ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Tumour archeology

Katharina Jahn

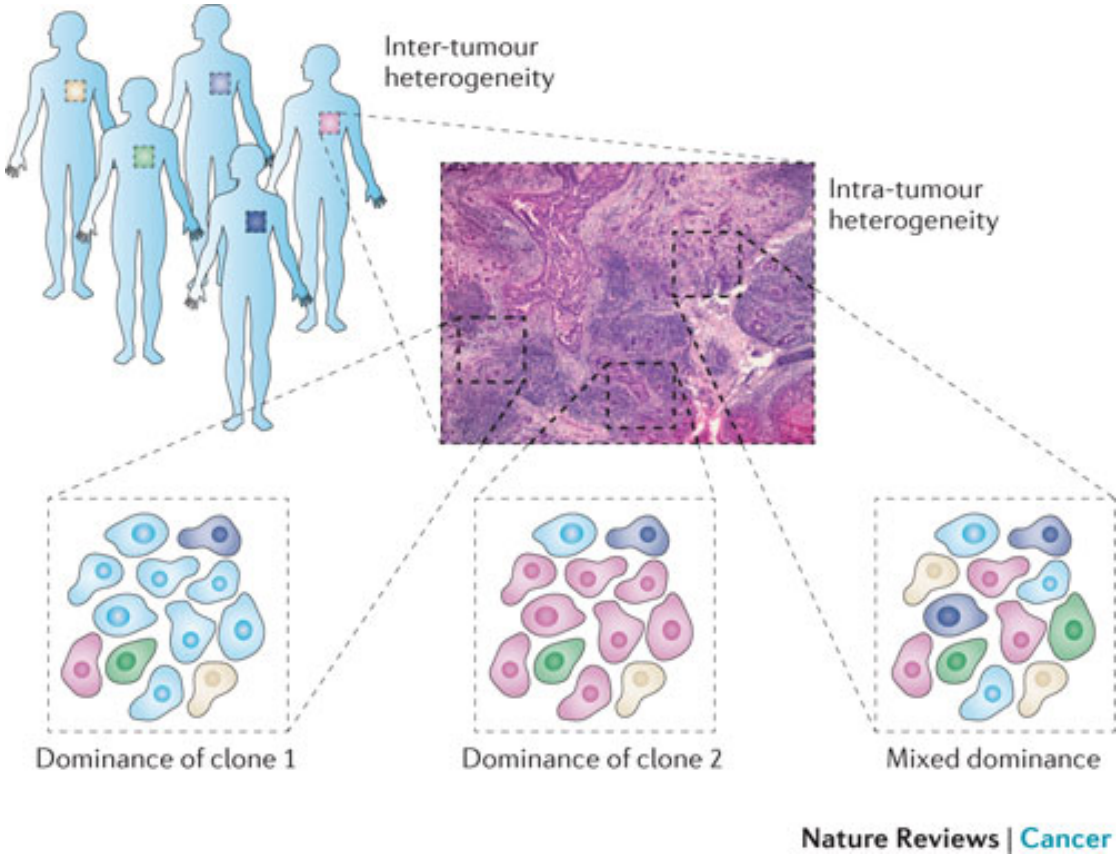
Goals

- Connect cancer sequencing data with evolutionary models
- Give an idea how we can infer parameters of cancer evolution (mutation rate, fitness advantage, tumour age, etc.) from sequencing data

Outline

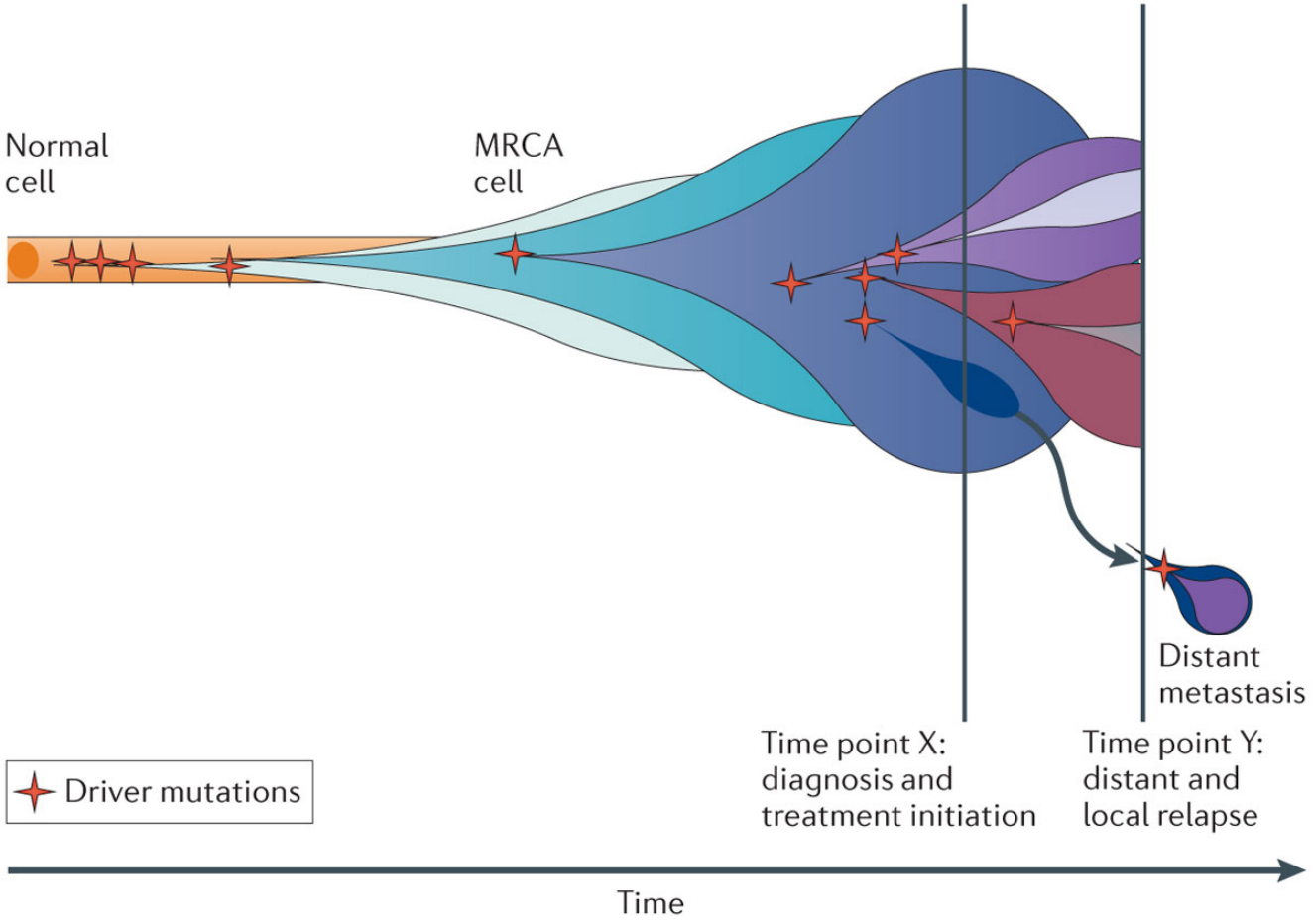
- Intra-tumour heterogeneity/models of tumour evolution
- The Variant-Allele-Frequency-Spectrum
- Inference under the neutral model
- Inference in the presence of selection
- Confounders of the Variant-Allele-Frequency-Spectrum

