

ABLE

Approximate Blockwise Likelihood Estimation

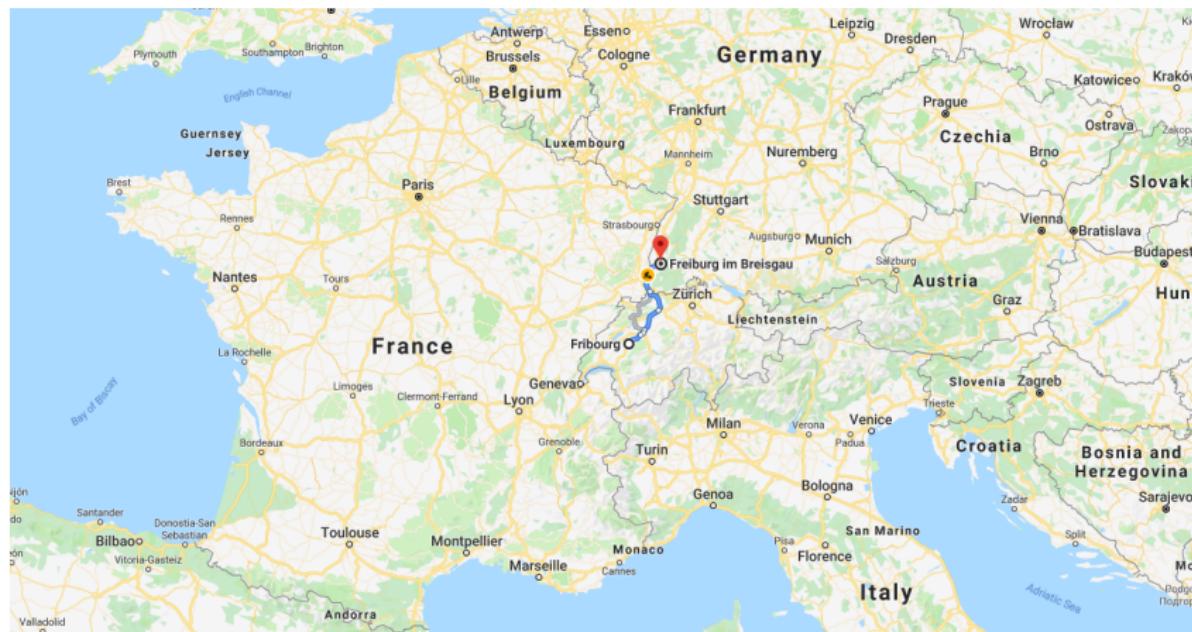
Champak Beeravolu Reddy

University of Fribourg

August 29, 2019

(From) Where am I !?

The University of "Fribourg"



Google Maps

The City of Fribourg

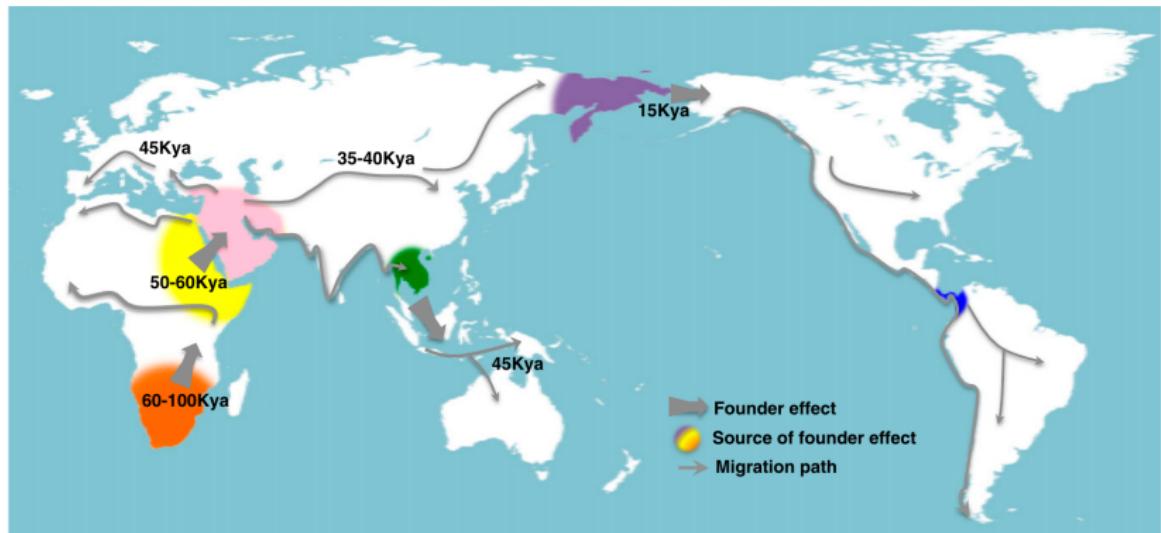


University of Fribourg

We move, and have been moving!



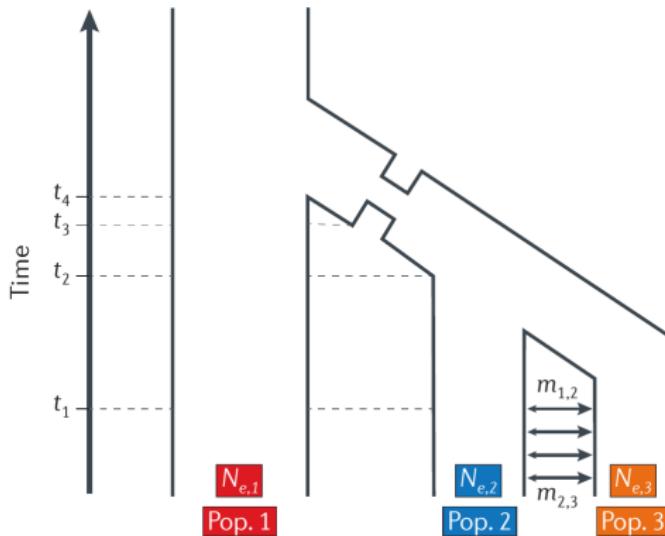
Picturing modern human migration



Henn et al. 2012

Simplifying a complex demography

modelling interacting panmictic units



Demographic inference using genomic data

The "two" major approaches

- Based on the SFS
 - No modelling of linkage
 - Usually no recombination
- Based on the haplotype structure
 - No intra-locus recombination
 - Tracts of IBD/IBS sharing
 - Recombination via the SMC

