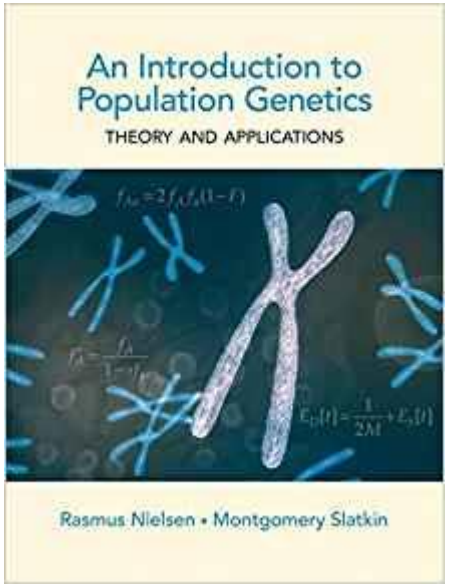
# 集団ゲノム学
# Population Genomics

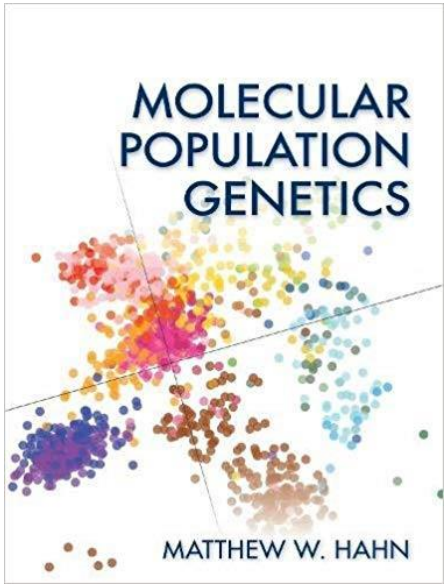
木村亮介
琉球大学大学院医学研究科
人体解剖学講座

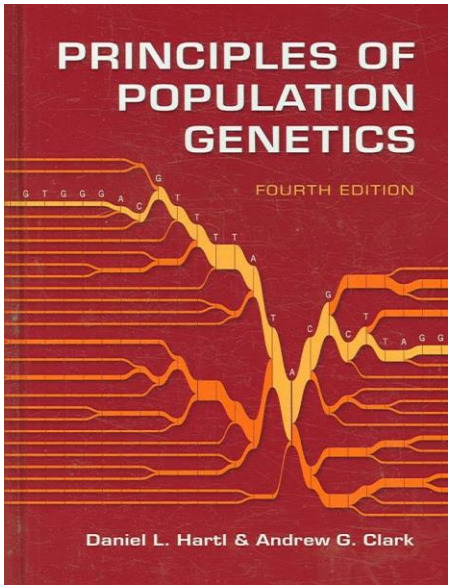
第3回進化セミナー2019
@御殿場高原時之栖

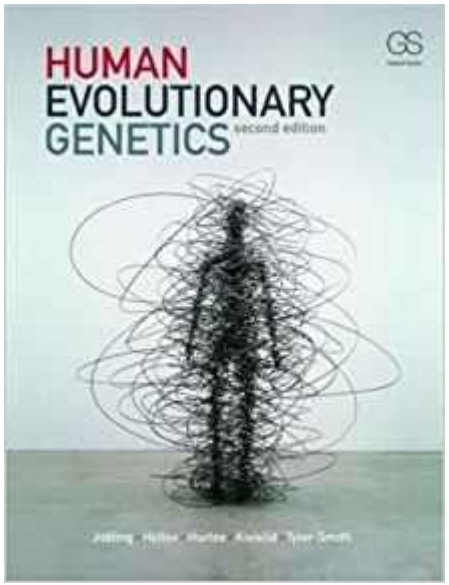
国立大学法人 琉球大学 University of the Ryukyus

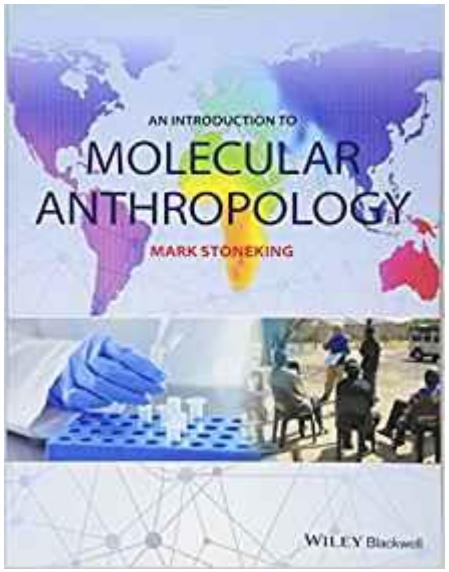
# 参考図書

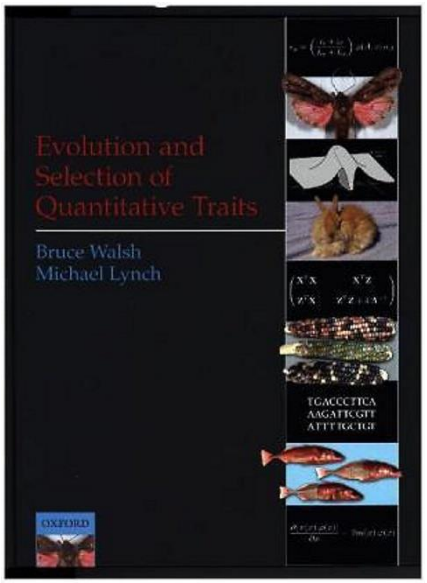

An Introduction to Population Genetics
THEORY AND APPLICATIONS
Rasmus Nielsen • Montgomery Slatkin


MOLECULAR POPULATION GENETICS
MATTHEW W. HAHN


PRINCIPLES OF POPULATION GENETICS
FOURTH EDITION
Daniel L. Hartl & Andrew G. Clark


HUMAN EVOLUTIONARY GENETICS
second edition
GS


AN INTRODUCTION TO
MOLECULAR ANTHROPOLOGY
MARK STONEKING
WILEY Blackwell


Evolution and Selection of Quantitative Traits
Bruce Walsh
Michael Lynch
OXFORD

# ゲノム解析の飛躍的進展とその人類学的応用

1. 疾患や形質に関連する遺伝子の同定
   Genome-wide association study (GWAS)

2. 過去の人口動態や移住（demography）、
   現在の集団構造（structure）の高解像解析

3. 自然選択が働いたゲノム領域の探索

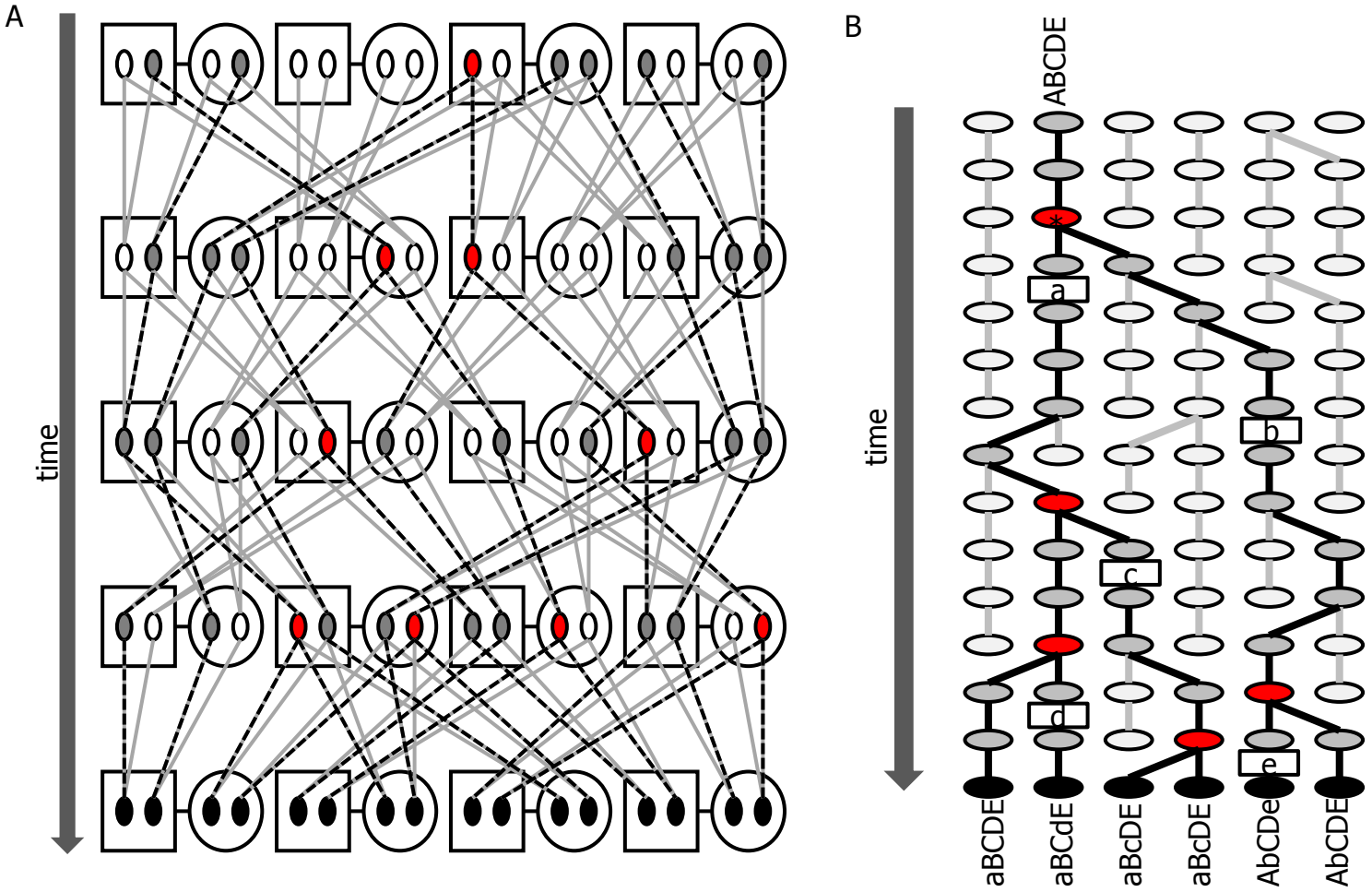
4. ネアンデルタール人などの古代ゲノム解析

など・・・

研究戦略の大転換！

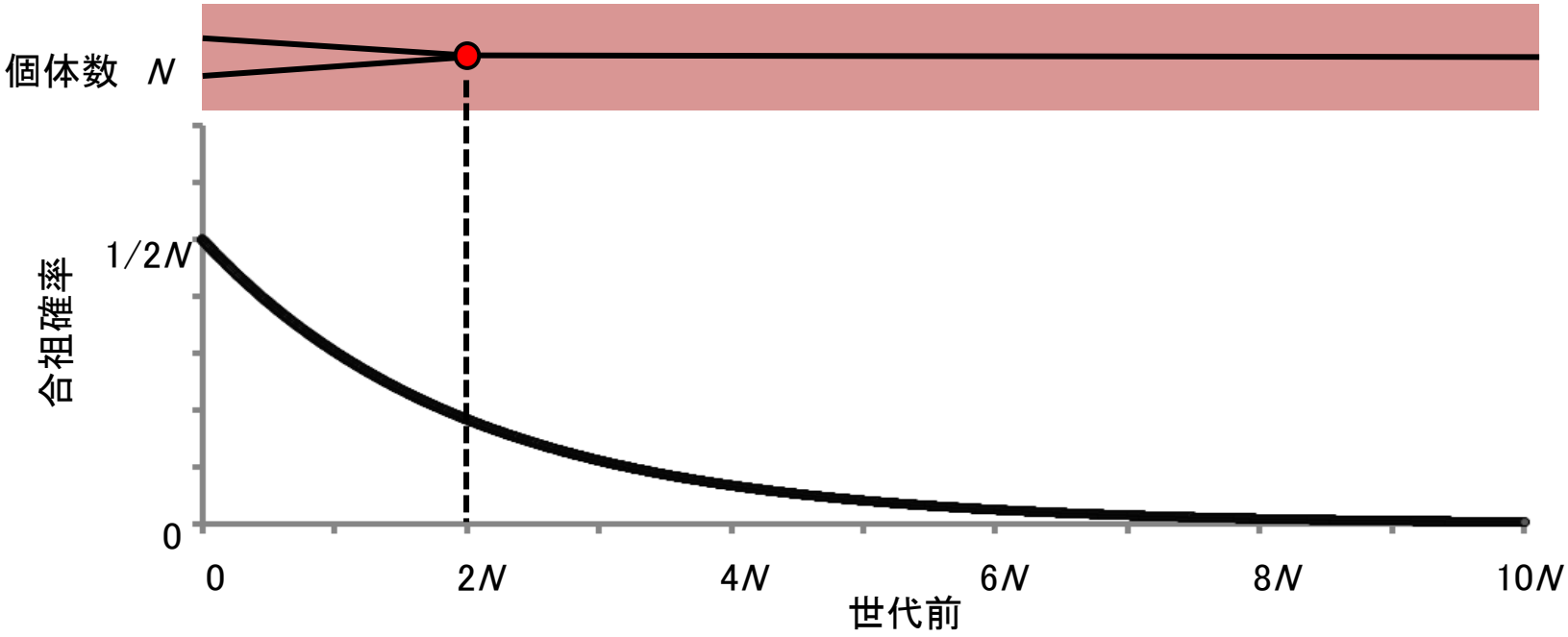
# 集団サイズを推定しよう

# Gene genealogy 遺伝子系図

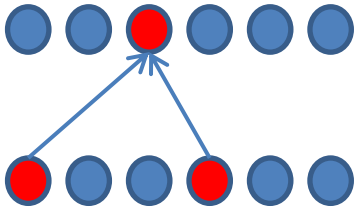



A：遺伝子の伝達（灰色：取りうる経路、点線：ある遺伝子座における実際の経路
B：集団中の全てのアリルは一つの共通祖先に遡ることができる

# Coalescence time　合祖時間



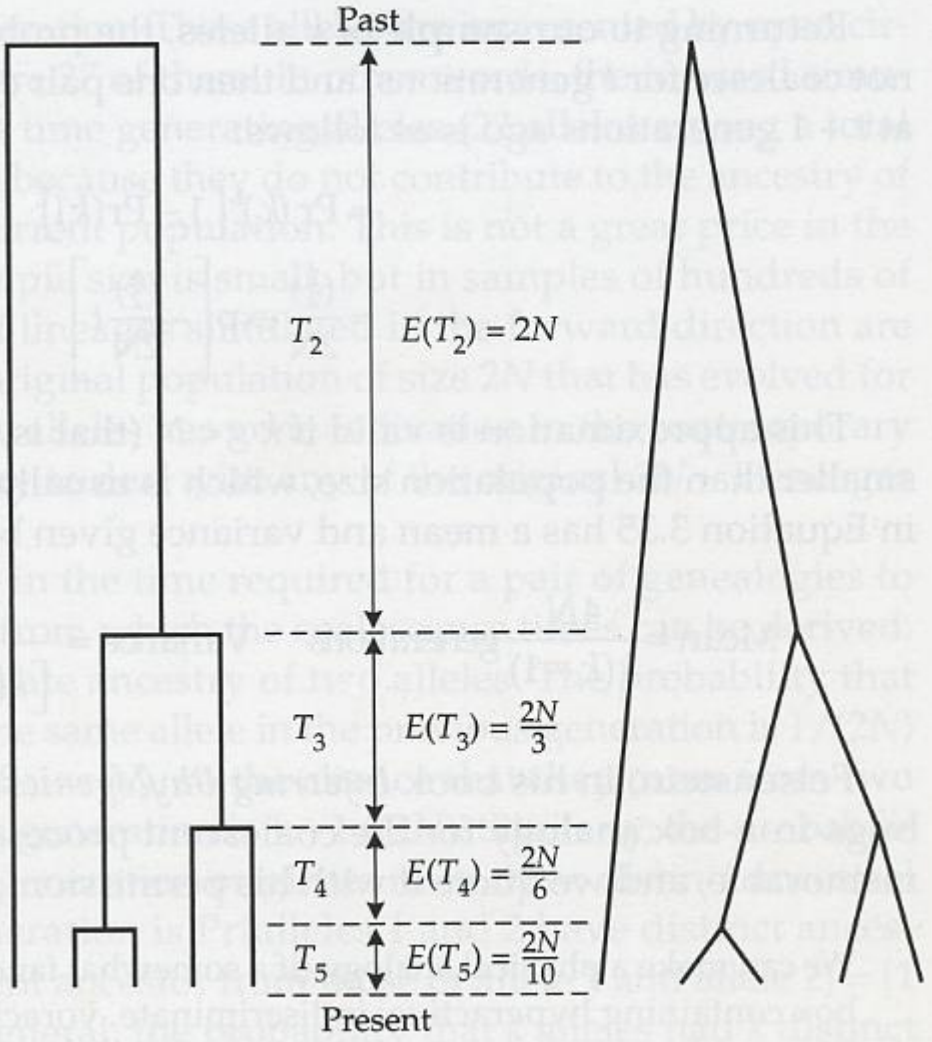

$$f_2(t) = (1/2N)(1-1/2N)^{t-1}$$
$$\simeq (1/2N)e^{-t/2N}$$



有効集団サイズ一定Nの集団中の合祖時間の平均値
任意の2つの配列:　2N　世代

# Coalescence theory

**FIGURE 3.15** Two completely equivalent ways of illustrating the coalescences in a gene tree. On the left, the coalescent events are represented as horizontal lines, on the left they are represented as nodes. In any each generation, if there are $k$ alleles present, the expected time back to the next coalescence is given by $4N/[k(1-k)]$. For example, starting with five alleles, the expected time back to the first coalescence is $4N/[(5)(4)] = 2N/10$. Note that the successive times get longer. When there are only two alleles, the time back to the final coalescence is $2N$ generations.




有効集団サイズ一定Nの集団中の合祖時間の平均値
抽出したn個の配列: 4N(1−1/n) 世代

# Bayesian Skyline Plot

**Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences**

*A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus*

Department of Zoology, University of Oxford, Oxford, United Kingdom

We introduce the Bayesian skyline plot, a new method for estimating past population dynamics through time from a sample of molecular sequences without dependence on a prespecified parametric model of demographic history. We describe a Markov chain Monte Carlo sampling procedure that efficiently samples a variant of the generalized skyline plot, given sequence data, and combines these plots to generate a posterior distribution of effective population size through time. We apply the Bayesian skyline plot to simulated data sets and show that it correctly reconstructs demographic history under canonical scenarios. Finally, we compare the Bayesian skyline plot model to previous coalescent approaches by analyzing two real data sets (hepatitis C virus in Egypt and mitochondrial DNA of Beringian bison) that have been previously investigated using alternative coalescent methods. In the bison analysis, we detect a severe but previously unrecognized bottleneck, estimated to have occurred 10,000 radiocarbon years ago, which coincides with both the earliest undisputed record of large numbers of humans in Alaska and the megafaunal extinctions in North America at the beginning of the Holocene.

Drummond et al. 2005

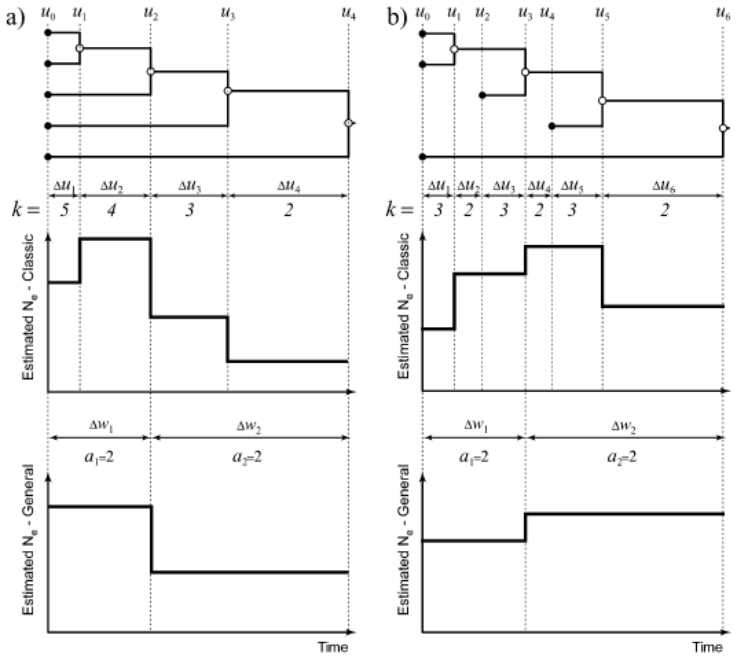



FIG. 1.—(*a*) A genealogy of five individuals sampled contemporaneously (top) together with its associated classic (middle) and generalized (bottom) skyline plots. (*b*) A genealogy of five individuals sampled at three different times (top) along with its associated classic (middle) and generalized (bottom) skyline plots. In the classic skyline plots, the changes in effective population size coincide with coalescent events, resulting in a stepwise function with $n - 2$ change points and $n - 1$ population sizes, where $n$ is the number of sampled individuals. In the generalized skyline plot, changes in effective population size coincide with some, but not necessarily all, coalescent events. The resulting stepwise function has $m - 1$ change points ($1 \leq m \leq n - 1$) and $m$ population sizes.

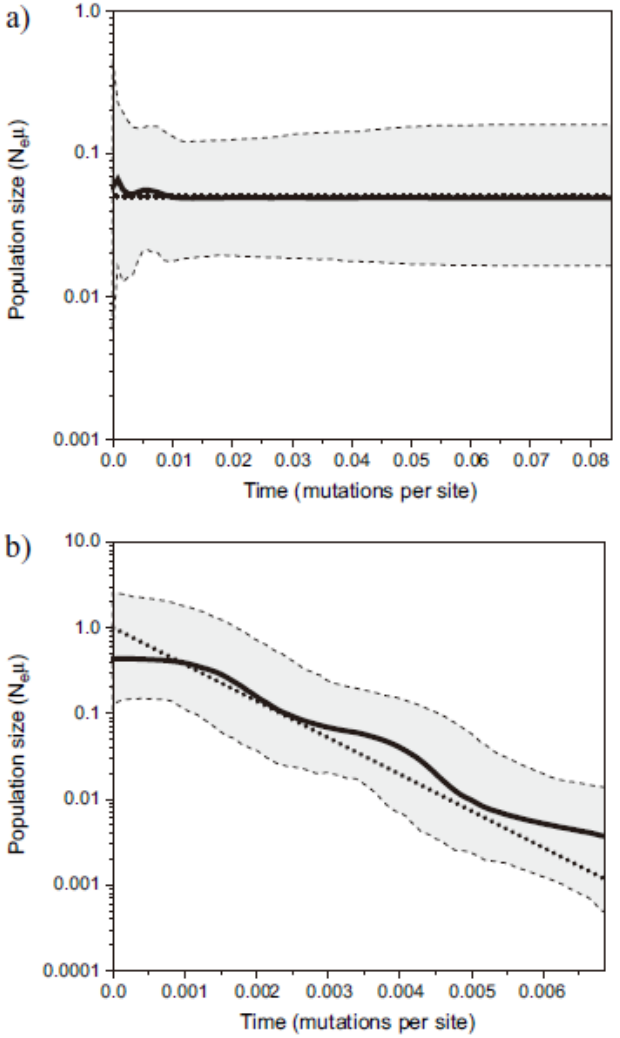
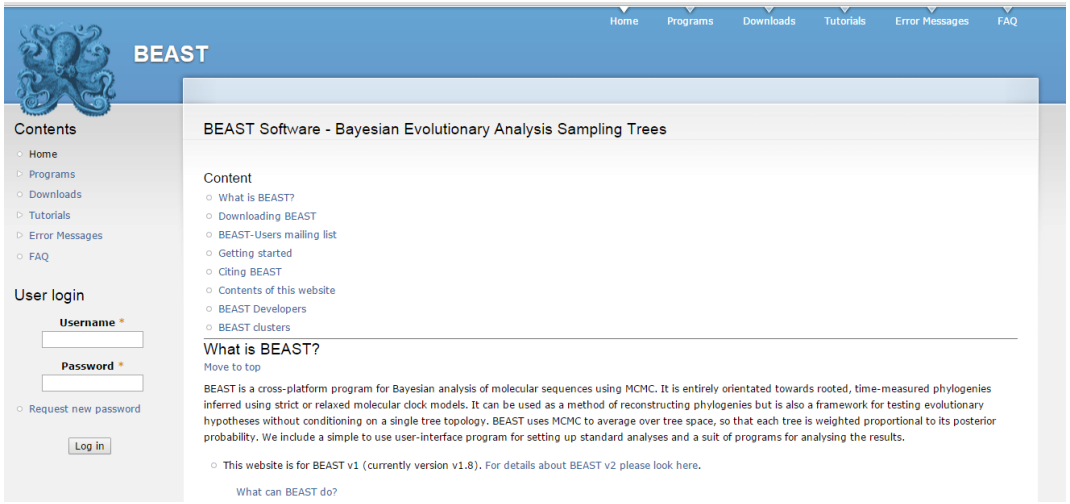



FIG. 2.—Performance of the Bayesian skyline plot on simulated data. Time is measured in units of mutations per site. The true demographic histories are shown as thick dotted lines, the median estimates are shown as thick solid lines, and the 95% HPD limits are shown by the gray areas bounded by thin dashed lines. (*a*) The Bayesian skyline plot ($m = 12$) calculated from a set of sequences that were simulated under a model of constant population size. (*b*) The Bayesian skyline plot ($m = 12$) calculated from a simulated data set for which the true demographic history was exponential growth (see text for details).

# BEAST Software - Bayesian Evolutionary Analysis Sampling Trees



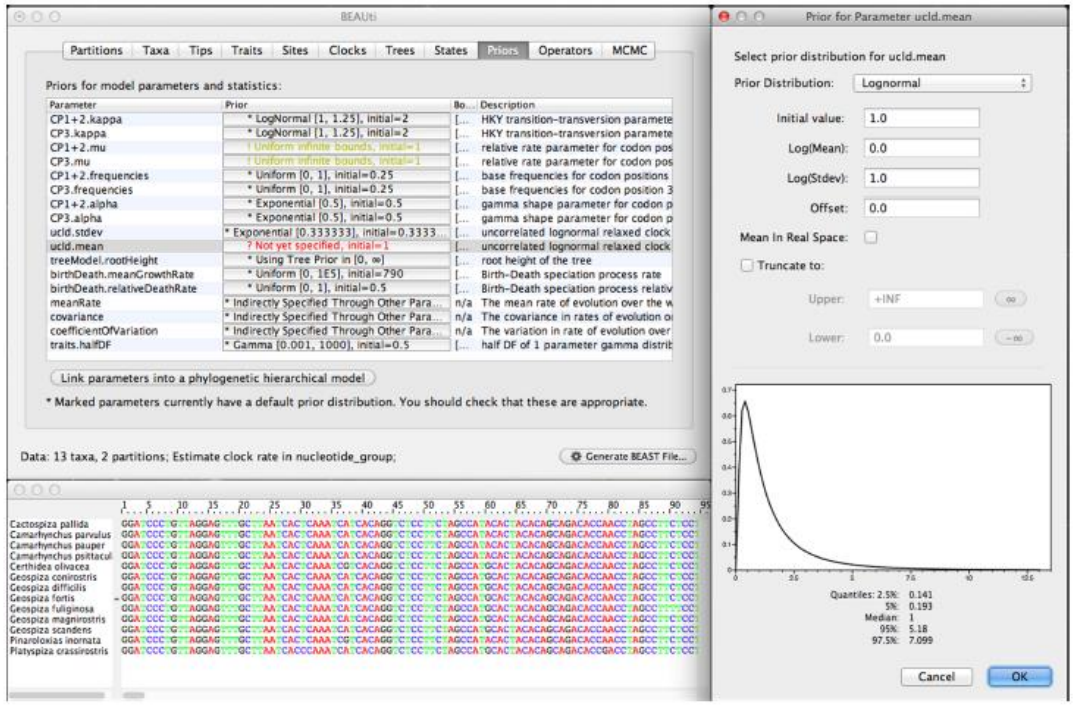




FIG. 1. BEAUti GUI for importing data and specifying the evolutionary model.

Drummond et al. 2012

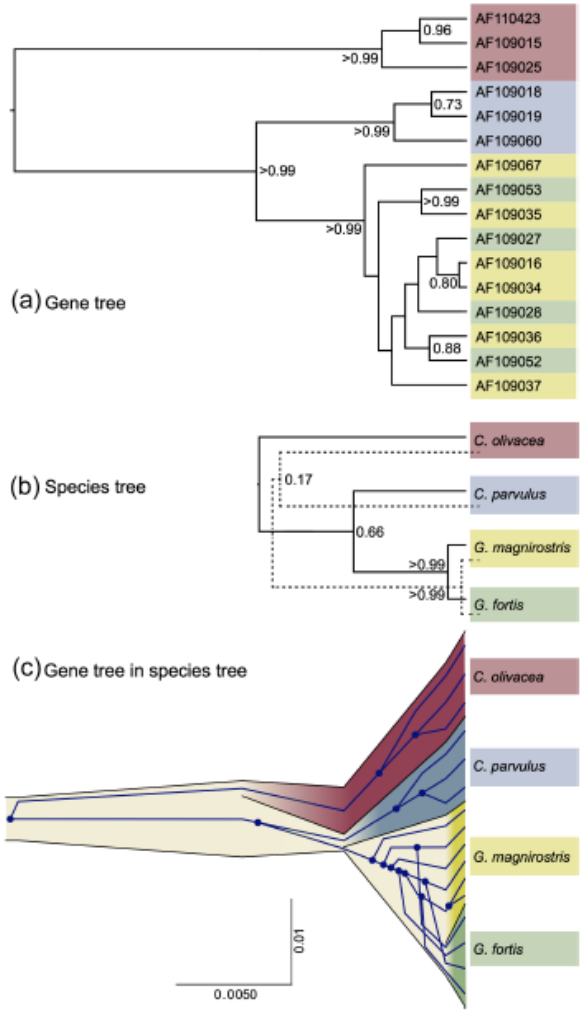
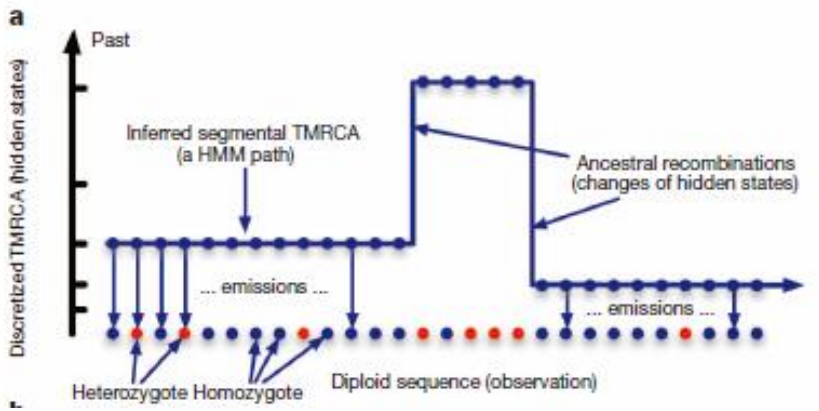
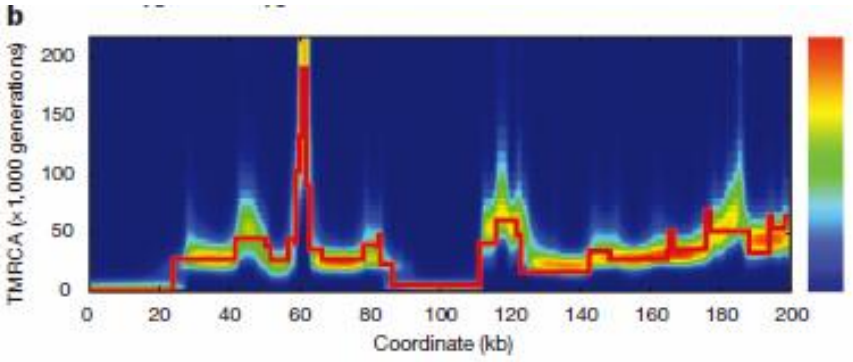



FIG. 3. (*a*) Representative gene tree of mitochondrial DNA fragment from 16 Darwin's finches of four species (*Geospiza fortis*, *G. magnirostris*, *Camarhynchus parvulus*, and *Certhidea olivacea*). Nodes that have posterior clade probabilities of greater than 0.5 are labeled with their posterior clade probability. (*b*) The two most probable species trees (solid line represents most probable species tree; dashed line is second most probable). (*c*) Gene tree embedded in a point estimate of the species tree, including divergence times and effective population sizes. The *x* axis is divergence time in units of substitutions per site and the *y* axis is proportional to effective population size.

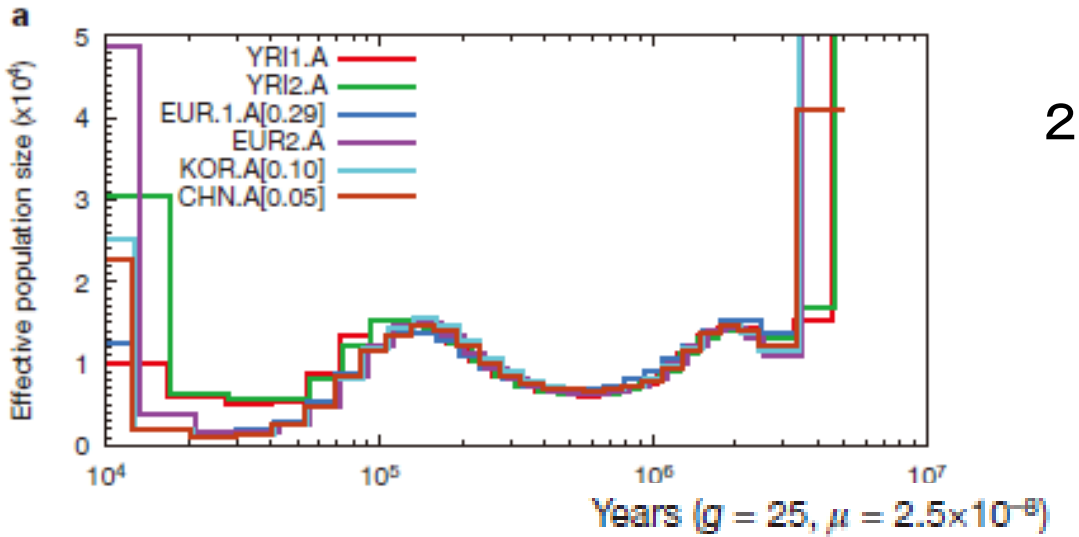
# ゲノムから有効集団サイズの変動を推定




1. 隠れマルコフをもちいた領域区分

2. TMRCAの分布から有効集団サイズの変遷を推定

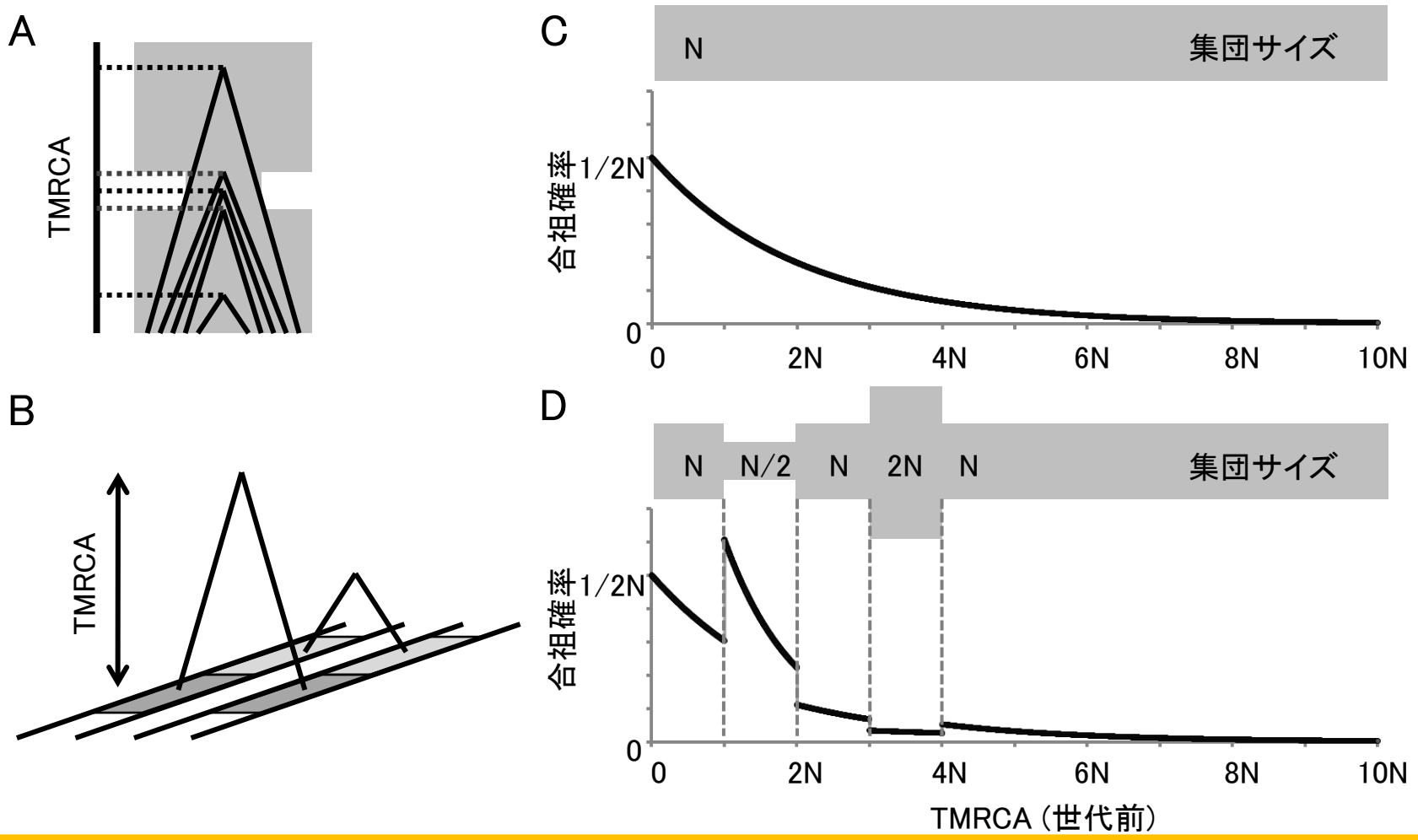



Li&Durbin 2010

1個体、つまり2本のゲノムから推定できる

pairwise sequential Markovian coalescent (PSMC) model

# 合祖時間の分布⇔集団サイズの変遷




集団サイズが減少すると合祖確率が増大
逆に合祖年代の分布をみることで集団サイズの変動を推定

| | Method [1] | $T_d$ [2] (kya) | $T_b$ [3] | $T_e$ [4] (kya) | $F$ [5] | Data | Comment |
|---|---|---|---|---|---|---|---|
| Marth et al. (2004) | AFS | N/A | 76 | 64 | 0.095 | $42 \times 33$k sites | |
| Adams and Hudson (2004) | coal-sim | N/A | 30 | N/A | N/A | Seattle SNPs | Size reduction followed by exponential growth. |
| Schaffner et al. (2005) | coal-sim | 52.5 | 52.5/30 | N/A | 0.085/0.067 | 3988 sites, genotyping | Two sharp bottlenecks mixed with reduced population size (Figure S6) |
| Plagnol and Wall (2006) | coal-sim | 130 | 60 | N/A | 0.20/0.24 | $34 \times 3.2$Mb, EGP | 0.20 for model without introgression; 0.24 with introgression; $F$ calculated from the *ms* command line |
| Liu et al. (2006) | | 56.1 | N/A | N/A | N/A | | |
| Garrigan et al. (2007) | coal-sim | 39.5 | 39.5 | N/A | N/A | $431 \times 16$kb from M,X,Y | Quoting the divergence time between Dogon and Mongolian. Reduction followed by exponential growth. |
| Fagundes et al. (2007) | coal-sim | 51.1 | 51.1 | N/A | N/A | $50 \times 25$kb from auto. | Reduction followed by exponential growth. |
| Keinan et al. (2007) | AFS | N/A | $23 \pm 2$ | N/A | $0.18 \pm 0.01$ | HapMap and Perlegen | One-bottleneck model. Bottleneck at $32 \pm 3$ for European. |
| Cox et al. (2008) | coal-sim | 27.7 | N/A | N/A | N/A | $100 \times 100$kb from X | Quoting the divergence time between Han and Biaka. Divergence time estimed on chrX with migration modeled. |
| Wall et al. (2009) | coal-sim | 80 | 30+1 | 30 | 0.27/0.33 | $58 \times 5.3$Mb, EGP | 0.27 without introgression; 0.33 with introgression; bottleneck duration fixed at 1kya; Eu-Af diverged at 120kya. |
| Gutenkunst et al. (2009) | diffusion | 140 | 140/21.2 | N/A | N/A | EGP, non-coding only | Two bottlenecks at the Af–Eu/As divergence and at the Eu–As divergence. |

Table S2: Inferred out-of-Africa event in current literature. Underlined numbers are fixed (not inferred) in the model. [1] 'coal-sim' denotes the category of methods that use a coalescent simulator to fit the observed data; 'AFS' denotes methods that explicitly compute the likelihood of an allele frequency spectrum given a piecewise constant population size history; 'diffusion' denotes the diffusion approximation on the joint allele frequency spectrum. [2] The divergence time of African and non-African populations. If the time is in unit of generations in the literature, it is scaled to years under the assumption of $10^{-9}$ mutation per site per year (equivalent to $\mu = 2.5 \times 10^{-8}$ and $g = 25$). [3] Time of the start of the bottleneck or population size reduction. [4] Time of the end of the bottleneck. [5] Inbreeding coefficient. $F$ can only be calculated when the model explicitly models the bottleneck.

Li&Durbin 2010

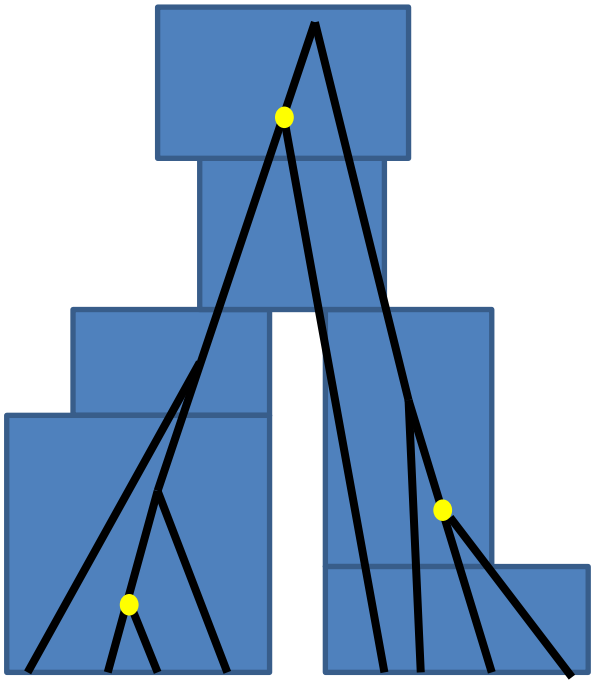
# Multiple sequential Markovian coalescent (MSMC) model

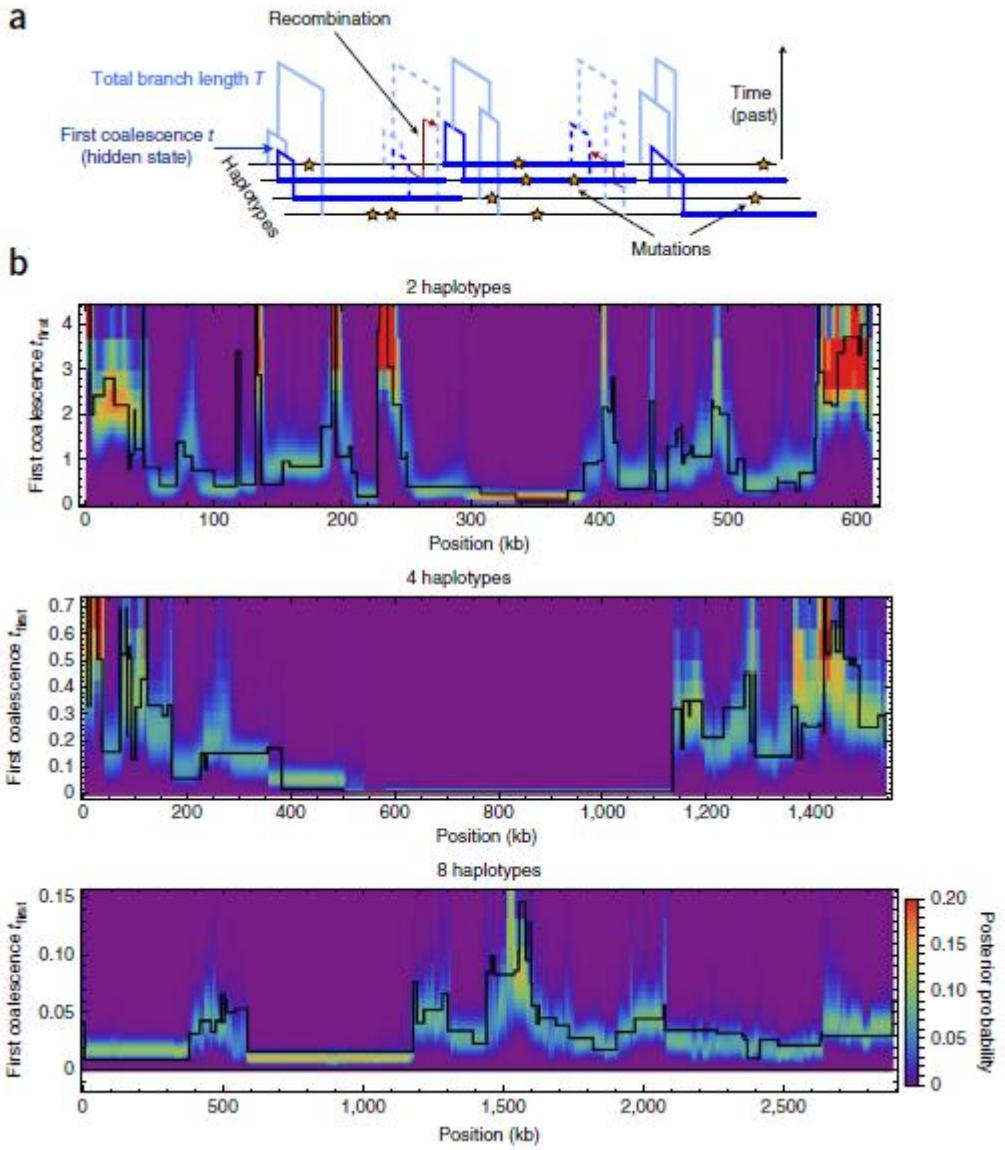
Inferring human population size and separation history from multiple genome sequences

Stephan Schiffels & Richard Durbin

Schiffels&Durbin 2014

PSMCを複数個体に拡張





Schiffels&Durbin 2014

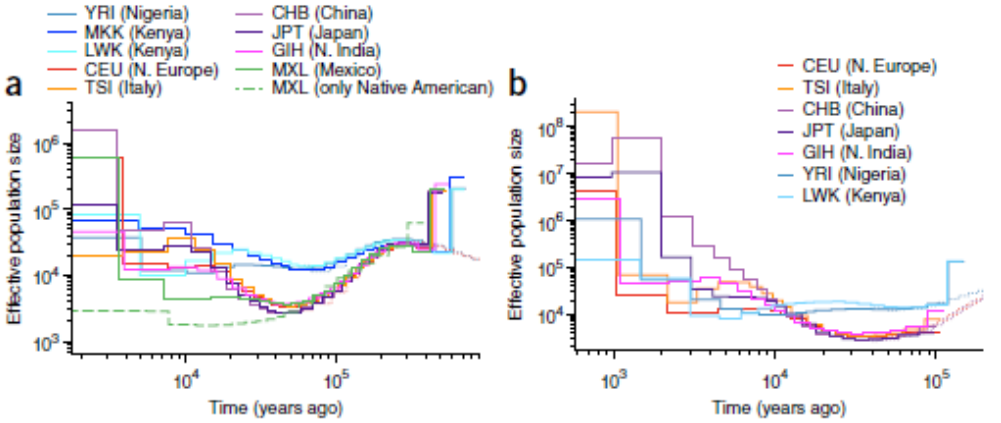



**Figure 3** Inference of population size from whole-genome sequences. (**a**) Population size estimates from four haplotypes (two phased individuals) from each of nine populations. The dashed line was generated from a reduced data set of only the Native American components of the MXL genomes. Estimates from two haplotypes for CEU and YRI are shown for comparison as dotted lines. N, Northern. (**b**) Population size estimates from eight haplotypes (four phased individuals) from the same populations as in **a** but excluding MXL and MKK. In contrast to estimates with four haplotypes, estimates are more recent. For comparison, we show the result from four haplotypes for CEU, CHB and YRI as dotted lines.

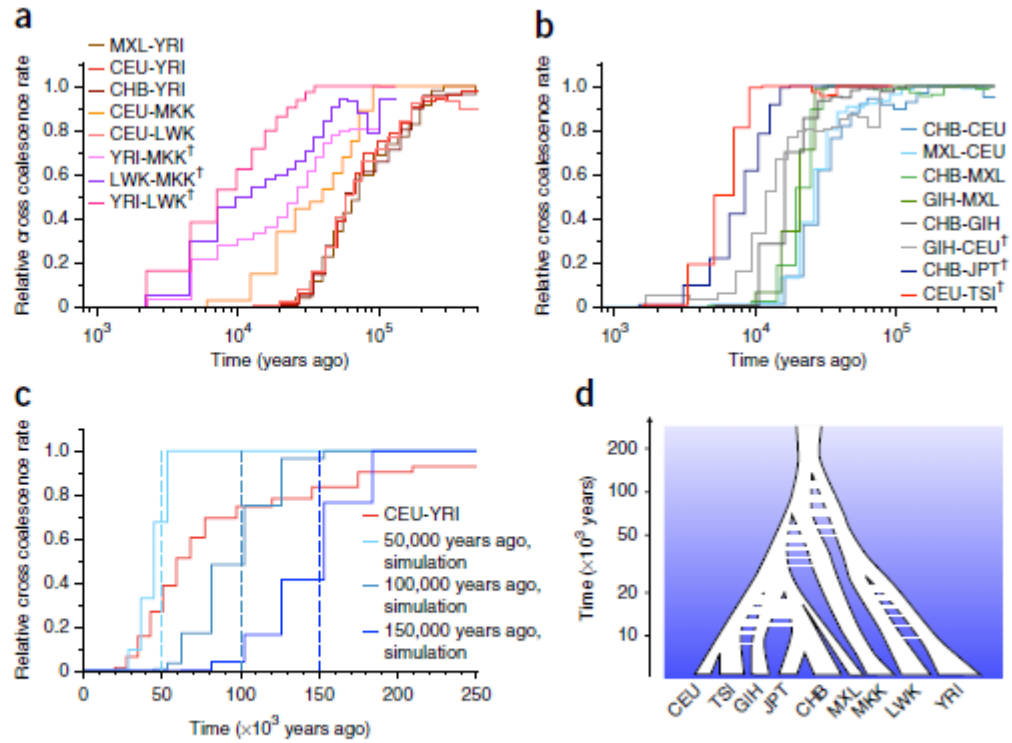
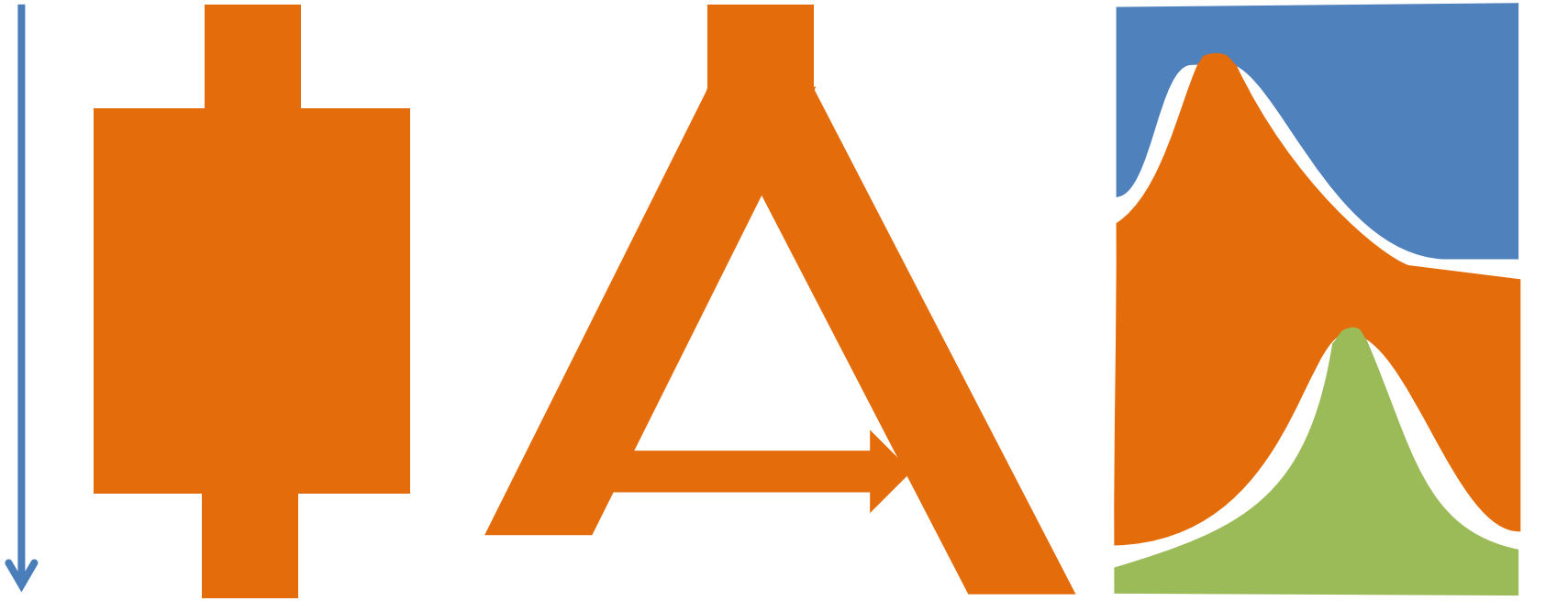



**Figure 4** Genetic separation between population pairs. (**a**) Relative cross coalescence rates in and out of Africa. African–non-African pairs are shown in red, and pairs within Africa are shown in purple. (**b**) Relative cross coalescence rates between populations outside Africa. European–East Asian pairs are shown in blue, Asian-MXL pairs are shown in green, and other non-African pairs are shown in other colors, as indicated. The pairs that include MXL are masked to include only the putative Native American components. In **a** and **b**, the most recent population separations are inferred from eight haplotypes, that is, four haplotypes from each population, and corresponding pairs are indicated by a cross. (**c**) Comparison of the African–non-African split with simulations of clean splits. We simulated three scenarios, at split times 50,000, 100,000 and 150,000 years ago. The comparison demonstrates that the history of relative cross coalescence rate between African and non-African ancestors is incompatible with a clean split model and suggests it progressively decreased from beyond 150,000 years ago to approximately 50,000 years ago. (**d**) Schematic of population separations. Timings of splits, population separations, gene flow and bottleneck are shown along a logarithmic axis of time.