Intra-tumour heterogeneity



Marusyk et al. 2012

Clonal tumour evolution

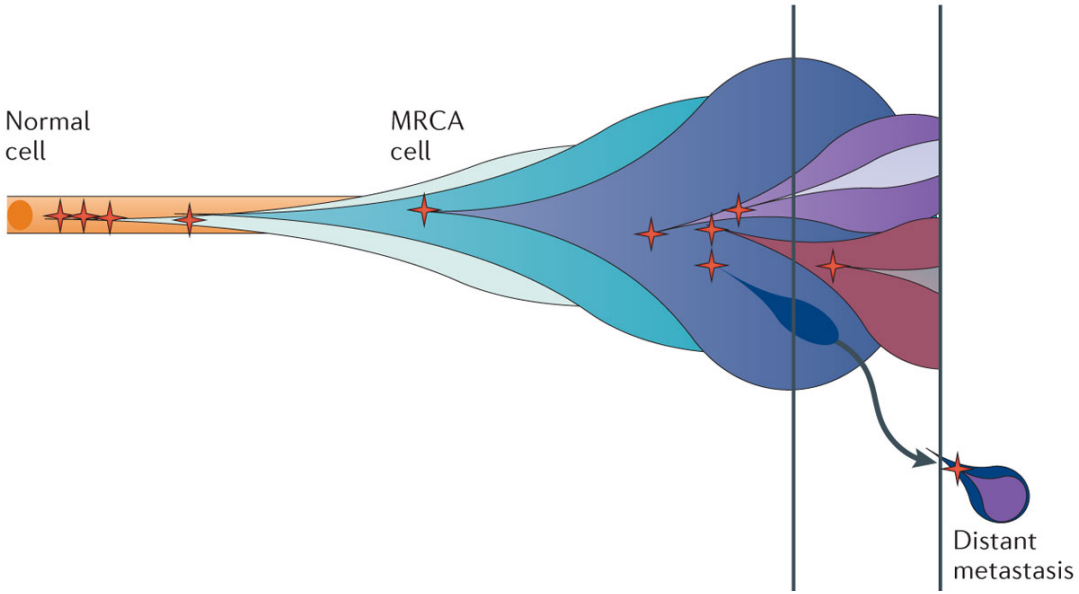


Yates et al. 2012

Nature Reviews | Genetics

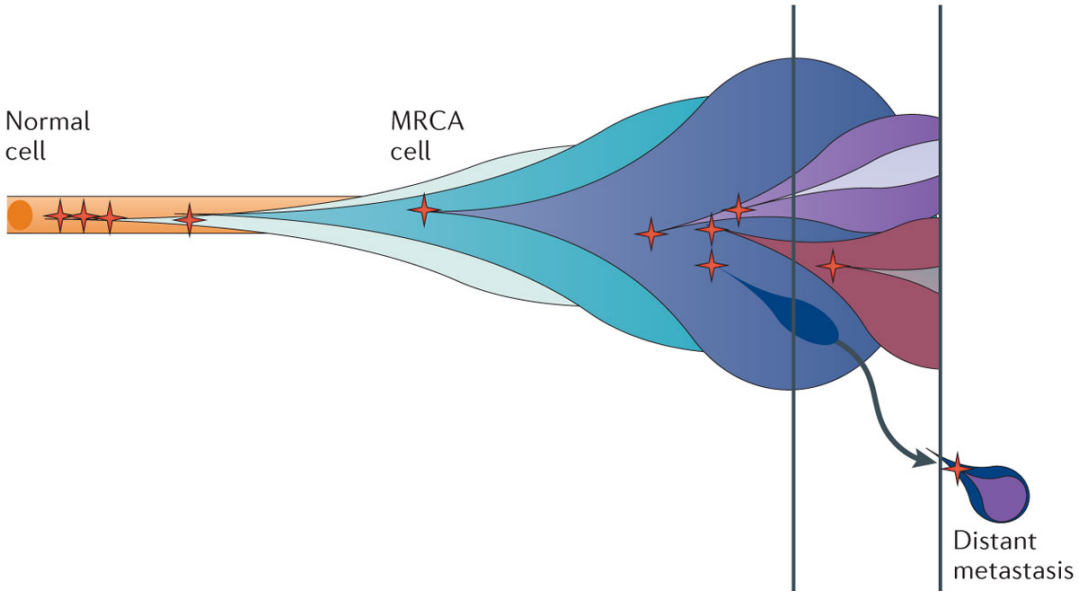
Evolutionary concepts and definitions

- **Clone**: A group of tumor cells that shares a highly similar genotype and mutational profile
- **Subclone**: A group of tumor cells that diverged from an ancestral clone by acquiring additional mutations



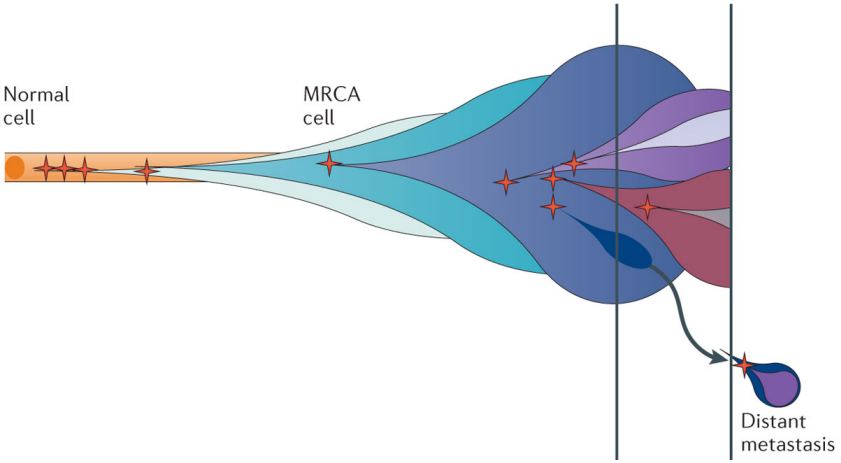
Evolutionary concepts and definitions

- **Clonal expansion**: Process in which one genotype with higher fitness expands in frequency in the tumor mass.
- **Selective sweep**: Process in which a genotype with a very high fitness emerges and outcompetes all other clones in the tumor



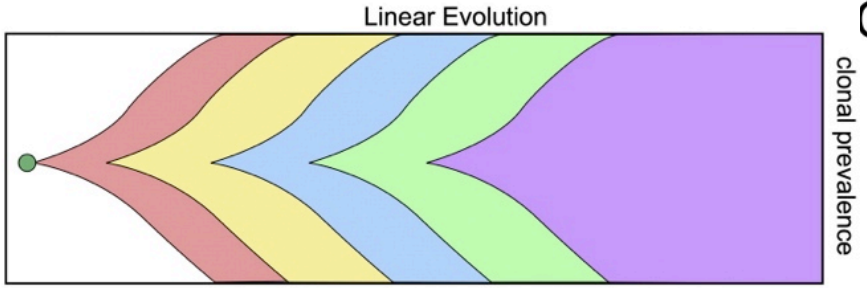
Types of mutations

- **Driver mutations** confer a fitness advantage
- **Passenger mutations** have no effect on fitness
- **Truncal mutations**: Ancestral mutations in the trunk of the phylogenetic tree that are shared by all clones
- **Subclonal mutations**: mutations in a lineage that has diverged from the trunk.



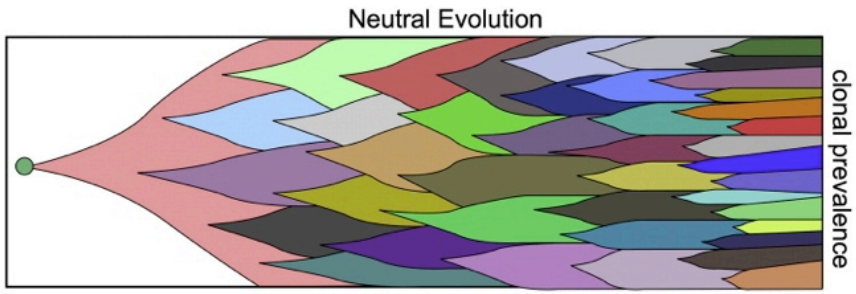
Models of tumour evolution

A



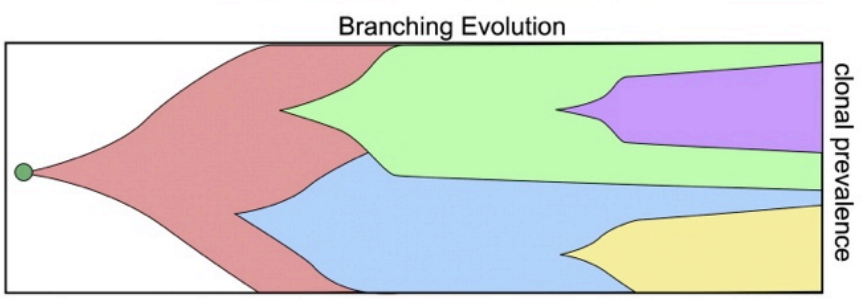
- Cancer progression models (e.g. colorectal cancer)

