Robust demographic inference from genomic and SNP data

Laurent Excoffier

Isabelle Duperret, Emilia Huerta-Sanchez, Matthieu Foll, Vitor Sousa, Isabel Alves

Computational and Molecular Population Genetics Lab (CMPG)
Institute of Ecology and Evolution
University of Berne
Swiss Institute of Bioinformatics

u b
b UNIVERSITÄT BERN

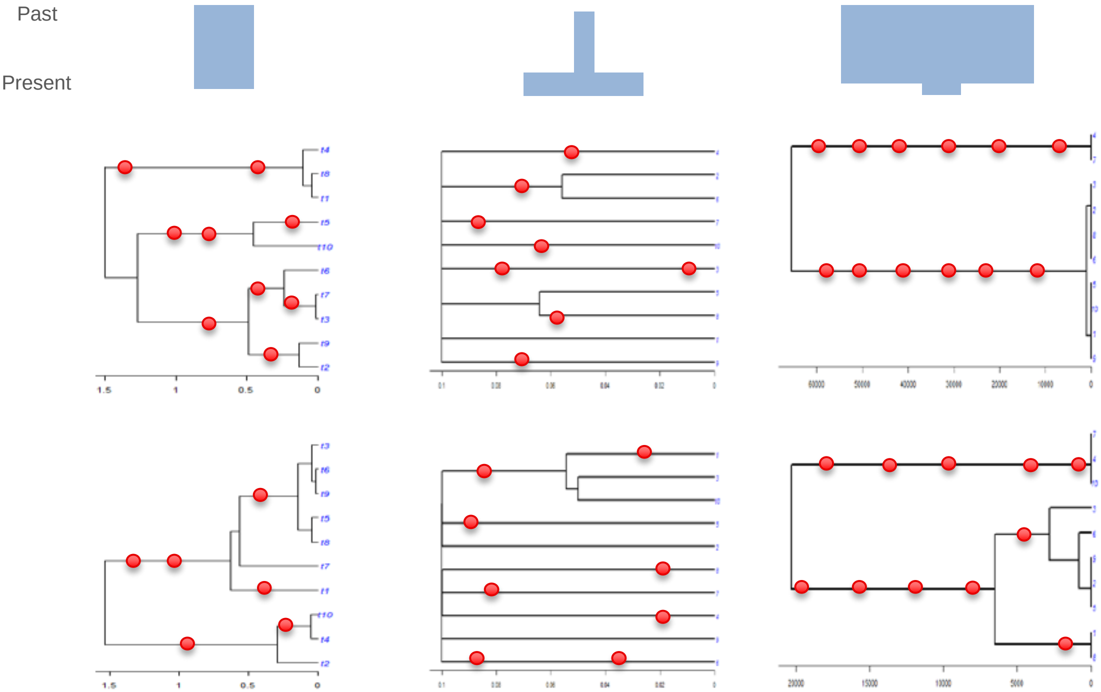
SIB Swiss Institute of Bioinformatics

Past demography affect genetic diversity

Stationary population

Recent expansion

Recent contraction

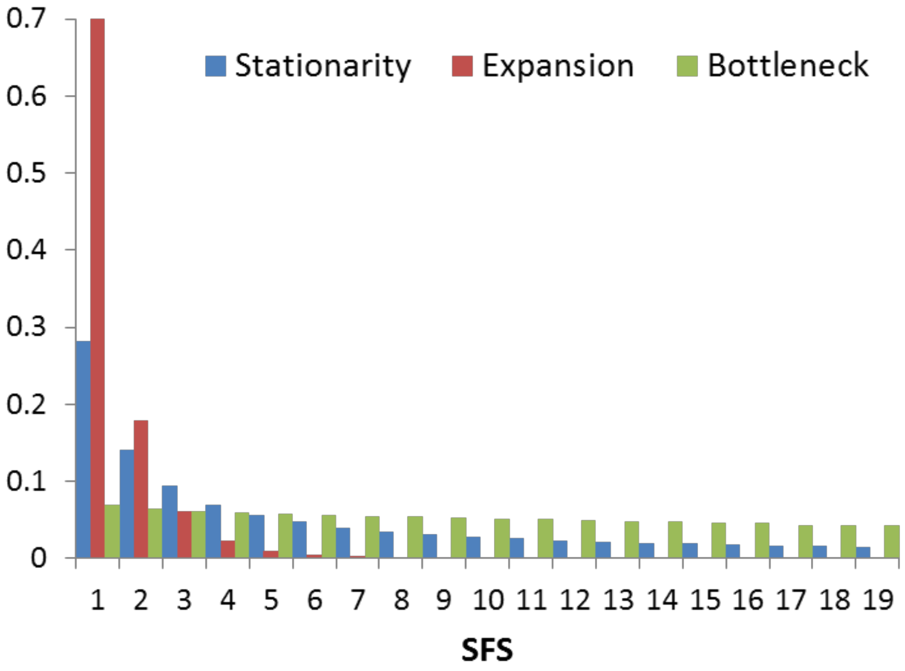


Mixture of rare and frequent mutations

Few and mostly rare mutations

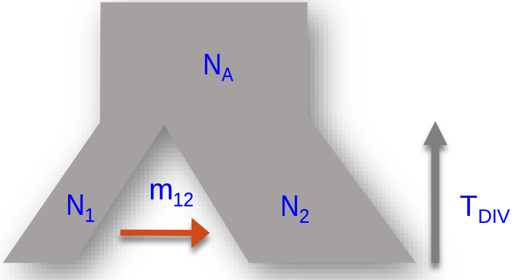
Very deep lineages separating little differentiated clades

