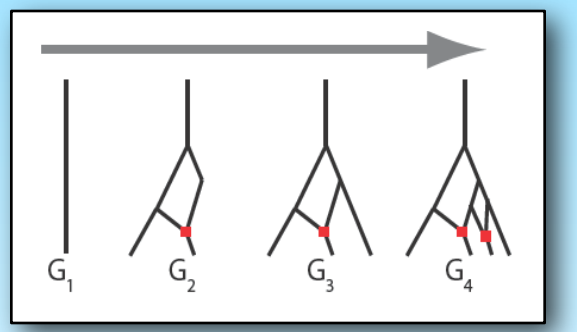
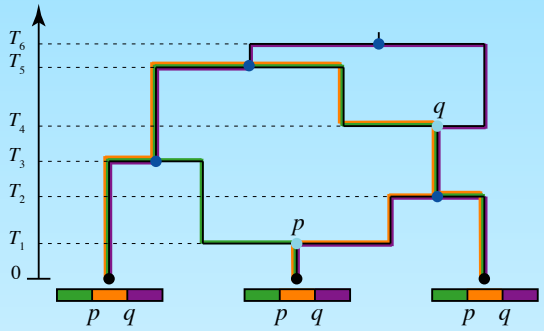
ARGweaver: Genome-wide Inference of Ancestral Recombination Graphs





Adam Siepel
Biological Statistics & Computational Biology
Center for Comparative and Population Genomics
Cornell University, Ithaca, NY

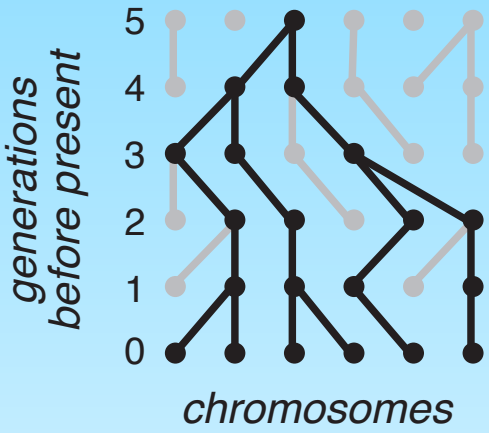
Joint work with Matthew Rasmussen, Melissa Hubisz, and Ilan Gronau

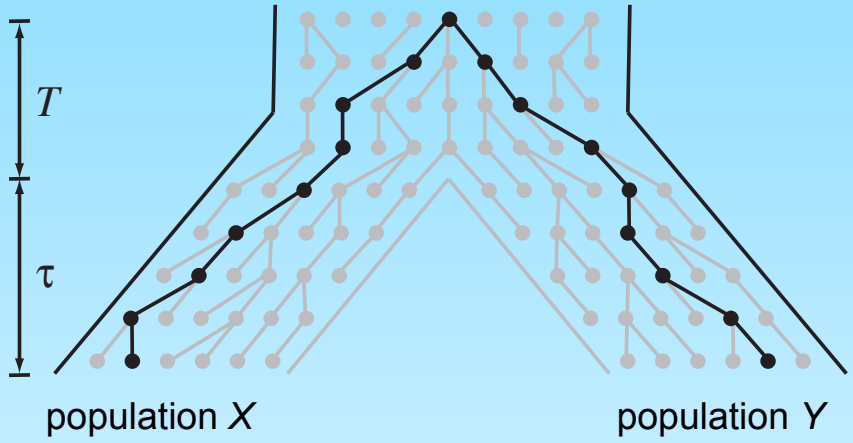
Comparative Genomics

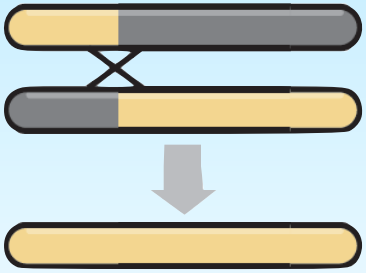
Comparative Genomics

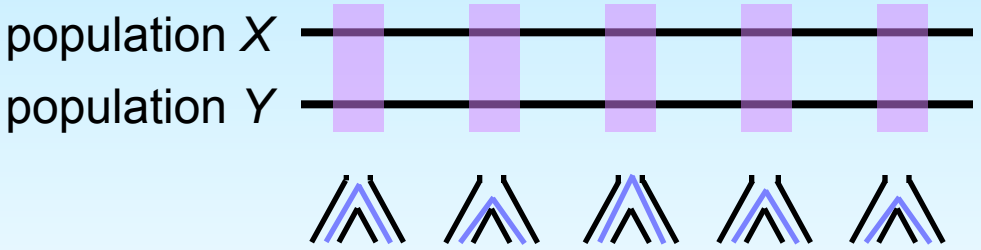
Evolution of Individual Human Genomes

Recombination and Genealogies

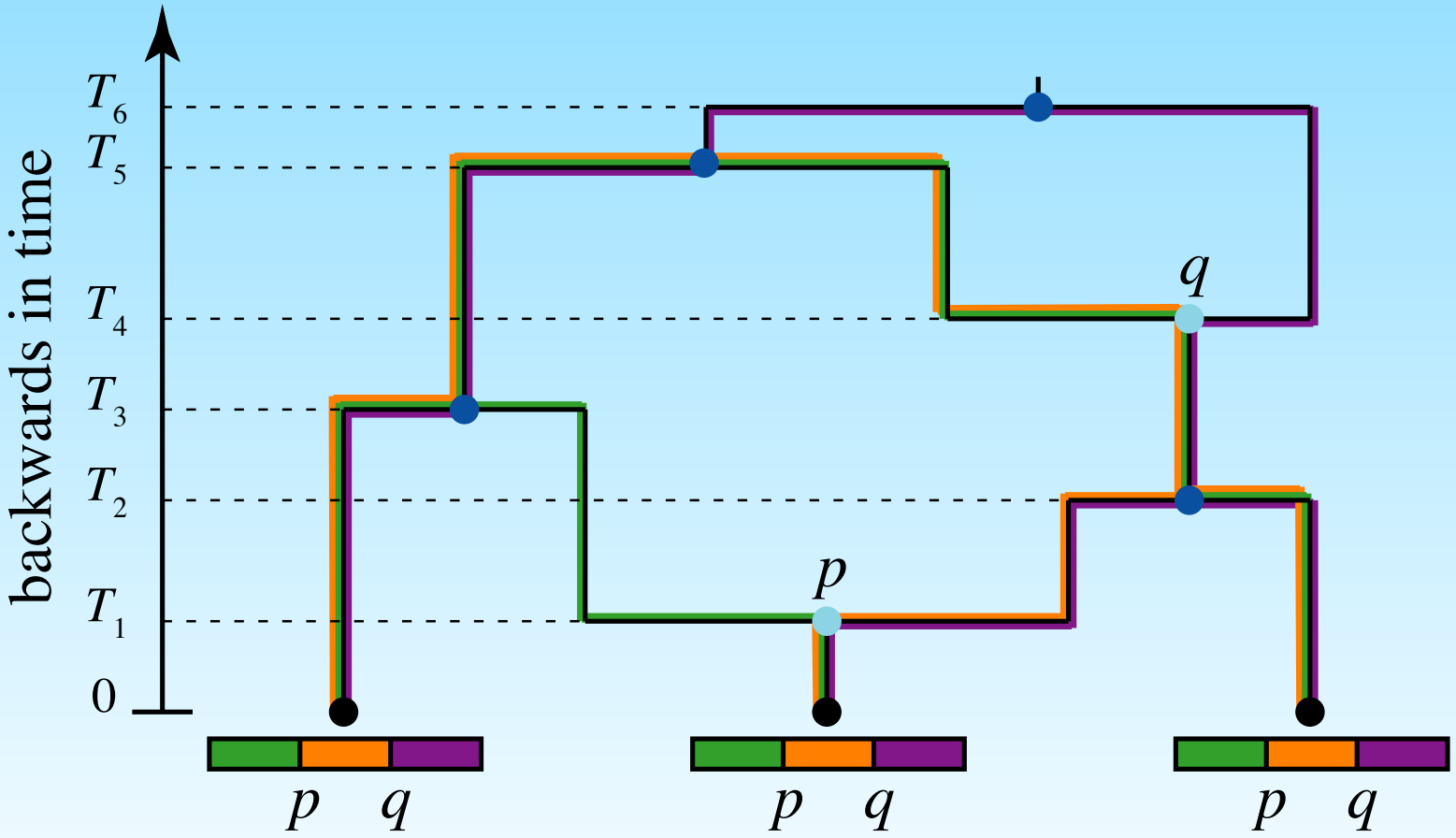








The Ancestral Recombination Graph



Griffiths and Marjoram, *J. Comput. Biol.*, 1996

Alas, if only we knew the ARG…

- Demography inference
- Inference of natural selection
- Recombination rate estimation
- Phasing/imputation
- Association mapping
- …

Alas, if only we knew the ARG…

- Demography inference
- Inference of natural selection
- Recombination rate estimation
- Phasing/imputation
- Association mapping
- …

ARG surrogates: IBD, IBS, haplotypes, local ancestry inference, site-frequency spectrum, PCA

Explicit ARG Inference

- **Importance sampling**
  - Griffiths and Marjoram, *J. Comput. Biol.*, 1996
  - Fearnhead and Donnelly, *Genetics*, 2001
- **Markov chain Monte Carlo sampling**
  - Kuhner, Yumato, and Felsenstein, *Genetics*, 2000
  - Nielsen, *Genetics*, 2000
- **Heuristics/Parsimony**
  - Hein, *J. Mol. Evol.*, 1993
  - Kececioglu and Gusfield, *Disc. Appl. Math.*, 1998
  - Song and Hein, *J. Comput. Biol.*, 2005
  - Minichiello and Durbin, *Am. J. Hum. Genet.*, 2006

Explicit ARG Inference

- Importance sampling
  - Griffiths and Marjoram,
  - Fearnhead and Donnelly,
- Markov chain Monte Carlo sampling
  - Kuhner, Yamato and Felsenstein,
  - Nielsen
- Heuristics/Parsimony
  - Hein,
  - Kececioglu and Gusfield,
  - Song and Hein,
  - Minichiello and Durbin,

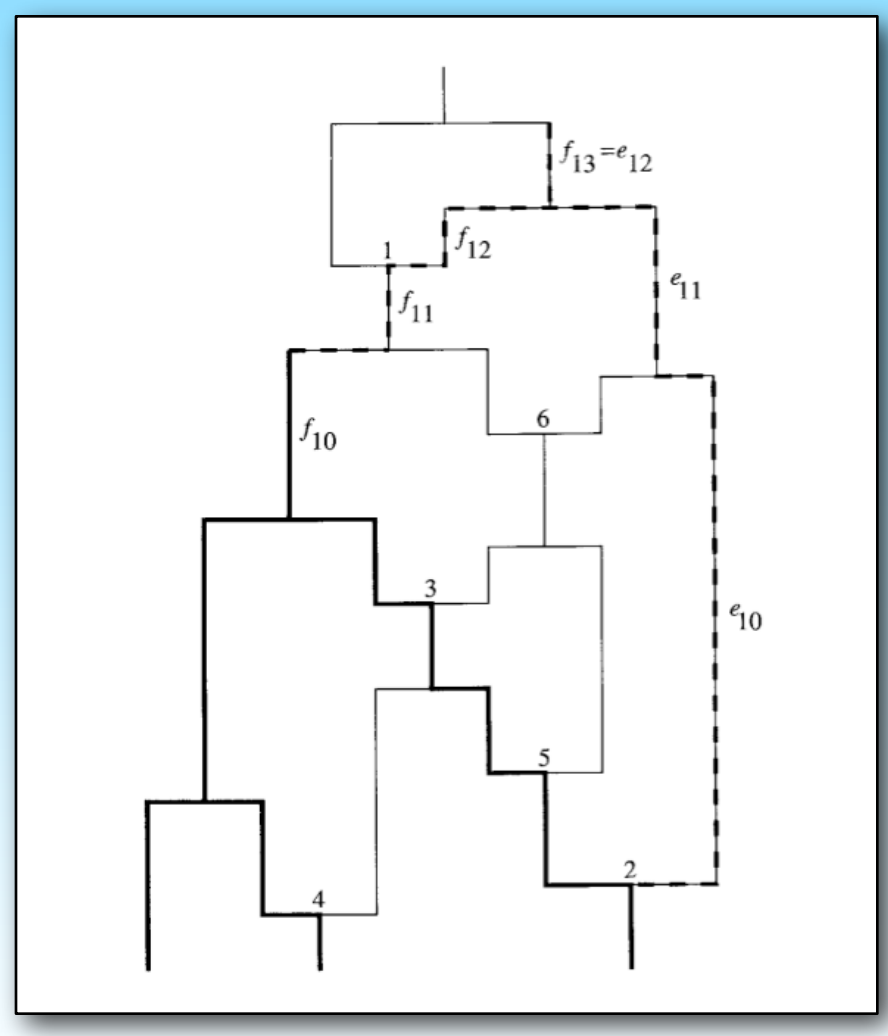
Computationally intensive

Limited to few samples

and/or

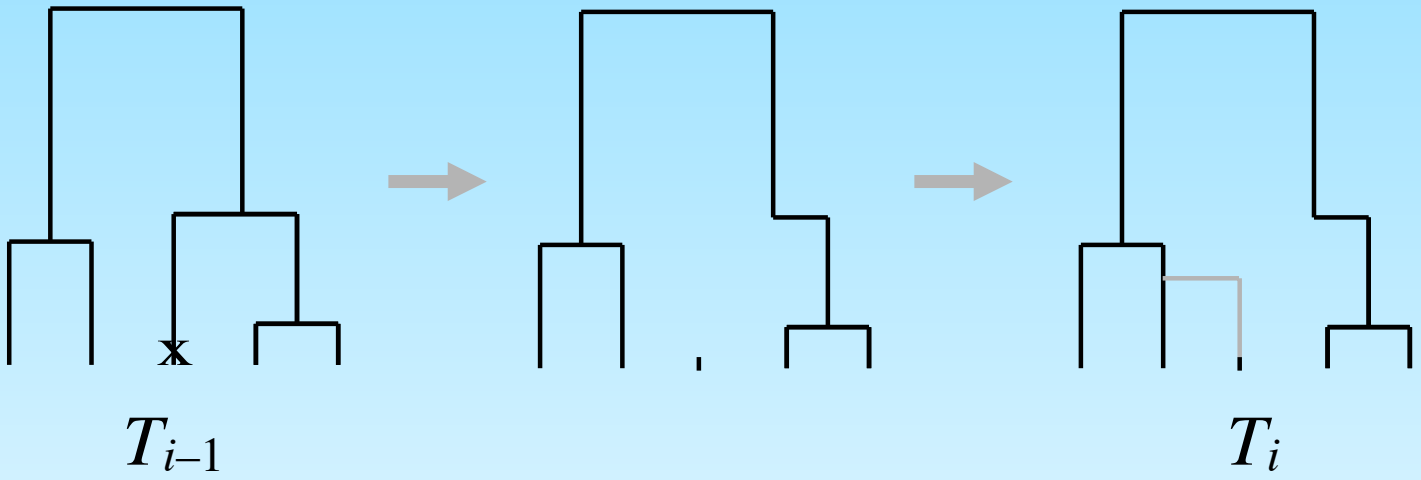
Depend on crude approximations

Sequential Coalescent with Recombination



Wiuf and Hein, *Theor. Popul. Biol.*, 1999

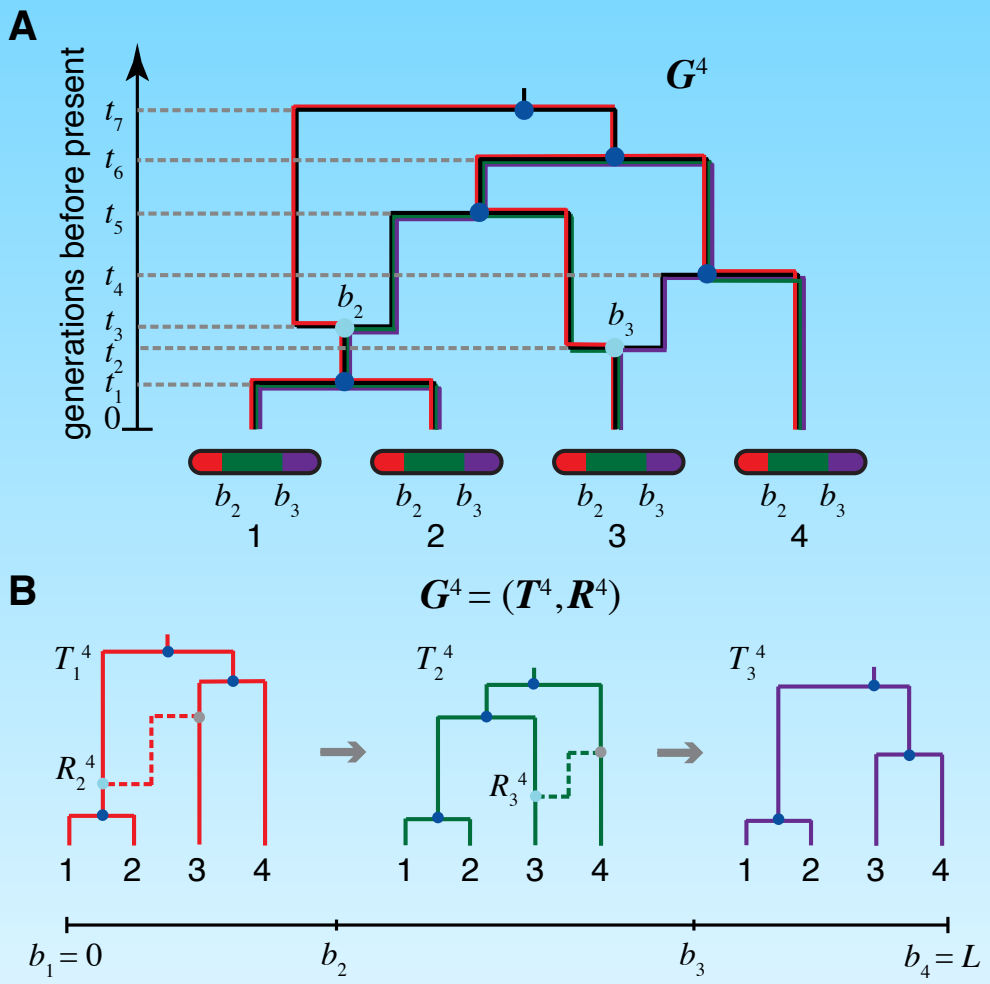
Sequentially Markov Coalescent (SMC)



$$P(T_i \mid T_1, \ldots, T_{i-1}) = P(T_i \mid T_{i-1})$$

$$T_{i-1} \perp T_{i+1} \mid T_i$$

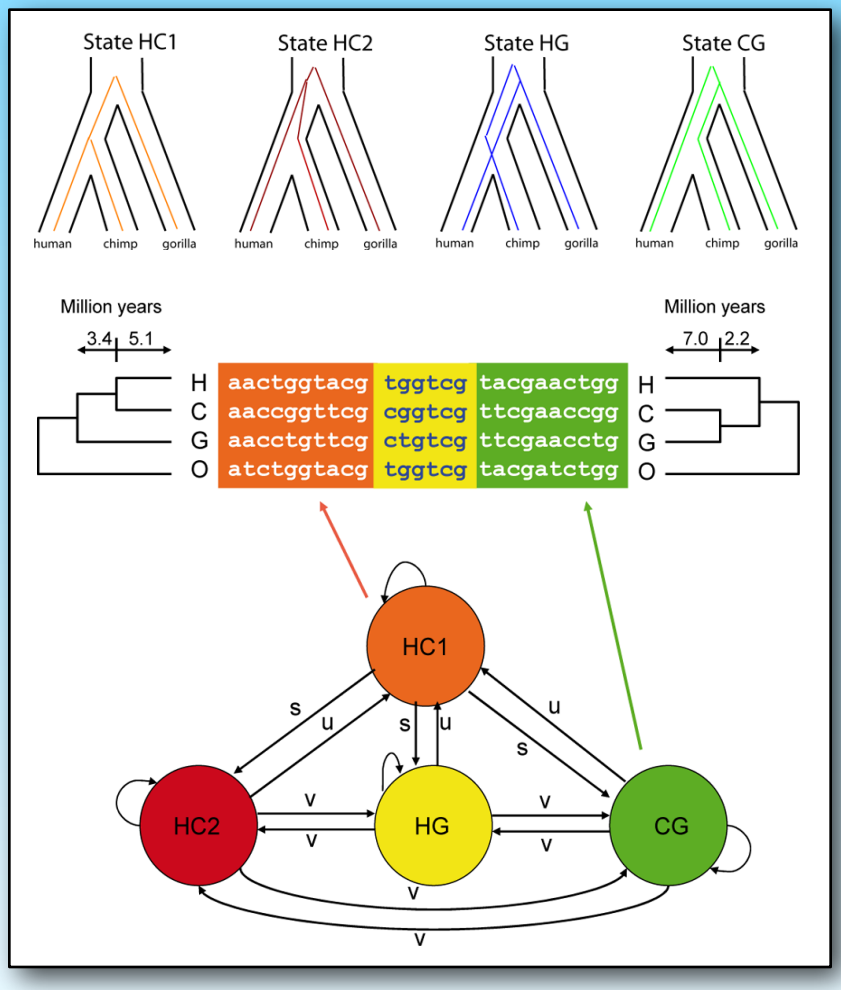
McVean and Cardin, 2005; Marjoram and Wall, 2006


A
G^4
generations before present
t_7
t_6
t_5
t_4
t_3
t_2
t_1
0
b_2
b_3
1
2
3
4
B
$G^4 = (T^4, R^4)$
T_1^4
R_2^4
T_2^4
R_3^4
T_3^4
$b_1 = 0$
b_2
b_3
$b_4 = L$
C
D^4
1 C G C G A
2 A C A C G A A
3 T G G T G T
4 T A A T G

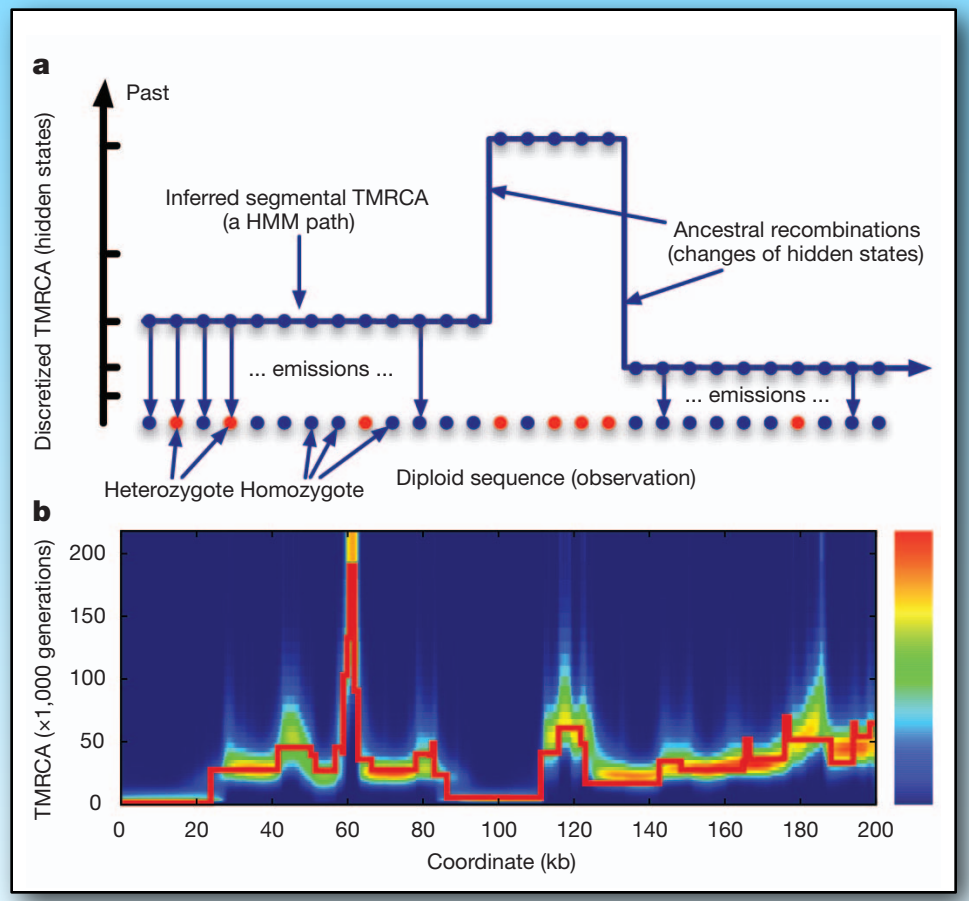
Discretized SMC and Hidden Markov Models

- By ***discretizing*** time and ***enumerating topologies***, the continuous state space of the SMC can be approximated by a finite set
- This opens up the possibility of using ***hidden Markov models*** (HMM) for inference
- Standard dynamic-programming algorithms for HMMs allow for ***exact ARG inference***, up to the SMC and discretization

