QUANTITATIVE METHODS FOR GENOMICS AND LINEAGE TRACING DATA

By

Weixiang Fang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

April, 2022

Abstract

The thesis consists of two parts. The first part discusses a new method for quantifying functional conservation of DNA elements. Evolutionary conservation is an important tool for identifying functional DNA elements in genomes and provides a foundation for studying human diseases using animal models. Conservation in DNA sequences does not necessarily imply conservation in dynamic functional activities. Quantifying functional conservation, however, has been constrained by limited availability of functional genomic data from well-matched samples across species. Here we present FUNCODE, a solution to scoring functional conservation of DNA elements by integrating data across species without requiring manually or exactly matched samples. By using in silico sample matching, FUNCODE more accurately scores functional conservation and offers scalability to new samples and ability to score different data modalities. Applying it to the Encyclopedia of DNA Elements (ENCODE), we systematically scored human-mouse conservation of DNA regulatory elements based on chromatin accessibility. We further demonstrate utility of FUNCODE in finding new cis-regulatory elements, identifying discoveries translatable across species, and cross-species single-cell genomic data integration. The second part of the thesis studies inference of cell state dynamics using lineage barcode data. Natural and induced somatic mutations that accumulate in the genome during development record the phylogenetic relationships of cells; however, whether these lineage barcodes can capture the dynamics of complex progenitor fields remains unclear. Here, we introduce quantitative fate mapping, an approach to simultaneously map the fate and quantify the commitment time, commitment bias, and population size of multiple progenitor groups during development based on a time-scaled phylogeny of their descendants. To reconstruct time-scaled phylogenies from lineage barcodes, we introduce Phylotime, a scalable maximum likelihood clustering approach based on a generalizable barcoding mutagenesis model. We validate these approaches using realistically-simulated barcoding results as well as

experimental results from a barcoding stem cell line. We further establish criteria for the minimum number of cells that must be analyzed for robust quantitative fate mapping. Overall, this work demonstrates how lineage barcodes, natural or synthetic, can be used to obtain quantitative fate maps, thus enabling analysis of progenitor dynamics long after embryonic development in any organism.

Readers:

Advisor: Dr. Hongkai Ji

Reader: Dr. Reza Kalhor

Reader: Dr. Don J. Zack

Reader: Dr. Alan L. Scott

Reader: Dr. Ni Zhao

Preface

Acknowledgements

I would like to thank my advisor Dr. Hongkai Ji for his mentorship, advice and continuous support over the years, as well as people in the Ji lab including Dr. Weiqiang Zhou, Dr. Zhicheng Ji, Boyang Zhang, Ruzhang Zhao, Yi Wang and Chaoran Chen. I would also like to acknowledge Dr. Don J. Zack's lab and Dr. Reza Kalhor for their collaboration and support during the program. I would also like to thank Claire M. Bell, Dr. Xitiz Chamling from the Zack lab as well as Abel Sapirstein and Kathleen Leeper from the Kalhor lab. Dr. Ni Zhao and Dr. Alan Scott have provided valuable feedback and discussion as the thesis committee members. Finally, I would like to thank many students, faculty members and staff in the biostatistics department, not to be fully listed here, for a cherished time spent in the department.

Table of Contents

List of Figures

Chapter 1

Scoring cross-species functional conservation of DNA elements via unsupervised sample matching

Other anchors: Chaoran Chen, Yi Wang, Ruzhang Zhao

Corresponding author: Hongkai Ji

Introduction

Comparative genomics, which involves comparing the genome sequences across different species, has been a powerful tool for studying evolution and identifying functional DNA elements in genomes [1,2]. Negative (purifying) selection on DNA elements with important functions often results in conserved DNA sequences across species, whereas positive selection may result in divergence of DNA sequence in a species from other species with a shared ancestor. While DNA sequence conservation has been widely used for identifying functional DNA elements, recent studies show that DNA sequence conservation does not necessarily imply conservation of a DNA element's functional activities which can be highly context (i.e. tissue or cell type) dependent. [3–5] The boom of functional genomic data in the past decade has enabled comparison of context-dependent biochemical activities of DNA elements across species, giving birth to comparative functional genomics and creating an unprecedented opportunity to examine the conservation of context-dependent activities, also known as functional conservation of DNA elements.

One important task of comparative functional genomics is to measure the degree of conservation of each DNA element with regard to each data modality (e.g., gene expression, chromatin accessibility, histone modification, DNA methylation, etc.). Measuring the

conservation of biochemical activities of DNA elements not only provides a foundation for understanding the evolution of functional genomes but also has important practical utility for studying human diseases using animal models. While animal models are widely used to study diseases when conducting experiments on human subjects is impractical, using such models relies on the assumption that the functional activities of human DNA elements (e.g. genes, regulatory elements, etc.) are conserved in the animal model when there is sequence homology, an assumption not always easy to check. Knowing which DNA elements have conserved functional activities across species therefore can inform the design of animal models and interpretation of findings from such models in terms of their translatability to human diseases.

Majority of the regulatory elements are context-dependent. Traditional conservation scores such as PhastCons [6] and PhyloP [7] derived from static DNA sequences are incapable of reflecting changes in tissue or cell type specificity of individual elements. While there are existing methods to systematically score functional conservation of DNA elements [8], the scores do not directly reflect concordances of functional activities.

As part of the Encyclopedia of DNA Elements (ENCODE) consortium, we present FUNCODE, a scalable computational framework for scoring functional conservation of DNA elements that does not require manual matching of samples across species using prior knowledge. Applying FUNCODE to ENCODE human and mouse DNase-seq and ATAC-seq data, we systematically scored the conservation of chromatin accessibility of human and mouse regulatory elements. Through comparisons with existing conservation scores, we demonstrate the unique value of FUNCODE and showcase its applications in annotating cis-regulatory elements, prioritizing candidates for animal studies, integrating single-cell genomic data, and interpreting disease associated genetic variants.

Results

A scalable computational framework for scoring functional conservation

In order to characterize functional conservation of DNA elements between two species, we developed a computational framework, FUNCODE, that does not rely on manually matched samples (**Fig. 1.1**). Given a set of functional genomic (e.g. DNase-seq) samples from each species, FUNCODE characterizes conservation of functional genomic signals for homologous DNA elements between the two species. First, DNA elements with sequence homology between the two species are identified based on the Blastz [9] pairwise genome alignment. Second, functional genomic data are extracted from each species for DNA elements that are aligned between species. Third, aligned elements with highly variable functional genomic signals across samples in both species are identified and used as features to integrate samples from both species by embedding them into a common low-dimensional space using canonical correlation analysis (CCA). Fourth, in the low-dimensional sample embedding, mutual nearest neighbors within a defined distance are identified between two species using Seurat to create *in silico* pseudo-matched samples [10]. Finally, the pseudo-matched samples are used to compute DNA elements' conservation scores (**Fig. 1.1a**).

A DNA element can have functional genomic signals that are variable across different contexts (tissues or cell types) as well as ubiquitous baseline signals that are context-independent. The conservation of context-dependent variable signals can be characterized via the correlation of the signals across pseudo-matched samples. However, the conservation of context-independent baseline signals may not be captured by correlation. For example, a housekeeping promoter may have high activity in all cell types. However, this activity may be a constant (or has low variability) across cell types, leading to zero (or low) correlation between two species even

though it represents conserved function (**Fig. 1.1b**). For this reason, FUNCODE computes two conservation scores for each element to characterize the conservation of context-dependent variable activities (CO-V) and the conservation of constitutive baseline activities (CO-B), respectively (**Fig. 1.1b**). DNA elements in a given species can then be categorized as: (i) elements without alignable sequence in the other species, (ii) elements with aligned sequence in the other species (i.e. alignable elements) but not functionally conserved (i.e. low CO-V and low CO-B scores), or (iii) alignable elements with functional conservation (i.e. high CO-V or high CO-B score) (**Fig. 1.1c**).

Unlike the conventional methods that evaluate conservation on the basis of manually matched samples between species, FUNCODE matches samples in a data-driven way. Such an approach has multiple advantages. First, accurate manual sample matching can be difficult due to incomplete knowledge. For example, it is not always clear how to map a mouse developmental time point to a specific human age, or whether a human sample and a mouse sample matched by tissue or cell type are truly equivalent as the experimental procedure of tissue or cell type isolation are often different. FUNCODE circumvents this issue as it does not rely on prior knowledge to match samples. Second, manual sample matching has difficulty scaling up to a large number of new samples. Our data-driven approach is easily scalable to new samples, making it more convenient to update conservation scores as more data become available. Third, with its scalability and ability to align samples without requiring exact matches, FUNCODE can include many more samples to compute conservation scores compared to methods that rely on a relatively small number of expert-annotated exact matches.

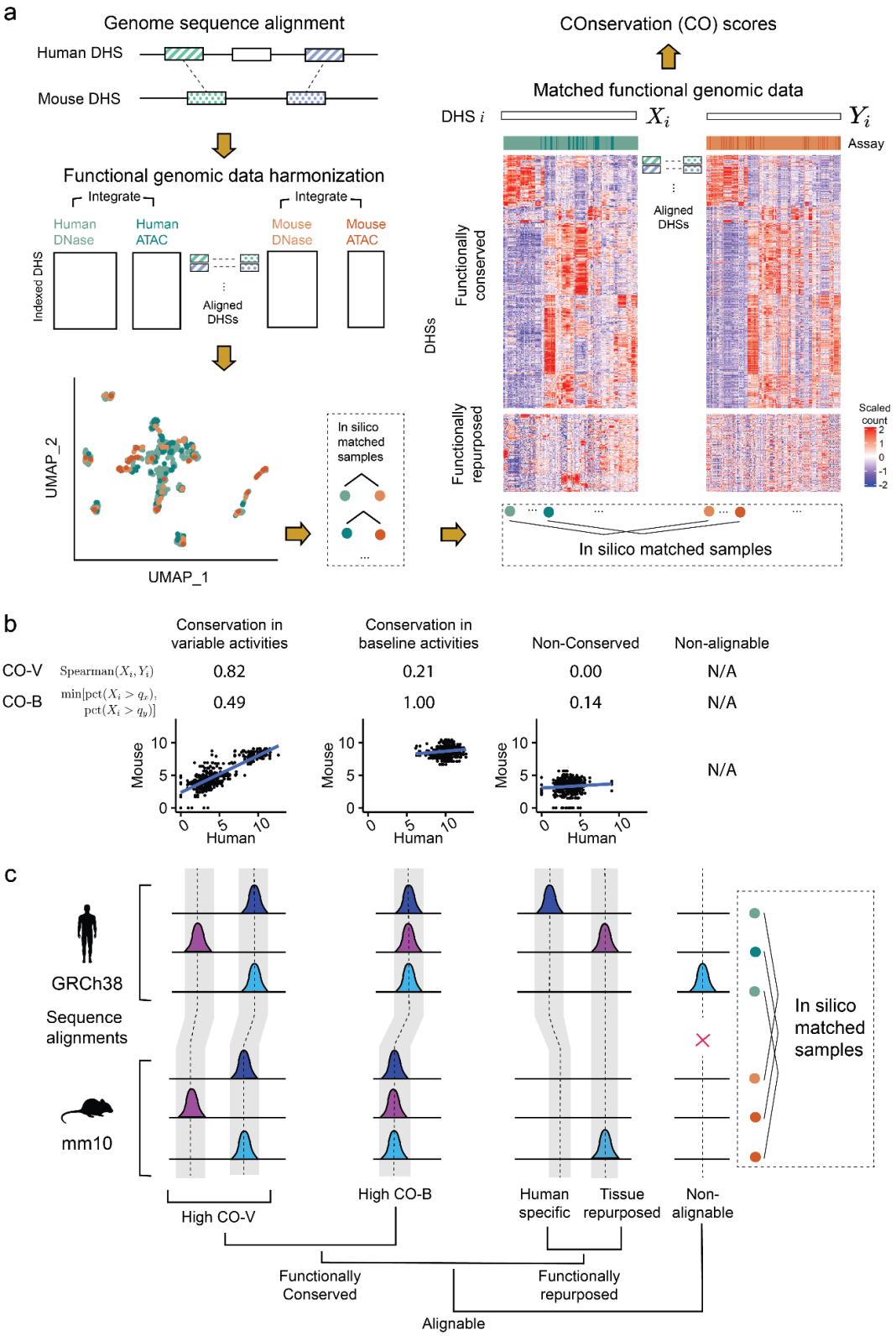
a
Genome sequence alignment
Human DHS
Mouse DHS
Functional genomic data harmonization
Integrate
Integrate
Human DNase
Human ATAC
Mouse DNase
Mouse ATAC
Indexed DHS
Aligned DHSs
UMAP_2
UMAP_1
In silico matched samples
COnservation (CO) scores
Matched functional genomic data
DHS i
X_i
Y_i
Assay
Aligned DHSs
DHSs
Functionally conserved
Functionally repurposed
Scaled count
2
1
0
-1
-2
In silico matched samples
b
Conservation in variable activities
Conservation in baseline activities
Non-Conserved
Non-alignable
CO-V
$\mathrm{Spearman}(X_i, Y_i)$
0.82
0.21
0.00
N/A
CO-B
$\min[\mathrm{pct}(X_i > q_x), \mathrm{pct}(X_i > q_y)]$
0.49
1.00
0.14
N/A
Mouse
Human
0
5
10
c
GRCh38
Sequence alignments
mm10
In silico matched samples
High CO-V
High CO-B
Human specific
Tissue repurposed
Non-alignable
Functionally Conserved
Functionally repurposed
Alignable

Figure 1.1. Overview of FUNCODE demonstrated using chromatin accessibility conservation. (a) FUNCODE workflow. Human regulatory elements or DNase I hypersensitive sites (DHSs) are aligned to the mouse genome based on pairwise sequence alignments. Functional genomic data are co-embedded into common low dimensional space from which in-silico matches are identified. Heatmaps show example DHSs (rows) with concordant (conserved) or discordant (repurposed) context-dependent activities. (b) Two types of functional conservation scores are computed for each DHS. Conservation score for context dependent variable activities (CO-V) and conservation scores in n constitutive baseline activity (CO-B). (c) Classification of DHSs. A pair of DHS is alignable if they can be mapped by pairwise sequence alignment. A pair of aligned DHS can be high CO-V or high CO-B (both classified as functionally conserved) or be non-conserved.

Systematic characterization of DNA elements with conserved chromatin accessibility between human and mouse

Using FUNCODE, we systematically scored the conservation of chromatin accessibility (CA) between human and mouse for the DNase I hypersensitive sites (DHSs) compiled by ENCODE. The conservation scores, referred to as CACO scores, will be made available via the ENCODE data portal. To derive CACO, 835 human (689 DNase-seq, 137 ATAC-seq) and 120 mouse (88 DNase-seq, 32 ATAC-seq) samples were integrated after removing assay specific effects (**Fig. 1.1a**, **Fig. S1.1 Supplementary data 1.1-3,** Methods). They were used to score 1.8 million human DHSs alignable to mouse among a total of 3.6 million human DHSs, and 1.1 million mouse DHSs alignable to human among a total of 1.8 million mouse DHSs [11] (**Fig. 1.2a**). As examples, **Fig. 1.2b-d** show three DHSs that contain human SNPs associated with multiple sclerosis based on genome-wide association studies (GWAS) . The CACO scores are shown as

two genome browser tracks. The DHS in **Fig. 1.2b** has highly conserved context-dependent activity (high CACO-V). The DHS in **Fig. 1.2c** has conserved constitutive baseline activity (high CACO-B). The DHS in **Fig. 1.2d** is not conserved (low CACO-V and low CACO-B).

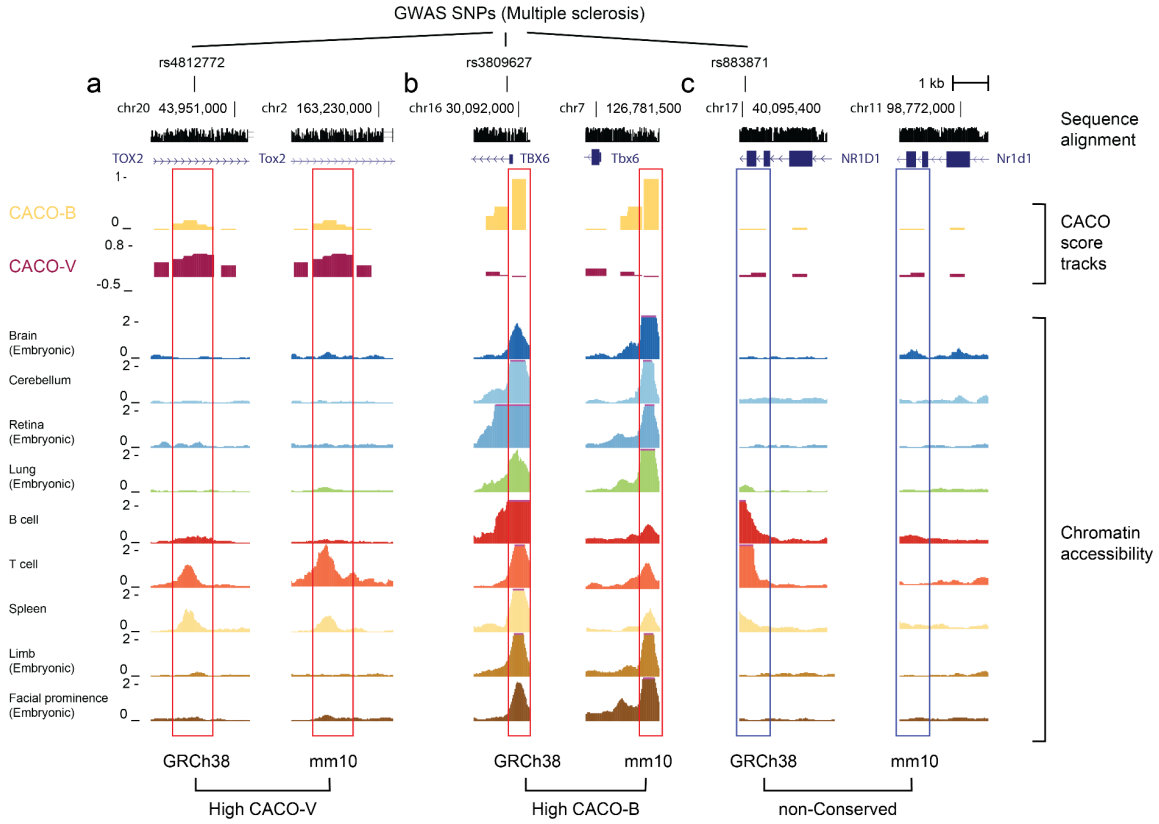

Figure 1.2. Examples of FUNCODE chromatin accessibility conservation (CACO) scores at three loci with GWAS SNPs associated with multiple sclerosis. (a) DHS with conserved context-dependent activities (high CACO-V elements), high activities are observed for T cell and spleen in both species. (b) DHS with conserved constitutive baseline activities (high CACO-B elements) near promoter of TBX6/Tbx6 gene. (C) DHS without functional conservation (low CACO-V and low CACO-B elements). Regulatory activities in B cell and T cell are missing in the mouse tracks.

To assess the statistical significance of conservation, we constructed null distributions of CACO scores by randomly pairing human and mouse DHSs (see Methods). We identified DHSs with statistically significant functional conservation at a false discovery rate (FDR) cutoff of 0.1 (**Fig. 1.3a,b**). For human, 129,718 DHSs with conserved variable activity (high CACO-V elements) and 88,585 DHSs with conserved baseline activity (high CACO-B elements) were identified, accounting for 7.07% and 4.83% of alignable DHSs respectively. 16,796 DHSs showed conservation in both variable and baseline activities (**Fig. 1.3c**). Using GREAT analysis, we found that DHSs with conserved variable activity enriched GO terms related to 'regulation of transcription' and DHSs with conserved baseline activities enriched GO terms related to chromatin silencing and chromosome organization (**Supplementary data 1.4**).

Cross-referencing with the ENCODE annotated candidate cis-regulatory element (cCRE) [12], we found that the majority of DHSs with conserved variable activity were distal and proximal enhancer-like elements (dELS: 51.3% - 37.7%, pELS:19.6% - 16.4%), and the majority of DHSs with conserved baseline activity were promoters and proximal enhancer-like elements (PLS: 33.1% - 31.8%, pELS: 41.5%-37.8%) (**Fig. 1.3c**). Different categories of cCRE showed different levels of functional conservation. For example, 24.4% (26.2%) of human (mouse) cCRE with promoter-like signature (PLS) had conserved variable activities and 56.1% (63.7%) had conserved baseline activities. By contrast, among alignable DHSs not annotated as human (mouse) cCREs, only 2.19% (2.93%) and 0.173% (0.375%) showed conserved variable and baseline activities respectively (**Fig. 1.3d**). Similarly, cCREs with proximal enhancer-like signature (pELS) or distal enhancer-like signature (dELS) also showed higher levels of

conservation compared to non-cCREs (**Fig. 1.3d**).

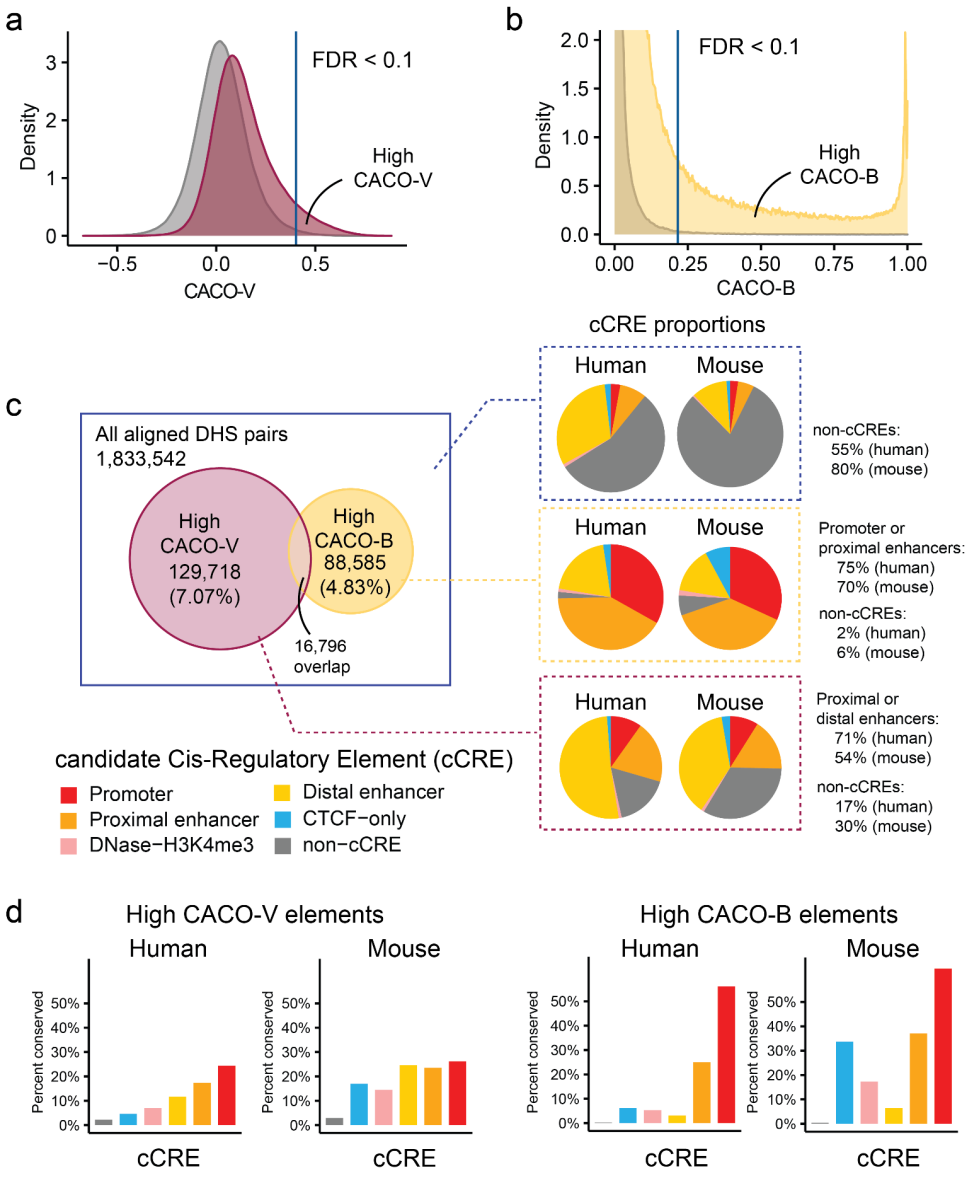

Figure 1.3. Functionally conserved regulatory elements enriches candidate Cis-Regulatory Elements (cCRE). (a) Distribution of CACO-V scores for aligned pairs and random pairs of DHS. High CACO-V elements were called with false discovery rate (FDR) cutoff of 0.1. (b) similar for CACO-B scores. (c) Breakdown of High CACO-V, high CACO-B and all regulatory elements by candidate cis-regulatory elements (cCRE) categories. High CACO-V elements enrich proximal

and distal enhancers. High CACO-B elements enrich promoters and proximal enhancers. (d) Different cCRE categories show different levels of functional conservation.

Functional conservation identifies new candidate cis-regulatory elements

Among the functionally conserved DHSs, a significant fraction (17.1% human and 33.3% mouse high CACO-V elements, 2.0% human and 6.2% mouse high CACO-B elements) have not been annotated as cCREs (**Fig. 1.3c**), likely due to lack of sufficient functional genomic data within each species to support the cCRE annotation. With the new information from cross-species conservation, we speculate that many of these conserved “non-cCRE” DHSs likely are real regulatory elements. Indeed, among human non-cCREs with low CACO-V scores and cCREs with low CACO-V scores, only 8.2% and 20.0% were aligned to mouse sequences that were annotated as mouse cCREs. By contrast, 72.6% human non-cCREs with high CACO-V scores were aligned to mouse cCREs (8.8 and 3.6 fold enchement) (**Fig. 1.4a**). Similar phenomena were observed for mouse and CACO-B scores (**Fig. 1.4a**), supporting our hypothesis that the conserved DHSs not currently annotated as cCREs are likely real cCREs.

As histone modification H3K4me1 data were not used to define cCREs, we used it as an independent data modality to further validate our hypothesis. We reasoned that if conserved “non-cCRE” DHSs were real cCREs, they should more likely carry H3K4me1 signals and such signals are likely also conserved (**Fig. 1,4b,c**). As FUNCODE can be conveniently generalized to other data modalities, we applied it to ENCODE H3K4me1 ChIP-seq data (**Supplementary data 1.5-7)** to compute conservation scores for this histone modification, referred to as HICO-H3K4me1-V and HICO-H3K4me1-B. Non-cCREs with high CACO-V scores and cCREs with high CACO-V scores indeed had comparable HICO-H3K4me1-V scores, both significantly higher than the HICO-H3K4me1-V scores of cCREs and non-cCREs with low CACO-V scores

(**Fig. 1.4d**). Similarly, non-cCREs with high CACO-B scores had higher HICO-H3K4me1-B scores compared to cCREs and non-cCREs with low CACO-B scores (**Fig. 4d**). These data indicate that many of the conserved “non-cCRE” DHSs are bona fide cCREs.

In total, we found 23,721 functionally conserved non-cCRE DHSs in human (21,944 with high CACO-V, 1,530 with high CACO-B, 220 with conservation in both) and 47,061 conserved non-cCRE DHSs in mouse (42,614 high CACO-V, 4,984 high CACO-B, 537 both). We label them as FUNCODE-detected new cCREs (**Supplementary data 1.8**). Based on the orthogonal HICO-H3K4me1 information at the non-conserved DHSs as the null, 84.6% of the new human cCREs and 62.5 of the new mouse cCRE were estimated to have q values smaller than 0.1 (CACO-V: human 86.8%, mouse% 0.609; CACO-B: human 57.7%, mouse: 74.6%). Thus, by borrowing information across species, FUNCODE can help identify new cCREs that were missed by analyses within a single species.

Functional rather than sequence conservation at regulatory elements more accurately predicts gene expression conservation

We next applied FUNCODE to ENCODE RNA-seq data (556 human and 186 mouse experiments) to score gene expression conservation (GECO) between the two species for 16,468 unique human-mouse homologous gene pairs, identifying 4,955 genes with conserved variable expression (high GECO-V) and 727 genes with conserved baseline expression (high GECO-B) at FDR 0.1 (**Fig. S1.2, Supplementary data 1.9-13**).

We asked whether conserved chromatin accessibility of regulatory elements predicts conserved gene expression of their target genes (**Fig. 1.4e**). We compared CACO with the recently developed LECIF conservation score [8] and three sequence-based conservation scores,

phastCons4Way (based on alignment of 4 genomes, Homo sapiens, Mus musculus, Galeopterus variegatus, Tupaia chinensis) [6], PhyloP4Way (based on alignment of 4 genomes, same as PhastCons) [7], and a human-mouse pairwise similarity score measured by the Percent Identical Base-pairs (PIB). We identified human-mouse DHS pairs whose candidate target genes were also homologous genes (See Methods). For these homologous DHS-gene pairs, GECO-V is correlated more with CACO-V than with LECIF and the three sequence-based conservation scores, PhastCons, PhyloP, and PIB (**Fig. 1.4f**). Taken together, CACO more accurately predicted the conservation of functional readout (gene expression) of regulatory elements.