# 「有効集団サイズ」の注意点

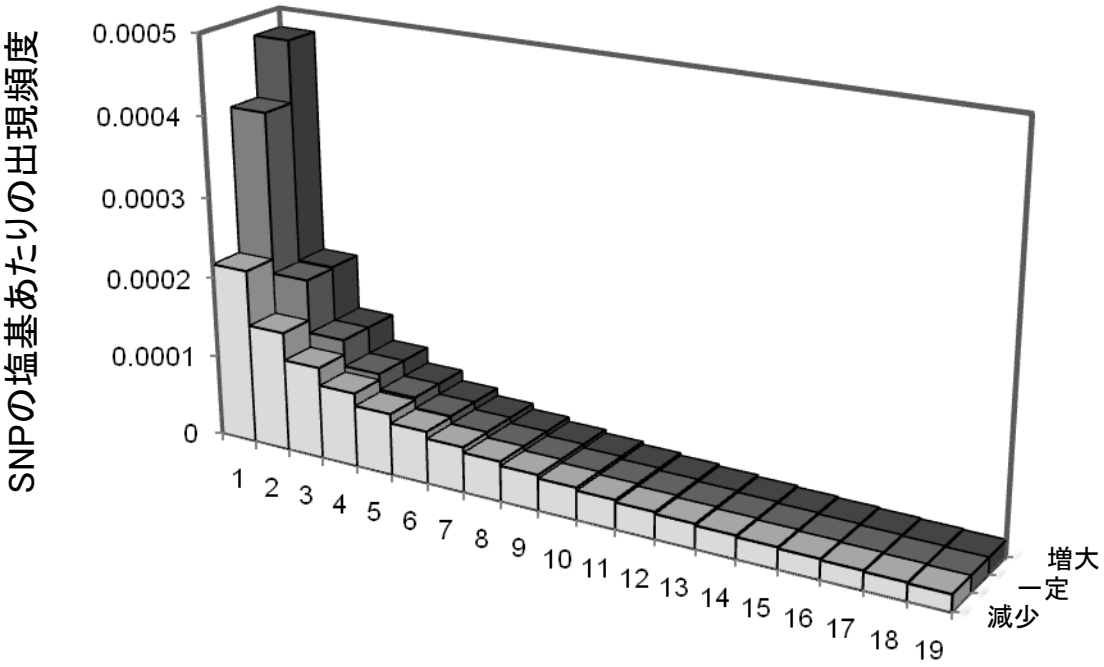
有効集団サイズ:
ランダム交配を仮定した場合の集団サイズ
≒後世に遺伝子を残している個体数
≠実際の個体数



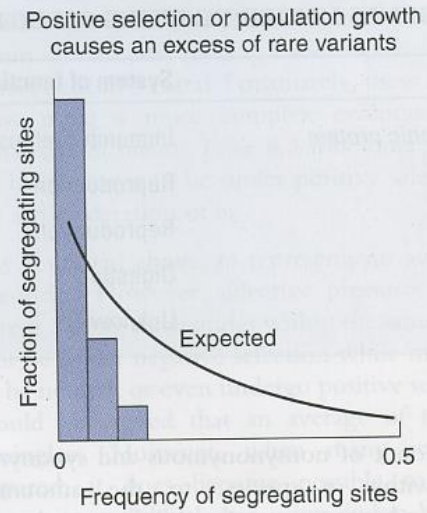
実集団サイズの変動　　分岐＆合流　　自然選択＆生殖隔離

# Allele frequency spectrum (AFS)

デモグラフィ
自然選択
の影響を受ける。




SNPにおける派生型アリルの頻度
（染色体20 本中）



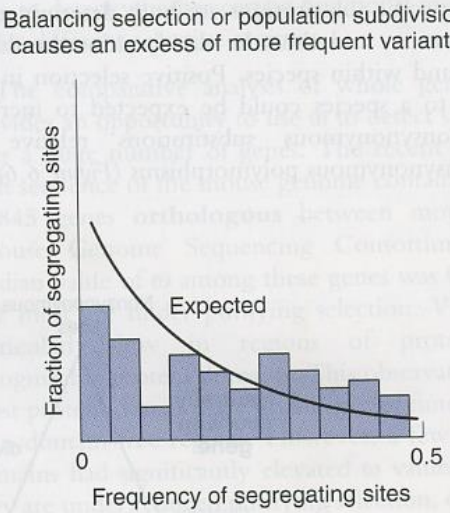




**Figure 6.7:** Frequency spectrum of segregating sites under different types of selection.

Two types of deviation from the site frequency spectrum expected in a constant size neutrally evolving population (smooth line) are shown. One spectrum shows rare alleles (those at low population frequency) being more prevalent than expected, and the other shows intermediate alleles being over-represented. Both scenarios can be caused by either selection or demographic factors.

# デモグラフィからASFを算出



## The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations

Gabor T. Marth,[1] Eva Czabarka, Janos Murvai and Stephen T. Sherry

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894*

Manuscript received April 15, 2003
Accepted for publication September 4, 2003

Marth et al. 2004

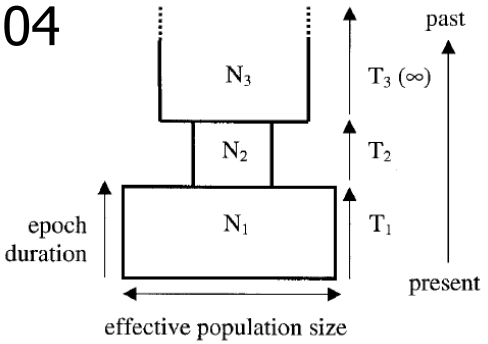



FIGURE 1.—Example of a three-epoch, piecewise constant, bottleneck-shaped population history profile. The ancestral effective population size ($N_3$) is followed by an instant reduction of effective size ($N_2$). The duration of this epoch is $T_2$ generations. This is followed by a stepwise increase of effective population size to $N_1$, $T_1$ generations before the present.

$$E(\Psi_i) = \frac{4\mu N_1}{i} + \sum_{m=1}^{M-1}\left\{4\mu\frac{N_{m+1}-N_m}{i}\binom{n-1}{i}^{-1} \times \sum_{k=2}^{n}\left[\binom{n-k}{i-1}\sum_{j=k}^{n}\left(e^{-\binom{j}{2}\tau_m^*}\prod_{\substack{k\le l\le n\\ l\neq j}}\frac{l(l-1)}{l(l-1)-j(j-1)}\right)\right]\right\}, \quad (1)$$

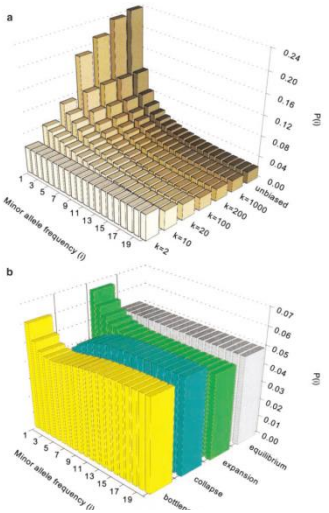



FIGURE 2.—Ascertainment bias. (a) Folded spectra under stationary history, at various values of "discovery sample" size $k$ (METHODS). (b) Allele frequency spectra predicted under competing scenarios of population history (conditioned on pairwise ascertainment $k = 2$). Equilibrium history, $N_1 = 10{,}000$; expansion, $N_1 = 20{,}000$, $T_1 = 3000$, $N_2 = 10{,}000$; collapse, $N_1 = 2000$, $T_1 = 500$, $N_2 = 10{,}000$; bottleneck history, $N_1 = 20{,}000$, $T_1 = 3000$, $N_2 = 2000$, $T_2 = 500$, $N_1 = 10{,}000$. (a and b) Sample size $n = 41$.

**TABLE 1**

**Results of fitting multi-epoch models of allele frequency spectrum to population-specific observed allele frequency data**

| Model structure | Model parameters | Resulting pairwise θ (units of $10^{-4}$) | ln $P$(data\|model) | Improvement over lower-epoch model |
|---|---|---|---|---|
| | | a. European data | | |
| One epoch | $N_1 = 10{,}000$ | 8.00 | −55.98 | — |
| Two epoch | $N_2 = 10{,}000$ | 8.74 | −38.11 | $2 \ln \lambda = 35.74$ |
| | $N_1 = 140{,}000$ | | | $P < 10^{-4}$ |
| | ($T_1 = 2{,}000$) | | | Highly significant |
| Three epoch | $N_3 = 10{,}000$ | 7.88 | −23.72 | $2 \ln \lambda = 28.78$ |
| | $N_2 = 2{,}000$ | | | $P < 10^{-4}$ |
| | ($T_2 = 500$) | | | Highly significant |
| | $N_1 = 20{,}000$ | | | |
| | ($T_1 = 3{,}000$) | | | |
| | | b. Asian data | | |
| One epoch | $N_1 = 10{,}000$ | 8.00 | −74.26 | — |
| Two epoch | $N_2 = 10{,}000$ | 8.63 | −31.95 | $2 \ln \lambda = 84.62$ |
| | $N_1 = 50{,}000$ | | | $P < 10^{-4}$ |
| | ($T_1 = 2{,}000$) | | | Highly significant |
| Three epoch | $N_3 = 10{,}000$ | 8.24 | −26.39 | $2 \ln \lambda = 11.12$ |
| | $N_2 = 3{,}000$ | | | $P = 0.0039$ |
| | ($T_2 = 600$) | | | Significant |
| | $N_1 = 25{,}000$ | | | |
| | ($T_1 = 3{,}200$) | | | |
| | | c. African-American data | | |
| One epoch | $N_1 = 10{,}000$ | 8.00 | −197.86 | — |
| Two epoch | $N_2 = 10{,}000$ | 9.20 | −28.69 | $2 \ln \lambda = 338.34$ |
| | $N_1 = 18{,}000$ | | | $P < 10^{-4}$ |
| | ($T_1 = 7{,}500$) | | | Highly significant |
| Three epoch | $N_3 = 10{,}000$ | 10.29 | −26.72 | $2 \ln \lambda = 3.94$ |
| | $N_2 = 16{,}000$ | | | $P = 0.1395$ |
| | ($T_2 = 15{,}000$) | | | Not significant |
| | $N_1 = 26{,}000$ | | | |
| | ($T_1 = 2{,}400$) | | | |

# SMC++

Terhost et al. 2017

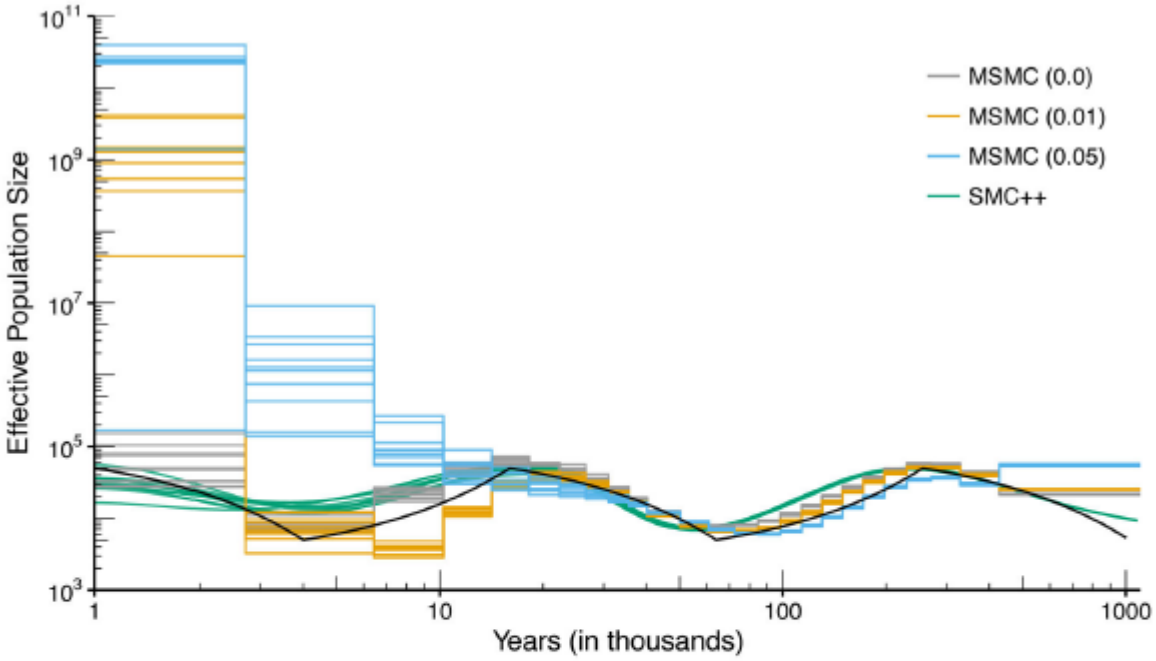



**Figure 1. The effect of phasing error**

The true population size history is indicated by a bold black line, while colored lines indicate inferred histories for ten simulations each with sample size $n = 4$. For MSMC, switch error was introduced at the rate of 0%, 1%, or 5%, indicated in parenthesis in the legend. SMC++ does not require phased data and its results are insensitive to phasing errors. With phasing error, MSMC estimates can be off by orders of magnitude in the recent past. In the absence of phasing error, the accuracy of MSMC is comparable to that of SMC++, with SMC++ producing higher resolution in the recent past.

# 連鎖不平衡係数$r^2$から過去の集団サイズを推定する

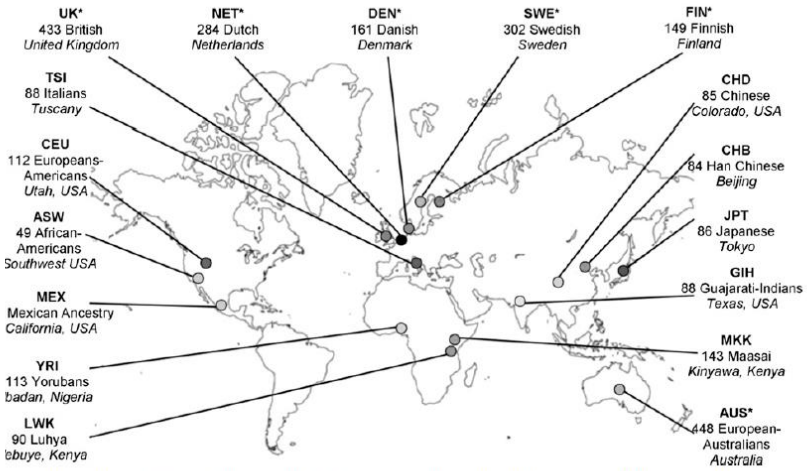
McEvoy et al. 2011

## Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs

Brian P. McEvoy,[1,4] Joseph E. Powell,[1,4,5] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

[1]*Queensland Institute of Medical Research, Brisbane 4006, Australia;* [2]*Department of Primary Industries Victoria, Bundoora 3083, Australia;* [3]*Faculty of Land and Environment, University of Melbourne, Parkville 3052, Australia*

Genetic and fossil evidence supports a single, recent (<200,000 yr) origin of modern *Homo sapiens* in Africa, followed by later population divergence and dispersal across the globe (the "Out of Africa" model). However, there is less agreement on the exact nature of this migration event and dispersal of populations relative to one another. We use the empirically observed genetic correlation structure (or linkage disequilibrium) between 242,000 genome-wide single nucleotide polymorphisms (SNPs) in 17 global populations to reconstruct two key parameters of human evolution: effective population size ($N_e$) and population divergence times ($T$). A linkage disequilibrium (LD)–based approach allows changes in human population size to be traced over time and reveals a substantial reduction in $N_e$ accompanying the "Out of Africa" exodus as well as the dramatic re-expansion of non-Africans as they spread across the globe. Secondly, two parallel estimates of population divergence times provide clear evidence of population dispersal patterns "Out of Africa" and subsequent dispersal of proto-European and proto-East Asian populations. Estimates of divergence times between European–African and East Asian–African populations are inconsistent with its simplest manifestation: a single dispersal from the continent followed by a split into Western and Eastern Eurasian branches. Rather, population divergence times are consistent with substantial ancient gene flow to the proto-European population after its divergence with proto-East Asians, suggesting distinct, early dispersals of modern *H. sapiens* from Africa. We use simulated genetic polymorphism data to demonstrate the validity of our conclusions against alternative population demographic scenarios.




s, codes, and sample sizes. The sampling location is indicated in italics. GenomEUtwin populations are marked *; otherwise

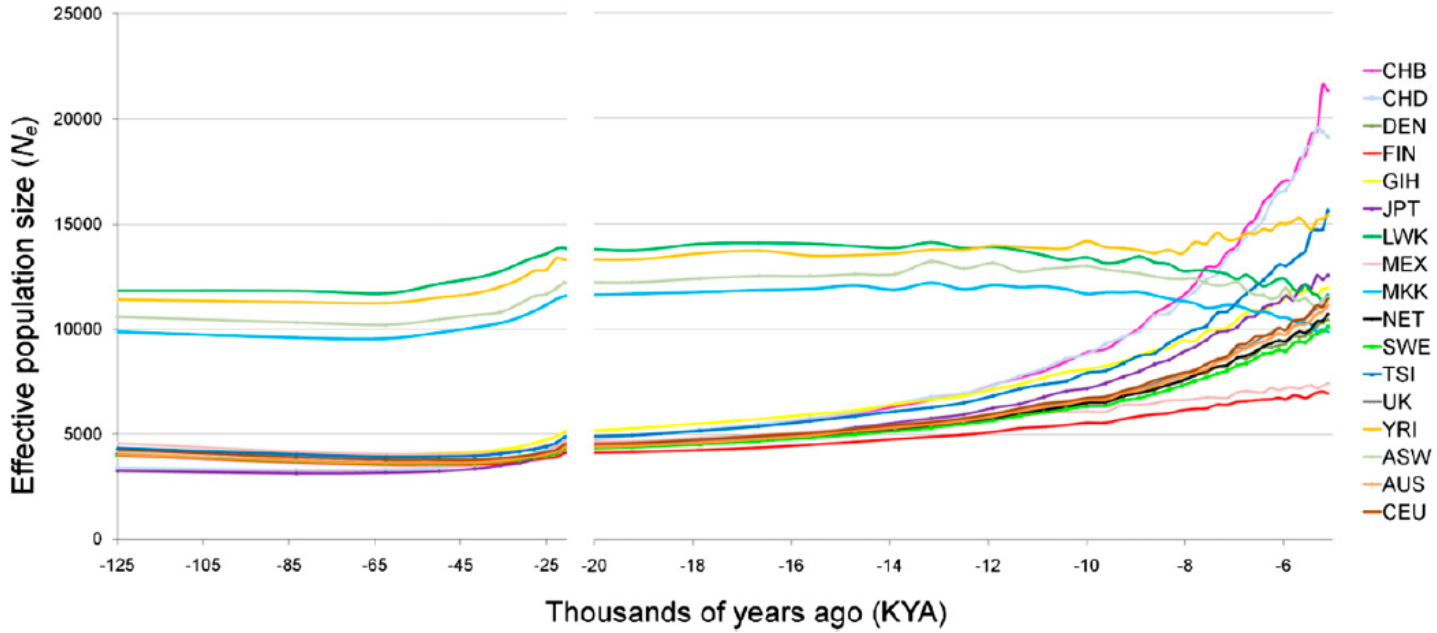



**Figure 2.** Spatial and temporal variation in $N_e$ estimates. $N_e$ was calculated from LD observations in each of 50 recombination distance classes for each population. Note the change in the time axis scale at 20 KYA. The underlying LD structure upon which these $N_e$ estimates are based can be seen in Supplemental Figure 5.

# 集団構造を観察しよう

# 従来のヒト集団研究

集団間の系統関係を明らかにするため

- mtDNA
- Y chromosome

が多く解析されてきた。

## 利点

- 組換えが起きない→突然変異が蓄積
- 突然変異率が高い（mtDNA）
- 有効集団サイズが小さいため（常染色体の1/4）、遺伝的浮動の効果が大きく、集団間に違いが生じやすい
- 女系・男系の祖先を遡れる

## 問題点

- 組換えが起きない
  →それぞれ1つの領域に過ぎない
  必ずしも集団の系統を反映しない
  1つの経路しか遡れない

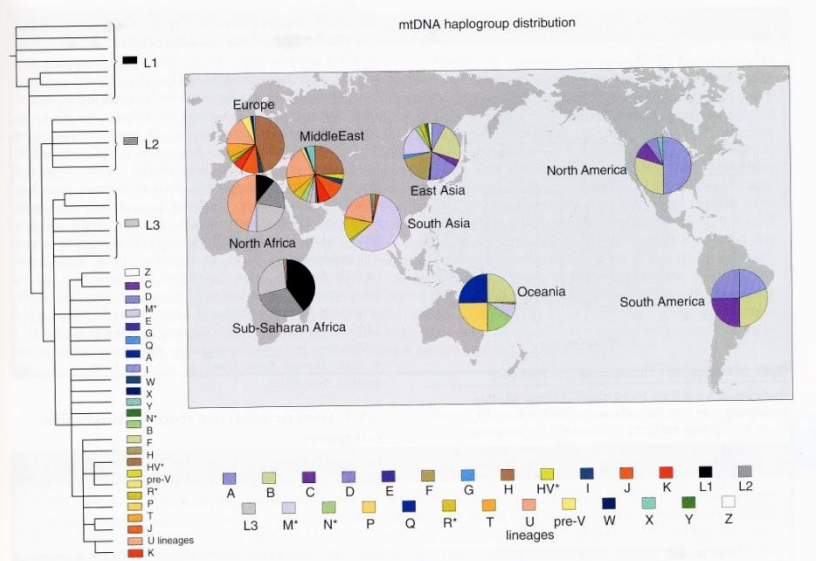



**Figure 9.16:** Geographical distribution of the major mtDNA clades.

Each major clade (haplogroup) is assigned a color reflecting its position in the phylogeny (left), and its frequency in population samples from broad geographical regions is shown in the pie charts. Note the most basal haplogroups (L) in Africa and the contrast between the frequencies of M (and subclades) compared to N (and subclades) in southern Asia compared with more northern parts of Asia and Europe. Color figure kindly sponsored by Oxford Ancestors Ltd.

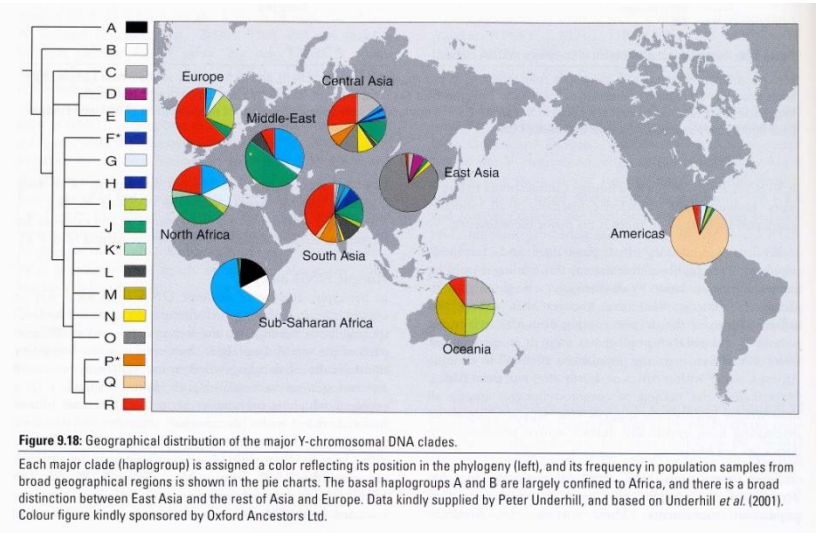
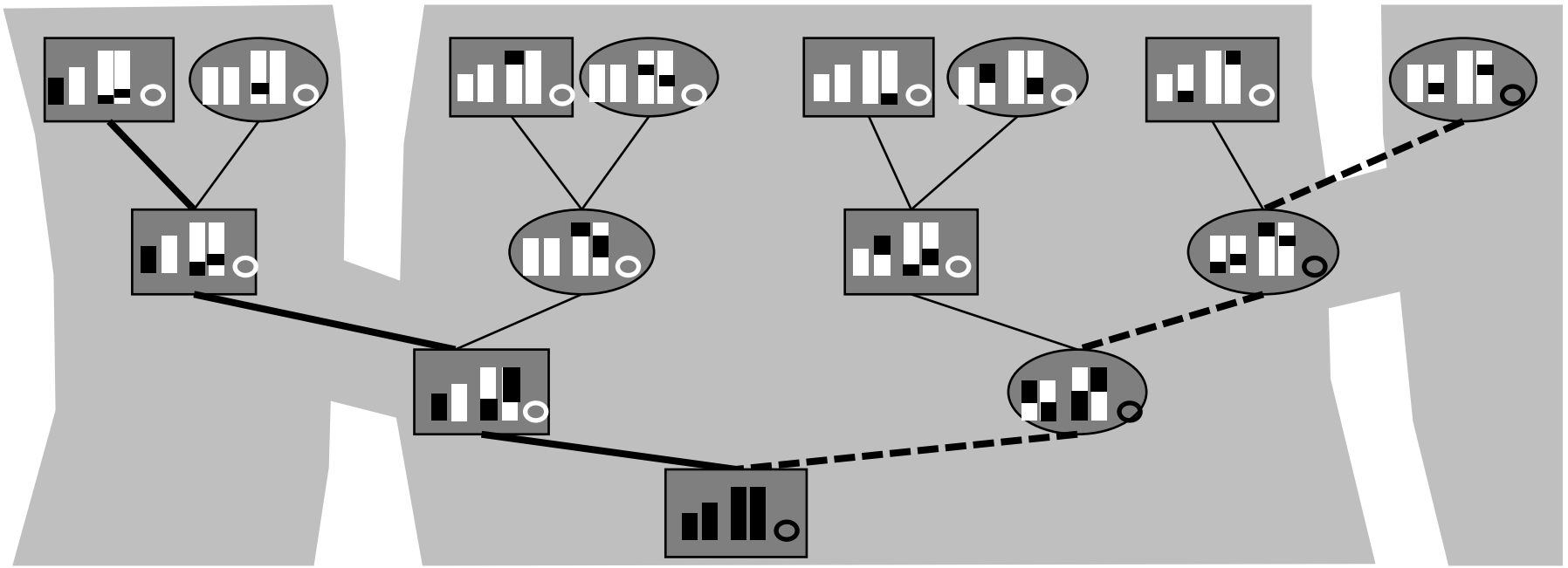



**Figure 9.18:** Geographical distribution of the major Y-chromosomal DNA clades.

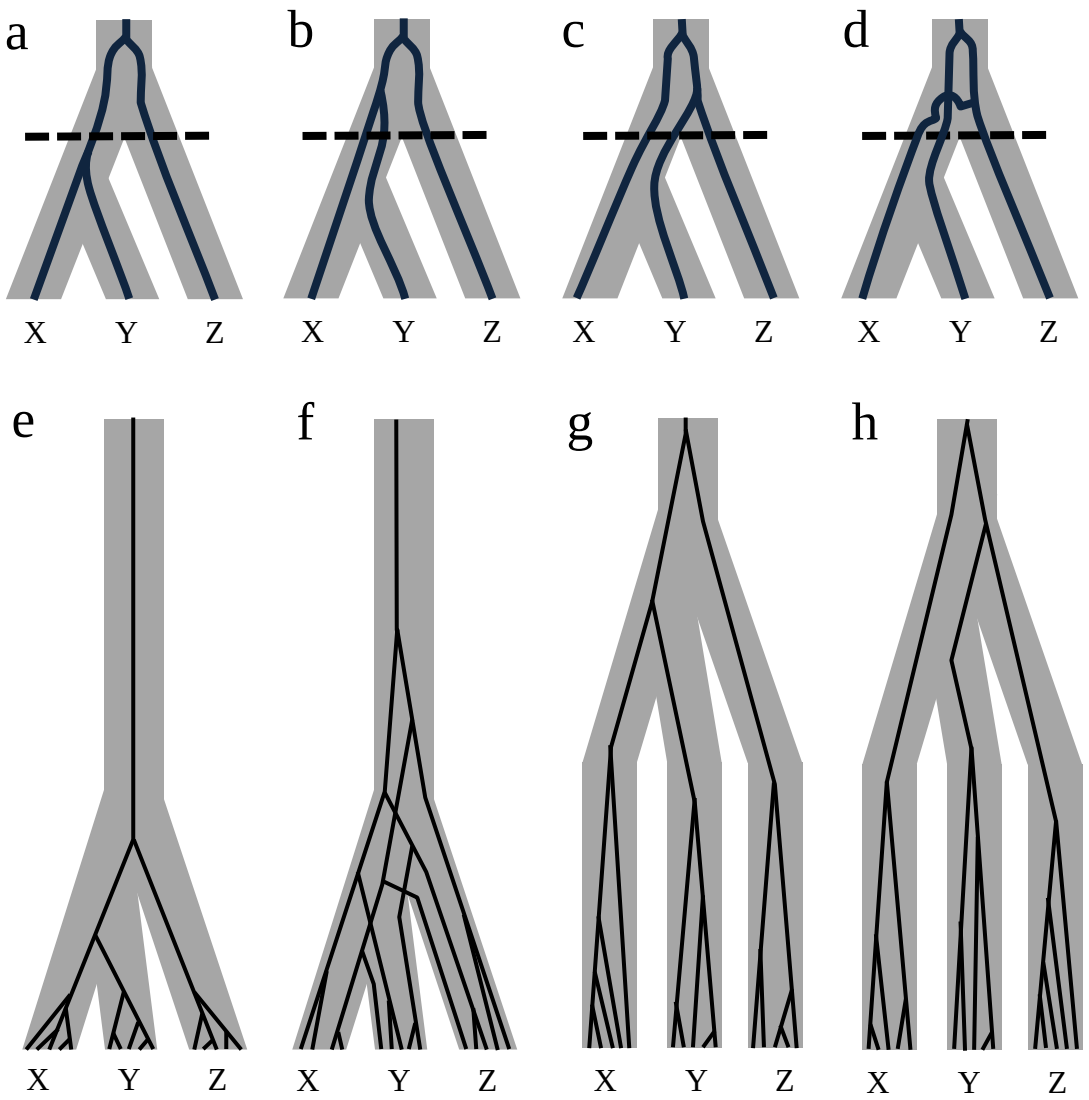
Each major clade (haplogroup) is assigned a color reflecting its position in the phylogeny (left), and its frequency in population samples from broad geographical regions is shown in the pie charts. The basal haplogroups A and B are largely confined to Africa, and there is a broad distinction between East Asia and the rest of Asia and Europe. Data kindly supplied by Peter Underhill, and based on Underhill *et al.* (2001). Colour figure kindly sponsored by Oxford Ancestors Ltd.

# 一個体への遺伝子の伝達



ひとりの現代人は多くの祖先から遺伝子を受け継いでいる。
10世代遡ると、重複が無ければ、その世代に祖先は$2^{10}$人。
父系（Y染色体）の伝達経路（太線）や母系（ミトコンドリアDNA）の伝達経路（点線）だけでなく、多くの経路を調べることが重要となる。

# Incomplete lineage sorting 不完全な系統仕分け




遺伝子と集団の樹形は必ずしも一致しない

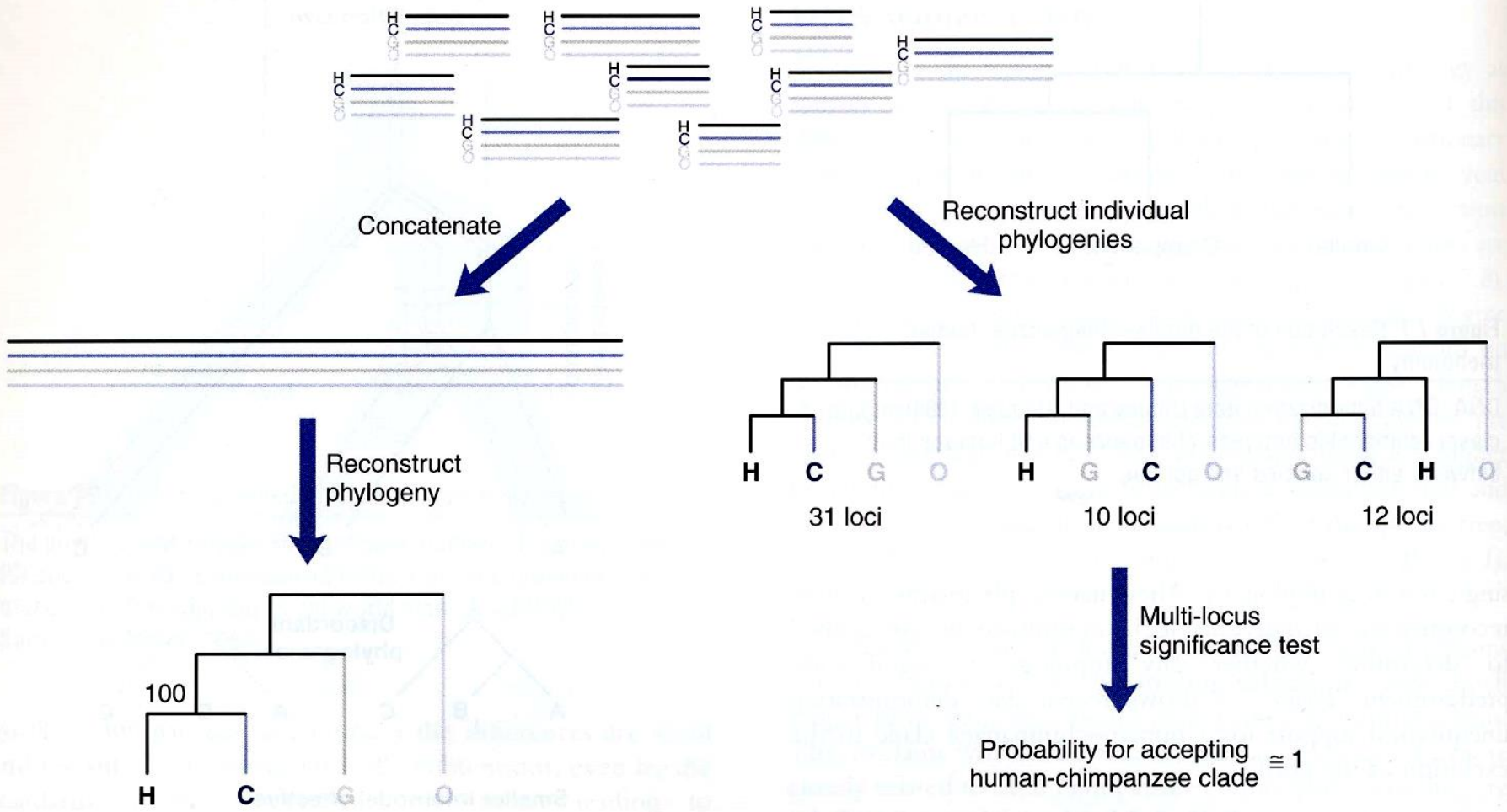
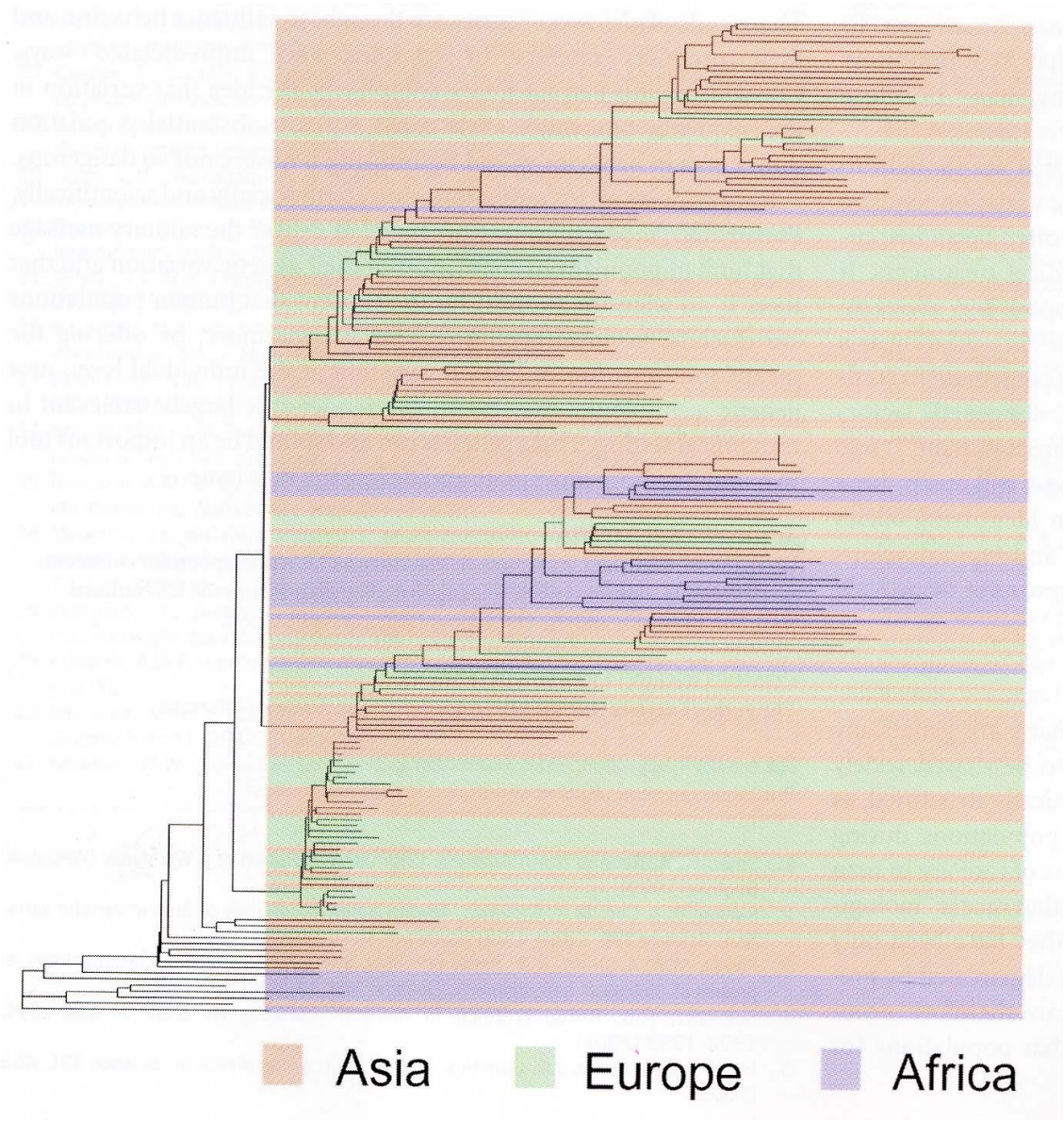



**Figure 7.9:** Combining data from multiple orthologous loci to resolve the gorilla–chimpanzee–human trichotomy.

The number (100) next to the human–chimpanzee clade indicates percentage bootstrap support. H, Human; C, chimpanzee; G, gorilla; O, orangutan. Data are from Chen and Li (2001).

# 常染色体の1つの遺伝子領域による個体間の系統樹



# 複数の遺伝子領域（190遺伝子座）による個体間の系統樹

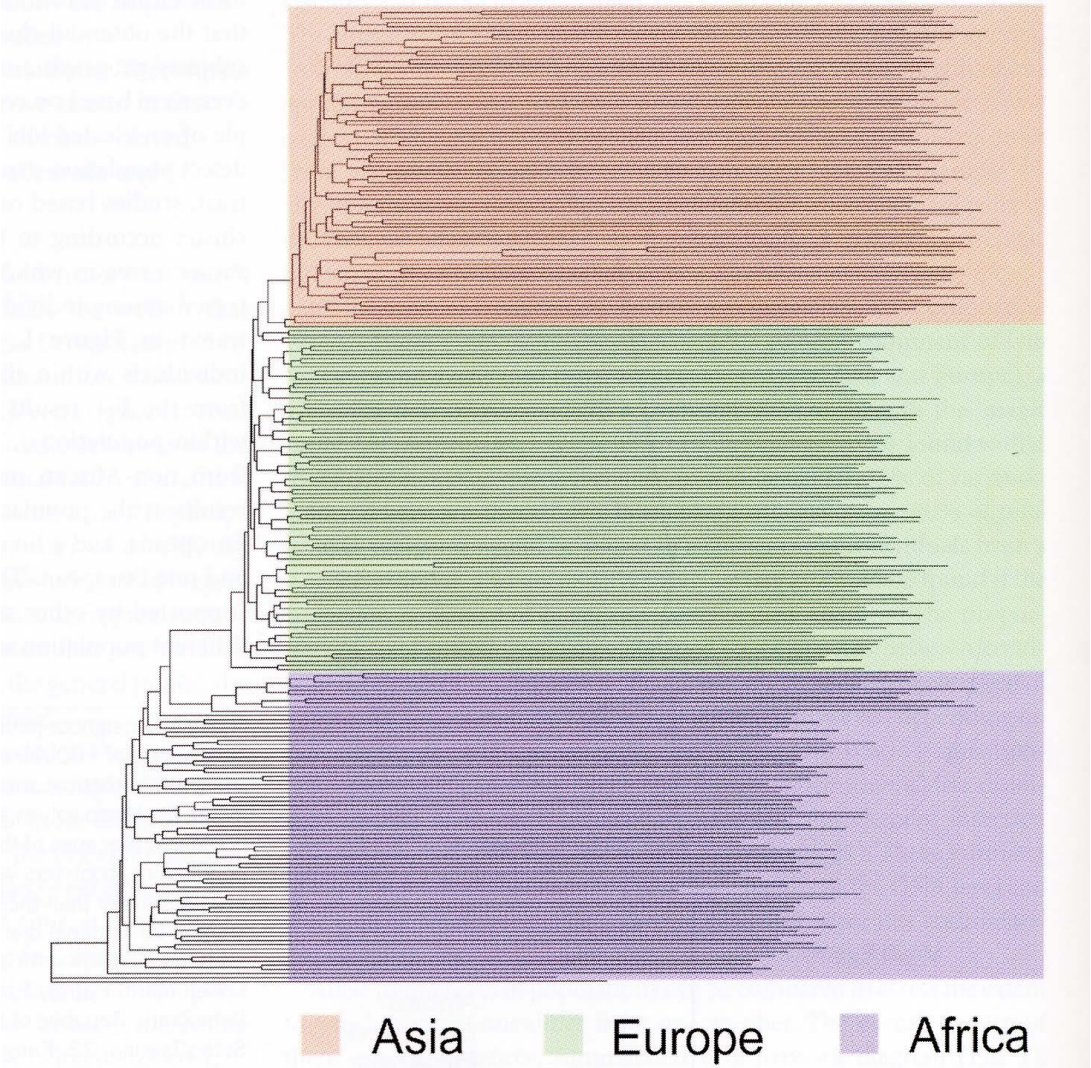



**Figure 2** A neighbor-joining tree of individual similarities, based on 60 STR polymorphisms[72], 100 *Alu* insertion polymorphisms[21] and 30 restriction site polymorphisms[73]. The percentage of shared alleles was calculated for all possible pairs of individuals, and a neighbor-joining tree was formulated using the PHYLIP software package[74]. African individuals are shown in blue, European individuals in green and Asian individuals in orange.

# 集団ゲノム解析で用いられる多変量解析

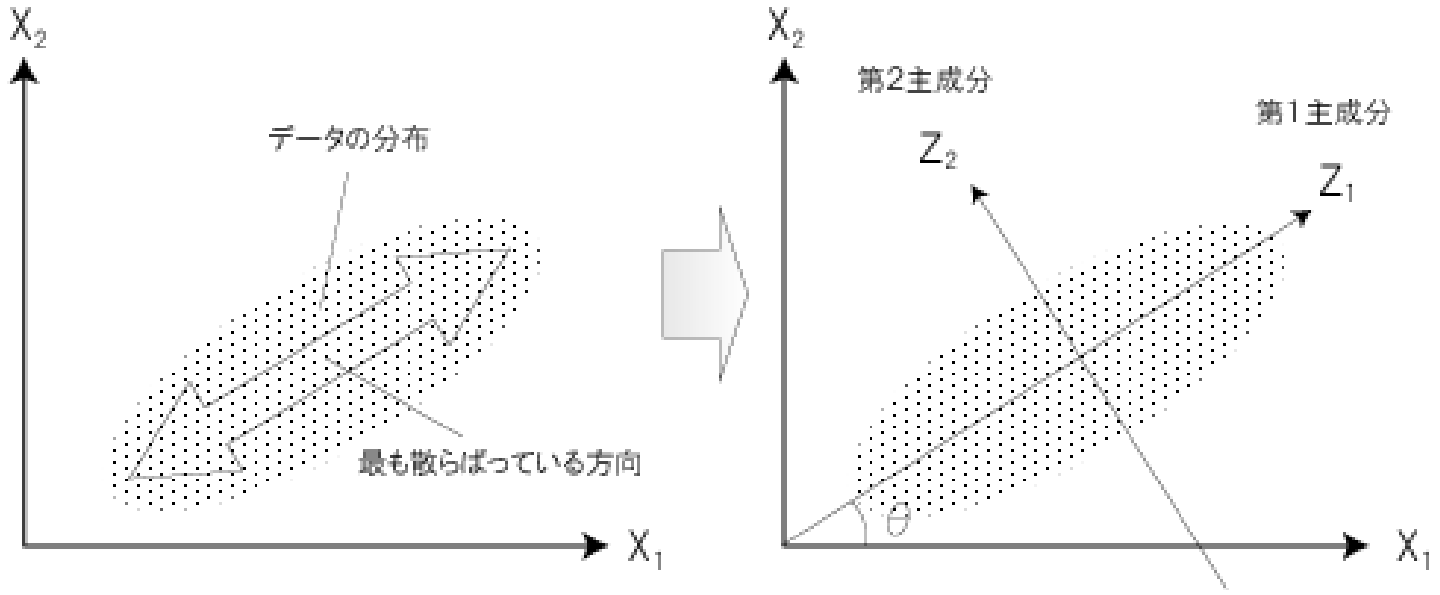
1. 遺伝距離マトリックス Genetic distance matrix
   ⇒系統樹解析 Phylogenetic analysis
   ⇒多次元尺度構成法 Multidimensional scaling (MDS)

2. 主成分分析 Principal component analysis (PCA)

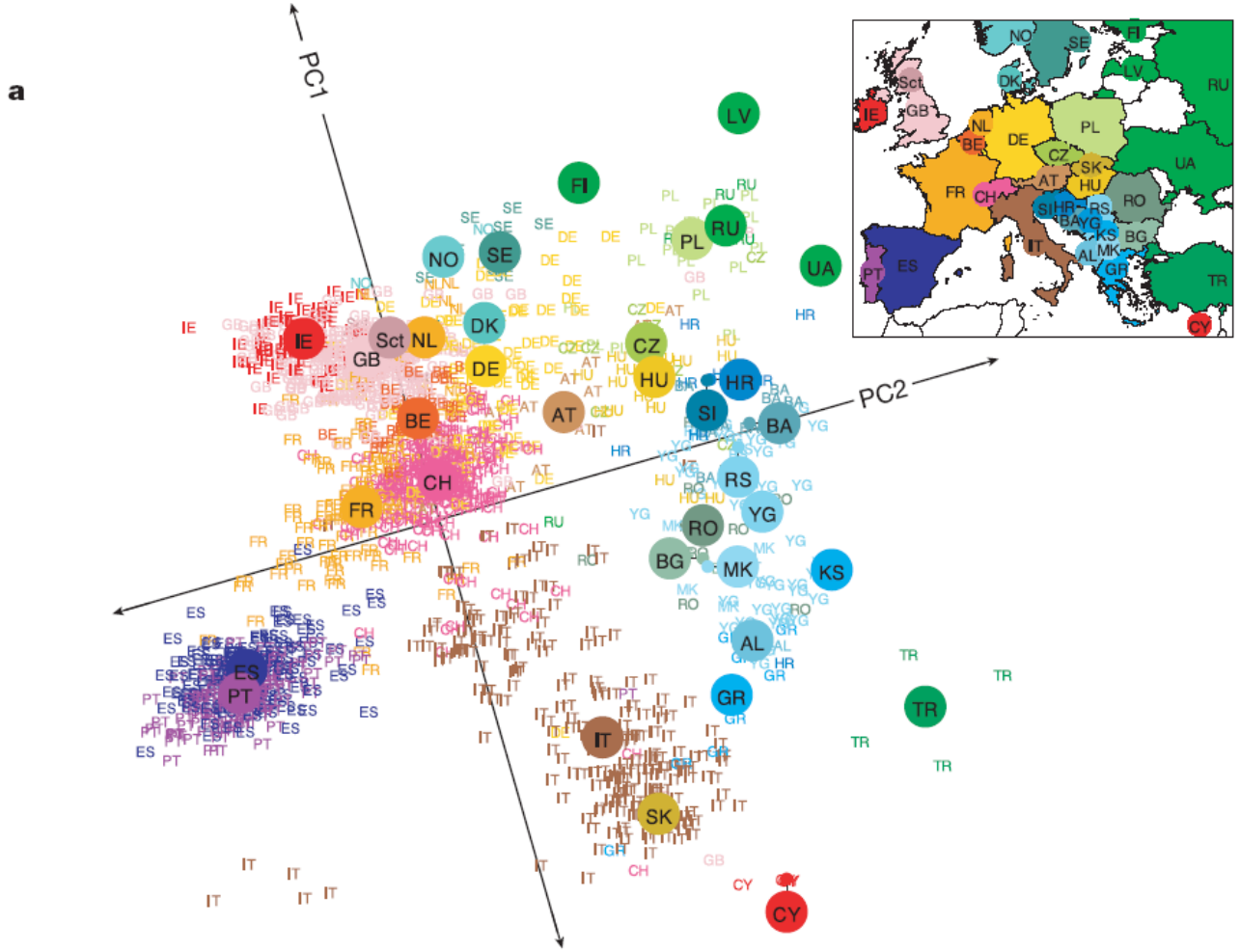
3. クラスタ分析 Clustering analysis
   (STRUCTURE, ADMIXTURE, FRAPPE)

# 主成分分析

- 多数項目（多次元）のデータを、互いに無相関な主成分として低次元に縮約する統計手法。
- ばらつきの大きい方向から第一主成分が生成され、最終的には項目数（あるいはサンプル数）と同じだけ主成分が生成される。
- 上位の主成分のみを扱うことで、データが低次元になる。



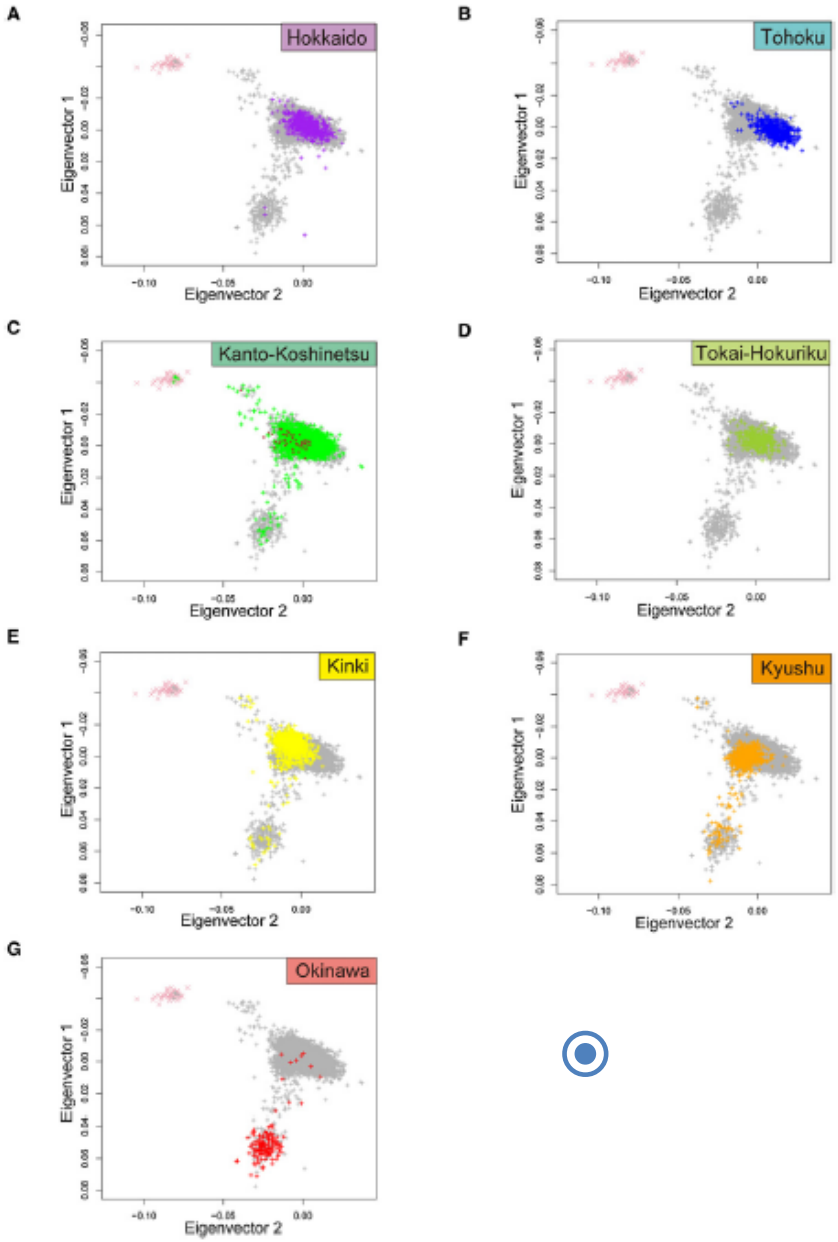
# ヨーロッパ人のゲノム構造：地理的分布と酷似

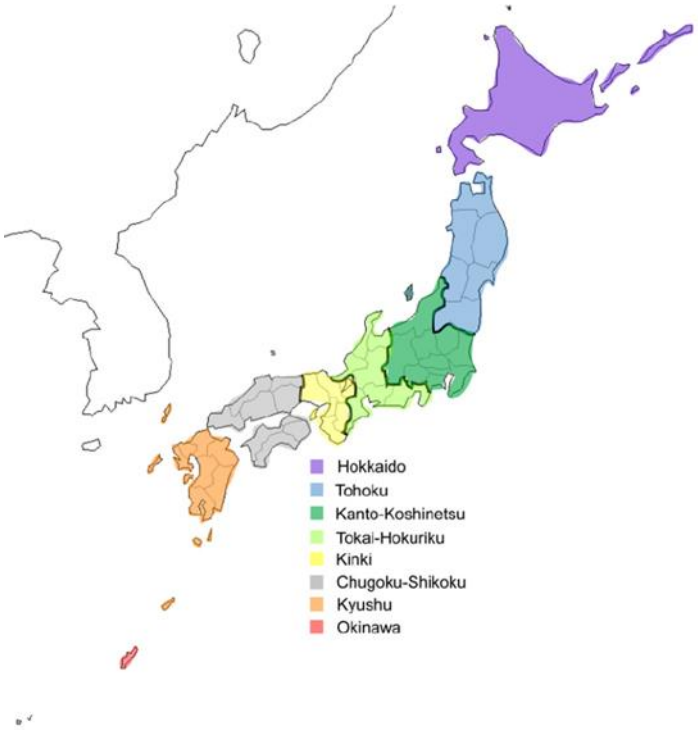



500K SNPによる主成分分析

Novembre et al. 2008

# ゲノムデータが示す日本人の二重構造




約14万SNP

Yamaguchi-Kabata et al. (2008)

# STRUCTURE解析

類似のクラスタ分析法
FRAPPE, ADMIXTURE

**Inference of Population Structure Using Multilocus Genotype Data**

Jonathan K. Pritchard, Matthew Stephens and Peter Donnelly

*Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom*

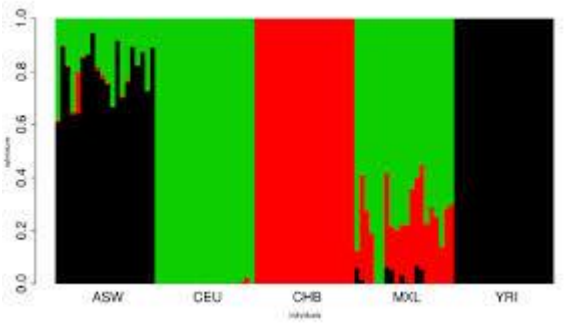
Manuscript received September 23, 1999
Accepted for publication February 18, 2000

**MCMC algorithm (with admixture):** The following algorithm may be used to sample from $\Pr(Z, P, Q|X)$.

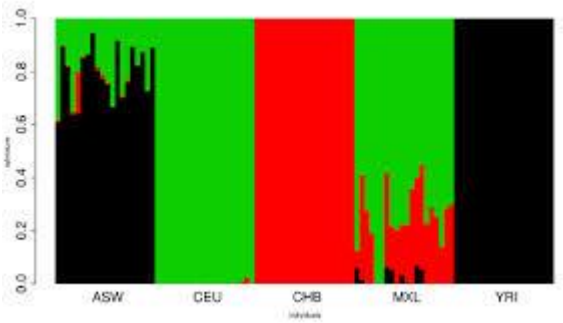
ALGORITHM 2: *Starting with initial values* $Z^{(0)}$ *for* $Z$ *(by drawing* $Z^{(0)}$ *at random using* (3) *for example), iterate the following steps for* $m = 1, 2, \ldots$.

