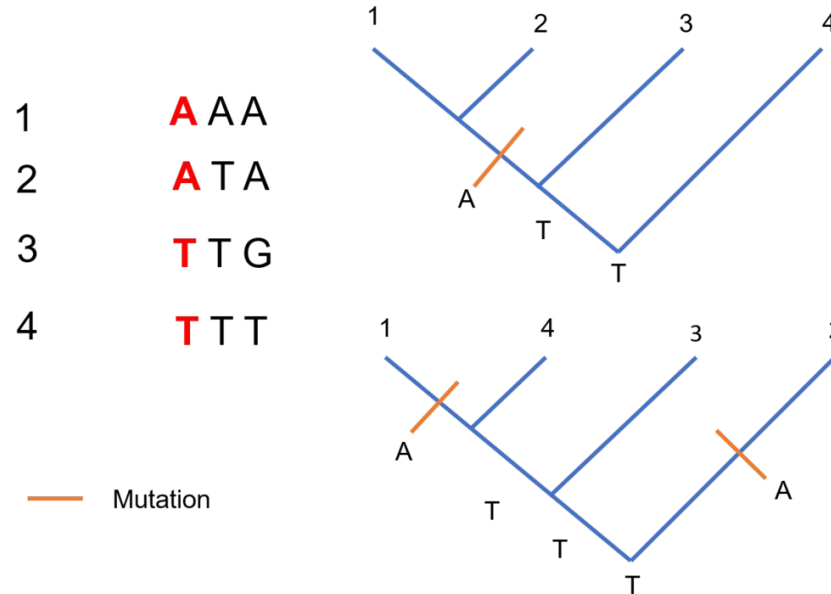# Hyperbolic point configurations for CRISPR lineage tracing
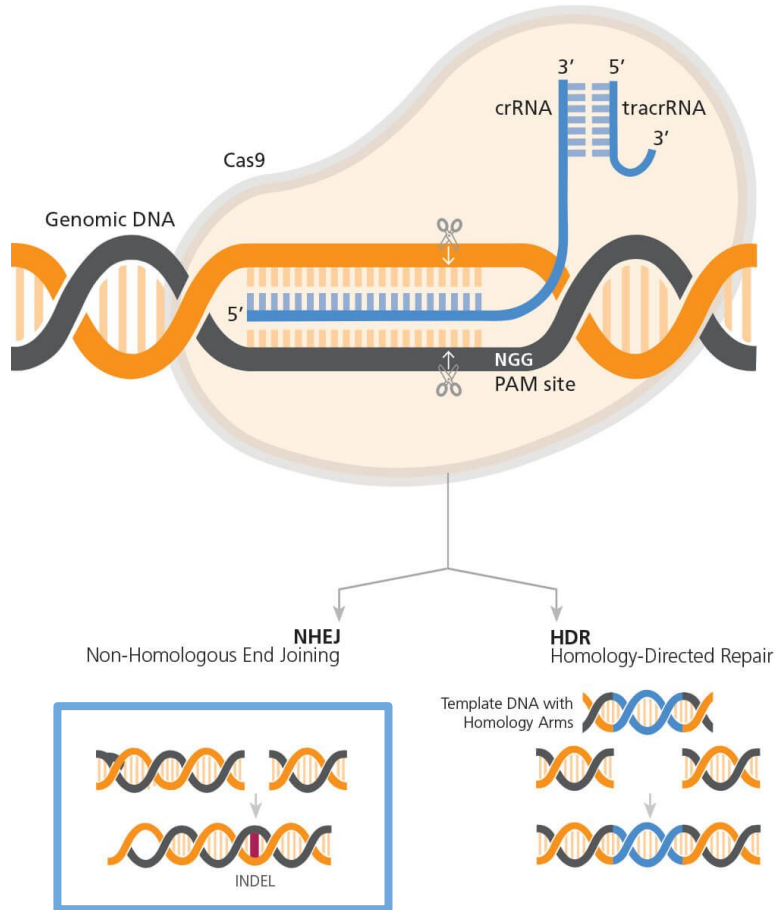
Anthony Ozerov

Supervised by Sitara Persad

Background - Genetic Approaches to Phylogeny

Inherited DNA alterations are used to infer phylogenetic relationships among samples.

CRISPR/CAS9

ATGAATCCGACCTAGA AGAGGGCACTTAAA
ATGAATCCGACCTAGATGAGGGCACTTAAA
ATGAATCCGA GCTAGATGAGGGCACTTAAA

ATGAATCCG TGGTTGATGCTCTAT CTTAAA
ATGAATCCG TGGTTGATGCTCTAT CTTAAA
ATGAATCCGACCTAGATGAGGGCACTTAAA

Rather than using random, small mutations, we can insert mutations at known, targeted sites.

