Applications of Native and Engineered Genetic Barcodes in Single-Cell RNA-Sequencing Data to Study Clonal Evolution and Cellular Phenotypic Diversity

by

Jideofor Agunwa Ezike

B.S. Chemical Engineering, Carnegie Mellon University, 2015

Submitted to the Computational and Systems Biology Graduate Program
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2025

Authored by:	Jideofor Agunwa Ezike Computational and Systems Biology April 18, 2025
Certified by:	Gad Getz Professor of Pathology, Harvard Medical School, Thesis Supervisor
Accepted by:	Chris Burge Professor of Biology, MIT Co-Director, Computational and Systems Biology Graduate Program

THESIS COMMITTEE

Thesis Supervisor

Gad Getz
Professor of Pathology

Thesis Readers

Jonathan Weissman
Professor of Biology

Caroline Uhler
Professor of Computer Science

Mario Suva
Professor of Pathology

Aviv Regev
Head, Executive Vice President, Genentech Research and Early Development

Applications of Native and Engineered Genetic Barcodes in Single-Cell RNA-Sequencing Data to Study Clonal Evolution and Cellular Phenotypic Diversity

by

Jideofor Agunwa Ezike

Submitted to the Computational and Systems Biology Graduate Program
on April 18, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

Cells are constantly altering their states, whether due to physiological stress or exogenous forces. Clonal expansion is a well-defined process that contributes to this alteration and indiscriminately occurs in all types of tissue throughout the body. Acquired mutations are thus at risk of being clonally expanded and ultimately propagated within cell lineages. Understanding how populations of cells relate to each other phylogenetically aids in the hypothesis generation of drivers of different developmental processes, such as cancer evolution and hematopoietic development. In this thesis, we describe a suite of computational and analytical approaches that enable one to study lineage trajectories within single cells and gene expression programs associated with said lineages. Each chapter describes a different single-cell modality from which either barcode markers or snapshot expression information is used to construct lineage trajectories or identify subclones with shared features.

Chapter 1 describes an atlas single-cell human hematopoiesis study that profiles HSPCs throughout human lifespan (gestation to 77yo). Here, we used various snapshot lineage tracing based methods to quantify the lineage fate biases across human lifetime and identify lineage-specific genes that are both consistently and variably expressed across human lifetime.

Chapter 2 describes a computational pipeline for identifying somatic mutations from full-length single-cell RNA-sequencing data, denoising various technical artifacts that plague mutation calling in RNA-sequencing data. This enables the uncovering of cancer associated mutational signatures from single-cell mutations and detection of clones that are orthogonally supported by shared copy number alterations.

Lastly, Chapter 3 describes a study where we build high resolution single-cell phylogenies, using a CRISPR-based lineage tracing system, to study cancer persistence potential. Devising a lineage informed "persistence potential score" per cell, coupled with phylogenetic statistical tests, enabled the identification of gene programs that potentially modulate lung cancer cells' potential to persist under targeted treatment. Upon inhibiting the pathways we found to be most associated with persistence potential in combination with osimertinib, a tyrosine kinase inhibitor (TKI), we observe synergistic killing of persisters relative to TKI-only treatment.

Collectively, these works demonstrate the utility in leveraging single-cell (native or engineered) barcoding information to identify single-cell lineages, and contribute advanced computational methods for hypothesis generation when using single-cell lineage or mutation data.

Thesis supervisor: Gad Getz
Title: Professor of Pathology, Harvard Medical School

Acknowledgments

I first want to thank God for making all of this possible and for seeing me through the ups and downs of the PhD experience.

I have a whole village of people to acknowledge for reaching this milestone in my life. I could not have achieved this feat without the support of key people that I have been blessed and lucky to have crossed paths with.

I want to, first, thank my advisor, Professor Gad (Gaddy) Getz, for investing in me as a scientist over the last 6 years. I have been lucky to learn from your extensive expertise and knowledge in cancer genomics and statistical modeling of genomics data. Thank you for giving me the space to intellectually explore multiple different interesting projects in the lab. In our meetings, whether it be one on one or small group meetings, Gaddy is always engaged, encouraging, curious and lighthearted, leaving me more energized and motivated to tackle the hurdles in my projects. I am very honored to have been trained and supported by Gaddy over the years and look forward to working together in different capacities in the future.

I also want to thank all of the members of the Getz lab for providing a great environment to conduct research and foster relationships. I want to thank my fellow graduate students for the fellowship and community they have provided over the years: Michael Vinyard, Ruitong Li, Chris Etienne and Rebecca Boiarsky. I also want to specially thank my postdoctoral collaborators, Binyamin (Benny) Zhitomirsky and Tim Coorens, for their invaluable feedback in the development of the respective projects. It has been an absolute pleasure working with such talented independent scientists on these projects. I also want to thank these amazing scientists for their invaluable conversations, assistance and feedback with my work: Chip

Stewart, Romanos Sklavenitis Pistofidis, Arvind Ravi, Julian Hess, Nicholas Haradhvala and Junko Tsuji. I also want to thank Mendy Miller for being instrumental in writing multiple important documents, from my F99-K00 NIH fellowship to project manuscripts. Working with you on different tasks over the years has been a pleasure and your positive spirit is always uplifting and greatly appreciated. Lastly, I want to thank Serene King for keeping the lab running smoothly, planning lab social events, and scheduling meeting updates, even on short notices :). Serene was absolutely instrumental to the maintenance of the Getz lab culture. I will miss being a regular member of the lab.

I want to give a huge thank you to Aviv Regev, my original MIT advisor, for also investing in me as a scientist and individual. It has been a great honor to be trained by such an esteemed scientist and equally great person. Learning from your extensive experience in all things single-cell genomics has been a great pleasure. Also, watching how you approach mentorship has been inspiring. You are always very engaged during our meetings and provide me with instrumental feedback that push my projects forward significantly. I am grateful for your presence in my academic journey and your continued support beyond graduate school.

I want to thank my entire Thesis Committee for consistently providing invaluable feedback and support: Jonathan Weissman, Caroline Uhler and Mario Suva. I feel privileged to have an esteemed committee of experts who each had a very tangible contribution to my thesis research. Thank you Mario for providing a lot of the full-length sequencing data for the early development of my mutation detection pipeline and providing intriguing biological applications to explore. Thank you, Jonathan, for helping us establish the cell line engineering for the Cassiopeia lineage tracing system and for posing important biological questions to consider when analyzing single-cell lineage tracing data in the context of cancer persistence. Lastly, thank you Caroline for providing invaluable guidance in the application of machine learning methodologies to my data. I could not have asked for a better fit and supporting committee than the one I had.

I also want to extend a special thank you to all of my extended research collaborators

who provided critical feedback to the development of my projects: Dana Silverbush (former Regev and Suva Lab postdoctoral associate; now UPenn Professor), Geoffrey Schiebinger (former Regev Lab postdoctoral associate; now UBC Professor), Mia Petljak (former Regev Lab postdoctoral associate; now NYU Professor) and Matthew Jones (former Weissman Lab PhD student; now Stanford postdoctoral associate), Masashi Nomura (former postdoctoral associate in Suva Lab) and John Lee (postdoctoral associate in Suva Lab). All of you were instrumental to the development of my work and I always enjoyed my meetings with all of these talented scientists. I look forward to maintaining professional relationships and friendships with all of you.

My academic journey has been quite the winding one, filled with recurrent happenstance encounters with key individuals who helped propelled me forward. I want to thank Reverend Ndubuisi Azubuine, who had been working at the Whitehead Institute of MIT for many years, for taking the effort to introduce me to David Pincus, a talented Whitehead fellow at the time (now Professor at the University of Chicago). Reverend was the catalyst that helped get my foot in the door of this world of biomedical and life sciences research and I am forever grateful for that. Thank you, David, for providing me my first research opportunity in studying the mechanisms of the heat shock response and for remaining a mentor, cheering and encouraging me along the way. Thank you for instilling the confidence in me that I can be a successful scientist. I also want to extend a huge thank you to Professor Harvey Lodish of the Whitehead Institute at MIT for investing in me as a scientist and providing me with transformative years as a research technician in his lab, studying mammalian erythropoiesis. I learned many skills during this time and received my first introduction to single-cell genomics in your lab, an area I went on to explore extensively in many different contexts after your lab. Thank you for being the intentional mentor that you are and for always asking the right questions that dare me to follow my dreams. I thank and appreciate you dearly, Harvey. I also want to thank all of the people in the Lodish and Pincus labs who were instrumental to my development: Xu Zheng, Joanna Krakowiak, Jai Pandey, Marko Knoll, Anirudh Natarajan

and Hojun Li.

I want to thank all of my life friends for their continued support and fellowship. While this is not an exhaustive list, I want to thank these individuals for being staples in my PhD journey: Daymanuel Sampson, Darlene Reid, Chimaobi Ahaneku, Randy Garcia, Delvis Taveras, the Baez Family, Sam Lobel, Will Johnson, Paul and Amber Lachaud, Churchill Isaac and Ntami Echeng. The fellowship, regular check-ins and emotionally support that you all provide have propelled me throughout my PhD.

I also want to thank all of my "MIT friends" who have made this esteemed place feel like home. Again, this is not an exhaustive list but I want to thank these staples in my MIT experience: Jude Safo, Manon Revel, Adedayo Aderibole, Michael West, Noah Jones, Sameed Siddiqui, Ifrah Tariq, Uyiosa Oriakhi, Corban Swain, Corshai Williams and Julius Francis. I appreciate you all for being humble and good character people. Thank you for all of the years of basketball playing, fun times, laughter and for sharing in this journey of the MIT PhD experience.

I have quite a large and connected family, so I have to acknowledge my extended family, both domestically and abroad, in Nigeria. I want to acknowledge and thank my cousins, the Iroche's: Jennifer, Lillian, Edward and Miriam. Thank you for being, effectively, extended siblings and constantly being a source of comfort, encouragement and support. Whenever I am around you guys, I automatically feel at home. We grew up together and I look forward to maintaining our close bonds as we get older and traverse different phases of adulthood. Special shout out to you, Eddie, for being like another brother, and always being present for events with the immediate family. Your funny, jovial presence is always greatly appreciated and enjoyed. I lastly want to thank their parents Chima, the late Christina (my mom's older sister), and Obioma Iroche for being great parental figures for both myself and their children.

I also want to thank the Okeke cousins: Nk, Casey, Gabriella and Daniella. You all are great kids and young adults and I enjoy watching you all grow up. I am proud of the people you are becoming and I will always be there to support you all as big cousin. I also want to

thank their parents, Ikenna and Ngozi, for being great parental figures, as well.

I want to thank the entire Egonu and Ezike families abroad, in Nigeria, for the regular prayers and check-ins. Despite, the distance, I always feel the love from all of you. I look forward to seeing and connecting more with many of you in person as I enter this next phase of life. I appreciate you all, dearly. I also want to use this opportunity to acknowledge all of the family members who have passed away and have contributed to my personal development; I specifically want to acknowledge my uncle Michael (Mike) Egonu, who was like a father figure, and passed away a few years into my PhD. I know he would have loved to have witnessed this milestone and would have been very proud of this achievement.

I also want to thank my aunty Theresa Nnodum, late uncle Valentine Nnodum and the entire Nnodum family. This family effectively provided the Nigerian immigrant blueprint that my parents adapted to survive and thrive in this country. Our families are tied at the hip; their kids are like extended siblings and aunty Theresa and uncle Val are like a second set of parents. I want to thank my aunty Theresa for being present at every major milestone in my life and for the invaluable phone calls we have periodically. Your support and love is always felt and I appreciate your presence in my life. I also want to thank all of the Nnodum children for supporting me and being there for me; I appreciate every single one of you.

I want to thank my parents, George and Miriam Ezike, for instilling in me the love of learning and the value of hard work for whatever you pursue. Watching them, as immigrants in this country, work tirelessly to build a better life for their kids inevitably instilled a drive in me to do the same for myself, my family and my larger community. To go from immigrants, starting from scratch in a foreign country in the Bronx, NY, to raising kids in an upper-middle class town who ended up attending the best universities in the country, is nothing short of remarkable. More importantly, beyond raising accomplished, productive children, they also managed to raise good, high-character children, who treat others with respect and aim to leave the world a better place than they found it. For this, I am grateful to them for helping to mold me into the person I am proud to be today. While, we have certainly had our fair

share of differences as I have matured into a young adult, at the time of this writing, I have seen a positive shift in dynamics and am hopeful and eager to witness the further evolution of our relationships. Thank you for everything!

I also want to acknowledge my siblings: Kayode (Ayo), Adaugo and Dioka Ezike for being great siblings. They made my life quite easy as the oldest child by just simply doing what was expected of them growing up. I am proud of the adults they have become. Ayo, it has been great to see our relationship evolve towards a deeper friendship over the years and I really appreciate our regular calls. Adaugo, I look forward to further building our relationship and I want you to know that I am always there for you if you need me. And Dee, you know we all love you immensely and are always rooting for you. You will always be our baby and we will always be there for you. Thank you for being great siblings and providing the necessary pressure for me to lead by example and constantly challenge myself to be a better person. Love you guys and thank you!

I want to thank my doggy son, Milo, for being a beautiful dog that brings joy to every room he enters. We got Milo in the middle of my PhD and he has been a pivotal aspect of my PhD experience. Milo has been by my side through many stressful moments in preparation for countless important deadlines and meetings. His presence has always been a grounding one for me and we have forged a deep bond through this experience. I love you, dearly, Milo.

Lastly, but certainly not least, I want to acknowledge and thank my beautiful wife, Ogechukwu (Oge) Ezike. You have supported me in countless ways on this PhD journey. Thank you for the encouragement along the way and the companionship you have provided in my most trying times. It has been quite the experience to go through the training phases of our careers together. You were an integral part of this PhD experience. I look forward to continuing to build a beautiful life with you, as we enter this new phase of our relationship. I look forward to reaping the fruits of our labor and loving each other, without the stressors of graduate school and medical training. I love you, Oge, and am forever grateful and lucky to have you as the ultimate staple in my life. Thank you for everything you do to enhance my

life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Clonal Expansion in Normal and Disease Tissue

Maintaining tissue homeostasis relies on a finely tuned balance of asymmetrical and symmetrical cell division. Stem cells, for instance, can asymmetrically divide to self-renew while producing differentiated progeny, or symmetrically divide to expand specific cell populations. This dynamic gives rise to clonal expansion, a process that occurs in both normal and diseased tissue contexts [1–5]. In healthy tissues, clonal expansion is essential—for example, in ensuring sufficient production of erythrocytes. However, when unchecked, it can give rise to precancerous lesions and ultimately malignant transformation. Increasing evidence supports the idea that many cancers emerge from such intermediate stages of clonal evolution. Studies have shown that histologically normal tissues can harbor abundant clones carrying canonical driver mutations. For instance, roughly 30% of cells in normal adult sun-exposed skin contain mutations in genes such as NOTCH1 and TP53—mutations that are frequently implicated in skin cancers [6].

This observation highlights an important feature of clonal expansion: cells may appear and function normally, yet harbor somatic mutations that confer a subtle fitness advantage, enabling them to outcompete neighboring cells. This stepwise continuum of cell states forms

the basis for viewing cancer as a disease of somatic evolution, driven by the accumulation of mutations over time [7, 8]. Among these, driver mutations confer selective growth advantages, while passenger mutations accumulate passively without functional consequence. Together, these somatic mutations serve as molecular 'breadcrumbs', marking the evolutionary history of distinct cellular lineages. Advances in next-generation sequencing (NGS), genomic characterization tools and lineage-tracing technologies have made it possible to identify these mutations, reconstruct clonal relationships, and analyze gene expression profiles across lineages [9–11]. By applying phylogenetic frameworks to genomics data, researchers can map out how clones diverge, expand, and acquire new traits. This not only deepens our understanding of the molecular mechanisms driving clonal expansion but also opens therapeutic avenues for intercepting or redirecting these trajectories toward less harmful outcomes.

1.2 Multi-Omic Profiling of Single Cells

Bulk sequencing assays—such as whole-genome sequencing (WGS), whole-exome sequencing (WES), and RNA sequencing (RNA-seq)—profile nucleic acid sequences from a large population of cells. These technologies have driven major biomedical discoveries, including the identification of commonly mutated cancer genes [12]. However, a more recent class of assays, broadly known as single-cell sequencing (sc-seq), enables tissue characterization at a resolution not achievable with bulk methods. These approaches can profile molecular features at the level of individual cells, such as gene expression (via mRNA count), chromatin accessibility, DNA methylation, and spatial organization. The granularity captured by these assays make them especially well-suited for dissecting the heterogeneity of human tissues and biological systems.

Single-cell RNA sequencing (scRNA-seq)—one of the most widely used sc-seq methods—captures the mRNA content of individual cells and identifies gene expression programs that distinguish cell types and niches. scRNA-seq has generated novel insights into human

development, such as hematopoietic stem cell (HSC) formation, and enabled the discovery of previously unrecognized cell types and states in lethal cancers, like glioblastoma [13, 14]. These technologies now provide powerful tools to study cell differentiation and trajectory in both normal and disease contexts [10, 11].

Sc-seq also opens new avenues for detecting genomic alterations such as single nucleotide variants (SNVs) and insertions/deletions (indels), many of which fall below the detection limit of bulk assays. Rare subclonal mutations, for example, may occur at frequencies indistinguishable from sequencing noise in bulk data. As a result, single-cell approaches have become increasingly appreciated as a complementary means of studying genomic heterogeneity—offering resolution that is often missed by bulk sequencing assays. Both experimental and computational strategies exist for this purpose [15–18].

Approaches like single-cell whole-exome sequencing (scWES) and single-cell whole-genome sequencing (scWGS) attempt to detect genomic variants by amplifying the limited DNA available in a single cell. However, these methods are challenged by technical issues such as uneven genome amplification (leading to dropout of loci) and polymerase-induced errors (leading to false positives). Additionally, due to high per-cell costs, they are often limited in throughput and scale compared to standard bulk sequencing.

More recent methods leverage single-cell assays not originally designed for mutation detection—such as scRNA-seq—to infer mutations from transcribed RNA molecules [15, 16]. These approaches take advantage of the cell's endogenous transcriptional machinery, which naturally amplifies the effects of DNA mutations through RNA expression. However, using scRNA-seq for variant detection presents its own challenges, including sparse coverage, allelic imbalance, and technical artifacts introduced by RNA polymerase fidelity or RNA editing.

Computational methods for this task must carefully model these sources of noise, often relying on heuristic filters or statistical models to enrich for genuine mutations. Yet even with such filtering, detecting somatic mutations—particularly in malignant samples without matched normals—remains difficult. Sequence context may provide an additional

layer of filtering, under the assumption that real and artifactual mutations differ in their sequence space. Importantly, the unique strength of these approaches lies in their ability to simultaneously profile both mutation and gene expression states within the same single cells. This dual information enables the identification of clonal populations and their associated transcriptional programs, providing valuable insights into intratumoral heterogeneity and potential therapeutic targets.

1.3 Lineage Tracing

Lineage tracing is a powerful class of experimental and computational methodologies used to track and developmentally order groups of cells. These methods rely on shared barcodes detected in individual cells to infer relationships among them. Barcodes can take multiple forms; broadly speaking, they are either naturally occurring or experimentally induced.

Retrospective lineage tracing methods utilize naturally occurring genetic barcodes, while prospective lineage tracing approaches introduce heritable barcodes experimentally. In this thesis, we leverage both types of methods. Common forms of natural barcodes include somatic nuclear mutations, mitochondrial mutations, and methylation markers. These mutations accumulate over the lifetime of primary human cells and enable developmental ordering that reflects underlying evolutionary processes within organisms. Moreover, primary human cells are often not amenable to experimental engineering or the artificial introduction of genetic barcodes in vitro, making naturally occurring barcodes especially valuable for lineage inference in these settings.

Prospective lineage tracing technologies broadly fall into two categories: stable clonal barcodes and evolving cell barcodes. In the former, individual cells are infected with lentiviral DNA barcodes that are stably propagated to the progeny of the founder cell. These clonal barcodes allow researchers to assign cells to clones and formulate hypotheses about the phenotypic fates of each clone. They have been widely used to track clones in various

biological contexts—from identifying distinct lineage trajectories of cycling cancer persister cells to tracking hematopoietic stem cells in mammalian models [19–22]. While these approaches have proven valuable, they are limited by the static nature of the barcodes: once introduced, they cannot be further modified. This prevents tracking additional changes that may occur as clones evolve over time.

In contrast, evolving barcode systems offer improved resolution by allowing individual cells within a clone to acquire unique mutations over time, enabling subclonal reconstruction. The most common systems in this class are CRISPR-based lineage tracing methods, which engineer cells to constitutively express Cas9 and integrate CRISPR target sites into the genome. As the experiment progresses, these target sites accumulate insertions and deletions (indels) induced by Cas9, which serve as lineage-defining barcodes. Evolving barcoding systems have been instrumental in unraveling biological heterogeneity—spanning from neural development in zebrafish to mapping the rates, routes, and drivers of tumor metastasis [23–25].

The stable barcode approach enables comparisons across clones, while the evolving barcode approach reveals how phenotypes diversify within a clone. Together, these approaches allow researchers to study the molecular features that drive cellular trajectories in a wide range of biological contexts, including cancer evolution, treatment response, tissue regeneration, and organismal development.

1.4 Phylogenetic Methods for Inferring Lineages/Phylogeny

While phylogenetics is a classical field originally focused on the evolutionary ordering of species based on shared morphological or genetic features, its core principles have been successfully adapted to genomics—particularly in building single-cell and clonal phylogenies. In the single-cell context, phylogenetic trees can model how mutations accumulate across

cells, revealing lineage relationships, clonal structure, and trajectories of development or disease progression. Broadly, there are three major classes of phylogenetic reconstruction methods: distance-based, maximum parsimony, and maximum likelihood.

Distance-based methods, such as Neighbor Joining (NJ) and UPGMA, construct phylogenetic trees by converting raw data into a pairwise distance matrix, where each entry reflects a chosen metric—commonly Euclidean distance, cosine similarity, or Hamming distance between mutation profiles [26]. The methods then build trees such that the distance between nodes in the tree reflects the distances in the input space. These approaches are fast and scalable, making them appealing for datasets with many cells, though they can be sensitive to noise and do not model the evolutionary process explicitly.

Maximum parsimony (MP) methods follow the principle of Occam's Razor: they attempt to reconstruct the tree that requires the fewest mutational events. In this framework, each possible tree topology is evaluated based on the minimum number of changes needed to explain the data observed at the leaves (i.e., individual cells) [23, 27, 28]. MP has been successfully applied in single-cell studies, such as lineage tracing experiments aimed at identifying drivers of metastasis or cellular plasticity in cancer [18, 25]. While MP is intuitive and relatively easy to interpret, it does not naturally account for sequencing error or missing data, which can be prevalent in single-cell assays. MP-based frameworks have been specifically adapted for single-cell lineage reconstruction, supporting both exact and heuristic tree-building strategies while accommodating for practical constraints in single-cell data, such as missingness.

Maximum likelihood (ML) methods, such as those implemented in tools like SCITE and SiFit, provide a probabilistic framework for phylogenetic inference [29, 30]. Instead of simply minimizing changes, ML methods evaluate tree topologies by calculating the likelihood of the observed data under a specific evolutionary model. These models can incorporate false positive and false negative rates, dropout, and other noise parameters common in single-cell sequencing. For example, SCITE models sequencing error explicitly and identifies the most likely tree structure given observed mutation presence/absence matrices. ML-based

approaches tend to be more robust in the presence of noise and missing data, but they are computationally more intensive, especially as the number of cells increases.

Each of these phylogenetic approaches comes with trade-offs. Distance-based methods are fast and scalable but may lack biological interpretability and are more susceptible to artifacts from high missingness. Maximum parsimony is simple and interpretable but can struggle with noisy or incomplete data. Maximum likelihood offers the most flexibility and accuracy in noisy datasets but may be computationally prohibitive for very large trees (e.g., thousands of cells). Therefore, the choice of method often depends on the size of the dataset, the level of missingness, and the desired balance between interpretability and accuracy.

Another class of methods gaining traction in lineage modeling involves unsupervised machine learning, particularly generative models like autoencoders. Autoencoders are neural network architectures designed for compressing high-dimensional, noisy input data into a lower-dimensional latent space and then reconstructing the original input data. This latent space, typically, captures the most informative aspects of the input while discarding noise and irrelevant variation. The architecture consists of an encoder that maps the input to the latent space, and a decoder that reconstructs the input from the latent representation.

Autoencoders, and particularly their probabilistic extension—variational autoencoders (VAEs)—have been applied across domains such as machine vision, natural language processing, and drug discovery, making them attractive for modeling noisy single-cell genomics data. In single-cell analysis, VAEs are often used for dimensionality reduction, denoising, and latent feature discovery. The latent space of a trained VAE can reveal meaningful biological structure, including clonal identity, cell types, and batch effects [31, 32]. These embeddings are increasingly being used as inputs to downstream tools for clone identification, trajectory inference, and integration with other omic modalities.

1.5 Summary of Thesis

Cells are continuously transitioning between states, influenced by both intrinsic processes and external pressures. Tracing how these cells relate to one another through lineage reconstruction provides a powerful lens for understanding important biological processes, including cancer evolution and hematopoietic development. In this thesis, we present a suite of computational and analytical approaches that enable the study of single-cell lineage trajectories and their associated gene expression programs. Each chapter focuses on a distinct single-cell modality, utilizing either barcoding markers or snapshot expression data to infer lineage relationships and identify subclones with shared molecular features.

Chapter 1 describes a single-cell human hematopoiesis atlas profiling 58,041 hematopoietic stem and progenitor cells (HSPCs) from 26 donors across the human lifespan, ranging from 10 post-conception weeks to 77 years of age. We implemented multiple snapshot-based lineage tracing methods to quantify lineage fate biases over time and to identify genes that drive commitment into specific hematopoietic lineages. One approach models cell trajectories as Markovian, biased random walks through neighboring transcriptional states, while another applies the principles of optimal transport to infer cell state transitions. These methods revealed how lineage outputs from HSPCs vary with age. Using inferred lineage fate probabilities from these methods, we further identified genes that both mark specific lineages and, within each lineage, show differential usage across the lifespan. Furthermore, utilizing cNMF on the expression matrix to find gene expression programs, two novel HSC states were found to be mid-gestation and elderly specific, respectively. This study contributes analytical approaches for inferring dynamics from atlas-scale single-cell data and uncovers novel genes and cell states specific to human hematopoietic lineages.

Chapter 2 details a computational pipeline for identifying somatic mutations from full-length single-cell RNA-sequencing data, with particular attention to denoising the mutations by removing the technical artifacts common to RNA-based mutation calling. Early analyses

revealed that many mutation candidates, even after filtering, did not reflect biologically plausible mutational processes by way of their mutation spectra. To address this, we developed an anomaly detection framework that considers sequence context to distinguish between likely true mutations ("outliers") and background noise ("inliers"). This approach revealed expected mutational signatures in various malignancies and demonstrated that outlier mutations were enriched for validated mutations found in matched tumor whole-exome sequencing. Furthermore, these outliers aligned with clones inferred from large-scale copy number alterations, suggesting they mark biologically meaningful clonal populations.

Chapter 3 presents a study in which we construct high-resolution single-cell phylogenies using a CRISPR-based lineage tracing system to investigate cancer persistence. By assessing the phylogenetic proximity of untreated cells to drug-persistent cells, we developed a "persistence potential score" to estimate a cell's likelihood of surviving treatment. We complemented this with a tree-traversal strategy that identifies clades with statistically significant enrichment for persisters. Differential expression analysis along the persistence potential axis revealed multiple pathways associated with persistence, with oxidative phosphorylation (OXPHOS) emerging as the most consistent signal. Functional experiments showed that co-inhibition of OXPHOS and EGFR (via osimertinib) synergistically eliminated persisters, supporting the predictive ability of our approach. This work introduces methods for generating mechanistic biological hypothesis from single-cell phylogenies and can be generalized to compare cellular phenotypes across a range of conditions or timepoints—such as treatment-naive vs. treated, or primary vs. metastatic tumors.

Collectively, these studies illustrate the power of leveraging single-cell barcoding—both natural and engineered—to reconstruct cell lineages. They also contribute computational innovations that facilitate hypothesis generation from single-cell mutation and lineage data, with applications in development, disease, and therapeutic response.

Chapter 2

The dynamics of hematopoiesis over the human lifespan

This chapter was adapted from a published study. I was one of the lead analysts and my main contributions were in inferring the lineage fate probabilities for all the cells, which was integral to the identification of the constant and variable gene expressions across lifetime.

Hojun Li was the lead and corresponding author and conceived of the idea with Grant Rowe.

2.1 Abstract

Over a lifetime, hematopoietic stem cells (HSCs) adjust their lineage output to support age-aligned physiology. In model organisms, stereotypic waves of hematopoiesis have been observed corresponding to defined age-biased HSC hallmarks. However, how the properties of hematopoietic stem and progenitor cells change over the human lifespan remains unclear. To address this gap, we profiled individual transcriptome states of human hematopoietic stem and progenitor cells spanning gestation, maturation and aging. Here we define the gene expression networks dictating age-specific differentiation of HSCs and the dynamics of fate decisions and lineage priming throughout life. We additionally identify and functionally validate a

fetal-specific HSC state with robust engraftment and multilineage capacity. Furthermore, we observe that classification of acute myeloid leukemia against defined transcriptional age states demonstrates that utilization of early life transcriptional programs associates with poor prognosis. Overall, we provide a disease-relevant framework for heterochronic orientation of stem cell ontogeny along the real time axis of the human lifespan.

2.2 Introduction

Over the human lifespan, tissue stem cells coordinate development and maintain tissue integrity in response to varying physiologic and pathologic demands [33]. Although stem cell ontogenies have been finely mapped at specific stages of life, how stem and progenitor cell properties change over a lifetime remains a fundamental but unanswered question [34–44]. To meet age-appropriate demands, the human hematopoietic system must adjust its relative prioritization of oxygen transport, hemostasis, innate and adaptive immunity and tissue regeneration. Over the preceding decades, insight has been gained into developmental differences in vertebrate hematopoietic stem and progenitor cells (HSPCs). For example, in mice, despite similar surface immunophenotypes compared with hematopoietic stem cells (HSCs) from mature bone marrow, HSCs from fetal liver (FL) possess higher self-renewal potential, show differential growth factor dependence and possess distinct programs of gene expression [45–48]. In human HSPCs, differential enhancer usage between adult and fetal cells plays a central role in erythroid maturation [49]. Additionally, human myeloid progenitors employ distinct differentiation cascades in fetal and postnatal life [50]. However, the heterochronic control of human HSPC developmental maturation and aging remains an important but largely unanswered question given the age skewing of many blood diseases such as hemoglobinopathies, bone marrow failure disorders and leukemia [51].

Advances in single-cell profiling technologies have enabled both the identification of stem and progenitor cell states and the assignment of differentiation trajectories in an

unbiased manner. Such approaches have been applied to investigate HSPC differentiation at specific human ages, such as FL, umbilical cord blood and adult bone marrow, but a single comprehensive and integrative analysis of hematopoiesis across the lifespan is lacking [52–54]. To better understand hematopoietic maturation and factors underlying age skewing of blood diseases, we combined single-cell RNA sequencing (scRNA-seq) of human HSPCs spanning early gestation into advanced adulthood with time-conscious computational analysis to provide an integrated map of how HSPC ontogeny itself matures and ages over the full lifespan. Our efforts overcome existing obstacles that obscure age-specific biology when integrating independent datasets and provide the first temporal continuum of physiologic stem and progenitor cell maturation and aging in humans at the single-cell level.

2.3 Results

2.3.1 An atlas of human HSPC maturation

In an effort to construct an atlas of human HSPCs to define age-dependent changes in differentiation over the lifespan, we aggregated nine accessible HSPC scRNA-seq datasets from various stages of life. Since each of these datasets covered select phases of life, we attempted to integrate them for comprehensive profiling of the entire lifespan. Following batch correction using multiple algorithms we applied a metric to measure disparity of individual specimens compared with the aggregated dataset and observed a high rate of rejection, indicating uncorrectable batch effects based on variability in data acquisition [55, 56]. Moreover, the aggregated dataset did not detect expected hallmarks of age-related blood formation, such as lymphoid predominance in childhood [57]. These findings diminished our confidence that aggregation of independent datasets could effectively define lifetime HSPC dynamics.

To address this gap, we procured CD34+ HSPCs at 22 timepoints from 10 postconception weeks (PCW) to 77 years of age (26 total donors) and used scRNA-seq to profile this spectrum

(Fig. 2.1). After quality controls (Methods), we retained 58,041 HSPCs with high-quality transcriptomes. We performed dimensionality reduction with a principal component analysis (PCA), followed by embedding via universal manifold approximation and projection (UMAP) to visualize the overall dataset and lineage trajectories (Fig. 2.2). Expression of lineage-specific marker genes revealed clear transcriptional priming of terminal lineages (Fig. 2.3). To begin to define temporal HSPC ontogeny, we visualized seven discrete phases (first trimester, second trimester, perinatal, childhood, adolescent, adult and elderly) of human definitive hematopoiesis based on recognized milestones and dynamics of multipotent HSPC and lineage-restricted progenitor usage (Fig. 2.4). We confirmed higher expression of fetal-associated genes in cells from fetal donors and also found sex chromosome genes contributed less to biologic variability than their 5% contribution to the genome [58]. Notably, batch correction performance metrics for this dataset were markedly improved compared with aggregation of previously published studies.

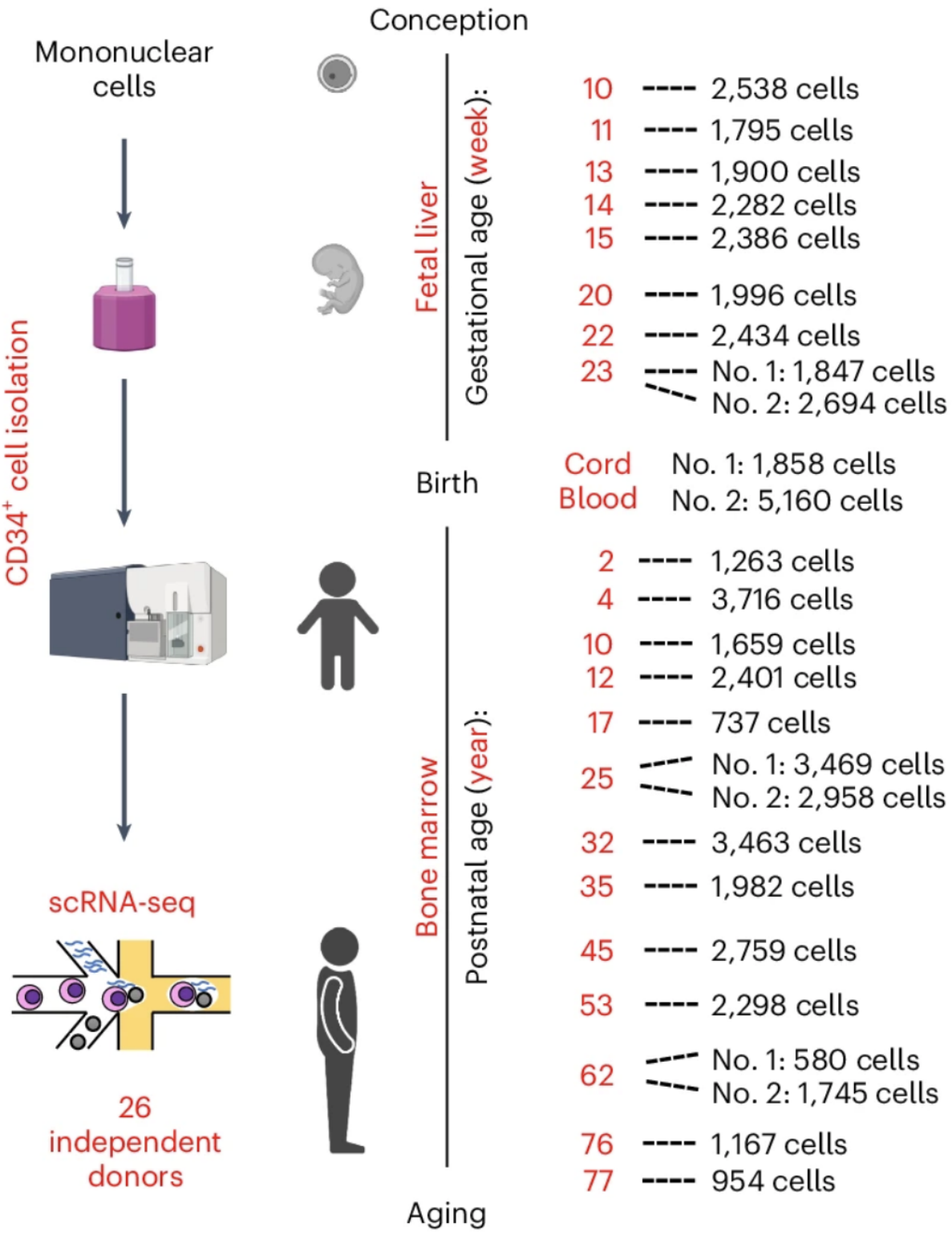

Figure 2.1: Schematic of cell processing, scRNA-seq workflow, timepoints and hematopoietic sites sampled (n=26 total).

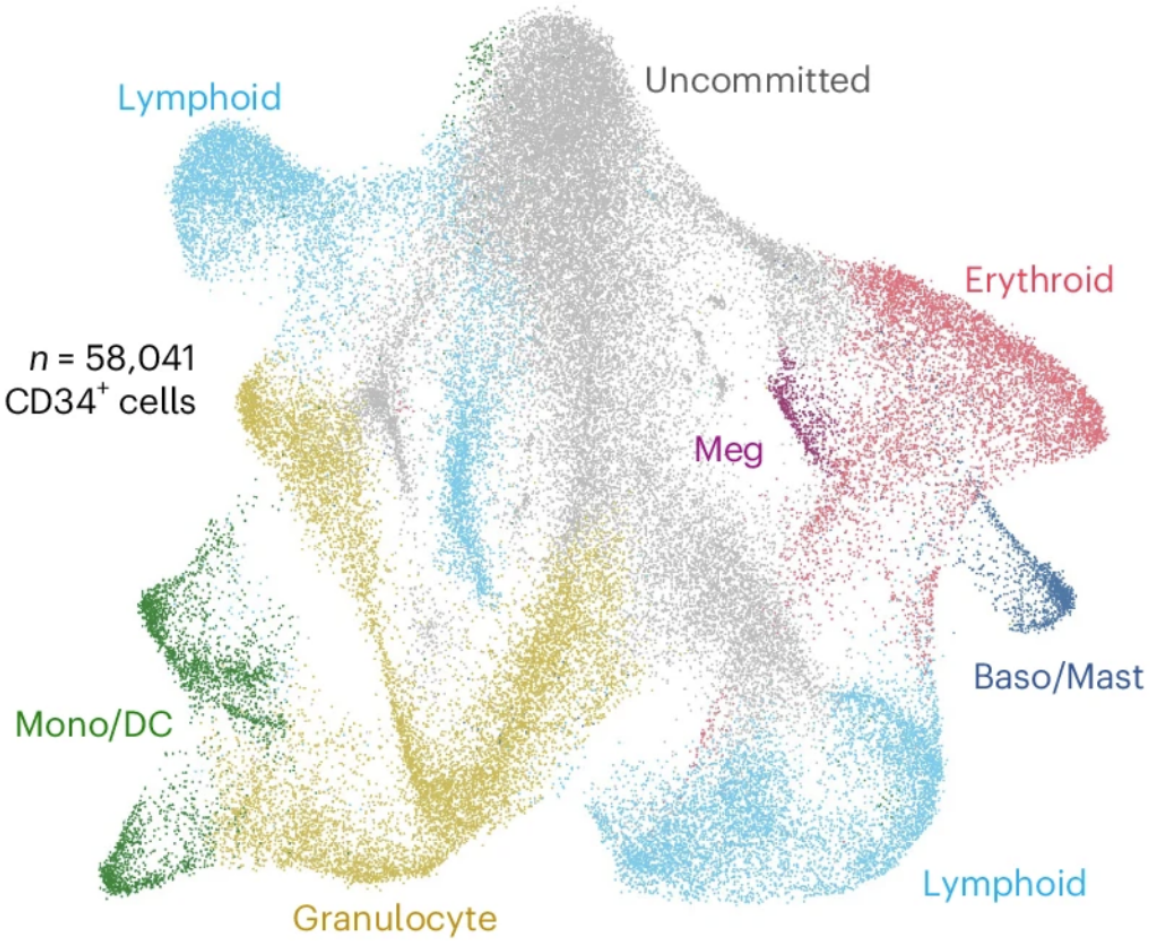

Figure 2.2: A UMAP of the entire dataset. The lineage branches were obtained by performing Louvain clustering and assignment of clusters as primed for the erythroid, granulocyte, lymphoid, monocyte/dendritic cell (Mono/DC), megakaryocyte (Meg) or basophil/mast cell (Baso/Mast) lineages or uncommitted to any lineage, based on gene expression state.

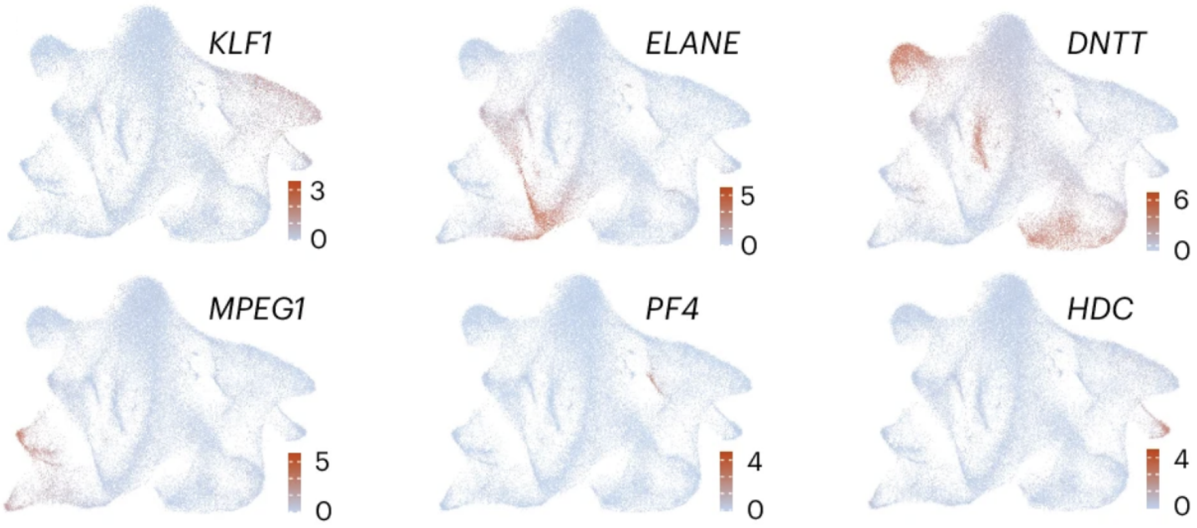

Figure 2.3: UMAP displays of the expression of characteristic marker genes KLF1 (erythroid), ELANE (granulocyte), DNTT (lymphoid), MPEG1 (Mono/DC), PF4 (Meg) and HDC (Baso/Mast). The marker gene expression legend indicates log-transformed normalized read counts per cell

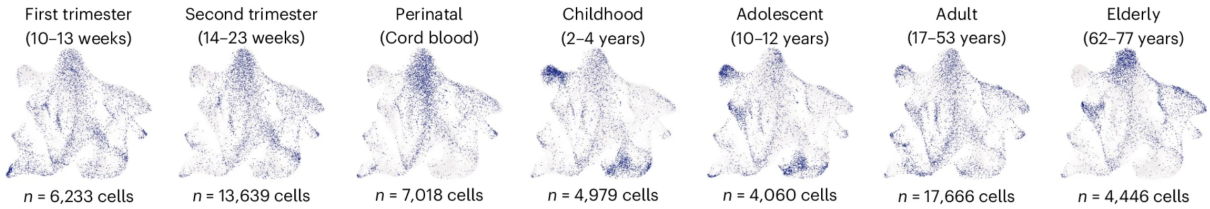

Figure 2.4: UMAP displays highlighting a random sampling of 4,000 total cells from each of the indicated phases of human life, with the total number of cells in each age group in the overall dataset indicated below.