Site Frequency Spectrum (SFS) depends on past demography

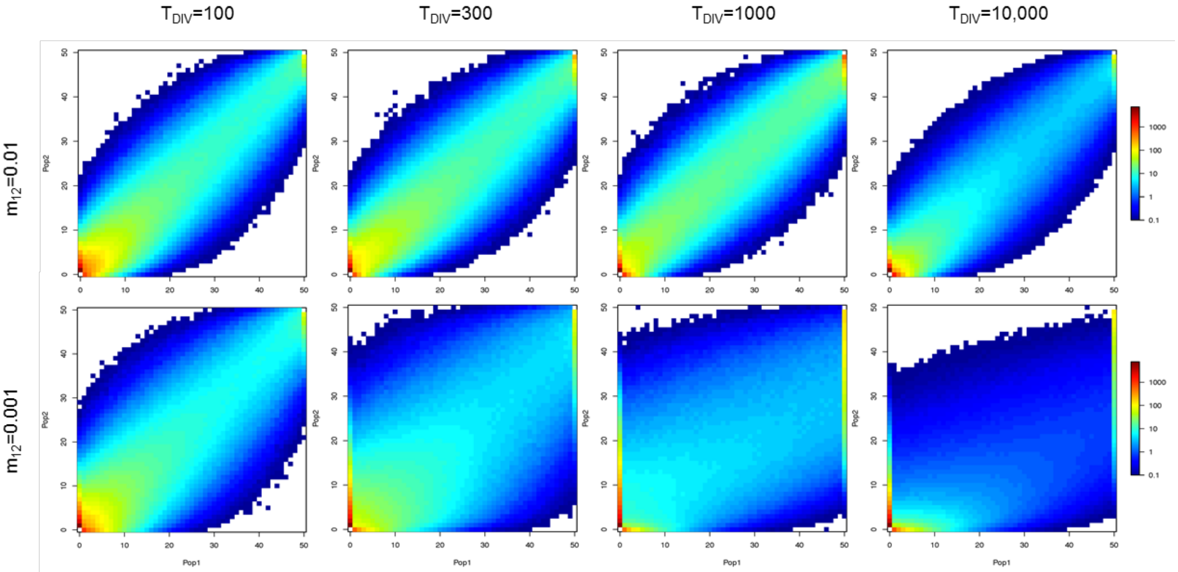


Joint SFS (2D-SFS)

UNIVERSITÄT BERN

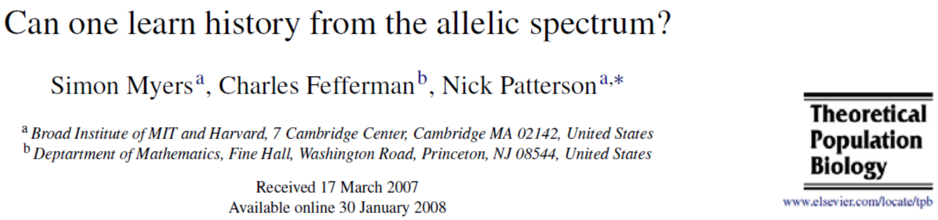


Model of Isolation with migration (IM)



SIB Swiss Institute of Bioinformatics

Problems with estimation of demographic parameters from SFS



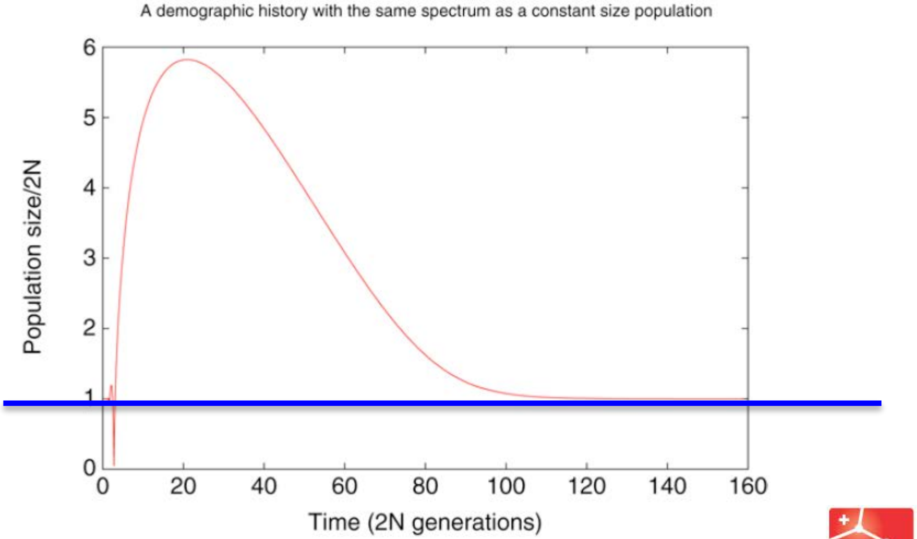
Can one learn history from the allelic spectrum?

Simon Myers[a], Charles Fefferman[b], Nick Patterson[a,*]

[a] *Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge MA 02142, United States*
[b] *Deptartment of Mathematics, Fine Hall, Washington Road, Princeton, NJ 08544, United States*

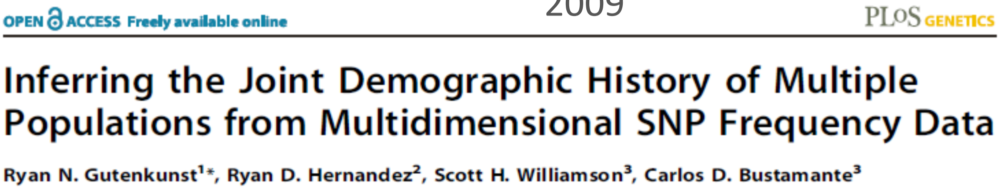
Received 17 March 2007
Available online 30 January 2008

