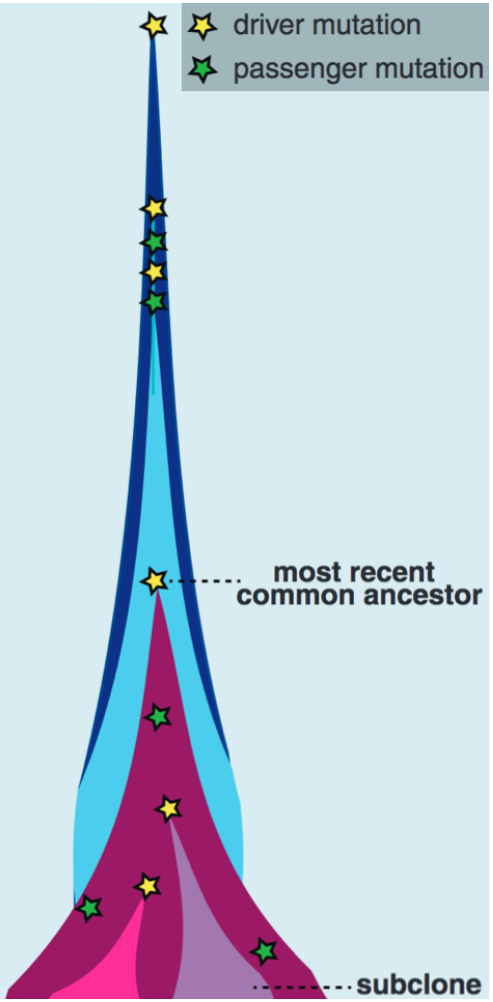
ETH zürich

Single-cell phylogenetics

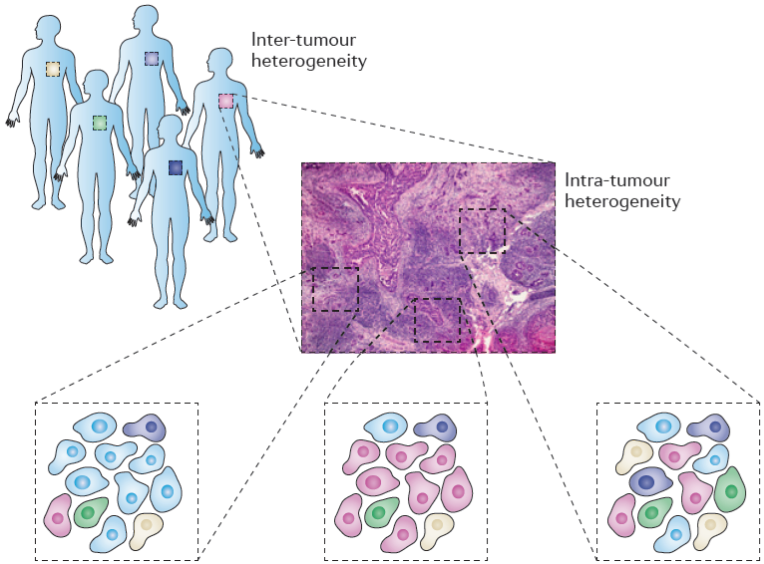
current challenges and future directions

Jack Kuipers, ETH Zürich, 25 January 2018
with Katharina Jahn and Niko Beerenwinkel