Then we can read the target sites and try to reconstruct the lineage.

Why Lineage Tracing?

We want to track the fates of individual cells over the course of biological processes, e.g. development, cancer.

Cell and tissue atlases using scRNA help understand cellular complexity, but provide:

- Only a static snapshot
- No information about genetic relationships

Genetic approaches for phylogenies provide:

- No information on transcriptional states
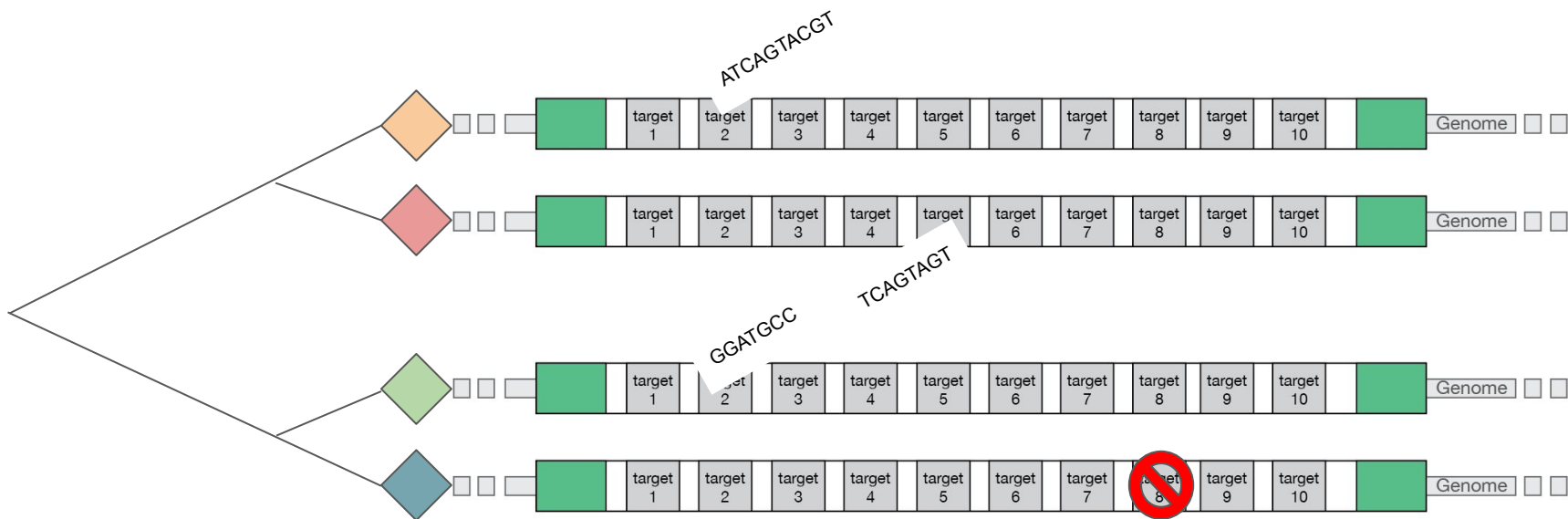- Resolution only as fine as the frequency of mutations

Ideally, lineage tracing provides:

- Dynamic view of cell trajectories
- Simultaneous recording of phylogenetic and transcriptional information at single-cell resolution

