Tethering distinct molecular profiles of single cells by their lineage histories to investigate sources of cell state heterogeneity

Anna Minkina

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Jay Shendure, Chair

Doug Fowler

Andrea Wills

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Tethering distinct molecular profiles of single cells by their lineage histories to investigate sources of cell state heterogeneity

Anna Minkina

Chair of the Supervisory Committee:
Jay Shendure
Genome Sciences

Gene expression heterogeneity is ubiquitous within single cell datasets, even among cells of the same type. Heritable expression differences, defined here as those which persist over multiple cell divisions, are of particular interest, as they can underlie processes including cell differentiation during development as well as the clonal selection of drug-resistant cancer cells. However, heritable sources of variation are difficult to disentangle from non-heritable ones, such as cell cycle stage, asynchronous transcription, and measurement noise. Since heritable states should be shared by lineally related cells, I sought to leverage CRISPR-based lineage tracing, together with single cell molecular profiling, to discriminate between heritable and non-heritable variation in gene expression. I show that high efficiency capture of lineage profiles alongside single cell gene

expression enables accurate lineage tree reconstruction and reveals an abundance of progressive, heritable gene expression changes. I find that a subset of these are likely mediated by structural genetic variation (copy number alterations, translocations), but that the stable attributes of others cannot be understood with expression data alone. Towards addressing this, I develop a method to capture cell lineage histories alongside single cell chromatin accessibility profiles, such that expression and chromatin accessibility of closely related cells can be linked via their lineage histories. I call this indirect “coassay” approach "THE LORAX" and leverage it to explore the genetic and epigenetic mechanisms underlying heritable gene expression changes. Using this approach, I show that we can discern between heritable gene expression differences mediated by large and small copy number changes, *trans* effects, and possible epigenetic variation.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Many, many people have helped shape my path as a human and scientist during my time in graduate school. From providing votes of confidence and words of encouragement when they were needed, to giving ample scientific guidance and assistance, to being eager to think deeply with me through complex analyses, to making it clear to me that I was not alone if/when I got stuck, I am immensely grateful to the dozens of colleagues, family, and friends who have contributed directly and indirectly to the work presented here.

I am grateful to the whole Shendure lab, especially Darren Cusanovich, Riza Daza, Jean-Benôit Lalanne, and Sanjay Srivatsan for experimental guidance and helpful conversation, and to my rotation mentor and co-author Junyue Cao for being the brainchild behind several experimental methods presented here. I am grateful to students and faculty across Genome Sciences who supported, encouraged, and advised me: to Phil Green for encouraging me to take Genome 540 despite very limited coding skills and showing me that algorithms are just a set of simple logical steps; to Cole Trapnell, for advising me to get really, really good at ggplot (I did!); to Bill Noble, Brian Beliveau, & Maitreya Dunham for being exceptional examples of thoughtful and effective teaching; to my designated go-to's – Mitchell Vollger for all things coding, Jean-Benoît Lalanne for all things statistics, and Sanjay Srivatsan for a wide variety of things – who made me feel like I could never be truly stuck; to Brian Giebel and Serena Newhall for making the logistics of graduate school unexpectedly simple (most notably Brian for making sure there was shrimp at recruitment dinner, just for me).

I am grateful to my committee – Doug Fowler, Kelley Harris, Erick Matsen, and Andrea Wills—for being not only being excellent scientific advisors but also for being supportive and encouraging mentors as I have navigated bridging my career and personal goals over the last few years. I am grateful to the wonderful colleagues and friends who offered feedback on drafts of my general audience scientific writing (Jay Shendure, Lea Starita, Jase Gehring, Sanjay Srivatsan, Trent Waterman, and Allison Pfeiffer).

I am grateful to my family and friends, of whom there are too many to name, for being both cheerleaders of and distractions from my work, whichever was needed at the time, and for reminding me to trust myself every step of the way. I am immensely grateful to the little family I have built while in graduate school – my husband, Trent Waterman, for moving across the country with me, for his often surprisingly enlightening out-of-the-box "non-scientist" perspectives on my scientific conundrums, and for accepting my annoying habit of talking loudly to myself when I think; and to my baby son, Elliott, for putting everything into perspective.

Finally, I am extremely grateful to Jay Shendure for giving me the resources, freedom, and support to shape my own scientific career in accordance with my professional and personal goals – in short, for allowing me to choose how I spend my time. In many ways I owe the completion of the projects presented here to his unwavering optimism. But I also owe my development as a scientist and writer to his encouragement of my individual aspirations, whether or not they were aligned with his. I deeply appreciate the genuine care and humanness with which Jay approaches mentorship; in the sometimes deeply stressful moments which seem to be inherent to graduate school, I have been immensely thankful to know that I am supported as both a scientist and a person.

Chapter 1. INTRODUCTION

1.1 Opening Remarks

The Ph.D. output we most often discuss is the tangible one, the breadth of scientific work produced during the years one spent in graduate school. But arguably, the most important output of graduate school is you, the scientist. My hope at the outset of graduate school, whether I could articulate it then or not, was to develop the confidence and technical flexibility to tackle the scientific problems which captivated me most. Though the scientific questions themselves evolved over the years, several common themes emerged. First, I wanted to have the technical skills to address any and all curiosities I had about my data; I learned quickly that using exclusively available tools restricts you to answering the limited set of questions they were built to address. Second, I continually found that I wanted to *see* my data -- to make any hidden structure in it clearly visible to the naked eye. Third, I wanted to make my work accessible to a broad audience. Often, seemingly complex analyses are at their core a series of simple logical steps; presenting them clearly enables others with a broad range of expertise to contribute meaningfully to the work.

In this dissertation, I describe my scientific contributions to the cell lineage tracing field. My choice of the broad scientific problem and the specific puzzles I chose to solve along the way were largely driven by the aspirations above: (1) to collect complex data in which hidden patterns are expected but not readily obvious; and (2) to develop the computational skills necessary to build from scratch any tool I need to access these patterns.

I have also included several articles I wrote for a general audience at the outset of the COVID19 pandemic in response to widespread public confusion about testing (distributed on medium.com). The first explains the biology behind the PCR test and systematic issues more broadly to address misconceptions about the slow rollout of testing in the U.S. The second responds to a New York Times article which, in an attempt to promote the widespread distributions of antigen tests, made dangerous assertions about the shortcomings of the PCR test. This work felt particularly significant and stimulating: it highlighted for me that by sharing our expertise in public-facing forums, we as scientists can have a truly meaningful impact on individual decision making and broader policy decisions.

1.2 THE RAPID EVOLUTION OF GENOME EDITING-BASED LINEAGE TRACING METHODS

Since John Sulston heroically and painstakingly reconstructed the entire cell lineage tree of *C. elegans* (consisting of about a thousand cells) by eye (Sulston et al., 1983), the quest to reconstruct cell lineages of more complex organism has been an ambitious, but elusive, aspiration. Classical approaches to lineage tracing in complex organisms involve marking cells early in development, originally via injectable dyes and later via genetically-encoded fluorescent markers, where fluorescence is induced by recombination events, and tracing progeny cells to determine which organs and/or cells types are lineally related (reviewed in Kester & van Oudenaarden, 2018; Kretzschmar & Watt, 2012; VanHorn & Morris, 2021). Though innovative and informative for their time, these approaches are limited by the number of unique labels available, allowing for tracing of only a small number of lineages within a single organism. Moreover, because the labels can only be evaluated by visual methods (e.g. by eye or FACS), cell types may be hard to pinpoint without additional targeted labeling.

Genetically-encoded labels – ones which can be read from DNA or RNA – theoretically provide a solution whereby both label and cell type information (say, an expression profile) can be captured together. Moreover, genetic labels can be diverse, theoretically enabling cells to acquire a series of progressive labels to mark temporally-asynchronous lineage relationships within one organism. The discovery of the CRISPR-Cas9 adaptive immunity system in bacteria (Jinek et al. 2012) revolutionized genome editing, providing a new foothold for the development of lineage tracing technologies. In the decade since, a number of innovative methods have emerged which use programmable genome editing to generate diverse, progressive, and permanent genetic changes which in some cases can be captured alongside other single cell features like expression. These methods have in turn brought forth a new set of challenges, broadly falling into three categories: (1) How to generate edits which are sufficiently diverse and progressively acquired to enable the theoretical reconstruction of high resolution lineage trees; (2) How to capture recorded lineage information alongside single cell technologies; and (3) How to reconstruct accurate cell lineage trees from often incomplete data. In this thesis, I address each of these challenges. I first present an overview of existing lineage recording and capture approaches, highlighting their successes and limitations. I then summarize various lineage reconstruction approaches, with an eye on the challenges inherent in developing computational methods for a quickly evolving technology. I will then describe the lineage tracing method I have developed and address the advantages it has over existing methods in each of the categories above. I will discuss the outstanding challenges which my work highlighted, how the concepts I present here can be used to enhance other approaches, and the recent developments in lineage tracing which may address some of these challenges. Finally, I discuss potential future applications of the approaches presented here.

1.2.1 *The inception of CRISPR -based lineage tracing methods: GESTALT*

The CRISPR-Cas9 system in theory addresses two important challenges within genome editing-based lineage tracing. First, edits can easily be directed to specific places in the genome, and thus more easily captured from DNA or RNA. Second, programmed genetic changes are not limited to a few outcomes as they had been using previous technologies. Diverse editing can theoretically be achieved in several ways. One can supply a cell with a set of repair templates, such that a diverse array of programmed insertions and/or deletions can result from a CRISPR-induced double strand break. Such an approach is practically challenging as template-based repair has historically not been particularly efficient (H. Yang et al., 2020). A simpler approach and one which underlies a number of existing lineage tracing technologies is to rely on the intrinsic deficiencies in double strand break (DSB) repair machinery: a subset of the time, DSBs are repaired imperfectly, leaving behind small insertions, mismatches, and/or deletions.

The inaugural lineage tracing method to make use of this phenomenon is called GESTALT (McKenna et al. 2016). McKenna, Findlay, Gagnon *et al.* engineered cultured cells and zebrafish genomes to contain multiple CRISPR targets, each of which can theoretically be a site for mutagenesis. The remaining CRISPR components (Cas9 & sgRNAs) were introduced transiently or stably into cells, and transiently (via injection) into zebrafish. To simplify capture of the entire set of targets within an individual cell, 10 targets were placed in tandem, with 3bp spacer sequences between them. Thus, the entire array could be amplified from DNA or RNA as a single molecule.

The benefits of this structure are obvious: capturing all lineage information as a simple amplicon allows for easy translation of editing information into a cell's complete lineage profile, and, as the authors show, enables one to infer a global set of lineage profiles in a tissue/group of cells from bulk RNA/DNA. But the experiments presented in the paper also illustrate a major drawback to this approach: there is a high degree of interaction between targets, resulting in frequent loss of editing information. This can occur in three ways. First, repair events which result in deletions which are more than a few bases long can remove neighboring targets, obscuring previously recorded information or removing an unedited locus, making it unavailable for subsequent recording. This problem can in principle be solved by increasing the distance between targets, but this solution is practically limited by sequencing constraints, with poor clustering of long amplicons using the Illumina short-read platforms. Second, the editing outcomes visualized in the paper show a substantial degree of "inter-target deletions," where deleted bases fall neatly between two CRISPR cut sites. Such an outcome suggests that the second cut occurred before the first was able to be repaired, and thus the repair removed the intervening sequence. Reducing the rate of editing may theoretically address this problem, though the rate at which cuts are repaired *perfectly* is an integral parameter here; if imperfect repair is exceedingly rare relative to the amount of cutting, reducing cutting rate sufficiently to address this phenomenon decrease editing frequency to unworkable levels. Finally, editing of the first and/or last target in the array may remove primer binding sites necessary for capture/amplification of the array, resulting in loss of the complete lineage profile associated with a cell. In fact, McKenna, Findlay, Gagnon et al. describe a reduction in the total number of uniquely-edited target arrays in zebrafish over time, a finding which is attributed to loss of clonal diversity over the course of development, but may perhaps also arise from full target array loss over the course of progressive editing.

McKenna, Findlay, Gagnon et al. also present an approach to control editing rate by generating mismatches between sgRNAs and CRISPR targets. In theory, targets with mismatches are cut much less frequently when mismatches are present, creating a set of loci where later developmental relationships may be recorded. Such an approach continues to be explored, along with other approaches to reduce editing rate described below.

Finally, to evaluate the relationship between lineage relationships and cell differentiation in zebrafish, McKenna, Findlay, Gagnon et al. dissected adult zebrafish organs and profiled the set of lineage barcodes associated with each using bulk PCR. They were thus able to describe organ-level clonal contributions but were limited in investigating cell type-level clonality distributions.

In the years since this seminal paper was published, other work has built upon these principles to address the main challenges described above: increasing editing capacity, reducing the rate of information loss during recording/editing, capturing lineage information alongside single cell molecular profiles to infer cell type-resolved clonal relationships, and reconstructing accurate lineage from this information (discussed in more detail below).

1.2.2 *The evolution of CRISPR-based lineage tracing method in zebrafish models*

Zebrafish emerged as an ideal organism for benchmarking CRISPR-based lineage tracing approaches, since the ability to inject CRISPR components directly into the one cell embryo allowed for quick iterations. Though the original approach generated a transgenic zebrafish harboring an editable target array, several groups made use of existing transgenic zebrafish lines with convenient editable loci. Alemany et al. (2018) (Alemany et al. 2018) developed ScarTrace,

using of a zebrafish line with eight genomic copies of GFP integrated in tandem. sgRNAs were designed to target a single position within the GFP gene, greatly increases the space between targets relative to GESTALT. This strategy addresses one major shortcoming of the original approach: repair events resulting in large deletions are less likely to remove neighboring targets. However, because targets are still located close together and editing rate is relatively uncontrolled, multiple cuts occurring close together in time can still result in inter-target deletions, removing previously recorded in formation. The frequency with which this occurs in the presented study was difficult to assess, as target sequences were identical and amplification of the entire array was not performed.

The distance between targets poses a challenge relative to GESTALT: it is not possible to amplify and sequence the entire array as a single amplicon. Therefore, it is not possible to infer a single cell's lineage profile (i.e. the complete set of edits associated with that cell) from bulk profiling. To this end, Alemany et al. adapted SORT-seq, developed for scRNA-seq, to capture transcriptomes alongside lineage profiles. Single cells are sorted into a 384-well plate, and reverse transcription is performed to capture transcriptomes. Alemany et al. note that "scars" made within GFP constructs can be captured from both RNA & DNA, but GFP expression is vulnerable to promoter shut-off. Thus, they perform an additional nested PCR step to amplify the sgRNA-targeted GFP loci from each single cell, enabling them to associate an expression profile with a set of clone-defining scars.

Profiling expression and lineage from the same single cells enables Alemany et al. (2018) to assess relative clonal contributions at the level of the cell type, without the need for dissection or cell type

labeling, but their approach has several drawbacks. First, their single cell profiling method is relatively low throughout, limiting them to assessing just hundreds of cells. Second, the eight targets available for editing are indistinguishable from one another. Thus, it is not possible to discern whether a shared edit between two cells is indicative of shared clonality (i.e. occurs at the same locus), or identical edits were independently introduced at two different loci. Though the authors find hundreds of unique editing patterns associated with a single targeted sequence, work presented in this dissertation and that of others (McKenna et al. 2016; W. Chen et al. 2019) shows that CRISPR repair outcomes at a single targeting sequence are not evenly distributed, with high likelihood that the same outcome occur more than once. Such a setup greatly complicates lineage tree reconstruction, and consistent with this, the authors focus on static clonal groups over progressive lineage divergence.

A similar approach, LINNEAUS (Spanjaard et al. 2018), used a zebrafish line with multiple RFP genes spread throughout the genome. Such a structure greatly reduces the likelihood of inter-target deletions plaguing both ScarTrace and GESTALT. Additionally, Spanjaard et al. captured edited RFP genes via targeted transcript capture alongside full transcriptomes via a droplet-based approach, greatly increasing the number of cells they were able to profile relative to ScarTrace. This approach, however, retains some of the drawbacks observed in ScarTrace, notably, that targets are indistinguishable from one another and thus common editing outcomes occurring at different targets may be interpreted erroneously as a sign of clonality. Spanjaard et al. address this problem more formally than previous methods. They evaluate the set of editing outcomes from multiple zebrafish embryos, pinpoint those outcomes which occur in more than one embryo as likely also having multiple intra-embryo origins, and removing them from downstream analyses

for all embryos. This approach likely improves the accuracy of tree reconstruction, but at the expense of lineage information loss.

A third method developed concurrently with those described above is scGESTALT (Raj et al. 2018), which improves upon the original approach in several important ways. First, the new construct is expressed off a heat shock promoter, such that expression can be induced prior to dissociation and target array captured alongside single cell expression profiles via a droplet-based approach. Second, Raj et al. introduce an innovative method to control the timing of editing. Previous methods introduced Cas9 and sgRNAs via injection at the one-cell stage, resulting in editing only during very early developmental stages. Raj et al. use such an approach to introduce edits in the first four targets in their array but introduce an inducible system for initiating editing of the remaining five targets at a later timepoint. Here, a construct containing constitutively expressed sgRNAs and heat shock-inducible Cas9 is integrated into the genome and meets the target construct only during fertilization. Low expression of these five sgRNAs relative to the concentration of the injected set early in development results in robust editing only of the first four targets until the heat shock-activate Cas9 is induced to be expressed. Since heat shock timing can be varied, this innovative approach allows for lineage recording at various developmental stages. Though the arrayed design still predisposes the construct to inter-target deletions, the authors find, as expected, that inter-target deletions occur much less frequently between the 'early' and 'late' edited targets, as expected if near-simultaneous cutting is a prerequisite for this phenomenon.

1.2.3 *CRISPR-based lineage tracing applied to mammalian systems*

Raj et al. showed that one can achieve progressive editing by introducing all lineage tracing components stably into the genome, paving the way for CRISPR-based lineage tracing in

mammalian systems. The work of Kalhor et al. (2018) (Kalhor et al. 2018) first suggested this was possible. Kalhor et al. introduced their homing CRISPR system (Kalhor, Mali, and Church 2017) into mice, in which sgRNAs simultaneously serve their traditional role and that of the target itself, resulting in continuous sequence diversification. In this system, the 60 targets/hgRNAs (i.e. homing guide RNAs) continuously edit themselves, miraculously not substantially interfering with normal development. Such a system is impractical for reconstructing complex lineage trees, as previously recorded information is continually being destroyed, but it does enable editing to continue well into development in a mammalian system, resulting in a diverse set of clonal labels from which clonal dynamics can be inferred. Importantly, this study shows that continuous double strand breaks are compatible with mammalian development.

The first more traditional CRISPR-based approach for lineage tracing applied in mice was presented by Chan et al. (2019) (Chan et al. 2019). Their target design is a hybrid between the array and dispersed individual targets; many short, 3-target arrays are stably integrated into the genome via oocyte injection of a piggyBac transposon vector, also expressing sgRNAs. Constitutively-expressed Cas9 is introduced during fertilization via sperm injection. Importantly, unlike previous methods which use the dispersed approach, individual target arrays contain a unique 8-bp barcode, making them distinguishable from one another irrespective of editing outcome. To mitigate the likelihood of inter-target deletions, Chan et al. introduced sequence mismatches between a subset of sgRNAs & targets. They noted that, with the exception of arrays which contained multiple perfect match target/sgRNA pairs, inter-target deletion rate was markedly reduced. This observation is promising, suggesting that for methods with arrayed targets, reducing editing rate in a variety of ways may address inter-target deletions.

To profile lineage concurrently with expression, Chan et al. designed their lineage constructs with a polyA capture sequence so that constructs are captured from mRNA via the 10x Genomics platform. Though some lineage information was captured alongside the majority of single cell expression profiles, lineage profiles were largely incomplete. Across 7 embryos with 3-15 targets per embryo, Chan et al. (2019) recovered at least one edited lineage array from 15-75% of cells per embryo, but just one target array was captured efficiently (>25% of cells) in 6 of 7 embryos (Chan et al. 2019). Such low recovery largely precludes accurate lineage tree reconstruction.

In a related lineage tracing method applied in mice, called CARLIN, Bowling et al. (2020) (Bowling et al. 2020) again record information to a single target array reminiscent of GESTALT, but the transgenic Cas9 construct is doxycycline-inducible, giving the authors control over when editing occurs. They show that administering doxycycline via drinking water to adult mice is sufficient to induce editing, allowing them to track clonal relationships between proliferating cells in the adult bone marrow. However, their target array design suffers from similar shortcomings to previous methods: they observe many large, inter-target deletions obscuring previously recorded information.

Together, these studies illuminate both the potential of CRISPR-based lineage tracing and the substantial challenges which remain to be addressed before truly accurate, high-resolution cell lineage tracing is achieved.

1.2.4 *Alternative, non-CRISPR, methods for progressive genome editing for lineage tracing*

Several methods introduce genomic changes for lineage reconstruction without the use CRISPR-mediated double strand break repair. Wagner et al. (2018) (Wagner et al. 2018) developed TracerSeq, a transposition-based approach where the Tol2 transposase system is used to integrate GFP reporter containing a diverse library of barcodes in its 3' UTR. All components of the transposase system are injected into the zebrafish embryo, requiring no prior genome engineering. Importantly, injected plasmids persist over multiple cell divisions and can thus integrate sequentially over time, theoretically enabling a reconstruction of a multi-tier tree. The GFP constructs are expressed such that the barcodes appear in a standard scRNA-seq library.

While novel, this approach has several limitations. First, there are only a few systems (e.g. zebrafish, frogs) which are amenable to injection to introduce editing component. Second, though transposition can occur over the span of multiple cell divisions during early development, such a system is not applicable to track lineage relationships beyond early development. Third, expression can theoretically occur from the transposon plasmid prior to integration, with mRNA molecules persisting in cells in which no integration has occurred. One can imagine this phenomenon generating a high degree of noise and complicating lineage reconstruction. In fact, Wagner et al. observe a transposition rate which far exceeds those previously reported for the Tol2 system (Urasaki, Asakawa, and Kawakami 2008), suggesting some degree of non-integrated plasmid expression.

Hwang et al. (2019) (Hwang et al. 2019) use a nickase Cas9 (nCas9) fused with a cytidine deaminase to introduce genomic edits via base editing into endogenous L1 loci. Cytidine

deaminase converts C:G base pairs to T:A base pairs within a small region (4-8 nucleotides) from the PAM protospacer. Importantly, double strand breaks are not induced as part of the editing process, a promising development for applying progressive lineage tracing to complex systems where the DSBs themselves may perturb cell fate decisions. However, given the relatively small number of bases available for conversion at each locus, multiple identical events are bound to occur independently many times, hindering accurate tree reconstruction.

Finally, Loveless et al. (2021) (Loveless et al. 2021) introduced CHYRON, an innovate approach where editing events are a series of short (1-7 base) insertions at a target locus. To achieve this, the authors use homing guide RNAs (hgRNAs) which serve both as locator and target site, and terminal deoxynucleotidyl transferase (TdT), a template independent DNA polymerase. While this approach has its own set of challenges (e.g. the algorithmic challenge of splitting a set of inserted bases of variable length into individual insertion events), it signals a shift towards methods where editing in more controlled, i.e. where large insertions and/or deletions are not a major sourse of information loss.

Together, these methods exhibit astoundingly quick developments in progressive, genome editing-based lineage tracing, and highlight a number of outstanding challenges. These include constructing a lineage recorder in which edits to not interfere with those at other targets, potentially better control of editing rate, and high efficiency capture alongside single cell expression profiles. Each of these advances would improve our ability to accurately reconstruct lineages, as outlined below.

1.3 COMPUTATIONAL METHODS FOR CELL LINEAGE TREE RECONSTRUCTION: A BRIEF OVERVIEW

With the rapid advances in experimental lineage tracing methods, computational methods for lineage inference from CRISPR-based data have had to evolve alongside. While cell lineage reconstruction is conceptually a similar problem to phylogenetic reconstruction, traditional phylogenetic tools have come up short to adequately address this problem. A number of assumptions (e.g. relative infrequency of genetic changes; low likelihood of convergence events, etc.) underlying phylogenetic algorithms do not hold true in cell lineage tracing data, and the sheer number of cells often precludes the use of algorithms which attempt to reconstruct all possible trees and evaluate them against each other. Moreover, CRISPR processes are arguably more challenging to model than evolutionary ones, and different methods require the consideration of different outcomes (e.g. are inter-target deletions possible?), making it impossible to develop a fully automated and globally-applicable algorithm for cell lineage reconstruction from CRISPR data. In this section, I present a brief overview of the methods which have been proposed and the challenges encountered.

The original GESTALT (McKenna et al. 2016) paper uses Camin-Sokal maximum parsimony, a phylogenetic approach implemented in the PHYLIP package (Felsenstein 2009). PHYLIP's implementation is written with evolutionary tree reconstruction in mind and thus has capacity to accommodate only a few variable options per genomic position. To accommodate the variety of editing outcomes observed in GESTALT, McKenna, Findlay, Gagnon et al. modified the input data such that each unique edit observed anywhere in the target array is given its own unique position in the lineage profile supplied to PHYLIP, with possible 'states' being the presence or

absence of the edit. Though this approach enables the encoding of multiple editing outcome at a single target array, it has the disadvantage of making all available editing patterns independent of one another, such that a cell's inferred lineage can in principle contain sequential, overlapping edits (Feng et al., 2021). Such a trajectory is not biologically valid, and highlights one challenge in applying phylogenetic algorithms to CRISPR-based cell lineage data.

The other early CRISPR-based approaches described above were less readily amenable to phylogenetic reconstruction methods because targets (e.g. multiple GRP or RFP loci) were not distinguishable from one another. For example, Spanjaard et al. (Spanjaard et al. 2018) found that Camin-Sokal parsimony produced inaccurate trees, and instead developed a graph-based approach. Edges are drawn between any edits found together in the same cell, and the most connected edit becomes the root. This edit is then removed, and the next most connected edit forms the following branch, iteratively constructing a tree such that each terminal branch is associated with a list of edits expected in that lineage. Cells can then be placed at the tips of this tree. The authors acknowledge that some cells cannot be placed at a terminal end of the tree due to missing data, but also point out that in some cases, missing data can be inferred using this approach. This edit abundance-based algorithm offers promise, but would clearly benefit from advances in experimental lineage tracing methods, including the ability to associate editing patterns with specific targets.

1.3.1 *Benchmarking algorithms using simulated data*

Benchmarking algorithms using experimental data is particularly challenging because the correct lineages are not known. Therefore, several groups have attempted to model CRISPR processes in order to generate simulated datasets which in theory reflect experimentally-collected data.

The largest comprehensive effort to solicit and benchmark algorithms was as part of a DREAM challenge (Gong et al., 2021). It consisted of several tasks, based either on the intMEMOIR (Chow et al., 2021) lineage recording strategy or on GESTALT. The intMEMOIR approach uses integrase to edit an array of 10 barcodes into two possible outcomes (the 'unedited' state serving as a third). Given the frequency with which the same event is bound to occur multiple times independently in different cells in a single experiment, reconstructing lineages from these data is algorithmically a different problem from CRISPR-based lineage inference, where much more edit diversity is expected. Thus, I will focus on the second challenge: reconstructing lineage relationship from simulated CRISPR-like editing data.

Two simulated datasets were generated. The first, modeling the development of *C. elegans*, postulated an editable array of 200 targets -- a number beyond what has been experimentally explored, but within a reasonable realm of possibility. The second challenge simulated the editing of a 1000-target array, with *M. musculus* development in mind. Simulations parameters were designed to be as consistent as possible with experimentally-observed phenomenon. Over many *in silico* cell divisions, targets stochastically accumulated edits, with a total of 32 possible states at each target (including unedited and deletion). All outcomes were not equally probable, but sampled from a gamma distribution. Inter-target deletions were also a possibility if edits occurred at two targets within 20 targets of each other in the same cell division and impacted 5-10% of targets in the simulation. Finally, drop-out events modeling loss of information during the capture stage were implemented in the *M. musculus* data.

Distance-based approaches which use hierarchical clustering dominated submissions (e.g. Liu, Guan, DCLEAR(Gong et al., 2022)). Because simulated training data was available, most strategies calculated "transition probabilities" associated with every edit and calculated distances between cells using these probabilities. This approach takes into account the fact that some edits are much more likely to occur multiple times independently than others. Thus, cells with rare edits are likely to be grouped together. Importantly, such probabilities may be calculated from experimental data if enough independent samples are collected, so such an approach is transferrable to non-simulated data. Hierarchical clustering approaches generally do well placing closely related cells next to each other, but the resulting tree does not necessarily reflect the set of editing events which generated the data.

One group (AMbeRland) used a machine learning approach to estimate parameters from the training data. Though such an approach can in principle work well to reconstruct test data trees, for it to work well in an experimental set, it is crucial that simulated parameters be consistent with real-world ones. As described above (and as we will see below), this is extremely challenging.

A stand-out approach, Cassiopeia (Jones et al., 2020), which was developed prior to the challenge, uses a maximum likelihood approach. Briefly, they first generate a Steiner Tree of all possible ancestral states and then use integrated linear programming (ILP) to solve this potential graph to arrive at the optimal solution. Because this approach was originally developed with relatively small (phylogenetic) datasets in mind, it is not possible to apply it to an entire dataset of thousands of cells. Thus, Jones et al. implement a greedy approach to split cells into related subgroups and reconstruct subtrees to be merged as a final step. The splitting algorithm is based on edit

abundance: subgroups are iteratively split based on the presence or absence of the most abundant edit. The problem with this approach arises from missing and/or erroneous data, an unfortunate but probably expected outcome of all lineage tracing methods. Such a splitting algorithm applied to raw data results in cells with missing/erroneous data being split from their close relatives early, without hope of reuniting in subsequent steps. In my work, I implement a similar greedy approach, but present a method to address missing/erroneous data such that close relative cells remain clustered together.

Finally, an innovative approach developed specifically with GESTALT data in mind (Feng et al., 2021), applies a penalized maximum likelihood estimation approach. Originally posted to bioRxiv in 2019, it paved the way for the challenge described above, considering carefully all the parameters which needed to be modeled to simulate life-like datasets. It is the only method to date which applies another phylogenetics-derived principle – time-estimation – to lineage reconstruction. Briefly, assuming constant rate of editing and enough editing capacity, one can in theory estimate the relative temporal properties associated with each lineage. This idea is extremely intriguing and particularly useful in complex organism development; it remains to be seen whether a system where editing rate is truly constant will be developed.

In all, accurate parameter estimation has proven to be an extremely challenging problem, in large part because (1) processes assumed to have a constant rate (e.g. Cas9 cutting) in practice appear not to; and (2) a number of complex and experiment-specific phenomenon underly missing data, and assuming it is randomly distributed creates some important asymmetries between simulated

and experimental data. In my work, I attempt to design an algorithm which is robust to missing data and has capacity to accommodate non-random missing data (e.g. target-specific loss).

1.4 AIM OF THIS WORK: APPLYING LINEAGE TRACING TO INVESTIGATE THE SOURCES OF CELL HETEROGENEITY

In the work presented here, I aim to address the challenges presented above, with regards to lineage recording, lineage capture alongside single cells molecular measurements (expression and chromatin accessibility), and computational reconstruction of lineage relationships. I apply the methods presented here to investigate the sources of heterogeneity in cultured cells. The work presented here is associated with a submitted manuscript, co-authored with Junyue Cao and Jay Shendure. Thus, the article "we" is used.

Single cell molecular profiling technologies have revealed extensive gene expression heterogeneity, even between cells of a single cell type (O'Leary et al. 2020; Patel et al. 2014; Y. H. Choi and Kim 2019; SoRelle et al. 2021; Muto et al. 2021; Li et al. 2022). Expression variation can arise from a number of sources, including transient phenomenon like cell cycle stage and transcriptional bursting (Tunnacliffe and Chubb 2020), as well as stable genetic (Ben-David et al. 2018) or epigenetic (Bonasio, Tu, and Reinberg 2010) differences within a cell population. Stable sources of variation are of particular interest as they are "heritable" over multiple cell divisions, and can thus serve as substrates for selection, altering a cell population over time. Such heritable phenomena may underlie differentiation during normal organismal development as well as the acquisition of drug resistance in cancer (Salgia and Kulkarni 2018). Yet within a set of single cell gene expression profiles, representing a population snapshot in time, it is difficult to distinguish between stable and transient expression variation. This is particularly challenging for cells of a

single cell type, where transient differences may mask heritable variation when performing clustering analysis to distinguish cell states (Kiselev, Andrews, and Hemberg 2019).

