Statistical inference for the evolutionary history of cancer genomes

Marek Kimmel, Rice University

kimmel@rice.edu

Khanh Dinh, Columbia U.

Roman Jaksik, Silesian Tech

Amaury Lambert, UPMC.

Simon Tavaré, Columbia U.

To appear in Statistical Science (available in Biorxiv)

Objective

Address two different approaches in modeling cancer cell proliferation and genome evolution:

Proliferation models based on **binary fission branching or linear birth and death** processes

Models based on population genetics paradigms: **Wright-Fisher or Moran**

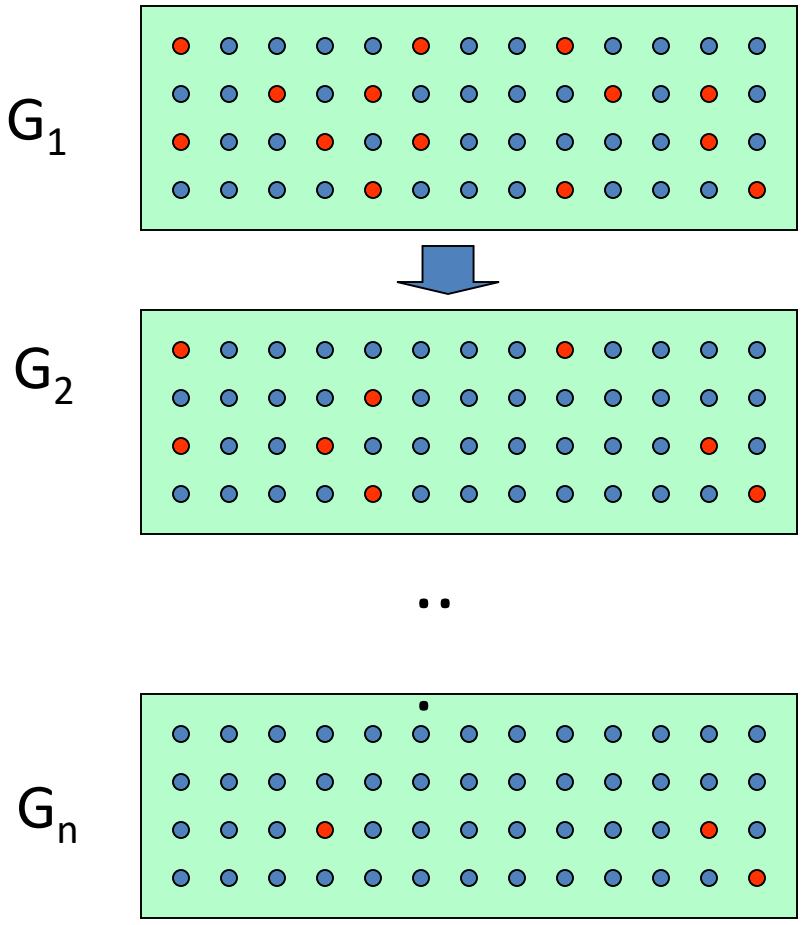
Genetic forces considered (in growing populations):

Drift

Mutation

Selective sweeps

Genetic drift



Alleles: A_1: A_2:

Replication = sampling with replacement

A_1 – becomes fixed

A_2 – becomes lost

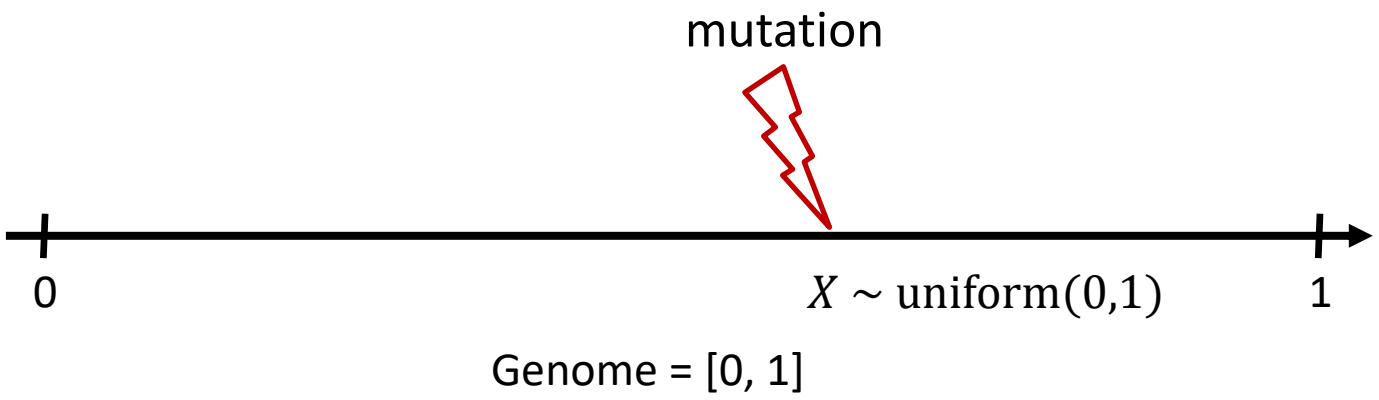
Mutation

Mutation times follow a **Poisson process** with intensity θ measured per locus (per site) per generation

We use:

- **Infinite Sites Model (ISM)**, where it is assumed that each mutation takes place at a DNA site that never mutated before

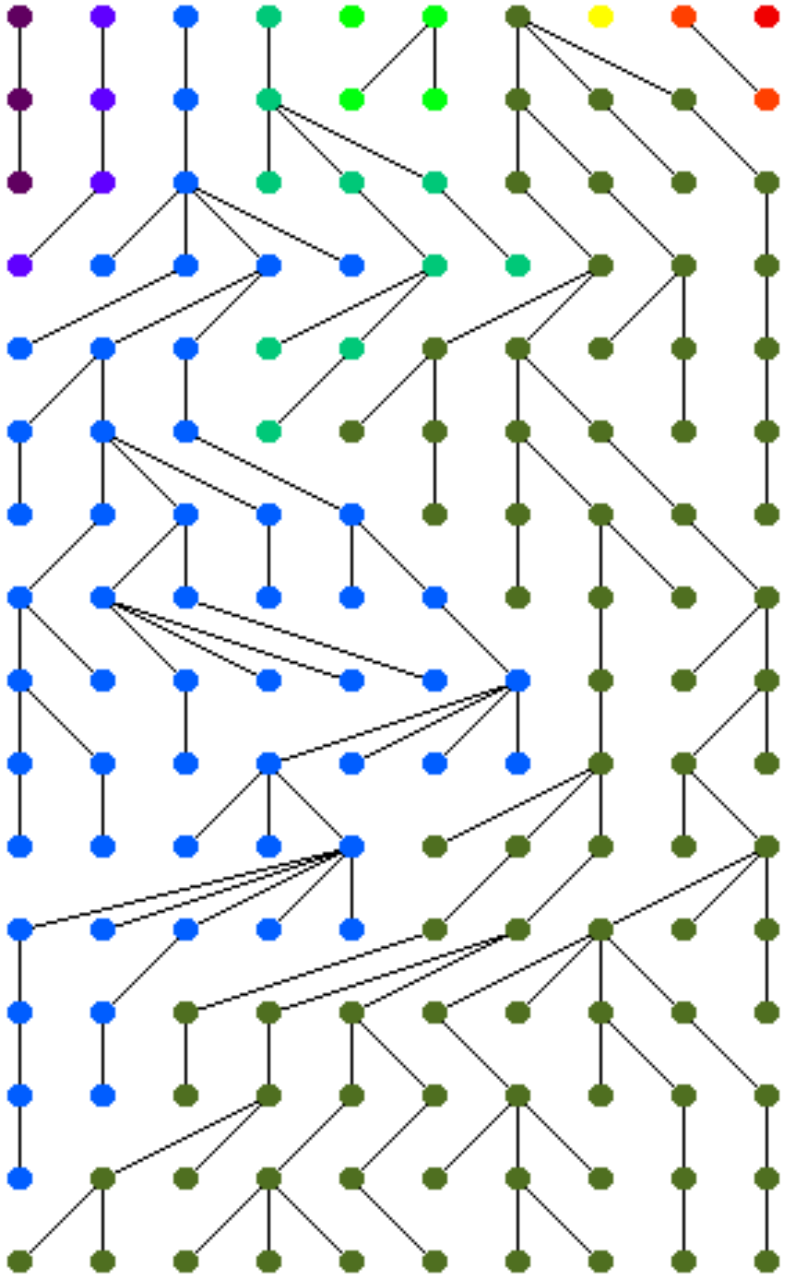
- Mutation sites can be represented as **iid uniform(0, 1) rv's**



Coalescence

Wright-Fisher seen in reverse time:

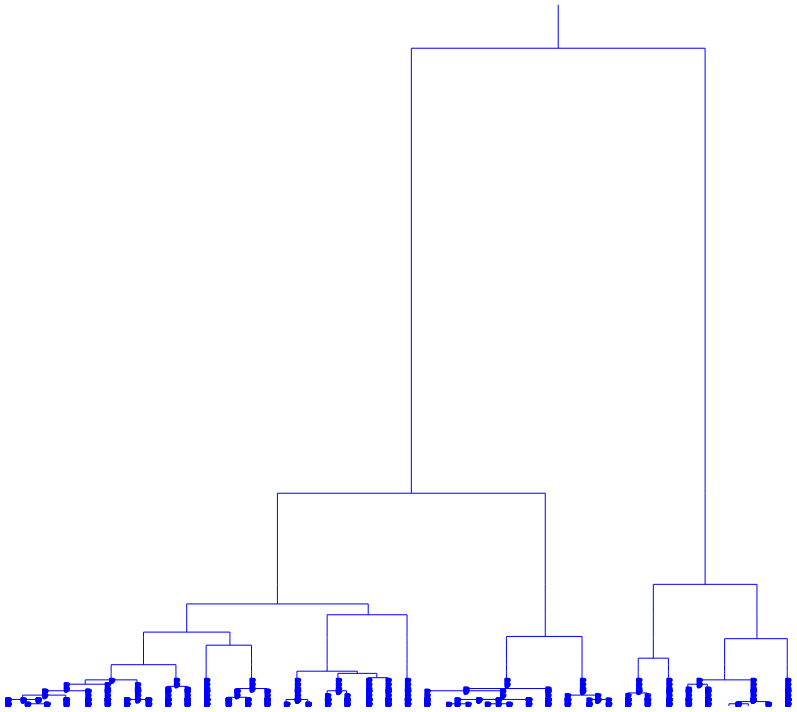
Eventually, all ancestry lines converge on common ancestor

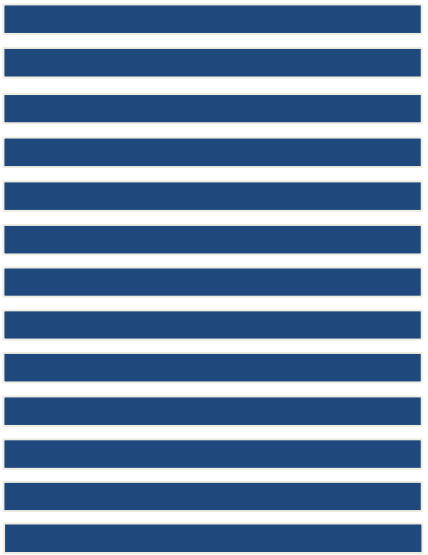


Constant population size scenario

Expected (reverse) time to first coalescence with sample size k is $\sim \binom{k}{2}^{-1}$

