Resurrecting Surviving Neandertal Lineages from Modern Human Genomes

Song Bin, Zhang Yuanwei, Jiang Lan
Group1, Topic 8
2021-10-23

Abstract

Resurrecting Surviving Neandertal Lineages from Modern Human Genomes

Benjamin Vernot and Joshua M. Akey*

Anatomically modern humans overlapped and mated with Neandertals such that non-African humans inherit ~1 to 3% of their genomes from Neandertal ancestors. We identified Neandertal lineages that persist in the DNA of modern humans, in whole-genome sequences from 379 European and 286 East Asian individuals, recovering more than 15 gigabases of introgressed sequence that spans ~20% of the Neandertal genome (false discovery rate = 5%). Analyses of surviving archaic lineages suggest that there were fitness costs to hybridization, admixture occurred both before and after divergence of non-African modern humans, and Neandertals were a source of adaptive variation for loci involved in skin phenotypes. Our results provide a new avenue for paleogenomics studies, allowing substantial amounts of population-level DNA sequence information to be obtained from extinct groups, even in the absence of fossilized remains.

Author

Benjamin Vernot

Max Planck Institute

Joshua M. Akey

Professor of Ecology and Evolutionary Biology at the Princeton University Lewis-Sigler Institute for Integrative Genomics

Outline

1. Research background and purpose

2. Source Data and method: A two-stage computational strategy

3. Results

- Recovering introgressed sequence
- Fitness costs to hybridization
- Refine admixture models
- Signature of adaptive introgression

4.Conclusions

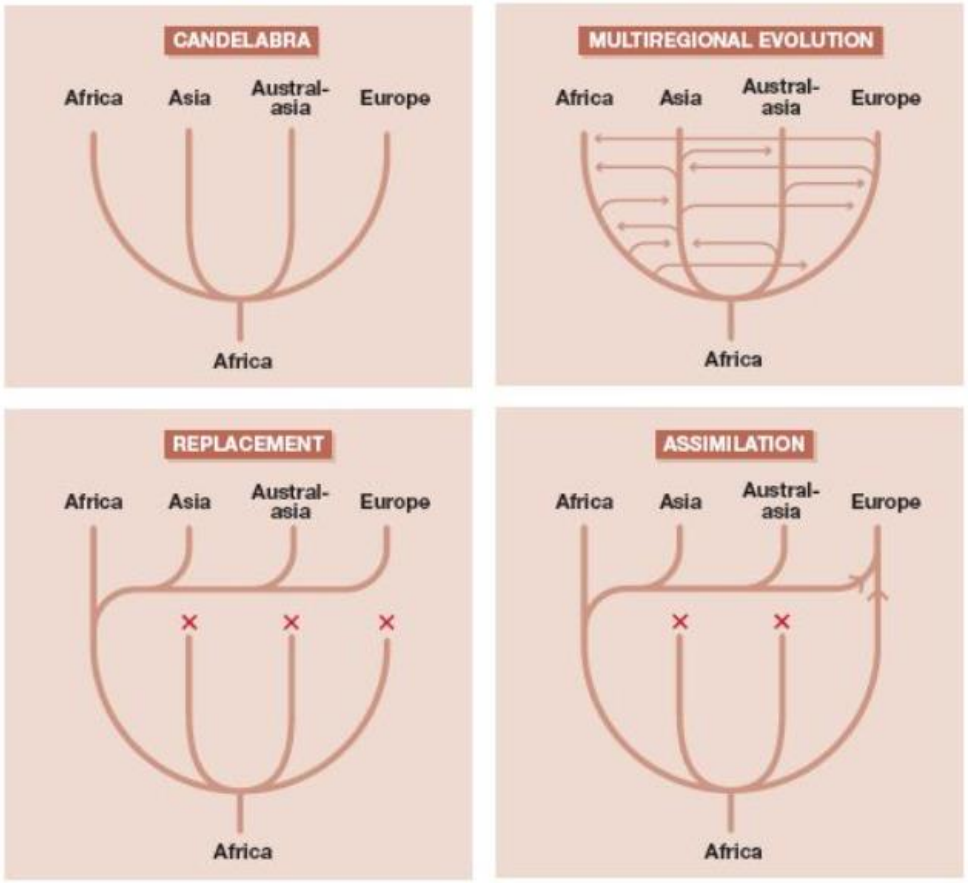
Hypotheses of human evolution

- Two competing hypotheses of human evolution were originally proposed.

(1) out-of-Africa model.

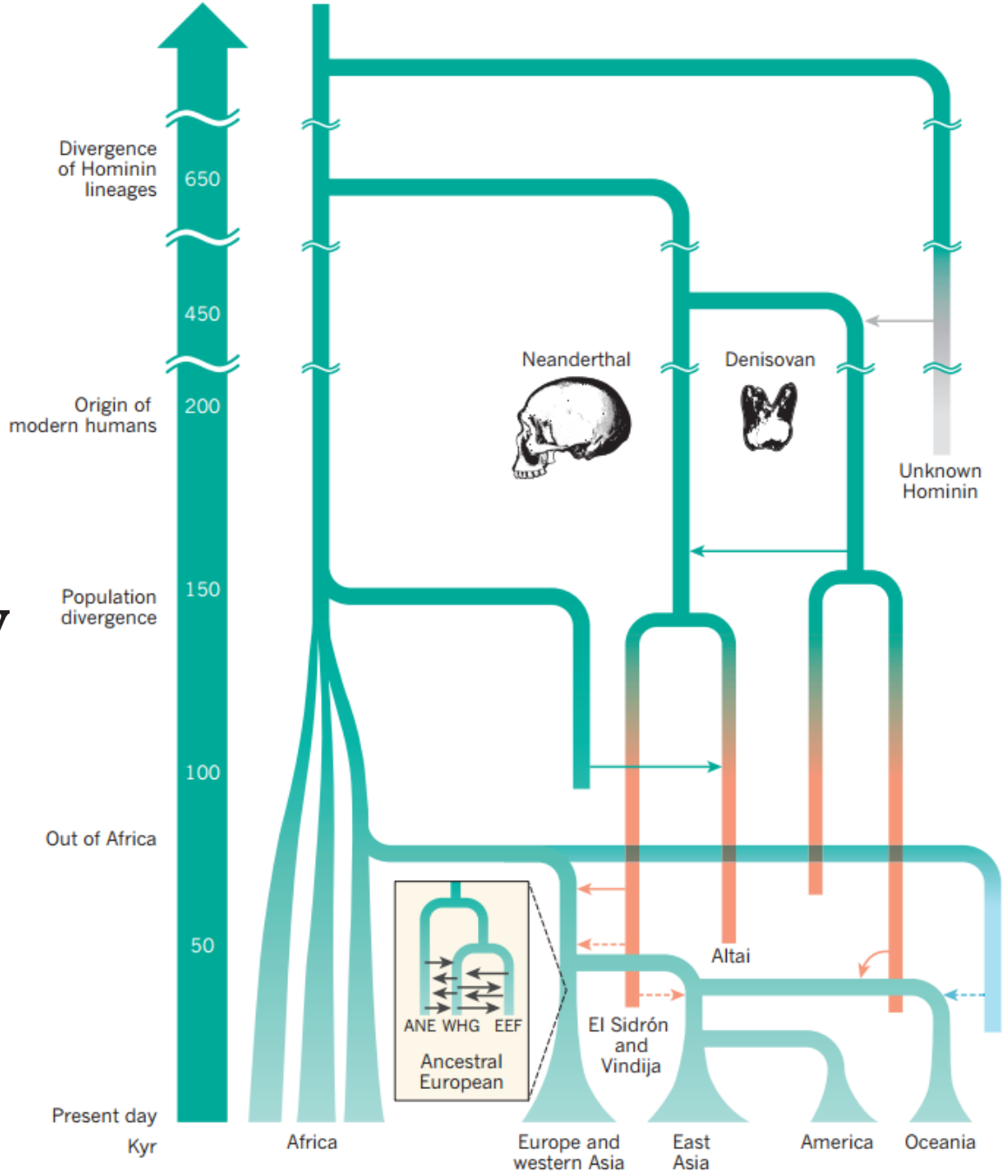
(2) multiregional model.

- Assimilation model: a intermediate model.



