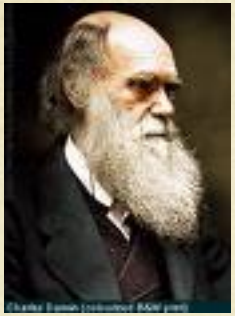
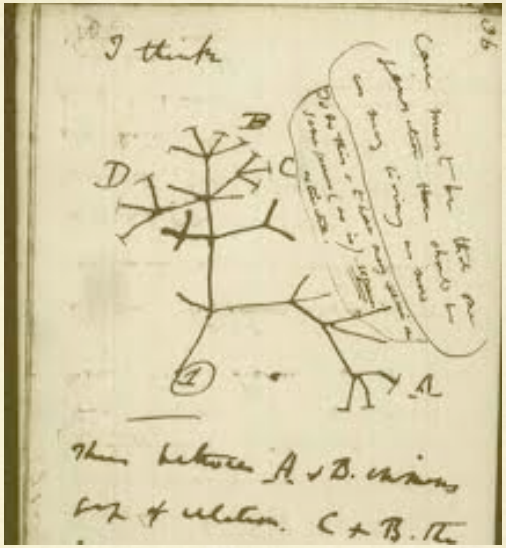
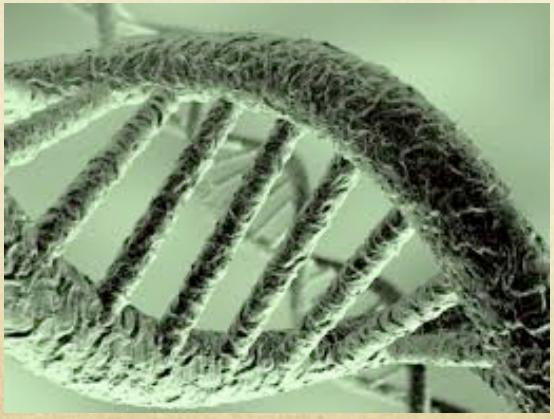
Detecting selection within genomes: an introduction

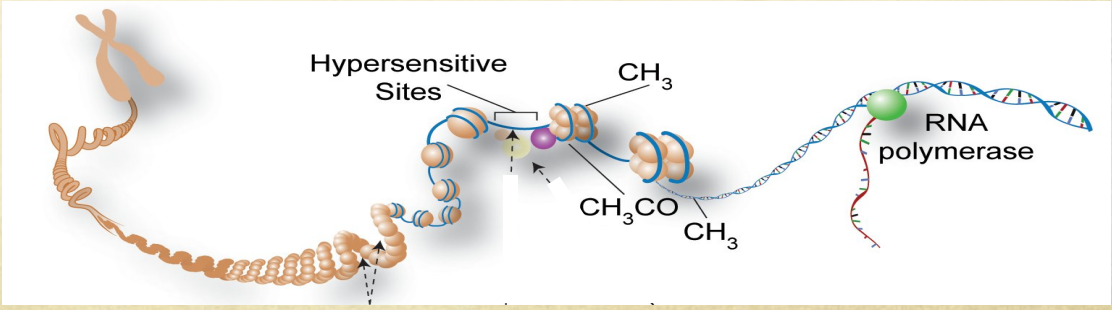
Laurent Duret
Laboratoire de Biométrie et Biologie Evolutive, CNRS, Université Lyon 1

Genome Function & Evolution









What's in our genome ?

- 3.1 10^9 bp
- Transposable elements (parasitic DNA): 45%

- About 20,000 protein-coding genes
- Protein-coding regions : 1.2%

- Non-coding functional elements: 5-10%

How to identify functional elements ?

How to identify functional elements ?

- **The ENCODE Project: ENCyclopedia Of DNA Elements**
- Large international consortium
- => systematic mapping of regions of transcription, transcription factor association, chromatin structure and histone modification (RNAseq, ChipSeq, ...)
- Result: **80% of the human genome associated to at least one « biochemical function »**

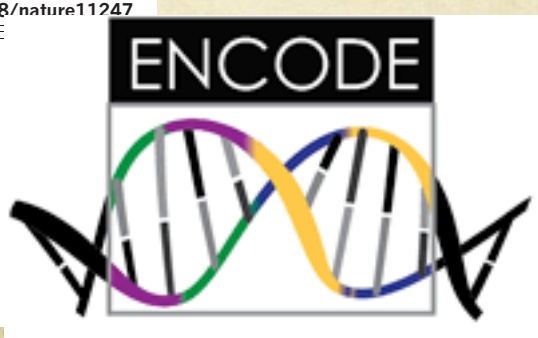
ARTICLE

6 SEPTEMBER 2012 | VOL 489 | NATURE | 57

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*



No more junk?

- « One of the more remarkable findings described in the ENCODE's paper is that 80% of the genome contains elements linked to biochemical functions, dispatching the widely held view that the human genome is mostly 'junk DNA'. » **J. Ecker (Nature,** News & Views)

No more junk?

HOME PAGE | TODAY'S PAPER | VID

The New York Times

WORLD | U.S. | N.Y. / REGION

Bits of Mystery DNA, Far From 'Junk,' Play Crucial Role

By GINA KOLATA

Published: September 5, 2012 | 574 Comments

theguardian

News | Sport | Comment | Culture | Business | Money | Life & style

News › Science › Genetics

Breakthrough study overturns theory of 'junk DNA' in genome

Biochemical activity = function ??

- Encode's definition of function is fuzzy
- 100% of the DNA has some « biochemical activity » (e.g. replication)
- DNA parasites (e.g. endogenous retroviruses) are associated to specific biochemical activities (e.g. transcription) => should they be considered as « functional elements »?

- **How to define a « functional element » ?**

Darwinian definition of function

- Functional genetic element = DNA segment that contributes positively to the fitness of the organism
- Function of a genetic element = phenotypic trait, determined by this element, that is under selective pressure

Non-functional genetic elements

- Neutral genetic element = DNA segment that has no impact on the fitness of the organism

- Intragenomic parasite = DNA segment that is able to replicate itself within a genome, at the expense of its host

Genome annotation by comparative genomics

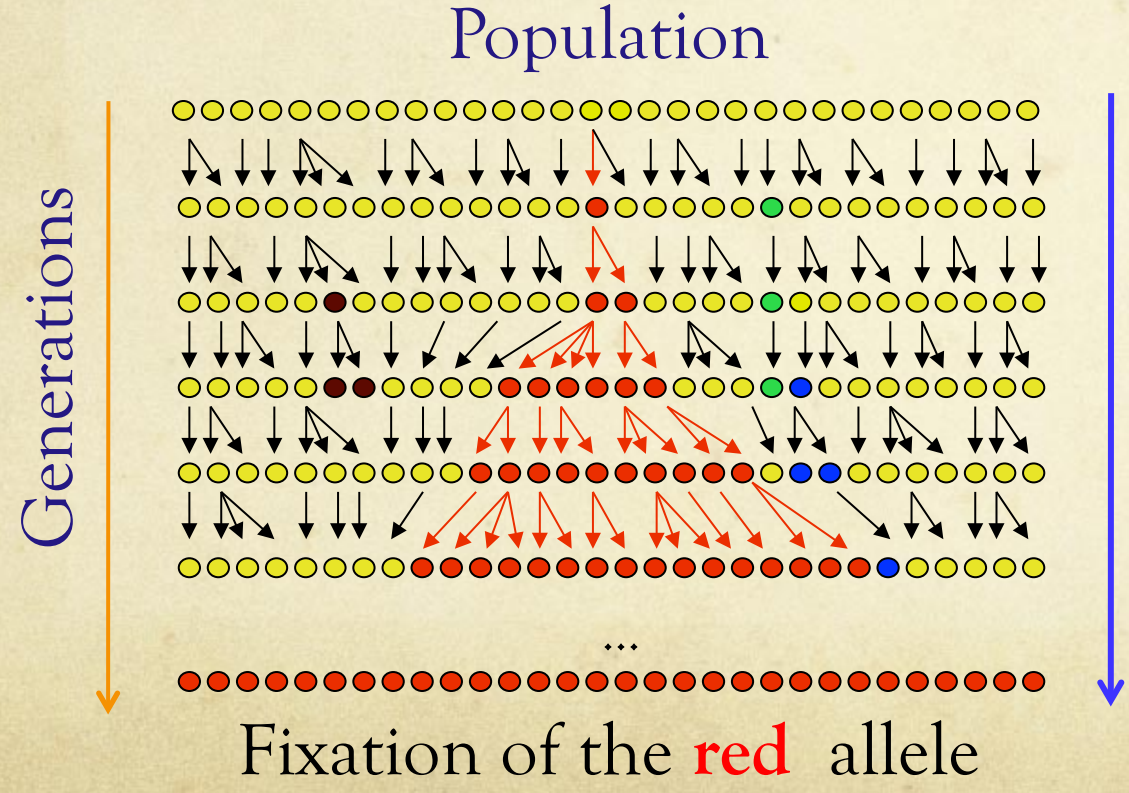
- Basic principle :
  - Functional element <=> constrained by natural selection
  - Detecting the hallmarks of selection in genomic sequences
    - Negative selection (conservation)
    - Positive selection (adaptation)

