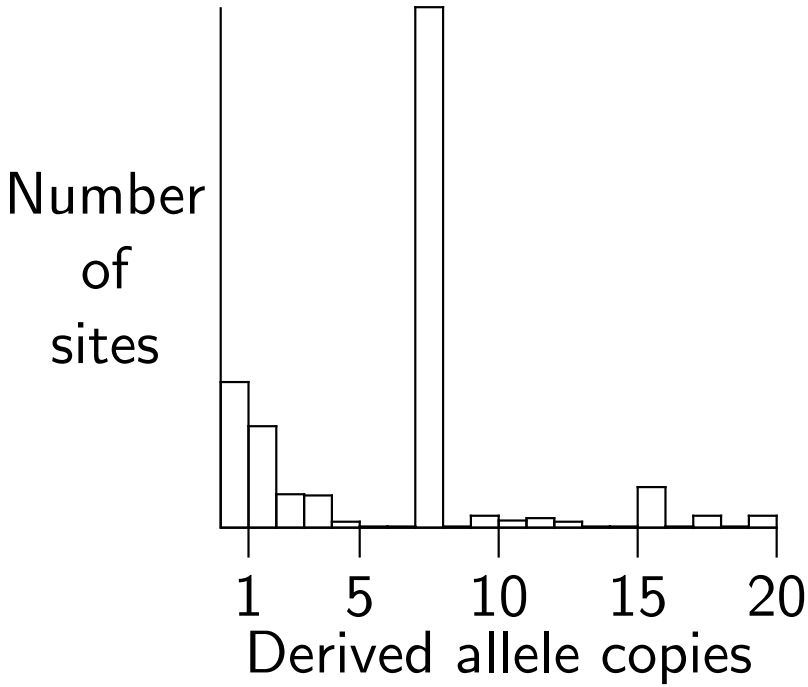
The Site Frequency Spectrum

Alan R. Rogers

February 7, 2023

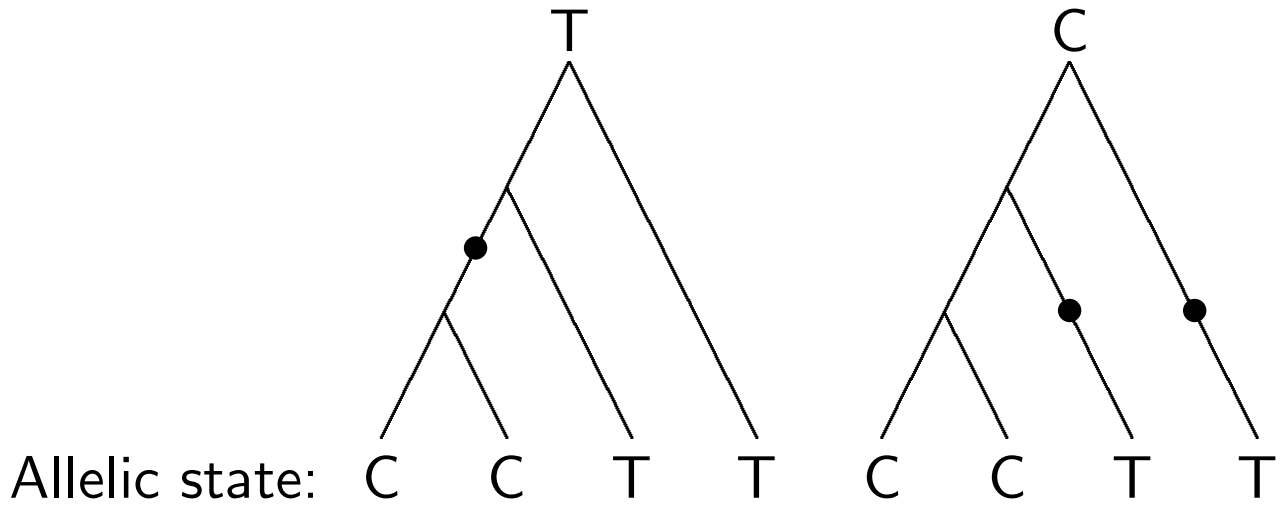
A site frequency spectrum



Y chromosome data from Underhill et al.