Theoretical Population Biology
www.elsevier.com/locate/tpb



Estimation of demographic parameters from SFS with dadi

2009



OPEN ACCESS Freely available online

PLoS GENETICS

Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data

Ryan N. Gutenkunst[1]*, Ryan D. Hernandez[2], Scott H. Williamson[3], Carlos D. Bustamante[3]

Program $\partial a \partial i$: Diffusion Approximation for Demographic Inference http://code.google.com/p/dadi/

dadi estimates the site frequency spectrum based on a diffusion approximation

Advantages of SFS for parameter inference

- Accuracy of estimates increases with data size, but computing time does not
- Can be used to study complex scenarios (e.g. as complex as ABC)
- Very fast estimations (as compared to ABC, or full likelihoods)

Potential problems

- Maximization of the CL is not trivial
  (precision of the approximation and convergence problems)
- Ignores (assumes no) LD
- Need to repeat estimations to find maximum CL
- Needs genomic data (several Mb)
  - difficult to have gene-specific estimates
- Next-generation sequencing data must have high coverage to correctly estimate SFS (likely to miss singletons or show errors).
- SFS needs to be estimated from the NGS reads
  (ML methods: Nielsen et al. 2013, Keightley and Halligan, 2011)

Estimating the SFS with coalescent simulations

The probability of a SFS entry *i* can be estimated under a specific model *θ* from its expected coalescent tree as (Nielsen 2000) a **ratio of expected branch lengths**

$$p_i = E(t_i \mid \theta) \,/\, E(T \mid \theta)$$

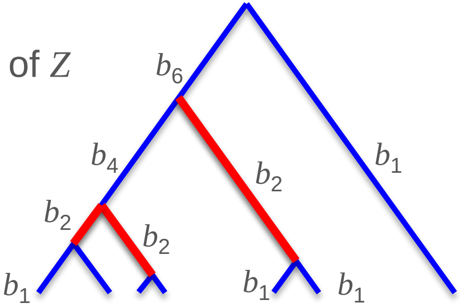
t_i : total length of all branches directly leading to i terminal nodes

T : total tree length.

This probability can then be estimated on the basis of Z simulations as

$$\hat{p}_i = \sum_{j}^{Z} \sum_{k \in \Phi_i} b_{kj} \Big/ \sum_{j}^{Z} T_j$$

where b_{kj} is the length of the k-th compatible branch in simulation j.



Likelihood

The (composite) likelihood of a model θ is obtained as a multinomial sampling of sites (Adams and Hudson, 2004)

$$CL = \Pr(SFS_{obs} \mid \theta) \propto P_0^M (1 - P_0)^S \prod_{i=1}^{n-1} \hat{p}_i^{\,m_i}$$

M : number of monomorphic sites

S : number of polymorphic sites

P_0 : probability of no mutation on the tree

p_i : probability of the *i*-th SFS entry

m_i: number of sites with derived frequency *i*

This can be generalized for the joint SFS of two or more populations

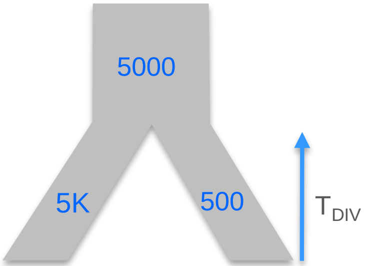
fastsimcoal2 program

- Uses coalescent simulations to estimate the SFS and approximate the likelihood
  - Large number of simulations per point (>50000)
- Uses a **conditional expectation maximization** (CEM) algorithm to find maxCL parameters
- Relatively fast and can explore wide and unbounded parameter ranges
- Can handle an arbitrary number of populations
- For more than 4 populations, we use a composite composite-likelihood

$$CL_{1234\ldots} = CL_{12} \times CL_{13} \times CL_{14} \times \ldots \times CL_{23} \times \ldots$$

Approximation of the SFS

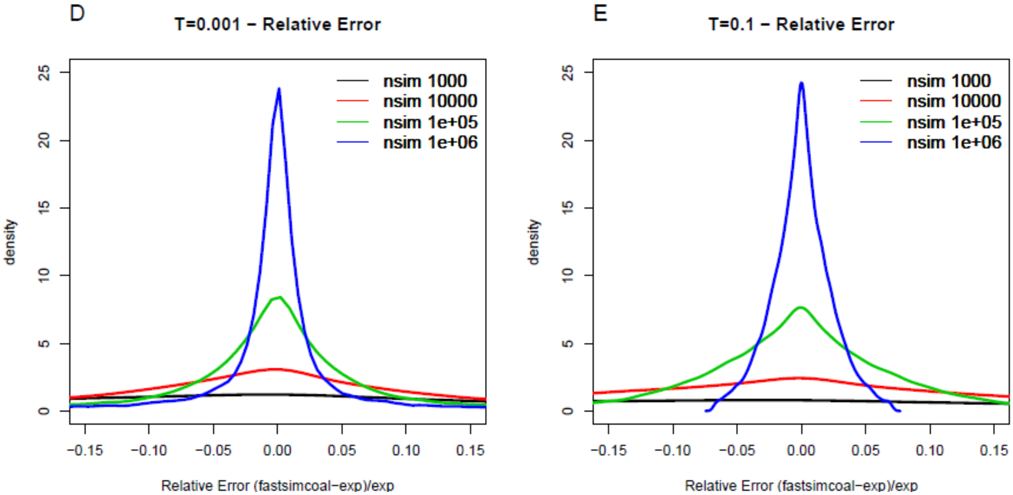
Divergence model



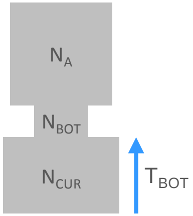
Chen (2012) TPB
Coalescent approach to infer the expected joint SFS numerically

