Human Genetic Variation

Section 3

(1.5 hours)

Learning objectives

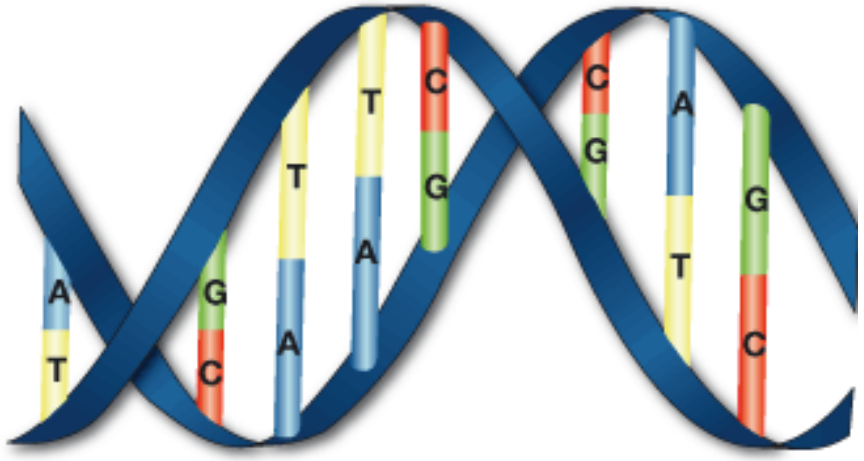
- Describe differences in types of genetic variation and how they affect phenotypes.
- Understand how variation perpetrates through generations.
- Calculate linkage disequilibrium between variants.

Our Genome in Numbers

23 chromosome pairs

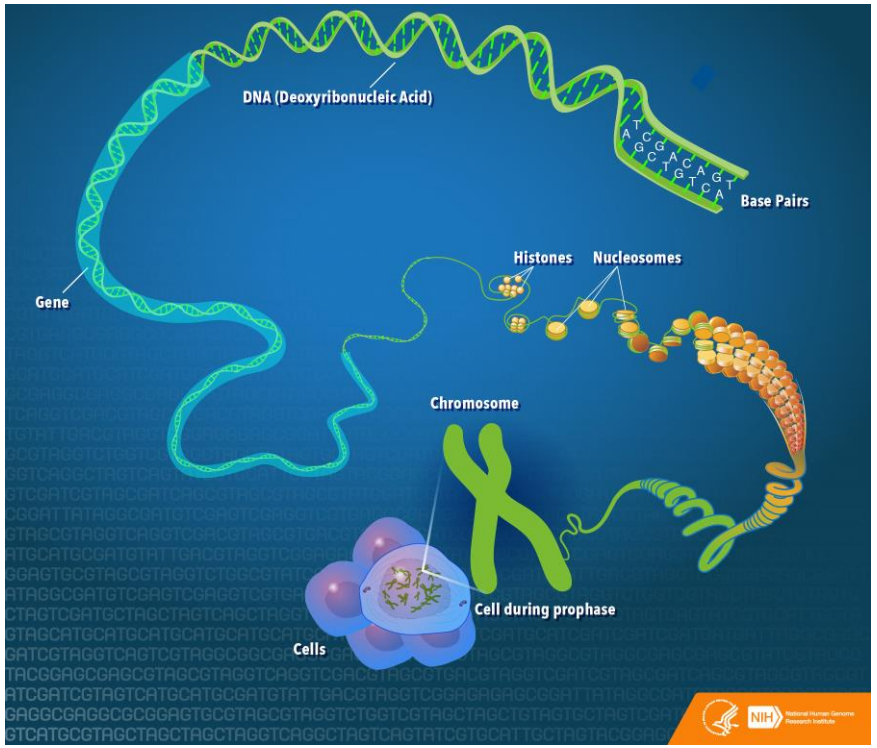
3.2 billion base-pairs (A,C,G,T)

~20,000 genes



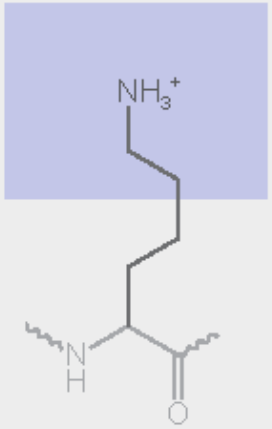
Thymine (Yellow) = T Guanine (Green) = G