Next, we compared CACO-V computed using in silico sample matching with a modified CACO-V computed using manually matched samples (CACO-V-Manual). For the latter, we manually matched ENCODE samples from the same tissue and cell type, yielding n=22 (covering 156 human and 37 mouse experiments) human-mouse matched sample pairs which were used to replace n=515 in silico matched sample pairs (covering 270 human and 126 mouse experiments) to recompute CACO scores (**Supplementary data 1.14**). Overall, CACO-V outperformed CACO-V-Manual in predicting gene expression conservation (**Fig. 1.4f**), demonstrating the advantage of unsupervised in silico sample matching over manual sample matching.

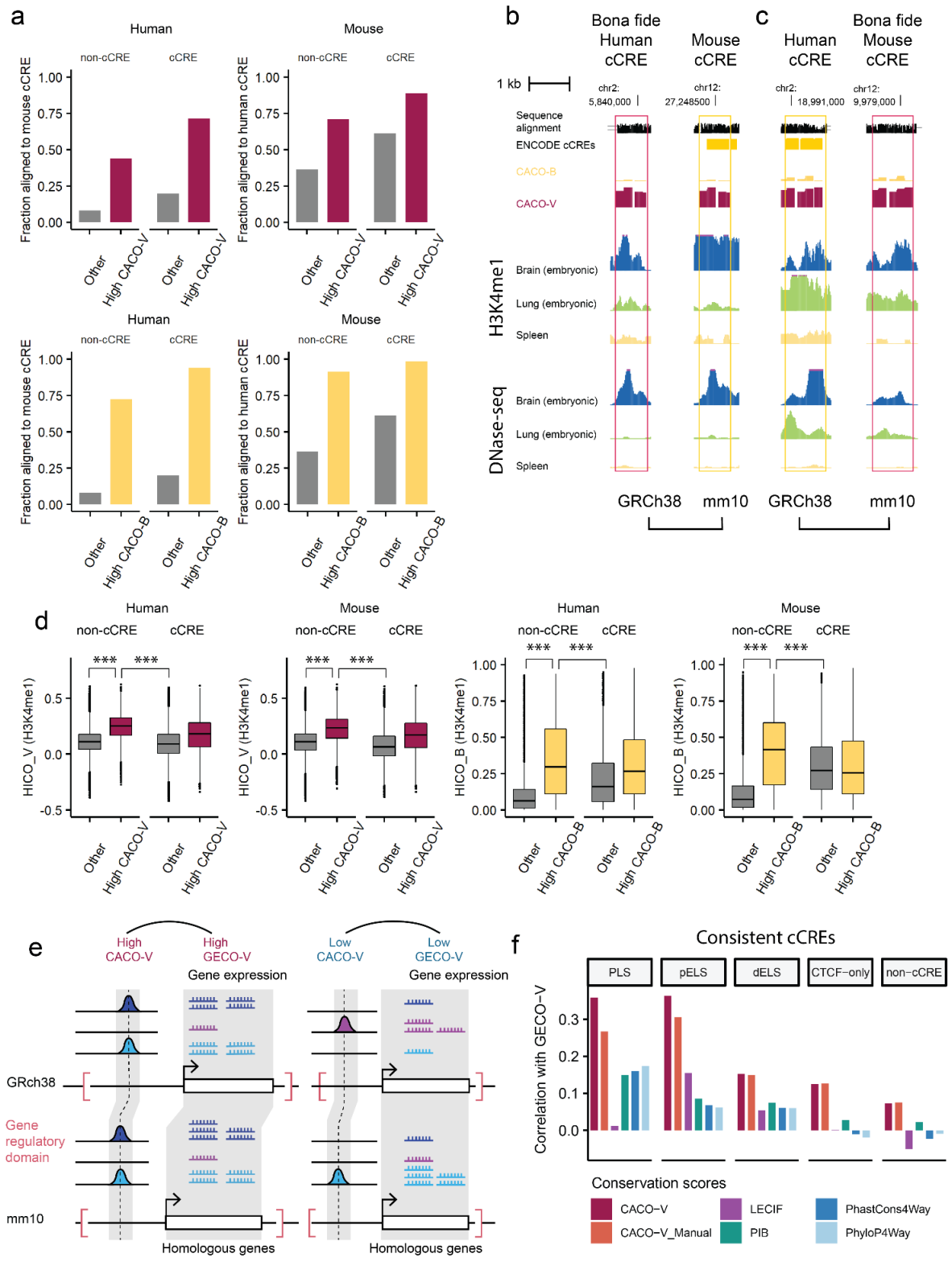
a
Human
Mouse
non-cCRE
cCRE
Fraction aligned to mouse cCRE
Fraction aligned to human cCRE
Other
High CACO-V
High CACO-B
b
Bona fide Human cCRE
Mouse cCRE
1 kb
Sequence alignment
ENCODE cCREs
CACO-B
CACO-V
H3K4me1
Brain (embryonic)
Lung (embryonic)
Spleen
DNase-seq
GRCh38
mm10
c
Human cCRE
Bona fide Mouse cCRE
d
HICO_V (H3K4me1)
HICO_B (H3K4me1)
***
e
High CACO-V
High GECO-V
Low CACO-V
Low GECO-V
Gene expression
GRch38
Gene regulatory domain
mm10
Homologous genes
f
Consistent cCREs
PLS
pELS
dELS
CTCF−only
non−cCRE
Correlation with GECO−V
Conservation scores
CACO−V
CACO−V_Manual
LECIF
PIB
PhastCons4Way
PhyloP4Way

Figure 1.4 Functional conservation scores identify functional non-cCRE and predict context-dependent conservation of gene expression. (a) non-cCRE with high CACO-V or high CACO-B are more likely to be aligned to a cCRE in the other species compared to non-conserved cCRE or non-cCRE. (b) An aligned region with high CACO-V score where the human DHS is cCRE but the aligned mouse DHS is non-cCRE. Example tracks show that tissue-specific activities of both chromatin accessibility and H3K4me1 ChIP-seq signals are highly correlated. (c) A similar example where the human element is annotated as cCRE but the aligned mouse element is not. (d) non-cCRE with high CACO-V or high CACO-B also have higher HICO-H3K4me1-V or HICO-H3K4me1-B, respectively, compared to non-conserved cCRE or non-cCRE. (e) Candidate target genes were assigned to a pair of aligned DHS they are in the regulatory domains of homologous genes. Correlation between conservation scores of regulatory elements and conservation scores of candidate target gene expression is expected. (f) Correlation of different conservation scores with GECO-V for each consistent cCRE category.

Functional rather than sequence conservation better identifies discoveries translatable across species

One utility of conservation scores is to evaluate how likely findings in one species can be translated to another species and hence guide the selection of candidate DNA elements for designing animal models for human diseases. For this task, we compared CACO with the other conservation scores via a cross-validation (CV) analysis.

We partitioned human-mouse sample pairs into training and test sets (**Fig. 1.5a, Supplementary data 1.15)**. In each fold of CV, conservation scores computed using the training set were used to rank human DHSs. To evaluate the performance for identifying DHSs with

conserved context-dependent variable activities, each test set contained two manually matched human-mouse sample pairs representing two different cell types not included in the training data. We asked how well the differential chromatin accessibility between the two test cell types correlates between human and mouse among the top ranked DHSs. The top conserved DHSs identified by CACO-V showed the highest human-mouse correlation in the test data, substantially outperforming the conserved DHSs identified by LECIF and the three sequence-based conservation scores (phastCons, PhyloP, PIB) (**Fig. 1.5a**)**.** Thus, CACO-V more accurately identified DHSs whose context-dependent activities in new test cell types are conserved between human and mouse. Moreover, in silico sample matching (CACO-V) substantially outperformed manual sample matching (CACO-V-Manual), demonstrating the advantage of unsupervised sample matching which yields more pseudo-matched sample pairs leading to increased power (**Fig. 1.5a**)**.**

Similarly, to evaluate the performance for identifying DHSs with conserved baseline activities, each test set contained one manually matched human-mouse sample pair representing one cell type not included in training data. We identified loci with high chromatin accessibility in the test cell type within each species and evaluated consistency of the high chromatin accessibility between human and mouse, measured using Jaccard index, among the top ranked DHSs identified by different conservation scores. CACO-B substantially outperformed LECIF, phastCons, PhyloP, and IBP. Thus, CACO-B most accurately identified DHSs whose constitutive baseline activities in new test cell types are conserved between human and mouse (**Fig. 1.5b**).

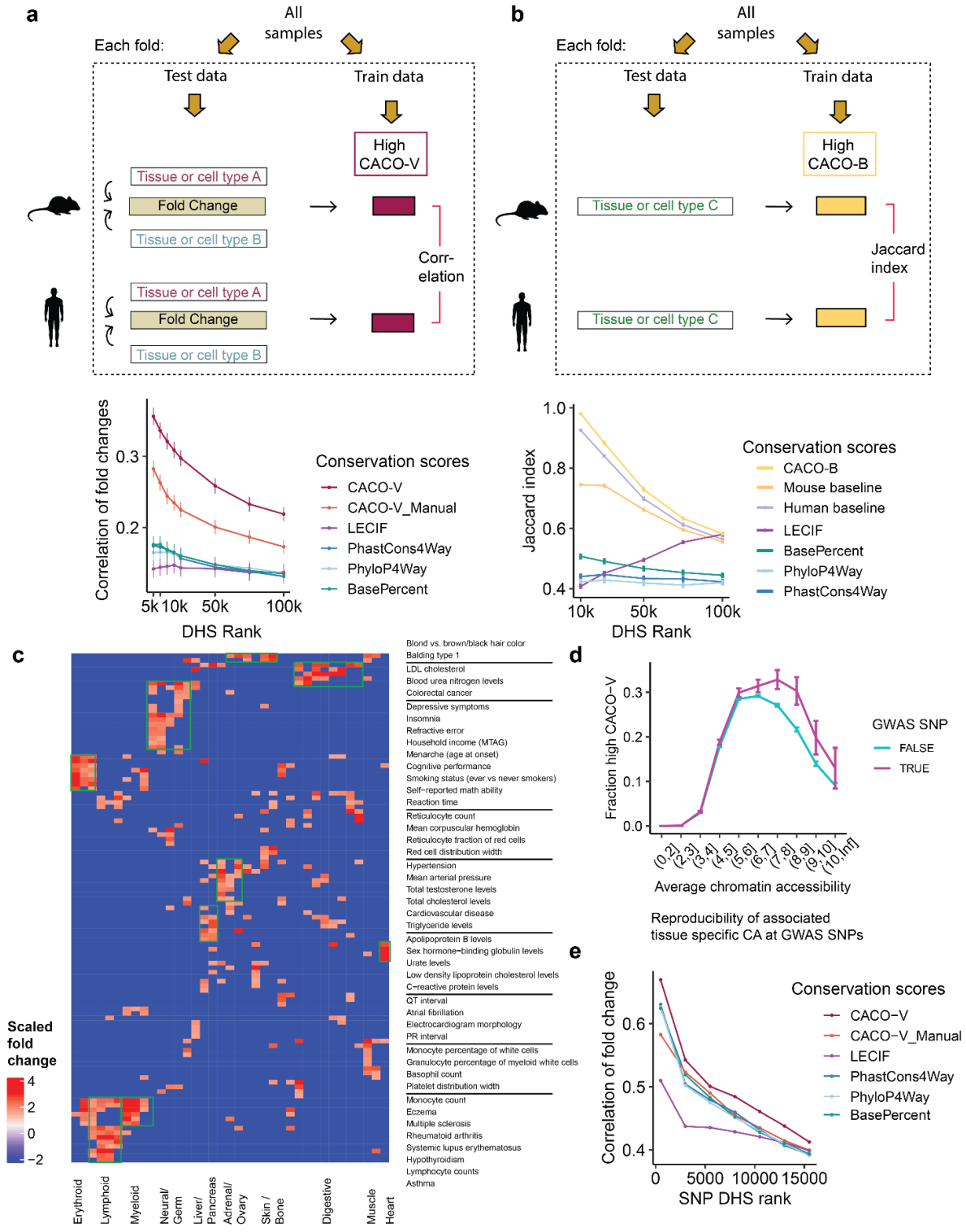
a
Each fold:
All samples
Test data
Train data
High CACO-V
Tissue or cell type A
Fold Change
Tissue or cell type B
Corr-elation
b
Each fold:
All samples
Test data
Train data
High CACO-B
Tissue or cell type C
Jaccard index
Correlation of fold changes
DHS Rank
Conservation scores
CACO-V
CACO-V_Manual
LECIF
PhastCons4Way
PhyloP4Way
BasePercent
Jaccard index
Conservation scores
CACO-B
Mouse baseline
Human baseline
LECIF
BasePercent
PhyloP4Way
PhastCons4Way
c
Scaled fold change
Erythroid
Lymphoid
Myeloid
Neural/ Germ
Liver/ Pancreas
Adrenal/ Ovary
Skin / Bone
Digestive
Muscle
Heart
Blond vs. brown/black hair color
Balding type 1
LDL cholesterol
Blood urea nitrogen levels
Colorectal cancer
Depressive symptoms
Insomnia
Refractive error
Household income (MTAG)
Menarche (age at onset)
Cognitive performance
Smoking status (ever vs never smokers)
Self−reported math ability
Reaction time
Reticulocyte count
Mean corpuscular hemoglobin
Reticulocyte fraction of red cells
Red cell distribution width
Hypertension
Mean arterial pressure
Total testosterone levels
Total cholesterol levels
Cardiovascular disease
Triglyceride levels
Apolipoprotein B levels
Sex hormone−binding globulin levels
Urate levels
Low density lipoprotein cholesterol levels
C−reactive protein levels
QT interval
Atrial fibrillation
Electrocardiogram morphology
PR interval
Monocyte percentage of white cells
Granulocyte percentage of myeloid white cells
Basophil count
Platelet distribution width
Monocyte count
Eczema
Multiple sclerosis
Rheumatoid arthritis
Systemic lupus erythematosus
Hypothyroidism
Lymphocyte counts
Asthma
d
Fraction high CACO−V
Average chromatin accessibility
GWAS SNP
FALSE
TRUE
Reproducibility of associated tissue specific CA at GWAS SNPs
e
Correlation of fold change
SNP DHS rank
Conservation scores
CACO−V
CACO−V_Manual
LECIF
PhastCons4Way
PhyloP4Way
BasePercent

Figure 1.5. Open chromatin conservation scores are predictive of reproducibility of signals between human and mouse at Human GWAS loci. (a) For two pairs of manually matched samples, highly variable DHS and high CACO-V score show higher correlation of fold change between human and mouse compared to CACO-V-Manual and other conservation scores. Line plot shows mean +/- SE of correlation. (b) For one pair of manually matched samples, DHS with high CACO-B scores are more likely to be open in both species when open in one of the species compared to scores based on one species only and other conservation scores. (c) Association of GWAS disease/trait with tissue or cell types. (d) Loci with GWAS SNPs are more likely to be high CACO-V after controlling for average chromatin accessibility. (e) Higher correlation of fold changes of associated tissue to baseline are observed between human and mouse for GWAS SNP DHS when ranked by CACO-V score compared to other conservation scores.

Functional conservation as a resource for annotating human phenotype-associated genetic variants

Using FUNCODE CACO scores, we annotated all human trait-associated SNPs in the NHGRI-EBI Catalog of human genome-wide association studies (GWAS) overlapping with alignable human DHSs, referred to as regulatory SNPs. Regulatory SNPs associated with GWAS traits showed significantly higher CACO-V scores than regulatory SNPs not currently associated with any GWAS trait after controlling for mean chromatin accessibility level (**Fig. 1.5d**).

For researchers who want to build mouse models to study human trait-associated SNPs, we reasoned that it makes more sense to study regulatory SNPs with high CACO-V scores. To verify this, we first associated GWAS traits with tissue-specific regulatory activities by testing the

enrichment of tissue-specific loci for each trait (**Fig. 1.5c**, Methods). All the associated tissue and traits were summarized in **Supplementary data 1.16**. For each GWAS trait, we then asked if the tissue-specific signals in human regulatory SNP loci can be reproduced in the same tissue in mice among conserved loci defined by different conservation scores. CACO-V scores were computed by leaving out the tissue type to be evaluated. CACO-V substantially outperformed the other conservation scores in predicting conserved tissue specificity (**Fig. 1.5e**). For example, in the top 1,000 most conserved regulatory SNP loci according to different scores, 70% of the loci with tissue-specific activities in human had the same tissue-specific activity in mouse if ranked by CACO-V, compared to only 50-65% if ranked by LECIF, PhastCons, PhyloP, PIB, or CACO-V-Manual. This shows FUNCODE allows better prioritization of GWAS SNPs for animal studies.

For example, we observe that GWAS SNPs associated with multiple sclerosis were enriched in DHS that were T cell specific. For those loci that had T cell specific activity, only a subset of them have T cell specific activity in mouse, and can be identified by selecting loci with high CACO-V scores. (**Fig. 1.2**) CACO-V score can be used to effectively prioritize a large number of candidate regions as it does not require knowing the particular tissue of interest a priori.

Functional conservation as a tool for cross-species integration of single-cell genomic data

FUNCODE can facilitate cross-species integration of single-cell genomic data. For instance, the standard approach for integrating human and mouse single cell ATAC-seq (scATAC-seq) data begins with selecting features (i.e., peaks or DHSs) informative for integration and then uses these features to embed cells from two species into a common space. This integration protocol

may be improved by additionally requiring features to be functionally conserved across species. To demonstrate, we integrated published human and mouse sciATAC-seq data from four tissues (lung, liver, heart, large intestine) [13,14]. Following the standard protocol [15], we ranked alignable DHSs in each species by the number of cells with non-zero read count and identified the top 300,000 DHSs as an initial feature set. We then filtered the initial feature set using different conservation scores to obtain the final feature set for integration. As a control, a final feature set was also constructed using the highest average rank in human and mouse in terms of the number of cells with non-zero read count but without applying any conservation filter. We applied Seurat [10] to integrate human and mouse data, using the same number of features selected by different methods.

To evaluate the integration performance, cell type annotations independently provided by the original publications were used as the gold standard. In the integrated low-dimensional space, we transferred cell type labels from human to mouse using a mutual nearest neighbor approach (See Methods) and evaluated the consistency between the transferred labels and the gold standard true mouse cell type labels using Adjusted Rand Index (ARI). A high ARI indicates high integration accuracy. Keeping the final feature number the same, the CACO-V filter yielded the highest ARI. Its ARI was substantially higher than the standard approach without the conservation filter and conservation filters by LECIF, sequence-based conservation scores (phastCons, PhyloP, PIB) and CACO-V-Manual (**Fig. 1.6a-d, Fig S1.3**). As an example, most mouse hepatocytes were correctly annotated as hepatocytes by CACO-V after label transfer, but the percentage of correctly annotated hepatocytes decreased substantially using other conservation scores or without using any conservation filter (**Fig. 1.6e**). Similarly, most mouse endothelial cells were correctly annotated by CACO-V, but the annotation accuracy decreased for other methods (**Fig. 1.6e**).

The improved integration enables more accurate comparative analysis of regulatory programs between species. For example, we analyzed differential chromatin accessibility between hepatocytes and endothelial cells and tried to identify differential signals conserved between human and mouse. As a ground truth, we first identified DHSs with conserved differential chromatin accessibility using true cell type labels provided in the original publications. We then asked how well one could recover this ground truth if the true mouse cell type labels were unknown. To this end, human cell type labels were transferred to mouse cells based on the integration, identifying mouse cells matching with the human cell types. Using the matching results, we computed differential pseudo-bulk chromatin accessibility between the two cell types in each species, and identified the conserved differential DHSs. We compared the list of conserved differential DHSs from the integrated cell labels to the list based on the true cell type labels, and found that conventional integration recovered a much lower fraction of the true conserved differential DHSs compared to the integration based on CACO-V. Moreover, CACO-V also outperformed LECIF, sequence-based conservation, and CACO-V-Manual (**Fig. 1.6f**). Thus, applying the FUNCODE filter helped one better identify conserved gene programs between human and mouse that would be missed by the conventional single-cell integration protocol.

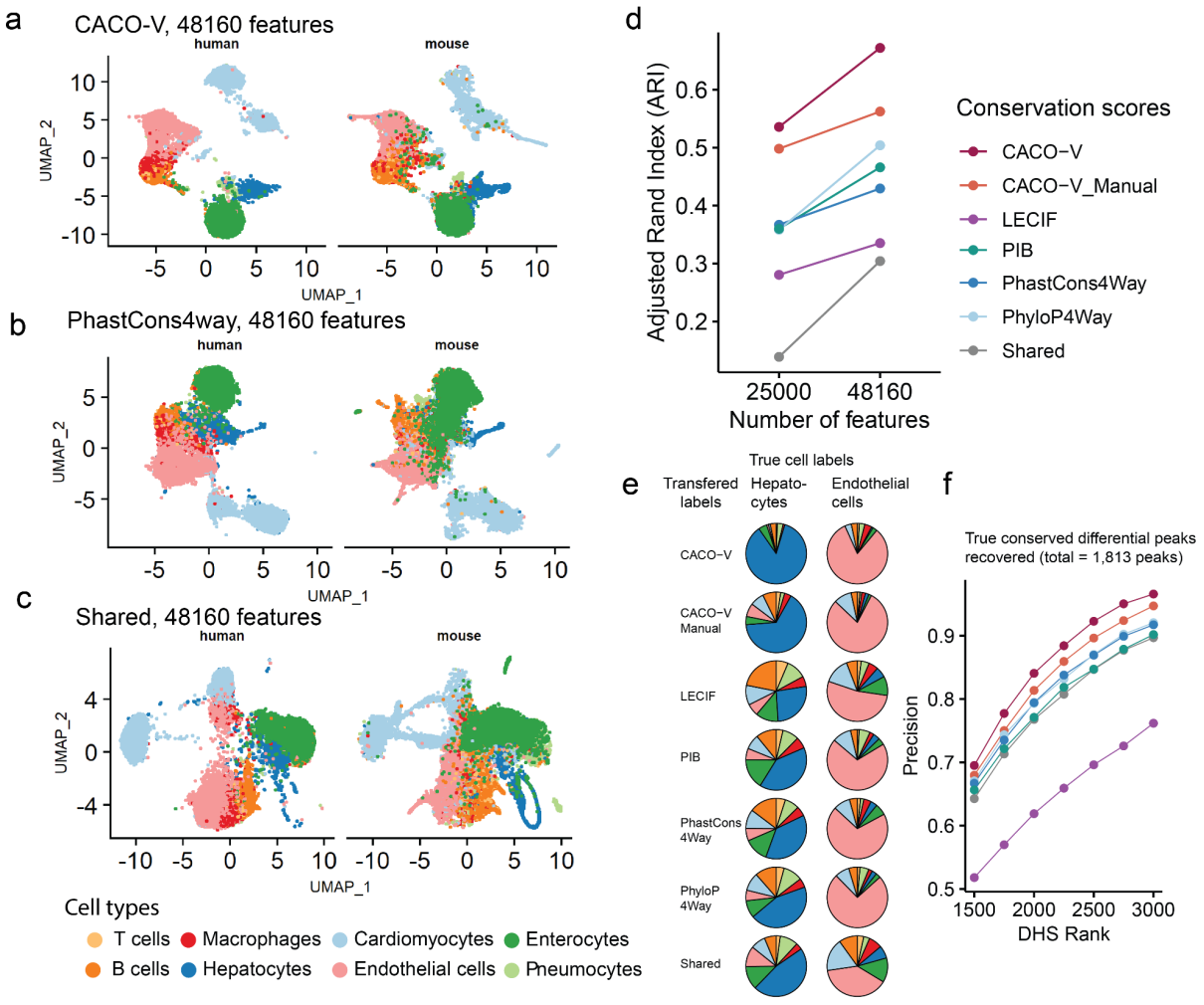

Figure 1.6. CACO-V score facilitates integration of single cell ATAC-seq data across species. (a-c) Visualization of Seurat integrated results when selecting top 48,160 features ranking by CACO-V, PhastCons4Way or Shared. (d) Integration performance of different feature selection methods as evaluated by Adjusted Rand Index (ARI) of label transfer accuracies. ARI=0 is completely random. ARI=1 is perfect accuracy. (e) Transferred label composition for human Hepatocytes and endothelial cells. Integration based on CACO-V transferred the most cells to the correct mouse labels. (f) Percentage of peaks differential in both human and mouse can be recovered the most based on integrated cell type labels using CACO-V scores.

Discussion

In summary, FUNCODE provides a new framework for scoring cross-species functional conservation of DNA elements without requiring manual sample matching. With its unsupervised sample matching, FUNCODE can be conveniently applied to new samples, making the updates of conservation scores easier when more data become available in the future. The data-driven sample matching also enabled more effective use of existing samples, yielding scores that better capture the functional conservation of DNA elements as demonstrated by our results.

One potential limitation of unsupervised sample matching is that it requires a diverse panel of samples in each species to cover a common space of major tissue or cell types. This limitation, however, will be mitigated by the continued growth of functional genomic data. The ENCODE consortium, in particular, has generated data from diverse sample types in both human and mouse, allowing this approach to be applied to a number of functional genomic data modalities. In this study, we demonstrated and benchmarked FUNCODE using chromatin accessibility data. For evaluating CACO scores, we also applied FUNCODE to H3K4me1 ChIP-seq and RNA-seq data, demonstrating its generalizability to other data types. Systematic scoring of other ENCODE data types using FUNCODE is ongoing and will be reported in a future ENCODE Consortium publication.

In this study, we focused on identifying the most conserved DNA elements which showed both DNA sequence conservation and functional conservation. As such, we only scored alignable DNA elements between human and mouse and did not consider evolutionary turnovers where functions of a DNA element in one species are transferred to a different DNA element without sequence homology (i.e., non-alignable) in another species. Using the FUNCODE aligned samples and by changing the input DNA element pairs from the alignable pairs to non-alignable pairs, our framework in principle can also be applied to score the functional conservation of two

non-alignable DNA elements. Systematic scoring of these turnover events will also be addressed in the future ENCODE Consortium publication.

The FUNCODE pipeline along with conservation scores generated in this study, including CACO, HICO-H3K4me1 and GECO scores, and CACO-annotated GWAS SNP catalog are made available as an public resource that can be downloaded from ENCODE portal and github.

Methods

Mapping of human DHS to mouse genome

List of DHSs for the GRCh38 genome was obtained from [11]. (ENCODE accession: ENCFF503GCK) List of DHSs for the mm10 genome was obtained from the ENCODE data portal. (ENCODE accession: ENCFF910SRW). To find regions of the mouse genome that align to each human DHS, the pairwise alignment results produced by BLASTZ were obtained from the UCSC genome browser.

https://hgdownload.soe.ucsc.edu/goldenPath/hg38/vsMm10/hg38.mm10.net.axt.gz

First, each summit of the human DHS as specified in the ENCODE annotation was mapped to the mouse genome. To ensure that mapped regions have the same width, the human and mouse submit were each extended to 200bp wide.

Computing existing conservation scores for DHSs

PhastCons and PhyloP: Human hg38 PhyloP and PhastCons files in bigwig format were collected from the UCSC genome browser (https://hgdownload.cse.ucsc.edu/goldenPath/hg38/). The UCSC tool 'bigWigAverageOverBed' was used to compute the average PhastCons and PhyloP score over each DHS region. The output gave the average PhastCons and PhyloP score over the DHSs. For comparisons in the paper, PhastCons4way and PhyloP4way were used as these were the smallest set that contained both human and mouse.

Percent Identical Base-pairs (PIB):

The percentage of identical bases were computed based on the reference genome sequences of the aligned regions. A fraction was computed by dividing the number of identical bases by the length of the region in human genome.

LECIF: LECIF score in bigwig format was available for hg19 and mm10 in the github repo (https://github.com/ernstlab/LECIF). The average LECIF score was computed for each DHSs. For DHS defined on hg38, the regions were first liftOver to hg19.

ENCODE chromatin accessibility data processing

All ENCODE human and mouse DNase-seq and ATAC-seq data available by September, 2020 were collected via the ENCODE data portal, including 689 DNase-seq and 137 ATAC-seq experiments from human, and 88 DNase-seq and 32 ATAC-seq experiments for mouse. Read alignment files to genome assembly GRCh38 and mm10 were downloaded. For each sample, the number of read centroids that overlap with each DHS were counted. For paired-end reads, the read centroids were taken to be the average of the ‘5 end of the forward read and the ‘3 end of the reverse read. For single end reads, the centroids were taken to be the midpoint of the read.

The library size for each sample was computed as the total number of reads that aligned to the genome. The counts and library sizes of the technical replicates were first averaged for each ENCODE experiment. Then, counts of each experiment are normalized by dividing by the library size, and subsequently multiplying by a factor of 1e8.

Tissue and cell type annotations of ENCODE experiments

Supplementary data 1.1-2 includes the tissue and cell type annotations for the ENCODE chromatin accessibility experiments collected.

Computing standardized variance for DHSs

Standardized variance for each DHS was computed by applying variance stabilizing transformation (vst) following a similar procedure to the Seurat 3.0. First, variance and mean were calculated for each DHS and log-transformed. A lowess curve was fitted to the log10(variance) to log(mean) relation, from which the fitted value of a lowess model was taken to be the expected standard deviation. Each feature was then standardized by subtracting the mean and dividing by expected standard deviation. Samples with a standardized value of larger than the square root of total sample size were clipped off. Finally standardized variance was taken to be the variance of the standardized values. Unlike Seurat, which applied the ‘vst’ procedure to raw count data, standardized variances were always computed using library size normalized data.