T_{DIV}=10

T_{DIV}=100



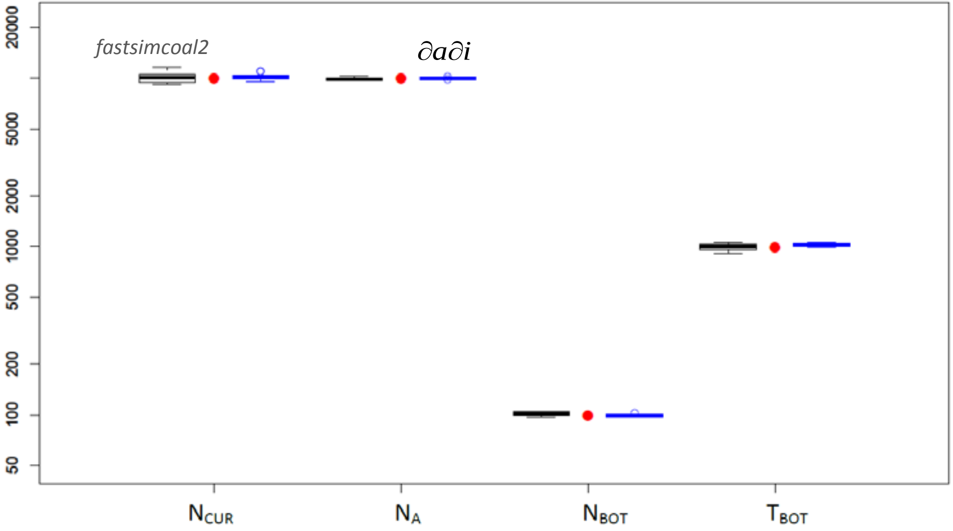
Bottleneck model



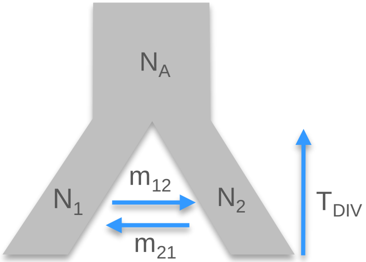
Simulation of 20 Mb data
10 cases, 50 runs/case

