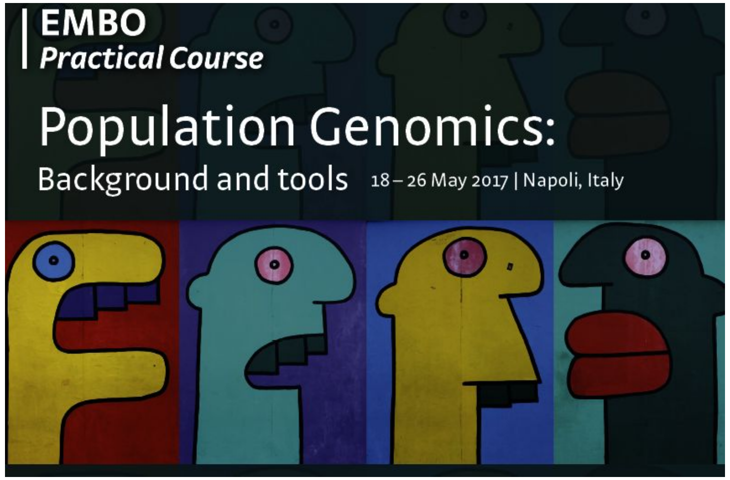
UNIVERSITY OF
WESTMINSTER

Measures of natural selection

EMBO practical course 2017

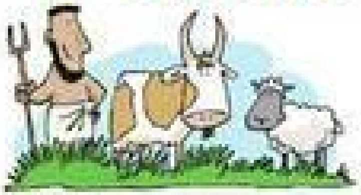
Pascale Gerbault

p.gerbault@westminster.ac.uk



EMBO
excellence in life sciences

A SHORT HISTORY OF DNA

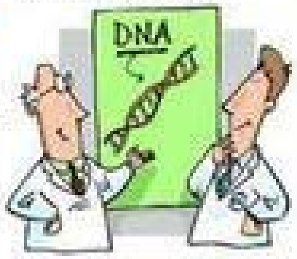


8000 BC
Long before the discovery of DNA, early farmers were using selective breeding to improve their crops and livestock. They kept back the best seed and offspring from their farms to begin the next generation.

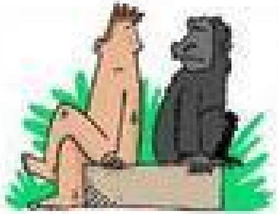
1859
Charles Darwin publishes his theory of evolution through natural selection. It was only long after his death that his ideas finally became widely accepted.

1863
Gregor Mendel, a monk in Austria, first documents hereditary traits in garden peas.

1953
James Watson and Francis Crick accurately describe the molecular structure of DNA as a double helix.

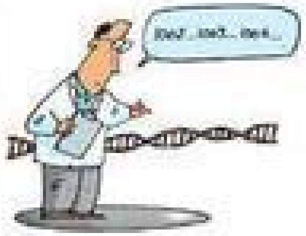


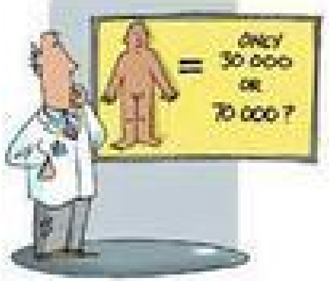
1966
The genetic code is revealed. It is established that a sequence of three nucleotide bases corresponds to each of 20 amino acids in the production of proteins.
Since then a further two amino acids have been discovered.



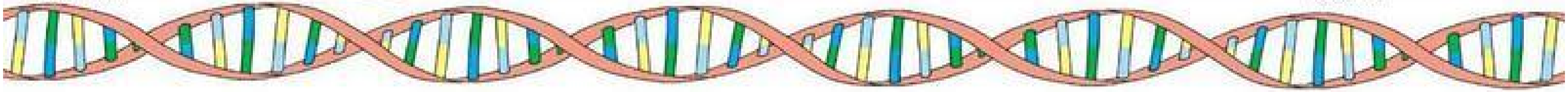
1972
The DNA composition of humans is found to be 99% similar to chimpanzees and gorillas.

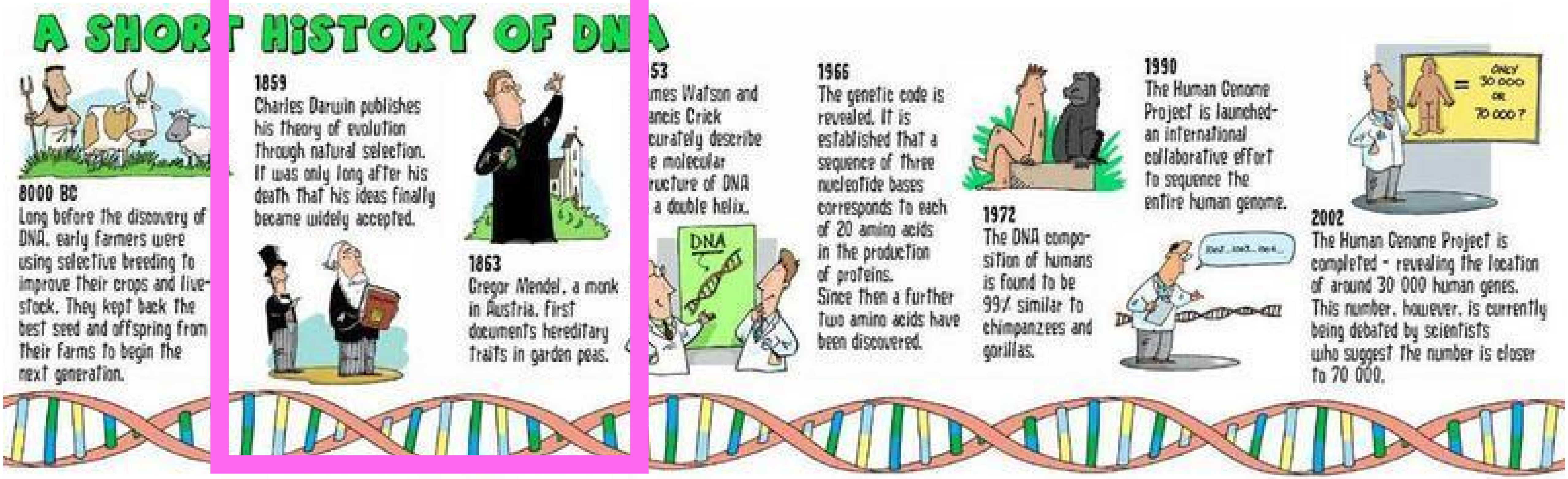
1990
The Human Genome Project is launched- an international collaborative effort to sequence the entire human genome.





