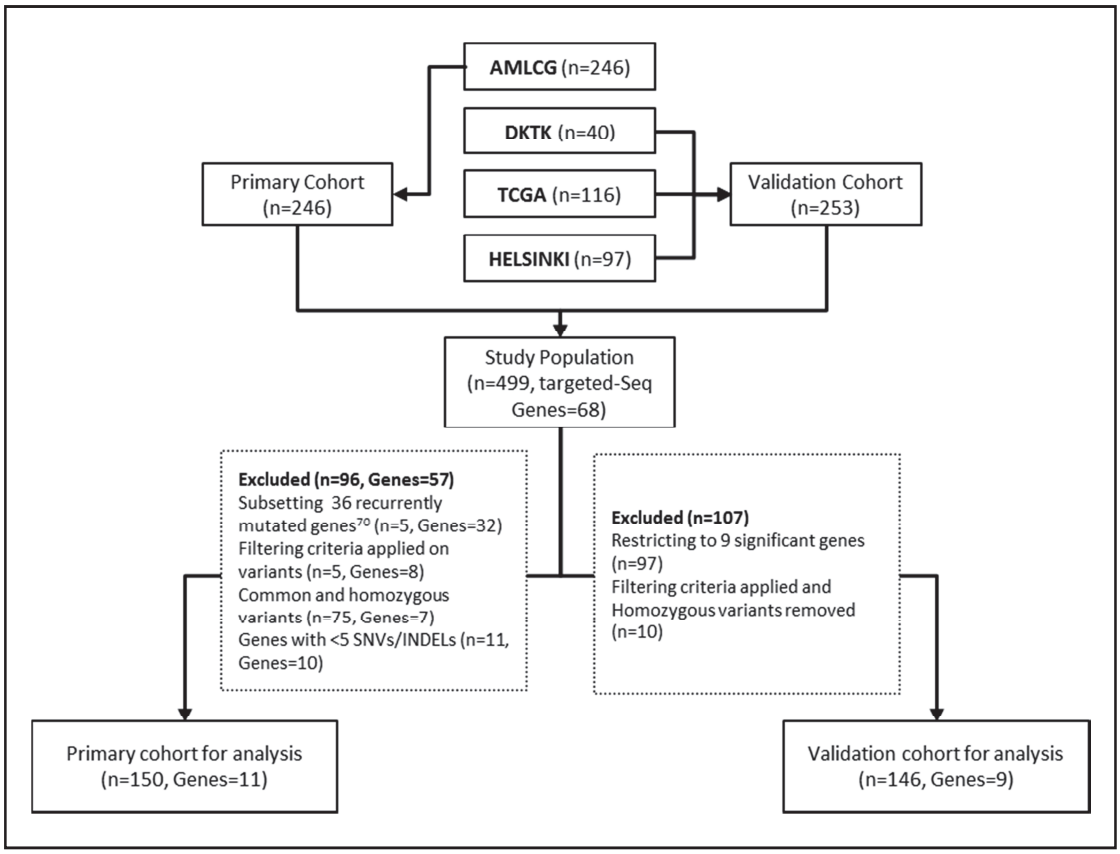

Figure 3: *Flow diagram of primary and validation cohorts.*

2.2 Sequencing and quality trimming in AMLCG cohort

2.2.1 Targeted DNA sequencing

In the primary cohort, a total of 68 genes, which are recurrently mutated in myeloid malignancies, were sequenced using a custom amplicon-based targeted enrichment assay (Haloplex, Agilent, Boeblingen, Germany). The entire coding region of 37 genes was sequenced along with recurrently mutated regions from 31 genes.[9,34] List of all 68 genes and the regions defined in targeted DNA-Seq are shown in the appendix (**Table 4**). All 246 samples were sequenced (250 base pairs (bp), paired-end) on an Illumina MiSeq instrument (Illumina, SanDiego, CA). The sequence information was obtained as a text-based FASTQ format files from multiple samples with unique infused barcodes (multiplexing). These sequences

were de-multiplexed using Je (v1.0) de-multiplexer in order to isolate individual samples.[109] After de-multiplexing, the Illumina adapters from the sequences were removed and were quality trimmed using Trimmomatic (v0.32).[110] The leading and trailing bases were cut off from the start and end of the reads, respectively, when the read quality threshold dropped below a score of 15. All reads below the read length of 45bp were filtered out as well. Further, a sliding window trimming (window size: 6bp) was also performed once the average quality within the window fell below 20. The quality of the sequence reads were determined using FastQC (v0.10.1) and MultiQC (v1.5) before and after the quality trimming procedure (**Figure 5**).[111,112]

2.2.2 RNA sequencing

Transcriptome Sequencing libraries were prepared using the Sense mRNA-Seq Library kit V2 (Lexogen). All the sequences were 100bp length, paired-end, strand-specific and were performed using a poly(A)-selected protocol.[105] Similar to DNA-Seq, RNA-Seq were also de-multiplexed to isolate individual samples.[109] In order to reduce the proportion of errors due to non-specific hybridization, nine and six bases were removed from the forward and reverse reads, respectively.[113] This was followed by adapter trimming using Cutadapt (v1.7.1).[114] The sequence reads were then quality-trimmed using adaptive quality-trimming with a minimum read length of 40bp and strictness of 0.5 to retain maximum number of reads possible.[110] The base quality statistics before and after quality trimming are shown in **Figure 6**.

2.3 Variant calling pipeline

Extraction of sequence variants from quality-trimmed reads is a multi-step process including alignment of sequences to the reference genome, post-processing aligned reads, selection of specific regions of interest when necessary, local re-alignment of INDELs, application of variant calling algorithm, annotation of called variants using external databases and filtering artefacts. Conventional sequence processing procedures were improved from available best practice guidelines.[97] Many sequence aligners and variant callers were considered for both DNA- and RNA-Seq to determine an optimal aligner and variant caller for both sequences.

2.3.1 Selection of sequence aligner

A random selection of a sample from the AMLCG cohort was carried out and its DNA- and RNA-Seq reads were aligned using different aligners for the comparison of speed and accuracy. Although there are a large number of sequence aligners available for DNA, three commonly used mappers were compared: BWA-MEM (v0.7.10), Bowtie2 (v2.2.6) and CUSHAW3 (v3.0.3).[45–47] In the case of RNA-Seq, three splice-aware aligners were considered: TopHat2 (v2.0.14), HISAT2 (v2.0.0) and STAR (v2.5.1b).[49–51] The sample reads were aligned to the human (hg19) reference genome build. In the case of RNA-Seq aligners, the gene transfer format (GTF) file containing the definitions of the gene structure was also provided based on the reference genome build used. For each aligner, the reference genome indices were built and the selected sample data was aligned using default parameters. The alignment summary metrics from the above mentioned aligners were observed along with the time taken for creating genomic indices and sample processing in order to determine their performance. The superior sequence aligner for DNA- and RNA-Seq was selected based on the observed factors. The chosen aligner parameters were further optimized to yield better alignment for targeted DNA-Seq and whole RNA-Seq (discussed in section 3.1).

2.3.2 Post-processing of the aligned sequences

Most of the aligners output the mapped reads in Sequence Alignment/Map format (SAM) or it's in its binary version (BAM).[55] The DNA- and the RNA-Seq were processed differently due to the inherent differences in their sequences as well as the sequencing methods used. The aligned reads of DNA-Seq were re-ordered and sorted based on their chromosomal co-ordinates. The duplicate sequence reads in amplcon-based targeted DNA-Seq have the same start and end positions as the initial read as they originate from the same amplicon and thus they were not removed. The sequencing reads which were mapped outside the recurrently mutated regions of interest were excluded. All the above processing was done using the Picard tool-box (v1.136).[54] The INDELs were then re-aligned using the GATK (v2.7.4).[62] In the case of RNA-Seq, the aligner generated properly paired and discordantly paired sequences separately. These were merged initially, followed by re-ordering sequence reads and sorting based on chromosomal co-ordinates.

The duplicate reads in RNA-Seq were removed to reduce sequence over-representation. The reads which were mapped outside the regions of interest were also excluded in the RNA-Seq. Due to the missing intronic regions in the RNA-Seq, a large number of "N" CIGAR strings were introduced by the aligners while mapping. These reads with "N"s in the middle were split into two reads and the CIGAR strings were converted into soft clips with the splitNRead tool.[115] All the INDELs were left aligned using bamleftalign from the Freebayes packages.[60] These processed sequences were used for further variant calling procedures.

2.3.3 Selection of variant caller

Similar to sequence aligners, there are several variant calling algorithms available for detecting SNVs and INDELs[60–62,116] This study focused on two variant callers: VarScan (v2.3.5) and VarDict.[61,116] VarScan requires the input in SAMtools mpileup format, in which it extracts coverage and quality information from sequence alignment SAM/BAM files.[55] The mpileup files are then used to detect SNVs and INDELs. In the case of VarDict, It takes the aligned BAM files directly to perform further local re-alignment to enrich INDEL detections, followed by variant calling. All 246 AMLCG samples were used to compare the variant callers. For DNA-Seq, loci with minimum read-depth cut-off of 30x was considered, along with mapping quality score of 10, base quality of 20 and minor allele frequency >2% for both variant callers. A VarScan p-value of 0.01 was also employed as a preliminary cut-off for DNA variants. Regarding RNA-Seq, a less stringent read-depth cut-off of 4x was used, along with a mapping quality score of 13 and a minor allele frequency >1% in order to avoid premature elimination of reputed variants. Subsequently, the called variants were functionally annotated using Annovar (vAug2013).[117] Publicly available databases such as COSMIC, dbSNP, ClinVar, RADAR and 1000 Genomes Project were used for annotating chromosomal position, clinical relevance and its frequency of occurrence in general population.[33–35,118,119] Mutation prediction scores such as SIFT, PolyPhen and MutationTaster were also calculated using Annovar.[120–122]

2.4 Application of variant filtering criteria

Several filtering criteria were employed on called raw variants, to improve the sensitivity and specificity of the variant calling pipeline (**Figure 4**a). The minimum read depth cut-off was raised to 10x based on the variant distribution across RNA read depths (**Figure 9**). All the variants, which were annotated in the RADAR database containing known RNA editing sites (post-transcriptional modification of RNA nucleotides), were removed.[118] Variants detected in the regions containing simple tandem repeats defined by UCSC were also excluded. To further determine the mapping quality bias, base quality bias, strand and tail-distance biases, all the detected variants were recalled using SAMtools (v0.1.19) and BCFtools (v0.1.19).[55] The p-value for base quality bias, mapping quality bias and tail-distance bias were calculated using t-test, whereas for strand bias, exact test was used.[55] Variant with a p-value ≤ 0.05 in any of the biases were filtered out. Furthermore, custom regions were defined to filter out artefacts and enrich the true positive variants (**Figure 4**b-d). They are listed below:

- A low mapping quality region was defined as a locus with more than 1/3rd of the supporting reads being of low mapping quality (<10).
- An error-prone region was defined as a 25bp upstream and 25bp downstream window around the variant of interest, in which the alternate allele frequency in the window (excluding the variant of interest) is > 1%.
- Position based filtering was carried out in a region which is defined as a 10bp upstream and 10bp downstream window around the variant of interest, in which the total number of bases (irrespective of the allele) on either left or right side of the variant of interest should be <50% of the total number of bases on the other side.

Variants which were annotated to be in one of the above defined regions were filtered out. The filtered alterations of transcribed and DNA-exclusive variants were used for VAF analysis, whereas those of RNA-exclusive variants were filtered further for detecting potential RNA edit sites. These sites were further filtered depending on their occurrence in at least 5% of the primary study population.

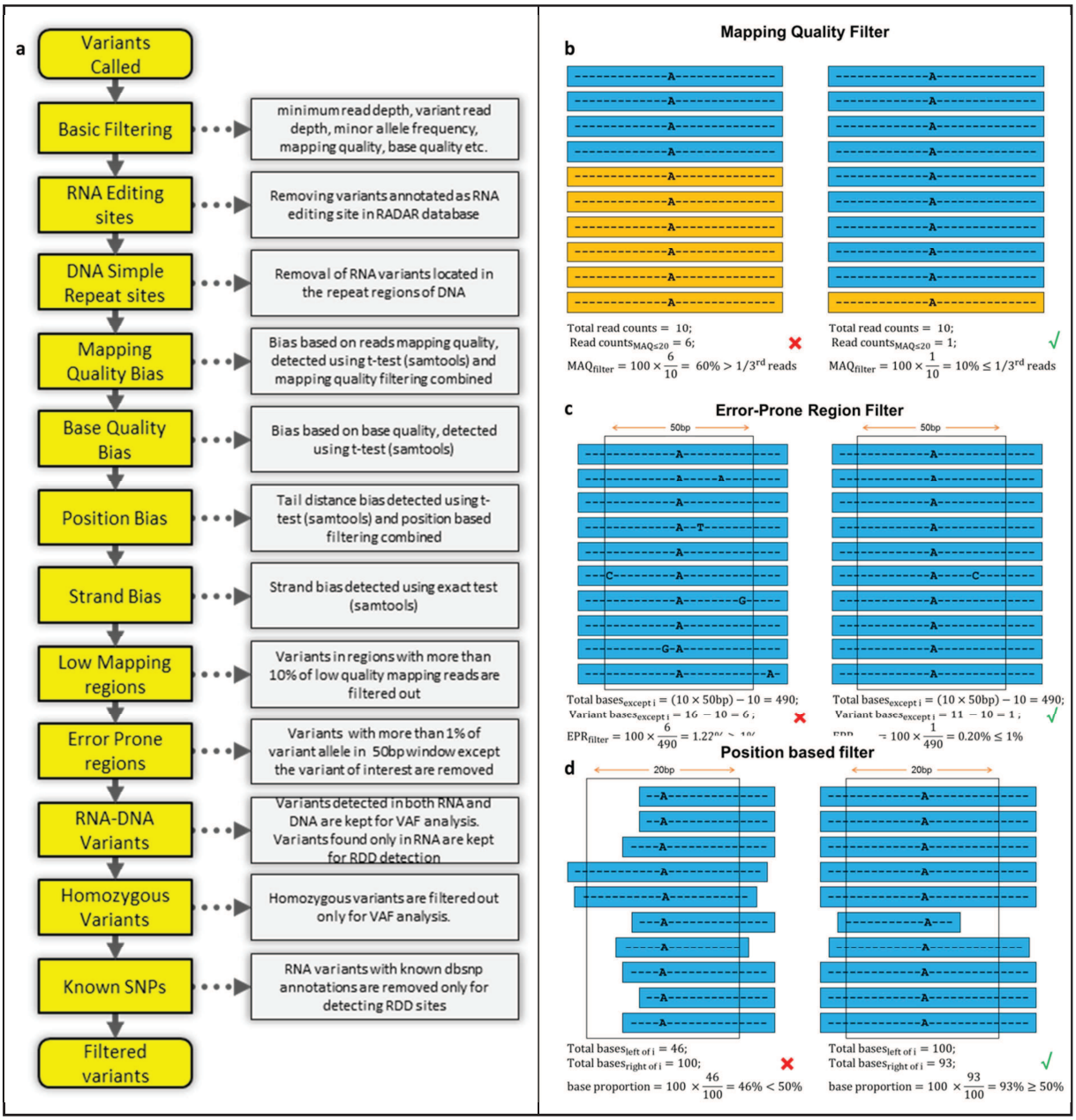

Figure 4: *Filtering criteria and visualization of criteria definitions. A) List of criteria applied. B) Low mapping quality region cut-off >=1/3rd reads. C) Error-prone region cut-off <=1% alternate allele frequency. D) Position based filtering cut-off >=50%.*

2.5 Sequence alignment, variant calling and filtering in validation cohort

All the DNA-Seq in the validation cohort were aligned using BWA/MAQ and variants were called using VarScan.[9,106,107] In the case of RNA-Seq, STAR aligner was used for mapping followed by variant calling using VarScan in both the DKTK and TCGA cohorts. In the case of the HELSINKI cohort, GATK best practice guidelines were followed.[97] All the above mentioned filtering criteria were applied on the called variants except for recalling variants using SAMtools and BCFtools.

