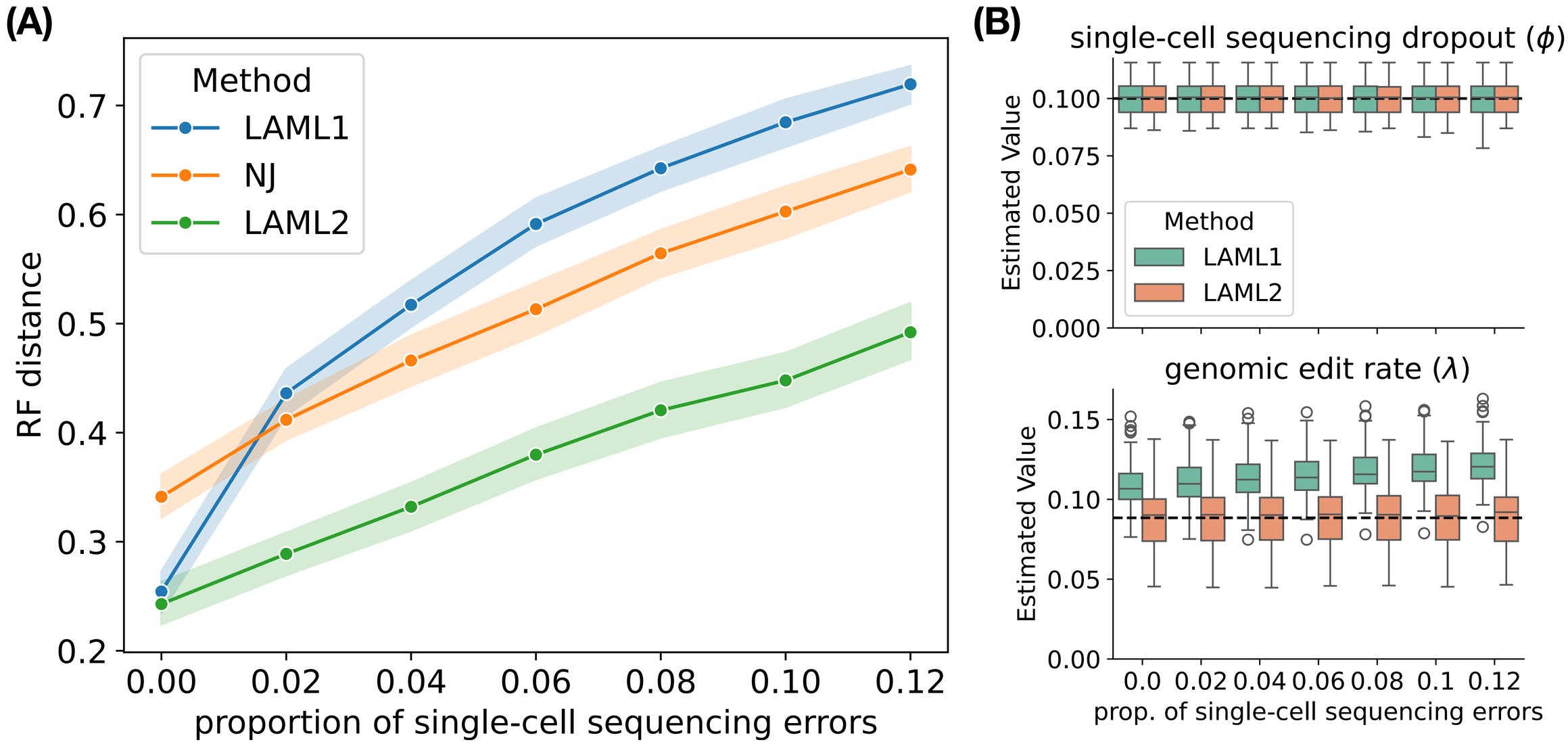

(A)
Method
LAML1
NJ
LAML2
RF distance
0.2
0.3
0.4
0.5
0.6
0.7
0.00
0.02
0.04
0.06
0.08
0.10
0.12
proportion of single-cell sequencing errors
(B)
single-cell sequencing dropout (ϕ)
Estimated Value
0.000
0.025
0.050
0.075
0.100
Method
LAML1
LAML2
genomic edit rate (λ)
Estimated Value
0.00
0.05
0.10
0.15
0.0
0.02
0.04
0.06
0.08
0.1
0.12
prop. of single-cell sequencing errors

Real data

TLS

Cell label	Cassette 1	Cassette 2	…
	[(Cassette state 0, count), (cassette state 1, count), …]	[(Cassette state 0, count), (cassette state 1, count), …]	

PMMC (counts) model
- dropout present (single-cell sequencing)
- Silencing present (continuously transcribed barcodes)

Take counts matrix as input

baseMemoir

Integration cassette index	Cell label	Site 1 Pr. Across 4 states	Site 2 Pr. Across 4 states	…
(1-66)	[1-73]	e.g. [0.95, 0.03, 0.01, 0.01]	…	

PMMN (noise) model
- dropout present (random FISH detection loss)
- No silencing (zombie transcription avoids silencing problems)

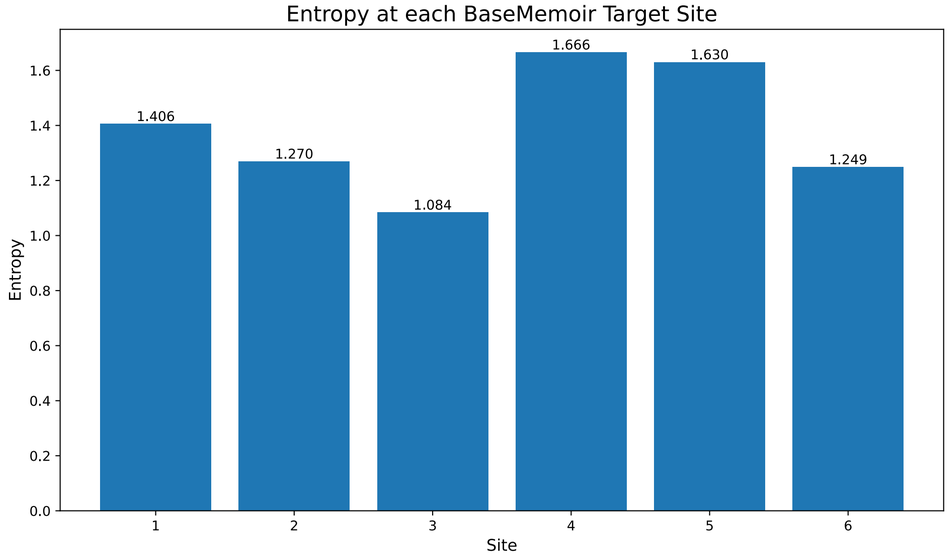
Take character matrix (max pr.) and emission matrix pr. of observing state C from state X at each site as inputs

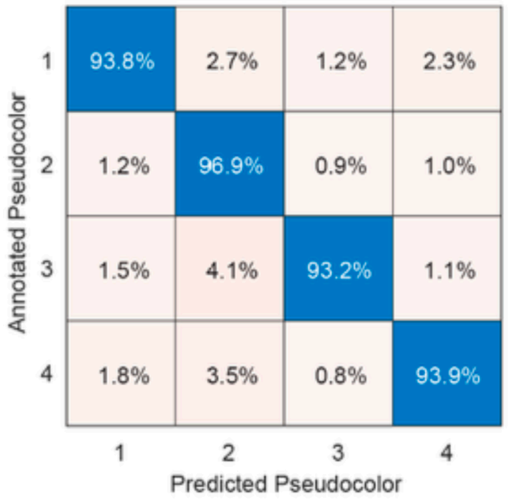
Current emission matrix is shared across all sites and cells, but this is trivial to change.