Heritable sources of expression variation have at least one property which distinguishes them from transient variation: because they are stable over multiple cell divisions, they should be shared by cells which are closely related by lineage. It follows that if all lineage relationships were known, we could discern heritable from non-heritable variation by assessing the distribution of variation across a lineage tree (Figure 1.1a) . While transient variation should be randomly distributed, stably maintained expression states should cluster together within the tree, *i.e.* tracking to a common "founder" event. Thus, lineage histories, coupled to gene expression profiling, could potentially enable the differentiation of heritable vs. non-heritable sources of expression variation.

Molecular methods for cell lineage history profiling compatible with concurrent expression profiling involve either static or progressive genetic barcoding. The static approach introduces short, transgenic barcodes to proliferating cells, such that closely related descendants share a barcode sequence (Rodriguez-Fraticelli et al. 2018; Weinreb et al. 2020; Biddy et al. 2018; Guo et al. 2019). Static barcoding might reveal heritable sources of gene expression that were acquired close to the time of labeling, but would presumably miss those occurring substantially earlier or later. In contrast, progressive lineage tracing methods (*e.g.* GESTALT and related methods described above), wherein cells accumulate sequence diversity at multiple genomic locations over time, facilitate reconstruction of multi-tier lineage trees, and might therefore be more sensitive with respect to detecting heritable gene expression variation (Alemany et al. 2018; McKenna et al. 2016; Wagner et al. 2018; Raj, Gagnon, and Schier 2018; Raj et al. 2018; Spanjaard

et al. 2018; Kalhor, Mali, and Church 2017; Kalhor et al. 2018; Chan et al. 2019; Loveless et al. 2021; Bowling et al. 2020; Perli, Cui, and Lu 2016; Hwang et al. 2019).

A high diversity of labels can be achieved via CRISPR/Cas9, where imperfect double strand break repair via NHEJ can generate a variety of outcomes (referred to here as "edits" or "indels") (Alemany et al. 2018; McKenna et al. 2016; Wagner et al. 2018; Raj, Gagnon, and Schier 2018; Raj et al. 2018; Spanjaard et al. 2018; Kalhor, Mali, and Church 2017; Kalhor et al. 2018; Chan et al. 2019; Loveless et al. 2021; Bowling et al. 2020; Perli, Cui, and Lu 2016). Over many cell divisions, the pattern of indels that accumulate at CRISPR/Cas9 targets are informative with respect to the lineage relationships amongst the cells in which they occur. Most strategies reported to date, whether implemented *in vitro* or *in vivo*, place several targets in tandem, such that the edits at these multiple targets can be recovered within a single DNA or RNA-derived sequencing read (Alemany et al. 2018; McKenna et al. 2016; Wagner et al. 2018; Raj, Gagnon, and Schier 2018; Raj et al. 2018; Spanjaard et al. 2018; Kalhor, Mali, and Church 2017; Kalhor et al. 2018; Chan et al. 2019; Loveless et al. 2021; Bowling et al. 2020; Perli, Cui, and Lu 2016).

In practice, however, there are a number of technical issues that limit this approach. First, arrays of CRISPR/Cas9 targets frequently acquire large deletions when concurrent DSBs at different targets within the array are joined, potentially excising previously recorded information at intervening targets. Second, read length limitations require targets to be placed close to one another, such that the editing of one target risks corrupting adjacent targets. Third, although it is possible to capture CRISPR/Cas9-edited lineage targets as part of a single cell RNA-seq (scRNA-seq) profile, this has usually been inefficient in practice. For example, using InDrops to capture a

tandem array of 10 CRISPR targets alongside single cell transcriptomes in juvenile zebrafish brains, Raj *et al.* (2018) recovered lineage profiles from just 6-28% of cells with expression profiles (Raj et al. 2018). Similarly, using 10X Genomics to capture arrays of 3 CRISPR targets from mouse embryos alongside scRNA-seq (3-15 array integrations per embryo), Chan *et al.* (2019) recovered at least one edited lineage array from 15-75% of cells per embryo, but just one target array was captured efficiently (>25% of cells) in 6 of 7 embryos (Chan et al. 2019). In each case, both target design and the method of capturing lineage targets during scRNA-seq likely contributed to the limited recovery.

Here, we introduce a CRISPR-based lineage tracing approach in which many distinct lineage recording loci are integrated independently throughout the genome. These targets can each accommodate relatively large deletions and insertions. We further show that, with targeted enrichment, they can be captured efficiently alongside transcriptomes via a combinatorial indexing approach (sci-RNA-seq) (Cao et al. 2017, 2019). To analyze data generated from a proof-of-concept *in vitro* monoclonal expansion, we developed a lineage tree reconstruction algorithm that is robust to missing data and recurrences (*i.e.* where identical edits occur independently), and validate the algorithm using copy number alterations (CNAs) that are evident in expression data. We show that incorporating lineage relationships into expression analysis reveals abundant heritable expression variation, including instances that are clearly explained by CNAs, but also many which are not.

Finally, towards investigating the mechanism(s) underlying expression heritability, we develop an approach to capture cell lineage relationships alongside single cell chromatin accessibility. We

show that we can link two distinct molecular features−gene expression and chromatin accessibility−via their lineage profiles (Figure 1.1b). We then use these lineage-tethered features to further distinguish between expression changes which can be explained directly by copy number alterations, ones likely mediated by *trans* effects of copy number alterations, and ones which are more likely to have resulted from a stable change in *cis* regulatory state. We term this approach THE LORAX: Tracking Heritable Events via Lineage-based Ordering of chRomatin Accessibility & eXpression profiles.

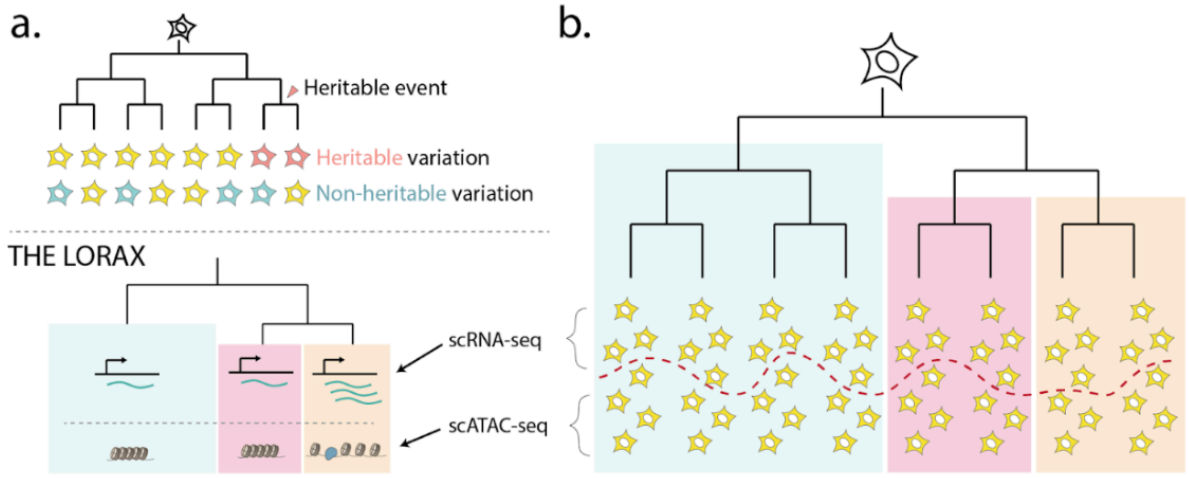


Figure 1.1. Tethering the molecular profiles of single cells by their lineage histories to investigate sources of cell state heterogeneity.

(**a**) A framework to distinguish heritable from non-heritable sources of gene expression variation using lineage relationships. (**b**) A framework for tethering single cell expression (scRNA-seq) and chromatin accessibility (scATAC-seq) measurements via lineage relationships to investigate the mechanisms underlying heritable expression variation (THE LORAX).

Chapter 2. DESIGNING & EVALUATING A LINEAGE RECORDING SYSTEM

2.1 LINEAGE TRACING CONSTRUCT AND EXPERIMENTAL DESIGN

We first set out to design a CRISPR/Cas9-based lineage tracing strategy that addresses outstanding technical challenges. Reconstructing an accurate, multi-tier lineage tree from progressively acquired edits requires the following: (a) multiple editable loci such that successive tagging can occur in a single lineage over time; (b) a high probability of diverse editing outcomes at a single target, such that identical edits at that target are unlikely to occur independently in different cells; (c) controllable editing machinery, such that target capacity is not exhausted quickly after editing onset; (d) permanence of edits, such that they are not likely to be overwritten or lost; and (e) a high rate of capture of editing information alongside single cell profiling of other features. Towards realizing these features, we designed a construct in which individual targets are integrated independently across the genome and captured as separate transcripts (Figure 2.2a-b). Each target contains a unique identifier sequence, which is positioned such that the target can accommodate up to a 70 bp deletion centered at the cut site without corrupting the identifier, as well as, assuming 300 bp read lengths, insertions of up to 105 bp. The sgRNAs are delivered on the same lentiviral construct as the targets, with targets expressed from a highly active EF-1α promoter to enable lineage capture from mRNA.

To generate cells with a high capacity for lineage recording, we transduced HEK293 cells at a high multiplicity-of-infection (MOI) with this construct and attempted to establish clonal populations. Even in the absence of editing, most clones grew poorly, with the lentiviral integrations themselves at this high MOI potentially contributing to toxicity. Across 26 clones, we observed integration

counts ranging from 2 to 53, with a median of 11 integrations (Figure 2.1a). We moved forward with a robust clone bearing 36 unique integrations, as evidenced by the diversity of unique identifier sequences ("target IDs"; Figure 2.1b). To induce editing, we transduced this clone again with a doxycycline-inducible Cas9 lentiviral construct, sorted single cells, and allowed a clonal population to grow from a single founder cell (such that all progeny cells comprise a single lineage tree). Interestingly, only 32 unique target IDs were observed after this second round of cloning, potentially due to karyotypic instability (discussed further below), while one integrant contained a mutation that corrupted its target site Figure 2.1b).

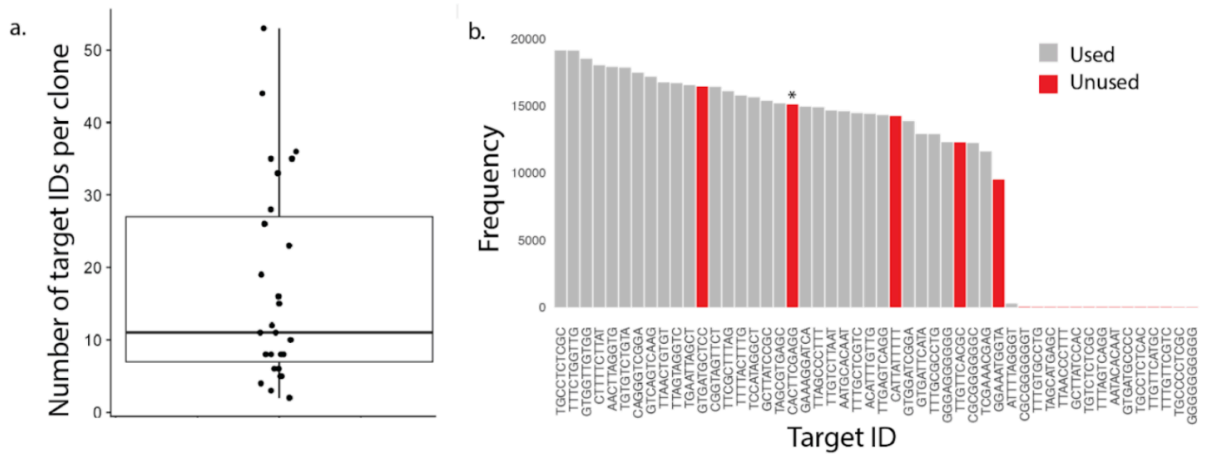


Figure 2.1. Evaluating lentiviral target integrations.

(**a**) Number of unique target IDs across 26 clones derived from high MOI transduction of HEK293 cells. Box shows median and encompasses counts in the second and third quartiles. Whiskers depict the interquartile range. (**b**) Frequency of each unique target ID within the unedited clone used for the main experiment. As discussed in the text, this clone was "re-cloned" following transduction with doxycycline-inducible Cas9 lentiviral construct, such that a single founder cell generated the tree. Four target IDs that were abundant after the first round of cloning were unobserved after this re-cloning step (red bars), while an additional one was corrupted by a mutation and therefore also excluded (red bar with asterisk). The remaining 31 abundant target IDs were carried forward in the analyses, with two of these "duplicated" *in silico* to account for their inferred duplication just before or during the clonal expansion.

After 35 days of expansion of this clone, with passaging as needed (**Methods**), a portion of the cells were harvested for single cell expression and lineage analysis, while the remaining cells were frozen down for subsequent profiling of chromatin accessibility. Of note, although doxycycline was not applied, we nonetheless observed diverse and progressive editing with this clone, presumably because of leaky expression of Cas9 (Costello et al., 2019). For concurrent acquisition of whole cell transcriptomes alongside lineage information, we performed 96 x 768 sci-RNA-seq, with processing of cells in eight batches during the second indexing step (Cao et al., 2017, 2019). To facilitate the efficient recovery of lineage targets from each cell, we introduced a supplemental set of reverse transcription primers during the first round of indexing, and split the material in half prior to indexed PCR during the second round of sci-RNA-seq2, with one half being used for the general transcriptome, and the other half for targeted recovery of the lineage profiles (**Methods**).

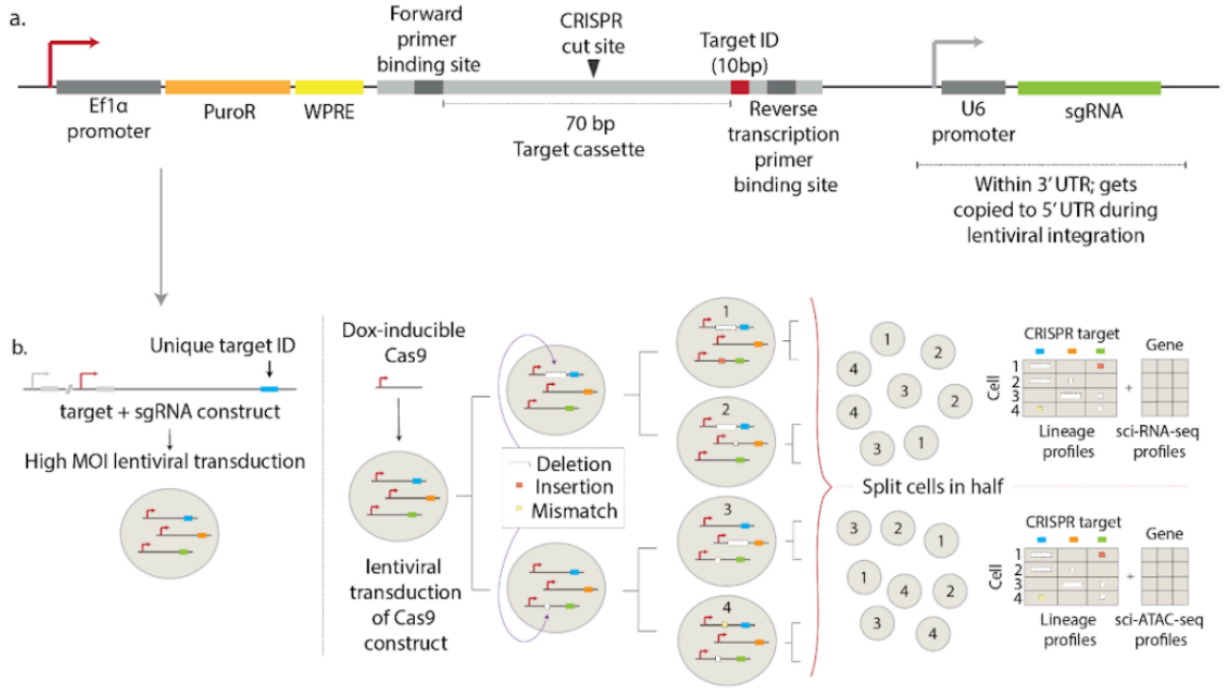


Figure 2.2. Lineage tracing construct and experimental design.

(**a**) Target vector design. A target cassette was integrated into the CROP-seq vector (Datlinger et al., 2017) as shown. (**b**) Schematic of experimental workflow. Cells were transduced at high MOI with constructs containing an sgRNA and barcoded target sequences, such that many integration events per cell were expected. A single clone was then transduced with a doxycycline-inducible Cas9 vector, single cells were sorted, and a single founder cell was allowed to divide for 35 days while editing occurred. The final cell population was split for either target capture alongside sci-RNA-seq or sci-ATAC-seq.

2.2 EVALUATING CONCURRENT CAPTURE OF LINEAGE AND SINGLE CELL EXPRESSION PROFILES

These libraries were sequenced, and the resulting reads were adaptor-trimmed, aligned to the reference human genome, and deduplicated. For the single cell transcriptomes, we observed a median of 13,212 UMIs per cell, across 15,525 cells (Figure 2.3c). For the 31 retained, uncorrupted lineage targets (Figure 2.1b), each bearing a unique target ID sequence in the resulting reads, we observed a high rate of capture, with ≥ 25 captured from 59% of cells, ≥ 20 from 85% of cells, and ≥ 10 from 99% (Figure 2.3b). Target capture rates were unevenly distributed across the eight batches of indexed PCR amplification, likely due to slight technical differences (**Methods**; Figure 2.4a-b). Recovery varied across the integrations as well, with each target ID recovered in a median of 80% of cells (range 50% to 93%) (Figure 2.3c), presumably due to position effect variegation and/or early karyotypic instability or large deletions associated with more frequently lost targets. Overall, these results indicate that a modified version of sci-RNA-seq can be used to efficiently recover transcriptomes alongside dozens of lineage target integrants from each of many single cells.

We next performed a series of filtration steps, removing cells with limited lineage information as well as those deemed likely to be doublets. First, cells were filtered to those with at least 10 lineage targets recovered, at least one of which was edited. In some cases, an edit could not be resolved, as more than one editing pattern seemed to exist for a given lineage target integrant (**Methods**). We termed these edits "ambiguous." Cells associated with more ambiguous than unambiguous edits, presumably doublets, were removed, as were cells with excessively high UMI counts

(**Methods**; Figure 2.4). The single cell transcriptomes and associated lineage targets of the remaining 10,234 cells were carried forward for all subsequent analyses.

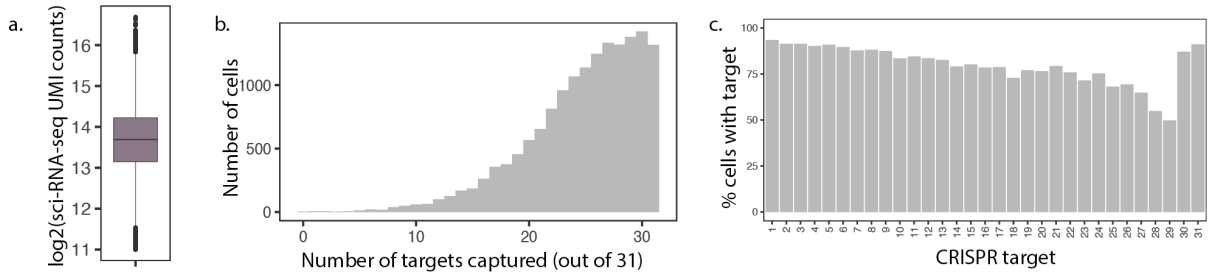


Figure 2.3. Evaluation of concurrent expression and lineage profile capture.

(**a**) Log-scaled boxplot of UMI counts for sci-RNA-seq (not including enriched target UMIs). Box shows median and encompasses counts in the second and third quartiles. Whiskers depict the interquartile range, with outliers shown. (**b**) Histogram of the number of targets captured per cell. (**c**) Percent of cells from which each individual target was captured. Targets 30 & 31 were duplicated (see text), and hence artificially appear to have a high rate of capture.

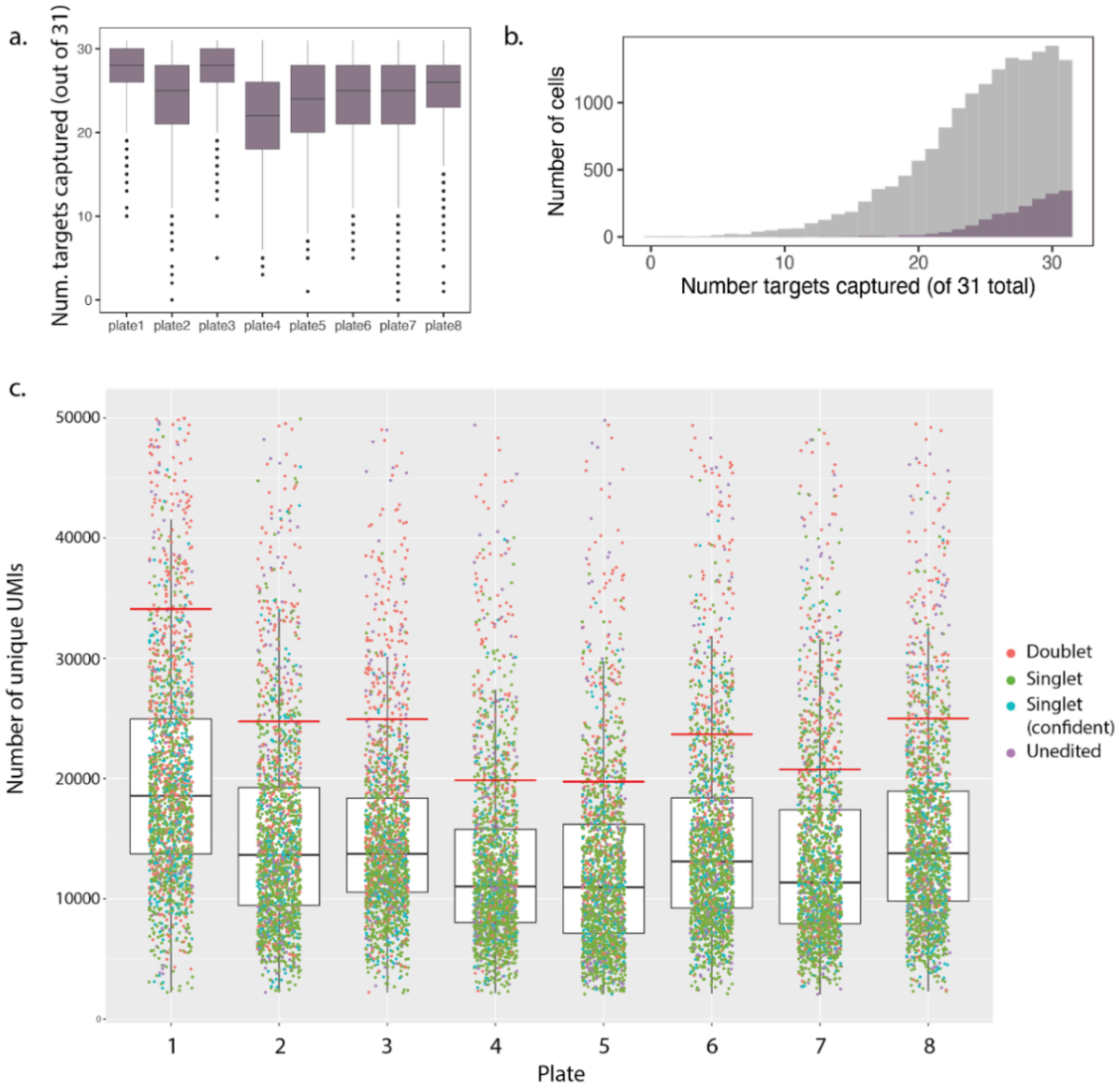


Figure 2.4. Batch-specific evaluation of target capture.

(**a**) Distribution of the number of targets captured per cell, per batch (out of 31). (**b**) Gray: Number of targets captured per cell across batches; Purple: number of targets captured per cell in batch #1. (**c**) Distribution of transcriptome UMIs per cell, per indexed PCR batch ("plate"), with UMI cut-off for doublet removal shown by red lines. Cells with UMI counts > 1.8X the median UMI count for each batch were removed from the analysis. Singlets and doublets inferred from collisions in

lineage data. “Singlet (confident)” corresponds to cells which can confidently be called as singlets based on the number of non-ambiguous editing events observed. In panels **a** & **c**, boxes show median and encompass counts in the second and third quartiles, while whiskers depict the interquartile range.

Across this entire dataset, we observed 461 unique editing patterns of the common target sequence, of which 182 were independently observed in at least 2 cells in association with the same target ID. The remainder may correspond to real events that occurred late in the expansion and were thus only sampled once, or alternatively PCR or sequencing errors. The 50 most frequently observed edits, across all cells and target IDs, are shown in Figure 2.5a. Of note, edits that recur independently as well as edits that occurred early during clonal expansion will both appear "common" by this measure. The three most frequently observed edits, together comprising 58% of all edits, appear to be recurrent: they occur in association with the majority of target IDs (Figure 2.5a), and furthermore correspond to outcomes anticipated to be favored by microhomology (Sfeir & Symington, 2015). Such frequent editing outcomes complicate tree construction, and can be avoided in the future through better target design (W. Chen et al., 2019). However, the clear majority of editing outcomes were only observed in association with a single target ID, consistent with their origination from a single event during the clonal expansion (Figure 2.5b).

Unexpectedly, two targets (#30 & #31) contained a large number of ambiguous editing calls−two distinct editing patterns convincingly present in association with the same target ID in the same single cell. This is consistent with a duplication event, *i.e.* in which the locus in which the target ID resides was duplicated early in the clonal expansion, or more likely during the second round of cloning. Additional evidence, discussed further below, of large-scale CNAs in the transcriptome data, corroborates this hypothesis. Rather than filtering out these targets, we "duplicated" them *in silico*, parsimoniously distributing the top two edits associated with these target IDs in a given cell,

while minimizing the number of independent editing events required to explain them (**Methods**). As such, in the end, single cell lineage profiles contained 33 unique targets.

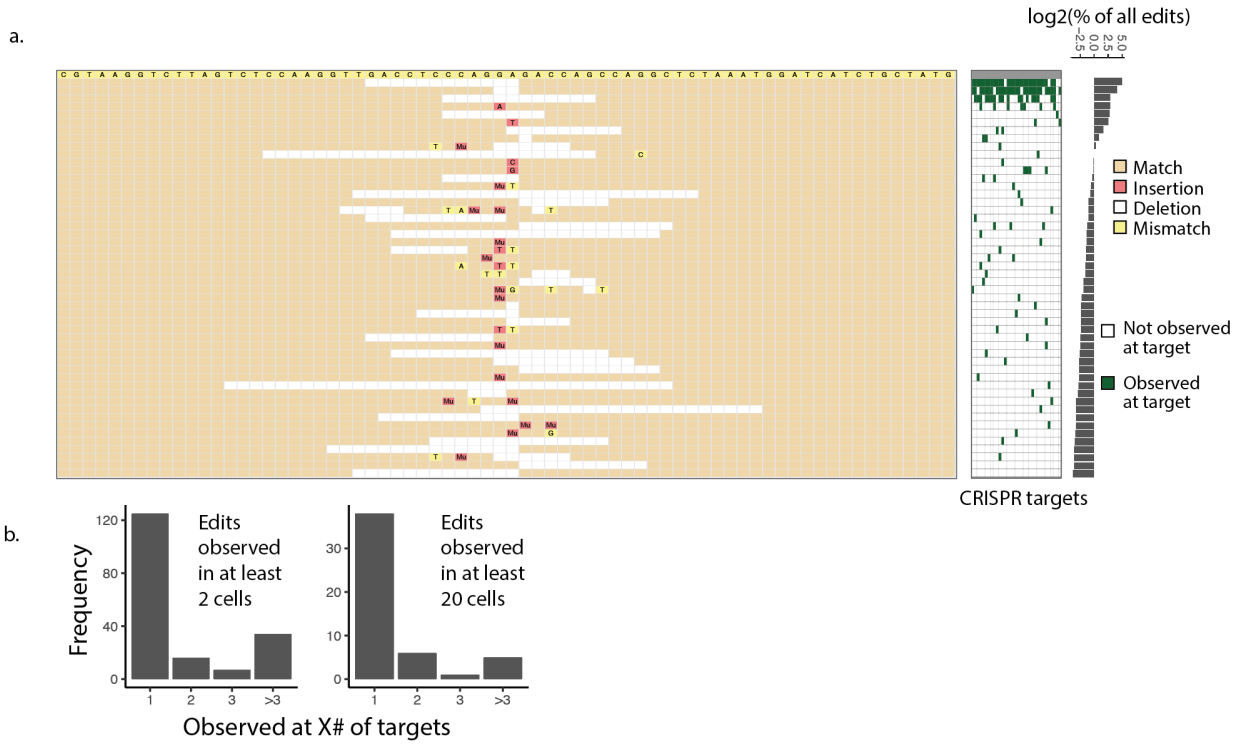


Figure 2.5. Evaluation of edit diversity.

(**a**) Left: Top 50 most abundant editing patterns. Insertions are shown one base left of the insertion site; "Mu": multi-base insertion. Middle: Targets at which the editing pattern is observed in at least 20 cells. Right: Log-scaled percentage of all edits represented by the top 50 editing patterns. (**b**) Proportion of editing patterns observed in 1, 2, 3, or more than 3 targets, if considering editing patterns appearing in at least 2 cells at a single target (left), or at least 20 cells (right).

Chapter 3. RECONSTRUCTION LINEAGE RELATIONSHIPS USING SINGLE CELL LINEAGE PROFILES

3.1 FEATURES OF LINEAGE DATA WHICH HINDER THE USE OF TRADITIONAL PHYLOGENETIC APPROACHES

The reconstruction of cell lineage trees from CRISPR-edited targets has proven to be a difficult problem (Gong et al., 2021; Salvador-Martínez et al., 2019). Although phylogenetic reconstruction methods can in principle be applied here, several factors make this practically challenging. First, the amount of information within a lineage profile is limited to the number of targets that are edited and successfully recovered; the inefficient recovery observed in most studies to date results in substantial "missing data". Second, recurrent events, *i.e.* the same edit occurring more than once independently at the same target, can be much more likely than in more conventional phylogenetic datasets, further complicating reconstruction. Third, it is computationally impractical to apply many popular phylogenetic algorithms to the large number of cells profiled with CRISPR-based lineage tracing, particularly those relying on generating a subset of all possible trees and choosing the most likely among them. To overcome this, one group employed a greedy approach to split cells into subgroups, generating subtrees of subgroups and merging them at the end (Jones et al., 2020). However, this approach was hindered by missing data in individual cell lineage profiles, which frequently split closely related cells across multiple subgroups.

On the other hand, CRISPR-based lineage tracing data has one feature which makes it more amenable to step-wise (rather than probabilistic) reconstruction strategies−the starting state of each target, *i.e.* unedited, is known. Given this, it is at least theoretically possible to employ a divisive, greedy approach described below to build a highly accurate tree (Figure 3.1c,d).

3.2 PROPOSED RECONSTRUCTION ALGORITHM

In the proposed algorithm, all cells begin as a single group, which is split into two groups based on the presence vs. absence of the most common editing pattern associated with a single target. This edit is inferred by its frequency to have occurred earlier than other edits in cells belonging to the group. This splitting step is iterated on each sub-group, and each sub-sub-group, etc., terminating when all unique lineage profiles are represented by individual branches. Subsequently, unsupported bifurcations (those wherein a branch is not defined by a specific editing event(s)) are collapsed, such that more than two branches can arise from a single inferred ancestor.

The success of this approach is dependent upon two important assumptions: erroneous or missing data is minimal, and convergence events−two or more identical edits occurring independently at a single target site—are rare. We thus set out to optimize the dataset to better fit these assumptions.