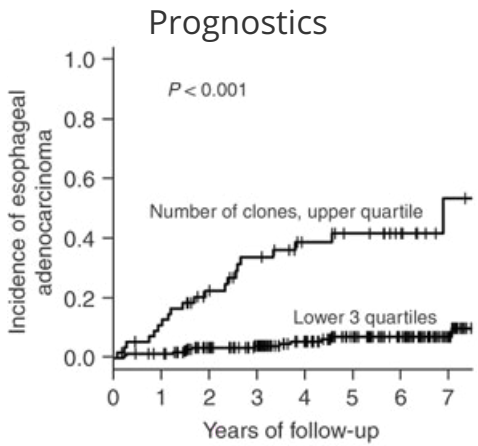
Tumour heterogeneity



van Loo and Voet, COGD 2014

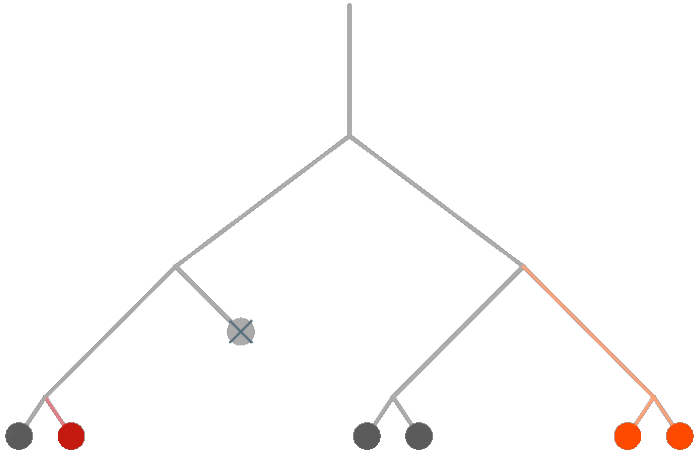


Marusyk, Almendro and Polyak, Nat Rev Gen 2012

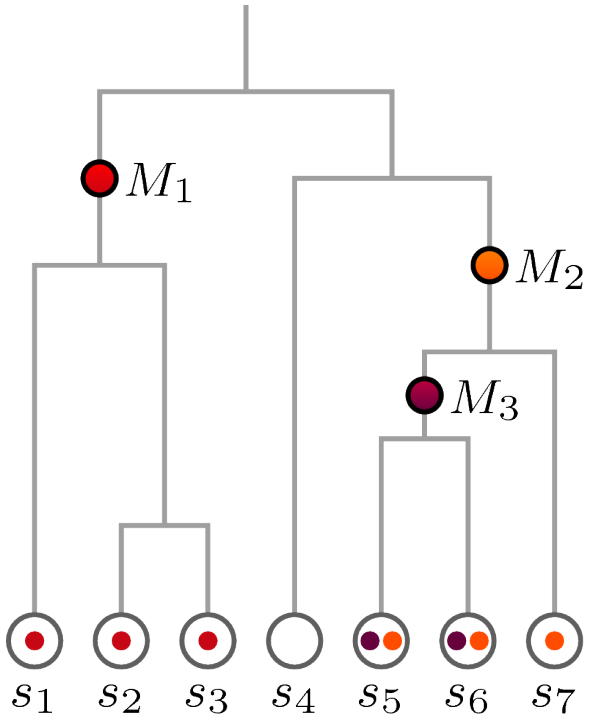


Maley et al, Nat Gen 2006

Cell evolution

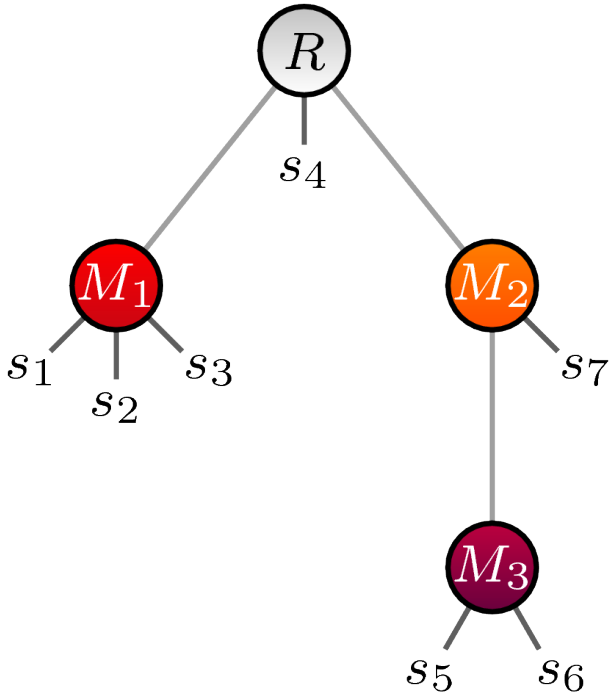


with phylogenetic tree

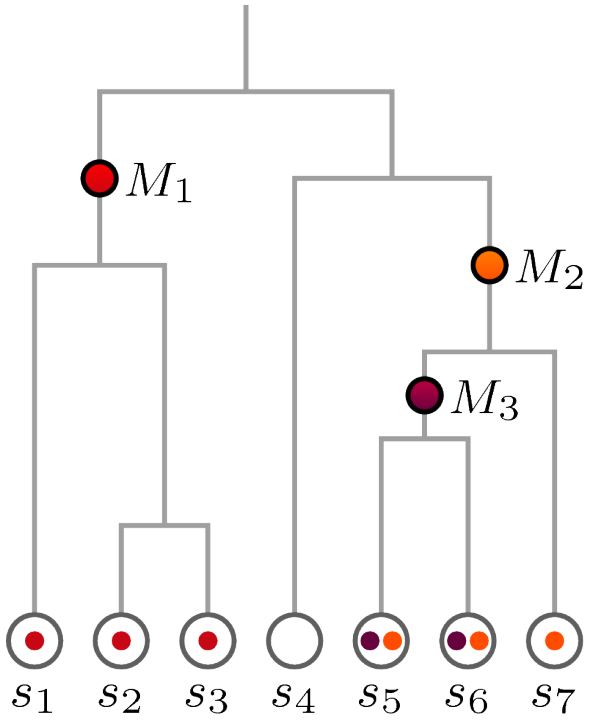


Mutation tree

Mutations also lie on a (rooted) tree

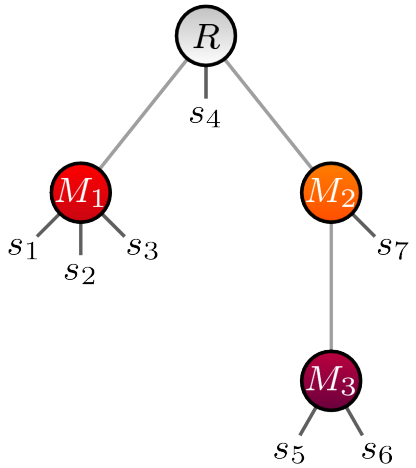


from phylogenetic tree



- samples attached as leaves
- inherit ancestral mutations

Observational errors



But we measure with errors

- false positive rate α
  - $0 \to 1$
- false negative rate β
  - $1 \to 0$

Expected mutation matrix

