Institute for Molecular Medicine Finland (FIMM),
Helsinki Institute of Life Science (HiLIFE),
Doctoral School in Health Sciences (DSHealth), Doctoral Program in Biomedicine (DPBM)
University of Helsinki, Helsinki, Finland.

Transcriptomic data integration for precision medicine in leukemia

Ashwini Kumar

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Biological and Environmental Sciences, University of Helsinki, for public examination in Lecture Hall 2, Biomedicum 1, Helsinki on Friday, October 11th 2019 at 12 noon.

Helsinki 2019

Supervised by

Caroline Heckman, PhD
Group leader,
Institute for Molecular Medicine Finland (FIMM),
Helsinki Institute of Life Science (HiLIFE), University of Helsinki,
Helsinki, Finland

Pirkko Mattila, PhD
Docent,
Institute for Molecular Medicine Finland (FIMM),
Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland

Thesis Advisory Committee

Sampsa Hautaniemi, PhD
Professor,
Research Program in Systems Oncology, Faculty of Medicine
University of Helsinki,
Helsinki, Finland

Antti Honkela, PhD
Associate Professor,
Department of Computer Science,
University of Helsinki,
Helsinki, Finland

Reviewed by

Throsten Zenz, MD, PhD
Professor,
Department of Medical Oncology and Hematology,
University Hospital Zurich,
Zurich Switzerland

Francesco Iorio, PhD
Group Leader,
Wellcome Sanger Institute,
Wellcome Genome Campus,
Hinxton, Cambridge, UK

Official opponent

Inge Jonassen, PhD
Professor,
Computational Biology Unit,
University of Bergen, Norway

ISBN 978-951-51-5516-0 (paperback)
ISBN 978-951-51-5517-7 (PDF)
https://ethesis.helsinki.fi
Unigrafia Oy
Helsinki 2019

Table of Contents

1 LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications, which are referred to in the text by their Roman numerals:

I. **Kumar A**, Kankainen M, Parsons A, Kallioniemi O, Mattila P, Heckman CA. The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia. *BMC Genomics*. 2017;18(1):629. Epub 2017/08/19. doi: 10.1186/s12864-017-4039-1. PubMed PMID: 28818039.

II. Kontro M, **Kumar A**, Majumder MM, Eldfors S, Parsons A, Pemovska T, Saarela J, Yadav B, Malani D, Floisand Y, Hoglund M, Remes K, Gjertsen BT, Kallioniemi O, Wennerberg K, Heckman CA*, Porkka K*. HOX gene expression predicts response to BCL-2 inhibition in acute myeloid leukemia. *Leukemia*. 2017;31(2):301-9. Epub 2016/08/09. doi: 10.1038/leu.2016.222. PubMed PMID: 27499136.

III. Karjalainen R*, Liu M*, **Kumar A**, Parsons A, Kontro M, Porkka K, Heckman CA. Elevated expression of the S100A8/S100A9 complex in AML correlates with reduced sensitivity to the BCL-2 inhibitor venetoclax. *Leukemia*. 2019 Jun 7. doi: 10.1038/s41375-019-0504-y. PMID: 31175323

* Equal contribution

The original articles (I-III) were reproduced with permission from the copyright holders.

ABBREVIATIONS

ABL	Abelson murine leukemia viral oncogene homolog
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
AUC	Area under the curve
BAD	BCL2 associated agonist of cell death
BAX	BCL2 associated X
BCL-2	B-cell lymphoma-2
BCL-XL	B-cell lymphoma-extra large
BCR	Breakpoint cluster region
BH-3	Bcl-2 homology domain 3
BM	Bone marrow
CCLE	Cancer Cell Line Encyclopedia
CPM	Counts per million
CR	Complete remission
DNA	Deoxyribonucleic acid
DNMT3A	DNA methyltransferase 3A
DMSO	Dimethylsulfoxide
DSRT	Drug sensitivity and resistance testing
DSS	Drug sensitivity score
ELN	European Leukemia Net
FDA	Food and Drug Administration
FHRB	The Finnish Hematology Registry and Biobank
FLT3	Fms-like tyrosine kinase 3
HOX	Homeobox
HSCT	Hematopoietic stem cell transplantation
IC50	Half-maximal inhibitory concentration
IDH	Isocitrate dehydrogenase
ITD	Internal tandem duplication
NGS	Next-generation sequencing
NRAS	Neuroblastoma RAS viral oncogene homolog
NPM1	Nucleophosmin gene 1
MAPK	Mitogen-activated protein kinase
MCL-1	BCL2 family apoptosis regulator (myeloid cell leukemia sequence 1)
NCI	National Cancer Institute
PA	Poly-A enrichment
PCR	Polymerase chain reaction
RD	Ribo-depletion
RNA	Ribonucleic acid
RT-qPCR	Reverse transcriptase quantitative polymerase chain reaction
TCGA	The Cancer Genome Atlas
TMM	Trimmed mean of M-values
WT	Wild type

2 ABSTRACT

This thesis is comprised of three studies demonstrating the application of different statistical and bioinformatic approaches to address distinct challenges of implementing precision medicine strategies for hematological malignancies. The approaches focus on the analysis of next-generation sequencing data, including both genomic and transcriptomics, to deconvolute disease biology and underlying mechanisms of drug sensitivities and resistance. The outcomes of the studies have clinical implications for advancing current diagnosis and treatment paradigms in patients with hematological diseases.

Study I, RNA sequencing has not been widely adopted in a clinical diagnostic setting due to continuous development and lack of standardization. Here, the aim was to evaluate the efficiency of two different RNA-seq library preparation protocols applied to cells collected from acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) patients. The poly-A-tailed mRNA selection (PA) and ribo-depletion (RD) based RNA-seq library preparation protocols were compared and evaluated for detection of gene fusions, variant calling and gene expression profiling. Overall, both protocols produced broadly consistent results and similar outcomes. However, the PA protocol was more efficient in quantifying expression of leukemia marker genes and drug targets. It also provided higher sensitivity and specificity for expression-based classification of leukemia. In contrast, the RD protocol was more suitable for gene fusion detection and captured a greater number of transcripts. Importantly, high technical variations were observed in samples from two leukemia patient cases suggesting further development of strategies for transcriptomic quantification and data analysis.

Study II, the BCL-2 inhibitor venetoclax is an approved and effective agent in combination with hypomethylating agents or low dose cytarabine for AML patients, unfit for intensive induction chemotherapy. However, a limited number of patients responding to venetoclax and development of resistance to the treatment presents a challenge for using the drug to benefit the majority of the AML patients. The aim was to investigate genomic and transcriptomic biomarkers for venetoclax sensitivity and enable identification of the patients who are most responsive to venetoclax

treatment. We found that venetoclax sensitive samples are enriched with *WT1* and *IDH1*/*IDH2* mutations. Intriguingly, *HOX* family genes, including *HOXB9*, *HOXA5*, *HOXB3*, *HOXB4,* were found to be significantly overexpressed in venetoclax sensitive patients. Thus, these *HOX*-cluster genes expression biomarkers can be explored in a clinical trial setting to stratify AML patients responding to venetoclax based therapies.

Study III, venetoclax treatment does not benefit all AML patients that demands identifying biomarkers to exclude the patients from venetoclax based therapies. The aim was to investigate transcriptomic biomarkers for *ex vivo* venetoclax resistance in AML patients. The correlation of *ex vivo* venetoclax response with gene expression profiles using a machine learning approach revealed significant overexpression of S100 family genes, *S100A8* and *S100A9*. Moreover, high expression of *S100A9* was found to be associated with birabresib (BET inhibitor) sensitivity. The overexpression of *S100A8* and *S100A9* could potentially be used to detect and monitor venetoclax resistance. The combination of BCL-2 and BET inhibitors may sensitize AML cells to venetoclax upon BET inhibition and block leukemic cell survival.

Taken together, we demonstrated the utility of transcriptomics and bioinformatics data analysis strategies for precision medicine in leukemia. The evaluation of RNA-seq library preparation protocols and identification of gene expression biomarkers for drug responses were investigated in patients with hematological malignancies.

3 INTRODUCTION

The molecular understanding of cancer has introduced novel approaches for routine clinical practice in diagnosis, prognosis, and treatment decisions making. The Human Genome Project resulted in the complete mapping of the human genome in 2003. Since then, the technological innovations, increasing speed and reducing the cost of next-generation sequencing (NGS) has facilitated an in-depth investigation of the molecular basis of cancer. It has also catalyzed the invention of newer technologies and computational tools that have transformed the cancer genomic research. Multiple types of cancer patients were sequenced in The Cancer Genome Atlas Program (TCGA) and the International Cancer Genome Consortium (ICGC) projects. The emerging genomic and transcriptomic information has facilitated biomarker discovery for disease monitoring, risk prediction and developing treatment modalities. The recent developments in cancer genomic and transcriptomic fields have built a platform for precision medicine.

The core components of current precision medicine include multi-omics studies, large-scale cohort trials, and big data integration. Standardization of sequencing protocols and the quality of data are key challenges to overcome in order to incorporate NGS-based tools for precision medicine. With a lack of standardized data processing, the outcomes of sequencing studies have been of low reliability. Therefore, establishing robust and standard protocols for clinical use are very crucial to implement precision medicine. Especially for RNA-sequencing (RNA-seq), subsequent data analysis pipelines must deliver accurate information with reproducible and robust performance. Moreover, establishment and standardization of methods for assessing reproducibility, accuracy and precision in a variety of clinically relevant conditions are needed to facilitate the adoption of RNA-seq data in the clinical laboratory.

Current precision medicine synonymizes genomics medicine to match the right drug to the right patient at the right time and dose. Although several breakthrough therapies have been discovered for cancer patients with specific genetic lesions, most cancer patients lack targeted therapies. Hence, to make precision medicine successful and applicable for cancer patients' treatment, it is essential to incorporate additional tools. Functional precision medicine offers an alternative solution by measuring signals from cancer cells upon therapeutic perturbation. Moreover, combining genomic profiling with functional testing provides promising precision medicine approaches not only to obtain a panoramic view of cancer cells but also to discover effective therapies to individual patients. This powerful approach helps to identify molecular denominators of drug sensitivity and resistance to stratify patients who are most likely to respond to the therapies. Utilizing modern statistical, bioinformatics and machine learning methods, play a crucial role in identifying robust genomic and transcriptomic biomarkers for drug responses. Therefore, it is crucial to combine multiple technologies and analytical tools to extract clinically relevant information from the complex biology of cancer cells.

4 REVIEW OF THE LITERATURE

4.1 Precision medicine in cancer

According to the president's council of advisors on science and technology the USA, precision medicine is defined as: "the tailoring of medical treatment to the individual characteristics of each patient to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment. Preventative or therapeutic interventions can then be focused on those who will benefit, sparing expense and side effects for those who will not"[1]. Precision medicine is an approach to stratify patients in order to improve diagnosis and treatment by integrating clinical and molecular information to understand the biological basis of human disease[2,3] as shown in Figure 1. It also considers the environmental exposures and additional traits of an individual and their lifestyle to create a tailor-made treatment[4,5]. As the definition suggests, the power of precision medicine lies in its ability to; i) optimize and improve health care by applying an innovative approach to disease prevention and treatment that takes into account individual differences in genetic make-up, environments, and lifestyles ii) help understand/discover mechanisms underlying the disease iii) provide relevant tools to better understand the complex mechanisms or disease condition to clinicians iv) predict which treatments will be most effective[6,7].

4.1.1 Overview of precision medicine

The term personalized medicine was re-coined as precision medicine after 2013 to recognize the shifting goals of modern medicine concerning the continuous development of technologies[8,9]. Personalized medicine refers to an approach for patients that considers their genetic make-up but with attention to their preferences, beliefs, attitudes, knowledge and social context, whereas precision medicine describes a systems model for health care delivery that relies heavily on data, analytics and information[8]. In 2011, the USA National Research Council (NRC) expressed concern with the term "personalized medicine" as it may be misunderstood to mean that completely individualized treatments are available for each unique patient.

The report defines precision medicine as "the tailoring of medical treatment to the individual characteristics of each patient." The report added, "it does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment."[1]

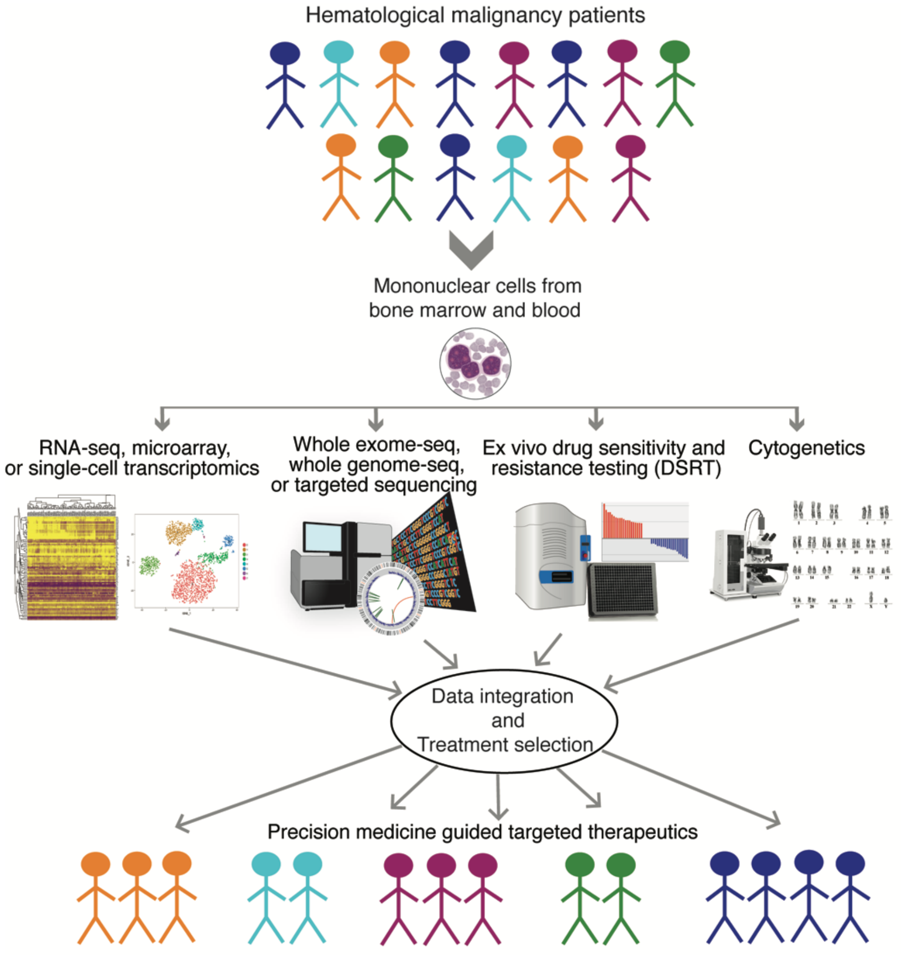

Hematological malignancy patients
Mononuclear cells from bone marrow and blood
RNA-seq, microarray, or single-cell transcriptomics
Whole exome-seq, whole genome-seq, or targeted sequencing
Ex vivo drug sensitivity and resistance testing (DSRT)
Cytogenetics
Data integration and Treatment selection
Precision medicine guided targeted therapeutics

Figure 1. An example of precision medicine workflow in patients with hematological malignancies: Technologies implemented for patient stratification and therapy recommendation.