Tracking the signatures of positive selection within genomes:

the basics

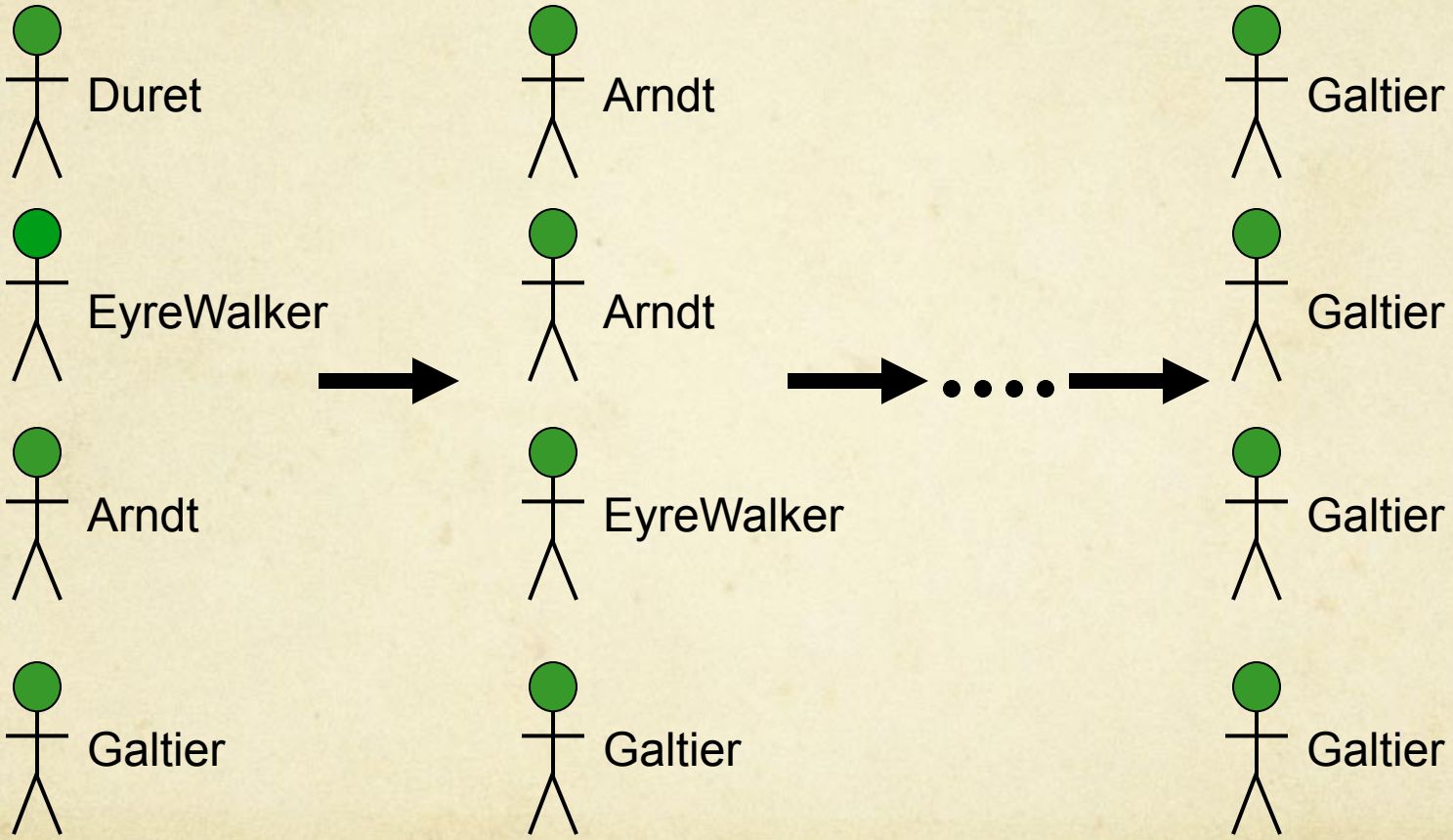
Evolution

- Mutation => new alleles
- Changes of allele frequencies over generations



- ✓ **Natural selection**
- ✓ **Genetic drift**

Last Names



Evolution : mutation, selection, drift

Probability of fixation:

$p = f(s, N_e)$

s : relative impact on fitness

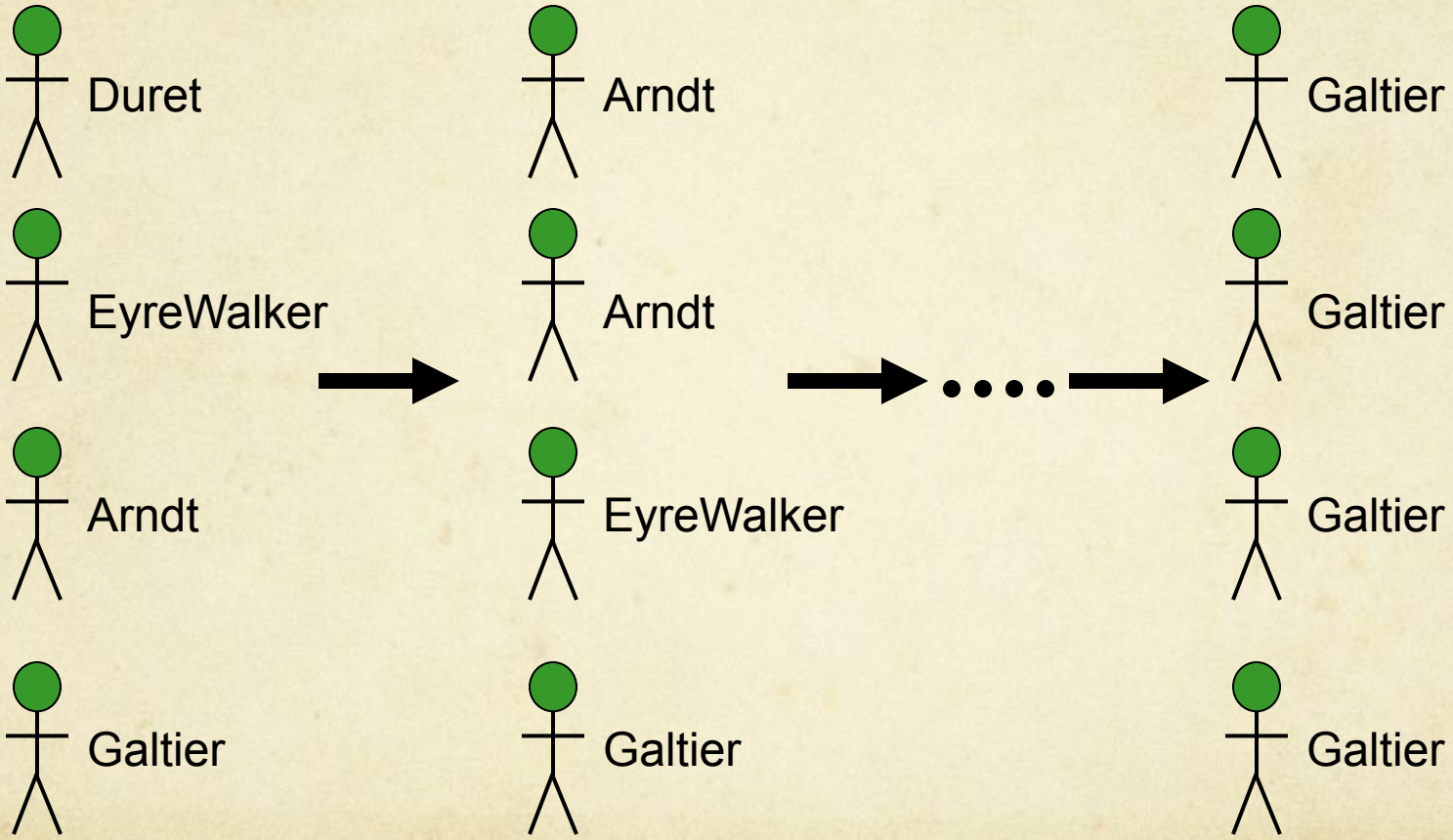
- $s = 0$: neutral mutation (random genetic drift)
- $s < 0$: disadvantageous mutation = negative (purifying) selection
- $s > 0$: advantageous mutation = positive (directional) selection

N_e : effective population size: stochastic effects are stronger in small populations

$|N_e s| < 1$: effectively neutral mutation

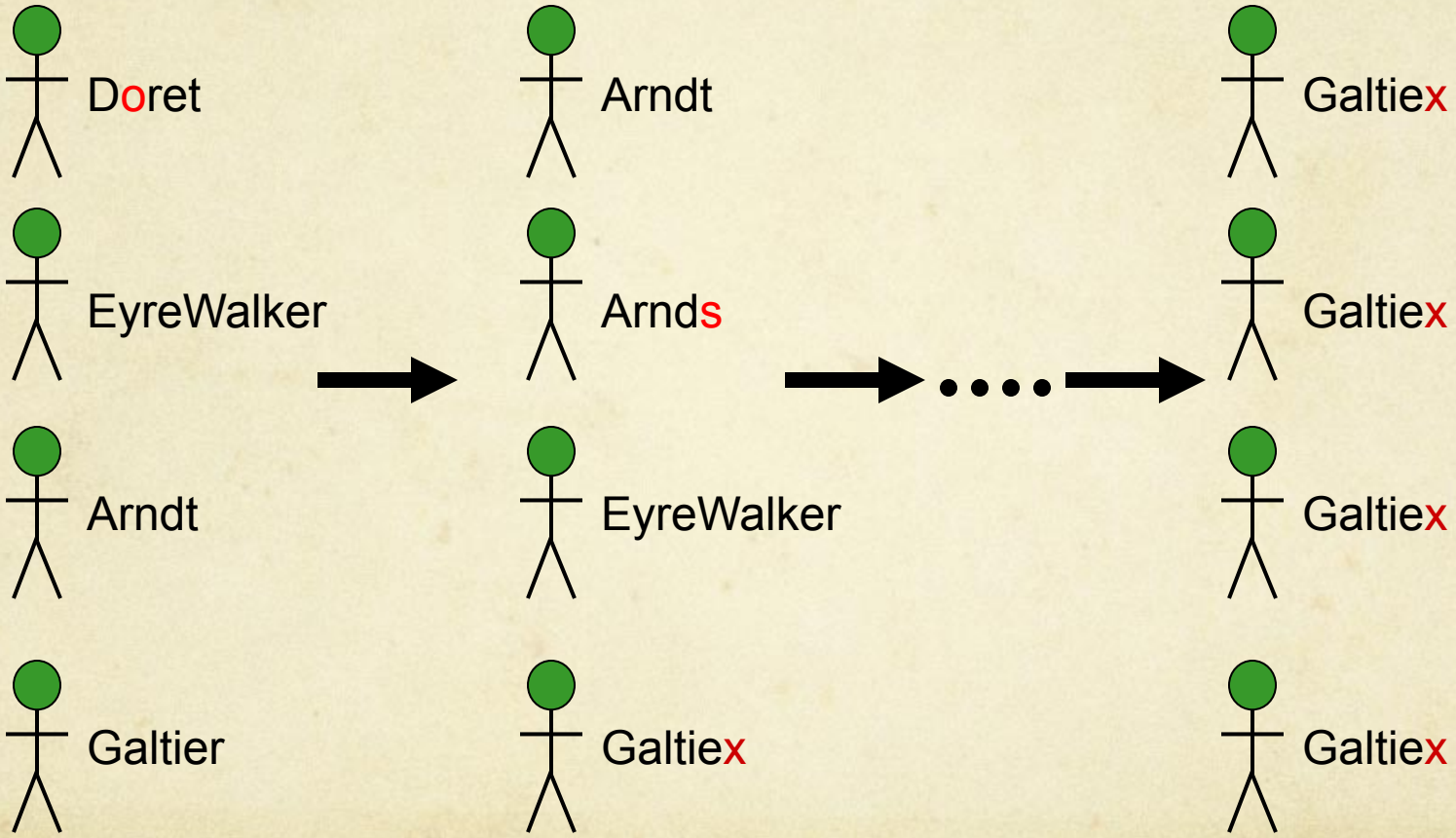
The rate of evolution of neutral sequences