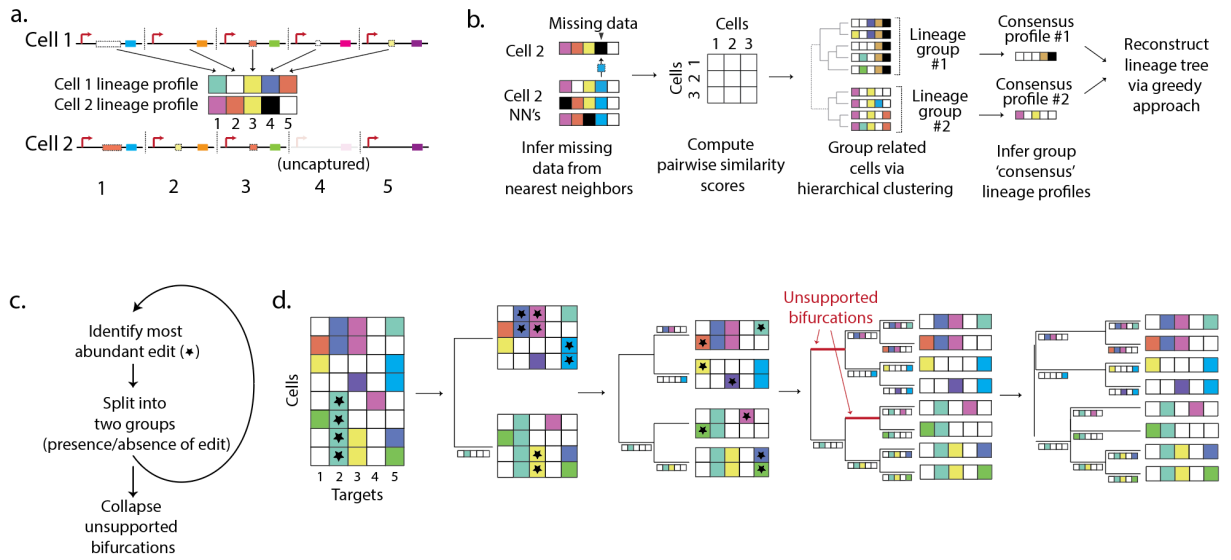


Figure 3.1. Cell lineage tree reconstruction algorithm.

(**a**) Visualization of cell lineage profiles. Each unique editing pattern is assigned a unique color.

(**b**) Preprocessing of lineage data. Missing data are imputed from nearest neighbors and pairwise

similarity scores are computed from corrected lineage profiles. Similarity scores are used to generate a hierarchically clustered tree, grouping related cells. This tree is subdivided into groups of related cells and consensus lineage profiles are generated for each lineage group. The consensus profiles are then used to reconstruct a preliminary cell lineage tree via a greedy approach. (**c,d)** Summary and example of a greedy approach to reconstruct a cell lineage tree. This greedy approach can be performed iteratively on groups of cells within a lineage group to generate a tree with individual cells at the leaves.

3.2.1 *Mitigating sequencing errors and inferring missing data*

Sources of erroneous data include PCR and sequencing errors within the target, where a single mismatch in the 70bp (unedited) amplicon would instead appear as a distinct edit. Defining edits is further complicated by the fact that an edit containing both deleted and inserted bases can appear discontinuous when aligned to the reference sequence (*e.g.* see examples within alignments shown in Figure 2.5a). To mitigate errors and misalignments, we required that an edit had to begin within 4 bases of the CRISPR cut site, and that all discontinuous segments be within a maximum of 4 bases from each other (**Methods**). To address missing data, we first defined a similarity metric between cells based on shared edits and used it to identify a set of nearest neighbors for each cell. We then imputed missing and ambiguous edits from these nearest neighbors (**Methods**). Individual cell lineage profiles for a group of closely related cells with missing and ambiguous data shown (black and red boxes, respectively) are plotted in Figure 3.2a.

3.2.2 *Mitigating molecular cross-talk between cells during single cell processing*

An additional source of error arises from cross-talk between cellular and target indices during PCR amplification, such that a target sequence derived from one cell becomes associated with the profile of another. A single such error might place a cell far from its true lineage via the algorithm described above. However, although these events are undetectable at the single cell level, they are often obvious when examining groups of closely related cells. To take advantage of this, we sought to pool closely related cells, infer a "consensus" lineage profile for each group (encompassing edits shared by the majority of the group), and generate a preliminary tree of these consensus profiles, such that cells with "contaminating" target sequences would be retained in the group via overall proximity to their neighbors. To identify groups of closely related cells, we again calculated all

pairwise similarity scores, and used these as input for hierarchical clustering using Ward's method. We visually determined the number of clusters into which to subdivide cells, using plots such as the one in Figure 3.2a (right), and computationally inferred a consensus profile for each group. In some cases, where we could explain why an edit did not reach the needed majority for inclusion, automatically inferred consensus profiles were manually corrected (**Methods**). Finally, we applied the algorithm above to the consensus profiles, generating a lineage tree of subgroups of closely related cells.

Since cells within each subgroup contain additional edits beyond the shared edits shown in the "consensus" profile, one can in theory iteratively apply this set of steps to each subgroup, and concatenate the resulting subtrees to derive a single cell-resolved lineage tree. Since our downstream intended application involved comparing pooled expression and chromatin accessibility profiles from groups of closely related cells, and we found that particularly small lineage groups were too noisy for meaningful gene expression and chromatin accessibility analysis, we performed such iterative subdivisions for only a subset of the groups.

3.3 BUILDING A LINEAGE TREE FROM OUR DATA

For several reasons, we generated an initial tree using only about a quarter of the filtered cells (n = 2,419). First, the hierarchical clustering algorithm used for initial subgrouping has $O(n^3)$ run time. Second, as described in the previous section, two out of eight batches (1 & 3, Figure 2.4) exhibited the most complete lineage profiles, and we reasoned that these would generate the most accurate cell lineage groups into which the remaining cells could be placed via a nearest neighbors approach. Provided that the terminal lineage groups we generate are large enough, we

can assume close cell relatives of every cell in the dataset are present within this subset of the overall data. Including all cells, the final tree used for downstream analyses contained 42 lineage groups, ranging in size from 34 to 1217 cells (Figure 3.2).

This iterative approach of building and concatenating subtrees from root to tip mitigates the probability that recurrent editing patterns at individual targets grossly impact tree structure. For example, if the same edit occurred in two cells independently at target #2, and if one of these events occurred early enough to define an early bifurcation, all descendants of the other cell would be misplaced early during tree reconstruction when employing a greedy approach. However, initial subgrouping of cells based on the full set of edits they contain prevents this problem when at least one of the edits occurs late enough that it does not define the group as part of its "consensus" lineage profile.

Nevertheless, CNAs inferred from expression data occurring over the course of this experiment (discussed in detail in the next section) signaled the presence of two convergence events within lineage data impacting our tree structure. In each case, the convergence events were mediated by a very common editing pattern (Figure 2.5a), and we manually resolved these events to come to the tree structure shown in Figure 3.2 (**Methods**). However, it should be emphasized that with the exception of these two manual changes, the tree shown in Figure 3.2 was reconstructed solely from lineage profiles, *i.e.* expression data was not used for lineage inference.

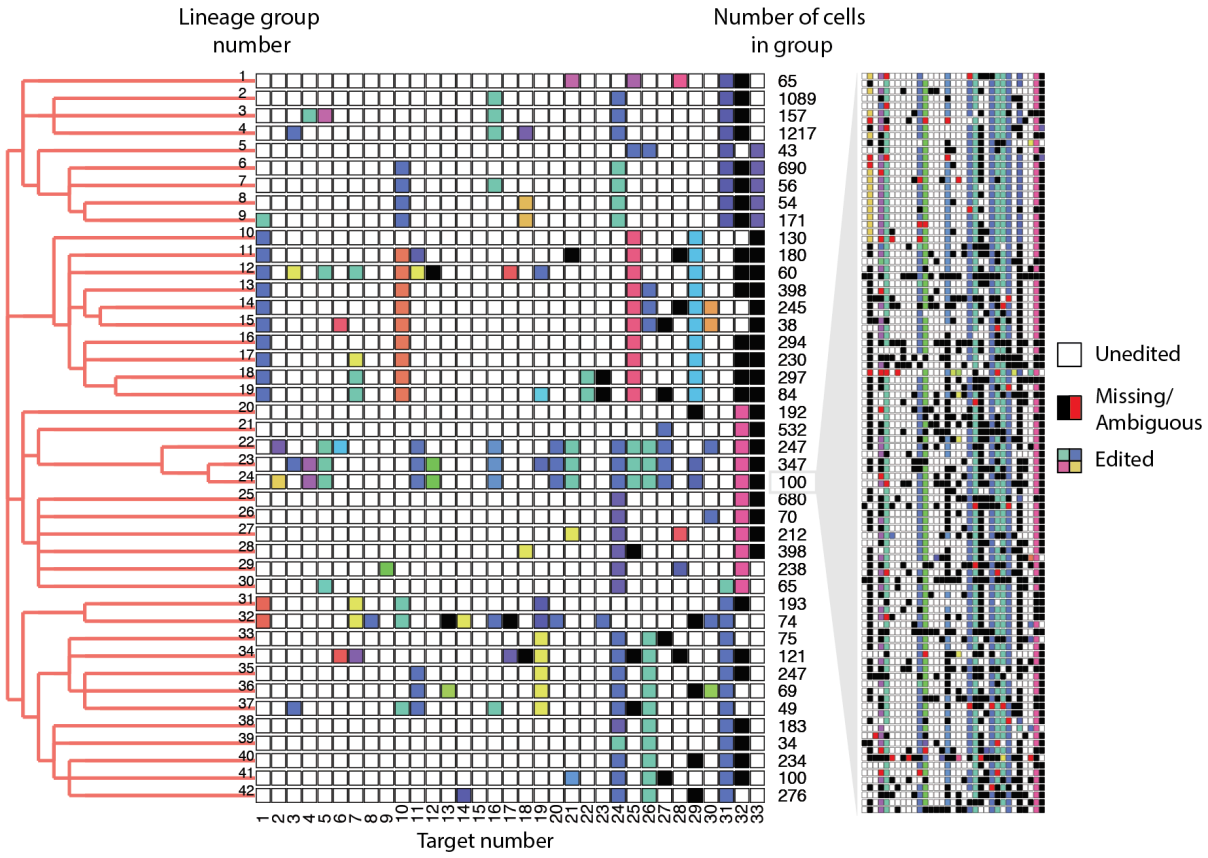


Figure 3.2. Reconstructed lineage tree.

Left: Tree of cell lineage groups ("consensus" editing patterns shown as rows; each column represents a unique target site). Each color represents a unique editing pattern. White: unedited target. Black: targets with missing data for a majority of cells in the group. Number of cells represented by each consensus cell is shown. Inset (right) shows the editing patterns for all 100 cells assigned to lineage group #24. Black: missing targets. Red: ambiguous targets.

Chapter 4. EVALUATING LINEAGE-ASSOCIATED DIFFERENTIAL EXPRESSION

4.1 CHROMOSOME COPY NUMBER ALTERATIONS INFERRED FROM SCI-RNA-SEQ RECAPITULATE THE LINEAGE-INFERRED TREE STRUCTURE

We reasoned that heritable variation in gene expression patterns should visually correlate with tree structure, whereas non-heritable variation should not (Figure 1.1a). To explore this, we aggregated single cell expression profiles within each of the 42 groups described above, and plotted relative group expression as a heatmap (Figure 4.1). Unexpectedly, when genes were arranged by their genomic location, we observed large, continuous stretches of down- or upregulated genes, strong evidence of partial or full chromosomal gain or loss events. HEK293s are pseudotriploid and known to be karyotypically unstable, and an active CRISPR/Cas9 system may also contribute to instability (Y.-C. Lin et al., 2014).

As CNAs are themselves heritable genomic events, we saw an opportunity to use them to validate our CRISPR-inferred tree structure. Strikingly, where present, CNAs were generally concordant with the tree structure inferred from lineage data. In particular, with the exception of full chromosome gains or losses, most CNAs appear to have arisen from a single founder event (Figure 4.1). As described in the previous section and **Methods**, on two occasions, CNAs were used to resolve ambiguity in the lineage data due to convergence events. However, the remaining CNAs shown in Figure 4.1 were not used for lineage reconstruction and, importantly, we observed no instances of CNAs contradicting CRISPR-derived lineage relationships.

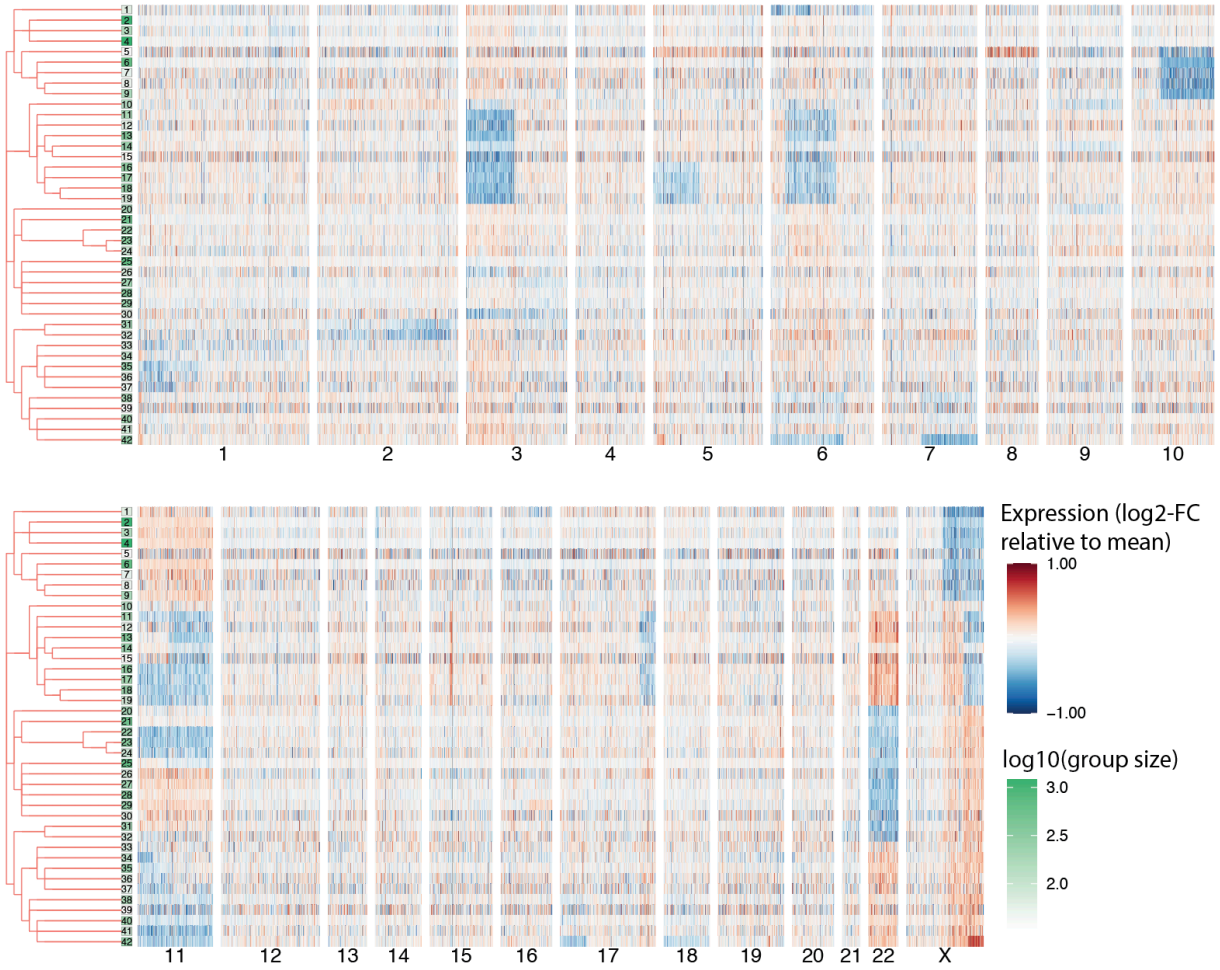


Figure 4.1. Gene expression in lineage groups arranged by genomic location. Heatmap shows log2-fold gene expression variation relative to the mean expression of each gene across cells. Genes are shown in the order in which they appear along chromosomes in the reference human genome. Log2 fold changes >1 & -1 were manually fixed at these maximum and minimum values for visualization. A minimum mean expression cutoff was applied to remove lowly-expressed genes, leaving 6,241 genes. Green shading of the boxes containing lineage group numbers at the tree leaves is based on the log-scale number of cells per group.

4.2 ALLELIC RATIOS FURTHER INFORM CHROMOSOME COPY NUMBER DYNAMICS ACROSS LINEAGES

4.2.1 *A method to infer allelic ratios from sci-RNA-seq data*

We next wondered whether we could use lineage-resolved expression data to investigate allele-specific copy number dynamics. Indeed, although we made no direct measurement of copy number, we found that in many cases we could infer copy number based on SNP ratios in sci-RNA-seq data (Figure 4.2a). For example, if a chromosome shows heterozygosity at known SNPs, and we observe allelic ratios of 1:2 across these positions, this chromosome is likely to be present in three copies, while a 1:1 allelic ratio would suggest two or four copies, and a 1:3 allelic ratio would suggest four copies. On the other hand, a paucity of SNPs would suggest regional or chromosome-wide loss-of-heterozygosity, in which case copy number could not be inferred by this method.

We first performed such an analysis on each chromosome using expression data from all cells. Since each genomic position is represented sparsely in sc-RNA-seq data, we divided the genome into 5Mb bins, identified coordinates which appeared to be heterozygous in our data (most frequent base present at in <85% of reads), subsetted these to include only those positions which overlapped known human SNPs (*i.e.* those appearing in dbSNP), and combined counts for SNPs within each 5Mb bin. For this last step, because phasing information was not available, we simply assumed the more abundant alleles at each SNP within a bin were on the same haplotype for binning purposes (as would be expected if homologs existed in unbalanced ratios, at least provided counts are sufficiently high). We then calculated a "major" (most abundant) allele frequency for each bin and plotted these by relative genomic position (Figure 4.2a,b). Figure 4.2a shows several examples of

this approach for chromosomes with stable copy number in our dataset, revealing there to be 3 copies of chr19, 4 copies of chr18, and 2 or 4 copies of chr17. Of note, because our heuristic always places the most abundant allele on the same haplotype, we expect a major allele frequency above 1/2 for cases where haplotypes exist in equal copies, *e.g.* as we infer for chr17. On the other hand, chr14 exhibited very low overall heterozygosity at known SNPs together with an unstable ratio, suggesting loss-of-heterozygosity. Consistent with this prediction, the "minor" alleles inferred in chr14 and other chromosomes which exhibit this unstable pattern (Figure 4.3a) often do not match known variants founds in the human population , in contrast with inferred minor alleles in chromosomes exhibiting heterozygosity (Figure 4.3b). Major allele frequency plots for all chromosomes are shown in Figure 4.3a.

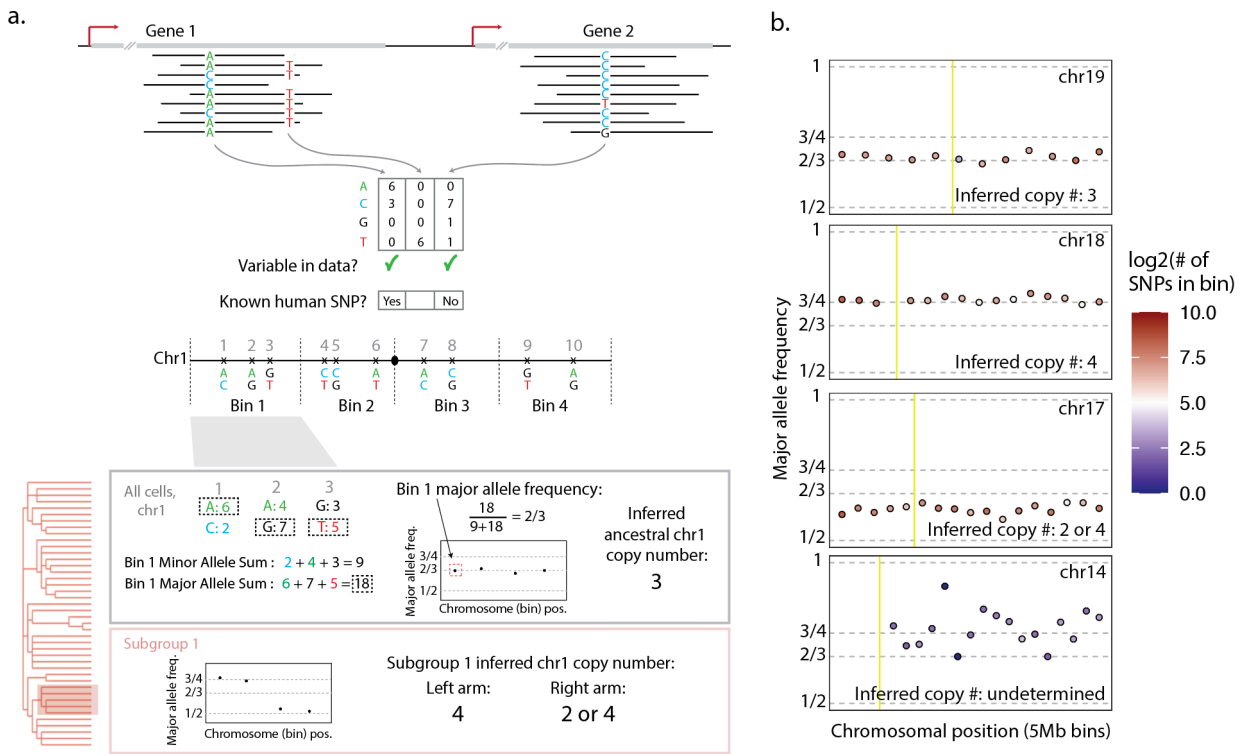


Figure 4.2. Inferring copy number using lineage-resolved allelic ratios.

.

(**a**) A strategy to infer copy number using SNPs from sc-RNA-seq data. First, haplotypic imbalance is assumed and haplotypes are inferred based on base abundance at known SNPs, using all cells. We can then use these to infer the ancestral (or most observed) copy number. Using these haplotypes, we can perform this analysis on subsets of the tree to infer whole or partial chromosome gains or losses. (**b**) Copy number analysis described in panel **a** for chr19, chr18, chr17, & chr14, using all cells. Point fill color represents the number of SNPs found to be heterozygous in that bin, signaling the reliability of this analysis at that location. Yellow line shows the centromere position.

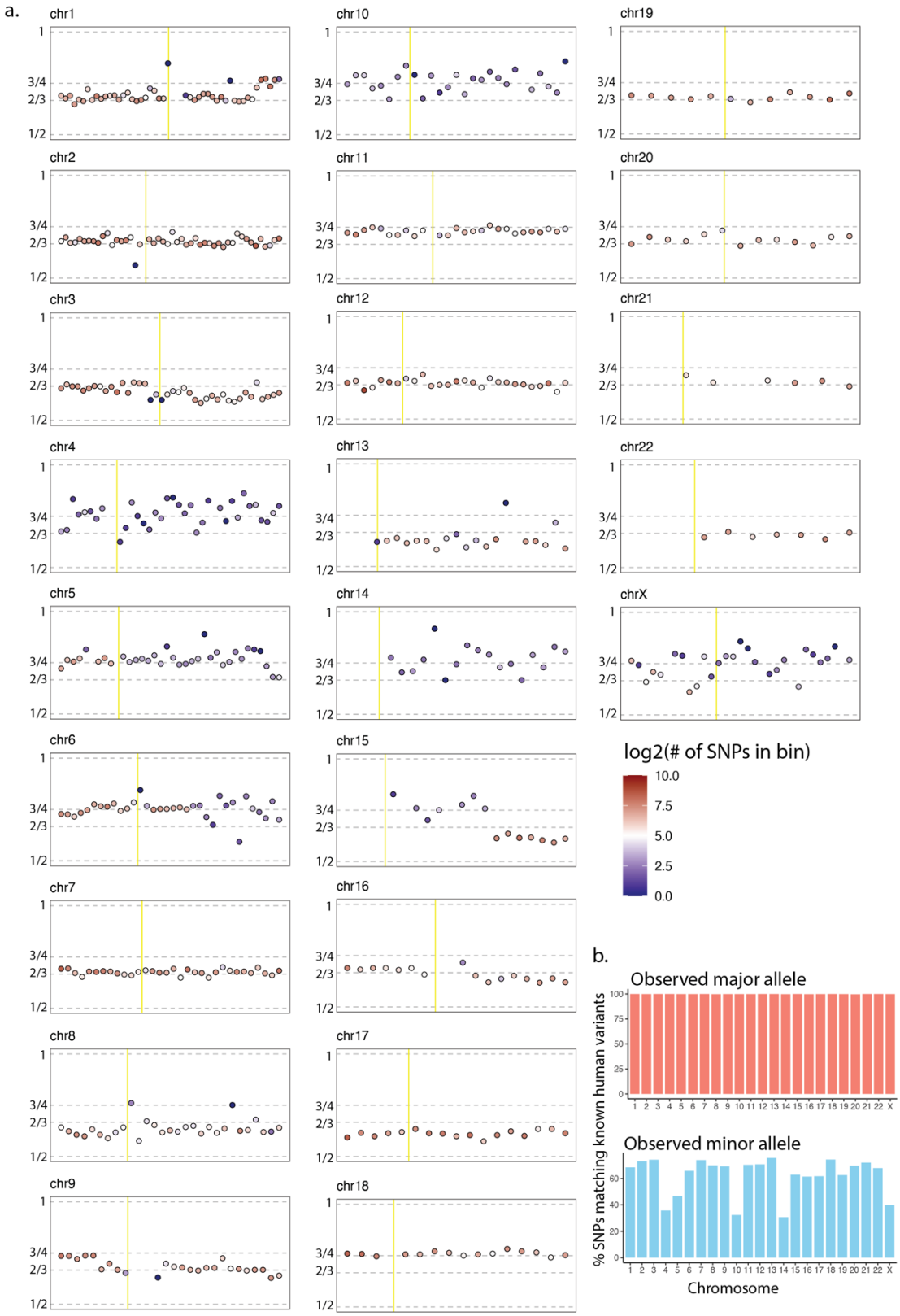


Figure 4.3. Allelic-ratio-based copy number analysis for all chromosomes.

(**a**) Analysis described in Figure 5a-b, performed on all cells for all chromosomes. Point fill color represents the number of SNPs found to be heterozygous in that bin, signaling the reliability of this analysis at that location. Yellow line indicates the centromere position. (**b**) Percent of inferred major and minor alleles at variable positions in the data (filtered as described in Figure 4.2**a**) which match SNP bases found in humans at those positions (dbSNPs). For simplicity, only single-base SNPs with at most two common alleles in the population were considered.

4.2.2 *Inferring lineage-informed allele dynamics*

We next applied this approach to subgroups of the tree to investigate copy number dynamics during the monoclonal expansion. For example, this analysis revealed a partial loss of an extra copy of the short arm of chr3 impacting only a subgroup of related cells (Figure 4.4a, left panel). Of note, the inferred breakpoint is slightly shifted from the centromere, such that several genes on the short arm are retained. We calculated a binned major allele frequency for the subgroups indicated in Figure 4.4a (left panel), using the major haplotypes we inferred from all cells (Figure 4.4a, right panel). Subgroup copy number analysis (Figure 4.4a, right panel) of groups 1-9 (top, purple) agrees with the predicted ancestral state, whereas the major allele frequency in groups 10-19 has dropped between 1/2 & 2/3 across the whole chromosome. Since heterozygosity appears preserved on the left arm, we infer that the partial chromosome (*i.e.* a copy of the short arm of chr3) was lost in groups 10-19, relative to the ancestral state.

A similar analysis suggested more complex copy number dynamics for chr11, for which multiple full and partial chromosome copy number changes appear to occur at different parts of the lineage (Figure 4.4b, left panel). Performing a subgroup analysis, we observe a pattern consistent with at least three independent full chromosomal losses (Figure 4.4b, middle panel). Intriguingly, these result in different allelic ratios, with loss-of-heterozygosity in two groups (Figure 4.4b, green & blue), and maintained heterozygosity in one (beige). Overall, these analyses highlight the potential of high-resolution, progressive lineage histories to disambiguate copy number alterations, including but not limited to recurrent gains and losses.

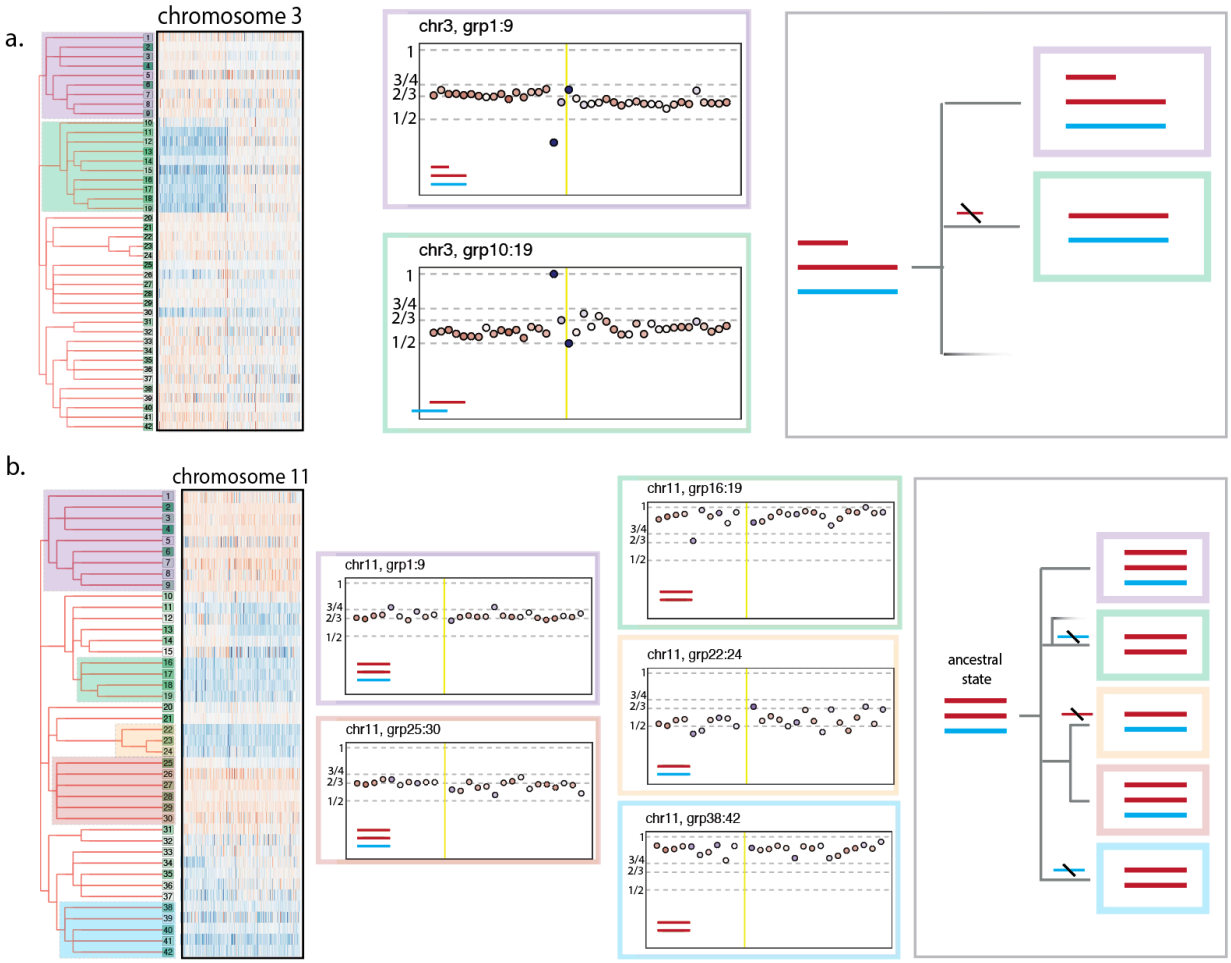


Figure 4.4. Lineage-resolved allelic ratios inform complex chromosome copy number dynamics.