Aim : Find common ground between the two approaches

Demographic inference using genomic data

The "two" major approaches

- Based on the SFS
 - No modelling of linkage
 - Usually no recombination
- Based on the haplotype structure
 - No intra-locus recombination
 - Tracts of IBD/IBS sharing
 - Recombination via the SMC

Aim : Find common ground between the two approaches

Demographic inference using genomic data

The "two" major approaches

- Based on the SFS
 - No modelling of linkage
 - Usually no recombination
- Based on the haplotype structure
 - No intra-locus recombination
 - Tracts of IBD/IBS sharing
 - Recombination via the SMC

Aim : Find common ground between the two approaches

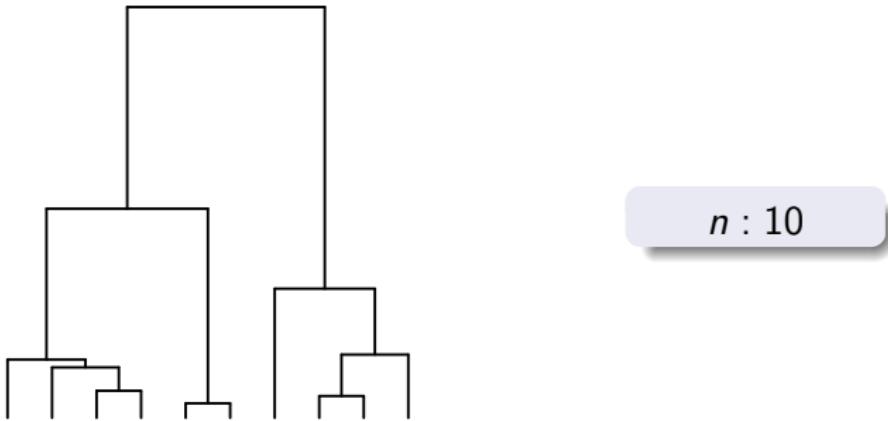
Demographic inference using genomic data

The "two" major approaches

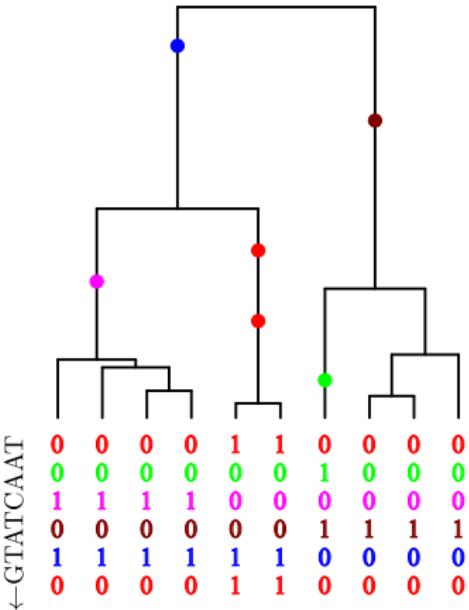
- Based on the SFS
 - No modelling of linkage
 - Usually no recombination
- Based on the haplotype structure
 - No intra-locus recombination
 - Tracts of IBD/IBS sharing
 - Recombination via the SMC