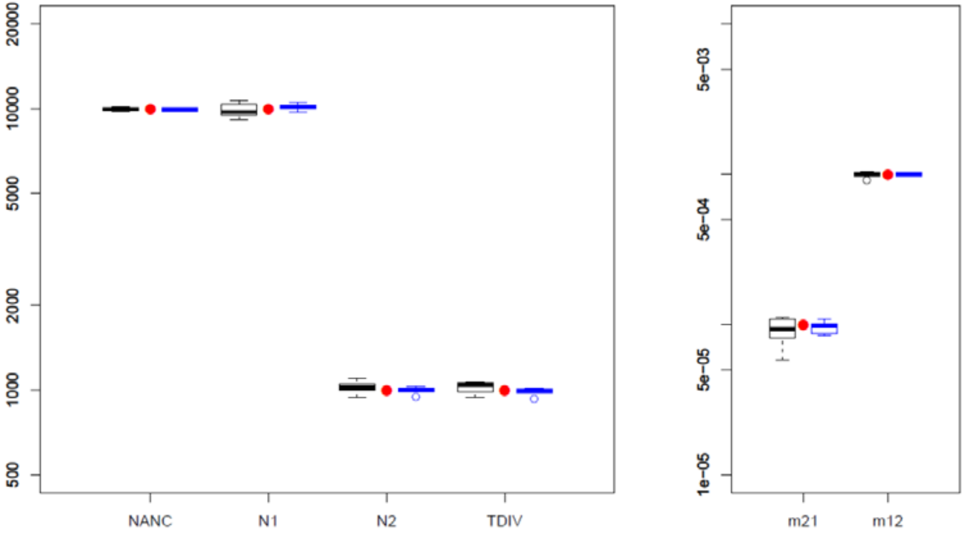
∂a∂i

9/10



IM model





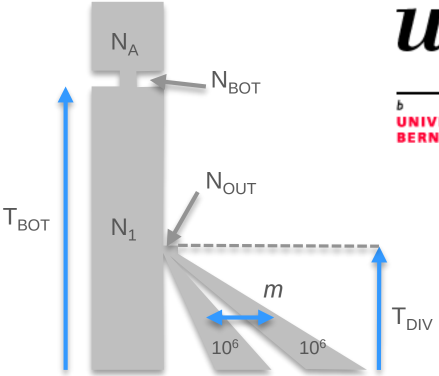
UNIVERSITÄT BERN

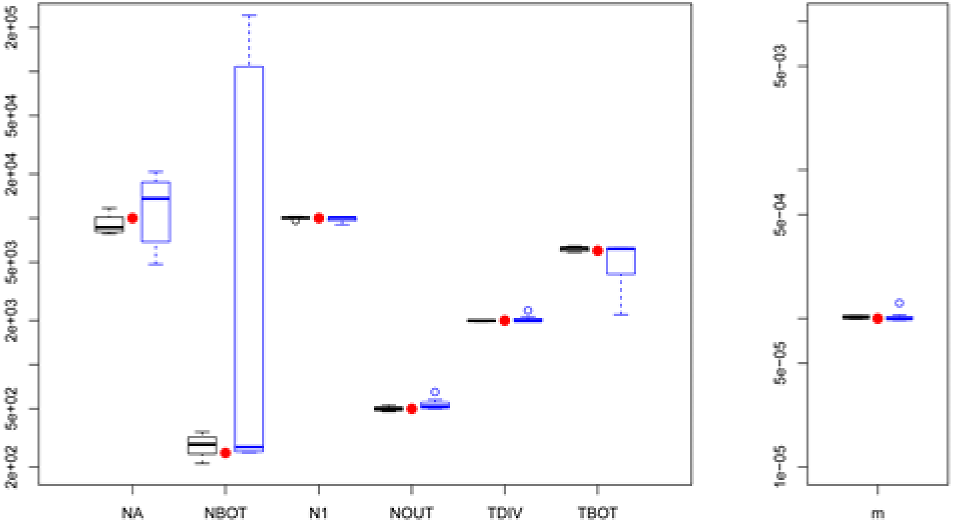
∂a∂i

8/10

SIB Swiss Institute of Bioinformatics

Pseudo human evolution model

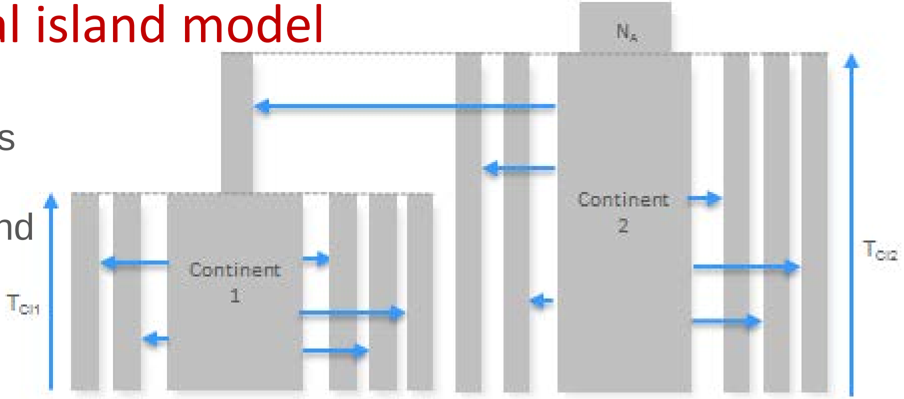


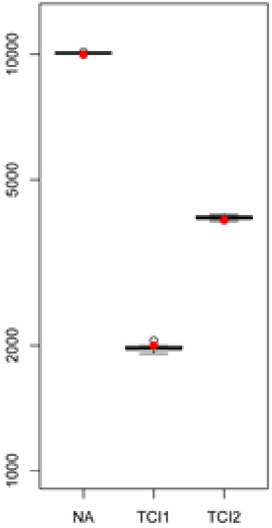


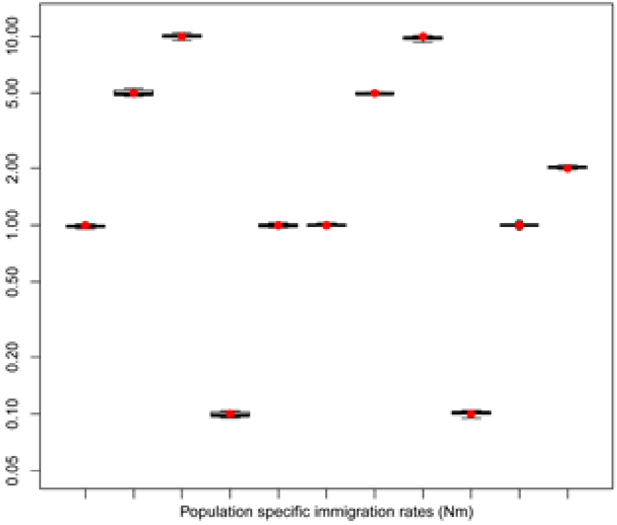
∂a∂i

Herarchical island model

12 populations in two continent-island models







Migration rates over 3 orders of magnitude are well recovered !!!

Application: Complete genomics data

Four sampled human populations:

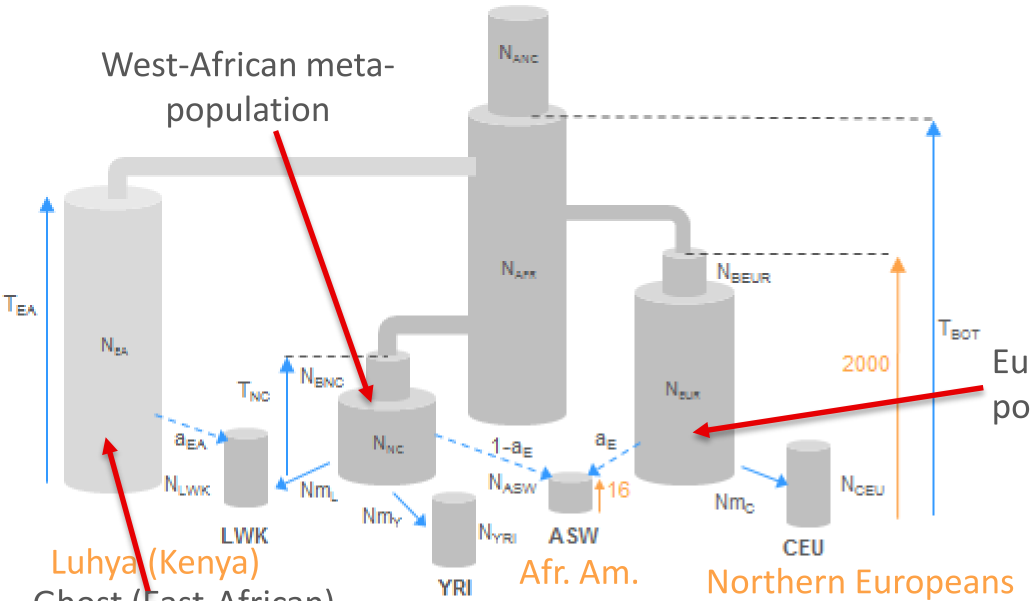
4 **Luhya** from Kenya (**LWK**)
9 **Europeans** (**CEU**)
9 **Yoruba** (**YRI**)
5 **African Americans** (**ASW**)

(sequenced at 51-89x per genome)

Data:

Multidimensional SFS estimated from :
239, 120 SNPs in non-coding and non CpG regions
Each SNP more than 5 Kb away from the other

