Inference of demographic histories of natural populations using sequence data: Coalescence, Mutation and Recombination

Sequencing experimental design: Population sampling, and reference genomes

Richard Durbin rd109@cam.ac.uk

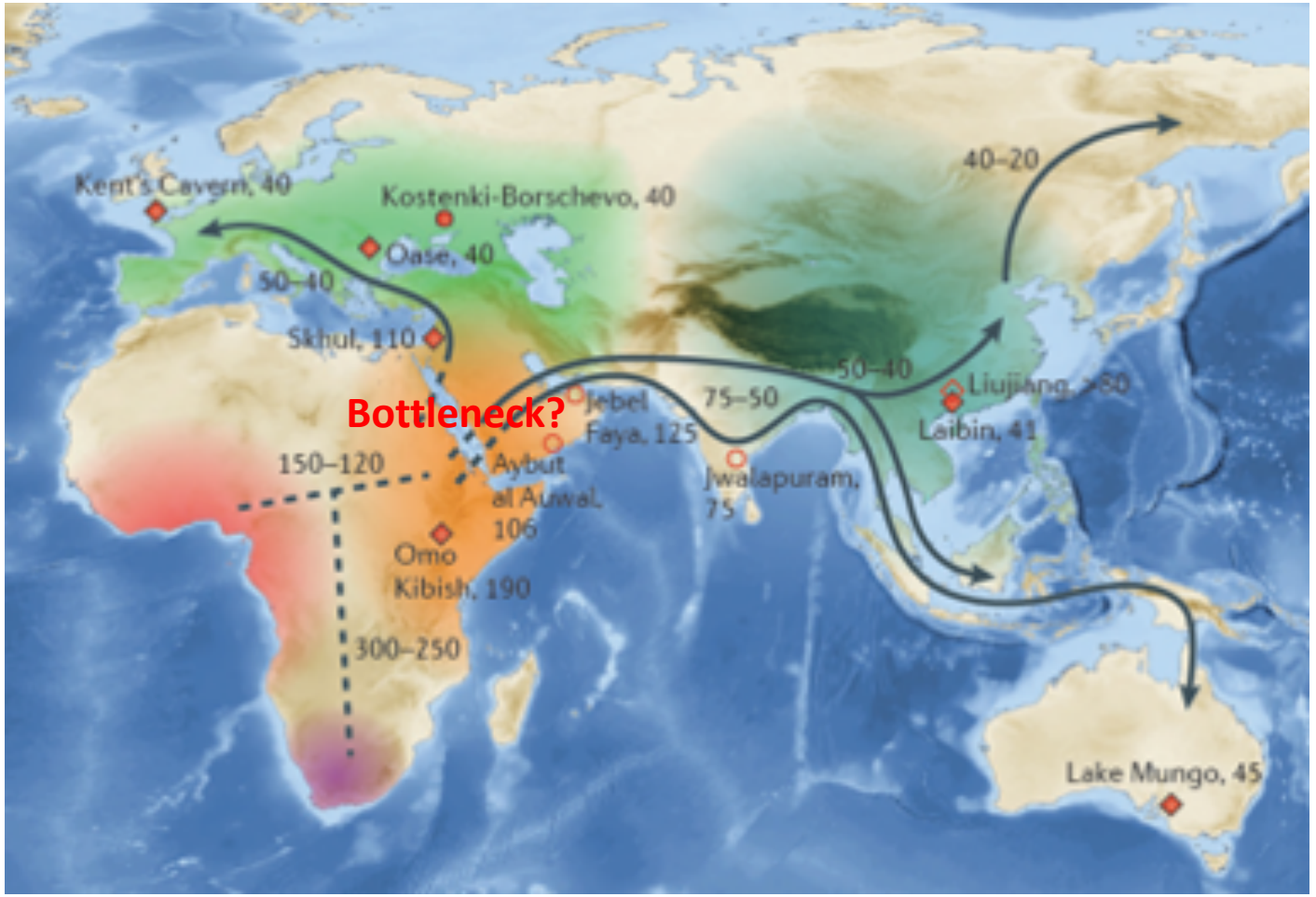
Cesky Krumlov 24/1/20

What I mean by demography

1. Population size going back in time
   - Actually "effective population size" N_e(t)
     - We will come back to what this means
   - Approximate time range 10k – 1M years ago
     - Again we will see why
2. Population structure
   - Subpopulations and when they split (and merged?)

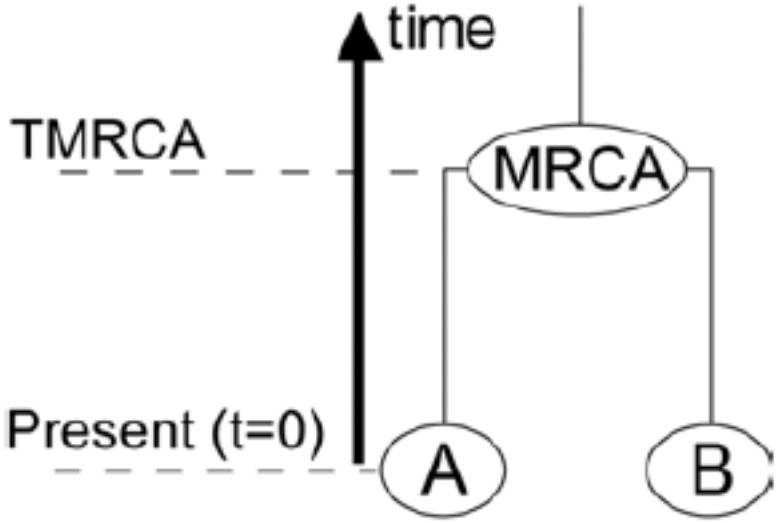
- Based on explicit evolutionary models
  - Relate patterns of (shared) genetic variation accumulated since a common ancestor to history

Example: human history



Tree on two sequences

- Gustave Malécot (1940s)



- *Coalescence* is joining together, in our case going backwards in time
- Chance of coalescence per generation is 1/N
- TMRCA is exponentially distributed with mean N

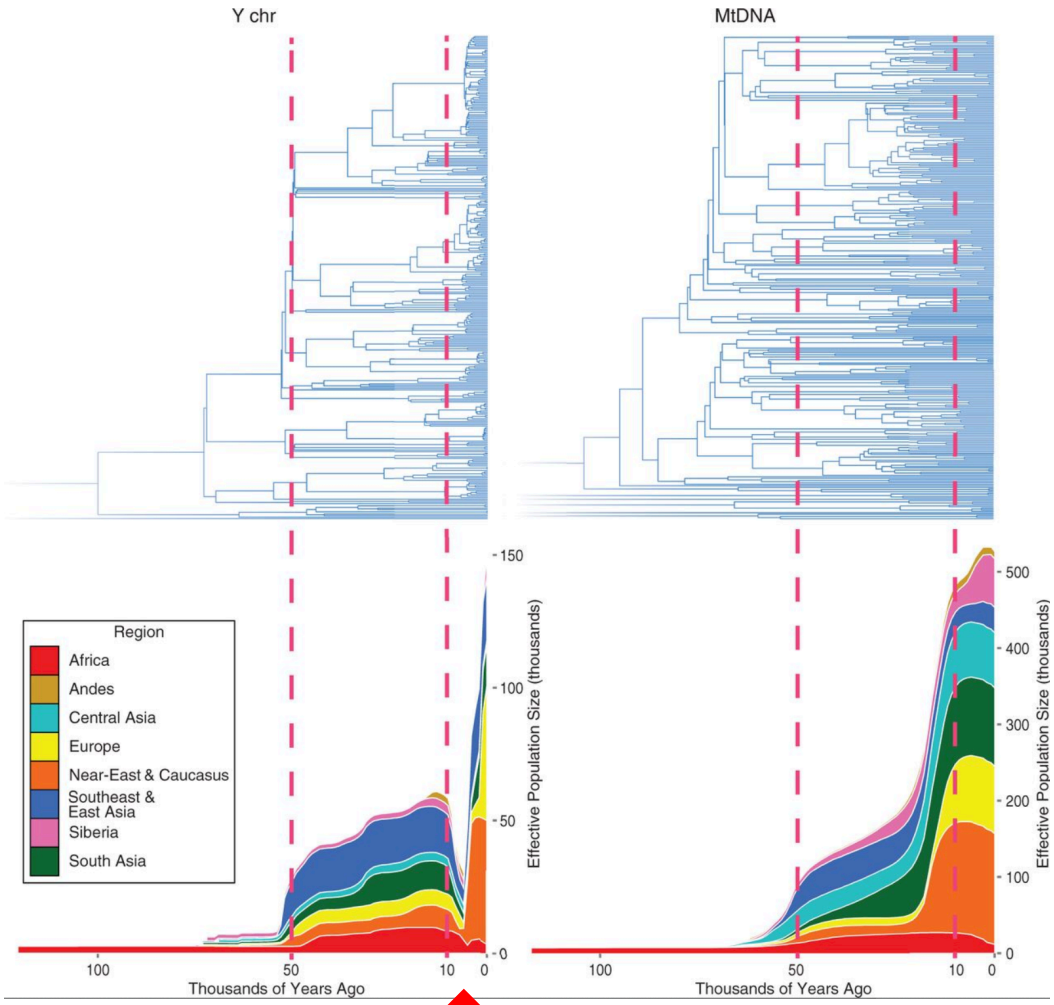
Probability of observing a mutation

- To see a mutation, it must have happened on one of the branches since the common ancestor
- P(observed mutation) = $2T\mu$
- E(observed difference rate) = $\theta_\pi = 2N\mu$
- Humans are diploid, so $\theta = 4N_e\mu$, where N_e is the *effective population size*
- For humans, θ_π =~ 0.001
  - 1/800 – 1/1200 depending on population
- Hard to measure N_e and μ independently…

Effective population size

- Lots of mystique/angst about this
  - Our definition is arguably at the core of the concept
    - the reciprocal of the probability of sharing a parent in the previous generation
    - = 1 / coalescence rate
- Why this is different from census population size:
  - Long term averaging: many consequences occur over large numbers of generations (often order of N_e)
  - *Population structure* generates non-random patterns of coalescence, and non-independence between generations
  - Maybe only a small percentage of individuals breed
  - Selection favours some individuals over others
- But it is always something of this form that we get at by population genetic analysis

For non-recombining sequence we can estimate $N_e(t)$ from a dated tree

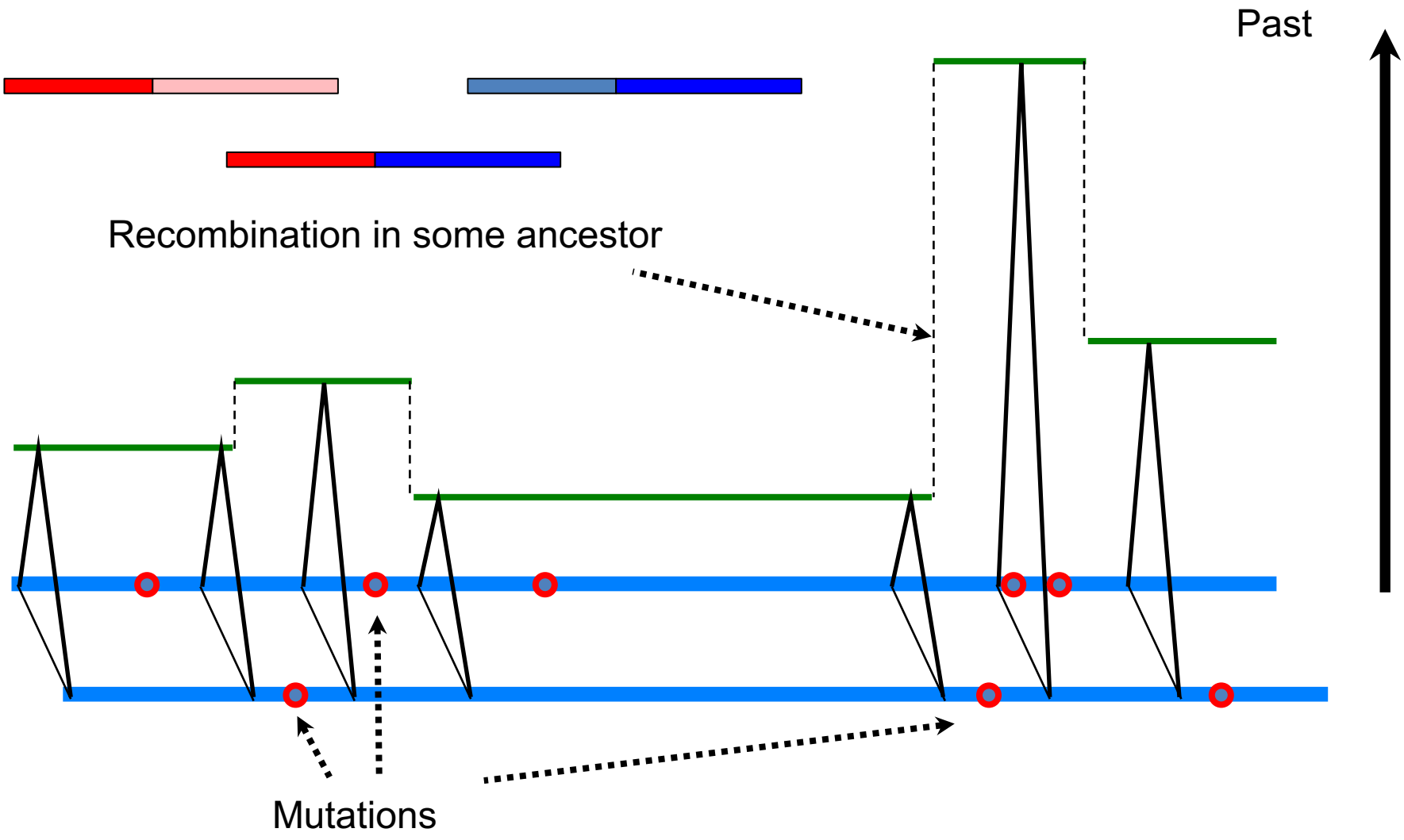


N_e estimated by "Bayesian skyline" method, essentially looking at 1/coalescence rate, and smoothing through the discrete events.

Note that there is a big dip in male but not female N_e around 8kya outside Africa, 5kya in Africa. This is the time of onset of agriculture. Dominant males?

Karmin, M. *et al.*. *Genome Res.* **25**, 459–466 (2015).

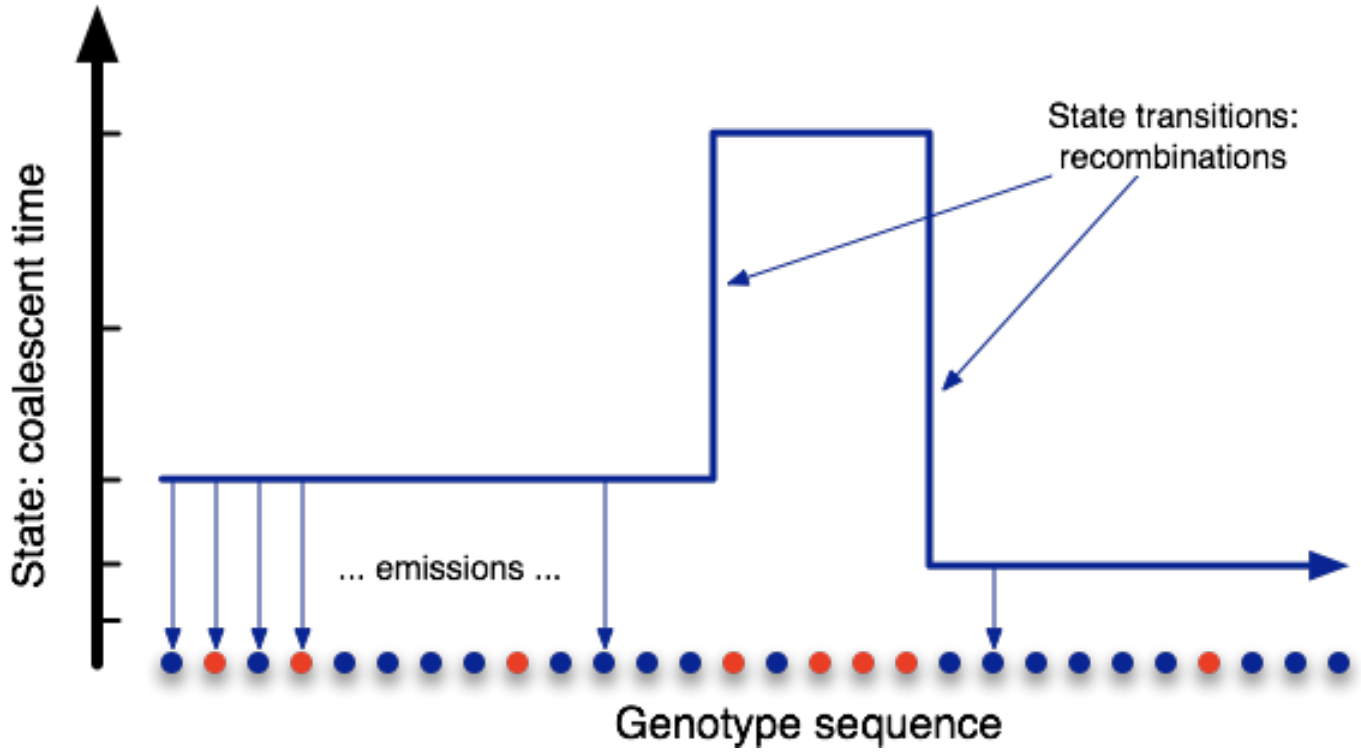
On autosomes segments of fixed trees are separated by ancestral recombination



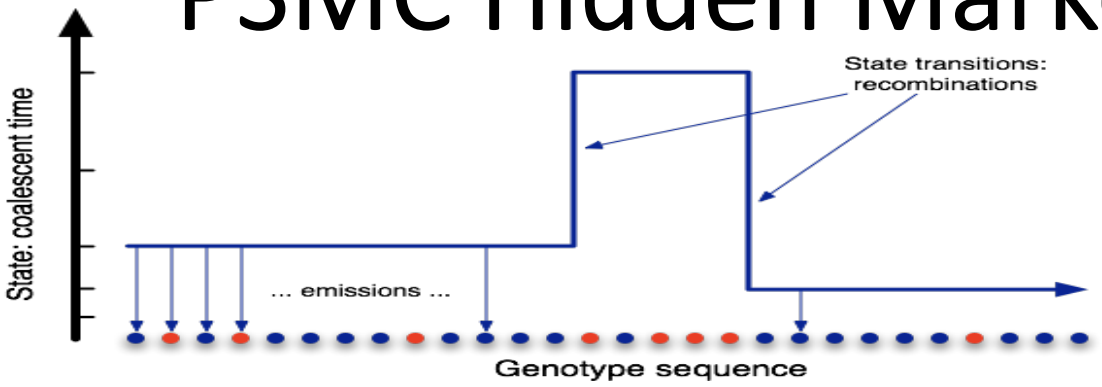
PSMC = Pairwise Sequentially Markovian Coalescent

Li and Durbin (2010): Inference of human population history from individual genome sequences

Hidden Markov Model



PSMC Hidden Markov Model



- Move from left to right in the genome
  - Let P($x|t$) = prob(data up to x|TMRCA at $x = t$)
  - Calculate P($x+1|t$) = ($\sum_s$ P($x|s$) $r(t|s)$) $e(x)$
- $e(x)$ = "emission at x" = $2\mu t$ if a het, else $(1\text{-}2\mu t)$
- $r(t|s)$ = prob(recombination from TMRCA s to t)

  = $2\rho s$ prob (coalesce back to t) ← Depends on $N(t')$ $t' < s,t$

  $+ (1\text{-}\ 2\rho s)$ if $t = s$

Markov assumption

- This model assumes that

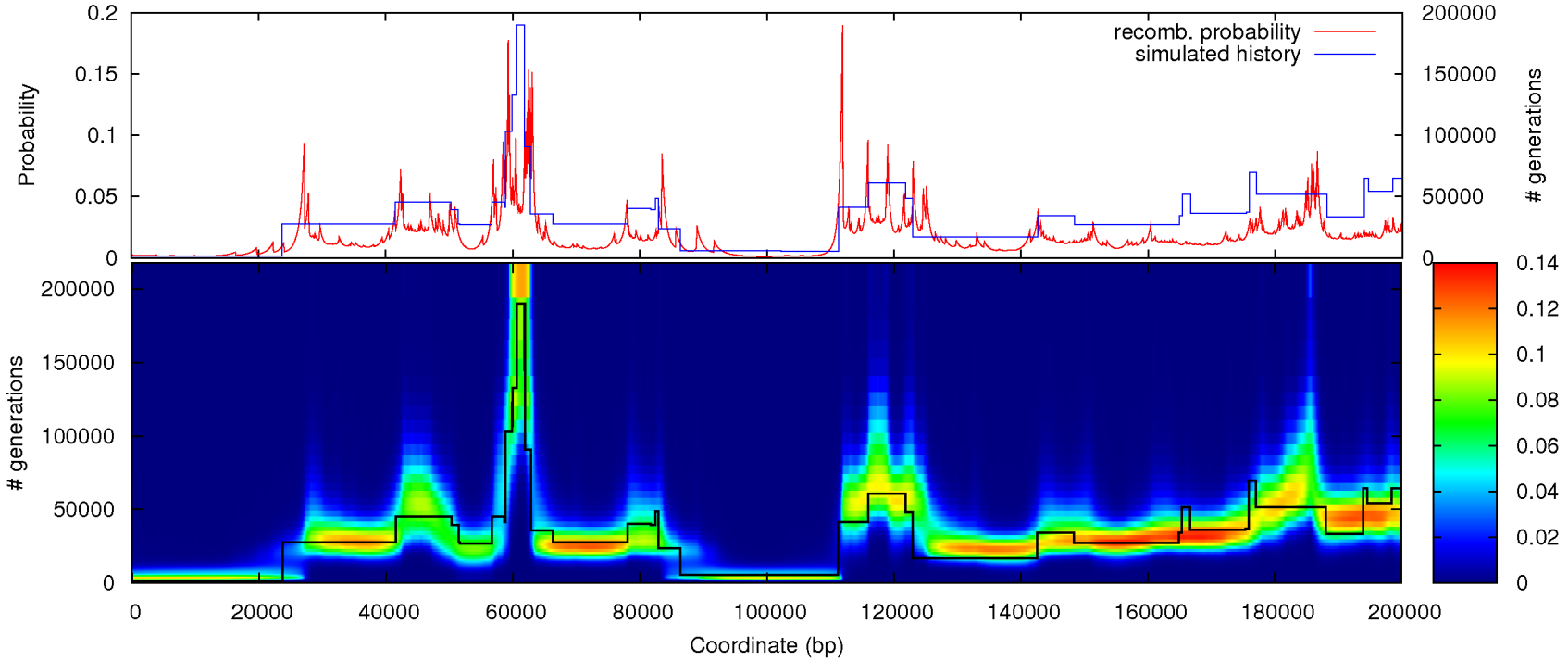
  data to the left of x|TMRCA at $x = t$

  is independent of

  data to the right of x|TMRCA at $x = t$

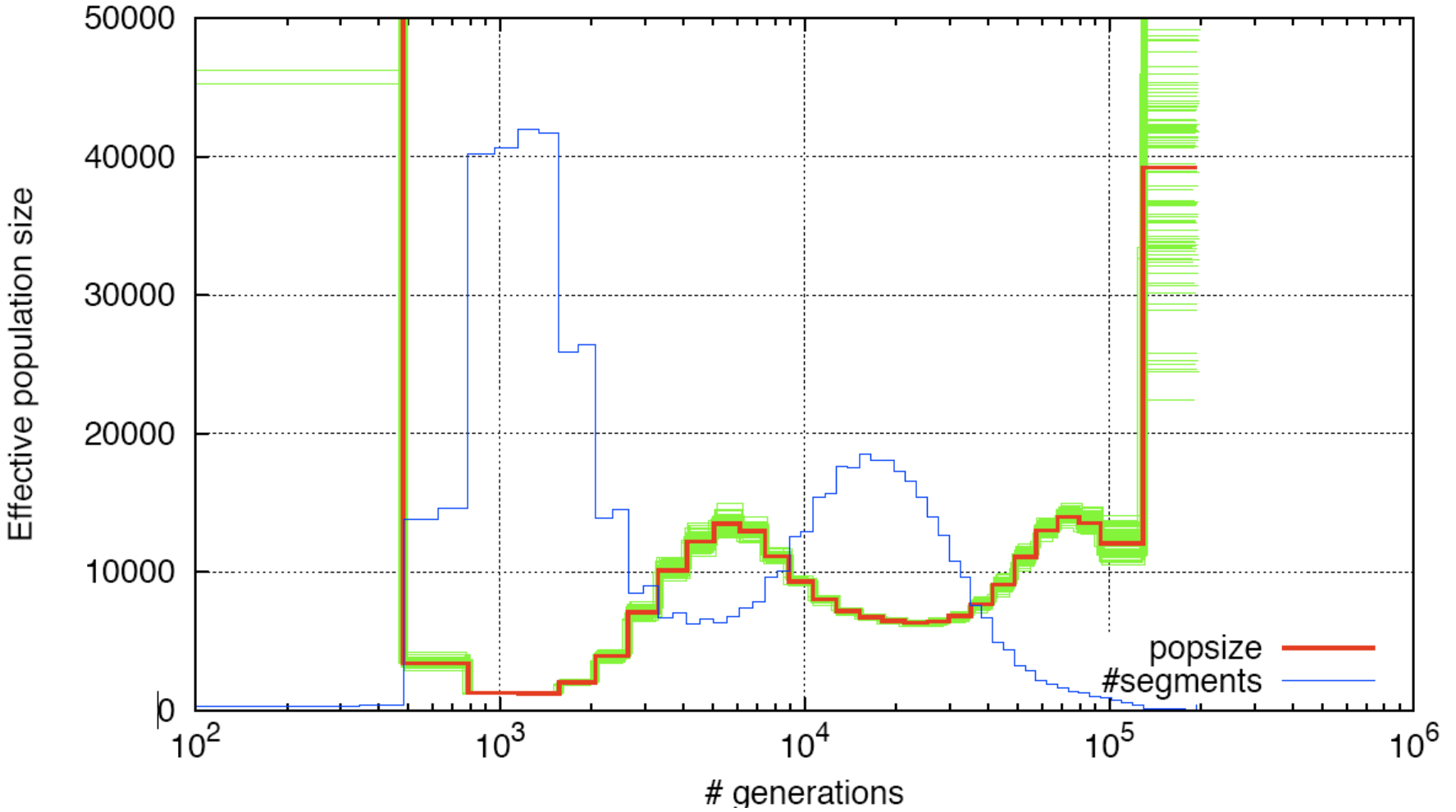
- For standard mixing populations this is a very good assumption
  - Sequentially Markovian Coalescent (SMC) approximation, McVean & Cardin 2005