*Step* 1. *Sample* $P^{(m)}$, $Q^{(m)}$ *from* $\Pr(P, Q|X, Z^{(m-1)})$.
*Step* 2. *Sample* $Z^{(m)}$ *from* $\Pr(Z|X, P^{(m)}, Q^{(m)})$.
*Step* 3. *Update* $\alpha$ *using a Metropolis-Hastings step.*

Z: populations of origin of the individuals
P: allele frequencies in all populations
Q: admixture proportions for each individuals
X: observed genotypes




m-1




m

# Human Genome Diversity Panel (HGDP)の650K SNP解析

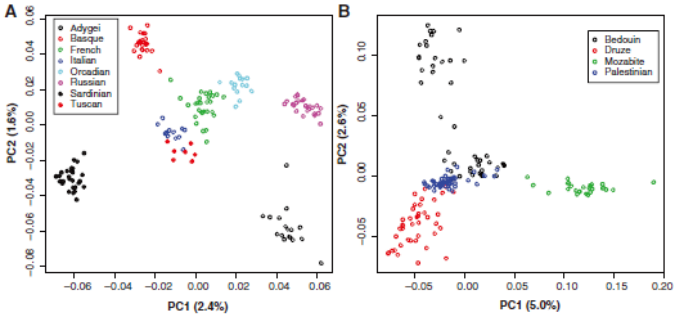



Fig. 2. Fine-scale population structure principal component analyses in two geographic regions, using all autosomal SNPs. (A) Europe. (B) The Middle East.

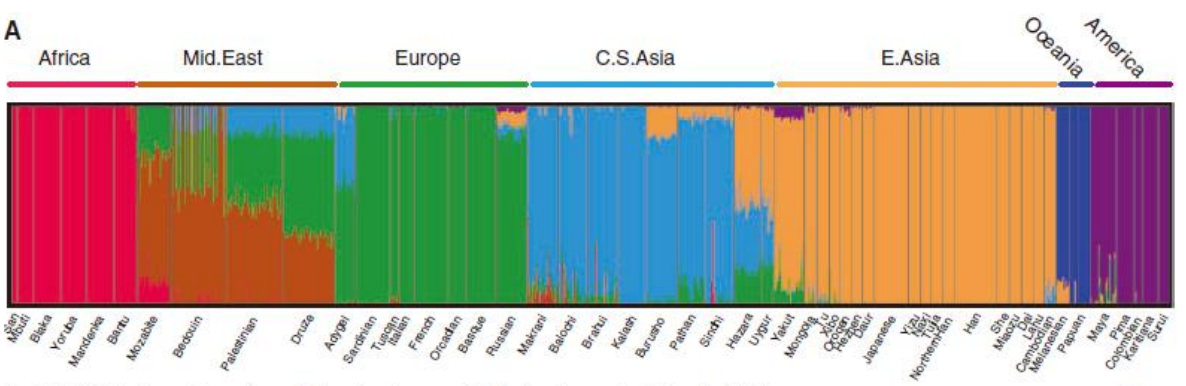



Fig. 1. Individual ancestry and population dendrogram. (A) Regional ancestry inferred with the *Frappe* program at $K = 7$ (*13*) and plotted with the Distruct program (*31*). Each individual is represented by a vertical line partitioned into colored segments whose lengths correspond to his/her ancestry coefficients in up to seven inferred ancestral groups. Population labels were added only after each individual's ancestry had been estimated; they were used to order the samples in plotting. (B) Maximum likelihood tree of 51 populations. Branches are colored according to continents/regions. * indicates the root of the tree, also where the chimpanzee branch is located.

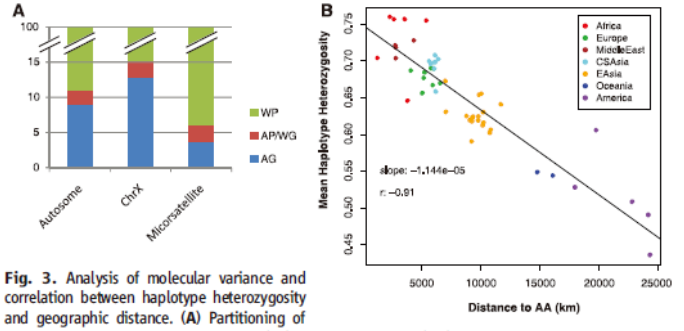
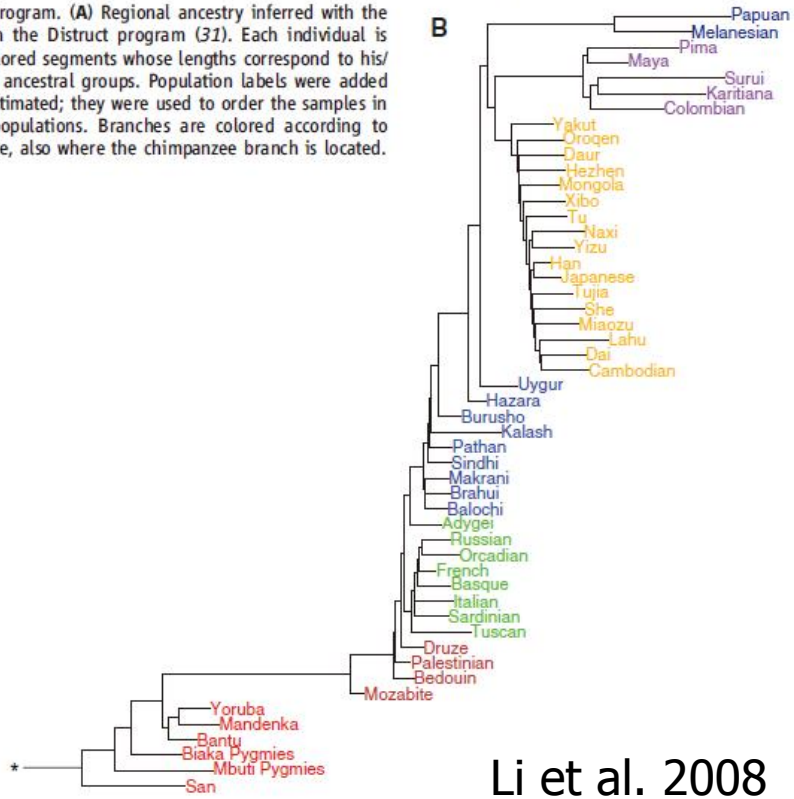



Fig. 3. Analysis of molecular variance and correlation between haplotype heterozygosity and geographic distance. (A) Partitioning of genetic variance into three components (*18*): Within-Population (WP), Among-Population-Within-Region (AP/WG), and Among-Region (AG), by using autosomal SNPs, microsatellite markers, and ChrX SNPs, respectively. (B) SNP haplotype heterozygosity versus geographic distances from Addis Ababa (AA), Ethiopia. The linear regression slope is indicated along with the Pearson correlation *r*.




# Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation


Jun Z. Li,[1,2]*† Devin M. Absher,[1,2]* Hua Tang,[1] Audrey M. Southwick,[1,2] Amanda M. Casto,[1] Sohini Ramachandran,[4] Howard M. Cann,[5] Gregory S. Barsh,[1,3] Marcus Feldman,[4]‡ Luigi L. Cavalli-Sforza,[1]‡ Richard M. Myers[1,2]‡


Li et al. 2008

# 50K SNPを用いたアジア諸国の遺伝的多様性

## Mapping Human Genetic Diversity in Asia

The HUGO Pan-Asian SNP Consortium*†

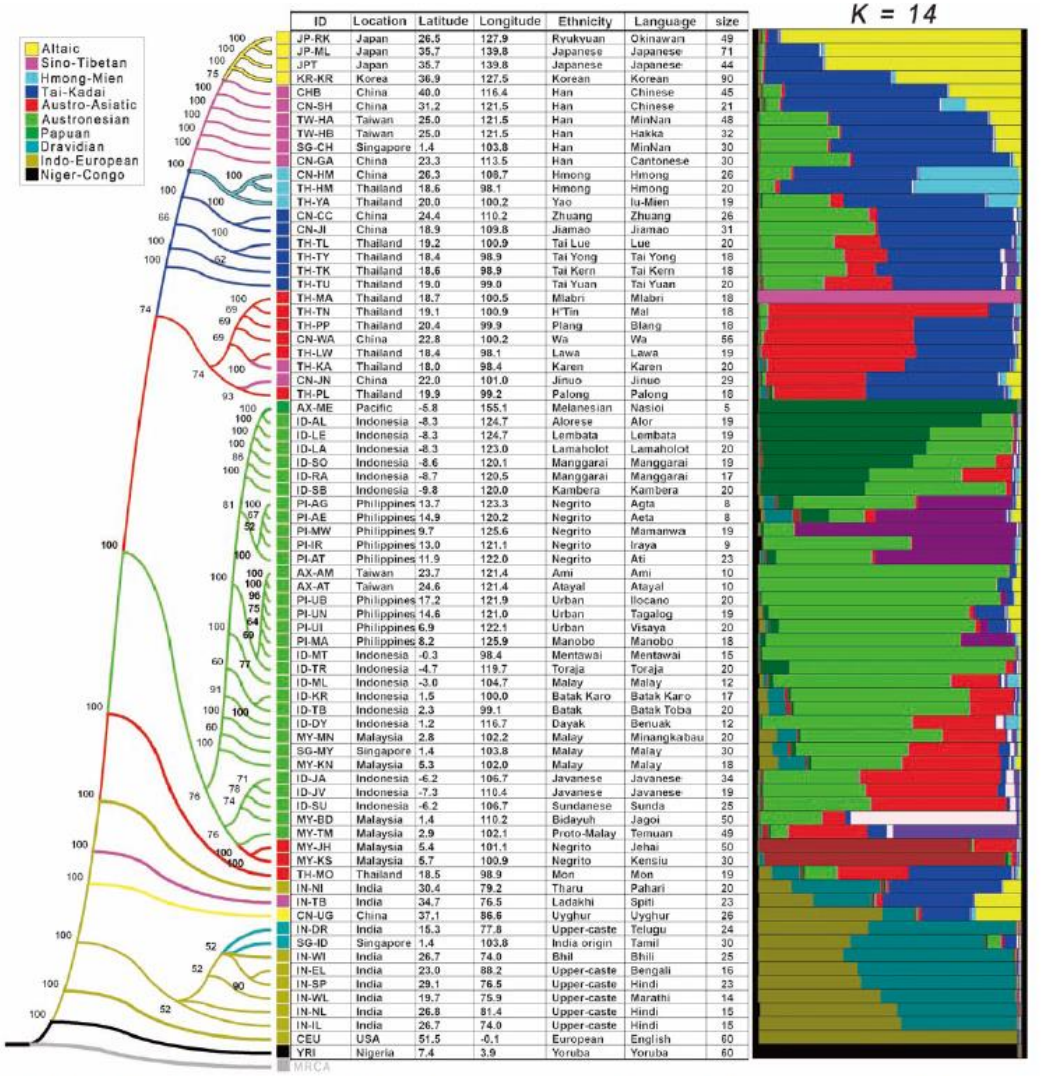
Asia harbors substantial cultural and linguistic diversity, but the geographic structure of genetic variation across the continent remains enigmatic. Here we report a large-scale survey of autosomal variation from a broad geographic sample of Asian human populations. Our results show that genetic ancestry is strongly correlated with linguistic affiliations as well as geography. Most populations show relatedness within ethnic/linguistic groups, despite prevalent gene flow among populations. More than 90% of East Asian (EA) haplotypes could be found in either Southeast Asian (SEA) or Central-South Asian (CSA) populations and show clinal structure with haplotype diversity decreasing from south to north. Furthermore, 50% of EA haplotypes were found in SEA only and 5% were found in CSA only, indicating that SEA was a major geographic source of EA populations.

HUGO Pan-Asian SNP Consortium　2009

- China
- Indonesia
- India
- Japan
- Korea
- Malaysia
- Philippines
- Singapore
- Thailand
- Taiwan

遺伝距離

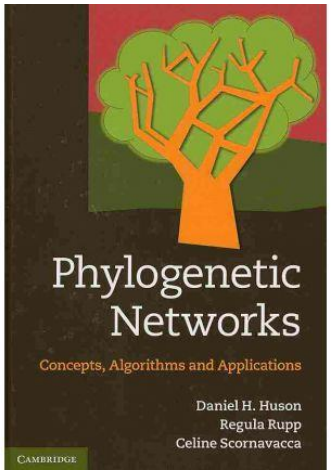
STRUCTURE解析

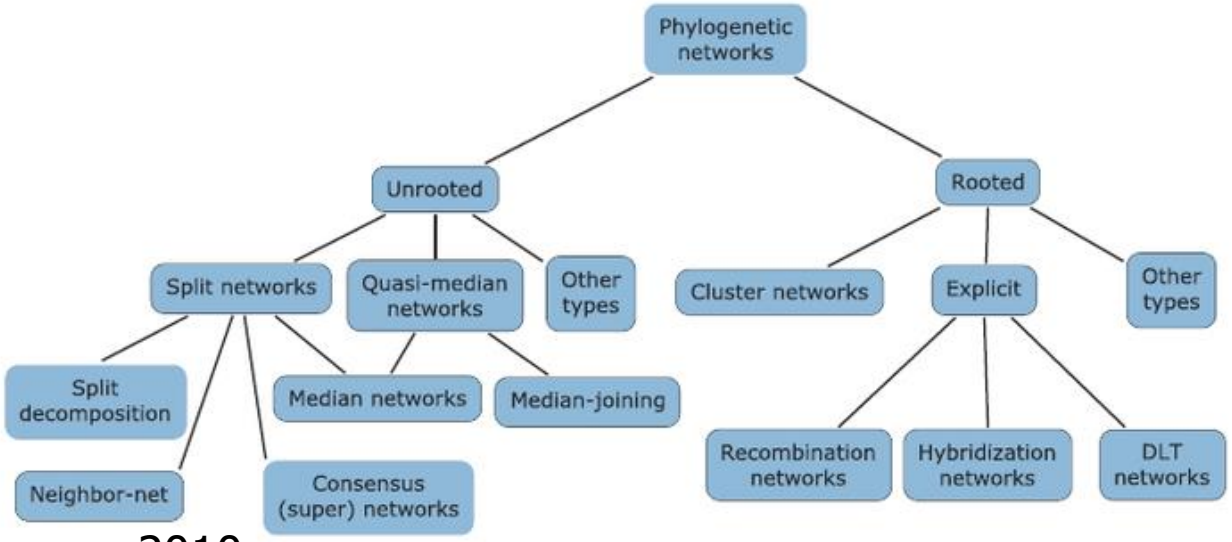



| ID | Location | Latitude | Longitude | Ethnicity | Language | size |
|---|---|---|---|---|---|---|
| JP-RK | Japan | 26.5 | 127.9 | Ryukyuan | Okinawan | 49 |
| JP-ML | Japan | 35.7 | 139.8 | Japanese | Japanese | 71 |
| JPT | Japan | 35.7 | 139.8 | Japanese | Japanese | 44 |
| KR-KR | Korea | 36.9 | 127.5 | Korean | Korean | 90 |
| CHB | China | 40.0 | 116.4 | Han | Chinese | 45 |
| CN-SH | China | 31.2 | 121.5 | Han | Chinese | 21 |
| TW-HA | Taiwan | 25.0 | 121.5 | Han | MinNan | 48 |
| TW-HB | Taiwan | 25.0 | 121.5 | Han | Hakka | 32 |
| SG-CH | Singapore | 1.4 | 103.8 | Han | MinNan | 30 |
| CN-GA | China | 23.3 | 113.5 | Han | Cantonese | 30 |
| CN-HM | China | 26.3 | 108.7 | Hmong | Hmong | 26 |
| TH-HM | Thailand | 18.6 | 98.1 | Hmong | Hmong | 20 |
| TH-YA | Thailand | 20.0 | 100.2 | Yao | Iu-Mien | 19 |
| CN-CC | China | 24.4 | 110.2 | Zhuang | Zhuang | 26 |
| CN-JI | China | 18.9 | 109.8 | Jiamao | Jiamao | 31 |
| TH-TL | Thailand | 19.2 | 100.9 | Tai Lue | Lue | 20 |
| TH-TY | Thailand | 18.4 | 98.9 | Tai Yong | Tai Yong | 18 |
| TH-TK | Thailand | 18.6 | 98.9 | Tai Kern | Tai Kern | 18 |
| TH-TU | Thailand | 19.0 | 99.0 | Tai Yuan | Tai Yuan | 20 |
| TH-MA | Thailand | 18.7 | 100.5 | Mlabri | Mlabri | 18 |
| TH-TN | Thailand | 19.1 | 100.9 | H'Tin | Mal | 18 |
| TH-PP | Thailand | 20.4 | 99.9 | Plang | Blang | 18 |
| CN-WA | China | 22.8 | 100.2 | Wa | Wa | 56 |
| TH-LW | Thailand | 18.4 | 98.1 | Lawa | Lawa | 19 |
| TH-KA | Thailand | 18.0 | 98.4 | Karen | Karen | 20 |
| CN-JN | China | 22.0 | 101.0 | Jinuo | Jinuo | 29 |
| TH-PL | Thailand | 19.9 | 99.2 | Palong | Palong | 18 |
| AX-ME | Pacific | -5.8 | 155.1 | Melanesian | Nasioi | 5 |
| ID-AL | Indonesia | -8.3 | 124.7 | Alorese | Alor | 19 |
| ID-LE | Indonesia | -8.3 | 124.7 | Lembata | Lembata | 19 |
| ID-LA | Indonesia | -8.3 | 123.0 | Lamaholot | Lamaholot | 20 |
| ID-SO | Indonesia | -8.6 | 120.1 | Manggarai | Manggarai | 19 |
| ID-RA | Indonesia | -8.7 | 120.5 | Manggarai | Manggarai | 17 |
| ID-SB | Indonesia | -9.8 | 120.0 | Kambera | Kambera | 20 |
| PI-AG | Philippines | 13.7 | 123.3 | Negrito | Agta | 8 |
| PI-AE | Philippines | 14.9 | 120.2 | Negrito | Aeta | 8 |
| PI-MW | Philippines | 9.7 | 125.6 | Negrito | Mamanwa | 19 |
| PI-IR | Philippines | 13.0 | 121.1 | Negrito | Iraya | 9 |
| PI-AT | Philippines | 11.9 | 122.0 | Negrito | Ati | 23 |
| AX-AM | Taiwan | 23.7 | 121.4 | Ami | Ami | 10 |
| AX-AT | Taiwan | 24.6 | 121.4 | Atayal | Atayal | 10 |
| PI-UB | Philippines | 17.2 | 121.9 | Urban | Ilocano | 20 |
| PI-UN | Philippines | 14.6 | 121.0 | Urban | Tagalog | 19 |
| PI-UI | Philippines | 6.9 | 122.1 | Urban | Visaya | 20 |
| PI-MA | Philippines | 8.2 | 125.9 | Manobo | Manobo | 18 |
| ID-MT | Indonesia | -0.3 | 98.4 | Mentawai | Mentawai | 15 |
| ID-TR | Indonesia | -4.7 | 119.7 | Toraja | Toraja | 20 |
| ID-ML | Indonesia | -3.0 | 104.7 | Malay | Malay | 12 |
| ID-KR | Indonesia | 1.5 | 100.0 | Batak Karo | Batak Karo | 17 |
| ID-TB | Indonesia | 2.3 | 99.1 | Batak | Batak Toba | 20 |
| ID-DY | Indonesia | 1.2 | 116.7 | Dayak | Benuak | 12 |
| MY-MN | Malaysia | 2.8 | 102.2 | Malay | Minangkabau | 20 |
| SG-MY | Singapore | 1.4 | 103.8 | Malay | Malay | 30 |
| MY-KN | Malaysia | 5.3 | 102.0 | Malay | Malay | 18 |
| ID-JA | Indonesia | -6.2 | 106.7 | Javanese | Javanese | 34 |
| ID-JV | Indonesia | -7.3 | 110.4 | Javanese | Javanese | 19 |
| ID-SU | Indonesia | -6.2 | 106.7 | Sundanese | Sunda | 25 |
| MY-BD | Malaysia | 1.4 | 110.2 | Bidayuh | Jagoi | 50 |
| MY-TM | Malaysia | 2.9 | 102.1 | Proto-Malay | Temuan | 49 |
| MY-JH | Malaysia | 5.4 | 101.1 | Negrito | Jehai | 50 |
| MY-KS | Malaysia | 5.7 | 100.9 | Negrito | Kensiu | 30 |
| TH-MO | Thailand | 18.5 | 98.9 | Mon | Mon | 19 |
| IN-NI | India | 30.4 | 79.2 | Tharu | Pahari | 20 |
| IN-TB | India | 34.7 | 76.5 | Ladakhi | Spiti | 23 |
| CN-UG | China | 37.1 | 86.6 | Uyghur | Uyghur | 26 |
| IN-DR | India | 15.3 | 77.8 | Upper-caste | Telugu | 24 |
| SG-ID | Singapore | 1.4 | 103.8 | India origin | Tamil | 30 |
| IN-WI | India | 26.7 | 74.0 | Bhil | Bhili | 25 |
| IN-EL | India | 23.0 | 88.2 | Upper-caste | Bengali | 16 |
| IN-SP | India | 29.1 | 76.5 | Upper-caste | Hindi | 23 |
| IN-WL | India | 19.7 | 75.9 | Upper-caste | Marathi | 14 |
| IN-NL | India | 26.8 | 81.4 | Upper-caste | Hindi | 15 |
| IN-IL | India | 26.7 | 74.0 | Upper-caste | Hindi | 15 |
| CEU | USA | 51.5 | -0.1 | European | English | 60 |
| YRI | Nigeria | 7.4 | 3.9 | Yoruba | Yoruba | 60 |

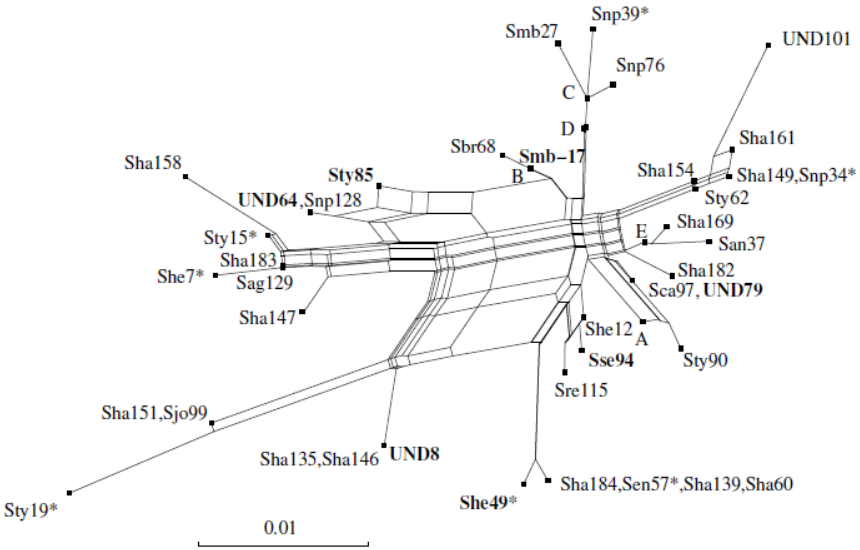
**Fig. 1.** Maximum-likelihood tree of 75 populations. A hypothetical most-recent common ancestor (MRCA) composed of ancestral alleles as inferred from the genotypes of one gorilla and 21 chimpanzees was used to root the tree. Branches with bootstrap values less than 50% were condensed. Population identification numbers (IDs), sample collection locations with latitudes and longitudes, ethnicities, language spoken, and size of population samples are shown in the table adjacent to each branch in the tree. Linguistic groups are indicated with colors as shown in the legend. All population IDs except the four HapMap samples are denoted by four characters. The first two letters indicate the country where the samples were collected or (in the case of Affymetrix) genotyped, according to the following convention: AX, Affymetrix; CN, China; ID, Indonesia; IN, India; JP, Japan; KR, Korea; MY, Malaysia; PI, the Philippines; SG, Singapore; TH, Thailand; and TW, Taiwan. The last two letters are unique IDs for the population. To the right of the table, an averaged graph of results from STRUCTURE is shown for $K = 14$.

# 系統ネットワーク

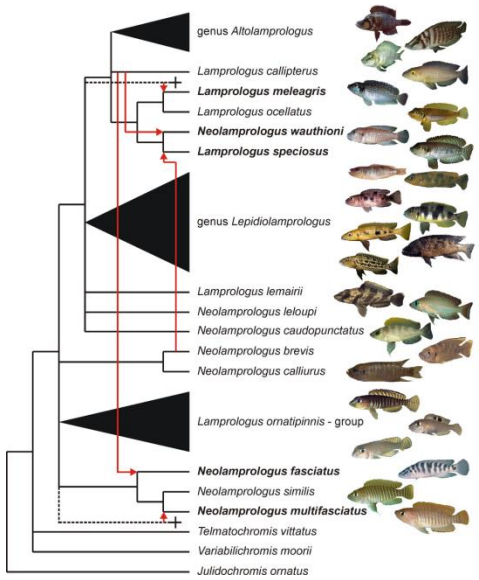



Huson, Rupp, Scornavacca 2010




Neighbor-net
Bryant & Moulton 2003




Hybridization network
Koblmuller et al (2007)

# 系統樹＋混血

OPEN ACCESS Freely available online

PLOS | GENETICS

## Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data

**Joseph K. Pickrell[1]*, Jonathan K. Pritchard[1,2]***

**1** Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America, **2** Howard Hughes Medical Institute, University of Chicago, Chicago, Illinois, United States of America

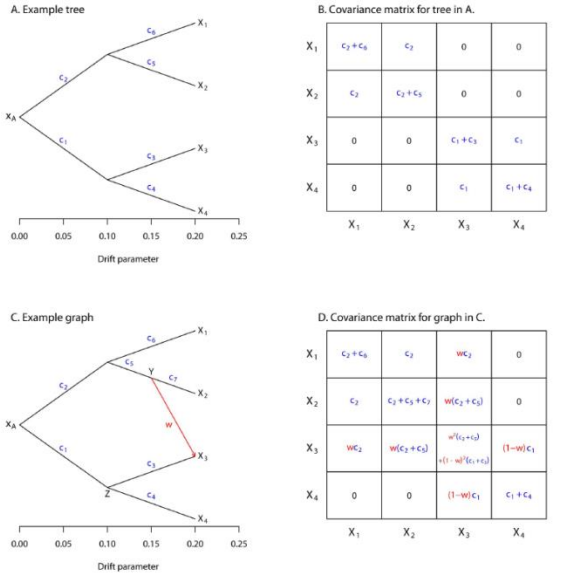



**Figure 1. Simple examples.** A. An example tree. B. The covariance matrix implied by the tree structure in A. Note that the covariance here is with respect to the allele frequency at the root, and that each entry has been divided by $x_A(1-x_A)$ to simplify the presentation. C. An example graph. The migration edge is colored red. Parental populations for population 3 are labeled 'Y' and 'Z'; see the main text for details. D. The covariance matrix implied by the graph in C; again, each entry has been divided by $x_A(1-x_A)$. The migration terms are in red, and the non-migration terms are in blue.
doi:10.1371/journal.pgen.1002967.g001

# TreeMix

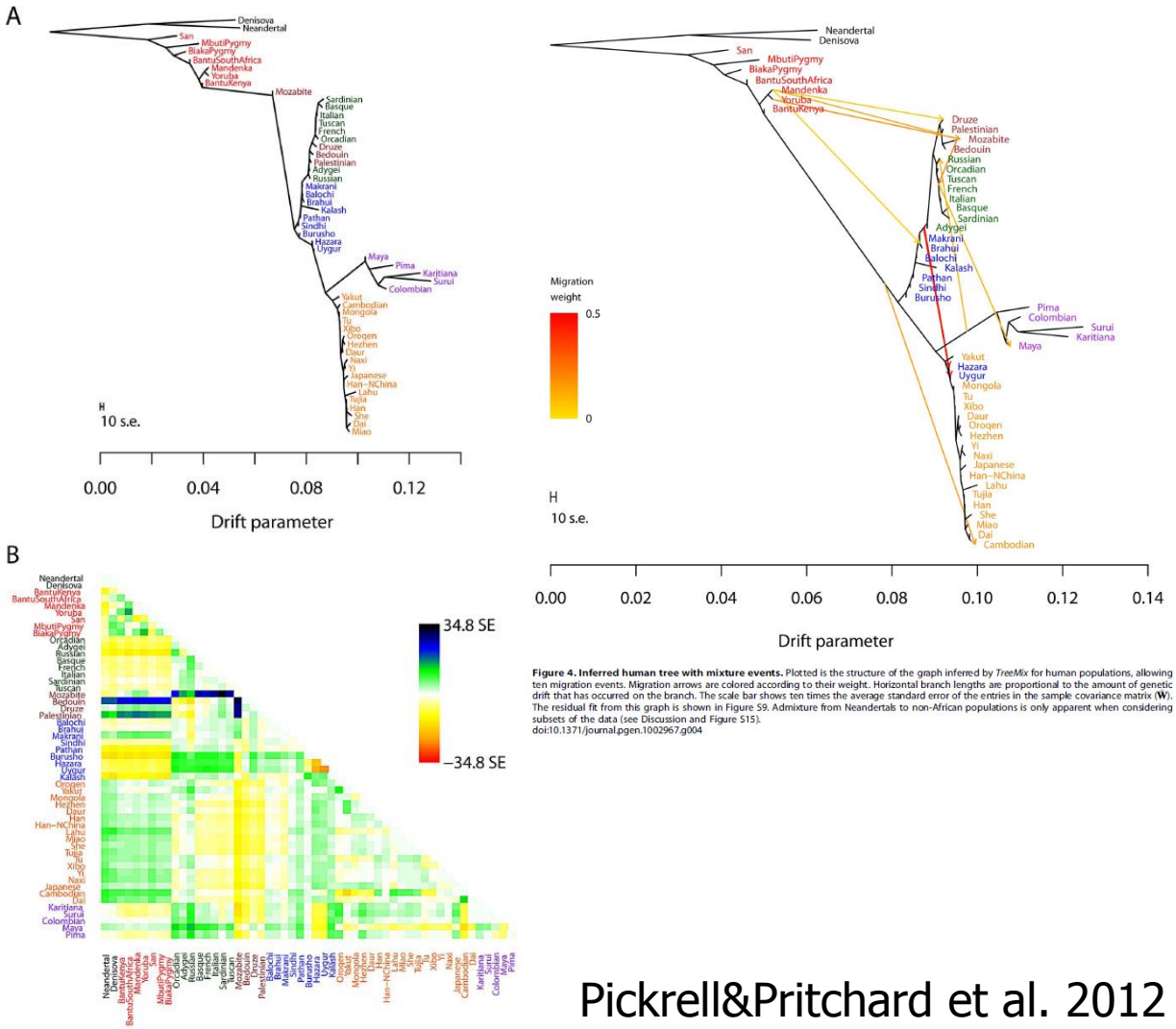
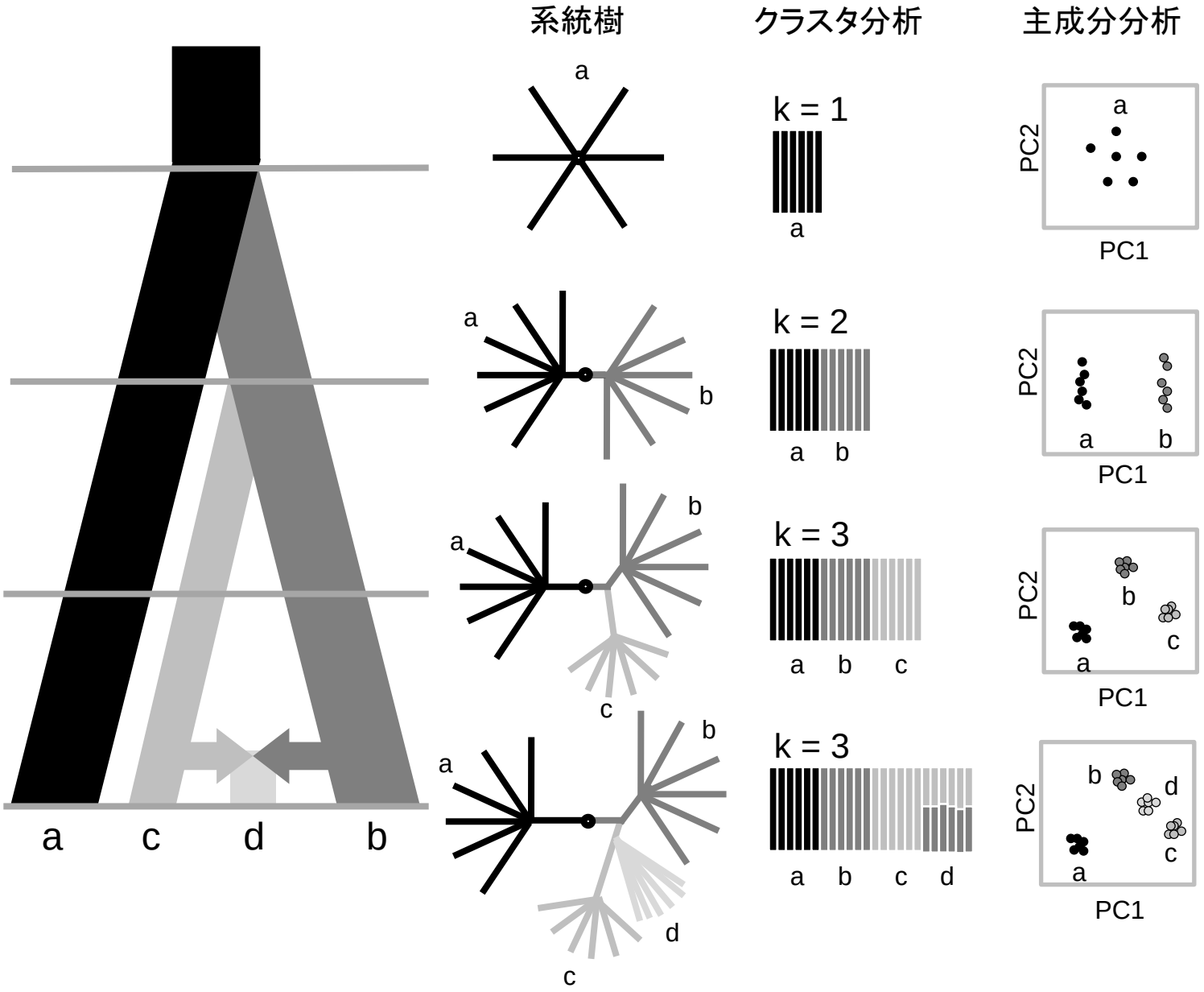


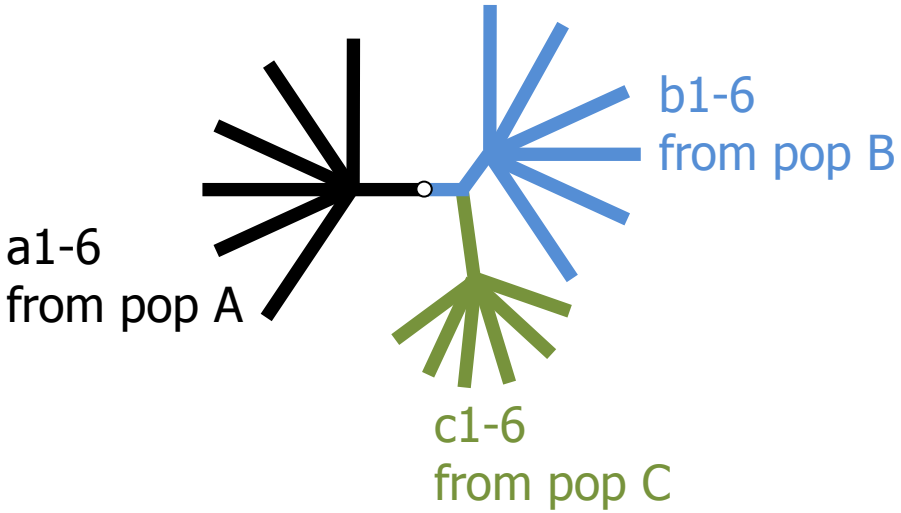

**Figure 4. Inferred human tree with mixture events.** Plotted is the structure of the graph inferred by *TreeMix* for human populations, allowing ten migration events. Migration arrows are colored according to their weight. Horizontal branch lengths are proportional to the amount of genetic drift that has occurred on the branch. The scale bar shows ten times the average standard error of the entries in the sample covariance matrix (**W**). The residual fit from this graph is shown in Figure S9. Admixture from Neandertals to non-African populations is only apparent when considering subsets of the data (see Discussion and Figure S15).
doi:10.1371/journal.pgen.1002967.g004

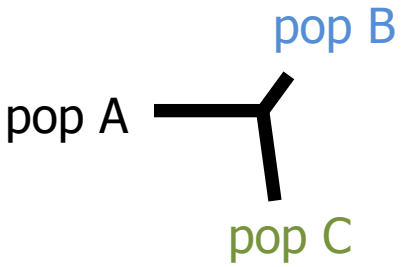
Pickrell&Pritchard et al. 2012

# 集団動態と様々な解析の結果



# 個体間系統樹と集団間系統樹の関係




Tree for individuals
from $D_{ij}$ matrix
($i$ and $j$: individuals)

Tree for populations
from $D_{mXY}$ matrix
($X$ and $Y$: populations)

◆Nei's minimum distance

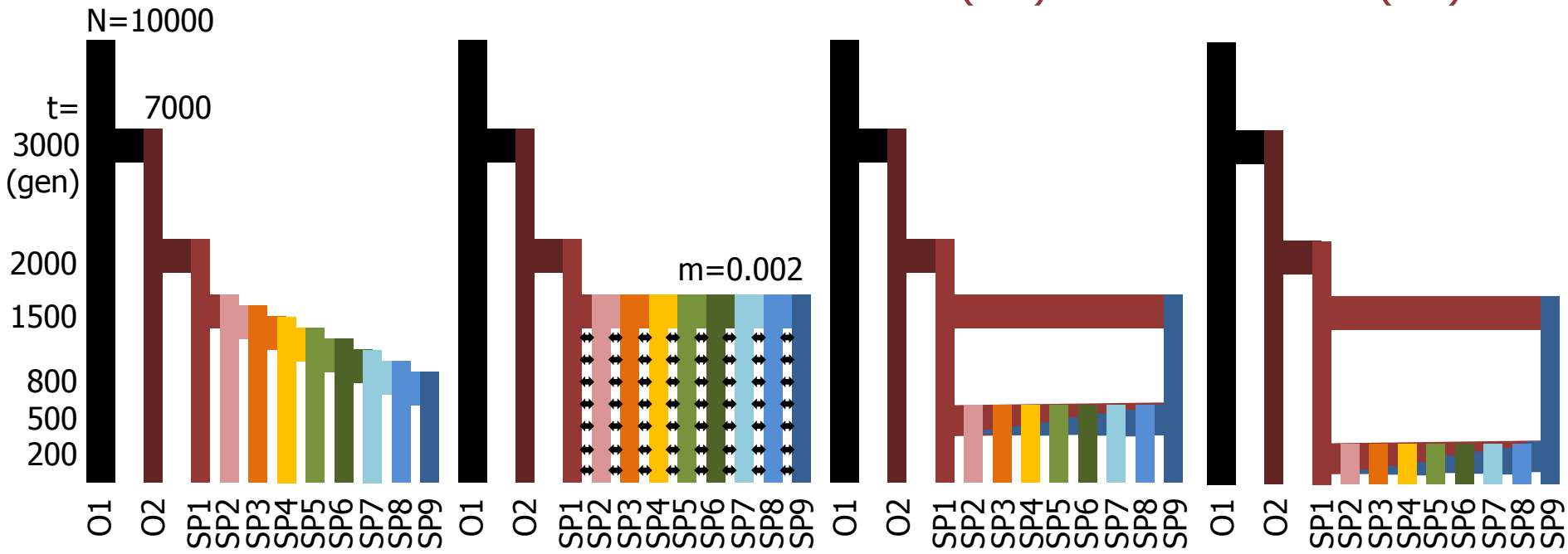
$$D_{mXY} = D_{XY} - \frac{(D_X + D_Y)}{2} = \frac{1}{L}\sum_{i=1}^{L}(p_{Xi} - p_{Yi})^2$$

遺伝距離のゲノム平均などに基づいた集団動態に関するパラメータ推定を
moment-based estimationと呼ぶ

# Simulation study

## Coalescent simulation (*msms*)

- Cascading splits(S)
- Migrations (M)
- Admixtures-Isolation:
  - Old (OA)
  - Recent(RA)

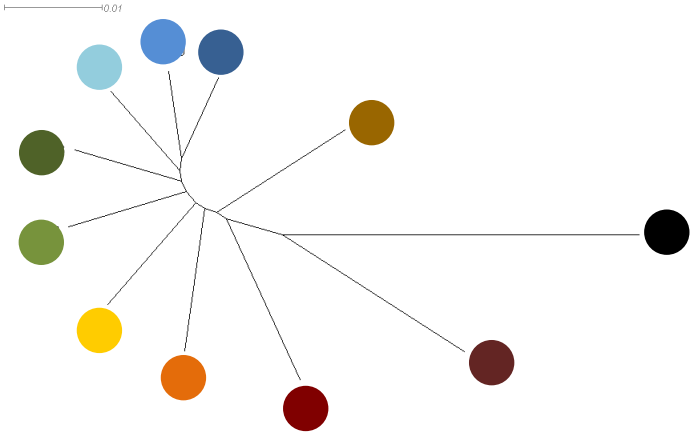


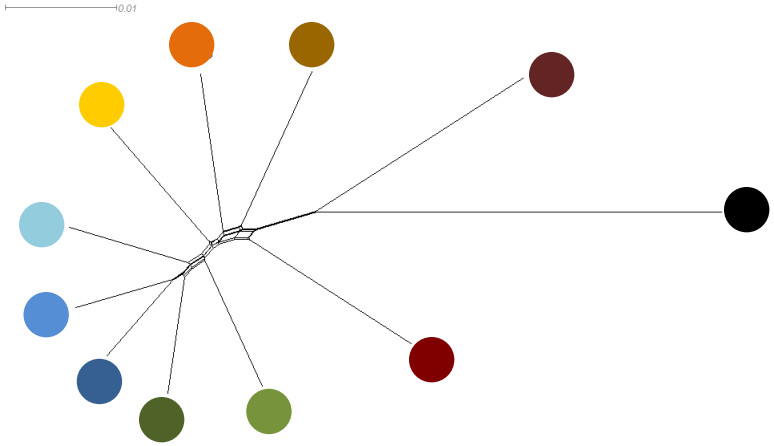

O: outgroup, SP: subpopulation

- 10,000 independent SNPs with MAF>10%
- 10 diploids/population

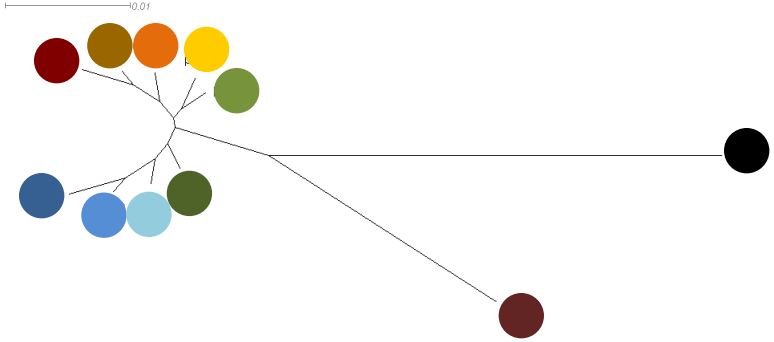
# Phylogenetic analysis (Splits Tree4)

