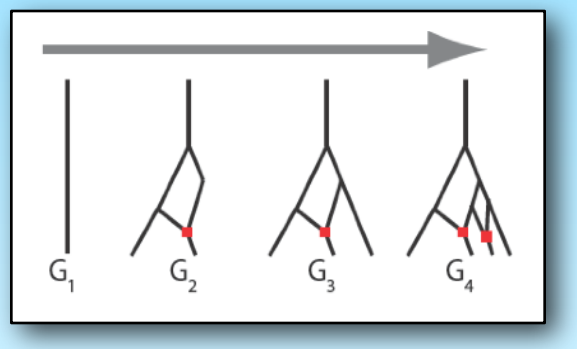
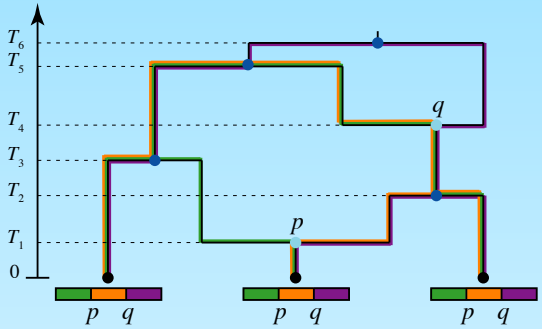
ARGweaver: Genome-wide Inference of Ancestral Recombination Graphs





Adam Siepel
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY

Cold Spring Harbor Laboratory

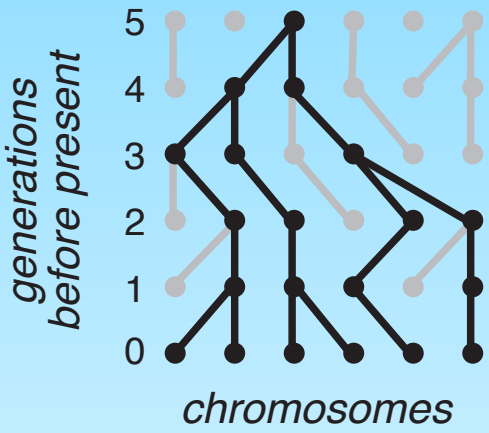
Joint work with Matthew Rasmussen, Melissa Hubisz, and Ilan Gronau

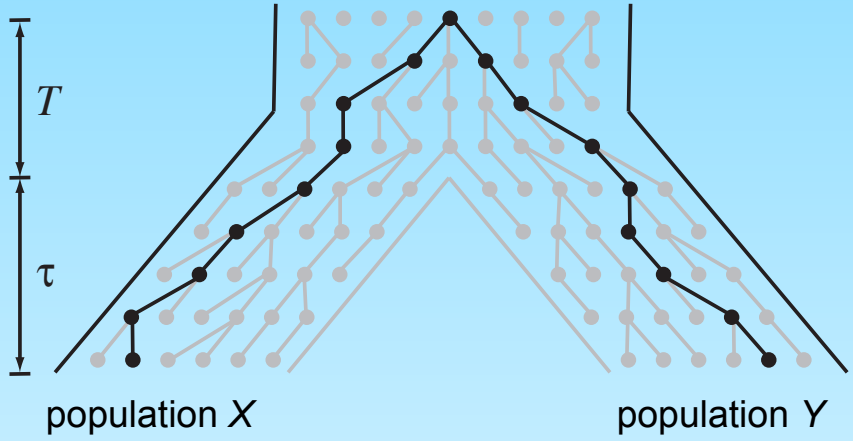
Cold Spring Harbor Laboratory

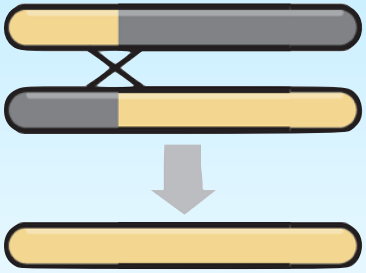
Simons Center for Quantitative Biology

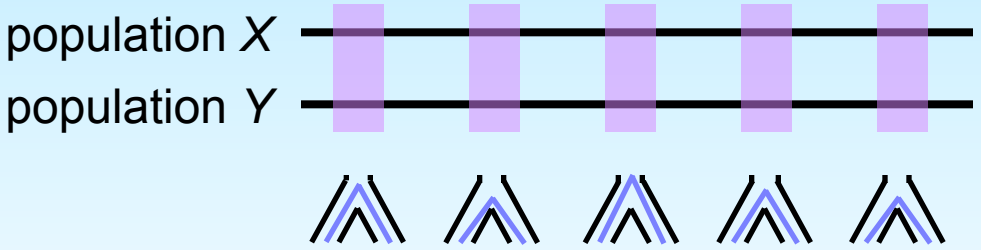
- Broad focus on theoretical and applied quantitative biology, including genomics, evolution, biophysics, cancer and neuroscience
- Founding $50M donation from Simons Foundation plus other gifts
- Ten current faculty members trained in physics, applied mathematics, and computer science
- Positions available at assistant professor, fellow, and postdoc levels

