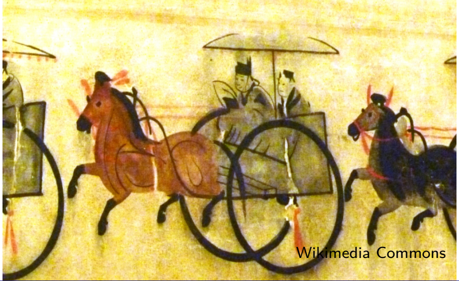
Approximate Blockwise Likelihood Estimation

Champak Beeravolu Reddy

University of Fribourg

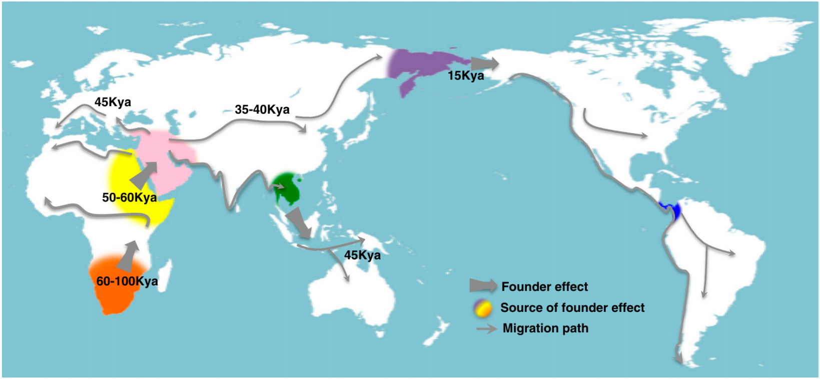
August 29, 2019

We move, and have been moving!



Wikimedia Commons

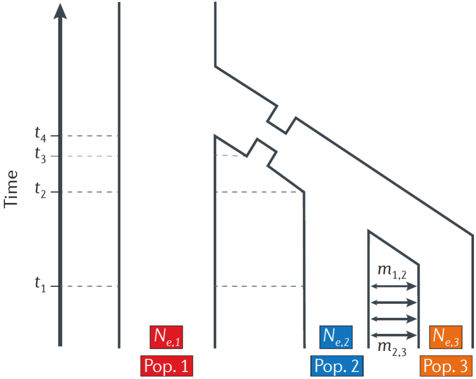
Picturing modern human migration



Henn *et al.* 2012

Simplifying a complex demography

modelling interacting panmictic units



Schraiber and Akey 2015

Demographic inference using genomic data

The two major approaches

- Based on the SFS
  → No modeling 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 modeling 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

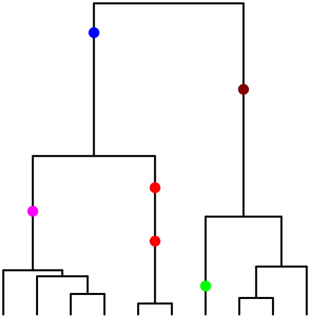
The two major approaches

- Based on the SFS
  → No modeling 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

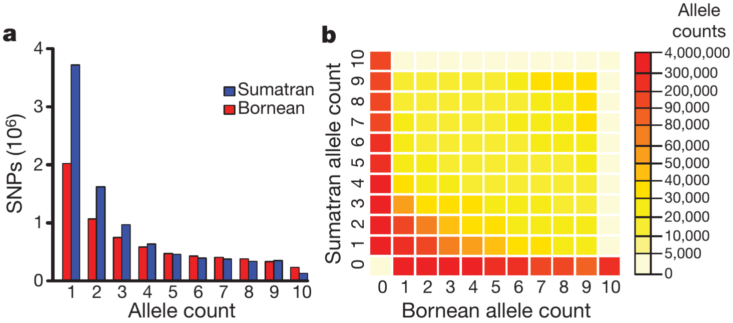
Gene genealogies and polymorphisms



←GTATCAAT

0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0
1	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	1	1	1	1
1	1	1	1	1	1	0	0	0	0
0	0	0	0	1	1	0	0	0	0