Probability of Fixation



Probability of fixation = 1/N = 1/4

Mutation Rate



Number of Mutations in the Pop = uN = 1/5 x 4 = 0.8

Neutral Rate

Population size = N = 4

Rate of mutation (per generation) = u = 1/5

Number of mutations in the population (per generation) = uN = 4/5

Probability of fixation = 1/N = 1/4

Rate of substitution = uN x 1/N = u = 1/5

Neutral Rate

Neutral substitution rate = mutation rate

The neutral substitution rate does not depend on population size

Non-neutral Rate

Number of mutations per generation (diploid) : 2uN

Probability of fixation (Kimura, 1962):

$$P(s) = \frac{1 - e^{-2s}}{1 - e^{-4Ns}}$$

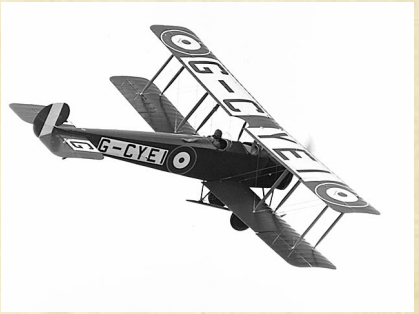
Rate of substitution = 2uN x P(s)

Tracking natural selection ...

- **Demonstrate the action of selection = reject the predictions of the neutral model**
- Compare substitution rate (K) to mutation rate (u) :
  - Neutral evolution => $K = u$
  - Negative selection => $K < u$
  - Positive selection => $K > u$

Searching for functional sequences under negative (purifying) selective pressure:

Phylogenetic Footprints



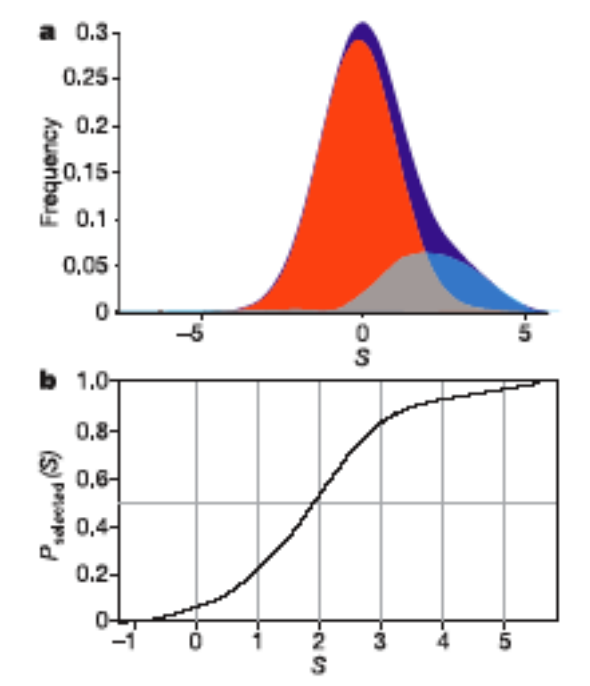
• Comparative genomics: when comparing sequences from different species, the mutations that are *not* observed are the ones that are deleterious (the others are neutral or beneficial)

Comparison of human and mouse genomes (MGSC 2002)

- Alignment of human and mouse genomes : 40% of the human genome can be aligned with the mouse genome

- How much of the human genome is under negative selective pressure ??

Comparison of human and mouse genomes



Distribution of substitution rates

- Ancient Repeats (neutral marker)
- Non-repeated sequences

Probability to be under negative selective pressure

MGSC (Nature, 2002)

- More than 5% of the genome of mammals is under negative selection
- NB: only 1.0% du genome is coding !! 4 times more functional non-coding regions than coding regions !!

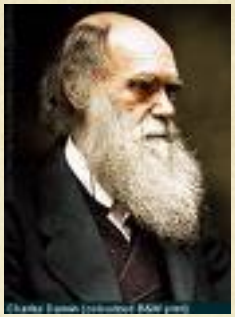
Phylogenetic footprints = genetic conservatism

- Phylogenetic footprints = functional elements conserved during evolution

- What about sequence elements that have been involved in functional innovation ?

- What are the functional elements responsible for adaptative evolution ?

What make chimps different from us ?



30×10^6 point substitutions + indels + duplications (copy number variations)

- Searching for functional elements subject to positive (directional) selection: *substitution rate* > *u*

Searching for positive selection in protein-coding genes

1. Align DNA sequences

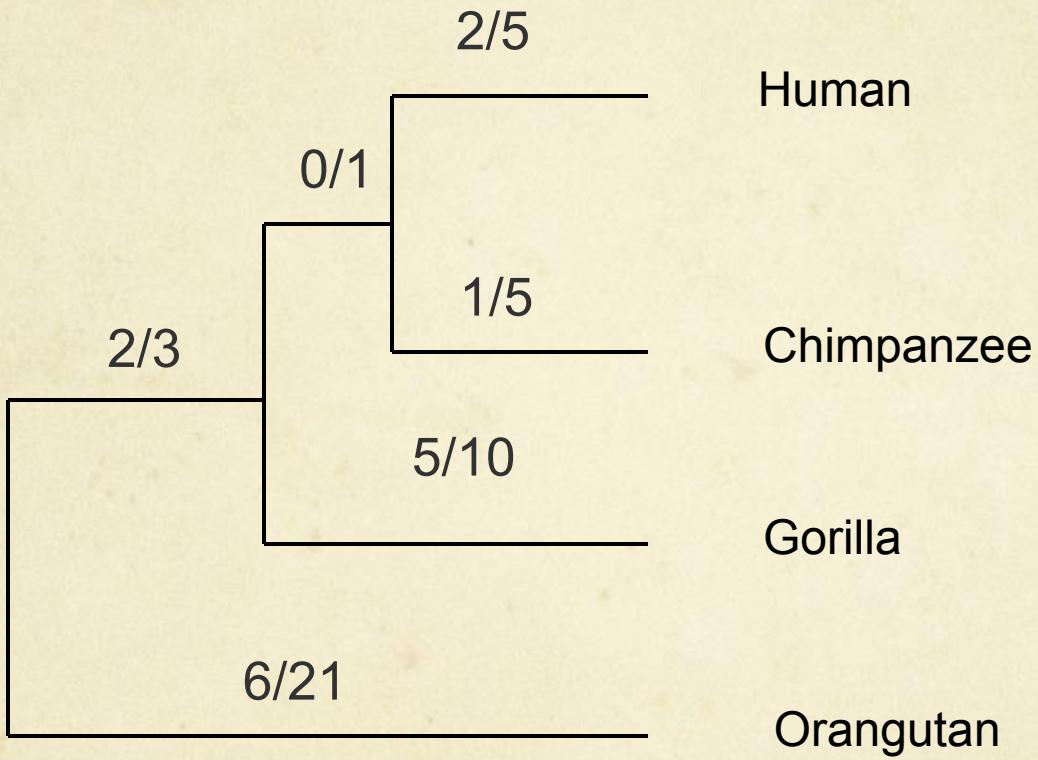
	Arg	Lys	Pro	_	Ile	Gln	Asn	Gly	Gln
Human	CGC	AAA	CCC	---	ATT	CAG	AAT	GGC	CAG
Mouse	CGC	AAG	GCC	CCG	ATG	CAG	AAT	GGT	CAG
	Arg	Lys	Ala	Pro	Met	Gln	Asn	Gly	Gln

2. Count changes

synonymous = 2
non-synonymous = 2

=> synonymous substitution rate (d_S)
=> non-synonymous substitution rate (d_N)