Hidden Markov Models



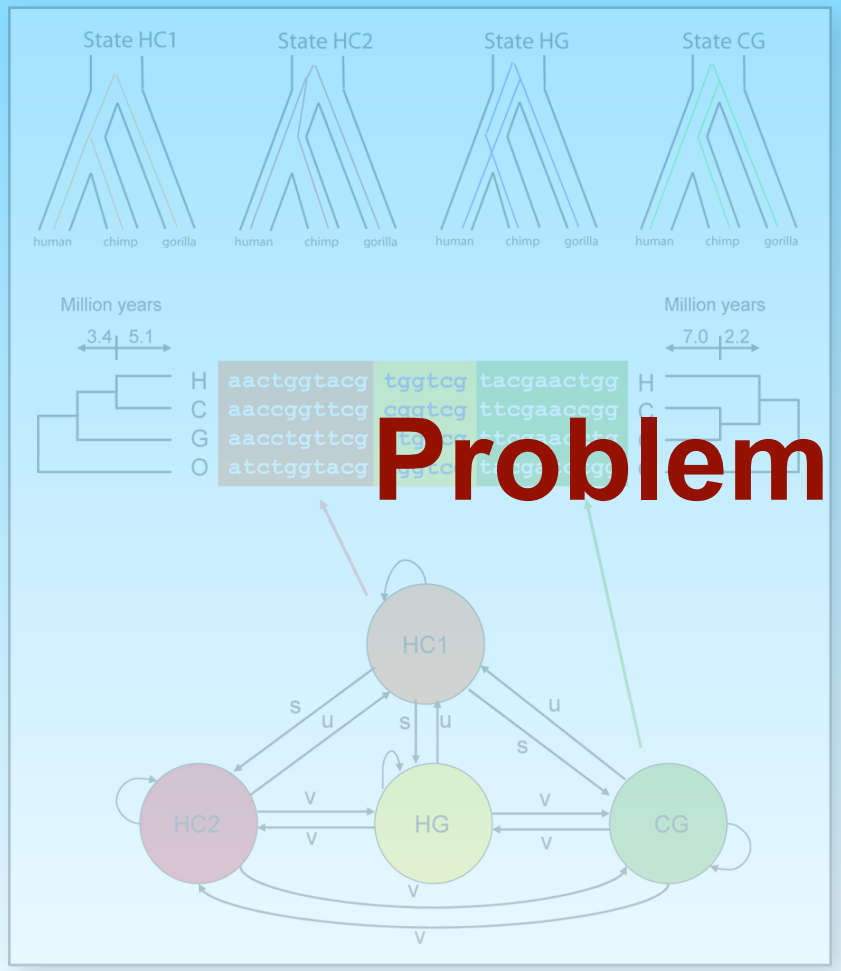
Coal-HMM



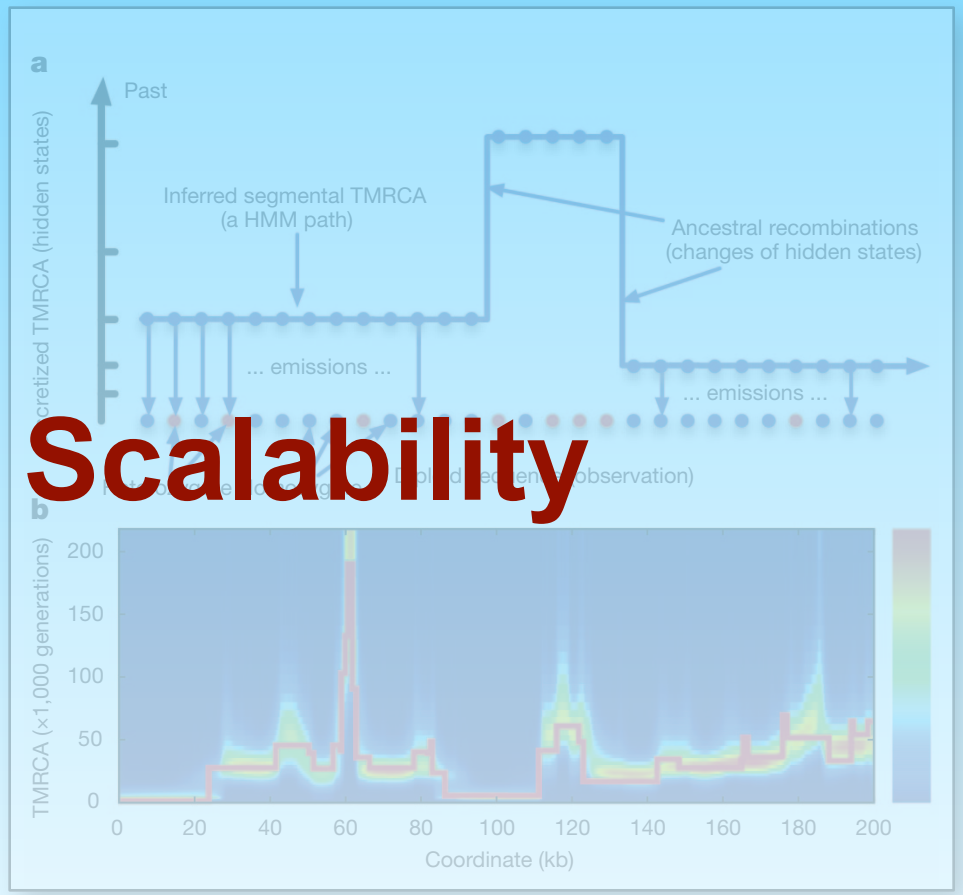
PSMC

Hobolth et al., *PLOS Genet.*, 2007;
Li and Durbin, *Nature*, 2011

Hidden Markov Models



Coal-HMM



PSMC

Problem: Scalability

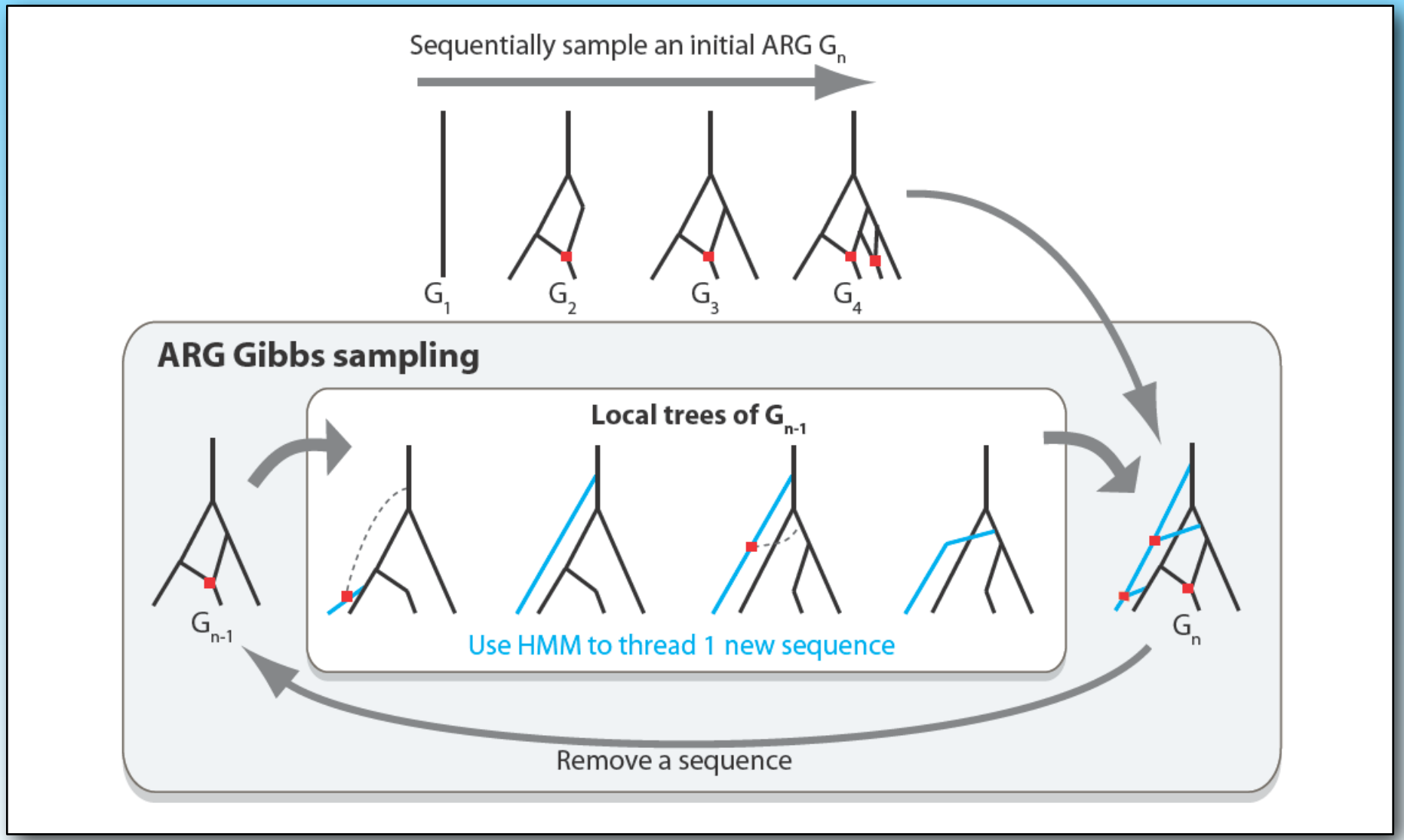
Hobolth et al., *PLOS Genet.*, 2007;
Li and Durbin, *Nature*, 2011

New Approach: Chromosome "Threading"

- Start with a data set of n sequences, D, and an ARG for n–1 of them, G_{n-1}
- Extend G_{n-1} to represent evolutionary history of n th sequence, obtaining G_n
- Sample this extension in a manner consistent with the conditional distribution, $P(G_n \mid G_{n-1}, D, \Theta)$, under the DSMC
- In repeated applications this operation is the basis of an ***ARG sampling*** algorithm

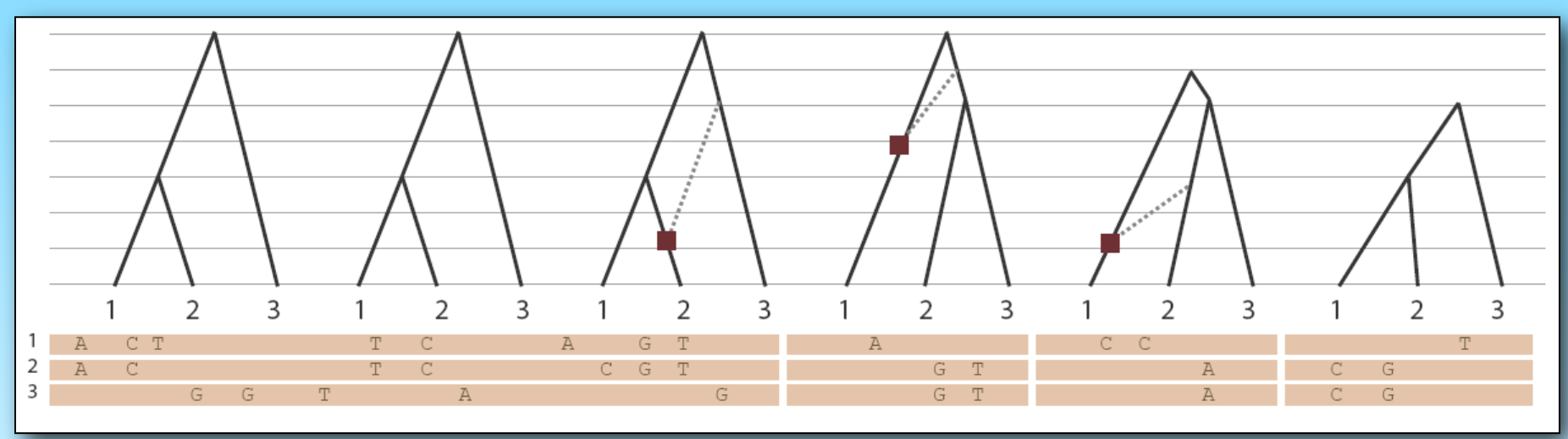
Rasmussen et al., *PLOS Genetics*, 2014

ARGweaver Sampling



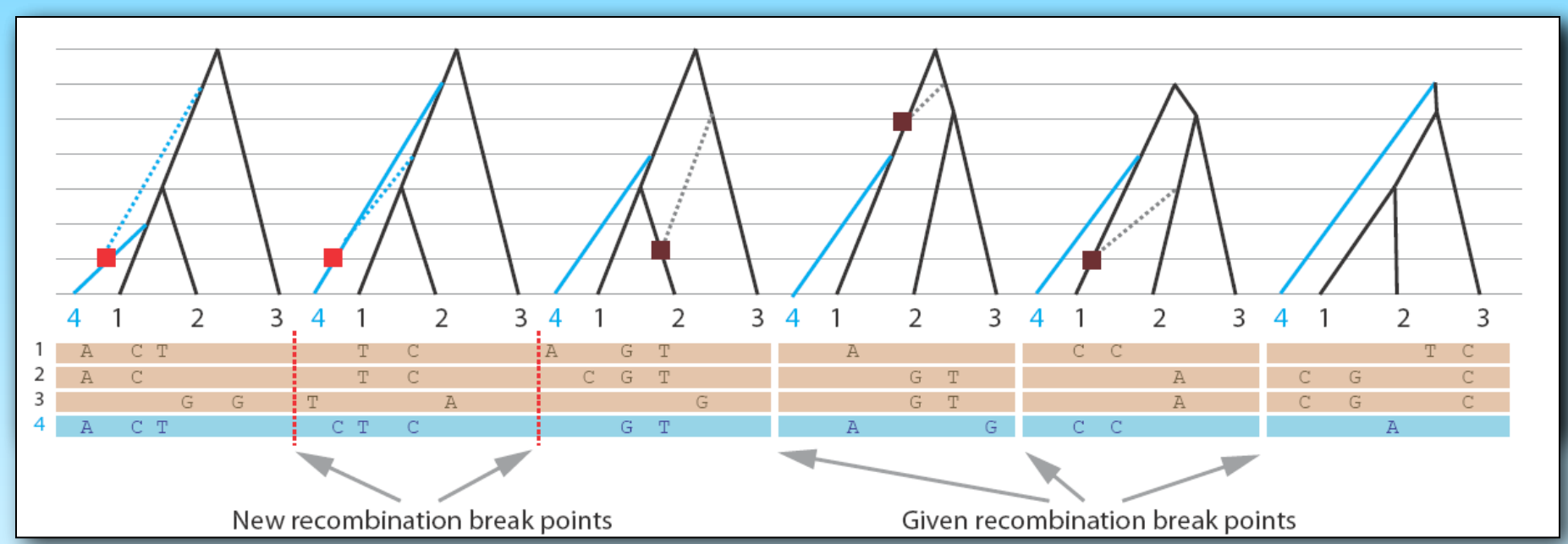
Rasmussen et al., *PLOS Genetics,* 2014

Threading



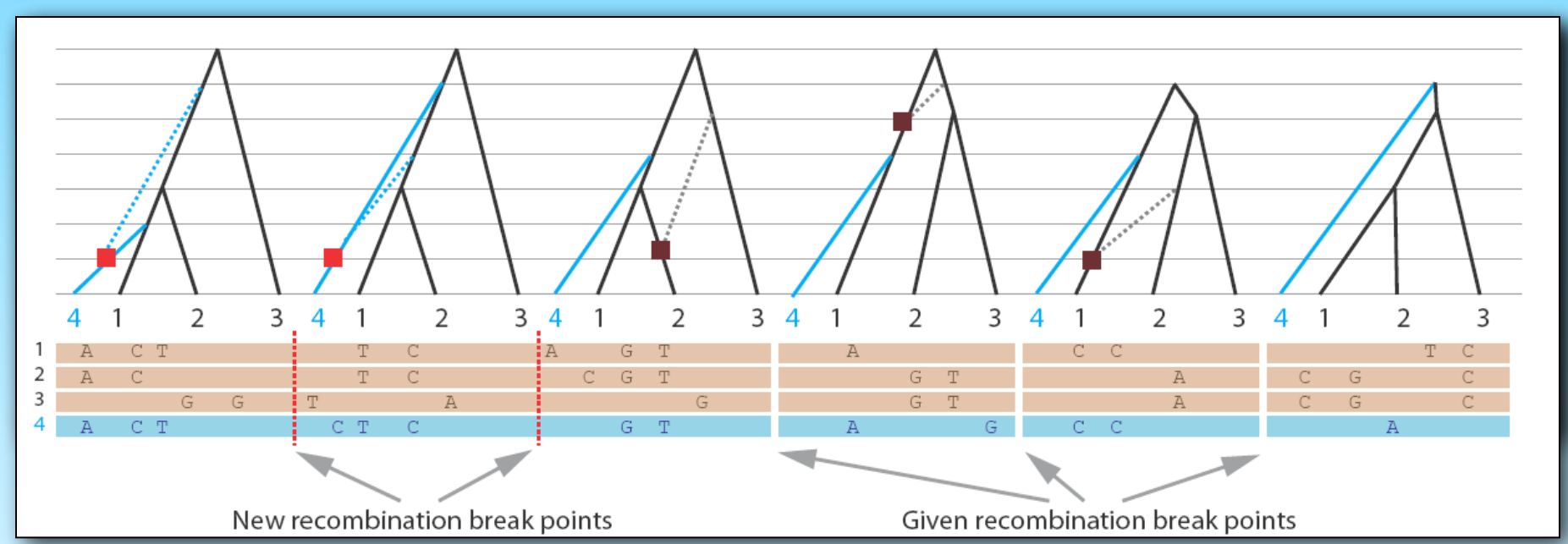
Rasmussen et al., *PLOS Genetics,* 2014

Threading



Rasmussen et al., *PLOS Genetics,* 2014

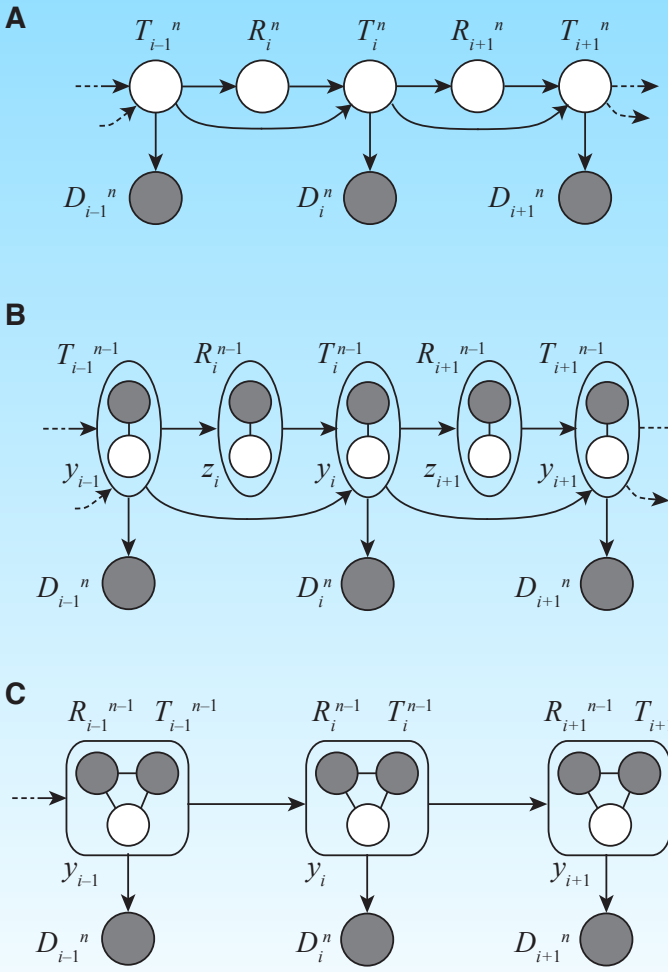
Threading



Solution: stochastic traceback with HMM

Rasmussen et al., *PLOS Genetics,* 2014

Graphical Models



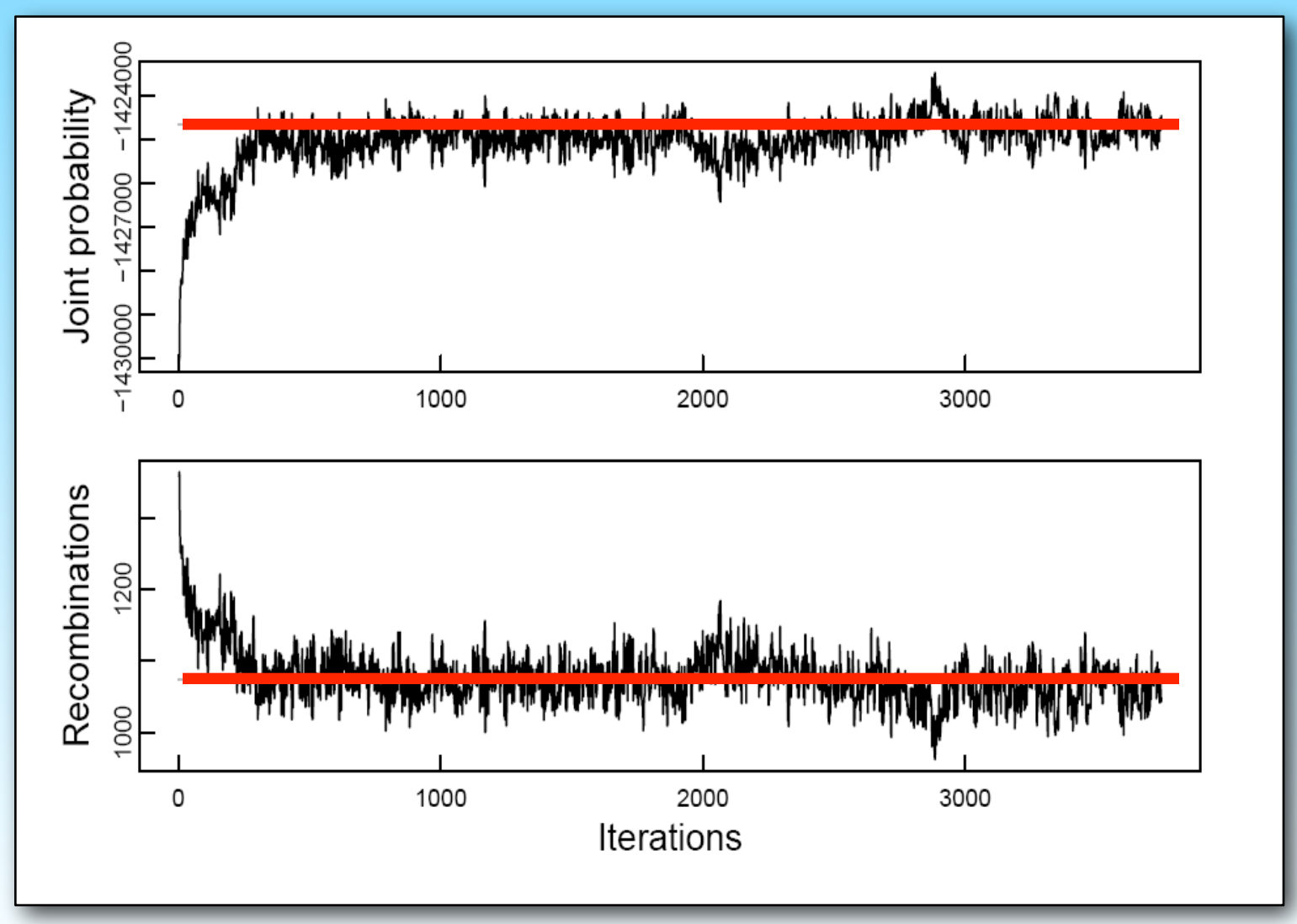
SMC (discretized)