$K = 718$ DNA sequences

Calling ancestral and derived alleles



- Two hypotheses about which allele is ancestral.
- "C" requires 2 mutations; "T" requires 1.
- Because mutations are rare, "T" is more likely.
- When the in-group is polymorphic, the ancestral allele is usually the one present in the out-group.

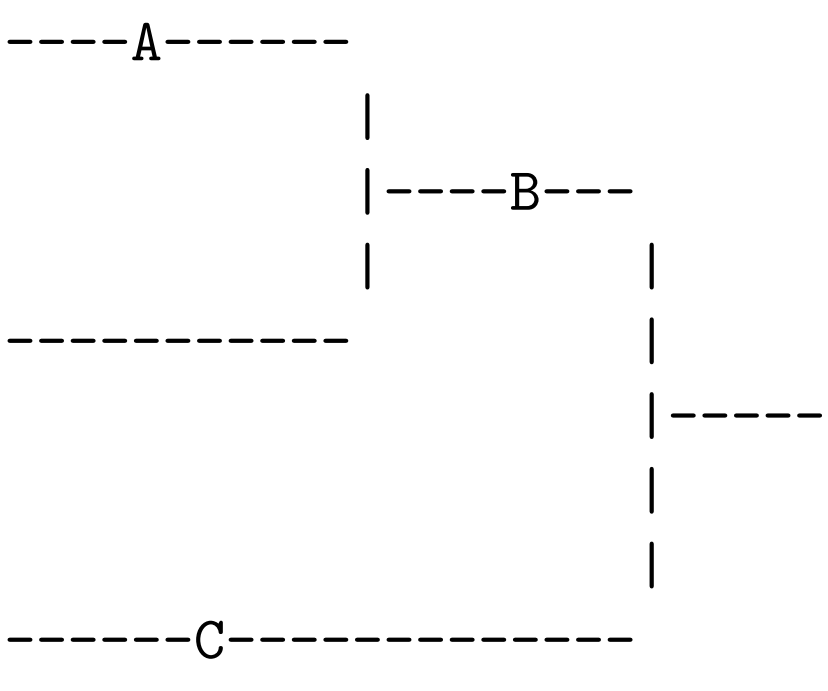
Unfolded site frequency spectrum

```
        123456
Human1  AATAGC
Human2  ..AC..
Human3  .TACT.
Human4  ..ACT.
------------
Chimp   AAAATC
```

1: fixed.
2: T derived; singleton.
3: T derived; singleton.
4: C derived; tripleton.
5: G derived; doubleton.
6: fixed.

Singletons	2
Doubletons	1
Tripletons	1

A site's position in spectrum depends on position in gene tree



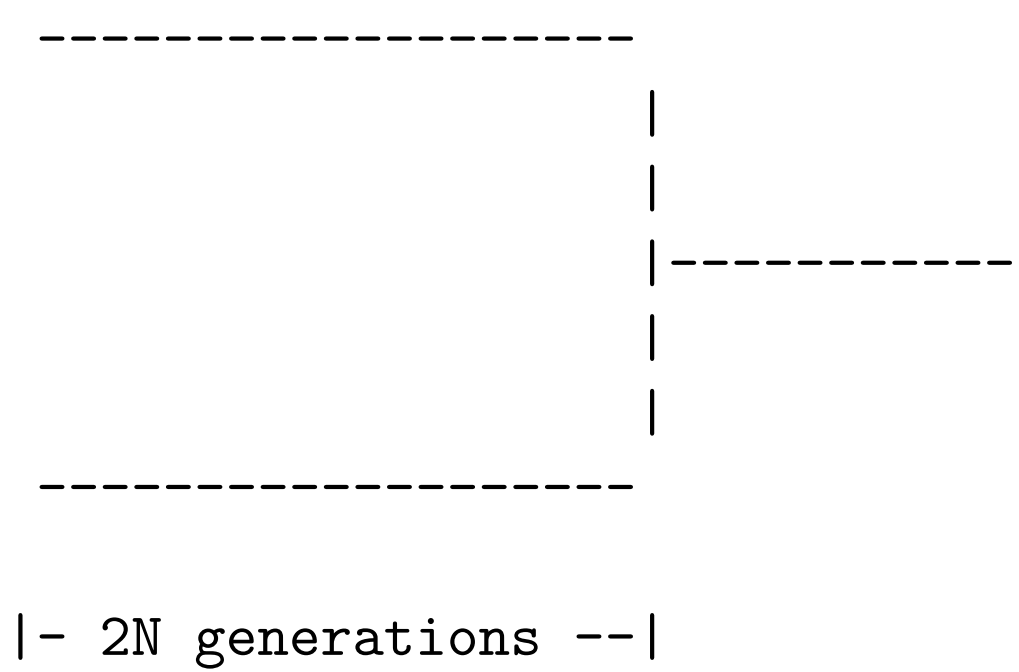
Mutations A and C are singletons; B is a doubleton