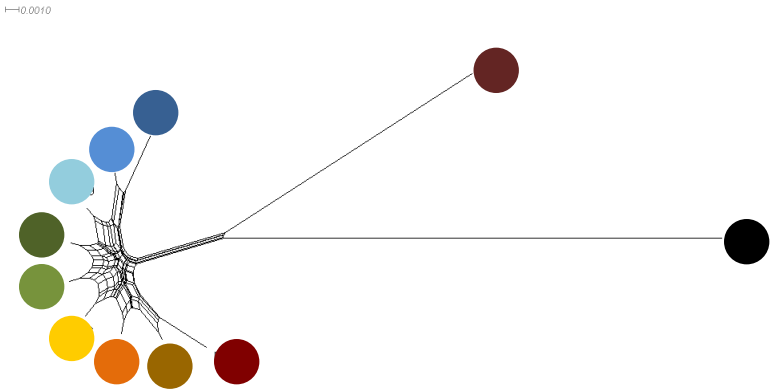
Model S





Model M

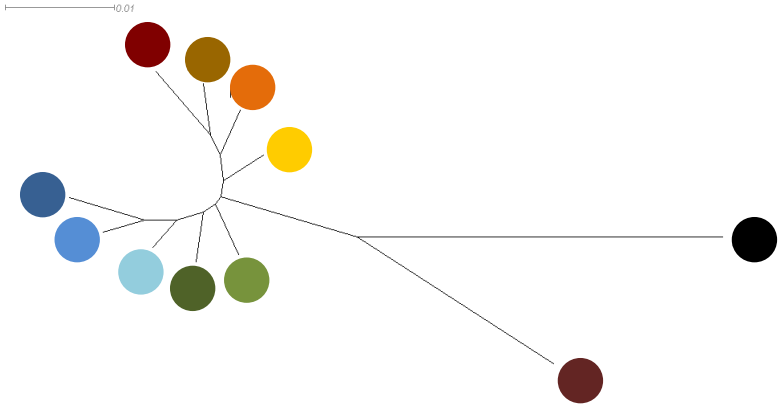


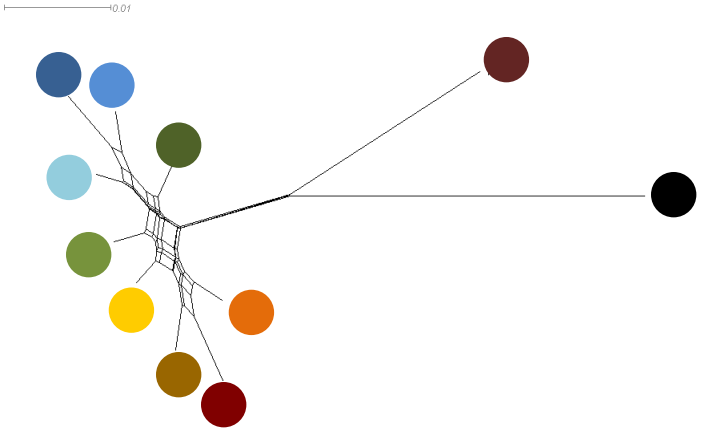


Neighbor-joining tree

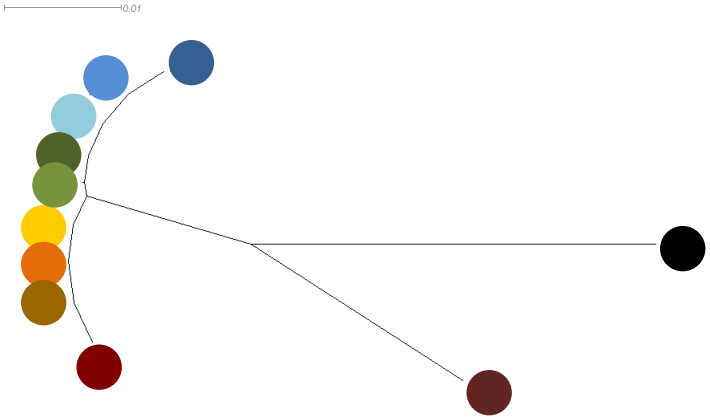
Neighbor net

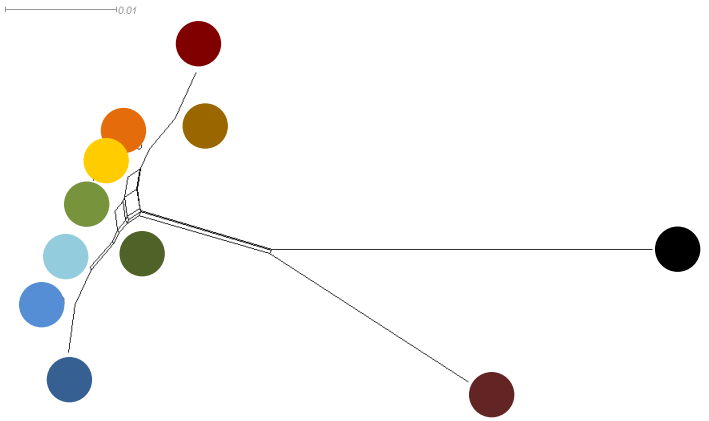
Model OA




Model RA

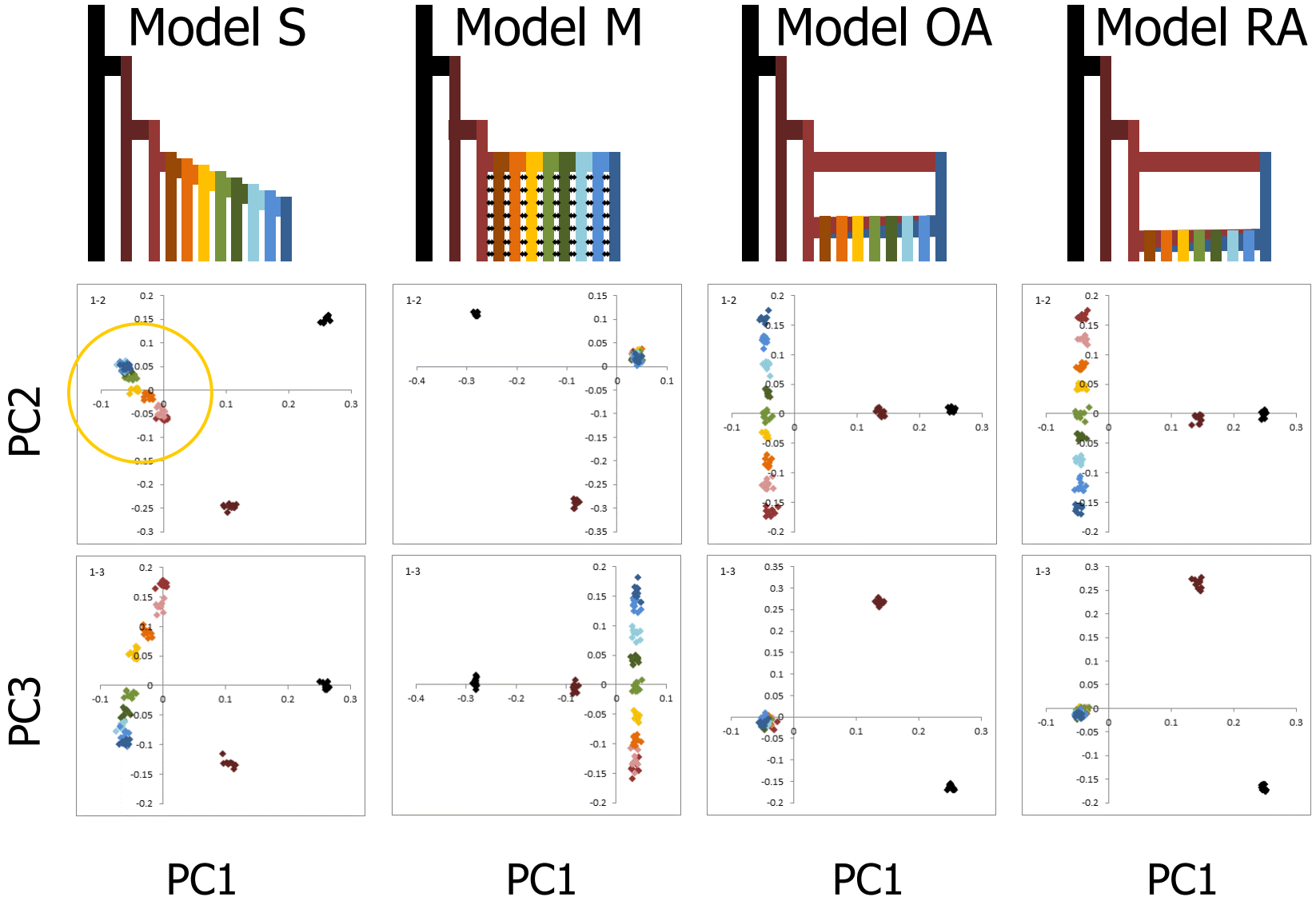


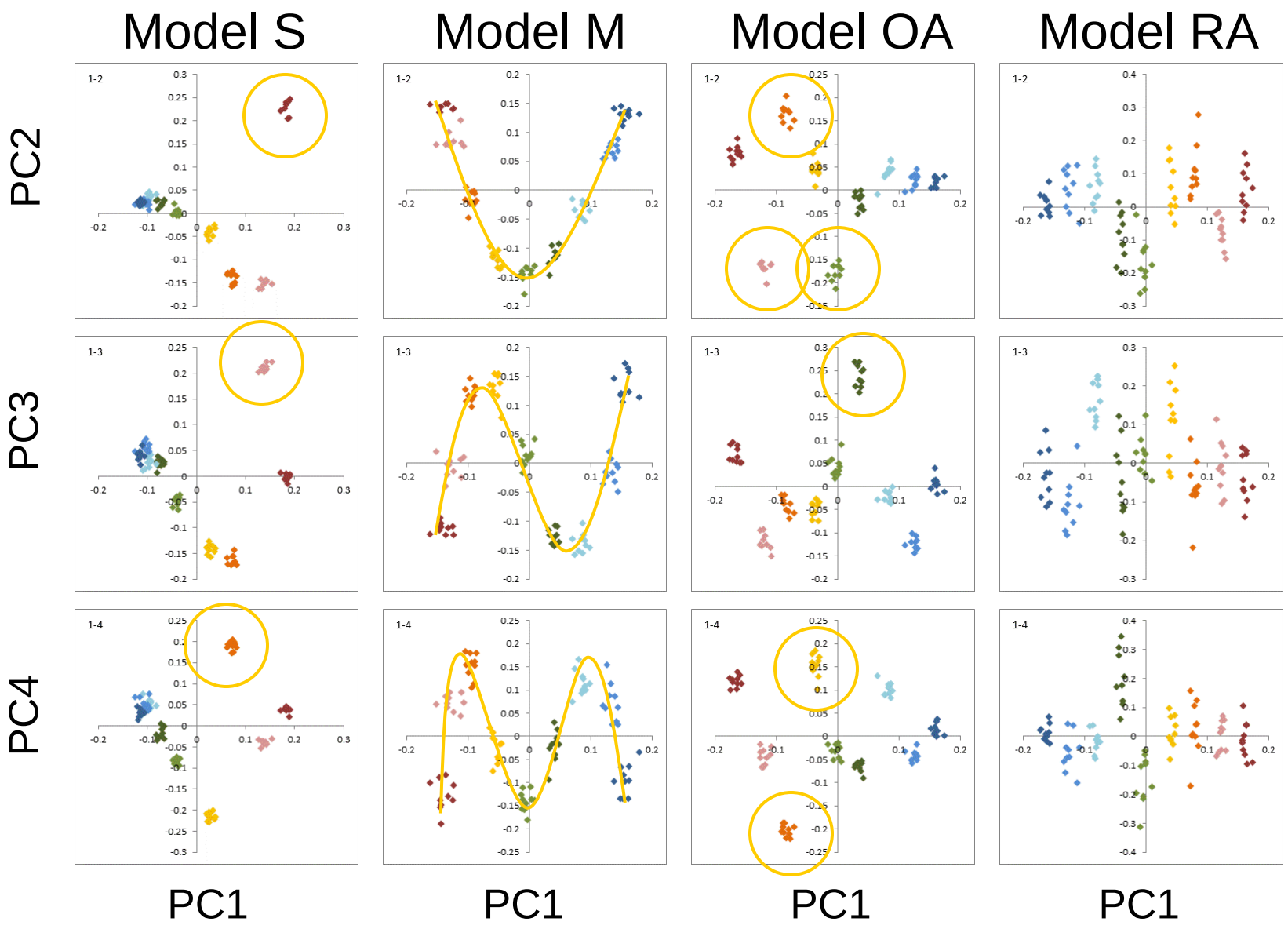

Neighbor-joining tree

Neighbor net

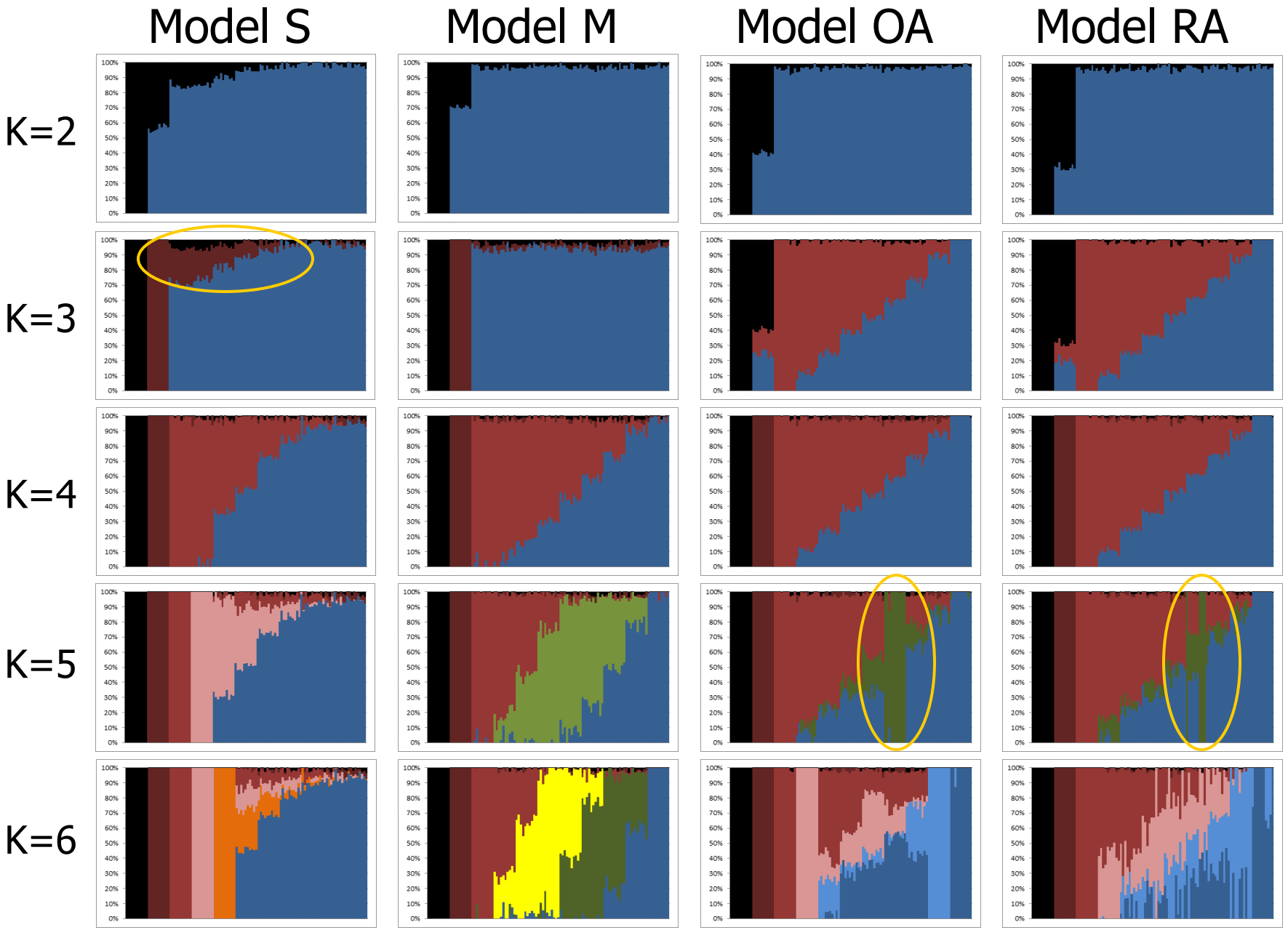
# Principal component analysis (EIGENSTRAT)



# Without outgroups

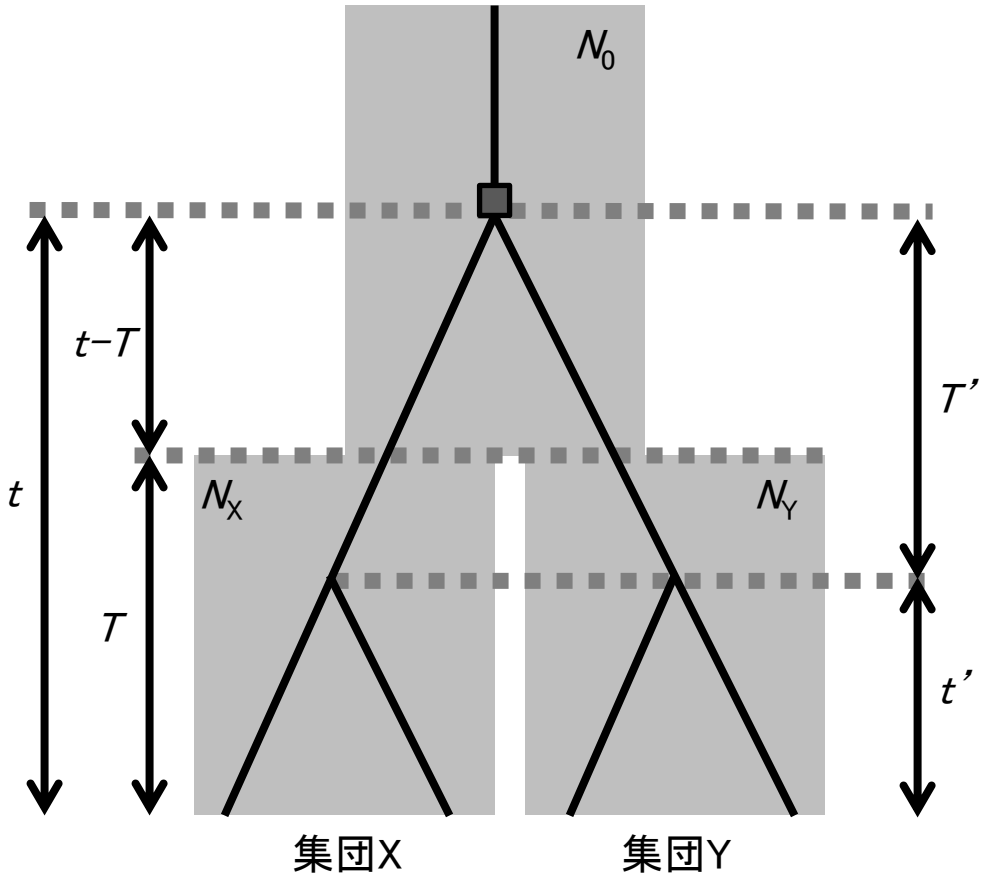


# Clustering method (ADMIXTURE)



# 集団間に関するパラメータを推定しよう

# 集団サイズと分岐年代を推定する
# —集団サイズ一定の場合—




$d_{XY}$： 集団から任意に1つずつ抽出した2つの分子で異なる塩基の割合

$t$： 分子の分岐時間（世代）

$\mu$： 塩基あたり世代あたりの突然変異率

$$E(d_{XY}) = 2\mu t$$
$$\hat{t} = d_{XY}/2\mu$$

$T$： 集団間の分岐時間（世代）

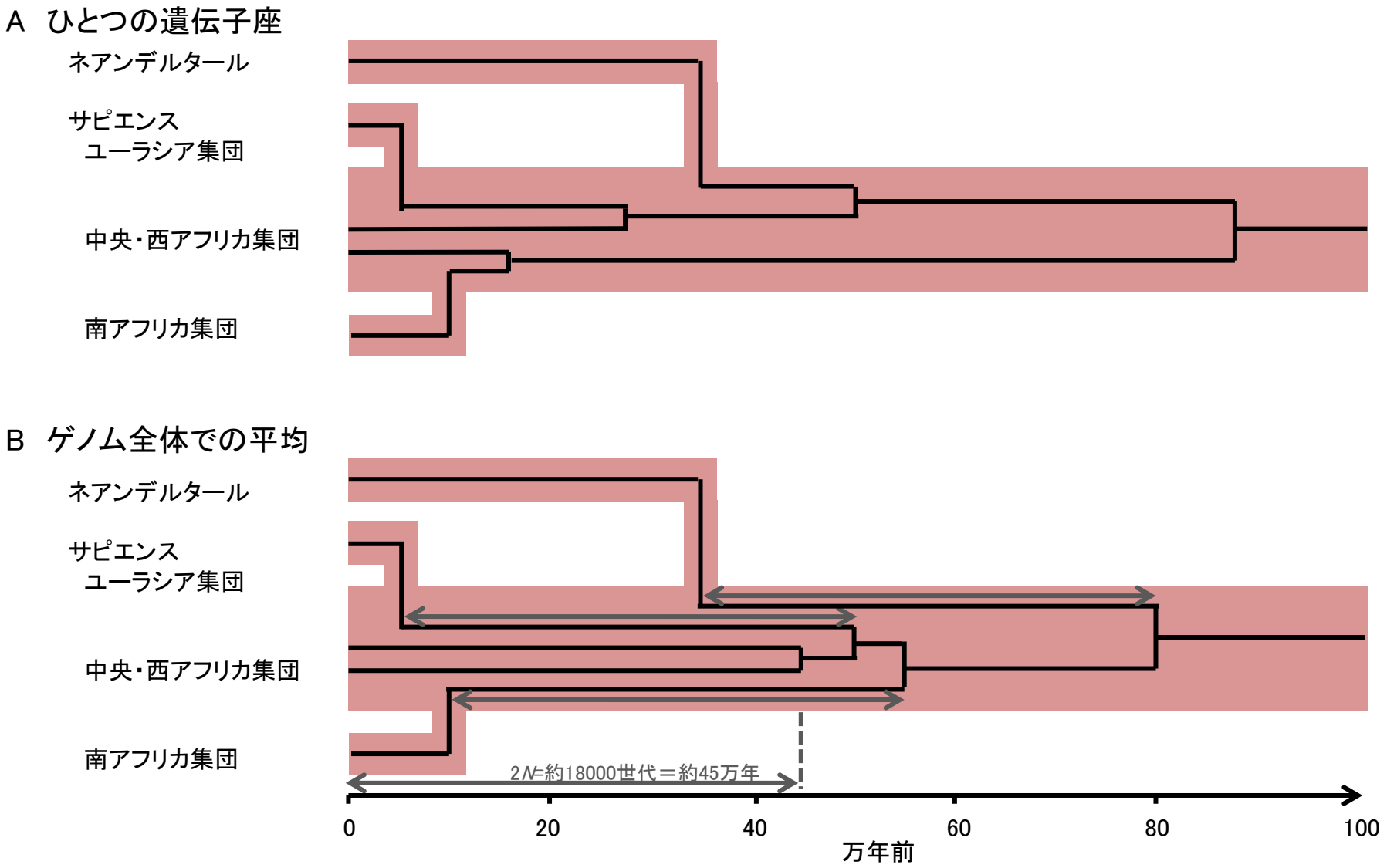
$N_0$： 共通祖先集団サイズ

$$E(d_{XY}) = 2\mu T + 4N_0\mu$$
$$\hat{T} = d_{XY}/2\mu - 2N_0$$

$N_X = N_Y = N_0$を仮定して、
$d = d_{XY} - (d_X + d_Y)/2$とすれば、

$$\hat{T} = d/2\mu$$

# 旧人と新人における遺伝子と集団の系統樹



# Admixture graph

See Reich et al. 2009 & 2012, Paterson 2012

ARTICLES

## Reconstructing Indian population history

David Reich[1,2]*, Kumarasamy Thangaraj[3]*, Nick Patterson[2]*, Alkes L. Price[2,4]* & Lalji Singh[3]

**India has been underrepresented in genome-wide surveys of human variation. We analyse 25 diverse groups in India to provide strong evidence for two ancient populations, genetically divergent, that are ancestral to most Indians today. One, the 'Ancestral North Indians' (ANI), is genetically close to Middle Easterners, Central Asians, and Europeans, whereas the other, the 'Ancestral South Indians' (ASI), is as distinct from ANI and East Asians as they are from each other. By introducing methods that can estimate ancestry without accurate ancestral populations, we show that ANI ancestry ranges from 39–71% in most Indian groups, and is higher in traditionally upper caste and Indo-European speakers. Groups with only ASI ancestry may no longer exist in mainland India. However, the indigenous Andaman Islanders are unique in being ASI-related groups without ANI ancestry. Allele frequency differences between groups in India are larger than in Europe, reflecting strong founder effects whose signatures have been maintained for thousands of years owing to endogamy. We therefore predict that there will be an excess of recessive diseases in India, which should be possible to screen and map genetically.**

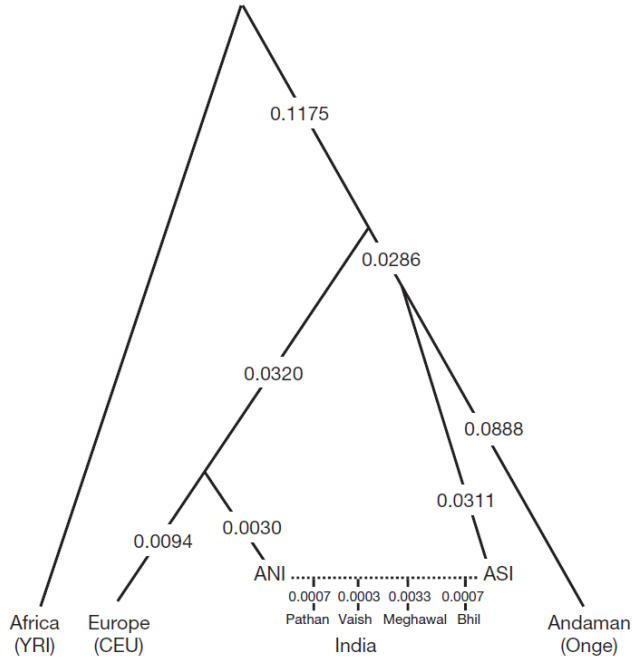



**Figure 4 | A model relating the history of Indian and non-Indian groups.**

LETTER

doi:10.1038/nature11258

## Reconstructing Native American population history

David Reich[1,2], Nick Patterson[2], Desmond Campbell[3,4], Arti Tandon[1,2], Stéphane Mazieres[3,5], Nicolas Ray[6], Maria V. Parra[3,7], Winston Rojas[3,7], Constanza Duque[3,7], Natalia Mesa[3,7], Luis F. García[7], Omar Triana[7], Silvia Blair[7], Amanda Maestre[7], Juan C. Dib[8], Claudio M. Bravi[3,9], Graciela Bailliet[9], Daniel Corach[10], Tábita Hünemeier[3,11], Maria Cátira Bortolini[11], Francisco M. Salzano[11], Maria Luiza Petzl-Erler[12], Victor Acuña-Alonzo[13], Carlos Aguilar-Salinas[14], Samuel Canizales-Quinteros[15,16], Teresa Tusié-Luna[15], Laura Riba[15], Maricela Rodríguez-Cruz[17], Mardia Lopez-Alarcón[17], Ramón Coral-Vazquez[18], Thelma Canto-Cetina[19], Irma Silva-Zolezzi[20]†, Juan Carlos Fernandez-Lopez[20], Alejandra V. Contreras[20], Gerardo Jimenez-Sanchez[20]†, María José Gómez-Vázquez[21], Julio Molina[22], Ángel Carracedo[23], Antonio Salas[23], Carla Gallo[24], Giovanni Poletti[24], David B. Witonsky[25], Gorka Alkorta-Aranburu[25], Rem I. Sukernik[26], Ludmila Osipova[27], Sardana A. Fedorova[28], René Vasquez[29], Mercedes Villena[29], Claudia Moreau[30], Ramiro Barrantes[31], David Pauls[32], Laurent Excoffier[33,34], Gabriel Bedoya[7], Francisco Rothhammer[35], Jean-Michel Dugoujon[36], Georges Larrouy[36], William Klitz[37], Damian Labuda[30], Judith Kidd[38], Kenneth Kidd[38], Anna Di Rienzo[25], Nelson B. Freimer[39], Alkes L. Price[2,40] & Andrés Ruiz-Linares[3]

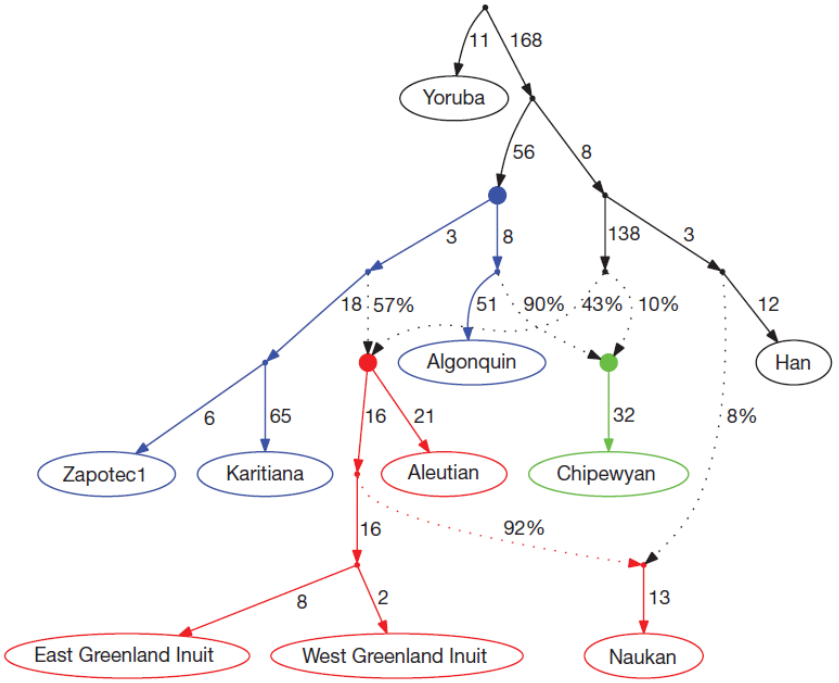



**Figure 2 | Distinct streams of gene flow from Asia into America.** We present

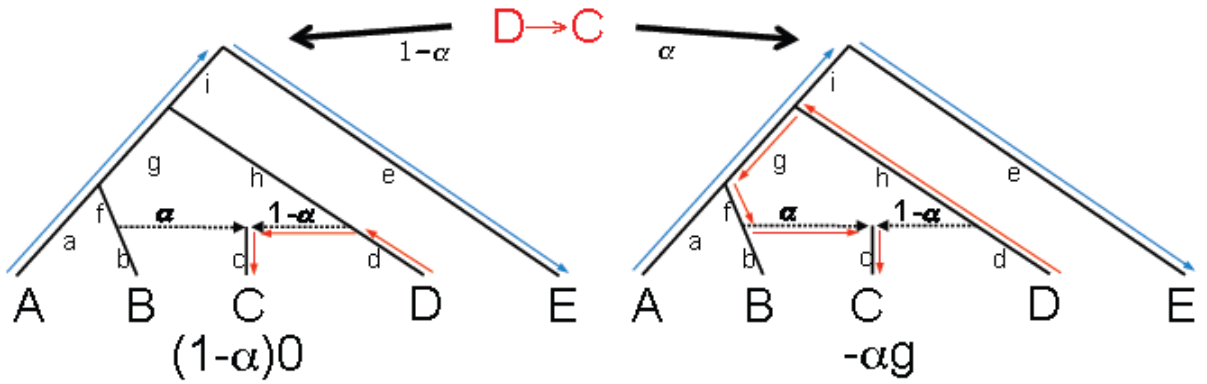
# Admixture graph

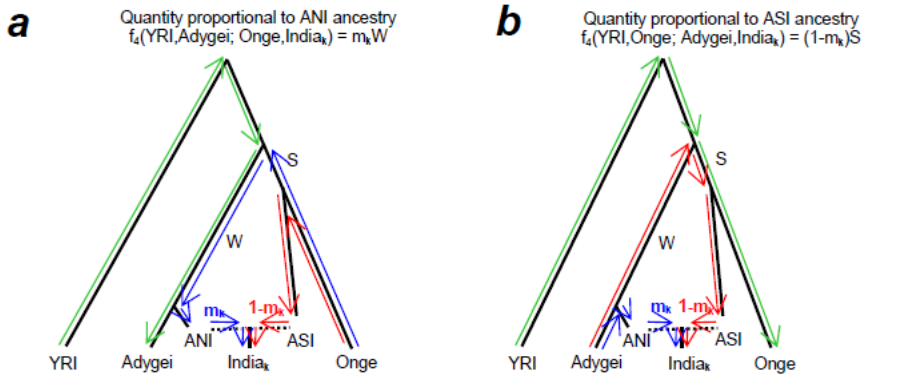
See Reich et al. 2009 & 2012, Paterson 2012

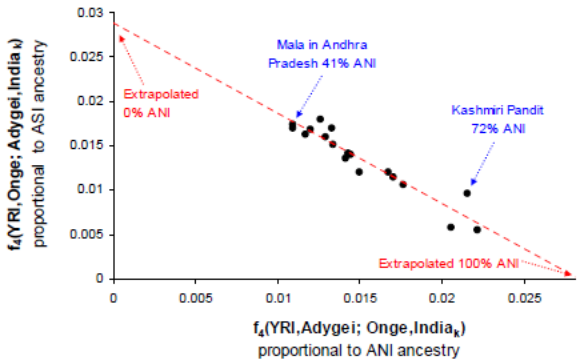
$$F_4(A,B;C,D) = E[(a'-b')(c'-d')]$$

共通の経路がない場合0

(c) $f_4(A,E;D,C) = -\alpha g$ $\quad f_4\ ratio = \dfrac{f_4(A,E;D,C)}{f_4(A,E;D,B)} = \dfrac{-\alpha g}{-g} = \alpha$



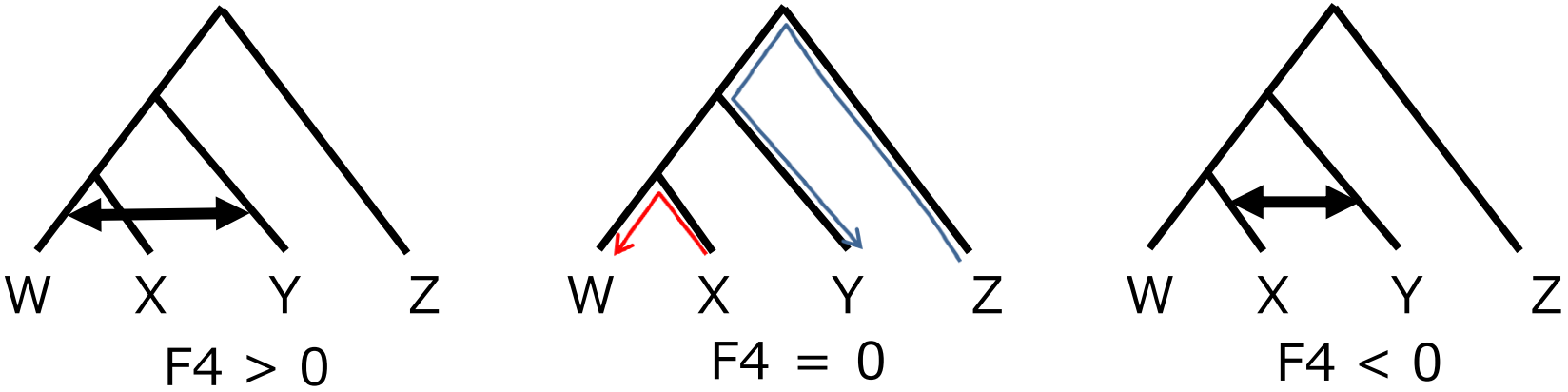




**Note S5 Figure 2:** *Regression Ancestry Estimation* along the Indian Cline. We use $f_4$ statistics to estimate a statistic $f_4$(YRI, Adygei; Onge, $India_k$) that is expected to equal $m_k W$ for each Indian group $India_k$, where W is the genetic drift that occurred ancestral to the divergence of Adygei and ANI. This value should be proportional to the ANI ancestry in each Indian Cline group. We similarly calculate $f_4$(YRI, Onge; Adygei, $India_k$), which we expect to equal $(1-m_k)S$ and should be proportional to the ASI drift in each Indian Cline group. By carrying out a least-squares fit to the 18 groups, we extrapolate the x- and y-intercepts, which correspond to the values expected for groups with entirely ANI and entirely ASI ancestry. We then interpolate the mixture proportions.

# F4 test (ABBA-BABA)によるadmixtureの検出

$F4(W, X; Y, Z) = \Sigma(P_W - P_X)(P_Y - P_Z)$

頻度の差＝ドリフト




$$D = \frac{n(BABA) - n(ABBA)}{n(BABA) + n(ABBA)}$$

# MixMapper

## Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow

Mark Lipson,[†,1] Po-Ru Loh,[†,1] Alex Levin,[1] David Reich,[2,3] Nick Patterson,[2] and Bonnie Berger[*,1,2]
[1]Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology
[2]Broad Institute, Cambridge, Massachusetts
[3]Department of Genetics, Harvard Medical School
[†]These authors contributed equally to this work.
***Corresponding author:** E-mail: bab@mit.edu.
**Associate editor:** John Novembre

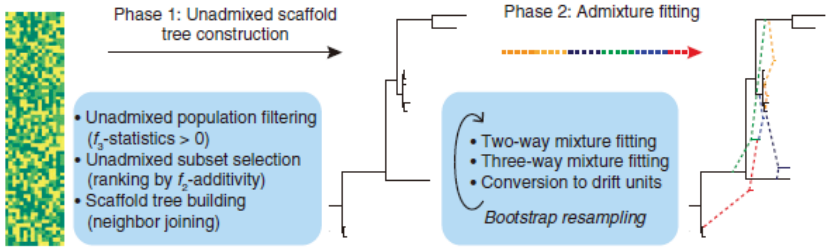



FIG. 1. *MixMapper* workflow. *MixMapper* takes as input an array of SNP calls annotated with the population to which each individual belongs. The method then proceeds in two phases, first building a tree of (approximately) unadmixed populations and then attempting to fit the remaining populations as admixtures. In the first phase, *MixMapper* produces a ranking of possible unadmixed trees in order of deviation from $f_2$-additivity; based on this list, the user selects a tree to use as a scaffold. In the second phase, *MixMapper* tries to fit remaining populations as two- or three-way mixtures between branches of the unadmixed tree. In each case, *MixMapper* produces an ensemble of predictions via bootstrap resampling, enabling confidence estimation for inferred results.

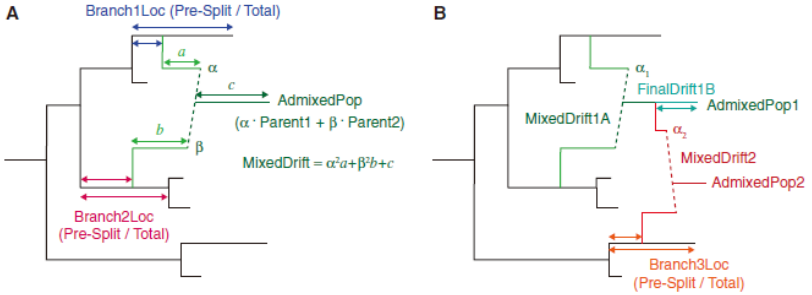



FIG. 2. Schematic of mixture parameters fit by *MixMapper*. (*A*) A simple two-way admixture. *MixMapper* infers four parameters when fitting a given population as an admixture. It finds the optimal pair of branches between which to place the admixture and reports the following: Branch1Loc and Branch2Loc are the points at which the mixing populations split from these branches (given as pre-split length/total branch length); $\alpha$ is the proportion of ancestry from Branch1 ($\beta = 1 - \alpha$ is the proportion from Branch2); and MixedDrift is the linear combination of drift lengths $\alpha^2 a + \beta^2 b + c$. (*B*) A three-way mixture: here AdmixedPop2 is modeled as an admixture between AdmixedPop1 and Branch3. There are now four additional parameters; three are analogous to the above, namely, Branch3Loc, $\alpha_2$, and MixedDrift2. The remaining degree of freedom is the position of the split along the AdmixedPop1 branch, which divides MixedDrift into MixedDrift1A and FinalDrift1B.

Lipson et al. 2013

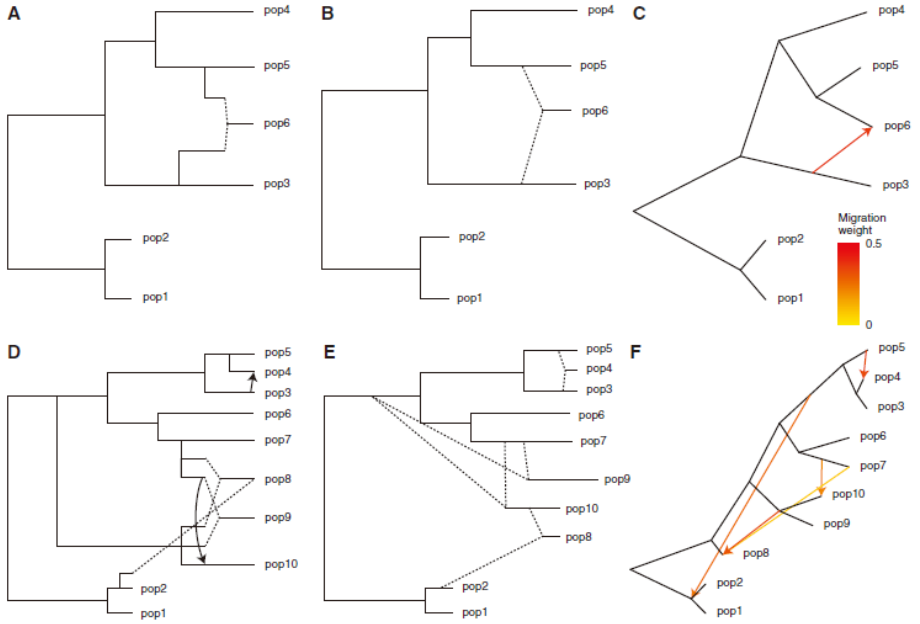



FIG. 3. Results with simulated data. (*A*–*C*) First simulated admixture tree, with one admixed population. Shown are (*A*) the true phylogeny, (*B*) *MixMapper* results, and (*C*) *TreeMix* results. (*D*–*F*) Second simulated admixture tree, with four admixed populations. Shown are (*D*) the true phylogeny, (*E*) *MixMapper* results, and (*F*) *TreeMix* results. In (*A*) and (*D*), dotted lines indicate instantaneous admixtures, whereas arrows denote continuous (unidirectional) gene flow over 40 generations. Both *MixMapper* and *TreeMix* infer point admixtures, depicted with dotted lines in (*B*) and (*E*) and colored arrows in (*C*) and (*F*). In (*B*) and (*E*), the terminal drift edges shown for admixed populations represent half the total mixed drift. Full inferred parameters from *MixMapper* are given in supplementary table S1, Supplementary Material online.

# MixMapper

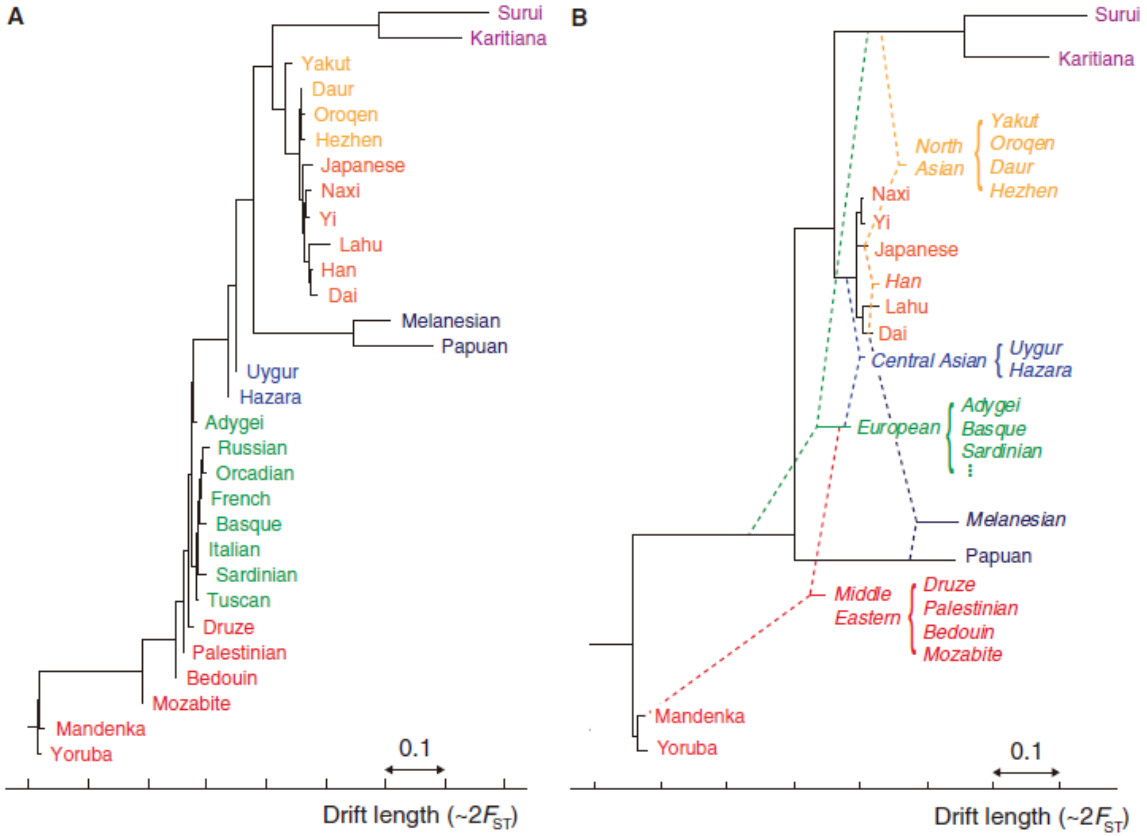



FIG. 4. Aggregate phylogenetic trees of HGDP populations with and without admixture. (*A*) A simple neighbor-joining tree on the 30 populations for which *MixMapper* produced high-confidence results. This tree is analogous to the one given by (Li et al. 2008, fig. 1B), and the topology is very similar. (*B*) Results from *MixMapper*. The populations appear in roughly the same order, but the majority are inferred to be admixed, as represented by dashed lines (cf. Pickrell and Pritchard 2012 and supplementary fig. S4, Supplementary Material online). Note that drift units are not additive, so branch lengths should be interpreted individually.

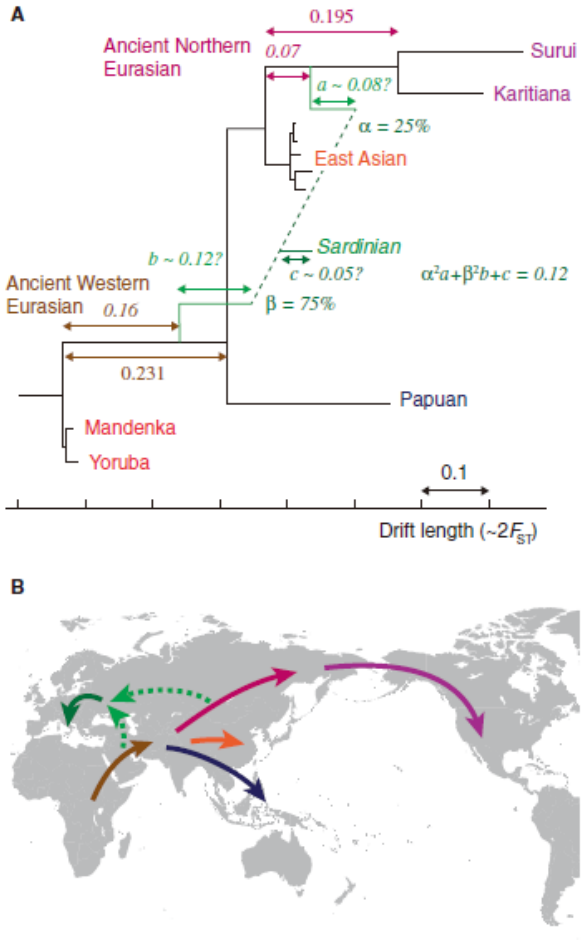
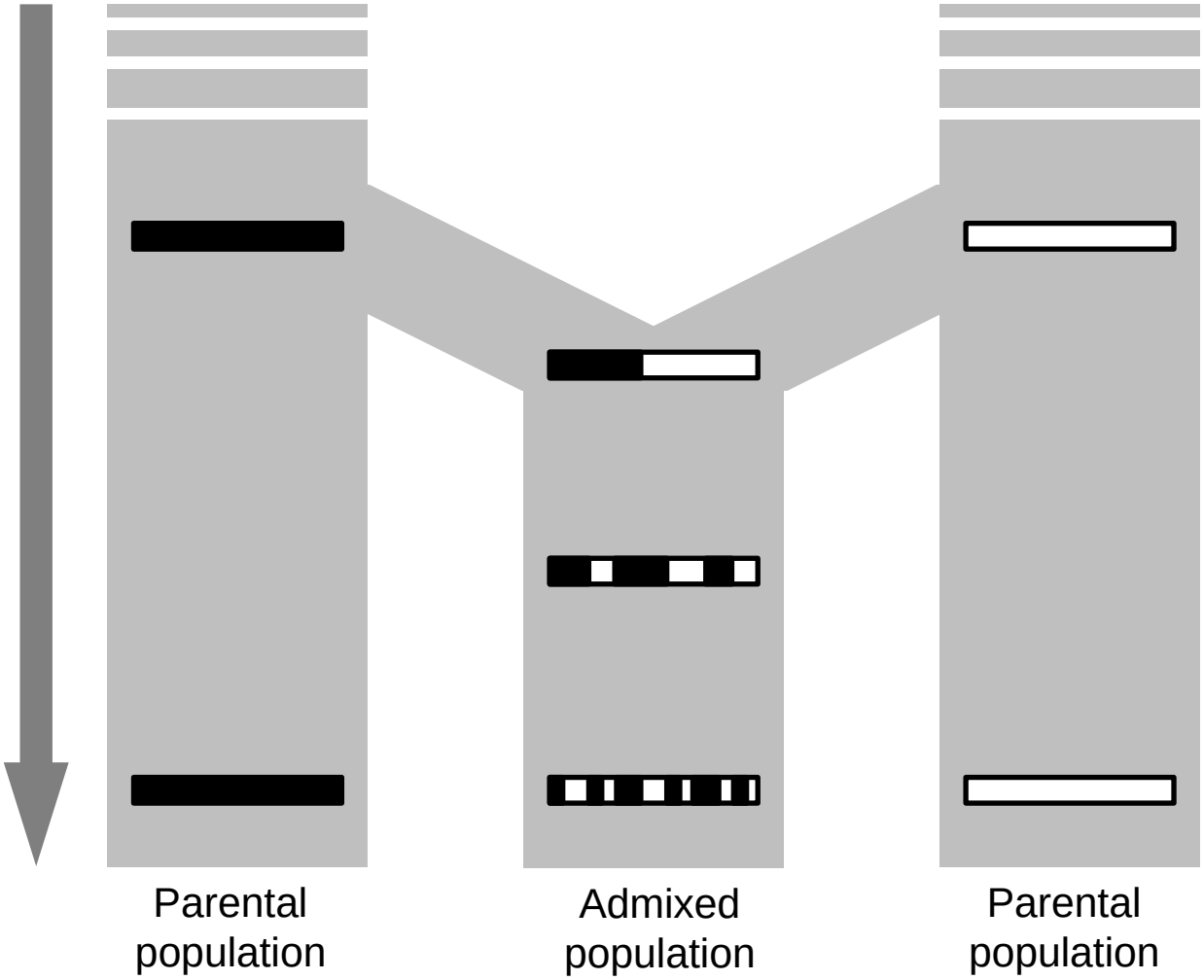



FIG. 5. Inferred ancient admixture in Europe. (*A*) Detail of the inferred ancestral admixture for Sardinians (other European populations are similar). One mixing population splits from the unadmixed tree along the common ancestral branch of Native Americans ("Ancient Northern Eurasian") and the other along the common ancestral branch of all non-Africans ("Ancient Western Eurasian"). Median parameter values are shown; 95% bootstrap confidence intervals can be found in table 1. The branch lengths *a*, *b*, and *c* are confounded, so we show a plausible combination. (*B*) Map showing a sketch of possible directions of movement of ancestral populations. Colored arrows correspond to labeled branches in (*A*).