Assay effect correction between DNase-seq and ATAC-seq data

Assay effects between DNase-seq and ATAC-seq were first removed before cross-species integration. We note here that it was important that this integration step adjusted the data in the DHS feature space as opposed to the features in the low dimensional space. Correcting in the DHS space allowed for assay effects on individual features to be corrected, making downstream evaluations of conservation possible at individual locus. Using DNase-seq data as the reference and ATAC-seq as the query batch, the normalized DHS data were corrected using the Seurat

anchor (v3.0) method. Anchors were identified across the two assays using 30 canonical variables following CCA of the 10,000 shared variable features across assays with anchor finding parameters ‘k.anchor=5’ and ‘k.filter=10’. Subsequently, the ‘IntegrateData’ function was called to calculate the correction vectors for the ATAC-seq samples, which were the Gaussian kernel weighted anchor differences. The weighting is set to use 20 canonical variables and can include up to 20 neighbors (**Figure S1.1**).

Unsupervised integration of chromatin accessibility data between human and mouse

Many unsupervised algorithms have been developed for integrating omics data across batches, biological replicates, technologies, and species. [10,16,17] The strategies that showed most success originated from the work of Haghverdi et al. [18]. The mutual nearest neighbor (MNN) criteria has proven successful in cases where the cell state compositions differ across sources of data to be integrated. To integrate two datasets, data were first mapped onto a common low dimensional space. Subsequently, matched pairs were identified by applying a mutual nearest neighbor criteria within the common space. Finally, the MNN pairs were used as reference points or anchors for data adjustments. We also use the MNN procedure to identify in silico matches of human and mouse samples. Instead of trying to batch correct, the MNN pairs were used as matched sample pairs for evaluating consistent tissue-specific activities across species. First, aligned DHS pairs with counts lower than 20 in any sample after normalization in either human or mouse were excluded. Next, standardized variances were computed for each DHS for human and mouse separately. To integrate between human and mouse, we focused on features that have high standardized variance in both species. For this purpose, each DHS is ranked by their standardized variance in both species. Features that ranked top 29% in both human and

mouse (212,904 (11.6%) DHSs) were used as the initial set for integration. Further, the Seurat v3 routine was employed to conduct dimension reduction and anchor finding. The Seurat v3 routine also computes an anchor score for each anchor pair with a value between zero and one that measures the quality of the anchor pairs. The function `FindIntegrationAnchors` was applied to further select top 10,000 variable features for the integration. The canonical correlation analysis (CCA) was used for dimension reduction into top 30 dimensions, and the parameters k.anchor = 5, k.filter = 20 and k.score = 30 were used for anchor finding and scoring. These parameters needed to be adjusted as the default were set for single cell analysis with a much larger number of cells compared to samples in our data. All anchor pairs identified were available in **Supplementary data 1.3**. It should be noted that downstream computation of conservation scores only make use of the anchors information and does not rely on actual integrated data or UMAP-reduced data. Integration and UMAP was conducted, however, for visualization in **Fig. 1.1a**. `IntegrateData` was applied to top 15 canonical variates with 'k.weight' set to 20, and UMAP is done on top 15 principle components (PCs).

Unsupervised integration of histone ChIP-seq data between human and mouse

Histone ChIP-seq data were downloaded via ENCODE data portal. (**Supplementary Data S1.5-6**) Unlike chromatin accessibility data, direct integration of H3K4me1 histone ChIP-seq data between human and mouse produced poor results due to the limited number of samples available (332 human experiments and 92 mouse experiments) and tissue types covered. However, we observed that it was relatively easy to computationally integrate histone ChIP-seq with chromatin accessibility data within each species as experiments were conducted on consistent biosamples within each species. Taking advantage of this fact, and the success of

unsupervised integration of chromatin accessibility, we first transferred Histone ChIP-seq data targeting H3K4me1 to chromatin accessibility space using the Seurat procedure, and subsequently integrated transferred data together with chromatin accessibility data across species. The H3K4me1 samples after transfer and integration were finally used to identify anchors between human and mouse.

To conduct the bridge integration, the CA is chosen as the reference data modality. A set of integration features were first selected for the reference data modality across species. In this case, the 10,000 features previously selected for CA integration were used. Histone ChIP-seq data were summarized on the alignable DHSs and normalized using the same pipeline as chromatin accessibility. Standardized variance was computed for each DHS and ranked. DHSs that ranked as the top 13% most variable in both data modality were selected for integration. Integration was conducted with the `FindTransferAnchors` function using CCA dimension reduction. `k.anchor=8`, `k.filter=20` and `k.score=40` were used. This integration was conducted for human and mouse histone data separately. To transfer all histone samples onto the CA space, `TransferData` function was called with `k.weight=10` on the cross species integration feature set. This procedure computes the equivalent of histone samples in the CA space, particularly for features that would be integrated across species in the ensuing step. Finally, to integrate both data modalities between human and mouse, transferred H3K4me1 data were concatenated with the DNase-seq data, and integrated between human and mouse using `k.anchor = 5`, `k.filter = 20`, `k.score = 30` and `k.weight = 20`. Using the integrated data across species, anchors were identified again with 'FindIntegrationAnhcors' with `k.anchor = 4`, `k.filter = 20`, `k.score = 25`. These anchor pairs and the normalized H3K4me1 data were used for downstream computation of HICO-H3K4me1 scores. Anchors identified are available in **Supplementary data 1.7**.

Computation of conservation scores

Conservation (CO) scores were computed using the in-silico matched samples (anchors). For aligned DHS i, let X_i be the vector of log normalized counts for human anchors and Y_i be the vector of log normalized counts for mouse anchors, and w be the anchor scores of each anchor pair computed by Seurat. The scores were used as weights for incorporating the anchor quality information. CO-V score is defined as the weighted Spearman's correlation between X_i and Y_i:

$$\text{CO-V}_i = \text{wSpearman}(X_i, Y_i; w)$$

CO-C score is defined as the minimum of the weighted fraction of anchors above some data derived species-specific cutoffs. The species specific cutoff is defined as the 80% quantile of all mean DHS activities among the aligned DHSs in human and mouse, respectively.

$$\text{CO-C}_i = \min\left[mean(X_i > c_h; w),\ mean(Y_i > c_m; w)\right]$$

If a sample has log-normalized count above the species-specific cutoff, it is considered a high signal at the DHS. Intuitively, CO-C scores close to one indicate that the DHS is open in most of the matched samples and is considered constitutively open.

To evaluate the significance of CO scores, human DHSs were randomly paired with a mouse DHS that is alignable to some human DHS. First, a total of 100k human DHSs were randomly sampled, and another 100k alignable mouse DHSs were randomly sampled and paired to the human DHSs. Next, CO-C and CO-V scores were computed for all of the random pairs, and empirical p values were computed based on the null distribution. The p values were further adjusted for multiple testing with the Benjamini Hochberg procedure. A false discovery rate cutoff of 0.1 was used for calling the conserved elements of each category. All scores and annotations for DHSs were summarized in **Supplementary Data 1.8**.

ENCODE candidate Cis-Regulatory Elements (cCRE) annotations of DHSs

ENCODE cCRE annotations were collected from the ENCODE data portal (human cCRE ENCODE accession: ENCFF535MKS, ENCFF196MIP, ENCFF036NSJ, ENCFF262LCI, ENCFF379UDA, mouse cCRE ENCODE accession: ENCFF116KIO, ENCFF810COU, ENCFF701CRY, ENCFF996ZSG, ENCFF404GUR). Overlaps between alignable DHSs with each of the cCRE categories were computed. A total of 13,083 aligned DHSs (1.59% of all annotated cCREs) were annotated with more than one cCRE category. PLS annotation was prioritized when a DHS overlaps both PLS and pELS (11,151 DHSs). Similarly, pELS was prioritized when a region overlaps both pELS and dELS (1,340 DHSs). Further dELS was given higher priority than DNase-H3K4me3, and CTCF-only had the lowest overall priority. As a result, 821,737 (44.8%) aligned human DHSs were annotated as one of the cCRE types, but only 362,645 (20.0%) of all the aligned mouse DHSs were annotated as one of the cCRE types. 913,965 (49.8%) were labeled as either human or mouse cCRE and 198,478 pairs (10.8%) were labeled as cCRE of the same category, referred to as the consistent cCREs.

Gene Expression COnservation (GECO) scores

ENCODE RNA-seq data were collected from the ENCODE data portal. All human and mouse polyA+ RNA-seq and total RNA-seq gene quantification data uniformly processed with the ENCODE pipeline were downloaded. (**Supplementary Data 1.9-10**) Human and mouse homologous gene annotations (HomoloGene) were downloaded from the Mouse Genome Informatics. (www.informatics.jax.org/downloads/reports/HOM_MouseHumanSequence.rpt) Only 16,468 unique homologous gene pairs were used. RNA-seq data were integrated using Seurat v3 with top 3,000 genes that were highly variable in both species. CCA was first performed, and top 30 dimensions were used. `k.filter` was set to 40 for anchor finding. For

visualization in **Figure S1.2a**, log-normalized count data were integrated and top 30 principal components from the integrated data were used for UMAP visualization. The anchors identified were shown in **Supplementary Data 1.11**. Conservation scores for all homologous gene pairs were available in **Supplementary Data 1.12**. The R package 'enirchR' was used for the gene ontology analysis based on the set of high GECO-V genes and high GECO-C genes. [19] Gene ontology results were available in **Supplementary Data 1.13**.

Gene context annotations

The GENCODE gene annotation was obtained from ENCODE (Human Accession: ENCFF159KBI, Mouse accession: ENCSR884DHJ) The 'GenomicFeatures' package was used for annotating exons, introns, CDS, 5' UTR and 3' UTR. Promoters were defined to be 1,000 bp upstream the transcription start site (TSS) [20]. A DHS was annotated with one of the gene features if it overlaps more than 50% with the feature. If a DHS overlapped with none of the gene features, it was annotated as intergenic.

Gene regulatory domain and DHS with consistent candidate target gene

Gene regulatory domains were defined similar to [21]. First, a basal domain was defined for each gene that is from 5kb upstream the TSS to 1kb downstream the TSS. Next, each gene's basal domain was extended in both directions until either the basal domain of a neighboring gene had been reached or until a maximum distance of 1MB was reached. The gene regulatory domain of a gene contains at least its basal domain. A DHS is associated with a gene if it is within its gene regulatory domain and a DHS can be associated with multiple genes. For a pair of aligned human and mouse DHSs, all pairs of unique homologous genes that are associated

with their respective DHS were first listed and distance of DHS to TSS were computed for each homologous gene pair and ranked for human and mouse. Finally, the consistent candidate target gene is chosen to be the gene with the smallest sum of human and mouse ranks (ties allowed). 93.17% of all aligned DHSs have at least one consistent candidate target gene identified, among which 6.1% had more than one consistent candidate target gene identified.

Manual matching of human and mouse tissue and cell types

Human and mouse samples were manually matched based on the experiment metadata. Two samples were considered a match if their biosample terms, lifestages (embryonic, child, or adult) and treatments were exactly the same. In total, 22 matched tissue or cell types were identified (**Supplementary data 1.14**).

Leave-out samples for cross validation

For evaluating the CACO-V score, a pair of samples were left out at each fold, the pairs were randomly selected from the set of all manually matched samples each time. **Supplementary data 1.15** shows the pairs of biosamples that were held out for each fold of the cross validation. For evaluating the CACO-C score, each manually matched sample was left out at a time.

GWAS catalog data processing

All GWAS associations (v1.0) were downloaded from the NHGRI-EBI GWAS catalog. The data was pruned following a previously reported procedure [22]. First, the GWAS SNPs were grouped according to study (PubMed ID) and disease/trait. For each study and disease/trait

combination, all SNPs were ranked by their p values. Next, each SNP was added to the final list by checking if there were a SNP within 5,000 bp in the current list, that is, a SNP is omitted if there existed a more significant SNP within 5,000 bp. Finally, the set that overlaps with HLA loci was removed. After pruning, 127,061 unique SNPs were left, covering 4,729 unique disease/trait.

GWAS disease/trait association with tissue and cell types

For each GWAS disease/trait, the associated SNPs were overlapped with alignable DHSs, and the set of DHSs is referred to as the associated DHS. We focused on disease or trait with at least 50 associated alignable DHSs. Then for each tissue a fold change (FC) and p-value (Wilcoxon rank sum test) is computed comparing the tissue-specific activity to average activities of all other tissues. The p values were corrected with the Benjamini Hochberg procedure for multiple testing. A DHS was classified as tissue-specific if it had logFC > 0.5 and FDR < 0.05. The fraction of associated DHSs that were specific to each tissue were then summarized into a matrix. Finally, to identify GWAS disease or trait associated with each tissue, the percentages for each tissue is standardized. The associated diseases or traits were selected as those with z-scores larger than 2 (**Figure 1.5c**).

Single cell ATAC-seq data processing and integration

Bam files from the mouse sciATAC atlas [13] were collected from the website (https://atlas.gs.washington.edu/mouse-atac/data/). The reads were aligned by the author to the mm9 genome, hence the aligned mm10 regions were first liftOver to mm9, and read centroids that overlapped each aligned GRCh38 region were counted. Fragment files from the human

sciATAC atlas [3] were collected from the website (http://catlas.org/humanenhancer/). Read centroids that overlapped each aligned GRCh38 region were counted. For the two atlas, four common tissues were identified, including lung, liver, heart and large intestine. Specifically, the following samples were included from the two studies:

Human samples:

"lung_SM-A62E9", "lung_SM-A8WNH", "liver_SM-A8WNZ", "liver_SM-A8WNZ", "heart_lv_SM-IOBHO", "heart_lv_SM-JF1NY", "heart_atrial_appendage_SM-JF1NX", "heart_atrial_appendage_SM-IOBHN", "colon_transverse_SM-A9HOW", "colon_transverse_SM-A9VP4", "colon_transverse_SM-ACCQ1", "colon_transverse_SM-BZ2ZS", "colon_transverse_SM-CSSDA", "colon_sigmoid_SM-AZPYO" "colon_sigmoid_SM-JF1O8"

Mouse samples:

"Lung1_62216", "Lung2_62216", "Liver_62016", "HeartA_62816", "LargeIntestineA_62816", "LargeIntestineB_62816"

The cell types as labeled by the respective publications were mapped between human and mouse according to **Supplementary data 1.17**. Signals were binarized by setting all DHSs with more than one read to one. For human samples, cells with fewer than 1,000 DHS detected were first dropped. Next, only cell types whose label could be matched across species were kept. (77.3% of all human cells, 79.8% of all mouse cells) Finally, only cells with more than 900 DHSs detected were kept. Hepatocytes accounted for a larger fraction of all cells in mouse data (22%) than in human data (13%). 2,500 mouse hepatocytes out of the total 4,913 were dropped randomly before integration.

In order to select an initial set of features as candidates for data integration, each DHS was ranked by the number of cells in which it was detected in, for human and mouse separately. Top 300,000 DHSs with the largest sum of human and mouse rank were selected. For the set of

300,000 initial DHSs, a median of 100 (0.14%) human cells and 99 (0.46%) of mouse cells were detected.

For the default integration (Shared), the initial features were further filtered with the overall rank of the number of cells detected to 25,000 and 48,160 features. For all other methods, the initial set of DHSs were ranked by the respective conservation scores. With each feature set, the data were integrated with top 30 dimensions of CCA, and arguments k.anchor = 40, k.filter = 100 and k.weight = 100.

Chapter 2

Quantitative fate mapping: Reconstructing progenitor field dynamics via retrospective lineage barcoding

Other authors: Claire M. Bell, Abel Sapirstein, Soichiro Asami, Kathleen Leeper, Donald J. Zack

Corresponding authors: Hongkai Ji, Reza Kalhor

Introduction

Embryonic development is the genesis of complex body plans in the animal kingdom. It starts with the zygote, a single cell in a totipotent state, and ends with millions of specialized terminal cells organized in tissues. In between, dividing cells assume increasingly diverse but decreasingly potent intermediate progenitor states. Each progenitor state specifies the ensuing states that its cells may take thus directing them toward their terminal fates. Collectively, progenitor states orchestrate the formation of complex tissues by ensuring the emergence of all terminal cell types in harmony. Therefore, delineating how progenitor states relate to each other and terminal fates—the cell fate map—is critical for our understanding of normal and dysregulated development as well as our ability to generate engineered tissues. However, mapping cell fate is challenging due to the number of progenitor states involved, their interdependent relationships, and the time lapse between terminal fates and most of the progenitor states that help create them.

The recent advances in genome engineering and sequencing have inspired a new approach for interrogating cell fates during development: retrospective lineage analysis using synthetic or natural somatic barcodes [23,24]. These approaches rely on the somatic accumulation of random mutations in the genome during development. Each mutation is inherited by the descendants of the cell in which it occurs; each descendant can add new mutations to the combination it inherited. This process marks each terminal cell with a combination of

mutations—a barcode—that encodes its phylogenetic relationship to the other cells [25]. Synthetic lineage barcoding, which relies on gene editing technologies to induce mutations, has been implemented in a variety of model systems including the zebrafish [26–29], fruit fly [30], and mouse [31–33]. Natural lineage barcoding, which relies on naturally-accumulating somatic mutations, has been primarily used in humans [34–37]. And while the mechanism, timing, and extent of mutagenesis differ between various implementations, the functional outcome remains the same: genomic barcodes that can retrospectively inform cell phylogeny within one organism at a single-cell resolution. As cells' fate decisions are also somatically inherited to their daughters through epigenetic mechanisms [38], these retrospective approaches hold unique promise for mapping cell fate because, unlike single-cell molecular profiling approaches, they can bridge time lapses between terminal cells and their progenitors that exist far earlier. Moreover, unlike prospective lineage tracing approaches, they allow parallel analysis of multiple progenitor groups and do not depend on the identification and manipulation of progenitors.

Despite this compelling potential, the full scope of the information that retrospective lineage barcoding can provide about the fate of the intermediate progenitor states remains unclear for several reasons. First, cell phylogeny is a function of cell divisions and most cell divisions in higher organisms do not accompany fate decisions. In the roundworm *C. elegans*, a unique model system in which almost all cell divisions give rise to daughters with different fates, the phylogeny of terminal cells is identical to the fate of their progenitors [39]. In more complex organisms, on the other hand, cell fate is often determined at the level of progenitor populations which also undergo cell divisions that are not associated with fate decisions, leading to divergences between phylogeny and fate [24,40,41]. As a result, adjacent liver hepatocytes of identical progenitor state history may have the maximum possible distance on the phylogenetic tree by being the descendants of different cells at the 2-cell stage. Second, while the progenitor states and their fate remain largely stereotyped within species, the phylogenetic histories of the

cell populations that assume those progenitor states can vary greatly from embryo to embryo due to stochasticity in fate decisions [36]. Third, single-cell lineage barcodes can be obtained for only a fraction of the cell population due to the practical limitations of single-cell sequencing. Whereas most mammals have millions of cells in each tissue, current technologies can only sequence thousands of single cells. Given the divergences between fate and phylogeny and the variable nature of the latter, it remains unclear how phylogenies derived from small samples can reliably inform organism-level fate maps. These considerations raise critical questions about the value of measuring cell phylogeny through barcoding approaches in complex organisms: What features of progenitor populations in embryonic development are reflected in the phylogeny of sampled terminal cells? Can retrospective barcodes, synthetic or natural, be used to extract these features and if so how?

To address these questions, here we systematically study the relationship between cell fate and cell phylogeny as derived using lineage barcodes. First, we examine how phylogeny of cells can be used to understand the dynamics of the progenitor states that gave rise to them. We define quantitative fate maps that specify the fate dynamics of a progenitor field—a collection of progenitor states [42] that give rise to a set of observed cells (**Figure 2.1**). We then establish a generative model of cell phylogenies based on predetermined quantitative fate maps (**Figure 2.1A**). Using generated phylogenies, we develop a strategy, ICE-FASE, to map the order of progenitor states and quantify their commitment time, commitment bias, and population size from phylogeny, thus reconstructing the original quantitative fate map (**Figure 2.1B**). We find that successful quantitative fate map reconstruction requires time-scaled cell phylogeny wherein branch lengths correlate with interdivision times. It further requires adequate representation of progeny of each progenitor state among the terminal cells of the phylogenetic tree. These results demonstrate that phylogeny of a small number of cells can inform progenitor dynamics at the organism level. Second, we examine whether phylogenies inferred from lineage barcodes

can reconstruct quantitative fate maps. We establish a barcoding mutagenesis model and parametrize it using experimental data from synthetic barcoding in mice. We then use the model to simulate realistic lineage barcodes based on time-scaled cell phylogenies (**Figure 2.1C**). To recover time-scaled phylogenies from lineage barcodes, we establish a generalizable and scalable method, Phylotime, that clusters cells based on temporal distances estimated from maximum likelihood of the mutagenesis model (**Figure 2.1D**). We demonstrate that Phylotime, coupled with ICE-FASE, enables quantitative fate map reconstruction from lineage barcodes (**Figure 2.1D,B**). Finally, we validate our methods in an experimental system with cultured stem cells grown based on preset fate parameters (**Figure 2.1E,D,B**). Overall, our results demonstrate how lineage barcodes from single time-point measurements can be used to decipher quantitative fate maps that capture the fate dynamics of progenitor populations long after their differentiation.

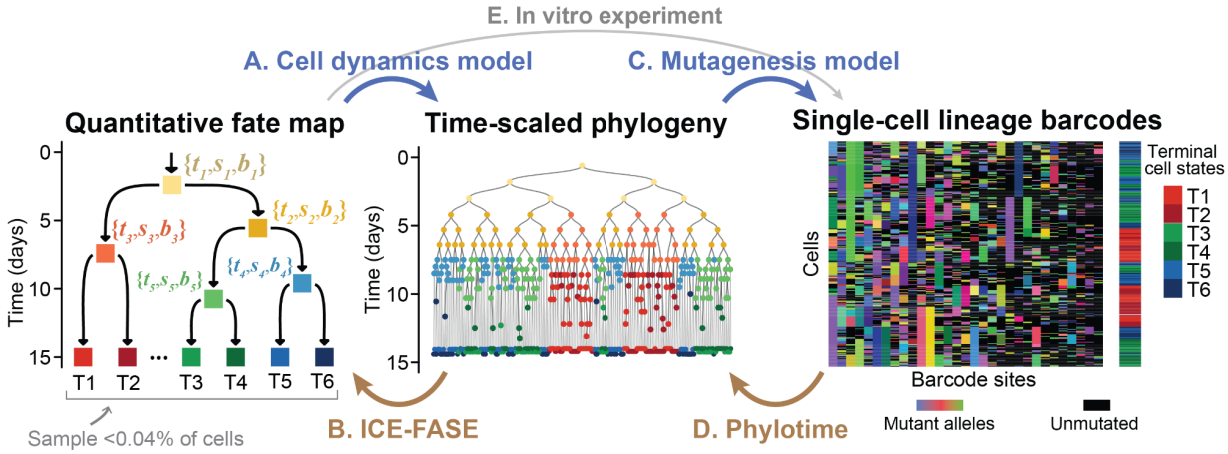

Figure 2.1. Graphical outline of this study, establishing quantitative fate mapping. Quantitative fate maps specify progenitor state dynamics and are used to generate time-scaled cell phylogenies and lineage barcoding results in terminal cells (blue arrows). Phylotime and ICE-FASE enable quantitative fate mapping by reconstructing first time-scaled phylogenies and then quantitative fate maps from lineage barcode data (brown arrows). $\{t_i, s_i, b_i\}$ are dynamic parameters of each progenitor state: commitment time, population size, and commitment bias.

An in vitro system is used to experimentally verify model assumptions (gray arrow and brown arrows).

Results

Modeling cell phylogeny based on quantitative fate maps

To study the relationship between cell phylogeny and fate dynamics, we began by establishing the quantitative fate map, a model that specifies a developmental process wherein dividing cells assume increasingly diverse but decreasingly potent progenitor states over time before arriving at their terminal fates. Each progenitor state is defined by its potency, which is the set of terminal states it is capable of producing, and is associated with a commitment event, when its cells transition to less potent downstream states. The commitment event confers each progenitor state three additional defining parameters: i) commitment time, defined as the time when a progenitor state's cells commit to the downstream states, ii) population size, defined as its number of cells at commitment time, and iii) commitment bias, defined as the proportions of its population committing to each downstream state. Progenitor states ultimately give rise to terminal states, which are the states of cells at the point they are sampled or observed. In summary, a quantitative fate map defines the fate dynamics of a progenitor field—a collection of progenitor states [42] that give rise to a set of observed cells.

We next constructed a panel of 125 test quantitative fate maps covering diverse developmental scenarios (**Figure 2.2**). Representing increasing field sizes, the maps are in three categories of 15, 31, or 63 progenitor states, producing 16, 32 or 64 terminal cell states, respectively. We label progenitor and terminal states with "P"s and "T"s followed by numerals, respectively (**Figure 2.3A**). Within each category, the topologies of the maps range from perfectly balanced to highly unbalanced (**Figure 2.2**) as measured by the BSUM imbalance index [43]. In more

unbalanced maps, progenitor states tend to split into increasingly unequal numbers of eventual terminal states (see Methods). Progenitor states in each map commit to two downstream states with prespecified ratios that were randomly drawn to cover a range of commitment biases within each map (**Figure 2.3A**, Methods). Commitment time for each progenitor state is also preset and randomly selected between $t = 1.8$ to 9.8 days subject to the topology constraints, based on the beginning of fate restrictions and the end of organogenesis in mouse development (**Figures 2.2 and 2.3A**, see Methods). Each fate map starts with a single founder cell at time $t = 0$ that divides according to the reported division rates during mouse development [44]. This cell division rate, which is the same for all progenitor states, together with their commitment biases determine the progenitor population size at different points in time (**Figure 2.3B, Figure S2.1**). Fate maps continue to $t = 15$ days when terminal cells can be sampled for observation. **Supplementary Data 2.1** provides a complete accounting of each map. Together, these fate maps represent a broad range of complex developmental scenarios.

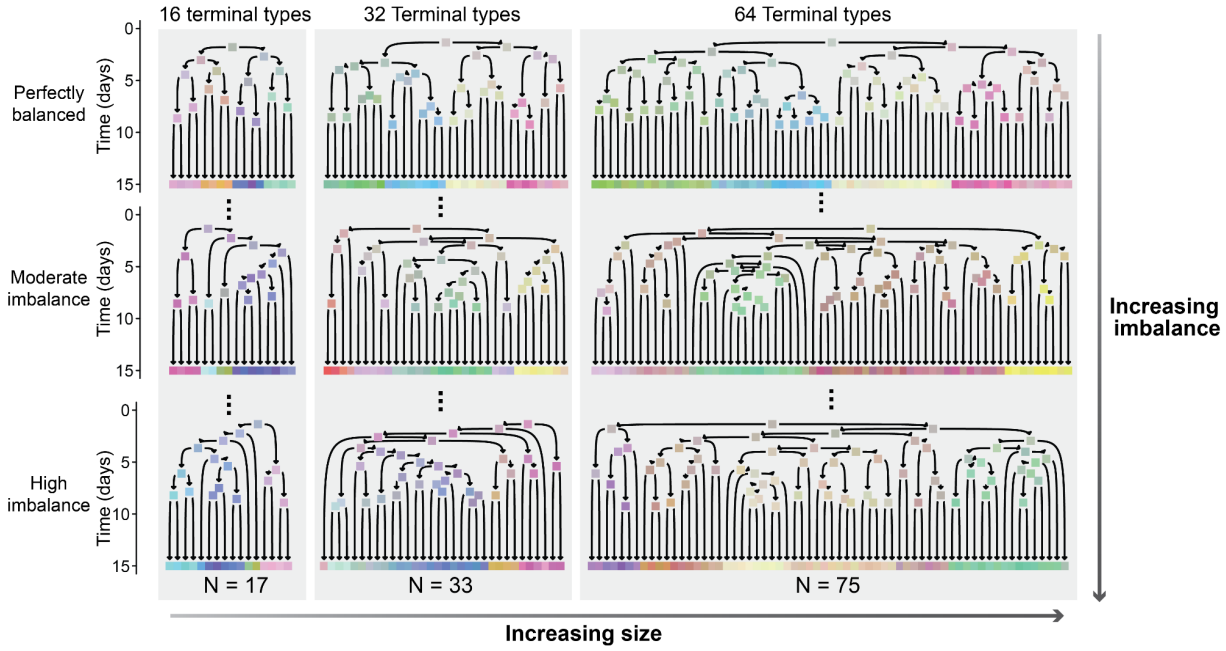

Figure 2.2. A test panel of 125 quantitative fate maps covering a broad range of developmental scenarios. The fate maps are categorized by three sizes of 16, 32, and 64 terminal cell types.

Three examples from each size are shown in each column. Within each size category, topologies range from perfectly balanced (top row) to highly unbalanced (bottom row). Arrows represent cell states, colored rectangles represent their commitment events. Rectangles at day 15 represent terminal states at the time of sample collection. Each map also specifies the cell commitment bias and division rate for each progenitor state, neither of which are shown in this figure.

