Demographic Inference

Ryan Gutenkunst
Molecular and Cellular Biology
University of Arizona

Name	Data type	Inference	Notes	Refs
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User-friendly GUI; can be computationally demanding	32
FRAPPE	Unlinked bi-allelic SNVs	Population structure, admixture	Alexander *et al.*[41] argue that convergence is not guaranteed	40
ADMIXTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Estimates the number of populations via cross-validation error	41
fastSTRUCTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Obtains variational Bayesian estimates of posterior probability distribution	42
Structurama	Unlinked multi-allelic genotypes	Population structure, admixture	Uses a Dirichlet process to estimate the number of populations	43
HAPMIX	Phased haplotypes; reference panel	Chromosome painting	Requires populations to be specified a priori	48
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	49
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Extends the fineSTRUCTURE approach to estimate unsampled ancestral populations and admixture times	7
LAMP	Phased haplotypes; reference panel	Chromosome painting	Identifies local ancestry in windows, rather than using an HMM, so is more discrete than other approaches	52
PCAdmix	Phased haplotypes	Chromosome painting, population structure	Uses PCA in small chunks followed by an HMM to estimate local ancestry	53
dadi	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Requires some Python-coding skills; applicable to up to three populations	60
Fastsimcoal2	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Can also be used to simulate data under the SMC	62,63
Treemix	Frequencies of unlinked bi-allelic SNVs	Admixture graph	Highly multimodal likelihood surface and heuristic search; redo inference from many starting points	64
fastNeutrino	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Applicable only to a single population; designed specifically for extremely large sample sizes	65
DoRIS	Lengths of IBD blocks between pairs of individuals	Demographic history	IBD must be inferred (for example, using Beagle or GERMLINE); specification of lower cut-off minimizes false-negative IBD tracts	71,72
IBS tract inference	Lengths of IBS blocks between pairs of individuals	Demographic	IBS can easily be confounded by missing data and/or sequencing errors	76
PSMC	Diploid genotypes from one individual	Demographic history	Best used in MSMC's PSMC mode, which uses the SMC to more accurately model recombination than the original PSMC; applicable to a single population	78
MSMC	Whole genome, phased haplotypes	Demographic history	Requires large amounts of RAM; cross-coalescence rate should not be interpreted as migration rate	82
CoalHMM	Whole genome, phased haplotypes	Demographic history	Multiple applications, including inference of population sizes, migration rates and incomplete lineage sorting	83–87
diCal	Medium-length, phased haplotypes	Demographic history	Uses shorter sequences than MSMC, but can be applied to multiple individuals in complex demographic models; infers explicit population genetic parameters for migration rates	89,92
LAMARC	Short, phased haplotypes	Demographic history	Requires Monte Carlo sampling of coalescent genealogies; very flexible	93
BEAST	Short, phased haplotypes	Species trees, effective population sizes	Used mainly as a method of phylogenetic inference. Can also infer population size history	94
MCMCcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP[131]	95
G-PhoCS	Short, (un)phased haplotypes	Demographic history	Incorporates migration into the MCMCcoal framework. Averages over unphased haplotypes	96
Exact likelihoods using generating functions	Short, phased haplotypes	Demographic history	Implemented in Mathematica; applicable only to specific classes of multi-population models	97,98

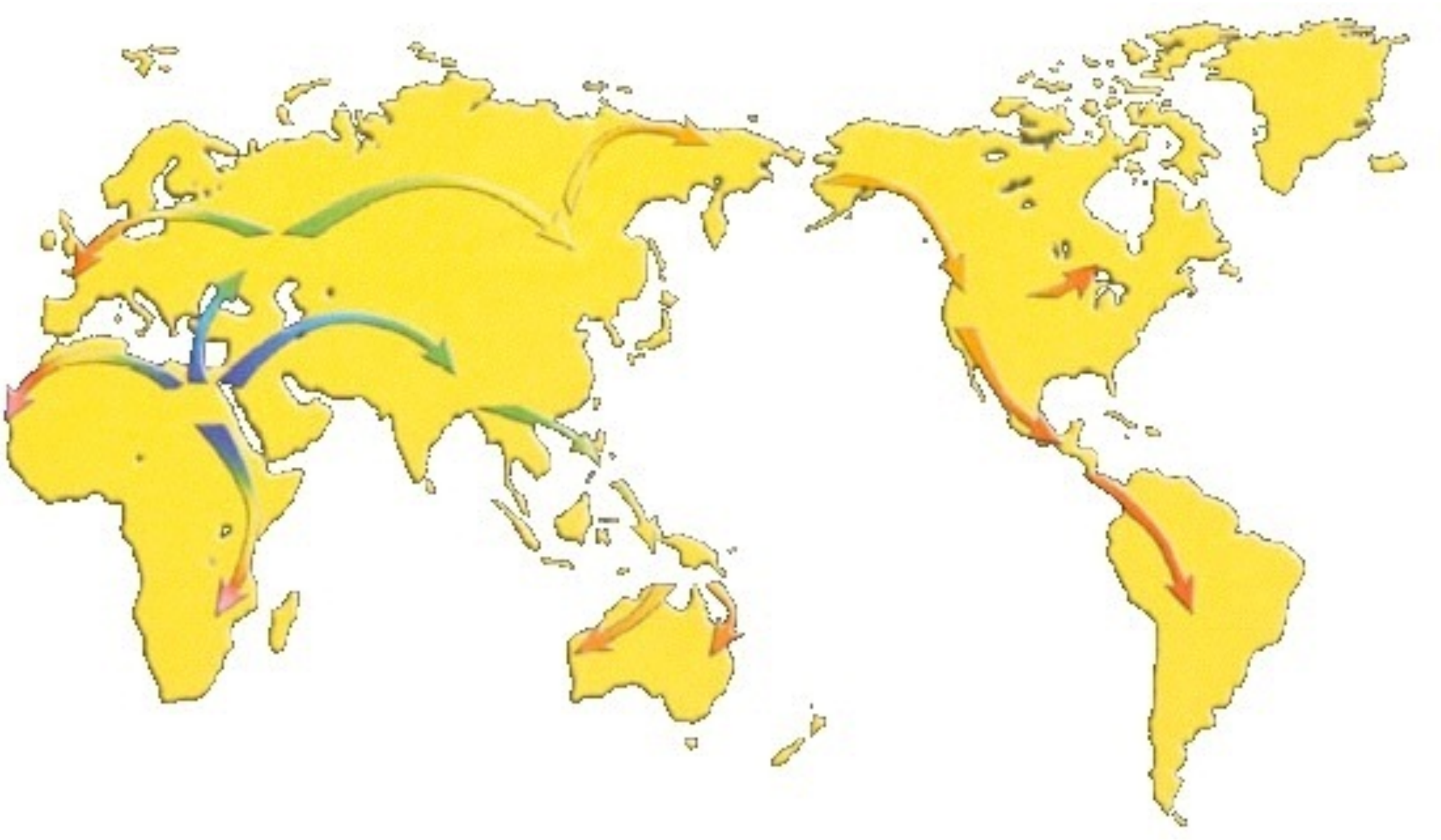
Why model demographic history?

- Essentially, all models are wrong,
  but some are useful.
  — George Box

Name	Data type	Inference	Notes	Refs
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User-friendly GUI; can be computationally demanding	32
FRAPPE	Unlinked bi-allelic SNVs	Population structure, admixture	Alexander *et al.*[41] argue that convergence is not guaranteed	40
ADMIXTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Estimates the number of populations via cross-validation error	41
fastSTRUCTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Obtains variational Bayesian estimates of posterior probability distribution	42
Structurama	Unlinked multi-allelic genotypes	Population structure, admixture	Uses a Dirichlet process to estimate the number of populations	43
HAPMIX	Phased haplotypes; reference panel	Chromosome painting	Requires populations to be specified a priori	48
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	49
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Extends the fineSTRUCTURE approach to estimate unsampled ancestral populations and admixture times	7
LAMP	Phased haplotypes; reference panel	Chromosome painting	Identifies local ancestry in windows, rather than using an HMM, so is more discrete than other approaches	52
PCAdmix	Phased haplotypes	Chromosome painting, population structure	Uses PCA in small chunks followed by an HMM to estimate local ancestry	53
dadi	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Requires some Python-coding skills; applicable to up to three populations	60
Fastsimcoal2	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Can also be used to simulate data under the SMC	62,63
Treemix	Frequencies of unlinked bi-allelic SNVs	Admixture graph	Highly multimodal likelihood surface and heuristic search; redo inference from many starting points	64
fastNeutrino	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Applicable only to a single population; designed specifically for extremely large sample sizes	65
DoRIS	Lengths of IBD blocks between pairs of individuals	Demographic history	IBD must be inferred (for example, using Beagle or GERMLINE); specification of lower cut-off minimizes false-negative IBD tracts	71,72
IBS tract inference	Lengths of IBS blocks between pairs of individuals	Demographic	IBS can easily be confounded by missing data and/or sequencing errors	76
PSMC	Diploid genotypes from one individual	Demographic history	Best used in MSMC's PSMC mode, which uses the SMC to more accurately model recombination than the original PSMC; applicable to a single population	78
MSMC	Whole genome, phased haplotypes	Demographic history	Requires large amounts of RAM; cross-coalescence rate should not be interpreted as migration rate	82
CoalHMM	Whole genome, phased haplotypes	Demographic history	Multiple applications, including inference of population sizes, migration rates and incomplete lineage sorting	83–87
diCal	Medium-length, phased haplotypes	Demographic history	Uses shorter sequences than MSMC, but can be applied to multiple individuals in complex demographic models; infers explicit population genetic parameters for migration rates	89,92
LAMARC	Short, phased haplotypes	Demographic history	Requires Monte Carlo sampling of coalescent genealogies; very flexible	93
BEAST	Short, phased haplotypes	Species trees, effective population sizes	Used mainly as a method of phylogenetic inference. Can also infer population size history	94
MCMCcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP[131]	95
G-PhoCS	Short, (un)phased haplotypes	Demographic history	Incorporates migration into the MCMCcoal framework. Averages over unphased haplotypes	96
Exact likelihoods using generating functions	Short, phased haplotypes	Demographic history	Implemented in Mathematica; applicable only to specific classes of multi-population models	97,98

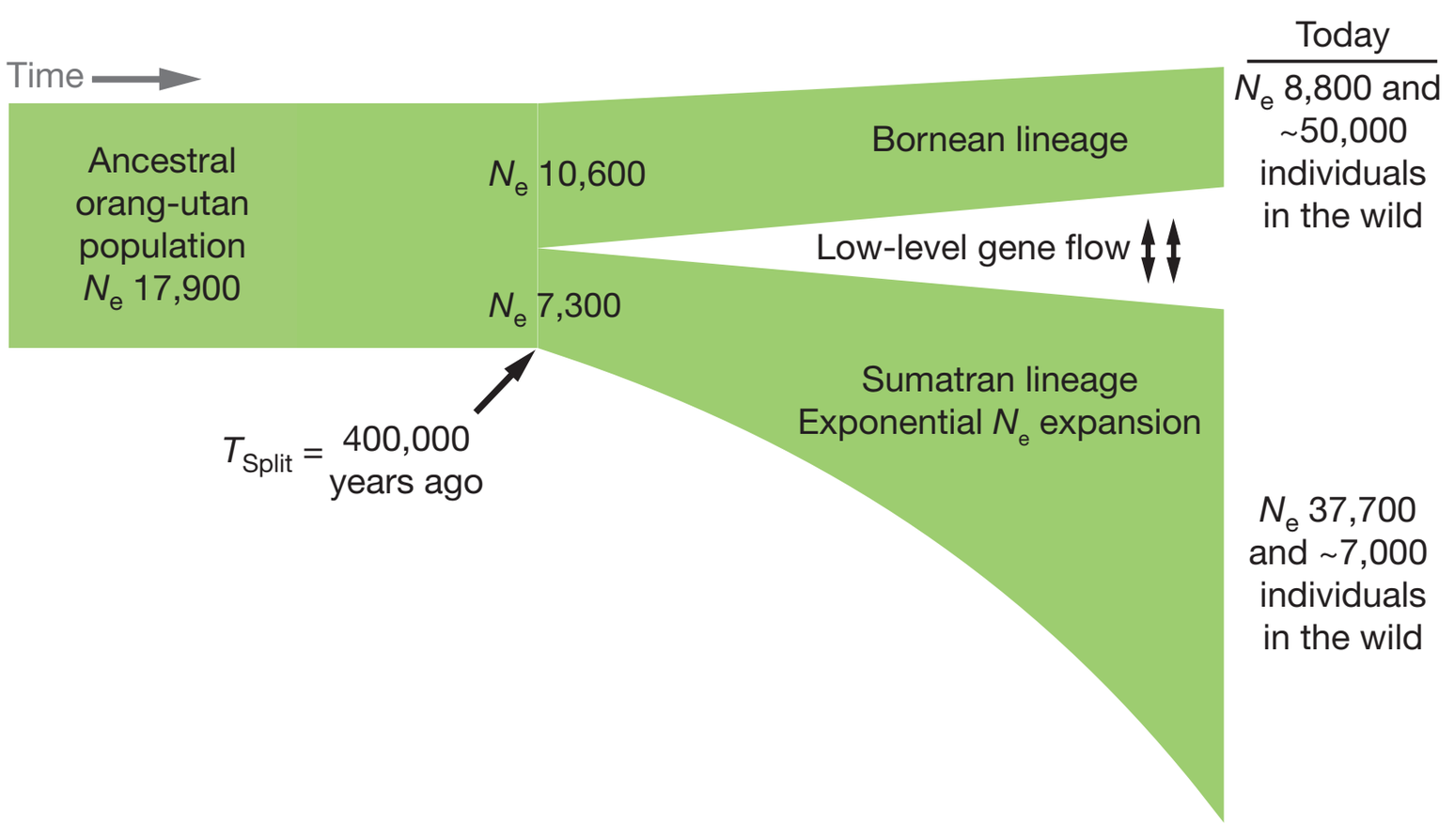
Why model demographic history?

- Understand population history
  Bottlenecks, gene flow, etc.



Why model demographic history?

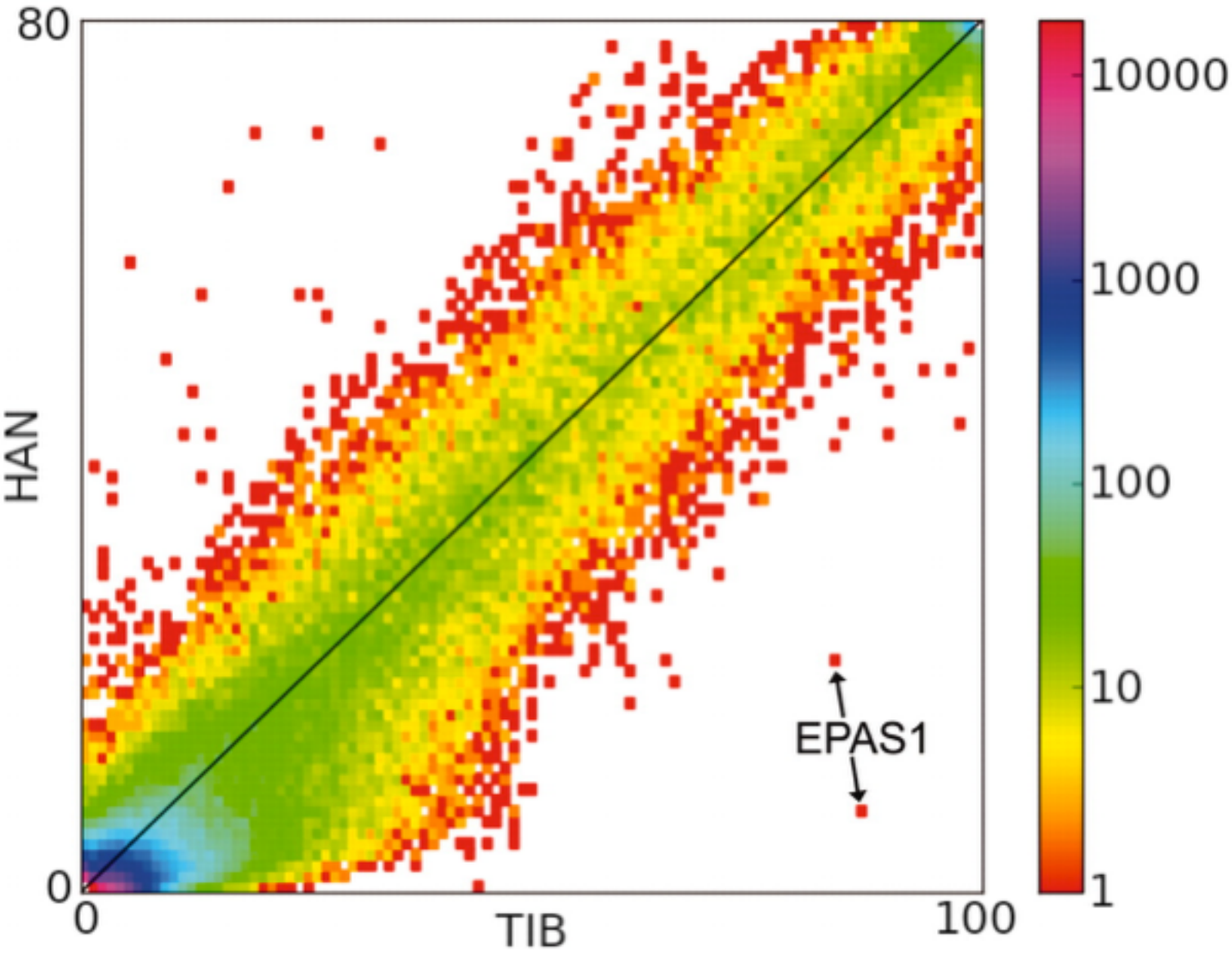
- Conservation
  Present versus historical genetic diversity



Locke et al. (2011) *Nature*

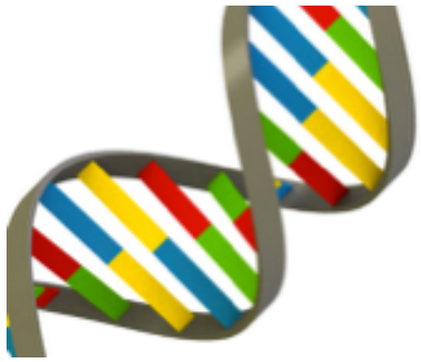
Why model demographic history?