Threading

Reduced Threading

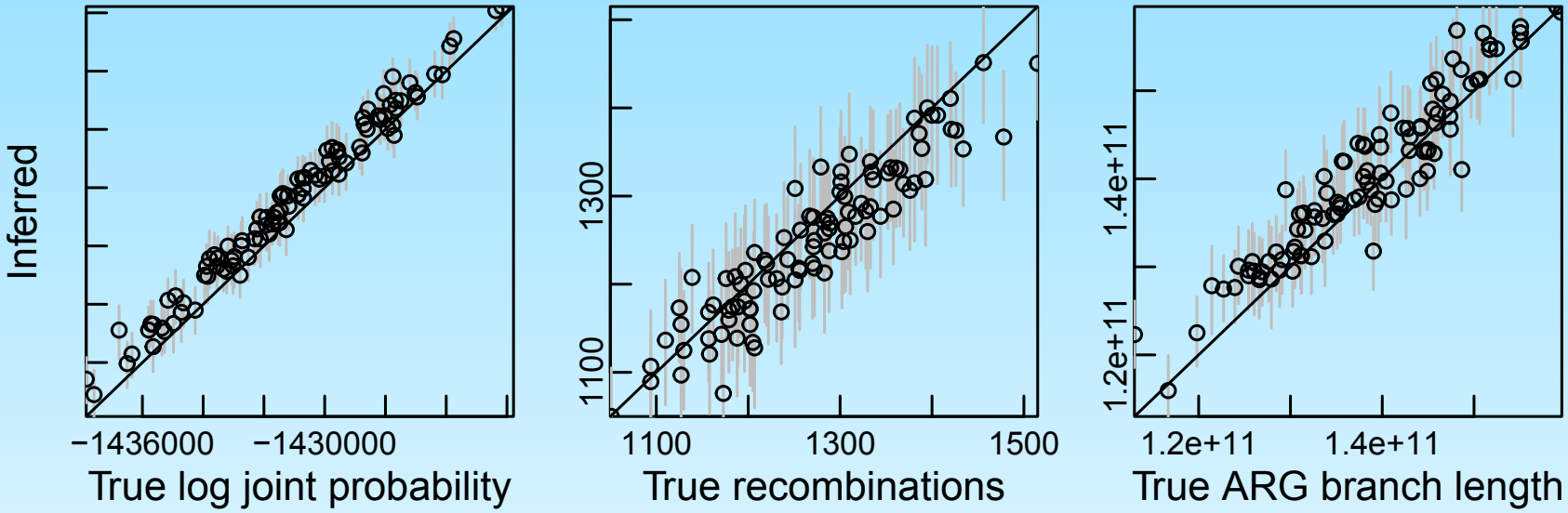
Rasmussen et al., *PLOS Genetics,* 2014

ARGweaver Converges Quickly



Rasmussen et al., *PLOS Genetics,* 2014

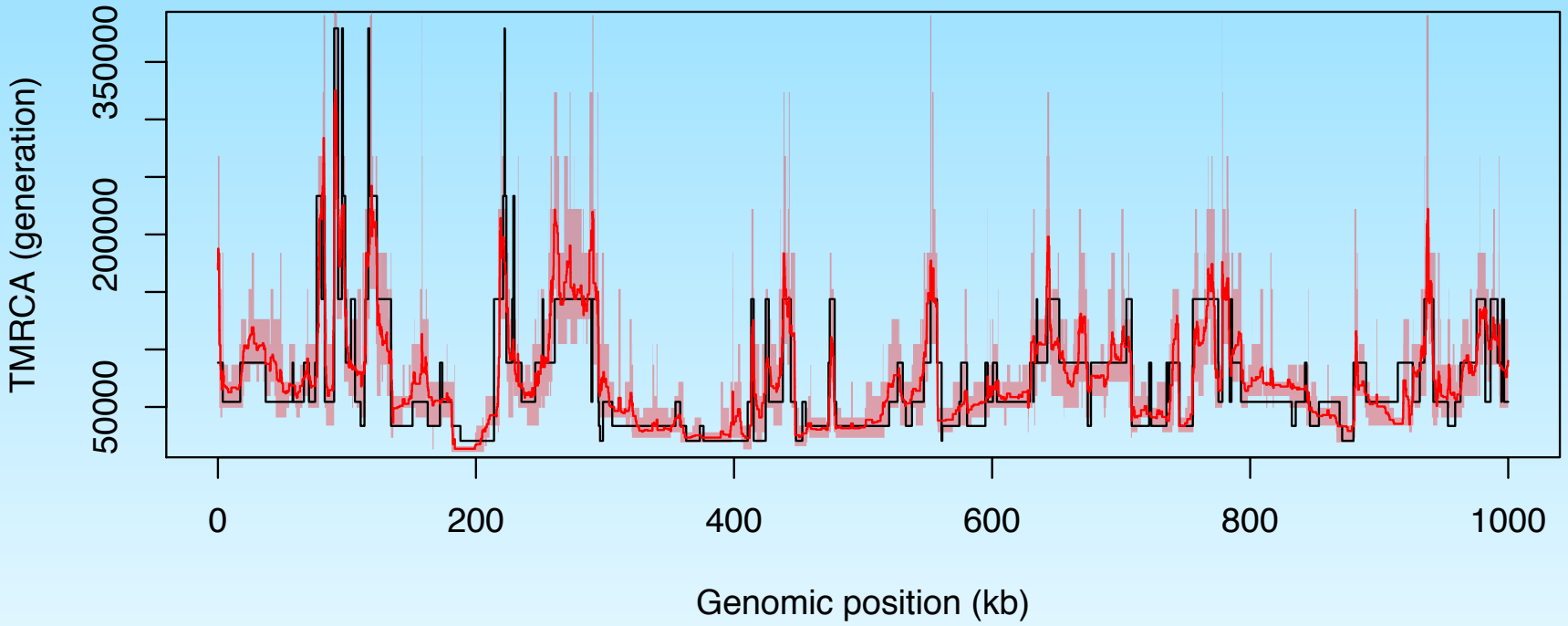
Recovery of Features of Simulated ARGs



Rasmussen et al., *PLOS Genetics,* 2014

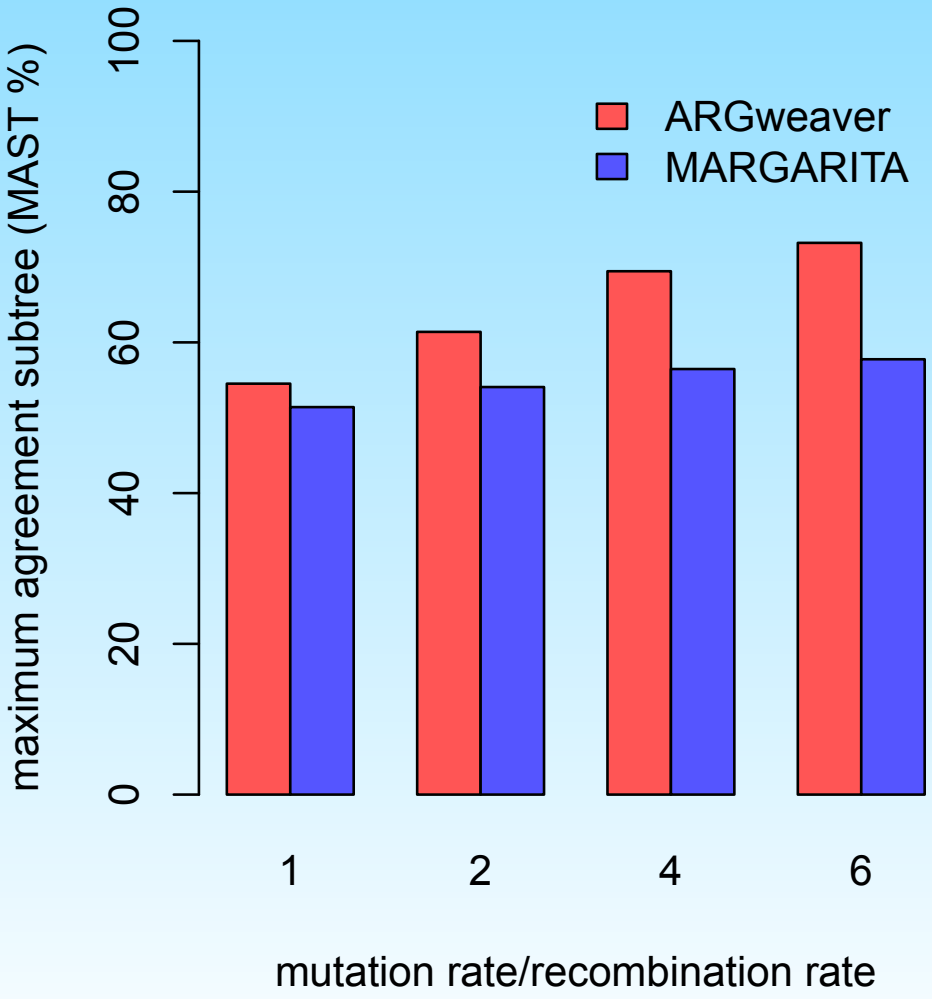
Part 2: *ARGweaver* Results

Recovery of Times to Most Recent Common Ancestry



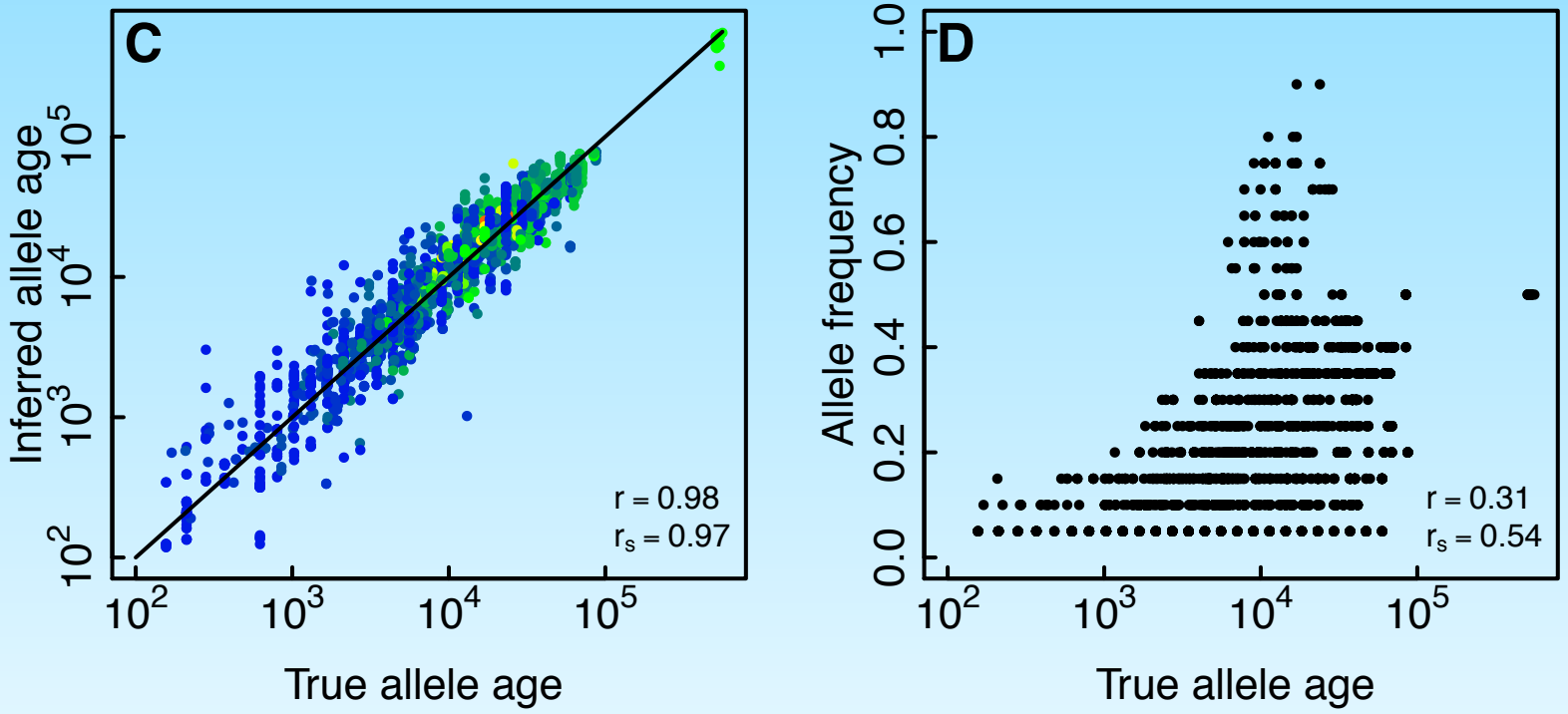
Rasmussen et al., *PLOS Genetics,* 2014

Accuracy of Inferred Trees



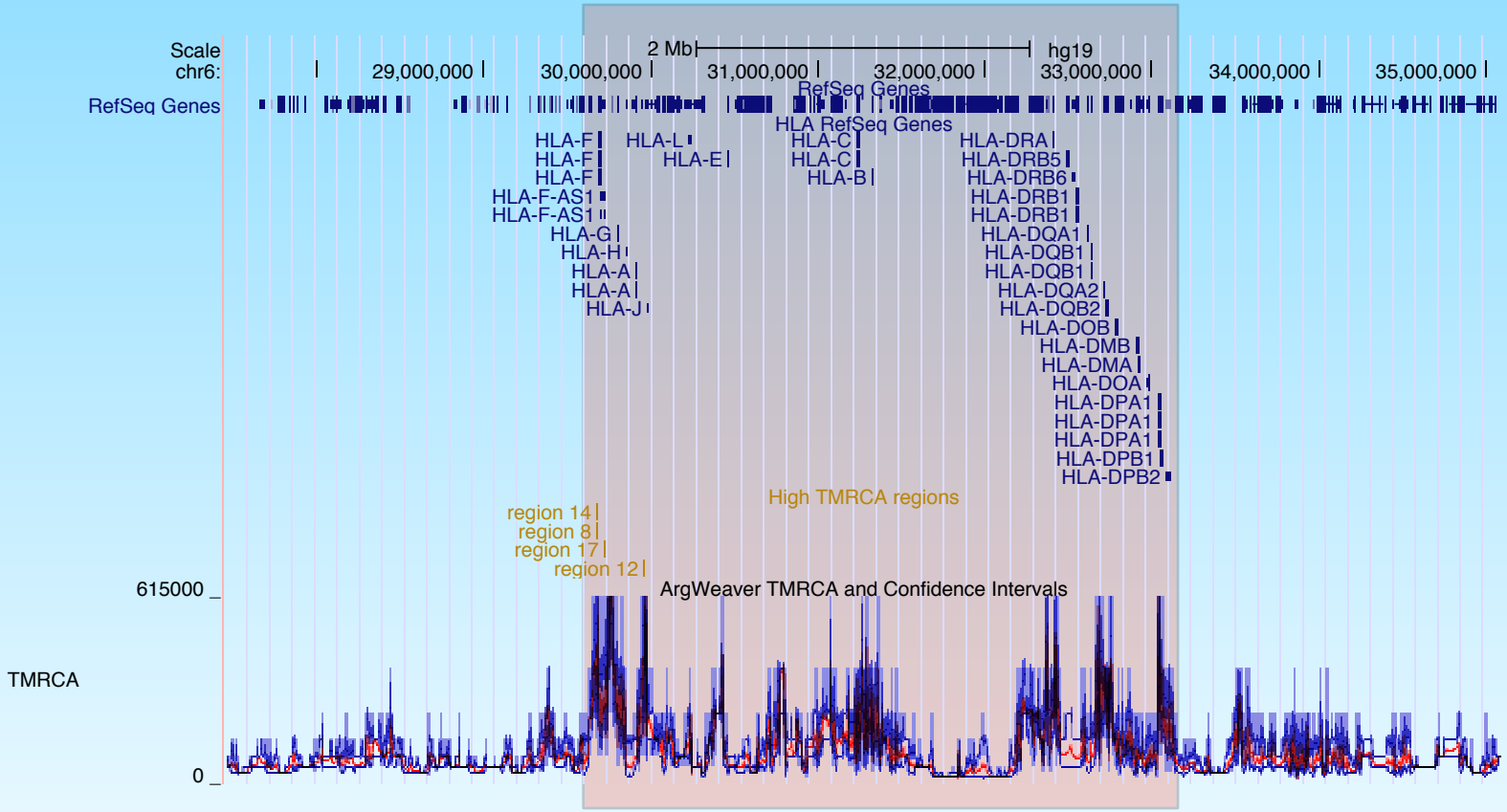
Rasmussen et al., *PLOS Genetics,* 2014

Recovery of Allele Age



Rasmussen et al., *PLOS Genetics,* 2014

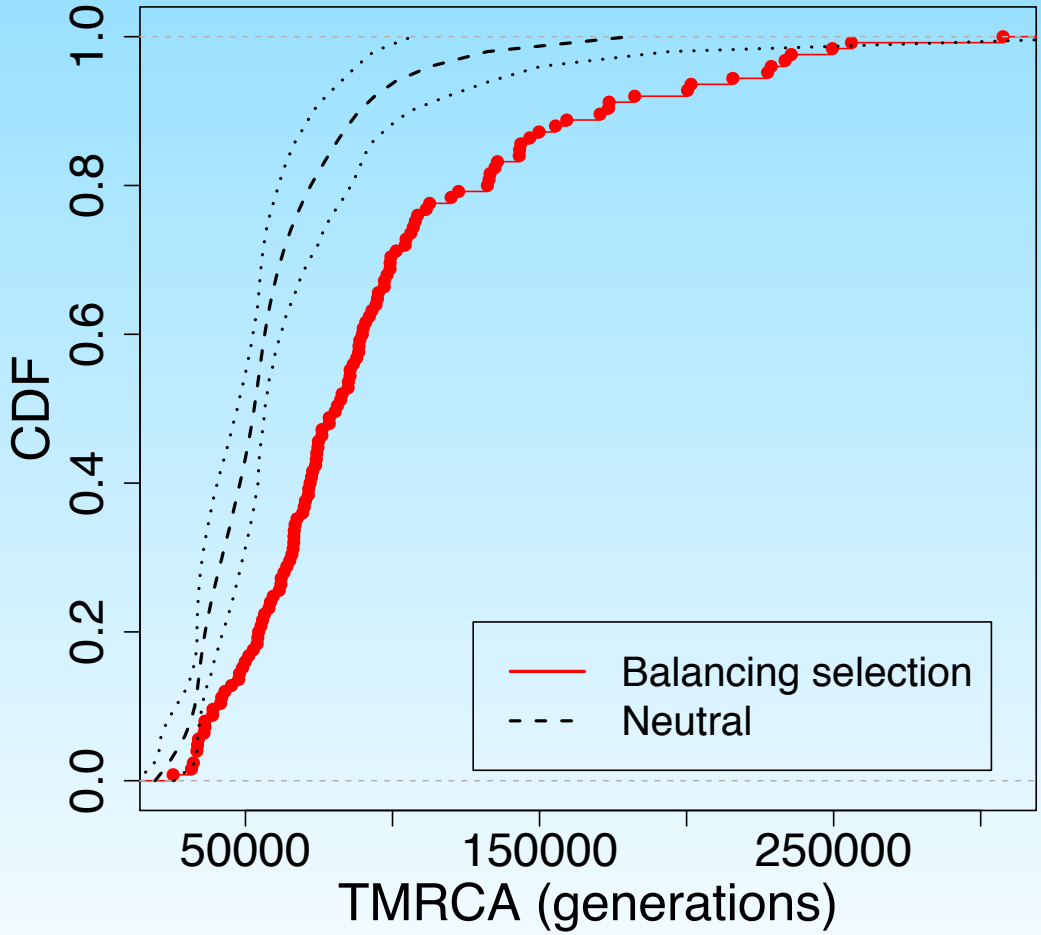
Real Data: Regions of High TMRCA



HLA Region

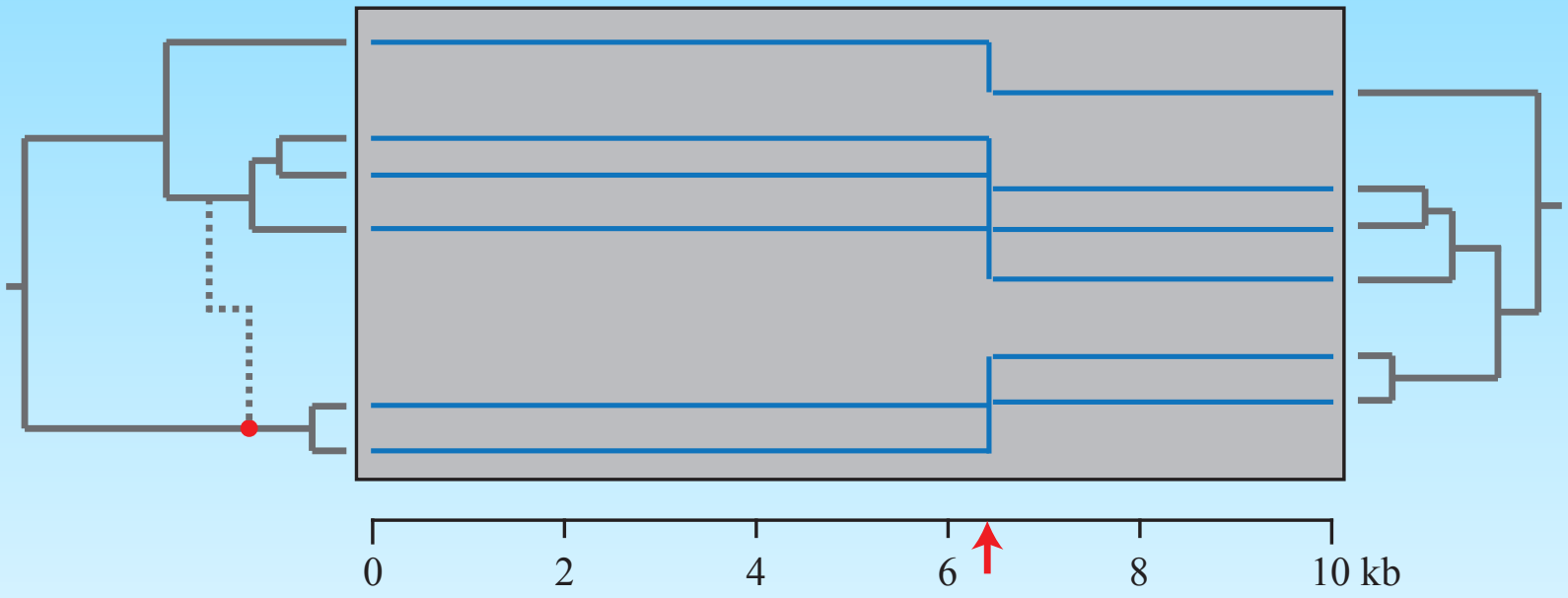
Rasmussen et al., *PLOS Genetics,* 2014

Regions of Shared Human/Chimp Polymorphism Have Old TMRCAs



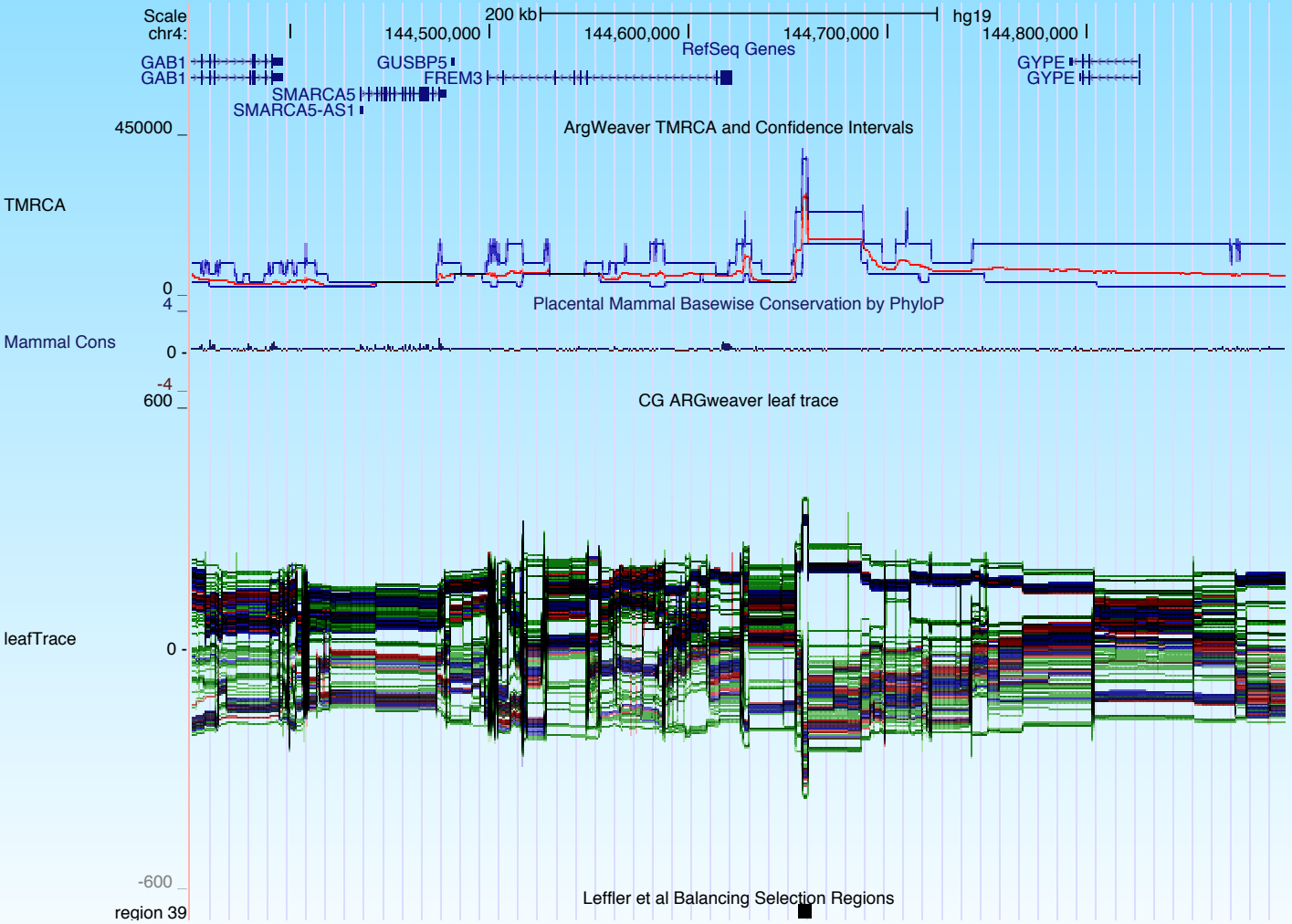
Leffler et al., *Science*, 2013

Leaf Trace



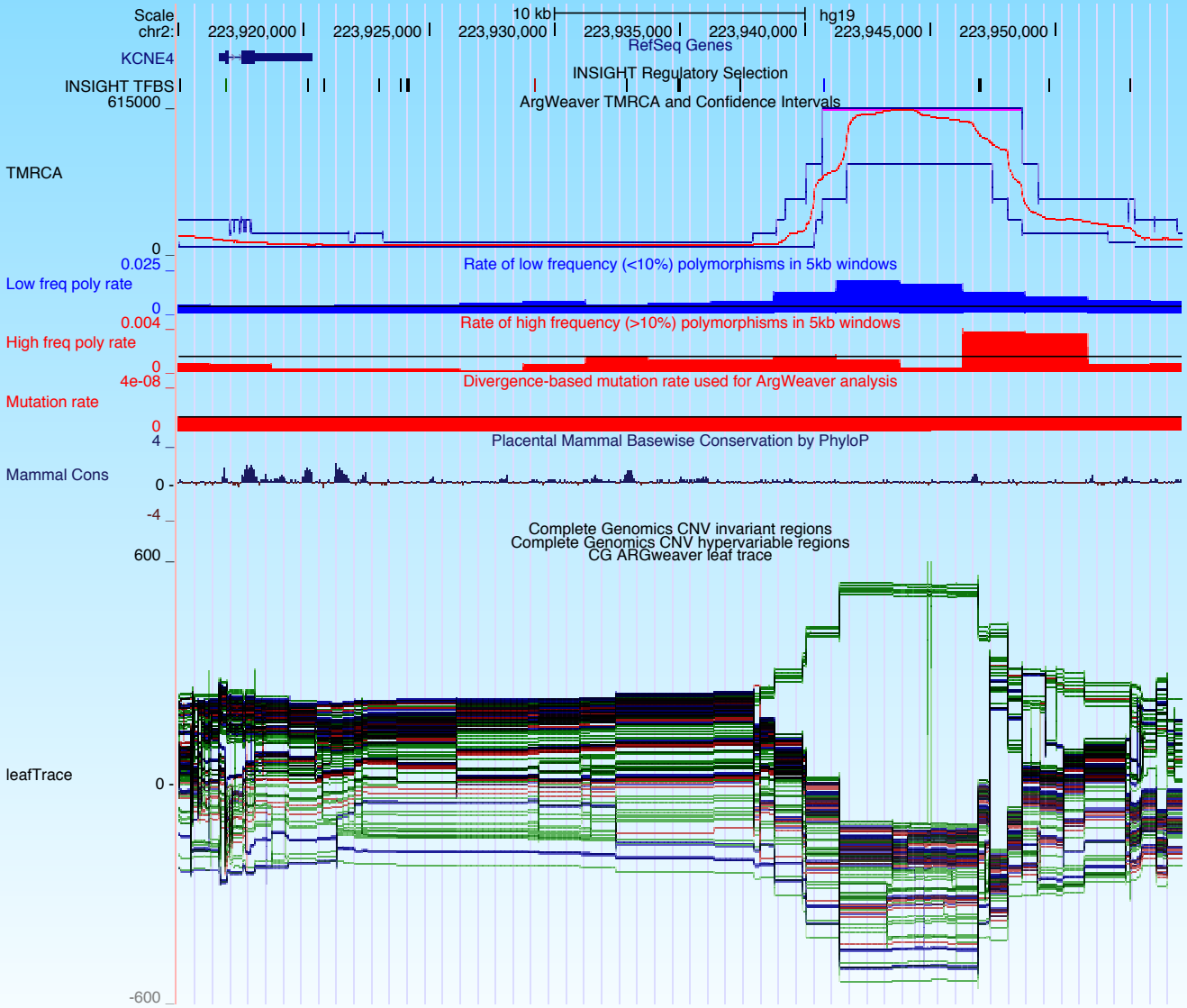
Rasmussen et al., *PLOS Genetics,* 2014

Putative Balancing Selection at FREM3



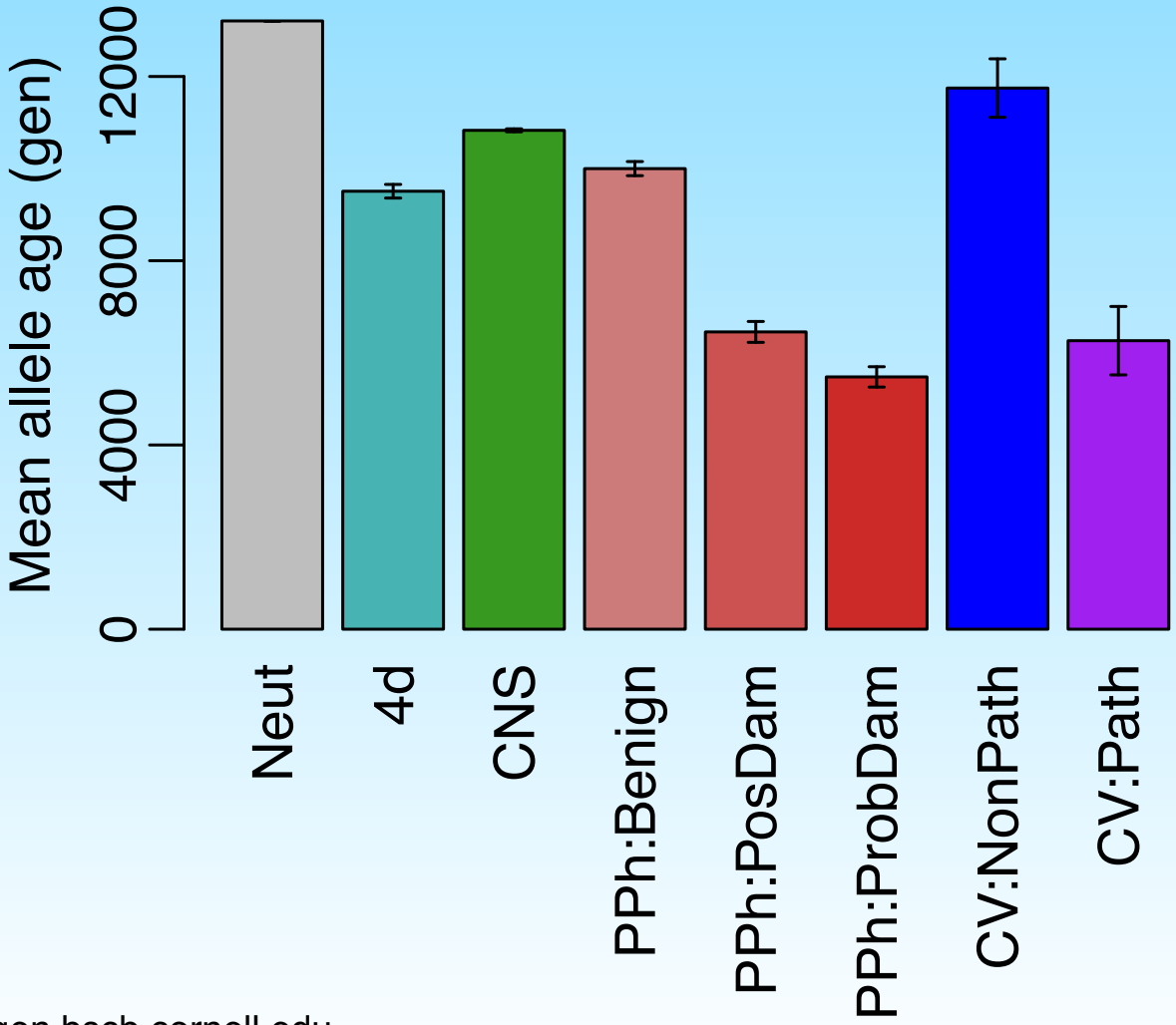
Leffler et al., *Science*, 2013

KCNE4 Promoter

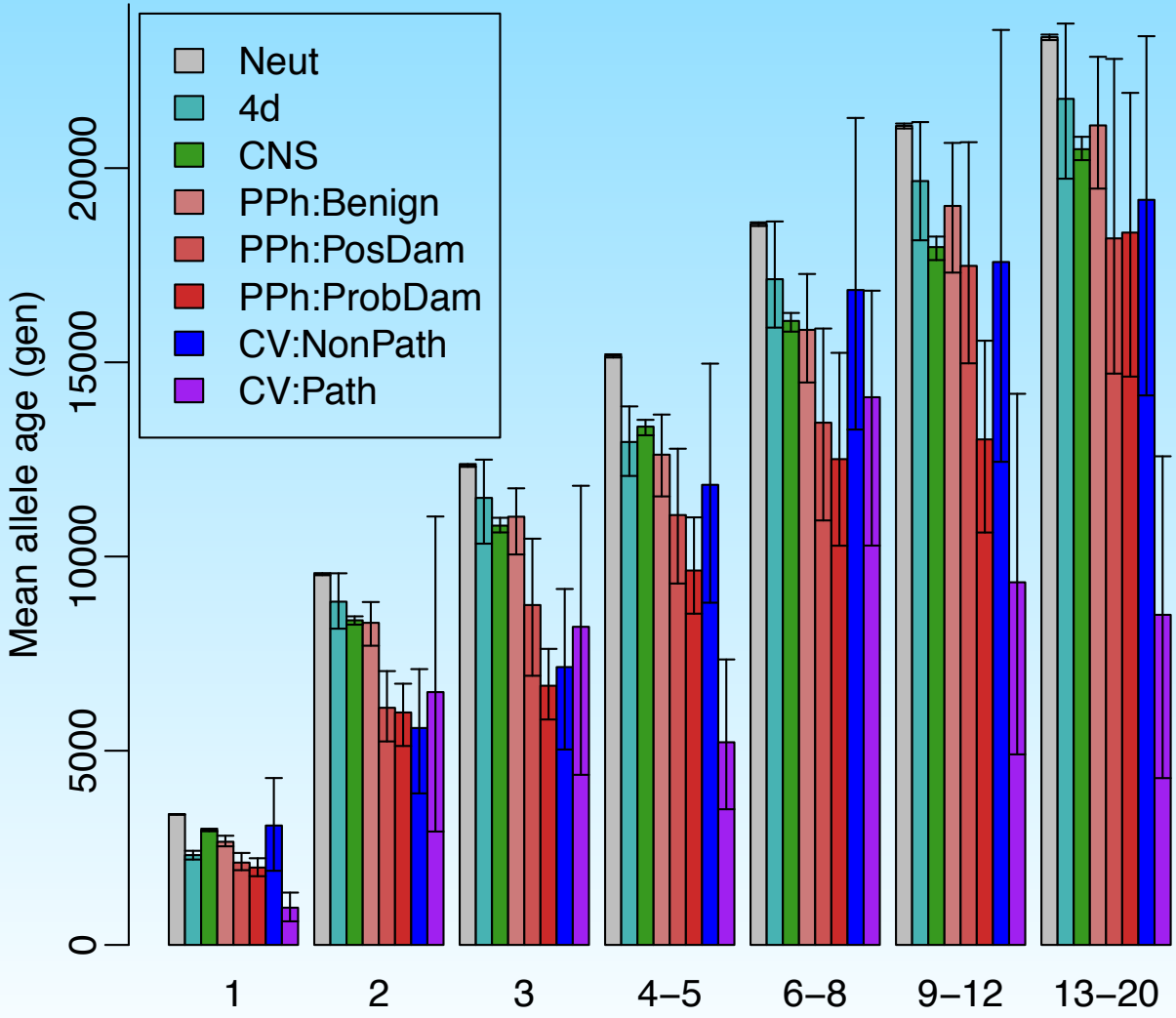