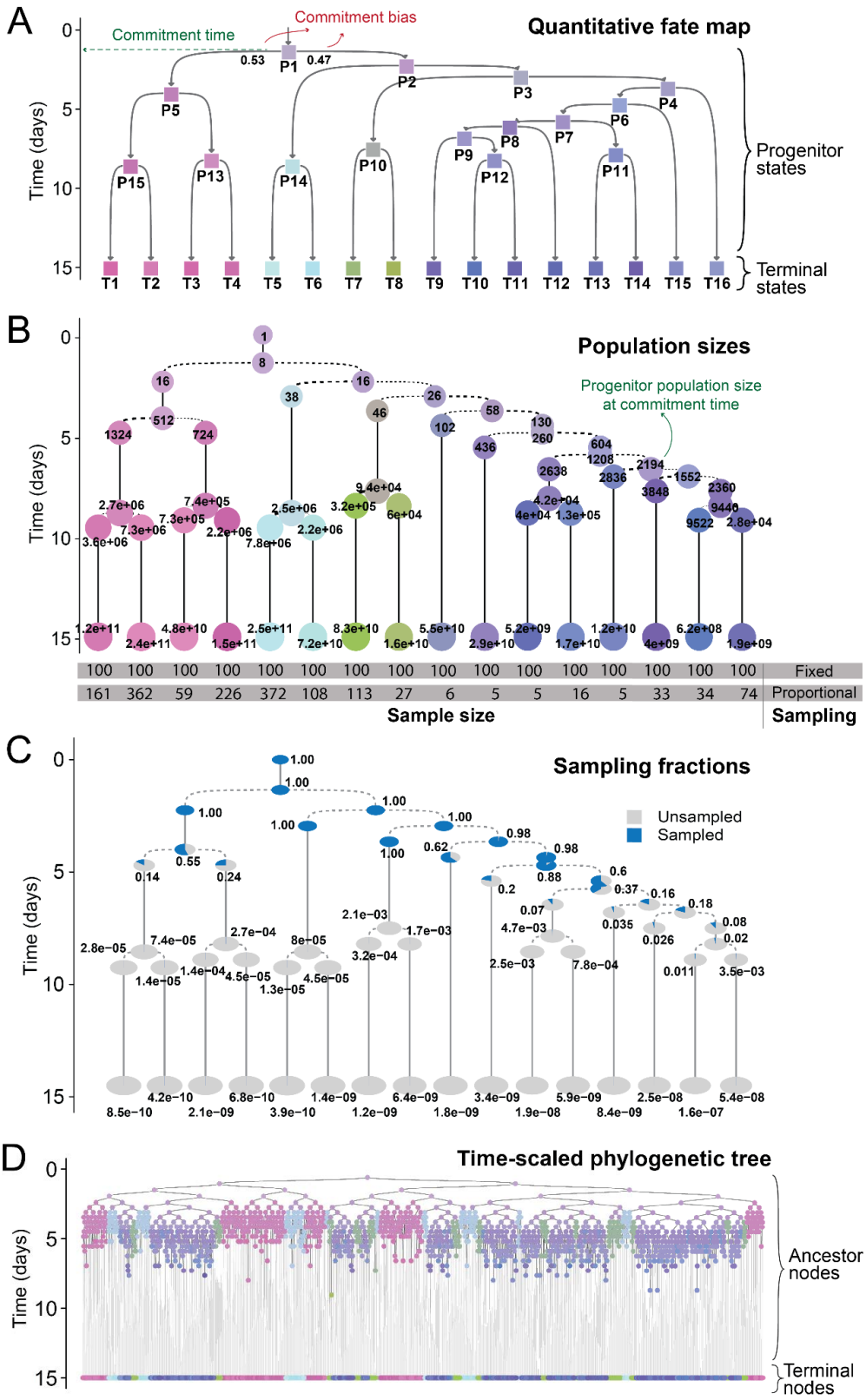
A
Quantitative fate map
Commitment time
Commitment bias
Time (days)
Progenitor states
Terminal states
B
Population sizes
Progenitor population size at commitment time
Fixed
Proportional
Sample size
Sampling
C
Sampling fractions
Unsampled
Sampled
D
Time-scaled phylogenetic tree
Ancestor nodes
Terminal nodes

Figure 2.3. Simulating time-scaled phylogenetic trees of sampled cells based on a known quantitative fate map. (**A**) Topology of a quantitative fate map specifies commitment time and bias for each progenitor state. (**B**) Quantitative fate map is populated with cell numbers at each time point based on cell division rates; these cells are apportioned to various states based on commitment biases. A number of cells are sampled from each terminal state based on the experimental sampling strategy as shown in the gray boxes on the bottom. (**C**) Sampling of terminal states dictates the fraction of progenitor population at each time point that are ancestral to sampled cells at all points in time. (**D**) Phylogeny of sampled cells is created by merging sampled terminal cells into ancestral nodes. Nodes are colored based on the terminal and progenitor states in A.

We next established a generative model of cell phylogenies for terminal cells based on a quantitative fate map. Generating the entire tree of cell divisions for billions of terminal cells (**Figure 2.3B**) is computationally impractical. To overcome this problem, our model draws inspiration from coalescent theory in population genetics [45,46]. It starts with a subset of terminal cells that are sampled from terminal states and merges them going backward in time to create their phylogeny in three steps (see Methods). First, terminal cells are randomly sampled in each state based on either fixed or proportional sampling (**Figure 2.3B**). Under fixed sampling, a fixed number of cells are sampled from each terminal state, similar to experiments where target terminal cell types are first identified and then collected using sorting or other methodology. Under proportional sampling, each terminal cell state is sampled based on its share of the total population, similar to experiments where cells are sampled without prior knowledge of their states (see Methods). This step establishes the number and state of terminal nodes in the phylogenetic tree. Second, the number of progenitor cells ancestral to sampled terminal cells at all earlier times is generated. To do so, starting from the number of sampled cells of each state in the terminal point in time ($t = 15$), the number of their ancestors in the

prior time point ($t - \Delta t$) is drawn based on the total cell number at each time point. This process is repeated recursively until reaching the founder at $t = 0$ (**Figure 2.3C,** see Methods for details). This step establishes the number of ancestors, or internal nodes, in the phylogenetic tree at each time point. Finally, nodes at each point in time are randomly connected to their progenitors in the earlier time point to create the branches of the phylogenetic tree. At the times of commitment events, cells of less potent states are connected with the combined pool of their ancestors. This leads to terminal cells gradually merging into increasingly common ancestors according to the numbers of ancestors over time established in the previous step. This sequence of merges forms the topology of the phylogeny and the times between merges form branch lengths (**Figure 2.3D)**. This approach generates time-scaled phylogenies for sampled cells based on their progenitors' fate map in a computationally efficient manner.

We applied our model to generate time-scaled phylogenies of sampled cells for each quantitative fate map in our test panel. For each map, we used both fixed and proportional sampling; for each sampling, we simulated two phylogenies. In all cases, an average of one hundred cells were sampled from each terminal state in each map. Together, these results represent 2,500 experiments (125 maps x 2 sampling strategies x 10 replicates) wherein phylogeny is determined for a small fraction of terminal cells derived from complex fate landscapes. With the benefit of knowing their underlying fate maps, we will use these 2,500 phylogenies in the ensuing sections to establish quantitative fate mapping algorithms and evaluate the reach and limitations of retrospective phylogenetic reconstruction approaches.

Reconstructing fate map topology using time-scaled cell phylogeny

These simulated time-scaled phylogenies, in their topologies and branch lengths, embody the order and timing of cell divisions that connect sampled cells (**Figure 2.4A**). Going from the root of a tree towards its terminal branches, nodes become gradually restricted in their potency, as can be observed by the diversity of their terminal progenies (**Figure 2.4A,B**). When and which

cell fate restrictions happen are clues to cell fate commitments. Therefore, to derive fate map topology from time-scaled phylogenies, we focused on putative cell fate restriction events. First, we annotated each node in the phylogenetic tree with the states of its observed terminal descendants, which we refer to as the node's observed fate (**Figure 2.4B**). Next, we compared the observed fate of each node with that of its two daughter nodes to identify nodes with both descendants having a more restricted fate (**Figure 2.4B**). For instance, if an internal node leads to terminal cell types T7, T8, and T9 but cells of type T7 and T8 are only seen in one of its branches and T9 only in the other, this node implies divergence of the terminal cell types T7 and T8 from T9. We refer to these nodes as FAte SEparation (FASE) between respective terminal fates. The prior example constitutes a T7–T9 and a T8–T9 FASE. The temporal distribution of FASEs that connect a pair of terminal cell types provides a measure of their developmental distance: cell types whose lineages separated early in development would be connected by FASEs that are closer to the root in the phylogenetic tree whereas those whose lineages separated later in development would be connected by FASEs that are closer to terminal branches (**Figure 2.4C**). Therefore, we defined the distance between terminal cell fates T1 and T2 as the mean temporal distance of T1–T2 FASEs to their terminal cells and compiled a pairwise distance matrix between all terminal cell types (**Figure 2.4D**, see Methods). Applying a distance-based clustering method (UPGMA) to this matrix, we obtained the topology of the fate map (**Figure 2.4E**). This fate map topology connects observed terminal cell states through a hierarchy of inferred progenitor states, each with the potency to give rise to two or more terminal cell types. We label these inferred progenitor states by "*iP"* followed by a numeral. To summarize, this strategy, which we call the FASE algorithm, produces a hierarchy of "inferred" progenitor states based on specific patterns of potency that result in the observed terminal states.

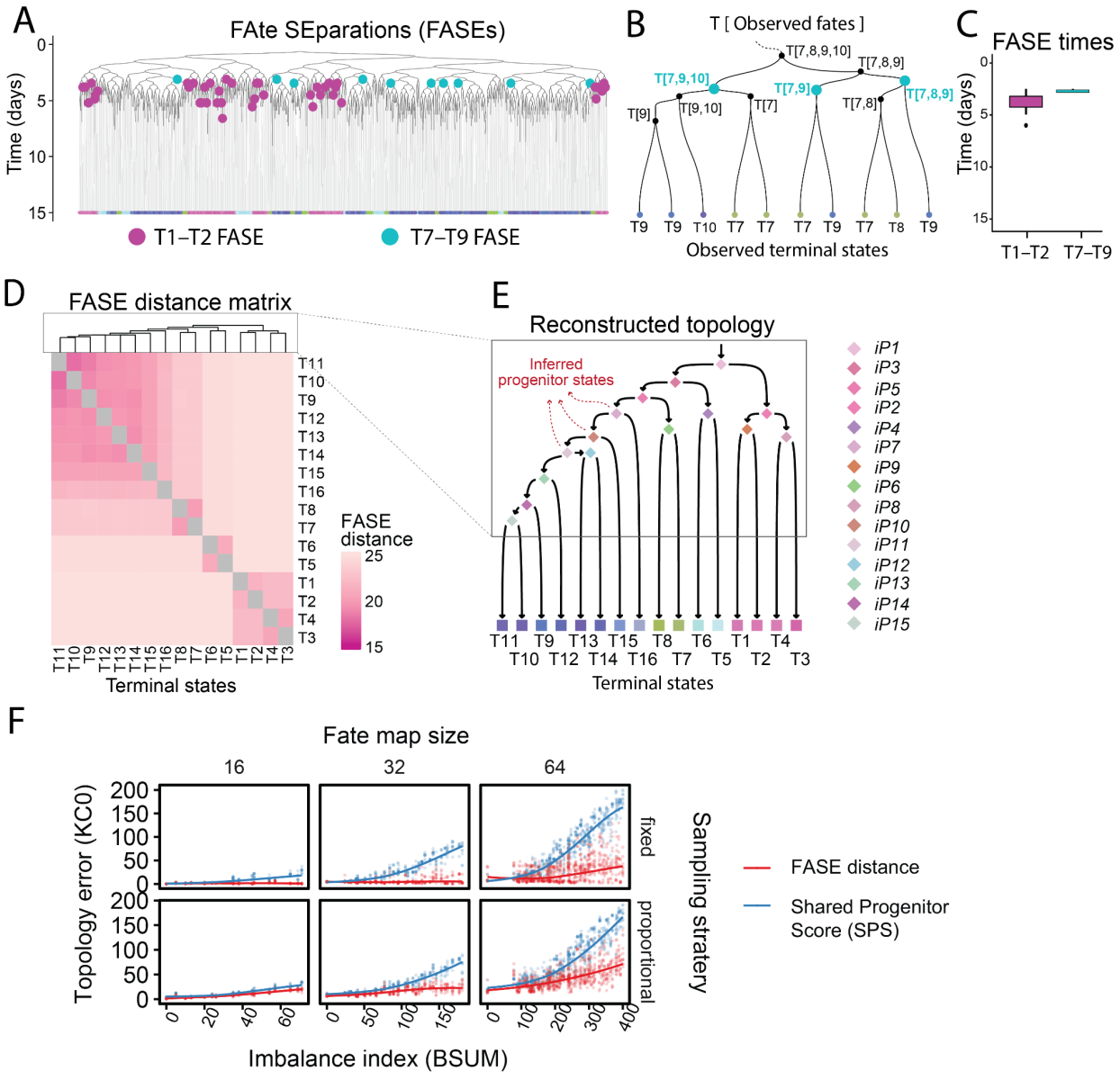

Figure 2.4. Reconstructing fate map topology from time-scaled phylogeny of sampled terminal cells. (**A**) A time-scaled phylogenetic tree of sampled terminal cells. Terminal cells of different states have been labeled by different colors. FAte SEparation (FASE) events between terminal states T1 and T2 (purple) as well as those between T7 and T9 (teal) have been labeled as examples. (**B**) A zoomed-in section of the tree in panel A where each internal node is labeled by its observed fate and T7–T9 FASEs have been labeled by teal circles. (**C**) Boxplot showing the distribution of T1–T2 and T7–T9 FASE times in the tree in A, from which the distance between

terminal states was derived. (**D**) Heatmap showing the FASE distance matrix for all pairs of terminal states in the phylogenetic tree in panel A. Dendrogram on top shows the hierarchical clustering results of the heatmap. (**E**) The fate map topology reconstructed by hierarchical clustering of the FASE distance matrix. Squares denote observed terminal states, diamonds denote inferred progenitor states. (**F**) Scatter plots showing the distance between fate map topologies inferred from the time-scaled phylogenetic tree and the true fate map topology in all 2,500 simulated phylogenies. The algorithm using FASE times for topology inference is compared to one using the Shared Progenitor Score (SPS) as a measure of similarity. Results are broken down in different plots by the total number of terminal states, the experimental sampling strategy, and by fate map imbalance on the x-axes. Solid lines show trend lines obtained through locally weighted smoothing (LOESS).

We applied the FASE algorithm to reconstruct fate map topology for each simulated phylogeny in our panel. For comparison, we also used the shared progenitor score (SPS) [32]. In each case, we compared the reconstructed topology to its corresponding true fate map using the Kendall-Colijin (KC) metric with its tuning parameter (λ) set to zero (KC0) [47]. A KC0 distance of zero indicates identical topologies; KC0 distances larger than zero indicate increasing differences between topologies. The results show that our FASE strategy faithfully reconstructs fate map topology in almost all tested phylogenies, outperforming SPS across the board (**Figure 2.4F**). Under fixed sampling, the FASE algorithm predicts perfectly accurate topologies in 16 and 32-terminal cell state fate maps irrespective of imbalance. It only shows modest decrease in accuracy with high degrees of imbalance under proportional sampling (**Figure 2.4F**). Moreover, fixed sampling outperforms proportional sampling, likely because it ensures better representation of small terminal populations. This result suggests that a fixed sampling strategy is more robust in lineage topology reconstruction using barcoding approaches than a proportional sampling strategy. Taken together, these results establish a robust and scalable

method to reconstruct fate map topology from the phylogeny of sampled cells that scales to complex fields.

Quantitative characterization of progenitor states using cell phylogeny

The fate map topology derived from the lineage of sampled cells effectively identifies a series of inferred progenitor states (*iPs*) each with a distinct potency and fate restriction pattern (**Figure 2.4E**). As the internal nodes of time-scaled phylogenies correspond to cells in these progenitor states, we sought to use them to further characterize the dynamics of these progenitor states at the cell population level. We therefore assigned each internal node in the phylogenetic tree to an inferred progenitor state or a terminal state based on fate map topology: given the observed fate of the internal node, we assigned it to the least potent progenitor state that contains the node's observed fate (**Figure 2.5A**). For example, a node with an observed fate of [T11, T13, T14] can be assigned a more potent inferred state (*iP11*) capable of [T9, T10, T11, T12, T13, T14] if the now-reconstructed topology of the fate map indicates that *iP11* differentiates into fates [T13, 14] and [T9, T10, T11, T12] (**Figure 2.5B**). To assess the fidelity of these assignments, we compared the inferred states of internal nodes in all phylogenies to their true states which are known, as these phylogenies were simulated. Where derived fate map topology differed from the truth, only correctly resolved progenitor states were considered. A progenitor state is correctly resolved if there exists a corresponding state in the true fate map with the same potency and commitment pattern (**Figure 2.5C**). The only error was assigning an internal node to a progenitor state less potent than its true state, which occurred, on average, for 19% of all nodes in each time-scaled phylogeny among the 2,500 experiments (**Figure 2.5C,D**). These errors are caused by failures to sample all the different terminal states a progenitor cell has led to and emerges when the number of sampled cells is small relative to the terminal population (**Figure 2.5E**). Since the sampling fraction of each progenitor state is known in simulated experiments (**Figure 2.3C**), we further examined its effect on resolving progenitor

states. We found that a great majority of progenitor states in our panel were correctly resolved if more than 25% of their population at the time of commitment were represented among sampled terminal cells. Conversely, only a minority of progenitor states with less than a 25% sampling fraction were correctly resolved. Going forward, we will distinguish progenitor states with more than 25% sampling as adequately-sampled and the rest as under-sampled.

To derive the commitment time of each progenitor state, we focused on when its inferred nodes transition to less potent states by defining a set of Inferred Commitment Events (ICEs). An ICE is a node whose inferred state is more potent than both of its descendants (**Figure 2.5B,F**). For example, in Figure 5B, when an internal node is assigned to *iP7* (capable of T9 through T16) and splits into nodes with assigned states of *T16* and *iP10*, we count this node as an ICE for *iP7*. The temporal distribution of each progenitor state's ICEs indicate when the progenitor commits to its downstream fates (**Figure 2.5G**). Unlike FASEs, which are defined for each pair of terminal fates, ICEs are defined with respect to inferred progenitor states. If a node is an ICE, it is also a FASE for at least some pair of terminal states, but the opposite is not necessarily true. ICE uses information from the fate map topology to identify a more confident set of nodes that represent state transitions. We thus defined the commitment time for a progenitor state as the mean of its ICE times (**Figure 2.5G**). Across all progenitor states, these ICE times captured the relative timing of commitment events as indicated by a high rank correlation (Spearman's ρ=0.850 for fixed sampling and ρ=0.928 for proportional sampling for progenitor states that were correctly resolved in fate map topology) (**Figure 2.5H**). However, across progenitor states that were adequately sampled, the ICE times captured the exact timing of commitments as indicated by a low root mean square error (RMSE=0.296 days for fixed sampling and RMSE=0.245 days for proportional sampling). Fixed and proportional sampling performed comparably, though fixed sampling performs better for under-sampled progenitors. These results establish ICE times as

estimates for the commitment times of the progenitor states from time-scaled phylogenies of sampled cells.

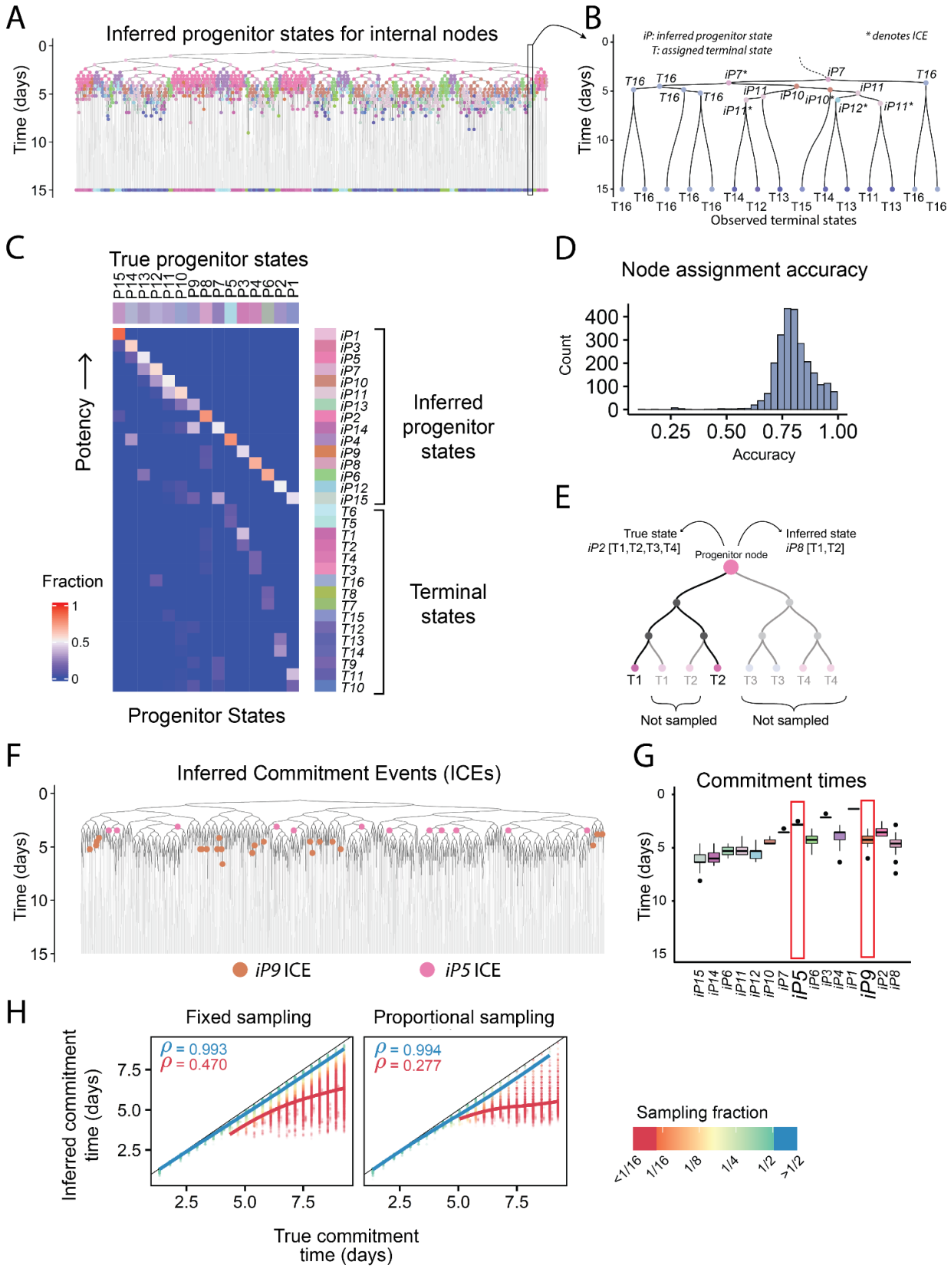
A
Inferred progenitor states for internal nodes
Time (days)
B
iP: inferred progenitor state
T: assigned terminal state
* denotes ICE
Observed terminal states
C
True progenitor states
Potency
Fraction
Inferred progenitor states
Terminal states
Progenitor States
D
Node assignment accuracy
Count
Accuracy
E
True state
iP2 [T1,T2,T3,T4]
Progenitor node
Inferred state
iP8 [T1,T2]
Not sampled
Not sampled
F
Inferred Commitment Events (ICEs)
iP9 ICE
iP5 ICE
G
Commitment times
H
Fixed sampling
Proportional sampling
ρ = 0.993
ρ = 0.470
ρ = 0.994
ρ = 0.277
Inferred commitment time (days)
True commitment time (days)
Sampling fraction

Figure 2.5. Obtaining progenitor state commitment times from phylogenies of sampled cells. (**A**) The time-scaled phylogenetic tree in Figure 4A where internal nodes are colored according to their inferred progenitor state. (**B**) A zoomed-in section of the tree in panel A where each internal node is labeled by its inferred progenitor state or terminal state. Asterisks signify ICE nodes. (**C**) Heatmap showing the agreement between inferred states and true states of all internal nodes in the tree shown in panel A. (**D**) Histogram showing the average accuracy of progenitor state assignment to internal nodes of the phylogenetic tree across the panel of 2,500 phylogenetic trees. On average, 81% of nodes were correctly assigned to a progenitor state, the rest were assigned to states with less potency. (**E**) Schematic showing how undersampling can lead to a node being assigned a progenitor state with less potency than its true state. (**F**) The tree in panel A where nodes corresponding to Inferred Commitment Events (ICEs) for inferred progenitor states iP9 and iP5 are shown in brown and pink, respectively. (**G**) Distribution of ICE times for all inferred progenitor states in the tree from panel A, representing each progenitor state's commitment time. The ICE times for *iP5* and *iP9* are boxed in red. (**H**) Scatterplots showing the correlation between actual commitment time of each progenitor state to the value inferred from the phylogenetic tree across all 2,500 simulated phylogenies broken down by experimental sampling strategy. Dots are colored based on progenitor sampling fraction and according to the key on the right. Trendlines (LOESS) for adequately-sampled and undersampled progenitor states are shown in blue and red, respectively. ρ indicates Spearman's correlation coefficient. The blue value corresponds to progenitors with sampling fraction better than 25%; the red value to those with sampling fraction equal to or less than 25%.

We next assessed whether population size and commitment bias of a progenitor state can be obtained from time-scaled phylogenies. We define progenitor population size as the number of cells in a progenitor population immediately before they commit to their downstream fates and commitment bias as the ratio of these cells that commit to each downstream fate. To estimate

population size for a progenitor state, we identified the subset of all branches in the phylogenetic tree that, first, are present at the progenitor state's inferred commitment time and, second, connect nodes assigned as either the progenitor state itself or any of its upstream or downstream states (**Figure 2.6A,B,** see Methods). These branches represent cells of the progenitor state that are present at its time of commitment. We then counted the number of incoming nodes to these branches as the population size (**Figure 2.6C**). For commitment bias, we calculated the ratio of branches that end in each of the downstream fates irrespective of their parental state (**Figure 2.6D,** see Methods). Applying this algorithm to our panel of time-scaled phylogenies, we found that the ability to estimate population size and commitment bias for a progenitor state depend heavily on its sampling fraction as well as the sampling method. Population size estimates of adequately-sampled progenitor states agree well with their actual size (Spearman's ρ=0.975 for fixed and ρ=0.943 for proportional sampling) (**Figure 2.6E**). For undersampled progenitor states, on the other hand, population size estimates are capped at the number of their sampled terminal progeny, thus performing poorly for the proportional sampling scheme (Spearman's ρ=0.594) and being completely non-informative for the fixed sampling scheme (Spearman's ρ=-0.06) (**Figure 2.6E**). Commitment bias estimates showed a different behavior: proportional sampling allowed for estimation with a minor effect from progenitor's sampling fraction (Spearman's ρ=0.929 and ρ=0.718 for adequately and undersampled progenitor states, respectively) (**Figure 2.6F**). Fixed sampling strategy, on the other hand, produced reasonable estimates only for adequately-sampled progenitor states' commitment biases (Spearman's ρ=0.815) and was uninformative for undersampled progenitor states (Spearman's ρ=0.050) (**Figure 2.6F**). These observations demonstrate the importance of sampling fraction for accurate characterization of progenitor states. They further indicate more effective estimation of population size and commitment bias with proportional sampling schemes, ostensibly due to the correlation between the size of terminal populations and that of their progenitor. Together, these results establish a strategy for estimating progenitor population

size and commitment bias from time-scaled phylogenetic trees and define critical parameters for robust estimation.

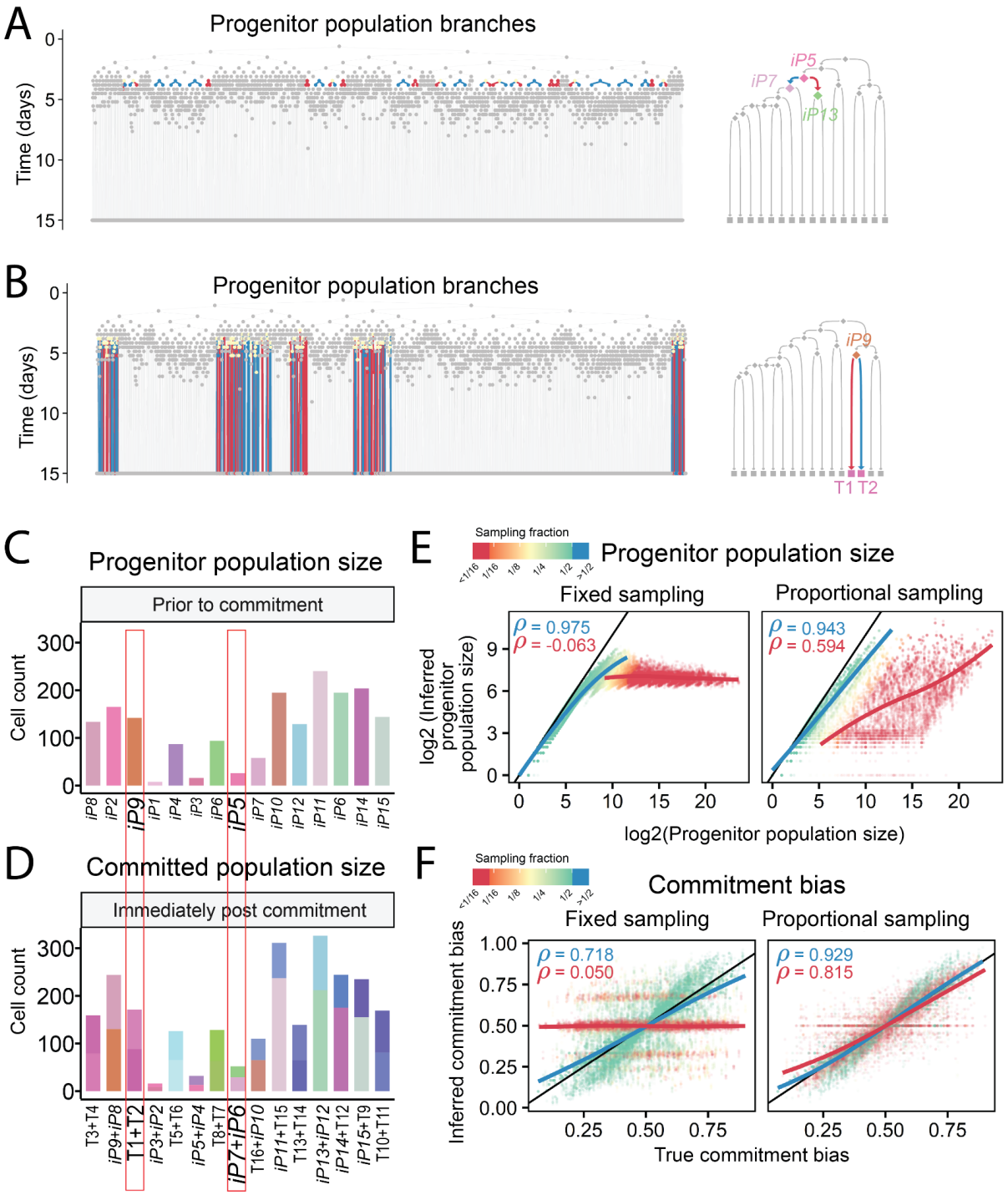

Figure 2.6. Obtaining progenitor population size and commitment bias from phylogeny of sampled cells. (**A**) The time-scaled phylogenetic tree from Figure 4A is shown and branches corresponding to inferred progenitor state *iP5* that are present at its commitment time are colored based on the state they commit to according to the key shown on the fate map to the right. (**B**) Same as A, but for *iP9*. (**C**) Barplots showing the inferred population size of each progenitor state in A based on the nodes before its branches at its commitment time. (**D**) Barplots showing the inferred post-commitment population size of each progenitor state in A immediately after its commitment time based on the nodes after its branches at its commitment time and stacked according to the downstream state they lead to. (**E**) Scatterplots showing the correlation between true population size of each progenitor state to the value inferred from the phylogenetic tree across all 2,500 simulated phylogenies broken down by experimental sampling strategy. Dots are colored based on progenitor sampling fraction and according to the key on the right. Trendlines (LOESS) for adequately-sampled and undersampled progenitor states are shown in blue and red, respectively. ρ indicates Spearman's correlation coefficient. (**F**) Scatterplots showing the correlation between actual commitment bias of each progenitor state to the value inferred from the phylogenetic tree across all 2,500 simulated phylogenies broken down by experimental sampling strategy. Dots are colored based on progenitor sampling fraction and according to the key on the right. Trendlines (LOESS) for adequately-sampled and undersampled progenitor states are shown in blue and red, respectively. ρ indicates Spearman's correlation coefficient.

