Statistical Methods and Machine Learning Techniques for Analyzing High-Dimensional Data in Cancer Biology: Modeling T-Cell Receptor Repertoire and Methylation Signatures in Acute Myeloid Leukemia

By

Burcu Gurun

B.Sc., Middle East Technical University, 2002

M.A., Columbia University, 2006

A DISSERTATION

Presented to the Department of Biomedical Engineering of the Oregon Health & Science University School of Medicine
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy

June 2023

Annem Güler, babam Ali İhsan ve oğlum Altan'a...

School of Medicine

Oregon Health & Science University

Certificate of Approval

_____

This is to certify that the PhD dissertation of

Burcu Gurun

has been approved

_____

Dr. Paul Spellman (Mentor/Advisor)

_____

Dr. Brian Druker (Co-Mentor/Advisor)

_____

Dr. Jeremy Goecks (Chair)

_____

Dr. Jeffrey Tyner (Member)

_____

Dr. Sadik Esener (Member)

1.1. Table of Contents

Table of Contents vii

Acknowledgements xi

List of Abbreviations xv

List of Tables xvii

List of Figures xix

List of Materials xx

Dissertation Abstract xxi

Chapter 1: Introduction 23

1.1 Utilizing Statistical Methods and Machine Learning to Advance Cancer Biology Research 23

1.2 Supervised and Unsupervised Machine Learning Methods 23

1.3 Statistical Normalization 24

1.4 Using Supervised and Unsupervised Machine Learning Models to Build AML Epigenetic Signatures 24

1.5 Using Statistical Normalization to Quantify TCR Repertoires 27

Chapter 2: Computational modeling of methylation impact of AML drivers reveals new pathways and refines AML risk-stratification 30

Abstract 31

2.1 Introduction 32

2.2 Methods 35

2.2.1 Dataset 35

2.2.2 Approach for supervised models 35

2.2.3 Approach for unsupervised models 36

2.2.4 Data sharing statement 37

2.3 Results 37

2.3.1 *DNMT3A* and *TET2* have robust downstream methylation signatures 37

2.3.2 DMRs selected by *DNMT3A* and *TET2* signatures highlight key downstream pathways 38
2.3.3 Factorization of methylation changes into Topics link genomic events into distinct methylation programs 41
2.4 Discussion 47
2.5 Acknowledgement 49
2.6 Authorship Contributions 50
Chapter 3: An open protocol for modeling T Cell Clonotype repertoires using TCRβ CDR3 sequences ..... Error! Bookmark not defined.
Abstract 52
3.1 Background 52
3.2 Results 54
3.2.1 Amplification bias due to multiplex PCR is reproducible 54
3.2.2 A negative binomial model fits the data 57
3.2.3 Scaling factors are relatively stable across experiments 58
3.2.4 Normalization considerably reduces amplification bias 59
3.2.5 Amplification bias reduction benefits from the dependence of primer pairs 61
3.2.6 A negative binomial model supports downstream analyses of T cell repertoire dynamics 63
3.3 Discussion 64
3.4 Conclusions 66
3.5 Material and Methods 67
3.5.1 Mouse handling 67
3.5.2 Multiplex primers and design of synthetic TCR templates 68
3.5.3 Amplification and deep sequencing of TCRβ genomic locus 69
3.5.4 TCR data analysis Pipeline 70
3.5.5 Data Sets 70
3.5.6 Batch mean scaling factors 72
3.5.7 Negative Binomial means 72

3.5.8 Normalization 73
3.6 Declarations 74
3.6.1 Ethics approval and consent to participate 74
3.6.2 Availability of data and materials 74
3.6.3 Funding 75
3.6.4 Authors' contributions 75
3.6.5 Acknowledgements 76
Chapter 4: Discussion 77
4.1 Contribution of Statistical Modeling of AML Signature Study 77
4.2 Contribution of Normalization of TCR Sequencing Study 78
4.3 Conclusions 79
Chapter 5: Future Applications 82
5.1 Potential Future Applications of Statistical Modeling of AML Signature Study..... 82
5.1.1 Using methylation signatures for subtyping for drug response studies 82
5.1.2 Using methylation signatures as a biomarker 85
5.2 Potential Future Applications of Normalization of TCR Sequencing Study..... 87
Appendix A: Chapter 2 Supplementary Information 86
Appendix B: Chapter 3 Supplementary Information 91
Supp. Material 1 99
Supp. Material 2 101
References 103

1.2. Acknowledgements

I would not have been able to achieve any of this work without the support of my husband, Emek Demir. He has been a rock for me, consistently encouraging me to pursue my aspirations, despite his demanding work schedule. I am grateful for his understanding as a mother following an MD-PhD program. I thank him for accompanying me on this journey and for his partnership in raising our child, in a foreign country, thousands of miles away from our home and the rest of the family. Without him by my side, I would not have come this far. Although my path towards being a physician-scientist is far from over, I feel empowered and excited to tackle the challenges ahead and welcome new opportunities with his steadfast support.

I would like to express my gratitude to my wonderful and brilliant child, Altan Demir, whose inspiration and motivation have pushed me to be the best version of myself and strive to make the world they will inherit a better place. Altan has taught me patience, tolerance, and the true value of life. I cannot thank him enough for his understanding when I had to prioritize my work and studies over spending time with him. I aspire to be a positive role model for him and his loved ones, encouraging him to pursue his dreams through hard work.

I am incredibly grateful to my parents, Guler Gurun and Ali Ihsan Gurun, for their unconditional love, constant support, and willingness to drop everything and fly across the world to help me whenever we need it. Without their guidance and encouragement, I never would have even considered the possibility of immigrating to the USA, attending graduate schools here, and pursuing a career as a physician-scientist. Their own dedication and perseverance in their endeavors have provided me with excellent role

models, and I am forever indebted to them for instilling in me a strong work ethic and commitment to my goals.

I am grateful to my brother, Utku Gurun, for being an amazing companion throughout my childhood and teenage years. He played a significant role in nurturing my love for music and medicine and never failed to lift my spirits whenever I felt down. I feel blessed to have him as my sibling and could not have asked for a better one. I also express my deepest gratitude to my eldest cousin, Siren Sezer, for being an incredible role model as a physician, who also has become my mother's nephrologist and takes care of her while I am writing my thesis. I also extend my thanks to Hanim Akkemik, my grandmother, who was one of the most visionary people of her time. Despite facing difficult circumstances, she made education a top priority for her children, and her efforts have had a profound impact on the character and achievements of her descendants.

I am deeply grateful to Irem and Volkan Yargici, Ayca and Emre Ozkumur, Gokce and Cagri Toruner, Zeynep and Bulent Omay and Cigdem Ataseven who have become our extended family in the US. I extend my gratitude to Mini Zhang, Adem and Gulsu Yildirim, Michelle Young, Cigdem and Ozcan Ertem, Derya Esener, JoAnn Takayabashi, Hamra and Hakan Bakircioglu, Emily and Nick Buccola, Duygu and Ali Kayaarslan, who made Portland feel like home. Their friendship has enriched our lives beyond measure.

I am grateful to both current and former members of Dr. Paul Spellman's lab as well as the Cancer Early Detection Advanced Research Center (CEDAR) at the Knight Cancer Institute. First and foremost, I want to thank Dr. Ece Eksi, for being an inspiring mentor and a lovely friend. Her unswerving support has been invaluable to me during the most challenging times, and her insights and feedback on my work have been immensely helpful in making it more scientifically sound. I am thankful for her presence in my life. I would also like to extend my thanks to Dr. Myron Peto, Dr. Michael Heskett,

Dr. Chris Boniface, Dr. Cigdem Ak, Dr. Aysegul Ors, Kami Chiotti and Carol Halsey for always taking the time to listen to my questions, offering insightful perspectives, and providing valuable feedback. Moreover, I am grateful to the CEDAR scientific board and reviewing committee for their useful remarks on my research, and to Dr. Sadik Esener for his ongoing support, guidance, and mentorship. His mentorship has been invaluable in shaping me both as a scientist and an individual. I would like to extend my gratitude to Drs. Funda Durupinar and Ozgun Babur, former members of the Computational Biology Department, for being not only wonderful friends but also lifelong companions and partners in all our adventures. They have become like family members and always been there for us when we needed them. Additionally, I want to give special thanks to Dr. Gulsu Sener Yildirim, a CEDAR member, for providing me with a sense of belonging in Portland through her kindness and support. Her friendship has been a great source of comfort and strength.

I would also like to express my gratitude to my friends and colleagues in the MD-PhD program for engaging in enriching discussions that have helped me grow professionally. I am grateful to our Director, Dr. Jacoby, for providing me with his continuous support, trust, and guidance throughout my journey. In particular, I would like to extend my thanks to Dr. Gavin Young, Dr. Kristen Stevens, Dr. Tony Zheng, Ashley Anderson, Will Yashar and Tetiana Korzun as well as Talia Ramos for their constant presence and support whenever I needed it.

I would like to thank Dr. Terry Speed for the guidance, support, and mentorship that played an important role in shaping my research. I also would like to extend my thanks to Dr. Lisa Coussens for her trust and support over the last seven years. Without them, I wouldn't have completed the third chapter of my thesis.

I am grateful to Drs. Jeff Tyner and Jeremy Goecks for their invaluable feedback, constructive criticism, and guidance which have significantly contributed to the improvement of my research. Their flexibility and trust in my work have been a great source of empowerment for me.

I would like to express my deep appreciation to my co-mentor, Dr. Brian Druker, for his unwavering confidence and support. Without him, I may not have pursued an MD-PhD program and embarked on this fulfilling path as a physician-scientist. His belief in my potential, encouragement to leverage my quantitative background, and continuous support have been instrumental in opening doors to various opportunities and shaping my journey. I cannot thank him enough for his ongoing guidance and mentorship.

Finally, I would like to extend my heartfelt gratitude to my mentor, Dr. Paul Spellman, who never ceased to challenge me to become a better scientist and kept me rooted in the pragmatic aspects of technical and scientific research. Dr. Spellman has been an exceptional role model and his steadfast support has contributed significantly to my personal and professional development. I am immensely thankful for his mentorship, and my appreciation knows no bounds.

1.3. List of Abbreviations

AML	acute myeloid leukemia
TCR	T cell receptor
OTSP	open TCR sequencing protocol
DMRs	differentially methylated regions
CDR3	complementarity determining region 3
AUROC	area under the receiver operating characteristics
ELN	European LeukemiaNet
LDA	latent dirichlet allocation
WT	wild-type
HOX	homeobox
NGS	next generation sequencing
ST	synthetic TCR template
gDNA	genomic DNA
MMTV	mouse mammary tumor virus
PyMT	polyomavirus middle T
PEAR	paired-end read merger
Indels	insertions and deletions
NB	negative binomial
SF	scaling factor
MLE	maximum likelihood estimation

1.4. List of Tables

Table 2-1 Gene annotations for the top ten most frequently selected features for the *DNMT3A* status prediction both for hypermethylated (green) and hypomethylated (yellow) sites 40

Table 2-2 Gene annotations for the top ten most frequently selected features for the *TET2* status prediction both for hypermethylated (green) and hypomethylated (yellow) sites. 41

Table 2-3. Annotation of the top 10 most enriched loci for all 15 topics..... 46

Supp. Table 1. Primer Sequences..... 96

1.5. List of Figures

Figure 1-1 Myeloid malignancies comprise a continuum of hematopoietic disorders: 26

Figure 1-2 V(D)J recombination determines T-cell receptor specificity 29

Figure 2-1 Overall Approach: Inferring methylation signatures that drives AML from methylation profiles. 34

Figure 2-2.A-B. Performance of classifiers measured with AUROC (Area under the ROC curve) 38

Figure 2-3. Topic Distribution for 220 post-diagnosis AML patients 43

Figure 2-4.Enrichment heatmap of topics vs. epigenetic factors 44

Table 2-3. Annotation of the top 10 most enriched loci for all 15 topics 46

Figure 3-1 Overview of the TCR sequencing analysis pipeline 55

Figure 3-2 ST proportion distributions 56

Figure 3-3. A negative binomial model fits the data 58

Figure 3-4.A-B. Stability of scaling factors 59

Figure 3-5. NB normalization reduces amplification bias 60

Figure 3-6. Amplification bias of spleen genomic DNA of P14 and OT-1 TCR transgenic mice 61

Figure 3-7. Cluster analysis revealing dependence between forward and reverse primers 63

Figure 3-8. Concordance analysis of T cell repertoire metrics 64

Figure 5-1. Epigenetic machinery and therapeutic agents 83

Supp. Figure 1. Top 10 most enriched loci for 15 topics 86

Supp. Figure 2. Mutations and events associated with ELN intermediate to adverse risk category 88

Supp. Figure 3. Cytogenetic events associated with ELN favorable risk category 89
Supp. Figure 4. Methylation signature of *FLT3* 90
Supp. Figure 5. ST Count Distribution in presence of DNA 91
Supp. Figure 6. Normalization reduces spread in presence of DNA 92
Supp. Figure 7. Concordance analysis of T cell repertoire metrics 93
Supp. Figure 8. Competition between gDNA and ST during TCR sequencing 94
Supp. Figure 9. Reproducibility analysis: Drop-outs are frequent even for the top clones 95
Supp. Figure 10. TCR sequencing pipeline schema 97
Supp. Figure 11. Monoclonal amplification check 98
List of Materials
Supp. Material 1 99
Supp. Material 2 101

Dissertation Abstract

This thesis is composed of two studies that utilizes different statistical methods to answer cancer research questions. The first study uses supervised and unsupervised statistical modeling on methylation array data from the BeatAML cohort to infer methylation signatures of AML. The findings demonstrate that computational modeling of methylation impact of AML drivers can reveal novel molecular pathways while also validating previously known associations and significantly enhancing AML risk-stratification. We extracted accurate and stable signatures of methylation impact of *DNMT3A* and *TET2* mutations, which are the most frequently mutated epigenetic regulators in AML, and revealed methylation pathways that are important in leukemogenesis. We also employed topic modeling to deconvolute methylation signatures of multiple drivers. We observed that this method can extract the signatures of relatively less frequent events that would have been obfuscated by high frequency events in a conventional hard label clustering approach and it has strong potential to improve methylation-based subtyping. Collectively, these signatures broaden our understanding of the impact of epigenetic mutations on leukemogenesis and may inform subsequent detection and drug response studies for AML patients.

The second study presents a non-commercial and inexpensive protocol for measuring and monitoring adaptive dynamics in TCR clonotype repertoire using genomic DNA-based bulk sequencing. The results show that the concordance between bulk clonality metrics obtained from using the commercial kits and that developed herein is high. The study describes the Open TCR Sequencing Protocol (OTSP) that efficiently corrects for amplification bias post-sequencing, provides a transparent protocol enabling clonality metrics, and is reproducible across samples. This study will be highly relevant to the scientific community, given the extensive interest in measuring and monitoring adaptive dynamics in patient TCR repertoires

and its potential significant impact on response and resistance monitoring for patients receiving various forms of immunotherapy in the treatment of cancer or auto-immune diseases. Overall, these two studies demonstrate the importance of statistical methods, bioinformatics, and machine learning in contemporary cancer biology research. They aid in integrating and interpreting a wide range of high-dimensional data types and identifying the molecular characteristics of tumors, predicting patient outcomes, and developing personalized treatments. In the future, these techniques will continue to play a critical role in advancing cancer research and improving patient outcomes.

Introduction

1.1. Utilizing Statistical Methods and Machine Learning to Advance Cancer Biology Research

Statistical methods and machine learning have transformed the field of cancer biology by providing researchers with powerful tools to analyze and interpret complex biological data at an unprecedented scale. These techniques support the integration of different layers of data from genomics, epigenomics, transcriptomics, proteomics, and imaging at the single-cell resolution. Importantly, these methods allow for the identification of molecular signatures that characterize tumors, predict patient outcomes, and guide the development of personalized therapies to name a few.

1.2. Supervised and Unsupervised Machine Learning Methods

Statistical, computational and machine learning techniques are essential to process, analyze, interpret and extract meaningful insights from the vast amount of data generated from high-throughput experiments. When class labels are present, supervised machine learning approaches such as support vector machines, random forests, regression and deep neural networks provide robust and efficient solutions to model the data, identify important variables, make predictions and classify samples. Often, however, we need to identify patterns and clusters in the high throughput data without pre-labeled information or prior knowledge of the structure of the data. After performing dimensionality reduction and including the most

informative variables, unsupervised machine learning models such as nonnegative matrix factorization, topic modeling and deep learning models can be used to represent the data in a latent space. These models, through different strategies, optimize the latent mapping such that a faithful approximation of the original data can be reconstructed from it. The difference of the reconstruction from the original data is called reconstruction error and is a key metric for assessing model performance. Latent representations have been shown to be more manageable and provide the hidden structures in the data with a higher resolution than the traditional clustering methods.

1.3. Statistical Normalization

High-throughput techniques in biology often exhibit substantial technical variability and batch effects that can confound the biological signal of interest. This variability can stem from multiple sources, including differences in sample preparation, instrument performance, and experimental conditions. By applying normalization techniques, researchers can mitigate these biases and ensure that the observed differences between samples primarily reflect the underlying biology rather than technical artifacts. Consequently, normalization is a critical step in the analysis of high-throughput data, enabling more reliable interpretations of the data, and ultimately leading to more robust and reproducible findings in biological research.

1.4. Using Supervised and Unsupervised Machine Learning Models to Build AML Epigenetic Signatures

In the first part of my thesis, I use supervised and unsupervised machine learning models on methylation profiles of AML patients to classify and categorize differentially

methylated regions to shed insights into AML etiology, dynamics, mechanisms and enable risk-stratification.

Acute myeloid leukemia (AML) is a type of cancer that can start developing in healthy tissue years before the onset of symptoms. The path to AML can differ between individuals, with hematopoietic stem cells following different paths and gaining different features over time (**Figure 1-1**). Once the disease progresses to an advanced stage, it develops substantial clonal heterogeneity and becomes difficult to treat, with only a small percentage of patients surviving for more than five years after diagnosis. The disease can be difficult to detect early on, and the early phase is primarily driven by a small set of mutations that can lead to large numbers of differentially methylated regions (DMRs) in the genome. These DMRs can be used to gain a better understanding of the mechanisms and dynamics that are important in leukemogenesis, as well as to characterize aberrant methylation in cancers, sub-classify tumors, distinguish the tissue of tumor origin and gain insight into the mechanisms underlying AML and use this knowledge to develop new diagnostic and treatment strategies.

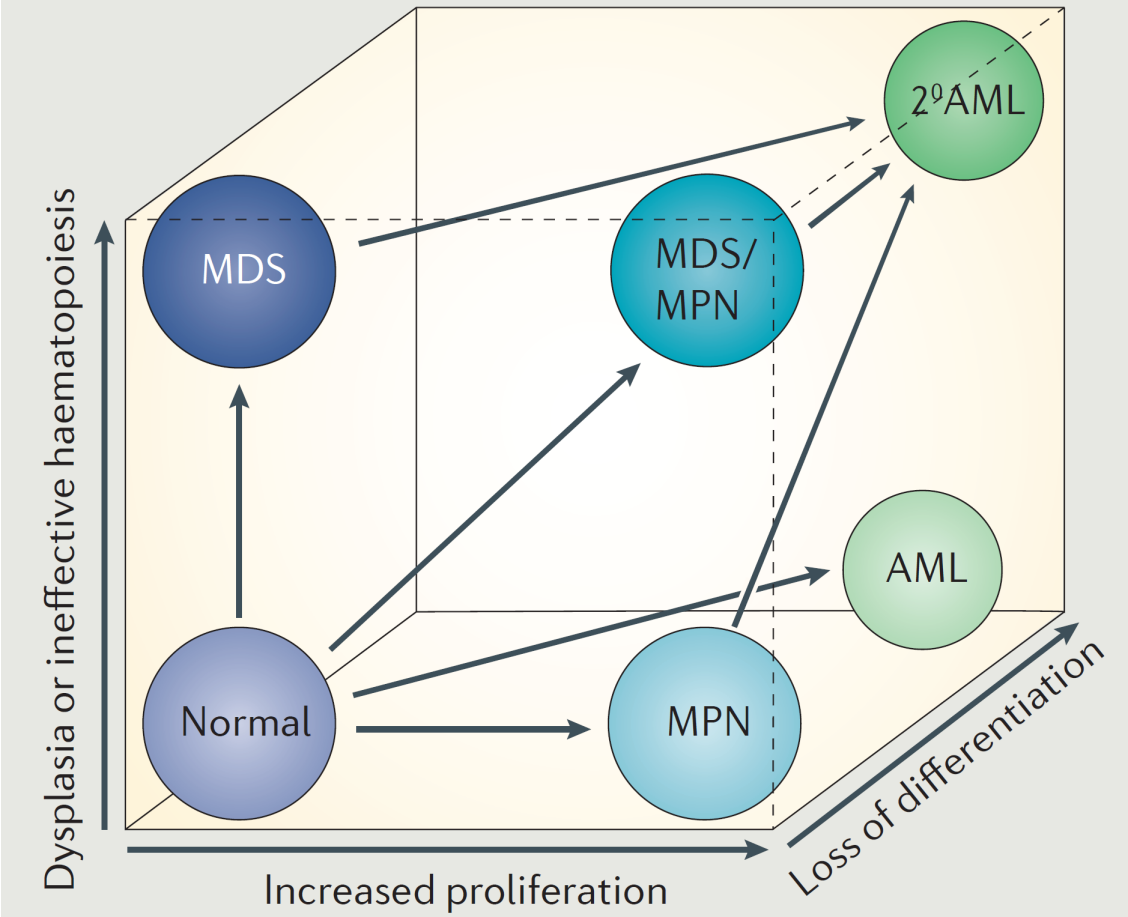


Figure 1-1 Myeloid malignancies comprise a continuum of hematopoietic disorders: Myeloid malignancies comprise a continuum of hematopoietic disorders that are characterized by increased proliferation, abnormal morphology (dysplasia) and impaired differentiation of myeloid lineage cells, or combinations thereof. Adapted with permission from Deininger et al., 2017 (1)

I grouped methylation changes into "epigenetic signatures" through two parallel statistical approaches: a supervised approach to find methylation signatures capturing the impact of driver mutations and an unsupervised Topic Modeling approach to factorize covarying epigenetic changes into few "topics".

Supervised machine learning models involve training a computer algorithm using a labeled dataset, where the inputs have corresponding outputs or target values. These models can then be used to predict the output for new, unlabeled data points. I use elastic net

regression models for inferring wild type vs mutant genotypes for *DNMT3A* and *TET2*, the most frequent epigenetic regulators, based on their downstream epigenetic impacts. This analysis reveals new methylation pathways of AML driver events important in leukemogenesis, uncovers new insights into methylation regulation in AML and validate previously known associations. This framework can be leveraged to build predictive models for AML-risk to inform detection and drug response studies for AML patients.

Unsupervised machine learning models are a type of artificial intelligence technique that can identify patterns and relationships in data without prior knowledge or guidance from humans. Unsupervised models like topic modeling are particularly beneficial for exploring the underlying patterns and structures in sparse and high-dimensional data, like data from methylation profiling. By analyzing the joint distribution of DNA methylation at multiple CpG sites across the genome, topic modeling reveals groups of CpG sites that tend to co-vary together, providing insights into potential regulatory mechanisms in AML. As such, our models uncover methylation signatures of infrequent mutations, identify topics that show high concordance with ELN-2017 criteria, reveals unique signatures of co-occurring and a group of adverse-risk mutations with convergent methylation impacts and provides a computational framework has potential to select methylation sites with high biomarker potential for liquid biopsy assays.

1.5. Using Statistical Normalization to Quantify TCR Repertoires

In the second part of my thesis, I implemented statistical methods to quantify and model T Cell Clonotype repertoires using TCRβ CDR3 sequences. The binding of wide range of antigens to receptors on the surface of T cells is a crucial factor in shaping the immune response in both healthy and diseased states. The T cell receptor (TCR) is a heterodimer

consisting of an alpha and a beta chain, encoded by the TCRα and TCRβ genes respectively. In order to recognize a vast array of antigens, the TCR genomic loci undergo somatic recombination of variable (V), diversity (D), and joining (J) gene segments, resulting in a diverse range of TCRs (**Figure 1-2**). The diversity of the TCRβ chain is particularly high in the complementarity determining region 3 (CDR3), located at the D segment of the recombined TCRβ gene – this region works as a natural barcode identifying the clone. Therefore, sequencing the recombined TCRβ gene or transcript TCR repertoire can help us understand the dynamics and diversity T-Cells.