PSMC-HMM reconstructs individual history

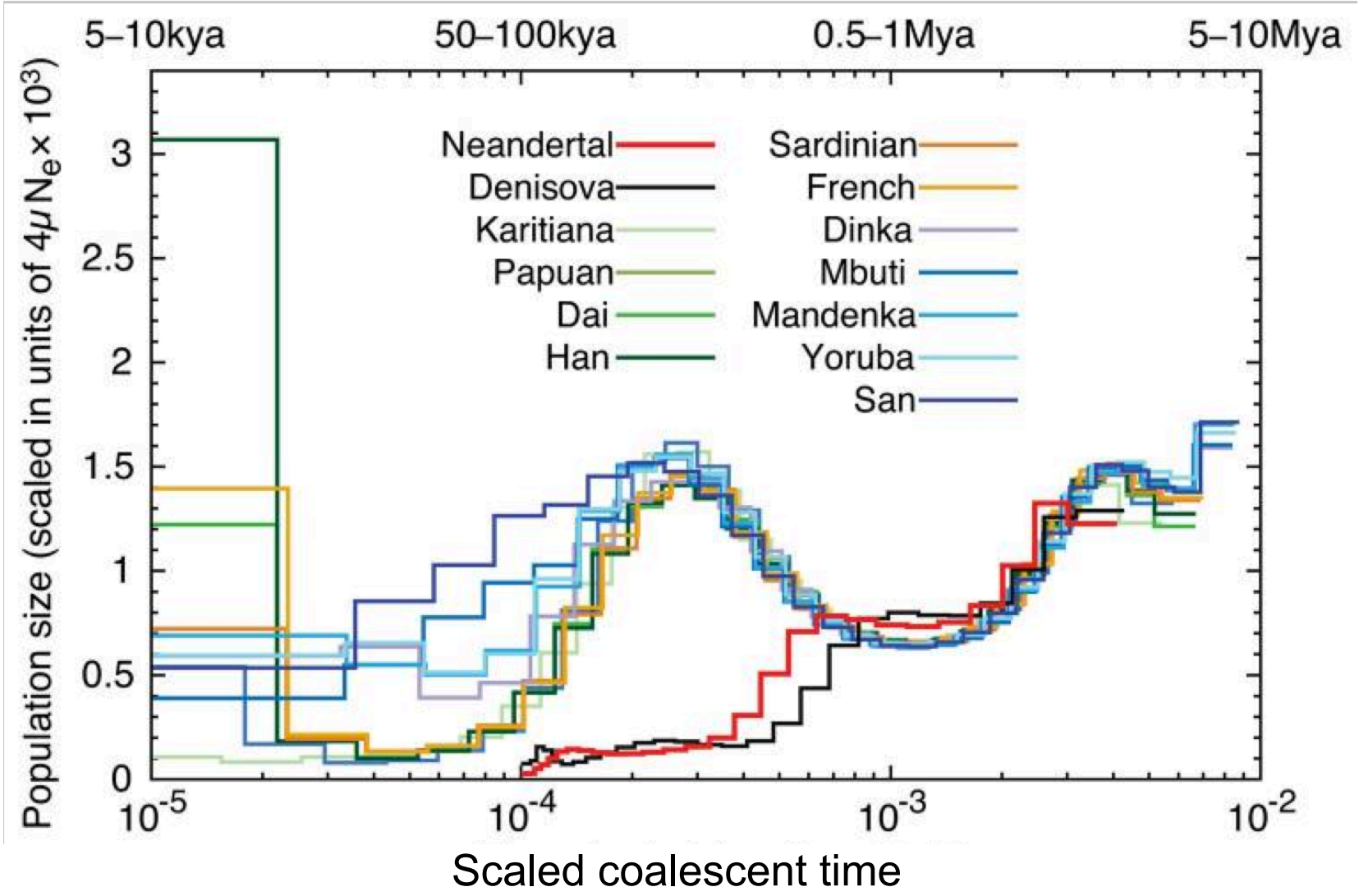


- Pairwise Sequentially Markovian Coalescent – Hidden Markov Model
- Data simulated using ms (Hudson)
- Model the coalescent time *t* by e.g. 50 discrete bins, spread logarithmically

Single human genome with bootstrap



Human population history, with Neanderthals



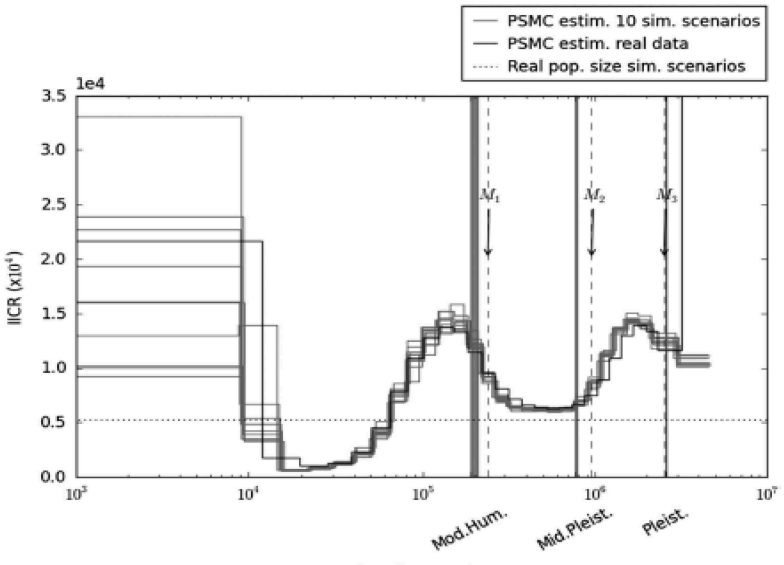
Heng Li

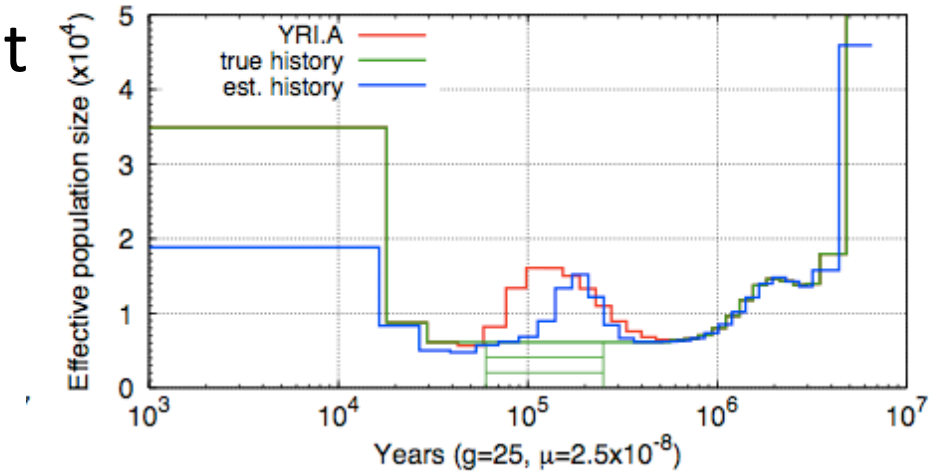
Advances since the original PSMC

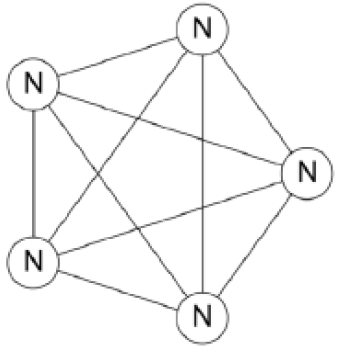
1. Use SMC' model which correctly handles recombinations coalescing back to the same ancestor (Stephan Schiffels, …)
   - Minor tweak to equations, but significant
   - Can now fit recombination:mutation ratio
   - Implemented in MSMC/MSMC2
2. Time speedup: linear not quadratic in number of time slices (Kelley Harris, ... Song, 2014)

Coalesecent $N_e(t)$ reflects ancestral structure as well as population size

- PSMC actually measures $\lambda = 1/\textit{coalescence rate}$
- Structure can also change coalescence rate
  - Li & Durbin supplement
  - Olivier Mazet…Chikhi

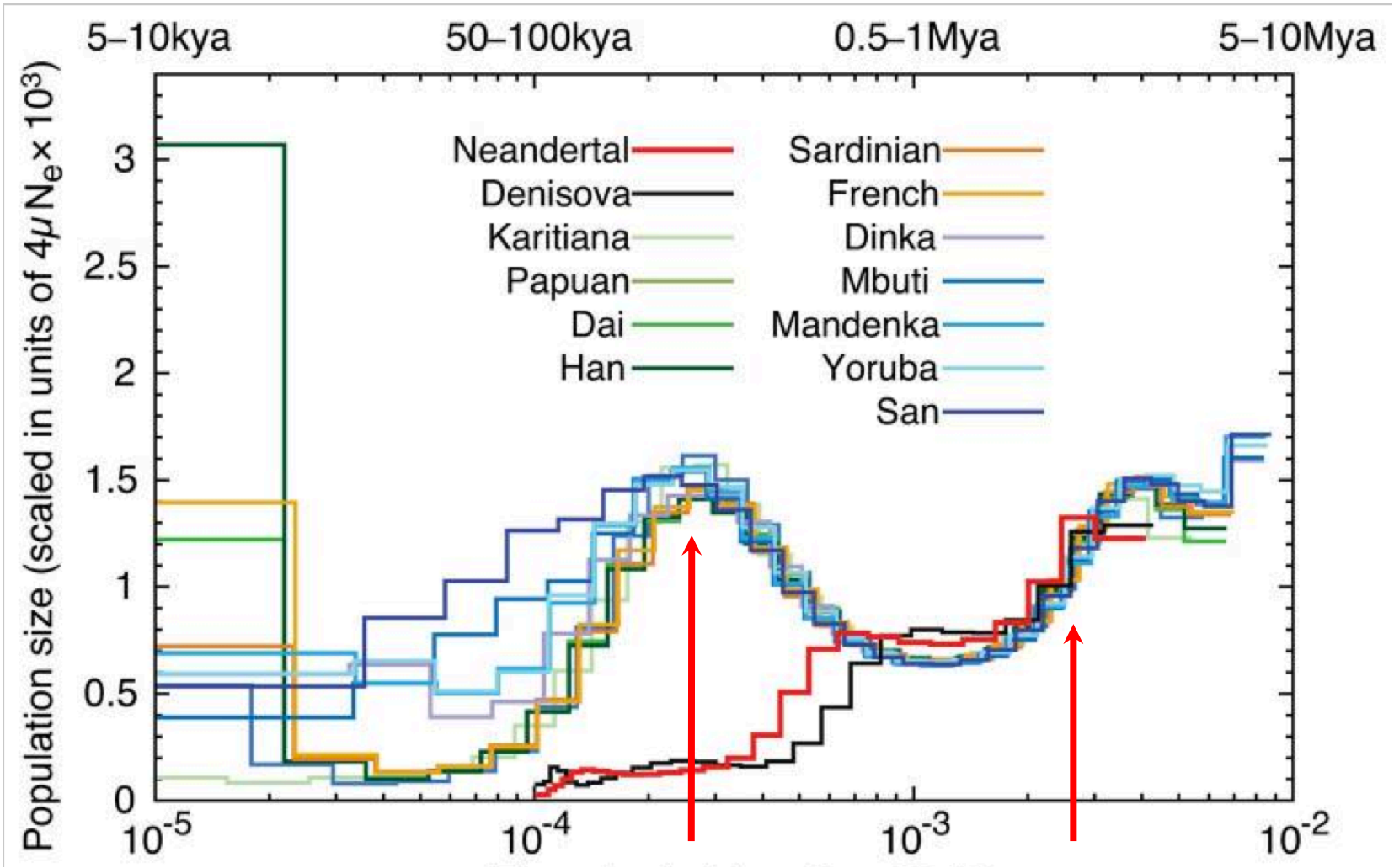






N-island model
Migration between islands controls coalescent rate

Human population history, with Neanderthals

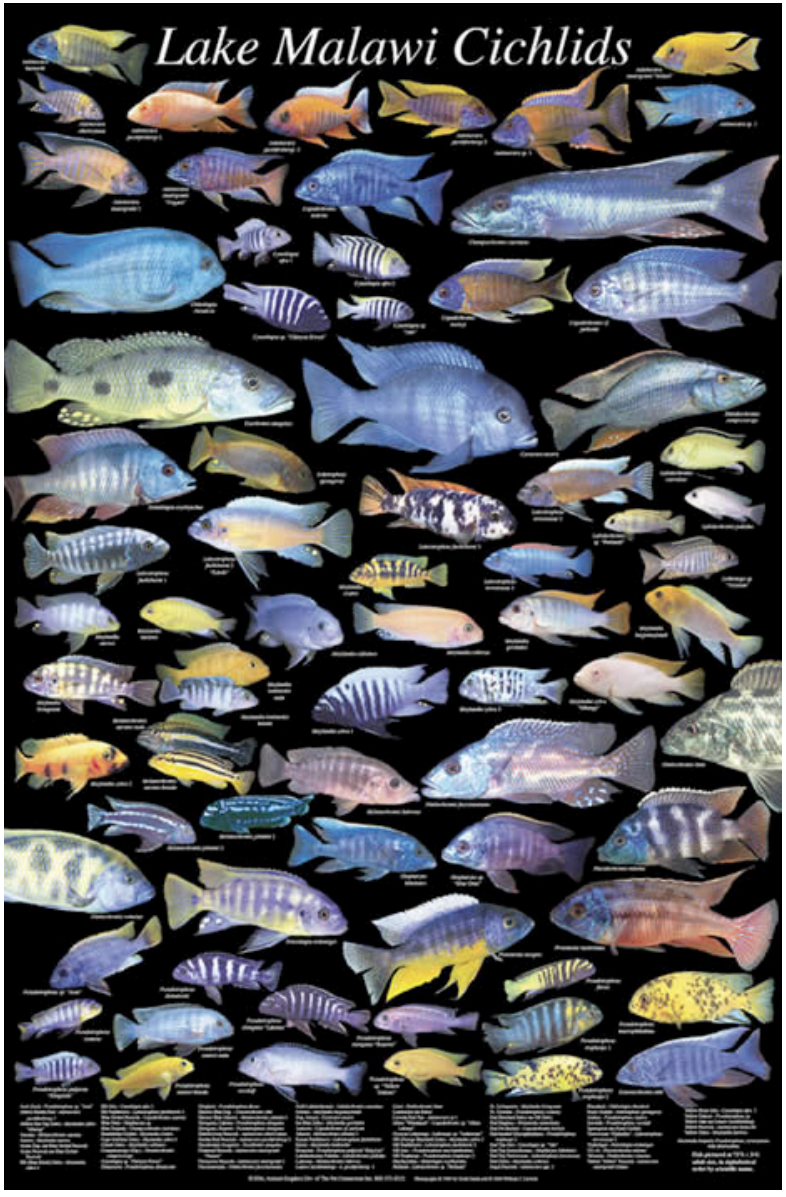


rise of anatomical modern humans 1-200kya

origin of *Homo* 1.5-2Mya

Heng Li

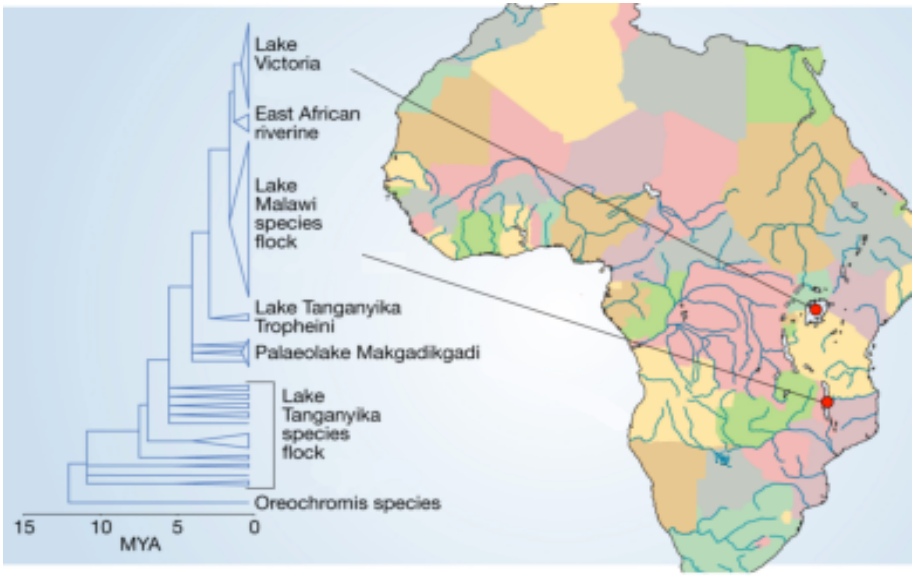
Brief introduction to another system



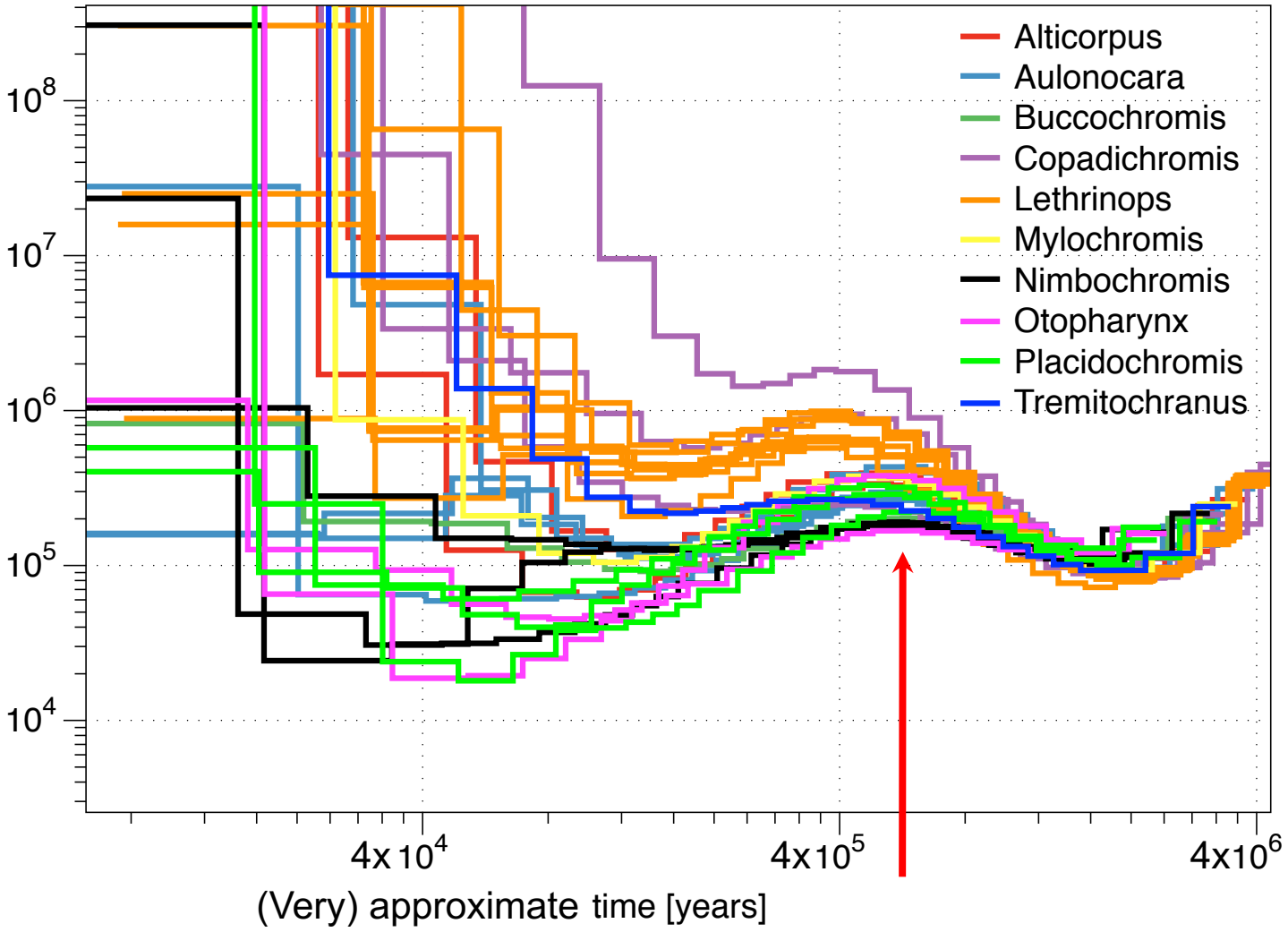
Dramatic recent radiations of haplochromime cichlids in the African rift valley great lakes

- Lake Malawi ~500 species within last 1M years

So far we have sequenced ~300 species at 15-20x coverage



Lake Malawi cichlid PSMC



Is structure associated with speciation?

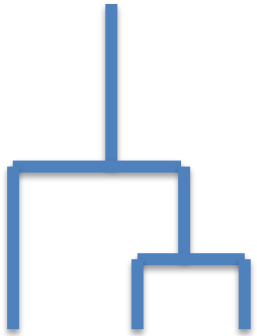
- There is increasing evidence that this is often the case
  - Ideas of hybrid speciation, reuse of alleles selected in different environments, hybrid swarms and gene flow
- But that is another talk…

Might structure be (partly) identifiable in the PSMC model?

- The inferred values N(t) have dimension T, the number of time bins
- But the transition matrix M has dimension T^2
- Currently we derive M from N by theory assuming panmixia
  - Is there a richer theory for structured populations?
  - How to parameterise structural complexity S(t) at time t, with associated theory for M(N,S)
- Or can we fit the transition matrix M unconstrained?
  - Then search for evidence of structure within it
  - And or do goodness of fit?