CRISPr/CAS9 Lineage Tracing System
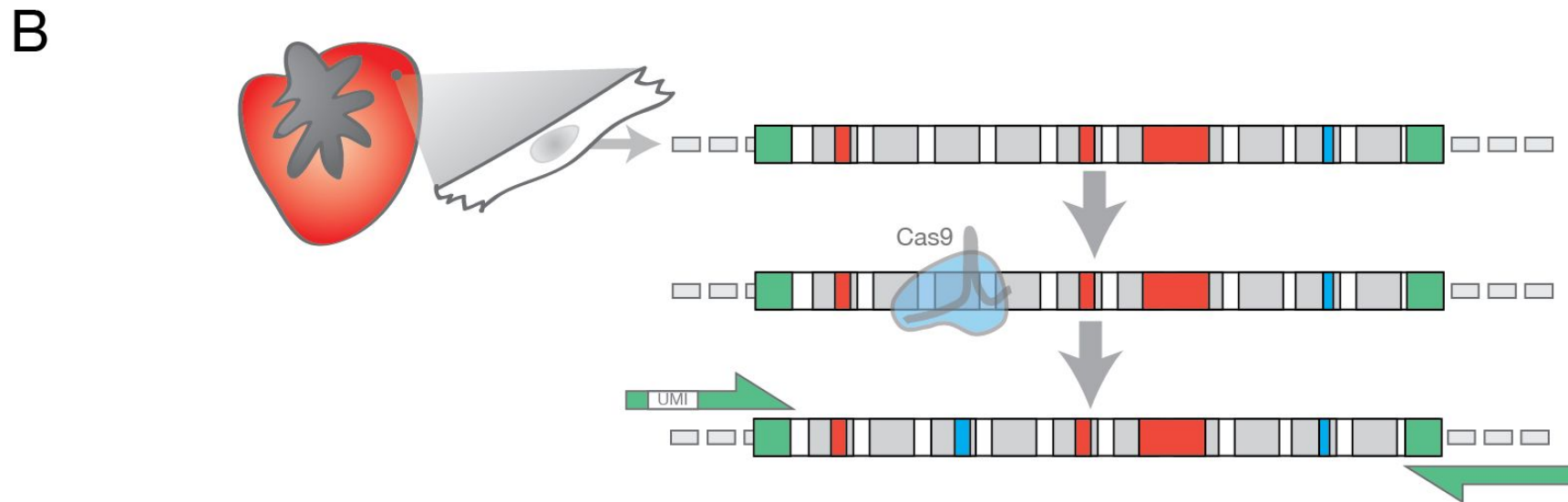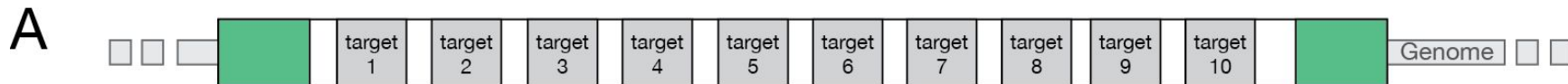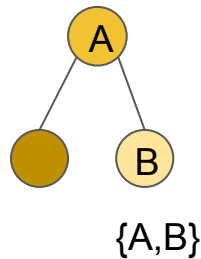
Character Matrix Representation

ATCAGTACGT

Genome

TCAGTAGT

GGATGCC

Genome

Targets

Cells

0	1	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	2	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	-	0	0

Current Tree Building Methods

Traditional Phylogenetics:

- Search tree topologies & branch lengths

Repeat until converged

A. Generate trees to explore ⟶ B. Compute likelihood of tree ⟶ C. Accept or reject proposed changes
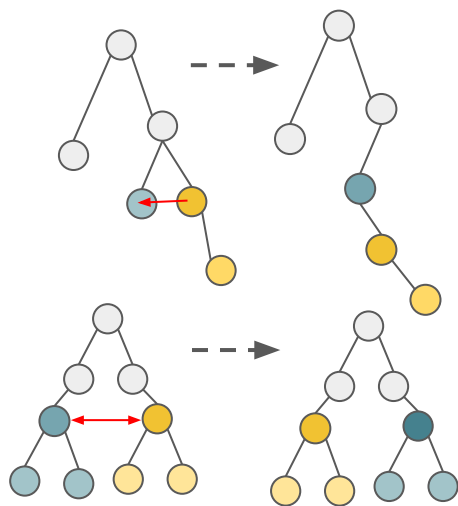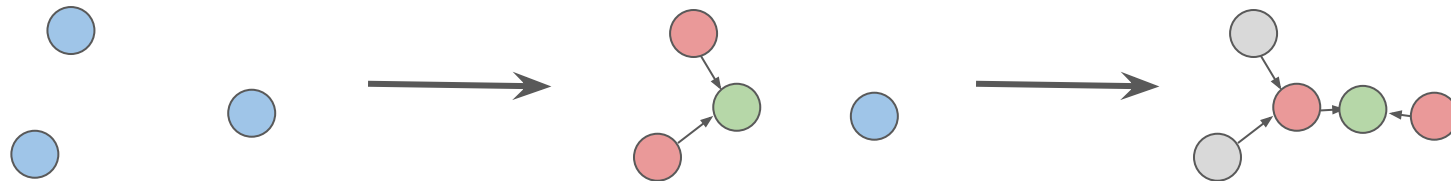

{A,B}

Current Tree Building Methods

Traditional Phylogenetics:

- Search tree topologies & branch lengths
- Neighbour Joining and Maximum Parsimony Approaches
    - Difficult to account for peculiarities in the data e.g. missing data
    - Doesn't incorporate knowledge about the biological process/ model by which data is generated

Points don't have to be embedded in some space. There just needs to be a distance metric.

