




The Big Picture of Y STR Patterns

Rob Spencer

The 14th International Conference on Genetic Genealogy
Houston March 22-24, 2019



Many Thanks

**Maurice Gleeson
David Langton
Iain McDonald**

and friends at the very active
England GB Groups EIJ project

and of course
Family Tree DNA

Zoom In for Details



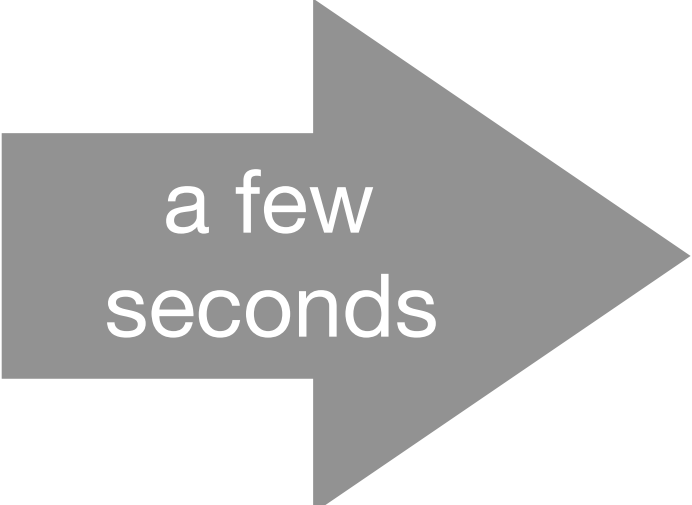
Zoom Out for Context



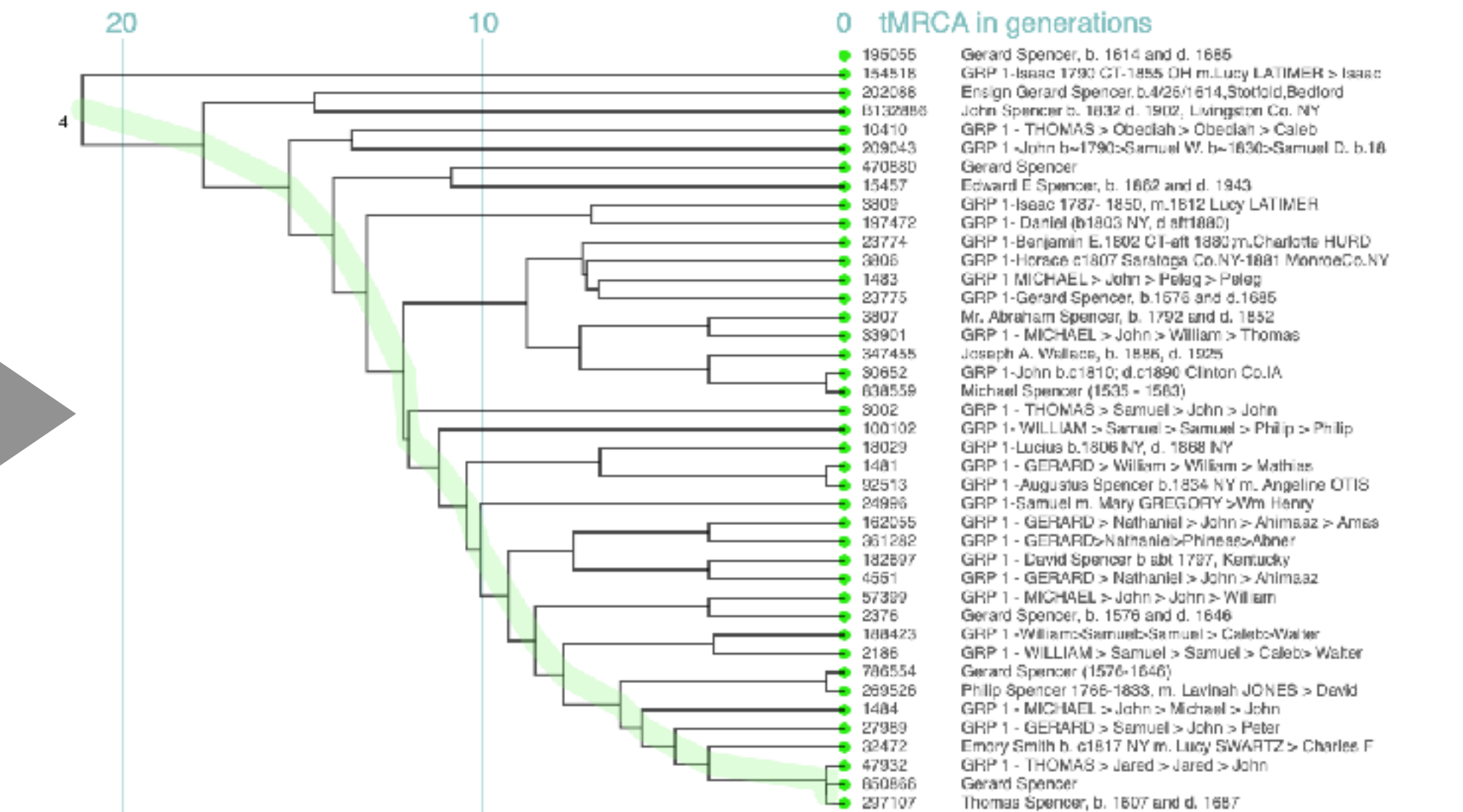
Methodology

FTDNA Y STR table

ID	Country	Y-STR	STR	Y-STR	STR	Y-STR	STR	Y-STR	STR	Y-STR	STR	Y-STR	STR	Y-STR	STR	Y-STR	STR				
114253	UK	D72	CT81	0028	86346	L22	Z2238	M1483	Young	Southam	Spalding	Northam	Spalding	England	I-M258	1322	15	1013-1411141112112019	8-9	1120162028	12-14-15
114253	UK	D72	CT81	0028	86346	L22	Z2238	M1483	Young	Southam	Spalding	Northam	Spalding	England	I-M258	1322	15	1013-1411141112112019	8-9	1120162028	12-14-15
114253	UK	D72	CT81	0028	86346	L22	Z2238	M1483	Young	Southam	Spalding	Northam	Spalding	England	I-M258	1322	15	1013-1411141112112019	8-9	1120162028	12-14-15
114253	UK	D72	CT81	0028	86346	L22	Z2238	M1483	Young	Southam	Spalding	Northam	Spalding	England	I-M258	1322	15	1013-1411141112112019	8-9	1120162028	12-14-15
114253	UK	D72	CT81	0028	86346	L22	Z2238	M1483	Young	Southam	Spalding	Northam	Spalding	England	I-M258	1322	15	1013-1411141112112019	8-9	1120162028	12-14-15



dendrogram



DIY at <http://scaledinnovation.com/gg/yClustering.html>

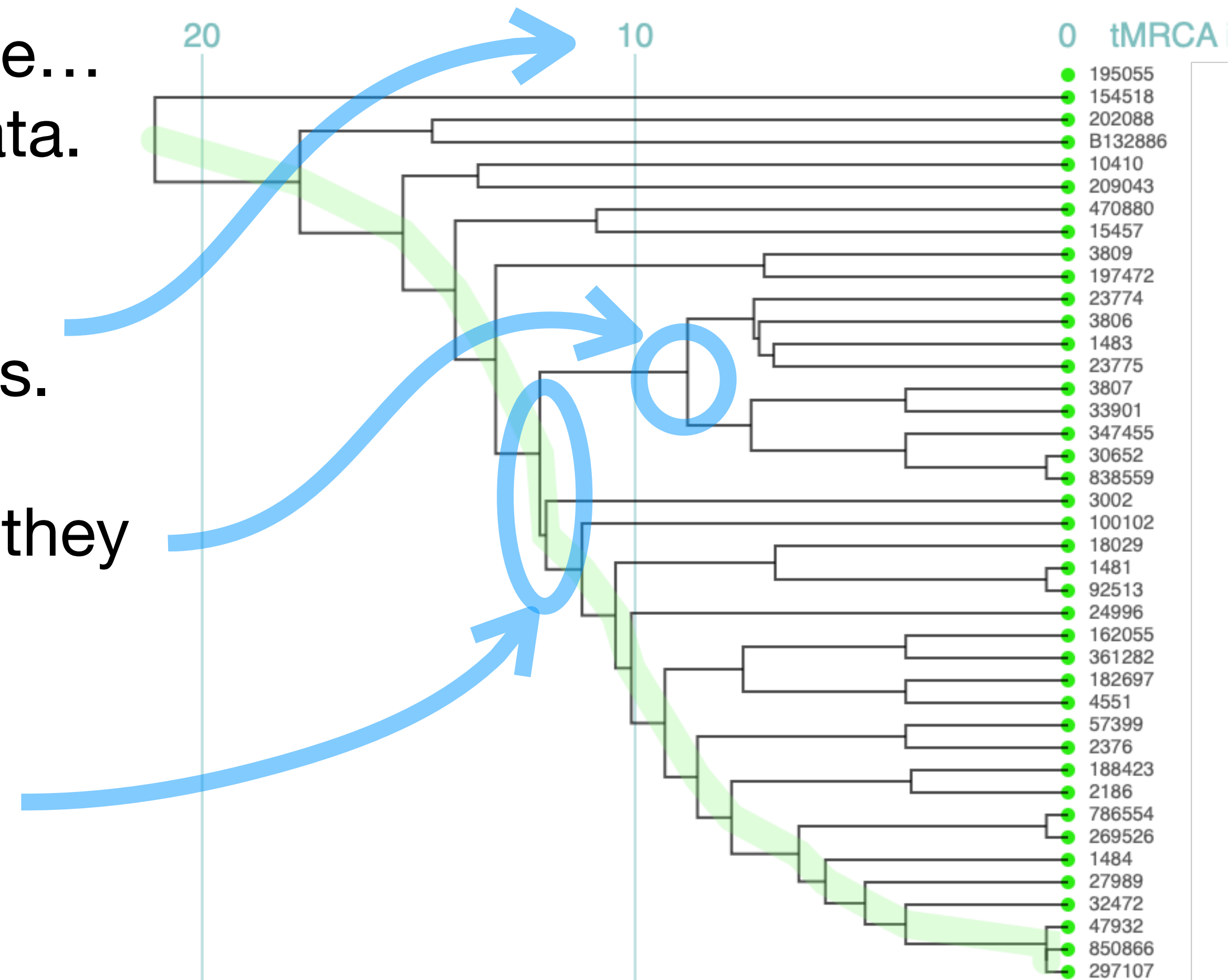
Dendrograms

A distance dendrogram is not a family tree... but it will converge to one with perfect data.

Its date scale is quite good — based on averages of hundreds of pairwise tMRCA.

Its branch-points aren't ancestors — but they indicate where they might be.

Its topology can be wrong — especially where branches are very close in time.



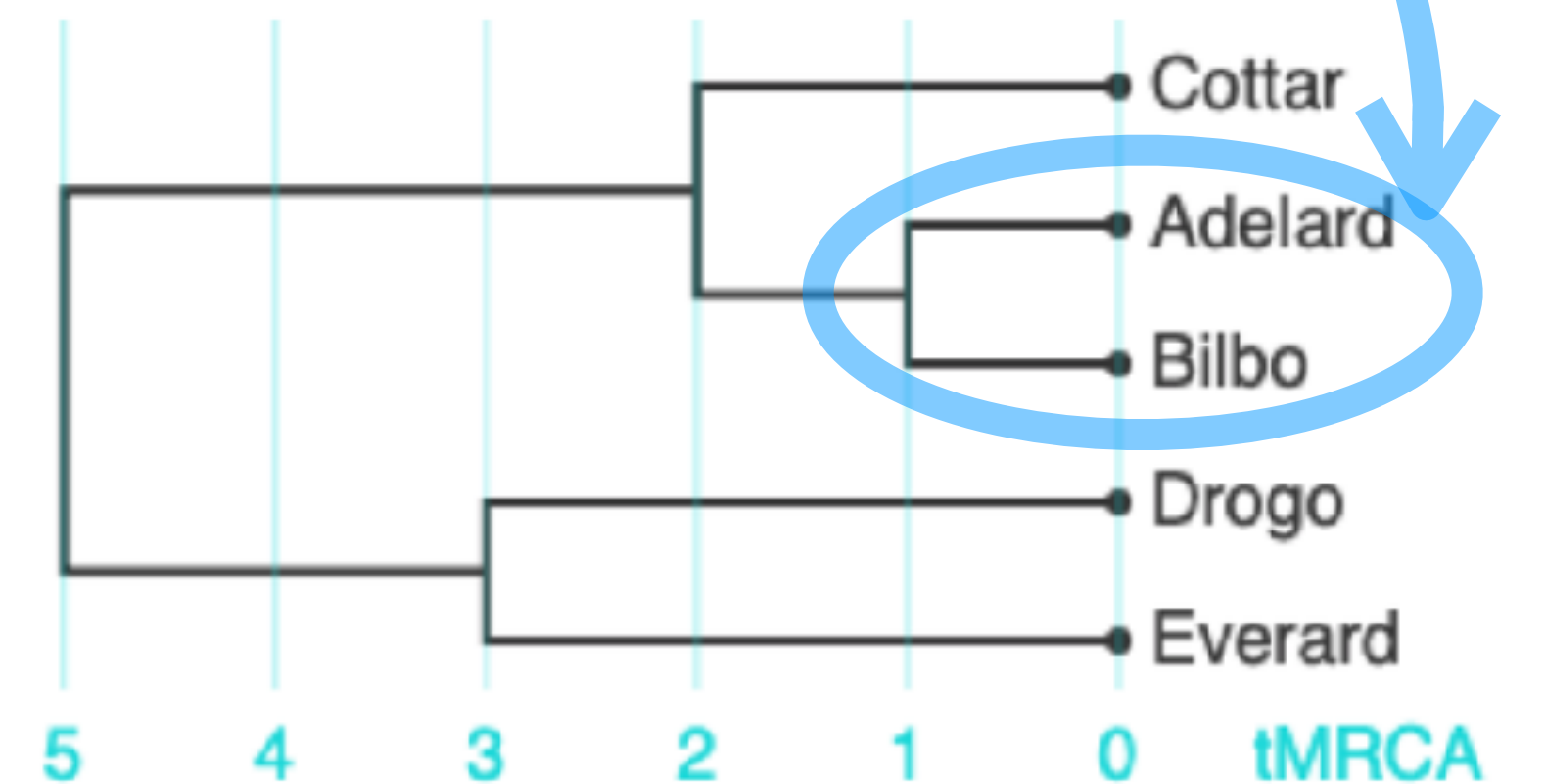
How to build a dendrogram

1. Compute all $n(n-1)/2$ pairwise distances between kits. I use tMRCA as a distance measure because it's linear and additive.
2. Find the two closest kits. Remove them from the list, combine them into a new node, and add this node to the list.
3. Repeat (2) until there's one node left, the root.

