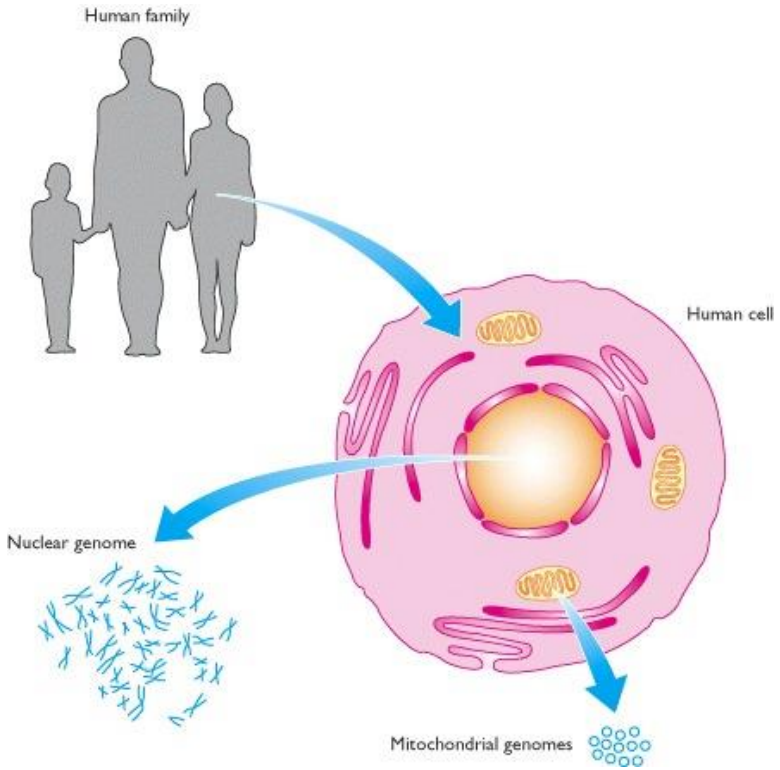
Evolutionary Genomics

University of Bern, 2020

Vitor Sousa
vmsousa@fc.ul.pt

Genome



In modern molecular biology and genetics, a **genome is the genetic material of an organism**. It consists of DNA (or RNA in RNA viruses). The genome includes both the genes (the coding regions) and the noncoding DNA, as well as the genetic material of the mitochondria and chloroplasts.

https://en.wikipedia.org/wiki/Genome

https://www.ncbi.nlm.nih.gov/books/NBK21134/

Genome composition

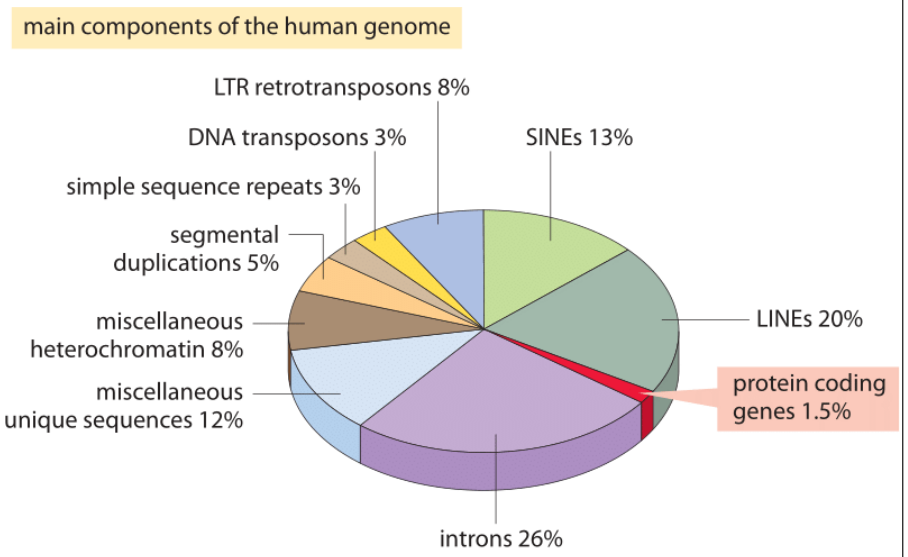
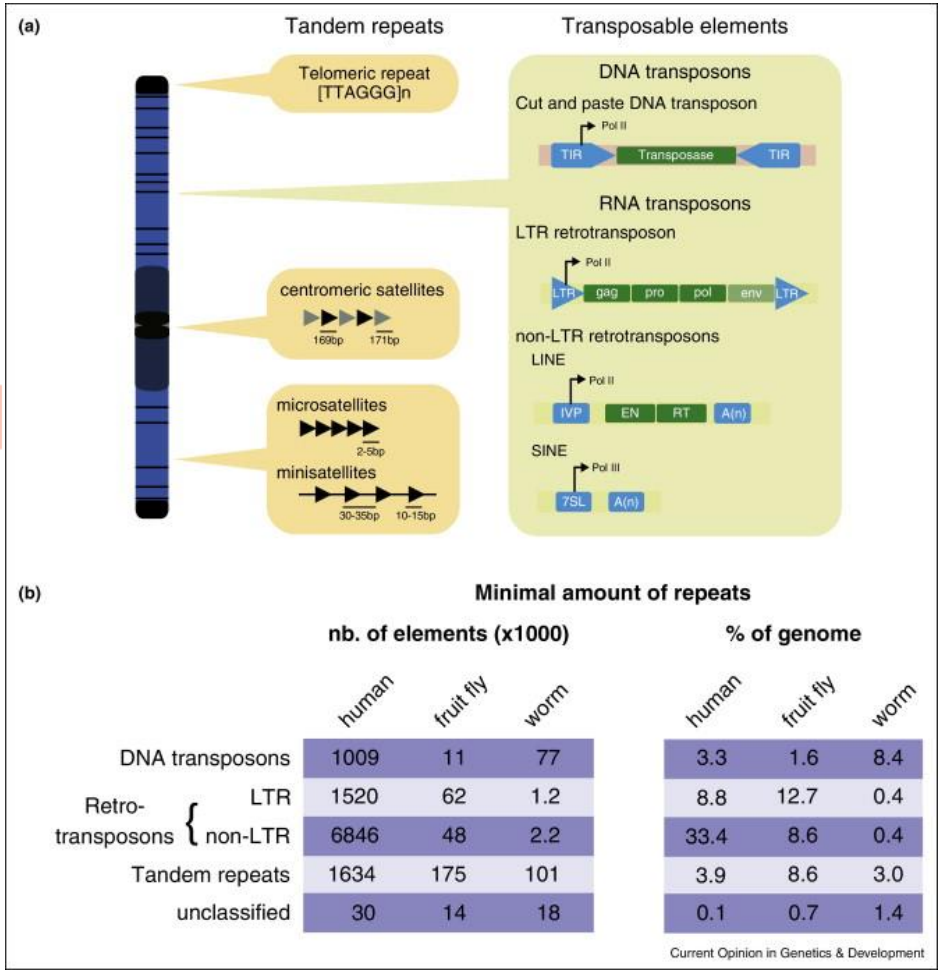


Figure 2: The different sequence components making up the human genome. About 1.5% of the genome consists of the ≈20,000 protein-coding sequences which are interspersed by the non coding introns, making up about 26%. **Transposable elements** are the largest fraction (40-50%) including for example **long interspersed nuclear elements** (LINEs), and **short interspersed nuclear elements** (SINEs). Most transposable elements are genomic remnants, which are currently defunct. (BNID 110283, Adapted from T. R. Gregory Nat Rev Genet. 9:699-708, 2005 based on International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409:860 2001.)

http://book.bionumbers.org/how-many-genes-are-in-a-genome/



(b) **Minimal amount of repeats**

	nb. of elements (x1000)			% of genome		
	human	fruit fly	worm	human	fruit fly	worm
DNA transposons	1009	11	77	3.3	1.6	8.4
Retro-transposons LTR	1520	62	1.2	8.8	12.7	0.4
Retro-transposons non-LTR	6846	48	2.2	33.4	8.6	0.4
Tandem repeats	1634	175	101	3.9	8.6	3.0
unclassified	30	14	18	0.1	0.7	1.4

Current Opinion in Genetics & Development

https://doi.org/10.1016/j.gde.2015.03.009
Padeken et al (2015)

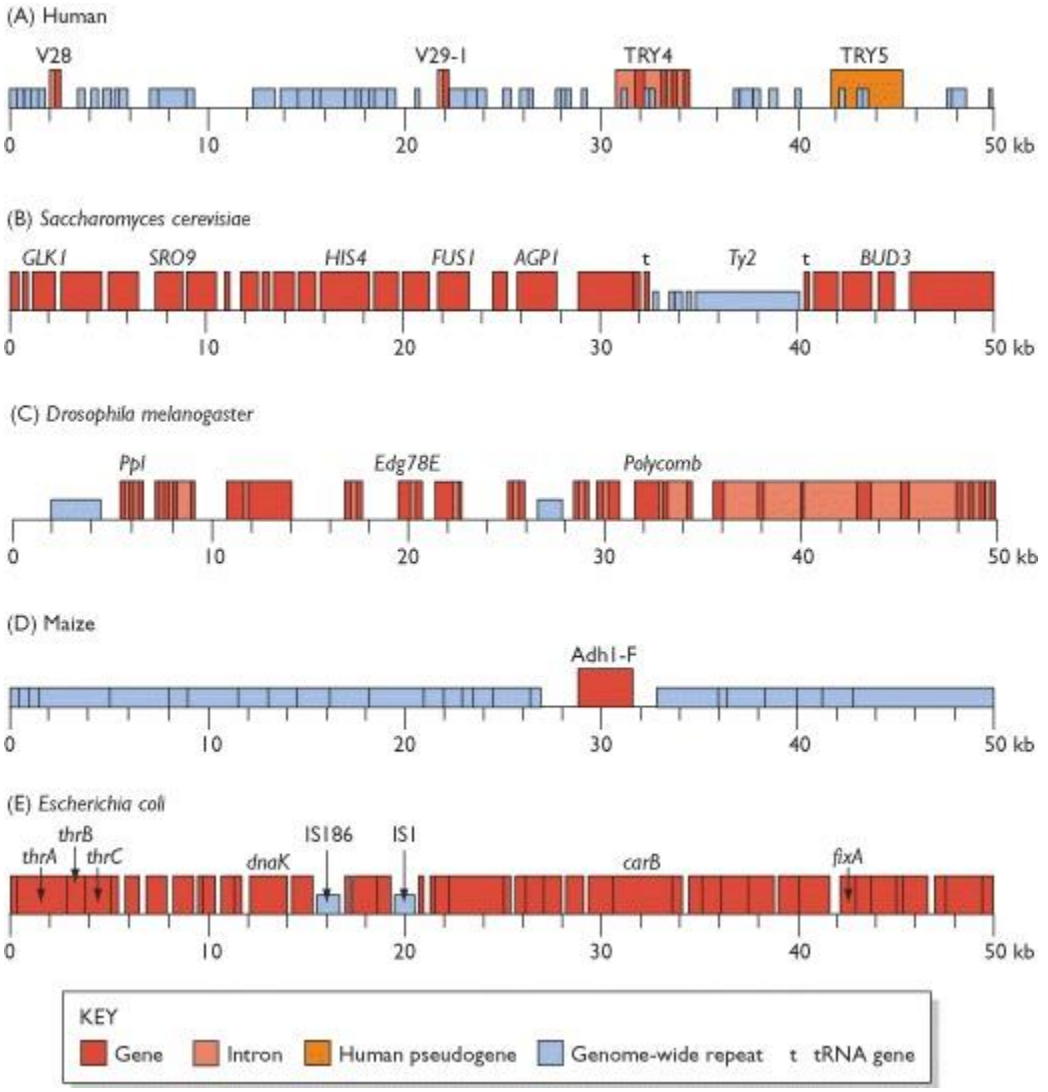
Only a small proportion of the genome codes for proteins

Genome structure varies among species

Repetitive elements may comprise over two-thirds of the human genome.

de Koning et al (2011) PLoS Genet
https://doi.org/10.1371/journal.pgen.1002384

But other species have more compact genomes, with less repetitive elements (e.g. E. coli), while other species can have even a higher proportion of repetitive elements (e.g. Maize).

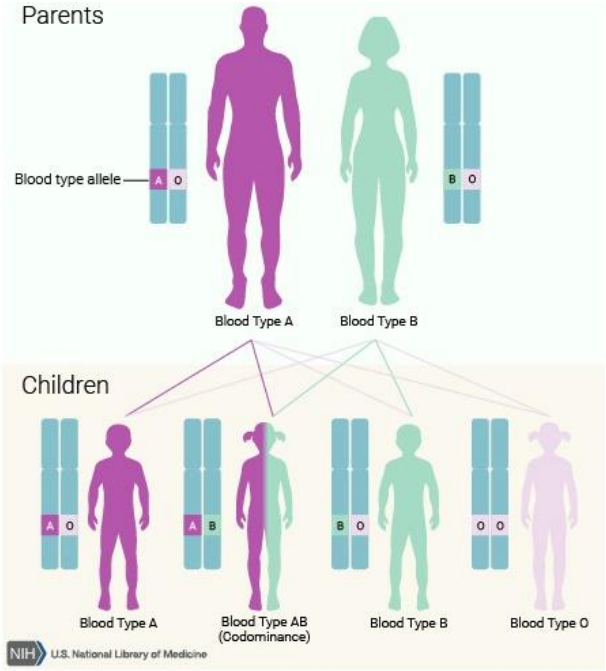


(A) is the 50-kb segment of the human β T-cell receptor locus shown in Figure 1.14. This is compared with 50-kb segments from the genomes of (B) Saccharomyces cerevisiae (chromosome III; redrawn from Oliver et al., 1992); (C) Drosophila melanogaster (redrawn from Adams et al., 2000); (D) maize (redrawn from SanMiguel et al., 1996) and (E) E. coli K12 (redrawn from Blattner et al., 1997).

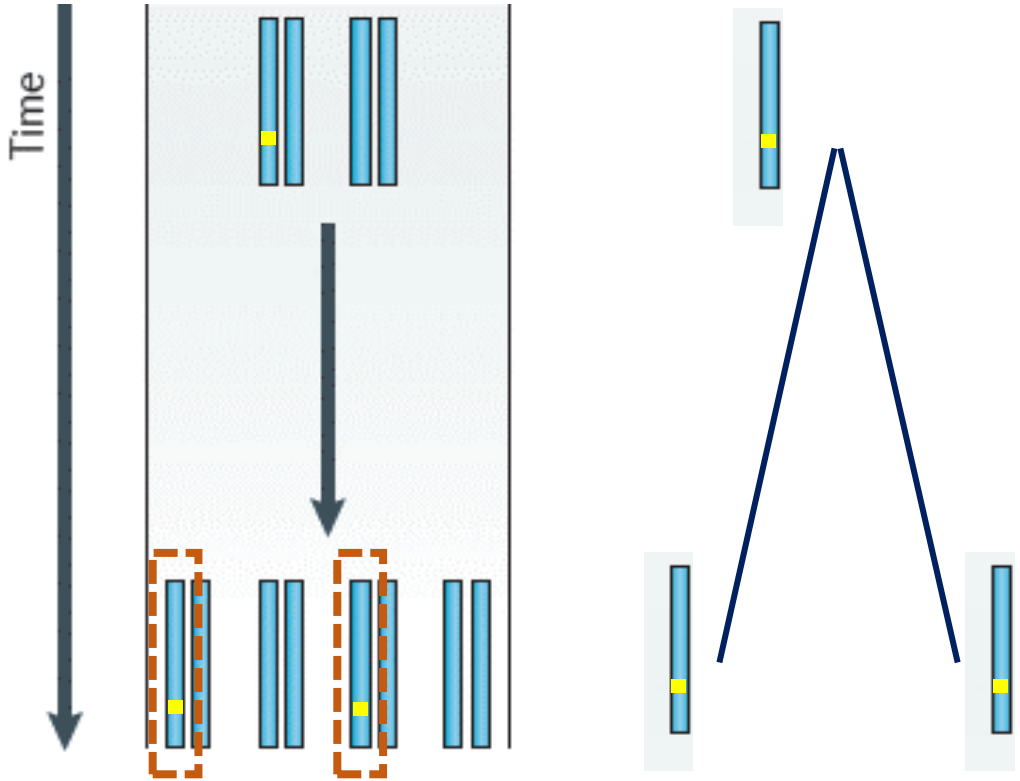
http://book.bionumbers.org/how-many-genes-are-in-a-genome/

Genomes as a history book

The genome is transmitted each generation

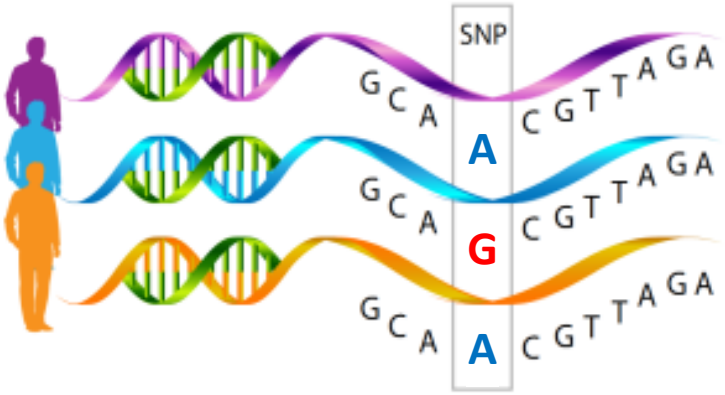


We can use mutations to reconstruct genealogies



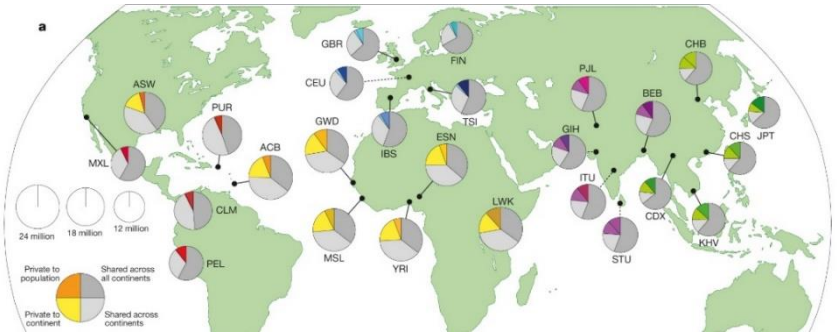
Racimo et al. (2015) *Nature Reviews Genetics*