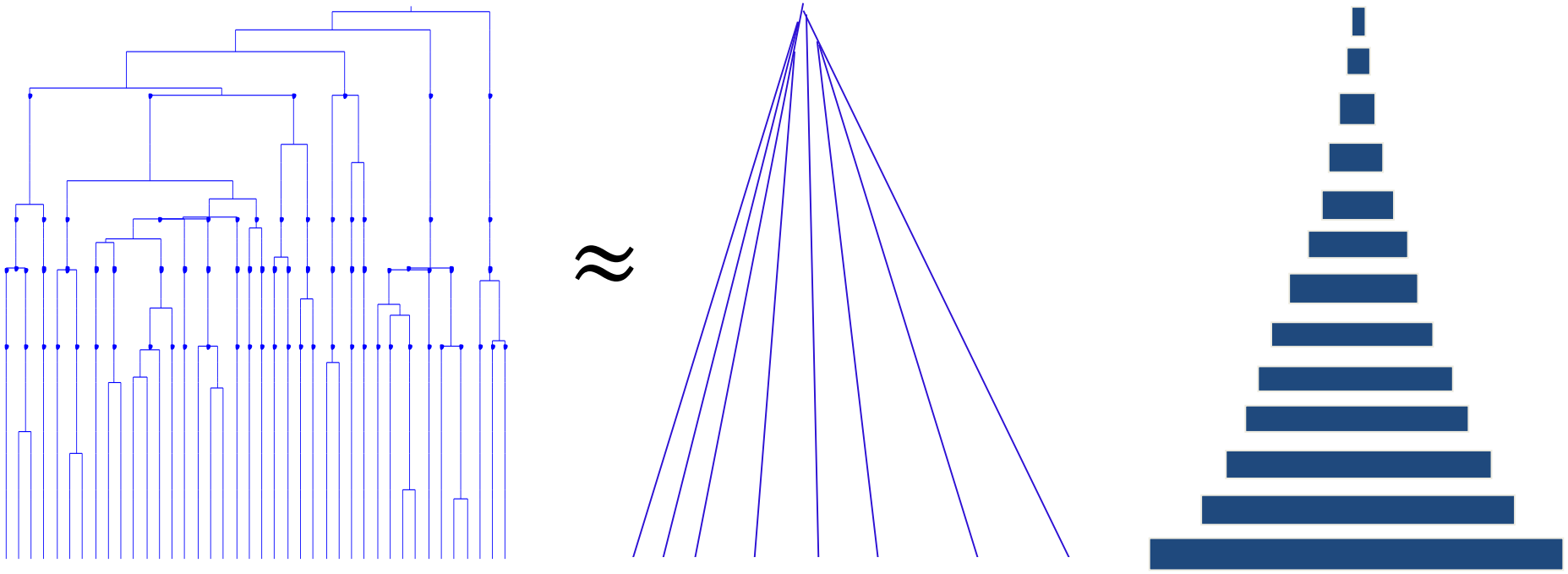
Hence, coalescent tree with deep “valleys”





Growing population scenario (≈ tumor growth case)

Coalescent tree more “star-like”



Coalescence method

- Genetic drift viewed in reverse time
- Estimating the past of an *n* - sample of sequences taken at present.
- Possible events that happen in the past are
  - **coalescences** (lineage merges) leading to common ancestors of sequences, and
  - **mutations** along branches of ancestral tree each at a new site ("Poisson rain")
  - **other**

GT-coalescent (Kingman, Tajima)

When population size is constant, the scaled pure-death time-continuous MC, counting the number of ancestors of the n-sample, at time t,

$$\{A_n(t), t \geq 0\}$$

has transition rates

$$q_{i,i-1} = \frac{i(i-1)}{2}$$

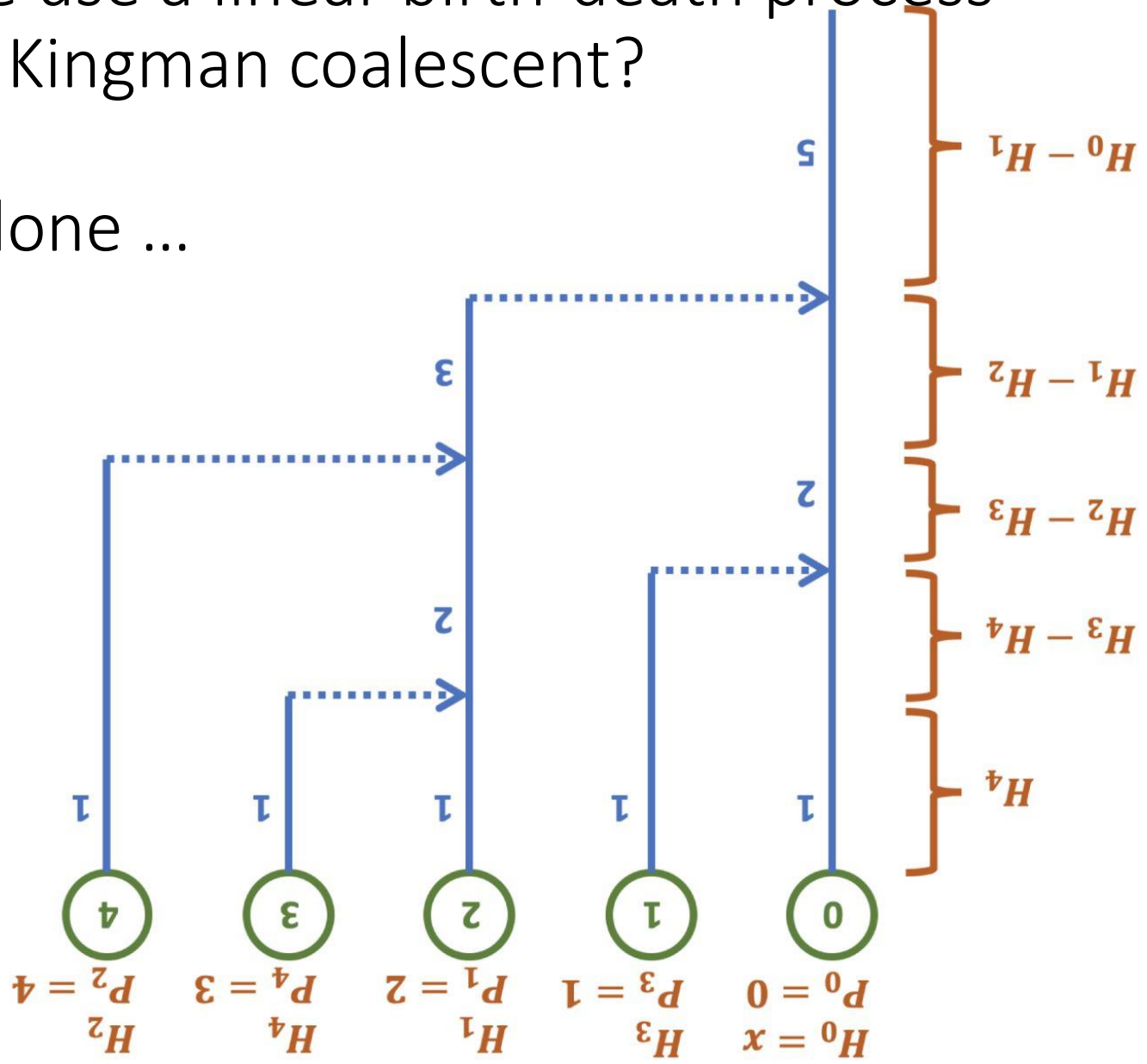
Inter-coalescence time intervals are independent and exponentially distributed.

Under exponential growth, it is enough to transform time deterministically

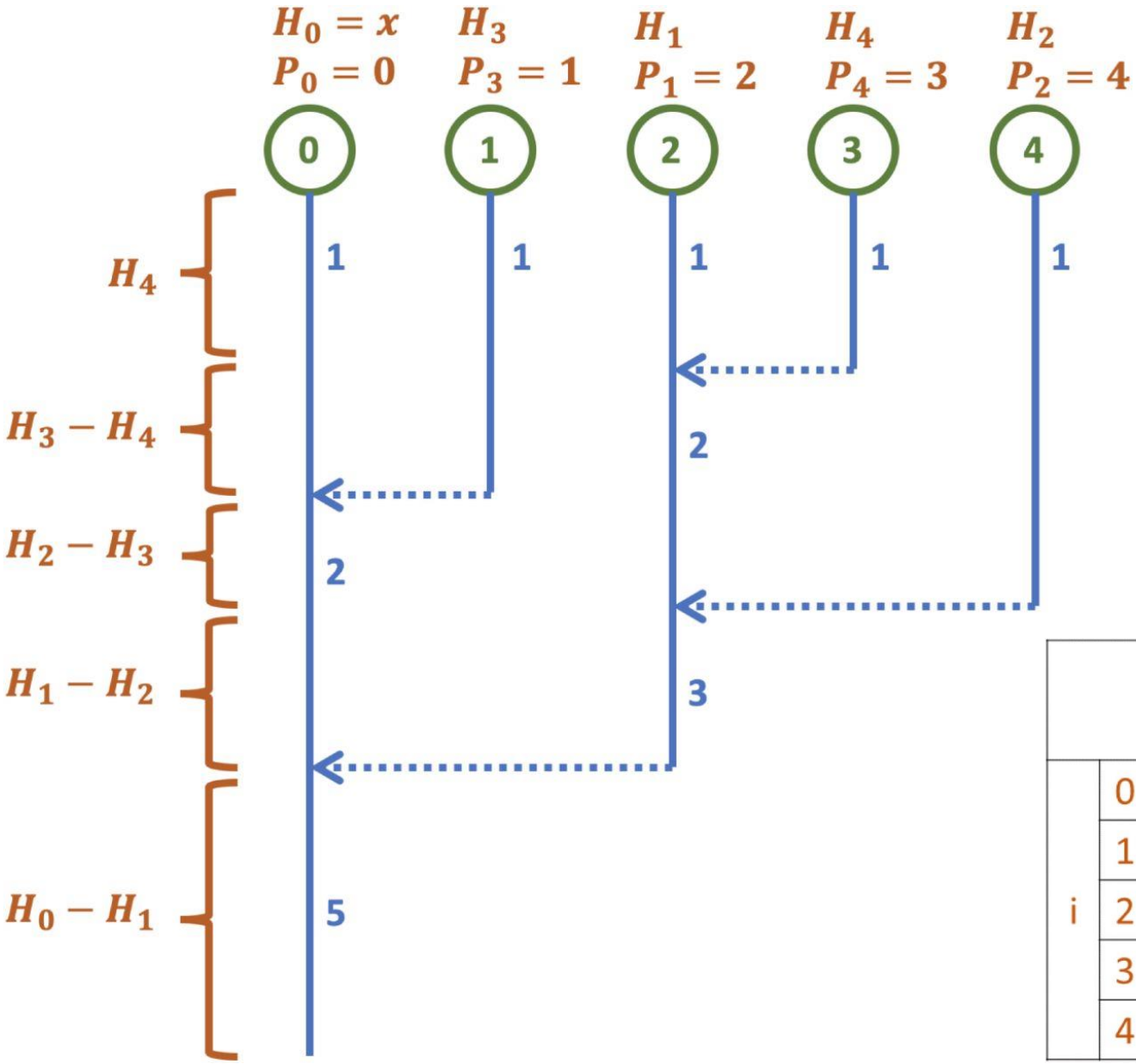
$$A_n^{\beta}(t) = A_n\left((e^{\beta t} - 1)/\beta\right), t \geq 0.$$

What if we use a linear birth-death process instead of Kingman coalescent?