By unsupervised clustering (Methods), we identified 22 HSPC states that we hierarchically ordered and annotated on the basis of UMAP localization and marker gene expression. To determine the consistency of these annotations, we used the singleCellNet random forest algorithm to classify each state against a human adult bone marrow scRNA-seq dataset [59]. The cells in HSC and multipotent progenitor (MPP) states were located centrally in UMAP space with branches terminating in oligo- or unilineage-primed progenitors; cluster relationships in the UMAP space were consistent with relative predicted lineage potential (Fig. 2.2). We observed two lymphoid trajectories in the UMAP space that were predominantly differentiated by cell cycle status and not due to doublets. We also observed that HSPC

states and lineage trajectories varied in their relative abundance. The first trimester FL prioritized myeloid output, shifting toward HSC/MPP expansion in the second trimester; erythroid production was maintained throughout gestation and diminished at birth (Fig. 2.5). Following the childhood surge in lymphopoiesis, myeloerythroid output increased over time in the second decade of life through adulthood, followed by expansion of HSCs and MPPs with aging (Fig. 2.5). These patterns of lineage prioritization are supported by changes that occur in peripheral blood leukocyte populations across human life and the myeloid-biased hematopoiesis and HSC expansion in mammalian FL and mature adulthood [60–62].

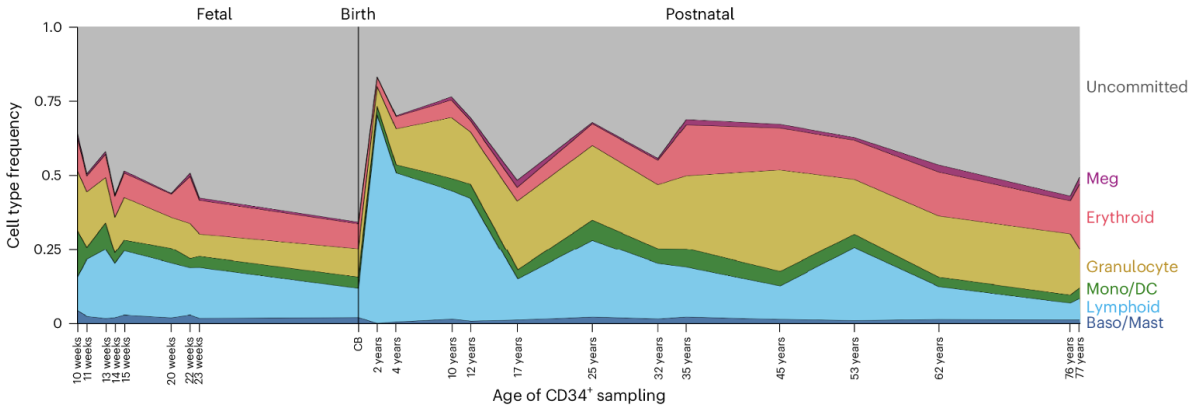

Figure 2.5: The relative frequency of each HSPC lineage state over the duration of the human lifespan from 10 weeks' gestation through age 77 years. CB, cord blood.

To gain further insight into time-dependent changes in fate, we dissected fate probabilities using both population balance analysis (PBA) and stationary optimal transport (StatOT) [63–65]. When comparing each lineage's predicted fraction of total output to the observed fraction of cells annotated as that lineage, predictions from PBA and StatOT agreed with observed output (Fig. 2.6). We found high myeloid differentiation probability in the earliest gestation FL HSCs and HSPCs overall, probably driven by macrophage bias, with a midgestation shift toward more balanced hematopoiesis; higher lymphoid probability predominated in childhood, shifting back to relatively higher myeloid probability in adulthood and increasing erythroid probabilities in elderly (Fig. 2.7-2.9). Taken together, these data provide an annotated atlas of human HSPC maturation spanning gestation through advanced adulthood and define

age-associated priorities of hematopoiesis.

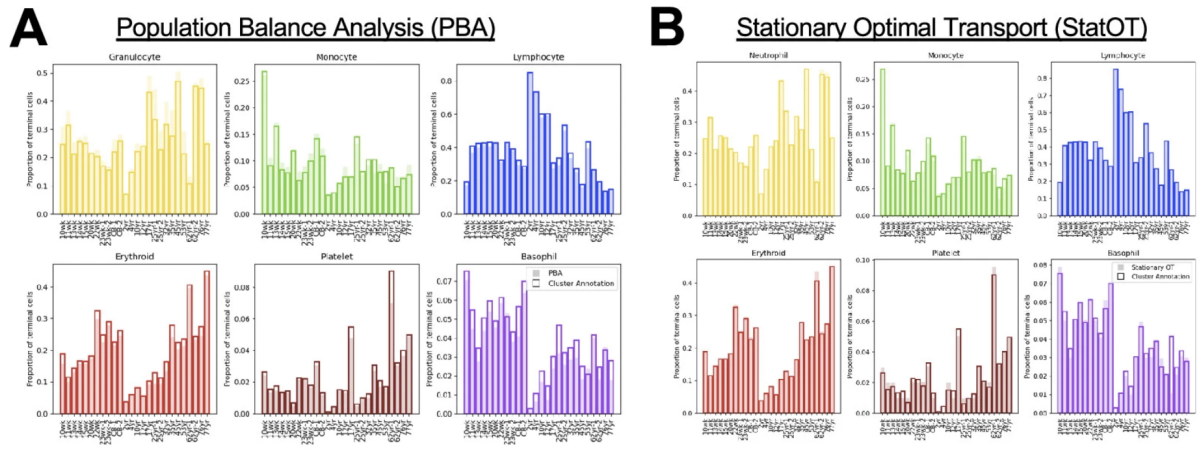

Figure 2.6: Diffusion-drift modeling was used to define commitment probabilities to lineage fates for HSPCs by PBA and StatOT. (a-b) We compare the proportion of mass (as measured by fraction of total predicted hematopoietic output) by lineage and age in terminal fates for (a) PBA and (b) StatOT to the proportion of cells that reside in annotated clusters of differentiated cell types (observed probability mass). The proportion of mass derived from PBA and StatOT represent the expected proportion of cells ending up with a particular lineage fate, though individual cells may have the potential to acquire multiple fates. Agreement between observed and expected probabilities were used as a metric for evaluating performance of the probabilistic algorithms.

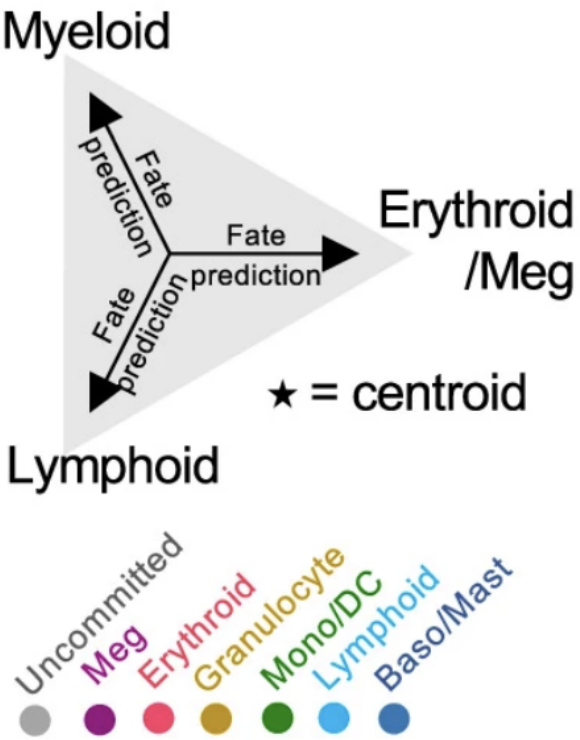

Figure 2.7: Schematic of triangle fate probability plots that serve as output from diffusion-drift modeling. Individual cell fate lineage probability is defined based on location relative to lymphoid, myeloid (including monocyte), and erythro-megakaryocytic fates. Centroid (*) indicates the expected average HSPC probability. Also displayed is color scheme for individual cells, which is based on cell type cluster category from Louvain clustering

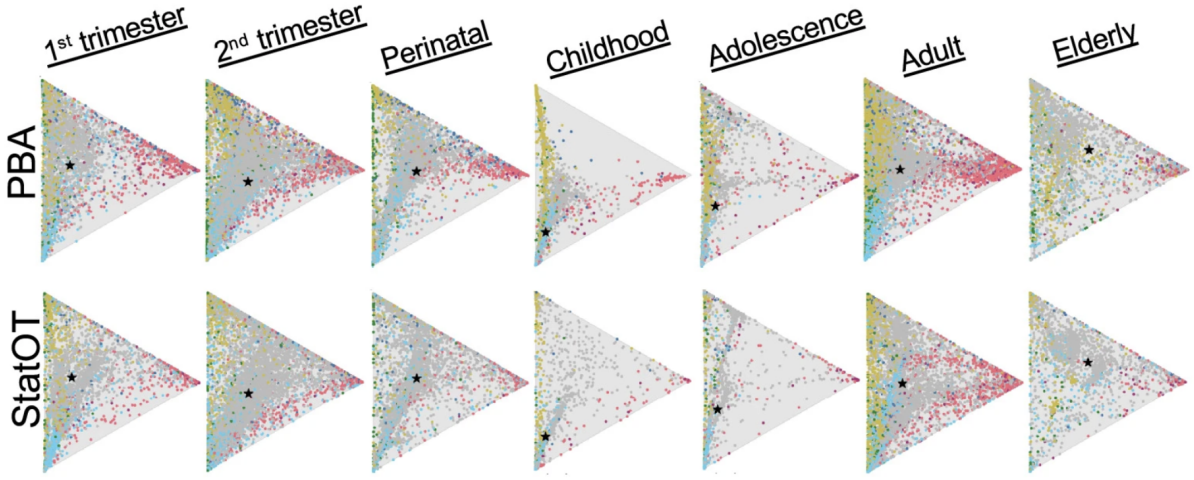

Figure 2.8: Individual cell lineage commitment probabilities to lymphoid, myeloid, and erythro-megakaryocytic fates for HSPCs in each age phase by population balance analysis (PBA) and stationary optimal transport (StatOT).

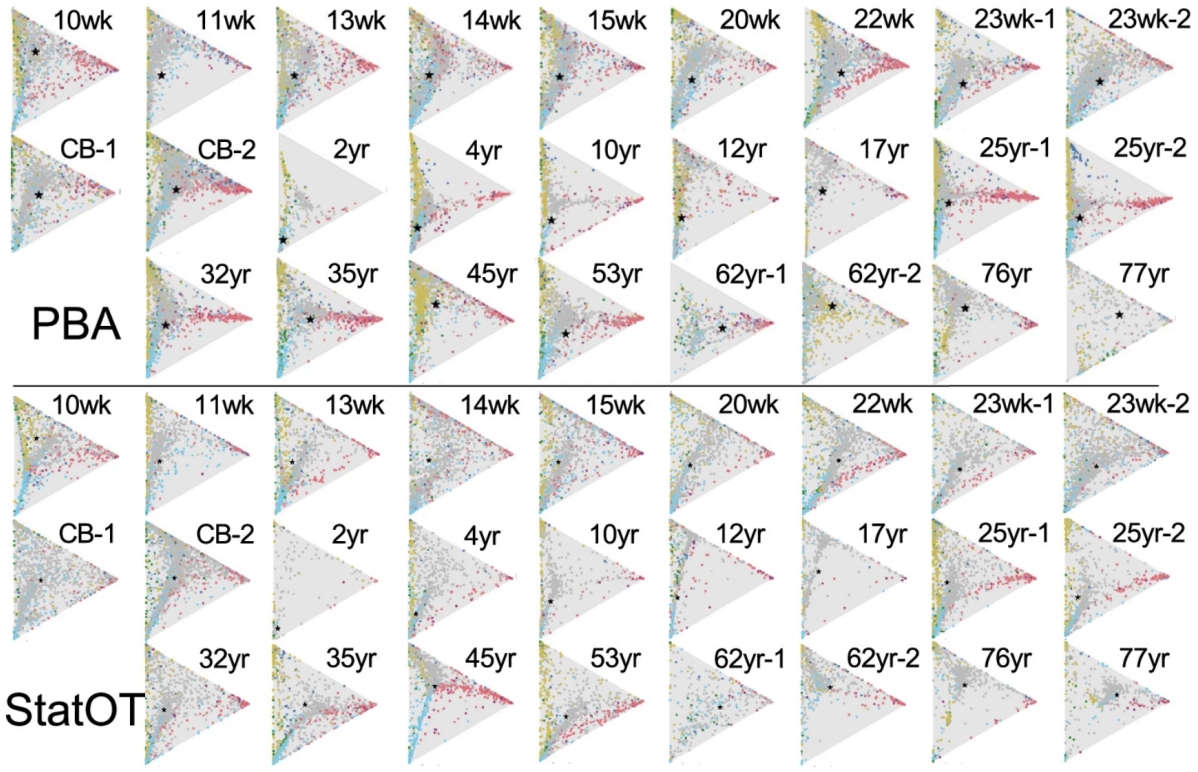

Figure 2.9: Individual cell lineage commitment probabilities for HSPCs in all 26 donors in the dataset.

2.3.2 Age-specific mechanisms of lineage commitment

To uncover mechanisms driving age-related shifts in hematopoiesis, we used consensus non-negative matrix factorization (cNMF) to define gene expression programs (GEPs) associated with lineage commitment (Fig. 2.10). These lineage GEPs included both known and previously unrecognized genes, validated by external datasets (Fig. 2.11). GEP usage varied across life stages, with lymphoid programs peaking in childhood, myeloid programs expanding in adulthood, and erythroid and megakaryocyte programs increasing in elderly HSPCs.

We next examined how lineage commitment programs varied with age by analyzing the expression dynamics of lineage-associated genes. While core GEPs were broadly used across lifespan, additional auxiliary programs were activated in an age-specific manner—such as a childhood-specific lymphoid program and a fetal-specific monocyte program (Fig. 2.12).

This suggests that progenitor lineage biases emerge through both shared and age-specific transcriptional programs.

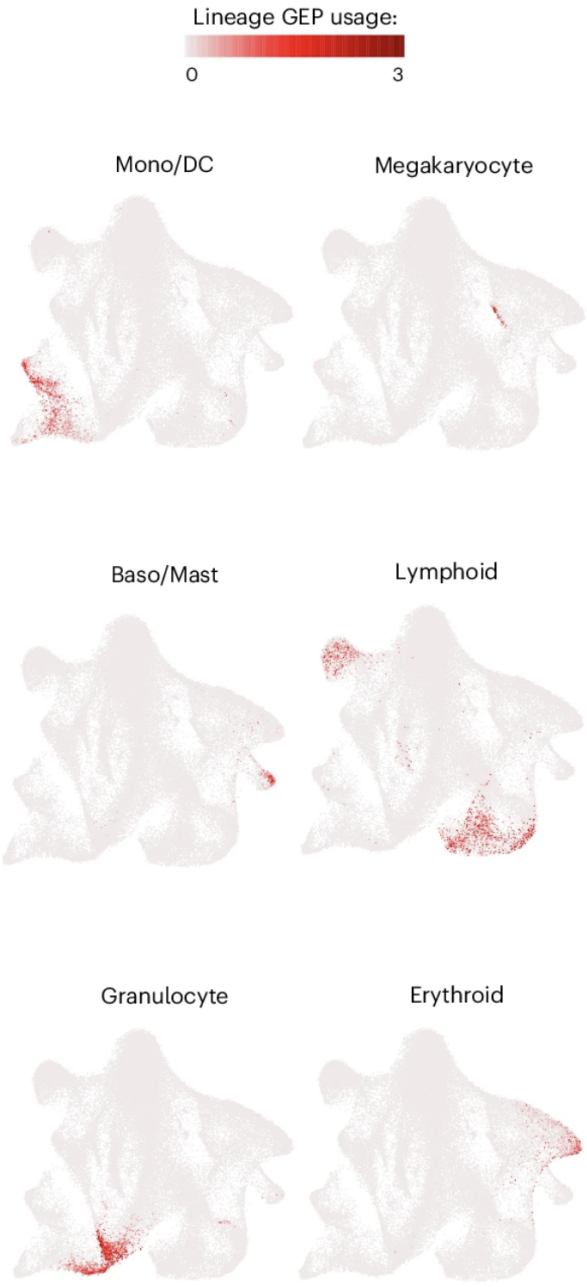

Figure 2.10: GEPs specific for the indicated lineages were identified by cNMF, and the usage of each GEP by each cell in the dataset was plotted in UMAP space. The relative expression color scale applies to all GEPs. Mono/DC, monocyte/dendritic cell; Baso/Mast, basophil/mast cell.

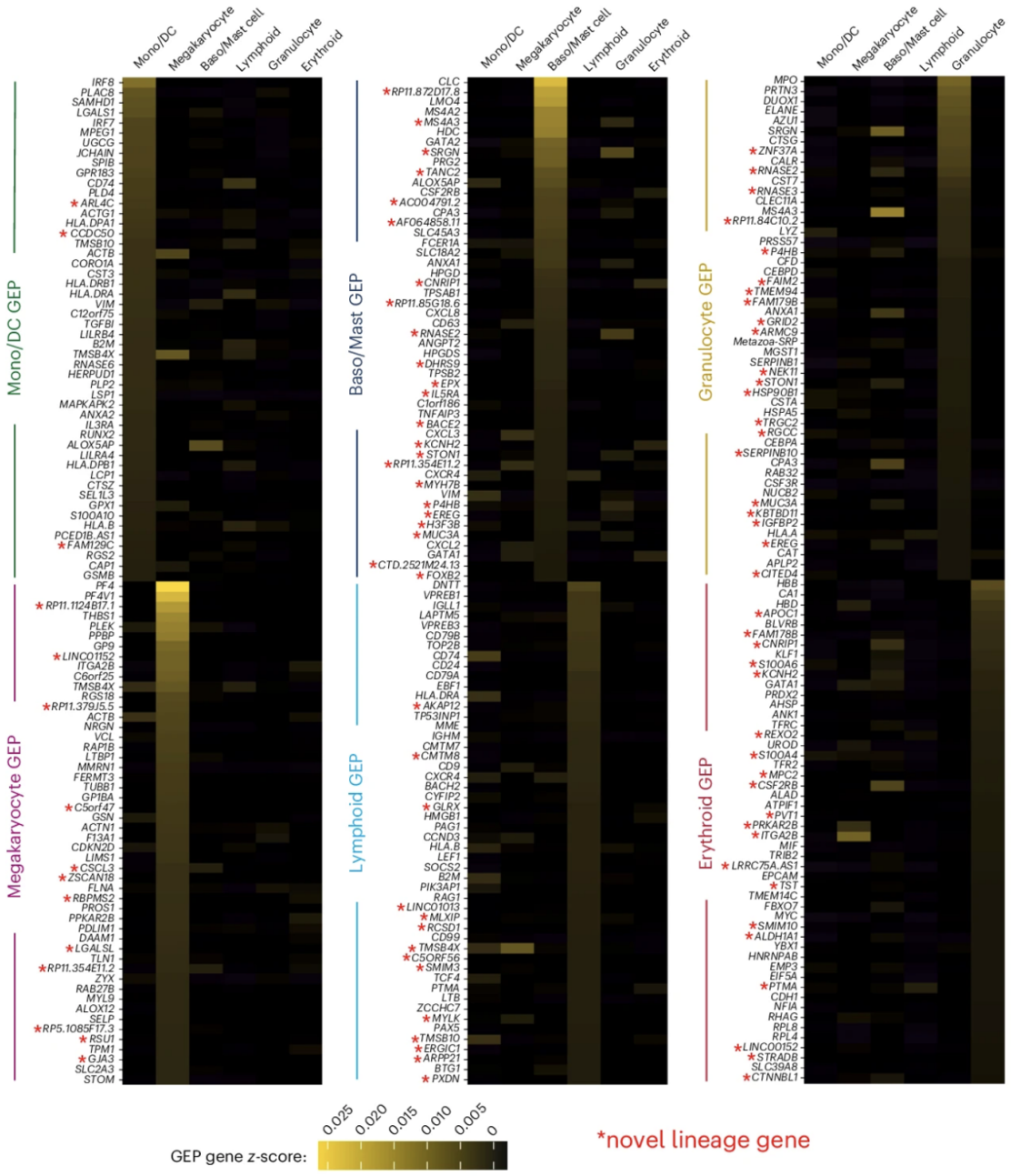

Figure 2.11: The top 50 genes contributing to each GEP are displayed as a heat map, with the color legend indicating the z-score for gene contribution to each GEP. The genes denoted with a red asterisk are novel genes associated with a given lineage, as defined by absence of any prior publications implicating that gene in lineage differentiation.

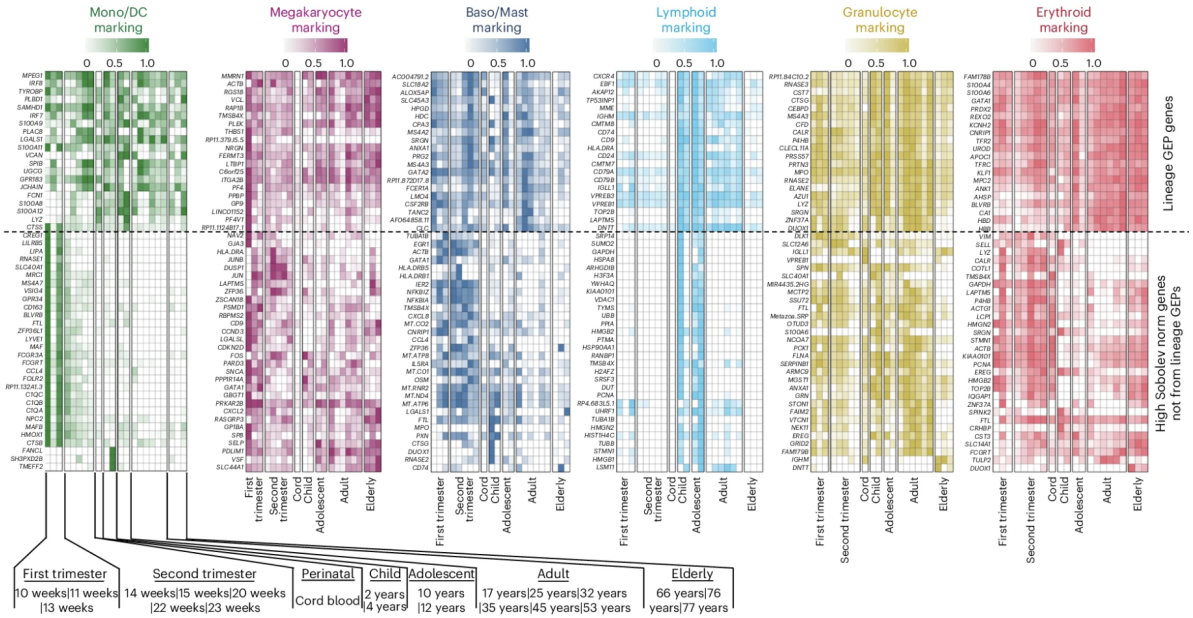

Figure 2.12: The relative strength with which genes mark commitment to the indicated lineages in each sample was determined by correlating gene expression with fate probability calculated from PBA and StatOT. Top: the top 20 genes from each lineage GEP are displayed. Bottom: the 30 genes with the highest Sobolev norms for variability in lineage-marking throughout human life. The color-coded heat maps for relative strength of commitment marking for each lineage is immediately under each lineage label.

To map lineage bifurcations over time, we calculated anticorrelations between lineage GEPs in multipotent progenitors (MPPs) (Fig. 2.13a). Early in gestation, monocyte/dendritic cell and basophil/mast cell programs diverged; postnatally, lymphoid and granulocyte programs bifurcated, aligning with the shift toward lymphoid output in childhood and myeloid output in adulthood (Fig. 2.13b). In elderly progenitors, significant lineage bifurcations were lost, suggesting reduced lineage restriction. Single-cell lineage assays confirmed that elderly progenitors exhibit more multilineage potential than adult progenitors (Fig. 2.13c).

To quantify lineage priming at the HSC level, we developed a "unilineage dominance" score based on GEP usage . Applying this to HSCs across ages, we observed a shift from monocyte/dendritic cell priming in fetal HSCs to lymphoid priming in childhood and adolescence, and a later shift toward myeloerythroid priming with age (Fig. 2.14a). Fetal HSCs exhibited the least lineage bias but highest overall lineage program usage, suggesting

efficient, unbiased lineage access. Conversely, elderly HSCs exhibited stronger single-lineage biases but lower overall lineage engagement (Fig. 2.14b,c).

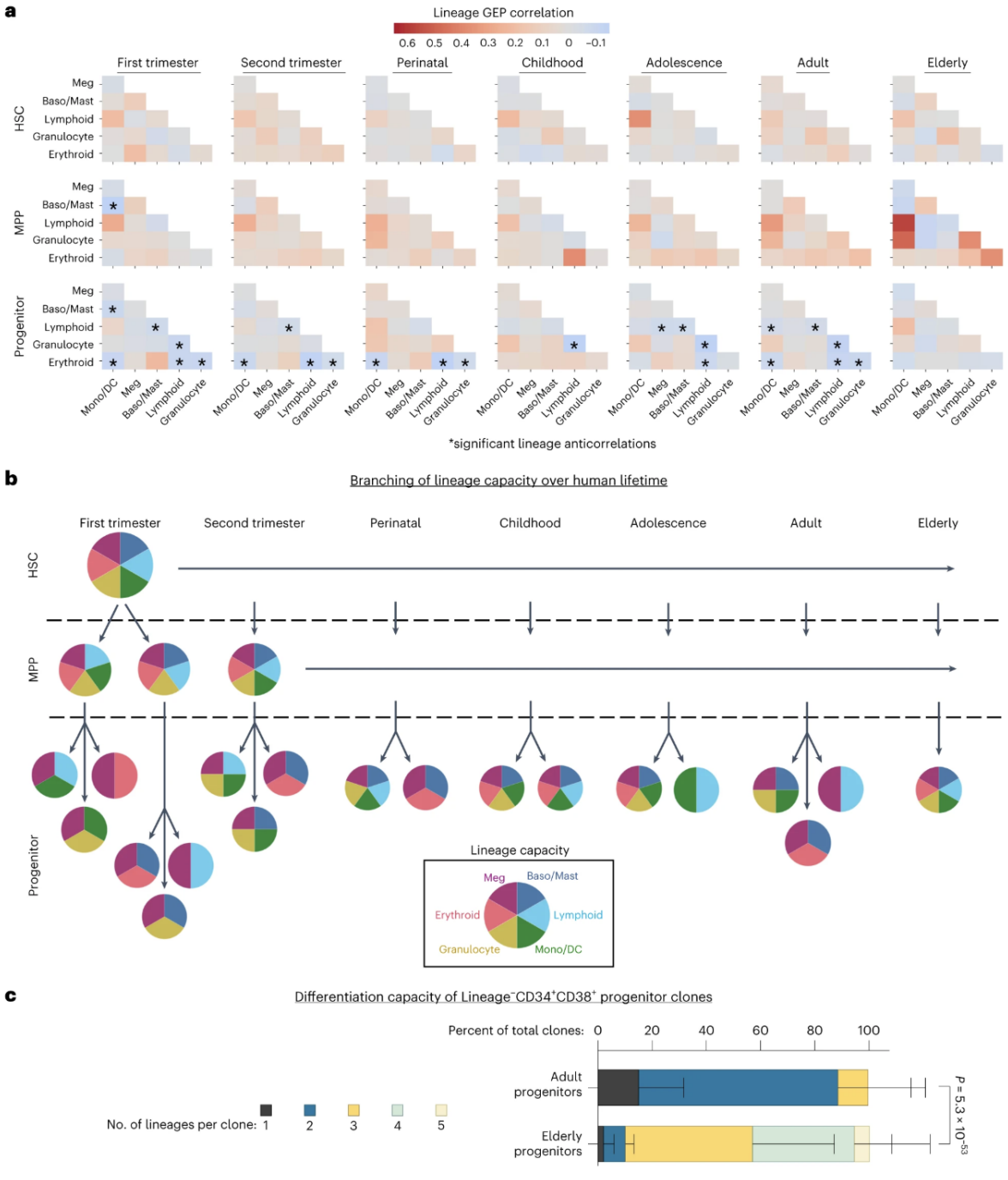

Figure 2.13: **a**, Based on Louvain clustering from Extended Data Fig. 3a, all cells in the study were categorized into the HSC cluster, MPP clusters and later stage cells (progenitor). Within each group of cells and each age group, Pearson correlation coefficients were calculated for all pairwise comparisons between the six lineage GEPs and displayed as comparison matrices with degree of correlation indicated by the indicated color intensity scale. Bonferroni-corrected P values of the two-tailed Pearson correlation coefficient were also calculated, and anticorrelations with a corrected P value < 0.05 are indicated with an asterisk. The correlations with a corrected P value > 0.05 are in gray, whereas correlations and anticorrelations with a corrected P value < 0.05 are in shades of red and blue, respectively. Meg, megakaryocyte; Mono/DC, monocyte/dendritic cell; Baso/Mast, basophil/mast cell.
b, Based on the significant lineage anticorrelations identified in a, a schema for differential hematopoietic lineage branching at various ages of human life is presented.
c, Functional assessment of decreased lineage restriction in progenitors during the adult-to-elderly transition was performed by placing individual Lineage$^-$CD34$^+$CD38$^+$ progenitor cells in clonal outgrowth assays capable of supporting five lineages (lymphoid, monocyte/dendritic, granulocyte, erythroid and megakaryocyte). The number of lineages present in each single-cell-derived clone was assessed by flow cytometry, and the fraction of clones with the indicated number of lineages presented is displayed. The results are the combination of two donors for each age. The error bars denote standard deviation for each fraction. Significance testing for the difference between the two distributions was performed using the upper one-tailed two sample χ^2 distribution test ($n = 73$ adult colonies and 75 elderly colonies), tested with two donors for each age over two independent experiments. The data are presented as mean $\pm$ standard deviation for each outcome.

2.3.3 Age-specific HSC states

Given the evidence for transcriptional reprogramming of HSCs with age, we identified two age-enriched HSC gene expression programs: a fetal-HSC-GEP and an elderly-HSC-GEP (Fig. 2.15a–c). The fetal program included activation markers such as *FOS* and *JUN*, while the elderly program featured ribosomal genes associated with proteostasis decline (Fig. 2.15d).

We found that CD69 marked the fetal-HSC-GEP and identified a subset of midgestation HSCs with enhanced clonogenicity and multilineage potential (Fig. 2.15f,g). Xenotransplantation confirmed that CD69+ fetal HSCs have superior long-term engraftment capacity compared to CD69- cells (Fig. 2.15h). Thus, we uncover a previously unrecognized midgestation-specific HSC state with functional significance.

2.3.4 Age-specific HSPC states in leukemia

We applied our HSPC atlas to classify leukemia single cells and bulk RNA-seq samples. AMLs predominantly resembled myeloid progenitors, while B-lymphoblastic leukemias resembled lymphoid progenitors.

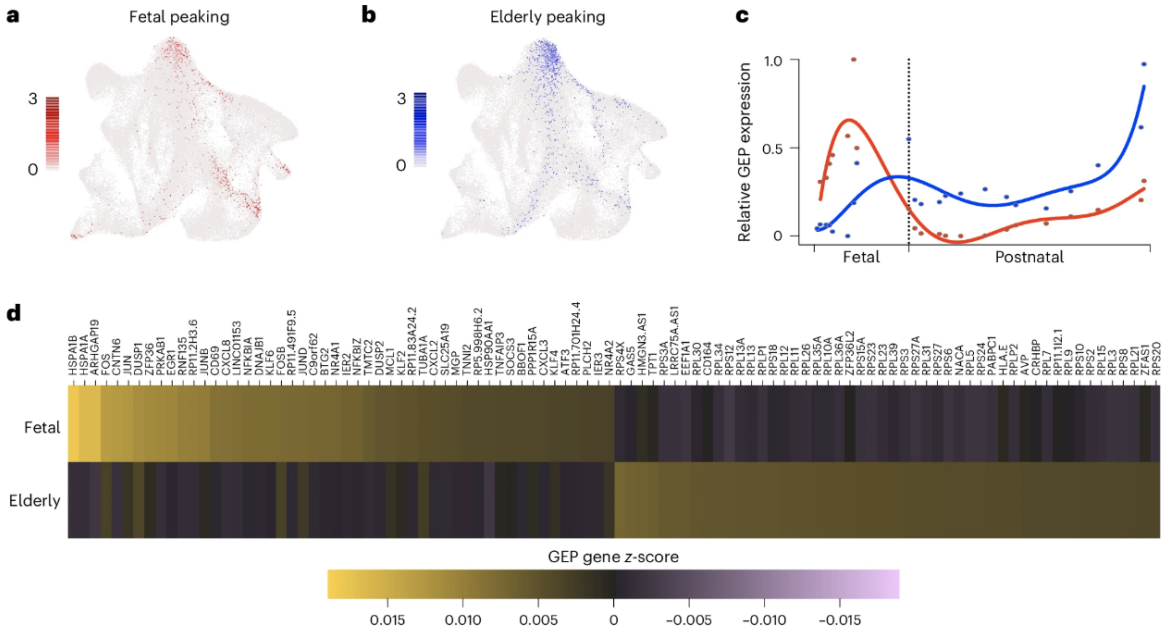

Figure 2.14: cNMF was used to identify GEPs biased toward HSC/MPP utilization, and their localization in UMAP space is displayed for a fetal-HSC-GEP (a) and an elderly-HSC-GEP (b). A UMAP color scale indicates log-transformed GEP utilization score in each cell. The temporal dynamics of fetal-HSC-GEP (red) and elderly-HSC-GEP (blue) utilization are depicted (c). A heat map depicting transcripts marking the fetal-HSC-GEP and elderly-HSC-GEP, with the color legend indicating the z-score for gene contribution to each GEP (d).

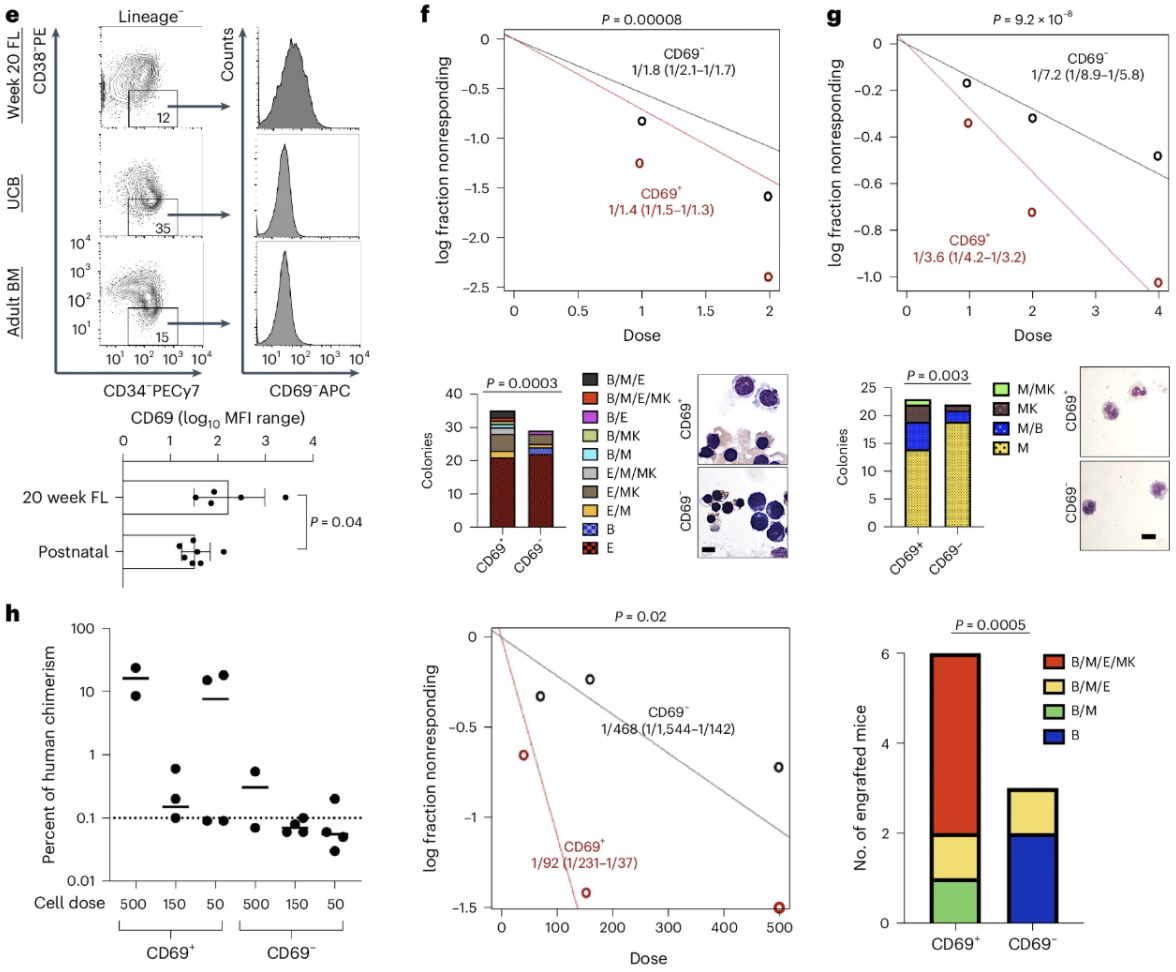
e
Lineage⁻
Week 20 FL
UCB
Adult BM
CD38⁻PE
Counts
CD34⁻PECy7
CD69⁻APC
CD69 (log10 MFI range)
20 week FL
Postnatal
P = 0.04
f
P = 0.00008
CD69⁻ 1/1.8 (1/2.1–1/1.7)
CD69⁺ 1/1.4 (1/1.5–1/1.3)
log fraction nonresponding
Dose
P = 0.0003
Colonies
B/M/E
B/M/E/MK
B/E
B/MK
B/M
E/M/MK
E/MK
E/M
B
E
g
P = 9.2 × 10⁻⁸
CD69⁻ 1/7.2 (1/8.9–1/5.8)
CD69⁺ 1/3.6 (1/4.2–1/3.2)
P = 0.003
M/MK
MK
M/B
M
h
Percent of human chimerism
Cell dose
500 150 50 500 150 50
CD69⁺
CD69⁻
P = 0.02
CD69⁻ 1/468 (1/1,544–1/142)
CD69⁺ 1/92 (1/231–1/37)
P = 0.0005
No. of engrafted mice
B/M/E/MK
B/M/E
B/M
B

Figure 2.15: **e**, Flow cytometry analysis for the indicated markers at each indicated age state with relative range of CD69 signal quantified (results comparing FL with postnatal cord blood (UCB) and bone marrow (BM) samples by a two-tailed Student's t-test with the P value shown. The data are presented as mean ± standard deviation; $n = 5$ for 20 week and seven postnatal biologic replicates). MFI, mean fluorescence intensity; PE, phycoerythrin; APC, allophycocyanin.
f,g, $CD69^+$ or $CD69^-$ $Lineage^-CD34^+CD38^-$ cells were sorted onto MS5 stroma with cytokines (recombinant human SCF, TPO, FLT3L and IL-7, with (f) or without (g) EPO) at 4, 2 or 1 cell(s) per well. After 4 weeks, the colony outgrowths were scored, and the clonogenic stem cell frequency calculated by limiting dilution analysis. The results pooled over two independent experiments from different donors (for f, $\chi^2 = 15.6$, d.f. of 1, $P = 0.00008$; for g, $\chi^2 = 28.3$, d.f. of 1, $P = 9.2 \times 10^{-8}$). Single-cell-derived colonies were picked and lineage outcomes analyzed by flow cytometry, with the representative morphology shown comparing observed and expected unilineage versus multilineage colony outcomes for $CD69^+$ compared with $CD69^-$ distribution (for f, $\chi^2 = 12.8$, d.f. of 1, $P = 0.0003$; scale, 10 μm; for g, $\chi^2 = 8.8$, d.f. of 1, $P = 0.003$; scale, 10 μm).
h, $CD69^+$ or $CD69^-$ $Lineage^-CD34^+$ $CD38^-$ cells were transplanted at indicated cell doses into NSG recipients, and human chimerism and lineage outcome were analyzed at 12 weeks (for engraftment, $\chi^2 = 5.2$, d.f. of 1, $P = 0.02$; note that the $CD69^+$ 500 cell data point is below the scale; for lineage outcomes, $\chi^2 = 12$, d.f. of 1, $P = 0.0005$). B, B cell; M, myeloid; E, erythroid; MK, megakaryocyte. The percent human chimerism for each cell dose is shown. All χ^2 testing is two-tailed. Horizontal dashed line represents 0.1% human chimerism cutoff to define positive engraftment.

Interestingly, AML transcriptional age signatures decoupled from patient age. AMLs segregated into two classes based on similarity to fetal, pediatric, or elderly HSPC programs, independent of patient age. AMLs classified as "elderly-like" had better outcomes, whereas those with mixed age signatures had worse prognosis.

Finally, we identified seven transcription factors enriched between age classes that predicted AML prognosis, including known regulators (*ZFX*, *DLX2*) and novel candidates (*NPAS1*, *RORB*, *ARNT2*, *BNC2*, *NPAS3*). Validation in an independent TCGA AML cohort supported these findings.

Thus, transcriptional maturation state of leukemia cells, independent of patient age, can influence disease outcome.

2.4 Discussion

To address the lack of a concerted study profiling hematopoiesis throughout human life, we sequenced individual transcriptomes of primary human HSPCs from donors spanning nearly the entire duration of definitive human hematopoietic maturation and aging from early gestation through advanced adulthood. While human studies always carry a concern of background genetic heterogeneity, a strength of our dataset is how robustly it replicates well-characterized age-associated and physiologically appropriate shifts in lineage output [66–68].