BaseMemoir

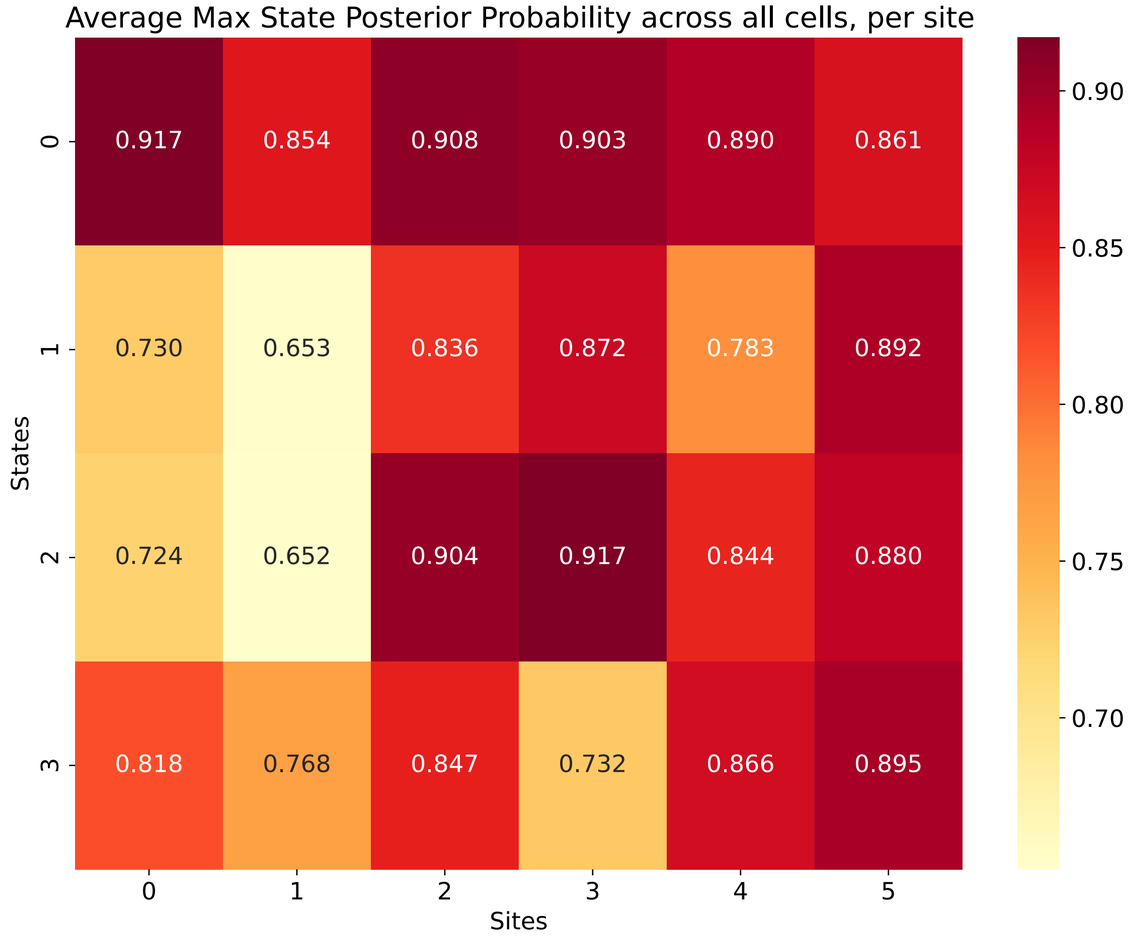
In order of least to most work

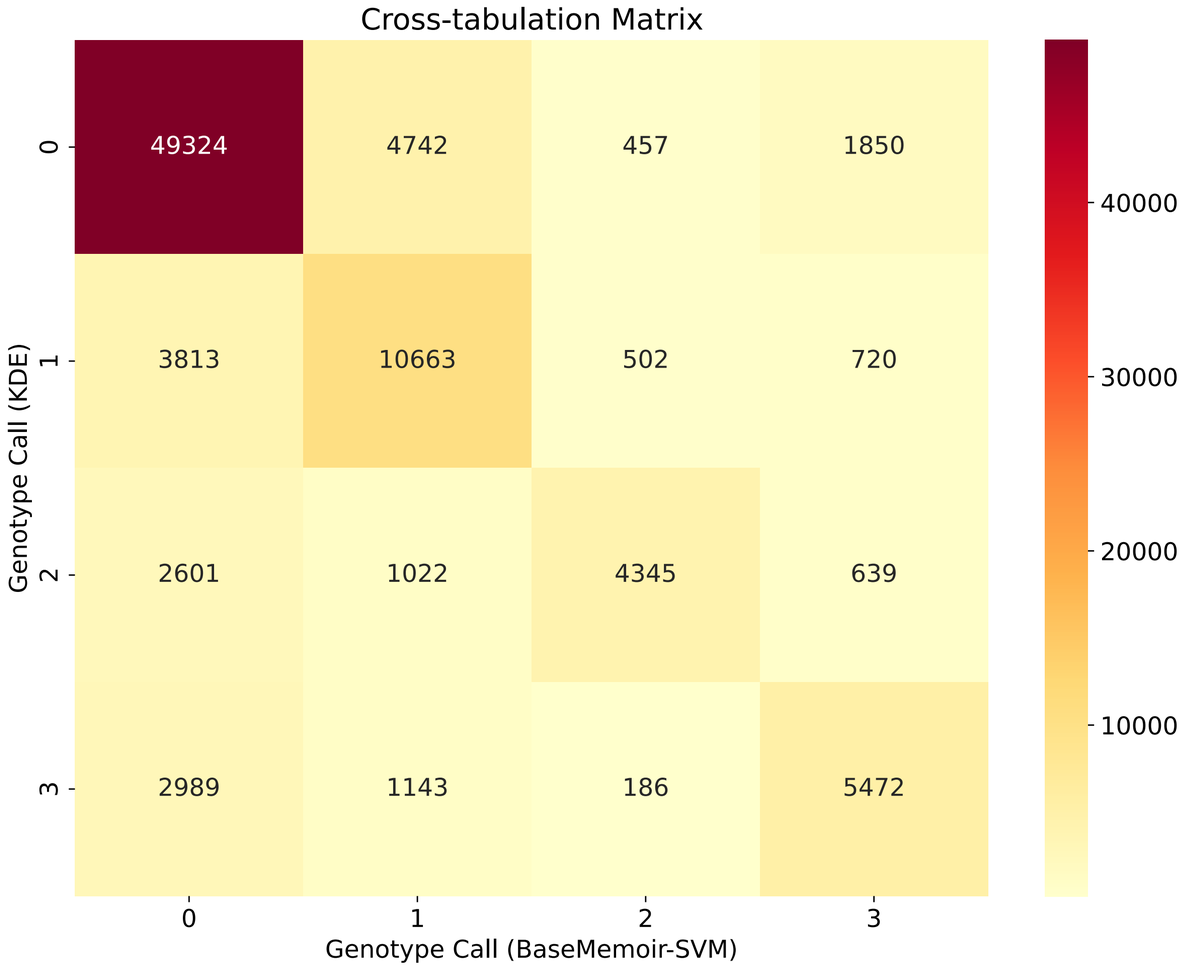
- Normalized posterior probability of observations (cells by sites by states)
- Labeled data: dots by intensities (in each of 16 hybridization rounds) and corresponding manual labels (into 1 of 4 states) - 9598 dots
- SVM classification probability for each unfiltered dot into 1 of 4 states
- 197 cells split amongst 7 colonies





Supp fig 2 (baseMemoir)


Average Max State Posterior Probability across all cells, per site
States
Sites
0
1
2
3
4
5
0.917
0.854
0.908
0.903
0.890
0.861
0.730
0.653
0.836
0.872
0.783
0.892
0.724
0.652
0.904
0.917
0.844
0.880
0.818
0.768
0.847
0.732
0.866
0.895
0.90
0.85
0.80
0.75
0.70


Cross-tabulation Matrix
Genotype Call (KDE)
Genotype Call (BaseMemoir-SVM)
0
1
2
3
49324
4742
457
1850
3813
10663
502
720
2601
1022
4345
639
2989
1143
186
5472
40000
30000
20000
10000

(A)

Fill in a multi-progenitor tree

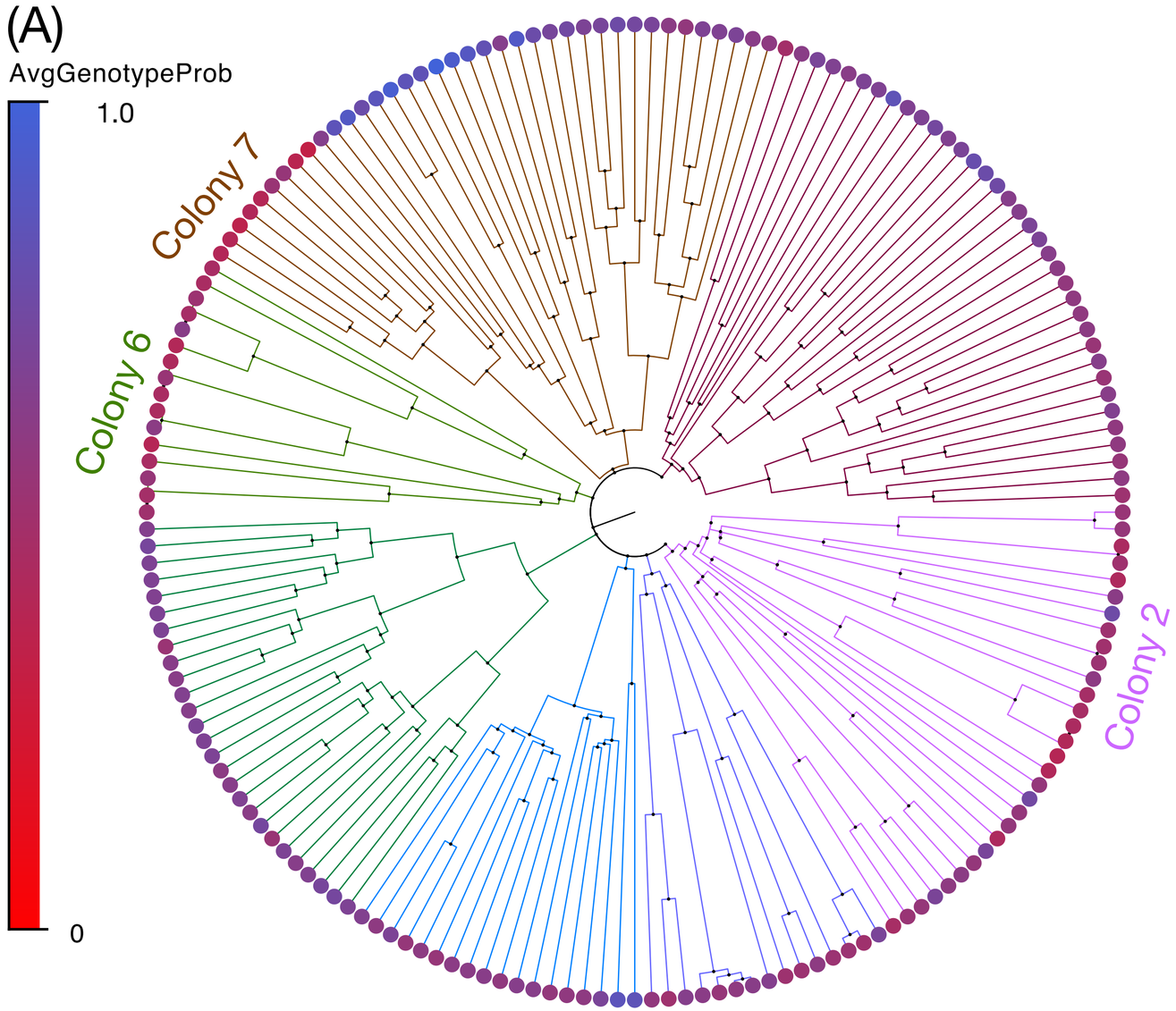
(B)

