

Workshop on Mathematical Population Genetics
ICMS Edinburgh

March 26-30, 2006

Neutral single-locus genealogies under genetic hitchhiking.
A few approximations and a closer look.

Anton Wakolbinger
Goethe-Universität Frankfurt

Joint work with
Alison Etheridge (Oxford) and Peter Pfaffelhuber (München)

At a selective locus,

consider a **beneficial allele** with **selective advantage** s
per individual and generation.

Assume constant population size N

“Selective sweep”:

initially only one copy of the beneficial allele

whose offspring fixates

in a time which is short compared to N generations.

Consider a **neutral locus** in the neighbourhood of the selective one

and a sample of n individuals taken from the population
at the end of the sweep.

Questions:

To **how many ancestors** does the sample trace back **at the neutral locus**?

What does the (random) ***ancestral sample partition*** at the neutral locus
look like?

Two extremal cases:

1. **Tight linkage** between the two loci:

only one neutral ancestor (the founder of the sweep)

2. **No linkage** between the two loci:

for large N_S , with high probability all neutral ancestors are different

What happens between these extremal cases?

This must depend on N , s and

r ... probability of recombination per individual per generation

$$\alpha := sN$$

$$\rho := rN$$

When N is large and time is measured in units of N generations, then the relative frequency P of the beneficial allele

follows the dynamics

$$dP = \alpha P(1 - P) dt + \text{genetic drift} \quad (*)$$

For large α , (*) should be well approximated by a logistic curve:

$$dP = \alpha P(1 - P) dt$$

$$dP = \alpha P(1 - P) dt$$

A slight conceptual difficulty:

Where to start the logistic curve?

A sort of remedy:

For large α

and a fixed level c ,

the time it takes to grow the logistic curve from c/α to $1 - c/\alpha$ is

$$\asymp \frac{\log \alpha}{\alpha}$$

Assume $\rho \sim \gamma \frac{\alpha}{\log \alpha}$

or equivalently $r \sim \frac{s}{\log N}$.

Then the probability
that a neutral line
is not hit by a recombination during the sweep is

$$e^{-\gamma} = \alpha^{-\gamma/\log \alpha} = (4Ns)^{-r/s} \left(1 + \mathcal{O}\left(\frac{1}{\log \alpha}\right) \right).$$

(Maynard-Smith and Haigh 1974)

Diffusion approximation

On the time scale Nt , with $N \rightarrow \infty$

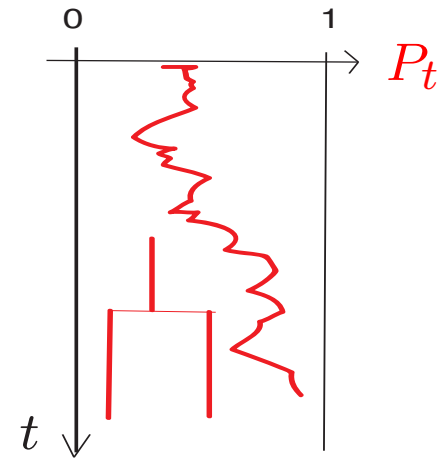
the relative frequency P of the beneficial allele

follows the dynamics

$$dP = \sqrt{P(1 - P)} \sigma dW + \alpha P(1 - P) dt.$$

given the path P
of this relative frequency,

the ancestral lines of a
sampled pair
coalesce at rate $1/P$



And all pairs of ancestral lines do this independently.

Kingman's coalescent in varying (sub-)population size P

Assume **one advantageous mutant** enters the stage at (say) time $t = 0$.

Condition its **offspring** to fixate. Let P_t be the **offspring size** at time t .

The dynamics of P conditioned to fixation is given by

$$dP = \sqrt{\sigma^2 P(1-P)} dW + \coth\left(\frac{\alpha}{\sigma^2} P\right) \alpha P(1-P) dt$$

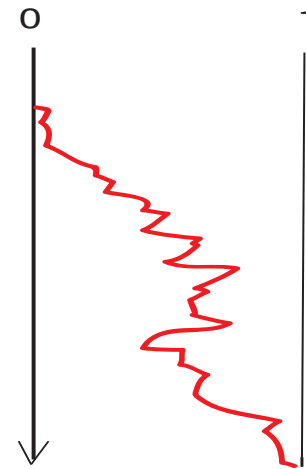


Assume **one advantageous mutant** enters the stage at (say) time $t = 0$.

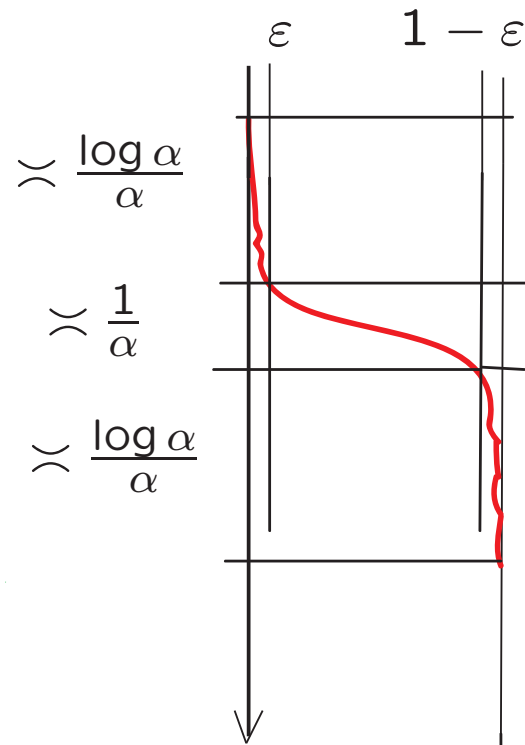
Condition its **offspring** to fixate. Let P_t be the **offspring size** at time t .

