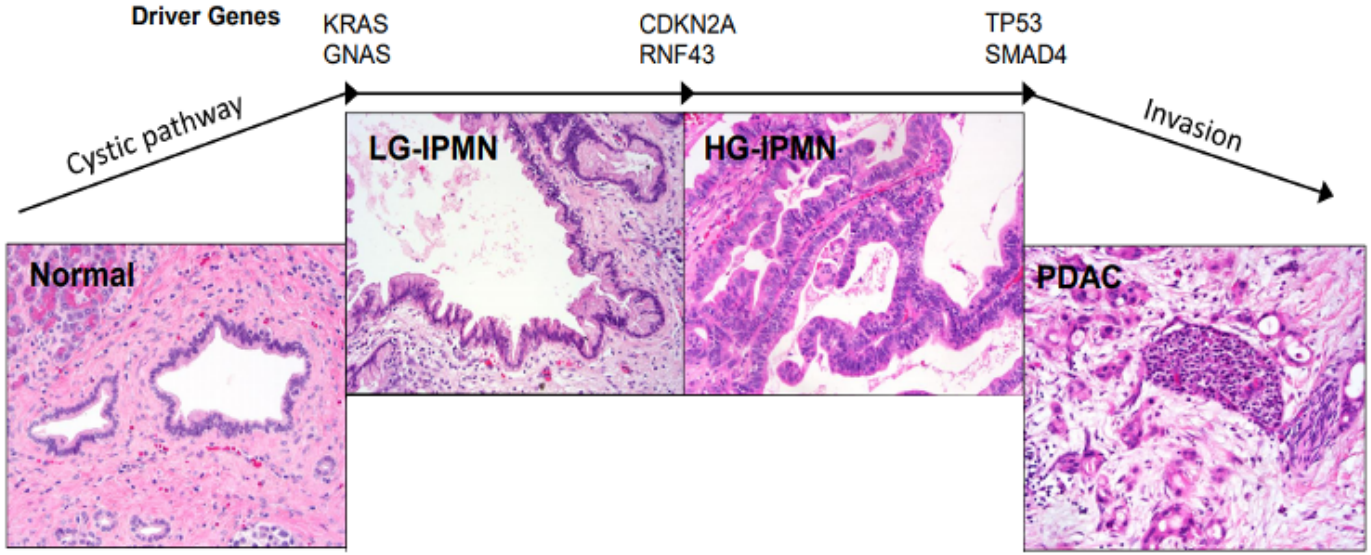
Quantifying the uncertain evolutionary history of a cancer through clone trees

Lily Zheng, Laura Wood, Rachel Karchin, Rob Scharpf

GI Spore Meeting

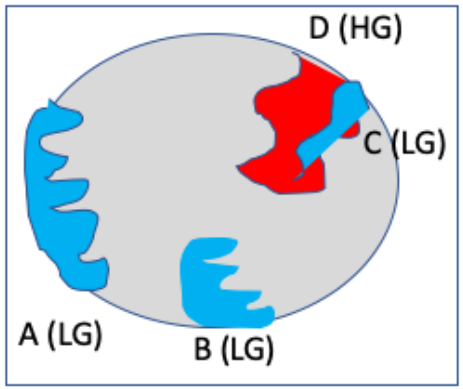
Feb. 10, 2020

The malignant progression of IPMNs is characterized by an accumulation of somatic driver gene alterations

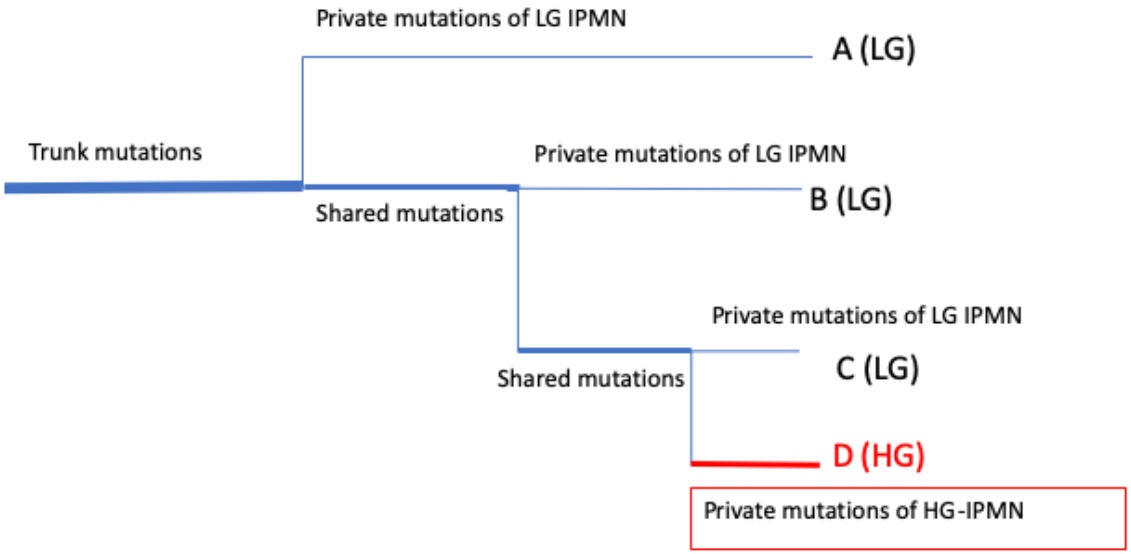


Felsenstein M. et. al. *Gut* (2018)

Phylogenetic (sample) trees are built by analyzing whether mutations are shared among or private to samples