Although the human hematopoietic system has been examined at single-cell resolution in dedicated studies of prenatal development, postnatal steady state hematopoiesis and aging, our work here highlights the challenges of attempting to integrate different studies profiling specific periods of human life. Indeed, batch effect from variable processing and sequencing methods proved persistent, leading to an inability to reproduce well-known lifetime hematopoietic trends. When looking individually at such prior studies, a survey of 7–17 PCW human hematopoietic tissues demonstrated a strong early erythroid bias that diminished during development with an increase in natural killer and B cell production and quiescence of HSCs. A multiomic approach to 17–22 PCW FL demonstrated that HSCs and MPPs showed a layer of lineage priming at the chromatin accessibility level not detectable at the transcriptional level. Interestingly, beyond 21 PCW, surface expression of CD69 has been detected on human FL HSPCs, although the specificity of this marker could not be concluded without other ages [69].

Postnatally, scRNA-seq was applied to neonatal HSPCs to resolve lineage potential in progenitor populations. Integration and comparative study of independently generated neonatal and adult scRNA-seq datasets suggested overall conservation of postnatal HSPC ontogeny [70]. Surveys of adult bone marrow datasets have challenged the concept that HSPCs are hierarchically organized by potency by suggesting that fate is acquired directly

from a pool of low-primed cells [71]. Additional studies in human adults have suggested a more classical HSPC ontogeny with progenitor-level lineage bifurcations [40, 72]. However, batch effects preclude combining studies spanning ages to resolve these discrepancies. Viewing our atlas from the perspective of transcriptional states, we observed sharp demarcation between fetal and postnatal specimens in terms of relative abundance of multipotent (HSCs and MPPs) and oligo/unipotent progenitors. By applying cNMF temporally, we uncovered lineage GEPs validated against independent benchmark signatures, which enabled reconstruction of temporal fate choice maps and models of fluid HSC potency and priming. Our data suggest temporal changes in dominant lineage priming in HSCs, though multipotency programs overall are generally preserved. Our data support a model that mechanisms of lineage commitment may differ over the lifetime, with a higher degree of lineage priming postnatally and more classical lineage bifurcations and maintenance of multipotency in HSCs and MPPs prenatally [73]. Accordingly, while we find that core lineage GEPs are consistently utilized during maturation and aging, we find that these GEPs cooperate with auxiliary programs of age-biased gene expression. This mechanism probably contributes to known age-related differences in mature blood cells.

Aging of the HSPC compartment is associated with enhancer remodeling as well as HSC and erythroid progenitor expansion and lymphoid progenitor diminishment [74]. We demonstrate that elderly HSCs are the least primed and lack discrete fate bifurcations in later progenitors. Moreover, the low level of priming in elderly HSCs is biased toward individual lineages. These data support a model where programming toward myeloerythroid output occurs at the level of HSCs with aging, but multipotency is maintained later in ontogeny, which may contribute to an increased risk of transformation. The lack of segregation of transcriptional signatures at this age also suggests a significant role for exogenous factors in lineage specification, such as inflammation and changes in niche composition [75]. These findings complement a study of age-related clonal dynamics in humans showing that HSC and MPP diversity decreases precipitously with age [76]. Age-related epigenetic changes have

also been suggested to parallel the epigenetic dysregulation of AML, providing a mechanism underlying the increasing incidence of AML with age [74]. By classifying AML transcriptomes against age signatures, we demonstrate that components of normal HSPC states are retained in leukemia. By delineating age-biased transcriptome signatures, we find that AMLs exist either in primarily elderly or mixed age states, a phenomenon with prognostic significance. Compared with existing studies comparing normal and malignant states, our consideration of development and maturation revealed that juvenile HSPC age states can be accessed in more aggressive AMLs. Together, these findings illustrate practical application of our dataset to blood disease.

Similar to all studies of human biology, our study is limited by the requisite use of donors with differing genetic backgrounds (with few childhood specimens) as well as variable numbers of single HSPCs captured. Despite these limitations potentially confounding comparisons of HSPC ontogeny across time [68], we observe gradual time-dependent shifts in HSPC abundance and consistency within age brackets despite comparisons across donors. Indeed, data extracted from independent studies support our finding that prenatal HSPCs can express CD69, and we validated the functional significance of this CD69-expressing population, which may play a key role in colonizing the nascent marrow [53].

When using our GEP expression data to define age-dependent branchpoints, a limitation is that our methodology may overestimate the true potency of progenitor states, as it does not account for predomination. However, our findings of anticorrelated lineage GEP usage coupled with our observation of age-variable lineage predominance in HSCs forms a strong premise for future confirmation of age-variable mechanisms of differentiation. This dataset can also serve as a basis for future trajectory analysis over real time to connect juvenile HSC states to progenitors at later stages of life.

Our finding that transcriptomic signatures of juvenile HSPCs correlates with unfavorable prognosis in AML raises additional questions for future study regarding the mechanisms by which normal HSPC signatures impact AML biology. It is important to note that we used

our dataset to define associations with AML prognosis in the BeatAML and TCGA cohorts, and further experimental validation will be required to mechanistically dissect what may drive causality. The TF associations we identified in our analysis may serve as a useful initial starting point for such investigations.

Overall, by profiling over 50,000 HSPCs, we define the changes that occur in human hematopoietic ontogeny over the course of the lifetime from development through maturation and aging. We demonstrate developmental stage- and age-specific mechanisms of HSC priming and differentiation and identify a fetal-specific human HSC state. We show that this dataset can identify ectopic age states in leukemia with potential impact on prognosis. Despite limitations of donor heterogeneity, our atlas serves as a foundation for future studies of human hematopoietic and immune development, aging and blood diseases, as well as a prototype for reconstructing the heterochronic control of the ontogeny of tissue stem cells.

2.5 Methods

2.5.1 scRNA-seq with InDrop

Cryopreserved FL, cord blood and bone marrow CD34-enriched cells were thawed and FACS-sorted for CD34+ cells. Sorted CD34+ cells were submitted to the Harvard University Single Cell Core for inDrop single-cell RNA library preparation. Single-cell RNA libraries were sequenced on an Illumina NextSeq 500. The raw sequencing reads were processed using the inDrop pipeline (https://github.com/indrops/indrops) using default parameters. The GRCh38 reference genome was used for alignment of sequencing reads.

2.5.2 UMAP visualization of scRNA-seq data

Using Seurat v3.0 (ref. 80), the UMAP embedding was generated on the basis of the same top 30 PCs of the integrated dataset40. The resulting cell coordinates were retrieved and used for

plotting gene expression intensities and other metrics of interest in the Seurat environment. The lineage scores were calculated for each lineage identified and for each cell in the integrated dataset by summing the normalized counts of genes within the functional signatures listed above. The cells with the highest scores in each lineage were colored according to lineage, with color intensity increasing with lineage score.

2.5.3 Cell cycle analysis and differential expression of lymphoid progenitors

To analyze differences cell cycle state between the lymphoid progenitor clusters, the integrated dataset was subset by cells assigned to these clusters. Using cell cycle markers derived from the literature44, each cell within the subset dataset was given continuous cell cycle scores and assigned a discrete cell cycle state (Seurat CellCycleScoring). The distribution of cells assigned to each cell cycle state was plotted for each lymphoid progenitor cluster. To generate the UMAP post cell cycle regression, we generated cell cycle scores using the CellCycleScoring function built into Seurat and regressed cell cycle scores from the dataset using the ScaleData function. Default parameters for the remainder of the standard Seurat pipeline were used to generate the UMAP. The original cluster calls were then overlayed on the post cell cycle regression UMAP coordinates.

2.5.4 Marker gene analysis in dataset subpopulations

To identify differentially expressed genes that positively marked specified subpopulations within the dataset, the cells within the specified population were compared with cells not within the subpopulation and tested for differentially expressed genes (FindMarkers) using a Wilcoxon rank-sum test.

2.5.5 Collection of the nine published datasets, batch correction using five methods, visualization with UMAP and Louvain clustering

After a thorough search of published research, we found nine studies with accessible scRNA-seq data of CD34+ cells from human FL, cord blood and/or adult bone marrow samples. We then chose five of the top performing published integration methods that could produce a batch-corrected count matrix to apply to the compiled set of published datasets31,32,33,34,35. The batch-corrected count matrix was then input into the standard Seurat pipeline with default parameters to generate visualize the cluster calls on the UMAP. To generate the lymphoid contribution to hematopoiesis over human lifetime plots, for each batch-corrected count matrix, we calculated the percent of cells at each age that fell into lymphoid clusters and fit a LOESS curve to the data.

2.5.6 SingleCellNet analysis

To assess the molecular similarity between HSPC states in our dataset and leukemia samples, we trained SingleCellNet classifiers on our scRNA-seq data43. Leukemia scRNA-seq datasets were imported into CellRouter for filtering83, and malignant cells were classified relative to our HSPC reference based on published annotations.

For transcriptome age classification, we annotated each HSPC cell by donor developmental stage and trained SingleCellNet models accordingly. These models were applied to bulk RNA-seq data from the BeatAML (257 samples) and TCGA AML (151 samples) cohorts to quantify transcriptional similarity to each age state.

BeatAML samples were hierarchically clustered by age similarity scores (Ward's method), and survival differences were assessed using Kaplan–Meier estimation. To refine prognostic prediction, we developed an age composite score by weighting each similarity score based on its association with survival in a Cox proportional hazards model. Samples were stratified

into low, medium, and high composite score groups, with survival differences evaluated by Kaplan–Meier analysis.

2.5.7 cNMF to identify GEPs

We used cNMF on the unintegrated dataset to identify and assign GEPs in our scRNA-seq data. cNMF is a technique that allows trends to be extracted from non-negative data, such as RNA count data51. We used the online implementation of cNMF available on GitHub (https://github.com/dylkot/cNMF), following the step-by-step guide available in the repository. By maximizing the combination of stability and error measures from cNMF implementation output, we used the parameters 35 components and a local-density-threshold of 0.15 (selected via inspection of the plots in Extended Data Fig. 5a). This generates 35 GEPs, with scores for each gene indicating how much they contribute to each of the GEPs. The scores can be accessed from the gene_scpetra_score file that is created when running the cNMF pipeline. The Euclidean distance between GEPs can also be accessed from the output files when running the cNMF pipeline. Furthermore, each cell is assigned 35 distinct usage scores indicating how much each of the GEPs contribute to that cell's expression pattern. To identify GEPs with significant biological contribution, we targeted our analysis on GEPs with normalized usage greater than 0.1 in at least 5% of cells in the dataset. We used the unintegrated dataset for GEP analysis throughout due to the need to access all detected genes. The full matrices of contribution of each gene to each GEP and usage of each GEP in each cell are available in Supplementary Data Table 1a,b. For cNMF analysis of the dataset in ref. 26, we performed the same procedure as on our dataset, but found that 21 components maximized the combination of stability and error.

2.5.8 Spearman correlation of GEPs

To compute the Spearman correlation between the GEPs, we used the gene scores output from the cNMF pipeline to construct a vector representation of each GEP. We then used this

vector representations to find the Pearson correlation between each pair of GEPs.

2.5.9 Visualization of lifetime dynamics of GEP utilization

Usage score for each GEP in each sample was normalized such that the sample with the highest GEP usage was set at 1. Normalized GEP usage score was then plotted as a function of time, and the polynomial regression of best fit for GEP usage versus time was plotted using Prism version 9 (GraphPad).

2.5.10 Pearson correlation coefficients to identify lineage bifurcation branching

To generate the plots of GEP correlations, the cells were first grouped based on the age of their source. Then, within each of these groups, each cell was treated as a datapoint with a GEP score for each of the GEPs. From this, the Pearson correlation coefficient was computed between pairs of GEPs for each age group individually using the cells' GEP score values. We identified all lineage GEP anticorrelations with significant (adjusted $P<0.05$) Bonferroni-corrected adjusted P values at the indicated levels of ontogeny and used those anticorrelations to signify branching bifurcations in lineage fate. We reconstructed the possible cell state conditions at each level based on these significant anticorrelations between two specific lineages. Although we could not exclude the possibility that other cell states might exist (that is, unilineage progenitors), we depicted the minimal cell states logically possible in this methodology to gain understanding of age-related patterns of lineage GEP segregation.

2.5.11 Lineage dominance score generation and statistics

Normalized usage scores for each of the six lineage defining GEPs were used to calculate a lineage dominance score in HSCs or progenitors, calculated in two possible ways: (1) the ratio of the maximum usage score in a cell to the average of the other five usage scores or

(2) the maximum usage score in a cell minus the average of the other five usage scores. A corollary data input type (instead of GEP usage) for the lineage dominance score was also used, consisting of the total number of cells for each lineage arising from a single HSC in a clonal output assay. The single dominant lineage in each cell was also recorded to classify HSCs on the basis of predominant lineage priming. When comparing the relationship of lineage dominance scores calculated by GEP usage and clonal output, lineage regression testing was performed on the median score for fetal, adult and elderly cells to generate a coefficient of determination. Lineage dominance scores in HSCs were compared across age groups using a Wilcoxon rank-sum test.

2.5.12 Total lineage priming score generation and statistics

The normalized usage scores for each lineage defining GEP were used to calculate a total lineage priming score, defined as the sum of all usage scores in a single cell. The total lineage priming scores in HSCs were compared across age groups using a Wilcoxon rank-sum test.

2.5.13 Definition of the AML age composite score threshold, identification of differentially expressed TFs in leukemia cohorts and identification of the TF score threshold

We iteratively tested values for an age composite score threshold to separate samples into groups of 'low' and 'high' age composite scores. Ultimately, we settled on the age composite score threshold that resulted in the most significant difference in survival outcomes according to the Kaplan–Meier estimator. A differential expression of 'high' age composite score samples relative to 'low' age composite score samples was performed on raw counts of bulk RNA-seq data using DESeq2. We filtered the lists of differentially expressed genes down to seven TFs that had not been shown to associate with AML prognosis, which we would use to define our TF score. The TF score is equal to the ratio of the sum of normalized expression of TFs over

expressed in the 'high' age composite score samples to the sum of normalized expression of TFs over expressed in the 'low' age composite score samples. We iteratively tested values for a TF score threshold to separate samples into groups of 'low' and 'high' TF scores. Ultimately, we settled on the TF score threshold that resulted in the most significant difference in survival outcomes according to the Kaplan–Meier estimator. The same analysis approach was used for both the BeatAML dataset and the TCGA dataset, with the exception that the TCGA dataset did not undergo differential expression testing and instead used the same seven TFs identified in the BeatAML analysis.

2.5.14 Computation of fate probabilities by PBA

PBA (*https://github.com/AllonKleinLab/PBA*) is a graph-based probabilistic methodology for inferring lineage trajectories from static snapshot data such as scRNA-seq [63]. PBA relies on a number of assumptions to allow for a constrained, unique output: cell expression dynamics are Markovian, absence of rotational gene-expression dynamics, and a large number of sampled cells. We apply PBA to each patient sample, from 10 weeks fetal gestation to 77 postnatal years old, to assess how the lineage trajectories or fate biases change with time.

The algorithm's primary inputs are the single-cell expression matrix (cell×gene), lineage-specific 'sink' matrix (cell×seven lineages) and the diffusion constant. From the expression profile, a kNN graph is formed (n-neighbors set at 20), which is used to propagate probability mass between cell states. In this framework, we define 'sinks' as the cells in the system that are the most terminally differentiated, where mass would be exiting the most. The lineage-specific 'sinks' are assigned on the basis of enrichment for well-known, predefined markers for each of our seven major lineages. The lineage 'sink' matrix ultimately defines the fluxes or the net rate of loss associated with terminally differentiated cell states for each lineage.

Lastly, the diffusion constant defines the level of stochasticity cell states have in traversing other more distant cell states. We chose a diffusion constant of 0.2, as the recommended

range from the PBA study was between 0.1–0.3, and we find this to produce the most stable results. Given the flux, the 'compute_fate_probability' function computes the probability that in a random walk a given cell state will be absorbed by nodes in the graph associated to each of the terminal sinks.

The initial fluxes for the sinks are set at 1/D, where D is the diffusion constant, and we iteratively modify the fluxes such that difference between the summed output fate probabilities per lineage and the observed fate probabilities, as defined by predefined lineage-specific cluster annotations, are minimized (Extended Data Fig. 4a,b, showing expected versus actual).

2.5.15 Optimal transport-based computation of fate probabilities by StatOT

Similarly to PBA, StatOT (https://github.com/zsteve/StationaryOT) models biological processes in equilibrium as a diffusion-drift process subject to cell division and death[49]. While both methods share a problem formulation, they take theoretically distinct approaches with stationary optimal transport using entropy-regularized optimal transport to infer dynamics.

As in the case of PBA, the inputs to StatOT include an expression matrix and assignments of sink cells and/or growth rates by lineage, and the output is a cell-by-cell transition matrix from which fate probabilities may be computed. Unlike the flux rates, which are not strictly enforced in the PBA method, the mass entering each sink in StatOT can be exactly controlled through sinks weights and does not rely on ad hoc flux tuning by the user in order for the overall fate probabilities to match the observed proportion of cells. The level of diffusion is controlled by the entropy regularization parameter ϵ, and the estimates of cell growth rates are provided to model the effect of cell death and division.

Using StatOT, we computed a transition matrix for each timepoint, assigning a total of 100 sinks proportional to the expected amount of terminal mass in each of the six lineages from a cluster-based annotation. For consistency, the same sinks were chosen as in the PBA analysis, and similar to PBA, we assumed no significant differences in growth rates in the

different lineages. The couplings were calculated in a 50-dimensional PCA space based on the top 2,000 highly variable genes. Based on empirical observations, we chose $\epsilon = \bar{C}$, the mean of the cost matrix.

2.5.16 Correlation coefficient of gene expression versus fate probability and quantification of variability via Sobolev norms

The genes with a greater than one read count in at least 3% of cells within lineage-marking clusters were selected for downstream analysis.

First, correlation coefficients of gene expression versus fate were computed for each measured stage of human lifetime. To calculate Sobolev norms, given a vector of correlation coefficients, with an entry for each measured stage of human lifetime, we first computed a cubic spline to provide a smooth approximation of the coefficients over time. We then computed a measure of the variability by numerically integrating the squared derivative of the spline over the time-interval. This approximates the integral of the squared first derivative, which is related to the Sobolev 2-norm. Note that the standard Sobolev 2-norm of a function is defined to be the sum of the L2 norm of the function's derivative and the L2 norm of the function itself. However, to measure variability, we use just the L2 norm of the derivative and refer to it as a 'Sobolev norm' in the main text.

For the marker gene tables, the first 20 consistent genes displayed were selected for having the highest z-scores for each lineage-marking GEP. The remaining 30 variable genes were selected for having the highest Sobolev norms within the lineage.

On the heat map, the first 20 consistent genes are ordered by their z-score within the lineage-marking GEP, while the remaining 30 variable genes are displayed in order of expected time. Here, the expected time is given by $\sum_t t c_{g,t}$, where $c_{g,t}$ is the normalized correlation coefficient for the gene and t is time. This can be interpreted as the average time at which the gene has a lineage-determining role.

Finally, the genes are colored at each time relative to the size of their own normalized

coefficients. As a result of this procedure, the heat maps show genes that may have a lineage-determining role at various periods of development.

Chapter 3

Somatic Mutation Denoising from full-length Single-Cell RNA-Sequencing Reveals known Cancer Associated Mutational Signatures and Clonal Markers

This chapter was adapted from a manuscript in prep for submission.

3.1 Abstract

Single-cell genomic technologies enable the identification and molecular characterization of clonal expansion in both normal and disease contexts. Robust profiling of somatic mutations in single cells is essential for characterization of evolution in diseases such as cancer that are riddled with mutations. That being said, somatic mutation profiling remains an underexplored area of single-cell genomics due to various technical challenges. Full length scRNA-seq assays,

such as Smart-seq2, are well suited for this task in comparison to other single-cell genomic assays, due to more even transcript read coverage and transcription-mediated amplification of the mutation loci. This data, however, still contains several technical artifacts that need to be filtered such as RNA-editing events, and recurrent sequencing error-prone loci. We employ a number of statistical and unsupervised machine learning filters to remove these technical artifacts. This method enables the de-novo detection of somatic mutations in single cells using full length scRNA-seq. Many of the mutations are validated using expected cancer mutational signatures as well as with matched WES data. We also demonstrate that mutations detected in scRNA-seq align with identified CNV states, as well as known clonal states in HSCs. Lastly, we demonstrate that we can study the relationship between different cell states in recurrent tumors and their exposure to certain mutation processes, using signatures detected in single cells.

3.2 Introduction

Identifying somatic mutations in individual cells is critical for studying clonal evolution, understanding cellular heterogeneity, and linking genotypic variation to transcriptional states in both normal development and disease progression [4, 6, 14, 77, 78]. Single-cell RNA-sequencing (scRNA-seq) offers a way to measure mutations alongside gene expression profiles at high throughput, but technical challenges, including sparse coverage, allelic dropout, and sequencing artifacts, complicate reliable mutation detection. Standard mutation calling approaches applied to scRNA-seq typically produce a large number of false positives, making it necessary to develop additional filtering strategies to extract biologically meaningful signals [79].

To address these challenges, we introduce a framework that denoises somatic mutation calls from full-length scRNA-seq data. Full-length protocols provide more uniform and complete transcript coverage compared to high-throughput droplet-based methods (e.g., 10x

Genomics), making them better suited for reliable mutation detection [80, 81]. Our approach combines initial blacklist-based filtering with anomaly detection methods to enrich for high-confidence somatic mutations. Candidate variants are scored based on their deviation from typical sequence contexts of mutations detected in normal cells, allowing the identification of anomalous variants that are more likely to represent true somatic events. Importantly, this approach enriches for mutations associated with real, expected cancer mutational signatures, rather than technical noise, as well as for "ground truth" mutations detected in matched WES. The resulting filtered mutations show strong concordance with known mutational processes and provide a robust set of mutations for downstream phylogenetic analyses.

In parallel, we contribute methods for assessing phylogenetic signal in single-cell mutation datasets. Although sophisticated likelihood-based phylogenetic methods exist, they often produce trees that are difficult to interpret in the presence of noisy data. To complement these approaches, we introduce fast, exploratory techniques based on modified distance metrics, enabling researchers to quickly evaluate whether meaningful clonal structure is present in their data and to select more phylogenetically informative mutations for downstream analysis. These methods provide a practical way to prioritize mutations and samples before applying more sophisticated phylogenetic reconstruction models. Together, these approaches advance efforts to recover high-confidence somatic mutations and uncover clonal structure from single-cell data, enabling integrative studies of genetic heterogeneity and transcriptional programs in both normal and disease contexts.

3.3 Results

3.3.1 scRNA Mutation Detection Pipeline (scRNA-MuTect2)

A number of mutation calling pipelines have been previously described for calling mutations from both single-cell and bulk RNA-sequencing data. RNA-MuTect was designed for detecting mutations from bulk RNA-sequencing, leveraging MuTect's logarithm of the odds (LOD)

based mutation detection algorithm, coupled with a suite of filtering steps for removing RNA-specific false positives [79]. Other methods (Monopogen, Scomatic) have been recently contributed for calling mutations specifically from single-cell sequencing data [15, 16]. However these approaches make different cell intrinsic or population-level based assumptions about the characterization of sequencing noise that do not always apply, resulting in sequencing calls that do not produce expected mutation signature spectra. For instance, Scomatic's core reliance on annotation of cell types is rooted in an assumption that mutations shared across cell types are either noise or germline mutations. However, this assumption becomes more nuanced when handling samples largely enriched for malignant cells, where malignant cell states do not necessarily correlate with mutation profile.

scRNA-MuTect2 was designed as an extension of RNA-MuTect, employing many of the same filters as in RNA-MuTect, but using the MuTect2 engine, to allow for detection of indels and cancer associated indel signatures in addition to SNVs (Fig. 3.1). We generated single-cell specific Panel of Normals (PoNs) using normal CD45+ immune cells, following the PoN procedure that uses the beta distribution to compare the similarity between artefactual mutations found in normals and candidate mutations found in cancer samples [82]. This procedure generates an empirical similarity score between the alternative read counts in normal cells and the malignant samples at each detected genomic loci (see Methods). We then applied a number of blacklisting filters to remove other common sources of false positives. We utilized the RADAR & DARNED databases to blacklist known RNA-editing loci. We also blacklisted sites that overlap with loci found in gnomAD to account for germline contamination [83]. Since, the majority of mutations that are found in 1 cell are likely to be sequencing noise, we applied a consensus filter that retained mutations that were found in at least 3 cells. We also applied a homopolymer filter, removing variant loci that with homopolymer length ≥ 4 (2 alt alleles on 5' & 3' side), to account for error-prone polymerase slippage.

We carefully selected cancer Smart-seq2 samples that were strongly expected to contain mutational processes that could be picked up from the mutation calls. Thus, we applied the

pipeline to a number of samples: two TMZ treated, hypermutated adult glioma samples (IDH-mut glioma & IDH-wt glioblastoma (GBM)), two lung adenocarcinomas (LUAD) samples with smoking history, one microsatellite instability (MSI) colon cancer cell line HCT116 and two blood colony samples with matched DNA-based phylogenies [78, 84]. Each of these Smart-seq2 samples were carefully chosen, as they were expected to have mutational signatures that we expect to find in their respective tumors and/or they contain ground truth information in matched WES variants or clonal assignment for cells. The decision to focus on full-length, Smart-seq2 samples relied on the fact that this library preparation protocol is better suited for mutation detection, given the even coverage of reads across transcripts compared to 10x's 5' or 3' transcript coverage bias. We demonstrate this difference in mutation detection capacity in matched 10x and smart-seq2 glioma samples (Fig. 3.2). In 10x samples from the same patient each cell captures $\sim$1 mutation on average, while the Smart-seq2 samples capture on the order of 100s of mutations per cell. Furthermore, the Smart-seq2 mutation calls captured 2371 of the 9301 somatic mutations called in WES and the 10x samples, despite capturing thousands of cells, only captured 155 matched somatic mutations (Fig. 3.3).

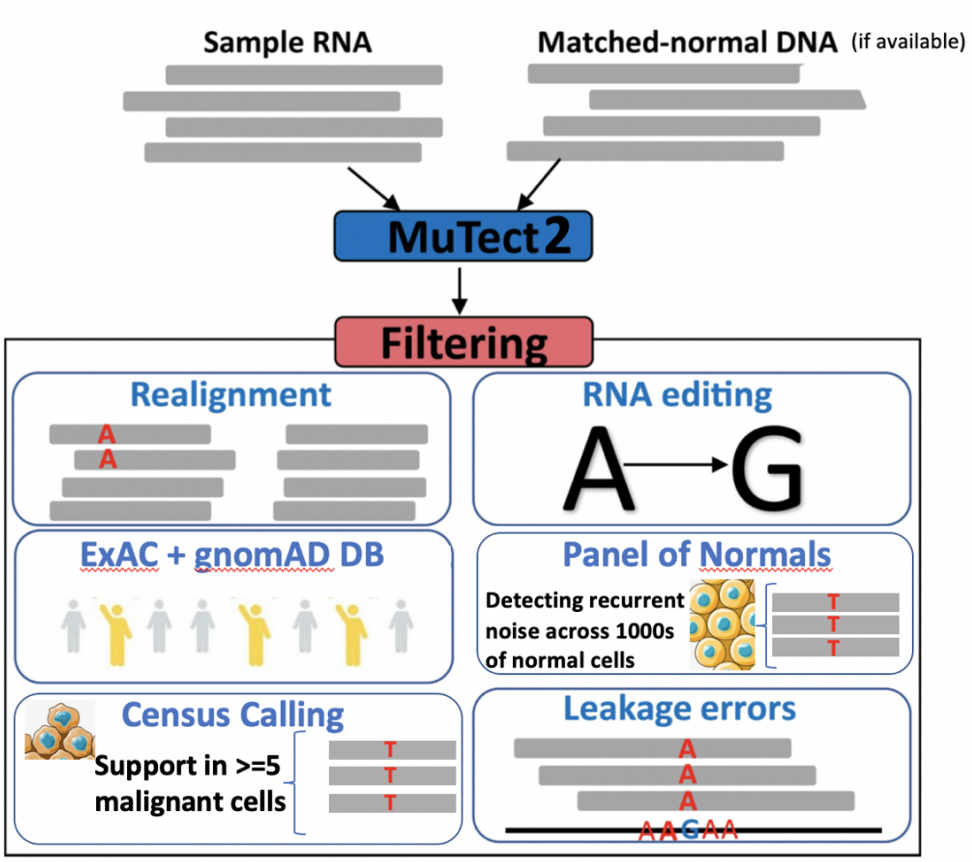

Figure 3.1: Initial set of analytical steps for getting starting set of candidate somatic mutations. Adapted from [79]

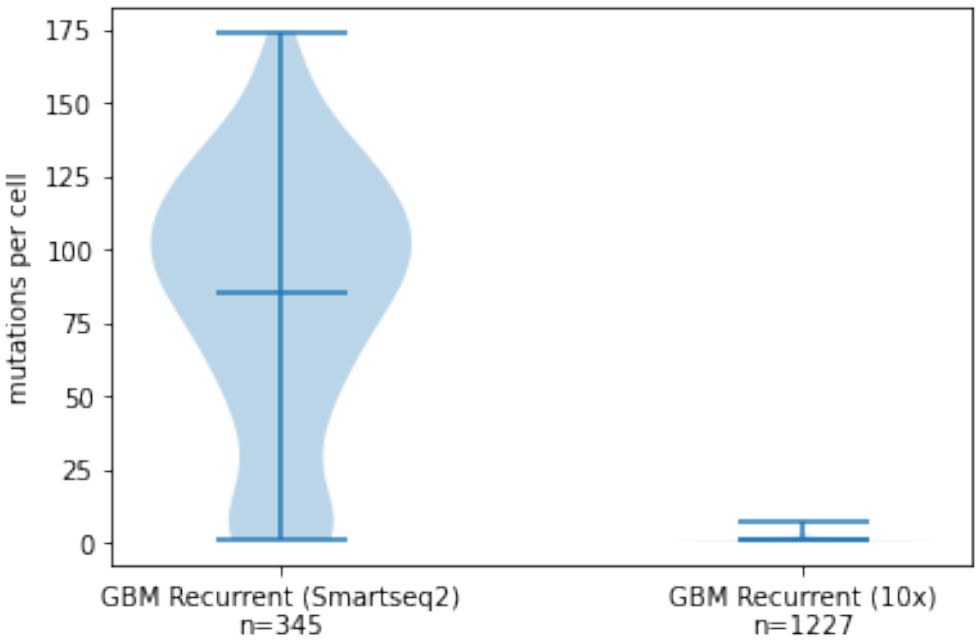

Figure 3.2: Smart-seq2 captures significantly more mutations per cell than 10x in matched GBM sample. Smart-seq2 captures more mutations per cell, despite capturing less cells.

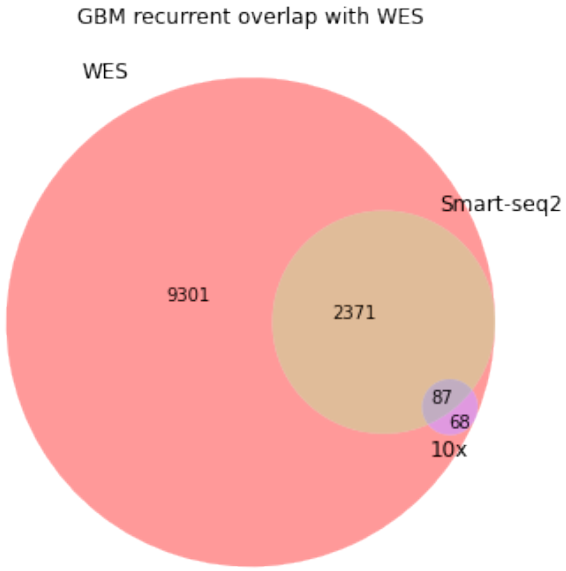

Figure 3.3: Smart-seq2 captures orders of magnitude more "ground truth" WES somatic mutations than matched 10x sample from GBM patient.

To capture the artefactual mutation profile from the Smart-seq2 library prep procedure, we called mutations from CD45+ immune that were processed similarly to other malignant

samples we analyzed. The spectra was a C>T, T>C dominant one that shared resemblance to other sample's spectra pre-filtering (Fig. 3.4). Upon applying the aforementioned filters to minimize technical noise in the LUAD scRNA-seq samples, we suspected that we were still being overwhelmed with technical noise. The profile shared resemblance to the suspected artefactual spectra we identified in the normal cells (Fig. 3.5). The mutation signatures we observed did not reflect the expected smoking signature, given the smoking history of the patients that were sampled (Fig.3.5). Given that mutations profiled in normal immune cells should be largely artefactual noise, coupled with the lack of smoking signature signal in the LUAD samples, we suspect that this profile is a Smart-seq2 library preparation specific noise profile. This suspected artefactual spectra features prominent C>T and T>C peaks primarily, with pronounced peaks at C>Ts in GpC contexts and T>Cs in CpC, CpT contexts.

Since we observe, from the mutation spectra, that the majority of the mutations called from single-cells are likely noise, we implemented a class of algorithms that are designed to separate outliers from the majority of data. Anomaly detection is a class of unsupervised learning approaches that learns the feature spaces that separate outliers from the majority of the data or "inliers". In this application, we assume that the majority of mutations called from scRNA-seq are noise that follow a distinct sequence context pattern from artefactual variants. We employ two different anomaly detection approaches, local outlier factory (LOF) and one-class support vector machines (one-class SVMs) to separate the candidate mutations based on their sequence contexts (see Methods). We use the surrounding base sequence as features for separating the data, where the majority of data that share patterns, or "inliers" , are flagged as noise, while the "outliers" that are distinct from the majority sequences are retained as candidate variants. This filtering step proved most effective in denoising our variant calls and enriching for the malignant mutational spectra we expect.

3.3.2 Validation (signatures, WES agreement, 10x vs. Smartseq2 comparison)

We validated our mutation calls primarily using the mutation spectra (single base substitutions: SBS, insertion/deletion spectra: IDS) that we expect to observe in the associated cancer as the "ground truth" to inform us as to whether we are enriching for real somatic mutations. In the cases where matched tumor WES was available, we checked for enrichment of the predicted outlier variants in the WES reads. Upon applying the set of filters described, in each of our samples, we were able to separate our mutational calls into two distinct sets of mutation spectra. The mutation spectra generated from the "inlier" set of mutations resembled more of an unknown artefactual mutational spectra signal, while the "outliers" resembled mutation signatures that we expected to see in the respective malignancies, particularly temozolomide (TMZ) signature (SBS11) in the glioma samples and smoking signature (SBS4) in the LUAD samples (Fig. 3.6, 3.7). We identified signs of smoking signature (SBS4) in both of our LUAD samples (TH226 & TH238) (with cosine similarities of 0.81) and signs of TMZ signature in our GBM samples (Fig. 3.6, 3.7). We also checked the matched GBM tumor WES reads for our mutation calls and found that the "outlier" mutation calls were significantly enriched for mutations that were supported in WES (at least 2 alternate allele reads in WES) (Fig. 3.8). The sensitivity or TPR, precision and F1 score for this sample was 0.698, 0.629 and 0.66, rivaling the performance of state-of-the-art methods, while further filtering for the expected spectra of TMZ (Fig. 3.8, 3.9).

Intrigued by its robust filtering procedure, which involves SVM read level filtering based on features learned from mutations found in population databases versus mutation that do not overlap, we implemented Monopogen on one of our LUAD samples [16]. The output spectra strongly resembled the artefactual spectra we saw prior to our anomaly detection filtering, despite using the recommended LD refinement and SVM prediction score thresholds. This could be due to a number of reasons, such as not explicitly taking into consideration sequence

context in the positive and negative sets. The read-level features may not be sufficient to separate population-supported SNV from de novo SNVs in various cancer types. Lastly, one could have a severely imbalanced training set of mostly noisy de-novo SNVs, skewing the prediction ability.

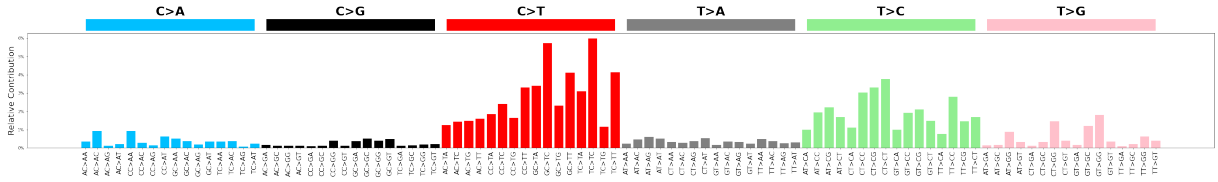

Figure 3.4: Mutation spectra of 5000 CD45+ normal cells processed with same Smart-seq2 protocol as malignant cells reveals suspected artefactual Smart-seq2 specific mutation profile

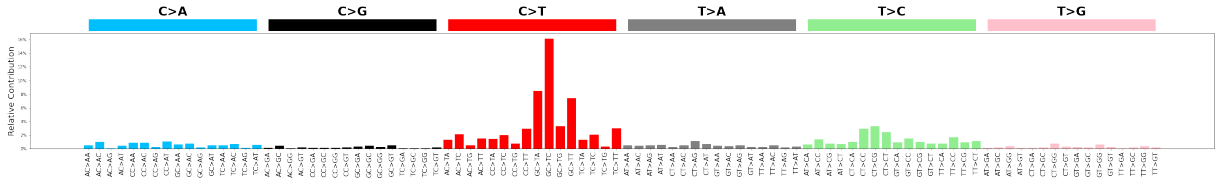

Figure 3.5: C>T dominant mutation spectra of LUAD sample with minimal filtering and prior to applying anomaly detection. Lack of expected SBS4 smoking signal.

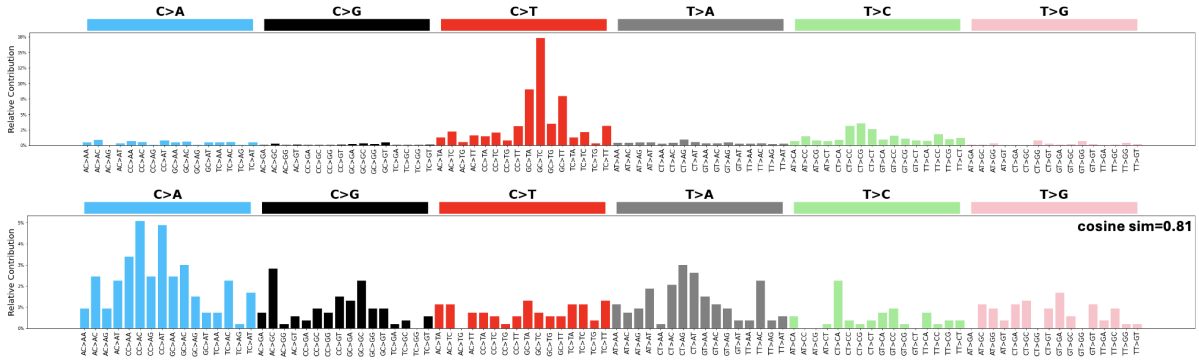

Figure 3.6: Application of anomaly detection to mutation candidates in LUAD samples separates mutations into distinct mutation spectra of "inliers" and "outliers." "Outliers" resemble SBS4 expected smoking signature with cosine similarity of 0.81.

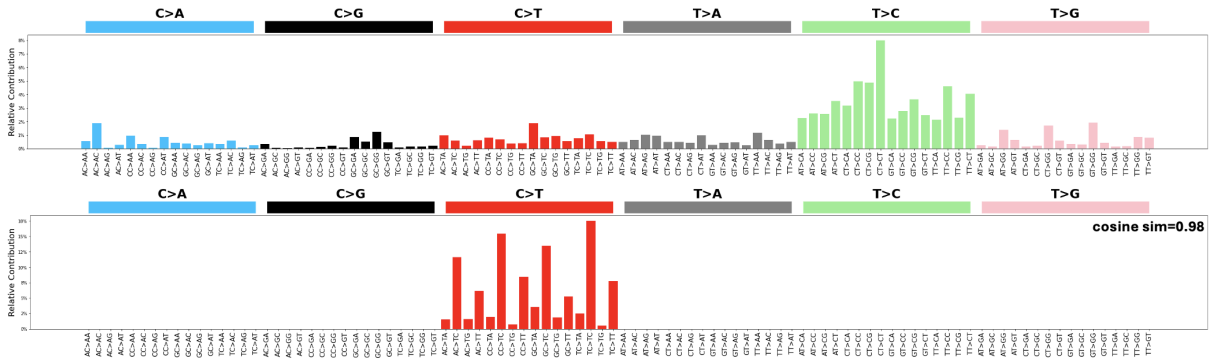

Figure 3.7: Application of anomaly detection to mutation candidates in GBM samples separates mutations into distinct mutation spectra of "inliers" and "outliers." "Outliers" resemble SBS11 expected TMZ signature with cosine similarity of 0.98

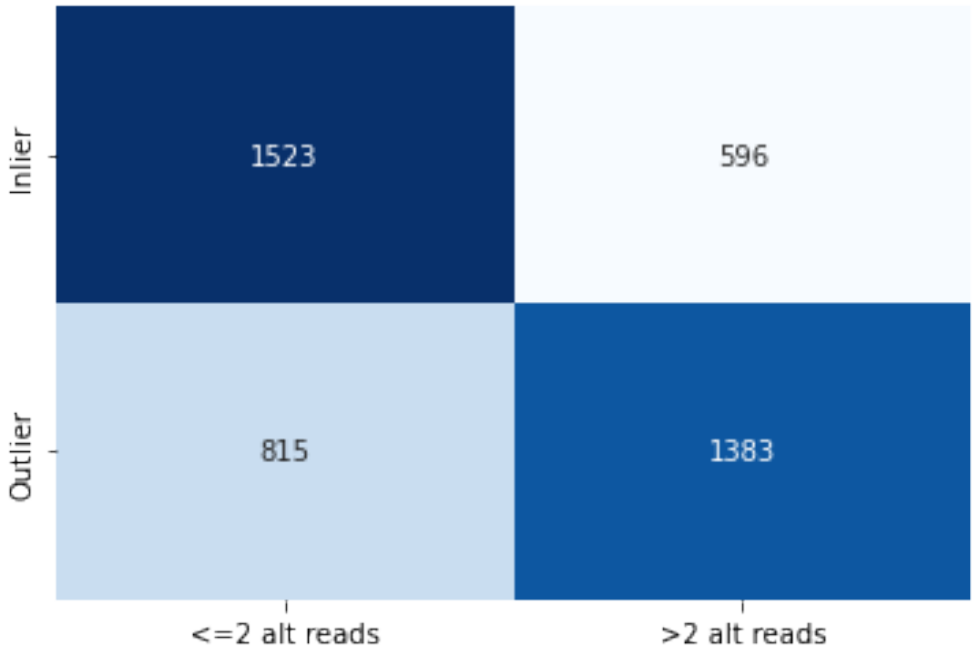

Figure 3.8: "Outlier" mutations are enriched for WES somatic mutations supported by alternate allele reads

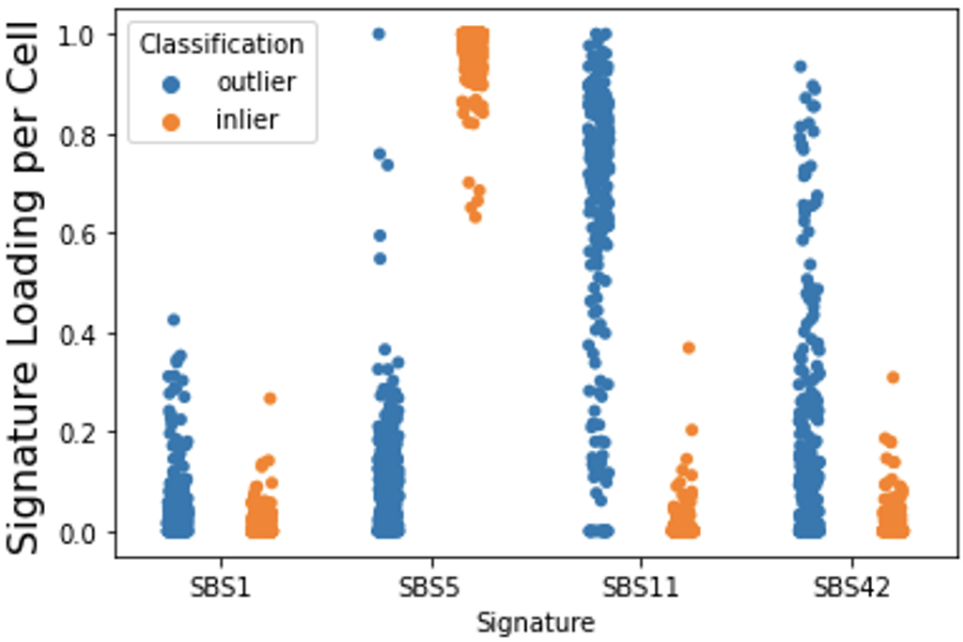

Figure 3.9: Supervised NMF reveals that "outlier" mutations have greater contribution of TMZ signature (SBS11) than the "inlier" mutations.