The result is a **tree structure** which lends itself directly to computer graphics.

all pairwise distances in generations

	Adelard	Bilbo	Cottar	Drogo	Everard
Adelard	-	1.00	2.00	5.00	5.00
Bilbo	1.00	-	2.00	5.00	5.00
Cottar	2.00	2.00	-	5.00	5.00
Drogo	5.00	5.00	5.00	-	3.00
Everard	5.00	5.00	5.00	3.00	-



See <http://scaledinnovation.com/gg/treeDemo.html>

Two types of computer-created trees

Maximum Parsimony

- Builds tree consistent with fewest mutations
- Better topology, often completely correct
- Pattern-matching method does not include dates
- Computationally complex, limited in scale

a microscope

Distance

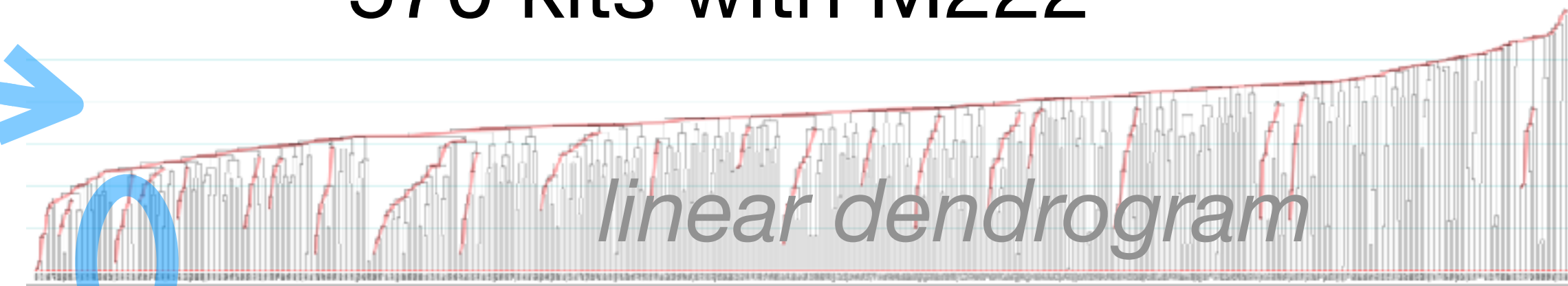
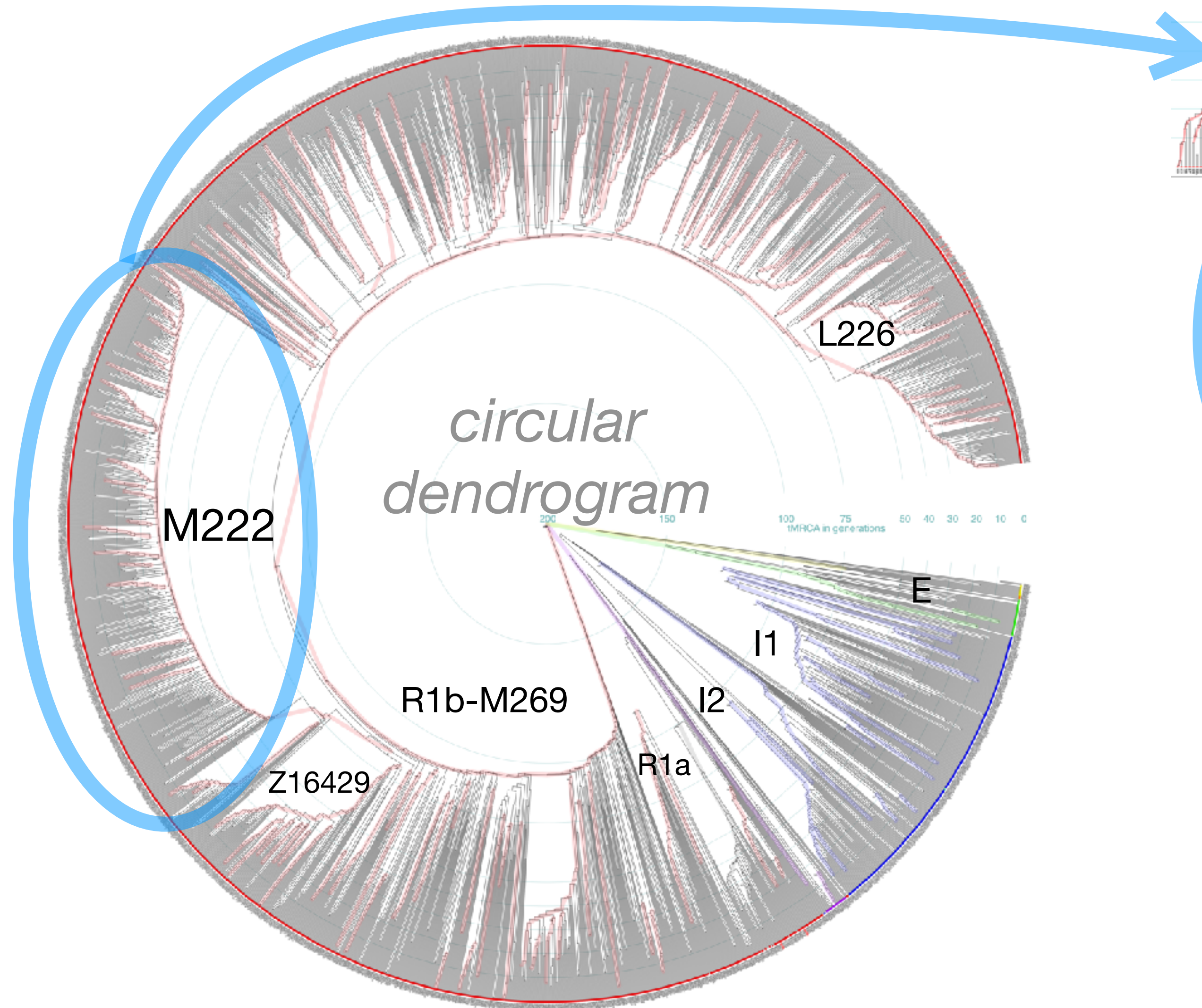
- Builds tree consistent with pairwise distances
- Topology may be incorrect in details
- Intrinsic date scale
- Computationally simple, scales to thousands of kits

a “macroscope”

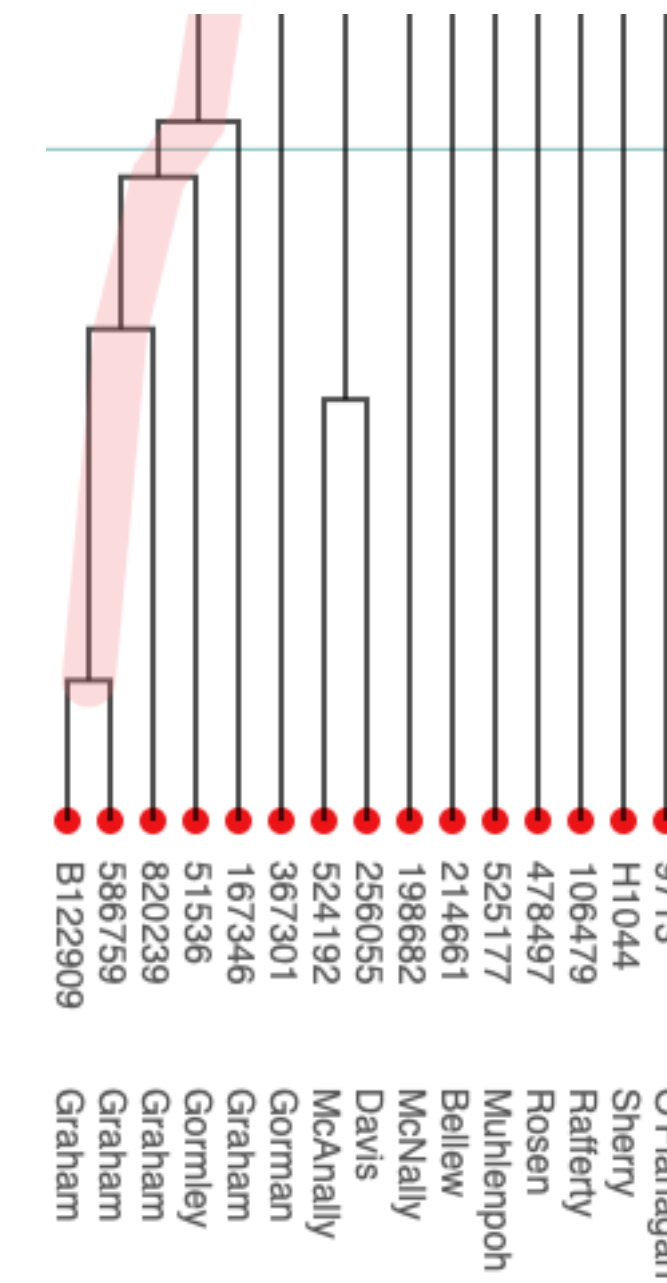
Examples

IrelandDNA 2850 kits at Y111

570 kits with M222



individual details



past
↑
time scale
|
present

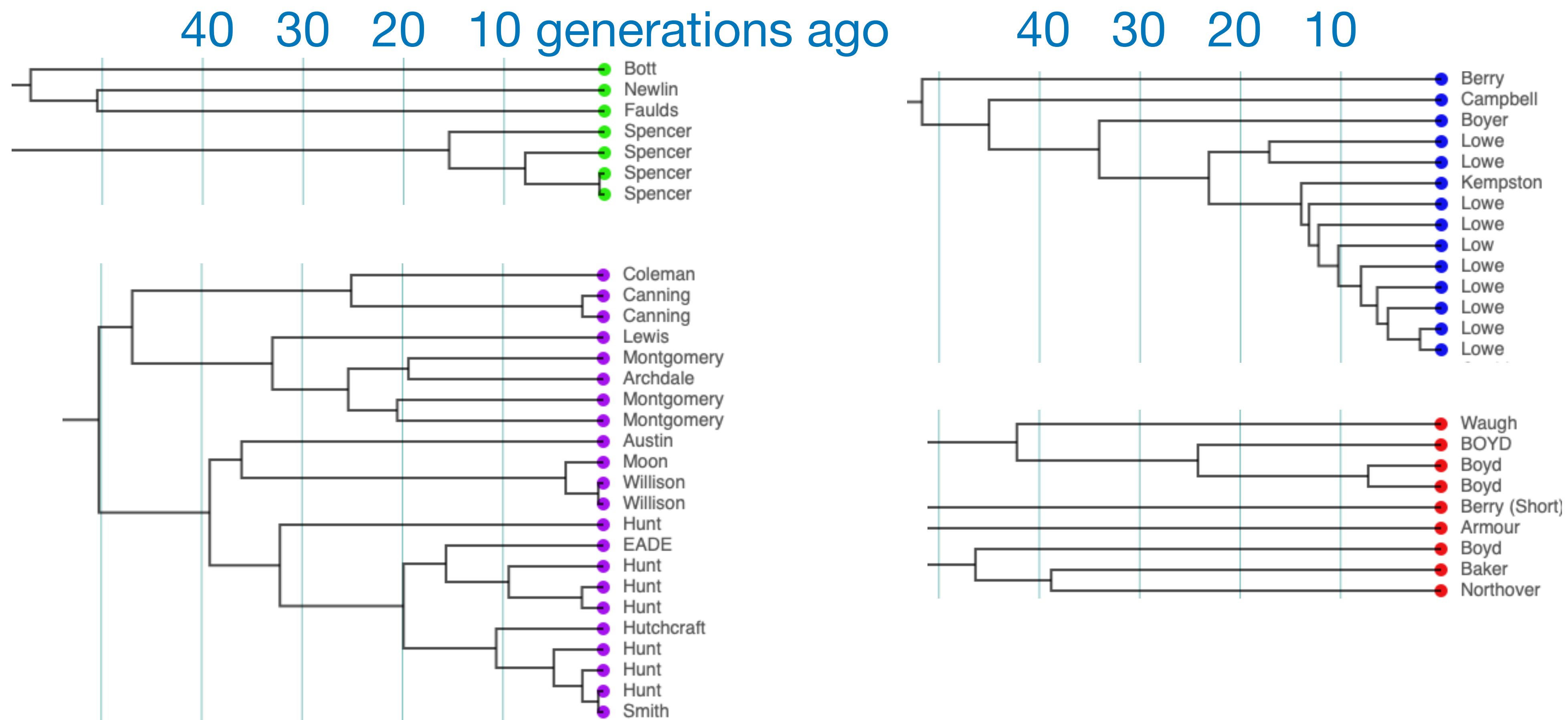
kit numbers, ancestor names

One tool, many
things to see.



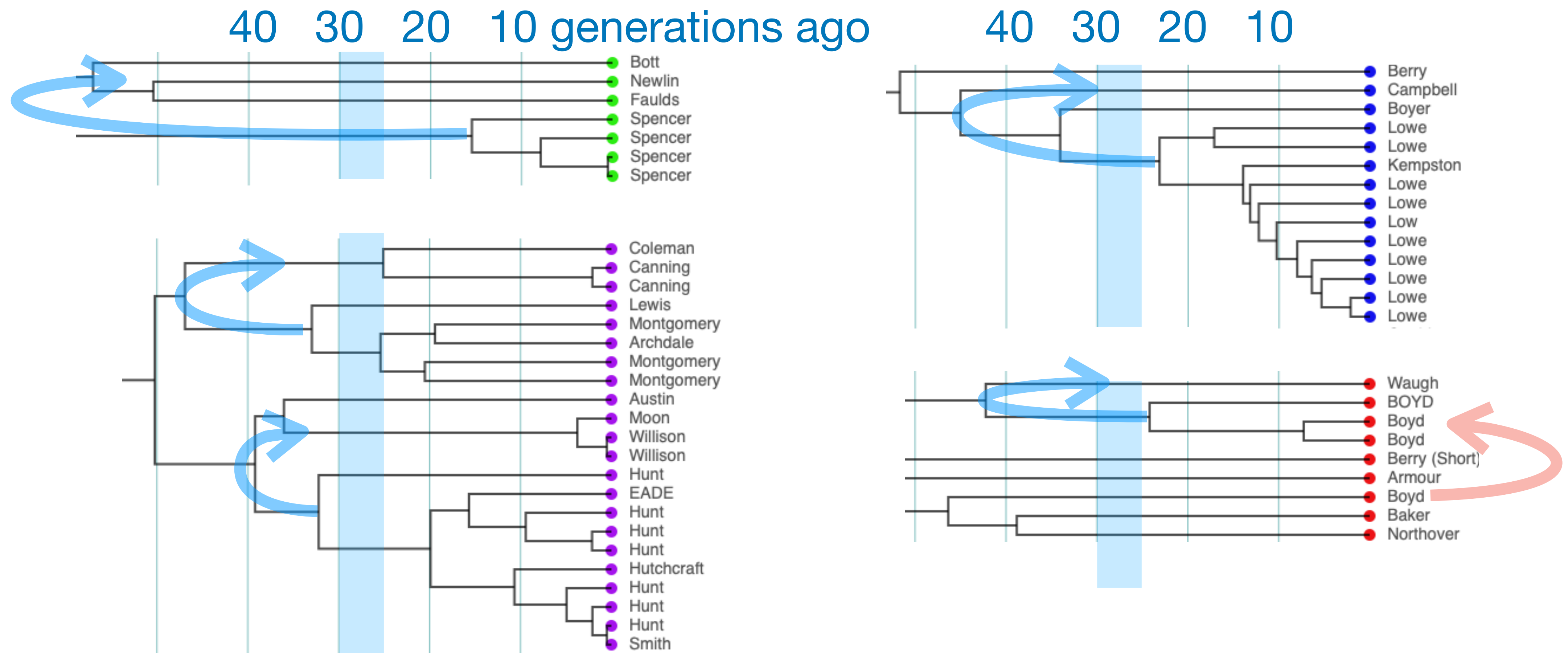
Spot the Onset of Surnames

Expect a common surname only for branches with tMRCA < 25-30 generations. Otherwise surnames are essentially **random**.