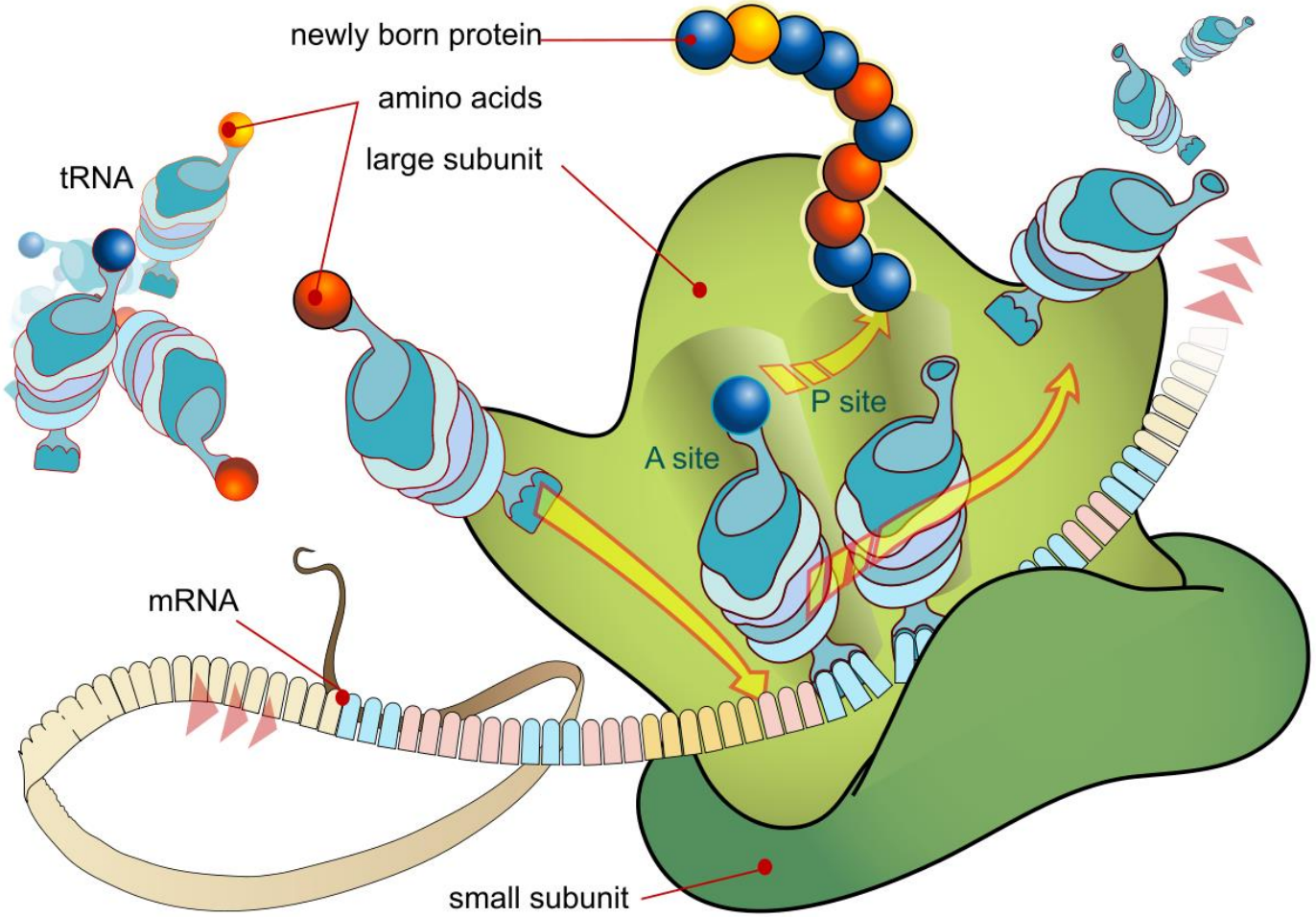
Adenine (Blue) = A Cytosine (Red) = C



Genotypes - phenotypes

DNA level	TTC
mRNA level	AAG
protein level	**Lys**





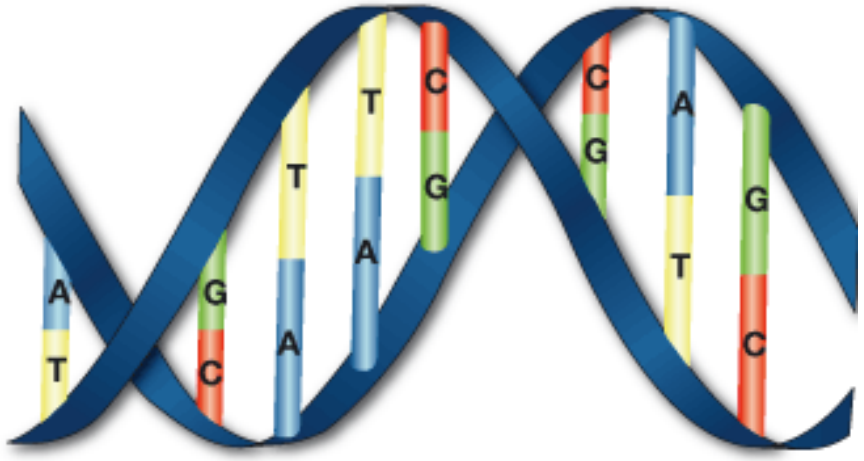
Our Genome in Numbers

23 chromosome pairs

3.2 billion base-pairs (A,C,G,T)