Aim : Find common ground between the two approaches

A gene genealogy

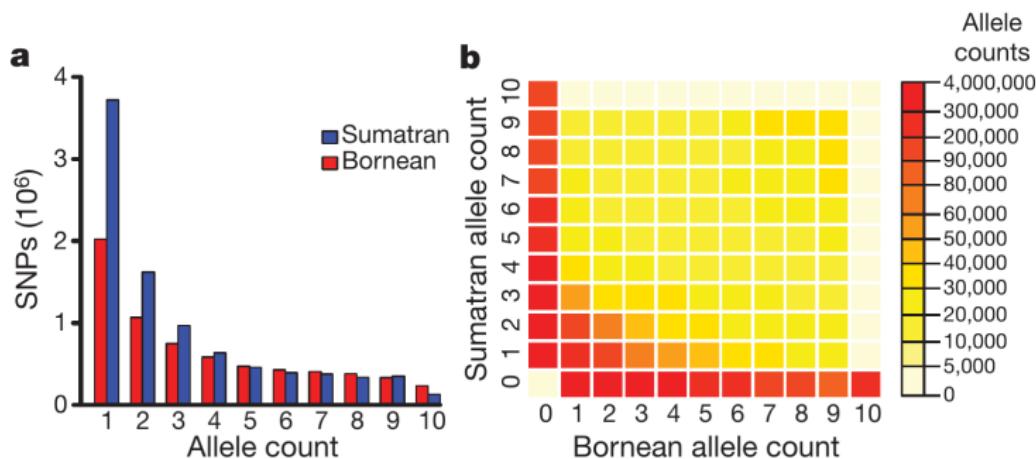


A gene genealogy with polymorphisms



$n : 10$
 $\mu : 6$

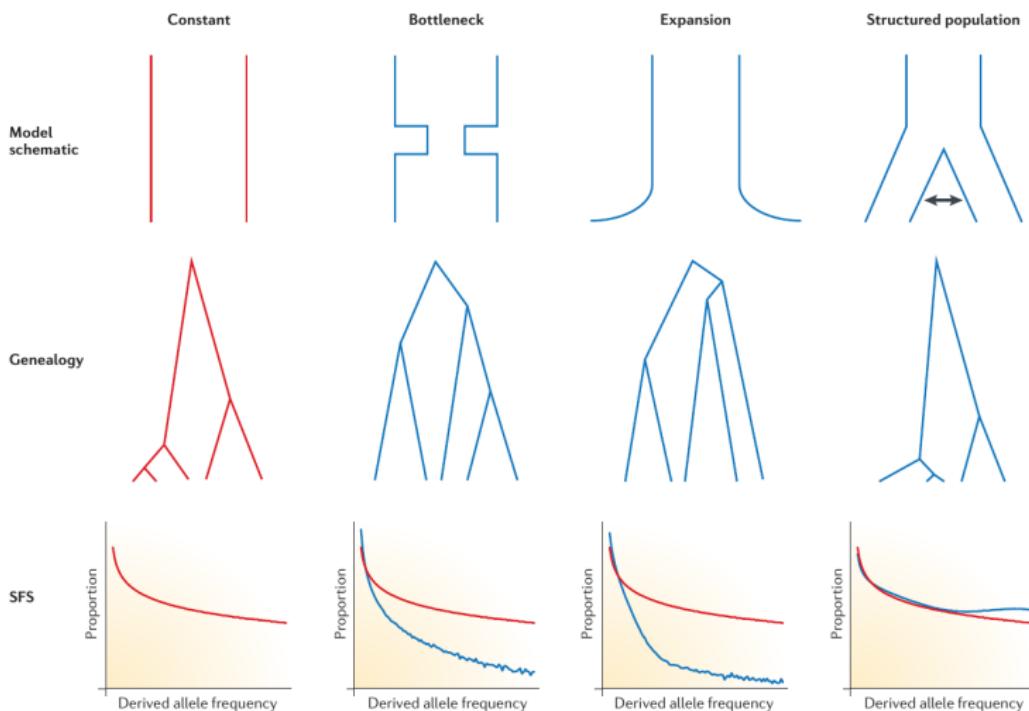
The Site Frequency Spectrum (SFS)



a. SFS

b. Joint SFS

Statistical identifiability and the SFS



Schraiber and Akey 2015

Extending the SFS

The Blockwise SFS (bSFS)

Short reads (*e.g.* RADSeq)



"Chopped" Genome



$$\mathcal{B}_{SFS} = (n_{\mathcal{B}_1}, n_{\mathcal{B}_2}, n_{\mathcal{B}_3}, n_{\mathcal{B}_4}, n_{\mathcal{B}_5}, n_{\mathcal{B}_6}, n_{\mathcal{B}_7}, \dots) = (1, 1, 3, 1, 2, 1, 1, \dots)$$

Calculating the likelihood for a given dataset

Felsenstein equation (discretized Chapman-Kolmogorov)

$$\begin{aligned}\mathcal{L}(\Theta) &\propto p(\mathcal{D} \mid \Theta) \\ &= \sum_{\mathcal{G}} p(\mathcal{D} \mid \mathcal{G}, \Theta) p(\mathcal{G} \mid \Theta)\end{aligned}$$

- ▶ \mathcal{D} : data
- ▶ \mathcal{G} : genealogy
- ▶ Θ : vector of demographic parameters

An exact analytical method makes use of the Generating Function of branch lengths

A General Method for Calculating Likelihoods Under the Coalescent Process

K. Lohse,^{*} R. J. Harrison,[†] and N. H. Barton^{*‡,†}

^{*}Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom, [†]East Malling Research, East Malling ME19 6BJ, United Kingdom, and [‡]Institute of Science and Technology, A-3400 Klosterneuburg, Austria

Neandertal Admixture in Eurasia Confirmed by Maximum-Likelihood Analysis of Three Genomes

Konrad Lohse^{*†} and Laurent A. F. Frantz[‡]

^{*}Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3JT, United Kingdom, and [†]Animal Breeding and Genomics Group, Wageningen University, De Elst 1, Wageningen, WD 6708, The Netherlands

Inferring Bottlenecks from Genome-Wide Samples of Short Sequence Blocks

Lynsey Bunnefeld,^{*†} Laurent A. F. Frantz,^{‡,2} and Konrad Lohse^{*}

^{*}Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom, and [‡]Animal Breeding and Genomics Centre, Wageningen University, Wageningen 6708 PB, The Netherlands

Lohse *et al.* 2011, Lohse & Frantz 2014, Bunnefeld *et al.* 2015

An exact analytical method makes use of the Generating Function (GF) of branch lengths

- Let μ be the overall mutation rate and t_b the expected length of a particular branch b
- The probability that there are k_b mutations on a specific branch b can be written as:

$$P[k_b] = E \left[e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} \right] = \int_0^{\infty} e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} P[t_b] dt_b$$

- This can be extended to the joint probability of observing a configuration \underline{k} of mutations oooooooooooo
- High cost of computation when $n > 6$ or $k_{max} > 8$ or $n_{pop} > 2$

An exact analytical method makes use of the Generating Function (GF) of branch lengths

- Let μ be the overall mutation rate and t_b the expected length of a particular branch b
- The probability that there are k_b mutations on a specific branch b can be written as:

$$P[k_b] = E \left[e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} \right] = \int_0^\infty e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} P[t_b] dt_b$$

- This can be extended to the joint probability of observing a configuration \underline{k} of mutations [Jump to genealogy](#)
- High cost of computation when $n > 6$ or $k_{max} > 8$ or $n_{pop} > 2$

An exact analytical method makes use of the Generating Function (GF) of branch lengths

- Let μ be the overall mutation rate and t_b the expected length of a particular branch b
- The probability that there are k_b mutations on a specific branch b can be written as:

$$P[k_b] = E \left[e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} \right] = \int_0^\infty e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} P[t_b] dt_b$$

- This can be extended to the joint probability of observing a configuration \underline{k} of mutations [Jump to genealogy](#)
- High cost of computation when $n > 6$ or $k_{max} > 8$ or $n_{pop} > 2$

An exact analytical method makes use of the Generating Function (GF) of branch lengths