Lipson et al. 2013

# Age of past admixture




親集団由来染色体のフラグメンテーションあるいは連鎖不平衡をみることで、
過去の集団間のadmixtureの年代を推定

# 染色体領域ごとに祖先集団を特定する

OPEN ACCESS Freely available online PLoS GENETICS

## Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations

Alkes L. Price[1,2,3], Arti Tandon[3,4], Nick Patterson[3], Kathleen C. Barnes[5], Nicholas Rafaels[5], Ingo Ruczinski[6], Terri H. Beaty[6], Rasika Mathias[7], David Reich[3,4]*, Simon Myers[3,8,9]*

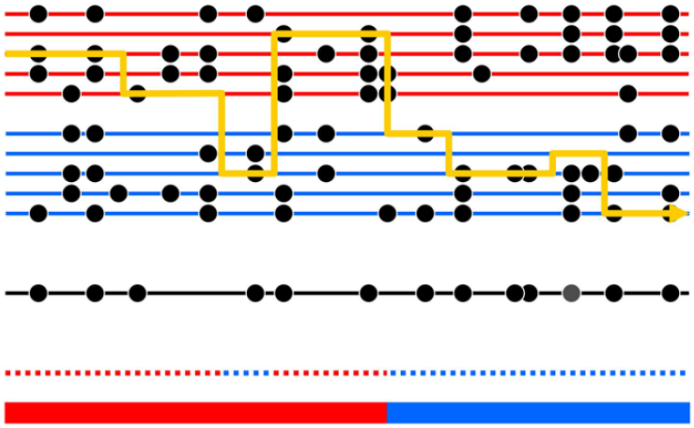


**Figure 1. Schematic of the Markov model we use for ancestry inference.** The black low

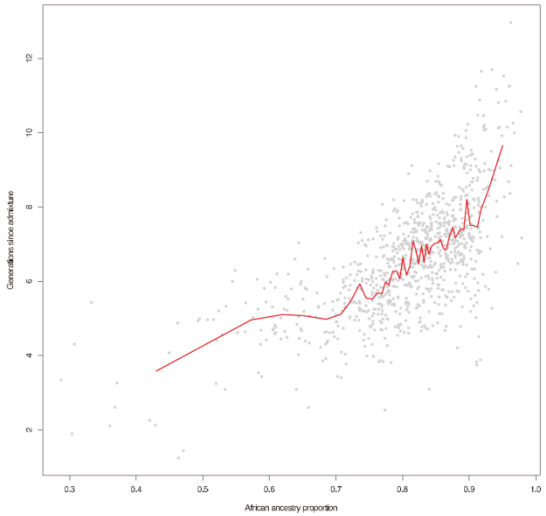



**Figure 5. Correlation between ancestry proportion and estimated time since admixture in African Americans.** Each grey point shows an estimate of the time $\lambda$ since admixture corresponding to one of 935 analysed African American individuals (Materials and Methods). The red line shows sliding averages of 20 individuals, binned according to increasing African ancestry proportions.
doi:10.1371/journal.pgen.1000519.g005

## HAPMIX

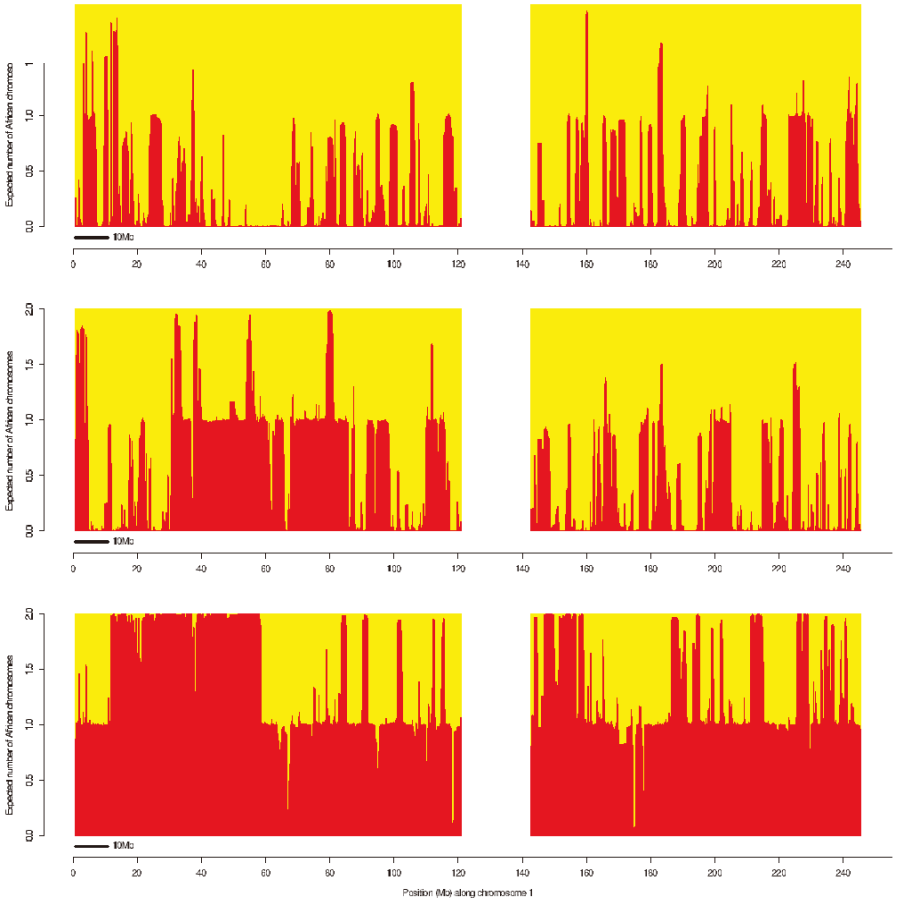



**Figure 7. Local ancestry estimates produced by HAPMIX for three real Mozabite individuals on chromosome 1.** The plots are constructed as for Figure 5, and show HAPMIX estimates of the number of sub-Saharan African copies across chromosome 1 for three individuals chosen for having different genome-wide African ancestries: 20% (top plot), 29% (middle plot) and 75% (bottom plot). The top plot looks similar to Figure 5, while the much longer segments seen in the two individuals with more African ancestry indicate more recent admixture with sub-Saharan Africans.
doi:10.1371/journal.pgen.1000519.g007

Price et al. 2009

# 連鎖不平衡係数から混血の年代を推定する

OPEN ACCESS Freely available online PLOS GENETICS

## The History of African Gene Flow into Southern Europeans, Levantines, and Jews

Priya Moorjani[1,2*], Nick Patterson[2], Joel N. Hirschhorn[1,2,3], Alon Keinan[4], Li Hao[5], Gil Atzmon[6], Edward Burns[6], Harry Ostrer[5], Alkes L. Price[7], David Reich[1,2,7*]

## Rolloff test

2SNP間の連鎖不平衡を集団間の頻度差で重みづけ

$$A(d) = \frac{\sum_{s_1,s_2 \in \mathcal{S}(d)} w(s_1)w(s_2)z(s_1,s_2)}{\left[\sum_{s_1,s_2 \in \mathcal{S}(d)} \left(w(s_1)w(s_2)\right)^2 \sum_{s_1,s_2 \in \mathcal{S}(d)} \left(z(s_1,s_2)\right)^2\right]^{1/2}}$$

$$A(d) \approx A_0 e^{-nd}$$

指数分布にフィットするように混血年代nを推定

複数回の混血があった場合には、新しい方を検出する傾向

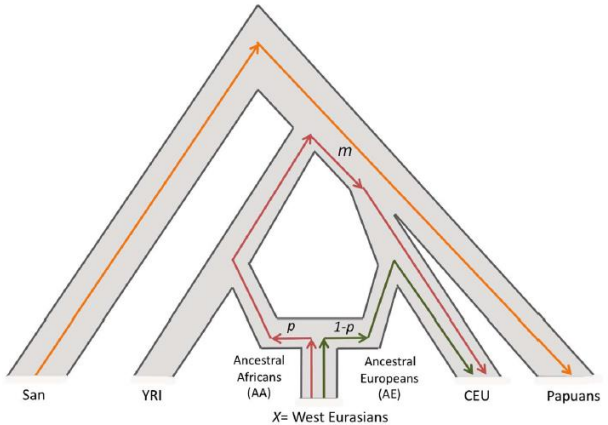



**Figure 2. Estimation of African ancestry using $f_4$ *Ancestry Estimation*.** $f_4$ *Ancestry Estimation* computes the quantity [(San-Papuan).(X-CEU)/[(San-Papuan).(YRI-CEU)]; where $X$ = any West Eurasian population. The denominator is proportional to the genetic drift $m$ that occurred in the ancestors of West or East Africans since their divergence from San but prior to their divergence from West Eurasians (intersection of red and orange lines). The numerator is proportion to $p$*(Ancestral Africans-YRI) + (1-$p$)*(Ancestral Europeans-CEU). Since the branches connecting (San, Papuan) and (CEU, $X$) do not overlap each other, the quantity (1-$p$)*(X-CEU) = 0 and hence the numerator is expected to equal $pm$. Thus, the ratio of the numerator and denominator is expected to equal $p$ (Ancestral African mixture proportion). This figure is adapted from reference [21], where we first developed $f_4$ *Ancestry Estimation*. and where we reported computer simulations demonstrating its robustness.

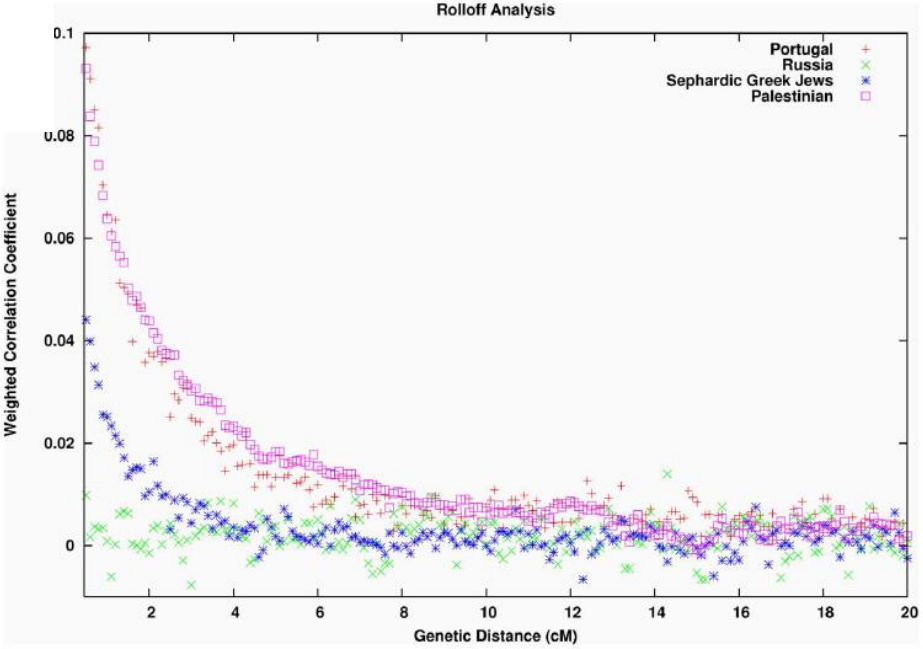



**Figure 3. Testing for LD due to African admixture in West Eurasians.** To generate these plots, we used the *ROLLOFF* software to calculate the LD between all pairs of markers in each population, weighted by their frequency difference between YRI and CEU to make the statistic sensitive to admixture LD. We plot the correlation as a function of genetic distance for Portuguese, Russians, Sephardic Greek Jews and Palestinians. We do not show inter-SNP intervals of <0.5cM since we have found that at this distance admixture LD begins to be confounded by background LD, and so inferences are not reliable (exponential curve fitting does not include inter-SNP intervals at this scale).
doi:10.1371/journal.pgen.1001373.g003

Moorjani et al. 2011

# 連鎖不平衡係数から混血の年代を推定する
# ―片方の親集団が現存しない場合

**Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium**

Po-Ru Loh,*,[1] Mark Lipson,*,[1] Nick Patterson,[†] Priya Moorjani,[†,‡] Joseph K. Pickrell,[‡] David Reich,[†,‡,2] and Bonnie Berger*,[†,2]
*Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, [†]Broad Institute, Cambridge, Massachusetts 02142, and [‡]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115

ALDER Loh et al. 2011

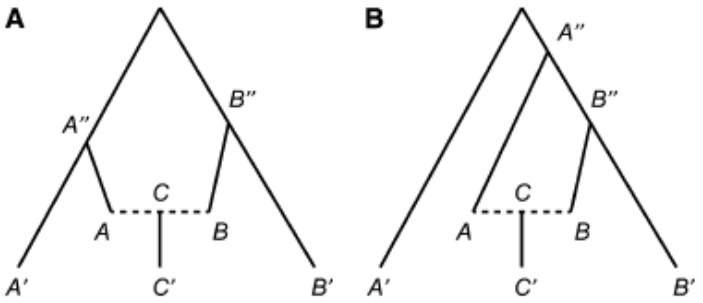



**Figure 1** Notational diagram of phylogeny containing admixed population and references. Population *C*′ is descended from an admixture between *A* and *B* to form *C*; populations *A*′ and *B*′ are present-day references. In practice, we assume that postadmixture drift is negligible; *i.e.*, the *C*–*C*′ branch is extremely short and *C*′ and *C* have identical allele frequencies. The branch points of *A*′ and *B*′ from the *A*–*B* lineage are marked *A*″ and *B*″; note that in a rooted phylogeny, these need not be most recent common ancestors (as in panel B; compare to panel A).

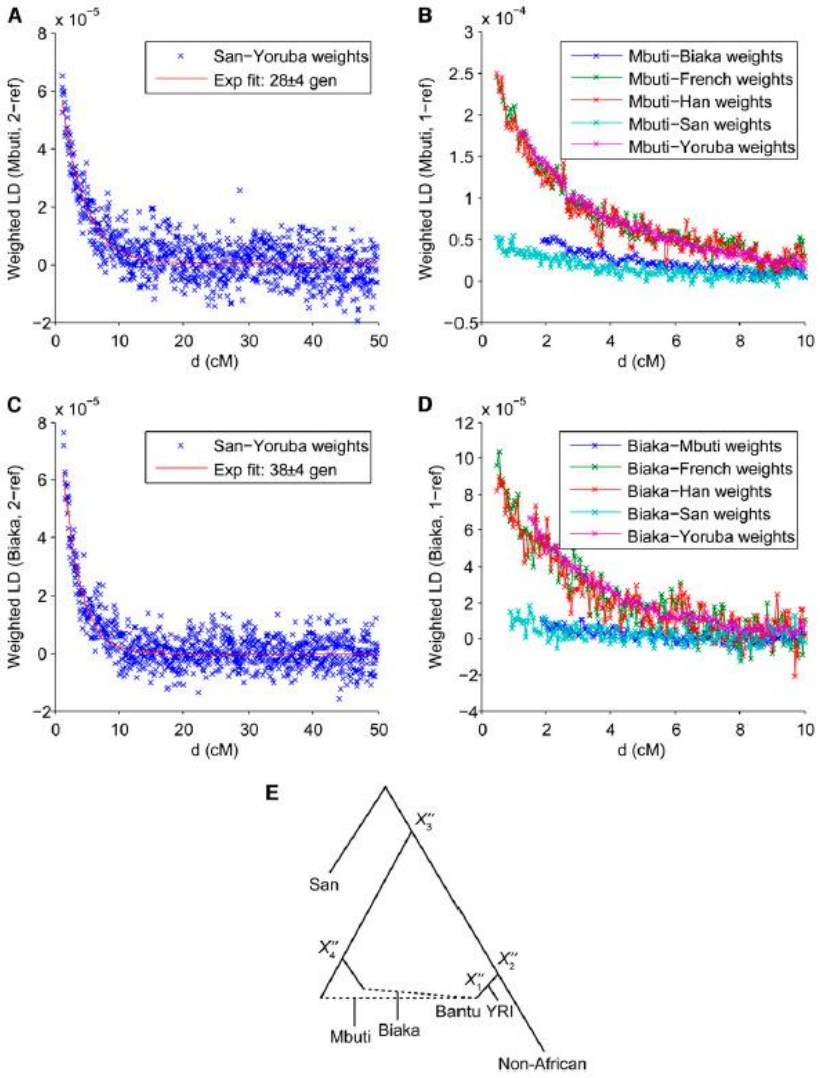



**Figure 3** Weighted LD curves for Mbuti using San and Yoruba as reference populations (A) and using Mbuti itself as one reference and several different second references (B), and analogous curves for Biaka (C and D). Genetic distances are discretized into bins at 0.05 cM resolution. Data for each curve are plotted and fit starting from the corresponding ALDER-computed LD correlation thresholds. Different amplitudes of one-reference curves (B and D) imply different phylogenetic positions of the references relative to the true mixing populations (*i.e.*, different split points $X_i''$), suggesting a sketch of a putative admixture graph (E). Relative branch lengths are qualitative, and the true root is not necessarily as depicted.

# Admixtools: Paterson et al. 2012

## Ancient Admixture in Human History

Nick Patterson[1], Priya Moorjani[2], Yontao Luo[3], Swapan Mallick[2], Nadin Rohland[2], Yiping Zhan[3], Teri Genschoreck[3], Teresa Webster[3], and David Reich[1,2]

[1]Broad Institute of Harvard and MIT, Cambridge, MA 02142
[2]Department of Genetics, Harvard Medical School, Boston, MA 02115
[3]Affymetrix, Inc., 3420 Central Expressway, Santa Clara, CA 95051

ADMIXTOOOLS version 3.0, 3/11/15
The package contains 6 programs:

convertf: programs for converting file formats.

qp3Pop: This test can be used as a format test of admixture with 3 populations.

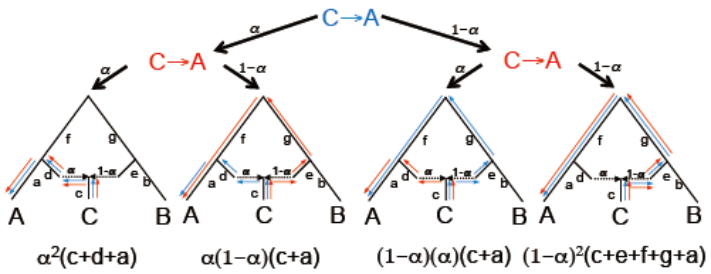
qpBound: This test can be used for estimating bounds on the admixture proportions, given 3 populations (2 reference and one target).

qpDstat: This is a formal test of admixture with 4 populations.

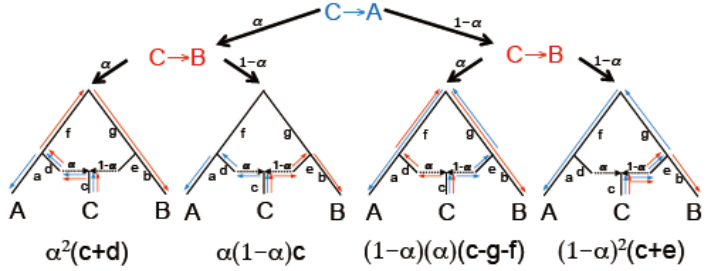
qpF4Ratio: This program computes the admixture proportion by taking the ratio of two f4 tests.

rolloff: This program can be used for dating admixture events.

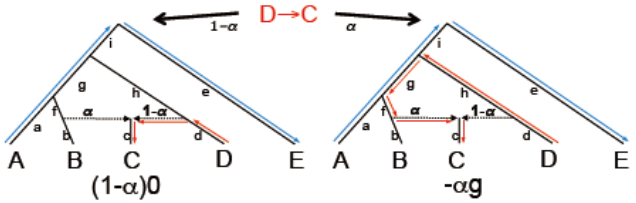
(a) $f_2(C,A) = a + c + \alpha^2 d + (1-\alpha)^2(e+g+f)$




(b) $f_3(C;A,B) = c + \alpha^2 d + (1-\alpha)^2 e - \alpha(1-\alpha)(g+f)$




(c) $f_4(A,E;D,C) = -\alpha g$ $f_4\ ratio = \frac{f_4(A,E;D,C)}{f_4(A,E;D,B)} = \frac{-\alpha g}{-g} = \alpha$



# 集団動態に関するパラメータの推定

## モーメントベースの推定

- 統計学において、母集団において成り立っているモーメント（積率：母平均，母分散など）と，観測される標本から作成するモーメント（標本平均，標本分散など）が一致するように未知パラメータを定める方法をモーメント法という。
- 仮定する集団動態が単純（集団サイズ一定、移住無し等）であれば、パラメータを算出可能。

例）コインを5回投げて表が3回、裏が2回出た。だから、表が出る確率pは3/5だね。

## 尤度ベースの推定

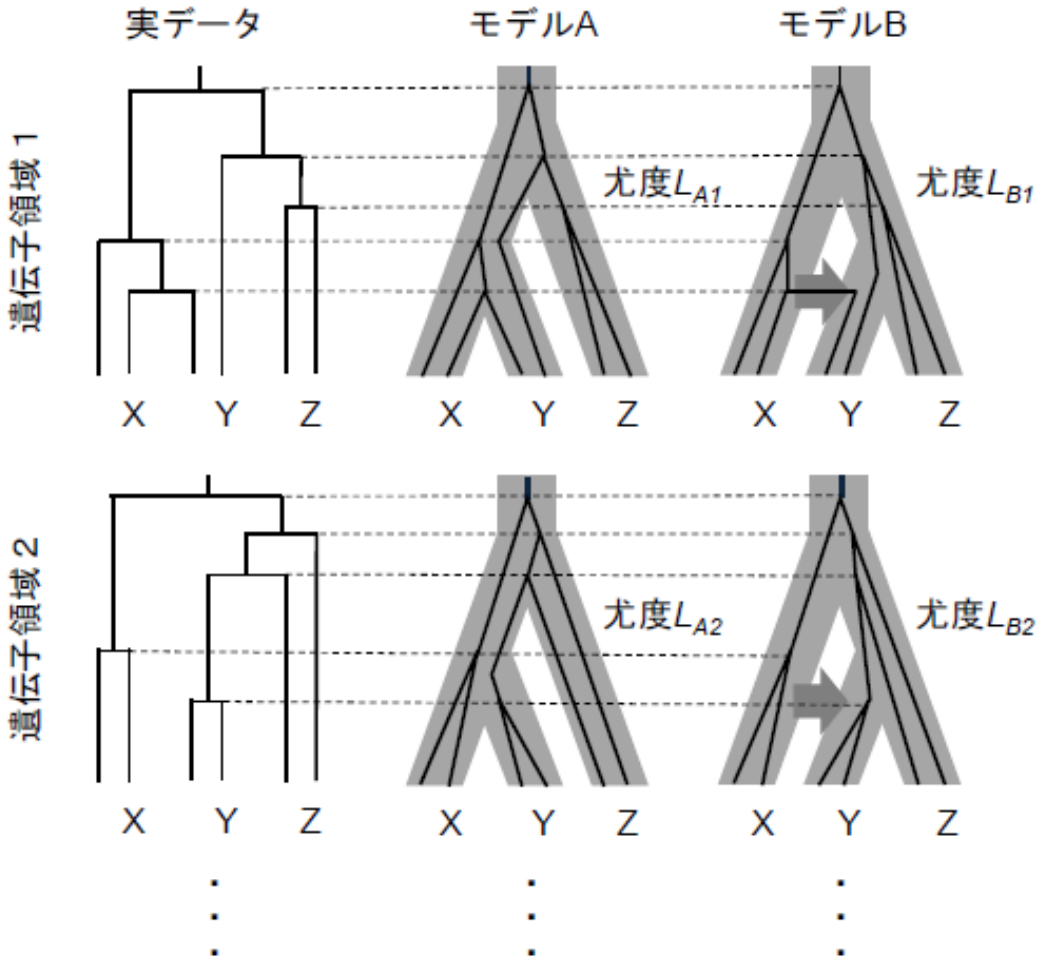
- 標本が得られる尤度が最大となるように未知パラメータの値を推定する方法。尤度関数から、最尤推定量を求める。
- 仮定する集団動態が複雑であっても、適用可能。

例） コインを5回投げて、表が3回、裏が2回出た。表が出る確率をpとすれば尤度は

$$L = {}_5C_3 p^3 (1-p)^2$$

で、Lを最大にするpは3/5だね。

# Likelihood-based estimation of population demography from molecular data




Likelihood-based estimationでは、モデルを設定し、それぞれのゲノム領域における遺伝子系図を尤もらしく説明できるようパラメータを推定する。
複数のモデルを比較する際して、尤もらしいモデルを選定することもできる

# Isolation with migration (IM) model

## Distinguishing Migration From Isolation: A Markov Chain Monte Carlo Approach

Rasmus Nielsen and John Wakeley

*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138*

Manuscript received September 5, 2000
Accepted for publication March 2, 2001

Nielsen&Wakeley 2001

## Isolation with Migration Models for More Than Two Populations

Jody Hey*

Department of Genetics, Rutgers University

***Corresponding author:** E-mail: hey@biology.rutgers.edu.

**Associate editor:** Asger Hobolth

**Abstract**

A method for studying the divergence of multiple closely related populations is described and assessed. The approach of Hey and Nielsen (2007, Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proc Natl Acad Sci USA. 104:2785–2790) for fitting an isolation-with-migration model was extended to the case of multiple populations with a known phylogeny. Analysis of simulated data sets reveals the kinds of history that are accessible with a multipopulation analysis. Necessarily, processes associated with older time periods in a phylogeny are more difficult to estimate; and histories with high levels of gene flow are particularly difficult with more than two populations. However, for histories with modest levels of gene flow, or for very large data sets, it is possible to study large complex divergence problems that involve multiple closely related populations or species.

**Key words:** divergence population genetics, coalescent, gene flow, speciation.

Hey 2010

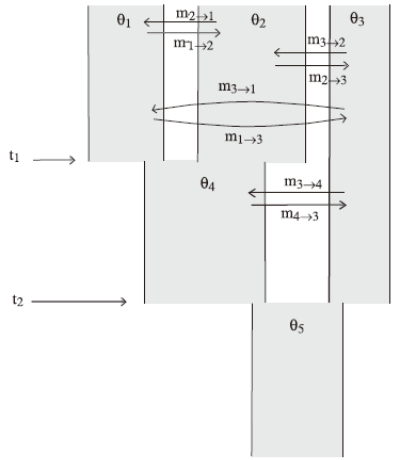



FIG. 1. An isolation-with-migration model for three sampled populations.

*Open access, freely available online* PLoS BIOLOGY

## On the Number of New World Founders: A Population Genetic Portrait of the Peopling of the Americas

Hey 2005

**Jody Hey**

Department of Genetics, Rutgers, the State University of New Jersey, Piscataway, New Jersey, United States of America

**The founding of New World populations by Asian peoples is the focus of considerable archaeological and genetic research, and there persist important questions on when and how these events occurred. Genetic data offer great potential for the study of human population history, but there are significant challenges in discerning distinct demographic processes. A new method for the study of diverging populations was applied to questions on the founding and history of Amerind-speaking Native American populations. The model permits estimation of founding population sizes, changes in population size, time of population formation, and gene flow. Analyses of data from nine loci are consistent with the general portrait that has emerged from archaeological and other kinds of evidence. The estimated effective size of the founding population for the New World is fewer than 80 individuals, approximately 1% of the effective size of the estimated ancestral Asian population. By adding a splitting parameter to population divergence models it becomes possible to develop detailed portraits of human demographic history. Analyses of Asian and New World data support a model of a recent founding of the New World by a population of quite small effective size.**

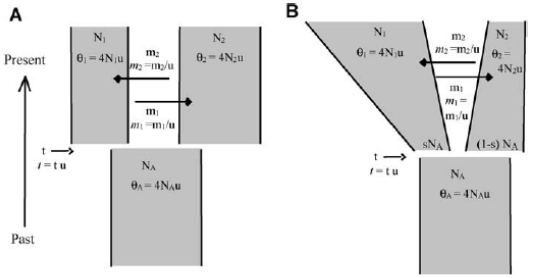



**Figure 1.** Isolation with Migration Models

(A) The basic IM model. The demographic terms are effective population sizes ($N_1$, $N_2$, and $N_A$), gene flow rates ($m_1$ and $m_2$), and population splitting time (t). Also shown are parameters scaled by the neutral mutation rate (u), as they are actually used in the model fitting. Terms for basic demographic parameters, including N, m, t, and u, are not italicized. Note that the migration parameters are identified by the source of migrants as time goes backward in the coalescent. In other words, the migration rate from population 1 to population 2 (i.e., $m_1$) actually corresponds to the movement of genes from population 2 to population 1 as time moves forward.
(B) The IM model with changing population size. An additional parameter, *s*, is the fraction of $N_A$ that forms $N_1$ (i.e., the fraction $1 - s$ gives rise to $N_2$)
DOI: 10.1371/journal.pbio.0030193.g001

**Table 4.** Estimates of Demographic Quantities

| Demographic Term | Population Size Change High $t_{upper}$ | Population Size Change Low $t_{upper}$ | Constant Population Size Low $t_{upper}$ |
|---|---|---|---|
| $N_1$ | 7,190 | 19,200 | 5,394 |
| $N_2$ | 480 | 830 | 770 |
| $N_A$ | 9,180 | 9,040 | 9,640 |
| $s\ N_A$ | 9,100 | 8,970 | Not applicable |
| $(1-s)\ N_A$ | 76 | 70 | Not applicable |
| t (years) | 7,130 | 6,350 | 44,400[a] (7,900) |
| $2N_1m_1 = \theta_1 m_1 /2$ | 11.8 | 10.5 | 3.2 |
| $2N_2m_2 = \theta_2 m_2 /2$ | 0.9 | 1.4 | 2.3 |

The conversion of model parameters to demographic terms is described in "Analyses" in Materials and Methods.
[a] The estimated time associated with the highest value of *t* which is at the right margin of the distribution. The estimated time associated with the secondary peak is given in parentheses.
DOI: 10.1371/journal.pbio.0030193.t004

# Approximate Bayesian Computation (ABC)

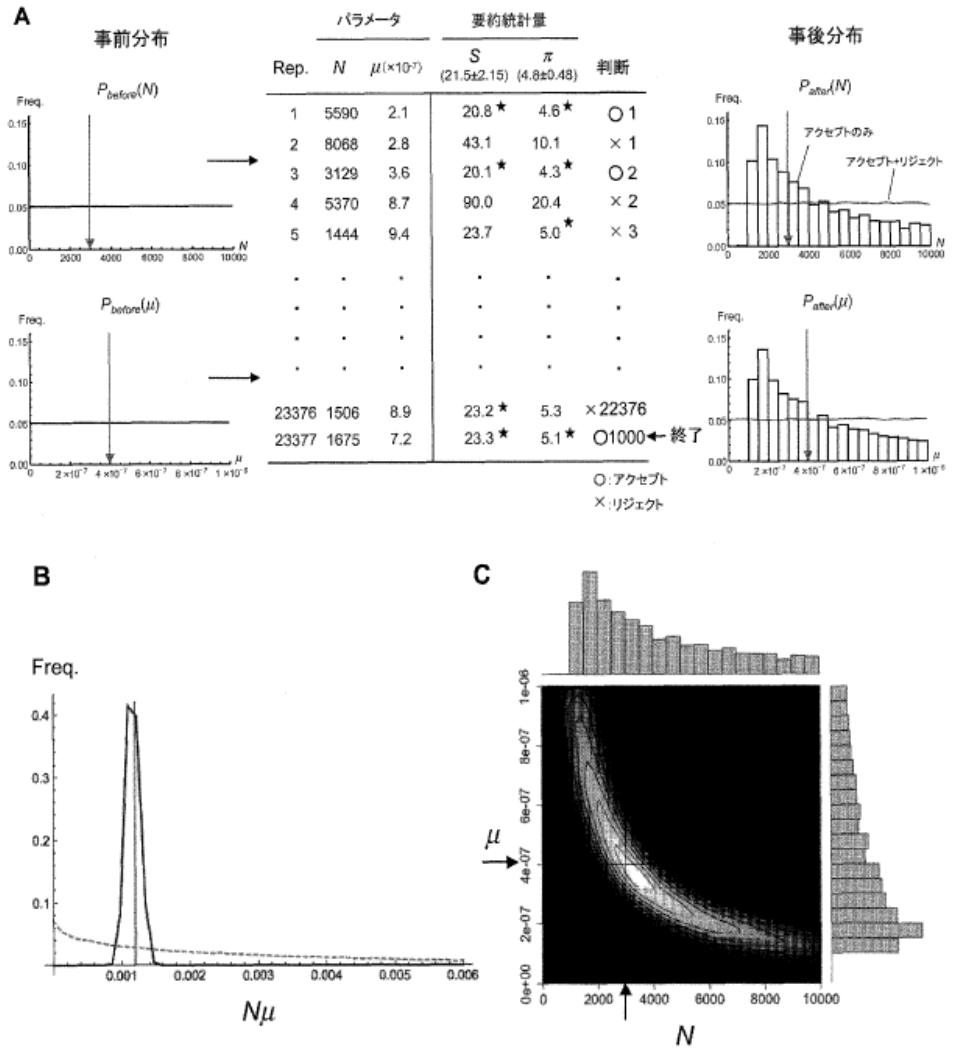



図3. 棄却サンプリングの方法の例と、2つの事後分布に相関がある場合の注意点。A. 事前分布からランダムに得られた値によって coalescent シミュレーションを行い、多型データを生成して要約統計量を計算する。採択されたサンプルのみで分布を描くと事後分布が得られる。このようにして得られた2つの別々の1次元の事後分布は、灰色の線で示した真の値からずれる。B. しかし、その積 $N\mu$ は真の値に合致する。事後分布を黒の実線で、事前分布からランダムに発生させた乱数を事前分布として灰色の破線で示した。C. 2つのパラメータの事後分布を2次元でプロットした場合。真のパラメータは $N = 3{,}000$、$\mu = 4.0 \times 10^{-7}$ に設定。それぞれ、矢印で示してある。

正確な尤度の算出を避け、要約統計量とシミュレーションを用いて近似的に尤度を算出する方法

山道&印南2010

OPEN ACCESS Freely available online PLoS GENETICS

多次元AFS

# 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]**

**1** Theoretical Biology and Biophysics and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America, **2** Human Genetics, University of Chicago, Chicago, Illinois, United States of America, **3** Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America

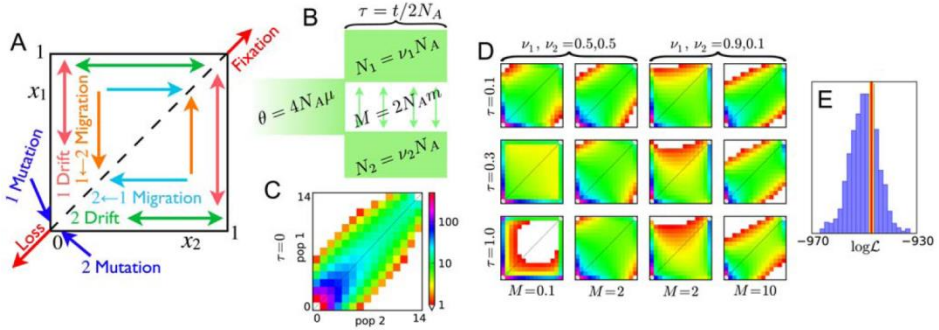



**Figure 1. Frequency spectrum gallery.** (A) Qualitative effects of modeled neutral genetic forces on $\phi(x_1, x_2, t)$, the density of alleles at relative frequencies $x_1$ and $x_2$ in populations 1 and 2. (B) For the spectra shown, an equilibrium population of effective size $N_A$ diverges into two populations $2N_A\tau$ generations ago. Populations 1 and 2 have effective sizes $\nu_1 N_A$ and $\nu_2 N_A$, respectively. Migration is symmetric at $m = M/(2N_A)$ per generation, and $\theta = 1000$. (C) The AFS at $\tau = 0$. Each entry is colored by the logarithm of the number of sites in it, according to the scale shown. (D) The AFS at various times for various demographic parameters, on the same scale as (B). (E) Comparison between coalescent- and diffusion-based estimates of the likelihood $\mathcal{L}$ of data generated under the model (A). Coalescent-based estimates of the likelihood, each of which took approximately 7.0 seconds, are represented in the histogram. The result from our diffusion approach, which took 2.0 seconds, is represented by the red line. For accuracy comparison, the yellow line indicates the likelihood inferred from $10^8$ coalescent simulations.
doi:10.1371/journal.pgen.1000695.g001

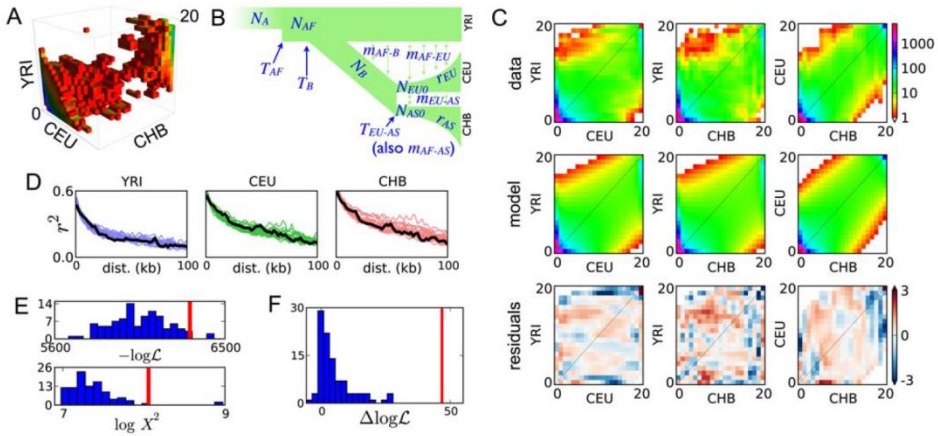



**Figure 2. Out of Africa analysis.** (A) AFS for the YRI, CEU, and CHB populations. The color scale is as in (C). (B) Illustration of the model we fit, with the 14 free parameters labeled. (C) Marginal spectra for each pair of populations. The top row is the data, and the second is the maximum-likelihood model. The third row shows the Anscombe residuals [61] between model and data. Red or blue residuals indicate that the model predicts too many or too few alleles in a given cell, respectively. (D) The observed decay of linkage disequilibrium (black lines) is qualitatively well-matched by our simulated data sets (colored lines). (E) Goodness-of-fit tests based on the likelihood $\mathcal{L}$ and Pearson's $X^2$ statistic both indicate that our model is a reasonable, though incomplete description of the data. In both plots, the red line results from fitting the real data and the histogram from fits to simulated data. Poorer fits lie to the right (lower $\mathcal{L}$ and higher $X^2$). (F) The improvement in likelihood from including contemporary migration in the real data fit (red line) is much greater than expected from fits to simulated data generated without contemporary migration (histogram). This indicates that the data contain a strong signal of contemporary migration.
doi:10.1371/journal.pgen.1000695.g002

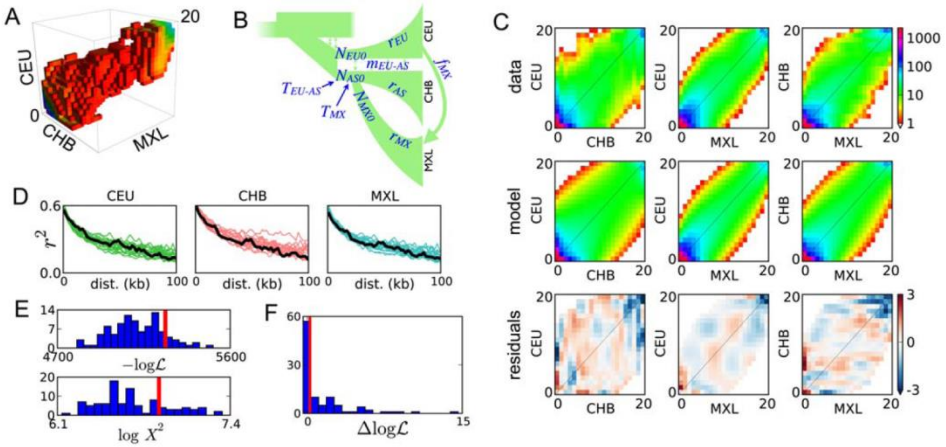



**Figure 3. Settlement of the New World analysis.** As in Figure 2, (A) is the data, (B) is a schematic of the model we fit, (C) compares the data and model AFS, and (D) compares LD. (E) The fit of our model to the real data is not atypical of fits to simulated data. (F) The improvement in real data fit upon including CHB-MXL migration (red line) is very typical of the improvement in fits to simulated data without CHB-MXL migration. Thus we have no evidence for CHB-MXL migration after divergence.
doi:10.1371/journal.pgen.1000695.g003

Gutenkunst et al. 2009

# ゲノム中の自然選択を検出しよう

# 自然選択の検出法（中立性の検定法）

1. 種間の置換にもとづく方法
   - dn/ds test
2. 種間の置換と種内の多型にもとづく方法
   - HKA test
   - McDonald-Kreitman test
3. 集団間の分化にもとづく方法
   - FST
4. 集団内の多型にもとづく方法
   - Ewens-Watterson test
   - Tajima’s D test
   - Fay and Wu’s H test
   - Long-range haplotype test (relative EHH)

など

# DNA配列比較とタイムスケール

Human

Chimp



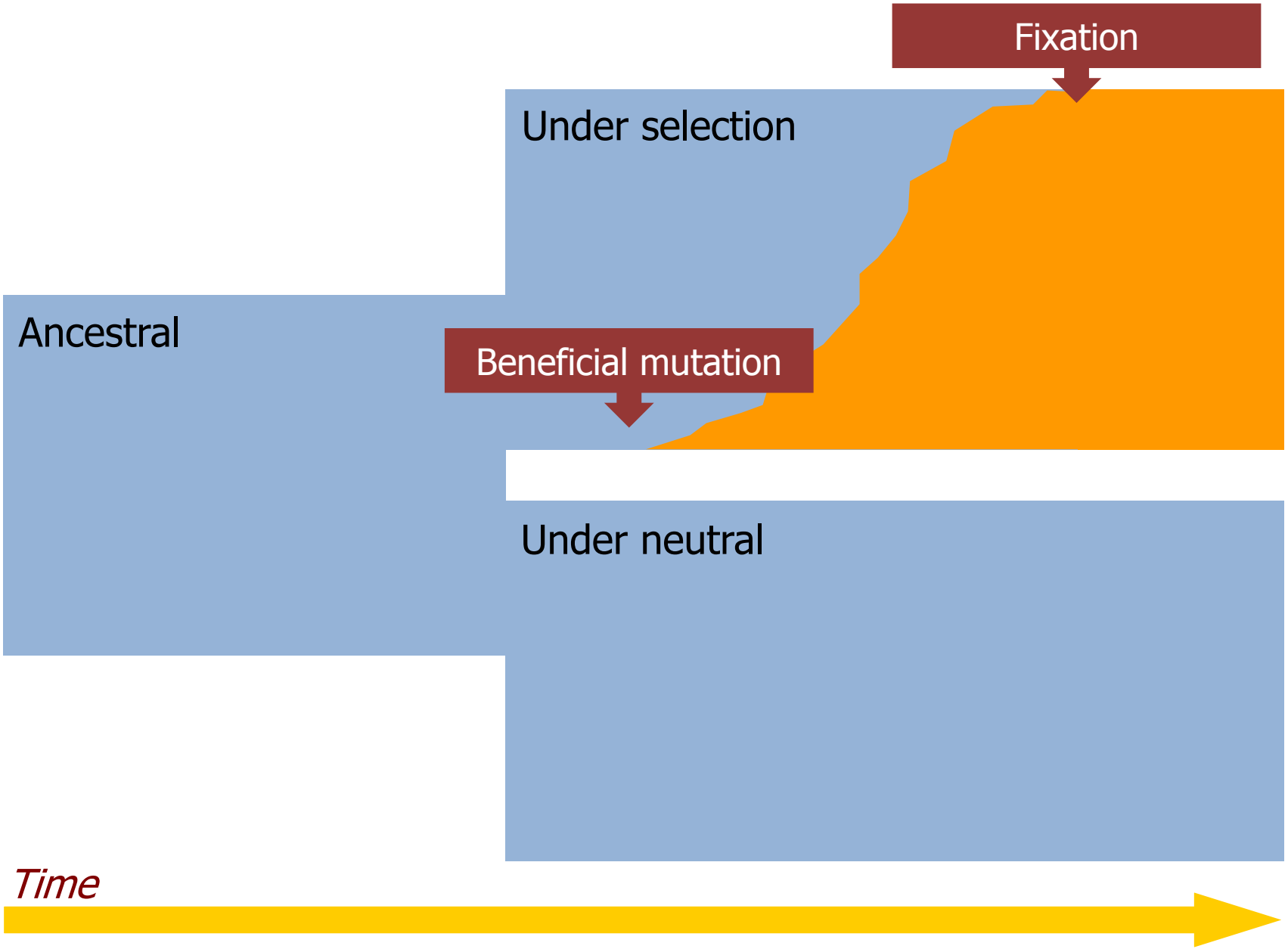

1. Between species
2. Within species
3. Between populations
4. Within populations

# *Human diversity*

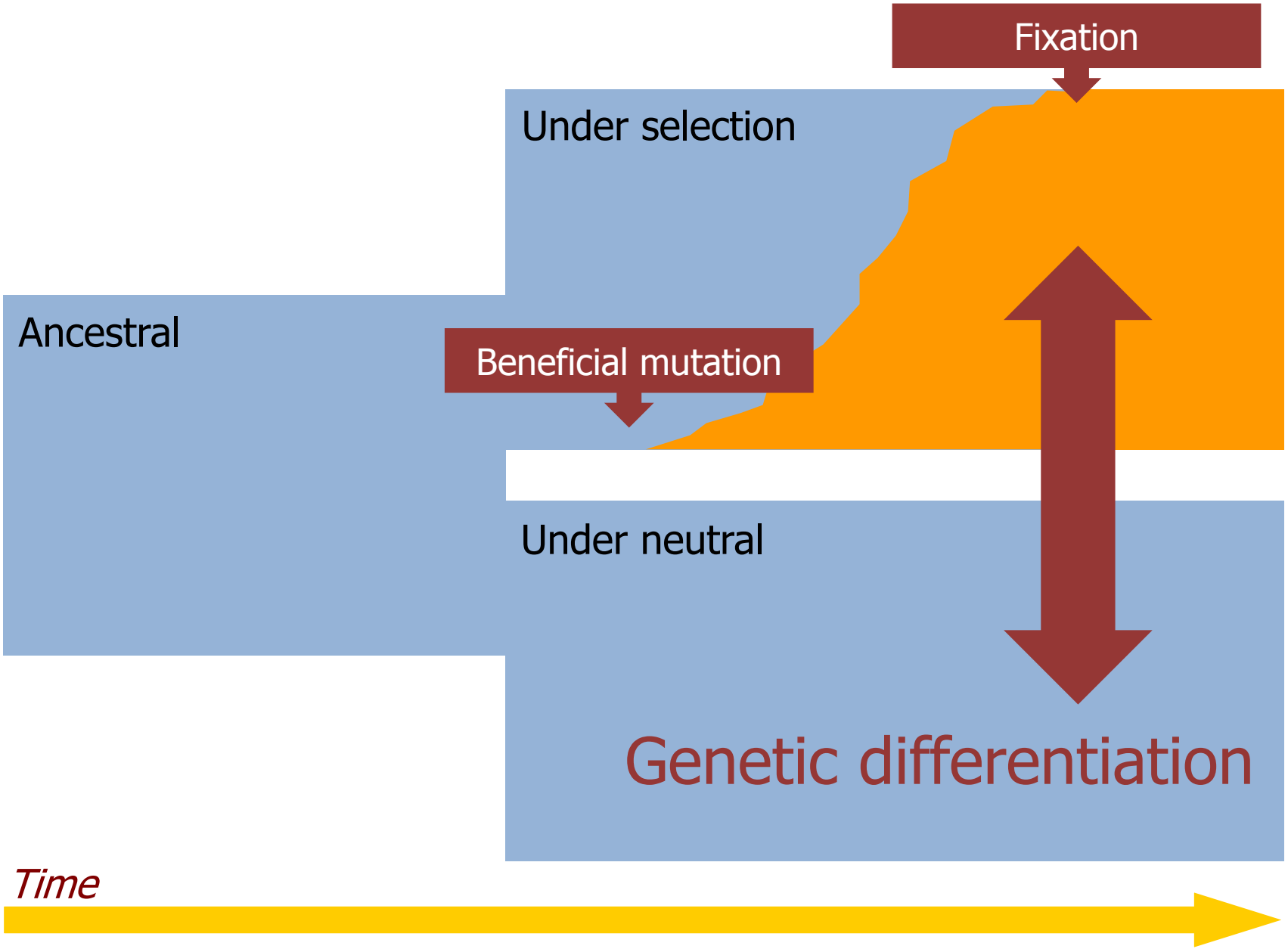

Meyers Konversations-Lexikon 1885

**Human biological diversity is likely to be caused by genetic drift and natural/sexual selection.**

# population-specific selection



# population-specific selection



# 集団間の遺伝的分化（チベット人の高地適応）

**Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude**

Xin Yi,[1,2]* Yu Liang,[1,2]* Emilia Huerta-Sanchez,[3]* Xin Jin,[1,4]* Zha Xi Ping Cuo,[2,5]* John E. Pool,[3,6]* Xun Xu,[1] Hui Jiang,[1] Nicolas Vinckenbosch,[3] Thorfinn Sand Korneliussen,[7] Hancheng Zheng,[1,4] Tao Liu,[1] Weiming He,[1,8] Kui Li,[2,5] Ruibang Luo,[1,4] Xifang Nie,[1] Honglong Wu,[1,9] Meiru Zhao,[1] Hongzhi Cao,[1,9] Jing Zou,[1] Ying Shan,[1,4] Shuzheng Li,[1] Qi Yang,[1] Asan,[1,2] Peixiang Ni,[1] Geng Tian,[1,2] Junming Xu,[1] Xiao Liu,[1] Tao Jiang,[1,9] Renhua Wu,[1] Guangyu Zhou,[1] Meifang Tang,[1] Junjie Qin,[1] Tong Wang,[1] Shuijian Feng,[1] Guohong Li,[1] Huasang,[1] Jiangbai Luosang,[1] Wei Wang,[1] Fang Chen,[1] Yading Wang,[1] Xiaoguang Zheng,[1,2] Zhuo Li,[1] Zhuoma Bianba,[10] Ge Yang,[10] Xinping Wang,[11] Shuhui Tang,[11] Guoyi Gao,[12] Yong Chen,[5] Zhen Luo,[5] Lamu Gusang,[5] Zheng Cao,[1] Qinghui Zhang,[1] Weihan Ouyang,[1] Xiaoli Ren,[1] Huiqing Liang,[1] Huisong Zheng,[1] Yebo Huang,[1] Jingxiang Li,[1] Lars Bolund,[1] Karsten Kristiansen,[1,7] Yingrui Li,[1] Yong Zhang,[1] Xiuqing Zhang,[1] Ruiqiang Li,[1,7] Songgang Li,[1] Huanming Yang,[1] Rasmus Nielsen,[1,3,7]† Jun Wang,[1,7]† Jian Wang[1]†

$$F_{ST} = (H_T - H_S)/H_T$$

遺伝的分化が進んでいない集団間で有効

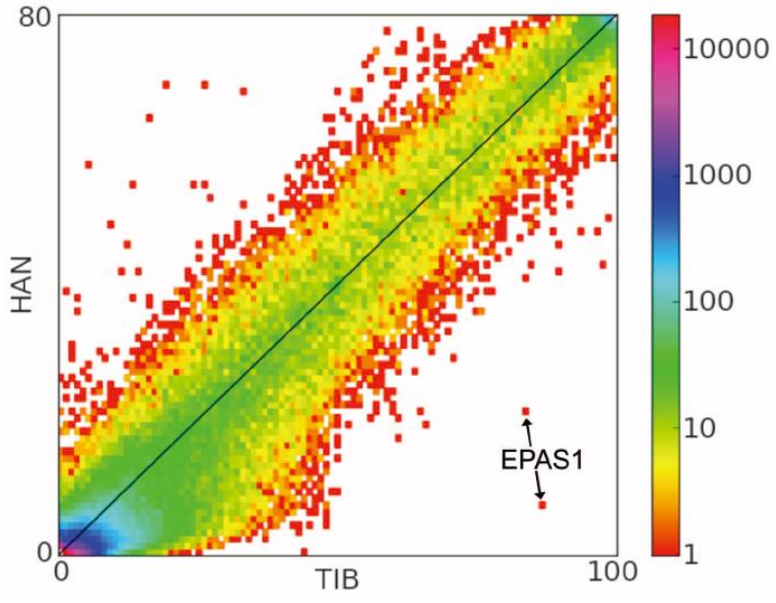



**Fig. 1.** Two-dimensional unfolded site frequency spectrum for SNPs in Tibetan (*x* axis) and Han (*y* axis) population samples. The number of SNPs detected is color-coded according to the logarithmic scale plotted on the right. Arrows indicate a pair of intronic SNPs from the *EPAS1* gene that show strongly elevated derived allele frequencies in the Tibetan sample compared with the Han sample.

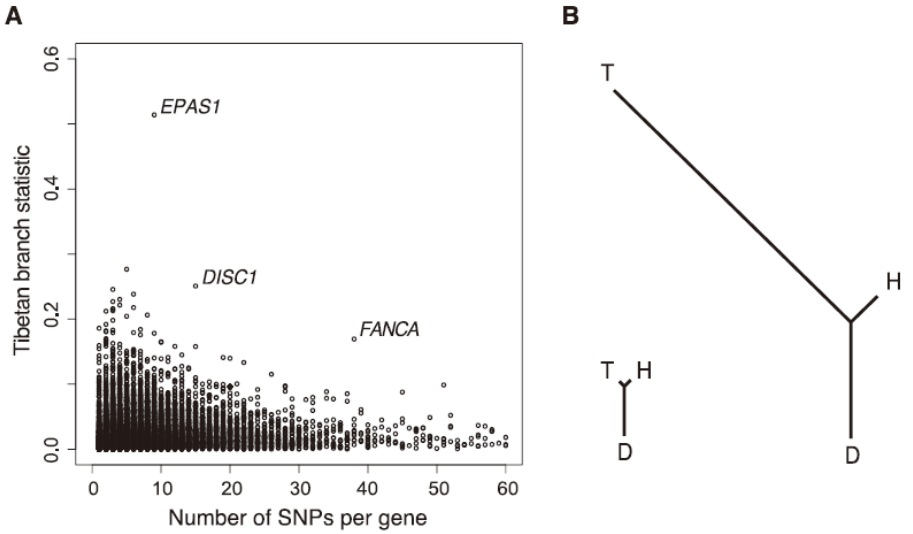



**Fig. 2.** Population-specific allele frequency change. (**A**) The distribution of $F_{ST}$-based PBS statistics for the Tibetan branches, according to the number of variable sites in each gene. Outlier genes are indicated in red. (**B**) The signal of selection on *EPAS1*: Genomic average $F_{ST}$-based branch lengths for Tibetan (T), Han (H), and Danish (D) branches (left) and branch lengths for *EPAS1*, indicating substantial differentiation along the Tibetan lineage (right).