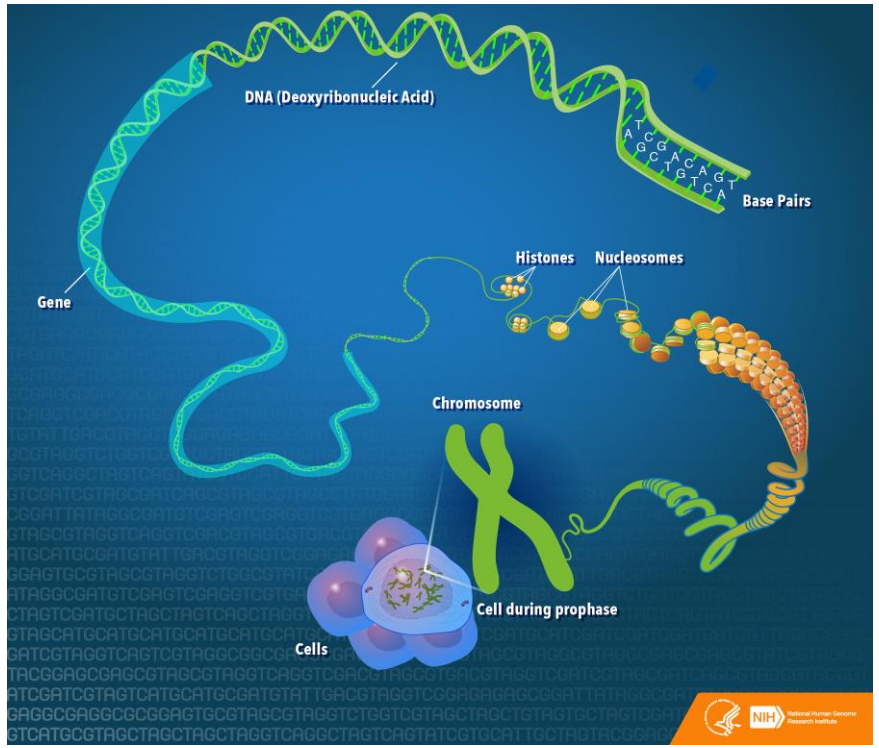
~20,000 genes

~1.5% of the genome is coding DNA



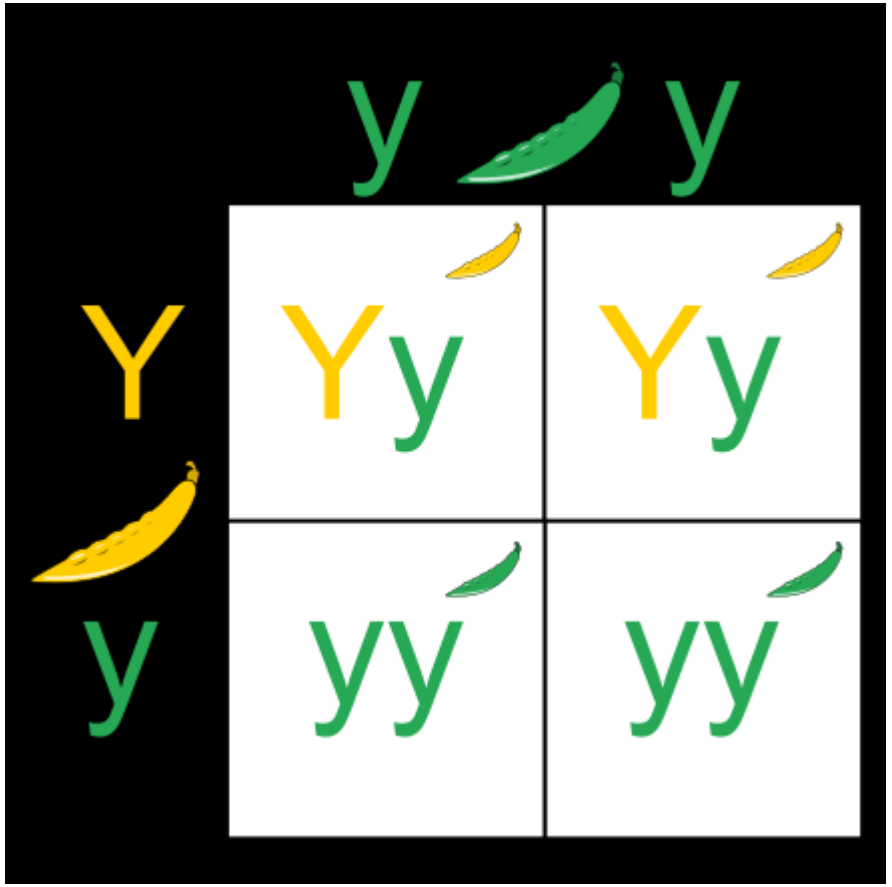
Thymine (Yellow) = T Guanine (Green) = G