Recombination and Genealogies

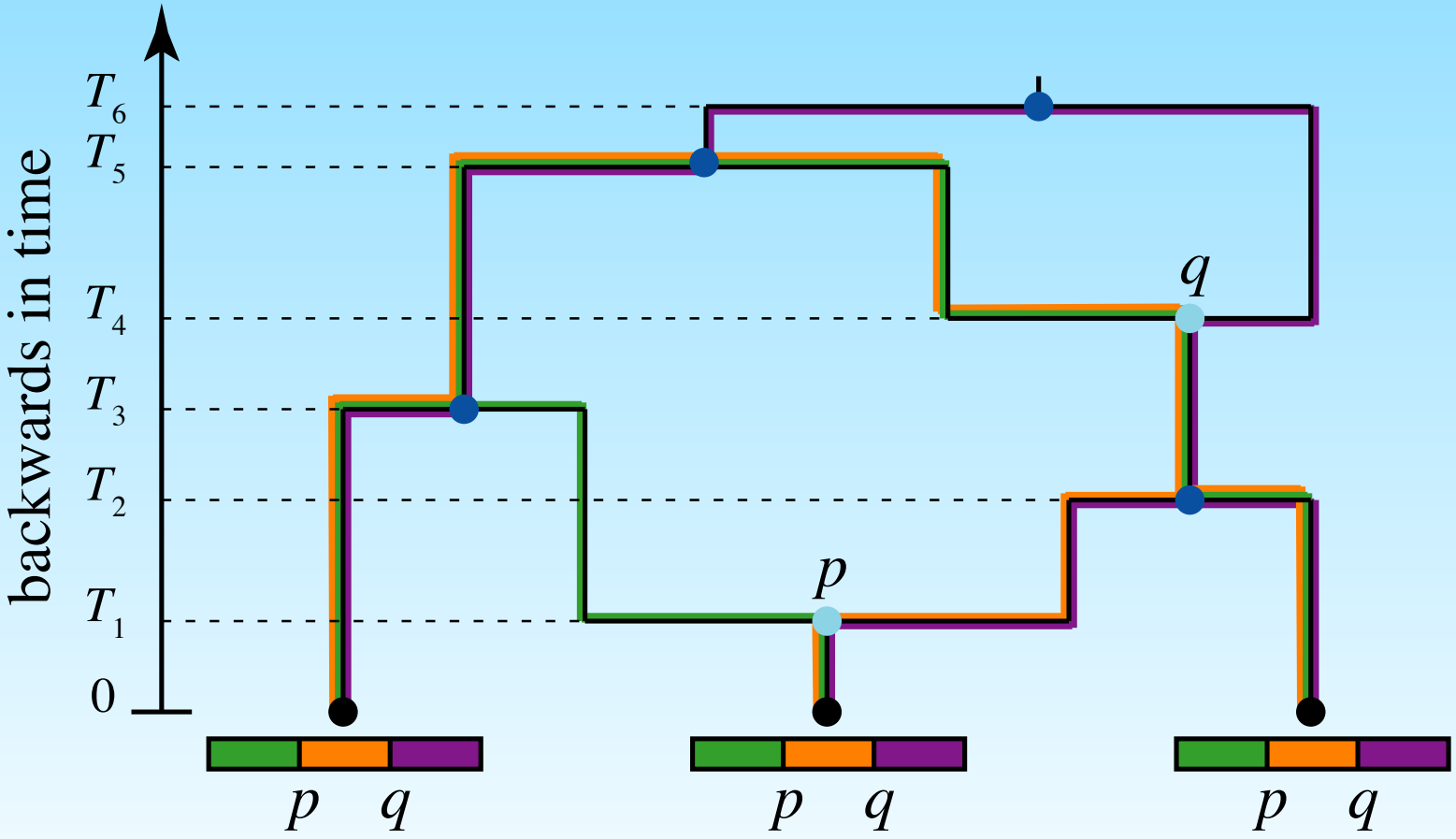








The Ancestral Recombination Graph



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
- …

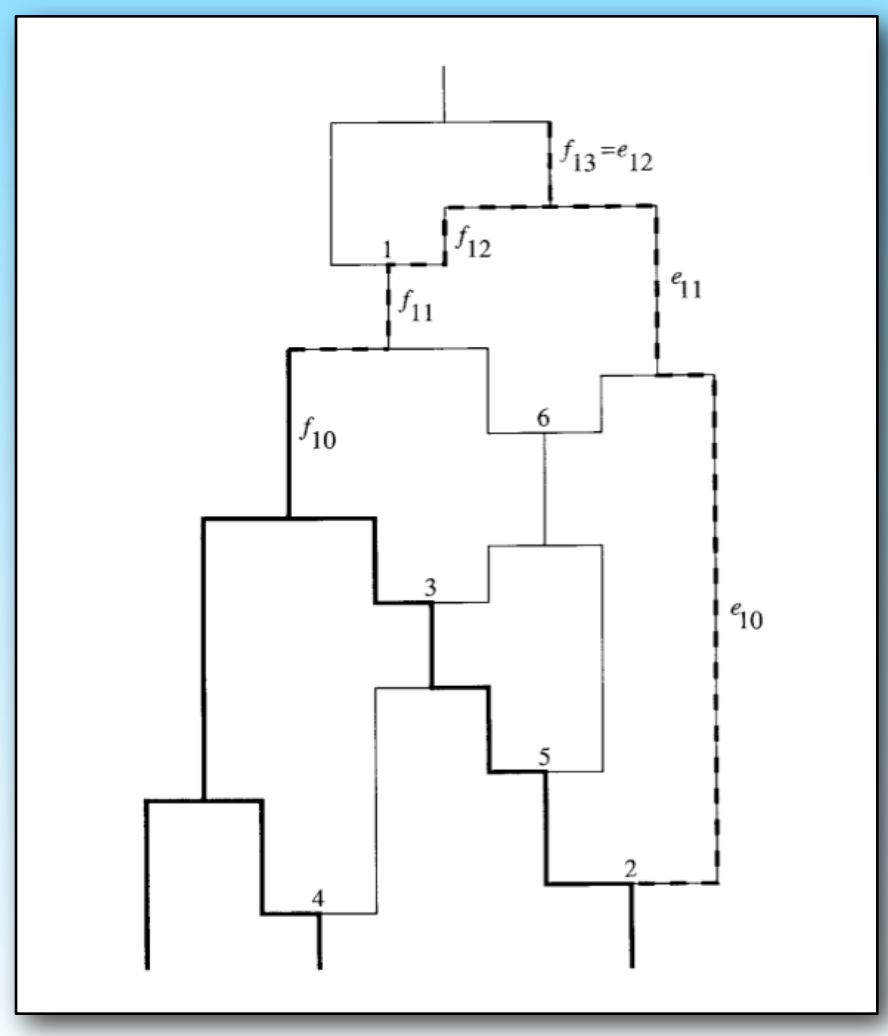
ARG surrogates: IBD, IBS, haplotypes, local ancestry inference, site-frequency spectrum, PCA

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.*, 1990
  - Kececioglu and Gusfield, *Disc. Appl. Math.*, 1998
  - Song and Hein, *J. Comput. Biol.*, 2005
  - Minichiello and Durbin, *Am. J. Hum. Genet.*, 2006

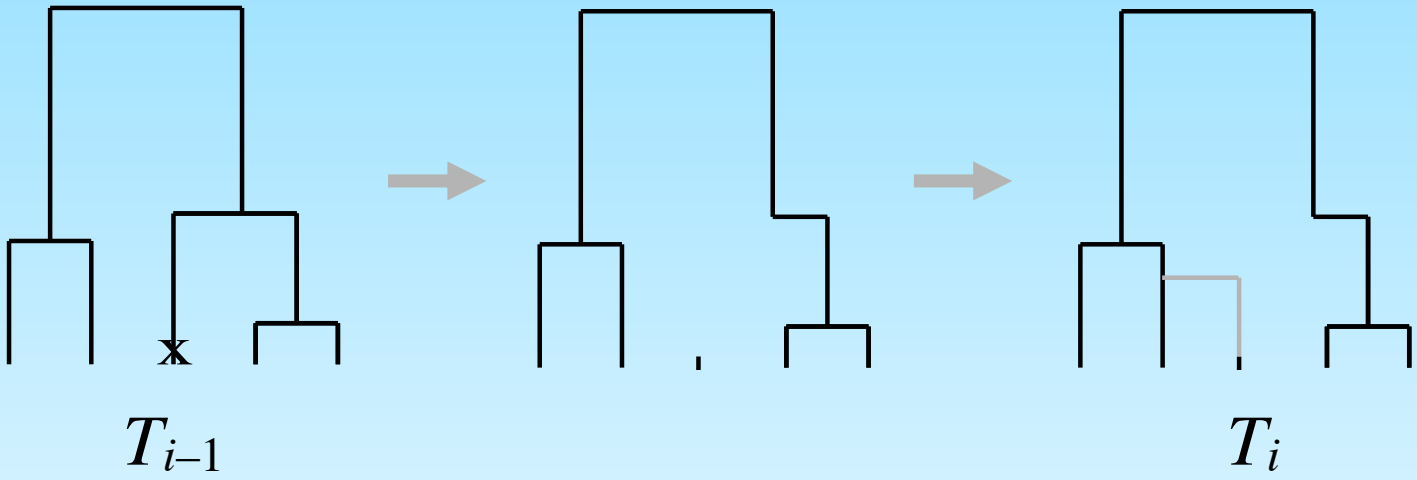
Computationally intensive
Limited to few samples
and/or
Depend on crude approximations

Sequential Coalescent with Recombination



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

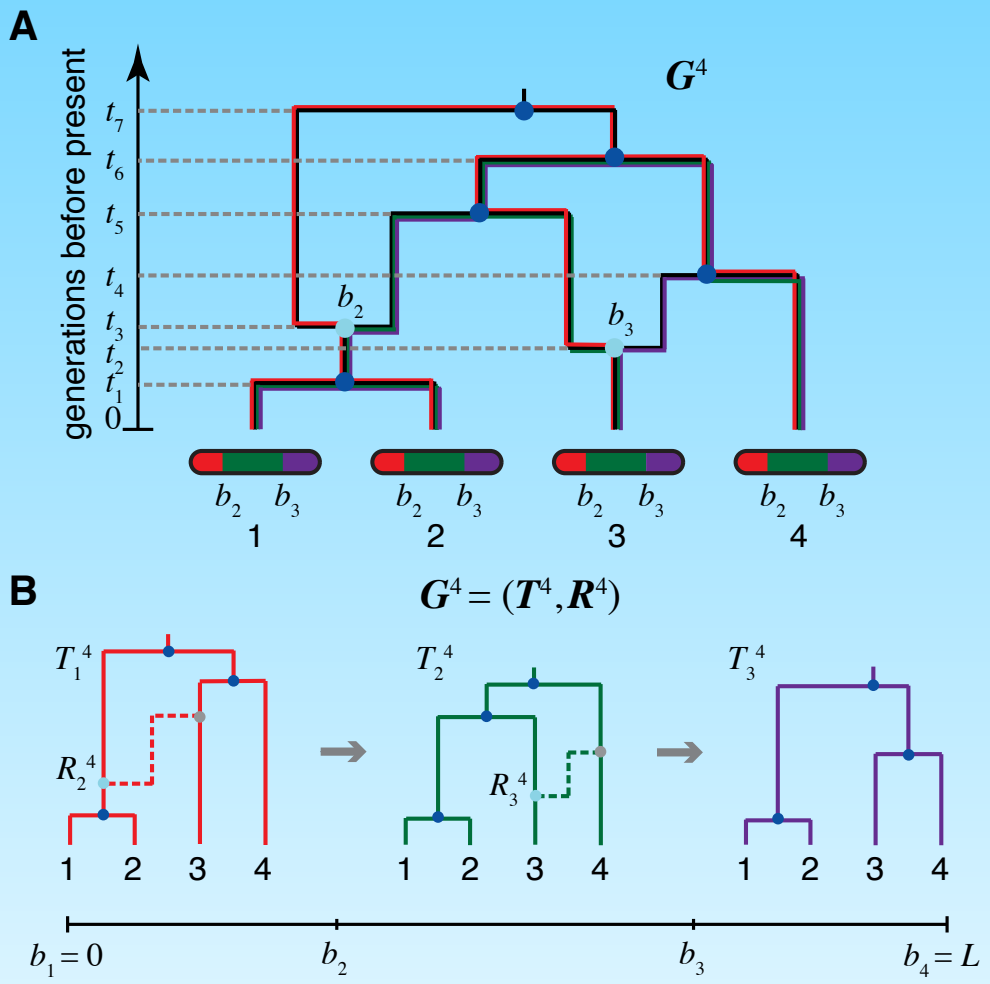
Sequentially Markov Coalescent (SMC)