It can be done …



It just has to be drawn correctly …



		j				
		0	1	2	3	4
i	0	1	1	1	1	1
	1	1	1	2	0	1
	2	2	0	2	0	1
	3	2	0	3	0	0
	4	5	0	0	0	0

Lambert (lbdp) coalescent

- We define random variables H_0, H_1, … as the consecutive coalescence times, contingent on representation of the tree in a **specific order** (**coalescent point process**).

- Following Lambert (2010), Theorem 5.4, $\{\boldsymbol{H_i}\}$, conditional on the tree having ***n*** **tips**, form a sequence of ***n*** **iid random** variables with tail $\boldsymbol{W(t)^{-1}}$ conditioned on being less than **tree depth *x***.

- Conversely, we can **generate the iid rv's and recreate the tree** using the drawing rules as in the small table earlier on

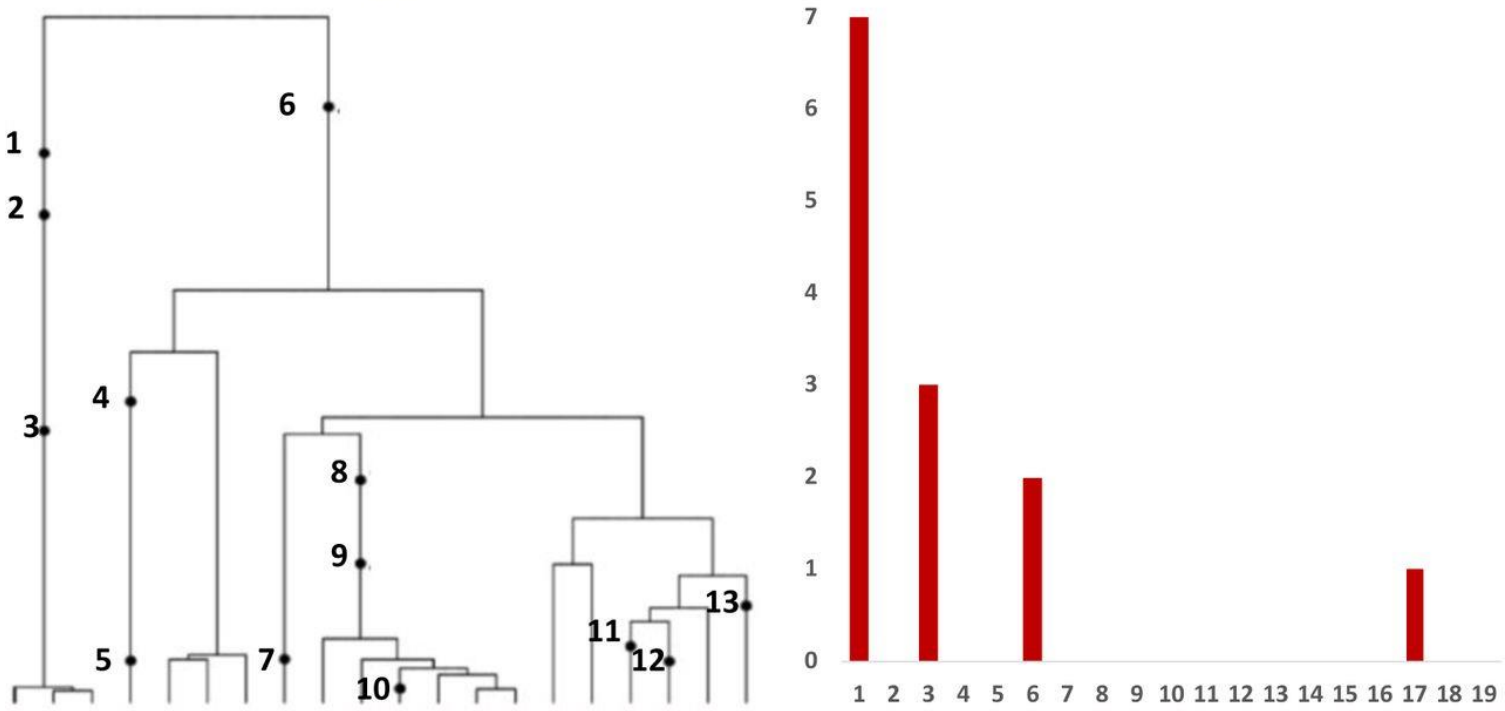
- For ordinary lbdp, $\boldsymbol{W(t) = \alpha + (1-\alpha)\exp(rt)}$, $\boldsymbol{\alpha = 1 - pb/r}$

Site Frequency Spectrum (SFS)
Most common statistic for equivalent mutation distribution
(neutral (?) "passenger" mutations in tumors)

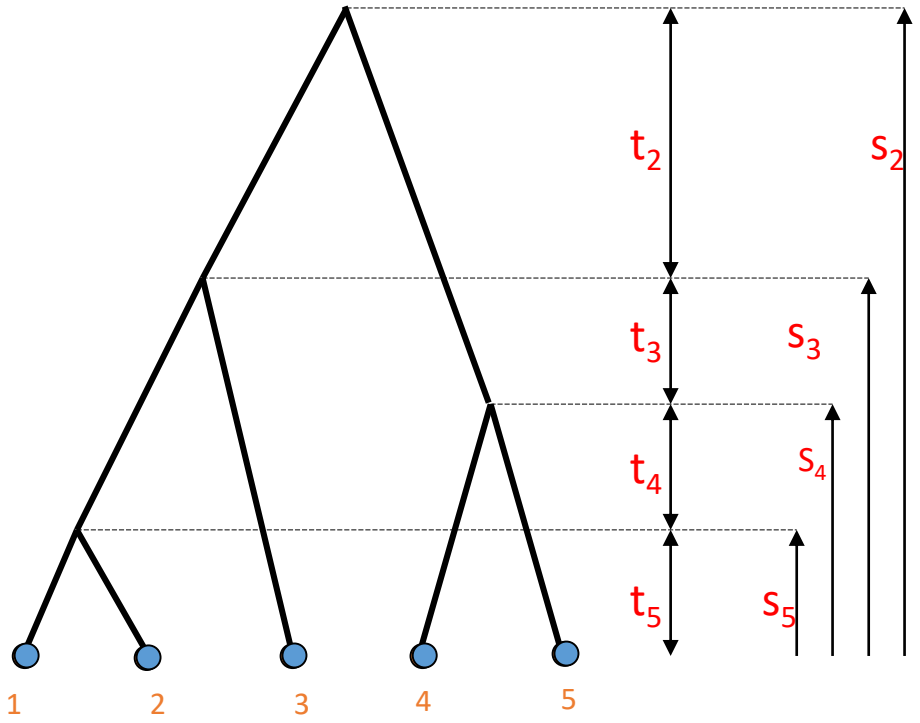
SFS is bar chart of $\boldsymbol{\eta_1}$ = # mutations represented in $\boldsymbol{i}$ out of $\boldsymbol{n}$ cells

$$\eta = \{\eta_1, \eta_2, \ldots, \eta_{20}\} = \{7, 0, 3, 0, 0, 2, 0, 0, 0, 1, \ldots, 0\}$$

$$\Sigma_{i=1}^{n-1} \eta_i = s = \text{\# segregating sites}$$



Deriving **E[SFS]** requires modeling of both topology and metrics of coalescence trees



Expected SFS based on GT-coalescent

q_b = probability that a mutation is present in

$b = 1, 2, \ldots, n$ sequences out of the sample of n sequences

Griffiths and Tavare, 1998

Depend on metrics of the tree: expectations

$$q_b = \frac{\sum_{k=2}^{n} p_k^n(b) k E(S_k)}{\sum_{k=2}^{n} k E(S_k)}$$

Probability that a mutation at the level where there are k ancestors will grow to b copies at the bottom of the tree

$$p_k^n(b) = \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}}$$

Durrett approximation under large N

Expected SFS under **scaled mutation and growth** rates

modo Durrett (heuristic, but simple and close to exact)

$$\vartheta = \theta N, \qquad \beta = r\frac{N}{2}$$

$$\mathbb{E}S_n(1) \sim \frac{\vartheta n \log \beta}{2\beta}.$$

$$\mathbb{E}S_n(k) \approx \frac{\vartheta}{2\beta}\frac{n}{k(k-1)},\ k = 2, \ldots, n-1,$$

How to reconstruct past dynamics ?

- Chief problem: how to measure **time** ?

- But are mutations **not** accumulating in/with **time** ?

- The difficulty is clear from the **asymptotic** Griffiths-Tavaré/Durrett formula

$$q_m \begin{cases} \sim \frac{n\theta}{r}\ln(Nr), & m = 1 \text{ (singletons)} \\ \to \frac{n\theta}{r}\frac{1}{m(m-1)}, & m = 2, \dots, n-1 \end{cases} \quad \text{as } N \to \infty$$

where $\boldsymbol{\frac{\theta}{r}\ln(Nr) = \theta t}$, since $N = N(t) = r\exp(rt)$

- So, time is associated with **singleton** count, but in genome data singletons are usually indistinguishable from **sequencing errors** and discarded

SFS for lbdp (Lambert et al.)

$$\mathbb{E}S_n(k) = \theta \int_0^\infty \left(1 - W(t)^{-1}\right)^{k-1} \left((n-k-1)W(t)^{-2} + 2W(t)^{-1}\right) dt$$

$$k = 1, \ldots, n-1$$

$$W(x) = 1/\,\mathbb{P}(H > x) \qquad x \in [0, \infty).$$

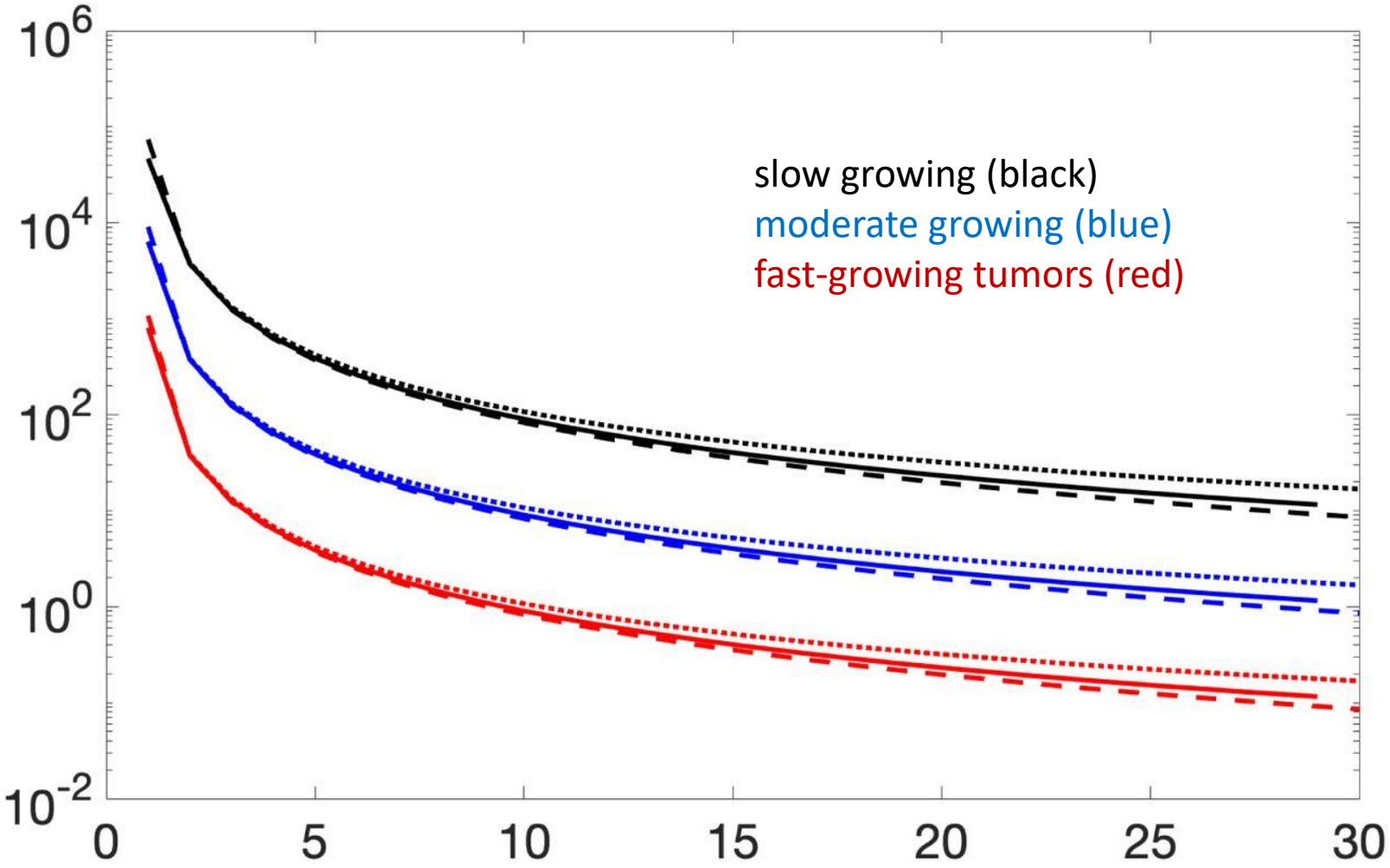
$$W(t) = \alpha + (1-\alpha)e^{rt},\ t \geq 0, \qquad r = b - d \text{ and } \alpha = 1 - pb/r$$