2002
The Human Genome Project is completed - revealing the location of around 30 000 human genes. This number, however, is currently being debated by scientists who suggest the number is closer to 70 000.





A SHORT HISTORY OF DNA

8000 BC
Long before the discovery of DNA, early farmers were using selective breeding to improve their crops and livestock. They kept back the best seed and offspring from their farms to begin the next generation.

1859
Charles Darwin publishes his theory of evolution through natural selection. It was only long after his death that his ideas finally became widely accepted.

1863
Gregor Mendel, a monk in Austria, first documents hereditary traits in garden peas.

53
mes Watson and
ancis Crick
curately describe
e molecular
ructure of DNA
a double helix.

1966
The genetic code is revealed. It is established that a sequence of three nucleotide bases corresponds to each of 20 amino acids in the production of proteins.
Since then a further two amino acids have been discovered.

1972
The DNA composition of humans is found to be 99% similar to chimpanzees and gorillas.

1990
The Human Genome Project is launched- an international collaborative effort to sequence the entire human genome.

2002
The Human Genome Project is completed - revealing the location of around 30 000 human genes. This number, however, is currently being debated by scientists who suggest the number is closer to 70 000.

- Why are Mendel's and Darwin's works still relevant today?
  - Examples of deliberate cross breeding…
  - => possible because most populations are genetically variable

- Darwinian concepts of selection have been rendered quantitative and measurable in real populations, thanks to methodological and technological advances.
- Evolutionary genetics contributed to the understanding of many adaptive traits, e.g. in humans lactase persistence, skin pigmentation, in mice coat color
- Approaches: (1st) phenotype hypothesised to be adaptive; (2nd) identification of underlying locus/loci
- Genomic advancements: test genomic evidence of selection on putative traits > uncovering candidate genetic regions through genome scans

What evidence is there for evolution?

- Change in allele frequencies > drift?
- Change in allele frequencies > natural selection? >> adaptation “meaningful variation”

What evidence is there for evolution?

- Change in allele frequencies > drift?
- Change in allele frequencies > natural selection? >> adaptation "meaningful variation"
- => selection affects the PHENOTYPE ~ genomic variation of *functional significance*
- One of the strongest selection acting on humans?

An example of ongoing natural selection that affects humans:

The evolution of drug-resistant bacteria

7:22min video on Natural selection and the bacterial resistance
https://www.youtube.com/watch?v=7VM9YxmULuo

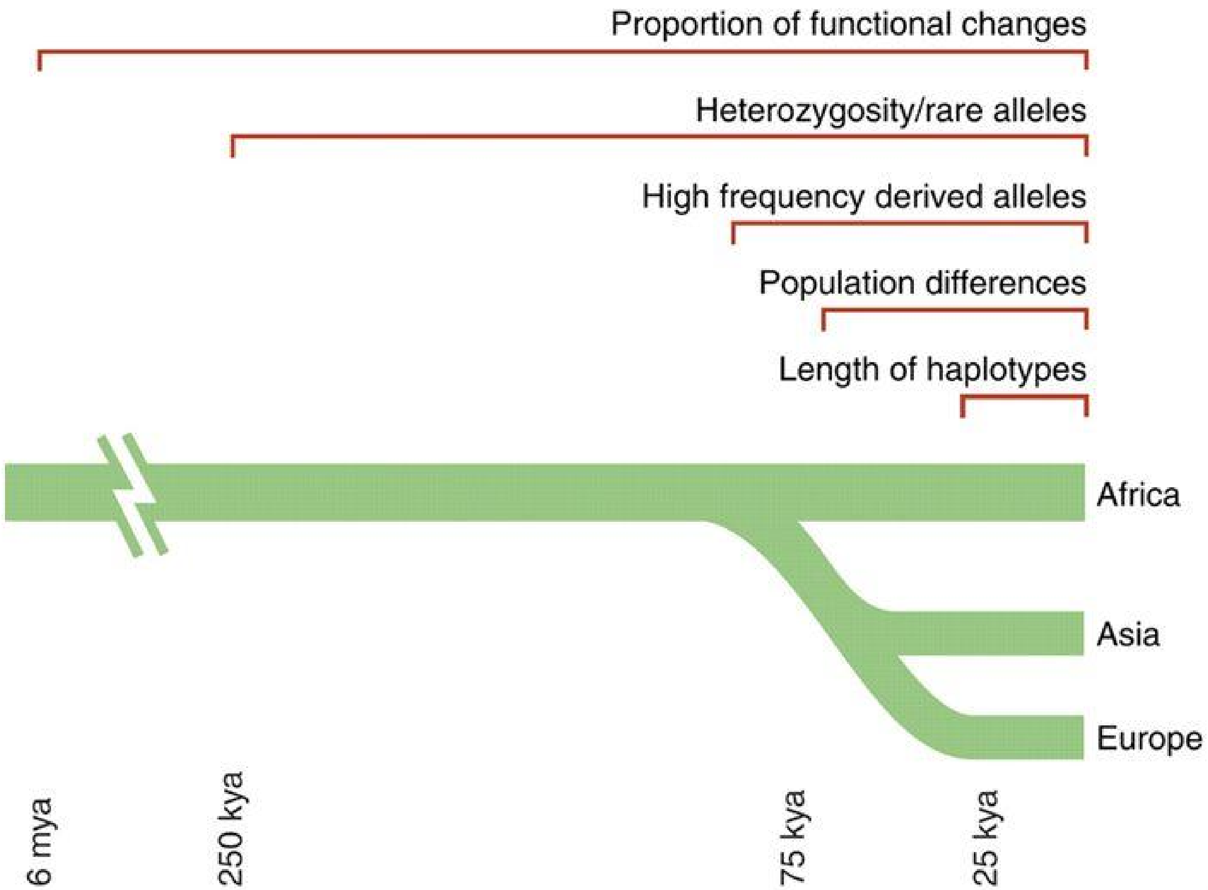
‘SELECTION’: One term, one process?