- Let μ be the overall mutation rate and t_b the expected length of a particular branch b
- The probability that there are k_b mutations on a specific branch b can be written as:

$$P[k_b] = E \left[e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} \right] = \int_0^\infty e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} P[t_b] dt_b$$

- This can be extended to the joint probability of observing a configuration \underline{k} of mutations [Jump to genealogy](#)
- High cost of computation when $n > 6$ or $k_{max} > 8$ or $n_{pop} > 2$

Lohse *et al.* (2011)

An exact analytical method makes use of the Generating Function (GF) of branch lengths

- Let μ be the overall mutation rate and t_b the expected length of a particular branch b
- The probability that there are k_b mutations on a specific branch b can be written as:

$$P[k_b] = E \left[e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} \right] = \int_0^\infty e^{-\mu t_b} \frac{(\mu t_b)^{k_b}}{k_b!} P[t_b] dt_b$$

- This can be extended to the joint probability of observing a configuration \underline{k} of mutations [Jump to genealogy](#)
- High cost of computation when $n > 6$ or $k_{max} > 8$ or $n_{pop} > 2$

Approximating the bSFS likelihood

Felsenstein equation (discretized Chapman-Kolmogorov)

$$\begin{aligned}\mathcal{L}(\Theta) &\propto p(\mathcal{D} \mid \Theta) \\ &= \sum_{\mathcal{G}} p(\mathcal{D} \mid \mathcal{G}, \Theta) p(\mathcal{G} \mid \Theta)\end{aligned}$$

Approximating the bSFS likelihood

Felsenstein equation (discretized Chapman-Kolmogorov)

$$\begin{aligned}\mathcal{L}(\Theta) &\propto p(\mathcal{D} \mid \Theta) \\ &= \sum_{\mathcal{G}} p(\mathcal{D} \mid \mathcal{G}, \Theta) p(\mathcal{G} \mid \Theta)\end{aligned}$$

Sampling genealogies $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_M$ from $p(\mathcal{G} \mid \Theta)$ yields a Monte Carlo estimator of the bSFS likelihood

$$p(\mathcal{B}_{SFS} \mid \Theta) \approx \frac{1}{M} \sum_{i=1}^M p(\mathcal{B}_{SFS} \mid \mathcal{G}_i, \Theta)$$

buildABLE

The pseudo-code:

- ① Simulate a random \mathcal{G}_i for a given demography Θ
- ② Calculate the probability of observing $1 \dots k_{max}$ mutations on every branch class of \mathcal{G}_i
- ③ Calculate the probability of observing every $\mathcal{B}_j \in \mathcal{B}_{SFS}$
- ④ Increment i and restart from (1) if $i < M$

buildABLE

The pseudo-code:

- ① Simulate a random \mathcal{G}_i for a given demography Θ
- ② Calculate the probability of observing $1 \dots k_{max}$ mutations on every branch class of \mathcal{G}_i
- ③ Calculate the probability of observing every $\mathcal{B}_j \in \mathcal{B}_{SFS}$
- ④ Increment i and restart from (1) if $i < M$

buildABLE

The pseudo-code:

- ① Simulate a random \mathcal{G}_i for a given demography Θ
- ② Calculate the probability of observing $1 \dots k_{max}$ mutations on every branch class of \mathcal{G}_i
- ③ Calculate the probability of observing every $\mathcal{B}_j \in \mathcal{B}_{SFS}$
- ④ Increment i and restart from (1) if $i < M$

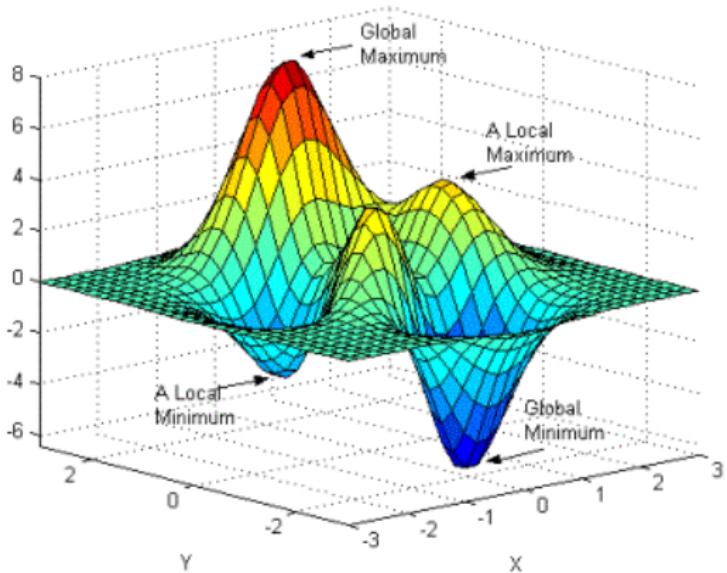
buildABLE

The pseudo-code:

- ① Simulate a random \mathcal{G}_i for a given demography Θ
- ② Calculate the probability of observing $1 \dots k_{max}$ mutations on every branch class of \mathcal{G}_i
- ③ Calculate the probability of observing every $\mathcal{B}_j \in \mathcal{B}_{SFS}$
- ④ Increment i and restart from (1) if $i < M$

inferABLE

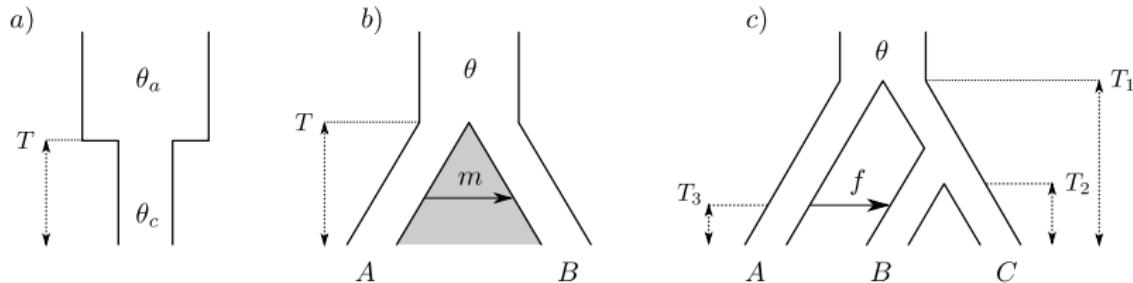
Maximum Likelihood Estimation (MLE)