C

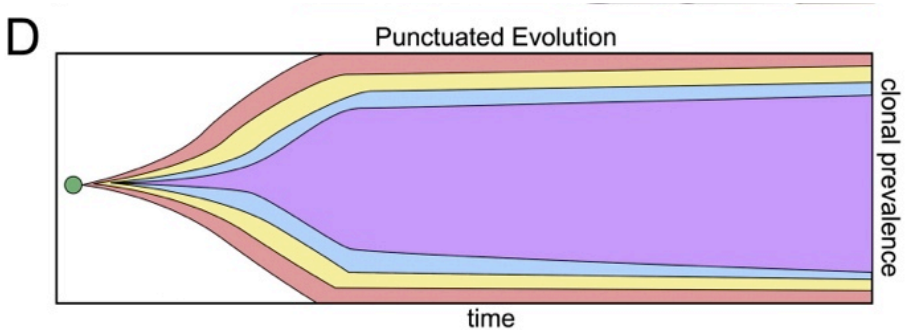


- E.g. Luria-Delbrück experiment, bacteria growth before infection

B

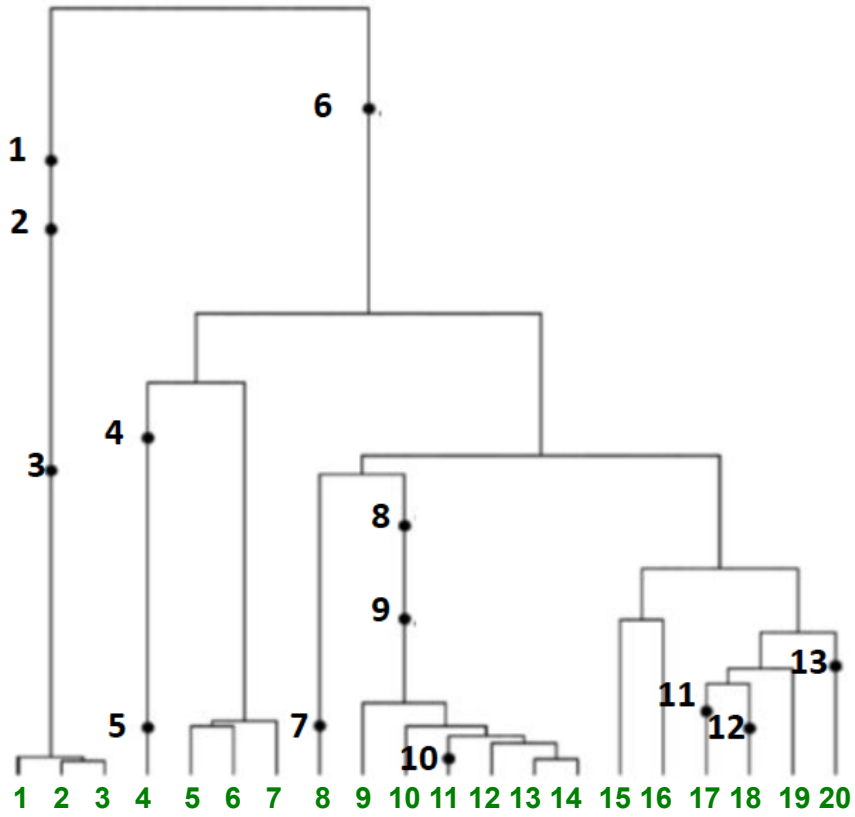


D



Figures adapted from Davis et al. 2017

The cells of a tumour form a genealogy



- Binary leaf-labeled tree T
- The leafs/tips represent cells
- Mutations occur at tree edges
- Relation between cells j and mutations i are described by a binary matrix M

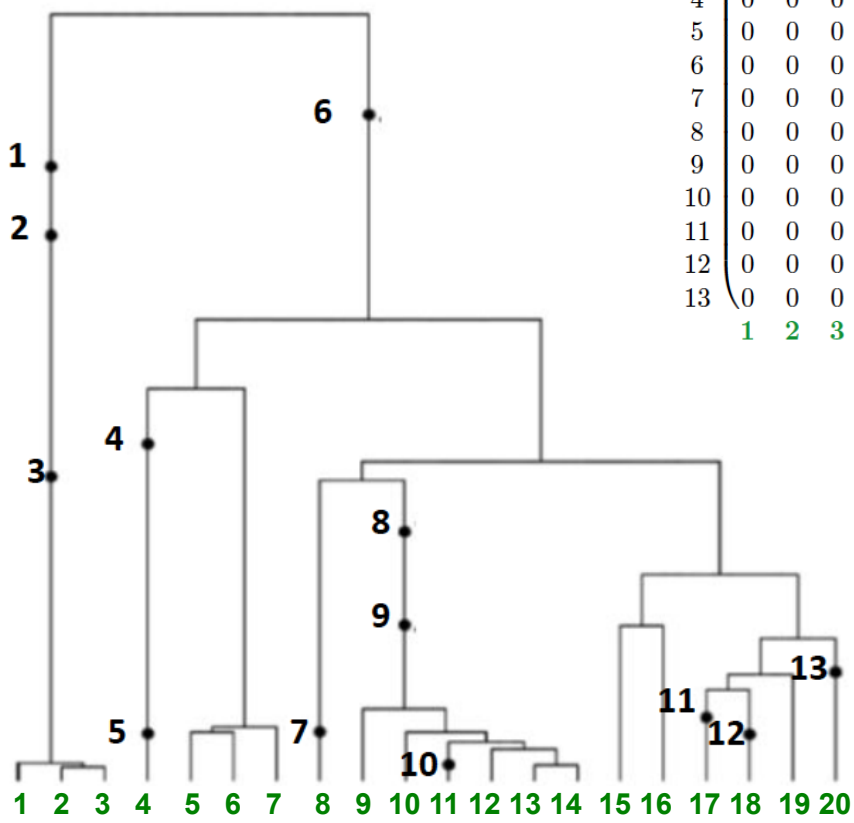
$$M_{ij} = \begin{cases} 1 & \text{if } j \text{ is located below } i \text{ in } T \\ 0 & \text{otherwise} \end{cases}$$

- The mutation frequencies correspond to the row sums

$$\text{freq}(i) = \sum_j M_{ij}$$

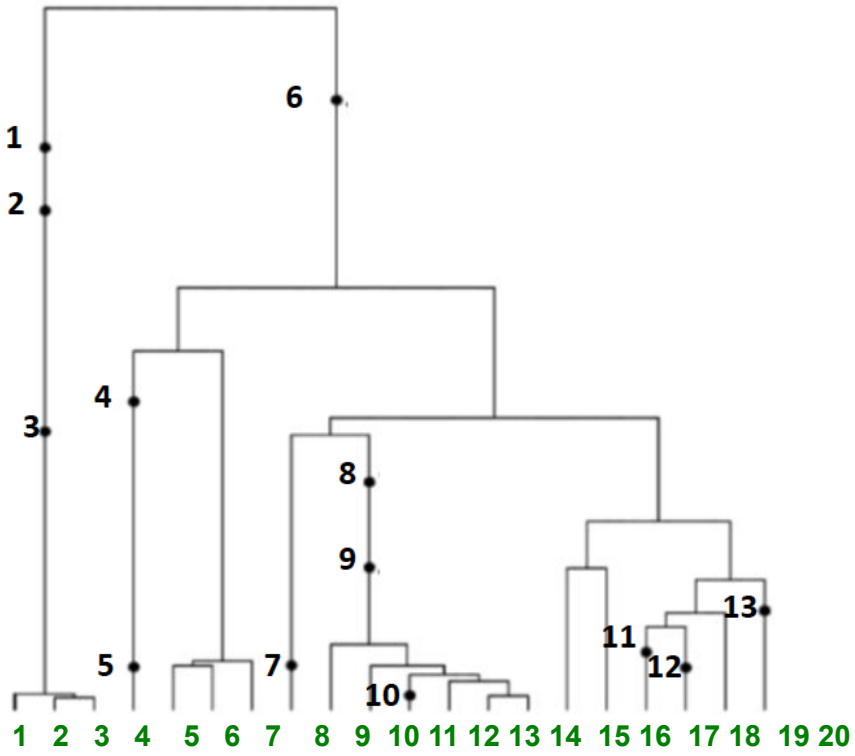
The cells of a tumour form a genealogy

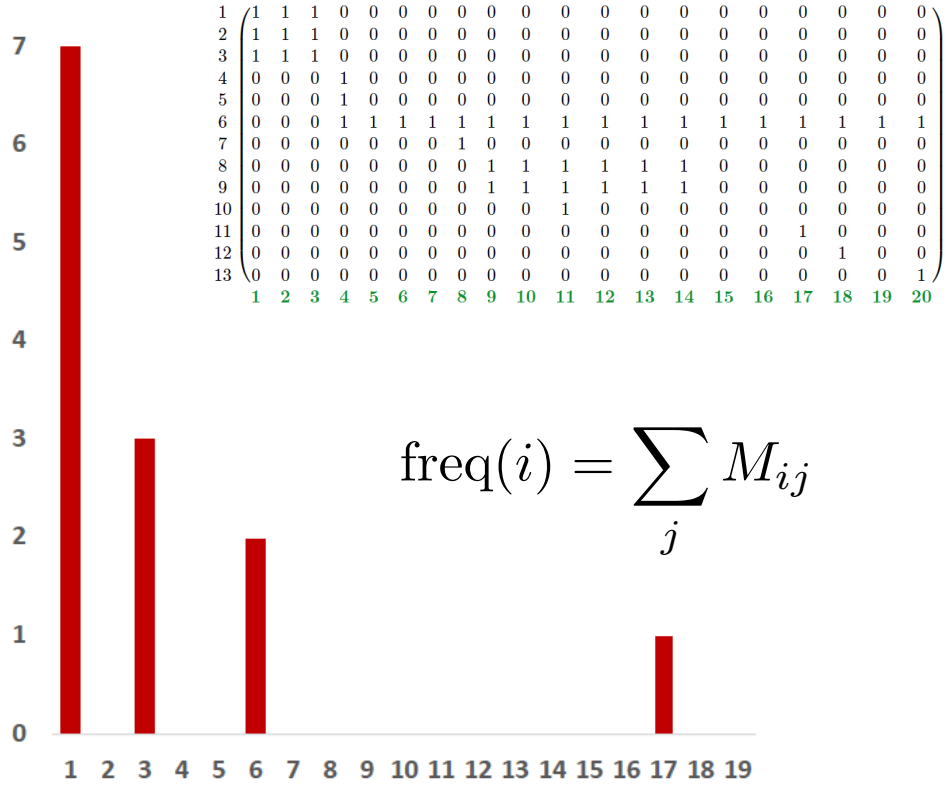
$$
\begin{array}{c}
1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13
\end{array}
\begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\begin{array}{cccccccccccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20
\end{array}
$$