2.6 Variant allele frequency comparison among recurrent mutations

All the recurrent mutations detected in the AMLCG cohort by Metzeler *et al.* were extracted from the filtered variants list and their VAFs were compared.[87,105] The mutations with alternate allele frequency between 2% and 75% were considered as heterozygous genotype and only those mutations were included in the analysis. All the other homozygous variants and the RNA-exclusive variants were dropped off. In order to determine AI among the genes, a linear regression model was used to compare the observed and expected RNA variant allele read depth in sequence fragments and the allele specific transcript abundance was calculated in the form of weighted allelic imbalance (WAI). The WAI is the estimation of AI by transforming the VAFs of DNA and RNA into the expected and observed variant read depth in RNA. The definition of expected variant read depth of RNA from the VAF of DNA is shown in **Box 2**. All the mutations were grouped for each genes^{MUT} separately and were adjusted for different mutation types (**Box 2**, model 1) to study the association of mutations and the VAF difference among DNA and RNA. The mutations were also grouped according to the mutation types^{MUT} and were adjusted for the genes (**Box 2**, model 2). The common SNPs were extracted from the filtered variants (based on dbSNP annotations, build 138) and the proposed model was applied on it in patients without recurrent mutations in their respective genes^{WT} to investigate the existence of allele-specific transcript abundance in general, irrespective of the mutation status (**Box 2**, model 3).

Expected Variant allele read depth Definition:

$$RNA\ Variant\ Depth_{i,Exp} = DNA\ VAF_{i,Obs} * RNA\ Total\ Depth_{i,Obs}/100$$

Linear regression model 1 applied on each gene^{MUT}:

$$RNA\ Variant\ Depth_{i,Obs} \sim RNA\ Variant\ Depth_{i,Exp} + Variant\ Type_i$$

Linear regression model 2 applied on each variant type^{MUT}:

$$RNA\ Variant\ Depth_{i,Obs} \sim RNA\ Variant\ Depth_{i,Exp} + Gene_i$$

Linear regression model 3 applied on each gene^{WT}:

$$RNA\ Variant\ Depth_{i,Obs} \sim RNA\ Variant\ Depth_{i,Exp} + Variant\ Type_i$$

Box 2: *Expected variant allele read depth definition and linear regression model employed. Gene denotes the recurrently mutated genes selected for the analysis and Variant Type denotes the annotated mutation types such as non-synonymous SNVs, stopgain SNVs, frameshift and non-frameshift insertions, deletions and substitutions. The 'Exp', 'Obs' and 'i' denotes the expected and observed values for every variant i.*

2.7 Differential expression based on the recurrent mutations

2.7.1 Gene and transcript isoform quantification

The conventional procedure for differential expression analysis uses the alignment based quantification of reads. In this method, the raw or quality trimmed reads are aligned using a splice-aware aligner as described above (section 2.3.1), followed by the quantification of read counts per gene using tools such as htseq-count or featureCounts.[123,124] The obtained read counts are then used for gene-level differential expression analysis using R packages such as DESeq2 or edgeR.[125,126] The procedure for the quantification of transcript isoforms using tools such as Cufflinks or RSEM, followed by transcript-level differential expression analyses.[127,128] However, in recent years, tools such as Kallisto and Salmon proposed to use the raw or quality trimmed reads directly to quantify the transcript isoforms.[129,130] These tools extract *k*-mers from the reads followed by the exact matching of them using the hash tables and thus greatly reducing the processing time when compared to the alignment-counting routines.[131] This methodology does not determine the exact alignment location within a transcript but rather provides a probabilistic measure of the transcript from which it could have been extracted. In our analysis, the transcript quantification was carried out using Salmon (v0.9.1) and the gene quantification was performed by aggregating the transcript counts.[130] Salmon uses a quasi-mappings method in which it

computes the mapping of reads to transcript positions without performing a base-to-base alignment of the reads to the transcript.[130]

2.7.2 Gene-level and transcript-level differential expression analysis

The differential expression of the genes and transcript isoforms were studied by grouping the patients with and without recurrent mutations in the respective genes to further investigate the AI and allele-specific transcript abundance. The quantified counts were used to conduct the gene- and the transcript-level differential expression analyses using limma (v3.34.9).[132] The total count of reads mapped to a gene or transcript depend on their own expression level, the length of the sequence reads, the read depth and the expression of other genes within the sample. In order to account for such variability and their systematic effects, the quantified read counts are filtered for low gene or transcript counts and normalization procedure. In our analyses, tool edgeR (v3.20.9) was used for filtering out all the genes or transcript isoforms with less than one count per million in 5 samples, followed by the Trimmed Mean of M-values (TMM) normalization.[126] The counts were then transformed with sample-specific quality weight adjustment in the experiment design (limma, voomWithQualityWeights) and fitted to linear model based on the mutation status in the genes with substantial WAI in the previous analyses (section 3.7.1).[132] The fold changes were calculated for both gene and transcript isoforms and were adjusted for multiple testing.

All the data processing of DNA- and RNA-Seq were carried out using an in-house Galaxy platform (v15.10.2).[133] All the statistical analyses were performed using R (v3.4.3) and were adjusted for multiple testing using Benjamini & Hochberg procedure.[134,135] The R session information is provided in the appendix (**Box 7**). We considered an adjusted p-value cut-off of ≤0.05 to determine the significance of the results obtained.

3 Results

A total of 499 AML patients were included in the analysis the study design is shown in **Figure 3**. The sequence coverage information and alignment statistics of the AMLCG cohort are shown in **Table 2**. One of the main difference between the targeted DNA- and RNA-Seq is the average coverage in the targeted regions (542x and 85x in the case of DNA-Seq and RNA-Seq, respectively). This conspicuous difference is due to the differences in the sequencing techniques used. The sequence per base quality of raw DNA-Seq reads (**Figure 5**a) showed almost 90% of low quality bases towards the end of the reads, whereas fewer reads in the RNA-Seq also showed a similar trend (**Figure 6**a). The application of adapter trimming and quality filtering eliminated a large number of low quality reads in both sequences (**Figure 5**b). The enrichment in read quality due to the removal of 9 and 6 bases from the forward and reverse reads in RNA-Seq is also clearly distinguishable (**Figure 6**b).

Table 2: *DNA and RNA Sequence Information*

Info	Targeted DNA-Seq	Total RNA-Seq
Sequencing kit	custom amplicon-based targeted enrichment assay (Haloplex)	Lexogen SENSE mRNA-Seq kit V2
Sequence length	250bp	100bp
Reference genome build	hg19 (Human)	hg19 (Human)
Average Total Aligned Reads (range)	722755 (301046–2208674)	57568431 (36418879–167296173)
Properly Paired Reads	99.6 %	99.4 %
Average Coverage In Target Regions (range)	542.3x (196.3 x – 2653.5 x)	85.32x (33.2x – 301.4x)
Average of Mean Insert Size (range)	178.6 (166.6 – 193.2)	308.0 (134.9 – 583.9)

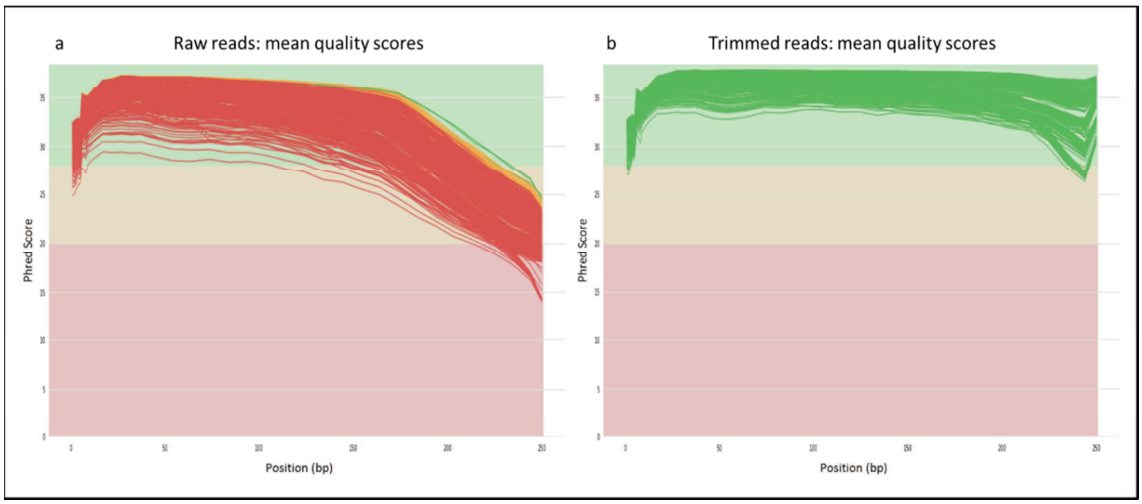

Figure 5: *DNA-Seq quality information of (a) raw and (b) quality trimmed reads per base level in AMLCG (n=246). The green, yellow and red colours indicate the high, medium and low quality reads.*

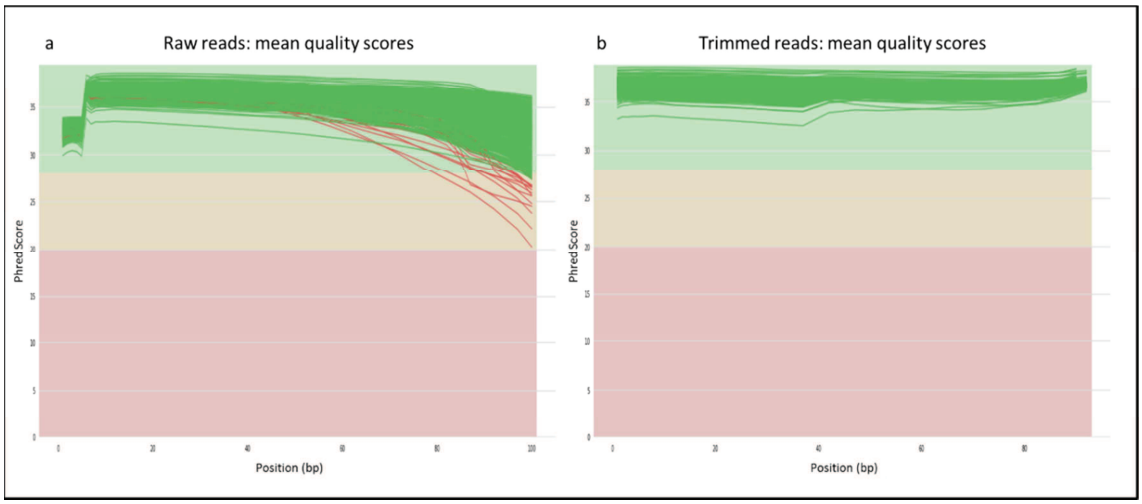

Figure 6: *RNA-Seq quality information of (a) raw and (b) quality trimmed reads per base level in AMLCG (n=246). The green, yellow and red colours indicate the high, medium and low quality reads.*

After the adapter clipping and quality trimming processes, both the processed sequences were considered to be valid for the subsequent analyses.

3.1 Sequence aligner comparison

The comparison of DNA-Seq aligners (**Table 3**) showed similar results when comparing the percentage of mapped and properly paired reads, whereas CUSHAW3 performed relatively well while indexing hg19 reference genome build (3 Gbp). However, BWA-MEM outperformed Bowtie2 and CUSHAW3 in the number of reads processed per second. In the case of RNA-Seq aligners, both HISAT2 and STAR excelled TopHat2 in most scenarios. Although STAR has inferior mapped and properly paired read proportions and requires substantially more

indexing time when compared to HISAT2, its processes almost three times the number of reads when compared to HISAT2. As the genomic indexing is done only once per reference genome build and the sequence alignment rate is better than its peers, BWA-MEM and STAR were preferred to be used for processing the DNA- and the RNA-Seq, respectively.

Table 3: *Aligner Comparison for DNA and RNA Sequencing*

Aligners	Mapped Reads (%)	Properly Paired Reads (%)	Singletons (%)	Indexing hg19 Genome (time)	Sequence Alignment Rate (reads/second)
DNA					
bowtie2	96.56 %	93.80 %	2.29 %	96m 11.577s	3960.79
BWA-MEM	99.25 %	98.67 %	0.55 %	73m 26.007s	7076.86
CUSHAW3	98.82 %	98.60 %	0.15 %	35m 3.980s	440.016
RNA					
HISAT2	97.64 %	91.02 %	1.83 %	62m 17.826s	11340.5
STAR	95.42 %	90.55 %	0.68 %	165m 32.375s	29270.5
TopHat2	88.92 %	70.23 %	9.09 %	96m 11.577s	1272.7

3.2 Variant caller comparison

The VAFs of recurrent mutations in AML were called using VarScan and VarDict and were plotted against each other for both the DNA- and the RNA-Seq (**Figure 7**). The VAF of the detected SNVs from both callers showed a similar trend. However, some differences were detected among the called INDELs. Even though a few INDEL VAFs were deviated from the expected values in the case of DNA, a large dispersion was observed among RNA variants. A handful of INDELs were not detected at all by VarScan. We also observe a trend in underestimation of INDEL VAFs by VarScan when compared to VarDict. This might be due to the in-built local realignment procedure in VarDict, enabling it to detect INDELs more accurately.