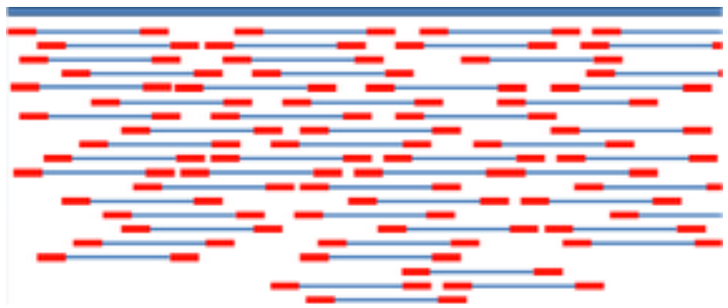
- Selection
  Demographic history sets neutral background

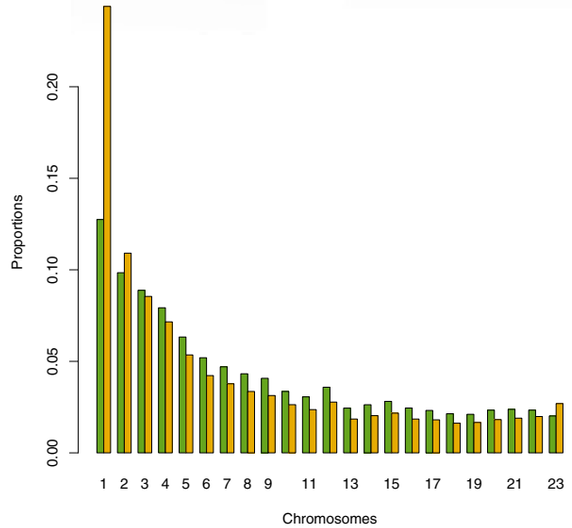


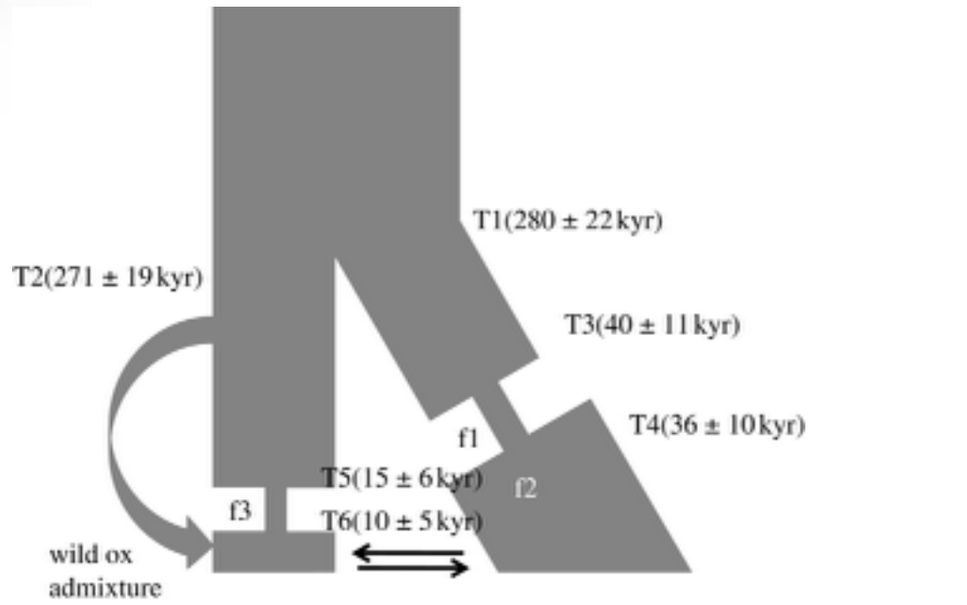
Yi et al. (2010) *Science*

Workflow









Beware: Almost all inference tools assume data is clean!

Modeling workflow

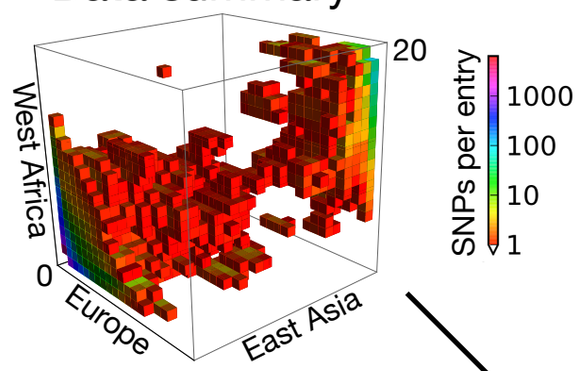
Data


TGGTCACTCTTAT
TGGTCACTCTTAT


TGGTCACTCTTAT
TGGTCACTCTTAT


TGGTCACTCTTAT
TGGTCACTCTTAT

Data summary

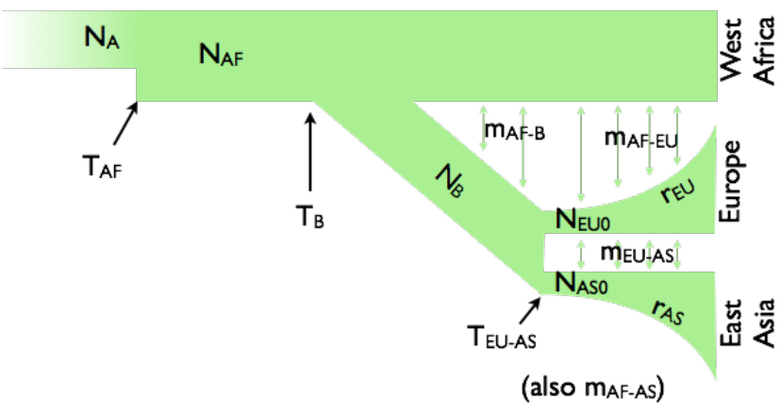


Compare simulations with data.

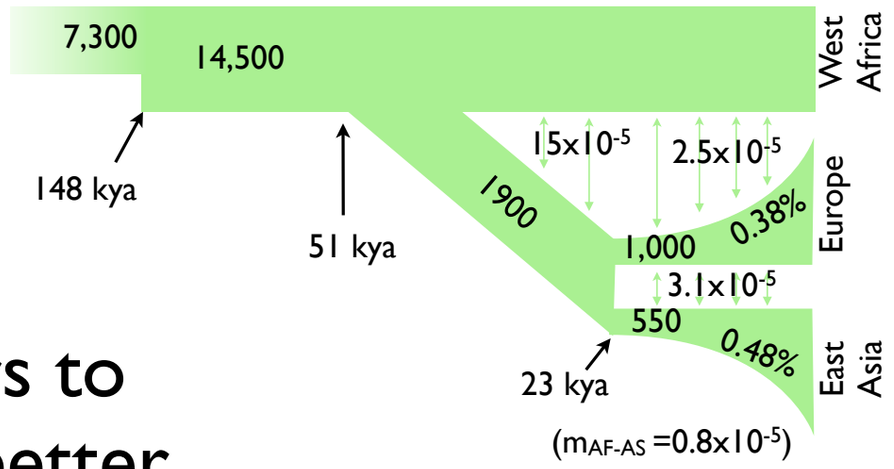
Choose candidate parameter values, simulate data.

Update parameters to (hopefully) fit data better.

Input parameterized model



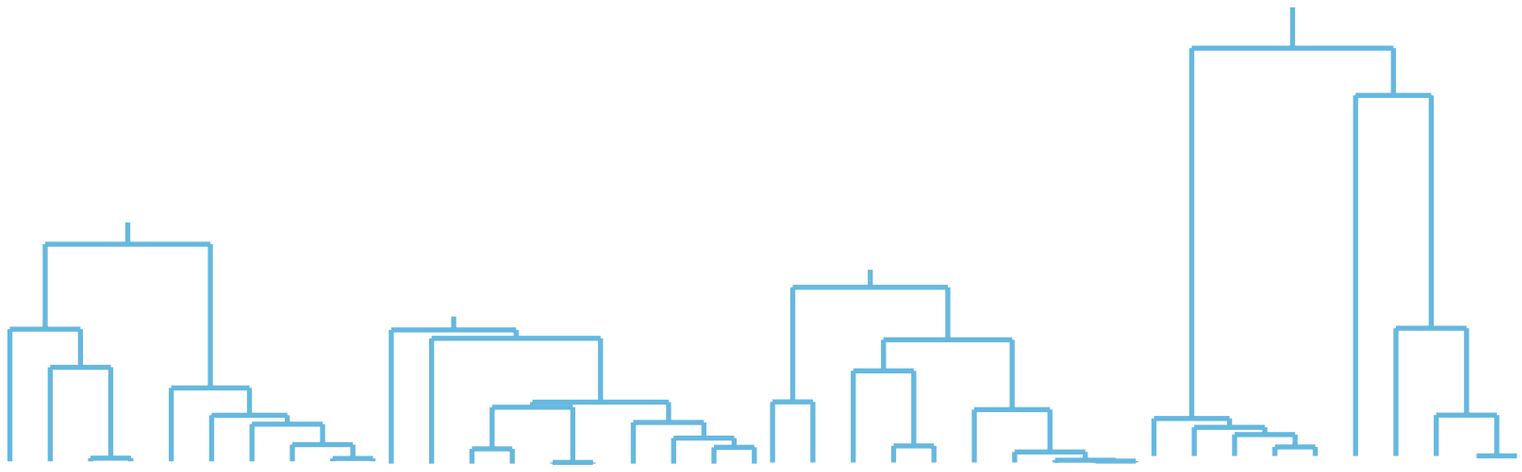
Converged output model



Ne: Effective population size

- The size of an idealized population (in individuals) that would give the same behavior in some regard as the real population of interest.
- Most commonly, variance effective population size, the population size in a Wright-Fisher model that has variance in allele frequencies over time equal to that of the real population.
- Almost always, Ne is less than the census size.
- Ne is affected by breeding ratio, historical demography, etc.
- Other definitions of Ne are possible.
- Arguably, in some populations (Drosophila), variation may be more strongly influenced by selection than drift.

Simulation via coalescent

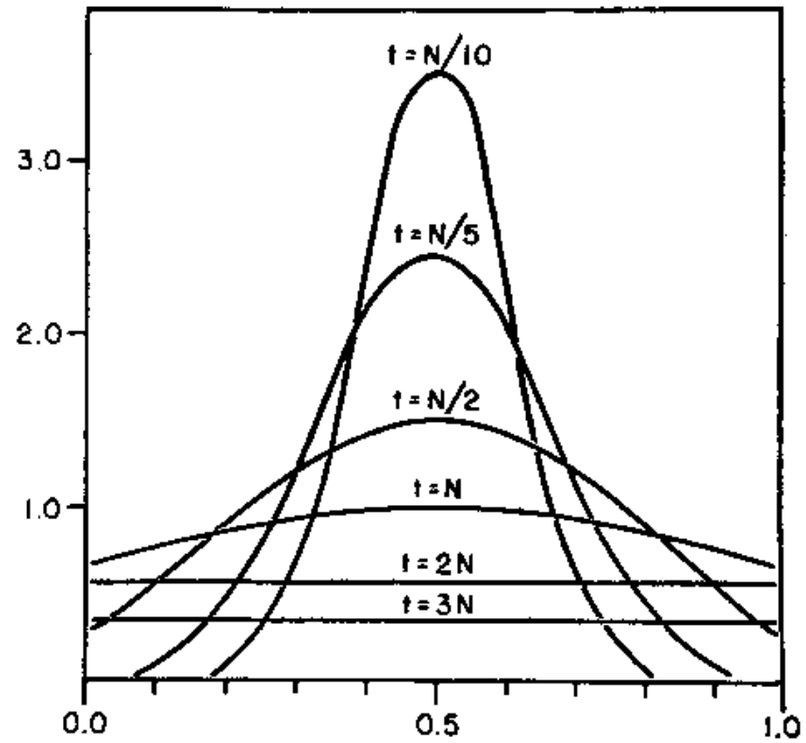


Rosenberg and Nordborg (2002) *Nat Rev Genet*

- Developed in the early 1980s, principally by Kingman.
- Approach is to model the genealogy of sampled sequences.
- Rate of coalescence is proportional to 1/Ne.
- Simulators first sample genealogies consistent with specified demographic history.
- To generate sampled sequences, mutations are then added to the genealogy via a Poisson process.
- Can model recombination with the Ancestral Recombination Graph. But selection is extremely challenging.

Simulation via diffusion

- Developed by Fisher and others in the 1930s. Further developed by Kimura in the 1960s.
- Approach is to model the distribution of allele frequencies in the population(s).
- Approximating allele frequencies as continuous allows partial differential equations to be applied.
- Simulation of selection is straightforward. Linkage is very challenging.



Kimura
(1964) *J Applied Prob*

Comparing model and data

- Likelihood: Probability of the data given the model (with specified parameter values).
- Frequentist approach: Maximize likelihood to find best-fit parameters, estimate confidence intervals, perform hypothesis tests.
- Bayesian approach: Sample posterior distribution of parameters based on likelihood function and prior distribution over parameters.

Composite likelihoods

- Often in population genetics, we can't calculate the likelihood of our data.
- But we can often calculate the likelihood for a single site.
- The composite likelihood function is the product of the likelihoods over all sites, implicitly assuming that sites evolve independently.
- Under neutrality, it can be shown that the composite likelihood approximation does not bias inferred parameters (Wiuf (2006) *J Theor Biol*).
- But composite likelihood does mean that many standard statistical inference approaches will be too liberal, because they effectively overestimate the amount of data.

Approximate Bayesian Computation

- Seminal work by Tavare et al. (1997) *Genetics*
- Simulate data by sampling from the prior distribution
- Calculate summary statistics from the simulation, and compare with summary statistics from the data.
- Accept the sampled parameter set if "distance" between summary statistics is less than some small threshold.
- The set of sampled parameters is an approximation to posterior distribution.
- Very computationally intensive, but very flexible, approach to model fitting, because you can chose summary statistics that are most sensitive to your particular problem.

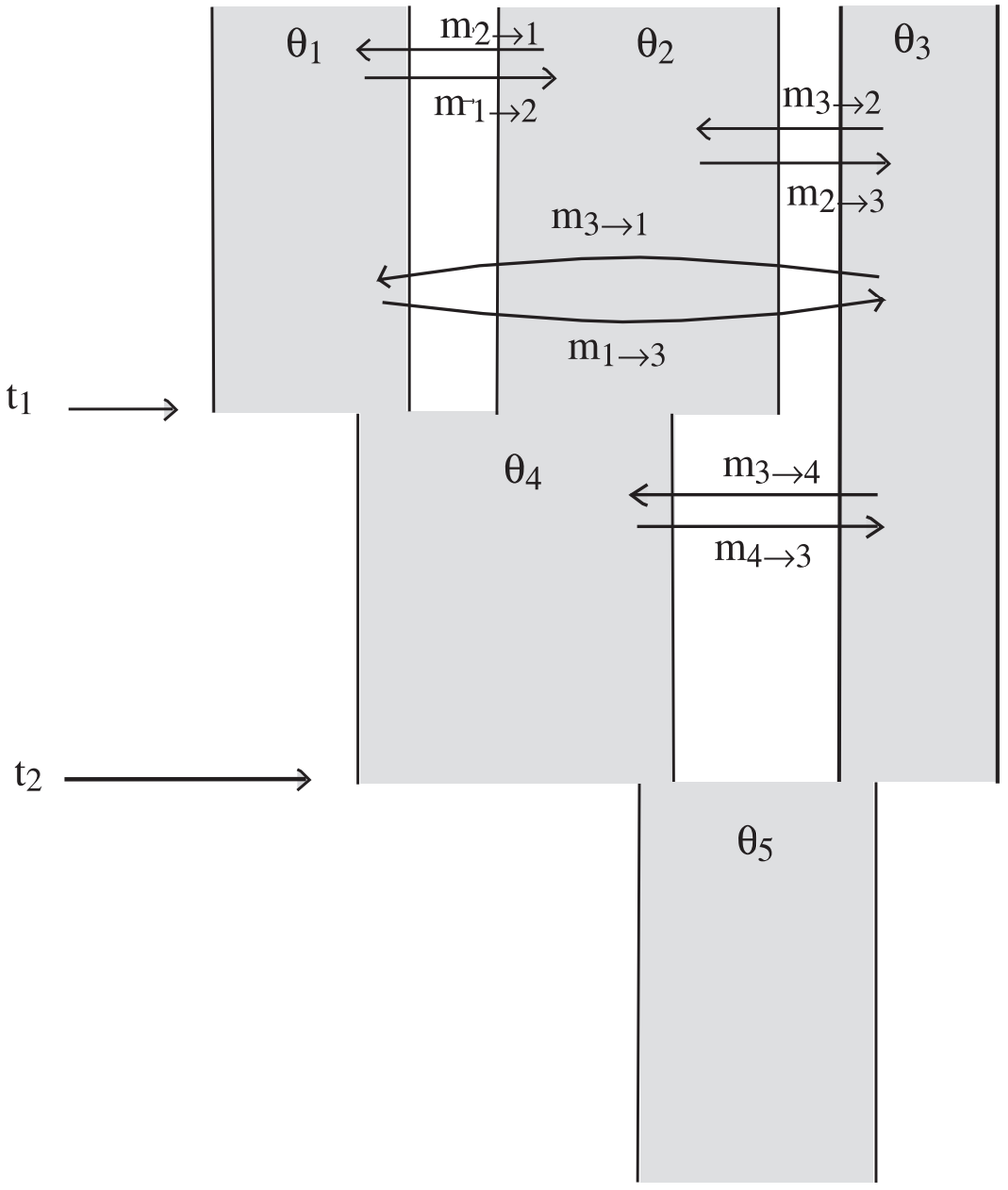
Demographic Inference Methods

- Many approaches
- My overview will be somewhat selective and historical
- See Schraiber and Akey (2015) *Nat Rev Genet* for a recent review.