Population genetic data

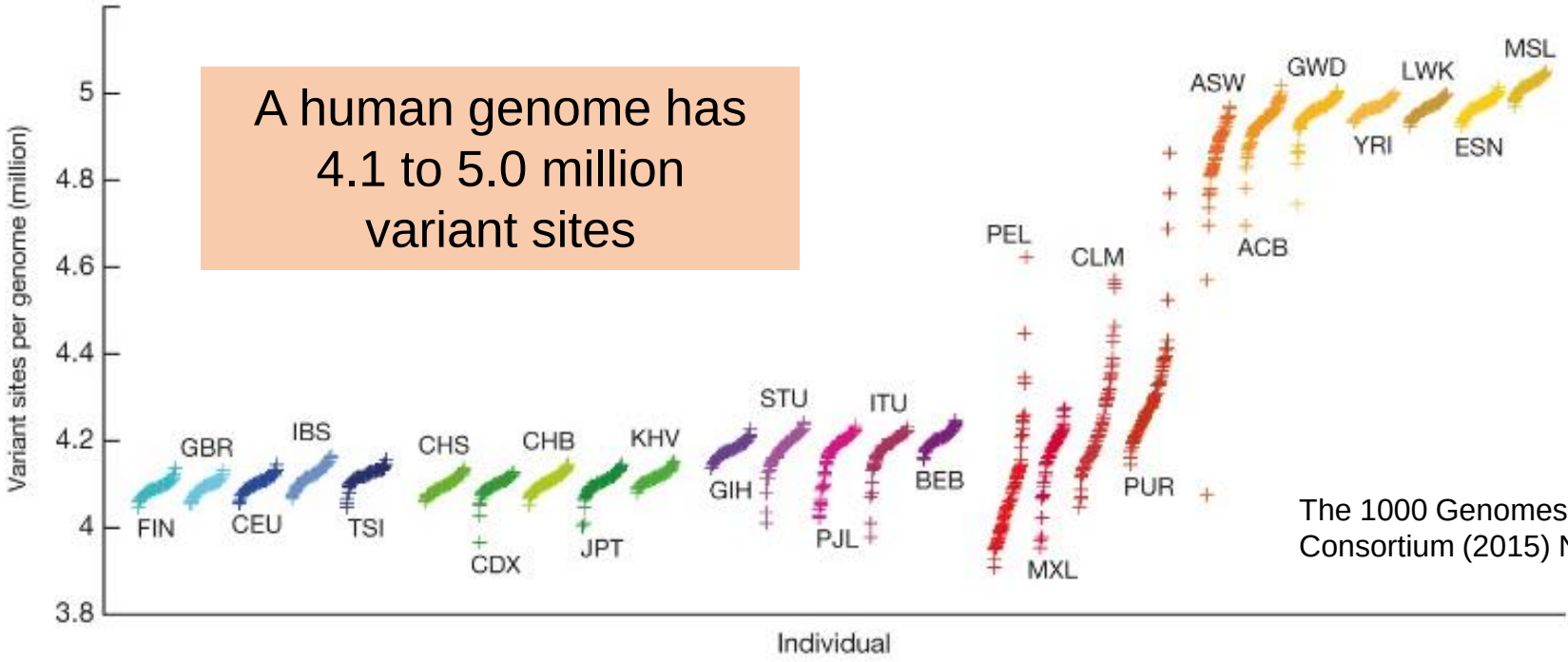


allele **A**

allele **G**

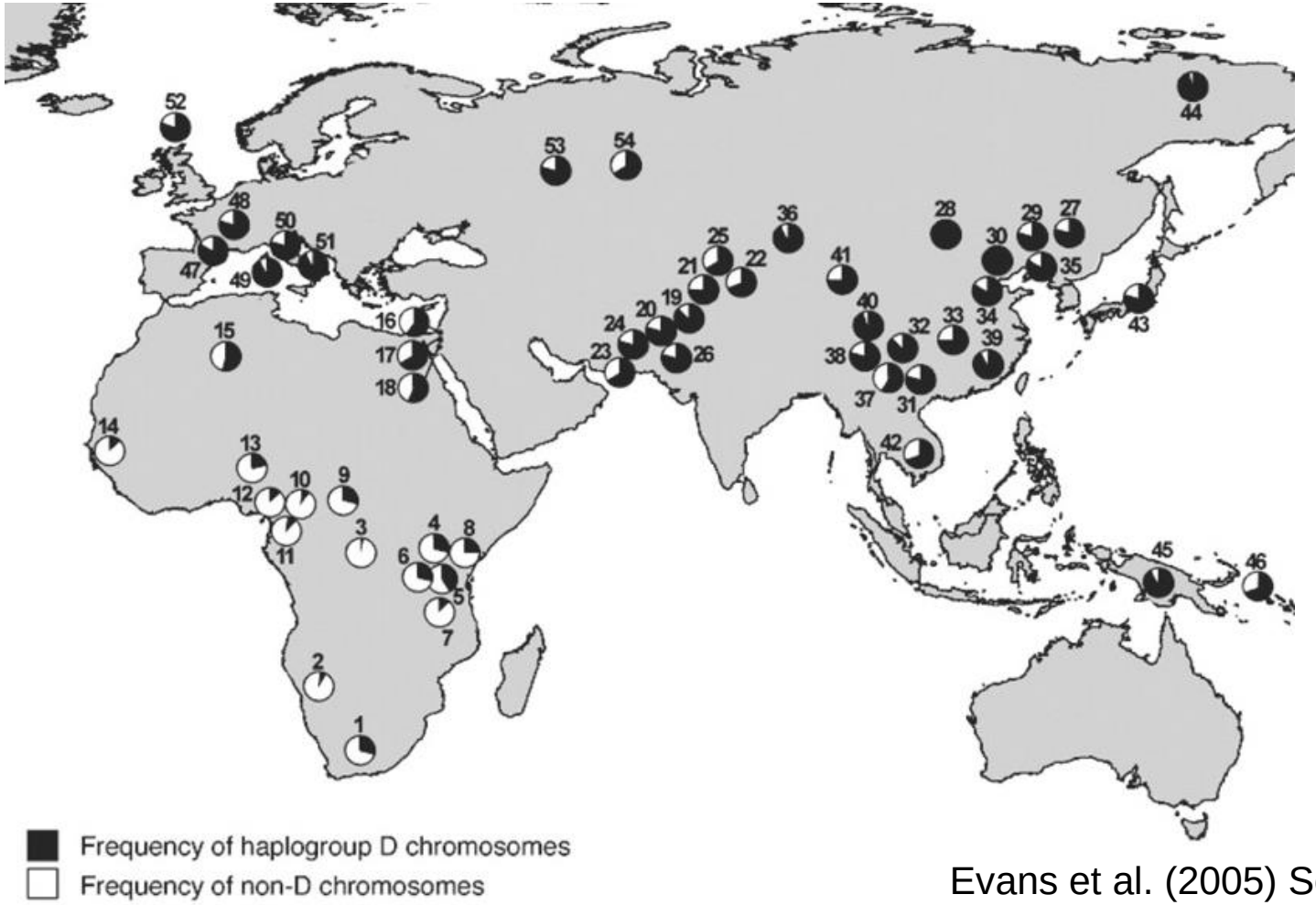


A human genome has 4.1 to 5.0 million variant sites



The 1000 Genomes Project Consortium (2015) Nature

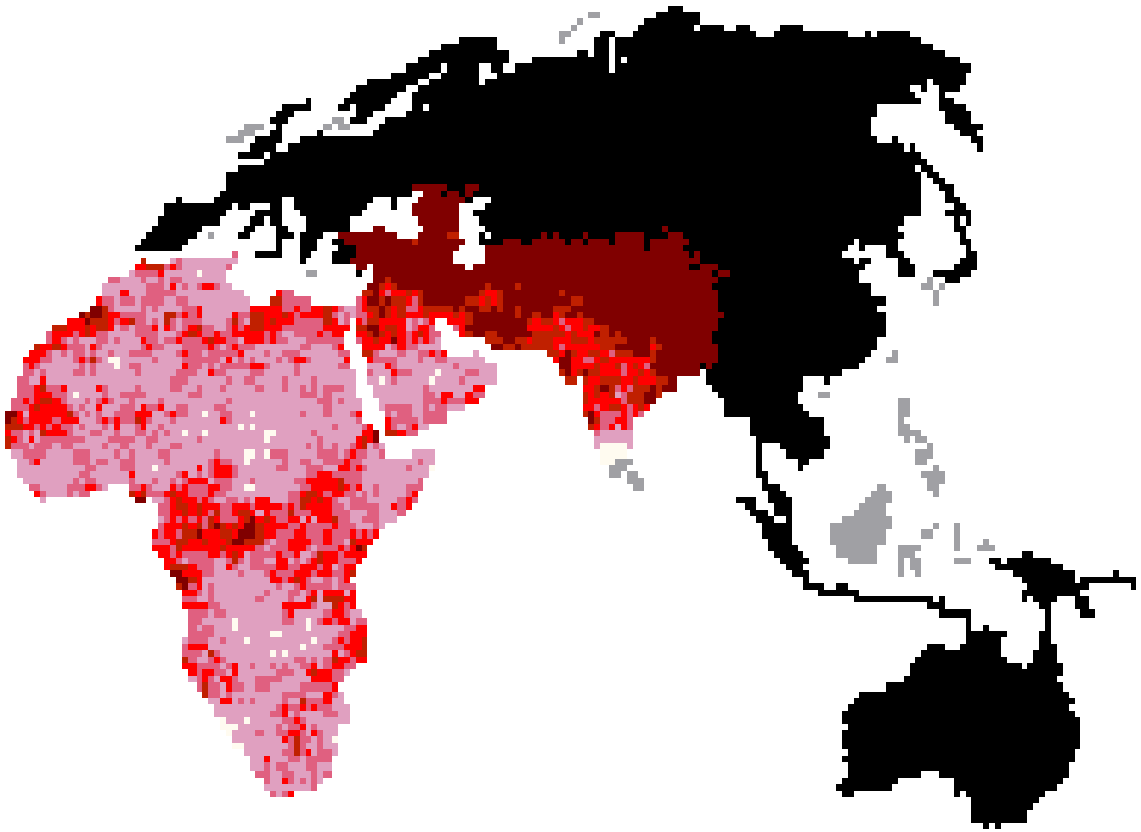
Example: Microcephalin gene in humans



Evans et al. (2005) Science

High frequency of one of the alleles in Europe and Asia interpreted as evidence of **positive selection** outside Africa

Example: Microcephalin gene in humans

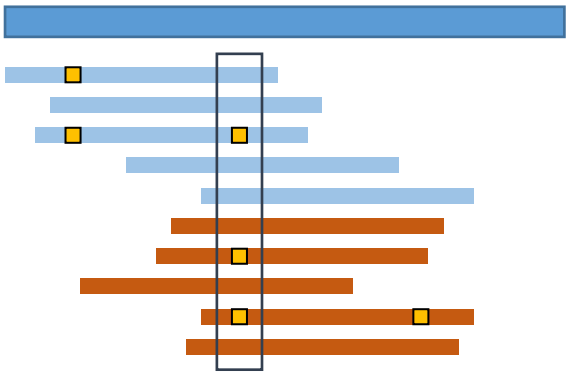


Currat et al. (2006) Science

Simulations of a **spatial expansion** (demographic factor) create similar allele frequency patterns

Genomic data with Next Generation Sequencing (NGS)

Mapping



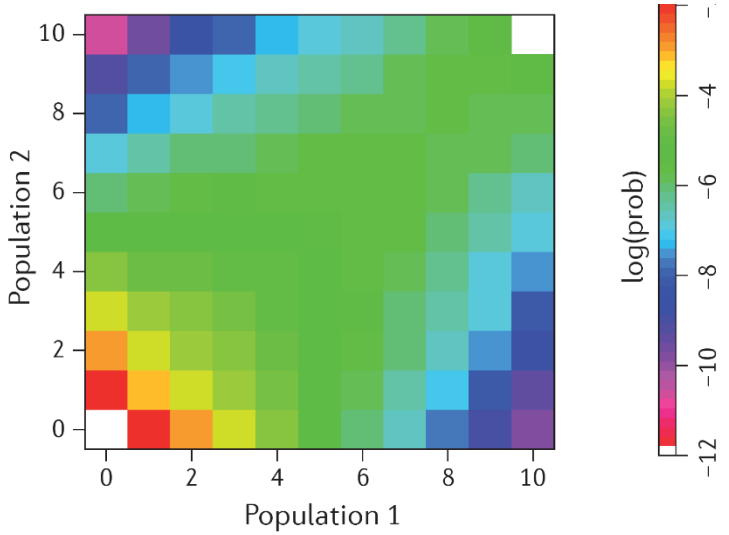
Genotypes

	SNP1	SNP2	SNP3	...	SNP *L*
Ind. 1	0	2	0	...	1
Ind. 2	0	0	1	...	2
Ind. 3	1	0	0	...	2
...	...	...	...	...	...
Ind. *n*	0	0	1	...	0

Information about:

- frequencies of variants
- linkage disequilibrium

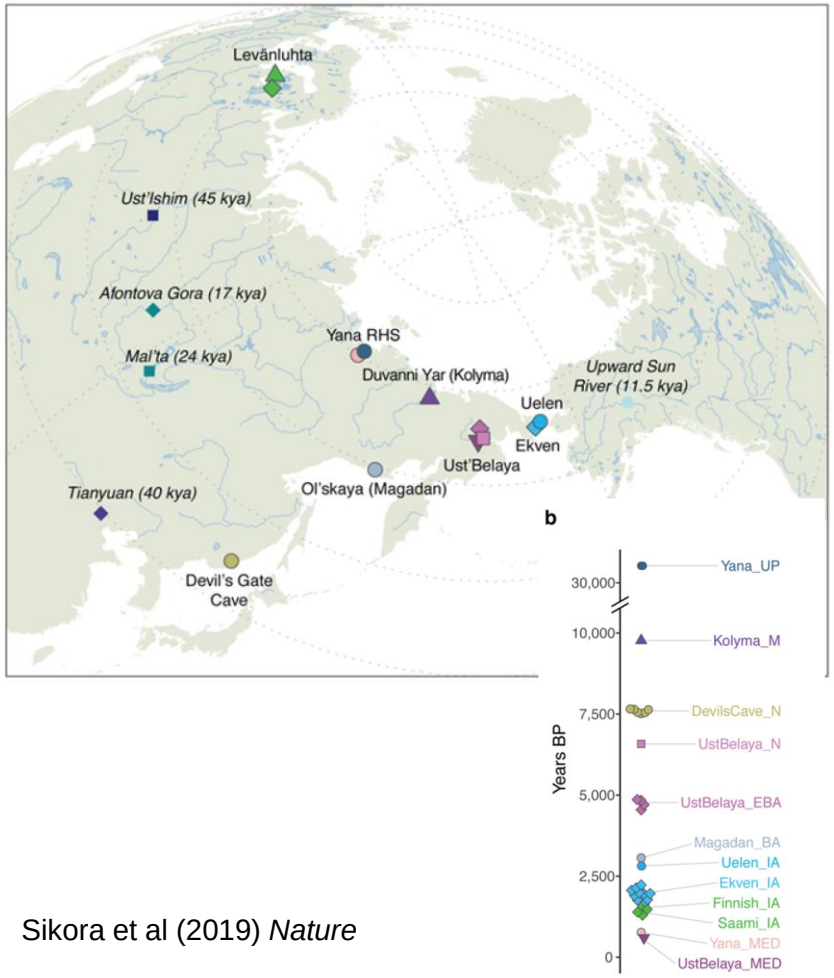
Site-frequency spectrum (SFS)



Genomic data across space and time

Natural populations

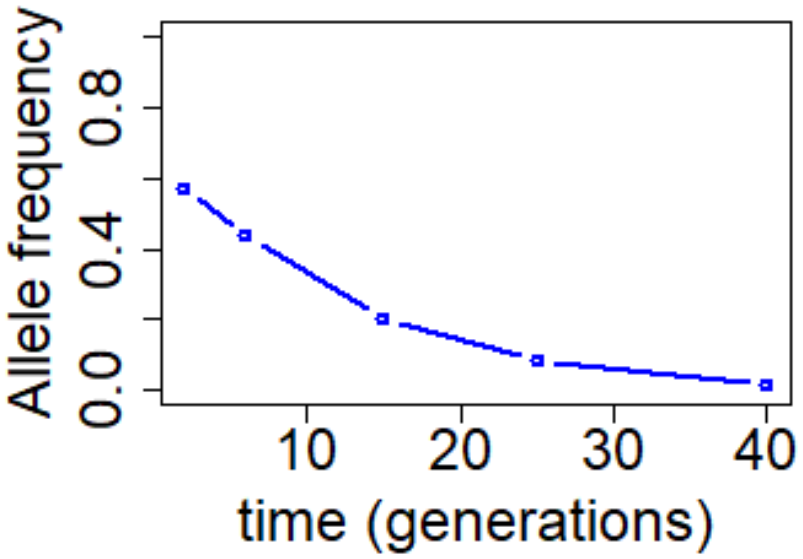
Example: modern humans



Sikora et al (2019) *Nature*

Experimental evolution

Example: *Drosophila* adaptation into lab



Fragata et al (2014) *JEB*

Example of a population genetics dataset Variant Call Format (VCF file)

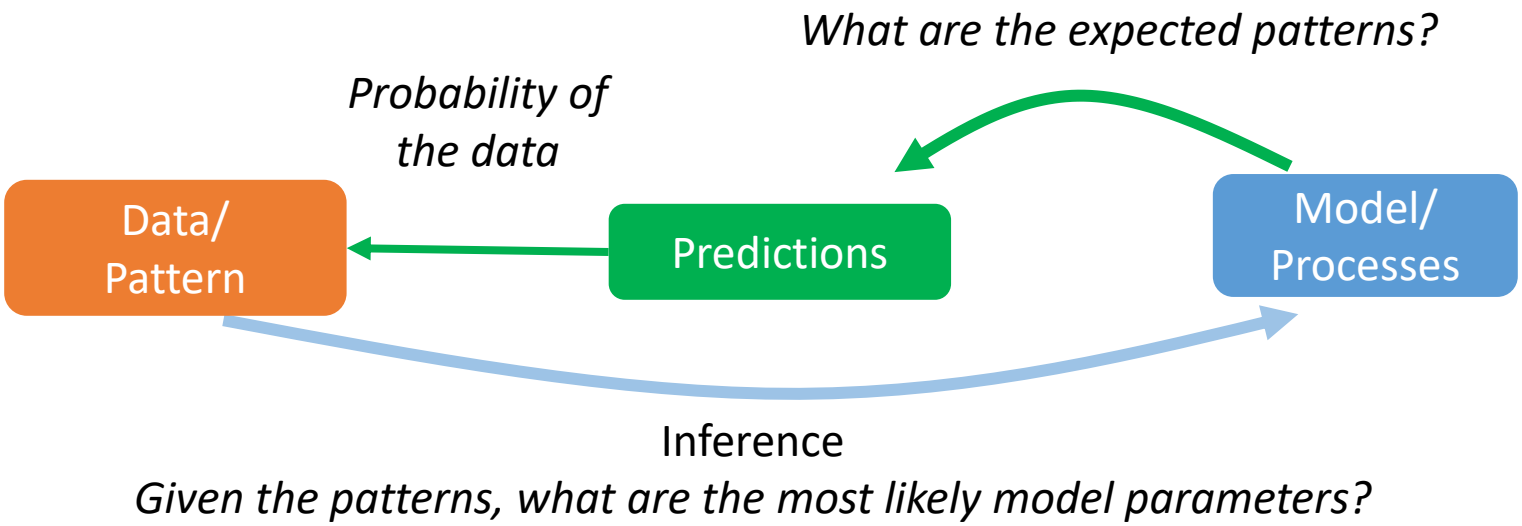
individuals

Positions along the genome

CHROM	POS	ID	REF	ALT	QUAL	FILTER	FORMAT	BL2009P4_us23
"Supercontig_1.50"	"2"	NA	"T"	"A"	"44.44"	NA	"GT:AD:DP:GQ:PL"	"0|0:62,0:62:99:0,190,2835"
"Supercontig_1.50"	"246"	NA	"C"	"G"	"144.21"	NA	"GT:AD:DP:GQ:PL"	"1|0:5,5:10:99:111,0,114"
"Supercontig_1.50"	"549"	NA	"A"	"C"	"68.49"	NA	"GT:AD:DP:GQ:PL"	NA
"Supercontig_1.50"	"668"	NA	"G"	"C"	"108.07"	NA	"GT:AD:DP:GQ:PL"	"0|0:1,0:1:3:0,3,44"
"Supercontig_1.50"	"765"	NA	"A"	"C"	"92.78"	NA	"GT:AD:DP:GQ:PL"	"0|0:2,0:2:6:0,6,49"
"Supercontig_1.50"	"780"	NA	"G"	"T"	"58.38"	NA	"GT:AD:DP:GQ:PL"	"0|0:2,0:2:6:0,6,49"

Data by itself does not say much…

Models: Historical science vs Experimental science



- Models have assumptions
- We can reject models, we get probabilities for the observed data
- We can re-consider assumptions, incorporating new knowledge

Cleland (2002) *Philosophy of Science;*
https://ncse.ngo/creationism/analysis/historical-science-vs-experimental-science

Let's think about it…

- How do populations adapt to different environments?
  - *What are the processes?*
  - *What is required for evolution to happen?*
  - *What do we need to know (and measure) to characterize these processes?*

Let's think about it…

- How do populations adapt to different environments?
  - *What are the processes?*
  - *What is required for evolution to happen?*
  - *What do we need to know (and measure) to characterize these processes?*

Transformation
(evolutionary processes)

time τ

$E(t)$

State of population at time t

$E'(t+\tau)$

State of population at time $t + \tau$

Building simple blocks that *can* explain very complex processes

Dynamic Sufficiency When we say that we have an evolutionary perspective on a system or that we are interested in the evolutionary dynamics of some phenomenon, we mean that we are interested in the change of state of some universe in time. Whether we look at the evolution of societies, languages, species, geological features, or stars, there is a formal representation that is in common to all. At some time t the system is in some state E, and we are interested in the state of the system, E', at a future time, or past time, τ time units away. We must then construct laws of transformation T that will enable us to predict E' given E. Formally, we may represent this as

$$E(t) \xrightarrow{T} E'(t + \tau)$$

Lewontin (1974) The genetic basis of evolutionary change

Thinking like a population geneticist

- The transformation *T* depends on a **sufficient description of state *E*** and some parameters θ

$$E(t) \xrightarrow{T(\theta, \tau)} E'(t + \tau)$$

For instance, if we want to predict the movement of a space ship, we need to know its current position ***E*(*t*)** in a 3D space, and we need to know the velocity and acceleration in each direction (parameters θ). With this information we can built laws of transformation ***T*** to predict the position of the space ship after some time τ, i.e. ***E'*(*t*+τ)**.

- How to define the *state* of a given population in evolution?
- How to define the *laws of transformation* in evolution?