Experimental pipeline is based on multiplex PCR followed by NGS. In the multiplex PCR, it is necessary to use multiple sets of primers to amplify the recombinant CDR3 region and each forward and reverse primers work at different efficiencies. This amplification bias needs to be statistically corrected to quantify the T-Cell Receptor (TCR) repertoire and abundance accurately and for comparisons between different samples or treatment conditions. Measuring and monitoring adaptive dynamics in patient TCR repertoires could have a significant impact on response and resistance monitoring for patients receiving various forms of immunotherapy in the treatment of cancer or auto-immune diseases.

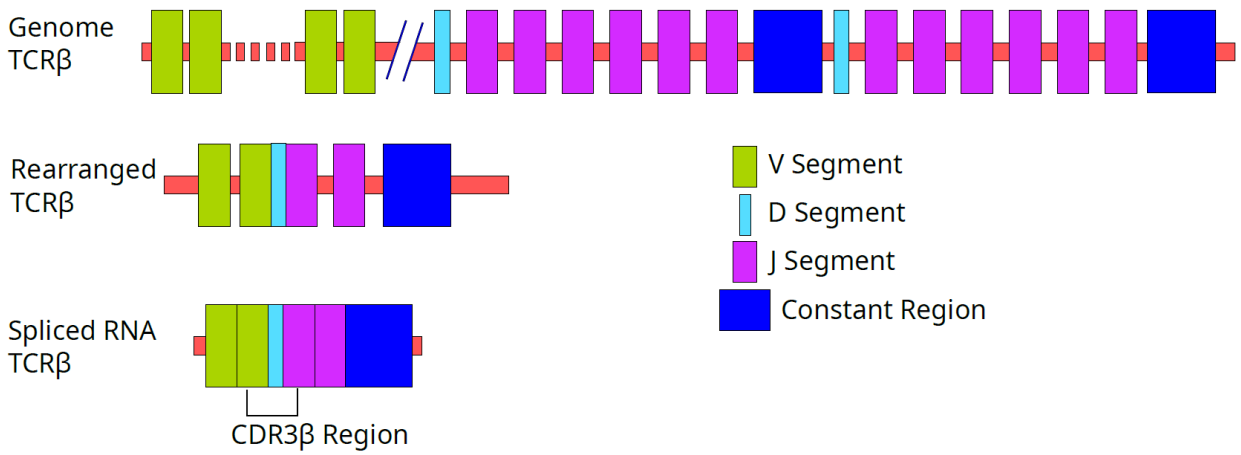


Figure 1-2 TCR specificity determined by V(D)J recombination: The T-cell receptor (TCR) specificity of αβ T cells is dictated by the distinctive events of V(D)J recombination that take place during the maturation of each T cell. In this process, V, D, and J gene segments are arbitrarily chosen and combined on the β chain, while the α-chain experiences a comparable rearrangement of the V–J gene segments. Within this progression, random nucleotide additions or deletions can happen at the junctions of these segments. The region known as the complementarity-determining region 3 (CDR3), which is encoded by sequences located in the V(D)J junction, exhibits the greatest diversity and ultimately determines the antigen specificity of each TCR. TCRβ: T-cell receptor beta; CDR3β: the gene sequence encoding the complementarity-determining region 3 of the TCR beta chain.

Computational modeling of methylation impact of AML drivers reveals new pathways and refines AML risk-stratification

SHORT TITLE: AML risk-stratification and subtyping by methylome

Burcu Gurun[1, 2, 3], Jeffrey W. Tyner[1,4], Emek Demir[1], Brian J. Druker[1,4,5]* and Paul T. Spellman[1, 3*]

[1] Knight Cancer Institute, Oregon Health & Science University, Portland, OR; [2] School of Medicine, Oregon Health and Science University, Portland, OR; [3] Cancer Early Detection Advanced Research Center, Knight Cancer Institute, Oregon Health & Science University, Portland, OR, United States, [4] Department of Cell, Developmental & Cancer Biology and Knight Cancer Institute, [5] Division of Hematology & Medical Oncology, Department of Medicine, Oregon Health & Science University, Portland, OR, USA, [6] Howard Hughes Medical Institute, Portland, OR, USA

*Paul Spellman: Email: spellmap@ohsu.edu

Correspondence may also be addressed to *Brian Druker Email: drukerb@ohsu.edu

---

Adapted from Gurun et al. 2023 [https://www.biorxiv.org/content/10.1101/2023.05.17.541249v1]

2.1. Abstract

Decades before its clinical onset, epigenetic changes start to accumulate in the progenitor cells of Acute Myelogenous Leukemia (AML). Delineating these changes can improve risk-stratification for patients and shed insights into AML etiology, dynamics and mechanisms. Towards this goal, we extracted “epigenetic signatures” through two parallel machine learning approaches: a supervised regression model using frequently mutated genes as labels and an unsupervised topic modeling approach to factorize covarying epigenetic changes into a small number of “topics”. First, we created regression models for *DNMT3A* and *TET2,* the two most frequently mutated epigenetic drivers in AML. Our model differentiated wild-type vs. mutant genotypes based on their downstream epigenetic impacts with very high accuracy: AUROC 0.9 and 0.8, respectively. Methylation loci frequently selected by the models recapitulated known downstream pathways and identified several novel recurrent targets. Second, we used topic modeling to systematically factorize the high dimensional methylation profiles to a latent space of 15 topics. We annotated identified topics with biological and clinical features such as mutation status, prior malignancy and ELN criteria. Topic modeling successfully deconvoluted the combined effects of multiple upstream epigenetic drivers into individual topics including relatively infrequent cytogenetic events, improving the methylation-based subtyping of AML. Furthermore, they revealed complimentary and synergistic interactions between drivers, grouped them based on the similarity of their downstream methylation impact and linked them to prognostic criteria. Our models

identify new signatures and methylation pathways, refine risk-stratification and inform detection and drug response studies for AML patients.

2.2. Introduction

In many Acute Myeloid Leukemia (AML) cases, precancerous clonal expansions start in healthy tissue decades before the cancer onset(2–4). A hematopoietic stem cell may follow different paths to AML gaining the necessary pathological features (e.g., increased proliferation, abnormal morphology, and impaired differentiation of myeloid lineage cells, or combinations thereof) at different rates,(1) dictated by the interplay between the type and order of mutations, epigenetic events and tumor micro environment(5).

Before the acute presentation of a myeloid malignancy with anemia, neutropenia and thrombocytopenia, the disease is clinically silent and hard to detect. Recent single-cell sequencing studies(6,7) showed that this early phase is primarily driven by a small set of mutations(8–10). Clones frequently harbor multiple co-occurring mutations including *DNMT3A*, *TET2* and *ASXL1*, which are epigenetic regulators that change the normal physiological methylation landscape when mutated (6,7,11), thus conferring a broad risk of progression to a myeloid malignancy with the acquisition of a cooperating mutation(12). Once the disease progresses to an advanced stage it develops substantial clonal heterogeneity and becomes difficult to treat; only 20% of the patients survive for more than 5 years after diagnosis and recurrence is common even after complete remission(13).

These early drivers lead to a large number of Differentially Methylated Regions (DMRs) in the genome(14). Their effects can be captured as distinct "signatures" to gain a better understanding of the mechanisms and dynamics that drive early disease. DMR

signatures could be used for characterizing aberrant methylation in cancers(15), tumor sub-classification and distinguishing the tissue of tumor origin(16–18). For example, Vosberg *et al.*(16) showed that the AML clinical risk stratification based on genetic mutations European LeukemiaNet (ELN-2017) has good concordance with the DNA methylation based clustering. The authors suggested that the DNA methylation profiling could be used for AML risk stratification as subgroups of epigenetically homogeneous AML patients differ significantly in clinical outcomes. Similarly, Cabezon *et al.*(19) demonstrated that different methylation signatures at the time of diagnosis could predict response to azacytidine, a hypomethylating agent.

Compared to genomic profiles, methylome profiles offer two key benefits for liquid biopsy applications. First, a single genomic alteration can be associated with thousands of methylation changes, leading to significant signal diversity if the methylation changes have selective advantage. Secondly, various low frequency genomic alterations that have similar downstream effects can be grouped into methylation factors. These advantages are critical in both cancer early detection and disease monitoring applications, as they increase the power to detect rare subclonal expansions of pre-malignant diseases.

In this study we investigated how well the upstream genomic events are reflected in the methylome, and whether common downstream effects of groups of genomic events can be detected. To achieve this objective, we systematically modeled methylation signatures in AML using methylation, genetic, and clinical profiling data from 220 AML patients in the BeatAML cohort(20). We used two statistical approaches to capture these signatures. First, we identified frequent epigenetic signatures linked to mutations in *DNMT3A* and *TET2* genes and used supervised regression models to deduce methylation activities. We demonstrated that epigenetic signatures in methylation pathways were associated with *DNMT3A* and *TET2* mutations (**Figure 2-1**). Next, we utilized unsupervised topic modeling(21,22) to develop a latent

representation of the methylation landscape and labeled methylation signatures using biological and clinical factors. We showed that topic modeling can accurately identify the impacts of all major drivers, and these signatures are highly correlated with ELN-2017 prognosis(23). Our findings suggest that our approach can substantially improve risk-stratification for AML patients by revealing the previously unknown methylation signatures of relatively rare cytogenetic events, which will also greatly improve the subsequent methylation-based subtyping of patients for drug response studies and provide a framework for the usage of methylation as a biomarker.

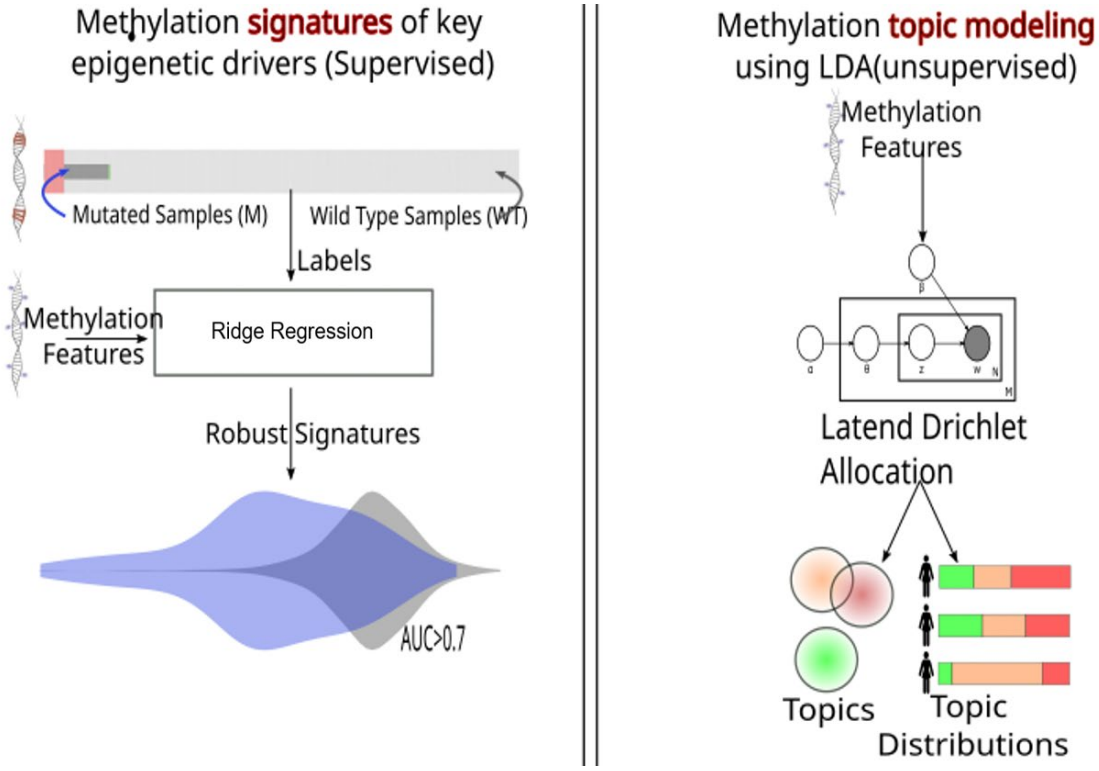


Figure 2-1 Overall Approach: Inferring methylation signatures that drives AML from methylation profiles. We use two parallel approaches: (a) regression models of epigenetic impact of epigenetic regulator mutations. (b) a latent representation of the epigenetic landscape using topic modeling.

2.3. Methods

2.3.1. Dataset

We used a subset of the BeatAML cohort that consists of 220 post-diagnosis AML samples profiled with three modalities: 1) Illumina Infinium MethylationEPIC assay(24) 2) matching exome sequencing and cytogenetic information and 3) matching clinical profiling(20).

Methylation values of each probe was normalized to follow an approximate Beta – valued distribution, with constrained to lie between 0 (unmethylated locus) and 1 (methylated locus), normalized based on the BMIQ method(25). Beta values indicate the probability that the corresponding locus is methylated.

2.3.2. Approach for supervised models

We used an Elastic Net based approach(26) to infer wild type vs mutant genotypes for *DNMT3A* and *TET2* based on their downstream epigenetic impacts. We trained regression models with different regularization parameters. Samples were divided into two sets: training and validation (80%), and test (20%). Training and validation set were used to optimize hyperparameter(s) with 10-fold cross validation. The test set was only used for obtaining performance measures of the model on new data. We used a variance filter of top 1% most variable methylation sites. We obtained the performance measures using receiver operator characteristic (ROC) curve analysis from the ROCR package(27). We evaluated 25 different 80% training and 20% test sets splits and observed the standard deviation of resulting models as well as hyperparameters to ensure that the models are highly robust to sampling bias. For further validation, we evaluated the biological relevance of the results by extracting the loci with the highest feature (CpG sites) coefficients and annotated them. For annotation purposes, we

used ridge regularized models to have more stable feature weights. In ridge regression, all coefficients are assigned with a nonzero value during regularization and coefficients with high absolute values are important in terms of predicting the mutation status, positive coefficients specifying hypermethylation and negative values specifying hypomethylation. We extracted a list of loci and annotated the genome based on their closest gene on the genome to reveal the relationship between the selected most important chromosome sites and the epigenetic regulator under investigation.

2.3.3. Approach for unsupervised models

We use topic modeling to reduce the feature space of 866,800 methylation loci to a latent space of small number of topics (factors) and represent each sample as a combination of a few topics. We used a Latent Dirichlet Allocation (LDA)(28) based approach implemented in the topicmodels package(29). This implementation provides explicit quantification of uncertainty, important for evaluating and comparing topics. While implementing the topic modeling, we optimized the model parameters including topic size as well as the size of the chromosome sites/probes provided as an input to the model. For validation with topic perplexity, the model evaluates how well the input matrix (methylation values for all chromosome sites of a new patient) is reconstructed from the output matrices, finding the hyperparameters minimizing the reconstruction error iteratively. We further evaluated model performance using topic coherence with biological and clinical information mapping. After inferring topics and assigning topic values to each patient with the specified optimal topic size of 15 and by taking the top 1% of the most variable genomic sites across patients, we annotated the topics with biological and clinical information. Then, we systematically tabulated the statistical

enrichment of known factors for each topic using –log p-values based on the student t-test.

2.3.4. Data sharing statement

All data used were from previously published studies(20,24). The source code for the analysis is available at https://github.com/gurundem/AML_methylation_signatures

2.4. Results

2.4.1. DNMT3A and TET2 have robust downstream methylation signatures

We computationally constructed signatures for the downstream impact of *DNMT3A* and *TET2*, and then, we used these signatures to shed light on their methylation pathways in AML.

Our rationale for modeling *DNMT3A* and *TET2* mutations has been four-fold: 1) For a given gene, our regression models require sufficient number of mutant cases in a given cohort to produce stable models. *DNMT3A* and *TET2* are the top two most frequently mutated epigenetic regulators. 2) *DNMT3A* and *TET2* are epigenetic regulator genes, directly impacting the methylation landscape and leaving a distinguishable signature on the DNA when mutated(17,30). 3) *DNMT3A* and *TET2* are known to confer a broad risk of converting to a myeloid malignancy with the acquisition of a cooperating mutation(12), making them important players of early AML etiology. 4) Mutations in these genes are found in almost all types of hematologic cancers(31), broadening the impact of our models.

Frequencies of *DNMT3A* and *TET2* mutations in AML are 22% and 11% respectively from prior work(20). The BeatAML dataset have similar frequencies: 22% (49) patients carry *DNMT3A* and 17% (37) carry *TET2* mutations.

We trained classifiers to infer the wild-type vs. mutant genotypes for *DNMT3A* and *TET2.* Regression models predict the mutation status of a given frequent epigenetic regulator from the downstream methylation changes. We calculated performance measures using receiver operator characteristic (ROC) curve analysis. Our results show that we trained highly robust and accurate classifiers with AUROC = 0.9 for *DNMT3A* and AUROC = 0.8 for *TET2* (**Figure 2-2**). This suggests that despite the complex genetic and epigenetic makeup of these post-diagnosis samples, the distinctive signatures left by epigenetic regulator mutations on DNA can be successfully extracted.

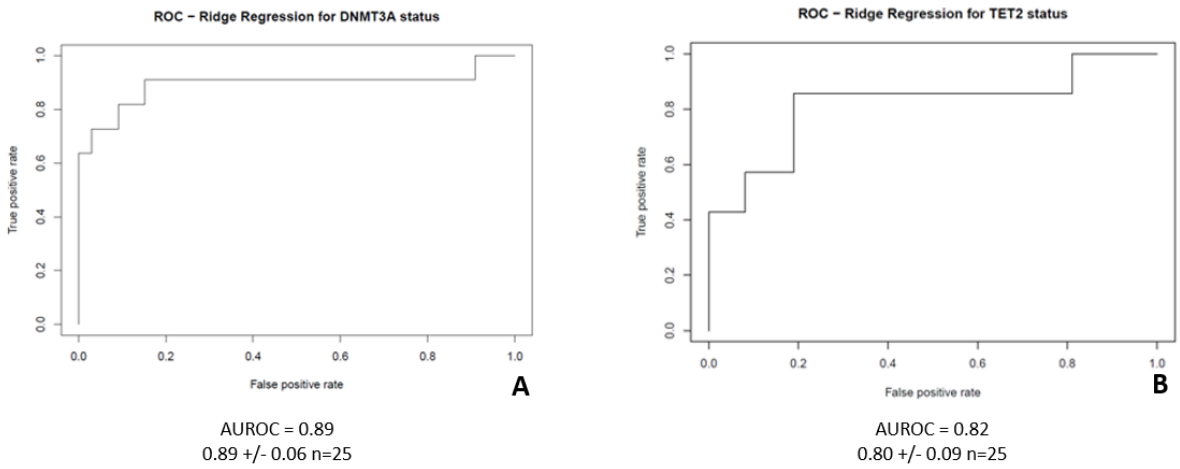


Figure 2-2.A-B. Performance of classifiers measured with AUROC (Area under the ROC curve) Regression models accurately classified wild type vs mutant genotypes based on their downstream impacts.

2.4.2. DMRs selected by DNMT3A and TET2 signatures highlight key downstream pathways

To gain more insight into the molecular processes that were indicative of the mutation status of *DNMT3A* and *TET2*, we annotated the genomic sites that has the highest coefficients in the model with the closest gene in the genome (**Table 2-1** and **Table 2-2**). We found that well-known targets of *DNMT3A*, such as *EVI1/MECOM*(32)

and AEBP2 (a member of the polycomb repressive complex), were often selected by the models. For example, hypomethylation of AEBP2 is strongly associated with *DNMT3A* mutations, and recent single-cell studies have suggested a link between mutated *DNMT3A* and preferential hypomethylation of targets of the polycomb repressive complex 2(33).

Known targets of *TET2* mutations were also selected by the models, including the *CEBPA-AS1* and ABR hypomethylation. Studies have shown that *CEBPA* and *TET2* mutations frequently co-occur and ABR is a transcriptional regulator of CEBPA and contributes to myeloid differentiation. In addition to known targets, we have also identified several novel loci that are strongly correlated with the mutation status of these genes. Further study of these sites could provide new insights into the role of methylation regulation in AML.

Gene	Potentially Relevant Associations
TM2D2	Found in 8p23.1 amp, linked to pediatric malignancies(34)
HYDIN	Indicated in familial MDS/AML(35)
N/A	
COX7A2	Part of a myeloid differentiation block regulated by homeobox TFs(36)
SAPCD2P2;VN1R28P	Pseudogene
WNK2	MEK pathway upstream regulator(37)
CHN2	Expression changes associated with lymphomas(38,39)
N/A	
N/A	
TENM2	Mutually exclusive with *MECOM* and *DNMT3A*, identified in BeatAML(20)
N/A	
HOXB2	Homoebox TF, known suppressor of *FLT3-ITD* driven AML(40)
C1orf87	
DYSF	Plays a role in monocyte differentiation(41,42)
AEBP2	Polycomb repressor complex member, frequently deleted in AML, associated with *DNMT3A* hypomethylation(33,43)
VKORC1L1	Next to *PIK3CG* – a potential tumor suppressor of AML.(44,45)
TXNRD1	Major regulator of metabolism in leukemia cells(46)
AGAP1	Part of a myeloid self-renewal block regulated by homeobox TFs(47)
RPS6KA2	rho gtpase(48)
MECOM	Translocations transform to AML/ mutually exclusive with *TENM2* and *DNMT3A*(32)

Table 2-1 Gene annotations for the top ten most frequently selected features for the *DNMT3A* status prediction both for hypermethylated (green) and hypomethylated (yellow) sites

Gene	Potentially Relevant Associations
FRMD6-AS2	
FOXK2;RP13-638C3.3	Genomic stability, DNA repair, cancer stem cell maintenance, cell proliferation, apoptosis and cell metabolism(49)
GTDC1	Glucosyltransferase high expression in blood leukocytes(50), 3′ MLL fusion partner in acute leukemia(51)
RP11-804A23.1	
SCUBE1;Z99756.1	Initiation and maintenance of MLL-AF9-induced leukemogenesis in vivo. Binds to *FLT3*(52)
TBCD	Shown to be differentially methylated during granulopoiesis(53)
PHACTR1	Cytoskeletal regulation(54)
EHD4	
DTHD1	Apoptotic control(55)
GDF7	
FAM155A	
IQSEC1	
TRAP1	Key mitochondrial regulator(56)
CEBPA-AS1;CTD-2540B15.9	*CEBPA* regulator, In *CEBPAdm* cases for concomitant mutations, *TET2* found to be most frequently mutated(57)
C1orf112;SELL	
ABR	*CEBPA* regulator, TF contributes to myeloid differentiation(58)
TPO	
RP11-191L9.4	
RGS17	*GPCR* regulator(59)
DMRTA1	Cell differentiation(60)

Table 2-2 Gene annotations for the top ten most frequently selected features for the *TET2* status prediction both for hypermethylated (green) and hypomethylated (yellow) sites.

2.4.3. Factorization of methylation changes into Topics link genomic events into distinct methylation programs

Using topic modeling, we reduced the methylation profiles to a latent space of 15 topics. For each patient, topic values represent probabilities of enrichment for each topic based on the

most co-variable DMRs across patients. We observe that 15 topics captured major mutations and cytogenetic events as well as having a prior MDS, gender and ELN-2017 risk stratification by genetics(23). Each patient has different combinations of driver events, but topic modeling successfully deconvoluted the impact of these drivers in the methylome, even for the relatively low frequency chromosomal alterations (**Figure 2-3**).

Figure 2-4 tabulates the statistical enrichment of known factors for each topic. For each topic and epigenetic factor pair, we calculated whether that topic's value is significantly higher in patients with the factor compared to patients without the factor to methodologically quantify enrichment.