The dynamics of P conditioned to fixation is given by

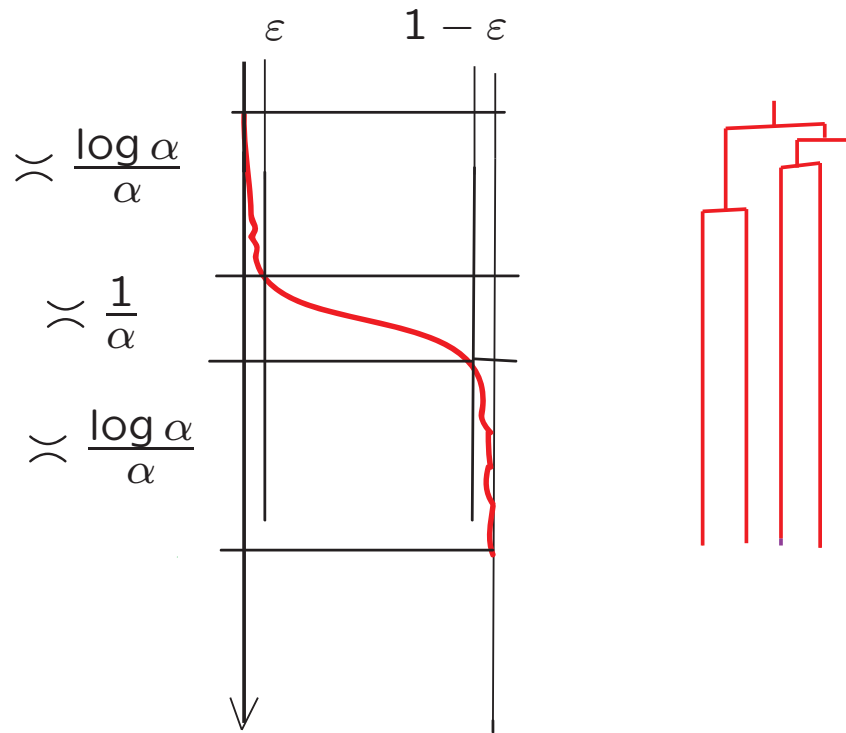
$$dP = \sqrt{\sigma^2 P(1-P)} dW + \coth\left(\frac{\alpha}{\sigma^2} P\right) \alpha P(1-P) dt$$



For large α
the **middle piece** of the curve
becomes almost deterministic,
with logistic drift $\alpha P(1 - P)$,
and lasts longer than the
initial and final piece.



In such a background, the coalescent is squeezed:

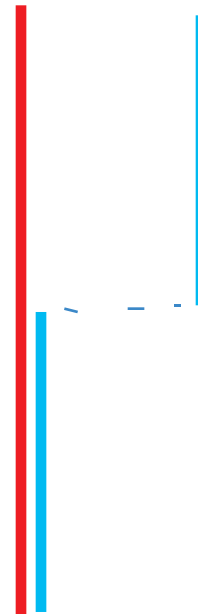


What tears the genealogie at the neutral and at the selective locus apart

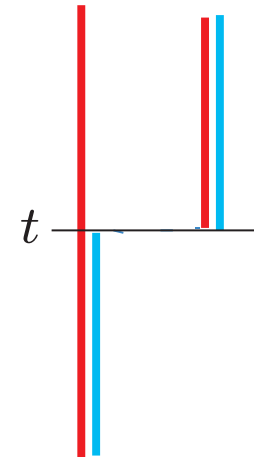
is *recombination*.

Along an individual ancestry,
recombination events
happen *at rate ρ* .

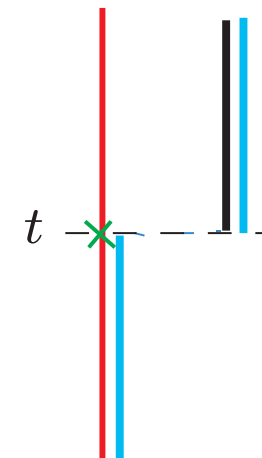
Recall: For population size
 N ,
the recombination probability
is $r = \rho/N$
per individual per generation.



Given P_t , and
 given a recombination event at time t
 of a **neutral lineage** travelling together
 with the **advantageous allele**,
 the probability that the neutral lineage
 only changes to another red driver
 is P_t .



Hence, given P ,
 for a **neutral lineage** the
effective recombination rate
out of the sweep is
 $\rho \cdot (1 - P_t)$



Structured coalescent in random background P

At any time, a **neutral lineage** finds itself either in the **red** or in the **black** compartment.

Given P ,

- a **neutral lineage** changes from **red** to **black** at rate $\rho(1 - P_t)$,
and from **black** to **red** at rate ρP_t
- the **pair coalescence** rate for **neutral lineages** is
 σ^2/P_t in the **red** compartment,
and $\sigma^2/(1 - P_t)$ in the **black** compartment.

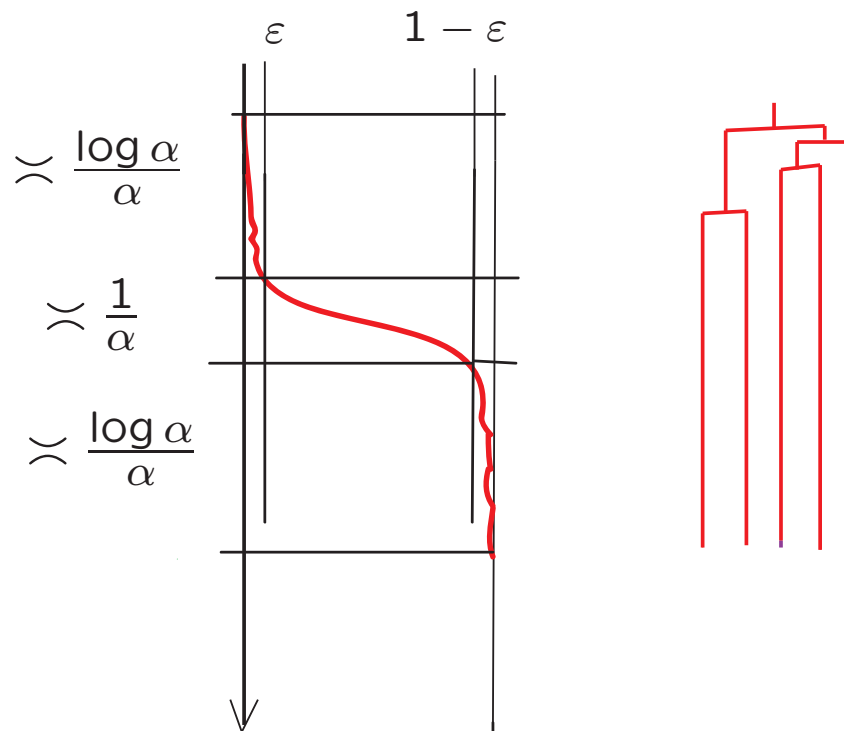
Take an n -sample at the end of the sweep. Trace back the **neutral lineages** to the beginning of the sweep.