Lewontin (1974) The genetic basis of evolutionary change

State at a given time

Variation + Heredity

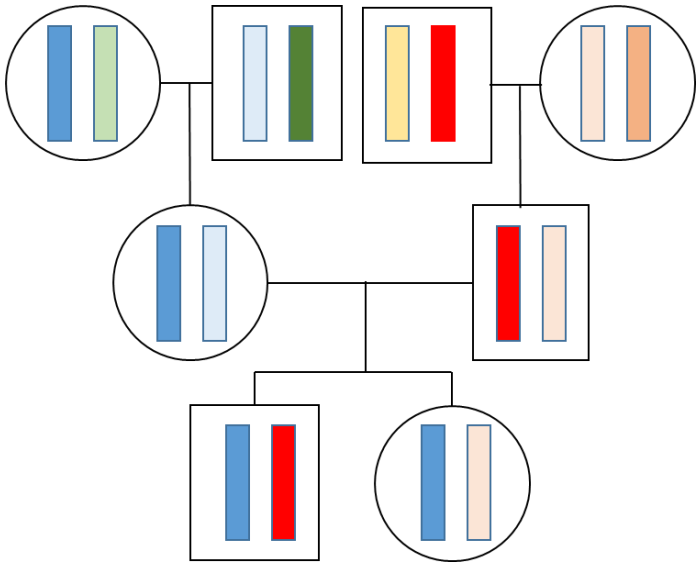
Laws of transformation

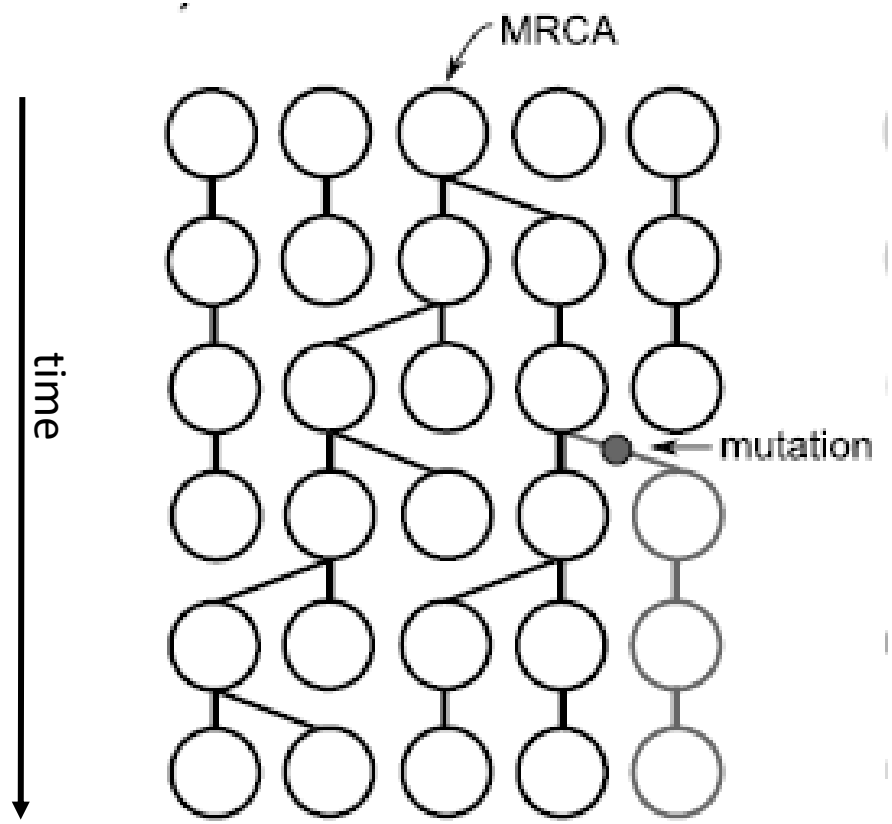
Evolutionary processes

- NEUTRAL: Genetic drift, migration, mutation, recombination
- SELECTIVE: Natural selection

Basis of population genetic models

- Mendel's principles of inheritance
- Reproduction between individuals in a population

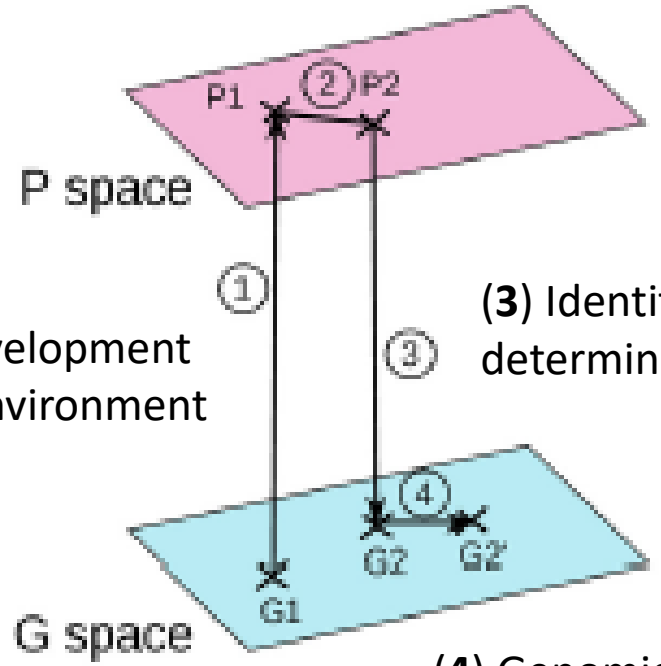




A way to think about how populations evolve: **Genotype-phenotype map**

(**2**) Migration, reproduction and natural selection acts on P1

Quantitative Genetics



Development Evo-Devo

(1) Development + Environment

(**3**) Identity of successful parents determines preserved genotypes

Population Genetics

(**4**) Genomic processes such as mutation and recombination modify position in G space

Orgogozo et al (2015) Frontiers in Genetics

Example of Genotype-Phenotype map

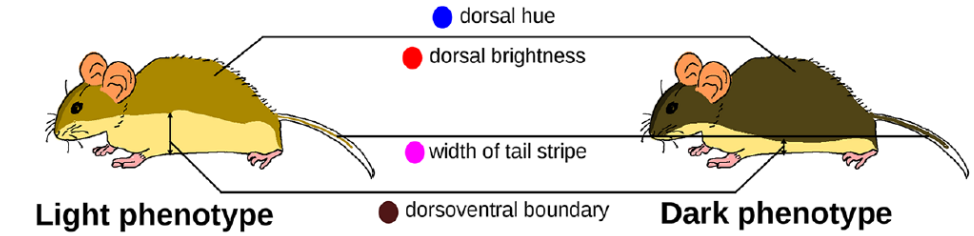


FIGURE 3 | Evolution of light-colored beach mice is caused by several mutations with distinct pigmentation effects in the *Agouti* locus. The dark and light phenotypes can be decomposed into four phenotypic traits, which are associated with different single nucleotide polymorphisms (SNPs, colored dots) located in the *Agouti* gene. Only SNPs with inferred selection coefficient of the light allele higher than 0.1 are shown. Coding exons are represented as dark boxes and untranslated exons as white boxes. Adapted from Linnen et al. (2013).

Orgogozo et al (2015) Frontiers in Genetics

Population genetics is central to Biology

"Nothing in biology makes sense except in the light of evolution"

Theodosius Dobzhansky

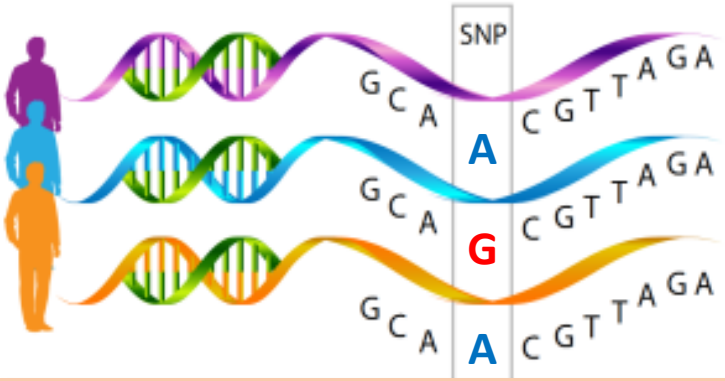
The American Biology Teacher, March 1973

"Nothing in evolution makes sense except in the light of population genetics"

Michael Lynch

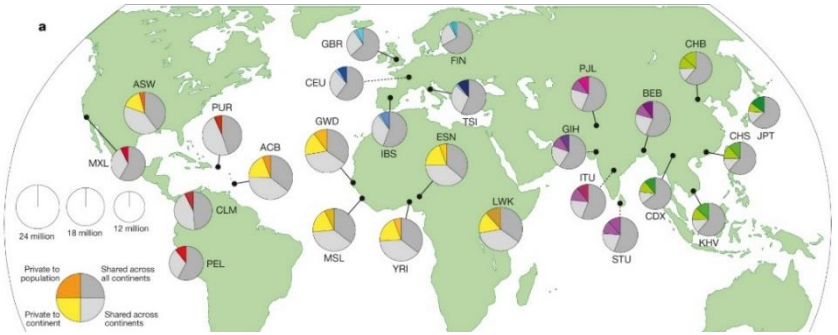
Xth Meeting of the European Society for the Study of Evolution, Krakow, August 2005

Population genetic data

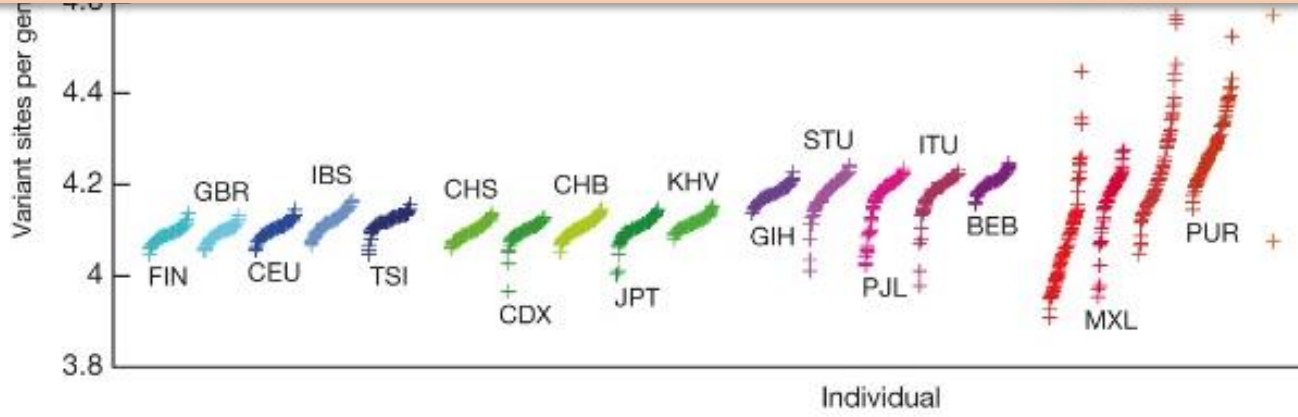


allele **A**

allele **G**



What processes lead to allele frequency differences among populations?



The 1000 Genomes Project Consortium (2015) Nature

Evolutionary history of populations is complex

Genomic processes

- Mutation
- Recombination

Demography

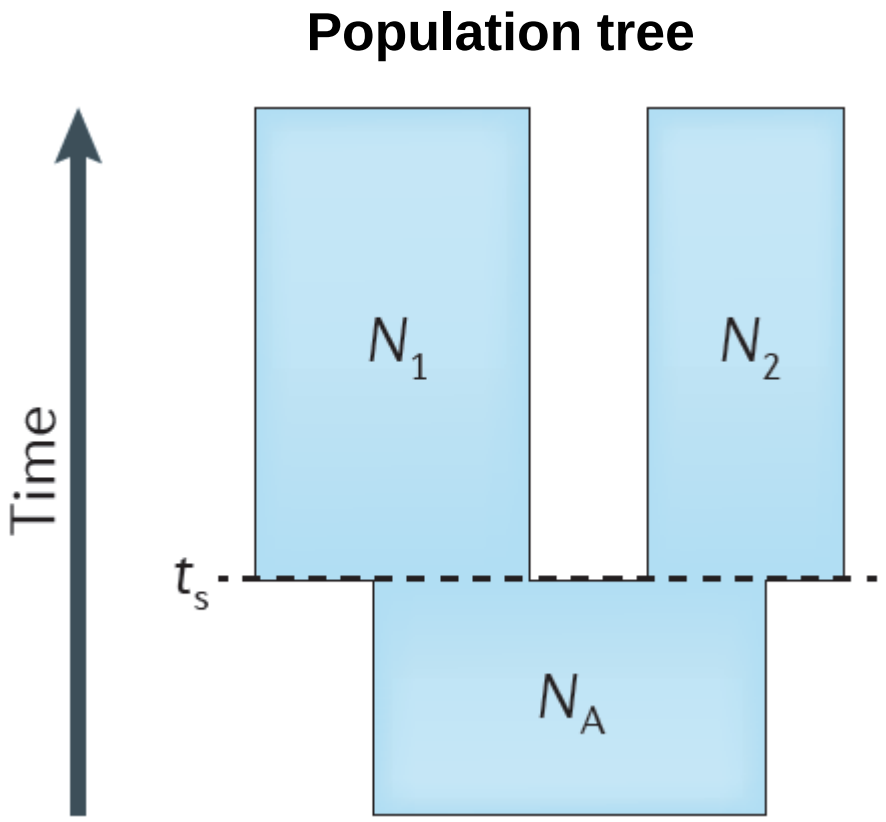
- Effective population sizes
- Population split times
- Migration rates

Selection

Natural selection:

- Beneficial mutations involved in adaptation
- Deleterious mutations with negative effects

The simple view of a population geneticist

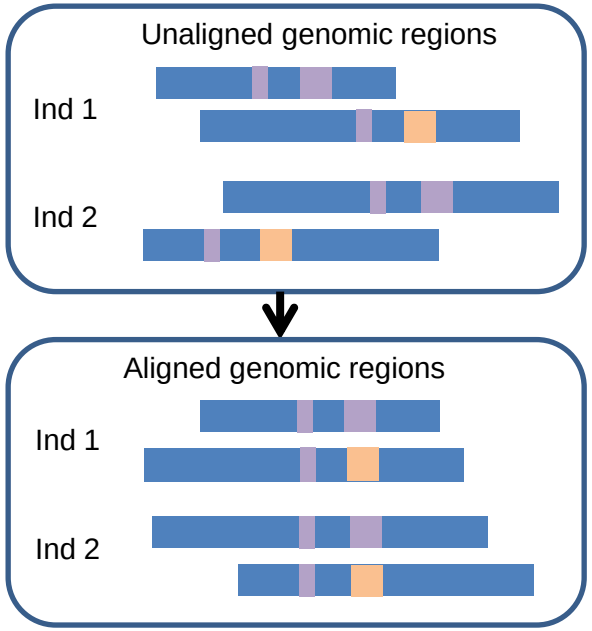


What is the effect of these processes on genetic diversity patterns?

How to measure genomic diversity?

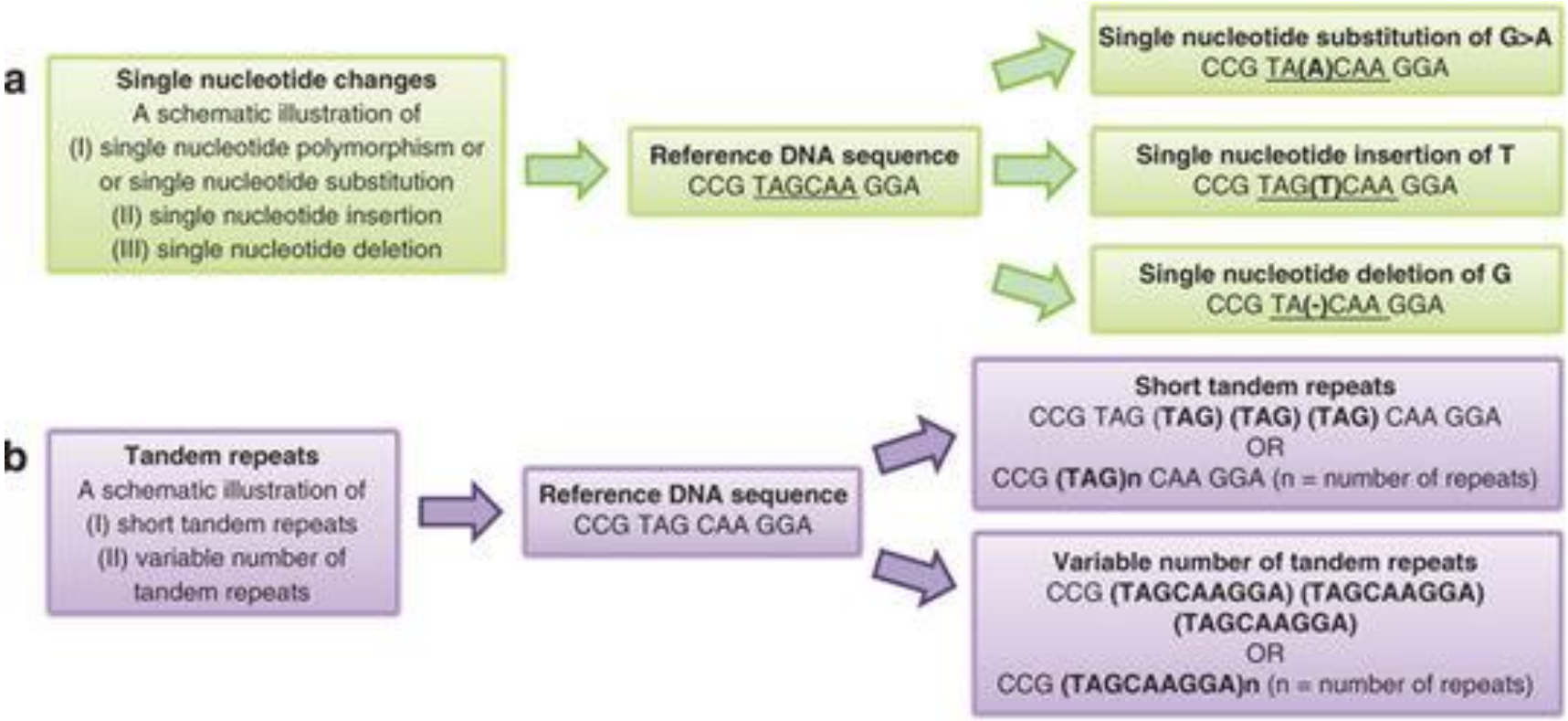
- Look at **homologous regions** of the genome of different individuals

- Find regions of the genome (locus) where there is variation among individuals – **polymorphism**



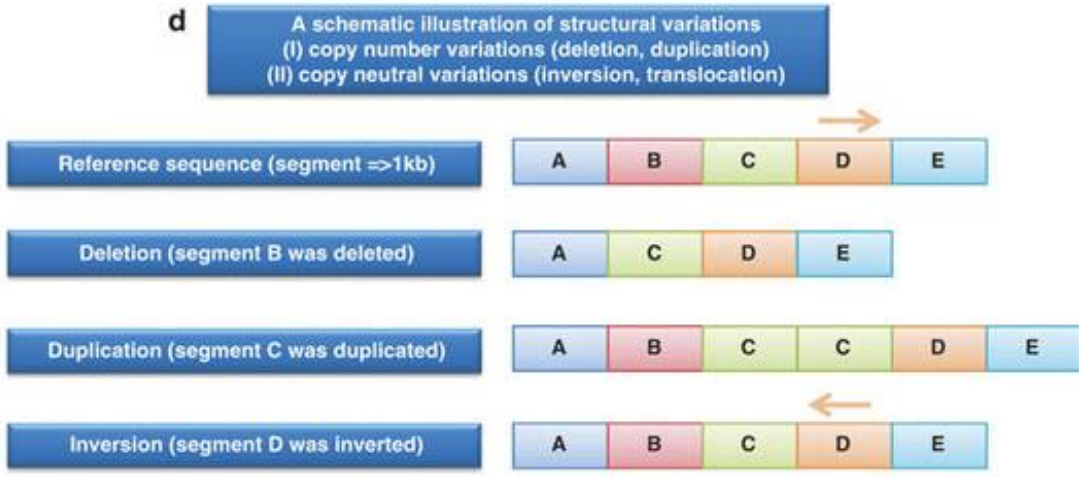
- We can characterize diversity by looking at **molecular markers:**
  - **single nucleotide polymorphisms** (SNPs)
    - the most used type of marker for population genomic analysis
  - **short tandem repeats** (STR or microsatellites)
    - still used when we cannot afford to obtain genome-wide data
  - **structural copy number variation** (CNV)
    - used to characterize diversity, but difficult to model with current population genetics models
  - **Older markers**: allozymes, AFLPs, RFLPs
    - still used for non-model organisms.

Genomic polymorphisms



Ku et al. *Journal of Human Genetics* (2010) **55**, 403–415

Genomic polymorphisms
Structural variation



Ku et al. *Journal of Human Genetics* (2010) **55**, 403–415

inversion

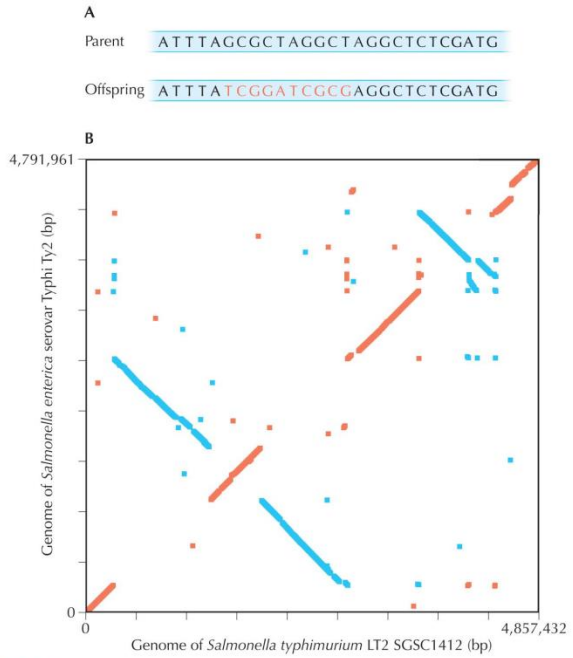
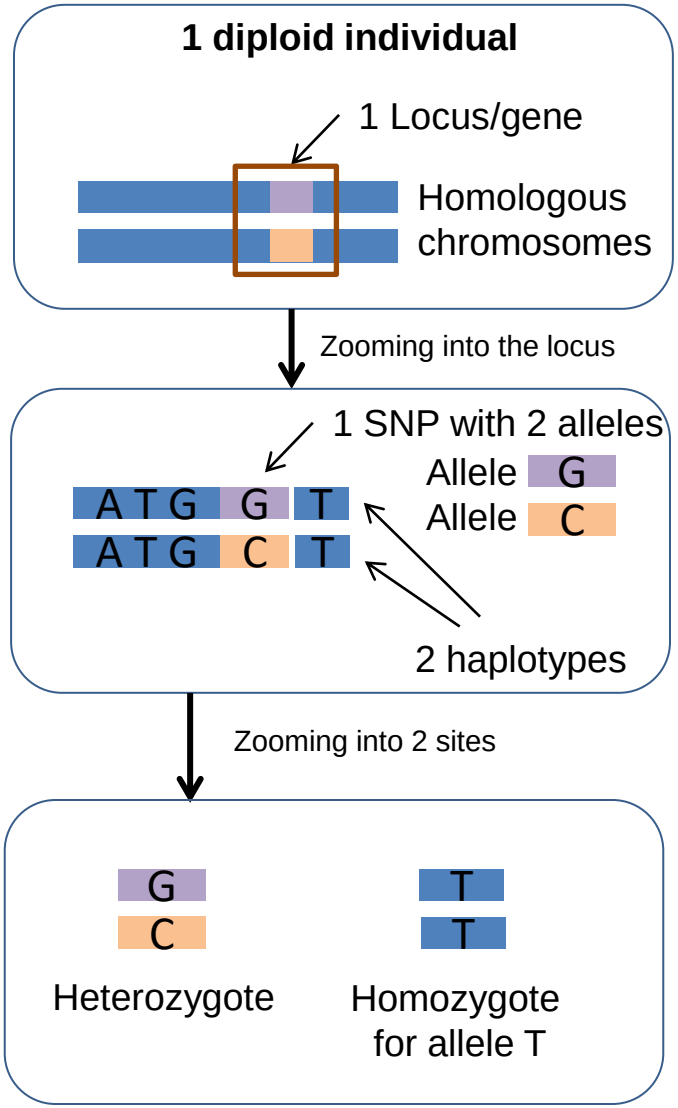


FIGURE 12.4. Inversions. (*A*) Hypothetical example of an inversion in a small section of DNA. Parental sequence is shown *above* and the offspring sequence is shown *below* with the inverted region highlighted in *red*. (*B*) Comparison of the genomes of two strains of the bacterial genus *Salmonella* showing the occurrence of multiple large inversions. The diagram shows a genome dot-plot (as in Box 7.1). The genome of one strain is on the *x*-axis and the other is on the *y*-axis (with the replication origins at $(x, y) = (1, 1)$). Conserved regions between the two genomes are indicated by a *dot*. If the two genomes showed the same total orientation, all the dots would be on the $y = x$ diagonal. The *blue* segments are inversions between the two strains.