$$\mathbb{E}S_n(k) = \frac{\theta}{r}\left(\frac{n-k-1}{k(k+1)}F([1,2]; k+2, \alpha) + \frac{2}{k}F([1,1]; k+1, \alpha)\right)$$

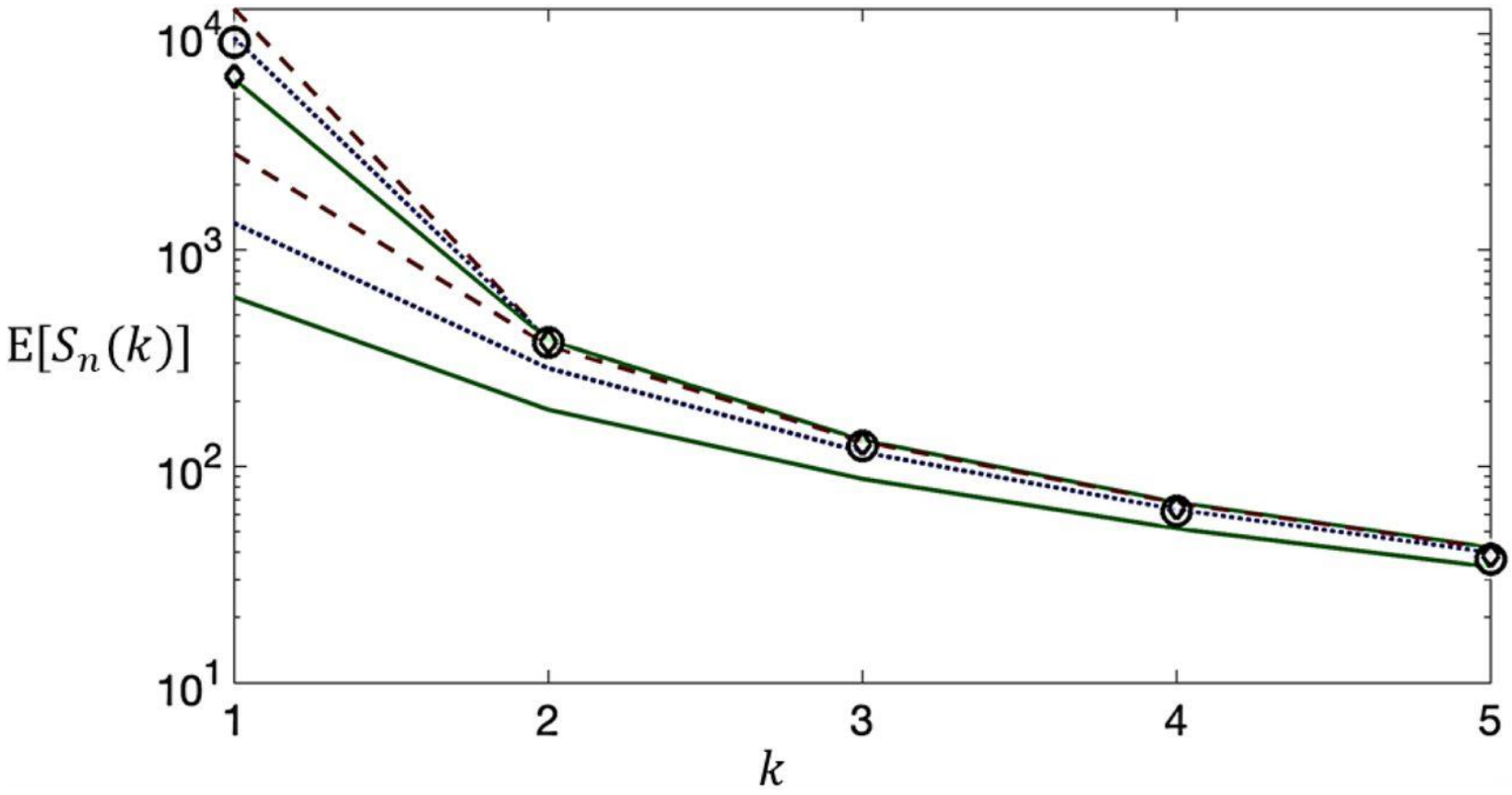
This latter formula can be expressed in the terms of algebraic combination of powers and logarithms (Kimmel, in preparation)

Comparison of expected SFS

Griffiths-Tavaré (continuous lines)
Durrett's approximation (dashed lines)
Lambert (dotted lines)

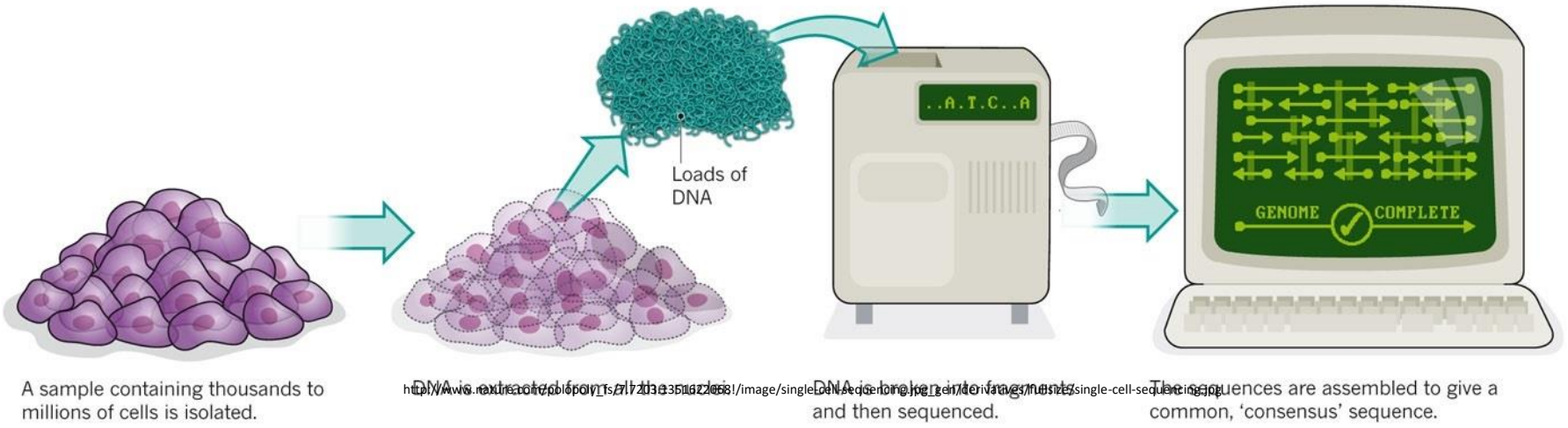


lbdp SFS is sensitive (faintly) to b/d pattern

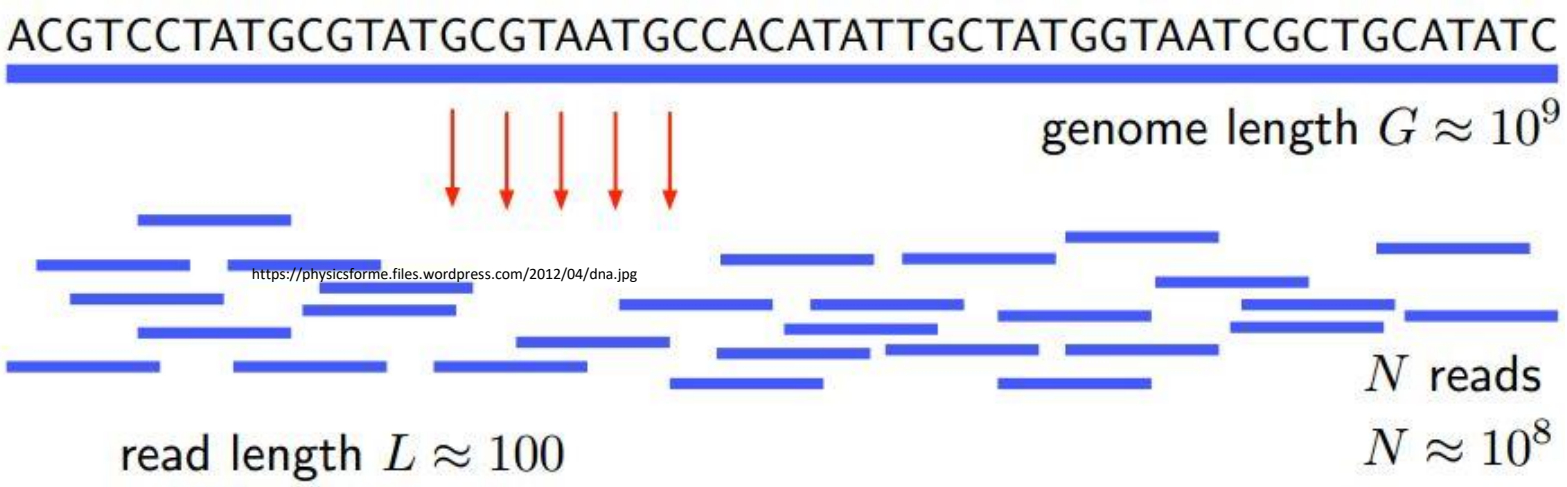


Expected SFS based on the lbdp, with $N = 10^7$, $n = 30$ and $r = 0.04029$, but with $1 - \alpha = 10^{-8}$, 10^{-6}, 0.0001, 0.01, 0.1, 0.5 (dashed, dotted, continuous, and again dashed, dotted and continuous lines), compared to GT SFS (diamonds) and Durrett approximation (circles).