Rasmussen et al., *PLOS Genetics,* 2014

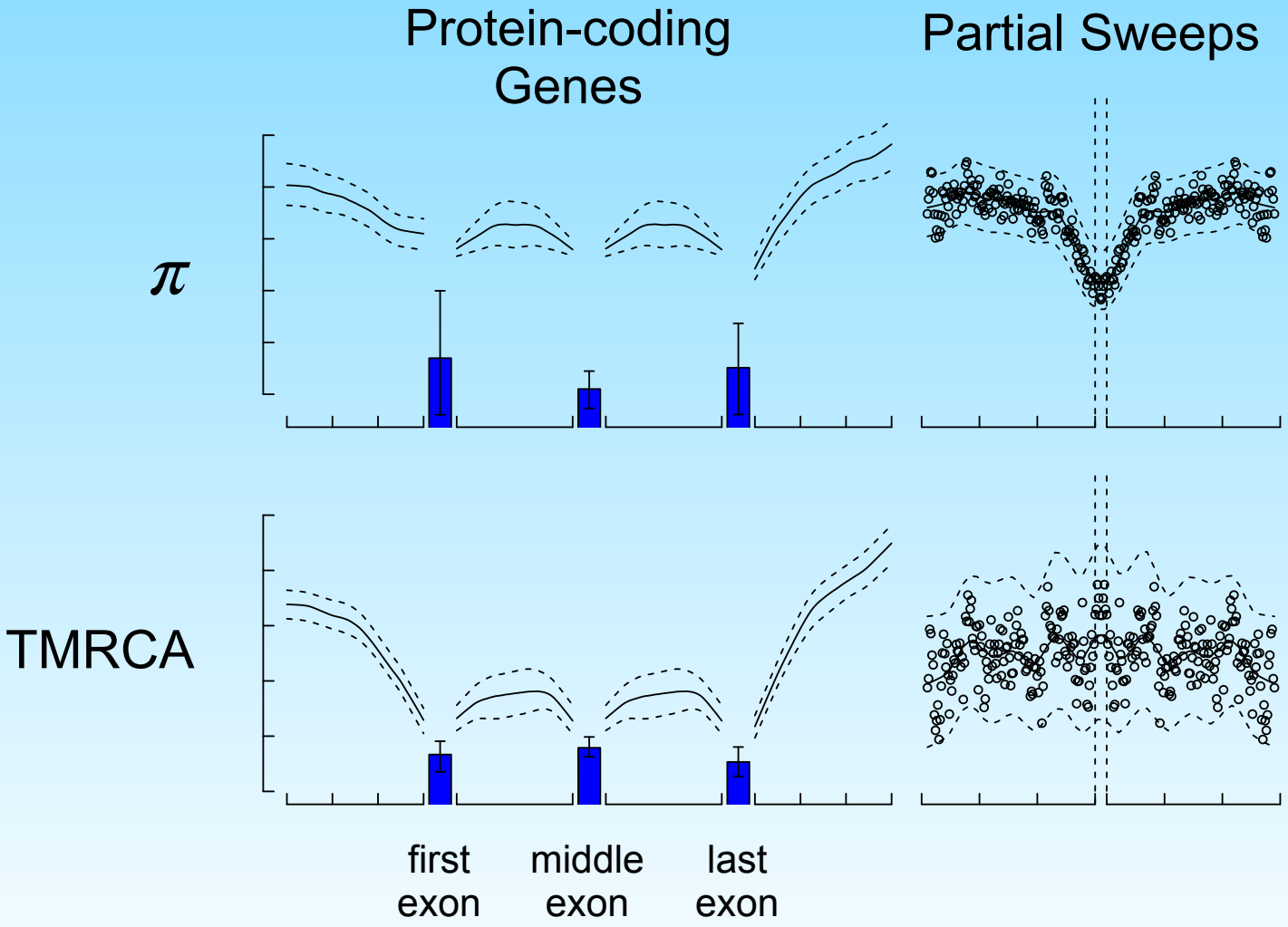
Sites Under Selection Have Decreased Allele Age



…Even After Accounting for Derived Allele Frequency



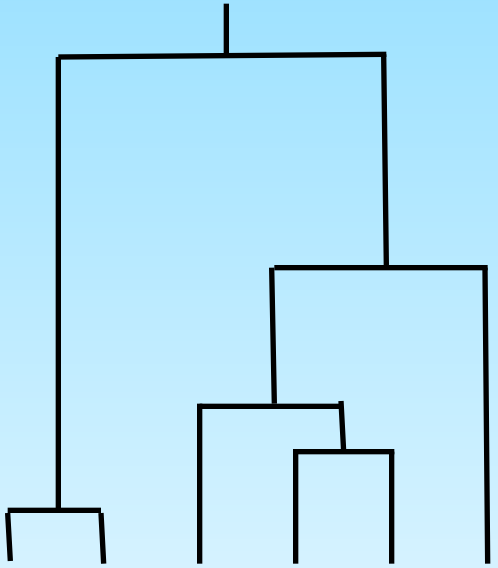
Genes and Sweeps (CEU)



Voight et al., *PLOS Biol*, 2006

Relative TMRCA Halflife (RTH)

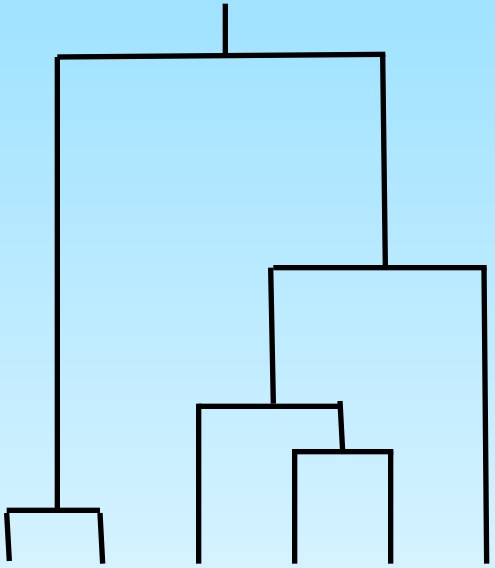
Neutral Drift



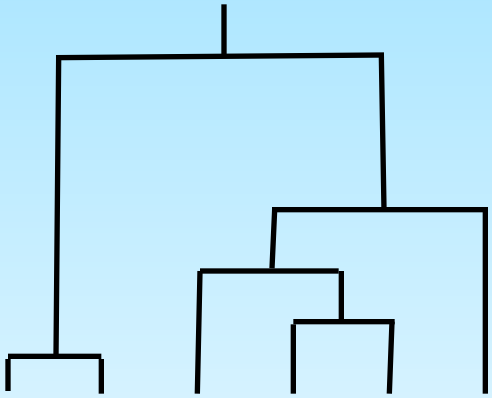
Rasmussen et al., *PLOS Genetics,* 2014

Relative TMRCA Halflife (RTH)

Neutral Drift



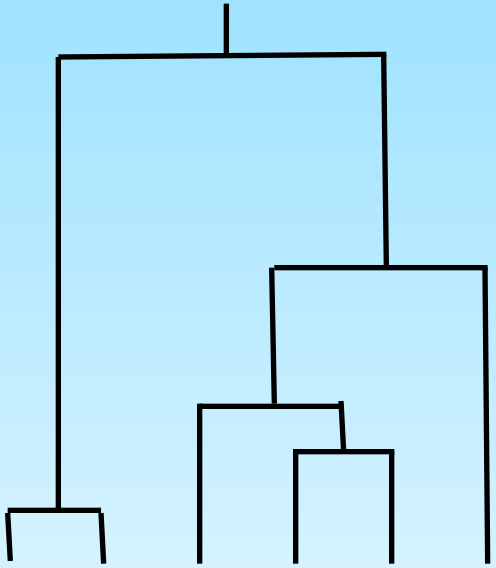
Background Selection



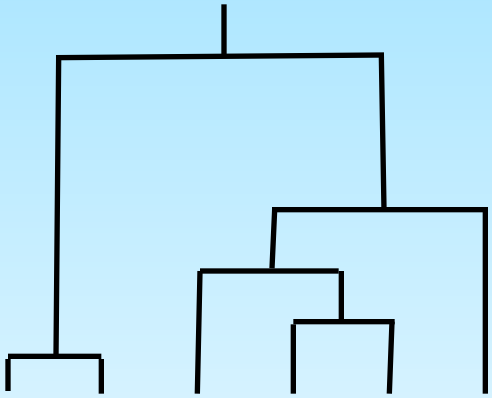
Rasmussen et al., *PLOS Genetics,* 2014

Relative TMRCA Halflife (RTH)

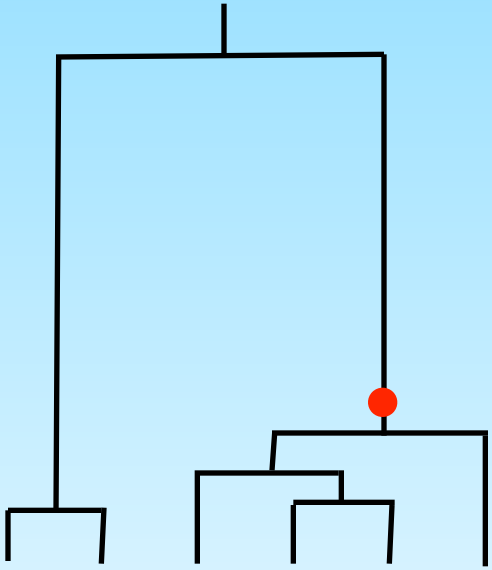
Neutral Drift



Background Selection

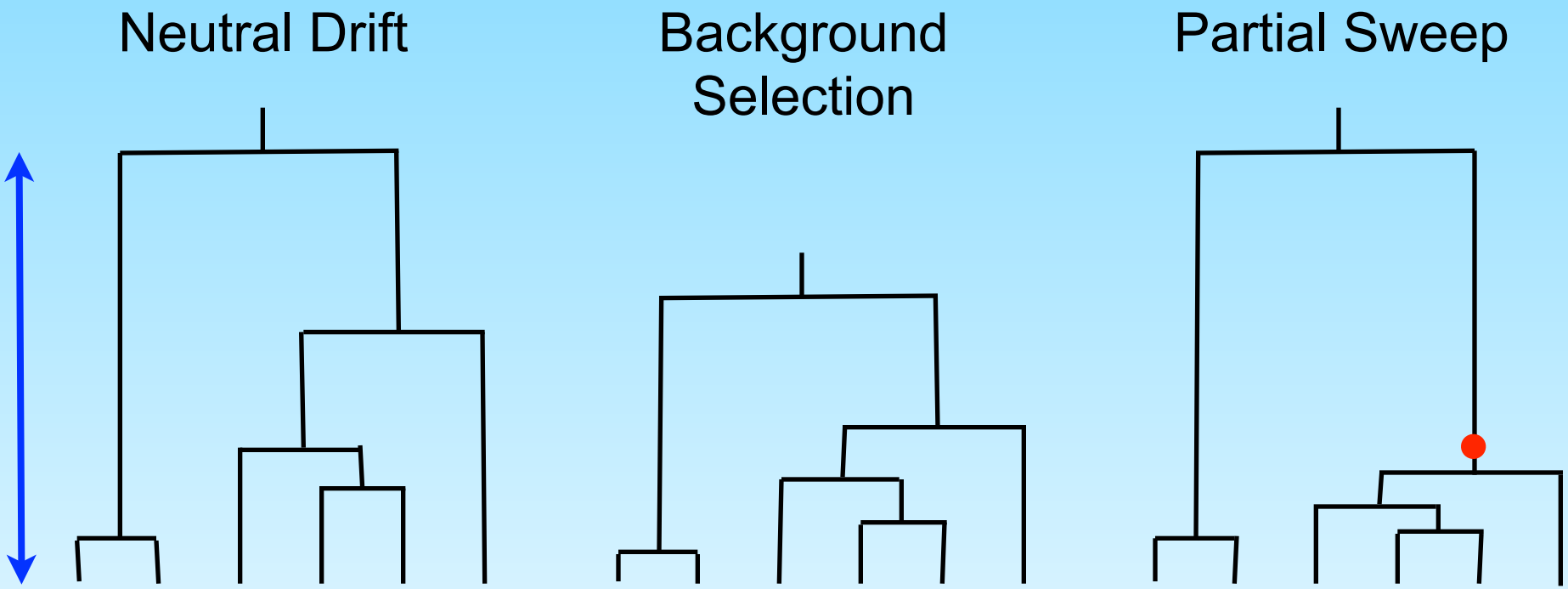


Partial Sweep



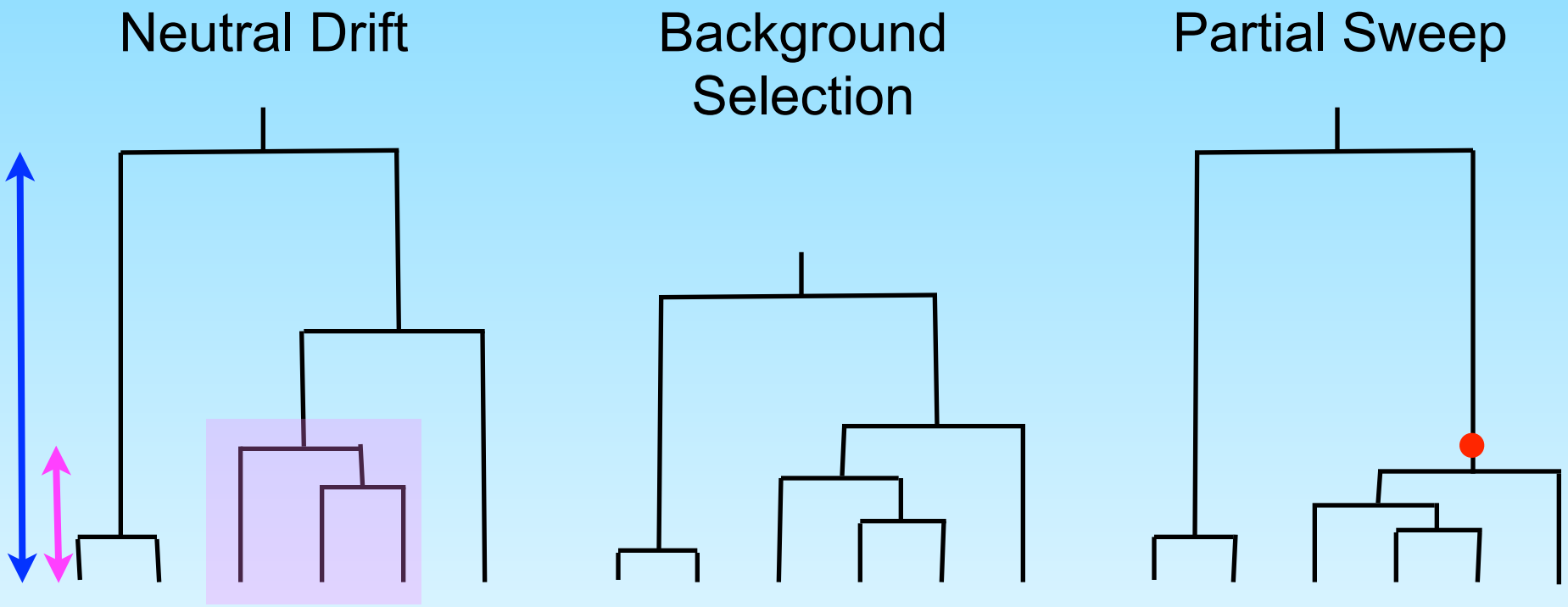
Rasmussen et al., *PLOS Genetics,* 2014

Relative TMRCA Halflife (RTH)



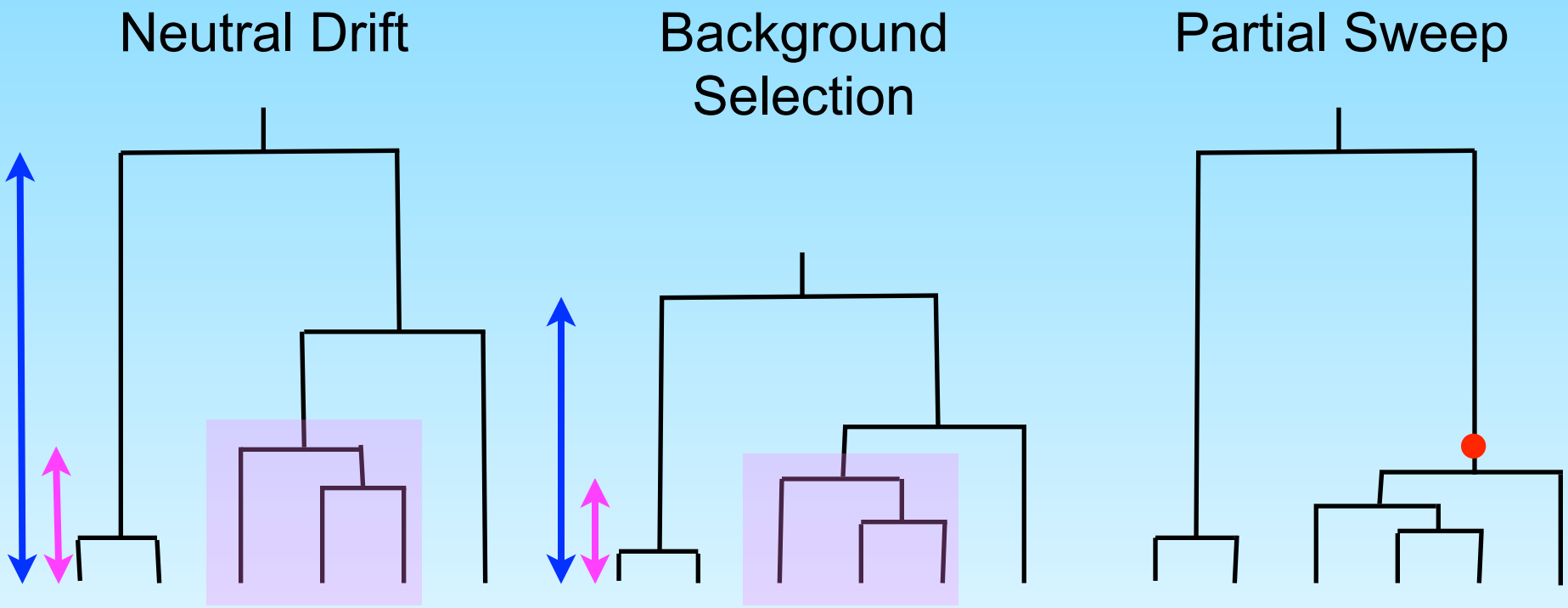
Rasmussen et al., *PLOS Genetics,* 2014

Relative TMRCA Halflife (RTH)



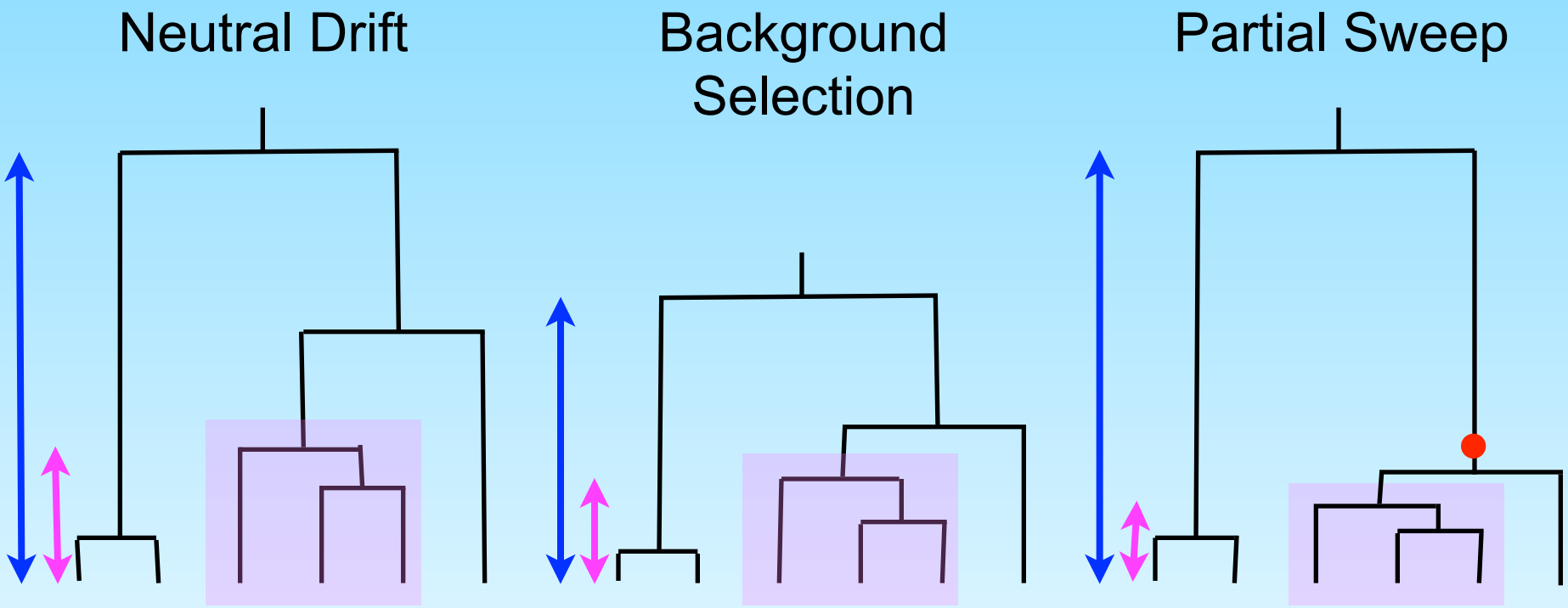
Rasmussen et al., *PLOS Genetics,* 2014

Relative TMRCA Halflife (RTH)



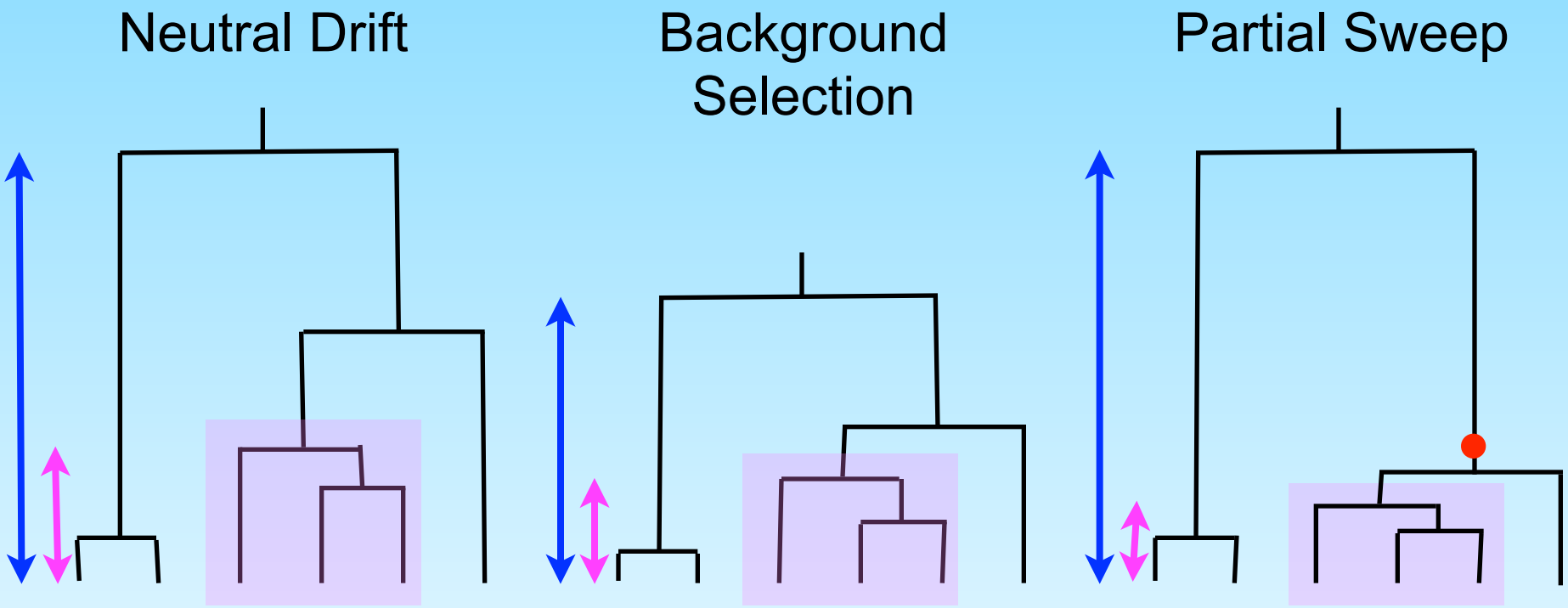
Rasmussen et al., *PLOS Genetics,* 2014

Relative TMRCA Halflife (RTH)