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

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

McVean and Cardin, 2005; Marjoram and Wall, 2006

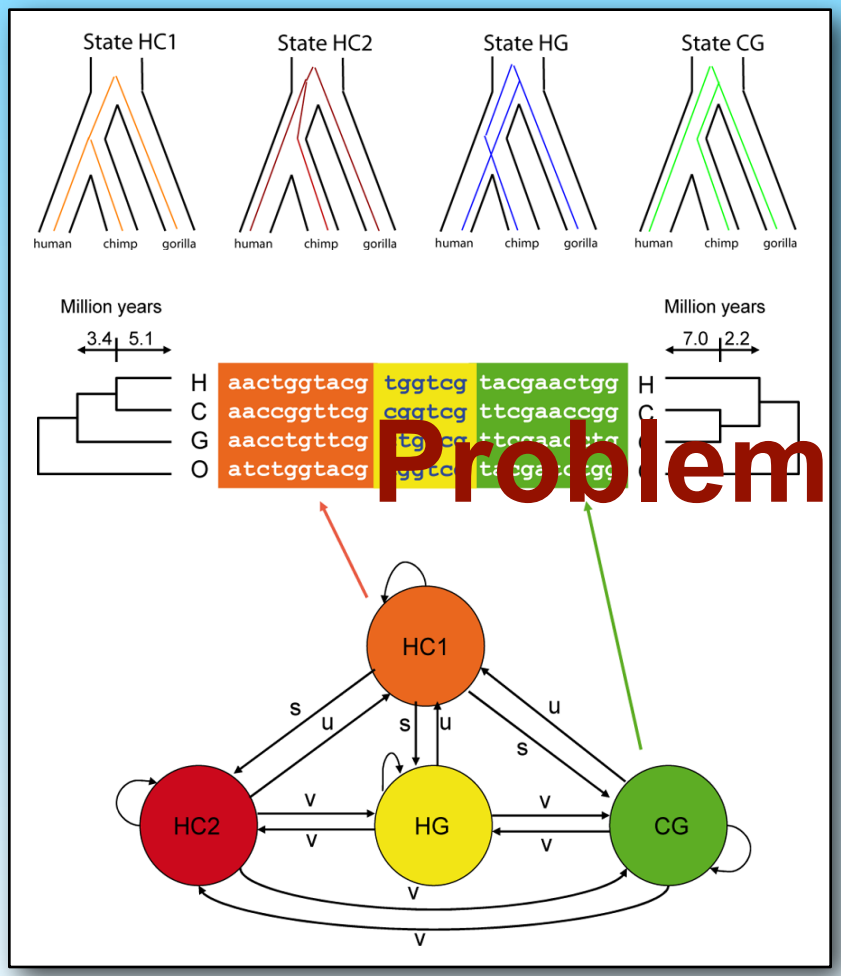


Rasmussen et al., *PLOS Genetics*, 2014

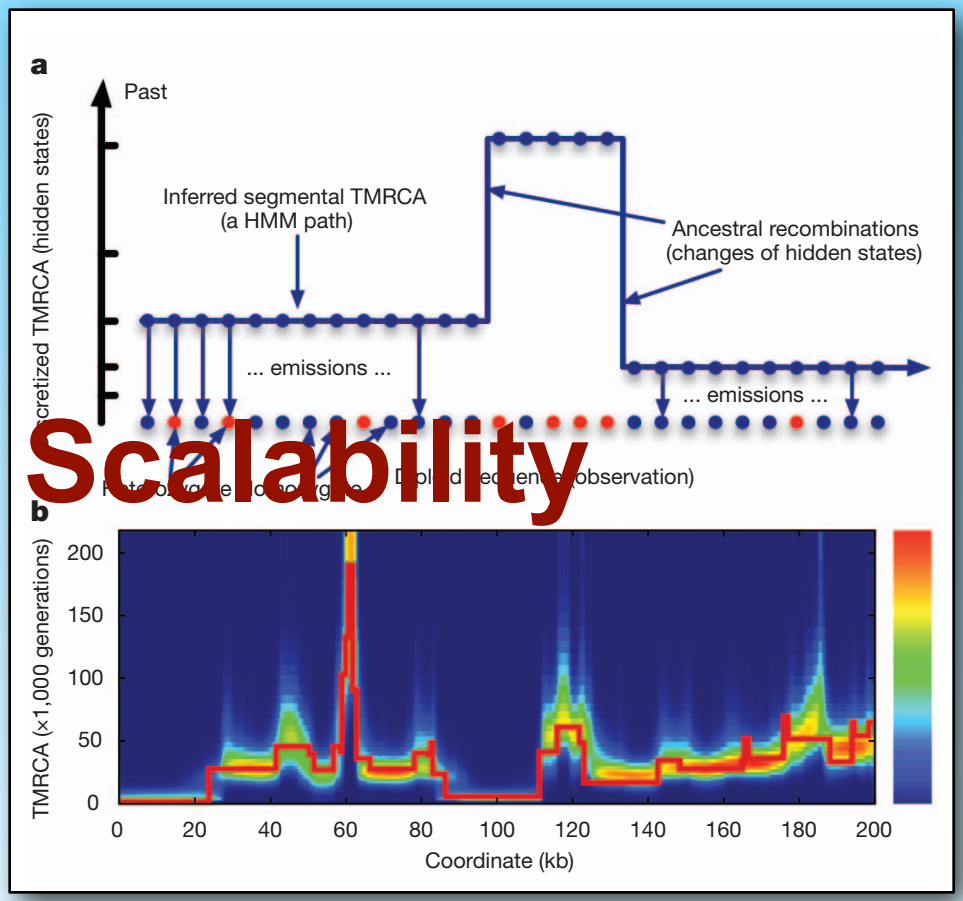
Discretized SMC and Hidden Markov Models

- 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

Problem: Scalability

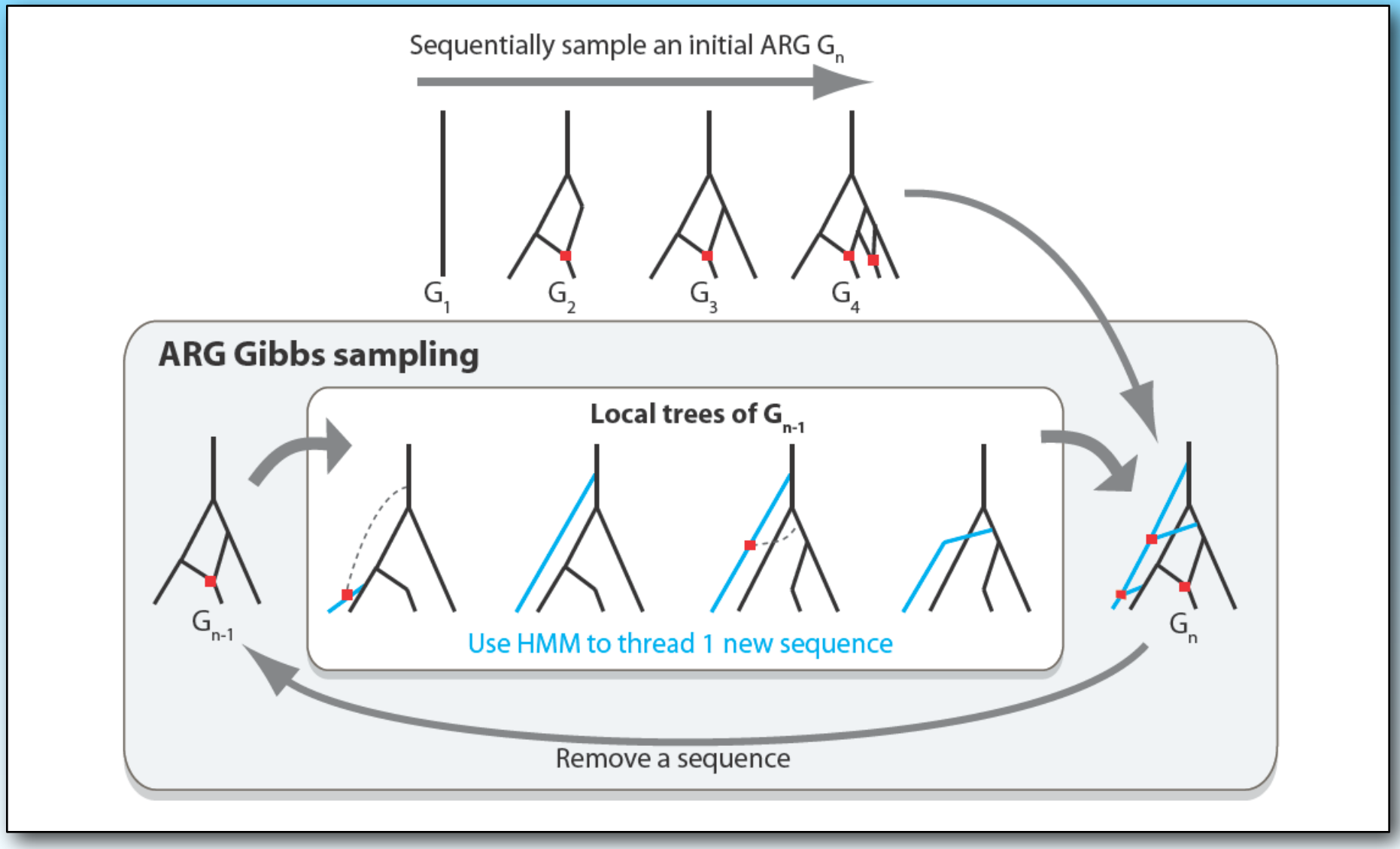
Hobolth et al., *PLOS Genet.*, 2007;
Li and Durbin, *Nature*, 2011