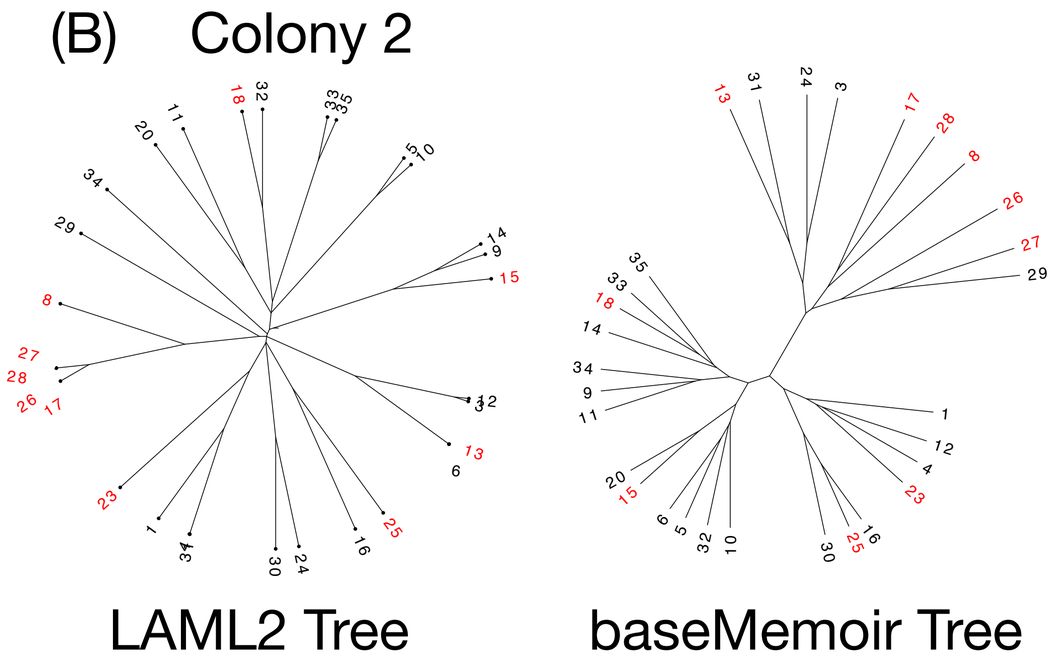
A single colony LAML2 tree, with leaf % missing marked

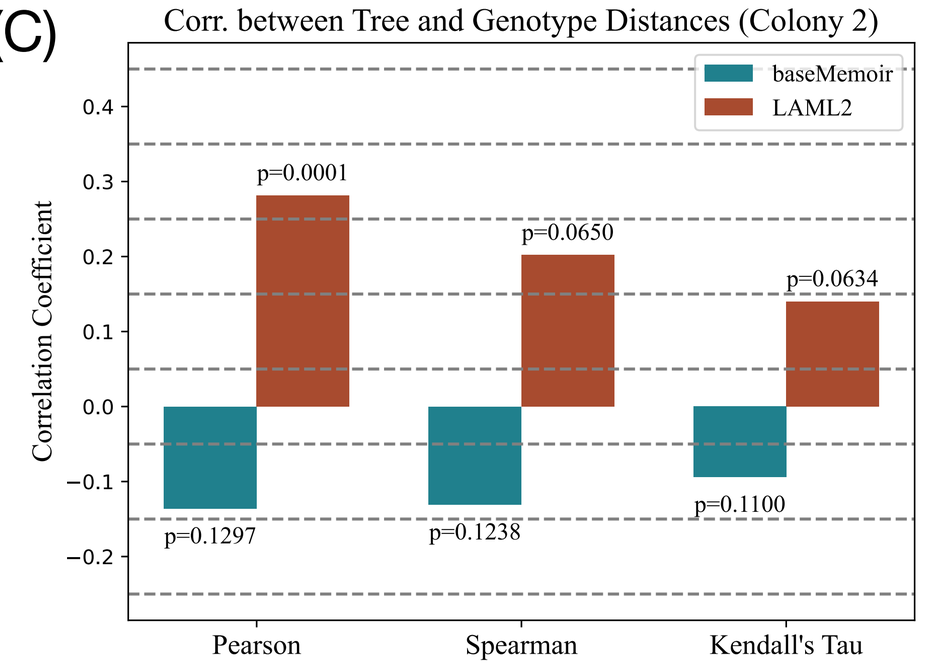
(C)

Show the concordance (Mantel test) between baseMemoir tree distance and genotype calls, and LAML tree distance and genotype calls

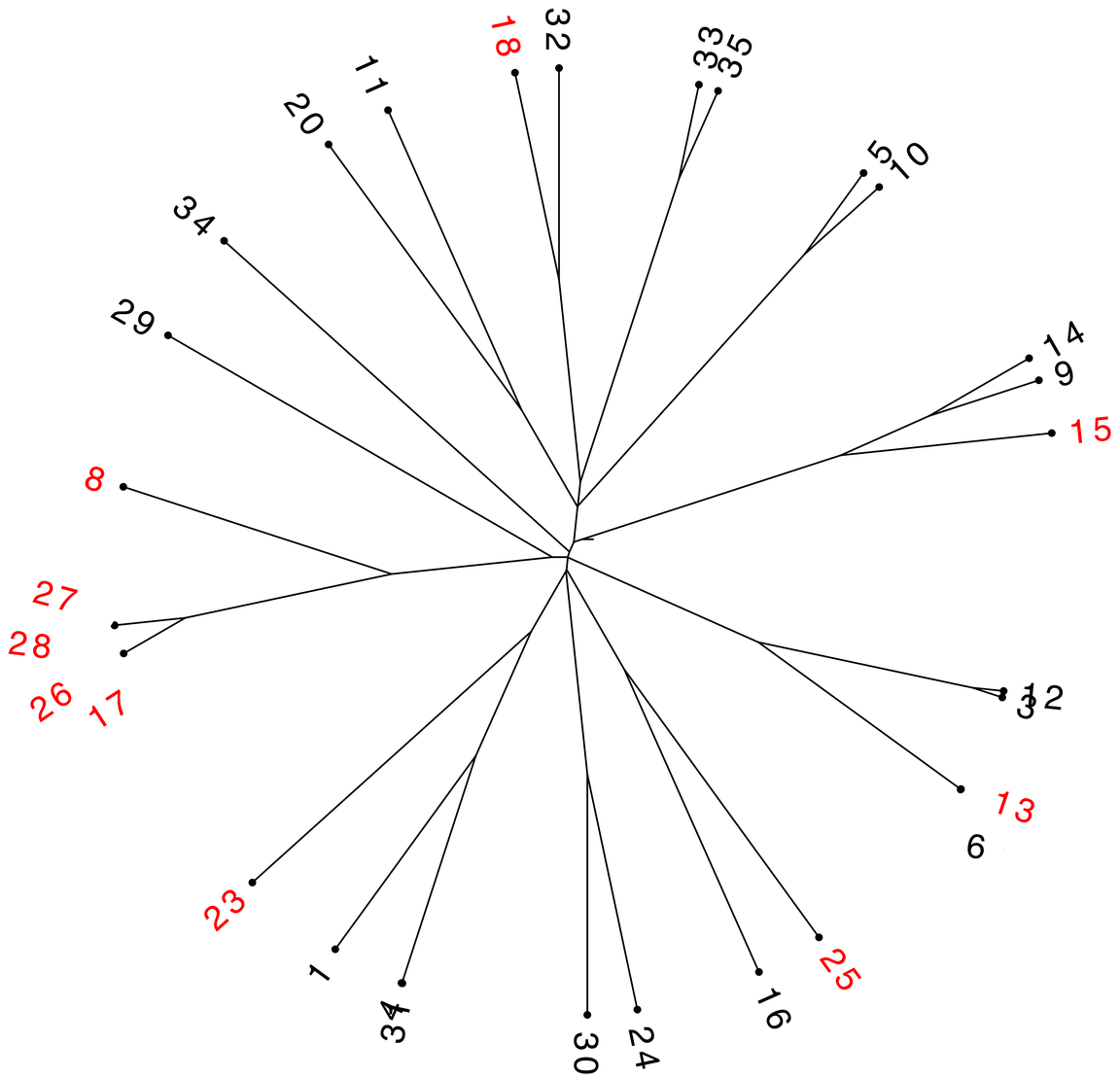
Single leaf imputed, show that posterior probabilities were low for these