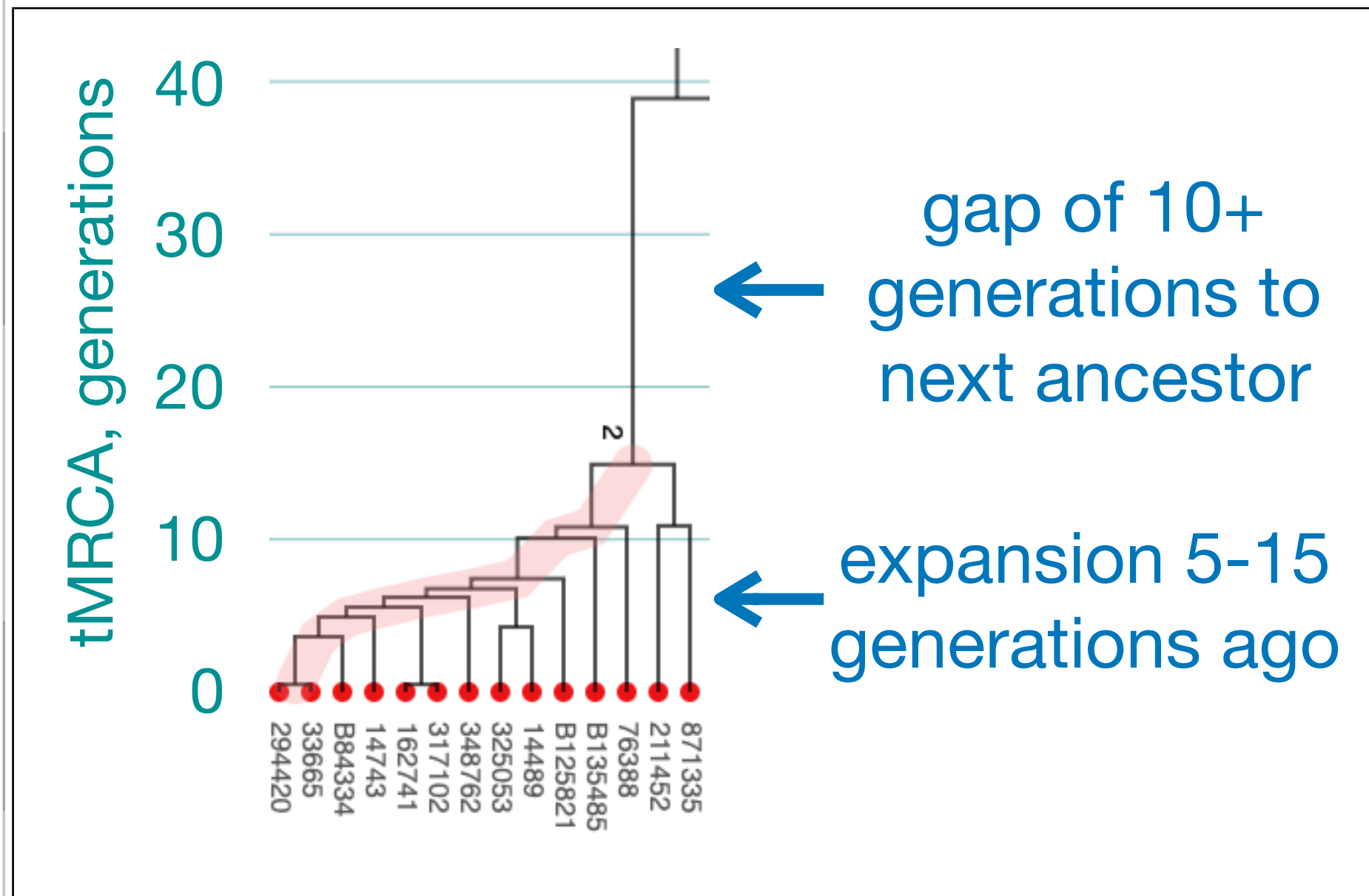


Spot the Onset of Surnames

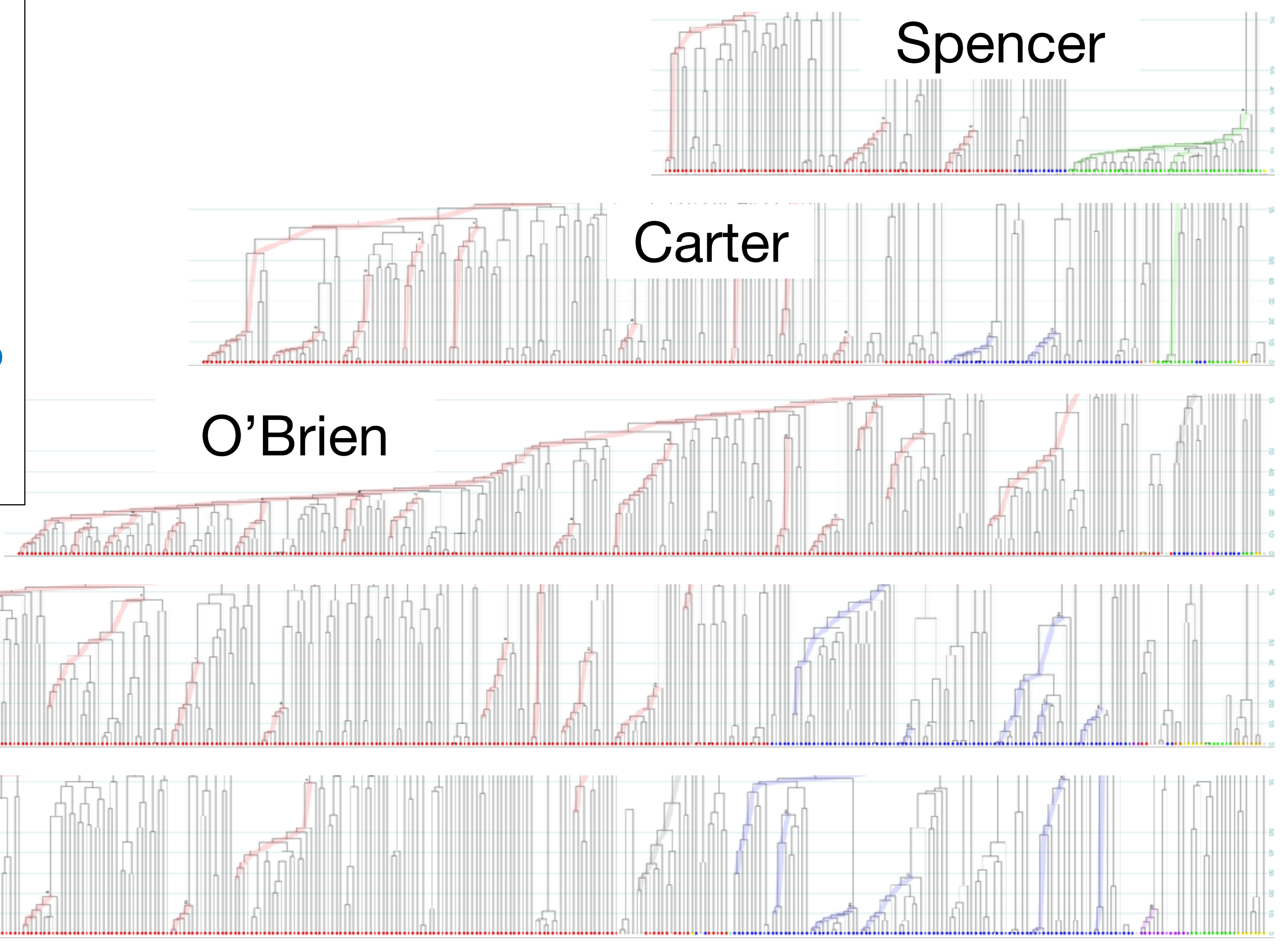
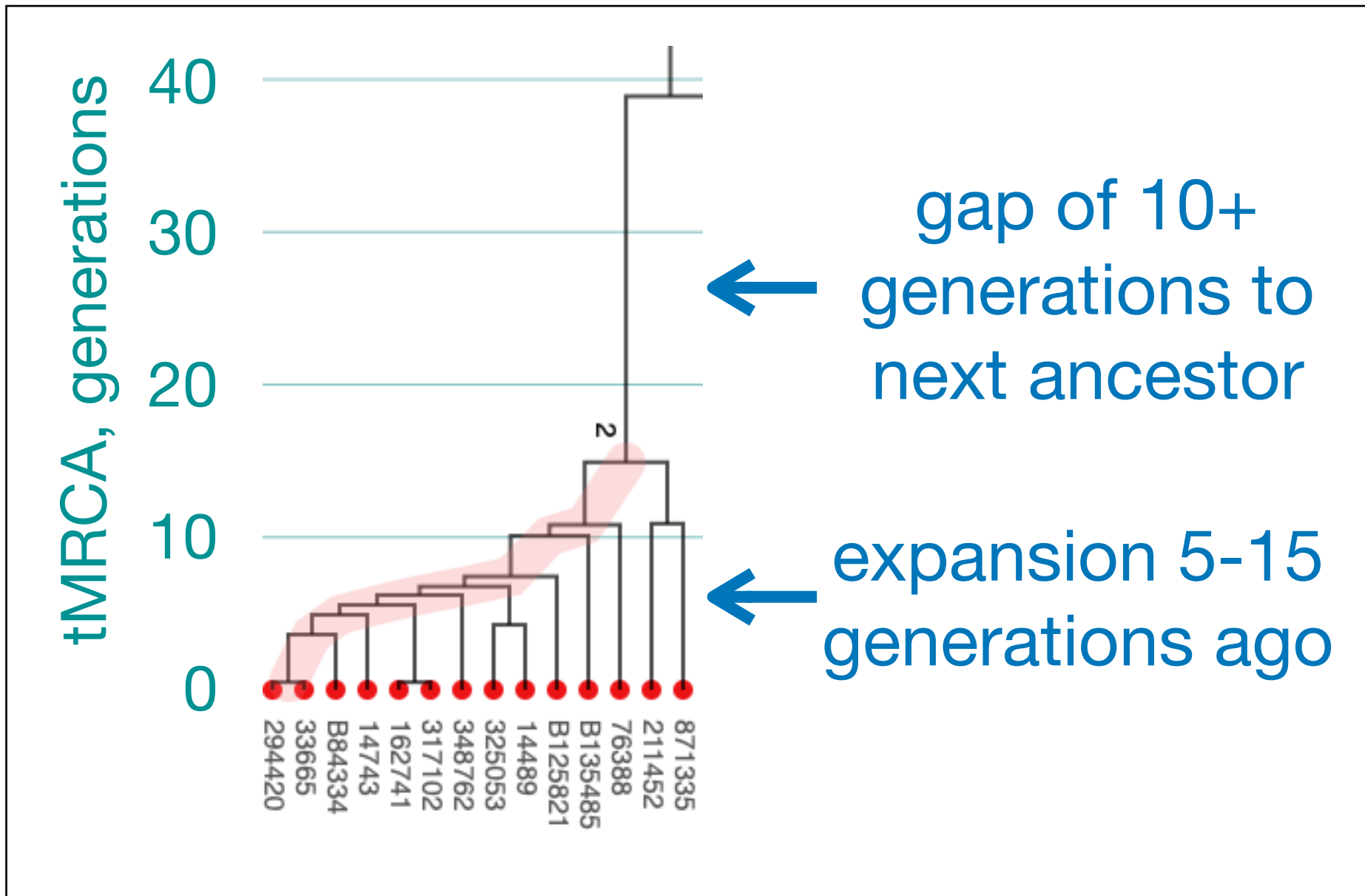
Expect a common surname only for branches with tMRCA < 25-30 generations. Otherwise surnames are essentially **random**.