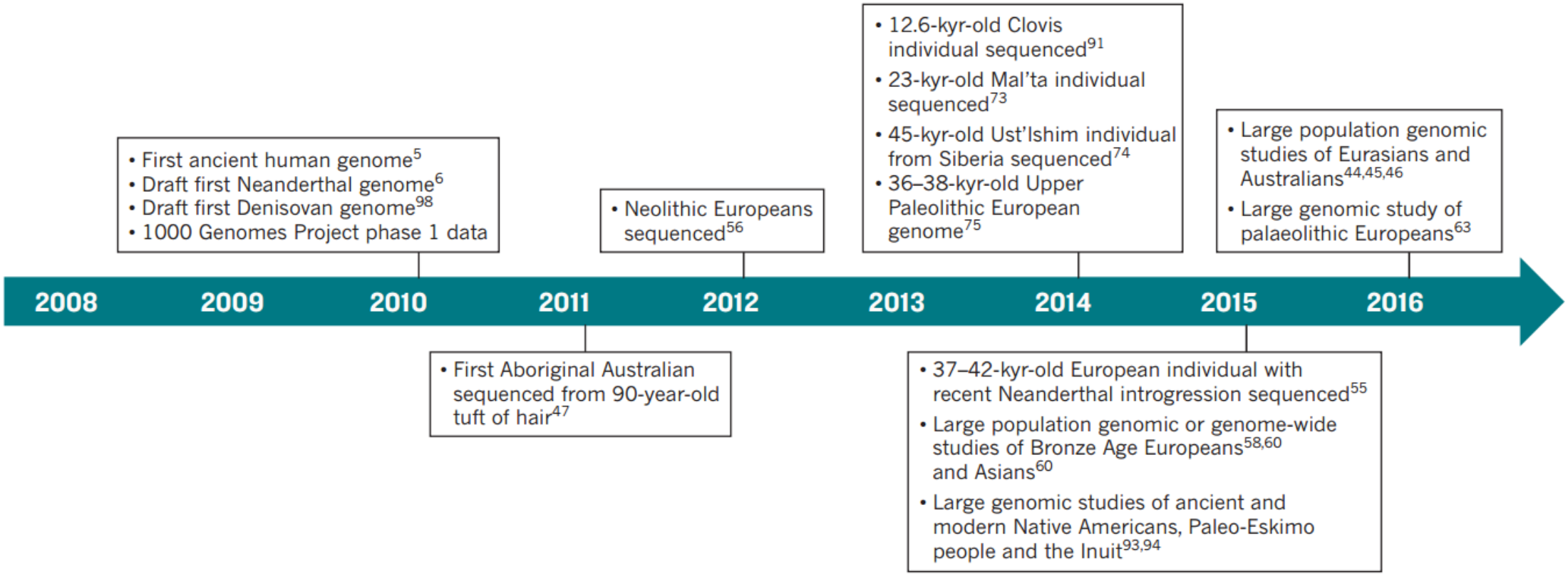
Four models of the origin of human

Simplified model of human evolutionary history



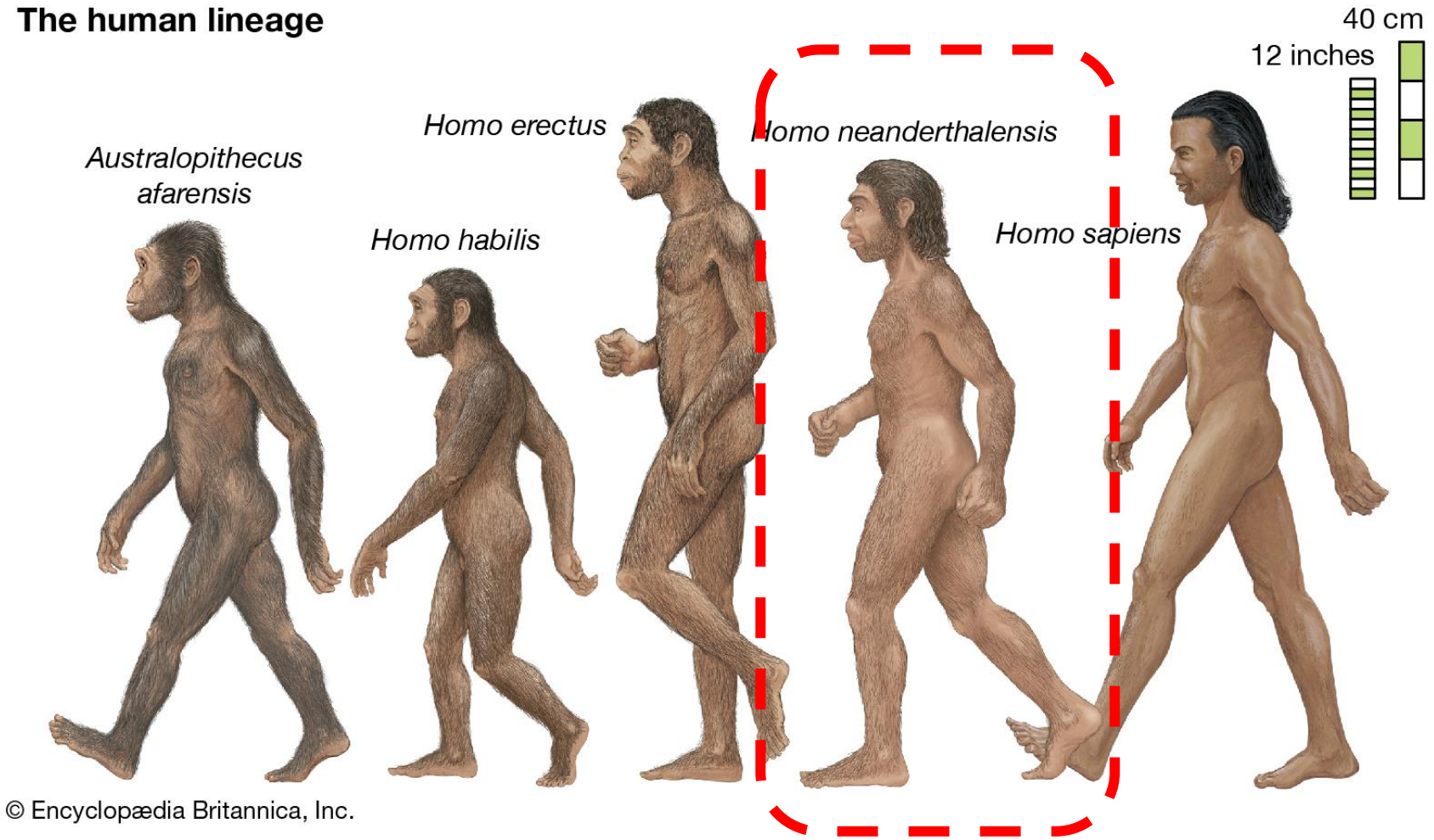
2017, Tracing the peopling of the world through genomics

Timeline of important milestones in human evolutionary genomics



2017, Tracing the peopling of the world through genomics

尼安德特人（尼人）



一群生存于旧石器时代的史前人类，1856 年，其遗迹首先在德国尼安德河谷被发现。

目前归类为人科人属，随着2010年的研究发现部分现代人是其混血后代后，也可能被归类于智人下的一个亚种。

The studying history of Neandertal DNA

Cell, Vol. 90, 19–30, July 11, 1997, Copyright ©1997 by Cell Press

Neandertal DNA Sequences and the Origin of Modern Humans

Summary

DNA was extracted from the Neandertal-type specimen found in 1856 in western Germany. By sequencing clones from short overlapping PCR products, a hitherto unknown mitochondrial (mt) DNA sequence was determined. Multiple controls indicate that this sequence is endogenous to the fossil. Sequence comparisons with human mtDNA sequences, as well as phylogenetic analyses, show that the Neandertal sequence falls outside the variation of modern humans. Furthermore, the age of the common ancestor of the Neandertal and modern human mtDNAs is estimated to be four times greater than that of the common ancestor of human mtDNAs. This suggests that Neandertals went extinct without contributing mtDNA to modern humans.

1997年，德国慕尼黑大学对尼人化石进行了线粒体DNA测序，并与来自非洲之外地区的现代人线粒体DNA样本进行对比。结果显示：尼安德特人对现代人在线粒体DNA上并无贡献。

7 MAY 2010 VOL 328 SCIENCE

A Draft Sequence of the Neandertal Genome

Neandertals, the closest evolutionary relatives of present-day humans, lived in large parts of Europe and western Asia before disappearing 30,000 years ago. We present a draft sequence of the Neandertal genome composed of more than 4 billion nucleotides from three individuals. Comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans, including genes involved in metabolism and in cognitive and skeletal development. We show that Neandertals shared more genetic variants with present-day humans in Eurasia than with present-day humans in sub-Saharan Africa, suggesting that gene flow from Neandertals into the ancestors of non-Africans occurred before the divergence of Eurasian groups from each other.

2010年，德国马普进化人类学研究所对3块尼人化石测序，获得第一份尼人全基因组草图。

（1）通过与五名现代人样本对比，认为现代非洲之外人群含有1%~4%的尼人基因，证实了尼安德特人与现代人发生过基因交流。

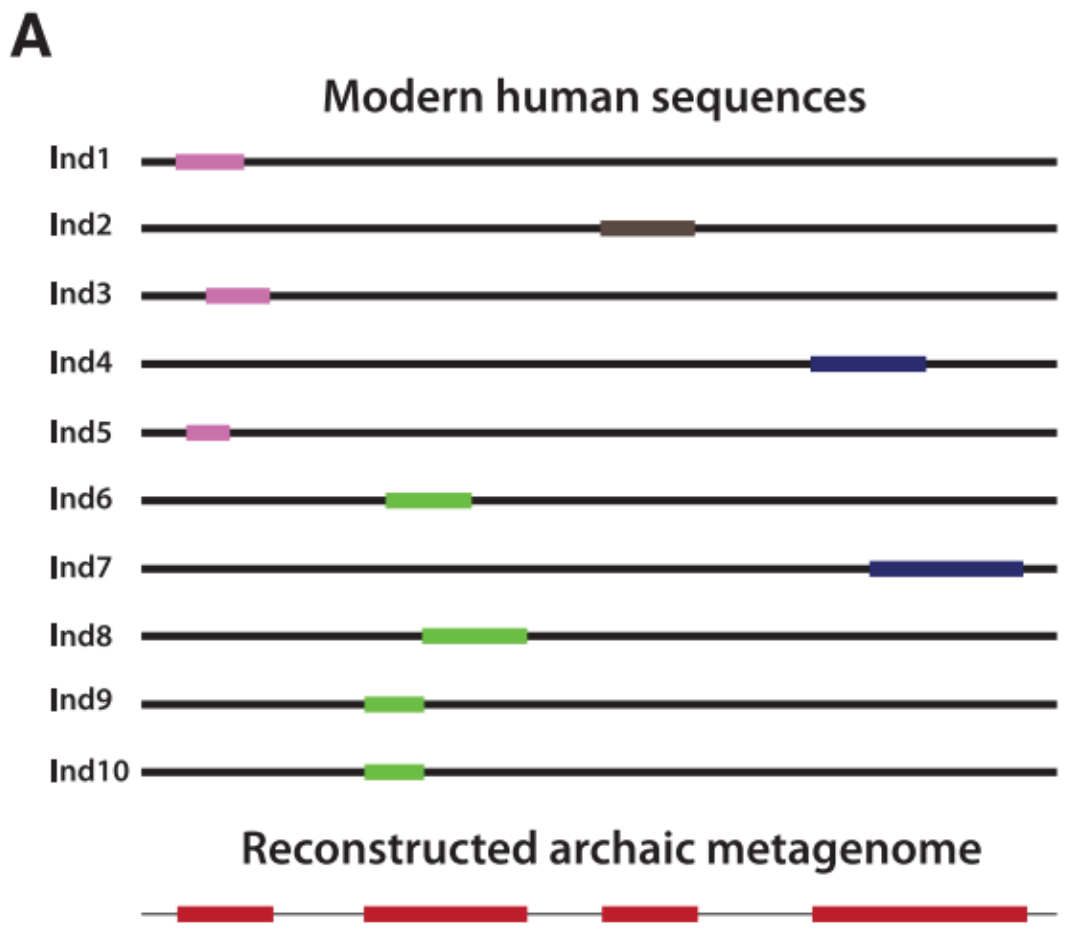
（2）认为尼安德特人与早期现代人的基因交流发生在早期现代人走出非洲之后，欧洲和亚洲的现代人分歧之前。

Research Purpose

- **Hypothesis**: a substantial amount of the Neandertal genome may be recovered from the analysis of contemporary humans despite the limited amounts of admixture

- **How**: By identifying Neandertal sequences from a large sample of modern humans

- **Purpose**: discover surviving lineages that may come from multiple archaic ancestors



Schematic representation illustrating that low levels of introgression may facilitate the recovery of substantial amounts of archaic sequence.

Source data

Table S1. Summary of 1000 Genomes Project samples.

Population	Description	Number Individuals	
TSI	Toscani in Italia	98	
IBS	Iberian population in Spain	14	
CEU	Utah Residents (CEPH) with Northern and Western European ancestry	85	379 European
GBR	British in England and Scotland	89	
FIN	Finnish in Finland	93	
CHS	Southern Han Chinese	100	
CHB	Han Chinese in Bejing, China	97	286 East Asian
JPT	Japanese in Tokyo, Japan	89	

The analysis of considering the Neanderthal sequence

Whole-genome Sequences (1000 Genome Project)

↓ Neanderthal alignments (a toe bone in Denisova Cave)

Call variants SAMtools mpileup, BCFtools

↓

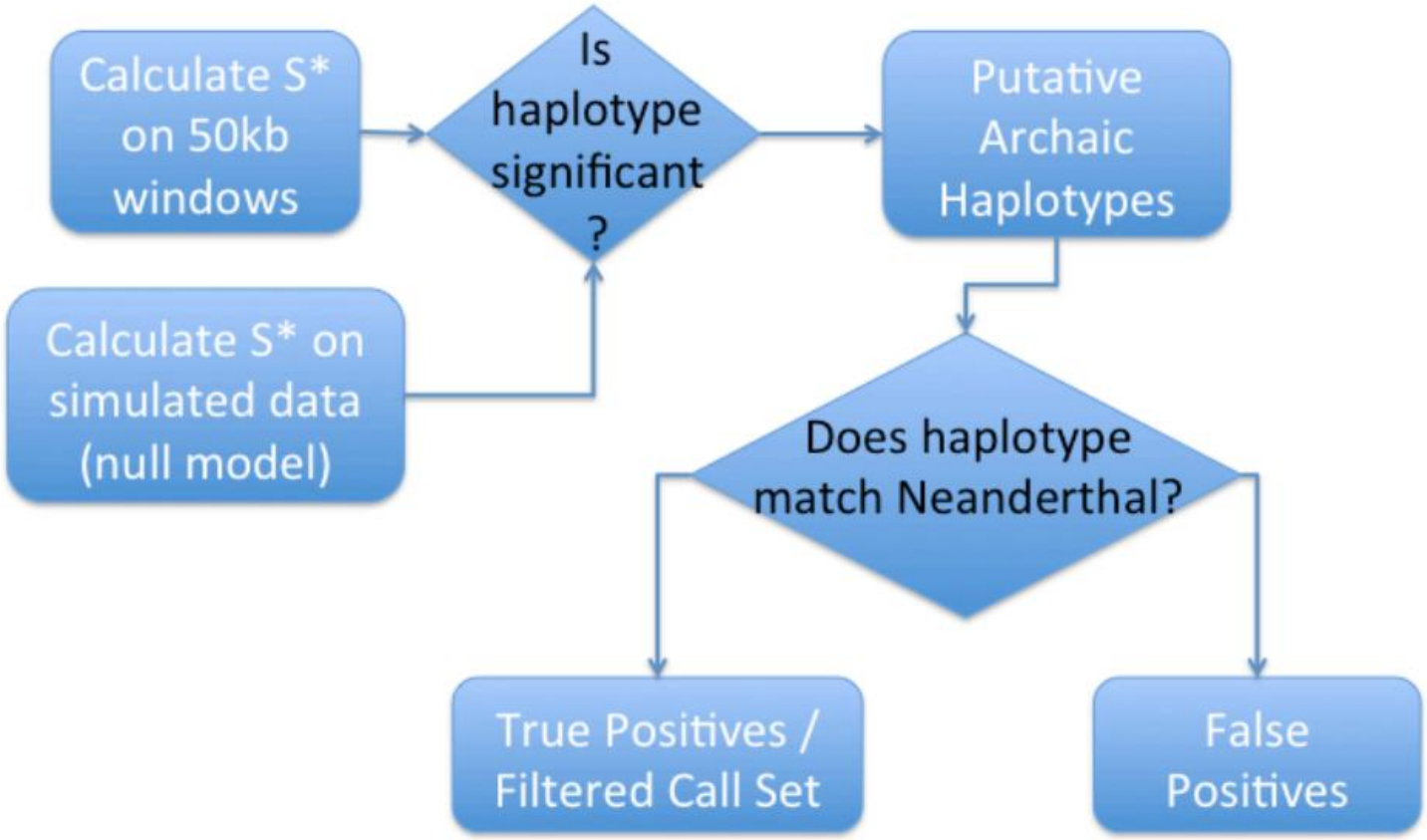
Remove INDELs

↓

Remove CpGs, genomics super dups, difficult to map bases → Neanderthal variants

have minimum RMS mapping quality of 25, minimum read depth of 10, and be further than 10bp from any INDELs.