(A)
AvgGenotypeProb
1.0
0
Colony 7
Colony 6
Colony 2

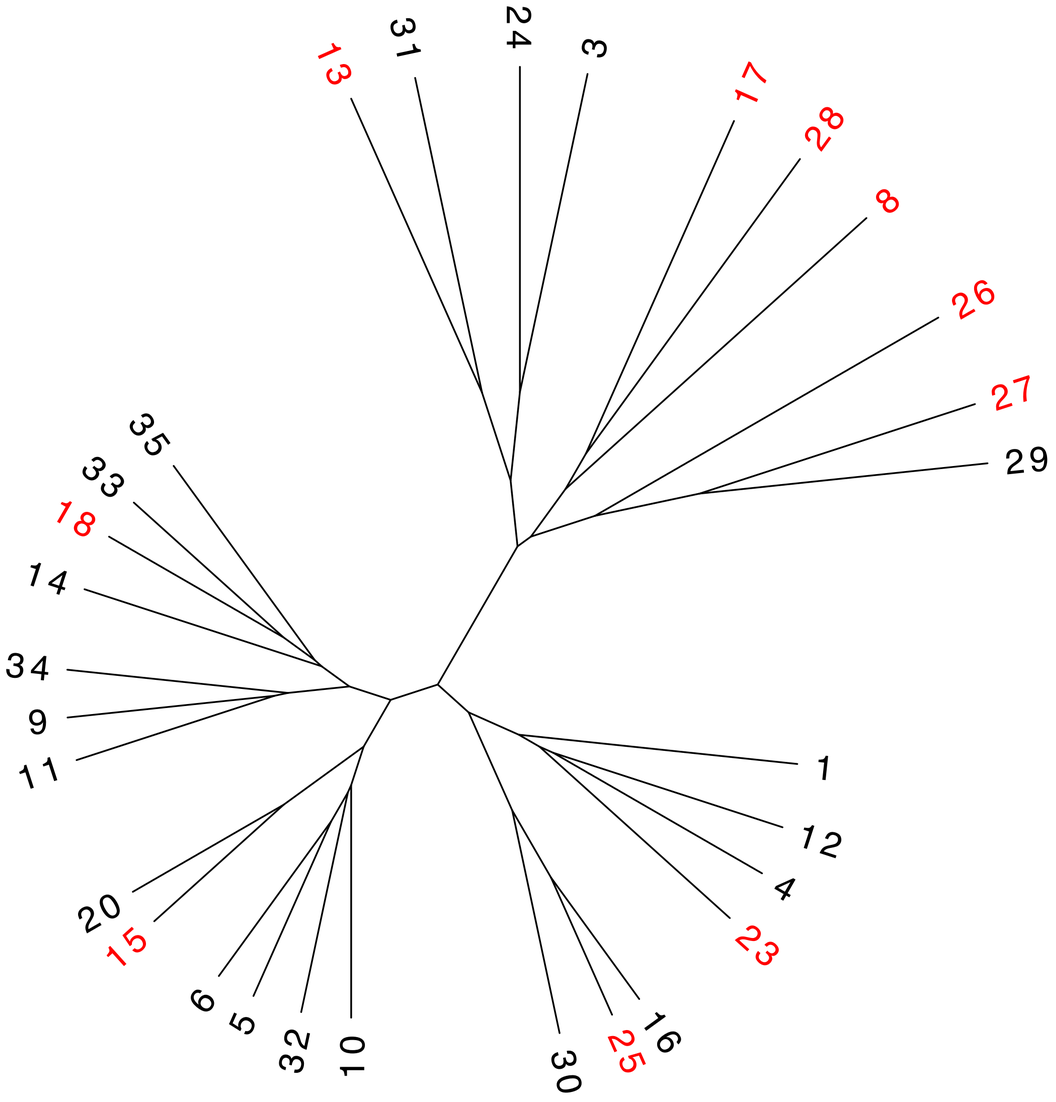

(B) Colony 2
LAML2 Tree
baseMemoir Tree


(C)
Corr. between Tree and Genotype Distances (Colony 2)
baseMemoir
LAML2
Correlation Coefficient
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
p=0.0001
p=0.0650
p=0.0634
p=0.1297
p=0.1238
p=0.1100
Pearson
Spearman
Kendall's Tau

(A) LAML2 Tree - Colony 2



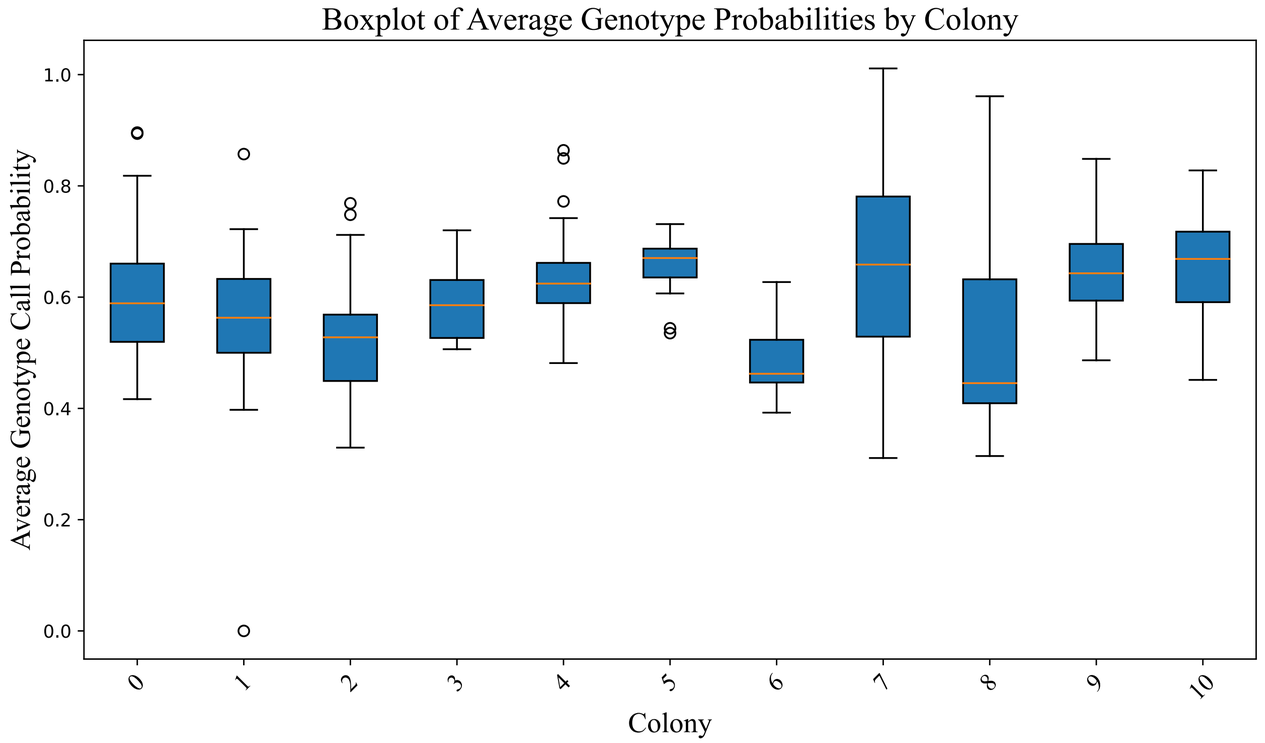
(B) Published baseMEMOIR Tree - Colony 2