Questions:

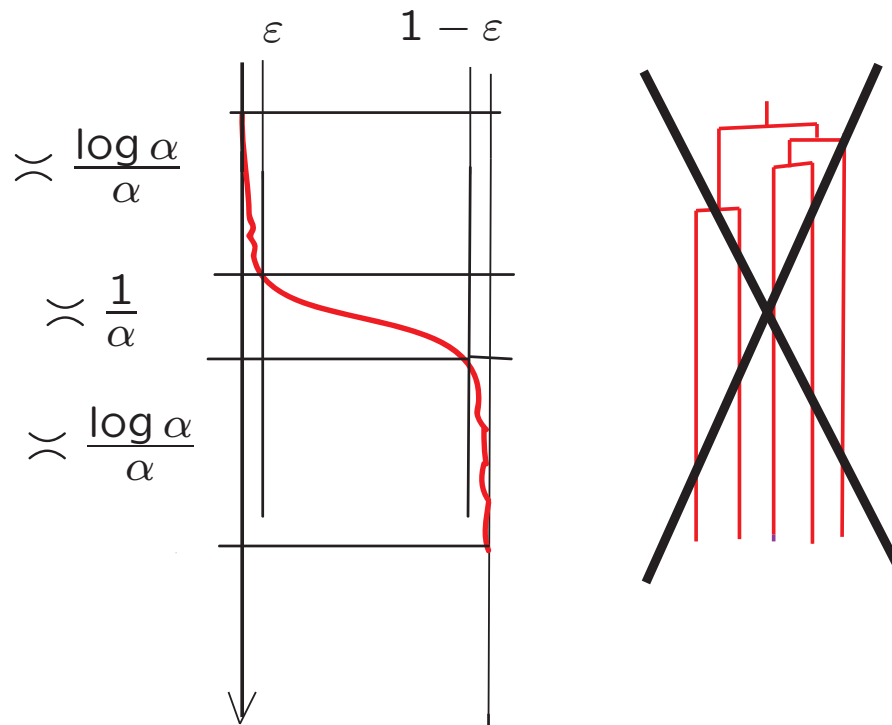
Averaged over P , what is the distribution

- of the number of neutral lineages that escape the sweep?
- of the **ancestral partition** from the beginning of the sweep?

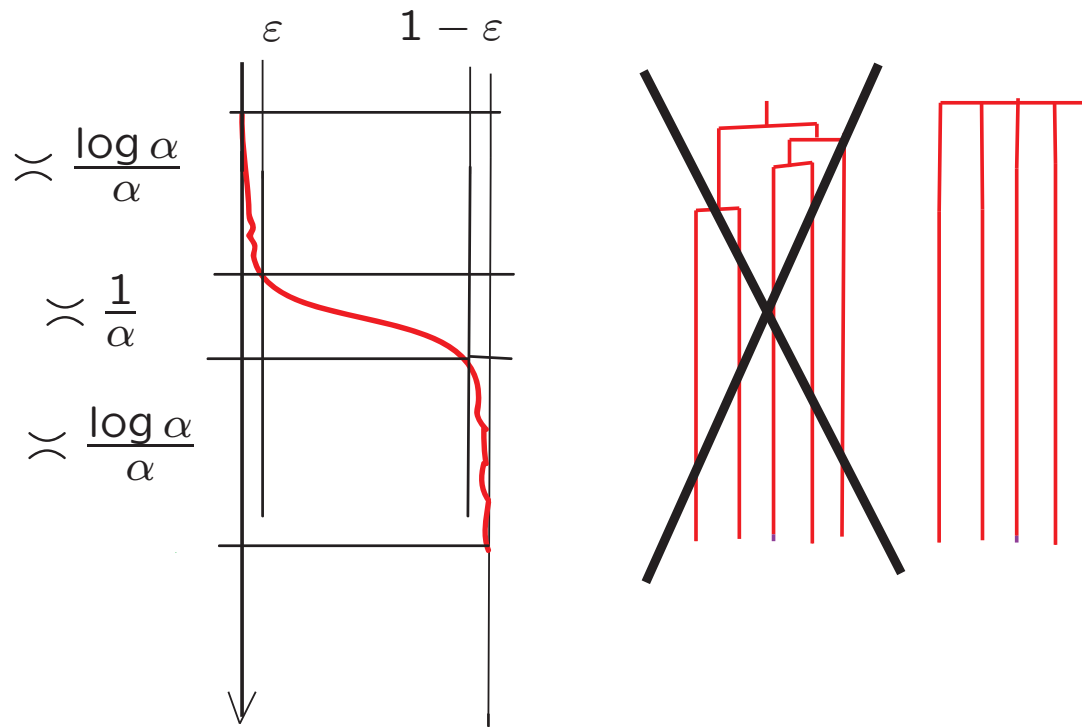
A first approximation: Assume the **coalescence tree** is *star-shaped*, i.e. not like this



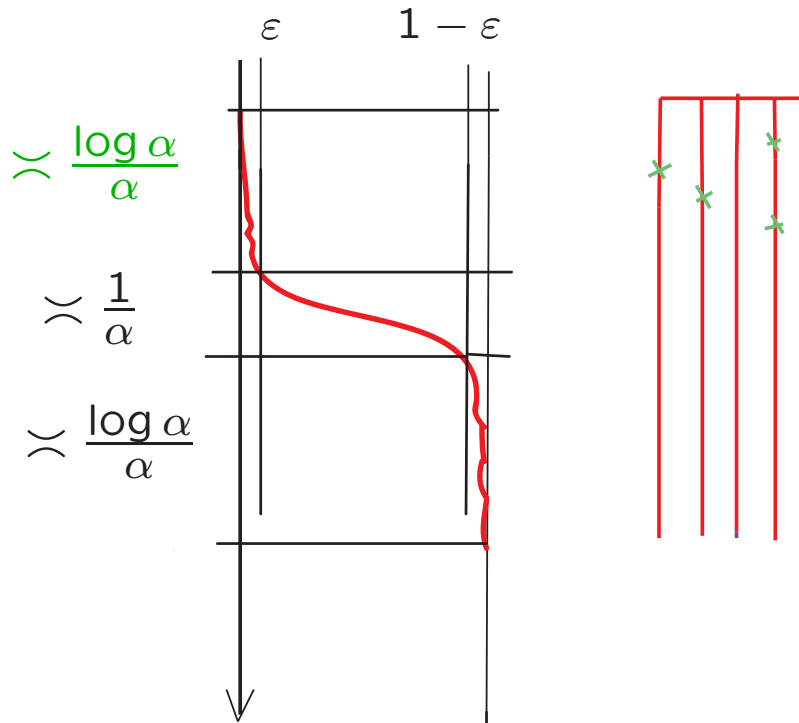
A first approximation: Assume the **coalescence tree** is *star-shaped*, i.e. not like this