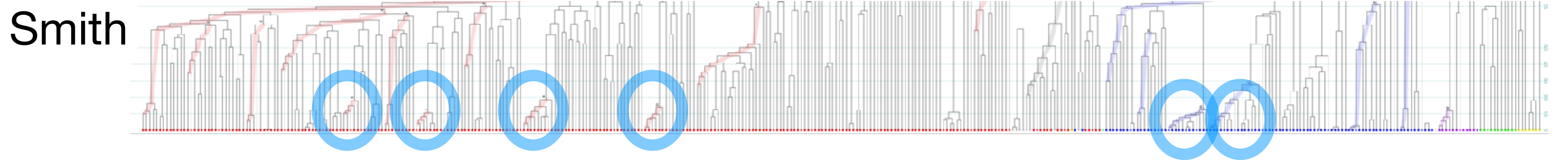
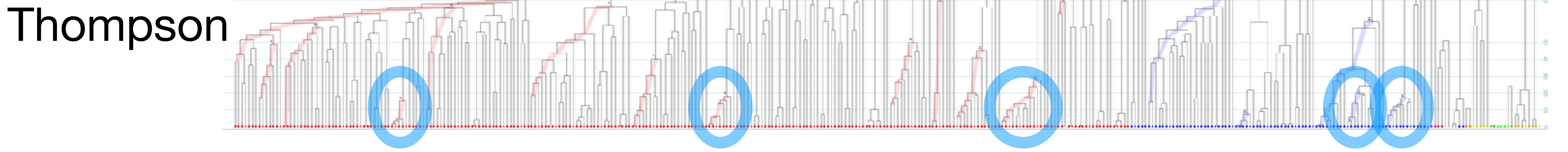
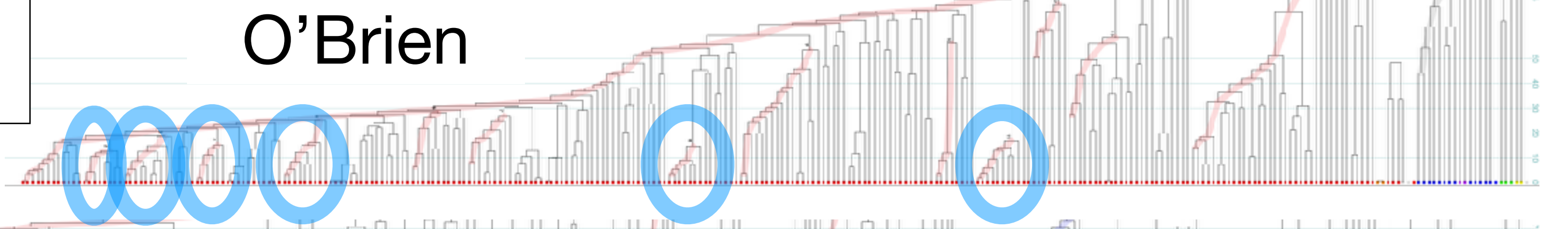
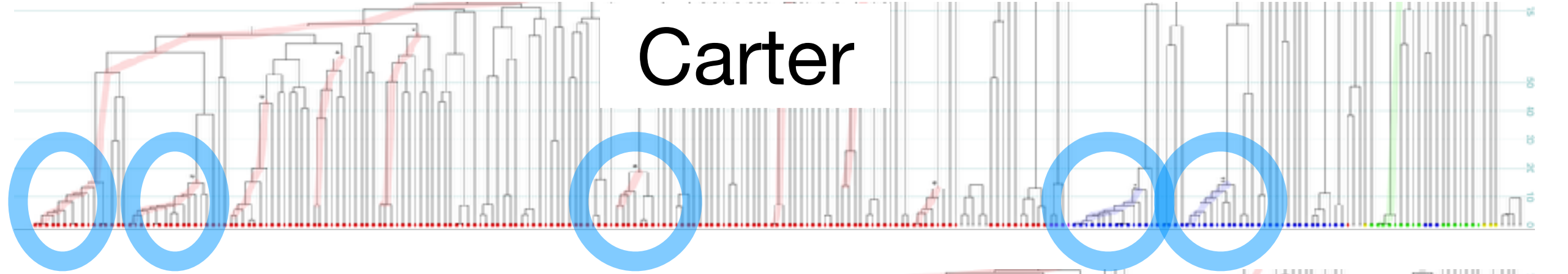
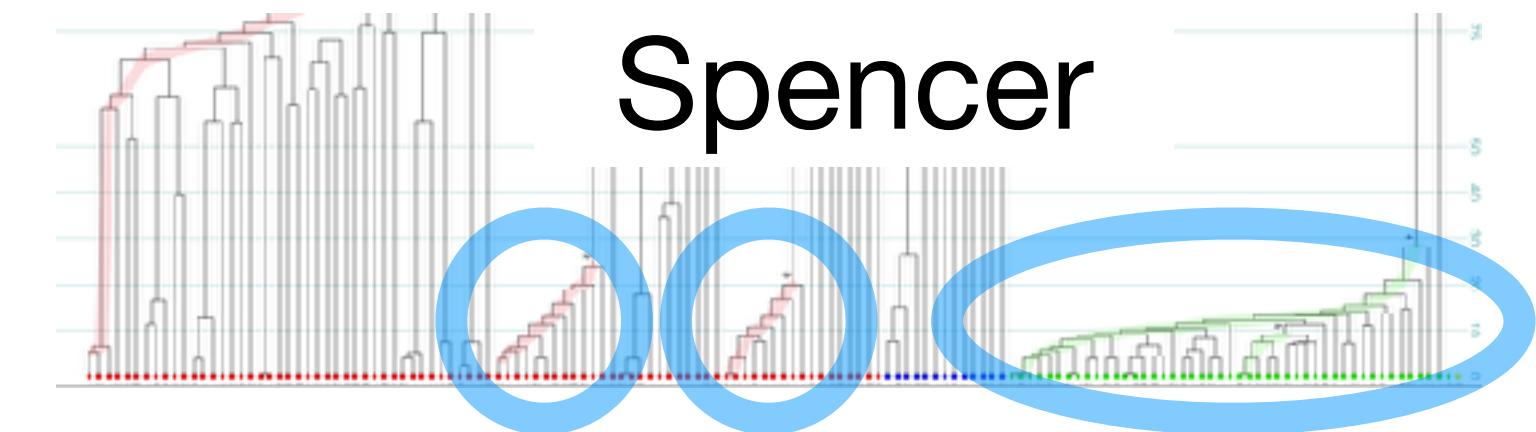
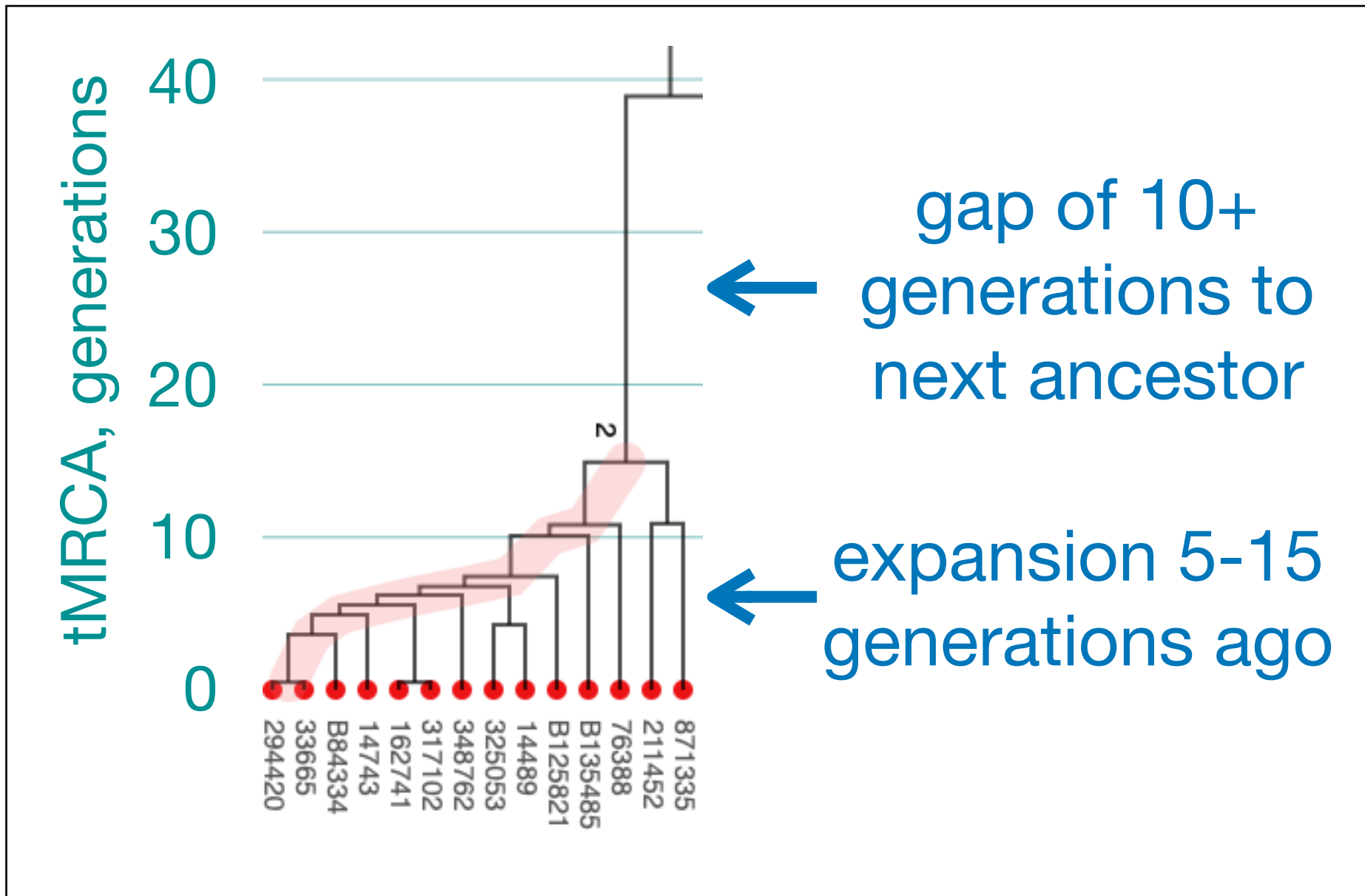
Spot the American Immigrant Families



Spot the American Immigrant Families



Spot the American Immigrant Families

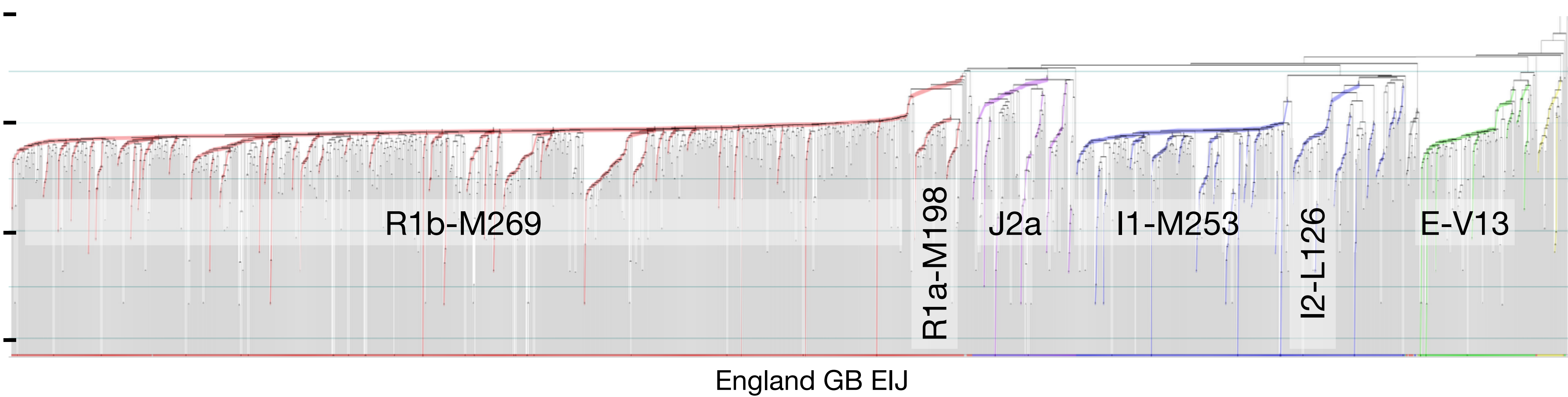


Spot the Bronze Age Expansion



generations before present

1000 -
100 -
10 -
1 -



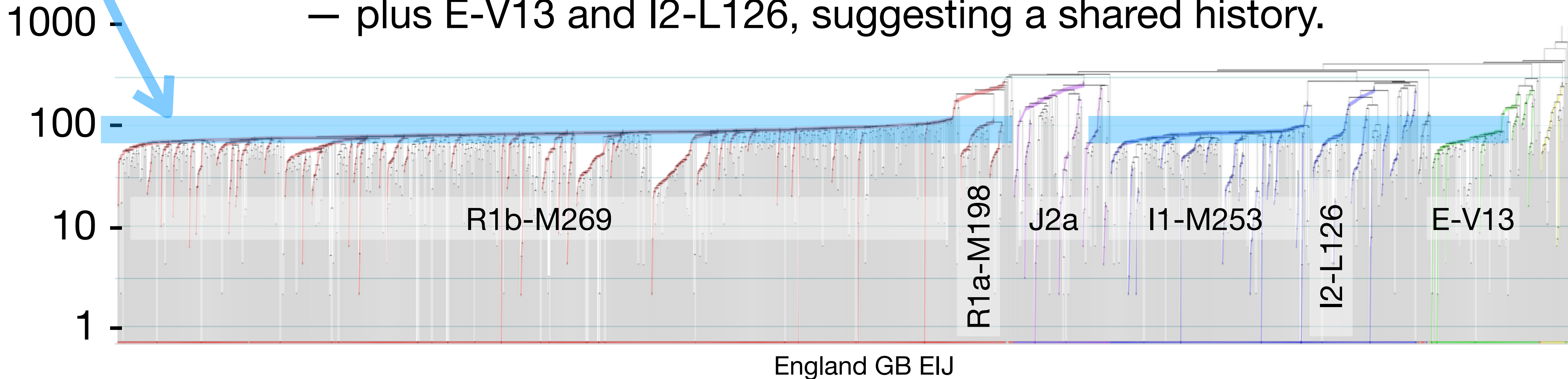
Spot the Bronze Age Expansion

“[Most European male] populations share similar histories featuring a demographic expansion starting ~2.1–4.2 KYA.”

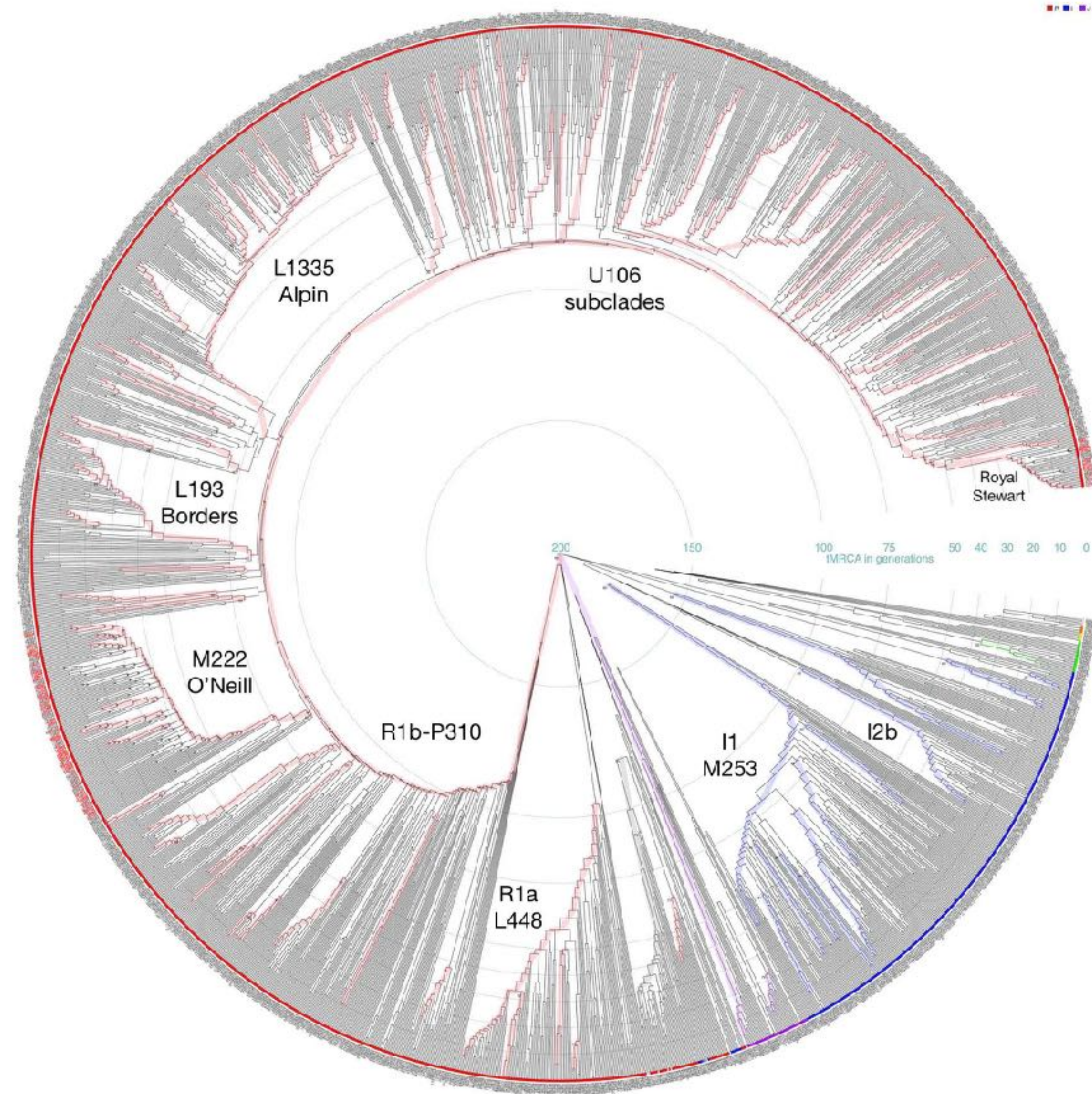
Batini et al, 2015. Nature comm 6, 7152

With one project's Y STRs, we can see what the Jobling group saw — plus E-V13 and I2-L126, suggesting a shared history.