$$E = \begin{array}{c|ccccccc} & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ \hline M_1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ M_2 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ M_3 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{array}$$

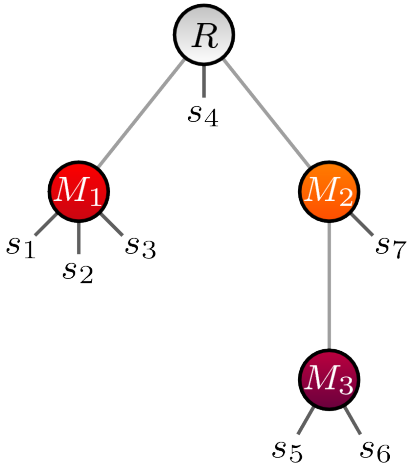
Observed data instead

$$D = \begin{array}{c|ccccccc} & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ \hline M_1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ M_2 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ M_3 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$$

Tree reconstruction?

$D =$

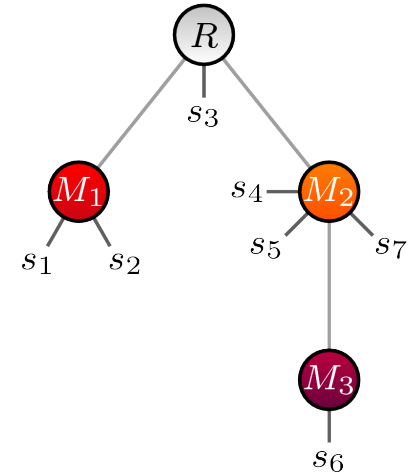
	s_1	s_2	s_3	s_4	s_5	s_6	s_7
M_1	1	1	0	0	0	0	0
M_2	0	0	0	1	1	1	1
M_3	1	0	0	0	0	1	0



- can we reconstruct the phylogeny?

Given a tree T with attachments $\boldsymbol{\sigma}$, we know the likelihood of the data