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

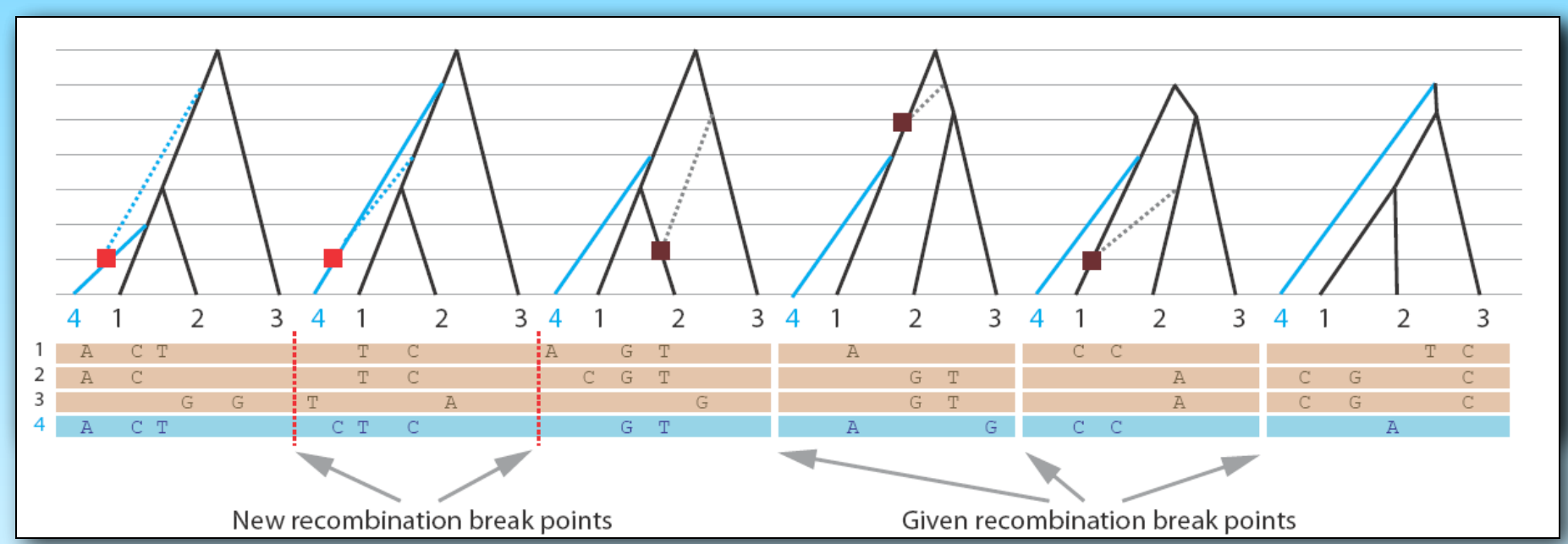
Rasmussen et al., *PLOS Genetics*, 2014

ARGweaver Sampling



Rasmussen et al., *PLOS Genetics,* 2014

Threading

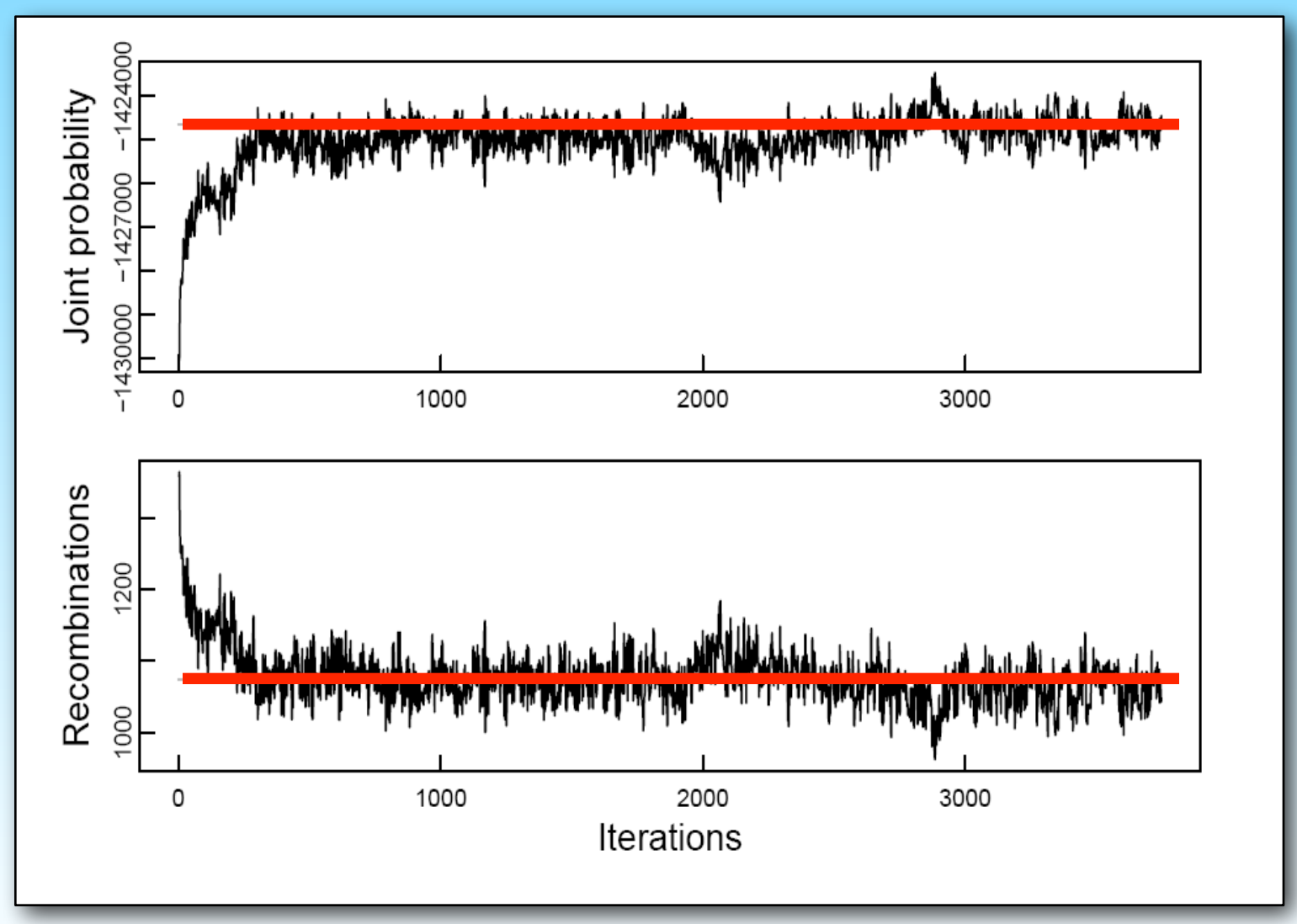


Solution: stochastic traceback with HMM

Cold Spring Harbor Laboratory

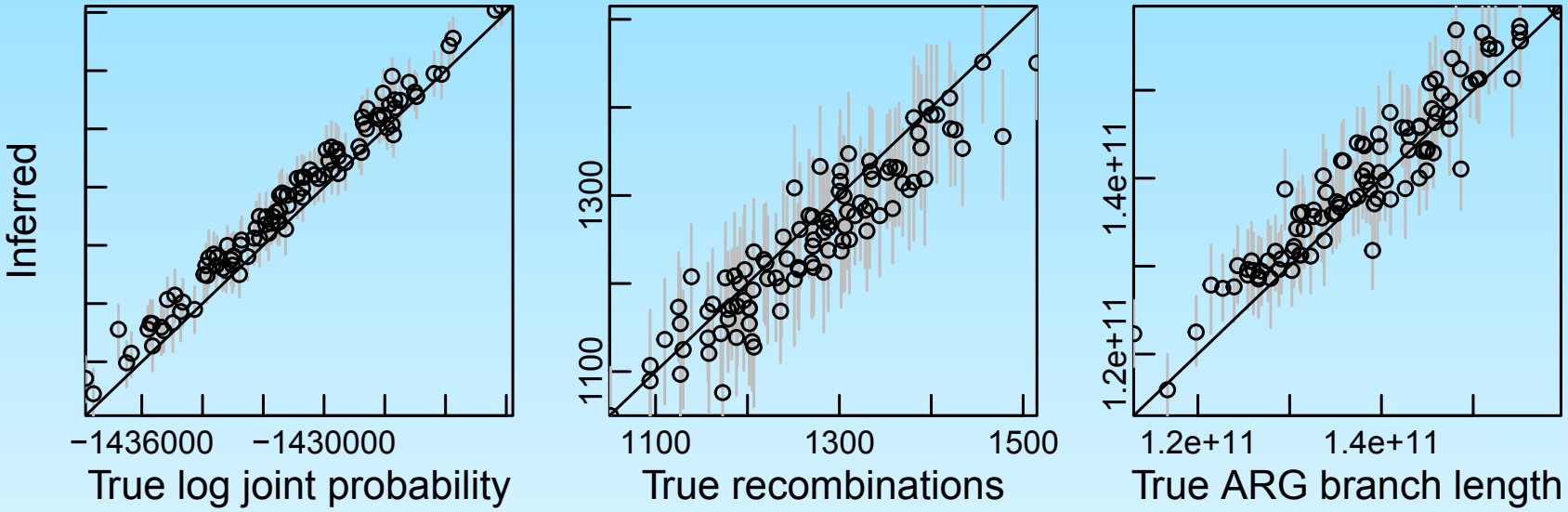