Multiple Sequences



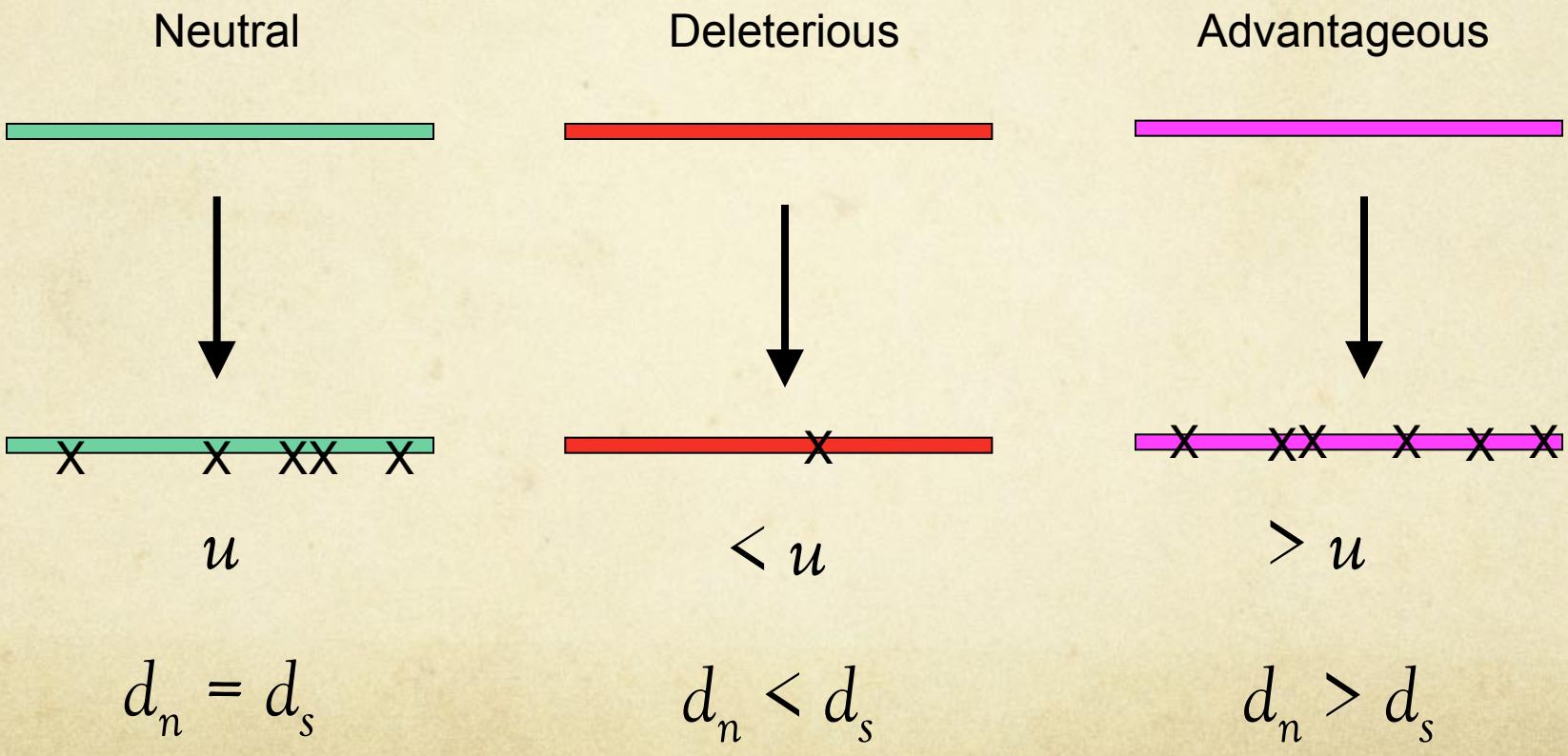
Assumptions

- Synonymous Mutations Are Neutral:
  - $d_S = u$

- Non-synonymous Mutations Are Neutral, Deleterious or Advantageous

Searching for positive selection in protein-coding genes

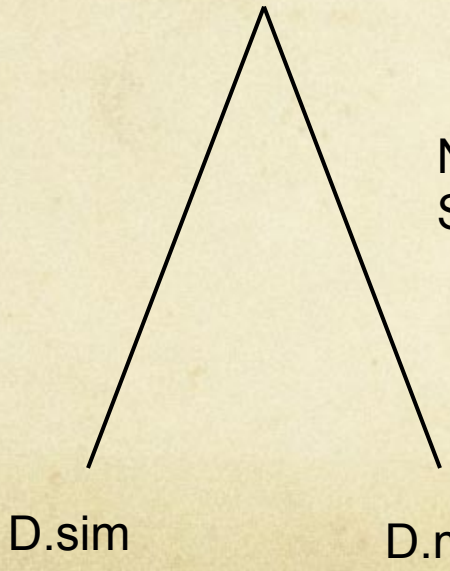
Rates of evolution :



Example 1 : *Odysseus*

Odysseus

- involved in hybrid sterility between *D.simulans* and *D.mauritiana*
- homeodomain protein



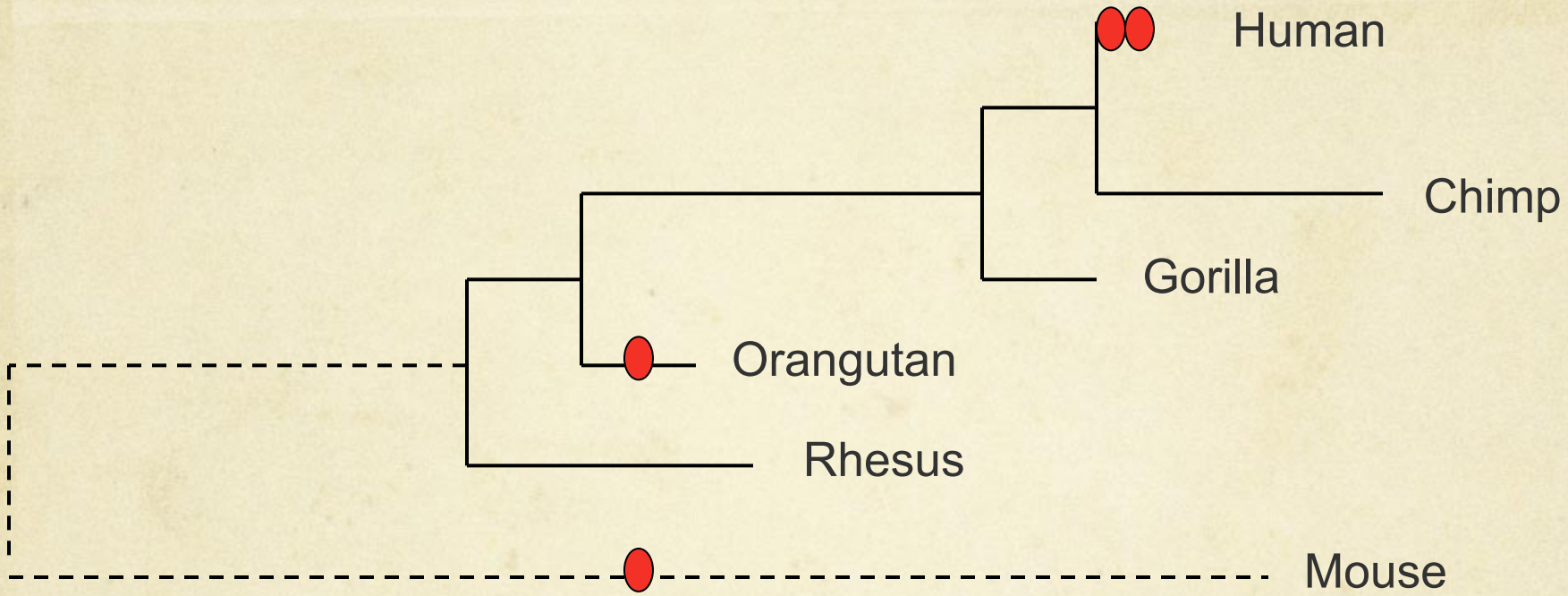
d_n	d_s	d_n/d_s
0.067	0.033	2.0

Example 2 : FOXP2

FOXP2
- transcription factor gene
- two mutations in humans lead to speech impairment

Enard et al. Nature 2002
- Sequenced FOXP2 in other primates

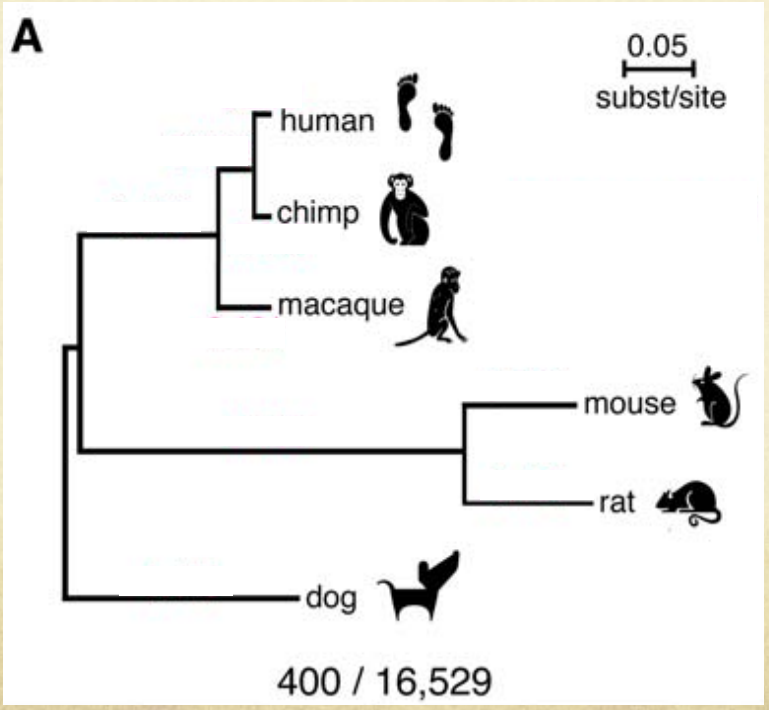
FOXP2 continued



- d_n/d_s significantly higher in humans
- $d_n > d_s$ in humans

Genome-wide scans for positively selected genes

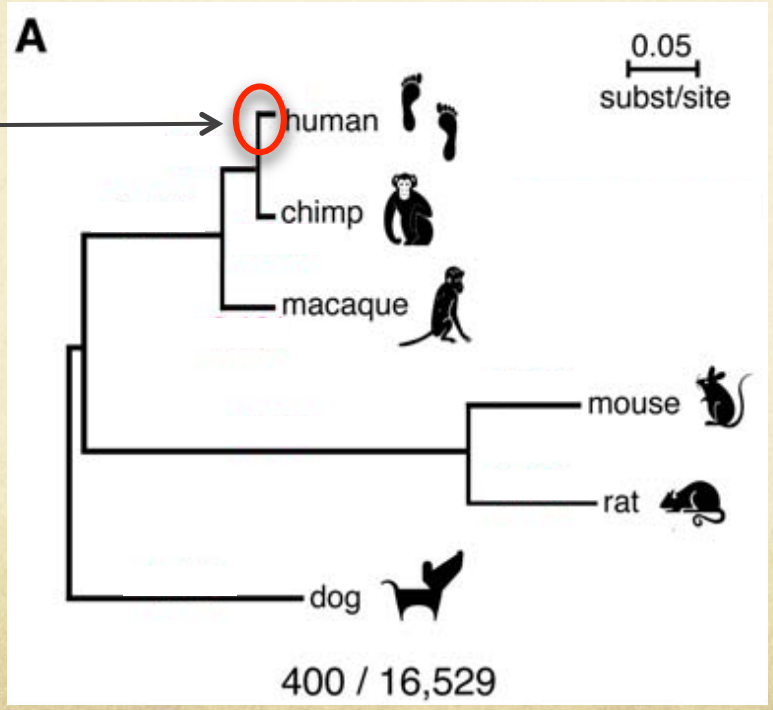
- Kosiol et al. (2008) PLoS Genet
- Complete genome from six mammalian species
- 16,529 human genes with orthologs in at least 2 other species