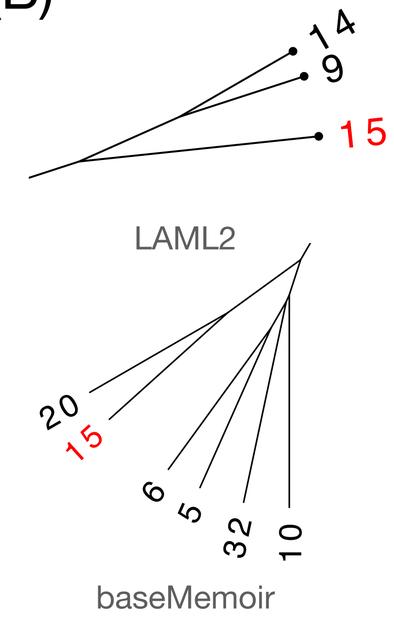
Marked in red are cells with 40-50% genotype call probability

colonies	% missing data	Avg. genotype call Pr (baseMemoir)	Lowest mean genotype Pr	Notes on Position
Position 0	71.4%	0.60	Cell 26: 0.41	Dropped
Position 1	71.0%	0.549	Cell 23: 0.0	Dropped
Position 2	30.0%	0.528	Cell 22: 0.329	-
Position 3	21.2%	0.588	Cell 11: 0.506	-
Position 4	33.5%	0.639	Cell 20: 0.481	-
Position 5	24.1%	0.661	Cell 3: 0.535	-
Position 6	23.6%	0.483	Cell 15: 0.392	-
Position 7	44.0%	0.651	Cell 38: 0.311	Merged into 7/8
Position 8	41.2%	0.529	Cell 36: 0.314	Merged into 7/8
Position 9	34.8%	0.653	Cell 25: 0.486	Merged into 9/10
Position 10	42.6%	0.655	Cell 29: 0.451	Merged into 9/10

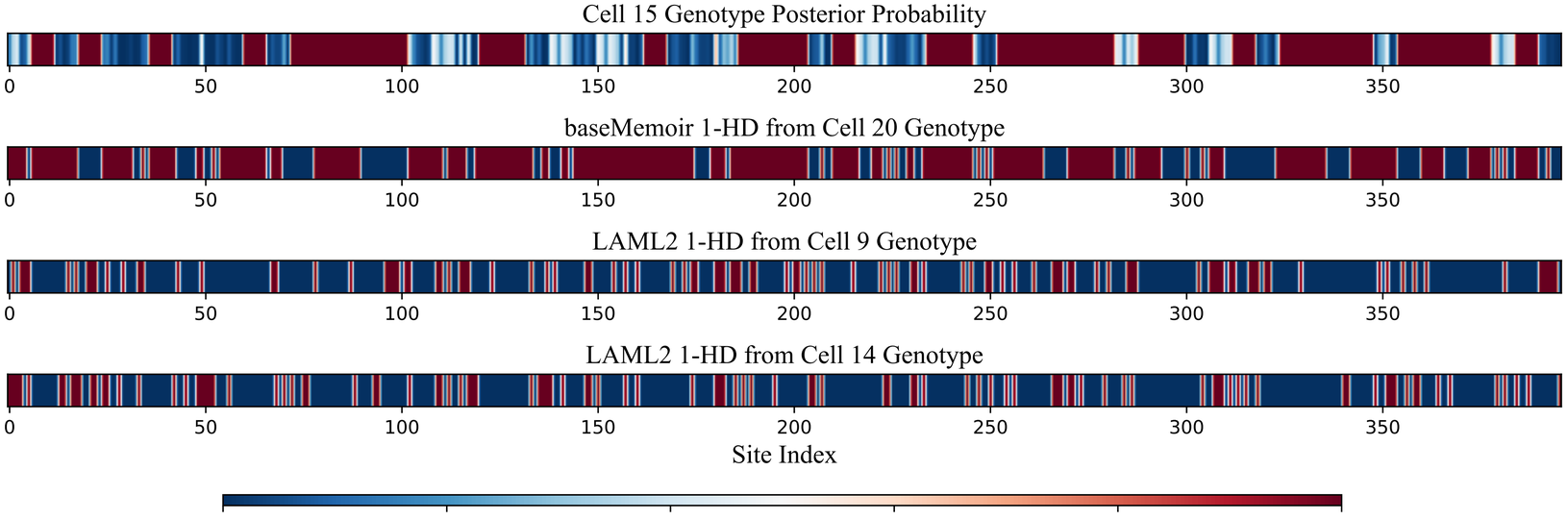
(A)


Boxplot of Average Genotype Probabilities by Colony
Average Genotype Call Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
Colony

(B)


14
9
15
LAML2
20
15
6
5
32
10
baseMemoir

(C)


Cell 15 Genotype Posterior Probability
0
50
100
150
200
250
300
350
baseMemoir 1-HD from Cell 20 Genotype
0.671 hamming distance
LAML2 1-HD from Cell 9 Genotype
0.268 hamming distance
LAML2 1-HD from Cell 14 Genotype
0.265 hamming distance
Site Index

Single leaf imputed, show that posterior probabilities were low for these

Compute the average Hamming distance between siblings' genotype calls in tree

Input

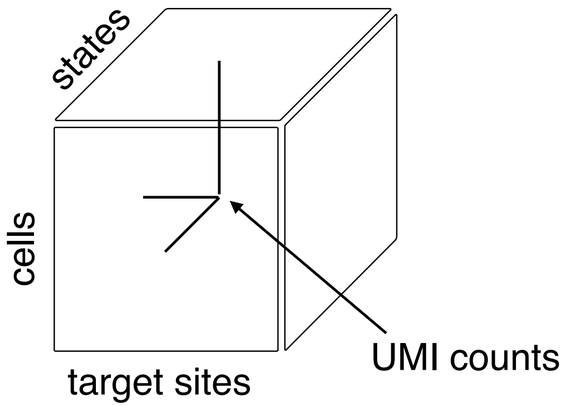
Target sites

1	0	4
3	?	3
3	1	?
?	?	1
?	?	1
2	?	1
2	2	2

cells

character matrix $\mathbf{D}$

OR



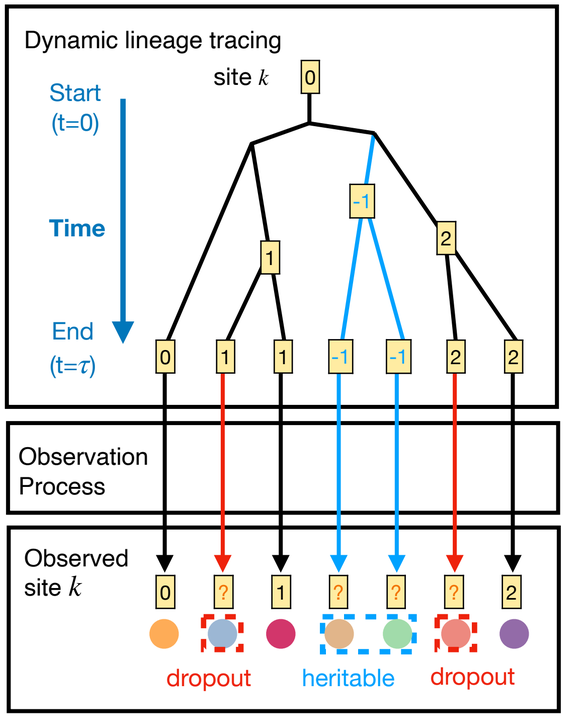
UMI counts matrix $\mathbf{C}$

LAML

maximum likelihood tree inference

$$\max_{T,\Theta} \log L(T, \Theta; \mathbf{D})$$

calculate $L(T, \Theta; \mathbf{D})$ under the PMM



Expand the single-cell sequencing part

Output

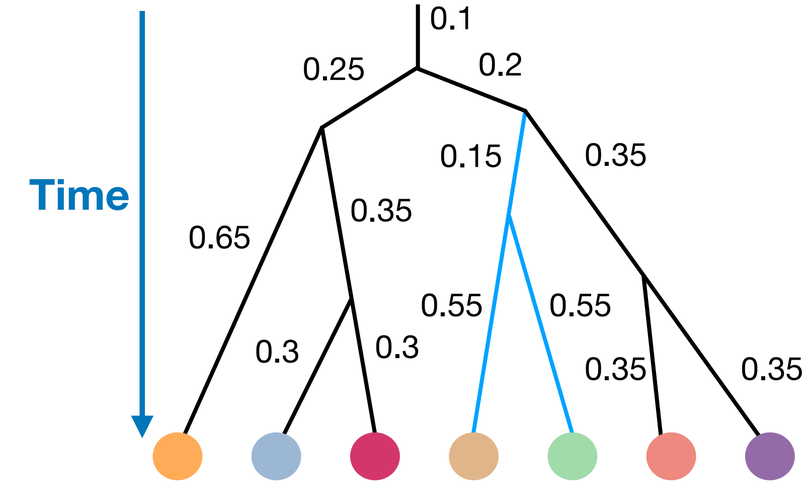
(1) Imputed character matrix

Target sites

1	0	4
3	1	3
3	1	3
0	0	1
0	0	1
2	2	1
2	2	2

cells

(2) Time-resolved lineage tree



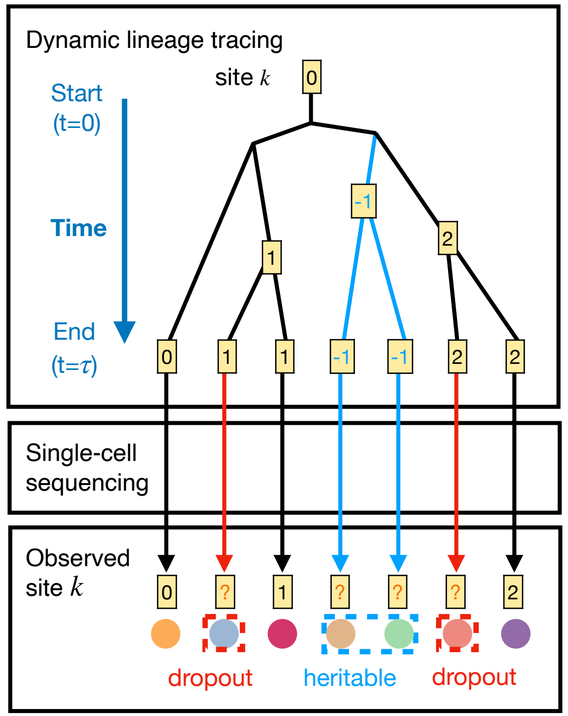
(3) Editing rate $\hat{\lambda}$, heritable missing rate $\hat{\nu}$, dropout probability $\hat{\phi}$, and observation accuracy $\hat{\rho}$