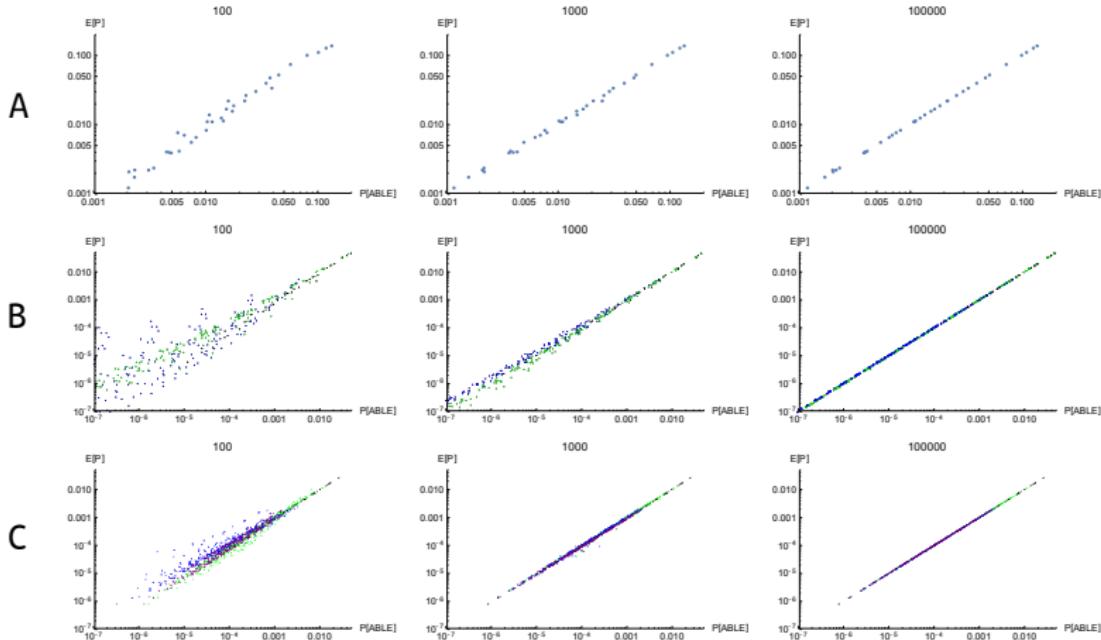
Wikimedia Commons

checkABLE

Expected bSFS : approximate vs. analytical



Asymptotic convergence of the bSFS 100, 1K & 100K genealogies



Every point represents a bSFS category

Beeravolu et al. 2018

buildABLE WITH intra-block recombination

The pseudo-code:

- ① Simulate a random \mathcal{G}_i for a given demography Θ
- ② Calculate the probability of observing $1 \dots k_{max}$ mutations on every branch class of \mathcal{G}_i
- ③ Calculate the probability of observing every $\mathcal{B}_j \in \mathcal{B}_{SFS}$
- ④ Increment i and restart from (1) if $i < M$

Sample from $p(\mathcal{A} | \Theta)$ instead of $p(\mathcal{G} | \Theta)$
(\mathcal{A} : Ancestral Recombination Graph)

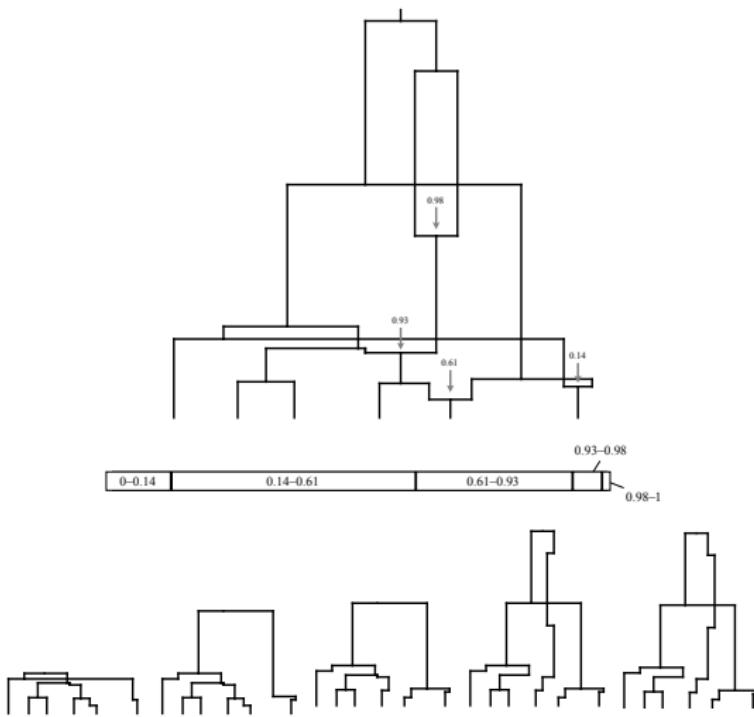
buildABLE WITH intra-block recombination

The pseudo-code:

- ① Simulate a random \mathcal{G}_i for a given demography Θ
- ② Calculate the probability of observing $1 \dots k_{max}$ mutations on every branch class of \mathcal{G}_i
- ③ Calculate the probability of observing every $\mathcal{B}_j \in \mathcal{B}_{SFS}$
- ④ Increment i and restart from (1) if $i < M$

Sample from $p(\mathcal{A} | \Theta)$ instead of $p(\mathcal{G} | \Theta)$
(\mathcal{A} : Ancestral Recombination Graph)

The Ancestral Recombination Graph (ARG) as a union of embedded marginal genealogies