...but like this:



The time interval when recombination is effective lasts $\asymp \frac{\log \alpha}{\alpha}$.
 Then lines are hit by recombination events at rate ρ .



The probability that a single line is not hit by a recombination is

$$\asymp e^{-\rho \frac{\log \alpha}{\alpha}}.$$

For $\alpha \rightarrow \infty$ this is different from 0 and 1 if

$$\rho = \gamma \frac{\alpha}{\log \alpha} \text{ for some } \gamma \in (0, \infty).$$

Then the probability that a single line is not hit by a recombination is

$$\asymp e^{-\gamma}$$

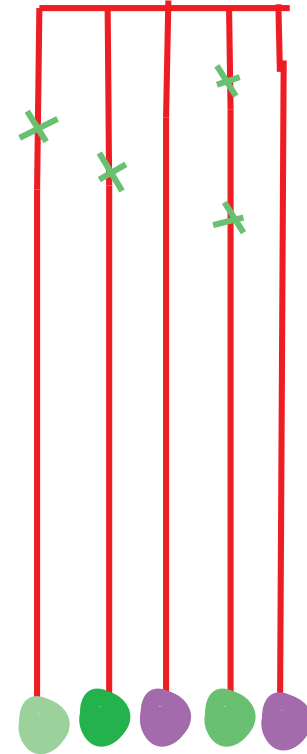
Proposition

In an n -sample taken at the end of the sweep,

the random number H of neutral lineages that trace back to the founder of the sweep is approximately Binomial with n trials and success probability $e^{-\gamma}$.

The other $n - H$ lineages all trace back to different neutral ancestors at the beginning of the sweep.

The error in probability is $\mathcal{O}\left(\frac{1}{\log \alpha}\right)$.

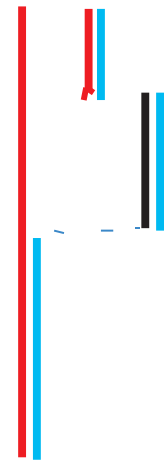


How to improve this $\mathcal{O}\left(\frac{1}{\log \alpha}\right)$ -approximation?

A key lemma:

The probability that a **neutral lineage** recombines out of the sweep and then recombines back into the sweep is

$$\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right).$$



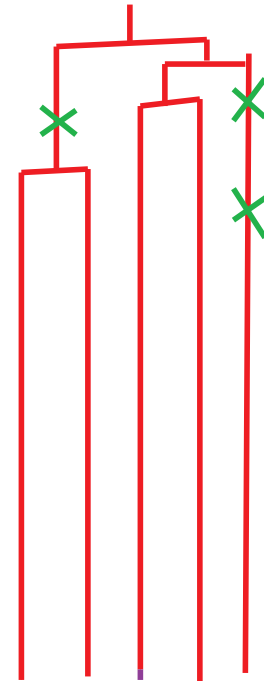
Corollary:

The random ancestral partition of the sample arises, up to an error in probability of $\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$,

through an n -coalescent

with pair coalescence rate $\frac{\sigma^2}{P_t}$,

marked with intensity $(1 - P_t)\rho$.



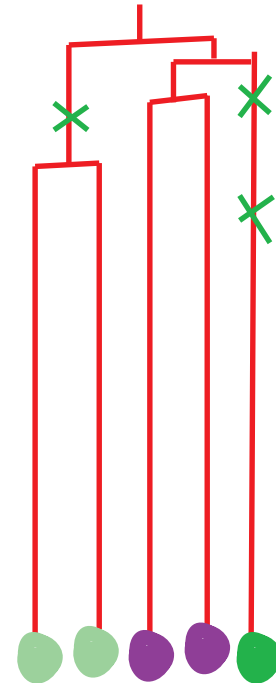
Corollary:

The random ancestral partition
of the sample arises,
up to an error in probability of
 $\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$,

through an n -coalescent

with pair coalescence rate $\frac{\sigma^2}{P_t}$,

marked with intensity $(1 - P_t)\rho$.



Corollary:

The random ancestral partition of the sample arises,

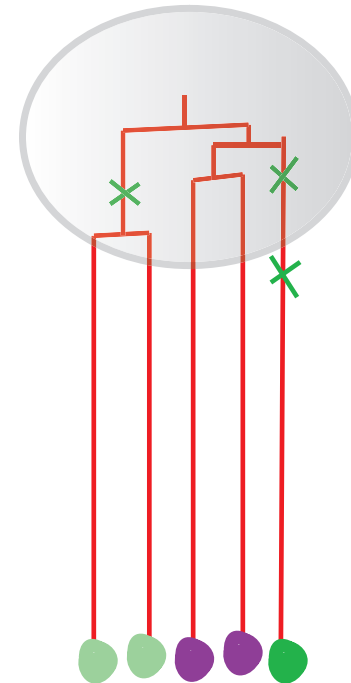
up to an error in probability of

$$\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right),$$

through an n -coalescent

with pair coalescence rate $\frac{\sigma^2}{P_t}$,

marked with intensity $(1 - P_t)\rho$.



To further approximate the ancestral partition of the sample

we zoom into the onset of the sweep.