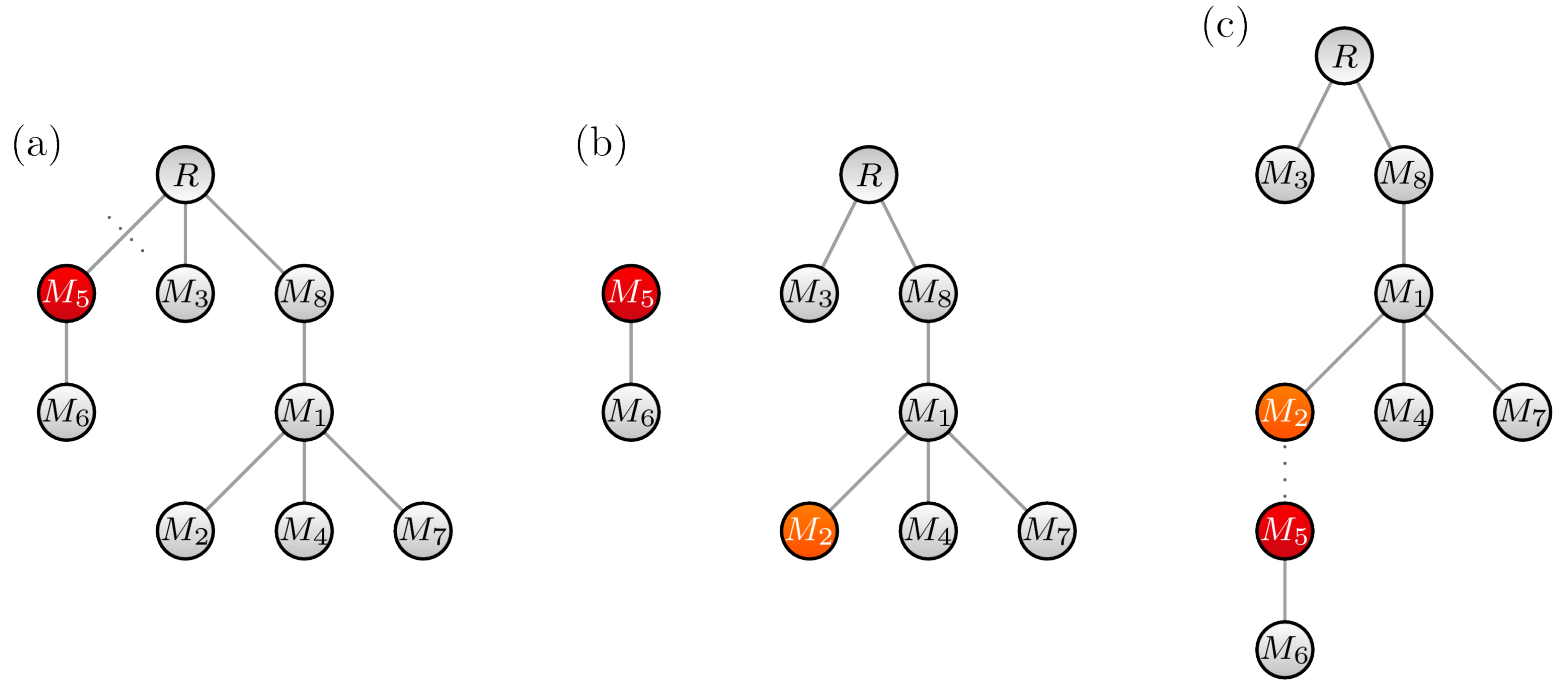
$$P(D|T, \boldsymbol{\sigma}, \alpha, \beta) = \alpha^{\sum_{ij} I_0(E_{ij})I_1(D_{ij})}(1-\alpha)^{\sum_{ij} I_0(E_{ij})I_0(D_{ij})} \beta^{\sum_{ij} I_1(E_{ij})I_0(D_{ij})}(1-\beta)^{\sum_{ij} I_1(E_{ij})I_1(D_{ij})}$$



- test all trees and attachments small n
- find maximum likelihood

Tree search

- Can marginalise/maximise sample attachment $\boldsymbol{\sigma}$ efficiently $O(mn)$
- Stochastic search through tree space