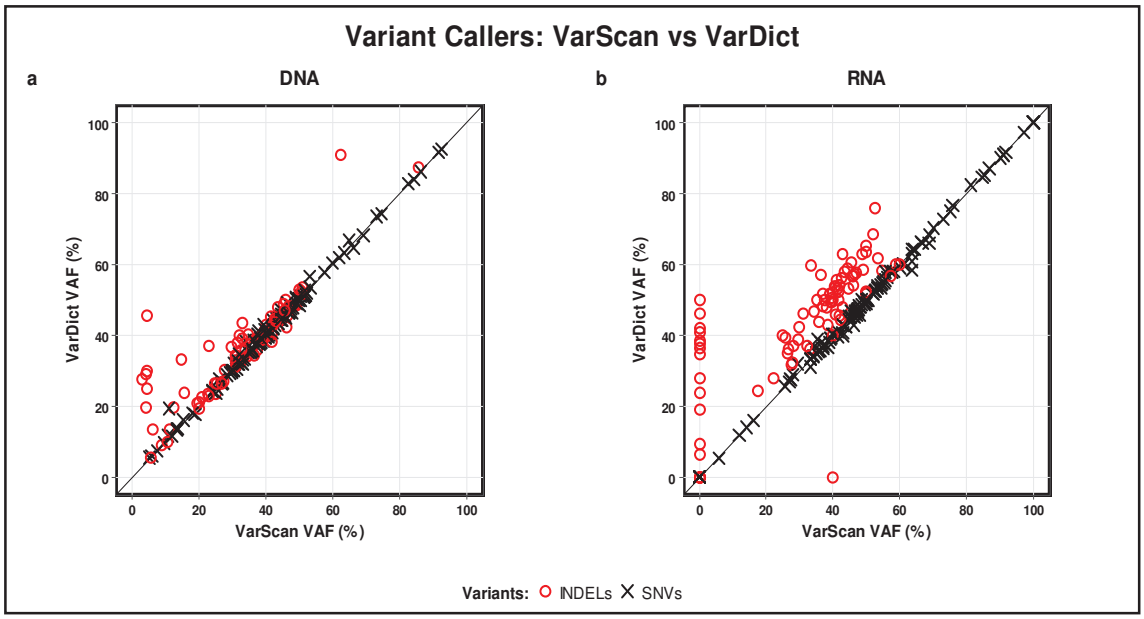

Figure 7: *Variant Allele Frequency differences of recurrent mutations between VarScan and VarDict in (a) DNA and in (b) RNA. The solid diagonal lines represent the expected VAF trend among the called variants.*

3.3 DNA and RNA variant calling pipeline

Both BWA-MEM (v0.7.10) and STAR (v2.5.1b) aligners were finalized for mapping the DNA- and the RNA-Seq, respectively, due to their superior performances. The optimized parameters used are shown in the appendix (**Box 3** and **Box 4**). The alignment performance of BWA-MEM on targeted DNA-Seq is improved by raising the penalty for mismatch of the sequence reads and reducing the limits of re-seeding. Although these parameters decrease the processing speed of the aligner, it greatly increases the accuracy. As the targeted DNA-Seq has large sequence coverage, accuracy of base-to-base alignment of reads is crucial for any downstream analysis. Also, the parameters to mark shorter split hits as secondary reads was also included to make it compatible with the Picard toolbox for downstream processing (**Box 3**).[54] In the case of RNA-Seq, the quality trimmed reads from all the samples in the primary AMLCG cohort were initially aligned to the previously indexed reference genome (with known gene definitions) using STAR aligner. The computed splice junctions output from all the samples were pooled, followed by filtering out previously annotated known splice junctions and mitochondrial regions. The un-annotated splice junctions were used to index the reference genome for the second time. This is followed by the alignment of all the samples to the newly indexed reference genome. This process is called the STAR

second pass alignment and this increases the accuracy of sequence alignment (**Box 4**). Regarding the variant callers, VarScan and VarDict were employed for calling SNVs and INDELs, respectively in both sequences. The parameters used for both variant callers were shown in the appendix (**Box 5** and **Box 6**) and discussed in the section 2.3.3. Thus the variant calling pipeline established in both sequences was exhibited in **Figure 8**. The called raw variants were further enriched by the application of filtering criteria defined (section 2.4).

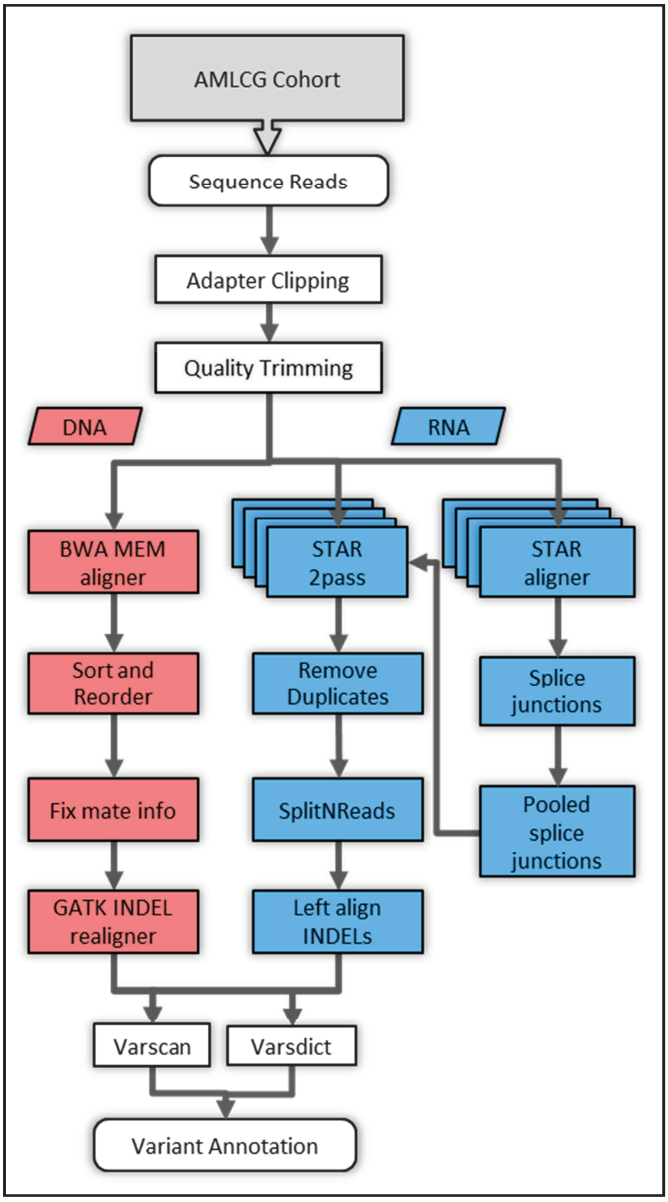

Figure 8: *Variant calling pipeline.*

3.4 Called variants in both sequences

The raw variants detected in both DNA and RNA were classified into three groups based on their detected sequences: transcribed (present in both DNA and RNA), DNA-exclusive (not detected in RNA with sufficient read depth) and RNA-exclusive (not detected in DNA with sufficient read depth) variants. The total number of variants called in both sequences distributed across different RNA read depths is shown in **Figure 9**. Variant calling in the recurrently mutated genes in both sequences resulted in a total of 8,052 variants (89.3% were SNVs and 10.7% were INDELs). Among the detected variants, 47.9% of them were RNA-exclusive, whereas 3.8% belong to DNA-exclusive variants. Some of the RNA-exclusive variants might be potential RNA editing sites, but most of them were suspected to be sequence artefacts.[103,136] The presence of DNA-exclusive variants might be due to a strong preferential allele-specific transcript abundance towards the wild-type allele or alternative mechanisms such as rapid RNA decay, epigenetic mechanisms of genomic imprinting or copy number alterations. The observed larger proportion among the exclusive variants also demonstrates the accuracy of variant calling in DNA-Seq when compared to RNA-Seq (**Figure 9**a-b).

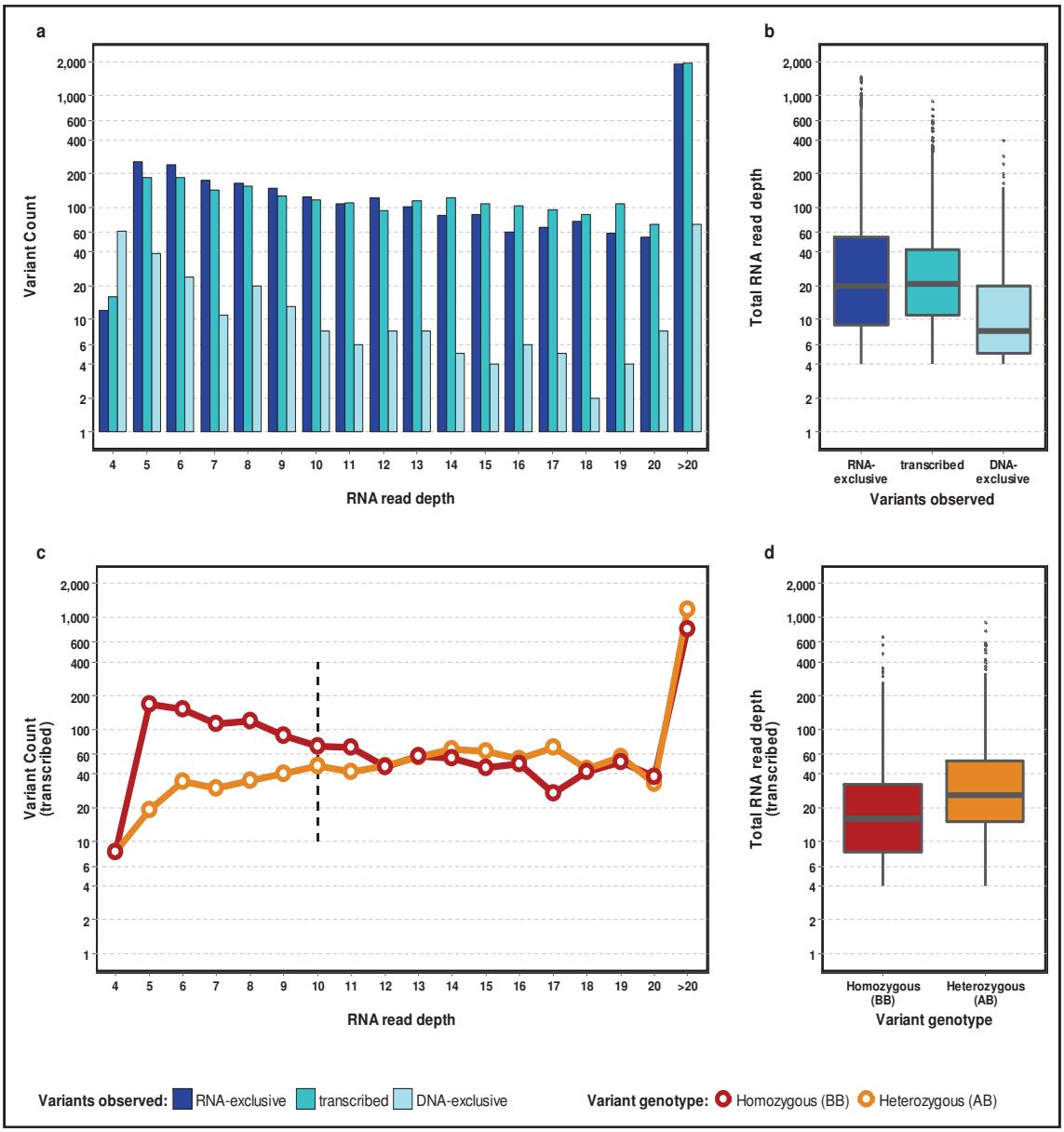

Figure 9: *RNA-Seq read depths of all detected variants. (a) RNA-Seq read depths grouped based on the different variant classes. (c) RNA-Seq read depth of transcribed variants grouped according to variant genotype information. (b,d) Read depth distribution based on variant groups.*

Initial variant calling in RNA-Seq was carried out with a minimum read depth cut-off of 4x. Thus to improve the reliability of the called variants, the proportion of the transcribed variants between homozygous (BB) and heterozygous (AB) genotypes were calculated (**Figure 9**c-d). The proportion of homozygous and heterozygous variants was observed to converge with the increase in RNA read depth. Observation of the differences between the proportions stabilized beyond a read depth of 10x. Strikingly, TCGA also employed a 10x read depth cut-off for detecting variants in RNA-Seq.[9] However, the distribution of SNVs and InDels separately showed that this cut-off might not be ideal for INDEL detection (**Figure 10**). The

cut-off was established based on the distribution of homozygous and heterozygous variants per RNA read depth and thus it required sufficient variant count in each read depth category. The average variant count calculated until read depth of 20 was found to be very low for the INDELs (mean INDEL count: 5.5) when compared to that of the SNVs (mean SNV count: 54.8). Due to such low INDEL counts per RNA read depth bin, it was not optimal to select an independent minimum read depth cut-off for INDELs (**Figure 10**d). Thus, an overall minimum read depth cut-off of 10x was defined to call the sequence variants.

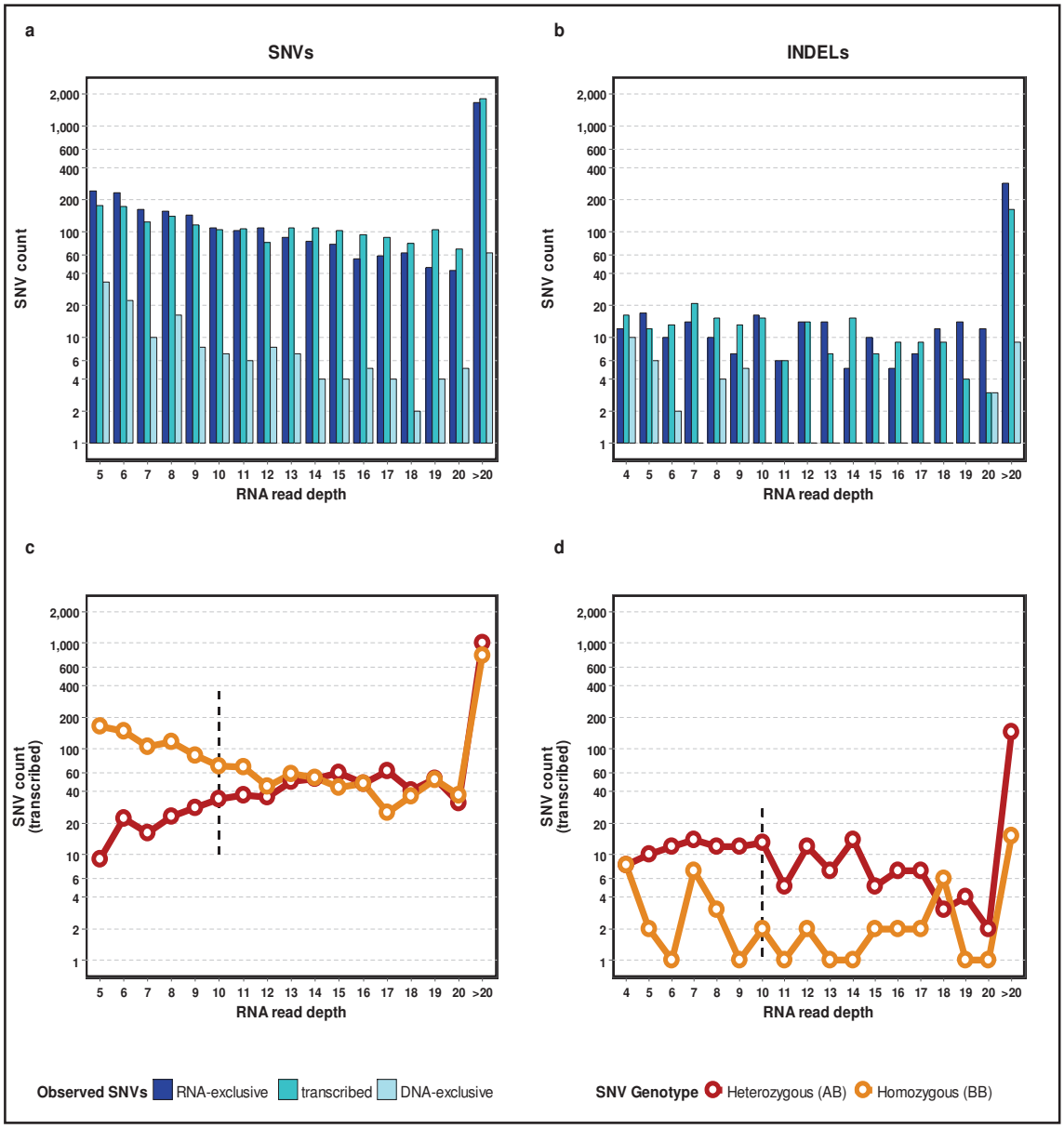

Figure 10: *RNA-Seq read depth of different variant classes for (a) SNVs and (b) INDELs. RNA-Seq read depth of transcribed SNVs (c) and INDELs (d) grouped according to variant genotype information.*