Name	Data type	Inference	Notes	Refs
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User-friendly GUI; can be computationally demanding	32
FRAPPE	Unlinked bi-allelic SNVs	Population structure, admixture	Alexander *et al.*[41] argue that convergence is not guaranteed	40
ADMIXTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Estimates the number of populations via cross-validation error	41
fastSTRUCTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Obtains variational Bayesian estimates of posterior probability distribution	42
Structurama	Unlinked multi-allelic genotypes	Population structure, admixture	Uses a Dirichlet process to estimate the number of populations	43
HAPMIX	Phased haplotypes; reference panel	Chromosome painting	Requires populations to be specified a priori	48
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	49
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Extends the fineSTRUCTURE approach to estimate unsampled ancestral populations and admixture times	7
LAMP	Phased haplotypes; reference panel	Chromosome painting	Identifies local ancestry in windows, rather than using an HMM, so is more discrete than other approaches	52
PCAdmix	Phased haplotypes	Chromosome painting, population structure	Uses PCA in small chunks followed by an HMM to estimate local ancestry	53
dadi	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Requires some Python-coding skills; applicable to up to three populations	60
Fastsimcoal2	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Can also be used to simulate data under the SMC	62,63
Treemix	Frequencies of unlinked bi-allelic SNVs	Admixture graph	Highly multimodal likelihood surface and heuristic search; redo inference from many starting points	64
fastNeutrino	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Applicable only to a single population; designed specifically for extremely large sample sizes	65
DoRIS	Lengths of IBD blocks between pairs of individuals	Demographic history	IBD must be inferred (for example, using Beagle or GERMLINE); specification of lower cut-off minimizes false-negative IBD tracts	71,72
IBS tract inference	Lengths of IBS blocks between pairs of individuals	Demographic	IBS can easily be confounded by missing data and/or sequencing errors	76
PSMC	Diploid genotypes from one individual	Demographic history	Best used in MSMC's PSMC mode, which uses the SMC to more accurately model recombination than the original PSMC; applicable to a single population	78
MSMC	Whole genome, phased haplotypes	Demographic history	Requires large amounts of RAM; cross-coalescence rate should not be interpreted as migration rate	82
CoalHMM	Whole genome, phased haplotypes	Demographic history	Multiple applications, including inference of population sizes, migration rates and incomplete lineage sorting	83–87
diCal	Medium-length, phased haplotypes	Demographic history	Uses shorter sequences than MSMC, but can be applied to multiple individuals in complex demographic models; infers explicit population genetic parameters for migration rates	89,92
LAMARC	Short, phased haplotypes	Demographic history	Requires Monte Carlo sampling of coalescent genealogies; very flexible	93
BEAST	Short, phased haplotypes	Species trees, effective population sizes	Used mainly as a method of phylogenetic inference. Can also infer population size history	94
MCMCcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP[131]	95
G-PhoCS	Short, (un)phased haplotypes	Demographic history	Incorporates migration into the MCMCcoal framework. Averages over unphased haplotypes	96
Exact likelihoods using generating functions	Short, phased haplotypes	Demographic history	Implemented in Mathematica; applicable only to specific classes of multi-population models	97,98

IM/IMa/IMa2

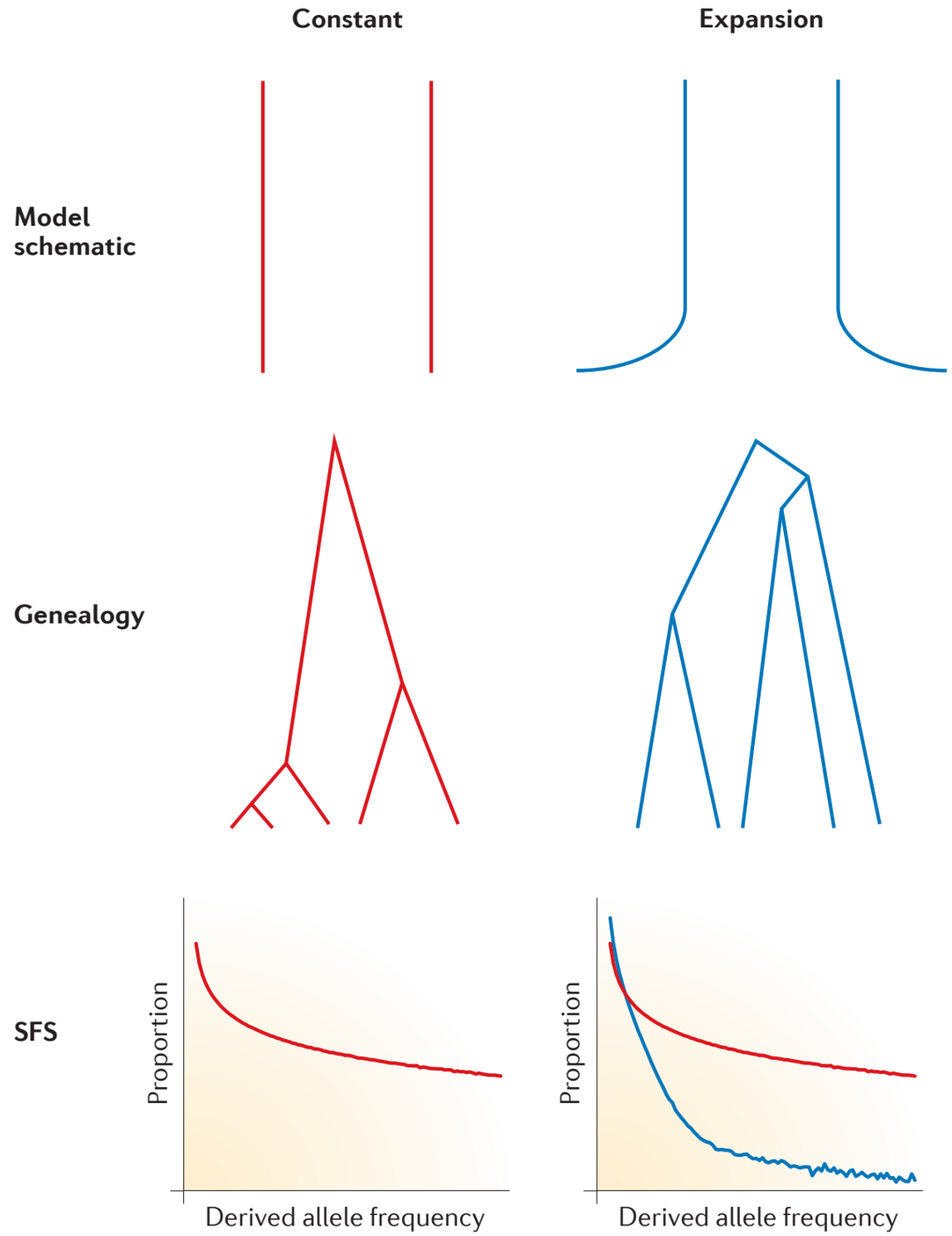
- Uses coalescent simulation to calculate the full likelihood of the data given the model, for non-recombining regions (mitochondria, Y chromosome, small autosomal regions).
- Bayesian inference based on MCMC walk through parameter space, can be computationally expensive.
- Handles arbitrary number of populations.



Hey and Nielsen (2004) *Genetics*
Hey (2010) *Mol Biol Evol*

Allele Frequency Spectrum

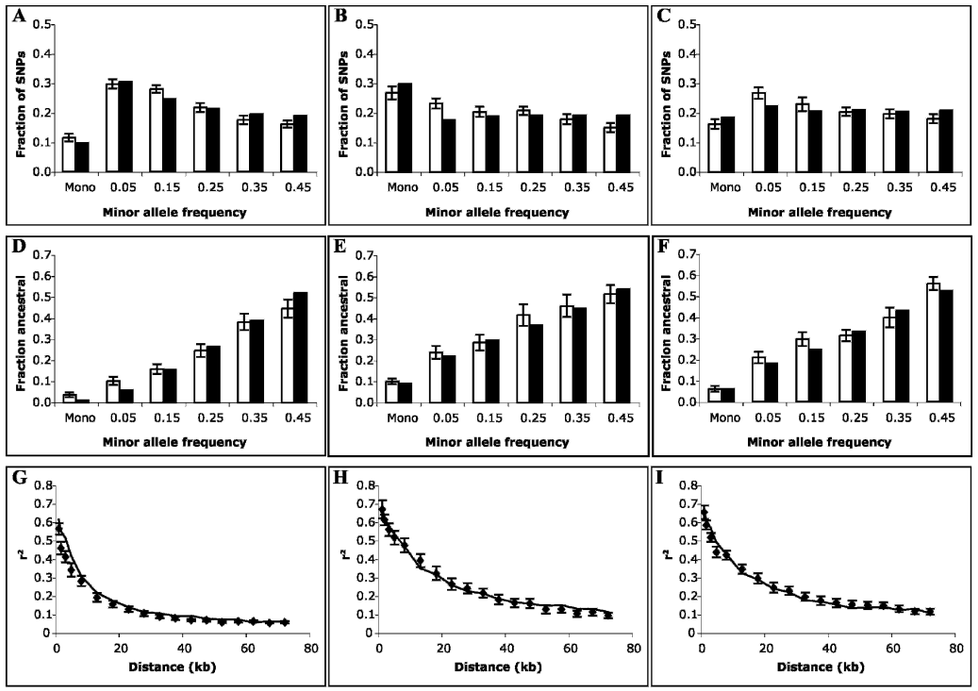
- In a population of constant size (the Standard Neutral Model), expectation number of SNPs at frequency i is proportional to 1/i.
- Population growth creates an excess of low-frequency alleles.
- Commonly quantified by Tajima's *D*.
- Selection generates a similar signal.

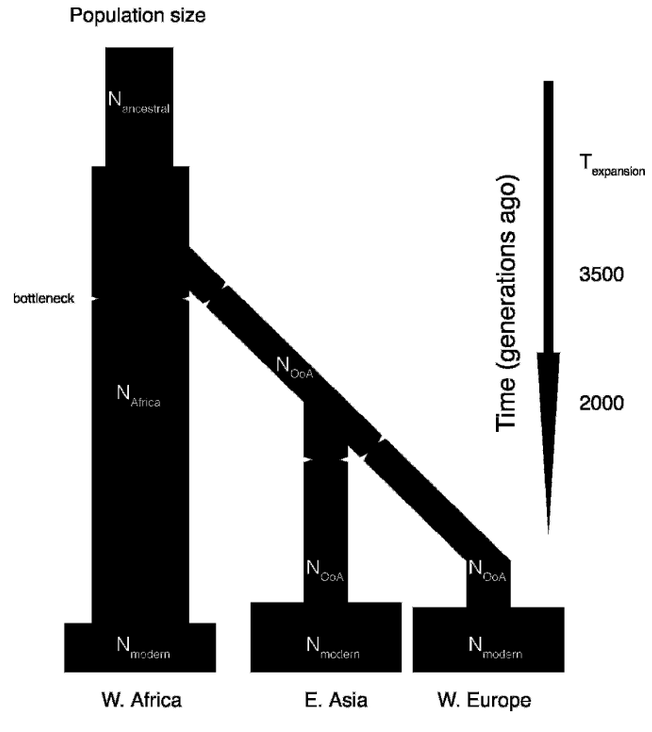


Schraiber and Akey (2015) *Nat Rev Genet*

Schaffner et al. (2005) *Genome Research*

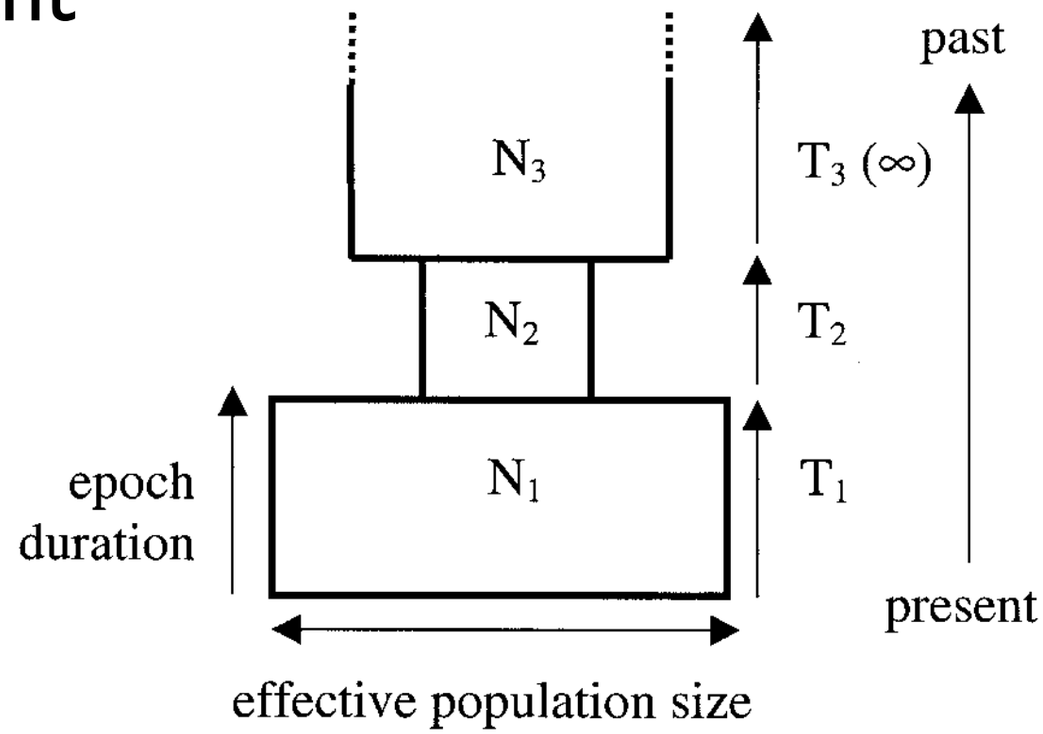
- Fit model to Africa, Asian, and European human data by using coalescent simulations to match several summary statistics (single-population frequency spectra, F_{ST}, and LD decay).
- Today, this would be called ABC.
- Still no packaged methods that integrate both allele frequencies and LD.