EPAS1 = HIF2A
低酸素誘導性因子
hypoxia-inducible factor: HIF

Yi et al. 2010

# Lewontin-Krakauer test

Hudson's $F_{ST}$ estimator is

$$\hat{F}_{ST} = 1 - \frac{\hat{h}_S}{\hat{h}_B}$$

where

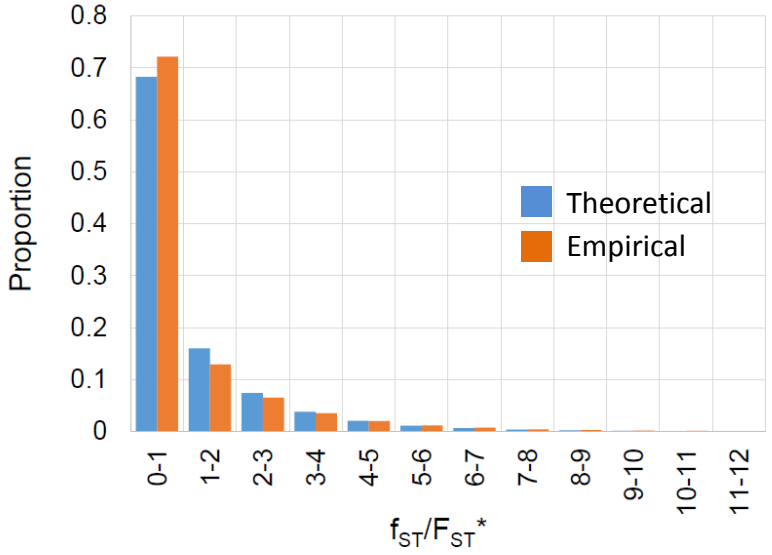
$$\hat{h}_S = \frac{1}{K}\sum_{k=1}^{K} 2\hat{p}_k(1-\hat{p}_k)\left(\frac{2n_k}{2n_k-1}\right)$$

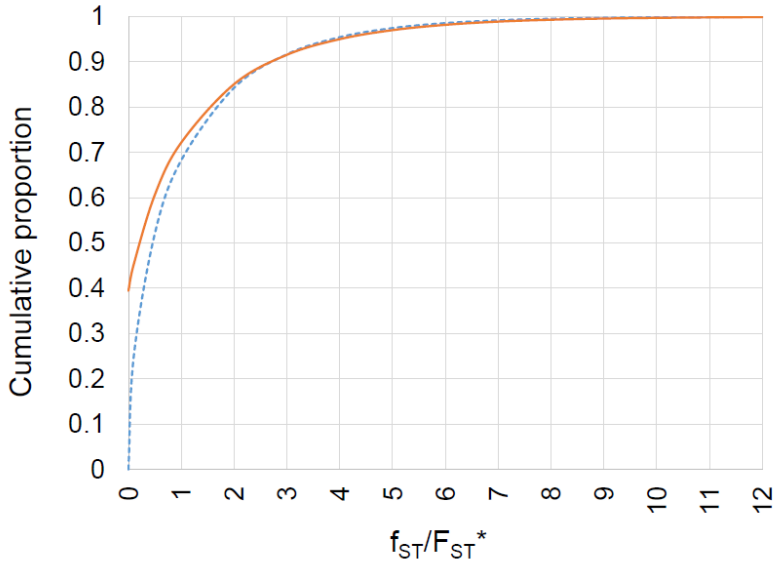
$$\hat{h}_B = \frac{1}{K(K-1)}\sum_{k\neq k'}^{K} 2\hat{p}_k(1-\hat{p}_{k'})$$

The mean of $\hat{F}_{ST}$ over loci can be calculated as a weighted mean

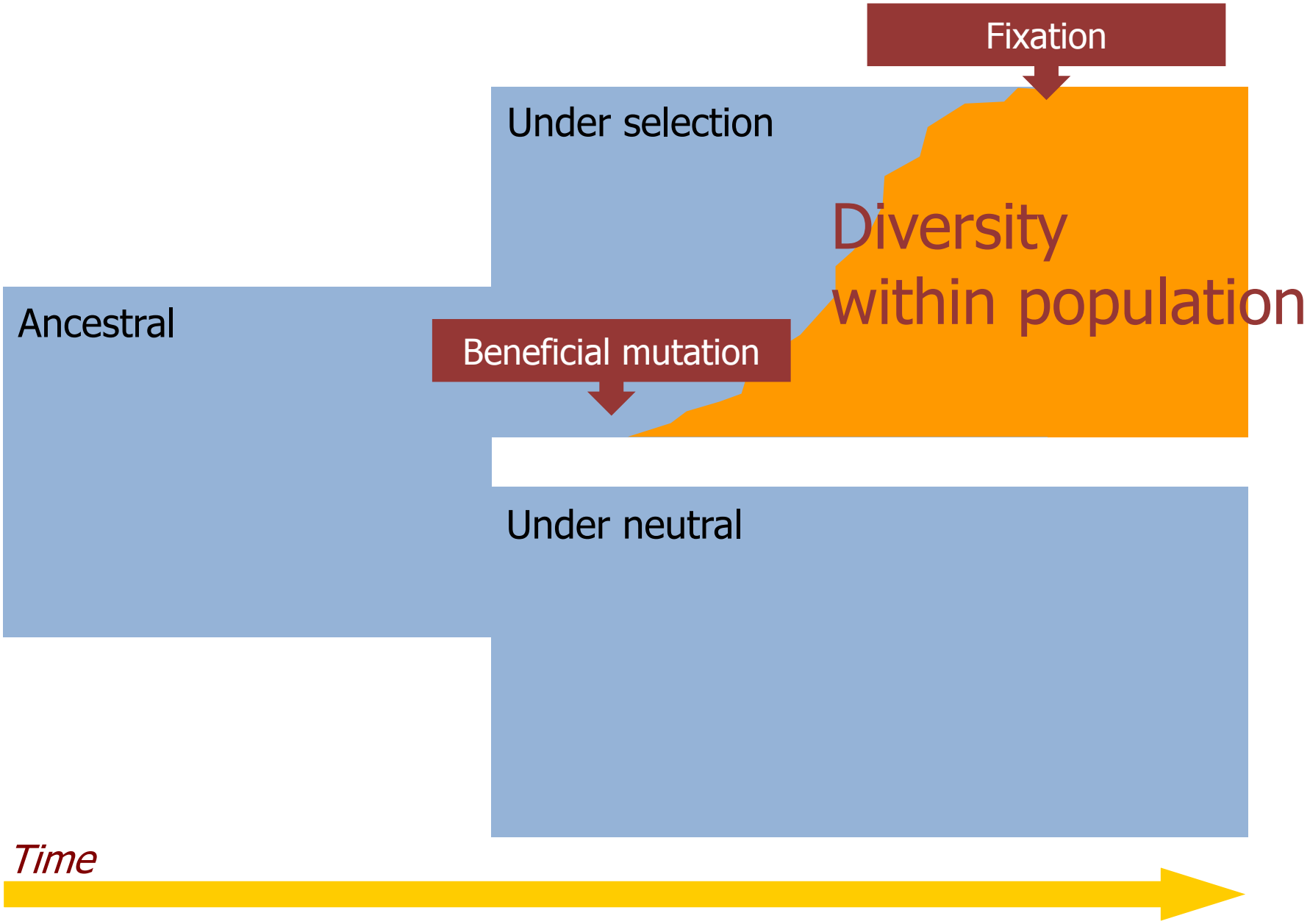
$$\bar{F}_{ST} = \frac{\sum_{i=1}^{M}\hat{h}_{Bi}\hat{F}_{STi}}{\sum_{i=1}^{M}\hat{h}_{Bi}} = 1 - \frac{\sum_{i=1}^{M}\hat{h}_{Si}}{\sum_{i=1}^{M}\hat{h}_{Bi}}$$

**Lewontin and Krakauer (1973)** pointed out that $(K-1)\hat{F}_{STi}/\bar{F}_{ST}$ has a $\chi^2$ distribution with *K-1* degrees of freedom

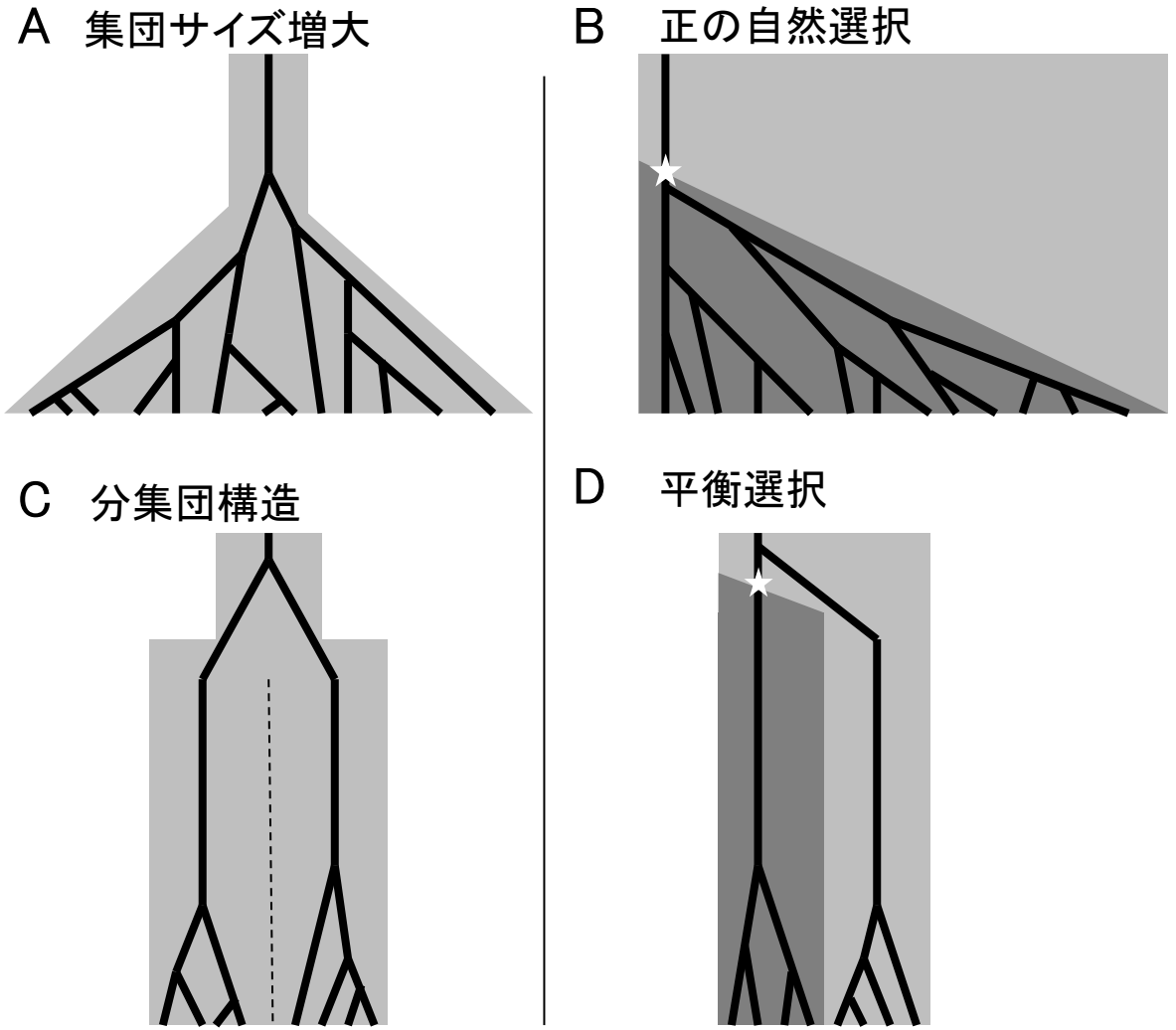


# population-specific selection



# デモグラフィおよび自然選択の遺伝子系図への効果



# Tajima's D test

$\theta = 4N\mu L$（ここでは、$\mu$が塩基あたりの突然変異率であるので、$\mu L$は配列全体の突然変異率）は塩基多様度$\pi$から推定することができ、

$$\theta_\pi = \pi L = \Pi$$

一方、集団からn個の配列をサンプリングして、シーケンシングしたとき、多型的なサイトの数の期待値は

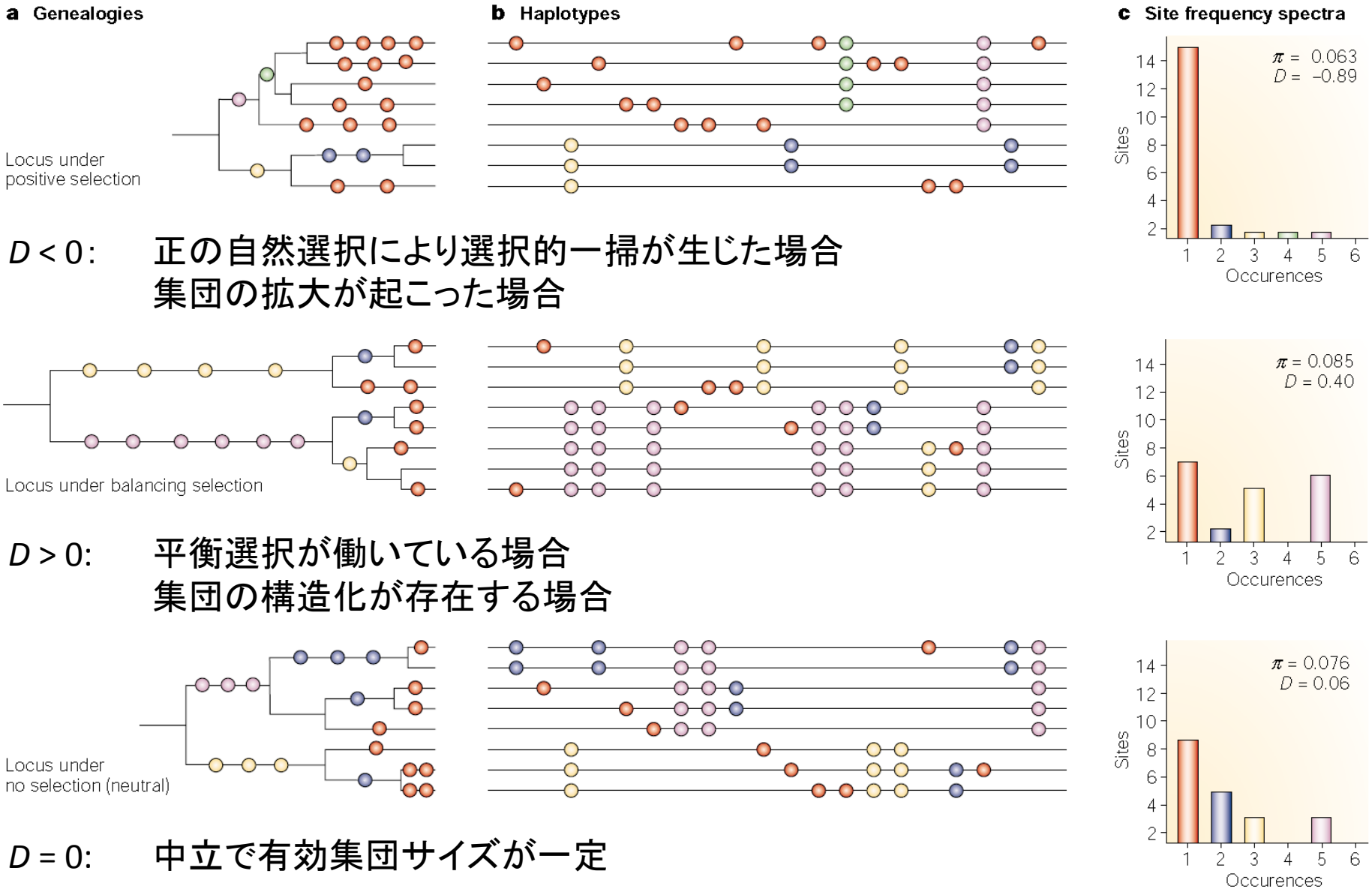
$$E(S) = 4N\mu L \sum_{i=1}^{n-1} (1/i)$$

ここで、$\theta$ の推定値は

$$\theta_S = S \Big/ \sum_{i=1}^{n-1} (1/i)$$

Tajima's D統計量は、

$$D = \frac{\theta_\pi - \theta_S}{\sqrt{V(\theta_\pi - \theta_S)}}$$




*D* < 0: 正の自然選択により選択的一掃が生じた場合
集団の拡大が起こった場合

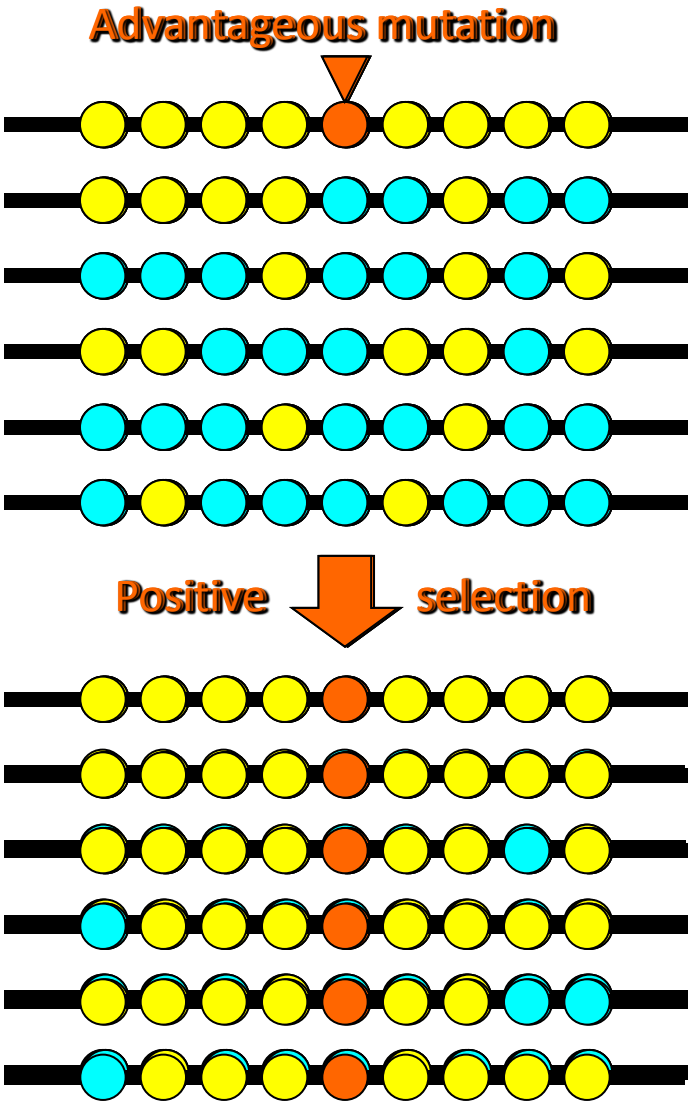
*D* > 0: 平衡選択が働いている場合
集団の構造化が存在する場合

*D* = 0: 中立で有効集団サイズが一定

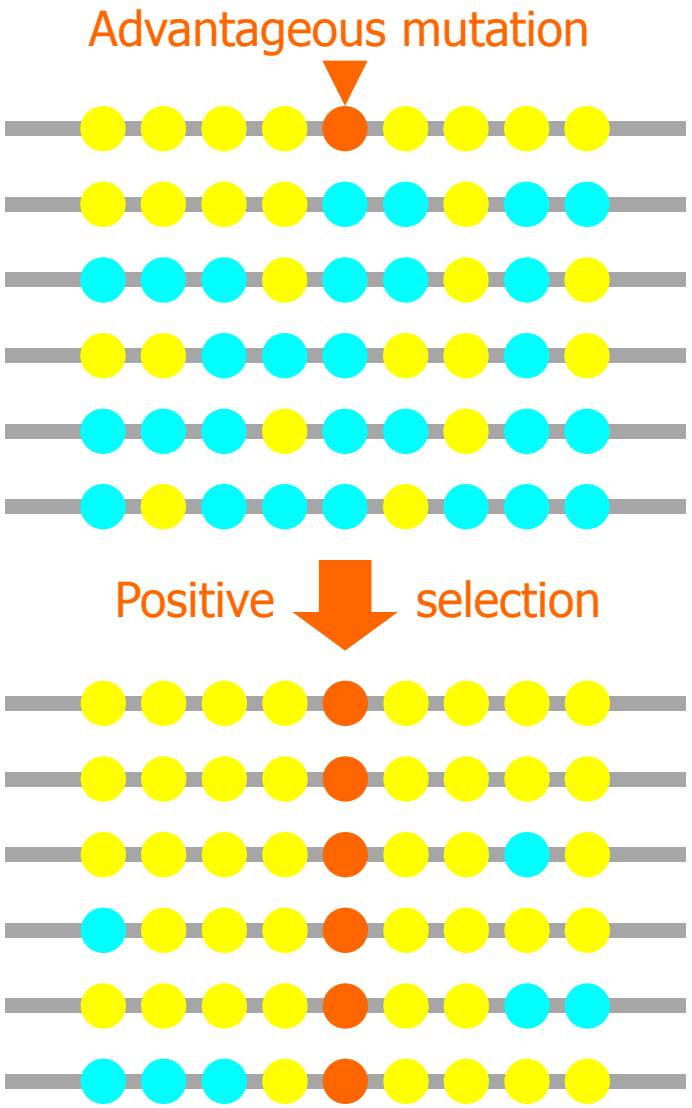
Tajima's Dでは組み替えは考慮されない

Bamshad&Wooding 2003

# Hitchhiking and selective sweep

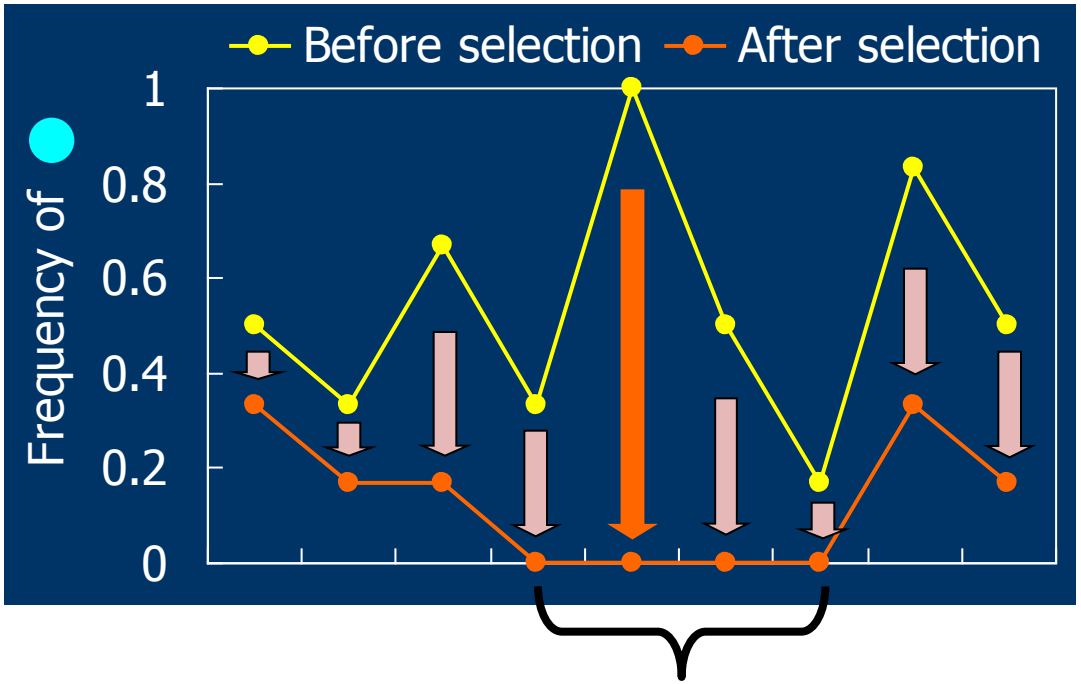


# Hitchhiking and selective sweep




**Hitchhiking**

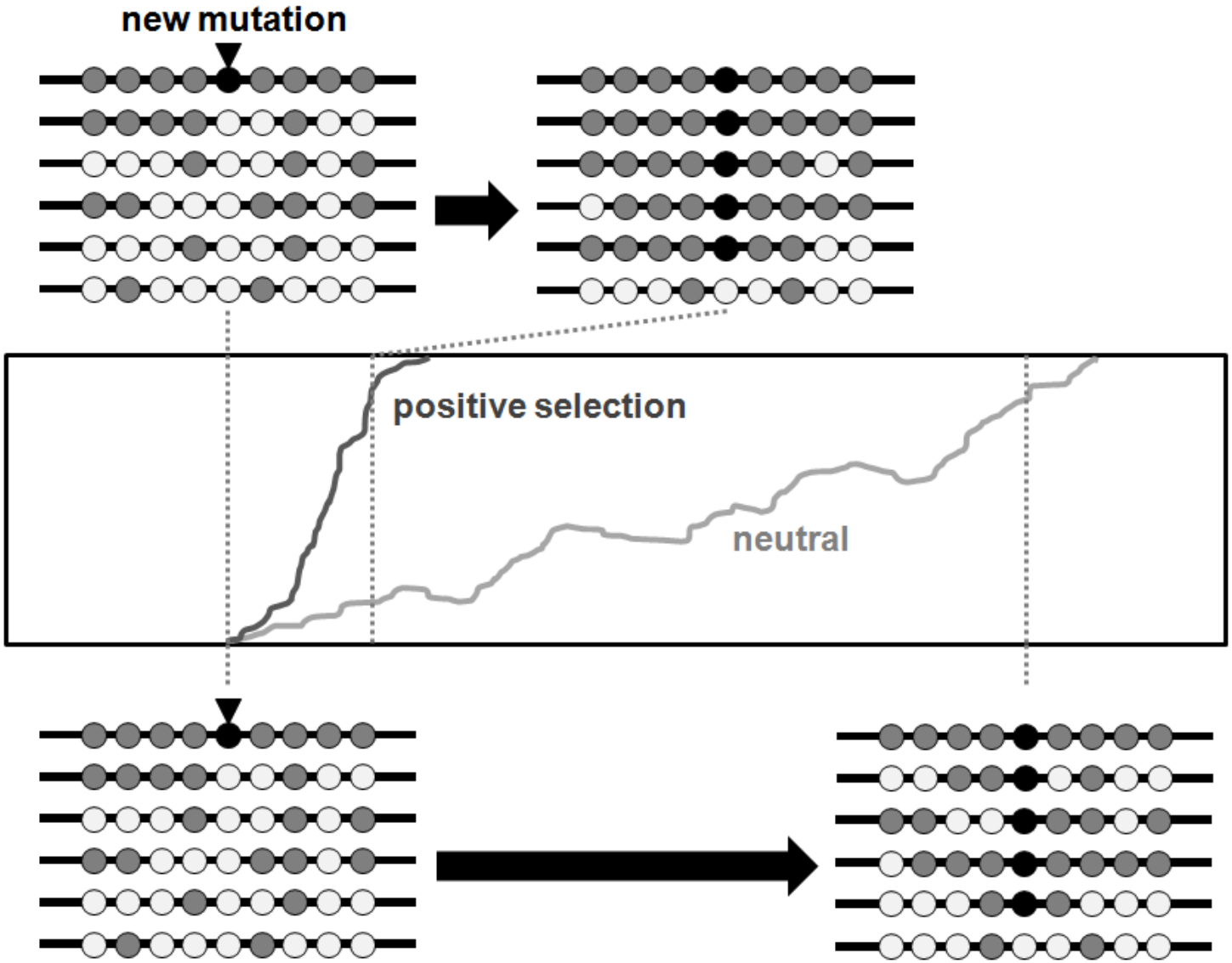
頻度がつられて動く




**Selective sweep**

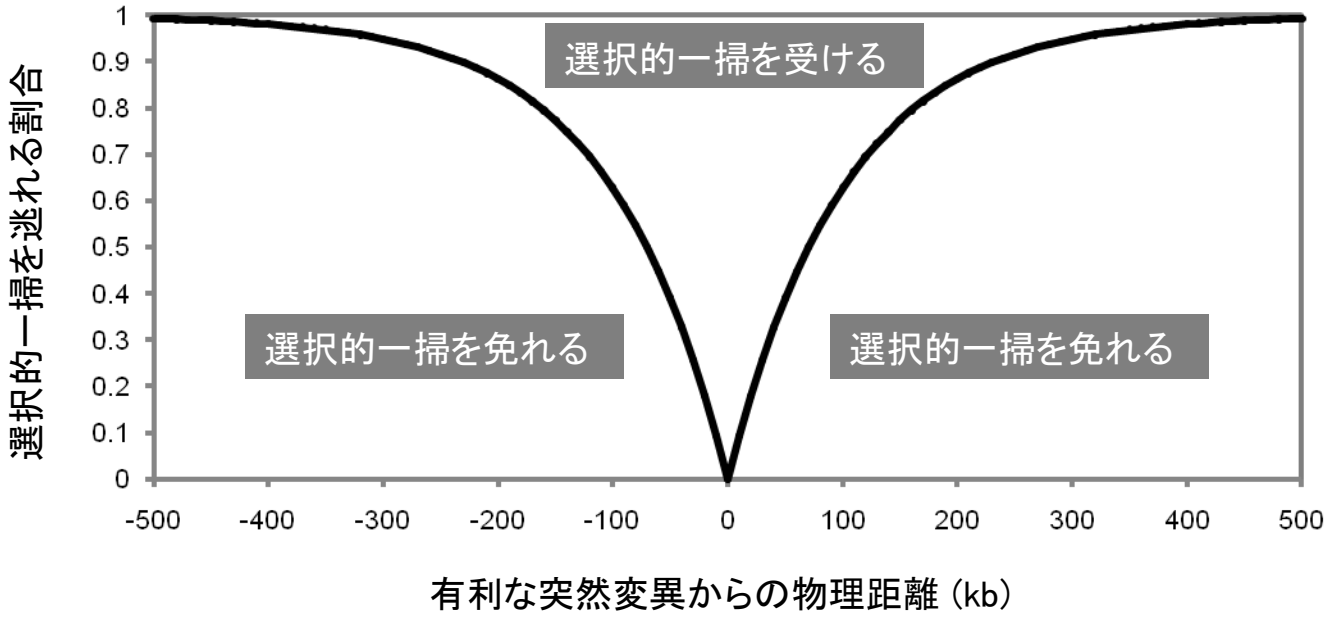
多様性が失われる

# Signatures of positive selection




Recombinationを考慮することにより、検出力が高くなる

# Selective sweepの領域




選択圧（s）と距離（d）もしくは組み換え率（r）の依存

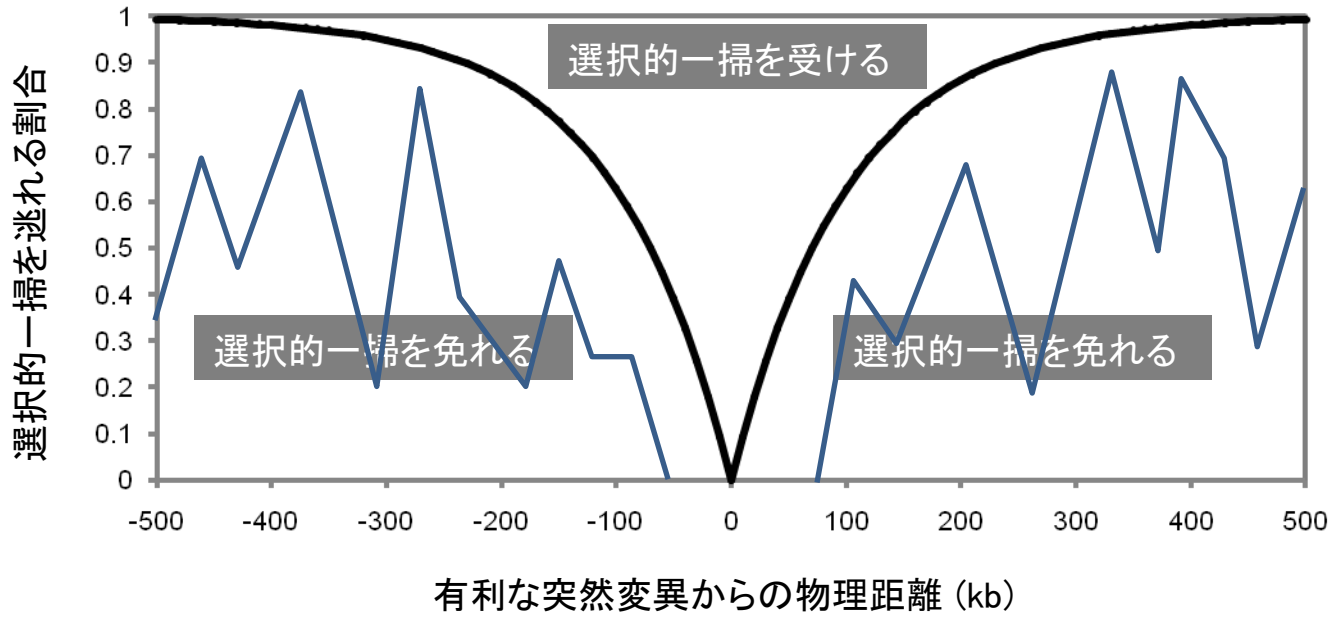
$$Q = (1 - p_0) \sum_{t=0}^{\infty} \frac{r(1-r)^t}{1 - p_0 + p_0(1+s)^{t+1}}$$

$p_0$ : Initial frequency of the beneficial allele

Q: Prop. of chrs under the effect of sweep

Maynard-Smith & Haigh 1974

# Composite likelihood ratio（CLR) test




SNPの位置と頻度の情報から、中立および自然選択それぞれの下でそのような頻度になる尤度を求め、composite likelihood（複合尤度）を算出し、その比で以って検定

Kim&Stephan 2002
Kim&Nielsen 2004
Nielsen et al 2005

# Composite likelihood ratio testによるゲノム探索

**Table 2.** Candidate Genes for Variation in Human Skin Pigmentation and Evidence of Population-Specific Selective Sweeps

| Gene | Chr | Position (Mb) | CLR *p* Value, African-American | CLR *p* Value, European-American | CLR *p* Value, Chinese |
|---|---|---|---|---|---|
| *POMC* | 2 | 25.36 | 0.654 (0.433) | 0.295 | 0.150 |
| *MITF* | 3 | 69.83 | 0.181 | 0.254 (0.182) | 0.658 (0.627) |
| *KIT* | 4 | 55.48 | 0.828 (0.813) | 0.618 | 0.301 |
| *F2r11* | 5 | 76.21 | 0.808 | 0.870 | 0.933 |
| *MATP* | 5 | 34.01 | 0.976 | **0.00014** | 0.658 |
| *DTNBP1*[a] | 6 | 15.70 | 0.913 (0.416) | 0.644 (0.599) | 0.037 |
| *TYRP1*[a] | 9 | 12.69 | 0.652 | 0.326 | 0.421 |
| *TYR* | 11 | 88.66 | 0.746 (0.725) | 0.145 (0.117) | 0.221 (0.209) |
| *SILV* | 12 | 54.64 | 0.092 | 0.050 | **0.007** |
| *KITLG* | 12 | 87.44 | **0.014** | **0.000007** | **0.00002** |
| *DCT* | 13 | 92.81 | 0.812 (0.796) | 0.335 | 0.305 |
| *OCA2* [a] | 15 | 25.77 | 0.400 **(0.046)** | 0.140 (0.055) | **0.020 (0.0023)** |
| *TRPM1* | 15 | 29.04 | 0.992 | 0.707 (0.689) | **0.00004 (0.00002)** |
| *SLC24A5*[a] | 15 | 46.14 | 0.287 | **0.0008** | 0.868 |
| *MYO5A*[a] | 15 | 50.43 | 0.382 | 0.492 (0.454) | 0.398 |
| *RAB27A* | 15 | 53.23 | 0.885 (0.814) | **0.0025** | **0.00020** |
| *MC1R* | 16 | 89.73 | 0.274 | 0.556 | 0.405 |
| *MC2R* | 18 | 13.88 | 0.839 | 0.125 | **0.0005** |
| *ATRN* | 20 | 35.19 | 0.613 | 0.608 (0.582) | **0.00020 (0.00006)** |
| *ASIP* | 20 | 33.57 | 0.518 | 0.749 | 0.375 |

Reported *p* values are from the genomic window with a midpoint nearest the midpoint of the gene.
Values in parentheses indicate the minimum *p* value of windows with a center between the start and stop codon of the gene, which is reported only if it is different from the midpoint *p* value. Bold typeface indicates *p* values with nominal significance below 5%.
[a]Genes previously identified as experiencing partial selective sweeps in the European population [15].
doi:10.1371/journal.pgen.0030090.t002

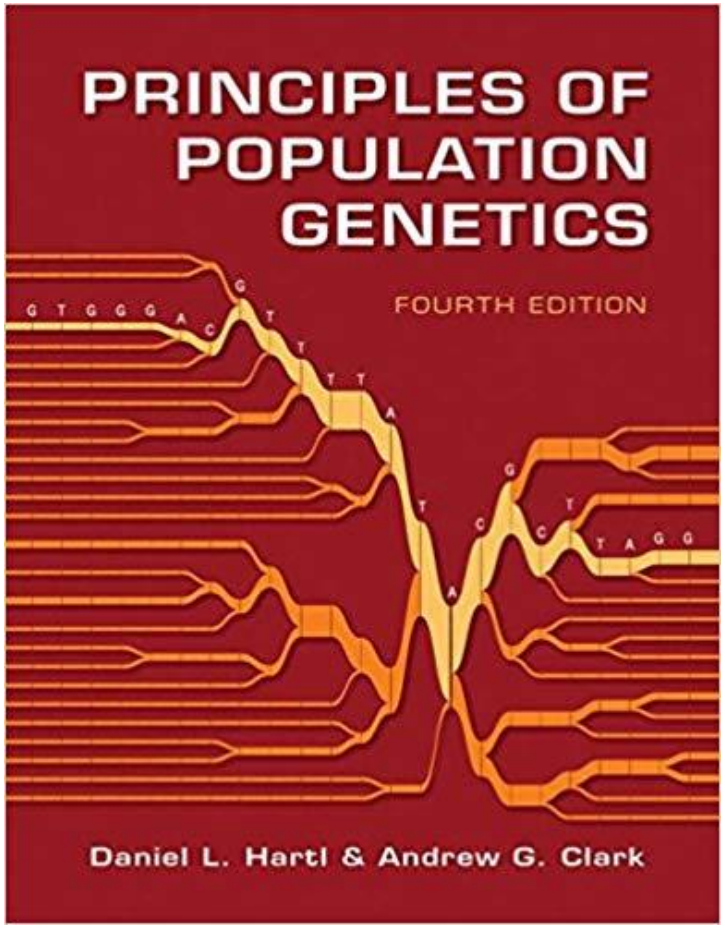
色素形成関連遺伝子における自然選択を検出

Williamson et al. 2007 PLoS Genet

# LRH test

## Detecting recent positive selection in the human genome from haplotype structure

Pardis C. Sabeti*†#, David E. Reich*, John M. Higgins*
Haninah Z. P. Levine*, Daniel J. Richter*, Stephen F. Schaffner*,
Stacey B. Gabriel*, Jill V. Platko*, Nick J. Patterson*, Gavin J. McDonald*,
Hans C. Ackerman‡, Sarah J. Campbell‡, David Altshuler*§,
Richard Cooper‖, Dominic Kwiatkowski‡, Ryk Ward† & Eric S. Lander*¶



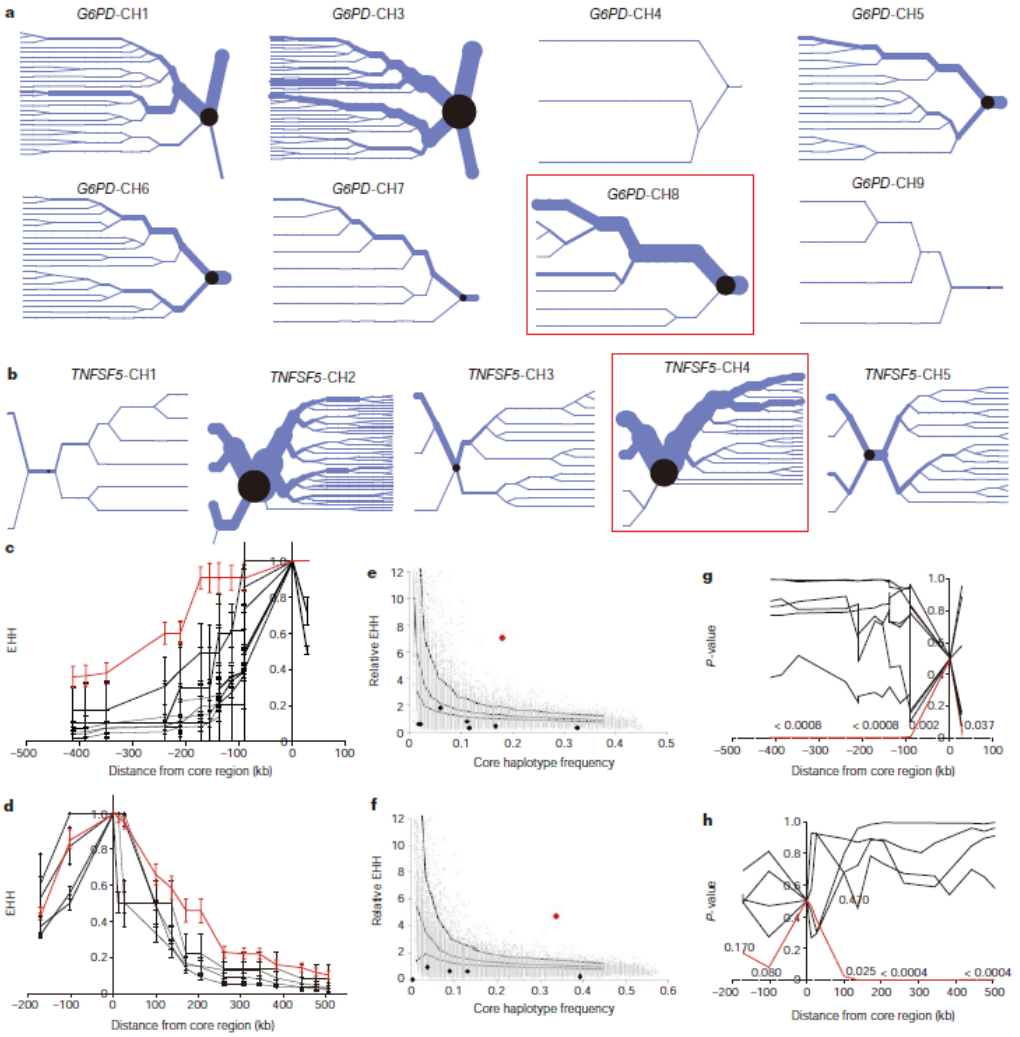




**Figure 2** Core haplotype frequency and relative EHH of *G6PD* and *TNFSF5*. **a**, **b**, Haplotype bifurcation diagrams (see Methods) for each core haplotype at *G6PD* (**a**) and *TNFSF5* (**b**) in pooled African populations demonstrate that *G6PD*-CH8 and *TNFSF5*-CH4 (boxed or labelled in red) have long-range homozygosity that is unusual given their frequency. **c**, **d**, The EHH at varying distances from the core region on each core haplotype at *G6PD* (**c**) and *TNFSF5* (**d**) demonstrates that *G6PD*-CH8 and *TNFSF5*-CH4 have persistent, high EHH values. **e**, **f**, At the most distant SNP from *G6PD* (**e**) and *TNFSF5* (**f**) core regions, the relative EHH plotted against the core haplotype frequency is presented and compared with the distribution of simulated core haplotypes (on the basis of simulation of 5,000 data sets; represented by grey dots and given with 95th, 75th and 50th percentiles). The observed non-selected core haplotypes in our data are represented by black diamonds. **g**, **h**, We calculated the statistical significance of the departure of the observed data from the simulated distribution at each distance from the core. *G6PD*-CH8 (**g**) and *TNFSF5*-CH4 (**h**) demonstrate increasing deviation from a model of neutral drift at further distances from the core region in both directions.

Sabeti et al. 2002

# Extended haplotype homozygosity（EHH）

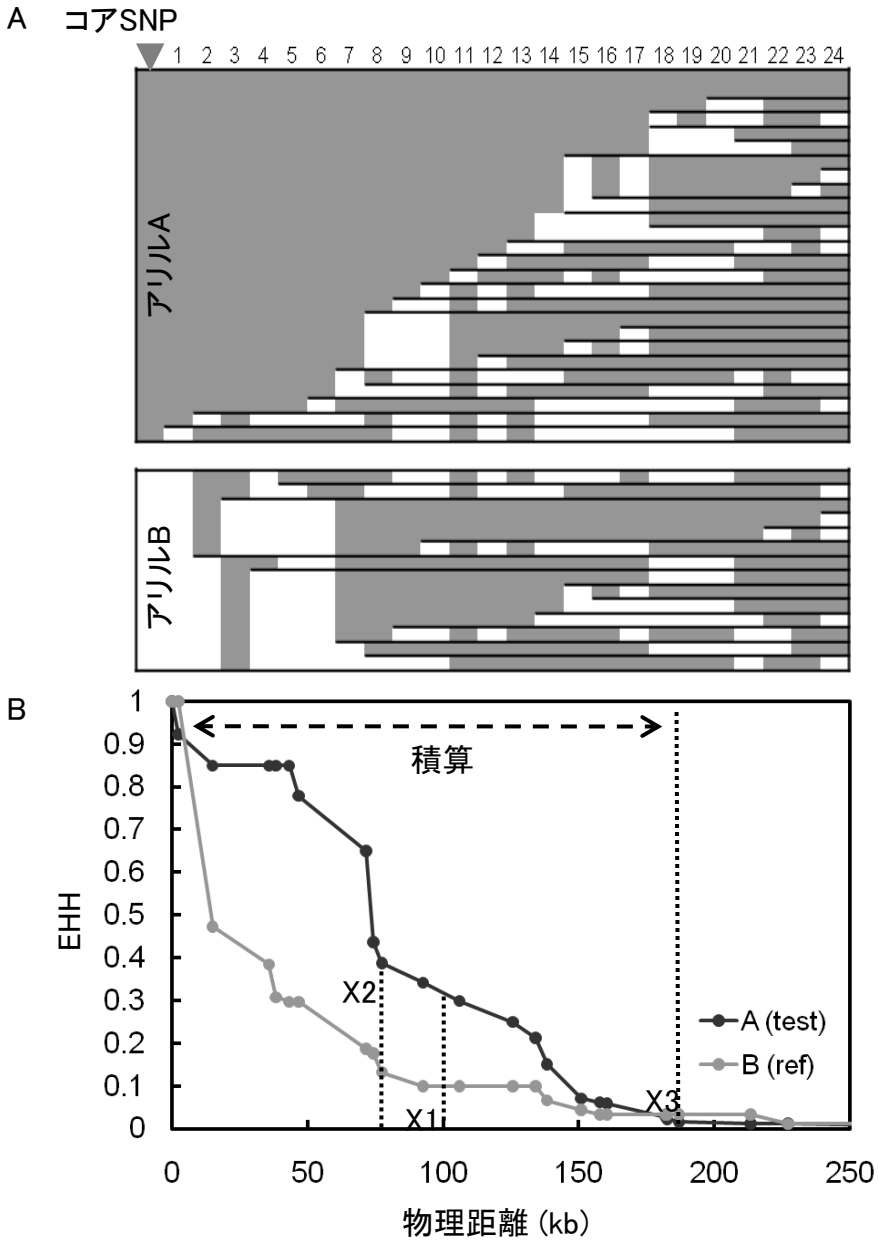
Extended haplotype homozygosity（EHH）の不偏推定値

$$\mathrm{EHH}_A = \frac{\sum_{i=1}^{k} e_{Ai}(e_{Ai}-1)}{n_A(n_A-1)}$$

- コアSNPのアリル*A*をもつ染色体の総数$n_A$
- ハプロタイプの数$k$
- *i*番目のハプロタイプをもつ染色体の数$e_{Ai}$
（$n_A = e_{A1} + e_{A2} + \cdots + e_{Ak}$）

EHHをテストアリル(EHHt)とリファレンスアリル(EHHr)の間で比較
⇒両者の比(REHH)
⇒ローカルな組み替え率をコントロール

REHHを算出するためのEHHの定義は様々
例えば
- コアSNPから100kbの位置（X1）
- 全体のHHが0.05を下回るところまでのEHHを積算（iHH） ⇒ iHS test
- EHHtが0.4以下に落ちる直前の位置（X2）
（物理距離bpや遺伝距離cMに依存しない）



# LRH testによるゲノムワイド探索

OPEN ACCESS Freely available online PLOS BIOLOGY

## A Map of Recent Positive Selection in the Human Genome

**Benjamin F. Voight, Sridhar Kudaravalli, Xiaoquan Wen, Jonathan K. Pritchard***

Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America

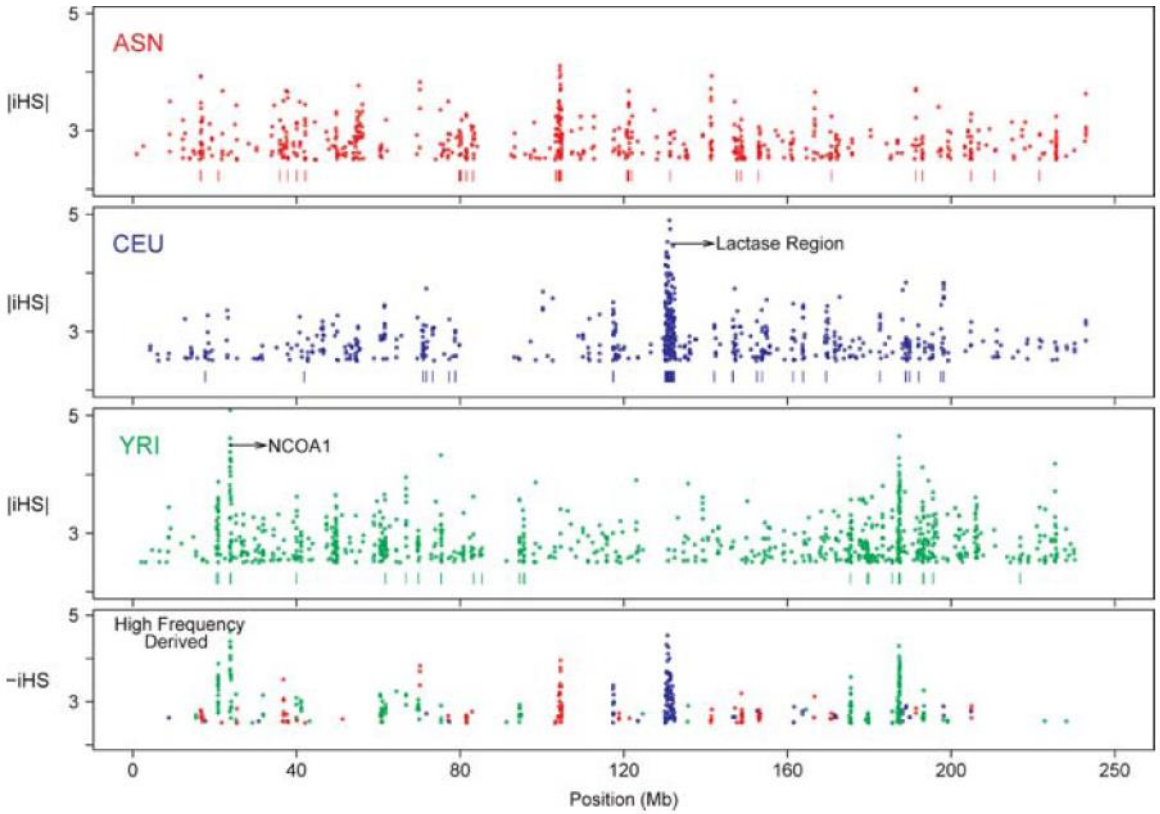



**Figure 3.** Plots of Chromosome 2 SNPs with Extreme iHS Values Indicate Discrete Clusters of Signals

Voight et al. 2006 PLoS Biol 4: e72

# Integrated Haplotype Score (iHS)

Voight et al. (2006)

- Core SNPからEHHが0.05に達するまでのEHH曲線の積分をIntegrated EHH (iHH)とする
- 祖先型アリルが乗るハプロタイプのiHHを$iHH_A$, 派生型アリルが乗るハプロタイプのiHHを $iHH_D$とすると

$$\text{unstandardized } iHS = \ln\left(\frac{iHH_A}{iHH_D}\right).$$

- 中立下では低頻度アリルは一般的に新しく、長いハプロタイプに乗っているため、iHHは頻度に依存。
- 同じ頻度のSNPにおける分布を用いて標準化

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{\mathrm{SD}_p\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}$$

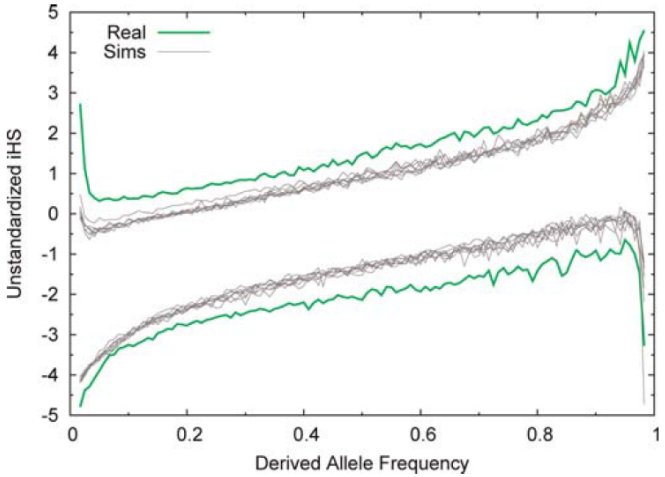



**Figure 4.** Central 99% Range of Unstandardized iHS for SNPs in the Yoruba Data and for SNPs in Matched Neutral Simulations

The upper and lower lines mark the boundaries of the central 99% distribution of the unstandardized iHS ratio, as a function of derived allele frequency. The gray lines plot results for a range of plausible demographic models. The fatter tails in the real data are consistent with the action of selection.