Mutation Frequency Spectrum

- The mutation frequency spectrum is very similar to the data we obtain from sequencing tumour samples



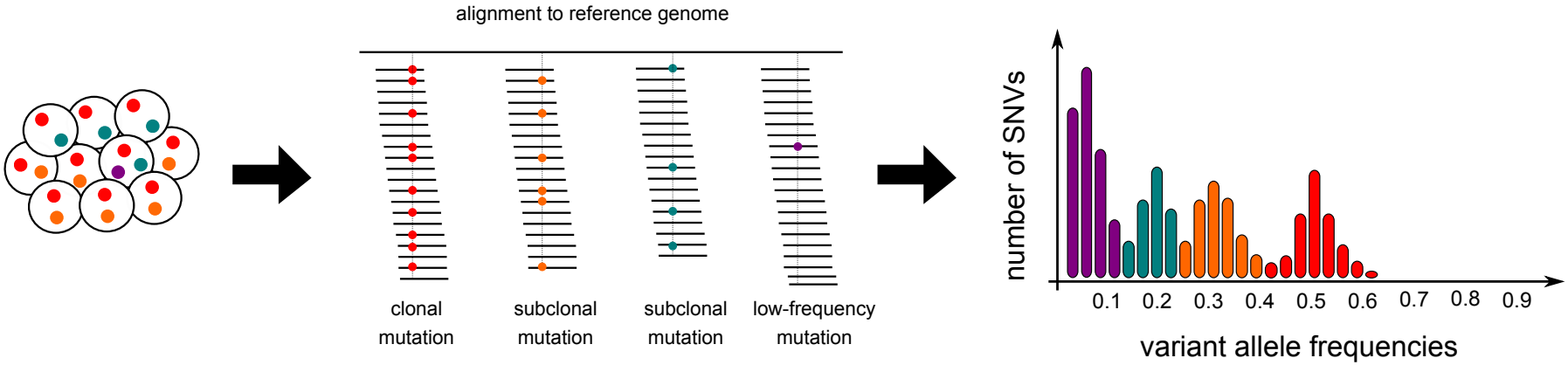


$$\begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

$$\text{freq}(i) = \sum_j M_{ij}$$

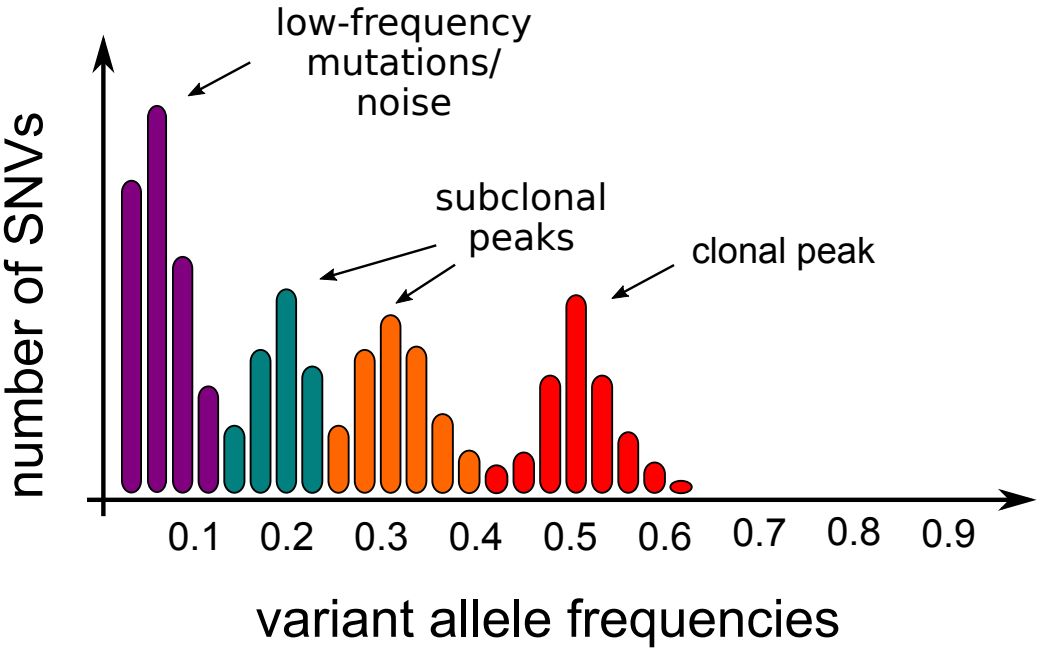
Bulk sequencing of tumour samples

- Most data is not from single-cells but from a mixture of cells
- Bulk sequencing: The DNA from all cells in a tumour sample (10^5 to 10^6 cells) is aggregated and sequenced together
- The sample can be a mixture of different clones/subclones



- A clonal heterozygous mutation can be expected to have a variant allele frequency of about 50%

Variant Allele Frequency spectrum

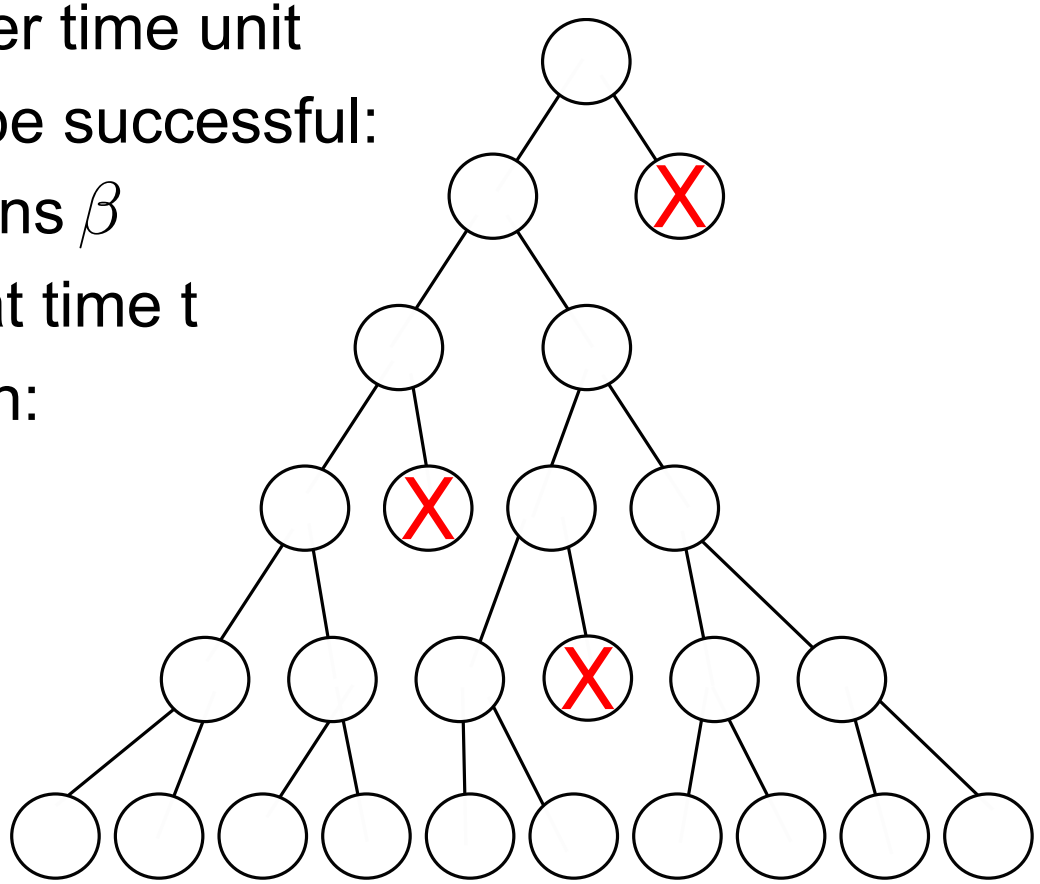


- Summary statistics of the mutation frequencies in a bulk sample (10^5 to 10^6 cells)

A simple model for tumour growth

- A tumour is founded by a single cell
- Cells divide with rate λ per time unit
- Not all cell divisions will be successful:
  rate of successful divisions β
- Number of tumour cells at time t
  under exponential growth:

$$N(t) = e^{\lambda \beta t}$$



Mutations under the infinite sites assumption

- Suppose mutations occur during cell division at rate μ
- If the genome is very long, we can assume that it has an infinite number of sites (loci)
- Then all mutations happen at a different nucleotide site

.... AGTTCTATGCGTAGCTGACATGCTGACATTAGCAAGTTCGAT …

- Mutations never get lost (no back mutations)
- The infinite sites model is appropriate for long DNA sequences under neutral evolution.
- Diploid human genome: $2 \times 3.2\ 10^9 = 6.4\ 10^9$ sites
- Ploidy $\pi = 2$ (i. e. #chromosome sets in cell)