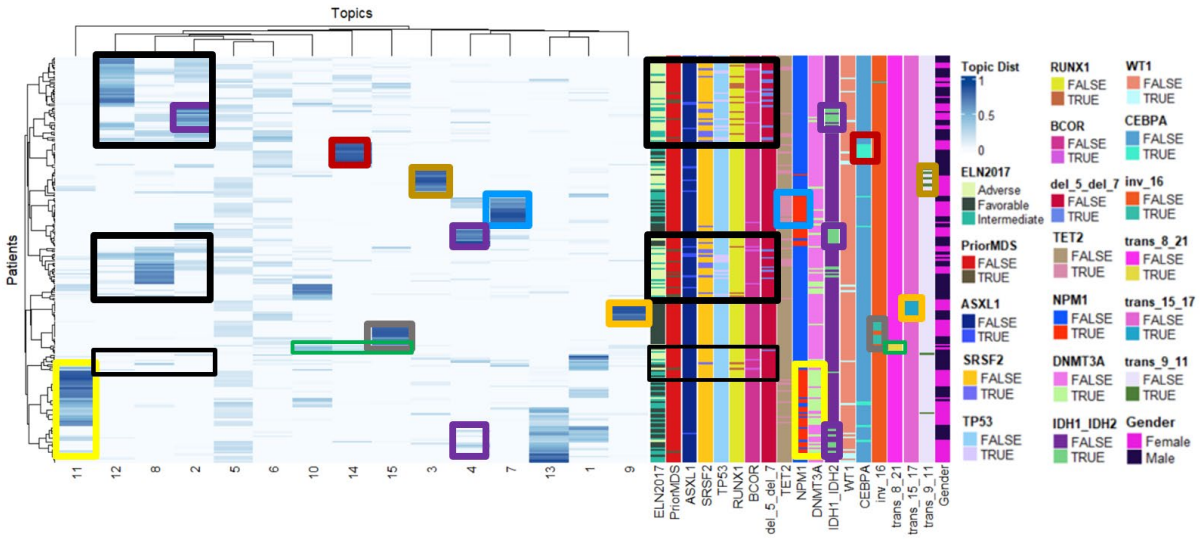


Figure 2-3. Topic Distribution for 220 post-diagnosis AML patients: Topic size is 15 (columns). For each patient, topic values represent probabilities of enrichment for each topic based on the most variable DMRs across AML patients. Topic 11 is enriched for *DNMT3A* and *NPM1* co-mutations (yellow rectangles). Topic 7 is enriched for *TET2* and *NPM1* co-mutations (blue rectangles). Topic 4 (primarily) and topic 2 are enriched for *IDH1* and *IDH2* (purple rectangles). *NPM1*'s effect is further distributed to topics 4 and 13. Topic 15 is enriched for inv16 (grey rectangles) and t(8,21) and topic 14 is enriched for *CEBPA* (red rectangles) and t(8,21). t(8,21)'s effect is further distributed to topic 10 (green rectangles). Topic 5 is enriched for gender. *ASXL1*, *SRSF2*, *TP53*, *RUNX1*, *BCOR* and 5q and 7q deletions associated topics are aligning with Prior MDS and ELN-2017 adverse category (black rectangles). The other common chromosomal events t(15, 17) and t(9,11) almost exclusively map to topics 9 and 3 respectively (orange and gold rectangles respectively).

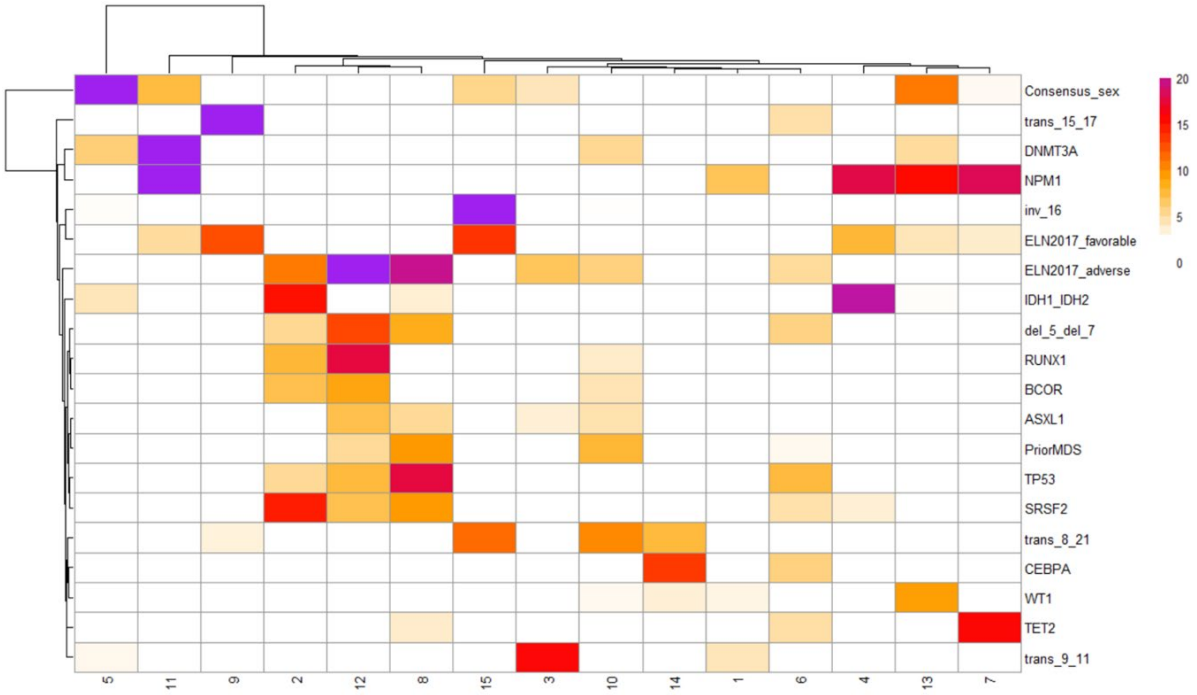


Figure 2-4.Enrichment heatmap of topics vs. epigenetic factors: Cell colors indicate the significance of whether the average topic values of the patients with the factor (e.g. *DNMT3A* mutant) is significantly higher than the average topic values of the patients without that factor (e.g. *DNMT3A* WT), reflecting –log p-values based on the one sided t-test. Gender is strongly associated with topic 5, *NPM1* and *DNMT3A* co-mutations, topic 11, t(15, 17), topic 9, (inv 16), topic 15, *IDH1* and *IDH2*, topic 4 (primarily) and 2, t(9, 11), topic 3, *CEBPA*, topic 14 (primarily), and *NPM1* and *TET2* co-mutations, topic 7. *NPM1*'s effect is further distributed to topics 4 and 13. t(8,21)'s effect is distributed to topic 15 (primarily), 10 and 14. *RUNX1*, *BCOR*, *ASXL1*, *TP53*, *SRSF2*, 5q and 7q deletions, having a prior MDS and ELN2017 adverse risk category associated topics are clustered together.

Strong statistical associations between eight topics and specific driving events including topics that represent gender (topic 5), *NPM1* and *DNMT3A* co-mutations (topic 11), translocation of chromosomes 15 and17 (topic 9), inversion of chromosome 16 (topic 15), *IDH1* or *IDH2* mutation (topic 4 and less strongly to topic 2), translocation of chromosomes 9 and 11 (topic 3), *CEBPA* mutation (topic 14) and *NPM1* and *TET2* co-mutations (topic 7) are identified (**Figure 2-3** and **Figure 2-4**). The effect of *NPM1* is also distributed to topics 4 and 13, while the effect of translocations of chromosome 8 and 21 is distributed to topics 15 (primarily), 10 and

14. Topics associated with *RUNX1*, *BCOR*, *ASXL1*, *TP53*, *SRSF2*, 5q and 7q deletions, having prior MDS and the ELN2017 adverse risk category are clustered together, , demonstrating their similar downstream methylation impacts.

We observe that these epigenetic topics can be broadly classified into four groups for AML **a)** gender associated, **b)** cytogenetic events with a distinct signature, **c)** co-mutations with *NPM1* **d)** A heterogenous set with *TP53, SRSF2, ASXL1, BCOR, RUNX1* mutations and 5q and 7q deletions that is associated with prior MDS and adverse prognosis.

We then asked if the DMRs that have high coefficients for a given topic recapitulate known downstream pathways or reveal novel biology. We annotated each locus with the closest gene in the genome (**Table 2-3** and **Supp. Figure 1**). We observe that topic group **a** (1, 5, 6) predictably have DMRs located on chromosomes X and Y. For other topics, multiple downstream pathways have been found to occur repeatedly across various topics, including pathways associated with homeobox (*HOX*) genes, histone deacetylases, lipid metabolism and maintenance of stem-cell state. Topic group **d** (2,8,12) had substantial enrichment in remodeling and differentiation pathways as well as expression of T-Cell receptors, potentially indicating a de-differentiation mechanism to other hematopoietic lineages.

Supp. Figure 2 and **Supp. Figure 3** show ELN-2017 adverse and ELN-2017 favorable risk category groups of mutations and events clustering together, respectively. In addition, we showed that there is no topic enriched in association with *FLT3* mutations—this is in agreement with previous findings(61) (**Supp. Figure 4**).

We also tested whether topic enrichment for cytogenetic events, such as t(15;17) are achieved through their downstream methylation impact or through the detection of their break points by the topic modeling algorithm. For example, topic number 9 is exclusively enriched for t(15; 17) (**Figure 2-3** and **Figure 2-4**). We annotated the top ten genomic sites with the highest

assigned topic values on the genome and none were mapped on chromosomes 15 or 17 (**Table 2-3 and Supp. Figure 4**), suggesting that cytogenetic aberrations are categorized based on their *trans*, downstream methylation impact rather than *cis* effects.

Topic number 1		**Topic number 2**		**Topic number 3**	
Chromosome	Gene	Chromosome	Gene	Chromosome	Gene
chrY	KALP/ANOS2P	chr2	PTH2R	chr7	AC110781.3;MAD1L1
chrY	NA	chr2	PTH2R	chr10	JMJD1C
chrY	ANKRD36P1	chr2	PTH2R	chr1	CD247
chrY	HDHD1P1/PUDPP1	chr1	RFWD2;RP11-195C7.1/COP1;RP11-195C7.1	chr3	TM4SF19;TM4SF19-TCTEX1D2
chrY	KALP/ANOS2P	chr17	C17orf64;USP32	chr17	MSI2
chrY	NLGN4Y	chr21	DIP2A	chr5	NA
chrY	CD24P4;TTTY14	chr14	OTUB2	chr8	NA
chrY	NA	chr2	PTH2R	chr4	NA
chrY	NA	chr13	NA	chr6	CNKSR3;IPCEF1
chrY	KALP/ANOS2P	chr2	FBLN7	chr3	PPARG
Topic number 4		**Topic number 5**		**Topic number 6**	
Chromosome	Gene	Chromosome	Gene	Chromosome	Gene
chr3	NFKBIZ;NXPE3	chrY	DDX3Y	chrY	NA
chr9	SUSD3	chrY	TMSB4Y	chrY	NA
chr9	SUSD3	chrY	TMSB4Y	chrY	HDHD1P1/PUDPP1
chr1	CD34	chrY	RPS4Y2	chrY	ANKRD36P1
chr15	NA	chrY	FAM41AY1	chrY	NLGN4Y
chr21	PDE9A	chrY	TTTY10	chrY	NA
chr12	TMCC3	chrY	PRKY;RN7SKP282	chrX	FIRRE
chr21	MAP3K7CL	chrY	NA	chrY	RBMY1A3P
chr5	NA	chrY	DDX3Y	chrY	NA
chr2	HDAC4	chrY	KDM5D	chrY	NA
Topic number 7		**Topic number 8**		**Topic number 9**	
Chromosome	Gene	Chromosome	Gene	Chromosome	Gene
chr21	PDE9A	chr13	LMO7;RP11-29G8.3	chr21	MIR99AHG
chr9	MOB3B;RP11-298E2.2	chr12	NA	chr2	LPIN1
chr7	AC007091.1	chr4	GALNT7	chr11	NA
chr2	TBC1D8	chr6	GCNT2	chr6	RPS6KA2
chr5	NA	chr10	TCF7L2	chr4	NA
chr1	NA	chr4	GALNT7	chr4	LRPAP1
chr6	JARID2	chr16	LA16c-444G7.2	chr2	LPIN1
chr1	GNG12-AS1;WLS	chr12	NA	chr15	NA
chr7	TRGC1;TRGC2;TRGV9	chr12	BCL7A	chr6	JARID2
chr7	NA	chr12	KDM2B;RP13-941N14.1	chr3	TM4SF19;TM4SF19-TCTEX1D2
Topic number 10		**Topic number 11**		**Topic number 12**	
Chromosome	Gene	Chromosome	Gene	Chromosome	Gene
chrY	NA	chr17	HOXB-AS3;HOXB3	chr8	NA
chr1	NA	chr17	HOXB-AS3;HOXB3	chr14	IGHG1;IGHG3;IGHJ3P;IGHJ6;IGHM
chrY	NA	chr17	HOXB-AS3;HOXB3	chr1	NA
chrY	ANKRD36P1	chr17	HOXB-AS3;HOXB3	chr3	RN7SL36P;XXYLT1;XXYLT1-AS2
chr17	ARHGAP23	chr17	HOXB-AS3;HOXB3	chr17	WNK4
chr19	NFIX	chr17	HOXB-AS3;HOXB3	chr7	GIMAP7;STRADBP1
chr19	NFIX	chr17	HOXB-AS3;HOXB3	chr12	DYRK4
chr11	VPS37C	chr17	HOXB-AS3;HOXB3	chr7	GIMAP7
chrY	PCDH11Y	chr17	HOXB-AS3;HOXB3	chr7	TRBC2;TRBJ2-2;TRBJ2-2P;TRBJ2-3;
					TRBJ2-4;TRBJ2-5;TRBJ2-6;TRBJ2-7
chrY	NA	chr17	HOXB-AS3;HOXB3	chr3	RN7SL36P;XXYLT1;XXYLT1-AS2
Topic number 13		**Topic number 14**		**Topic number 15**	
Chromosome	Gene	Chromosome	Gene	Chromosome	Gene
chr19	OSCAR	chr3	OPA1	chr3	NA
chr7	TCAF1	chr2	C2orf42	chr8	DENND3
chr8	NA	chr2	C2orf42	chr2	MEIS1
chr19	OSCAR	chr7	HOXA7	chr3	OPA1
chr22	MIRLET7BHG	chr7	SKAP2	chr6	CNKSR3;IPCEF1
chr17	HOXB-AS3;HOXB3	chr3	CD96	chr3	TM4SF19;TM4SF19-TCTEX1D2
chr17	HOXB-AS3;HOXB3	chr5	LINC01183/CCDC192	chr4	LRPAP1
chr18	NA	chr15	RP11-507J18.1	chr2	MEIS1
chr22	MIRLET7BHG	chr22	ZMAT5	chr7	HOXA9
chr7	CLIP2	chr2	AC098617.1;SDPR/AC098617.1;CAVIN2	chr2	MEIS1

Table 2-3. Annotation of the top 10 most enriched loci for all 15 topics

2.5. Discussion

We systematically modeled the epigenetic impact of genomic events using supervised and unsupervised approaches. Individual *DNMT3A* and *TET2* signatures were detected with high accuracy and robustness, as they yielded high AUROCs with a testing error that was consistent with training error across multiple training rounds, even amidst the complex genetic and epigenetic landscape of post-diagnosis samples. Methylation loci commonly chosen by the models includes well-known downstream targets such as *EVI1/MECOM* and *AEBP2* for *DNMT3A*, and *CEBPA-AF1* for *TET2*, all of which exhibited hypomethylation. A hypomethylated loci near *HOXB2* was a strong predictor of mutant *DNMT3A* status although there are conflicting reports in the literature of the overall effect of *DNMT3A* mutation on the methylation of the *HOX* promoters(62–64). This observation strongly suggests that specific loci might have stronger signal for the upstream regulator status as opposed to methylation patterns over a genomic region. Further investigation of these sites can lead to new mechanistic insight on methylation pathways important in leukemogenesis. As the cohorts become larger, these approaches can be extended beyond *DNMT3A* and *TET2* to other potential AML epigenetic drivers including cytogenetic events and related methylation pathways can be investigated.

Our regression models require labeled data (such as mutation status) and allow us to model the impact of each driving factor (e.g. *DNMT3A* mutation status) independently. However, we need substantially larger cohorts for building models of infrequent drivers such as cytogenetic events. Furthermore, the analysis may be confounded by co-occurring mutations. To complement this approach, we used topic modeling on the same dataset to factorize methylation profiles. This yielded a representation of each patient's methylation pattern as a combination of multiple topics. For each topic we identified mutations and clinical classes that are statistically enriched. We observe that identified topics fell into 4 broad categories: Gender

associated, cytogenetic events, *NPM1* co-mutations and a heterogeneous set of mutations in *TP53*, *SRSF2*, *ASXL1*, *BCOR*, *RUNX1*, and deletions of chromosomes 5q and 7q. The latter group is associated with adverse-risk prognosis and topic modeling reveals that they have converging downstream methylation impacts.

We have shown that DNA methylation-based categorization achieved by topic modeling resulted in a good concordance with the ELN-2017 risk stratification by genetics and having a prior MDS. Vosberg *et al.*(16) and Figueroa *et al*(61) previously pointed to the potential of DNA methylation profiling to refine AML risk stratification and our study reinforces these observations. However, we improve upon these hard-label clustering techniques by representing each patient's pattern as a combination of several factors as opposed to belonging to a single subtype. This factorization and deconvolution are crucial, given the highly combinatorial nature of the disease. We, in fact see that less frequent factors such as cytogenetic events crosscut previously defined subtypes, methylation effects of which have been obfuscated by the effects of more frequent drivers in a hard-clustering approach. Through topic modeling, we can deconvolute the impact of overlapping factors, creating an improved latent representation of the disease state. As a result of this, our models have made substantial improvements in methylation-based risk-stratification by uncovering previously unknown signatures associated with relatively rare cytogenetic events. This may greatly enhance the subtyping of patients for drug response studies and provide a framework for using methylation as a biomarker in subsequent studies.

We reviewed the frequently selected methylation loci to gain insight on the downstream methylation impact. We observed multiple downstream pathways re-occurring across multiple topics including *HOX* genes, stem-cell and de-differentiation associated pathways as well as histone deacetylases. This implies that although different genetic regulators have distinct

methylation impact, they might be converging on common downstream processes and these convergence points might be excellent therapeutic targets. We also observed several downstream pathways that were not canonically associated with upstream genetic events such as parathyroid hormone receptor, lipid homeostasis and protein glycosylation. Perhaps more importantly, we observed that the topic modeling was able to systematically capture the impact of key epigenetic regulating events and simultaneously grouped events that have similar or synergistic impacts together.

As these signatures help us quantify the relationship between a driver event and its downstream methylation impact, they can be leveraged to build predictive models for AML-risk. Methylation changes can be tracked throughout leukemogenesis in early detection cohorts through simple liquid biopsies. Risk-stratification based on downstream epigenetic signatures can help us determine which patients are at risk and who should be monitored more frequently. We can also use these models to test if the function of a regulator is restored or disrupted following treatment, such as hypomethylating agents targeting epigenetic regulator functioning, and stratify who can benefit from the treatment. Since modifications by epigenetic mutations are reversible with therapy, broadening our understanding of the impact of epigenetic mutations on leukemogenesis and therapeutic response will be essential for advancing treatment of myeloid malignancies(31).

2.6. Acknowledgement

BG and PTS acknowledges support from the Cancer Early Detection Advanced Research Center (CEDAR) at the Knight Cancer Institute.

JWT acknowledges support from the Aqcuired Resistance to Therapy Network (ARTNet), National Institutes of Health (NIH), and National Cancer Institute (NCI) grant

U54CA224019. JWT also received support from NCI award R01CA262758 (JWT, SEK), the V Foundation for Cancer Research (JWT), the Gabrielle's Angel Foundation for Cancer Research (J.W.T.), the Anna Fuller Fund (J.W.T.), the Mark Foundation for Cancer Research (J.W.T.), and the Silver Family Foundation (J.W.T.).

The authors thank the Knight Cancer Institute (NCI P30 CA069533), Bioinformatics, and CEDAR. We are grateful to members of the Spellman Lab, Blundell Lab and to members of the CEDAR; Ece Eksi, Gurkan Yardimci, Hisham Mohammed, Stefanie Lynch, Ruslan Strogantsev and Jose Luis Montoya Mira for rich discussions.

2.7. Authorship Contributions

BG built the statistical models and methods, analyzed and interpreted the data, plotted the figures. ED contributed to outlining the aim of the project and to our understanding of the computational methods and their applications around biology. JWT built collaborations across different labs, advised BG on the AML aspects of the analyses. BJD provided resources, mentored and advised BG. PTS supervised and mentored BG and reviewed all preceding and analyses herein and provided feedback. All authors read and approved the final manuscript.

An open protocol for modeling T Cell Clonotype repertoires using TCRβ CDR3 sequences

Burcu Gurun[1, 2*], Wesley Horton[1], Dhaarini Murugan[3], Biqing Zhu[4], Patrick Leyshock[1], Sushil Kumar[3], Katelyn T. Byrne[3, 5], Robert H. Vonderheide[5], Adam A. Margolin[6], Motomi Mori[7], Paul T. Spellman[1*], Lisa M. Coussens[1,3*] and Terence P. Speed[8,9*]

[1] Knight Cancer Institute, Oregon Health & Science University, Portland, OR; [2] School of Medicine, Oregon Health and Science University, Portland, OR; [3] Department of Cell, Developmental & Cancer Biology and Knight Cancer Institute, Oregon Health & Science University, Portland, OR; [4] Computational Biology and Bioinformatics Program, Yale University, New Haven, CT; [5] Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [6] NextVivo, Palo Alto, CA [7] Department of Biostatistics, St. Jude's Children's Research Hospital, Memphis, TN; [8] Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia; [9] School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia; [10] Massively parallel Sequencing Core, Oregon Health & Science University, Portland, OR

*Burcu Gurun: Tel: +1-503-494-8220 Email: gurundem@ohsu.edu

Correspondence may also be addressed to

*Paul Spellman: Tel: +1-503-494-9895 Email: spellmap@ohsu.edu

*Lisa Coussens: Tel: +1-503-494-9336 Email: coussenl@ohsu.edu

*Terence P Speed: Tel: +61 3 9345 2555 Email: terry@wehi.edu.au

---

Adapted from Gurun et al. 2023

[https://www.biorxiv.org/content/10.1101/2022.03.30.486449v1]

3.1. Abstract

T cell receptor repertoires can be profiled using next generation sequencing (NGS) to measure and monitor adaptive dynamical changes in response to disease and other perturbations. Genomic DNA-based bulk sequencing is cost-effective but necessitates multiplex target amplification using multiple primer pairs with highly variable amplification efficiencies. Here, we utilize an equimolar primer mixture and propose a single statistical normalization step that efficiently corrects for amplification bias post sequencing. Using samples analyzed by both our open protocol and a commercial solution, we show high concordance between bulk clonality metrics. This approach is an inexpensive and open-source alternative to commercial solutions.

3.2. Background

The receptors on the surface of T cells bind to an enormous array of antigens that play a pivotal role in shaping immune response during health and disease. The T cell receptor (TCR) is a heterodimer composed of one alpha and one beta chain which are encoded by the *TCRα* and *TCRβ* genes, respectively. To recognize an extremely large antigen space, the TCR genomic loci undergo somatic recombination of variable (V), diversity (D), and joining (J) gene segments, and generate a diverse repertoire of TCRs. The complementarity determining region 3 (CDR3) region present at the D segment of the recombined *TCRβ* gene is highly diverse in TCR beta chains. Therefore, surveying the recombined *TCRβ* gene or transcript as a proxy for overall TCR repertoire diversity has emerged as a rational approach to study TCR repertoire dynamics.

Over the last ten years it became possible to obtain comprehensive profiles of TCR through an array of next generation sequencing (NGS) based approaches(65–72). These vary based on the experiment type (bulk or single-cell), sample type (RNA or DNA), library preparation

(multiplex PCR, bead-based enrichment, 5'RACE) and sequencing platform, each choice presenting a different, and fascinatingly interlocked trade-off. Single cell approaches compared to bulk can be very accurate and unbiased for high frequency clones, but have lower resolution for low frequency clones(73,74). They are also substantially more expensive and require intact cells. RNA-based approaches are affected by the variability in TCR RNA expression levels, but may better reflect diversity when the sample size is limited(72). Genomic DNA (gDNA)-based approaches require either multiplex PCR or target enrichment during library preparation which introduces biases(75). A recent comparison of these two approaches confirmed these issues for both RNA- and DNA-based methods but also found the methodological variability to be smaller than the biological variability(76). New innovations are being introduced rapidly for all of these approaches but currently there is no established gold standard.

Multiplexed PCR-based bulk sequencing approaches using gDNA, however, have become the standard approach for translational and even clinical applications due to reasonable sample requirements and moderate costs(77). This is reflected in the fact that all currently available commercial TCR sequencing products offer this as their major DNA option (Adaptive Biotechnologies(78), BGI(79), iRepertoire). Multiplex PCR refers to the usage of multiple forward primers specific for the V segments and multiple reverse primers specific for the J segments in combination during the initial amplification for target enrichment. Since each primer pair will have a different efficiency multiplexing will distort the relative abundances of the VDJ segment combinations(80). Correcting for this amplification bias is a key challenge for accurate quantification. One approach (78) is using spiked-in oligomers (synthetic templates) for each primer pair to measure primer efficiencies and to carefully control the design of the primers and their concentrations accordingly. Assuming that there is no interaction between the efficiency of primer pairs, amplification bias is reduced by iteratively calibrating the primer concentrations to

find the optimal primer mix, and then removing any remaining amplification bias computationally using spiked-in oligomer counts.