3.5 The effect of filtering criteria on called variants

The filtering criteria applied on SNVs and INDELs were plotted separately to understand their effect on both variants (**Figure 11**). Even though 24% of the transcribed variants failed due to the minimum read depth cut-off, overall application of the filtering criteria resulted in excluding 36.2% of all transcribed variants (2,302 SNVs and 182 INDELs). Among DNA-exclusive variants, 59.7% of them were filtered out leaving 106 SNVs and 16 INDELs. In the case of RNA-exclusive variants, around 67% of them failed due to position bias alone and thus resulted in excluding 91% variants as potential sequence artefacts. Further filtering for the occurrence of potential RNA editing sites in at least 5% of the study population and visualizing the variants using IGV removed all RNA-exclusive variants.[137] The proportion of variants removed was also similar for SNVs, whereas, larger differences were found among INDELs.

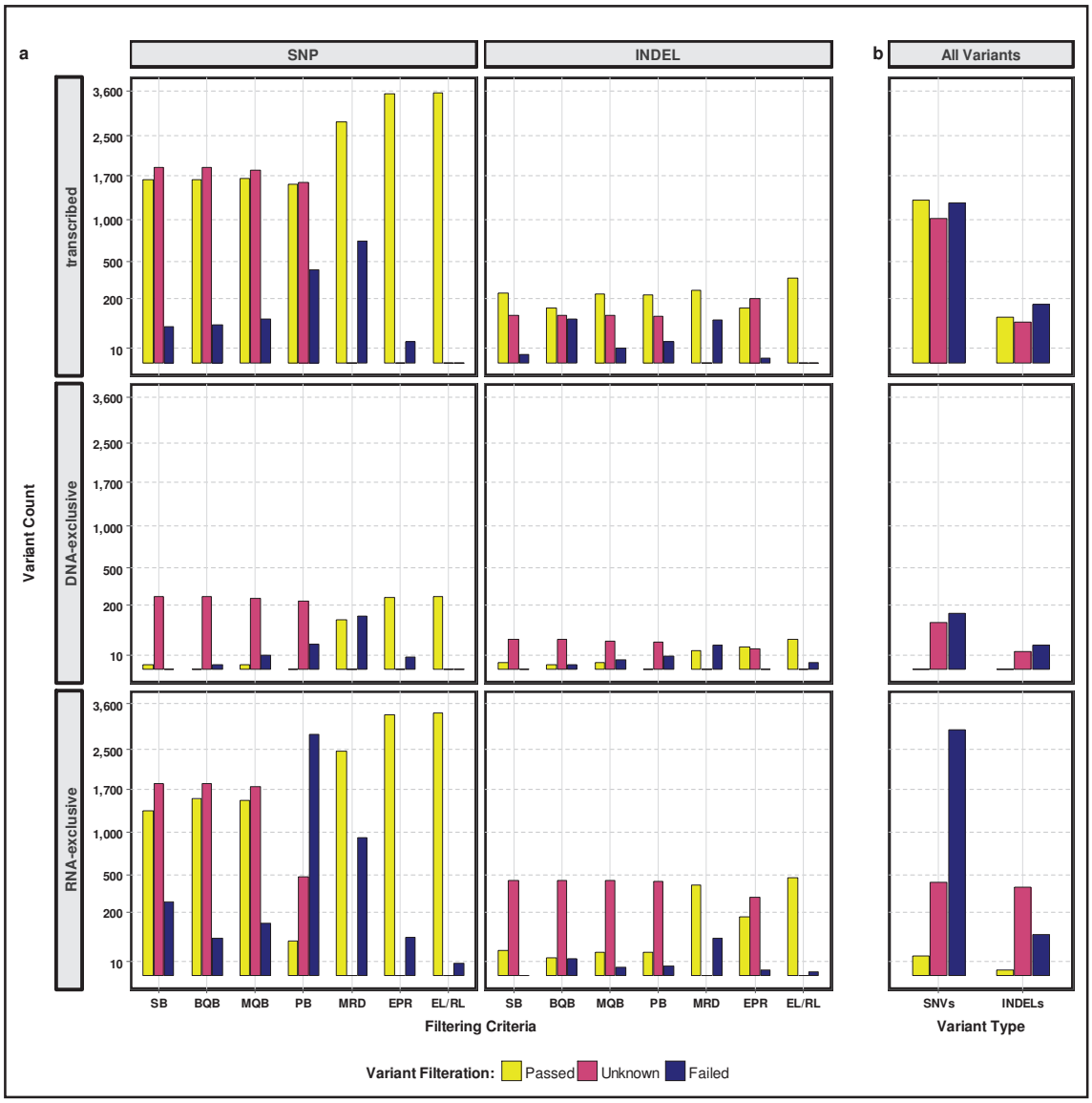

Figure 11: *Filtering criteria applied on called variants. (a) Filtering categories: SB – Strand Bias; BQB – Base Quality Bias; MQB – Mapping Quality Bias; PB – Position Bias; MRD – Minimum Read Depth; EPR – Error-prone Region; EL/RL – Edit Loci/Repeat Loci. (b) Summary of filtered variants for SNVs and INDELs separately. Variant filtration status was defined as 'Unknown' when the variants were unable to be subjected to defined filtering criteria mostly due to the absence of either forward or reverse stranded reads supporting each allele.*

3.6 DNA and RNA variant comparison

Due to the inherent differences in the sequencing techniques employed for both DNA and RNA, a huge difference in their mean coverage was observed in the recurrently mutated regions (**Table 2**). After excluding the artefacts, the VAFs among the transcribed and DNA-exclusive variants were compared (2606 variants). Based on the genotype information alone, 92.3% of the filtered variants showed no noticeable VAF change (**Figure 12**). The observed dispersion of

heterozygous DNA variants along the RNA VAF axis was due to the relatively low read coverage of the RNA-Seq when compared with the targeted DNA-Seq. A similar trend was found among the recurrent mutations in genes^{MUT} (83.5%). A 5.3% over-representation of mutated alleles in RNA-Seq was also observed in comparison to the DNA-Seq. However, 9.9% of recurrent mutation in RNA-Seq were unable to be detected in the DNA-Seq, indicating a lack of transcription, DNA degradation etc. (**Figure 13**).

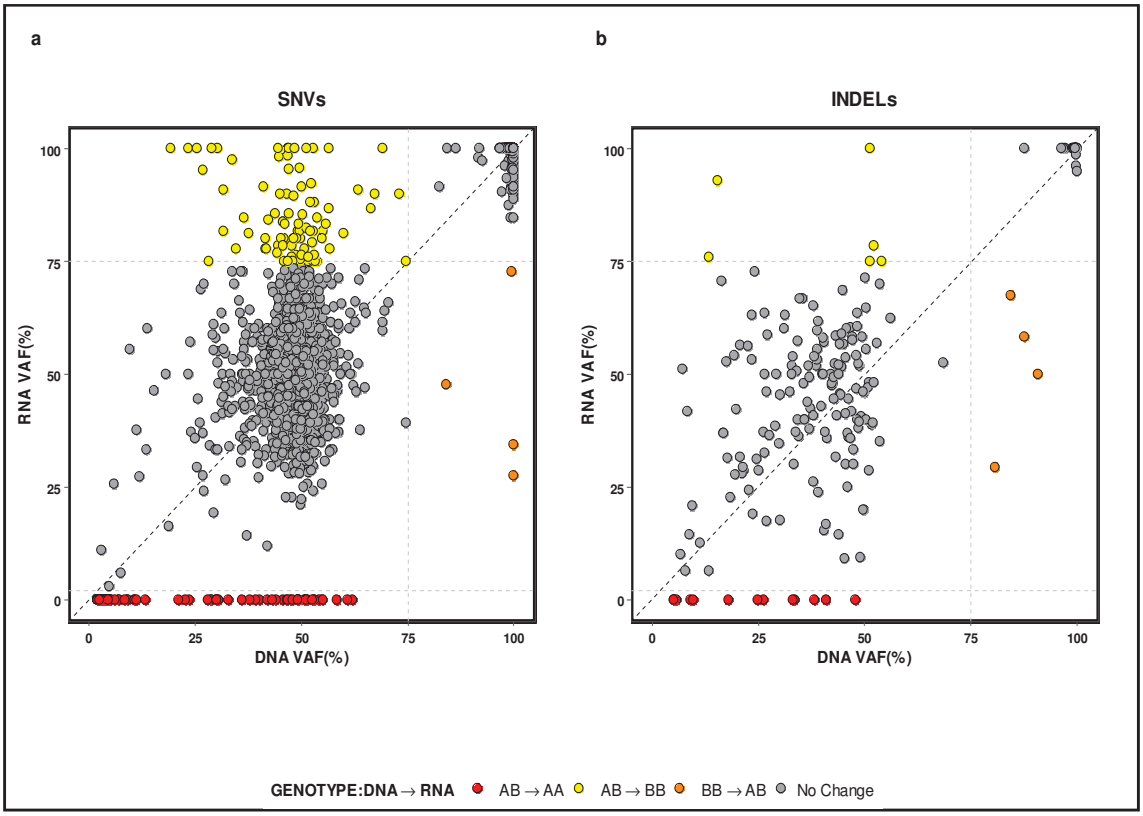

Figure 12: *Variant allele frequency differences of all variants between DNA and RNA for SNVs (a) and INDELs (b). The dotted diagonal lines represent the expected DNA vs RNA trend.*

3.7 Regression Analysis

The VAFs of DNA and RNA were transformed into expected and observed RNA variant read depths in order to account for the differences in their read coverage. The VAF distributions of SNVs and INDELs are shown in **Figure 13**. The regression model applied on both primary and validation cohorts are presented below.

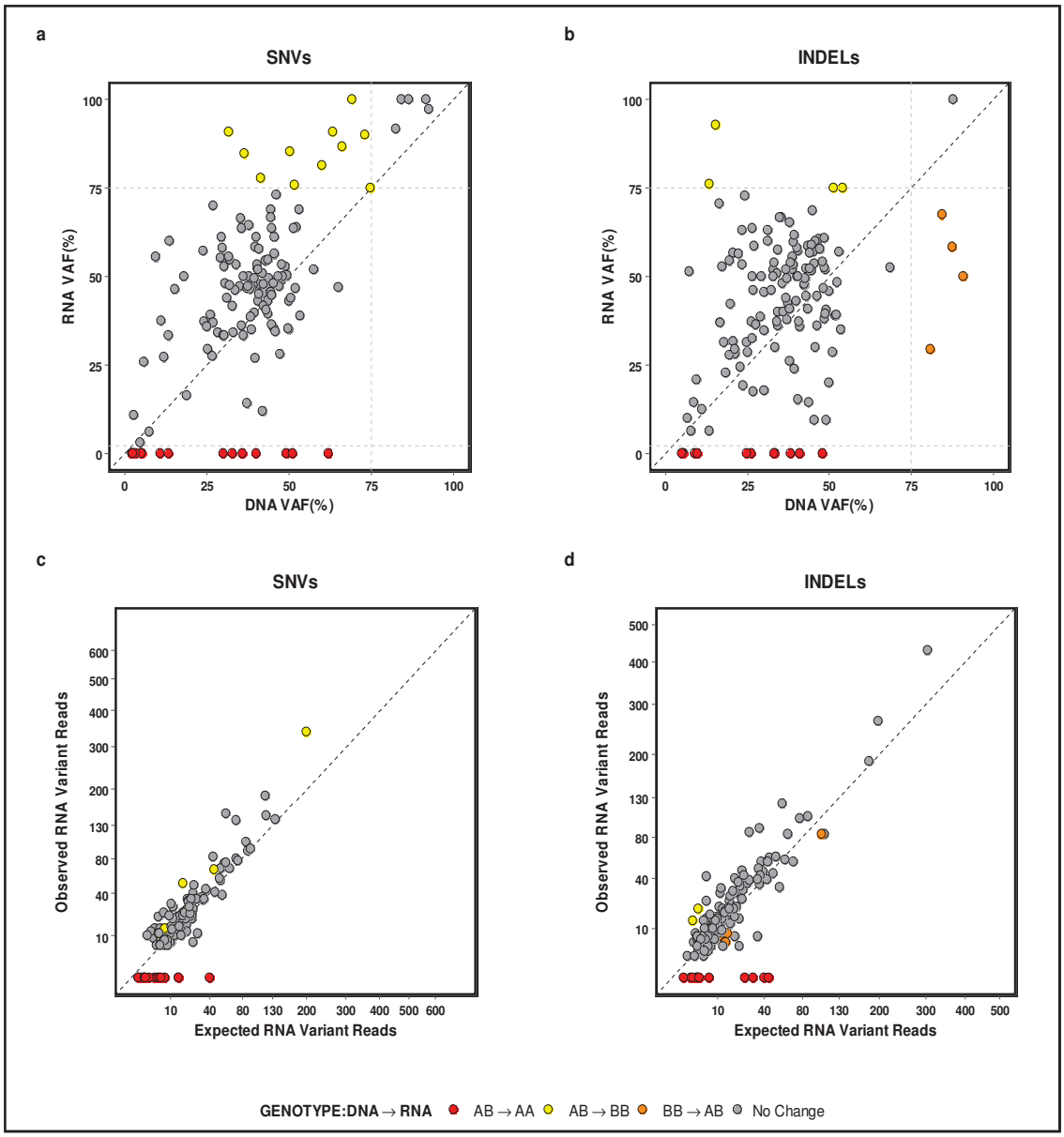

Figure 13: *Variant allele frequency differences of recurrent mutation between DNA and RNA for SNVs (a) and INDELs (b). Expected and observed RNA variant read depths of SNVs (c) and INDELs (d). The dotted diagonal lines represent the expected DNA vs. RNA trend. The genotype conversion of AB->AA and AB->BB represent the allele specific transcript abundance of wild-type and mutant allele, respectively. The observation of BB->AB genotype change might be due to the arbitrary definition of homozygous and heterozygous variants.*