Going beyond two sequences

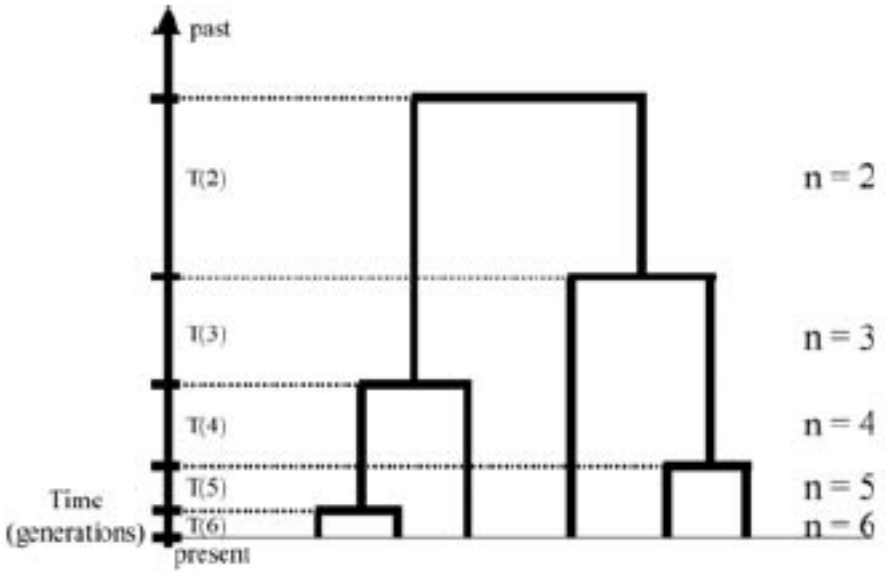
- Chance of coalescence per generation from three sequences is 3/N



- Once we have a coalescence we are back to the situation with two sequences

- From *i* sequences chance is $i(i-1)/2N$

Digression: "The coalescent" model (Kingman, 1980) A distribution on trees



- T(i) ~ exponential with mean 2N/i(i-1)

	E[*T*(*i*)]			E[*T*(*i*)/*N*]		
i	*N*=100	*N*=200	*N*=1000	*N*=100	*N*=200	*N*=1000
6	6.7	13	67	0.07	0.07	0.07
5	10	20	100	0.10	0.10	0.10
4	17	33	167	0.17	0.17	0.17
3	33	67	333	0.33	0.33	0.33
2	100	200	1000	1.00	1.00	1.00

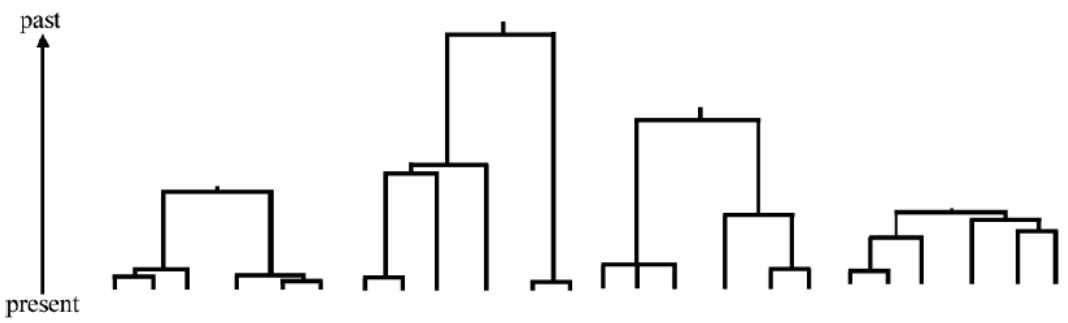
Properties of the coalescent

- As we add extra sequences, they are increasingly likely to coalesce very fast, and increasingly unlikely to affect the full TMRCA

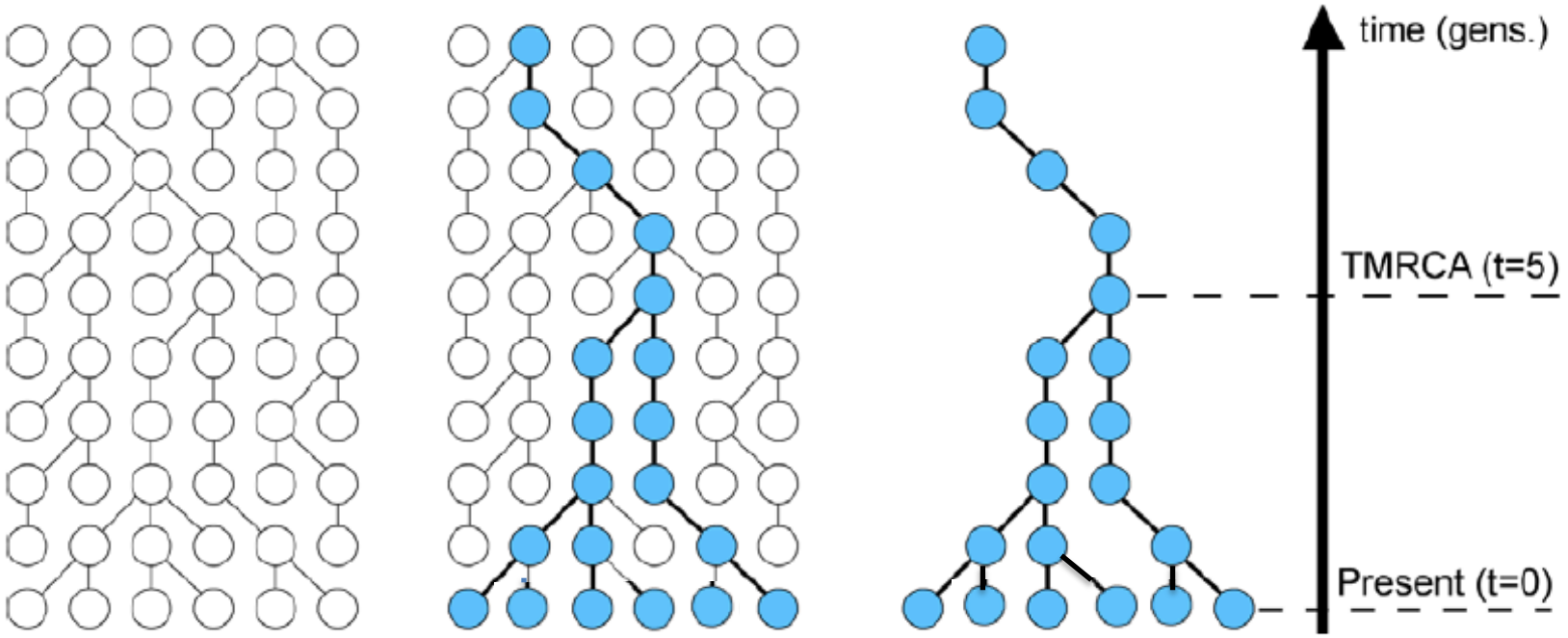
$$E[TMRCA] = \sum_{i=2}^{n} E[T(i)] = 2\left(1 - \frac{1}{n}\right)$$

The expected height of the tree for many samples is only twice that with two samples

- Trees are very variable
  - E.g. 4 samples on 6 leaves



Relationship between forwards in time (Wright-Fisher) and backwards in time (Coalescent) models



Population evolution forwards

Coalescent tree backwards

The coalescent tree describes a *sample* from the forward process

Kingman coalescent generates an “exact” sample from Wright-Fisher

Genetic variation in a sample

- Mutations occur at random on the tree
  - Separation of sources of randomness
    - Random *demography* tree structure from coalescent
    - Random *sampling* of mutations on the tree

Let S be the number of mutations = segregating sites

$$E[S] = 2\mu \sum_{i=2}^{n} iT(i)$$

$$E[S] = \frac{\theta}{2} \sum_{i=2}^{n} iT(i)$$

$$E[S] = \frac{\theta}{2} \sum_{i=2}^{n} i \frac{2}{i(i-1)}$$

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

Watterson's theta

$$\hat{\theta}_S = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

$E(S) \sim \theta \log n$

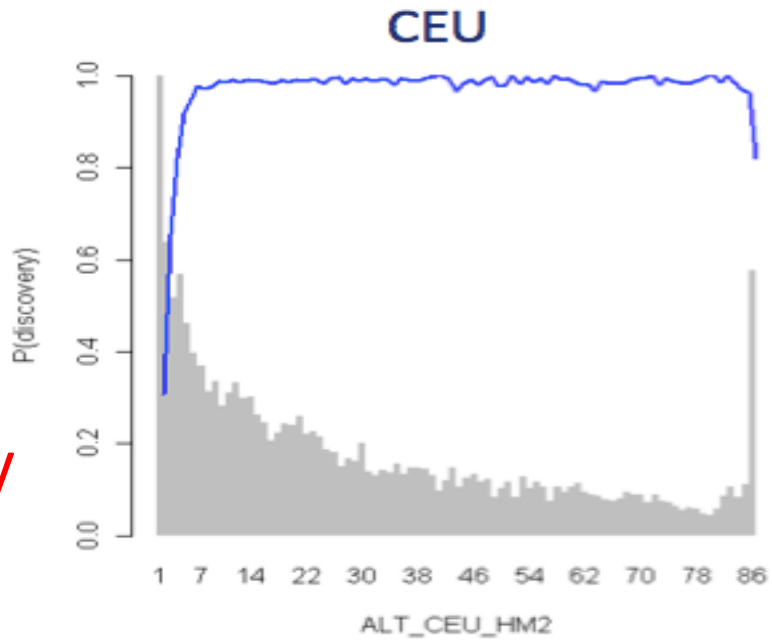
Distribution of variant allele frequencies

- Density of mutations with frequency *i* in a sample of *n* is θ/i

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

- *1/f* distribution of population allele frequencies

site frequency spectrum SFS

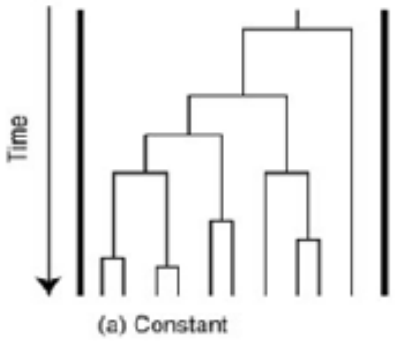


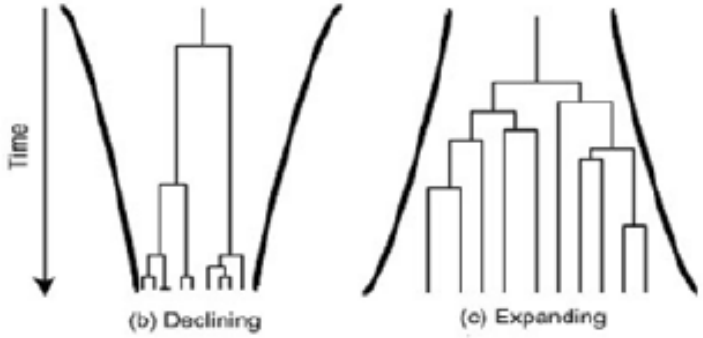
- Population minor allele frequency distribution of a difference observed between two sequences is flat
  - Probability $(1/f).2f(1-f) = 2(1-f)$, folded at ½ is 2

Relaxation of assumptions (1)

- E.g. change in population size changes the site frequency spectrum (SFS)

  This is the basis of SFS-based demography inference





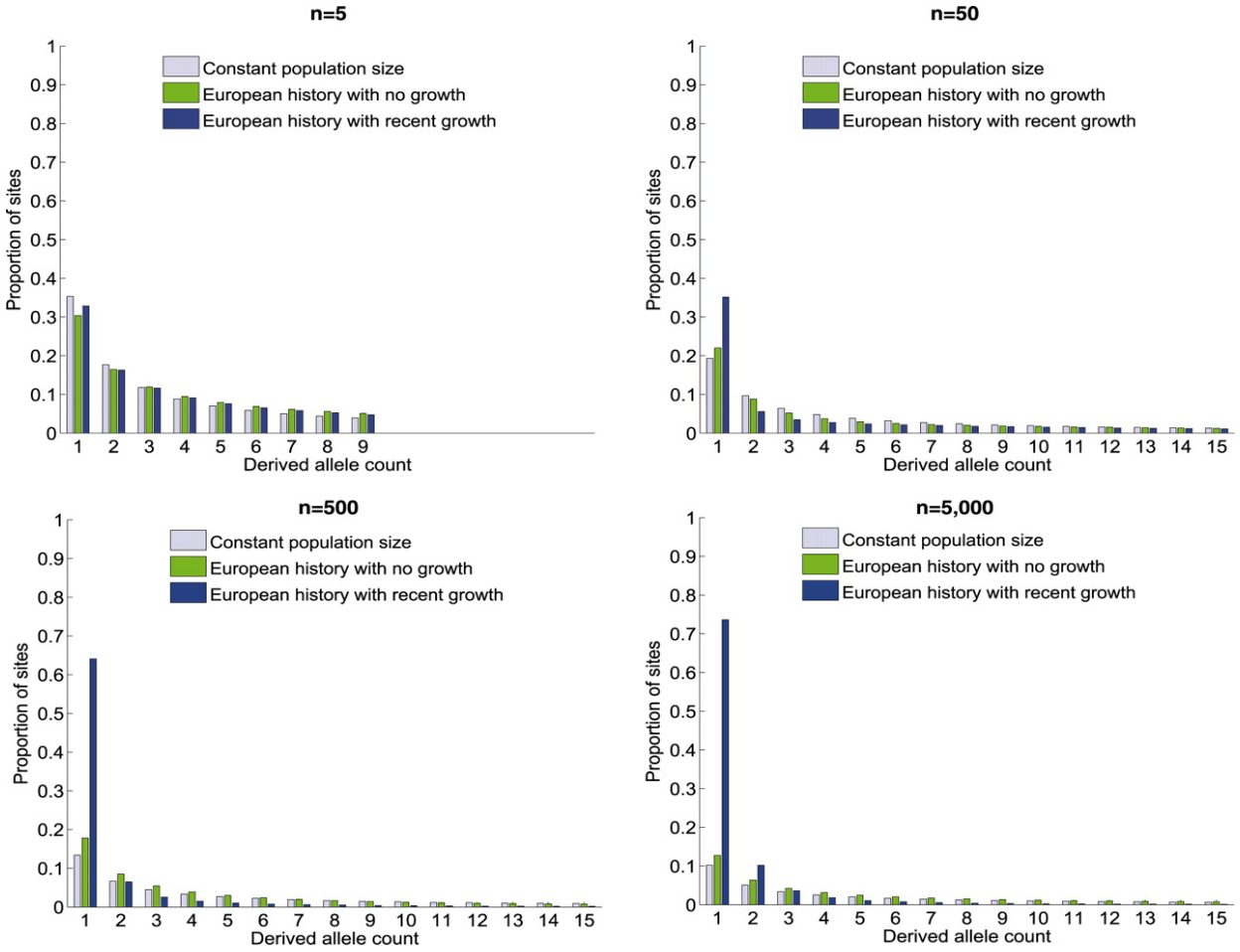
D > 0 D < 0

- Tajima's D

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{Var(\hat{\theta}_\pi - \hat{\theta}_S)}}$$

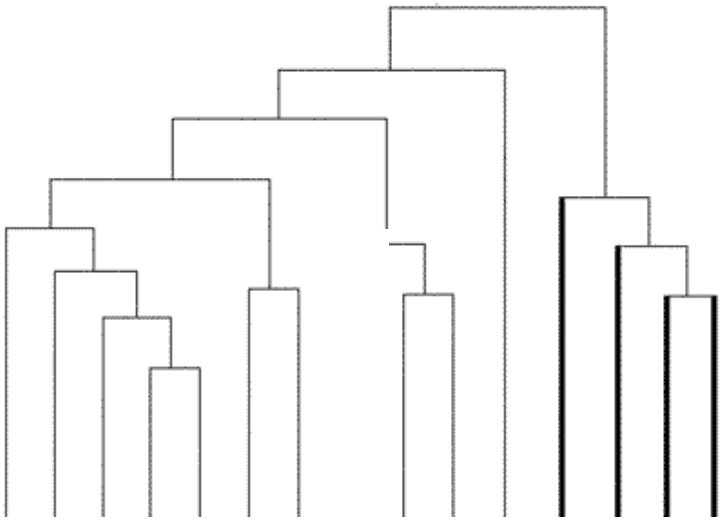
  - Sensitive to number of rare mutations, so change in N_e
  - If D is positive there is a deficiency of rare mutations
    - Excess recent coalescences, recent small N_e - *selection*

Fig. 2 The expected site frequency spectrum (SFS) of the derived allele (the new mutation arisen in the population) for three different demographic models: (i) a population that has been of constant size throughout history; (ii) a model previously fit to the derived allele frequency spectrum of Europeans, which includes an out-of-Africa population bottleneck and a second, more recent, population bottleneck (21); and (iii) the same two-bottleneck model of European history with the addition of recent exponential growth from a population size of 10,000 at the advent of agriculture to an extant effective population size of 10,000,000, which amounts to 1.7% growth per generation during the last 400 generations.



A Keinan, A G Clark Science 2012;336:740-743

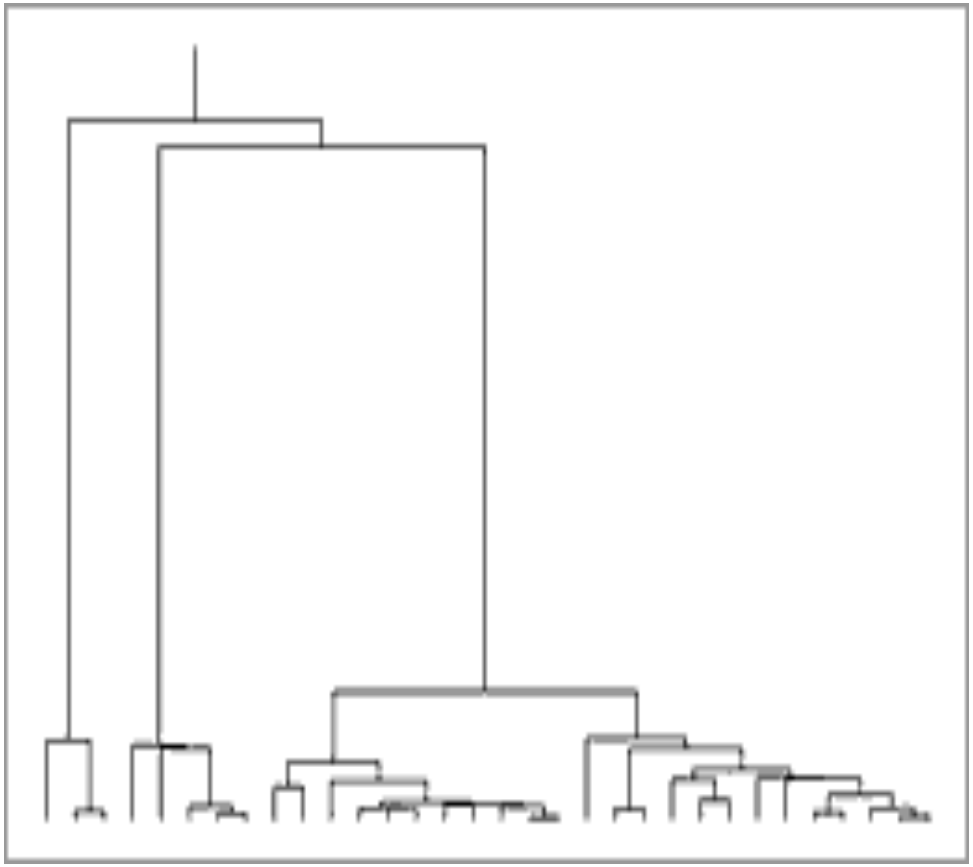
Individuals in human outbred populations still carry many variants not in the large sequence data sets (1000 Genomes etc.)



- Exponential population growth in last 10,000 years gives long tips to the tree
- In “big” populations, tips are hundreds of generations long, so tens of thousands of private variants per sample, hundreds functional

This behaviour is very dependent on population structure.

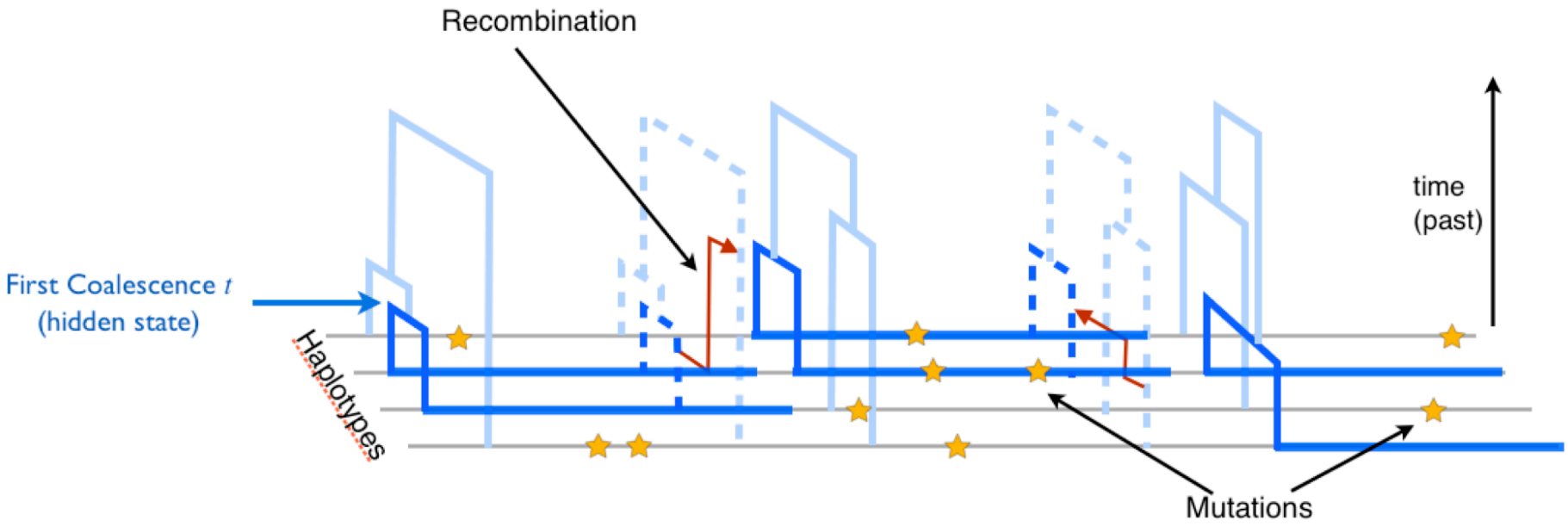
In genetic isolates the recent effective population size is smaller, and the tips are shorter



What about recombination?

- If points on the genome are very close, e.g. adjacent, they share the same tree
- If points are very far, their trees are sampled from the coalescent independently
- What happens in between?
- A recombination in the ancestor of a modern sequence made it out of two separate sequences, one contributing to the left and one to the right

Recombination changes the tree as you move along the sequence



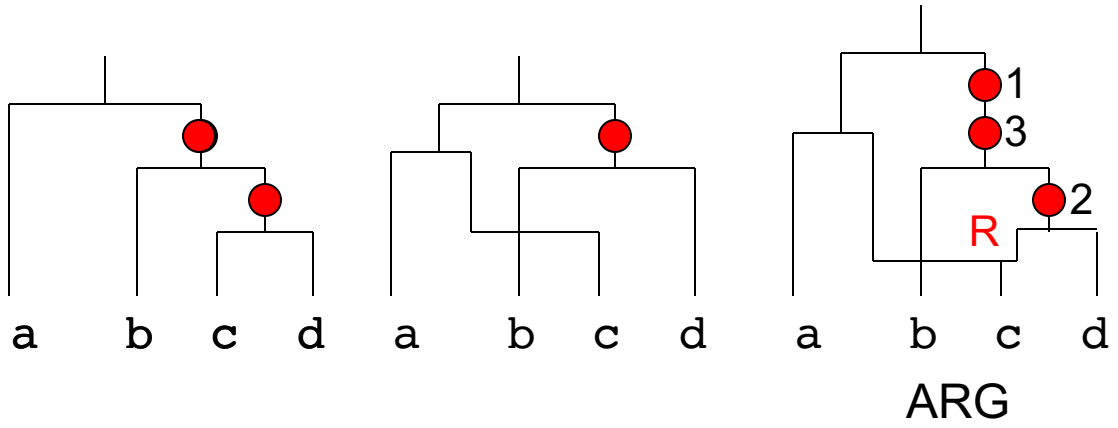
Typically recombination rate is comparable to or larger than the mutation rate: both ~10^{-8} /bp /gen in human

So "gene tree" varies every site in mixing populations

Ancestral Recombination Graph (ARG)

- The *Ancestral Recombination Graph* describes the way that individual sequences in a population are related
  - At a locus, sequences are related by a tree
  - Ancestral recombinations change the tree as you move along the chromosome

```
a  ..C..G..A..   0 0 0
b  ..T..G..C..   1 0 1
c  ..T..A..A..   1 1 0
d  ..T..A..C..   1 1 1
```



"Prune and graft" operation going left to right

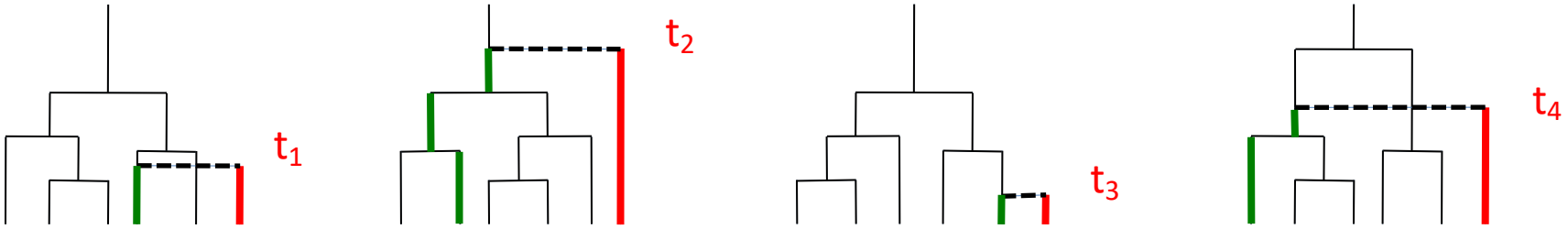
Coalescent with recombination

- ARG is a structure (data type)
- The probability distribution over ARGs that arises when recombination is added to the standard (Wright-Fisher) model is called the *Coalescent with Recombination*
  - Hudson's *ms* software is the classic simulator
  - *msprime* from Jerome Kelleher is MUCH faster
- Now two possible events going backwards in time
  - Coalescence: which merges two sequences
    - For i sequences, rate is $i(i\text{-}1)/2N$
  - Recombination: which splits a sequence into two
    - For i sequences, rate is $iL\rho$