4.1.2 Genomic precision medicine

The ultimate goal of genomic precision medicine for cancer is to identify somatic genetic alterations (point mutations, amplifications, translocations) and match them with effective the treatments. The tyrosine kinase inhibitor imatinib was the first example of genomic precision medicine to treat chronic myeloid leukemia (CML) patients the carrying *BCR-ABL1* fusion gene[10]. Imatinib helped to improve the overall survival rates of CML patients to 90% over five years and 88% over eight years[11]. Another example is trastuzumab which is approved for treatment of breast cancer patients with amplification or overexpression of human epidermal growth factor receptor 2 (*HER2*). Compared to chemotherapy alone, the addition of trastuzumab to chemotherapy significantly slowed the disease progression (i.e., median, 4.6 vs. 7.4 months), prolonged survival time (i.e., median, 20.3 vs. 25.1 months), and reduced the risk of death by 20%[12]. Gefitinib[13,14] was approved for epidermal growth factor receptor (*EGFR*) mutant non-small-cell lung cancers and crizotinib was approved for patients with *EML4-ALK* fusion gene[15,16]. Furthermore, vemurafenib and dabrafenib were approved for advanced-stage *BRAF* V600E mutant melanoma[17].

Recent genomic precision medicine studies conducted on different cancer types with a heavy focus on NGS of tumor samples[18-26]. As an example, Lagana *et al*. demonstrated a unique approach where genomic and transcriptomic features were integrated of 64 multiple myeloma patients to generate treatment recommendations[27]. In this study, patients were assigned therapies based on both gene expression and somatic mutation findings. For acute myeloid leukemia (AML), the FMS like tyrosine kinase 3 (FLT3) inhibitor midostaurin was approved for the treatment of patients with a *FLT3* mutation[28,29]. Ivosidenib and enasidenib were approved for relapsed or refractory AML patients with isocitrate dehydrogenase 1 and 2 (*IDH1/2*) mutations, respectively[30,31]. Although genomic precision medicine approaches have been accepted in the clinic for treatment decision making, the response rates have remained modest[25,32].

Recently, large-scale genomic precision medicine approaches were systematically applied for individual solid tumor patients in clinical trial settings. In the systematics analysis, molecular profiling was performed for recurrent metastatic cancer patients to guide clinical treatment decision making. Within the TARGET study, therapy with experimental targeted treatments was guided by sequencing results from circulating tumor DNA and genomic DNA[33]. In the I-PREDICT study, drug combinations were designed based on DNA sequencing[34]. Interestingly, RNA-seq was coupled with DNA sequencing from tumor and adjacent tissues to select targeted therapies in combination with immunotherapies in the WINTHER trial[35].

4.1.3 Functional precision medicine

Precision medicine field has been considered synonymous to genomics medicine. Most cancer patients lack benefits from genomic data-driven precision treatment strategies in terms of long-lasting remission or lengthened survival. Hence, it is essential to broaden the scope of precision medicine by exploring functional characteristics besides genomics and transcriptomics[32]. Investigating functional features could help to identify additional targetable vulnerabilities and effective therapies matched to patient-specific phenotypes. A recent study demonstrated a workflow of a drug testing assay for clinical referral in solid tumors and hematological malignancies[36]. The development of protocols to grow primary tumor cells for drug testing was optimized towards precision medicine implications to identify patient-specific effective drugs in lung cancer patients[37]. In a clinical trial with 769 patients, drug testing identified effective drug treatments and combinations for metastatic and primary tumors[38]. To accelerate precision medicine efforts for solid tumor patients, similar methods have been developed for culturing primary solid tumor cells by academic research groups[39,40] and pharma industry[41].

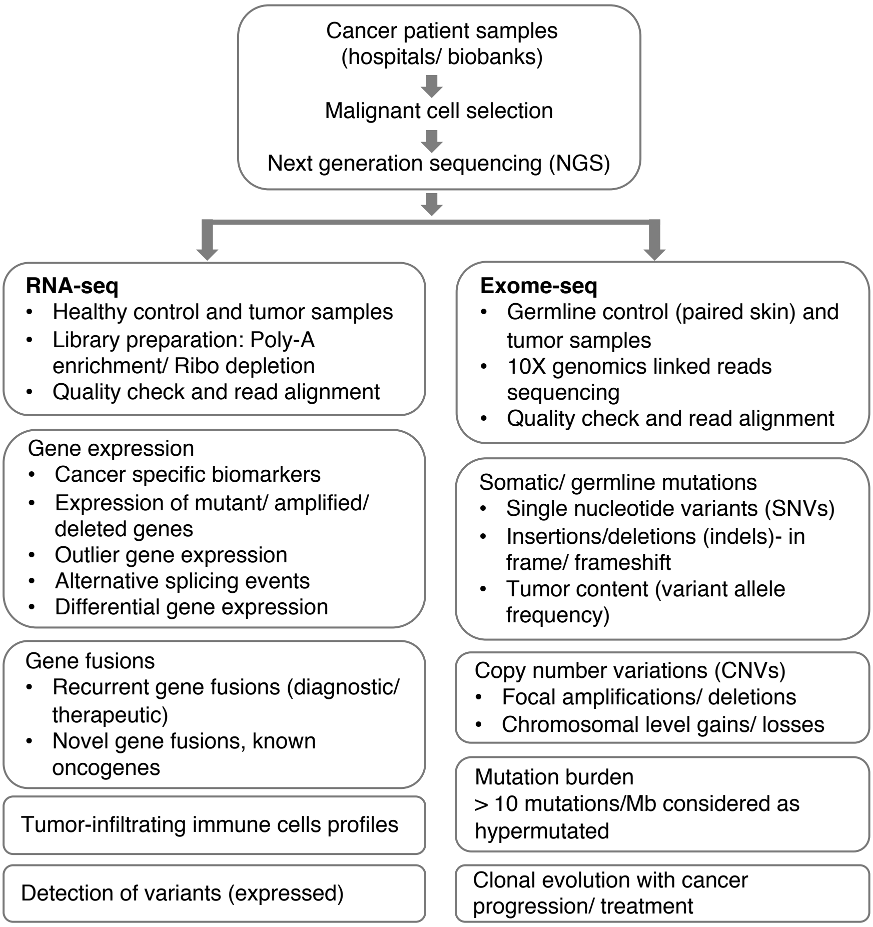
Snijder *et al.* utilized *ex vivo* imaging of drug responses to recommend drugs for clinical treatment in patients with hematological malignancies[42]. Survival benefit with selected therapeutics was reported using drug testing on primary cells over standard regimen for AML patients where non-targeted chemo drugs were used for clinical translation[43]. The analyses of large-scale data leading to a systematic exploration of targeted drug vulnerabilities associated with molecular subsets of AML patients can help establish precision medicine practice[44]. Also, there are emerging public datasets that can facilitate precision medicine efforts. Iorio *et al.* provided genomic and functional data from 1,001 molecularly annotated human cancer cell lines from 29 tissues. The study generated a large dataset including somatic mutations, copy number alterations, DNA methylation, gene expression and correlated with sensitivity to 265 drugs[45]. Genetic perturbation screens (CRIPSR/siRNA) have been used to identify novel cancer therapeutic targets as well as biomarkers using data integration efforts in pan-cancer cell lines[46,47]. Recently, the Beat AML program provided *ex vivo* drug responses data (122 inhibitors) from 562 AML patients with paired whole exome-seq and RNA-seq data on bulk cells [48]. Thus, functional precision medicine approaches may lead to better treatment outcomes but can be further improved by integrating molecular profiling with functional assays for clinical response prediction.

4.2 Tools for precision oncology

The increasing sequencing speed, analysis, accuracy, and affordability of NGS has helped spur the advent of precision oncology[49,50]. Tools facilitating precision medicine include transcriptomics, genomics and functional assays. The implications of genomic and transcriptomic sequencing of tumor specimens have been applied to improve the diagnosis and treatment of cancer patients, as shown in Figure 2. Advancements in high-throughput drug testing technology made functional profiling, as one of the emerging tools of precision medicine. The integrative analysis includes multi-dimensional data layers and application of machine learning algorithms, which has the potential to improve the clinical management of cancer patients.

4.2.1 RNA-sequencing

RNA sequencing (RNA-seq) detects expression changes by capturing quantitative gene expression patterns and describes the underlying phenotypes in great detail. Compared with microarray-based transcriptome profiling, RNA-seq covers a wider dynamic range and avoids certain technical limitations, for example, varying probe performance and cross-hybridization[51]. The primary outcomes of cancer transcriptomics can be broadly classified as genetic and functional readouts. The functional phenotypes that can be interrogated through transcriptome profiling are very broad and include quantitative estimates of expression levels and the detection of transcript isoforms, chimeric RNAs and RNA-editing sites. Similarly, the genotypes that can be interrogated by RNA sequencing include structural variants (e.g., gene fusions), copy number variants (CNVs) (e.g., amplifications) and somatic mutations (e.g., single nucleotide variants (SNVs)).



Cancer patient samples (hospitals/ biobanks)
Malignant cell selection
Next generation sequencing (NGS)
RNA-seq
• Healthy control and tumor samples
• Library preparation: Poly-A enrichment/ Ribo depletion
• Quality check and read alignment
Exome-seq
• Germline control (paired skin) and tumor samples
• 10X genomics linked reads sequencing
• Quality check and read alignment
Gene expression
• Cancer specific biomarkers
• Expression of mutant/ amplified/ deleted genes
• Outlier gene expression
• Alternative splicing events
• Differential gene expression
Somatic/ germline mutations
• Single nucleotide variants (SNVs)
• Insertions/deletions (indels)- in frame/ frameshift
• Tumor content (variant allele frequency)
Gene fusions
• Recurrent gene fusions (diagnostic/ therapeutic)
• Novel gene fusions, known oncogenes
Copy number variations (CNVs)
• Focal amplifications/ deletions
• Chromosomal level gains/ losses
Mutation burden
> 10 mutations/Mb considered as hypermutated
Tumor-infiltrating immune cells profiles
Detection of variants (expressed)
Clonal evolution with cancer progression/ treatment

Figure 2. Bioinformatic workflow. Integrative clinical next-generation sequencing and its applications for precision oncology.

Advances in experimental and computational tools have dynamically revolutionized transcriptome profiling over the past four decades[52,53]. Using RNA-seq, it has now become possible to sequence and quantifies the gene expression patterns at a single cell level[48,54]. These transcriptomes provide an opportunity to dissect the complexity and heterogeneity of tumors and to discover new biomarkers or therapeutic approaches for translational and precision medicine strategies[55,56]. For example, the RNA-seq technique has been particularly insightful in understanding the drug sensitivity and resistance patterns of malignant cells in AML and classifying the disease[48,57]. It has enabled identification of a wide variety of clinically relevant predictive expression biomarkers[58-60], fusion-genes including structural variants and amplifications[61-64], as well as alternative splicing events[65,66] in different cancer types. The high coverage of RNA-seq allows detecting SNPs and somatic mutation in the genes with average to high expression levels[67,68]. However, the highest sensitivity and specificity to detect genomic alterations can be achieved by combining both genomic and transcriptomic sequencing[69].

RNA-seq includes a sequence of related methodologies[70]. Typical RNA-seq experiment involves sample processing, library preparation, sequencing and downstream computational data analysis[71]. The first step is the disruption of cells and isolation of RNA molecules. The protocols have been adapted for a wide range of materials, including body fluids (e.g., blood, bone marrow biopsies), solid tissues and cell cultures. The second step is library preparation (mRNA selection). The third step is RNA fragmentation, cDNA synthesis and addition of sequencing adaptors and the final step is sequencing itself. The sequencing of the RNA-seq libraries are commonly performed using Illumina sequencers that utilize sequencing by synthesis chemistry[72].

Although ribosomal RNA (rRNA) is the most abundant (>80% of total RNA) RNA molecules in a cell[73], it has limited potential for clinical applications. Hence, depleting rRNA is an essential step to save sequencing bandwidth[70,71]. A number of rRNA removal (depletion) methods exist based on i) hybridization[74-76] followed by depleting the bound targets using immobilized streptavidin; ii) duplex digestion[77] involving heat-denaturing followed by re-annealing and removal of rRNA using duplex-specific nucleases; iii) pseudo-random or not-so-random priming[78], relying on a collection of short, computationally selected oligonucleotides, called 'not-so-random' primers.

Standard approaches for RNA-seq library preparation include either enrichment of polyadenylated (PA) RNA transcripts using oligo (dT) primers or rRNA depletion through hybridization followed by magnetic bead separation. However, the PA enrichment and RD method each have unique advantages and limitations, respectively. The PA enrichment method is currently the most popular protocol in cancer transcriptomics[79]. However, this approach requires intact RNA to avoid technical biases and artifacts. Protocols that utilize ribodepletion[80] or hybridization are therefore more suitable for clinical use, where RNA material is limited or obtained from frozen or variable quality tissue. Over the past decade, many comparative studies between PA and RD methods have been performed[81-87] but mostly using non-clinical samples. This challenge emphasizes the need for systematic comparison of library preparation protocols for cancer patient transcriptomic studies in a precision medicine setting. Our comparative analysis provides recommendations for the application of RNA-seq in clinical or pre-clinical settings with a limited number of samples.

4.2.2 Whole exome-sequencing

Whole exome sequencing (WES) also referred to as exome sequencing, analyzes the coding region of the genome and offers a comprehensive genomic profile of aberrations in protein-coding genes. The Encyclopedia of DNA Elements (ENCODE) project in 2012 reported that human exons of protein-coding genes cover 2.94% of the genome[88]. The latest version of the human reference genome “GRCh38” has a complete set of protein-coding regions and constitutes ~3.09% (over 90 million nucleotides)[89]. On a larger scale, whole-genome sequencing (WGS) provides the most comprehensive view of the entire human genome[90] that is ~3 billion bases for a single human sample sequenced. WGS also provides a better resolution of structural variations and CNVs compared to WES.

The overall goal of WES is to measure inter-personal variability in genomic DNA by comparing an individual’s DNA sequence to the reference human genome. The approach is useful for both clinical as well as research applications since it covers actionable areas of the genome. It determines the variations in the exonic regions to help identify various cancer-associated mutations[91]. With the improvement in sequencing technologies and standardization of data analysis pipelines, the WES has been employed for real-time clinical applications[92,93].

4.2.3 Functional assays

In addition to sequencing technologies, functional assays hold promise to advance current precision medicine approaches. One approach gaining popularity is high-throughput drug testing, which accesses the impact of drugs on cell viability, cell differentiation, or other cellular phenotypes. High-throughput drug testing allows testing of thousands of drugs at multiple doses and has been broadly implemented to identify cancer effective drugs based on response to the tested drugs[94]. Early systematic high-throughput drug testing has used established human cancer cells lines to identify potential targeted drugs for further clinical development. A National Cancer Institute (NCI) study, screened FDA approved drugs with 60 human cancer cell lines (NCI60) generated a widely used dataset resource[95]. Later in 2012, Barretina *et al*. at the Broad Institute[96] and Garnet *et al.* at the Sanger institute[97] published studies on high-throughput drug testing of 1000 human cell lines covering major cancer types. The *ex vivo* drug testing approach was extended to primary patient material, including relapsed and refractory AML patients to facilitate therapy selection for individual patient cases[48,98]. A similar approach was adapted for chronic lymphocytic leukemia (CLL) patients to identify potential targeted drugs and associated patterns of molecular features by Dietrich *et al.*[99]. Tzelepis *et al*. applied a genome-wide CRISPR screening platform and reported that inhibition of *KAT2A* gene leads to differentiation and apoptosis of human AML cells[100]. Also, drug testing analysis using flow cytometry to distinguish the drug response based on different cell populations has been demostrated[101]. A novel BH3 mimetic assay was reported to predict chemotherapy resistance in leukemia patients and was promoted to advance current functional precision medicine efforts[102].