12.4, source generated by author, using the MUMMER program, Comprehensive Microbial Resource

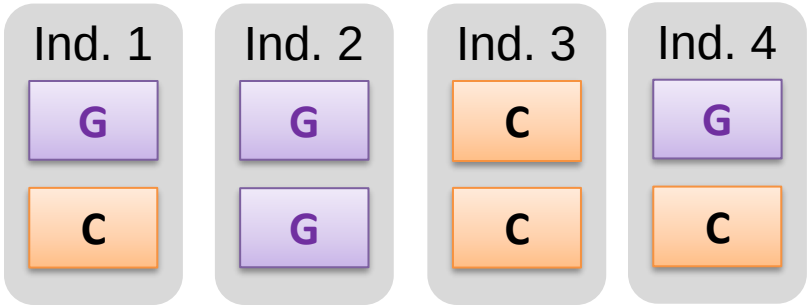
Population genomics – some definitions

- **Gene**: A unit of inheritance, a non-recombining segment of DNA. A given location on a chromosome
- **Locus**: A given location on a chromosome, a non-recombining segment of a chromosome. We shall sometimes use **gene** when we are meaning a given locus on a chromosome.
- **Allele**: A possible form of a gene. Alleles usually differ from each other by mutations.
- **Genotype**: The combination of the two homologous **alleles** carried on the two chromosomes of a diploid individual at a given locus. Genotypes can also describe the allelic constitution at several loci (multi-locus genotypes)
- **Haplotype**: A particular combination of alleles at different loci on a chromosome
- **Polymorphism**: Describes the fact that there exists different alleles at a given locus in a population.
- **Population:** A group of interbreeding individuals living together in time and space. It is usually a subdivision of a species.
- **Sample**: A collection of individuals or of genes drawn from a population



Allele frequencies *vs* Genotypic frequencies

- Genotype determines how alleles occur within individuals
- Allele frequencies describe the frequency in the population or sample



1. What are the genotype frequencies?
2. What are the allele frequencies?
3. Can we predict the allele frequencies from genotype frequencies?
4. Can we predict the genotype frequencies from the allele frequencies?

Hardy-Weinberg equilibrium

Godfrey Harold Hardy (English mathematician) and **Wilhelm Weinberg** (German Physiologist) enunciated independently in 1908 what is now known as the **Hardy-Weinberg principle**, or the principle of **Hardy-Weinberg equilibrium**:

Godfrey Harold Hardy
1877-1947

Wilhelm Weinberg
1862-1937

Hardy-Weinberg equilibrium

Consider a locus with 2 alleles *A* and *a* with frequencies of respectively *p* and *q* in the population.

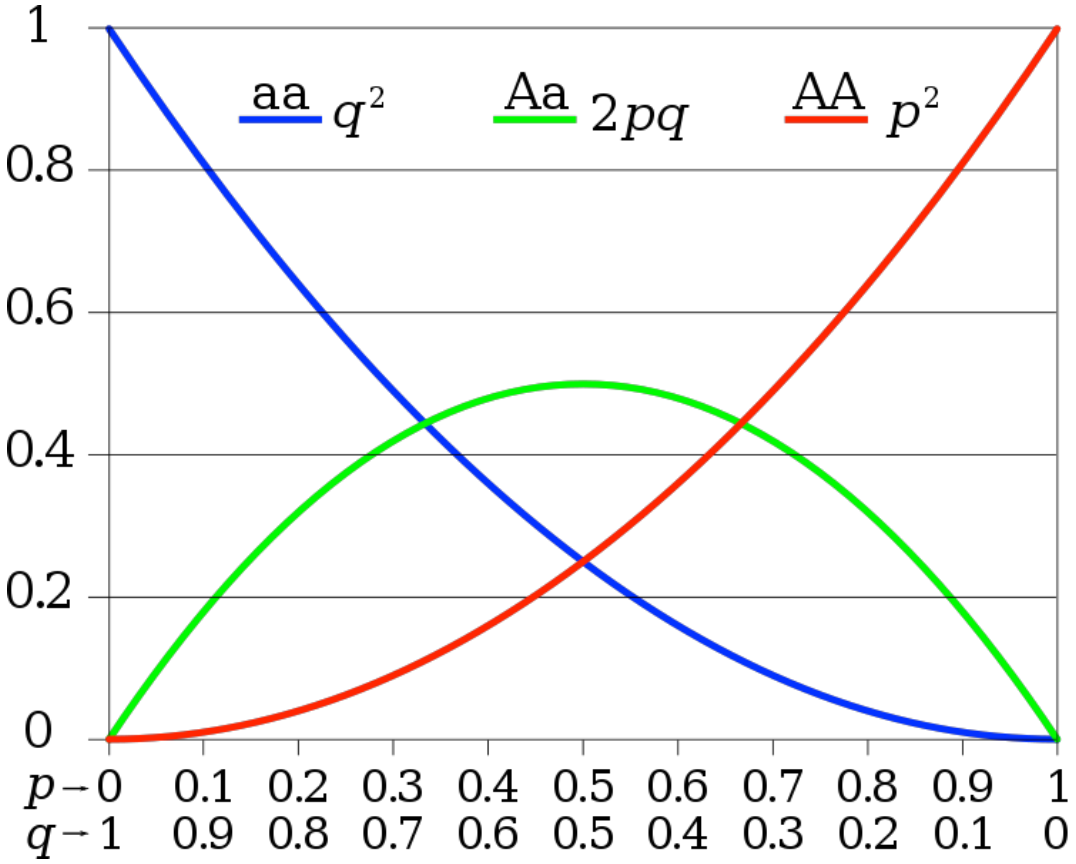
In a idealized diploid population where **reproduction occurs at random**, the relationship between the allele and the genotype frequencies at a bi-allelic locus is given by the relationship:

Genotypes	*AA*	*Aa*	*aa*
Observed frequency	*P*	*H*	*Q*
Expected frequency	p^2	$2pq$	q^2

The Hardy–Weinberg equilibrium refers to the

equilibrium between expected and observed genotype frequencies

Genotype frequencies as a function of allele frequencies



Implications of HWE

1. Allele frequencies are stable over time in an ideal population with random reproduction and infinite population size
2. Since genotypic frequencies can be predicted from allele frequencies, they also remain stable over time
3. Because of Mendelian inheritance, **Hardy-Weinberg equilibrium is reached after a single generation of random mating** (in case of externally driven change of allele frequencies).
4. Allele frequencies can only change due to departure from assumption of ideal population
    1. **Genetic drift** (finite size of the population)
    2. **Mutation**
    3. **Migration**
    4. **Selection**

Evolutionary forces

Note that departure from random mating do not change allele frequencies but only affect genotype frequencies

Allele frequencies vary among populations

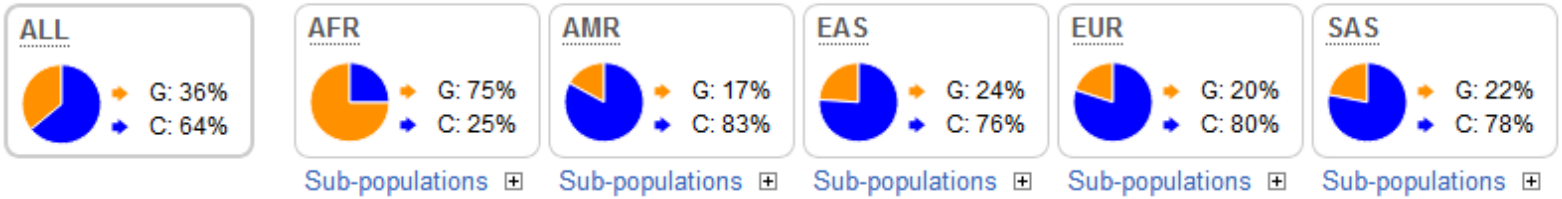
riant: rs930557

rs930557 SNP

Most severe consequence	missense variant | See all predicted consequences
Alleles	G/C | Ancestral: G | MAF: 0.36 (G) | Highest population MAF: 0.37
Location	Chromosome 8:6302183 (forward strand) | VCF: `8 6302183 rs930557 G C`
Evidence status	
Clinical significance	
HGVS names	This variant has 19 HGVS names - Show
Synonyms	This variant has 8 synonyms - Show
Genotyping chips	This variant has assays on 7 chips - Show
Original source	Variants (including SNPs and indels) imported from dbSNP (release 150) | View in dbSNP
About this variant	This variant overlaps 3 transcripts, has 3154 sample genotypes, is associated with 3 phenotypes
Description from SNPedia	Description not available [More information from SNPedia]

Population genetics

1000 Genomes Project Phase 3 allele frequencies



Evolutionary processes

Genomic processes

- Mutation
- Recombination

Demography

- Effective population sizes
- Population split times
- Migration rates

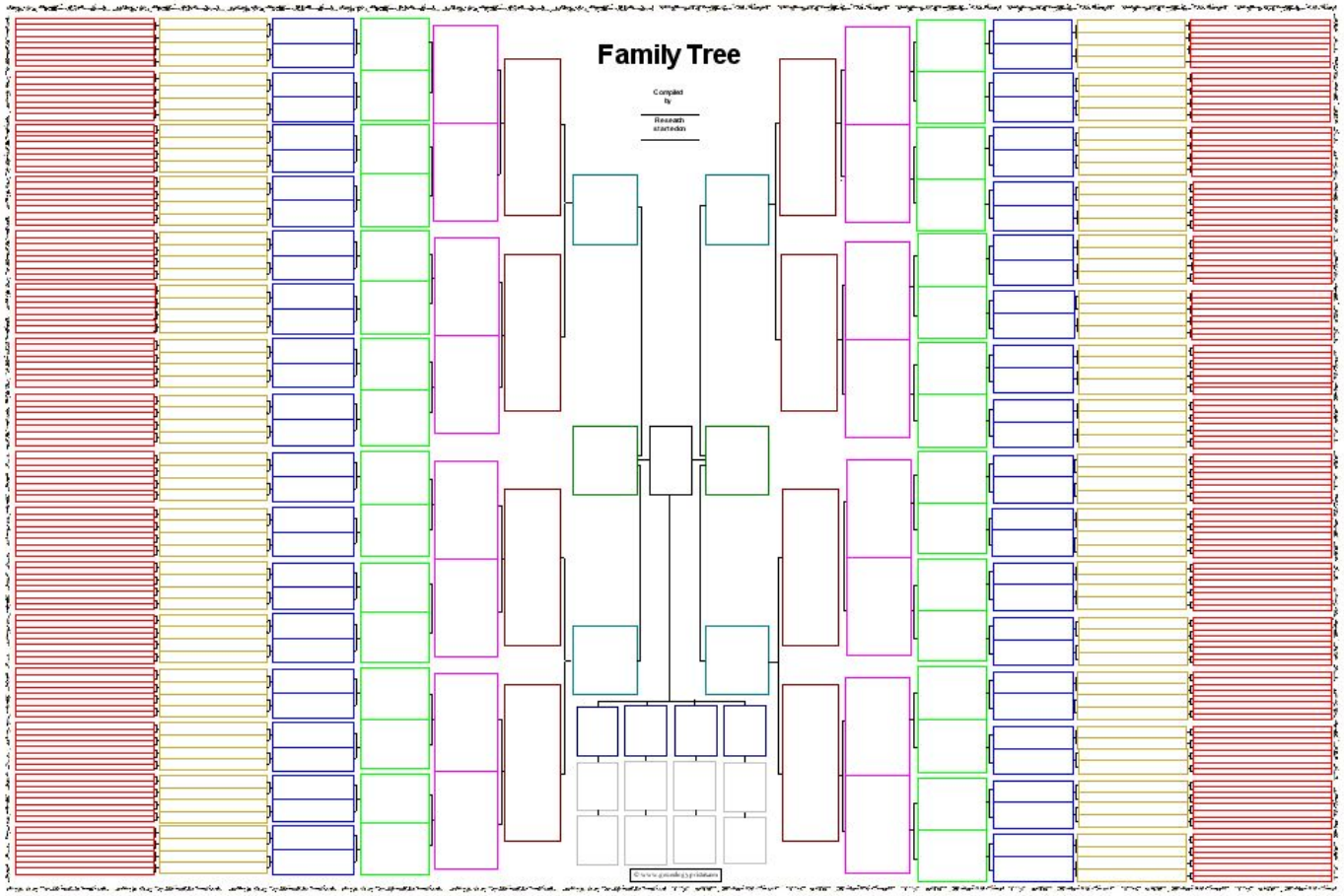
Selection

Natural selection:

- Beneficial mutations involved in adaptation
- Deleterious mutations with negative effects

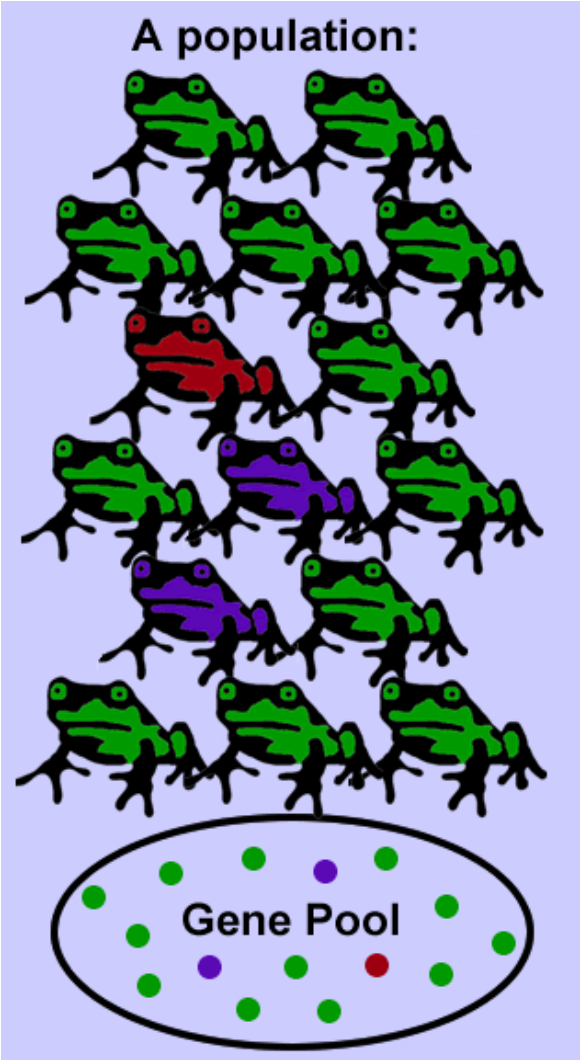
What is the effect of demographic neutral processes (i.e. without selection) on genetic diversity patterns?

Inbreeding and genetic drift



- Imagine you are building your genealogical tree
- How many ancestors do you have after 20 generations if no one was related?
  - 1,048,576 = 2^{20}
  - Those are too many in just a few generations. Hence, there must be repetitions of ancestors, some of our ancestors were related!
  - This is genetic drift and its related inbreeding

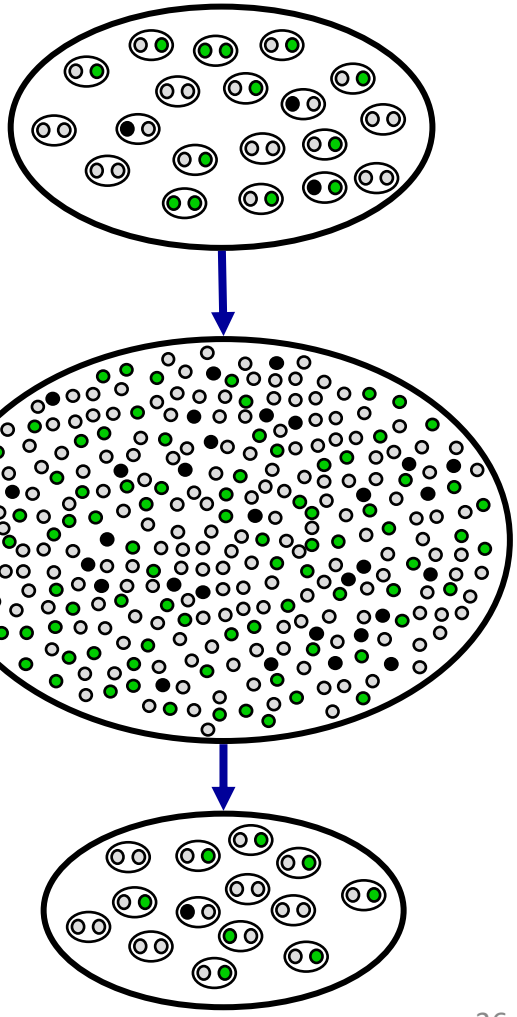
A simple model of population



Life cycle of a diploid population

Non-overlapping generations

Reproducing individuals
generation *t*
$2N = 40$



Gametic pool
$2N \rightarrow \infty$

Reproducing individuals
generation *t+1*
$2N = 30$

An ideal population

Properties of an ideal diploid population studied at a single autosomal locus with Mendelian inheritance

1. Discrete non-overlapping generations
2. Allele frequencies are identical in males and females
3. Panmictic population: Individuals randomly choose their partners (random union of gametes)
4. Population size is very large (infinite)
5. There is no migration (closed population)
6. Mutations can be ignored
7. Selection does not affect allele frequencies (neutral alleles)

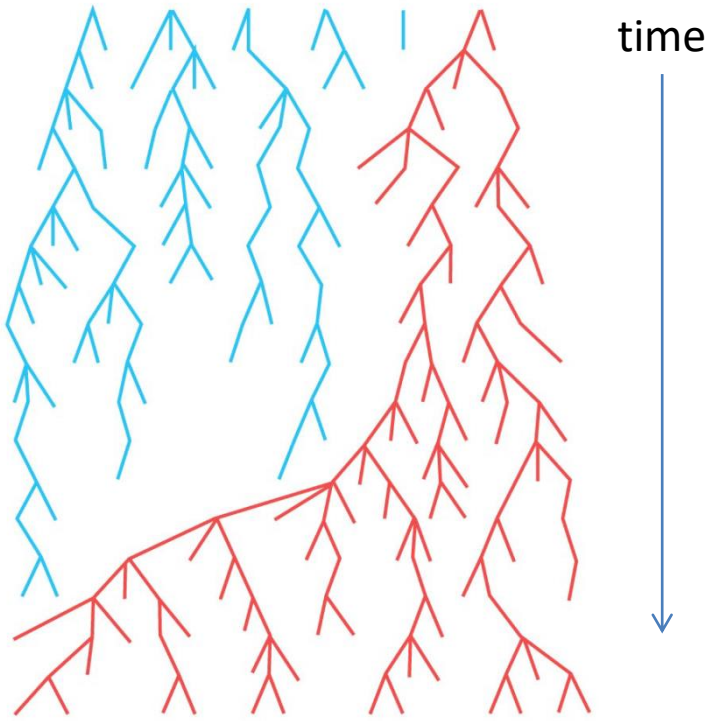
Genetic drift

An important evolutionary force

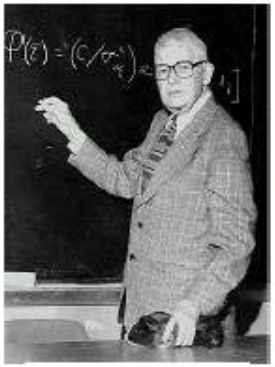
Genetic drift describes the **random fluctuations of allele frequencies occurring over time in a population of finite size**, due to the fact that the individuals of a populations will have by chance different numbers of children.

NOTE: **It is not due to natural selection!** All individuals have the same chance of having offspring, but by chance some do others do not.

It follows that some genes will be more or less transmitted to the next generation and thus allele frequencies will change over time.