Genome-wide scans for positively selected genes

- dN/dS test: 500/16,529 genes with evidence of positive selection
- Genes under positive selection: enriched for roles in defense/ immunity, odor/taste perception, and reproduction

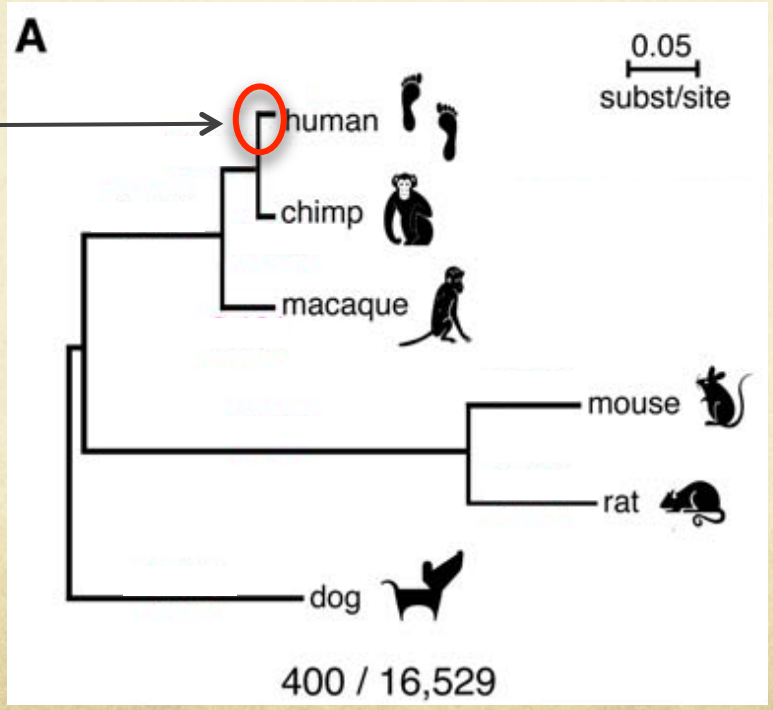
Only 10 genes with signal of positive selection specifically in the human lineage



Genome-wide scans for positively selected genes

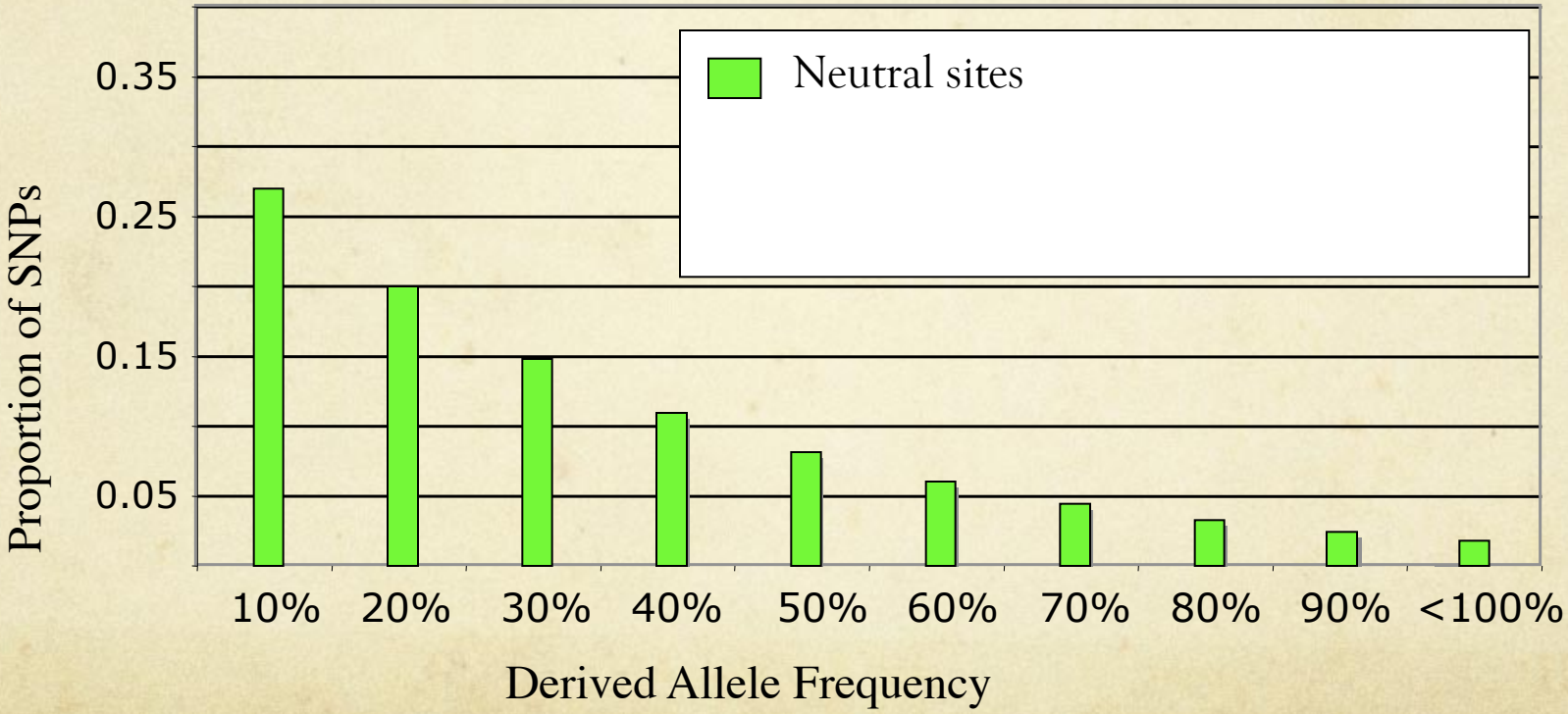
- Short phylogenetic branch
- Few substitutions (~1% divergence at synonymous sites)
- => Limited power to detect selection

Only 10 genes with signal of positive selection specifically in the human lineage



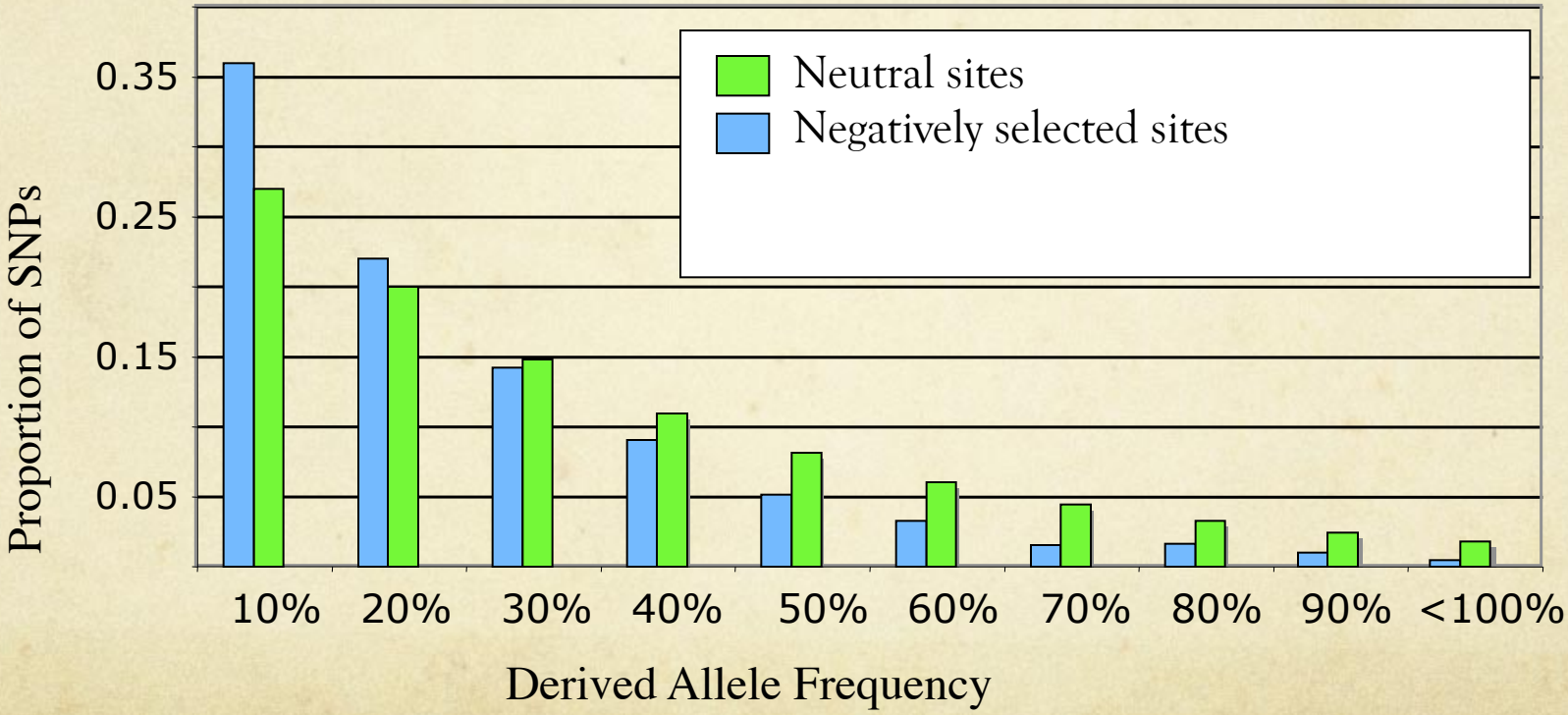
Tracking natural selection ... by analysis of polymorphism data