The Site Frequency Spectrum (SFS)

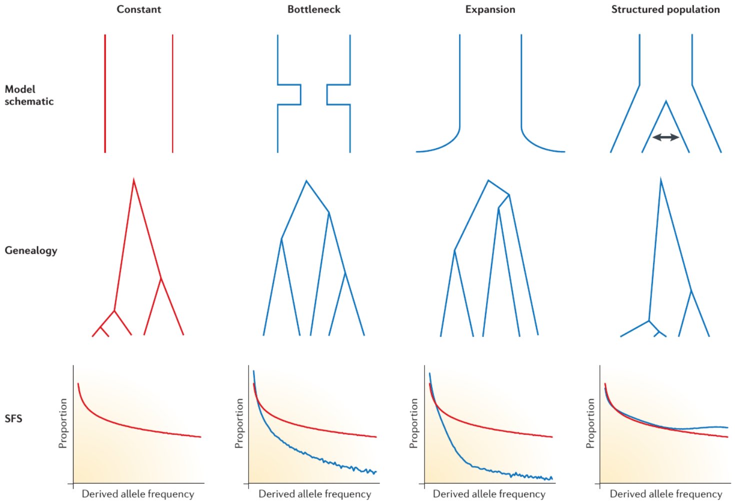


a. SFS

b. Joint SFS

Locke *et al.* 2011

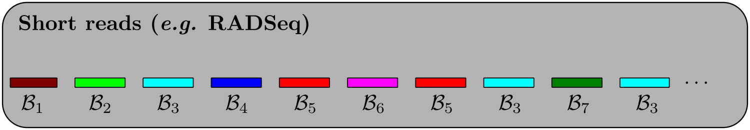
Statistical identifiability and the SFS



Schraiber and Akey 2015

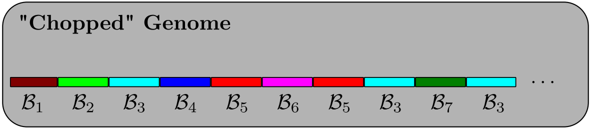
Extending the SFS

The Blockwise SFS (bSFS)



Short reads (*e.g.* RADSeq)

$\mathcal{B}_1$ $\mathcal{B}_2$ $\mathcal{B}_3$ $\mathcal{B}_4$ $\mathcal{B}_5$ $\mathcal{B}_6$ $\mathcal{B}_5$ $\mathcal{B}_3$ $\mathcal{B}_7$ $\mathcal{B}_3$ . . .



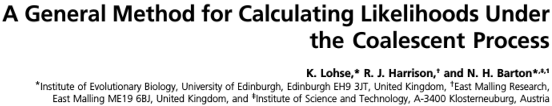
"Chopped" Genome

$\mathcal{B}_1$ $\mathcal{B}_2$ $\mathcal{B}_3$ $\mathcal{B}_4$ $\mathcal{B}_5$ $\mathcal{B}_6$ $\mathcal{B}_5$ $\mathcal{B}_3$ $\mathcal{B}_7$ $\mathcal{B}_3$ . . .

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

An exact analytical method

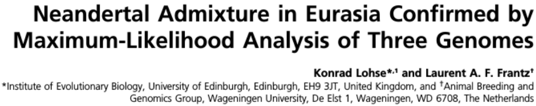
makes use of the Generating Function of branch lengths



A General Method for Calculating Likelihoods Under the Coalescent Process

K. Lohse,* R. J. Harrison,† and N. H. Barton*,‡,1

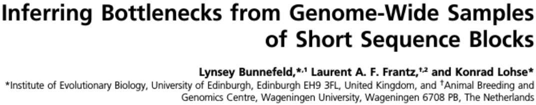
*Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom, †East Malling Research, East Malling ME19 6BJ, United Kingdom, and ‡Institute of Science and Technology, A-3400 Klosterneuburg, Austria



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

Konrad Lohse*,1 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,*,1 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