15.0, Genetic drift

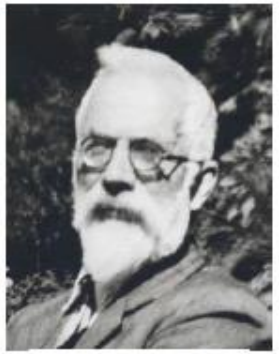


Sewall Wright

Wright-Fisher model

W-F population assumptions :

- finite number *N* of individuals



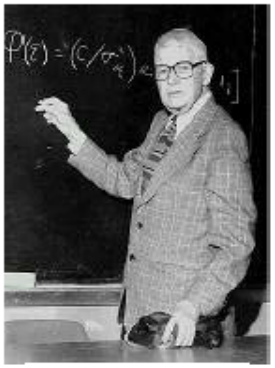
Ronald Fisher

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Let's draw 16 numbers at random (say in R)

```
> sample(1:16, 16, replace=T)
 [1]  9  3 14  4 13 12 12 12 12  7  5  7 15 16  3  8
```

This amounts at drawing 16 gametes at random from the gametic pool

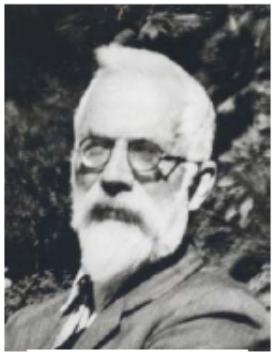


Sewall Wright

Wright-Fisher model

W-F population assumptions:

- finite number *N* of individuals
- non-overlapping generations
- monoecious (selfing is possible)



Ronald Fisher

Now examine the ancestry of the individuals

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Gen 1																
Gen 2	9	3	14	4	13	12	12	12	12	7	5	7	15	16	3	8

These two gene copies are derived from the same gene copy in the previous generation

They are identical by descent

The probability that any pair of gene is identical by descent from the previous generation is 1/16

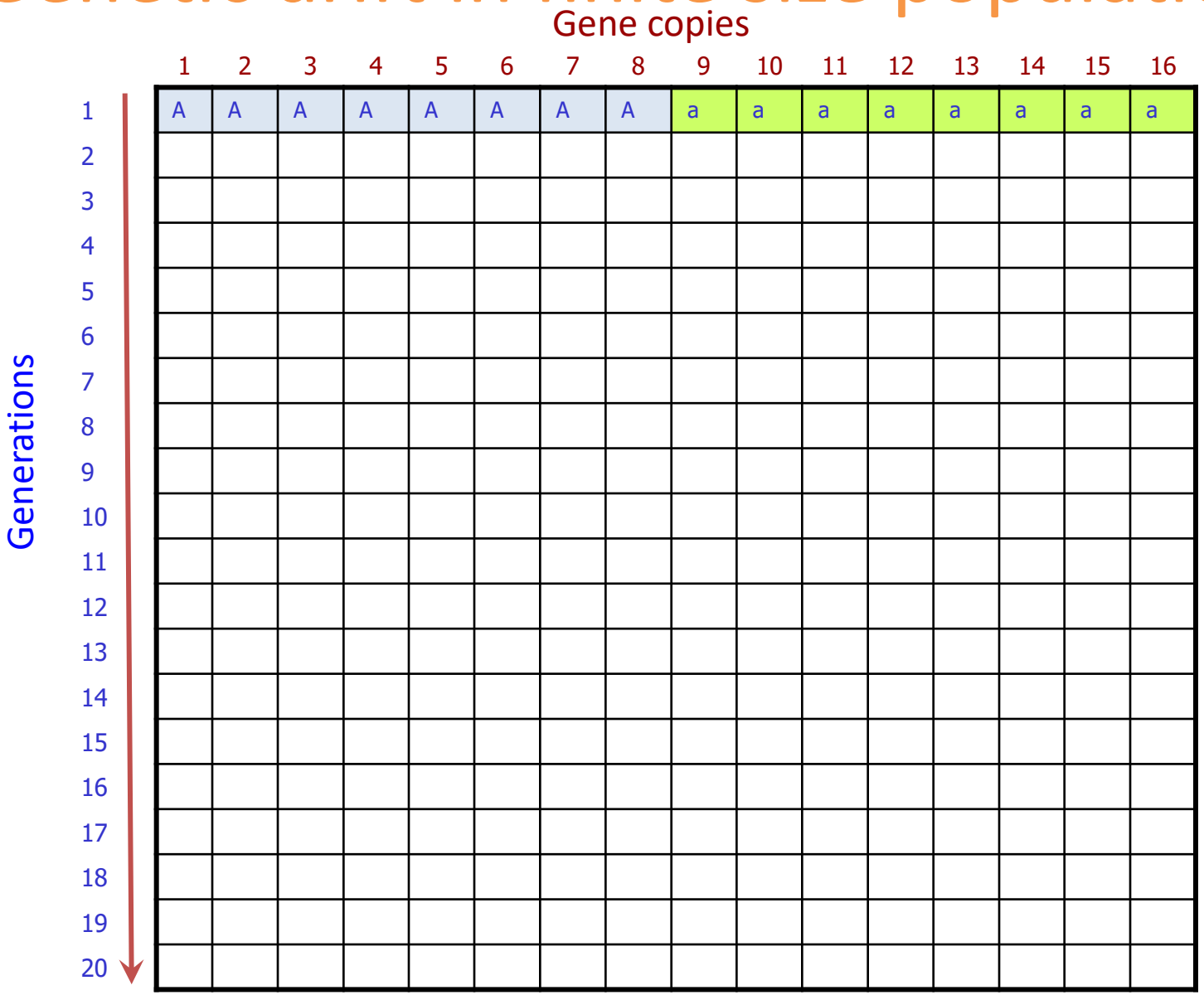
The probability that any individual is inbred is 1/16

The inbreeding coefficient of a random individual is 1/16

The inbreeding coefficient *f* of a random individual in a W-F population of size *N* is 1/(2*N*)

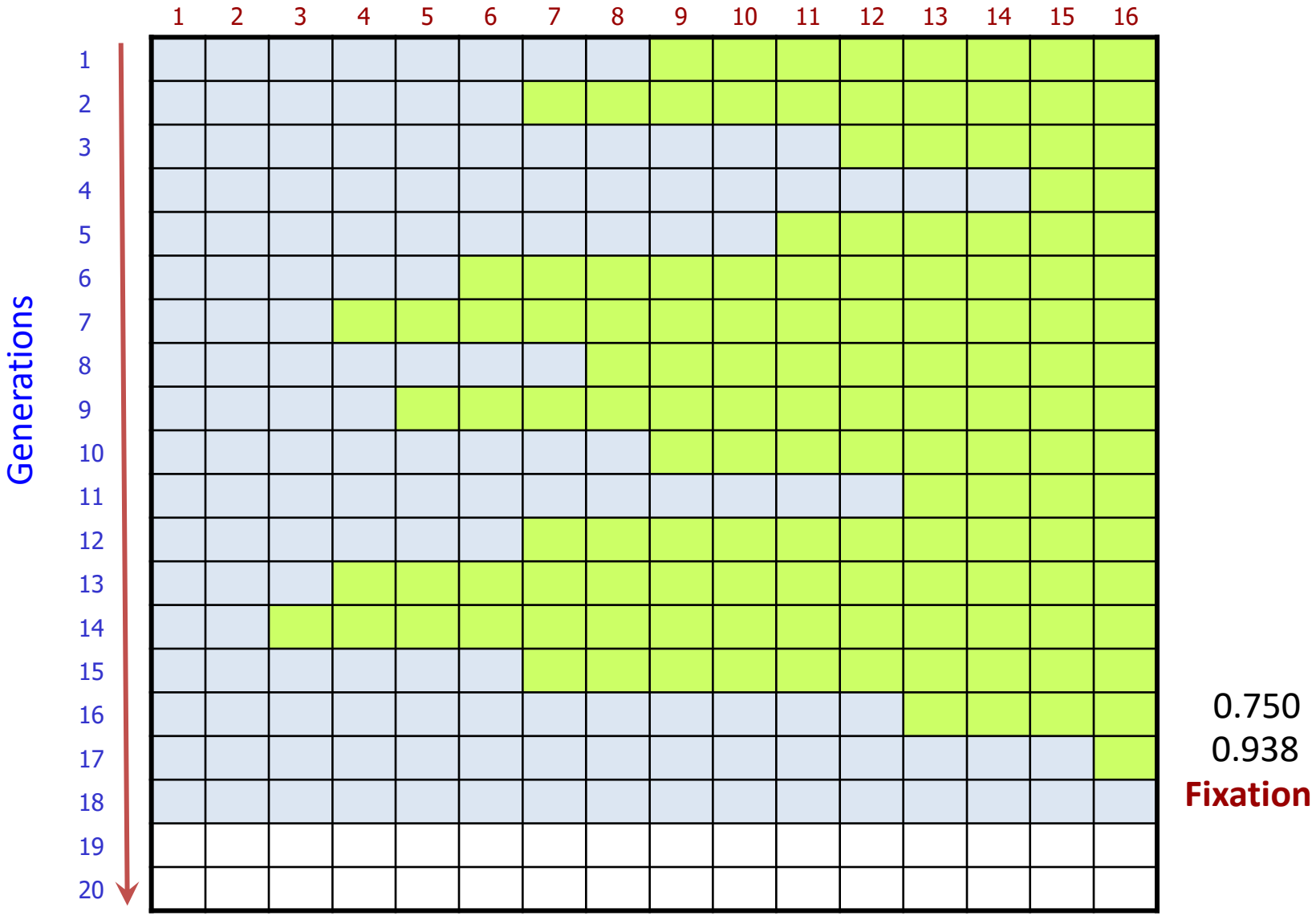
Laurent Excoffier

Genetic drift in finite size populations

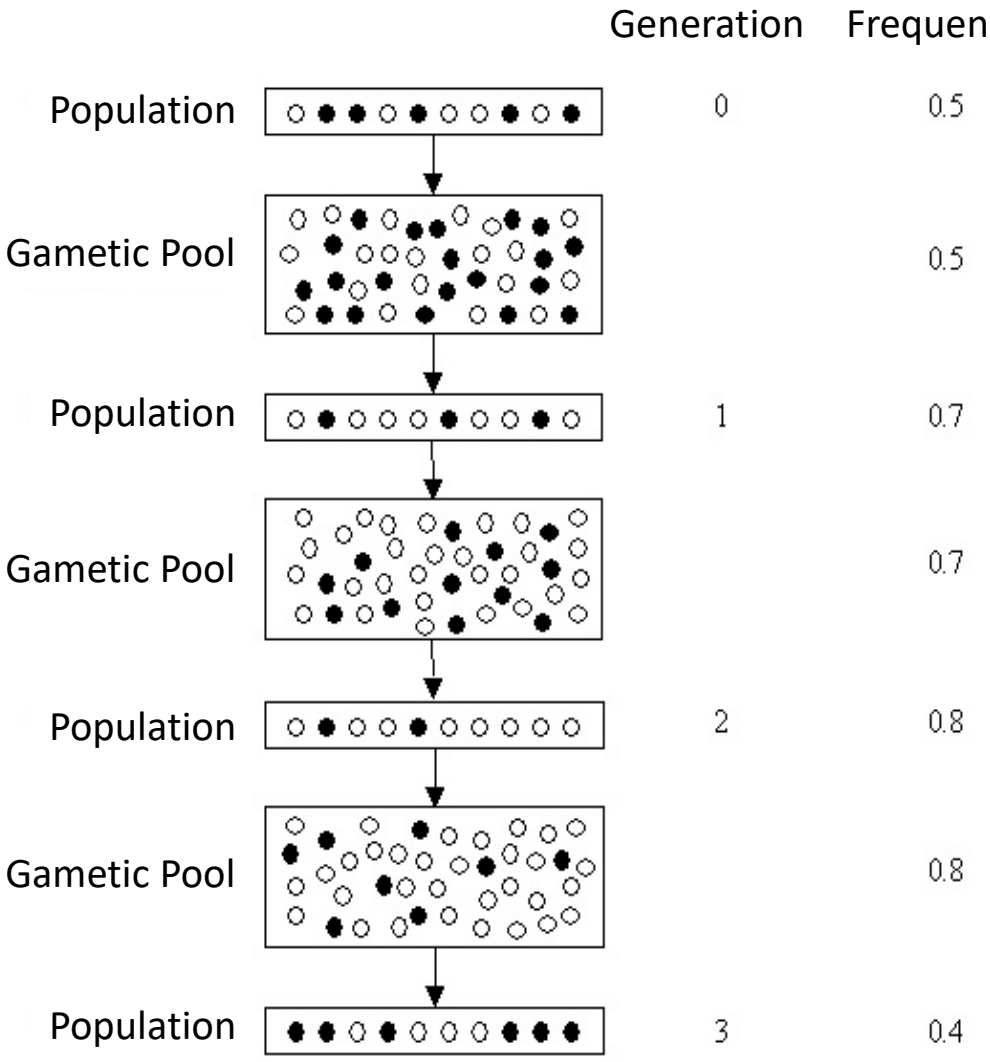


L. Excoffier

Genetic drift in finite size populations



Random change of allele frequencies in finite size populations



Let p be the frequency of the allele A at generation t

Let X be the number of alleles of type A at generation $t+1$, X follows a Binomial distribution with parameters $2N$ and p

$$X \sim b(2N, r; p)$$

$$P(X = r) = \frac{2N}{r!(2N-r)!} p^r (1-p)^{2N-r}$$

$$E(X) = 2Np \quad V(X) = 2Npq$$

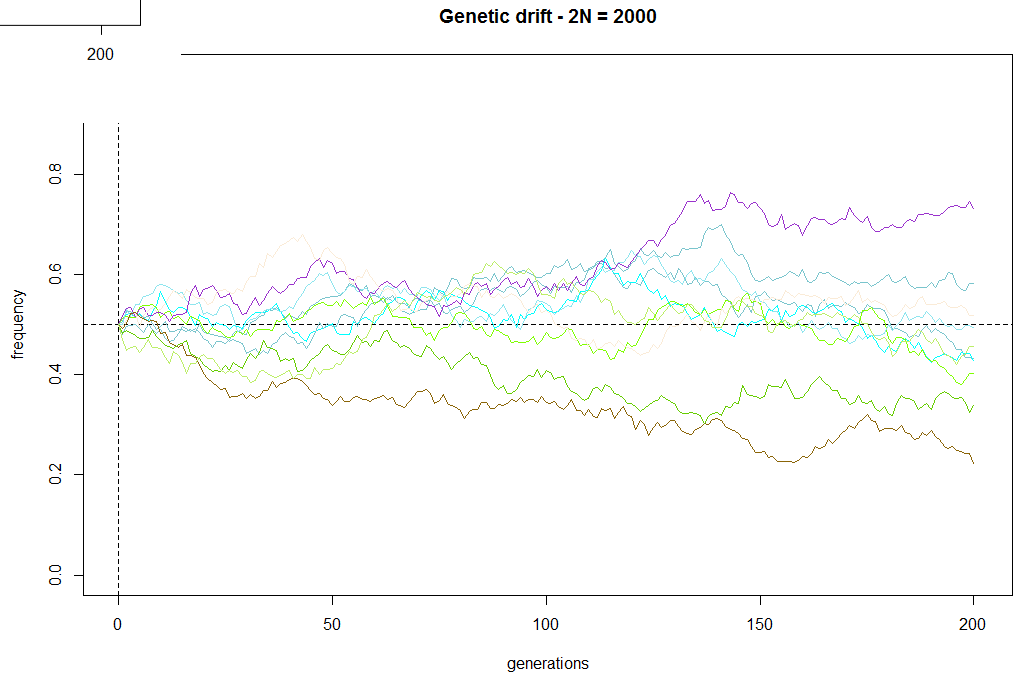
Rate of drift is stronger in small populations

The expected value (average) across many replicates is expected to remain the same as the initial allele frequency value (i.e. genetic drift has no direction).

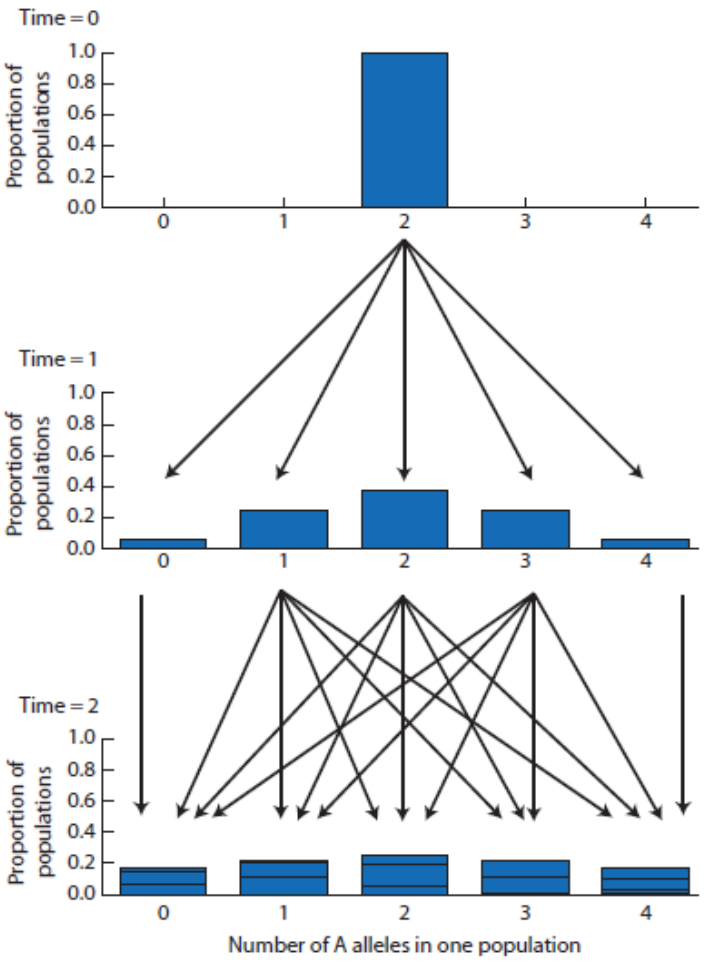
$$E\left[p(t)\right] = p_0$$

But the variance among replicates increases. Genetic drift increases the dispersion around the mean.

$$V\left[p(t)\right] = p_0 \quad (1-p_0)\left[1-\left(1-\frac{1}{2N}\right)^t\right]$$



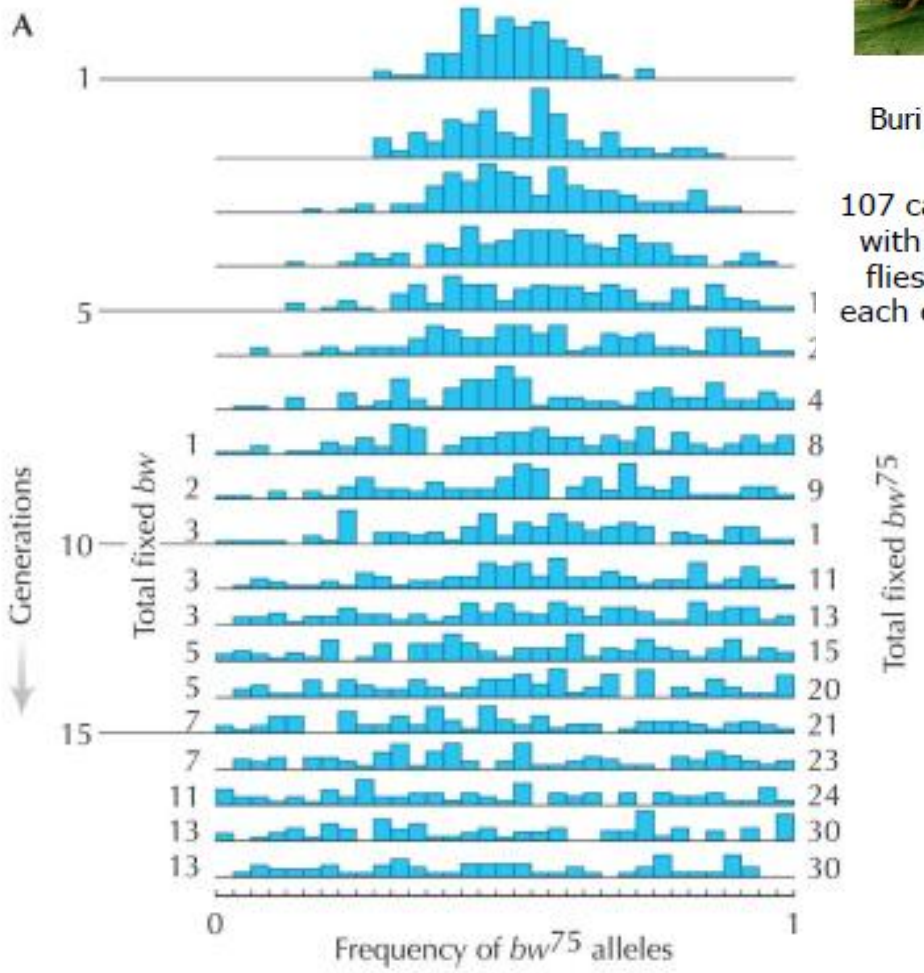
Predictions of genetic drift



Experimental evidence of genetic drift

Buri 1956

107 cages with 16 flies in each cage



Consequences of genetic drift

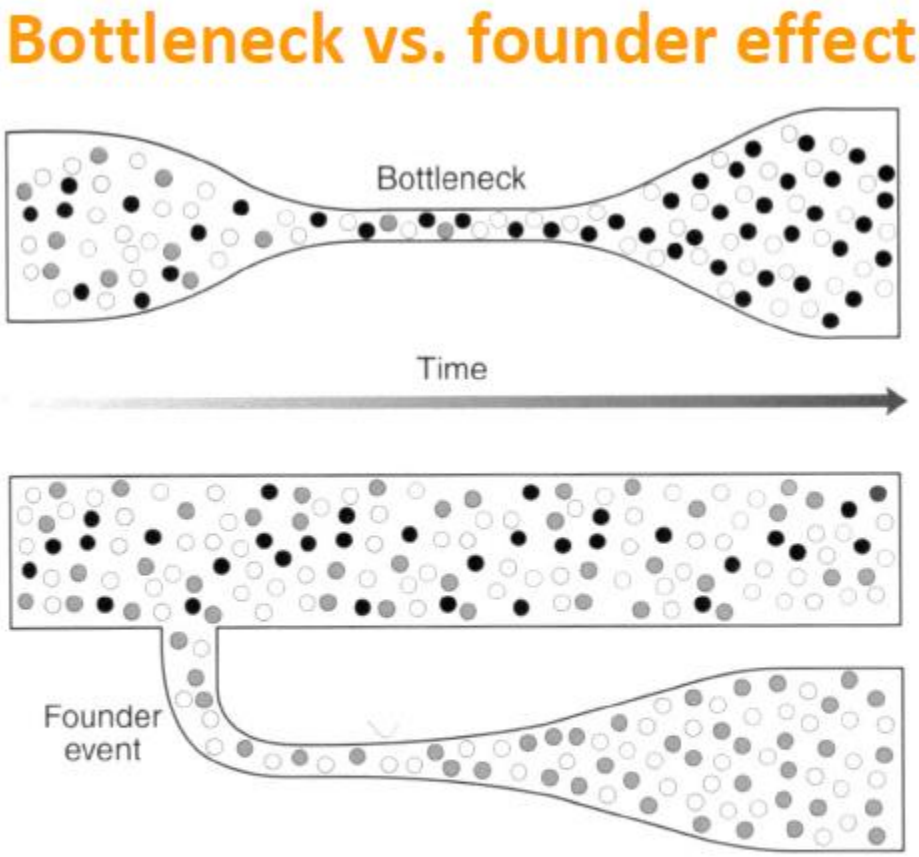
Within populations

1. **Allele frequencies will change over time, by chance alone even in absence of natural selection**
2. These **changes are stronger in small than in large populations**
3. In absence of new mutations:
    a. **All alleles except one will be lost by genetic drift**
    b. The genetic **diversity of the population decreases**
4. Large populations are expected to show more diversity than small populations
5. Inbreeding will increase in the population