Most recent interval: singletons only

2nd most recent: singletons and doubletons

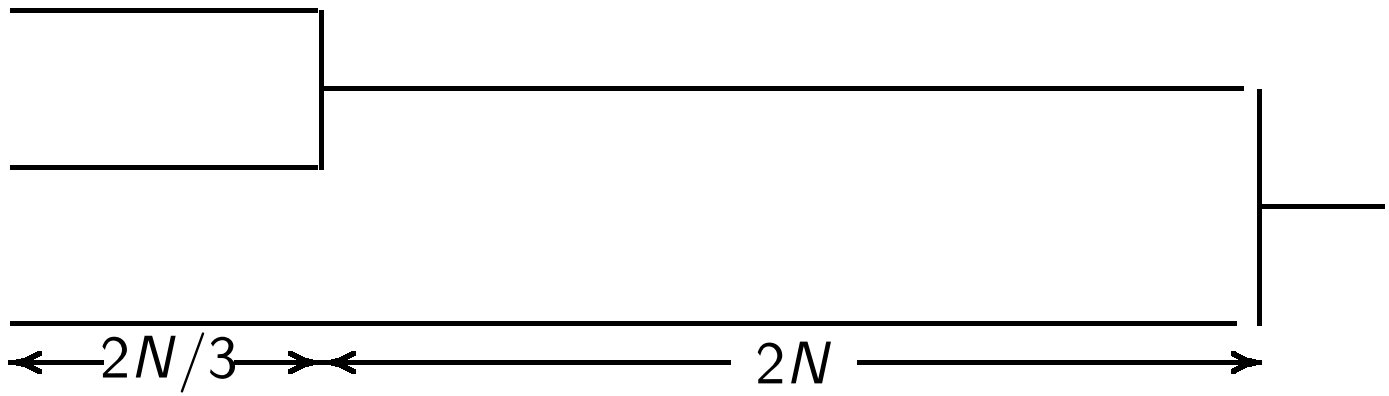
3rd most recent: singletons, doubletons, and tripletons

A tree with 2 leaves has only singletons



We expect $4Nu = \theta$ mutations, all singletons.

3 leaves $\Rightarrow \theta$ singletons and $\theta/2$ doubletons



Expect θ singletons to arise during oldest interval.

Half of these become doubletons at coalescent event: we end up with $\theta/2$ doubletons.

New singletons in recent interval: $\frac{2N}{3} \times 3 \times u = 2Nu = \theta/2$. This gain exactly compensates for the singletons that became doubletons.