Method | Identifying surviving Neandertal lineages



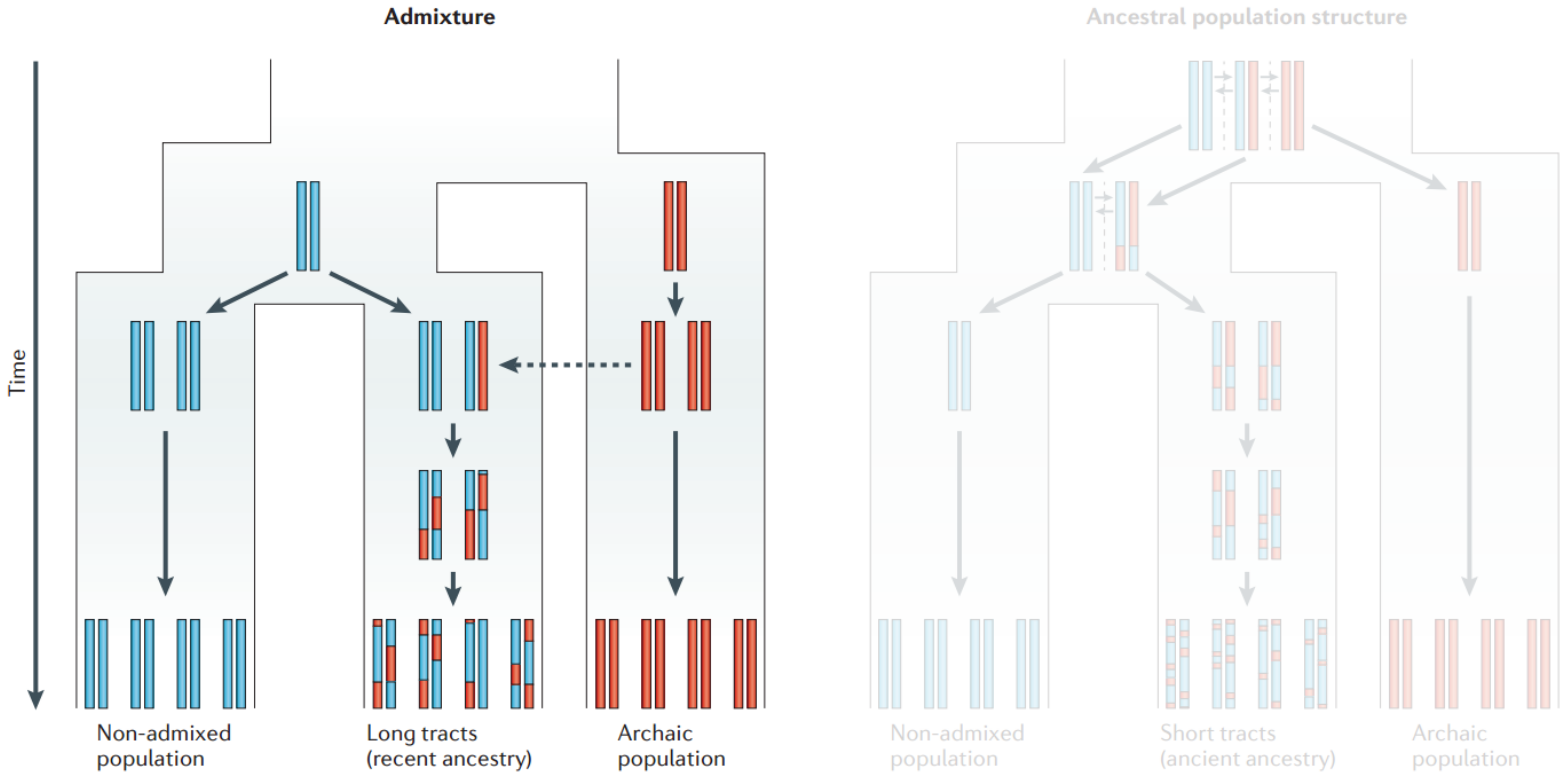
A two-stage computational strategy

1, identify candidate introgressed Sequences (perfect LD) by Calculation of S*

2, identify true introgressed Sequences by comparing them to the Neandertal reference genome

Figure S3. Schematic of computational strategy to identify introgressed lineages

Theory of infering introgressed segments



- An introgressed haplotype should have high sequence divergence to other present-day human individuals
- DNA from recent introgression events should fall into longer contiguous tracts than DNA resulting from old introgression events

Expected length of archaic tracts under admixture and ancestral population structure scenarios

2015-Evidence for archaic adaptive introgression in humans

Stage1, Calculation of S* for individuals (1)

- Challenge

(1) How to distinguish introgressed haplotypes from non-introgressed lineages ?

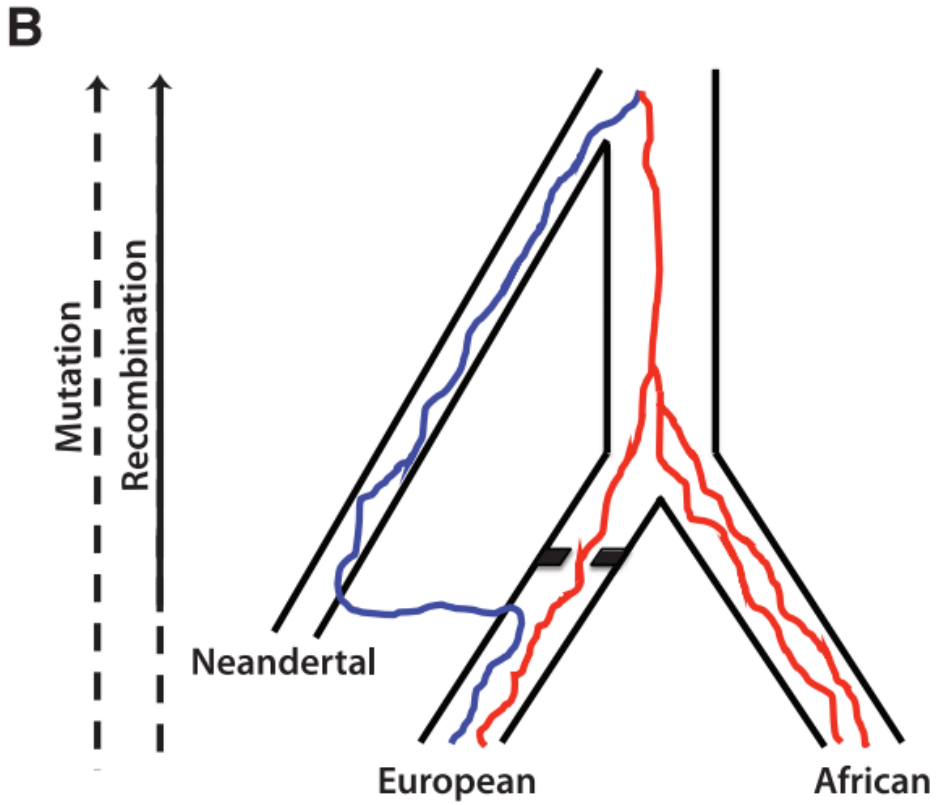
(2) How to distinguish true introgression from shared ancestral genetic variation ?

- Expected signature of an introgressed lineage

(1) high levels of divergence

(2) admixture occurred relatively recently, long haplotype blocks (~50kb)

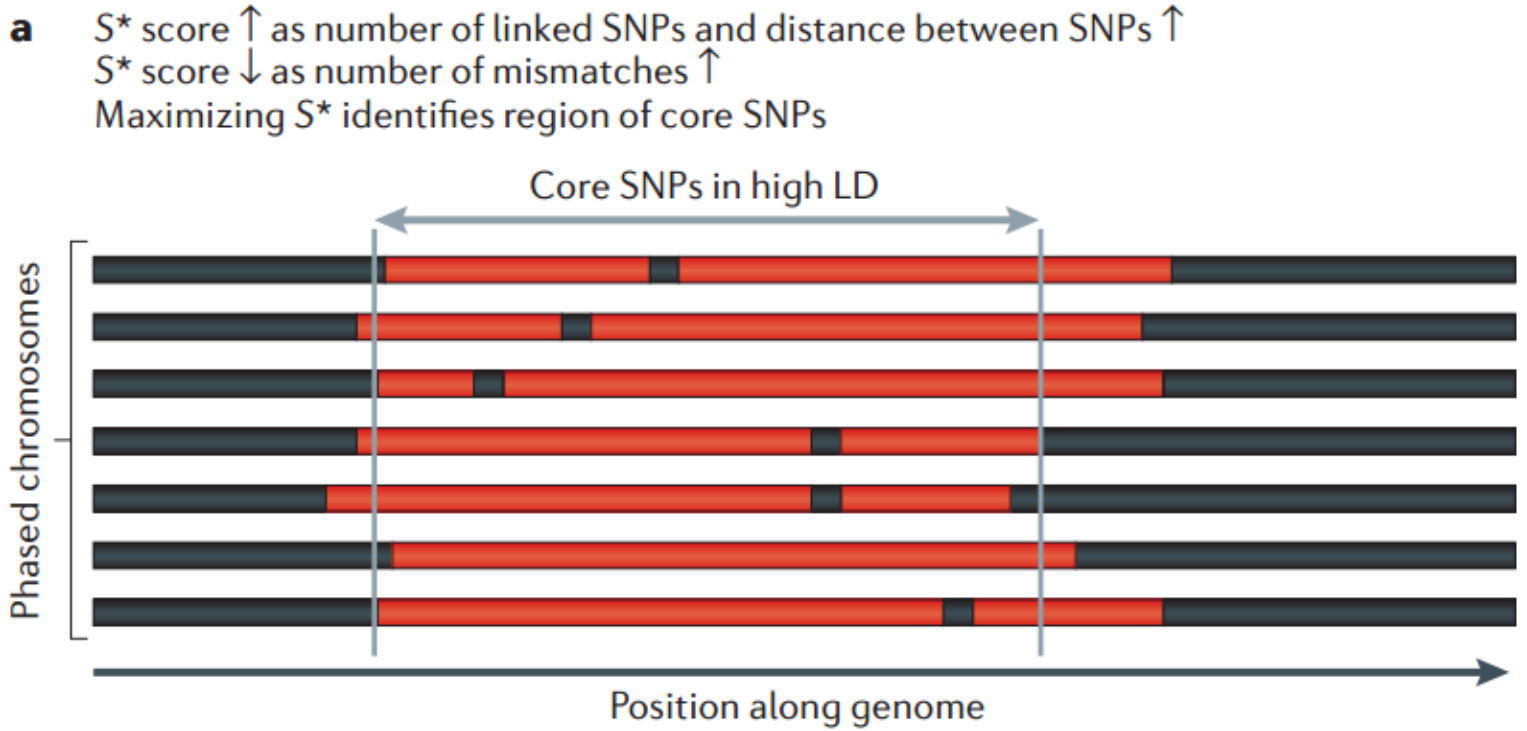
(3) Neanderthal admixture is expected to have occurred only in non-African populations, not found in a "reference" population



F1B. Genealogies of loci in Europeans and Africans in the presence of introgression

Stage1, Calculation of S* for individuals (2)

S* is designed to detect **divergent haplotypes** whose variants are in **strong linkage disequilibrium** and are **not found in a "reference" population**.