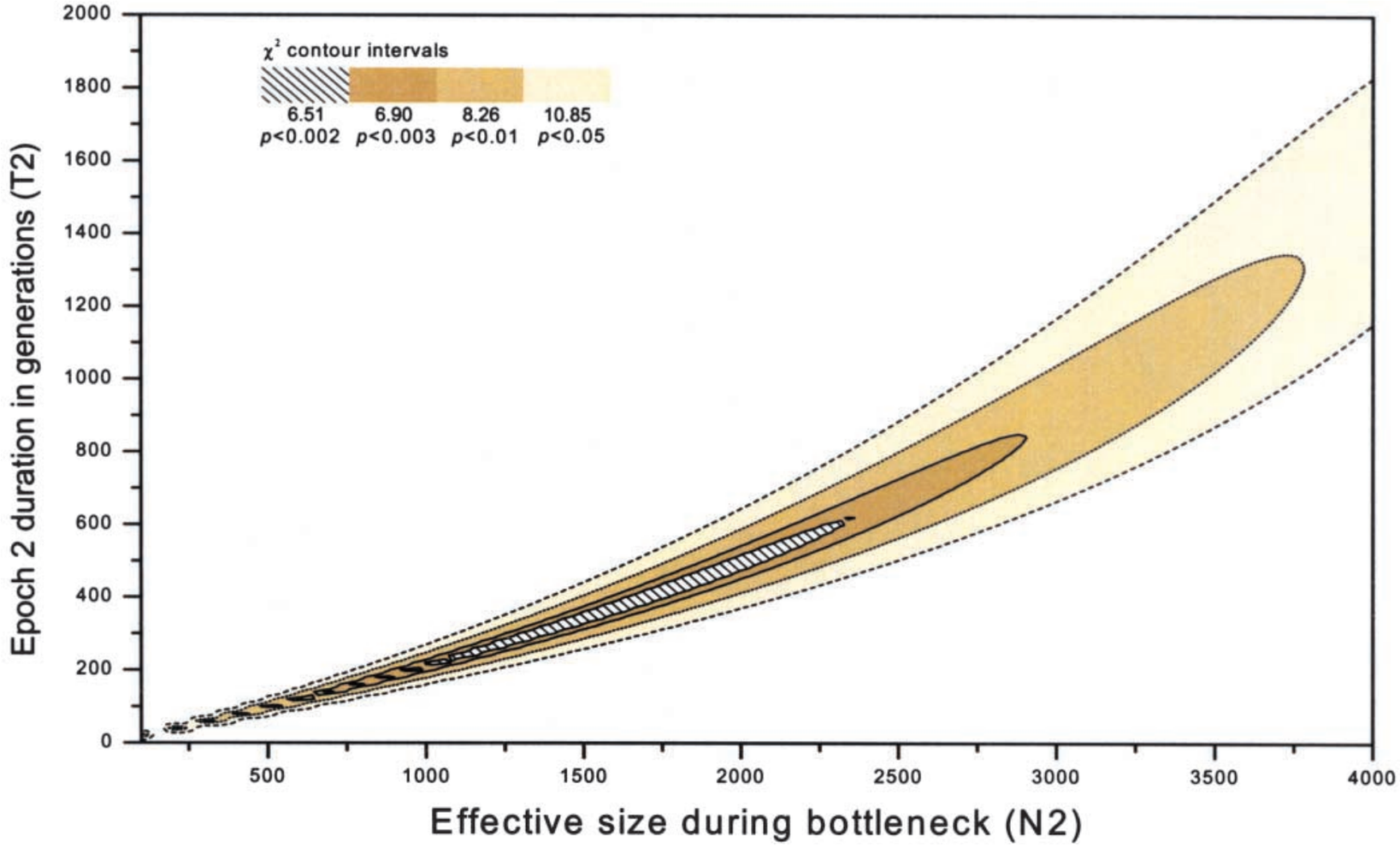


Inference from the AFS

- Marth et al. (2004) *Genetics* calculated expected frequency spectrum under piecewise constant histories using (big) summation formulas.
- Implemented projection of data down to smaller sample size (for handling incomplete calling) and correction for ascertainment bias (for handling genotype data).
- Fit growth and bottleneck models to human populations from Africa, Europe, and Asia.



Parameter identifiability

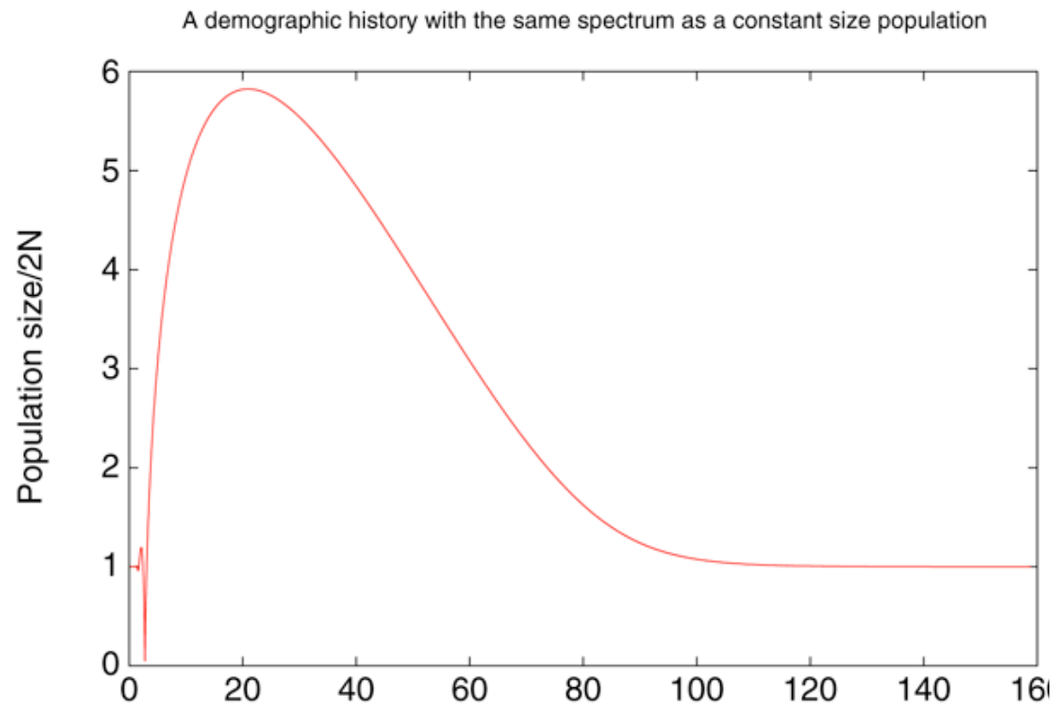


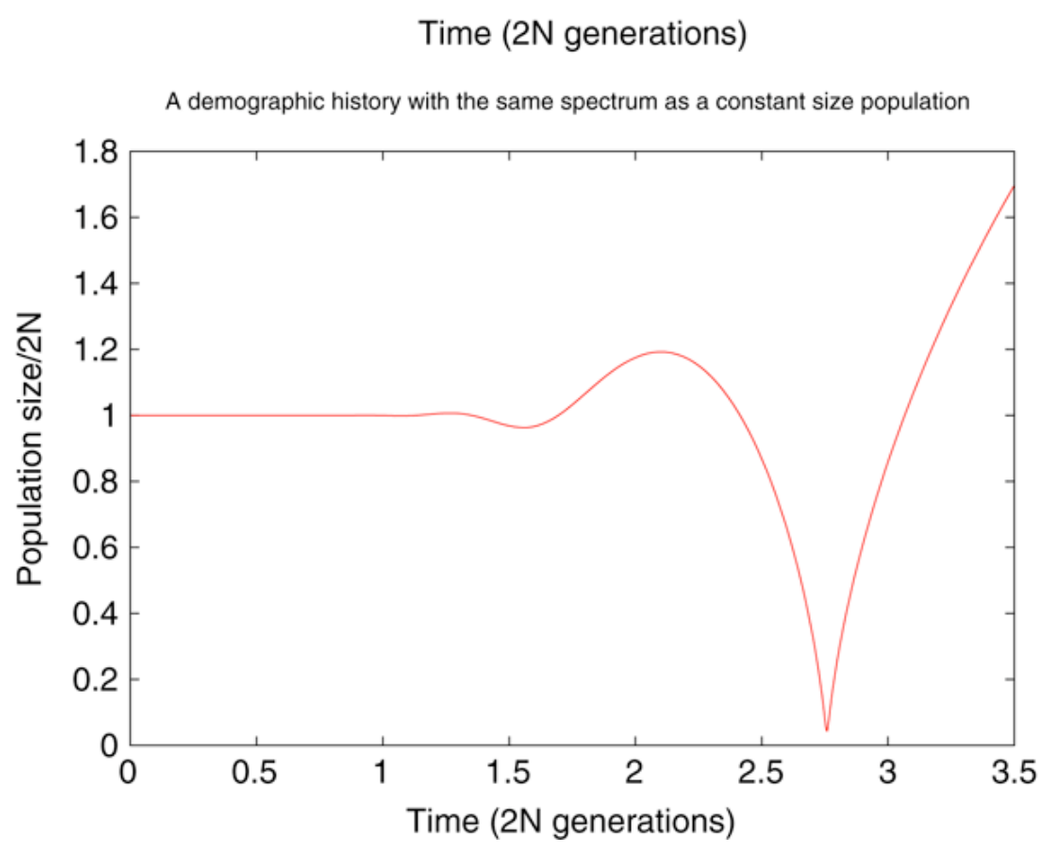
- Often the likelihood surface has "ridges", correlated sets of parameters that give very similar high likelihoods.
- For example, depth and duration of bottleneck.

Marth et al. (2004) *Genetics*

Absolute limits to inference

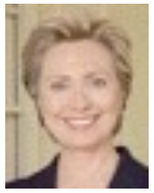
- Myers, Fefferman, and Patterson (2008) *Theor Pop Biol*
- Can show analytically that even an infinite amount of frequency spectrum data does not uniquely determine population history.
- Recent results from Song show that can uniquely determine piecewise constant histories.



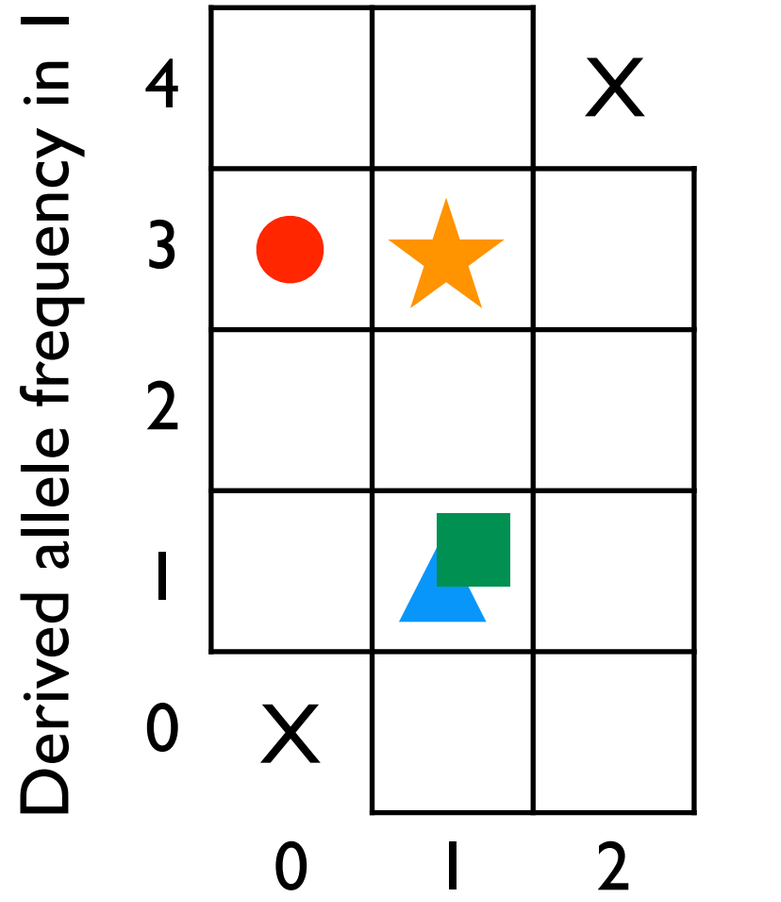


Joint Allele Frequency Spectrum

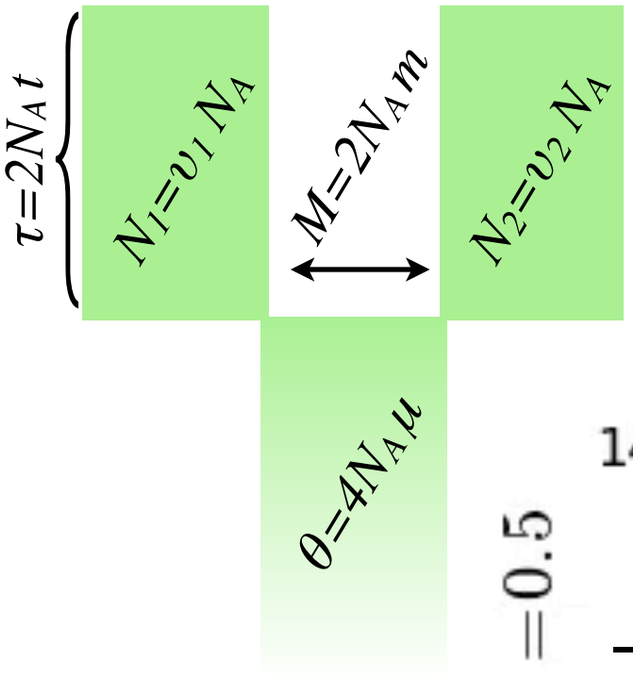
Single nucleotide polymorphisms (SNPs)

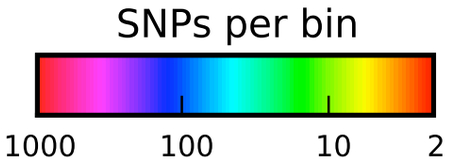


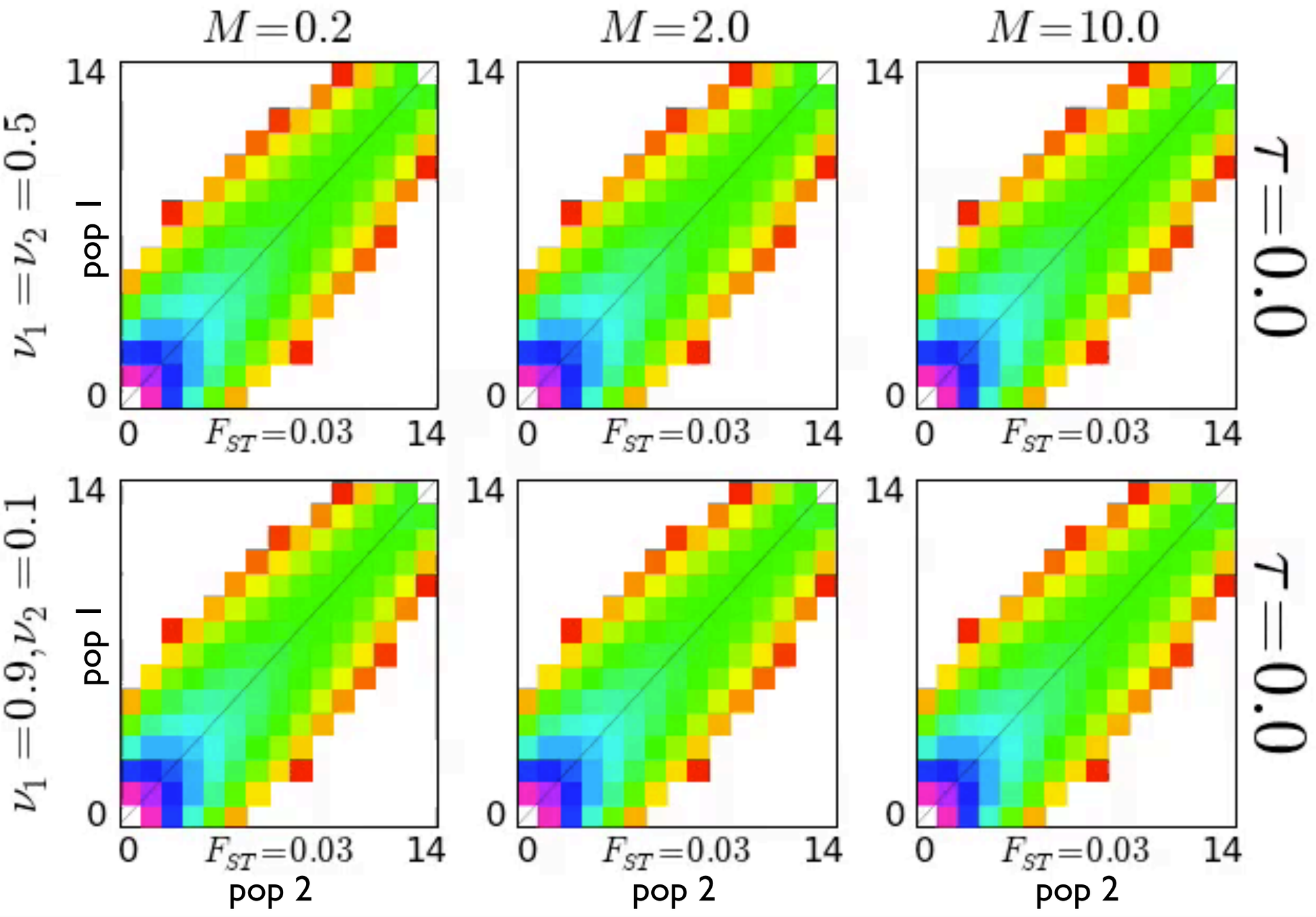

```
TGGTCACTCTTATCATTATGT
TGGTCACTCTTATCATTATGT
TGGTCACTCTTAACATTCTGT
TGCTCATTCTTATCATTATGT
TGCTCATTCTTATCATTCTGT
TGCTCACTCTTAACATTATGT
AGCTCACTCGTATCATTCTGT
```