Between populations

1. Populations recently separated **will progressively diverge genetically** by chance alone
2. Populations separated for a long time will have distinct sets of alleles (**different alleles will fix in the population**)

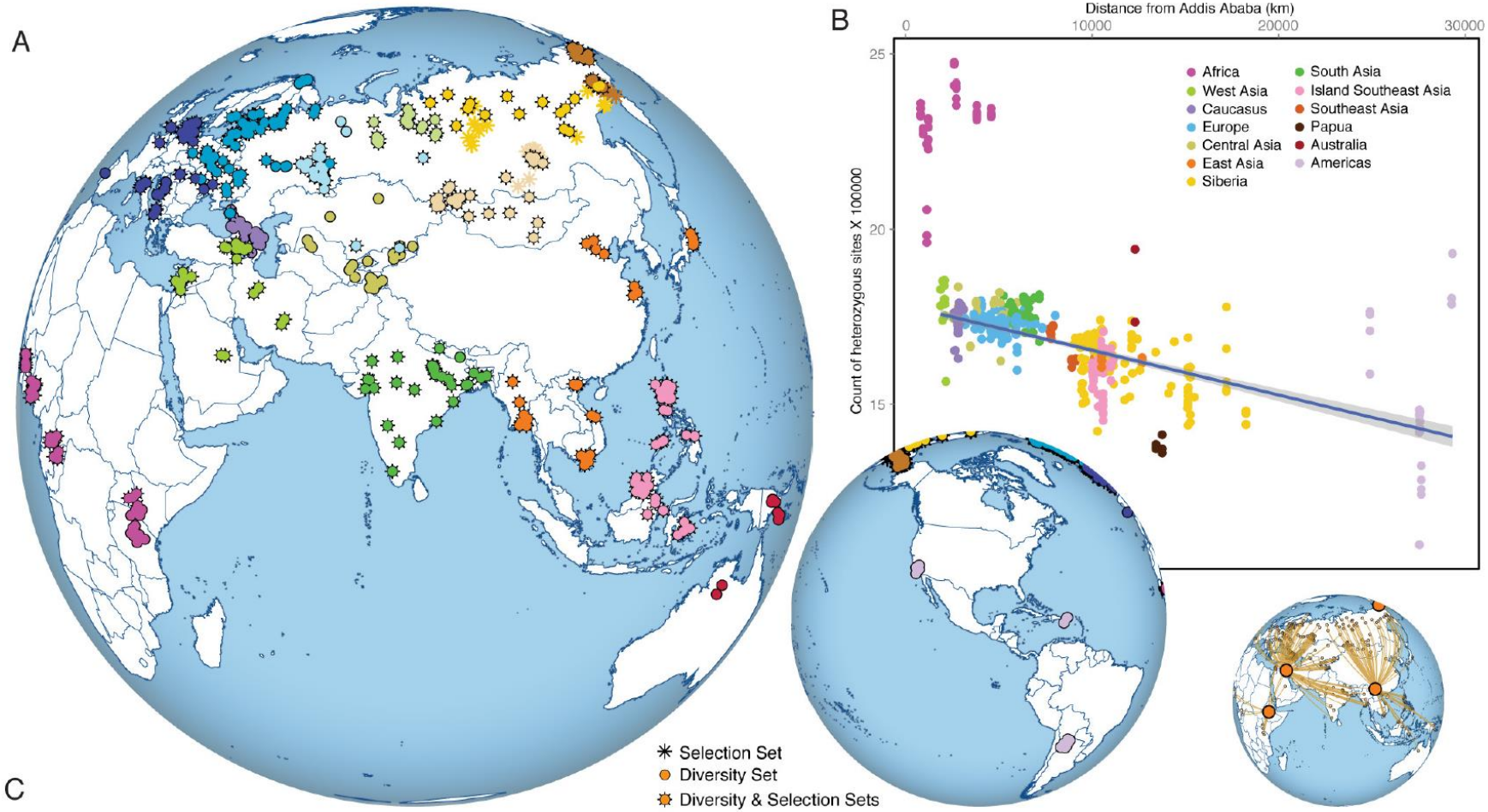
Demographic events and genetic drift



Both phenomena lead to increased changes in allele frequencies: either temporal changes (for bottleneck) or spatial changes (for founder effect)

L. Excoffier

Evidence for modern human range expansion based on heterozygosity



- Decrease in genetic diversity with distance from Africa

Pagani et al (2016)

Definition of effective population size

Effective population size N_e

The size of a Wright-Fisher population having the same pattern of polymorphism, rate of genetic drift, or level of inbreeding as the natural population

- Not equal to the census size *N* (total number of individuals)
- Not necessarily equal to the number of breeding individuals

Different between census and effective size

Census population size (N) The number of individuals in a population; the head count size of a population.

Effective population size (N_e) The size of an ideal Wright–Fisher population that maintains as much genetic variation or experiences as much genetic drift as an actual population regardless of census size.

Hamilton (2009) Population Genetics.

Factors affecting effective size

Effective size is influenced by:

- The number of breeding individuals in a population
- Time fluctuations of the population size (seasonal, climatic change)
- Sex ratio
- Variance in the number of offspring among individuals (polygyny, polyandry, sexual selection)
- Inbreeding
- Overlapping generations
- Level of population subdivision
- Gene flow (migrations)
- Natural selection

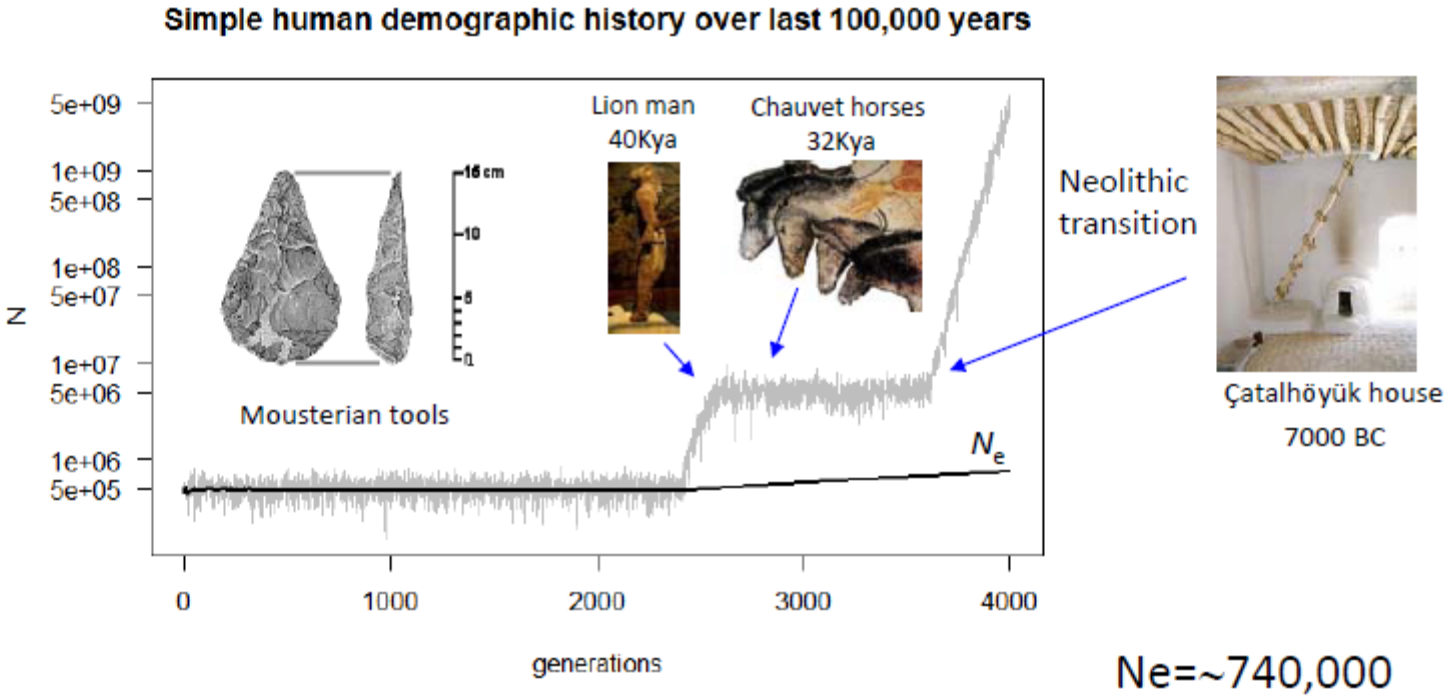
Effective size of the human species

The human species is believed to have increased dramatically in the last 100,000 years.

Very simplistically there could have been two major transitions:

Upper Paleolithic transition: ~500,000 to ~ 5 millions

Neolithic transition: exponential growth until now (7 billions in 2013)



Ne=~740,000

What would be the effective size of the human population?

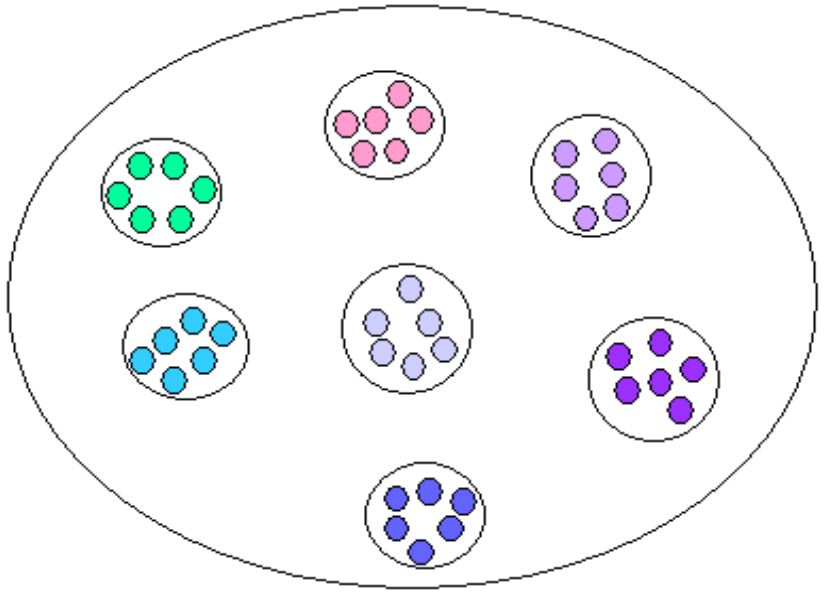
In reality the human effective size is around 10,000! Why?

Laurent Excoffier

Population subdivision

Natural populations are usually subdivided

- Discontinuous habitats:
  - Ponds, lakes, trees
  - Resource limitations
  - Host-parasite systems
  - Seasonality …
- Behavior (social or mating system)



Subdivisions (demes) can maintain genetic cohesion through an exchange of reproducing migrants

Pattern of migration and timing of separation between demes will have a profound effect on the degree of differentiation between demes

Population differentiation (F_{ST})

F_{ST} is a measure of the degree of genetic differentiation among demes

It was originally defined by Sewall Wright (1943) as

$$F_{ST} = \frac{\mathrm{var}(p)}{\bar{p}(1-\bar{p})} = \frac{\mathrm{var}(p)}{p_0(1-p_0)} \qquad \text{(since } \bar{p} = p_0\text{)}$$

It is therefore a standardized measure of the variance in allele frequencies among populations

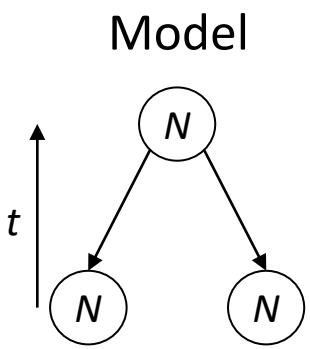
The notation *F* has been chosen, because it is a measure of where populations are in their ***F***ixation process due to genetic drift.

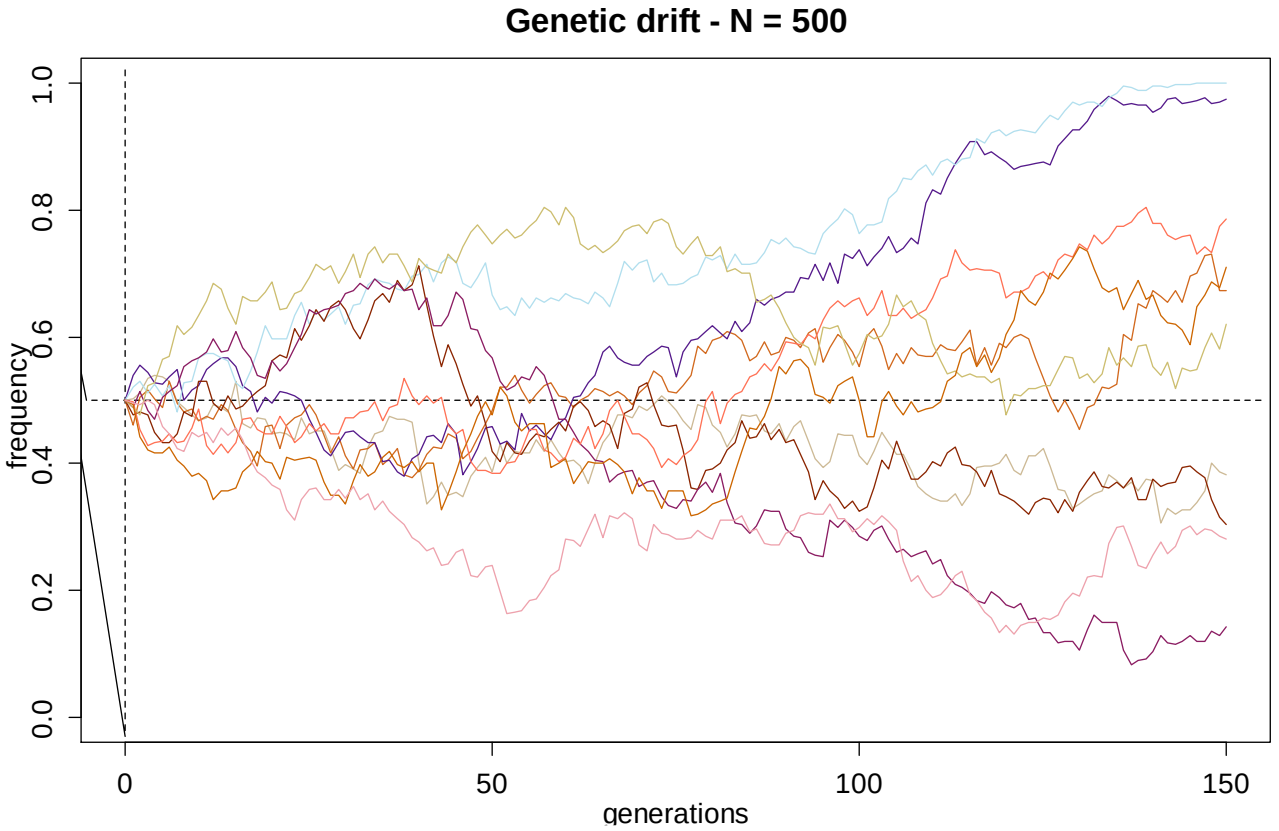
It is often used as a measure of the **genetic distance between populations**

F_{ST} =0, implies that two or more populations have identical allele frequencies

F_{ST} =1, implies that two or more populations have fixed different alleles

Variance of gene frequencies increases over time between diverged populations



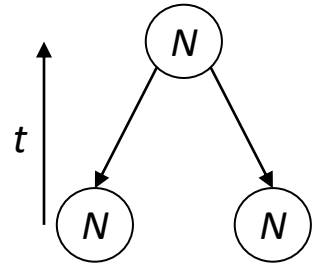


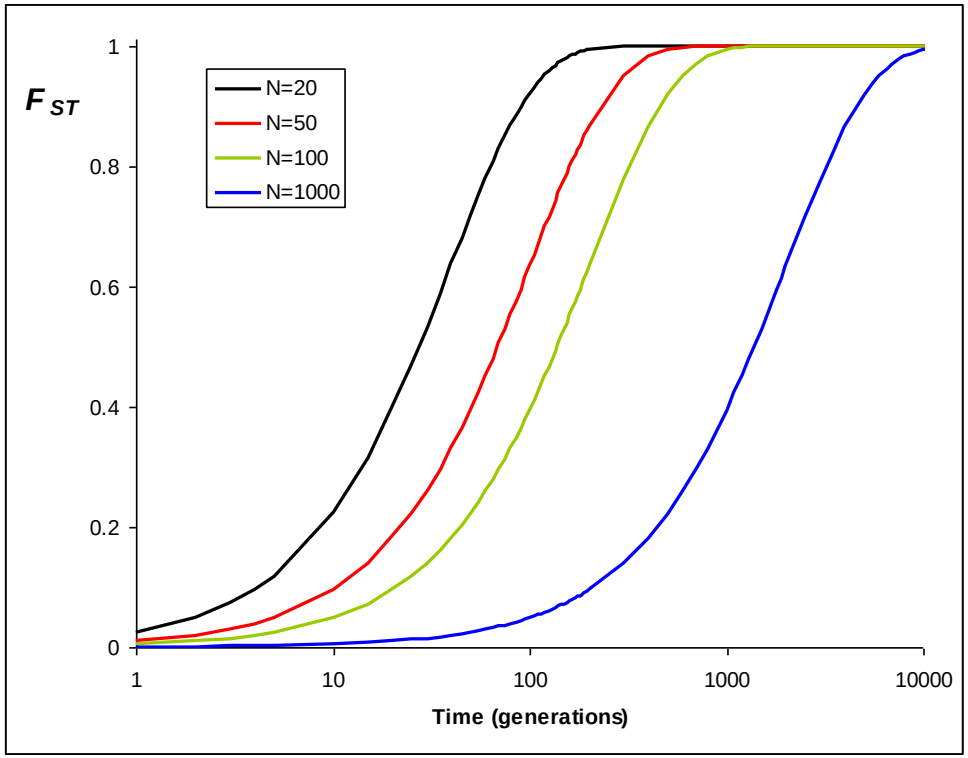
$$\mathrm{var}\left[p(t)\right] = p_0(1-p_0)\left[1-\left(1-\frac{1}{2N}\right)^t\right]$$

Laurent Excoffier

F_{ST} under population divergence

When measured between a pair of populations separated some *t* generations ago, F_{ST} increases over time as





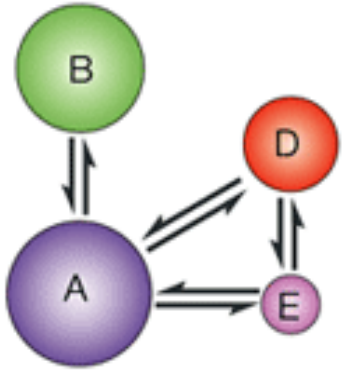
$$F_{ST} = 1 - \left(1 - \frac{1}{2N}\right)^{t}$$

For the same divergence time, two large populations will be more similar than two small populations

Models of migration

Quite generally, migration between populations:

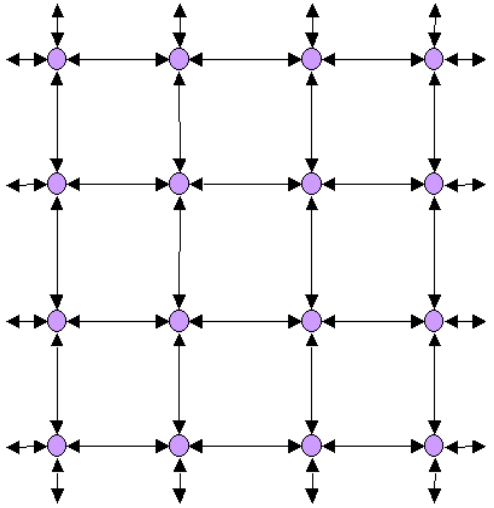
- Maintains genetic similarity
- Changes allele frequencies



Spatially explicit models:
Geographically close populations exchange more migrants

1-D stepping-stone model

2-D stepping-stone model



Laurent Excoffier

Evolutionary processes

Genomic processes

- Mutation
- Recombination

Demography

- Effective population sizes
- Population split times
- Migration rates

Selection

Natural selection:

- Beneficial mutations involved in adaptation
- Deleterious mutations with negative effects

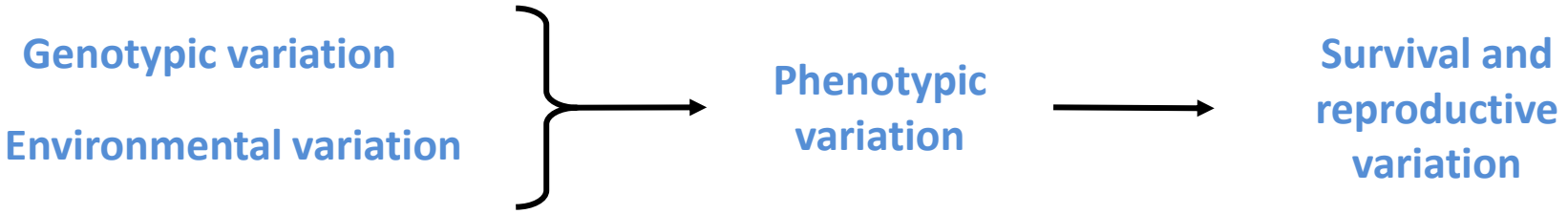
What is the effect of natural selection on genetic diversity patterns?