Robustness of phylogeny-based quantitative fate map estimates

So far, we have established the central role of progenitor state sampling fraction as an indicator of the ability to infer quantitative fate map parameters. However, this parameter cannot be directly observed from cell phylogeny or other data collected from sampled terminal cells

(**Figure 2.3C**). Hence, it is important to find a proxy for the true sampling fraction of each progenitor state that can be derived from sampled cells alone. Here, we introduce one such proxy: estimated progenitor state coverage (PScov), which is defined as the terminal sample size of a progenitor state divided by its estimated progenitor population size. By terminal sample size, we refer to the sum of the number of sampled cells of all the terminal states that the progenitor state is capable of. Intuitively, this statistic indicates how many terminal descendents are being sampled per each progenitor cell. We found that a high PScov is indeed predictive of high sampling fraction, auROC = 0.973 (Confidence Interval: 0.972 – 0.974) (**Figure S2.2A**). For example, the majority of states that are sampled more than 25% also have a PScov larger than 2.5 (**Figure S2.2B**). Therefore, this estimated progenitor state coverage makes it possible to assess the robustness of quantitative fate map parameters for each progenitor state based solely on the phylogeny of terminal cells.

Modeling and simulating lineage barcoding in development

The results above indicate that time-scaled phylogenies of sampled cells can identify progenitor states of common fate and their dynamics. The time-scaled phylogenies that were used thus far were known—representing the exact sequence and timing of events as simulated. In actual experiments, phylogeny must be inferred from lineage barcodes. However, such inferred phylogenetic trees are inherently error-prone for two reasons. First, exhaustive phylogenetic tree search to guarantee optimality is not computationally practical for hundreds of terminal cells, and practical heuristic algorithms do not guarantee optimality [48–50] despite the recent advances in distributed computing [51]. Second, barcoding strategies employ a limited number of mutation sites which encode a limited amount of information; how close any inferred tree—including the optimal tree—is to its true phylogeny remains uncertain. We therefore asked: can phylogenies inferred from lineage barcodes be accurate enough to recreate quantitative fate maps despite their errors? To address this question, in this and the ensuing two sections,

we will: i) describe a model to generate realistic barcoding outcomes in cells, ii) establish a scalable strategy for inferring time-scaled phylogenies from barcodes, and iii) evaluate the feasibility of quantitative fate mapping using barcode-inferred phylogenies.

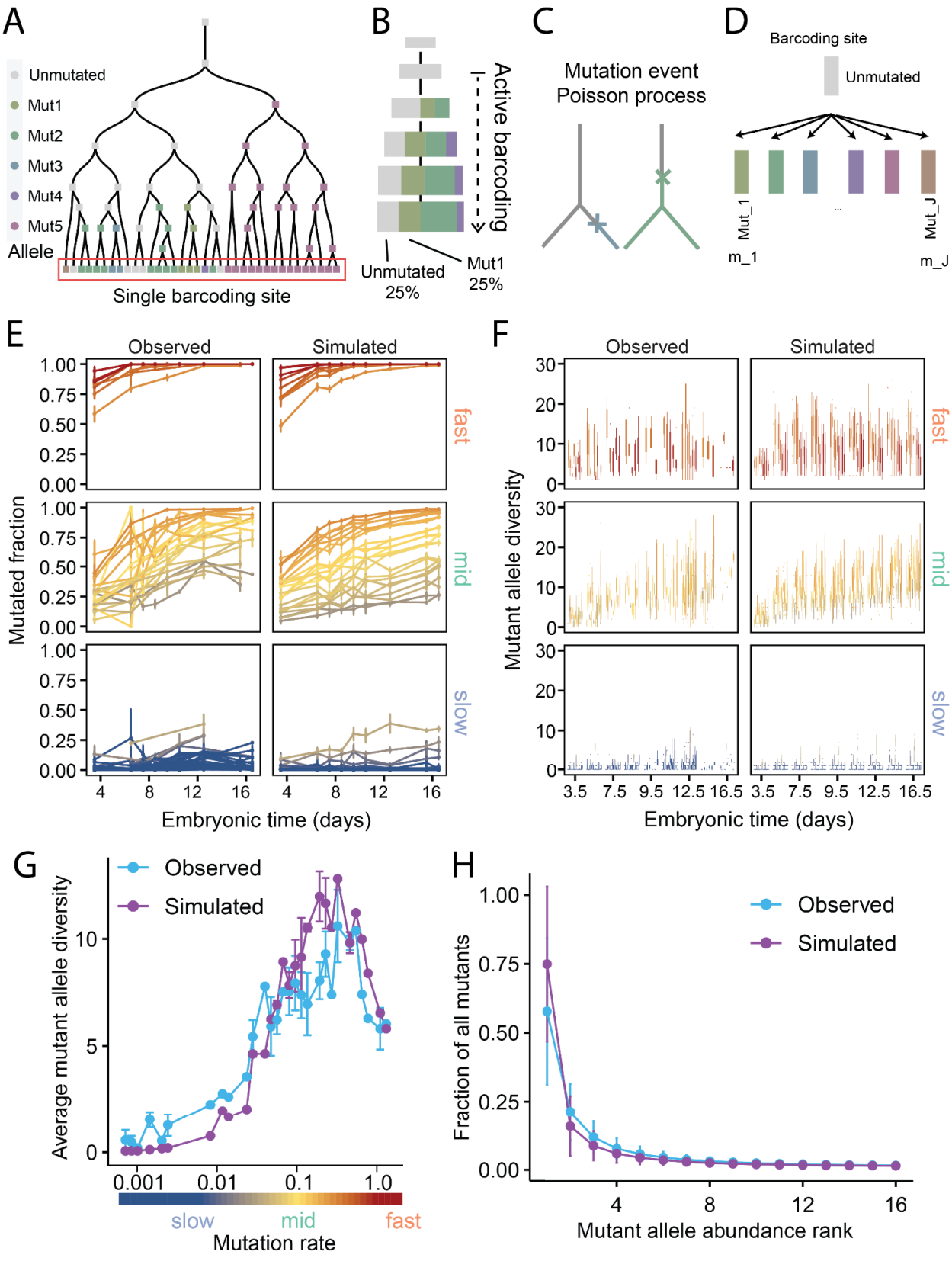
A
Unmutated
Mut1
Mut2
Mut3
Mut4
Mut5
Allele
Single barcoding site
B
Active barcoding
Unmutated
25%
Mut1
25%
C
Mutation event
Poisson process
D
Barcoding site
Unmutated
Mut_1
Mut_J
m_1
m_J
E
Observed
Simulated
Mutated fraction
fast
mid
slow
Embryonic time (days)
F
Observed
Simulated
Mutant allele diversity
fast
mid
slow
Embryonic time (days)
G
Observed
Simulated
Average mutant allele diversity
slow
mid
fast
Mutation rate
H
Observed
Simulated
Fraction of all mutants
Mutant allele abundance rank

Figure 2.7. Lineage barcoding model and agreement of simulated barcoding outcomes to those observed in MARC1 mice. (**A**) A phylogenetic tree with a single barcoding site shown as a box and its allelic state denoted with color according to the key on the right. The barcoding site is inherited from each cell by its daughters. Only the unmutated allele (gray) can change, mutated alleles (colored), once formed, are inherited from a cell by all its daughters. (**B**) Accumulation of mutations over time upon barcoding activation across a population of cells. (**C**) Mutation events (crosses) take place according to a Poisson point process. (**D**) Mutation events convert the unmutated allele to one of many possible mutated alleles ($Mut_1, \ldots, Mut_J$) (Mut_1 to Mut_j), each with an emergence probability (m_1 to m_j). (**E**) Line-plots showing the mutated fraction of MARC1 barcoding sites over time in observed (left) and simulated (right) embryos for fast, mid, and slow categories of mutation rates. Means ± SEM are shown. Each barcoding site is colored according to its mutation rate using the color scale shown in panel G. (**F**) Boxplots showing the number of mutant alleles for each MARC1 barcoding site over time in observed (left) and simulated (right) embryos for fast, mid, and slow categories of mutation rates. Each barcoding site is colored according to its mutation rate using the color scale shown in panel G. (**G**) Line-plots showing the average mutant allele diversity as a function of barcoding site mutation rate for observed (blue) and simulated (purple) embryos. Means ± SEM are shown. (**H**) Line-plots showing the prevalence of a mutant allele among all mutant alleles as a function of its rank in observed (blue) and simulated (purple) embryos. For example, a rank of one denotes the most abundant allele. Means ± SE are shown.

First, we established a general mutagenesis model. The model comprises barcoding sites that are present in each cell and inherited to its daughters (**Figure 2.7A**). Barcoding sites are unmutated in the founder cell at the beginning (i.e., $t = 0$) and once activated, start accumulating heritable mutations over time (**Figure 2.7A,B**). Each site mutates independently

from other sites in the system according to a Poisson point process with a constant rate after activation. Each mutation event converts an unmutated active copy of the site into one of many possible mutated inactive alleles, each with a distinct emergence probability and unable to mutate further (**Figure 2.7C,D** and **Methods**). Therefore, the parameters of the barcoding model are the number of sites, their mutation rates, and each site's mutant allele emergence probabilities.

This model is broadly applicable to synthetic and natural barcoding systems; here, we parameterize it based on the MARC1 (Mouse for Actively Recording Cells!) system [31] wherein extensive embryonic barcoding data are available [52]. In MARC1 mice, somatic mutations are induced in tens of independent homing guide RNA loci (hgRNAs) [53] starting at the 2-cell stage. We estimated the mutation rates of MARC1 hgRNAs (i.e., λ, rate of the Poisson process) using embryonic time course data [52] (**Figure S2.3A,B**, see Methods). We estimated emergence probabilities of mutant alleles for each hgRNA by adapting the inDelphi algorithm that predicts CRISPR-Cas9 mutations [54]. We compared and verified inDelphi's predictions against published MARC1 data [52] (**Figure S2.3B, Supplementary Data 2.2**, see Methods). To test this MARC1 barcoding model, we simulated barcoding in whole-mouse embryos for E3.5 to E16.5 in samples of 2,000 cells (or fewer when there were fewer than 2,000 cells in the organism) and compared the results to that of experiments (see **Methods**). Overall, we observed broad agreement between experimental and simulated barcoding results (**Figure 2.7E–H**). First, the distribution of mutated fractions over the course of embryogenesis agrees between simulated and experimental results for hgRNAs with a range of mutation rates (**Figure 2.7E**). Second, the total number of distinct mutant alleles (i.e. the mutant allele diversity) during embryogenesis were consistent between experiments and simulations (**Figure 2.7F**). Third, in both systems, as the hgRNA mutation rates increase, the diversity of mutated hgRNA alleles increase because there are more mutagenesis events; however, at the fastest mutation rates,

barcoding sites reach 100% mutated when there are fewer total cells, and the total diversities drop as a result (**Figure 2.7G**). Fourth, the composition of mutant alleles within embryos agree: for both simulated and observed embryos, after ranking all mutant alleles based on their frequencies, alleles of similar rank account for similar percentages of the mutated cells (**Figure 2.7H**). Taken together, these results suggest that our simulation is comparable to actual lineage barcoding experiments and produces realistic barcoding results.

Finally, based on the above stochastic barcoding model and the MARC1 system's parameters, we simulated mutagenesis in our panel of 2,500 phylogenies assuming 50 hgRNA sites per cell (see Methods). The results are 2,500 simulated barcoding experiments where, similar to real-life experiments, the barcode and terminal state information is known for the sampled panel of single cells (**Figure 2.8A**; **Supplementary data 2.4**).

Reconstructing time-scaled phylogenies from single-cell lineage barcodes

To infer quantitative fate maps in each simulated barcoding experiment, single-cell barcodes must be converted to a time-scaled phylogenetic tree. However, many current methods for phylogenetic reconstruction based on lineage barcodes lack a mutagenesis model specific to lineage barcoding. As a result, their inferred phylogram branch lengths are often in arbitrary units. Those strategies that do involve a barcoding mutagenesis model [55] depend on optimization techniques that are not scalable to thousands of sampled cells as we have simulated in each experiment. To address this gap, we developed a method to infer phylogenies with branch lengths measured in actual time units that readily scales to thousands of cells. We first compute a maximum likelihood estimate of the time that separates a pair of cells from their most recent common ancestor (time to MRCA) for all pairs of terminal cells (**Figure 2.8B**, see Methods). We apply UPGMA hierarchical clustering [56] to the pairwise temporal distance matrix to obtain a time phylogenetic tree (**Figure 2.8C**). We call this approach, which scales in polynomial time, PHYlogeny reconstruction using Likelihood Of TIME (Phylotime).

To evaluate Phylotime's performance in reconstructing time-scaled phylogenies, we first compared MRCA times estimated in our simulated barcoding experiments to that derived from the corresponding true trees (**Figure 2.8D**) and found that the two are highly correlated (Pearson's R = 0.98). We then applied Phylotime to all 2,500 simulated barcoding experiments to obtain inferred-phylogenetic trees (**Figure 2.8C**; **Supplementary Data 2.4** for all trees) and compared the topology and branch lengths of the Phylotime-reconstructed trees to the true trees, using KC0 distance for topology and KC1 distance for combined topology and branch length (**Figure 2.8E,F**). KC1 is the Kendall-Colijin metric weighted by branch length by setting its parameter to one [47]. A KC1 distance of zero between two trees indicates identical topology and branch lengths and distances larger than zero indicate increasing differences in branch length and topology. The results show that Phylotime's solutions converge to the true phylogeny with an increasing number of barcoding sites(**Figure 2.8E,F**) and produce time-scaled phylogenies that are by far the most accurate according to KC1 error (**Figure 2.8F**). Only a Hamming distance based method and Cassiopeia [57], a heuristic approach based on maximum parsimony, were compared because other common methods do not scale to this number of terminal cells and barcoding sites. Expectedly, given their scale, none of the trees were perfectly reconstructed compared to the truth, recapitulating errors that are inherent to tree inference. Despite this, Phylotime's results provide a test panel of inferred time-scaled phylogenetic trees, with generally accurate topology and branch length for fate map reconstruction.

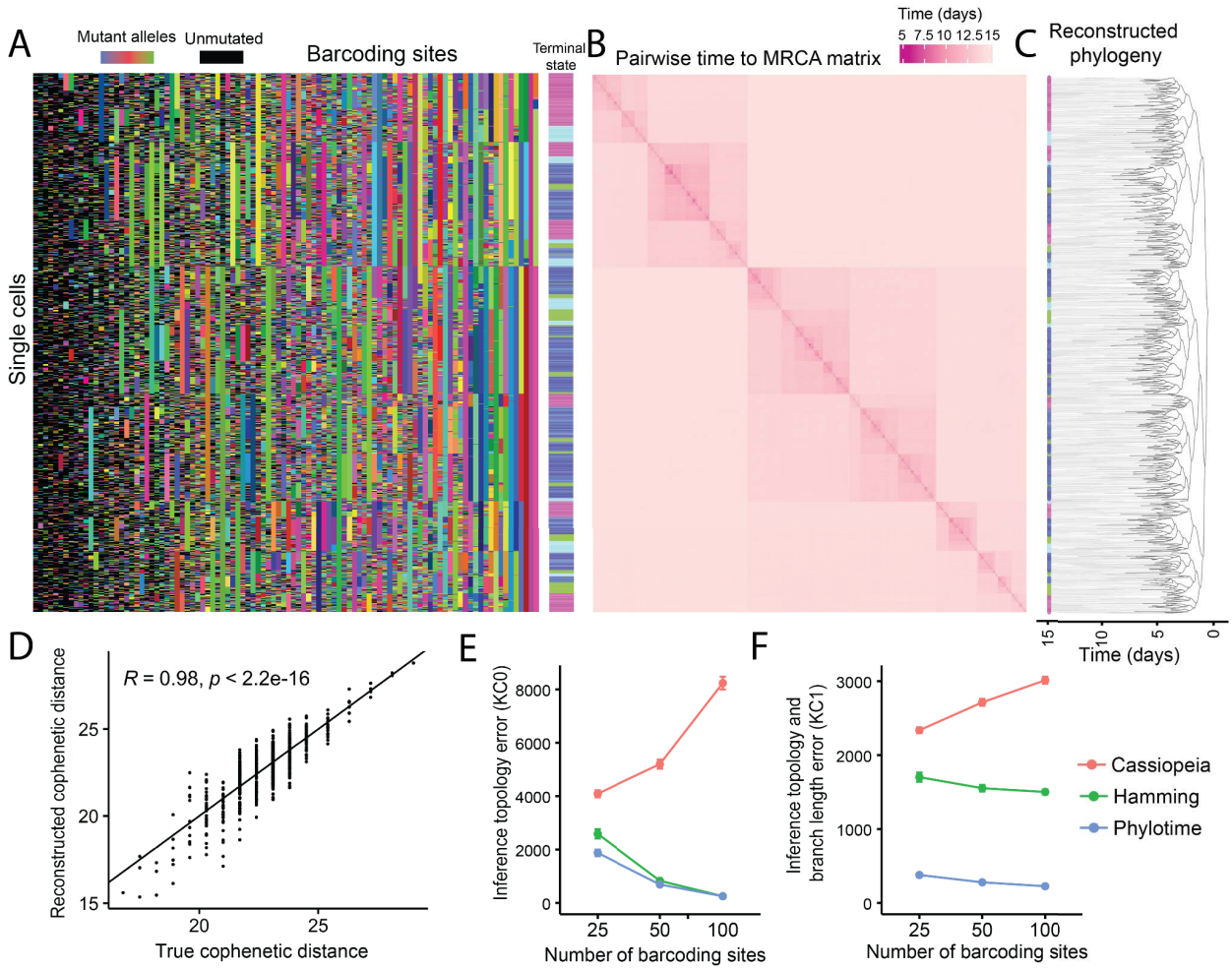

Figure 2.8. Accurate reconstruction of time-scaled phylogenetic trees using Phylotime. (**A**) The output of a simulated barcoding experiment shown as a barcode matrix with each hgRNA barcoding siteas a column and each cell as a row. Colors correspond to mutant alleles; black is an unmutated allele. The color bar on the right shows the state of each terminal cell. (**B**) Heatmap showing the pairwise time to most recent common ancestor (MRCA) for all cells in A. (**C**) Time-scaled phylogenetic tree reconstructed by applying a clustering algorithm to the matrix in B. Colors on the terminal branches signify the observed state of the cell. (**D**) Scatter plot showing the correlation between the Phylotime-inferred and true cophenetic distances between all pairs of cells in A. Trendline is shown. (**E**) Error of phylogenetic reconstruction using Phylotime, Hamming distance with UPGMA, and Cassiopeia, with 25, 50, or 100 barcoding sites

when considering only tree topology (KC0 distance) across the panel of 2,500 simulated barcoding experiments. Means ± SEM are shown (N=2,500). (**F**) Error of phylogenetic reconstruction using Phylotime, Hamming distance with UPGMA, and Cassiopeia, using 25, 50, or 100 barcoding sites when considering both tree topology and branch length. Cassiopeia phylogenies were scaled to the same total time as the reference trees. Means ± SEM are shown (N=2,500).

Quantitative fate map inference based on lineage barcodes

We next assessed if time-scaled phylogenetic trees inferred from lineage barcodes can faithfully reproduce quantitative fate maps despite their inherent uncertainties. We applied the ICE-FASE algorithm to all 2,500 time-scaled phylogenetic trees that were inferred using Phylotime from simulated barcoding experiments to derive quantitative fate maps. We then compared various parameters of these fate maps against those obtained by applying ICE-FASE to the true phylogeny. We found that inferred phylogenies perform almost as well as true phylogenies at estimating fate map topology, regardless of map complexity and imbalance, or the sampling strategy (**Figure 2.9A**). Similarly, for commitment times, population sizes, and commitment biases of progenitor states that were properly identified in the reconstructed topologies, we found that Phylotime-inferred phylogenies perform similar to true phylogenies with proportional sampling in most conditions (**Figure 2.9B–D**), though a drop off was observed in population size and commitment time detection with fixed sampling (**Figure 2.9C–D**). Taken together, these results show that the ICE-FASE algorithm can faithfully reconstruct quantitative fate maps and infer parameters of the progenitor states using barcode-inferred phylogenies. This finding indicates that quantitative fate mapping is feasible despite errors inherent to phylogenetic reconstruction. More broadly, these results show that quantitative fate mapping may be accomplished with current lineage barcoding technologies.

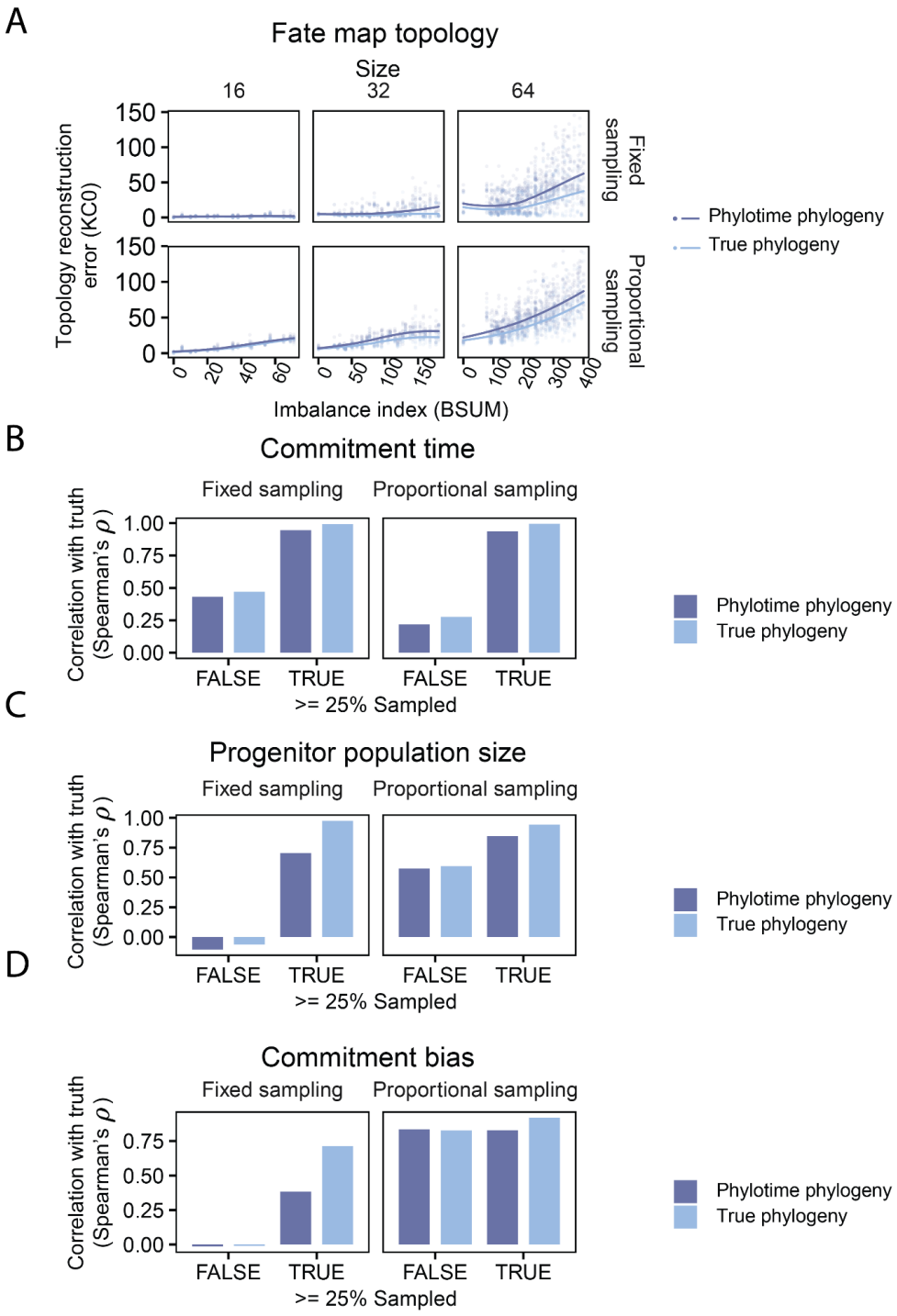

Figure 2.9. Successful quantitative fate mapping using barcode-reconstructed time-scaled phylogenetic trees. (**A**) Scatter plots showing fate map topology reconstruction error as a function of its imbalance in all the fate maps described in Figure 1 broken down by fate map

size and experimental sampling strategy. Trendlines (LOESS) are shown. (**B**) Barplots showing Spearman's correlation between true commitment time and inferred commitment time of all progenitor states in 2,500 simulated experiments for progenitor states that were adequately sampled (TRUE) and undersampled (FALSE), broken down by fate map size and experimental sampling strategy. Results using true simulated phylogenies (light blue) or Phylotime-inferred phylogenies (dark blue) are shown. (**C**) Barplots showing Spearman's correlation between true population size and inferred population size of all progenitor states in 2,500 simulated experiments for progenitor states that were adequately sampled (TRUE) and undersampled (FALSE), broken down by fate map size and experimental sampling strategy. Results using true simulated phylogenies (light blue) or Phylotime-inferred phylogenies (dark blue) are shown. (**D**) Barplots showing Spearman's the correlation between true commitment bias and inferred commitment bias of all progenitor states in 2,500 simulated experiments for progenitor states that were adequately sampled (TRUE) and undersampled (FALSE), broken down by fate map size and experimental sampling strategy. Results using true simulated phylogenies (light blue) or Phylotime-inferred phylogenies (dark blue) are shown.

In vitro validation of quantitative fate map inference using lineage barcodes

The above results prove in principle that lineage barcodes from a subset of terminal single cells can be used to retrospectively reconstruct quantitative fate maps of their progenitors. These quantitative fate maps not only define progenitor states based on their fate but also provide information about the times they were present, their numbers, and their commitment biases. This quantitative fate map reconstruction strategy constitutes a new method in the developmental biology toolbox to assess developmental systems, such as mammals, wherein multipotent progenitor populations dynamically commit to downstream fates. However, in silico models inevitably make simplifying assumptions. Our models have assumed that cells

proliferate at fixed intervals with constant rate and without death, or that all the lineage barcodes can be measured without error. Therefore, we sought to validate our quantitative fate mapping strategy on an in vitro experimental model. However, there is no known biological system where these quantitative parameters for a set of progenitor states are established, making it difficult to experimentally test this novel approach. To address this gap, we created an experimental system in which quantitative fate map parameters can be controlled and then interrogated using lineage barcodes. We established a clonal human induced pluripotent stem cell (iPSC) line with 32 hgRNA barcoding sites distributed in its genome as a non-tandem array (**Supplementary data 2.1**). The line also includes doxycycline inducible Cas9 expression for barcoding activation (see Methods). 24 of the 32 hgRNAs were active and accumulated random mutations upon doxycycline induction (**Figure S2.4**).