Frequency spectrum gallery







∂a∂i: Diffusion Approximations for Demographic Inference

- Up to three interacting populations, with arbitrary parameter time courses

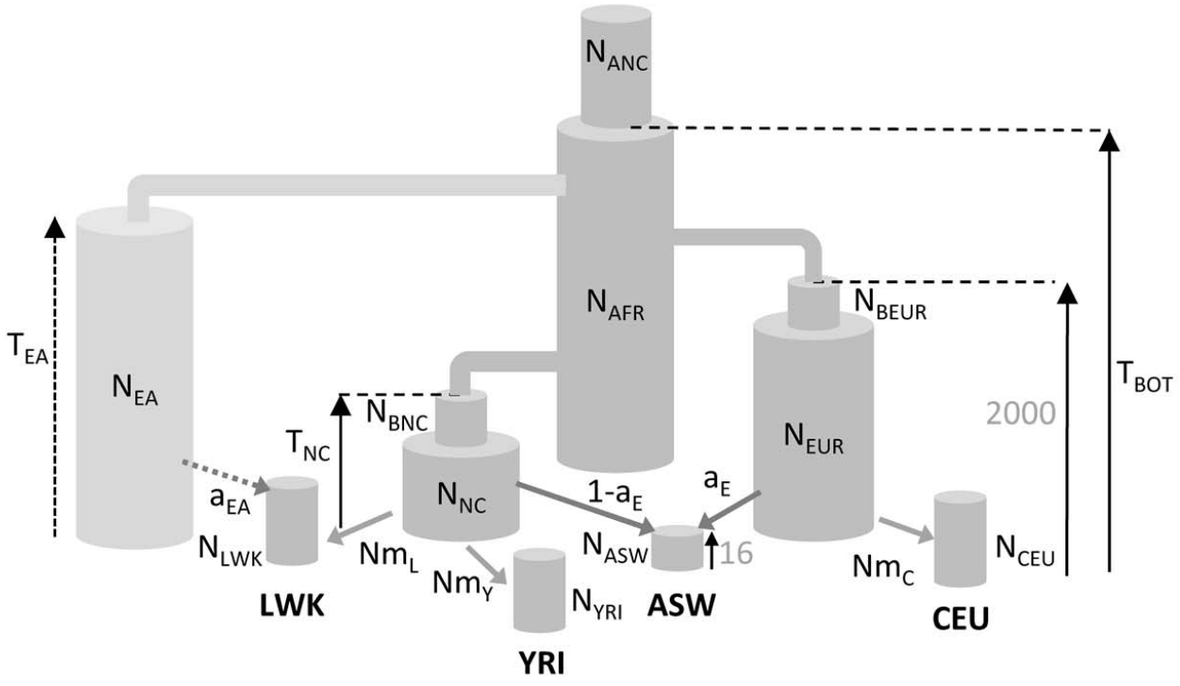
- 1 pop, 20 samples, ~3 params: ~1 minute to fit
  2 pops, 20 samples each, ~6 params: ~10 minutes to fit
  3 pops, 20 samples each, ~12 params: ~3 hours to fit

- Computational cost independent of SNP count, but exponential in number of populations.

Gutenkunst et al.
PLoS Genet (2009)

fastsimcoal2

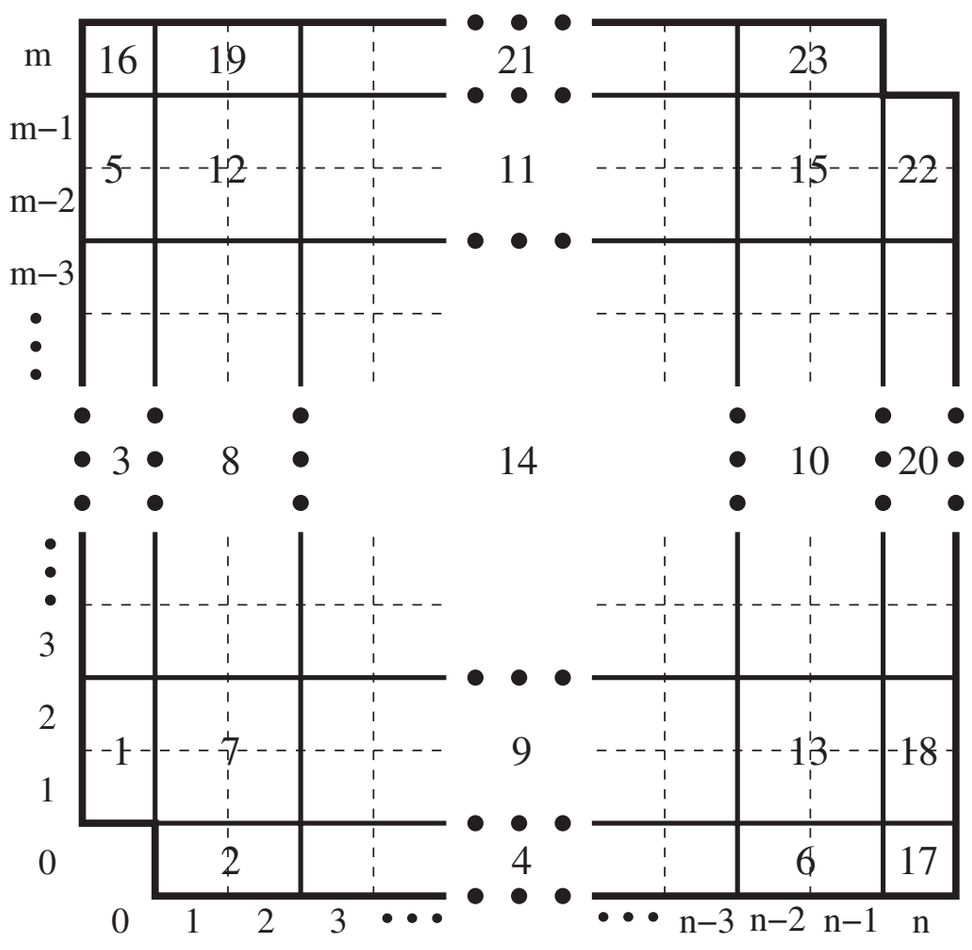
- Estimate pairwise joint frequency spectra using coalescent simulations.
- Scales to arbitrary number of populations.
- Estimate parameters by maximum composite likelihood.
- Optimization may be more robust than ∂a∂i.



Excoffier et al.
PLoS Genet (2013)

Jaatha

- Partition joint AFS into 23 distinct regions and use simulations to fit to these summary statistics.



- Recently applied to simultaneously infer demography and selection. (Mathew and Jensen (2015) *Frontiers Genet*).

Naduvilezhath et al. (2011) *Molec Ecol*

MOMI

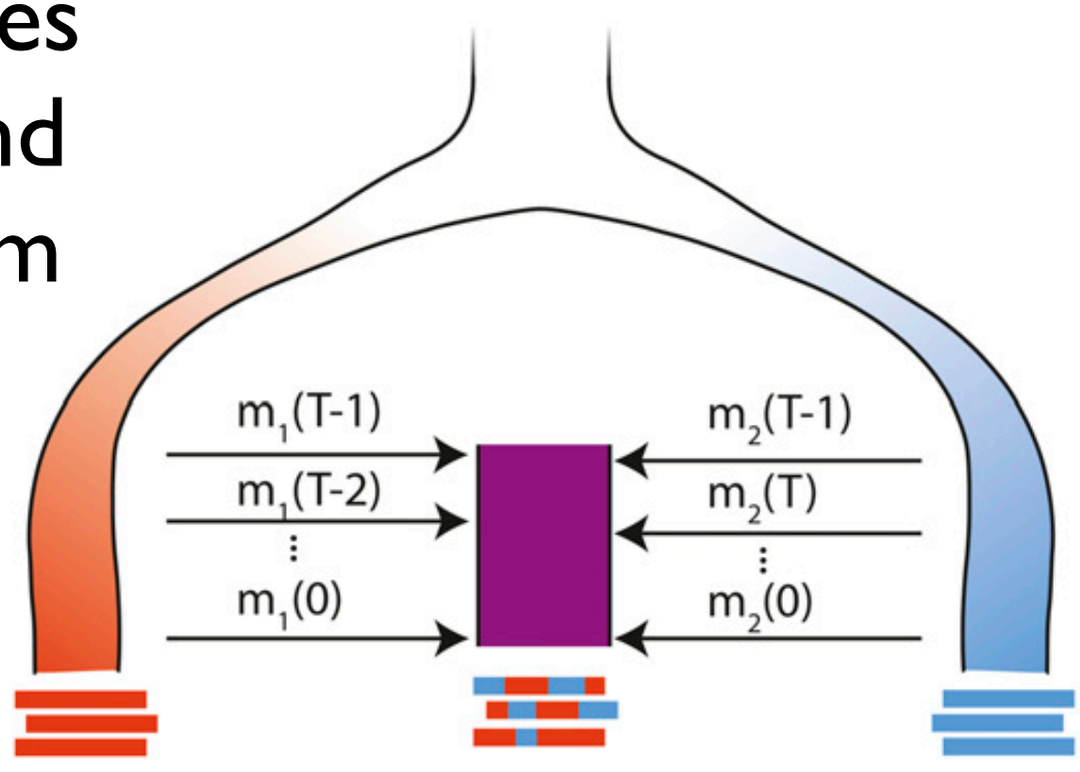
- Kamm, Terhorst, and Song (2015) arXiv
- Use coalescent theory to calculate expected joint frequency spectra for arbitrary number of populations and demography that piecewise constant or exponential.
- Should be faster and more numerically stable than diffusion or coalescent simulation methods.
- Software forthcoming.

Sequentially Markovian Coalescent

- Introduced by McVean and Cardin (2005) *Phil Trans R Soc B* as approximation to standard coalescent with recombination.
- Essential assumption is that when recombination occurs the genealogy to the right of the recombination event depends only on the genealogy to the left of the event. (Hence the name Sequentially Markovian Coalescent.)
- Often an excellent approximation to the full coalescent, while being much faster to computer, and more amenable to analysis.

Haplotype lengths

- The genomes of admixed individuals will be mosaics of the source populations.
- As time passes since admixture, recombination breaks up admixture tracts.
- TRACTS infers admixture times (potentially multiple pulses) and proportions from the spectrum of haplotype lengths.



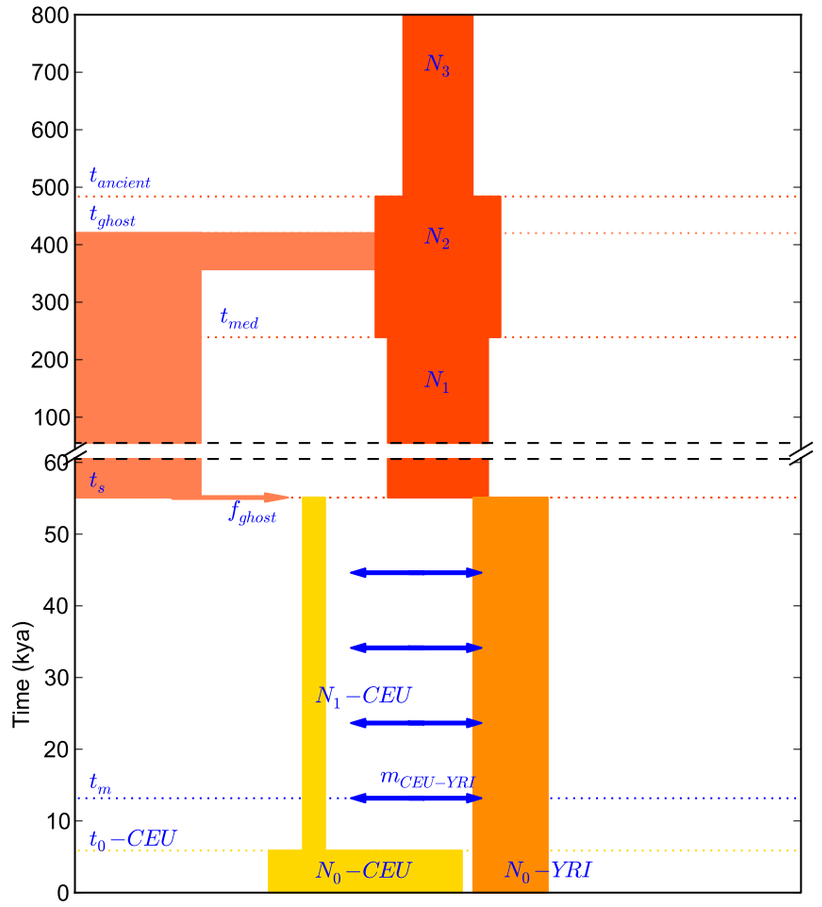
Gravel et al.
Genetics (2012)

IBS tracts

- Sequences that are Identical By State (IBS) with and between populations are informative about demographic history.
- Calculate expected spectrum of IBS tract lengths using coalescent theory.
- Can fit very complex models.