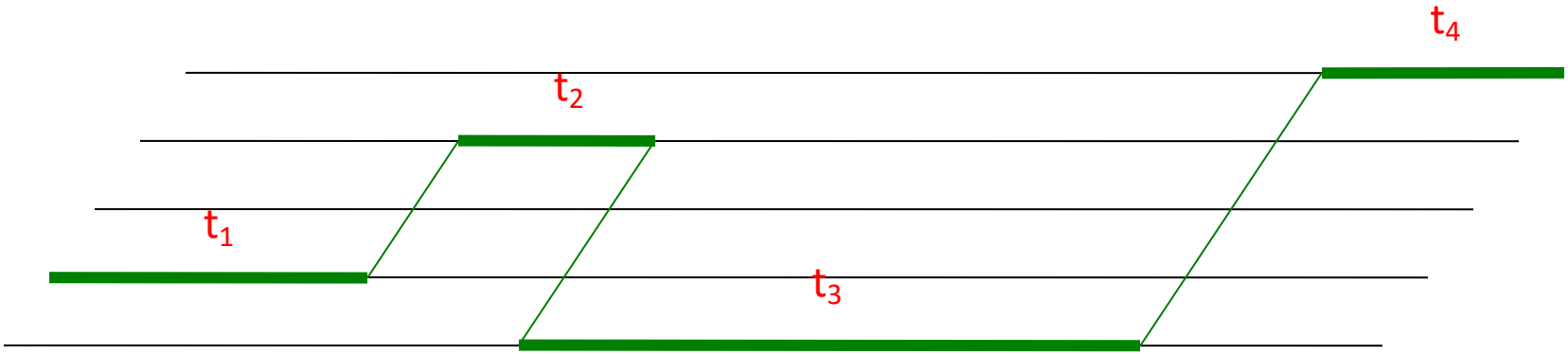
Extending to multiple sequences

- The recent time limit of ~20kya for PSMC is set because we run out of recent coalescences between two haplotypes
- If we add more haplotypes, then there are more recent coalescences and we could look at more recent history
- But, … the hidden state is then a tree (with branch lengths): impractical to model fully
  - MCMC is notoriously difficult

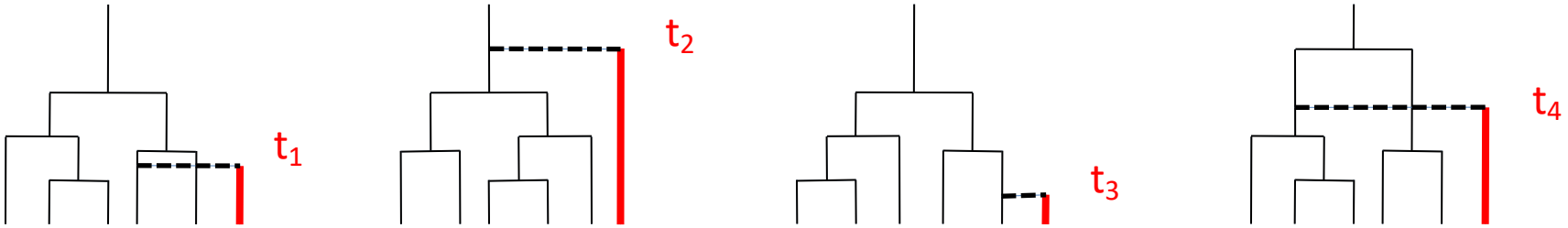
Option 1: First coalescence of one sequence to the tree of the others



- This is related to the Li and Stephens model (or Stephens and Donnelly) – chromopainter



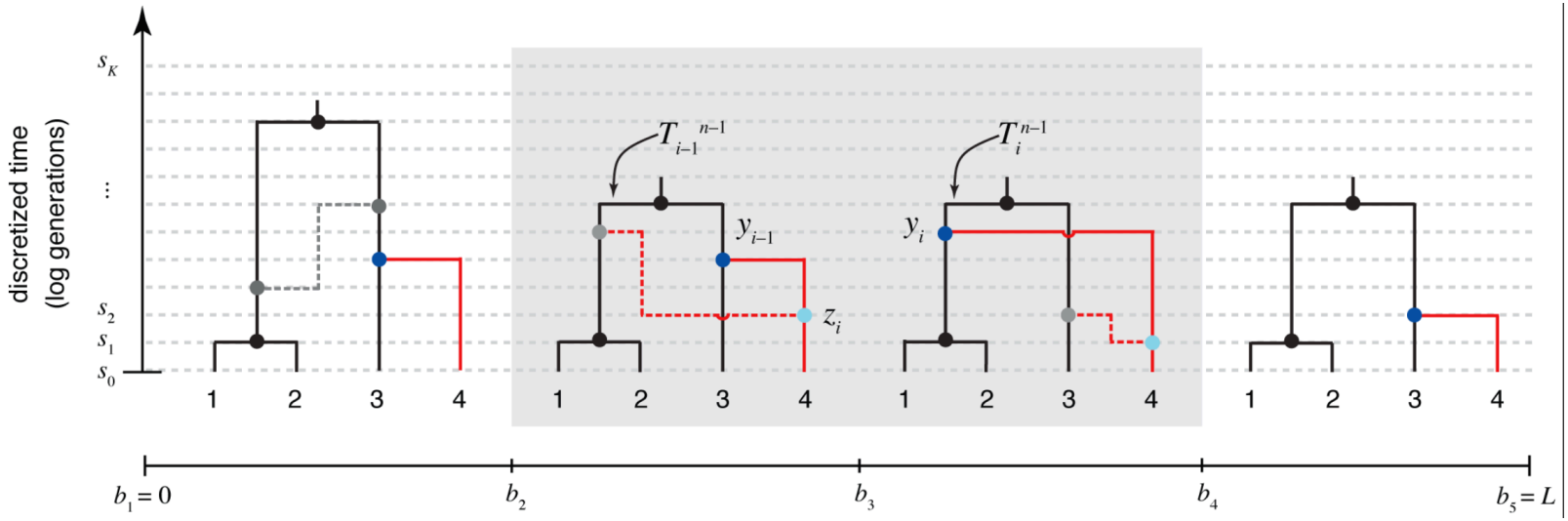
Problem: Coalescence of chosen sequence to the others depends on the number of lineages M(t) remaining at time t



- M(t) is a random variable, and we need the entire history of M(t) to calculate transition probabilities q
- Huge increase in state space and/or this breaks Markov assumptions

MCMC approach: ARGweaver

- Repeatedly remove a sequence* and add it back, sampling conditional on remaining ARG
- HMM: sample with forward-backward algorithm



- Costly – use for inference given history

Genome-wide inference of ancestral recombination graphs

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. *PLoS Genet.* 10:e1004342 (2014)

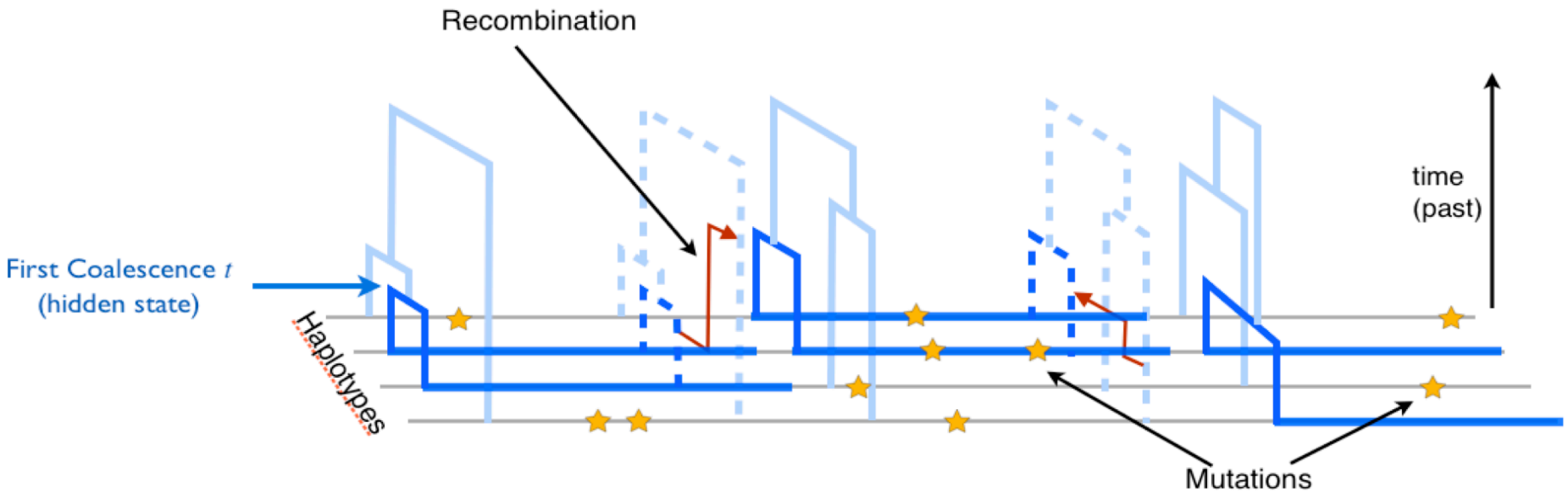
Heuristic approximations

- Relate – Speidel and Myers this morning
  - Sequence of trees, not full ARG structure
  - Not good for inference about recombination, but can be used for N_e estimation and other things
- tsinfer – Wong, Kelleher, … McVean
  - Current released version gives topology of tree sequence only – closer to ARG
  - Unpublished tsdate will allow demography analysis

Option 2: first coalescence between any pair

- This remains (approximately) Markov
- State space is $O(M^2T)$ – pair of states and time they coalesce
  - But transition updates are only $O(M^2T^2)$, because transitions are memoryless
- Emissions from X_{ij} are singletons on *i* or *j*
  - Non-singletons that are discrepant between *i* and *j* wipe out density at X_{ij}

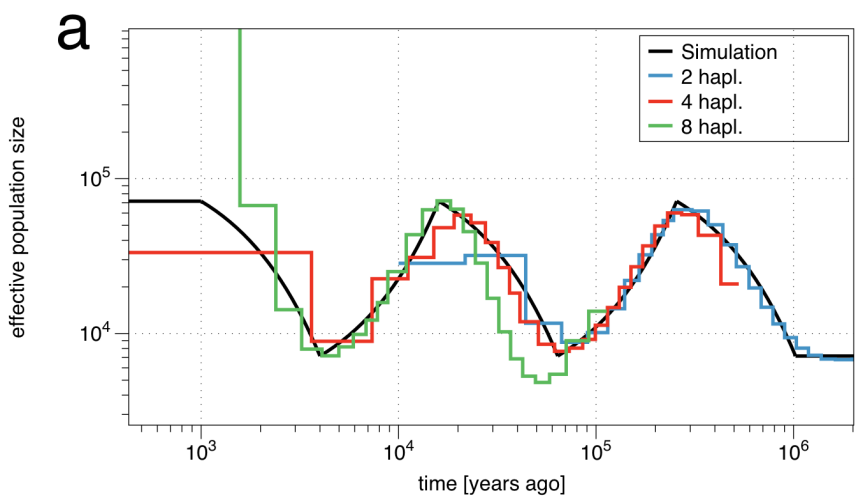
MSMC

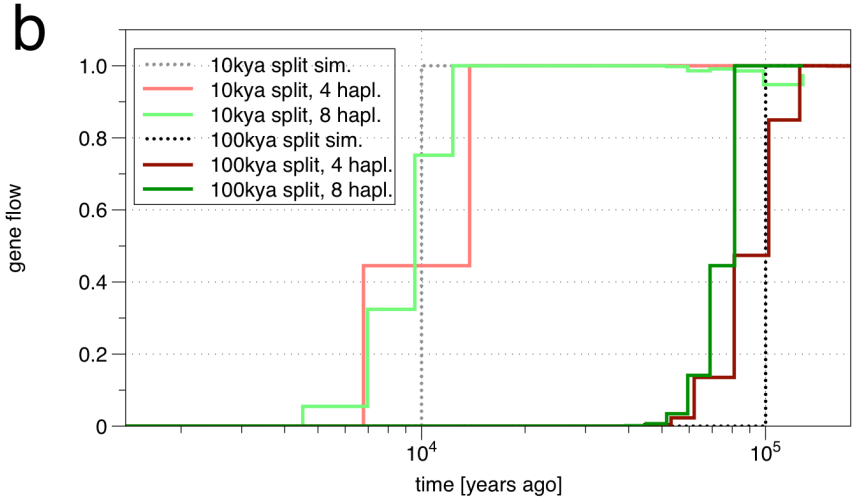


Stephan Schiffels and Durbin (Nature Genetics, 2015)

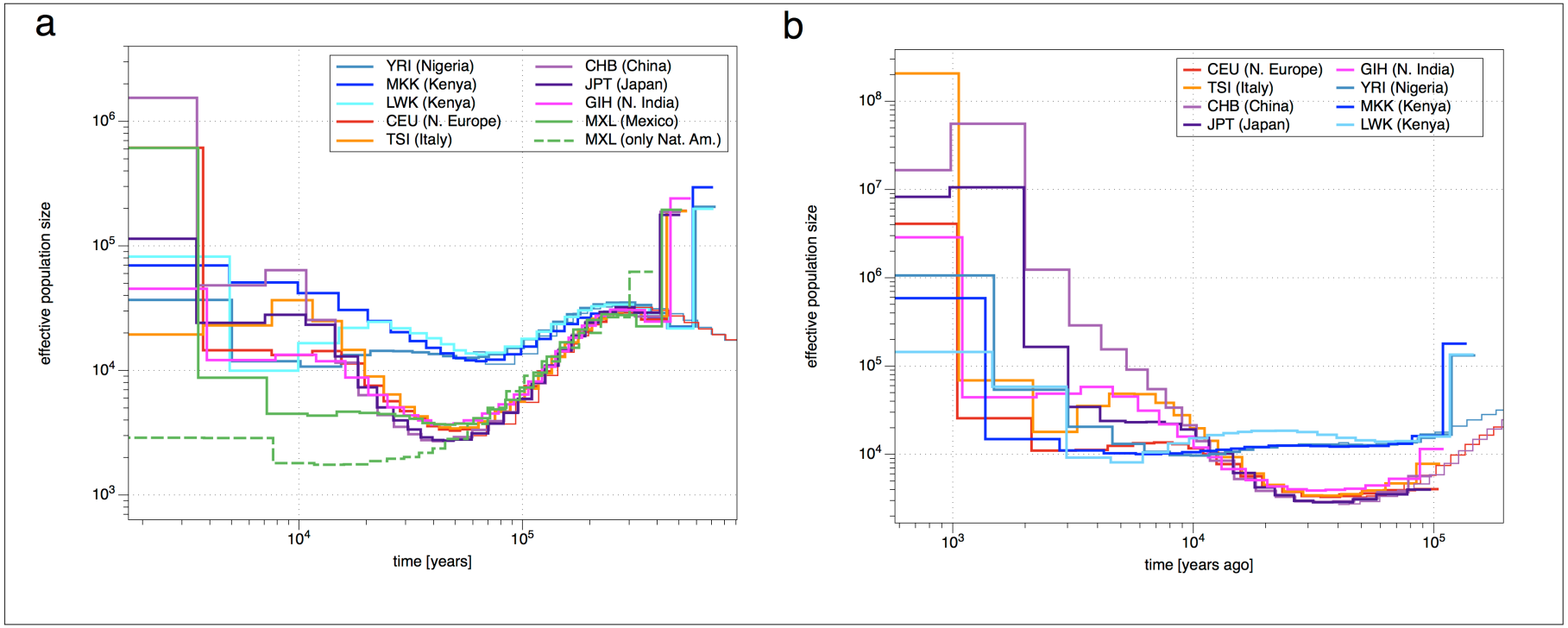
MSMC can fit both population size history and separation history

- Separation via the (scaled) ratio of coalescence between and within populations





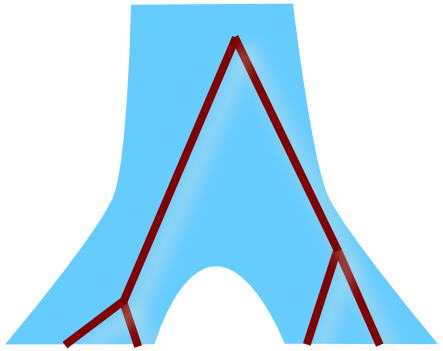
Access more recent history



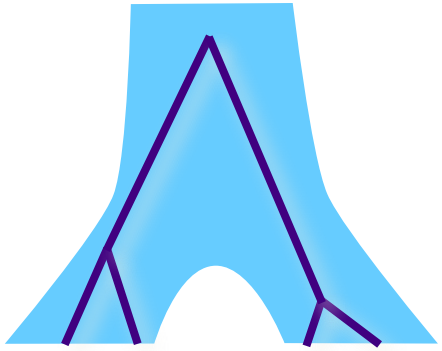
Use lower mutation rate here ~ 0.5×10^{-9}/year

Divergence between populations

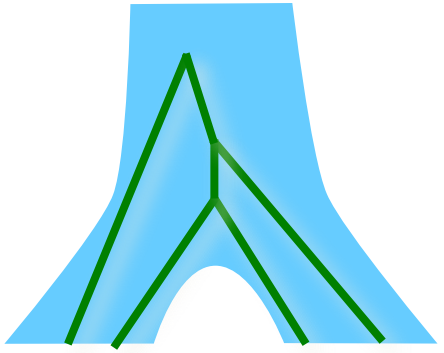
- Idea: Infer separate coalescence rates within and between populations:



First Coalescence within Population 1



First Coalescence within Population 2

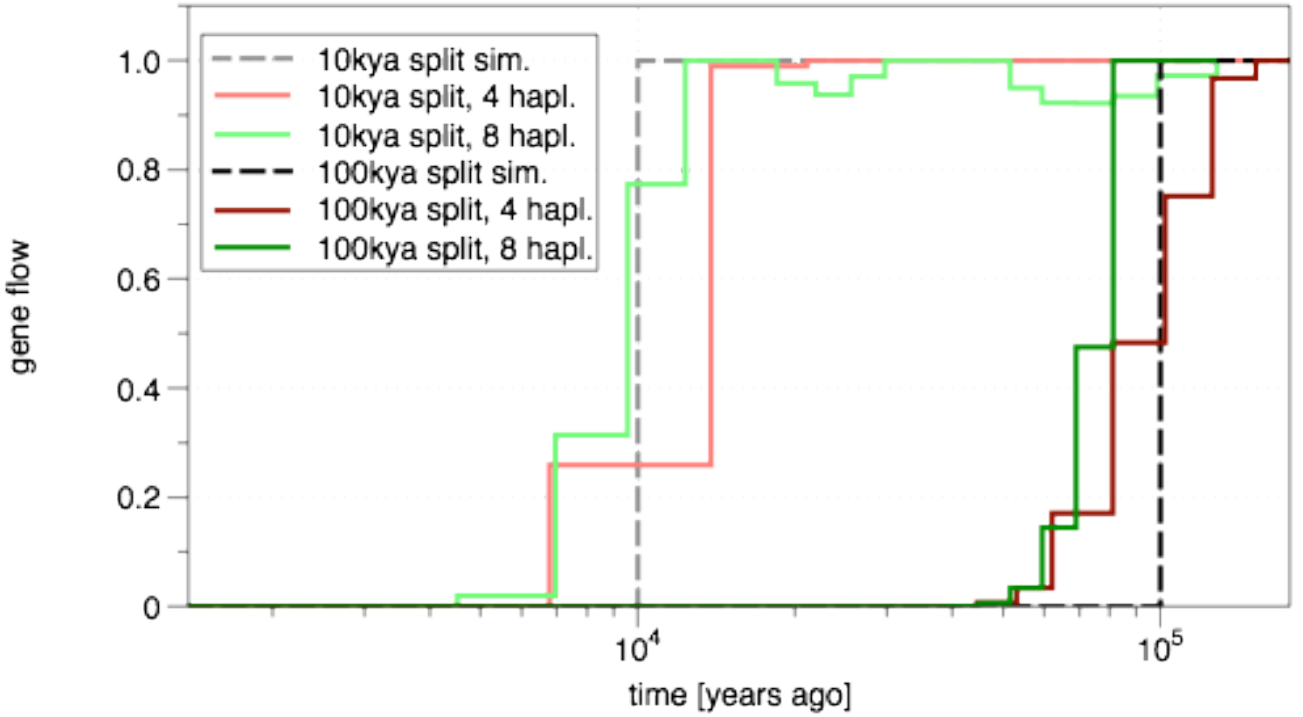


First Coalescence across both populations

- MSMC can infer separate coalescence rates within and between populations
- Given rates within populations, $\lambda_{11}(t)$ and $\lambda_{22}(t)$, and across populations, $\lambda_{12}(t)$, compute relative gene flow as ratio

$$m(t) = \frac{\lambda_{12}(t)}{[\lambda_{11}(t) + \lambda_{22}(t)] / 2}$$

Testing gene flow inference with simulated split



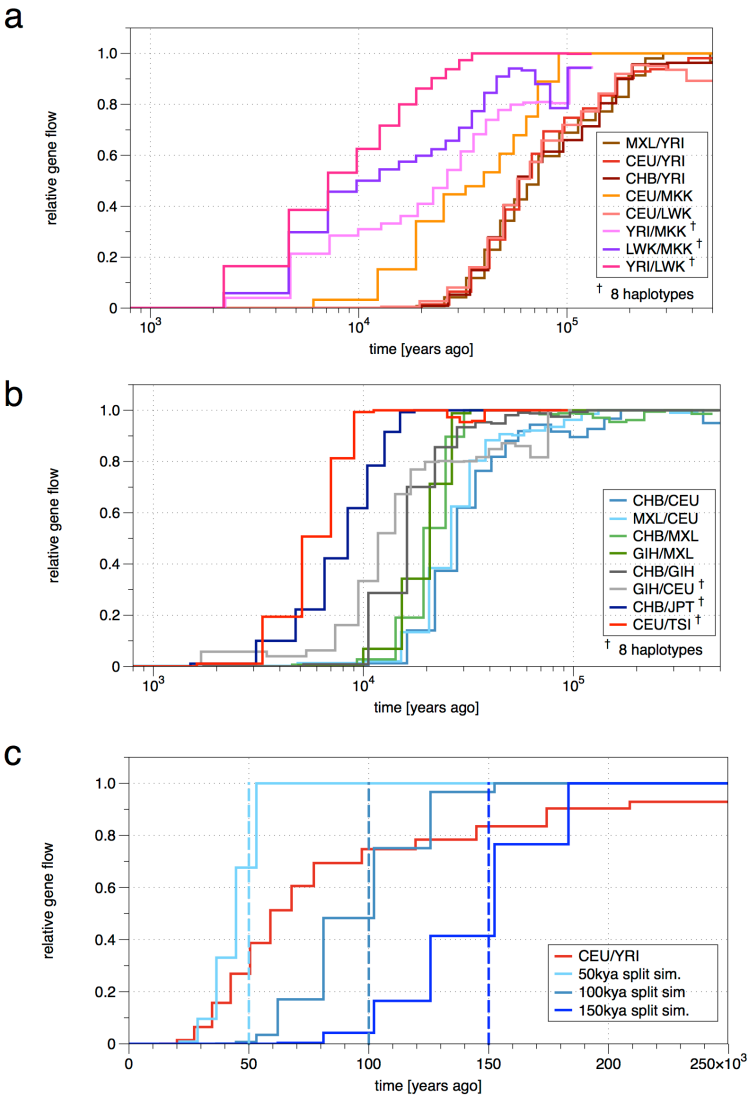
☜ m(t)=1: perfectly mixed

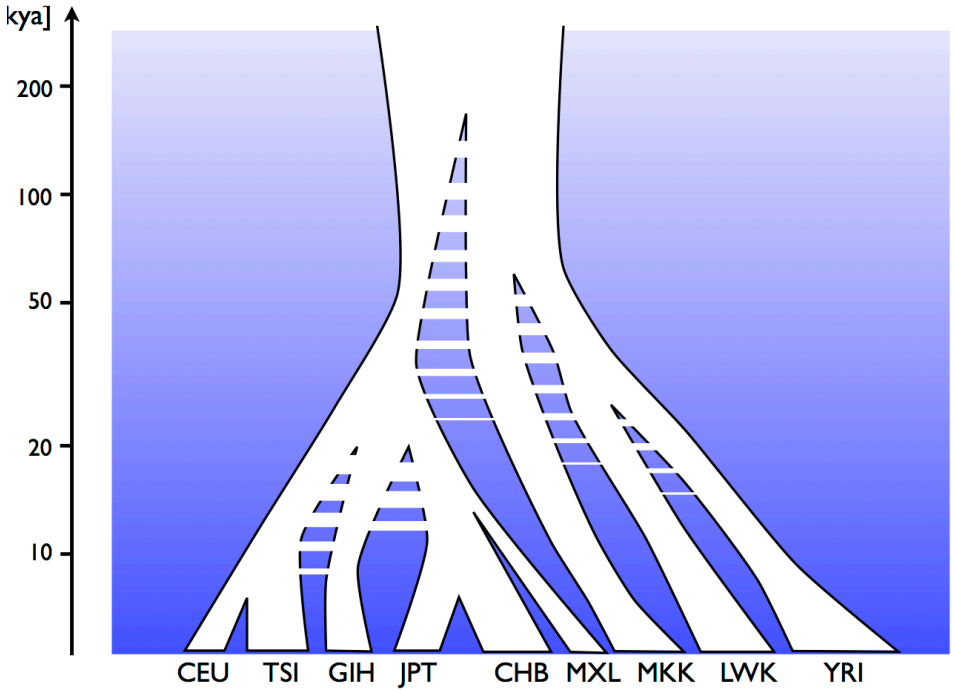
☜ m(t)=0: perfectly split

4 haplotypes: good for splits 50-200kya.

8 haplotypes: good for splits 5-50kya.