Rasmussen et al., *PLOS Genetics,* 2014

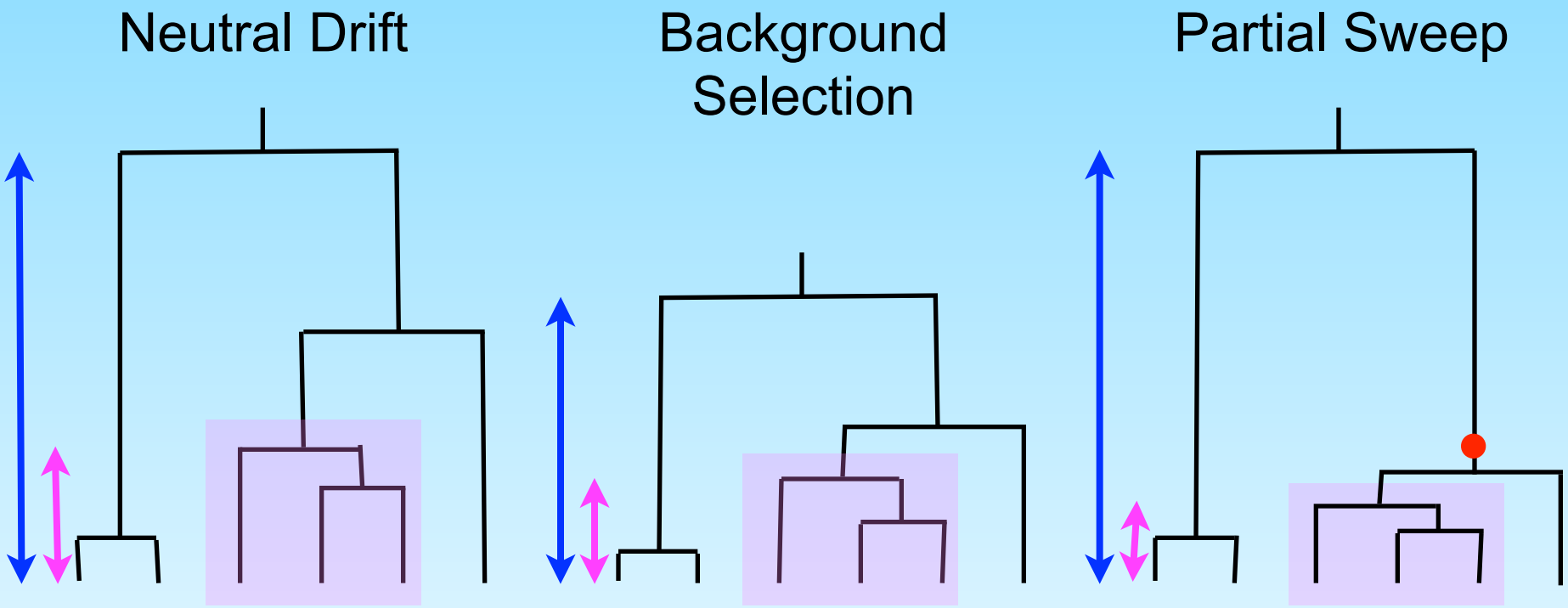
Relative TMRCA Halflife (RTH)



RTH = half-TMRCA / TMRCA

Rasmussen et al., *PLOS Genetics,* 2014

Relative TMRCA Halflife (RTH)

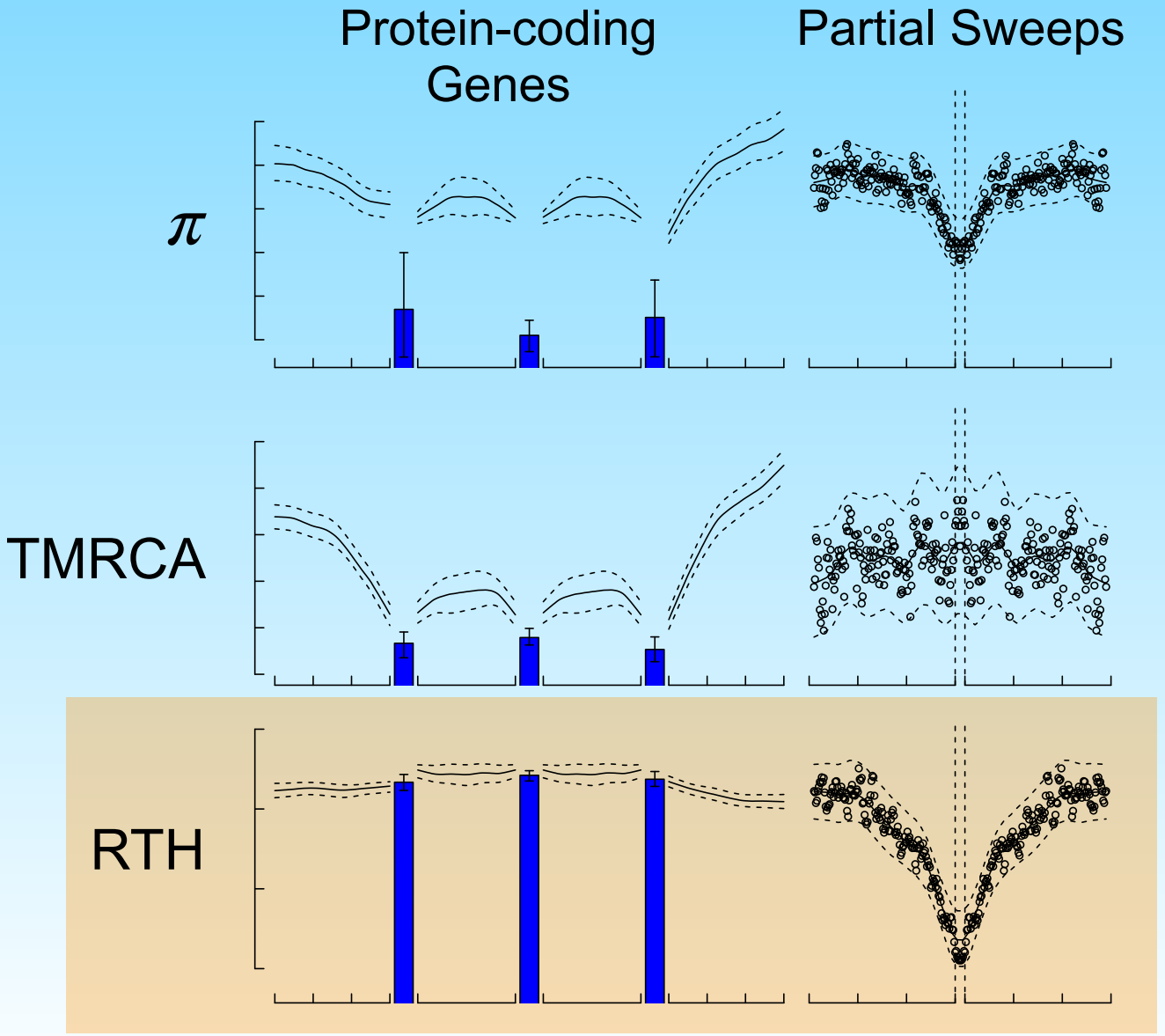


RTH = half-TMRCA / TMRCA

RTH = 1/3 RTH = 1/3 RTH = 1/6

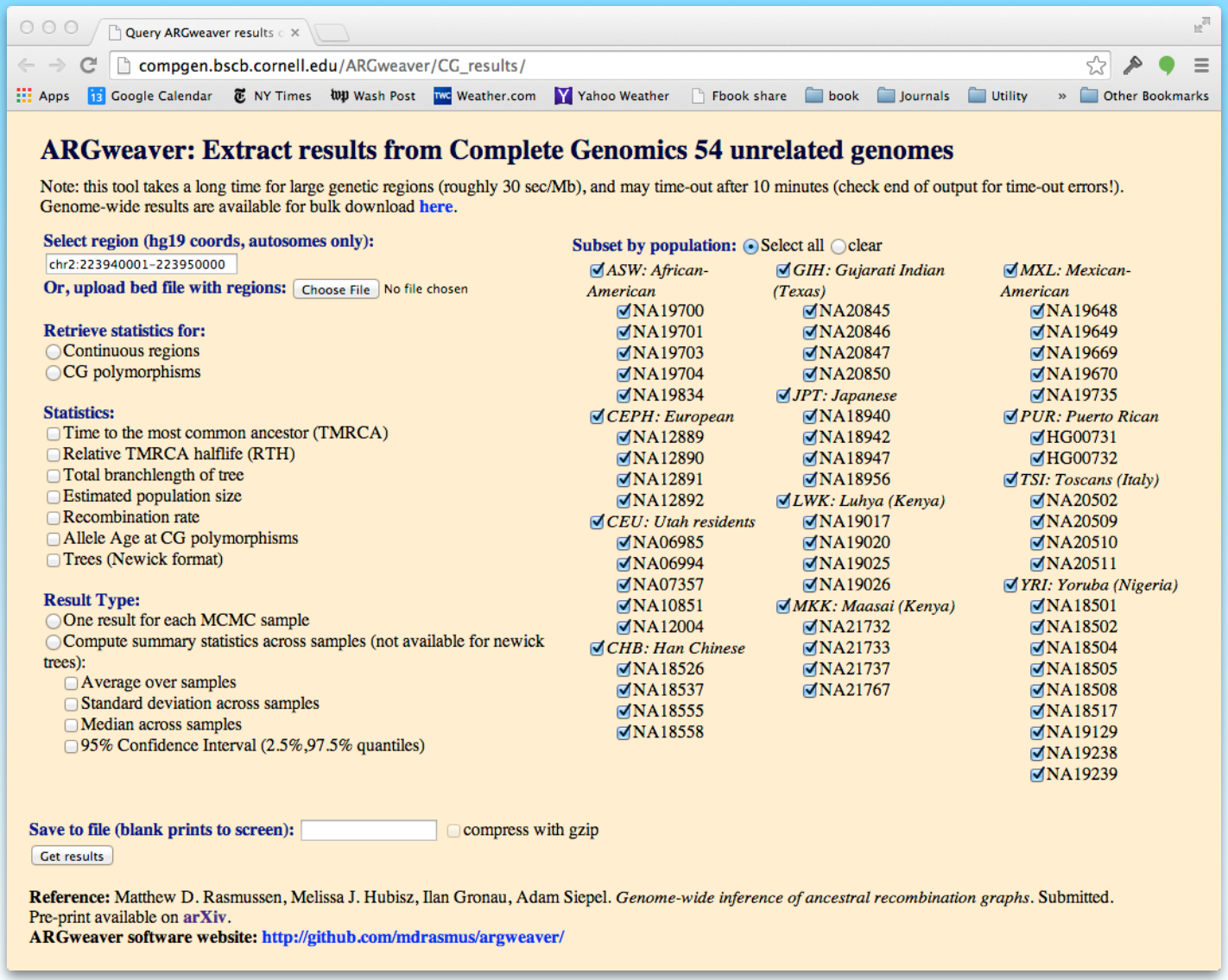
Rasmussen et al., *PLOS Genetics,* 2014

Genes and Sweeps (CEU)



Rasmussen et al., *PLOS Genetics,* 2014

Query Interface



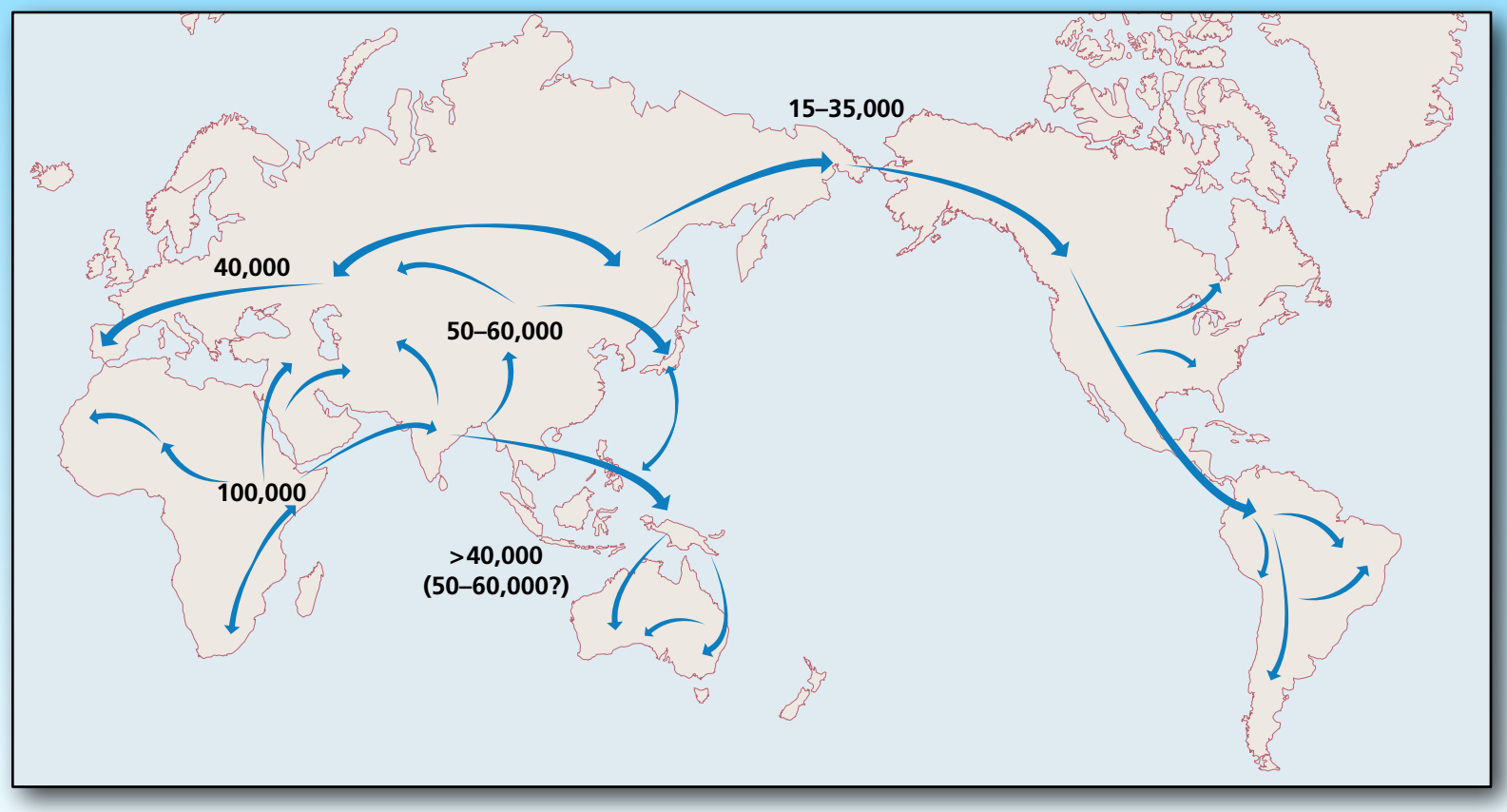
http://compgen.bscb.cornell.edu

Future Work

- Simultaneous phasing and ARG inference
- Demography inference
- Community resources
  - Extended dynamic querying of ARGs
  - On-the-fly threading
- Association mapping
- Any problem addressed by Li & Stephens model!

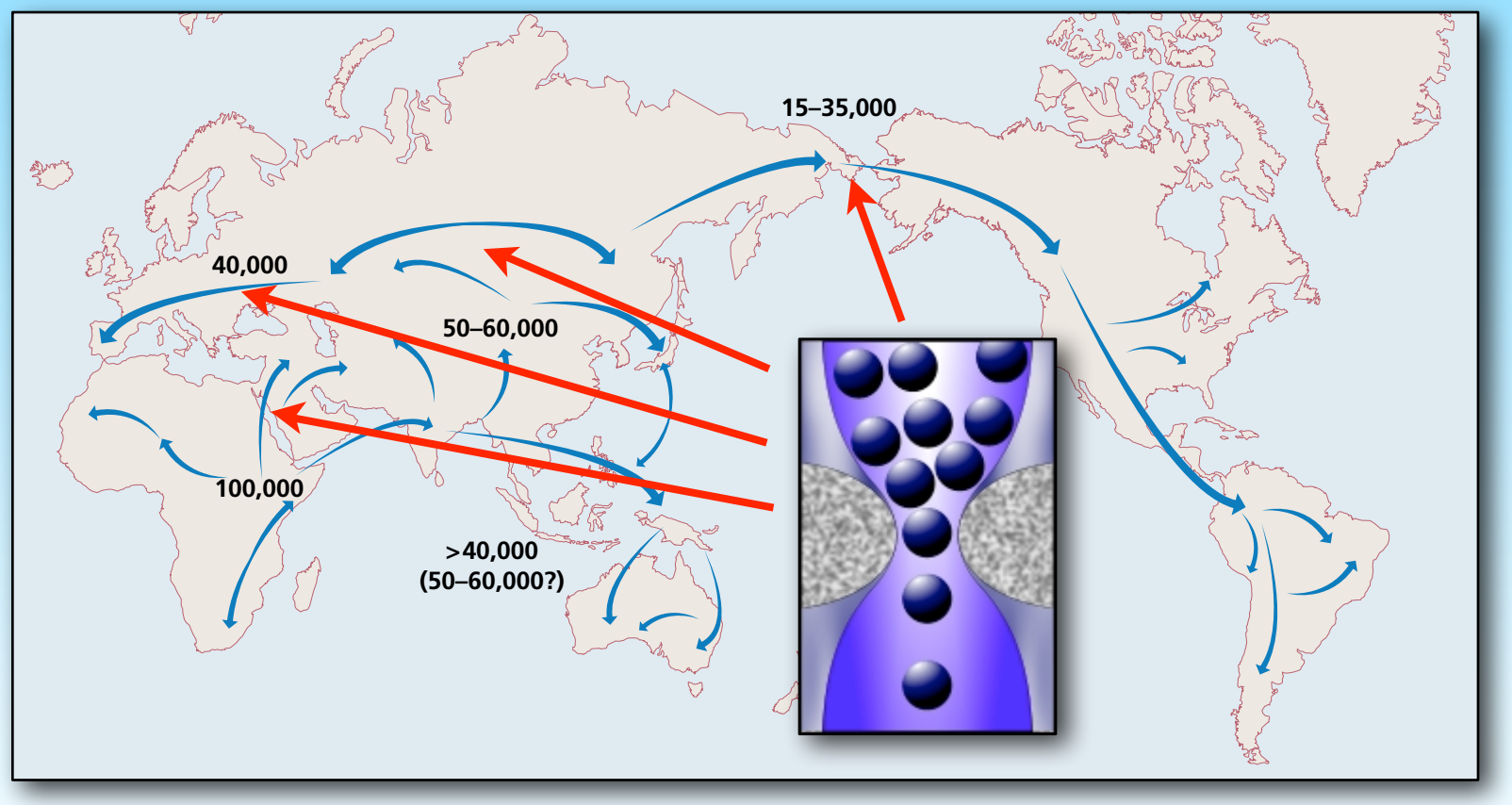
Part 3: Demography Inference

Origins of Human Populations



Cavalli-Sforza and Feldman, *Nat Genet*, 2003

Origins of Human Populations

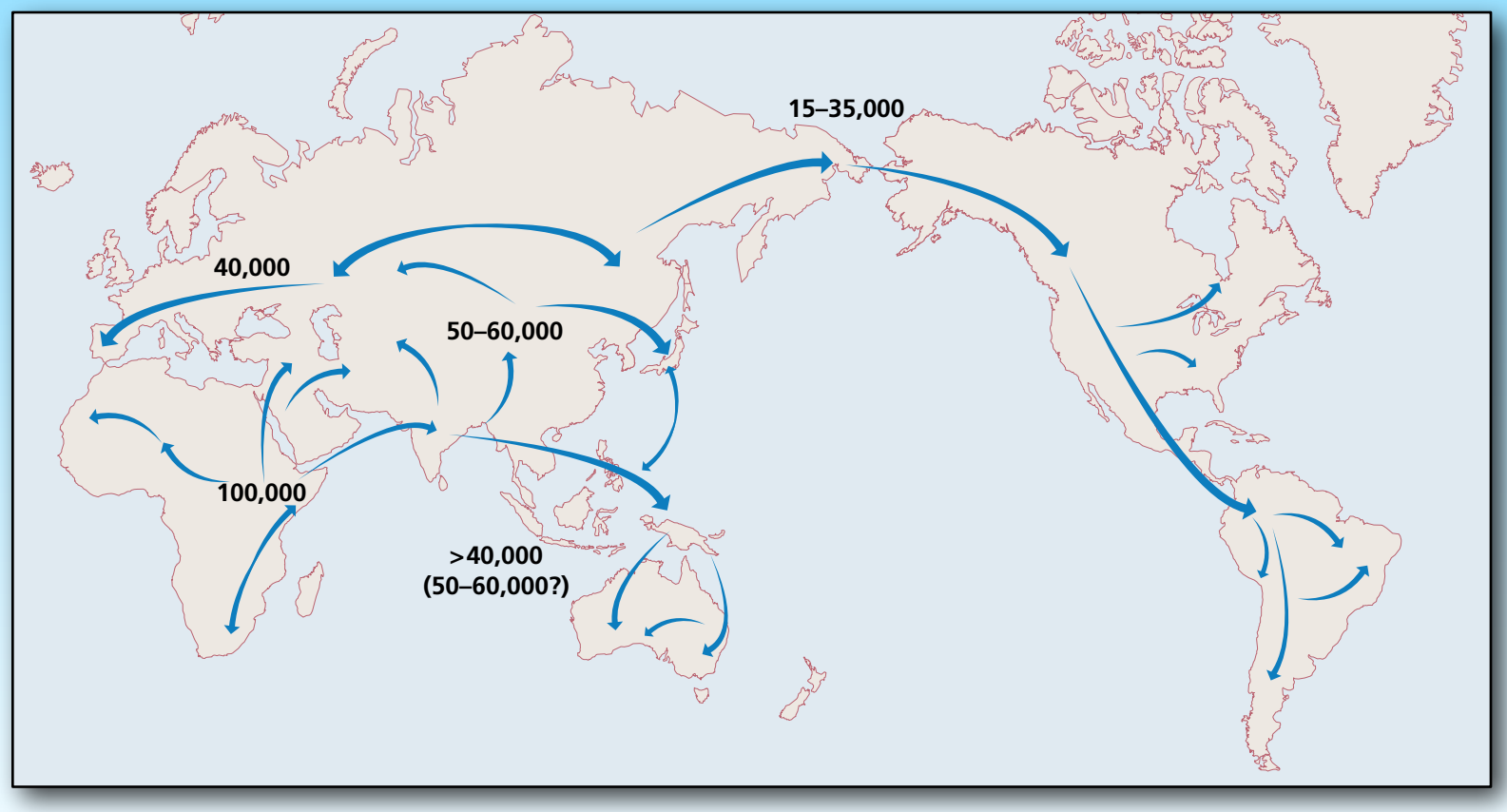


Cavalli-Sforza and Feldman, *Nat Genet*, 2003

Origins of Human Populations

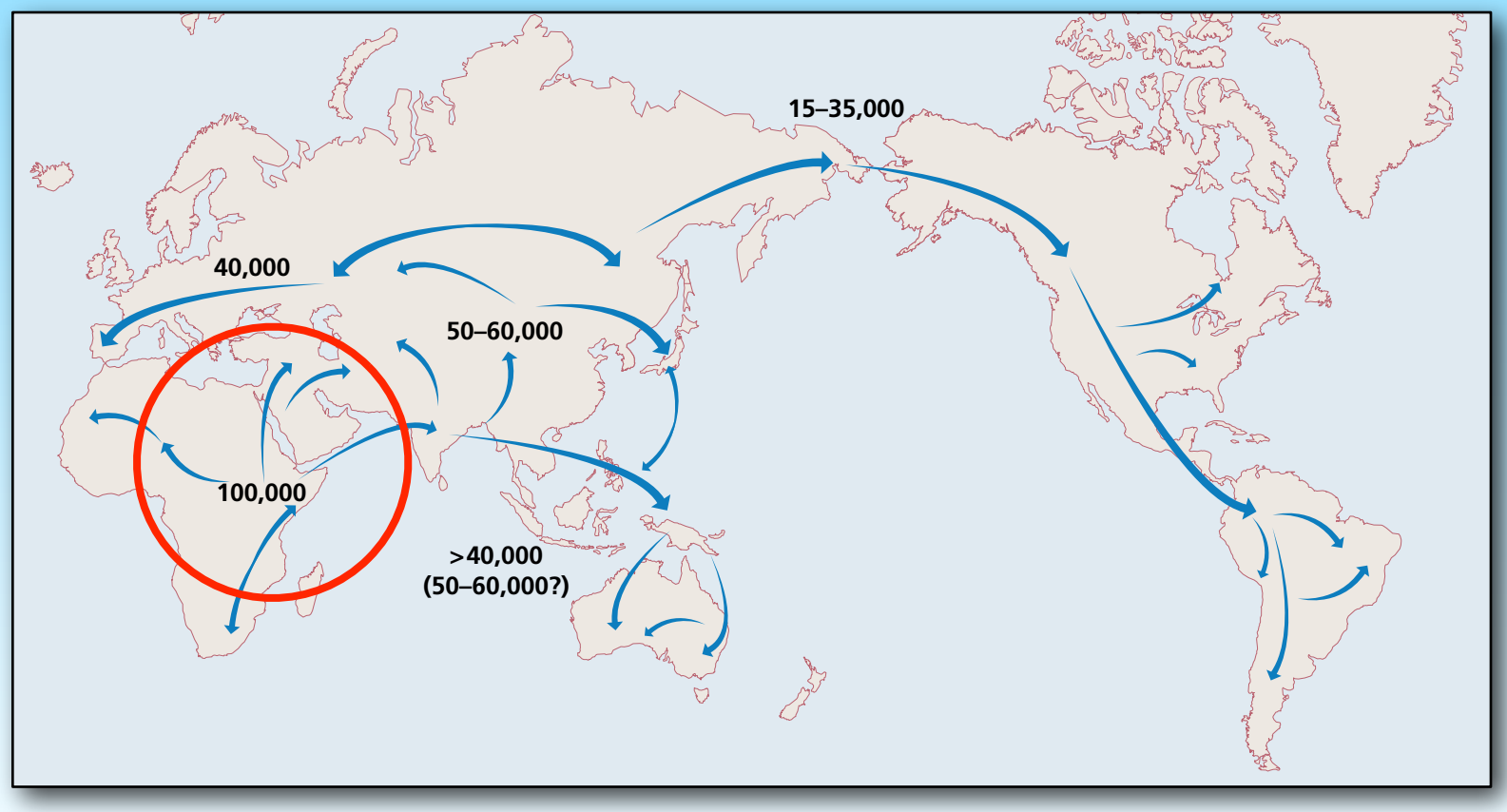
Cavalli-Sforza and Feldman, *Nat Genet*, 2003

Origins of Human Populations



Cavalli-Sforza and Feldman, *Nat Genet*, 2003

Origins of Human Populations



Schuster et al, *Nature*, 2010

Origins of Human Populations

Schuster et al, *Nature*, 2010

Integrated Statistical Model *(MCMCcoal)*

```
AATGAACCGTTTCTGAGGCCATT
AGTGAACCGTTACTGACGCCATT
AATGAATCGTTACTGAGGCTATT
```