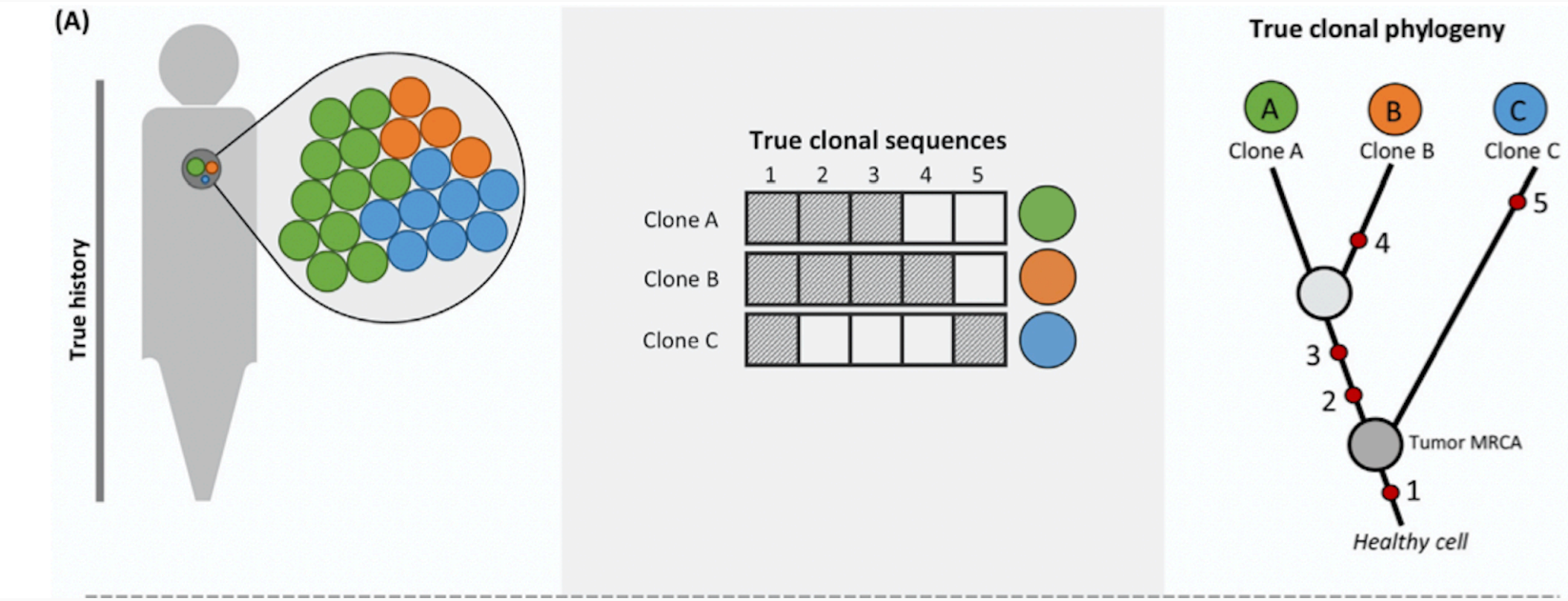
Block T1



4

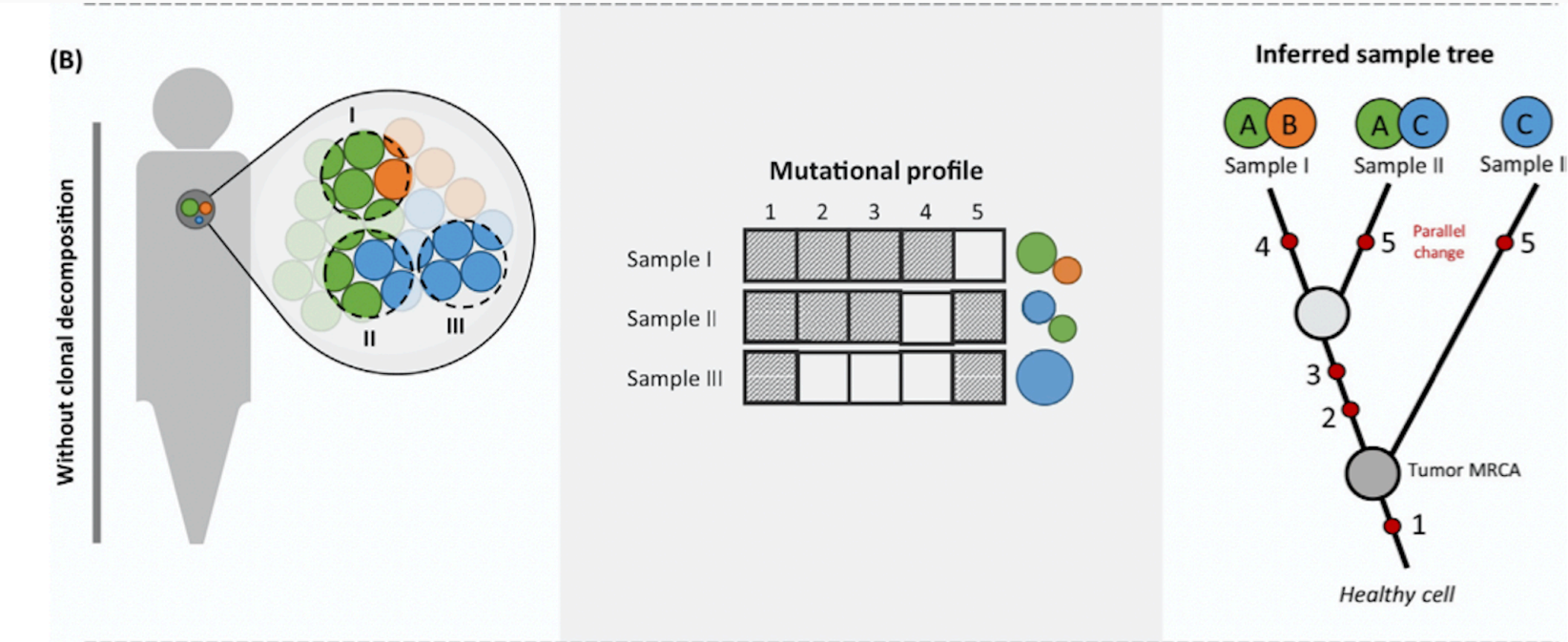
Figures courtesy of Kohei Fujikura

True clonal phylogeny



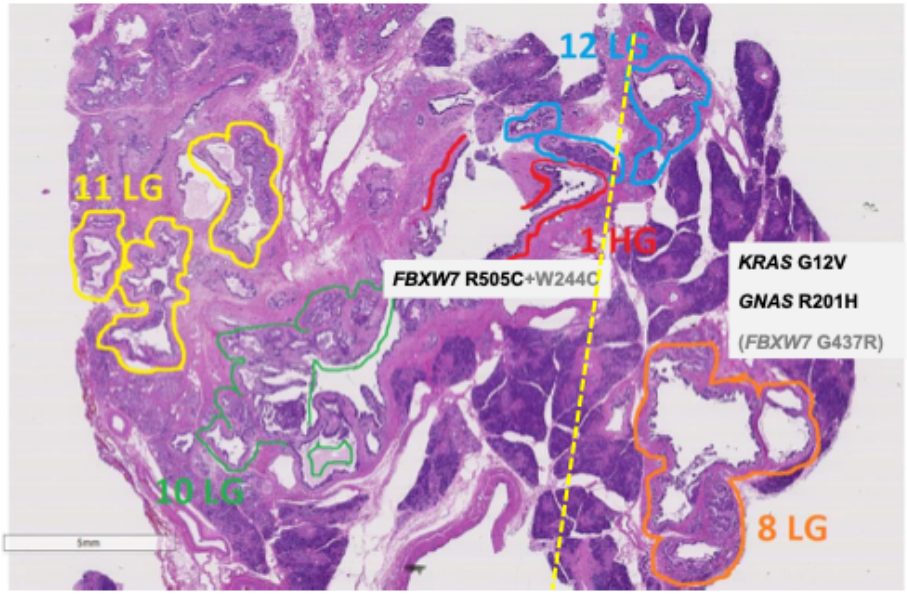
- clone C is an early branch with sequence 5 as a private mutation

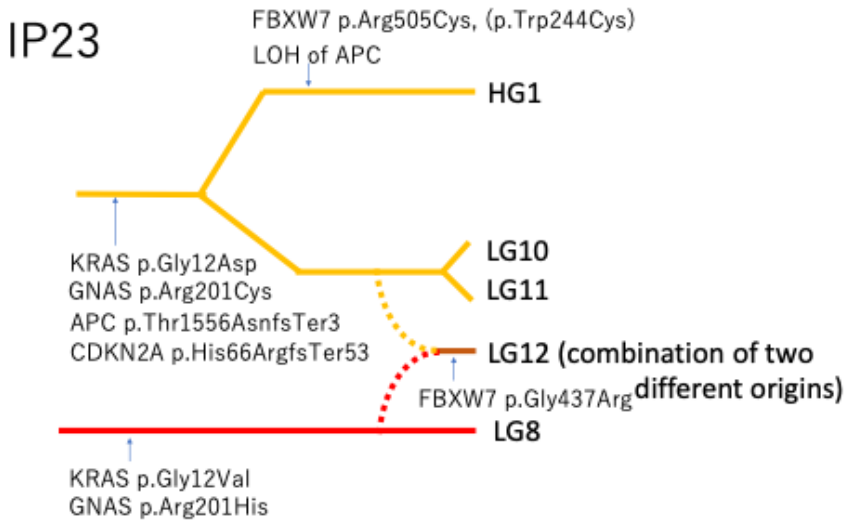
Inferred sample tree



- sample I appears to have two different origins, and the 5th sequence mutation is recorded as a parallel change between samples I and II

Sample trees do not reflect within sample heterogeneity





Shared mutation (except LG8): *APC CDKN2A GNAS KRAS MUC16*

5

Figures courtesy of Kohei Fujikura

Bayesian probabilistic inference of clone trees

For y_{is} denote the number of reads containing variant i in sample s. The sampling distribution for y is binomial:

$$[y_{is}|\text{VAF}_{is}, n_{is}] \sim \text{Binomial}(\text{VAF}_{is}, n_{is})$$

$$[\text{VAF}_{is}|Z_{is} = z, m_{is}, c_{is}, \omega_{zs}] = \frac{m_{is} \times \text{MCF}_{zs}}{c_{is} \times \text{MCF}_{zs} + 2 \times (1 - \text{MCF}_{zs})}$$