DOI: 10.1371/journal.pbio.0040072.g004

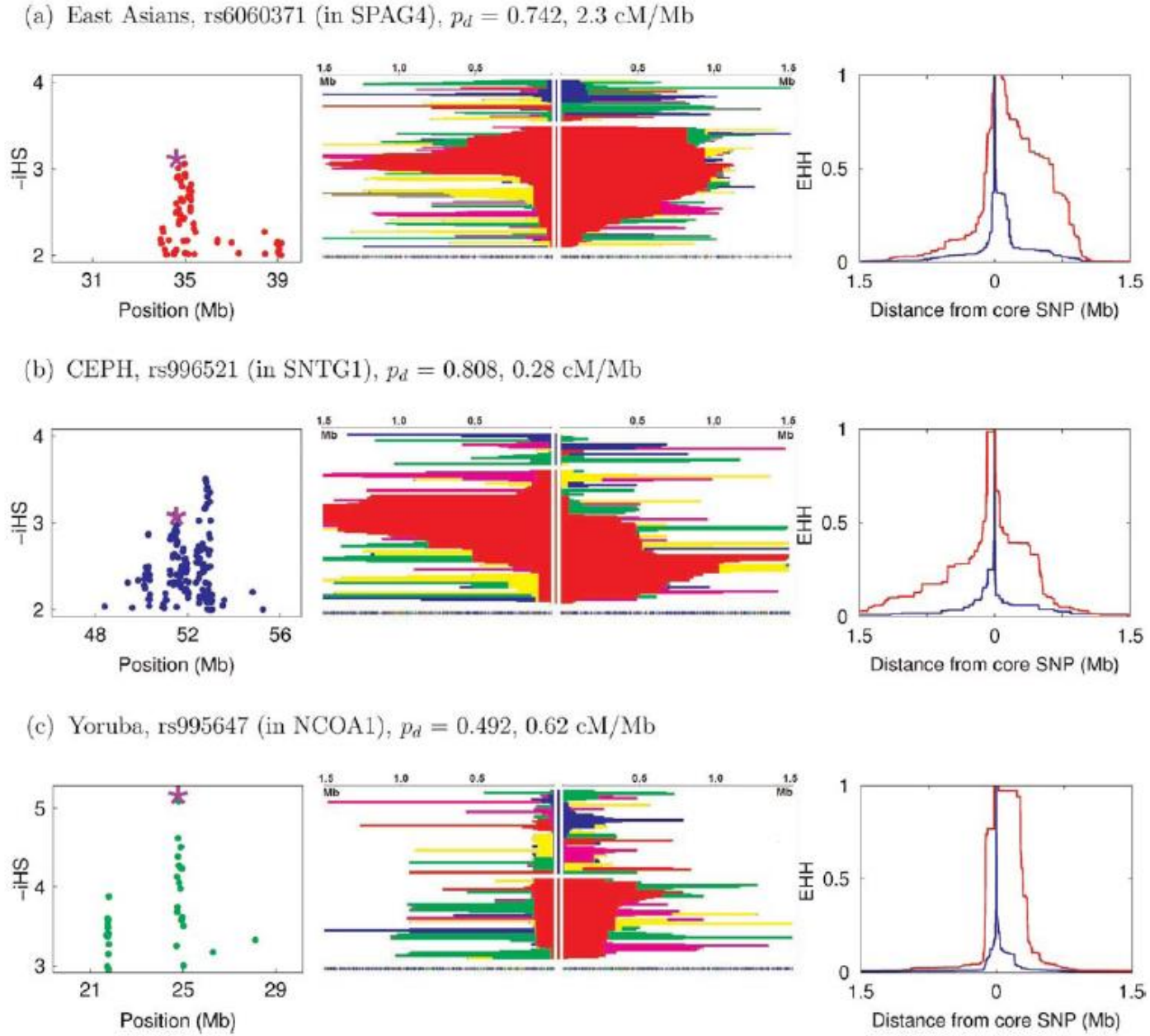



**Figure 6.** Signals of Selection for Three Candidate Selection Regions Discussed in the Text

The columns show (left) scatter plots of negative iHS scores, (center) haplotype plots, and (right) decay of haplotype homozygosity. In each case the core SNP for the center and right-hand plots was chosen as a SNP with high negative iHS score (starred in the scatter plots); the allele marked in red is derived. For each signal, values are listed for the derived allele frequency ($p_d$) and the local deCode recombination rate estimate.

DOI: 10.1371/journal.pbio.0040072.g006

# GO enrichment

**Table 2.** *p*-Values for Enrichment of GO Categories among Genes Showing Evidence for Partial Sweeps

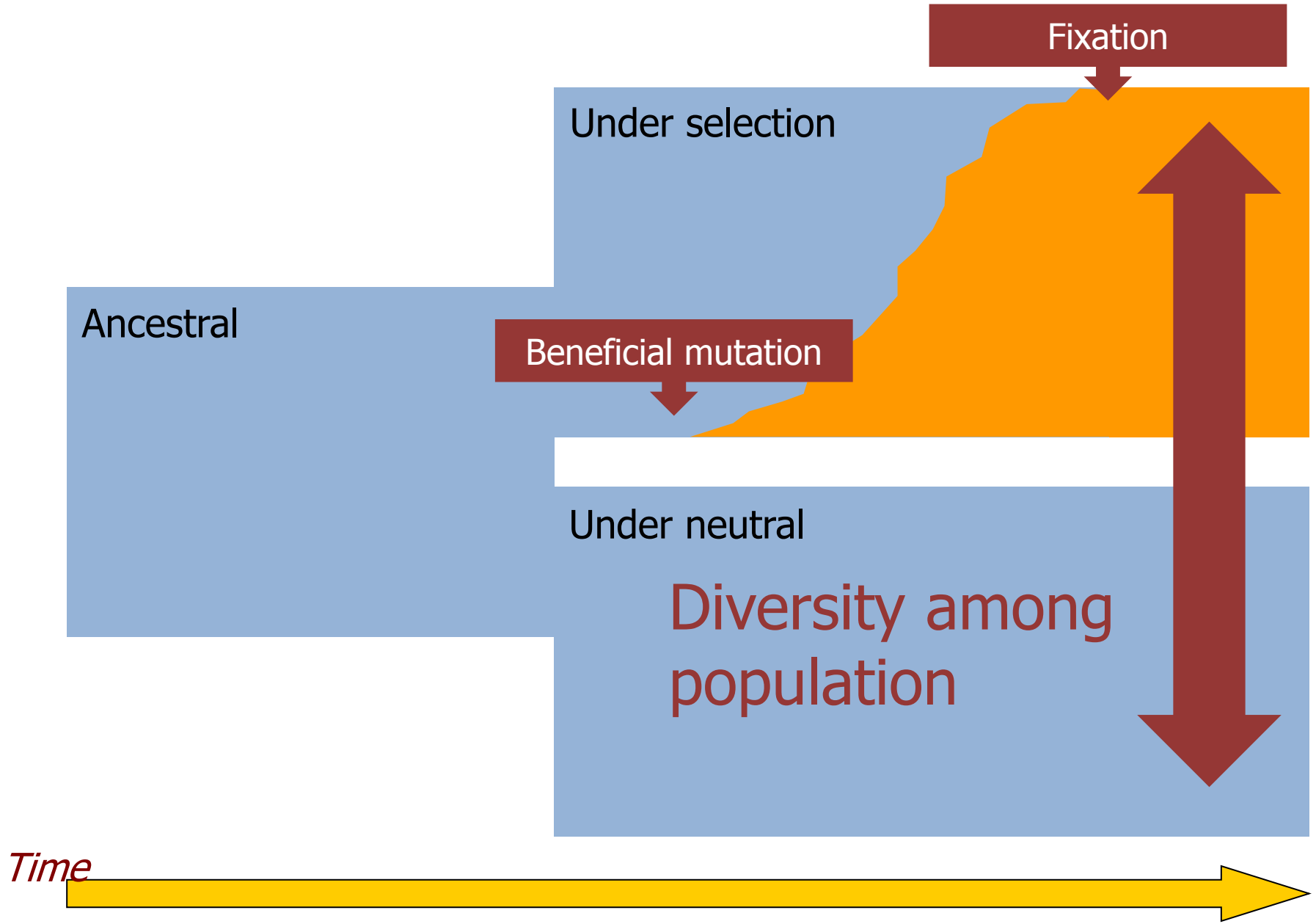
| GO Nesting | GO Category | ASN | CEU | YRI |
|---|---|---|---|---|
| 21–1 | Chemosensory perception | – | 0.0006 | 0.0004 |
| 21–1–1 | Olfaction | – | 0.0006 | 0.0008 |
| 22–2 | Gametogenesis | 0.008 | – | – |
| 22–2–2 | Spermatogenesis and motility | 0.02 | 0.03 | – |
| 22–3 | Fertilization | 0.004 | 0.003 | – |
| 1–11 | Other carbohydrate metabolism | **0.0002** | – | – |
| 6 | Electron transport | – | **0.0002** | – |
| 4–13 | Chromatin packaging/remodeling | **<0.0001** | 0.01 | – |
| 16–1–1 | MHC-I–mediated immunity | – | **<0.0001** | 0.02 |
| 3–2 | Steroid metabolism | – | – | **<0.0001** |
| 3–5 | Lipid and fatty acid binding | 0.001 | – | – |
| 4–4–2 | mRNA transcription initiation | – | 0.002 | – |
| 5–3 | Protein modification | 0.002 | – | – |
| 7–5 | Vitamin/cofactor transport | 0.002 | – | – |
| 9 | Phosphate metabolism | 0.002 | 0.03 | – |
| 13–4 | Peroxisome transport | – | – | 0.002 |

All *p*-values are one-sided, testing for enrichment of categories in each population; "–" indicates that the *p*-value is >0.05.

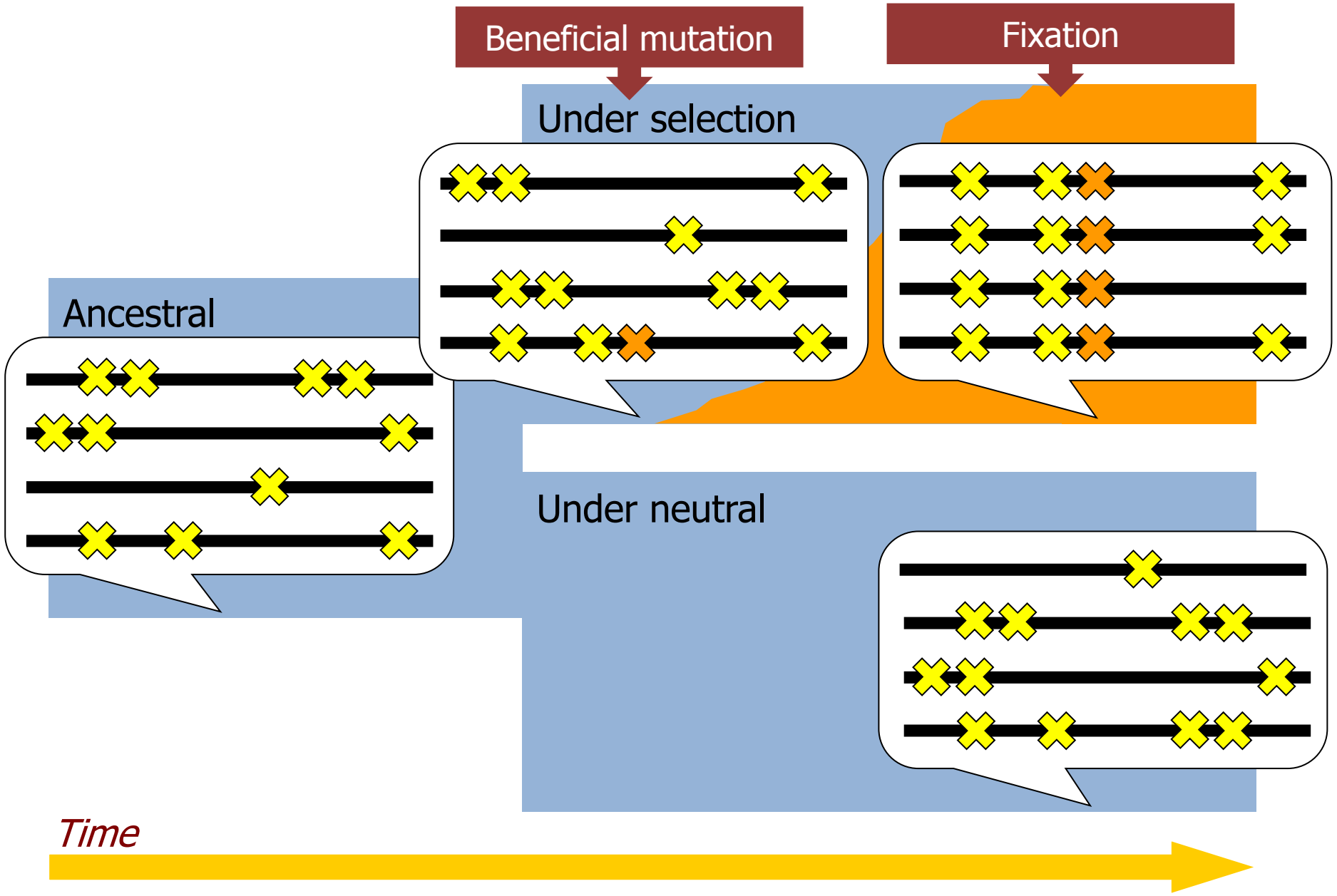
Boldfaced text indicates *p*-values that are significant after a conservative Bonferroni correction for multiple testing.

DOI: 10.1371/journal.pbio.0040072.t002

# population-specific selection



# population-specific selection

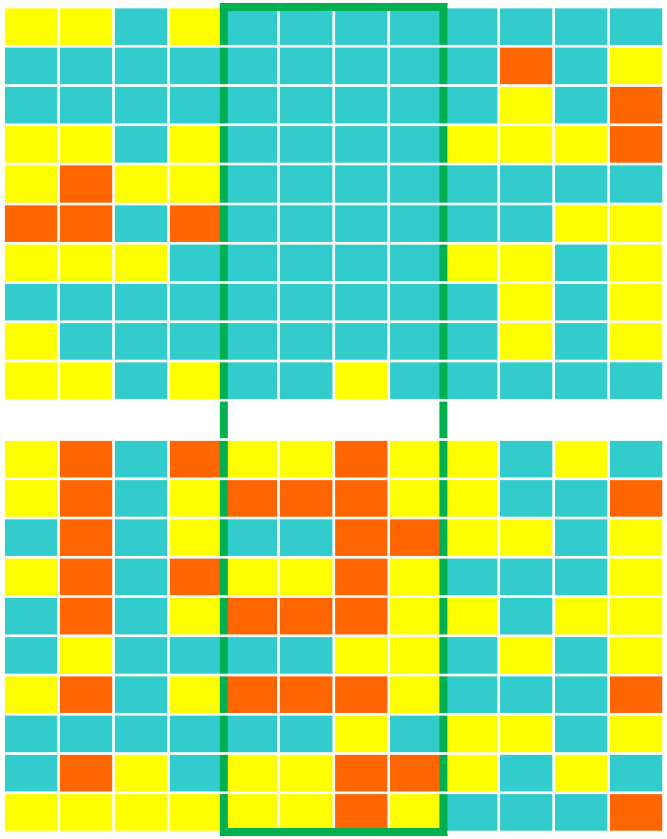


# 集団間でハプロタイプ多様性を比較

SNPs

Pop A under selection

Pop B under neutral



Homo for allele1

Hetero

Homo for allele2

OPEN ACCESS Freely available online PLoS one

## A Practical Genome Scan for Population-Specific Strong Selective Sweeps That Have Reached Fixation

Ryosuke Kimura[1,2]*, Akihiro Fujimoto[1], Katsushi Tokunaga[1], Jun Ohashi[1]

1 Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, 2 Japan Society for the Promotion of Science, Tokyo, Japan

Phenotypic divergences between modern human populations have developed as a result of genetic adaptation to local environments over the past 100,000 years. To identify genes involved in population-specific phenotypes, it is necessary to detect sig
(LD) has b
fixed loci.
this meth
between p
statistics u
density of
determine
method w
more effic
selection
likely to c

Citation: Ki
Reached Fix

1 Department of Evo
of California Irvine, I

OPEN ACCESS Freely available online PLoS BIOLOGY

## A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome

Kun Tang[1]*, Kevin R. Thornton[2], Mark Stoneking[1]

Vol 449|18 October 2007|doi:10.1038/nature06250 nature

Genome-wide
understanding
aimed at thos
the extended
single nucleot
detected wide
sweeps. A cha
neutral effect
and the lack
robust across
genome clear
towards stron
revealed seve
phenotypic tra

Citation: Tang K, The
doi:10.1371/journal.

LETTERS

## Genome-wide detection and characterization of positive selection in human populations

Pardis C. Sabeti[1]*, Patrick Varilly[1]*, Ben Fry[1], Jason Lohmueller[1], Elizabeth Hostetter[1], Chris Cotsapas[1,2], Xiaohui Xie[1], Elizabeth H. Byrne[1], Steven A. McCarroll[1,2], Rachelle Gaudet[3], Stephen F. Schaffner[1], Eric S. Lander[1,4,5,6] & The International HapMap Consortium†

With the advent of dense maps of human genetic variation, it is now possible to detect positive natural selection across the human genome. Here we report an analysis of over 3 million polymorphisms from the International HapMap Project Phase 2 (HapMap2)[1]. We used 'long-range haplotype' methods, which were developed to identify alleles segregating in a population that have undergone recent selection[2], and we also developed new methods that are based on cross-population comparisons to discover alleles that have swept to near-fixation within a population. The analysis reveals more than 300 strong candidate regions. Focusing on the strongest 22 regions, we develop a heuristic for scrutinizing these regions to identify candidate targets of selection. In a complementary analysis, we identify 26 non-synonymous, coding, single nucleotide polymorphisms showing regional evidence of positive selection. Examination of these candidates highlights three cases in which two genes in a common biological process have apparently undergone positive selection in the same population: *LARGE* and *DMD*, both related to infection by the Lassa virus[3], in West Africa; *SLC24A5* and *SLC45A2*, both involved in skin pigmentation[4,5], in Europe; and *EDAR* and *EDA2R*, both involved in development of hair follicles[6], in Asia.

few alternative alleles in the population (Supplementary Fig. 2 and Supplementary Tables 1–2).

We next developed, evaluated and applied a new test, Cross Population Extended Haplotype Homozygosity (XP-EHH), to detect selective sweeps in which the selected allele has approached or achieved fixation in one population but remains polymorphic in the human population as a whole (Methods, and Supplementary Fig. 2 and Supplementary Tables 3–6). Related methods have recently also been described[10–12].

Our analysis of recent positive selection, using the three methods, reveals more than 300 candidate regions[1] (Supplementary Fig. 3 and Supplementary Table 7), 22 of which are above a threshold such that no similar events were found in 10 Gb of simulated neutrally evolving sequence (Methods). We focused on these 22 strongest signals (Table 1), which include two well-established cases, *SLC24A5* and *LCT*[5,13], and 20 other regions with signals of similar strength.

The challenge is to sift through genetic variation in the candidate regions to identify the variants that were the targets of selection. Our candidate regions are large (mean length, 815 kb; maximum length, 3.5 Mb) and often contain multiple genes (median, 4; maximum, 15). A typical region harbours ~400–4,000 common SNPs (minor allele frequency >5%), of which roughly three-quarters are represented in

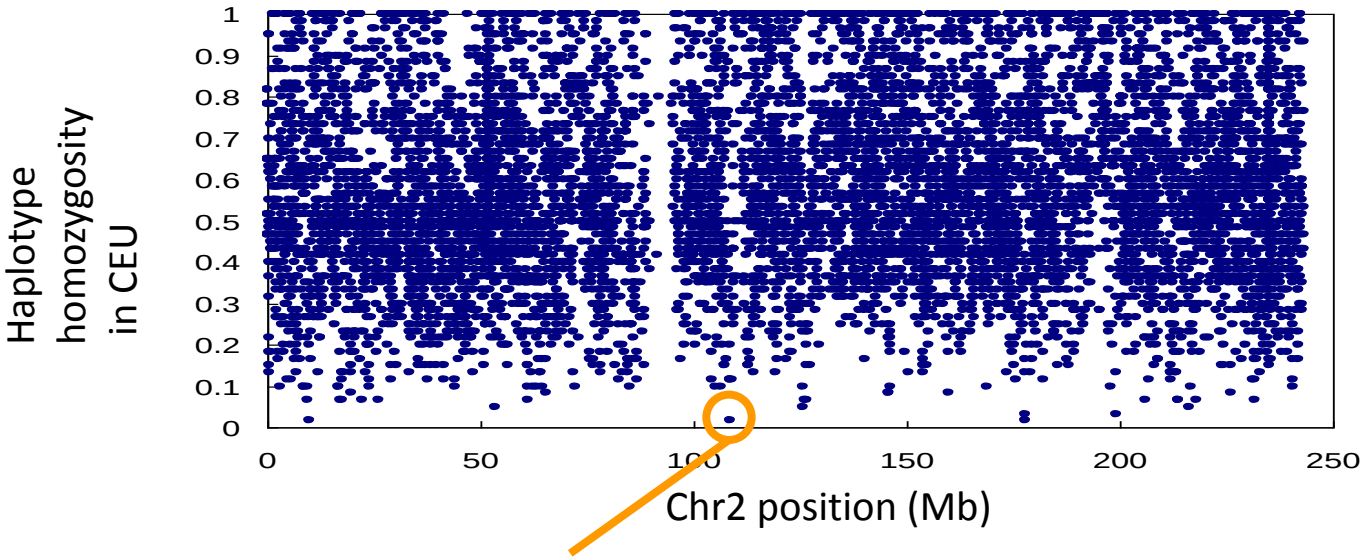
**Kimura et al. 2007**

Tang et al. 2007

Sabeti et al. 2007

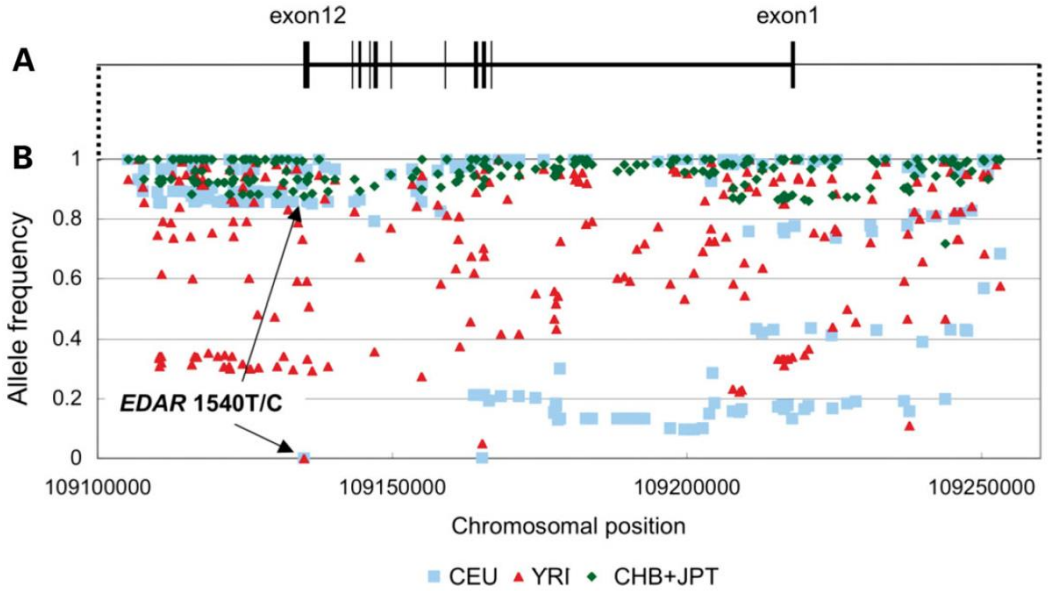
# *EDAR* 遺伝子領域のselective sweep

## Test population: EAS (Chr 2)　HH>0.5




*EDAR region*

アジアにおける
メジャーアリル頻度



# XP-EHH



## Genome-wide detection and characterization of positive selection in human populations

Pardis C. Sabeti[1]*, Patrick Varilly[1]*, Ben Fry[1], Jason Lohmueller[1], Elizabeth Hostetter[1], Chris Cotsapas[1,2], Xiaohui Xie[1], Elizabeth H. Byrne[1], Steven A. McCarroll[1,2], Rachelle Gaudet[3], Stephen F. Schaffner[1], Eric S. Lander[1,4,5,6] & The International HapMap Consortium†

同一アリルのEHHを
集団間で比較

**Table 1 | The twenty-two strongest candidates for natural selection**

| Region | Chr:position (MB, HG17) | Selected population | Long Haplotype Test | Size (Mb) | Total SNPs with Long Haplotype Signal | Subset of SNPs that fulfil criteria 1 | Subset of SNPs that fulfil criteria 1 and 2 | Subset of SNPs that fulfil criteria 1, 2 and 3 | Genes at or near SNPs that fulfil all three criteria |
|---|---|---|---|---|---|---|---|---|---|
| 1 | chr1:166 | CHB + JPT | LRH, iHS | 0.4 | 92 | 39 | 30 | 2 | *BLZF1, SLC19A2* |
| 2 | chr2:72.6 | CHB + JPT | XP-EHH | 0.8 | 732 | 250 | 0 | 0 | |
| 3 | chr2:108.7 | CHB + JPT | LRH, iHS, XP-EHH | 1.0 | 972 | 265 | 7 | 1 | *EDAR* |
| 4 | chr2:136.1 | CEU | LRH, iHS, XP-EHH | 2.4 | 1,213 | 282 | 24 | 3 | *RAB3GAP1, R3HDM1, LCT* |
| 5 | chr2:177.9 | CEU,CHB + JPT | LRH, iHS, XP-EHH | 1.2 | 1,388 | 399 | 79 | 9 | *PDE11A* |
| 6 | chr4:33.9 | CEU,YRI, CHB + JPT | LRH, iHS | 1.7 | 413 | 161 | 33 | 0 | |
| 7 | chr4:42 | CHB + JPT | LRH, iHS, XP-EHH | 0.3 | 249 | 94 | 65 | 6 | *SLC30A9* |
| 8 | chr4:159 | CHB + JPT | LRH, iHS, XP-EHH | 0.3 | 233 | 67 | 34 | 1 | |
| 9 | chr10:3 | CEU | LRH, iHS, XP-EHH | 0.3 | 179 | 63 | 16 | 1 | |
| 10 | chr10:22.7 | CEU, CHB + JPT | XP-EHH | 0.3 | 254 | 93 | 0 | 0 | |
| 11 | chr10:55.7 | CHB+JPT | LRH, iHS, XP-EHH | 0.4 | 735 | 221 | 5 | 2 | *PCDH15* |
| 12 | chr12:78.3 | YRI | LRH, iHS | 0.8 | 151 | 91 | 25 | 0 | |
| 13 | chr15:46.4 | CEU | XP-EHH | 0.6 | 867 | 233 | 5 | 1 | *SLC24A5* |
| 14 | chr15:61.8 | CHB + JPT | XP-EHH | 0.2 | 252 | 73 | 40 | 6 | *HERC1* |
| 15 | chr16:64.3 | CHB + JPT | XP-EHH | 0.4 | 484 | 137 | 2 | 0 | |
| 16 | chr16:74.3 | CHB + JPT, YRI | LRH, iHS | 0.6 | 55 | 35 | 28 | 3 | *CHST5, ADAT1, KARS* |
| 17 | chr17:53.3 | CHB + JPT | XP-EHH | 0.2 | 143 | 41 | 0 | 0 | |
| 18 | chr17:56.4 | CEU | XP-EHH | 0.4 | 290 | 98 | 26 | 3 | *BCAS3* |
| 19 | chr19:43.5 | YRI | LRH, iHS, XP-EHH | 0.3 | 83 | 30 | 0 | 0 | |
| 20 | chr22:32.5 | YRI | LRH | 0.4 | 318 | 188 | 35 | 3 | *LARGE* |
| 21 | chr23:35.1 | YRI | LRH, iHS | 0.6 | 50 | 35 | 25 | 0 | |
| 22 | chr23:63.5 | YRI | LRH, iHS | 3.5 | 13 | 3 | 1 | 0 | |
| | | Total SNPs | | 16.74 | 9,166 | 2,898 | 480 | 41 | |

Twenty-two regions were identified at a high threshold for significance (Methods), based on the LRH, iHS and/or XP-EHH test. Within these regions, we examined SNPs with the best evidence of being the target of selection on the basis of having a long haplotype signal, and by fulfilling three criteria: (1) being a high-frequency derived allele; (2) being differentiated between populations and common only in the selected population; and (3) being identified as functional by current annotation. Several candidate polymorphisms arise from the analysis including well-known *LCT* and *SLC24A5* (ref. 2), as well as intriguing new candidates.

Sabeti et al 2007

# いろんな統計量をまとめて評価

## A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection

Sharon R. Grossman,[1,2*†] Ilya Shylakhter,[1,2*†] Elinor K. Karlsson,[1,2] Elizabeth H. Byrne,[1,2] Shannon Morales,[1,2,3] Gabriel Frieden,[1] Elizabeth Hostetter,[1,2] Elaine Angelino,[1,4] Manuel Garber,[2] Or Zuk,[2] Eric S. Lander,[2,4,5] Stephen F. Schaffner,[2] Pardis C. Sabeti[1,2,4†]

The human genome contains hundreds of regions whose patterns of genetic variation indicate recent positive natural selection, yet for most the underlying gene and the advantageous mutation remain unknown. We developed a method, composite of multiple signals (CMS), that combines tests for multiple signals of selection and increases resolution by up to 100-fold. By applying CMS to candidate regions from the International Haplotype Map, we localized population-specific selective signals to 55 kilobases (median), identifying known and novel causal variants. CMS can not just identify individual loci but implicates precise variants selected by evolution.

$$CMS = \prod_{i=1}^{n} \frac{P(s_i|selected) \times \pi}{\left( \begin{array}{c} P(s_i|selected) \times \pi \\ + P(s_i|unselected) \times (1-\pi) \end{array} \right)} \quad (1)$$

A prior probability $\pi = 1/N_{SNP}$

$P(s_i|selected)$, $P(s_i|unselected)$はそれぞれシミュレーションから算出

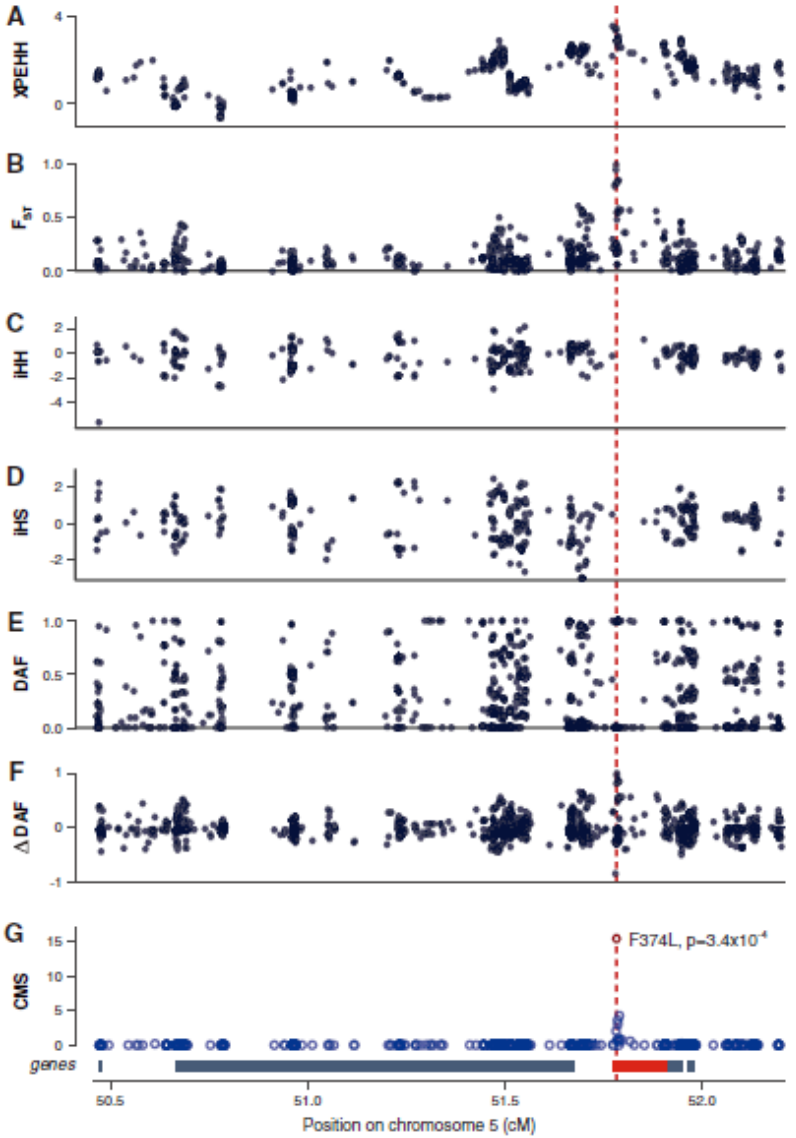
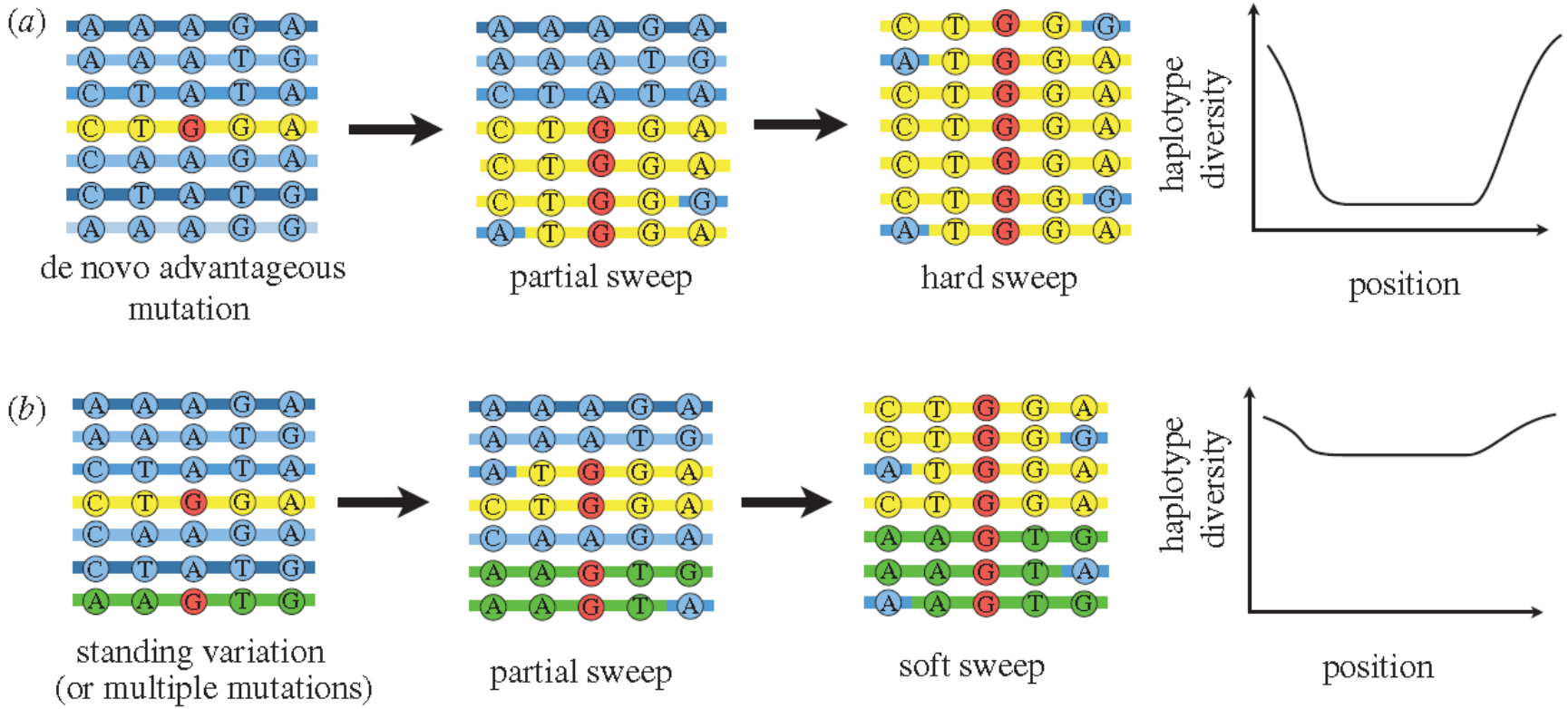



**Fig. 2.** Localizing selection at *MATP*. Scores of six individual tests (**A** to **F**) and CMS (**G**) for a region containing *MATP*. A nonsynonymous SNP [rs16891982, F374L ($Phe^{374}$→Leu), red dotted line] associated with pigmentation is believed to be the mutation under selection.

Grossman et al. 2010

# Hard Sweep & Soft Sweep




Novembre & Han 2012

# Empirical distribution vs Simulated distribution

Sweep候補領域についてP値やFDRを算出したい場合、用いた統計量の中立の下での分布を知る必要がある。

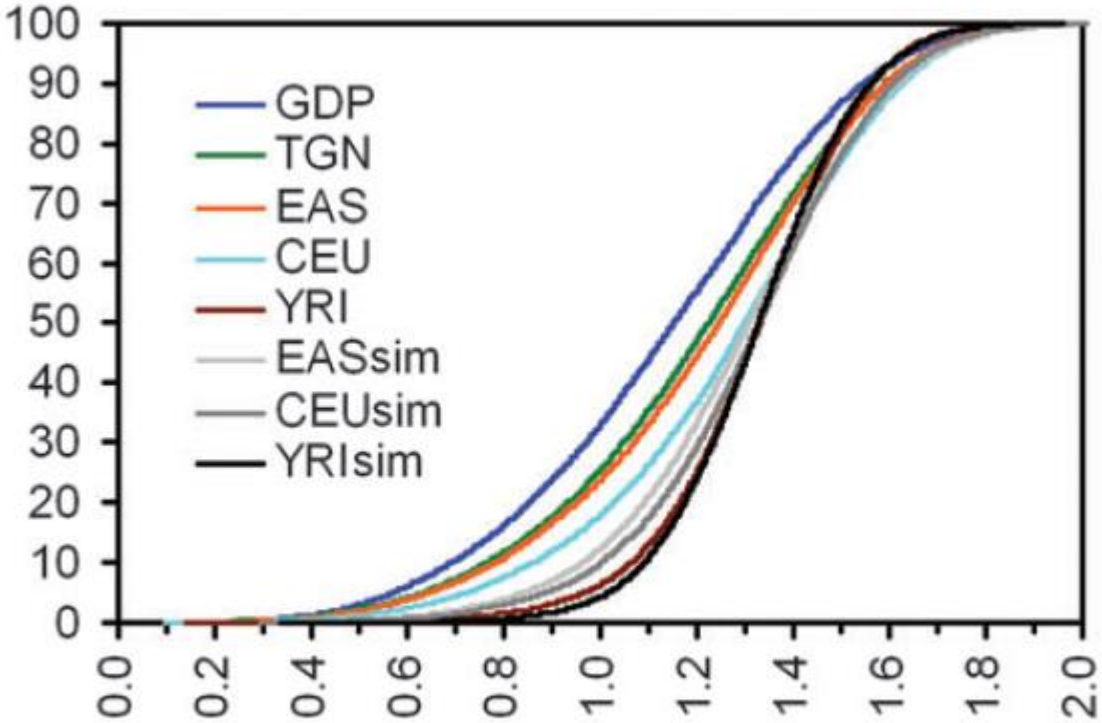
しかし、統計量の分布は分集団構造も含むデモグラフィに左右されるため、再現が非常に難しい。

そのため、

Outlier approach

を取らざる得ないことが多い。

有意水準は？




Average $EHH_R/EHH_T$ over adjacent 20 SNPs

# *The most famous examples of genes under strong local positive selection*

| Population | Gene | Chr | Function |
|---|---|---|---|
| East Asia | *ABCC11* | 16 | Ear wax type |
| | *ADH1B* | 20 | Alcohol metabolism |
| | *EDAR* | 2 | Ectodermal development |
| Europe | *SLC24A5* | 15 | Skin pigmentation |
| | *SLC45A2* | 5 | Skin pigmentation |
| | *LCT* | 2 | Lactose tolerance |
| Out-of-Africa | *OCA2* | 15 | Skin pigmentation |
| | *KITLG* | 12 | Skin pigmentation |

# GWSSの有効性

1．自然選択が働いた痕跡のある遺伝子を探索

2．集団間の表現型の差を担う遺伝子を同定

→場合によってはGWASより効率的

3．知られていない集団特異的適応形質を発掘

→一見でわからない耐病性や生理的形質など

→過去の感染症の流行などの手掛かりに

# 多遺伝子形質における中立性テスト

# 環境因子と遺伝子多型との相関

OPEN ACCESS Freely available online PLoS GENETICS

## Adaptations to Climate-Mediated Selective Pressures in Humans

Angela M. Hancock[1], David B. Witonsky[1], Gorka Alkorta-Aranburu[1], Cynthia M. Beall[2], Amha Gebremedhin[3], Rem Sukernik[4], Gerd Utermann[5], Jonathan K. Pritchard[1,6], Graham Coop[1,7], Anna Di Rienzo[1]*

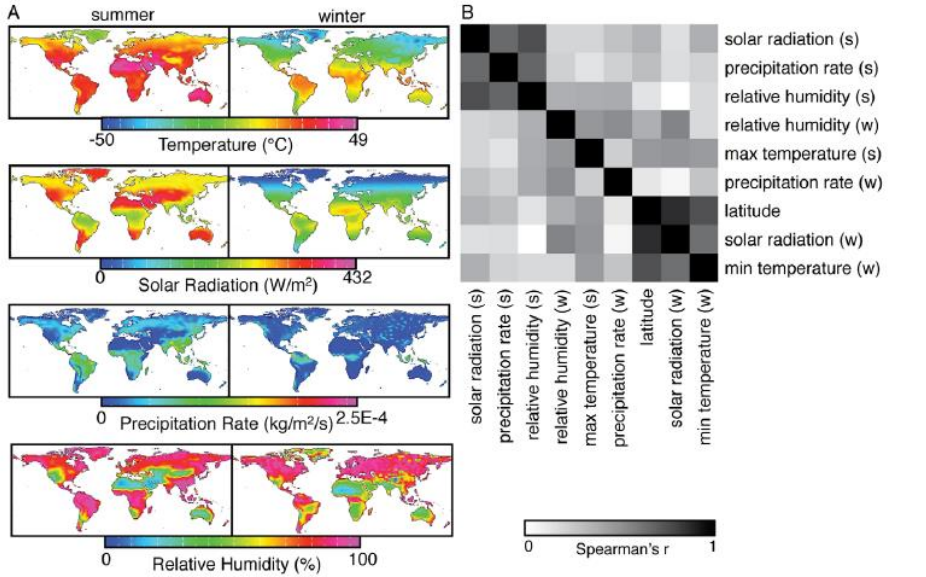



**Figure 1. Climate variables used for the analysis.** (A) Maps show the distributions of summer and winter climate variables: maximum summer temperature, minimum winter temperature and solar radiation, precipitation rate and relative humidity in the summer and winter. (B) A heatmap shows the absolute values of Spearman rank correlation coefficients between pairs of climate variables.
doi:10.1371/journal.pgen.1001375.g001

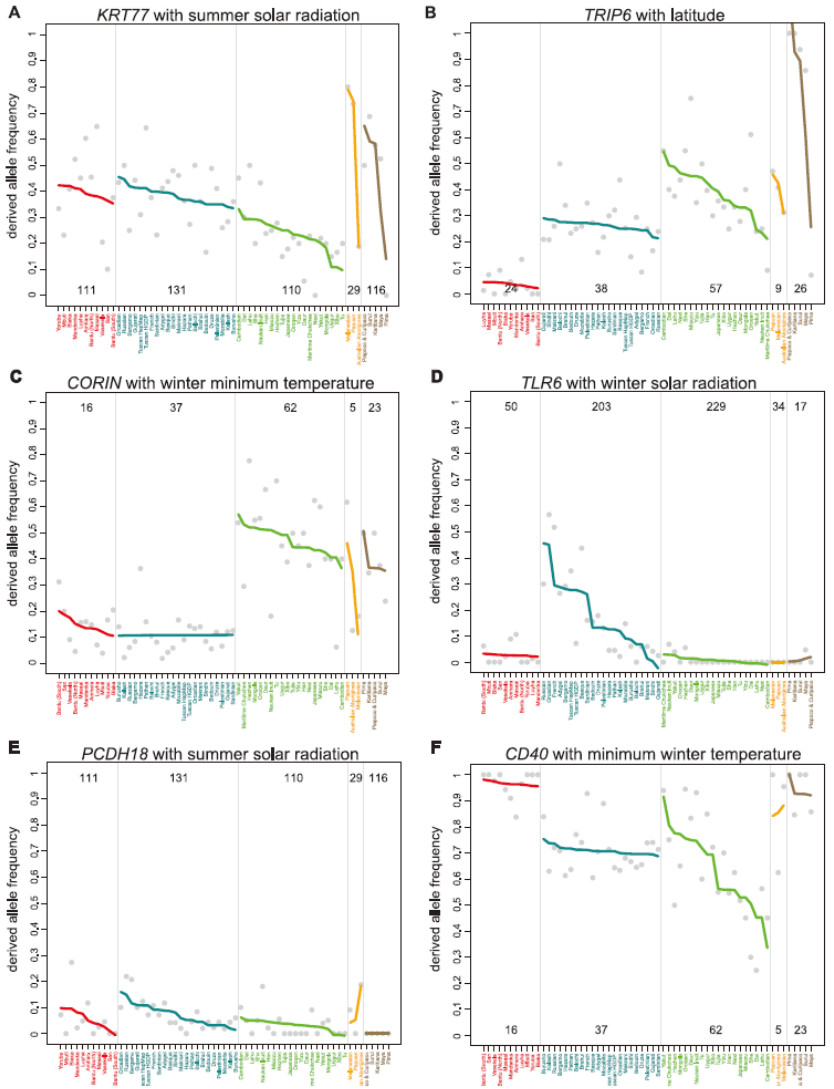



**Figure 3. Global variation in allele frequencies for SNPs with strong signals with climate.** Two NS SNPs from the worldwide analysis: (A) A SNP (rs3782489) in keratin 77 (*KRT77*), is strongly correlated with summer solar radiation, and (B) a SNP (rs2075756) in the thyroid receptor interacting protein (*TRIP6*) is strongly correlated with absolute latitude. Two SNPs from the population subset analysis: (C) A SNP (rs4558836) in *CORIN* has a signal in the AEA population subset with winter minimum temperature, but not in the AWE subset, and (D) a NS SNP (rs5743810) in TLR6 has a signal in the AWE population subset with winter solar radiation, but not in the AEA subset. Two SNPs that are associated with autoimmune disease from GWAS: (E) A SNP (rs2313132) upstream of *PCDH18* that is associated with SLE is strongly correlated with summer solar radiation, and (F) a SNP (rs6074022) upstream of *CD40* that is associated with multiple sclerosis is strongly correlated with minimum winter temperature. For each plot, gray points represent individual SNPs and colored lines represent fitted lines (obtained using the lm function in R) for each region. The ranges of the climate variable values for each region are shown at the bottom of the corresponding segment of the plot.
doi:10.1371/journal.pgen.1001375.g003

Hancock et al. 2011

# $Q_{ST}$ and $P_{ST}$

Additive genetic variances assuming infinite number of subpopulations

$$\sigma_{AT}^2 = (1+F)\sigma_o^2$$

$$\sigma_{AW}^2 = (1-F)\sigma_o^2$$

$$\sigma_{AB}^2 = \sigma_{AT}^2 - \sigma_{AW}^2 = 2F\sigma_o^2$$

$$Q_{ST} = \frac{\sigma_{AB}^2}{\sigma_{AB}^2 + 2\sigma_{AW}^2}$$

(Wright 1951; Lande 1992; Spitze 1993)

Using phenotypic variances,

$$Q_{ST} = \frac{a\sigma_{PB}^2}{a\sigma_{PB}^2 + 2h^2\sigma_{PW}^2}$$

$h^2$: heritability

$$\hat{\sigma}_{PW}^2 = \hat{s}_{PW}^2 = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{(n_k-1)}\sum_{j=1}^{n_k}\left(x_{kj} - \bar{x}_k\right)^2$$

$$\hat{\sigma}_{PB}^2 = \frac{K}{K-1}\hat{s}_{PB}^2 = \frac{1}{K-1}\sum_{k=1}^{K}(\bar{x}_k - \bar{x})^2$$

$K$: the number of subpopulations
$n_k$: the number of samples from $k$

Under the assumption that the dominance and epistatic effects and the environmental factors are the same between subpopulations (a = 1)

$$P_{ST} = \frac{\hat{\sigma}_{PB}^2}{\hat{\sigma}_{PB}^2 + 2h^2\hat{\sigma}_{PW}^2}$$

# QST/FSTテスト

OPEN ACCESS Freely available online

PLOS | GENETICS

## A Population Genetic Signal of Polygenic Adaptation

**Jeremy J. Berg[1,2,3]*, Graham Coop[2,3]***

**1** Graduate Group in Population Biology, University of California, Davis, Davis, California, United States of America, **2** Center for Population Biology, University of California, Davis, Davis, California, United States of America, **3** Department of Evolution and Ecology, University of California, Davis, Davis, California, United States of America

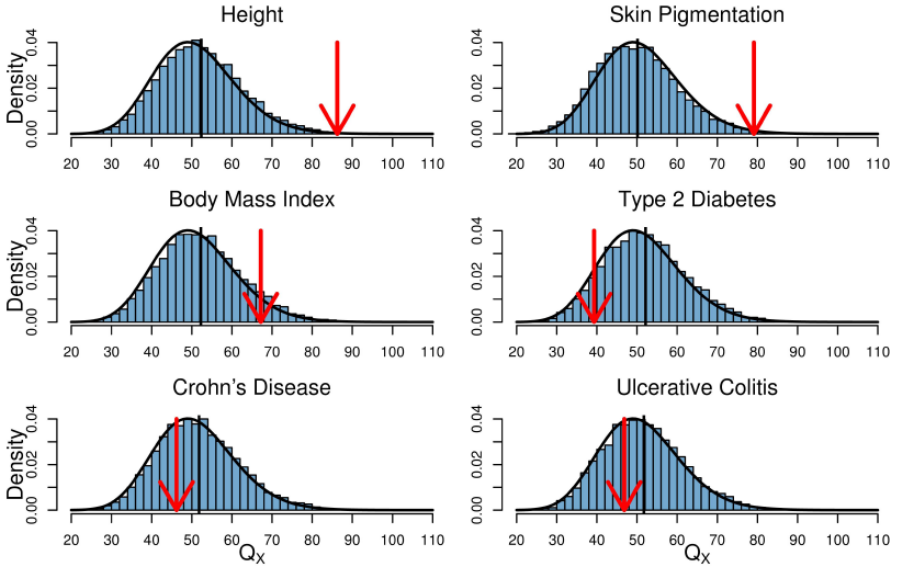



**Figure 3. Histogram of the empirical null distribution of $Q_X$ for each trait, obtained from genome-wide resampling of well matched SNPs.** The mean of each distribution is marked with a vertical black bar and the observed value is marked by a red arrow. The expected $\chi^2_{M-1}$ density is shown as a black curve.
doi:10.1371/journal.pgen.1004412.g003