Adenine (Blue) = A Cytosine (Red) = C



Genotypes and Phenotypes

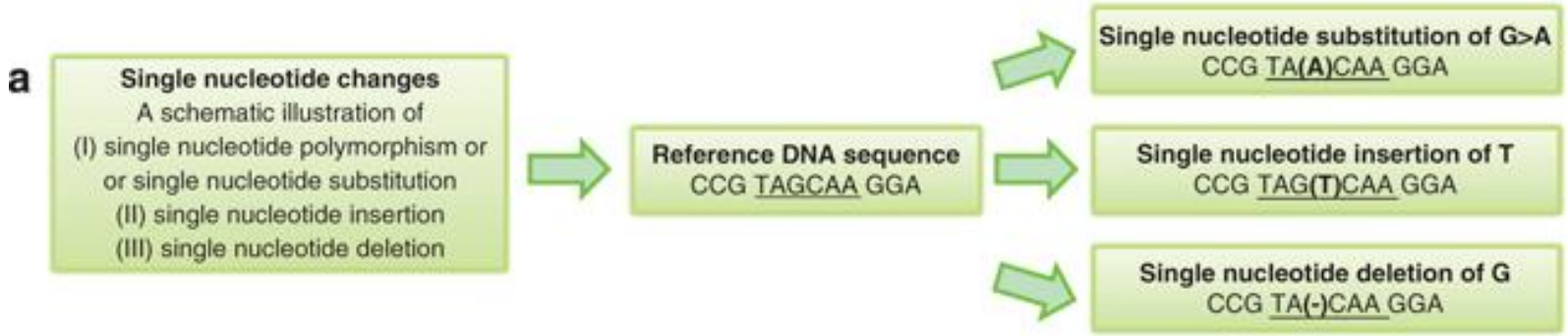
- **Mendelian phenotype** is one driven by variation at a single genetic locus.
- **Complex phenotype** does not show such simple patterns of inheritance.
  - oligogenic (a few genetic loci)
  - polygenic (many genetic loci)

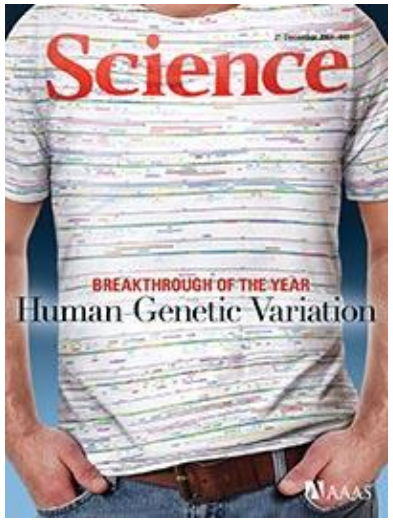


Genotypes and Phenotypes

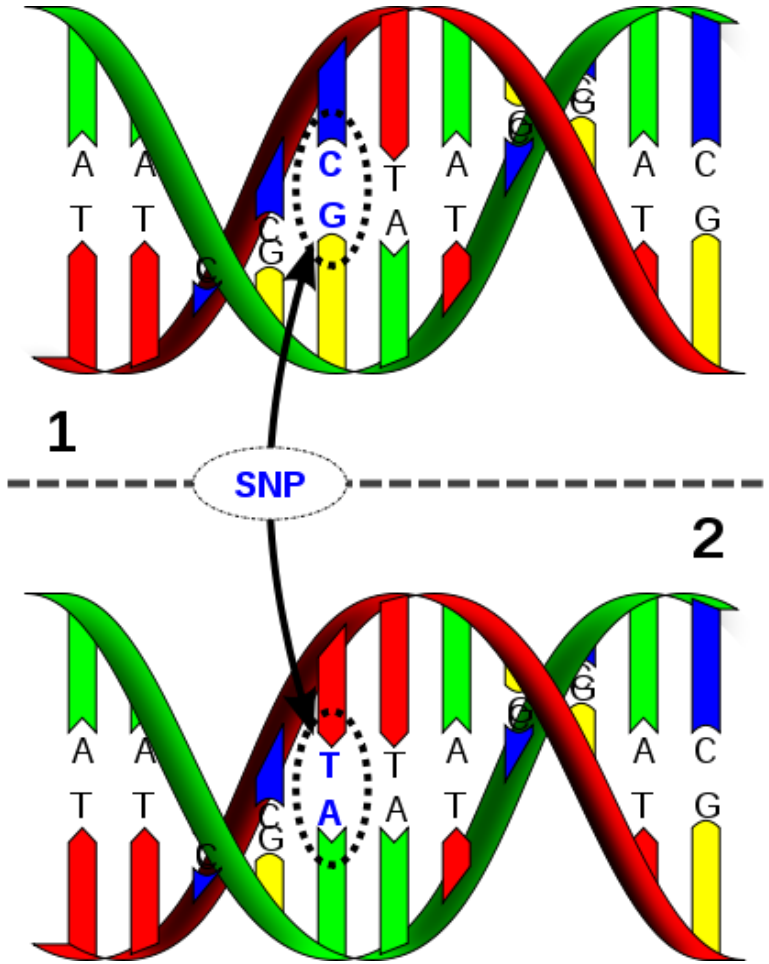
- **Mendelian phenotype** is one driven by variation at a single genetic locus.
- **Complex phenotype** does not show such simple patterns of inheritance.
  - oligogenic (a few genetic loci)
  - polygenic (many genetic loci)
- Binary outcomes (yes/no, i.e. disease status)
- Quantitative outcomes (continuous, i.e. height)