LAML

maximum likelihood tree inference

$$\max_{T,\Theta} \log L(T, \Theta; \mathbf{D})$$

calculate $L(T, \Theta; \mathbf{D})$ under the PMM

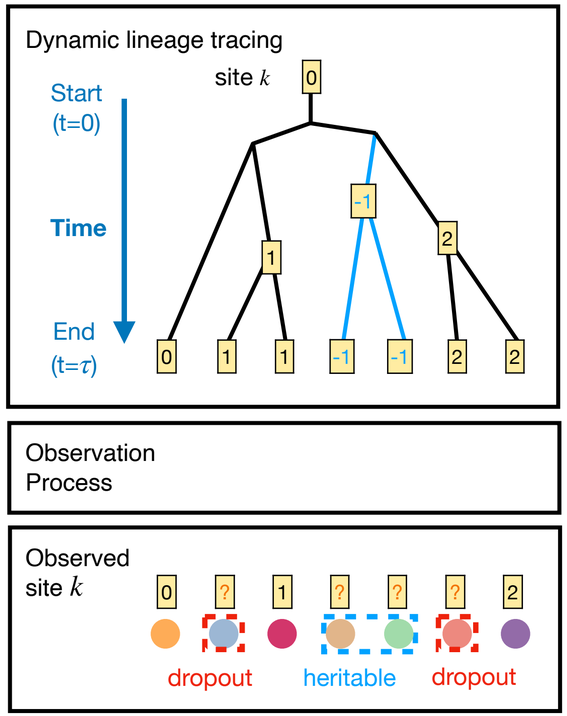


LAML2

maximum likelihood tree inference

$$\max_{T,\Theta} \log L(T, \Theta; \mathbf{D})$$

calculate $L(T, \Theta; \mathbf{D})$ under the PMM



Expand the single-cell sequencing part

LAML2

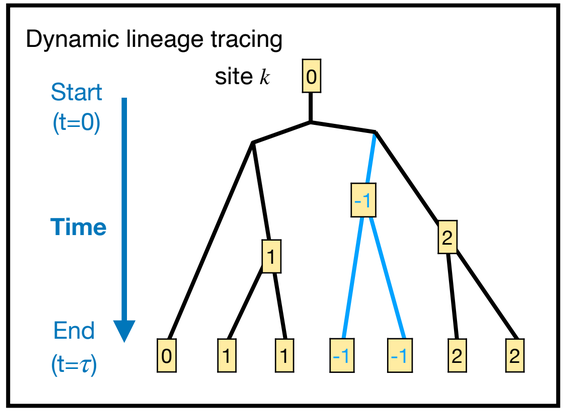
Sequencing errors

Imaging errors

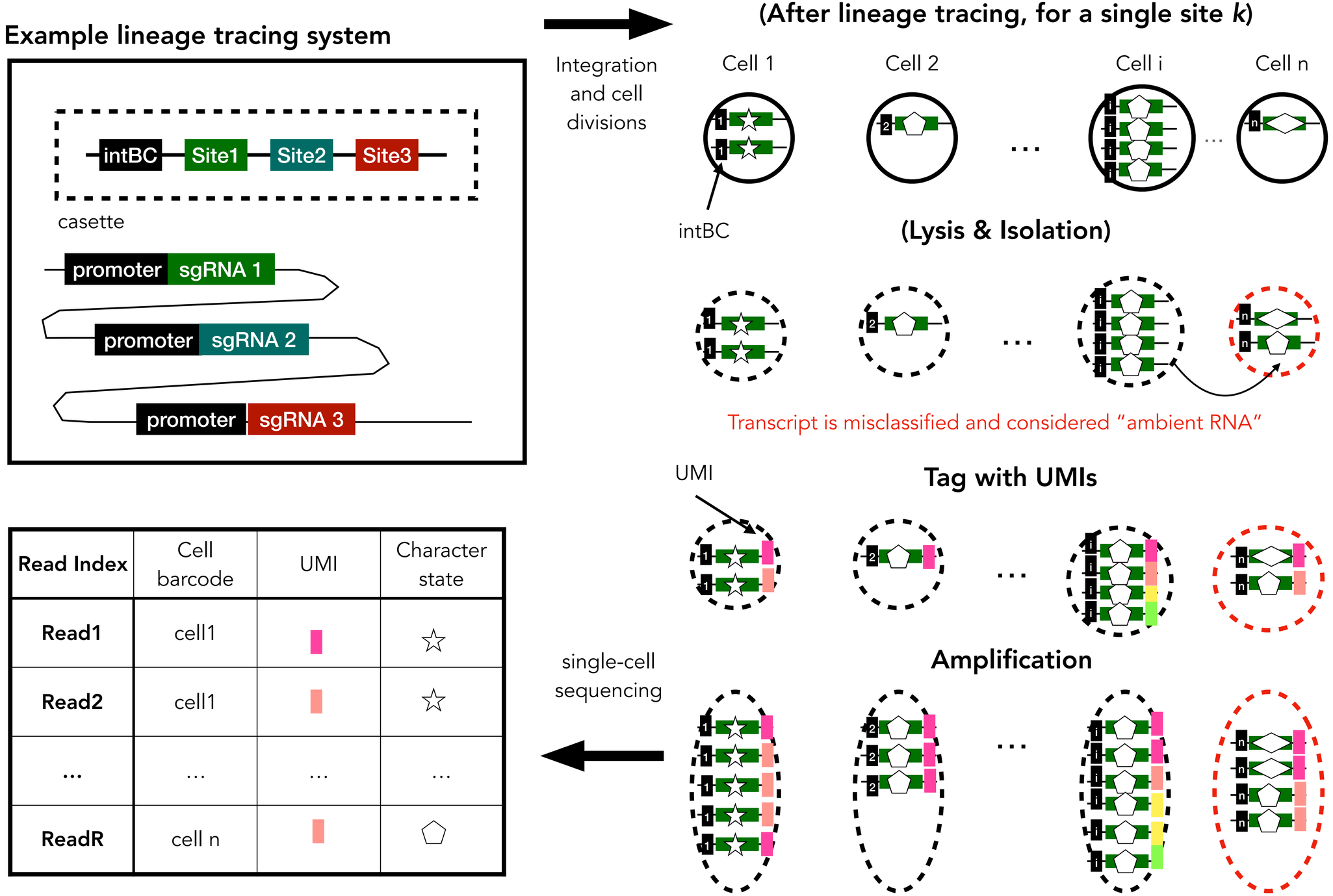
maximum likelihood tree inference

$$\max_{T,\boldsymbol{\Theta}} \log L(T, \boldsymbol{\Theta}; \mathbf{D})$$

calculate $L(T, \boldsymbol{\Theta}; \mathbf{D})$ under the PMM



Expand the single-cell sequencing part



Read Index	Cell barcode	UMI	Character state
Read1	cell1	▮	☆
Read2	cell1	▮	☆
...	...	...	...
ReadR	cell n	▮	⬠

Example image-based lineage tracing system

Relevant figures

Figures

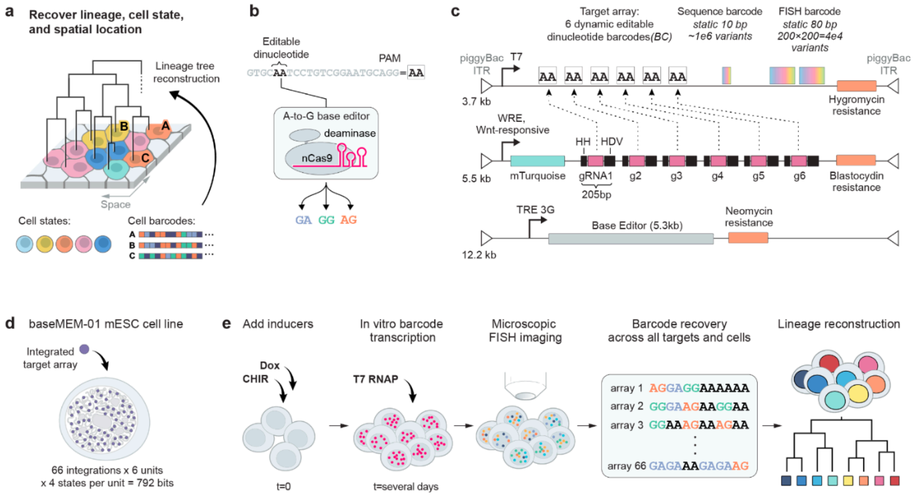


Figure 1: Multiplexed, genomically dispersed, editable barcodes enable detailed recording of lineages over many generations with in situ readout.