3.3.3 Clonal Concordance with Filtered SNVs

Next, we wanted to demonstrate some of the downstream capabilities of being able to enrich for high-confidence somatic mutations in single cells. The most obvious application of utility would be to leverage these mutations to identify specific clones in different tissue samples. In the blood colony data, where we have the ground truth DNA phylogeny, we found informative variants in RNA that aligned with clonal annotation. This demonstrates that, while difficult due to sparsity, lineage-informative mutations can be found in scRNA-seq, even in low mutational burden regimes like in normal cells (Fig. 3.10). In the IDH-wt GBM samples, we utilized the expression profile to infer large copy events and CNV-defined clones, using inferCNV [85, 86] (Fig. 3.11). We then identified 20 outlier mutations that aligned with these identified CNV clones, demonstrating that these mutations were capable of tracking orthogonally identified clones (Fig. 3.12). We also found a smaller set of 3 "inlier" mutations that aligned with the CNV clones, which is to be expected given the unsupervised nature of

the filtering and the lack of theoretical guarantees that all artifacts must follow a specific sequence context and/or prevalence. There could in theory be artifacts that are similar in their density to "real" somatic mutations. That being said, enrichment of clonally tracked mutations in outliers relative to inliers, coupled with the de-novo identification of mutation signatures, makes this a novel approach for enriching for informative somatic mutations in single cells.

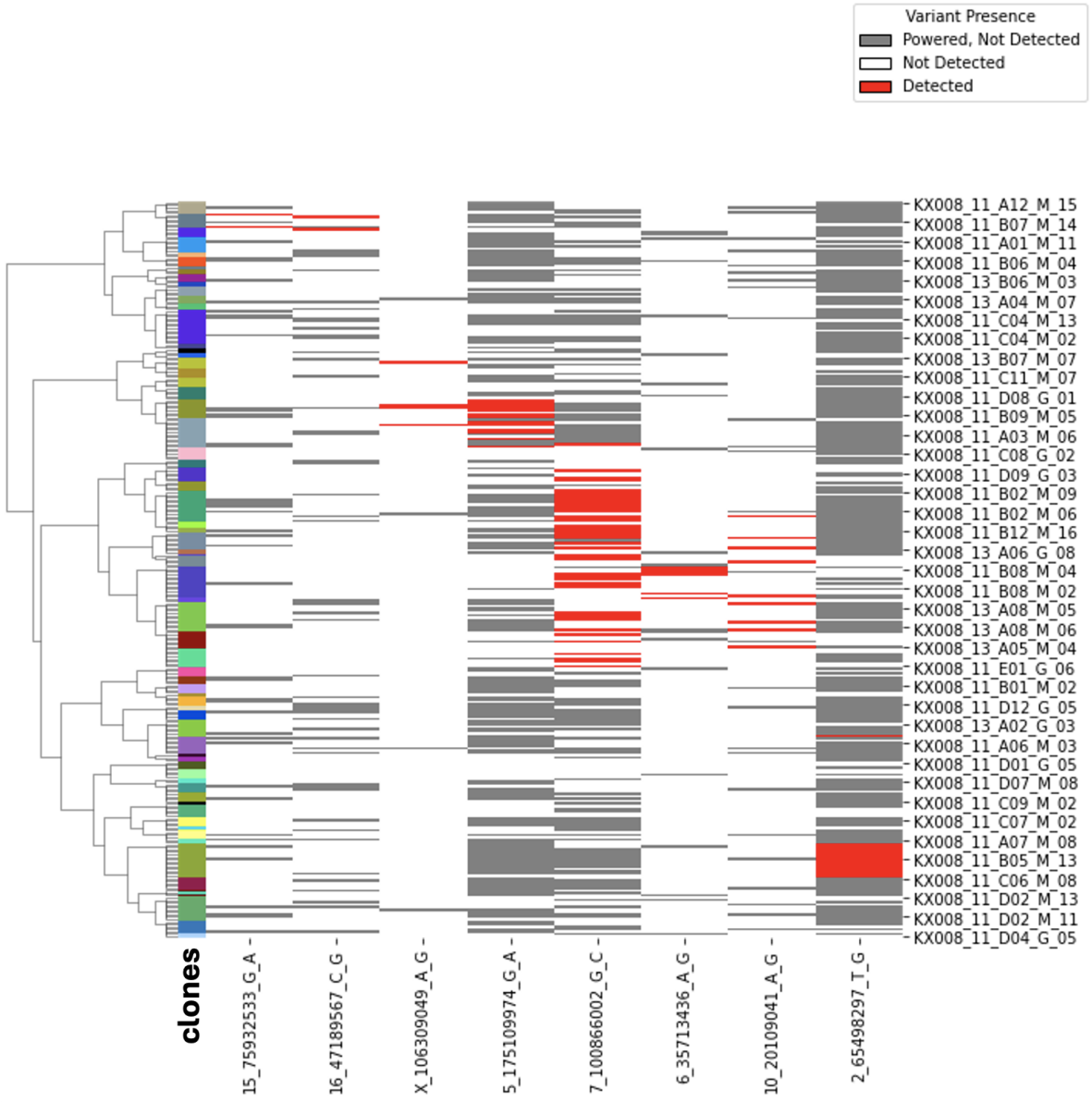

Figure 3.10: Mutations called from scRNA-seq data from blood colonies align with clades from "ground truth" DNA-based phylogeny. The dendrogram is the "ground truth" phylogeny derived from low coverage whole genome sequencing of colonies. Each colored row marks the colony from which the individual cell came from.

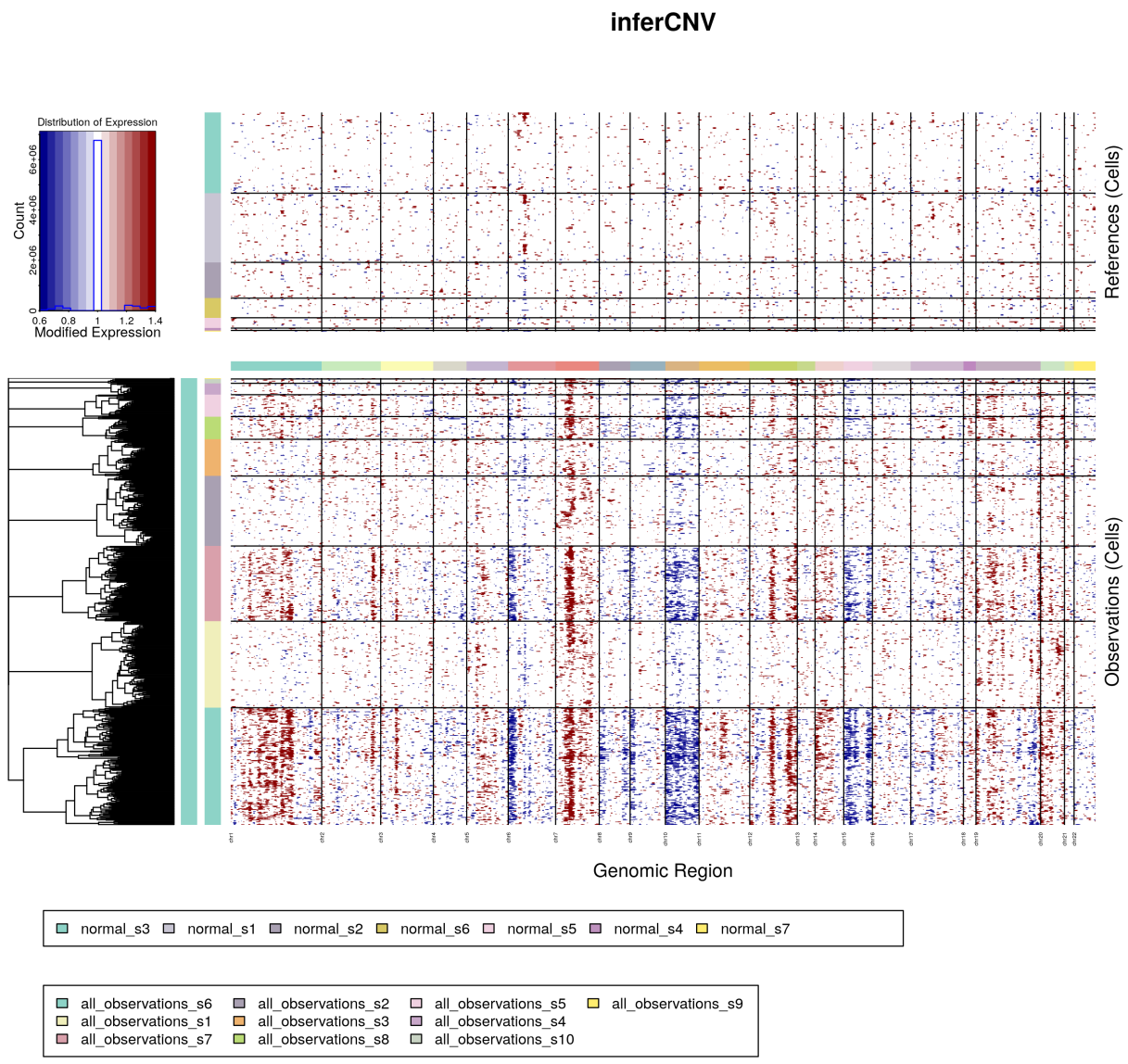

Figure 3.11: Inferred copy number profile from GBM Smart-seq2 data. Reveals canonical chromosome 7 gain and 10 loss. Hierarchical clustering on CNV profiles reveals "CNV-clones" on the basis of shared CNV states, with chromosome 1 gains uniquely defining certain clusters.

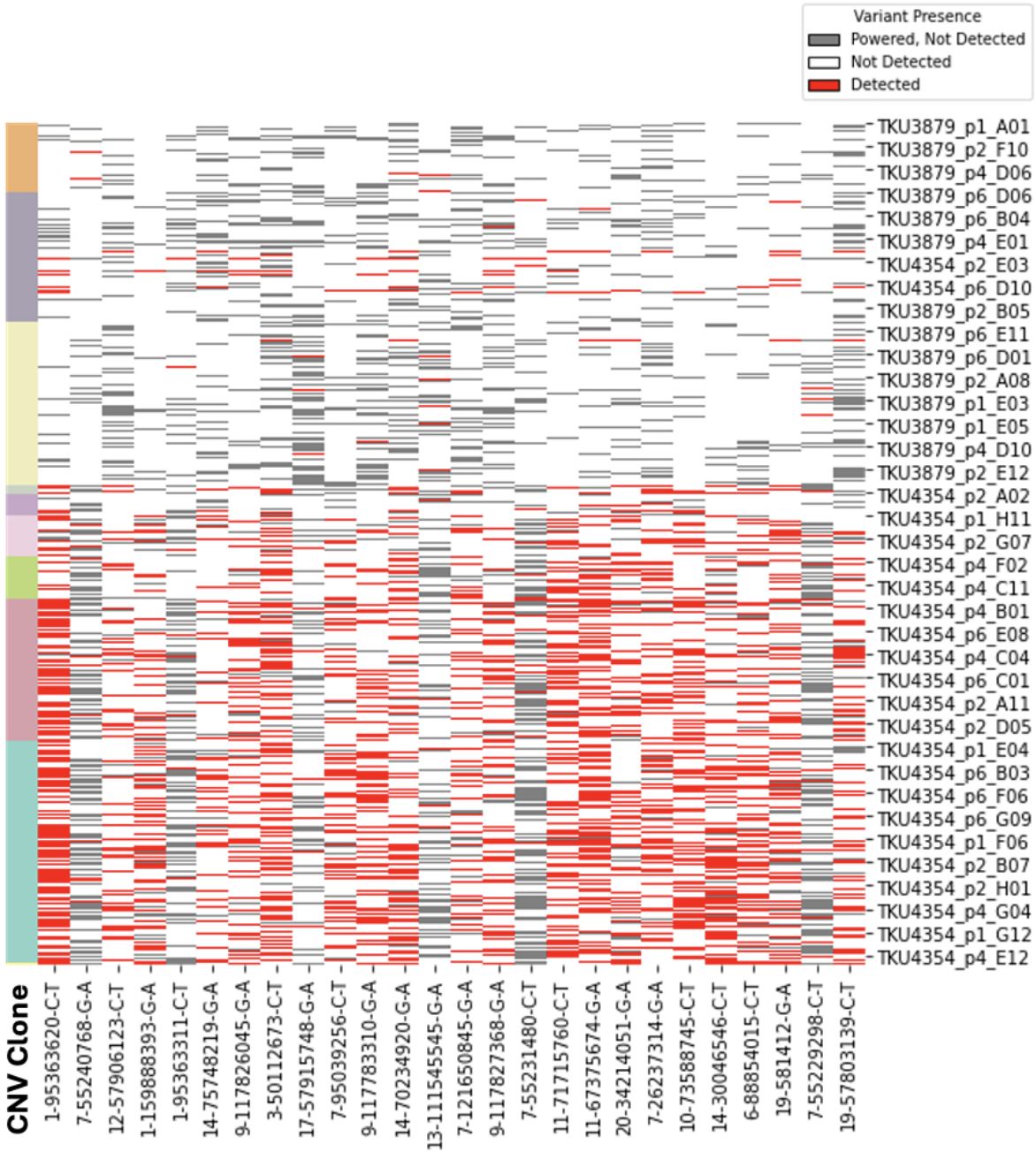

Figure 3.12: Oulier mutations are enriched for mutations that align with CNV-defined clones. 19 somatic mutations found to exclusively align with clones with shared CNV states. Clone color scheme is the same as the CNV-clone color scheme in Fig. 3.11.

3.3.4 De-Novo identification of cells with shared mutations

Single-cell data is notoriously wrought with missingness, resulting in a sparse single-cell mutation matrix that requires sophisticated analytical tools to identify any phylogenetic signal. While this was a secondary focus of this study, we highlight a few approaches for exploratory analysis of single-cell mutation data. While there are a variety of distance-based phylogenetic reconstruction methods (see Introduction) that one can explore to uncover phylogenetic signal, one must carefully select the distance metric appropriate for the context. In our approach we feature our data as ternary: 0 for sites that do not have a mutation ('missing'), but powered to detect mutations ('powered'), 0.5 for missing and not powered to detect and 1 for detected. Jaccard metric is a commonly used metric for assigning similarity within sparse data, where there is no penalty for comparisons with missing entries. However, its natural formulation handles binary data and only rewards shared detected values or 1s but not shared informative missing values or 0s. We modified the Jaccard metric in our case to reward both 0s and 1s, and normalized only by comparisons that contain zeroes or ones, as 0.5s are phylogenetically uniformative (see Methods). True missing mutations contribute to phylogenetic signal and should be considered in the distance formulations. Furthermore, with this formulation we find it easier to identify clusters of cells with shared mutations than with other distance metrics, including standard Jaccard (Fig. 3.13).

Given autoencoders' adept handling of missing data and common use for representing single-cell data, we explored this class of unsupervised machine learning approaches for also identifying clusters of cells with shared mutations [31, 32]. We kept the standard implementation of the Variational Autoencoder (VAE), but used a modified reconstruction loss function where we minimize the negative log likelihood of the beta distribution, given the data representation of 0, 0.5, and 1s (Fig. 3.14). We masked the contribution of entries with 0.5s in the loss update step. We applied this model to the blood colony data, given its clonal ground truth, and used a generalized mixture model to cluster cells in the latent space to

identity "mutation clusters," revealing an interesting bifurcating structure (Fig. 3.15). When aligning the cell x mutation matrix by cluster assignment, there appears to be mutations that uniquely define identified clusters (Fig. 3.16).

We propose these exploratory approaches not as formal phylogenetic reconstruction methods, but more as preprocessing steps for checking for the presence of phylogenetic signal in sparse single-cell mutation data. This is an active area of research and these proposed formulations can form as inputs into more sophisticated phylogenetic reconstruction algorithms. For instance, one can go beyond identifying mutation clusters in the latent space and explore contrastive learning methods for phylogenetically ordering the clusters identified in the latent space. Future approaches could also be grounded by a known ground truth phylogenetic tree, such as from a lineage tracing experiment, such that the latent space is constrained by this known pairwise distance between the trees.

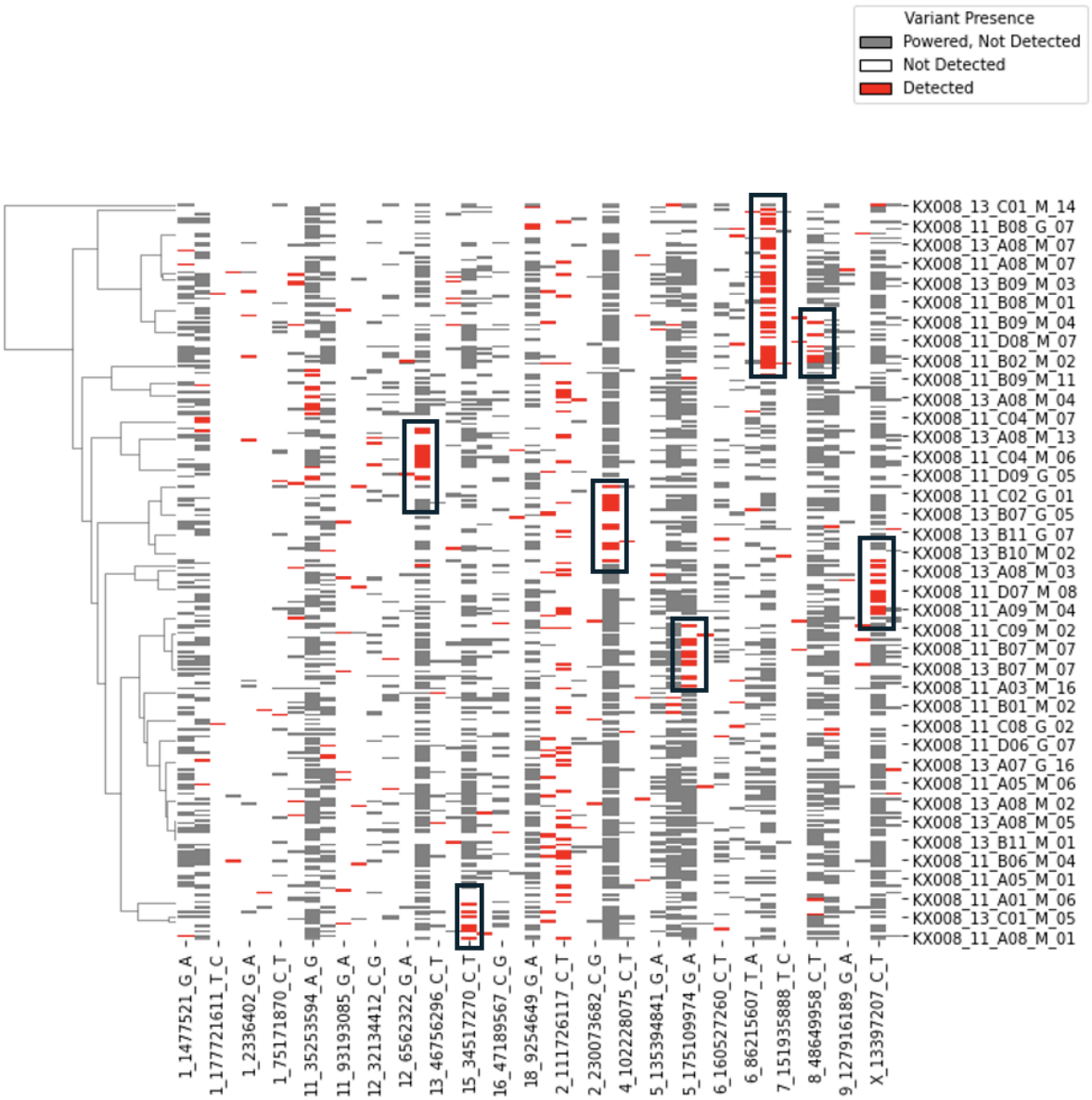

Figure 3.13: Hierarchical clustering using modified Jaccard-distance reveals clusters of cells with shared mutations, suggesting evidece of phylogenetic structure in blood colony sample, KX008.

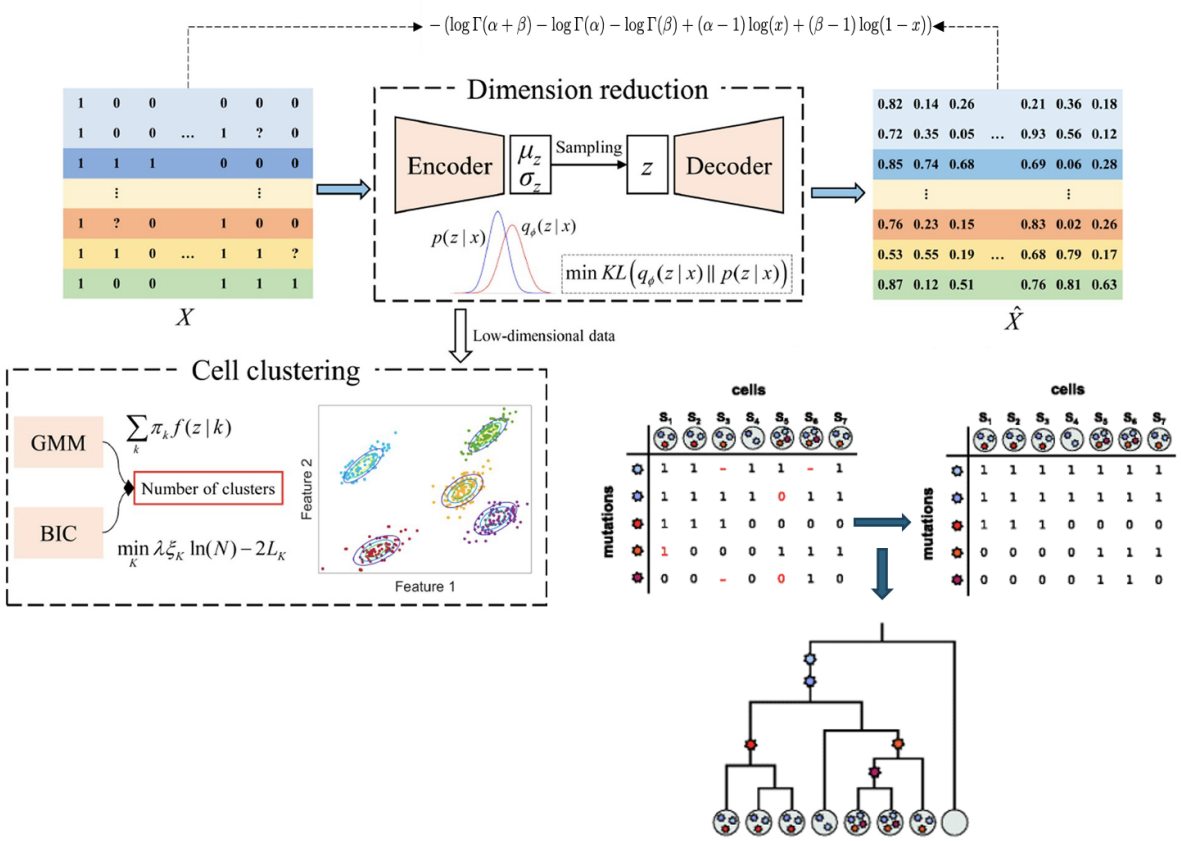

Figure 3.14: Beta-VAE architecture for modeling single-cell mutation data with custom beta log-likelihood loss function; adapted from[87]

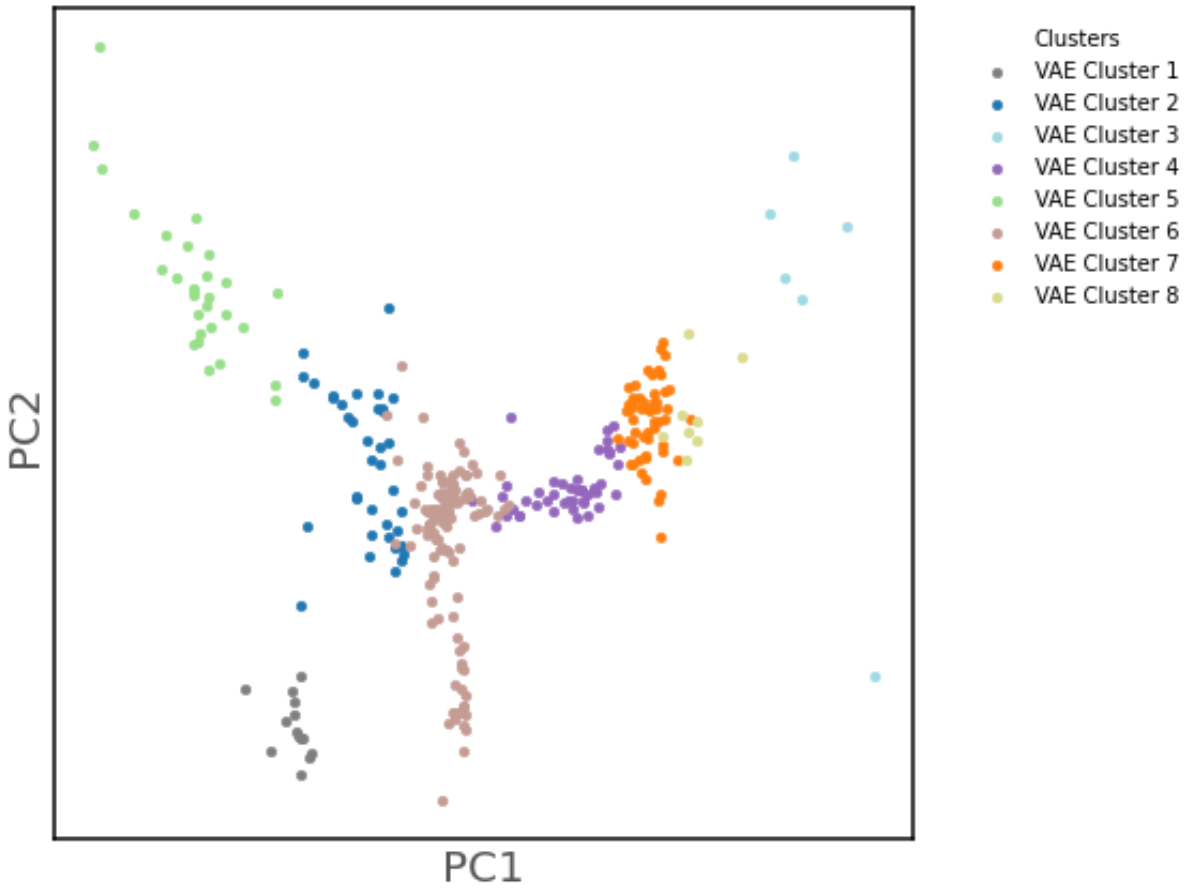

Figure 3.15: Performing PCA and Clustering of single-cells in low dimensional space of the beta-VAE reveals interesting bifurcating structure. VAE clusters appear to align with individual ground truth clones or larger clades containing multiple clones.

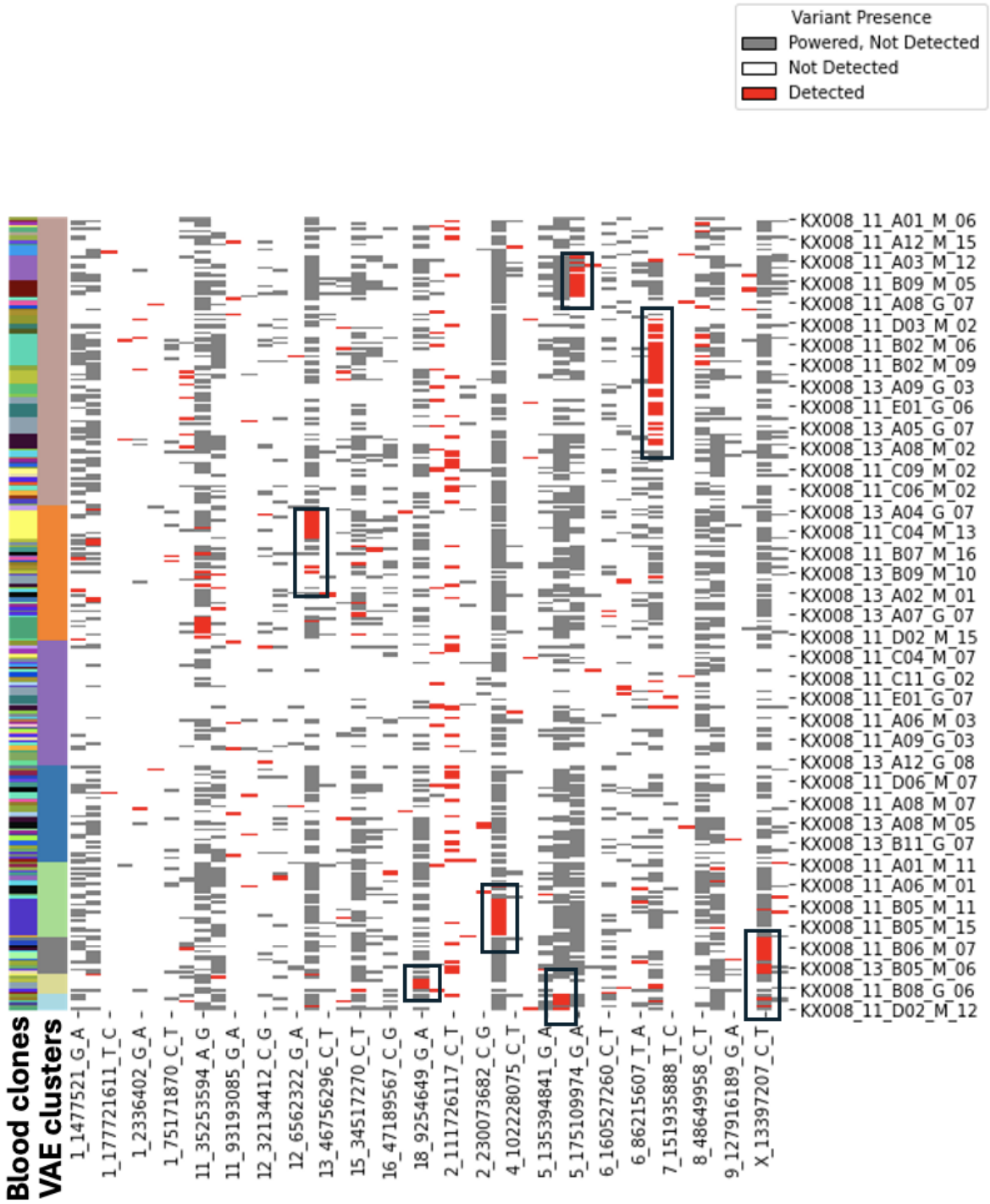

Figure 3.16: Performing PCA and Clustering of single-cells in low dimensional space of the beta-VAE reveals interesting bifurcating structure. VAE clusters appear to align with individual ground truth clones or larger clades containing multiple clones.

3.3.5 Associations with Mutation Spectra and Cell States

In an attempt to integrate cellular genotypic information with phenotypic states, we implemented supervised NMF to identify the relative contribution of mutation signatures in each cell. We then looked for an association between the 4 dominant cell states in GBM (OPC, NPC, AC, MES) and the relative contribution of TMZ (Fig 3.17) [14]. There appeared to be a slightly higher contribution of SBS11 in the AC in comparison to the MES populations of cells (Fig. 3.18). However, this analysis could benefit from greater sample sizes to confidently assess the contribution of signatures to expression states, as this analysis was performed on a single sample in a single patient. Furthermore, this analysis could be better suited for more diverse subclonal and episodic mutational processes such as APOBEC, because when correcting for expression in the cells, the difference in SBS11 contribution is less significant. Regressing out the effects of expression when comparing mutation clusters is a challenge, and an area for improvement in future work.

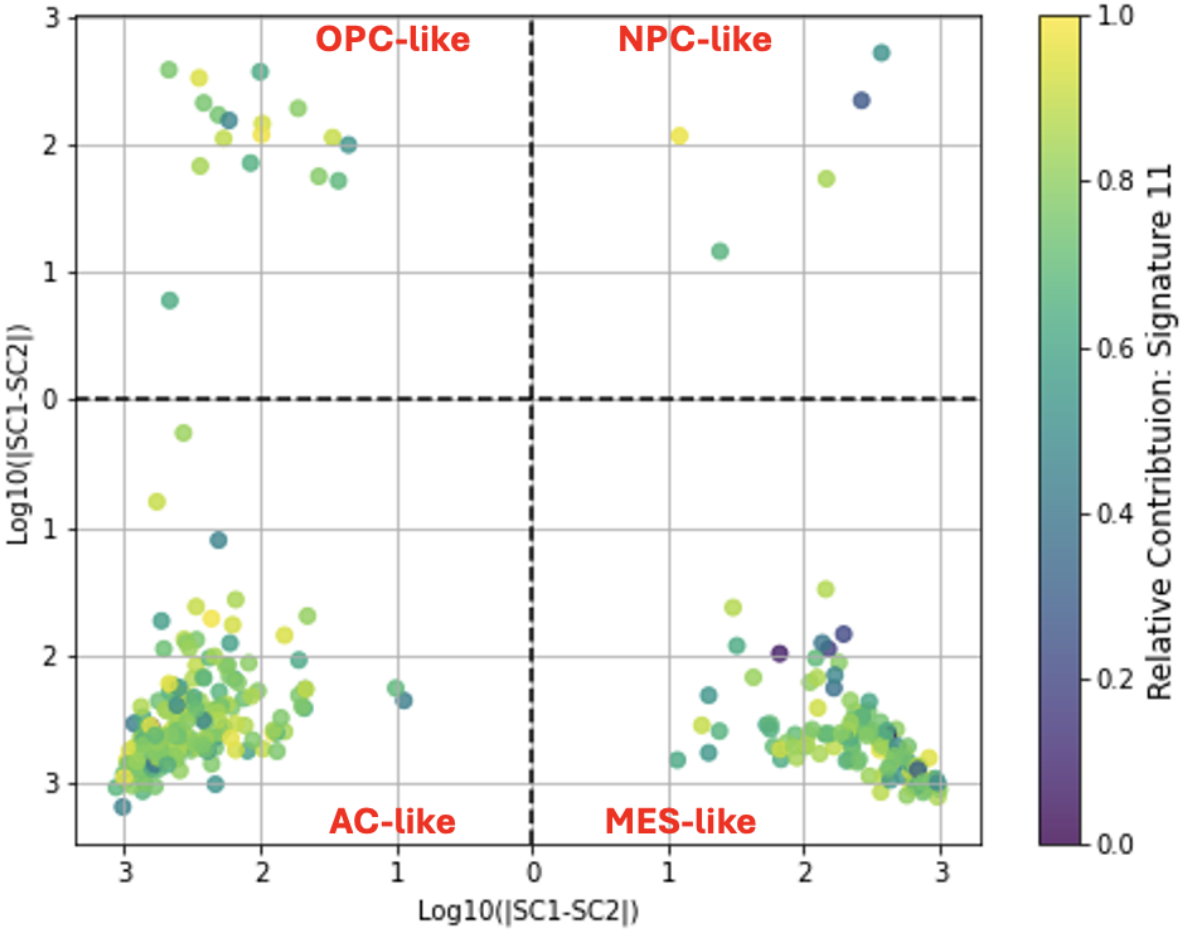

Figure 3.17: Mapped TMZ signature loading onto individual cells in GBM sample. Each quadrant defines one of the four dominant GBM cell states: AC, MES, OPC and NPC. Plotting scheme adapted from [14].

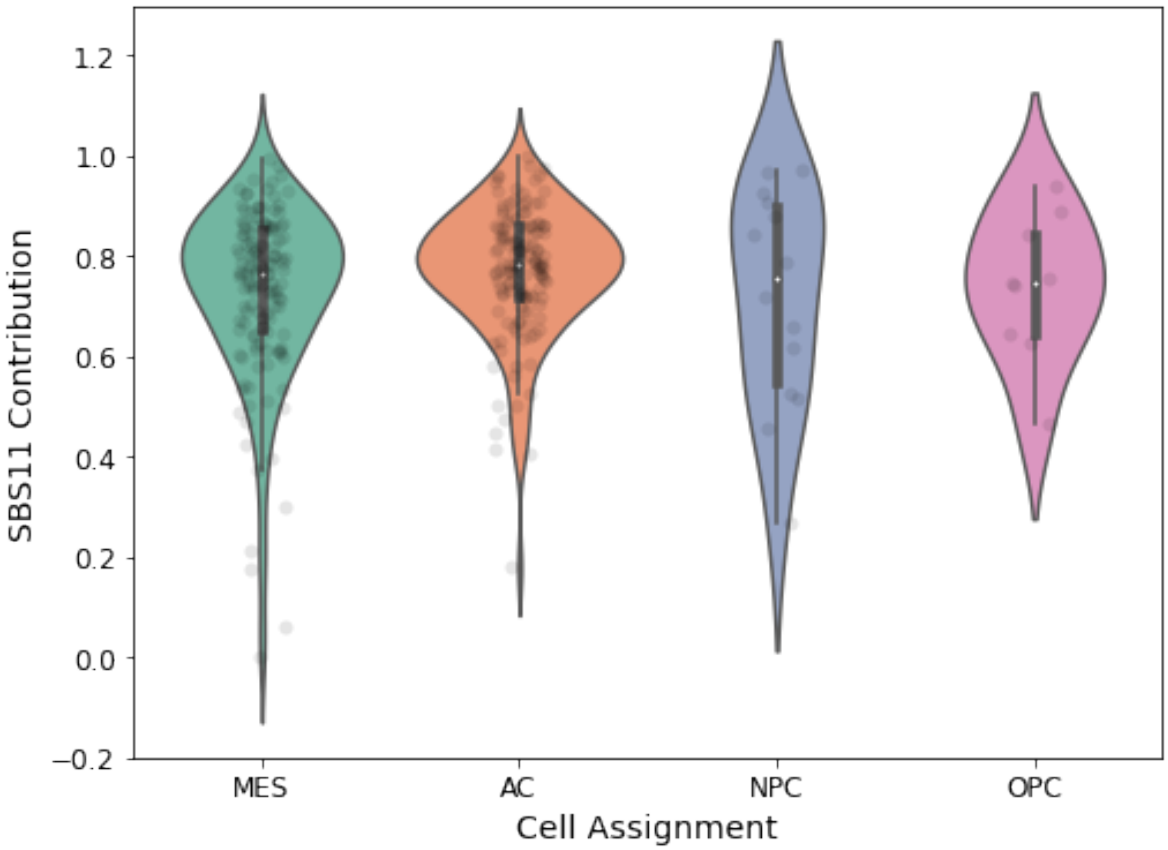

Figure 3.18: Comparing TMZ signature Loading Across GBM Cell States

3.4 Discussion

Calling somatic mutations from RNA-seq and scRNA-seq comes with a number of challenges, due to false positive technical artifacts and the missingness. In this work, we propose a number of filtering approaches to manage the large number of putative false positives we observe when initially calling mutations from full-length scRNA-seq. Anomaly detection approaches proved effective in separating mutation calls that appeared to generate artefactual mutation spectra from "outlier" mutations that more closely resembled real, expected mutation spectra of the respective profiled cancers. Furthermore, when judging performance using enrichment in matched WES, the performance rivals that of recently published single-cell mutation calling pipelines, while also enriching for known signatures. Lastly, we demonstrate that

high confidence variants called from scRNA-seq data can serve clonal markers, using either DNA-phylogenies or CNV states as "ground truths" of clonality.

We also contribute approaches for assessing phylogenetic signal in single-cell mutation data. While there exists a number of sophisticated likelihood-based methods for phylogenetic reconstruction for single-cell data, in our hands, we found the phylogenies built from these methods challenging to interpret and struggled to find obvious mutations that supported branches in the tree [29, 30]. We reasoned that having approaches that can quickly assess the phylogenetic signal in single-cell mutation data would prove useful for researchers prior to advancing to more classical phylogenetic approaches. One can both identify evidence of phylogenetic signal and also single out mutations that define mutation clusters for better feature selection, which can then serve as high confidence inputs for downstream sophisticated phylogenetic reconstruction methods. That being said, there are ways to infer likelihoods of mutations given a phylogeny (https://github.com/NickWilliamsSanger/treemut) [88]. One could build upon this and develop an iterative phylogenetic inference approach that both infers trees and then iteratively prunes non-informative (low likelihood) mutations and repeats inference with more highly confident (high likelihood) mutations.

Future studies integrating the genotypic information from called mutations with the paired expression information from the transcriptome, holds promise in uncovering heterogeneity in different developmental processes in both normal and disease tissues. These kinds of studies can help identify mutations that influence downstream expression programs that drive cancers. A recurrent issue worth addressing, in future work, is the effect of detection ability in given cells (ie: number genes in a given cell) in the downstream assessment of differential expression between cells with different sets of mutations. One will need to develop ways to regress out the effect of variability in overall expression across cells to confidently assess the expression differences between groups of cells with different mutation profiles.

Overall, this work contributes to and complements the growing body of work involving mutation calling in single cells and downstream phylogenetic reconstruction by providing an

approach for uncovering high confidence mutations supported by known mutational signatures and methodologies for identifying evidence of phylogenetic structure in single-cell mutation data.