AGGTCGAGCTTG
ACGTCGAGCTGG

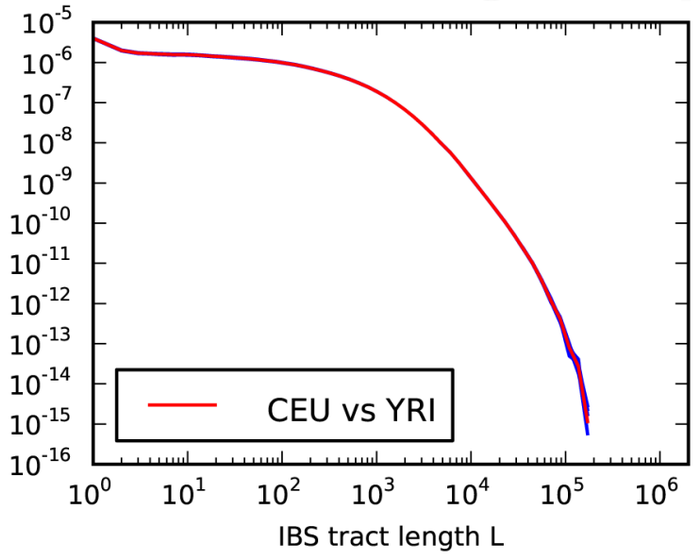
L bases of IBS

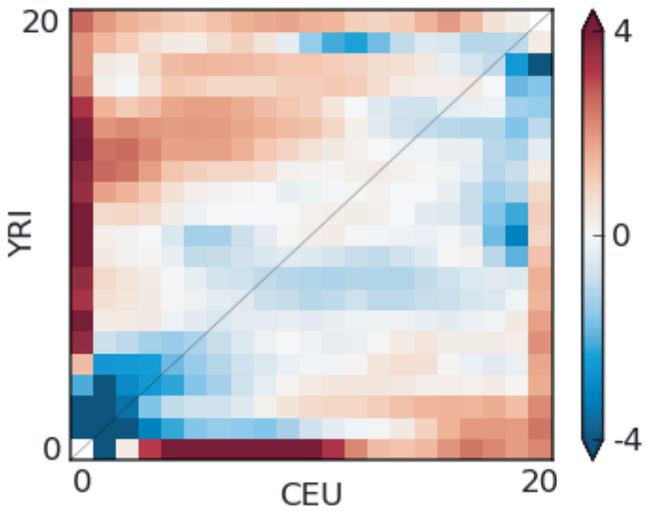


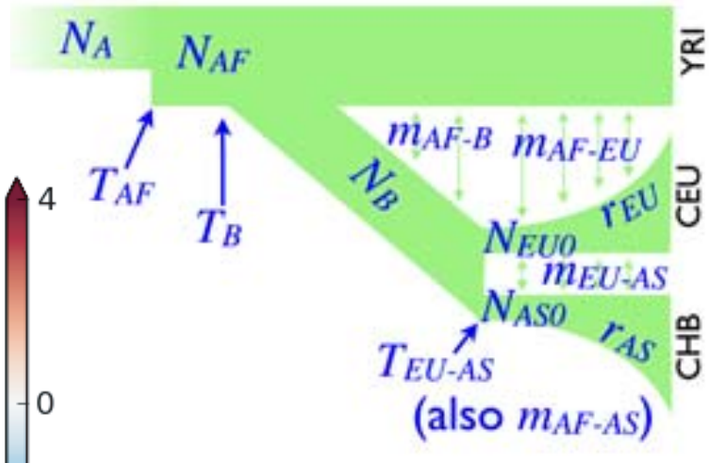
Harris and Nielsen (2013) *PLoS Genet*

Same model, same data, different summary, different results

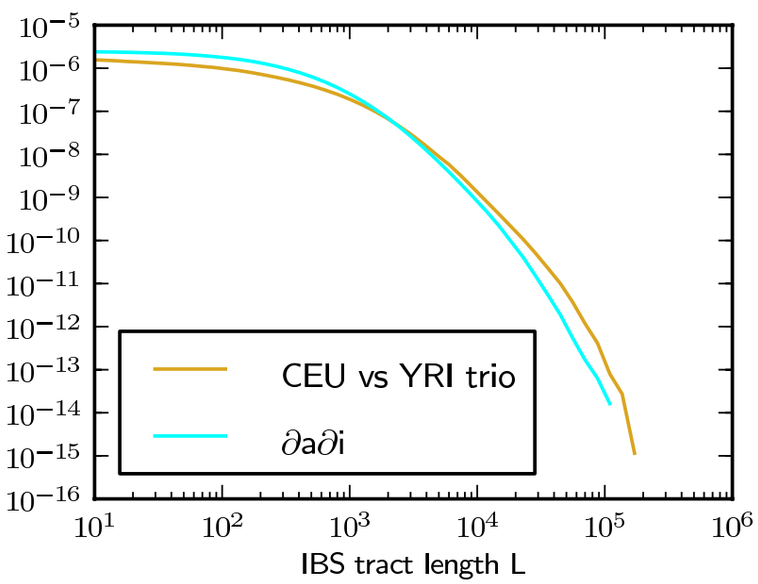
- Fit IBS tract lengths, predict AFS

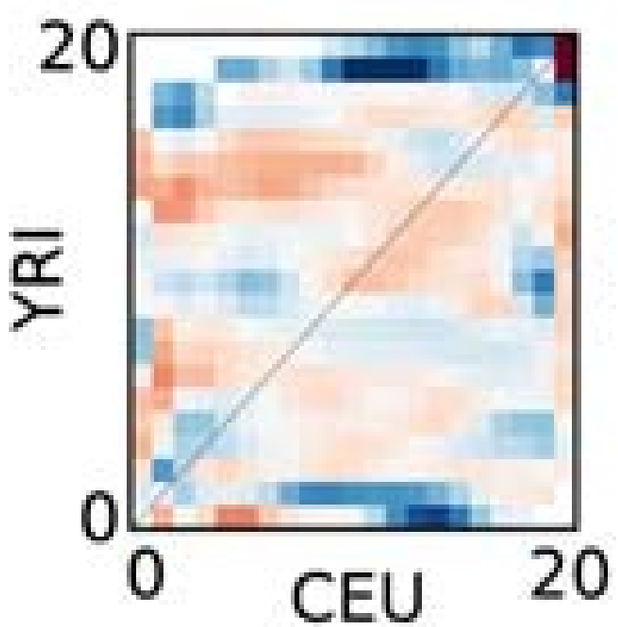






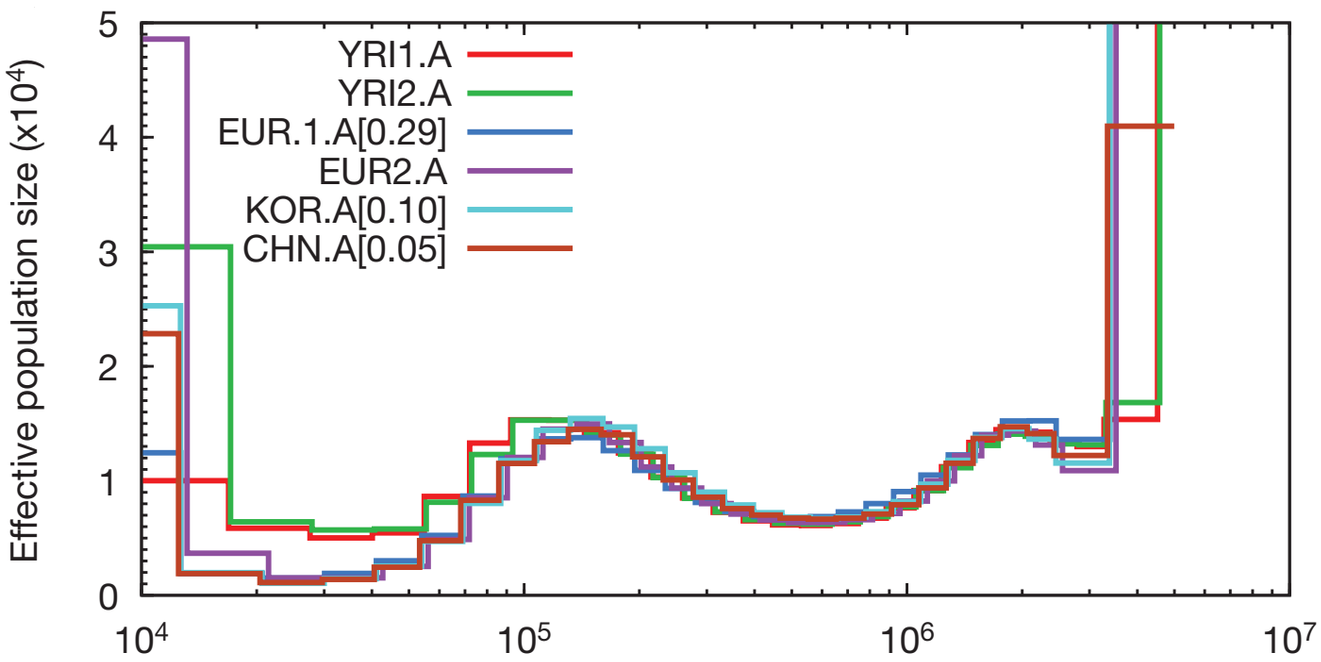
- Fit AFS, predict IBS tract lengths





Gutenkunst et al. (2009) *PLoS Genet*
Harris and Nielsen (2013) *PLoS Genet*

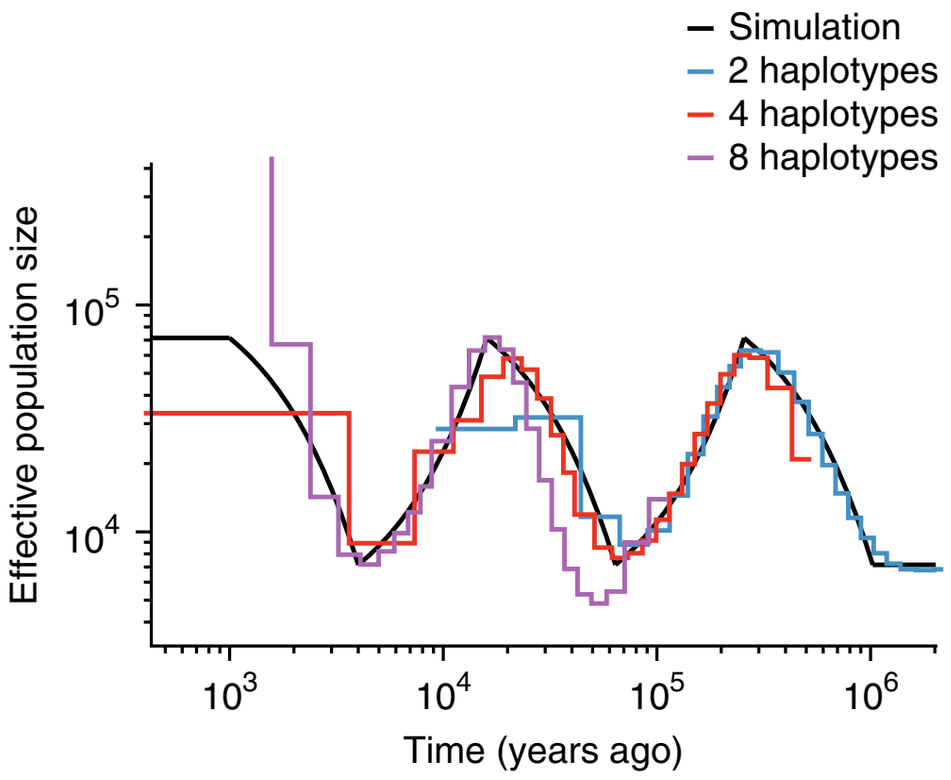
PSMC

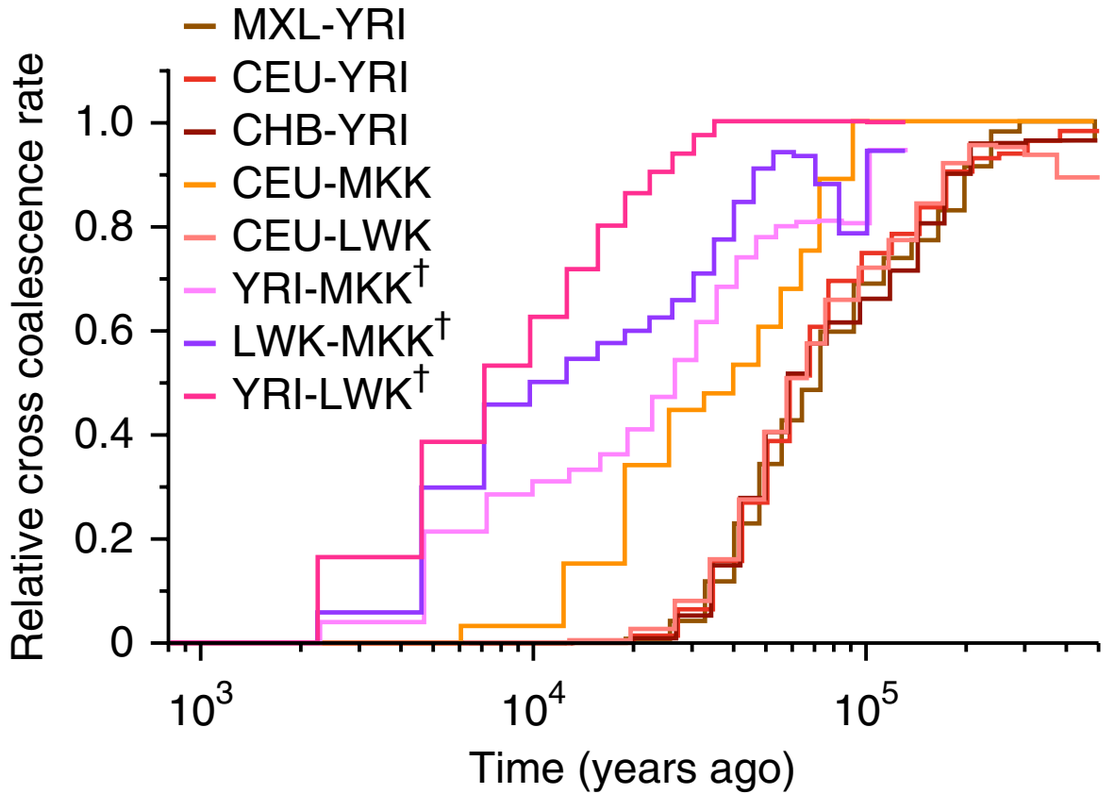


- Estimate effective population size over time from a single unphased genome.
- No parametric model (e.g. exponential growth) assumed.

Li and Durbin
(2011) *Nature*

MSMC





- SMC model for multiple phased sequences
- Inferences of population sizes for more recent times than PSMC.
- Inferences of cross-coalescent rates between populations, which are indicative of population divergence and migration.

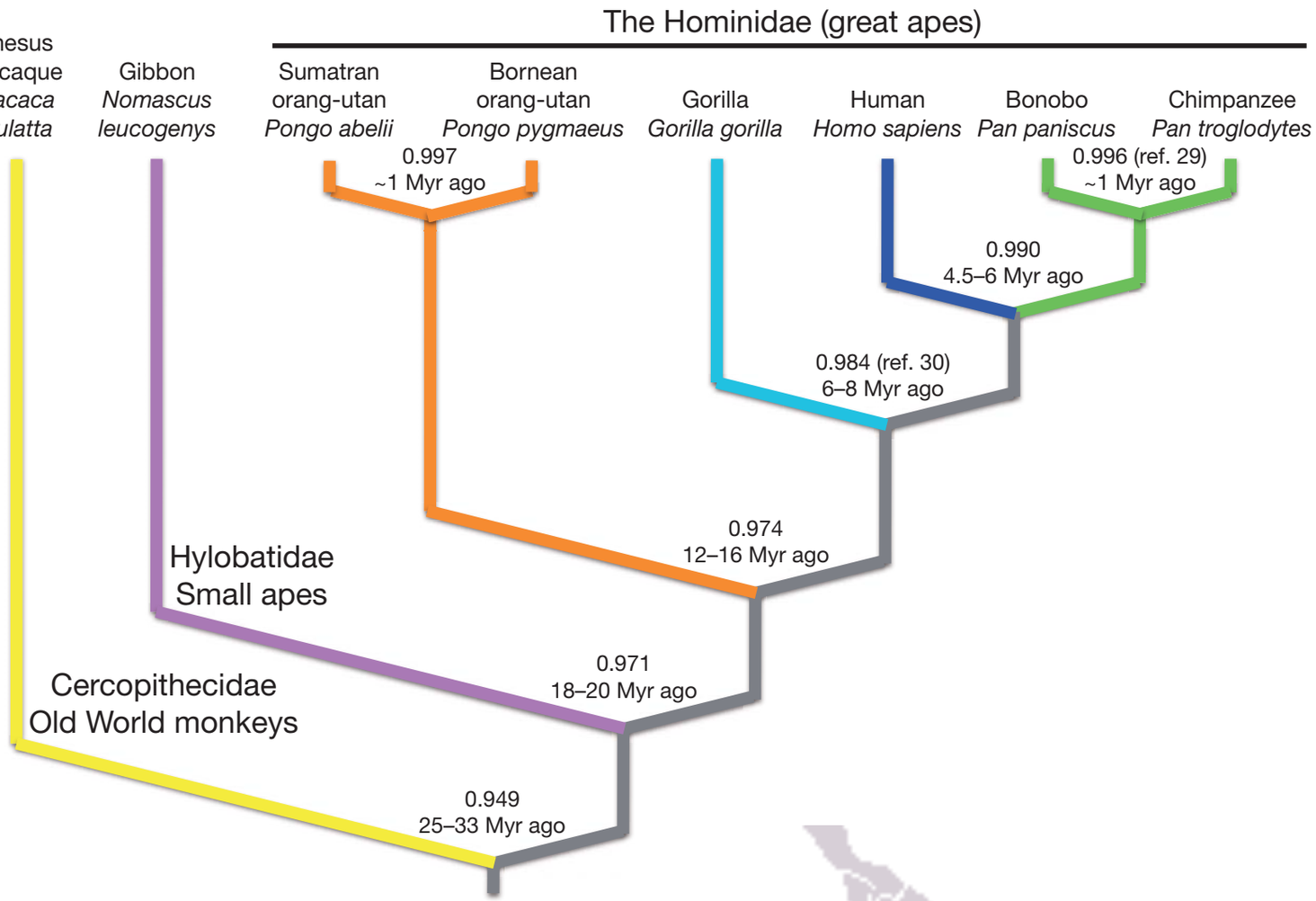
Schiffels and Durbin
Nature Genet (2014)

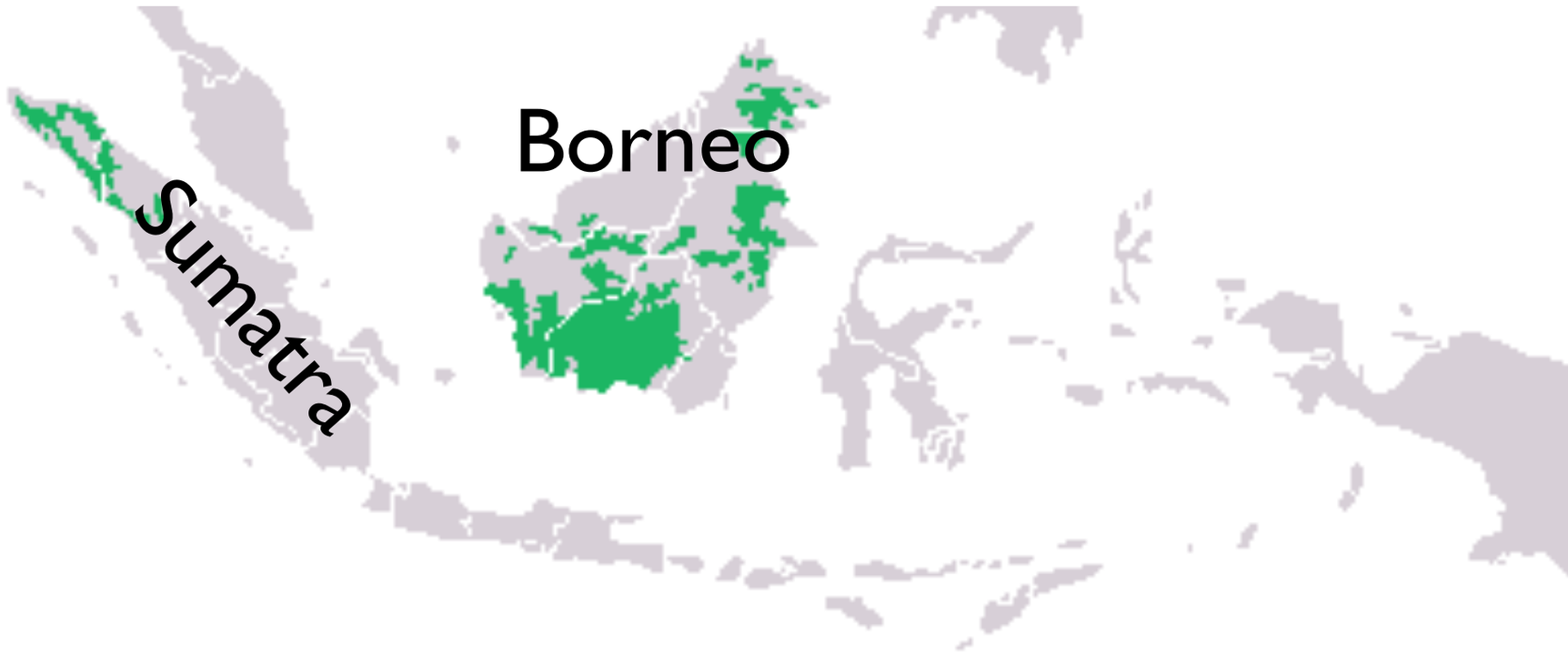
Demographic Inference Methods

- Many options…
- For inference from non-recombining regions, IMa2 is most powerful.
- For inference from many short sequences (RAD-seq, transcriptomes), frequency spectrum methods are most powerful.
- If you can reliably phase and align your data, haplotype methods are very powerful.