Based on implementation specifics, two points are chosen and joined into a new point with different distances to the rest.
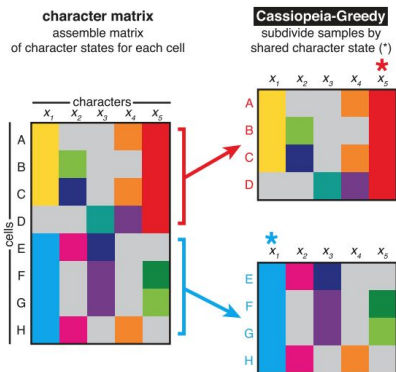
Current Tree Building Methods
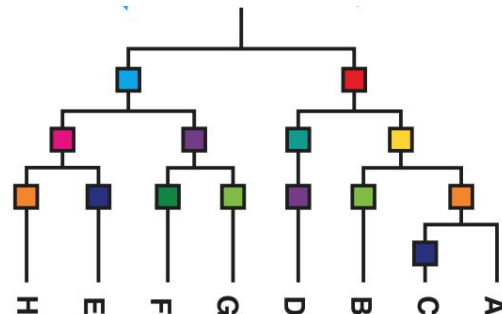
Traditional Phylogenetics:

- Search tree topologies & branch lengths
- Neighbour Joining and Maximum Parsimony Approaches

State of the Art in Lineage Tracing:

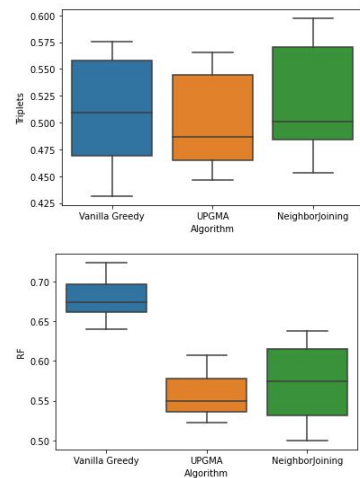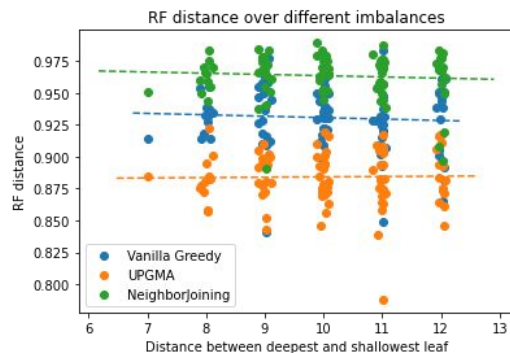- Greedy partitioning based on character frequencies

Recursively split samples by most prevalent character-state to get final tree.

Concerns

- Do the above methods work for highly imbalanced trees (*e.g*, possibly from a tumor)?
- Do the above methods work when the character matrix encodes CRISPR-Cas9 edits?

Evaluation of Cassiopeia methods in this setting:
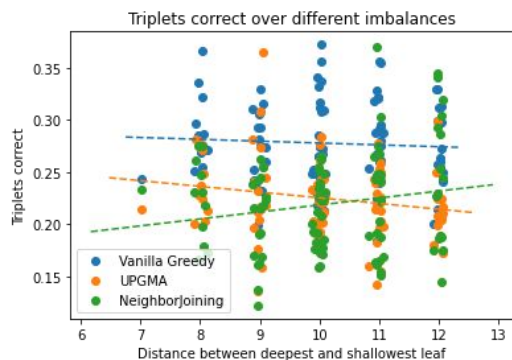
No evidence of worsening performance as imbalance increases, but performance is quite poor nonetheless (perhaps due to small number of sites)