Approximating the bSFS

Approximate Blockwise Likelihood Estimation (ABLE)

Felsenstein equation (discretized Chapman-Kolmogorov)

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

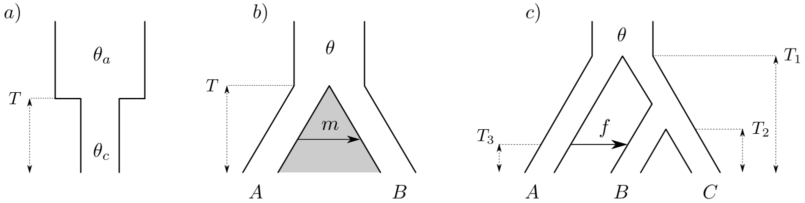
Sampling genealogies $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_M$ from $p(\mathcal{G} \mid \Theta)$ yields a Monte Carlo estimator of the bSFS likelihood

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

Felsenstein 1988

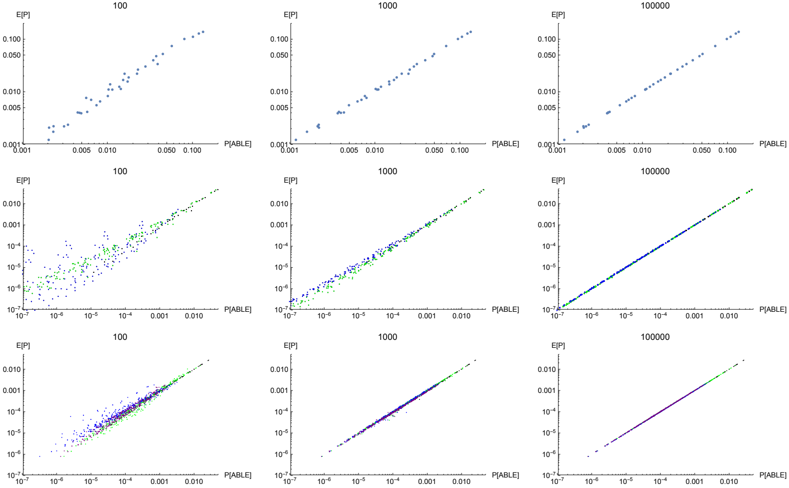
checkABLE

Expected bSFS : approximate vs. analytical



Asymptotic convergence of the bSFS