Demography and selection in orang-utans

Orang-utans


The Hominidae (great apes)
Rhesus macaque *Macaca mulatta*
Gibbon *Nomascus leucogenys*
Sumatran orang-utan *Pongo abelii*
Bornean orang-utan *Pongo pygmaeus*
Gorilla *Gorilla gorilla*
Human *Homo sapiens*
Bonobo *Pan paniscus*
Chimpanzee *Pan troglodytes*
0.997
~1 Myr ago
0.996 (ref. 29)
~1 Myr ago
0.990
4.5–6 Myr ago
0.984 (ref. 30)
6–8 Myr ago
0.974
12–16 Myr ago
Hylobatidae
Small apes
0.971
18–20 Myr ago
Cercopithecidae
Old World monkeys
0.949
25–33 Myr ago


Borneo
Sumatra

Sequencing

Locke et al.
Nature (2011)

Reference genome sequencing

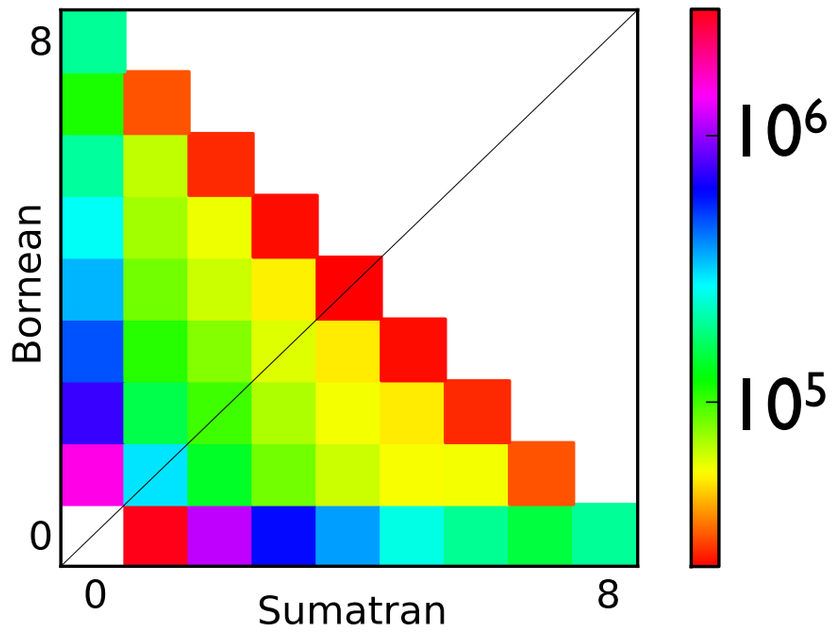
5.6-fold Sanger coverage of Sumatran female
~2.5% divergence from human

Population genomic sequencing

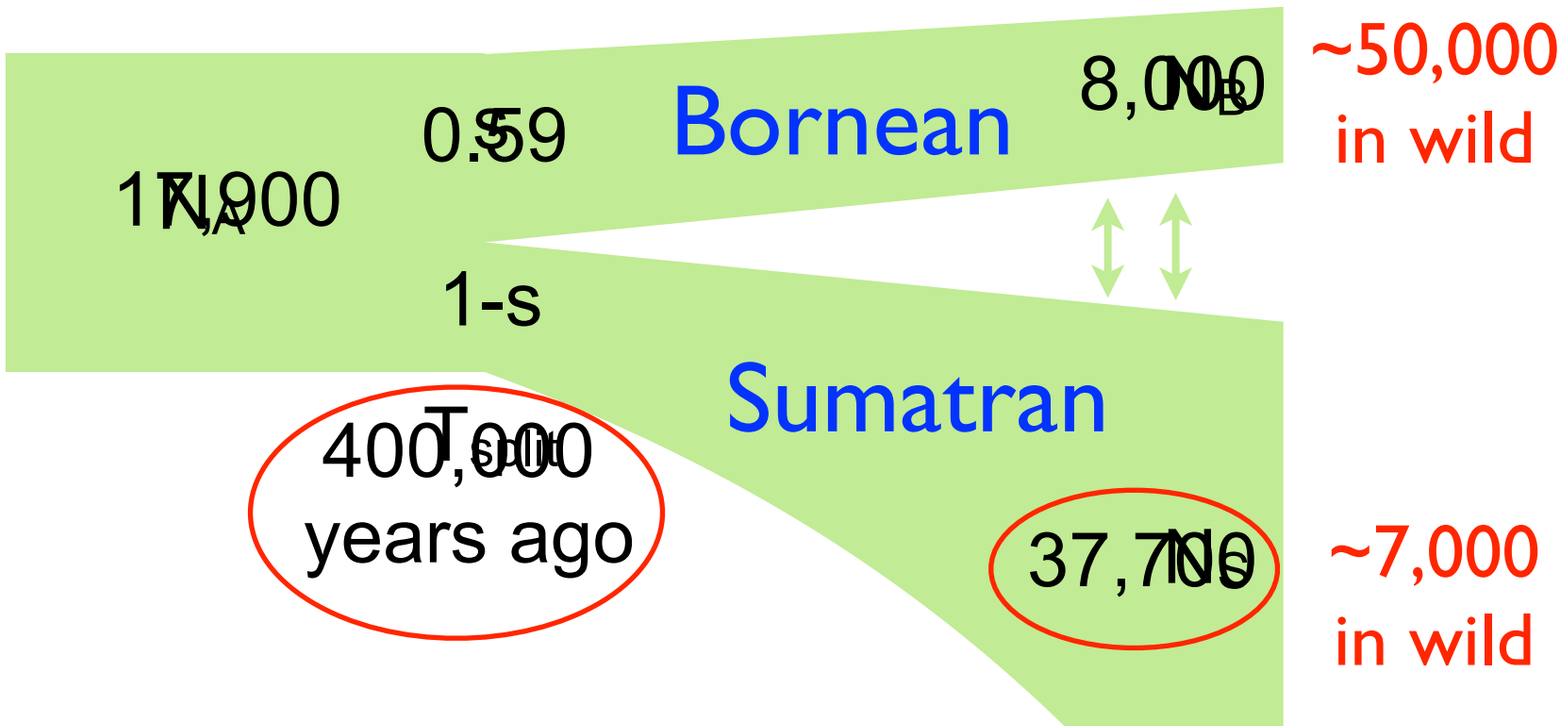
5 Sumatran and 5 Bornean individuals on Illumina GAII
one Bornean individual to 20-fold, rest to ~8-fold
~0.3% divergence between Bornean and Sumatran

Custom Bayesian SNP caller
overall 99% concordance with Sanger validation
~8% false positive rate for singletons

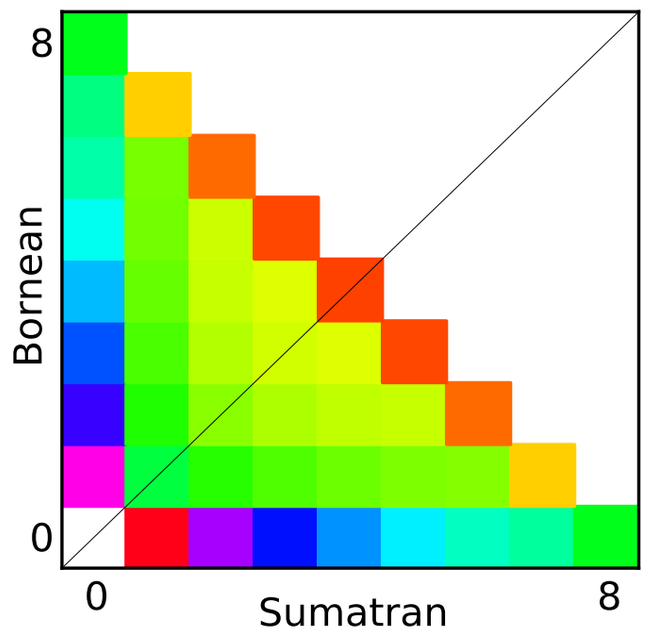
Demographic inference



11.8 million SNPs in folded spectrum



Max likelihood model



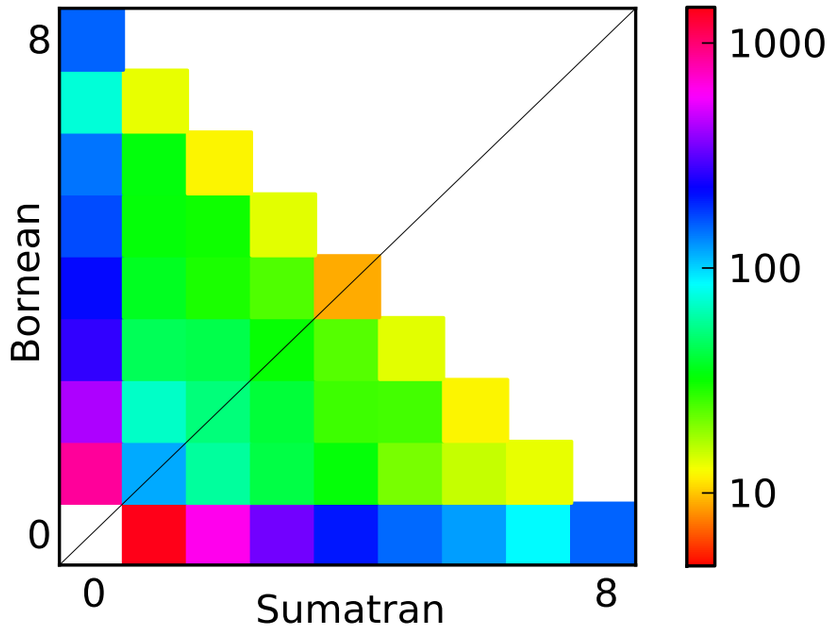
Locke et al. *Nature* (2011)

T_{split} more recent than earlier estimates, but consistent with coalescent HMM inference

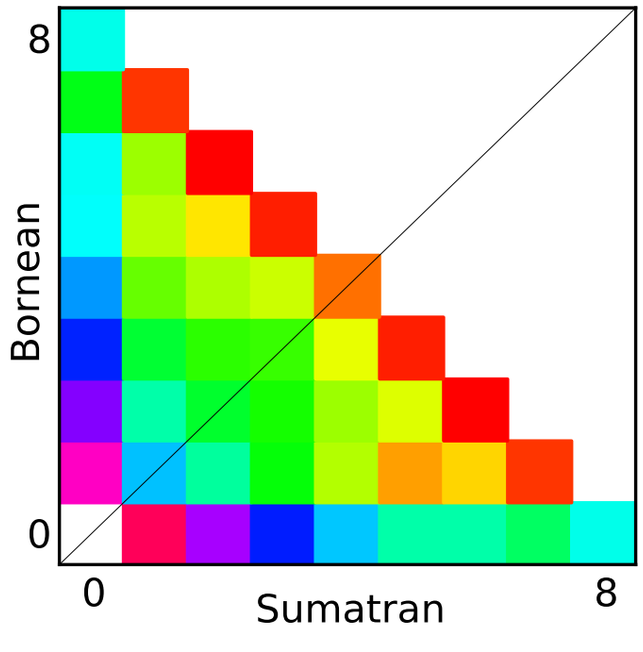
Mailund et al. *PLoS Genet* (2011)

Selection against non-synonymous mutations

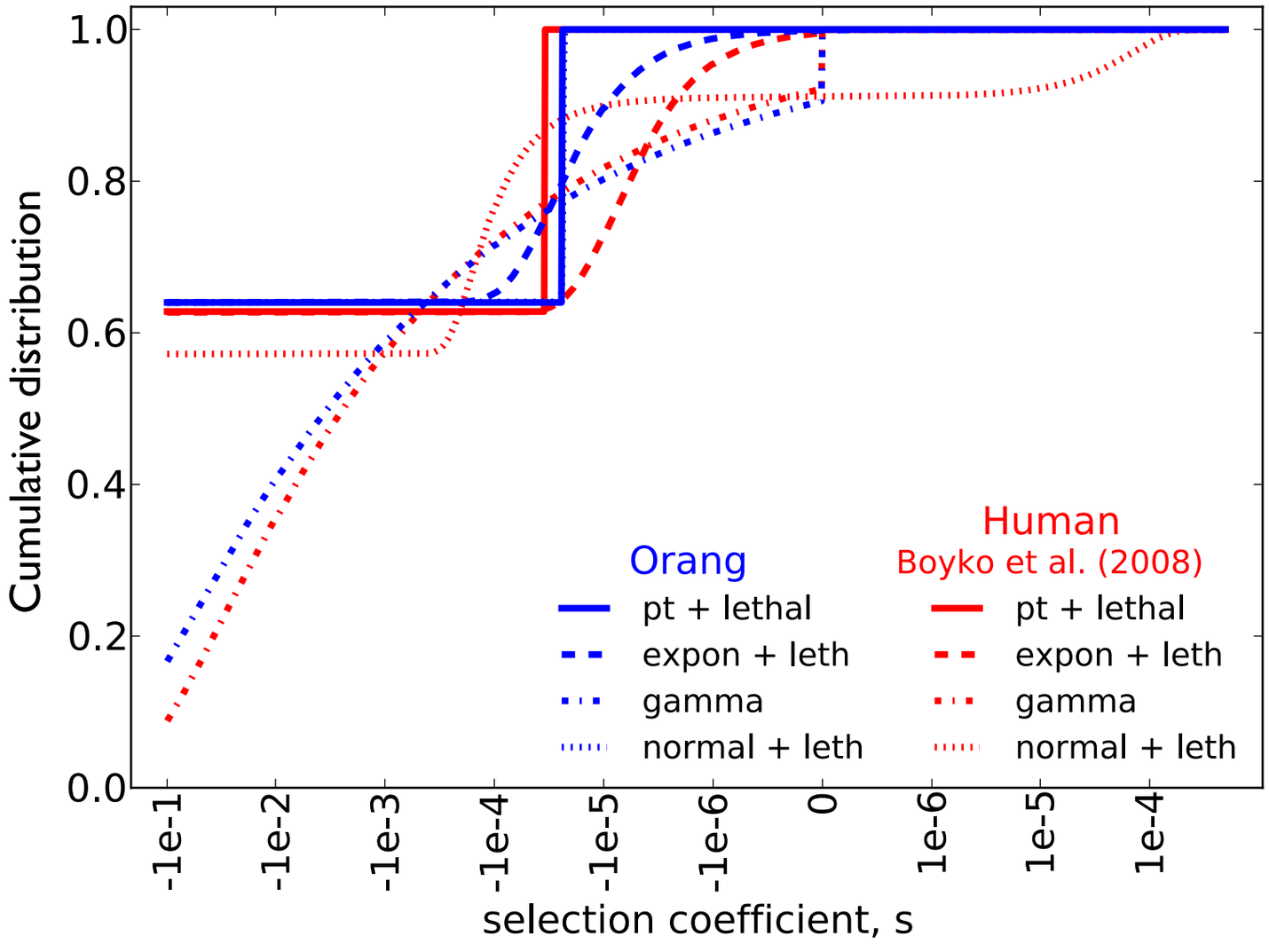
6,434 synonymous SNPs



4,497 nonsynonymous SNPs



Fit distribution of selection coeff for newly arising nonsynon. mutations



(No clear difference between Bornean and Sumatran orangs)

Ma et al.
PLoS ONE (2013)