- Test for best tree which maximises $P(D|T, \alpha, \beta)$

Single cell data

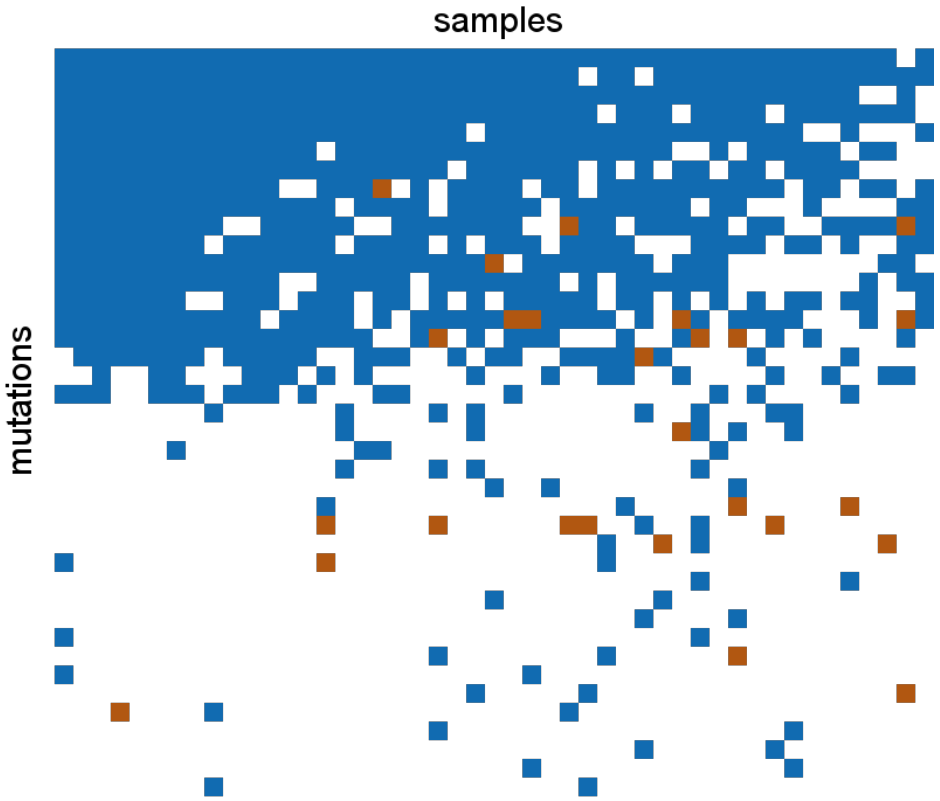
Breast tumour Wang … Navin, Nature 2014

- 47 cells columns
- 40 mutations rows

Error rates

$\alpha = 1.24 \times 10^{-6}$

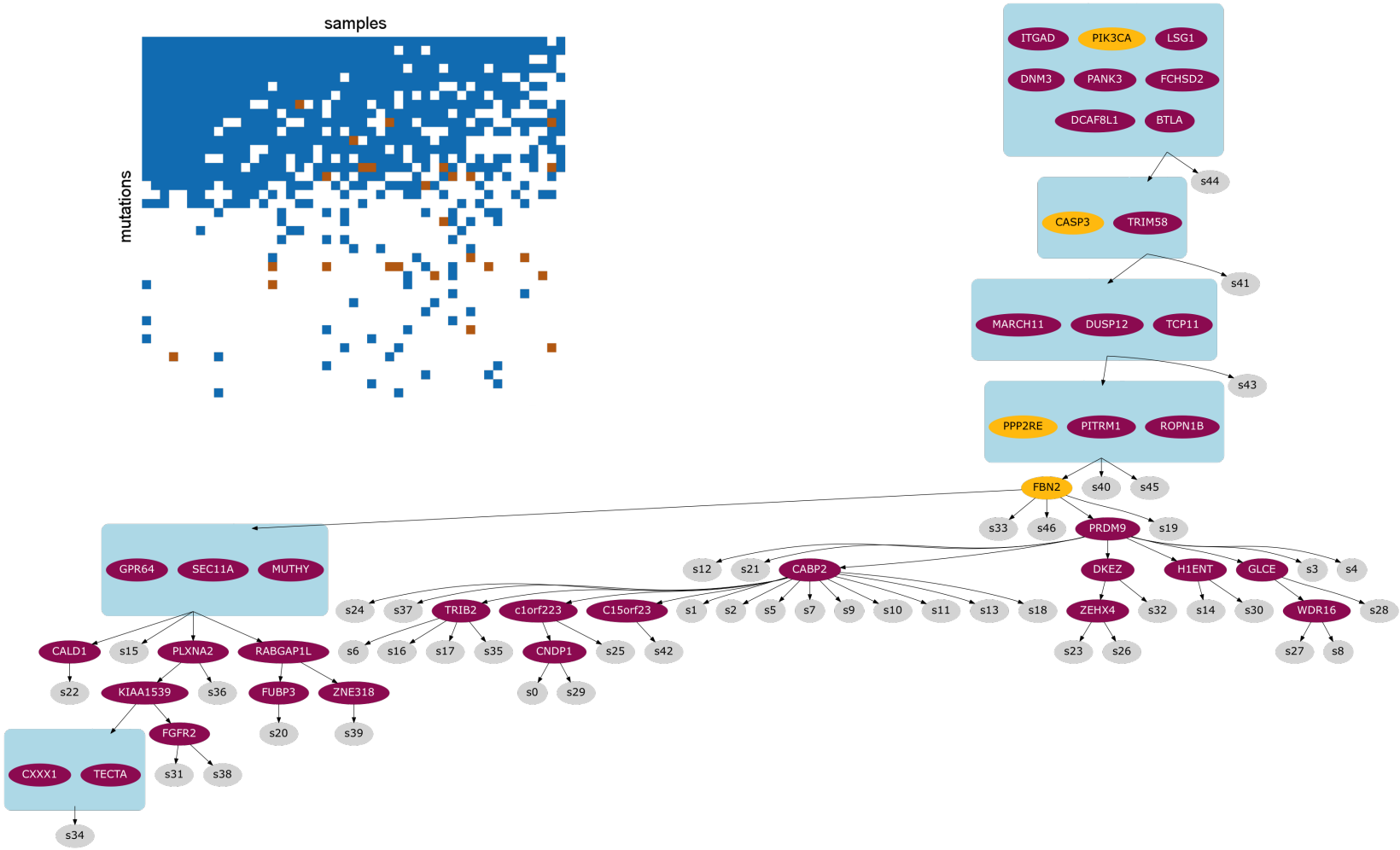
$\beta = 9.73\%$



mutation

missing data

Tree inferred from single cell data



Data Wang … Navin, Nature 2014 Inference Jahn, Kuipers and Beerenwinkel, Genome Biology 2016