(**a**) Subgroup copy number analysis of chr3. Left: expression heatmap as described in Figure 4.1. Middle: Copy number analysis of chr3 for indicated subgroups. Right: schematic of inferred haplotype dynamics. Point fill color represents the number of observed heterozygous SNPs per bin detected when pooling all cells, not just subgroup cells. Yellow line shows the centromere position. (**b**) Subgroup copy number analysis of chr11. Left: expression heatmap as described in Figure 4. Middle: Copy number analysis of chr11 for indicated subgroups. Right: Schematic of inferred

haplotype dynamics. Point fill color represents the number of observed heterozygous SNPs per bin detected when pooling all cells, not just subgroup cells. Yellow line shows the centromere position.

4.3 HERITABLE EXPRESSION CHANGES UNEXPLAINED BY CNAS ARE OBSERVED THROUGHOUT THE TREE

Within genomic regions exhibiting large-scale CNAs, copy number change is the obvious mechanism for differential expression of genes in the impacted region. But other phenomena−−*e.g.* epigenetic changes, changes in the levels of upstream regulators, focal CNAs and translocations−might induce heritable expression changes as well. To explore contributions from such sources, we set out to systematically identify examples of heritable expression variation across the tree that were not obviously explained by CNAs.

4.3.1 *A permutation-based approach for differential expression analysis*

To this end, we first inferred the boundaries of CNA events between every pair of sister branches (defined as those that share an immediate common ancestor in the tree) using a combination of expression heatmaps (as shown in Figure 4.1, Figure 4.8a), and pairwise log-fold change plots, where stretches of differential expressed (DE) genes are visible (Figure 4.6d; Figure 4.9). We then sought to evaluate DE between every pair of sister branches, using DE within CNAs as ground truth for sensitivity. Applying DEseq2, which models data as a negative binomial distribution, we observed a substantial number of false negatives−genes within CNAs which were not detected as DE−even between large groups of cells (Figure 4.6a, top panel; Figure 4.7a). We thus sought to develop a strategy which would be sensitive to small-magnitude expression changes, while also being robust to large differences in the number of cells between the groups being compared (Figure 4.5a; **Methods**). As a first step, cells from each pair of sister branches are permuted 10,000 times, in each instance creating two groups of the original sizes. For each permuted set, we calculate the log2-fold change for each gene. We then use permuted expression ratios to (a) generate an expected

distribution which we can use to calculate a z-score associated with the observed fold change; and (b) rank against the observed expression ratio to assign significance. For a set of genes evaluated for a pair of groups, if none are significantly DE, the distribution of observed ranks is expected to be uniformly distributed; on the other hand, if there are DE genes, we expect to observe their enrichment at the extremes of the rank list. Using an FDR of 5%, we can calculate a set of "significant" ranks (and thus genes) for each pair of groups being compared.

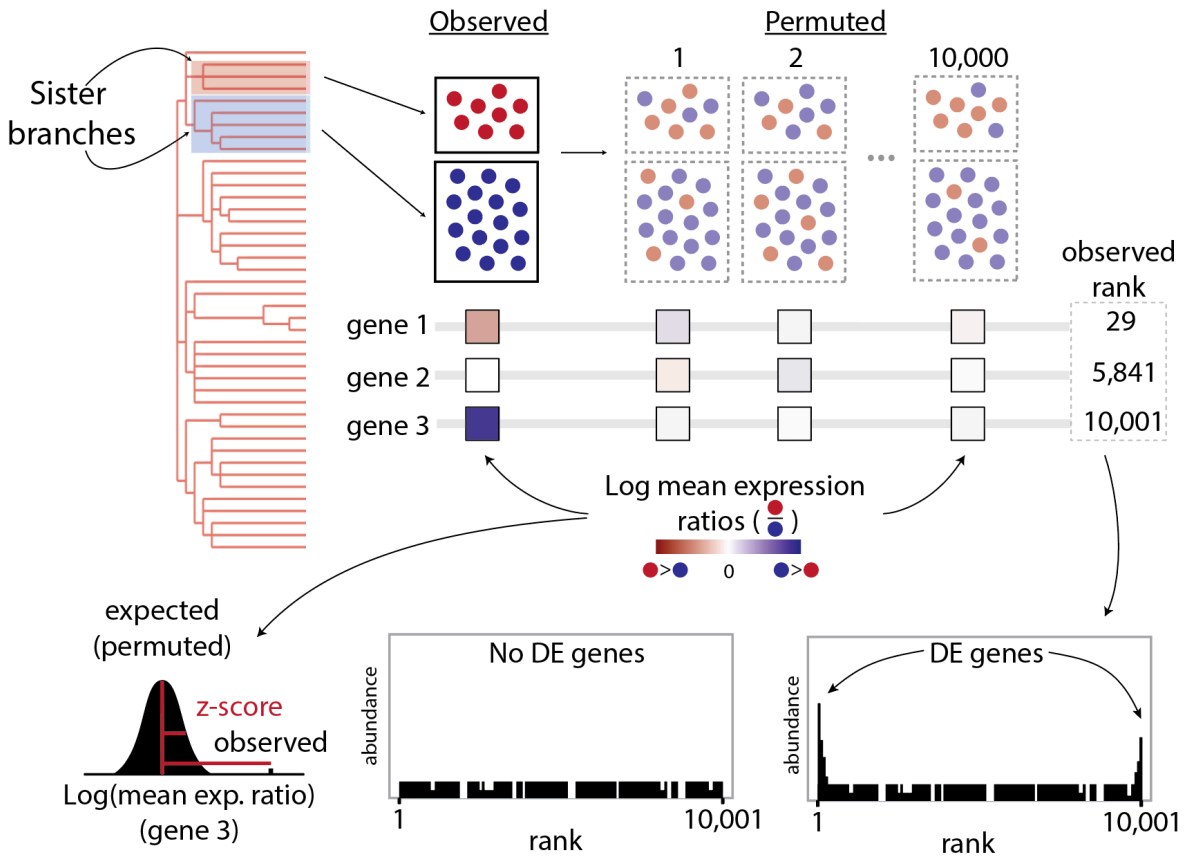


Figure 4.5. A permutation-based approach for detecting heritable differential expression within lineage-resolved sci-RNA-seq data

.

4.3.2 *Evaluating the permutation-based approach to detect differential expression across the lineage tree*

This permutation strategy detected a substantial fraction of genes within CNA regions as differentially expressed (Figure 4.6a,b; Figure 4.10). Genes within CNAs across all pairwise comparisons were more likely to be identified by our approach, with lowly-expressed genes within CNAs more likely to be missed by DESeq2 (Figure 4.7a; Figure 4.10). For example, between groups A & B, 85% of expressed genes (see **Methods** for filtering criteria) within the CNA region on chromosome 3 were identified as DE using our approach, compared with 49% detected by DESeq2 (Figure 4.7a; Figure 4.10). Unless otherwise stated, here we will refer to DE genes as those identified by the permutation approach at an FDR of 5%.

As expected, statistical power decreases with group size, but we nonetheless detected some DE genes within CNAs even between smaller groups (Figure 4.6**d**; comparisons G/H; J/K). For example, between group J & K (as labeled in Figure 4.6**d**), containing 234 and 276 cells, respectively, we detect a subset of CNA-associated genes across several chromosomes (Figure 4.9**a**), including *TRIO, SRPK2, & FGF13* (log2-fold changes of -.22, .36, & -.59, respectively). The allelic chromosome copy number analysis presented in Figure 4.2 suggests a copy number change from 4 to 5 on chr5 (*TRIO*) & from 3 to 2 on chr7 (*SRPK2*) between these two groups. Since no heterozygosity is observed on chrX, and thus we cannot infer absolute copy number change for *FGF13*.

In total, across 66 pairwise comparisons, we detected 11,454 DE genes using the permutation approach. Of these, 4,810 (42%) were detected using DESeq2, which detected an additional 520 genes not detected by our approach (Figure 4.6**b**; Figure 4.10). Surprisingly, 48% of DE genes

detected by permutation analysis could not be directly explained by large-scale CNAs (Figure 4.10). The heritable nature of these expression changes may be a product of smaller scale copy number changes , focal genetic or epigenetic differences, or *trans*-effects mediated by heritable events elsewhere in the genome (*e.g.* CNAs or other). Interestingly, when quantified by sister branch pair comparisons, the number of DE genes that we detected outside CNA regions was well correlated with the number of genes within CNAs (Pearson's r of log-transformed numbers of genes within vs. outside of CNAs = .90, Figure 4.6**c**), suggesting CNA-mediated expression changes might contribute to heritable gene expression variation through *trans*-acting effects. However, this relationship may largely be explained by the increased statistical power to detect DE genes in larger groups (Pearson's r of log-transformed number of genes outside of CNAs vs. group size = .76, Figure 4.6**c**; Figure 4.7**b**).

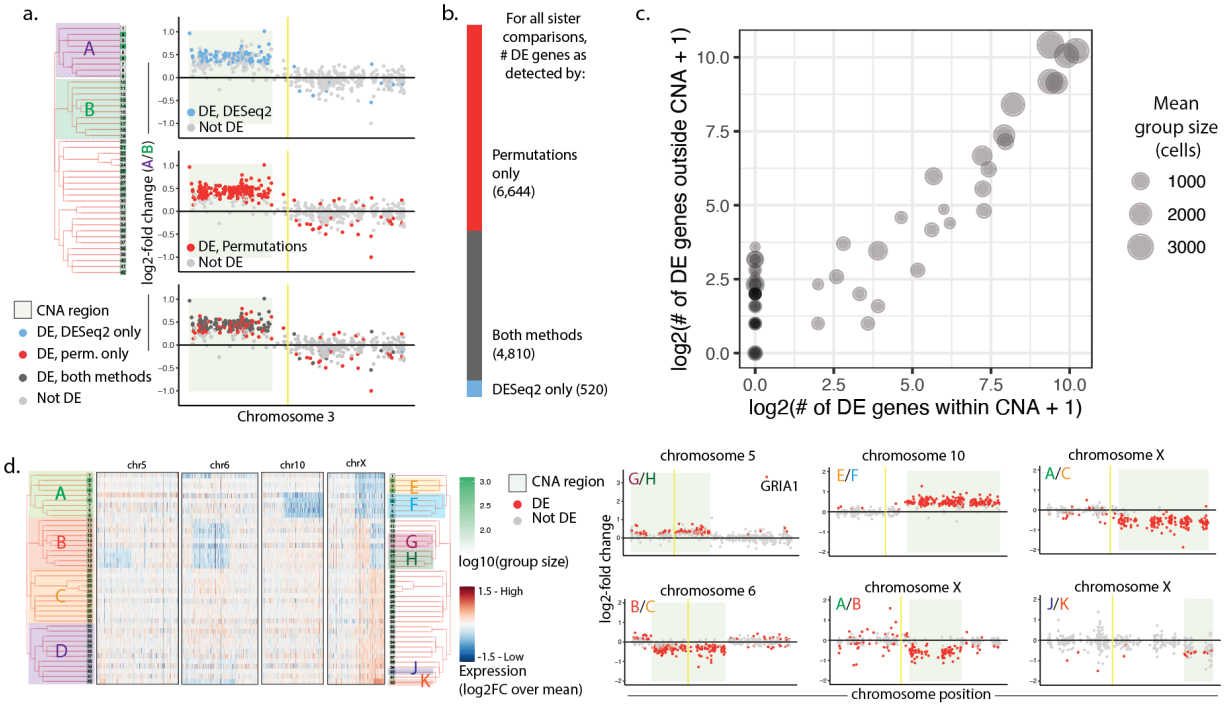


Figure 4.6. Evaluating permutation approach for detecting heritable differential expression within lineage-resolved sci-RNA-seq data

(**a**) Comparison of DE genes identified by the permutation method and/or DESeq2, showing log2-fold change expression on chr3 between indicated groups A & B. Yellow bar indicates centromere position. (**b**) Relationship between the log-scale number of detected DE genes within CNAs and DE gene falling in non-CNA regions per each sister pair comparison. Size of points represents the mean number of cells in the sister pair. (**c**) Number of DE genes identified using permutations, DESeq2, or both, across every pairwise comparison (66 total) of sister lineage groups (*i.e.* branches sharing an immediate common ancestor in the tree). (**d**) Left: Heatmaps as described in Figure 4.1 depicting CNAs on chrs 5,6,10, & X, with lineage groups indicated on tree. Right: Log2-fold changes of genes on indicated chromosomes between indicated groups, depicting the power to detect DE genes within CNA regions via the permutation approach across groups of different sizes.

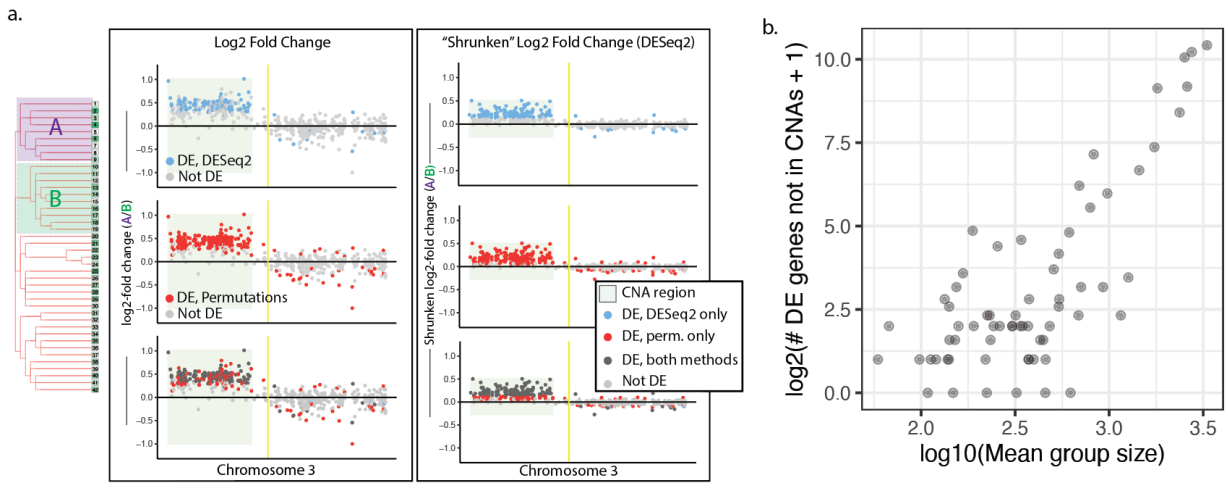


Figure 4.7. Evaluating permutation approach relative to DESeq2 and the contribution of group size to DE gene detection efficiency

(**a**) DE genes detected by the permutation approach vs. DESeq2. The left plots show log2-fold changes, while the right plots show the "shrunken" log2-fold changes calculated by DESeq2, which takes absolute expression level into account, and corrects for higher variance at low expression levels. (**b**) Relationship between group size (mean of the two groups being compared) and DE genes not associated with a CNA.

4.3.3 *Lineage-associated differential expression which cannot be explained by copy number changes: notable examples*

The most striking heritable expression change which cannot be explained by an obvious CNA was observed in *GRIA1*, a glutamate receptor subunit on chr5 (Figure 4.8, z-score = 28.2, log2 fold-change (FC) = 3.32, between the indicated groups). Markedly elevated expression is observed in lineage groups 11-15 relative to the rest of the tree (with elevated expression in group 16 likely due to misplaced cells). Though we cannot conclusively determine from this data alone whether this expression change is caused by genetic (e.g. focal amplification) or epigenetic factors, it is notable that *GRIA1* is located in a replication transition zone in various cell lines, potentially predisposing it to structural instability (Watanabe et al., 2014). Additional examples of genes exhibiting differential gene expression patterns that track closely with the lineage-derived tree structure appear throughout the tree (Figure 4.9b).

Another intriguing example, where multiple expression levels appear to have been stably inherited is observed in *CSMD3*, on chr8 (Figure 4.8). Group B expression is markedly elevated over its sister group A (A/B z-score = -7.30, log2FC = -0.57), while in the branch encompassing both groups A & B, *CSMD3* is even more highly expressed relative to group C (A&B/C z-score = 27.8, log2FC = 2.02). A weaker, but similarly heritable relationship appears between groups D & E (z-score = 3.8, log2FC = 0.34). Such a heritable but labile expression pattern might indicate flexible but relatively stable regulation at this locus. Interestingly, such graded but clone-specific expression patterns were observed with cell type groups in both *Apoe* and *Lmo4* in mouse neurons (Mold et al., 2022). Alternatively, this lability might be explained by local genomic instability. In fact, translocations at a breakpoint near *CSMD3* have been associated with autism in multiple *de novo* cases (Floris et al., 2008), and the *CSMD3* locus is implicated in a wide range of diseases

including epilepsy & non-small cell lung carcinoma (Floris et al., 2008; P. Liu et al., 2012; Shimizu et al., 2003). CNAs are particularly common in branch C (Figure 4.1), bolstering the likelihood that a translocation event explains reduced expression in that group.

Even within CNAs, we observe single gene expression changes which deviate strongly from the expected copy number ratios. An intriguing example is the transcript *AC090518.1*, which normally exhibits testis-specific expression, and is located within a short stretch of genes with modestly elevated expression on chr15 consistent with a CNA (Figure 4.8b,c; *AC090518.1* is located between *MNS1* & *ZNF280D*). This transcript's markedly increased expression well beyond that of its neighbors (log2-fold change (A/B) = -3.82, A/B z-score = -28.67), points to a possible translocation (or tandem duplication) event, exposing it to a new regulatory context. Chromosomal rearrangements are a hallmark of cancer progression, and tracking such small-scale events may reveal the mechanism behind biologically-meaningful expression changes. The genes *GNGT1*, *C6orf14*, and *NEAT1*, all lie within CNA regions but show heritable expression changes in the opposite direction of surrounding genes (z-scores -7.40, -4.10, -8.84, respectively, Figure 4.8**c**). Such patterns may indicate expression compensation or selection for particular expression levels. In fact, both *GNGT1* & *C6orf141* have been associated with cancer prognosis (C.-M. Yang et al., 2019; J.-J. Zhang et al., 2021), with *C6orf141* playing a direct role in cell proliferation. *GNGT1* was designated a hub gene in non-small-cell lung cancer, suggesting its misexpression may have widespread downstream consequences which would also appear heritable. *NEAT1*, a long non-coding RNA with a known epigenetic role in a variety of cell types, may also stably modify expression of multiple downstream target genes (Wang et al., 2020).

Here, lineage relationships enabled us to identify stably-inherited expression changes which may not otherwise be obvious among non-heritable expression fluctuations. In most cases, however, it is not possible with this data alone to determine the mechanistic basis for this differential gene expression (*e.g. cis*-genetic, *trans*-genetic vs. epigenetic). We next sought to distinguish between these possibilities by additionally tethering chromatin accessibility information to this same lineage tree.

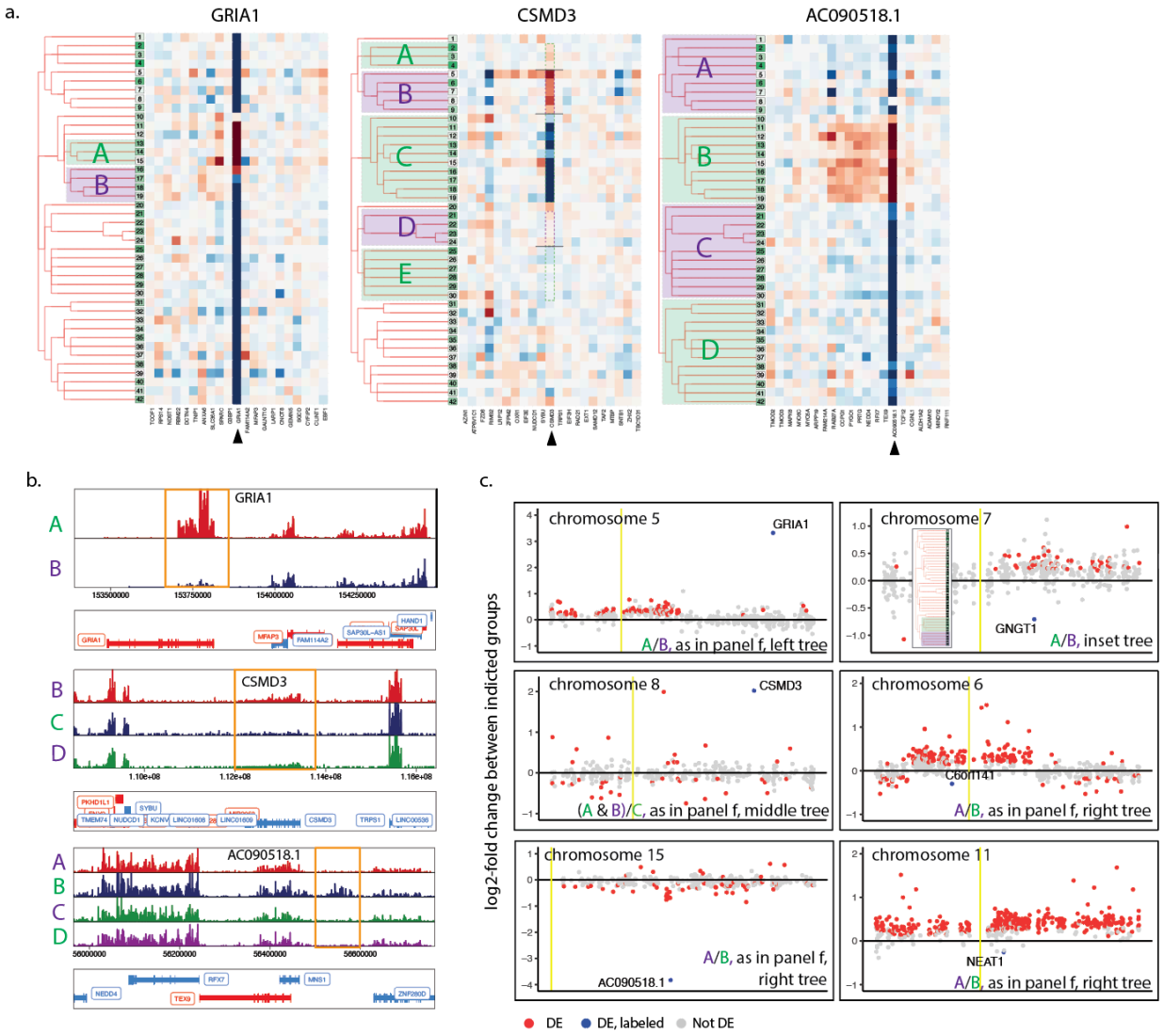


Figure 4.8. Non CNA-mediated heritable differential expression: notable examples

(**a**) Heatmaps showing DE expression of *GRIA1*, *CSMD3*, *AC090518.1,* and surrounding genes. (**b**) Pileup visualizations of *GRIA1*, *CSMD3*, *AC090518.1* in groups indicated on the trees in panel **a.** *AC090518.1* is positioned between *MNS1* & *ZNF280D*. (**c**) DE genes showing heritable expression patterns which cannot be explained by detected CNAs. The pair of groups being compared for each plot is indicated on the bottom right, with groups indicated on the trees in panel **a** (except for top-right sub-panel, for which pair of groups is shown in inset tree).

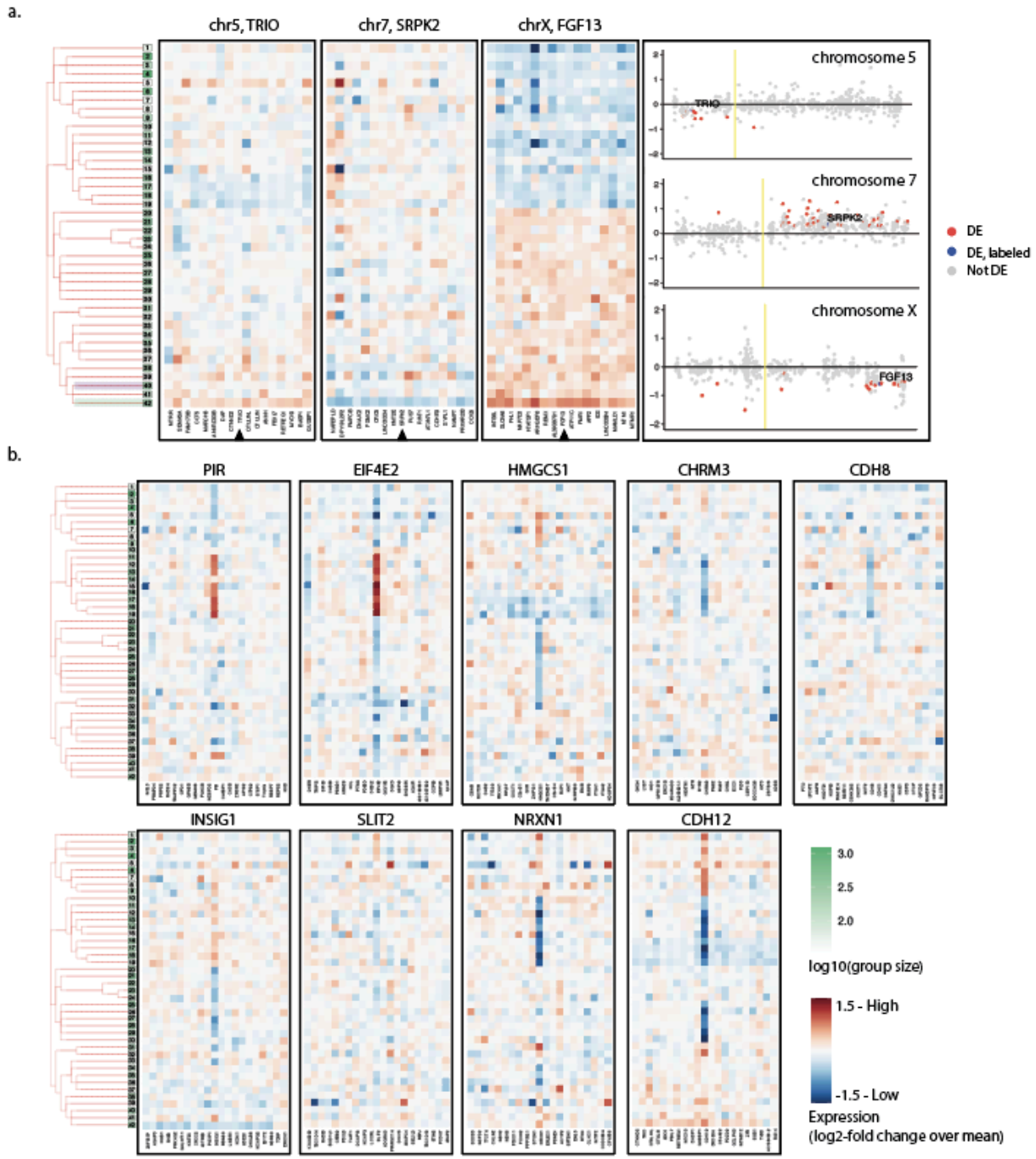


Figure 4.9 Differentially expressed genes within and outside of detected CNAs observed across sister lineage group comparisons.

(**a**) DE genes detected within CNA regions on chrs 5, 7, and X, between the indicated groups (234 and 276 cells, respectively). (**b**) Heatmaps showing single genes (middle of each plot) which

exhibit heritable expression patterns consistent with the tree structure. Surrounding genes are not DE, suggesting these patterns are not due to CNAs, although we cannot rule out highly focal amplifications with gene expression data alone.

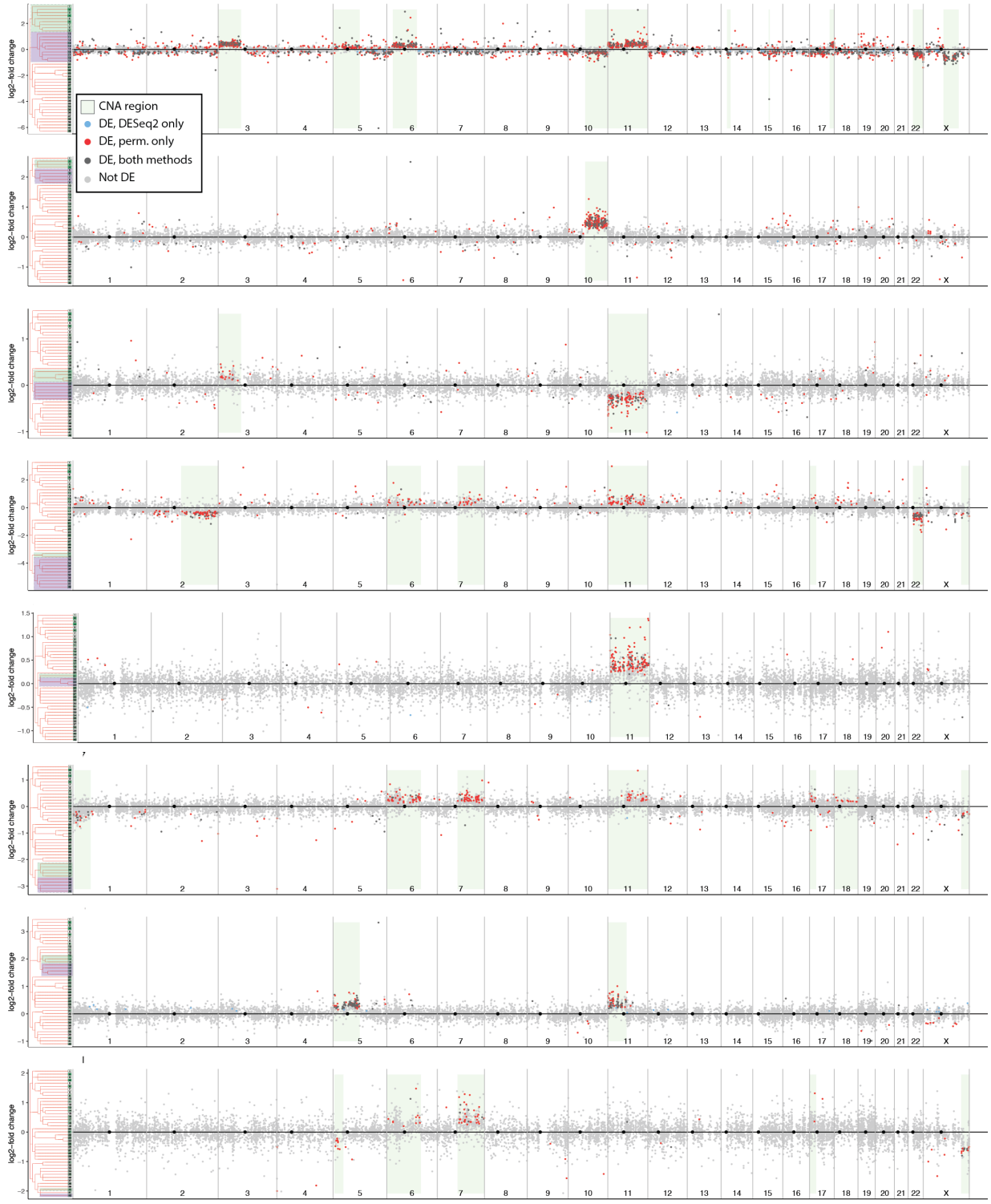


Figure 4.10 Global DE between select pairs of sister groups

Log2-fold change for expressed genes across all chromosomes between select pairs of sister lineage groups. Groups that are compared in each plot are indicated on the trees at the left with

green and purple boxes. Colors indicate by which method (if any) a gene was found to be differentially expressed. Inferred CNAs are shown as light green boxes.

Chapter 5. COLLECTING LINEAGE DATA ALONGSIDE SINGLE CELL CHROMATIN ACCESSIBILITY TO EVALUATE SOURCES OF HERITABLE DIFFERENTIAL EXPRESSION

5.1 COLLECTING LINEAGE INFORMATION ALONGSIDE SINGLE CELL CHROMATIN ACCESSIBILITY PROFILES ENABLES TETHERING OF GENE EXPRESSION AND CHROMATIN ACCESSIBILITY

5.1.1 *A novel method for collecting lineage information alongside single cell chromatin accessibility profiles*

Both genetic and epigenetic phenomena can potentially underlie what we observe as heritable expression changes, and measuring expression alone is often not sufficient to disentangle these from one another. Coassays of single cell expression and chromatin accessibility may provide more insight, but contemporary methods result in relatively sparse profiling in any given cell. However, since heritable states are presumably shared by cells with similar lineage histories, we can theoretically measure these features independently in clonally related cells and link them retrospectively based on lineage relationships (Figure 5.2a). Furthermore, pooling of single cell chromatin accessibility profiles of closely related cells, as we did with expression profiles, increases the power to detect changes. To this end, we developed a method to capture lineage-associated transcripts alongside sci-ATAC-seq (Cusanovich et al., 2015, 2018), *i.e.* to concurrently profile single cell lineage relationships and chromatin accessibility states (Figure 5.1). sci-ATAC-seq is a pool-split approach where genetic material undergoes two rounds of molecular indexing, such that DNA from each cell is ultimately associated with a unique pair of indexes. To associate

lineage information with sci-ATAC-seq profiles, we devised a strategy to concurrently index mRNA transcripts containing recorded lineage information at each sci-ATAC-seq indexing round, via reverse transcription and PCR, such that both features can be retroactively linked to a single cell via index combinations (Figure 5.1; **Methods**).