4.3 Bioinformatic approaches for precision medicine

In recent years, high-throughput technologies have been used to generate a vast amount of multi-omics and functional data. Previously, the main focus was on the analysis of a single layer of data type including gene expression, somatic mutations, CNVs and DNA methylation, independently. Since the molecular complexity and heterogeneity of cancer exists at all levels, integrative bioinformatic analyses of multiple layers of data simultaneously offer an effective and robust strategy to achieve a better understanding of pathogenic mechanisms. The data-driven analyses of multi-omics data in addition to data from functional assays can result in more profound insights into cellular functionality.

4.3.1 Transcriptomics data analysis

The primary and fundamental goal of transcriptomics data analysis is to identify genes with significantly altered expression level changes between given conditions. For example, frequent comparisons include drug-sensitive versus drug-resistant or mutation-negative (wild type) versus mutation-positive. More complicated experimental designs include extra experimental factors to account for covariates (such as experimental batch, age, gender, library preparation methods, disease etc). The standard RNA-seq work-flow includes i) experimental design based on the biological question and determination of the appropriate sample size for sequencing; ii) obtaining the sequence data, which include sample collection and processing, RNA-extraction, library preparation and sequencing; iii) preprocessing the data, which includes performing quality control, adapter trimming, and alignment; iv) analyzing data, which involves normalizing the read counts, identifying differentially expressed genes, identifying fusion genes, visualizing the results, correlate with phenotype if data available, and validate outcomes. The commonly used tools used for pre-processing the RNA-seq data are explained in Table 1.

Tools	Descriptions	Ref
Quality control		
FastQC	Rapid assessment of sequence data	Andrews
RNA-SeQC	Read mapping summary statistic, coverage	DeLuca (2012)[103]
RSeQC	Read distribution over genome, read depth	Wang (2012)[104]
PRINSEQ	Summary statistics, trim adaptor sequences	
Trimmomatic	Performs trimming for Illumina platforms	Bolger (2014)[105]
Cutadapt	Removes adapter sequences	Martin (2011)[106]
Alignment		
TopHat v2	Candidate exon pairing, implanted bowtie	Kim (2013)[107]
Subread	Seed-and-vot	Lioa (2013)[108]
STAR	Maximal mappable prefix	Dobline (2013)[109]
HISAT2	Spliced alignment program	Kim (2015)[110]
Read counting		
featureCounts	Gene-level quantification	Lioa (2014)[111]
htseq-count	Gene-level quantification	Anders (2015)[112]
Rcount	Reads aligning with multiple locations	Schmid (2015)[113]
Differential gene expression		
edgeR	Negative binomial distribution	Robinson (2010)[114]
DEseq2	Negative binomial distribution	Love (2014)[115]
Fusion genes detection		
EricScript	Recalibrates junction reference	Benelli (2012)[116]
SOAPfuse	Can detect low fusion over coverage	Jia (2013)[117]
FusionCatcher	Detects both known and novel fusions	Nicorici (2014)[118]

Table 1: Tools for RNA-seq data analysis.

4.3.2 Machine learning approaches for data integration

Application of machine learning tools in genomics has massively increased in recent years and proved to be very valuable in providing novel insights[119,120]. For example, machine learning can be used to identify the location of transcription start sites, promoters, splice sites, or enhancer sites in the genome[119]. In the past, single-layer analysis has been extensively conducted at different levels, including mRNA, microRNA, CNV, DNA methylation and somatic mutations were analyzed independently[121,122]. As the molecular complexity of disease etiology exists many different levels, integrative analysis approaches offer an efficient way to join forces across multi-level omics data. Since the diverse layers of patient-derived "big data" are being generated, new bioinformatic approaches need to be developed to integrate multi-dimensional data[123]. High-throughput *ex vivo* drug sensitivity testing read-outs have the potential to become one of the significant components of precision oncology. The machine learning approaches have been proven to be enormously useful in predicting drug responses by integrating multi-omics data[124]. The resultant molecular denominator for *ex vivo* drug responses could assist clinicians to make decisions on patient treatment, including the selection of the most effective therapies. Machine learning is a data-driven field that involves applying algorithms and building models with the ability to 'learn' to make accurate predictions with experience. Machine learning methods can primarily be categorized into supervised learning or unsupervised learning. Supervised learning requires known examples or established patterns to train the models, which is then used to predict the respective labels. In contrast, unsupervised learning is concerned with finding patterns or clusters without any prior knowledge[125,126].

Several drug sensitivity prediction algorithms have been proposed to characterize the relationship between gene expression profiles and drug responses[96,127-132]. Liu *et al.* applied linear regression models to identify gene expressions, co-expressions, and co-expression modules associated with drug sensitivity in CCLE (Cancer Cell Line Encyclopedia) data by considering relevant confounding factors such as age, sex, batch, cancer and tissue types[133]. Masica *et al.* developed a novel approach named multivariate organization of combinatorial alterations (MOCA), combining many genomic alterations into biomarkers of drug response. Outcomes of the MOCA approach suggested that multi-gene features correlation with drug response substantially better compared to individual genes[134].

Over the last decade, many machine learning models have been used for the big data integration and drug response prediction, including linear regression, elastic net regression, support vector machines, neural networks and random forest as reviewed by Azuaje[135]. Emad *et al.* proposed a gene prioritization method called Prioritization of Genes Enhanced with Network Information (ProGENI) to rank genes that are closely related to a phenotype[136]. With the ranked genes, the authors employed a kernel support vector machine (SVM) for drug sensitivity prediction and reported that ProGENI–identified genes can better predict drug response compared to genes identified by other widely used prioritization methods such as Pearson correlation and elastic net regression[136]. A collaborative effort between the NCI and the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project, performed a comparison of 44 different drug response prediction methods and found that Bayesian multitask multiple kernel learning exhibited the best predictive performance[59]. Also, gene expression was found to have more predictive power in drug response prediction compared to other features such as mutations or CNVs[59,136].

Lee *et al.* developed the MERGE (mutation, expression hubs, known regulators, genomic CNV, and methylation)[137] algorithm, which integrates multi-omic data to identify statistically correlated gene markers of drug sensitivity in AML. MERGE learns the weight of each unique driver features to successfully predict known drug sensitivity using a wide variety of input data including mutations, CNVs, and DNA methylation, gene expression and regulatory annotations. The MERGE model was able to identify an association between high FLT3 expression and sensitivity to FLT3 inhibitors midostaurin, ponatinib, sunitinib, and tandutinib[137]. Aben *et al.* implemented elastic net regression models and developed TANDEM method. The method is a two-stage approach where the first stage explains drug response using upstream features (mutations, CNVs, methylation and cancer type) and the second stage explains the remainder using downstream features (gene expression, pathways). Jang *et al.* applied elastic net regression and found it to be one of the best-performing modeling strategies for drug response prediction in the CCLE and GDSC (Genomics of Drug Sensitivity in Cancer) cancer cell line datasets[138]. Likewise, Ding *et al.* applied elastic net regression combining genomic data for drug sensitivity prediction through deep learning in the CCLE and GDSC datasets[139]. The broad applications of machine learning have tremendously advanced the goals of precision cancer medicine. Importantly, standardization of the methods and uniform data analysis strategies across various biological modalities can yield more advantages in the future. Moreover, the increasing scale of data obtained from cancer patients will sharpen machine learning tools in terms of improved efficiency and robust outcomes.

4.4 BCL-2 inhibition in AML

4.4.1 Acute myeloid leukemia

Acute myeloid leukemia (AML) is a heterogeneous malignancy of the hematopoietic system, characterized by massive proliferation and accumulation of undifferentiated leukemic blasts[140,141]. Even though the occurrence of AML is relatively rare (1.2% of all cancers) compared to other systemic cancer types, the five-year survival rate of AML patients is poor as recorded in cancer registry data from the Surveillance, Epidemiology and End Results Program (SEER) database. According to the SEER database, AML occurrence and mortality rates are higher in elderly patients compared to children as the median age at diagnosis is 67 years. Although almost 70% of adults with AML achieve a complete remission with conventional chemotherapy, the long-term survival rate has remained only 30%[142]. Another challenge for successful AML treatment is the complexity and molecular heterogeneity of the disease[143] (Figure 3).

Extensive efforts to characterize AML genome has deconvoluted the recurrence and interaction patterns of mutations[144,145]. Investigating epigenetic events has provided insights in DNA methylation patterns in AML[146] . Sequencing of 200 *de novo* adult AML patients revealed major genomic and epigenomic drivers of the disease[147]. The aberrantly regulated gene expression signatures have been reported in molecular subtypes of AML. For example, unique gene expression signatures and regulatory networks were identified in mutation subtypes of AML[148,149]. Papaemmanuil *et al.* demonstrated the utility of combining cytogenetics and molecular taxonomy as prognosis schema in AML patient cases[121]. Furthermore, the cytogenetics and mutation information was used to design the European Leukemia Net (ELN) 2017 classification system by a panel of international experts[150]. Additionally, clonal heterogeneity studies have revealed the complexity of the disease progression and emphasize the need for strategies to target this progressive disease efficiently [151-153]. Furthermore, potential targeted drugs were identified for specific molecular features from genomics, transcriptomics[154,155], and methylation profiling[99] in leukemia patients.

Conventional treatment options for AML patients include chemotherapy and subsequent allogeneic hematopoietic stem cell transplantation (allo-HSCT). The primary goal of chemotherapy is to eradicate the leukemic blasts cells or to induce differentiation of immature leukocytes in some instances. The widely used induction (first line of therapy) treatment regimen, which massively kills the majority of leukemic blasts, consists of cytarabine (nucleoside analog) in combination with daunorubicin, idarubicin or mitoxantrone (anthracyclines, also known as topoisomerase inhibitors). The induction regimen is followed by a consolidation regimen to eradicate the remaining leukemic blasts, where the selection of drugs may vary from patient to patient. Allo-HSCT is performed to replenish healthy hematopoietic progenitors in patients achieving complete remission. In the case of acute promyelocytic leukemia (APL) patients, carrying *PML-RARA* gene fusion, tretinoin treatment is cornerstone therapy. The tretinoin treatment leads to differentiation of leukemic blasts into mature and functional leukocyte cells.

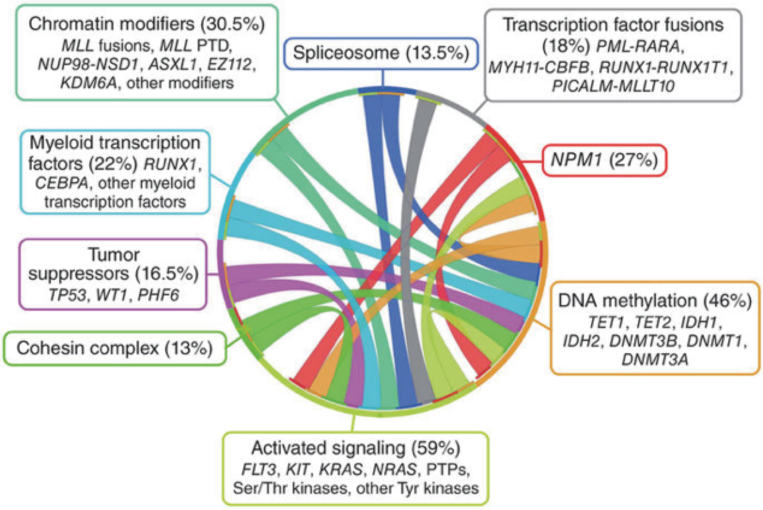

Chromatin modifiers (30.5%)
MLL fusions, *MLL* PTD, *NUP98-NSD1*, *ASXL1*, *EZ112*, *KDM6A*, other modifiers
Spliceosome (13.5%)
Transcription factor fusions (18%) *PML-RARA*, *MYH11-CBFB*, *RUNX1-RUNX1T1*, *PICALM-MLLT10*
Myeloid transcription factors (22%) *RUNX1*, *CEBPA*, other myeloid transcription factors
NPM1 (27%)
Tumor suppressors (16.5%) *TP53*, *WT1*, *PHF6*
DNA methylation (46%) *TET1*, *TET2*, *IDH1*, *IDH2*, *DNMT3B*, *DNMT1*, *DNMT3A*
Cohesin complex (13%)
Activated signaling (59%) *FLT3*, *KIT*, *KRAS*, *NRAS*, PTPs, Ser/Thr kinases, other Tyr kinases

Figure 3. Mutation and cytogenetic landscape of acute myeloid leukemia. Adapted from Chen and Chen *et al*., Nature Genetics, 2013[143].

The conventional chemotherapy regimens are not suitable for unfit elderly AML patients. Therefore, the treatment options were confined to hypomethylating agents decitabine and azacytidine for the patients until recently. FDA approval of the combination of the BCL-2 inhibitor venetoclax and a hypomethylating agent or low dose cytarabine has revolutionized treatment for the elderly AML patients. The treatment has improved median overall survival from 11.3 to 17.5 months in AML patients with a median age of 75 years of age and older, was reported in a clinical trial[156]. Similarly, recent FDA approvals of other targeted agents have made new treatment options available for patients with specific molecular lesions. Approval of FLT3 inhibitors midostaurin and gilteritinib for AML patients with activating *FLT3* mutations including internal tandem duplication (*FLT3*-ITD) or point mutation of the tyrosine kinase domain (*FLT3*-TKD) at diagnosis bring new treatment options for 30% of the AML population. Two IDH inhibitors, ivosidenib and enasidenib, were approved for AML patients carrying *IDH1* and *IDH2* mutations, respectively for relapsed/refractory AML patients. Smoothened inhibitor glasdegib targeting the hedgehog pathway was also approved for the treatment of AML patients. Furthermore, the CD33 antibody gestuzumab ozogamicin was approved for *CD33* expressing AML patients. Approval of targeted drugs has dramatically changed the treatment paradigm for the mutation-specific subgroup of AML patients. However, limited response rates with targeted drugs in mutation stratified patient populations demands the development of advanced strategies and robust biomarkers to identify patients most likely to respond.

4.4.2 Apoptosis

Apoptosis blockade is one of the hallmarks of the cancer[157]. Apoptosis is known as programmed cell death, was reported for the first time in mammalian tissue by Kerr *et al.* in 1970s[158]. In this cellular process, the cell receives death signal results in cell disintegration into small apoptotic bodies that are eventually phagocytosed by white blood cells[159]. Apoptosis pathway is characterized by two distinct mechanisms (intrinsic and extrinsic) depending on the source of death signals. In the case of the extrinsic pathway, the death signal bind to cell surface receptors, including Fas cell surface death receptor (FAS), Tumor necrosis factor receptor (TNFR), and TNF Receptor Superfamily Member 25 (WSL) and activate downstream caspase cascade. The activated caspase 8 cleaves and activates BH3 protein BID to induce intrinsic pathway[160]. On the contrary, the intrinsic pathway occurs in mitochondria upon cellular damage and is activated through B-cell lymphoma 2 (BCL-2) family proteins. The proteins regulate outer membrane permeabilization (MOMP) of cytochrome c and SMAC proteins to the cytosol, thereby activating downstream caspases and leading to apoptosis[161].