Separation history





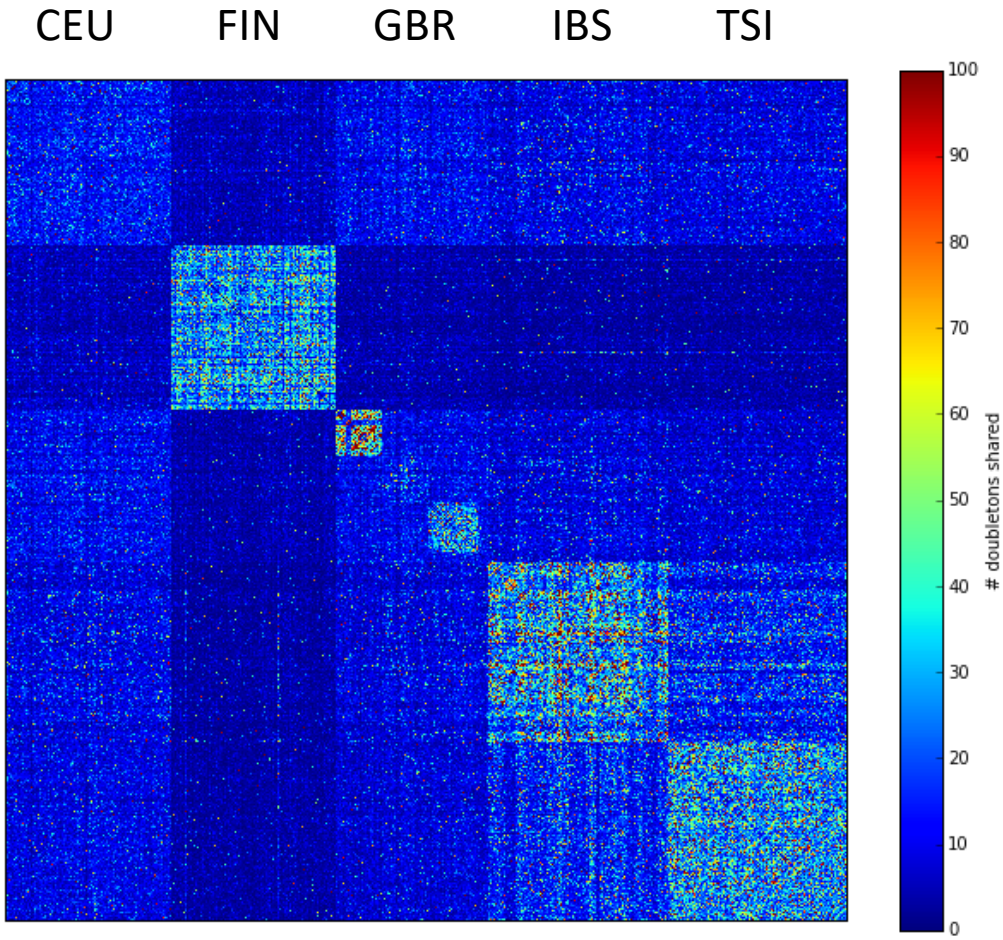
Alternatives to MSMC

- MSMC2 (Schiffels: in Malaspinas 2016/unpub.)
  - Run PSMC' on all pairs of sequences independently
  - Multiply the likelihoods – **Composite likelihood**
    - Assumes the pairs are independent, which is false
    - But gives unbiased estimation (though overconfident)
- SMC++ (Terhorst, Kamm, Song: Nat Gen 2017)
  - Pair, with p(het | **other sequences**)
  - Very cool – works even on genotype data!
  - But there are approximation problems analogous to those in MSMC – not a panacea

Using rare variants to infer demographic history

- Rare variants contain information about recent population history and structure
- Shown here: number of doubletons shared among European samples
- We would like to estimate population split times and population sizes from the frequency of rare variants

Compare to ChromoPainter data



[1000 Genomes Project, Phase3]

Ancient samples from Hinxton

12885A, Saxon

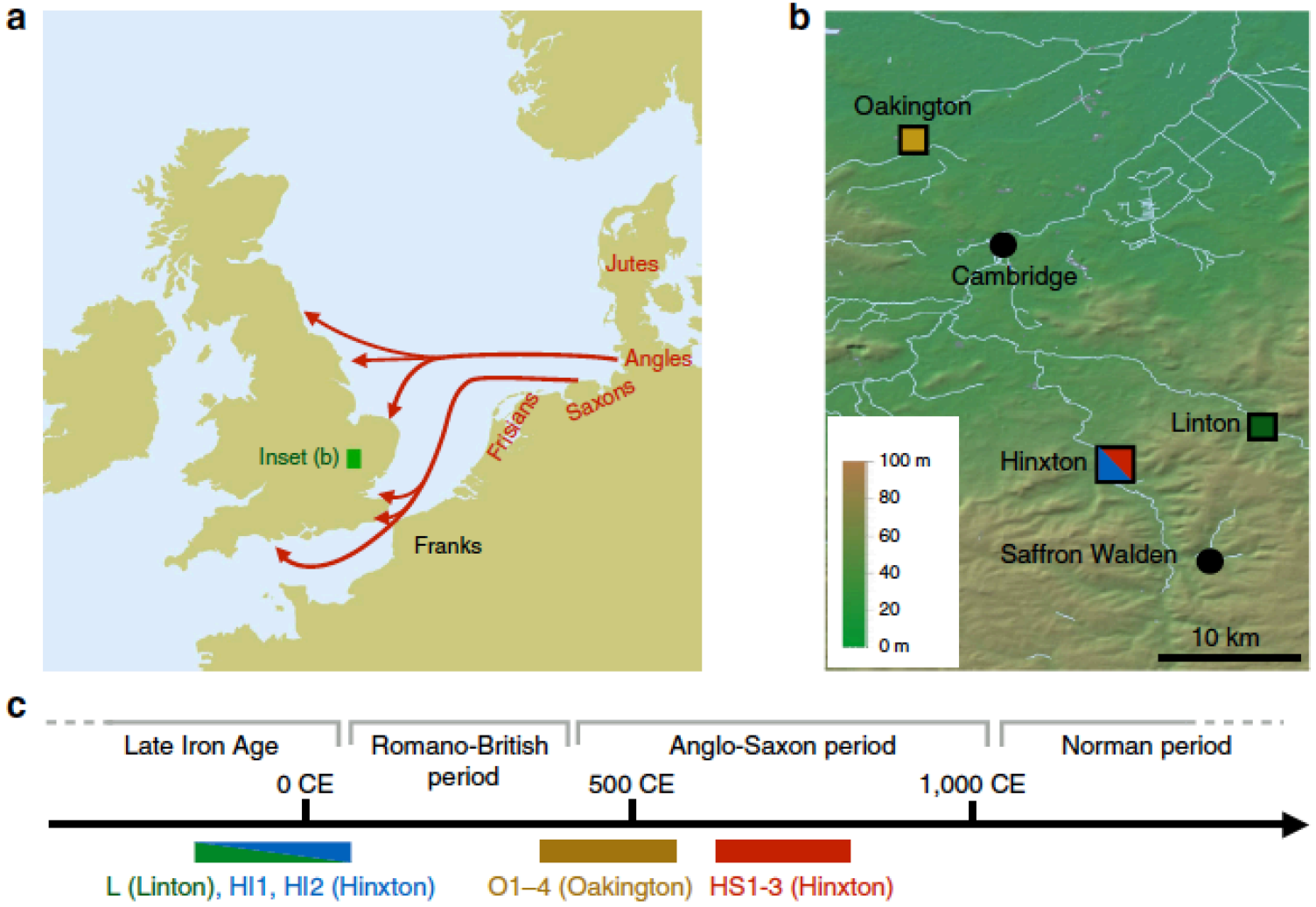
12881A, Saxon

12884A, Iron age

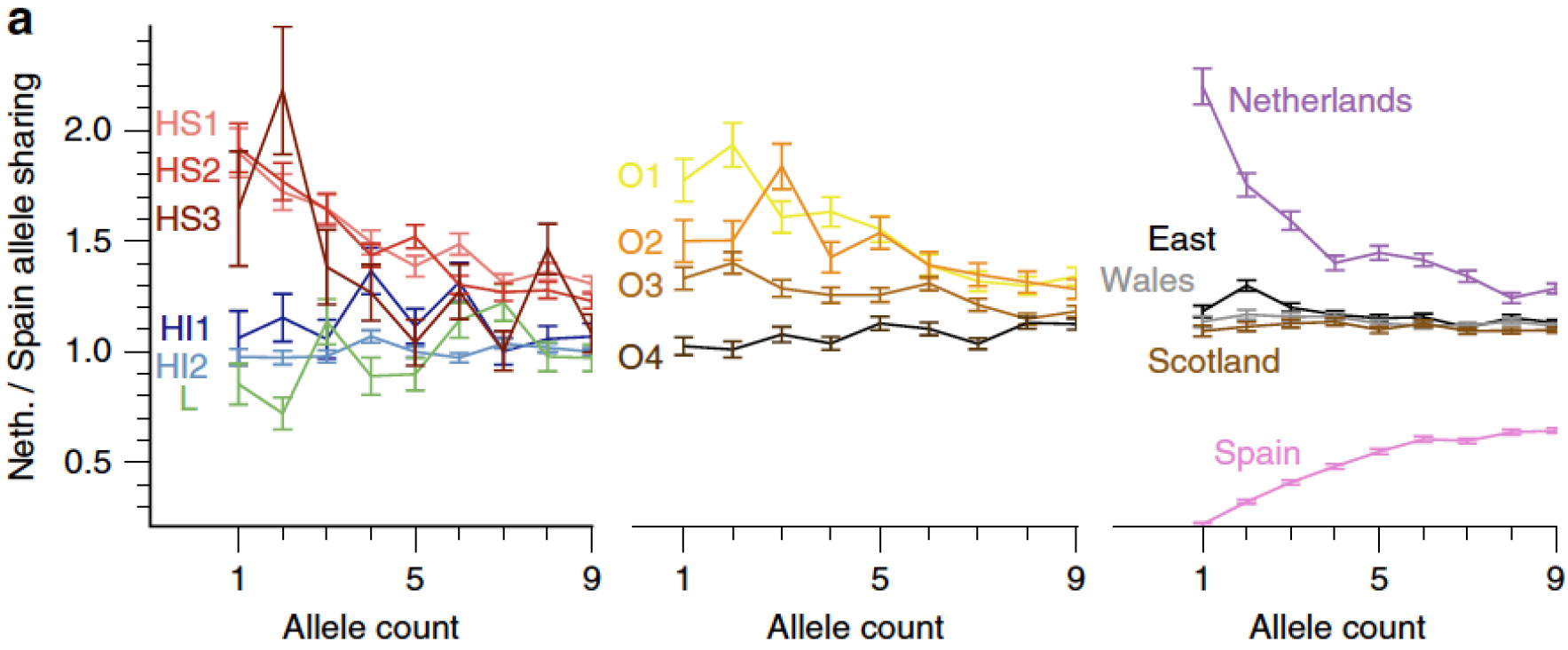
12883A, Saxon

12880A, Iron age

More samples from Linton/Oakington

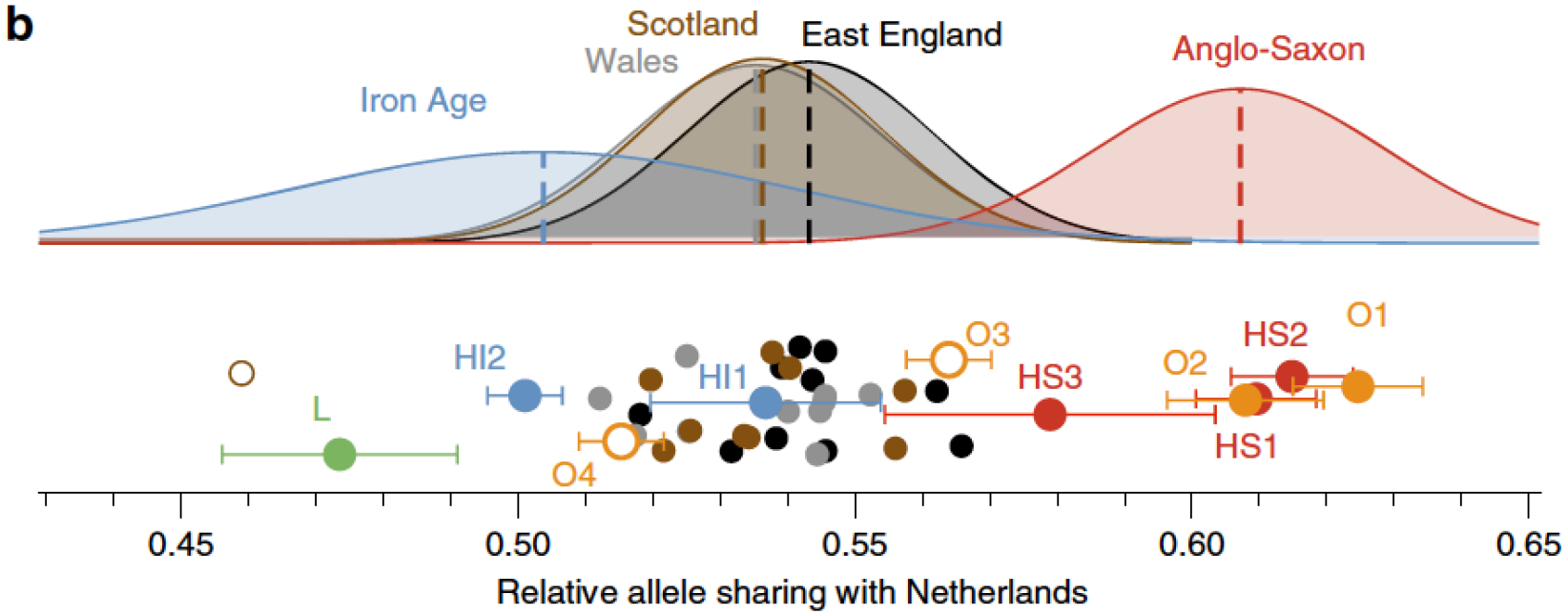


Sharing patterns between ancient and modern samples



- Small but significant differences also within modern Britain (UK10K): Samples from Wales and Scotland share fewer rare variants with Dutch people

Estimates of Anglo-Saxon contribution to modern British genomes



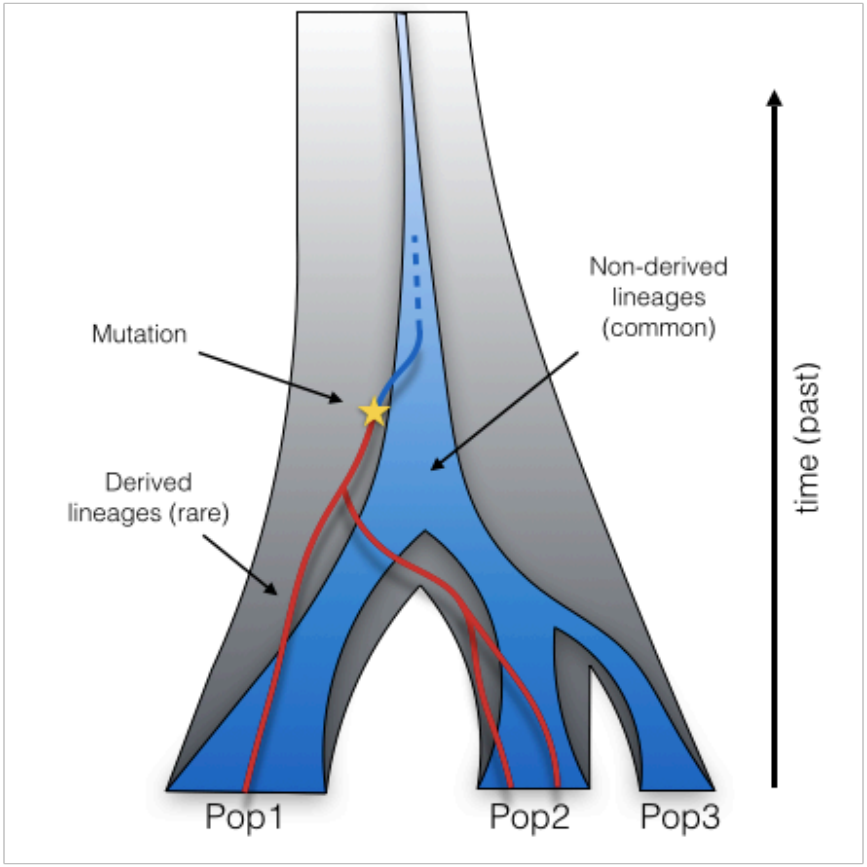
- Suggests ~30% Saxon contribution to samples in East of England, and ~20% to UK10K samples from Wales and Scotland

- Consistent with 20-40% indirect estimate from POBI (Peoples of the British Isles) study

The rare allele coalescent

- Goal: Estimate demographic history (population sizes and split times) from rare variants

- Compute likelihood of demographic model given a distribution of rare variants



RareCoal model

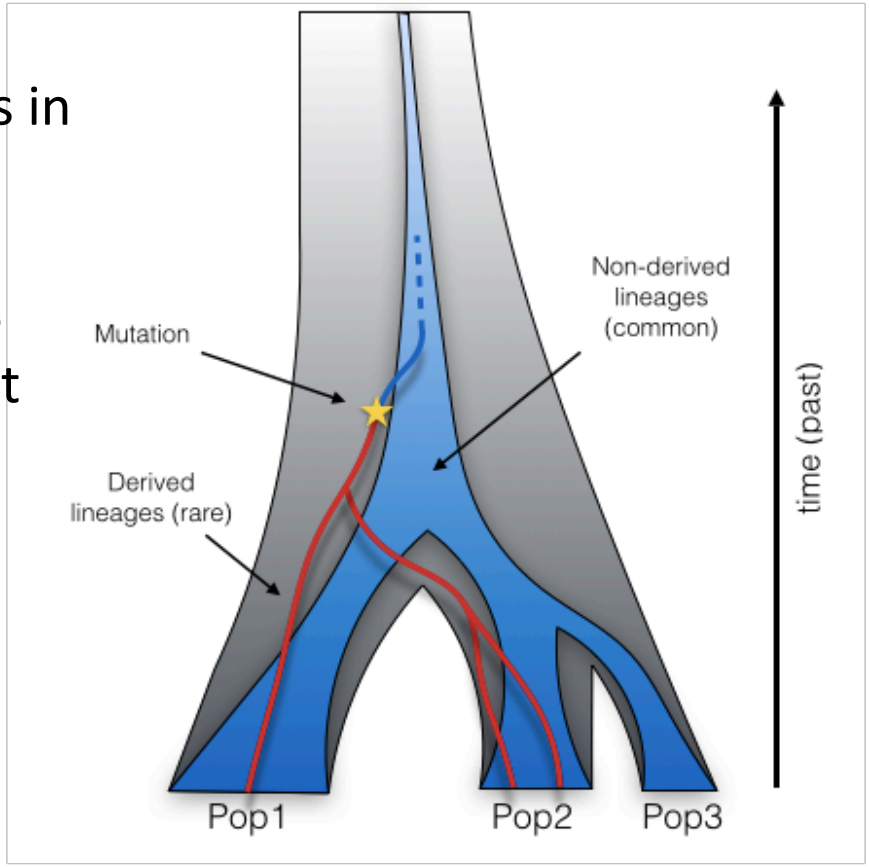
- Idea: Define recursion equations for probability of observing ***i*** derived alleles in population ***k***:

$$b_k^i(t)$$

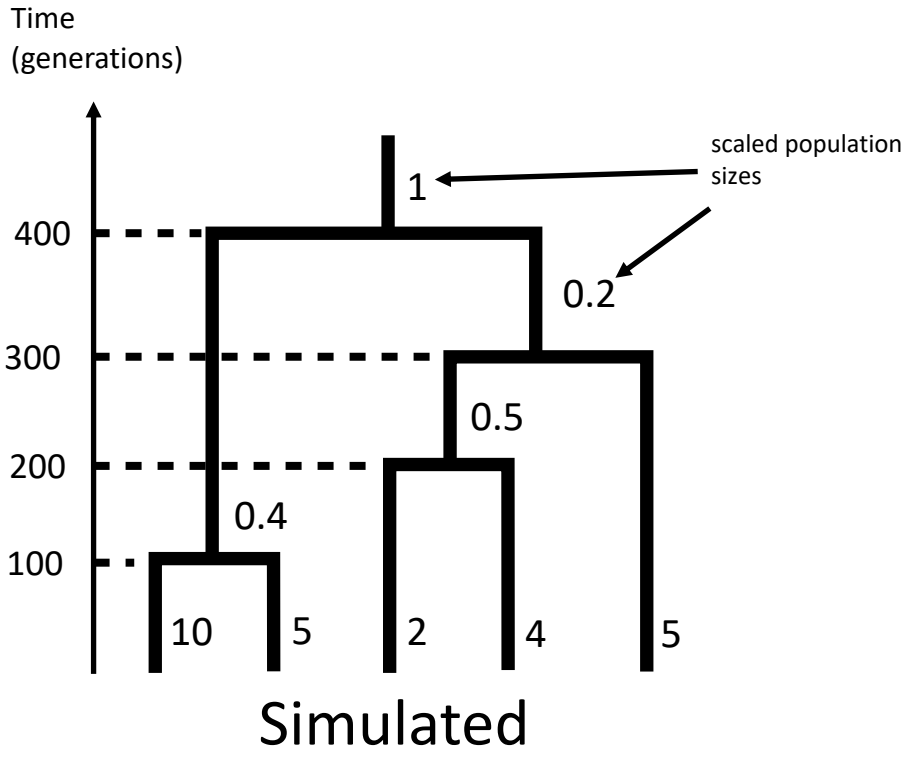
- Given a demographic model, propagate this probability backwards in time to get full likelihood of the data.