It turns out that

up to an error in probability of $\mathcal{O}\left(\frac{1}{(\log \alpha)^2}\right)$,

the coalescent tree in the random background P ,
marked at rate $\rho \cdot (1 - P)$

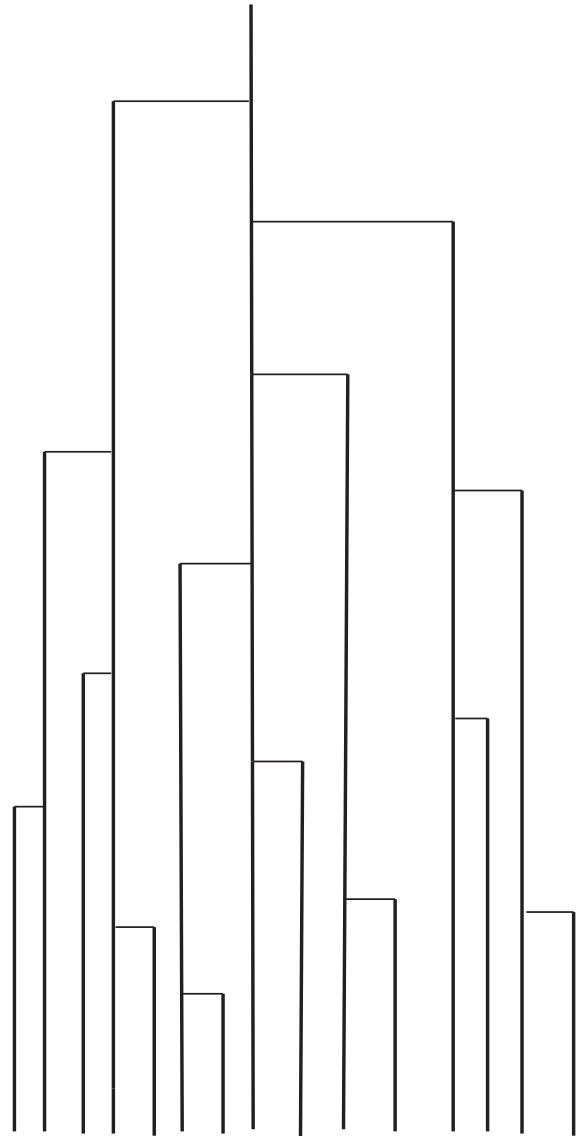
can be replaced by a

binary tree \mathcal{Y} splitting at rate α ,

truncated when its number of branches reaches $\frac{2}{\sigma^2}\alpha$,

and marked at rate ρ .

a Yule tree \mathcal{Y}

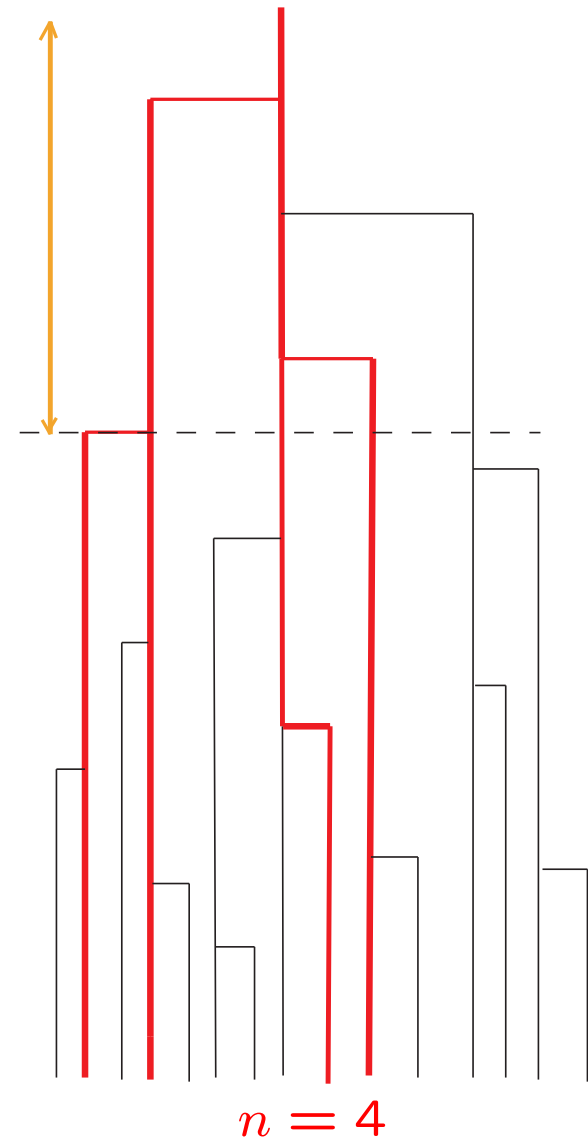




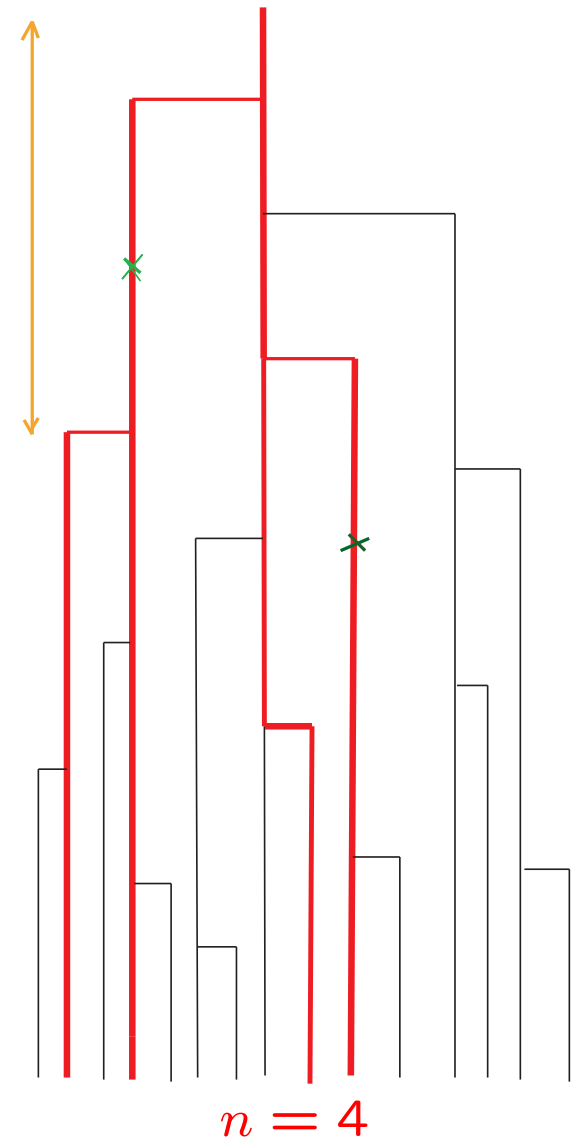
George Udny Yule, 1871-1951

Let \mathcal{Y}_n be that **subtree** of \mathcal{Y} , which belongs to n randomly chosen leaves of \mathcal{Y} .