3.5 Methods

3.5.1 Mutect2

Mutect2 is a somatic variant caller that identifies tumor-specific SNVs and indels by locally assembling sequencing reads, comparing tumor and normal samples, and applying a series of statistical filters that evaluate features such as mapping quality, strand bias, base quality, read orientation artifacts, and contamination; these filters are used to distinguish true somatic mutations from sequencing errors, germline variants, and technical artifacts. Unlike MuTect, Mutect2 is capable of calling indels mutations.

Mutect2 was ran on STAR aligned scRNA-seq merged samples with matched normal WES samples if available and in "tumor only" mode if we didn't have access to matched normals. The filtered VCF outputs from Mutect2 were converted to MAF files where each row in the MAF file constitutes a cell's record of given mutation. Remaining downstream filtering steps were performed on this file.

3.5.2 Blacklist Filters

To account for additional sources of technical noise, I implemented a number of blacklisting filters using genomic databases. To account for common germline SNPs, especially when a matched normal is missing, I removed variants that overlapped with sites in gnomAD [83]. I also utilized the DARNED, RADAR & Rediportal RNA-editing databases, to blacklist known RNA-editing sites [89–91]. To retain high-confidence mutations, I applied a consensus filter such that I only considered mutations that were found in at least three malignant cells.

3.5.3 scPoN

I employ an in-house panel of normal (PoN) samples procedure, as previously described [82]. The procedure is a pile-up based approach that bins the alternative counts in different allele frequency ranges within a large set of normal CD45+ immune cells. For each candidate variant, it computes a dot product between the discrete bins and the respective allele frequency slices of the tumor variant's beta distribution, parameterized by its reference and alternate allele frequencies. PoN procedure:

$$\begin{aligned} &\mathbf{h:} \\ &1: \quad 0 < VAF < 0.1\% \\ &2: \quad 0.1 < VAF < 0.3\% \\ &3: \quad 0.3 < VAF < 1\% \\ &4: \quad 1 < VAF < 3\% \\ &5: \quad 3 < VAF < 20\% \end{aligned} \tag{3.1}$$

$$\mathbf{f} \sim \text{Beta}(n_{\text{alt}} + 1, n_{\text{ref}} + 1) \tag{3.2}$$

$$\textbf{PoN Score:} \quad S = \vec{f} \cdot \vec{h} \tag{3.3}$$

This calculation yields a PoN score for each variant. Thresholds for filtering are determined based on empirical testing of different score ranges, typically chosen when common SNPs begin to be filtered or when enrichment of expected mutation signatures are achieved.

3.5.4 Local Outlier Factor (LOF)

LOF is a density-based anomaly detection method that identifies outliers by comparing the local density of a point to that of its neighbors. In low-dimensional feature space, LOF

assigns an anomaly score to each point based on how isolated it is relative to its surrounding neighborhood. Points that have significantly lower density than their neighbors are labeled as outliers. The LOF score for a point x is computed as:

$$\text{LOF}(x) = \frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \frac{\text{lrd}(y)}{\text{lrd}(x)} \tag{3.4}$$

where:

- $N_k(x)$ is the set of k nearest neighbors of x,
- $\text{lrd}(x)$ is the **local reachability density** of point x, defined as:

$$\text{lrd}(x) = \left(\frac{\sum_{y \in N_k(x)} \text{reach-dist}_k(x, y)}{|N_k(x)|} \right)^{-1} \tag{3.5}$$

and the **reachability distance** is given by:

$$\text{reach-dist}_k(x, y) = \max \{\text{k-distance}(y), \text{distance}(x, y)\} \tag{3.6}$$

Here, d(x,y) is the distance between points x and y, and k-distance(y) is the distance from y to its k-th nearest neighbor. In our application, k=7 neighbors were chosen as the neighborhood size, and Hamming distance was used as the distance metric between sequence-encoded feature vectors.

3.5.5 One-Class Support Vector Machine (One-Class SVM)

One-Class SVM is a non-linear classification method that constructs a boundary to encompass the majority of the training data, treating it as a single class of "inliers." Specifically, the one-class SVM solves the following optimization problem:

$$\min_{w,\rho,\xi} \quad \frac{1}{2}\|w\|^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i - \rho \tag{3.7}$$

subject to:

$$(w \cdot \phi(x_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0 \tag{3.8}$$

where:

- n is the number of points,
- w defines the decision boundary,
- ρ is the offset,
- ξ_i are slack variables allowing some margin violations,
- ν controls the fraction of outliers allowed,
- $\phi(\cdot)$ is a non-linear mapping to a higher-dimensional space via a kernel function.

A radial basis function (RBF) kernel was used to capture non-linear relationships in the mutation sequence context space.

3.5.6 Feature Representation

Both LOF and One-Class SVM were trained using the **heptamer sequence context** surrounding each single nucleotide variant (SNV). The sequence context was represented as a flattened one-hot encoded vector:

- Each nucleotide (A, T, C, G) was represented by a 4-dimensional one-hot encoded vector (e.g., A = [1, 0, 0, 0], T = [0, 1, 0, 0], etc.).
- Each base substitution (mutation) was represented by a 6-dimensional vector corresponding to the six possible single-base changes (e.g., C>A = [1, 0, 0, 0, 0, 0], C>G = [0, 1, 0, 0, 0, 0], etc.).

Thus, each SNV was encoded by concatenating the one-hot encoding of its heptamer sequence context with its mutation type encoding, producing a feature vector of dimension 30 suitable for anomaly detection.

Different training schemes were tested for filtering anomalous sequencing contexts in different settings:

-Models were trained exclusively on mutations detected in normal samples, under the assumption that most of these are technical artifacts.

-Models were trained on a mixture of the normal samples' mutation and the malignant candidate mutations, to spike in "inlier" mutations that would cluster with the noisy mutations in the malignant samples.

-Models were trained solely on the malignant samples with the assumption that the majority of the mutations in cancer samples are noise with distinguishable sequence contexts with respect to real mutations.

3.5.7 Inferring Copy Number and Identifying CNV Clones

`inferCNV` is a computational tool for identifying large-scale copy number variation from scRNA-seq data. `inferCNV` identifies CNV events from scRNA-seq by calculating a moving average of relative expression of cells against a reference baseline of cells, using a sliding window of 101 genomically adjacent genes:

$$\mathrm{CNV}_k(i) = \frac{\sum_{j=i-50}^{i+50} E_k(o_j)}{101} \tag{3.9}$$

where CNV_k is the copy number in cell k in the i^{th} ordered gene, o_j is the j^{th} gene in the genomically ordered list, and $E_k(o_j)$ is the relative normalized expression of that gene in cell k.

The mean value for each gene across normal (reference) cells is *subtracted* from the smoothed expression values in the putative tumor cells, representing fold change of expression

of genomic regions with respect to normals.

The relative average expression values are hierarchically clustered to identify groups of cells with shared CNA events. If adequate admixed normal cells can be detected in the sample, these matched cells are utilized as our reference. Otherwise, normal $CD45^+$ immune cells from previous GBM studies are utilized as reference normal cells.

`inferCNV` was run with the suggested expression cutoff for Smartseq-2 (cutoff = 1), HMM set to `true` to enable CNV prediction, and analysis mode = `'subclusters'` to enable clustering of CNA events.

3.5.8 Modified Jaccard Distance Metric

For exploratory analysis, we represented the mutation matrix as a **ternary matrix**, with entries encoded as follows: 0 for "powered but mutation not detected," 0.5 for "unpowered and missing," and 1 for "mutation detected." Traditional distance metrics, such as the **Jaccard index**, are commonly used for sparse binary data, as they naturally disregard missing entries when computing similarity. However, the standard Jaccard index rewards only shared detected values (i.e., shared 1s) and does not account for informative absence (shared 0s), which in our context contains valuable phylogenetic information.

To better capture similarity in ternary single-cell mutation matrices, we implemented a **customized Jaccard distance**. In our formulation, both shared 1s (mutations detected in both cells) and shared 0s (mutations absent in both cells but confidently called) contribute positively to similarity, while entries with a value of 0.5 (unpowered/missing) are masked and excluded from both the numerator and denominator. This approach ensures that only phylogenetically informative comparisons influence the distance computation.

The **custom Jaccard similarity** between two cells A and B is given by:

$$\text{Custom Jaccard Similarity} = \frac{J_{11} + J_{00}}{J_{11} + J_{00} + J_{10} + J_{01}} \tag{3.10}$$

where:

- J_{11} = number of sites where both A and B have 1 (mutation detected),
- J_{00} = number of sites where both A and B have 0 (mutation absent but powered to detect),
- J_{10} = number of sites where A has 1 and B has 0,
- J_{01} = number of sites where A has 0 and B has 1.

The corresponding **custom Jaccard distance** is defined as:

$$\text{Custom Jaccard Distance} = 1 - \left(\frac{J_{11} + J_{00}}{J_{11} + J_{00} + J_{10} + J_{01}} \right) \tag{3.11}$$

Entries corresponding to unpowered (0.5) sites are ignored in all J_{ij} counts. This adjustment ensures that missing, non-informative entries do not bias similarity calculations while rewarding both shared presences and shared informative absences.

3.5.9 Beta VAE

To model the noisy and partially observed nature of single-cell mutation data, we developed a Beta-Variational Autoencoder (Beta-VAE) framework. In this setup, the input matrix consisted of mutation calls across cells, with entries encoded as 1, 0.5, or 0. A value of 1 indicated a detected mutation, 0 indicated a site with sufficient power but no mutation detected, and 0.5 represented sites that were unpowered for detection (i.e., missing or ambiguous data).

The Beta-VAE decoder parameterized the reconstruction of each input entry through a Beta distribution, outputting site-specific $\alpha(z)$ and $\beta(z)$ parameters for each cell. The model was trained by minimizing a two-term loss function consisting of a reconstruction loss and a regularization term. Specifically, the objective function was:

$$\mathcal{L}(\theta, \phi; x) = -\mathrm{E}_{q_\phi(z|x)} \left[\log \mathrm{Beta}(x; \alpha(z), \beta(z)) \right] + D_{\mathrm{KL}} \left(q_\phi(z|x) \,\|\, p(z) \right) \tag{3.12}$$

where:

- **The reconstruction loss** is the negative log-likelihood of the observed mutation states under the Beta distribution:

$$\begin{aligned} -\log \mathrm{Beta}(x; \alpha, \beta) = \\ -\big(\log_e \Gamma(\alpha + \beta) - \log_e \Gamma(\alpha) - \log_e \Gamma(\beta) \\ +(\alpha - 1) \log_e(x) + (\beta - 1) \log_e(1 - x) \big) \end{aligned}$$

- **The KL divergence** term regularizes the approximate posterior $q_\phi(z|x)$ toward a standard normal prior $p(z)$, encouraging smoothness and disentanglement in the latent space.

To handle unpowered or ambiguous sites (entries with value 0.5), we implemented a masking strategy: these entries were excluded from the reconstruction loss calculation during training, ensuring that model updates were based solely on confident observations (0s and 1s). This masking helped the model remain robust to technical artifacts such as dropout and uneven coverage.

After training, each cell was represented by a latent vector summarizing its mutation profile. To identify groups of cells with shared mutational features, we applied a Gaussian Mixture Model (GMM) to the latent space. The number of Guassians are selected by fitting GMMs with varying numbers of components (fromm 1 to 10) and selected the component with the lowest BIC as previously described [87]. The GMM probabilistically assigned cells to clusters, corresponding to putative clonal populations or sublineages. This combination of Beta-VAE modeling and latent space clustering enabled the unsupervised identification of clonal structures while appropriately handling detection uncertainty inherent to single-cell mutation data.

3.5.10 Code Availability

The reproducibility code for this work, while not public yet, will be made public upon publication and can be found here: *https://github.com/jidezike3/scRNA_Mutect2_Denoising_Pipeline*

Chapter 4

Single-cell phylogenies for the study of cancer drug persistence potential

This chapter was adapted from a manuscript in preparation for submission, in collaboration with Binyamin Zhitomirsky, a postdoctoral associate in the Getz Lab. We will be co-first authors in this study.

4.1 Abstract

Drug-tolerant Persister cells (DTPs) are a subpopulation of cancer cells that survive initial anticancer drug treatment despite lacking genetic resistance-driving mutations. DTPs have been shown to serve as an intermediate stage in the evolution of acquired drug resistance, facilitating the adaptation and survival of cancer cells under sustained therapeutic pressure.

Recent work has shed light on the biology of persister cells during and following drug treatment. However, little is known about the biology of persistence potential - the intrinsic cellular and molecular characteristics that prime certain drug-naïve cells to persist upon initial exposure to anticancer drugs. The main challenge in studying the persistence potential of drug-naïve cells is rooted in the inability to determine, prior to treatment, which cells will ultimately survive (i.e., persist), coupled with the fact that post-treatment survivors have

undergone drug-induced alterations that obscure their original pre-treatment state.

To address this issue, we adapted a CRISPR-based high resolution lineage tracing technology for the study of persistence potential and persistence potential altering events [18]. We also contributed a set of phylogenetic analytical methodologies for scoring "potential" in untreated cells, identifying events in a phylogeny and, lastly, for generating hypotheses of phenotypic drivers that associate with these phylogenetic events. Utilizing this technology to study persistence to EGFR inhibitors in EGFR-positive non-small cell lung cancer cells, we demonstrated evidence for the heritability of persistence potential. We also identified events on the lineage tree that significantly alter the persistence potential of specific clades of cells. By pairing phylogenetic information with single-cell RNA-sequencing (scRNA-seq) expression profiles, we identified multiple pathways and genes that were associated with persistence potential. We found oxidative phosphorylation (OXPHOS) and related mitochondrial genes to be consistently positively associated with persistence potential, while ribosomal and translation related genes were found to be consistently negatively associated with persistence potential. We also identified the proteasomal gene, *PSMA7*, as being significantly positively correlated with persistence potential in multiple clades. Mapping of cycling persisters onto single-cell phylogenetic trees demonstrated that certain clades are more likely to generate cycling persisters while others will generate persisters that will remain in cell cycle arrest. Pairing lineage information with full length scRNA-seq using MAS-iso-seq allows to study the role of differential isoform expression in driving persistence potential.

Our findings of persistence potential associated pathways and genes were validated in functional in-vitro pharmacological experiments, resulting in synergistic killing of persister cells. Furthermore, we demonstrate that, in a cohort of EGFR-positive lung cancer patients, expression of persistence potential drivers (e.g., *PSMA7*) correlates with shortened progression-free survival and overall survival.

4.2 Introduction

Cancer drug-tolerant persister (DTP) cells represent a distinct subpopulation of cancer cells that survive anticancer drug exposure by entering a transient persister cell state. Unlike drug-resistant cancer cells, which harbor genetic mutations that confer the ability to thrive under continuous drug pressure, DTPs evade cytotoxicity through non-genetic mechanisms [92–94]. The transient DTP state is characterized by quiescence or slowed proliferation, metabolic rewiring, and epigenetic adaptations [95–98]. This state enables DTPs to tolerate high concentrations of anti-cancer drugs that eliminate the bulk of the tumor population. Upon withdrawal of therapy, DTPs can revert to a drug-sensitive state or evolve into genetically resistant clones. Studies have demonstrated that DTPs can acquire diverse resistance mechanisms, including mutations in key oncogenes like *EGFR*, *KRAS*, and *BRAF*, following prolonged drug exposure [93]. The ability of cancer cells to enter a persister state upon drug exposure and the ability of DTPs and DTP-derived cells to acquire resistance mediating mutations, underscores the critical role of DTPs in the dynamic interplay between cancer cell plasticity and therapeutic pressure, positioning them as a primary target for novel intervention strategies to improve the efficacy of anti-cancer treatment protocols.

The biology of persister cells is shaped by a wide array of cellular and molecular mechanisms that enable their survival during drug treatment. These mechanisms include metabolic rewiring, epigenetic reprogramming, transcriptional adaptation, and activation of stress response pathways. Persister cells often shift from glycolysis to oxidative phosphorylation (OXPHOS), allowing them to meet energy demands while maintaining in redox balance in the face of therapeutic stress [98]. Epigenetic modifications, such as histone methylation and acetylation, play a critical role in establishing a drug-tolerant state by altering chromatin structure and gene expression [95]. For example, upregulation of histone demethylases like *KDM5B* has been implicated in persister cell survival across multiple cancer types [99]. Transcriptionally, persister cells rewire gene regulatory networks to promote survival

under stress, with key transcription factors such as YAP and TEAD orchestrating adaptive responses. In addition, persister cells activate integrated stress responses (ISRs), including the unfolded protein response (UPR), to mitigate damage caused by endoplasmic reticulum stress and reactive oxygen species [100]. These stress response pathways enable persister cells to suppress apoptotic signals and sustain viability. Collectively, these diverse adaptations confer a remarkable plasticity to persister cells, positioning them as a formidable obstacle to effective cancer treatment and a focal point for therapeutic innovation.

Recent findings suggest that not all cancer cells possess an equal likelihood of transitioning into a DTP state, a property referred to as persistence potential [101]. This non-stochastic probability, which varies between individual cells within a tumor, indicates that some drug-naïve cells are inherently more likely than others to survive therapeutic stress and become DTPs. The mechanisms underlying persistence potential remain poorly understood, yet this trait is likely influenced by pre-existing cellular properties, including signaling dynamics, epigenetic states, and metabolic configurations. As it has been established that cells undergo significant alterations to their cell state and gene expression profile upon treatment and transition into a DTP state, we postulate that studying cells that are already in the DTP state will provide a partial understanding at best of the traits governing the persistence potential of drug-naïve cells. Understanding the determinants of persistence potential could inform strategies to prevent the transition of drug-naïve cells into a persister state, thereby reducing the emergence of drug tolerance and resistance. By addressing this knowledge gap, researchers could develop interventions to preemptively target high-persistence potential cells, offering a new avenue for improving therapeutic efficacy and patient outcomes. Despite its importance, persistence potential has been insufficiently studied due to the inherent challenges of studying cell fates without inducing transition into said fate by drug treatment.

In recent years, novel CRISPR-based lineage tracing technologies have facilitated new insights into questions regarding cell fate. Findings related to developmental trajectories as well as cancer progression, metastasis potential and tumor plasticity and evolution, were

reported using evolving barcodes that are engineered to accumulate mutations over time, allowing for the accurate reconstruction of phylogenies at single-cell scale [18, 25]. We herein report the adaptation of such a lineage tracing method, with both experimental and analytical advancements, to study the molecular mechanisms underlying persistence potential.

4.3 Results

4.3.1 Reconstructing Single-Cell Phylogenies of Untreated and Persister Cells

PC-9, EGFR-positive non-small-cell lung cancer cells, are sensitive to treatment with EGFR inhibitors (EGFRi). However a subpopulation of PC-9 cells has been shown to persist under EGFRi concentrations as high as ten times the IC50, by entering a reversible DTP cell state. For this reason PC-9 cells treated with EGFRi have been often used as a model cell line for the study of cancer drug persistence. We and others have previously reported that single-cell-derived clones of PC-9 cells have stable yet variable persistence potential, suggesting that persistence potential is a quantitative and inherited trait that is propagated from the original single cell to its progeny (Fig. 4.1) [101, 102]. To study alterations in persistence potential during clonal expansion and to gain insight into the molecular mechanisms underlying persistence potential, we have introduced a high-resolution CRISPR-based lineage tracing technology into PC-9 cells, as previously described [18]. Briefly, PC-9 cells were stably transduced with constitutively expressed Cas9, and multiple copies of expressed evolving barcodes. Finally, to start the lineage recording, cells were transduced with three guide RNAs, each targeting one of three target sites on each evolving barcode (Fig 4.2). Following initiation of lineage recording, single cells were plated, and allowed to expand into clones. Clones were split into two arms: the drug naïve arm was untreated, while the persister arm was treated with 300 nM osimertinib for 7 days (Fig. 4.3). At the end of treatment, 2,000

untreated cells and 1,000 persister cells were collected from each of three clones and processed using a 10X Genomics 3' kit, to generate scRNA-seq libraries. Amplicon libraries of the recording barcodes were generated using custom primers, as previously described (Fig 4.3).

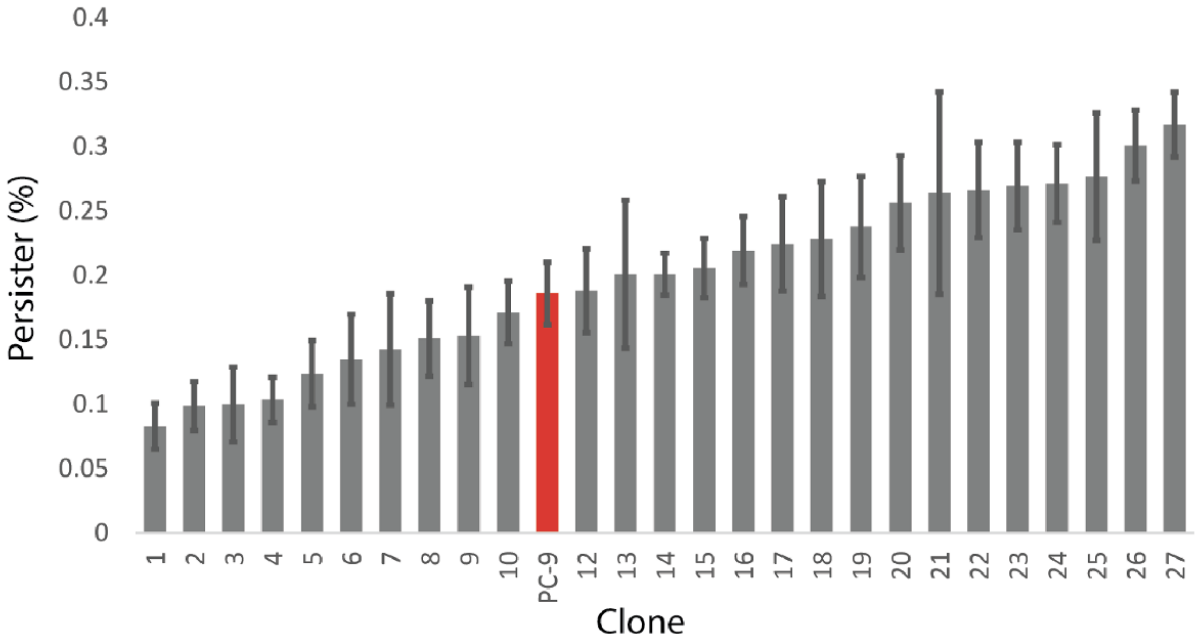

Figure 4.1: Different single-cell derived PC-9 clones demonstrate varying persistence capacity. The persistence capacity of each clone is retained across repeated experiments, suggesting that there is a memory of persistence in populations of cancer cells.

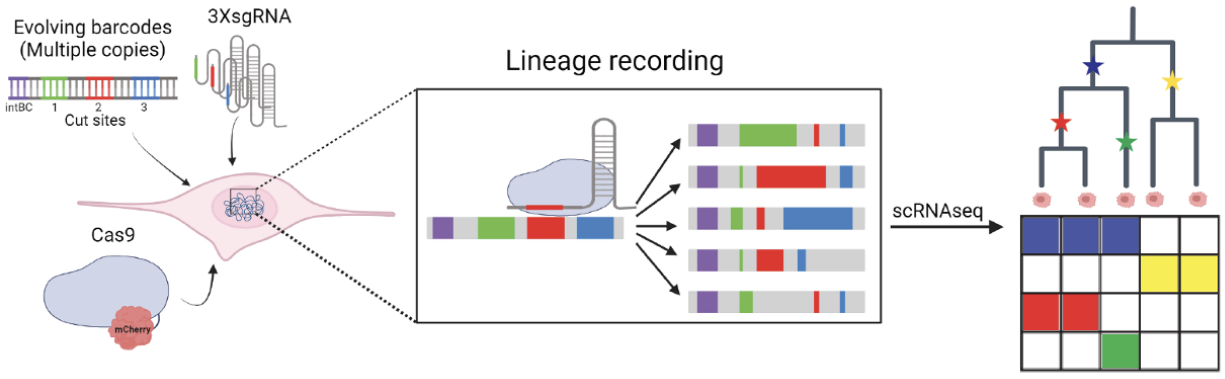

Figure 4.2: Description of Cassiopeia CRISPR Lineage Tracing System. **Left**: PC-9 cells are engineered to constitutively express CRISPR-Cas9. Multiple target-sites are integrated into the host genome as "scratchpads" for CRISPR-induced mutations and complementary sgRNAs that target the the target-site's cut-sites are introduced to the cells. **Middle**: Recruitment of the CRISPR machinery to the target site cut-sites leads to cutting of the DNA and acquired indels upon NHEJ repair. **Right**: Shared, inherited indels are used to build high resoultion single-cell phylogenies.

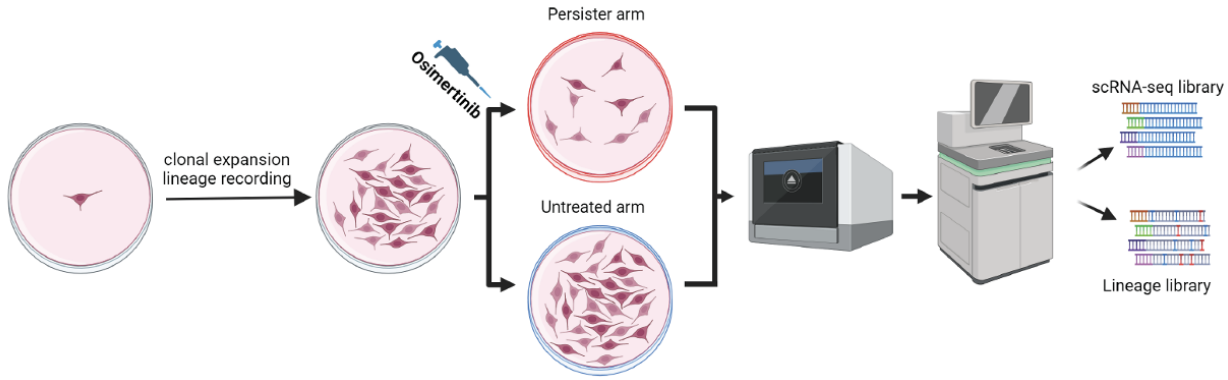

Figure 4.3: **Left**: Single-cell derived clones with the aforementioned engineering (target-sites), with a unique set of intBC marking that clone, expand for 3 weeks while lineage recording occurs. **Middle**: the cells are split to a persister arm and an untreated, where the persister arm receives 300nM osimertinib and expands/records for 1 week. **Right**: the scRNA-seq and lineage tracing libraries are generated from the cells from each arm and sequenced.

We utilize the Cassiopeia pipeline, as previously described, to map both untreated PC-9 cells and persister cells onto the same lineage tree for three different clones (clones Clone I, Clone II & Clone III) (Fig. 4.4) with varying levels of persistence. Briefly, Cassiopeia is a maximum parsimony phylogenetic reconstruction pipeline that utilizes CRISPR indel

character states to build single-cell phylogenies. Cells are engineered to constitutively express CRISPR-Cas9 and guide RNAs that target the cut sites within each target-site construct. Each unique target-site is randomly integrated in different genomic loci in the clone's founder cell and is marked by a respective integration barcode (intBC). In achieving maximum parsimony, the method minimizes the number of character state (e.g., CRISPR-induced indels) changes as it traverses from the root nodes to the leaf nodes that are the cells. We employed the "Cassiopeia-Hybrid" variant of the method, which combines both a greedy iterative split of groups of cells based on shared indel states with a more exhaustive approach that considers all possible ancestral node states when resolving the most parsimonious tree with the minimal number of state changes across nodes. In order to place cells from different timepoints (ie: untreated & persister cells) onto the same tree, we made slight modifications to the pipeline. We ran the Cassiopeia preprocessing steps in parallel for both the untreated and persister cell libraries, generating an allele table for each of the groups of cells. We then only considered the intBCs that were shared by the untreated and persister cells for each clone. We classified the clones based on the bulk intBC library we generated from the clone, which was possible due to the single-cell plating experimental approach. Lastly, to avoid introducing artificial clustering to the tree due to potential CRISPR recording while the persisters are undergoing treatment, we did not consider any character states that were unique to the persister cells. We performed multiple "sanity checks" to validate the reliability of our phylogenetic trees. First, we assessed the correlation between phylogenetic distance and allelic distance, observing Pearson correlation coefficients of 0.897, 0.42, and 0.71 for Clones I, Clone II, & Clone III, respectively—comparable or better than the 0.52 pearson correlation reported in the original Cassiopeia study (Fig. 4.5) [18]. Second, we confirmed that specific character states consistently aligned with clades in the inferred phylogeny, supporting the biological coherence of the tree structure (Fig. 4.6).

Utilizing the resolved trees for containing both untreated and persister cells, we aimed to use the phylogenies to confirm the previous observation that persistence potential is a

heritable trait. To this end, we suggested that a heritable trait would result in spatial autocorrelation and non-random clustering of persister cells on the lineage tree. To this end, we calculated a significantly positive Moran's I for each of our trees (Fig. 4.7 - Moran's I: Clone I: 0.018, Clone II: 0.0027, Clone III: 0.0117). We then generated simulated trees by shuffling the persister and untreated labels while preserving the structure of the trees, and compared the Moran's I from these simulated trees to the values from the real trees. The results show a significant spatial autocorrelation and clustering of persisters on the lineage trees across all clones, validating the notion that persistence potential is a heritable, non-randomly distributed trait within our trees (Fig. 4.7).

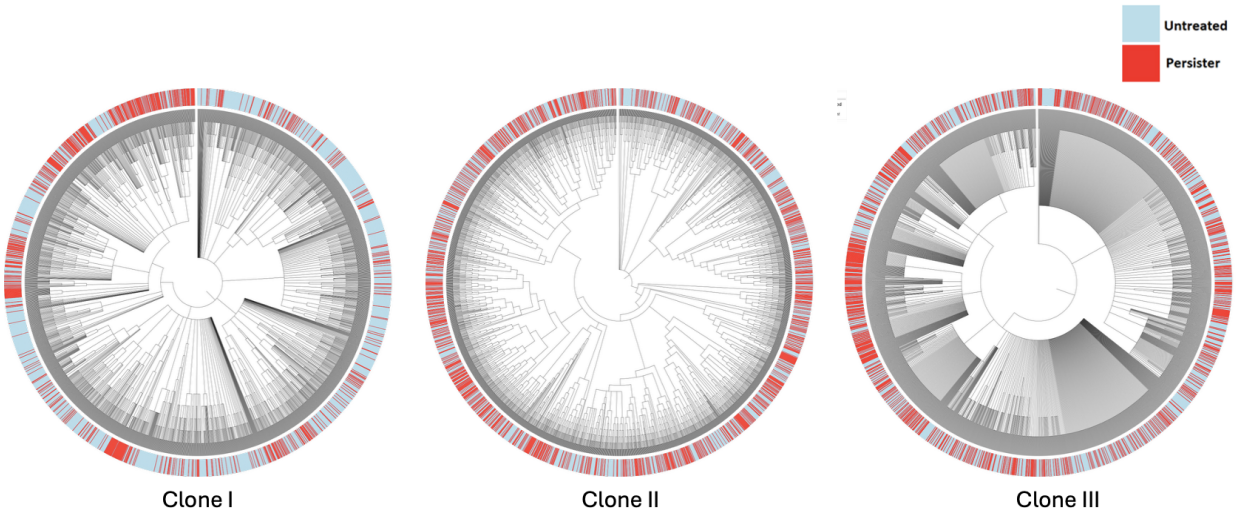

Figure 4.4: High resolution single-cell phylogenies are generated for 3 clones, leveraging the shared indels across cells. Cassiopeia Hybrid, an algorithm that integrates both a greedy and exhaustive maximum parsimony phase of tree building, is used to build the trees.

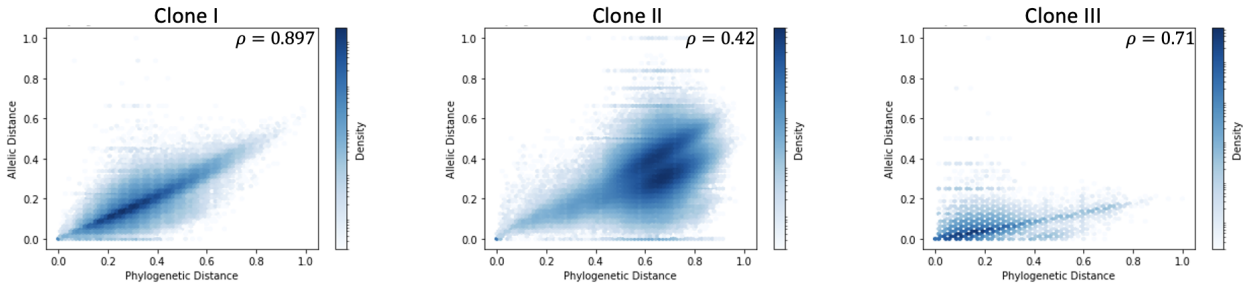

Figure 4.5: There is a positive association between the node distances inferred from the reconstructed trees and the mutational allelic distances for all three clones.

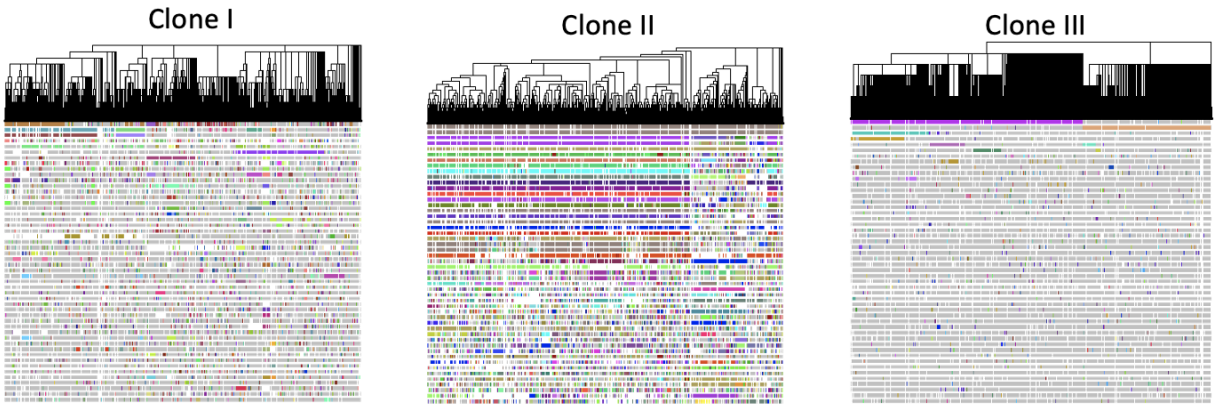

Figure 4.6: The reconstructed phylogenies are defined by indel character states for all three clones.

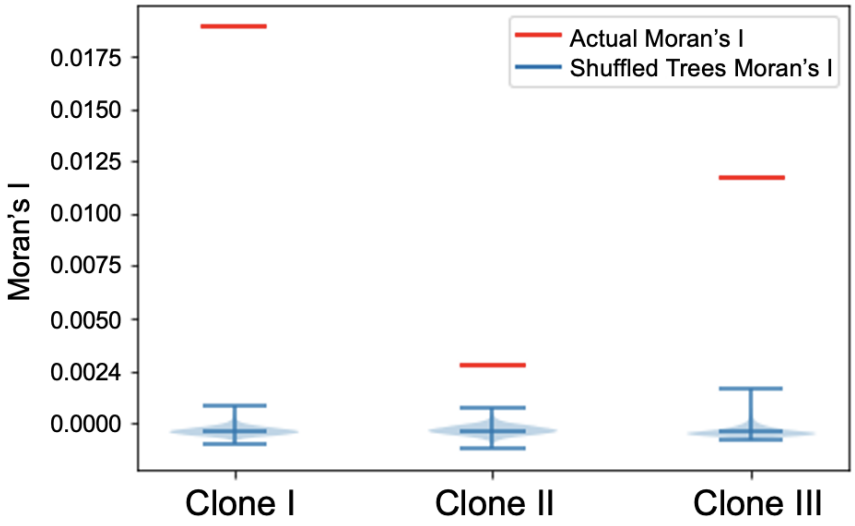

Figure 4.7: Binary Moran's I demonstrates the heritability of persister positioning along the phylogeny. The actual Moran's I values calculated from the observed trees are shown in red, while the violin plots represent the distribution of Moran's I values obtained from 10,000 random shufflings of the tree.

4.3.2 Identification of Persistence Potential Differences across Untreated Cells

Utilizing the paired lineage and expression information per cell, we then sought to identify the gene expression changes between untreated cells with high and low persistence potential. We devised approaches for exploring two possible modes for evolution of persistence potential; rare,

discrete & deterministic versus frequent, continuous, transient changes. Taking advantage of the single-cell nature of the data, we first sought to define a single-cell score that reflects the propensity for an untreated cancer cell to persist under treatment. We reasoned that cells that share a similar evolutionary trajectory as many persisters are more likely to persist under treatment relative to cells that are more phylogenetically distant to persister cells. As a result, we assigned persistence potential scores to each of the untreated cells in the trees, reflecting their phylogenetic distances between persister cells in the tree. For each untreated cell, we defined the persister score as the sum of the inverse node distances to every persister cell in the tree (see Methods). Untreated cells that are phylogenetically closer to more persisters will receive a higher persistence potential score and those that are more distant from persisters will have a lower score (Figures 4.8, 4.9).

To identify larger discrete events, we then applied a bottom-to-top "clade-level" analytical approach using iterative fisher exact tests that identifies clades that are significantly enriched or depleted for persister cells relative to their closest "sister clades" (see Methods). We identified 9 such events across the 3 clonal trees (Fig 4.10).

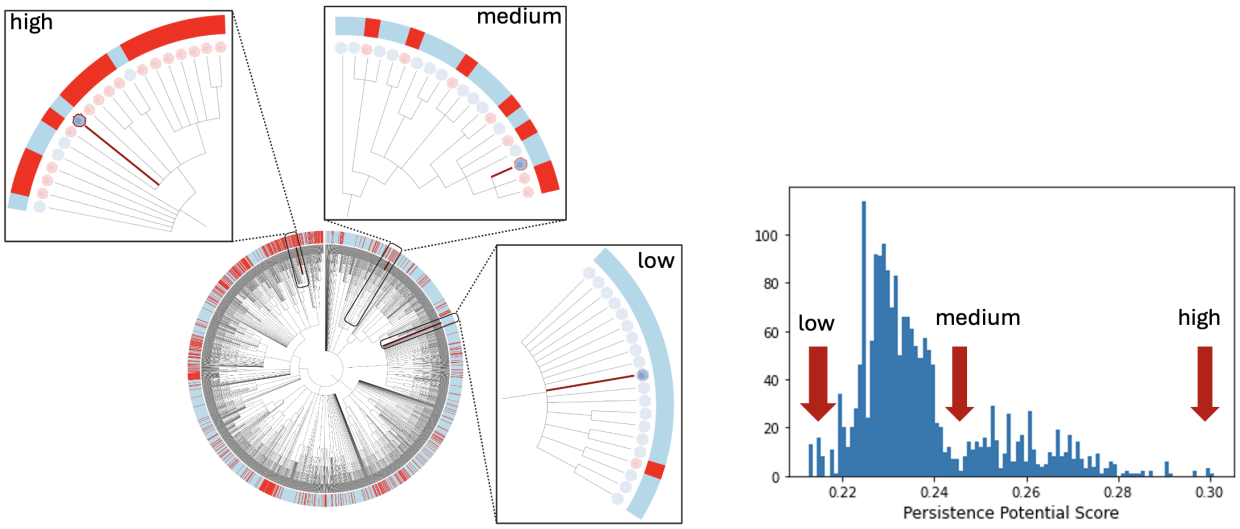

Figure 4.8: Intuitive representation of persistence potential score formulation (see Methods). Cells that are phylogenetically closer to more persisters will receive a higher score, while cells that are phylogenetically distant to persisters will receive lower scores.

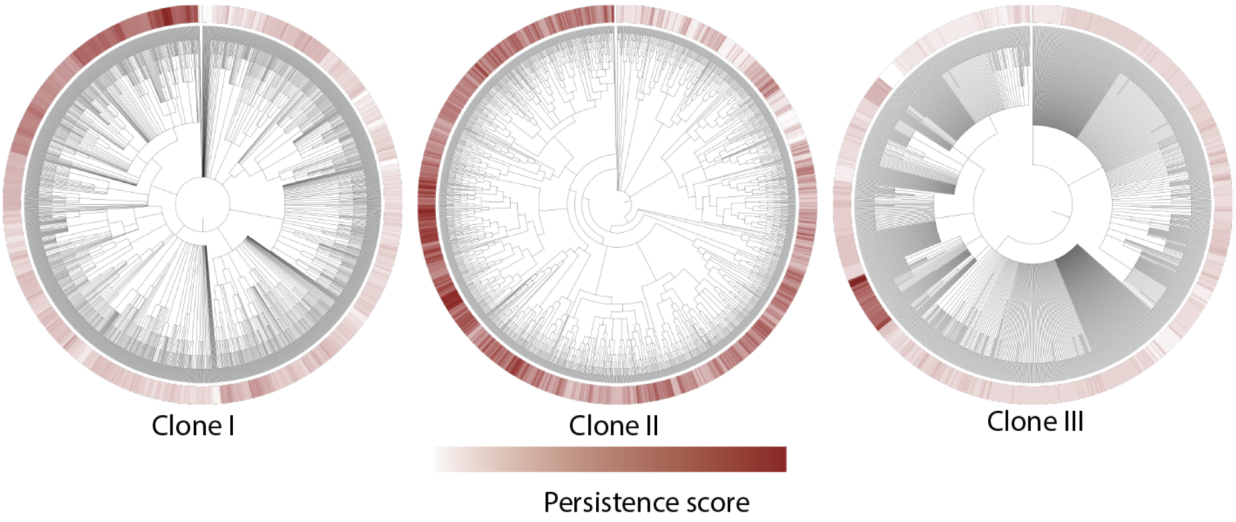

Figure 4.9: Clonal phylogenies shown are pruned to only contain untreated cells. Each cell has its respective persistence potential score mapped onto it. Untreated cells that are closer to more persisters (see Fig. 4.4) receive higher scores.