Using this barcoding cell line, we designed a growing and splitting scheme in culture that mimics the cell fate commitment hierarchies that we have so far only simulated in silico (**Figure 2.10A**). In two experiments, starting from single cells, we initiated barcoding and passaged the growing cell population into an increasing number of branches at known times, numbers, and split ratios (**Figure 2.10B, Figures S2.5,S2.6**, see **Methods**). The experiments were similar except that in one experiment (E1), progenitor state #3 (P3) was split two days before progenitor state 4 (P4), whereas in the other (E2), P4 was split two days before P3. In effect, these experiments represent quantitative fate maps with the last set of cultured populations as the terminal cells and their prior ancestors as the progenitor populations. Finally, we sequenced barcodes from 192 single cells in each terminal population (see **Methods**). After data processing and filtering out low quality cells, we obtained on average 158 cells per terminal population (**Figure S2.7A,B**). E1 and E2 had medians of 27 and 26 hgRNAs detected per cell, respectively (**Figure 2.10C**). We imputed the alleles of undetected hgRNAs using the machine learning software “xgboost” [58] (see **Methods**) As a reference, we conducted parallel simulations on the two fate

maps that represent the ground truths, with cell division rates derived from the progenitor population sizes at each split. The lineage barcodes are simulated with hgRNA mutation rates and allele emergence probabilities obtained from time-course measurements and inDelphi predictions, respectively. We then applied Phylotime and ICE-FASE to both simulated and experimental data. The experimental data reconstructed the topology correctly in both E1 and E2 (**Figure 2.10D**), and so we refer to the inferred states by their true state names hereafter. The PScov ranged from 1.68 to 2.36 in all non-founder (P5) progenitor states, indicating that the progenitor states were generally undersampled in these experiments. Nevertheless, in addition to the correct topology, the inferred fate maps recovered the correct orders of fate commitment in both experiments (**Figure 2.10D**). Moreover, the inferred maps successfully recovered the difference between E1 and E2, which was the relative order of commitment for P3 and P4 progenitor states (**Figure 2.10D**). This result suggests that our strategy can identify specific quantitative fate map differences in different systems. However, due to undersampling, we did not expect to recover the exact times of commitment and progenitor population size. Nevertheless, we wanted to evaluate if these estimates would approach the truth with increased sampling. We thus subsampled the experimental data to lower numbers of cells per terminal state, repeating the process 50 times at varying sample sizes. In parallel, we carried out simulations with the same sample sizes. We then classified inferred fate maps based on their topology and correctness of relative ordering (**Figure 2.10E**) and found that the fraction of correct topologies and relative order of commitment increased in a similar fashion as more cells were sampled in both simulated and experimental sets. Additionally, commitment times and population sizes of P3 and P4 approached the actual amount with increasing sampling in simulated and experimental sets alike (**Figure 2.10F,G**). Together, these observations validate our barcoding models used for simulation and indicate that our quantitative fate mapping strategy, ICE-FASE, and Phylotime are robust to the simplifying assumptions made in their

models. These results also suggest that developmental differences, ostensibly caused by genetic or environmental factors, can be detected using quantitative fate mapping.

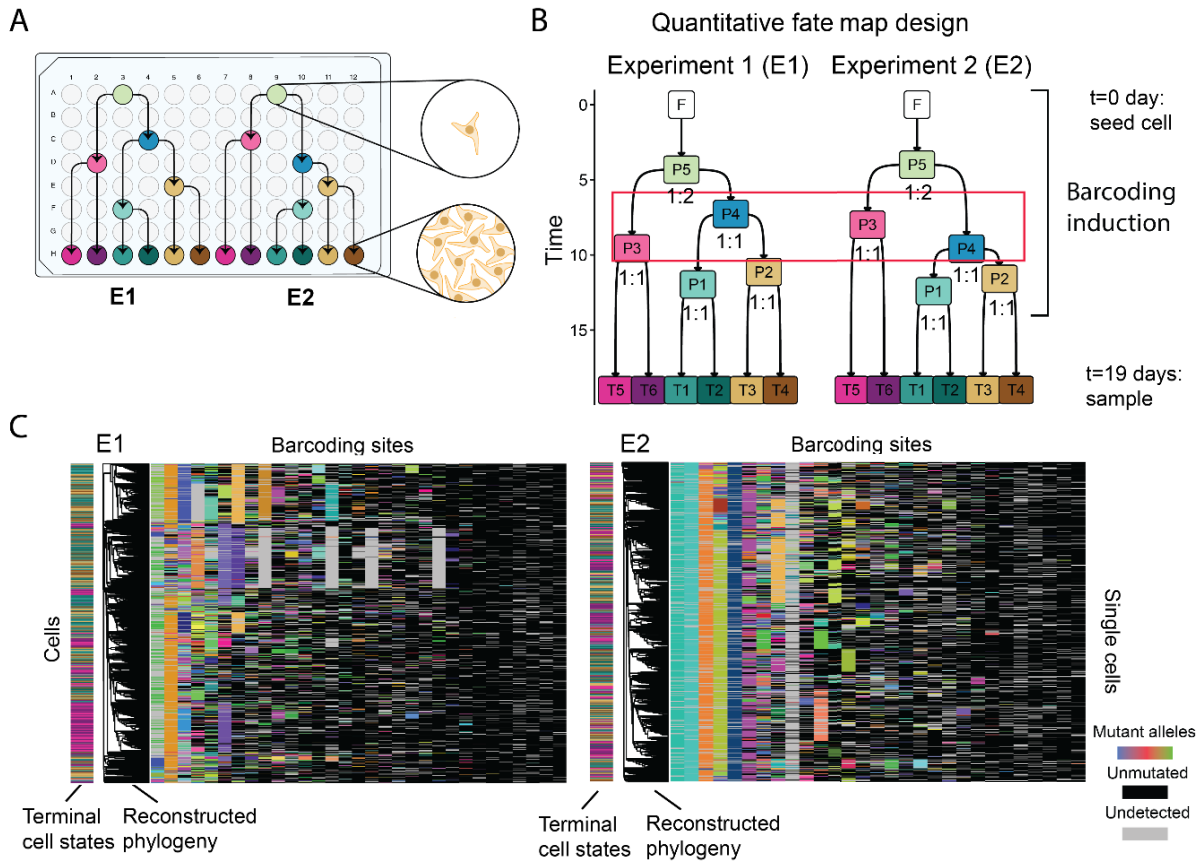
A
E1
E2
B
Quantitative fate map design
Experiment 1 (E1)
Experiment 2 (E2)
Time
t=0 day: seed cell
Barcoding induction
t=19 days: sample
C
E1
Barcoding sites
E2
Barcoding sites
Cells
Single cells
Mutant alleles
Unmutated
Undetected
Terminal cell states
Reconstructed phylogeny
D
Inferred fate map topology and commitment times
Time (days)
E
# of replicates
Sample size per teriminal state
F
Commitment time
P3 (E1)
P3 (E2)
P4 (E1)
P4 (E2)
Time (days)
Ground truth
Experiment
Simulation
Sample size per terminal state
G
Progenitor population size
Truth: 530
Truth: 380
Truth: 1625
Population size
P4 early
P4 early (P1-P2 reversed)
P3 early
P3 early (P1-P2 reversed)
Wrong ordering
Wrong topology

Figure 2.10. Validation of quantitative fate mapping strategy using an in vitro system. (**A**) Implementation of the split-passage scheme in cultured iPSCs. (**B**) The two quantitative fate maps that were implemented using the passage-split scheme in A. The red box highlights the differences between progenitor state order. (**C**) Heatmap of lineage barcodes from hundreds of sequenced single cells in each experiment are shown (E1 shown on the left, E2 on the right) and used to reconstruct a time-scaled phylogeny which is shown to the left of the map. The state of each cell is marked on the bar to the left of the phylogram and the color code is according to panel B. (**D**) Reconstructed quantitative fate maps from E1 and E2 single-cell barcode results. The red box highlights the differences between progenitor state ordering. (**E**) Barplot showing the fraction of correctly ordered topologies among a number of subsampled replicates in both simulation and experiment at a number of different sample sizes. **(F)** Boxplot for comparing commitment time estimates and true commitment times for progenitor states 3 and 4 at different sample sizes in simulation and experiment. The yellow lines indicate the ground truth. **(G)** The same comparison for progenitor population size. The yellow lines indicate the ground truth.

Discussion

In this study, we set out to determine how cell phylogeny—as estimated by lineage barcoding approaches—can be used to understand the dynamics of progenitor states that drive the development of animals. This is a timely problem as advances in genome engineering and sequencing technologies make it possible to estimate cell phylogeny with high throughputs using synthetically-induced or naturally-occurring somatic mutations. However, phylogeny of terminal cells is inherently stochastic and has a complex relationship with the fate of progenitor states. Moreover, retrospective strategies can only sample a fraction of the cells in an organism and are subject to inherent and complex errors in phylogenetic reconstruction.

To address these challenges, we established a robust and practical approach for phylogeny-based analysis of cell fate called quantitative fate mapping. This approach uses terminal cells' lineage barcode and type information to quantitatively characterize the progenitor field—the collection of progenitor states [42]—that gave rise to them. Our approach first estimates a time-scaled phylogenetic tree based on single-cell barcodes using the Phylotime algorithm. It then combines the phylogenetic tree with the terminal fate of the sampled cells to identify nodes associated with fate decisions and uses their timing to reconstruct an initial map of progenitor states. Finally, it infers the progenitor state of all internal nodes of the phylogenetic tree and uses these assignments to estimate the commitment time, population size, and commitment bias of progenitor states (ICE-FASE algorithm). While other strategies have been described to infer the lineage relationship of terminal cells in specific biological systems, our quantitative fate mapping approach is unique in that it evaluates the dynamics of progenitor states, can be applied to any retrospective strategy, and scales to large and complex fate and lineage landscapes. Other strategies to infer progenitor dynamics from cell lineage have focused on reversible cell state transitions or single progenitor states [59–61]. Moreover, we have demonstrated that this strategy can tolerate the errors that are inherent to phylogenetic reconstruction and used in silico experiments based on realistic barcoding parameters to validate its performance. Importantly, we have used experiments with cultured stem cells to show that our strategy is robust to our simplifying assumptions.

A key finding of our quantitative fate mapping strategy is that time-scaled phylogenies—phylogenetic trees wherein branch length corresponds to actual time—can be used as chronometers of progenitor population dynamics. The few currently available lineage reconstruction strategies that can create time-scaled phylogenetic trees require NP-hard searches in the tree space that become computationally untenable for adequately large sampling depths. To reconstruct time-scaled phylogeny at scale, we developed the Phylotime

algorithm which obtains the time since the most recent common ancestor of each pair of cells with maximum likelihood estimation and uses the estimates to reconstruct time-scaled phylogenies. Phylotime performs in polynomial time making it readily applicable to large trees with thousands of terminal branches. Phylotime can be readily adopted for other systems where multiple barcoding sites create recurring mutant alleles in parallel..

This work also establishes a foundation for obtaining meaningful estimates of fate dynamics in a single animal from phylogeny of a small number of cells. Unlike the phylogenetic tree itself, these estimates can be directly combined to create species-level insights or make comparisons between different species. Critically, our results show that only when a progenitor population's progeny is adequately sampled can its existence, potency, and quantitative parameters be meaningfully estimated from a single time-scaled phylogeny; estimates in severely undersampled progenitor states are not meaningful and will not be improved by combining multiple animals. To assess if estimates are robust, we defined progenitor population sampling fraction as the fraction of the actual progenitor population whose descendants are observed among terminal cells. We observed that statistically meaningful characterization typically requires sampling fractions larger than 25%. To meet this criterion, the number of terminal descendants that are sequenced should have the same order of magnitude as the progenitor population size (PScov larger than 1). We propose this as a fundamental rule for retrospective lineage analysis approaches. Accordingly, the growing capacity of single-cell sequencing technologies bodes well for quantitative fate mapping since the average progenitor state is less than 1,000 cells during and prior to organogenesis [44,62]. For larger progenitor populations the sampling barrier may be overcome by strategically bottlenecking the number of sampled progenitors. For example, in the neocortex where progenitors develop into known and limited anatomical positions, bottlenecking can be accomplished by sampling a limited number of cortical columns. Alternatively, prospective lineage tracing approaches may be employed to

label a subset of a progenitor population, for instance with fluorescence, and only sample their terminal progeny [61].

Sampling strategy is also a critical parameter in experimental design. Our results show that analyzing a fixed number of cells from each terminal state can be more beneficial for characterizing the topology of the fate map especially with more unbalanced topologies, and, to a lesser extent, determining the commitment time of progenitor states. Sampling in accordance with the abundance of each terminal state can be beneficial for determining commitment bias and population size of progenitor states.

Quantitative fate maps obtained in this fashion represent a fate landscape that can capture emergent features of the developmental system that gave rise to the sampled cells. They complement the state manifolds obtained using direct molecular analysis of progenitor populations (e.g., single-cell RNA sequencing) [62–64] in multiple ways. Firstly, they provide information about the long-term fate of progenitors. Secondly, they can capture variation between embryos of the same species. Finally, they enable analyzing progenitor populations with respect to a specific subset of their progeny which may be of particular interest, for example due to relevance to a specific genetic or environmental signal.

The limitations of this study are born out of the assumptions of its models. For instance, progenitor states examined here differentiate into only two downstream states. While this assumption does not alter the general conclusions of our study, our strategy can be modified to accommodate other scenarios when necessary. A trifurcation of a progenitor state into three downstream states, for example, can be resolved as two progenitor states with no distance between them on the fate map. Additionally, our models assume parameters of cell division and barcoding mutagenesis tailored to mouse development. These parameters can be adjusted to analyze other organisms, other stages of development, and other developmental systems such as organoids. Importantly, we have not made any assumptions about the mechanisms of

differentiation and commitment (e.g., asymmetric versus symmetric cell divisions). As such, the models are agnostic to these parameters and quantitative fate mapping performance may even be enhanced should information about these mechanisms be incorporated. Another limitation of the current model is that it does not consider cell death in progenitor or terminal populations. However, cell death rates would be indirectly reflected in fate map parameters such as population size. Moreover, given some priors about its rate, cell death rates can be incorporated into the model.

In summary, this work provides a framework for high-throughput quantitative fate mapping that can be used both with synthetic lineage barcoding technologies and with naturally occurring somatic mutations. These quantitative fate maps describe progenitor populations that gave rise to sampled cells and characterize their fate dynamics. Robust fate mapping relies on adequate representation of each progenitor population among sampled cells as well as the ability to infer phylogenetic trees with branch lengths that represent time. This work expands the scope of barcoding approaches beyond lineage tracing to quantitative fate mapping which can allow for the characterization of genetic and environmental effects during development.

Methods

Definition of quantitative fate map (QFM)

The quantitative fate map describes cell division and fate commitment dynamics at the cell population level. At t = 0, there is one cell of the root (most potent) state. In between fate commitments, at doubling times specific to each progenitor state, cells double to give rise to the next set of cells at the next time point. At each time point when there is a fate commitment event, cells transition to one of the two downstream states that are less potent. They do so by first doubling, and have the resulting daughter cells randomly assigned to one of the less potent states according to a commitment bias of p. Suppose N cells commit to downstream states PX

and PY. At the commitment time, the N cells of the more potent type double once again and $[2Np]$ cells become type PX and $2N - [2Np]$ cells become type PY. Cells of types PX and PY continue to divide with their respective doubling times until another fate commitment event occurs or the target time of sample collection has been reached.

In summary, parameters of the quantitative fate map include: topology of fate commitment and the following parameters for each progenitor state: doubling time, commitment time and commitment bias. Progenitor population size, defined as the number of progenitor cells at a fate commitment event can be derived from the fate map topology and doubling time. In this work, our interest is in inferring fate map topology, along with commitment time, commitment bias and progenitor population size for each progenitor state.

Definition of time-scaled phylogeny

A time-scaled phylogeny is defined as a rooted, ultrametric phylogenetic tree where branch lengths are in the unit of time. Nodes in the time-scaled phylogeny represent cell division. The root node represents the most recent common ancestor (MRCA) of all terminal cells. The length of the root edge is the time until the cell division of the root MRCA. Cophenetic distance is defined for each pair of terminal cells, which is the distance between the cells along the phylogenetic tree. The depth of a node in the phylogenetic tree is defined as the distance of a node to the root plus the length of the root edge. The ultrametric property requires that all tips are equidistant from the root, that is, have the same depth. The total time of a time-scaled phylogeny is defined as the depth of its tips.

Model of sampled cell phylogenies based on quantitative fate maps

To generate a phylogeny of a set of sampled cells based on the quantitative fate map, the following steps were employed. First, the number of sampled progenitor cells at each time point were drawn by propagating backward in time, from the terminal cells to the single zygote. At

each step, from some time point j to the next time point $j + 1$, suppose n cells have divided into $2n$ cells, and S_{j+1} cells were sampled from the time point $j + 1$. C_j is defined as the number of merges among the sampled progenitor cells at time point $j + 1$, then it can be shown that C_j follows a hypergeometric distribution with the following probability mass function (pmf):

$$p(C_j = z \mid S_{j+1} = k) = (\binom{n}{z}\binom{n-z}{k-2z}2^{(k-2z)})/\binom{2n}{k}.$$

To get the number of sampled progenitor cells in generation j, we compute $S_j = S_{j+1} - C_j$.

Recursively drawing the number of sampled progenitor cells at the previous time point from the above distribution, the progenitor sample sizes at each time point could be generated. At each time point when one cell type commits to two different downstream types, the sample sizes of the first time point of the downstream types were summed, which becomes the sample size of the last time point of the more potent cell type. This process is repeated until the sample size of the zygote is drawn, which is always one.

Next, the exact cell phylogeny is generated based on the progenitor sample sizes from the last step. Given that there were $C_j = S_j - S_{j+1}$ merges at time point j, C_j cells are chosen at random from cells at time point $j + 1$ to be merged. Each merge gives rise to a sampled cell at time j. This process allows cells to be recursively merged at each time from terminal population to the zygote. The merges alongside their timing makes up a time-scaled phylogeny.

Realistic cell division rates during early mouse developement

Before t = 4.2 days, interdivision time was set to 0.6 days. After day 4.2 interdivision time was set to 0.35 day. The number of cells resulting from these division rates were compared to the previously reported numbers in [44]. **Figure S2.1** compares the number of cells in our simulation based on these doubling times to the estimates of the number of cells in a developing mouse.

Construction of fate map panel

To generate a series of fate map topologies with varying levels of imbalance, 10,000 random tree topologies were generated using *rtree* function from the “ape” package in R [65], and computed for each category of 16, 32 and 64 terminal states the BSUM index of each tree. Next, the trees were classified into groups with incremental values of BSUM 5. Finally, one tree from each group was randomly selected. The most balanced tree was also appended to each category. The most unbalanced tree for a fixed number of terminal states is pectinate, which is included with this procedure.

Next, the timing of commitment events (bifurcations) given fate map topologies was generated. To make commitment times comparable across the different categories of fate maps, commitment events were spaced out across a common time window (1.8 - 9.8 days), subject to the constraint of each topology. To do so, a random node order on the topology was first drawn. Specifically, a permutation of the nodes of each daughter was generated at each bifurcation, and subsequently combined [66]. The ordering ranked each commitment event from earliest to latest. Next, the intervals between consecutive commitment events were taken as the parameters, and we minimized the values of these parameters, subject to the constraints of fate map topology and allowing for at least one cell doubling in between two consecutive commitments, using linear programming optimization. The remaining duration of the interval in addition to the minimum length was then distributed into each interval according to a symmetric Dirichlet distribution with $\alpha = 5$. Finally, the commitment biases in each fate map were drawn from a Beta distribution with $\alpha = \beta = 5$.

Fate map topology reconstruction with FASE

To determine if a node is a FAte SEparation (FASE), all unique terminal types of the progeny of each node were listed. The node was classified as a FASE (for at least some pair of terminal

types) if any of its daughter nodes had observed fates that are less potent. We identified FASE nodes across all nodes in the phylogenetic tree. Next, for each FASE, all pairs of terminal states that a FASE separated were listed. Now for a pair of terminal fates, the mean depth of the FASEs that separated the terminal fates were computed, referred to as the FASE time. If no FASE existed for a pair of terminal states, the FASE time was taken to be zero. Finally, the distance between a pair of terminal states was equal to two times the difference of total time and the FASE time. To get the topology from the full distance matrix, the *upgma* function from the "phangorn" package [67] was applied, which is a wrapper of *hclust* in base R.

Node state assignment for time-scaled phylogeny

Each bifurcation in the fate map topology corresponds to the commitment event of a progenitor state. One characteristic of the progenitor state is its potency: the set of terminal states it can lead to. Each node in the time-scaled phylogeny also had an observed potency determined by the set of states its progeny covers. The progenitor state of a node of the phylogeny was assigned based on its potency: the node was assigned a state that had the same potency as itself. If no such state exists in the fate map, then it was assigned a least potent state in the fate map that was more potent than the node.

Commitment time inference with Inferred Commitment Events (ICE)

To infer commitment time of a progenitor state, a set of ICEs were identified. A node in the time-scaled phylogeny was considered an ICE if both daughters had a different assigned state than itself. Each ICE was associated with a progenitor state. The depths of all ICEs associated with a state defined the ICE times. The mean of ICE times was used as an estimate for the commitment time. In cases where inferred commitment times of the downstream state is earlier than that of the upstream state, the commitment time of the downstream state was set to that of the upstream state.

Progenitor population size and commitment ratio inference

To identify the population present at the commitment time of a progenitor state, a set of branches in the time-scaled phylogeny needed to be identified. First, a set of extended states was defined. The extended states included the state itself, its upstream states up to root and its downstream states down to the terminal states. Next, a state path was constructed for each branch that spanned the commitment time of the progenitor state: the state path included a number of connected states on the fate map topology, which started at the state of the incoming node and ended at the outgoing node of the edge. A branch was considered associated with the progenitor state if its state path overlapped with any of the extended states. To estimate the progenitor population size, the collection of incoming nodes of the associated branches were counted.

Next, to quantify the bias of a progenitor state's commitment, each associated branch was further classified into one of the three categories based on if it was committed to one or the other immediate downstream states, or if it was uncommitted. The classification was made based on the state path; if the state path covered one of the two immediate downstream states, it was classified accordingly. Otherwise, it was classified as uncommitted. To estimate the commitment bias, the ratio of associated branches committing to one side versus the other was used.

Estimation of mutagenesis parameters in MARC1 mice

To get posterior estimates of mutation rates of MARC1 hgRNAs (i.e., λ, rate of the Poisson process), a grid search was conducted to match empirical distributions of mutated fractions among simulated and observed data across a number of embryonic time points. Previously reported hgRNAs formed three classes: the 'slow' class generated mutations on the order of 0.001 mutations/day, 'mid' class generated ~0.1 mutations/days, and fast class generated ~1.0

mutations/day during early mouse development. The ‘slow’ and ‘fast’ estimates expectedly had large uncertainties as most observed fractions are close to 0 or 100 percent mutated. Alternatively, a naive estimate of mutation rate can also be used. If mutated fractions F_i were observed at time T_i in animal i for$i = 1, \ldots, N$, then $\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} -\frac{1}{T_i}\log(1 - F_i)$ is a naive estimate.

For mutant alleles of a barcoding site, estimating the probabilities of individual repair outcomes created by Cas9 DNA break-repair (mutant emergence probabilities) was challenging. Normally, the fraction of cells carrying a particular mutant allele among all cells with a mutated allele (within-animal estimates) is a good estimator of the allele emergence probabilities. However, when cells divide and mutate starting from a small field size, these fractions are largely affected by the time of the mutagenesis events, as early events result in larger clones carrying the same exact mutation. On the other hand, when hgRNA genotypes are observed for multiple animals, the fraction of animals that carry a particular genotype, once normalized, and when the probability is small, can be good estimates to the mutation probabilities (across-animal estimates) (**Figure S2.3B,C**). In this case, the estimation accuracy depends on the number of animals analyzed. From the MARC1 time course data, the within-animal estimates were calculated for each animal and averaged, and the across-animal estimates were calculated based on 173 embryos from 2 mouse lines. To get a more complete profile of possible mutant alleles and their occurrence probabilities for each hgRNA, we adapted the inDelphi machine learning algorithm to predict CRISPR-Cas9 mutation results [54] for hgRNAs. We observed that the inDelphi-predicted probabilities agreed well with the across-animal estimates from MARC1, but poorly with within-animal estimates (**Figure S2.3B**). Further, the fact that the majority of the low probability mutations were not observed in any mouse suggests that the limited number of hgRNA mutation events during mouse development does not sufficiently cover a large portion of

the mutational profiles. These conclusions were further validated by simulating multiple animal lineage barcode data based on inDelphi-predicted mutational profiles and comparing the within- and across-animal estimates from the simulated data of the true parameters (**Figure S2.3C**).

InDelphi predictions of hgRNA allele emergence probabilities

The emergence probabilities of hgRNA mutant alleles were computed by inDelphi. inDelphi is a machine learning algorithm to predict heterogeneous insertions and deletions resulting from CRISPR/Cas9 double-strand break. [54] In this study, inDelphi model trained with the mouse embryonic stem cell mutation dataset was used to predict the probabilities of hgRNA mutants from MARC1 mice. The original 64 hgRNA sequences in MARC1 mice were used as inputs. Since Cas9 nuclease cuts 3 bp upstream of the Protospacer Adjacent Motif (PAM, NGG sequence)[1], the possible mutations from the cut site at -3 bp from the PAM sequence were computed. To take into account the repeated targeting of hgRNAs, inDelphi is first applied to predict a set of first-round mutations. Subsequently, the resulting first-round mutations were used as inputs to the next round of inDelphi predictions. Notably, only mutant sequences with >16 bp protospacer and PAM were subject to the second round analysis as gRNA without >16 bp spacer sequence loses its activity [68]. Here, the probabilities of the next generation mutants were computed by multiplying the probabilities of the mutant in the current round by the probabilities of the mutant in the previous round. Repetitive application of inDelphi produces exponentially growing numbers of potential mutant alleles. Therefore, the analysis was limited to three cycles, resulting in first to third generations of mutants. The same mutation can be created in multiple rounds, in such cases, the probabilities from multiple rounds were summed. Finally, probabilities of all mutant alleles were normalized to have a sum of one.

Simulation of lineage barcodes from time-scaled phylogeny

To simulate lineage barcodes mimicking barcoding in MARC1 mice, 50 hgRNAs that were of the 'mid' or 'fast' fast category were randomly sampled from the MARC1 pool of hgRNAs for each simulation (**Supplementary Data 2.4**). The corresponding estimated mutation rates and inDelphi predicted allele emergence probabilities were used as input to the mutagenesis model.

Phylotime for reconstructing time-scaled phylogeny from lineage barcodes

Our approach to reconstructing time-scaled phylogenetic trees for thousands of cells is based on a maximum likelihood estimation of pairwise temporal distances between cells. Given a pair of terminal cells (**Figure S2.8**), the branch length parameter t_{N_1} was estimated, which is the time since the most recent common ancestor (MRCA) of the two cells. For a barcoding site i with a mutation rate of λ_i, and probabilities $a_i = (a_1, a_2, \ldots, a_J)_i$ of mutating into alleles $(A_1, A_2, \ldots, A_J)_i$, the likelihood of observing the given allele M_{iC_1} and M_{iC_2} in a single barcoding site in two cells (c_1, c_2) is the sum of two terms:

$$p_i = p_{i,N_{1,1}} + p_{i,N_1,0}\left(p_{i,C_1}p_{i,C_2}\right) \quad \text{(Eq. 1)}$$

The first term, $p_{i,N_{1,1}}$, is the probability that a mutation has occurred before the MRCA, leading to identical alleles in both cells. The second term is the probability of observing the allele in each terminal cell respectively (p_{i,C_1} and p_{i,C_2}) conditional on no mutation occurring before the MRCA ($p_{i,N_1,0}$).

For the first term, for each allele, $p_{i,N_1,A_{ij}} = \left(1 - e^{-\lambda_i t_{N_1}}\right)a_{ij}$ if A_{ij} is common to all progenies of N_1, and $p_{i,N_1,A_{ij}} = 0$ otherwise. Then

$$p_{i,N_1,1} = \sum_{j=1}^{J_i} p_{i,N_1,A_{ij}}$$

is the sum over all probable mutations, where at most one term can be non-zero.

For the second term, we have

$$p_{i,N_1,0} = e^{-\lambda_i t_{N_1}}$$

and

$$p_{i,c} = \left(1 - e^{-\lambda_i (T - t_{N_1})}\right)^{1 - Ind(M_{ic} = \text{"0"})} \prod_j a_{ij}^{Ind(M_{ic} = A_{ij})} \left(e^{-\lambda_i (T - t_{N_1})}\right)^{Ind(M_{ic} = \text{"0"})}$$

where M_{ic} denote the allele observed for $c = C_1, C_2$ and "0" denotes an unmutated allele.

Because barcoding sites in our model were independent, the likelihood for the set of alleles observed in all barcoding sites was then the product of their individual likelihoods:

$$p\left(\{\lambda_i, a_i\}_i, t_{N_1}\right) = \prod_i p_i \qquad \text{(Eq. 2)}$$

To get estimates of pairwise distance, or equivalently, time until MRCA between two cells (t_{N_1}), we first plugged in estimates of mutation rates and allele emergence probabilities. Here, naive estimates of mutagenesis parameters were plugged in. The estimates of mutation rates were obtained by $\hat{\lambda} = -\frac{1}{T}\log(1 - F)$, where F is the mutated fraction and T is the total time from the start of the experiment to the sample collection. Mutant allele emergence probabilities were set according to a uniform prior by default, that is $a_j = 1/J$ for $j = 1, \ldots, J$. Alternatively, better estimates such as the across-animal estimates or inDelphi predictions may be used and are

expected to improve Phylotime performance. To get the optimal value of t_{N_1}, the following score equation was solved using the Newton-Raphson method:

$$\frac{d\log(p(t_{N_1}))}{dt_{N_1}} = 0 \quad \text{(Eq. 3)}$$

The distance between the two cells is $2(T - t_{N_1})$. Once all pairwise distances were computed with the above method, we applied UPGMA hierarchical clustering [56] to derive a phylogenetic tree wherein branch lengths represent actual time. We called this approach Phylotime.