We define the *early phase* as the time interval it takes \mathcal{Y}_n to grow to its full number n of branches.

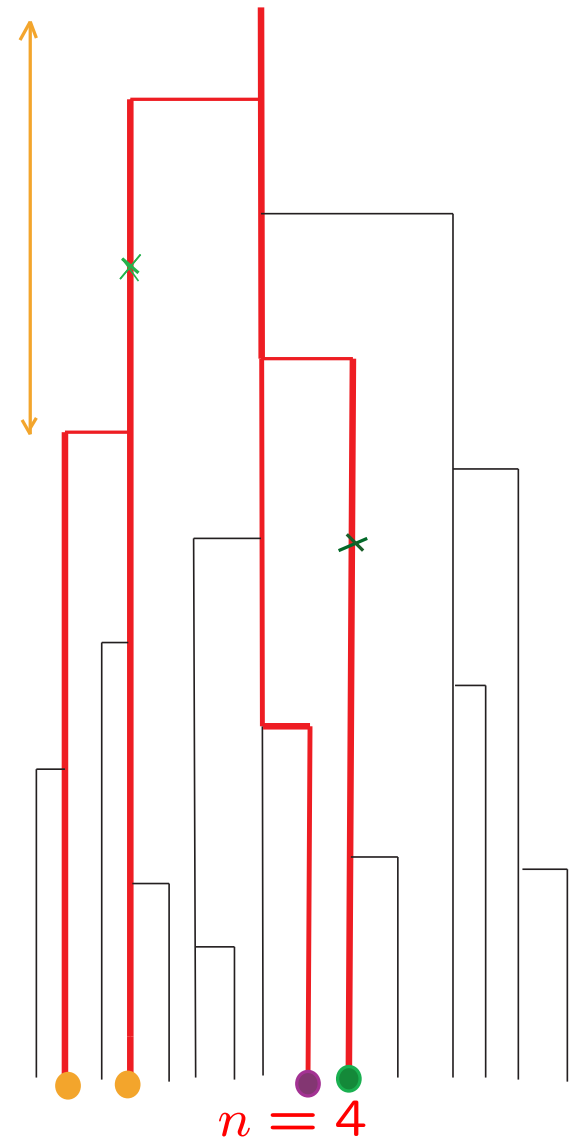


The branches of \mathcal{Y}_n
are marked
with a Poisson point process
with rate $\rho = \gamma \frac{\alpha}{\log \alpha}$.



The branches of \mathcal{Y}_n
are marked
with a Poisson point process
with rate $\rho = \gamma \frac{\alpha}{\log \alpha}$.

In this way,
the n -sample is partitioned.



Proposition.

The distribution of the sample's partition into

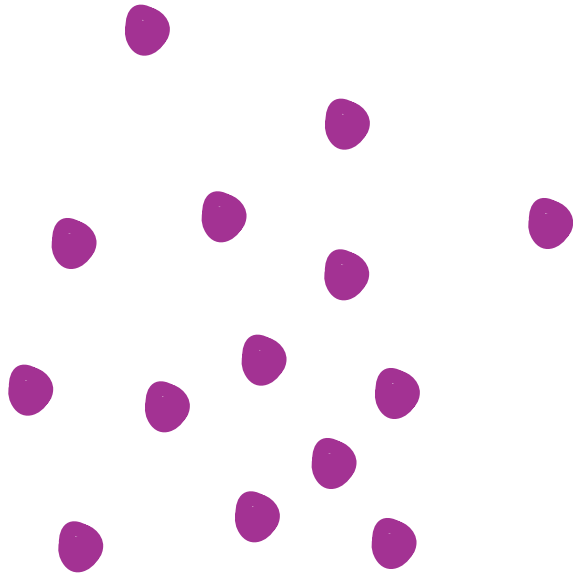
the founder's family,

the early recombinant families

and the late recombinant singletons

is approximated by the Yule Modell

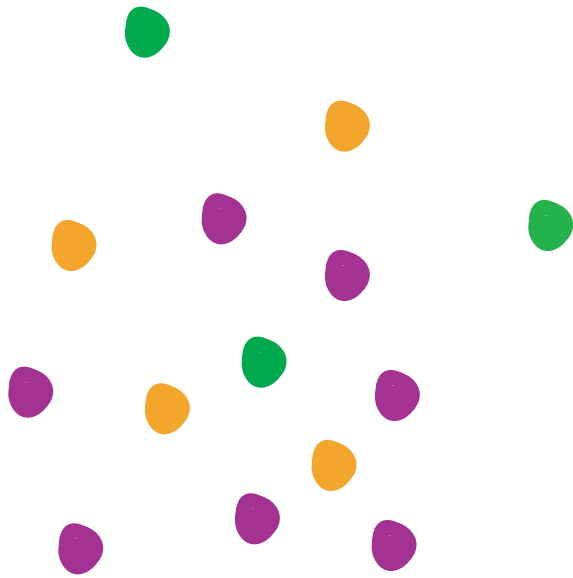
with an error $\mathcal{O}((\log \alpha)^{-2})$ in probability.



Without recombination
everybody in the sample would
be a **hitchhiker**



At the end of the early phase
some potential hitchhikers
are replaced by a number S
of early recombinants



Later a number
 L of late recombinants invades
the sample.

How can we compute the distribution of the sample partition
in the Yule model?