Nordborg 2000

Orangutans : a tale of two islands

Pongo pygmaeus

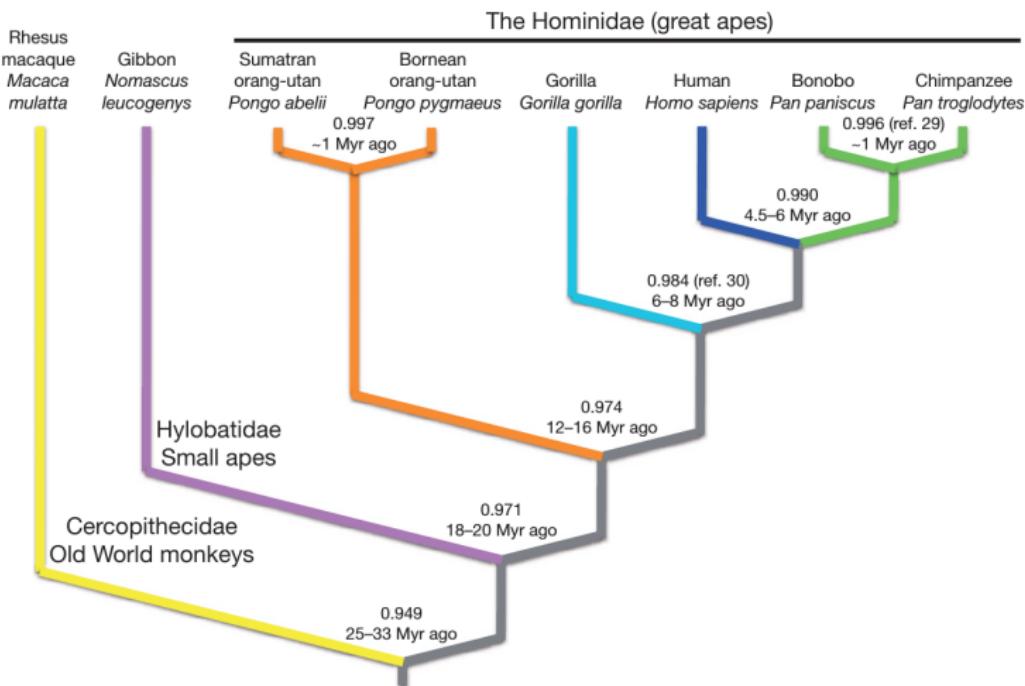


Pongo abelii



Wikimedia Commons

Orangutans : a tale of two islands



Locke et al. 2011

Orangutans : a tale of two islands

LETTER

doi:10.1038/nature09687

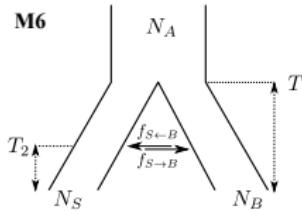
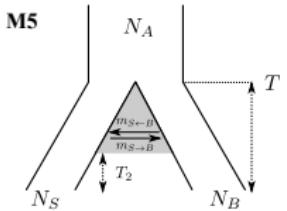
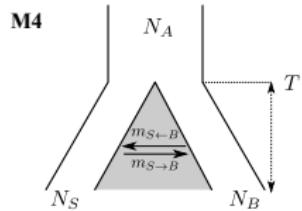
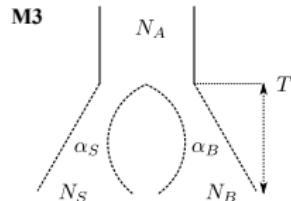
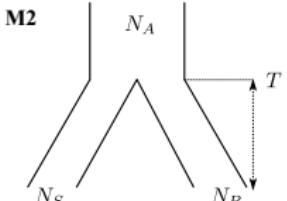
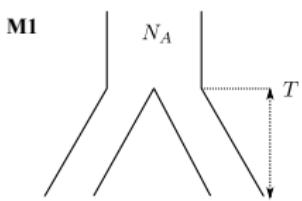
Comparative and demographic analysis of orang-utan genomes

Devin P. Locke¹, LaDeana W. Hillier¹, Wesley C. Warren¹, Kim C. Worley², Lynne V. Nazareth², Donna M. Muzny², Shiaw-Pyng Yang¹, Zhengyuan Wang¹, Asif T. Chinwalla¹, Pat Minx¹, Makedonka Mitreva¹, Lisa Cook¹, Kim D. Delehaunty¹, Catrina Fronick¹, Heather Schmidt¹, Lucinda A. Fulton¹, Robert S. Fulton¹, Joanne O. Nelson¹, Vincent Magrini¹, Craig Pohl¹, Tina A. Graves¹, Chris Markovic¹, Andy Cree², Huyen H. Dinh², Jennifer Hume², Christie L. Kovar², Gerald R. Fowler², Gerton Lunter^{3,4}, Stephen Meader³, Andreas Heger³, Chris P. Ponting³, Tomas Marques-Bonet^{5,6}, Can Alkan⁵, Lin Chen⁵, Ze Cheng⁵, Jeffrey M. Kidd⁵, Evan E. Eichler^{5,7}, Simon White⁸, Stephen Searle⁸, Albert J. Vilella⁹, Yuan Chen⁹, Paul Flicek⁹, Jian Ma¹⁰, Brian Raney¹⁰, Bernard Suh¹⁰, Richard Burhans¹¹, Javier Herrero⁹, David Haussler¹⁰, Rui Faria^{6,12}, Olga Fernando^{6,13}, Fleur Darré⁶, Domènec Farre⁶, Elodie Gazave⁶, Meritxell Oliva⁶, Arcadi Navarro^{6,14}, Roberta Roberto¹⁵, Oronzo Capozzi¹⁵, Nicoletta Archidiacono¹⁵, Giuliano Della Valle¹⁶, Stefania Purgato¹⁶, Mariano Rocchi¹⁵, Miriam K. Konkel¹⁷, Jerilyn A. Walker¹⁷, Brygg Ullmer¹⁸, Mark A. Batzer¹⁷, Arian F. A. Smil¹⁹, Robert Hubley¹⁹, Claudio Casola²⁰, Daniel R. Schrider²⁰, Matthew W. Hahn²⁰, Victor Quesada²¹, Xose S. Puente²¹, Gonzalo R. Ordóñez²¹, Carlos López-Otin²¹, Tomas Vinar²², Brona Brejova²², Aakrosh Ratan¹¹, Robert S. Harris¹¹, Webb Miller¹¹, Carolin Kosiol²³, Heather A. Lawson²⁴, Vikas Taliwal²⁵, André L. Martins²⁵, Adam Siepel²⁵, Arindam RoyChoudhury²⁶, Xin Ma²⁵, Jeremiah Degenhardt²⁵, Carlos D. Bustamante²⁷, Ryan N. Gutenkunst²⁸, Thomas Mailund²⁹, Julien Y. Dutheil²⁹, Asger Hobolth²⁹, Mikkel H. Schierup²⁹, Oliver A. Ryder³⁰, Yuko Yoshinaga³¹, Pieter J. de Jong³¹, George M. Weinstock¹, Jeffrey Rogers², Elaine R. Mardis¹, Richard A. Gibbs² & Richard K. Wilson¹