generations before present

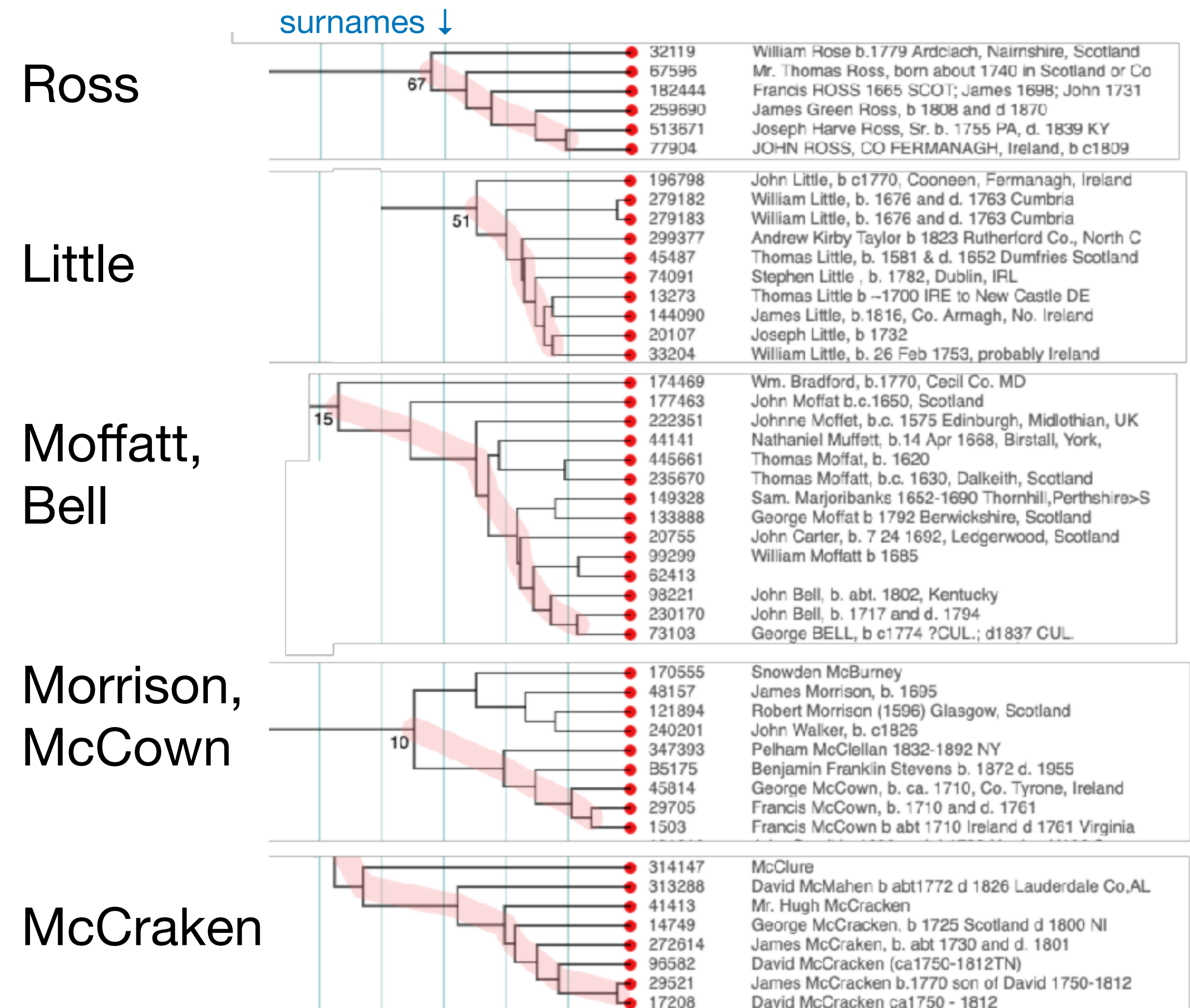
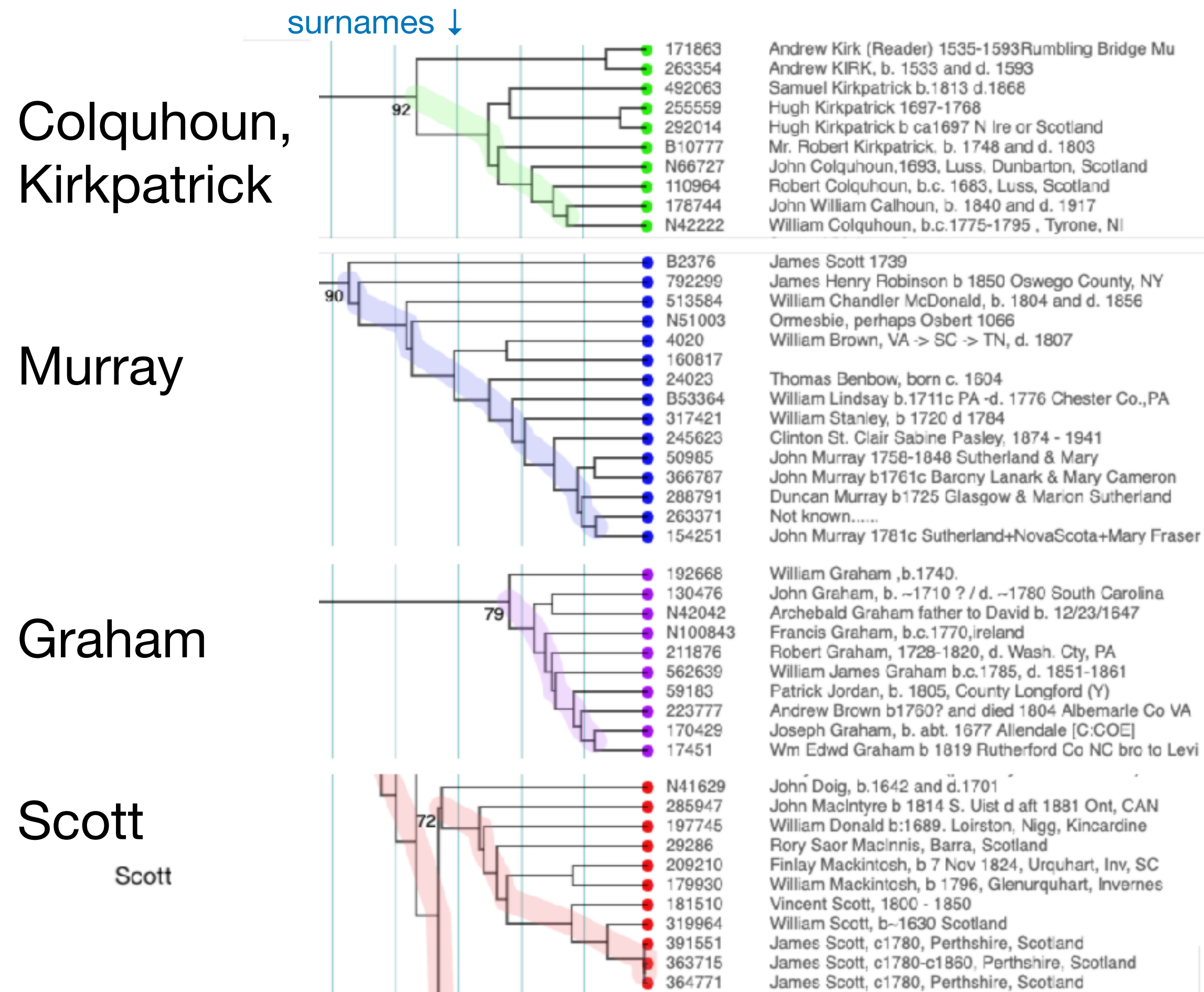


Spot the Clans and Septs

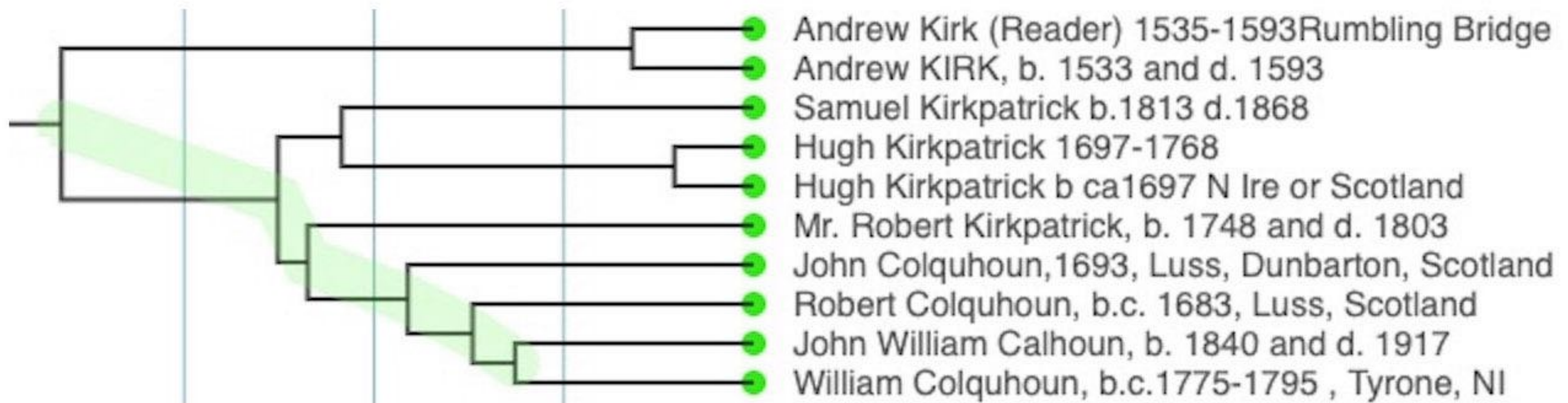


Spot the Clans and Septs

Most clan expansions occurred 20-30 generations ago, distinct from American immigrations 10-15 generations ago.



Spot the Clans and Septs



“Convergence”

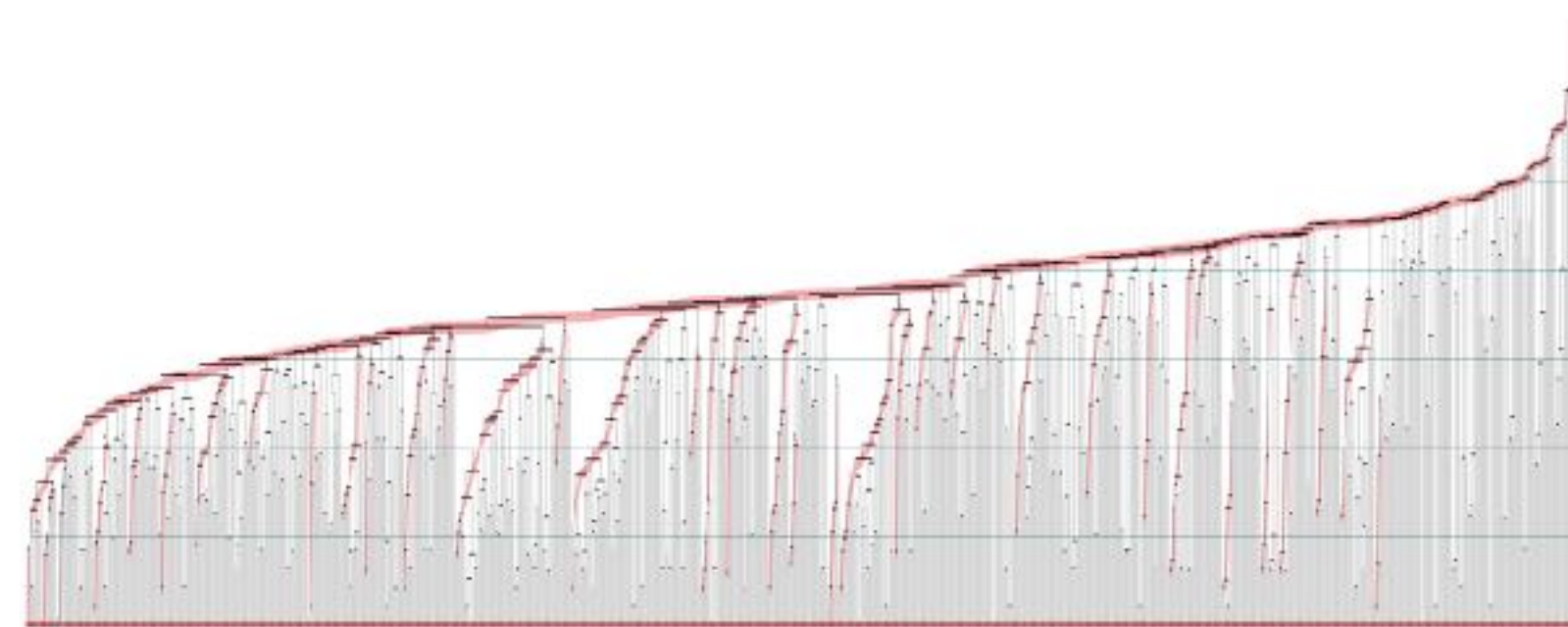
Come to the
workshop!

Getting Quantitative

We can fit experimental tMRCAs to a distribution, separating the statistical noise of the STR process from experimental error (mutation model and rates).

If we use the same model and rates for all analyses, we can compare different datasets with accuracy.