Simulation Process
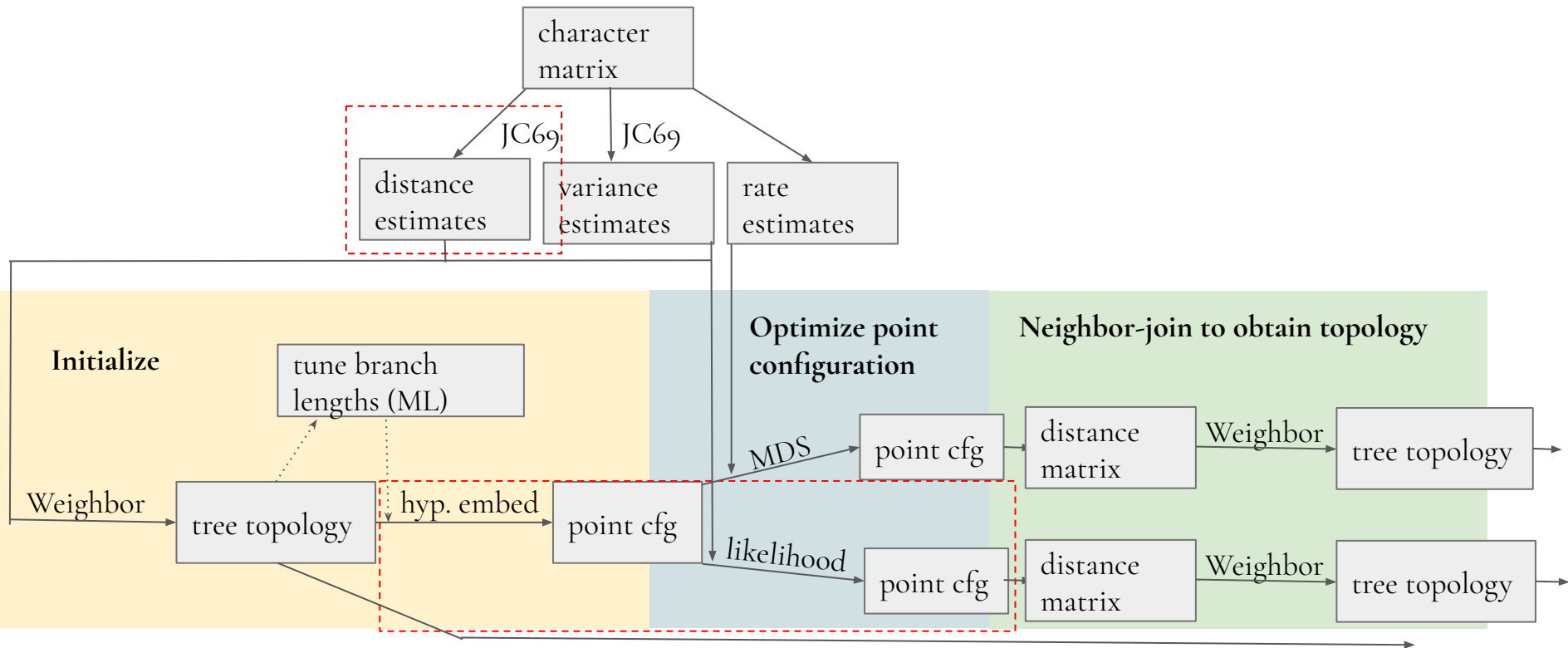
Simulate for 15 generations $\sim 2^{15} = 2$ million cells.

Start with a single cell

Draw probability of an indel: Bin(p)

Draw indel from empirical distribution

Cells replicate
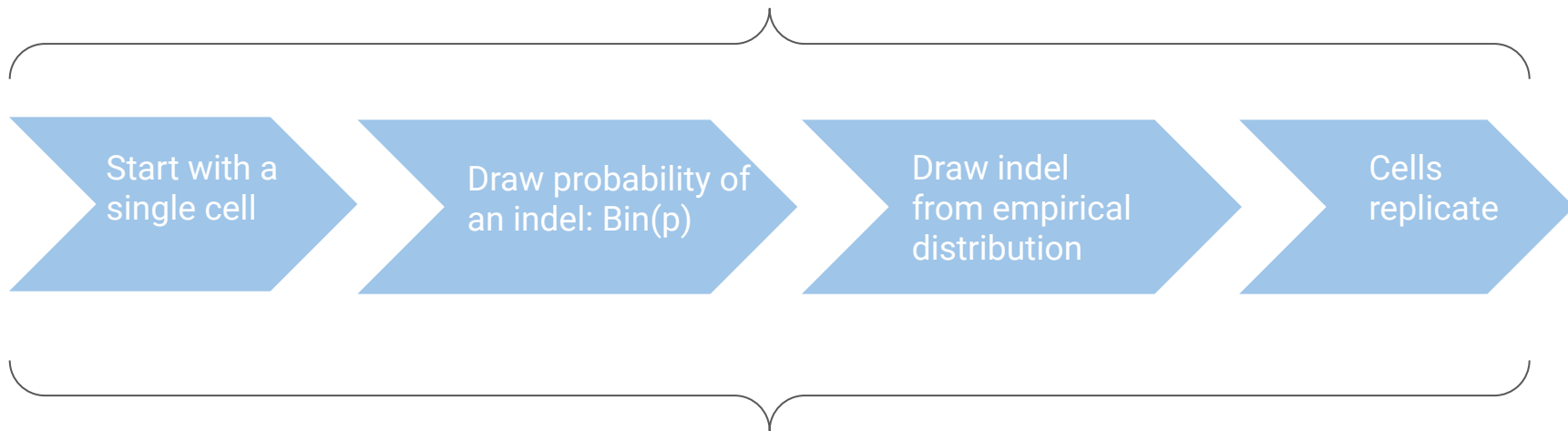
Subsample 1- 2000 cells

Technique overview [Wil21]