Sequencing cancer DNA



http://www.nature.com/polopoly_fs/7.72031!/image/single-cell-sequencing.jpg_gen/derivatives/full/2/single-cell-sequencing.jpg



https://physicsforme.files.wordpress.com/2012/04/dna.jpg

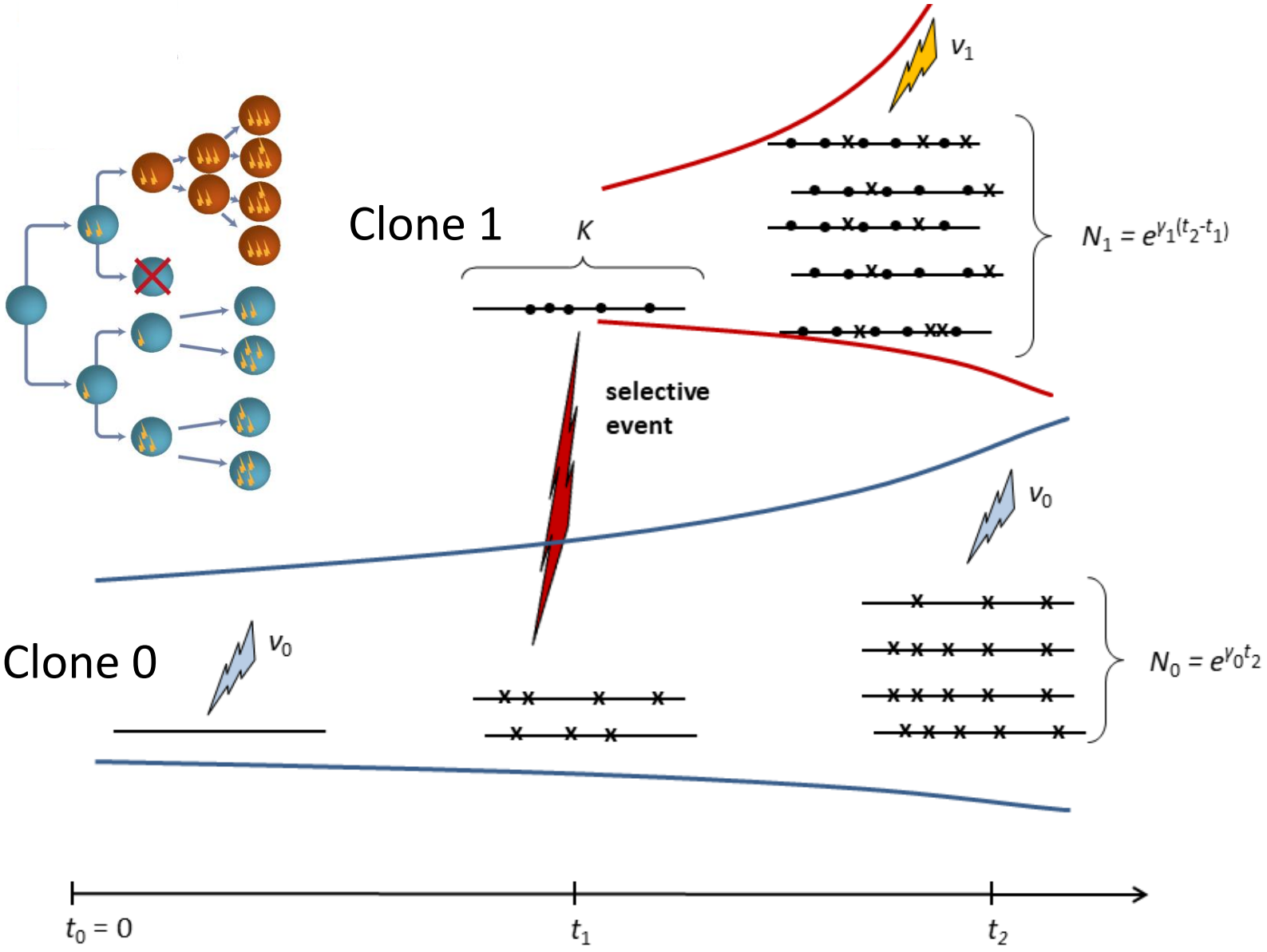
Mutations

…TATATGCTAGCTAGCTACGGCGCGCTG…

…TATATGCTAGCTAGATACGGCGCGCTG…

O. Morozova, M. A. Marra, Genomics, 2008

Model of neutral evolution with a selective sweep in a tumor



Selective sweep caused by appearance of a faster growing clone

- At time $t_0 = 0$, corresponding to the unknown age of the individual, the initial malignant cell population (Clone 0) arises
- Clone 0 grows at rate γ_0, with cells acquiring mutations at the rate ν_0 per time unit per genome
- At time $t_1 > 0$, a secondary clone (Clone 1) arises.
  - This is the "selective event"
  - Clone 1 arises at the background of a haplotype already harboring $\boldsymbol{K}$ mutational hits
  - It grows and mutates at generally different rates, γ_1 and ν_1, respectively

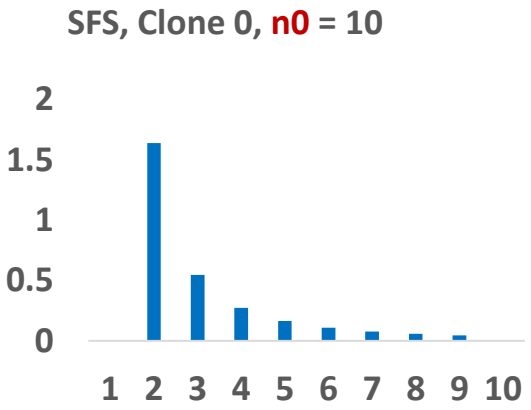
Sampling

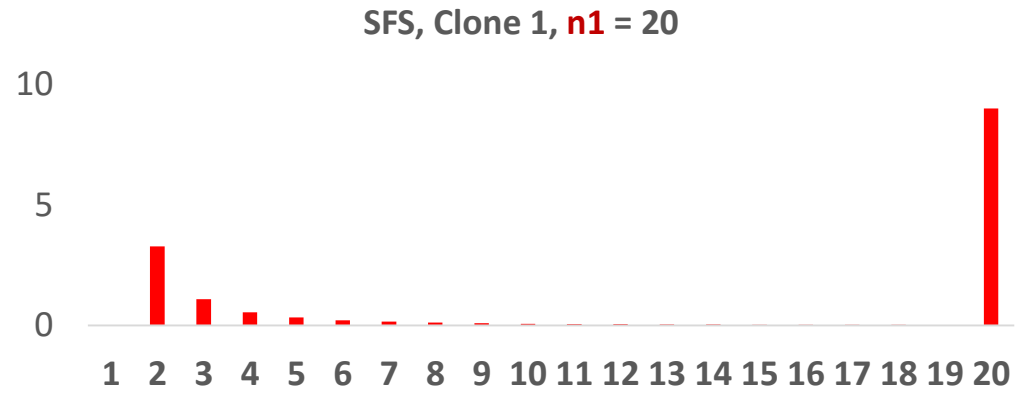
- At $t_2 > t_1 > 0$, tumor is diagnosed and a sample of DNA is available for sequencing
- The resulting sequencing reads represent a mixture from all extant clones
- For each site of the genome, a sample of size n is drawn from $N_0 + N_1$ cells where

$$N_0 = \exp(\gamma_0 t_2), N_1 = \exp(\gamma_1 (t_2 - t_1))$$

so that probability a read is drawn from Clone i is equal to

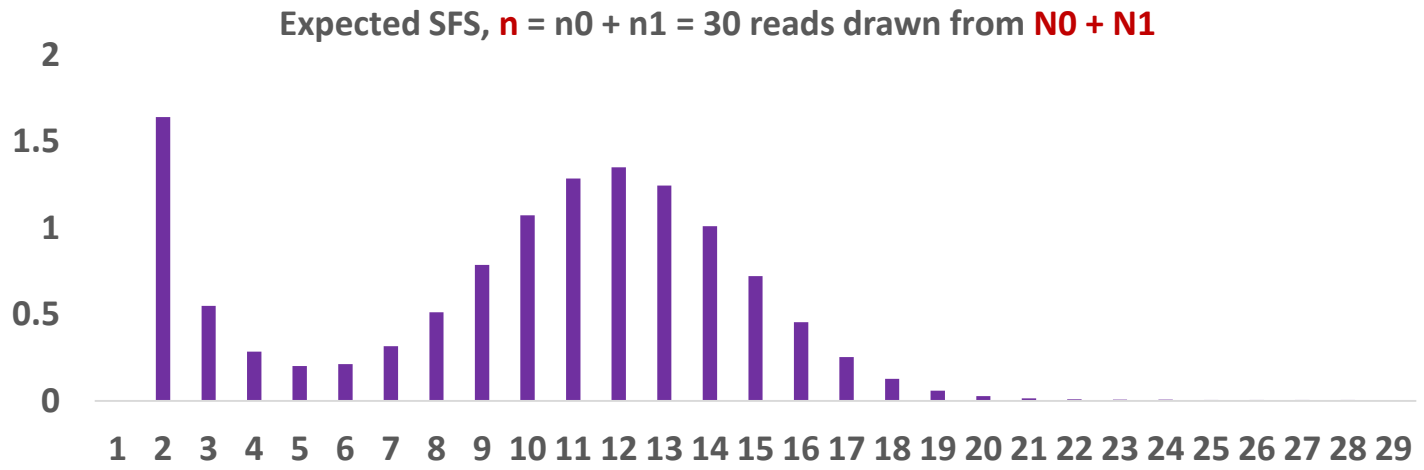
$$p_i = \frac{N_i}{N_0 + N_1}, \qquad i = 0, 1$$



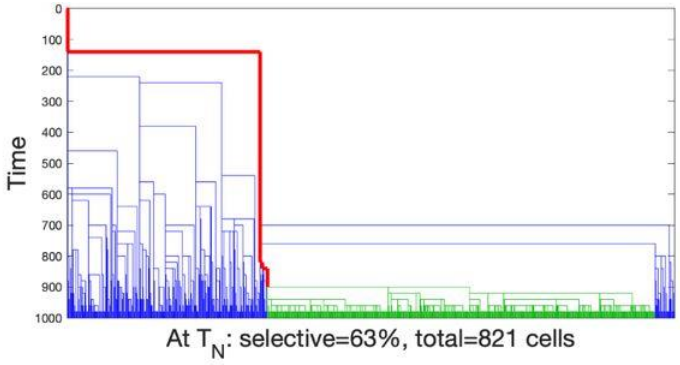


$$q_m^0 = \frac{p_0 n_0 \upsilon_0}{\gamma_0} \frac{1}{m(m-1)}, m = 2, \dots, n_0 - 1$$

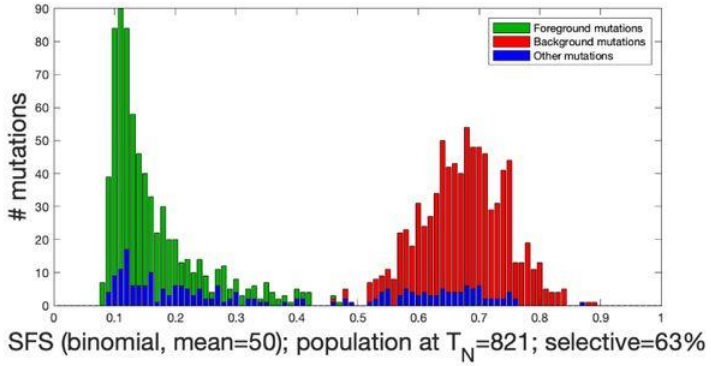
$$q_m^1 = \frac{p_1 n_1 \upsilon_1}{\gamma_1} \frac{1(m < n_1)}{m(m-1)} + K\delta_{m,n_1}, m = 2, \dots, n_1$$