Natural selection and adaptation

I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection

Charles Darwin, The Origin of Species (1859)

Natural selection is the **differential reproductive success** of different **phenotypes** resulting from the **interaction of the organisms with their environment**



Natural selection and adaptation

Mechanism of natural selection

- Natural selection arises from the **competition** between the relative capacity of survival (**viability**) and reproduction (**fecundity**) of different organisms **in a specific environment**

Consequence of natural selection

- Natural selection **changes the relative frequencies of alleles in the population**
- Natural selection leads to a **better adaptation of individuals to their environment**

Concept of fitness

Selection is often quantified by considering the **fitness** of individuals

Darwinian fitness (*W*): Measures the **ability of individuals** to survive and reproduce and thus **to transmit their genes to the next generation**

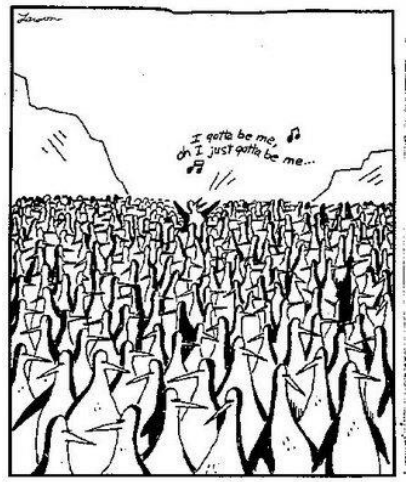
Simplified view

$$W_i = V_i \times f_i$$

Mean number of descendants in the next generation

Probability to survive to adulthood

Mean number of offspring per adult



Can we use genomic data from natural populations to find genes under positive selection, involved in adaptation?

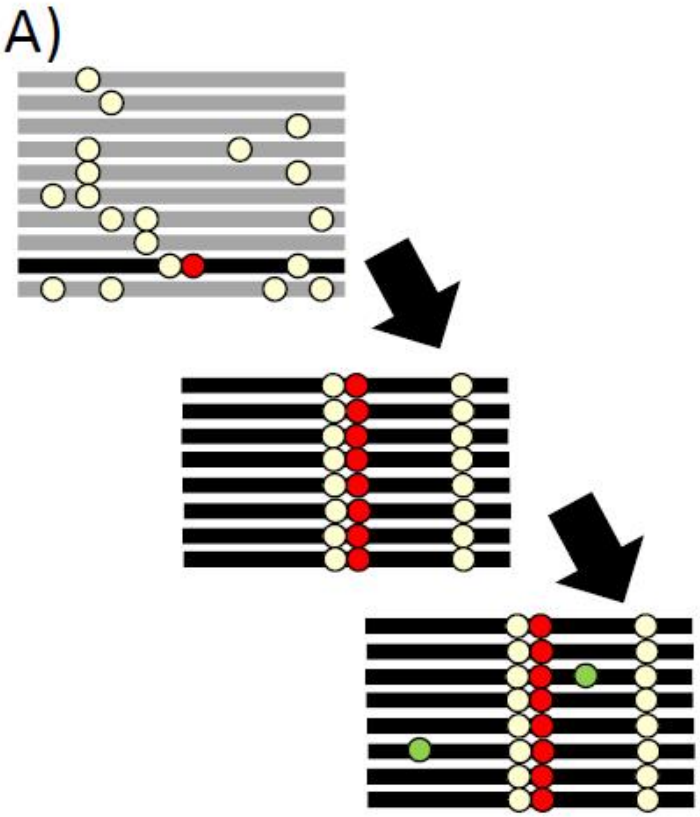
- What are the genes/mutations involved in resistance to antibiotic?

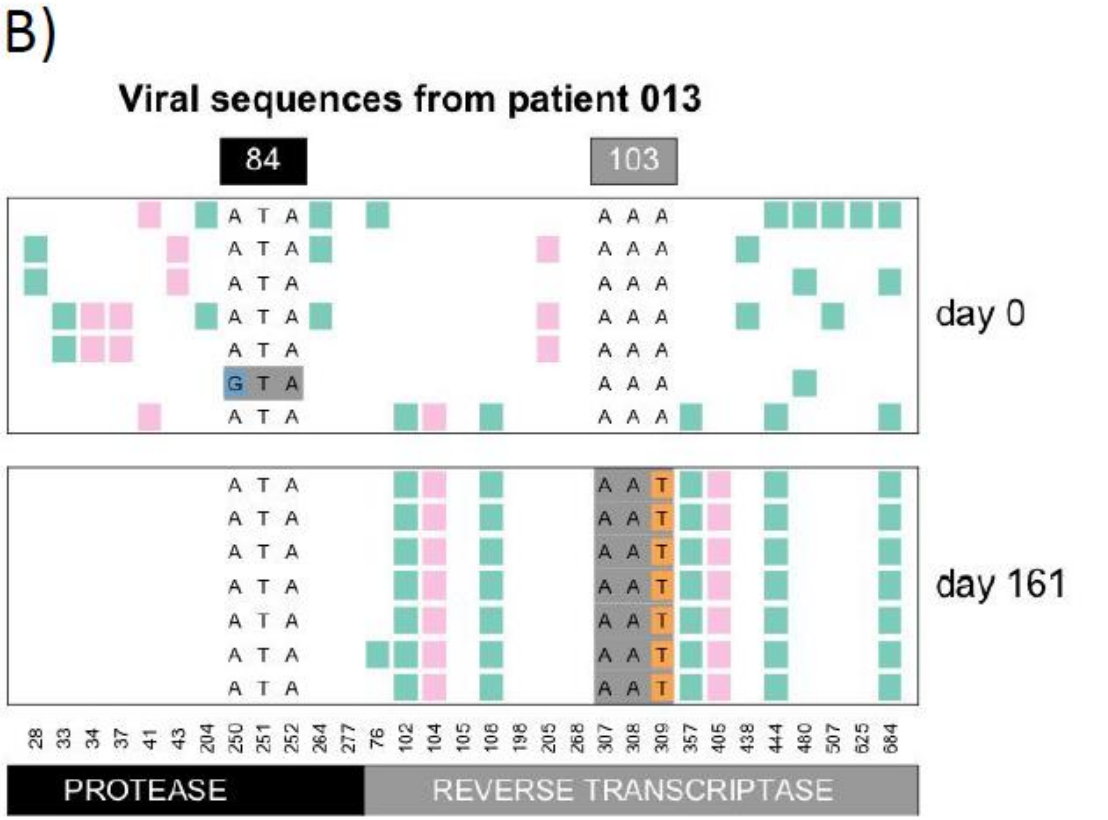
Can we use genomic data from natural populations to find mutations with a deleterious effect?

- What are the mutations associated with susceptibility to disease?

What are the **expected genomic patterns** when there is selection?

Example HIV





Patterns of HIV genetic diversity in samples taken from patient before and after taking a retrotransposase inhibitor (Wlliams and Pennings, 2019)

Relative fitness

Haploid model

1 locus, two alleles A_1 and A_2

It is simpler to express the change of frequencies in terms of **relative fitness** w_1 instead of the absolute fitness W_1

	Genotypes	
	A_1	A_2
Absolute fitness	W_1	W_2
Relative fitness	$w_1 = W_1 / W_1$	$w_2 = W_2 / W_1$
Relative fitness	$w_1 = 1$	$w_2 = 1 - s$

Phenotype A_1 is taken arbitrarily as a reference

***s* : selection coefficient**
(measures the difference in fitness relative to the reference genotype)

As with absolute fitness $p' = p\frac{w_1}{\bar{w}}$ if $w_1 > \bar{w} \Rightarrow p \nearrow$ if $w_1 < \bar{w} \Rightarrow p \searrow$

Deterministic diploid selection model (3)

Relative fitness for different types of selection

	A_1A_1		A_1A_2		A_2A_2
Fitness	w_{11}		w_{12}		w_{22}
Directional selection					
Recessive deleterious A_2	1	$=$	1	$>$	1-s
Dominant deleterious A_2	1	$>$	1-s	$=$	1-s
Dominant advantageous A_1	1	$=$	1	$>$	1-s
General dominance	1	$>$	1-hs	$>$	1-s
Heterozygote advantage	$1 - s_1$	$<$	1	$>$	$1 - s_2$
Heterozygote disadvantage	$1 + s_1$	$>$	1	$<$	$1 + s_2$

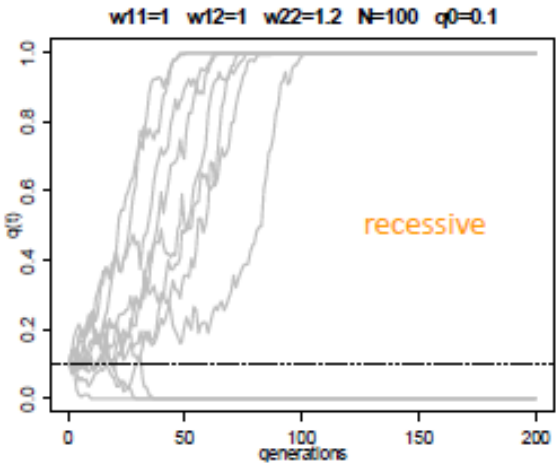
***s*: selection coefficient** . Usually, $(0 \le s \le 1)$ but this is **rather arbitrary**
It is just a convenient way to express the fact that the relative fitness of a given genotype differs from 1

***h*: level of dominance between alleles** $(0 \le h \le 1)$

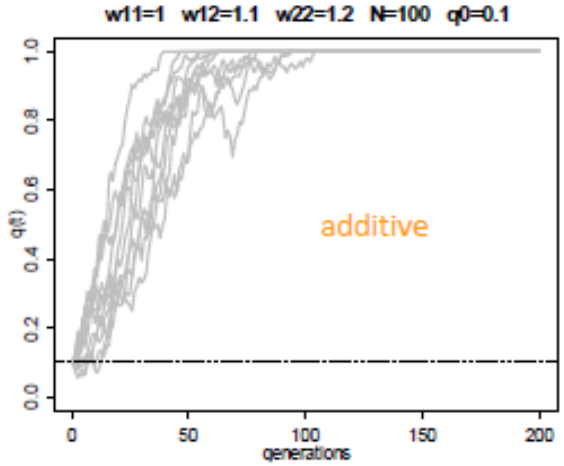
***hs*: measures the amount of selection on heterozygotes**

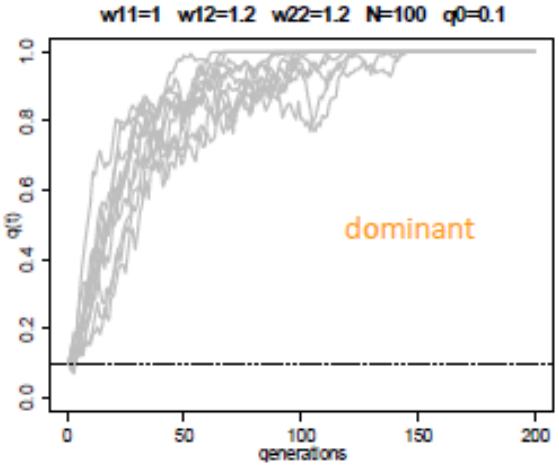
Selection in finite populations

In finite populations, genetic drift will interact with selection



10 replicates of directional selection with s=0.2 q_0=0.1



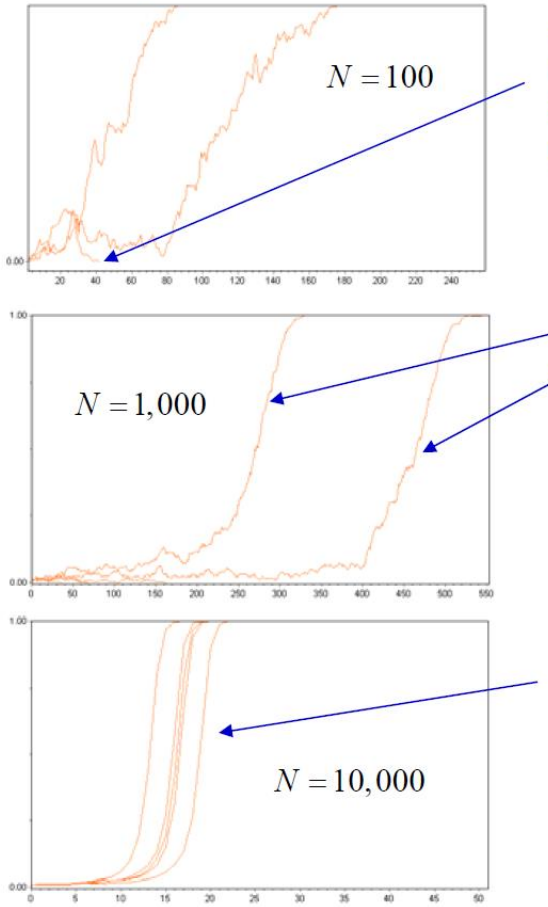


Laurent Excoffier

Selection in finite populations

Directional selection

$w_{AA} = 1 \quad w_{Aa} = 1 \quad w_{aa} = 1.1 \quad p_0 = 0.01$



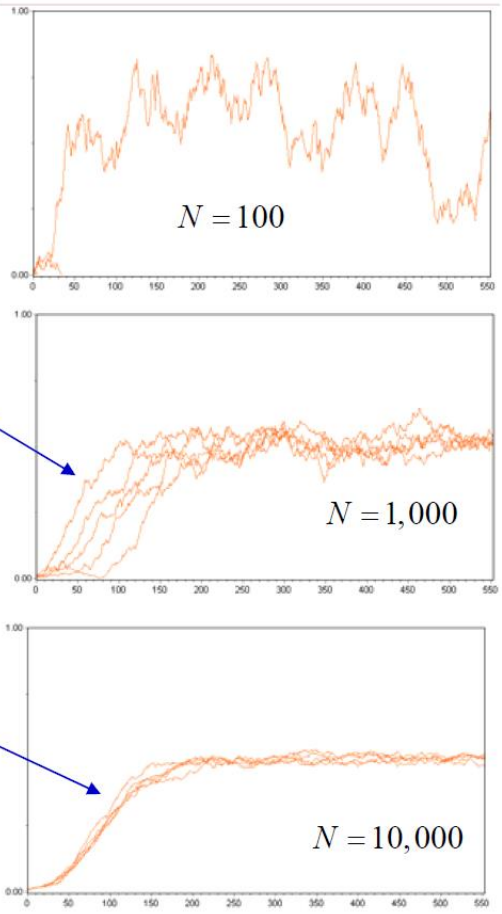
Favorably selected mutants can disappear due to genetic drift

There is a large variability in the dynamics of allele frequency change, even in medium-sized populations

Only in very large populations do we observe an almost deterministic dynamics

Heterozygote advantage

$w_{AA} = 1 \quad w_{Aa} = 1.05 \quad w_{aa} = 1 \quad p_0 = 0.01$



An example of directional selection: Industrial melanism

The British peppered moth (*Biston betularia*)

Wild type

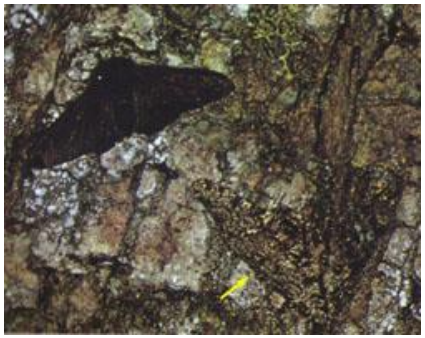
Common

Melanic form

Now very rare

Melanic and wild-type form on two different birch trees

Natural tree color

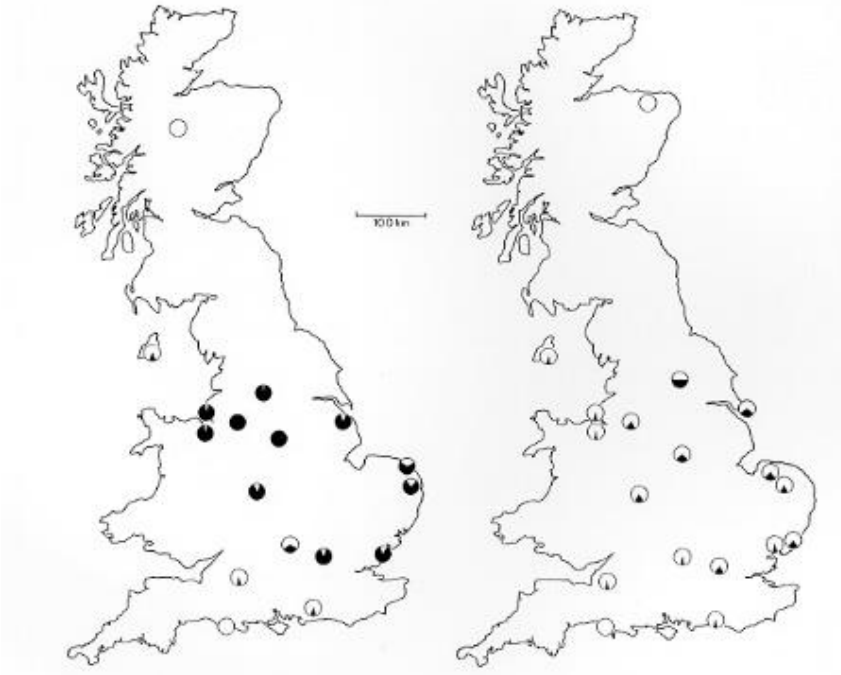


Polluted tree

Viability selection

1956 — Coal pollution

1996 — Clean air



Percentage of melanic form has dramatically changed in 40 years

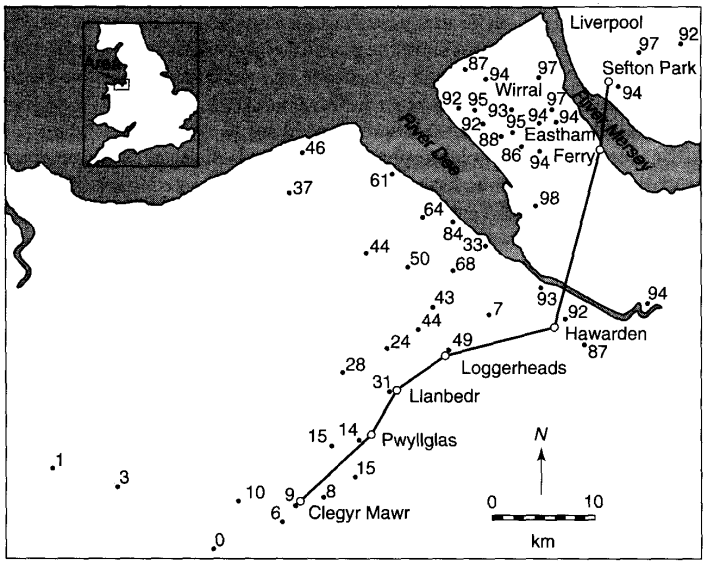
Laurent Excoffier

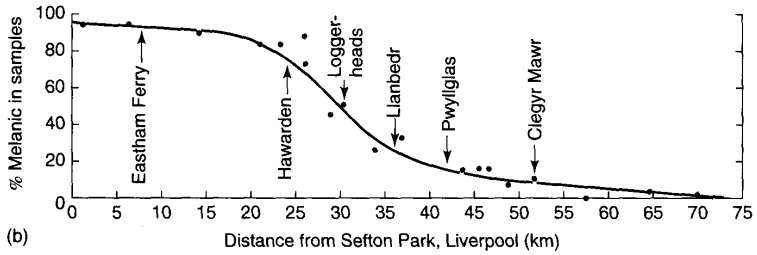
An example of directional selection: Industrial melanism

The British peppered moth (*Biston betularia*)

	Genotypes		
	Melanic form		*Wild-type*
	MM	*Mm*	*mm*
Relative fitness	w_{MM}	w_{Mm}	w_{mm}
Pollution	1 =	1 >	$1 - s_1$
Clean air	$1 - s_2$ =	$1 - s_2$ <	1