Model of admixture in African Americans



West-African meta-population

European meta-population

Luhya (Kenya)

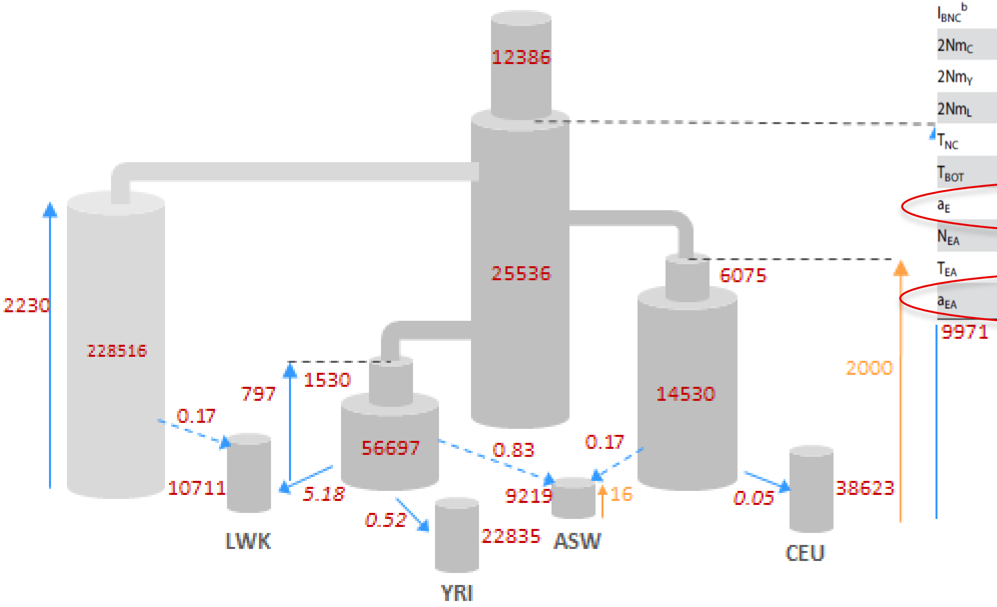
Ghost (East-African) meta-population

Yoruba (Nigeria)

Afr. Am.

Northern Europeans

Model of admixture in African Americans

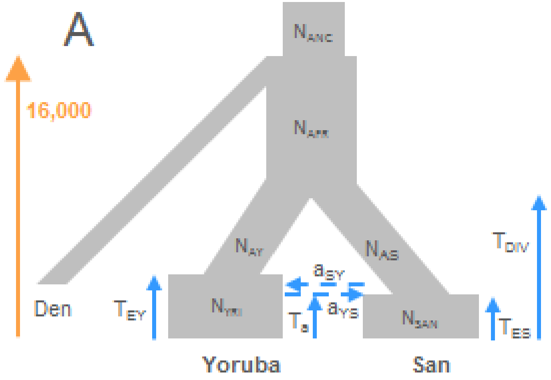


Parameters	Point estimation	95% CI[a] Lower bound	Upper bound
N_{ANC}	12386	10986	14875
N_{AFR}	25536	22054	35939
N_{ASW}	9219	9906	44026
N_{CEU}	38623	8842	43883
N_{LWK}	10711	13288	41103
N_{YRI}	22835	14809	44010
N_{EUR}	14530	11792	25615
I_{BEUR}[b]	0.418	0.375	0.450
N_{NC}	56697	33872	414434
I_{BNC}[b]	0.027	0.011	0.040
$2Nm_C$	0.05	0.04	26.57
$2Nm_Y$	0.52	0.04	22.83
$2Nm_L$	5.18	0.03	35.68
T_{NC}	797	509	1981
T_{BOT}	9971	8900	12834
a_E	0.17	0.16	0.18
N_{EA}	228516	95844	451516
T_{EA}	2230	1479	3386
a_{EA}	0.17	0.08	0.19

Models of African population divergence

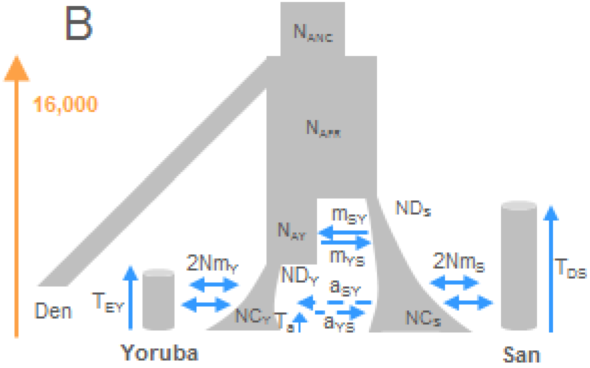
Two models with different degrees of realism and complexity