The proprietary setup described in (78) is currently available for human and murine samples exclusively through commercial kits. These are difficult and labor-intensive to adapt if requiring different settings, e.g. a different mammal or application. Furthermore, synthetic templates when added to all of the samples with the kits can substantially increase preparation and sequencing costs, as well as decrease coverage for clonotypes.

3.3. Results

3.3.1. Amplification bias due to multiplex PCR is reproducible

To amplify all possible TCR somatic recombination products, we performed a multiplex PCR with 20 different V-specific forward primers and 13 different J-specific reverse primers. Since the differences in efficiency of primer pairs can produce significant amplification bias in TCR clonotypes, we spiked-in 260 synthetic TCR templates (ST) in equimolar concentration as internal controls to the multiplex PCR reaction in order to measure and control bias (**Figure 3-1**).

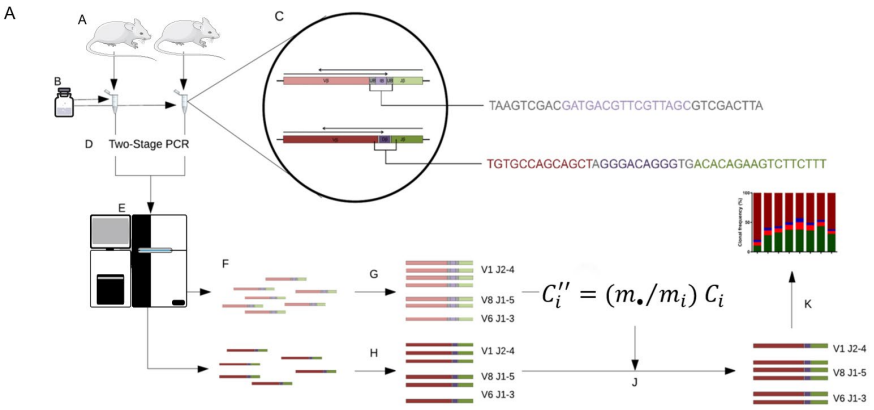


Figure 3-1 Overview of the TCR sequencing analysis pipeline : **A.** gDNA extracted from freshly resected murine peripheral blood, peritoneal orthotopic mesotheliomas, and spleen tissues. **B.** Equimolar mixture of synthetic TCR templates (ST) were then added, and **C.** followed by addition of forward and reverse sequences of ST (top) and TCRβ (bottom). For ST, universal 9-bp barcode (grey), unique 16-bp barcode (purple). For TCRβ, V-region (red), VD/DJ-junctions (grey), D-region (purple), J-region (green). **D.** Samples amplified with multiplex PCR followed by second-stage barcoding PCR. **E.** Samples were then pooled for sequencing. **F.** ST and TCRβ were then separated using universal barcodes, and **G.** ST was quantified using unique barcodes. **H.** TCRβ clonotypes were quantified with the MiXCR tool suite. **I.** Negative binomial normalization was used to remove amplification bias. **J.** Scaling factors were applied to counts, and **K.** then used to normalize counts for diversity analyses.

We hypothesized that the estimated mean counts of the 260 ST would scale proportionally across samples and experiments. To test this, we measured the distribution and variation of ST counts across 20 ST-only samples in the absence of genomic DNA **(Figure 3-2).**

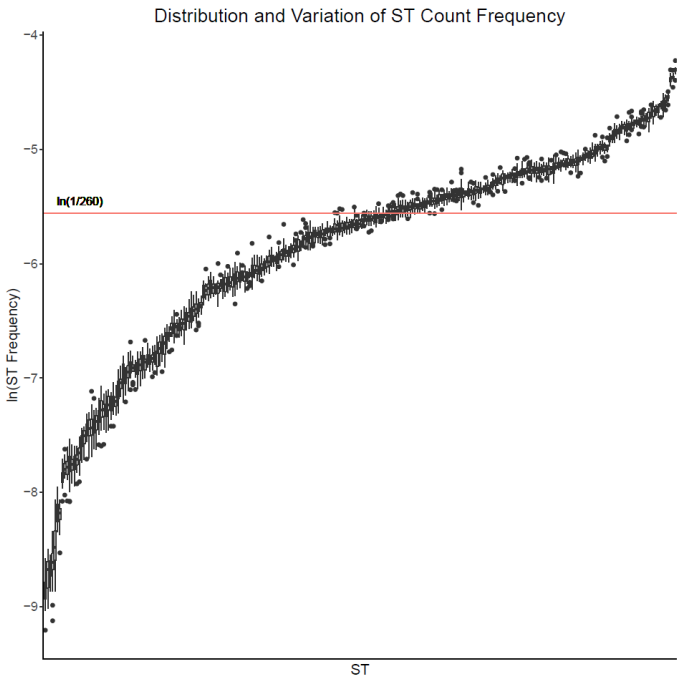


Figure 3-2 ST proportion distributions: Plot showing stability for relative frequencies of ST counts within individual samples, and amplification bias based on the reproducibility of the multiplex PCR. Median IQR of ST-to-ST variation was ~20-fold greater than the median IQR of experiment-to-experiment relative frequencies of ST counts. Target values aim to be at the ln (1/260) line in the absence of amplification bias. Data derived from twenty ST-only samples described in the *ST-only data sets.*

Within each sample, ST counts were converted to relative frequencies to reduce the effect of random sample-to-sample variation on the comparisons. For a given ST, deviation of the median relative frequency across all samples from 1/260 is a measure of the PCR amplification bias for the corresponding primer pair, while the spread of relative frequencies is an indication of the random sample-to-sample variation. We observed that the ST-to-ST variation was much larger than the sample-to-sample variation within an individual ST (~20-fold difference in respective median IQRs). These differences indicate that the observed differences in ST counts are primarily caused by amplification bias of different primer pairs. The ln (1/260) line indicates the (log) expected ST relative frequency in the absence of amplification bias. To demonstrate that the above observations apply in the presence of TCR clonotypes, we obtained

ST counts from samples with genomic DNA extracted from P14 and OT-1 TCR transgenic mice where CD8 T cells primarily recognize $OVA_{257-264}$ when presented by the MHC I molecule. A similar relationship was found between ST relative frequencies in the presence of TCR clonotypes across samples, though the sample-to-sample variation is noticeably larger (**Supp. Figure 5**).

3.3.2. A negative binomial model fits the data

The fit of count data on 260 ST from each of 20 ST-only samples to the negative binomial (NB) with ST-specific means and a common dispersion parameter *d* can be informally assessed by examining an empirical variance (*v*) vs mean (*m*) plot with the line $v = m + dm^2$ displayed, all with log-log scales.

At the high end, we expect an approximate linear relationship between log mean and log variance for the ST counts, with a slope of about *2*, since $\log v \approx \log d + 2\log m$ for large m. In **Figure 3-3,** we took $d=0.125$, as this was the median value of the dispersion estimates found by fitting separate NB distributions to the 260 sets of 20 ST counts.

The fact that the ST counts fit NB distributions with approximately similar overdispersion parameters reassures us the variation we are seeing is in some sense natural and that the system is in control(81).

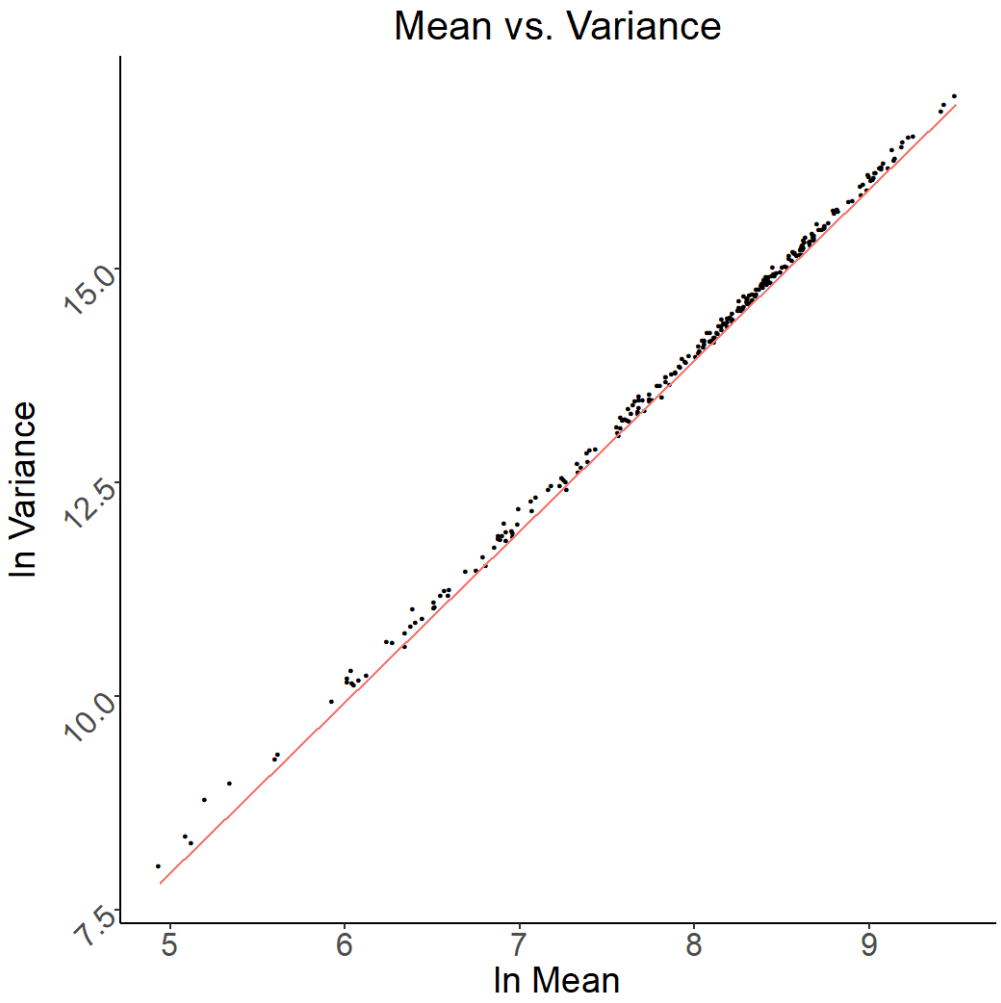


Figure 3-3. A negative binomial model fits the data : Observed mean-variance relationships fit as compared to the negative binomial (NB) model. Dots are observed values derived from 260 ST count distributions with a separately fitted dispersion parameter. The red line is the relationship predicted by the NB model with fixed d (0.125). Data derived from twenty ST-only samples described in the *ST-only data sets.*

3.3.3. Scaling factors are relatively stable across experiments

We fit the NB model to three different sets of ST-only observations at different equimolar concentrations, coming from two different batches: (1) a set of 20 samples from one batch, (2) two sets of 10 samples from another batch. To demonstrate that scaling factors were relatively stable across ST-only samples, we compared the sets of 260 $(m_i/m_\bullet)$ values based on 20 ST-

only samples, to those based on the sets of 10 ST-only samples. We observed the ($m_i/m_\bullet$) values to be highly correlated on a *log-log* scale (Pearson r = 0.83 and 0.97, p-value < 2×10^{-16} for both comparisons) (**Figure 3-4**), indicating that scaling factors are relatively stable across experiments.

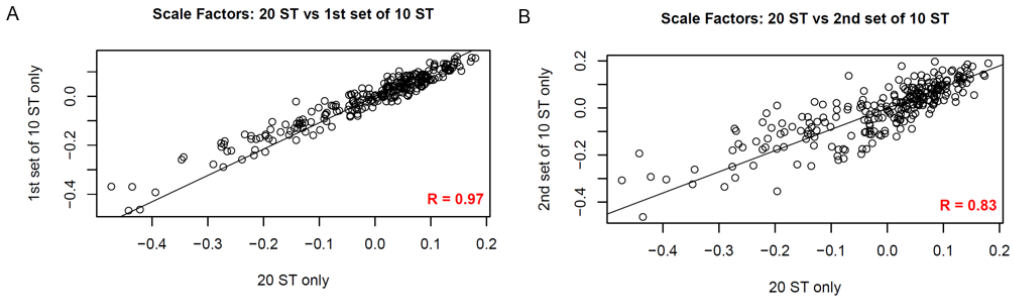


Figure 3-4.A-B. Stability of scaling factors : Scatter plots (*log-log* scales) comparing the sets of 260 ($m_i/m_\bullet$) values based on 20 ST-only samples to those based on the two sets of 10 ST-only samples (described in the Data Sets section). The ($m_i/m_\bullet$) values are observed to be highly correlated across different batches and samples. (Pearson r = 0.97 and 0.84, p-value < 2×10^{-16} for both comparisons).

We then computed estimates of the (m_i) by pooling data across these sets, adjusting for the concentration differences, calling the results the combined estimates. Pooling also deals with errors introduced to the ST-only counts by processing samples in different batches along with different samples. The combined estimates of the (m_i) are therefore used for normalization below (**Supp. Material 1**).

3.3.4. Normalization considerably reduces amplification bias

We normalized the 20 ST-only measurements with the combined estimates of the (m_i). The spread of the ST counts for 20 samples was considerably reduced after normalization (**Figure 3-5**). A similar comparison of ST counts for 20 samples, in the presence of genomic

DNA derived from mesothelioma tumors, revealed that the reduction in spread of ST normalized counts was present, although less pronounced (**Supp. Figure 6**).

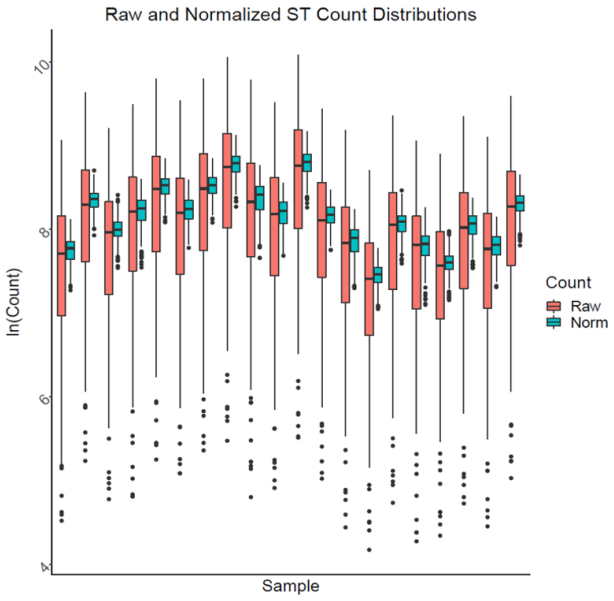


Figure 3-5. NB normalization reduces amplification bias: Variation in ST counts within samples is apparent before normalization (red), however, the ST-to-ST differences within each sample are reduced to less than two-fold after normalization (cyan). Data derived from twenty ST-only samples described in the *ST-only data sets.*

To validate the normalization procedure further, we assessed the observed ratio of monoclonal counts before and after normalization utilizing the 50:50 mixture of P14 and OT-1 TCR transgenic monoclonal DNA. After normalization, the differences between the transgenic TCR counts were substantially reduced for all samples, as represented by the smaller deviations of the proportion of the dominant clonotype from ½ following normalization (**Figure 3-6**).

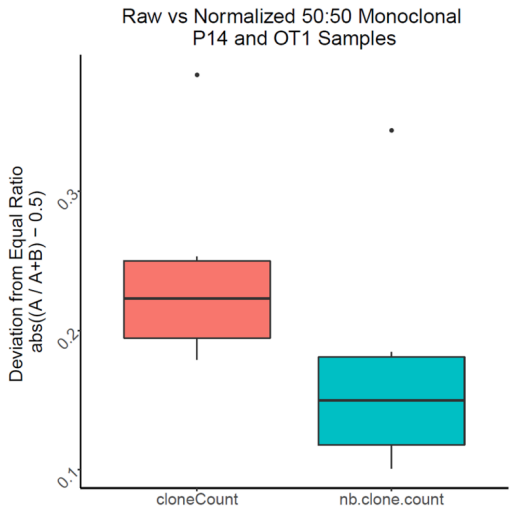


Figure 3-6. Amplification bias of spleen genomic DNA of P14 and OT-1 TCR transgenic mice : To further evaluate normalization parameters, a 50:50 mixture of P14 and OT1 TCR transgenic monoclonal DNA was utilized to examine differences between transgenic TCR counts (red) that were reduced for all samples following normalization (cyan), as observed by the smaller deviations from ½ of the proportions of the dominant clonotype. Red color indicates values from clone counts before normalization, green color indicates values from normalized clone counts. Data derived from 12 50:50 mixture of P14 and OT-1 transgenic mice samples described in the *Transgenic TCR data sets.*

3.3.5. Amplification bias reduction benefits from the dependence of primer pairs

A question from data presented in **Figures 3.5 and 3.6** that arose had to do with determining how great a reduction in the spread of the 260 ST counts would be possible, given the variation, even in the absence of amplification bias. A theoretical analysis is presented in **Supp. Material 2** under the assumption that counts from equimolar concentrations of the 260 ST are *independent* NB distributions with the same mean and dispersion parameters gives a lower bound. Seeing that counts from the 20 ST-only samples were well approximated by NB distributions with the same dispersion parameters (albeit with quite different means), our conclusion was that the different ST counts were likely not independent. Further, this result

supports the normalization scheme, as the dependence aids reduction of the amplification bias below the level that would be expected under independence. Since each primer pair shares one primer with 32 other primer pairs, it is not surprising that the different ST counts are not independent; indeed, patterns of dependence in counts using chi-squared statistics are observed (**Figure 3-7**). The interactions revealed as patches of red and blue colors demonstrate that groups of V primers exhibit positive or negative dependence together with groups of J primers, that is, they interact to become over or under-represented in groups.

In their experimental setup, Carlson et al., using an ANOVA based approach, concluded that primer pairs can be treated independently, and employed primer iteration experiments to find the optimal primer mix(79). Based on our normalization approach, however, the non-random, non-zero interaction terms revealed by chi-squared statistics substantially complicate preparation of an optimal primer mix through primer iteration experiments, and indicate a limit on the extent to which amplification bias can be addressed experimentally. We note that Carlson et al. also employed a second, computational normalization step, which we believe is primarily due to this limit.

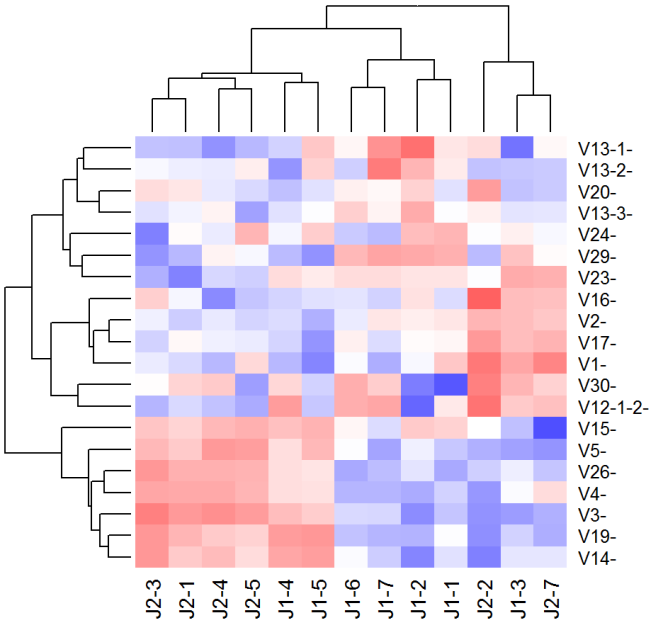


Figure 3-7. Cluster analysis revealing dependence between forward and reverse primers: A heatmap of interaction terms between forward and reverse primers reveal widespread and reproducible deviations from expected efficiencies under independence. Interaction terms were calculated as signed Pearson residuals (O-E)/√E where O is the observed ST count and E is the expected ST count under independence, calculated as E = row total x column total / grand total. Blue and red colors indicate positive and negative deviation from independence between forward and reverse primers respectively, and white indicates no deviation. Data derived from twenty ST-only samples described in the *ST-only data sets.*

3.3.6. A negative binomial model supports downstream analyses of T cell repertoire dynamics

With the TCR clonotype counts normalized, we wanted to determine if results from these analyses were concordant with results from a commercially available platform (Adaptive Biotechnologies) described by Carlson et al(78). To achieve this, we utilized gDNA generated from pancreatic ductal adenocarcinoma specimens described in Byrne *et al*(82) containing 17 samples previously sequenced by Adaptive Biotechnologies' platform. Aliquots of samples were sequenced and processed using the protocol described above. We utilized in-house software (see methods), tcR package, and VDJ tools to compute TCR repertoire metrics such as the diversity, clonality, and clonal distribution and refer to this pipeline as Open TCR Sequencing Protocol (OTSP). The 16 samples with enough DNA were sequenced in parallel and results were evaluated for concordance using Spearman correlation analyses for all combinations of: 1) Adaptive Biotech. platform sequences run by OTSP; 2) OTSP sequences run by OTSP; and 3) Adaptive Biotech. platform sequences run by Adaptive Biotech. **Figure 3-8** show concordance between results of OTSP sequences run by OTSP and Adaptive Biotech. sequences run by OTSP for Clonal index (r=0.7) and Shannon diversity index (r=0.8). The concordance for other combinations for Clonal Index and Shannon Diversity, as well as concordance between results of

OTSP sequences run by OTSP and Adaptive Biotech. platform sequences run by Adaptive Biotech. for the frequency of hyperexpanded clones are shown in **Supp. Figure 7 (A-E).** Overall, results from the OTSP pipeline demonstrates a strong association with results from the Adaptive Biotech. TCR sequencing platform ($p<0.001$ for all comparisons).

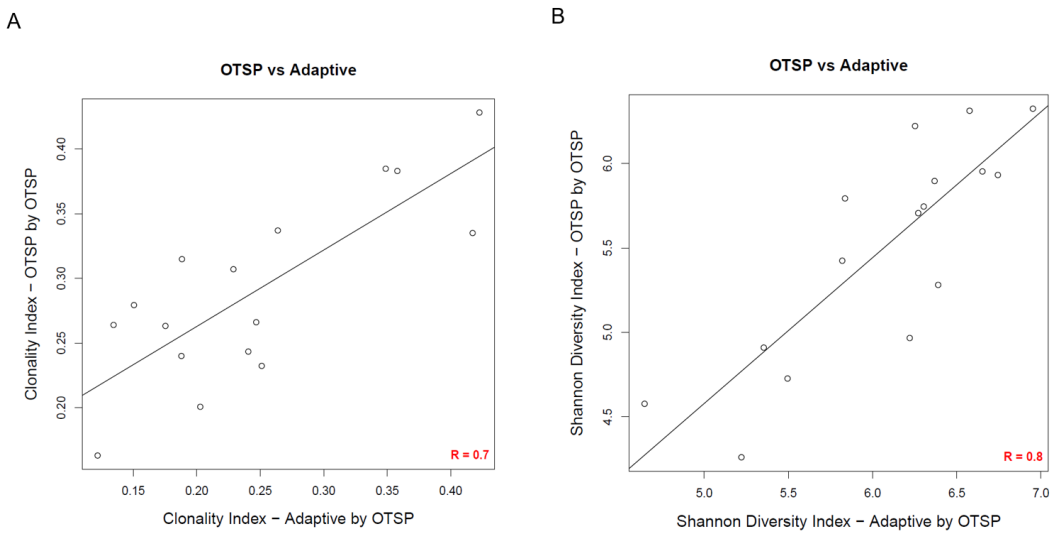


Figure 3-8. Concordance analysis of T cell repertoire metrics: Concordance analysis comparing commercial and in-house pipelines where samples were evaluated based on the results of OTSP sequences run by OTSP and the Adaptive Biotech platform sequences run by OTSP for Clonal Index (A) and Shannon Diversity Index (B). Spearman r=0.8 for Clonal Index and r=0.7 for Shannon Diversity Index. ($p<0.001$ for both comparisons). Data derived from samples described in the *Byrne et al. data set.*

The descriptions and formulas related to Clonal index, Shannon diversity index and frequency of hyperexpanded clones are included in (83), which deploys the NB mean normalization methodology described here in a biological context.

3.4. Discussion

Measuring and monitoring adaptive dynamics in patient TCR repertoires could have a significant impact on response and resistance monitoring for patients receiving various forms of

immunotherapy in the treatment of cancer or auto-immune diseases. To achieve the goal of capturing the diversity, quantifying the abundance of T-cell clones and performing longitudinal comparisons, TCR sequencing has emerged as an approach to monitor T cell responses to therapy and disease progression.