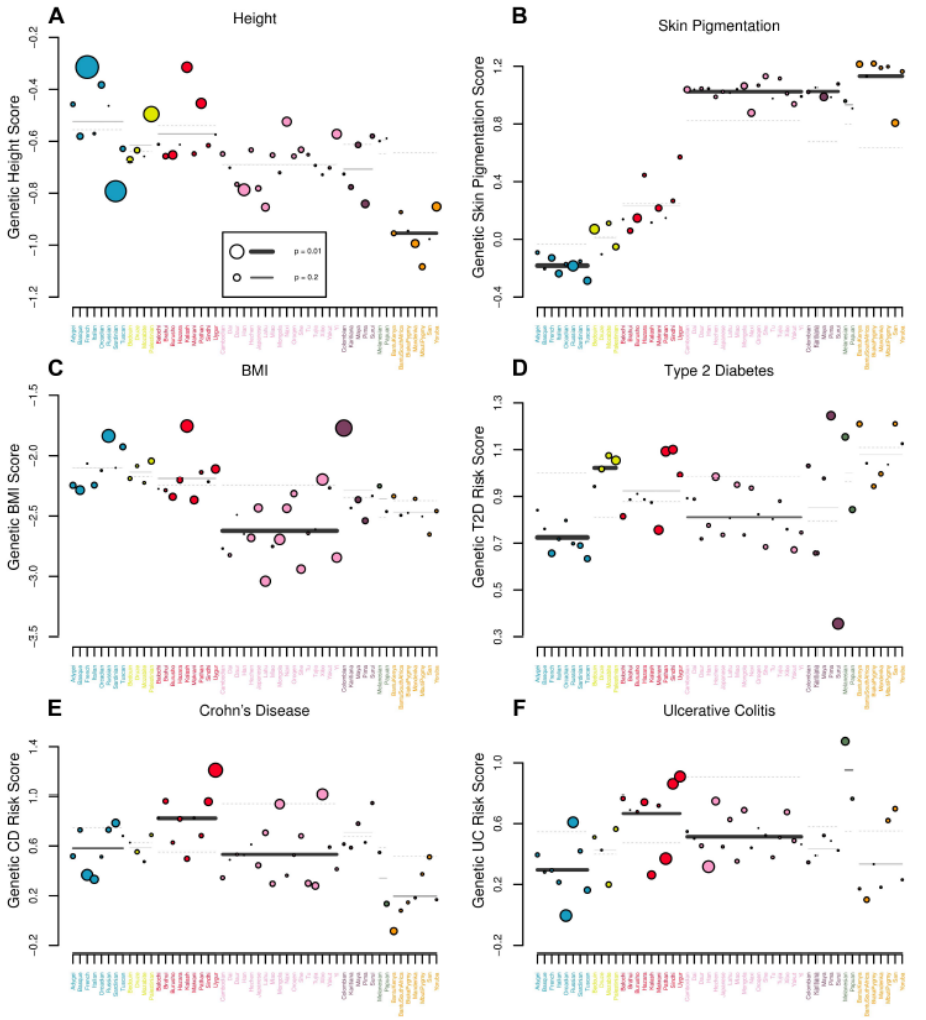
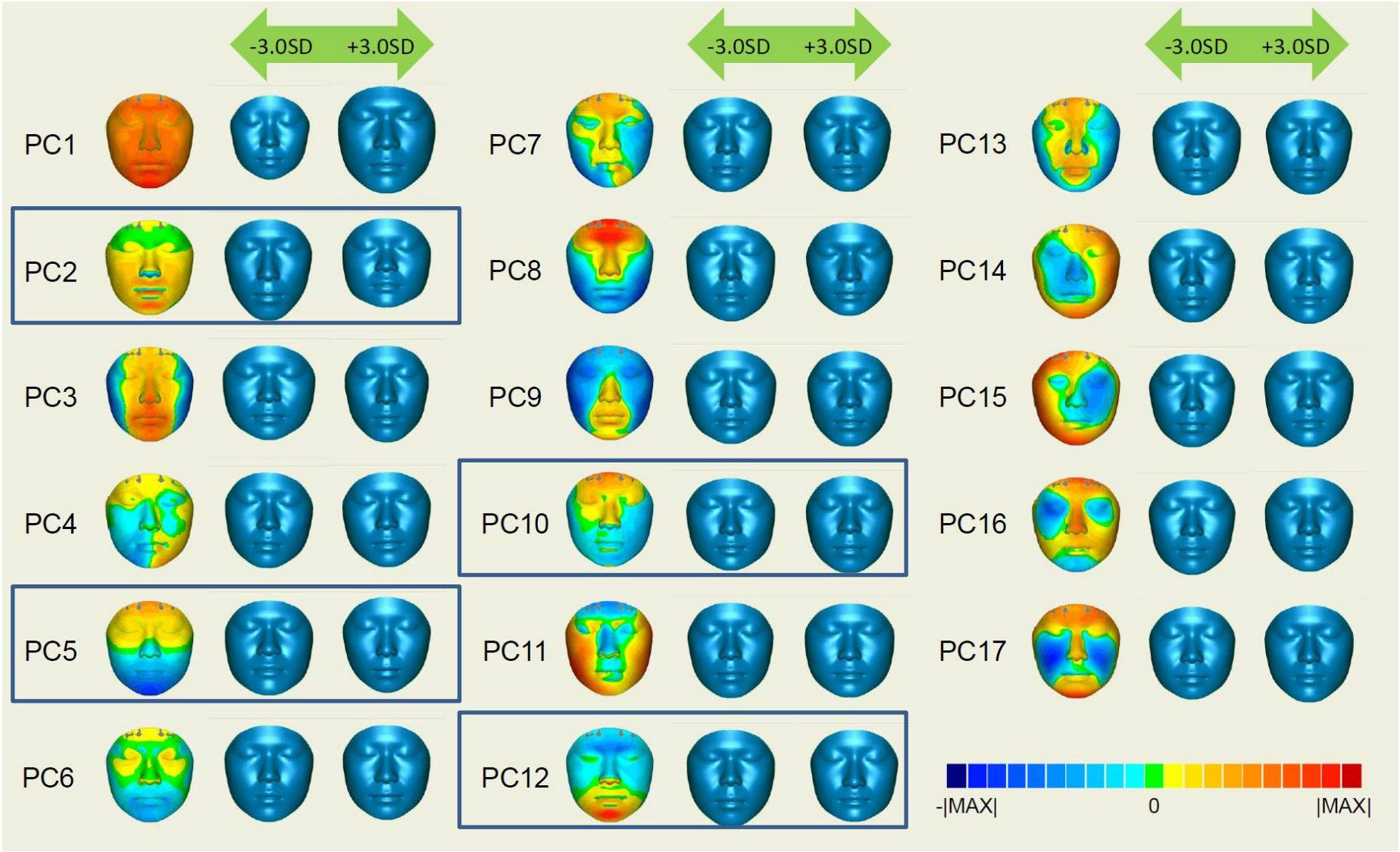



**Figure 5. Visual representation of outlier analysis at the regional and individual population level for (A) height, (B) skin pigmentation, (C) body mass index, (D) type 2 diabetes, (E) Crohn's disease and (F) ulcerative colitis.** For each geographic region we plot the expectation of the regional average, given the observed values in the rest of the dataset as a grey dashed line. The true regional average is plotted as a solid bar, with darkness and thickness proportional to the regional Z score. For each population we plot the observed value as a colored circle, with circle size proportional to the population specific Z score. For example, in (A), one can see that estimated genetic height is systematically lower than expected across Africa. Similarly, estimated genetic height is significantly higher (lower) in the French (Sardinian) population than expected, given the values observed for all other populations in the dataset.
doi:10.1371/journal.pgen.1004412.g005

Berg&Coop 2014

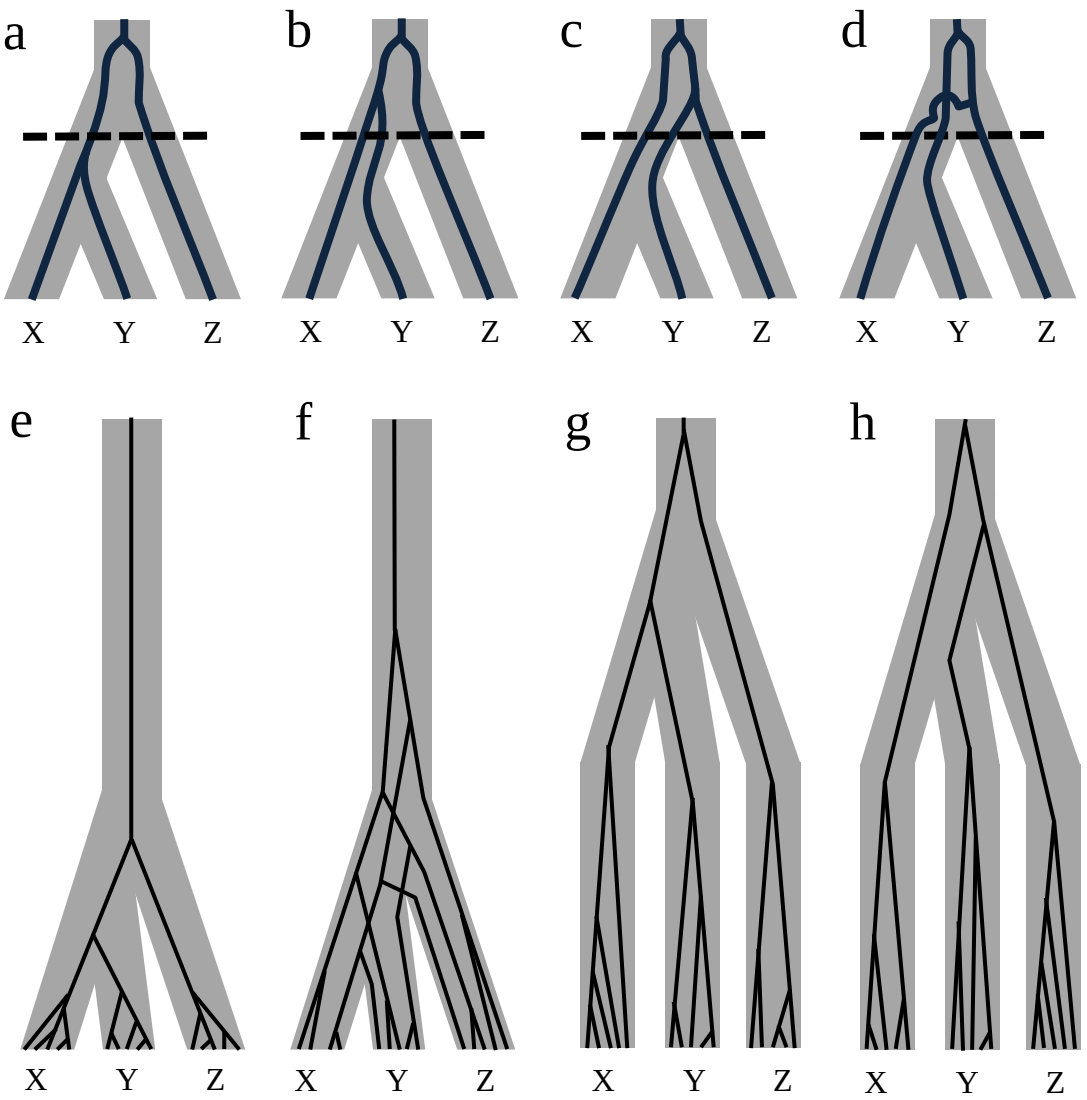
# PST/FSTテスト




本土ー沖縄間で中立の下では説明できない顔面形態成分

# 収斂進化とincomplete lineage sorting

# Incomplete lineage sorting 不完全な系統仕分け




遺伝子と集団の樹形は必ずしも一致しない

# 色素形成関連遺伝子からみた遺伝的近縁性

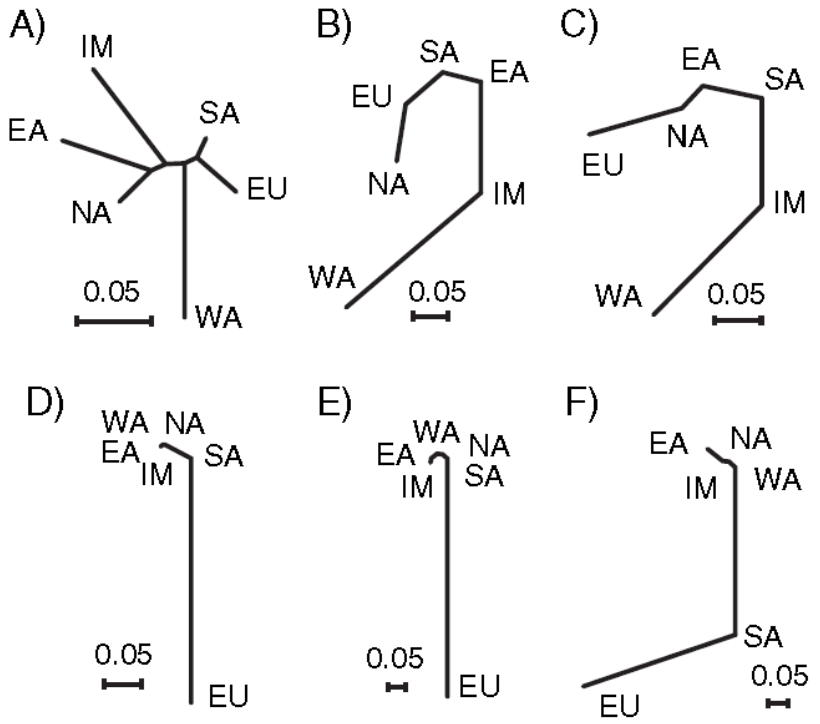
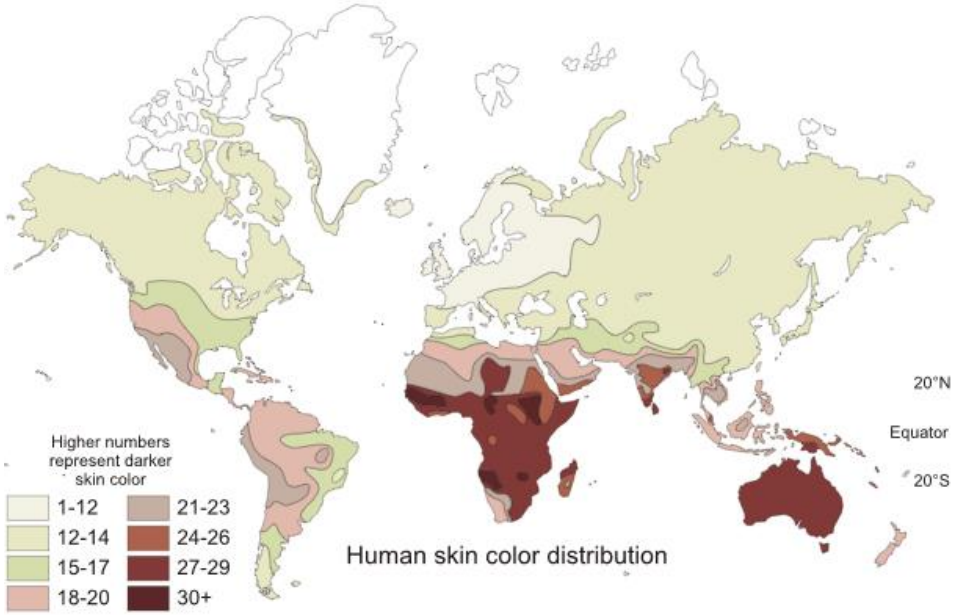
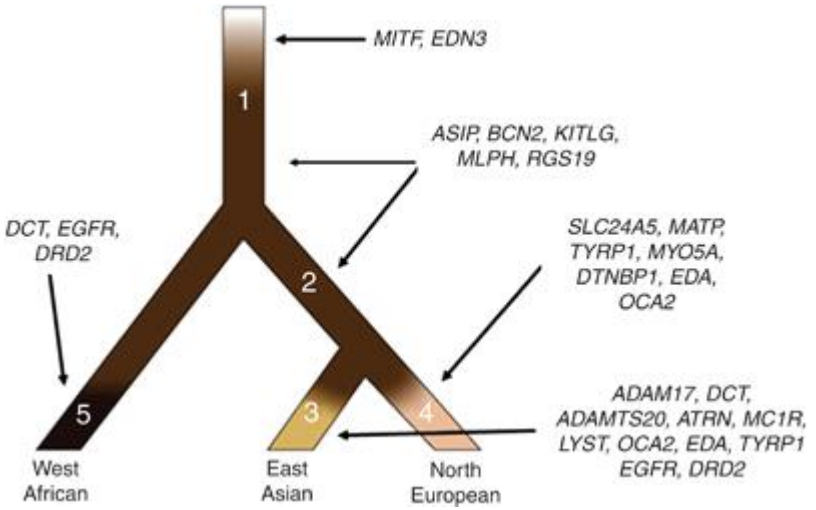
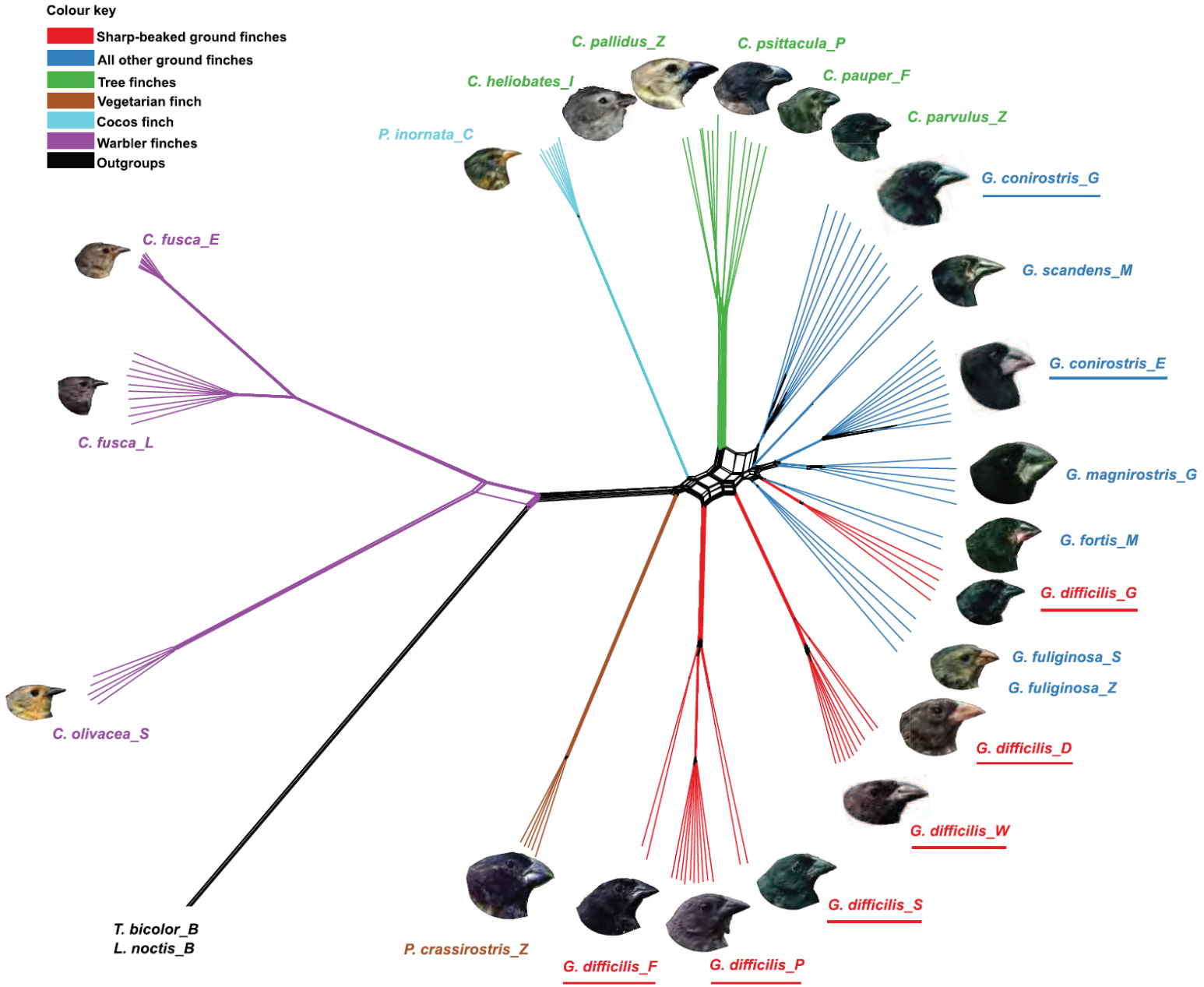



FIG. 1.—Neighbor-Joining trees based on (*a*) average $F_{ST}$ values among the 6 populations typed on the Affymetrix 10K WGSA chip and locus-specific $F_{ST}$ values at (*b*) *ASIP* A8818G, (*c*) *OCA2* A355G, (*d*) *TYR* A192C, (*e*) *MATP* C374G, and (*f*) *SLC24A5* A111G. Populations are abbreviated as follows: WA, West African; SA, South Asian; NA, Native American; EU, European; EA, East Asian; IM, Island Melanesian.




Norton et al. 2007




**Extended Data Figure 3 | Network tree for the Darwin's finches on the basis of all autosomal sites.** Taxa that showed deviations from classical taxonomy are underscored. Finch heads are reproduced from ref. 5. *How and Why Species Multiply: The Radiation of Darwin's Finches* by Peter R. Grant & B. Rosemary Grant. Copyright © 2008 Princeton University Press. Reprinted by permission.

# 集団間交雑と選択

# プエルトリコ人にみられるHLA領域アリルの偏り

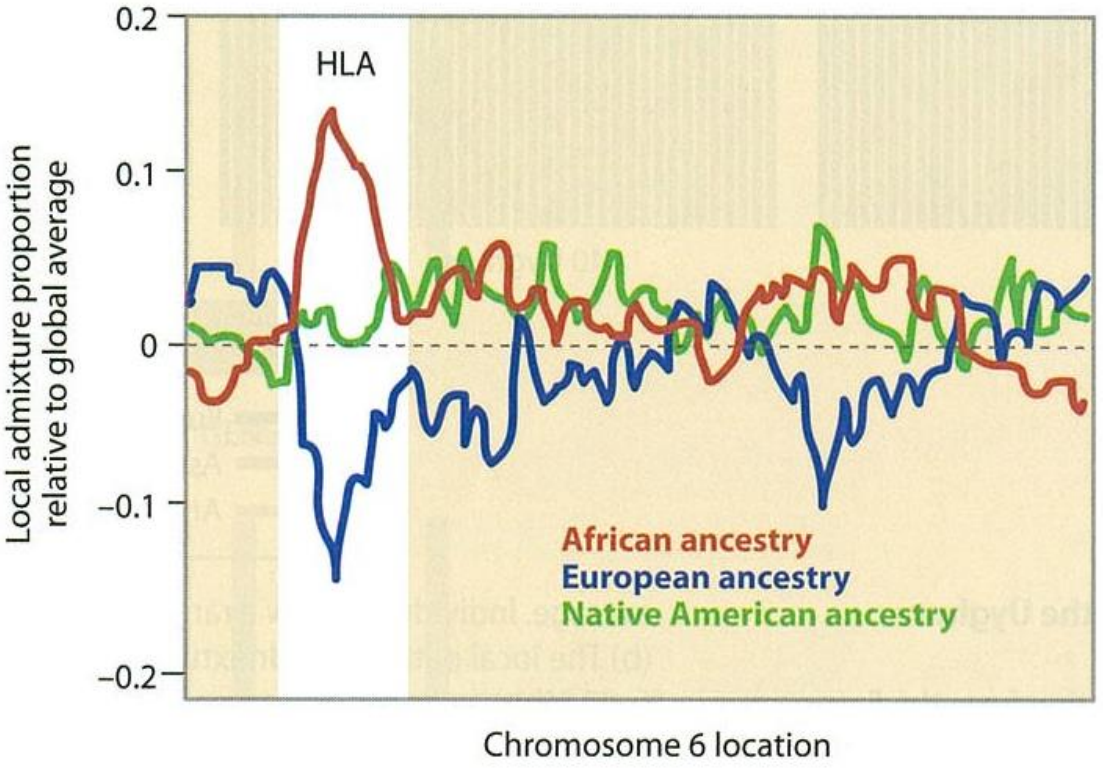
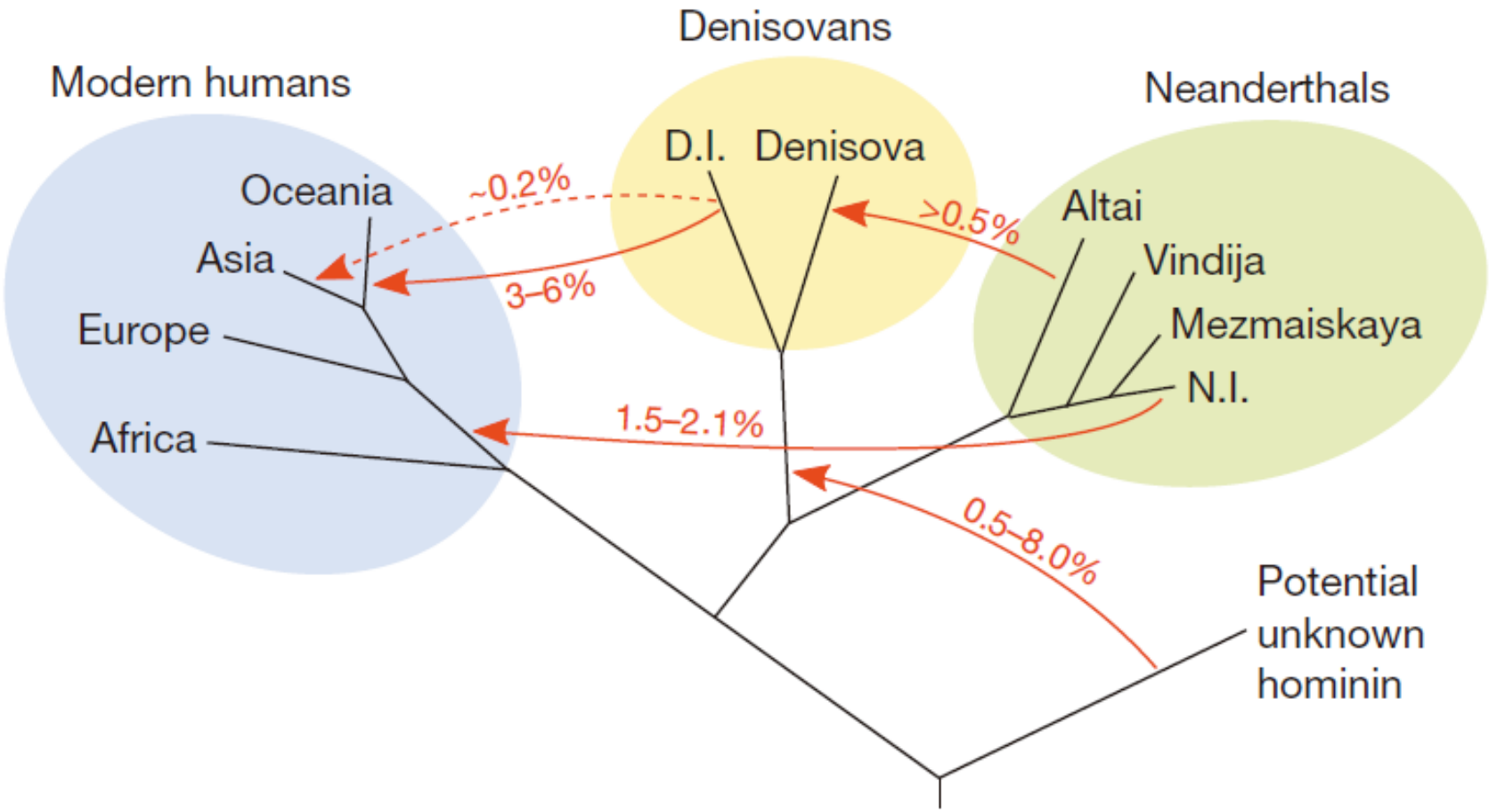


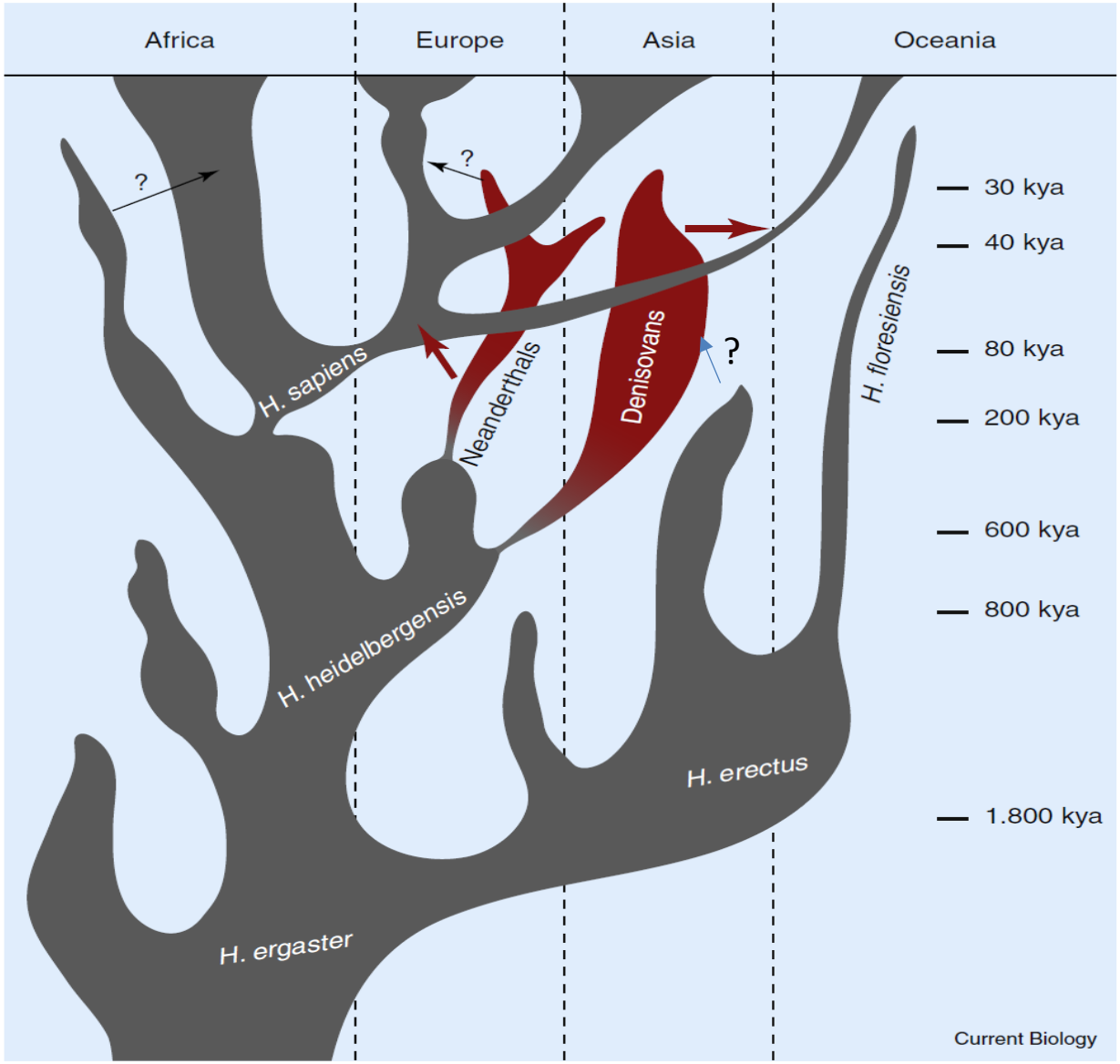

**Figure 14.10: Deviations of admixture proportions due to natural selection on HLA alleles.**
An excess of African and deficiency of European ancestry at the HLA locus (*white* bar) on chromosome 6. [Adapted from Tang H et al. (2007) *Am. J. Hum. Genet.* 81, 626. With permission from Elsevier; and Oleksyk TK et al. (2010) *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 185. With permission from Royal Society Publishing.]

# *Introgression between modern and archaic humans*




Prufer et al. 2014

# *Evolutionary history of genus Homo*

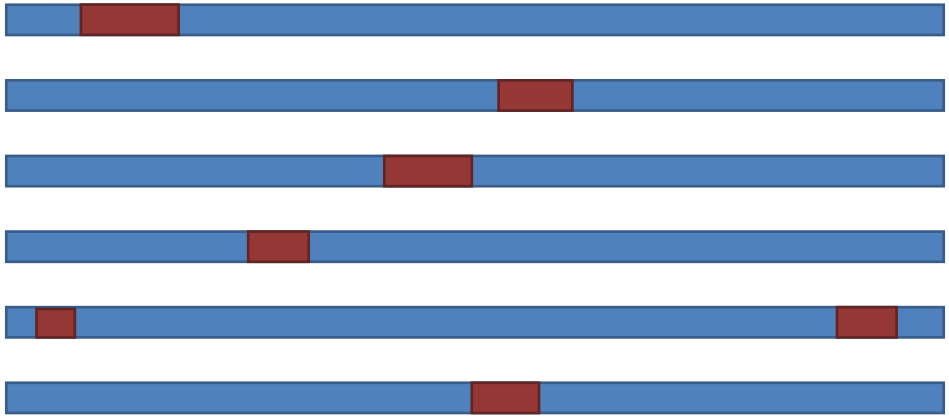



Lalueza-Fox and Gilbert, Current Biology 2011

# *Genomic regions derived from the Neanderthals are different among individuals*

Neanderthal

Modern human



Every non-African modern human have 1-4% of genome sequences derived from the Neanderthals.

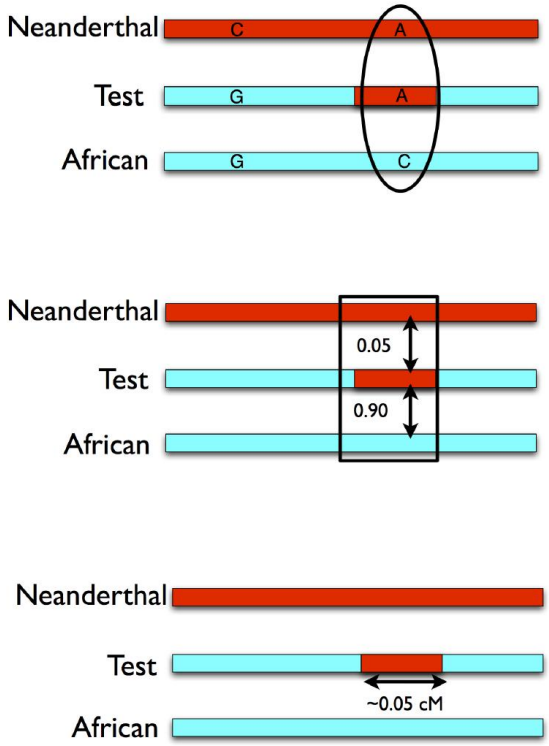
# Genomic regions derived from the Neanderthals

LETTER

doi:10.1038/nature12961

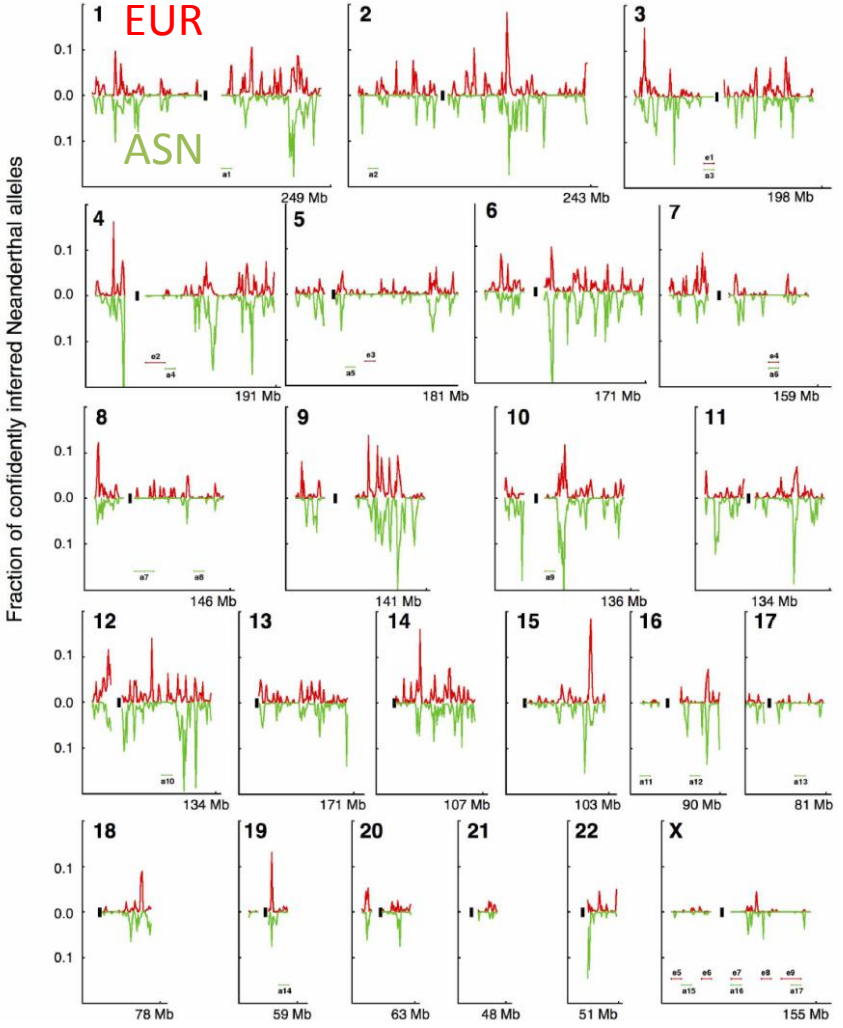
## The genomic landscape of Neanderthal ancestry in present-day humans

Sriram Sankararaman[1,2], Swapan Mallick[1,2], Michael Dannemann[3], Kay Prüfer[3], Janet Kelso[3], Svante Pääbo[3], Nick Patterson[1,2] & David Reich[1,2,4]




**Extended Data Figure 1 | Three features used in the Conditional Random Field for predicting Neanderthal ancestry.** Top (feature 1), patterns of variation at a single SNP. Sites at which a panel of sub-Saharan-African individuals carry the ancestral allele and in which the sequenced Neanderthal and the test haplotype carry the derived allele are likely to be derived from Neanderthal gene flow. Middle (feature 2), haplotype divergence patterns. Genomic segments in which the divergence of the test haplotype to the sequenced Neanderthal is low, whereas the divergence to a panel of sub-Saharan-African individuals is high, are likely to be introgressed. Bottom (feature 3), we searched for segments that have a length consistent with what is expected from Neanderthal-to-modern-human gene flow approximately 2,000 generations ago, corresponding to a size of about 0.05 cM = (100 cM per Morgan)/(2,000 generations).

Sankararaman et al. 2014




**Extended Data Figure 2 | Map of Neanderthal ancestry in 1000 Genomes European and east-Asian populations.** For each chromosome, we plot the fraction of alleles confidently inferred to be of Neanderthal origin (probability >90%) in non-overlapping 1-Mb windows in Europeans (red) and in east Asians (green). Black bars denote the coordinates of the centromeres. We plot traces in non-overlapping 10-Mb windows that pass filters. We label 10-Mb-scale windows that are deficient in Neanderthal ancestry (e1–e9 (e, European), a1–a17 (a, Asian)) (see Supplementary Information section 8 for details).

# Genomic regions derived from the Neanderthals

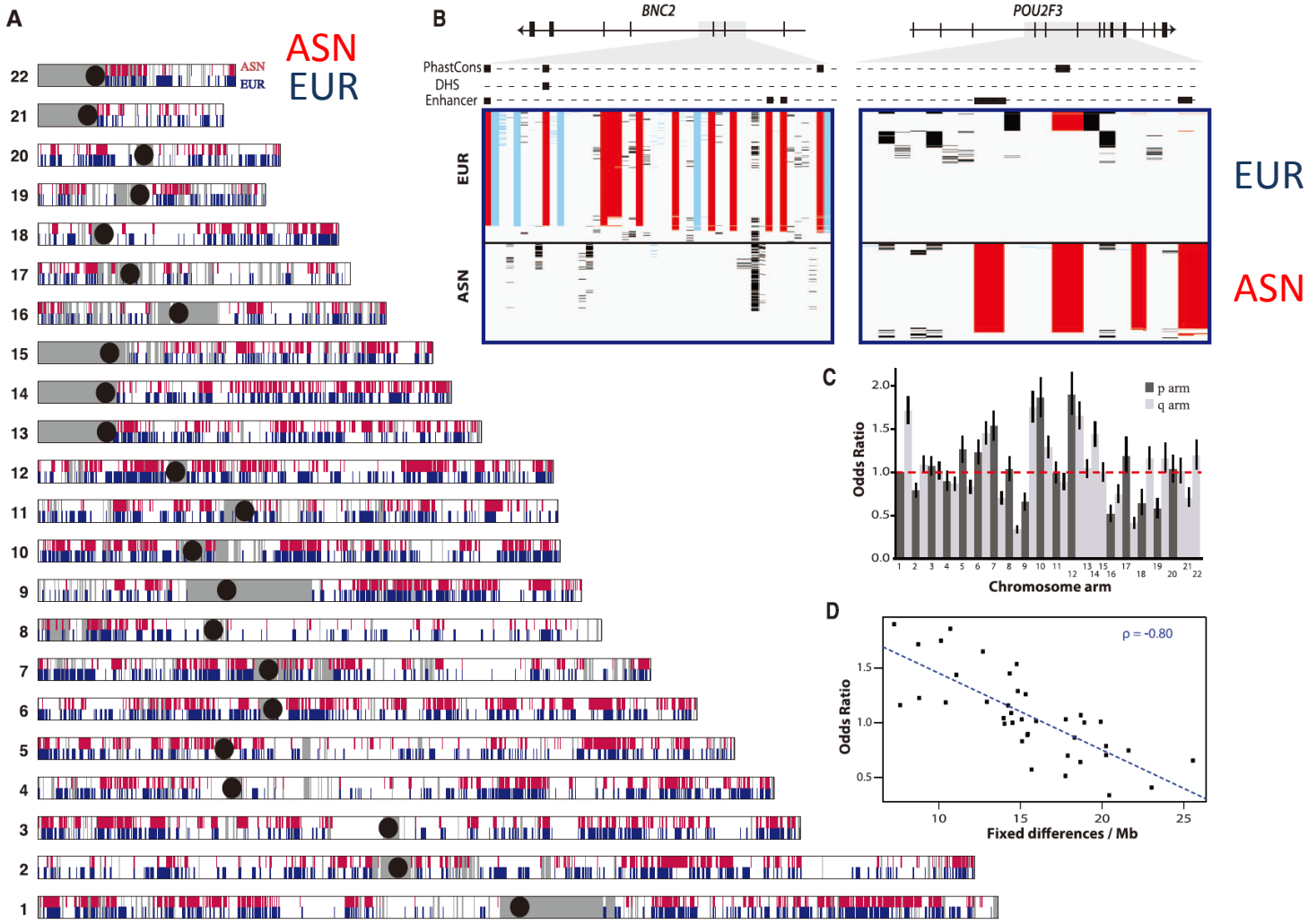



**Fig. 2. Genomic distribution of surviving Neandertal lineages.** (**A**) Neandertal lineages identified in East Asians (ASN, red) and Europeans (EUR, blue). Gray shading denotes regions that did not pass filtering criteria (*10*); black circles represent centromeres. (**B**) Visual genotype illustrations of introgressed sequences identified in the *BNC2* and *POU2F3* genes. Rows denote individuals, columns indicate variant sites, and rectangles are colored according to genotype (red, predicted Neandertal variant that matches the allele present in the Neandertal reference genome; blue, predicted Neandertal variant that does not match the allele present in the Neandertal reference genome; black, other variants). Introgressed variants that overlap a PhastCons conserved element, DNaseI hypersensitive site (DHS), or putative enhancer elements are shown as boxes (*10*). (**C**) Odds of finding an introgressed lineage on each chromosomal arm calculated from a logistic regression model (*10*). Odds ratios (ORs) are expressed using chromosome 1p as the baseline level. Horizontal bars represent 95% CIs. (**D**) Relation between the OR and the number of fixed differences per megabase between humans and Neandertals. ρ, Spearman's rank correlation coefficient.

Vernot et al. 2014

# Adaptive introgression

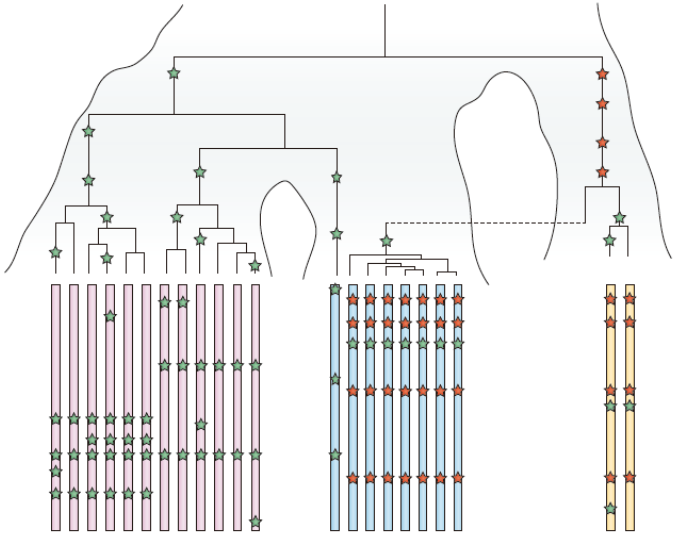


Figure 2 | **Example coalescent genealogy of uniquely shared mutations.** Several DNA fragments from two modern populations (pink and blue chromosomes) are sequenced. A diploid sequence is also obtained from an extinct archaic population (yellow chromosomes) that split from the population tree more anciently than the two modern populations split from each other. Uniquely shared mutations (red stars) occur in the archaic population but are passed on to the ancestors of the blue modern population via admixture (dashed line). These are then swept to high frequency by selection, producing a shallow local coalescent genealogy. This process results in sites with high-frequency derived alleles in the blue samples that are present in the archaic sample but not in the pink samples from the other modern population. Mutations in the genealogy that are not uniquely shared are shown as green stars.

Racimo et al. 2015

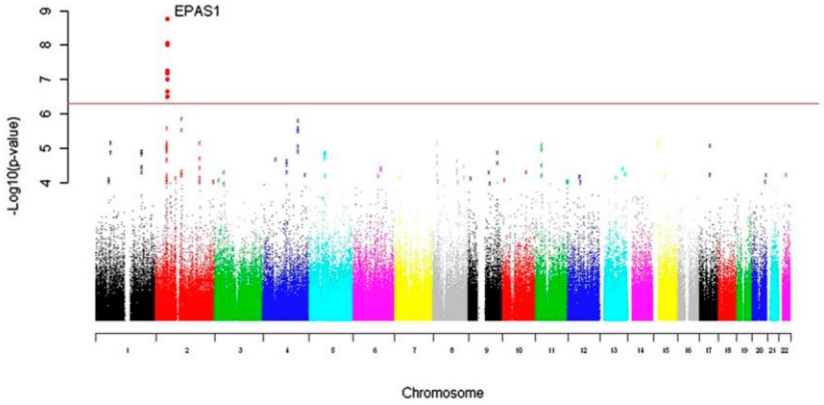
Table 1 | **Candidates for adaptive introgression**

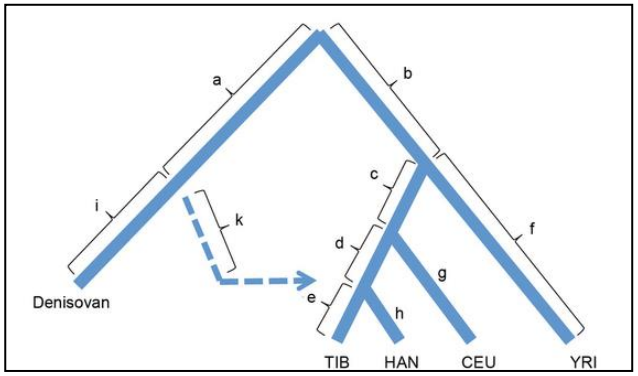
| Introgressed haplotype | Closest putative archaic source population | Modern populations showing evidence of introgression | Evidence for positive selection and selection tests performed | Most likely population in which selection occurred |
|---|---|---|---|---|
| *HLA-A*, *HLA-B* and *HLA-C* | Neanderthal and Denisovan | Europeans, East Asians and Melanesians | Extreme allelic and haplotypic diversity in the *HLA* region, which is indicative of balancing selection[58] | Europeans, East Asians and Melanesians |
| *HLA-DPB1* | Neanderthal (?), but this was questioned in REF. 73 | Europeans (?) | No formal test of neutrality performed[72], but a phylogenetic analysis suggests that the haplotype is not introgressed[73] | – |
| *STAT2* (haplotype N) | Neanderthal | Non-Africans | One-tailed test for elevated frequency of diagnostic SNP in introgressed haplotype based on empirical distribution of SNP frequencies[41] | Melanesians |
| *STAT2* (haplotype D) | Denisovan (?) | Melanesians | No formal test of neutrality performed, but haplotype is only present in Papuans at a low frequency[41] | – |
| *OAS1* | Denisovan, but extremely ancient coalescence with human reference suggests that the direct source was a different archaic group | Melanesians | • No evidence for positive selection in REF. 75<br>• Evidence for positive selection in REF. 44<br>• Suggestive signal of balancing selection based on genetic differentiation of haplotypes across continents[74] | – |
| *OAS* gene cluster | Neanderthal | Non-Africans | | |
| *HYAL2* (3p21.31)[60] | Neanderthal | East Asians | iHS[46], EHH[47] and CMS[51] | East Asians |
| *MC1R*[78] | Neanderthal | Non-Africans | Tajima's D[49], Fu and Li's test[82], and iHS[46] | Taiwanese (?) |
| *SLC16A11* and *SLC16A13* | Neanderthal | Native Americans | Genotype–phenotype association[90] | Native Americans (?) |
| *DMD* | Neanderthal | Non-Africans | No tests performed[91,92] | – |
| *EPAS1* (REF. 57) | Denisovan | East Asians or Tibetans only | • High population differentiation[55]<br>• Genotype–phenotype association[55]<br>• High archaic haplotype frequency[57] | Tibetans |
| Various regions identified via *S** that contain genes involved in the integumentary system | Neanderthal | Europeans and East Asians | • High $F_{ST}$ between Europeans and Asians in variants identified to be introgressed (local adaptation)[19]<br>• Introgressed haplotypes with frequencies in both Europeans and Asians that are higher than expected under neutrality (shared adaptation)[19] | Europeans and East Asians |
| Various regions identified via CRF that contain genes involved in keratin filament, sugar metabolism, muscle contraction and oocyte meiosis | Neanderthal | Europeans and East Asians | Windows in which the population frequency of the Neanderthal genetic material is too high to be explained by neutral drift[44] | Europeans and East Asians |
| Various regions identified via HMM | Neanderthal and Denisovan | Europeans, East Asians and Melanesians | No claims about selection and no formal tests of neutrality performed[9,31] | – |

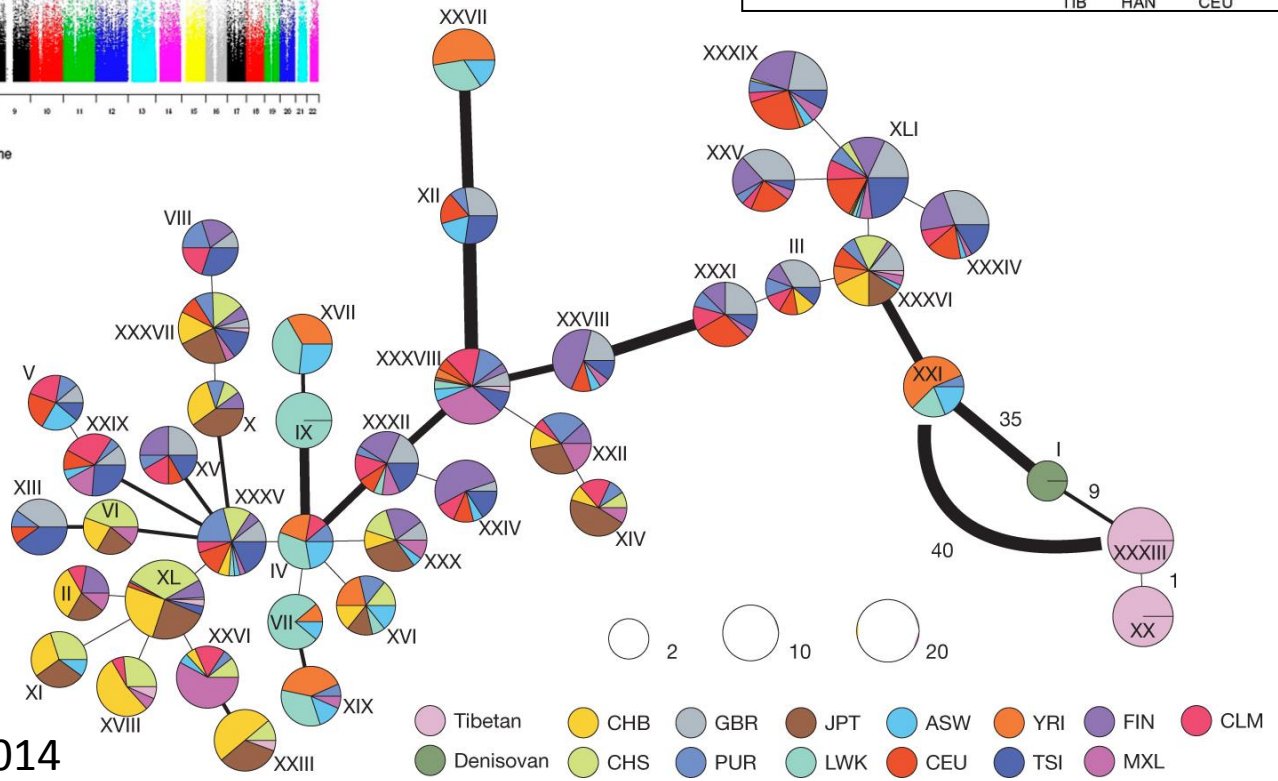
CMS, composite of multiple signals; CRF, conditional random field; *DMD*, dystrophin; EHH, extended haplotype homozygosity; *EPAS1*, endothelial PAS domain protein 1; *HLA*, human leukocyte antigen; HMM, hidden Markov model; *HYAL2*, hyaluronoglucosaminidase 2; iHS; integrated haplotype score; *MC1R*, melanocortin 1 receptor; *OAS*, 2′-5′-oligoadenylate synthetase; *SLC*, solute carrier; SNP, single-nucleotide polymorphism; *STAT2*, signal transducer and activator of transcription 2. The table provides a summary of recent studies reporting adaptive introgression with a comparison of the evidence provided in support of positive selection acting on the introgressed regions. A question mark denotes that the source population, the receiving population or the population in which selection occurred remains unresolved and/or has conflicting reports from different studies.

# 低酸素誘導因子*EPAS1*

漢人とチベット人の遺伝的分化($F_{ST}$)




Huerta-Sanchez et al. 2014

チベット人のみがもつ高地適応型ハプロタイプはデニソワ人に由来する。

Thank you very much
for your attention !