Locke et al. 2011

Inferring demography AND recombination rates

The model choice pipeline



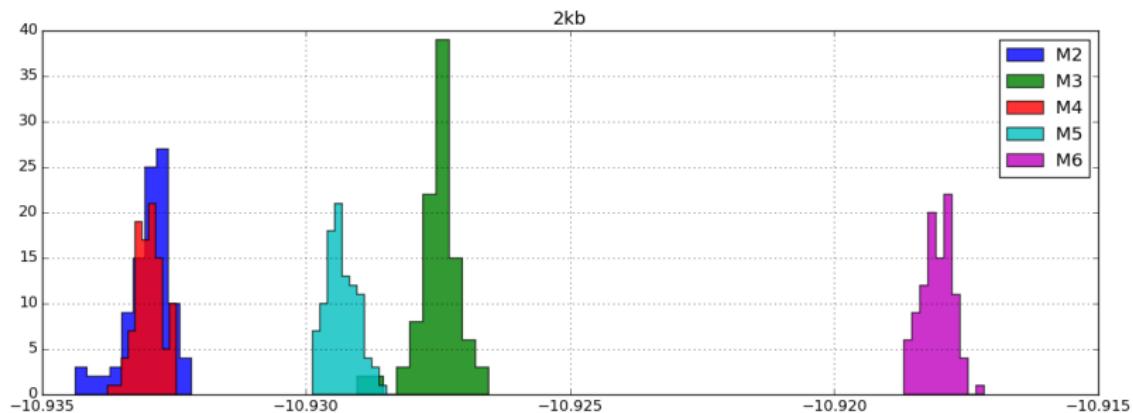
Results from 2kb blocks

Total spliced length : 163 Mbp

Model	N_A	c	T	N_S	N_B	α_S	α_B	$4N_A m_{S \rightarrow B}$	$4N_A m_{S \leftarrow B}$	T_2	$f_{S \rightarrow B}$	$f_{S \leftarrow B}$	$\ln L$
M1	18 200	1.58×10^{-8}	387 000										-907 477
M2	1 380	2.06×10^{-8}	294 000	22 100	8 610								-891 341
M3	2 180	2.09×10^{-8}	306 000	21 800	5 490	-0.003	-0.728						-891 308
M4	1 260	2.11×10^{-8}	320 000	22 300	8 210			0.025	0.000				-892 423
M5	1 280	1.87×10^{-8}	1 807 000	21 600	8 850			1.568	2.202	274 000			-892 225
M6	1 420	2.73×10^{-8}	816 000	22 400	8 910					295 000	0.121	0.267	-891 139

$\mu : 1 \times 10^{-8} / bp/generation$
20 yrs/generation
2 diploid genomes per pop.

Relative model fit for 2kb blocks distribution of 100 LnLs using 1M ARGs



Direction of best fit

▶ Jump to model pipeline

Beeravolu *et al.* 2018

ABLE : a quick summary

- Uses the **bSFS**, a very rich summary of genomic data
- **Does not require polarized data** (*i.e.* no outgroups)
- **Does not require phased data** and accounts for linkage
- Can **infer recombination rates** along with demography
- Is **computationally efficient** (coded in C/C++)
- Uses ***ms*** for sampling from $p(\mathcal{G}_i | \theta)$ or $p(\mathcal{A} | \Theta)$
- Runs on **parallel threads** using OpenMP

More info from <https://github.com/champost/ABLE>

publishABLE!

Beeravolu *et al.* *Genome Biology* (2018) 19:145
<https://doi.org/10.1186/s13059-018-1517-y>

Genome Biology

METHOD

Open Access

ABLE: blockwise site frequency spectra for inferring complex population histories and recombination

Champak R. Beeravolu^{1,2*} , Michael J. Hickerson^{1,3,4}, Laurent A. F. Frantz^{5,7} and Konrad Lohse⁶