Rasmussen et al., *PLOS Genetics,* 2014

ARGweaver Converges Quickly



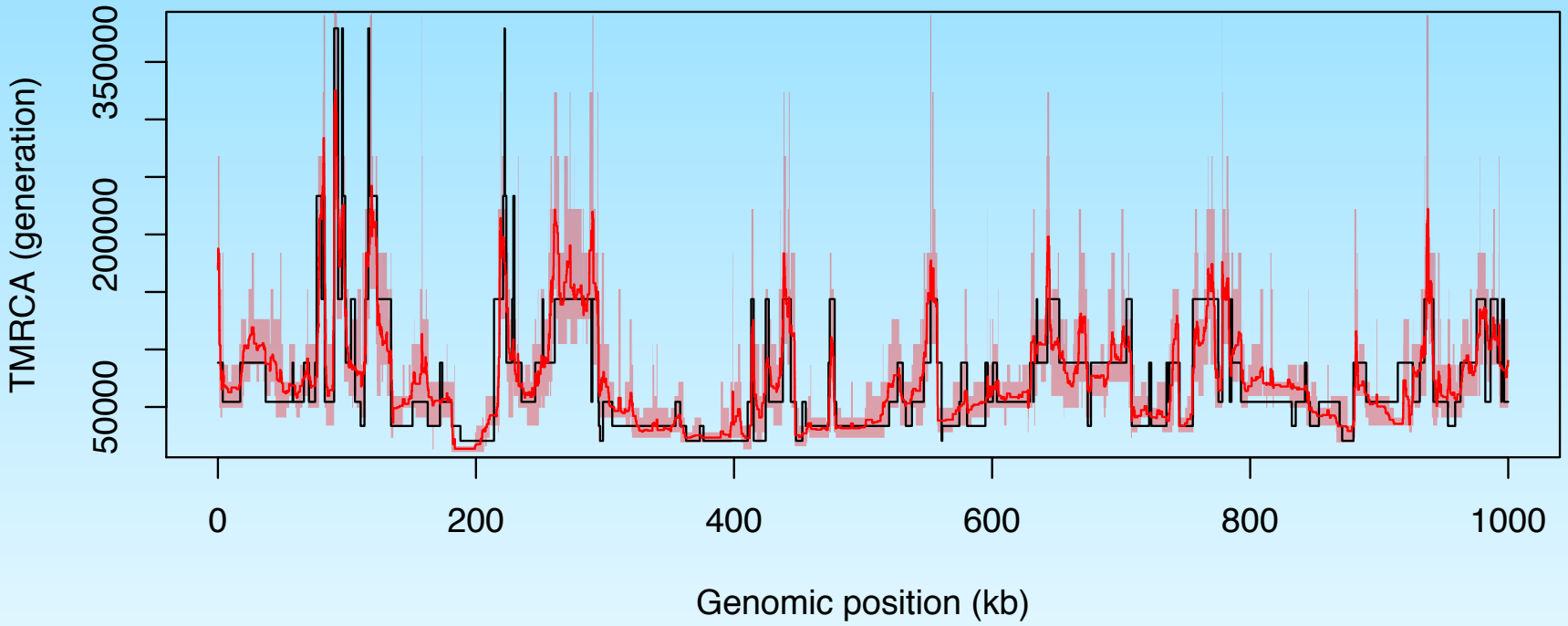
Rasmussen et al., *PLOS Genetics,* 2014

Recovery of Features of Simulated ARGs



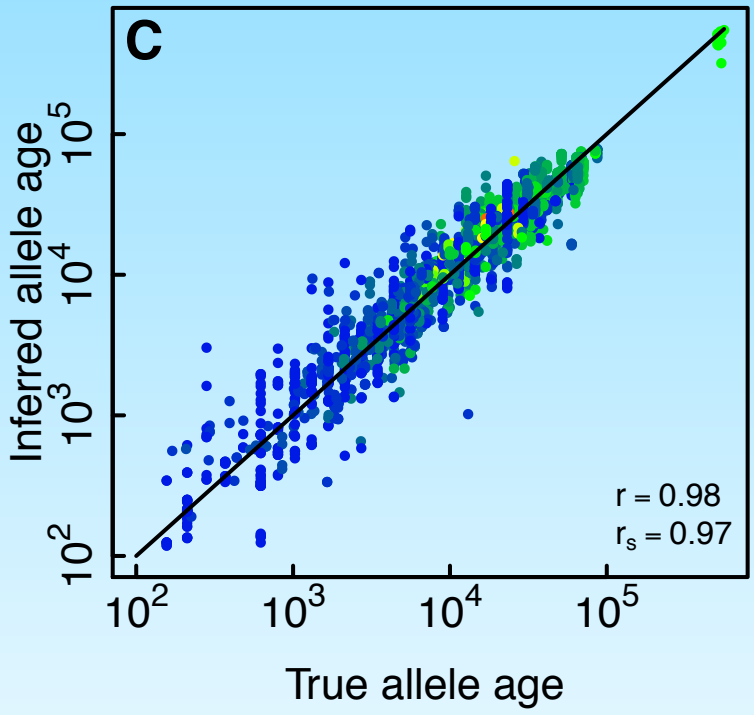
Rasmussen et al., *PLOS Genetics,* 2014

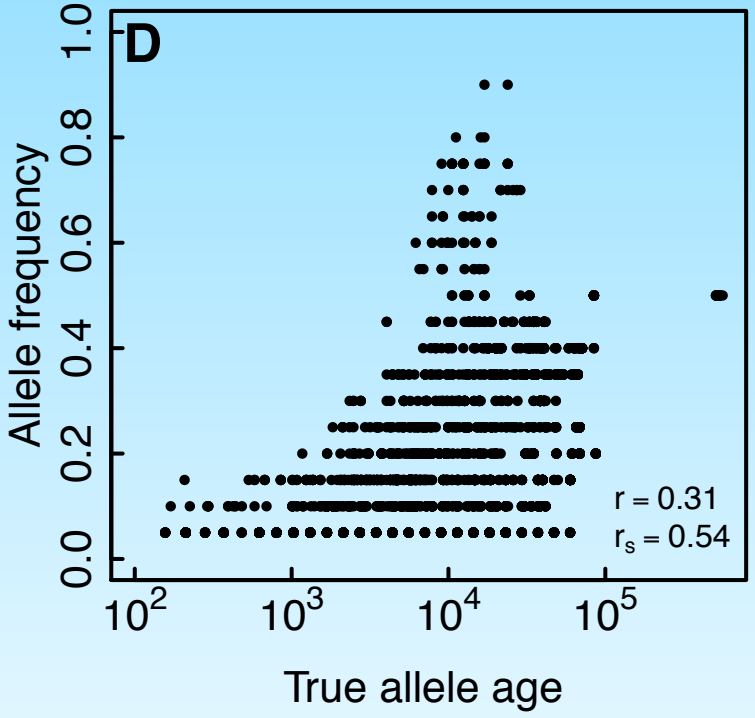
Recovery of Times to Most Recent Common Ancestry