After unrevealing the crucial role of BCL-2 family members in regulating apoptosis, many studies have unfolded important roles of distinct pro-and anti-apoptotic BCL-2 family proteins [162-165]. The anti-apoptotic BCl-2 family proteins include BCL-2, BCL-X_L, BCL-W, MCL-1 and BFL-1 and carry four BH-domains. The pro-apoptotic proteins were divided into two categories based on protein structure BH domains carrying proteins and BH3 only proteins. BAK, BAX and BOK proteins contain three to four BH domains whereas BAD, BIK, BIM, PUMA, NOXA, HRK and BMF proteins contain BH3 domain[166-173]. The interaction between anti-apoptotic and pro-apoptotic proteins occurs through binding of the BH3 domain of pro-apoptotic proteins to BH1, BH2 and BH3 domains of anti-apoptotic proteins. Upon activation, BAK and BAX oligomerize on the mitochondrial outer membrane to form pores and cause MOMP and cytochrome c release in the cytosol for progression of apoptosis[166,167,174]. Therefore, BCL-2 family members and especially BH3 related proteins have become an exciting target for therapy development.

4.4.3 Venetoclax as effective BCL-2 inhibitor

Several drugs have been developed to target BCL-2 and induce apoptosis in leukemic blast cells that dependent on BCL-2 for survival[175]. BCL-2 inhibitors are BCL-2 homology domain 3 (BH3) like small molecules, also known as BH3 mimetics. These drugs bind to anti-apoptotic molecules, including BCL-2, BCL-X_L and BCL-W and BCL2-A1. BH3 mimetic drug navitoclax (ABT-263) binds with a strong affinity to BCL-2 and BCL-X_L and with weaker affinity to MCL-1 and BCL2-A1[176]. The target binding affinity of navitoclax with BCL-2 and BCL-X_L is <1nM. Despite the promising pharmacokinetic properties, the navitoclax treatment produced severe thrombocytopenia due to survival dependence of thrombocytes on BCL-X_L[177]. The next BH3 mimetic in the drug development pipeline was venetoclax (ABT-199)[178]. Venetoclax was approved for CLL patients in April 2016. In the case of AML, the clinical trials resulted in a remarkable response in elderly AML patients in combination with hypomethylating agents[156]. The venetoclax in combination with azacitidine or decitabine or low-dose cytarabine received FDA approval for elderly patients diagnosed with AML in November 2018. Venetoclax has a strong and selective binding affinity for BCL-2 (Figure 4) and had no impact on platelet survival[178]. However, it has been challenging to select AML patients for venetoclax treatment and monitor treatment response in the clinic.

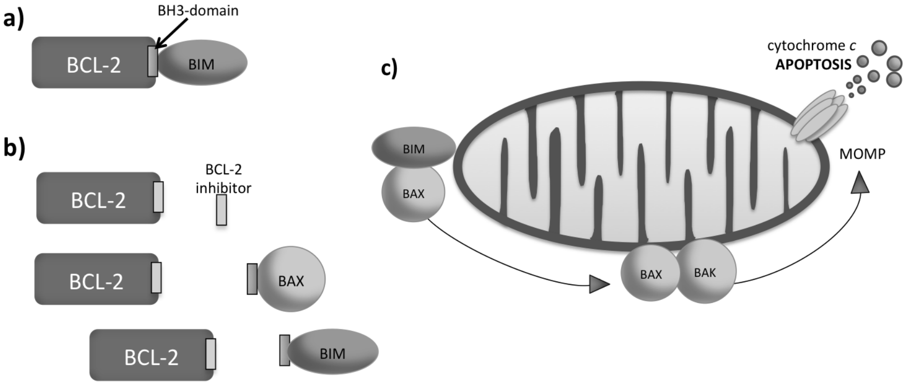


Figure 4. Venetoclax mode of action. a) the balance between anti-apoptotic and pro-apoptotic proteins in healthy cells. b) BCL-2 inhibitor prohibits binding of pro-apoptotic proteins BAK and BAX to anti-apoptotic BCL2 on BH3 domain to promote apoptosis. c) Activated pro-apoptotic proteins trigger permeabilization of mitochondrial membrane to releases cytochrome c to execute the final step of apoptosis.[179] Abbreviation: MOMP: mitochondrial outer membrane permeabilization. Adopted from Konopleva *et al*. Cancer Discovery, 2016[180] with permission from AACR.

4.4.4 Biomarkers for venetoclax response

Genetic biomarkers for venetoclax sensitivity and resistance have been reported in solid tumors and hematological malignancies. As an apparent biological correlate, *BCL-2* gene overexpression was reported to be associated with venetoclax sensitivity in small cell lung cancer[181]. Chyla *et al*. reported that *IDH1*, *IDH2* and *SRSF2* mutations correlate with clinical response to venetoclax in AML patients[182].

Due to the lack of stability of the genetic and gene expression biomarkers for venetoclax responses, Konopleva M. and Letai A. have developed a functional assay known as BH3 mimetic assay. The assay measures the apoptotic potential of the cells to determine the corresponding response to BH3 mimetic venetoclax. The assay was reported as a robust biomarker which can be implemented in the clinic to predict venetoclax sensitivity by Pan *et al*. [102]. Later, Konopleva *et al*. reported phase II clinical trial results of venetoclax monotherapy in 32 AML patients. The BH3 profiling assay results were consistent with clinical response in the patients. The 800 mg clinical dosage of venetoclax was well tolerated in AML patients without causing any severe side effects[180]. The functional BH3 mimetic assay can be a promising biomarker for venetoclax response. However, current molecular profiling dependent clinical practices need robust genomic and transcriptomic biomarkers for predicting sensitivity and resistance to venetoclax.

5 AIMS OF THE STUDY

This thesis addresses the challenge of utilizing transcriptomics information from leukemia patients to aid precision medicine strategies. The specific aims of this study were:

- To evaluate the advantages and limitations of two mainstream RNA-seq library preparation methods for optimal selection (Study I)

- To identify gene expression biomarkers for BCL2 inhibitor venetoclax sensitivity in AML patients (Study II)

- To identify biomarkers and therapeutic strategies to counteract BCL-2 inhibitor venetoclax resistance in AML patients (Study III)

6 MATERIALS AND METHODS

6.1 Patient material

Bone marrow (BM) aspirates or peripheral blood samples were collected from AML patients. Skin biopsies were also collected (non-malignant cells for germline genomic information) from AML patients. All samples were collected with the approval of Helsinki University Hospital Ethics Committee (permit numbers 239/13/03/00/2010, 303/13/03/01/2011, Helsinki University Hospital Ethics Committee) and after signed informed consent in accordance with the Declaration of Helsinki. All of the AML patients included in studies I-III were venetoclax treatment naïve. Mononuclear cells (MNCs) were isolated by Ficoll-Paque PREMIUM density gradient separation (GE Healthcare). MNCs were further used for drug sensitivity and resistance testing (DSRT) and extraction of nucleic acids (DNA and RNA).

6.2 RNA-sequencing

Total RNA (2.5-5 µg) was extracted from BM MNCs using the miRNeasy kit (Qiagen). The Qubit fluorometer (Thermo Fisher) was used for RNA quantification and RNA quality was measured using Bioanalyzer with RNA nanochips (Agilent). Next, RNA-seq libraries were prepared using Dynabeads® mRNA Purification Kit (Thermo Fisher) and using the Ribo-ZeroTM rRNA Removal Kit (Epicentre) as per the manufacturer's instructions. The RNA was further reverse transcribed to double-stranded cDNA (SuperScript™ Double-Stranded cDNA Synthesis Kit, Thermo Fisher). RNA sequencing libraries were prepared with Illumina compatible Epicentre Nextera™ Technology and ScriptSeq v2™ Complete kit (Illumina). RNA sequencing libraries were purified with SPRI beads (Agencourt AMPure XP) and library QC was evaluated on high sensitivity chips using the Agilent Bioanalyzer (Agilent Technologies). Paired-end sequencing with 100 bp read length was performed using HiSeq 2000 (Illumina).

6.3 Exome-sequencing

Genomic DNA was isolated from both BM MNCs and skin biopsies from AML patients using the DNeasy Blood & Tissue Kit (Qiagen). 3 µg of DNA was used for the sequencing. Exome capture was performed with the Nimblegen SeqCap EZ v2 capture Kit (Roche NimbleGen, Madison, WI, USA) and sequencing was done on HiSeq1500, 2000 or 2500 instruments[98]. Data preprocessing (QC, alignment) and somatic mutation calling were done as described previously[183].

6.4 *Ex vivo* drug sensitivity and resistance testing (DSRT)

The drug sensitivity and resistance testing (DSRT) was performed in a high-throughput setting with five-point doses of each drug as described previously[98]. The drugs were dispensed in nanoliter volumes to 384 well plates using Echo 500 and Echo 550 acoustic liquid dispensing system (Labcyte). The freshly isolated MNCs from healthy donors and AML patients were resuspended in mononuclear cell medium. The cells suspension was dispensed in pre-drugged plates using Multidrop (Thermo Scientific) and incubated for 72 hours at 37° C. The cell viability was measured as a surrogate of ATP production by live cells in terms of luminescence using Cell Titre Glow® reagent (Promega). Benzonthenium chloride as positive and dimethyl sulphoxide as negative control were added to multiple wells in every plate. To read the plates PHERAstar FS plate reader (BMG LABTECH) was used. The controls were used to calculate percent inhibition for each of five doses of a given drug. These values were used to fit four-parameter non-linear regression curves and to calculate half inhibitory concentration (IC50). The curve fitting criteria and IC50 values were used to calculate drug sensitivity scores (DSS), which is a modified area under the curve. The cancer-specific drug responses as selective DSS were calculated by subtracting responses from healthy controls as described previously[184].

6.5 RNA-sequencing data analysis

RNA sequencing data were pre-preprocessed as described previously[185]. Briefly, Trimmomatic[105] was used to correct reads for low quality, Illumina adapters, and short read-length. Filtered paired-end reads were aligned to the human genome (GRCh38) using STAR aligner[109] with the guidance of EnsEMBL v82 gene models using 2-pass per-sample parameters were used with the overhang on each side of the splice junctions set to 99. The alignments were sorted and PCR duplicates were marked using Picard, feature counts were computed using SubRead[108] and converted to expression estimates using Trimmed Mean of M-values (TMM) normalization[186]. Fusioncatcher[118] tool was used to call fusion genes using RNA-seq fastq files. In study II, to find out the differentially expressed genes between venetoclax sensitive and resistant groups, DESeq2[115] package was used. In study III, we applied a linear regression model[187], assuming that gene expression is affected by confounding factors including age, gender, sequencing batch, RNA extraction method and RNA-sequencing library preparation protocols. To find the relationship between gene expression change and drug sensitivity, we corrected the confounding factors in the linear regression model. Genes at false discovery rate (FDR) < 0.05 were considered significant.

6.6 Pathway and network analysis

Pathway analysis was performed using QIAGEN's Ingenuity® Pathway Analysis (www.qiagen.com/ingenuity). The method was applied to genes with ≥ 2-fold change in expression. Z-scores > abs (2) was considered as significant. In addition to IPA, the Gorilla[188] web-server was used to identify enriched terms across all three-gene ontologies. In study III, network analysis was performed for negatively associated genes (n=349) with venetoclax response using the Enrichr[189,190] tool. Outputs from KEGG 2016 and Reactome 2016 cell signaling pathway databases were considered for further analysis. GeneMANIA[191] was used for visualizing a sub-cluster of 29 genes associated with venetoclax resistance.

6.7 Quantitative reverse transcription-PCR (RT-qPCR)

For the RT-qPCR validation experiments, RNA was isolated either from AML patients and from cells lines. The Qubit fluorometer was used for the RNA quantification. The SuperScript III Reverse Transcriptase (Thermo Fisher) was used for the cDNA synthesis. The reaction was run on the

CFX96 Real-Time System instrument (Bio-rad) using 10ng cDNA from each sample, including the iQ SYBR Green SuperMix (Bio-Rad, Hercules, CA, USA). Data were analyzed by applying $\Delta\Delta C_t$ method using reference genes.

6.8 Statistical analyses

R version 3.3.3 was used for all the statistical analyses and for generating plots. Pearson's correlation coefficient assessed statistical dependence between two variables. The Mann Whitney U test was used for analyzing differences between drug responses. Two-sided P-values below 0.05 and false discovery rate (FDR) below 0.05 were considered statistically significant. The Wilcoxon signed-rank test was applied to find significant differences in drug response between wild type and mutated AML patient samples for a given mutation.

7 RESULTS

Detailed results from published articles are presented in their original communications and briefly summarized here.

7.1 Impact of RNA-seq protocols (Publication I)

We compared two major library preparation protocols, poly-A enrichment (PA) and ribo-depletion (RD) used for clinically relevant molecular features. The protocols applied to the total RNA isolated from mononuclear cells from the bone marrow of leukemia patients. Eight libraries were generated from two AML and two acute lymphoblastic leukemia (ALL) patient samples, including experimental replicates and technical replicate from the same total RNA sample (Figure 5).

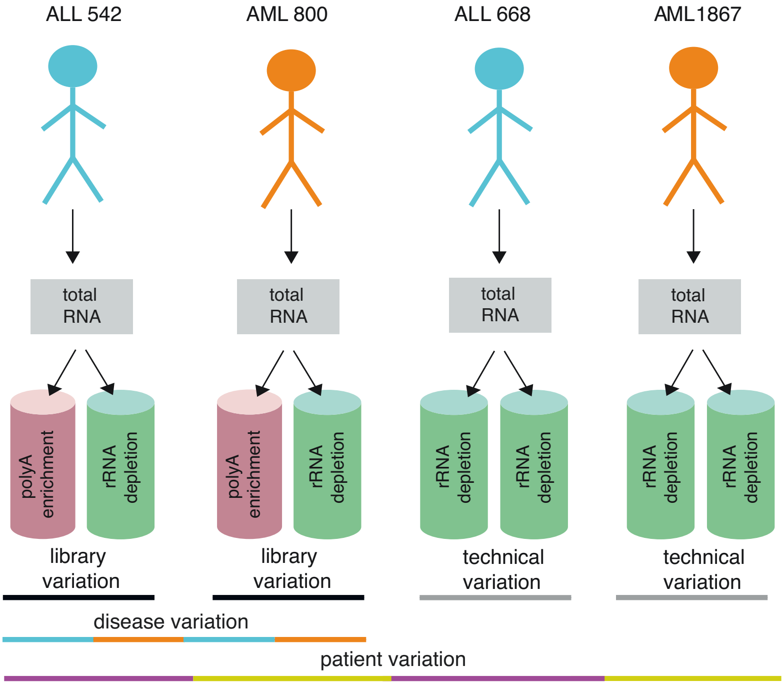


Figure 5. A systematic overview of the experimental design. Total RNA was isolated from mononuclear cells and used for RNA-seq library preparation.