Mutations in the neutral model of tumour evolution

- Assumptions:
  - Founding cell has acquired all mutations that give fitness advantage
  - Subclonal mutations are neutral
- Expected number of new mutations per time interval

$$\frac{dM(t)}{dt} = \mu\pi\lambda N(t)$$

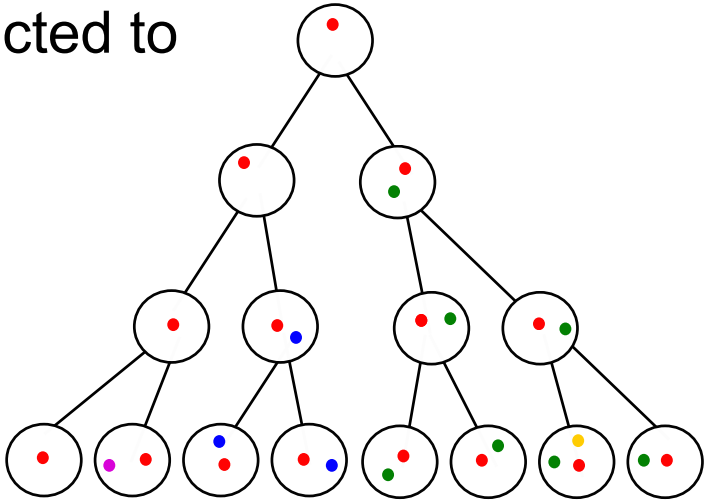
- Total number of subclonal mutations accumulating in time interval [t_0,t]

$$M(t) = \mu\pi\lambda \int_{t_0}^{t} N(t)dt = \frac{\mu\pi}{\beta}\left(e^{\lambda\beta t} - e^{\lambda\beta t_0}\right)$$

- For t_0=0 this corresponds to the Luria-Delbrück model

Connecting mutation age to mutation frequency

- Parameters μ, λ, β and tumour age t cannot be directly measured
- We only observe mutation frequencies
- A mutation arising in a tumour of 100 cells will have a cellular fraction of f = 1/100
- In absence of selection (and substantial genetic drift) the allelic fraction of the mutation can be expected to remain constant
- After expansion to 1000 cells:
  f= 10/1000 = 1/100



Connecting mutation age to mutation frequency

- Allelic frequency f of a mutation arising at time point t is the inverse of the number of alleles in the population at time t:

$$f = \frac{1}{\pi N(t)} = \frac{1}{\pi e^{\lambda \beta t}} \qquad f_{max} = \frac{1}{\pi N(t_0)} = \frac{1}{\pi e^{\lambda \beta t_0}}$$

- In a diploid tumour, t_0=0 corresponds to f_{max} = 0.5 (expected variant allele frequency of clonal variants)
- We can express N(t) and N(t_0) in terms of f

$$N(t) = e^{\lambda \beta t} = \frac{1}{\pi f} \qquad N(t_0) = e^{\lambda \beta t_0} = \frac{1}{\pi f_{\max}}$$

Estimating the mutation rate from the VAF spectrum

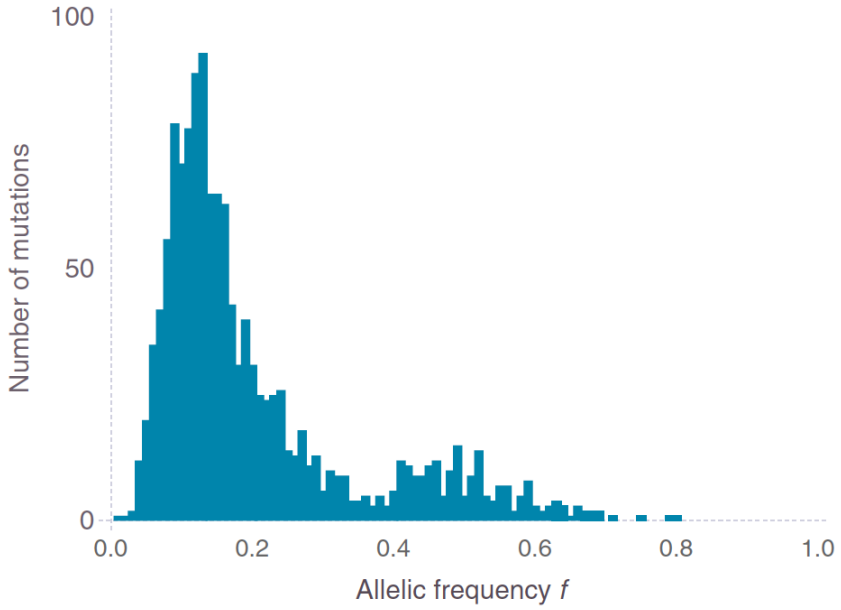
- Substituting t for f gives the cumulative number of mutations with frequency f or higher:

$$M(t) = \frac{\mu\pi}{\beta}(e^{\lambda\beta t} - e^{\lambda\beta t_0})$$

$$M(f) = \frac{\mu\pi}{\beta}(\frac{1}{\pi f} - \frac{1}{\pi f_{\max}})$$

$$= \frac{\mu}{\beta}(\frac{1}{f} - \frac{1}{f_{\max}})$$

- M(f) can be obtained from the VAF spectrum from bulk sequencing
- Then we can estimate the mutation rate per effective cell division $\mu_e = \frac{\mu}{\beta}$



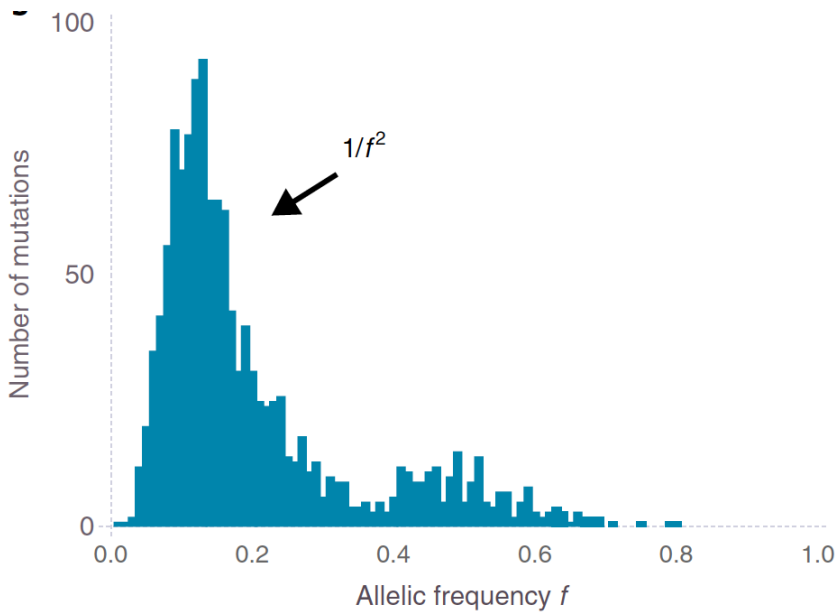
Estimating the mutation rate from the VAF spectrum

- With a change of parameter from t to f in dM/dt, we obtain for the expected number of mutations per frequency interval

$$\frac{dM}{dt} = \mu\pi\lambda\, e^{\lambda\beta t}$$

$$\frac{dM}{df} = \mu\pi\lambda\, \frac{1}{f}$$

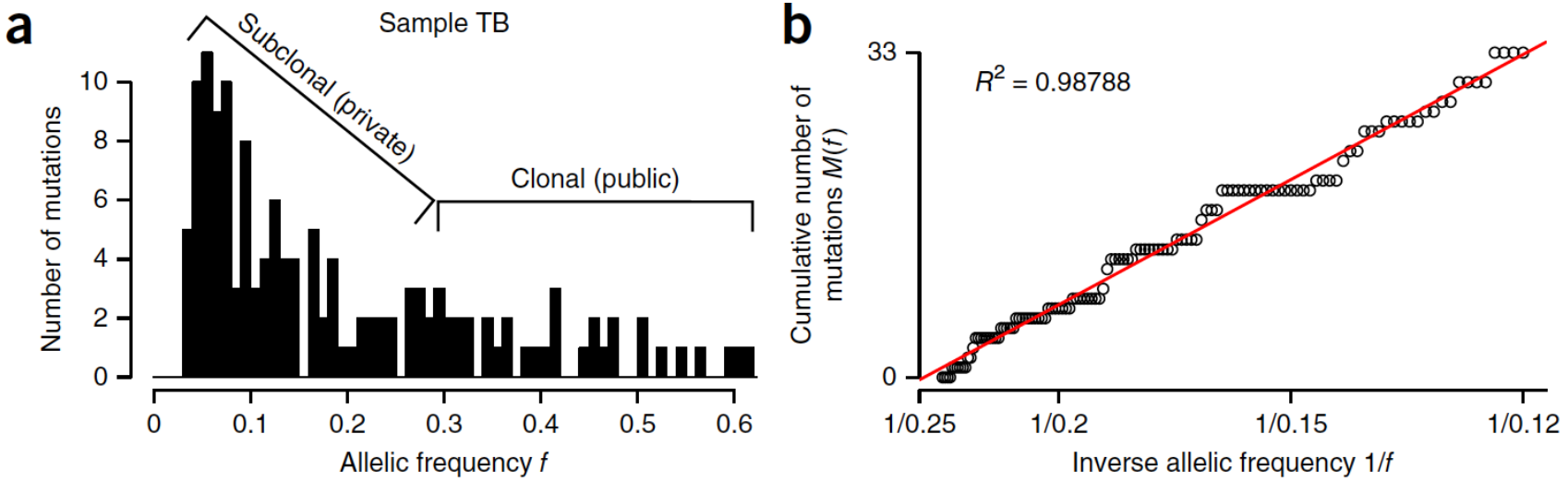
$$= (-1)\, \mu\pi\lambda\, \frac{1}{f^2}$$



- This corresponds to the slope in the neutral part of the VAF spectrum

How can we know if a tumour evolves neutrally?