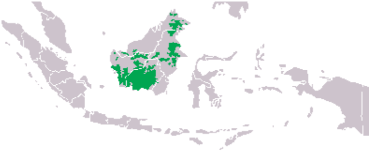
100, 1K & 100K genealogies



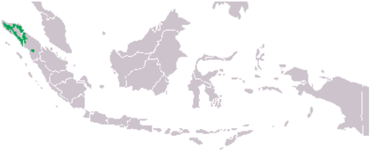
Every point represents a bSFS category

Orangutans : a tale of two islands

Pongo pygmaeus

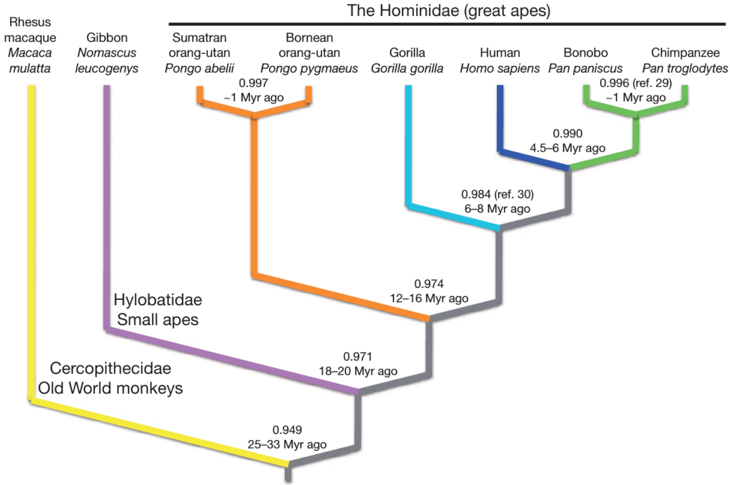


Pongo abelii



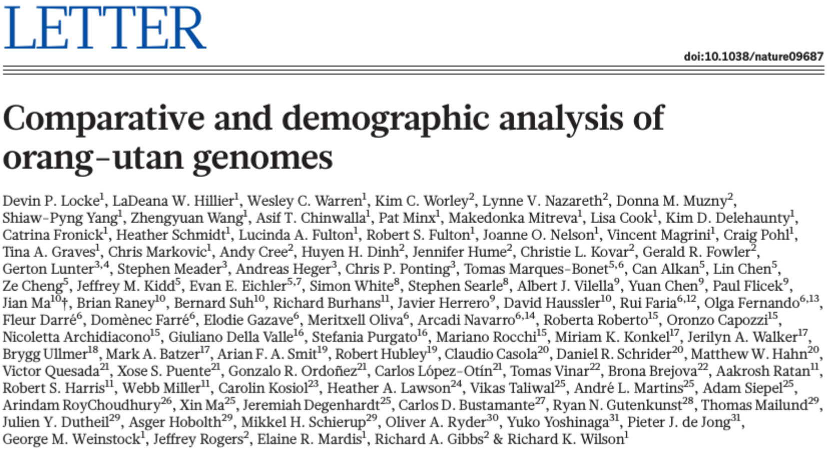
Wikipedia

Orangutans : a tale of two islands


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

Locke *et al.* 2011

Orangutans : a tale of two islands



LETTER

doi:10.1038/nature09687

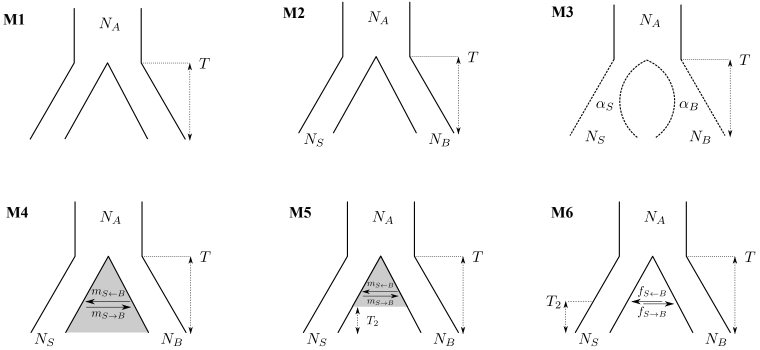
Comparative and demographic analysis of orang-utan genomes