We evaluated the utility of two mainstream protocols to detect clinically relevant molecular characteristics and assessed their effects on based on different analyses including;

i. Read mapping; the PA protocol provided a higher number of exon mapping reads (75.2-76.9%) than the RD libraries (52.0-72.6%). Reads mapping to intronic regions were higher in RD (33.8%) than in PA (21%) (Figure 6a).

ii. Expression of protein-coding and non-coding RNAs; the RD protocol detected 20.8 to 26.3% more features altogether compared to PA libraries. The list of genes includes both protein-coding and non-coding (processed pseudogene, lincRNA, snRNA, antisense and miRNA) RNAs. In the case of protein-coding genes, 1380 of them were, discordantly called between the matched PA and RD libraries. The PA protocol overlooked 55 histone genes and on the contrary, many cancer-related genes e.g., *TGF-β1*, *BCL3*, *BRD4* were overlooked by RD protocol.

iii. rRNA removal efficiency; the PA libraries had higher rRNA mapping read rates than RD libraries (1.8% vs. 0.6%) (Figure 6b).

iv. The hierarchical clustering of highly variable genes depicted the groups driven by disease biology instead of protocols/technical variation (Figure 6c).

v. Differential gene expression; to find out the impact of library preparation protocol on differential gene expression, an independent RT-qPCR experiment was performed. The expression analysis of randomly selected five oncogenes (*POLR1B*, *TUBB*, *SRM*, *TGFB1*, *NABP1*) revealed that the PA protocol captured target mRNAs more efficiently than the RD protocol. The PA protocol captures target mRNA to a greater extent compared to RD protocol. *STAT3*, *NABP1* and *TET2* were depleted significantly in the PA enriched library and *NRAS*, *STAT3*, *TET2*, *EMD*, *SRM*, *TGFB1*, *ZFP36L2* showed a significant difference between the RD library and total RNA.

vi. Fusion gene detection; FusionCatcher[118] tool was applied to the evaluation of the efficiency of PA and RD protocols in detecting fusion genes. The clinically relevant fusions with well-known roles in leukemia diagnosis and prognosis. For example, *BCR-ABL1* in-frame fusion gene that was supported by 184 and 188 spanning pair-end reads in PA and RD.

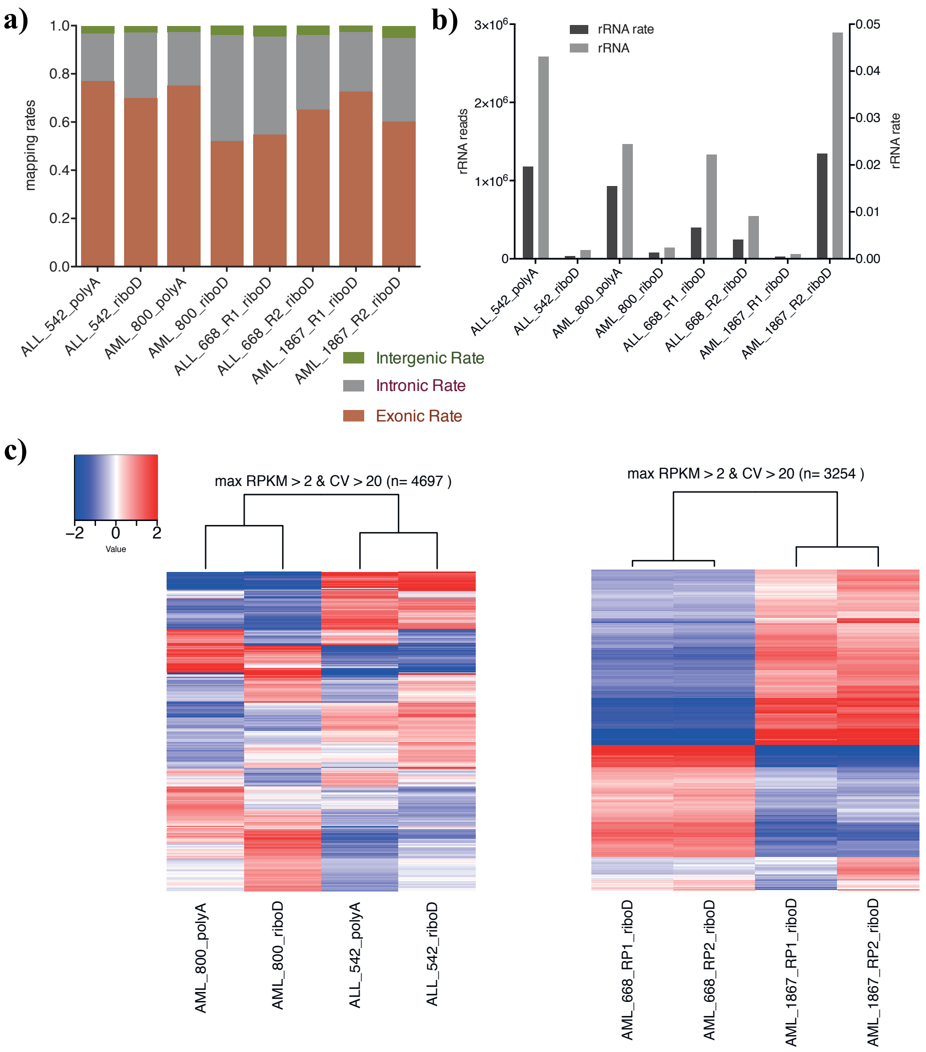


Figure 6. a) The percentage of reads mapped to intragenic, intronic and exonic regions, read mapping rates are on Y-axis. b) Y-axis on the left represents number of rRNA reads and Y-axis on the right represents rRNA mapping rates. c) The first heatmap represent the disease and library specific variations, the second heatmap represents the technical variation. For the hierarchical clustering genes (log2 RPKM >2 and CV >20) were selected using Euclidean distance and complete linkage.

Overall, both protocols produced similar results with consistent outcomes. We found that RD protocols capture whole transcriptome information, detects a higher number of ncRNA features (snRNA, processed pseudogene, lincRNA, miRNA). Also, it had higher alignment and gene coverage efficiency, depleted some protein-coding mRNA and MT genes and removed rRNA effectively. On the other hand, PA protocol captured more protein-coding regions, closely represented gene expression values with total RNA, depleted histone mRNA and lost targeted RNAs of interest if lack poly-A tails, efficient in differential gene expression analysis.

7.2 Biomarkers for venetoclax response (Publication II and III)

To identify genomic and transcriptomic biomarkers associated with venetoclax sensitivity and resistance, DSRT was performed with mononuclear cells isolated from the bone marrow of AML patients. Also, WES (n=42) and RNA-seq (n=35) was performed with same MNCs.

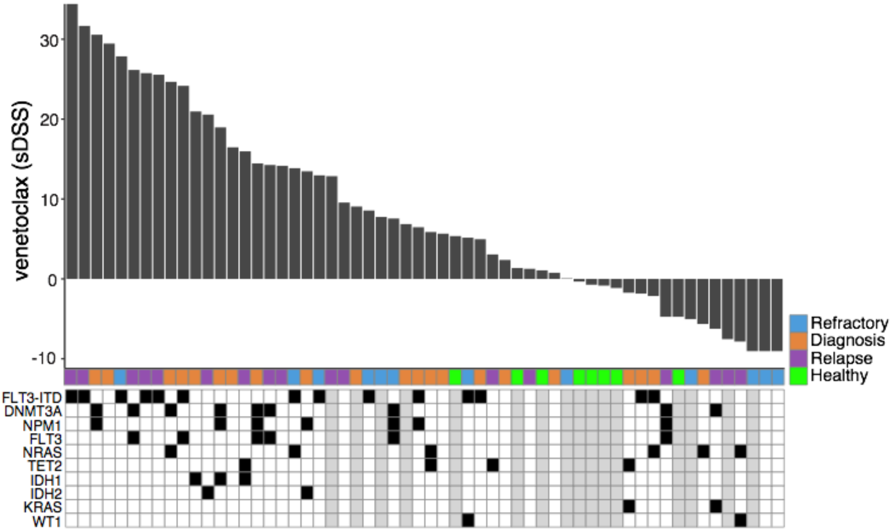


Figure 7. Waterfall plot illustrates the selective venetoclax response profile in 50 samples from AML patients and eight healthy controls (green). The sDSS represents leukemia-selective responses compared to healthy controls, where high sDSS represents strong sensitivity. Each sample is annotated for the disease type and presence of key AML somatic mutations.

In study II, Venetoclax produced heterogeneous response across AML patient samples (Figure 7). The differential gene expression analysis was performed using RNA-seq read counts data. The analysis was applied between samples that were highly resistant (n=4) and highly sensitive (n=3) to venetoclax response (Figure 8a). The analysis resulted in 322 significant differentially expressed genes (FDR <0.05). Out of 322 genes, 41 of them were overexpressed in the sensitive group (Figure 8b) and 281 genes were overexpressed in the resistant group. The genes were further analyzed for their biological function and class. The analysis revealed several *HOX* family genes with significantly higher expression in venetoclax-sensitive compared to resistant samples. Furthermore, we also confirmed the overexpression of the *HOX* family genes using RT-qPCR in AML patient samples. Moreover, *IDH1/2* and *WT1* mutations were found to be enriched in venetoclax sensitive patient samples compared to resistant samples.

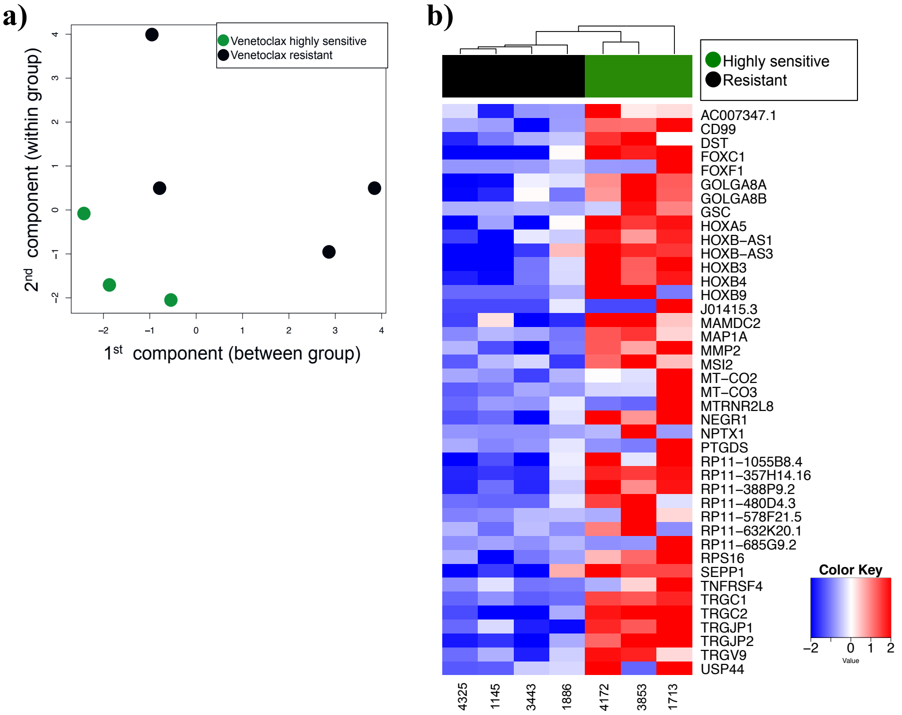


Figure 8. a) Multi-dimensional scaling plot exhibits differences between the expression profiles of three venetoclax sensitive and four venetoclax resistant samples in two dimensions. b) 41 significantly with false discovery rate (FDR <

0.05) overexpressed (>log 2-fold) genes in the sensitive group as compared to the resistant group.

In study III, we aimed to identify gene expression biomarkers for venetoclax resistance. Availabilityof a larger sample set, especially gene expression data, allowed to apply machine learning analysis. The analysis was performed between gene expression and venetoclax responses.

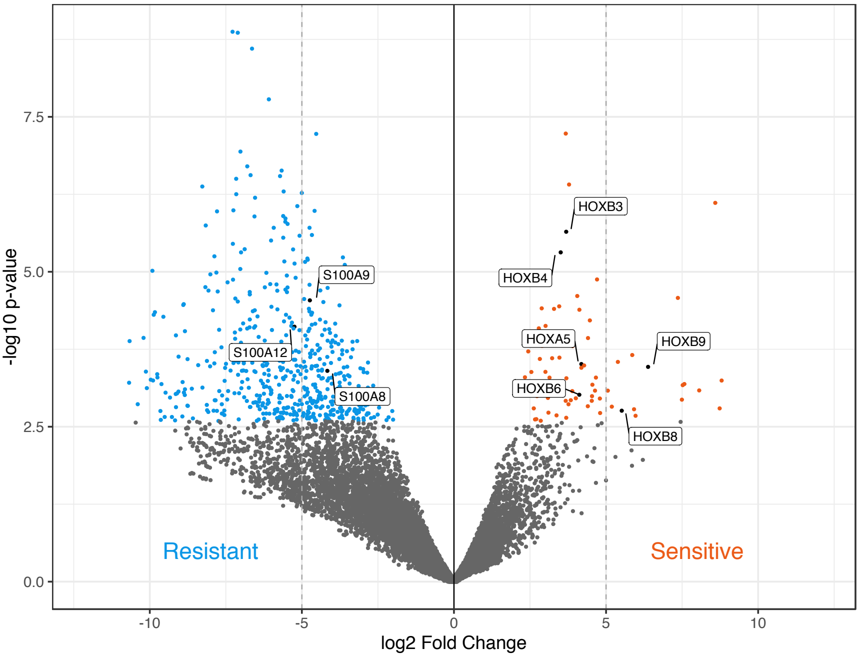


Figure 9. Volcano plot highlighting differentially expressed genes in venetoclax sensitive and resistance samples, respectively. The linear regression analysis was performed between venetoclax response (sDSS) and protein-coding genes (n=19,220) by correcting for the possible technical covariates (e.g., gender, RNA-seq library preparation method, RNA extraction kits). The *S100* genes were further validated at expression (RT-qPCR) and protein (Western blot) level in venetoclax resistant AML patient samples.

The analysis resulted in 601 significantly associated genes (FDR < 0.05). Of these genes, 252 were positively and 349 negatively associated with venetoclax response. The positively associated genes included five *HOX* family genes, namely *HOXB5*, *HOXB6*, *HOXB7*, *HOXB8* and *HOXB9*, confirming findings from study II[58]. In contrast, amongst the most significantly associated genes were three *S100* family genes, namely *S100A6*, *S100A8*, and *S100A9* (FDR < 0.05). Taken together, HOX and S100 family genes were overexpressed in venetoclax sensitive and resistance AML samples, respectively (Figure 9).

Furthermore, we aimed to identify possible drugs that effectively target venetoclax resistant patient samples. The expression of *S100A8* and *S100A9* genes was correlated with response to 349 approved drugs and emerging investigational chemical compounds. The analysis resulted in a positive correlation between *S100A9* gene and BET inhibitor birabresib. Considering that birabresib re-sensitizes AML cell to venetoclax, the combination of birabresib and venetoclax was tested. Intriguingly, we found strong synergy between birabresib and venetoclax in *S100A8/A9* overexpressing, venetoclax resistant AML patient samples and cell lines.

8 DISCUSSION

Evolution of technologies to study gene expression has opened new avenues to understand disease biology and identify biomarkers to monitor therapy responses. The increased sensitivity and specificity of technologies measuring gene expression changes now reveal the subtle changes in transcriptomes and disease mechanisms at unprecedented precision. Several sophisticated methods and protocols are emerging to quantify gene expression patterns. Therefore, it is essential to evaluate and compare each method and protocol for optimal selection of the method suitable for clinical application under the modern-day precision medicine setting.

It is widely accepted and biologically meaningful to study gene expression signatures at transcriptome level compared to studying expression changes in individual genes to understand the changes upon treatment better. This speculation is supported by the fact that genes coordinate and function in networks rather than individually. For example, genes belonging to the same family can co-operatively activate or deactivate certain cellular processes. Thus, detecting gene expression changes can lead to the identification of novel biomarkers for diagnosis, prognosis and drug response. Notably, the biomarkers for drug responses are most useful to monitor therapeutic effect and development of drug resistance in patients receiving molecularly guided treatments. Identifying disease biomarkers and associated gene expression changes not only provides insights on drug responses but may also help to understand the mechanism of resistance development, enabling us to build strategies to avoid the development of resistance. Despite the extensive efforts to discover biomarkers, the gene expression biomarkers used in routine clinical practice are negligible. This indicates the need to implement powerful bioinformatics and statistical approaches to investigate robust biomarkers for drug responses systematically.