2015-Evidence for archaic adaptive introgression in humans

Stage1, Calculation of S* for individuals (3)

[redacted] The statistic for the i^{th} individual in a region is calculated as $S_i^* = \max_{J \subseteq V_i} S(J)$, where:

$$S(J) = \sum_{j \in J} \begin{cases} -\infty, & d(j, j+1) > 5 \\ -10000, & d(j, j+1) \in \{1 \dots 5\} \\ 5000 + bp(j, j+1), & d(j, j+1) = 0 \\ 0, & j = max(J) \end{cases}$$

d(j,j+1) is the genotype distance between two variants
bp(j,j+1) is the distance in base pairs between two variants
the final line allows the last variant to be added

Stage1, Calculation of S* for individuals (4)

Table S2. Genotypes for six hypothetical individuals. Two individuals (1-4; blue) are from the target population, in which S* is being calculated. Two individuals (5 & 6; red) are from the reference population – variants which are present in these individuals are excluded from the analysis (variants 3,7,12 & 15; gray).

snp	pos	ind_1	ind_2	ind_3	ind_4	ind_5	ind_6
snp_0	2309	1	0	0	0	0	0
snp_1	7879	0	1	0	0	0	0
snp_2	11484	0	0	1	0	0	0
snp_3	16249	2	1	2	2	2	2
snp_4	17324	0	1	0	0	0	0
snp_5	19064	0	0	1	0	0	0
snp_6	19124	0	1	0	0	0	0
snp_7	23559	2	1	2	2	2	2
snp_8	25354	1	0	0	0	0	0
snp_9	26654	2	1	2	2	0	0
snp_10	29724	2	1	2	2	0	0
snp_11	30769	0	1	0	0	0	0
snp_12	31319	0	0	0	0	2	2
snp_13	37199	1	0	0	0	0	0
snp_14	38009	0	1	0	0	0	0
snp_15	39444	0	0	0	0	2	2
snp_16	40809	2	1	2	2	0	0
snp_17	45079	2	1	2	2	0	0
snp_18	48989	0	1	0	0	0	0

Table S3. Genotype distance for all pairs of snps present in individual 1. Red cells (lower triangle) mark pairs of snps with distance greater than 5; these variants cannot both be present in the optimal set of variants. Note that these cells contain "-∞" in Table S4 and Table S5.

.	snp_0	snp_8	snp_9	snp_10	snp_13	snp_16	snp_17
snp_8	0	-	6	6	0	6	6
snp_9	6	6	-	0	6	0	0
snp_10	6	6	0	-	6	0	0
snp_13	0	0	6	6	-	6	6
snp_16	6	6	0	0	6	-	0
snp_17	6	6	0	0	6	0	-

Table S4: Results of S* calculation on individual 1 using dynamic programming algorithm. The dark blue cell is the maximum value of S*. Blue cells mark variant pairs that are present in the optimal set of variants, i.e., 0, 8 and 13.

.	snp_0	snp_8	snp_9	snp_10	snp_13	snp_16	snp_17
snp_8	28045	-	-	-	-	-	-
snp_9	-∞	-∞+28045	-	-	-	-	-
snp_10	-∞	-∞+28045	8070	-	-	-	-
snp_13	39890	44890	-∞	-∞+8070	-	-	-
snp_16	-∞	-∞+28045	19155	24155	-∞+44890	-	-
snp_17	-∞	-∞+28045	23425	28425	-∞+44890	33425	-

1, genotypes are coded as 0, 1, and 2

2, variants which are present in reference population are excluded

3, genotype distance calculation

4, S* calculation on each region

5, get maximum value of S*

Stage1, Null model coalescent simulations (1)

Q： Is the Introgressed Haplotypes significant (from random generation)?
A： To determine statistical significance through coalescent simulations

溯祖理论(Coalescent theory)：多个样本(或者染色体)回溯到最近共同祖先(Most recent common ancestor, MRCA)的过程。而回溯过程可以看做为逐步构建系谱树(Genealogy tree)的过程。

溯祖模拟（Coalescent simulations）：依时间向后模拟遗传重组事件。

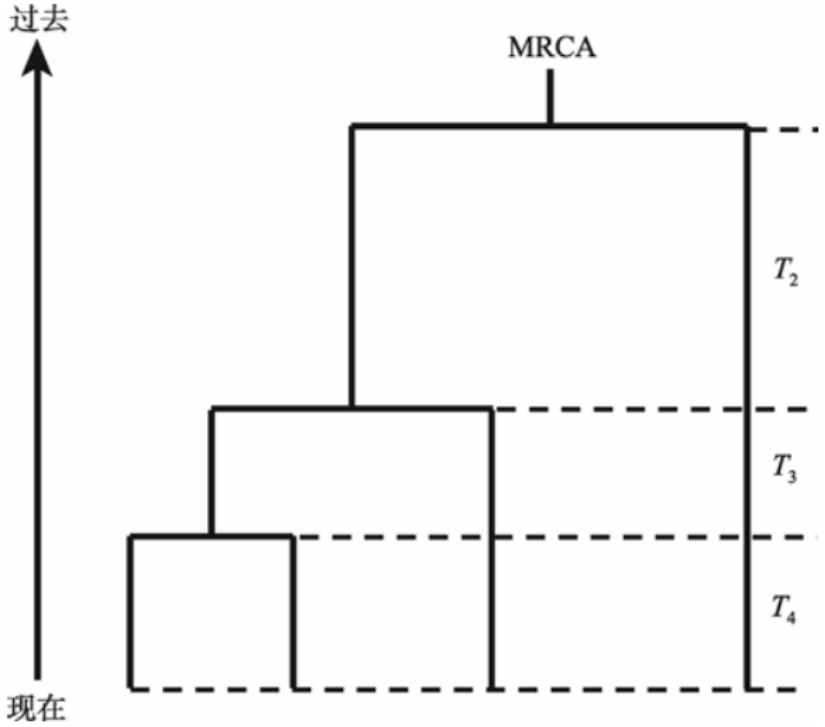


图 1　4 个个体溯祖过程

Stage1, Null model coalescent simulations (2)

To determine the expected distribution of S* under the null model of no introgression

(1) simulated sequences on a grid for all appropriate values (32*27=864)

(2) calculated S* on 20,000 50kb regions per grid point

(3) obtain the expected distribution of S* and p-value ≤ 0.01 thresholds per grid point (S6A)

(4) estimate null distribution quantiles for arbitrary windows by fitting a generalized linear model (S6B)

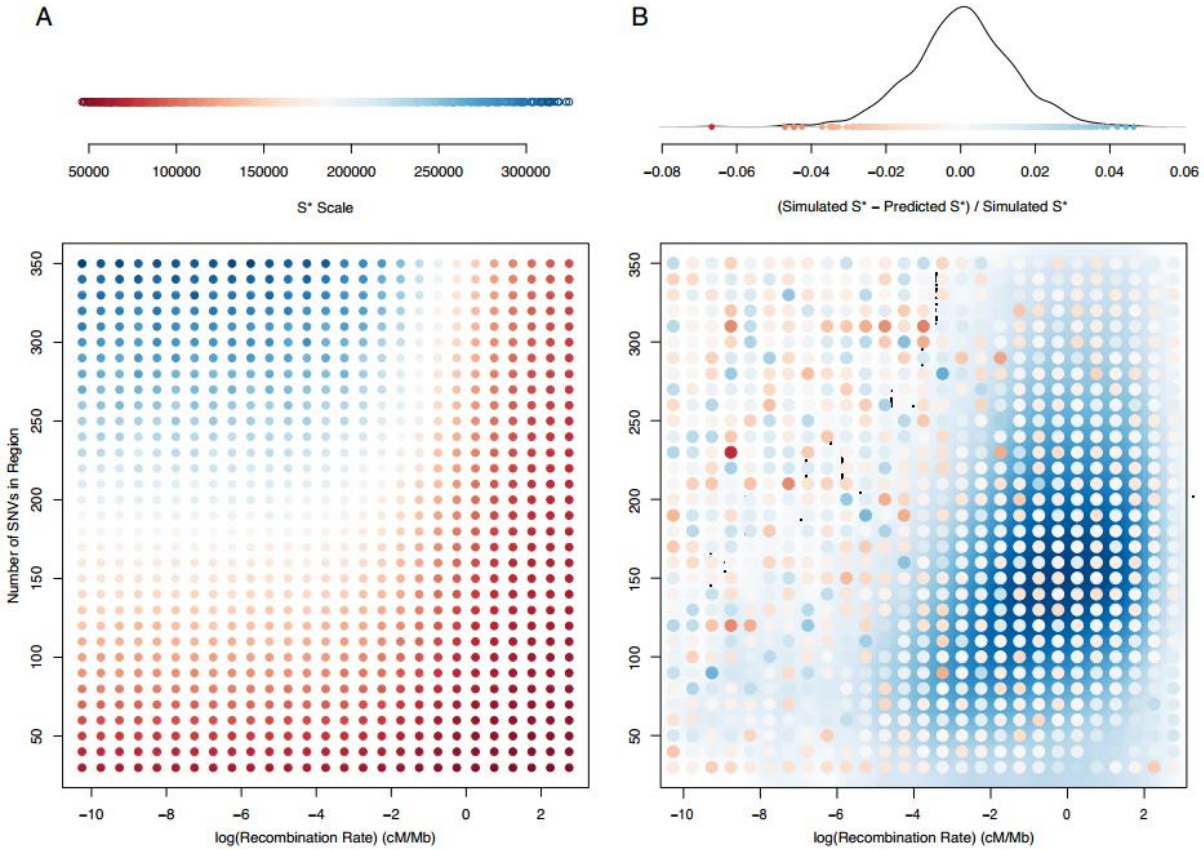


Figure S6. Computationally efficient strategy to estimate *S statistical significance**

Stage1, Evaluating the performance of S*

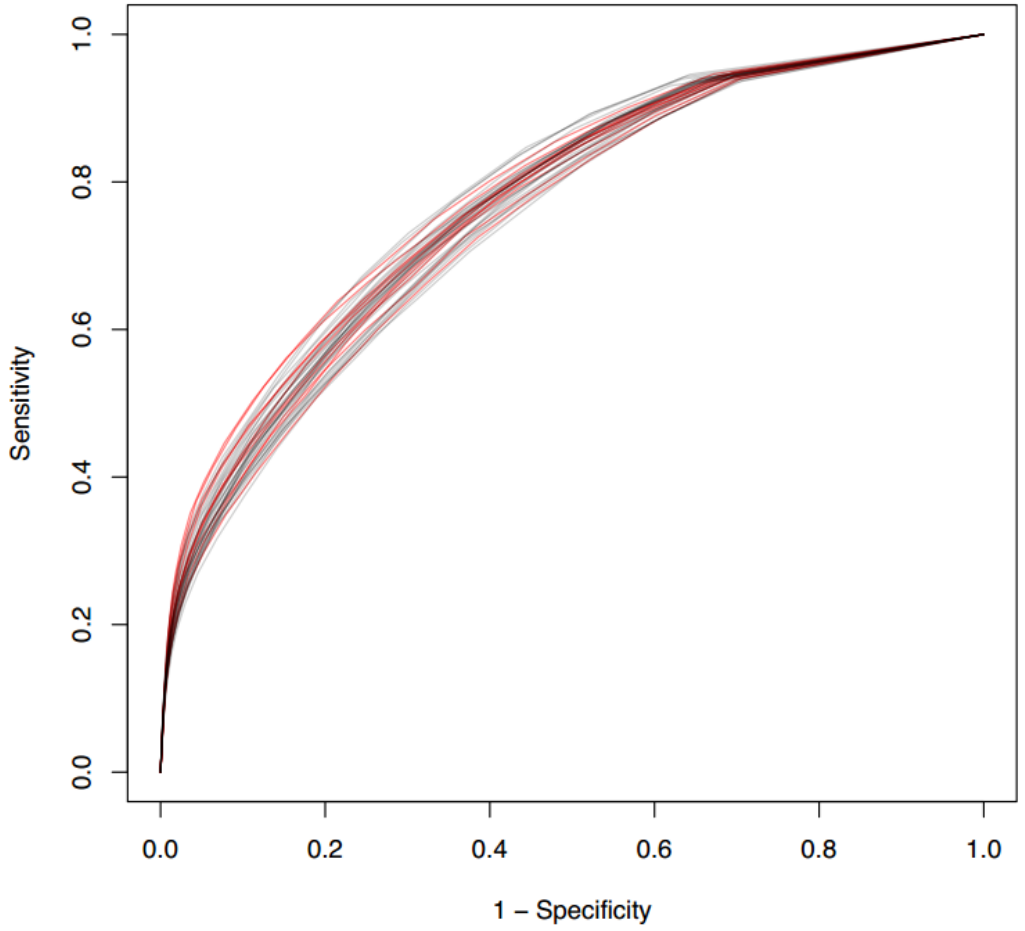


Figure S4. ROC curves for *S for different demographic models.** ROC curves for "Likely" models are shown in red.