The OTSP pipeline described herein provides a transparent protocol enabling clonality metrics, including evaluating amplification bias specific to each primer pair, and is reproducible across samples. Count variability approximates the negative binomial distribution that can be exploited using estimated NB means. The NB distribution tells us the variation anticipated in ST counts is indicative of a controlled process.

OTSP pipeline results were compared to data generated from the commercial platform by fixing 260 ST-specific scaling factors derived from ST-only measurements (without the presence of genomic DNA). We observed a high concordance between bulk clonality metrics across the two platforms. This observation indicates that the OTSP pipeline can be integrated across batches, samples and platforms, further improving utility of TCR clonality measurements, something not generally possible when using commercial platforms as ST-counts are typically not provided.

OTSP is open and freely available; we anticipate this will allow scaling up the number of measurements substantially. Since we only use computational normalization, rather than addressing differing primer efficiencies at the bench level, the OTSP methodology is also less labor-intensive than previous methods. Notably, OTSP avoids the primer iteration experiments needed to address amplification bias problems. Porting of the platform to new organisms or designs will require new calibration. Utilizing the designs presented here for mouse are likely to require minimal optimization.

Since PCR amplification bias is repeatable across samples we have demonstrated the possibility of conducting analyses without the addition of ST beyond the initial calibration. This is achieved by using ST-specific normalization scale factors obtained from independent, ST-only measurements. The idea of using 260 universal ST-specific scaling factors to address amplification bias in mouse models could be explored further, for if deployed this would substantially decrease cost of this methodology as ST are costly. An additional advantage stems from the fact that ST reduce sequencing depth due to competition between genomic DNA and the ST (**Supp. Figure 8**). We observed that when gDNA amount was kept constant at 600 ng, increasing concentrations of ST led to decreasing detectability of clonotypes, indicating a competition between the ST and clonotypes during the process. This is especially important as indicated by our results revealing that drop-outs can be frequent, even for the most abundant clonotypes (**Supp. Figure 9**). For example, when the most frequent 0.3% clonotypes from Wild Type spleen tissue were used, we observed that only 119 distinct clonotypes were detected in the 5 replicates, with 63 clonotypes detected in 4 samples, 69 in 3, and 90 in 2, respectively. These drop-outs stem from the stochastic nature of the sampling and could be reduced by increasing the read coverage by not using ST.

3.5. Conclusions

Measuring and tracing changes in the abundance of clones is important for observing the immune response to different therapies such as cancer immunotherapy, leukemia monitoring and predicting patient outcomes.

A high concordance between bulk clonality metrics across Adaptive Biotech., one of the leading commercial companies in the field, and OTSP is observed. Our approach is a less laborious (as OTSP avoids the primer iteration experiments needed to address amplification bias

problems), reproducible and an inexpensive alternative for understanding relative abundances of clonotypes. Utilization of STs for every experiment in a batch beyond the initial ST calibration may not be necessary and their utilization in every batch decreases the sequencing depth.

We propose the OTSP as an open-source, transparent protocol that efficiently corrects for amplification bias post sequencing for accurate, reproducible measurement of clonality metrics.

3.6. Material and Methods

The aim of the methodology is modeling T Cell Clonotype repertoires using TCRβ CDR3 sequences utilizing NGS, multiplex PCR and an NB mean normalization strategy. Some of the experiments were performed in the context of genomic DNA derived from mice.

3.6.1. Mouse handling

To generate the Byrne et al data set, wild-type C57BL/6 mice were purchased from The Jackson Laboratory and housed at the University of Pennsylvania. Animal protocols were reviewed and approved by the Institute of Animal Care and Use Committee at the University of Pennsylvania. Mice were euthanized in a CO2 chamber using a flow meter to ensure CO2 was displaced at a rate of 30-70% of the chamber volume per minute and maintained for at least 1 minute after the loss of righting reflex is observed. Euthanasia was confirmed by bilateral thoracotomy. Animal handling information regarding tumor injections, drug prep and injection are included in Byrne et al(82).

To generate transgenic TCR data sets, spleen tissue from P14 and OT1 TCR transgenic mice (Nolz lab) were obtained. The spleen tissue was homogenized, treated with RBC lysis buffer and the resulting single-cell suspension was pelleted. Genomic DNA was extracted from the cell pellet using DNeasy blood and tissue kit (Qiagen).

To generate all the other data using animals, wild-type C57BL/6 mice were purchased from The Jackson Laboratory and maintained within the UCSF or OHSU laboratory for animal care barrier facility according to IACUC procedures. To generate the ST dilution series data set with mammary tumor and the Mesothelioma data set, we used mammary tumors from mouse mammary tumor virus (MMTV)- Polyomavirus middle T (PyMT) transgenic FVB/N mice and 40L orthotopic mesothelioma tumors respectively. Mammary tumors were resected from day 95 mice. For 40L orthotopic mesothelioma tumors, 2×10^6 cancer cells were injected i.p. into wild-type male C57BL/6 that were 6-12 weeks of age. For both tumor models, all mice were euthanized at a pre-defined end-stage for tissue harvest by cardiac puncture followed by cervical dislocation. These mice were cardiac perfused, under anesthesia using 1-5 % isoflurane, with 20 mL solution of heparin in PBS to clear tissues of residual blood followed by tissue harvest for further analysis including TCR sequencing. Tumor tissue was excised and flash frozen in liquid nitrogen and stored at -80 degree Celsius until further use for extracting genomic DNA for TCR sequencing. Murine SCC tumors were obtained from a previously published study, Medler et al(83).

3.6.2. Multiplex primers and design of synthetic TCR templates

Multiplex PCR primers previously described by Faham *et al.* (US patent 8,628,972 B2), for the amplification of murine *TCRβ* genomic loci were utilized **(Supp. Table 1. Primer Sequences)**. The 20 Vβ segment specific primers amplify all the 21 functional Vβ segments, and the 13 Jβ specific primers amplify all the 13 functional Jβ segments. As previously described by Carlson et al., we designed 260 (20V x 13J) synthetic TCR templates (ST) to minimize amplification bias due to multiplexing with 20 Vβ forward and 13 Jβ reverse primers(78). Briefly, ST are 200 bp long double stranded DNA segments that contain partial V segment and J segment

sequences encompassing a set of internal barcodes for post-sequencing identification. The internal barcode region contains a 16 bp barcode specific for each VJ combination. This specific barcode is further flanked by a 9 bp barcode that is common for all ST. An equimolar mixture of the 260 ST is added to the genomic DNA samples during PCR as internal controls.

3.6.3. Amplification and deep sequencing of TCRβ genomic locus

Genomic DNA from freshly resected mouse peripheral blood, peritoneal orthotopic mesotheliomas(84), and spleen was isolated using the Qiagen DNeasy Blood and tissue kit. Utilizing the method described in Robins et al(85), we performed a 2-stage PCR using genomic DNA for TCRβ deep-sequencing library preparation. The 1st stage involved amplifying the gDNA and the ST using 35 cycles of multiplex PCR with 20 Vβ forward and 13 Jβ reverse primers using the Qiagen multiplex PCR kit. The multiplex PCR primers contain a common 5' overhang, allowing amplification by a single primer pair in the 2nd stage PCR. Using 2.0% of the purified PCR product from stage 1 as template, a 2nd stage PCR, including 8 cycles, was performed with universal and indexed Illumina adaptors. Of note, the indexed adaptors contained an 8-base index sequence, providing each sample with a unique sample barcode. Equal volumes of all samples were pooled. Each pool concentration, typically containing PCR mixtures from 70 samples, was measured with a 2200 TapeStation (Agilent), and the concentration determined by real time PCR using a StepOne Real Time Workstation (ABI/Thermo) with a commercial library quantification kit (Kapa Biosystems). Paired-end sequencing was performed with a 2 x 150 protocol using a Midoutput 300 sequencing kit on a NextSeq 500 (Illumina). Target clustering was ~ 160 million clusters per run. Following the run, base call files were converted to fastq format and demultiplexed by a separate barcode read using the most current version of Bcl2Fastq software (Illumina).

3.6.4. TCR data analysis Pipeline

Fastq files were assessed for initial read quality using the FASTQC public tool(86), including the per-base quality scores. Quality paired-end sequences were combined using the PEAR (Paired-End reAd mergeR) algorithm(87). Merged sequences were then separated into ST and non-ST sequences. ST sequences were identified by searching for the common flanking 9-bp internal barcodes allowing a one-nucleotide mismatch or indel. Sequences flagged as ST via this search were removed from downstream clonotype analyses. The individual ST sequences were distinguished and quantified by searching for the specific 16-bp barcode sequences unique to each ST, again allowing a one-nucleotide mismatch or indel (**Supp. Figure 10**). Clonotypes were identified from purified (ST-removed) sequences utilizing the MiXCR pipeline(88), which is a two-step alignment and assembly process. First, reads were aligned to reference V, D, and J sequences, using the align module. Next, the assemble module grouped alignments into distinct clonotypes using a hierarchical clustering method based on sequence similarity and relative abundance. Finally, the export module exported alignments as well as assembled clones in tabular format. Raw clonotype counts were normalized using the NB mean normalization strategy described below. Normalized clonotype counts were exported in tabular format for use in downstream analysis. A number of TCR repertoire metrics, including clonality, maximum clonal frequency, and the Shannon diversity index were calculated. Quality control data was recorded in an overall summary table.

3.6.5. Data Sets

ST-only data sets: Twenty samples of an equimolar mixture of ST were sequenced. Two sets of ten samples of equimolar mixtures of ST at different concentrations were also sequenced at a later date. No genomic DNA was present in these samples.

Transgenic TCR data sets: A dilution series was created from spleen genomic DNA of P14 and OT-1 TCR transgenic mice. Three technical replicates of 300, 600, 900, and 1200 ng of DNA from P14, OT1, and a 50:50 mixture of P14 and OT1 DNA were sequenced along with an equimolar mixture of ST for a total of 36 samples. 4 mice were used to create the DNA from P14, another 4 mice were used to create the DNA from OT1 and 8 mice were used to create 50:50 mixture of P14 and OT1 DNA. In total, 16 mice were used to create this data set. Appropriate amplification of the transgenic clones was assessed. (**Supp. Figure 11**).

ST dilution series data set: A dilution series was created from ST at six levels as below:

* Stock: 3.2 ng/ul (equimolar mixture of all 260 spikes)

Dil1= 0.1 of stock

Dil2= 0.1 of Dil1

Dil3= 0.01 of Dil1

Dil4= 0.01 of Dil3

Dil5= 0.001 of Dil3

Dil6= 0.001 of Dil5

Three technical replicates of 600 ng of murine blood DNA were added to all levels of dilutions and for no ST samples for a total of 21 samples and three technical replicates of 600 ng of mammary tumor murine DNA were added to all levels of dilutions and for no ST samples for a total of 21 samples. In total, 14 mice were used to create this data set.

WT spleen data set: Wild-type mouse spleen genomic DNA were sequenced for a total of five technical replicates. A single mouse was used to create this data set.

Byrne et al. data set(82)*:* gDNA from 17 murine pancreatic ductal adenocarcinoma specimens were previously sequenced by a commercially available TCR beta platform. Aliquots of these gDNA samples were obtained and sequenced along with an equimolar mixture of ST using the protocol described above. In total, 17 murine samples were used to create this previously published data set.

Mesothelioma data set: Peritoneal mesothelioma tumors derived from the 40L cell line (84)derived genomic DNA samples derived from 12 syngeneic mice were sequenced along with equimolar mixture of ST. In total, 12 mice were used to create this data set.

In total, 60 mice were used to create above datasets to assess methodological approaches to normalizing TCR repertoires. No groups of animals were compared to each other as the purpose of the study has no biological study endpoint. Therefore, no *a priori* sample size calculations performed.

3.6.6. Batch mean scaling factors

This method creates a *scaling factor* for each ST (and therefore for each primer pair) based on that ST's counts among all samples, relative to all ST counts within a batch. Given a matrix (C_{ij}) of ST counts from a batch, where *i=1,…,260 labels* ST and *j=1,…,n labels* samples in the batch, we denote the batch mean of the counts for ST *i* by $C_{i\bullet} = n^{-1}\sum_j C_{ij}$ and the batch mean of all ST means by $C_{\bullet\bullet} = (260)^{-1}\sum_i C_{i\bullet} = (260n)^{-1}\sum_{i,j} C_{ij}$. The scaling factor (SF) for ST *i* in that batch is $SF_i = C_{i\bullet}/C_{\bullet\bullet}$.

3.6.7. Negative Binomial means

The above idea of a scale factor is distribution free, but for its use in normalizing counts, would require a full set of ST in every sample. We explored the use of a ST-specific negative

binomial model to dispense with the use of synthetic templates. Consider the set *(C_1, …., C_n)* of counts for single, fixed ST across a batch of *n* replicate ST-only samples. A plausible model for these counts is the negative binomial (NB) distribution. We write $C \sim NB(m,d)$ for this distribution, where *m>0* is the *mean* parameter and *d≥0* is the (over-) *dispersion* parameter, and refer to (89) for an explicit formula for the *NB* probability mass function. For present purposes, *S* has expected value $\boldsymbol{E}(C) = m$ and variance $\boldsymbol{var}(C) = m + dm^2$. When *d=0*, the negative binomial reduces to the Poisson distribution, for which $\boldsymbol{E}(C) = \boldsymbol{var}(C)$, and thus the use of the term over-dispersion here is relative to the Poisson. Using the methods of generalized linear models(89), we can obtain the maximum likelihood estimates (MLE) $\hat{m}$ and $\hat{d}$ of *m* and *d* from a replicate set of ST such as *(C_1, …., C_n)*. In the notation of the previous paragraph, the MLE $\hat{m}_i = C_{i\bullet}$, that is, the MLE of the *i*th mean parameter m_i of an NB fitted to *(C_{ij})* is the arithmetic mean of the *i*th set of observed counts (assumed independent and identically distributed across *j=1,…,n* with common mean m_i and common dispersion parameter. Where no confusion will result in what follows, we will not distinguish the parameters *(m_i)* from their (maximum likelihood) estimates $(\hat{m}_i)$. These ST or primer-pair-specific means estimated from ST-only data can be used as scaling factors for normalization, even when there are no ST present in samples. Consistent with the notation in the previous paragraph, we write $m_\bullet = (260)^{-1}\sum_i m_i$ for the average of the 260 (estimated) mean parameters. The 260 NB mean scaling factors are $(m_i/m_\bullet)$.

3.6.8. Normalization

To normalize a set of clonotype counts from a single sample, we first calculated the primer-pair totals *(C_i),* where C_i denotes the total count of all clonotypes amplified with primer-pair *i,* where *i=1,…,260.* We then normalize the 260 counts *(C_i)* using the batch scaling factors *(SF_i)* by dividing by the corresponding scaling factor: $C_i' = (SF_i)^{-1}C_i = (C_{\bullet\bullet}/C_{i\bullet})\,C_i$. Similarly, we normalize the *(C_i)* using the (estimated) NB mean scaling factors $(m_i/m_\bullet)$ by dividing: $C_i'' =$

$(m_{\bullet}/m_i)\ C_i$. After these primer-pair totals were normalized, the counts for distinct clonotypes sharing the same primer-pair were normalized: if one such clonotype accounts for a proportion *p* of the total count *C* corresponding to its primer pair, then it will be assigned a normalized value equal to the same proportion *p* of the normalized primer-pair total *C'* or *C"*.

The same normalization could be used for ST counts if available. That is, divide the observed count of the fragments arising from primer pair *i* by the SF_i or $m_i/m_{\bullet}$ for primer pair *i* for both ST and clonotype counts alike. Since the mean count of ST *i* will be proportional to m_i , normalization should preserve the total count of ST, exactly for batch mean normalization, on average for *NB* normalization. As long as the observed clonotype counts exhibit the same relative over- and under-representation after amplification as that exhibited by the ST, any bias will be reduced by this normalization.

3.7. Declarations

3.7.1. Ethics approval and consent to participate

All animal experiments were performed in compliance with the National Institutes of Health guidelines and were approved by the Institutional Animal Care and Use Committees (IACUC) of the University of California, San Francisco and Oregon Health & Science University.

3.7.2. Availability of data and materials

The source code for the analysis is available at https://github.com/burcudem/TCRSeqNormalization

The datasets used and/or analyzed during the study are available from the corresponding author on reasonable request.

3.7.3. Funding

LMC acknowledges support from the NIH/NCI (CA130980, CA155331, CA163123), a DOD BCRP Era of Hope Scholar Expansion Award (W81XWH-08-PRMRP-IIRA), the Susan G. Komen Foundation (KG110560), and the Brenden-Colson Center for Pancreatic Health. AM, RHV, and LMC acknowledge support from a Stand-Up-To-Cancer Lustgarten Foundation Pancreatic Cancer Convergence Dream Team Translational Research Grant (SU2C-AACR-DT14-14), as well as R01 CA217176 (to RHV), and the American Cancer Society (125403-PF-14-135-01-LIB) to KTB.

3.7.4. Authors' contributions

BG analyzed and interpreted all of the ST and clonotype count data described under "Data Sets", and built the statistical models and methods which were used to address the central amplification bias problem resulting from the utilization of TCR sequencing with multiplex PCR. WH processed raw sequencing fastq files and deployed MIXCR software to achieve ST and clonotype counts. DM performed lab experiments to generate *ST-only data sets, Transgenic TCR data sets, ST dilution series data set* and *WT spleen data set* samples described under "Data Sets". BZ contributed to our understanding of the count dynamics and the platforms by analyzing several datasets. PL supervised WH, and contributed to the initial set up of the fastq file conversion and deployment of software to achieve counts. SK generated the *Mesothelioma data set* samples. KTB and RHV generated and shared the *Byrne et al. data set*(82). AAM outlined the aim of the project and supervised BG, PL and WH. MM built the multi-disciplinary effort setup to achieve statistical methods and analytically reviewed the ms. PTS supervised and mentored BG and reviewed analyses and provided feedback. LMC provided the ultimate goal and resources, built collaborations across different labs and supervised DM. TPS advised BG, WH and BZ on

data processing, analysis, interpretation and modeling and devised the NB normalization. All authors read and approved the final manuscript.

3.7.5. Acknowledgements

The authors thank the Knight Cancer Institute (NCI P30 CA069533), Bioinformatics, and OHSU Massively Parallel Sequencing shared resources. We are grateful to members of the Coussens Lab, Spellman Lab and Speed Lab for critical discussions, to Meghan Lavoie for technical assistance, and Justin Tibbitts and Teresa Beechwood for research regulatory oversight and animal husbandry.

Discussion

4.1. Contribution of Statistical Modeling of AML Signature Study

Using supervised and unsupervised statistical modeling on methylation array data from the BeatAML cohort, we developed methylation signatures of AML patients. Our findings demonstrate that computational modeling of methylation impact of AML drivers can reveal novel pathways while also validating previously known associations, and enhancing AML risk-stratification.

In this study, we systematically annotated the genes involved in leukemogenesis to elucidate the methylation pathways, using statistical techniques. Additionally, we utilized topic modeling to identify methylation signatures of infrequent mutations, and improve methylation-based subtyping. To the best of our knowledge, this is the first study to systematically annotate the genes important in leukemogenesis and use topic modeling to enhance methylation-based subtyping.

Our study resulted in several key contributions:

(i) Supervised and unsupervised models reveal new methylation pathways of AML driver events and validate previously known associations.

(ii) Individual *DNMT3A* and *TET2* signatures are precise and robust—They yielded high AUROCs and testing error was consistent with training error across multiple training rounds. This performance, despite the complex genetic and epigenetic make-up of post-diagnosis AML samples and relatively small cohort size, was highly encouraging for future applications.

(iii) Unsupervised topic modeling factorizes covarying methylation changes and isolates methylation signatures caused by rare mutations.

(iv) Topic modeling reveals a group of mutations with similar downstream methylation impacts and mapped to adverse-risk class by ELN.

(v) Topic modeling uncovers methylation signatures of infrequent cytogenetic events, significantly improving methylation-based subtyping.

(vi) Our models can be leveraged to build predictive models for AML-risk.

(vii) Our models show that cytogenetic events, such as t(15;17) have widespread *trans* downstream methylation impacts.

Our study will be highly useful to the scientific community due to the significant interest in using methylation for various applications, including characterizing aberrant methylation in cancers, sub-classifying tumors, distinguishing the tissue of tumor origin, developing methylation-based early detection tools, as well as subtyping patients in drug response studies.

4.2. Contribution of Normalization of TCR Sequencing Study

We developed a non-commercial and inexpensive protocol for measuring and monitoring adaptive dynamics in TCR clonotype repertoire using genomic DNA-based bulk sequencing. Our results show that the concordance between bulk clonality metrics obtained from using the commercial kits and that developed herein is high. For the first time, an open, publicly available protocol to process and analyze raw sequencing data generated by genomic DNA-based bulk sequencing, which remains the most cost-effective method to profile TCRs, is described.

Our study has several key contributions:

(i) We describe the Open TCR Sequencing Protocol (OTSP) that efficiently corrects for amplification bias post sequencing.

(ii) The OTSP pipeline provides a transparent protocol enabling clonality metrics and is reproducible across samples.

(iii) A high concordance between bulk clonality metrics across a commercial platform and OTSP is observed.

Given extensive interest in measuring and monitoring adaptive dynamics in patient TCR repertoires and its potential significant impact on response and resistance monitoring for patients receiving various forms of immunotherapy in the treatment of cancer or auto-immune diseases, our work is strong utility to the scientific community.

4.3. Conclusions

In contemporary cancer biology research, statistical methods, bioinformatics, and machine learning have become indispensable tools. They help in the integration and interpretation of a wide range of high-dimensional data types, and aid in identifying the molecular characteristics of tumors, predicting patient outcomes, and developing personalized treatments. In the future, these techniques will continue to play a critical role in advancing cancer research and improving patient outcomes.

Future Applications

5.1. Potential Future Applications of Statistical Modeling of AML Signature Study

Aberrant epigenetic control has clinical implications in diagnostics, prognostics and therapy. Our models revealing the methylation impact of AML drivers can be used for methylation-based subtyping of patients for drug response studies as well as as a biomarker for early detection and risk-stratifying.

5.1.1. Using methylation signatures for subtyping for drug response studies

In recent years, there has been a strong interest in targeting particular mutations or epigenetic machinery for therapeutics.

Epigenetic alterations are reversible. This facet underscores the considerable potential of therapeutic strategies that aim to target and correct these alterations. Such interventions predominantly include drugs that specifically interact with the epigenetic machinery, often through enzymatic regulators. Enzymes, as catalytic proteins, offer a substantial scope for interaction with small molecule inhibitors due to their intricate structure and functional activity. Therefore, they present an attractive and viable target for drug design and development.

However, caution is necessary when administering these drugs systemically. Epigenetic mechanisms play a pivotal role in the functioning of all tissues, and as such, they have the potential to cause a wide range of side effects if not appropriately regulated. It is crucial to balance the therapeutic efficacy of these drugs against potential systemic complications. The

primary aim of developing these epigenetic drugs, ultimately, is to improve patient survival. Therefore, rigorous research is ongoing to develop therapeutic strategies that maximize this objective while minimizing adverse effects.

Another pivotal consideration in the successful design of targeted therapies involves the contemplation of the distinct functions of each epigenetic modifier in diverse cell types. Epigenetic modifiers may target different genes in different cell types. The target genes, in turn, determine whether a particular epigenetic modifier serves as a tumor suppressor or an oncogene in that particular cell type and state. For example, we can't universally say inhibiting Enhancer of Zeste Homolog 2 (*EZH2*) would be beneficial because inhibiting *EZH2* might be advantageous in treating some solid malignancies and lymphomas at certain stages(90) but might be a poor choice in treating myelodysplastic syndrome. Therefore, it is essential to carefully evaluate the context-dependent function of these epigenetic modifiers.