Rasmussen et al., *PLOS Genetics,* 2014

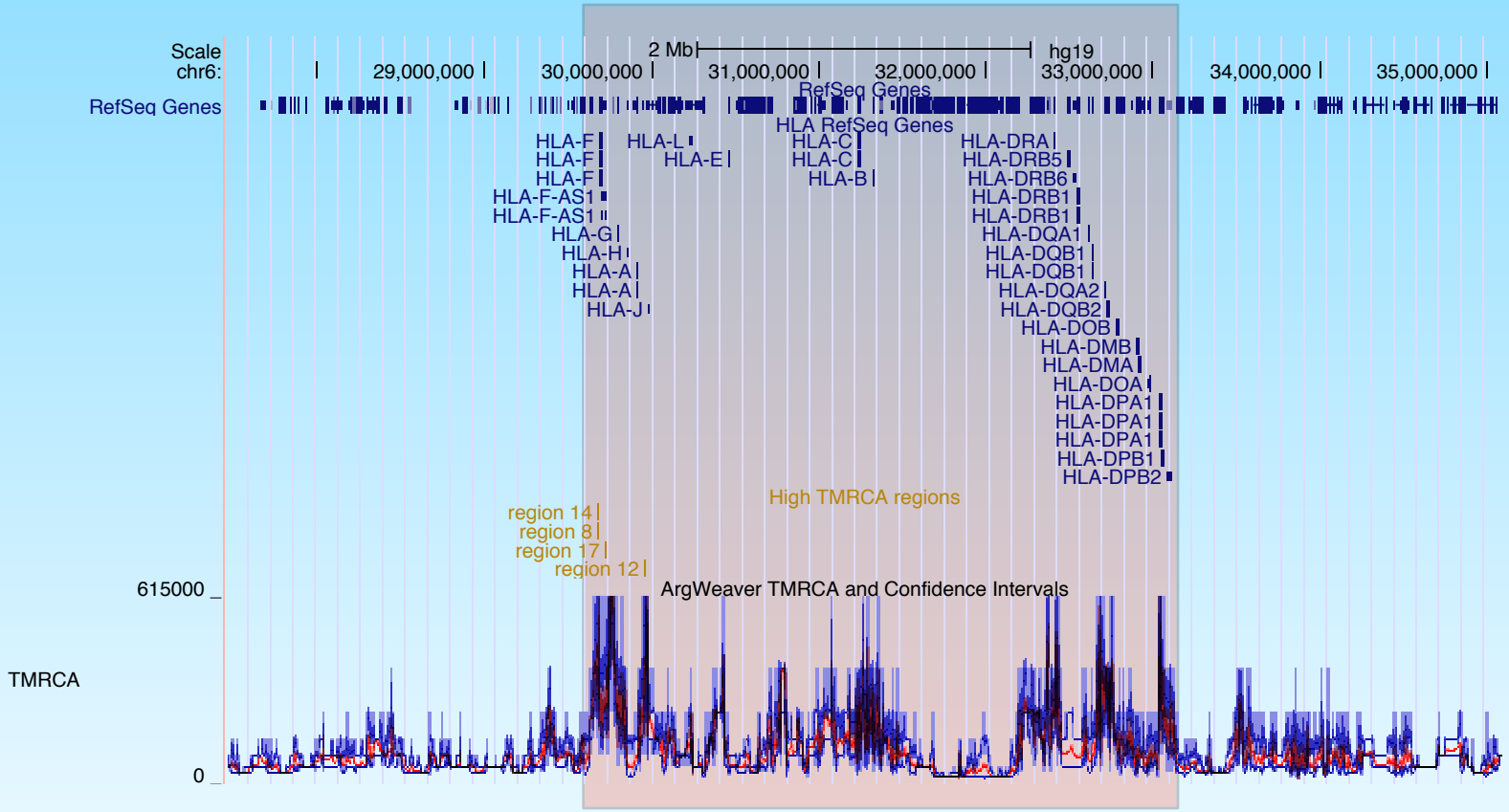
Recovery of Allele Age





Rasmussen et al., *PLOS Genetics,* 2014

Real Data: Regions of High TMRCA

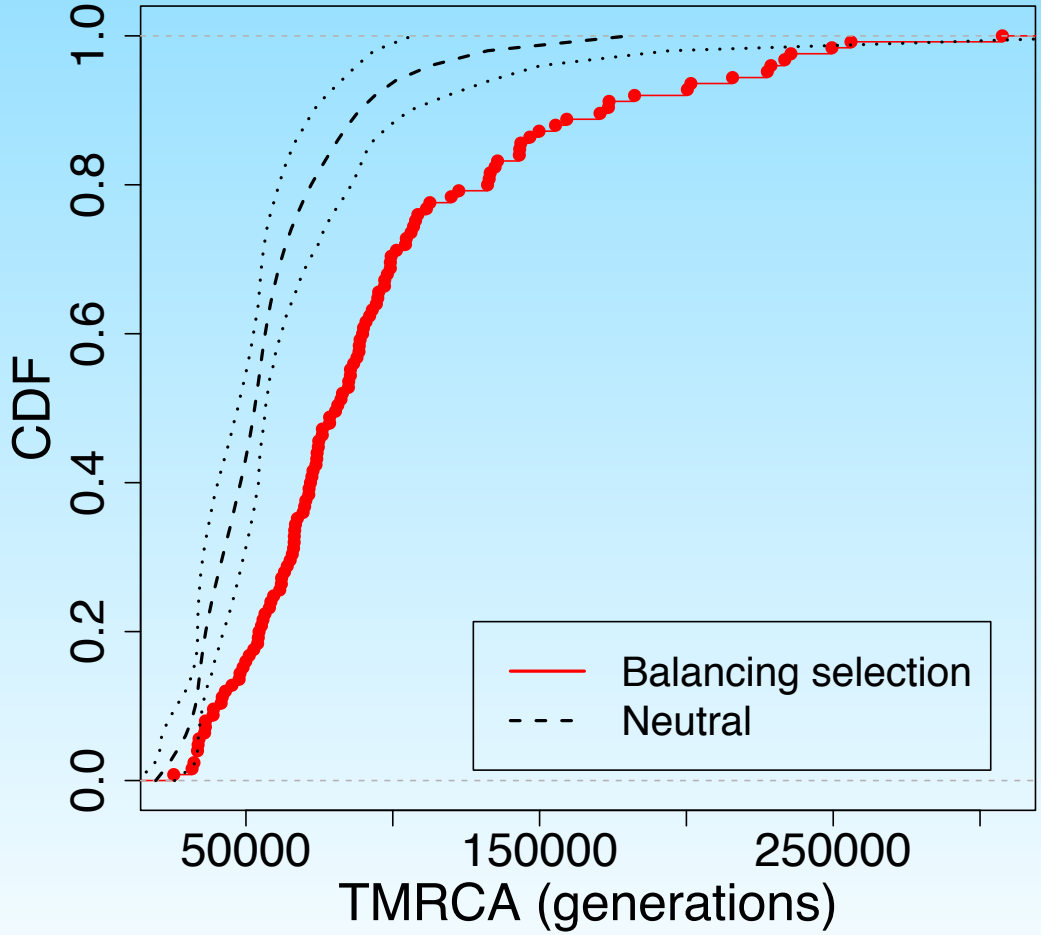


HLA Region

Cold Spring Harbor Laboratory

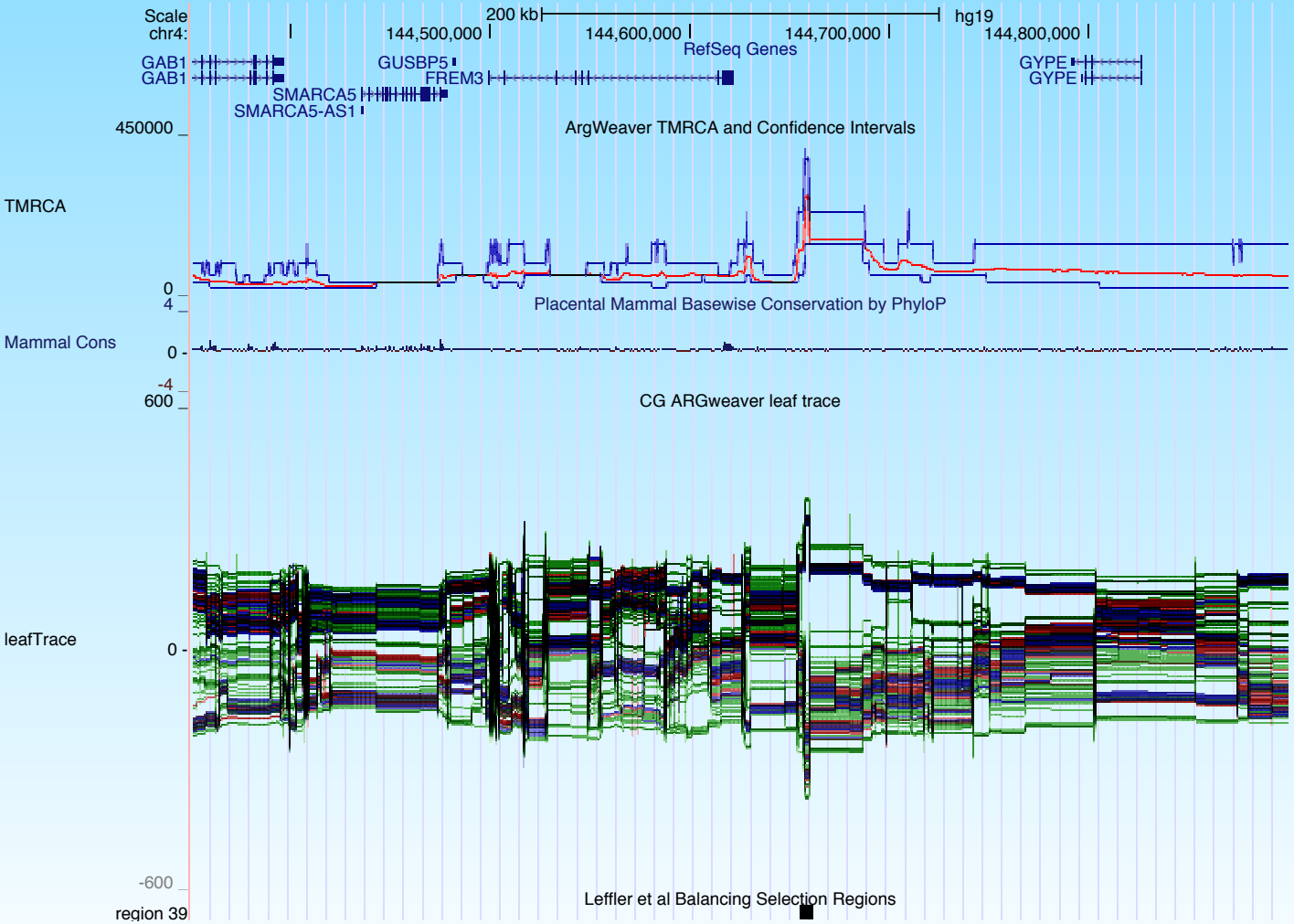
Rasmussen et al., *PLOS Genetics,* 2014

Regions of Shared Human/Chimp Polymorphism Have Old TMRCAs



Leffler et al., *Science*, 2013