X_i

G_i

$\boldsymbol{\theta}, \boldsymbol{\tau}$

Yang, 2002; Rannala & Yang, 2003;
Burgess & Yang, 2008

Integrated Statistical Model *(MCMCcoal)*

```
AATGAACCGTTTCTGAGGCCATT
AGTGAACCGTTACTGACGCCATT
AATGAATCGTTACTGAGGCTATT
```

X_i G_i $\boldsymbol{\theta}, \boldsymbol{\tau}$

$$P(\mathbf{X}, \mathbf{G}, \boldsymbol{\theta}, \boldsymbol{\tau}) = P(\boldsymbol{\theta})\, P(\boldsymbol{\tau}) \prod_i P(G_i \mid \boldsymbol{\theta}, \boldsymbol{\tau})\, P(X_i \mid G_i)$$

Gamma priors

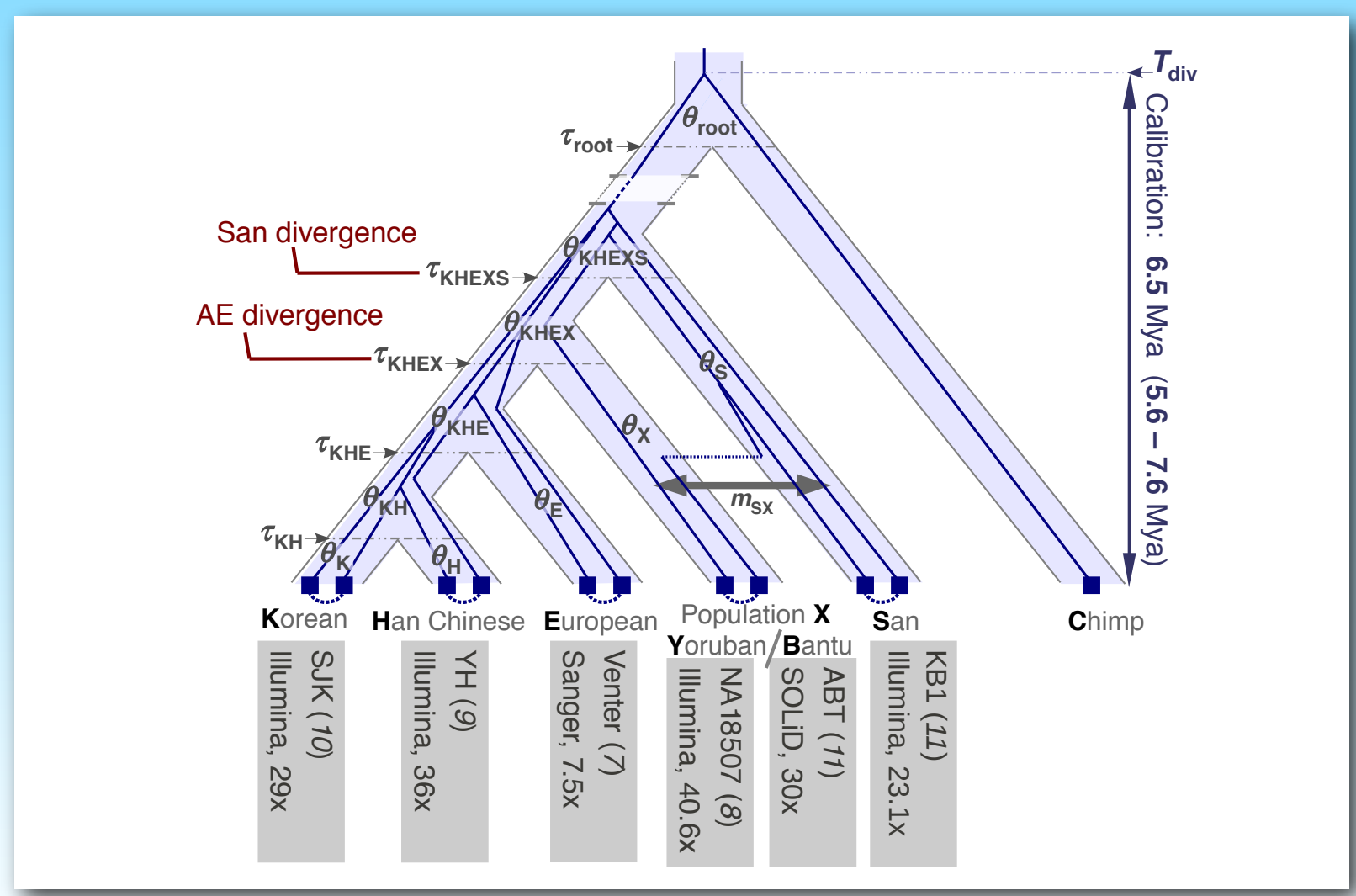
Censored coalescent model

Finite sites model, Felsenstein's pruning algorithm

Goal: $P(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{X})$

Yang, 2002; Rannala & Yang, 2003; Burgess & Yang, 2008

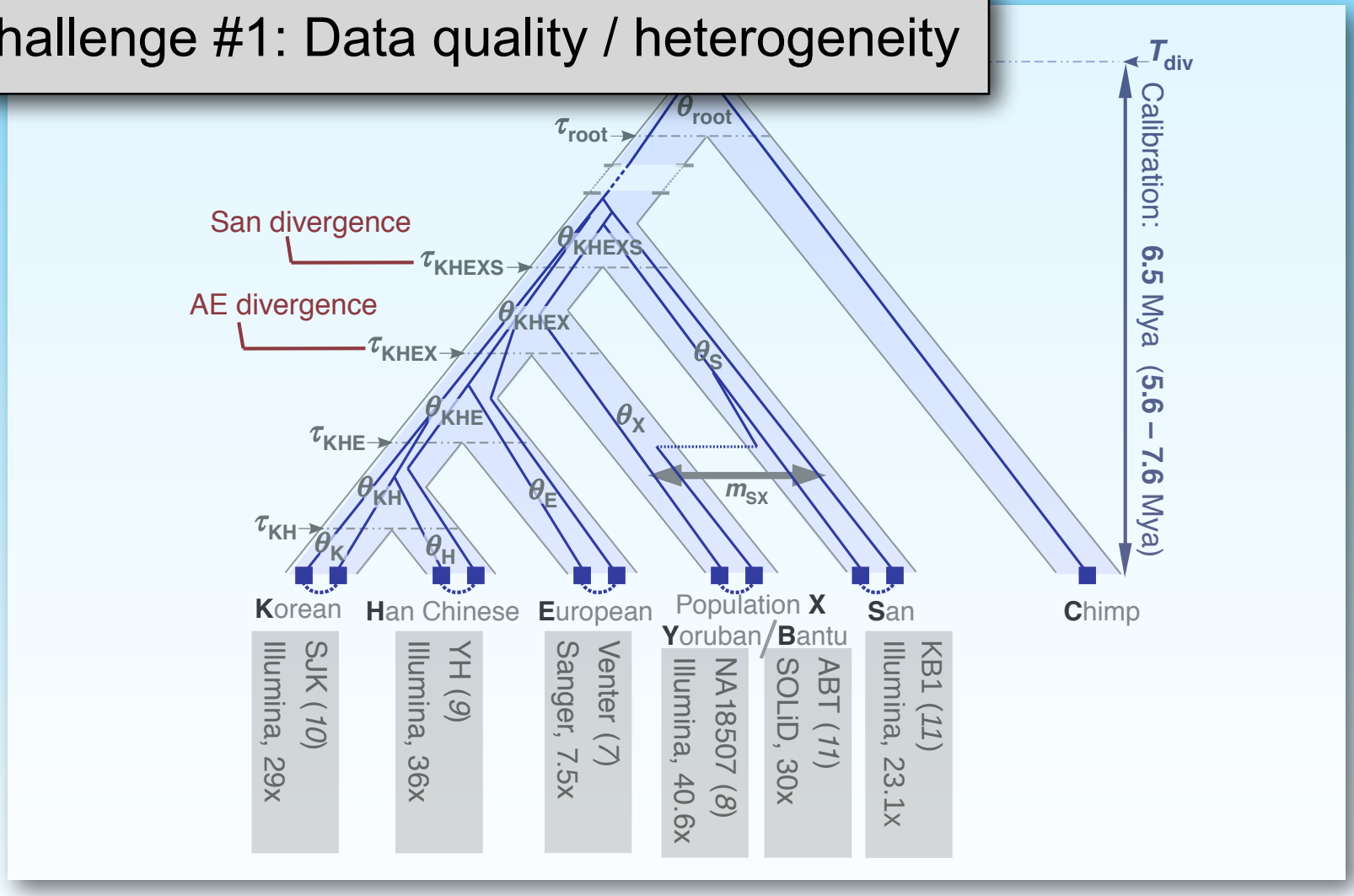
Data and Phylogeny



Gronau et al., *Nature Genetics*, 2011

Data and Phylogeny

Challenge #1: Data quality / heterogeneity

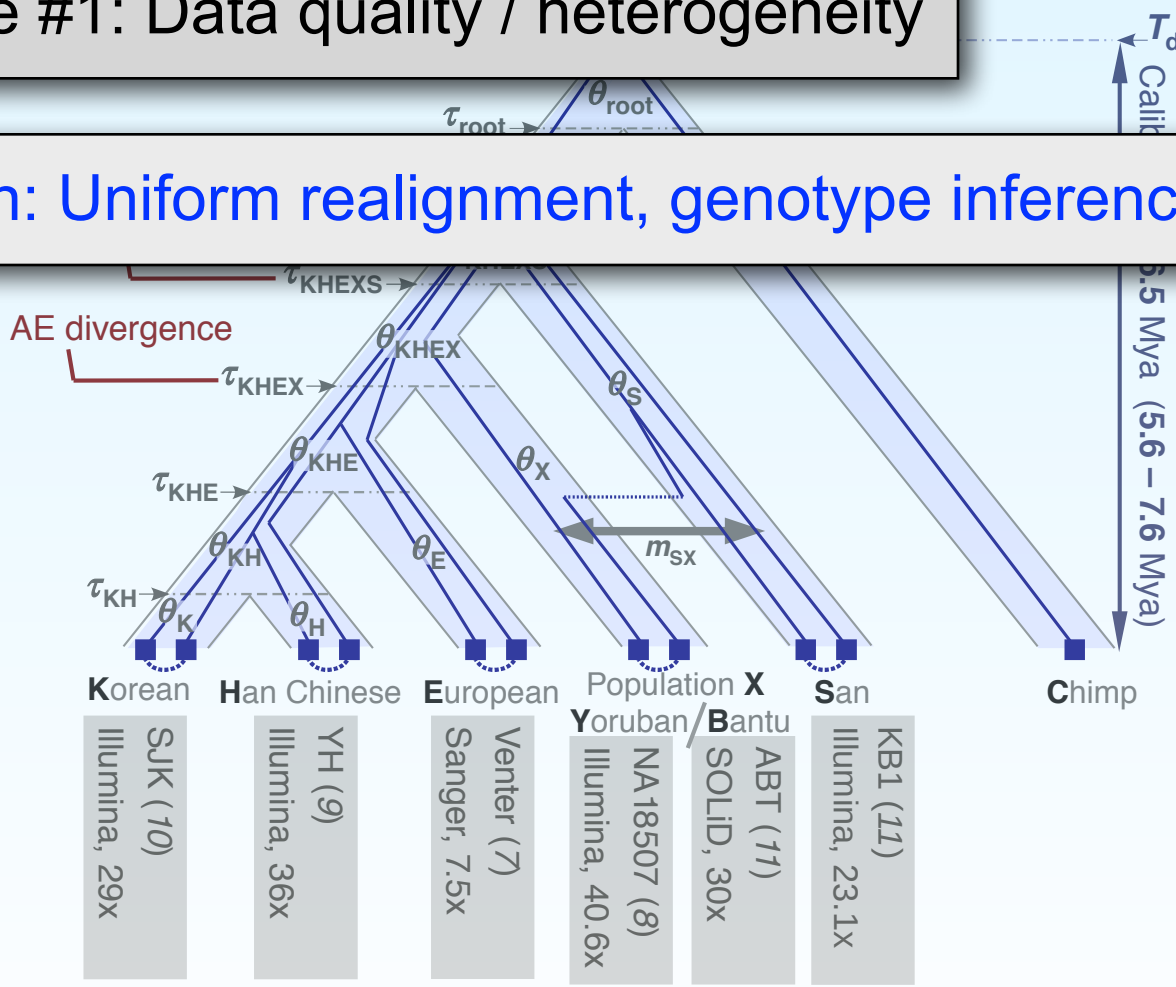


Gronau et al., *Nature Genetics*, 2011

Data and Phylogeny

Challenge #1: Data quality / heterogeneity

Solution: Uniform realignment, genotype inference



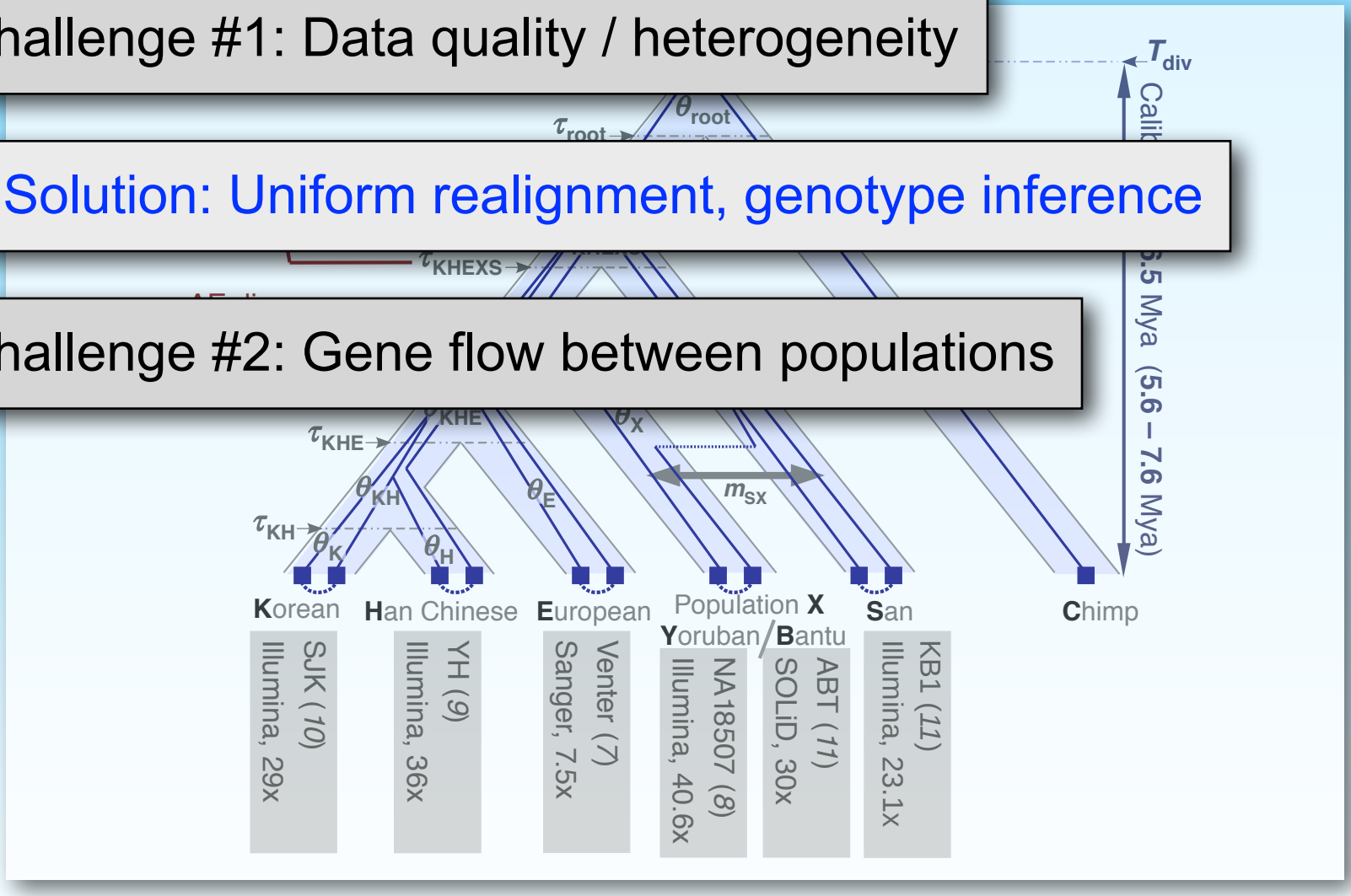
Gronau et al., *Nature Genetics*, 2011

Data and Phylogeny

Challenge #1: Data quality / heterogeneity

Solution: Uniform realignment, genotype inference

Challenge #2: Gene flow between populations



Gronau et al., *Nature Genetics*, 2011

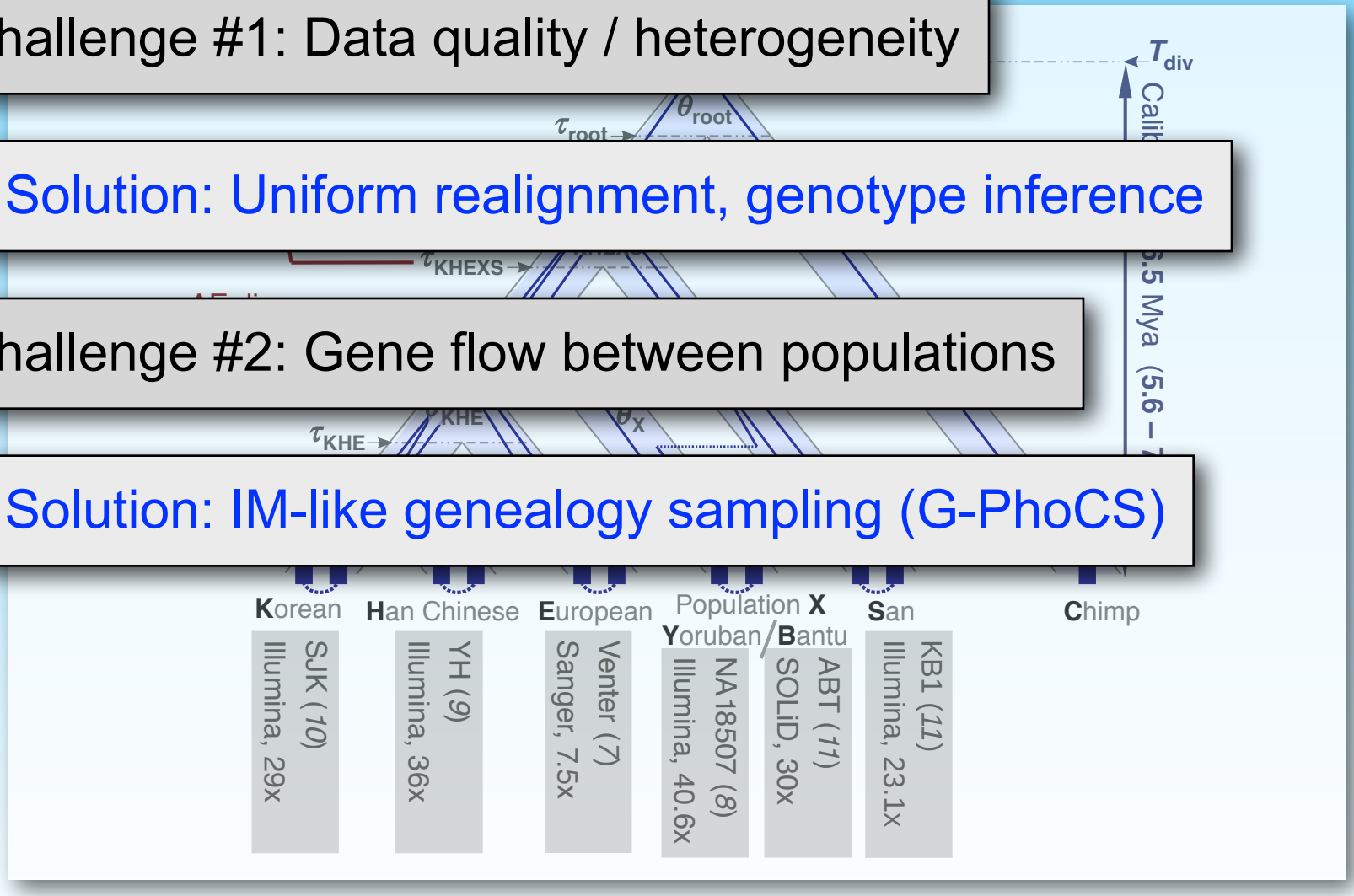
Data and Phylogeny

Challenge #1: Data quality / heterogeneity

Solution: Uniform realignment, genotype inference

Challenge #2: Gene flow between populations

Solution: IM-like genealogy sampling (G-PhoCS)



Gronau et al., *Nature Genetics*, 2011

Data and Phylogeny

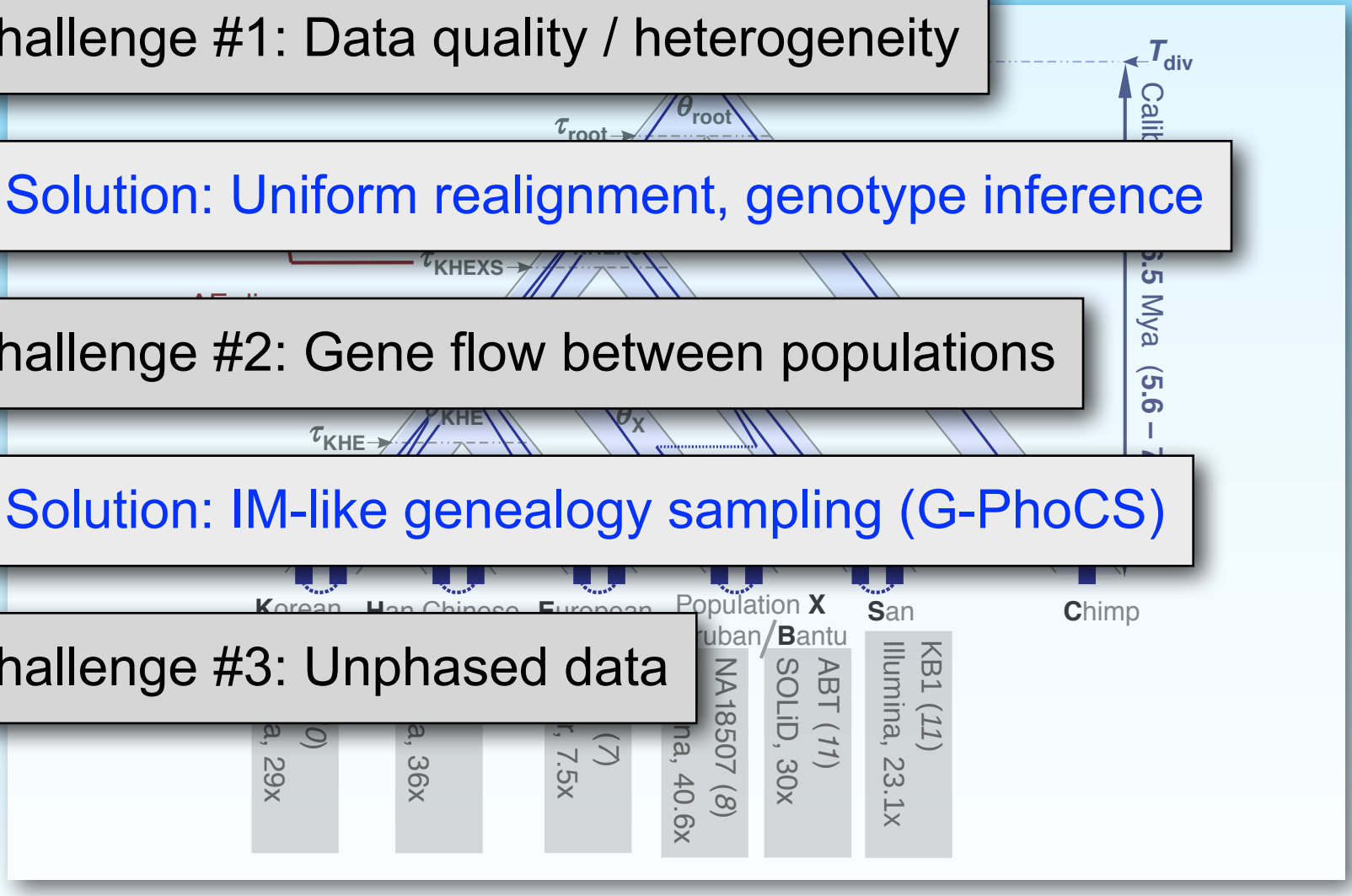
Challenge #1: Data quality / heterogeneity

Solution: Uniform realignment, genotype inference

Challenge #2: Gene flow between populations

Solution: IM-like genealogy sampling (G-PhoCS)

Challenge #3: Unphased data



Gronau et al., *Nature Genetics*, 2011

Data and Phylogeny

Challenge #1: Data quality / heterogeneity

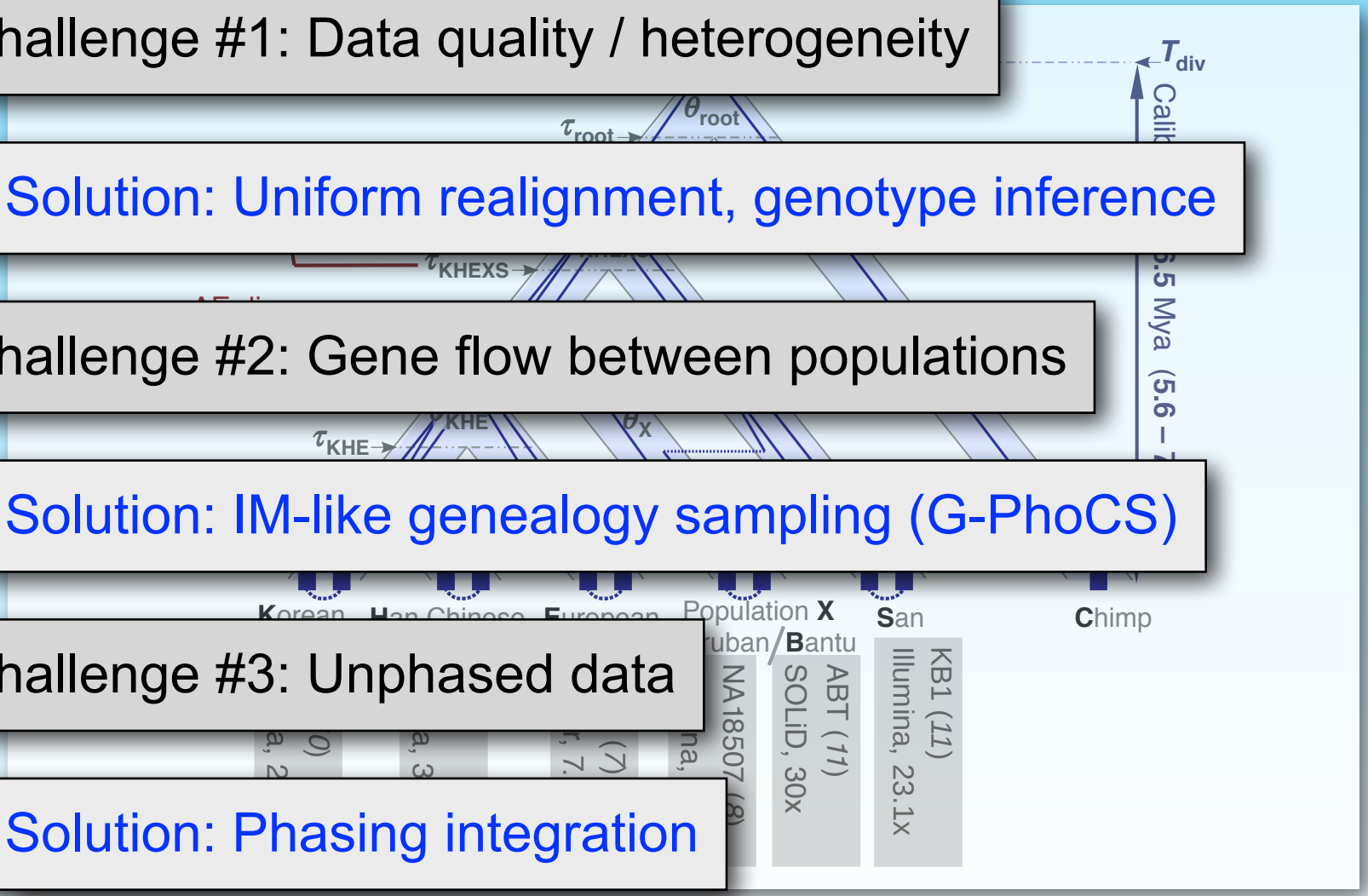
Solution: Uniform realignment, genotype inference