Study I – RNA-seq library preparation protocol comparison

Library preparation is one of the most critical steps of RNA-seq analysis. The selection of library preparation protocol is known to affect downstream analysis and can hinder the interpretation of RNA-seq findings[71,85,87]. Also, inconsistencies between RNA-seq data generated from different library preparation protocols have been reported in many studies[86,192]. So far, the performance of RNA-seq library preparation protocols has mostly been reported from non-clinical samples[82,86,87]. To find out the impact of library preparation protocols on leukemia patients, we applied two mainstream library preparation protocols. We systematically compared the performance of the protocols to detect clinically relevant gene expression changes. We also conducted independent RT-qPCR experiments and analyzed which protocol produced mRNA expression estimates more similar to those observed in total RNA. The small number of patient cases is a limitation of our study towards getting statistically significant conclusions e.g., differential gene expression and pathway analysis.

The PA libraries provided a higher fraction of exome mapping reads and lower number of intronic reads compared to RD libraries. The intronic reads originated from immature, mostly non-spliced transcripts[193]. Immature transcripts are pre-mRNAs meaning that the RNA polymerase has not yet attached to the 3′ end of the gene. Recently, it was shown that in case of the RD protocol, intronic reads come primarily from the immature transcripts such as nuclear RNAs[194,195]. A small portion of intronic reads in case of PA protocol might represent background oligo(dT) priming of adenines in primary transcripts, rather than true polyadenylated transcripts[195]. Wetterbom *et al.* have also suggested that mRNA purification by PA protocol was not completely efficient[196]. Since the PA protocol covered more exonic regions than the RD protocol, it positively affects differential expression analysis[197]. Hence, the PA protocol is the preferred method to identify differentially expressed genes between two or more conditions. However, a higher fraction of intergenic and intronic reads gives information on pre-mRNA dynamics and novel transcripts[198].

In our study, many protein-coding and non-polyadenylated genes that were missed by the PA protocol were detected by RD protocol. In accordance with a previous study there gene included histone genes[85]. While RD protocols captured more coding and non-coding transcripts, many known oncogenes were missed, including *TGF-β1*, *BCL3* and *BRD4,* reported to be associated with leukemia development[199-201]. PA protocol can be better suited for characterization of leukemia transcriptomes. However, the results of the study I were acquired from a small number of patient cases and therefore, the conclusion is drawn here needs to be interpreted with caution.

Finally, which RNA-seq protocol should one select? The choice of the optimal RNA-seq protocol will strongly depend on the quality and quantity of input material[202,203]. In general, if the input RNA is intact and extracted from fresh biopsies for examples bone marrow or peripheral blood, most protocols will produce high-quality RNA-seq data. However, in most cases, the PA protocol is recommended as they will provide the best interoperability. Moreover, PA outperforms RD in many clinically relevant assessments, including gene expression analysis, classification of leukemia patients, quantification of leukemic marker genes, and variant analysis. If RNA integrity is compromised or PA protocol is not possible and the main experimental focus is on non-coding transcripts, then the RD protocol is recommended. Thus, the objectives of the study should guide the selection of the RNA-seq library protocol.

Study II and III - Biomarkers for sensitivity and resistance to BCL-2 inhibitor venetoclax

In study II, we sought to identify gene expression biomarkers for BCL-2 inhibitor venetoclax response in AML patients. The differential gene expression analysis between venetoclax sensitive and resistant samples revealed *HOX* family gene overexpression in venetoclax sensitive samples[58]. Homeobox (*HOX*) are evolutionarily conserved genes encodes for transcription factors. These factors are known to play a fundamental role in embryonic development, including cell differentiation[204-206]. Early hematopoietic progenitor cells highly express *HOX* genes during hematopoietic maturation but the expression levels eventually halt in differentiated cells[207]. The known upstream deregulators of *HOX* gene overexpression are fusions involving *HOX* genes (*NUP98-HOX*) and *MLL* rearrangements[208], often found in leukemia patients. The survival of undifferentiated leukemic stem cells was shown to be dependent on BCL-2[209]. The deregulation of *HOX* gene expression was found to be restricted in progenitor cells[210]. Venetoclax treatment can effectively target the leukemic stem/progenitor cell population considering *HOX* gene deregulation and BCL-2 dependence found in this cell population. This is in consistence with our finding revealing that *HOX* gene overexpression is associates positively with venetoclax sensitivity, representing BCL-2 dependence in AML patients. Hence, *HOX* gene expression can be developed as a biomarker for selecting AML patients for effective venetoclax therapy outcomes with further testing under clinical settings.

In study III, we investigated possible gene expression markers associated with *ex vivo* resistance to BCL-2 inhibitor venetoclax. We applied a machine learning approach, since a large data-set of RNA-seq from AML patient was available. We observed that *S100* family genes were associated with venetoclax resistance. *S100* family genes encode for 21 calcium-binding proteins described to play crucial roles in cell proliferation, differentiation and inflammation[211,212]. S100 protein family are the largest group of calcium-binding proteins and the S100 name was given based on the fact that solubility of S100 proteins occurs in a 100% saturated ammonium sulfate solution at pH 7[213]. We observed that *S100A8* and *S100A9* genes were significantly associated with venetoclax resistance amongst other family members. Expression of *S100A8* was reported to be a poor prognostic factor as well as to be associated with chemo-resistance[214]. Similarly, mitochondrial priming for apoptosis predicts response to cytotoxic chemotherapy[215]. The interaction analysis using GeneMania webtool showed co-expression patterns between *S100A8/A9* and *BCL-2* gene family gene *BCL2A1*. This may indicate that co-overexpression of *BCL-2* and *S100* family genes confer resistance to venetoclax in AML cells. However, the exact mechanism of interaction between these genes needs further mechanistic investigation.

BET family inhibitor mivebresib (ABBV-075) was reported to modulate the apoptosis and synergize with venetoclax efficacy in AML cell lines, patient cells and mouse xenograft model[216,217]. A phase I clinical trial (NCT02391480) has been set to test a combination of venetoclax and mivebresib in patients with solid tumors and AML. In this study, we observed that birabresib (OTX-015) mediated BET inhibition reduced protein levels of BCL-X_L and BCL-2 protein levels in cell lines overexpressing *S100A8* and *S100A9*. This finding is in line with the published study suggesting that BET inhibitor treatment downregulates the expression of anti-apoptotic proteins BCL-X_L, BCL-2 and MCL-1[217]. Furthermore, we observed that birabresib treatment led to upregulation of pro-apoptotic protein BIM in AML patient samples. Moreover, upregulation of *BIM* has been shown upon JQ1 and mivebresib treatment in AML cell lines and other cancers[217-220]. Thus, with further investigations, the *S100* gene expression pattern could potentially be used to identify patients that would benefit from a combination of BET and BCL-2 inhibitors.

Moreover, mutation patterns were also investigated in venetoclax sensitive and resistant AML patient samples. *IDH1/2* mutations were found to be enriched in samples with high *ex vivo* sensitivity to venetoclax in study II. The *IDH* mutations related to *BCL-2* dependence was also reported by other studies[221,222], which were designed to target *IDH1* and *IDH2* mutated AML patients with BCL-2 inhibitors. Isocitrate dehydrogenase (*IDH*) 1 and 2 enzymes, upon mutation, produce oncometabolite 2-HG that subsequently causes hyper-methylation of several downstream genes[223-225], including *HOX* gene expression regulation and impaired metabolism that may lead to BCL-2 dependence[221]. Thus, *IDH1* and *IDH2* mutation may become biomarkers for venetoclax sensitivity with further investigation.

Taken together, this thesis aimed to utilize gene expression information for advanced precision medicine outcomes in patients with hematological malignancies. In study I, the contemporary mainstream library preparation protocols, ribo-depletion and polyA enrichment used for RNA-seq, were compared in order to select an optimal protocol that suffices the goal of the experiment, especially in patients with acute leukemias. In study II, we applied bioinformatics approaches to identify *IDH1/2* mutation and *HOX* family gene expression correlated with *ex vivo* sensitivity to BCL-2 inhibitor venetoclax in AML patients. In study III, statistical and machine learning methods were implemented to identify *S100A8/A9* gene expression biomarkers for *ex vivo* resistance to venetoclax in AML patients. In summary, this thesis addresses the challenges of utilizing gene expression information to stratify patients based on biomarkers to promote precision medicine practice in hematological malignancies.

9 FUTURE PERSPECTIVES

The precision medicine field has been advancing with the development of new technologies and these advancements continue to improve overall patient outcomes. Also, the declining cost of sequencing and rapidly growing computational resources for big data analysis enables better implementation of these techniques for real-time diagnosis and treatment decision making in the clinic. Moreover, where current clinical practices are limited to sequencing efforts, an adaptation of more sophisticated analyses, including circulating tumor DNA sequencing, immuno-profiling, functional assays to predict and monitor treatment responses and cancer progression. In the future, it will be possible to collect complete molecular profiling data from the same time points to explore various angles of complex disease information. Emerging deep learning and artificial intelligence models have been bringing forth cutting-edge tools to combine data from multiple biological sources to deconvolute information on molecular profiles, survival prognosis, drug dose selection and phenotypes of cancer cells.

The transcriptomic studies have primarily contributed to increased understanding of functional relationship and gene regulation aspects of cancer cells, thereby bridging the gap between genotype and phenotype. Current application of bulk RNA-sequencing has enabled us to discover therapeutic fusion genes (e.g., *BCR-ABL*), epigenetic deregulation pattern in molecular subsets, gene expression patterns associated with drug responses and prognosis. It has been demonstrated that gene expression signatures can efficiently proximate the cellular processes compared to single genes. However, application of single-cell RNA-seq can capture heterogeneity at molecular, functional and phenotype levels to underpin the detailed but comprehensive information. The clinical utility of single-cell sequencing to aid therapy selection or monitor treatment response needs well-designed strategies to utilize the potential of the information from the data. Translating cancer genome and transcriptome for patients will require continued multi-disciplinary collaboration between oncologists, pathologists, basic scientists, and computational biologists. Routine molecular profiling of cancer patients for basic genomics research, tumor sequencing in the clinic, and big data sharing networks would be essential to enable precision cancer medicine in practice.

10 ACKNOWLEDGEMENTS

This thesis work is carried out at the Institute for Molecular Medicine Finland (FIMM) as part of the Individualized System Medicine (ISM) Grand Challenge programme. I want to thank all colleagues at FIMM for making it an amazing and multi-cultural workplace.

My deepest gratitude goes to my supervisor Caroline Heckman for her continuous support and faith in me during these years. I admire her for giving constructive scientific inputs and freedom to learn computational skills at my own pace. Your open-mindedness for giving me a platform to work on my projects as well as on collaborative projects, ease of availability and your outgoing attitude has been very helpful. Your excellent writing skills and kindness have been inspiring, that I would like to take with me after my PhD. I am immensely grateful for giving me the freedom to explore my potential in research. I appreciate your efforts for creating a friendly environment within our group for easy collaboration and mutual growth.

My PhD work was part of personalized medicine (PM) for hematological malignancies collaborative effort. I had an excellent opportunity to interact and learn from PM group leaders; Caroline Heckman, Kimmo Porkka, Olli Kallioniemi, Tero Aittokallio and Krister Wennerberg. I am grateful to be part of this unique set-up impacting the patients′ lives in real-time. I am deeply inspired by visionary scientists Kimmo and Olli for my future career endeavors. I extend my thanks to Kimmo for the positive vibes at every interaction and great encouragement, which required to fuel team-work and great science.

I want to express my gratitude to my co-supervisor Pirkko Mattila for your support, availability and essential inputs into my studies. I owe a great deal of thanks to Simon Anders for valuable scientific guidance. Interacting with you has been very crucial for me to get a better perspective on computational and quantitative thinking, in most scientific way possible. Thank you for hosting me as an EMBO fellow at your lab in Heidelberg. I want to extend my heartful gratitude to Matti Kankainen for intriguing scientific discussions and writing, teaching the basics of exploratory data analysis and spending time on improvising my computational skills.

I am grateful to my thesis advisory committee members, Sampsa Hautaniemi and Antti Honkela for their critical inputs as well as encouragement throughout my thesis work. I thank Francesco Iorio and Throsten Zenz for reviewing my thesis and providing constructive insights.

I sincerely thank all my co-authors, without whom this work would have been impossible. I thank Mika Kontro for the collaborative work and for introducing me to the world of venetoclax and apoptosis. Thank you, Mika for providing invaluable patient samples and excellent clinical expertise. Heartfelt thanks to other co-authors; Matti Kankainen, Riikka Karjalainen, Minxia Liu, Alun Parsons, MM Majumder, Samuli Eldfors, Bhagwan Yadav, Disha Malani, Krister Wennerberg, Olli Kallioniemi, Kimmo Porkka and Caroline Heckman. I thank Alun for helping in RT-qPCR experiments and friendly gesture. I am grateful to the FIMM technology center and IT support; especially Pirkko Mattila, Pekka Ellonen and Henrikki Almusa.

It has been a pleasure to work with present and former members of the Heckman group. Special thanks to Mamun, Jarno, Riikka, Heikki, Dimitrios for sharing the office and having delightful scientific and philosophical conversations. Many thanks to my lab mates Alun, Minna, Siv, Samuli, Komal, Juho, Romika, Minxia, Joseph, Aino-Maija, and Alina for making the lovely atmosphere at work.

I was lucky to have fellow bioinformatician friends at FIMM; Guru, Bhagwan, Vishal, Shabbeer, Himanshu, Swapnil and John Patrick Mpindi for countless technical discussions and support. I am grateful to have Bhagwan as a colleague and friend, who accompanied in every science, sports and social activities. I thank Abhilash for all the advice and encouragement during challenging situations in my PhD. I thank Amit Singh for discussing machine learning concepts and your help during my stay at Heidelberg. I am delighted to be part of the Desi lunch community at FIMM; Swapnil, Himanshu, Sarang, Jatin, Ashwini Nagraj, Vishal, Poojitha, Khalid, Mahesh, Prson, Alok, Om, Sushil, Raghu, Ram and Bharat for the entertaining but useful day-to-day life conversations and social get-togethers.

I express my gratitude to Megha Purohit for her mentorship and motivating me for higher studies. Life is incomplete without good friends. I am fortunate to have friends; Gajraj, Maneesh, Surana, Preetam, Sandeep and Dharmveer. Thank you for sticking together in all my ups and downs throughout life.

I am thankful for all of the patients and healthy donors who participated in the PM research initiative. Without them, it would not have been possible to perform this study.

I am pleased to have received financial support from EMBO Short-term fellowship, Finnish Hematology Association, Cancer Society of Finland, Ida Montin Foundation, Doctoral Programme in Biomedicine (DPBM), American Society of Hematology (ASH), University of Helsinki, and FIMM. I am also thankful to DPBM for organizing exciting courses and symposiums to provide excellent doctoral training.

I sincerely thank my parents Saroj Devi and Rameshwar Dayal for love and care. They have always been supportive and allowed me to follow my dreams and ambitions. I am thankful to my grandfather, O. R. Yadav for friendly support at every stage of my life and nurturing my social skills. I thank my brother Kunal for good food, family-like atmosphere in Finland and accompanying me for sweating in the gym. I extend my gratitude to in-law family for moral support and care, especially Dharmik.