Percentage of melanic form around Liverpool in the 60's

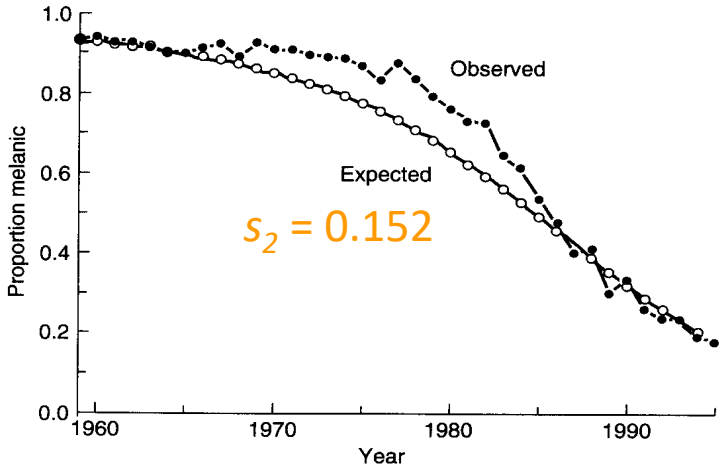




Industrial area

Countryside

Evolution of melanic form frequencies after 1960



Laurent Excoffier

The effects of selection can be seen in experimental evolution

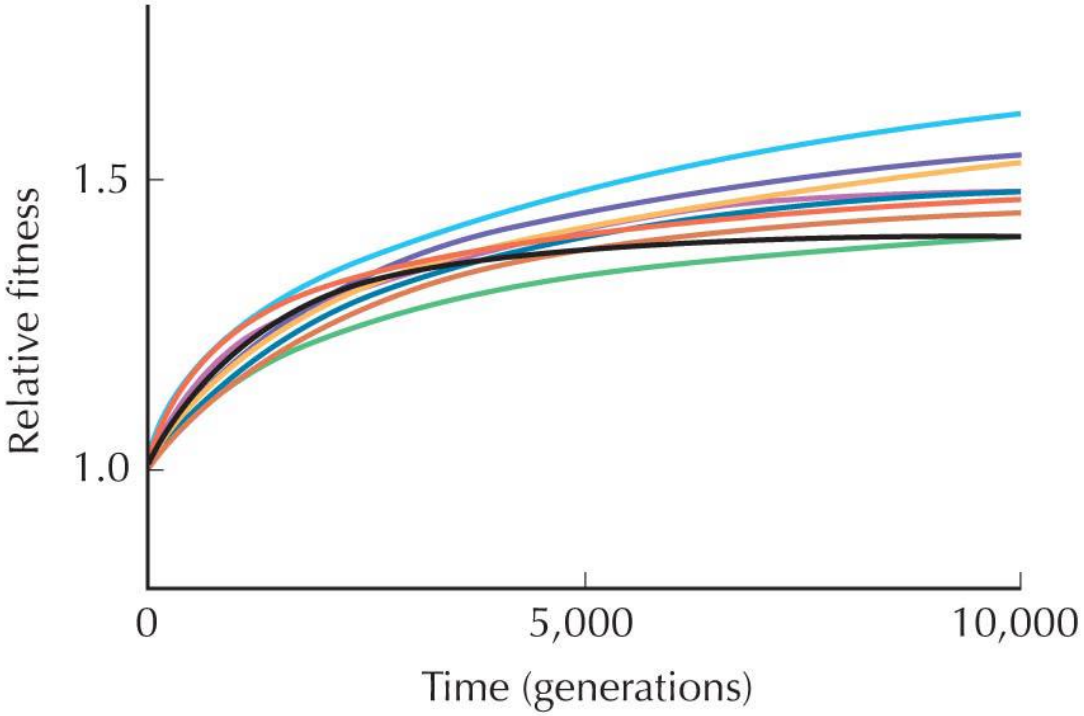
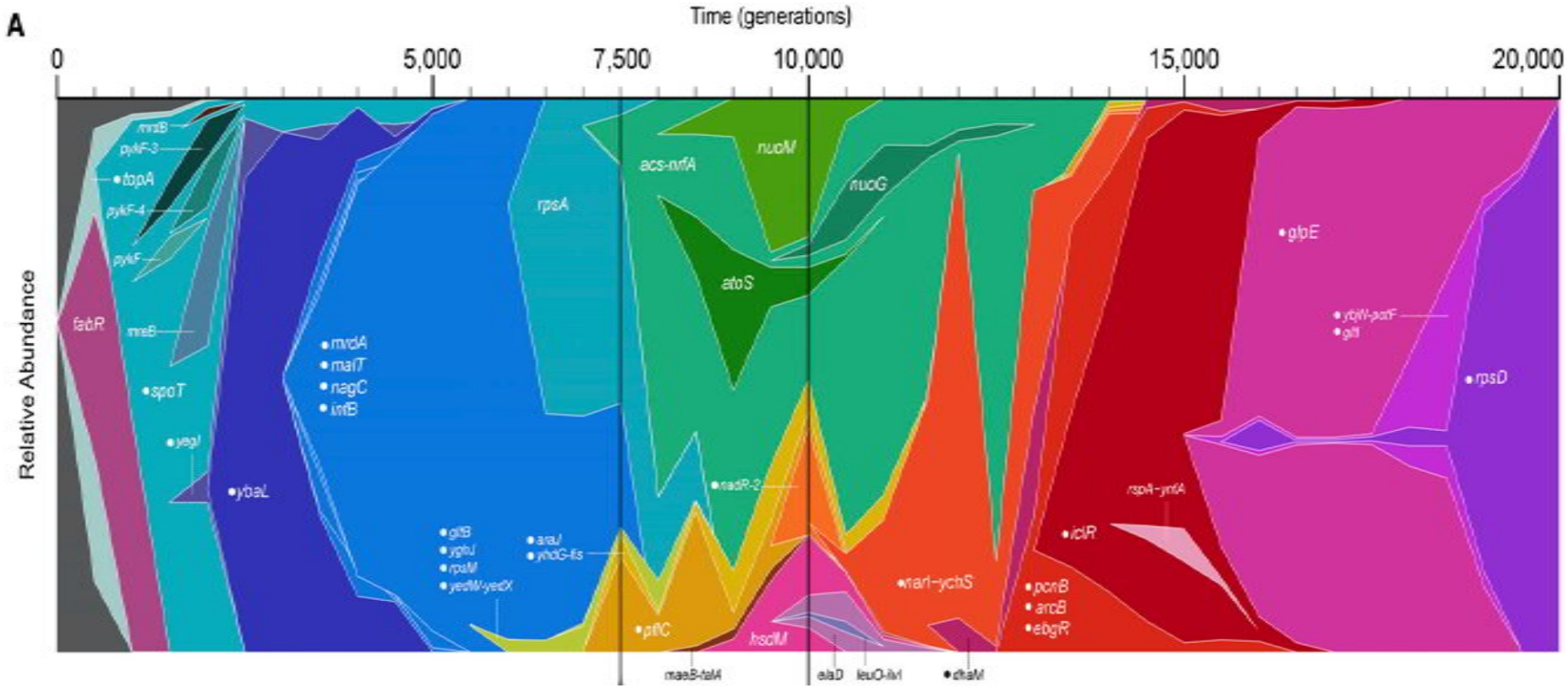


FIGURE 17.32. Average fitness of nine replicate *Escherichia coli* populations relative to the ancestral strain during 10,000 generations of adaptation to a glucose-limited medium.

17.32, redrawn from Lenski R.E. et al., *Proc. Natl. Acad. Sci.* **91:** 6608–6618, © 1994 National Academy of Sciences, U.S.A.

R. Lenski long term experimental evolution of *E. coli*



Maddamsetti et al (2015) Genetics

Different forms of selection

At the genetic level, we can think as if alleles are:
- neutral
- positively selected (beneficial
- negatively selected (deleterious)

This can be described by the **distribution of fitness effects (DFE).**

Note that the effect of an allele depends on the environment and genetic background!

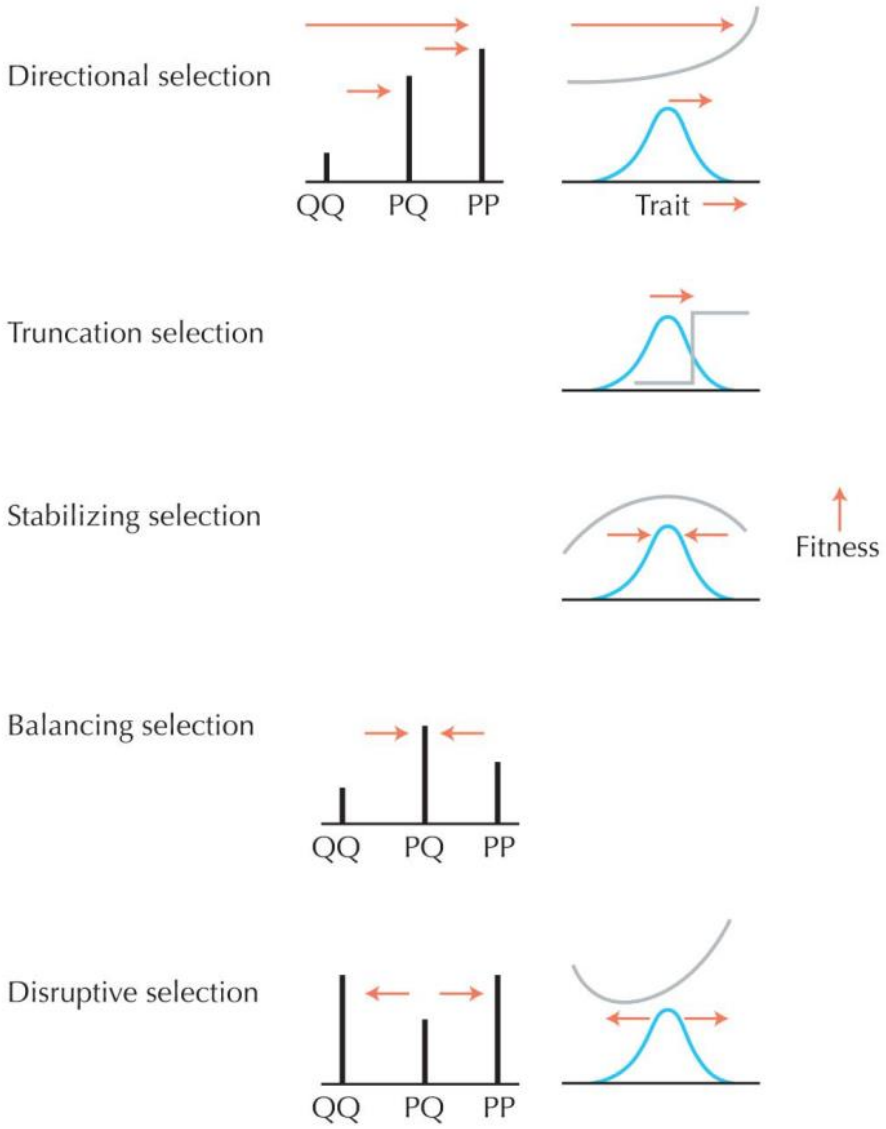


FIGURE BOX 17.2. Modes of Selection

Selection coefficients of new mutations

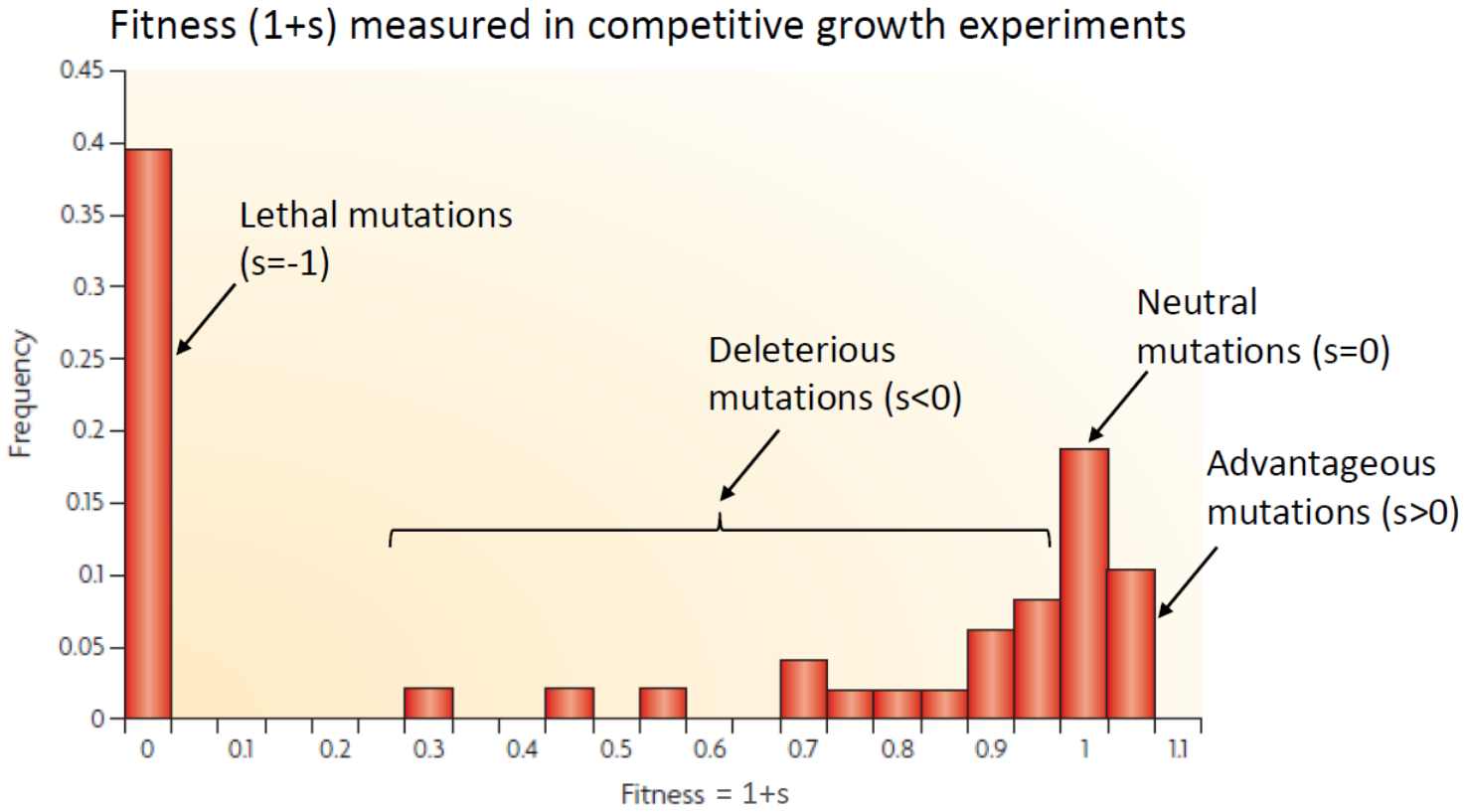


Figure 1 | **The distribution of fitness effects of random mutations in vesicular stomatitis virus.** In this experiment, random mutations were introduced into the virus, and the fitnesses of the mutants were compared against the unmutated wild type.

Eyre-Walker and Keyghtley, 2007 NRG

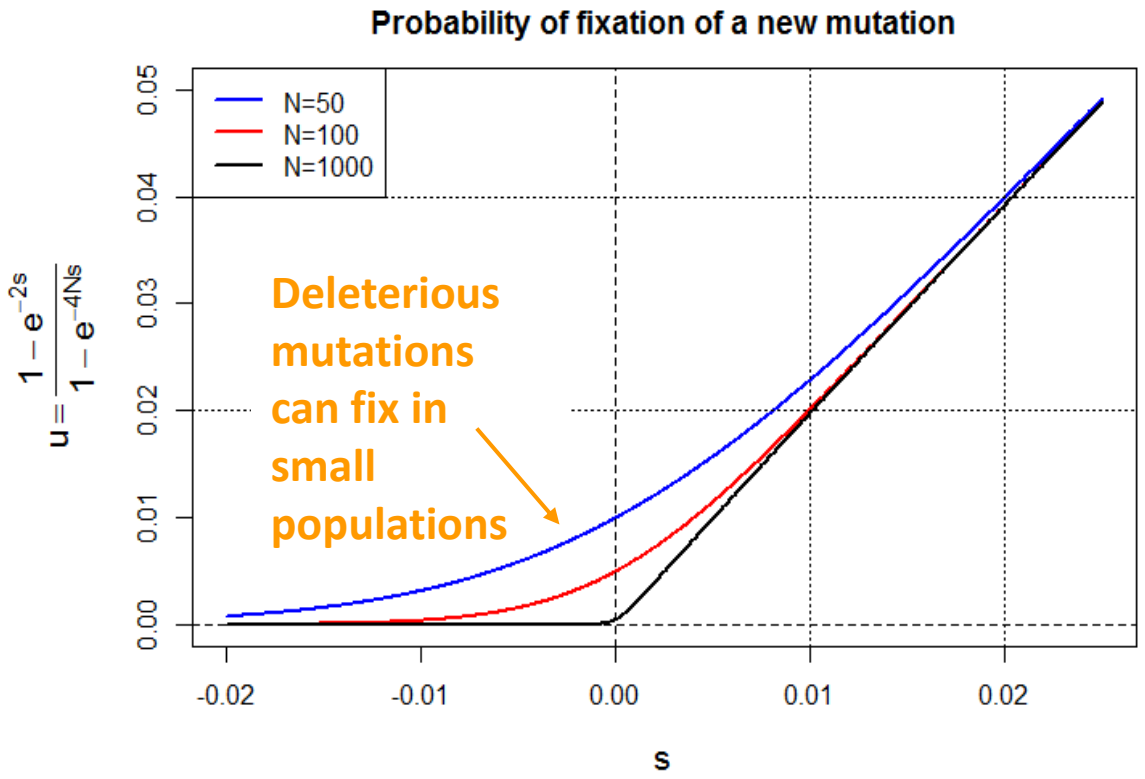
What do the relative frequencies of the different mutations tell you?

Laurent Excoffier

Fate of a new selected mutation

Kimura (1962) that the probability of fixation of a new (additive) selected mutant in a diploid population of size *N* was

$$P_{fixation}\left(\frac{1}{2N}\right) = \frac{1-e^{-2s}}{1-e^{-4Ns}}$$



Fate of a new selected mutation

For mildly beneficial mutations ($s<1/2$) and **in large populations** (***Ns*>>1**), then the probability of fixation simplifies to

$$P_{fixation}(\frac{1}{2N}) \approx \frac{1-(1-2s)}{1-e^{-4Ns}} \approx 2s \qquad \left(e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots\right)$$

It implies that:

1) not all beneficial mutations will go to fixations
2) the vast majority of beneficial mutations will be quickly lost by drift

But **in small populations** where $4Ns<1$, the probability of fixation simplifies to

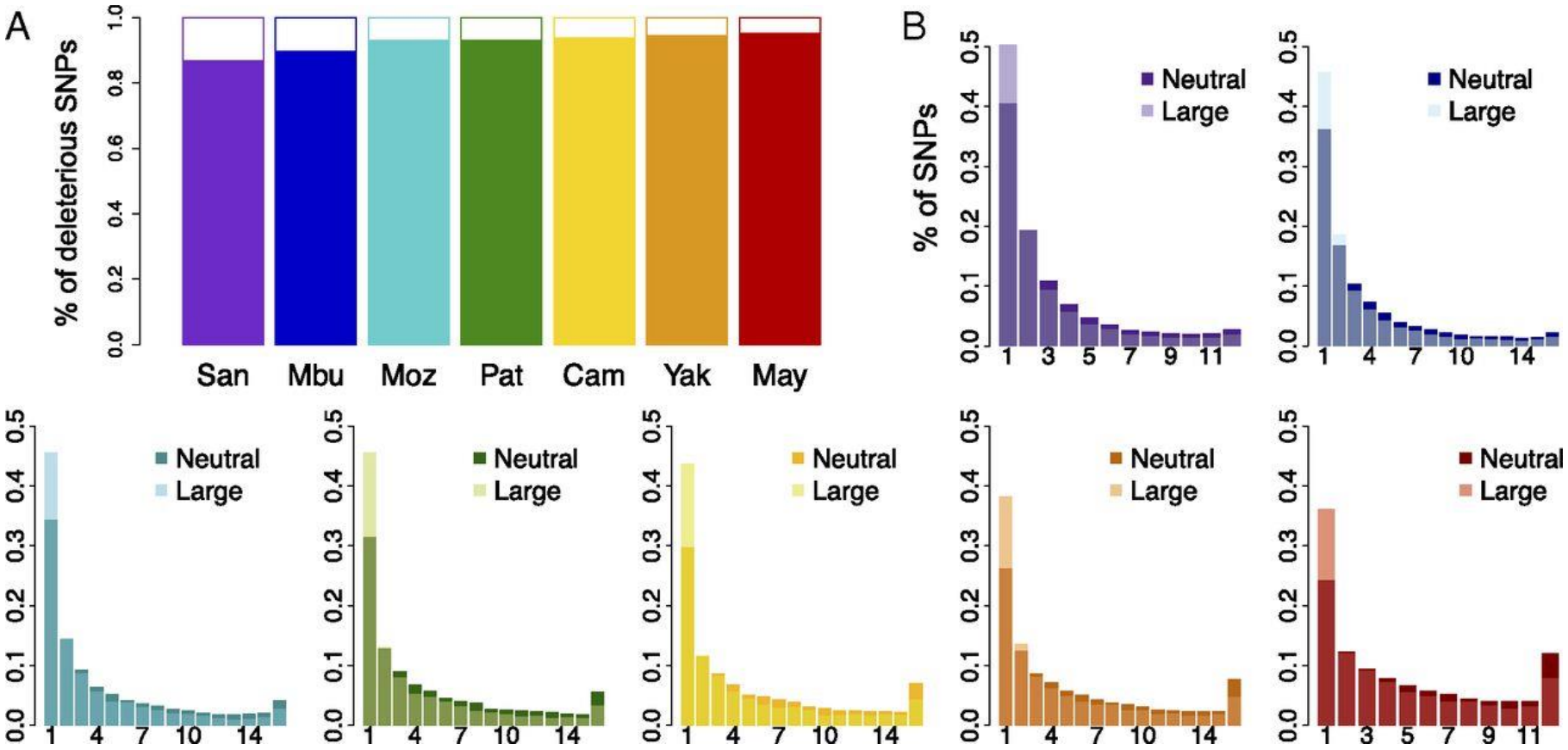
$$P_{fixation}(\frac{1}{2N}) \approx \frac{2s}{1-(1-4Ns)} \approx \frac{1}{2N}$$

like for neutral mutations, showing that **genetic drift becomes stronger than selection** for these beneficial mutations.

The **same mutation can be selected in a large population (if *4Ns*>1)** and **neutral in a small population (if *4Ns*<1)**

Interaction of demography with selection

Brenna M. Henn et al. PNAS 2016;113:E440-E449



Proportion of **deleterious mutations** in human populations is higher in smaller populations.

Recall the gray zone where drift is stronger than selection.

If one simply cannot measure the state variables or the parameters with which the theory is constructed, or if their measurement is so laden with error that no discrimination between alternative hypotheses is possible, the theory becomes a vacuous exercise in formal logic that has no points of contact with the contingent world. The theory explains nothing because it explains everything. It is my contention that a good deal of the structure of evolutionary genetics comes perilously close to being of this sort.

Lewontin (1974) The genetic basis of evolutionary change