Infinite sites assumption

Each mutation can only occur once

- Whole genome $\approx 3 \times 10^9$
- Whole exome $\approx 3 \times 10^7$

Is this infinity?

Birthday paradox:

- with only 23 people
- probability of shared birthday $> \frac{1}{2}$

In general with Z 'days'

- require $\approx \sqrt{2Z}$ people

Even for slowest dividing cancer (osteosarcoma)

- 10^7 lifetime mutations

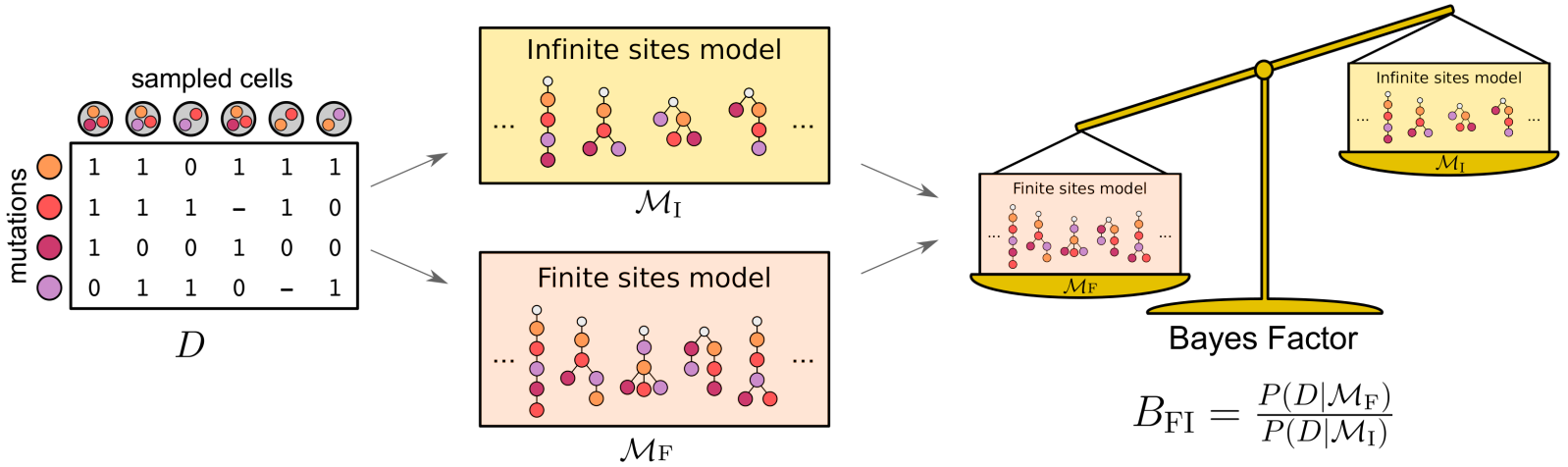
  Tomasetti and Vogelstein, Science 2015