S* can distinguish introgressed from nonintrogressed sequences

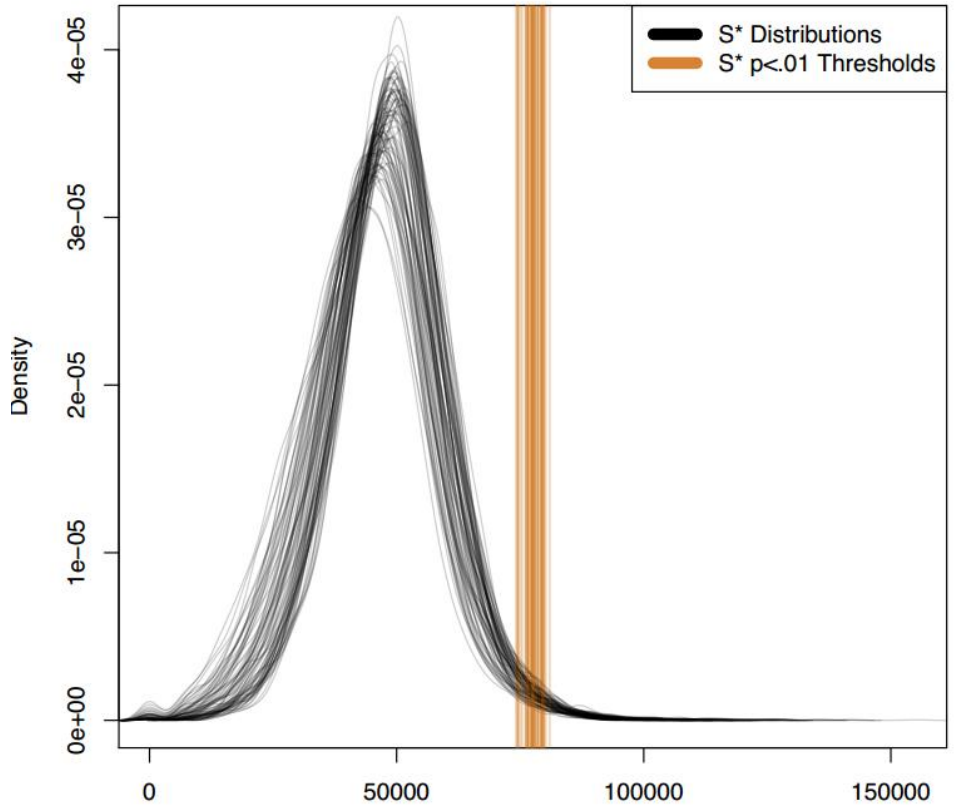


Figure S7. *S* p-values* are robust to demographic uncertainty. *S** null distribution for 10,000 50kb regions of simulated sequence data under ~700 different demographic models of modern human history (black lines). Orange lines represent *S* p-value* ≤ 0.01 thresholds for each model.

S* P values are robust to demographic uncertainty

Stage2, Calculation of Neanderthal match p-values (1)

Q: whether they match significantly more than expected by chance ?
A: calculated p-values of how well that haplotype matched the Neanderthal sequence compared to random similar haplotypes in non-introgressed portions of the genome

- Data: a call set that consists of all haplotypes with S* pvalue ≤ 0.01
- First method: SVM filtering
- Result: recover 74 Mb of sequence at FDR = 20%
- Second method: Neanderthal p-values
- Result: recover 240 Mb at FDR 5%

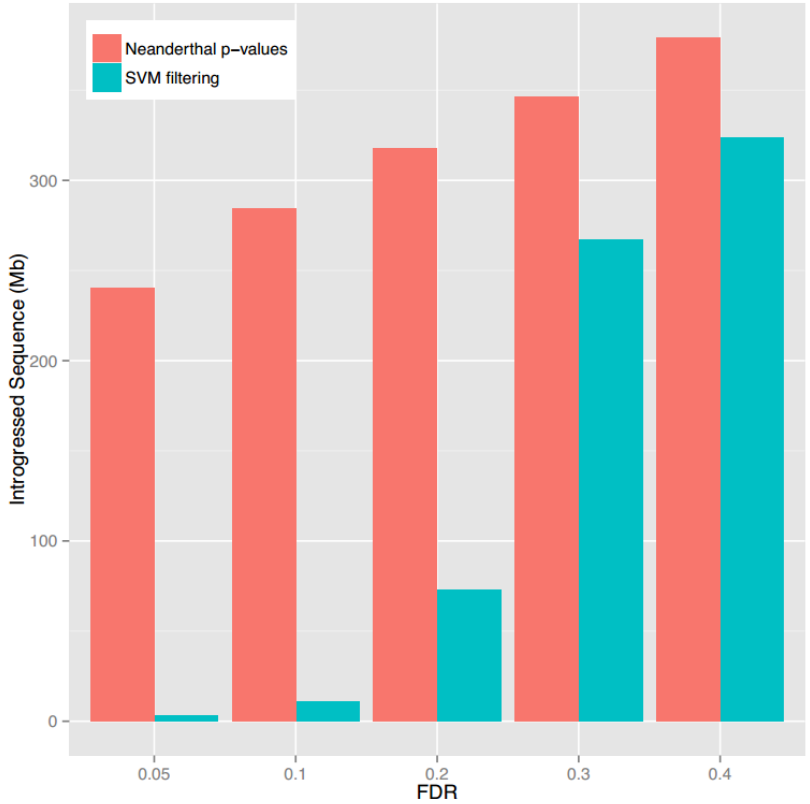


Figure S5. FDR vs Mb of introgressed sequence recovered for two methods of identifying introgressed Neanderthal sequence

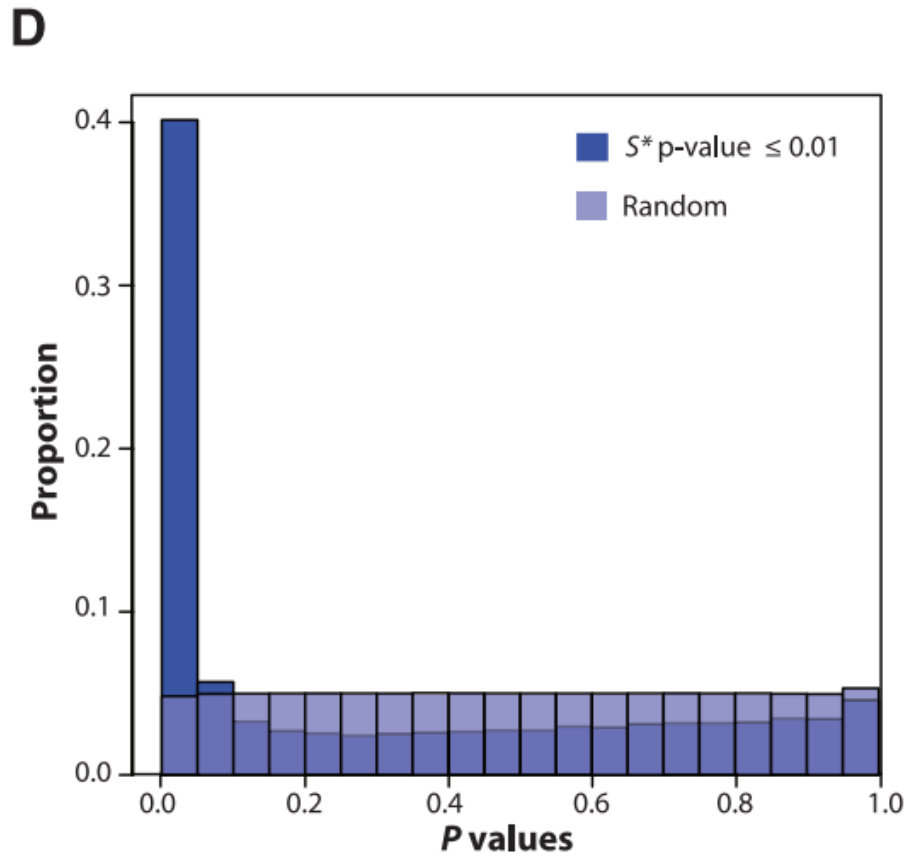
Stage2, Calculation of Neanderthal match p-values (2)

The distribution of Neandertal-match P values

(1) For candidate introgressed sequences: a strong skew toward zero

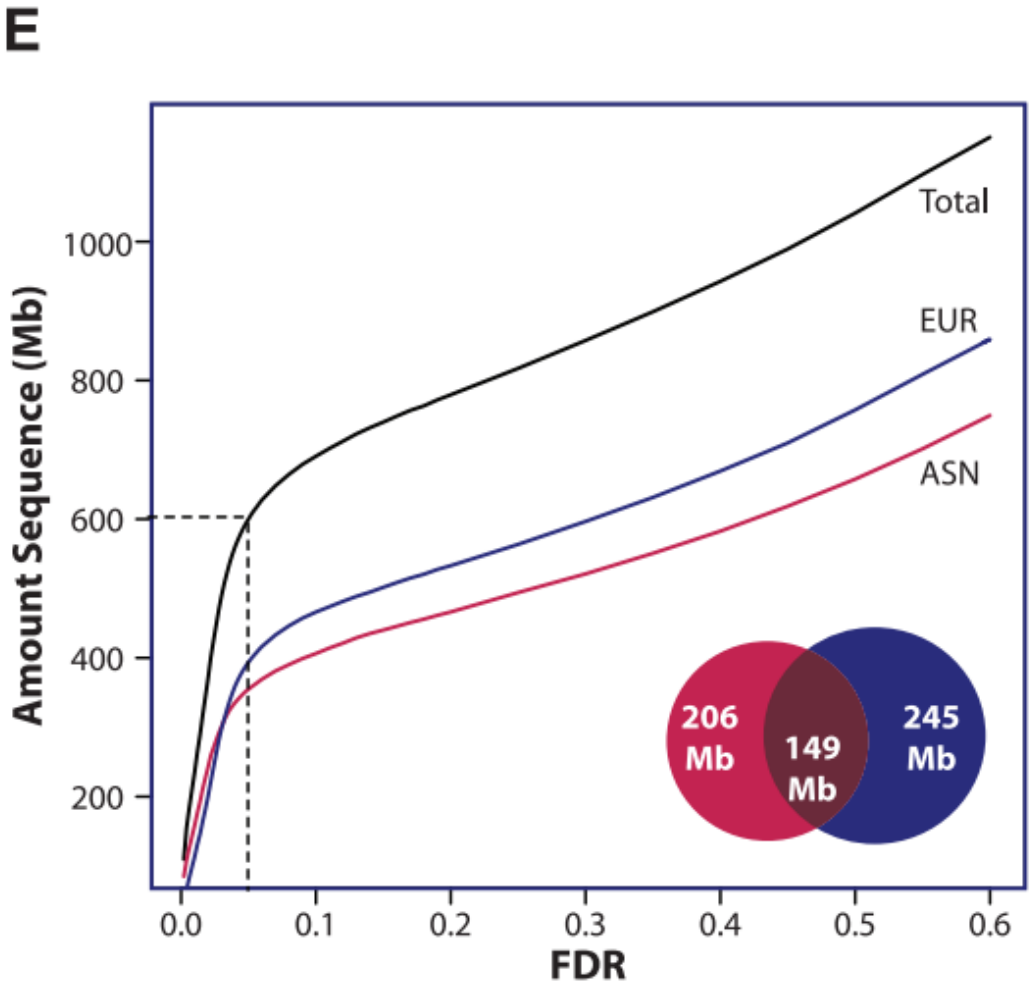
(2) For random sequences :approximately uniform

Conclusion: The statistical approach is able to distinguish between introgressed and nonintrogressed lineages



Distribution of P values testing for an enrichment of Neandertal variants for S* candidate and randomly selected regions