- Key simplification: Treat number of ancestral alleles over time as average (mean-field approximation):

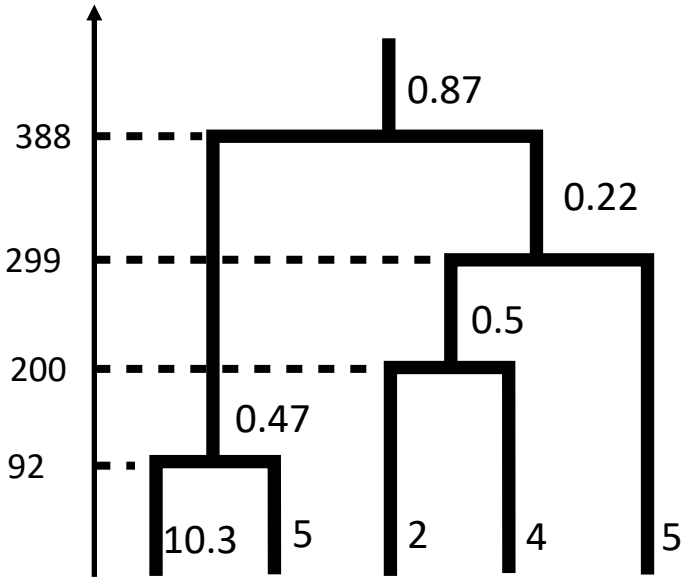
$$a_k(t)$$



Test inference with simulated data



Simulated



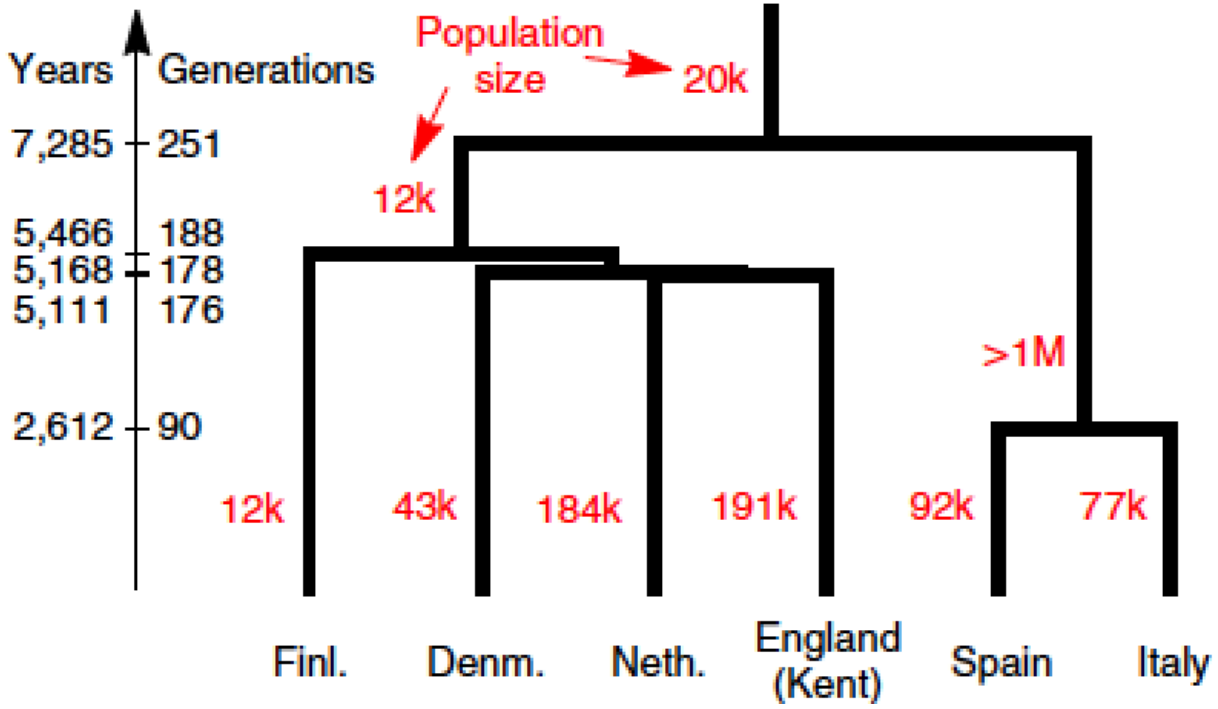
Estimated

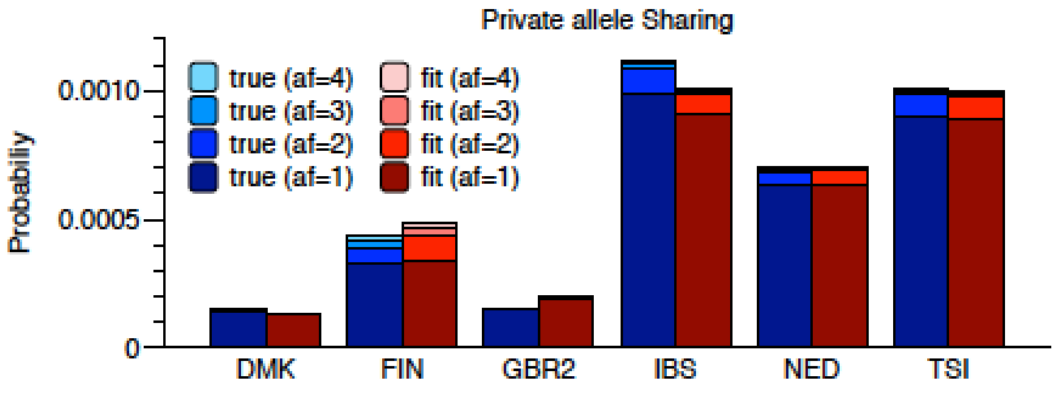
Fits (100 samples per pop.):

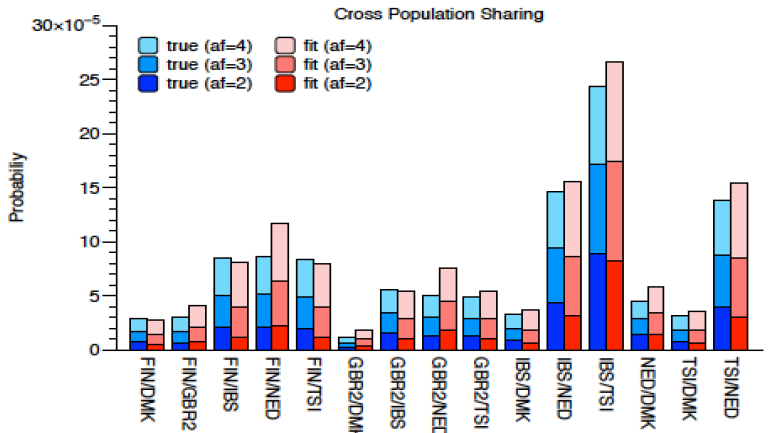
0,1,0,2,1	1114	1159
2,1,0,0,0	140585	139657
1,0,2,0,0	1138	1205
thousands of rows …		

Fitting population sizes and split times separates **drift** from **divergence -> different from Treemix, qpGraph etc.**

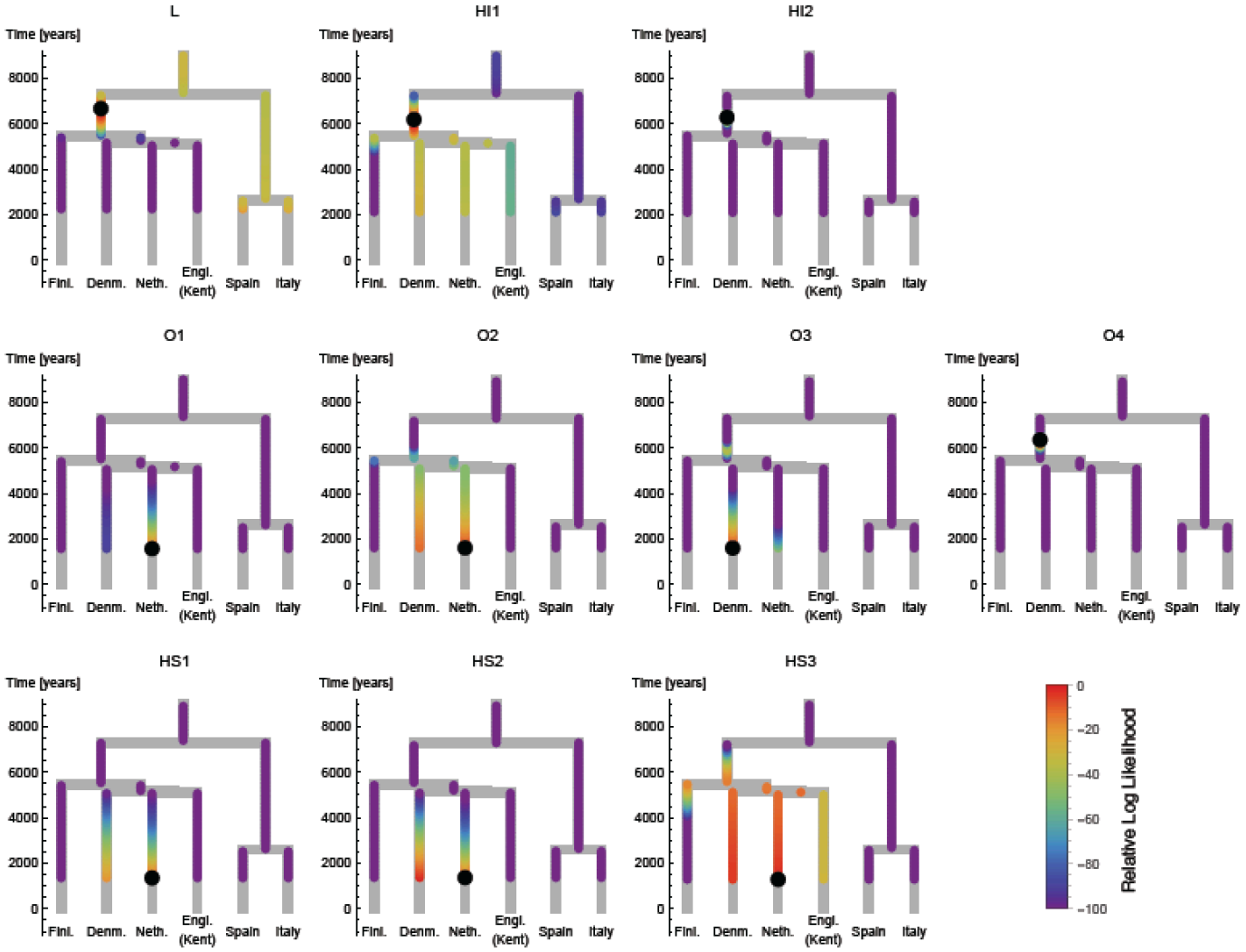
European Tree (Fits)





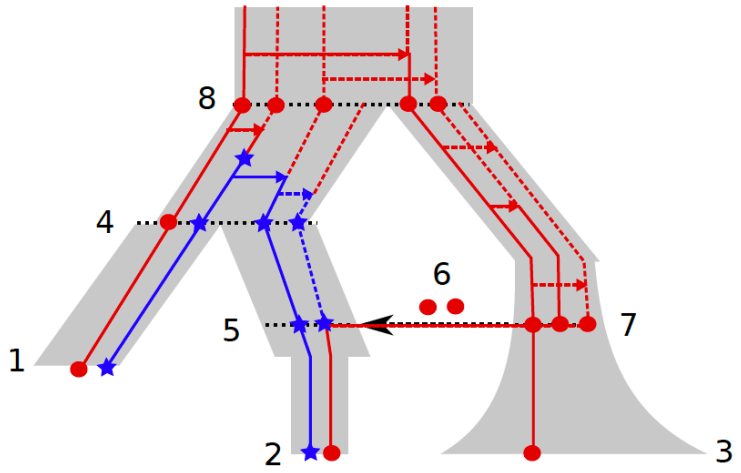


Placing ancient samples on the tree



- Plots show the likelihood for merging the population N=1 sample onto the tree as a heat map

More direct calculation of the likelihood of the joint site frequency spectrum with **momi**



- Complexity of ancestral allele state is reduced by using Moran model
- Use Automatic Differentiation to calculate gradients to maximise likelihood over demography with (limited) gene flow

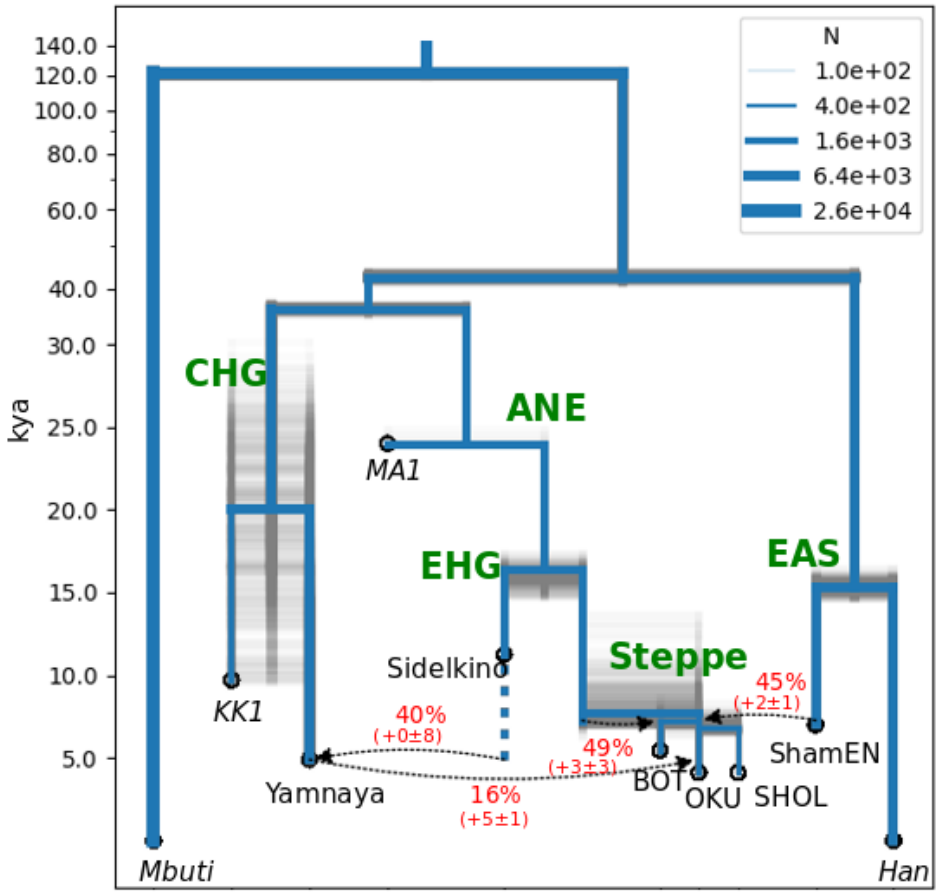
momi

Compute SFS using Moran model & Bayesian graph

Jack Kamm...Durbin, Song (2019)

Momi applied to central Asian data

- Include ancient samples
  - Condition ascertainment on modern/deep samples
    - Total branch length on these
  - Random allele sampling for low coverage samples
- Estimates split times
- Bootstrap for confidence intervals
  - But beware model misspecification



de Barros Damgaard, Martiniano, Kamm …. Durbin, Willerslev (2018)

Momi calculations

- To calculate $P(x_1,x_2,x_3,\ldots)$
  - Set leaves to $\text{Indicator}(x_i)$, e.g. $[0,0,1,0\ldots0]$ for $x_i=2$
  - Propagate likelihoods up tree ("tree-peeling")
- Can correspondingly calculate the expectation of any multi-linear function of allele counts
  - $\mathbb{E}[f_1(x_1)f_2(x_2)f_3(x_3)\ldots]$
- by setting leaf i to $[f_i(0), f_i(1), \ldots, f_i(n_1)]$
  - Works because propagation is linear

Examples

- Total branch length $\propto$ chance of any mutation
  - $f_i(j) = 1$, vector is $[1,1,1\ldots1]$
- TMRCA for pop i (i arbitrary unless ancient model)
  - $f_i(j) = j/n_i$, vector is $[0,0.2,0.4,0.6,0.8,1]$ for n_i=5
  - $f_k(j) = 1$, $k \neq i$
- $f_3 = \mathbb{E}[(X_1\text{-}X_3)(X_2\text{-}X_3)]$, $f_4 = \mathbb{E}[(X_1\text{-}X_2)(X_3\text{-}X_4)]$
  - Requires terms such as $\mathbb{E}[X_1X_2]$ for which
  - $f_1(j) = j/n_1$, $f_2(j) = j/n_2$, $f_k(j) = 1$, $k > 2$
- Also numerators, denominators of F_{ST}, Tajima's D

Summary

- PSMC(‘) estimates demography from a single pair of sequences
  - Sample size is in length not number
  - Quite a clean model
  - Major issue is population structure
- MSMC, MSMC2, SMC++ use additional samples to get at more recent times
- RareCoal/Momi use coalescent modelling of the SFS on more samples to estimate trees
  - With limited modelled gene flow for Momi

Activity on Monday afternoon

Experimental design

- (Sequence) data collection costs money
- We always need to make decisions in how to sample and sequence
  - Number of samples
  - Number of populations
  - Depth of sequencing
  - Whole Genome Shotgun or RADseq or Exomes…
- **Population sequencing** and **Genome assembly**

1000 Genomes Project

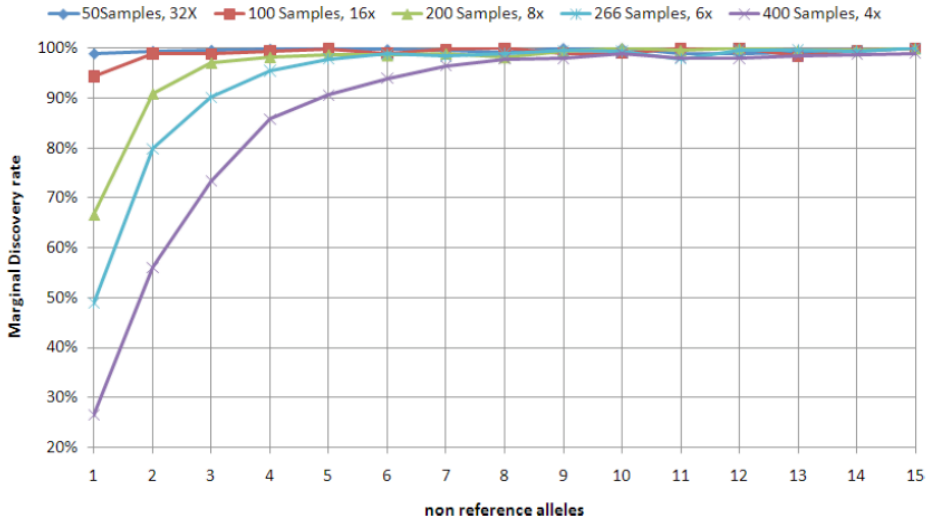
- Pilot (a **very** long time ago!)
  - 2 trios at high depth 30x
    - Phasing, accurate single-sample genotype calling, mutation rates
  - 3 populations x 60 samples at low depth 2-4x + exomes
- Main project
  - 26 populations of ~100 (2504 total) at 6-8x (+exomes)
  - (150 trios at high depth – but who remembers them?)

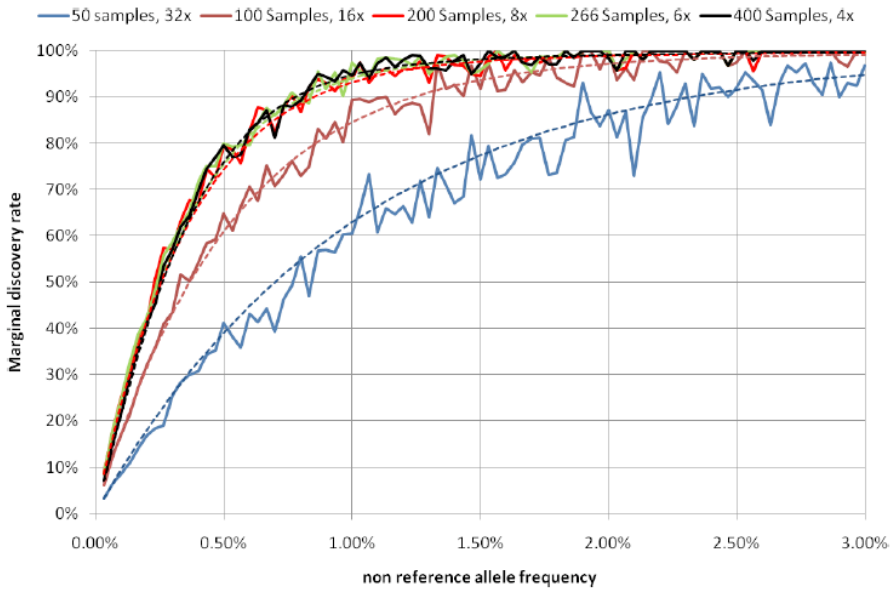
Malawi cichlid sequencing

- Phase 1
  - Three trios at 30x: mutation rate estimation, controls
  - ~70 species at 15-20x, additional samples for some at 8-12x
- Phase 2
  - 7 sets of 20 at 15x
  - More species
  - Some sets of 24 or 48 to address specific questions
- Massoko GWAS (Turner)
  - 200 samples at 4x + 100 samples for replication
  - Increased to ~600 samples more recently

Low coverage sequencing strategy

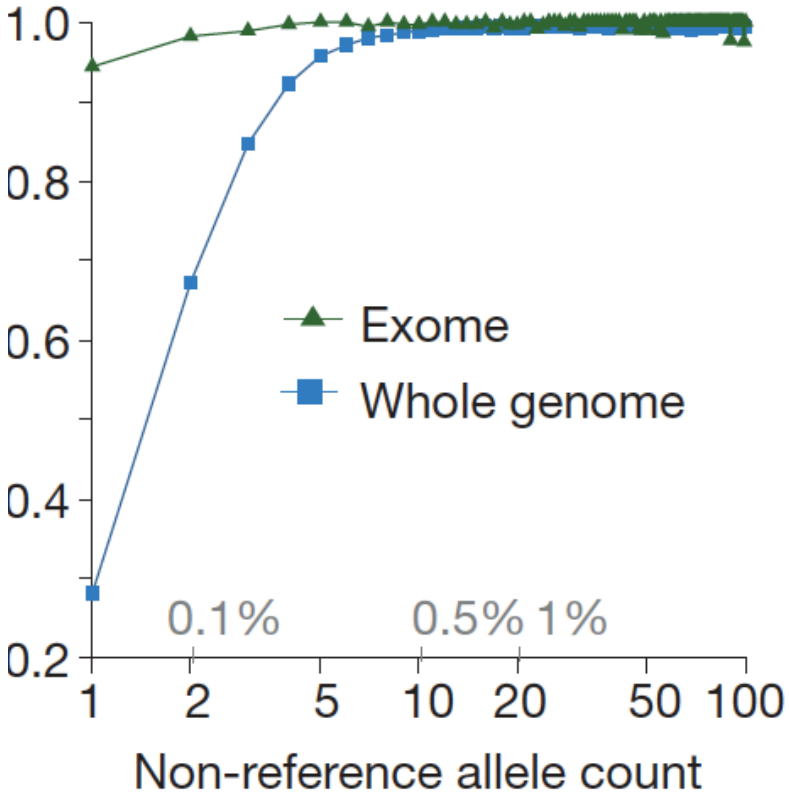
- Typically one needs to sequence at ~30x depth to find (almost) all variants in a sample
- To find low frequency variants we want to sequence many samples
- Spread sequence across more samples



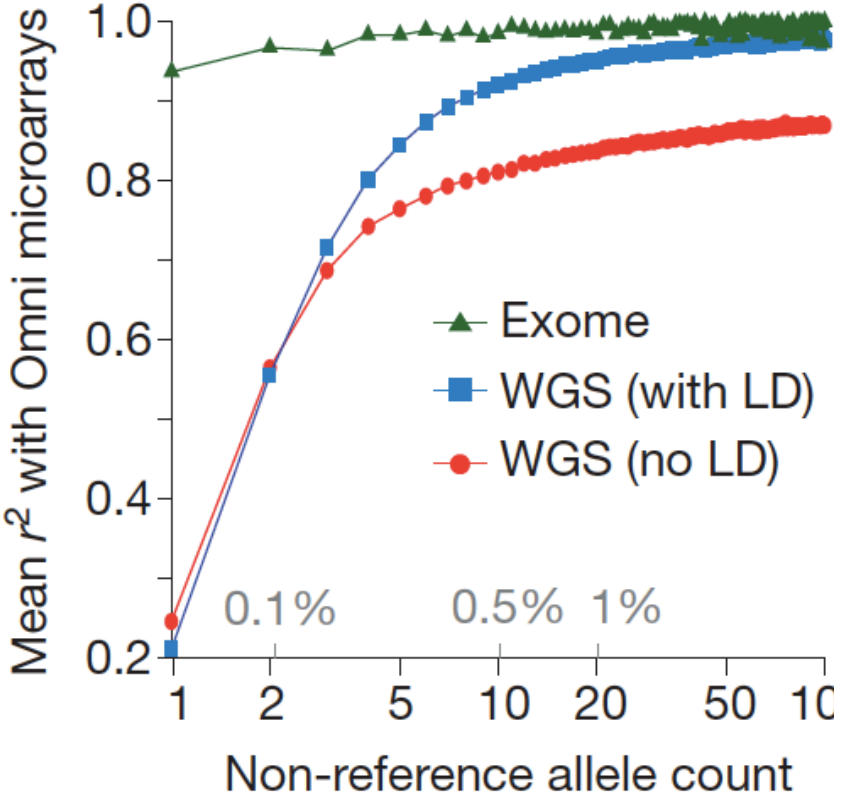


Phase 1 power and genotyping accuracy

SNP detection



Genotyping accuracy



Hyun Min-Kang (U Michigan)

Calling from low coverage sequence

- Multi-sample call **sites** with samtools or GATK
- Obtain **genotype likelihoods** at each site in each same (also samtools or GATK)
  - Likelihood = P(data | genotype)
- Combine in an imputation framework using BEAGLE (Browning), or MINIMAC (Abecasis), or perhaps STITCH (Mott)?
- Phase using SHAPEIT2 (Marchini) or EAGLE2 (Loh)

Sequencing depth

- 30x is standard for near-complete accuracy
  - Sufficient to estimate mutation rates in trios (need several trios for many species)
- 15x is good enough for SNPs (~97%), not quite so good for indels (perhaps 90-95%)
- 4-8x gives good low coverage imputation as in previous slides
- People have used 1-2x, but this is hard work…
- 60x + is necessary for subclonal structure, e.g. cancer, high ploidy
- In a cross, sequence the founders to high depth, and the F2/F3 to low depth (1x or less is fine) and impute using STITCH or other Richard Mott tools

Genome reference assembly

- To work on a species it makes a big difference to have a good reference genome
- Historically this was difficult/expensive, but recent technical developments have dramatically improved quality and decreased cost
  - A trend that will continue
- Primary reason is long single molecule reads
  - Also other long range data: Hi-C, linked read clouds…

The era of sequencing genomes

- DNA sequencing is a transformative technology for biology
- We are still in the middle of its development and application
- Phase 1: reference genomes for key organisms
  - Sanger technology through a gel
- Phase 2: population resequencing and genomic assay by sequence
  - Cluster technology on a surface (Illumina) – short reads
- Phase 3: *ab initio* sequencing of arbitrary genomes, nucleic acids
  - Single molecule technology (PacBio, Nanopore, …) – >100x longer reads