Expected spectrum in population of constant size

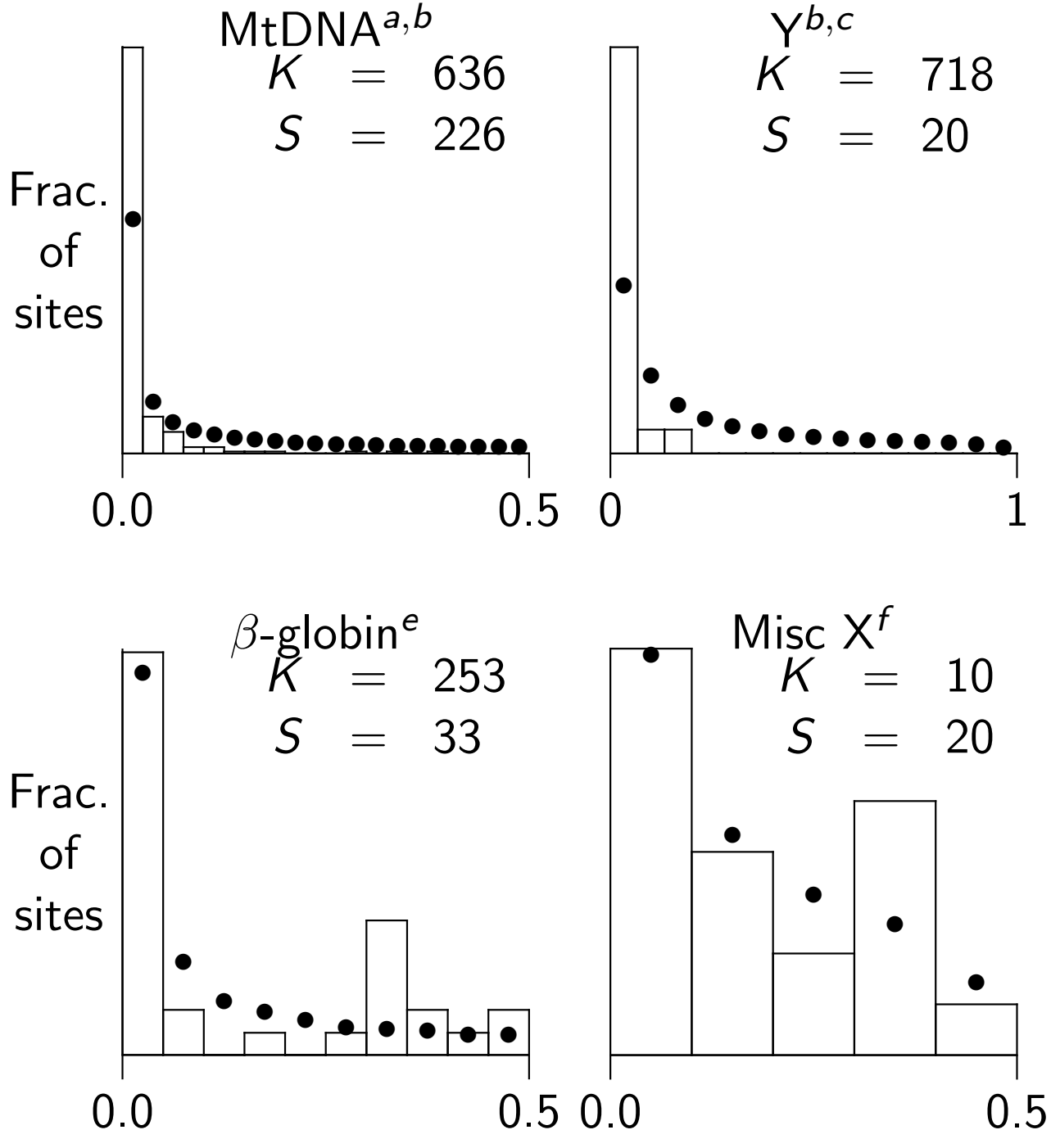
Sample size	Expected spectrum (singletons, doubletons, ...)
2	θ
3	$\theta, \quad \theta/2$
4	$\theta, \quad \theta/2, \quad \theta/3$
5	$\theta, \quad \theta/2, \quad \theta/3, \quad \theta/4$
	Etcetera