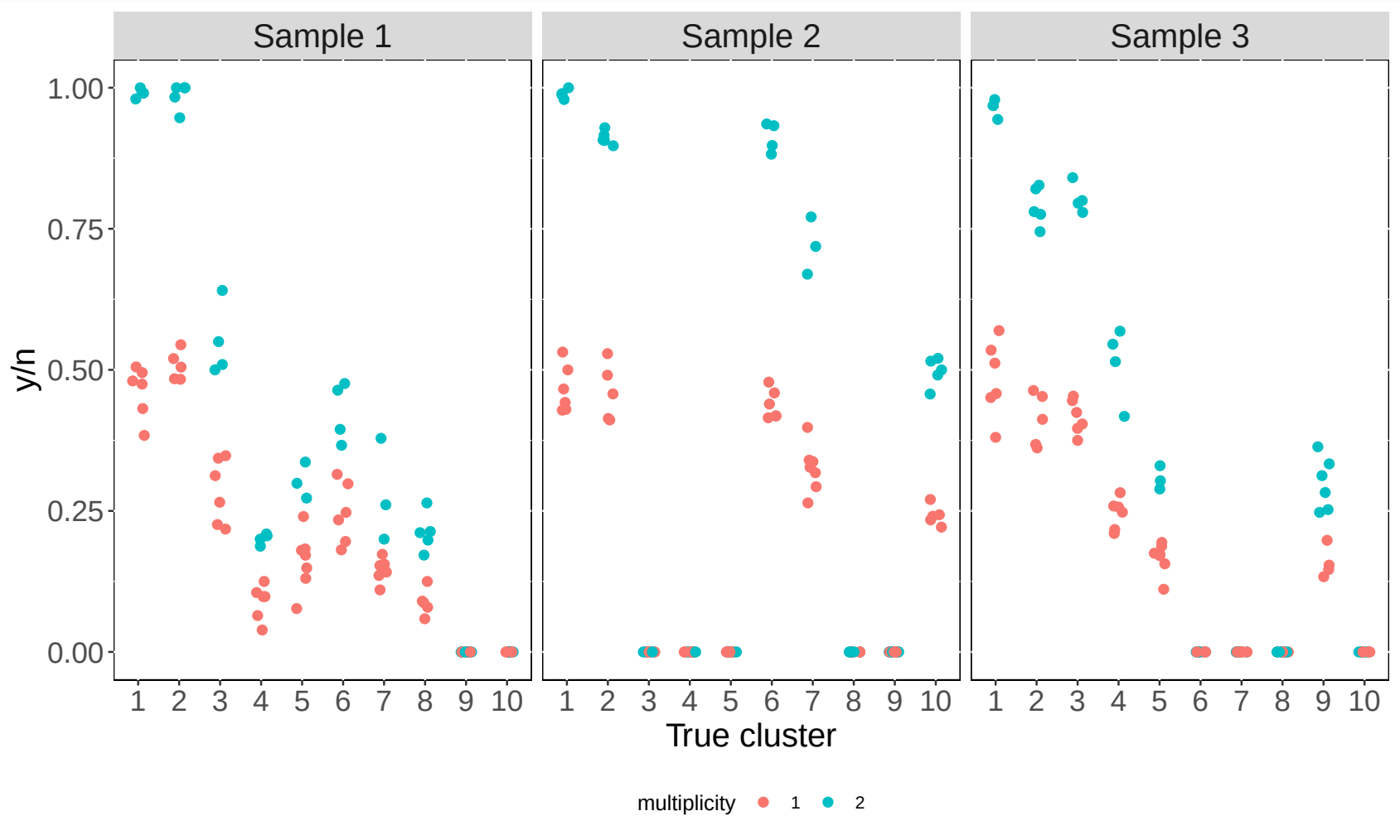
$$[z_i|\pi_1, \dots, \pi_K, K] \sim \text{Multinomial}(\pi_1, \dots, \pi_K)$$