Probability of a mutation occuring twice ≈ 1

Bayes factors

with Ben Raphael

Perform model comparison with Bayes factor



- $\mathcal{M}_{\mathrm{I}}$ all trees with no recurrent mutations
- $\mathcal{M}_{\mathrm{F}}$ all trees with a single recurrence

- $B_{\mathrm{FI}} > 1$ favours violation of infinite sites assumption

Posterior ratio

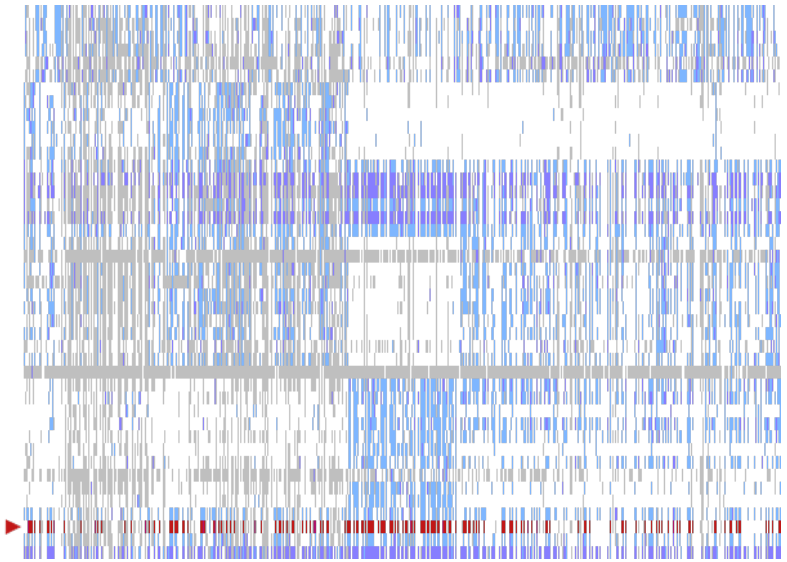
$$\frac{P(\mathcal{M}_{\mathrm{F}} \mid D)}{P(\mathcal{M}_{\mathrm{I}} \mid D)} = B_{\mathrm{FI}} \frac{P(\mathcal{M}_{\mathrm{F}})}{P(\mathcal{M}_{\mathrm{I}})}$$

Genome Research, 2017

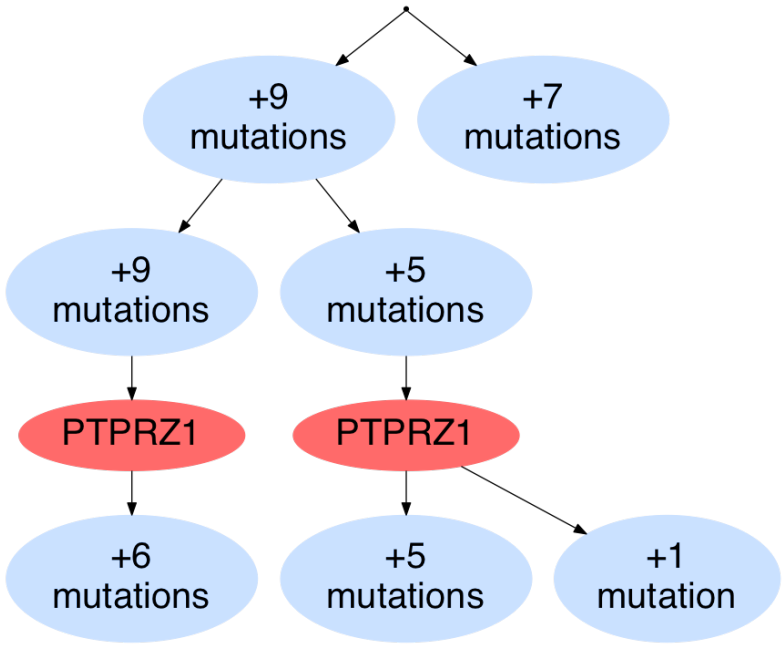
Single cell sequencing data

Tested 3 whole exome and 9 panel sequencing datasets

- find evidence for violations in 11 out of 12
- 4 examples of parallel mutations