- Macro/micro evolutionary scale
- Negative, diversifying, positive selection
- Selective sweep, hard/soft sweep

Methods to infer selection

- Different methods have different power according to the time and strength of selection

Fig. 1. Time scales for the signatures of selection.


Proportion of functional changes
Heterozygosity/rare alleles
High frequency derived alleles
Population differences
Length of haplotypes
Africa
Asia
Europe
6 mya
250 kya
75 kya
25 kya

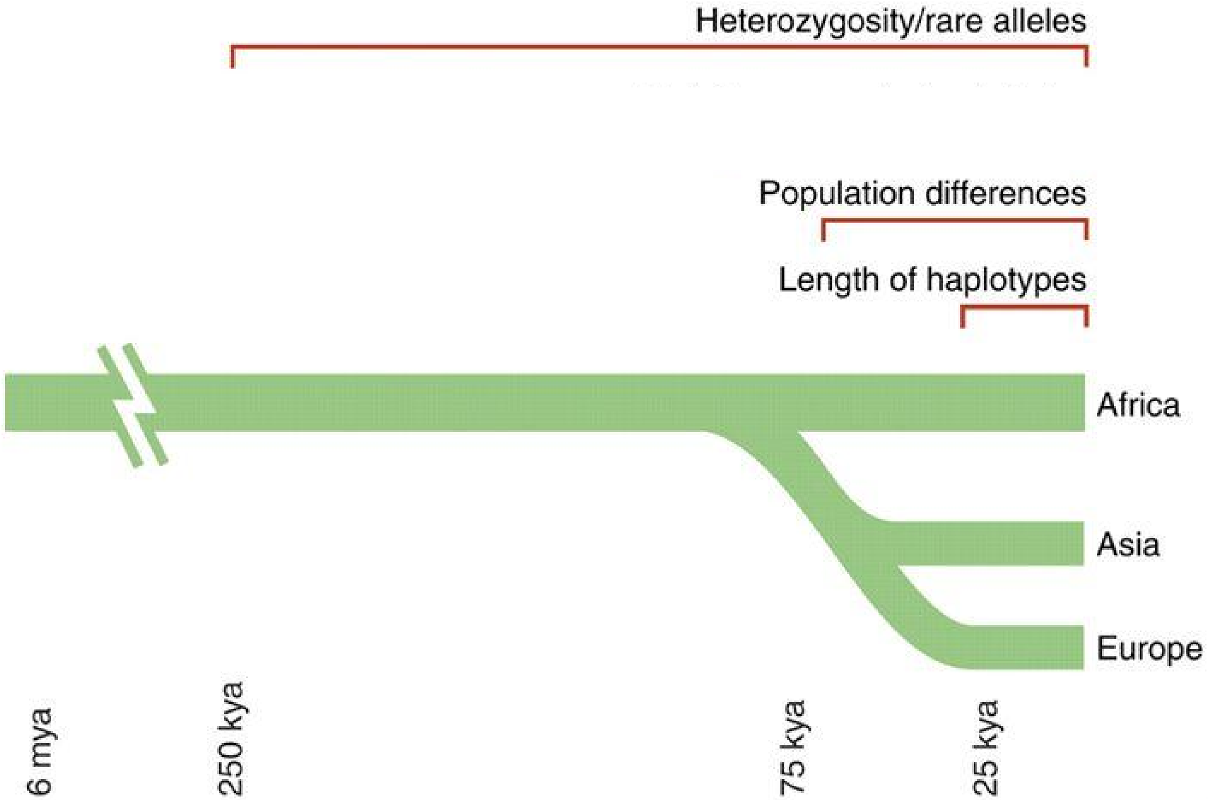
P C Sabeti et al. Science 2006;312:1614-1620

Published by AAAS

Methods to infer selection

- Different methods have different power according to the time and strength of selection