$$\text{MCF}_{zs}|G \sim \text{Beta}(a_G, b_G)$$

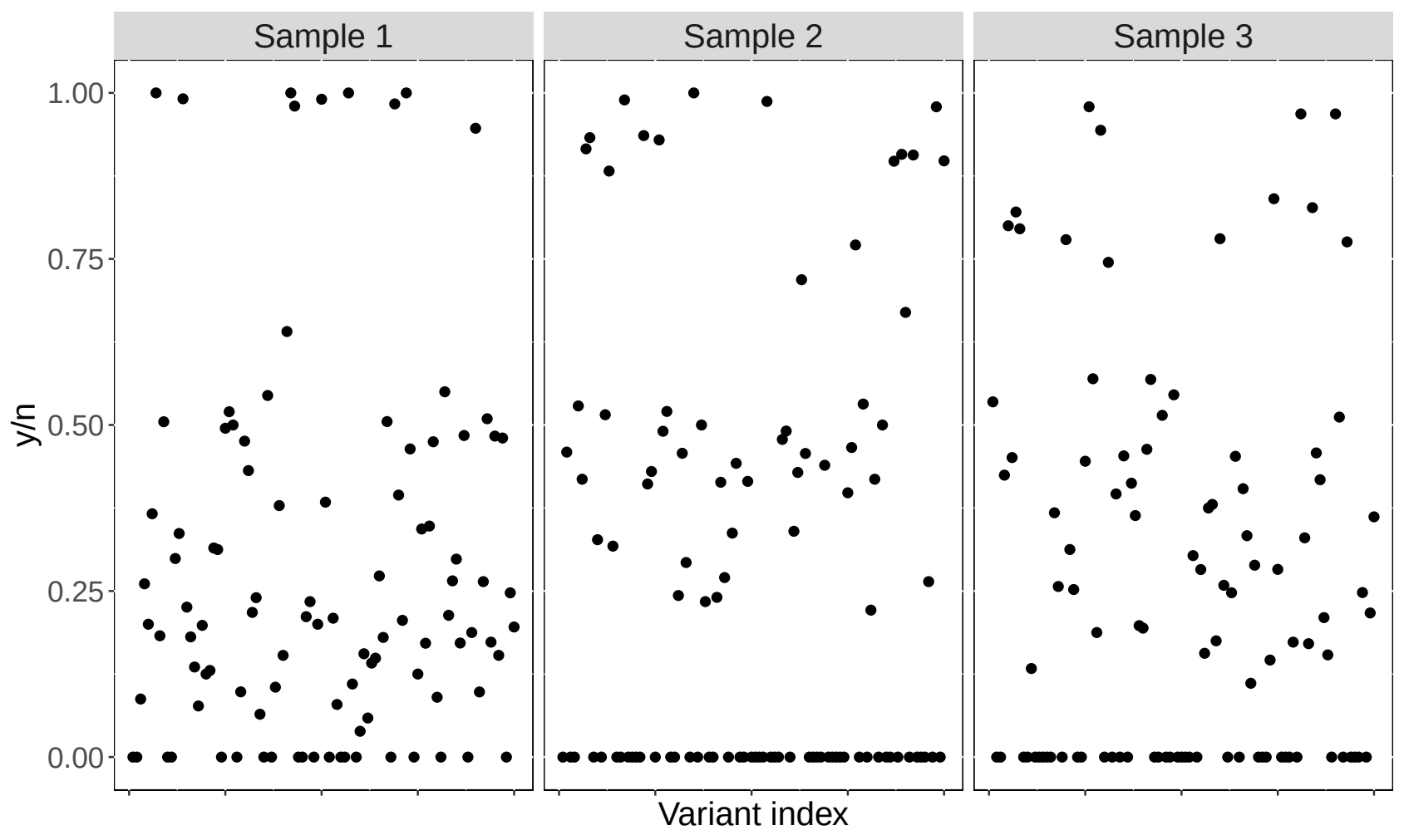
Given K, find z and the MCF

- Variants in the same clone z have the same unobserved MCF
- Neither z or the MCF are observed
- Adopting a Bayesian hierarchical model for the observed VAF, we obtain a posterior distribution for the unobserved MCF and the clonal membership for each variant
    - As the counts are modeled directly, uncertainty reflected in the posteriors for these parameters reflects depth of coverage and the degree to which clones have distinguishable MCFs