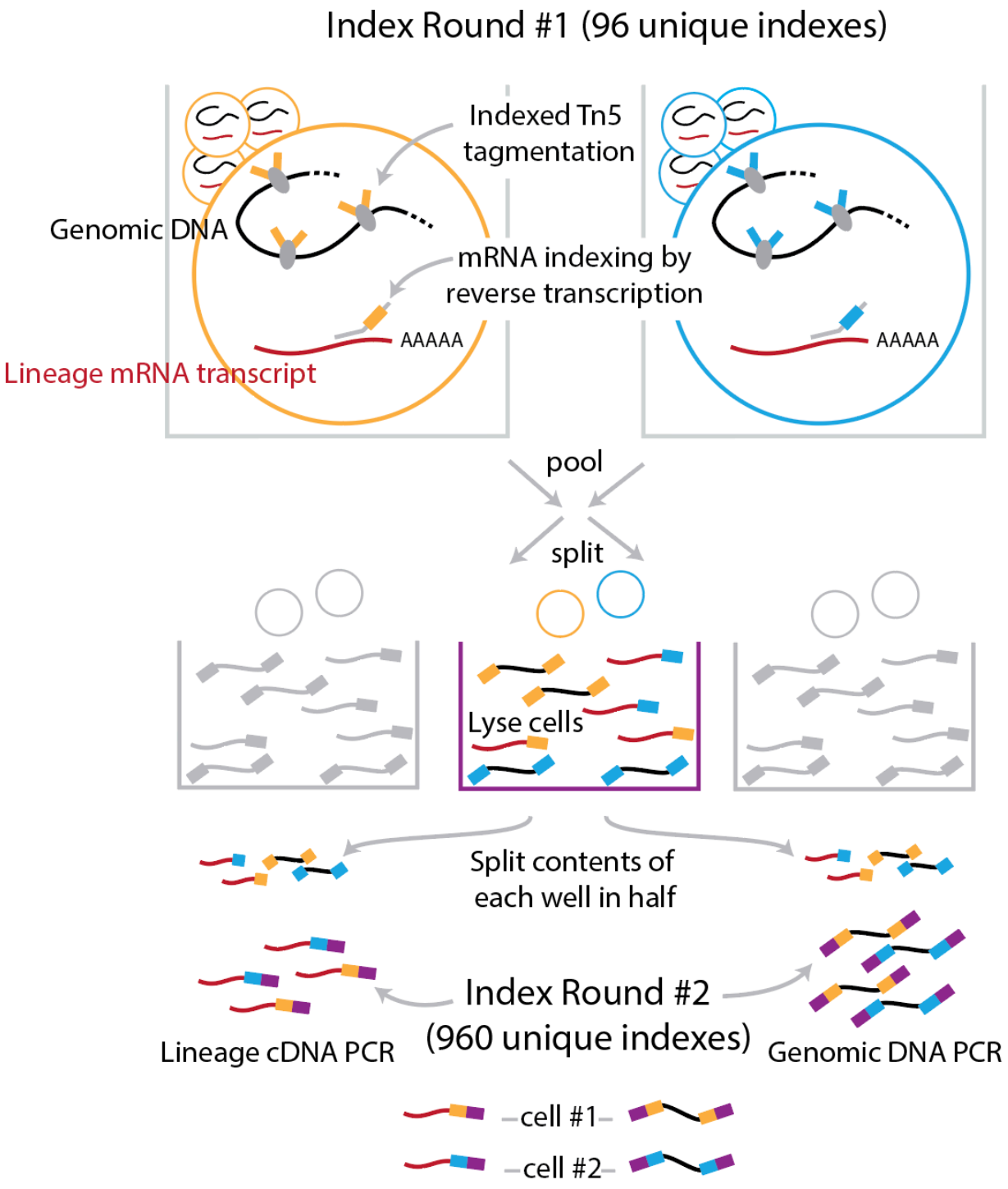


Figure 5.1. A combinatorial indexing strategy to concurrently capture chromatin accessibility and lineage mRNA from the same single cell.

5.1.2 *Tethering chromatin accessibility and expression profiles via lineage information*

We applied this method to the remaining cells from the lineage/expression capture experiment, and filtered cells to those for which we collected both chromatin accessibility profiles and suitable lineage information. Since a lineage tree has already been built, lineage profiles captured alongside sci-ATAC-seq need only be complete enough to accurately place them into existing lineage groups. Keeping cells with at least 5 captured targets of which at least one was edited, with more unambiguous than ambiguous editing events (the latter likely representing doublets), we retained 12160 cells with lineage information. In this group of cells, a median of 20 unique targets were captured per cell (Figure 5.3b). We next filtered on chromatin accessibility profiles. Chromatin fragment lengths exhibited the expected nucleosomal peaks (Figure 5.3a), and filtering on UMI counts yielded a total of 9014 cells (median non-mitochondrial UMI count: 1601; mean UMI count: 6491; minimum 32 UMIs, Figure 5.3a).

To place these cells into existing clonal groups, we first computed a weighted similarity score based on lineage profiles for each ATAC-associated cell with each RNA-associated cell. We then placed cells into existing groups based on nearest neighbors (Figure 5.2a). Encouragingly, the relative group sizes of ATAC-associated cells correlated well with the original group sizes (Figure 5.3c). Moreover, lineage profiles collected alongside accessibility were visually consistent with those collected alongside expression within tethered groups (Figure 5.2b). Together, these data suggest that cells were accurately placed into lineage groups, and thus we can expect analogous heritable states to be reflected in expression and accessibility measurements within a group.

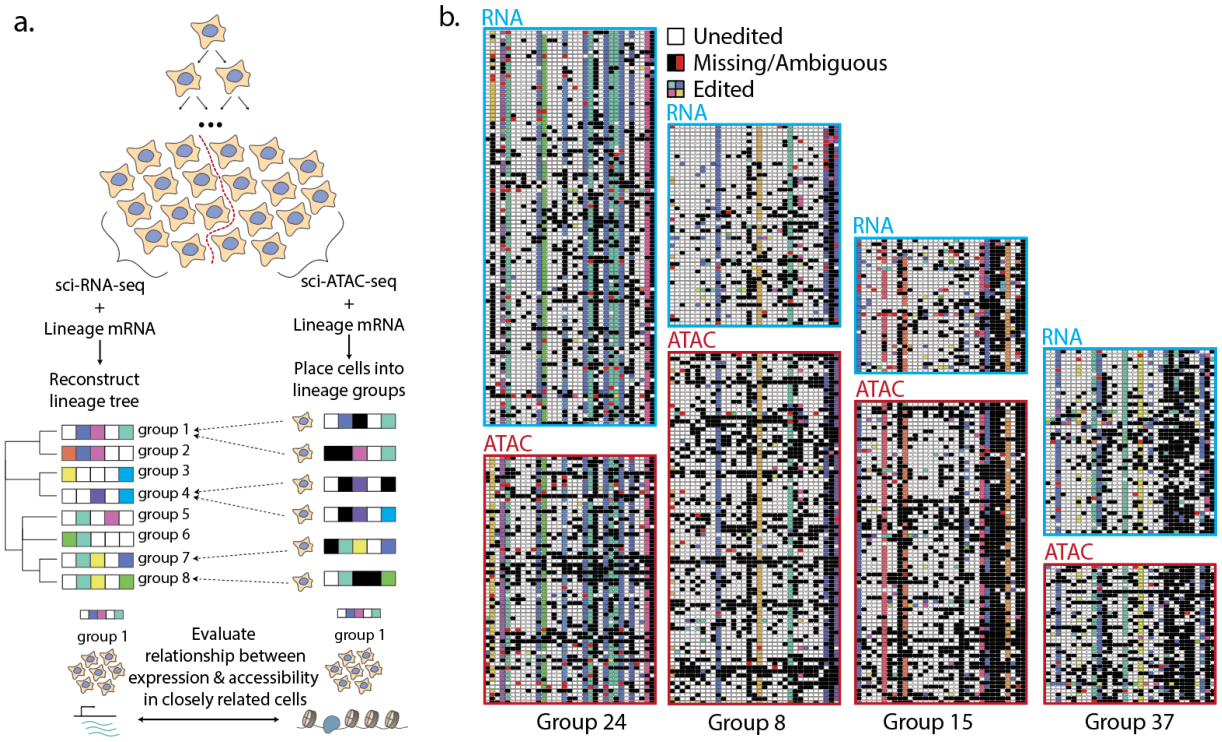


Figure 5.2. Tethering chromatin accessibility and expression profiles via lineage information.

(**a**) Schematic depicting how expression (sci-RNA-seq) and accessibility (sci-ATAC-seq) are linked via lineage information. Lineage-traced cells are split in half, and lineage profiles are captured separately alongside each single cell feature. A lineage tree was reconstructed from cells with concurrently profiled expression, and lineage profiles of cells with concurrent accessibility profiling were used to place cells into previously defined lineage groups via nearest neighbors. The relationship between expression and accessibility of closely related cells could then be evaluated. (**b**) Lineage profiles of individual cells within four clonally related groups collected alongside either sci-RNA-seq or sci-ATAC-seq.

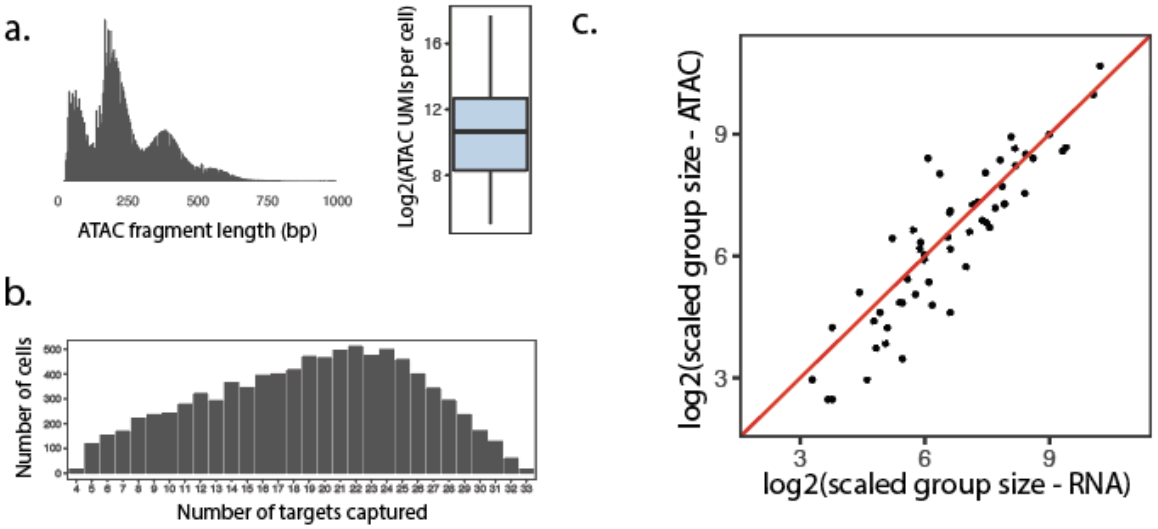


Figure 5.3. Evaluating linked chromatin accessibility and lineage profiles.

(**a**) Histogram of sci-ATAC-seq fragment lengths across all cells (left) and a boxplot of sci-ATAC-seq reads per cell (right). (**b**) Histogram of the number of targets captured per cell included in the analysis. (**c**) Correlation of group sizes collected along sci-RNA-seq and sci-ATAC-seq. Each point represents a single lineage group. Group sizes were normalized to a total cell count of 10,000 for each feature.

5.2 USING LINEAGE-TETHERED CHROMATIN ACCESSIBILITY AND EXPRESSION PROFILES TO INVESTIGATE MECHANISM OF HERITABLE EXPRESSION

5.2.1 *Using lineage-associated chromatin data to infer small copy number changes likely mediating differential expression*

Although sci-ATAC-seq is primarily used to measure local chromatin accessibility changes, copy number changes should also be apparent since they affect the amount of DNA available for tagmentation. Thus, if paired expression and accessibility measurements truly capture closely related cells, CNAs observed in expression data should also appear in accessibility data. To visually evaluate CNA concordance, we quantified relative sci-ATAC-seq read counts across 1MB windows of the genome for each lineage group and generated heatmaps analogous to those shown in Figure 4.1. Indeed, we observed striking agreement in CNA patterns between expression and accessibility data (Figure 5.4a), further confirming lineage profiles do link close cell relatives. To determine if CNAs were measurable in accessibility data at the gene level, we evaluated accessibility within gene bodies, including 5kb upstream of the TSS, once again using the permutation strategy described in Figure 4.5. We found that within CNAs, RNA and ATAC z-scores are strongly correlated at genes which are DE, DA, or both (Pearson's r = .73, Figure 5.4c, left panel), while much more limited correlation is observed outside of CNA regions (Pearson's r = .16, Figure 5.4c, right panels).

Since copy number differences are often observable at the gene level in ATAC data, we wondered if we could use gene body accessibility outside of large CNAs to identify genes whose DE status is likely due to small genomic amplifications or deletions, affecting one or a few genes. Correlated

DE and DA status may alternatively indicate a regulatory change, but such DA is more likely to be promoter-specific; in this case, we would expect a higher promoter-specific signal, while evaluating DA across the whole gene body could dampen such localized signal (Nair et al., 2021). 21 genes outside of CNAs are both DE and DA (Figure 5.4c, middle panel), making them good candidates for residing in short CNAs. In fact, three of these−*MREG, PECR, XRCC5*−−are adjacent genes on chr2, with higher expression in group B relative to group A, despite similar expression outside of this region (Figure 5.4b; Figure 5.5a). This pattern strongly suggests that a focal amplification occurred at this locus, explaining the increase in transcript abundance. Similarly, *AC016205.1* on chr18 & *TAF1* on chrX are both DE and DA between the groups indicated in Figure 5.4b, and also appear within short stretches of genes with elevated expression. A pileup of ATAC data, showing the positions of Tn5 insertions across *TAF1*, shows elevated signal across the whole gene body as well as the neighboring gene *OGT*, validating our prediction. A small CNA is also likely on chr3, where elevated expression is observed in DA gene *SERPINI1* and nearby *PDCD10* (Figure 5.4b; Figure 5.5a, *SERPINI1* does not appear on the heatmap due to low expression level.). Although *PDCD10* is not significantly DA by our metrics, it lies in the vicinity of genes which are (Figure 5.4c, middle panel). Pileup of ATAC reads in this region supports this prediction, with denser coverage of reads across the gene body of *SERPINI1* in group B (Figure 5.5a, right panel). These data suggest that paired expression and accessibility data can help identify small copy number changes.

5.2.2 *Evaluating sources of differential expression unlikely to be mediated by copy number changes*

We next sought to use accessibility data to identify genes whose expression changes are unlikely to be mediated by copy number changes. If a heritable expression change is triggered by a simple

gene copy number change, we expect a linear fold-change concordance between expression and gene body accessibility. If, on the other hand, an expression change is due to other factors, such as abundance of an upstream regulator or change in its regulatory context, these features are not necessarily expected to be linearly correlated. Though log2-fold changes at single genes between variable size groups are inherently noisy, especially in ATAC data, outlier DE genes are especially likely candidates for non-copy number mediated heritable states. We thus further inspected several such outliers, where expression change greatly exceeds accessibility change (Figure 5.4d). Between groups A & B as indicated in Figure 5.4d, the expression change in *GRIA1* is 19 times greater than its gene body accessibility change (RNA log2-fold change = -6.04; ATAC log2-fold change = -.32), suggesting genomic amplification is very unlikely to be the cause of this expression change. Similarly, the expression changes observed in *BCKDHB, CDH12, and ADGRB3* (Figure 5.4e, Figure 5.5b,c) between the indicated groups greatly exceed gene body accessibility changes (log2-fold change shown in figure or legend). The absence of significant accessibility change in *BCKDHB* in particular allows us to rule out a focal amplification of the 3' end of the gene as an explanation for high RNA read coverage specifically in that region in groups E & F (Figure 5.4e). A more likely explanation is that a different transcription termination site was used.

Beyond copy number changes, heritable changes in accessibility at regulatory regions would signal an epigenetic origin to expression variation. We thus identified peaks in ATAC data, both in the entire dataset as well as in lineage-specific subgroups internal to the tree, and looked for DA peaks within 5kb of TSSs or within the gene body between every pair of sister groups near genes found to be DE. We did not observe any DA peaks in these regions. Consistent with this, Kiani *et al.* recently showed that accessibility and expression changes are not well correlated in single gene

perturbation experiments (Kiani et al., 2022). Others have observed a similar lack of concordance between accessibility changes and expression level (Hota et al., 2020; Y. Zhang et al., 2020).

Together, these data illustrate the potential of lineage-based coupling of expression and accessibility data to help distinguish between potential mechanistic explanations for heritable expression changes.

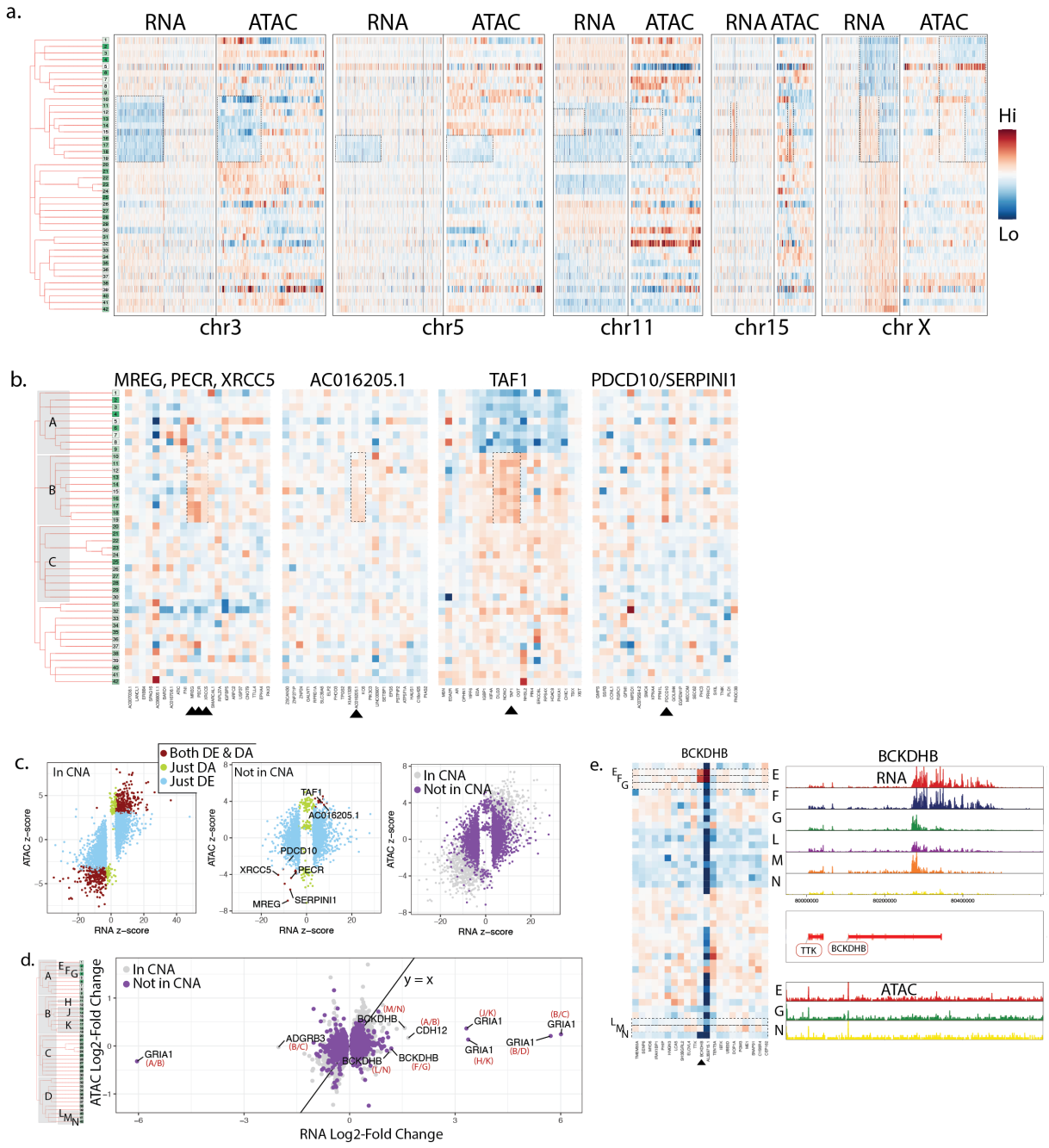


Figure 5.4. Investigating sources of heterogeneity using chromatin accessibility and expression profiles tethered by lineage information.

(**a**) Heatmaps showing the relative expression (RNA) and accessibility (ATAC) across the 42 lineage groups, calculated for each gene (RNA), and for each 1MB bin (ATAC) for five selected chromosomes. Genes & bins are ordered by their chromosomal position. Dashed boxes indicate

chromosomal regions with visually consistent copy number changes across the tree. (**b**) Heatmaps showing relative expression for a subset of genes which are both DE and DA, and including 10 positionally adjacent genes on either side. Associated RNA & ATAC read pileups are shown in Figure 5.5**a**. (**c**) Left: Relationship between expression and accessibility changes evaluated within gene bodies plus 5kb upstream of the TSS, calculated using the permutation approach described in Figure 4.5. Only genes within CNAs are shown. Each point represents an expression/accessibility change at a single gene for a pair of sister lineage groups (and thus a gene may be represented more than once). Points are colored by their DE and DA status. Middle: Analogous to the left plot, except including only genes *outside* of CNAs. Labeled genes are referenced in the text. Right: Overlay of left and middle plots. 10 outlier genes, where noise was likely due to low expression/accessibility, were removed from the middle and right plots. (**d**) Relationship between RNA and ATAC log2-fold change (as opposed to z-score). Each point represents an expression/accessibility change at a single gene for a pair of sister lineage groups (and thus a gene may be represented more than once). Outliers discussed in the text are labeled with gene name and pair of sister groups as indicated on the tree. Because small groups result in noisy data, comparisons involving at least one small group (<100 cells) were removed. An expression cut-off was also applied to reduce visual noise, leaving the 45% of comparisons with the highest expression. (**e**) Left: Heatmap of relative expression of *BCKDHB* and surrounding genes, respectively. Right: Pileup of expression and chromatin accessibility data for the indicated groups (as labeled on tree in panel **d**). Log2-fold change between groups F&G: 1.18 (RNA), -0.06 (ATAC); groups L&N: 1.13 (RNA), -0.12(ATAC).

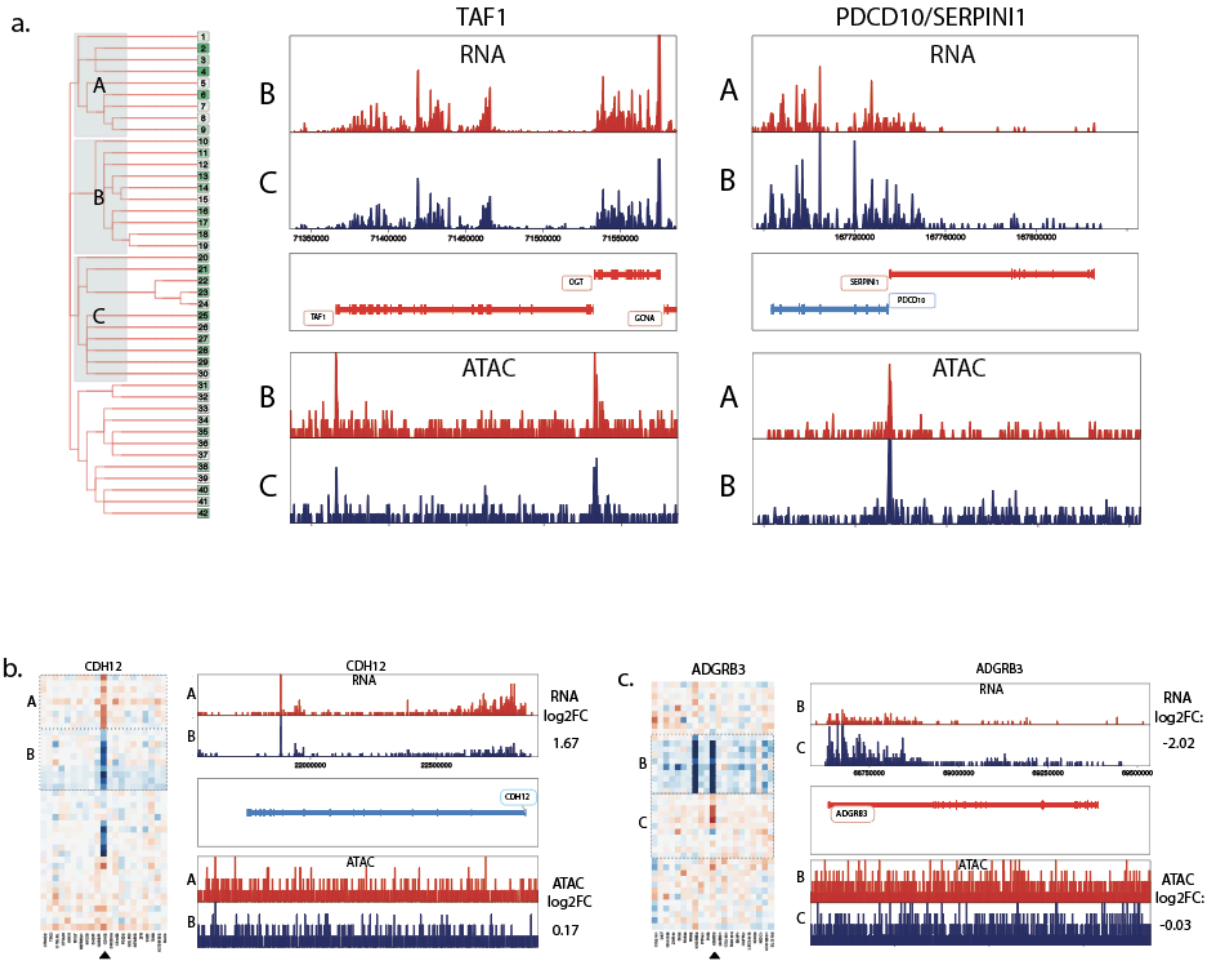


Figure 5.5. Investigating sources of heterogeneity using chromatin accessibility and expression profiles tethered by lineage information (continued).

(**a**) Read pileups for RNA (top) and ATAC (bottom) data for the lineage groups and genes indicated on the tree. Associated heat maps shown in Figure 5.4**b**. (**b**) Left: Heatmap of relative expression of *CDH12* and surrounding genes Right: Pileup of expression and chromatin accessibility data for the indicated groups (as labeled on tree in Figure 5.4**b**) at the *CDH12* locus. (**c**) Same as panel **b**, but for *ADGRB3*.

Chapter 6. MATERIALS & METHODS

6.1 EXPERIMENTAL METHODS: LINEAGE RECORDING

6.1.1 *CRISPR lentiviral target construct & Cas9 construct generation*

Target/sgRNA construct: In order to integrate CRISPR targets and sgRNAs into the genome, we modified the CROPseq vector (Datlinger et al., 2017) (Addgene ID 86708), which expresses an sgRNA and a PolII transcript. We integrated a CRISPR target construct after the WPRE, such that it is expressed off the PolII promoter (sequence and location shown below). Target constructs were identical except for a unique 10bp barcode. sgRNAs matched the targets and were thus identical across all uniquely-barcoded constructs. A primer binding site was placed 35bp upstream of the CRISPR cut site, such that the target accommodates a 70bp deletion. The sequencing and computational processing scheme enables capture of insertions of >105 bp. (see Computational processing and edit calling from lineage target sequencing data))

Target insert:

TCCAAGCTCCATAGGTCCAACTCAAGCTTAGTTCCTATACTGATTCCAAGCCATGGT

ACCATAGCAGATGATCCATTTAGAGCCTGGCTGGTCTCCTGGGAGGTCAACCTTGGA

GACTAAGACCTTACGNNNNNNNNNN

Unique target barcode

gRNA binding site

Forward primer binding site

Position of insertion after WPRE between sequences shown:

TCCCCGCGTCGACTT[INSERTION SITE]TAAGACCAATGACTT

Primer binding sites:

Forward: CTGATTCCAAGCCATGGTAC

Reverse: GACTTACAAGGCAGCTGTAG

A modified version of the doxycycline-inducible SpCas9 lentiviral plasmid (https://www.addgene.org/50661/) was used in this experiment. This construct contains an auxin inducible mAID sequence (cloned from pMK288 (mAID-Bsr), Plasmid #72826, Addgene) This degron sequence was not used in this experiment. Doxycycline was not used to induce this construct -- instead, we relied on known leaky expression to achieve a low level of editing. The full construct sequence is available on Benchling (tinyurl.com/24pbhfth).

6.1.2 *Cell line generation*

HEK293 (ATCC, CRL-1573) were first transduced with the barcoded target/sgRNA modified CROPseq vector at high MOI and single cells were sorted to grow clonal populations. Targets were counted by PCR amplifying and sequencing the unique barcodes. A clone containing 31 unique barcodes was chosen.

To induce editing, cells were transduced with the doxycycline-inducible Cas9 lentiviral construct described above, selected for Cas9 integration using Blastocidin, and single cell sorted such that all profiled cells arose from a single founder cell. The Cas9 construct was not induced with doxycycline; instead, we made use of its known propensity for leaky expression without induction to produce slow editing. After 35 days in culture (DMEM), passaged every 2-3 days using trypsin, editing efficiency was evaluated by bulk PCR of the target regions, and a single

clonal edited population was chosen for further exploration. A portion of the resulting cells were collected and processed immediately in a target+sci-RNA-seq capture experiment, and a portion was frozen in liquid nitrogen for later target+sci-ATAC-seq processing.

6.2 Experimental methods: Capturing lineage profiles alongside molecular profiles

6.2.1 *Concurrent capture with sci-RNA-seq*

The sci-RNA-seq 2-level protocol for methanol-fixed cells described in Cao *et al.* 2017 (Cao et al., 2017) was modified to concurrently capture CRISPR target mRNAs. A single 96 well plate was used for the first round of indexing, and 8 96-well plates were used in the second round, with 25 cells sorted into each well.

The following modifications were made:

(1) To index the lineage target mRNA during the first round of indexing, we added a 1um of 10uM indexed target-specific reverse transcription primer in addition to the oligo-dT primers.

Reverse transcription primer sequence:

ACGACGCTCTTCCGATCTNNNNNNNNTTGGTAGTCG ctacagctgccttgtaagtc

UMI

RT index (well-specific sequence)

(2) After Tn5 tagmentation, lysis, and ampure bead purification, cDNA was eluted in 10ul of buffer EB (Qiagen). Then half of the contents of each well were transferred to a second 96 well plate. In one plate, PCR and sequencing of the transcriptome was performed as described. The other plate

was used for amplification of the lineage targets, with well-specific primers indexed to match well-specific transcriptome indices.

Lineage targets were PCR amplified using the KAPA HiFi HotStart ReadyMix (Roche, KK2602) with primer sequences below and elongation time of 1 minute and an annealing temperature of 65°C. All other steps were consistent with the KAPA protocol provided by manufacturer.

PCR primers:

Forward (unindexed):

CAAGCAGAAGACGGCATACGAGATTTGGTAGTCGGTGACTGGAGTTCAGACGTGTG

CTCTTCCGATCTCTGATTCCAAGCCATGGTAC

Reverse (indexed):

AATGATACGGCGACCACCGAGATCTACACTTCTACCTCAACACTCTTTCCCTACACG

ACGCTCTTCCGATCT

PCR index (well-specific sequence)

PCR index (plate-specific sequence)

After PCR, all wells were pooled and a 0.8x AMPureXP bead cleanup was performed prior to sequencing.