Likelihood method of optimizing point configuration [Wil21]

Jukes-Cantor model

$$P_{ab}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4t/3} =: P_{\text{diag}}, & \text{if } a = b \\ \frac{1}{4} - \frac{1}{4}e^{-4t/3} =: P_{\neg\text{diag}}, & \text{otherwise,} \end{cases}$$

= conditional probability of observing b at the site t time after observing a

t, the time, is effectively evolutionary distance

Likelihood function of distance between 2 points

$$\mathcal{L}(t) = \prod_{\text{sites } \sigma} \pi_{\sigma_i} P_{\sigma_i \sigma_j}(t),$$

Log-Likelihood function of distance between 2 points

$$\log \mathcal{L}(t) = \sum_{\text{sites } \sigma} \log P_{\sigma_i \sigma_j}(t) + C,$$

Log-Likelihood function of point configuration (with constant terms removed)

$$\mathbf{l(x)} = \frac{1}{L} \sum_{i \neq j} \sum_{\text{sites } \sigma} \log P_{\sigma_i \sigma_j} \left(\text{d}(x^i, x^j) \right) \cdot$$

Maximize this

What hyperbolic space changes

Smooth! No reference to tree topology (like in maximum-likelihood tree-search) or discrete topology moves needed. Wil21 further shows that maximizing **l(x)** roughly maximizes the log-likelihood of the tree

Hyperbolic space: rationale [Wil21]

4-point condition: finite metric space (X, d) defines tree with inter-leaf distances given by a pairwise distance matrix iff:

$d(x, w) + d(y, z) \leq \max\{d(x,y) + d(z,w), d(x,z) + d(y,w)\}$ for all w, x, y, z in X

Relaxed 4-point condition: metric space (X, d) is δ-hyperbolic if:

$d(x, w) + d(y, z) \leq \max\{d(x,y) + d(z,w), d(x,z) + d(y,w)\} + 2\delta$ for all w, x, y, z in X

Theorem: For any $m \geqslant 2$ and $\rho > 0$, there exists $\delta > 0$ such that $\mathbb{H}_\rho^m$ is δ-hyperbolic.

So by embedding the points in a hyperbolic space, and optimizing their configuration there, we can get arbitrarily close to satisfying the 4-point condition
(in Euclidean space, we might end up with a configuration that cannot reasonably describe a tree)