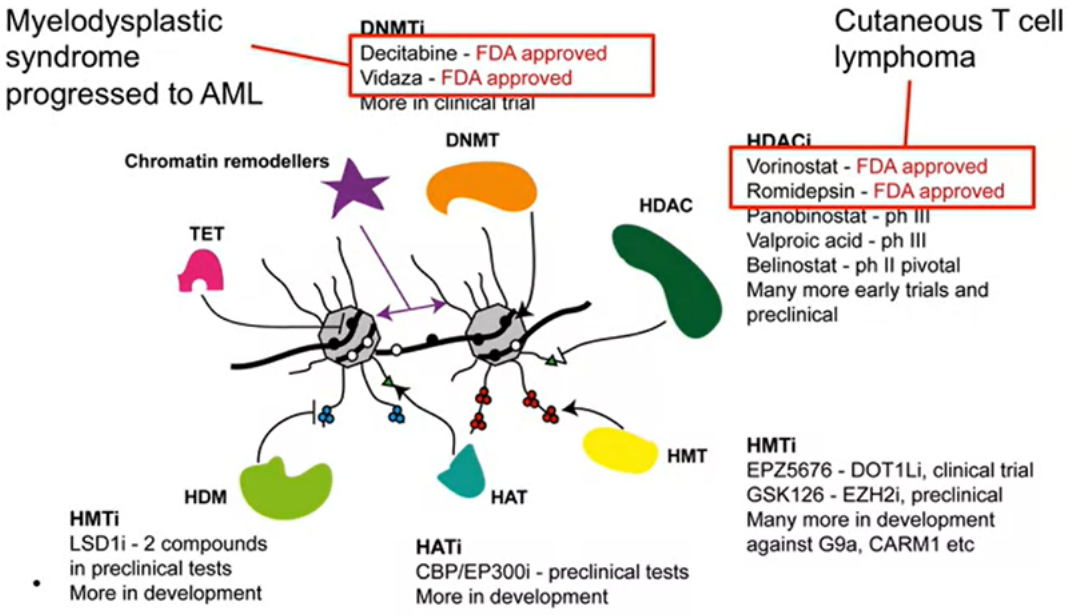


Figure 5-1. Epigenetic machinery and therapeutic agents: Epigenetic machinery along with therapeutic agents at various clinical trial stages and with FDA approval are shown. Adapted from Dr. Marnie Blewitt's Fall 2021 lecture notes from Epigenetic Control of Gene Expression class offered by the University of Melbourne.

Over recent years, pharmaceutical entities have been intensely mobilizing resources towards the exploration of epigenetic machinery, particularly with the utilization of small molecule inhibitors. **Figure 5-1** elucidates this machinery along with the current therapeutic agents at various clinical trial stages, extending to those that have received FDA approval. These agents, with the potential to target distinct components of the epigenetic machinery, encompass DNA methyltransferase inhibitors (DNMTi), histone deacetylase inhibitors (HDACi), histone methyltransferase inhibitors (HMTi), histone acetyltransferase inhibitors (HATi), and histone demethylase inhibitors (HDMi), all derived from publicly available information.

Focusing on DNA methyltransferase inhibitors, Decitabine and Vidaza, in particular, have procured approval for treating myelodysplastic syndrome that has progressed to AML. Vidaza and Decitabine are nucleoside analogs that bind irreversibly to DNMTs post-incorporation into the DNA, thereby rendering their action replication-dependent. Hence, cancer cells, due to their rapid replication, are more susceptible than normal cells. However, to optimize anti-neoplastic effects accompanied by DNA demethylation, the dosage of DNMTi should be cautiously maintained at lower levels to preclude nonspecific and toxic consequences. The efficacy of DNMTi is markedly observed in myelodysplastic syndrome, potentially due to its reliance on CpG island hypermethylation, a prognostic indicator linked to adverse outcomes.

In the light of our model-generated methylation signatures, we can investigate potential disruptions in DNMT function post-treatment. Furthermore, our topic models can be leveraged to categorize secondary AML and high-grade MDS patients based on their methylation profiles, aiding decision-making processes regarding patient selection, optimal dosing strategy, latency to optimal response, and therapy duration post-disease progression. Ongoing clinical trials involving combinations with conventional therapeutics or other epigenetically active agents, and in concert with bone marrow transplantation, continue to offer hope for the optimization of these agents for patients with myeloid disease. Despite controversies surrounding the mechanisms

responsible for the proven efficacy of these agents—whether they induce DNA hypomethylation, direct DNA damage, or possibly even immune modulation—it remains indisputable that they have secured their position in the therapeutic arsenal against myeloid neoplasms(91). Given that there are multiple epigenetic modulators in the drug development pipeline, it will be important to have computational means to match patients to correct epigenetic drugs or drug combinations – and our results from regression models and topics suggest a strong potential for using methylation profiles for guiding these decisions and subtyping patients.

5.1.2. Using methylation signatures as a biomarker

Relative to genomic profiles, the use of methylome profiles conveys two primary advantages that are particularly beneficial for liquid biopsy applications. Initially, a single genomic alteration has the potential to lead to myriad methylation changes, thereby launching substantial signal diversity provided the methylation alterations confer a selective advantage. Subsequently, a range of low-frequency genomic alterations that culminate in similar downstream impacts can be collectively categorized into methylation factors. These advantages are integral for applications involving early detection of cancer and disease monitoring as they heightened the ability to detect rare subclonal expansions associated with pre-malignant conditions.

The impact of a genetic mutation changes through tumor progression and methylation marks are not always mirrored by gene expression. Recent studies suggested that for assessing the impact of a mutation, threshold definitions of clonal hematopoiesis are not sufficient and we need to know clonal fitness or growth speed in addition to clone size(90,91). Therefore, directly monitoring the downstream methylation impact of epigenetic regulators, rather than the clonal size of the upstream mutation provides key benefits. It is especially important in AML as the most frequently mutated early drivers are epigenetic regulators. Our

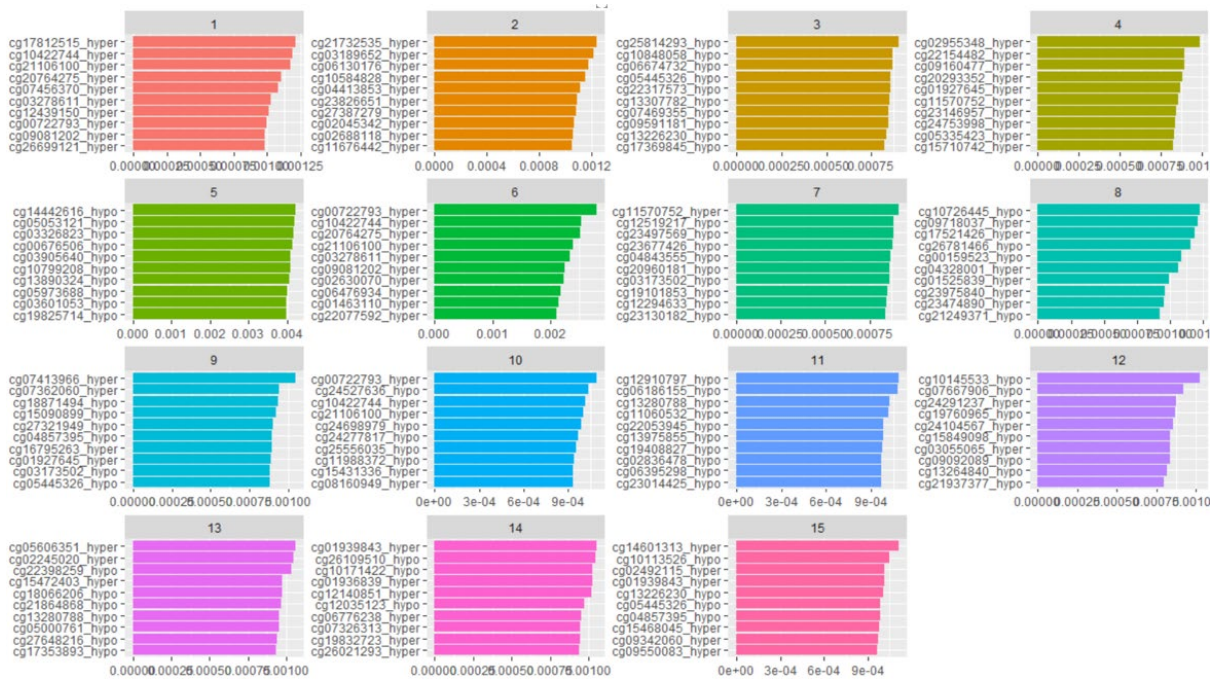
supervised models find methylation signatures of epigenetic regulators and they quantify the relationship between a mutation and its downstream methylation impact. Implementations of our regression models to early detection cohorts can reveal whether these regulator mutations maintain their importance in all stages of the disease and the certain shifts in epigenetic states leading to malignant transformation. Similarly, our models could be implemented to build time-to-AML models to uncover a predictive signature which is indicative of early AML before the usual time of diagnosis. This way we could determine which patients are prime candidates for frequent monitoring and early intervention. We could also use mutation specific signatures to do a risk stratification for each patient. For instance, if a person is carrying a *DNMT3A* mutation but his/her inferred methylation signature is similar to WT methylation patterns, it might imply relatively less AML risk or in the opposite scenario, it might signify an unprofiled mutation on the same pathway with *DNMT3A* leaving similar methylation signature and thus pose a similar risk to *DNMT3A* mutant tumors.

Our unsupervised topic models can be used to build a pan-AML signature using larger cohorts with matched normal. Historical cohorts with longitudinal blood samples, such as the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS(92)) can be analyzed using these topics to assess the clinical value of these signatures as early biomarkers of the disease. Additionally, we can learn about the progression trajectories of different subtypes by studying the methylation changes in those cohorts. By surveying the whole genome with bisulfite sequencing methods, more information about the variances between normal and disease states can be gathered. We can use the pan-AML signature identified by the matched models and perform risk stratification indicated by topics to develop a methylation based liquid biopsy assays for early detection purposes.

5.2. Potential Future Applications of Normalization of TCR Sequencing Study

Measuring T-Cell Clonotype Repertoires is vital for studying immune response dynamics. Our NB mean normalization methodology and the OTSP were deployed in studies with various biological context by Coussens Lab(83,93,94). Some other examples of clinical applications utilizing TCR sequencing are for monitoring the impact of treatments to immunomodulators(95,96) and monitoring Minimal Residual Disease (MRD) in T-ALL patients(78) to name a few. Recently, Adaptive Tech.'s T-detect has been approved for Covid-19 response monitoring(97). Considering the profound interest in assessing and tracking the adaptive dynamics within patient TCR repertoires undergoing diverse forms of immunotherapy for conditions such as cancer and autoimmune diseases, our research presents substantial utility to the scientific community by providing an open, reproducible and affordable protocol.

Appendix A: Chapter 2 Supplementary Information

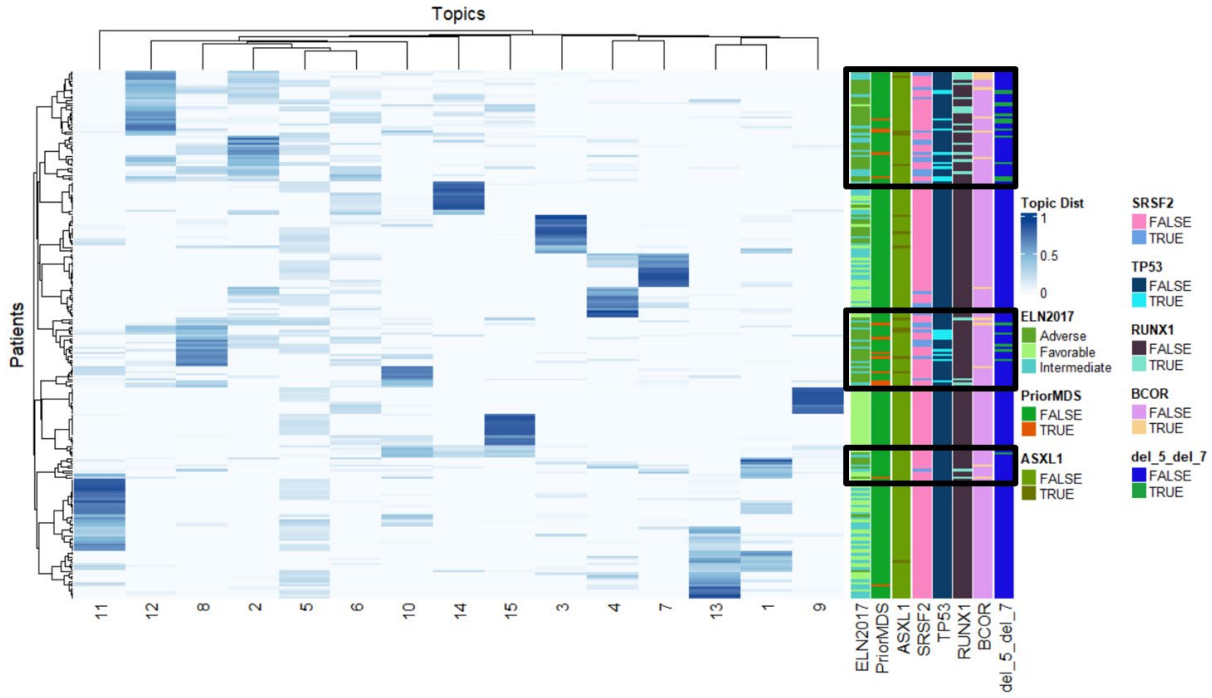


Supp. Figure 1. Top 10 most enriched loci for 15 topics

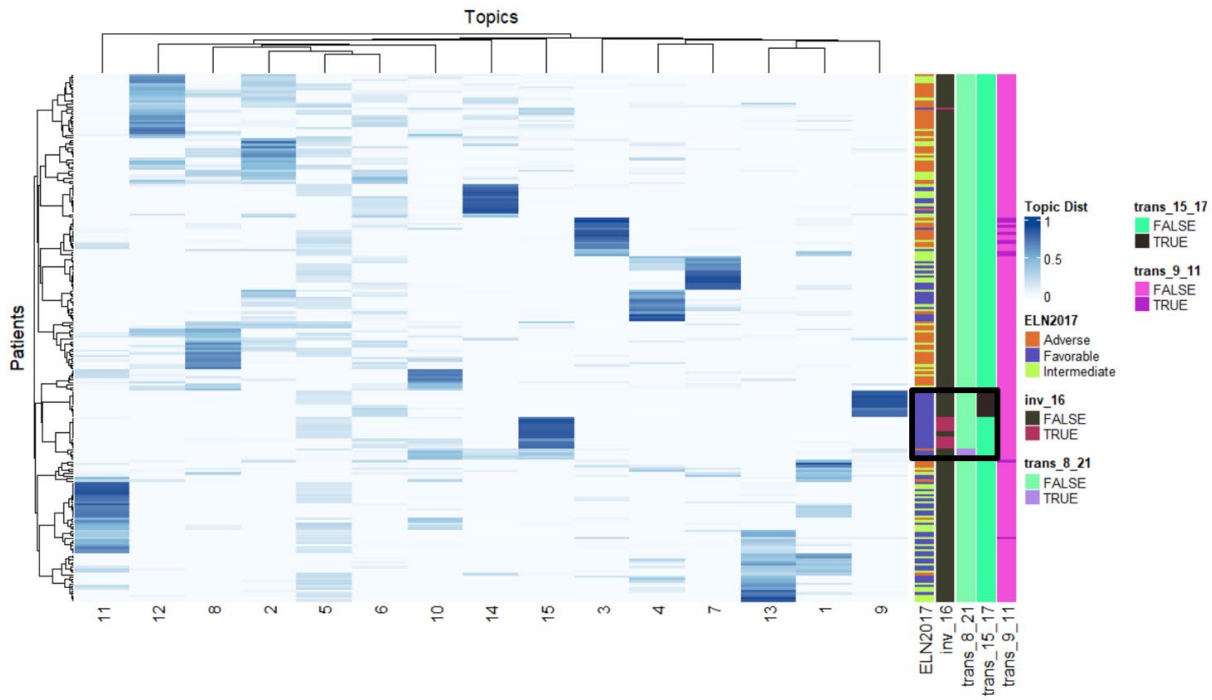
The top 10 probes with the highest assigned topic values and their methylation status (whether they are hyper or hypo methylated) are indicated for all 15 topics. We observe that topics 1, 5 and 6 map to gender and predictably have DMRs located on chromosomes X and Y. Topic 3 is exclusively associated with t(9;11). Notable enriched DMRs are located near: *MSI2*, a translocation partner with *HOXA9*, *EVI1*, *TTC40*, and *PAX5* in leukemias(98–101); *PPARG* – a key regulator for apoptosis and survival and *CNKSR3* – a gene that was identified as a commonly upregulated target in all *AF9/AF10* rearranged AMLs that includes t(9;11)(102). Topic 9, enriched for t(15; 17) driver events have DMRs near *JARID2*, a known tumor suppressor in AML(103), *LRPAP1* and *LPIN1*, regulators of lipid hematopoiesis implicated in AML progression(104). Topic 15, enriched for inv (16) driver events have DMRs around *HOXA9* and *MEIS1*, a frequently implicated pathway in AML progression. Another gene was *OPA1* that is known to be upregulated in AML and a mechanistic component for venetoclax resistance(105). Topic 11, enriched for co-mutations in *DNMT3A* and *NPM1* has a highly specific signature composed of multiple DMRs centered around *HOXB3* gene – another key homeobox family protein in AML progression(106). Topic 4 is strongly associated with *NPM1* and *IDH1/IDH2* mutations and interestingly include a DMR close to *CD34*, the definitive marker for hematopoietic stem/progenitor cells. Other AML associated genes selected in Topic

4 include *HDAC4*, a key epigenetic regulator in AML(107,108). Topic 13, enriched for *NPM1* and *WT1* mutations, also has DMRs near *HOXB3*. Other notable genes include *OSCAR*, a regulator of osteoclast differentiation(109) and *MIRLET7BHG*, an autophagy regulating lncRNA implicated as an AML survival marker(110). Topic 7, associated with co-occurring mutations in *NPM1* and *TET2,* has a DMR near *JARID2* a known tumor suppressor in AML(103), and *TBC1D8*, a known target of *HDAC2* in AML(111). Topics 2, 8 and 12 have shared components and is associated with ELN adverse category and a complex set of associated mutations in *RUNX1*, *BCOR*, *ASXL1*, *TP53*, *SRSF2* and 5q and 7q deletions. Topic 2 has multiple DMRs near *PTH2R*, parathyroid hormone receptor, which was shown to be the most upregulated gene in MDS and AML(112) and differentially expressed in patients with *IDH2* mutations(113). Topic 8 has DMRs in close proximity to *KDM2B*, a key lysine demethylase in AML; *BCL7A*, a *BAF* remodeling tumor suppressor(114), and *TCF7L2*, a *WNT* pathway transcription factor implicated in regeneration of hematopoietic stem cells(115). Topic 12 has DMRs near immunoglobulin heavy constant gamma as well as T-Cell Receptor Beta locus. Other notable genes include *WNK4*(116),

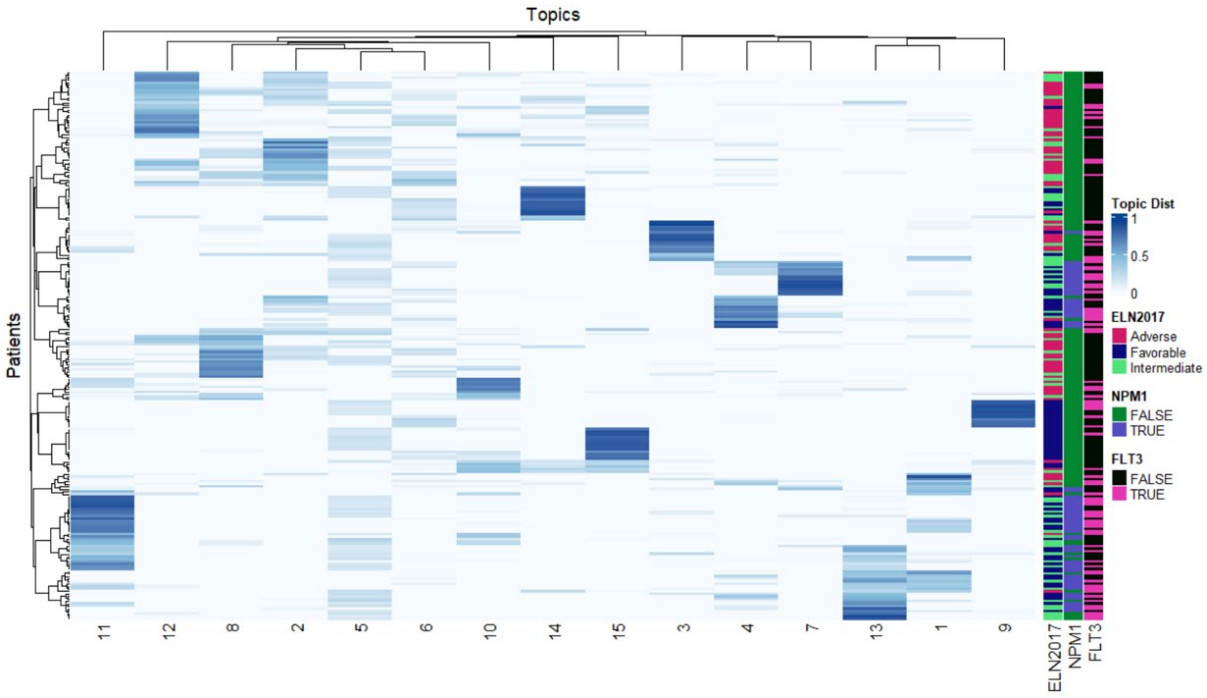
DYRK1(117) and *GIMAP7*(118) all indicated in stem cell-like signatures. Weak enrichments for *CEBPA* mutations and t(8;21) exist for topics 14 and 10 respectively.



Supp. Figure 2. Mutations and events associated with ELN intermediate to adverse risk category: *RUNX1*, *BCOR*, *ASXL1*, *TP53*, *SRSF2*, 5q and 7q deletions, having a prior MDS and ELN2017 adverse risk category associated topics are clustered together, shown in black rectangles.

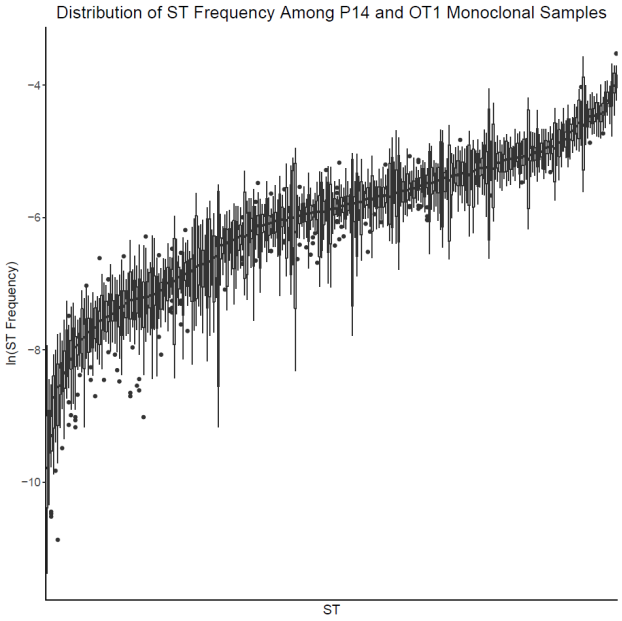


Supp. Figure 3. Cytogenetic events associated with ELN favorable risk category: t(15, 17), (inv 16) and t(8, 21) and ELN2017 favorable risk category associated topics are clustered together, shown in the black rectangle.

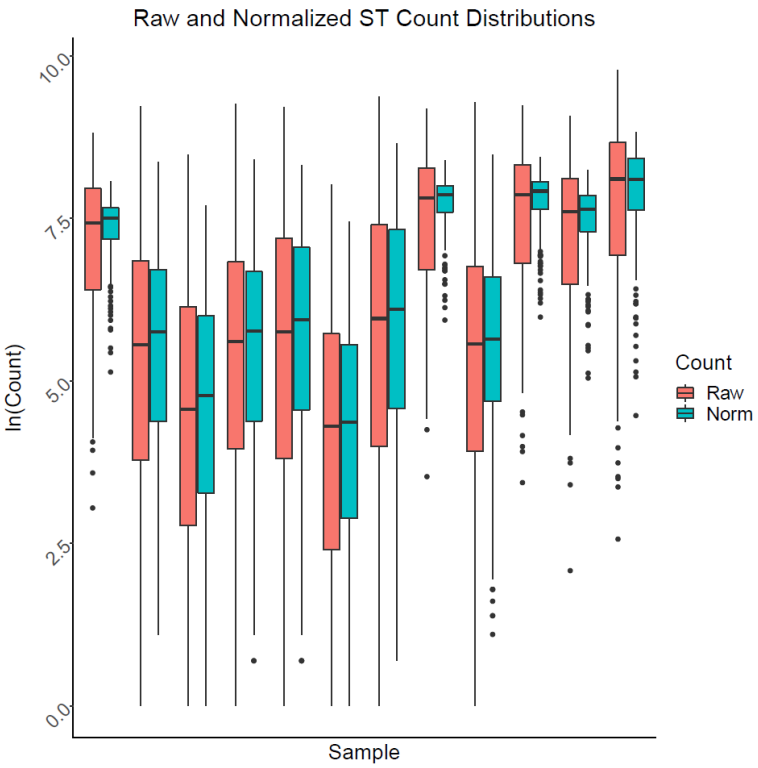


Supp. Figure 4. Methylation signature of *FLT3*: *FLT3* mutant samples (in pink) doesn't have a distinguisahble methylation signature.