- *Derived allele frequency spectrum*



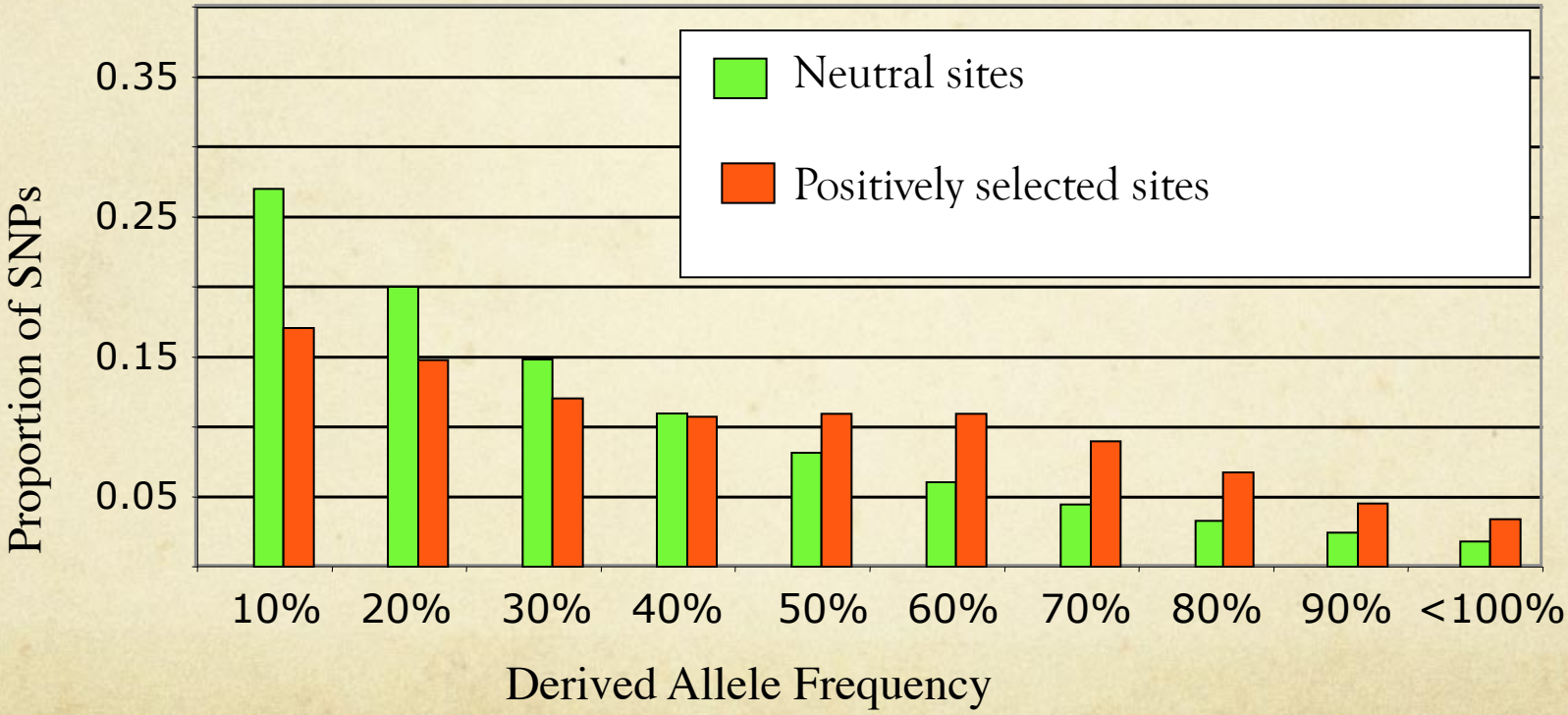
Tracking natural selection … by analysis of polymorphism data

- *Derived allele frequency spectrum*



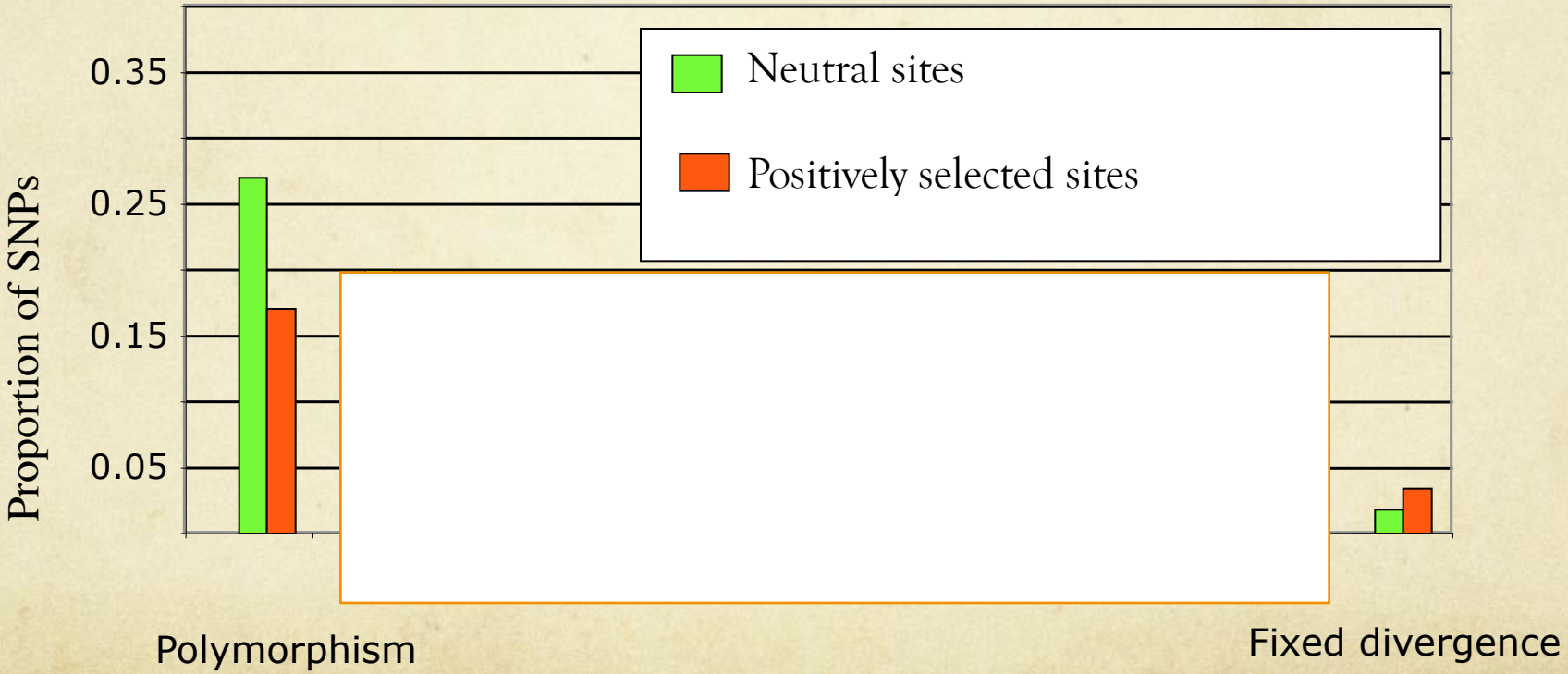
Tracking natural selection ... by analysis of polymorphism data