(3) Paired-end sequencing of the lineage target PCR products was performed using a 300bp Illumina sequencing kit (Miseq), with 148 bases sequences from each end (along with standard 10bp index reads, which are associated with the second round of indexing). The first index as well as the UMI appear in R1 and are parsed during downstream computational processing. 10% PhiX was added for sequencing to address sequence homogeneity.

6.2.2 *Concurrent capture with sci-ATAC-seq*

The concurrent lineage target + chromatin accessibility capture protocol builds upon the 2-level sci-ATAC-seq protocol presented in Cusanovich *et al. (*2015) (Cusanovich et al., 2015). The following modifications were made:

(1) Lysis buffer was supplemented with SuperaseIN (ThermoFisher AM2694).

(2) Reverse transcription of lineage target mRNA:For s first round of lineage target indexing, reverse transcription was performed prior to tagmentation in the first set of wells. After lysis, 5000 nuclei (2ul) were distributed per well of a 96 well plate, along with reagents for the first step of reverse transcription: 0.25ul dNTPs (10mM) & 1ul of indexed the reverse transcription primer described above (at 2uM). The plate was then incubated at 55C for 5 minutes, and immediately chilled on ice. Reagents from the SuperScriptIV (ThermoFisher, 18090010) kit were then added to each well (1ul buffer, .25ul DTT, .25ul SSIV enzyme, .25ul RNAseOUT (ThermoFisher, 10777019). The plate was then incubated at 55C for 10 minutes, and immediately chilled on ice.

(3) Buffer exchange following reverse transcription: 60ul of nuclei lysis buffer was added to each well. Nuclei were then pelleted by centrifugation at 300g for 5 minutes in 4°C. 57ul were then carefully removed from each well, taking care not to disturb the pellet.

(4) After sorting nuclei (25 nuclei per well) into a solution containing SDS & inclubating to insure Tn5 inactivation and lysis, the contents of each well are split in half across two plates. One plate

underwent indexed DNA PCR amplification in accordance with the sci-ATAC-seq protocol; the other underwent a 2X AmpureXP bead purification to remove SDS, followed by PCR amplification as described above. Primer cleanup and sequencing of lineage target amplicons was performed as described above.

6.3 COMPUTATIONAL PROCESSING: EXPRESSION ANALYSIS

6.3.1 *Initial computational processing of sci-RNA-seq data*

Sequencing was performed as previously described (Cao et al., 2017). Reads were adapter-trimmed using trim_galore and aligned to the reference genome (hg38) using STAR. Non-unique mappers were removed. Reads were then deduplicated using a custom script (190223_sciRNA_remove_duplicates.cpp), taking into account both UMIs and cell indices to call a duplicated read. Only cells with at least 2048 deduplicated non-mitochondrial UMIs were used for subsequent analyses.

A custom script (190704_process_sciRNA_mapped_file.cpp) was used to map reads to genes. Reads which overlapped multiple genes but only fell in an exon in one gene were counted towards that gene.

RNA processing to generate the cell by gene raw counts file is implemented in script 190807_sciRNA_wrapper_ALL.txt, with user-defined UMI cutoff of 2^11.

6.3.2 *Permutation Analysis for DE gene identification*

DE genes were identified using the following procedure.

First, raw counts were scaled to 10,000 reads per cell. Then, for each pair of sister groups within the tree (defined as those that share an immediate common ancestor branch), cells were permuted into two groups of the original sizes 10,000 times and the log-fold change for each gene was calculated. Only genes which were expressed in at least 10% of cells in either group were kept for downstream analysis. The measured (real) mean expression ratio for each gene was ranked against the permuted values, for a total of 10,001 values. Z-scores are calculated here as the distance of the real log ratio from the mean divided by the standard deviation of the permuted values.

To account for differences in group sizes across the tree, as well as large CNVs, we evaluated genes on each chromosome in each pair of groups separately to determine the rank cutoff values associated with significant DE. We chose a false discovery rate cutoff of 5%.

Rank cutoff values for each chromosome-group pair combination were determined as follows. If no genes on a chromosome were differentially expressed, we would expect a uniform rank distribution for 1 to 10,001. Thus, the expected number of genes observed at any given rank value is the total number of filtered genes on chromosome/10,001, referred to here as the baseline value. If true DE genes are present, we should observe an enrichment of genes at either or both ends of the distribution, manifesting as higher counts and denser coverage.

An FDR value for each rank position can be determined simply by subtracting the baseline value from the total gene count at each rank. Since those genes of rank 1 or 10001 are most likely to be

true positives, we begin at the ends and move inward to identify a group of ranks which together produce an FDR of <= 5%.

The procedure to determine significant ranks is implemented as follows. We begin at rank 1 or 10001, choosing the one with the highest observation count, and calculate the FDR associated with that rank. If it is smaller than 5%, we compare the next most extreme ranks (2 or 10001 if rank 1 was already used), and again choose the one with the highest gene count. We calculate the total FDR encompassing both rank positions and continue this procedure iteratively, until the FDR reaches 5%. All genes with the ranks identified by this procedure are considered DE.

Genes which were lowly expressed in both groups being compared (defined as those for which the percent of cells expressing the gene, calculated separately and then summed between the two groups, is <10%) were removed from the final analysis.

Procedure implemented in A_210327_perm_qsub_script.sh & 210330_process_permutation_table_log_version.cpp.

6.3.3 *SNP-based copy number analysis*

To identify variable genomic positions from expression data, a 4 column file was generated for each chromosome from the STAR alignment output file, including cell name, mapping position, CIGAR string, and sequence, and the frequency of each base was calculated as implemented in 201114_wrapper_for_ASEs_for_lineage_groups.txt. Counts were generated for all cells as well as subsets of groups. Variable positions were retained and SNP info was added to via code

191018_add_snp_info_to_ASE_file.cpp, using as input a tab-delimited file generated from a vcf file, containing five files: chromosome, position, rs_id, major allele, minor allele. Plots were generated in 200203_ASE_calc_major_freq.R.

6.1 Computational processing: Chromatin Accessibility Analysis

For processing sci-ATAC-seq sequencing reads, we first compare observed and expected lists of single cell indices, correcting any indices with a likely off-by-one error. All reads are then adaptor trimmed using trimmomatic (parameters: TRAILING:3 SLIDINGWINDOW:4:10 MINLEN:20), and all reads associated with a single cell are then aligned to the genome using bowtie2 (hg38 genome build). Reads are then deduplicated by UMIs using a custom script (191226_CROPt_process_atac_bedfile.cpp). Both cell by gene and cell by interval counts were generated using a custom script (191226_CROPt_make_cell_by_interval_count_file.cpp). During analysis, count files were converted into the 10X Genomics format for compatibility with other analysis tools.

For heatmap plotting, counts per gene/interval were pooled by lineage group, and a mean was calculated for each gene using total number of UMIs (as opposed to total number of cells) as the denominator to account for a large spread of total observed UMIs per cell. Each value was then scaled by the median of the total read count for all genes/bins. Genes & bins with low total counts across the dataset were removed (those whose scaled values were below 120 per 1MB bin, or below 5 per gene, in all groups). For each retained gene/interval, the lineage group mean was divided by the mean accessibility of all cells at that gene/interval, and a log was taken to center around 0. For visualization scaling purposes, values above or below .9 & -.9 respectively

were changed to those values. This was implemented in 210222_ATAC_process_bin_counts_by_groups_play_w_scaling.R.

Differential accessibility was evaluated using the permutation approach described above, with mean counts per a group again calculated with total number of UMIs (as opposed to total number of cells) as the denominator.

Pileups were plotted using ArchR (Granja et al., 2021). For DA analysis at peaks, a set of peaks was determined using ArchR, using both the whole dataset as well as successive subgroups moving across the tree. The union of these peaks was then overlapped with DE genes (including 5kb upstream) and DA at these peaks was again evaluated using the permutation approach.

6.2 COMPUTATIONAL METHODS: LINEAGE PROFILE CALLING AND CELL FILTERING

6.2.1 *Computational processing and edit calling from lineage target sequencing data*

Targets were enriched from the cDNA as described above and sequences on the Illumina Nextseq or Miseq 300 cycle kit, with paired end sequencing. Read pairs (150b from each end on Miseq; 148 from each end on Nextseq) were merged using PEAR. Since large insertions can possibly result in pairs which do not overlap, we took reads which were unable to be merged and looked for features (common sequence near barcode, primer binding sites) which indicated reads from the correct location. We then pasted the pairs into a single read, and used the combined insertion sequence in our analysis. Thus, insertions of >105bp could be captured, as long as the amplicon could cluster efficiently on the sequencer chip.

Merged reads contain UMIs (first 8bp), reverse transcription index (index #1 of combinatorial indexing - next 10bp), and a target ID (obtained by searching for flanking sequences). These features were first extracted from the reads (191203_CROPt_make_UMI_RT_BC_seq_output_file.cpp, within wrapper script 191203_CROPt_Step2_collapse_UMIs_wrapper.txt), and the remaining sequences were collapsed by UMIs (191203_CROPt_collapse_by_UMIs.cpp, run within 191203_CROPt_Step2_collapse_UMIs_wrapper.txt) and aligned to the reference sequence using needleall (http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/needleall.html) with default settings. To remove PCR amplification or sequencing errors being interpreted as a CRISPR edit, we devised a strategy to disentangle likely editing from technical errors in sequences where indels or mismatches appeared discontinuous and/or did not overlap the CRISPR cut site. Beginning at the cut site and moving in either direction, each part of a real "edit" had to be within 4 bases of the last position of an edit. This reduces the possibility that a technical error will be counted towards an edit, while allowing for some edits which appear discontinuous. These likely result from complex events in which bases were both deleted and inserted, with small fragments of insertions mapping to the reference sequence of the deleted region.

Editing at each target in each cell was then evaluated. An unambiguous target was defined as one which either contained no discrepant editing patterns, or if multiple editing patterns were observed, had more than one UMI (unique transcript) associated with the "real" editing pattern, and no more than 1 of the other (assumed to be either a stray transcript picked up during processing or a product of template switching during PCR). If more than one edit was associated with more than one UMI,

the target was termed "ambiguous." If each edit was only associated with one UMI, the target also was termed "ambiguous." For the two duplicated targets, if ambiguous editing patterns were distributed in silico as described below.

The above steps are implemented in wrapper script 191205_local_target_analysis_all_UPDATED.txt.

6.2.2 *Evaluating CRISPR target capture rates and filtering cells based on target capture and expression*

The dual sci-RNA-seq + target capture was performed in eight batches. The median number of targets captured varied by batch (Figure 2.4). This discrepancy was traced to the batch of Tn5 buffer used in each batch: more recently made batches as well as the commercial batch (as opposed to older buffer made internally) produced more efficient Tn5 integration into cDNA (readily observed in difference of sci-RNA-seq median library size). Since Tn5 integration occurs prior to separating the samples for separate RNA and target processing, a smaller cDNA fragment size means that Tn5 is more likely to integrate within a target region (downstream of the 5' primer binding site), thus preventing that target from being captured. Thus, optimization of buffer composition might address this issue.

To filter out presumed doublets, both target editing and expression data were used (Figure 2.4). Cells were called "Singlet" of "Doublet" based on fraction of ambiguous targets (those with more ambiguous than non-ambiguous targets were considered doublets). For doublet cells, the sci-RNA-seq UMI count distributions were shifted, indicating that high count cells are likely doublets. In addition to removing cells defined as doublets by target editing patterns, we thus additionally

removed cells which were above 1.8x the median sci-RNA-seq UMI count for each batch (Figure 2.4c).

6.3 Computational Methods: lineage tree reconstruction

The final tree was reconstructed via the following steps:

6.3.1 *(1) Computationally split duplicated targets*

Two targets (#30 & 31) were consistently associated with two editing patterns within a single cell, strongly suggesting that the section of chromosome on which these targets reside underwent a duplication event in an early cell division (or in an ancestor of the founder cell of this population). Because editing patterns at these targets clearly contained early editing events which were informative of tree structure, we decided to computationally split each target into two separate targets. For each target, we first generated a list of pairs of edits which were commonly found together in a single cell. Since these had to have occurred at two different targets, we constrained a set of editing patterns to one target and a set to the other. Editing patterns which were frequently found alongside an unedited target (indicating that just a single target of the pair was editing in that subset of cells) or on their own (indicating no duplication or a loss of the duplicated target) were randomly assigned to the first target of the pair. Thus, a list of allowed "edits" was generated for each target in the pair. If a cell contained edits on either list, they were distributed accordingly between the pair of targets. The final dataset thus contains a total of 33 targets per cell.

6.3.2 *(2) Infer missing data*

While a subset of missing data reflects true loss of either the target itself (due to a large deletion or a CNA) or an editing pattern which makes the target hard to capture (e.g. a very large insertion), some targets are stochastically not captured during mRNA processing. We thus attempted to infer these edits using a nearest neighbors approach. Since batch 1 had the most complete lineage data, for correcting missing data from other batches we combined them with batch 1 cells and performed the following steps. We first calculated similarity scores between every pair of cell lineage profiles using an additive approach. For each target with matching editing patterns a score of 5 would be added to the total; for each target that was unedited in both lineage profiles, a score of 1 would be added. Targets which did not match (or contained missing or ambiguous data in either cell) received a score of zero. Based on these similarity scores, we defined a set of "nearest neighbor" cells for each cell, and used these to computationally infer missing data for each cell. Specifically, for each cell, for each missing or ambiguous target, we used the most common editing pattern in its closest set of neighbors at that target to infer the missing edit. If the majority of neighbors also had missing data at this target, this likely reflects a true loss at this target, and thus was left uncorrected.

Steps 1 & 2 above are implemented in
200713_wrapper_for_wrapper_for_AMBcorr_Xcorr_step.txt.

6.3.3 *(3) Generate initial groups of related cells using hierarchical clustering.*

We generated a similarity matrix using the similarity score described above, and hierarchically clustered cells via Ward's method (Ward2 in "hclust" package in R). Duplicated targets described

in "Computationally split duplicated targets" (targets 30-33) were not used for similarity calculations as they were found to bias groupings. Trees generated via hierarchical clustering are not consistent with progressive CRISPR-based editing events, but do a reasonable job of placing similar cells next to one another. Hierarchically clustered trees can be split automatically into a desired number of groups, but we found that for downstream applications, it was best to manually determine how to split the tree since in some cases groups of very different sizes were desired. We thus generated plots of the hierarchically clustered tree (resembling the inset in Figure 3.2 but containing the full tree) and manually chose the break points at which groups should be split. We generated plots of both the lineage profile in which we had inferred missing data as in step 2, and of the raw data, and consulted both to ensure missing data inference appeared accurate. Importantly, these groups were chosen with the intention that some would be split further in a subsequent step: as long as cells appeared confidently as close relatives, they were kept in a single group at this stage. This procedure generated 94 groups. Groups with less than three cells were removed to be placed into larger groups at a subsequent step, leaving 45 groups remaining.

Groups were evaluated visually as implemented in 200811_combine_like_cells_for_loop.R, 200219_make_LG_group_plots_for_combined_cell_groups.R, & 200225_plot_many_LG_on_one_plot_from_Refcell_list.R.

6.3.4 *(4) Generate a "consensus" lineage profile for each group.*

A consensus editing pattern at a target was defined as one which appeared in at least 75% of cells in that group. A single consensus lineage profile was first generated automatically using this definition for each group. We then manually corrected these profiles to account for known sources

of missing data which may contribute to an editing pattern being captured at fewer than 75% of cells. For example, large insertions and deletions are captured less efficiently, and thus a target in which contains >25% of missing data, but the remaining cells contain a consistent large insertion or deletion, we can plausibly infer that that editing pattern is likely present in all cells.

6.3.5 *(5) Generate a preliminary lineage tree of consensus cells via an iteratively applied greedy approach*

If no data were missing and no convergence (identical edits occuring at a single target independently) were present, one could theoretically build a perfect tree using the greedy approach shown in Figure 3.1. First, we identify the most abundant editing pattern at a single target in the tree, and split the consensus cells into two groups based on the presence or absence of this editing pattern. This defines the first branch point. We then apply this approach to the two new subgroups, and iteratively apply it to all subsequent groups to generate a bifurcating tree with leaves being defined by a single consensus lineage profile (implemented in 201109_building_a_tree_3_record_all_changes.cpp). We then collapse any bifurcations which are not supported (when a branch is formed which is not defined by a specific editing event), such that greater than two branches can arise from a single node (201109_AUTO_collapse_bifurcations.R).

Though the consensus editing patterns are not perfect with regards to the above algorithm (there are several instances of convergence, and some missing data), the pooling of related cells to increase confidence of consensus editing patterns makes the algorithm above a viable approach. We thus applied it to the preliminary group of consensus lineage profiles to generate a preliminary tree.

As described above, some groups could be subdivided further. We thus applied the above algorithm to subgroups of the tree, by taking all cells within a single consensus lineage profile, subdividing them into smaller "consensus" groups (beginning with hierarchical reclustering), and generating a subtree as described above. These subtrees were then combined to form the larger tree.

Importantly, this approach of successive tree and subtree generation allows us to deliberately leave out potentially problematic targets, and to choose different sets of targets for each subtree reconstruction. For example, since targets 30-33 contained missing data which may have been the product of edit pattern distribution to resolve target duplication, we removed these for the initial hierarchical clustering which generated cell groups, but used this information for consensus lineage profile calling and greedy tree generation.

Though branching order correctly describes the order of editing events, the depth of the branching events shown in Figure 3.2 does not necessarily indicate a true temporal relationship. Depth on the tree correlates with the number of edits which occurred over the course of that branch's formation but should not be interpreted as temporal relationships as a consistent editing rate cannot be assumed.

6.3.6 *(6) Visualizing preliminary trees for manual correction of missing data and resolution of convergence events.*

Visualizing these trees at various stages allowed us to refine the trees further by helping to resolve previously unclear editing patterns within some consensus cells. For example, the edit at target 26 in groups 33-42 is a large insertion which is not efficiently captured. The majority of cells within groups 33-40 contained missing data at this target, while a subset contained the insertion. But based on the edit in target 31, it appears most likely that all cells actually did contain the insertion at target 26, but it was not captured well. We thus manually corrected targets at which events like these appeared to be the case.

Visualization of intermediate trees also helped to resolve convergence events. Though few convergence events (defined as the same edit occurring multiple times independently at the same target) impacted the automatically-generated tree structure as earlier subdivisions isolated these events from one another, this was not the case in a few places in the tree. In these cases, a group which visually appears to be closely related to another group because of subsequent shared edits is separated from it in early divisions. These events were manually corrected as well.

In two instances, several convergence events were also resolved by shared CNAs between groups. This was rare; with the exception of the instances described below, expression data was not used for tree reconstruction.

Change 1: A single discrepancy (copy number pattern on chromosomes 5 & 11) revealed a convergence event whereby a common editing pattern occurring independently (target #7, teal edit) forced groups together improperly. Instead, a common CNV pattern at chromosomes 5 & 11

strongly suggested that groups 16-19 shared a common ancestor. A change was made accordingly, slightly increasing tree resolution.

Change 2: CNVs on chromosomes 6 & 11 also allowed for better resolution of groups 38-42, where a combination of factors including a convergence event of a commonly observed edit and a large insertion event frequently manifesting as missing data made it challenging to resolve tree structure.

We found for downstream analysis that small groups reduced power below the level at which meaningful expression and accessibility differences could be detected. We thus recombined some closely related groups such that the minimum number of cells per group is 34.

In the end, the final tree contained 42 lineage groups.

6.3.7 *(7) Integrating remaining cells into pre-defined consensus lineage groups*

About a quarter of the cells (batches 1 & 3) were used to construct the original tree. Some of these which formed a group of 1 or two cells in step 3 were removed to be placed into larger groups later, along with the remaining three quarters of the cells w/ lineage profiles. We placed cells into their most closely related groups by calculating similarity scores described above(see (2) Inferring missing data above) on uncorrected lineage profiles with cells already in the tree, and placing new cells into the group in which they had the highest similarity scores. If a cell had identical similarity scores w/ cells from multiple groups, it was placed into the group in which it had the most neighbors.

Final lineage groups were evaluated visually, by plotting lineage profiles of all cells in a single group and visually confirming shared editing patterns.

6.4 ORIGINAL VISUALIZATIONS

6.4.1 *Tree lineage profile visualizations*

Tree visualizations were generated using custom code (200807_AUTO_tree_custom_visualization_organized.R, internally running 200806_make_coordinates_for_tree_plot.cpp), which converted tree structure into line segment coordinates which can be plotted in a ggplot space alongside visual lineage profiles. Input files are provided (tree_file_LinRNA, lineage_profiles_wRNA.txt).

Visualizing single lineage groups (**Figure 7c**) implemented in 211129_CopyForFigsRNA_Uncorr_AUTO_tree_custom_visualization_organized.R (RNA) & 211129_CopyForFigsATAC_Uncorr_AUTO_tree_custom_visualization_organized.R (ATAC).

6.4.2 *sci-RNA-seq visualization*

For heatmap plotting, counts per gene were pooled by lineage group, and a mean was calculated for each gene using the total number of cells as the denominator. Genes with low total counts across the dataset were removed. Specifically, a lowly expressed gene was defined as one which was expressed at a mean of .5 counts per cell or less in all lineage groups. For each retained gene, the lineage group mean was divided by the mean expression in all cells of that gene, and a log2 was taken to center around 0. For visualization scaling purposes, values above or below 1 & -1 (Figure 4) and 1.5 and -1.5 (all other figures), respectively, were changed to those values.

Visualization implemented in 201117_AUTO_NewGroups_BETTER_long_AllChr_heatmap_plot.R.

Pileups were plotted using ArchR (Granja et al. 2021).

Chapter 7. SCIENCE WRITING

Four years into my graduate studies, the coronavirus pandemic began. As I spoke with friends and family as they considered how to respond appropriately in the early days of the pandemic, it became clear to me that my status as a scientist, for better or for worse, gave me credibility in these conversations. While the media was rife with competing messages, calling for everything from strict lockdown to business as usual, I saw people comforted to hear from a scientist: *No, you are not overreacting by canceling that event. This is a big deal. You are making the responsible choice*.

So when the slow rollout of testing in the U.S. relative to other countries spawned conspiracy theories, I set out to write a general audience piece to explain the most likely possibility: that PCR primers designed by the C.D.C. less efficiently amplified the viral genome than those used in other countries. This piece (below) snowballed into a larger discussion of diverse factors impacting testing, including regulatory barriers and supply chain bottlenecks.

Several months later, another threat to public decision making regarding testing occurred: a New York Times article claimed that the majority of positive PCR tests were false! The epidemiologist whose ideas were promoted in the article presented an overly simplistic view of infection which supported his campaign to increase rapid testing, but with potentially grave consequences for our collective decision making following test results. The second article presented here was written as a response, to clarity the much more nuanced science behind his assertions, as well as highlight the complex interplay of motivations and emotions which have drawn us to welcome simple stories.

Both pieces were originally self-published on Medium. The first was ultimately republished on Medium's Elemental Blog. I was encouraged by the positive response to both pieces, and I plan to continue writing for a general audience in the future.

7.1 The Science Behind Coronavirus Testing, and Where the U.S. Went Wrong

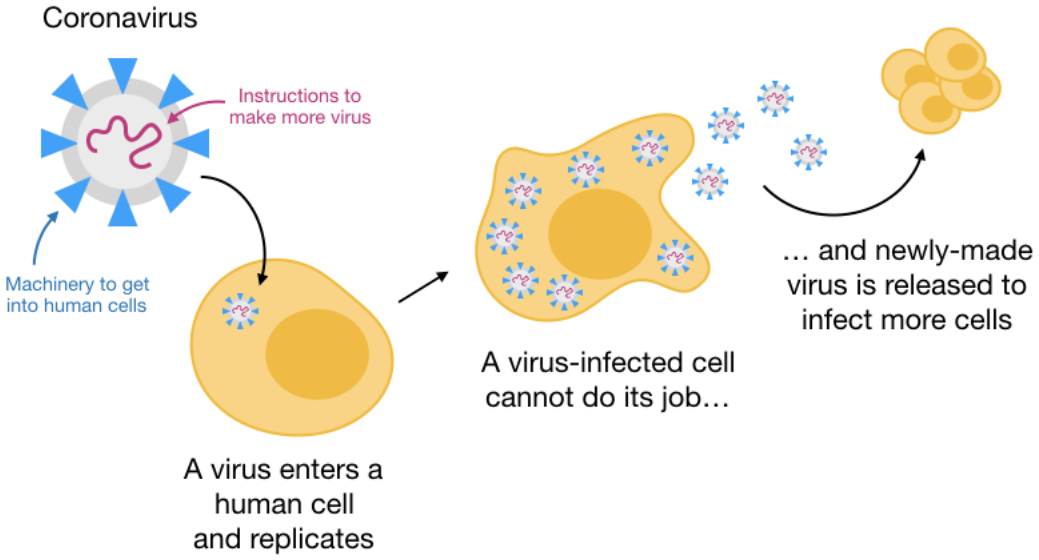
This piece was published on Medium's Elemental Blog in March 2020. It can be found at https://tinyurl.com/44hz9xv5.

As has been widely reported, a major bottleneck in addressing the novel coronavirus outbreak in the U.S. is the extremely limited testing capacity. While South Korea has tested over a quarter million people, the U.S. has performed only 33,000 tests to date*, just three times South Korea's *daily* testing capacity. Speculations about negligence, incompetence, and deliberate conspiracy have been floating around to explain this discrepancy. In reality, as is almost always the case, the factors impacting the U.S.'s ability to ramp up testing are incredibly complex. Complications include regulatory hurdles at the federal and local levels, shortages of supplies, equipment, and certified personnel, as well as technical challenges associated with the test itself.

Given these challenges, it may surprise you to learn that the coronavirus test is relatively simple, and operationally the same in every country. As a molecular biologist, I have run similar procedures hundreds of times. So with South Korea performing an order of magnitude more tests than the U.S., where did we go wrong? If the test is so simple that it can be performed in any molecular biology lab, why aren't we routinely testing thousands of people every day? Why aren't results available for days after testing? Will curbside and at-home testing help?

How does the coronavirus test work?

To answer these questions, let's first consider the culprit the test aims to detect: the virus itself. Viruses, at their core, are surprisingly simple entities: capsules with machinery to penetrate a cell, containing genetic information with instructions to make more viruses. Once a virus enters a cell, the instructions are read and more viral parts are made and assembled. Newly made viruses have mechanisms to escape their host cells and, in the case of coronavirus, travel further down the respiratory tract, eventually reaching the lung cells. When infected, lung cells can no longer perform their normal jobs, leading to the respiratory symptoms of Covid-19 (the disease caused by the novel coronavirus).



The novel coronavirus enters and multiplies inside our cells

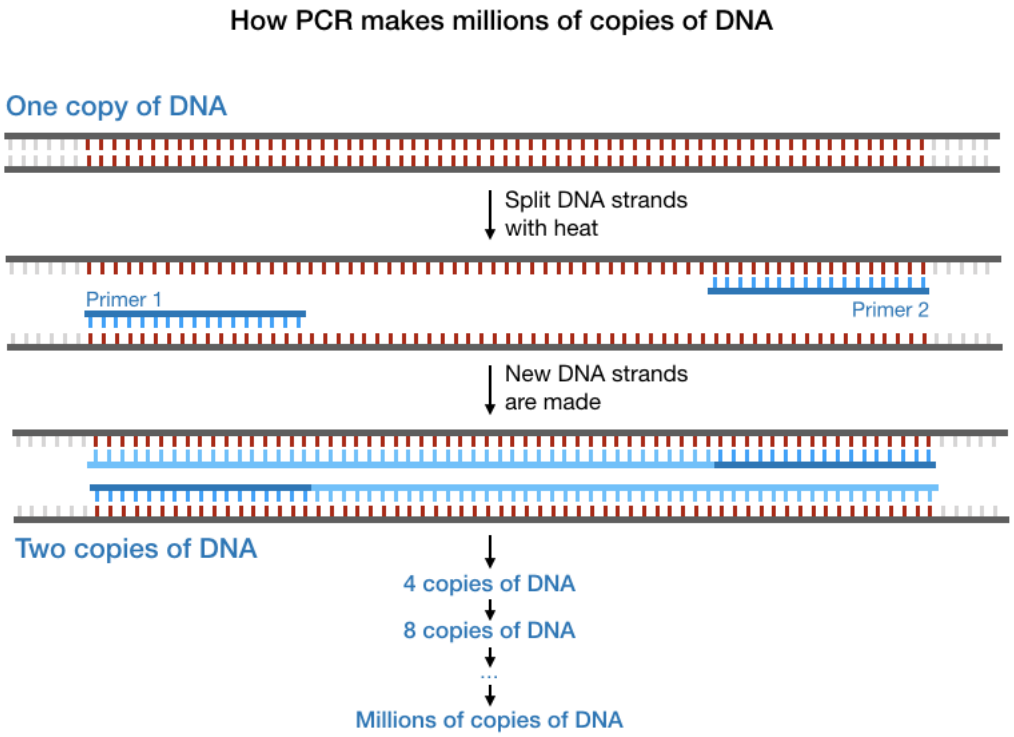
There are a number of ways to detect the presence of a virus in a clinical sample. We can look for an indication that our immune system is reacting to it, or we can look for the virus itself. The latter turns out to be much more straightforward because we have a tried-and-true, easily

adaptable method to detect one feature of the virus: the “how to make more virus” instructions that it carries.

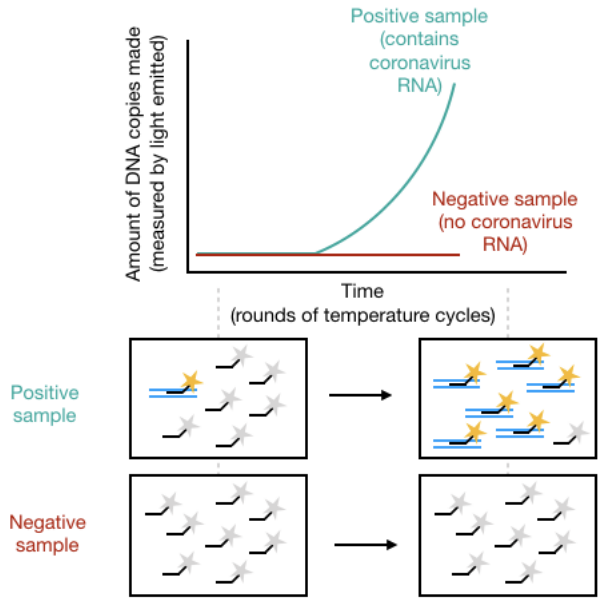
These instructions are the virus’ genetic information, similar to our DNA. They are a chain of molecular building blocks—abbreviated as As, Cs, Gs, and U’s—strung together in an order we understand—they are a manual we can open and read, letter by letter. To detect a virus, we just have to look for its instruction manual.

To do so, we use a standard molecular biology technique called polymerase chain reaction, or PCR. PCR allows us to make copies of a small, specific section of DNA—say, a sentence from the instruction manual—as long as we know the first and last word of that sentence. These flanking words are called primers, and as you will see, it is important to choose them wisely. By combining primers with the patient sample and a few other components in a machine that can cycle through a few different temperature settings, we can produce millions of copies of our chosen DNA “sentence” relatively quickly (in 30–45 minutes).

(To be totally accurate, coronavirus is actually an RNA virus. RNA is similar to DNA, but this method looks a little different in practice and is referred to as RT-PCR. The outcome is the same: Many, many copies of the DNA are made from the viral RNA instructions.)



But wait, which sentence in the instruction manual are we trying to copy…and why? The sentence itself turns out not to matter much, as long as it exists exclusively in the coronavirus' instruction manual; it must be absent from the genetic code of other common viruses like flu. We make copies not so we can read them, but purely to see they can be made. Think of it this way: A million copies of a document are easier to see than a single sheet of paper. To "see" these copies, we use a fluorescent probe, which only emits light when it binds to DNA—the more DNA copies are made, the more light is emitted from the sample. In a sample from an uninfected individual—where no viral material exists in the first place—no copies will be made, and no light will be emitted. A sample containing viral RNA will get brighter over time. We can track the amount of light emitted by each sample in real time, as shown below.