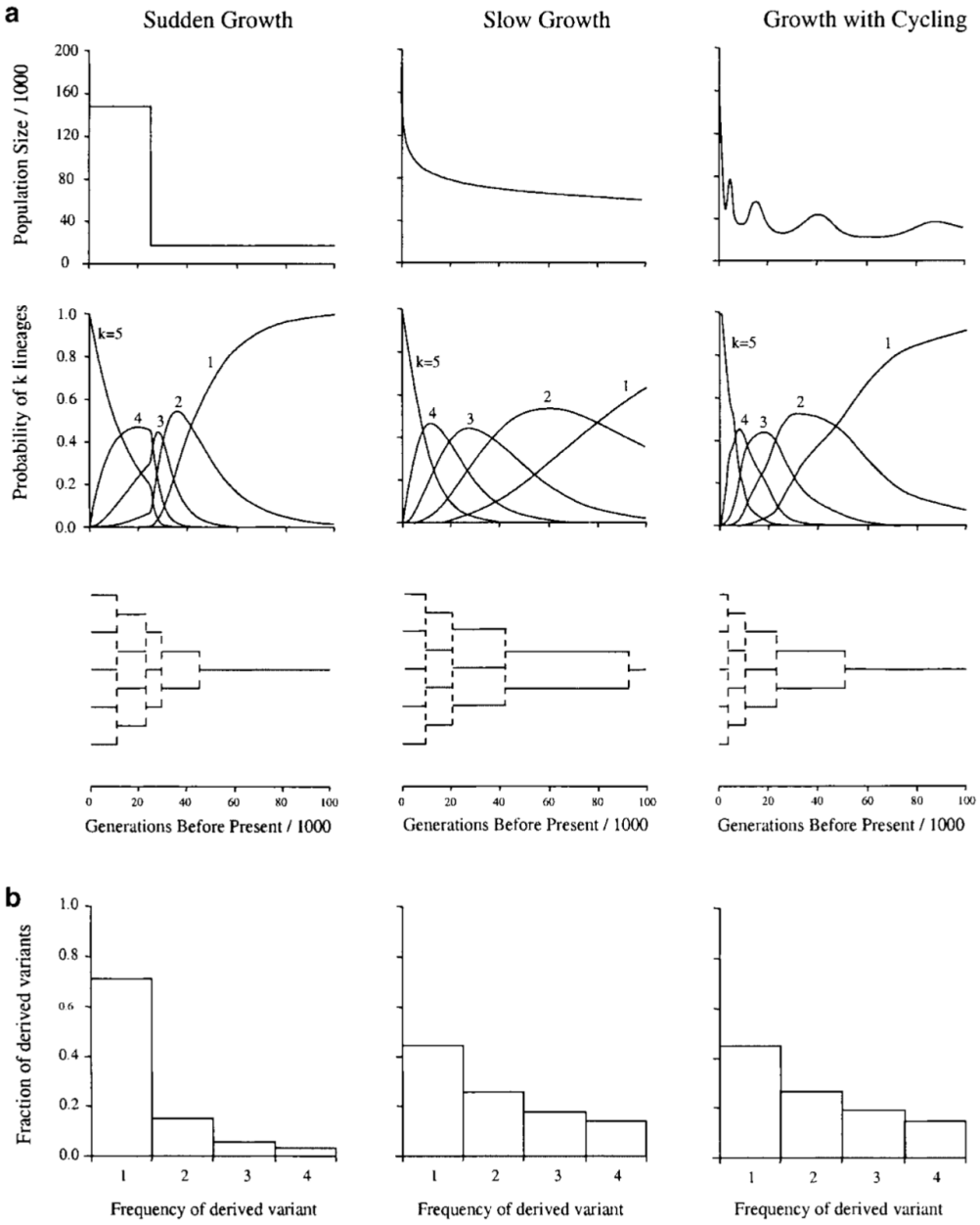
It is remarkable that as we increase sample size, the number of mutants in each category doesn't change. We merely add a new category at the right side of the spectrum.

Details: Rogers & Wooding 2021

How to estimate θ?

To use this formula with data, we need an estimate of θ.
The sum of the observed spectrum is the number, S, of segregating sites.
This is true of the theoretical spectrum only if you use $\hat{\theta}_S$ to estimate θ.


MtDNA[a,b]
K = 636
S = 226
Y[b,c]
K = 718
S = 20
β-globin[e]
K = 253
S = 33
Misc X[f]
K = 10
S = 20
Frac. of sites
Frac. of sites
0.0
0.5
0
1
0.0
0.5
0.0
0.5



When population size varies

Summary

- Under neutrality and constant population size, θ/i is the expected number of sites at which the derived allele is present in i copies.
- This does not depend on sample size.
- Some human loci conform to this model; many do not.
- Departures imply something interesting: selection or changes in population size.