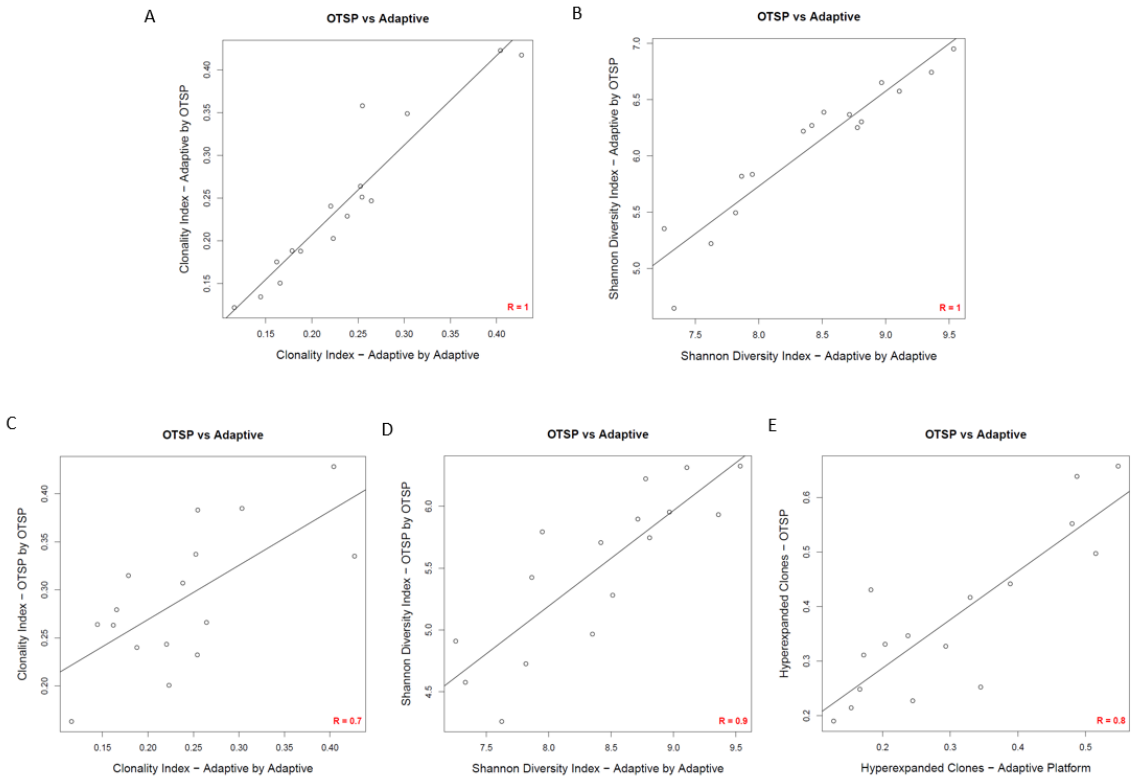
Appendix B: Chapter 3 Supplementary Information



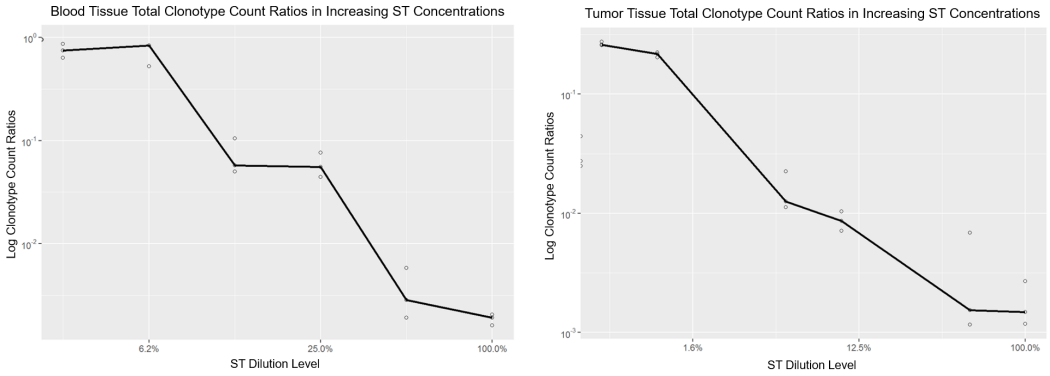
Supp. Figure 5. ST Count Distribution in presence of DNA: ST counts were obtained from samples described in the *Transgenic TCR data sets,* where 24 samples of gDNA from P14 and OT1 TCR transgenic mice were processed along with an equimolar mixture of ST. OT1 was amplified by the primer pair (V12-1,2, J2.7), and P14 was amplified by the primer pair (V13-3, J2-4). As with the ST-only samples, when TCR clonotypes were present in the samples along with the ST, the observed variation in the ST counts was caused by the amplification biases of the different primer pairs, rather than by sample to sample variation.



Supp. Figure 6. Normalization reduces spread in presence of DNA: gDNA from 12 murine mesothelioma specimens were amplified along with ST (described in the Data Sets section) where the reduction in ST count spread in the presence of DNA before (red) and after (cyan) normalization was plotted.

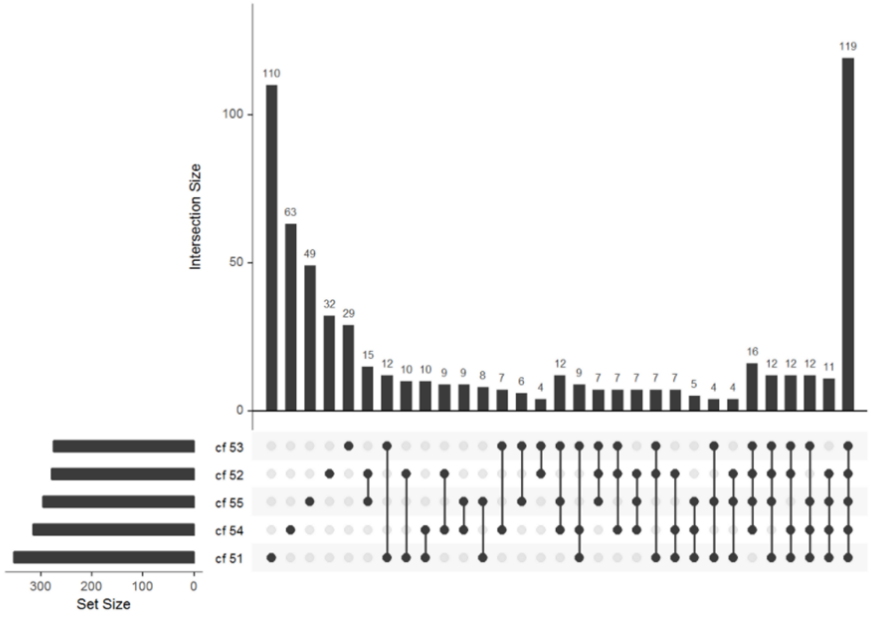


Supp. Figure 7. Concordance analysis of T cell repertoire metrics: Concordance analysis comparing commercial and in-house pipelines. gDNA from PDAC tumor samples were evaluated based on a commercial platform (Adaptive Biotech.) where sequences were compared based on output from Adaptive Biotechnology versus OTSP for Clonal index (R=1) and Shannon diversity index (R=1) **(A-B),** concordance between the Adaptive Biotech platform versus OTSP for Clonal index (R=0.7) and Shannon diversity index (R=0.9) **(C-D),** and concordance between the two pipelines for the frequency of hyperexpanded clones (R=0.8) **(E).** $p<0.001$ for all comparisons with Spearman correlation analysis. Data derived from samples described in the *Byrne et al. data set.*



Supp. Figure 8. Competition between gDNA and ST during TCR sequencing: Three replicates of 600 ng of gDNA isolated from peripheral blood leukocytes was added to all levels of dilutions (described in the Data Sets section) and for no ST samples for a total of 21 samples plus three replicates of 600 ng of mouse mesothelioma tumor DNA were added to all levels of dilutions and for no ST samples for a total of 21 samples. When the gDNA amount was kept constant at 600 ng, the increasing (relative) concentration of ST lead to decreasing detectability of clonotypes for both type of tissues, showing the competition between DNA and ST occurring during TCR sequencing.

The detectability of distinct clones in replicates for the most frequent top 0.3% clones

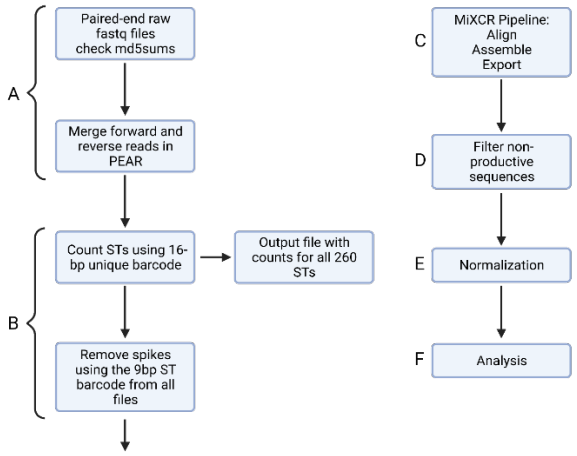


Supp. Figure 9. Reproducibility analysis: Drop-outs are frequent even for the top clones: The detectability of distinct clones in five replicates for the most frequent 0.3% clonotypes from wild type spleen tissue is shown. (Data derived from samples described in the *WT spleen data set*.) Vertical bars indicate the frequency of distinct clones detected in replicates, where the number on top of the vertical bars indicates the total number of distinct clones detected in the replicates. On the bottom left, the five replicates (samples cf51-cf55) and the number of distinct clones detected in each replicate is represented by horizontal bars with set size scale. The round dots are black if a particular clone was detected in the corresponding replicate shown at the very left. The connected black dots indicate how many and in which replicates distinct clones were detected out of five replicates. For example, going from right to left, only 119 distinct clonotype were detected in all 5 replicates, 63 clonotypes were detected in 4 samples, 69 in 3 and 90 in 2, respectively. These drop-outs come from the stochastic nature of the sampling.

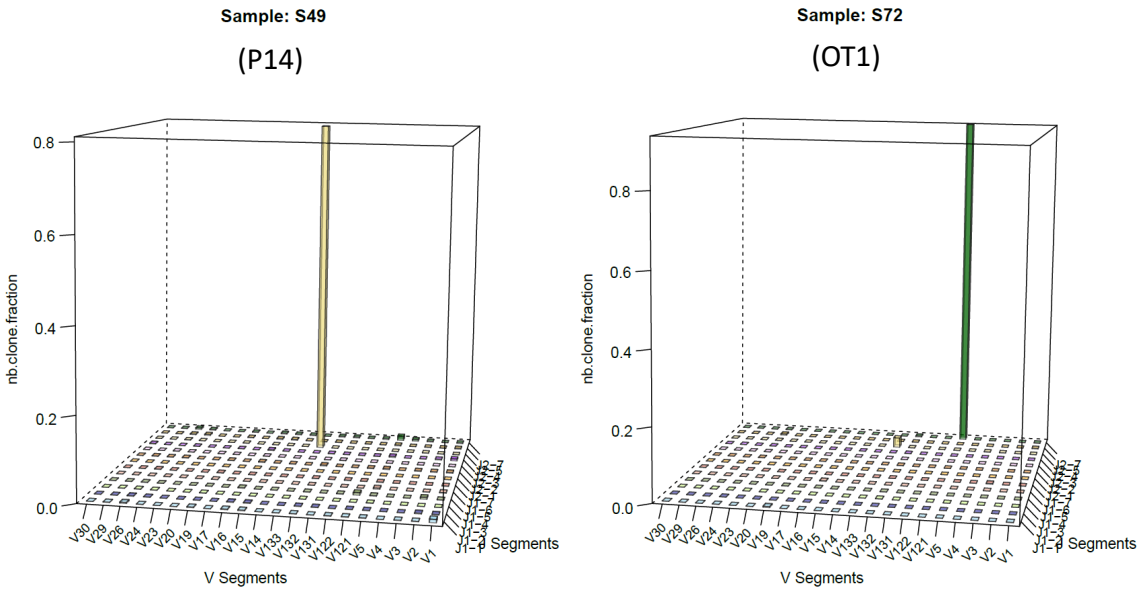
V segment	5' to 3'
V1	CAAAGAGGTCAAATCTCTTCCCG
V2	CTTATGGACAATCAGACTGCCTCA
V3	GTCATGGAGAAGTCTAAACTGTTTAAGG
V4	GTAAACGAAACAGTTCCAAGGCG
V5	GGTGCCCAGTCGTTTTATACCTGAAT
V 12-1, V 12-2	CCCAGCAGATTCTCAGTCCAACAGT
V 13-2	AGATATCCCTGATGGATACAAGGC
V 13-3	AGATATCCCTGATGGGTACAAGGC
V 13-1	AGATGTCCCTGATGGGTACAAGGC
V 14	GATAATTCACAGTTGCCCTCGGAT
V 15	GATGGTGGGGCTTTCAAGGATC
V 16	CAAGCTCCTATAGATGATTCAGGG
V 17	CTATGATAAGATTTTGAACAGGGAAGC
V 19	GATCTACTATTCAATAACTGAAAACGATCTTC
V 20	TAGCACTTTCTACTGTGAACTCAGCA
V 23	CTTGATCAAATAGACATGGTCAAGG
V 24	AGAGATTCTCAGCTAAGTGTTCCTCG
V 26	GTTCTTCAGCAAATAGACATGACTG
V 29	AGCGAAGGAGACATCCCTAAAGGAT
V 30	CGAGAGTGGATTCACCAAGGACAAG

J segment	5' to 3'
J1-1	ACTGTGAGTCTGGTTCCTTTACC
J1-2	AAGGCCTGGTCCCTGAGCCGAAG
J1-3	CTTCCTTCTCCAAAATAGAGC
J1-4	GACAGCTTGGTTCCATGACCG
J1-5	GAGTCCCCTCTCCAAAAAGCG
J1-6	TCACAGTGAGCCGGGTGCCTGC
J1-7	ATACCTAAGTTCCTTTCCAAGACC
J2-1	GTGAGTCGTGTTCCTGGTCCGAAG
J2-2	CCAGCACTGTCAGCTTTGAGC
J2-3	GTTCCTGAGCCAAAATACAGCG
J2-4	GTGCCCGCACCAAAGTACAAG
J2-5	GTGCCTGGCCCAAAGTACTGG
J2-7	CTAAAACCGTGAGCCTGGTGC

Supp. Table A and B:. Forward Primer Sequences



Supp. Figure 10. TCR sequencing pipeline schema: Depiction of TCR sequencing pipeline constructed from both extant software tools (configured for use within the pipeline), and dedicated programs written in-house. Multiple samples are processed in parallel, and quality-control checks provide visibility into the pipeline's operation. Computation is performed on the 5000 cores using ExaCloud computing cluster. Steps in the pipeline include: **A**. Verification of file integrity and merging of paired-end reads; **B**. ST reads are then identified, quantified, and removed; **C**. Clonotypes are then aligned to reference segments, clustered, and quantified; **D**. Clonotypes containing frameshifts and stop codons flagged, and output converted for use by visualization and analysis software **E**. Clonotype frequencies are then adjusted to account for PCR amplification; **F**. Analytic metrics computed (diversity, clonal expansion and other as applicable) using various tools indicated within the text. Figure created with BioRender.com



Supp. Figure 11. Monoclonal amplification check: Appropriate amplification and identification of clonal TCR segments was verified using OT1 and P14 monoclonal samples, where OT1 was amplified by the primer pair (V12-1,2, J2.7), and P14 amplified by the primer pair (V13-3, J2-4). One example of each monoclonal sample for appropriate amplification is shown.

Supp. Material 1

Suppose that we have three equimolar spike-in only datasets with possibly different concentrations consisting of 20, 10 and 10 replicate libraries, respectively. Denote their ST counts by $\{C_{ij}^{(k)}: i = 1, \ldots 260, j = 1, \ldots, n^{(k)}\}$, where $i = 1, \ldots 260$ labels ST, $j = 1, \ldots, n^{(k)}$ labels replicate libraries within datasets, $k = 1, 2, 3$, and $n^{(1)} = 20$, $n^{(2)} = n^{(3)} = 10$.

Our basic assumption is that for any given ST (i.e. primer-pair) the expected values of the counts for are essentially the same, i.e. that we have

$$\boldsymbol{E}\left(C_{ij}^{(k)}\right) = c^{(k)} m_i, i = 1, \ldots, 260, j = 1, \ldots, n^{(k)}, k = 1,2,3,$$

up to the concentrations $c^{(1)}, c^{(2)}$, and $c^{(3)}$. Within dataset *k*, natural unbiased estimates of the $c^{(k)} m_i$ are the averages $C_{i\bullet}^{(k)} = \left(n^{(k)}\right)^{-1} C_{i+}^{(k)}$, where $C_{i+}^{(k)} = \sum_{j=1}^{n^{(k)}} C_{ij}^{(k)}$.

These are maximum likelihood estimates (MLE) under the assumption that all the counts are mutually independent Poisson or Negative Binomial random variables with a common overdispersion parameter for each ST. Our goal here is to show how to combine the three estimates of m_i for any given i taking into account the possibly different concentrations. Without loss of generality we can take $c^{(1)} = 1$, and we will write $c^{(2)} = c$ and $c^{(3)} = d$. Here are two approaches to combining the estimates.

Assuming independent Poisson or Negative Binomial distributions. In this case it is a straightforward calculation to show that the MLE of μ_i based on all the counts is

$$\widehat{m}_i = (n^*)^{-1} C_{i+}^{(+)} \text{ where } n^* = 20 + 10\frac{C_{++}^{(2)}}{C_{++}^{(1)}} + 10\frac{C_{++}^{(3)}}{C_{++}^{(1)}}.$$

This makes sense. We sum *all* the counts observed for ST i and divide that by the sum of the effective number of replicates in each dataset, relative to the concentration for dataset 1.

Avoiding strong independence and distributional assumptions. Here we begin by noting that $log\boldsymbol{E}\left(C_{i\bullet}^{(k)}\right) = log\ c^{(k)} + log\ m_i,\ i = 1, \ldots, 260, k = 1,2,3$ and make our goal the linear combination of the three approximately unbiased estimates of $\mu_i = log\ m_i$, namely the quantities $l_i^{(k)} = logC_{i\bullet}^{(k)}, k = 1,2,3$, correcting for the two offsets $\gamma = log\ c$ and $\delta = log\ d$ of the second and third datasets relative to the first, and taking into account the fact that the first dataset has twice their number of observations. A straightforward weighted least squares estimation process leads to the combined estimate of μ_i as

$$\tilde{\mu}_i = \frac{1}{4}[2l_i^{(1)} + \left(l_i^{(2)} - \tilde{\gamma}\right) + (l_i^{(3)} - \tilde{\delta})]$$

where $\tilde{\gamma} = \frac{1}{260}(l_+^{(2)} - l_{i+}^{(1)})$ and $\tilde{\delta} = \frac{1}{260}(l_i^{(3)} - l_i^{(1)})$. Once we have a combined estimate of $\mu_i = log\ m_i$, we antilog to obtain our estimate $\tilde{m}_i$ of m_i.

Although the individual ST counts were plausibly negative binomial, they seemed far from independent. As a result, we used the second method to combine the three sets of estimated count means. Recall that in practice, all we need are the estimates of ratios $m_i/m_\bullet$, so that the concentration terms cancel.

Supp. Material 2

Let C_i be the count for spike (primer-pair) *i, i=1,…,n=260.* Our basic assumption is that the $\{C_i\}$ are mutually independent, with $C_i \sim NB(m_i, d_i)$. Thus $\boldsymbol{E}(C_i) = m_i$ and $\boldsymbol{V}ariance(C_i) = \boldsymbol{V}(C_i) = m_i + d_i m_i^2$. For the moment, we assume that the parameters $\{m_i, d_i : i = 1, \dots, n\}$ are all known, with $m_\bullet$ and $d_\bullet$ being the average of the $\{m_i\}$ and $\{d_i\}$ respectively.

The normalized counts are $N_i = \left(\frac{m_\bullet}{m_i}\right) C_i$. Clearly $\boldsymbol{E}(N_i) = m_i$ for all *i,* i.e. the normalized values have the same expected value as the original counts.

How variable are they?

Their average is $N_\bullet$ and their empirical variance is $s^2 = (n-1)^{-1} \sum (N_i - N_\bullet)^2$.

We give a lower bound to the expected value of s^2, which implies that we cannot use normalization to produce values that are guaranteed to be arbitrarily close together.

Assertion. $\boldsymbol{E}(N_\bullet) = m_\bullet$, $\boldsymbol{E}(s^2) \geq m_\bullet + d_\bullet m_\bullet^2$.

Note. The last expression is the variance of an $NB(m_\bullet, d_\bullet)$.

Proof. The equality $\boldsymbol{E}(N_\bullet) = m_\bullet$ follows by averaging both sides of $\boldsymbol{E}(N_i) = m_i$.

Now $\boldsymbol{V}(N_i) = \left(\frac{m_\bullet}{m_i}\right)^2 \boldsymbol{V}(C_i) = \frac{m_\bullet^2}{m_i} + d_i m_\bullet^2$, while

$\boldsymbol{V}(N_\bullet) = n^{-2} \sum V(N_i) = n^{-2} \sum \left(\frac{m_\bullet^2}{m_i} + d_i m_\bullet^2\right) = n^{-1} m_\bullet^2 \left\{ n^{-1} \sum \left(\frac{1}{m_i}\right) + d_\bullet \right\}$.

The *harmonic mean* of the $\{m_i\}$ is $H = n / \sum \left(\frac{1}{m_i}\right)$, and so $H^{-1} = n^{-1} \sum \left(\frac{1}{m_i}\right)$, and we can write $n\boldsymbol{V}(N_\bullet) = m_\bullet^2 (H^{-1} + d_\bullet)$. .

We now expand $\sum (N_i - N_\bullet)^2$ in a familiar way as

$\sum (N_i - N_\bullet)^2 = \sum ((N_i - m_\bullet) - (N_\bullet - m_\bullet))^2 = \sum (N_i - m_\bullet)^2 - n(N_\bullet - m_\bullet)^2$

as the cross term vanishes. Taking $\boldsymbol{E}$ of both sides, we get

$$\boldsymbol{E}\sum(N_i - N_\bullet)^2 = \sum \boldsymbol{V}(N_i) - n\boldsymbol{V}(N_\bullet)\,.$$

The rest is algebra. The right-hand side above is

$$\sum \boldsymbol{V}(N_i) - n\boldsymbol{V}(N_\bullet) = \sum(\frac{m_\bullet^2}{m_i} + d_i m_\bullet^2) - m_\bullet^2(H^{-1} + d_\bullet) = m_\bullet^2(n-1)(H^{-1} + d_\bullet).$$

Hence $\boldsymbol{E}\{(n-1)^{-1}\sum(N_i - N_\bullet)^2\} = m_\bullet^2(H^{-1} + d_\bullet) \geq m_\bullet + d_\bullet m_\bullet^2$, since $m_\bullet \geq H$,

with equality if and only if the $\{m_i\}$ are all equal.

References

1. Deininger MWN, Tyner JW, Solary E. Turning the tide in myelodysplastic/myeloproliferative neoplasms. Nat Rev Cancer. 2017 Jul;17(7):425–40.

2. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015 Sep 25;349(6255):1483–9.

3. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. N Engl J Med. 2014 Dec 25;371(26):2488–98.

4. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. bioRxiv. 2018 Sep 12;161562.

5. Mazor T, Pankov A, Song JS, Costello JF. Intratumoral Heterogeneity of the Epigenome. Cancer Cell. 2016 Apr 11;29(4):440–51.

6. Miles LA, Bowman RL, Merlinsky TR, Csete IS, Ooi AT, Durruthy-Durruthy R, et al. Single-cell mutation analysis of clonal evolution in myeloid malignancies. Nature. 2020 Nov;587(7834):477–82.

7. Morita K, Wang F, Jahn K, Hu T, Tanaka T, Sasaki Y, et al. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. Nature Communications. 2020 Oct 21;11(1):5327.

8. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. Science [Internet]. 2019 Nov 1 [cited 2021 Jan 1];366(6465). Available from: https://science.sciencemag.org/content/366/6465/eaan4673

9. Steensma DP, Bejar R, Jaiswal S, Lindsley RC, Sekeres MA, Hasserjian RP, et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood. 2015 Jul 2;126(1):9–16.