(a) Detailed lineage trees can be measured alongside transcriptional cell states while maintaining spatial context through phylogenetic barcoding. **(b)** Predicted stochastic editing of AA dinucleotides results in one of three terminal outcomes. **(c)** An inducible barcode editing system can be integrated into cells at high copy number via piggyBac transposase. Target arrays (top) contain 6 AA dinucleotides flanked by unique protospacer sequences as well as sequencing and imaging-readable static barcodes which serve to uniquely mark different genomic integrations of the array. Editing is induced by expression of guide RNAs (middle), controlled by a Wnt-responsive element, and base editor (bottom), controlled by the TRE3G tet-on promoter. **(d)** We engineered a monoclonal mESC cell line containing 66 uniquely labeled target array copies (396 editable dinucleotides, or 792 bits of information) alongside the inducible editing machinery. **(e)** This cell line enables genomic lineage recording and recovery through FISH imaging and phylogenetic tree inference.

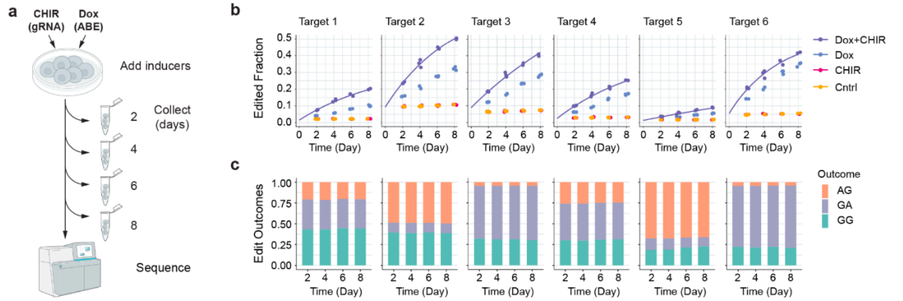


Figure 2: Dinucleotide targets accumulate edits over time in engineered mESCs.

(a) Next generation amplicon sequencing quantified editing over time after induction of gRNAs and ABE. **(b)** All targets edited over time in the presence of the two inducers together, although at distinct rates **(b, purple)**. Dox induction alone drives editing at a slower rate **(b, blue)**. In the absence of dox, editing does not proceed at an appreciable rate **(b, red and gold)**. Three biological replicates were collected for each time point. The solid purple line shows the fit for a probabilistic model of editing over time (**Methods**). **(c)** Each target has a unique distribution of editing outcomes that remains constant as editing progresses.

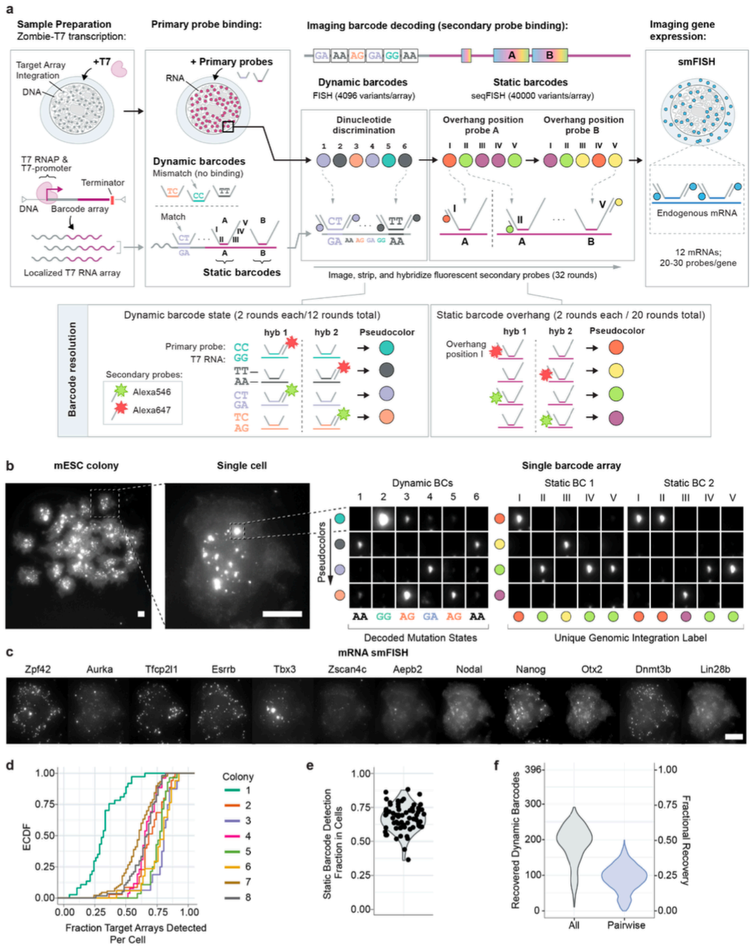
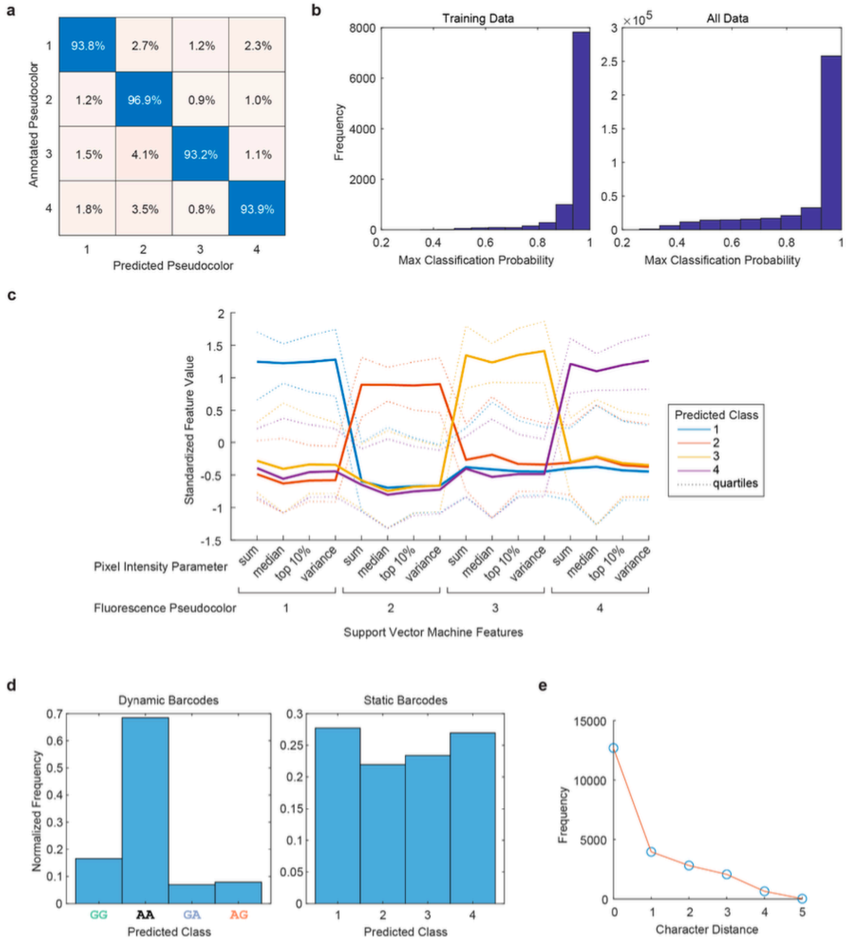


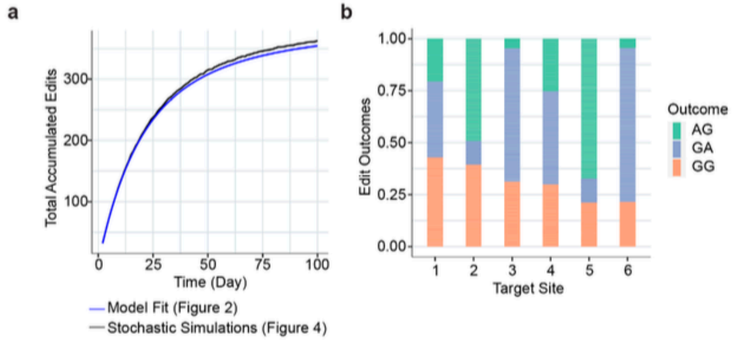
Figure 3: Multiple rounds of Zombie-FISH recover dynamic and static barcode states.