We are just at the start of phase 3 – will dominate next decade

494 Vertebrate genome assemblies (55 human)

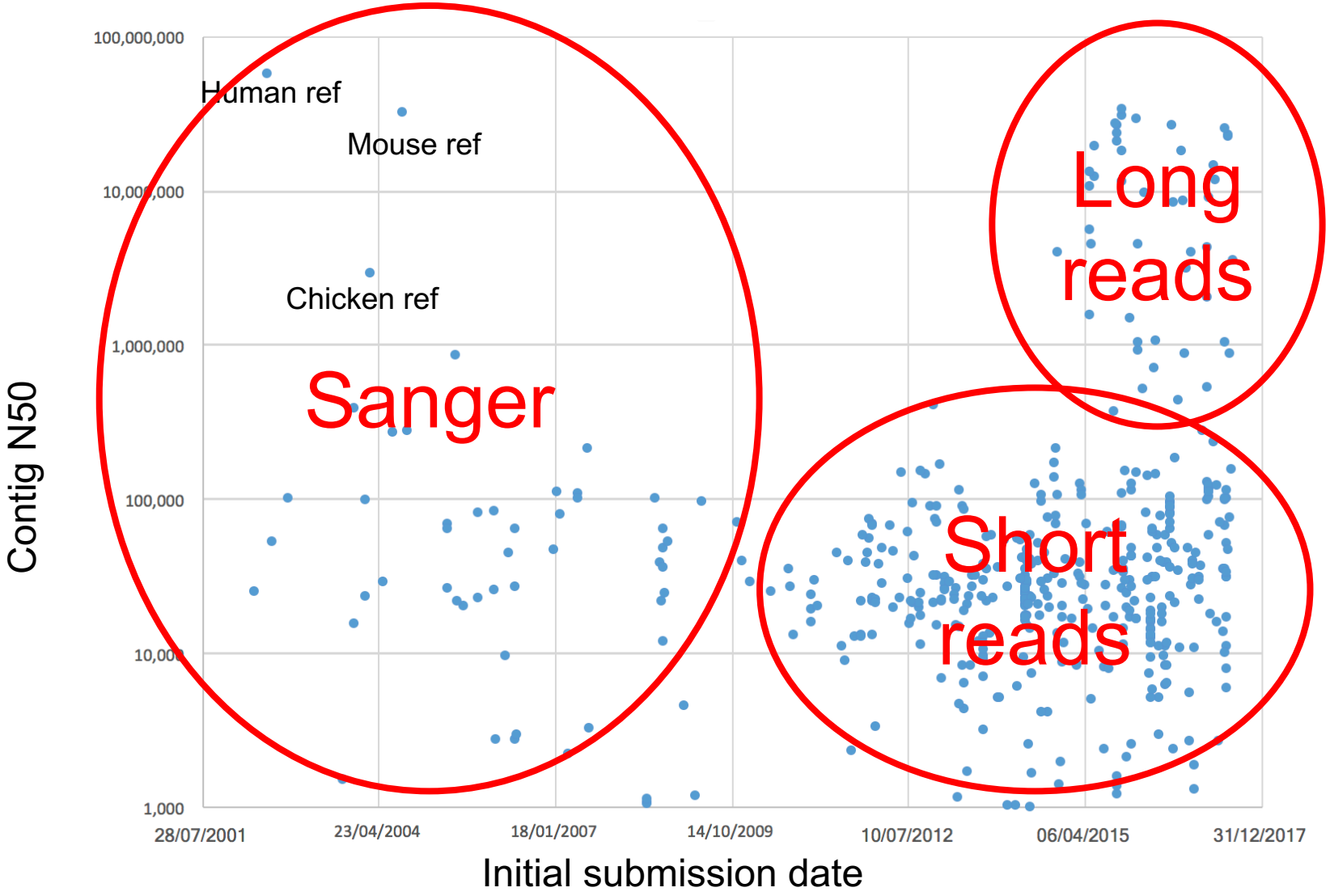
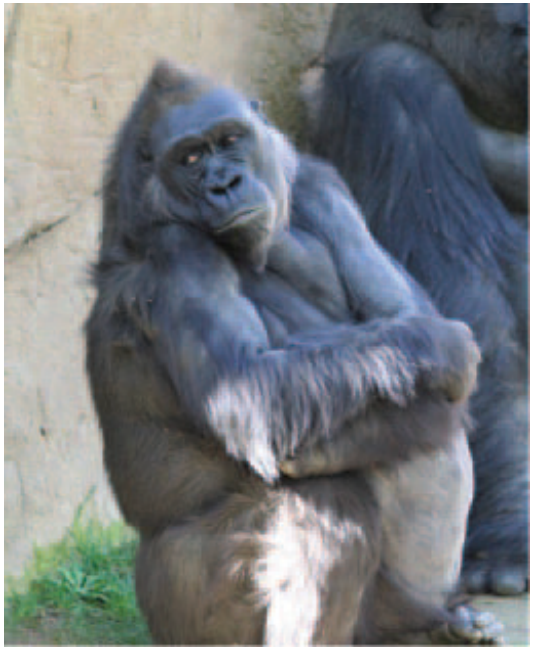
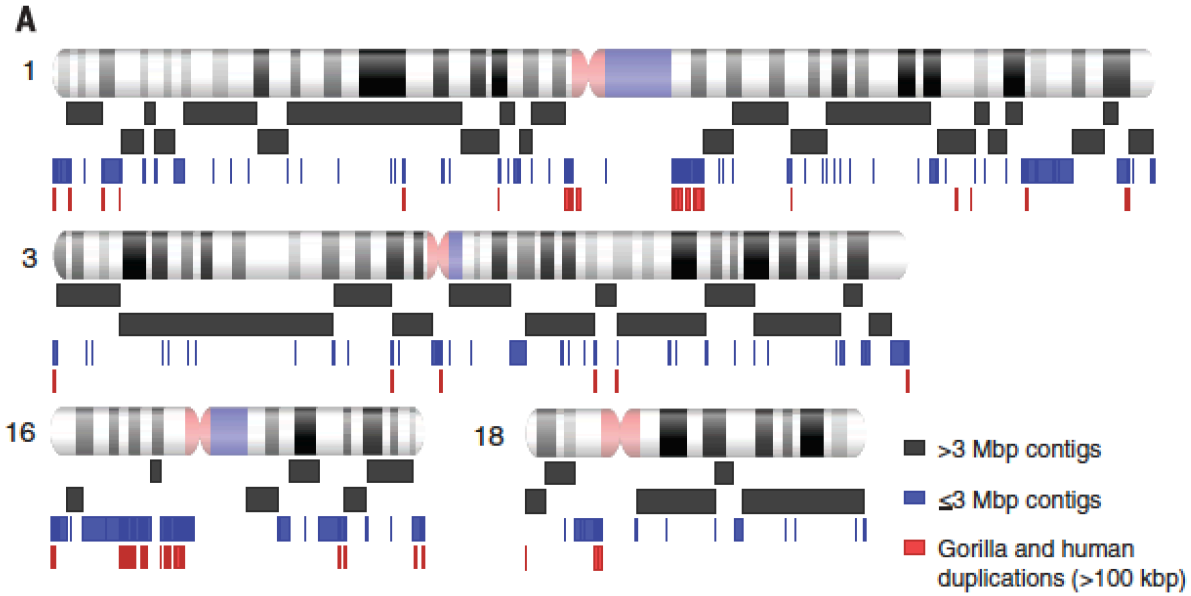


Illustration: Gorilla genome





Gordon et al. Science 2016

2010 Sanger assembly

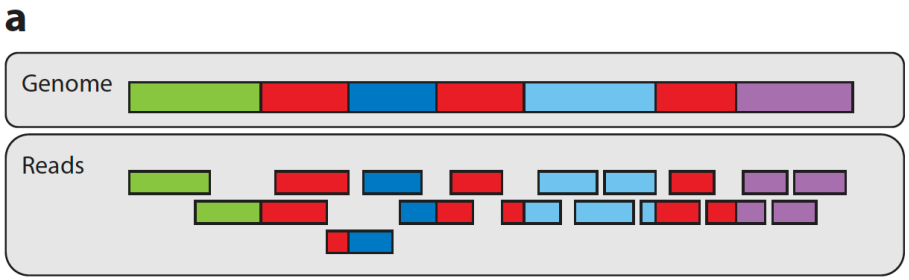
- Illumina + 3x Sanger
- Fosmids and mate pairs
- 12kb contig N50

2016 75x PacBio assembly

- 800x increase in N50
- 200Mb new DNA, +many fixes
- 3700 → 220 incomplete genes

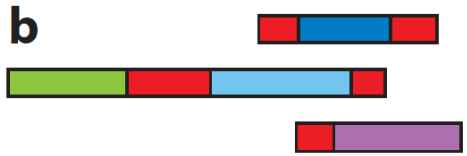
Main problem in assembly is repeats

Typical genome (e.g. human) is ~50% transposons and segmental duplications

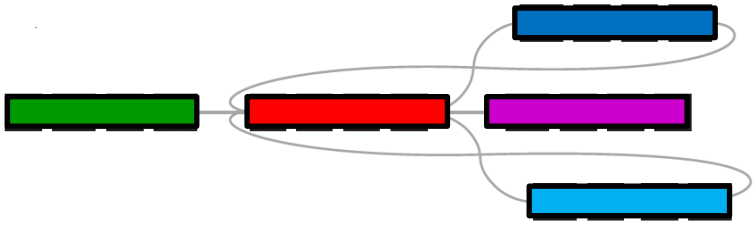


Where do the red reads go?

Greedy approach can make errors



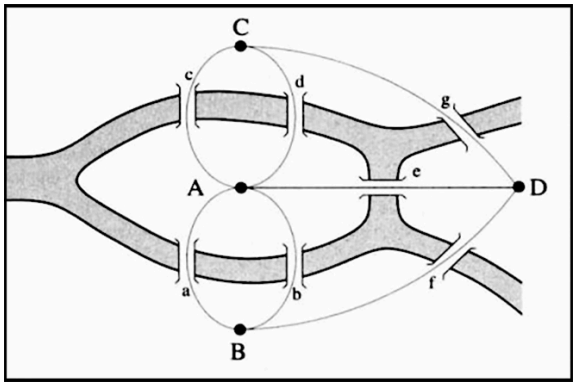
Generate contigs and possible links



Simpson JT, Pop M. The Theory and Practice of Genome Sequence Assembly. Annu Rev Genomics Hum Genet. 2015.

Scaffolding contigs

In principle the connectivity and (modified) **Euler path** requirement constrains order and orientation. But this is little used in modern assemblers.



The bridges of Königsberg

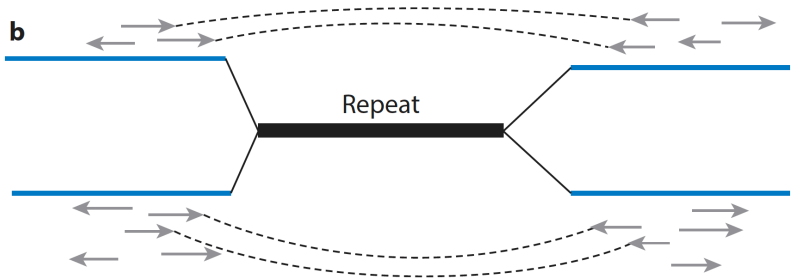
Math Stack Exchange

In practice we use additional data types to order and orient contigs into **scaffolds**

- Read pairs, genetic/physical maps, read "clouds" from long molecules, Hi-C

Read pairs that span repeats can pull them apart:

- Standard Illumina inserts ~500bp
- Large insert mate pairs 2-10kb
- Fosmid/BAC clone pairs 40-150kb

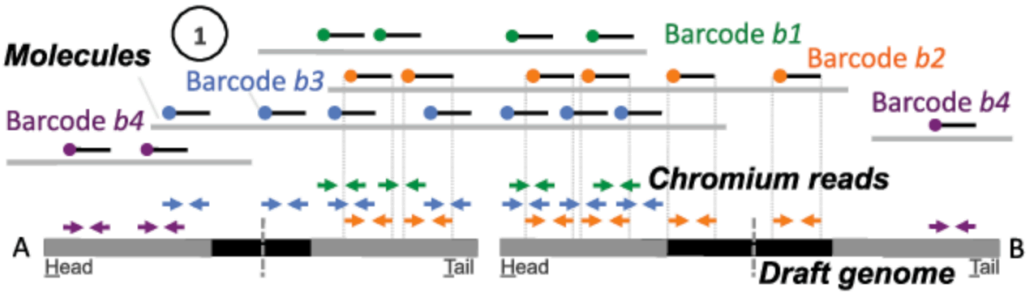


Simpson JT, Pop M. Annu Rev Genomics Hum Genet. 2015.

Read clouds and optical maps

Longer range “read clouds” e.g. from 10X Genomics Chromium, can bridge bigger repeats/gaps

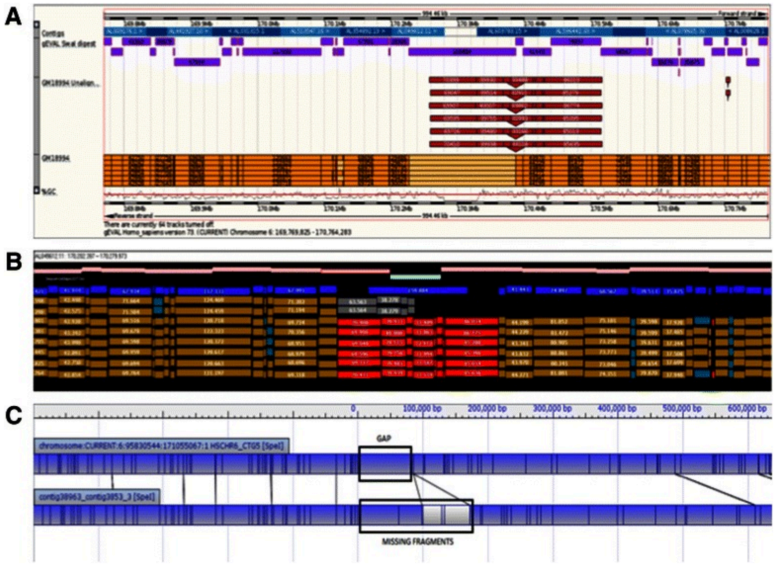
- Multiple reads from a long ~100kb template sharing a barcode



Yeo S, Coombe L, Warren RL, Chu J, Birol I. ARCS: scaffolding genome drafts with linked reads. *Bioinformatics*. 2018.

Restriction digest maps give approximate spacing of restriction sites (short sequence motifs), skipping over repeats

- Now scaled up by dedicated optical imaging machines (BioNano)



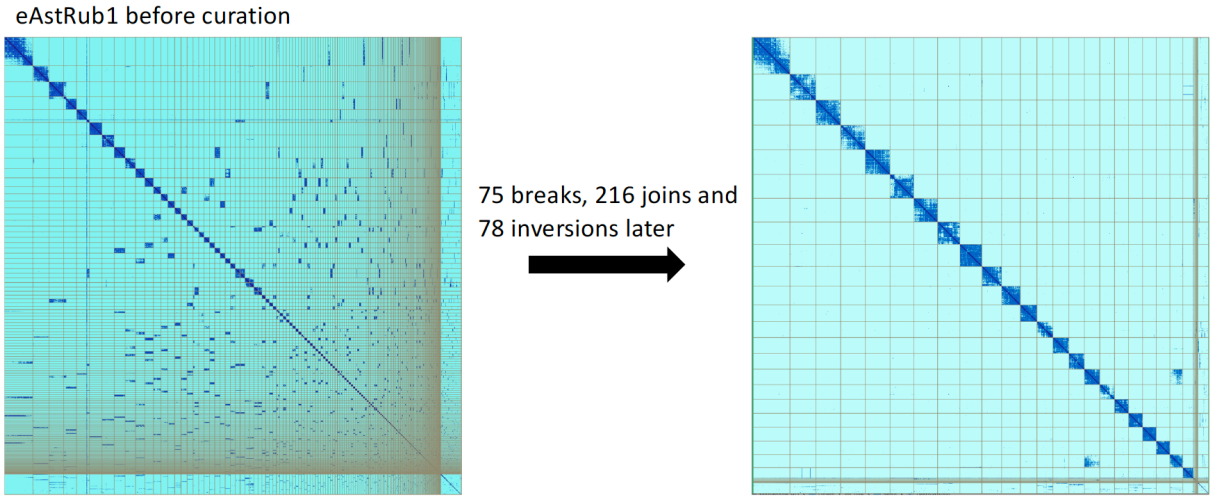
Howe K, Wood JMD. Using optical mapping data for the improvement of vertebrate genome assemblies. *Gigascience*. 2015.

Genetic maps and Hi-C give chromosomal scale

Genetic linkage maps are naturally on the scale of chromosomes.
Although typically low resolution (e.g. a few thousand markers per genome) they can place large scaffolds and contigs onto chromosomes by matching marker sequences.

Hi-C “proximity ligation” is used to look at local chromatin organisation, but globally most Hi-C links are on the same chromosome, and more are close than distant

- Automated tools (e.g. SALSA)
- Visualisers (e.g. Hi-Glass)
- Editors



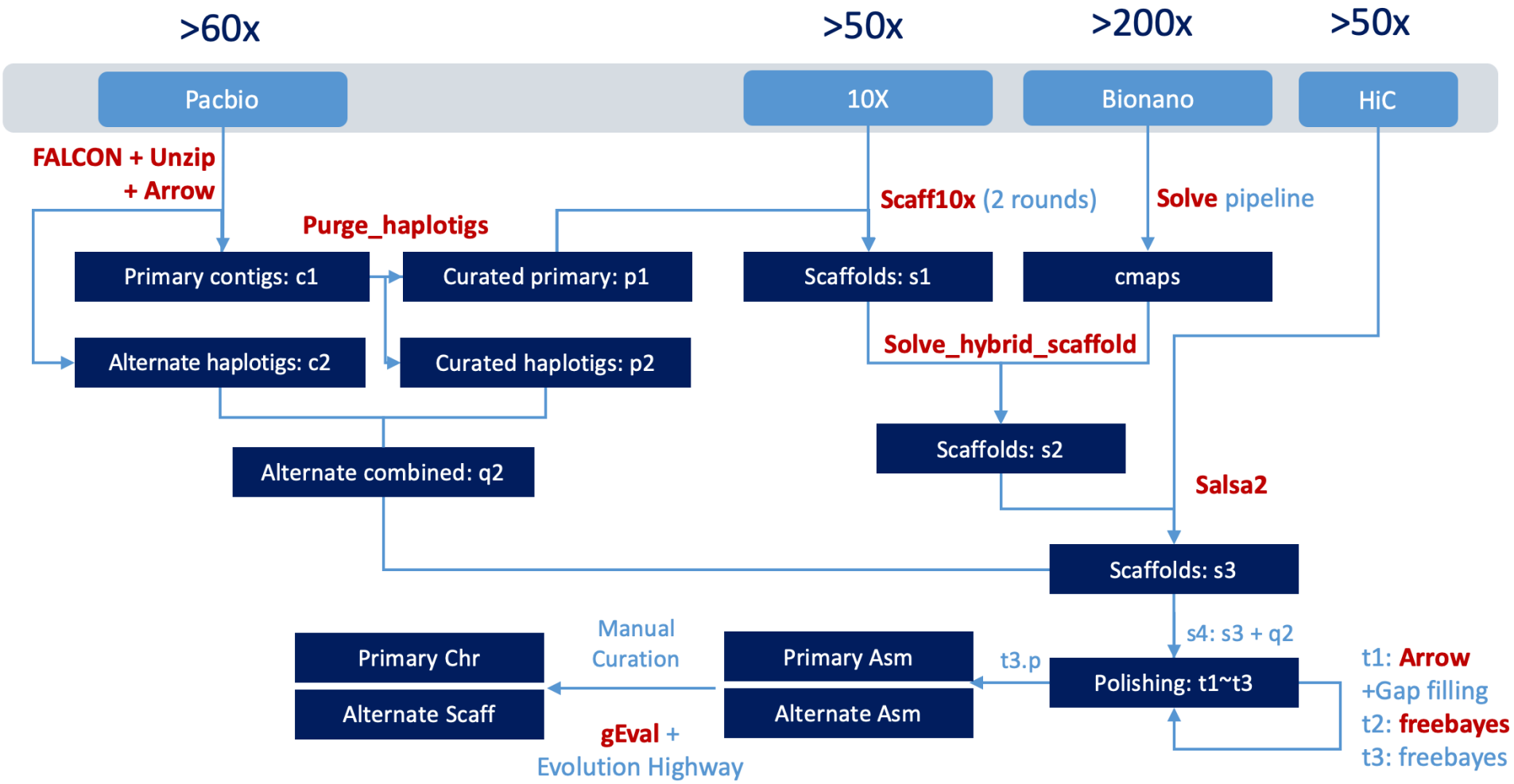
So, what recipe to recommend?

- Long reads
  - 60x PacBio CLR
    - ~10% error, ~30kb N50, ~120Gb / SMRT cell (~€1k)
  - Or 25x PacBio CCS
    - ~0.1% error, ~15-20kb N50, ~20-25Gb / SMRT cell
  - Or 80x Oxford Nanopore
    - ~10% error, 10kb-100kb+ N50, ~50-150Gb / Promethion
- Hi-C: this is the most useful for scaffolding
  - Dovetail (support), Arima, Phase, Qiagen, in-house
  - BioNano or 10x
- “Polishing” – Illumina or CCS ideally

Example thorough assembly pipeline

VGP standard 1.5 pipeline

https://github.com/VGP/vgp-assembly

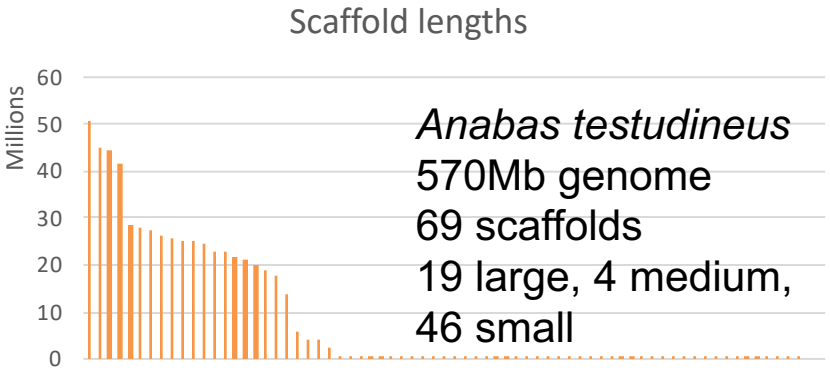


How good is an assembly?

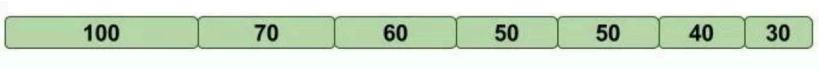
- Length / contiguity
  - The whole point of sequence assembly is to reconstruct long sequences
  - So, *all other things being equal*, longer is better
  - Measurement: N50 and its relatives
- Accuracy
  - Ultimately this should be more important
  - Base pair accuracy: qv score
  - Misassembly: alignment consistency
- Completeness
  - Raw data coverage
  - BUSCO score for complete gene representation

N50 and its relatives

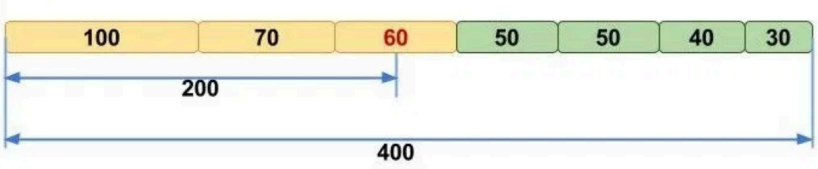
- How to characterize the typical fragment size of an assembly?
- Often there are very many small fragments, contributing little
- N50 asks “what is the size of pieces that contribute half the data?”
- Contigs, or scaffolds, or …



Mean scaffold size 8.3Mb
N50 scaffold size 26.1Mb



1a. Contigs, sorted according to their lengths.



1b. Calculation of N50 using sorted contigs.

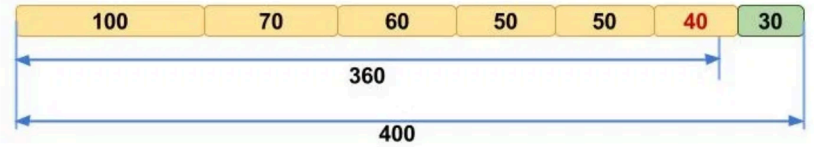


Fig. 2. Example of calculating N90 for the same set of seven contigs. Here N90 equals 40 kbp.

Introduced in the 2001 Human Genome Project paper (Nature, 2001)

https://www.molecularecologist.com/2017/03/whats-n50/ and follow-on posts

Problems with N50 and proposed solutions

1. It depends on total length
   - Different assemblies of the same species may have different lengths
   - Filtering out small “junk” changes the N50
   - Solution: NG50 – divide by fixed genome size
   - Good for comparison of methods – can use a single size estimate

Fig. 5. Example assembly of a 500 kbp genome consisting of seven contigs. NG50 = 50 kbp, N50 = 60 kbp.

2. It doesn’t pay attention to accuracy
   - Assemblers can gamble and join together contigs when not certain, making false joins to increase length
   - Solution: NA50 – count aligned lengths
   - Needs a (correct!) reference

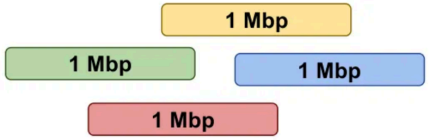


Fig 6a. Correct Assembly with N50 = 1 Mbp and 0 misassemblies.

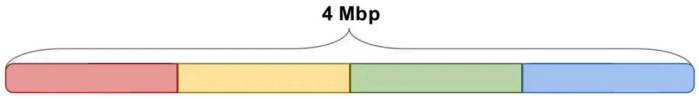


Fig 6b. Incorrect Assembly obtained by contig merging of the correct Assembly. N50 = 4 Mbp, 3 misassemblies.

3. Combine the two: NGA50
4. Etc…

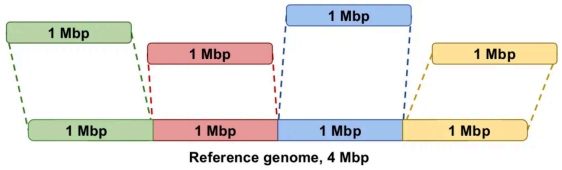


Fig 6c. Alignment of the Correct Assembly to the reference genome provides four alignment blocks, 1 Mbp each, resulting in NA50 = 1 Mbp.