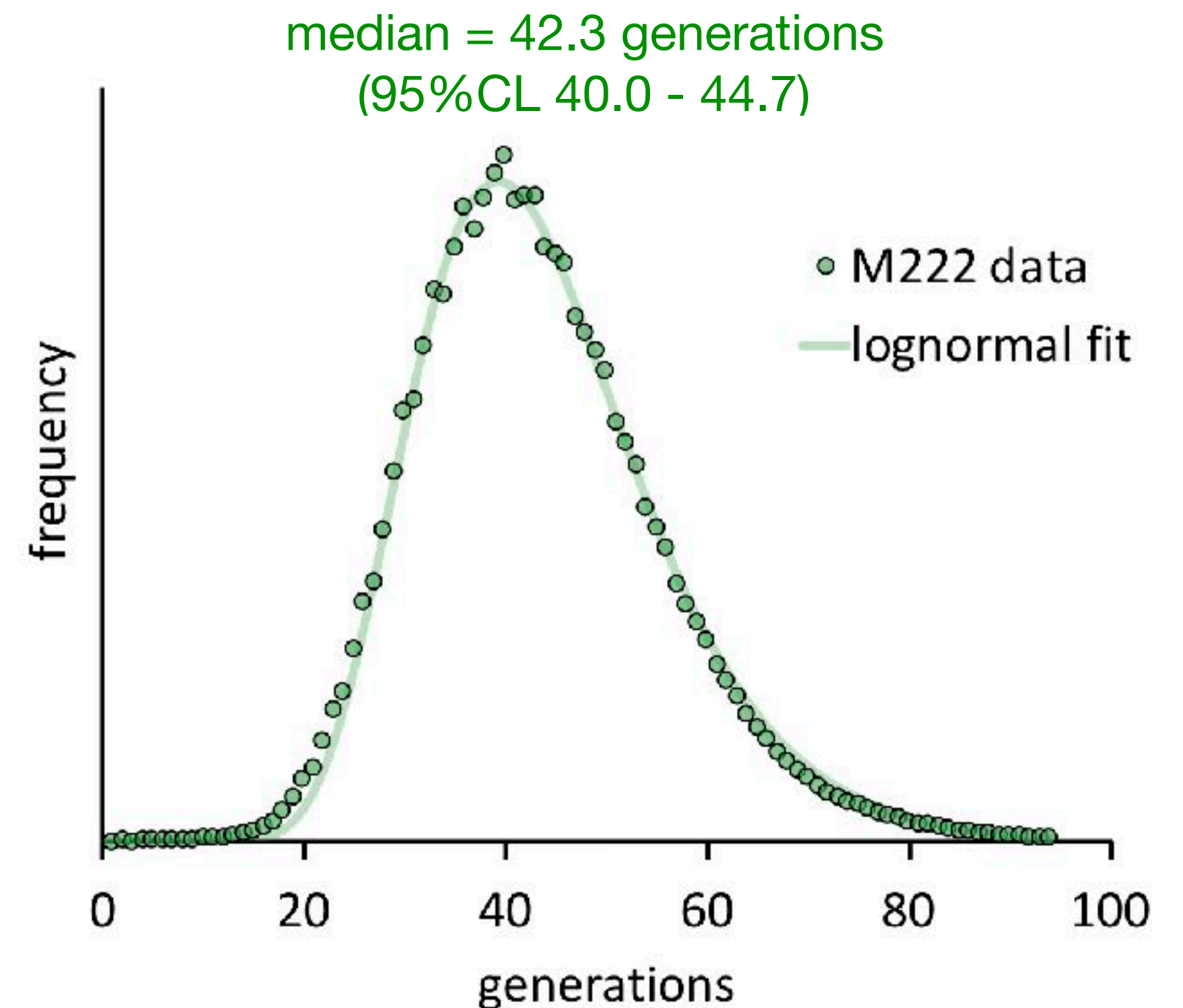
NB: These “STR clade tMRCAs” are dates of **expansion**, not founding.
The nearest corresponding founder SNP may be centuries earlier.



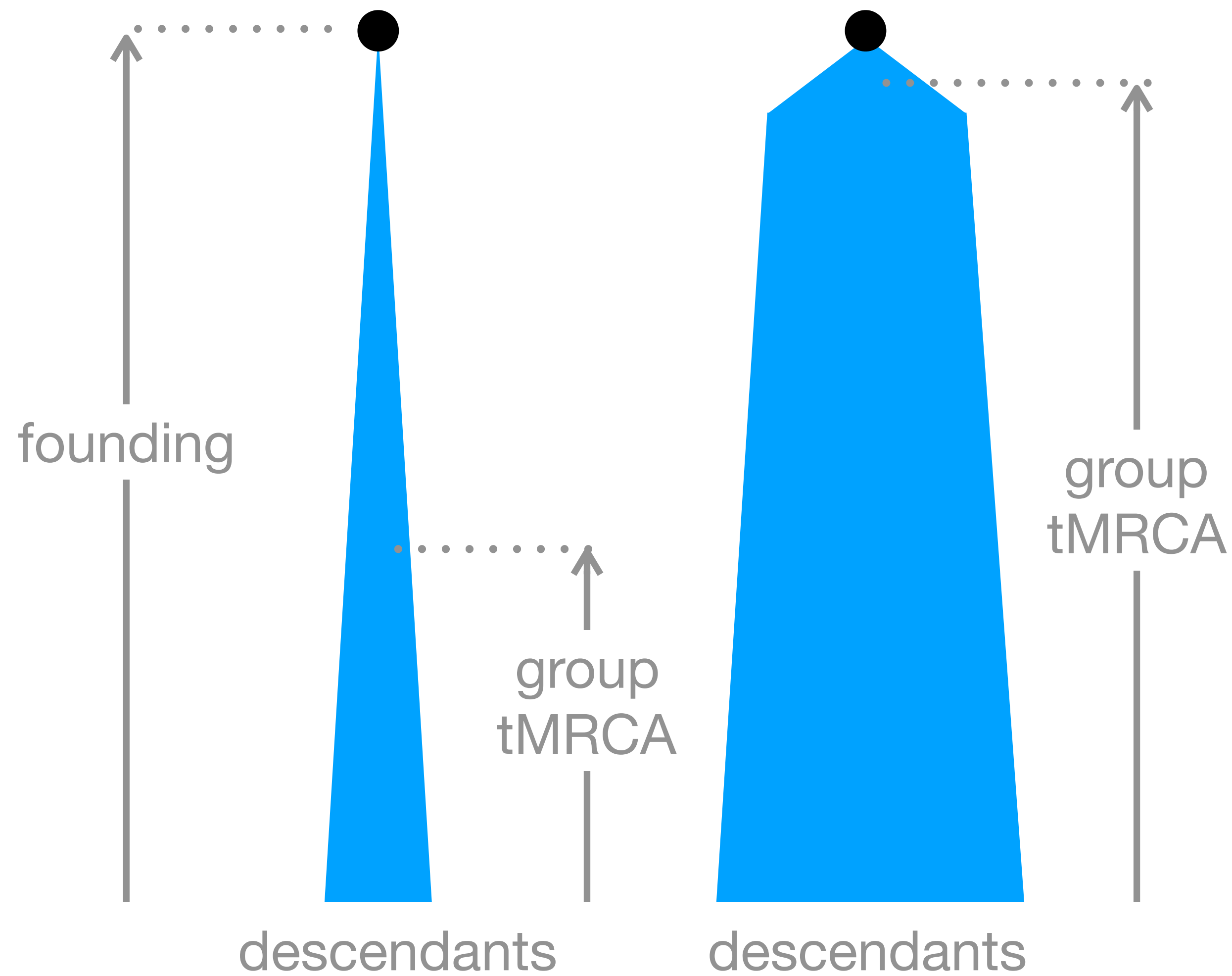
DNA testers

tMRCAs

math & theory



Founding vs Expansion

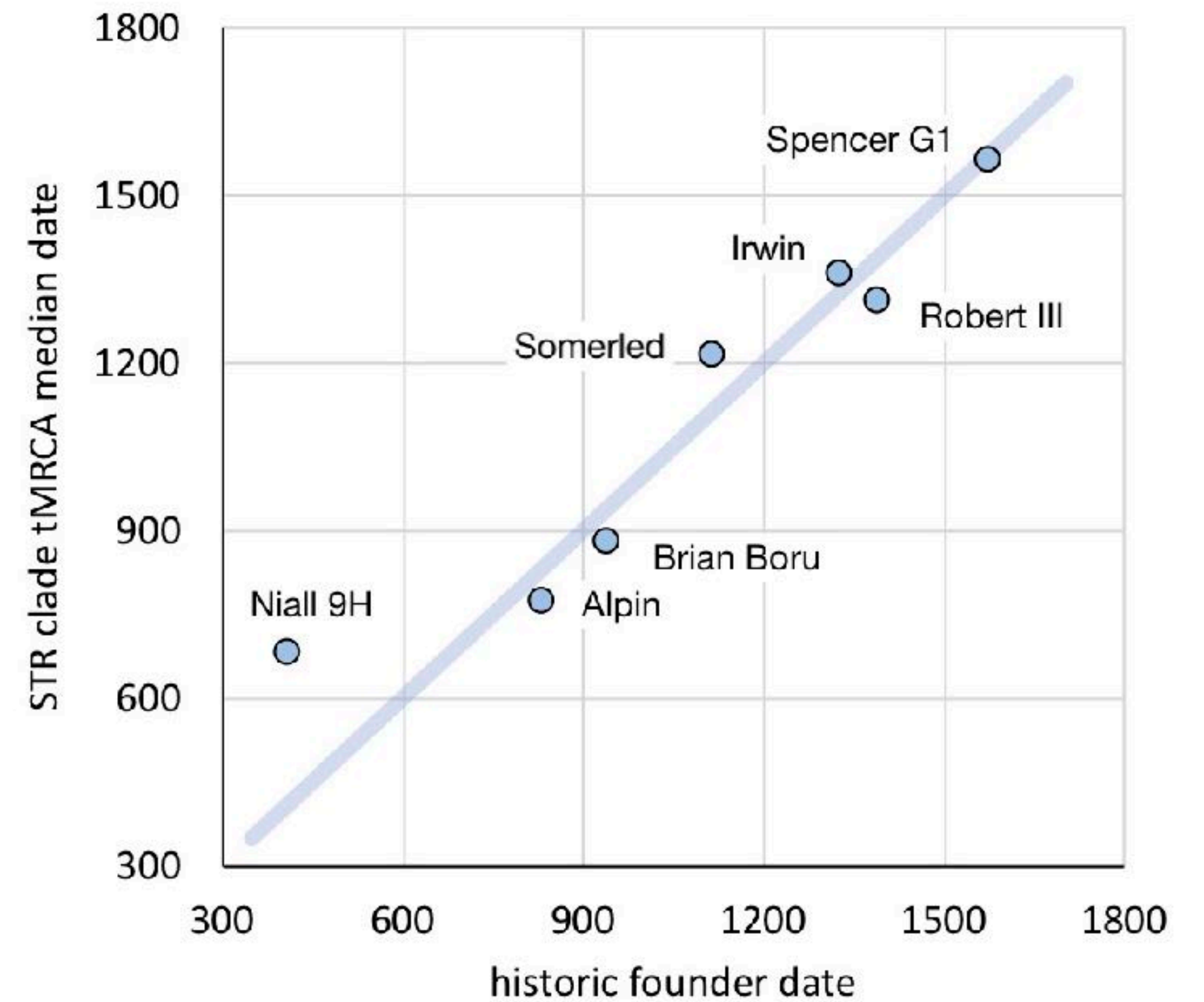
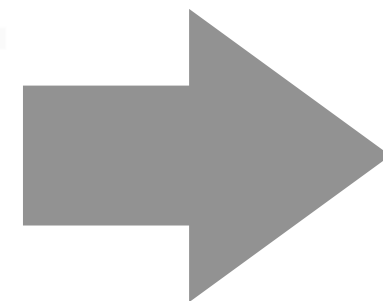
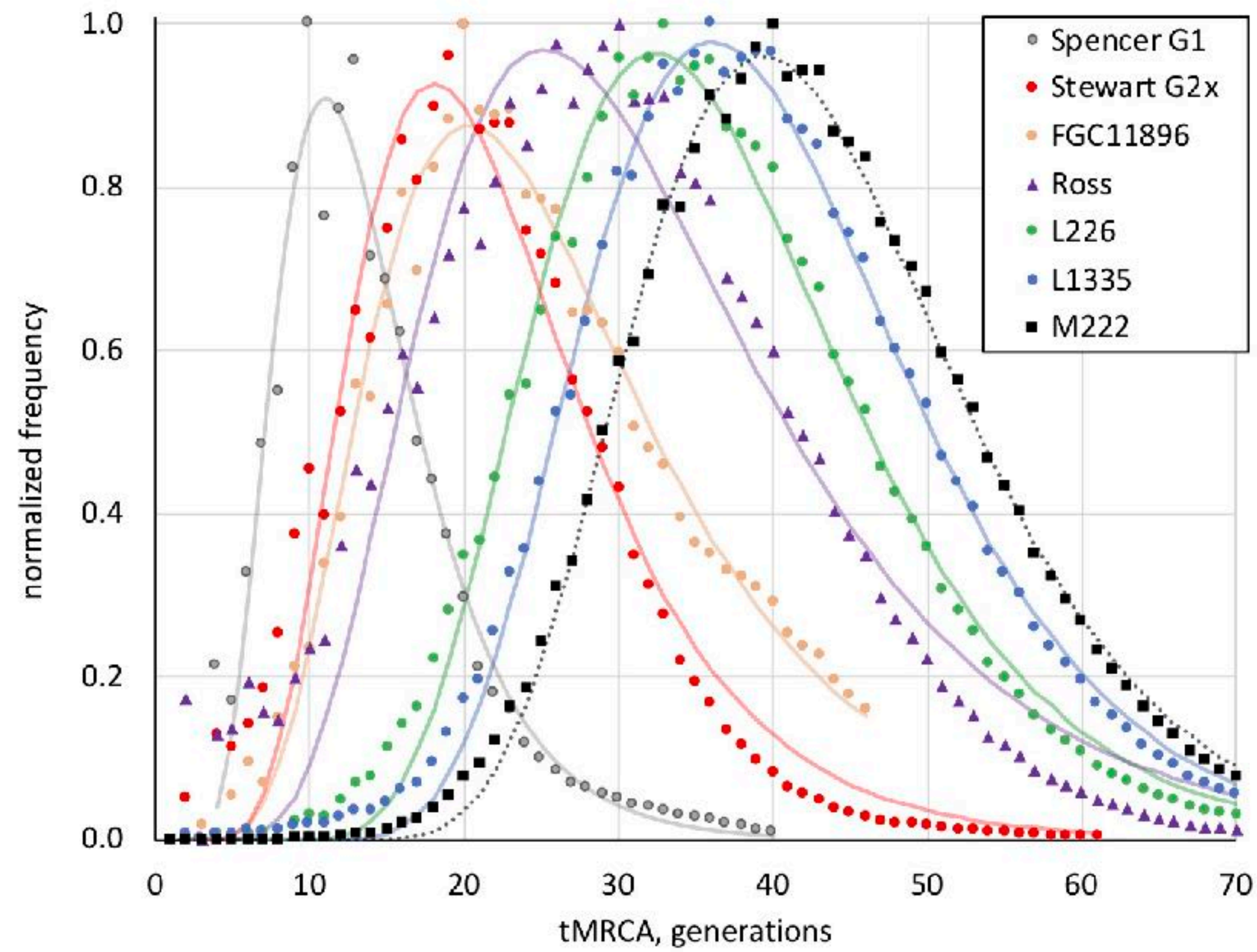


At low fertility rates, the difference between founding and STR group tMRCA can be hundreds of years.

In the “fertile patriarch” scenario, the difference can be less than a generation.