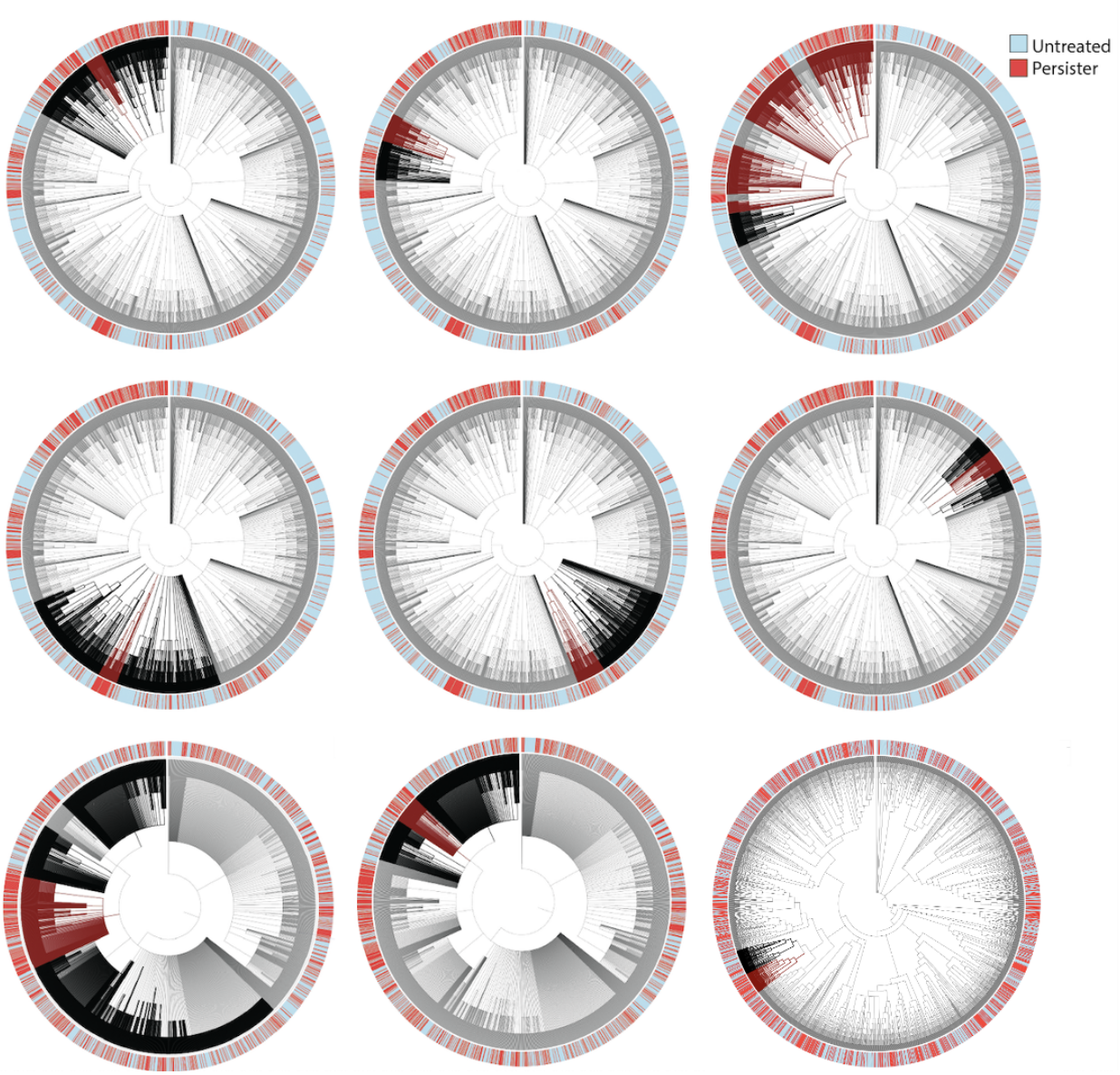

Figure 4.10: Nine identified events that are enriched or depleted for persisters identified using bottom-to-top tree traversal Fisher's exact tests looking for differences between closely related clades.

4.3.3 Expression Modules Driving Persistence Potential

Genes Associated with Persistence Potential

To identify expression changes associated with persistence potential, we employ differential expression analysis strategies catered to the two approaches for identifying differential potential

in the tree, mentioned above. First, differential expression analysis, using one-sided Wilcoxon rank-sum tests, was performed for each of the 9 clade events, comparing the untreated cells across the clades (high versus low persister clades) within each event. We identified genes that are consistently differentially expressed across the events by calculating the combined fisher p-value across the 9 events, combining the results of the 9 statistical tests into one combined p-value per gene. Comparing these Fisher combined p-values to a uniform distribution of p-values, revealed genes that are consistently differentially expressed more than what would be expected by chance. Several mitochondrial genes, such as *MT-ATP6* & *MT-CO3*, were identified as positively associated with persistence potential, while several translation-related genes, such as *RPL29* & *RPL14*, were identified as negatively associated with persistence potential (Figures 4.11, 4.12). This suggests that hallmarks that are known to be associated with DTPs can preexist in cancer cells prior to treatment, as both OXPHOS dependence and global repression of translation (decreased protein synthesis) are known features of DTPs [98, 103].

Using our persister potential scores, we then performed a poisson regression analysis for each clone to identify genes that associate along the continuous persistence potential axis. We also identified that mitochondrial genes were consistently positively associated with persistence potential (Fig. 4.13). Furthermore *HSP90* (*HSP90AA1* & *HSP90AB1*) consistently appears as positively associated with persistence potential in both the single-cell poisson regression as well as in the "clade-level" differential expression analysis (Fig. 4.14). This result agrees with another known feature of DTPs in which heat shock proteins can serve to protect tumor cells from harmful effects of protein denaturation caused by stressful drug conditions [104, 105].

When comparing the Fisher p-values from both differential expression strategies for identifying drivers of persistence potential, we see in both approaches that mitochondrial and ribosomal genes are both significantly associated with persistence potential (Fig. 4.14). We also see genes that are uniquely significant to either clade-level approach or the single-cell

approach.

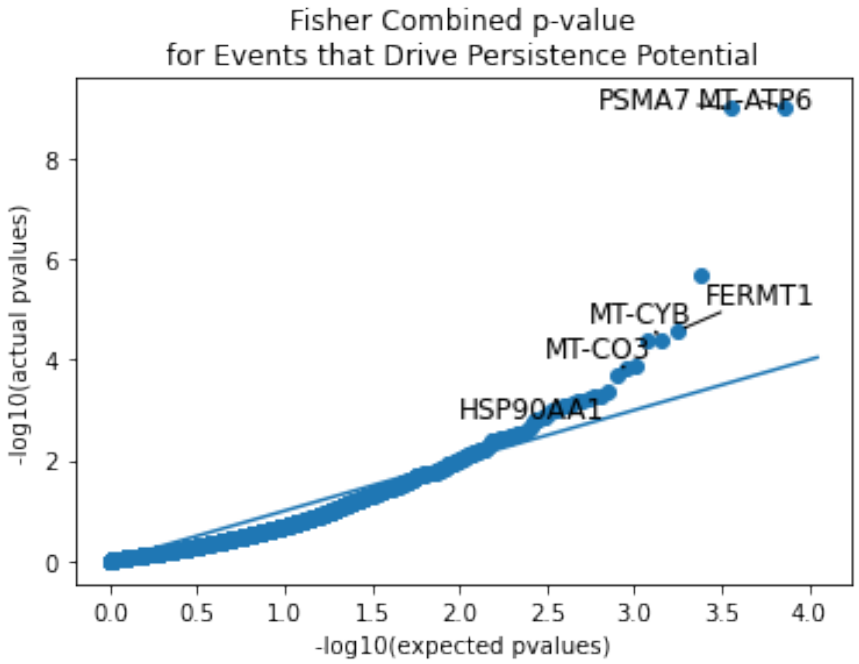

Figure 4.11: QQ plot demonstrating genes with greater differential expression than expected by chance, where combined Fisher's p-values were derived from separate differential expression tests conducted for each identified clade event. The tests here were one-sided tests for genes that are greater in persister enriched clades relative to its depleted "sister clade." Mitochondrial genes are consistently higher in persister enriched clades.

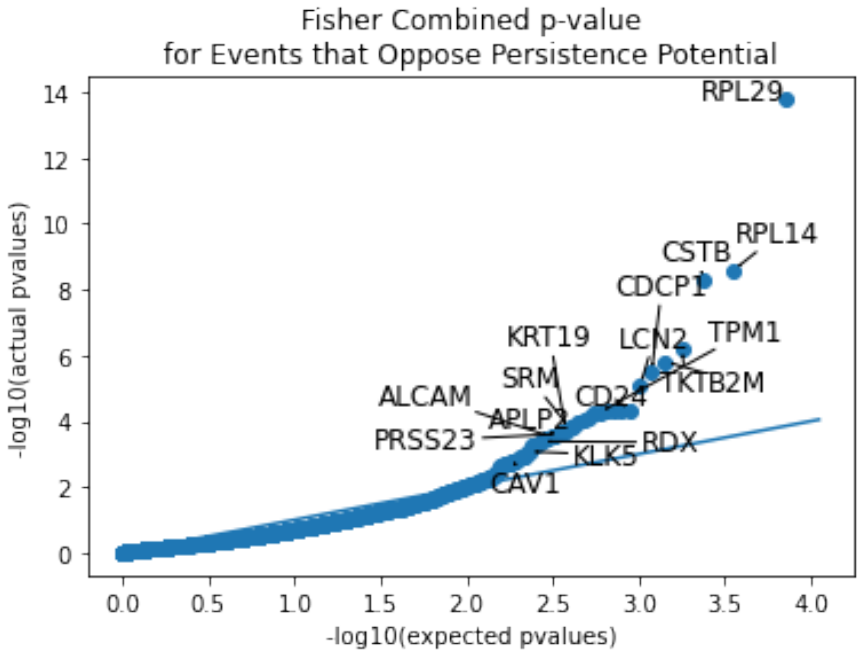

Figure 4.12: QQ plot demonstrating genes with greater differential expression than expected by chance, where combined Fisher's p-values were derived from separate differential expression tests conducted for each identified clade event. The tests here were one-sided tests for genes that are lower in persister enriched clades relative to its depleted "sister clade." Ribosomal genes are consistently lower in persister enriched clades.

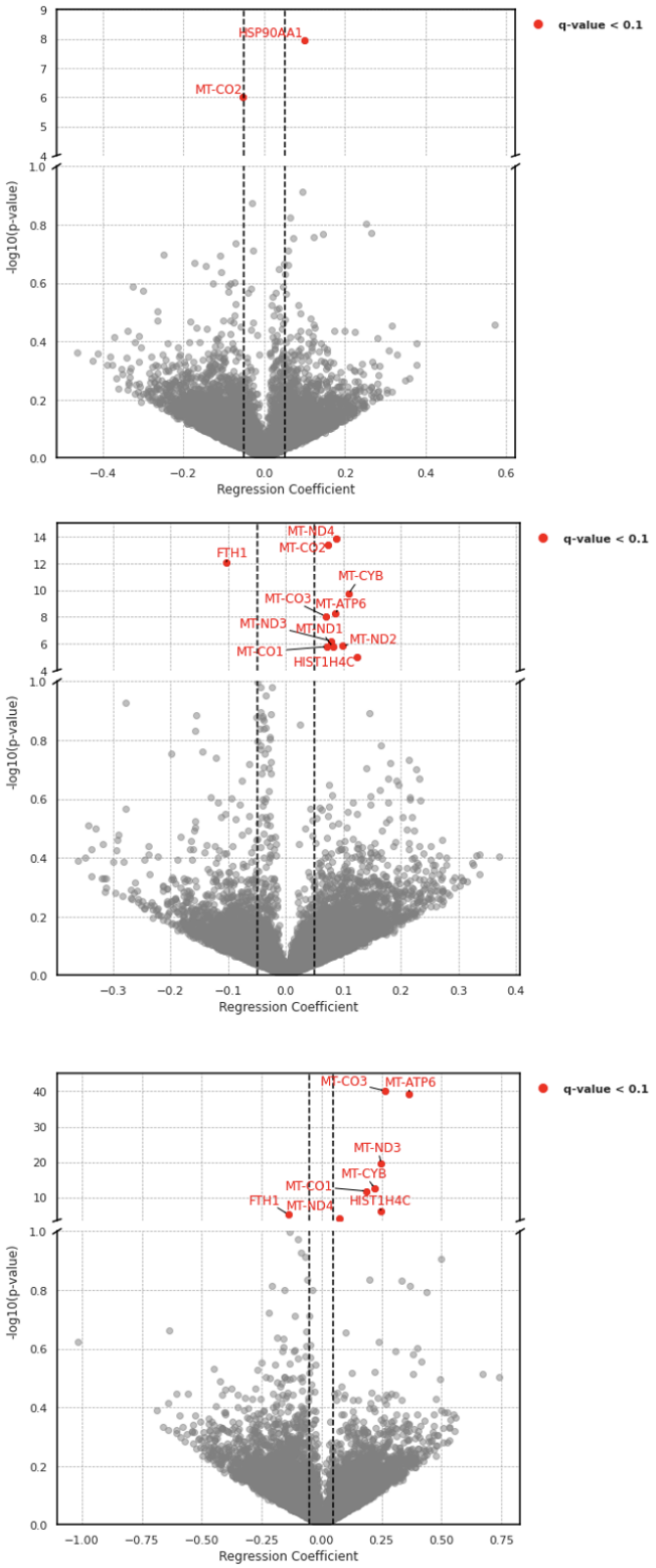

Figure 4.13: Genes differentially expressed along the persistence potential score axis, identified using p-values from the Poisson regression analysis. Mitochondrial genes are consistently positively correlated with persistence potential across clones, although with a relatively small effect size.

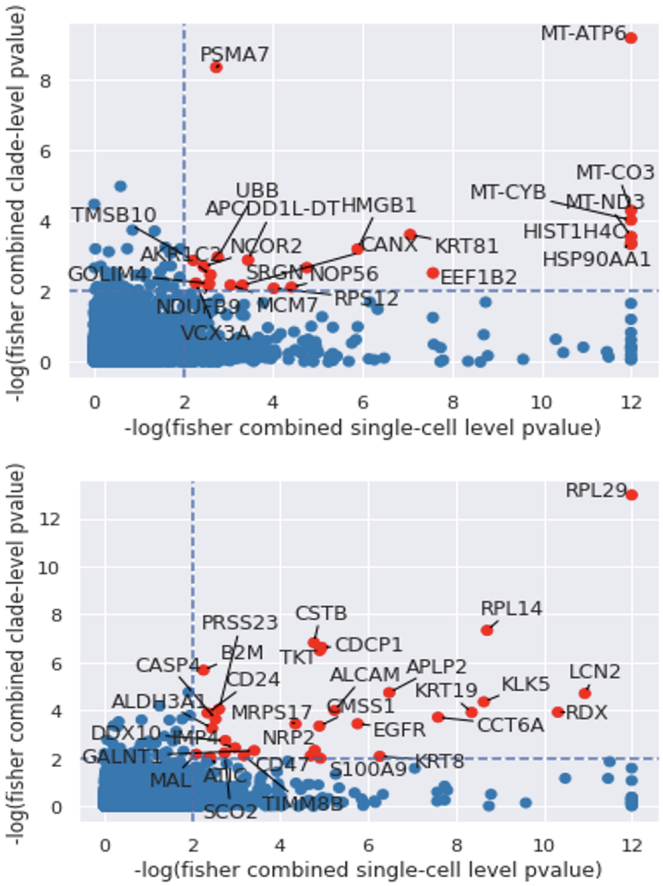

Figure 4.14: Comparing most consistently differentially expressed genes between clones in the clade-level analysis and the single-cell poisson regression analysis. There is both consistency and uniqueness in the identified genes between the two approaches.

Pathways Associated with Persistence Potential

Next, we sought to study gene modules that associate with persistence potential based on the distribution of expression of these modules within the trees. We first utilized AUCell to assign "AUC scores" for select gene programs to each cell and then adopted the "clade-level" analysis strategy to find the gene programs that are differentially expressed across untreated cells in our 9 identified significant persister enriched/depleted clades [106]. Interestingly, we identified cancer hallmark pathways that are both positively and negatively associated with persistence potential, many of which are consistent with the known literature of DTPs. The OXPHOS Hallmark Pathway was identified as positively associated with persistence

potential (Fig. 4.15). This result further aligns with previous literature, which discuss that DTPs have been shown to rely more on OXPHOS for energy production [98]. This finding also orthogonally supports our gene-level findings of mitochondrial genes' positive association with persistence potential, as OXPHOS occurs in mitochondria and the genes included in the OXPHOS pathway do not include mitochondrial genes.

We then applied Hotspot to discover de-novo gene modules that possess phylogenetic signals within the trees [107]. The method identifies genes & modules that have non-random expression patterns within the tree. The modules contain genes that co-occur within the phylogeny, as demonstrated in the Z-Score standardized autocorrelation matrix for Clone I (Fig. 4.16). For each of the clones, Hotspot strikingly identified modules enriched for translation related genes that were consistently anti-correlated with the persistence potential scores (Figures 4.17, 4.18). The enrichment of translation related genes in the Hotspot modules anti-correlated with persistence potential can also be seen in the summary statistics from the Hotspot analysis (Tables 4.1-4.3). Given the phenomenon of global repression of translation within DTPs, this provides yet another line of evidence of untreated cells being primed with known features of DTPs. Furthermore, proteosomal subunit PSMA7, appears again in this analysis as belonging to a module that is strongly positively with persistence potential, consistent with findings from the "clade-level" analysis (Figures 4.11, 4.17, 4.18, Table 4.4). These findings strongly suggest that cells that have a higher propensity to persist under treatment have reduced protein production, potentially as a means of being primed to resist stress and minimize energy expenditure.

Isoforms Associated with Persistence Potential

Since we reasoned that the mechanisms that drive preexisting persistence potential can exist on multiple levels of biological control (i.e., expression, epigenetic, etc.), we wanted to investigate the role of differential isoform usage in driving persistence potential using single-cell Mas-Iso seq [108]. For two of our biggest clades from our "clade-level" analysis, we looked

for differential isoform usage between the persister enriched and depleted clades, using the Jensen Shannon metric, a metric for assessing differences between distributions (see Methods). We observed that in both clones, the isoforms from genes implicated in ubiquitination and ubitiquine-proteasome pathways were imbalanced between the persister high and low clades (Figures 4.19, 4.20). Given our other observations of suppressed translation and upregulation of the proteasomal subunit, PSMA7, in high persistence potential cells, we suspect that these isoforms could be playing a role in working with the proteasome to clean up corrupted proteins, helping to maintain the health of the viability of the cell. While it would be of great interest to validate some of these findings, perturbing individual isoforms is not experimentally trivial. However, this analysis serves as a template for exploring differential isoform usage in single-cell phylogenies, and to our knowledge, is the first analysis of its kind, combining high resolution single-cell phylogenies with isoform expression.

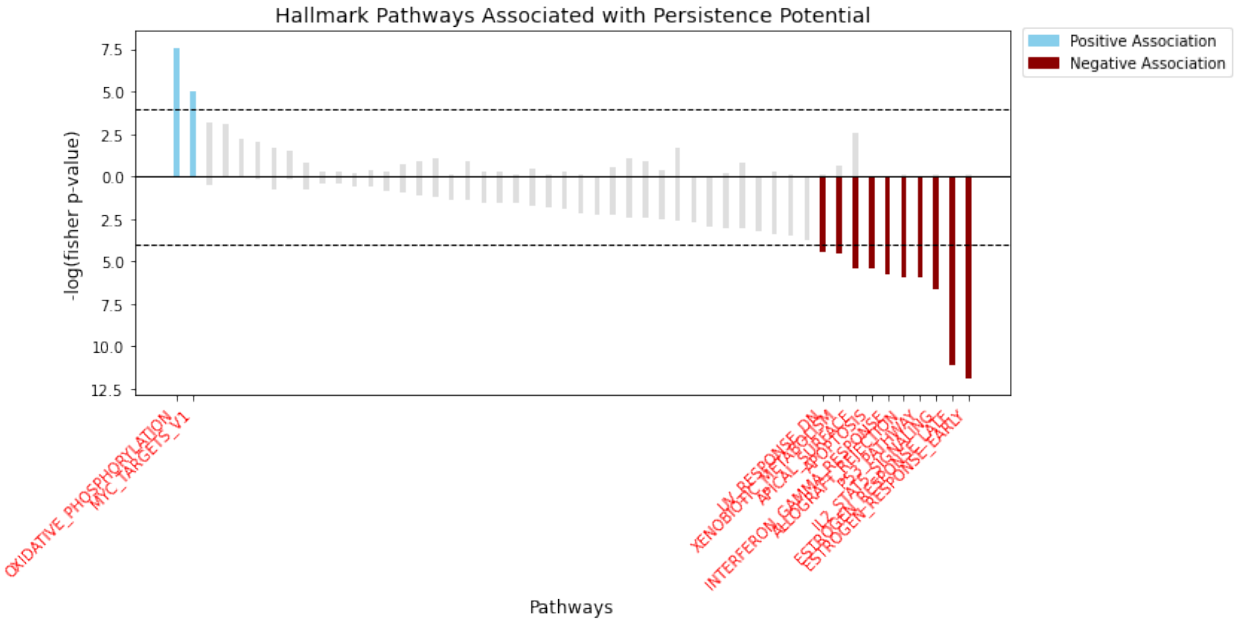

Figure 4.15: Hallmark Pathways that are differentially expressed across clade pairs in identified clade events. Oxidative phosphorylation was strongly associated with persistence potential, and estrogen response pathways were negatively associated with persistence potential.

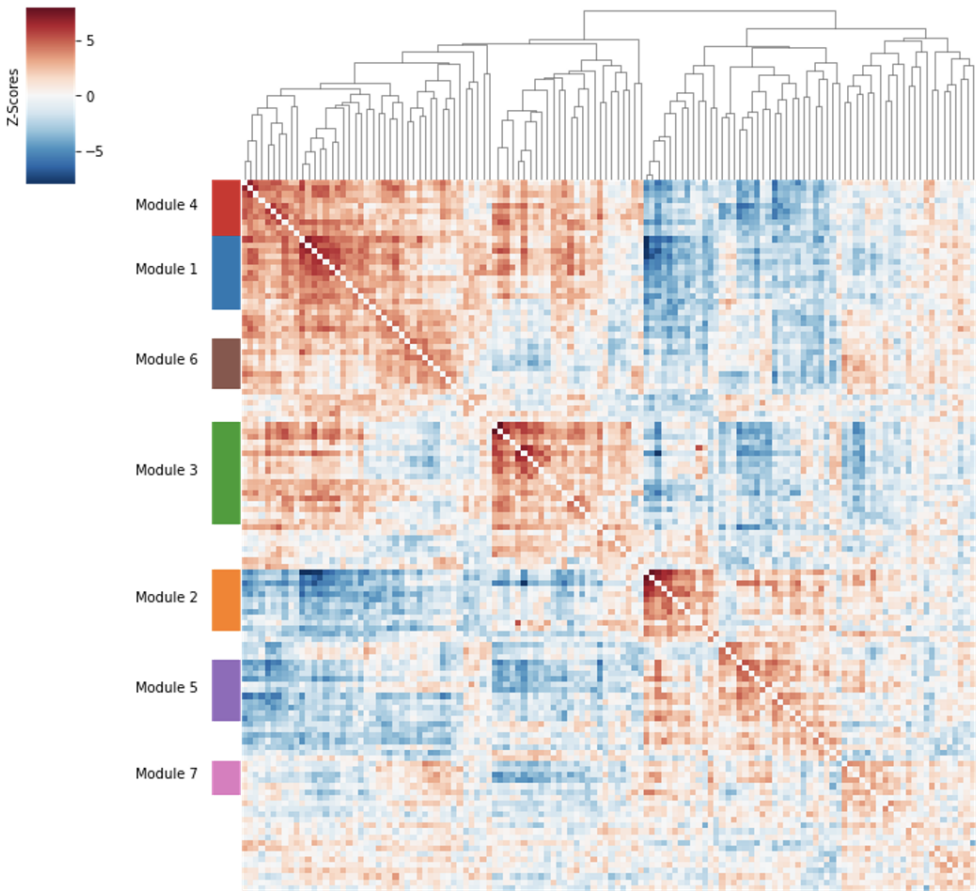

Figure 4.16: De-Novo identification of gene modules that associate non-randomly along the phylogeny using Hotspot [107]. Pairwise, genexgene autocorrelation matrix, highlighting modules of genes that phylogenetically co-occur in Clone I.

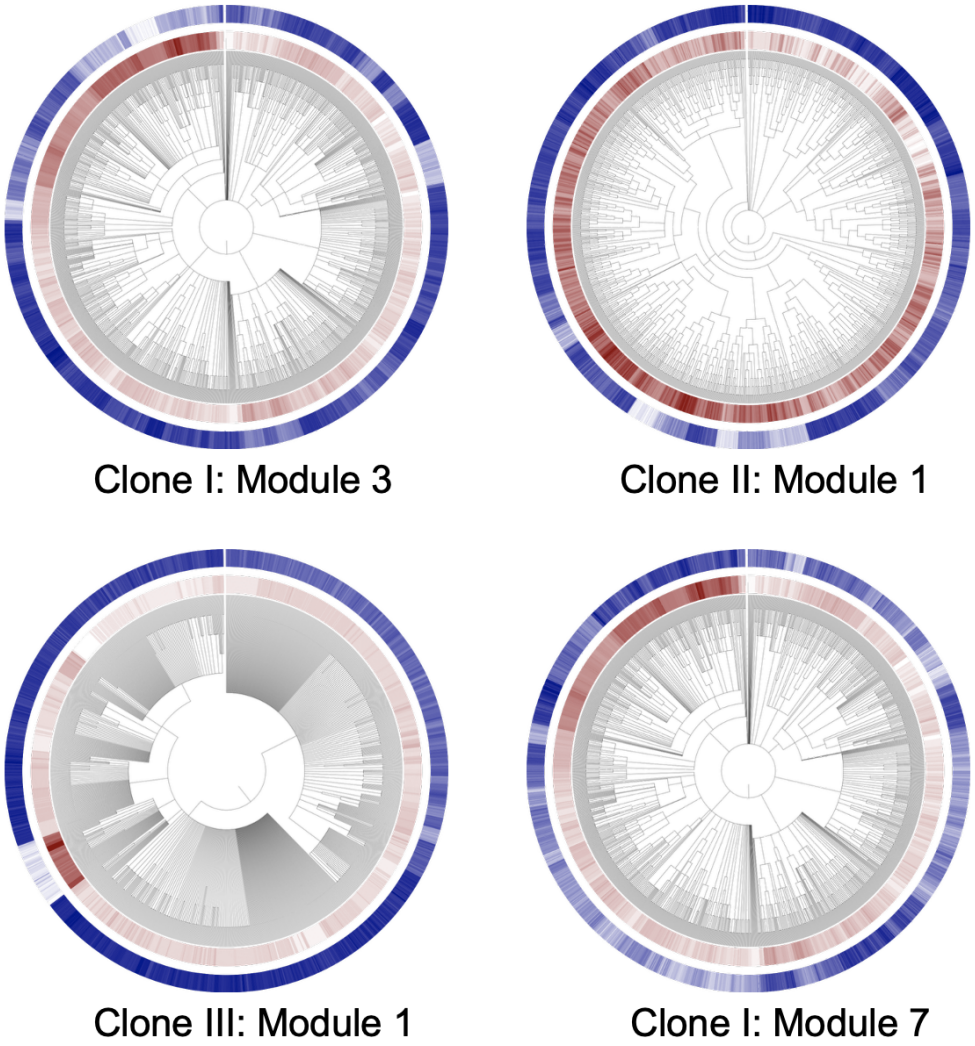

Figure 4.17: De-Novo identification of gene modules that associate non-randomly along the phylogeny using Hotspot [107]. Mapping the Module scores alongside the persistence potential score onto phylogeny visually demonstrate correlations between the two scores. The white-to-blue color gradients represent the modules scores and the white-to-red color gradients represent the persister scores.

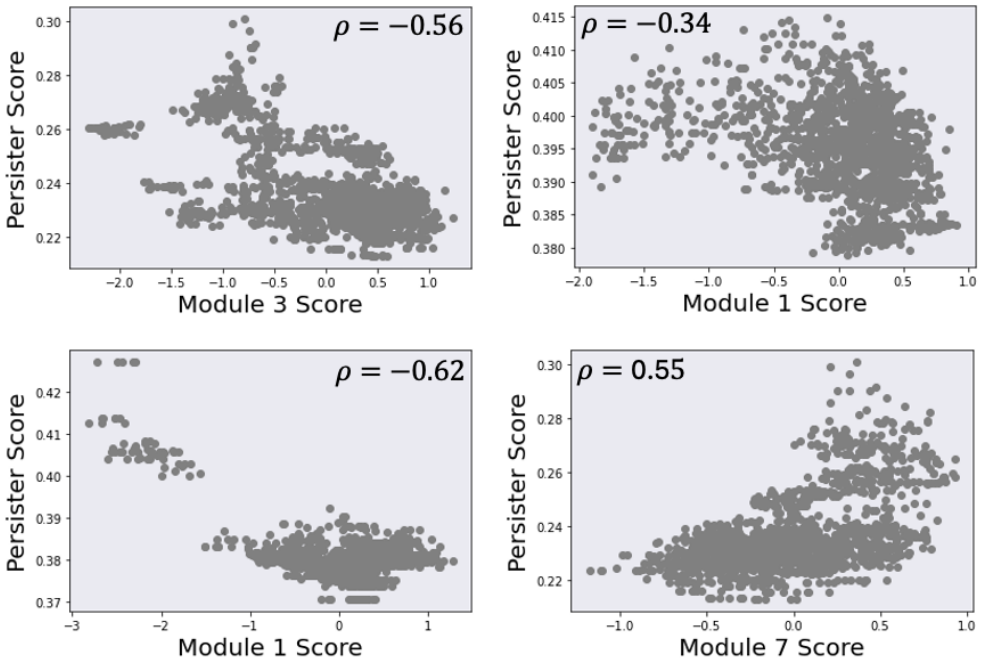

Figure 4.18: Hotspot Module scores positively and negatively correlate with persistence potential scores. The negatively correlated modules are strongly enriched for ribosomal genes. The one positively correlated module contains the*PSMA7* genes identified from "clade-level" differential expression analysis (Fig. 4.11)

Gene	Z	Pval	FDR	Module
RPL37A	15.800067	1.540665e-56	7.169484e-53	3.0
RPS25	15.371056	1.279807e-53	3.970388e-50	3.0
RDX	11.179251	2.567160e-29	9.960654e-26	3.0
APLP2	10.217716	8.259356e-25	9.608729e-22	3.0
EEF1B2	9.493093	1.121396e-21	1.048638e-18	3.0
TMSB10	6.103792	5.179042e-10	2.008289e-07	3.0
NRP2	6.003465	9.657517e-10	3.457020e-07	3.0
RAMP1	4.954738	6.321309e-07	2.699873e-05	3.0
APCDD1L-DT	4.758978	9.778206e-07	2.626347e-04	3.0
EI24	4.547682	2.709860e-06	5.374424e-04	3.0
ARL4C	4.115889	1.928449e-05	9.291346e-03	3.0
SEPTIN2	3.858940	5.673074e-05	7.650278e-03	3.0
ARHGEF12	3.807827	7.024994e-05	8.853535e-03	3.0
LRRFIP1	3.389821	3.496914e-04	1.341188e-02	3.0
TGFB1	3.196957	4.505764e-04	3.918947e-02	3.0
REXO8	3.268933	5.397695e-04	4.368387e-02	3.0
TIMP2	3.248283	6.081967e-04	4.773197e-02	3.0
SMYD3	3.225863	6.279617e-04	4.785823e-02	3.0

Table 4.1: Negatively Associated Hotspot Gene Module for Clone I

Gene	Z	Pval	FDR	Module
RPS27	8.746592	1.099461e-18	2.422112e-15	1.0
EEF1A1	6.343395	1.123781e-10	1.237844e-07	1.0
RPS25	5.729468	5.037297e-09	4.035333e-06	1.0
RPL24	4.776916	8.900223e-07	5.537489e-04	1.0
S100A6	4.693858	1.340504e-06	7.105156e-04	1.0
S100A11	4.689298	1.370718e-06	7.105156e-04	1.0
RPS21	4.447283	4.348170e-06	2.016638e-03	1.0
RPS12	4.129022	1.821547e-05	6.489265e-03	1.0
FTH1	3.969149	3.606494e-05	1.177053e-02	1.0

Table 4.2: Negatively Associated Hotspot Gene Module for Clone II

Gene	Z	Pval	FDR	Module
RPL24	12.431634	8.799942e-36	7.921708e-32	1.0
RPL29	8.772591	8.730056e-19	3.923998e-15	1.0
CHCHD2	7.166152	3.856755e-13	8.679626e-10	1.0
NIT2	6.766593	6.592535e-12	1.186920e-08	1.0
CSTB	6.718176	9.200691e-12	1.384018e-08	1.0
RPSA	6.607628	1.952635e-11	2.511008e-08	1.0
RPL14	5.924494	1.566299e-09	1.762478e-06	1.0
RPL17	5.812823	2.910456e-09	2.191112e-06	1.0
CCT6A	4.545891	2.577331e-06	2.320113e-03	1.0
CDCP1	4.519232	3.198880e-06	2.055594e-03	1.0
FHIT	4.017776	2.332803e-05	1.314392e-02	1.0
LCN2	3.961905	3.717713e-05	1.968632e-02	1.0
CCT8	3.740033	9.189940e-05	4.136392e-02	1.0
ASCC3	3.654826	1.286782e-04	4.852127e-02	1.0

Table 4.3: Negatively Associated Hotspot Gene Module for Clone III.

Gene	Z	Pval	FDR	Module
GSTM3	4.613796	0.000002	0.000428	7.0
RTRAF	4.526752	0.000003	0.000581	7.0
TCP1	4.450579	0.000004	0.000797	7.0
RPS12	3.607649	0.000154	0.016719	7.0
FRMD6	3.519333	0.000216	0.021883	7.0
PSMA7	3.494109	0.000238	0.023800	7.0

Table 4.4: Positively Associated Hotspot Gene Module for Clone I.

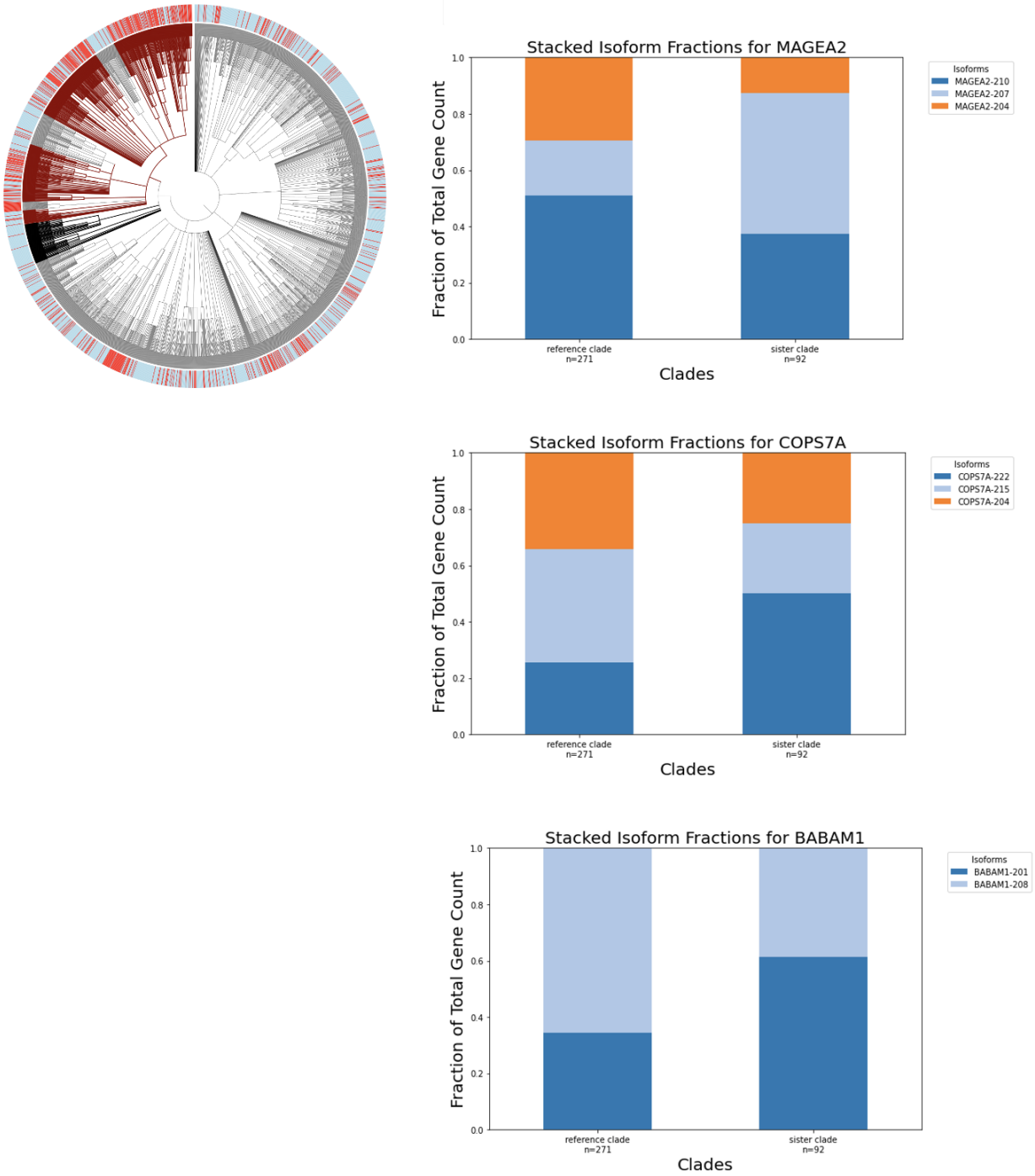

Figure 4.19: Identification of imbalanced isoform expression between persister enriched and persister depleted clades in largest clade-event in Clone I. Multiple ubiquitine pathway related genes containing imbalanced isoform expression.

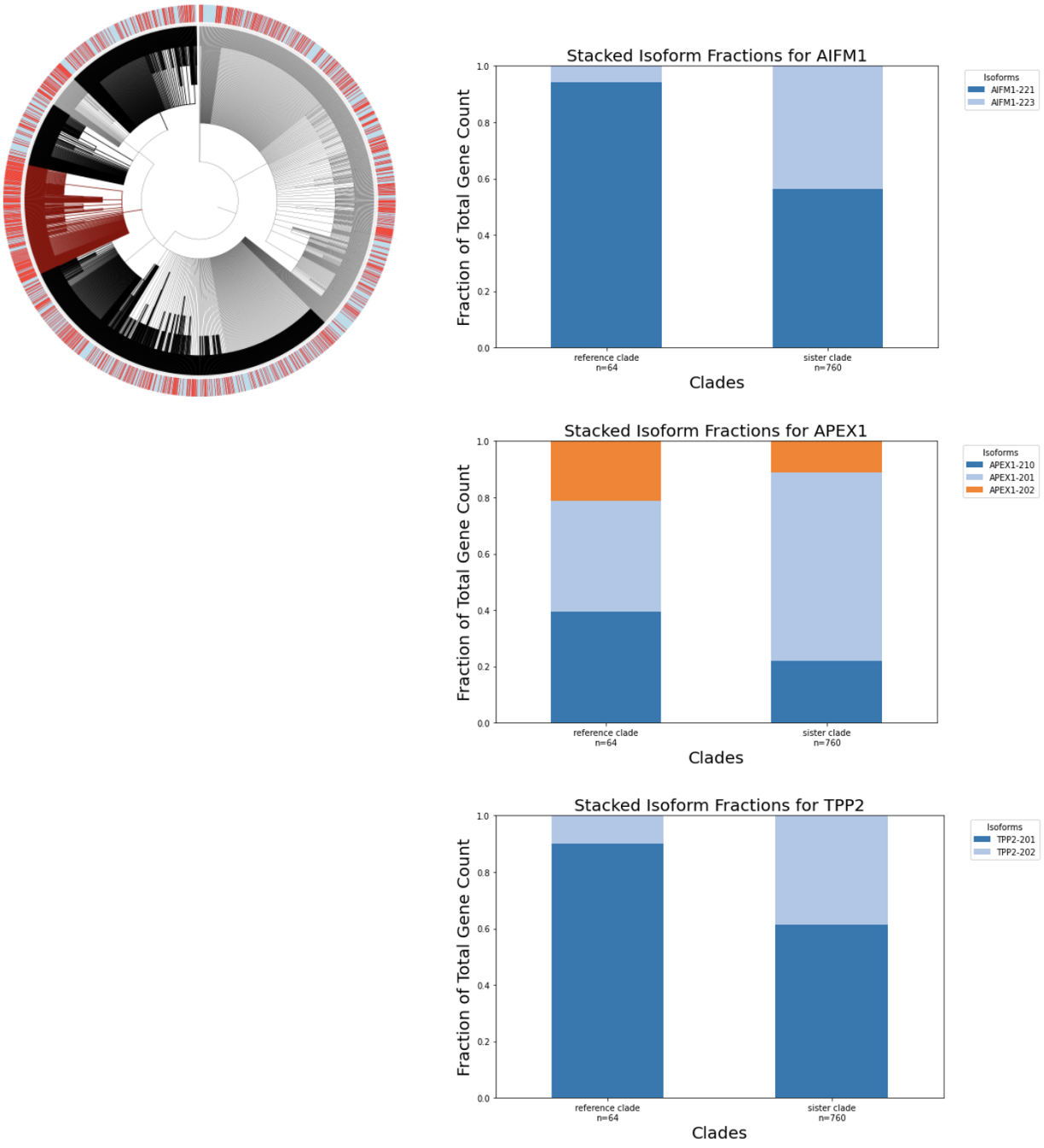

Figure 4.20: Identification of imbalanced isoform expression between persister enriched and persister depleted clades in largest clade-event in Clone I. Multiple ubiquitine pathway related genes containing imbalanced isoform expression.

4.3.4 Validation of Persistence Potential Associated Genes and Pathways

Functional Studies

We performed a number of validation experiments to confirm the different hypotheses generated from our analysis. Specifically, we sought to test the hypothesis of our most consistent "hits": mitochondrial/OXPHOS gene programs & ribosomal/translation gene programs. We performed pharmacological functional assays using ribosomal and OXPHOS inhibitors (Homoharringtonine (HHT) & bedaquiline, respectively) to test whether these interventions could further alter the prevalence of persisters in combination with osimertinib. We identified that the combination of osimertinib with bedaquiline reduced persisters more than expected by the null hypothesis of simple additivity of effect of two independent interventions (Figures 4.21, 4.22, 4.23). The combination treatment of bedaquiline and osimertinib synergistically reduced persisters in both PC-9 and HCC827 cell lines (Figures 4.21, 4.22, 4.23; synergy ratios: 4.11, 3.35 in PC-9 and HCC287, respectively). This result suggests that OXPHOS is a driver of persistence potential, a hypothesis supported by our differential expression analysis. Interestingly, we performed, in parallel, a similar combination treatment experiment with HHT and osimertinib, and found evidence of antisynergy (Fig. 4.23: synergy ratios of 0.74, 0.73 in PC-9 and HCC287). It appears that inhibition of ribosomal activity opposes the effects of osimertinib and actually increases the number of persisters. This result also complements our hypothesis that reduced ribosomal activity is a marker and potential driver of persistence potential.