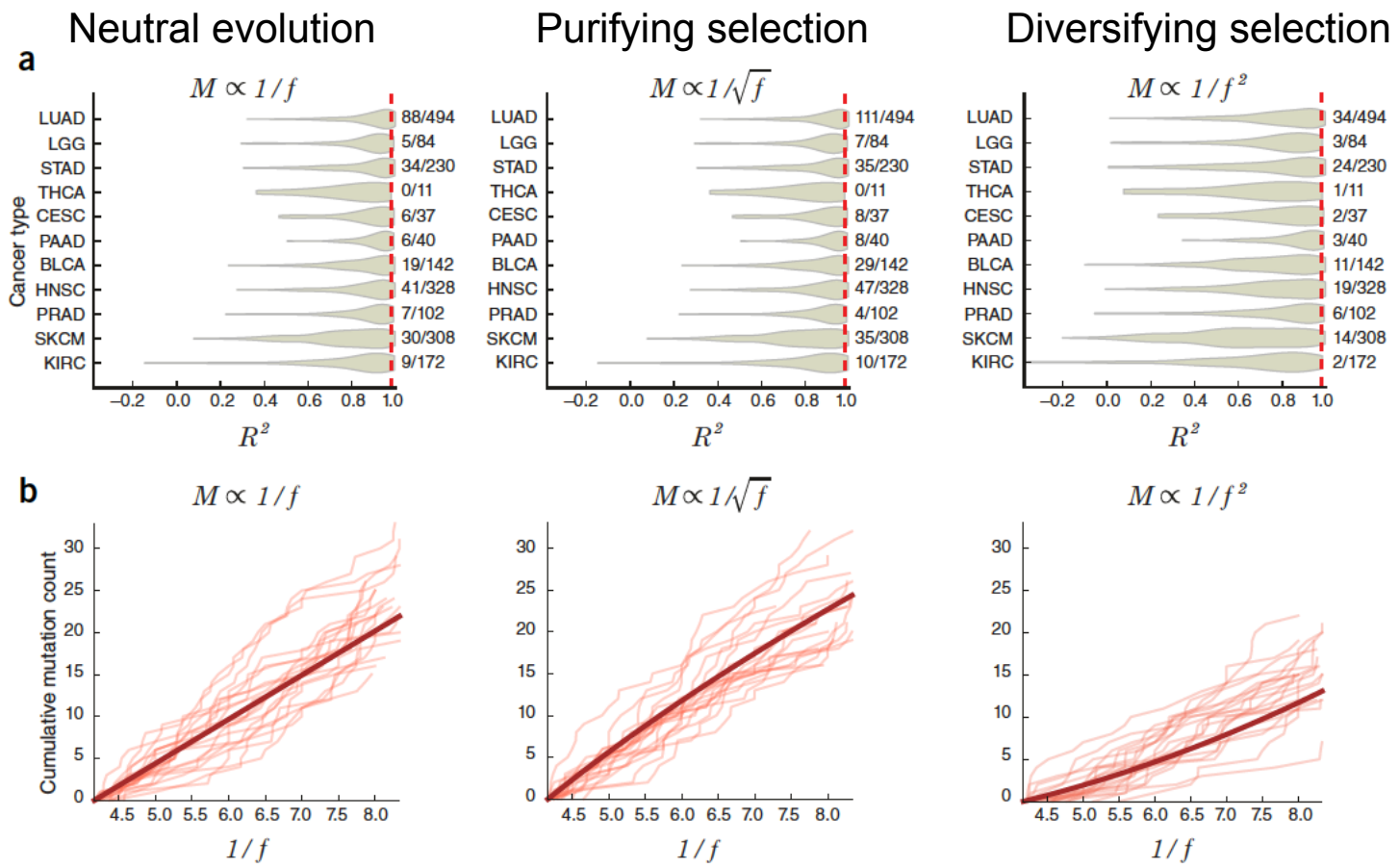
- Neutral evolution: Null Hypothesis of tumour evolution
- We know when a tumour does not evolve neutrally
- What if the VAF spectrum matches neutral evolution?
- Example: Case of colorectal cancer from TCGA database



The neutral evolution controversy

- Williams et al. 2016: studied 904 cancers, 1/3 have 1/f tail in VAF spectrum ➔”*Neutral evolution common in cancer*”
- Objections by fellow researchers:
  - Logical fallacy: The data matching the null hypothesis does not equate null hypothesis being true
  - Showed empirically that other evolutionary models can create similar results
  - M(f) cannot be accurately estimated from VAF spectra
  - ….

High R^2 values are consistent with other evolutionary models



Noorbakhsh et al., 2018

The neutral evolution controversy

Neutral tumour evolution debate

- *Reply to 'Neutral tumor evolution?'* Heide T.*, Zapata L.*, Williams M.J.*, Werner B.*, Barnes C.P., Graham T.A.§, Sottoriva A.§ **Nature Genetics**, 2018, doi:10.1038/s41588-018-0256-z. *Equal contribution. §Co-correspondent.
- *Neutral tumor evolution?* Tarabichi M., Martincorena I., Gerstung M., Markowetz F., Spellman P.T., Morris Q.D., Lingjaerde O.C., Wedge D.C., van Loo P. **Nature Genetics**, 2018, 10.1038/s41588-018-0258-x.
- *Reply to 'Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution'.* Werner B., Williams M.J., Barnes C.P., Graham T.A.§, Sottoriva A.§ **Nature Genetics**, 2018, 10.1038/s41588-018-0235-4. §Co-corresponding.
- *Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution.* McDonald T.O., Chakrabarti S., Michor F. **Nature Genetics**, 2018, 10.1038/s41588-018-0217-6.
- *Reply to 'Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data'.* Williams M.J.*, Werner B.*, Heide T., Barnes C.P., Graham T.A.§, Sottoriva A.§ **Nature Genetics**, 2018, 10.1038/s41588-018-0210-0. *Equal contribution. §Co-corresponding.
- *Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data.* Balaparya A., De S. **Nature Genetics**, 2018, 10.1038/s41588-018-0219-4.
- *Reply: Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures.* Williams M.J.*, Werner B.*, Barnes C.P., Graham T.A.§, Sottoriva A.§ **Nature Genetics**, 2017, 49:1289-1291. *Equal contribution. §Co-corresponding.
- *Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures.* Noorbakhsh J., Chuang J.H. **Nature Genetics**, 2017, 49:1288-1289.
- *Reply: Is the evolution in tumors Darwinian or non-Darwinian?* Williams M.J.*, Werner B.*, Barnes C.P., Graham T.A.§, Sottoriva A.§ **National Science Review**, 2018, 0:1-3, doi: 10.1093/nsr/nwx131. *Equal contribution. §Co-corresponding.
- *Is the evolution in tumors Darwinian or non-Darwinian?* Wang H., Chen Y., Tong D., Ling S., Hu Z., Tao Y., Lu X., Wu C. **National Science Review**, 2018, 0:1-3, doi: 10.1093/nsr/nwx076.

From other groups:

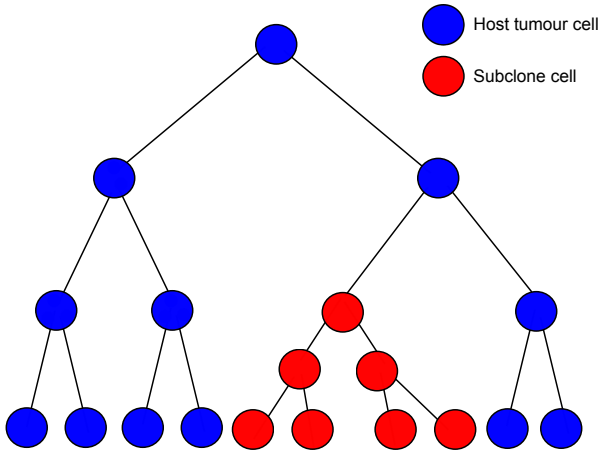
- *Neutral theory and the somatic evolution of cancer.* Cannataro V.L., Townsend J.P. **Molecular Biology and Evolution**, 2018, 36(6):1308-1315.
- *Neutral theory in cancer cell population genetics.* Niida A., Iwasaki W.M., Innan H. **Molecular Biology and Evolution**, 2018, 36(6):1316-1321.