Challenge #2: Gene flow between populations

Solution: IM-like genealogy sampling (G-PhoCS)

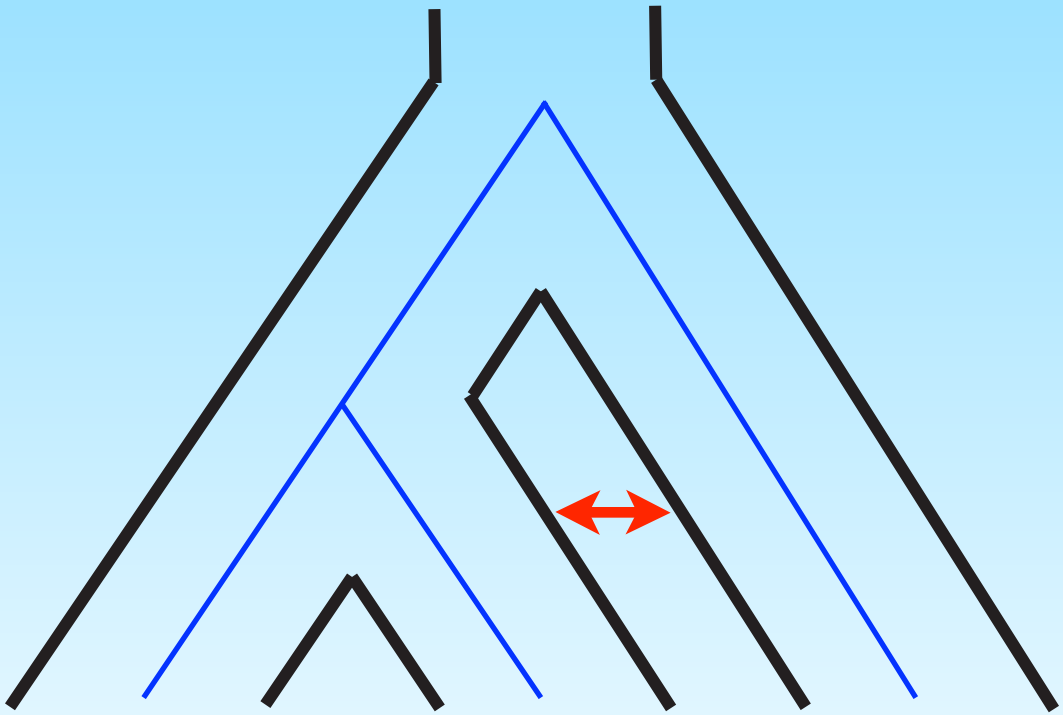
Challenge #3: Unphased data

Solution: Phasing integration

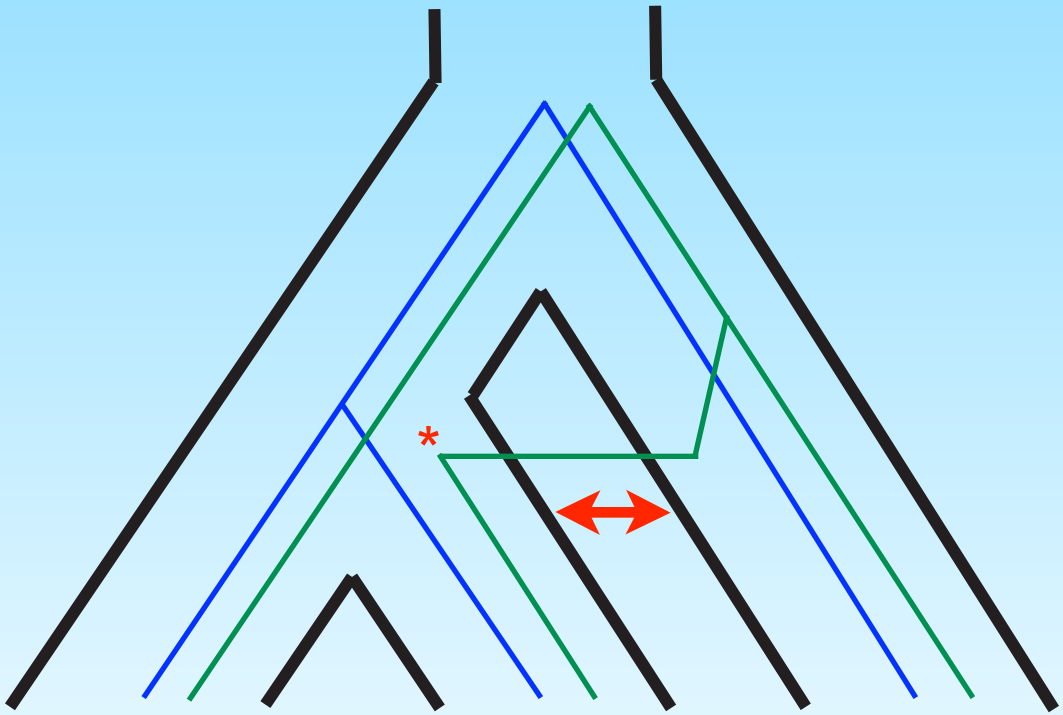


Gronau et al., *Nature Genetics*, 2011

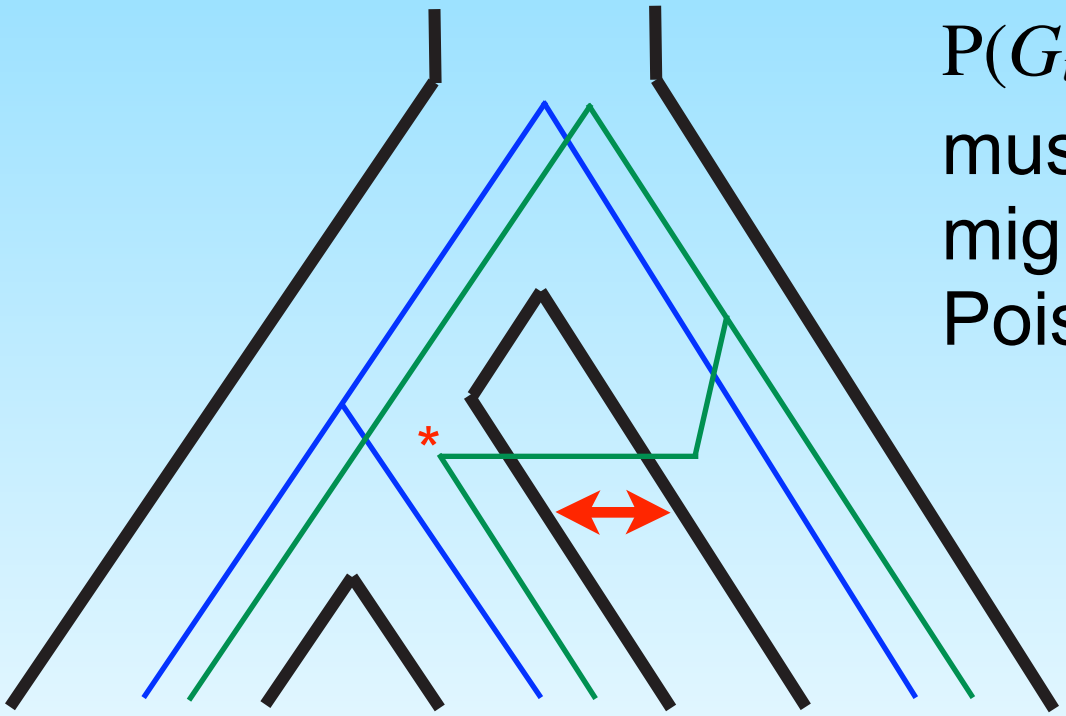
Allowing for Migration



Allowing for Migration

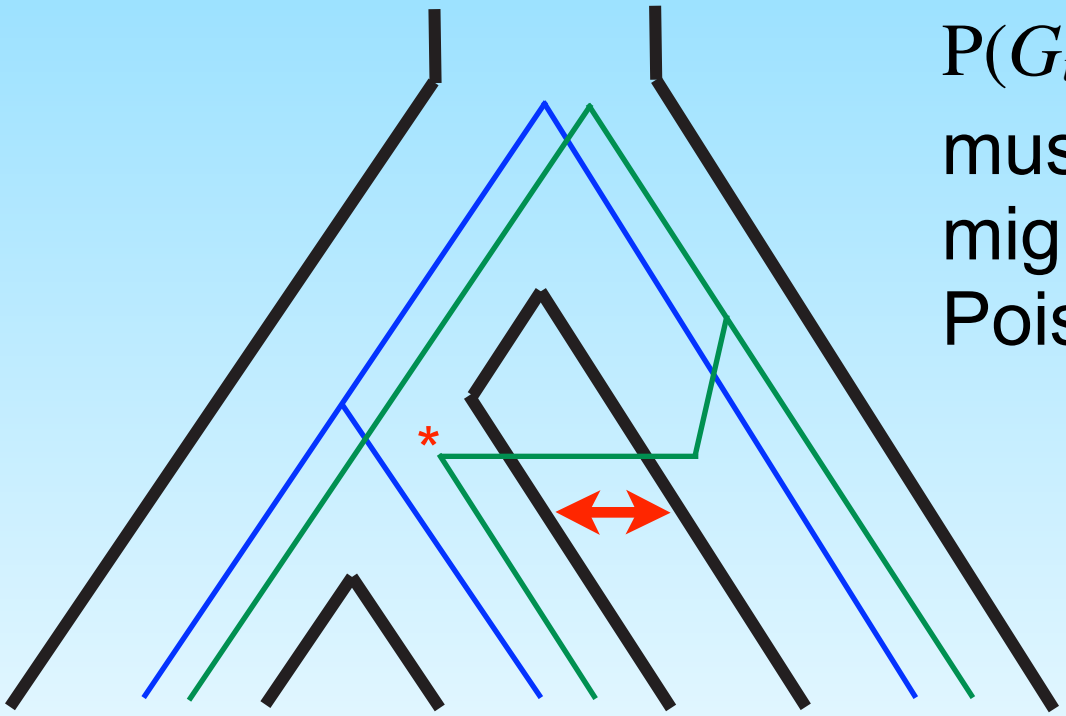


Allowing for Migration



The *genealogy prior*, $\mathrm{P}(G_i \mid \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{m})$, must allow for migration (another Poisson process)

Allowing for Migration



The *genealogy prior*,
$P(G_i \mid \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{m})$,
must allow for migration (another Poisson process)

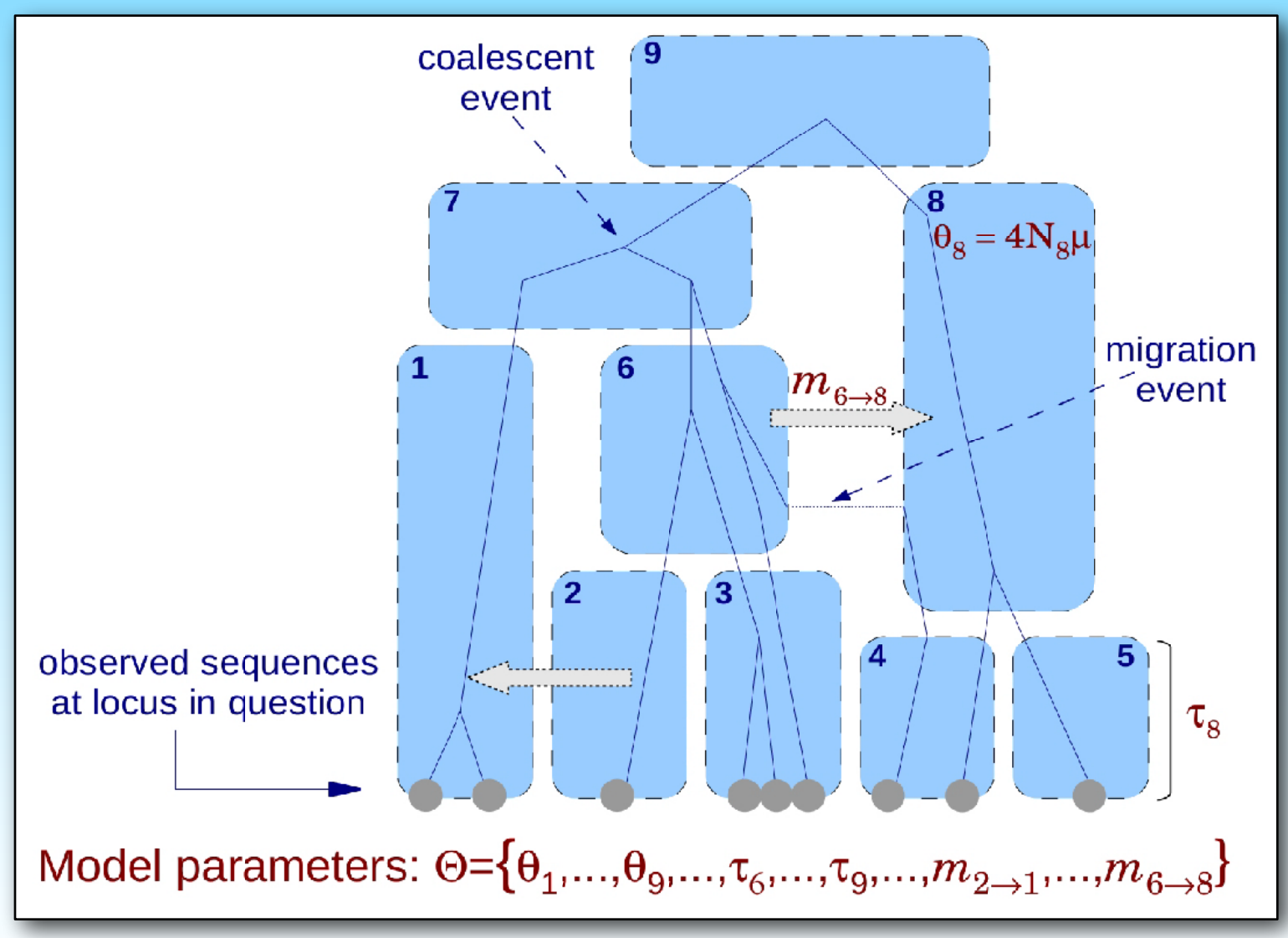
The *locus data likelihood*,
$P(X_i \mid G_i)$,
is unchanged

Phasing Integration

$$P(X|G) = \sum_{\mathcal{P} \in \{0,1\}^{k \times n}} P(X|\mathcal{P}, G)\, P(\mathcal{P})$$

$$= \prod_{j} \left(\frac{1}{2^{|\mathcal{H}_j|}} \sum_{\mathcal{P}^j \in \{0,1\}^{|\mathcal{H}_j|}} P(X^j \mid \mathcal{P}^j, G) \right)$$

General Model



MCMCcoal Algorithm

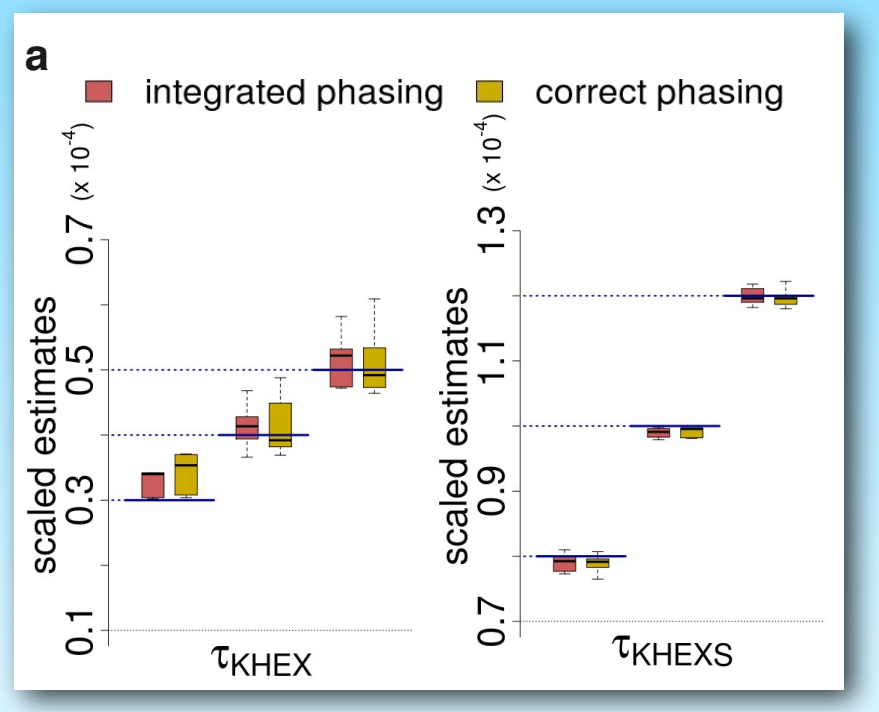
- **Step 1**: Update coalescent times
- **Step 2**: Subtree pruning and regrafting of genealogy
- **Step 3**: Update θ's
- **Step 4**: Update τ's, adjusting associated coalescent times via "rubber band"
- **Step 5**: Global scaling of θ's and τ's (mixing step)

Rannala and Yang, *Genetics*, 2003

Changes to Sampler

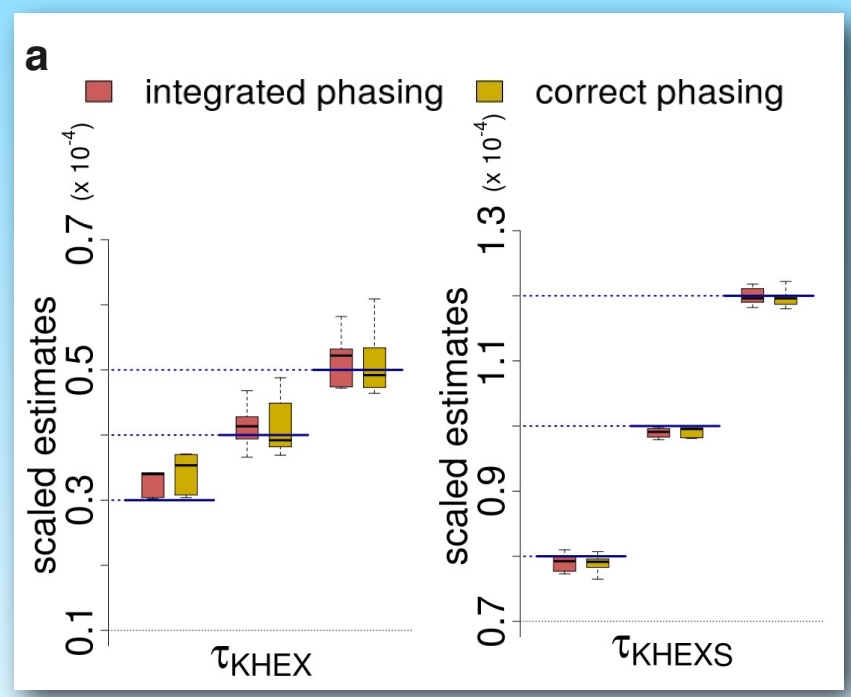
- The “regrafting” operation now allows lineages to pass through migration bands
- New steps are needed to update the individual migration times and the global migration parameters
- Certain additional conflicts must be considered when updating population divergence times

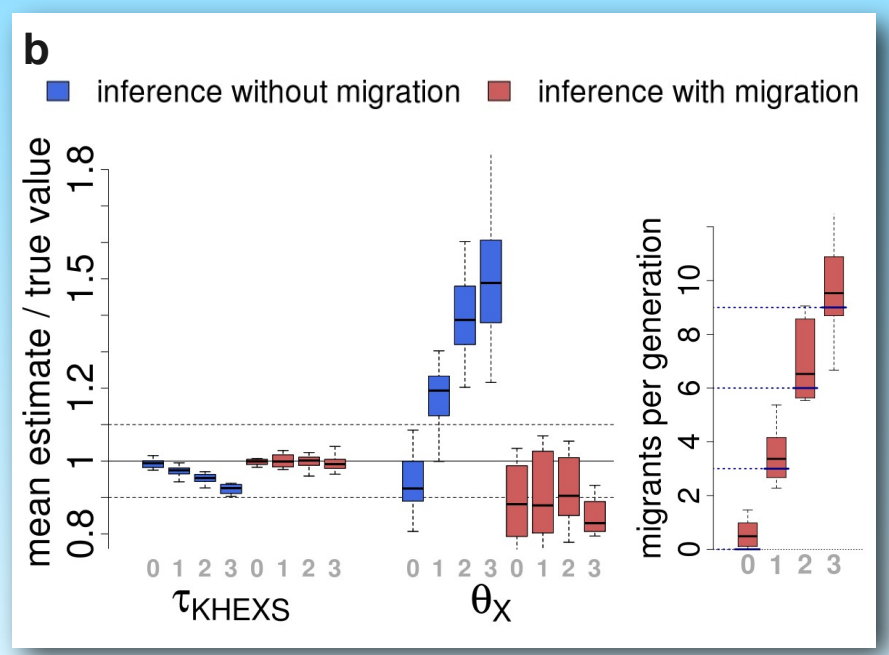
Simulation Results



Gronau et al., *Nature Genetics*, 2011

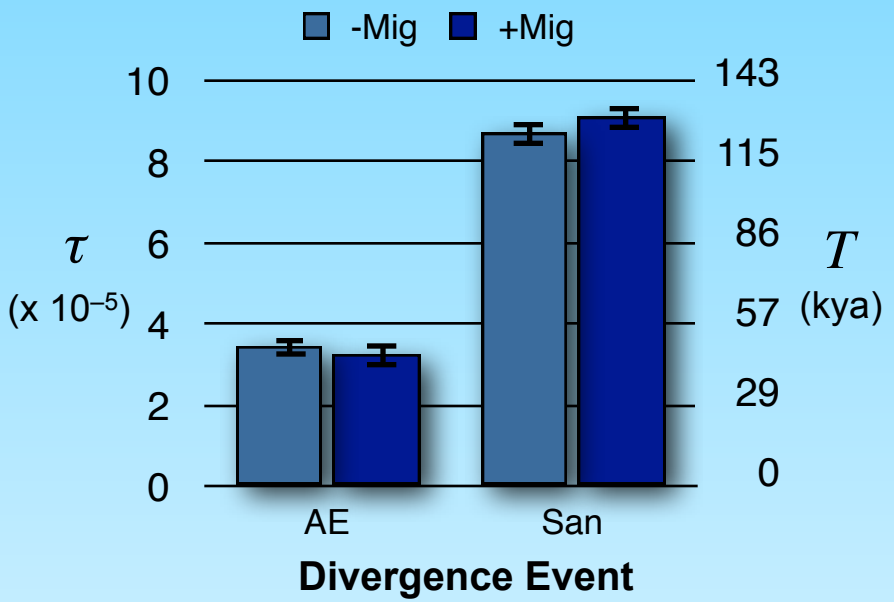
Simulation Results

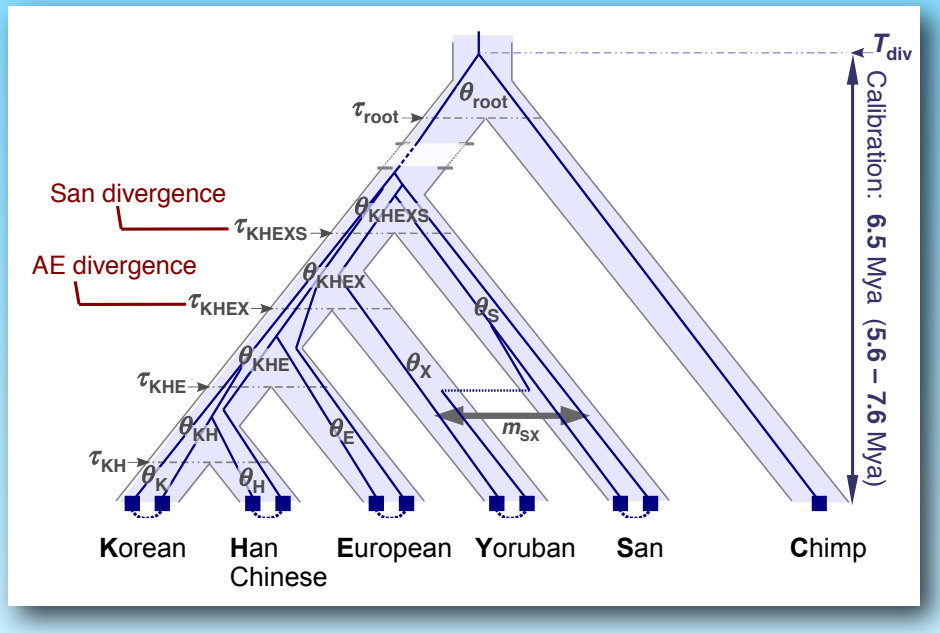




Gronau et al., *Nature Genetics*, 2011

Main Results

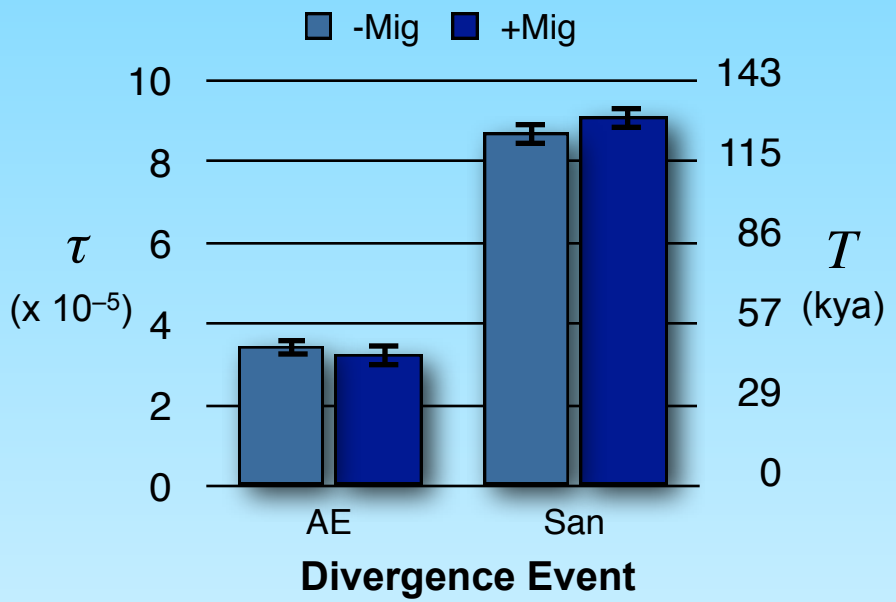


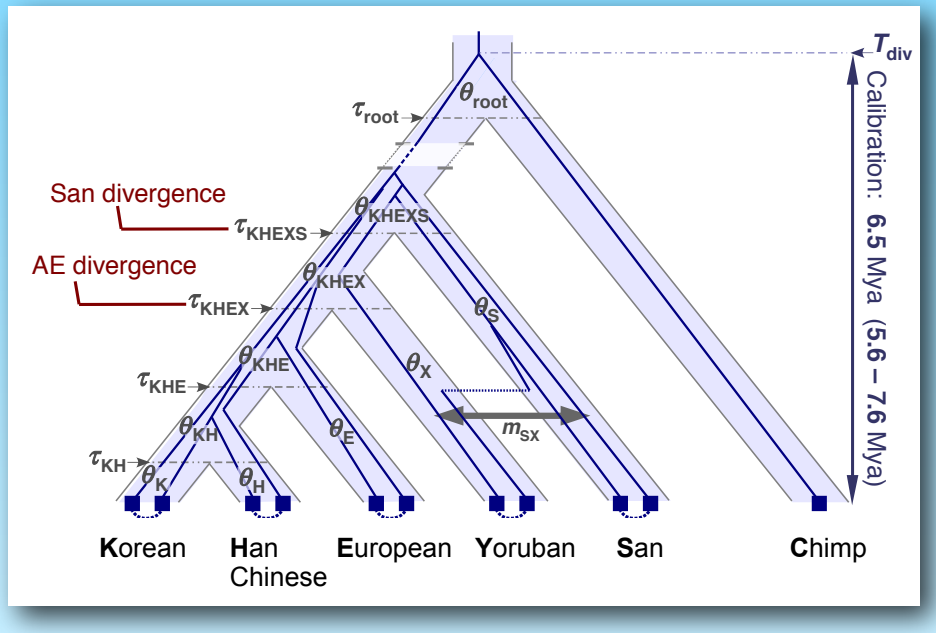


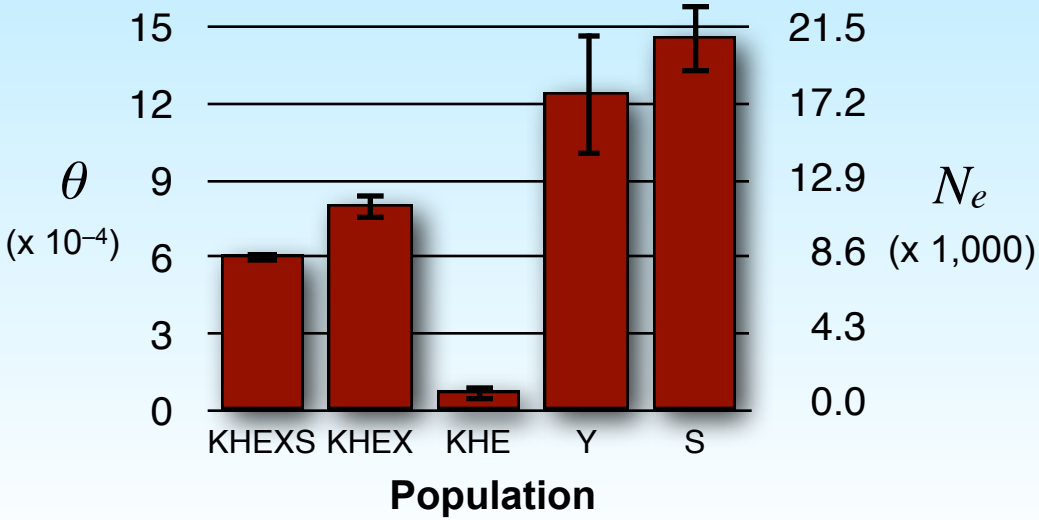
37,574 "neutral" loci, each 1 kbp in length