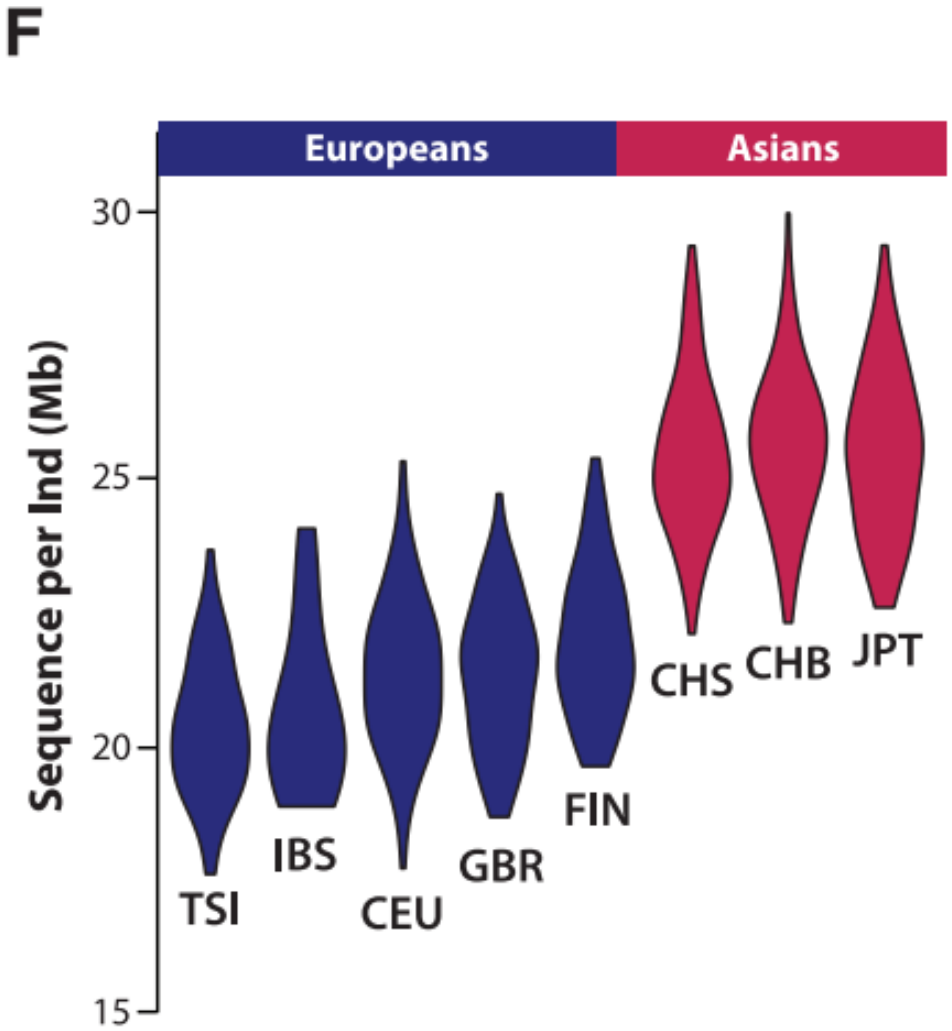
Result 1, recovering introgressed sequence



- At FDR = 5%, we identified more than 15 Gb of introgressed sequence across all individuals, spanning ~20% (600 Mb) of the Neandertal genome.

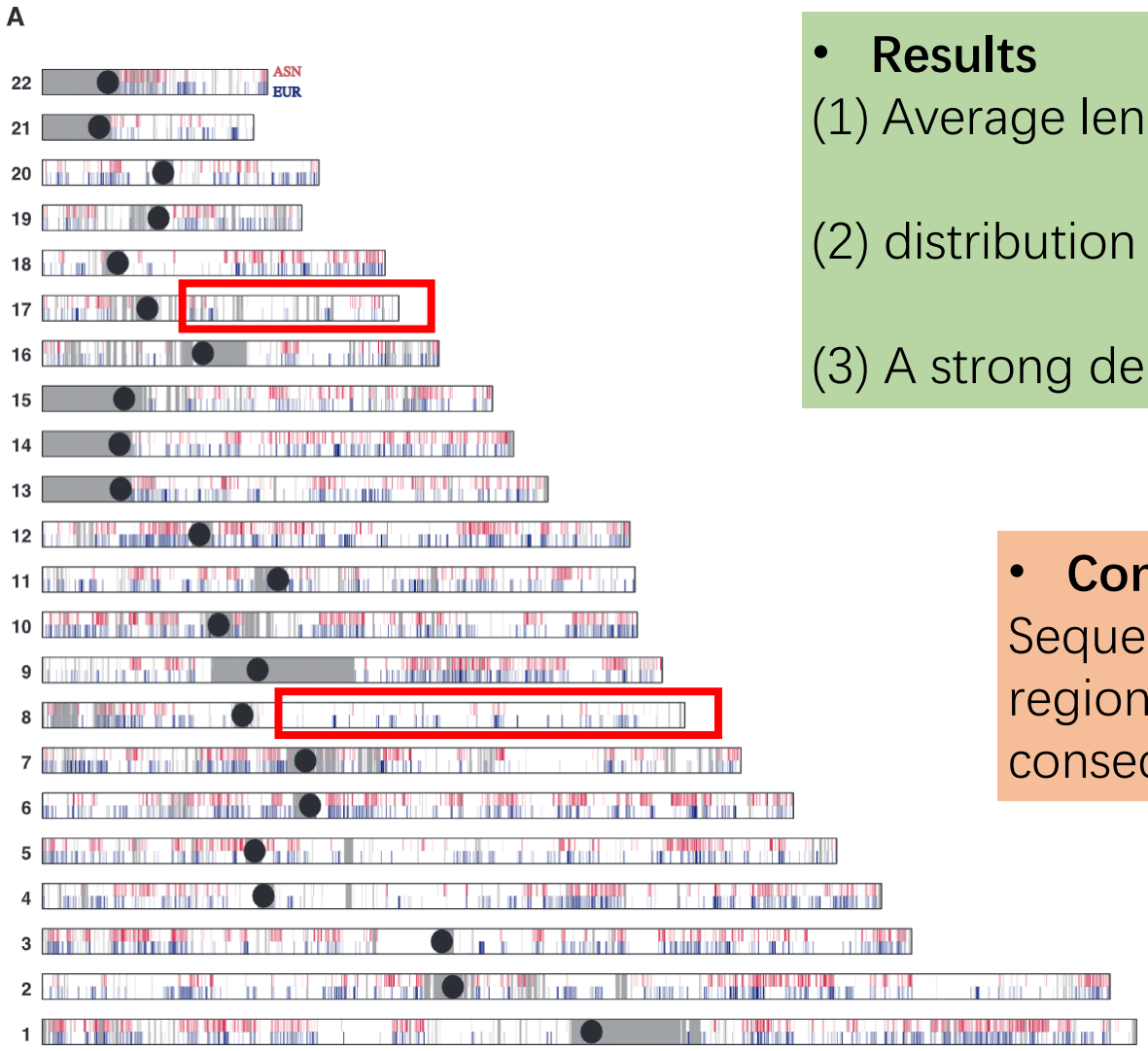
- Of the 600 Mb of distinct sequence, ~25% (149 Mb) was shared between Europeans and East Asians.

Result 1, recovering introgressed sequence



- On average, we found 23 Mb of introgressed sequence per individual
- with East Asian individuals inheriting 21% more Neandertal sequence than Europeans.

Result 2, fitness costs to hybridization



- **Results**

(1) Average length: 57 kb

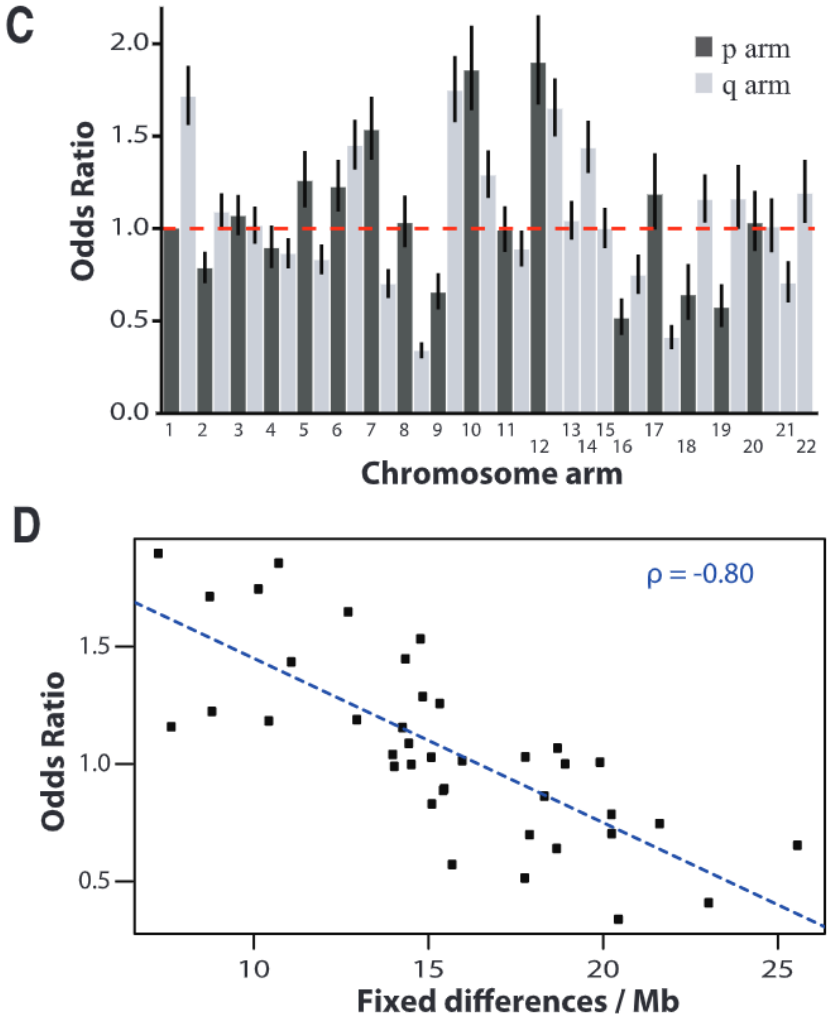
(2) distribution with heterogeneity (e.g. 8q, 17q).

(3) A strong depletion on 7q encompasses the FOXP2 locus

- **Conclusion**

Sequence divergence was a barrier to gene flow in some regions and associated with deleterious fitness consequences.