Survival Analysis

We analyzed patient survival RNA-seq data from the CPTAC cohort to assess whether the genes and pathways we found to be associated with persistence potential have clinical significance. We looked at both progression free survival (PFS) and overall survival associa-

tions (OS). Strikingly, PSMA7 expression, which was found to be positively associated with persistence potential from the clade-level analysis, was strongly negatively associated with both PFS (p=0.0046) and OS (p=0.08) (Fig. 4.24). Furthermore, RPL14, one of the ribosomal genes strongly negatively associated with persistence potential, had a strong positive association with both PFS (p=0.0046) and OS (p=0.039) (Fig 4.25). The Hallmark Oxidative Phosphorylation and MTORC1 pathways, which we showed to have positive associations with persistence potential, had slight negative associations with PFS (p=0.2, p=0.11, respectively) (Fig. 4.25). The directionality of these implicated pathways across our different analyses appear to be in agreement, suggesting that genes and pathways that influence persistence potential prior to any treatment can have an effect on patient survival.

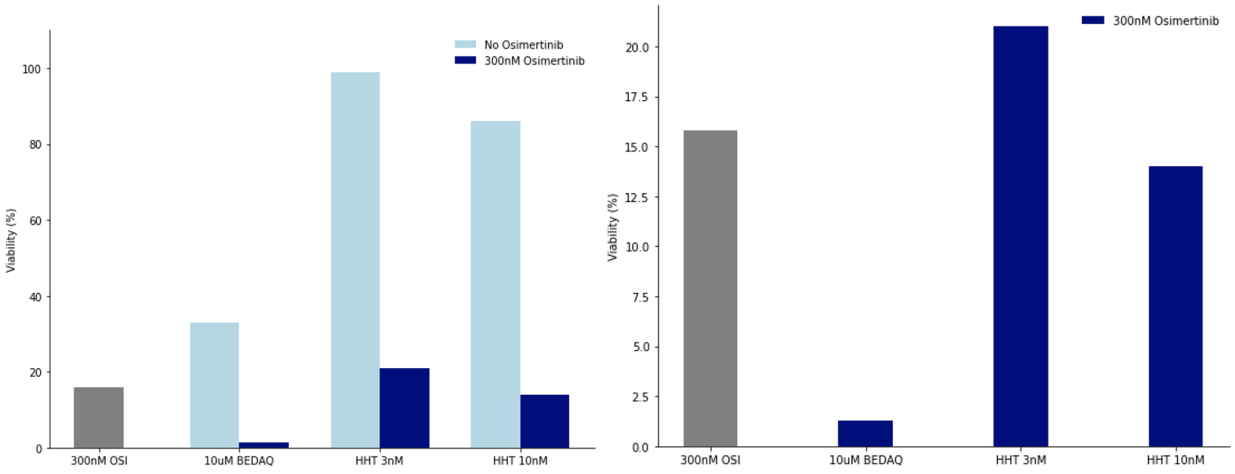

Figure 4.21: Effects of combination therapy of osimertinib plus inhibitor of different pathways on persistence in PC-9 cells. Left panel shows the effect of drug alone (light blue) and the effects of combination therapy (dark blue). Right panel shows data with just combination therapy to better visualize the effect with respect to the baseline of osimertinib alone. Bedaquilline (OXPHOS inhibition) further reduces persisters with respect to baseline, while HHT (ribosomal inhibitor) increases persisters slightly.

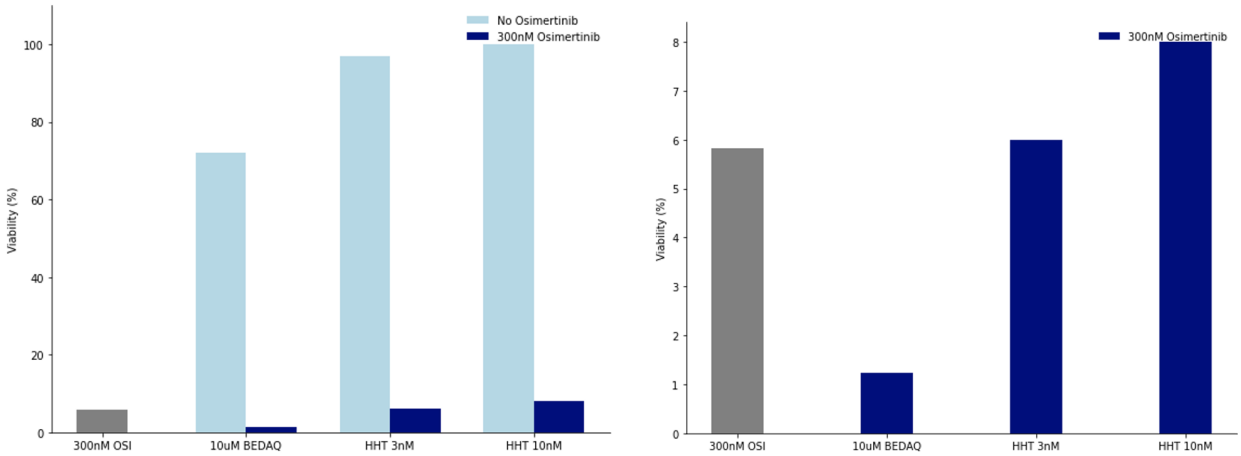

Figure 4.22: Effects of combination therapy of osimertinib plus inhibitor of different pathways on persistence in HCC827 cells. **Left panel** shows the effect of drug alone (light blue) and the effects of combination therapy (dark blue). **Right panel** shows data with just combination therapy to better visualize the effect with respect to the baseline of osimertinib alone. Bedaquilline (OXPHOS inhibition) further reduces persisters with respect to baseline, while HHT (ribosomal inhibitor) increases persisters slightly.

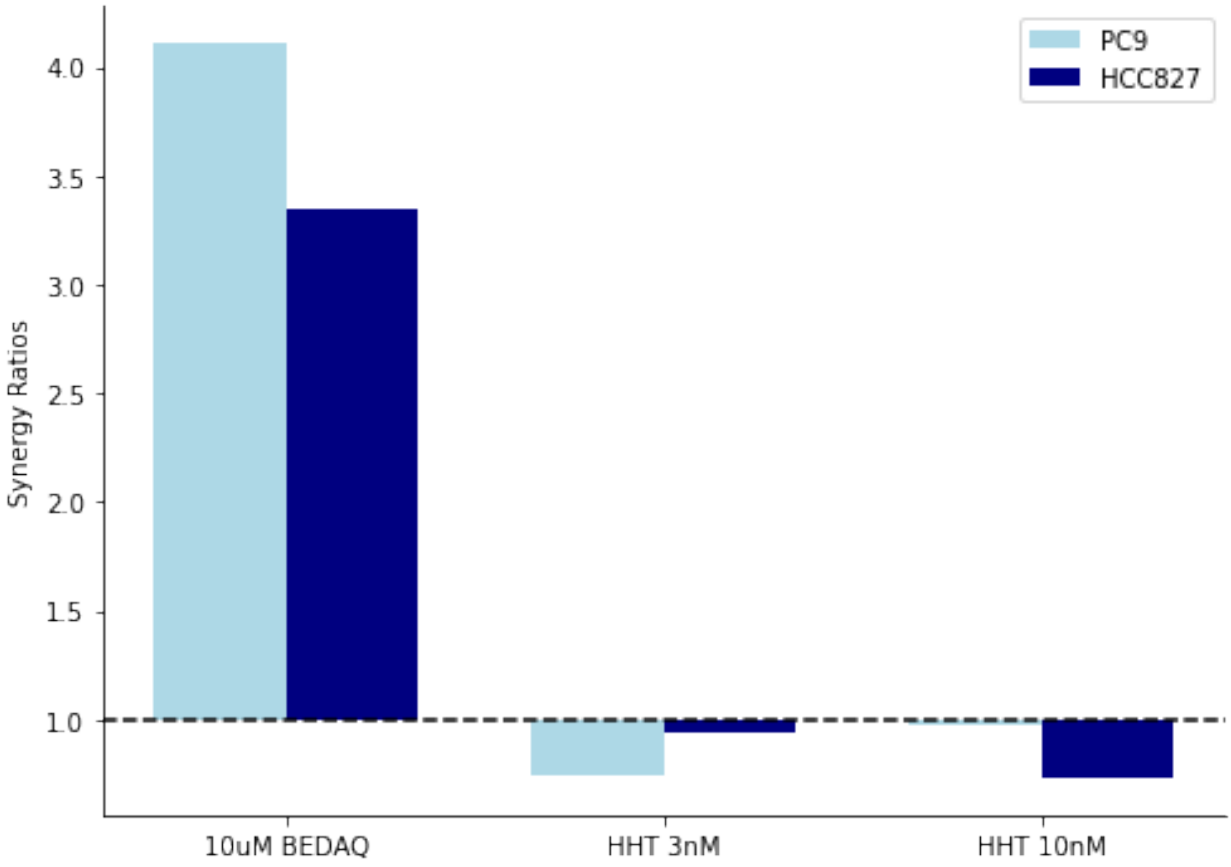

Figure 4.23: Synergy ratios in PC-9 and HCC827 validation experiments demonstrates synergistic killing of persister cells with bedaquilline and antisynergism with HHT. Effects were observed in both PC-9 and HCC827 cell lines.

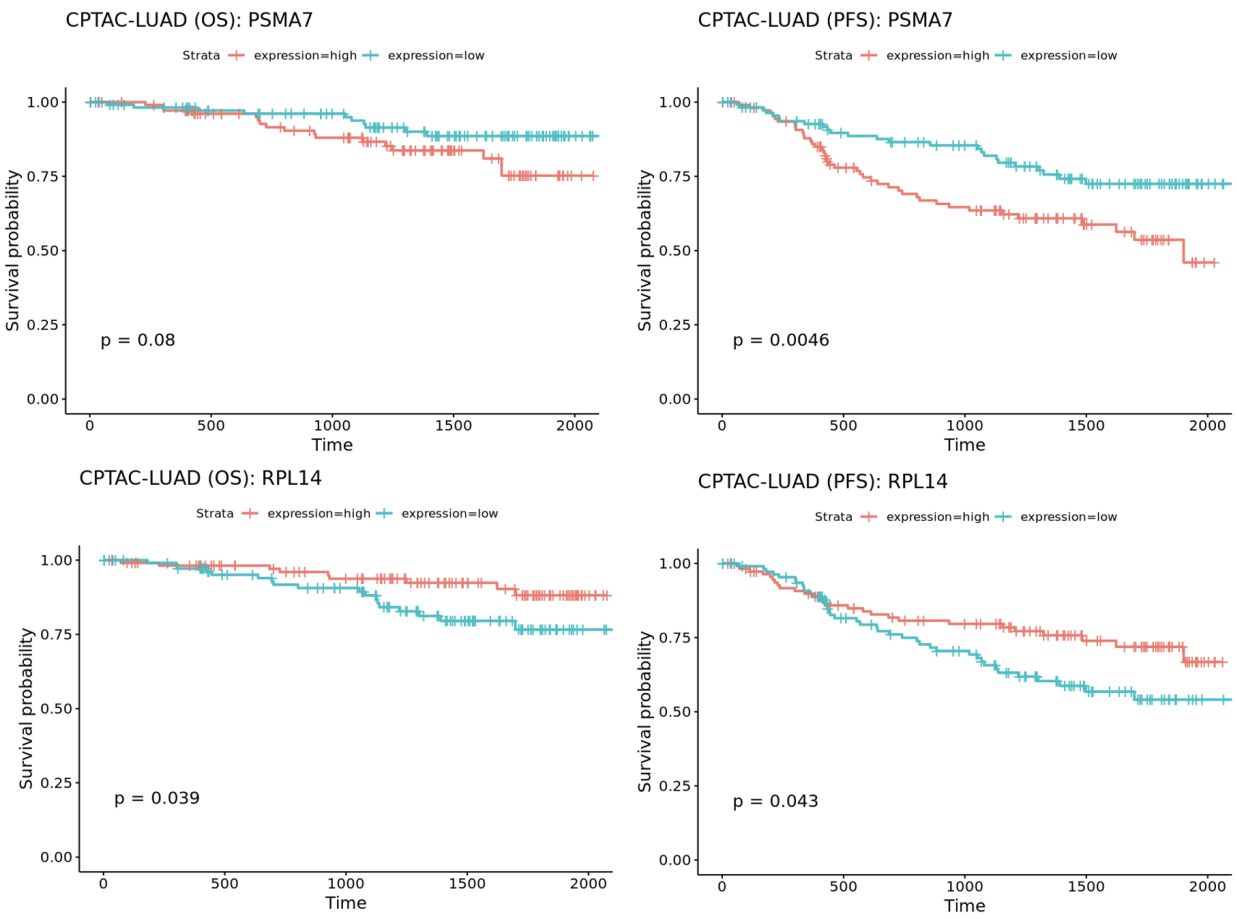

Figure 4.24: Identified associations between top persistence potential associated genes and survival in CPTAC LUAD patient cohort with EGFR driver mutations. Higher expression of PSMA7, a gene identified to be positively associated with persistence potential, leads to worse overall and progression free survival. Higher expression of RPL14, a gene identified to be negatively associated with persistence potential, leads to better overall and progression free survival.

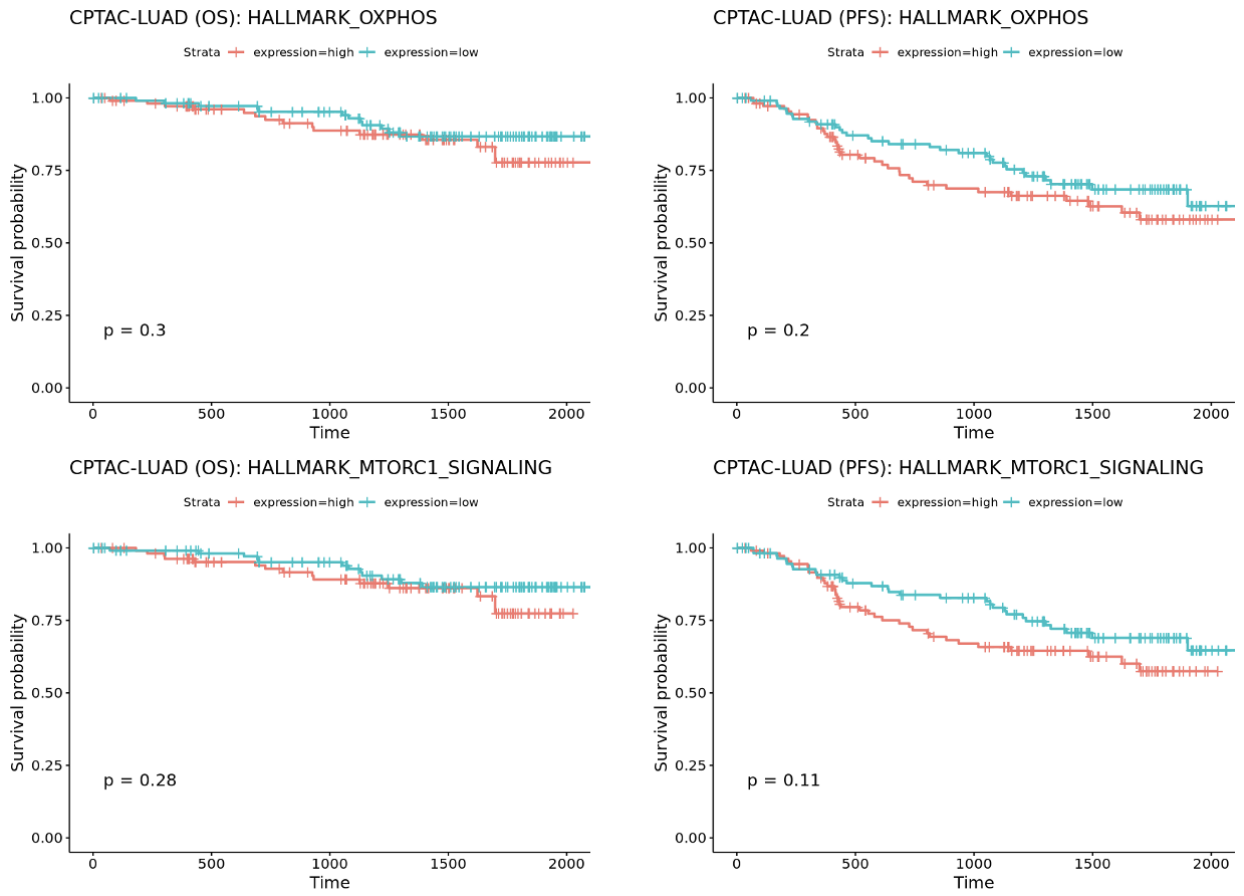

Figure 4.25: Identified associations between top persistence potential associated gene modules and survival in CPTAC LUAD patient cohort with EGFR driver mutations. Higher expression of Oxidative Phosphorylation and MTORC, gene modules identified to be positively associated with persistence potential, trends towards worse progression free survival.

4.3.5 Trajectory Analysis: Inferring Transition and Gene Expression Response along Untreated to Persistence Trajectory

The primary focus of this study was to investigate the evolution of persistence potential, prior to introducing cancer cells to drug treatment. However, we further speculated that there could be differential response to drug treatment detected within a phylogeny, potentially driven by persistence potential. To study this, we devised an analytical approach utilizing PCA to identify the genes that have the most differential response to treatment within the tree. As opposed to a typical cell x gene expression matrix, we created an untreated cell x

gene-response matrix as input to PCA. In order to generate the untreated cell x gene-response response matrix, for each element in the matrix, we sum the differences between the expression of the untreated cell and each persister, weighted by a phylogenetic weighting factor (see Methods). The weighting factor for each persister cell, with respect to a given untreated cell, is the inverse of the node distance between the untreated and persister cell of interest. This allows persister cells that are very distant to contribute very little to the difference to untreated expression and those that are close to contribute more. The essential logic was to devise an approach that captured the gene expression differences between untreated cells and the persisters it would most likely resemble if they were to persist.

Upon generating this matrix, we ran PCA and investigated the gene loadings that most strongly contributed to PC1 and PC2. Intriguingly, we found in all 3 clones that ribosomal gene responses and mitochondrial gene responses were strongly anticorrelated (Figures 4.26, 4.27). This finding suggests that there are multiple, potentially opposing, paths by which an untreated cell can respond to become a persister cell, relying on a translation-mediated and a mitochondria-mediated axis. This finding further compliments our findings that ribosomal & mitochondrial genes were negatively and positively correlated with persistence potential, respectively. Lastly, we also, interestingly, identified a number of genes that were consistently within the top 10 genes that were most differentially responsive along one of the PC axis: *GADD45A*, *GAS5* and *ATF3* (Tables 4.5, 4.6, 4.7). ZFAS1 was also identified within the top 15 differential responsive genes across all 3 clones (Tables 4.5, 4.6, 4.7). *GAS5* has been shown to regulate apoptosis in cancer cells, *ATF3* is a transcription factor involved in multiple stress response pathways and *ZFAS1* is a lincRNA implicated in multiple cancers and shown to have a negative association with patient survival in NSCLC.

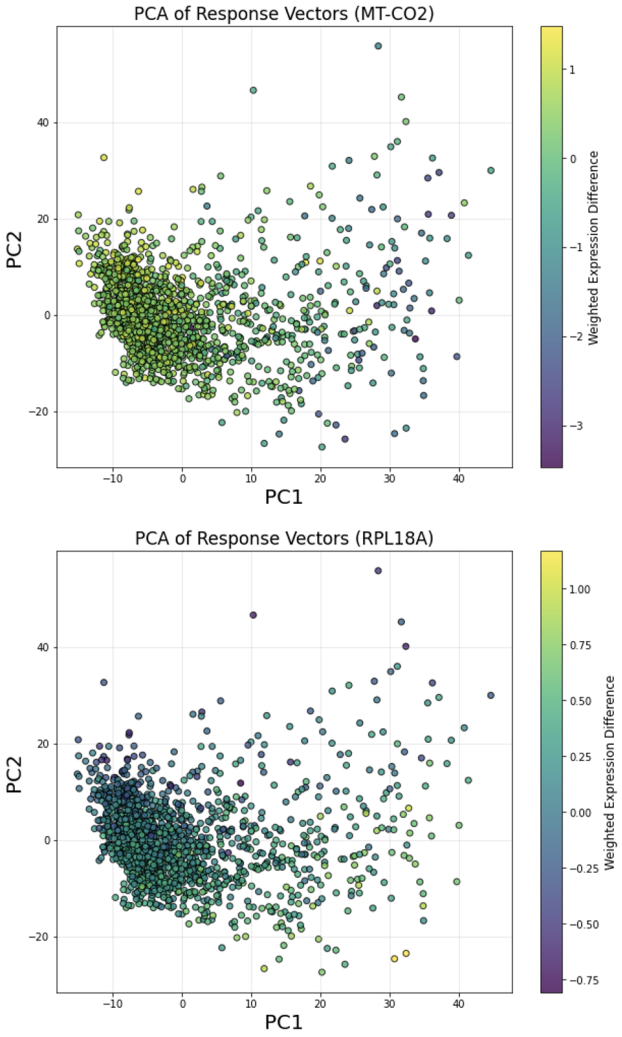

Figure 4.26: Visualization of the most differentially responsive genes response values in PC-space. Responses in mitochondrial genes and ribosomal genes are anti-correlated consistently across clones. Representative example in Clone I.

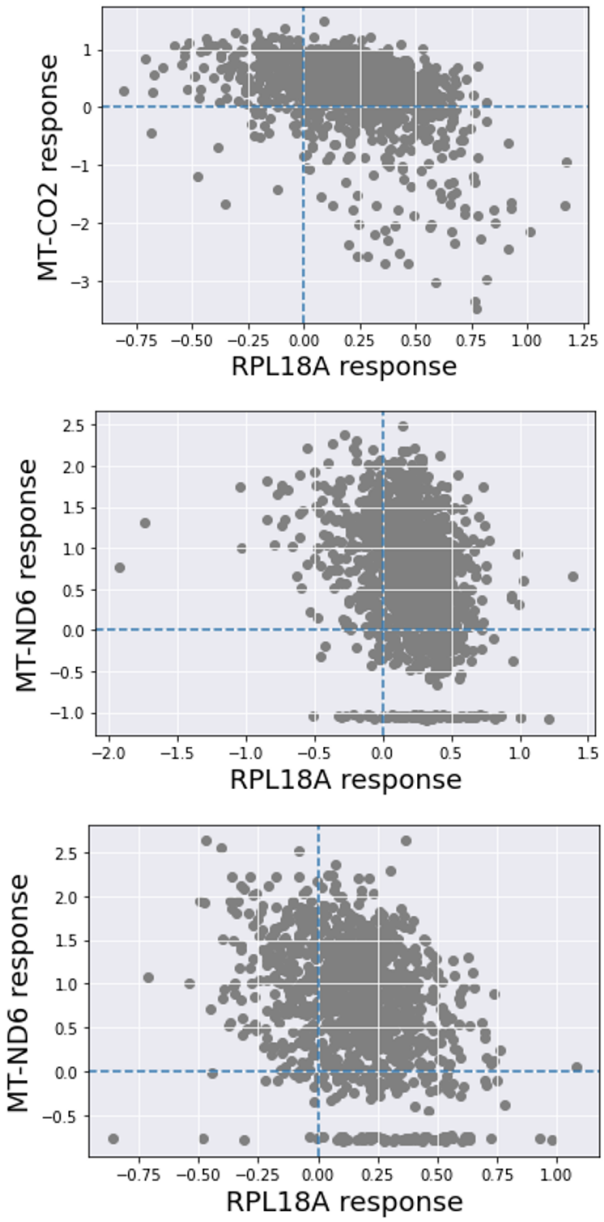

Figure 4.27: Correlation between most differentially responsive genes identified from trajectory analysis. There appears to be anti-correlation between ribosomal response and mitochondrial response, repeated across all three clones. (**Top Panel**: Clone I, **Middle Panel**: Clone II, **Bottom Panel**: Clone III,

Gene	PC1
GADD45A	0.060969
MT-CO2	-0.060429
MT-CO3	-0.060388
MT-ND4	-0.056828
TUBA1B	-0.055878
MT-ATP6	-0.055479
MT-ND3	-0.054948
PPP1R15A	-0.054190
MT-CO1	-0.054112
TUBB	-0.054062
PMEPA1	-0.052897
LDHA	-0.051303
GABARAPL1	0.050387
GAS5	0.049806
PKM	-0.049506
HMGB1	-0.049299
MT-ND2	-0.049010
MT-CYB	-0.048541
ACTB	-0.047808
MT-ND5	-0.045780
ATF3	0.045423
MT-ND1	-0.045225
TNFAIP3	0.045186
SAT1	0.045096
ZFAS1	0.044708

Gene	PC2
RPL39	-0.062801
GAPDH	-0.062072
RPL10	-0.059885
RPL18A	-0.059877
RPL8	-0.058868
RPS12	-0.058297
RPS4X	-0.057816
RPS3	-0.057688
RPL28	-0.056856
RPS14	-0.055574
RPL41	-0.054797
RPL37	-0.054734
RPS28	-0.054498
RPS23	-0.053908
RPS15	-0.053322
RPL13	-0.053077
RPL18	-0.053077
RPL30	-0.052719
RPS27	-0.052540
RPS19	-0.051867
GSTP1	-0.051832
RPL15	-0.051650
RPL29	-0.051559
RPS2	-0.051156
FTL	-0.050450
RPS29	-0.050264

Table 4.5: Top genes contributing to PC1 (left) and PC2 (right) for Clone I.

Gene	PC1
GADD45A	0.063056
GAPDH	-0.062480
MT-CO3	-0.061413
PPP1R15A	0.060732
DDIT3	0.058641
ACTB	-0.058181
EIF5	0.057395
MT-CO2	-0.056851
SNHG12	0.056364
MT-ND4	-0.056358
SNAPC1	0.055419
TUBB	-0.054938
IL32	0.054802
TMEM265	0.054770
MT-ATP6	-0.054293
MT-CO1	-0.053120
LDHA	-0.053009
ATF3	0.052906
MT-CYB	-0.052774
HSPA8	-0.052669
GAS5	0.051044
MT-ND5	-0.050892
SLC3A2	0.050115
MIF	-0.049916
MT-ND3	-0.049414

Gene	PC2
RPL18A	0.076018
RPL37	0.070192
RPS18	0.069511
RPS12	0.069229
RPS27	0.068506
RPS19	0.067877
RPL18	0.065776
RPL28	0.065117
RPL8	0.063251
RPS29	0.063059
RPS2	0.062851
RPL10	0.062476
RPS15	0.060941
RPL30	0.059450
RPS27A	0.058904
RPL13	0.058419
RPL19	0.057879
RPS15A	0.057178
RPS14	0.056935
RPL39	0.056915
RPS23	0.056090
RPL15	0.054651
RPS9	0.053675
RPL37A	0.055463
RPLP1	0.055431

Table 4.6: Top genes contributing to PC1 (left) and PC2 (right) for Clone II.

Gene	PC1
MT-CO2	-0.067747
GAPDH	-0.066474
MT-ND3	-0.060760
KRT7	-0.059845
GADD45A	0.058409
PPP1R15A	0.057704
EEF1A1	0.057426
MT-CO3	-0.057230
MT-ND4	-0.056777
ATF3	0.056118
MIF	-0.054636
CFL1	-0.053636
GAS5	0.052152
PRDX2	-0.052103
GSTP1	-0.051580
TUBA1B	-0.051128
MT-CYB	-0.050669
TPI1	-0.050342
NME2	-0.050336
KRT18	-0.049870
PHB	-0.048813
MALAT1	0.048117
TAF1D	0.047844
NDUFS6	-0.047555
EIF4A2	0.046712

Gene	PC2
RPL18A	0.073031
RPS19	0.071181
RPS18	0.070961
RPL28	0.069561
RPS12	0.069255
RPL8	0.068740
RPL10	0.067837
RPL18	0.067689
RPS27A	0.065822
RPL37	0.063223
RPL19	0.062697
RPL32	0.061988
RPS9	0.061363
RPL13	0.060386
RPS27	0.058627
RPS16	0.057655
RPL15	0.057582
RPL30	0.056892
RPLP1	0.056440
HNRNPU	-0.056190
MT-ND6	-0.055901
RPS15A	0.055034
RPS21	0.054897
RPL39	0.054856
RPL37A	0.054591

Table 4.7: Top genes contributing to PC1 (left) and PC2 (right) for Clone III.

4.3.6 Persistence Potential Genes and Pathways Association with Progression in scRNA LUAD Patient Cohort

We wanted to investigate the role of the genes and pathways implicated in our analysis in human patient data applied to studying cancer persistence. We found the closest clinical analog to our study in a large scRNA-seq study that profiled cells from LUAD patients and samples, spanning different treatment regimes: treatment naïve (TN), residual disease (RD) & progressive disease (PD). The RD cells, the cells that survive right after treatment, are the closest analog to our PC-9 persisters. The PD cells could either represent a more aggressive persister state (potentially cycling), or a fully resistant state that acquired driver mutations. While we could not identify significant differences between TN & RD states, we found some persistence potential associated genes and modules to be differentially expressed between the RD & PD states. OXPHOS was strongly overexpressed in PD in comparison to RD (p=1.23e-120) (Fig. 4.28). Furthermore, the ribosomal signature, consisting of the ribosomal genes implicated in our analysis, was underexpressed in PD in comparison to RD (p=3.4e-59) (Fig. 4.28). Lastly, PSMA7 is overexpressed in PD relative to RD (p=4.96e-17) (Fig. 4.28). These results suggest that the genes associated with persistence potential could also play a role in progressing lung cancers from a less aggressive persister state (RD) to a, potentially, more aggressive persister state (PD) in patients.

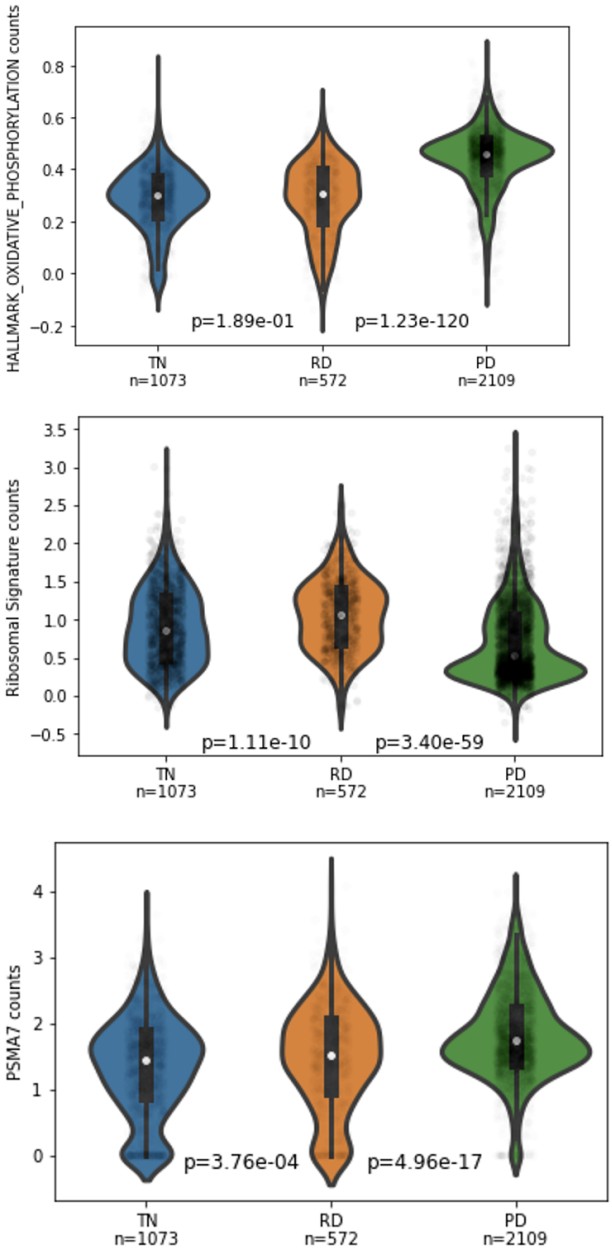

Figure 4.28: Comparing the expression of genes found to be positively associated with persistence across different timepoints in LUAD patients: Treatment Naïve (TN), Residual Disease(RD) and Progressive Disease (PD). Oxidative Phosphorylation expression is greater in PD relative to RD. Ribosomal signature expression is lower in PD relative to RD. Lastly, PSMA7 s greater in PD relative to RD.

4.3.7 Transcriptional Drivers of Cycling Persistence Potential

While persister cells are thought to be slow-dividing groups of cells that barely thrive in the presence of treatment, there can exist a smaller fraction of cells in this population that persist long enough to start cycling and ultimately repopulate a tumor. Other works have highlighted hallmarks of cycling persisters but no study to date has investigated the mechanisms that influence cycling persistence potential [19]. We next wanted to investigate whether there are subclonal differences in cycling persistence potential within a clone and if so, identify drivers that dictate an untreated cell or persister cell to be more likely to persist & cycle. For each of the persister cells, we assigned a mapping score for each cycling persister based on the shared characters the two cells shared. We devised a custom hamming distance that rewarded shared characters and penalized characters that differed (see Methods) (Fig. 4.29). Lastly, common indel states were downweighted, while more rare states were upweighted. We then, for each persister cell, averaged the scores across all of the cycling persisters and used these scores to perform a "clade-level" analysis similar to what we described above to identify clades with differential persistence potential. In this particular instance, since we're handling continuous values and not counts, we performed a bottom-to-top comparison of cycling scores between clades and their closest "sister clades." We did a Wilcoxon rank sum test between the pairs of clades to identify clades that were differentially enriched for high versus low mapping of cycling persisters.

After performing differential gene expression analysis across cycling mapping potential, we identified "hits" that overlapped with genes that were previously identified to be upregulated in cycling persisters [19] (Fig. 4.30). For instance, we also found *AKR1B1*, *AKR1C2*, *GSTM3* and *GSTM4* to be among the top hits of genes associated with cycling potential in persisters. *AKR1B10*, *AKR1C2*, and *GSTM3* were previously identified to be correlated with cycling persistence [19]. The GSTM genes, in particular, are part of the glutathione metabolism, which have been reported to be upregulated in cycling persisters. Furthermore,

we observe that multiple heat shock proteins (e.g., *HSPH1*, *HSP90AA1*, *HSPD1*, *HSP90AB1*) are associated with persisters that have higher cycling potential. Interestingly, this result suggests that persisters that manage stress better by way of the heat shock response, are more likely to cycle.

We performed the same analysis with the untreated cells in the trees by assigning cycling mapping scores to the untreated cells in the tree and identifying groups of untreated cells that had significantly high or low mappability to cycling persisters. Upon performing differential expression of the untreated cells within the identified events, we uncovered intriguing genes that associate with cycling potential. *EGFR* appears to be positively associated with cycling potential among untreated cells. *S100A9*, which has been associated with metastasis in LUAD and tumor progression in other malignancies, was the most significant gene to be positively associated with cycling potential. Lastly, *AKR1B1*, a member of the glutathione metabolism pathway, is also upregulated among untreated cells that have higher cycling potential.

Lastly, *ALCAM* was a top hit for being negatively associated with cycling potential in both high cycling potential untreated cells and high cycling potential persister cells. Together, these findings further support the recurrent theme in this study that features that exist in a more progressed, malignant state appear to pre-exist in their previous states, in untreated cells for persister cells and non-cycling persisters for cycling persister cells.

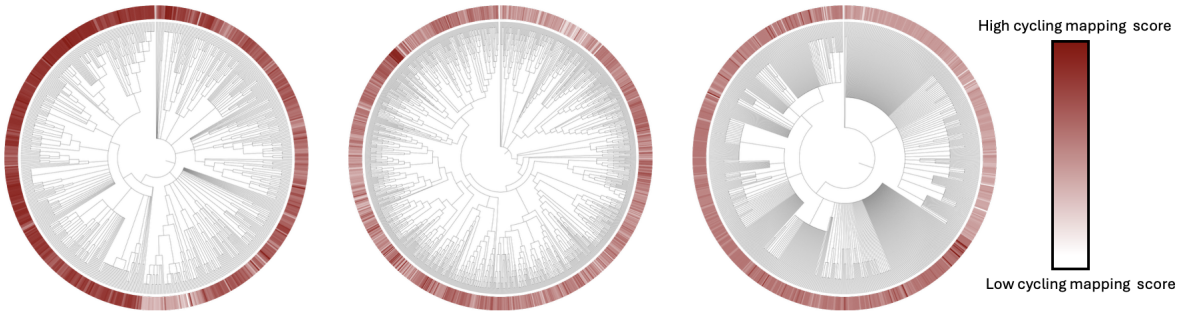

Figure 4.29: Visualization of Cycling Scores mapped onto respective short-term persisters.

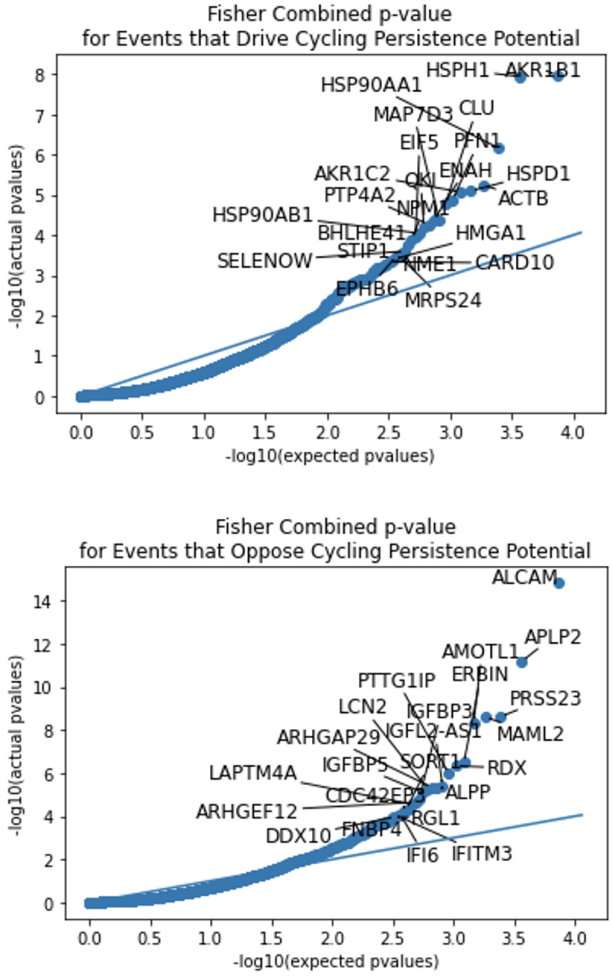

Figure 4.30: Differentially expressed genes across cycling persistence potential. Several heat Shock genes are positively associated with cycling potential.

4.4 Discussion

Previous studies of cancer persistence typically highlight differences between cancer clones in response to treatment. While these studies have yielded key insights into different pathways to resistance, they don't address the features that preexist treatment that drive differential persistence. Additionally, they also do not study the phenotypic changes that can occur within clones. We generated high resolution single-cell phylogenies of lung cancer clones to study the subclonal diversity that informs persistence potential.

Applying tailored approaches for studying different modes of evolution of persistence

potential, we identified key genes and pathways that were associated with persistence potential across clones. We show that mitochondrial oxidative phosphorylation is strongly associated with persistence potential. Furthermore, upon inhibiting OXPHOS in combination with EGFR inhibition with osimertinib, we demonstrate synergistic killing of persisters. This approach could be disproportionately killing the cells that are most likely to persist, resulting in a significant reduction in persisters. This combination therapeutic strategy could be a promising approach to reduce recurrence of persisters and potentially increase survival in EGFR positive LUAD patients. We also demonstrate a consistent negative correlation between persistence potential and ribosomal activity, suggesting that cells that are primed to persist downregulate protein production. When inhibiting ribosomal activity, we show a synergistic increase in persisters. While this result is not of much therapeutic significance, it complements one of our main hypotheses, that high persistent potential cells rely on suppression of translation. Dependence on OXPHOS for energy and global suppression of translation are hallmark features of drug tolerant persisters, potentially as means for tolerating physiological stress. The implication of both of these modules in untreated cancer cells suggests that the same stress tolerant states that define DTPs can preexist to varying degrees in untreated cells and serve as pathways towards full persistence or resistance.

This work serves the basis for a number of intriguing follow-up studies regarding cancer persistence and resistance. It would be of interest to investigate the subclonal dynamics driving persistence potential in the context of different treatment strategies such as chemotherapeutics, radiotherapy, or other targeted therapies. Are persistence potential mechanisms shared across different strategies and malignancies or are they dependent on malignancies and treatment types? Future studies applying this technology to these different context could address these questions. Similarly to this paper, we could extend this to mammalian models (ie: mice xenograft) to address the concern that the results we observe are cell culture artifacts. Future studies could also be modified by including more clones, as we accepted the trade-off of less clones for more cells per clone.

This study serves as another example of the power of using single-cell CRISPR lineage tracing systems to study subclonal diversity in cancer. These types of studies enable the investigation of how individual clones evolve in the context of persistence and resistance. We contributed a persister score formulation that proved useful in identifying genes differentially expressed across the axis of persistence potential. We also contribute an approach for identifying significant events in phylogenies containing cells from different conditions or timepoints. These strategies can be extended to any evolutionary process where one is interested in how the potential of a phenotype of interest evolves within a clone. These phylogenetic strategies studying phenotypic potential can enable the identification of novel therapeutic targets that can be missed by less granular lineage tracing assays that study the effects post intervention.

4.5 Methods

4.5.1 Tissue Culture

PC-9, HCC827, and NCI-H1975 EGFR-mutated non-small-cell lung cancer cells used in the current study were maintained in RPMI-1640 medium, supplemented with 10% fetal bovine serum, 2 mM glutamine, 100 µg/ml penicillin and streptomycin, in a humid atmosphere containing 5% CO_2 at 37°C.

4.5.2 Lentivirus Preparation

Lentivirus used for transduction during cell line engineering were produced in HEK293T cells, using TransIT-LTI as per manufacturer's instructions, as previously described [109]. PCT48 lentiviral preparations containing the target site were concentrated tenfold using Lenti-X Concentrator (Takara Bio).

4.5.3 Cell Line Engineering

To generate PC-9 lineage tracing cells (PC9-LT), PC-9 cells were transduced by serial lentiviral infections, as previously described [18].

Briefly, cells were first transduced with a construct containing constitutively expressed Cas9-mCherry, followed by flow-sorting to obtain mCherry-positive cells. Then, cells were transduced using a concentrated lentivirus containing a GFP-labeled target site to maximize genomic target site integrations per cell. Following transduction with the target site lentivirus, cells were sorted by FACS to obtain the top 25% of cells with the highest GFP intensity. These cells were transduced again using a concentrated lentivirus containing target sites and sorted again. This process was repeated three times to obtain a sufficient number of target sites per cell.

Lastly, cells with Cas9 and target sites were transduced with the triple-sgRNA lentivirus (initiating lineage recording) and single-cell plated in 96-well plates to generate single-cell-derived lineage-recording clones.

For cytotoxicity experiments, PC-9 cells and HCC827 cells were transduced with a lentiviral vector expressing nuclear GFP (pHR_dSV40-H2B-GFP).

4.5.4 Lineage Tracing Experiments

Following activation of lineage recording and single-cell plating of PC9-LT cells, clones were expanded for 2–3 weeks. During clonal expansion, cells were transferred to appropriate tissue culture plates to accommodate clone size. When clones reached 5–10 million cells, aliquots were frozen, and cells were plated in 10 cm plates for both the untreated and persister arms of the experiment.