3.7.1 Weighted allelic imbalance of genes^{MUT} and mutation types^{MUT} in the AMLCG cohort

The linear regression model used in the analyses estimate WAI and the substantial increase or decrease of WAI infer to preferential allelic transcript abundance of mutant and wild-type allele, respectively. The model was restricted to 11 genes from the initial consideration of 36 genes of interest due to the following reasons:

- Variants removed due to the application of filtering criteria (excluded: 8 genes)
- Common variants (dbSNP build 138 NonFlagged) were removed along with homozygous variants (excluded: 7 genes)
- Genes with less than five SNVs or INDELs were dropped off (excluded: 10 genes)

The exclusion of genes and samples from the analysis is depicted in the **Figure 3**. The model applied on SNVs showed a significant decrease in the WAI of mutant allele reads in *PTPN11*^{MUT}, whereas among genes such as *GATA2*^{MUT}, *RUNX1*^{MUT}, *TET2*^{MUT}, *SRSF2*^{MUT}, and *IDH2*^{MUT}, a substantial increase of the WAI was observed when compared to the expected values (**Figure 14**). In the case of INDELs, *CEBPA*^{MUT} and *WT1*^{MUT} demonstrated a considerable WAI decrease, whereas *NPM1*^{MUT} and *RUNX1*^{MUT} showed a significant increase in the number of reads supporting the mutant allele. In spite of many genes showing a significant change, the effect size was observed to be for *GATA2*^{MUT}, *CEBPA*^{MUT} and *WT1*^{MUT}. There were no AI observed among *U2AF1*^{MUT} and *FLT3*^{MUT} (which includes both *FLT3*-ITD and *FLT*-TKD variants). The analyses on the mutation types^{MUT} showed a significant increase in the WAI among non-synonymous SNV^{MUT} and frameshift INDELs^{MUT} and a reversal effect among non-frameshift insertions^{MUT}. Interestingly, stopgain SNVs^{MUT} exhibited no sign of AI among the variants.

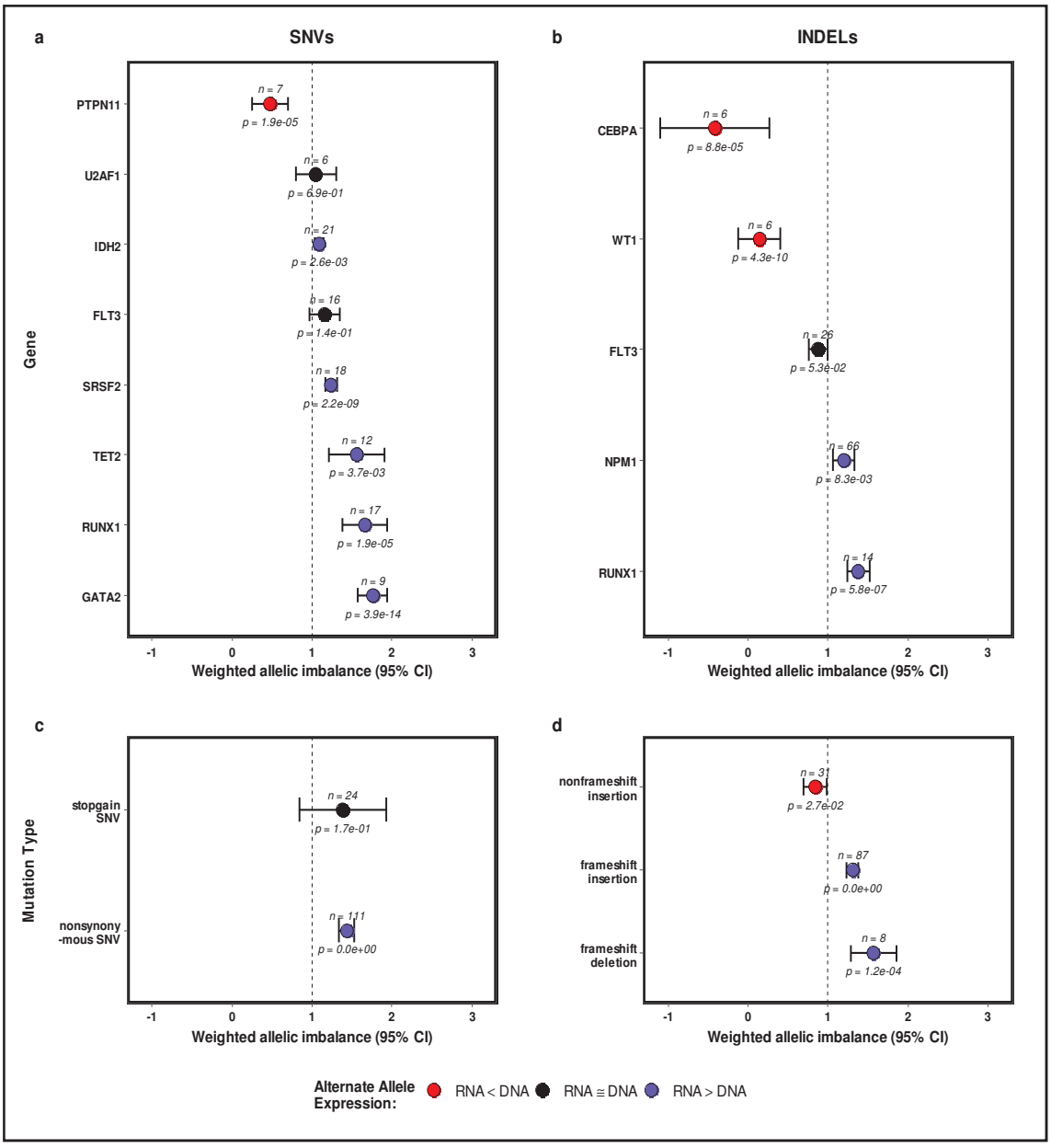

Figure 14: *WAI of recurrent mutations per gene^{MUT} in the AMLCG cohort for SNVs (a) and INDELs (b). WAI of recurrent mutations per mutation type^{MUT} in the AMLCG cohort for SNVs (c) and INDELs (d)*

3.7.2 Weighted allelic imbalance of genes^{MUT} in validation cohort

External data sets were used to independently validate the main findings. Due to the unavailability of large and comprehensive data sets at the current time point, I pooled all the mutations in those nine genes^{MUT} of interest from the DKTK, TCGA and the HELSINKI cohorts. The regression model was modified to account for the difference in the cohorts and applied on the pooled cohort. I was able to validate a significant increase in the WAI of *GATA2*^{MUT} (p-value = 4.9×10^{-7}) and a significant decrease in their WAIs of *WT1*^{MUT} (p-value = 3.8×10^{-3}) and *CEBPA*^{MUT} (p-value = 2.7×10^{-12}). These observations were consistent with the obtained results in

AMLCG, indicating preferential allelic transcript abundance (**Figure 15**). However, *NPM1*^{MUT} showed a substantial decrease in the WAI and thus preferring wile-type allelic transcript abundance. This in turn is in contradiction with the initial findings. In addition, genes such as *SRSF2*^{MUT}, *RUNX1*^{MUT}, *IDH2*^{MUT}, *TET2*^{MUT} and *PTPN11*^{MUT} were unable to be validated, most of which might be due to their relatively small effect size.

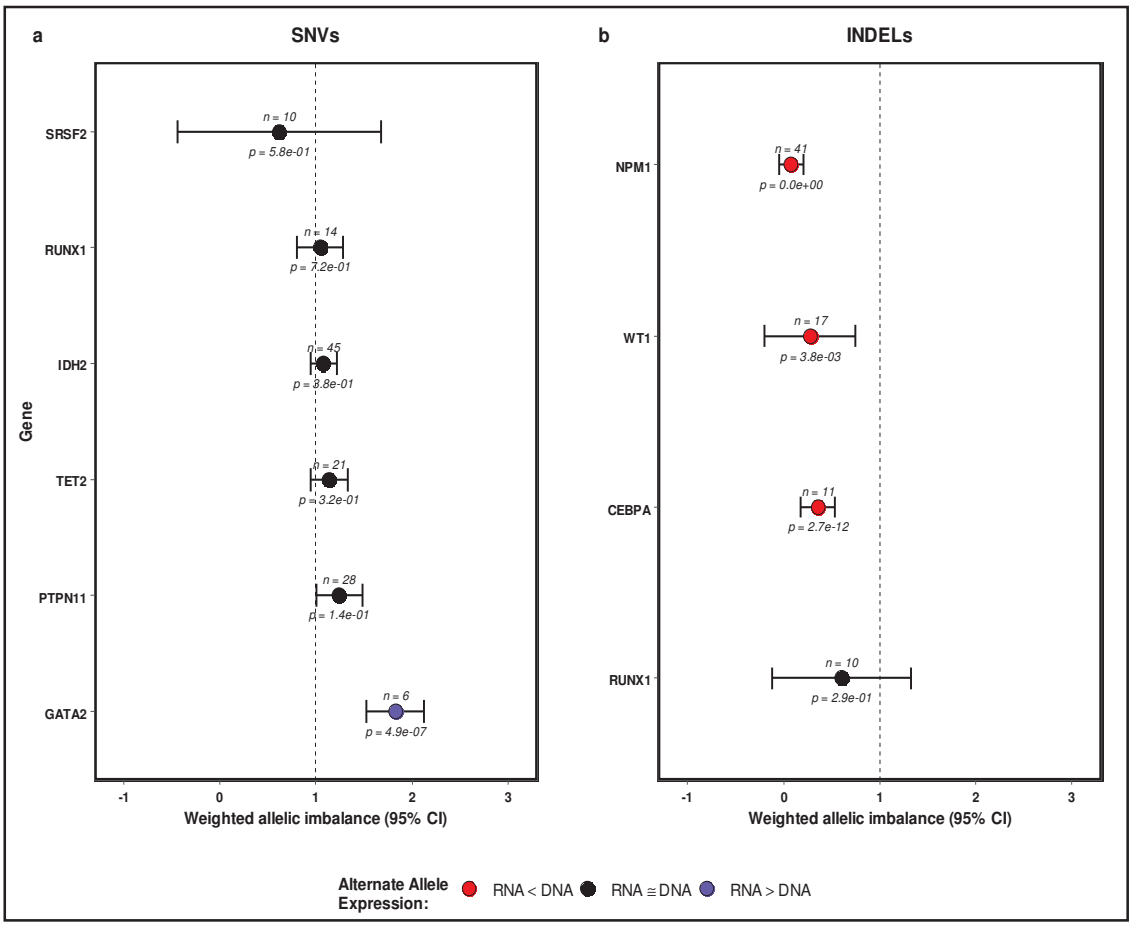

Figure 15: *Weighted allelic imbalance of recurrent mutations per gene*^{MUT} *among DKTK, TCGA and HELSINKI cohorts for SNVs (a) and INDELs (b).*

3.7.3 Weighted allelic imbalance of genes^{WT} based on SNPs

In order to determine if the allele-specific transcript abundance generally exist among the variants in the nine significant genes of interest from the previous section (*SRSF2*, *RUNX1*, *IDH2*, *TET2*, *PTPN11*, *GATA2*, *NPM1*, *WT1* and *CEBPA*), I performed the WAI analysis based on the common SNPs in those genes. All the SNPs in these genes were pooled from the AMLCG, the DKTK and TCGA cohorts which did not have recurrent mutations in the respective genes^{WT}. The potential sequence artefacts were filtered out as described previously. All the SNPs which

were dbSNP annotated (build 138, NonFlagged) in the genes^{WT} of interest were filtered out and the WAI analysis was carried out (**Figure 16**). The analysis was restricted to genes with ≥5 SNPs, similar to the previous analyses. As expected, no evidence of AI was observed for *WT1*^{WT}, *TET2*^{WT} and *SRSF2*^{WT}, whereas SNPs in *GATA2*^{WT}, *RUNX1*^{WT} and *IDH2*^{WT} showed significant WAI indicating the existence of AI in general among the AML patients. Interestingly, all three genes^{WT} showed preferential major allelic transcript abundance, which was reversed when compared to the genes^{MUT} with recurrent mutations. However, the effect of genes^{WT} was rather small in comparison to the genes^{MUT}.

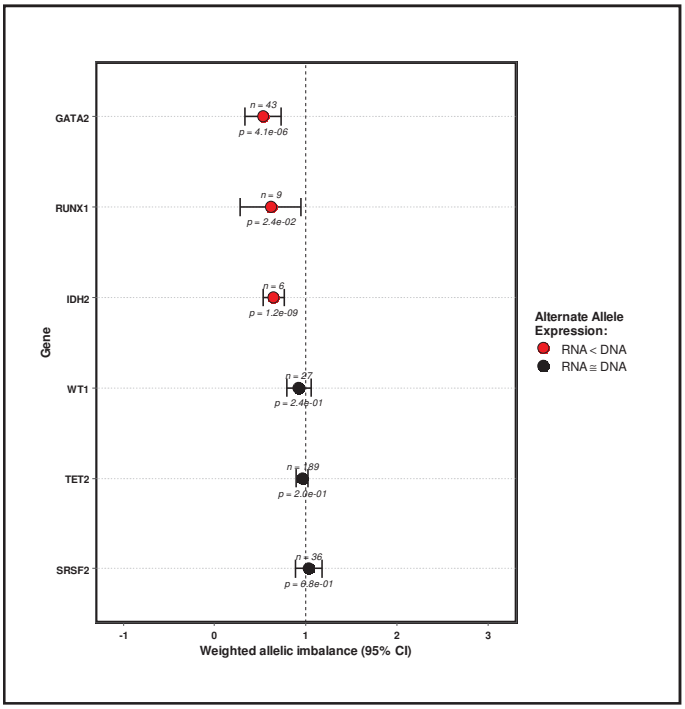

Figure 16: *Weighted allelic imbalance of common SNPs in AMLCG, DKTK and HELSINKI cohorts without recurrent mutations in the respective genes.*

3.8 Internal validation for allelic imbalance

The differential gene expression between patients harbouring recurrent mutations in respect to genes^{MUT} and patients with wild-type status (genes^{WT}) showed no significant difference in the nine genes of interest with the exception of *CEBPA* (**Figure 17**). In the case of profiling transcript isoforms, one transcript isoform exhibited differentially expression in each of *CEBPA*, *WT1* and *SRSF2*. However, the presence of mutations was observed not only in these transcript isoforms. Other transcript isoforms, which did not show any noticeable difference in their expression levels, also harboured the recurrent mutations in the case of *WT1* and

SRSF2. Thus, there was no conclusive evidence of the effect of recurrent mutations on differential expression of *WT1* and *SRSF2*. We were unable to compare the presence of mutations among other *CEBPA* transcript isoforms because of their low read counts (filtered out).