Implementation Problems
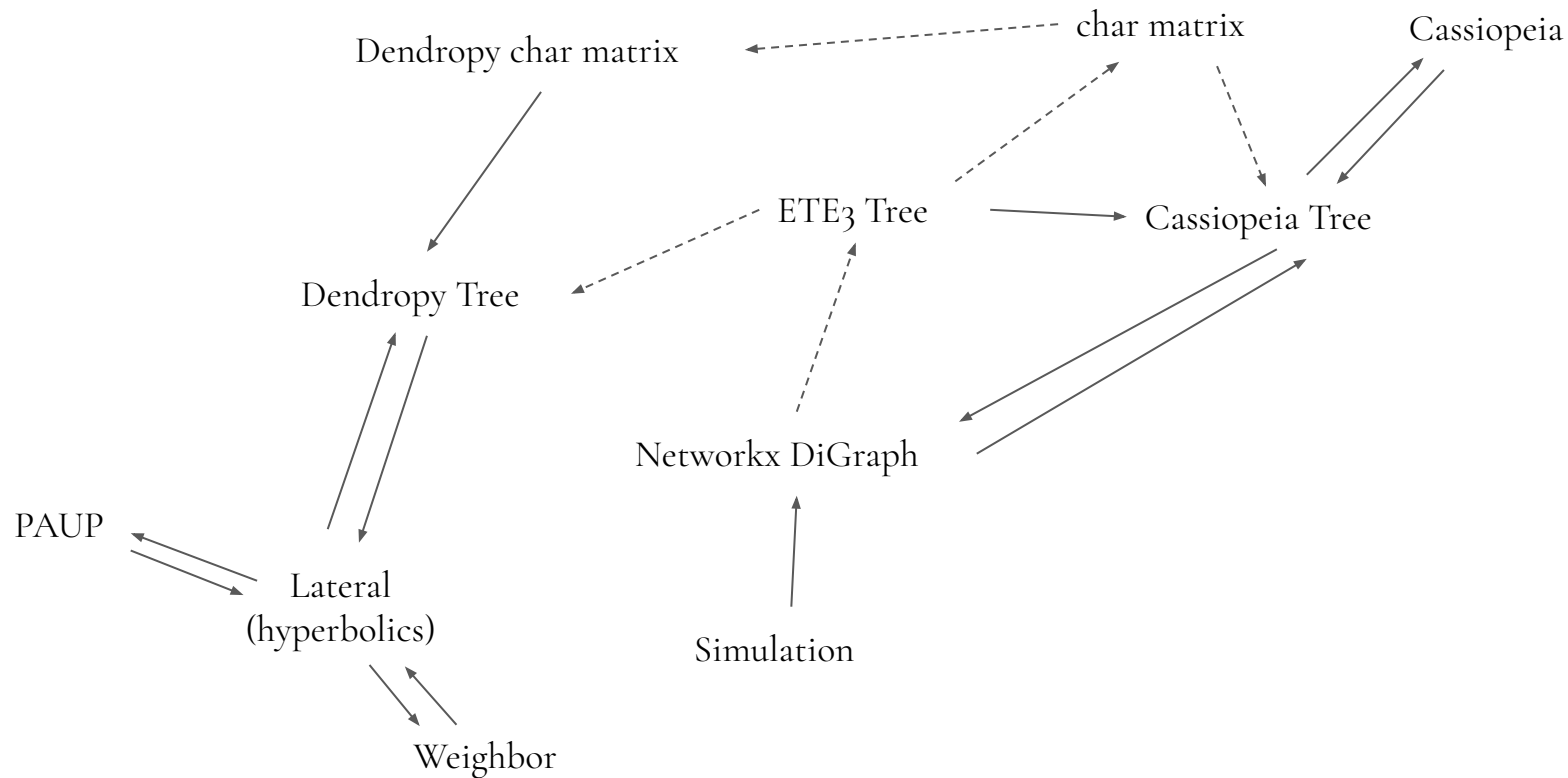
Implementation Problems

States must be encoded as
single characters

Dendropy char matrix

char matrix

Cassiopeia

ETE3 Tree

Cassiopeia Tree

Only accepts
small subset of
characters (58) as
states

Dendropy Tree

Networkx DiGraph

PAUP (branch
length tuning)

Lateral
(hyperbolics)