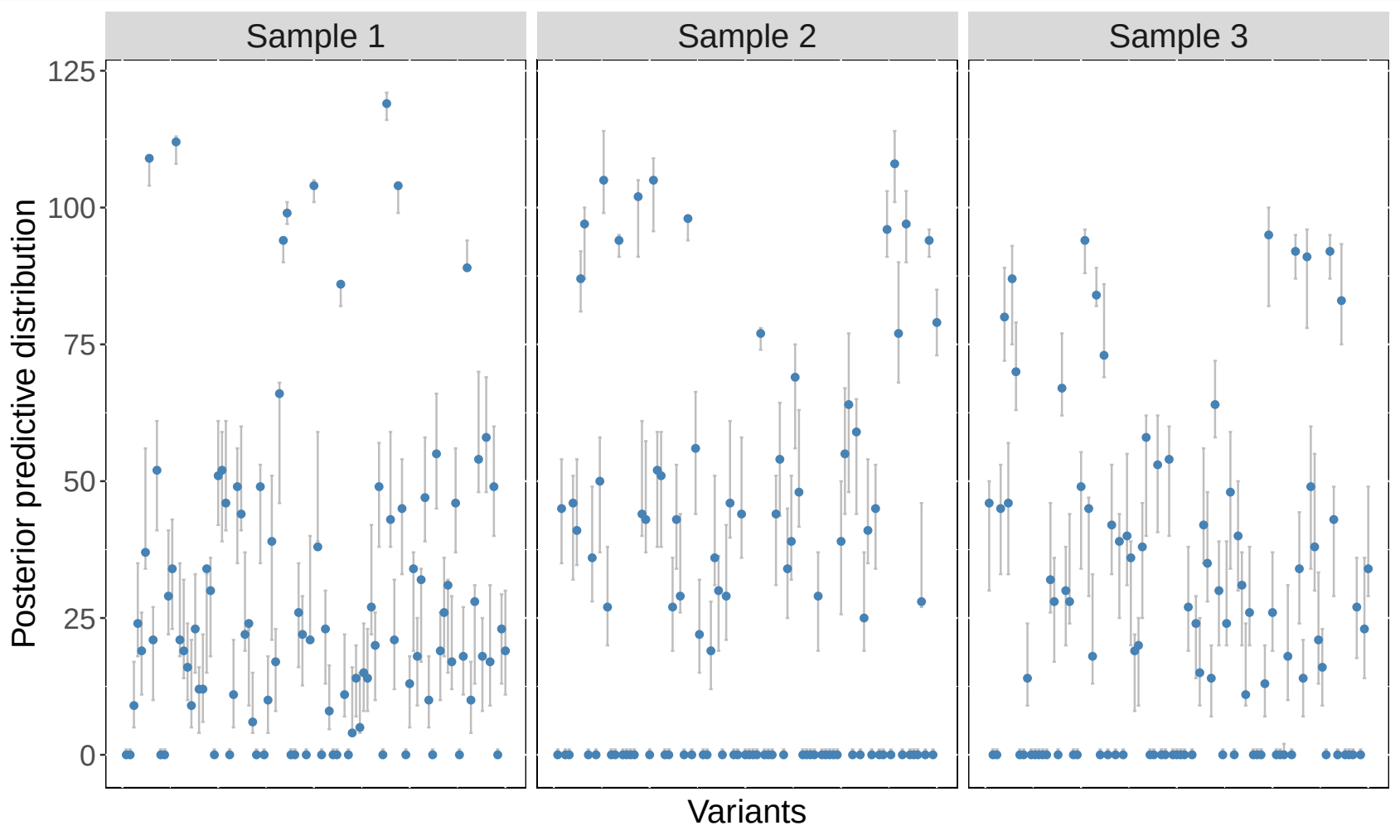
Simulated data grouped by true cluster



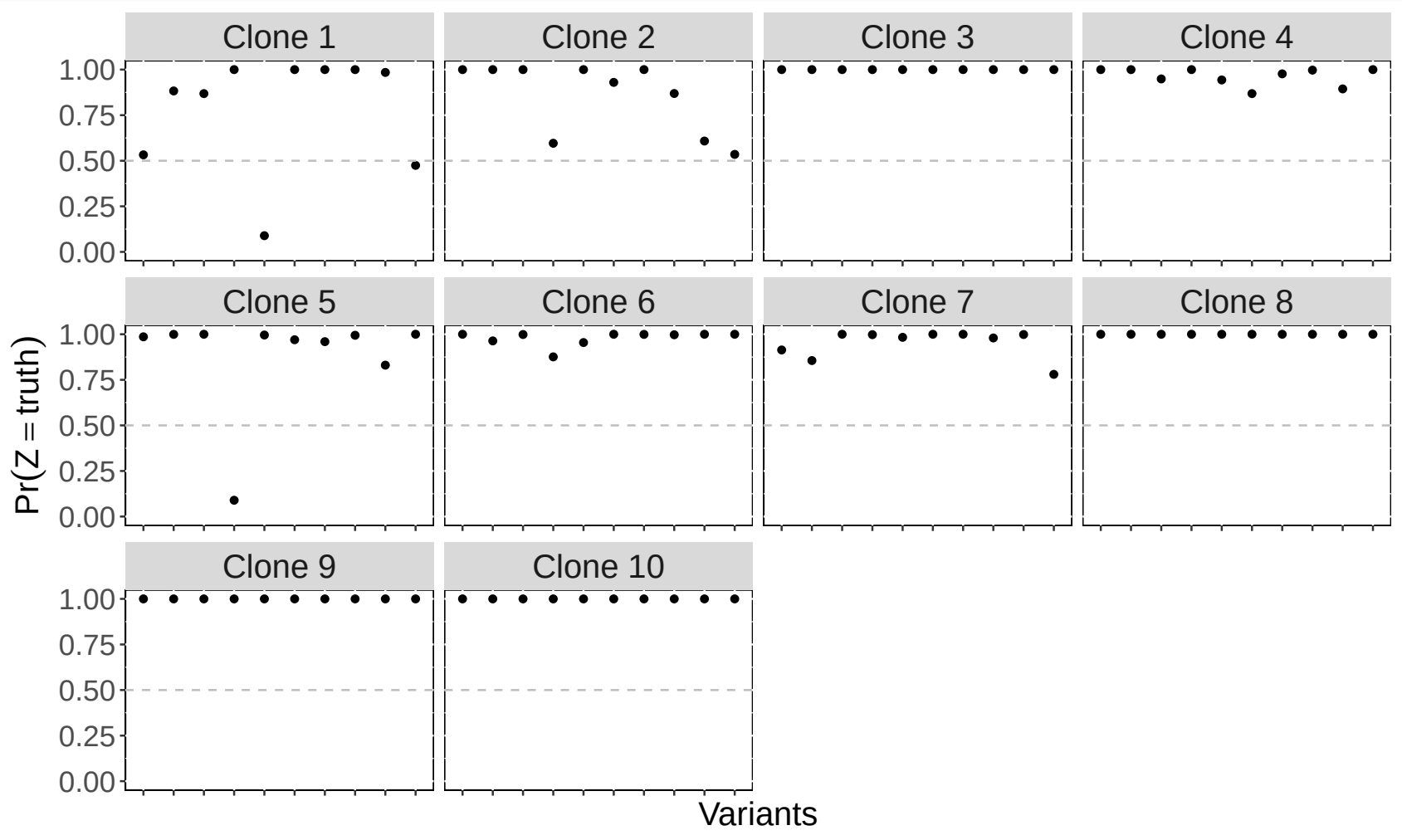
Same data


Sample 1
Sample 2
Sample 3
1.00
0.75
0.50
0.25
0.00
y/n
Variant index

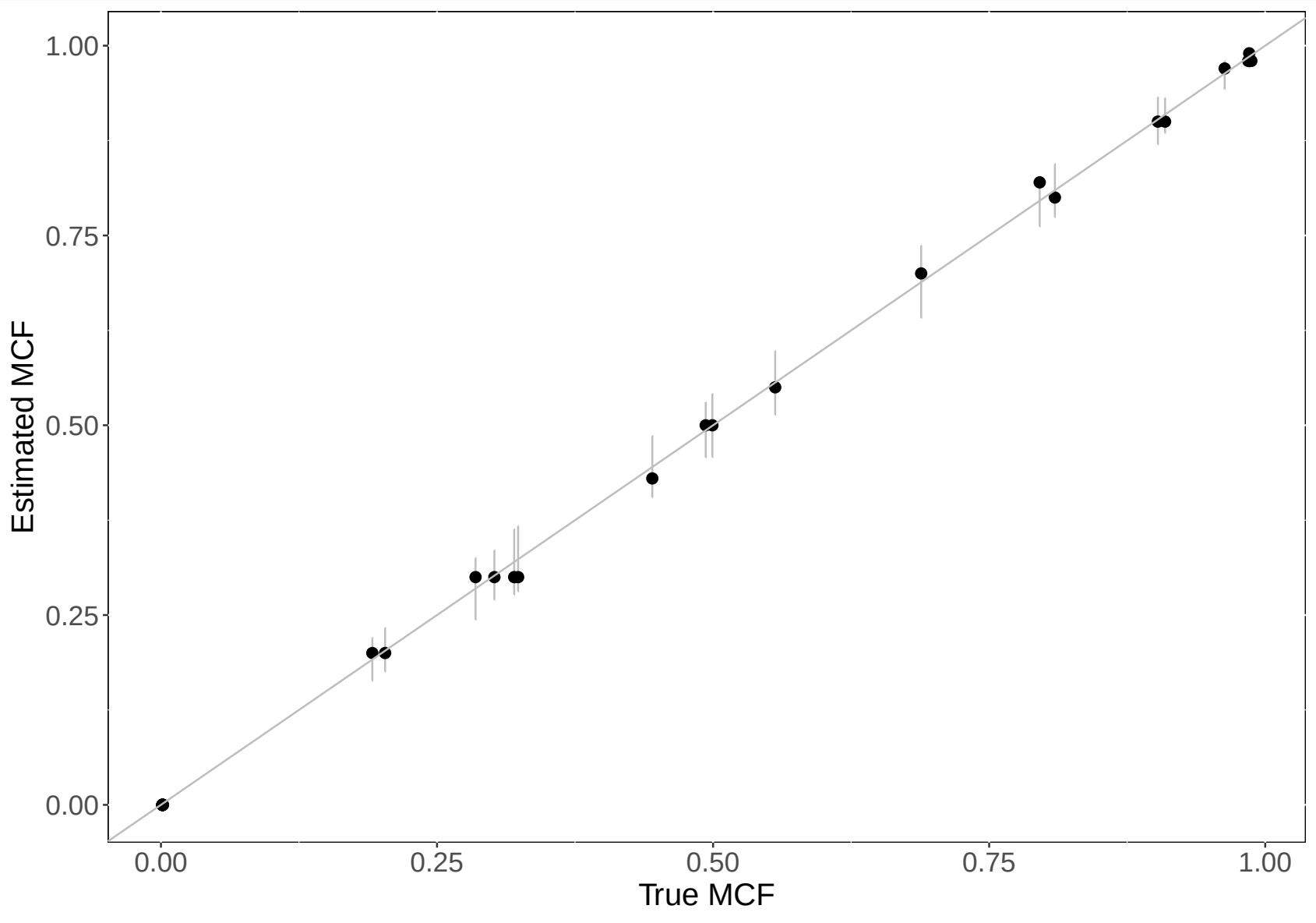
Goodness of fit



Posterior probability of clone assignments



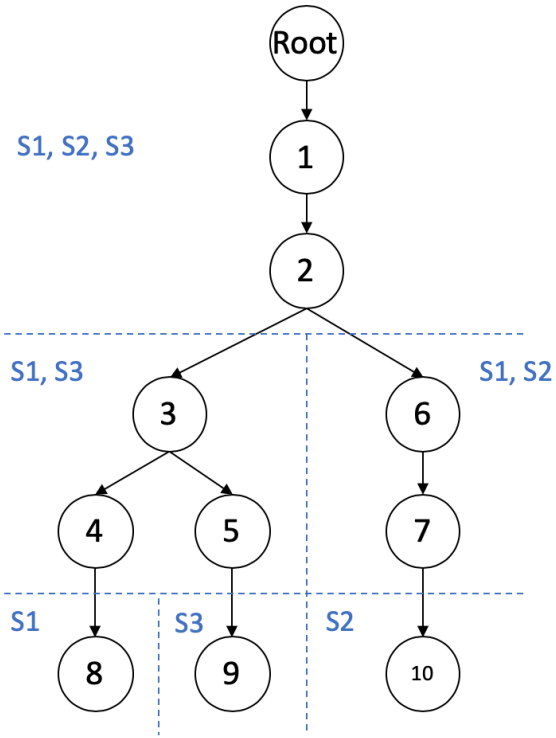
MCFs