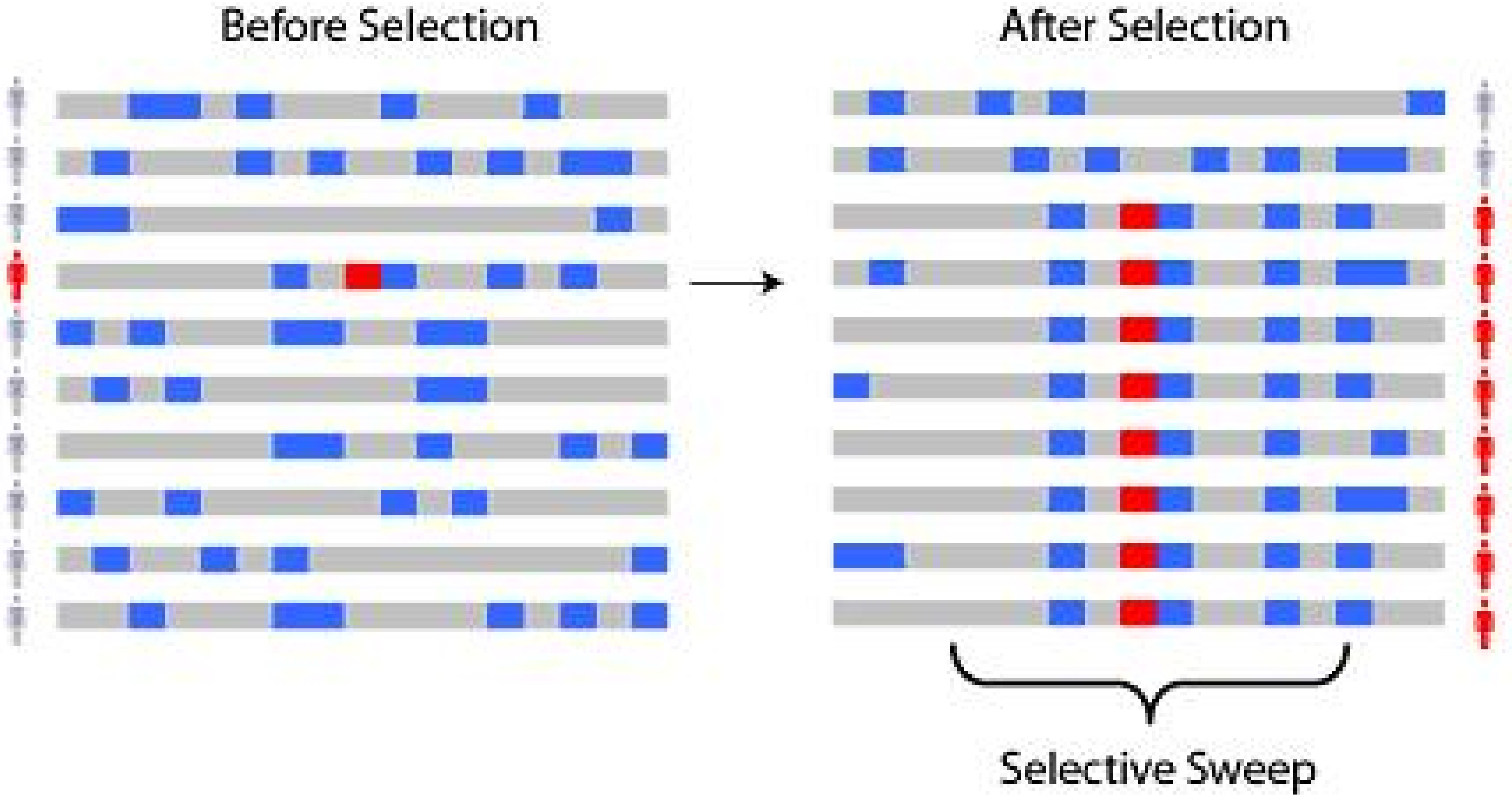
Fig. 1. Time scales for the signatures of selection.



P C Sabeti et al. Science 2006;312:1614-1620

Published by AAAS

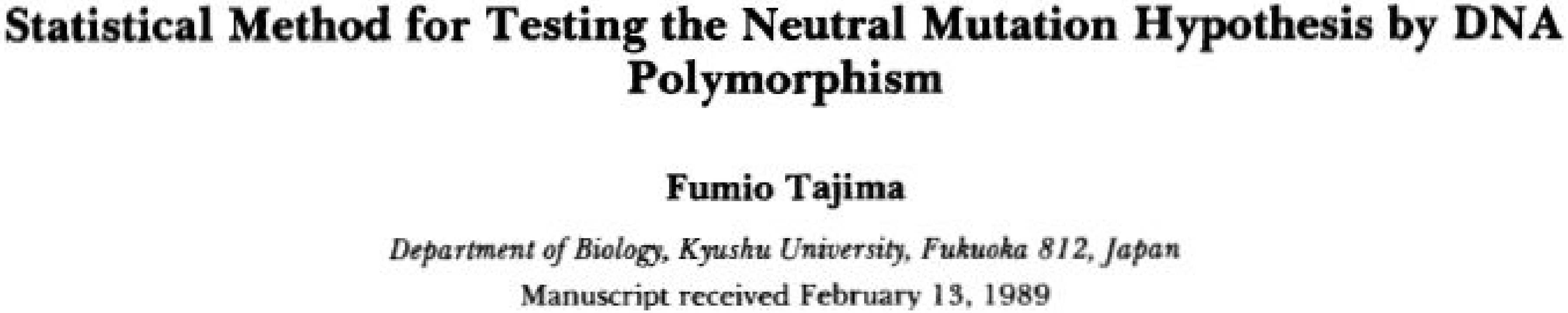
Heterozygosity/rare alleles



- As a new positively-selected allele (red) rises to high frequency, nearby linked alleles on the chromosome 'hitchhike' along with it to high frequency, creating a 'selective sweep.'

Ancestral alleles
Derived alleles
Selected allele

Schaffner & Sabeti 2008 Nature Education



Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism

Fumio Tajima

Department of Biology, Kyushu University, Fukuoka 812, Japan

Manuscript received February 13, 1989

- Tajima’s D compares two estimators of genetic diversity: the average number of nucleotide differences θ_T and the number of segregating sites θ_W
- If the population is at equilibrium: Tajima’s D ~ 0
- Rare variants contribute little to the number of pairwise differences
- After a selective sweep, the number of segregating sites will be >> than the average number of pairwise differences
- Because after a selective sweep, most sequences will be the same -> when mutations occur => rare
- rare mutations -> low value of the average number of nucleotide differences in comparison to the number of segregating sites
- If number of segregating sites >> average number of pairwise differences => Tajima’s D < 0
- Smaller values of Tajima’s D ~ positive selection... or population expansion

$$\hat{\theta}_T = \frac{\sum_{i<j} d_{ij}}{n(n-1)/2} \qquad \hat{\theta}_W = \frac{S}{\sum_{i=1}^{n-1} 1/i}$$

$$D = \frac{\hat{\theta}_T - \hat{\theta}_W}{\sqrt{\hat{V}(\hat{\theta}_T - \hat{\theta}_W)}}$$