Stem Cell Culture

Human induced pluripotent stem cells (iPSCs) from the EP1 line [69] were cultured in mTeSR1 (STEMCELL Technologies) on plates coated with Matrigel Growth Factor Reduced Basement Membrane Matrix (Corning). Cells were maintained at 37°C and 10% CO_2/5% O_2 conditions with daily media changes. When up to 80% confluent, cells were passaged by dissociation with Accutase (Sigma Aldrich) and seeded in mTeSR1 media supplemented with 5 μM blebbistatin (Sigma Aldrich).

Knock-in of an Inducible Cas9 Cassette

EP1 iPSCs were modified to express Cas9 protein under doxycycline induction. CRISPR/Cas9 was used to target and insert both a reverse tetracycline-controlled transactivator (rtTA) construct and a tetracycline-dependent Cas9 construct into each of the two copies of the AAVS1 safe harbor locus. The following plasmids were used to generate the cell line:

Cas9/AAVS1 gRNA- modified pSpCas9(BB)-2A-Puro (PX459) V2.0 (Addgene Plasmid #62988 [70] with T2A replaced with P2A) with gRNA sequence caccGGGGCCACTAGGGACAGGAT

Cas9 donor- modified Puro-Cas9 donor (Addgene Plasmid #58409 [71] with puromycin resistance replaced with blasticidin resistance)

AAVS1 donor- AAVS1-Neo-M2rtTA (Addgene Plasmid #60843 [72])

Cells were grown to 80% confluency and then dissociated with Accutase for 13 minutes to generate a single cell suspension. 50,000 cells were resuspended in mTeSR1 media with 5 μM blebbistatin and seeded into one well of a 24-well plate coated in Matrigel. The following day, 350 ng of Cas9/AAVS1 gRNA plasmid and 500 ng each of Cas9 donor and AAVS1 donor plasmids were combined and added to 48 ul Opti-MEM (Gibco). 2 ul of Lipofectamine Stem Transfection Reagent (Thermo Fisher) were added to the transfection mix, which was then vortexed and incubated for ten minutes at room temperature. The entire transfection mix was added to one well of cells. Media was replaced the following day. 40 hours after transfection, cells were transiently selected for 24 hours with 0.95 ug/mL puromycin, 5 ug/mL blasticidin, and 200 ug/mL G418 sulfate. Surviving cells were cultured to at least 30% confluency, and then dissociated to a single cell suspension for clonal expansion. 500-1000 cells were seeded in one well of a 6-well plate and cultured for 7-10 days before clonal colonies were picked and screened for the intended insertions. PCR was performed with the following primers to confirm knock-in of the Cas9 and rTTA constructs:

Cas9_F: CACCTTGTACTCGTCGGTGA

rtTA_F: GCTGATTATGATCCTGCAAGC

AAVS1_RHA_R: GGAACGGGGCTCAGTCTGA

Positive colonies were cultured and clonally expanded once more, with a second round of colony picking and PCR screening to ensure clonality of the final cell line.

Lentiviral Infection with an Array of hgRNAs

50,000 cells from the doxycycline-inducible Cas9 line were seeded into one well of a 24-well plate. The following day, 300 ng of Super PiggyBac Transposase (SBI System Biosciences), 700 ng of PiggyBac-hgRNA L21 library ([31]), and 50 ng of PiggyBac-hgRNA L21 + puro library were combined and added to 48 ul Opti-MEM (Gibco). 2 ul of Lipofectamine Stem Transfection Reagent (Thermo Fisher) were added to the transfection mix, which was then vortexed and incubated for ten minutes at room temperature. The entire transfection mix was added to one well of cells. edia was replaced the following day. Transfected cells were selected with 0.95 ug/mL puromycin for one week.

Selected cells were dissociated to single cell suspensions, and 500-1000 cells were seeded in one well of a 6-well plate and cultured for 7-10 days before clonal colonies were picked. Colonies were screened for hgRNA insertions using qPCR. Genomic DNA from each colony was amplified with the following pairs of primers:

Sox11_F: TGATGTTCGACCTGAGCTTG

Sox11_R: TAGTCGGGGAACTCGAAGTG

hgRNA_F: ATGGACTATCATATGCTTACCGT

hgRNA_R: TTCAAGTTGATAACGGACTAGC

For each colony, the cycle threshold value for hgRNA amplification was subtracted from that of Sox11 amplification. Colonies with the largest cycle threshold value difference, indicating the highest number of hgRNA insertions, were cultured and clonally expanded once more. Colonies were picked one additional time to ensure clonality of the final cell line.

Determining cell line hgRNA array identity and function

Cell line hgRNA identities and functions were determined by performing a doxycycline time course experiment. Genomic DNA was extracted from cells after 0, 4, 8 and 11 days of doxycycline treatment and Cas9 induction. hgRNA sequencing libraries were prepared, sequenced, and analyzed as per the published pipeline [52]. The percent of reads for each hgRNA identifier sequence that were mutated was calculated at each time point (**Figure S2.4**), determining the relative activity for every hgRNA.

In vitro quantitative fate map experiments

Single cells from the Cas9-hgRNA iPSC cell line were FACS sorted into a 96-well plate coated with Matrigel and containing mTeSR plus medium supplemented with 5 μM blebbistatin, 10% CloneR (StemCell Technologies), and 1 μM Pifithrin-α hydrobromide (Tocris), 1X Antibiotic-Antimycotic (Gibco), and 0.2 ug/uL doxycycline. Supplemented media was replaced every other day. Wells were assigned to follow specific quantitative fate maps, and were passaged at times, sizes, and ratios as determined by the fate maps. To passage small numbers of cells, media was aspirated from wells and 30-100 ul of Accutase (depending on the size of the well) was added and incubated for 8 minutes. The Accutase and cell suspension was directly added to supplemented media in the wells the cells were passaged to. Passaged cells were incubated for 2 hours to allow cells to attach to the Matrigel coating, after which a 50% media exchange was performed to decrease the amount of Accutase remaining with the cells. At the end of the experiments when the cells were passaged into their terminal wells, doxycycline treatment was ended so that barcode editing would discontinue.

Determining progenitor population size from in vitro experiment

Brightfield images were taken of cells at P5, P4, and P3 for E1 and E2 before passaging (**Figure S2.6**) To estimate the cell numbers at each state, images were analyzed in ImageJ.

Outlines were drawn for 5 different cells, and the average area noted. The total area of the colony or colonies was then measured, and divided by the average cell area to estimate the total number of cells present in each image.

Sequencing single cell lineage barcodes

Cells were dissociated into single cell suspensions and diluted in PBS pH 7.4. Single cells were FACS sorted into 1 ul each of QuickExtract DNA Extraction Buffer (Lucigen). DNA was extracted by incubating cells for 10 minutes at 65°C followed by 5 minutes at 98°C. Single cell barcode sequencing libraries were generated using serial PCR reactions as per the published protocol ([52]) with each cell treated as an individual sample. Libraries were sequenced on a MiSeq System (Illumina).

Data processing for in vitro experiment data

Identifier and spacer sequence pairs were extracted for each sample using the initial step (BLAST search) detailed in the published pipeline. [52] For cells that were sequenced more than once, pair counts for unique “identifier+spacer” were first merged for each cell. The merged data were then provided as input to the remainder of the pipelines for sequencing error correction and filtering.

A total of 32 hgRNAs were identified from the filtered results, each observed in more than 921 cells. Sequencing errors among identifier sequences were first corrected. For each identifier sequence that was not one of the 32 hgRNAs observed in the unmutated sample, if the identifier sequence was within a hamming distance of 1 to any true identifiers, its spacer counts were merged with that of the known hgRNA’s. After the correction, no other identifiers other than the 32 known hgRNAs were observed in more than 3 cells.

Spacer sequencing errors were corrected next. First, the error reads within each cell and hgRNA combination were corrected. In one of our sequencing runs, one cycle of sequencing returned ‘N’ for all spacer sequences. These errors were computationally corrected: if there existed another spacer sequence for the same identifier and cell that was exactly the same except for the ‘N’ base pair, the count of the error spacer with the ‘N’ base was merged with the other spacer. Next, the spacer sequencing errors across different cells were corrected. Again, the error involving ‘N’ base pairs were further corrected across the cells using the same criteria as the within cell correction.

Each allele was labeled as unmutated or mutated by comparing the spacer sequences to that of the reference sequencing result, that is, if a spacer sequence was observed in the parent for the same identifier, it was labeled as unmutated. One identifier “GCCAAAAGCT” did not amplify in the parent data, and the sequence “GAAACACCGGTGGTCGCCGTGGAGAGTGGTGGGGTTAGAGCTAGAAATAG” was identified as the unmutated spacer based on alignments of its different observed alleles.

Noisy reads were further filtered for cell+hgRNA combinations that had more than one spacer observed. If the most abundant spacer was at least four times more abundant than all the other spacer reads observed, only the most abundant spacer was kept. All the spacer counts with fewer than five total reads were also excluded.

After processing, 1197 cell+hgRNA combinations out of the total 54,012 (2.2%) still had more than one spacer observed. If more than two spacers were observed for more than two hgRNAs in a single cell, the cell was likely a doublet. 33 such cells were identified and filtered. Each cell had a median 25 out of 32 hgRNAs detected. Each hgRNA was detected in a median of 83% of all cells (**Figure S2.7**). Before reconstruction, non-informative cells and hgRNAs were filtered out. Any hgRNA with a diversity of one, that is, all cells in which an hgRNA was observed had

the same allele, was considered non-informative and was excluded. An allele was considered informative if it was mutated and observed in more than one cell in each group, and cells with less than three informative alleles were filtered out. In all, 970 out of 1051 cells for 31 hgRNAs passed the filters for E1 and 943 out of 1032 cells for 29 hgRNAs passed filters for E2.

Imputation of missing hgRNA alleles with xgboost

As an intermediate step, the missing alleles were imputed before ICE-FASE could be applied to the in vitro data. First, the missing percentages of all hgRNAs were computed, and hgRNAs were imputed one by one going from the one with fewest cells missing to the most cells missing. To impute missing alleles for a single hgRNA site, an ‘xgboost’ model with multinomial softmax objective was trained using all the observed cells as training [58]. The model classified each missing cell as one of the observed alleles. The design matrix was constructed where each column is a mutant allele observed in one of the other hgRNAs. The parameters ‘max_depth=4’ and ‘nrounds = 20’ were used for the xgboost models.

Simulation and ground truth fate map of in vitro experiment

To conduct simulations that best resembled the in vitro experiment, the effective cell division rates during the experiment were first determined. First, the division rate of P5 was chosen so that the population size at the first split is the most consistent with what was observed. Next, we assumed that cells were split in proportion to the volume of the suspension as P5 was split into P3 and P4. The division rate of P3 and P4 were set so that their respective population sizes at the split agreed with what was observed. The division rate for P1, P2 and all the terminal cells were set to once every 20 hours. The exact division rate chosen and the population size at each stage are detailed in **Figure S2.5**. Notice that the commitment happens one cell division prior to the well split, so the ground truth commitment time is one cell division earlier than the spit time, and the progenitor field size is half of what was observed in the well at the split.

To simulate hgRNA barcodes from ground truth fate map, mutation rates were estimated from the time course data of bulk mutated fractions from the iPSC line. For mutation profiles, predictions from inDelphi (**Supplementary Data 2.3**) were used.

Appendix

Chapter 1 Supplementary Figures

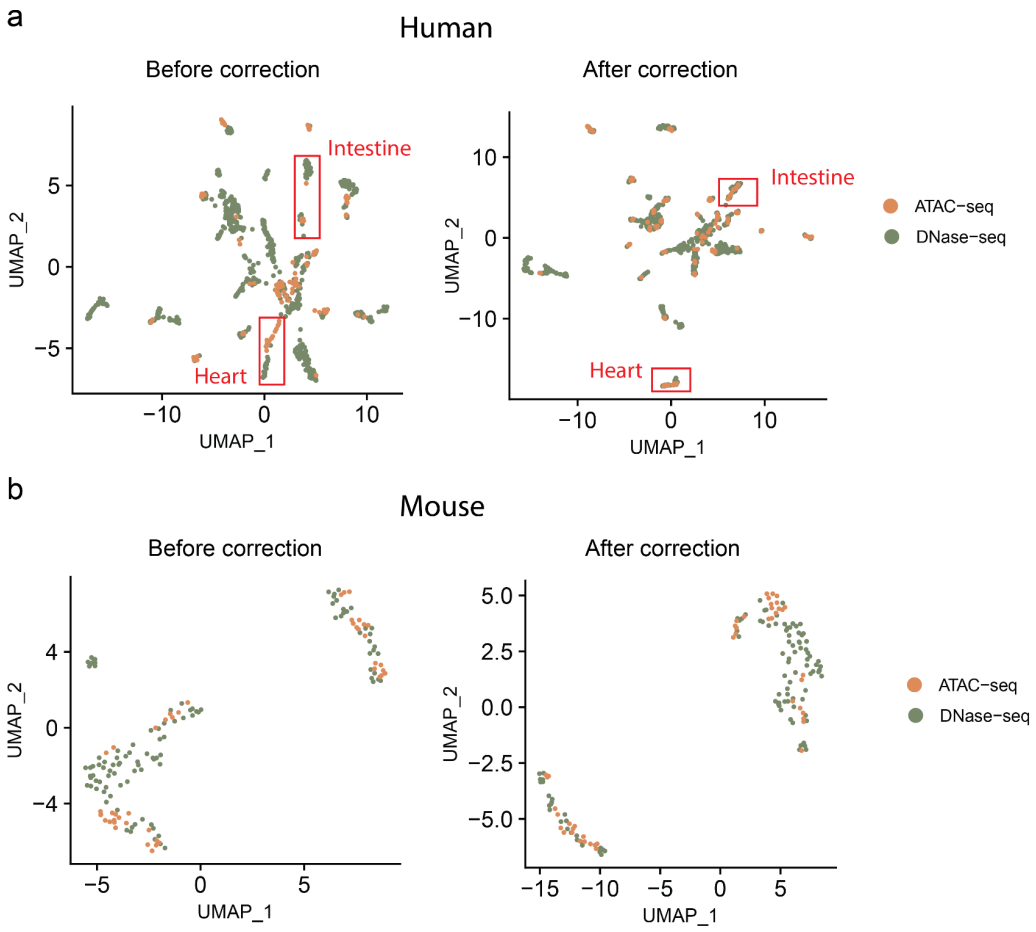

Figure S1.1. Assay effect correction between human DNase-seq and ATAC-seq. (a) UMAP embedding of human DNase-seq and ATAC-seq samples before and after assay effect correction. Heart and intestine tissues are highlighted as examples, samples were observed to clustered more according to tissue than according to assay. (b) (a) UMAP embedding of mouse

DNase-seq and ATAC-seq samples before and after assay effect correction. No significant change in cell type clustering were observed.

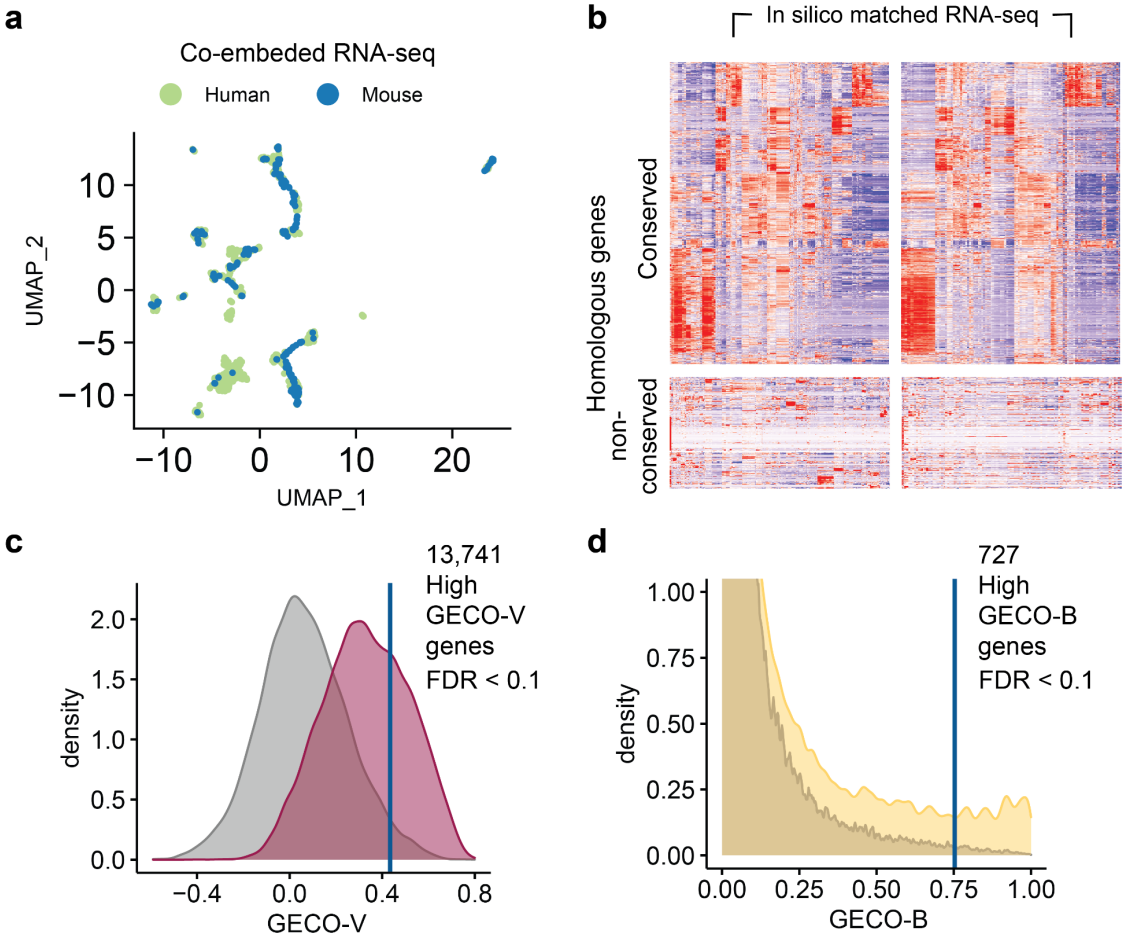

Figure S1.2. Computation of gene expression conservation scores. (a) UMAP visualization of integrated human and mouse RNA-seq samples. (b) Scaled homologous gene expression signals for in-silico matched RNA-seq samples. (c-d) Distribution of GECO-V and GECO-B scores (red and yellow) for homologous gene pairs versus the null distribution based on randomly paired genes (gray).

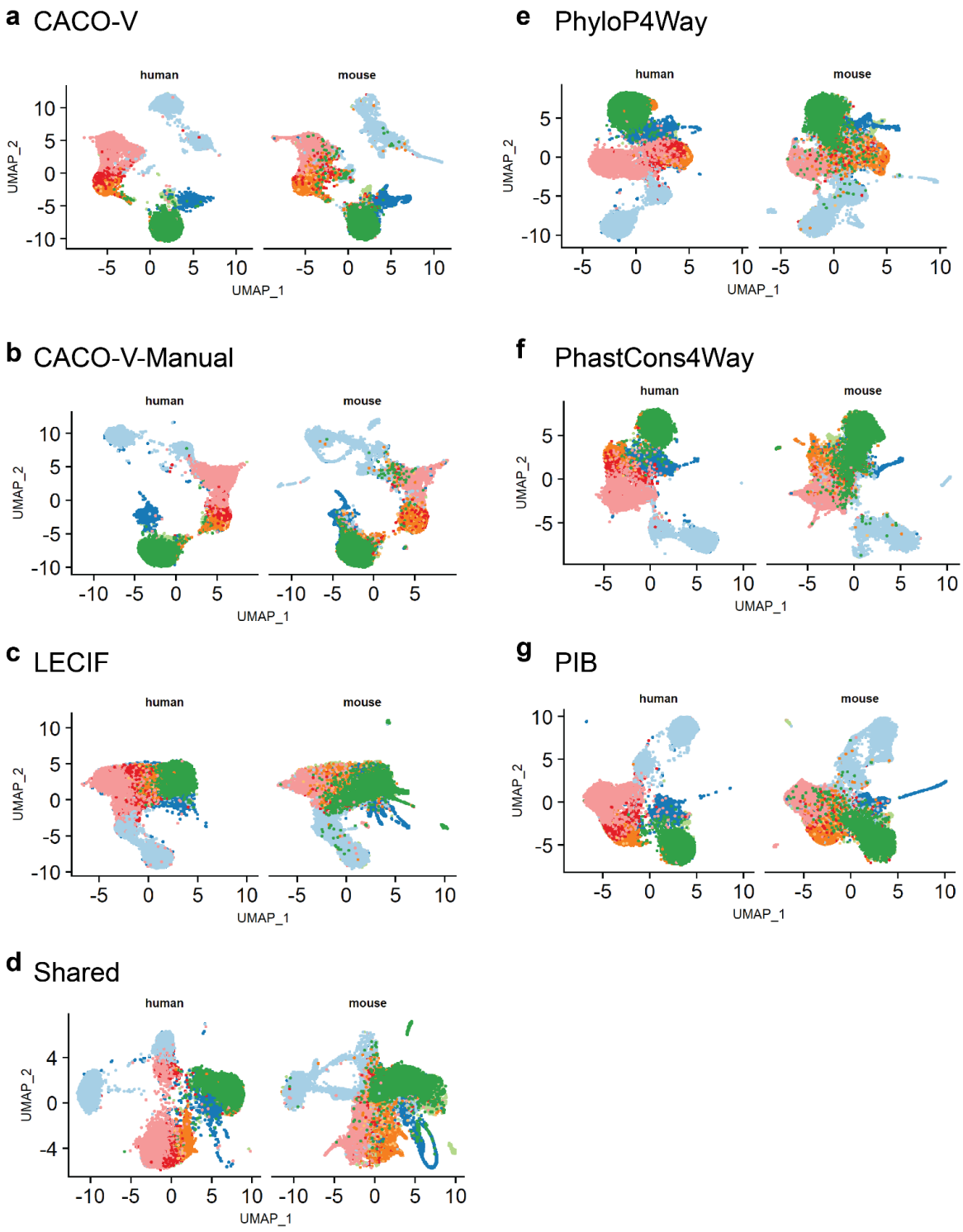
a CACO-V
human
mouse
UMAP_2
UMAP_1
b CACO-V-Manual
c LECIF
d Shared
e PhyloP4Way
f PhastCons4Way
g PIB

Figure S1.3. Visualization of Seurat integrated results with different conservation scores. top 48,160 features were selected ranking by (a) CACO-V (b) CACO-V-Manual (c) LECIF (d) Shared (e) PhyloP4Way (f) PhastCons4Way or (g) PIB

Chapter 2 Supplementary Figures

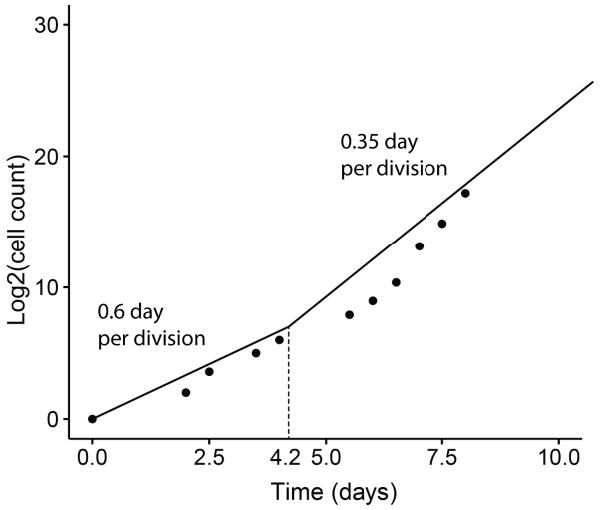

Figure S2.1. Comparison between the total number of cells in simulation and the reported number of cells in early mouse embryogenesis. Dots represent the number of cells from [44]. The line represents the number of cells in the simulation. The two cell division rates used for simulation are shown.

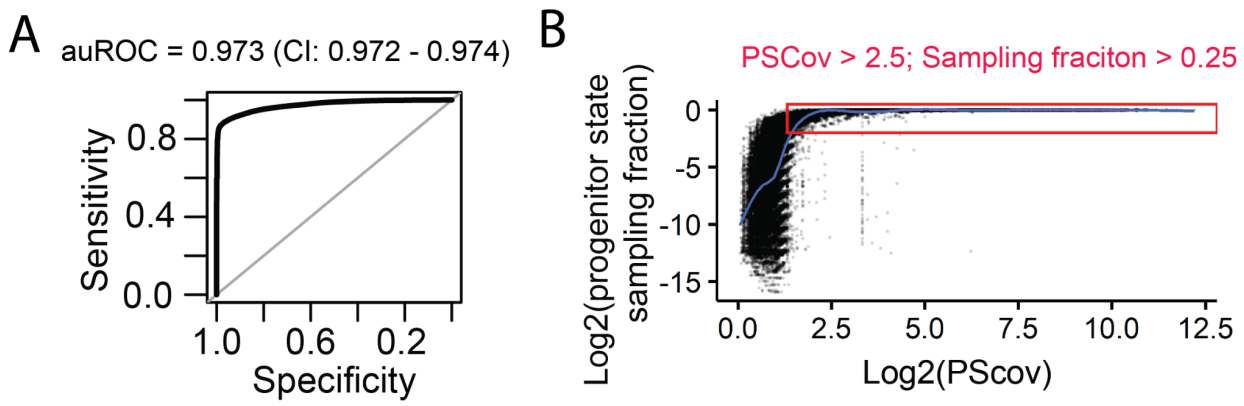

Figure S2.2. Progenitor state coverage statistics (PScov) reveal robustness of obtained quantitative fate map parameters. (**A**) Area Under the Receiver Operating Characteristics (auROC) curve for progenitor state coverage shows a high sensitivity and specificity in detecting adequately sampled progenitor states. CI: confidence interval. (**B**) Scatter plot of progenitor state sampling fraction as a function of its PScov among 2,500 simulated experiments. Progenitor states with sampling fraction better than 0.25, which produced robust estimates of progenitor states, tend to have coverage above 2.5, as highlighted by the red box. Trendline (LOESS) is shown in blue.

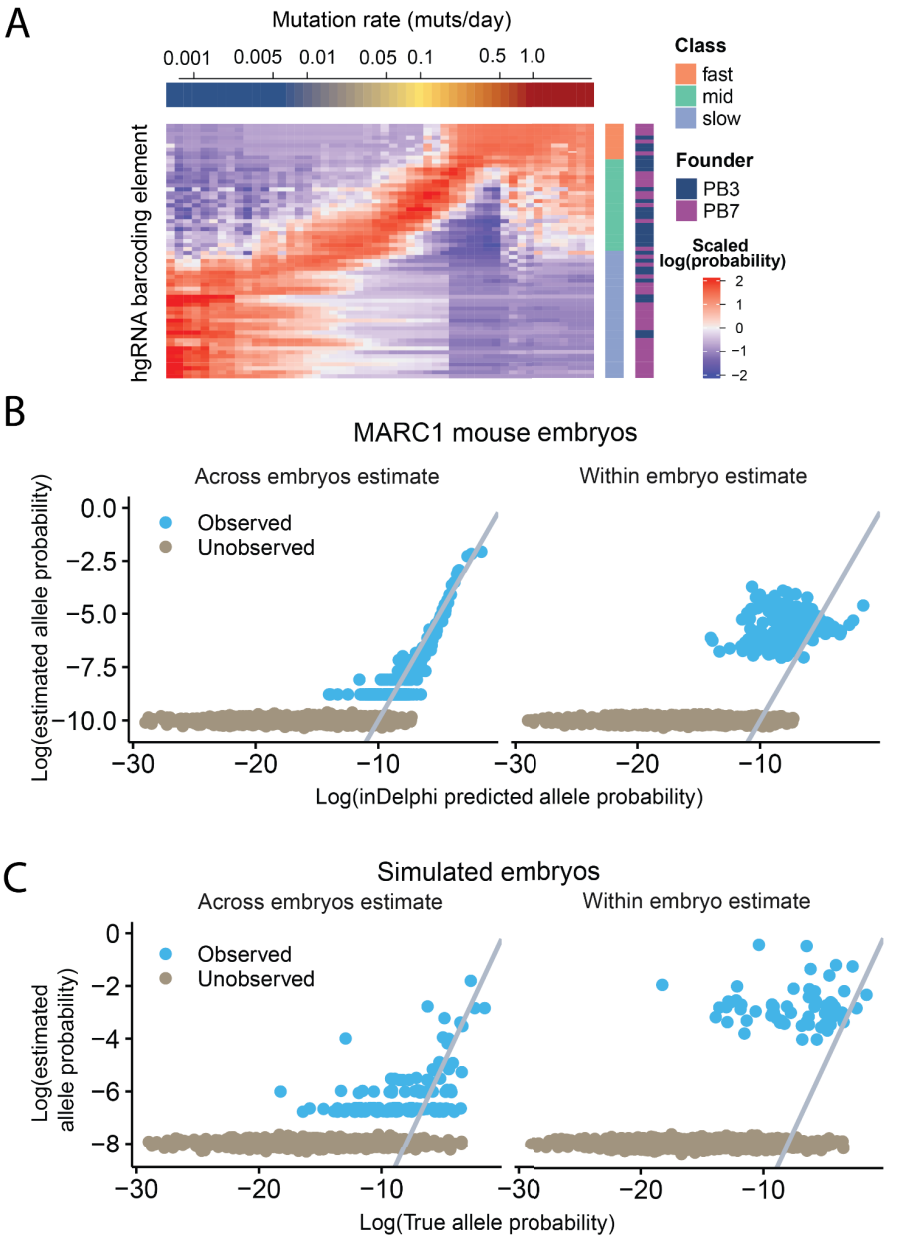

Figure S2.3. Agreement between inDelphi-based allele predictions and those observed in simulation and mouse experiments. (**A**) Posterior probabilities of mutation rates for all MARC1 hgRNAs which originate from either the PB3 founder mouse or the PB7 founder mouse (Leeper et al., 2021) as shown in either dark navy or purple on the color bar to the right. Also shown is the initial characterization label of the hgRNA as either fast, mid, or slow (Leeper et al., 2021).