Simulation

Weighbor

misc. divide-by-zero and
overflow errors

Only accepts 4-state encoding,
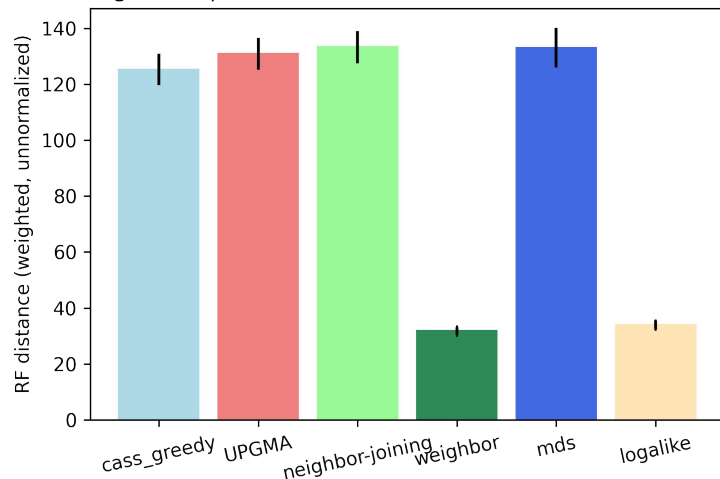does not accept missing values
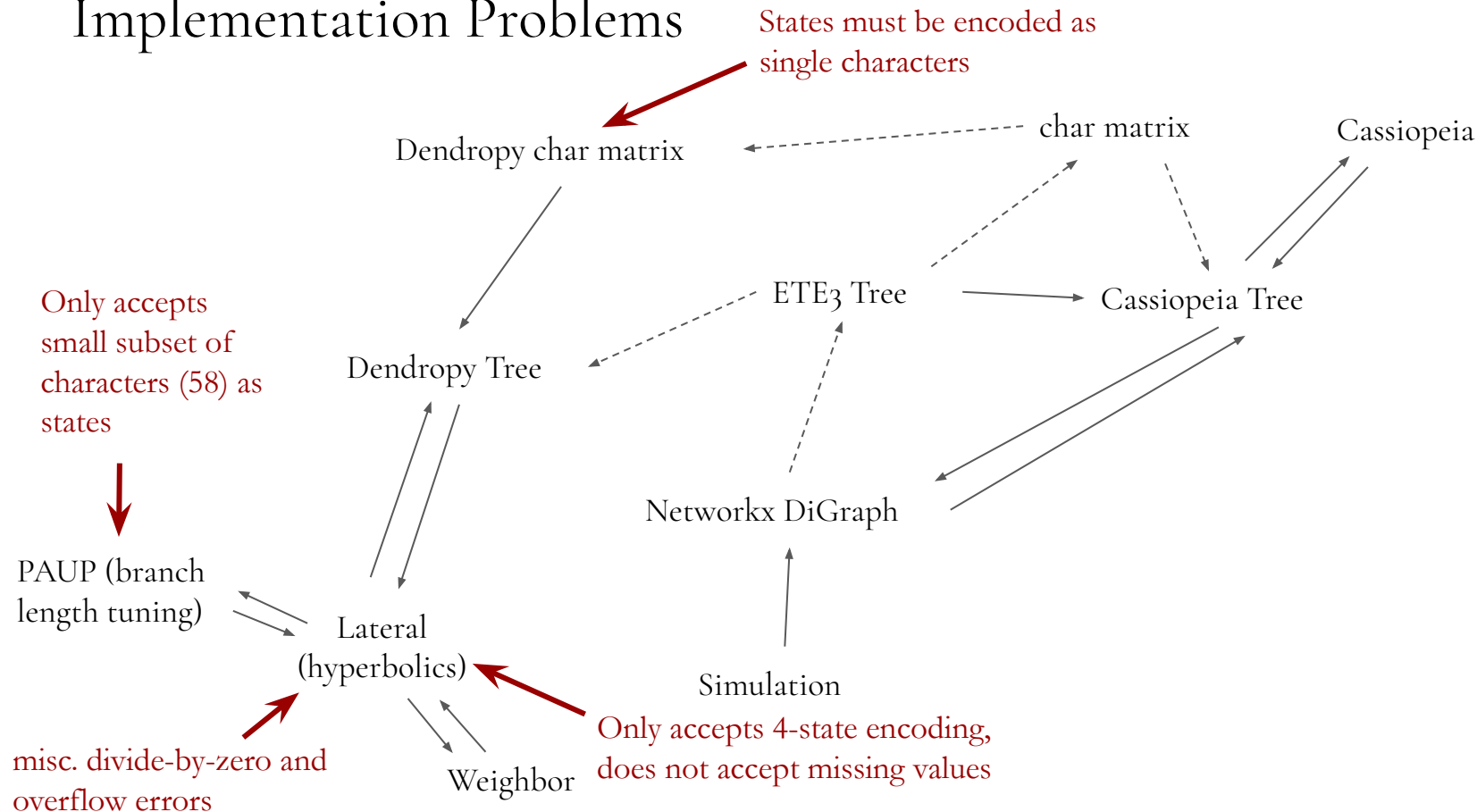
Results

Tuning branch lengths before embedding points offers no improvement
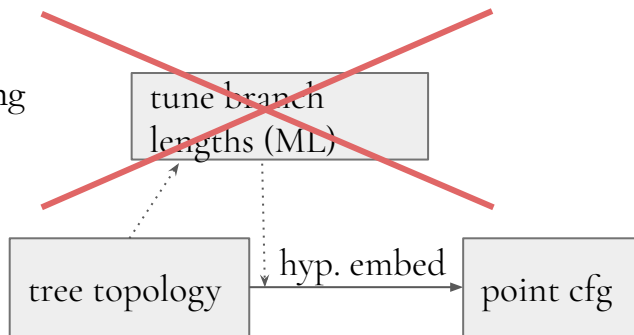
Mean algorithm performance (RF distance) over 20 simulated trees

Both the logalike hyperbolic embedding method and the Weighbor neighbor-joining algorithm outperform all others when tested on simulated CRISPR-edited lineage tracing data.
Optimizing point configuration in hyperbolic space offers no improvement over just using the Weighbor neighbor-joining algorithm.

Going forwards:

- Weighbor neighbor-joining method very promising for building phylogenetic trees from CRISPR lineage-tracing data. Outperforms traditional neighbor-joining and Cassiopeia
- Wil21 did find an improvement when combining Weighbor with point configuration tuning in hyperbolic space. Possible reasons this was not seen here:
  - Need to explore parameter space more: dimension and curvature of hyperbolic space
  - JC69 model to estimate distances does not incorporate prior information on transition rates caused by CRISPR edits
- Should evaluate along more metrics, not just Robinson-Foulds distance