Genetic Variation – sequence variation

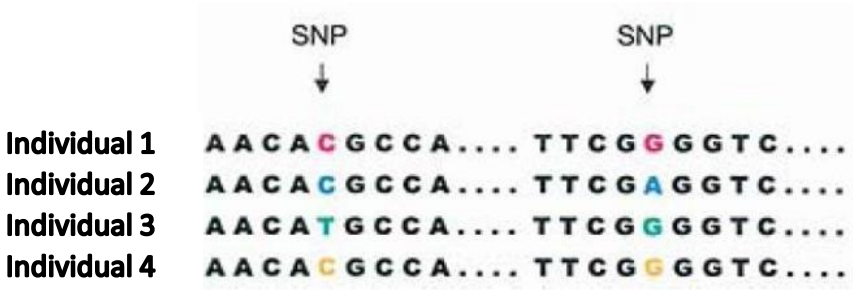


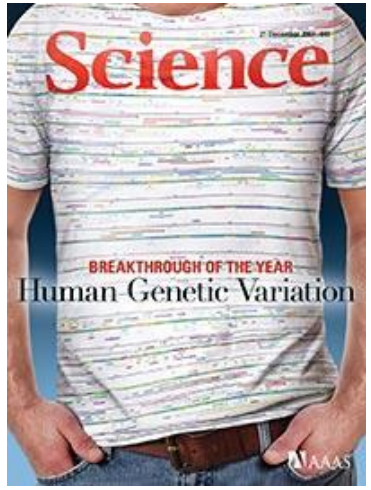


Genetic variation - SNPs



Single Nucleotide Polymorphism (SNP)





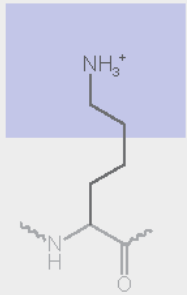
A recent study sequenced 2,504 individuals and identified 84.7 million SNPs

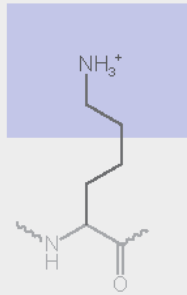
On average, each individual carried 3.5-4.3 million SNPs. 21,400-26,000 (~0.6%) of those are in coding regions (cf. 1.5% of coding DNA in the genome)

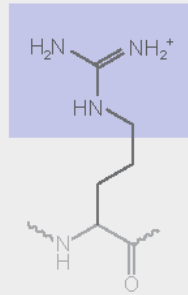
1000 Genomes Project, *Nature 2015*

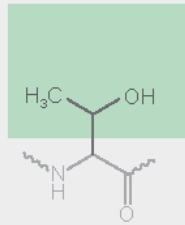
Genetic variation – SNP effects

	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	**Lys**	**Lys**	**STOP**	**Arg**	**Thr**