https://arundurvasula.wordpress.com/2015/02/18/interpreting-tajimas-d/

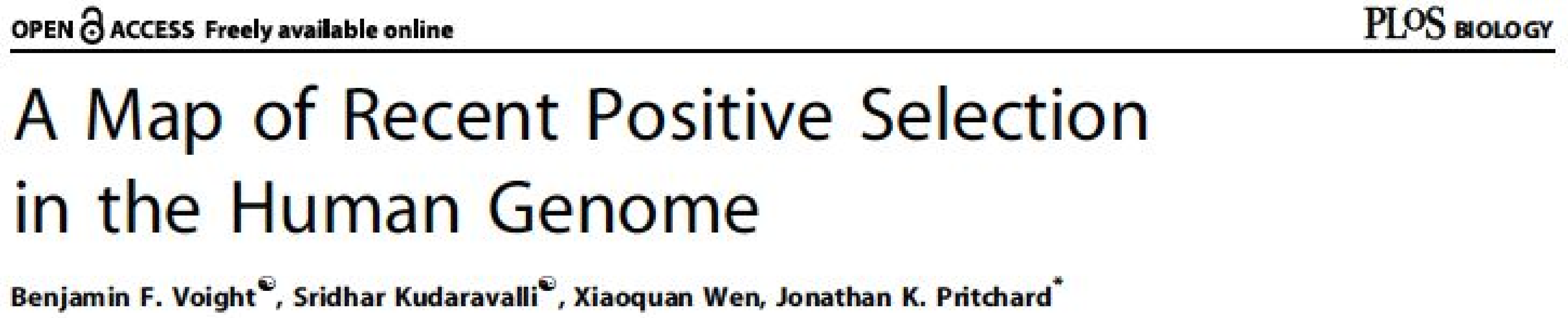
Population differences

- Different populations ~ different environments
- Different environments ~ distinct adaptive traits
- Selection acting in an environment on a locus in a local population but not all populations
- Population differentiation can be measured with F_{ST} (Wright's fixation index): compares the variance of allele frequencies ***between*** populations
  - Large F_{ST} values strong differentiation between populations: directional, positive selection
  - Small F_{ST} values ~ populations are homogeneous: balancing selection

Length of haplotype

- Natural selection leaves footprints on genomes

- ~ Tajima’s D: selective sweep affects whole haplotypes -> extension
- Extended haplotype homozygosity (EHH); long range haplotype(LRH); integrated haplotype score (iHS)



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[*]

Length of haplotype: the composite iHS statistic

1/ measure EHH ~ the decay of identity as a function of distance of haplotypes that carry a specific (‘core’) allele at one end
EHH varies between 0 and 1, where haplotype homozygosity for the core SNP starts at 1 and decays to 0 with increasing distance from the core SNP
EHH is computed on haplotypes with the ancestral allele and with the derived allele and the area under the curve (EHH decay over distance) is kept ~ integrated EHH (iHH_A and iHH_D)

2/ $iHS = iHH_A/iHH_D$, where

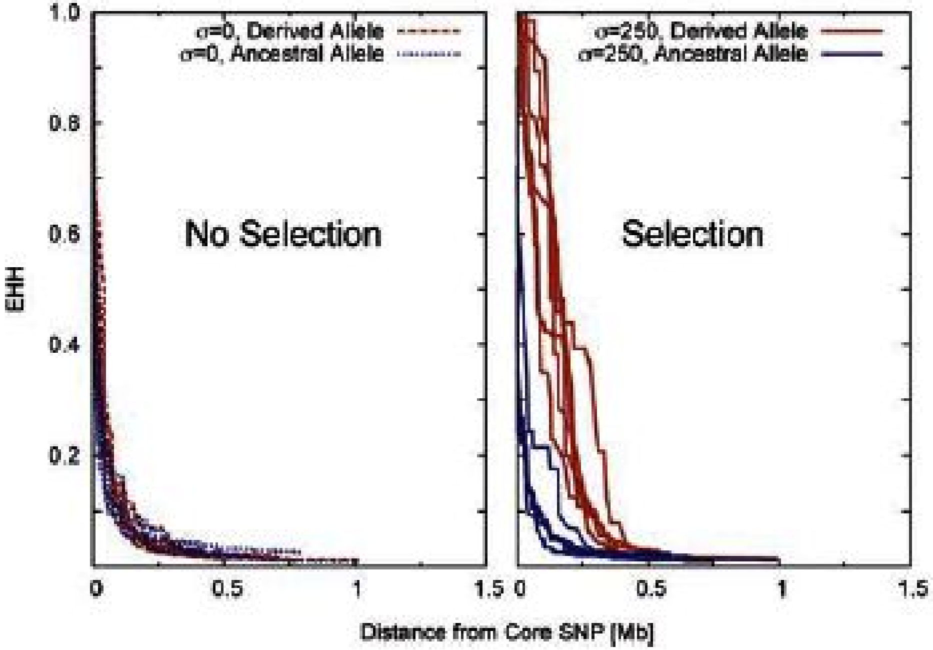
- iHS=1 if decay is similar between haplotypes carrying either alleles
- iHS << 0 if haplotypes carrying the derived allele are longer
- iHS >> 0 if haplotypes carrying the ancestral allele are longer