For a nice recent perspective and overview of the neutral evolution debate in evolutionary biology see recent issue of Molecular Biology and Evolution 2018, Volume 35, Issue 6.

http://www.sottorivalab.org/neutral-evolution-debate.html

A model of tumour evolution with selection

- Assume we have two cell populations (host tumour and subclone)
- Both populations grow exponentially at rates $\lambda_{sub} \geq \lambda_{host}$
- The relative fitness advantage of the subclone is $s = \frac{\lambda_{sub} - \lambda_{host}}{\lambda_{host}}$
- s=1 means the subclone grows twice as fast as the host tumour
- s=0 means no selective advantage

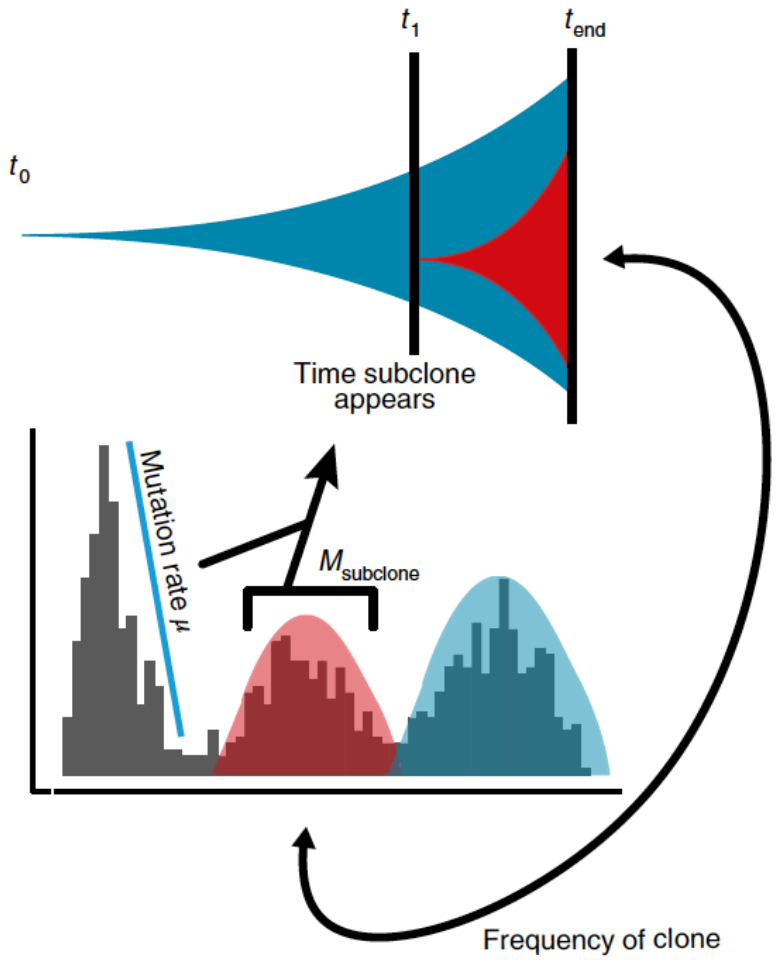


- Then the fitness of the subclone is

$$1 + s = \frac{\lambda_{sub}}{\lambda_{host}}$$

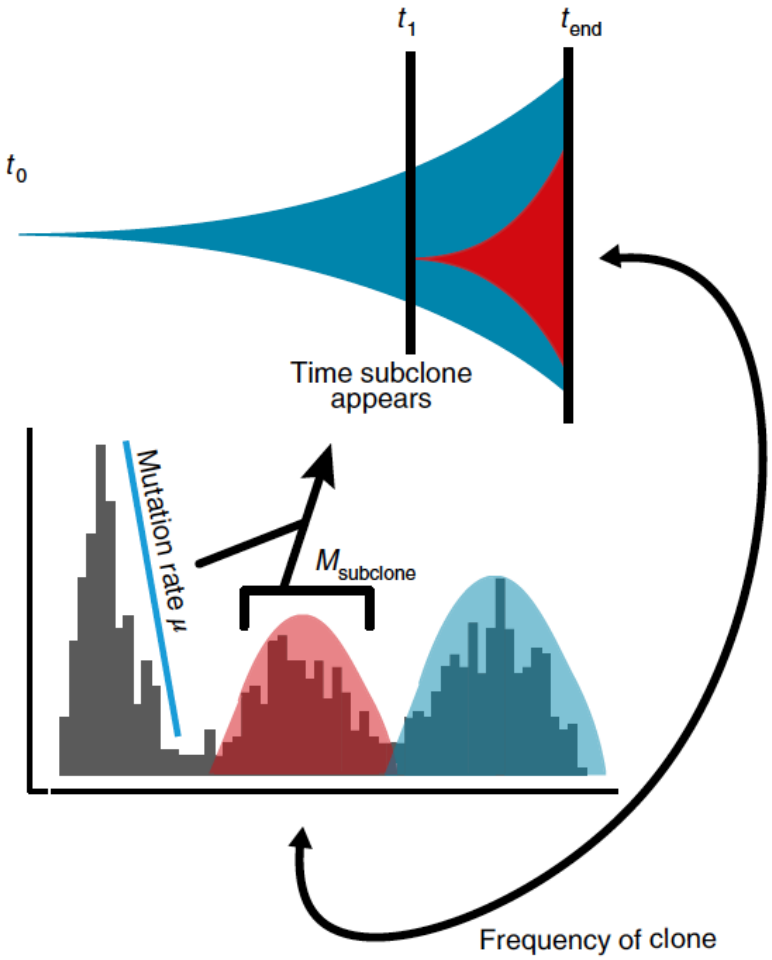
- We cannot directly measure s

Measuring properties of the subclone from the VAF spectrum



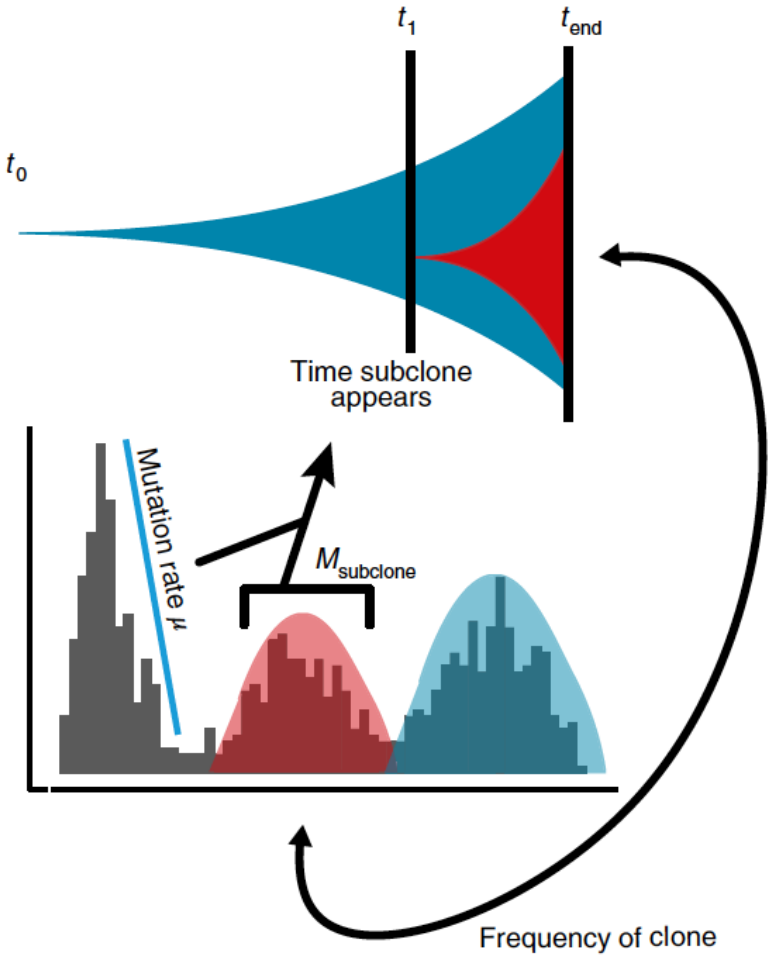
- As before the mutation rate μ can be estimated from the neutral peak
- The subclone frequency f_{sub} can be estimated from the mean of the subclone peak
- Example: A peak at 0.2 indicates that 40% of the tumour cells are from the subclone
- The number of mutations in the subclone M_{sub} at time t_1 can be estimated from the area of the VAF cluster

Measuring properties of the subclone from the VAF spectrum



- Why does M_{sub} denote the number of mutations the subclone acquired between t_0 and t_1?
- The subclone cluster in the VAF spectrum consists primarily of these mutations
- Mutations arising later in the subclone have a lower frequency, as they will not be shared by all subclone cells
- Note: The variant allele frequency of the M_{sub} mutations is higher than expected under the neutral model