Result 2, fitness costs to hybridization

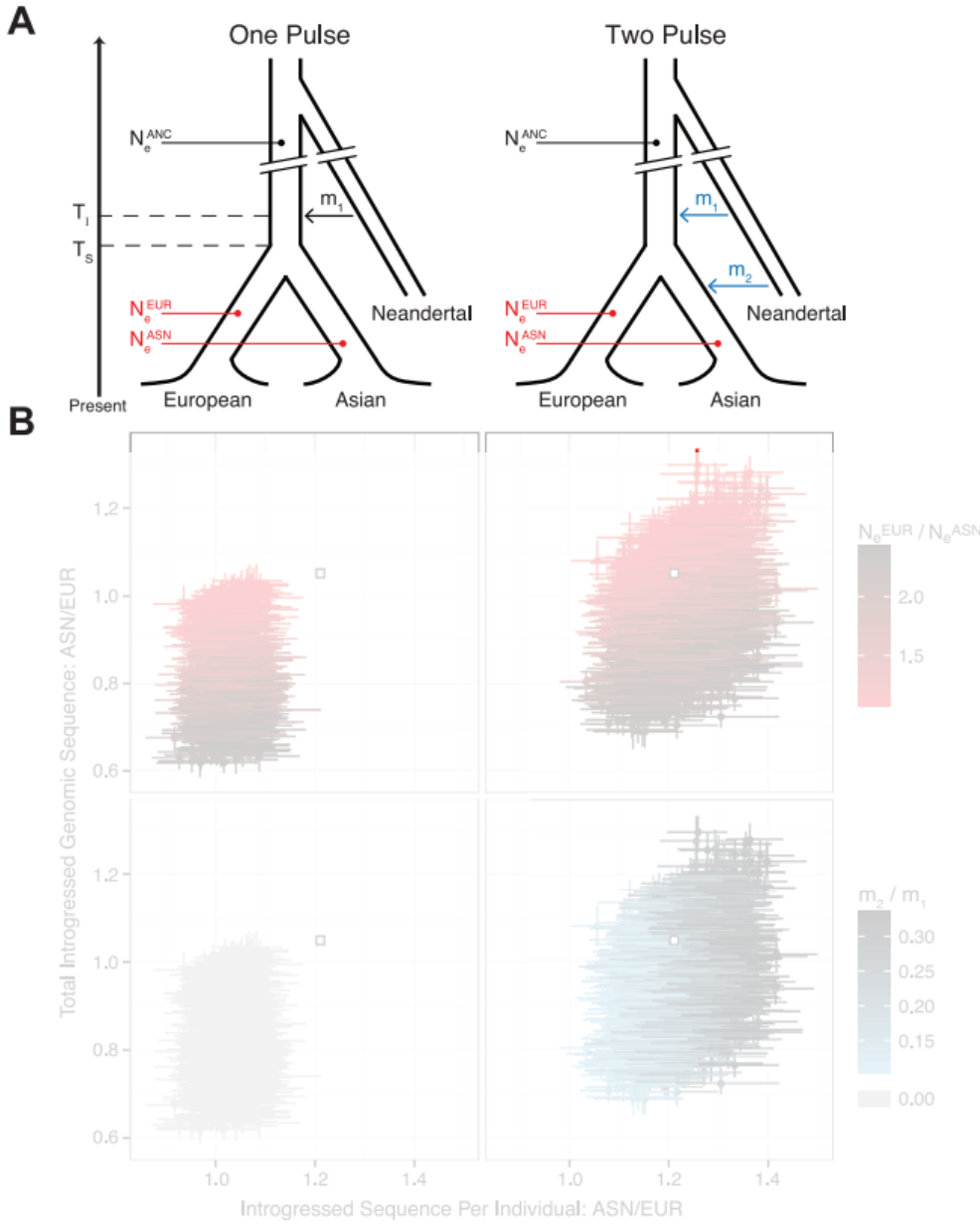


Odds of finding an introgressed lineage on each chromosomal arm calculated from a logistic regression model.
Odds ratios (ORs) are expressed using chromosome 1p as the baseline level.

Odds ratios were negatively correlated with **fixed differences** between modern humans and Neandertals.
ρ, Spearman's rank correlation coefficient.

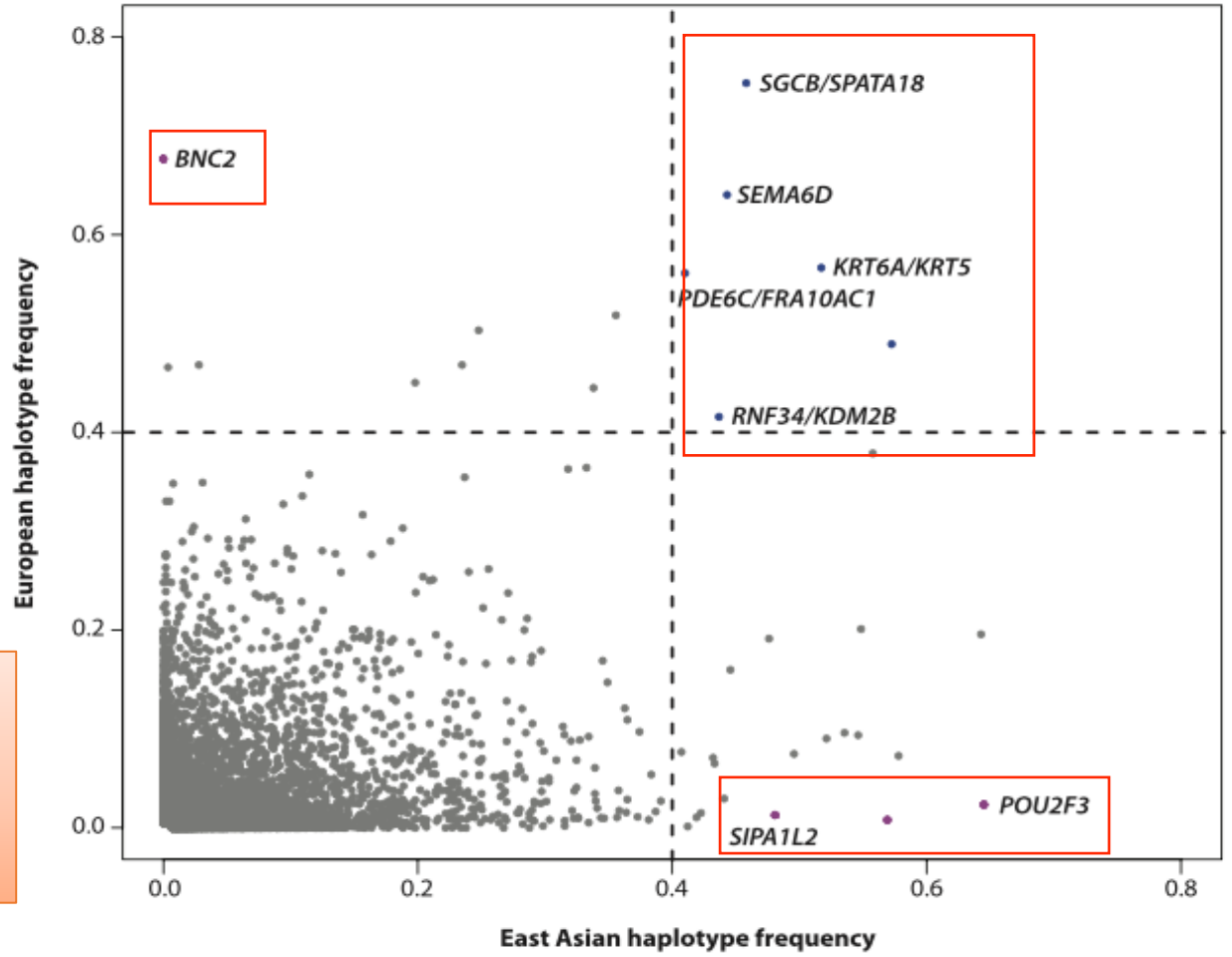
Result 3, refine admixture models

- Phenomenon: East Asians contained more introgressed sequence than Europeans.
- Hypothesis: this imbalance is possible under a single, ancestral pulse of introgression.
- Method: To simulate 2,000 random demographic models and then compare the simulated values of these summary statistics to the observed values.
- Result: observations are incompatible with the single pulse models.
- Conclusion: admixture occurred both before and after divergence of non-African modern humans.



Result 4, signature of adaptive introgression

Fig. 4. Signatures of adaptive introgression. A scatter plot of introgressed haplotype frequency in Europeans and East Asians is shown. Significantly differentiated and common shared haplotypes are represented in magenta and blue, respectively. Protein-coding genes that overlap candidate adaptively introgressed loci are also shown.



Result: identify four significantly differentiated regions and six distinct regions
Conclusion: Neandertals were a source of adaptive variation for loci involved in skin phenotypes

Result 4, signature of adaptive introgression

Q: what's the specific part of the two population?
A: BNC2 : has a frequency of ~70% in Europeans and is completely absent in East Asians;
POU2F3 : has a frequency of ~66% and is found at less than 1% frequency in European.

Table S10. Summary of population specific signatures of adaptive introgression.

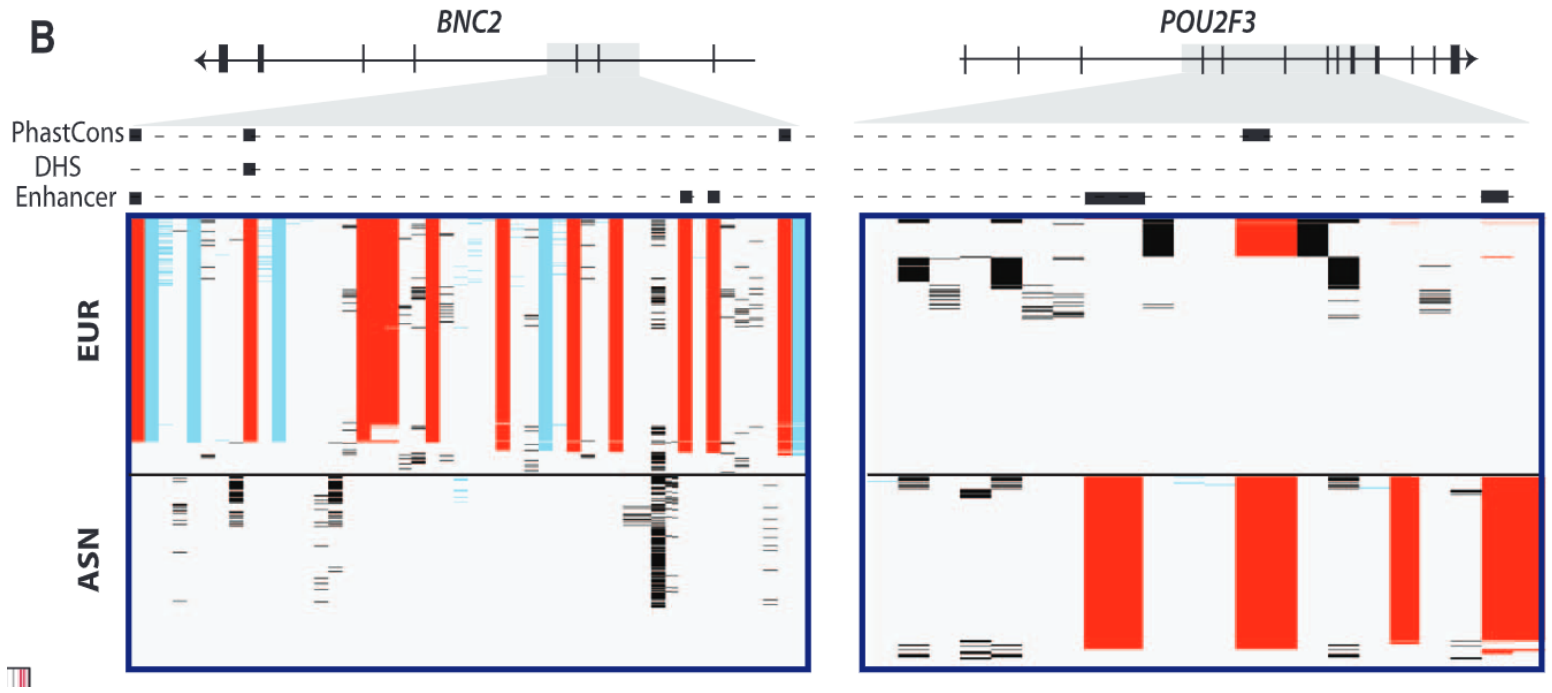
Chr	Coordinates	Length (kb)	Number SNPs $F_{ST} \geq 0.40$	Average Freq ASN[a]	Average Freq EUR[a]	Genes[b]
1	232,603,040-232,643,331	40.3	9	0.569	0.008	*SIPA1L2*
4	38,424,899-38,548,737	123.8	4	0.580	0.009	*Intergenic*
9	16,720,121-16,786,930	66.8	10	0.00	0.691	*BNC2*
11	120,154,630-120,178,414	23.8	5	0.639	0.003	*POU2F3*

[a] Average frequency denotes the mean derived allele frequency for variants with $F_{ST} \geq 0.40$.
[b] A gene name is listed if any of the highly differentiated introgressed variants overlaps a gene.

associated with skin pigmentation levels in Europeans

mediates keratinocyte proliferation and differentiation.

Result 4, signature of adaptive introgression



No coding introgressed variants were found in BNC2 or POU2F3, although several highly differentiated introgressed variants were located in functional noncoding elements suggesting that modern humans acquired adaptive regulatory sequences at these loci.

Result 4, signature of adaptive introgression

Q: What is shared part of the two population?
A1: Identifying **six** distinct regions that have introgressed haplotype frequencies greater than 40% in both populations.
A2: Neandertals provided modern humans with adaptive variation for skin phenotypes (*KRT6A*, *KRT5*).

Table S11. Summary of signatures of adaptive introgression shared between East Asians and Europeans.