- Error bars are 95% posterior credible intervals

Sampling Directed Acyclic Graphs (DAGs)

Simulated data: 100 variants total, 10 per cluster



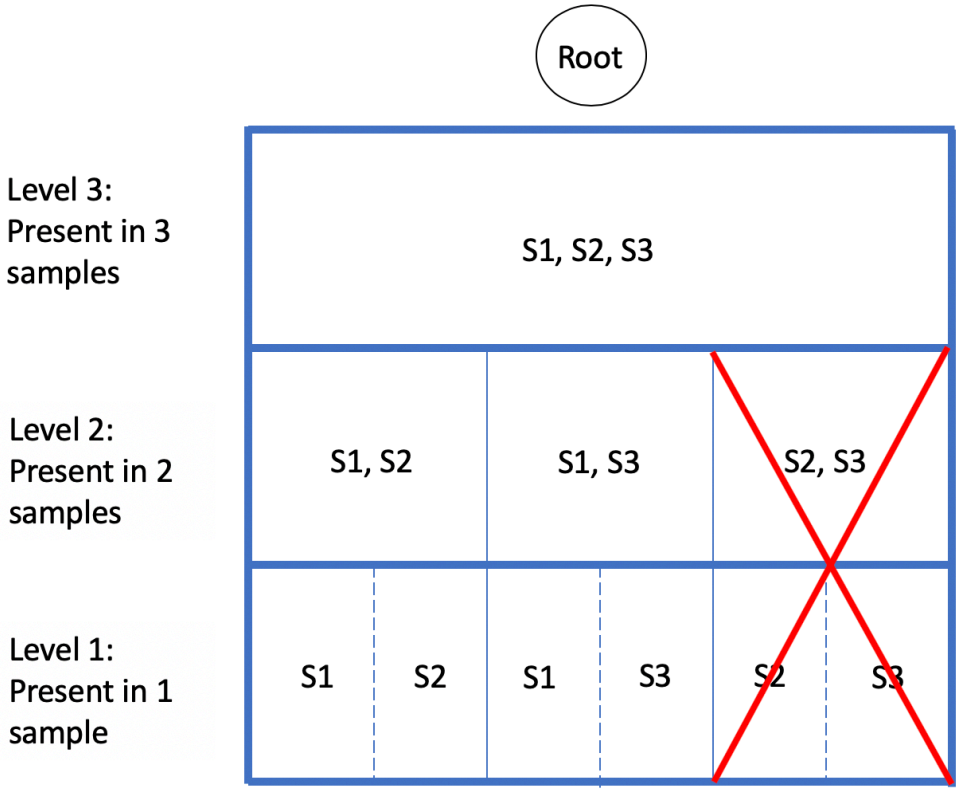
MCF, ω

k	S1	S2	S3
1	0.98	0.99	0.97
2	0.98	0.90	0.82
3	0.55	0.00	0.80
4	0.20	0.00	0.50
5	0.30	0.00	0.30
6	0.43	0.90	0.00
7	0.30	0.70	0.00
8	0.20	0.00	0.00
9	0.00	0.00	0.30
10	0.00	0.50	0.00