(**B**) Comparison of mutant allele probability estimates of our modified inDelphi algorithm to those estimated from MARC1 mouse embryos by taking either abundance within the embryo when the allele is present (right) or fraction of embryos that present the allele (left). Blue dots represent predicted alleles that were observed in MARC1 embryos, gray ones were predicted by modified inDelphi but not observed in embryos. Alleles observed in embryos but not predicted by inDelphi, which were a very small fraction, are not shown. Gray line has intercept 0, and slope 1. (**C**) Comparison of true mutant allele probability as estimated (x-axis) with their average observed fraction in simulated mouse embryos (right) and fraction of simulated embryos that showed the allele (left). Blue dots represent alleles that were observed in simulated embryos, gray ones were possible in simulation by not observed in simulated embryos. Simulations were conducted 100 times for 9 time points. Gray line has intercept 0, and slope 1. Panels B and C show that the estimator based on allele occurrences across multiple samples outperforms average allele fraction in both simulation and MARC1 mouse data.

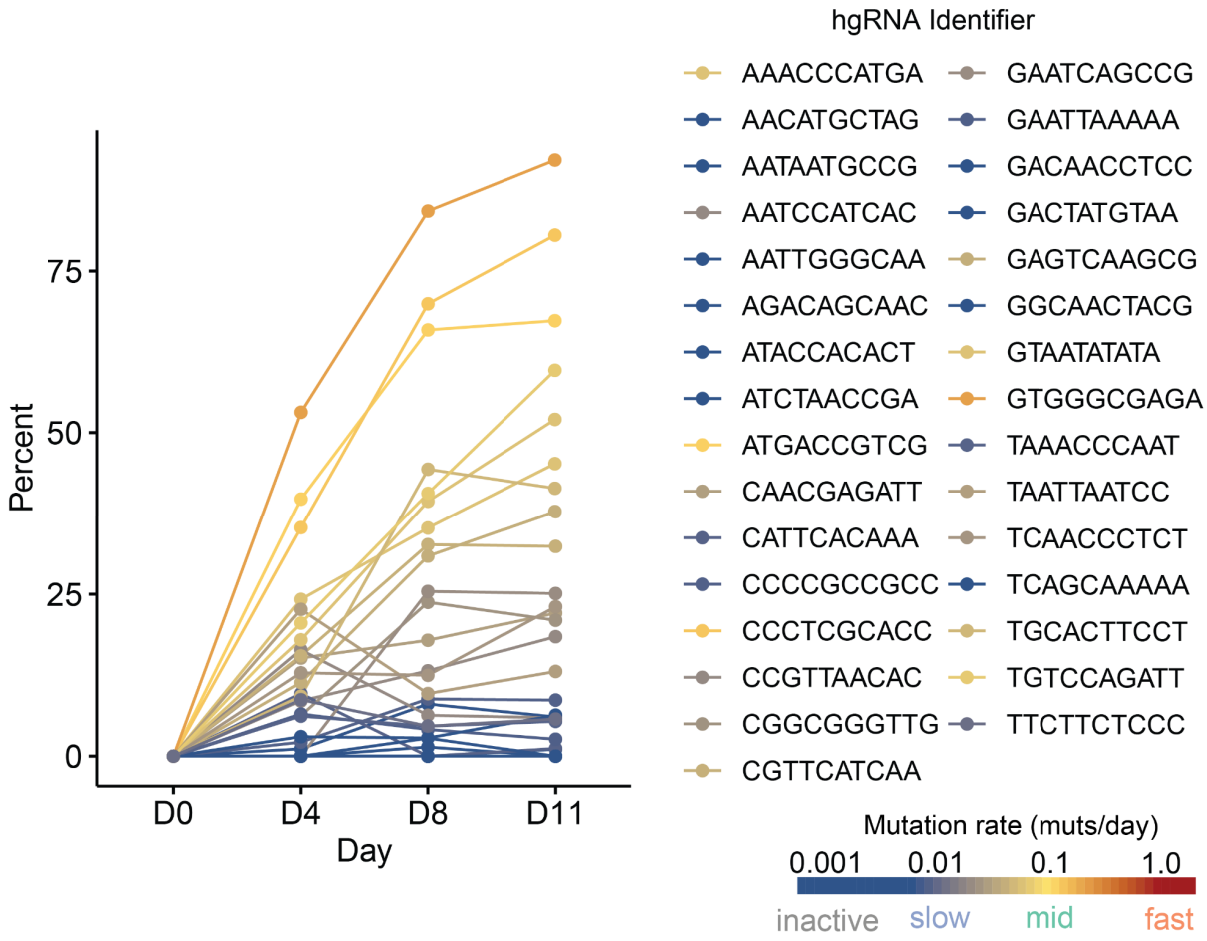

Figure S2.4. Mutated fraction of hgRNAs over time in the iPSC line. Color shows the estimated mutation rate for each hgRNA according to the key on bottom right. The list on top right denotes the color of each hgRNA on the plot.

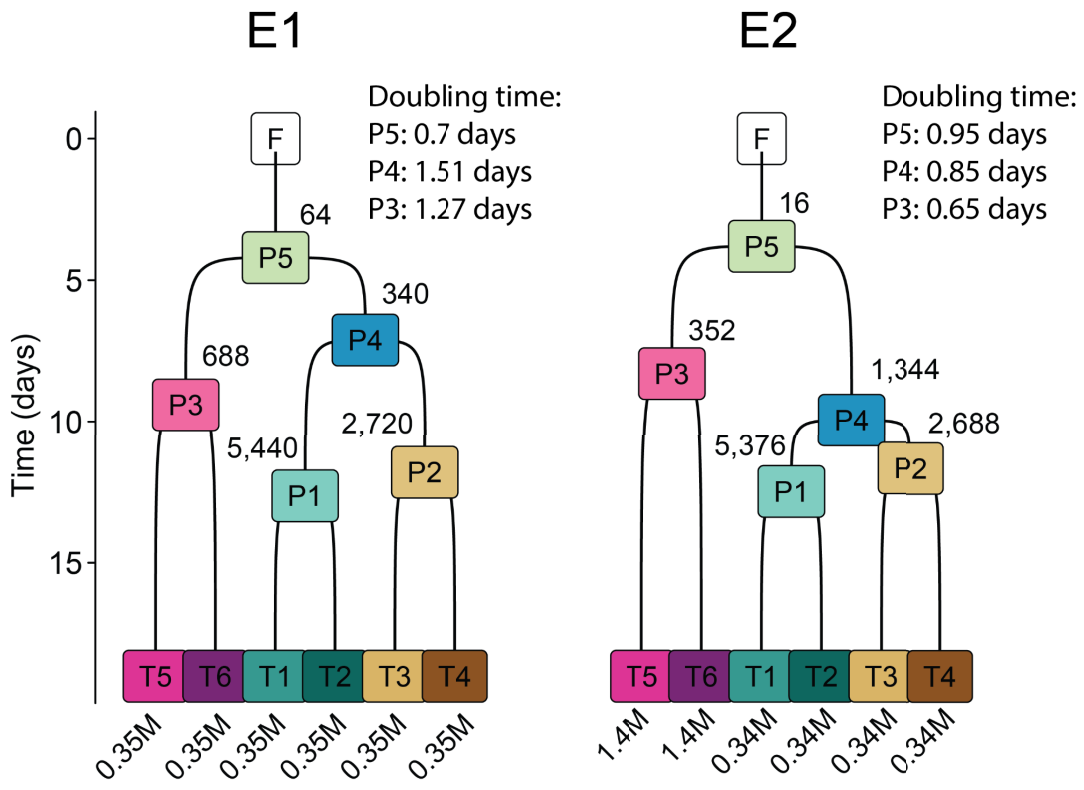

Figure S2.5. Cell division rates and progenitor population sizes in the ground truth fate maps for in vitro experiments. Numbers on the top right corner of each progenitor state show its total population size estimated from bright field images. Numbers on the bottom of each terminal state show its estimated population size from bright field images. The estimated doubling times, which represent cell division rates, estimated from population size at different times, are shown on the top right of each map.

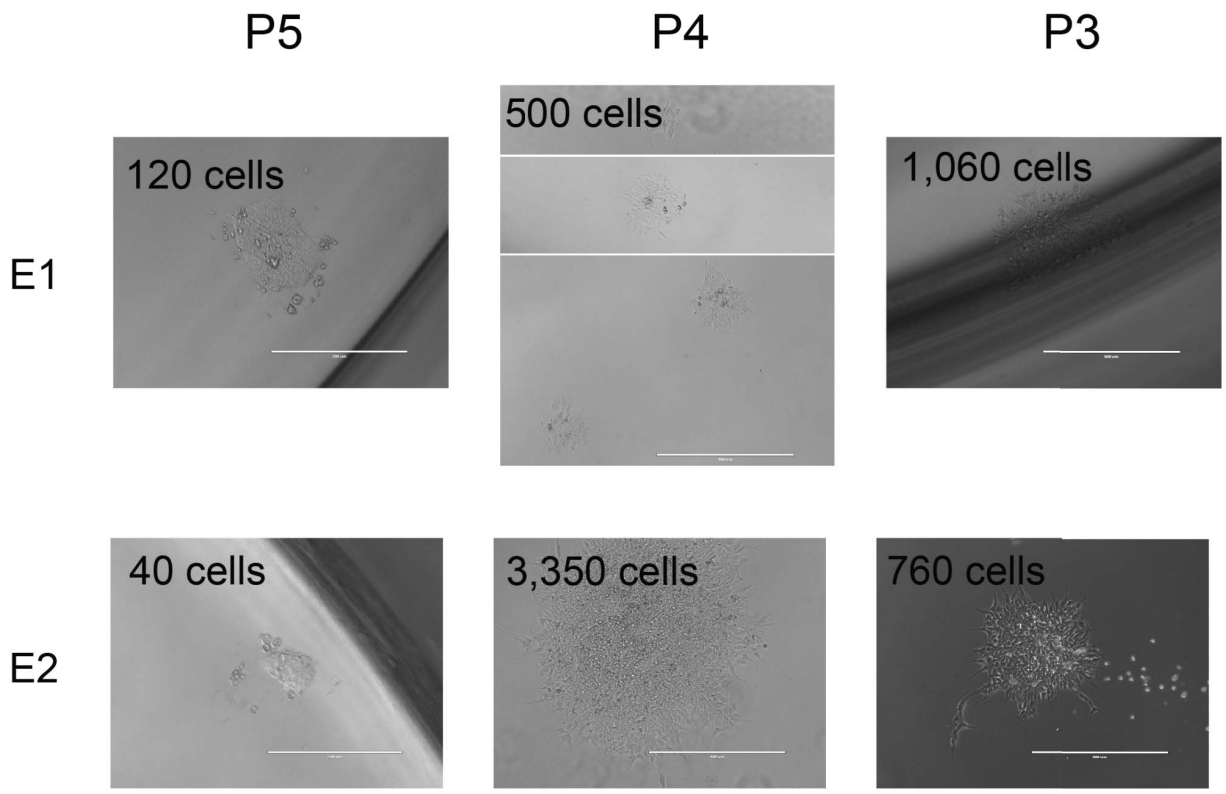

Figure S2.6. Bright field images showing the P3, P4, and P5 progenitor population size estimates (columns) for E1 and E2 experiments (rows). Scale bars for P5 images are 200 microns. Scale bars for P4 and P3 images are 400 microns.

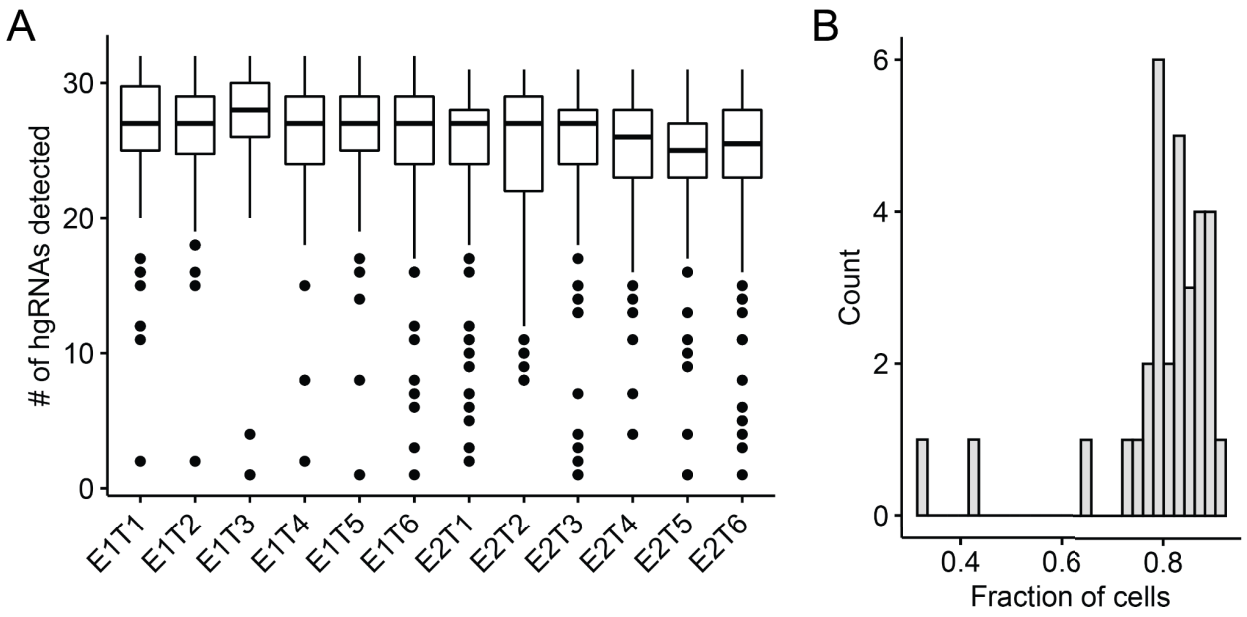

Figure S2.7. Amount of undetected hgRNA alleles. (a) Boxplot showing number of hgRNAs detected (out of the total 32) in each terminal well of each experiment. (b) Histogram of the fraction of cells in which each hgRNA was detected.

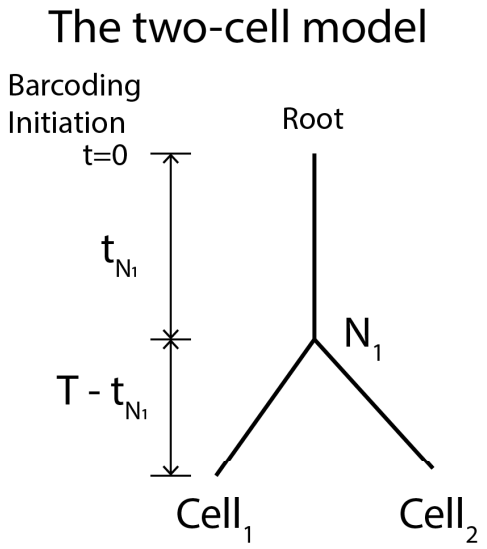

Figure S2.8. Diagram showing a two-cell phylogram illustrating Phylotime likelihood computation. N_1 is the most recent common ancestor (MRCA) of terminal $Cell_1$ and $Cell_2$.

hgRNA Identifier Sequence	hgRNA Spacer Sequence
AAACCCATGA	GGTTTAGTATGGAGGGAAGTG
AACATGCTAG	GGTTTTCGATCATAGGTCGTG
AATAATGCCG	GGTGCCCTTTATGGGACCGTG
AATCCATCAC	GGTGACAAACGTATGTCAGTG
AATTGGGCAA	GGTTGTGGGGAACTCGGTGTG
AGACAGCAAC	GGTGTTCTTTGCTCGAGGGTG
ATACCACACT	GGTTGTAAGCGATGGTTTGTG
ATCTAACCGA	GGTGTGCGGAGATTATGCGTG
ATGACCGTCG	GGTCCTAGAAGCTGAGTTGTG
CAACGAGATT	GGTGCTAGCGGTTTCGAAGTG
CATTCACAAA	GGTGGCAATTCGATCTGAGTG
CCCCGCCGCC	GGTAGAATGGATCCACGGGTG
CCCTCGCACC	GGTTCAACCGGCGTCTTTGTG
CCGTTAACAC	GGTTGGCTTTACTCCTTTGTG
CGGCGGGTTG	GGTCGTCGTTCGTAGGGCGTG
CGTTCATCAA	GGTGAGAACAGAACGTTTTGG
GAATCAGCCG	GGTTTTAGTAAATGGTGAGTG
GAATTAAAAA	GGTGCTAATAGTTAGCTCGTG
GACAACCTCC	GGTAATCTAAAGATCCCCGTG
GACTATGTAA	GGTACTAGTTACTTACGGGTG
GAGTCAAGCG	GGTTATCGTTACGGATTTGTG
GCCAAAAGCT	GGTGGTCGCCGTGGAGAGGTG
GGCAACTACG	GGTGGCAGCCAGCCACTTGTG
GTAATATATA	GGTGAGCATGATAACGTCGTG
GTGGGCGAGA	GGTGAGATGCCTCAAGTGGTG
TAAACCCAAT	GGTTTACTTAGTTAACTAGTG
TAATTAATCC	GGTTGAGATAATCAAAAAGTG
TCAACCCTCT	GGTGACGCGAGGACGGTGGTG
TCAGCAAAAA	GGTG
TGCACTTCCT	GGTTCCCGTTAGCTCGTAGTG
TGTCCAGATT	GGTCTTCCTGGAATCTACGTG
TTCTTCTCCC	GGTCGCGAAAATGTGCCGGTG

Table S2.1. hgRNA identifier and spacer sequences amplified from the iPSC line.

Chapter 1 List of supplementary data

The supplementary data are available from the github repo:

https://github.com/wefang/hg38_mm10_OpenChromatinCnsv

Supplementary data 1.1. List of DNase-seq and ATAC-seq samples for human.

Supplementary data 1.2. List of DNase-seq and ATAC-seq samples for mouse.

Supplementary data 1.3. List of Anchors identified for chromatin accessibility data.

Supplementary data 1.4. GREAT analysis for conserved elements.

Supplementary data 1.5. List of H3K4me1 Histone ChIP-seq data for human.

Supplementary data 1.6. List of H3K4me1 Histone ChIP-seq data for mouse.

Supplementary data 1.7. Summary table for all alignable DHS.

Supplementary data 1.8. List of Human RNA-seq experiments.

Supplementary data 1.9. List of mouse RNA-seq experiments.

Supplementary data 1.10. List of anchors for RNA-seq data.

Supplementary data 1.11. Gene ontology analysis for genes with conserved expression.

Supplementary data 1.12. Summary table for all homologous gene pairs.

Supplementary data 1.13. List of manually matched samples.

Supplementary data 1.14. List of samples held out for validation.

Supplementary data 1.15. GWAS disease and trait associations with tissue or cell types.

Supplementary data 1.16. Matched labels for human and mouse sciATAC-seq data.

Chapter 2 List of supplementary data

Supplementary data and raw data is available from the github repo:

https://github.com/Kalhor-Lab/QFM-Data

Supplementary data 2.1. All quantitative fate maps.

Supplementary data 2.2. Mutation rate estimates for hgRNAs in MARC1 mice and iPSC line.

Supplementary data 2.3. inDelphi predicted mutant allele probabilities for hgRNAs in MARC1 mice and iPSC line.

Supplementary data 2.4. Simulated phylogenies, single cell lineage barcodes, Phylotime reconstructed trees, fate map and set of MARC1 hgRNAs used for all experiments.

Bibliography

1. Consortium MGS, Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002. pp. 520–562. doi:10.1038/nature01262

2. Filipski A, Kumar S. Comparative Genomics in Eukaryotes. The Evolution of the Genome. 2005. pp. 521–583. doi:10.1016/b978-012301463-4/50011-5

3. Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, et al. Principles of regulatory information conservation between mouse and human. Nature. 2014;515: 371–375.

4. Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Scott Hansen R, et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science. 2014 [cited 13 Oct 2021]. Available: https://science.sciencemag.org/content/346/6212/1007.summary

5. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer evolution across 20 mammalian species. Cell. 2015;160: 554–566.

6. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15: 1034–1050.

7. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20: 110–121.

8. Kwon SB, Ernst J. Learning a genome-wide score of human-mouse conservation at the functional genomics level. Nat Commun. 2021;12: 2495.

9. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human–Mouse Alignments with BLASTZ. Genome Res. 2003;13: 103–107.

10. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177: 1888–1902.e21.

11. Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature. 2020;584: 244–251.

12. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583: 699–710.

13. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell. 2018;174: 1309–1324.e18.

14. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A single-cell atlas of chromatin accessibility in the human genome. Cell. 2021;184: 5985–6001.e19.

15. Adorf CS, Dodd PM, Ramasubramani V, Glotzer SC. Simple data and workflow management with the signac framework. Comput Mater Sci. 2018;146: 220–229.

16. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park J-E. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020;36: 964–965.

17. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell. 2019;177: 1873–1887.e17.

18. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature Biotechnology. 2018. pp. 421–427. doi:10.1038/nbt.4091

19. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14: 128.

20. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9: e1003118.

21. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28: 495–501.

22. Boix CA, James BT, Park YP, Meuleman W, Kellis M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature. 2021;590: 300–307.

23. Baron CS, van Oudenaarden A. Unravelling cellular relationships during development and regeneration using genetic lineage tracing. Nat Rev Mol Cell Biol. 2019;20: 753–765.

24. Wagner DE, Klein AM. Lineage tracing meets single-cell omics: opportunities and challenges. Nat Rev Genet. 2020;21: 410–427.

25. Stadler T, Pybus OG, Stumpf MPH. Phylodynamics for cell biologists. Science. 2021;371. doi:10.1126/science.aah6266

26. McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science. 2016. p. aaf7907. doi:10.1126/science.aaf7907

27. Alemany A, Florescu M, Baron CS, Peterson-Maduro J, van Oudenaarden A. Whole-organism clone tracing using single-cell sequencing. Nature. 2018;556: 108–112.

28. Spanjaard B, Hu B, Mitic N, Olivares-Chauvet P, Janjuha S, Ninov N, et al. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. Nat Biotechnol. 2018;36: 469–473.

29. Raj B, Wagner DE, McKenna A, Pandey S, Klein AM, Shendure J, et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat Biotechnol. 2018;36: 442–450.

30. Liu K, Deng S, Ye C, Yao Z, Wang J, Gong H, et al. Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. Nat Methods. 2021;18: 1506–1514.

31. Kalhor R, Kalhor K, Mejia L, Leeper K, Graveline A, Mali P, et al. Developmental barcoding of whole mouse via homing CRISPR. Science. 2018;361. doi:10.1126/science.aat9804

32. Chan MM, Smith ZD, Grosswendt S, Kretzmer H, Norman TM, Adamson B, et al. Molecular recording of mammalian embryogenesis. Nature. 2019;570: 77–82.

33. Bowling S, Sritharan D, Osorio FG, Nguyen M, Cheung P, Rodriguez-Fraticelli A, et al. An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. Cell. 2020;181: 1410–1422.e27.

34. Bizzotto S, Dou Y, Ganz J, Doan RN, Kwon M, Bohrson CL, et al. Landmarks of human embryonic development inscribed in somatic mutations. Science. 2021;371: 1249–1253.

35. Fasching L, Jang Y, Tomasi S, Schreiner J, Tomasini L, Brady MV, et al. Early developmental asymmetries in cell lineage trees in living individuals. Science. 2021;371: 1245–1248.

36. Coorens THH, Moore L, Robinson PS, Sanghvi R, Christopher J, Hewinson J, et al. Extensive phylogenies of human development inferred from somatic mutations. Nature. 2021;597: 387–392.

37. Spencer Chapman M, Ranzoni AM, Myers B, Williams N, Coorens THH, Mitchell E, et al. Lineage tracing of human development through somatic mutations. Nature. 2021;595: 85–90.

38. Cheedipudi S, Genolet O, Dobreva G. Epigenetic inheritance of cell fates during embryonic development. Front Genet. 2014;5: 19.

39. Sulston JE, Schierenberg E, White JG, Thomson JN. The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol. 1983;100: 64–119.

40. Salipante SJ, Horwitz MS. Phylogenetic fate mapping. Proc Natl Acad Sci U S A. 2006;103: 5448–5453.

41. Salipante SJ, Horwitz MS. A phylogenetic approach to mapping cell fate. Curr Top Dev Biol. 2007;79: 157–184.

42. Davidson EH. Later embryogenesis: regulatory circuitry in morphogenetic fields. Development. 1993;118: 665–690.

43. Shao K-T, Sokal RR. Tree Balance. Systematic Biology. 1990. pp. 266–276. doi:10.2307/2992186

44. Kojima Y, Tam OH, Tam PPL. Timing of developmental events in the early mouse embryo. Semin Cell Dev Biol. 2014;34: 65–75.

45. Kingman JFC. On the genealogy of large populations. Journal of Applied Probability. 1982. pp. 27–43. doi:10.2307/3213548

46. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol. 1983;23: 183–201.

47. Kendall M, Colijn C. Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution.

Mol Biol Evol. 2016;33: 2735–2743.

48. Yang Z. Molecular Evolution: A Statistical Approach. OUP Oxford; 2014.

49. Salvador-Martínez I, Grillo M, Averof M, Telford MJ. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? Elife. 2019;8. doi:10.7554/eLife.40292

50. Benchmarked approaches for reconstruction of in vitro cell lineages and in silico models of C. elegans and M. musculus developmental trees. Cell Systems. 2021;12: 810–826.e4.

51. Konno N, Kijima Y, Watano K, Ishiguro S, Ono K, Tanaka M, et al. Deep distributed computing to reconstruct extremely large lineage trees. Nat Biotechnol. 2022. doi:10.1038/s41587-021-01111-2

52. Leeper K, Kalhor K, Vernet A, Graveline A, Church GM, Mali P, et al. Lineage barcoding in mice with homing CRISPR. Nat Protoc. 2021;16: 2088–2108.

53. Kalhor R, Mali P, Church GM. Rapidly evolving homing CRISPR barcodes. Nat Methods. 2017;14: 195–200.

54. Shen MW, Arbab M, Hsu JY, Worstell D, Culbertson SJ, Krabbe O, et al. Predictable and precise template-free CRISPR editing of pathogenic variants. Nature. 2018;563: 646–651.

55. Feng J, DeWitt WS III, McKenna A, Simon N, Willis AD, Matsen FA IV. Estimation of cell lineage trees by maximum-likelihood phylogenetics. Ann Appl Stat. 2021;15. doi:10.1214/20-aoas1400

56. Sokal RR, Michener CD, University of Kansas. A Statistical Method for Evaluating Systematic Relationships. 1958.

57. Jones MG, Khodaverdian A, Quinn JJ, Chan MM, Hussmann JA, Wang R, et al. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. Genome Biol. 2020;21: 92.

58. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. pp. 785–794.

59. Hormoz S, Singer ZS, Linton JM, Antebi YE, Shraiman BI, Elowitz MB. Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. Cell Syst. 2016;3: 419–433.e8.

60. Seidel S, Stadler T. TiDeTree: A Bayesian phylogenetic framework to estimate single-cell trees and population dynamic parameters from genetic lineage tracing data. bioRxiv. 2022. p. 2022.02.14.480422. doi:10.1101/2022.02.14.480422

61. Blanpain C, Simons BD. Unravelling stem cell dynamics by lineage tracing. Nat Rev Mol Cell Biol. 2013;14: 489–502.

62. Mittnenzweig M, Mayshar Y, Cheng S, Ben-Yair R, Hadas R, Rais Y, et al. A single-embryo, single-cell time-resolved model for mouse gastrulation. Cell. 2021;184: 2825–2842.e22.

63. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell

transcriptional landscape of mammalian organogenesis. Nature. 2019;566: 496–502.

64. Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, Calero-Nieto FJ, et al. A single-cell molecular map of mouse gastrulation and early organogenesis. Nature. 2019;566: 490–495.

65. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35: 526–528.

66. Ford D, Matsen FA, Stadler T. A method for investigating relative timing information on phylogenetic trees. Syst Biol. 2009;58: 167–183.

67. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27: 592–593.

68. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat Biotechnol. 2014;32: 279–284.

69. Bhise NS, Wahlin KJ, Zack DJ, Green JJ. Evaluating the potential of poly(beta-amino ester) nanoparticles for reprogramming human fibroblasts to become induced pluripotent stem cells. Int J Nanomedicine. 2013;8: 4641–4658.

70. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nat Protoc. 2013;8: 2281–2308.

71. González F, Zhu Z, Shi Z-D, Lelli K, Verma N, Li QV, et al. An iCRISPR platform for rapid, multiplexable, and inducible genome editing in human pluripotent stem cells. Cell Stem Cell. 2014;15: 215–226.

72. DeKelver RC, Choi VM, Moehle EA, Paschon DE, Hockemeyer D, Meijsing SH, et al. Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. Genome Res. 2010;20: 1133–1142.