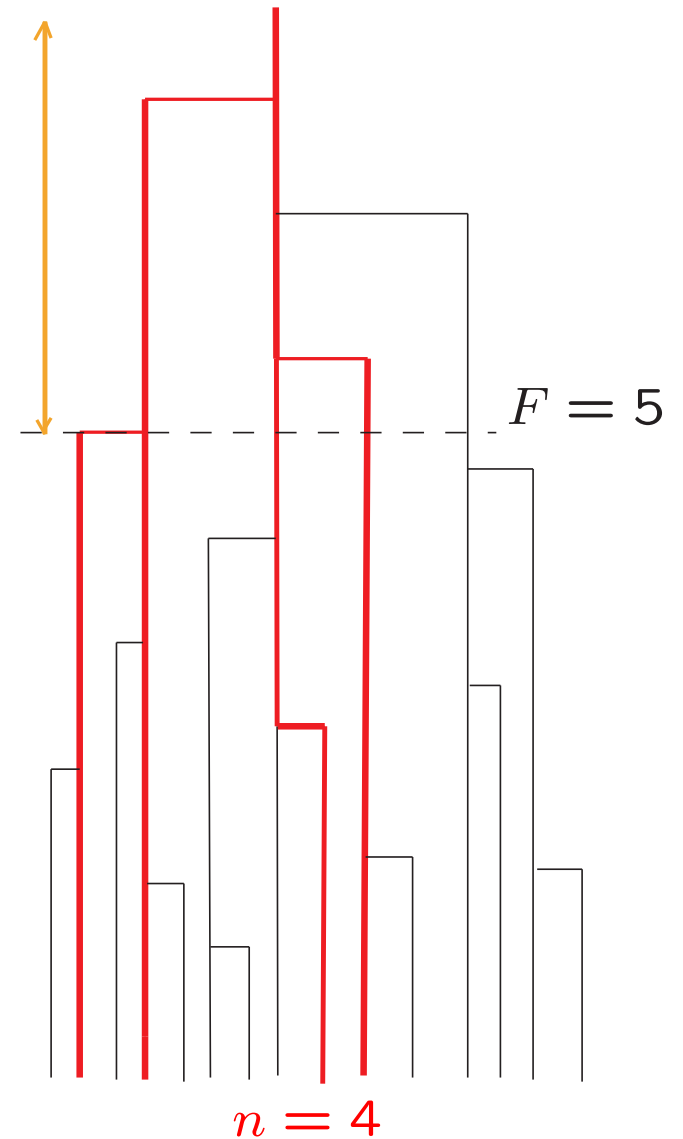
A crucial role is played by the random variable

$F :=$

the width of the Yule-tree \mathcal{Y} at the end of the early phase

Lemma:

$$\mathbf{P}[F \leq i] = \frac{(i-(n-1)) \cdots (i-1)}{(i+(n-1)) \cdots (i+1)}.$$



Given F , after the early phase
each of the n lines experiences a “late recombination”

with probability

$$\pi_F := 1 - \exp\left(-\frac{\gamma}{\log \alpha} \sum_{i=F}^{\alpha} \frac{1}{i}\right).$$

In other words:

The number L of *late recombinant singletons*

is mixed binomial,

with n trials

and random success probability π_F .

And what happens in the early phase?

Proposition

Up to an error of $\mathcal{O}((\log \alpha)^{-2})$ in probability
at the end of the early phase
there is no more than **one recombinant family**.

Its **size S** has the distribution

$$\mathbf{P}[S = 1] = \frac{\gamma n}{\log \alpha} \sum_{i=2}^{n-1} \frac{1}{i}, \quad \mathbf{P}[S \geq 2] = \frac{\gamma n}{\log \alpha};$$

$$\mathbf{P}[S = s | S \geq 2] = \frac{1}{s(s-1)}, \quad s = 2, \dots, n-1;$$

$$\mathbf{P}[S = n | S \geq 2] = \frac{1}{n-1}.$$

Theorem (Approximate sample partition at the neutral locus)

Let S and L be independent, with the above described distributions.

Let \mathcal{G} be a random choice of S elements from $\{1, \dots, n\}$

and independently let

\mathcal{L} be a random choice of L elements from $\{1, \dots, n\}$.

Up to an error of probability $\mathcal{O}((\log \alpha)^{-2})$ in probability
the sample,
partitioned after the neutral ancestors at the beginning of the sweep,
consists of

- L late recombinant singletos
- an early recombinant family $E = \#(\mathcal{G} \setminus \mathcal{L})$
- the founder's family of size $H = n - (L + E)$.

Resumé

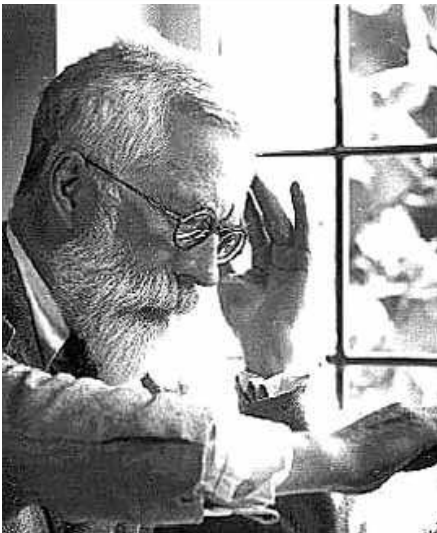
Three steps in the analysis of the random sample partition:

1. From the structured to the marked coalescent.
2. From the marked coalescent to the marked Yule tree,
3. Analysis of the sample partition in the marked Yule tree.

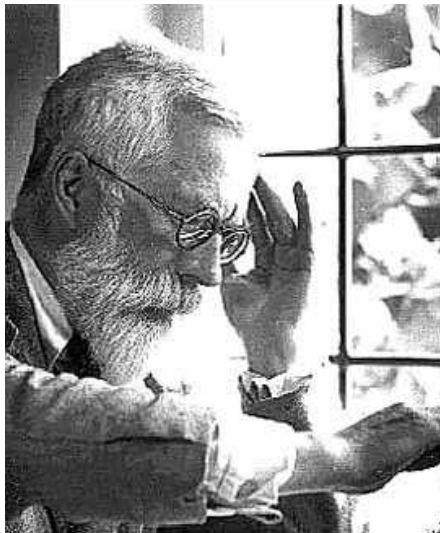
What is behind the passage
from the marked coalescent in the background *P*
to the Yule tree?

What's the background of the Yule tree?