Chr	Coordinates	Length (kb)	Genes
4	52,886,169-52,969,221	83.1	*SGCB, SPATA18*
7	110,141,435-110,209,984	68.5	*Intergenic*
10	95,422,244-95,466,847	44.6	*PDE6C, FRA10AC1*
12	52,880,370-52,929,370	49.0	*KRT6A, KRT5*
12	121,842,267-121,908,509	66.2	*RNF34, KDM2B*
15	47,615,219-47,646,349	31.1	*SEMA6D*

type II cluster of keratin genes

KRT6A, *KRT5* gene: Keratin-16 and keratin-17 mutations cause pachyonychia-congenita；

further suggesting that Neandertals provided modern humans with adaptive variation for skin phenotypes

Conclusions

About study

- recovering more than 15 gigabases of introgressed sequence that spans ~20% of the Neandertal genome
- fitness costs to hybridization;
- admixture occurred both before and after divergence of non-African modern humans;
- Neandertals were a source of adaptive variation for loci involved in skin phenotypes.

About approach

- working in the lack of an archaic reference sequence (fossil-free);
- finding and characterizing of previous unknown hominids that interbred with modern humans.

Significances

- revealing insights into hominin evolution;
- the population genetics characteristics of archaic hominins;
- how introgression has influenced extant patterns of human genomic diversity;
- narrowing the search for genetic changes that endow distinctly human phenotypes.

Limitations

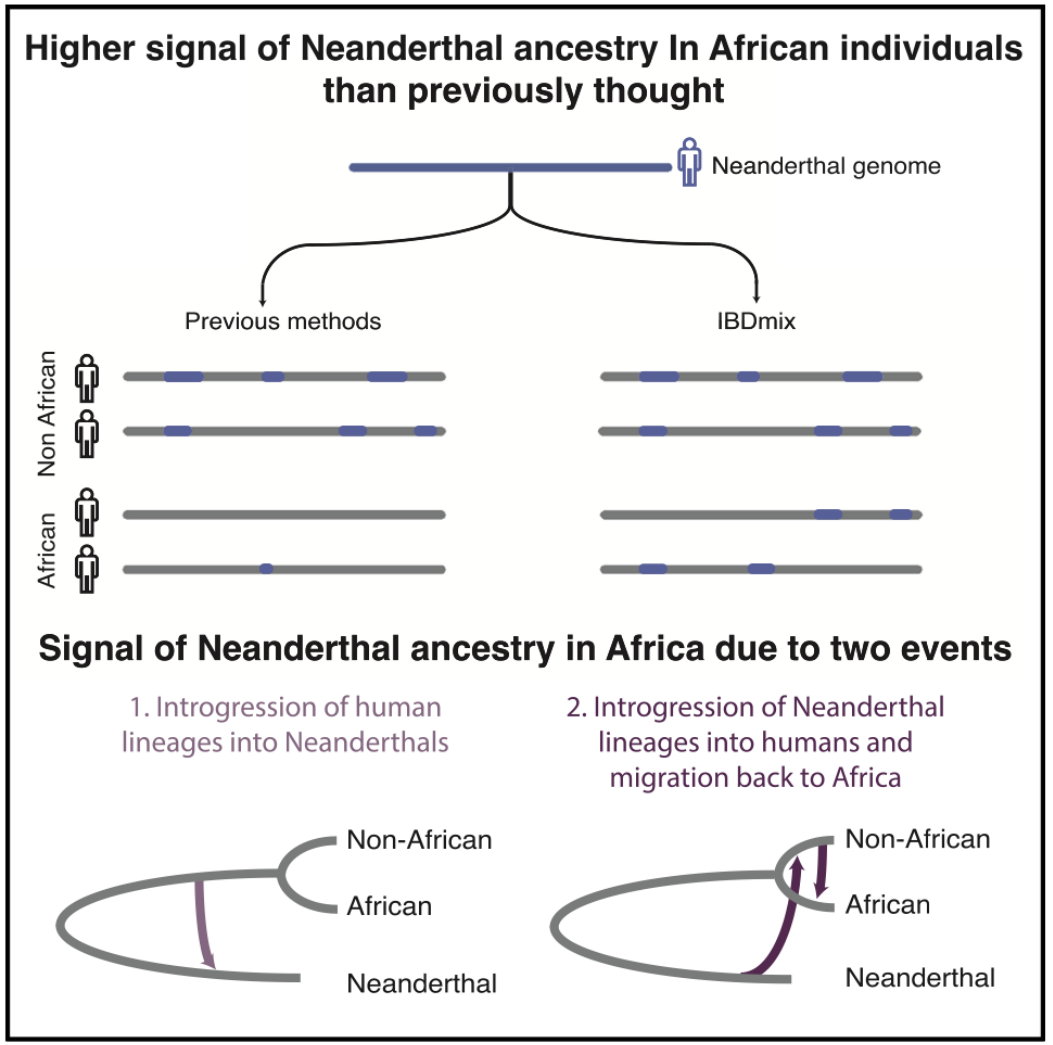
- additional unexplored models may provide a better fit to the data;
- refining demographic models of hominin evolution is an important area for future work.

What's new about this area?

Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals

Lu Chen,[1,4] Aaron B. Wolf,[1,2,4] Wenqing Fu,[3] Liming Li,[1] and Joshua M. Akey[1,5,*]
[1]The Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540, USA
[2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[3]1 Microsoft Way, Redmond, WA 98052, USA
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: jakey@princeton.edu
https://doi.org/10.1016/j.cell.2020.01.012

提出一种全新的基因组鉴定方法（不使用"非杂交"群体作为参考基因组）IBDmix，对千人基因组2504个不同族群的现代人数据进行分析，首次在非洲人中鉴定尼安德特序列，同时对非非洲人群体重新分析，拓展、更新现代人群的尼安德特成分图谱，并探索发现适应性杂交对人类进化及现代人表型变异的作用和影响。



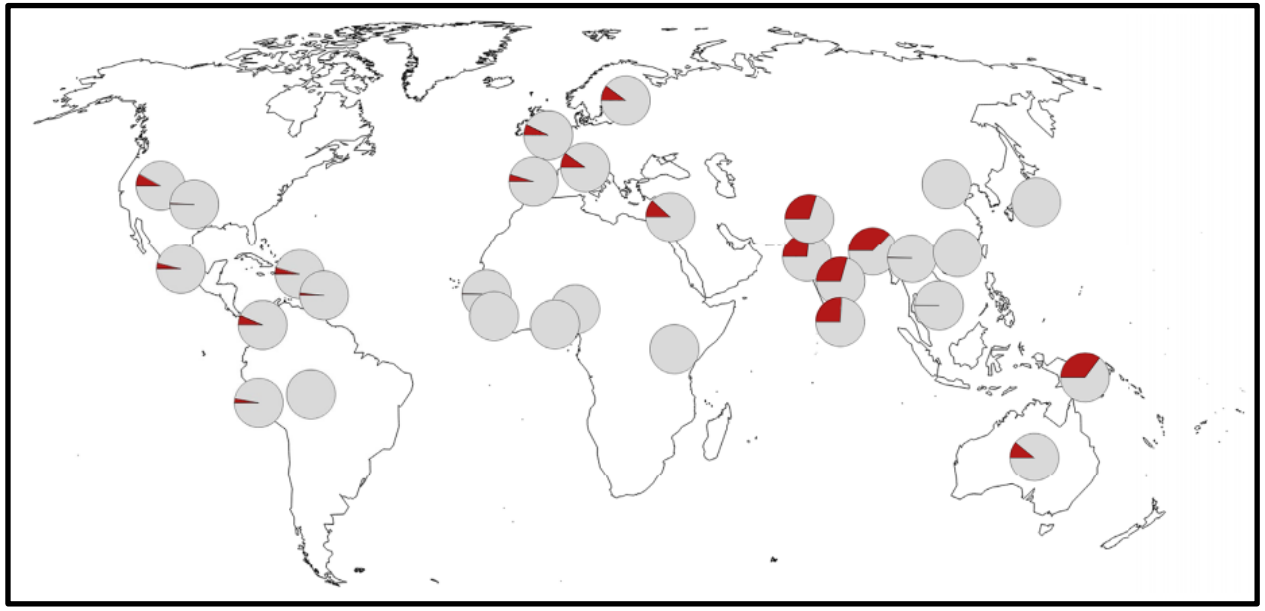
What's new about this area?

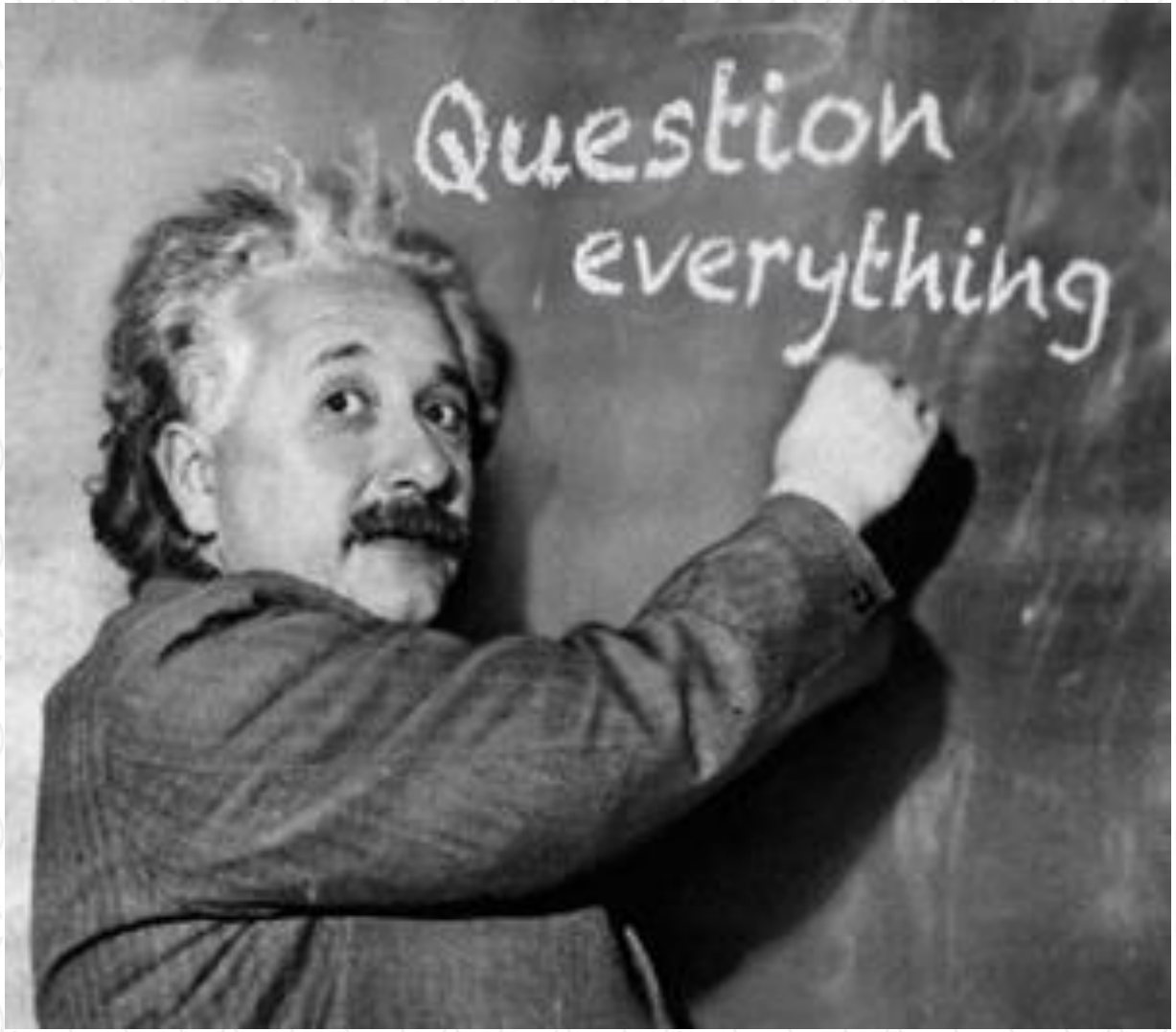
Article

The major genetic risk factor for severe COVID-19 is inherited from Neanderthals

研究了一组与感染冠状病毒SARS-CoV-2的患者住院和呼吸衰竭的风险较高有关的基因，发现一些人从尼安德特人祖先（Neanderthal ancestor）那里遗传的基因可能会增加他们罹患重症COVID-19的可能性。

这种单倍型在欧洲约有16%的人口和南亚一半的人口中发现，而在非洲和东亚则不存在。





THANKS