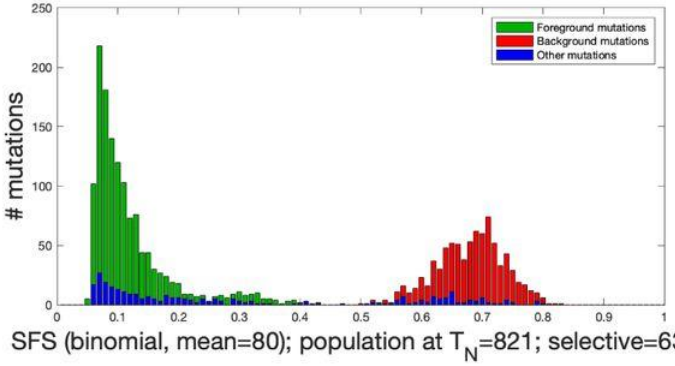
$$Q_m = n\left(\frac{p_0 \upsilon_0}{\gamma_0} + \frac{p_1 \upsilon_1}{\gamma_1}\right)\frac{1}{m(m-1)} + K\binom{n}{m} {p_0}^{n-m} {p_1}^m, \qquad m = 2, \dots, n-1$$



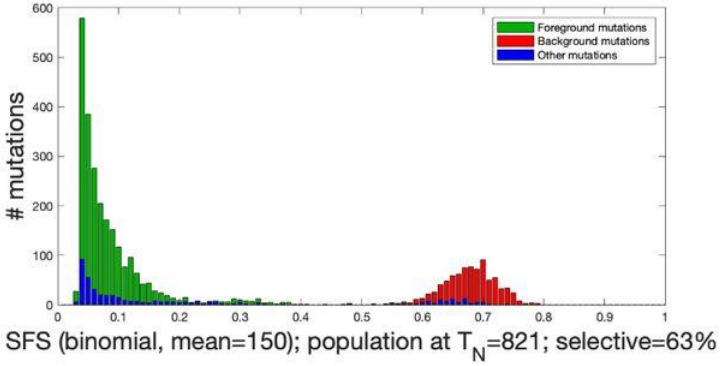
(a) A



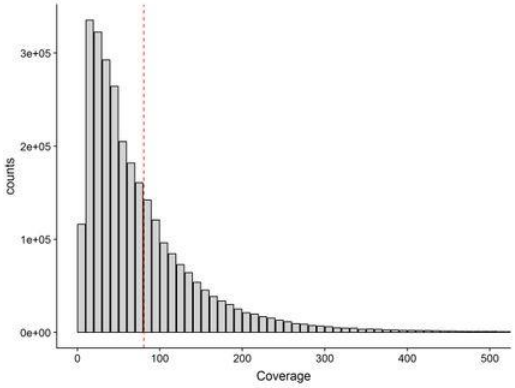
(b) B



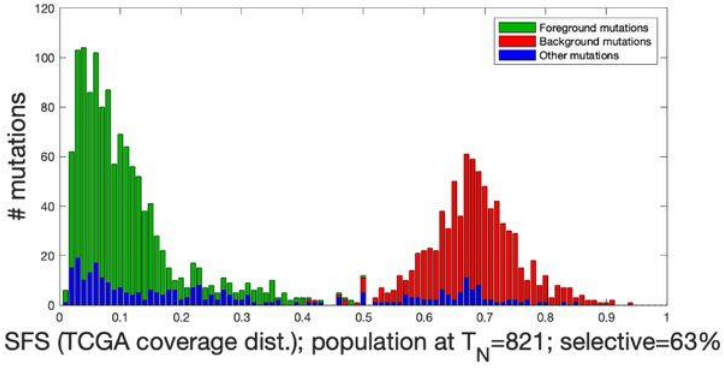
(c) C



(d) D



(e) E



(f) F

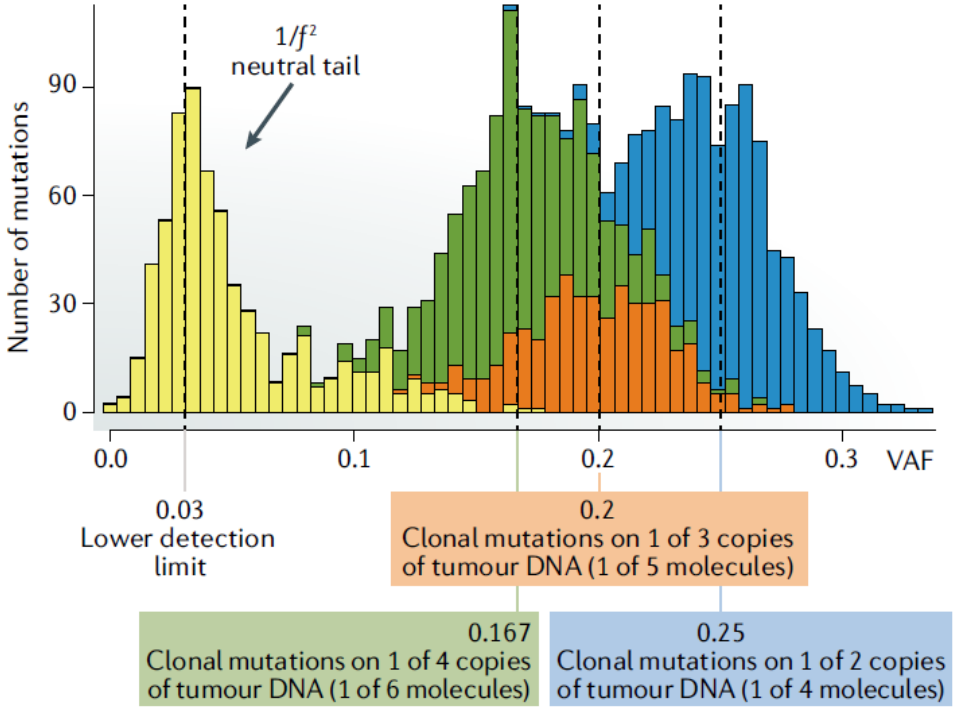
Reality more intricate than models?

Resolving genetic heterogeneity in cancer

Samra Turajlic[1,2,7], *Andrea Sottoriva*[3,7], *Trevor Graham*[4*] and *Charles Swanton*[1,5,6*]

NATURE REVIEWS | GENETICS

VOLUME 20 | JULY 2019 | 405



Generalization of the single sweep model to multiple sweeps

SFS consists of a neutral part (GT or Lambert type)

and number of binomial humps corresponding to emergent clones

$$Q_{nk} = \sum_{\{\sum_h n_h = n,\, n_h \geq 0\}} \binom{n}{n_0, n_1, \ldots, n_H} \prod_{s=0}^{H} p_s^{n_s} \times$$
$$\times \left[\frac{1}{k(k-1)} \sum_{\sigma=0}^{H} \left(n_\sigma \frac{\vartheta_\sigma}{2\beta_\sigma} \right) + \sum_{\sigma=1}^{H} K_\sigma \delta_{n_\sigma m} \right]$$

$$= \frac{A}{k(k-1)} + \binom{n}{k} \sum_{\sigma=1}^{H} K_\sigma p_\sigma^k (1 - p_\sigma)^{n-k}$$

SFS transformation under sampling

- DNA fragments (**reads**) originate from **sampling** from a population of **cells**.
- For each particular mutation site, **each read** covering this site originates from a **different cell**. Number of reads is usually of order 10^2, while there are around at least 3-5 orders of magnitude more tumor cells in a cubic millimeter of tumor tissue (Del Monte 2009).
- For a given mutation site, g**iven coverage *R*, the count *Z* of variant reads has a binomial distribution** Binomial (R, ϕ), where ϕ is the relative frequency of this mutation among the tumor cells.

$$Z_{ki}|R_{ki} \sim \text{Binomial}\left(R_{ki}, \frac{k}{n}\right).$$

- For a given mutation site, the numbers of **reads covering it is considered a random variable** (generically named *R*) drawn from a distribution which is estimated from coverage data.

SFS transformation under sampling

We obtain the "expected sampled SFS", # mutations with VAF between x_1 and x_2

$$\begin{aligned}\mathbb{E}[\Omega(x_1,x_2)] &= \mathbb{E}\left(\sum_{k=1}^{n}\sum_{i=1}^{S_n^u(k)}\mathbb{1}\left(\frac{Z_{ki}}{R_{ki}}\in(x_1,x_2]\right)\right)\\ &= \mathbb{E}\left(\sum_{k=1}^{n}S_n^u(k)\cdot\mathbb{P}\left(\frac{Z_k}{R_k}\in(x_1,x_2]\right)\right)\\ &= \sum_{k=1}^{n}\mathbb{E}S_n^u(k)\cdot\sum_r\varphi_r\cdot\text{Binomial}\left(z\in(x_1r,x_2r];\,r,\frac{k}{n}\right)\end{aligned}$$

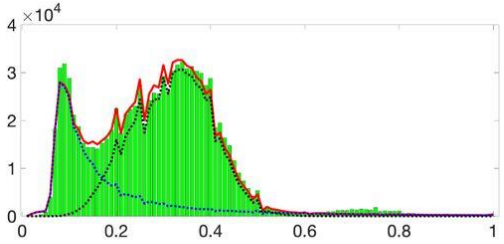
The data can be additionally "corrected" by rejecting (to eliminate sequencing errors)

Sites with coverage $\boldsymbol{R < M}$

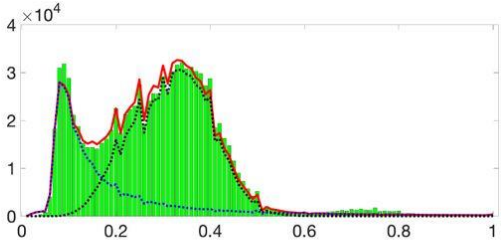
Mutant sites with variant count $\boldsymbol{Z < L}$

Transformation can account for this, and also help estimate the total (unobservable) mutation count from the tally

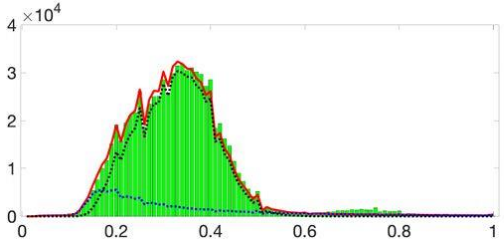
$$\Sigma = A + \sum_{\sigma=0}^{H} K_\sigma,$$



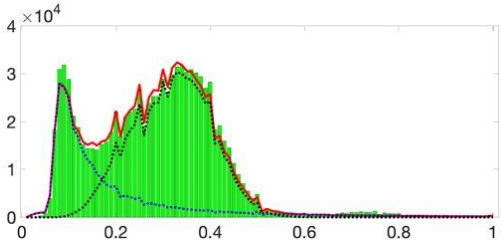
(a) A



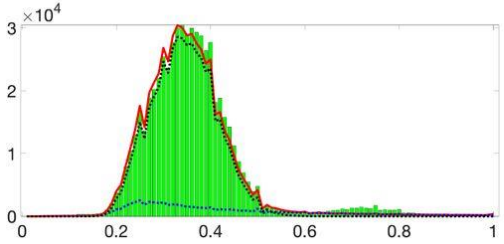
(b) E



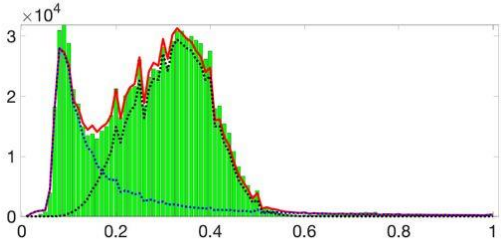
(c) B



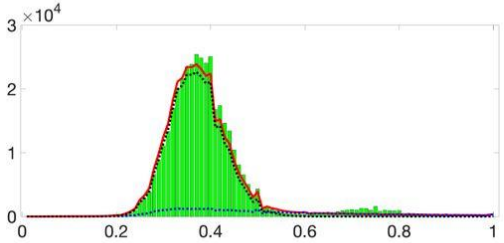
(d) F



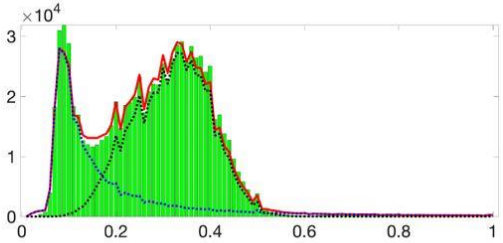
(e) C



(f) G



(g) D



(h) H

Fitting the SFS of case TCGA-A6-6141 (colon cancer).