(a) Barcode states can be recovered across multiple rounds of microscopic imaging. Ectopic application of T7 polymerase generates localized RNA clusters. Primary DNA probes are bound to the dynamic and static barcodes as well as to endogenous transcripts, competing primary probes against each other for binding to the different possible dynamic barcode variants. Each primary probe has an overhang sequence allowing for binding of one or more fluorescently labeled secondary probes, which are hybridized, imaged, and stripped away sequentially to recover barcoding **(b)** and transcriptional **(c)** information. **(d)** Across 8 colonies, we recovered 50-80% of target arrays per cell. One colony had dramatically lower barcode recovery and was excluded from further analysis (**d**, colony 1). **(e)** Each unique target array is recovered in a similar fraction of cells. **(f)** We recovered approximately 200 dinucleotide dynamic targets with high confidence per cell, with around 100 of these measured jointly between any pair of cells. Scale bars are 20 μm.



Supplemental Figure 2: A support vector machine classifies barcode states based on fluorescence measurements.

(a) Manually annotated barcodes are correctly classified by a quadratic kernel support vector machine (SVM) approximately 94% of the time. **(b)** Classification probability estimates are very high within the training dataset **(b, left)**. Outside of the training sample, most classification probabilities are still high but with a subset of predictions that are less certain **(b, right)**. The support vector machine predicts classes based on 16 fluorescence measurements corresponding to each pseudocolor as defined in **Figure 3b**. **(c)** Each class is well separated based on these features. **(d)** After 3 days of editing induction, many dynamic barcodes are identified as class 2, corresponding to the unedited state **(d, left)**. Static barcode classifications are more evenly distributed, as anticipated **(d, right)**. **(e)** Static barcodes decoded by FISH typically perfectly match the 66 image readable barcode sequences identified by sequencing (**Supplemental Figure 1b**), although a fraction of barcodes are recovered with one or more character differences relative to their closest match.



Supplemental Figure 3: Stochastic simulations closely recapitulate the empirical editing process.

(a) We developed a stochastic editing simulator based on the Gillespie algorithm that closely recapitulates the average edit accumulation model developed in **Figure 2b** (**Methods**). **(b)** The simulated edit outcome distributions for each target site match the observed distributions from **Figure 2c**.

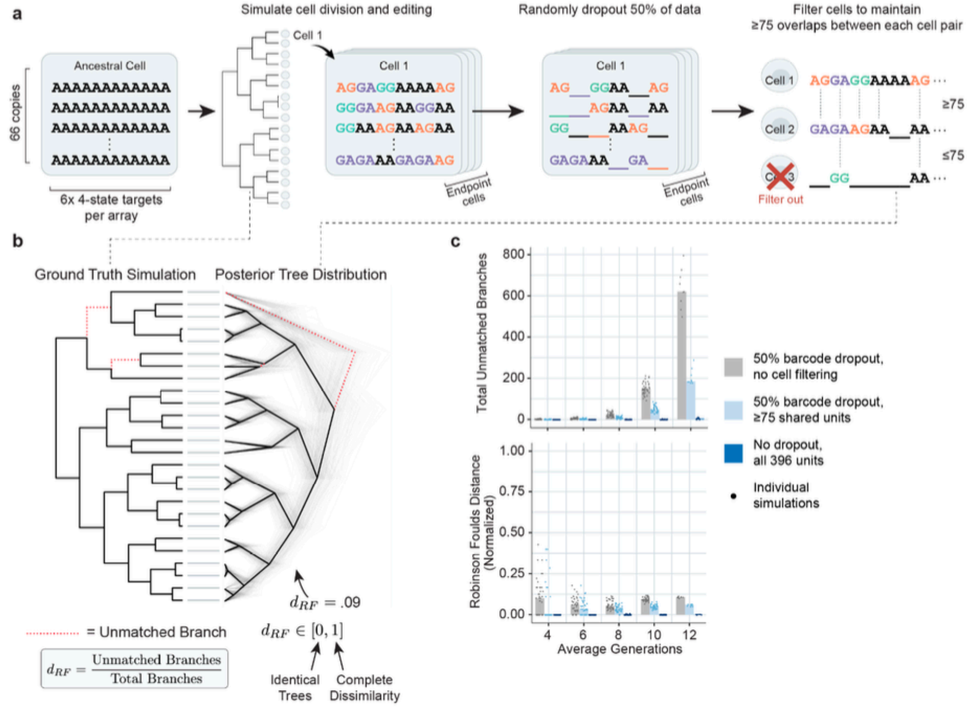


Figure 4: Lineage can be accurately reconstructed for at least 12 generations in simulation.

(a) To estimate the expected accuracy of reconstruction, we simulated cell division and stochastic editing starting with unedited barcodes, represented as sets of AA dinucleotides (**left**) over time to produce heterogeneous edit patterns. We then either retained all sequences or dropped 50% of the data to represent random FISH detection losses, and filtered out cells that had few barcode characters overlapping with those measured in other cells (**right**). **(b)** Based on these ground truth simulations, we reconstructed lineage relationships and computed the Robinson-Foulds distance between the ground truth input (**left**) and reconstructed output (**right**) trees. **(c)** Reconstruction accuracy was nearly perfect without barcode dropout (**dark blue dots**). With dropout, we observed ~ 10% error rates with tree depths up to 12 cell generations (**c, gray dots**). In the presence of dropout, filtering cells with few shared units moderately improved the reconstructed tree (**c, light blue dots**).

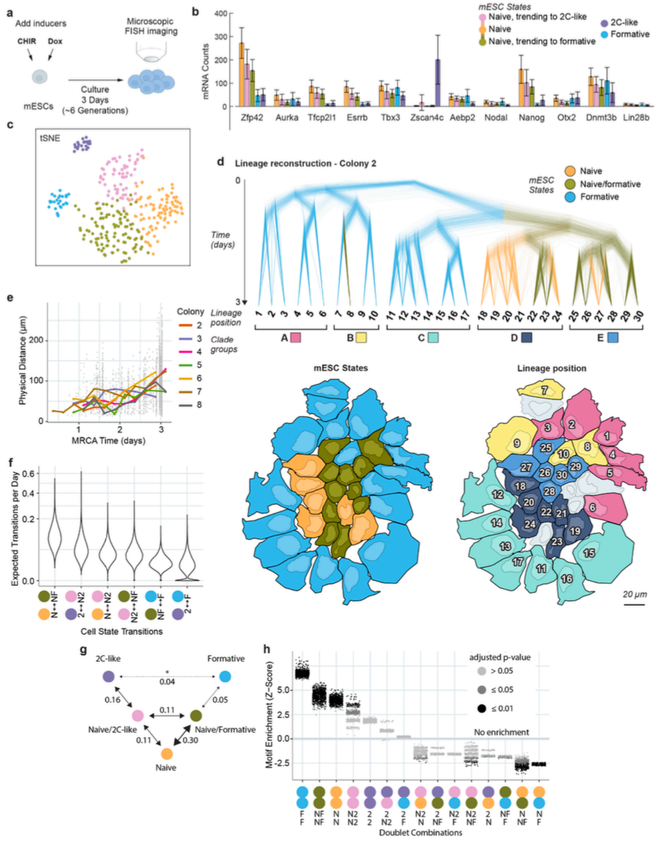


Figure 5: Joint measurements of lineage, gene expression, and spatial position reveal cell state transition dynamics.

(a) We recorded lineage relationships in mESC cells cultured in serum-LIF media over a 3 day period, inducing editing with 3 μM CHIR and 1 μM Dox. **(b,c)** Cells clustered into 5 states based on gene expression as measured by smFISH. Two clusters were well separated from the other groups while three clusters appeared continuously related and expressed different levels of key marker genes (**see Supplementary Figure 4**). **(d-g)** Lineage reconstruction infers topological lineage tree relationships, cell division timing, ancestral cell states, and transition rates between those states. Uncertainty in lineage tree measurements is visualized by overlaying trees sampled from the posterior distribution of trees generated by Markov chain Monte Carlo for each colony (**d, top; Supplemental Figure 5**). Cell states and clade groups from the lineage tree can be mapped to the spatial colony images to qualitatively inspect the relationships between cell state, lineage, and spatial location (**d, bottom; Supplemental Figure 5**). **(e)** Spatial distance is larger between cells with more distant common ancestors. **(f)** Several cell state transitions were inferred to have nonzero median values across all posterior samples. **(g)** These state transitions predict a restricted cell state transition graph. One transition (denoted by *) contained a high fraction of posterior samples with a transition rate of 0. Numbers indicate the median expected number of transitions per day for cells of the given type. **(h)** Several doublet motifs are significantly over or underrepresented across the lineage tree posterior samples. N: Naive; 2: 2C-like; F: Formative; N2: Naive, trending to 2C-like; NF: Naive, trending to formative; MRCA: Most recent common ancestor.