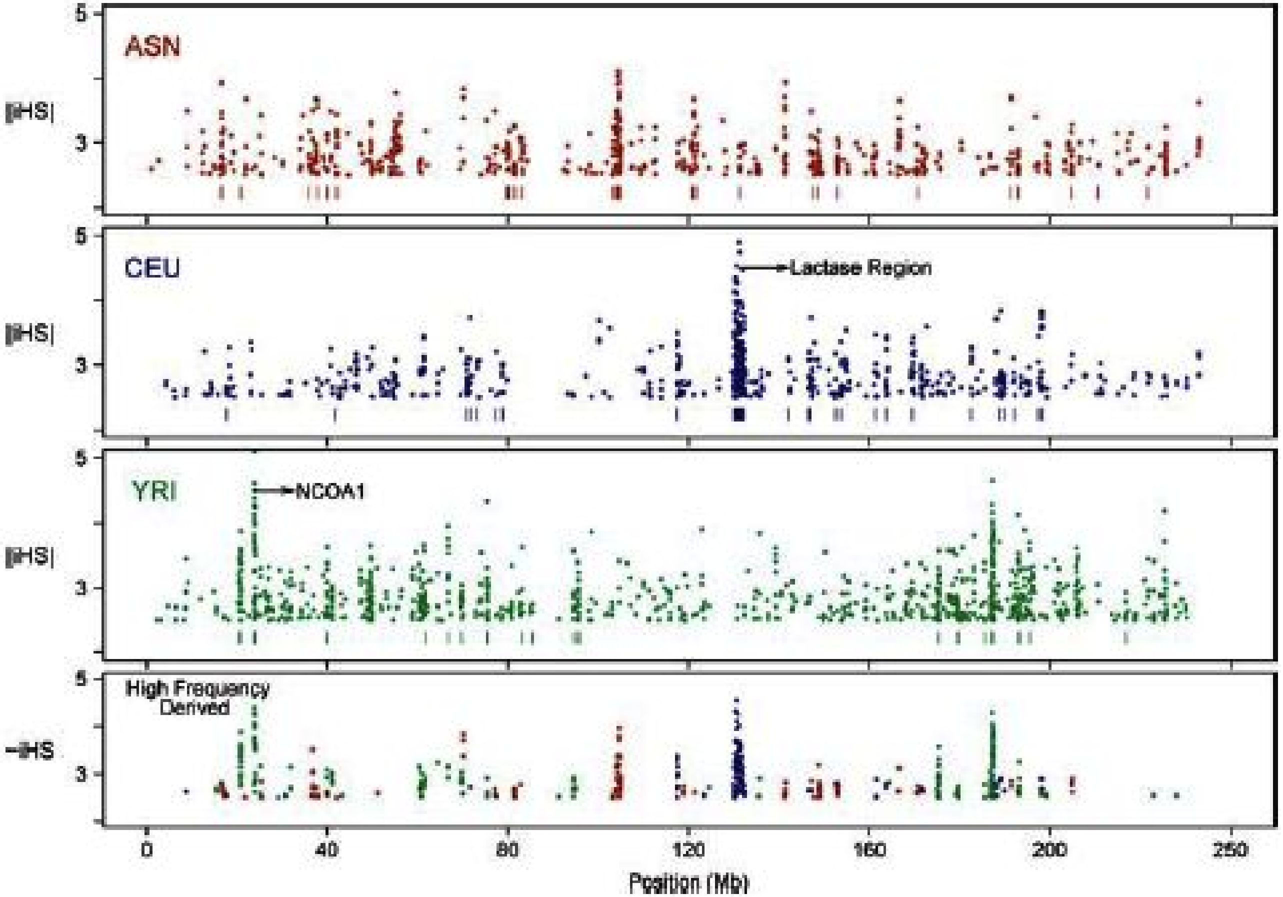
!! Sensitive to allele frequency: adjust/allele frequency and iHS mean and variance

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*

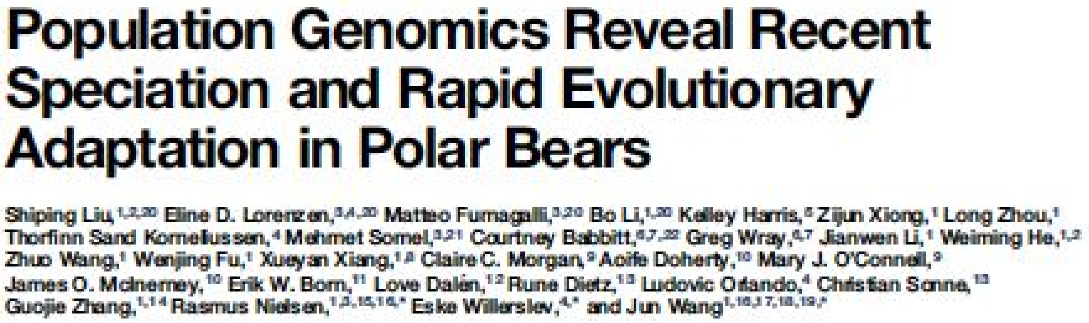




Many tests exist…

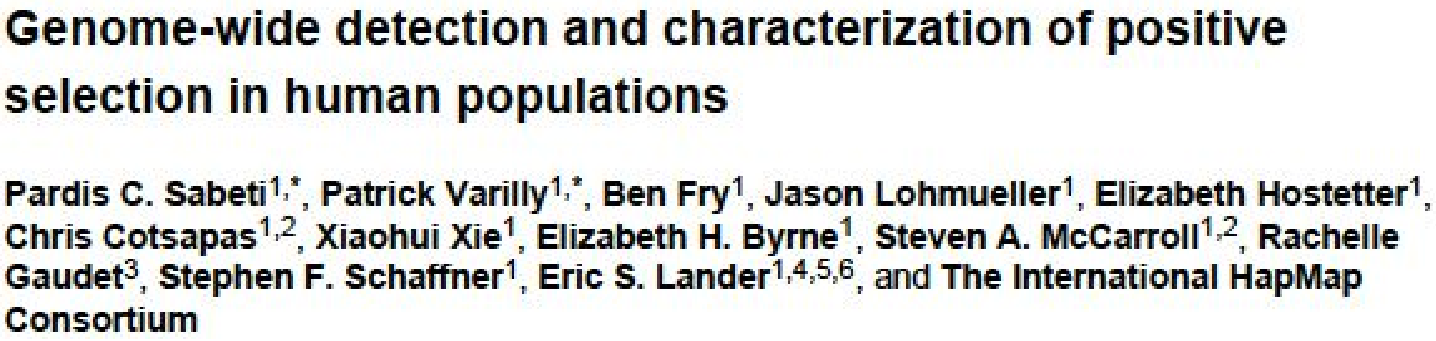
- PBS and XPEHH ~ population differences
- EHH ~ long range haplotypes
- Reviews:
  - Nielsen 2005 Ann Rev Genet
  - Vitti et al. 2013 Ann Rev Genet