Idea: Replace Wright-Fisher by



Idea: Replace Wright-Fisher by



William Feller, 1906-1970

by a random time change depending on the sweep path P

$$d\tau = (1 - P) dt$$

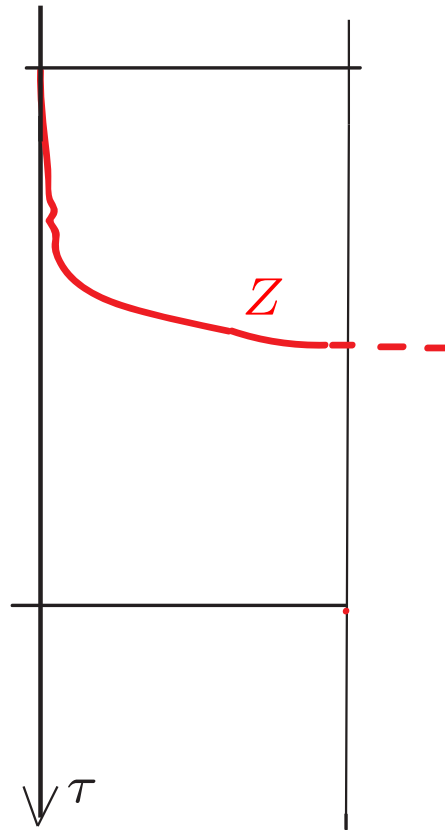
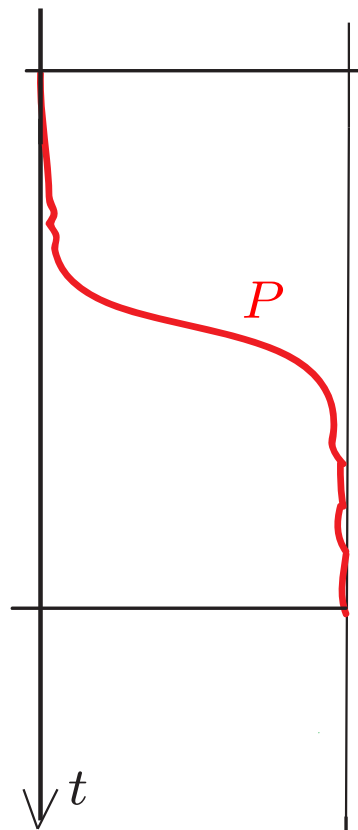
This translates the dynamics of the sweep path

$$dP = \sigma \sqrt{P} \sqrt{(1 - P)} dW + \coth\left(\frac{\alpha}{\sigma^2}\right) \alpha P(1 - P) dt$$

into

$$dZ = \sigma \sqrt{Z} d\tilde{W} + \coth\left(\frac{\alpha}{\sigma^2}\right) \alpha Z d\tau$$

that is the excursion of a Feller diffusion starting in $Z_0 = 0$ and conditioned to non-extinction:



This time change transforms

- the marking rate $\rho (1 - P)dt$ into $\rho d\tau$,

- the pair coalescence rate $\frac{1}{P} dt$
into $\frac{1}{Z} \frac{1}{1-Z} d\tau$.

The Yule tree \mathcal{Y} can be seen as a
coalescent in the varying population size Z ,
that is with pair coalescence rate $\frac{1}{Z}d\tau$,
averaged over Z .

And the coalescences of the sample ancestral lineages happen
when P and Z are small
– which makes the factor $\frac{1}{1-Z}$ asymptotically negligible.

Nick Barton (1998):*

“... a selective sweep can be divided into four phases:

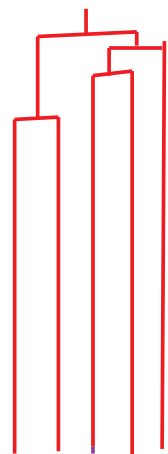
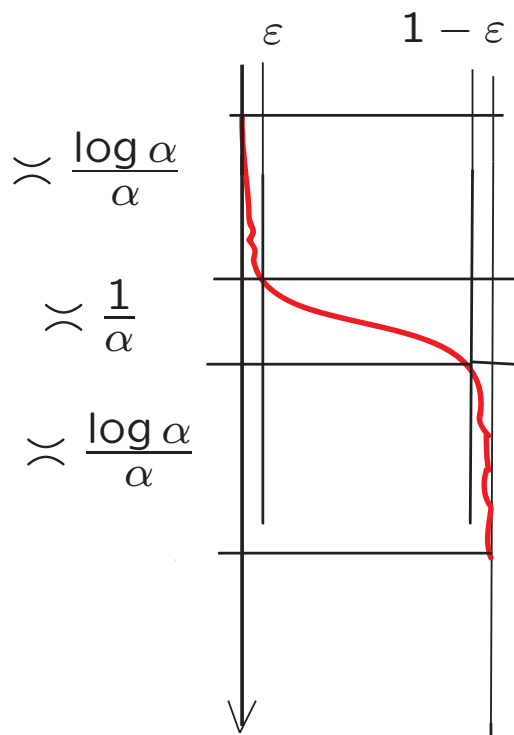
After a favourable mutation arises, its numbers, k , fluctuate in a branching process (1st phase).

After some randomly distributed time, these will reach a large enough number to increase deterministically; in this, the 2nd phase, numbers are large enough for genetic drift to be negligible, but the allele frequency is still low.

In the 3rd phase, which lasts $\sim 1/s (= 2N/\alpha)$ generations, the new allele sweeps to high frequency.

Finally, the original allele is eliminated, leading to fixation (4th phase).”

*The effect of hitch-hiking on neutral genealogies, Gen. Res. 72: 123-133



Some references:

J. Maynard Smith, H.M. Haigh, The hitch-hiking effect of a favorable gene. *Genetic Research*, 23: 23-35, 1974.

N. L. Kaplan, R. Hudson, C.H. Langley, The 'Hitchhiking effect' revisited. *Genetics* 123: 887 -899, 1989.

W. Stephan, T. Wiehe, M. Lenz, The effect of strongly selected substitutions on neutral polymorphism: analytic results based on diffusion theory. *Theoret. Pop. Biol.* 41: 237-254, 1992.

N. Barton, The effect of hitch-hiking on neutral genealogies, *Gen. Res.* 72: 123-133, 1998.

J. Schweinsberg, R. Durrett, Random partitions approximating the coalescence of lineages during a selective sweep, *Ann. Appl. Probab.* 15: 1591-1651, 2005

A. Etheridge, P. Pfaffelhuber, A. Wakolbinger, An approximate sampling formula under genetic hitchhiking, <http://arxiv.org/abs/math.PR/0503485>, to appear in *Ann. Appl. Probab.*