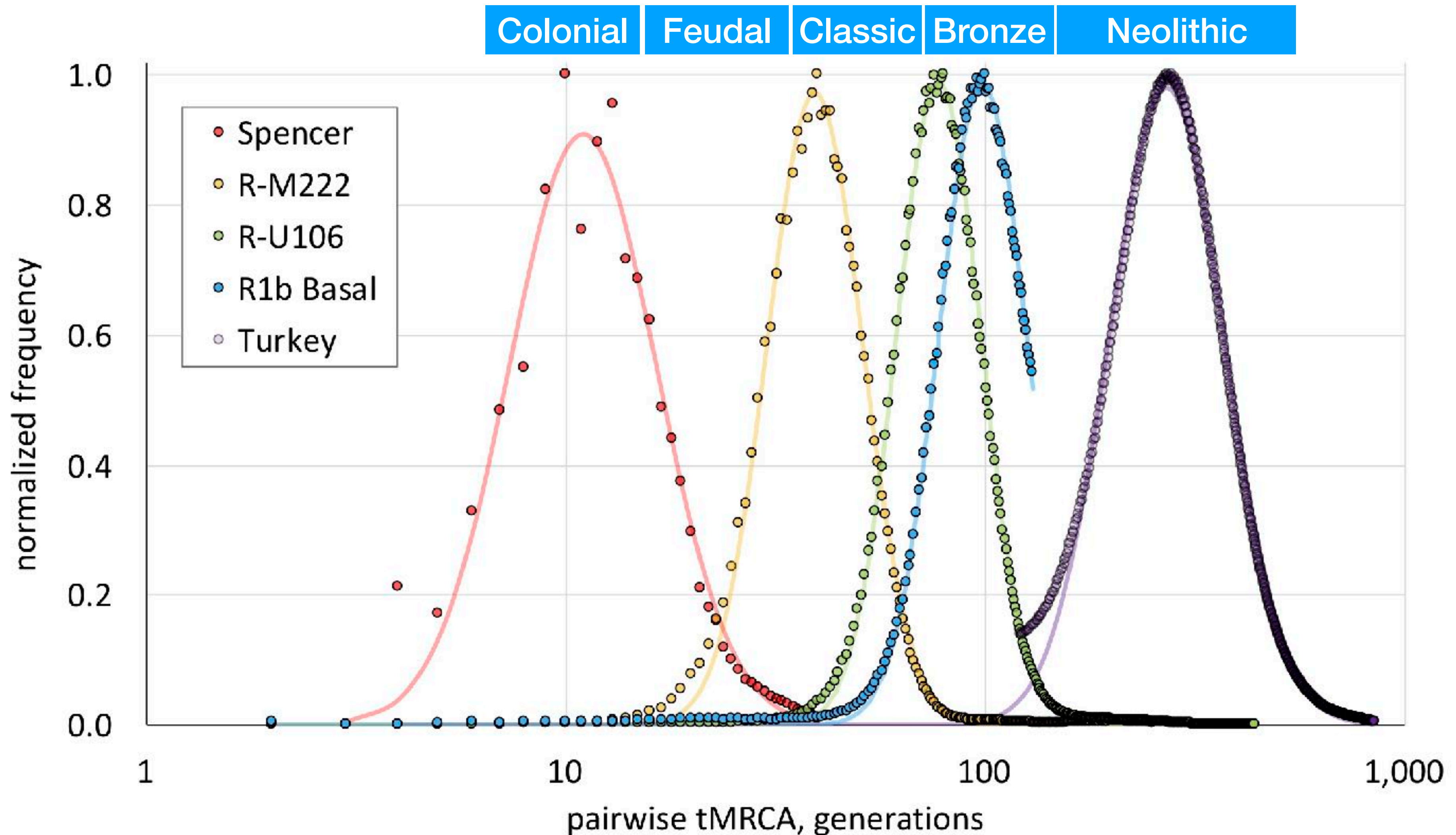
Dating Clans, Finding Founders

Known founders and dates (Spencer, Irwin, Boru) let us interpolate for other clades. We can then try to identify possible historic figures where lines and dates intersect.

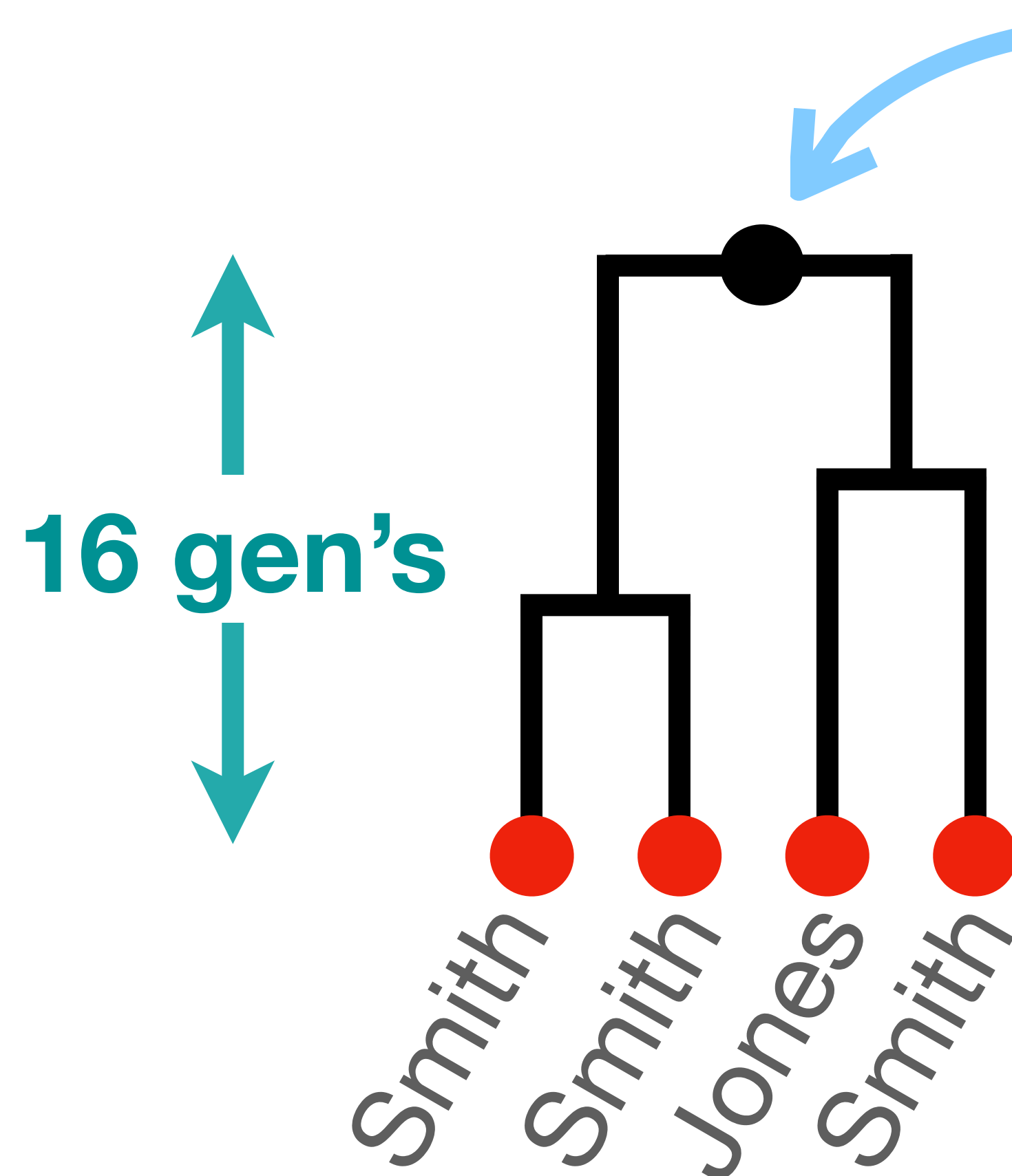


Enormous Range

The founder-expansion lognormal fit method has enormous range.
Y111 STR dates reach back 1000 generations to the Paleolithic Ice Ages



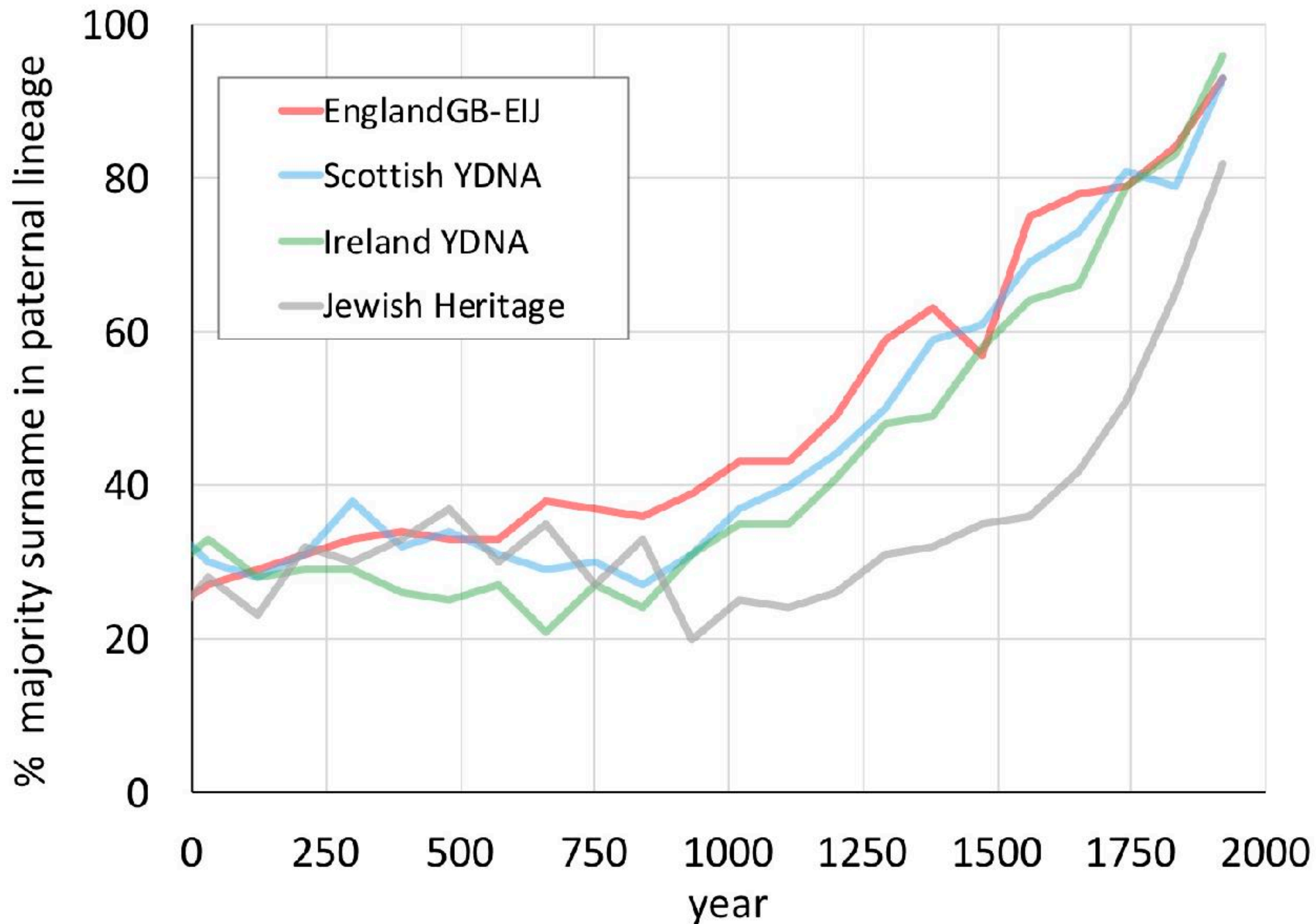
The Onset of Surnames



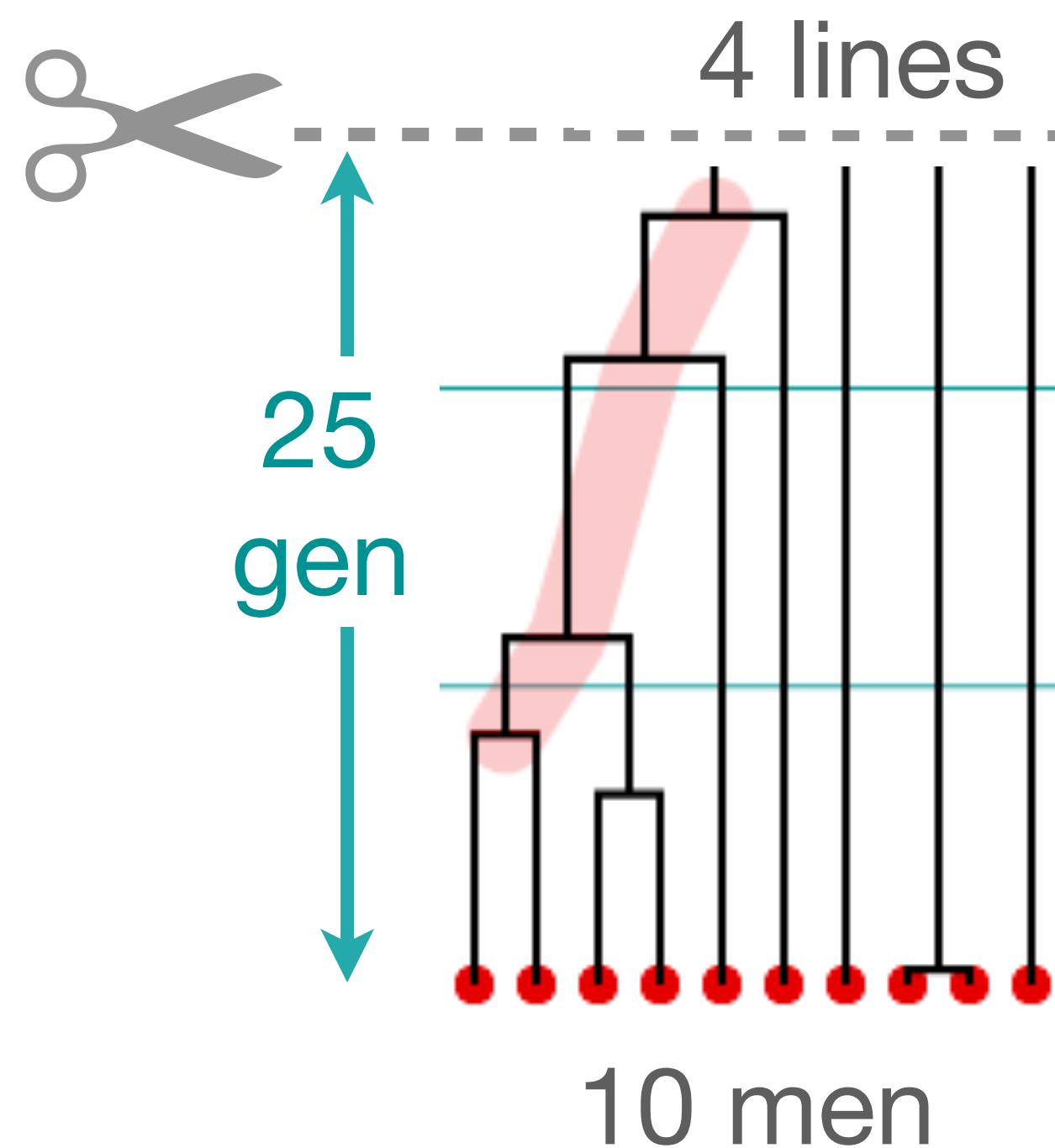
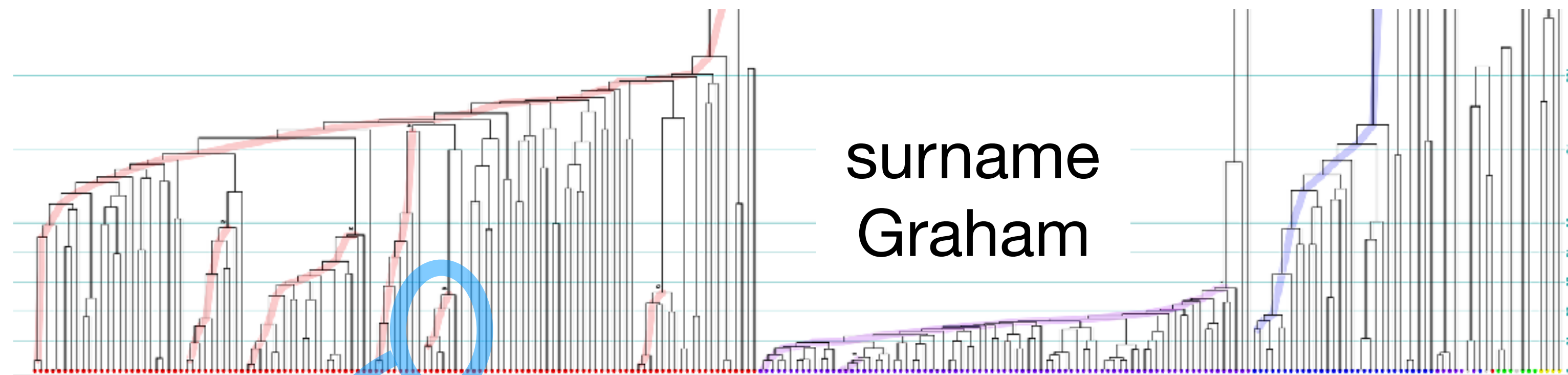
At this point, the time is 16 generations ago and the majority surname applies to 75% of the descendants.