Gronau et al., *Nature Genetics*, 2011

Main Results







37,574 "neutral" loci, each 1 kbp in length

Gronau et al., *Nature Genetics*, 2011

Application to Domestic & Wild Canids


Boxer

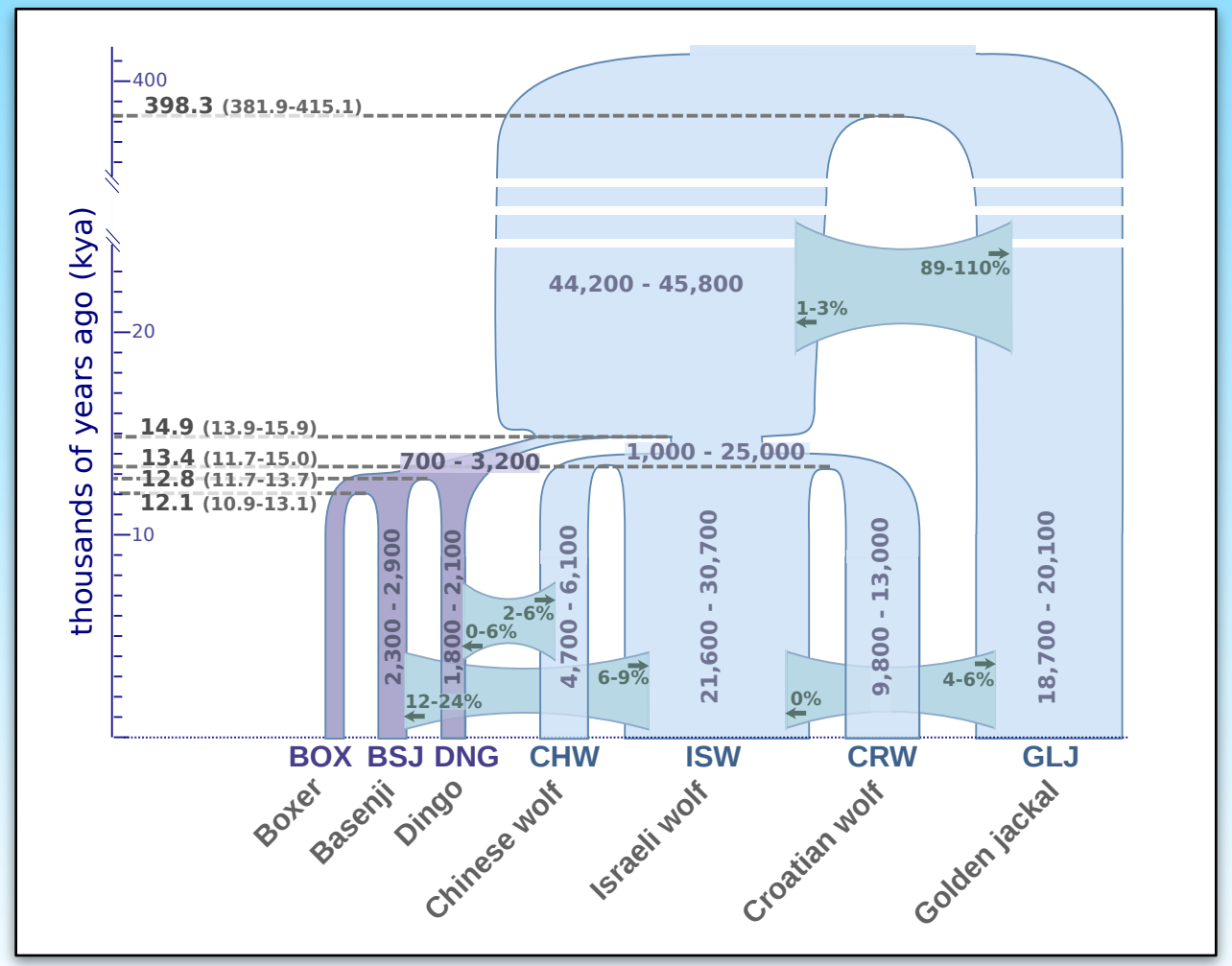

Basenji


Dingo


Gray wolf x 3

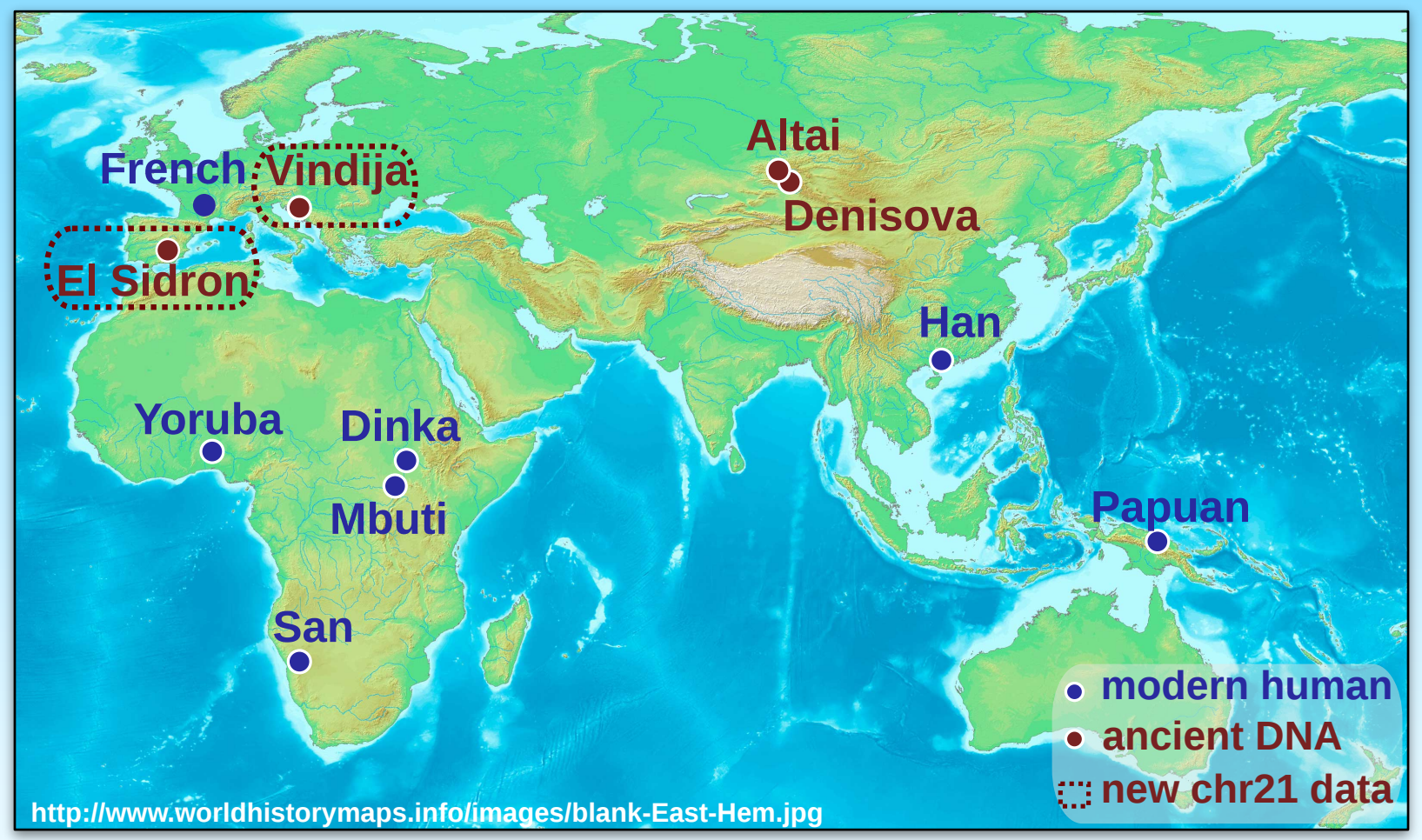

Golden jackal

Best Model



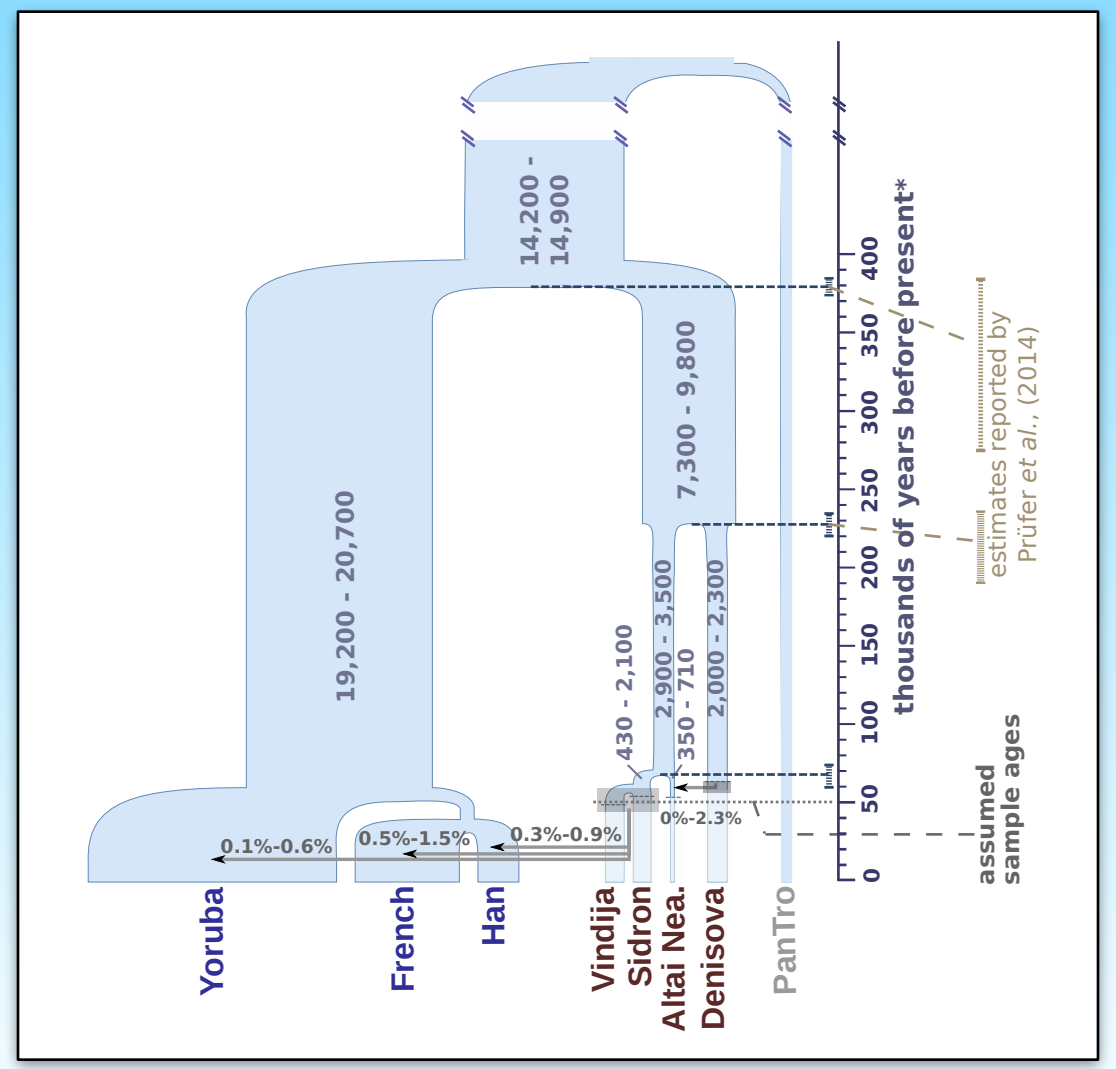
Freedman, Gronau et al., *PLOS Genetics*, 2014

Archaic Hominin Analysis



Sergi Castellano, Martin Kuhlwilm, Svante Paabo

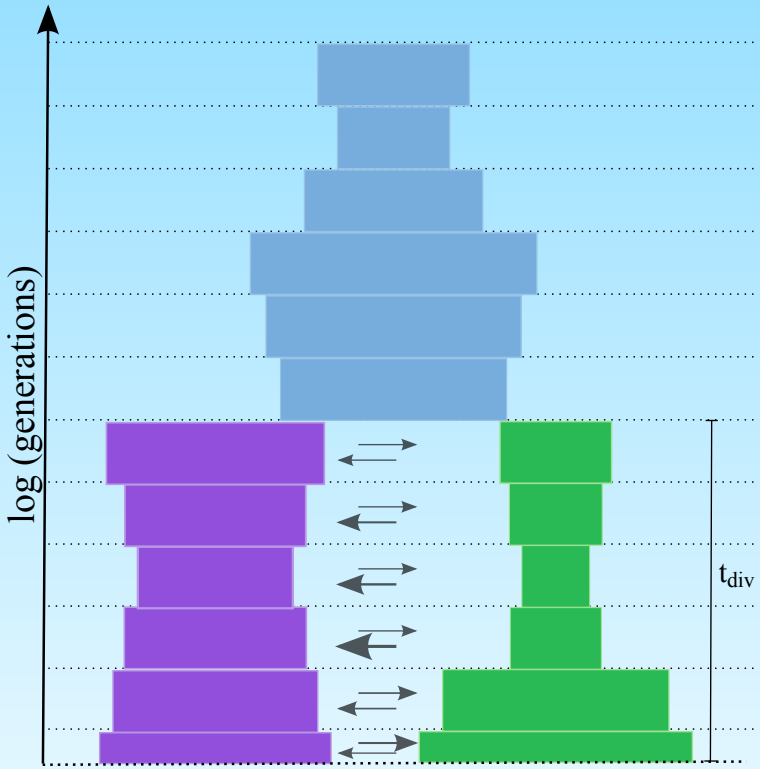
Best Model (So Far)



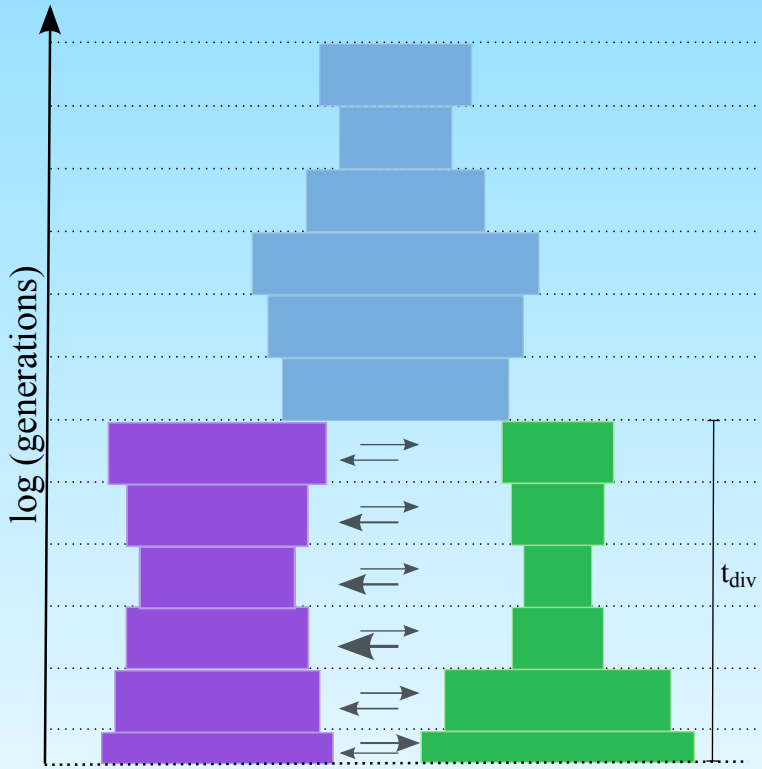
Limitations of G-PhoCS

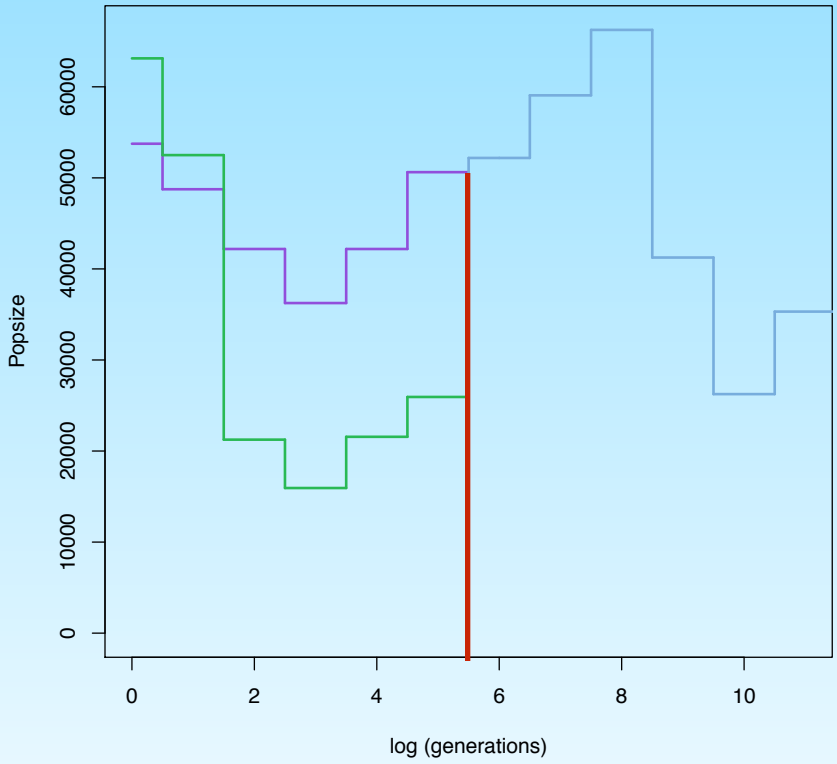
- Discards most of the data
- Must use short loci due to restrictive assumption of no intralocus recombination
- Fails to benefit from demographic information in LD structure
- Want to use full ARG!

First Goal: IM + *ARGweaver*

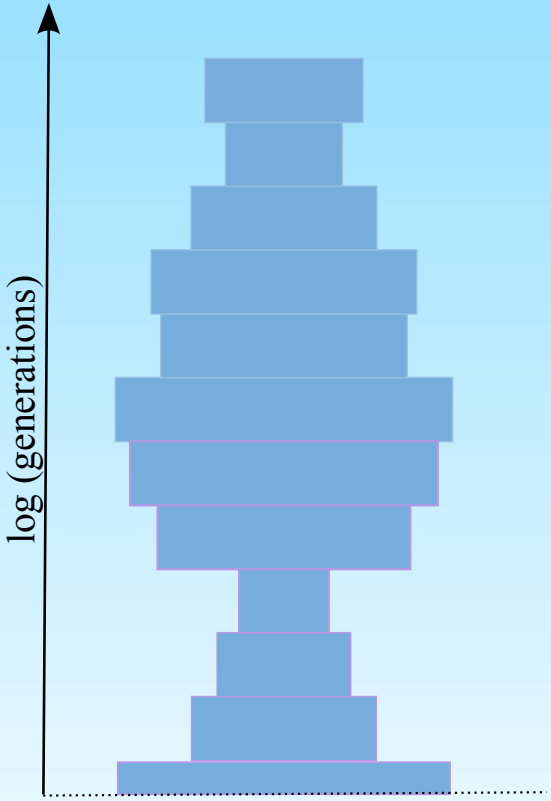


First Goal: IM + *ARGweaver*

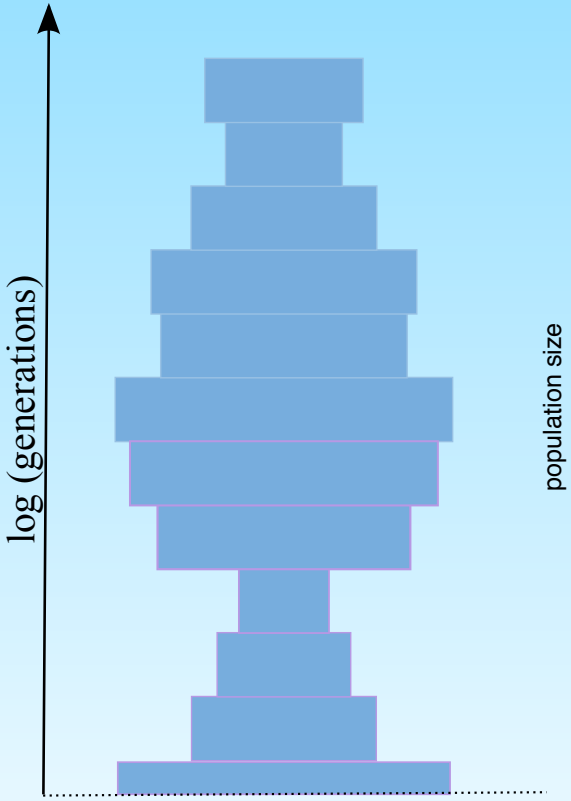


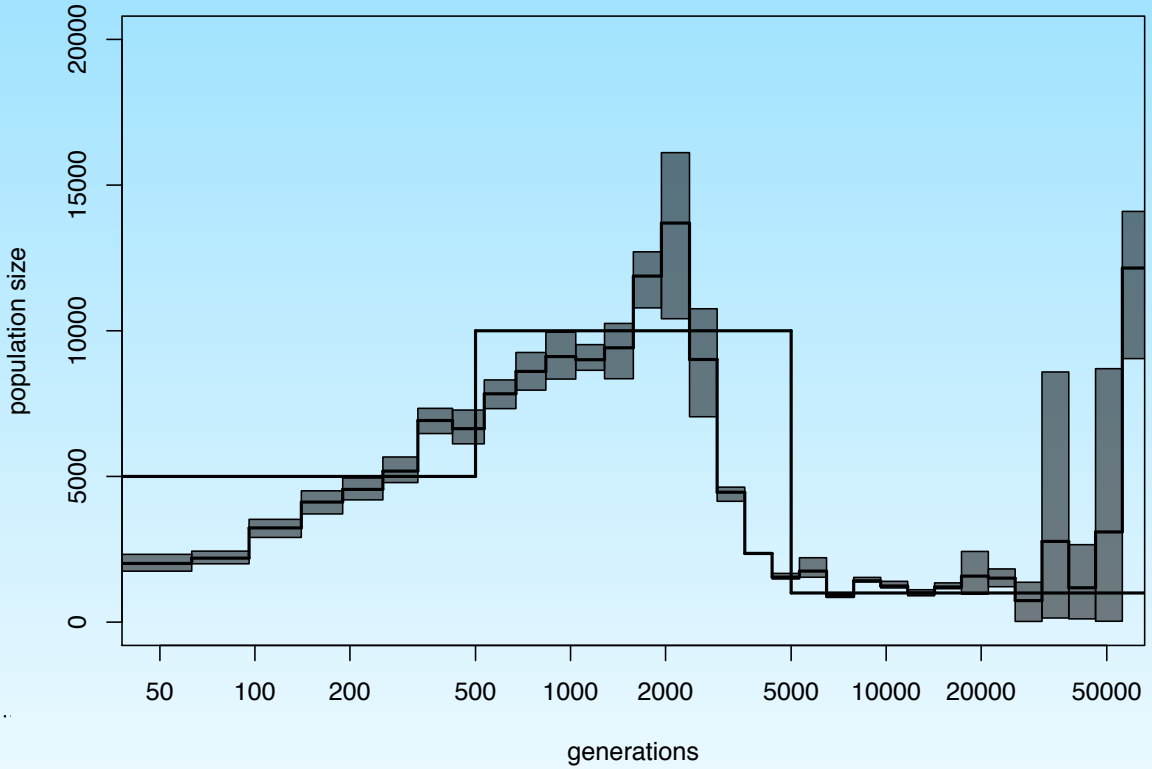


Preliminary Results



Preliminary Results





Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory

- Recently launched Center, with founding donation from Simons Foundation
- Focus on several areas of QB, including genomics, gene regulation, cancer biology, and neuroscience
- Faculty & fellow positions opening soon
- Several postdoc positions in my group
- See me if interested!

Acknowledgments

Contributors:, Matthew Rasmussen, Melissa Hubisz, Ilan Gronau

Other Group Members: Charles Danko, Andre Martins, Lenore Pipes, Brad Gulko, Jaaved Mohammed

Collaborators: John Novembre, Adam Freedman, Bob Wayne, Sergi Castellano, Martin Kuhlwilm, Svante Paabo

the David & Lucile Packard FOUNDATION

NATIONAL INSTITUTES OF HEALTH

Microsoft Research

NSF

ALFRED P. SLOAN FOUNDATION