Estimating the subclone age from M_{sub}



- We can express M_{sub} in terms of μ and Γ the mean number successful cell divisions between t_0 and t_1

$$M_{\text{sub}} = \mu\Gamma$$

- We can further relate Γ to the time in terms of number of population doublings

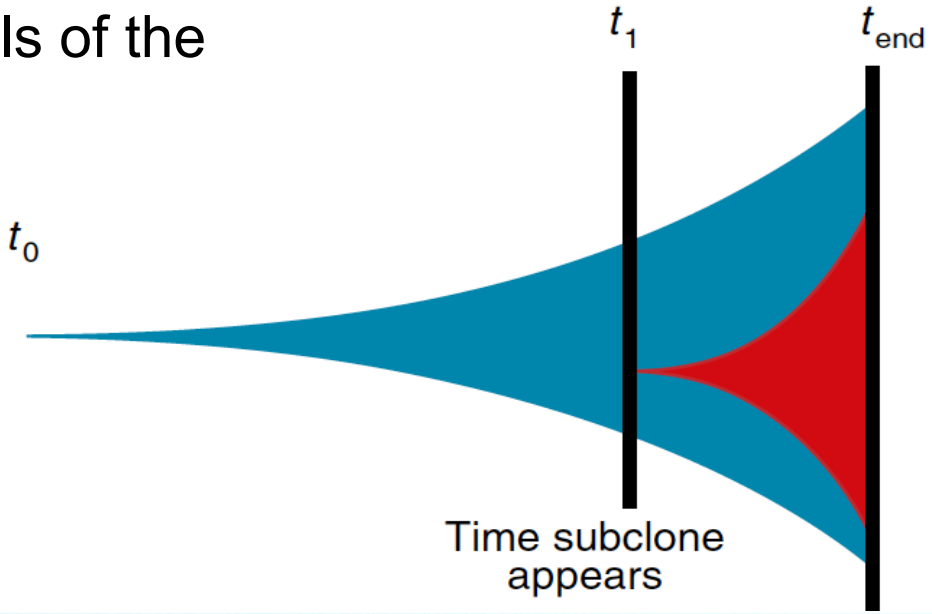
$$\Gamma = 2\log(2)t_1$$

(See Williams et al. 2018 for formal derivation)

- Since M_{sub} and μ can be measured from the VAF spectrum, this gives an estimate of the subclone age (in terms of population doublings)

Estimating the fitness advantage of a subclone

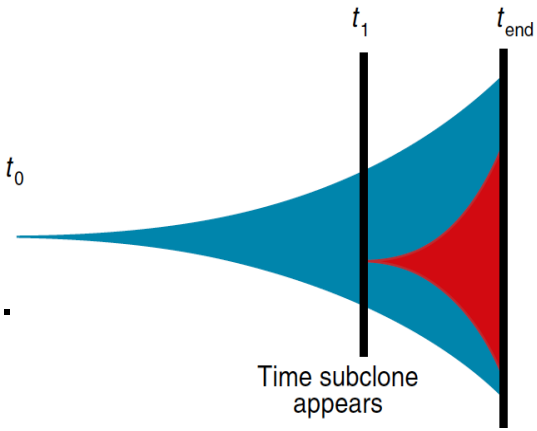
- We estimate s based on the observed subclone frequency at t_{end} and the estimated subclone age
- Assume mutant subclone was founded by a single cell at t_1
- For s>1, the frequency of the subclone will increase over time
- Let $N_{host}(t)$ be the number of cells of the host tumour population at t
- Let $N_{mut}(t)$ be the number of cells in the subclone at t



Estimating the fitness advantage of a subclone

- The frequency of the subclone at t_{end} will be

$$f_{\mathrm{sub}}(t_{\mathrm{end}}) = \frac{N_{\mathrm{sub}}(t_{\mathrm{end}}-t_1)}{N_{\mathrm{sub}}(t_{\mathrm{end}}-t_1)+N_{\mathrm{host}}(t_{\mathrm{end}})}$$



- We use the shorthand $\lambda = \lambda_{host}$ in the following.
- Assuming exponential growth, we get

$$f_{\mathrm{sub}}(t_{\mathrm{end}}) = \frac{e^{\lambda(1+s)(t_{\mathrm{end}}-t_1)}}{e^{\lambda(1+s)(t_{\mathrm{end}}-t_1)}+e^{\lambda t_{\mathrm{end}}}-e^{\lambda(t_{\mathrm{end}}-t_1)}}$$

- The term $-e^{\lambda(t_{\mathrm{end}}-t_1)}$ corrects for the host cell that founded the fitter subclone

Estimating the fitness advantage of a subclone

$$f_{\text{sub}}(t_{\text{end}}) = \frac{e^{\lambda(1+s)(t_{\text{end}}-t_1)}}{e^{\lambda(1+s)(t_{\text{end}}-t_1)}+e^{\lambda t_{\text{end}}}-e^{\lambda(t_{\text{end}}-t_1)}}$$

- Taking $e^{\lambda t_{\text{end}}}$ out of each term, we obtain

$$f_{\text{sub}}(t_{\text{end}}) = \frac{e^{\lambda s(t_{\text{end}}-t_1)}e^{-\lambda t_1}}{e^{\lambda s(t_{\text{end}}-t_1)}e^{-\lambda t_1}+1-e^{-\lambda t_1}}$$

- Since $e^{-\lambda t_1} \ll 1$ even for moderate t_1, we neglect the term

$$f_{\text{sub}}(t_{\text{end}}) = \frac{e^{\lambda s(t_{\text{end}}-t_1)}e^{-\lambda t_1}}{e^{\lambda s(t_{\text{end}}-t_1)}e^{-\lambda t_1}+1}$$

- Solving for s gives an expression for the fitness advantage

Estimating the fitness advantage of a subclone

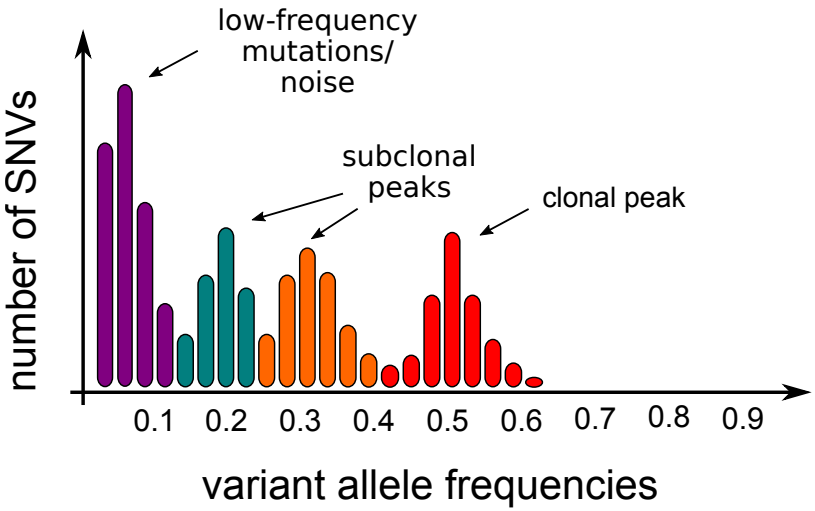
- Expression for the fitness advantage

$$s = \frac{\log(\frac{f_{\text{sub}}}{1-f_{\text{sub}}}) + \lambda t_1}{\lambda(t_{\text{end}} - t_1)}$$

- To calculate s, we need estimates for f_{sub}, t_1, t_{end} and λ
- We have estimates for f_{sub}, t_1 (in terms of population doublings)
- t_{end} can be estimated from the tumour size: Assume a tumour with 10^{10} cells (typical size for colon cancer), then $2^{t_{\text{end}}} = (1 - f_{\text{sub}})10^{10}$
- Finally since we measure time in population doubling time, we can simply set $\lambda = log(2)$

Confounders of the Variant-Allele-Frequency-Spectrum

- We measure quantities from the VAF spectrum
- We estimate model parameters from it using strong assumptions
- Our estimates will be affected if the assumptions are wrong



- Contamination with normal cells
  - Shifts VAF spectrum to the left
- Copy number changes
  - Gains of mutated alleles, losses of normal alleles shift mutations to the right (>0.5 possible)
- Mutation losses destroy connection between mutation frequency and age

- Using single-cell instead of bulk data circumvents many of these issues

Summary

- Intra-tumour heterogeneity/models of tumour evolution
- Our data comes from the Variant-Allele-Frequency-Spectrum
- Inference of the effective mutation rate under the neutral model
- Inference of the selective advantage of a fitter subclone
- Confounders of the Variant-Allele-Frequency-Spectrum

References

- Williams, Marc J., et al. "Identification of neutral tumor evolution across cancer types." *Nature genetics* 48.3 (2016): 238.
- Williams, Marc J., et al. "Quantification of subclonal selection in cancer from bulk sequencing data." *Nature genetics* 50.6 (2018): 895.
- Dinh, K. N., et al. "Statistical inference for the evolutionary history of cancer genomes." *bioRxiv* (2019): 722033.
- Noorbakhsh, Javad, and Jeffrey H. Chuang. "Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures." *Nature genetics* 49.9 (2017): 1288.
- Davis, Alexander, Ruli Gao, and Nicholas Navin. "Tumor evolution: Linear, branching, neutral or punctuated?." *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1867.2 (2017): 151-161.
- Yates, Lucy R., and Peter J. Campbell. "Evolution of the cancer genome." *Nature Reviews Genetics* 13.11 (2012): 795.
- Marusyk, Andriy, Vanessa Almendro, and Kornelia Polyak. "Intra-tumour heterogeneity: a looking glass for cancer?." *Nature Reviews Cancer* 12.5 (2012): 323.