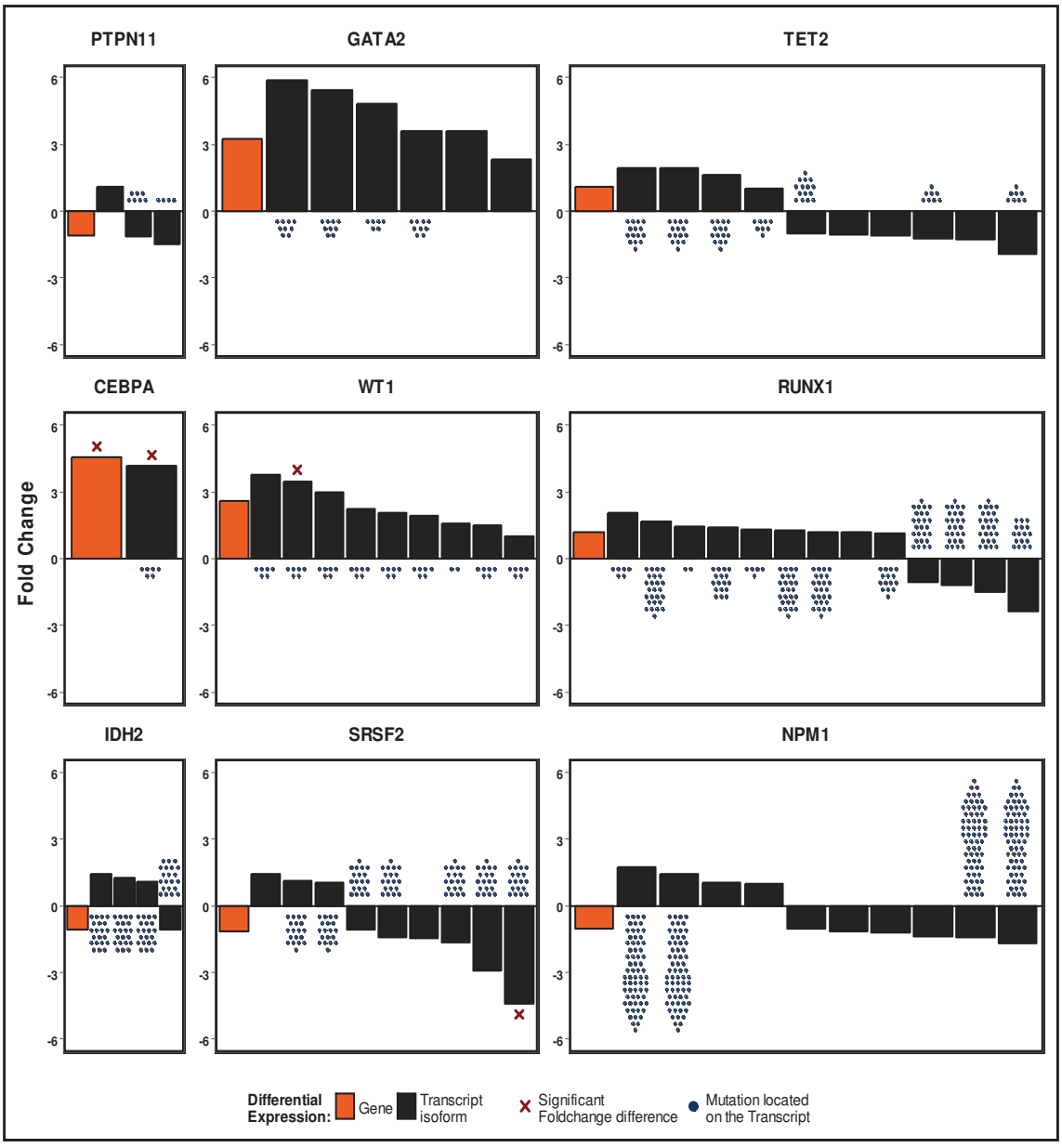

Figure 17: *Gene-level and transcript-level differential expression calculated with limma after precision-weighting with voom for all recurrently mutated genes with a significant WAI in the AMLCG cohort. Dots below or above the bars represent recurrent mutations present inside the transcripts. Crosses represent significant fold change differences.*

4 Discussion

The aim of my thesis project was the systematic comparison and characterisation of the transcription of variants from DNA to RNA. To this aim, I improved the conventional variant calling pipeline for the processing both targeted DNA-Seq and whole transcriptome RNA-Seq and defined a set of variant filtering criteria to eliminate potential sequence artefacts. I applied this pipeline on a large AML cohort (N=499) with the matched DNA- and RNA-Seq, and analysed the extent to which the recurrent mutations in AML are transcribed from DNA to RNA.

Several studies have proposed workflows for the alignment of sequencing reads and calling variants.[3,11,97,138] I examined different alignment and variant calling algorithms for their performance and constructed an optimized pipeline for DNA- and RNA-Seq. Default parameters were used to evaluate the aligners' performance in terms of accuracy and the time taken for the process. Despite the observations on BWA-MEM and STAR's better performance, optimizing aligner parameters based on sequence coverage, read length, bin range etc. might greatly increase the overall performance of any aligner.[139] The aligned reads from DNA- and RNA-Seq were processed separately due to the missing intron regions in RNA-Seq. In the case of the variant caller comparison, many previous studies showed the proportion of commonly called INDELs to be very low when compared to the SNVs detected.[140–142] Hasan *et al.* analysed seven INDEL callers including VarScan and showed a large number of known INDELs were still remained undetected in their analysis.[140] In this study, VarDict performed well when compared to VarScan in detecting INDELs which is consistent with Lai *et al.*'s systematic comparison of different variant callers, although there was not much difference in the case of SNV detection.[116] The variant calling parameters were optimized for DNA- and RNA-Seq, respectively, based on the alignment accuracy, read depth, mapping and base quality followed by variant filtration.

The comparison of weighted VAFs between DNA and RNA showed AI in nine genes^{MUT}, recurrently mutated in the homogenously treated primary cohort. Among the six genes with substantial increase in the WAI ($p < 0.05$), *GATA2*^{MUT} showed the largest effect size. This in turn infer to an increased mutant-allele specific transcript abundance of *GATA2*^{MUT}. A similar trend of mono-allelic

expression towards the mutant allele in *GATA2* was also observed by Al Seraihi and colleagues.[143] Celton *et al.* studied the normal karyotype AML samples and observed preferential mutant allelic expression among low-*GATA2*-expressing specimens.[104] In addition, they demonstrated that the hypermethylation of the silenced allele can be reversed by the exposure to demethylating agents and thus suggesting the requirement of DNA methylation for ASE of *GATA2*.[104] Another transcriptional analysis in mouse model demonstrated the down-regulation of *GATA2* to be a crucial step in the progression of leukaemia.[144] The term minimal residual disease refers to a small number of leukaemic cells that remains in the person during treatment or after in the remission stage. The differences in the preferential allelic transcript abundance in some cases might suggest RNA to be a better source for the measurement of minimal residual disease than DNA, especially in the case of *NPM1*.[145] Krönke *et al.* defined the time points for monitoring *NPM1* transcript levels and MRD assessment for the identification of high risk of relapse among the AML patients.[146] I was able to independently validate the effect of *GATA2*^{MUT}, *WT1*^{MUT} and *CEBPA*^{MUT} using the pooled validation cohort, irrespective of the differences in the sequencing techniques and differences in the population under study. Unfortunately, large and homogenous cohorts of AML patients with matched DNA- and RNA-Seq data were not available at this point. Therefore, the direct comparison of the results of the primary discovery cohort and the pooled validation cohorts was difficult. The existance of AI was not observed for *SRSF2*^{MUT}, *RUNX1*^{MUT}, *IDH2*^{MUT}, *TET2*^{MUT} and *PTPN11*^{MUT} in the validation cohort. However, *NPM1*^{MUT} in our validation cohort showed the AI towards the wild-type allele, which is in contradiction to our finding in the primary cohort. This does not prove the absence of AI as 'no evidence for a difference is not an evidence for no difference'. Most of these discrepancies might be due to their smaller effect size in the primary cohort and thus would require a larger number of mutations to be analysed in order to gain enough power for the analysis. Also bias might be introduced due to the different sequencing techniques used in the validation cohort.

The largest study regarding this topic in AML was conducted by Ley and colleagues in TCGA.[9] The allelic expression biases were detected by them among *RUNX1*, *TET2* and *WT1* mutated patients, along with three other genes in AML samples.[9]

Although some of my results might be associated with copy number alterations, genomic imprinting, uniparental disomy or differences in RNA half-life between the mutated and wild type allele, we claim the significant change in the WAI might be due to allele-specific transcript abundance which might be a common phenomenon among genes frequently altered by mutations in AML. At least, some of the findings could be validated in external independent data sets which points toward an additionally biologic regulative mechanism which might be associated with leukaemogenesis.[147] It is difficult to simulate these small expression changes in an *in vitro* model and larger studies including other disease are necessary to draw a conclusive picture. However, it is highly likely that this phenomenon contributes to leukaemogenesis. Further investigation of WAI based on common SNPs in *WT1*^{WT}, *TET2*^{WT} and *SRSF2*^{WT} revealed no AI among AML samples without any recurrent mutations in the respective genes^{WT}. This in turn implies an indirect association of AI with the recurrent mutations harboured in these genes. We also observed a significant WAI among genes^{WT} towards major allele transcript abundance and an AI shift in the opposite direction, towards mutant allele among genes^{MUT}. One interpretation might be due presence of functional variations in the *cis*-regulatory regions of these genes.[148] Adoue et al.' work on multiple cell lines, identified *cis*-regulatory variants based on mapping the differential allelic expression and reported 40-60% of these variants are shared across all cell types.[148] Despite the observations of the WAI of recurrent mutations and common SNPs among genes^{MUT} and genes^{WT}, respectively, I was unable to detect any consistent impact in the differential expression of genes and transcript isoforms between patients with and without recurrent mutations. Especially among the significant differentially expressed transcript isoforms, the mutations harboured in them were also found in the isoforms which did not show any significant expression. Thus, it was unclear to interpret the sole attribution of the harboured mutations and the transcript-level differential expression. However, a differential expression of transcript isoforms could be observed in three genes, which might suggest an inadequate increase of mutant alleles or, in the case of SRSF2, an effect of counteracting mechanisms regarding preferential wild-type allelic transcription. Further analysis by grouping the mutations based on mutation type showed frameshift INDELs^{MUT} to have an enhanced mutant allele expression. This in turn is in contradiction with Rhee *et al.*'s demonstration of a negative allelic fraction

difference.[8] In addition, we were also unable to validate their negative allelic fraction difference among stopgain SNV^{MUT}.[8] However, Rhee *et al.* determined the AI of somatic mutations using five different tumour sample types (BRCA, HNSC, KIRC, LUAD and STAD) from TCGA (excluding AML). Thus the discrepancy might be due to the difference in the tumour types under study. In addition, they compared the RNA-Seq with the WES whereas we used the targeted DNA-Seq and this might have also been contributed to the inconsistancies.[8] The discovery cohort consists of homogeneously treated AML patient samples and the differences in the findings might suggest the observed variation in the allele-specific transcript abundance might depend on the tumour entity under study.

This analysis on quantifying imbalances in a uniform cohort reduces ascertainment bias and thus improves the validity of the results.[149,150] Previous studies compared the AI in terms of allele fraction difference (AI = RNA VAF minus DNA VAF).[7,8] Thus, the effect was determined when the difference is not equal to zero with a defined cut-off. It was not suitable to use this approach directly, since I compared the targeted DNA-Seq (with relatively high sequence coverage) to the RNA-Seq. In order to overcome this issue, the VAFs of both sequences were transformed into expected and observed mutant allele reads. In this approach, the AI was determined when the difference is not equal to 1 (RNA VAF = $\alpha \times$ DNA VAF). This weighted approach ensures comparability between two data sets with huge differences in read coverage (542 fold in the DNA-Seq vs. 85 fold in the RNA-Seq). The exploratory analysis was based on classical linear regression (excluding the intercept), assuming normal homoscedastic distribution of residuals. It was not possible to prove a non-normal distribution of residuals due to the low number of sample points per gene. The exact false detection rate was not known and due to the assumption it fell below the significance level (5% in our case). I addressed this issue by performing validation of our results on independent cohorts using the same algorithm.

One of the major limitations in using whole RNA-Seq for variant calling is the inherent low coverage in the regions of interest when compared to the targeted DNA-Seq. We were able to detect the majority of the genomic variants (95.4%) which is in accordance with the previous publication, although the frequencies of the variants varied considerably.[99] Nevertheless, employing variant discovery

solely in RNA-Seq could not be recommended due to the large number of potential false positive variant calls (>52% in our analysis). Also, the variant discovery in RNA-Seq also depends on the expression of those regions in the genes. Less stringent parameters were used for calling variants in RNA-Seq to avoid premature filtration of putative variants. One of the important aspects of this study is to eliminate potential artefacts and achieve enriched variants. Selection of ideal read depth cut-off is essential for any variant filtering procedure. I optimized the parameters for variant calling in DNA-Seq and then visualized the concordance rate of homozygous and heterozygous variants with respect to incremental variant read depth. By assuming similar proportions of homozygous and heterozygous variants for any given read depth, I observed a convergence of proportions at the read depth of 10x. The proportions remain stable for higher read depths, showing a reliable cut-off for RNA-Seq, which stands in agreement with the Ley *et al.*'s defined cut-off (TCGA).[9] Similarly, Quinn and colleagues showed 89% specificity in calling SNPs in RNA-Seq with a read cut-off of 10x, although different sequence aligner (TopHat) and variant caller (GATK and SAMtools) was employed.[5] Defining a reliable read cut-off was possible among the SNV due to the detection of large number of SNVs per RNA read depth. However, same was not possible among INDELs due to their lower numbers per read depth in the RNA-Seq. The application of other filtering criteria and the proportions of potential artefacts suggested the importance of the defined criteria. Variant calling and variant discovery in RNA-Seq are hampered by several factors including read coverage, expression of the genes, RNA-editing sites, repeat regions and regions nearby splice junctions. However, application of the proposed filtering criteria might assist in enriching the called variants in the RNA-Seq.

5 Conclusion

In this study, I compared commonly employed sequence aligners and variant callers to construct a variant calling pipeline for targeted DNA- and RNA-Seq. Several filtering criteria were defined to remove potential artefacts by considering varying read depths, error-prone regions, edit and repeat loci, biases due sequence strand, position of the variant, mapping and base quality. I determined the extent of allelic proportion of recurrently mutated genes^{MUT} being transcribed from DNA to RNA and the existence of AI among them in AML. Allele-specific transcript abundance of *GATA2*^{MUT}, *WT1*^{MUT} and *CEBPA*^{MUT} were detected and validated in independent cohorts. The presence of AI in general was also demonstrated among some genes^{WT} in patients who did not harbour any recurrent mutations in the respective genes. This study acquaints the notion of preferential wild-type or minor allelic transcript abundance to be a common and underestimated mechanism in the pathogenesis of AML. Further research will be required to determine any association of allele-specific transcript abundance with the epigenetic mechanisms inside the intricate pathomechanisms of recurrent mutations in AML.