Threshold combinations of variant and total read counts:

[A]: $L = 5$, $M = 0$,

[B]: $L = 10$, $M = 0$,

[C]: $L = 15$, $M = 0$,

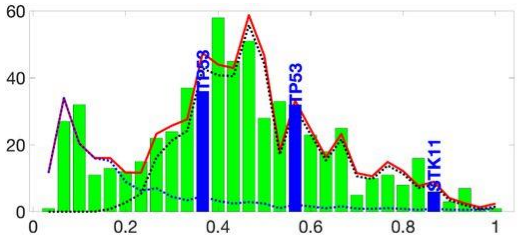
[D]: $L = 20$, $M = 0$,

[E]: $L = 5$, $M = 20$,

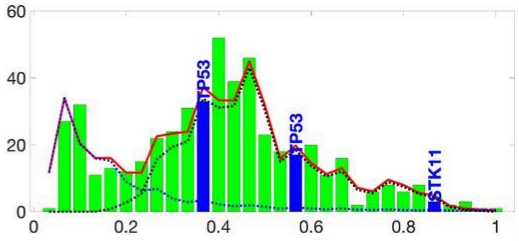
[F]: $L = 5$, $M = 50$,

[G]: $L = 5$, $M = 80$,

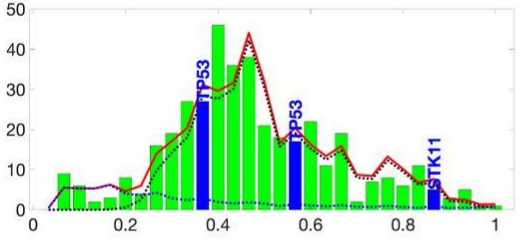
[H]: $L = 5$, $M = 100$.



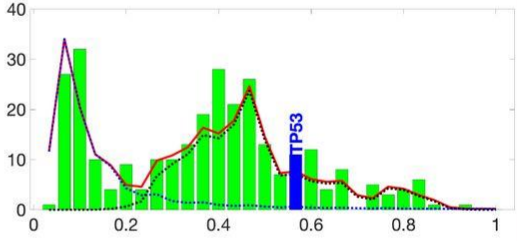
(a) A



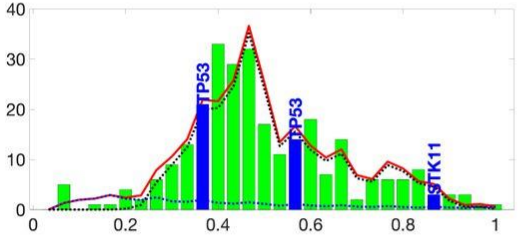
(b) E



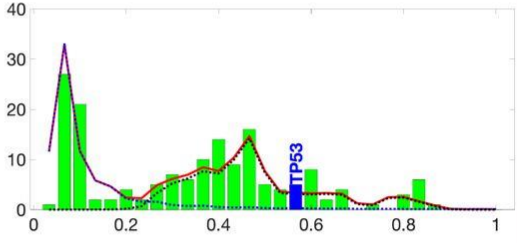
(c) B



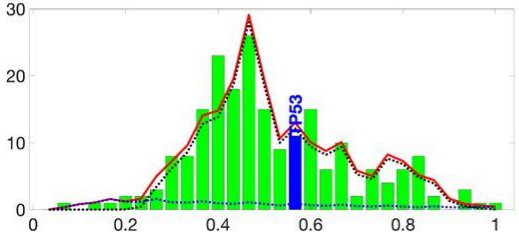
(d) F



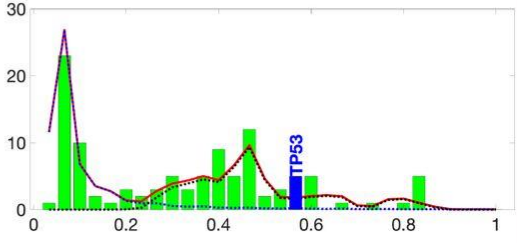
(e) C



(f) G



(g) D



(h) H

Fitting the SFS of case TCGA-86-A4D0 (lung cancer). Threshold combinations of variant and total read counts:

[A]: $L = 5$, $M = 0$.

[B]: $L = 10$, $M = 0$.

[C]: $L = 15$, $M = 0$.

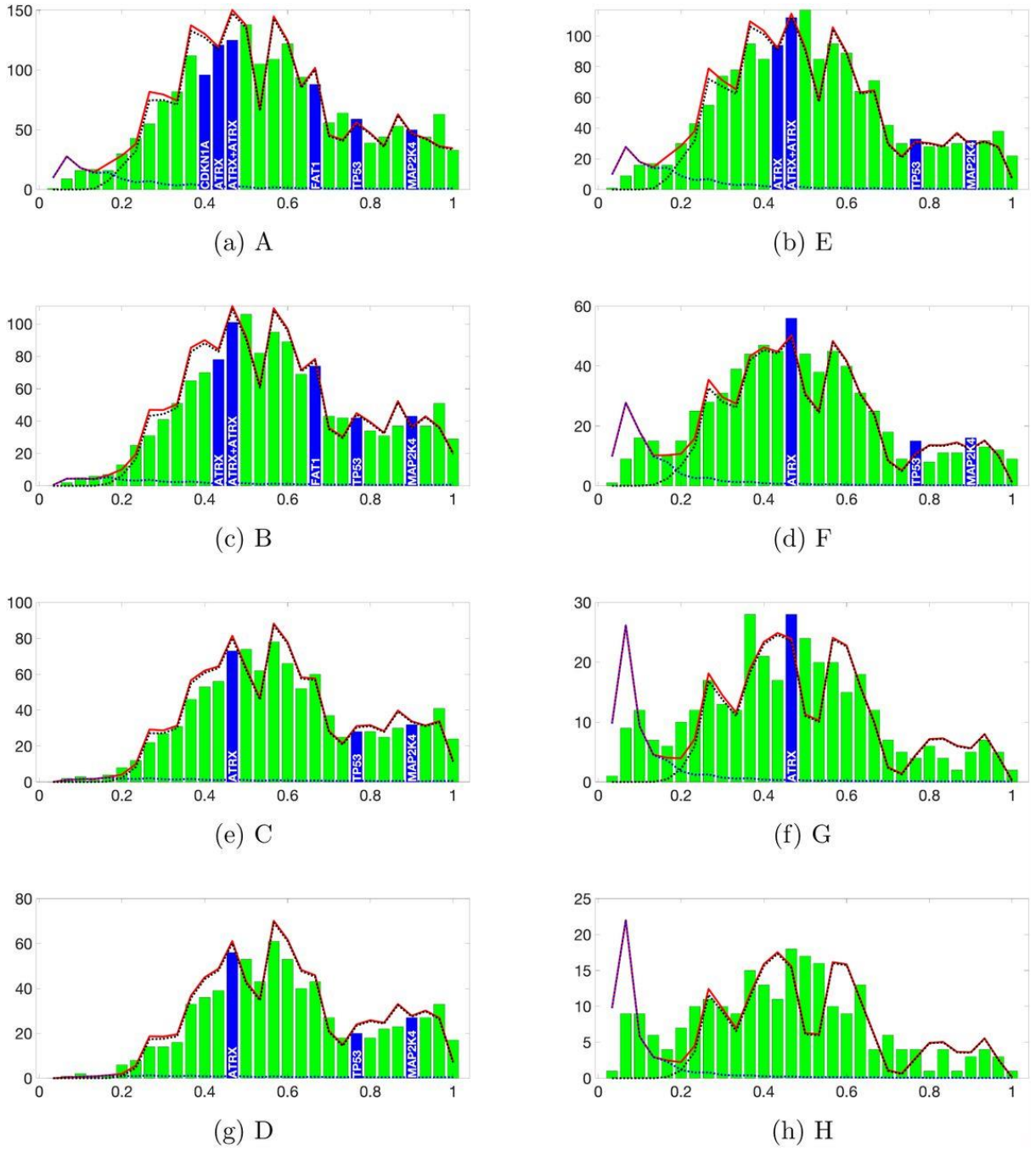
[D]: $L = 20$, $M = 0$.

[E]: $L = 5$, $M = 20$.

[F]: $L = 5$, $M = 50$.

[G]: $L = 5$, $M = 80$.

[H]: $L = 5$, $M = 100$.



(a) A (b) E (c) B (d) F (e) C (f) G (g) D (h) H

Fitting the SFS of case TCGA-62-A46O (lung cancer

Threshold combinations of variant and total read counts:

[A]: $L = 5$, $M = 0$,

[B]: $L = 10$, $M = 0$,

[C]: $L = 15$, $M = 0$,

[D]: $L = 20$, $M = 0$,

[E]: $L = 5$, $M = 20$,

[F]: $L = 5$, $M = 50$,

[G]: $L = 5$, $M = 80$,

[H]: $L = 5$, $M = 100$

Some conclusions

- Fitting allows to determine several synthetic coefficients
  - $A = n(\frac{p_0\theta_0}{r_0} + \frac{p_1\theta_1}{r_1} + \cdots)$ combined reduced mutation rate of clones
  - $p_0, p_1, \ldots$ frequency of clones 0, 1 ,… in the bulk sample
  - $K_1, K_2, \ldots$ underlying haplotype size of clones 1, 2 ,…
- Selective events in history of tumor may have complicated nature (ploidy changes)
- Confidence intervals are mainly simulation-based; some theory exists but it is quite complicated
- Using Wright-Fisher or Moran leads to similar conclusions as using LBDP

Some formulae

Number of mutations in the funder of Clone 1 ≈ $K = \theta_0 t_1$

Mass of neutral muts $\frac{A}{n} = A' = \frac{p_0\theta_0}{r_0} + \frac{p_1\theta_1}{r_1} = \frac{\theta}{r}(p_0 + p_1\frac{\varphi}{\alpha})$

where $r_0 = r,\ r_1 = \alpha r,\ \theta_0 = \theta,\ \theta_1 = \varphi\theta,\ t_1 = t,\ t_2 = \beta t,$ ....