McPherson, Roth … Shah, Nature Genetics 2016

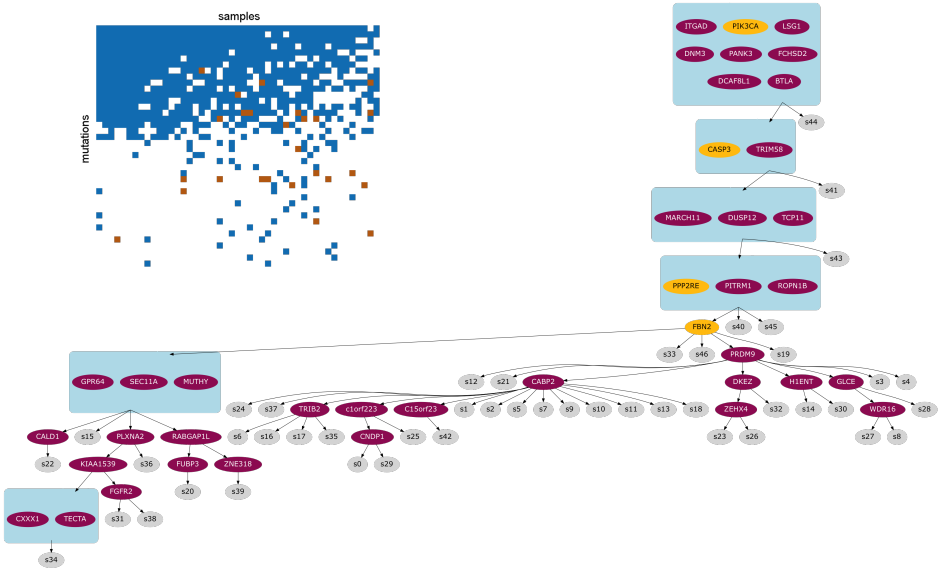


- panel of 43 mutations
- 588 cells
- Bayes Factor: 7.2×10^{14}

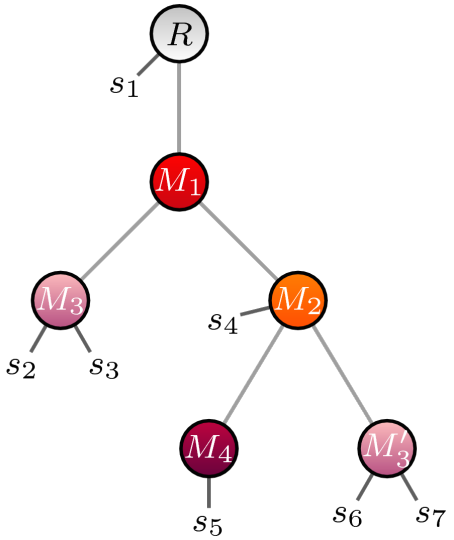
Summary

Mutation tree

- can be inferred from single cell data



Genome Biology 2016



Can extend to test infinite sites assumption

- doublet samples need to be modelled

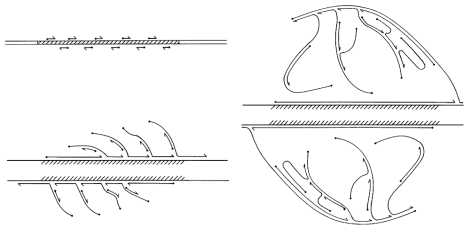
Assumption violated

- often by mutational loss
- occasionally by parallel mutations

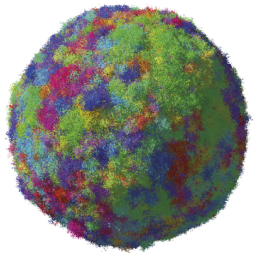
Genome Research, 2017

Current and future challenges

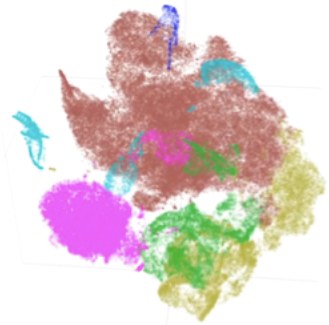
- Modelling sequencing data



- Copy number abberations
  - mutational loss
  - overlap
- Inferring evolutionary parameters



- Integrating bulk sequencing bioRxiv:234914
- Connecting to gene expression and drug response



- Faster inference
  - better MCMC moves
  - SMC samplers
  - ILP schemes