The tests developed by the Center for Disease Control (CDC) in the U.S., the World Health Organization (WHO), and in-house at labs around the U.S., as well as those used in South Korea, all employ this basic strategy: using RT-PCR to detect viral RNA.

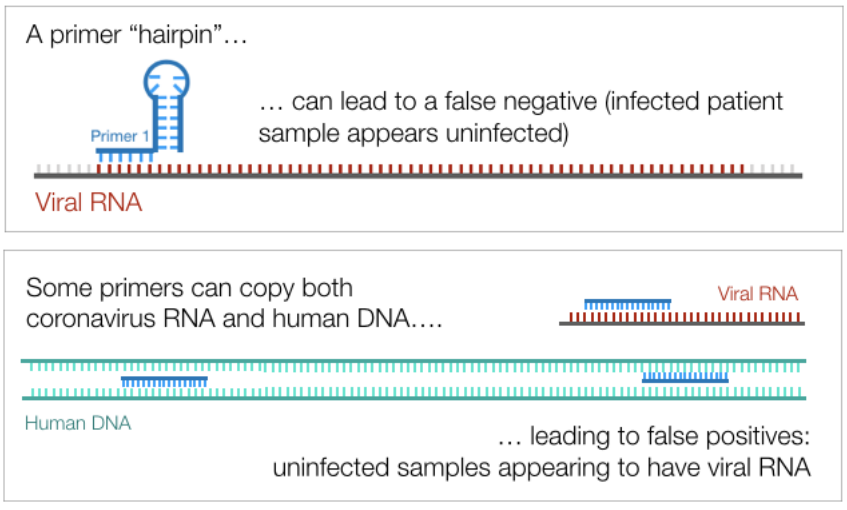
If the test is so simple, why is the U.S. having trouble getting it to work?

The U.S. initially mandated the use of CDC-developed test kits for all coronavirus testing, but labs reportedly had trouble getting them to work. The CDC was criticized for not using test kits developed in Germany, which were successfully detecting coronavirus around the world and were backed by WHO. U.S. labs responded by developing their own tests, and in some cases reporting quicker turnaround of results. This prompts the question: What are the differences between these tests and why do some work better than others?

The answer is relatively simple: Each test chose a different "sentence" to copy from the viral RNA. Effectively, this changes the primers (first and last word of the sentence), the fluorescent

probe (corresponding to some word in the middle), and the "positive control" (a tube of DNA containing the "sentence" that is used to make sure the test is working properly). In fact, the CDC's test kits are really just this: a few tubes containing the three components above.

Choosing primers for any PCR experiment turns out to be tricky and sometimes unpredictable. Primers are just short pieces of DNA themselves, and some DNA has a tendency to fold in on itself, creating a "hairpin" structure which inhibits PCR. (This is a bit like the matching letters in a palindrome finding one another). These "palindrome" primers can produce a false negative—an infected patient whose sample appears to lack the virus. Alternatively, the primers can work just fine to make copies of coronavirus RNA, but might *also* be capable of copying some part of human DNA. Because patient samples (most often nasal swabs) contain both viral particles and human cells, these primers can produce a false positive—an uninfected individual testing positive for the virus. Other potential sources of RT-PCR failure are temperature issues, low primer or sample concentration, and contamination, among others.



How a coronavirus test can fail. A "palindrome" primer can cause a false negative. Primers which can recognize human DNA can lead to a false positive.

Whether the CDC's primers were inferior to others' has been debated, but U.S. labs have another critical reason to develop their own tests: The CDC's test kits are in short supply. In principle, most molecular biology labs can develop such a test in a week or two, but those who have done so have come against another major hurdle: FDA regulations barring them from testing patient samples and returning results.

Federal regulations complicate in-house testing

Before we get into the weeds here, it is important to remind ourselves why FDA regulations exist: to protect the consumer—us—from being given incorrect medical information. Typically, there is regulatory oversight both of the laboratories where clinical tests are performed and of the tests themselves (though as this article points out, prior to this outbreak, FDA oversight of clinical tests under the current administration has been alarmingly slim).

In response to this outbreak, the FDA initially mandated use of the CDC's tests exclusively, but loosened restrictions when it became clear that the need severely outpaced the CDC's supply. Currently, CLIA-certified labs (those already certified to perform clinical testing) are permitted to develop and use their own tests, but must submit an Emergency Use Authorization application, which involves careful (and time-consuming) validation of in-house tests and sample collection procedures. Labs with extensive experience and infrastructure to perform viral PCR assays of clinical samples, but whose primary focus has historically been research rather than clinical diagnostics, must obtain CLIA certification as well. This added requirement temporarily prevented at least one well-equipped lab which played a central role in uncovering the extent of Seattle's outbreak from doing further testing.

It is a complicated matter that pits individual consumer protections against the country's immediate need to ramp up testing quickly in the face of a public health emergency. Stringent and time-consuming FDA requirements are preventing academic and clinical labs around the country, with capacity and willingness to develop and deploy testing within their communities, from being able to do so.

Though a path to certification exists, it is neither quick nor painless, and in a crisis that grows exponentially worse every day, where widespread testing has proven effective in curbing transmission, it is crucial that the federal government do everything in its power to allow willing labs to pick up the slack where it has ultimately failed: being prepared to mobilize quickly in the event of a public health crisis—a job that was formerly held by the pandemic preparedness team which Trump disbanded in 2018.

Massive supply shortages require creative solutions

Labs that manage to get proper certification to run clinical testing face another hurdle: a massive shortage of supplies. Patient samples are most commonly collected as nasal swabs, and before RT-PCR, viral RNA must be separated from mucous, human cells, and other debris. Commercially available RNA extraction kits are by far the quickest and safest way to process many samples at once, but unsurprisingly, demand has quickly outpaced supply, forcing testing labs to seek donations locally via social media.

PCR machines are another major bottleneck, as well as trained individuals needed to run them at capacity. Although some researchers argue that FDA restrictions should be lifted to allow all academic labs to fire up their idle PCR machines and start testing immediately, even they

recognize this would substantially increase the rate of erroneous results, leading us into uncharted and unpredictable territory. It seems more prudent for the FDA to move quickly to certify a set of well-equipped labs, which can train volunteer scientists to perform approved in-house tests at scale. On the West Coast, it's already all hands on deck. Both the University of Washington's School of Medicine and UC Berkeley's Innovative Genomics Institute are training willing academic researchers, graduate students, and postdocs to perform coronavirus testing.

Will drive-thru & at-home testing help?

First, let's clear up some confusion here. When it comes to coronavirus testing, "drive-thru" and "at-home" do not describe the test itself, which requires training and specialized equipment. These terms refer instead to *how* and *where* nasal swabs are collected. Though these strategies may not substantially increase the speed of testing, there may be immense public health benefits to performing sample collection via mail or a drive-thru point. Why? Because those who fear they are ill need not travel to a clinic, risking infecting others while there or in transit. Plans to implement at-home sample collection are already in progress, and drive-thru testing is already available for UW Medicine patients and staff. But regulatory hurdles exist in this domain as well. To process at-home tests, labs must provide substantial evidence that these samples are reliable relative to those collected by trained individuals, further hampering labs' ability to quickly roll out these operations.

Where do we go from here?

The challenges outlined here all converge around one conclusion: The U.S. was completely unprepared for a public health emergency of this scale. South Korea revamped its emergency

preparedness plans after the MERS outbreak of 2015, recognizing that early detection and isolation were effective to mitigate an outbreak, and putting resources and procedures into place which could be mobilized quickly.

Hopefully, the U.S. government learns from this catastrophe and diverts more resources toward emergency preparedness in the future. In the meantime, scientists are heroically doing what they can to pick up the slack, and it should be the government's immediate priority to simplify and accelerate regulatory procedures to permit qualified and well-equipped labs to scale up testing. To paraphrase Trevor Bedford, a virologist and leading voice reporting on the virus's predicted prevalence and expected trajectory, increasing testing capacity is crucial to reduce rates of transmission and get the coronavirus outbreak under control in the U.S.

As a scientist, I feel proud and encouraged by the quick and selfless responses of fellow scientists to extend testing, communicate important information to the public, and begin developing and testing vaccines at record speed. Ultimately, we are all in this together.

7.2 COVID-19 Testing & The Danger of a Quick Fix Narrative

This piece was published on Medium in October 2020. It can be found at https://tinyurl.com/y6vnsuek.

For months now, you have woken each day to a world that is both maddening and heartbreaking. You sip coffee and scroll through the day's stories, scanning for hope, bracing for despair, at times uncertain which one you were meant to feel.

One day you read a headline : "*Your Coronavirus Test Is Positive. Maybe It Shouldn't Be.*"

Huh.

You read on. A huge number of people testing positive carry "relatively insignificant amounts of the virus," it says. Most are "not likely to be contagious." On Twitter, an alarming message from the author: "90% (!!) of people who get a positive result are no longer contagious and don't need to isolate. Strap in, this is important."

This feels important, you think.

Your mom calls, excited. "Did you hear?" she says. "We've been measuring it all wrong. It's all way overblown. Can you make it home next month?" You vaguely recall a conversation from April: "No, mom, this isn't *no worse than the flu*."

But you are not sure what to make of this story. Our standard coronavirus test is too sensitive, it says. With a straightforward policy change, infection rates & needless isolations would plummet. But maybe we should downgrade this test altogether, it muses. Replace it with a less accurate test which returns fast results. Such tests already exist. And their poor sensitivity should be welcomed, not feared.

Huh.

You are skeptical. And you are not alone. As has almost always been the case over the last nine months, the story behind the story is a complicated one.

The Glossary (more pertinent than the average glossary; do read)

PCR test: The standard coronavirus test used across the world. Quite accurate—can detect even small amounts of virus. *Not* easily scalable—shortages of tests & testing facilities in the U.S.

Antigen test: A type of rapid test. Results in as little as 15 minutes. Can be cheap to make & simple to administer. Not nearly as accurate as PCR.

FDA: The regulatory agency responsible for evaluating new tests. Compares tests' accuracy to PCR for approval.

Dr. Michael Mina: Epidemiologist. Strong supporter of antigen tests. Critical of FDA's accuracy requirements.

Quick fixes: Proposed solutions which seem easy to implement & promise game-changing results. Easy to sell. Rarely pan out as promised.

Federal officials: Exasperated. Heavily criticized for pandemic response. Looking for novel solutions. The simpler, the better.

You: Exasperated. Skeptical of national pandemic response. Hoping for novel solutions. The quicker, the better.

The U.S.: A country on the precipice of a monumentally important election.

COVID-19 testing: where are we today?

In March, I wrote about the technical, logistical, and regulatory hurdles plaguing coronavirus testing in the U.S. As other countries established coordinated testing and contact tracing programs, the U.S. was just waking to the reality of a local pandemic for which it had no plan. States and communities were left scrambling to contain local outbreaks, largely shaping our haphazard testing landscape of today, where ill-equipped public health labs, universities, and private companies struggle together to scale up PCR testing against insurmountable demand. Half a year later, there are still long lines for testing, huge backlogs at private labs, and little national coordination of contact tracing. It is clear that the status quo isn't enough—that something needs to change.

In recent months, vast hope has gathered around the antigen test, a type of rapid test which delivers results in as little as 15 minutes and can be performed outside of a lab. It evokes a future where children are tested on their way into school, adults at work, the gym, the grocery store. Those who test positive are sent home; everyone else resumes a much more normal way of life. This vision is so enticing that XPrize is offering $5 million to anyone who can develop such a test.

But existing antigen tests cannot yet sustain this vision. Those which are fast and simple enough for such routine, ubiquitous use are not yet accurate enough for the job, returning both false positives and false negatives at concerningly high rates. Experts caution that test results would understandably be taken with a grain of salt if unreliable tests were delivered at such vast scales. Those testing positive but showing no symptoms might assume a false positive; those testing

negative would be given a false sense of security and perhaps forgo precautions like distancing and wearing masks.

But if one believes in this vision despite the tests' flaws, as does Dr. Mina, one faces another barrier to widespread adoption of existing tests: the FDA. To receive federal approval, a test must meet sensitivity and specificity standards which are set relative to PCR, the workhorse test used across the world. A test need not be *as* sensitive as PCR, but must reach a threshold which simple antigen tests have not yet been able to meet.

So, if one believes we should implement this vision *now*, they may need to convince the FDA to change its approval process. And one way to do so is by questioning the sensitivity of its gold standard: PCR.

Is PCR too sensitive?

Is PCR a sensitive test? Yes. It can detect very small amounts of virus.

Is this a good thing? In most cases, also yes. It can detect virus soon after infection, reducing the likelihood of transmission. It can detect virus from nasal samples, even with poor collection technique. It can detect virus in saliva, where less virus resides. It can detect virus in samples which have partially degraded during storage. It can detect virus transported in a variety of media. It can detect virus in samples containing large amounts of human cells and debris.

And, occasionally, it can detect the vestiges of a recent infection—residual bits of viral genetic information which cannot infect others.

Ah. You read about this as you drank your morning coffee. The story implied that the vast majority of people testing positive are in this post-infectious state.

But is this correct? Probably not.

Let me explain.

You carry little virus at two points during infection: at the onset and the tail end. The late infection period lasts comparatively longer than early infection, and thus the article reasons that those who carry little virus are likely beyond their contagiousness peak. Inspecting PCR data from several labs, they infer that many samples do in fact contain little virus.

But can we reliably estimate the amount of virus a person carries from PCR at all? This is the sticking point and one that has confused even scientists, particularly those who are used to well-controlled research experiments as opposed to finicky clinical samples.

The raw data from a PCR test is not a crisp positive or negative, but a number. Low numbers are associated with lots of virus. Very high numbers indicate no virus—a negative result. It is the middle ground numbers which are the source of confusion here. While a *relatively* high middle ground number *may* indicate little virus, it can also arise from intrinsic clinical variables like the time between sampling & processing, sampling technique, and sample consistency, to name a few. Generally, in clinical diagnostics, PCR's numerical output is not a great indicator of the amount of virus an individual carries.

But the story relies on these numbers anyway, designating a cutoff above which it infers "insignificant amounts of virus." It states neither the cutoff nor the reasoning behind it. This is

important, because it prevents comparison with data from other labs. If others *do not* observe similar rates of "high" PCR outputs, features specific to the surveying labs or geographic areas may be to blame. These include equipment, differences in sampling, transport, storage, and processing protocols, and perhaps most importantly, the local availability of testing. How long after an individual requested testing were they able to get it? If wait times are long, then a bias towards late infections may be expected.

In short, we cannot reliably infer "rates of contagiousness" from PCR's numerical output, let alone generalize "nationwide" from these data. The 90% of individuals the story deems non-contagious likely span all stages of infection. Suggesting they need not isolate can have devastating consequences.

The danger of a quick fix narrative

Simple solutions sell. Against a backdrop of a global pandemic and a monumentally important election, the illusion of a quick fix can be immensely alluring—and enormously dangerous.

Dr. Mina calls for policy change to place a stringent cap on the PCR values associated with a positive result. This change would substantially reduce reported infection rates in the short term, leading to the illusion of a pandemic under control. In the long term, viral transmission would spike, as infected individuals who receive the all clear forgo isolation and contact tracing.

But despite an absence of data linking these values to coronavirus transmissibility and a cautionary statement from the College of American Pathologists, federal officials at the CDC are considering this proposal "for policy decisions."

Simple solutions sell.

Referencing Dr. Mina, other media sources have presented the sweeping adoption of antigen tests as a no-brainer. Rapid tests suitable for everyday at-home testing already exist, they write, referring to paper strip antigen tests as "the wand that will accomplish this feat," and declaring there is "no technical obstacle…only a dearth of political will." Dr. Mina himself writes that strip tests can be "mass produced in a matter of weeks and freely supplied by the government to everyone in the country."

But many experts warn that existing antigen tests are not yet suitable for the sort of widespread surveillance testing Dr. Mina aspires to. One study found that testing all grade school students three times per week would produce nearly 800,000 false positives weekly. With tens of millions of people tested per day, positive results may become obsolete as people forgo what they perceive to be needless isolation. Negative results are also dubious, with the CDC recommending that anyone with symptoms or known exposure "confirm negative antigen test results with an RT-PCR test." Supply chain issues, which plague PCR testing, are also bound to arise at the scale of testing Dr. Mina proposes. Though ubiquitous rapid testing is an enticing idea, we do not yet have the tests to make it a reality.

Yet despite mounting reports of problems with Abbott's antigen tests, most recently tied to the outbreak at the White House, the government has purchased 150 million rapid tests from the company, broadcasting this move as "President Trump's all-of-America approach to constructing our world-leading COVID-19 testing capacity," and "fully leveraging America's industry and innovative spirit." In July, the government distributed rapid point-of-care tests across nursing homes, the use of which was recently discontinued due to inaccuracies.

Simple solutions sell, not just to federal officials, but to us. With the election weeks away, the current administration has a strong incentive to invest in quick fixes which have been sold to the public as panaceas—and we have little time to see them fall short of their promises.

Fast, frequent, and ubiquitous testing *would* be a game-changer, as Dr. Mina widely proclaims. It is a worthwhile target to aim for, even if we never fully reach it. But striving to get there by eroding public trust in the only reliable test we currently have is short-sighted. And using misleading messaging to unite the public around an untried alternative will only deepen apathy in the long run.

Simple solutions sell.

But simple solutions distract us from coordinating a complex and multifaceted response befitting such a complex and multifaceted problem. The real solution places at least as much emphasis on public trust as it does on "techno-fixes." It makes use of multiple forms of testing, carefully considering the interplay between sensitivity, frequency, and the integral role of public confidence and buy-in. It invests in contact tracing and promotes consistent, science-backed messaging around mask wearing and social distancing. It keeps us informed about novel treatments & vaccines, without cultivating false hope.

But simple solutions sell. And so do simple stories.

Chapter 8. DISCUSSION

Here, I have shown how tethering single cell expression and chromatin accessibility profiles via lineage relationships facilitates the detection and characterization of heritable gene expression

changes. Surprisingly, even in a non-differentiating cell line, I observed abundant, progressively-acquired heritable expression changes. Some differentially expressed genes had an obvious genetic origin−copy number changes impacting multiple adjacent genes, while many others showed stable, lineage-associated expression but with less clear origins. The explanations for this latter category might include epigenetic changes within nearby regulatory sites, changes in abundance of upstream regulators, the acquisition of new regulatory contexts via genomic rearrangements, and/or focal genetic changes, amplifications, or deletions. Above, I have shown that our approach of profiling multiple features in closely related cells can, at least in some cases, be used to distinguish between these possibilities.

8.1 PROPOSED APPLICATIONS OF THE LORAX ACROSS BIOLOGICAL SYSTEMS

Clonal tracking, achieved via various methods across diverse systems, has revealed the presence of biologically important heritable states. For example, combining Luria-Delbrück fluctuation analysis with RNA-seq, Shaffer *et al.* found rare, but clonally stable expression states which predisposed cancer cells to drug resistance (Shaffer et al., 2020). Intriguingly, these states were in some cases reversible, suggesting an epigenetic origin. Goyal *et al.* confirmed the presence of clone-specific responses of cancer cells to various drug treatments using a clonal barcoding approach (FateMap) (Goyal et al. 2021). Mold *et al.* made use of 'natural' clonal barcodes−−T-cell receptors in lymphocytes−and found that clonal lymphocytes responded more similarly to vaccination than more distantly related cells (Mold et al., 2022). Using an *in vivo* transgenic barcoding strategy (TREX, (Ratz et al. 2021)), they found that in mouse neurons, gene expression states mimicked clonal structure, even among different clones of the same cell type. Finally, He *et al.* investigated the timing of cell fate restriction in organoids with iTracer, a system which includes

an initial and an induced round of clonal barcoding (He et al., 2021). These studies present intriguing examples of heritable expression but are limited in terms of fully distinguishing between potential underlying causes.

I envision that THE LORAX may be applied to such systems, enhancing our ability to detect heritable events and explain their mechanistic origins. First, progressive lineage labeling increases the likelihood of detecting rare heritable events, as finer-scale, temporally-resolved clonal labeling produces more homogenous clones. Progressive labeling may be particularly useful for detecting events which are stable over multiple cell divisions but reversible, since both the acquisition and reversal may be captured via a finely-tuned lineage recording system. Second, the addition of a chromatin accessibility measurement alongside clonal labels may help resolve the mechanisms behind clonal expression stability. Genetically-mediated expression variation is likely during cancer progression, where copy number changes ((Harbers et al., 2021), loss of heterozygosity (Nichols et al., 2020), and chromothripsis (Cortés-Ciriano et al., 2020)−widespread fragmentation and reassembly of genetic material−are commonly observed. I have shown above that such events may be inferred using our approach. On the other hand, myriad epigenetic changes accompany cell fate commitment during organoid and organism development, and concurrent lineage tracing and RNA and ATAC profiling in closely related cells may illuminate the order of events which give rise to progressive cell type divergence (Thomas et al., 2011). In these systems and others where cell state diversification is taking place, it is likely that lineage-resolved ATAC-seq will show clone-specific enhancer and promoter accessibility changes beyond what I observed here, which may explain heritable expression variation. In fact, profiling clonal T-cell populations expanded *in vitro* using bulk ATAC- and RNA-seq, Mold *et al.* found clone-specific accessibility

changes at regulatory regions, with enrichment near clonally differentially expressed genes (Mold et al., 2022).

8.2 ADVANCES IN LINEAGE TRACING PRESENTED IN THIS WORK

Our work presents some advances in CRISPR-based lineage tracing, and also highlights some fresh challenges. First, encoding lineage at many independently-integrated loci rather than at tandem loci expressed as a single transcript eliminates the chance that a large deletion removes neighboring CRISPR targets, supports larger deletions, and enables efficient capture of larger insertions. These features in turn reduce both the rate of missing lineage information and the probability of convergence events. Second, I show that NN-based inference of missing data in individual cells and subsequent pooling of cells to generate "consensus" profiles prior to lineage reconstruction (and iteratively generating subtrees from these consensus groups) reduces the likelihood of misplaced cells early in the reconstruction process. Though I demonstrate the usefulness of this approach when a "greedy" algorithm is used for reconstruction, it is applicable even to methods which primarily use traditional phylogenetic reconstruction approaches (*e.g.* maximum likelihood) (Gong et al., 2021; Jones et al., 2020; Konno et al., 2022), since the sheer number of cells often makes early "greedy" subgrouping necessary. Finally, these lineage recording and analysis approaches are compatible with other recent advances in lineage recording technology, like DNA Ticker Tape (Choi et al., 2021), where successive insertions as a single locus greatly simplify ordering of lineage-encoding events. Integrating multiple such loci would enable higher resolution trees, and the approaches presented here can be used to order events occurring at distinct recording loci, where event ordering is not so straightforward.

8.3 OUTSTANDING CHALLENGES IN LINEAGE TRACING

Our work also highlights some unresolved challenges within the CRISPR-based lineage tracing field. First, fine control of editing rate remains elusive; I observed abundant editing in some lineages, while most targets in others remained unused. Loss or silencing of the Cas9-expressing genomic locus might explain lineage-specific reductions in editing efficiency, while position effect variegation in cutting or editing rates might explain variation in usage or recovery across targets. Second, though I observed a great diversity of editing patterns, they are not evenly distributed, with the top three edits frequently occurring independently. This phenomenon can in part be addressed with careful target design to avoid regions of microhomology (W. Chen et al., 2019; Sfeir & Symington, 2015). Third, though the design of our construct allows for large indels relative to other methods, relying on double strand break repair for editing diversity still presents a risk that a recorded event will not be reliably captured due to indel size. Finally, frequent DSBs (which may themselves be contributing to the CNAs observed here), and the persistently high expression of transgenes (which are prone to silencing) may not be compatible with organismal or ES cell-derived systems. Excitingly, these challenges are addressed in large part by DNA Ticker Tape, which leverages prime editing to introduce diverse insertional edits to a target site in an ordered manner, without requiring double-stranded breaks (Choi et al., 2021).

8.4 THE FUTURE OF LINEAGE TRACING AND RECONSTRUCTION METHODS

The last few years has brought rapid advances in genome editing tools which will pave the way for more refined lineage tracing technologies. CRISPR-Cas9 emerged as a frontrunner in lineage tracing due to its ability to induce a diversity of outcomes at a single locus. The downside to this approach, of course, is the lack of control over this set of outcomes, with some outcomes (e.g.

large insertions and deletions) being frequent sources of missing data and thus inaccurate lineage reconstruction.

Prime editing (Anzalone et al., 2019) offers incredible promise in the lineage tracing field. This method uses catalytically inactive Cas9 fused to reverse transcriptase to enable high-efficiency template-directed repair. With respect to lineage recording, this method has several important advantages. First, it does not induce double strand breaks, reducing the likelihood that the editing itself may influence a cell's biological state. Second, it enables *controlled* editing outcomes, which can be programmed to accommodate capture constraints (i.e. size limits).

Choi et al. 2021 (Choi et al., 2021) devises a beautiful lineage tracing strategy which highlights another less obvious benefit of prime editing: predictable editing enables engineered conditional editing, such that an edit cannot be made prior to another one. Choi et al. does just this. They design a system in which 3 base pair insertions are continually made to a target array, but an insertion cannot be made at target #2 if one has not already occurred at target #1.

This structure is ingenious for the purpose of lineage reconstruction, as order of events within a single array is known. Choi et al. integrate multiple such arrays into the genome and capture them at high efficiency using droplet-based capture. High editing rates mean that each cell has more information available for reconstruction.

Such an approach asks again for the continued evolution of reconstruction methods. An ideal reconstruction approach would account for known ordering within a target array, but unknown

ordering of edits between different arrays. Traditional phylogenetics approaches (e.g. as implemented in Cassiopeia (Jones et al. 2020)) require modifications in order to account for this new feature. The reconstruction approach I present here can in theory be easily modified to encode this feature, by considering only the earliest unused edit in each target array for the next splitting step.

A drawback to my approach is that erroneous data and convergence events may easily cause the misplacement of cells. Though this is somewhat mitigated by the strategy described above where only a subset of targets are considered at each split, it remains a problem as long as editing diversity is limited and/or distributed unevenly. Neighbor joining approaches hold a lot of promise as lineage profiles become more complex and complete. Though resulting trees do not immediately track with specific editing events, one can imagine reconstructing an order of events retroactively, by automatically inferring commonly shared editing patterns are one iteratively moves up the tree.

8.4.1 *Improved methods for edit calling*

In my work, I took great care to make sure PCR and sequencing errors did not result in an artificial diversity of editing patterns. A recently developed package, TraceQC (Hu et al., 2020), formally addresses these considerations. It introduces improved indel calling in linage data, including gap opening and extension penalties which bias towards insertion and deletions, as opposed to substitutions. Read counts and alignment scores are also considered to reduce technical noise. Like our approach, they incorporate consensus sequence calling to further reduce noise.

8.4.2 *Modeling lineage tracing processes: editing rate*

Evaluating reconstruction methods remains a challenge, as simulated datasets often do not reflect true variables in the data. The data presented here highlights a few of these challenges. First, it is not necessarily true that editing occurs at a consistent rate across all lineages. In fact, our data and that of others (the other tree paper) suggests, at least in cell lines, that editing accumulates much quicker in some lineages than others. I observed some lineages which acquired very few edits overall, which, in our system, may be due to the loss of the Cas9 genomic locus as part of a karyotypic changes. One can imagine that in a cell line that can tolerate karyotypic changes, those lineages which lose early Cas9 (and thus are no longer experiencing Cas9-induced double strand breaks) will be selected for over others. This may explain the large number of cells in groups 25-30 in our tree, where a relatively small number of edits define the lineages. On the other hand, I observe lineages where a large number of targets are edited (e.g. 12, 22-24). Were editing rate consistent and selection not present, I would expect a gradation of editing, where one or a few edits define a group which accumulates additional, diverse edits over time. Though this occurs to some degree in groups 22-24 (additional edits are observed beyond the shared ones within each group), the ten edits which unite these groups do not appear elsewhere in the tree, which would be expected if they accumulated gradually. Instead, such a pattern is suggestive of a process in which Cas9 activity was high in a single founder cell, generating multiple edits in a single generation. Such high activity may result from an elevated concentration of Cas9 in the cell, due perhaps to higher expression or a genomic duplication of the Cas9 locus, or it may result from a weakening of the DSB repair machinery, increasing the likelihood that a cut will be repaired imperfectly.

These examples illustrate that a complex interplay of factors underpin editing rate. Further factors may need to be considered in systems where differentiation occurs (e.g. during embryonic or

organoid development). For example, Cas9 expression and repair dynamics may not be consistent across cell types, potentially influenced by factors such as cell division rate, and primary method of repair (Featherstone and Jackson 1999).

8.4.3 *Modeling lineage tracing processes: information loss*

Perhaps the most complicated and experiment-specific challenges lie in modeling the rate of information loss, both during recording and recovery of lineage profiles. Some methods assume a random distribution of missing data, stochastically removing subsets of the simulated data to evaluate the how robust the method is to incomplete recovery (Gong et al., 2021; Konno et al., 2022). However, the distribution of missing data is highly dependent on the process which led to its loss, and experiment-specific factors may not be obvious until after the data has been collected. In our dataset, I observed data loss occurring at several points in the recording and recovery process.

First, during the CRISPR/Cas9 DSB repair process, large deletions may remove either the primer binding site(s), or the barcode necessary to associate an edit with a particular target. In the latter case, even if the transcript associated with the target can be captured, it will be discarded. If this deletion occurs early, missing data will be observed at as particular target, specific to a particular lineage.

Second, large deletions and insertions which do not remove necessary capture and target-associated information may be more difficult to capture via the combinatorial indexing method developed here, due to size selection (large deletions) and clustering limitations during sequencing

(large insertions). Such events will appear as incomplete loss specific to a single target within a lineage. Importantly, these events can be computationally distinguished from other (stochastic) types of loss, since a large insertion or deletion will be observed in the minority of cells in which the target information is retained. This property was used in our study to manually correct a subset of the missing data where large insertions or deletions were likely the culprit, but one can imagine automating this process, by quantifying a relationship between insertion/deletion size and capture efficiency.

Third, as described above, karyotypic changes over the course of the experiment result in non-random information loss.

8.4.4 *Other unsolved problem in lineage tracing*

A still unsolved challenge is the control of editing rate. Several groups have found promise in making mismatches between gRNAs and their targets, creating a set of early and late editors (McKenna et al. 2016; Simeonov et al. 2021; Chan et al. 2019). Others have used inducible systems to control Cas9 expression (Raj, Gagnon, and Schier 2018). As observed in my work and those of others (Simeonov et al., 2021), doxycycline-inducible systems alone are insufficient, as even small amounts of drug result in a very high level of editing. Combining such a system with another approach which modulates Cas9 protein levels may be of use. I have begun work on several such approaches (unpublished work), including modifying the transcription start site of Cas9, such that the transcript is not made as efficiently, and tethering Cas9 to an auxin-inducible degron (Yesbolatova et al., 2020), creating multiple control switches at which Cas9 levels can be

controlled. One can also imagine engineering a system where Cas9 expression or availability is tied to cell cycle, such that edits can be used to track the rate of lineage expansion.

A second challenge not observed in my experimental system but reported frequently in more complex systems is transgene silencing. In order to capture lineage information from mRNA, a high level of expression must be maintained across cell types in a system where cell fate divergence occurs. Several modification to the transgene delivery system and construct hold promise, including using piggyBac as a delivery vector (X. Chen et al. 2015) and flanking transgenes with insulator sequences (Pérez-González and Caro 2019) to prevent epigenetic modification.

8.5 APPLYING THE LOGICAL CORE OF THE LORAX OUTSIDE OF LINEAGE TRACING

8.5.1 *Compatibility of THE LORAX with other barcoding systems*

The logical core of THE LORAX–pooling cells based on genetically-encoded labels captured alongside multiple genomic and/or epigenetic features to evaluate the relationship between those features–is broadly applicable to any system amenable to genetic barcoding. Systems where static barcodes (*e.g*. CellTag (Guo et al., 2019)) are used to interrogate clone-specific heterogeneity, are an obvious candidate, but labels need not necessarily mark clonal populations. For example, sgRNAs in CRISPR perturbation screens can be used to tether multiple single cell molecular measurements. Importantly, combinatorial indexing approaches are not required here, as both short barcode integrants and sgRNAs can now be captured alongside scRNA-seq (Cao et al., 2018; S. Chen et al., 2019; L. Liu et al., 2019; Ma et al., 2020; Xing et al., 2020; Zhu et al., 2019) and scATAC-seq (Pierce et al., 2021; Replogle et al., 2020; Rubin et al., 2019) via droplet-based methods.