Scientific activities and collaboration projects

Manuscript in preparation

Aarif M. N. Batcha, Stefan A. Bamopoulos, Paul Kerbs, Vindi Jurinovic, Maja Rothenberg-Thurley, Bianka Ksienzyk, Julia Phillippou-Massier, Stefan Krebs, Helmut Blum, Stephanie Schneider, Nikola Konstandin, Stefan K Bohlander, Wolfgang Hiddemann, Karsten Spiekermann, Jan Braess, Ashwini Kumar, Mika Kontro, Klaus H Metzeler, Philipp A. Greif, Ulrich Mansmann and Tobias Herold. "Weighted Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia"
This thesis work will be submitted for publication

Published papers

Li, J., **Batcha, A.M.N.**, Gaining, B. and Mansmann, U.R., 2015. An NGS workflow blueprint for DNA sequencing data and its application in individualized molecular oncology. *Cancer informatics, 14*, pp.CIN-S30793.

Herold, T., Jurinovic, V., **Batcha, A.M.**, Bamopoulos, S.A., Rothenberg-Thurley, M., Ksienzyk, B., Hartmann, L., Greif, P.A., Phillippou-Massier, J., Krebs, S. and Blum, H., 2017. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. *Haematologica*, pp.haematol-2017.

Presentation paper

Aarif M. N. Batcha, Stefanos A. Bamopoulos, Bianka Ksienzyk, Vindi Jurinovic, Nikola P. Konstandin, Stefan Krebs, Helmut Blum, Carola Christ, Wolfgang Hiddemann, Michael von Bergwelt-Baildon, Michael Walter, Stephanie Schneider, Karsten Spiekermann, Klaus H. Metzeler, Nicola Gökbuget and Tobias Herold. "Custom targeted RNA sequencing for the classification of BCR-ABL1-like acute lymphoblastic leukemia and the identification of druggable mutations and fusions"
Presented at the EHA-SWG Scientific Meeting on New Molecular Insights and Innovative Management Approaches for Acute Lymphoblastic Leukemia 2018

Submitted papers

Roman Hornung, Vindi Jurinovic, **Aarif M. N. Batcha**, Stefanos A. Bamopoulos, Maja Rothenberg-Thurley, Susanne Amler, Maria Cristina Sauerland, Wolfgang E. Berdel, Bernhard J. Woermann, Stefan K. Bohlander, Jan Braess, Wolfgang Hiddemann, Sören Lehmann, Sylvain Mareschal, Karsten Spiekermann, Klaus H Metzeler, Tobias Herold and Anne-Laure Boulesteix. "Mediation analysis reveals common mechanisms of RUNX1 point mutations and RUNX1/RUNX1T1 fusions influencing survival of patients with acute myeloid leukemia".
Under revision in Scientific Reports

Katrin Schranz, Max Hubmann, Egor Harin, Sebastian Vosberg, Tobias Herold, Klaus H. Metzeler, Maja Rothenberg-Thurley, Hanna Janke, Kathrin Bräundl, Bianka Ksienzyk, **Aarif M. N. Batcha**, Sebastian Schaaf, Stephanie Schneider, Stefan K. Bohlander, Dennis Görlich, Wolfgang E. Berdel, Bernhard J. Wörmann, Jan Braess, Stefan Krebs, Wolfgang Hiddemann, Ulrich Mansmann, Karsten Spiekermann, Philipp A. Greif. "Clonal heterogeneity of FLT3-ITD detected by high throughput amplicon sequencing correlates with adverse prognosis in acute myeloid leukemia"
Accepted for publication by Oncotarget

References

1. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkologia* **1A,** A68–A77 (2015).
2. Zhang, J. *et al.* International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database* **2011,** (2011).
3. Pirooznia, M. *et al.* Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum. Genomics* **8,** 14 (2014).
4. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17,** 13 (2016).
5. Quinn, E. M. *et al.* Development of Strategies for SNP Detection in RNA-Seq Data: Application to Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. *PLoS One* **8,** e58815 (2013).
6. Chakravarthi, B. V. S. K., Nepal, S. & Varambally, S. Genomic and Epigenomic Alterations in Cancer. *American Journal of Pathology* **186,** 1724–1735 (2016).
7. Castle, J. C. *et al.* Mutated tumor alleles are expressed according to their DNA frequency. *Sci. Rep.* **4,** 4743 (2015).
8. Rhee, J.-K., Lee, S., Park, W.-Y., Kim, Y.-H. & Kim, T.-M. Allelic imbalance of somatic mutations in cancer genomes and transcriptomes. *Sci. Rep.* **7,** 1653 (2017).
9. Ley, T. J. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).
10. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74,** 5463–5467 (1977).
11. Li, J., Nazeer Batcha, A. M., Grüning, B. & Mansmann, U. R. An NGS workflow blueprint for DNA sequencing data and its application in individualized molecular oncology. *Cancer Informatics* **15,** 87–107 (2016).
12. Mardis, E. R. DNA sequencing technologies: 2006-2016. *Nat. Protoc.* **12,** 213–218 (2017).
13. Thermes, C. Ten years of next-generation sequencing technology. *Trends in genetics : TIG* **30,** 418–426 (2014).
14. Sur, I. & Taipale, J. Transcription regulation and animal diversity. *Nat Rev. Cancer* **424,** 147–151 (2016).
15. Pastinen, T. & Hudson, T. J. Cis-acting regulatory variation in the human genome. *Science* (2004). doi:10.1126/science.1101659
16. Jones, P. A. *et al.* Moving AHEAD with an international human epigenome project. *Nature* (2008). doi:10.1038/454711a
17. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* (2005). doi:10.1038/nature03479
18. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science (80-.).* **318,** 1136–1140 (2007).
19. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science (80-.).* **343,** 193–196 (2014).
20. Yu, Y. *et al.* NOEY2 (ARHI), an imprinted putative tumor suppressor gene in ovarian and breast carcinomas. *Med. Sci.* **96,** 214–219 (1999).
21. Yu, Y. *et al.* Biochemistry and Biology of ARHI (DIRAS3), an Imprinted Tumor Suppressor Gene Whose Expression Is Lost in Ovarian and Breast Cancers. *Methods in Enzymology* (2005). doi:10.1016/S0076-6879(05)07037-0
22. Chávez, S., García-Martínez, J., Delgado-Ramos, L. & Pérez-Ortín, J. E. The importance of controlling mRNA turnover during cell proliferation. *Current Genetics* (2016). doi:10.1007/s00294-016-0594-2
23. Heck, A. M. & Wilusz, J. The interplay between the RNA decay and translation machinery in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **10,** 1–20 (2018).

24. Den Dunnen, J. T. & Antonarakis, E. Nomenclature for the description of human sequence variations. *Human Genetics* **109,** 121–124 (2001).
25. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
26. Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome An initial map of insertion and deletion (INDEL) variation in the human genome. 1182–1190 (2006). doi:10.1101/gr.4565806
27. Rahim, N. G., Harismendy, O., Topol, E. J. & Frazer, K. A. Genetic determinants of phenotypic diversity in humans. *Genome Biology* **9,** (2008).
28. Freeman, J. L. *et al.* Copy number variation: New insights in genome diversity. *Genome Research* (2006). doi:10.1101/gr.3677206
29. Fernandes, I. R. *et al.* Genetic variations on SETD5 underlying autistic conditions. *Developmental Neurobiology* (2018). doi:10.1002/dneu.22584
30. Masoodi, T. A., Banaganapalli, B., Vaidyanathan, V., Talluri, V. R. & Shaik, N. A. Computational Analysis of Breast Cancer GWAS Loci Identifies the Putative Deleterious Effect of STXBP4 and ZNF404 Gene Variants. *J. Cell. Biochem.* **118,** 4296–4307 (2017).
31. Brown, D. K. & Tastan Bishop, Ö. Role of Structural Bioinformatics in Drug Discovery by Computational SNP Analysis: Analyzing Variation at the Protein Level. *Global Heart* **12,** 151–161 (2017).
32. Bhattacharya, R., Rose, P. W., Burley, S. K. & Prlić, A. Impact of genetic variation on three dimensional structure and function of proteins. *PLoS One* **12,** (2017).
33. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
34. Forbes, S. A. *et al.* COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805–D811 (2015).
35. Landrum, M. J. *et al.* ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–D985 (2014).
36. Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biology* (2010). doi:10.1186/gb-2010-11-5-r52
37. Ni, X. *et al.* Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci. U. S. A.* (2013). doi:10.1073/pnas.1320659110
38. Garcia-Perez, J. L. *et al.* The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. *Microbiol. Spectr.* (2015). doi:10.1128/microbiolspec.MDNA3-0061-2014
39. Harshey, R. M. The Mu story: how a maverick phage moved the field forward. *Mob. DNA* (2012). doi:10.1186/1759-8753-3-21
40. Burrows, M. & Wheeler, D. J. A block-sorting lossless data compression algorithm. *Syst. Res.* **Research R,** 24 (1994).
41. Ferragina, P. & Manzini, G. Opportunistic data structures with applications. *Proceeding FOCS '00 Proc. 41st Annu. Symp. Found. Comput. Sci. FOCS '00 Proc. 41st Annu. Symp. Found. Comput. Sci.* 390–398 (2000). doi:10.1109/SFCS.2000.892127
42. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: Short oligonucleotide alignment program. *Bioinformatics* **24,** 713–714 (2008).
43. Liu, Y., Schmidt, B. & Maskell, D. L. Cushaw: A cuda compatible short read aligner to large genomes based on the burrows-wheeler transform. *Bioinformatics* **28,** 1830–1837 (2012).
44. Eaves, H. L. & Gao, Y. MOM: Maximum oligonucleotide mapping. *Bioinformatics* **25,** 969–970 (2009).
45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013). doi:arXiv:1303.3997 [q-bio.GN]
46. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

47. Liu, Y., Popp, B. & Schmidt, B. CUSHAW3: Sensitive and accurate base-space and color-space short-read alignment with hybrid seeding. *PLoS One* **9,** (2014).
48. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5,** (2016).
49. Kim, D. *et al.* TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** (2013).
50. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–60 (2015).
51. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
52. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–72 (2010).
53. Meacham, F. *et al.* Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12,** (2011).
54. BroadInstitute. Picard Tools - By Broad Institute. (2016). Available at: http://broadinstitute.github.io/picard/. (Accessed: 23rd April 2018)
55. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
56. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6,** 291–295 (2009).
57. Pireddu, L., Leo, S. & Zanetti, G. Seal: A distributed short read mapping and duplicate removal tool. *Bioinformatics* **27,** 2159–2160 (2011).
58. Xu, H. *et al.* FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS One* **7,** (2012).
59. DePristo, M. a. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (2011).
60. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012). doi:arXiv:1207.3907 [q-bio.GN]
61. Koboldt, D. C., Larson, D. E. & Wilson, R. K. Using varscan 2 for germline variant calling and somatic mutation detection. *Curr. Protoc. Bioinforma.* (2013). doi:10.1002/0471250953.bi1504s44
62. McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
63. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).
64. Chen, K. *et al.* BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–681 (2009).
65. Wang, J. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8,** 652–654 (2011).
66. Howlader, N. *et al.* Cancer Statistics Review, 1975-2014 - SEER Statistics, National Cancer Institute. *SEER Cancer Statistics Review, 1975-2014* http://seer.cancer.gov/csr/1975_2014/ (2016).
67. Shysh, A. C. *et al.* The incidence of acute myeloid leukemia in Calgary, Alberta, Canada: A retrospective cohort study. *BMC Public Health* **18,** (2017).
68. Hartmut Döhner, M.D., Daniel J. Weisdorf, M.D., and Clara D. Bloomfield, M. D., Döhner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N. Engl. J. Med.* **373,** 1136–52 (2015).
69. Li, S., Mason, C. & Melnick, A. Genetic and epigenetic heterogeneity in acute myeloid leukemia. *Current Opinion in Genetics and Development* **36,** 100–106 (2016).
70. Horton, S. J. & Huntly, B. J. P. Recent advances in acute myeloid leukemia stem cell biology. *Haematol.* | **97,** (2012).

71. Döhner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: Recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **115,** 453–474 (2010).
72. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* (2017). doi:10.1182/blood-2016-08-733196
73. Zarrinkar, P. P. *et al.* AC220 is a uniquely potent and selective inhibitor of FLT3 for the treatment of acute myeloid leukemia (AML). *Blood* **114,** 2984–2992 (2009).
74. Kayser, S. & Levis, M. J. Advances in targeted therapy for acute myeloid leukaemia. *Br. J. Haematol.* **180,** 484–500 (2018).
75. Poh, S. L. & Linn, Y. C. Immune checkpoint inhibitors enhance cytotoxicity of cytokine-induced killer cells against human myeloid leukaemic blasts. *Cancer Immunol. Immunother.* **65,** 525–536 (2016).
76. Laszlo, G. S. *et al.* Cellular determinants for preclinical activity of a novel CD33/CD3 bispecific T-cell engager (BiTE) antibody, AMG 330, against human AML. *Blood* (2014). doi:10.1182/blood-2013-09-527044
77. Bennett, J. *et al.* Proposals for the Classification of the Acute Leukaemias. *Br. J. Haematol.* **33,** 451–458 (1976).
78. Bennett, J. M. *et al.* Proposed revised criteria for the classification of acute myeloid leukemia. A report of the French-American-British Cooperative Group. *Ann. Intern. Med.* **103,** 620–625 (1985).
79. Swerdlow, S. H. *et al.* WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. Lyon, France. *World Heal. Organ. Calssification Tumours Haematop. Lymphoid Tissue* **4th,** 326 (2008).
80. Nishii, K. *et al.* Characteristics of t(8;21) acute myeloid leukemia (AML) with additional chromosomal abnormality: Concomitant trisomy 4 may constitute a distinctive subtype of t(8;21) AML. *Leukemia* **17,** 731–737 (2003).
81. Delaunay, J. *et al.* Prognosis of inv(16)/t(16;16) acute myeloid leukemia (AML): A survey of 110 cases from the French AML intergroup. *Blood* **102,** 462–469 (2003).
82. Wang, Y., Wu, N., Liu, D. & Jin, Y. Recurrent Fusion Genes in Leukemia: An Attractive Target for Diagnosis and Treatment. *Curr. Genomics* **18,** 378–384 (2017).
83. De Kouchkovsky, I. & Abdul-Hay, M. 'Acute myeloid leukemia: A comprehensive review and 2016 update'. *Blood Cancer Journal* **6,** (2016).
84. Gaidzik, V. I. *et al.* RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features. *Leukemia* (2016). doi:10.1038/leu.2016.126
85. Naoe, T. & Kiyoi, H. Gene mutations of acute myeloid leukemia in the genome era. *Int. J. Hematol.* **97,** 165–174 (2013).
86. Woods, B. A. & Levine, R. L. The role of mutations in epigenetic regulators in myeloid malignancies. *Immunol. Rev.* **263,** 22–35 (2015).
87. Metzeler, K. H. *et al.* Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. doi:10.1182/blood-2016-01
88. Lindsley, R. C. *et al.* Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. *Blood* **125,** 1367–1376 (2015).
89. Wouters, B. J. *et al.* Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* (2009). doi:10.1182/blood-2008-09-179895
90. Taskesen, E. *et al.* Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: Further evidence for CEBPA double mutant AML as a distinctive disease entity. *Blood* **117,** 2469–2475 (2011).
91. Godley, L. A. Inherited Predisposition to Acute Myeloid Leukemia. *Semin. Hematol.* **51,** 306–321 (2014).