Finally, I want to thank my wife, Disha, who has exceptionally supported and encouraged me throughout this journey. Without her scientific, intellectual and motivational support, I may never have completed this thesis work.

Helsinki, October 2019
Ashwini Kumar

11 REFERENCES

1. National Research Council Committee on, A.F.f.D.a.N.T.o.D. The National Academies Collection: Reports funded by National Institutes of Health. in *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* (National Academies Press (US). National Academy of Sciences., Washington (DC), 2011).
2. Haendel, M.A., Chute, C.G. & Robinson, P.N. Classification, Ontology, and Precision Medicine. *New England Journal of Medicine* **379**, 1452-1462 (2018).
3. Desmond-Hellmann, S. Toward precision medicine: a new social contract? in *Sci Transl Med*, Vol. 4 129ed123 (United States, 2012).
4. Yates, L.R., *et al.* The European Society for Medical Oncology (ESMO) Precision Medicine Glossary. *Ann Oncol* **29**, 30-35 (2018).
5. Garraway, L.A., Verweij, J. & Ballman, K.V. Precision oncology: an overview. *J Clin Oncol* **31**, 1803-1805 (2013).
6. Collins, F.S. & Varmus, H. A New Initiative on Precision Medicine. *New England Journal of Medicine* **372**, 793-795 (2015).
7. Ashley, E.A. Towards precision medicine. *Nat Rev Genet* **17**, 507-522 (2016).
8. Katsnelson, A. Momentum grows to make 'personalized' medicine more 'precise'. in *Nature medicine*, Vol. 19 249 (United States, 2013).
9. Dolsten, M. & Sogaard, M. Precision medicine: an approach to R&D for delivering superior medicines to patients. *Clin Transl Med* **1**, 7 (2012).
10. Druker, B.J., *et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N Engl J Med* **344**, 1031-1037 (2001).
11. Saussele, S., *et al.* Impact of comorbidities on overall survival in patients with chronic myeloid leukemia: results of the randomized CML study IV. *Blood* **126**, 42-49 (2015).
12. Slamon, D.J., *et al.* Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med* **344**, 783-792 (2001).
13. Fukuoka, M., *et al.* Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non-small-cell lung cancer in Asia (IPASS). *J Clin Oncol* **29**, 2866-2874 (2011).
14. Rosell, R., *et al.* Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol* **13**, 239-246 (2012).
15. Cui, J.J., *et al.* Structure based drug design of crizotinib (PF-02341066), a potent and selective dual inhibitor of mesenchymal-epithelial transition factor (c-MET) kinase and anaplastic lymphoma kinase (ALK). *J Med Chem* **54**, 6342-6363 (2011).
16. Kwak, E.L., *et al.* Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N Engl J Med* **363**, 1693-1703 (2010).
17. Hauschild, A., *et al.* Dabrafenib in BRAF-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial. *Lancet (London, England)* **380**, 358-365 (2012).

18. Belin, L., *et al.* Randomized phase II trial comparing molecularly targeted therapy based on tumor molecular profiling versus conventional therapy in patients with refractory cancer: cross-over analysis from the SHIVA trial. *Annals of Oncology* **28**, 590-596 (2017).
19. Le Tourneau, C., *et al.* Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *The lancet oncology* **16**, 1324-1334 (2015).
20. Massard, C., *et al.* High-throughput genomics and clinical outcome in hard-to-treat advanced cancers: results of the MOSCATO 01 trial. *Cancer discovery* **7**, 586-595 (2017).
21. Stockley, T.L., *et al.* Molecular profiling of advanced solid tumors and patient outcomes with genotype-matched clinical trials: the Princess Margaret IMPACT/COMPACT trial. *Genome medicine* **8**, 109 (2016).
22. Von Hoff, D.D., *et al.* Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. *J Clin Oncol* **28**, 4877-4883 (2010).
23. Hyman, D.M., Taylor, B.S. & Baselga, J. Implementing Genome-Driven Oncology. *Cell* **168**, 584-599 (2017).
24. Nakagawa, H. & Fujita, M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci* **109**, 513-522 (2018).
25. Prasad, V. Perspective: The precision-oncology illusion. *Nature* **537**, S63 (2016).
26. Voest, E.E. & Bernards, R. DNA-Guided Precision Medicine for Cancer: A Case of Irrational Exuberance? *Cancer discovery* **6**, 130-132 (2016).
27. Lagana, A., *et al.* Precision Medicine for Relapsed Multiple Myeloma on the Basis of an Integrative Multiomics Approach. *JCO Precis Oncol* **2018**(2018).
28. Levis, M. Midostaurin approved for FLT3-mutated AML. *Blood* **129**, 3403 (2017).
29. Stone, R.M., *et al.* Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *N Engl J Med* **377**, 454-464 (2017).
30. Stein, E.M., *et al.* Enasidenib in mutant IDH2 relapsed or refractory acute myeloid leukemia. *Blood* **130**, 722-731 (2017).
31. DiNardo, C.D., *et al.* Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML. *N Engl J Med* **378**, 2386-2398 (2018).
32. Letai, A. Functional precision cancer medicine-moving beyond pure genomics. *Nature medicine* **23**, 1028-1035 (2017).
33. Rothwell, D.G., *et al.* Utility of ctDNA to support patient selection for early phase clinical trials: the TARGET study. *Nature medicine* (2019).
34. Sicklick, J.K., *et al.* Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. *Nature medicine* (2019).
35. Rodon, J., *et al.* Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. *Nature medicine* (2019).
36. Blom, K., Nygren, P., Alvarsson, J., Larsson, R. & Andersson, C.R. Ex Vivo Assessment of Drug Activity in Patient Tumor Cells as a Basis for Tailored Cancer Therapy. *J Lab Autom* **21**, 178-187 (2016).
37. Kodack, D.P., *et al.* Primary Patient-Derived Cancer Cells and Their Potential for Personalized Cancer Patient Care. *Cell Rep* **21**, 3298-3309 (2017).
38. Pauli, C., *et al.* Personalized In Vitro and In Vivo Cancer Models to Guide Precision Medicine. *Cancer discovery* **7**, 462-477 (2017).

39. Clevers, H. Modeling Development and Disease with Organoids. *Cell* **165**, 1586-1597 (2016).

40. Drost, J.*, et al.* Organoid culture systems for prostate epithelial and cancer tissue. *Nat Protoc* **11**, 347-358 (2016).

41. Majumder, B.*, et al.* Predicting clinical response to anticancer drugs using an ex vivo platform that captures tumour heterogeneity. *Nat Commun* **6**, 6169 (2015).

42. Snijder, B.*, et al.* Image-based ex-vivo drug screening for patients with aggressive haematological malignancies: interim results from a single-arm, open-label, pilot study. *The Lancet. Haematology* **4**, e595-e606 (2017).

43. Swords, R.T.*, et al.* Ex-vivo sensitivity profiling to guide clinical decision making in acute myeloid leukemia: A pilot study. *Leukemia research* **64**, 34-41 (2018).

44. Prasad, V. & Gale, R.P. Precision medicine in acute myeloid leukemia: Hope, hype or both? *Leukemia research* **48**, 73-77 (2016).

45. Iorio, F.*, et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754 (2016).

46. Behan, F.M.*, et al.* Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**, 511-516 (2019).

47. Tsherniak, A.*, et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e516 (2017).

48. Tyner, J.W.*, et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526-531 (2018).

49. Bode, A.M. & Dong, Z. Precision oncology- the future of personalized cancer medicine? in *NPJ Precis Oncol*, Vol. 1 2 (England, 2017).

50. Morash, M., Mitchell, H., Beltran, H., Elemento, O. & Pathak, J. The Role of Next-Generation Sequencing in Precision Medicine: A Review of Outcomes in Oncology. *J Pers Med* **8**(2018).

51. Zhao, S., Fung-Leung, W.P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9**, e78644 (2014).

52. Carninci, P.*, et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563 (2005).

53. van Dijk, E.L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet* **30**, 418-426 (2014).

54. Cao, J.*, et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).

55. Byron, S.A., Van Keuren-Jensen, K.R., Engelthaler, D.M., Carpten, J.D. & Craig, D.W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* **17**, 257-271 (2016).

56. Van Keuren-Jensen, K., Keats, J.J. & Craig, D.W. Bringing RNA-seq closer to the clinic. *Nat Biotechnol* **32**, 884-885 (2014).

57. Network, C.G.A.R. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine* **368**, 2059-2074 (2013).

58. Kontro, M.*, et al.* HOX gene expression predicts response to BCL-2 inhibition in acute myeloid leukemia. *Leukemia* **31**, 301-309 (2017).

59. Costello, J.C.*, et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* **32**, 1202-1212 (2014).

60. Chen, P.L.*, et al.* Analysis of Immune Signatures in Longitudinal Tumor Samples Yields Insight into Biomarkers of Response and Mechanisms of Resistance to Immune Checkpoint Blockade. *Cancer discovery* **6**, 827-837 (2016).

61. Lavallee, V.P., *et al.* RNA-sequencing analysis of core binding factor AML identifies recurrent ZBTB7A mutations and defines RUNX1-CBFA2T3 fusion signature. in *Blood*, Vol. 127 2498-2501 (United States, 2016).
62. Lilljebjorn, H., *et al.* RNA-seq identifies clinically relevant fusion genes in leukemia including a novel MEF2D/CSF1R fusion responsive to imatinib. in *Leukemia*, Vol. 28 977-979 (England, 2014).
63. Maher, C.A., *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97-101 (2009).
64. Kivioja, J.L., *et al.* Dasatinib and navitoclax act synergistically to target NUP98-NSD1(+)/FLT3-ITD(+) acute myeloid leukemia. *Leukemia* (2018).
65. Ding, L., Rath, E. & Bai, Y. Comparison of Alternative Splicing Junction Detection Tools Using RNA-Seq Data. *Curr Genomics* **18**, 268-277 (2017).
66. Zhao, S. Alternative splicing, RNA-seq and drug discovery. *Drug Discov Today* (2019).
67. Deelen, P., *et al.* Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med* **7**, 30 (2015).
68. Lopez-Maestre, H., *et al.* SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Res* **44**, e148 (2016).
69. Wilkerson, M.D., *et al.* Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res* **42**, e107 (2014).
70. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).
71. Li, S., *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* **32**, 915-925 (2014).
72. Bentley, D.R., *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
73. Lindberg, J. & Lundeberg, J. The plasticity of the mammalian transcriptome. *Genomics* **95**, 1-6 (2010).
74. Benes, V., Blake, J. & Doyle, K. Ribo-Zero Gold Kit: improved RNA-seq results after removal of cytoplasmic and mitochondrial ribosomal RNA. *Nature Methods* **8**(2011).
75. O'Neil, D., Glowatz, H. & Schlumpberger, M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. *Current protocols in molecular biology* **103**, 4.19. 11-14.19. 18 (2013).
76. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M. & Weissman, J.S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* **7**, 1534-1550 (2012).
77. Yi, H., *et al.* Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Res* **39**, e140 (2011).
78. Armour, C.D., *et al.* Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods* **6**, 647-649 (2009).
79. Cieslik, M. & Chinnaiyan, A.M. Cancer transcriptome profiling at the juncture of clinical translation. *Nat Rev Genet* **19**, 93-109 (2018).
80. Eikrem, O., *et al.* Transcriptome Sequencing (RNAseq) Enables Utilization of Formalin-Fixed, Paraffin-Embedded Biopsies with Clear Cell Renal Cell

Carcinoma for Exploration of Disease Biology and Biomarker Development. *PLoS One* **11**, e0149743 (2016).

81. Alberti, A.*, et al.* Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* **15**, 912 (2014).
82. Cui, P.*, et al.* A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* **96**, 259-265 (2010).
83. Guo, Y.*, et al.* RNAseq by Total RNA Library Identifies Additional RNAs Compared to Poly(A) RNA Library. *Biomed Res Int* **2015**, 862130 (2015).
84. Kissopoulou, A., Jonasson, J., Lindahl, T.L. & Osman, A. Next generation sequencing analysis of human platelet PolyA+ mRNAs and rRNA-depleted total RNA. *PLoS One* **8**, e81809 (2013).
85. Sultan, M.*, et al.* Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* **15**, 675 (2014).
86. Sun, Z.*, et al.* Impact of library preparation on downstream analysis and interpretation of RNA-Seq data: comparison between Illumina PolyA and NuGEN Ovation protocol. *PLoS One* **8**, e71745 (2013).
87. Zhao, W.*, et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**, 419 (2014).
88. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
89. Guo, Y.*, et al.* Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **109**, 83-90 (2017).
90. Ng, P.C. & Kirkness, E.F. Whole genome sequencing. *Methods Mol Biol* **628**, 215-226 (2010).
91. Weinstein, J.N.*, et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120 (2013).
92. Niguidula, N.*, et al.* Clinical whole-exome sequencing results impact medical management. *Mol Genet Genomic Med* **6**, 1068-1078 (2018).
93. Matias, M.*, et al.* Comparison of medical management and genetic counseling options pre- and post-whole exome sequencing for patients with positive and negative results. *J Genet Couns* **28**, 182-193 (2019).
94. Broach, J.R. & Thorner, J. High-throughput screening for drug discovery. *Nature* **384**, 14-16 (1996).
95. Holbeck, S.L., Collins, J.M. & Doroshow, J.H. Analysis of Food and Drug Administration-approved anticancer agents in the NCI60 panel of human tumor cell lines. *Mol Cancer Ther* **9**, 1451-1460 (2010).
96. Barretina, J.*, et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).
97. Garnett, M.J.*, et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575 (2012).
98. Pemovska, T.*, et al.* Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer discovery* **3**, 1416-1429 (2013).
99. Dietrich, S.*, et al.* Drug-perturbation-based stratification of blood cancer. *J Clin Invest* **128**, 427-445 (2018).
100. Tzelepis, K.*, et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep* **17**, 1193-1205 (2016).

101. Hernandez, P., *et al.* Drug Discovery Testing Compounds in Patient Samples by Automated Flow Cytometry. *SLAS Technol* **22**, 325-337 (2017).
102. Pan, R., *et al.* Selective BCL-2 inhibition by ABT-199 causes on-target cell death in acute myeloid leukemia. *Cancer discovery* **4**, 362-375 (2014).
103. DeLuca, D.S., *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530-1532 (2012).
104. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184-2185 (2012).
105. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
106. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011).
107. Kim, D., *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
108. Liao, Y., Smyth, G.K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**, e108 (2013).
109. Dobin, A., *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
110. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).
111. 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).
112. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
113. Schmid, M.W. & Grossniklaus, U. Rcount: simple and flexible RNA-Seq read counting. *Bioinformatics* **31**, 436-437 (2015).
114. 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 (2010).
115. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
116. Benelli, M., *et al.* Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics* **28**, 3232-3239 (2012).
117. Jia, W., *et al.* SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol* **14**, R12 (2013).
118. Nicorici, D., *et al.* FusionCatcher-a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *BioRxiv*, 011650 (2014).
119. Libbrecht, M.W. & Noble, W.S. Machine learning applications in genetics and genomics. *Nat Rev Genet* **16**, 321-332 (2015).
120. McPadden, J., *et al.* A scalable data science platform for healthcare and precision medicine research. *arXiv preprint arXiv:1808.04849* (2018).
121. Papaemmanuil, E., *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209-2221 (2016).
122. Campbell, B.B., *et al.* Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042-1056. e1010 (2017).
123. Vigilante, K., Escaravage, S. & McConnell, M. Big Data and the Intelligence Community - Lessons for Health Care. *N Engl J Med* **380**, 1888-1890 (2019).