History and selection in African pygmies

Ping-Hsun Hsieh

Collaborators
Michael Hammer
Sarah Tishkoff
Krishna Veeramah

Hsieh et al.
(In press)
Genome Research

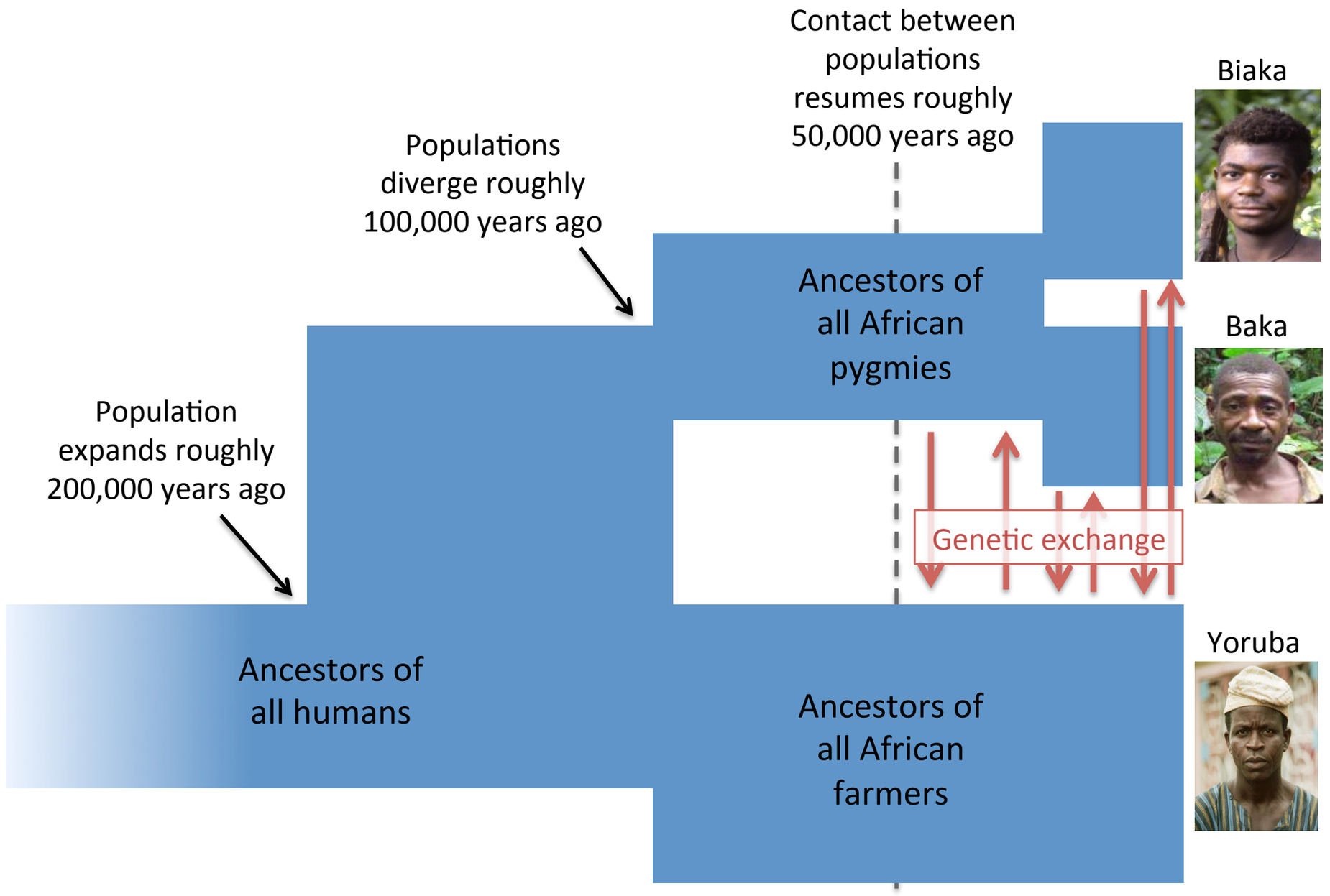
Data and demographic history

Whole Genome Data

Biaka pygmy: 4 individuals from Hammer lab

Baka pygmy: 3 individuals from Tishkoff lab

Yoruba farmer: 9 individuals



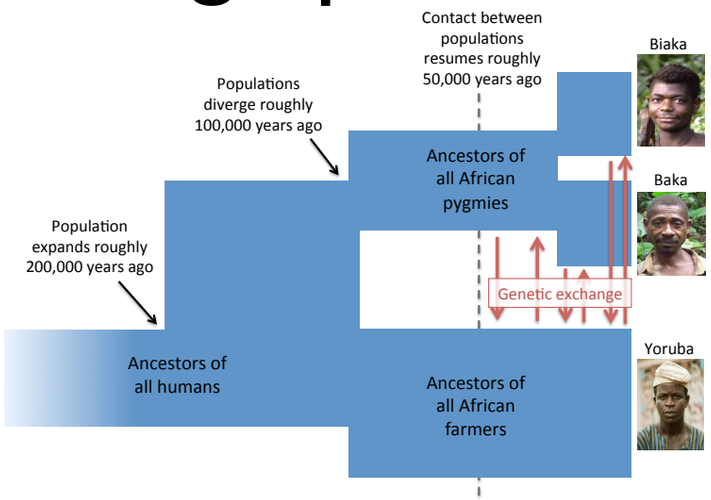
Inferred using ∂a∂i

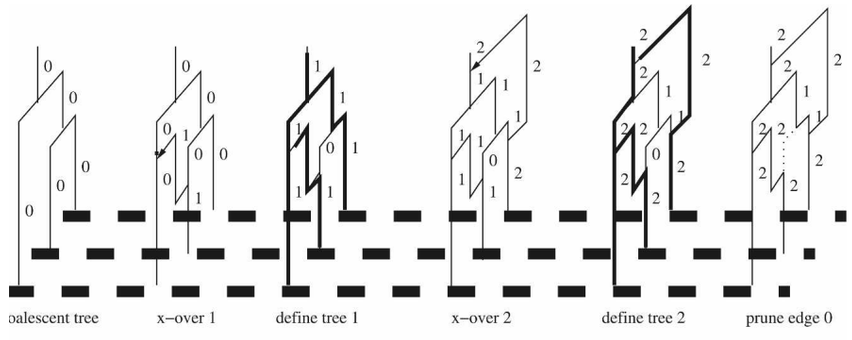
Gutenkunst et al.
(2009) PLoS Genet

Hsieh et al.
(In press)
Genome Research

Genomic null model

Demographic history

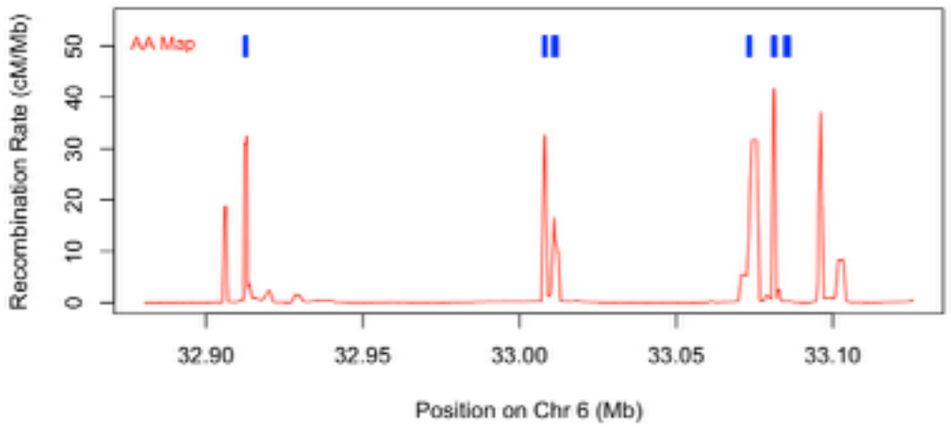




Simulate in MaCS

Chen, Marjoram & Wall
(2009) *Genome Research*

Recombination heterogeneity



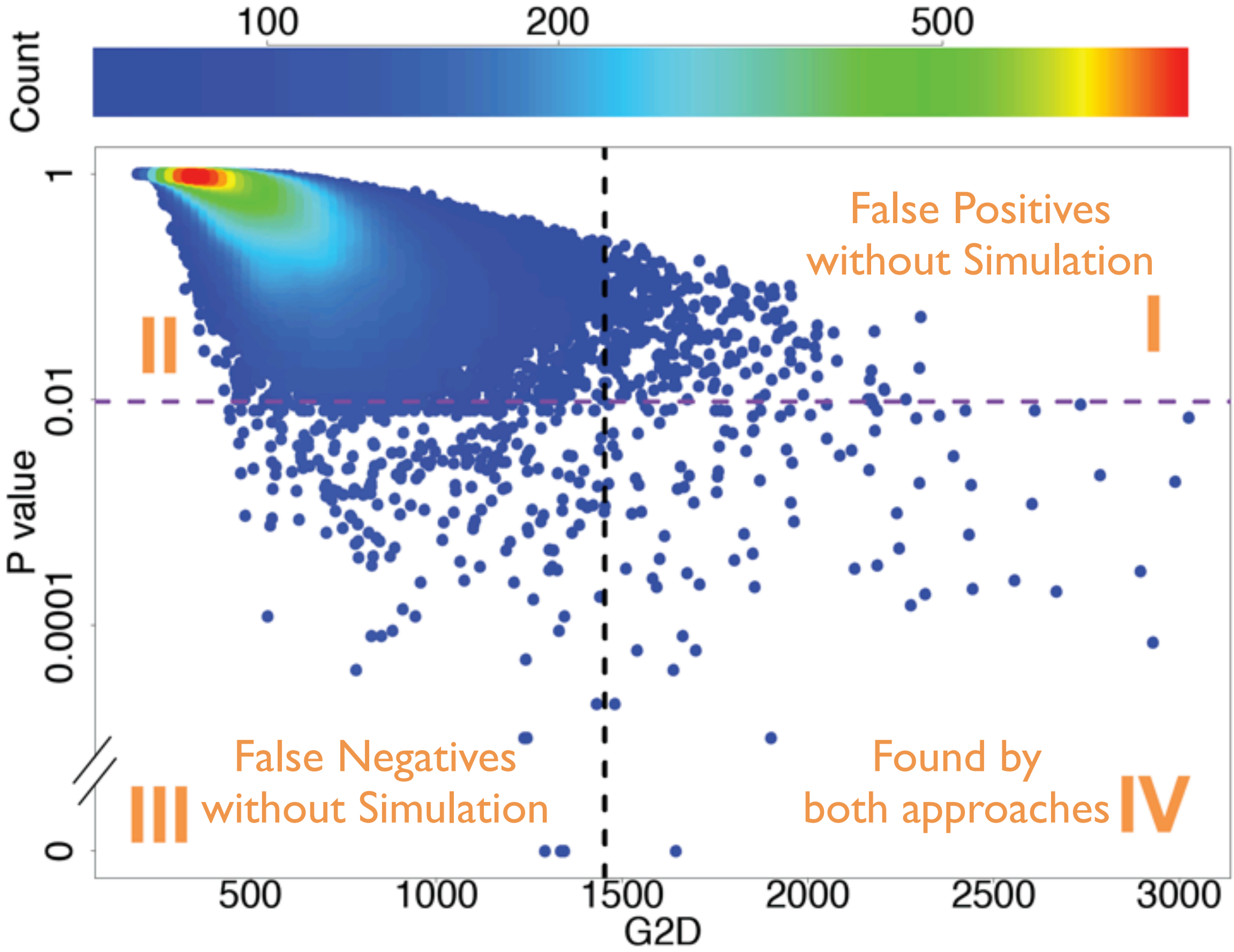
SNP density heterogeneity

Gene structure

Compare real data with neutral simulations to test for selection

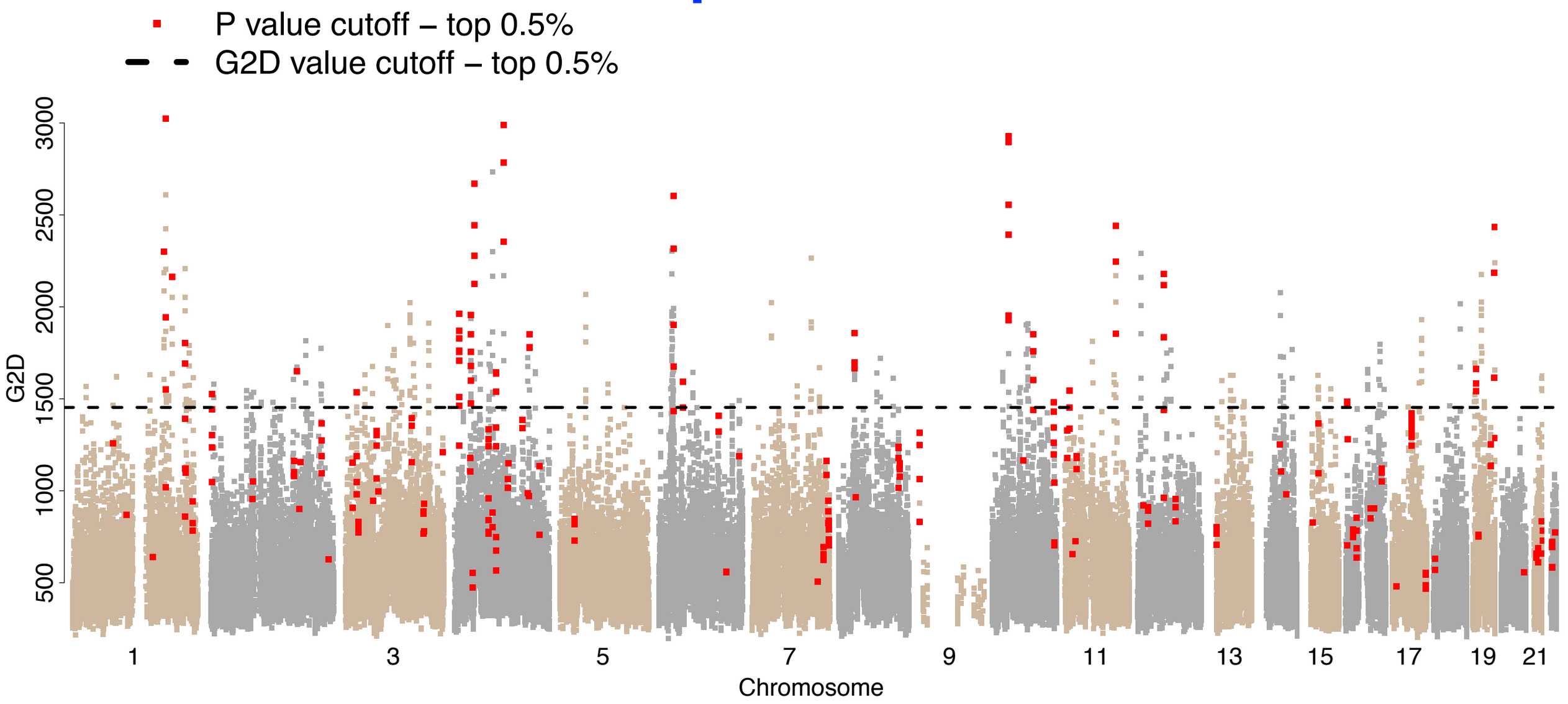
Hsieh et al.
(In press)
Genome Research

Scanning for adaptive loci



Hsieh et al.
(In press)
Genome Research

Adaptive loci



- Bone synthesis: FLNB, AXDND1, EPHB1, TSPAN5, ZBTB38, GAREM
- Muscle development: OBSCN, COX10, LARGE
- Immunity: 3 HLA genes

Demographic Inference

Ryan Gutenkunst
Molecular and Cellular Biology
University of Arizona

Name	Data type	Inference	Notes	Refs
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User-friendly GUI; can be computationally demanding	32
FRAPPE	Unlinked bi-allelic SNVs	Population structure, admixture	Alexander *et al.*[41] argue that convergence is not guaranteed	40
ADMIXTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Estimates the number of populations via cross-validation error	41
fastSTRUCTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Obtains variational Bayesian estimates of posterior probability distribution	42
Structurama	Unlinked multi-allelic genotypes	Population structure, admixture	Uses a Dirichlet process to estimate the number of populations	43
HAPMIX	Phased haplotypes; reference panel	Chromosome painting	Requires populations to be specified a priori	48
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	49
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Extends the fineSTRUCTURE approach to estimate unsampled ancestral populations and admixture times	7
LAMP	Phased haplotypes; reference panel	Chromosome painting	Identifies local ancestry in windows, rather than using an HMM, so is more discrete than other approaches	52
PCAdmix	Phased haplotypes	Chromosome painting, population structure	Uses PCA in small chunks followed by an HMM to estimate local ancestry	53
dadi	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Requires some Python-coding skills; applicable to up to three populations	60
Fastsimcoal2	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Can also be used to simulate data under the SMC	62,63
Treemix	Frequencies of unlinked bi-allelic SNVs	Admixture graph	Highly multimodal likelihood surface and heuristic search; redo inference from many starting points	64
fastNeutrino	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Applicable only to a single population; designed specifically for extremely large sample sizes	65
DoRIS	Lengths of IBD blocks between pairs of individuals	Demographic history	IBD must be inferred (for example, using Beagle or GERMLINE); specification of lower cut-off minimizes false-negative IBD tracts	71,72
IBS tract inference	Lengths of IBS blocks between pairs of individuals	Demographic	IBS can easily be confounded by missing data and/or sequencing errors	76
PSMC	Diploid genotypes from one individual	Demographic history	Best used in MSMC's PSMC mode, which uses the SMC to more accurately model recombination than the original PSMC; applicable to a single population	78
MSMC	Whole genome, phased haplotypes	Demographic history	Requires large amounts of RAM; cross-coalescence rate should not be interpreted as migration rate	82
CoalHMM	Whole genome, phased haplotypes	Demographic history	Multiple applications, including inference of population sizes, migration rates and incomplete lineage sorting	83–87
diCal	Medium-length, phased haplotypes	Demographic history	Uses shorter sequences than MSMC, but can be applied to multiple individuals in complex demographic models; infers explicit population genetic parameters for migration rates	89,92
LAMARC	Short, phased haplotypes	Demographic history	Requires Monte Carlo sampling of coalescent genealogies; very flexible	93
BEAST	Short, phased haplotypes	Species trees, effective population sizes	Used mainly as a method of phylogenetic inference. Can also infer population size history	94
MCMCcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP[131]	95
G-PhoCS	Short, (un)phased haplotypes	Demographic history	Incorporates migration into the MCMCcoal framework. Averages over unphased haplotypes	96
Exact likelihoods using generating functions	Short, phased haplotypes	Demographic history	Implemented in Mathematica; applicable only to specific classes of multi-population models	97,98