Some examples



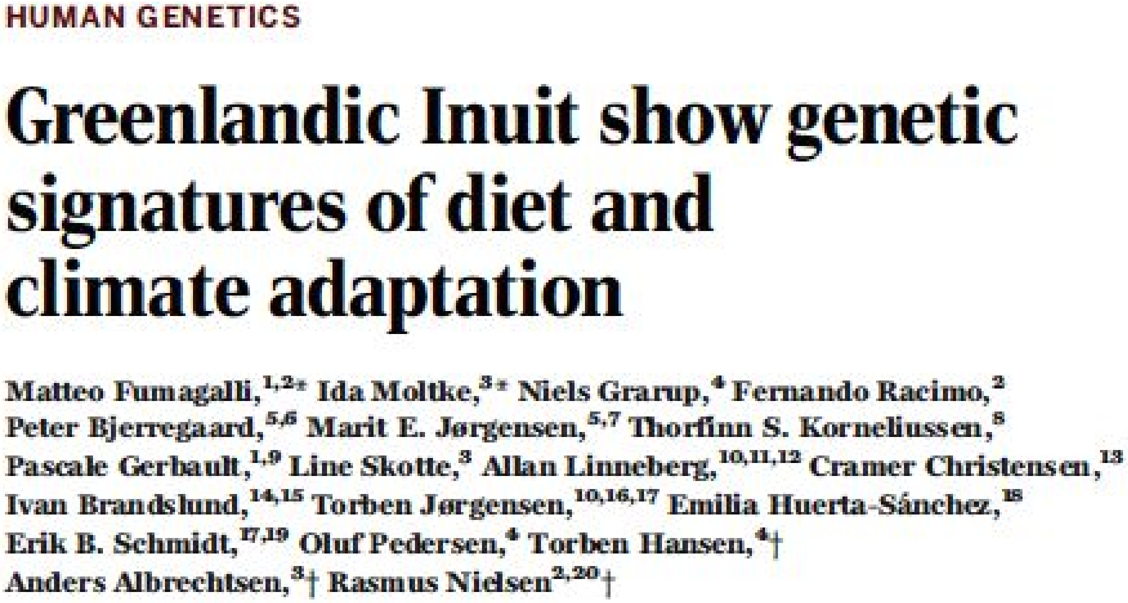
Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears

Shiping Liu,[1,2,20] Eline D. Lorenzen,[3,4,20] Matteo Fumagalli,[3,20] Bo Li,[1,20] Kelley Harris,[5] Zijun Xiong,[1] Long Zhou,[1] Thorfinn Sand Korneliussen,[4] Mehmet Somel,[3,21] Courtney Babbitt,[6,7,22] Greg Wray,[6,7] Jianwen Li,[1] Weiming He,[1,2] Zhuo Wang,[1] Wenjing Fu,[1] Xueyan Xiang,[1,8] Claire C. Morgan,[9] Aoife Doherty,[10] Mary J. O'Connell,[9] James O. McInerney,[10] Erik W. Born,[11] Love Dalén,[12] Rune Dietz,[13] Ludovic Orlando,[4] Christian Sonne,[13] Guojie Zhang,[1,14] Rasmus Nielsen,[1,3,15,16,*] Eske Willerslev,[4,*] and Jun Wang[1,16,17,18,19,*]



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], and The International HapMap Consortium