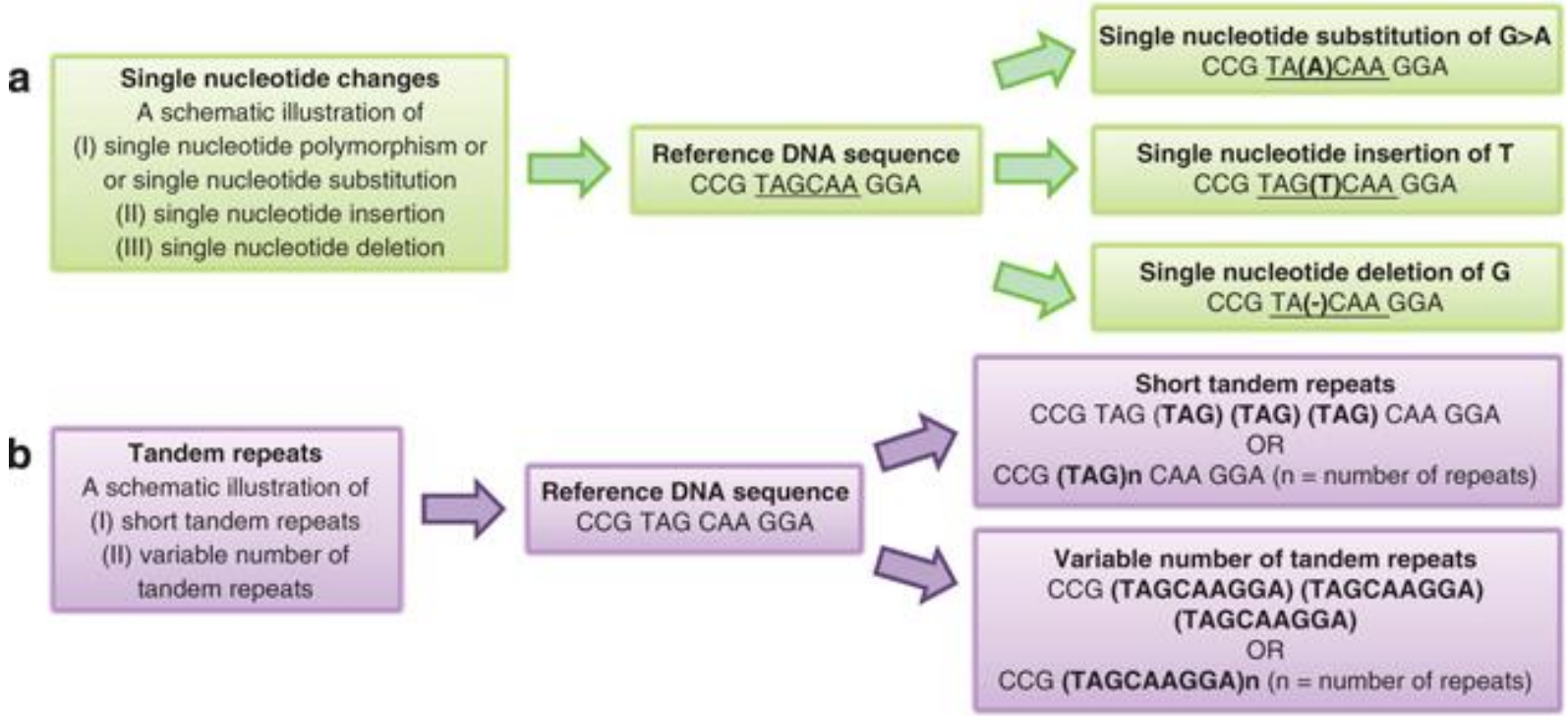


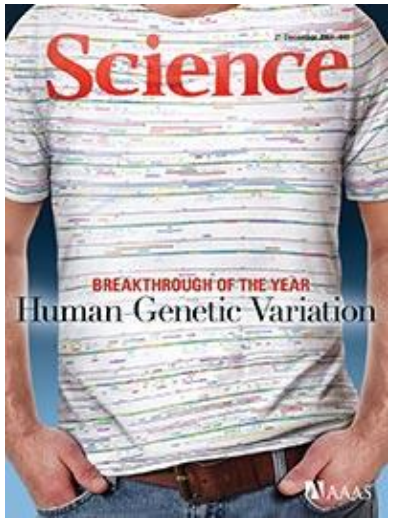




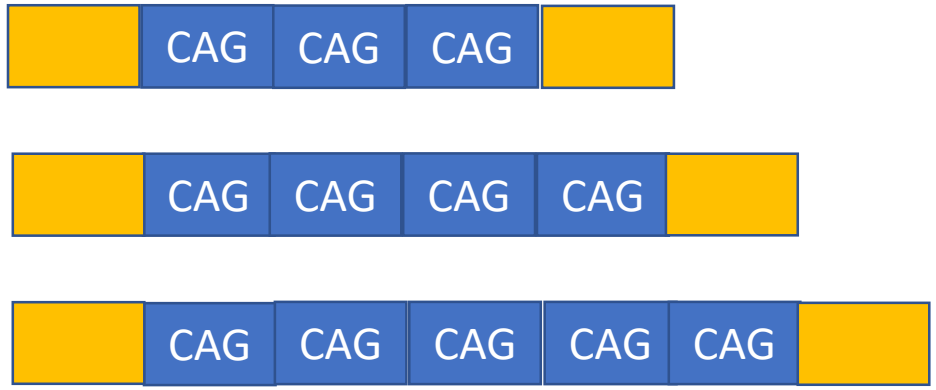
basic
polar

Genetic Variation – sequence variation

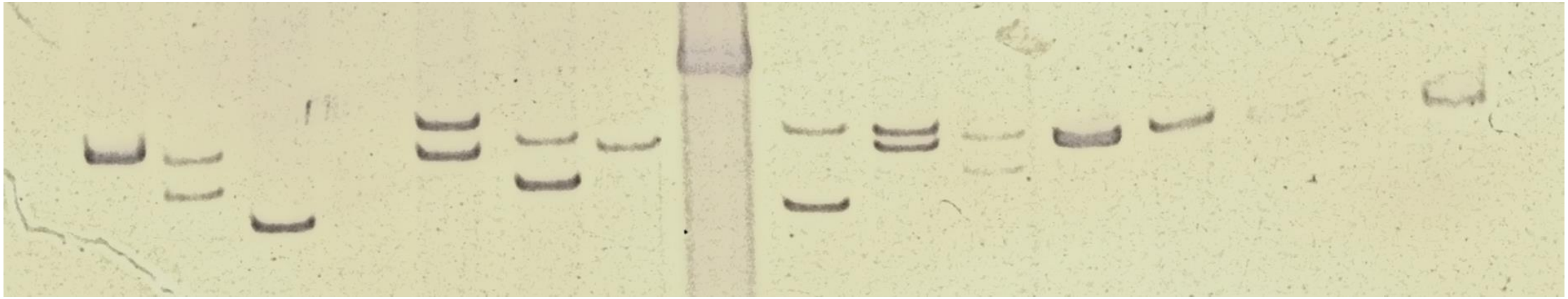




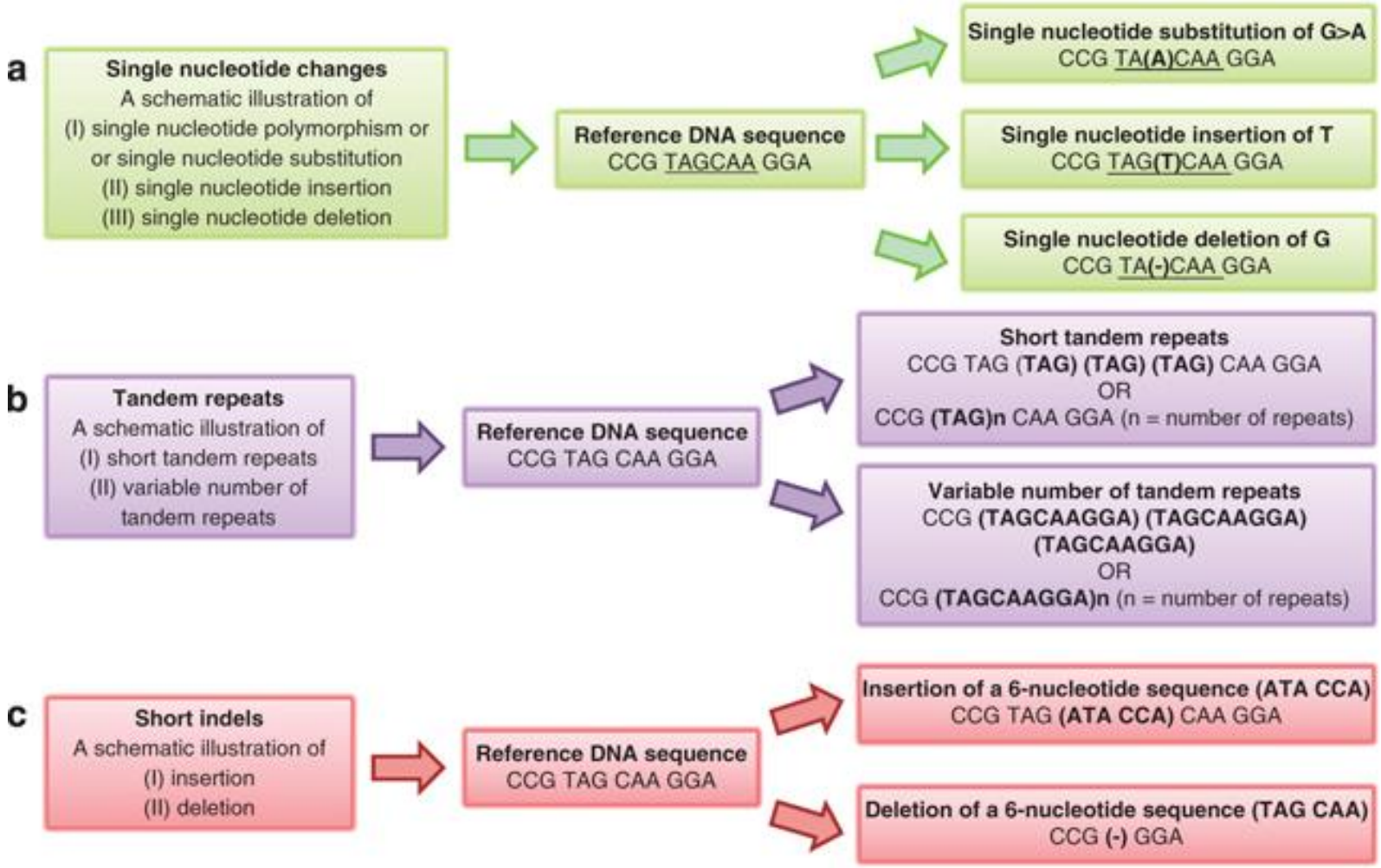
Tandem repeats – Huntington’s disease

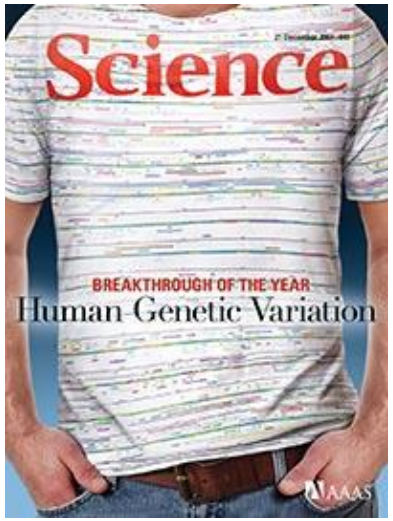


Repeat count	Classification	Disease status
<28	Normal	Unaffected
28–35	Intermediate	Unaffected
36–40	Reduced-penetrance	May be affected
>40	Full-penetrance	Affected