10. Papaemmanuil E, Gerstung M, Bullinger L, Gaidzik VI, Paschka P, Roberts ND, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia [Internet]. https://doi.org/10.1056/NEJMoa1516192. Massachusetts Medical Society; 2016 [cited 2021 Apr 6]. Available from: https://www.nejm.org/doi/10.1056/NEJMoa1516192

11. Li S, Chen X, Wang J, Meydan C, Glass JL, Shih AH, et al. Somatic Mutations Drive Specific, but Reversible, Epigenetic Heterogeneity States in AML. Cancer Discov. 2020 Dec 1;10(12):1934–49.

12. Link DC, Walter MJ. 'CHIP'ping away at clonal hematopoiesis. Leukemia. 2016 Aug;30(8):1633–5.

13. Survivorship in AML – a landmark analysis on the outcomes of acute myelogenous leukemia patients after maintaining complete remission for at least 3 years: Leukemia & Lymphoma: Vol 61, No 13 [Internet]. [cited 2021 Apr 3]. Available from: https://www.tandfonline.com/doi/abs/10.1080/10428194.2020.1802450

14. Docherty LE, Rezwan FI, Poole RL, Jagoe H, Lake H, Lockett GA, et al. Genome-wide DNA methylation analysis of patients with imprinting disorders identifies differentially methylated regions associated with novel candidate imprinted genes. J Med Genet. 2014 Apr;51(4):229–38.

15. Eric Tang MH, Varadan V, Kamalakaran S, Zhang MQ, Dimitrova N, Hicks J. Major Chromosomal Breakpoint Intervals in Breast Cancer Co-Localize with Differentially Methylated Regions. Front Oncol. 2012 Dec 27;2:197.

16. Vosberg S, Kerbs P, Jurinovic V, Metzeler KH, Amler S, Sauerland C, et al. DNA Methylation Profiling of AML Reveals Epigenetic Subgroups with Distinct Clinical Outcome. Blood. 2019 Nov 13;134(Supplement_1):2715–2715.

17. Mao SQ, Cuesta SM, Tannahill D, Balasubramanian S. Genome-wide DNA Methylation Signatures Are Determined by DNMT3A/B Sequence Preferences. Biochemistry. 2020 Jul 14;59(27):2541–50.

18. Szyf M. DNA methylation signatures for breast cancer classification and prognosis. Genome Medicine. 2012 Mar 30;4(3):26.

19. Cabezón M, Malinverni R, Bargay J, Xicoy B, Marcé S, Garrido A, et al. Different methylation signatures at diagnosis in patients with high-risk myelodysplastic syndromes and secondary acute myeloid leukemia predict azacitidine response and longer survival. Clinical Epigenetics. 2021 Jan 14;13(1):9.

20. Tyner JW, Tognon CE, Bottomly D, Wilmot B, Kurtz SE, Savage SL, et al. Functional genomic landscape of acute myeloid leukaemia. Nature. 2018 Oct;562(7728):526–31.

21. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. SpringerPlus. 2016 Sep 20;5(1):1608.

22. Valle F, Osella M, Caselle M. A Topic Modeling Analysis of TCGA Breast and Lung Cancer Transcriptomic Data. Cancers. 2020 Dec;12(12):3799.

23. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood. 2017 Jan 26;129(4):424–47.

24. Giacopelli B, Wang M, Cleary A, Wu YZ, Schultz AR, Schmutz M, et al. DNA methylation epitypes highlight underlying developmental and disease pathways in acute myeloid leukemia. Genome Res. 2021 Mar 11;gr.269233.120.

25. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013 Jan 15;29(2):189–96.

26. Friedman J, Hastie T, Tibshirani R, Narasimhan B, Tay K, Simon N, et al. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models [Internet]. 2022 [cited 2023 Mar 22]. Available from: https://CRAN.R-project.org/package=glmnet

27. Sing T, Sander O, Beerenwinkel N, Lengauer T, Unterthiner T, Ernst FGM. ROCR: Visualizing the Performance of Scoring Classifiers [Internet]. 2020 [cited 2023 Mar 22]. Available from: https://CRAN.R-project.org/package=ROCR

28. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003 Mar 1;3(null):993–1022.

29. Grün B, Hornik K, CTM) DMB (VEM estimation of L and, CTM) JDL (VEM estimation of, LDA) XHP (MCMC estimation of, RNG) MM (Mersenne T, et al. topicmodels: Topic Models [Internet]. 2022 [cited 2023 Mar 22]. Available from: https://CRAN.R-project.org/package=topicmodels

30. Asmar F, Punj V, Christensen J, Pedersen MT, Pedersen A, Nielsen AB, et al. Genome-wide profiling identifies a DNA methylation signature that associates with TET2 mutations in diffuse large B-cell lymphoma. Haematologica. 2013 Dec;98(12):1912–20.

31. Shih AH, Abdel-Wahab O, Patel JP, Levine RL. The role of mutations in epigenetic regulators in myeloid malignancies. Nat Rev Cancer. 2012 Sep;12(9):599–612.

32. Paredes R, Kelly JR, Geary B, Almarzouq B, Schneider M, Pearson S, et al. EVI1 phosphorylation at S436 regulates interactions with CtBP1 and DNMT3A and promotes self-renewal. Cell Death Dis. 2020 Oct 20;11(10):1–14.

33. Nam AS, Dusaj N, Izzo F, Murali R, Myers RM, Mouhieddine TH, et al. Single-cell multi-omics of human clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation. Nat Genet. 2022 Oct;54(10):1514–26.

34. Rajan A, Tonk S, Bose C, Singh S, Swarup S, Reddy T, et al. Chromosomal alterations of pediatric malignancy in a West Texas population. The Southwest Respiratory and Critical Care Chronicles. 2020 Feb 9;8(33):7–28.

35. Churpek JE, Pyrtel K, Kanchi KL, Shao J, Koboldt D, Miller CA, et al. Genomic analysis of germ line and somatic variants in familial myelodysplasia/acute myeloid leukemia. Blood. 2015 Nov 26;126(22):2484–90.

36. Trompouki E, Piragyte I, Clapes T, Klein-Geltnik R, Yin N, Polyzou A, et al. A Metabolic Interplay Coordinated By Hlx Balances Hematopoietic Stem Cell Maintenance and Differentiation. Blood. 2017 Dec 8;130:3771.

37. Costa AM, Pinto F, Martinho O, Oliveira MJ, Jordan P, Reis RM. Silencing of WNK2 is associated with upregulation of MMP2 and JNK in gliomas. Oncotarget. 2014 Dec 22;6(3):1422–34.

38. Nishiu M, Yanagawa R, Nakatsuka S ichi, Yao M, Tsunoda T, Nakamura Y, et al. Microarray Analysis of Gene-expression Profiles in Diffuse Large B-cell Lymphoma: Identification of Genes Related to Disease Progression. Japanese Journal of Cancer Research. 2002;93(8):894–901.

39. Finalet Ferreiro J, Rouhigharabaei L, Urbankova H, van der Krogt JA, Michaux L, Shetty S, et al. Integrative genomic and transcriptomic analysis identified candidate genes implicated in the pathogenesis of hepatosplenic T-cell lymphoma. PLoS One. 2014;9(7):e102977.

40. Lindblad O, Chougule RA, Moharram SA, Kabir NN, Sun J, Kazi JU, et al. The role of HOXB2 and HOXB3 in acute myeloid leukemia. Biochem Biophys Res Commun. 2015 Nov 27;467(4):742–7.

41. de Morrée A, Flix B, Bagaric I, Wang J, van den Boogaard M, Grand Moursel L, et al. Dysferlin regulates cell adhesion in human monocytes. J Biol Chem. 2013 May 17;288(20):14147–57.

42. Zhang X, He D, Xiang Y, Wang C, Liang B, Li B, et al. DYSF promotes monocyte activation in atherosclerotic cardiovascular disease as a DNA methylation-driven gene. Translational Research. 2022 Sep 1;247:19–38.

43. Puda A, Milosevic JD, Berg T, Klampfl T, Harutyunyan AS, Gisslinger B, et al. Frequent deletions of JARID2 in leukemic transformation of chronic myeloid malignancies. Am J Hematol. 2012 Mar;87(3):245–50.

44. Kratz CP, Emerling BM, Bonifas J, Wang W, Green ED, Beau MML, et al. Genomic structure of the PIK3CG gene on chromosome band 7q22 and evaluation as a candidate myeloid tumor suppressor. Blood. 2002 Jan 1;99(1):372–4.

45. Mack EKM, Marquardt A, Langer D, Ross P, Ultsch A, Kiehl MG, et al. Comprehensive genetic diagnosis of acute myeloid leukemia by next-generation sequencing. Haematologica. 2019 Feb;104(2):277–87.

46. Karunanithi S, Liu R, Hou Y, Gonzalez G, Oldford N, Roe AJ, et al. Thioredoxin Reductase is a major regulator of metabolism in leukemia cells. Oncogene. 2021 Aug;40(33):5236–46.

47. Yan L, Davé UP, Engel M, Brandt SJ, Hamid R. Loss of TG-Interacting Factor 1 decreases survival in mouse models of myeloid leukaemia. J Cell Mol Med. 2020 Nov;24(22):13472–80.

48. Rettig EM, Talbot CC, Sausen M, Jones S, Bishop JA, Wood LD, et al. WHOLE-GENOME SEQUENCING OF SALIVARY GLAND ADENOID CYSTIC CARCINOMA. Cancer Prev Res (Phila). 2016 Apr;9(4):265–74.

49. Wang Z, Liu X, Wang Z, Hu Z. FOXK2 transcription factor and its roles in tumorigenesis (Review). Oncology Letters. 2022 Dec 1;24(6):1–21.

50. Zhao E, Li Y, Fu X, Zhang JY, Zeng H, Zeng L, et al. Cloning and expression of human GTDC1 gene (glycosyltransferase-like domain containing 1) from human fetal library. DNA Cell Biol. 2004 Mar;23(3):183–7.

51. Meyer C, Hofmann J, Burmeister T, Gröger D, Park TS, Emerenciano M, et al. The MLL recombinome of acute leukemias in 2013. Leukemia. 2013 Nov;27(11):2165–76.

52. Sahoo BK, Lin YC, Tu CF, Lin CC, Liao WJ, Li FA, et al. Signal peptide-CUB-EGF-like repeat-containing protein 1-promoted FLT3 signaling is critical for the initiation and maintenance of MLL-rearranged acute leukemia. Haematologica. 2022 Aug 25;108(5):1284–99.

53. Rönnerblad M, Andersson R, Olofsson T, Douagi I, Karimi M, Lehmann S, et al. Analysis of the DNA methylome and transcriptome in granulopoiesis reveals timed changes and dynamic enhancer methylation. Blood. 2014 Apr 24;123(17):e79–89.

54. Fedoryshchak RO, Přechová M, Butler AM, Lee R, O'Reilly N, Flynn HR, et al. Molecular basis for substrate specificity of the Phactr1/PP1 phosphatase holoenzyme. Hunter T, Cooper JA, Hunter T, Barford D, editors. eLife. 2020 Sep 25;9:e61509.

55. DTHD1 death domain containing 1 - NIH Genetic Testing Registry (GTR) - NCBI [Internet]. [cited 2023 May 8]. Available from: https://www.ncbi.nlm.nih.gov/gtr/genes/401124/

56. Xie S, Wang X, Gan S, Tang X, Kang X, Zhu S. The Mitochondrial Chaperone TRAP1 as a Candidate Target of Oncotherapy. Front Oncol. 2021 Jan 26;10:585047.

57. Grossmann V, Haferlach C, Nadarajah N, Fasan A, Weissmann S, Roller A, et al. CEBPA double-mutated acute myeloid leukaemia harbours concomitant molecular

mutations in 76·8% of cases with TET2 and GATA2 alterations impacting prognosis. Br J Haematol. 2013 Jun;161(5):649–58.

58. Namasu CY, Katzerke C, Bräuer-Hartmann D, Wurm AA, Gerloff D, Hartmann JU, et al. ABR, a novel inducer of transcription factor C/EBPα, contributes to myeloid differentiation and is a favorable prognostic factor in acute myeloid leukemia. Oncotarget. 2017 Oct 26;8(61):103626–39.

59. Hayes MP, Roman DL. Regulator of G Protein Signaling 17 as a Negative Modulator of GPCR Signaling in Multiple Human Cancers. AAPS J. 2016 Feb 29;18(3):550–9.

60. Kikkawa T, Osumi N. Multiple Functions of the Dmrt Genes in the Development of the Central Nervous System. Front Neurosci. 2021 Dec 9;15:789583.

61. Figueroa ME, Lugthart S, Li Y, Erpelinck-Verschueren C, Deng X, Christos PJ, et al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. Cancer Cell. 2010 Jan 19;17(1):13–27.

62. Koya J, Kataoka K, Sato T, Bando M, Kato Y, Tsuruta-Kishino T, et al. DNMT3A R882 mutants interact with polycomb proteins to block haematopoietic stem and leukaemic cell differentiation. Nat Commun. 2016 Mar 24;7:10924.

63. Xu J, Wang YY, Dai YJ, Zhang W, Zhang WN, Xiong SM, et al. DNMT3A Arg882 mutation drives chronic myelomonocytic leukemia through disturbing gene expression/DNA methylation in hematopoietic cells. Proceedings of the National Academy of Sciences. 2014 Feb 18;111(7):2620–5.

64. Ribeiro AFT, Pratcorona M, Erpelinck-Verschueren C, Rockova V, Sanders M, Abbas S, et al. Mutant DNMT3A: a marker of poor prognosis in acute myeloid leukemia. Blood. 2012 Jun 14;119(24):5824–31.

65. Baum PD, Venturi V, Price DA. Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. European journal of immunology. 2012/10/31 ed. 2012 Nov;42(11):2834–9.

66. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology. 2012 Mar;135(3):183–91.

67. Calis JJ, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. Trends Immunol. 2014 Oct;35(12):581–90.

68. Han Y, Li H, Guan Y, Huang J. Immune repertoire: A potential biomarker and therapeutic for hepatocellular carcinoma. Cancer letters. 2015/07/19 ed. 2016 Sep;379(2):206–12.

69. Hou XL, Wang L, Ding YL, Xie Q, Diao HY. Current status and recent advances of next generation sequencing techniques in immunological repertoire. Genes and immunity. 2016/03/11 ed. 2016 Apr;17(3):153–64.

70. Nikolich-Zugich J, Slifka MK, Messaoudi I. The many important facets of T-cell repertoire diversity. Nature reviews Immunology. 2004/03/26 ed. 2004 Feb;4(2):123–32.

71. Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP, et al. The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. Frontiers in immunology. 2013/12/19 ed. 2013 Nov 27;4:413.

72. Woodsworth DJ, Castellarin M, Holt RA. Sequence analysis of T-cell repertoires in health and disease. Genome medicine. 2013/11/01 ed. 2013;5(10):98.

73. Han A, Glanville J, Hansmann L, Davis MM. Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nature biotechnology. 2014/06/24 ed. 2014 Jul;32(7):684–92.

74. Redmond D, Poran A, Elemento O. Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. Genome Medicine. 2016 Jul 27;8(1):80.

75. Dziubianau M, Hecht J, Kuchenbecker L, Sattler A, Stervbo U, Rodelsperger C, et al. TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell-related pathology. American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons. 2013/09/12 ed. 2013 Nov;13(11):2842–54.

76. Liu X, Zhang W, Zeng X, Zhang R, Du Y, Hong X, et al. Systematic Comparative Evaluation of Methods for Investigating the TCRbeta Repertoire. PloS one. 2016/03/29 ed. 2016;11(3):e0152464.

77. Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. Overview of methodologies for T-cell receptor repertoire analysis. BMC Biotechnology. 2017 Jul 10;17(1):61.

78. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun. 2013 Oct 25;4(1):2680.

79. Zhang W, Du Y, Su Z, Wang C, Zeng X, Zhang R, et al. IMonitor: A Robust Pipeline for TCR and BCR Repertoire Analysis. Genetics. 2015/08/25 ed. 2015 Oct;201(2):459–72.

80. Okino ST, Kong M, Sarras H, Wang Y. Evaluation of bias associated with high-multiplex, target-specific pre-amplification. Biomolecular detection and quantification. 2016/04/15 ed. 2016 Jan;6:13–21.

81. Shewhart WA, Deming WE. Statistical method from the viewpoint of quality control. Courier Corporation; 1986.

82. Byrne KT, Vonderheide RH. CD40 Stimulation Obviates Innate Sensors and Drives T Cell Immunity in Cancer. Cell Rep. 2016/06/09 ed. 2016;15(12):2719–32.

83. Medler TR, Murugan D, Horton W, Kumar S, Cotechini T, Forsyth AM, et al. Complement C5a Fosters Squamous Carcinogenesis and Limits T Cell Response to Chemotherapy. Cancer Cell. 2018 Oct 8;34(4):561-578.e6.

84. Goodglick LA, Vaslet CA, Messier NJ, Kane AB. Growth factor responses and protooncogene expression of murine mesothelial cell lines derived from asbestos-induced mesotheliomas. Toxicologic pathology. 1997;25(6):565–73.

85. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. Blood. 2009 Nov 5;114(19):4099–107.

86. Andrews S. FASTQC. A quality control tool for high throughput sequence data. FASTQC A quality control tool for high throughput sequence data [Internet]. 2010; Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

87. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014 Mar;30(5):614–20.

88. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015 May;12(5):380–1.

89. McCullagh P, Nelder J. Generalized Linear Models II. 1989;

90. Straining R, Eighmy W. Tazemetostat: EZH2 Inhibitor. J Adv Pract Oncol. 2022 Mar;13(2):158–63.

91. Griffiths EA, Gore SD. Epigenetic Therapies in MDS and AML. Adv Exp Med Biol. 2013;754:253–83.

92. Menon U, Gentry-Maharaj A, Burnell M, Singh N, Ryan A, Karpinskyj C, et al. Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. The Lancet. 2021 Jun 5;397(10290):2182–93.

93. Steele MM, Jaiswal A, Delclaux I, Dryg ID, Murugan D, Femel J, et al. T cell egress via lymphatic vessels is tuned by antigen encounter and limits tumor control. Nat Immunol. 2023 Apr;24(4):664–75.

94. Galliverti G, Wullschleger S, Tichet M, Murugan D, Zangger N, Horton W, et al. Myeloid Cells Orchestrate Systemic Immunosuppression, Impairing the Efficacy of Immunotherapy against HPV+ Cancers. Cancer Immunol Res. 2020 Jan;8(1):131–45.

95. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJM, Robert L, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature. 2014 Nov 27;515(7528):568–71.

96. Twyman-Saint Victor C, Rech AJ, Maity A, Rengan R, Pauken KE, Stelekati E, et al. Radiation and dual checkpoint blockade activate non-redundant immune mechanisms in cancer. Nature. 2015 Apr;520(7547):373–7.

97. Commissioner O of the. Coronavirus (COVID-19) Update: FDA Authorizes Adaptive Biotechnologies T-Detect COVID Test [Internet]. FDA. FDA; 2021 [cited 2023 Jun 14]. Available from: https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-adaptive-biotechnologies-t-detect-covid-test

98. Barbouti A, Höglund M, Johansson B, Lassen C, Nilsson PG, Hagemeijer A, et al. A novel gene, MSI2, encoding a putative RNA-binding protein is recurrently rearranged at disease progression of chronic myeloid leukemia and forms a fusion gene with HOXA9 as a result of the cryptic t(7;17)(p15;q23). Cancer Res. 2003 Mar 15;63(6):1202–6.

99. De Weer A, Speleman F, Cauwelier B, Van Roy N, Yigit N, Verhasselt B, et al. EVI1 overexpression in t(3;17) positive myeloid malignancies results from juxtaposition of EVI1 to the MSI2 locus at 17q22. Haematologica. 2008 Dec;93(12):1903–7.

100. Saleki R, Christensen T, Liu W, Wang X, Chen QC, Aakre M, et al. A novel TTC40-MSI2 fusion in de novo acute myeloid leukemia with an unbalanced 10;17 translocation. Leuk Lymphoma. 2015 Apr;56(4):1137–9.

101. Wang K, Sanchez-Martin M, Wang X, Knapp KM, Koche R, Vu L, et al. Patient-derived xenotransplants can recapitulate the genetic driver landscape of acute leukemias. Leukemia. 2017 Jan;31(1):151–8.

102. Chen BR, Deshpande A, Barbosa K, Kleppe M, Lei X, Yeddula N, et al. A JAK/STAT-mediated inflammatory signaling cascade drives oncogenesis in AF10-rearranged AML. Blood. 2021 Jun 17;137(24):3403–15.

103. Celik H, Koh WK, Kramer AC, Ostrander EL, Mallaney C, Fisher DAC, et al. JARID2 Functions as a Tumor Suppressor in Myeloid Neoplasms by Repressing

Self-Renewal in Hematopoietic Progenitor Cells. Cancer Cell. 2018 Nov 12;34(5):741-756.e8.

104. Zhang H, Song G, Song G, Li R, Gao M, Ye L, et al. Identification of DNA methylation prognostic signature of acute myelocytic leukemia. PLoS One. 2018 Jun 22;13(6):e0199689.

105. Chen X, Glytsou C, Zhou H, Narang S, Reyna DE, Lopez A, et al. Targeting Mitochondrial Structure Sensitizes Acute Myeloid Leukemia to Venetoclax Treatment. Cancer Discov. 2019 Jul;9(7):890–909.

106. Nagy Á, Ősz Á, Budczies J, Krizsán S, Szombath G, Demeter J, et al. Elevated HOX gene expression in acute myeloid leukemia is associated with NPM1 mutations and poor survival. J Adv Res. 2019 Jun 11;20:105–16.

107. Huang F, Sun J, Chen W, He X, Zhu Y, Dong H, et al. HDAC4 inhibition disrupts TET2 function in high-risk MDS and AML. Aging (Albany NY). 2020 Jul 1;12(17):16759–74.

108. Zhang J, Gao X, Yu L. Roles of Histone Deacetylases in Acute Myeloid Leukemia With Fusion Proteins. Frontiers in Oncology [Internet]. 2021 [cited 2023 Apr 19];11. Available from: https://www.frontiersin.org/articles/10.3389/fonc.2021.741746

109. Bär I, Ast V, Meyer D, König R, Rauner M, Hofbauer LC, et al. Aberrant Bone Homeostasis in AML Is Associated with Activated Oncogenic FLT3-Dependent Cytokine Networks. Cells. 2020 Nov 9;9(11):2443.

110. Zhao C, Wang Y, Tu F, Zhao S, Ye X, Liu J, et al. A Prognostic Autophagy-Related Long Non-coding RNA (ARlncRNA) Signature in Acute Myeloid Leukemia (AML). Front Genet. 2021 Jun 30;12:681867.

111. Conte M, Dell'Aversana C, Sgueglia G, Carissimo A, Altucci L. HDAC2-dependent miRNA signature in acute myeloid leukemia. Febs Letters. 2019 Sep;593(18):2574.

112. Zhang Z, Zhao L, Wei X, Guo Q, Zhu X, Wei R, et al. Integrated bioinformatic analysis of microarray data reveals shared gene signature between MDS and AML. Oncol Lett. 2018 Oct;16(4):5147–59.

113. Marcucci G, Maharry K, Wu YZ, Radmacher MD, Mrózek K, Margeson D, et al. IDH1 and IDH2 Gene Mutations Identify Novel Molecular Subsets Within De Novo Cytogenetically Normal Acute Myeloid Leukemia: A Cancer and Leukemia Group B Study. J Clin Oncol. 2010 May 10;28(14):2348–55.

114. Yu N, Shin S, Choi JR, Kim Y, Lee KA. Concomitant AID Expression and BCL7A Loss Associates With Accelerated Phase Progression and Imatinib Resistance in Chronic Myeloid Leukemia. Ann Lab Med. 2017 Mar;37(2):177–9.

115. Reister S, Mahotka C, Grinstein E. Nucleolin as activator of TCF7L2 in human hematopoietic stem/progenitor cells. Leukemia. 2021 Dec;35(12):3616–8.

116. Mabrey FL, Chien SS, Martins TS, Annis J, Sekizaki TS, Dai J, et al. High Throughput Drug Screening of Leukemia Stem Cells Reveals Resistance to Standard Therapies and Sensitivity to Other Agents in Acute Myeloid Leukemia. Blood. 2018 Nov 29;132(Supplement 1):180.

117. Liu Q, Liu N, Zang S, Liu H, Wang P, Ji C, et al. Tumor Suppressor DYRK1A Effects on Proliferation and Chemoresistance of AML Cells by Downregulating c-Myc. PLoS One. 2014 Jun 5;9(6):e98853.

118. Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA. A Leukemic Stem Cell Gene Expression Signature is Associated with Clinical Outcomes in Acute Myeloid Leukemia. JAMA. 2010 Dec 22;304(24):2706–15.