124. Ali, M. & Aittokallio, T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical Reviews* **11**, 31-39 (2019).
125. Obermeyer, Z. & Emanuel, E.J. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* **375**, 1216-1219 (2016).
126. Azuaje, F. Artificial intelligence for precision oncology: beyond patient stratification. *NPJ Precis Oncol* **3**, 6 (2019).
127. Chang, J.C*., et al.* Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet (London, England)* **362**, 362-369 (2003).
128. Yang, W*., et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* **41**, D955-961 (2013).
129. Rees, M.G*., et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* **12**, 109-116 (2016).
130. Lamb, J*., et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-1935 (2006).
131. Menden, M.P*., et al.* Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS One* **8**, e61318 (2013).
132. Basu, A., Mitra, R., Liu, H., Schreiber, S.L. & Clemons, P.A. RWEN: response-weighted elastic net for prediction of chemosensitivity of cancer cell lines. *Bioinformatics* **34**, 3332-3339 (2018).
133. Liu, X*., et al.* A systematic study on drug-response associated genes using baseline gene expressions of the Cancer Cell Line Encyclopedia. *Scientific reports* **6**, 22811 (2016).
134. Masica, D.L. & Karchin, R. Collections of simultaneously altered genes as biomarkers of cancer cell drug response. *Cancer Res* **73**, 1699-1708 (2013).
135. Azuaje, F. Computational models for predicting drug responses in cancer research. *Brief Bioinform* **18**, 820-829 (2017).
136. Emad, A., Cairns, J., Kalari, K.R., Wang, L. & Sinha, S. Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance. *Genome Biol* **18**, 153 (2017).
137. Lee, S.I*., et al.* A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun* **9**, 42 (2018).
138. Jang, I.S., Neto, E.C., Guinney, J., Friend, S.H. & Margolin, A.A. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. in *Biocomputing 2014* 63-74 (World Scientific, 2014).
139. Ding, M.Q., Chen, L., Cooper, G.F., Young, J.D. & Lu, X. Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics. *Mol Cancer Res* **16**, 269-278 (2018).
140. Dohner, H. & Gaidzik, V.I. Impact of genetic features on treatment decisions in AML. *Hematology Am Soc Hematol Educ Program* **2011**, 36-42 (2011).
141. Estey, E. Are immunoconjugates approaching "standard of care" in AML? *Best Pract Res Clin Haematol* **26**, 261-268 (2013).
142. Burnett, A., Wetzler, M. & Lowenberg, B. Therapeutic advances in acute myeloid leukemia. *J Clin Oncol* **29**, 487-494 (2011).
143. Chen, S.J., Shen, Y. & Chen, Z. A panoramic view of acute myeloid leukemia. in *Nat Genet*, Vol. 45 586-587 (United States, 2013).

144. Gerstung, M., *et al.* Precision oncology for acute myeloid leukemia using a knowledge bank approach. *Nat Genet* **49**, 332-340 (2017).

145. Patel, J.P., *et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N Engl J Med* **366**, 1079-1089 (2012).

146. Silva, P., *et al.* Acute myeloid leukemia in the elderly is characterized by a distinct genetic and epigenetic landscape. in *Leukemia*, Vol. 31 1640-1644 (England, 2017).

147. Ley, T.J., *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059-2074 (2013).

148. Assi, S.A., *et al.* Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nat Genet* **51**, 151-162 (2019).

149. Huang, L., *et al.* FLT3-ITD-associated gene-expression signatures in NPM1-mutated cytogenetically normal acute myeloid leukemia. *Int J Hematol* **96**, 234-240 (2012).

150. Dohner, H., *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424-447 (2017).

151. Ding, L., *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506-510 (2012).

152. Klco, J.M., *et al.* Functional heterogeneity of genetically defined subclones in acute myeloid leukemia. *Cancer Cell* **25**, 379-392 (2014).

153. Quek, L., *et al.* Clonal heterogeneity of acute myeloid leukemia treated with the IDH2 inhibitor enasidenib. *Nature medicine* **24**, 1167-1177 (2018).

154. Lavallee, V.P., *et al.* The transcriptomic landscape and directed chemical interrogation of MLL-rearranged acute myeloid leukemias. *Nat Genet* **47**, 1030-1037 (2015).

155. Lavallee, V.P., *et al.* Chemo-genomic interrogation of CEBPA mutated AML reveals recurrent CSF3R mutations and subgroup sensitivity to JAK inhibitors. *Blood* **127**, 3054-3061 (2016).

156. DiNardo, C.D., *et al.* Venetoclax combined with decitabine or azacitidine in treatment-naive, elderly patients with acute myeloid leukemia. *Blood* **133**, 7-17 (2019).

157. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).

158. Kerr, J.F., Wyllie, A.H. & Currie, A.R. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br J Cancer* **26**, 239-257 (1972).

159. Elmore, S. Apoptosis: a review of programmed cell death. *Toxicol Pathol* **35**, 495-516 (2007).

160. Green, D.R. Cell death and the immune system: getting to how and why. *Immunol Rev* **277**, 4-8 (2017).

161. Zou, H., Li, Y., Liu, X. & Wang, X. An APAF-1.cytochrome c multimeric complex is a functional apoptosome that activates procaspase-9. *J Biol Chem* **274**, 11549-11556 (1999).

162. Boise, L.H., *et al.* bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell* **74**, 597-608 (1993).

163. Kozopas, K.M., Yang, T., Buchan, H.L., Zhou, P. & Craig, R.W. MCL1, a gene expressed in programmed myeloid cell differentiation, has sequence similarity to BCL2. *Proc Natl Acad Sci U S A* **90**, 3516-3520 (1993).

164. Lin, E.Y., Orlofsky, A., Wang, H.G., Reed, J.C. & Prystowsky, M.B. A1, a Bcl-2 family member, prolongs cell survival and permits myeloid differentiation. *Blood* **87**, 983-992 (1996).
165. O'Connor, L.*, et al.* Bim: a novel member of the Bcl-2 family that promotes apoptosis. *Embo j* **17**, 384-395 (1998).
166. Carpio, M.A.*, et al.* BCL-2 family member BOK promotes apoptosis in response to endoplasmic reticulum stress. *Proc Natl Acad Sci U S A* **112**, 7201-7206 (2015).
167. Gross, A., Jockel, J., Wei, M.C. & Korsmeyer, S.J. Enforced dimerization of BAX results in its translocation, mitochondrial dysfunction and apoptosis. *Embo j* **17**, 3878-3885 (1998).
168. Inohara, N., Ding, L., Chen, S. & Nunez, G. harakiri, a novel regulator of cell death, encodes a protein that activates apoptosis and interacts selectively with survival-promoting proteins Bcl-2 and Bcl-X(L). *Embo j* **16**, 1686-1694 (1997).
169. Nakano, K. & Vousden, K.H. PUMA, a novel proapoptotic gene, is induced by p53. *Mol Cell* **7**, 683-694 (2001).
170. Oda, E.*, et al.* Noxa, a BH3-only member of the Bcl-2 family and candidate mediator of p53-induced apoptosis. *Science* **288**, 1053-1058 (2000).
171. Wang, K., Yin, X.M., Chao, D.T., Milliman, C.L. & Korsmeyer, S.J. BID: a novel BH3 domain-only death agonist. *Genes Dev* **10**, 2859-2869 (1996).
172. Yang, E.*, et al.* Bad, a heterodimeric partner for Bcl-XL and Bcl-2, displaces Bax and promotes cell death. *Cell* **80**, 285-291 (1995).
173. Yu, J., Zhang, L., Hwang, P.M., Kinzler, K.W. & Vogelstein, B. PUMA induces the rapid apoptosis of colorectal cancer cells. *Mol Cell* **7**, 673-682 (2001).
174. Wei, M.C.*, et al.* tBID, a membrane-targeted death ligand, oligomerizes BAK to release cytochrome c. *Genes Dev* **14**, 2060-2071 (2000).
175. Leverson, J.D.*, et al.* Found in Translation: How Preclinical Research Is Guiding the Clinical Development of the BCL2-Selective Inhibitor Venetoclax. *Cancer discovery* **7**, 1376-1393 (2017).
176. Tse, C.*, et al.* ABT-263: a potent and orally bioavailable Bcl-2 family inhibitor. *Cancer Res* **68**, 3421-3428 (2008).
177. Schoenwaelder, S.M.*, et al.* Bcl-xL-inhibitory BH3 mimetics can induce a transient thrombocytopathy that undermines the hemostatic function of platelets. *Blood* **118**, 1663-1674 (2011).
178. Souers, A.J.*, et al.* ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nature medicine* **19**, 202-208 (2013).
179. Hotchkiss, R.S., Strasser, A., McDunn, J.E. & Swanson, P.E. Cell death. *N Engl J Med* **361**, 1570-1583 (2009).
180. Konopleva, M.*, et al.* Efficacy and Biological Correlates of Response in a Phase II Study of Venetoclax Monotherapy in Patients with Acute Myelogenous Leukemia. *Cancer discovery* **6**, 1106-1117 (2016).
181. Lochmann, T.L.*, et al.* Venetoclax Is Effective in Small-Cell Lung Cancers with High BCL-2 Expression. *Clin Cancer Res* **24**, 360-369 (2018).
182. Chyla, B.*, et al.* Genetic biomarkers of sensitivity and resistance to venetoclax monotherapy in patients with relapsed acute myeloid leukemia. *American Journal of Hematology* **93**, E202-E205 (2018).
183. Sulonen, A.M.*, et al.* Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* **12**, R94 (2011).
184. Yadav, B.*, et al.* Quantitative scoring of differential drug sensitivity for individually optimized anticancer therapies. *Scientific reports* **4**, 5193 (2014).

185. Kumar, A.*, et al.* The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia. *BMC Genomics* **18**, 629 (2017).

186. Robinson, M.D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, R25 (2010).

187. Ritchie, M.E.*, et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

188. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).

189. Chen, E.Y.*, et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

190. Kuleshov, M.V.*, et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97 (2016).

191. Warde-Farley, D.*, et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* **38**, W214-220 (2010).

192. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* **10**, 48 (2009).

193. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific reports* **8**, 4781 (2018).

194. Zaghlool, A.*, et al.* Efficient cellular fractionation improves RNA sequencing analysis of mature and nascent transcripts from human tissues. *BMC Biotechnol* **13**, 99 (2013).

195. Ameur, A.*, et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature Structural &Amp; Molecular Biology* **18**, 1435 (2011).

196. Wetterbom, A., Ameur, A., Feuk, L., Gyllensten, U. & Cavelier, L. Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing. *Genome Biol* **11**, R78 (2010).

197. Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res* **21**, 2213-2223 (2011).

198. Gaidatzis, D., Burger, L., Florescu, M. & Stadler, M.B. Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. *Nat Biotechnol* **33**, 722-729 (2015).

199. Chen, C.Y.*, et al.* Bcl3 Bridges LIF-STAT3 to Oct4 Signaling in the Maintenance of Naive Pluripotency. *Stem Cells* **33**, 3468-3480 (2015).

200. Gilan, O.*, et al.* Functional interdependence of BRD4 and DOT1L in MLL leukemia. *Nat Struct Mol Biol* **23**, 673-681 (2016).

201. Rouce, R.H.*, et al.* The TGF-beta/SMAD pathway is an important mechanism for NK cell immune evasion in childhood B-acute lymphoblastic leukemia. *Leukemia* **30**, 800-811 (2016).

202. Adiconis, X.*, et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods* **10**, 623-629 (2013).

203. Cieslik, M.*, et al.* The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res* **25**, 1372-1381 (2015).

204. Mallo, M., Wellik, D.M. & Deschamps, J. Hox genes and regional patterning of the vertebrate body plan. *Dev Biol* **344**, 7-15 (2010).

205. Vitiello, D., Kodaman, P.H. & Taylor, H.S. HOX genes in implantation. *Semin Reprod Med* **25**, 431-436 (2007).

206. Wellik, D.M. Hox patterning of the vertebrate axial skeleton. *Dev Dyn* **236**, 2454-2463 (2007).

207. Skvarova Kramarzova, K., *et al.* Homeobox gene expression in acute myeloid leukemia is linked to typical underlying molecular aberrations. *J Hematol Oncol* **7**, 94 (2014).

208. Hess, J.L. MLL: a histone methyltransferase disrupted in leukemia. *Trends Mol Med* **10**, 500-507 (2004).

209. Lagadinou, E.D., *et al.* BCL-2 inhibition targets oxidative phosphorylation and selectively eradicates quiescent human leukemia stem cells. *Cell Stem Cell* **12**, 329-341 (2013).

210. Spencer, D.H., *et al.* Epigenomic analysis of the HOX gene loci reveals mechanisms that may control canonical expression patterns in AML and normal hematopoietic cells.

211. Bresnick, A.R., Weber, D.J. & Zimmer, D.B. S100 proteins in cancer. *Nat Rev Cancer* **15**, 96-109 (2015).

212. Donato, R., *et al.* Functions of S100 proteins. *Curr Mol Med* **13**, 24-57 (2013).

213. Moore, B.W. A soluble protein characteristic of the nervous system. *Biochem Biophys Res Commun* **19**, 739-744 (1965).

214. Nicolas, E., *et al.* Expression of S100A8 in leukemic cells predicts poor survival in de novo AML patients. *Leukemia* **25**, 57-65 (2011).

215. Ni Chonghaile, T., *et al.* Pretreatment mitochondrial priming correlates with clinical response to cytotoxic chemotherapy. *Science* **334**, 1129-1133 (2011).

216. Bui, M.H., *et al.* Preclinical Characterization of BET Family Bromodomain Inhibitor ABBV-075 Suggests Combination Therapeutic Strategies. *Cancer Res* **77**, 2976-2989 (2017).

217. Fiskus, W., *et al.* Superior efficacy of cotreatment with BET protein inhibitor and BCL2 or MCL1 inhibitor against AML blast progenitor cells. *Blood Cancer J* **9**, 4 (2019).

218. Li, G.Q., *et al.* Suppression of BRD4 inhibits human hepatocellular carcinoma by repressing MYC and enhancing BIM expression. *Oncotarget* **7**, 2462-2474 (2016).

219. Patel, A.J., *et al.* BET bromodomain inhibition triggers apoptosis of NF1-associated malignant peripheral nerve sheath tumors through Bim induction. *Cell Rep* **6**, 81-92 (2014).

220. Peirs, S., *et al.* Targeting BET proteins improves the therapeutic efficacy of BCL-2 inhibition in T-cell acute lymphoblastic leukemia. *Leukemia* **31**, 2037-2047 (2017).

221. Buege, M.J., DiPippo, A.J. & DiNardo, C.D. Evolving Treatment Strategies for Elderly Leukemia Patients with IDH Mutations. *Cancers (Basel)* **10**(2018).

222. Chan, S.M., *et al.* Isocitrate dehydrogenase 1 and 2 mutations induce BCL-2 dependence in acute myeloid leukemia. *Nature medicine* **21**, 178-184 (2015).

223. Dang, L., *et al.* Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739-744 (2009).

224. Montalban-Bravo, G. & DiNardo, C.D. The role of IDH mutations in acute myeloid leukemia. *Future Oncol* **14**, 979-993 (2018).

225. Ward, P.S., *et al.* The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. *Cancer Cell* **17**, 225-234 (2010).