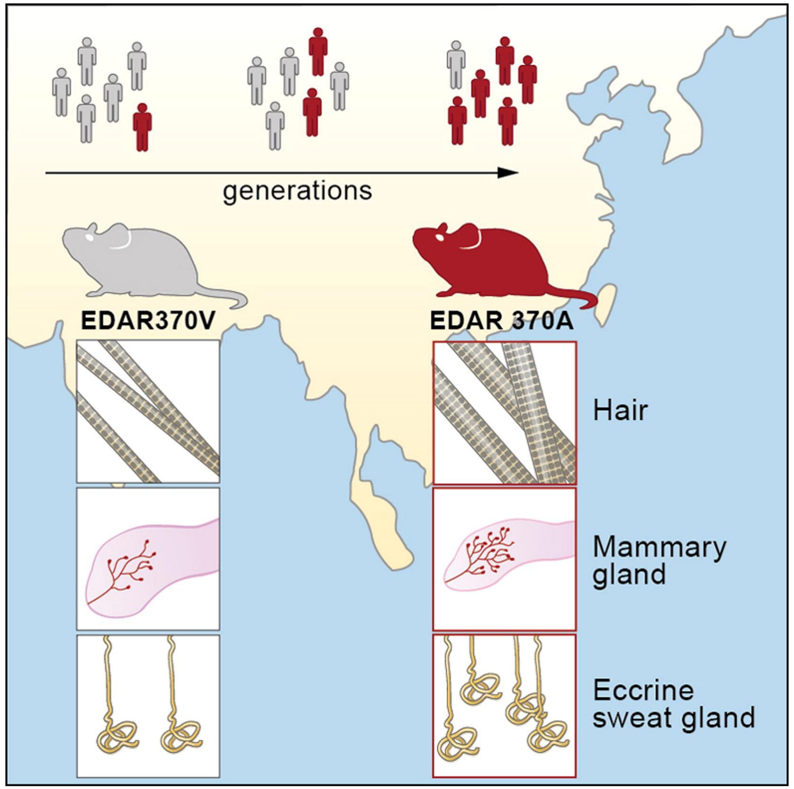


HUMAN GENETICS

Greenlandic Inuit show genetic signatures of diet and climate adaptation

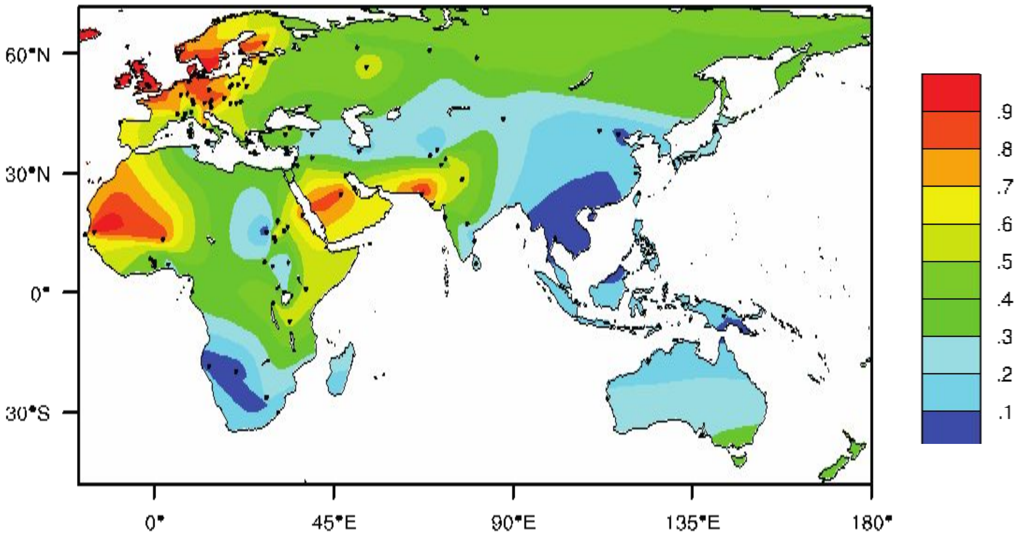
Matteo Fumagalli,[1,2*] Ida Moltke,[3*] Niels Grarup,[4] Fernando Racimo,[2] Peter Bjerregaard,[5,6] Marit E. Jørgensen,[5,7] Thorfinn S. Korneliussen,[8] Pascale Gerbault,[1,9] Line Skotte,[3] Allan Linneberg,[10,11,12] Cramer Christensen,[13] Ivan Brandslund,[14,15] Torben Jørgensen,[10,16,17] Emilia Huerta-Sánchez,[18] Erik B. Schmidt,[17,19] Oluf Pedersen,[4] Torben Hansen,[4†] Anders Albrechtsen,[3†] Rasmus Nielsen[2,20†]

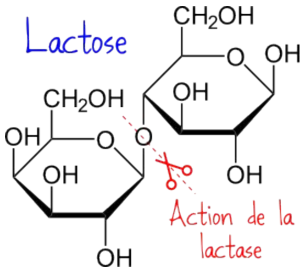
Specific examples for the practical today

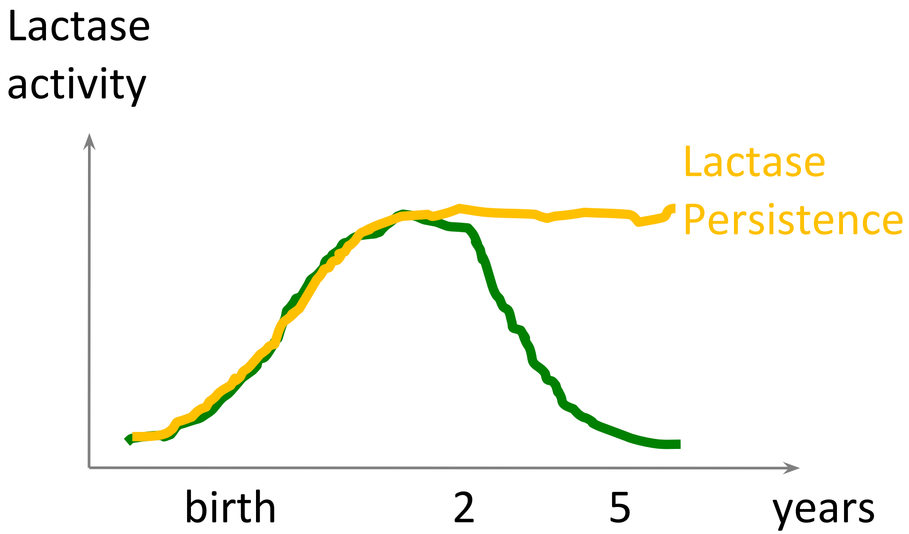


Kamberov et al. 2013 Cell

(a) Lactase persistence frequency







Gerbault et al. 2011 Phil Trans Roy Soc

Aims for the morning

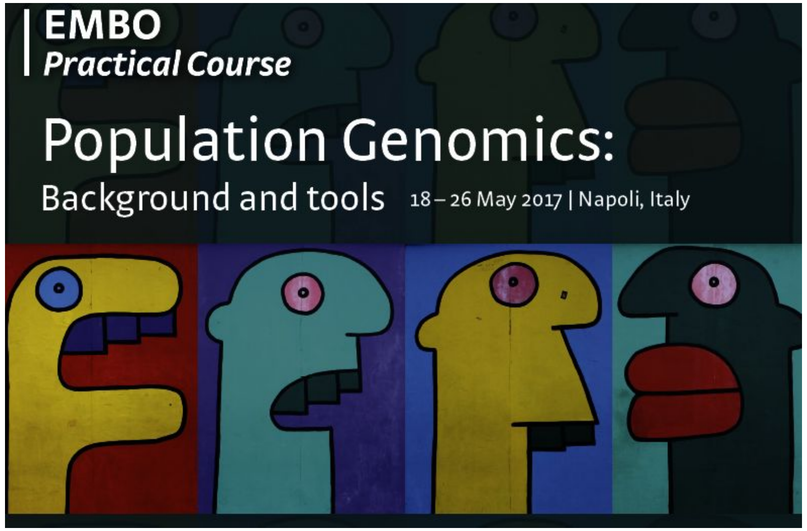
- THINK… together ;-)
- Series of commands used to generate the files from 1000 genomes data > what do they mean?
- Text files with statistics for EDAR~ SNP in Asians and LCT ~ SNP in Europeans > should be read/analysed in R: understand the commands and the statistic
- If you're fast: keep playing or have a break
- Any questions: PLEASE ASK :-)

With thanks to… YOU ☺

* Matteo Fumagalli, Imperial College London
* Garrett Hellenthal, University College London
* Andrea Manica, University of Cambridge

Saki Raheem, Linda Percy, Caroline Smith, University of Westminster

Enza and Chiara



EMBO
excellence in life sciences