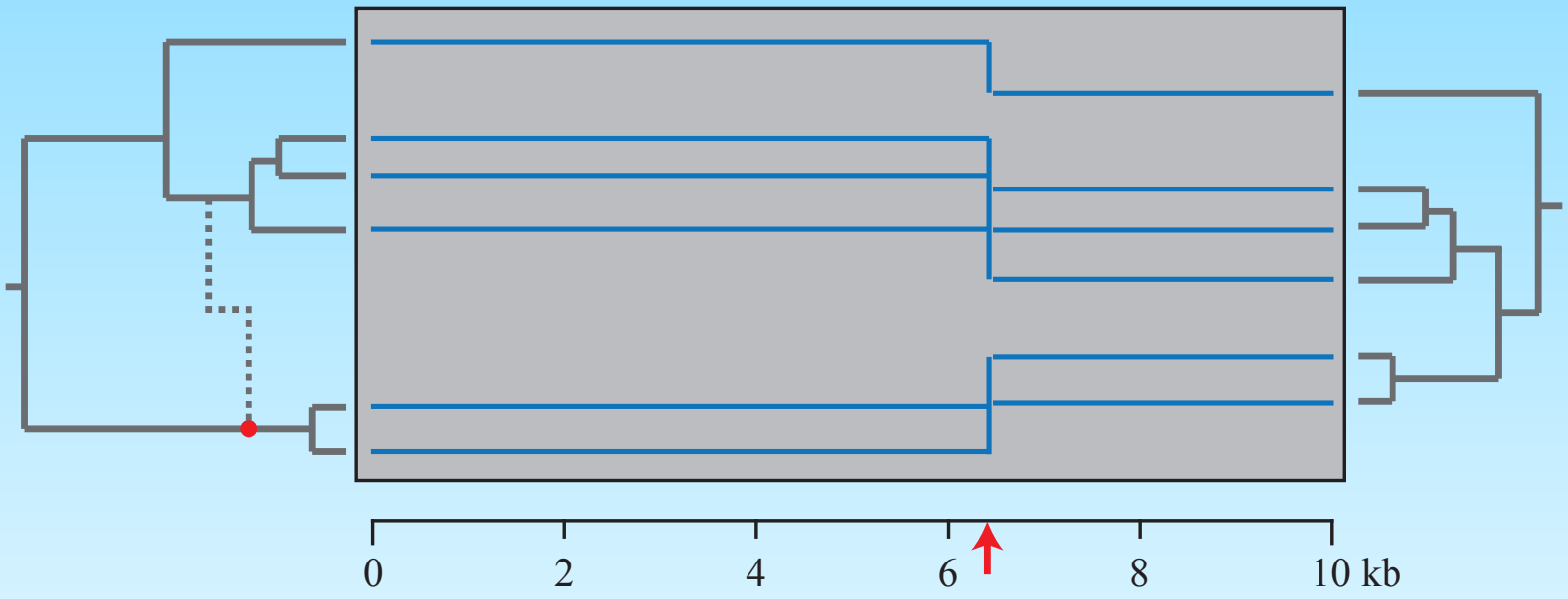
Putative Balancing Selection at FREM3



Cold Spring Harbor Laboratory

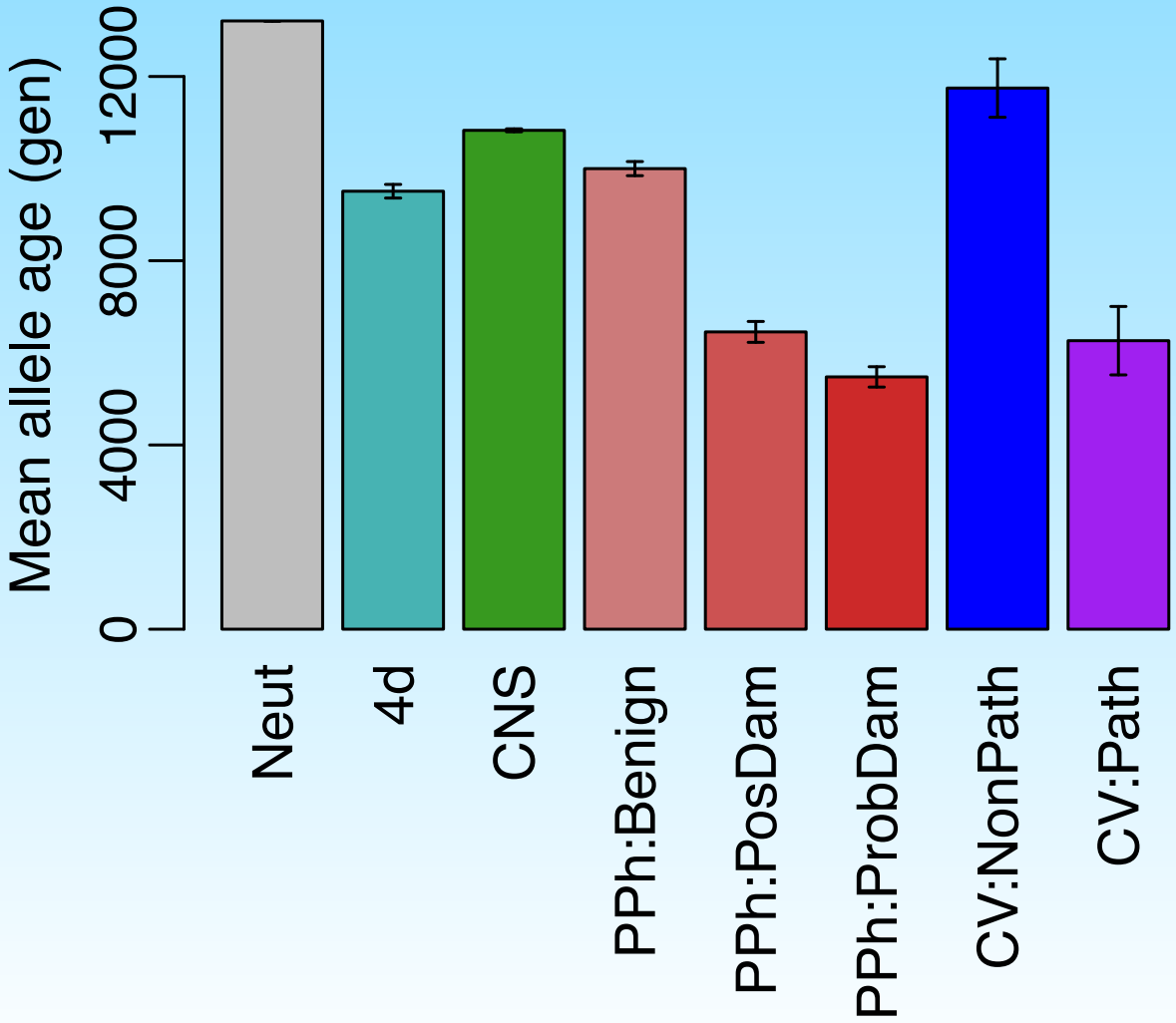
Leffler et al., *Science*, 2013

Leaf Trace

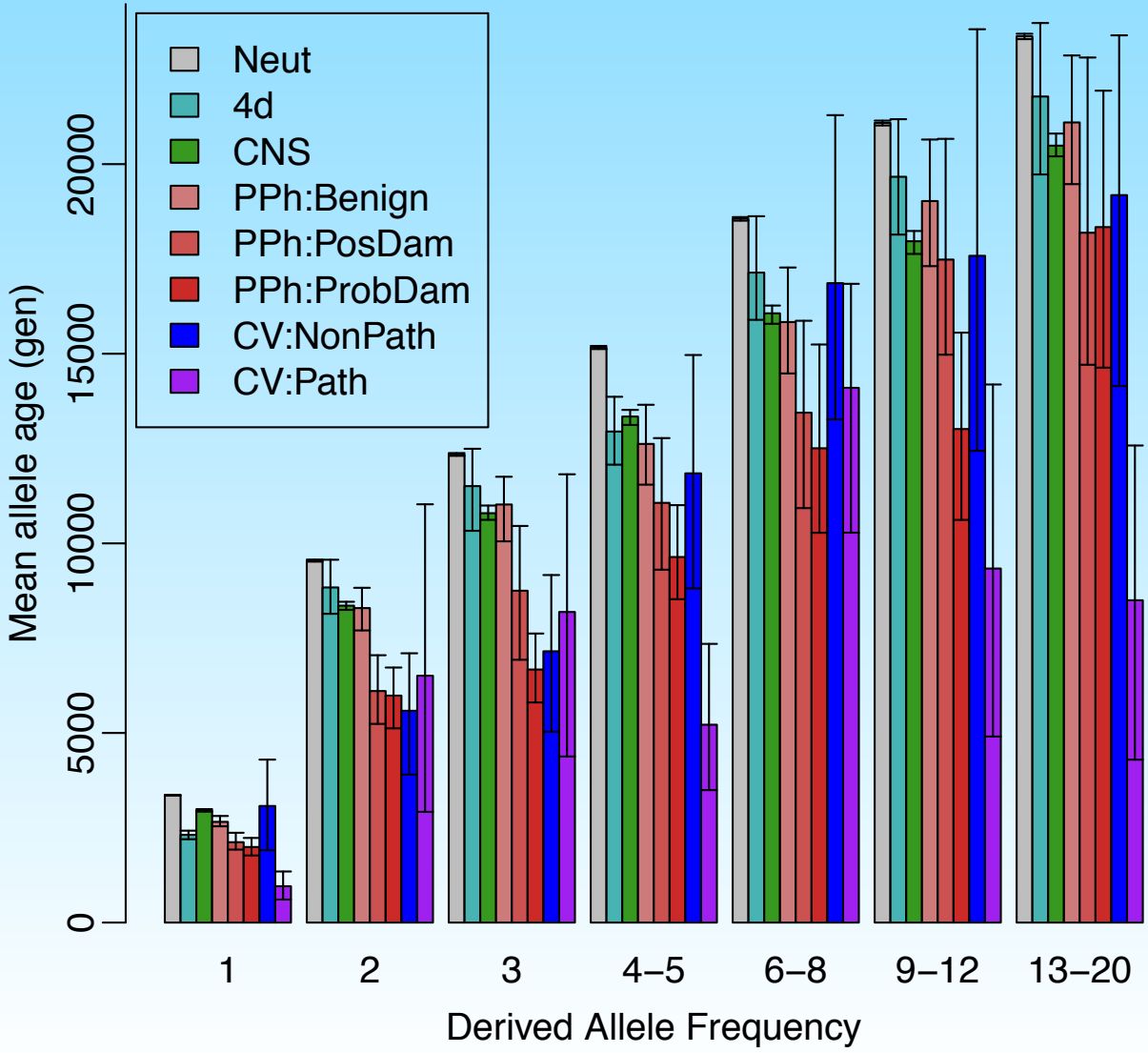


Rasmussen et al., *PLOS Genetics,* 2014

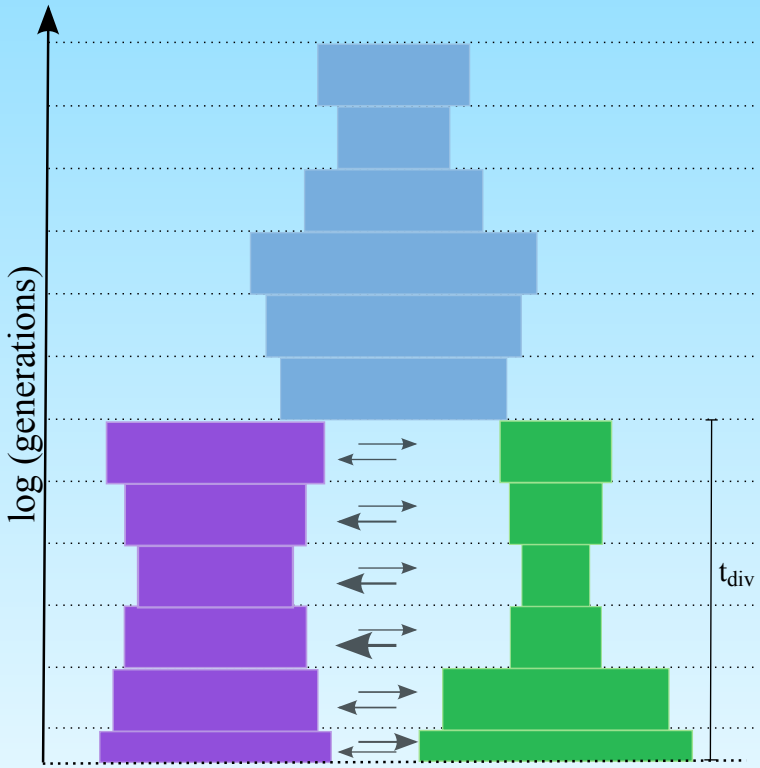
Sites Under Selection Have Decreased Allele Age

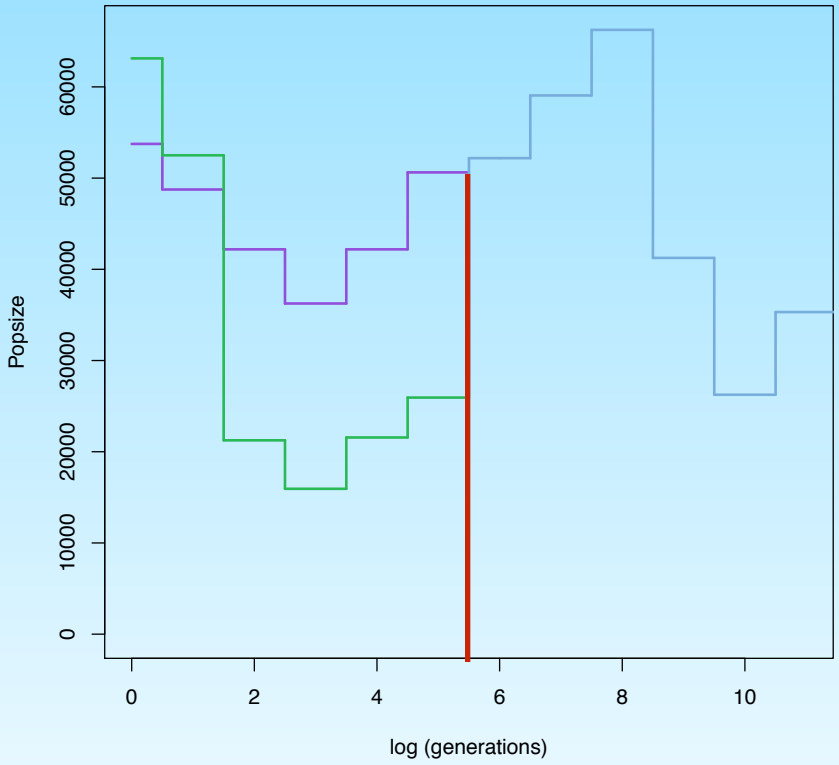


…Even After Accounting for Derived Allele Frequency



Current Work: IM + *ARGweaver*





Other Advances

- Joint phasing and threading—allows analysis of unphased genomes
- Clean-up of theory: rounding issues, “active branches”, etc.
- Watch for new paper from Melissa Hubisz

Acknowledgments

Contributors:, Matthew Rasmussen, Melissa Hubisz, Ilan Gronau

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

the David & Lucile Packard FOUNDATION

NATIONAL INSTITUTES OF HEALTH

Microsoft Research

NSF

ALFRED P. SLOAN FOUNDATION