After an overnight incubation to allow cell adhesion to the plate, the persister arm cells were treated with 300 nM Osimertinib for 7 days. The cells in the untreated arm were passaged 3 days after plating. After 7 days, untreated and persister cells were collected and used

to generate 3' scRNA-seq libraries using a 10X Genomics kit, following the manufacturer's instructions.

During clonal expansion and the drug treatment stage, single-cell-derived clones were cultured separately. Cells from four single-cell-derived clones were used. Cells were loaded into the 10X Genomics controller such that each clone contained approximately 2,000 untreated cells and ∼1,000 persister cells.

Amplicon libraries, containing single-cell lineage barcode data, were amplified from the single-cell cDNA, and both scRNA-seq and amplicon libraries were sequenced using the Illumina NovaSeq platform, as previously described [18].

4.5.5 Drug Combination Experiments

PC9-GFP, HCC827-GFP, and NCI-H1975 cells were plated in triplicates in 96-well plates at 2,000 cells per well. After an overnight incubation, cells were treated with 300 nM Osimertinib, with or without other drugs as described above, for 72 hours. The plates were imaged and counted using the IncuCyte S3 system.

4.5.6 Preprocessing of Raw Target Site Library using Cassiopeia

The target site library was largely processed using the default parameters of the Cassiopeia preprocessing pipeline as defined (cite) with minor modifications. Briefly, an error-corrected consensus sequence is established for each cellBC-UMI pair. The consensus sequences are then aligned to the reference target site sequence and indels and intBCs are then called from this alignment. Utilizing the genomic DNA we sequenced for each clone, the top 20 high confidence, clone-specific intBCs were identified based on read support and used to assign cells to clones based on relative enrichment of these high confidence intBCs. Specifically, cells that had >80% of its reads matched to the set of high confidence intBCs for a given clone were assigned to that clone. Once the high confidence mapped cells are retained, there are molecule filter steps (filter_molecule_table step) that remove low quality or lowly supported

UMIs and cellBCs. The default parameters were used for this step. Lastly, despite having intBCs from gDNA we still ran, for each clone-specific allele table, the call_lineage_groups steps to infer the full set of intBCs that define a given clone. We ran this step with its default parameters with the exceptions of the min_intbc_thresh and min_avg_reads_per_umi, which we set to 0.3 and 1 respectively. This resulted in 45-55 intBCs per clone.

Upon generating final allele tables for both the untreated and persister cell libraries, these tables were combined and only the intersection of intBCs between the untreated and persister cells were retained. This final combined allele table is then used for downstream phylogenetic reconstruction.

4.5.7 Phylogenetic Reconstruction using Cassiopeia Hybrid

The Cassiopeia-Hybrid algorithm was implemented for each clone to reconstruct clonal lineage trees, utilizing the filtered clonal-specific allele tables as input. This module was implemented as described with subtle modifications to correct for potential clustering of persister cells due to technical bias [18]. In short, Cassiopeia-Hybrid takes as input a cellxcharacter character matrix, representing the indel states recorded within the target site's cut sites. In the Greedy mode of Cassiopeia-Hybrid, in a top-down fashion, it iteratively splits cells into binary bins based on presence or absence of highly prevalent characters. Once the switch criteria has been met, the Hybrid algorithm switches to its more exhaustive Steiner-Tree-based Cassiopeia-ILP mode to order the lower depths of clades and cells. We set the switch parameter to be LCA of 30, meaning that at the LCA distance of 30, the subproblem will be solved with the Steiner-based approach. The character matrix is a binary matrix with '-' for missing data. In order to correct for technical bias, any character that was unique to persister cells due to continued recording during the treatment phase was set to missing ('-'). This was done to correct for potential artefactual clustering of the cells due to indels that were acquired in only one group of cells.

4.5.8 Moran's I Calculation: Assessing Phylogenetic Signal of Persistence in Clonal Trees

We apply the standard Moran's I formulation for assessing nonrandom phylogenetic signal of features such as binary persister state or expression:

$$I = \mathbf{z}^T \mathbf{W} \mathbf{z} \tag{4.1}$$

where $\mathbf{z}$ is the standardized feature vector of length n, with n representing the number of cells. $\mathbf{W}$ is the $n \times n$, globally normalized proximity matrix. We derived the proximity matrix from the pairwise node distance matrix by taking the inverse of each element in the distance matrix. We then normalized the proximity matrix such that the sum of the elements in $\mathbf{W}$ equals 1.

4.5.9 Persistence Potential Score Calculation

The persistence potential score for each untreated cell was defined as the sum of the inverse node distance between the untreated cell of interest to every persister cell in the tree. We normalized this score by the sum of the inverse node distance between untreated cell of interest and all cells in the tree to account different neighborhood size, potentially driven by clonal expansions:

$$\text{Persistence\_Potential\_Score}(\text{cell}_i) = \frac{\sum_{j=1}^{n_{\text{PER}}} \frac{1}{\text{node\_distance}(\text{cell}_i, \text{PER\_cell}_j)}}{\sum_{k=1}^{n_{\text{cells}}} \frac{1}{\text{node\_distance}(\text{cell}_i, \text{cell}_k)}} \tag{4.2}$$

where $nPER$ is the number of persister cells in the tree and $ncells$ is the number of total cells in the tree.

4.5.10 Differential Expression across Single-Cell Persistence Potential Score by Poisson regression

Genes that are differentially expressed along the continuum of persistence potential were identified using a Poisson regression model (implemented with the Python package `statsmodels`). The models for identifying genes associated with persistence potential were implemented as follows:

$$\begin{aligned} H_0: &\quad Y \sim \text{constant} \\ H_a: &\quad Y \sim \text{persistence\_potential\_score} + \text{constant} \end{aligned}$$

where Y represents the UMI-normalized counts (counts per 10,000 UMIs). A likelihood ratio test was implemented to assess whether the data better fit the full model (H_a), with persistence score as an explanatory variable, compared to the null model (H_0) where the constant alone explains the expression. FDR-adjusted p-values were calculated using the Benjamini–Hochberg false discovery rate procedure, and genes were deemed significant if they had an adjusted p-value less than 0.01.

The regression coefficients capture the direction of the effect of persistence potential on gene expression. Positive coefficients suggest a positive association between persistence score and expression, whereas negative coefficients imply a negative association. Log_2 fold changes ($\log_2$FC) between the top 25% and bottom 25% of cells ranked by persistence scores were also calculated to capture the magnitude and directionality of the effect for each gene.

This differential expression approach was implemented independently for each clone. For each clone, only cells with at least 2,000 detected genes were retained, and only genes expressed in at least 100 cells were considered.

4.5.11 Identification of Clade Events Enriched or Depleted for Persisters

A bottom to top clade identification approach was performed to identify clades that were significantly enriched or depleted for persisters. At each depth in the tree, a fisher exact test is performed between the categorical variables of clade identity ("reference clades" vs. "sister clades" and cell identity (untreated cells vs. persister cells). In an attempt to only compare clades that are phylogenetically near each other, sister clades clades were defined as clades that share a most recent common ancestor (MRCA) with a given clade that was just one depth above the depth of consideration. Events that were identified at each depth were pruned from future comparisons for higher depths. For instances where there are perfect binary splits between the "reference clade" and sister clade, the smaller of the 2 events were retained as the "event." At each depth, FDR corrections were performed, where the number of hypotheses were the number of possible comparisons at a given depth of interest. Events that had at least 8 untreated cells in both reference and sister clades and adjusted p-values less than 0.01 were retained for downstream differential expression analysis.

4.5.12 Differential Expression across Clade Events

For each of the retained clade events, a one-sided Wilcoxon rank sum test was used to perform differential expression between the untreated cells in the persister enriched clades and the untreated cells in the persister depleted clades. Two one-sided tests were performed for each of the events, to identify the set of genes that positively associate / potentially drive persistence potential and genes that negatively associate/potentially oppose persistence potential. In order to integrate the signal across the events, we calculated a combined fisher p-value for both one sided tests across all events.

For each event, cells that have greater than 20% of its counts attributed to the mitochondrial genome are filtered out. Cells that contained at least 2000 detected genes were retained

and genes that were found in at least 100 cells were retained. For each event's filtered set of cells and genes, highly variable genes were identified using Scanpy's highly_variable_genes function. Lastly, only the union of the highly variable genes for each event were considered for differential expression analysis.

This analysis similarly extended for identification of differentially expressed pathways across clade events. AUCell, (featured in single-cell best practices), is utilized to calculate enrichment scores for pathways of interest in each cell by calculating whether the top ranked genes of a given cell are enriched for a critical subset of genes of the pathway of interest [106]. These scores are used in the Wilcoxon rank sum tests to test differential expression of pathways.

4.5.13 Identification of De-Novo Lineage Dependent Modules

Hotspot was used to identify de-novo gene modules that have phylogenetic signal in our clones [107]. In short, Hotspot is a graph based approach that works in a two-step fashion to identify lineage Dependent Modules. First, it identifies genes that have nonrandom distribution of expression within the phylogeny, using a slightly modified form of Moran's I that only considers KNN neighbors in its calculation. Secondly, from this filtered set of genes, it generates an autocorrelation matrix that captures how well pairs of genes colocalize, also using a modified form of Moran's I that incorporates the gene expression of both genes in a KNN cell pair of interest. We input our clonal phylogeny structures for Hotspot to generate its KNN graph. When running, we only considered the expression of the untreated cells in each clone, as we are primarily interested in persistence potential in our study. Cells that contained at least 2000 detected genes were retained and genes that were found in at least 100 cells were retained. Lastly, we used the normal model for the background gene expression model, and set K to 30 for the KNN graph formulation step.

4.5.14 Trajectory Analysis and Identification of Lineage Dependent, Differential Response Genes

A CxG "response matrix" was generated to capture the weighted difference in expression between untreated cells and all persisters in the tree, where C is the number of untreated cells and G is the number of genes being considered. The elements of the matrix are calculated as such:

$$\text{Weighted\_response}_{\text{untreated},g} =$$

$$\frac{\sum_{\text{persister}} \left(\text{Expression}_{\text{untreated},g} - \text{Expression}_{\text{persister},g}\right) \times \left(\frac{1}{\text{Dist(untreated,persister)}}\right)}{\sum_{\text{persister}} \left(\frac{1}{\text{Dist(untreated,persister)}}\right)} \tag{4.3}$$

Standard PCA is then ran on this "response matrix" and the gene loadings that contribute the most in magnitude to PC1 and PC2 are identified as the most differentially responsive genes. The response values for the most differentially responsive genes are compared using Pearson correlation.

4.5.15 Mas-Isoseq Preprocessing & Differential Isoform Expression Analysis

Processing of the Mas-Iso seq data follows a similar protocol to that of short-read transcriptomic data, such as 10x. We used best practices tools for processing long read data: Skera & Isoseq (https://isoseq.how/getting-started.html). As input to the processing steps, we start with an unmapped bam file (uBAM). The skera split command splits the concatenated cDNA reads at the adaptors to generate segmented reads. The isoseq lima command then removes the 10x cDNA primers from the reads. We then ran the isoseq tag, isoseq refine, isoseq correct, isoseq groupdedup commands to extract UMI and cell barcodes, remove polyA tail and artificial concatemers, correct cell barcodes and tag reads that are real cells, and deduplicate reads, respectively. We then run minimap2 to map the reads to the reference genome (hg38).

We then use Isoquant to assign reads to annotated isoforms based on their exon and intron structure and quantifies the isoforms. To remove artefactual or non-informative isoforms, we run squanti3 ss, which uses a rules based method for filtering out isoform candidates that are a result of intra-priming events (from internal polyA events), are mono-exonic, or those with non-canonical splice junctions. With the corrected fasta containing a clean set of isoforms,

we realign the data to this custom transcriptome using minimap2. Lastly, we quantify the reads using oarfish to generate a cellxisoform matrix.

Next, we sought to identify genes that have imbalance isoform expression in two of our largest identified clade events. Genes that had isoforms that accounted for greater than 80% of that gene's expression were removed from downstream differential expression analysis. For each gene, we employed the Jensen Shannon metric between the pseudobulked counts of isoforms in the untreated cells of the "reference clade" and the untreated cells of the "sister clade." The genes with the highest Jensen Shannon scores, had the largest divergence in their distribution of relative isoform counts.

4.5.16 Mapping of Cycling Cells to Short-term Persisters and Untreated Cells

Utilizing the character matrix for both the cycling persisters and short-term persisters, we assigned a mapping score for each short-term persister, cycling persister pair. The mapping score for any pair of cells is determined by a custom weighted hamming distance metric defined by this algorithmic scoring scheme, where differences in character states are penalized and similarity is rewarded.

For each short-term persister, we average the score across all the mapped cycling persisters for downstream identification of clades that are enriched for cycling mapping. Below is the mapping scheme we used:

```
def custom_weighted_hamming_combined_similarity_distance(
    s1: List[int],
    s2: List[int],
    missing_state_indicator=-1,
    weights: Optional[Dict[int, Dict[int, float]]] = None,
) -> float:
    similarity = 0
```

```
    num_present = 0
    for i in range(len(s1)):
        if s1[i] == missing_state_indicator or s2[i] ==
           missing_state_indicator:
            continue

        num_present += 1

        if s1[i] != s2[i]:
            if s1[i] == 0 or s2[i] == 0:
                if weights:
                    if s1[i] != 0:
                        similarity -= weights[i][s1[i]]
                    else:
                        similarity -= weights[i][s2[i]]
                else:
                    similarity -= 1
            else:
                if weights:
                    similarity -= weights[i][s1[i]] +
                       weights[i][s2[i]]
                else:
                    similarity -= 2
        else:
            if s1[i] != 0:
                if weights:
                    similarity += 2 * (weights[i][s1[i]])
                else:
                    similarity += 2
```

```
35
36      if num_present == 0:
37          return 0
```

Listing 4.1: Custom Weighted Hamming Combined Similarity Distance

4.5.17 Identification of Cycling Clade Events and Differential Expression of Events

The clade identification scheme is almost identical to that of the clade-level analysis for studying persistence potential, mentioned above. Since we are handling continuous values per cell here and not integer counts, for each "reference clade" / "sister clade" comparison we perform a Wilcoxon Rank Sum test to identify differential enrichment of cycling mappability scores between clades. If significant, we prune the event in upstream comparison between clades, and iteratively repeat comparisons until reaching the root of the tree to identify remaining events.

4.5.18 Kaplan Meier Survival Analysis of LUAD Cohort

We utilized RNA-seq data from the CPTAC LUAD cohort for performing a survival analysis on the top gene and pathway "hits." We only considered patients that were annotated to contain driver somatic mutations in the EGFR gene. We then utilized the progression free survival (PFS) time (days_to_last_contact_or_death_or_recurrence) and overall survival (OS) time since diagnosis to perform Kaplan-Meier analysis. We used the Survfit package in R to generate the Kaplan-Meier curves for genes and pathways of interest, where patients were stratified by "high" and "low" based on positioning of their gene expression relative to the average expression of the gene or pathway of interest. Pathway signature scores were generated using Scanpy's score_genes module, which takes the average expression of the genes that define a pathway of interest and normalizes this expression by a control set of

genes of similar expression levels.

4.5.19 Code Availability

The reproducibility code for this work, while not public yet, will be made public upon publication and can be found here: *https://github.com/jidezike3/Single_Cell_Phylogenetic_Persistence_Analysis*

Chapter 5

Conclusions and Future Directions

Cells continuously transition between states under the influence of intrinsic programs and external pressures. Reconstructing their lineage relationships offers a powerful framework for studying dynamic biological processes such as cancer evolution and hematopoietic development. In this thesis, we present a suite of computational and analytical methods designed to infer clonal lineage trajectories and associated gene expression programs from single-cell data.

In Chapter 1, we leveraged a comprehensive single-cell atlas of human hematopoiesis to trace lineage fate biases across lifespan. Through methods based on markovian random walks along graphical embeddings of single-cells and optimal transport, we inferred lineage fate probabilities and differentiation trajectories and identified lineage-specific gene programs that were both constant and variable along lifetime. We also uncovered novel, age-associated hematopoietic stem cell gene expression programs, revealing a novel fetal HSC marker that marks cells with greater proliferative capacity. These approaches demonstrate the ability to conduct lineage tracing of cells without the presence of genetic barcodes when you have a well-defined hierarchical system with known phylogenetic endpoints, such as hematopoiesis. However, these systems have recently started to integrate genetic barcodes in the form of mitochondrial mutations, resulting in confirmation of expected hematopoietic lineage trajectories and clarifying branch-points that are still uncertain, as previously described [110].

Chapter 2 introduced a denoising procedure, which includes an anomaly detection framework, for distinguishing true somatic mutations from technical artifacts in full-length single-cell RNA-sequencing data, revealing biologically meaningful mutational signatures and clonal structures across malignancies. These approaches are well suited for contexts where the majority of mutations are expected to be noise, and the sequence context of the respective artefactual mutations follow a distinct pattern from real somatic mutations. There was also an introduction of different de-novo approaches for identifying phylogenetic signal in single-cell mutation data, such as Jaccard-based hierarchical clustering and betaVAE latent-space clustering of the single-cell mutation profiles. These analyses are useful exploratory approaches for identifying whether there are groups of cells that have shared mutations and are capable of finding groups of cells that agree with known clones or clades. Future work can build on the VAE approach by applying contrastive learning methods to the latent space, which pushes dissimilar points apart and brings similar points closer. This can be an intriguing machine learning approach to infer a phylogenetic tree from low-dimensional space of single-cell mutation data. Upon finding groups of cells with similar mutation profiles, one could also apply the "pigeonhole principle," using the frequencies of the mutations found across the cells, to order cells and mutations phylogenetically.

In chapter 3, we applied CRISPR-based lineage tracing to construct high-resolution single-cell phylogenies in a cancer persistence model. By quantifying phylogenetic proximity to persister cells, we developed a single-cell persistence potential score and identified clade events that have significant differences in persistence levels. Combination therapy of genes and gene modules (oxidative phosphorylation and ribosomal activity) that are differentially expressed across persistence potential, both at the single-cell and clade levels, has a synergistic effect on the survival of persister cells. This pipeline contributes a set of analytical approaches that phylogenetically relates cells from different time-points or conditions and infers potential (and associated gene expression changes) in transitioning between conditions. This type of analyses can be extended to any other developmental process or malignancy, across numerous

different conditions. It would be intriguing to conduct this type of study with standard-of-care chemotherapeutics to see whether the gene pathways associated with potential to persist differ from those found in the targeted therapy setting (the setting for our work, with EGFR inhibition). Overall, these types of studies are very useful for generating hypothesis for drivers of persistence and resistance potential to various types of malignant treatments or, more generally, for finding drivers of potential in transitioning from one cell state to another.

Together, these studies demonstrate the value of integrating single-cell barcoding—both natural and engineered—with computational approaches for lineage reconstruction. The methods and insights presented here provide a foundation for investigating developmental processes, clonal evolution, and treatment response potential at single-cell resolution.

References

[1] I. Martincorena, J. C. Fowler, A. Wabik, A. R. J. Lawson, F. Abascal, M. W. J. Hall, et al. "Somatic mutant clones colonize the human esophagus with age". In: *Science* 362 (Nov. 2018), pp. 911–917. DOI: *10.1126/science.aau3879*.

[2] S. Jaiswal, P. Fontanillas, J. Flannick, A. Manning, P. V. Grauman, B. G. Mar, et al. "Age-related clonal hematopoiesis associated with adverse outcomes". In: *New England Journal of Medicine* 371 (Dec. 2014), pp. 2488–2498. DOI: *10.1056/NEJMoa1408617*.

[3] K. Suda, H. Nakaoka, K. Yoshihara, T. Ishiguro, R. Tamura, Y. Mori, et al. "Clonal expansion and diversification of cancer-associated mutations in endometriosis and normal endometrium". In: *Cell Reports* 24.7 (Aug. 2018), pp. 1777–1789. DOI: *10.1016/j.celrep.2018.07.031*.

[4] I. Martincorena and P. J. Campbell. "Somatic mutation and clonal expansions in human tissues". In: *Nature Genetics* 47.10 (Oct. 2015), pp. 1053–1060. DOI: *10.1038/ng.3394*.

[5] E. Mitchell, S. Chapman, N. Williams, et al. "Clonal dynamics of haematopoiesis across the human lifespan". In: *Nature* 606 (2022), pp. 343–350. DOI: *10.1038/s41586-022-04780-3*.

[6] I. Martincorena, J. C. Fowler, A. W. Gerstung, L. J. Y., K. L. M., et al. "High burden and pervasive positive selection of somatic mutations in normal human skin". In: *Science* 348.6237 (2015), pp. 880–886. DOI: *10.1126/science.aaa6806*.

[7] J. Cairns. "Mutation selection and the natural history of cancer". In: *Nature* 255 (1975), pp. 197–200. DOI: *10.1038/255197a0*.

[8] P. C. Nowell. "The clonal evolution of tumor cell populations". In: *Science* 194.4260 (1976), pp. 23–28. DOI: *10.1126/science.956949*.

[9] V. G. Sankaran, J. S. Weissman, and L. I. Zon. "Cellular barcoding to decipher clonal dynamics in disease". In: *Science* 355.6325 (2017), pp. 1133–1137. DOI: *10.1126/science.aaf6404*.

[10] L. Kester and A. van Oudenaarden. "Single-cell transcriptomics meets lineage tracing". In: *Cell Stem Cell* 23.2 (2018), pp. 166–179. DOI: *10.1016/j.stem.2018.06.015*.

[11] D. E. Wagner and A. M. Klein. "Lineage tracing meets single-cell omics: opportunities and challenges". In: *Nature Reviews Genetics* 21 (2020), pp. 410–427. DOI: *10.1038/s41576-020-0215-1*.

[12] M. S. Lawrence et al. "Mutational heterogeneity in cancer and the search for new cancer genes". In: *Nature* 499.7457 (2013), pp. 214–218. DOI: *10.1038/nature12213*.

[13] H. Li, P. Côté, M. Kuoch, J. Ezike, K. Frenis, et al. "The dynamics of hematopoiesis over the human lifespan". In: *Nature* 626 (2024), pp. 743–751. DOI: *10.1038/s41586-024-07194-w*.

[14] C. Neftel, J. Laffy, M. G. Filbin, T. Hara, M. E. Shore, et al. "An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma". In: *Cell* 178.4 (2019), 835–849.e21. DOI: *10.1016/j.cell.2019.06.024*.

[15] F. Muyas, C. M. Sauer, J. E. Valle-Inclán, R. Li, R. Rahbari, et al. "De novo detection of somatic mutations in high-throughput single-cell profiling datasets". In: *Nature Biotechnology* 55.10 (2024), pp. 1623–1633. DOI: *10.1038/s41588-023-01469-0*.

[16] J. Dou, Y. Tan, K. H. Kock, J. Wang, X. Cheng, et al. "Single-nucleotide variant calling in single-cell sequencing data with Monopogen". In: *Nature Biotechnology* (2024). DOI: *10.1038/s41587-024-02190-8*.

[17] M. A. Lodato, M. B. Woodworth, S. Lee, G. D. Evrony, B. K. Mehta, et al. "Somatic mutation in single human neurons tracks developmental and transcriptional history". In: *Science* 350.6256 (2015), pp. 94–98. DOI: *10.1126/science.aab1785*.

[18] M. G. Jones, A. Khodaverdian, J. J. Quinn, M. M. Chan, J. A. Hussmann, et al. "Inference of single-cell phylogenies from lineage tracing data using Cassiopeia". In: *Genome Biology* 21.1 (2020), p. 92. DOI: *10.1186/s13059-020-01991-0*.

[19] Y. Oren, M. Tsabar, M. S. Cuoco, L. Amir-Zilberstein, H. F. Cabanos, et al. "Cycling cancer persister cells arise from lineages with distinct programs". In: *Nature* 596.7873 (2021), pp. 576–582. DOI: *10.1038/s41586-021-03836-0*.

[20] R. Lu, N. F. Neff, S. R. Quake, and I. L. Weissman. "Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding". In: *Nature Biotechnology* 29.10 (2011), pp. 928–933. DOI: *10.1038/nbt.1977*.

[21] S. H. Naik, L. Perié, E. Swart, C. Gerlach, N. van Rooij, R. J. de Boer, and T. N. Schumacher. "Diverse and heritable lineage imprinting of early haematopoietic progenitors". In: *Nature* 496.7444 (2013), pp. 229–232. DOI: *10.1038/nature12013*.

[22] C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, and A. M. Klein. "Lineage tracing on transcriptional landscapes links state to fate during differentiation". In: *Science* 367.6479 (2020), eaaw3381. DOI: *10.1126/science.aaw3381*.

[23] B. Raj, D. E. Wagner, A. McKenna, S. Pandey, A. M. Klein, J. Shendure, J. A. Gagnon, and A. F. Schier. "Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain". In: *Nature Biotechnology* 36.5 (2018), pp. 442–450. DOI: *10.1038/nbt.4103*.

[24] B. Raj et al. "Emergence of neuronal diversity during vertebrate brain development". In: *Neuron* 108.6 (2020), 1058–1074.e6. DOI: *10.1016/j.neuron.2020.09.010*.

[25] J. J. Quinn, M. G. Jones, R. A. Okimoto, S. Nanjo, M. M. Chan, N. Yosef, T. G. Bivona, and J. S. Weissman. "Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts". In: *Science* 371.6532 (2021), eabc1944. DOI: *10.1126/science.abc1944*.

[26] N. Saitou and M. Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees". In: *Molecular Biology and Evolution* 4.4 (1987), pp. 406–425. DOI: *10.1093/oxfordjournals.molbev.a040454*.

[27] J. F. Weng, I. Mareels, and D. A. Thomas. "Probability Steiner trees and maximum parsimony in phylogenetic analysis". In: *Journal of Mathematical Biology* 64.7 (2012), pp. 1225–1251. DOI: *10.1007/s00285-011-0442-4*.

[28] B. Kolaczkowski and J. W. Thornton. "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous". In: *Nature* 431.7011 (2004), pp. 980–984. DOI: *10.1038/nature02917*.

[29] H. Zafar, A. Tzen, N. Navin, K. Chen, and L. Nakhleh. "SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models". In: *Genome Biology* 18.1 (2017). DOI: *10.1186/s13059-017-1242-4*. URL: *http://creativecommons.org/licenses/by/4.0/*.

[30] K. Jahn, J. Kuipers, and N. Beerenwinkel. "Tree inference for single-cell data". In: *Genome Biology* 17 (2016), p. 86. DOI: *10.1186/s13059-016-0936-x*.

[31] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis. "Single-cell RNA-seq denoising using a deep count autoencoder". In: *Nature Communications* 10 (2019), p. 390. DOI: *10.1038/s41467-018-07931-2*.

[32] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. "Deep generative modeling for single-cell transcriptomics". In: *Nature Methods* 15 (2018), pp. 1053–1058. DOI: *10.1038/s41592-018-0229-2*.

[33] Y. Li et al. "Single-cell analysis of neonatal HSC ontogeny reveals gradual and uncoordinated transcriptional reprogramming that begins before birth". In: *Cell Stem Cell* 27 (2020), 732–747.e7.

[34] Z. Bian et al. "Deciphering human macrophage development at single-cell resolution". In: *Nature* 582 (2020), pp. 571–576.

[35] J. D. Buenrostro et al. "Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation". In: *Cell* 173 (2018), 1535–1548.e16.

[36] J. Cao et al. "A human cell atlas of fetal gene expression". In: *Science* 370 (2020), eaba7721.

[37] X. Han et al. "Construction of a human cell landscape at single-cell level". In: *Nature* 581 (2020), pp. 303–309.

[38] Y. Lu et al. "Single-cell analysis of human retina identifies evolutionarily conserved and species-specific mechanisms controlling development". In: *Developmental Cell* 53 (2020), 473–491.e9.

[39] J. E. Park, L. Jardine, B. Gottgens, S. A. Teichmann, and M. Haniffa. "Prenatal development of human immunity". In: *Science* 368 (2020), pp. 600–603.

[40] D. Pellin et al. "A comprehensive single cell transcriptional landscape of human hematopoietic progenitors". In: *Nature Communications* 10 (2019), p. 2395.

[41] A. M. Ranzoni et al. "Integrative single-cell RNA-seq and ATAC-seq analysis of human developmental hematopoiesis". In: *Cell Stem Cell* (2020). DOI: *10.1016/j.stem.2020.11.015*.

[42] Q. Weng et al. "Single-cell transcriptomics uncovers glial progenitor diversity and cell fate determinants during development and gliomagenesis". In: *Cell Stem Cell* 24 (2019), 707–723.e8.

[43] H. Wang et al. “Decoding human megakaryocyte development”. In: *Cell Stem Cell* 28 (2021), 535–549.e8.

[44] S. F. Stras et al. “Maturation of the human intestinal immune system occurs early in fetal development”. In: *Developmental Cell* 51 (2019), 357–373.e5.

[45] M. R. Copley et al. “The Lin28b–let-7–Hmga2 axis determines the higher self-renewal potential of fetal haematopoietic stem cells”. In: *Nature Cell Biology* 15 (2013), pp. 916–925.

[46] V. I. Rebel, C. L. Miller, C. J. Eaves, and P. M. Lansdorp. “The repopulation potential of fetal liver hematopoietic stem cells in mice exceeds that of their adult bone marrow counterparts”. In: *Blood* 87 (1996), pp. 3500–3507.

[47] C. L. Miller, V. I. Rebel, C. D. Helgason, P. M. Lansdorp, and C. J. Eaves. “Impaired steel factor responsiveness differentially affects the detection and long-term maintenance of fetal liver hematopoietic stem cells in vivo”. In: *Blood* 89 (1997), pp. 1214–1223.

[48] A. E. Beaudin et al. “A transient developmental hematopoietic stem cell gives rise to innate-like B and T cells”. In: *Cell Stem Cell* 19 (2016), pp. 768–783.

[49] H. Li et al. “Efficient CRISPR–Cas9 mediated gene disruption in primary erythroid progenitor cells”. In: *Haematologica* 101 (2016), e216–e219.

[50] F. Notta et al. “Distinct routes of lineage development reshape the human blood hierarchy across ontogeny”. In: *Science* 351 (2016), aab2116.

[51] R. G. Rowe, J. Mandelbaum, L. I. Zon, and G. Q. Daley. “Engineering hematopoietic stem cells: lessons from development”. In: *Cell Stem Cell* 18 (2016), pp. 707–720. DOI: *10.1016/j.stem.2016.05.018*.

[52] S. Triana et al. “Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states”. In: *Nature Immunology* 22 (2021), pp. 1577–1589.

[53] K. Vanuytsel et al. “Multi-modal profiling of human fetal liver hematopoietic stem cells reveals the molecular signature of engraftment”. In: *Nature Communications* 13 (2022), p. 1103.

[54] J. M. Granja et al. “Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia”. In: *Nature Biotechnology* 37 (2019), pp. 1458–1465.

[55] M. Buttner, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis. “A test metric for assessing single-cell RNA-seq batch correction”. In: *Nature Methods* 16 (2019), pp. 43–49.

[56] M. D. Luecken et al. “Benchmarking atlas-level data integration in single-cell genomics”. In: *Nature Methods* 19 (2022), pp. 41–50.

[57] A. J. MacKinney. “Effect of aging on the peripheral blood lymphocyte count”. In: *Journal of Gerontology* 33 (1978), pp. 213–216.

[58] V. Calvanese et al. “Mapping human haematopoietic stem cells from haemogenic endothelium to birth”. In: *Nature* 604 (2022), pp. 534–540.

[59] Y. Tan and P. Cahan. “SingleCellNet: a computational tool to classify single cell RNA-seq data across platforms and across species”. In: *Cell Systems* 9 (2019), 207–213.e2.

[60] A. J. MacKinney. “Effect of aging on the peripheral blood lymphocyte count”. In: *Journal of Gerontology* 33 (1978), pp. 213–216.

[61] W. W. Pang et al. “Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age”. In: *Proceedings of the National Academy of Sciences of the USA* 108 (2011), pp. 20012–20017.

[62] I. Beerman et al. “Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion”. In: *Proceedings of the National Academy of Sciences of the USA* 107 (2010), pp. 5465–5470.

[63] C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein. "Fundamental limits on dynamic inference from single-cell snapshots". In: *Proceedings of the National Academy of Sciences of the USA* 115 (2018), E2467–E2476.

[64] G. Schiebinger et al. "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming". In: *Cell* 176 (2019), p. 1517.

[65] S. Zhang, A. Afanassiev, L. Greenstreet, T. Matsumoto, and G. Schiebinger. "Optimal transport analysis reveals trajectories in steady-state systems". In: *PLoS Computational Biology* 17 (2021), e1009466.

[66] S. C. van den Brink et al. "Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations". In: *Nature Methods* 14 (2017), pp. 935–936.

[67] M. Slyper et al. "A single-cell and single-nucleus RNA-seq toolbox for fresh and frozen human tumors". In: *Nature Medicine* 26 (2020), pp. 792–802.

[68] C. DeBoever et al. "Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells". In: *Cell Stem Cell* 20 (2017), 533–546.e7.

[69] K. Vanuytsel et al. "Multi-modal profiling of human fetal liver hematopoietic stem cells reveals the molecular signature of engraftment". In: *Nature Communications* 13.1 (2022), p. 1103. DOI: *10.1038/s41467-022-28770-x*.

[70] S. Zheng et al. "Molecular transitions in early progenitors during human cord blood hematopoiesis". In: *Molecular Systems Biology* 14 (2018), e8041.

[71] L. Velten et al. "Human haematopoietic stem cell lineage commitment is a continuous process". In: *Nature Cell Biology* 19 (2017), pp. 271–281.

[72] P. van Galen et al. "Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity". In: *Cell* 176 (2019), 1265–1281.e24.

[73] F. Notta et al. “Distinct routes of lineage development reshape the human blood hierarchy across ontogeny”. In: *Science* 351 (2016), aab2116.

[74] E. R. Adelman et al. “Aging human hematopoietic stem cells manifest profound epigenetic reprogramming of enhancers that may predispose to leukemia”. In: *Cancer Discovery* 9 (2019), pp. 1080–1101.

[75] K. Young et al. “Decline in IGF1 in the bone marrow microenvironment initiates hematopoietic stem cell aging”. In: *Cell Stem Cell* 28 (2021), 1473–1482.e1477.

[76] E. Mitchell et al. “Clonal dynamics of haematopoiesis across the human lifespan”. In: *Nature* 606 (2022), pp. 343–350.

[77] T. H. H. Coorens et al. “Extensive phylogenies of human development inferred from somatic mutations”. In: *Nature* 597.7876 (2021), pp. 387–392. DOI: *10.1038/s41586-021-03883-y*.

[78] A. Maynard et al. “Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing”. In: *Cell* 182.5 (2020), 1232–1251.e22. DOI: *10.1016/j.cell.2020.07.017*.

[79] K. Yizhak et al. “RNA Sequence Analysis Reveals Macroscopic Somatic Clonal Expansion Across Normal Tissues”. In: *Science* 364.6444 (2019), eaaw0726. DOI: *10.1126/science.aaw0726*.

[80] F. Liu, Y. Zhang, L. Zhang, Z. Li, Q. Fang, R. Gao, and Z. Zhang. “Systematic Comparative Analysis of Single-Nucleotide Variant Detection Methods from Single-Cell RNA Sequencing Data”. In: *Genome Biology* 22.1 (2021), p. 284. DOI: *10.1186/s13059-021-02462-6*.

[81] S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. “Smart-seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells”. In: *Nature Methods* 10 (2013), pp. 1096–1098. DOI: *10.1038/nmeth.2639*.

[82] K. Ellrott et al. "Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines". In: *Cell Systems* 6.3 (Mar. 2018), 271–281.e7. DOI: *10.1016/j.cels.2018.03.002*.

[83] K. J. Karczewski, L. C. Francioli, G. Tiao, et al. "The mutational constraint spectrum quantified from variation in 141,456 humans". In: *Nature* 581 (2020), pp. 434–443. DOI: *10.1038/s41586-020-2308-7*.

[84] E. Mitchell et al. "Clonal Dynamics of Haematopoiesis Across the Human Lifespan". In: *Nature* 606 (2022), pp. 343–350. DOI: *10.1038/s41586-022-04784-4*.

[85] A. Patel, I. Tirosh, J. Trombetta, A. Shalek, S. Gillespie, H. Wakimoto, D. Cahill, B. Nahed, W. Curry, R. Martuza, et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". In: *Science* 344.6190 (2014), pp. 1396–1401. DOI: *10.1126/science.1254257*.

[86] I. Tirosh, B. Izar, S. Prakadan, M. Wadsworth, D. Treacy, J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, et al. "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq". In: *Science* 352.6282 (2016), pp. 189–196. DOI: *10.1126/science.aad0501*.

[87] J. Yan, M. Ma, and Z. Yu. "bmVAE: a variational autoencoder method for clustering single-cell mutation data". In: *Bioinformatics* 39.1 (2023). DOI: *10.1093/bioinformatics/btac790*.

[88] T. H. Coorens, M. Spencer Chapman, N. Williams, I. Martincorena, M. R. Stratton, J. Nangalia, and P. J. Campbell. "Reconstructing phylogenetic trees from genome-wide somatic mutations in clonal samples". In: *Nature Protocols* 19.6 (2024). Epub 2024 Feb 23, pp. 1866–1886. DOI: *10.1038/s41596-024-00962-8*.

[89] A. Kiran and P. V. Baranov. "DARNED: a DAtabase of RNa EDiting in humans". In: *Bioinformatics* 26.14 (July 2010), pp. 1772–1776. DOI: *10.1093/bioinformatics/btq285*.

[90] G. Ramaswami and J. B. Li. "RADAR: a rigorously annotated database of A-to-I RNA editing". In: *Nucleic Acids Research* 42.Database issue (Jan. 2014), pp. D109–D113. DOI: *10.1093/nar/gkt996*.

[91] L. Mansi, M. A. Tangaro, C. Lo Giudice, T. Flati, E. Kopel, A. A. Schaffer, T. Castrignanò, G. Chillemi, G. Pesole, and E. Picardi. "REDIportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments". In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D1012–D1019. DOI: *10.1093/nar/gkaa916*.

[92] A. Brauner, O. Fridman, O. Gefen, and N. Balaban. "Distinguishing between resistance, tolerance and persistence to antibiotic treatment". In: *Nature Reviews Microbiology* 14.5 (2016), pp. 320–330. DOI: *10.1038/nrmicro.2016.34*.

[93] A. Hata et al. "Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition". In: *Nature Medicine* 22.3 (2016), pp. 262–269. DOI: *10.1038/nm.4040*.

[94] S. Shaffer et al. "Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance". In: *Nature* 546.7658 (2017), pp. 431–435. DOI: *10.1038/nature22794*.

[95] D. Rosano et al. "Long-term multimodal recording reveals epigenetic adaptation routes in dormant breast cancer cells". In: *Cancer Discovery* 14.5 (2024), pp. 866–889. DOI: *10.1158/2159-8290.CD-23-1161*.

[96] W. C. Zhang et al. "miR-147b-mediated TCA cycle dysfunction and pseudohypoxia initiate drug tolerance to EGFR inhibitors in lung adenocarcinoma". In: *Nature Metabolism* 1.4 (2019), pp. 460–474. DOI: *10.1038/s42255-019-0052-9*.

[97] P. Falletta, C. R. Goding, and Y. Vivas-García. "Connecting Metabolic Rewiring With Phenotype Switching in Melanoma". In: *Frontiers in Cell and Developmental Biology* (2020). DOI: *10.3389/fcell.2020.00030*.

[98] P. Karki, V. Angardi, J. C. Mier, and M. A. Orman. "A Transient Metabolic State in Melanoma Persister Cells Mediated by Chemotherapeutic Treatments". In: *Frontiers in Molecular Biosciences* (2022). DOI: *10.3389/fmolb.2022.841081*.

[99] A. Roesch et al. "Overcoming intrinsic multidrug resistance in melanoma by blocking the mitochondrial respiratory chain of slow-cycling JARID1B^{high} cells". In: *Cancer Cell* 23.6 (2013), pp. 811–825. DOI: *10.1016/j.ccr.2013.05.003*.

[100] R. Vendramin et al. "Activation of the integrated stress response confers vulnerability to mitoribosome-targeting antibiotics in melanoma". In: *Journal of Experimental Medicine* 218.9 (2021), e20210571. DOI: *10.1084/jem.20210571*.

[101] J. A. Berger et al. "IRS1 phosphorylation underlies the non-stochastic probability of cancer cells to persist during EGFR inhibition therapy". In: *Nature Cancer* 2.10 (2021), pp. 1095–1109. DOI: *10.1038/s41571-021-00517-2*.

[102] A. Noronha, N. Belugali Nataraj, J. S. Lee, B. Zhitomirsky, Y. Oren, and et al. "AXL and Error-Prone DNA Replication Confer Drug Resistance and Offer Strategies to Treat EGFR-Mutant Lung Cancer". In: *Cancer Discovery* 12.11 (Nov. 2022), pp. 2666–2683. DOI: *10.1158/2159-8290.CD-22-0111*.

[103] S. Shen et al. "An epitranscriptomic mechanism underlies selective mRNA translation remodelling in melanoma persister cells". In: *Nature Communications* 10.1 (2019), p. 5713. DOI: *10.1038/s41467-019-13360-6*.

[104] K. Pakos-Zebrucka, I. Koryga, K. Mnich, M. Ljujic, A. Samali, and A. M. Gorman. *The integrated stress response*. In: *EMBO Reports* vol. 17 (2016), pp. 1374–1395. DOI: *10.15252/embr.201642195*.

[105] C. Hu, J. Yang, Z. Qi, H. Wu, B. Wang, F. Zou, H. Mei, J. Liu, W. Wang, and Q. Liu. "Heat shock proteins: Biological functions, pathological roles, and therapeutic opportunities". In: *MedComm* 3.3 (2022), e161. DOI: *10.1002/mco2.161*.

[106] S. Aibar, C. González-Blas, T. Moerman, et al. "SCENIC: single-cell regulatory network inference and clustering". In: *Nature Methods* 14 (2017), pp. 1083–1086. DOI: *10.1038/nmeth.4463*.

[107] D. DeTomaso and N. Yosef. "Hotspot identifies informative gene modules across modalities of single-cell genomics". In: *Cell Systems* 12.5 (May 2021), 446–456.e9. DOI: *10.1016/j.cels.2021.04.005*.

[108] A. Al'Khafaji, J. Smith, K. Garimella, M. Babadi, V. Popic, et al. "High-throughput RNA isoform sequencing using programmed cDNA concatenation". In: *Nature Biotechnology* 42.4 (Apr. 2024), pp. 582–586. DOI: *10.1038/s41587-023-01815-7*.

[109] B. Adamson et al. "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response". In: *Cell* 167.7 (2016), 1867–1882.e21. DOI: *10.1016/j.cell.2016.11.048*.

[110] C. Weng, F. Yu, D. Yang, et al. "Deciphering cell states and genealogies of human haematopoiesis". In: *Nature* 627 (2024), pp. 389–398. DOI: *10.1038/s41586-024-07066-z*.