92. Polprasert, C. *et al.* Inherited and Somatic Defects in DDX41 in Myeloid Neoplasms. *Cancer Cell* **27,** 658–670 (2015).
93. Engström, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10,** 1185–1191 (2013).
94. Liu, X., Han, S., Wang, Z., Gelernter, J. & Yang, B. Z. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS One* **8,** (2013).
95. Sandmann, S. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci. Rep.* **7,** (2017).
96. Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P. & Kocher, J.-P. A. Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. *Brief. Bioinform.* **18,** bbw069 (2016).
97. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* (2013). doi:10.1002/0471250953.bi1110s43
98. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **5,** 17875 (2015).
99. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93,** 641–651 (2013).
100. Tuch, B. B. *et al.* Tumor Transcriptome Sequencing Reveals Allelic Expression Imbalances Associated with Copy Number Alterations. doi:10.1371/journal.pone.0009317
101. Soh, J. *et al.* Oncogene mutations, copy number gains and mutant allele specific imbalance (MASI) frequently occur together in tumor cells. *PLoS One* (2009). doi:10.1371/journal.pone.0007464
102. O'Brien, T. D. *et al.* Inconsistency and features of single nucleotide variants detected in whole exome sequencing versus transcriptome sequencing: A case study in lung cancer. *Methods* **83,** 118–127 (2015).
103. Schrider, D. R., Gout, J.-F. & Hahn, M. W. Very few RNA and DNA sequence differences in the human transcriptome. *PLoS One* **6,** e25842 (2011).
104. Celton, M. *et al.* Epigenetic regulation of GATA2 and its impact on normal karyotype acute myeloid leukemia. *Leukemia* **28,** 1617–1626 (2014).
105. Herold, T. *et al.* A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. *Haematologica* haematol.2017.178442 (2017). doi:10.3324/haematol.2017.178442
106. Greif, P. A. *et al.* Evolution of cytogenetically normal acute myeloid leukemia during therapy and relapse: 1 An exome sequencing study of 50 patients 2 3. (2018). doi:10.1158/1078-0432.CCR-17-2344
107. Pemovska, T. *et al.* Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discov.* **3,** 1416–1429 (2013).
108. Kumar, A. *et al.* The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia. *BMC Genomics* **18,** (2017).
109. Girardot, C., Scholtalbers, J., Sauer, S., Su, S. Y. & Furlong, E. E. M. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics* **17,** (2016).
110. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
111. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/* http://www.bioinformatics.babraham.ac.uk/projects/ (2010). doi:citeulike-article-id:11583827
112. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32,** 3047–3048 (2016).
113. Barreto, H. & Howland, F. Lexogen. *Introd. Econom.* **001,** 1–9
114. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

115. Ahdesmaki, M. SplitNRead. Available at: https://github.com/mjafin/splitNreads.
116. Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44,** e108 (2016).
117. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
118. Ramaswami, G. & Li, J. B. RADAR: A rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.* **42,** (2014).
119. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
120. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* (2003). doi:10.1093/nar/gkg509
121. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* (2013). doi:10.1002/0471142905.hg0720s76
122. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods* (2010). doi:10.1038/nmeth0810-575
123. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
124. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).
125. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** (2014).
126. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).
127. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* (2012). doi:10.1038/nprot.2012.016
128. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** (2011).
129. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3519
130. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14,** 417–419 (2017).
131. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32,** 462–464 (2014).
132. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** (2014).
133. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44,** W3–W10 (2016).
134. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria* **0,** {ISBN} 3-900051-07-0 (2017).
135. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).
136. Bazak, L. *et al.* A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* **24,** 365–376 (2014).
137. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).
138. Ruffalo, M., Laframboise, T. & Koyutürk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27,** 2790–2796 (2011).

139. Zhang, J., Lin, H., Balaji, P. & Feng, W. C. Optimizing Burrows-Wheeler Transform-based sequence alignment on multicore architectures. in *Proceedings - 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, CCGrid 2013* 377–384 (2013). doi:10.1109/CCGrid.2013.67
140. Hasan, M. S., Wu, X. & Zhang, L. Performance evaluation of indel calling tools using real short-read data. *Hum. Genomics* **9,** 20 (2015).
141. Ghoneim, D. H., Myers, J. R., Tuttle, E. & Paciorkowski, A. R. Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. *BMC Res. Notes* **7,** (2014).
142. Neuman, J. A., Isakov, O. & Shomron, N. Analysis of insertion-deletion from deep-sequencing data: Software evaluation for optimal detection. *Brief. Bioinform.* **14,** 46–55 (2013).
143. Al, A. F. *et al.* GATA2 monoallelic expression underlies reduced penetrance in inherited GATA2 -mutated MDS / AML. *Leukemia* 2–7 doi:10.1038/s41375-018-0134-9
144. Bonadies, N. *et al.* Genome-Wide Analysis of Transcriptional Reprogramming in Mouse Models of Acute Myeloid Leukaemia. *PLoS One* (2011). doi:10.1371/journal.pone.0016330
145. Schnittger, S. *et al.* Minimal residual disease levels assessed by NPM1 mutation-specific RQ-PCR provide important prognostic information in AML. *Blood* **114,** 2220–2231 (2009).
146. Krönke, J. *et al.* Monitoring of Minimal Residual Disease in NPM1 -Mutated Acute Myeloid Leukemia: A Study From the German-Austrian Acute Myeloid Leukemia Study Group. *J. Clin. Oncol.* **29,** 2709–2716 (2011).
147. Chen, J., Odenike, O. & Rowley, J. D. Leukaemogenesis: More than mutant genes. *Nature Reviews Cancer* **10,** 23–36 (2010).
148. Adoue, V. *et al.* Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs. *Mol. Syst. Biol.* (2014). doi:10.15252/msb.20145114
149. Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15,** 1496–1502 (2005).
150. Haasl, R. J. & Payseur, B. A. Multi-locus inference of population structure: A comparison between single nucleotide polymorphisms and microsatellites. *Heredity (Edinb).* **106,** 158–171 (2011).

Appendix

Table 4: *List of genes and the regions of interest analysed using targeted DNA-Seq.*[87] *CDS represent the coding sequence of the gene.*

Gene	Region	Gene	Region	Gene	Region
ABCB1	entire CDS	*GATA1*	exons 2,3	*PTEN*	entire CDS
ABCG2	entire CDS	*GATA2*	exons 4-6	*PTPN11*	exons 3,13
ADA	entire CDS	*GATA3*	exons 3-5	*PTPRT*	entire CDS
ASXL1	exon 12	*HNRNPK*	entire CDS	*RAD21*	entire CDS
BCOR	entire CDS	*HRAS*	exons 2,3	*RUNX1*	entire CDS
BCORL1	entire CDS	*IDH1*	exon 4	*SETBP1*	exon 4
BRAF	exons 11,12,15	*IDH2*	exon 4	*SF1*	entire CDS
CBL	exons 8,9	*IL7R*	exon 6	*SF3A1*	entire CDS
CDA	entire CDS	*JAK1*	exons 13-15	*SF3B1*	exons 14-16
CDKN2A	entire CDS	*JAK2*	exons 12-16	*SMC1A*	entire CDS
CEBPA	entire CDS	*JAK3*	entire CDS	*SMC3*	entire CDS
CSF3R	exons 12-16	*KDM6A*	entire CDS	*SRSF2*	exon 1
CSFR1	exons 7, 22	*KIT*	exons 8,9,11,17	*STAG2*	entire CDS
DAXX	entire CDS	*KMT2A*	exons 1,3,4,33	*TERC*	entire CDS
DCK	entire CDS	*KRAS*	exons 2,3	*TERT*	exons 1,15
DCLK1	entire CDS	*MIR-142*	entire CDS	*TET2*	entire CDS
DIS3	entire CDS	*MPL*	exon 10	*TP53*	entire CDS
DNMT3A	exons 7-23	*MYD88*	exons 3-5	*U2AF1*	exons 2,6
ETV6	entire CDS	*NOTCH1*	exons 26-28,34	*U2AF2*	entire CDS
EZH2	entire CDS	*NPM1*	exons 10,11	*WAC*	entire CDS
FAM5C	entire CDS	*NRAS*	exons 2,3	*WT1*	entire CDS
FBXW7	exons 8-12	*NT5C2*	entire CDS	*ZRSR2*	entire CDS
FLT3	exons 13-16, 20	*PHF6*	entire CDS		

```
## BWA-MEM parameters on DNA-Seq
$ bwa mem \
  -t "7" -v 1 -k "19" -w "150" -d "100" -r "1.4" -y "20" -c "500" \
  -D "0.5" -W "0" -m "50" -A "1" -B "5" -O "6,6" -E "1,1" -L "3" \
  -U "17" -T "30" -h "5" -M “$reference" "$read1" "$read2" \
  > "$output"
```

Box 3: *Optimized BWA-MEM parameters for targeted DNA sequencing alignment.*

```
## STAR first pass parameters on RNA-Seq
$ STAR \
  --runThreadN "7" --genomeLoad NoSharedMemory --genomeDir
"$reference"
  --readFilesIn "$read1" "$read2"

## Pooling splice junction from all samples
$ cat splice_junctions_* | grep -v ^chrM | sort -V | uniq | \
  awk -F"\t" '$6=="0"{ print $1"\t"$2"\t"$3"\t"$4; }' | sort | \
  uniq > pooled_splice_junctions

## STAR genomic index using pooled splice junctions are created

## STAR second pass parameters on RNA-Seq with new indices
$ STAR --runThreadN "7" --genomeLoad NoSharedMemory
  --genomeDir "$reference_new" --readFilesIn "$read1" "$read2" \
  --quantMode "TranscriptomeSAM" --outReadsUnmapped Fastx \
  --chimSegmentMin 12 --chimJunctionOverhangMin 12 \
  --alignSJDBoverhangMin 10 --alignMatesGapMax 200000 \
  --alignIntronMax 200000
```

Box 4: *Optimized STAR parameters for total RNA sequencing alignment*

```
## DNA – VarScan parameters
$ samtools mpileup \
  -f "$reference" -l "$regionbed" -A -C "0" -d "1000000" -q "10" \
  -Q "20" "$input" | java –jar Varscan.jar mpileup2snp \
  --min-coverage 30 --min-reads2 6 --min-avg-qual 20 \
  --min-var-freq 0.02 --min-freq-for-hom 0.75 --p-value 0.01 \
  --strand-filter 1 --output-vcf 1 --variants 1 > $output

## DNA – VarDict parameters
$ vardict -G "$reference" -b "$input" -k 1 -c 1 -S 2 -E 3 -g 4 \
  "$regionbed" | teststrandbias.R | var2vcf_valid.pl -E -f 0.01 \
  > $output
```

Box 5: *Optimized variant calling parameters for DNA-Seq.*

```
## RNA – VarScan
$ samtools mpileup -f "$reference" -A -B -C "0" -d "250" -q "0" \
  -Q "13" "$input" | java –jar Varscan.jar mpileup2snp \
  --min-coverage 4 --min-reads2 1 --min-avg-qual 15 \
  --min-var-freq 0.01 --min-freq-for-hom 0.75 --p-value 0.01 \
  --strand-filter 0 --output-vcf 1 --variants 0 > $output

## RNA – VarDict
$ vardict -G "$reference" -b "$input" -k 1 -c 1 -S 2 -E 3 -g 4 \
  "$regionbed" | var2vcf_valid.pl -E -f 0.01 > $output
```

Box 6: *Less stringent variant calling parameters used for RNA-Seq*

```
> sessionInfo()
R version 3.4.3 (2017-11-30)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

Matrix products: default

locale:
[1] LC_COLLATE=German_Germany.1252
[2] LC_CTYPE=German_Germany.1252
[3] LC_MONETARY=German_Germany.1252
[4] LC_NUMERIC=C
[5] LC_TIME=German_Germany.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets
[6] methods   base

other attached packages:
 [1] edgeR_3.20.9       limma_3.34.9       viridis_0.5.0
 [4] viridisLite_0.3.0  gdata_2.18.0       RColorBrewer_1.1-2
 [7] wesanderson_0.3.2  reshape2_1.4.3     cowplot_0.9.2
[10] gridExtra_2.3      ggplot2_2.2.1

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.15     magrittr_1.5     munsell_0.4.3
 [4] lattice_0.20-35  colorspace_1.3-2 rlang_0.2.0
 [7] stringr_1.3.0    plyr_1.8.4       tools_3.4.3
[10] grid_3.4.3       gtable_0.2.0     gtools_3.5.0
[13] lazyeval_0.2.1   digest_0.6.15    tibble_1.4.2
[16] labeling_0.3     stringi_1.1.6    compiler_3.4.3
[19] pillar_1.2.1     scales_0.5.0     locfit_1.5-9.1
```

Box 7: *R session information including all additional packages used for the analysis and plotting.*

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to **Prof. Dr. Ulrich Mansmann** for giving me the opportunity to proceed with my Ph.D. Study. Thank you so much for all the guidance, support and the freedom to explore my research interests and grow as a person. Whenever, I or any other fellow students get frustrating results with insignificant p-values ($p > 0.05$), we always console ourselves with your famous quote "*Such is Life*".

I am very much obliged to **PD Dr. med. Tobias Herold** for steering me in the right direction in finishing my Ph.D. study. I am very grateful for all your inputs and insights regarding this work and providing me the opportunities to work in other collaborative projects. Your constant motivation and moral support helped me through tough times.

Another big thanks to my fellow researcher, soon to be a Dr. **Stefanos Alexandros Bamopoulos,** for all the discussions we had regarding our project works and sharing the frustrations that comes with it. I am thankful to my colleague, Herr. **Sebastian Schaaf** for his teachings and advices regarding the computational work and helping me out with all the technical issues. Furthermore, I would like to thank all my friends from India and in Germany, especially Ashok Varadharajan, Gopalakrishnan Dhandapani, Lien Le, Guokun Zhang for their persistent motivational and emotional support.

I owe a big thanks to my parents, Mr. Nazeer Batcha and Mrs. Nazeema begum, for being the best parents ever. Thank you Sheerin for being an awesome sister and Thanks a lot Benazir for being an understanding wife.