IM model



3 populations

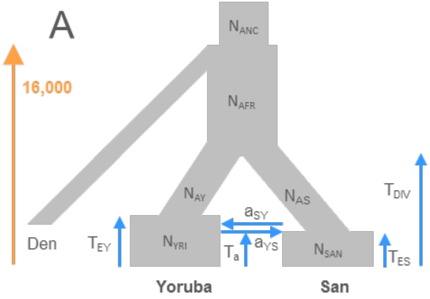
2 continent-island model



5 populations

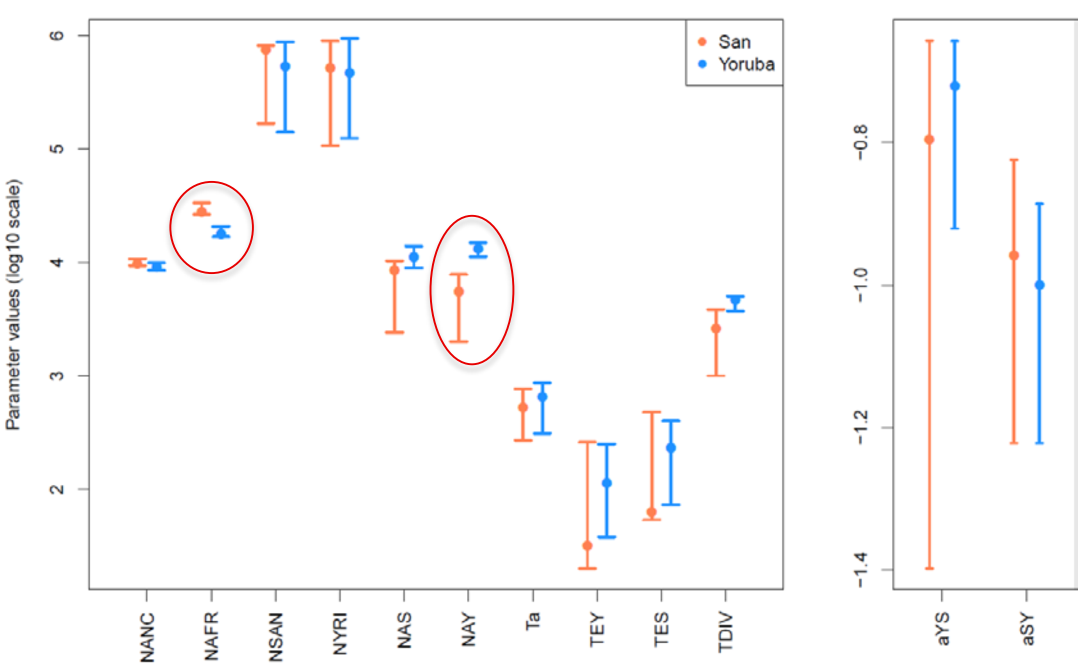
The estimation of each model were performed separately for the San (109,020 SNPs) and the Yoruba (81,383 SNPs) SNP panels