- *Derived allele frequency spectrum*

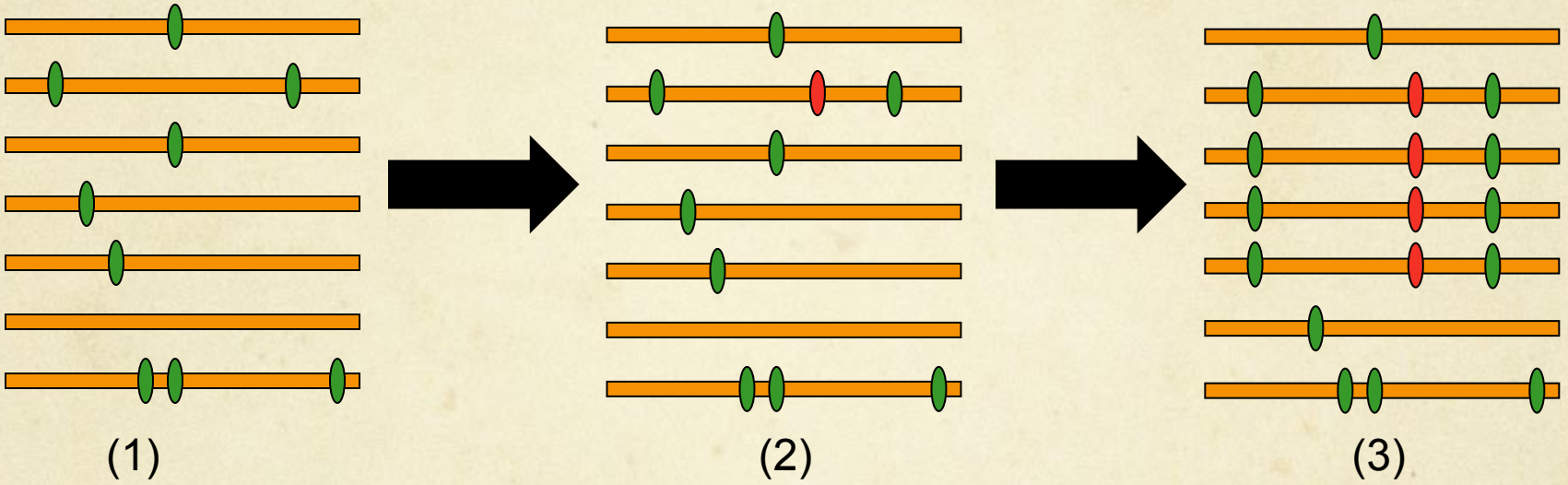


Tracking natural selection ... by comparison of divergence and polymorphism

- *Mc Donald-Kreitman test, HKA test*



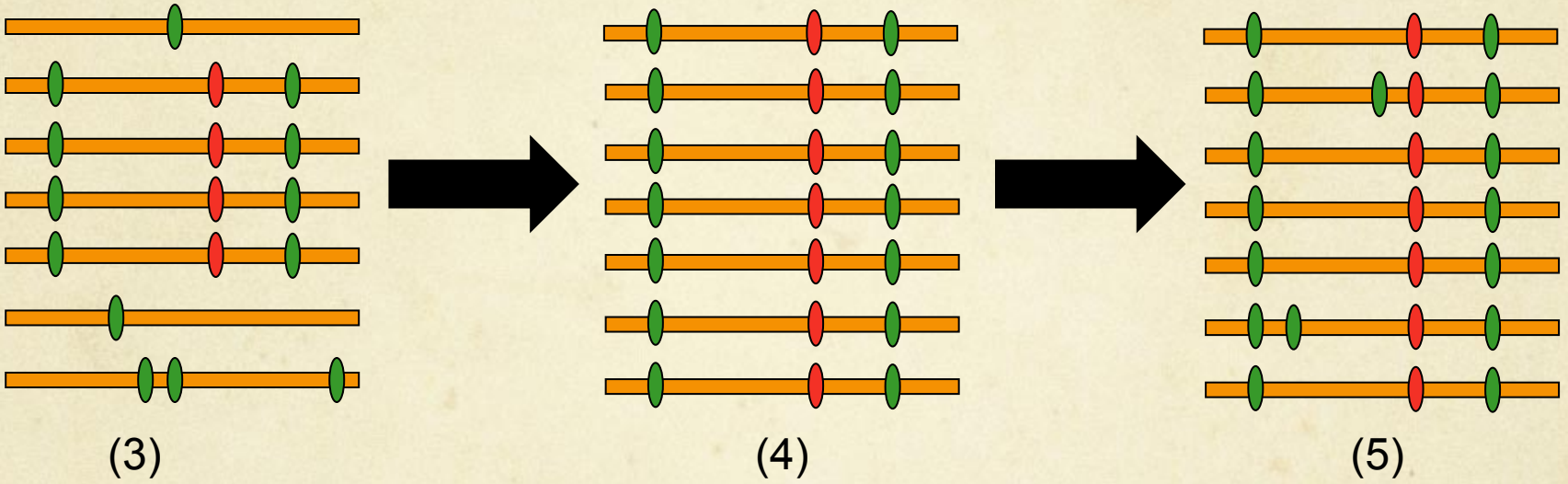
Pattern of polymorphism at linked sites



Effects :

- long blocks of strong linkage disequilibrium (long haplotypes)

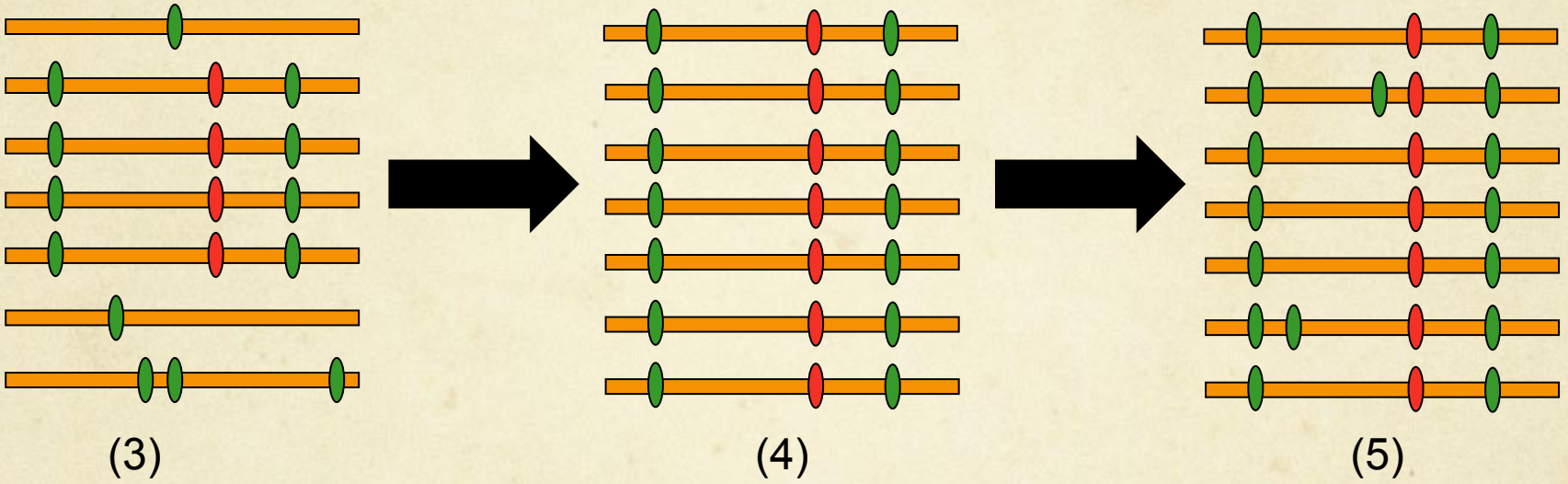
Pattern of polymorphism at linked sites



Effects :

- long blocks of strong linkage disequilibrium (long haplotypes)
- reduced level of polymorphism in the neighborhood of the selected allele
- skew towards rare allele

Pattern of polymorphism at linked sites



Selection tests:

- linkage disequilibrium: iHS, IBD, LRH, ...
- allele frequency: Tajima's D, Fay & Wu's H, ...

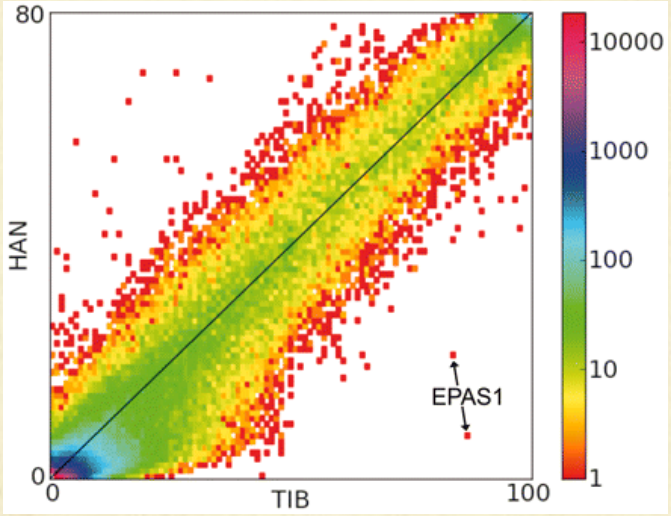
Composite methods

Population differentiation

- Different populations are subject to different environmental pressure => local adaptation
- If a locus is subject to selection in one population but not in another => differences in allele frequency among populations
- => searching for alleles with unexpectedly strong differences in allele frequency among populations
- E.g.: adaptation to high altitude

Population differentiation

- Sequencing exomes of 50 Tibetans (4300 m in altitude)
- Comparison of allele frequencies with 40 Han individuals (Beijing)



EPAS1, a transcription factor involved in response to hypoxia

Yi *et al.*, *Science (2010) 329:75-78*

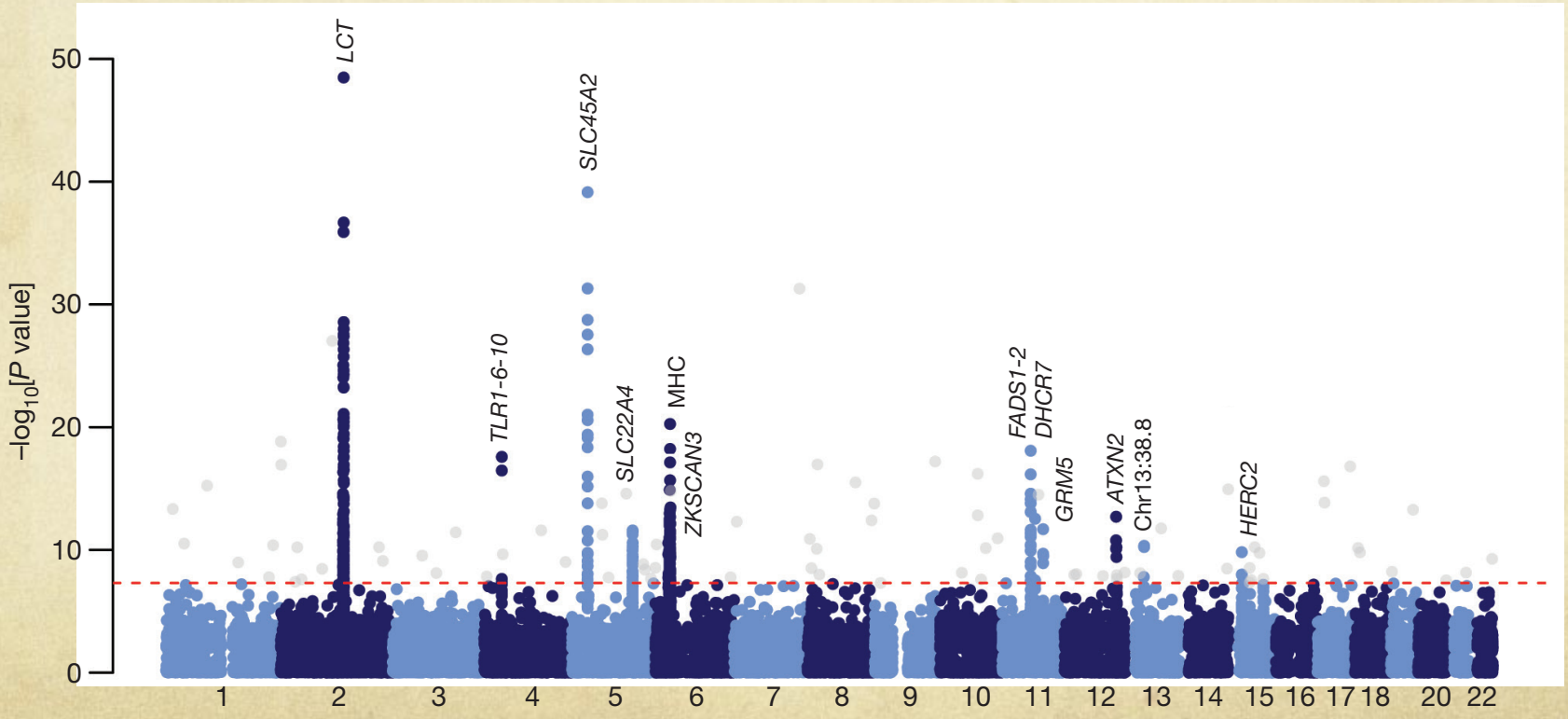
Huerta-Sánchez et al. (2014) *Nature.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA.