- For 10 clones, there are 1.59×10^{10} possible DAGs!
- Searching the space of all possible trees is computationally prohibitive

Heuristic

Crude tree structure based on sample presence



ω

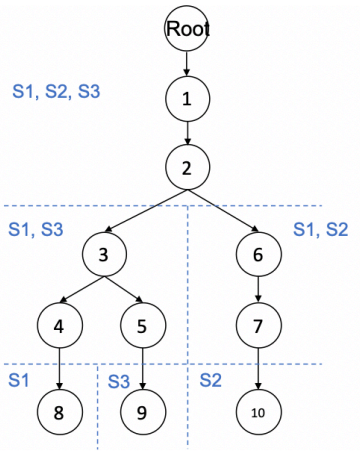
	sample1	sample2	sample3
[1,]	0.98	0.99	0.97
[2,]	0.98	0.90	0.82
[3,]	0.55	0.00	0.80
[4,]	0.20	0.00	0.50
[5,]	0.30	0.00	0.30
[6,]	0.43	0.90	0.00
[7,]	0.30	0.70	0.00
[8,]	0.20	0.00	0.00
[9,]	0.00	0.00	0.30
[10,]	0.00	0.50	0.00

Present in sample
if ω > 0.01

- ≈ 2 million DAGs compatible with these constraints

Adjacency matrix of edges

Crude tree structure places restraints on possible adjacency matrices

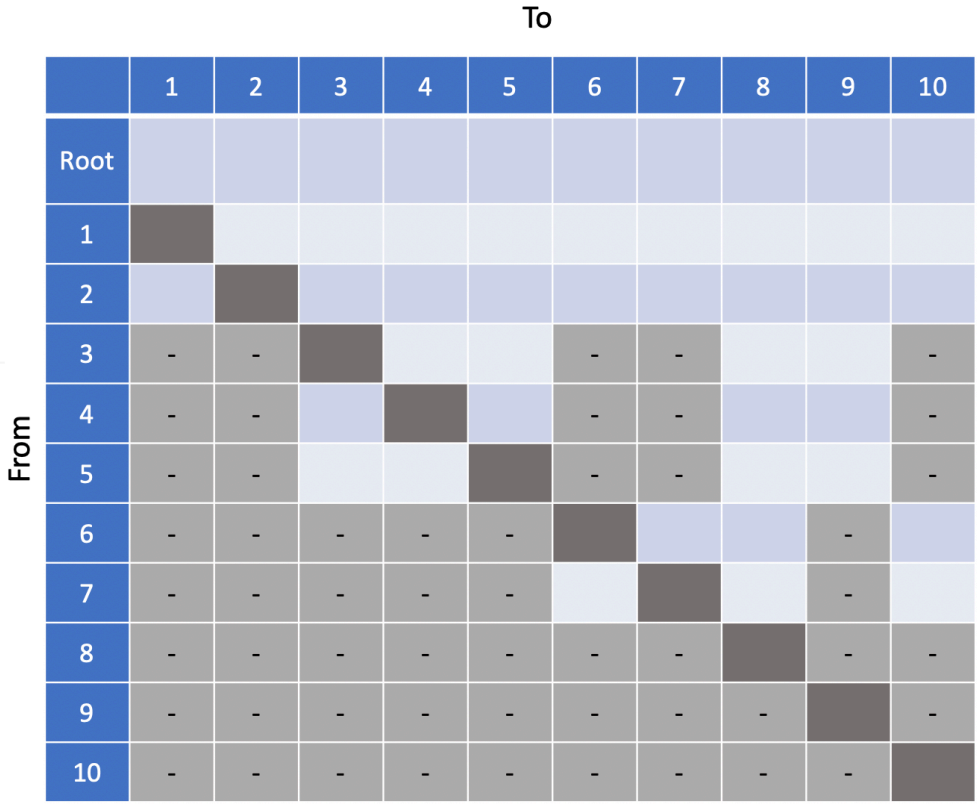


ω

	sample1	sample2	sample3
[1,]	0.98	0.99	0.97
[2,]	0.98	0.90	0.82
[3,]	0.55	0.00	0.80
[4,]	0.20	0.00	0.50
[5,]	0.30	0.00	0.30
[6,]	0.43	0.90	0.00
[7,]	0.30	0.70	0.00
[8,]	0.20	0.00	0.00
[9,]	0.00	0.00	0.30
[10,]	0.00	0.50	0.00

Constraints:

- Cannot go to self
- No costraints if present in same set of samples
- # from.samples < # to.samples
- to.samples must be a subset of from.samples



- gray boxes indicate edges that are prohibited

Conclusions and ongoing efforts

- Inferential goals are modest:
    - to derive posterior probability distributions of the MCFs for each clone in each available sample
    - to derive posterior probability distributions for the DAGs depicting the evolutionary relationship among these clones
    - the probability of a specific clone tree or a set of similar clone trees is easily obtained
- An efficient MCMC sampler for DAGs and application to experimental datasets are in progress