Models of African population divergence



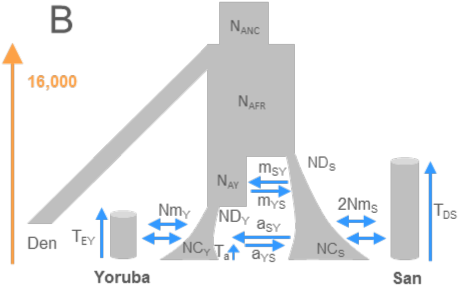
IM model

Model A – San panel vs. Yoruba panel



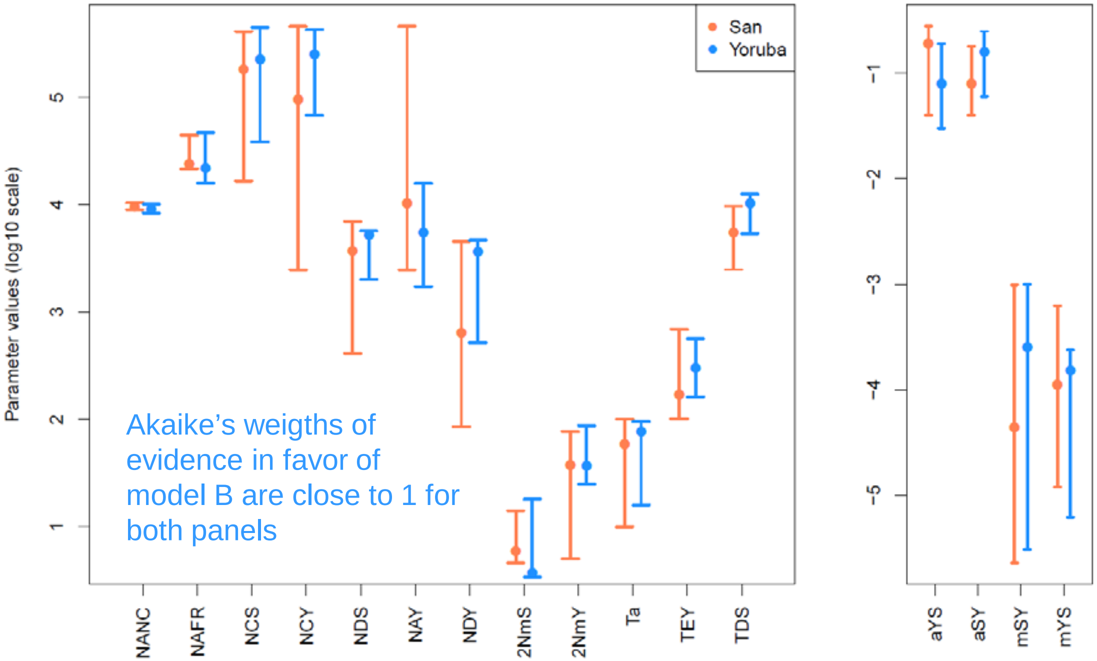
Good agreement between panels

Models of African population divergence



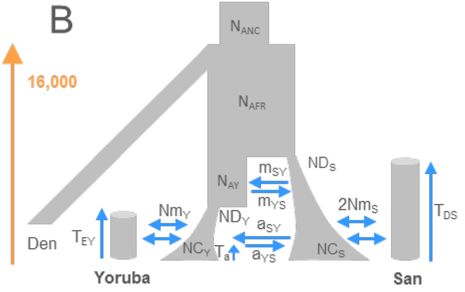
2 continent-island model

Model B – San panel vs. Yoruba panel



Akaike's weigths of evidence in favor of model B are close to 1 for both panels

Models of African population divergence

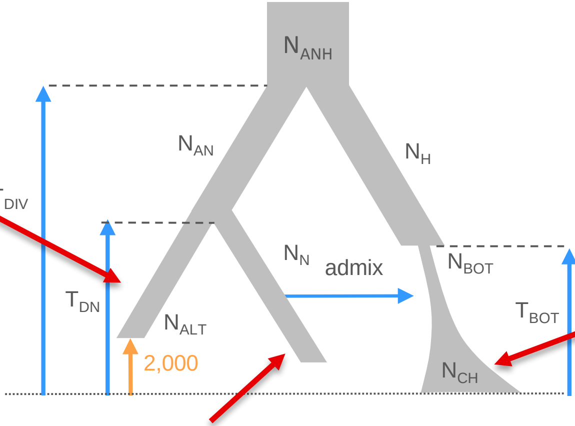


Model B

	Panel 4 (San)		Panel 5 (Yoruba)	
Parameters	**Point estimation**	**95% CI[ab]**	**Point estimation**	**95% CI[ab]**
N_{ANC}	9612	8977–10424	9013	8384–10146
N_{AFR}	23849	21634–44081	21762	15867–46813
NC_S	180,771	16598–411442	224,695	38694–446151
NC_Y	96,071	2464–461785	251,150	67722–428360
ND_S	3,704	412–6996	5187	2,000–5,700
N_{AY}	10251	2456–461785	5480	1730–15823
ND_Y	644	85–4553	3654	517–4680
$2Nm_S$	5.9	4.6–14	3.7	3.4–18
$2Nm_Y$	37.4	5–77	36.8	25–88
T_a	1,475 y	10–100	1,925 y	16–95
a_{YS}	0.19	0.04–0.28	0.08	0.03–0.19
a_{SY}	0.08	0.04–0.18	0.16	0.06–0.25
m_{SY}	4.45E-05	2.3E-06–9.9E-04	2.56E-04	3.1E-06–1.0E-03
m_{YS}	1.11E-04	1.2E-05–6.3E-04	1.53E-04	6.2E-06–2.4E-04
T_{EY}	4,250 y	101–691	7,450 y	162–567
T_{DS}	138,250 y	2482–9710	258,250 y	5358–12561

Inference of archaic admixture in modern humans

Simple model (proof of concept)



Altai Neandertal

Complete genomics CHB or TSI samples (4 inds / pop)

Other unsampled Neandertal

Data set:

Non coding DNA and non CpG sites.

Altai Neandertal (Prüfer et al. 2013), unfiltered vcf

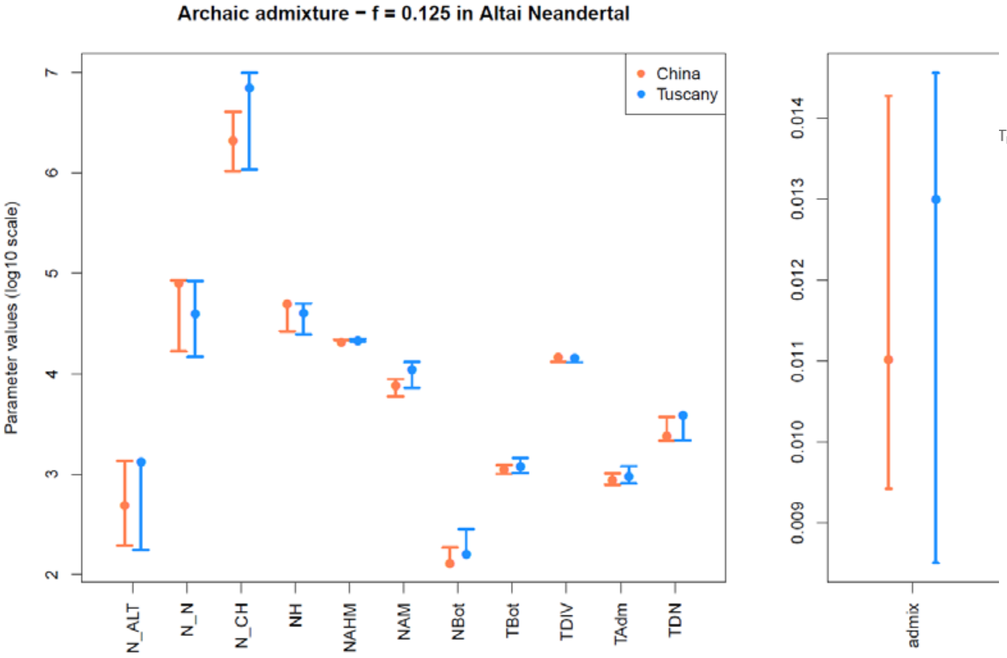
271,994 regions of 100 bp in non-coding DNA

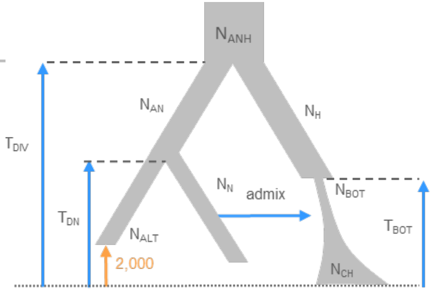
Ancestral state deduced by 1000G for 26,466,040 bp (26.5Mb)

All regions are at least 5 Kb apart from each other

Inference of archaic admixture in modern humans

Very preliminary results





Admixture level

CHB: 1.2% (0.94-1.43)

TSI: 1.3% (0.85-1.45)

Recent admixture ! !

TSI: 875 gen (790-1030)

CHB: 950 gen (810-1200)

<25,000 y

(assuming u=2e-8)

Possible extensions

- Multiprocessor version of fsc
- MCMC (Beaumont 2004, Garrigan 2009)
- Multilocus SFS
- Coalescent simulations through pedigrees

Thanks to:

Isabelle Duperret
Emilia Huerta-Sanchez
Isabel Alves
Vitor Sousa
Matthieu Foll

David Reich
Nick Patterson

Rasmus Nielsen

CMPG lab

FNSNF
Swiss National Science Foundation

SIB
Swiss Institute of
Bioinformatics