Temporal variation in allele frequency

- An allele subject to positive selection tends to increase in frequency across generations, at a higher rate than alleles subject to random genetic drift
- Comparison of allele frequencies within populations, sampled at different time points
- Ancient DNA
- E.g.: Mathieson et al. (2015) *Nature* 528: 499–503
  - Genome-wide SNP data from 230 ancient Eurasians (from 6500 to 300 bc), including 26 Anatolian Neolithic farmers
  - Comparison of allele frequency in modern and ancient populations

Temporal variation in allele frequency

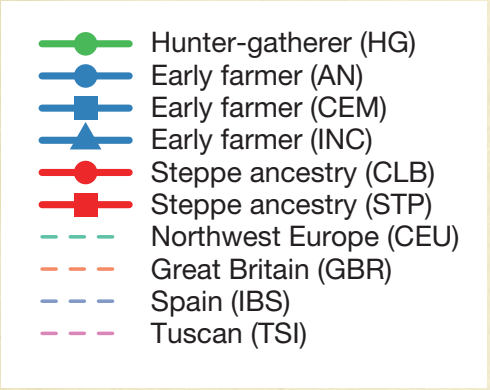
- The strongest signal of selection is at the SNP (rs4988235) responsible for lactase persistence in Europe

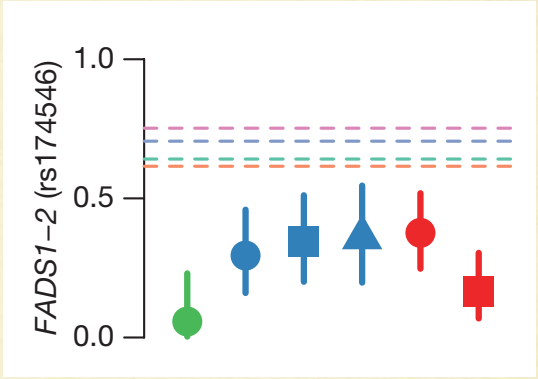


Mathieson et al. (2015) *Nature* 528: 499–503

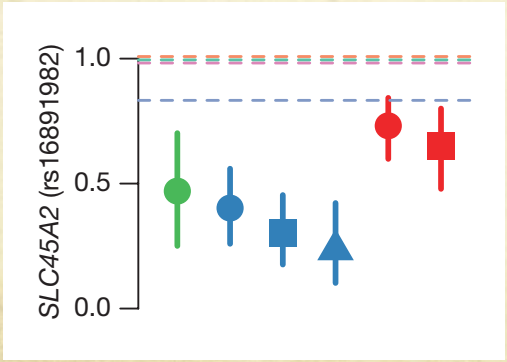
Temporal variation in allele frequency

- Comparison of allele frequency in modern and ancient European populations

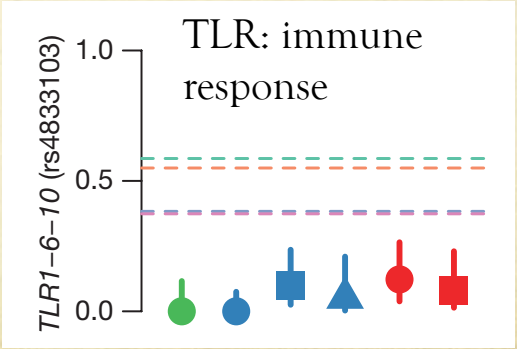




FADS1, FADS2: fatty acid metabolism



SLC45A2: light skin pigmentation



TLR: immune response

Mathieson et al. (2015) *Nature* 528: 499–503

Summary: tracking natural selection ...

- Macroevolution: Rate-based methods
  - Interspecies (e.g. dN/dS)
  - Polymorphism/divergence comparison (e.g. MK-test, HKA)
- Microevolution: analysis of polymorphism
  - Allele frequency spectra (e.g. Tajima's D)
  - Linkage disequilibrium (e.g. iHS, IBD)
  - Population differentiation
  - Temporal variation in allele frequency

For a review, see: Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. *Annu Rev Genet*. 2013;47: 97–120. doi:10.1146/annurev-genet-111212-133526

Tomoko
Ohta

Motoo
Kimura

The neutralist /selectionist controversy

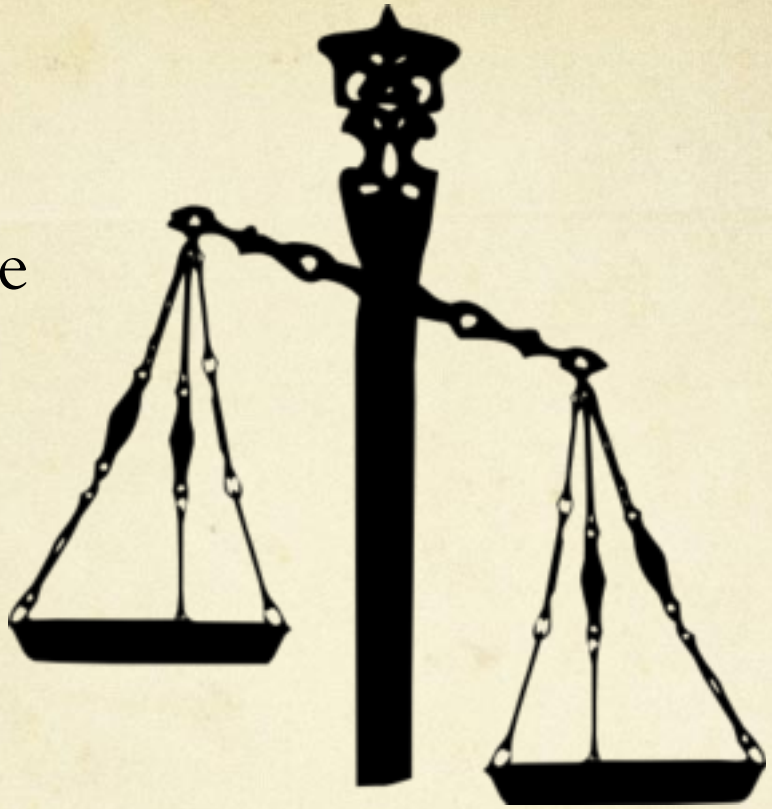
To what extent is the organization and content of genomes driven by selection or by non-adaptive evolutionary processes ?

Non-adaptive processes

Selection

Contingency
Random genetic drift
Intragenomic parasites
Genetic conflicts
...

Non-adaptive processes



Selection

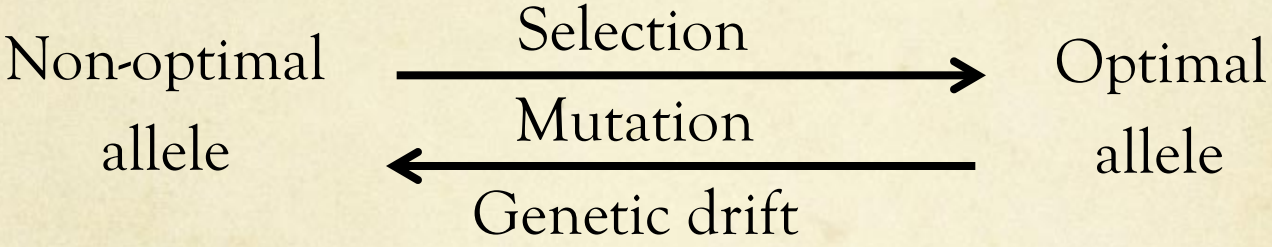
Pan-adaptionism : the best of all possible genomes (Pangloss)

A pragmatic view...

- If we want to demonstrate that selection is acting, we have to reject the alternative hypothesis
  - Neutral evolution = null hypothesis
- To be able to detect selection, **it is essential to identify all non-adaptive evolutionary processes that contribute to genome evolution**

The (nearly) neutral theory

- The efficacy of selection has some limits:



$$P(s) = \frac{1 - e^{-2s}}{1 - e^{-4Ns}}$$

- A genome is necessarily sub-optimal…

Don't forget selection levels !

- Selection at the individual level
- Selection at the species level
  - ± robustness to extinction
- Selection at the intragenomic level
  - Selfish genetic elements: elements that are able to replicate, without contributing positively to the fitness of their host

Conflicts between different levels of selection