Beeravolu *et al.* 2018

Collaborators

Konrad Lohse

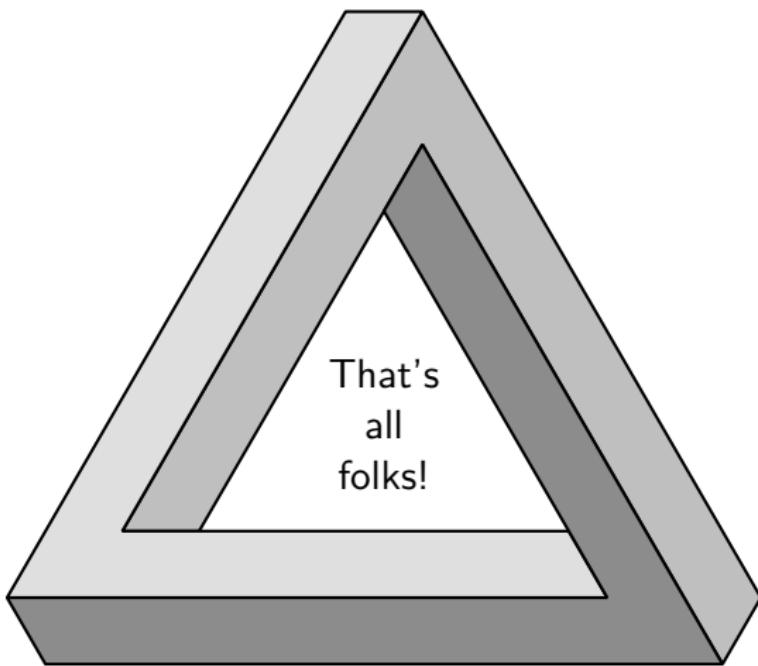
University of Edinburgh

Laurent A.F. Frantz

Queen Mary University of London

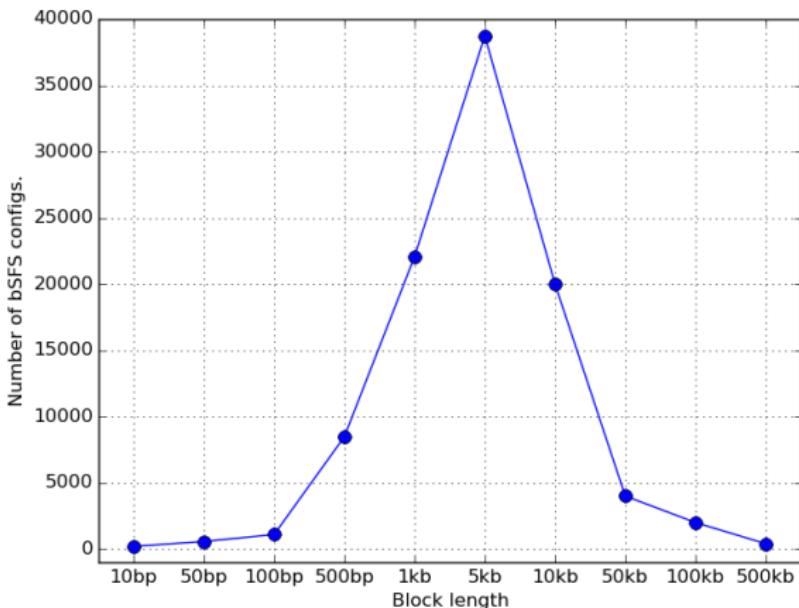
Michael J. Hickerson

The City College of New York



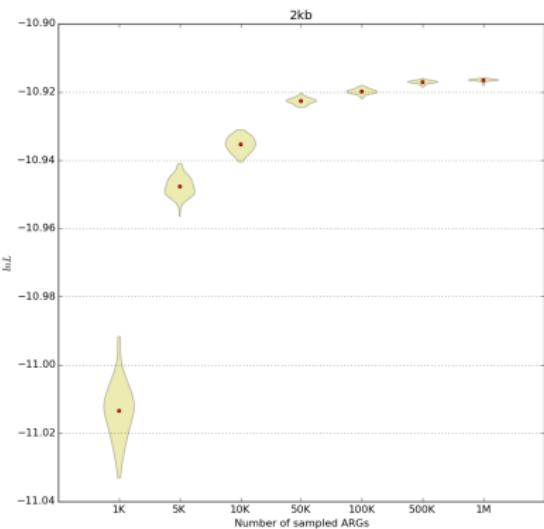
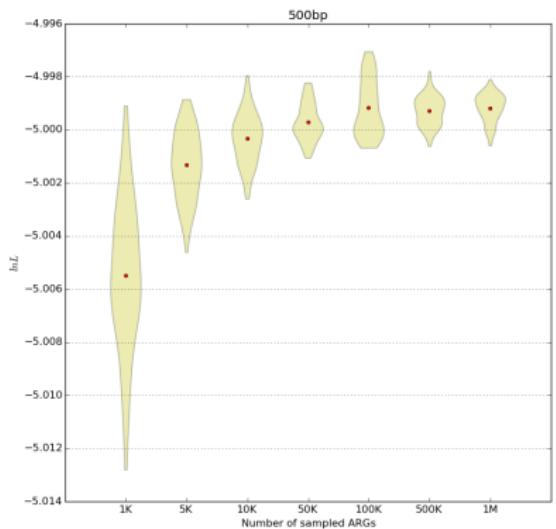
"Optimum" block size

Total genome size : 200Mb

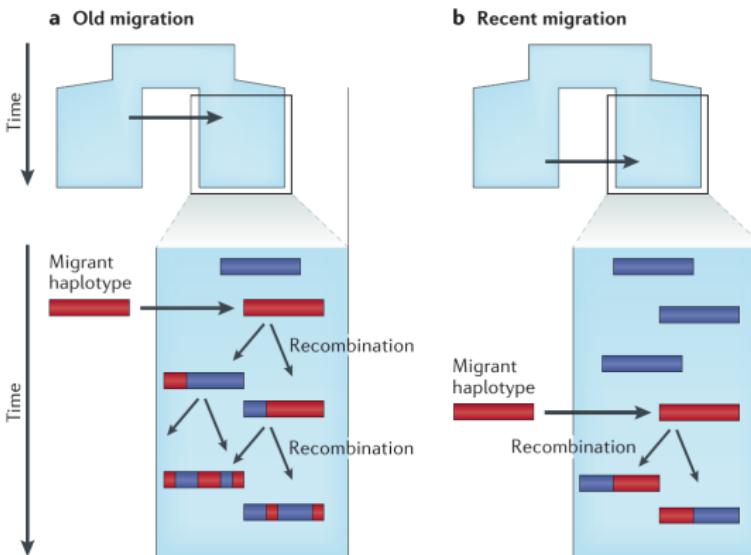


Likelihood convergence for complex demographies

An example of model M6



Migration AND Recombination



Sousa and Hey, 2011