Now repeat 10,000 times...

The Onset of Surnames

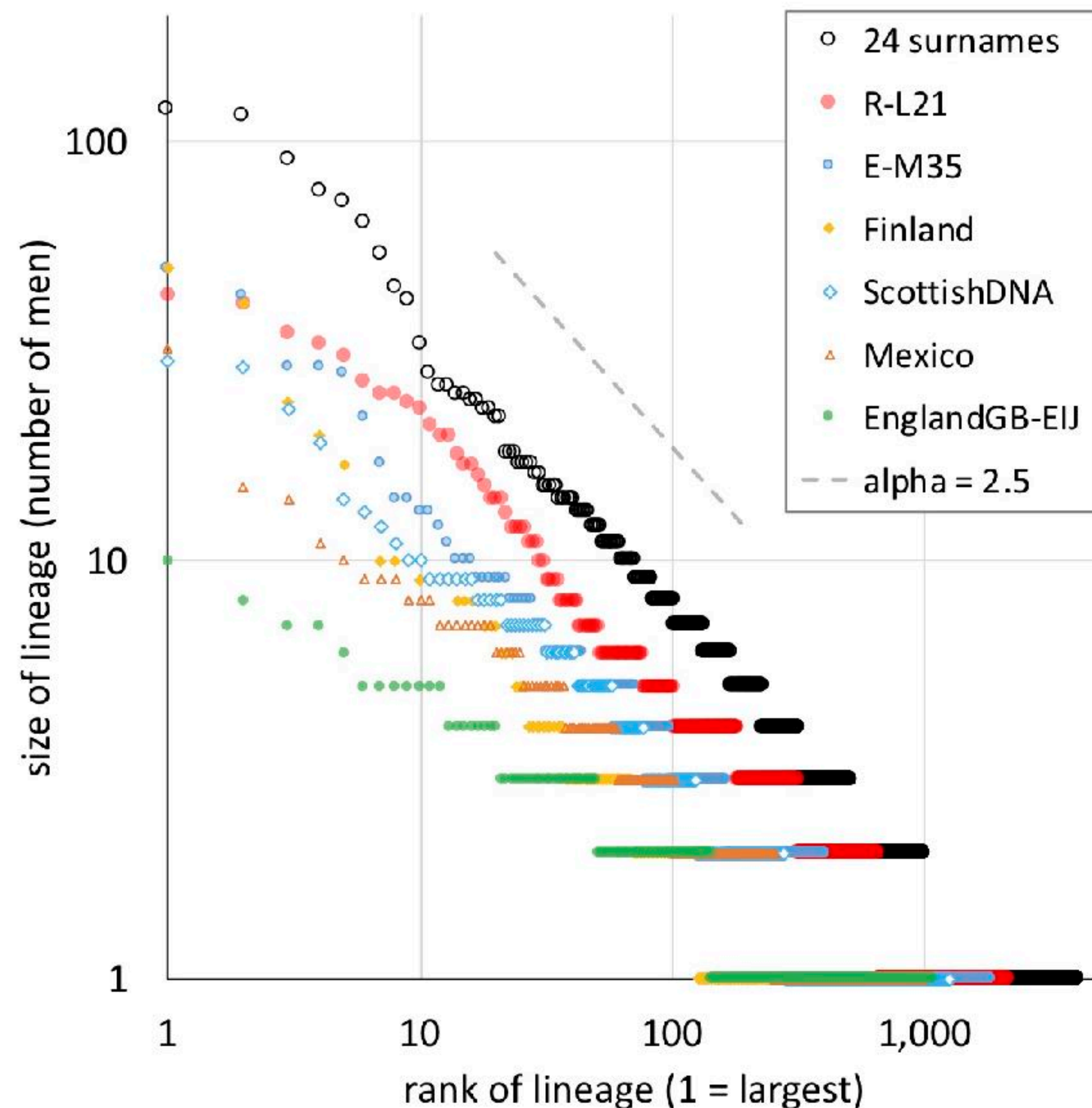


How many Y lineages are there?



1. Select kits
2. Slice at 25 generations before present
3. Count lineages and descendants
4. Repeat 20,000x

Y Lineages Obey Power Laws



Every collection of Y STRs has a few large lineages and many small ones.

This result does not depend on the level of Y testing (Y111..) or type of project.

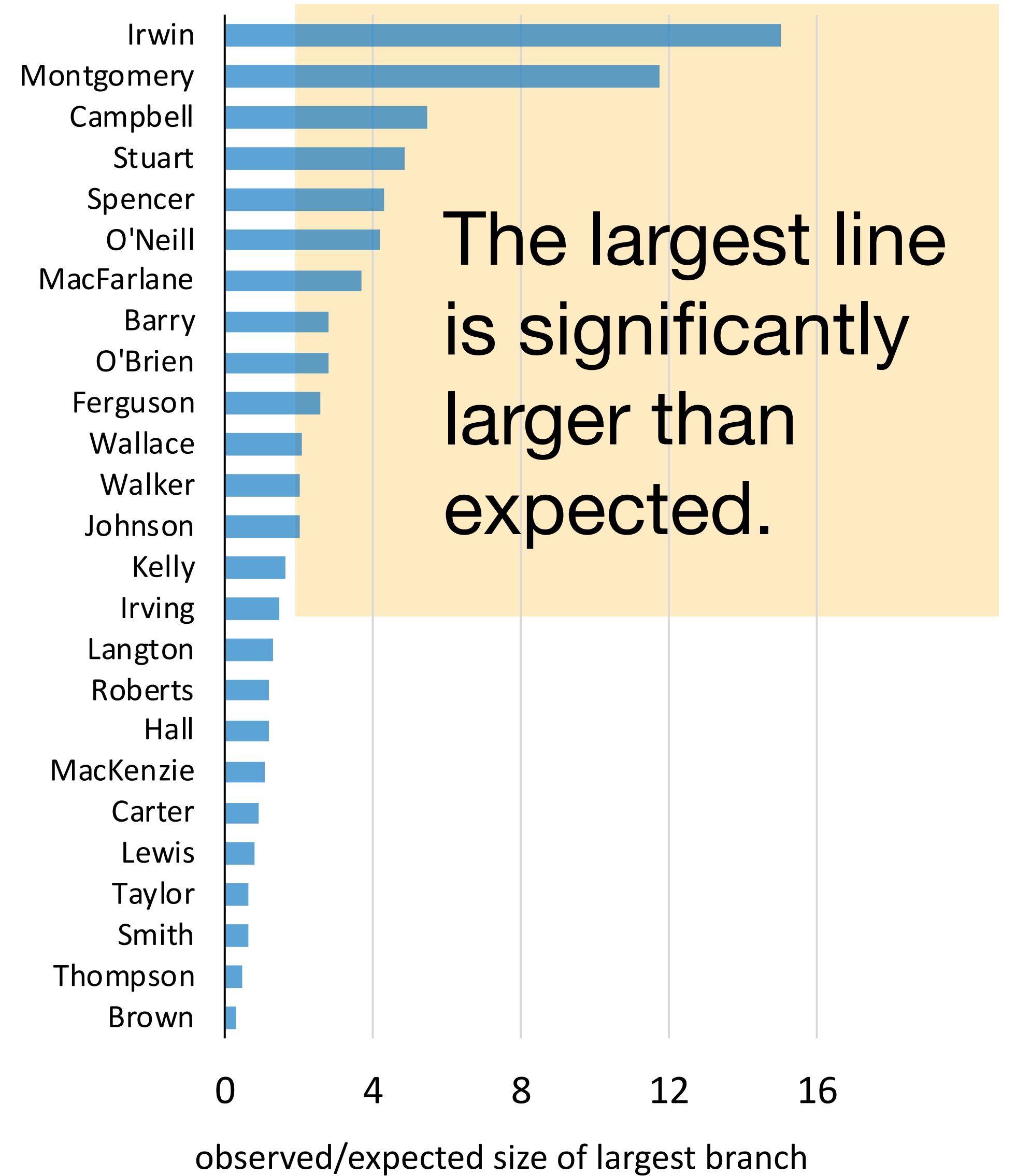
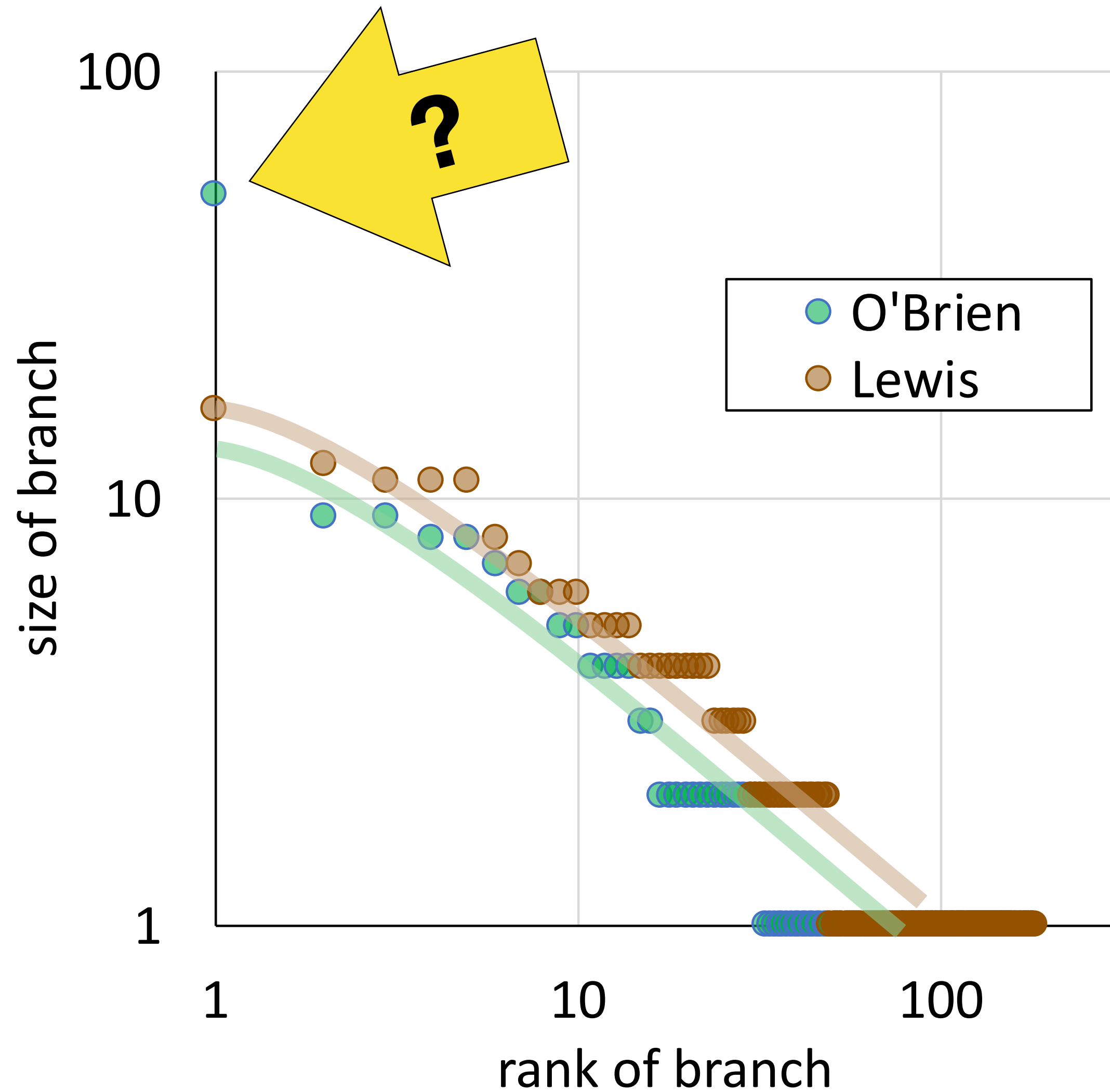
Biased selection projects will be exceptions (e.g. Royal Stewarts).

Rough rules of thumb:

For any set of N men, you will find $0.6 \cdot N$ lineages (defined as within surname/Y match distance).

A third of men will have no matches.


Which Surnames Have Unusual History?



Conclusions

- You may think that your projects are special — but with a wide-angle lens, the unusual becomes commonplace.
- Founder effects are pervasive and hierarchic with consequences far downstream. Observations like convergence or sparse Y hits are direct consequences of robust large-scale patterns.
- Y STRs measure expansions after bottlenecks with high precision and can be calibrated against historic events. They have data density for recent events that SNPs cannot approach.
- Step back from the trees and look at the forest!
- Documentation and source code at <http://scaledinnovation.com/gg/gg.html>



A photograph of a roller coaster with red tracks and white supports, set against a sunset sky with scattered clouds. The coaster features several loops and drops. The text is overlaid on the right side of the image.

Hope you enjoyed the ride!

Rob Spencer
co-admin at England GB Groups EIJ
spencerrw@alum.mit.edu
<http://scaledinnovation.com/gg/gg.html>



Abstract

The Big Picture of Y STR Patterns

Most of us use Y STR data locally to explore personal matches and to help in building family trees. But STRs can tell us much more when we sit back and take a long look. In this talk we use an efficient way to visualize thousands of kits at once. The large-scale patterns explain “convergence”, illuminate ancient, feudal, and colonial expansions, pick apart Scottish clans, identify American immigrant families, allow accurate relative clade dating, let us see the onset of surnames, and reveal the power law distribution of lineages.