Devin P. Locke[1], LaDeana W. Hillier[1], Wesley C. Warren[1], Kim C. Worley[2], Lynne V. Nazareth[2], Donna M. Muzny[2], Shiaw-Pyng Yang[1], Zhengyuan Wang[1], Asif T. Chinwalla[1], Pat Minx[1], Makedonka Mitreva[1], Lisa Cook[1], Kim D. Delehaunty[1], Catrina Fronick[1], Heather Schmidt[1], Lucinda A. Fulton[1], Robert S. Fulton[1], Joanne O. Nelson[1], Vincent Magrini[1], Craig Pohl[1], Tina A. Graves[1], Chris Markovic[1], Andy Cree[2], Huyen H. Dinh[2], Jennifer Hume[2], Christie L. Kovar[2], Gerald R. Fowler[2], Gerton Lunter[3,4], Stephen Meader[3], Andreas Heger[3], Chris P. Ponting[3], Tomas Marques-Bonet[5,6], Can Alkan[5], Lin Chen[5], Ze Cheng[5], Jeffrey M. Kidd[5], Evan E. Eichler[5,7], Simon White[8], Stephen Searle[8], Albert J. Vilella[9], Yuan Chen[9], Paul Flicek[9], Jian Ma[10]†, Brian Raney[10], Bernard Suh[10], Richard Burhans[11], Javier Herrero[9], David Haussler[10], Rui Faria[6,12], Olga Fernando[6,13], Fleur Darré[6], Domènec Farré[6], Elodie Gazave[6], Meritxell Oliva[6], Arcadi Navarro[6,14], Roberta Roberto[15], Oronzo Capozzi[15], Nicoletta Archidiacono[15], Giuliano Della Valle[16], Stefania Purgato[16], Mariano Rocchi[15], Miriam K. Konkel[17], Jerilyn A. Walker[17], Brygg Ullmer[18], Mark A. Batzer[17], Arian F. A. Smit[19], Robert Hubley[19], Claudio Casola[20], Daniel R. Schrider[20], Matthew W. Hahn[20], Victor Quesada[21], Xose S. Puente[21], Gonzalo R. Ordoñez[21], Carlos López-Otín[21], Tomas Vinar[22], Brona Brejova[22], Aakrosh Ratan[11], Robert S. Harris[11], Webb Miller[11], Carolin Kosiol[23], Heather A. Lawson[24], Vikas Taliwal[25], André L. Martins[25], Adam Siepel[25], Arindam RoyChoudhury[26], Xin Ma[25], Jeremiah Degenhardt[25], Carlos D. Bustamante[27], Ryan N. Gutenkunst[28], Thomas Mailund[29], Julien Y. Dutheil[29], Asger Hobolth[29], Mikkel H. Schierup[29], Oliver A. Ryder[30], Yuko Yoshinaga[31], Pieter J. de Jong[31], George M. Weinstock[1], Jeffrey Rogers[2], Elaine R. Mardis[1], Richard A. Gibbs[2] & Richard K. Wilson[1]

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	$f_{S\rightarrow B}$	$f_{S\leftarrow B}$	lnL
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

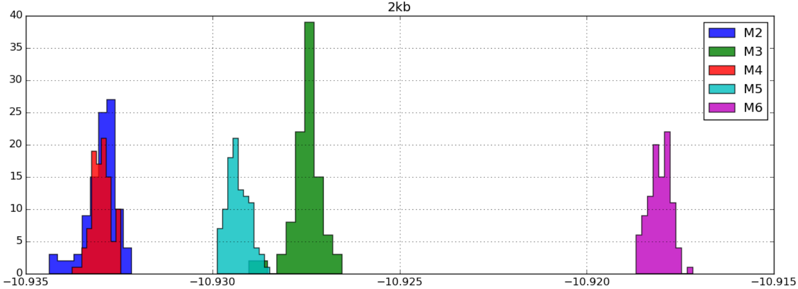
$\mu : 1 \times 10^{-8} / bp / generation$

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 →

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 $p(\mathcal{G}_i \mid \theta)/p(\mathcal{A} \mid \Theta)$
- Runs on **parallel threads** using OpenMP

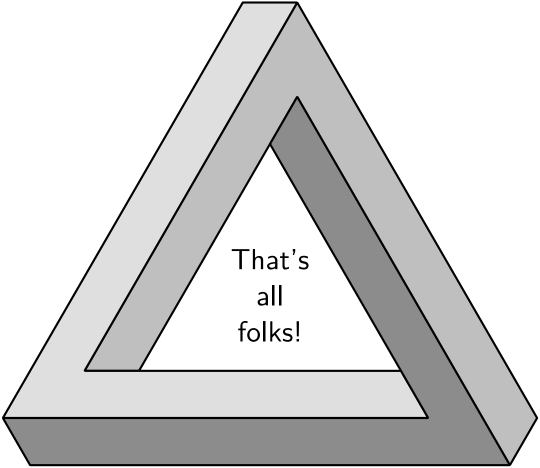
Download **v0.1** from **https://github.com/champost/ABLE**

Collaborators

Konrad Lohse
University of Edinburgh

Laurent A.F. Frantz
Queen Mary University of London

Michael J. Hickerson
The City College of New York


That's
all
folks!