8.5.2 *Lineage tethering of multiple features as an alternative to co-assays*

In some applications, THE LORAX has several advantages over traditional co-assays of expression and accessibility where both features are measures in the same single cells (Cao et al., 2018; S. Chen et al., 2019; L. Liu et al., 2019; Ma et al., 2020; Xing et al., 2020; Zhu et al., 2019), as well as computational integration methods which merge single cell expression and accessibility datasets (Y. Lin et al., 2021; Stuart et al., 2019). First, existing co-assay methods are relatively low resolution compared with methods which profile each feature separately; thus, associating single-feature profiles via lineage relationships improves resolution at the single cell level. Second, by aggregating profiles of closely related cells, I achieve higher statistical power to detect even rare, heritable events. Third, though computational integration is possible in datasets composed of a variety of cell *types*, it is less feasible in ones composed of different cell *states* where well-separated clusters are not expected and stochastic factors often drive within-cluster positioning. THE LORAX enables overlaying of expression and accessibility datasets without making *a priori* assumptions about their relationship, as is necessary during computational integration.

8.6 CONCLUSION

In summary, I have shown that (a) progressive recording of lineage information across distinct genomic loci, and their high rate of recovery alongside sci-RNA-seq, enables accurate reconstruction of cell lineage trees; (b) aggregating expression profiles of closely related cells reveals abundant, and progressively acquired heritable expression variation, even in non-differentiating cells; and (c) we can investigate the relationship between multiple features−like

expression and chromatin accessibility−by tethering them via concurrently captured lineage profiles.

8.7 CLOSING REMARKS

Reflecting on the work I have produced in graduate school, I note that an inclination towards simplicity has been a driving force in my development as a scientist. I leave you with a few closing remarks outlining my hopes for the development of our field.

When extracting meaning out of large datasets, visualization is often key. Across scientific disciplines, tools which enable direct observation of previously unobservable phenomenon (e.g. microscopes, telescopes, X-rays) have transformed entire fields. Our desire to *see* our object of study is unsurprising: our eyes are the most sophisticated pattern finding tool we have. With the explosion of high throughput sequencing and single cell methods, our ability to collect large amounts of data quickly has skyrocketed. Visualization methods – UMAP, pseudotime trajectories, RNA velocity, to name a few – have followed. But the interpretability of these visual outputs relative to the data is a major challenge in the field (Lange et al. 2022). Yet it has become all too easy – expected, even—to perform a standard set of analyses on single cell datasets, generating beautiful but often information-limited (or, in the worst case, misleading) visualizations, which are highly dependent on a set of complex and subjective parameter choices. Our fine-tuned ability to see patterns has become a liability in the face of tools so vulnerable to visual artifacts.

As our data has grown more complex, I fear that we have placed undue value on complex visual outputs. I believe the ability to envision and construct simple visualizations from scratch holds

incredible power – it gives us the confidence to start with the questions we have rather than relying on the questions previously asked. It is my hope that we see growing emphasis on the development of these skills during scientific training and stive for simplicity and interpretability in our visuals.

My second hope is that, in the field of genomics technology development, we do not lose sight of the power of simple systems to optimize complex technologies. Rapid advances in genome and epigenome manipulation have paved the way for ingenious *ideas* (e.g. mutation-based lineage tracing, massively parallel assays), astonishingly beautiful in their logical simplicity and tremendous potential, but exceedingly difficult to actualize fully. I strongly suspect that the potent incentives to rapidly apply nascent technologies to increasingly complex systems, and the relative de-emphasis on technical optimization in simple systems, ultimately lengthens our path to fully realizing their potential.

It is an incredibly exciting time to be working in the field of genomics. I anticipate continued and rapid refinement of technologies in lineage tracing and single cell profiling, among others, and look forward to seeing the breadth of biological insight to which they will inevitably give rise.

BIBLIOGRAPHY

Alemany, Anna, Maria Florescu, Chloé S. Baron, Josi Peterson-Maduro, and Alexander van Oudenaarden. 2018. "Whole-Organism Clone Tracing Using Single-Cell Sequencing." *Nature* 556 (7699): 108–12.

Anzalone, Andrew V., Peyton B. Randolph, Jessie R. Davis, Alexander A. Sousa, Luke W. Koblan, Jonathan M. Levy, Peter J. Chen, et al. 2019. "Search-and-Replace Genome Editing without Double-Strand Breaks or Donor DNA." *Nature* 576 (7785): 149–57.

Ben-David, Uri, Benjamin Siranosian, Gavin Ha, Helen Tang, Yaara Oren, Kunihiko Hinohara, Craig A. Strathdee, et al. 2018. "Genetic and Transcriptional Evolution Alters Cancer Cell Line Drug Response." *Nature* 560 (7718): 325–30.

Biddy, Brent A., Wenjun Kong, Kenji Kamimoto, Chuner Guo, Sarah E. Waye, Tao Sun, and Samantha A. Morris. 2018. "Single-Cell Mapping of Lineage and Identity in Direct Reprogramming." *Nature* 564 (7735): 219–24.

Bonasio, Roberto, Shengjiang Tu, and Danny Reinberg. 2010. "Molecular Signals of Epigenetic States." *Science* 330 (6004): 612–16.

Bowling, Sarah, Duluxan Sritharan, Fernando G. Osorio, Maximilian Nguyen, Priscilla Cheung, Alejo Rodriguez-Fraticelli, Sachin Patel, et al. 2020. "An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells." *Cell* 181 (6): 1410-1422.e27.

Cao, Junyue, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, et al. 2018. "Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells." *Science* 361 (6409): 1380–85.

Cao, Junyue, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, et al. 2017. "Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism." *Science* 357 (6352): 661–67.

Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M. Ibrahim, Andrew J. Hill, Fan Zhang, et al. 2019. "The Single-Cell Transcriptional Landscape of Mammalian Organogenesis." *Nature* 566 (7745): 496–502.

Chan, Michelle M., Zachary D. Smith, Stefanie Grosswendt, Helene Kretzmer, Thomas M. Norman, Britt Adamson, Marco Jost, et al. 2019. "Molecular Recording of Mammalian Embryogenesis." *Nature* 570 (7759): 77–82.

Chen, Song, Blue B. Lake, and Kun Zhang. 2019. "High-Throughput Sequencing of the Transcriptome and Chromatin Accessibility in the Same Cell." *Nature Biotechnology* 37 (12): 1452–57.

Chen, Wei, Aaron McKenna, Jacob Schreiber, Maximilian Haeussler, Yi Yin, Vikram Agarwal, William Stafford Noble, and Jay Shendure. 2019. "Massively Parallel Profiling and Predictive Modeling of the Outcomes of CRISPR/Cas9-Mediated Double-Strand Break Repair." *Nucleic Acids Research* 47 (15): 7989–8003.

Chen, Xiang, Jing Cui, Zhengjian Yan, Hongmei Zhang, Xian Chen, Ning Wang, Palak Shah, et al. 2015. "Sustained High Level Transgene Expression in Mammalian Cells Mediated by the Optimized PiggyBac Transposon System." *Genes & Diseases* 2 (1): 96–105.

Choi, Junhong, Wei Chen, Anna Minkina, Florence M. Chardon, Chase C. Suiter, Samuel G. Regalado, Silvia Domcke, et al. 2021. "A Temporally Resolved, Multiplex Molecular

Recorder Based on Sequential Genome Editing." *BioRxiv*. https://doi.org/10.1101/2021.11.05.467388.

Choi, Yoon Ha, and Jong Kyoung Kim. 2019. "Dissecting Cellular Heterogeneity Using Single-Cell RNA Sequencing." *Molecules and Cells* 42 (3): 189–99.

Chow, Ke-Huan K., Mark W. Budde, Alejandro A. Granados, Maria Cabrera, Shinae Yoon, Soomin Cho, Ting-Hao Huang, et al. 2021. "Imaging Cell Lineage with a Synthetic Digital Recording System." *Science* 372 (6538). https://doi.org/10.1126/science.abb3099.

Cortés-Ciriano, Isidro, Jake June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L. Jung, Lixing Yang, Dmitry Gordenin, et al. 2020. "Comprehensive Analysis of Chromothripsis in 2,658 Human Cancers Using Whole-Genome Sequencing." *Nature Genetics* 52 (3): 331–41.

Costello, Alan, Nga T. Lao, Clair Gallagher, Berta Capella Roca, Lourdes A. N. Julius, Srinivas Suda, Jens Ducrée, et al. 2019. "Leaky Expression of the TET-On System Hinders Control of Endogenous MiRNA Abundance." *Biotechnology Journal* 14 (3): e1800219.

Cusanovich, Darren A., Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. 2015. "Multiplex Single Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing." *Science* 348 (6237): 910–14.

Cusanovich, Darren A., Andrew J. Hill, Delasa Aghamirzaie, Riza M. Daza, Hannah A. Pliner, Joel B. Berletch, Galina N. Filippova, et al. 2018. "A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility." *Cell* 174 (5): 1309-1324.e18.

Datlinger, Paul, André F. Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C. Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. 2017. "Pooled CRISPR Screening with Single-Cell Transcriptome Readout." *Nature Methods* 14 (3): 297–301.

Dixit, Atray, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, et al. 2016. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens." *Cell* 167 (7): 1853-1866.e17.

Featherstone, C., and S. P. Jackson. 1999. "DNA Double-Strand Break Repair." *Current Biology: CB* 9 (20): R759-61.

Felsenstein, J. 2009. "PHYLIP (Phylogeny Inference Package) Version 3.7a." *Distributed by the Author. Department of Genome Sciences, University of Washington, Seattle.*

Feng, Jean, William S. DeWitt III, Aaron McKenna, Noah Simon, Amy D. Willis, and Frederick A. Matsen IV. 2021. "Estimation of Cell Lineage Trees by Maximum-Likelihood Phylogenetics." *The Annals of Applied Statistics* 15 (1): 343–62.

Floris, Chiara, Stefania Rassu, Loredana Boccone, Daniela Gasperini, Antonio Cao, and Laura Crisponi. 2008. "Two Patients with Balanced Translocations and Autistic Disorder: CSMD3 as a Candidate Gene for Autism Found in Their Common 8q23 Breakpoint Area." *European Journal of Human Genetics: EJHG* 16 (6): 696–704.

Gong, Wuming, Alejandro A. Granados, Jingyuan Hu, Matthew G. Jones, Ofir Raz, Irepan Salvador-Martínez, Hanrui Zhang, et al. 2021. "Benchmarked Approaches for Reconstruction of in Vitro Cell Lineages and in Silico Models of C. Elegans and M. Musculus Developmental Trees." *Cell Systems* 12 (8): 810-826.e4.

Gong, Wuming, Hyunwoo J. Kim, Daniel J. Garry, and Il-Youp Kwak. 2022. "Single Cell Lineage Reconstruction Using Distance-Based Algorithms and the R Package, DCLEAR." *BMC Bioinformatics* 23 (1): 103.

Goyal, Yogesh, Ian P. Dardani, Gianna T. Busch, Benjamin Emert, Dylan Fingerman, Amanpreet Kaur, Naveen Jain, et al. 2021. "Pre-Determined Diversity in Resistant Fates Emerges from Homogenous Cells after Anti-Cancer Drug Treatment." *BioRxiv*. https://doi.org/10.1101/2021.12.08.471833.

Granja, Jeffrey M., M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. 2021. "ArchR Is a Scalable Software Package for Integrative Single-Cell Chromatin Accessibility Analysis." *Nature Genetics* 53 (3): 403–11.

Guo, Chuner, Wenjun Kong, Kenji Kamimoto, Guillermo C. Rivera-Gonzalez, Xue Yang, Yuhei Kirita, and Samantha A. Morris. 2019. "CellTag Indexing: Genetic Barcode-Based Sample Multiplexing for Single-Cell Genomics." *Genome Biology* 20 (1): 90.

Harbers, Luuk, Federico Agostini, Marcin Nicos, Dimitri Poddighe, Magda Bienko, and Nicola Crosetto. 2021. "Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas." *Frontiers in Oncology* 11 (July): 700568.

He, Zhisong, Ashley Maynard, Akanksha Jain, Tobias Gerber, Rebecca Petri, Hsiu-Chuan Lin, Malgorzata Santel, et al. 2021. "Lineage Recording in Human Cerebral Organoids." *Nature Methods*, December. https://doi.org/10.1038/s41592-021-01344-8.

Hota, Swetansu K., Andrew P. Blair, Kavitha S. Rao, Kevin So, Aaron M. Blotnick, Ravi V. Desai, Leor S. Weinberger, Irfan S. Kathiriya, and Benoit G. Bruneau. 2020. "Chromatin Remodeler Brahma Safeguards Canalization in Cardiac Mesoderm Differentiation." *BioRxiv*. https://doi.org/10.1101/2020.06.03.132654.

Hu, Jingyuan, Rami Al-Ouran, Xiang Zhang, Zhandong Liu, and Hyun-Hwan Jeong. 2020. "TraceQC: An R Package for Quality Control of CRISPR Lineage Tracing Data." *BioRxiv*. https://doi.org/10.1101/2020.06.05.137141.

Hwang, Byungjin, Wookjae Lee, Soo-Young Yum, Yujin Jeon, Namjin Cho, Goo Jang, and Duhee Bang. 2019. "Lineage Tracing Using a Cas9-Deaminase Barcoding System Targeting Endogenous L1 Elements." *Nature Communications* 10 (1): 1234.

Jinek, Martin, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. 2012. "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity." *Science* 337 (6096): 816–21.

Jones, Matthew G., Alex Khodaverdian, Jeffrey J. Quinn, Michelle M. Chan, Jeffrey A. Hussmann, Robert Wang, Chenling Xu, Jonathan S. Weissman, and Nir Yosef. 2020. "Inference of Single-Cell Phylogenies from Lineage Tracing Data Using Cassiopeia." *Genome Biology* 21 (1): 92.

Kalhor, Reza, Kian Kalhor, Leo Mejia, Kathleen Leeper, Amanda Graveline, Prashant Mali, and George M. Church. 2018. "Developmental Barcoding of Whole Mouse via Homing CRISPR." *Science* 361 (6405). https://doi.org/10.1126/science.aat9804.

Kalhor, Reza, Prashant Mali, and George M. Church. 2017. "Rapidly Evolving Homing CRISPR Barcodes." *Nature Methods* 14 (2): 195–200.

Kester, Lennart, and Alexander van Oudenaarden. 2018. "Single-Cell Transcriptomics Meets Lineage Tracing." *Cell Stem Cell* 23 (2): 166–79.

Kiani, Karun, Eric M. Sanford, Yogesh Goyal, and Arjun Raj. 2022. "Changes in Chromatin Accessibility Are Not Concordant with Transcriptional Changes for Single-Factor Perturbations." *BioRxiv*. https://doi.org/10.1101/2022.02.03.478981.

Kiselev, Vladimir Yu, Tallulah S. Andrews, and Martin Hemberg. 2019. “Challenges in Unsupervised Clustering of Single-Cell RNA-Seq Data.” *Nature Reviews. Genetics* 20 (5): 273–82.

Konno, Naoki, Yusuke Kijima, Keito Watano, Soh Ishiguro, Keiichiro Ono, Mamoru Tanaka, Hideto Mori, et al. 2022. “Deep Distributed Computing to Reconstruct Extremely Large Lineage Trees.” *Nature Biotechnology*, January, 1–10.

Kretzschmar, Kai, and Fiona M. Watt. 2012. “Lineage Tracing.” *Cell* 148 (1–2): 33–45.

Lange, Marius, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, et al. 2022. “CellRank for Directed Single-Cell Fate Mapping.” *Nature Methods* 19 (2): 159–70.

Li, Zhenhua, Quanli Yang, Xin Tang, Yiming Chen, Shanshan Wang, Xiaojie Qi, Yawen Zhang, et al. 2022. “Single-Cell RNA-Seq and Chromatin Accessibility Profiling Decipher the Heterogeneity of Mouse Γδ T Cells.” *Science Bulletin of the Faculty of Agriculture, Kyushu University* 67 (4): 408–26.

Lin, Yao-Cheng, Morgane Boone, Leander Meuris, Irma Lemmens, Nadine Van Roy, Arne Soete, Joke Reumers, et al. 2014. “Genome Dynamics of the Human Embryonic Kidney 293 Lineage in Response to Cell Biology Manipulations.” *Nature Communications* 5 (September): 4767.

Lin, Yingxin, Tung-Yu Wu, Sheng Wan, Jean Y. H. Yang, Wing H. Wong, and Y. X. Rachel Wang. 2021. “ScJoint: Transfer Learning for Data Integration of Atlas-Scale Single-Cell RNA-Seq and ATAC-Seq.” *BioRxiv*. https://doi.org/10.1101/2020.12.31.424916.

Liu, Longqi, Chuanyu Liu, Andrés Quintero, Liang Wu, Yue Yuan, Mingyue Wang, Mengnan Cheng, et al. 2019. “Deconvolution of Single-Cell Multi-Omics Layers Reveals Regulatory Heterogeneity.” *Nature Communications* 10 (1): 470.

Liu, Pengyuan, Carl Morrison, Liang Wang, Donghai Xiong, Peter Vedell, Peng Cui, Xing Hua, et al. 2012. “Identification of Somatic Mutations in Non-Small Cell Lung Carcinomas Using Whole-Exome Sequencing.” *Carcinogenesis* 33 (7): 1270–76.

Loveless, Theresa B., Joseph H. Grotts, Mason W. Schechter, Elmira Forouzmand, Courtney K. Carlson, Bijan S. Agahi, Guohao Liang, et al. 2021. “Lineage Tracing and Analog Recording in Mammalian Cells by Single-Site DNA Writing.” *Nature Chemical Biology*, March. https://doi.org/10.1038/s41589-021-00769-8.

Ma, Sai, Bing Zhang, Lindsay M. LaFave, Andrew S. Earl, Zachary Chiang, Yan Hu, Jiarui Ding, et al. 2020. “Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin.” *Cell* 183 (4): 1103-1116.e20.

McKenna, Aaron, Gregory M. Findlay, James A. Gagnon, Marshall S. Horwitz, Alexander F. Schier, and Jay Shendure. 2016. “Whole-Organism Lineage Tracing by Combinatorial and Cumulative Genome Editing.” *Science* 353 (6298): aaf7907.

Mold, Jeff E., Martin H. Weissman, Michael Ratz, Michael Hagemann-Jensen, Joanna Hård, Carl-Johan Eriksson, Hosein Toosi, et al. 2022. “Clonally Heritable Gene Expression Imparts a Layer of Diversity within Cell Types.” *BioRxiv*. https://doi.org/10.1101/2022.02.14.480352.

Muto, Yoshiharu, Parker C. Wilson, Nicolas Ledru, Haojia Wu, Henrik Dimke, Sushrut S. Waikar, and Benjamin D. Humphreys. 2021. “Single Cell Transcriptional and Chromatin Accessibility Profiling Redefine Cellular Heterogeneity in the Adult Human Kidney.” *Nature Communications* 12 (1): 2190.

Nair, Venugopalan D., Mital Vasoya, Vishnu Nair, Gregory R. Smith, Hanna Pincas, Yongchao Ge, Collin M. Douglas, Karyn A. Esser, and Stuart C. Sealfon. 2021. "Differential Analysis of Chromatin Accessibility and Gene Expression Profiles Identifies Cis-Regulatory Elements in Rat Adipose and Muscle." *Genomics* 113 (6): 3827–41.

Nichols, Caitlin A., William J. Gibson, Meredith S. Brown, Jack A. Kosmicki, John P. Busanovich, Hope Wei, Laura M. Urbanski, et al. 2020. "Loss of Heterozygosity of Essential Genes Represents a Widespread Class of Potential Cancer Vulnerabilities." *Nature Communications* 11 (1): 2517.

O'Leary, Timothy P., Kaitlin E. Sullivan, Lihua Wang, Jody Clements, Andrew L. Lemire, and Mark S. Cembrowski. 2020. "Extensive and Spatially Variable Within-Cell-Type Heterogeneity across the Basolateral Amygdala." *ELife* 9 (September). https://doi.org/10.7554/eLife.59003.

Patel, Anoop P., Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, et al. 2014. "Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma." *Science* 344 (6190): 1396–1401.

Pérez-González, Ana, and Elena Caro. 2019. "Benefits of Using Genomic Insulators Flanking Transgenes to Increase Expression and Avoid Positional Effects." *Scientific Reports* 9 (1): 8474.

Perli, Samuel D., Cheryl H. Cui, and Timothy K. Lu. 2016. "Continuous Genetic Recording with Self-Targeting CRISPR-Cas in Human Cells." *Science* 353 (6304). https://doi.org/10.1126/science.aag0511.

Pierce, Sarah E., Jeffrey M. Granja, and William J. Greenleaf. 2021. "High-Throughput Single-Cell Chromatin Accessibility CRISPR Screens Enable Unbiased Identification of Regulatory Networks in Cancer." *Nature Communications* 12 (1): 2969.

Raj, Bushra, James A. Gagnon, and Alexander F. Schier. 2018. "Large-Scale Reconstruction of Cell Lineages Using Single-Cell Readout of Transcriptomes and CRISPR-Cas9 Barcodes by ScGESTALT." *Nature Protocols* 13 (11): 2685–2713.

Raj, Bushra, Daniel E. Wagner, Aaron McKenna, Shristi Pandey, Allon M. Klein, Jay Shendure, James A. Gagnon, and Alexander F. Schier. 2018. "Simultaneous Single-Cell Profiling of Lineages and Cell Types in the Vertebrate Brain." *Nature Biotechnology* 36 (5): 442–50.

Ratz, Michael, Leonie von Berlin, Ludvig Larsson, Marcel Martin, Jakub Orzechowski Westholm, Gioele La Manno, Joakim Lundeberg, and Jonas Frisén. 2021. "Cell Types and Clonal Relations in the Mouse Brain Revealed by Single-Cell and Spatial Transcriptomics." *BioRxiv*. https://doi.org/10.1101/2021.08.31.458418.

Replogle, Joseph M., Thomas M. Norman, Albert Xu, Jeffrey A. Hussmann, Jin Chen, J. Zachery Cogan, Elliott J. Meer, et al. 2020. "Combinatorial Single-Cell CRISPR Screens by Direct Guide RNA Capture and Targeted Sequencing." *Nature Biotechnology* 38 (8): 954–61.

Rodriguez-Fraticelli, Alejo E., Samuel L. Wolock, Caleb S. Weinreb, Riccardo Panero, Sachin H. Patel, Maja Jankovic, Jianlong Sun, Raffaele A. Calogero, Allon M. Klein, and Fernando D. Camargo. 2018. "Clonal Analysis of Lineage Fate in Native Haematopoiesis." *Nature* 553 (7687): 212–16.

Rubin, Adam J., Kevin R. Parker, Ansuman T. Satpathy, Yanyan Qi, Beijing Wu, Alvin J. Ong, Maxwell R. Mumbach, et al. 2019. "Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks." *Cell* 176 (1–2): 361-376.e17.

Salgia, Ravi, and Prakash Kulkarni. 2018. “The Genetic/Non-Genetic Duality of Drug ‘Resistance’ in Cancer.” *Trends in Cancer Research* 4 (2): 110–18.

Salvador-Martínez, Irepan, Marco Grillo, Michalis Averof, and Maximilian J. Telford. 2019. “Is It Possible to Reconstruct an Accurate Cell Lineage Using CRISPR Recorders?” *ELife* 8 (January). https://doi.org/10.7554/eLife.40292.

Sfeir, Agnel, and Lorraine S. Symington. 2015. “Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway?” *Trends in Biochemical Sciences* 40 (11): 701–14.

Shaffer, Sydney M., Benjamin L. Emert, Raúl A. Reyes Hueros, Christopher Cote, Guillaume Harmange, Dylan L. Schaff, Ann E. Sizemore, et al. 2020. “Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors.” *Cell* 182 (4): 947-959.e17.

Shimizu, Atsushi, Shuichi Asakawa, Takashi Sasaki, Satoru Yamazaki, Hidehisa Yamagata, Jun Kudoh, Shinsei Minoshima, Ikuko Kondo, and Nobuyoshi Shimizu. 2003. “A Novel Giant Gene CSMD3 Encoding a Protein with CUB and Sushi Multiple Domains: A Candidate Gene for Benign Adult Familial Myoclonic Epilepsy on Human Chromosome 8q23.3-Q24.1.” *Biochemical and Biophysical Research Communications* 309 (1): 143–54.

Simeonov, Kamen P., China N. Byrns, Megan L. Clark, Robert J. Norgard, Beth Martin, Ben Z. Stanger, Jay Shendure, Aaron McKenna, and Christopher J. Lengner. 2021. “Single-Cell Lineage Tracing of Metastatic Cancer Reveals Selection of Hybrid EMT States.” *Cancer Cell* 39 (8): 1150-1162.e9.

SoRelle, Elliott D., Joanne Dai, Emmanuela N. Bonglack, Emma M. Heckenberg, Jeffrey Y. Zhou, Stephanie N. Giamberardino, Jeffrey A. Bailey, Simon G. Gregory, Cliburn Chan, and Micah A. Luftig. 2021. “Single-Cell RNA-Seq Reveals Transcriptomic Heterogeneity Mediated by Host-Pathogen Dynamics in Lymphoblastoid Cell Lines.” *ELife* 10 (January). https://doi.org/10.7554/eLife.62586.

Spanjaard, Bastiaan, Bo Hu, Nina Mitic, Pedro Olivares-Chauvet, Sharan Janjuha, Nikolay Ninov, and Jan Philipp Junker. 2018. “Simultaneous Lineage Tracing and Cell-Type Identification Using CRISPR-Cas9-Induced Genetic Scars.” *Nature Biotechnology* 36 (5): 469–73.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. “Comprehensive Integration of Single-Cell Data.” *Cell* 177 (7): 1888-1902.e21.

Sulston, J. E., E. Schierenberg, J. G. White, and J. N. Thomson. 1983. “The Embryonic Cell Lineage of the Nematode Caenorhabditis Elegans.” *Developmental Biology* 100 (1): 64–119.

Thomas, Sean, Xiao-Yong Li, Peter J. Sabo, Richard Sandstrom, Robert E. Thurman, Theresa K. Canfield, Erika Giste, et al. 2011. “Dynamic Reprogramming of Chromatin Accessibility during Drosophila Embryo Development.” *Genome Biology* 12 (5): R43.

Tunnacliffe, Edward, and Jonathan R. Chubb. 2020. “What Is a Transcriptional Burst?” *Trends in Genetics: TIG* 36 (4): 288–97.

Urasaki, Akihiro, Kazuhide Asakawa, and Koichi Kawakami. 2008. “Efficient Transposition of the Tol2 Transposable Element from a Single-Copy Donor in Zebrafish.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (50): 19827–32.

VanHorn, Sadie, and Samantha A. Morris. 2021. “Next-Generation Lineage Tracing and Fate Mapping to Interrogate Development.” *Developmental Cell* 56 (1): 7–21.

Wagner, Daniel E., Caleb Weinreb, Zach M. Collins, James A. Briggs, Sean G. Megason, and Allon M. Klein. 2018. “Single-Cell Mapping of Gene Expression Landscapes and Lineage in the Zebrafish Embryo.” *Science* 360 (6392): 981–87.

Wang, Ziqiang, Kun Li, and Weiren Huang. 2020. “Long Non-Coding RNA NEAT1-Centric Gene Regulation.” *Cellular and Molecular Life Sciences: CMLS* 77 (19): 3769–79.

Watanabe, Yoshihisa, Kiyoshi Shibata, and Masato Maekawa. 2014. “Cell Line Differences in Replication Timing of Human Glutamate Receptor Genes and Other Large Genes Associated with Neural Disease.” *Epigenetics: Official Journal of the DNA Methylation Society* 9 (10): 1350–59.

Weinreb, Caleb, Alejo Rodriguez-Fraticelli, Fernando D. Camargo, and Allon M. Klein. 2020. “Lineage Tracing on Transcriptional Landscapes Links State to Fate during Differentiation.” *Science* 367 (6479). https://doi.org/10.1126/science.aaw3381.

Xing, Qiao Rui, Chadi A. El Farran, Ying Ying Zeng, Yao Yi, Tushar Warrier, Pradeep Gautam, James J. Collins, et al. 2020. “Parallel Bimodal Single-Cell Sequencing of Transcriptome and Chromatin Accessibility.” *Genome Research* 30 (7): 1027–39.

Yang, Cheng-Mei, Hao-Sheng Chang, Hung-Chih Chen, Jyun-Jie You, Huei-Han Liou, Su-Chen Ting, Luo-Ping Ger, Sung-Chou Li, and Kuo-Wang Tsai. 2019. “Low C6orf141 Expression Is Significantly Associated with a Poor Prognosis in Patients with Oral Cancer.” *Scientific Reports* 9 (1): 4520.

Yang, Han, Shuling Ren, Siyuan Yu, Haifeng Pan, Tingdong Li, Shengxiang Ge, Jun Zhang, and Ningshao Xia. 2020. “Methods Favoring Homology-Directed Repair Choice in Response to CRISPR/Cas9 Induced-Double Strand Breaks.” *International Journal of Molecular Sciences* 21 (18). https://doi.org/10.3390/ijms21186461.

Yesbolatova, Aisha, Yuichiro Saito, Naomi Kitamoto, Hatsune Makino-Itou, Rieko Ajima, Risako Nakano, Hirofumi Nakaoka, et al. 2020. “The Auxin-Inducible Degron 2 Technology Provides Sharp Degradation Control in Yeast, Mammalian Cells, and Mice.” *Nature Communications* 11 (1): 5701.

Zhang, Jia-Jia, Jiang Hong, Yu-Shui Ma, Yi Shi, Dan-Dan Zhang, Xiao-Li Yang, Cheng-You Jia, et al. 2021. “Identified GNGT1 and NMU as Combined Diagnosis Biomarker of Non-Small-Cell Lung Cancer Utilizing Bioinformatics and Logistic Regression.” *Disease Markers* 2021 (January): 6696198.

Zhang, Yusheng, Ho Lam Chan, Liliana Garcia-Martinez, Daniel L. Karl, Natalia Weich, Joyce M. Slingerland, Ramiro E. Verdun, and Lluis Morey. 2020. “Estrogen Induces Dynamic ERα and RING1B Recruitment to Control Gene and Enhancer Activities in Luminal Breast Cancer.” *Science Advances* 6 (23): eaaz7249.

Zhu, Chenxu, Miao Yu, Hui Huang, Ivan Juric, Armen Abnousi, Rong Hu, Jacinta Lucero, M. Margarita Behrens, Ming Hu, and Bing Ren. 2019. “An Ultra High-Throughput Method for Single-Cell Joint Analysis of Open Chromatin and Transcriptome.” *Nature Structural & Molecular Biology* 26 (11): 1063–70.

APPENDIX A: DATA AND CODE AVAILABILITY

Raw and processed data and code are available on GEO (GSE201339) & Github (https://github.com/minkinaa/TheLorax). See README on Github for further details.

VITA

Originally from Ukraine, Anna Minkina immigrated to the U.S. a child. She earned her Bachelor of Arts degree in Biology with a Neuroscience concentration from Carleton College in 2010. Prior to beginning graduate school, she worked as a research scientist in the lab of Dr. David Zarkower at the University of Minnesota, studying mammalian sex determination, while seriously considering careers in science communication and education. After discovering a love for high-throughput data and computational biology, she joined the Genome Sciences department at the University of Washington in 2016, where she pursued her Ph.D. in the lab of Dr. Jay Shendure.

Her first research experience was in the lab of Dr. Stephan Zweifel, a University of Washington Genetics Department alumnus, who offered to hit her over the head with a textbook upon the completion of her undergrad thesis so she would never think about it again. She is pleased that producing a physical document thick enough to accomplish this task is a Ph.D. requirement.