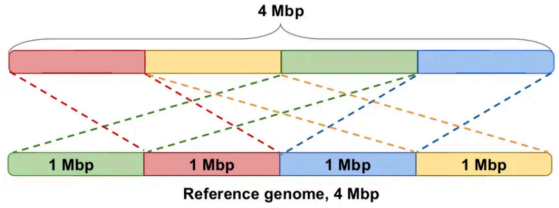


Fig 6d. Alignment of the Incorrect Assembly to the reference genome provides the same four alignment blocks of 1 Mbp each, resulting in NA50 = 1 Mbp.

https://www.molecularecologist.com/2017/03/whats-n50/ and follow-on posts

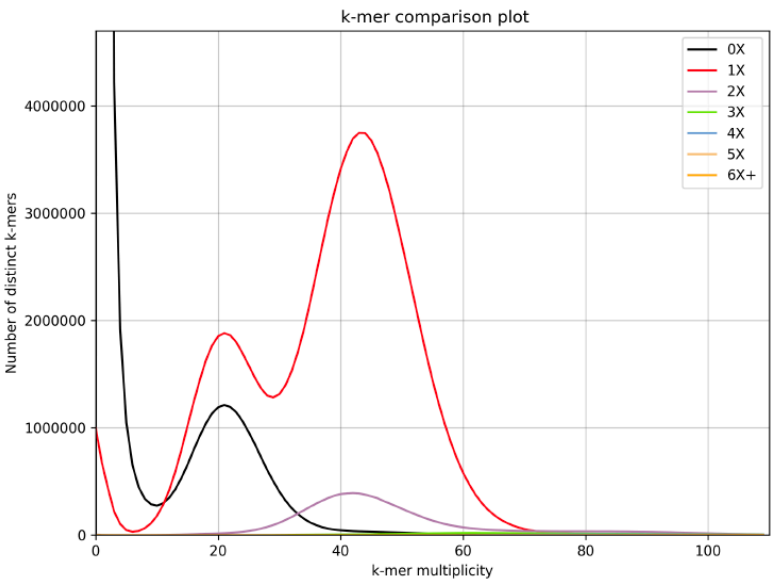
QV Quality score (Phred score)

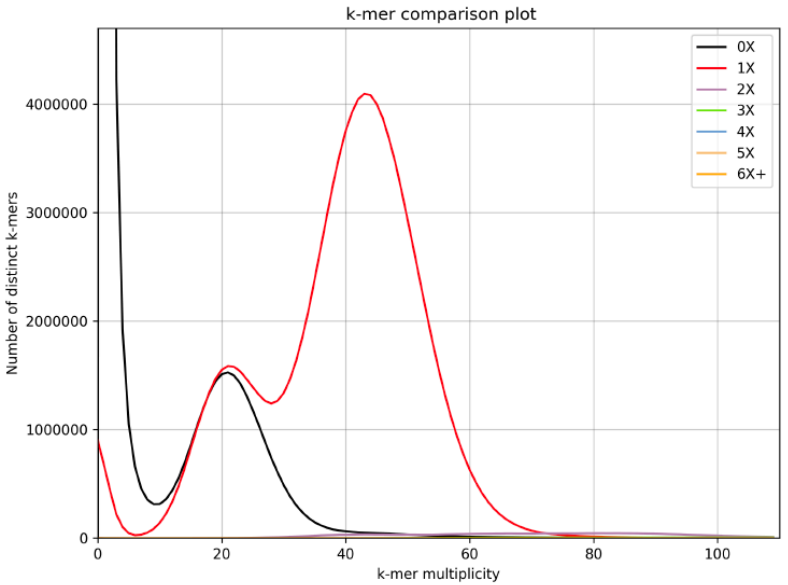
- Base pair error rate
  - 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, …
  - Cumbersome to write, so use log space
  - Q20, Q30, Q40, Q50
- Formally $Q = -10 \log_{10} p(\text{error})$
  - Like decibels for sound
- Can apply to whole sequences, or to error estimates per base
  - Introduced by Phil Green for sequencing reads

Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 1998

Data consistency

- Are the primary data consistent with the assembly, and vice versa?
  - But beware data errors, and circular reasoning…
- E.g. depth of coverage



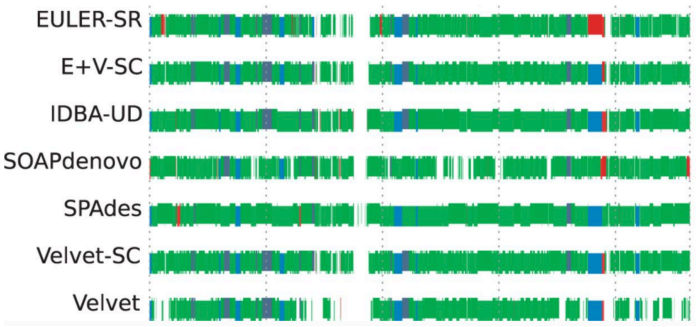


Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *bioRxiv*. August 2019:729962. doi:10.1101/729962.

Data consistency

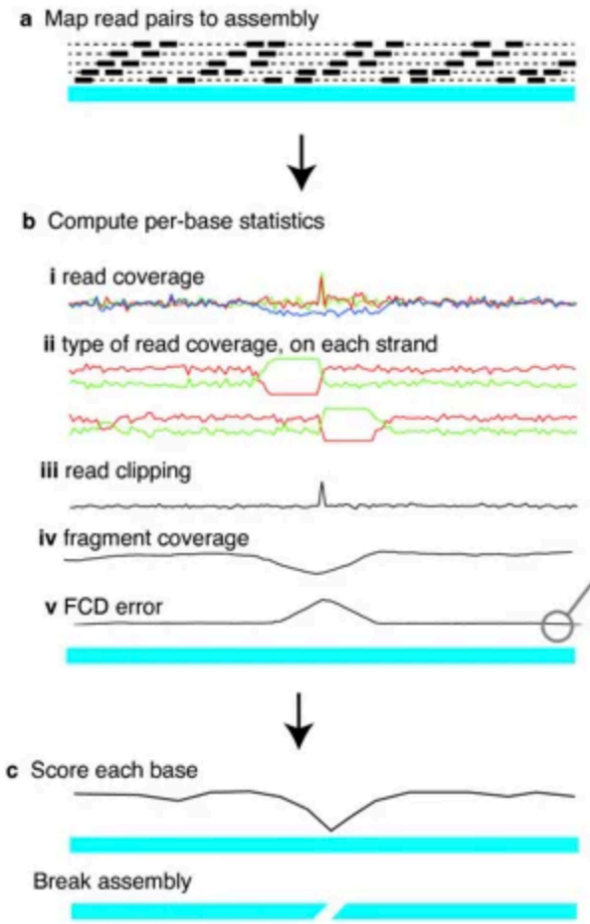
- QUAST: http://bioinf.spbau.ru/quast
  - Many measures
  - Including cross-assembly consistency

7 *E. coli* assemblies from the same data assessed in QUAST



- REAPR: https://www.sanger.ac.uk/science/tools/reapr
  - Read pair consistency
  - Also alignment orientation, clipping…

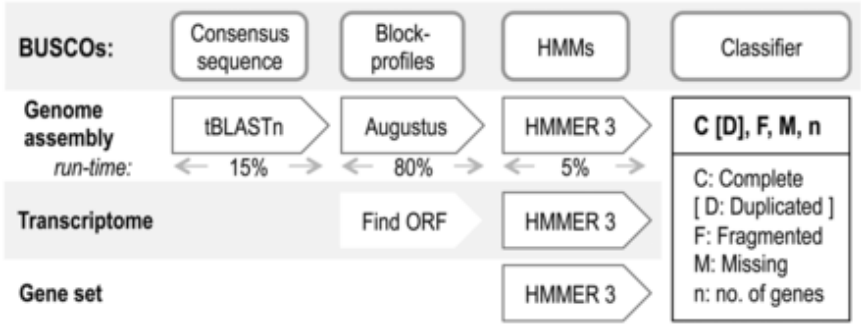
REAPR workflow

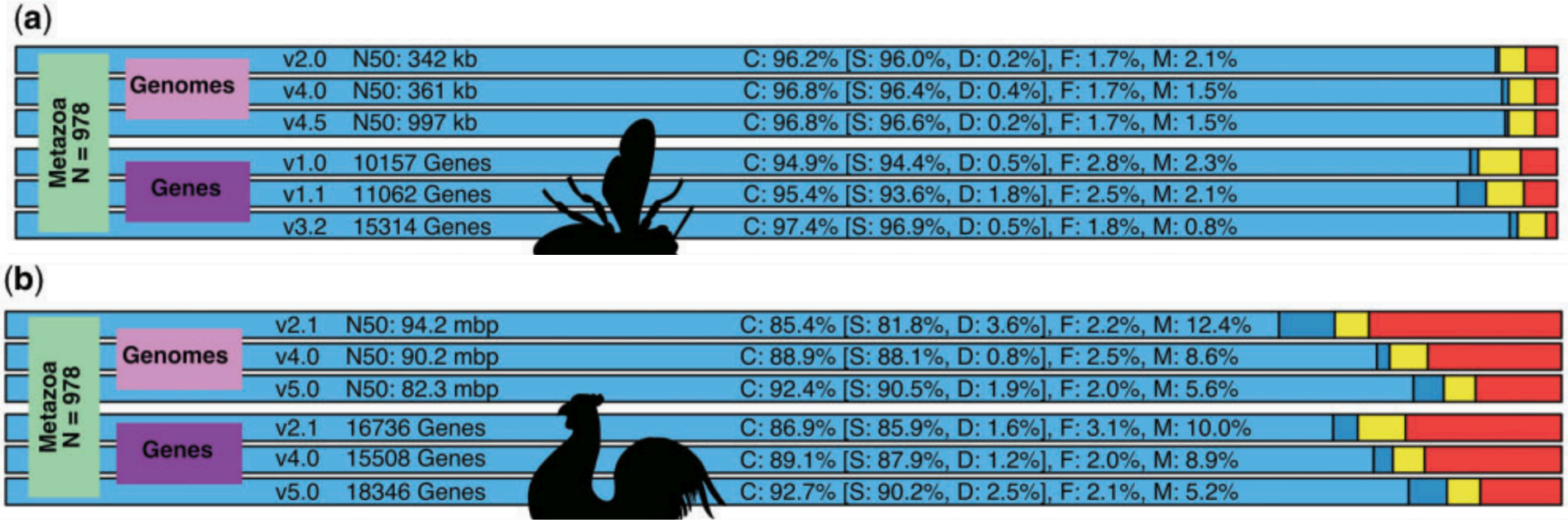


BUSCO score: completeness (and accuracy)

Benchmarking Universal Single-Copy Orthologs

- Start from set of conserved genes
- Find them in the assembly
- Assess whether they are Complete (C), Single copy (S) or Duplicated (D), Fragmented (F) or Missing (M)





Simão FA et al.. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015, Waterhouse et al. *Mol. Biol. Evol.* 2017

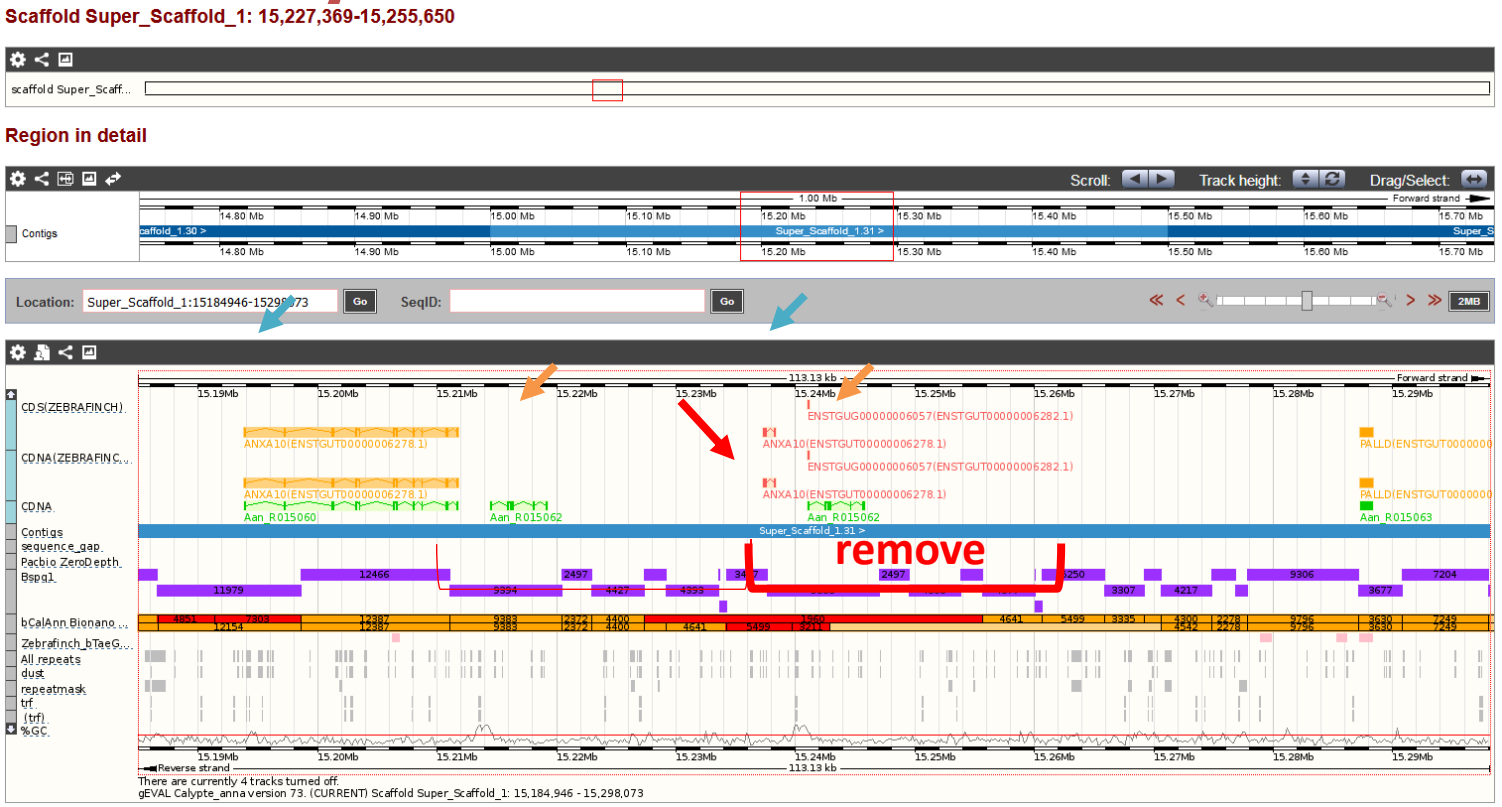
Problems with heterozygosity

- Heterozygosity levels vary
  - 0.1-1% for vertebrates
  - 1-5 % typical for insects, 0-5% for plants
  - up to 15% (?) for some invertebrates (*Caenorhabditis. brenneri*)

- Almost all current assemblers aim to (initially) squash haplotypes
  - They can over-collapse, losing paralogs, or under-collapse, putting haplotigs into the main assembly
  - Structural heterozygosity gives rise to misassembly
  - Major problems above ~1% heterozygosity, but problems at all levels…
- The only good historical genomes are haploid/clonal, or assembled from cloned fragments

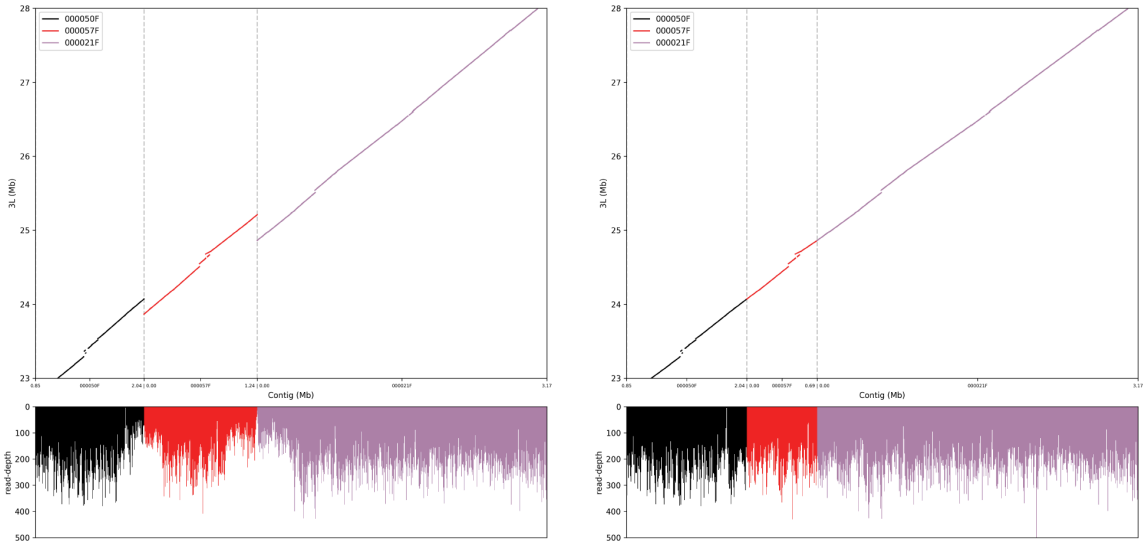
Even in good new assemblies, most errors and contig breaks are caused by heterozygosity

Internal duplication in Hummingbird assembly



Kirsten Howe and Sanger Institute gEVAL team

Strategy 1: Remove the duplications



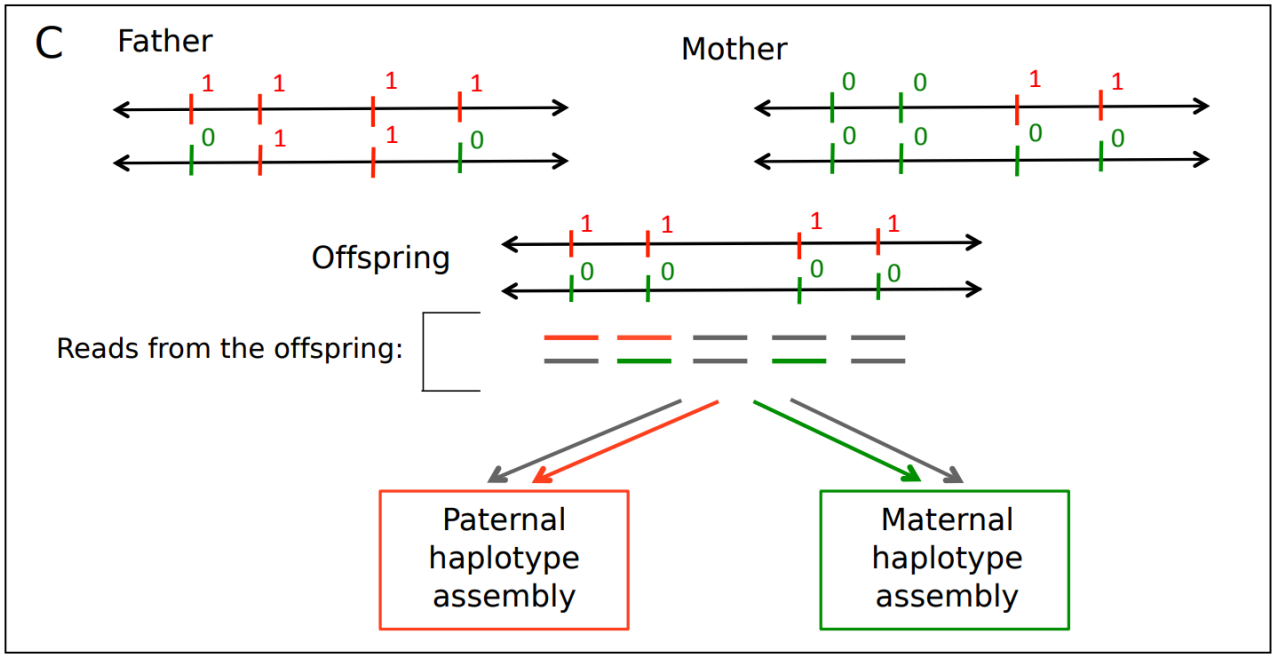
	BUSCO scores[1] (%)					Assembly size (Mb)	Num. Contigs	Ctg N50 (Mb)
	C	C(S)	C(D)	F	M			
At-FU[2]	**98.1**	91.9	6.2	**0.3**	**1.6**	140	172	7.96
At-PD	97.7	**96.6**	**1.1**	0.6	1.7	121	95	7.98
Mm-FU	**95.8**	79.0	16.8	**2.0**	**2.2**	1250	1290	2.63
Mm-PD	94.4	90.9	3.5	2.7	2.9	838	559	3.41
Mm-PDS	94.7	**91.3**	**3.4**	2.6	2.7	840	**322**	**14.50**

Dengfeng Guan et al. *Biorxiv* (2019)

Strategy 2: Use trios (pedigrees)

trio-sga: facilitating *de novo* assembly of highly heterozygous genomes with parent-child trios

Milan Malinsky[1,2,*], Jared T. Simpson[3,4] and Richard Durbin[1,*]



Idea: Use parental unique 31mers to assign offspring reads to haplotypes, and assemble them separately.

In practice limited by short Illumina reads

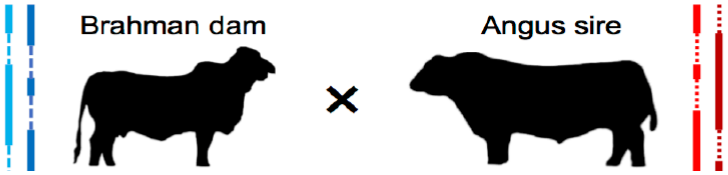
Milansky et al. *Biorxiv* (2016)

Koren et al. Nature Biotechnology (2018)

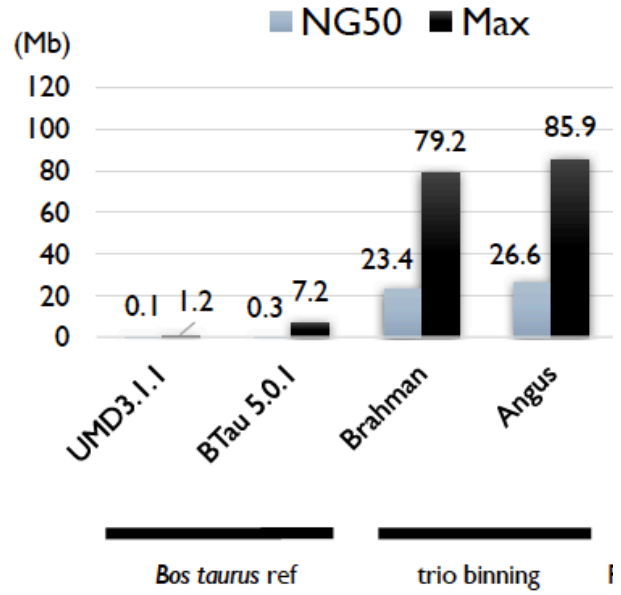
Complete assembly of parental haplotypes with trio binning

Sergey Koren, Arang Rhie, Brian P. Walenz, Alexander T. Dilthey, Derek M. Bickhart, Sarah B. Kingan, Stefan Hiendleder, John L. Williams, Timothy P.L. Smith, Adam Phillippy

- K-mer profiling of each parent (Illumina, 60x)



- K-mer profiling of the F1 (PacBio, 120x)



VGP application to zebrafish *Danio rerio* SAT strain = Tu x AB cross

Initial joint Falcon assembly 0.5Mb N50, 1.9Gb total size

fDreABH1	*Danio rerio*	zebrafish (AB strain)	5.77	1,920	8.17	1,699	37.46	1,354	QC
fDreTuH1		zebrafish (Tübingen strain)	4.06	2,382	5.78	2,112	24.84	1,370	QC

Better than current GRCz11 reference assembly!

Strategy 3: use intrinsic data

1. Build graph of all the sequence including both haplotypes, then thread two paths through it most consistent with the data (Garg, …)
2. Allocate reads into (local) bins, by kmer phasing combining 10X Genomics and HiC Illumina data, long reads (Heaton, …)
3. Separate overlaps between reads based on differences at single-copy (heterozygous) sites (Myers, …)

Shilpa Garg (Tobias Marschall) ISMB 2018 (yeast), biorxiv 2019

Proposed VGP quality metrics

Quality	Metric	Finished	High Quality Reference	Reference	Draft
Continuity	Contig (NG50)	= Chr. NG50	>10 Mb	>1 Mb	>10 kb
	Scaffolds (NG50)	= Chr. NG50	= Chr. NG50*	>10 Mb	>100 kb
	Gaps (num.)	No gaps	< 200	< 1,000	< 10,000
Correctness	Reliable blocks	= Chr. NG50	>90% of Scaffold NG50	>75% of Scaffold NG50	>50% of Scaffold NG50
	Curated	Yes	Yes	Not required	Not required
Accuracy	Basepair QV	50	40	30	30
	k-mer completeness	100% complete	>95%	>90%	>80%
Phasing	Phased block (NG50)	= Chr. NG50	>1 Mb	>100 kb	Not required
Functional completeness	Genes (ex. BUSCO)	100% complete	>95%	>90%	>80%
	Transcript mappability	TBD	TBD	TBD	TBD
Chromosome	Assigned %	100% assigned	>90% assigned	Not required	Not required
	Sex chromosomes	Present, right order, no gaps	Present, localized hom pairs	Present, at least 1 longer chr (ex. X or Z)	Fragmented
	MT Presence	1 Complete major allele	1 Complete allele	Not required	Not required