$p_0 = \frac{e^{r_0 t_2}}{e^{r_0 t_2} + e^{r_1(t_2 - t_1)}} \Rightarrow \beta(\alpha - 2)rt = \ln\frac{p_1}{1 - p_1}$

$\frac{K}{A'} = \frac{rt}{p_0 + p_1\frac{\varphi}{\alpha}}$ hence $\frac{K}{A'} > rt \Rightarrow \frac{\varphi}{\alpha} < 1$

Growth accelerates faster than mutation rate

(Assume cumulative growth rate $rt = \ln 10^{11} = 25.33$)

Estimates of SFS parameters based on colon cancer genomes from TCGA

	A'	p1	K1	p2	K2	A+K1+K2	(K1+K2)/A'	ln(p/(1-p))
Colon-mutator								
TCGA-A6-6141	1300	0.12	42000	0.28	180000	1522000	170.77	-0.41
TCGA-AA-3555	2000	0.35	230000			2230000	115.00	-0.62
TCGA-AA-3977	20000	0.35	450000	0.23	200000	20650000	32.50	0.32
TCGA-AA-A00N	20000	0.23	450000			20450000	22.50	-1.21
TCGA-AG-A002	5000	0.41	700000			5700000	140.00	-0.36
TCGA-AZ-4315	3300	0.13	30000	0.31	1000000	4330000	312.12	-0.24
TCGA-BS-A0TC	1000	0.39	17000			1017000	17.00	-0.45
TCGA-CA-6717	40000	0.21	1000000	0.09	1500000	42500000	62.50	-0.85
TCGA-CA-6718	10000	0.28	600000	0.50	40000	10640000	64.00	1.27
TCGA-EI-6917	2500	0.27	800000			3300000	320.00	-0.99
TCGA-F5-6814	80000	0.38	800000	0.25	400000	81200000	15.00	0.53
Colon non-mutator								
TCGA-CA-6718	100	0.27	6500	0.44	600	107100	71.00	0.90
TCGA-AZ-4315	50	0.13	900	0.33	22000	72900	458.00	-0.18

Mutation rate: We note that $\boldsymbol{K = \theta_0 t_1}$, number of mutation in clone 1's backbone.
Suppose the $\boldsymbol{t_1 = 10\ yr}$.
Non-mutator cases show **1-2** mutations per genome per day, almost normal range.
Mutator cases show **10 – 1000** time more.

Estimates of SFS parameters based on colon cancer genomes from TCGA

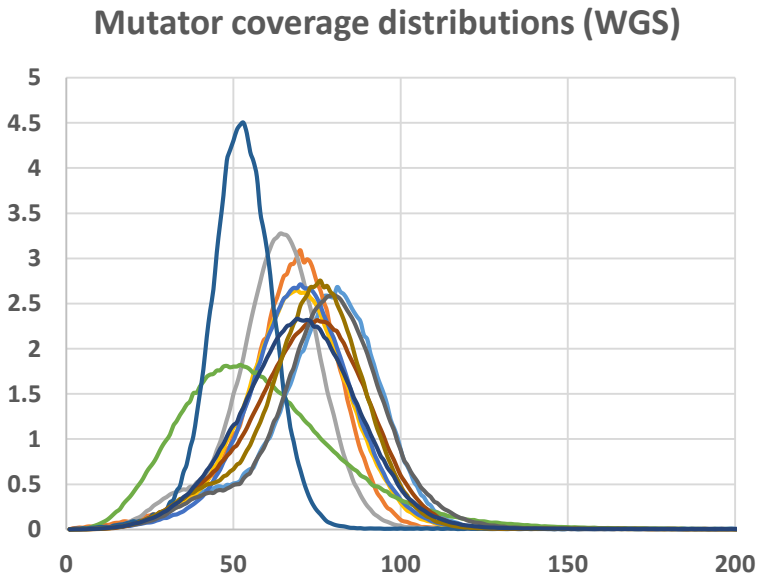
	A'	p1	K1	p2	K2	A+K1+K2	(K1+K2)/A'	ln(p/(1-p))
Colon-mutator								
TCGA-A6-6141	1300	0.12	42000	0.28	180000	1522000	170.77	-0.41
TCGA-AA-3555	2000	0.35	230000			2230000	115.00	-0.62
TCGA-AA-3977	20000	0.35	450000	0.23	200000	20650000	32.50	0.32
TCGA-AA-A00N	20000	0.23	450000			20450000	22.50	-1.21
TCGA-AG-A002	5000	0.41	700000			5700000	140.00	-0.36
TCGA-AZ-4315	3300	0.13	30000	0.31	1000000	4330000	312.12	-0.24
TCGA-BS-A0TC	1000	0.39	17000			1017000	17.00	-0.45
TCGA-CA-6717	40000	0.21	1000000	0.09	1500000	42500000	62.50	-0.85
TCGA-CA-6718	10000	0.28	600000	0.50	40000	10640000	64.00	1.27
TCGA-EI-6917	2500	0.27	800000			3300000	320.00	-0.99
TCGA-F5-6814	80000	0.38	800000	0.25	400000	81200000	15.00	0.53
Colon non-mutator								
TCGA-CA-6718	100	0.27	6500	0.44	600	107100	71.00	0.90
TCGA-AZ-4315	50	0.13	900	0.33	22000	72900	458.00	-0.18

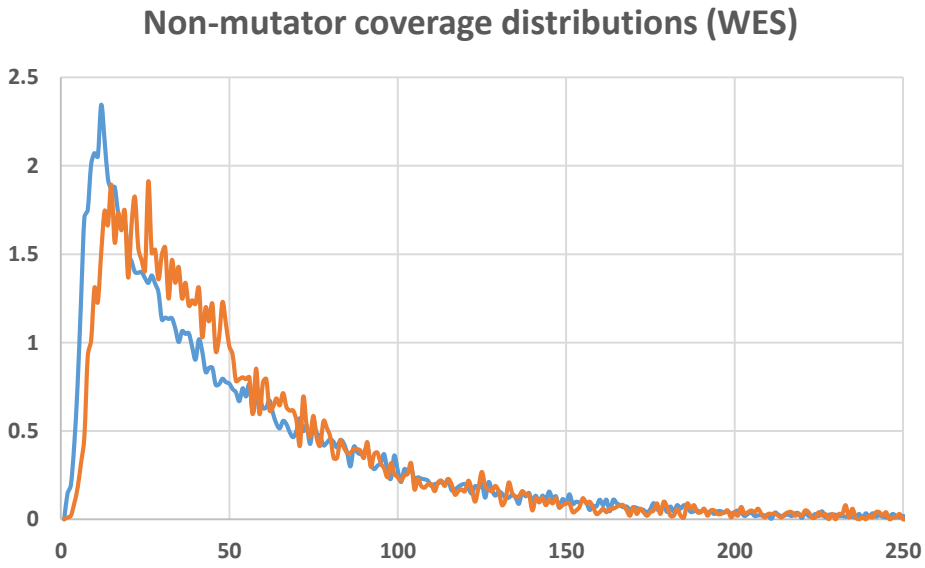
We note that $\frac{K}{A\prime} = \frac{rt}{p_0 + p_1\frac{\varphi}{\alpha}}$ and assume $rt = ln\,10^{11} = 25.33$. We note that observed values of $\frac{K}{A\prime}$ are generally higher than those of rt, indicating growth rate acceleration over mutation rate acceleration in clone 1.

Discussion

- Based on TCGA colon cancer subsample, parameters obtained from the SFS are crudely consistent with know epidemiology of colon cancer and and human mutation rates.

- Ploidy and CNV variation may alter conclusions, but consider coverage distributions.





Discussion

- Williams et al. (Nature Genet.) propose that most humps are simply diploid genome signature of the ancestral cell of the tumor ("truncal mutations").
  - This is unlikely in mutator cases, since it would indicate the hypermutation rate in premalignant cells.
  - In other cases it may be te case if humps are forced to 0.5 VAF.
  - This may be right, if contamination with normal cells is serious.

- On the other hand, hematologic examples indicate that drivers of malignancy appear late in the natural course of disease (Wojdyla et al. PLoS CB; Kimmel and Corey, Front. Immunol.)

Tug of war between passenger and driver mutations

Kimmel, Bobrowski, Dinh and Kurpas

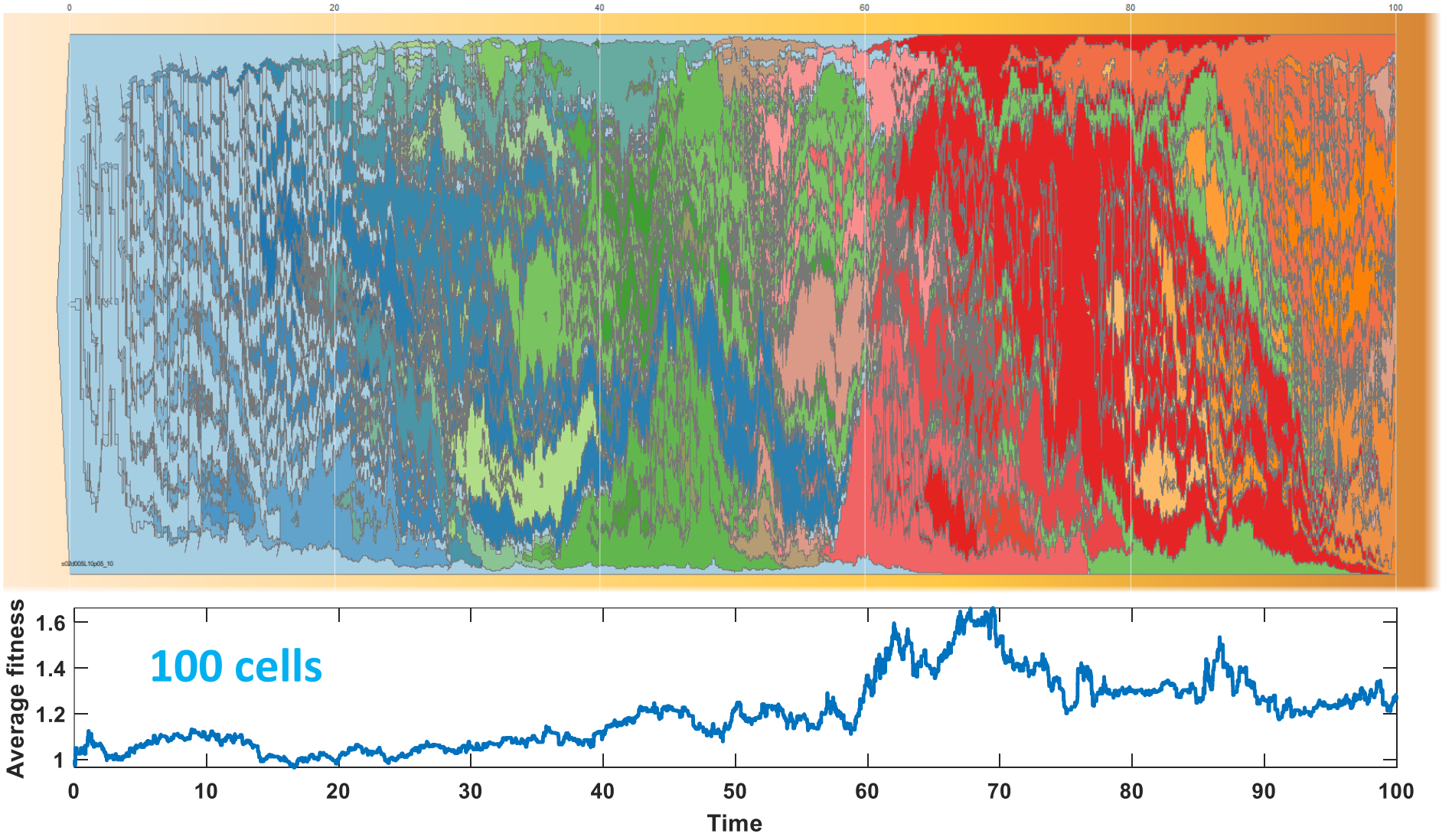
Cells sequentially acquire

- Less frequent Driver mutations (advantageous)
- More frequent Passenger mutations (disadvantageous)

Individual cell fitness increases with the number of drivers but decreases with the number of passengers

Cells compete by the rules of Moran process (or critical bp)

As a result, tumor fitness fluctuates and aggressive clones are transiently established


0
20
40
60
80
100
100 cells
Average fitness
1.6
1.4
1.2
1
0
10
20
30
40
50
60
70
80
90
100
Time