Genetic Variation – sequence variation





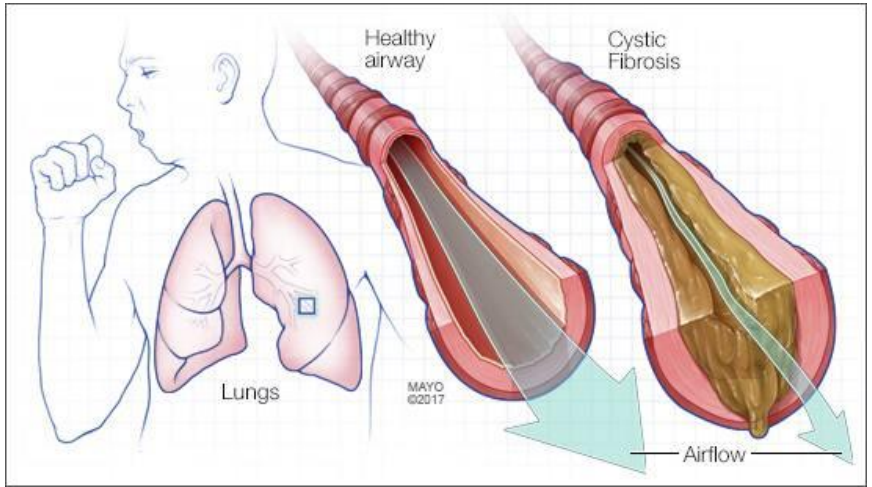
Deletion – cystic fibrosis

Functioning CFTR sequence:

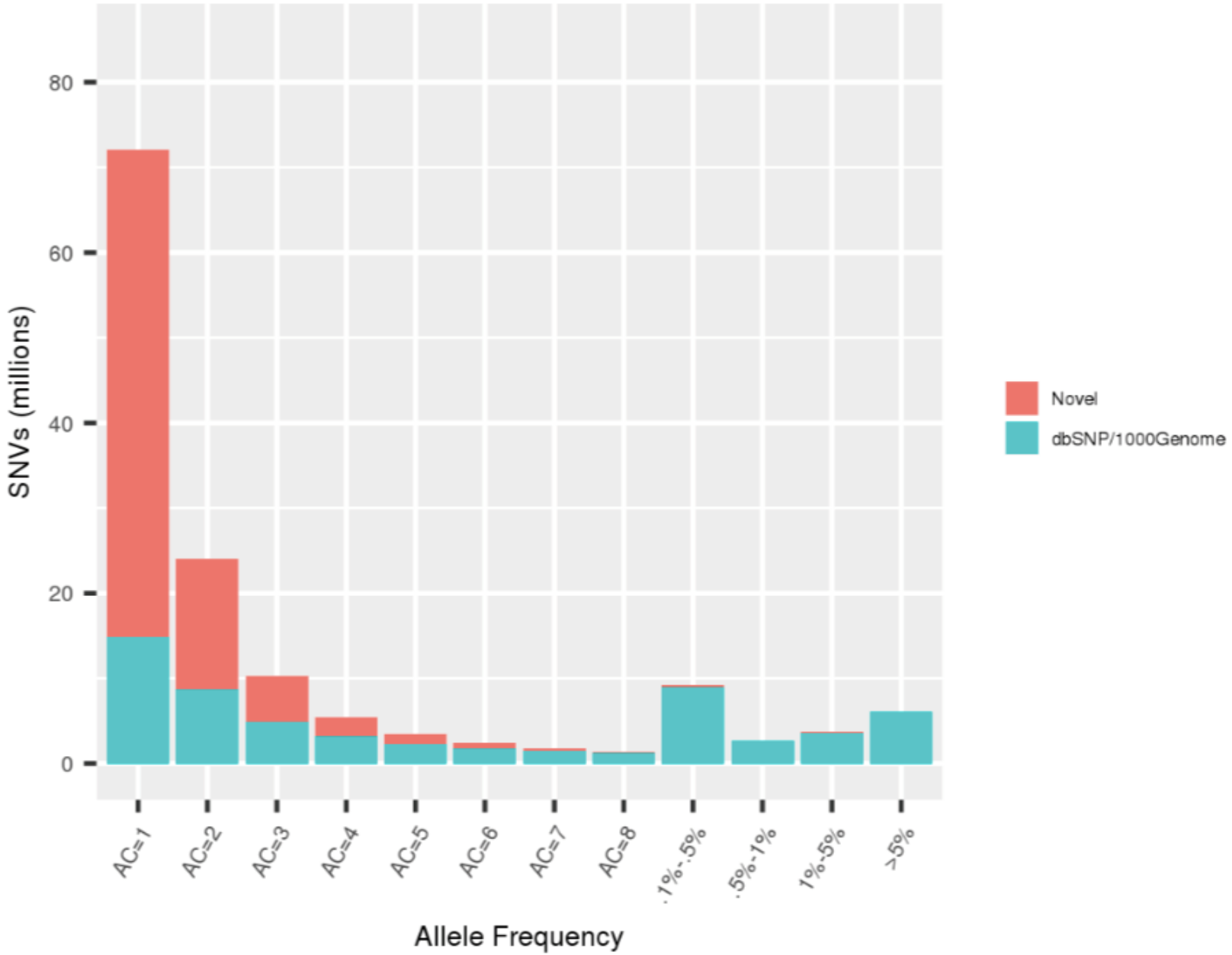
Nucleotide	ATC	ATC	TTT	GGT	GTT
Amino acid	Ile	Ile	Phe	Gly	Val

F508Del variant inactivating chloride channel:

Nucleotide	ATC	ATT	GGT	GTT
Amino acid	Ile	Ile	Gly	Val

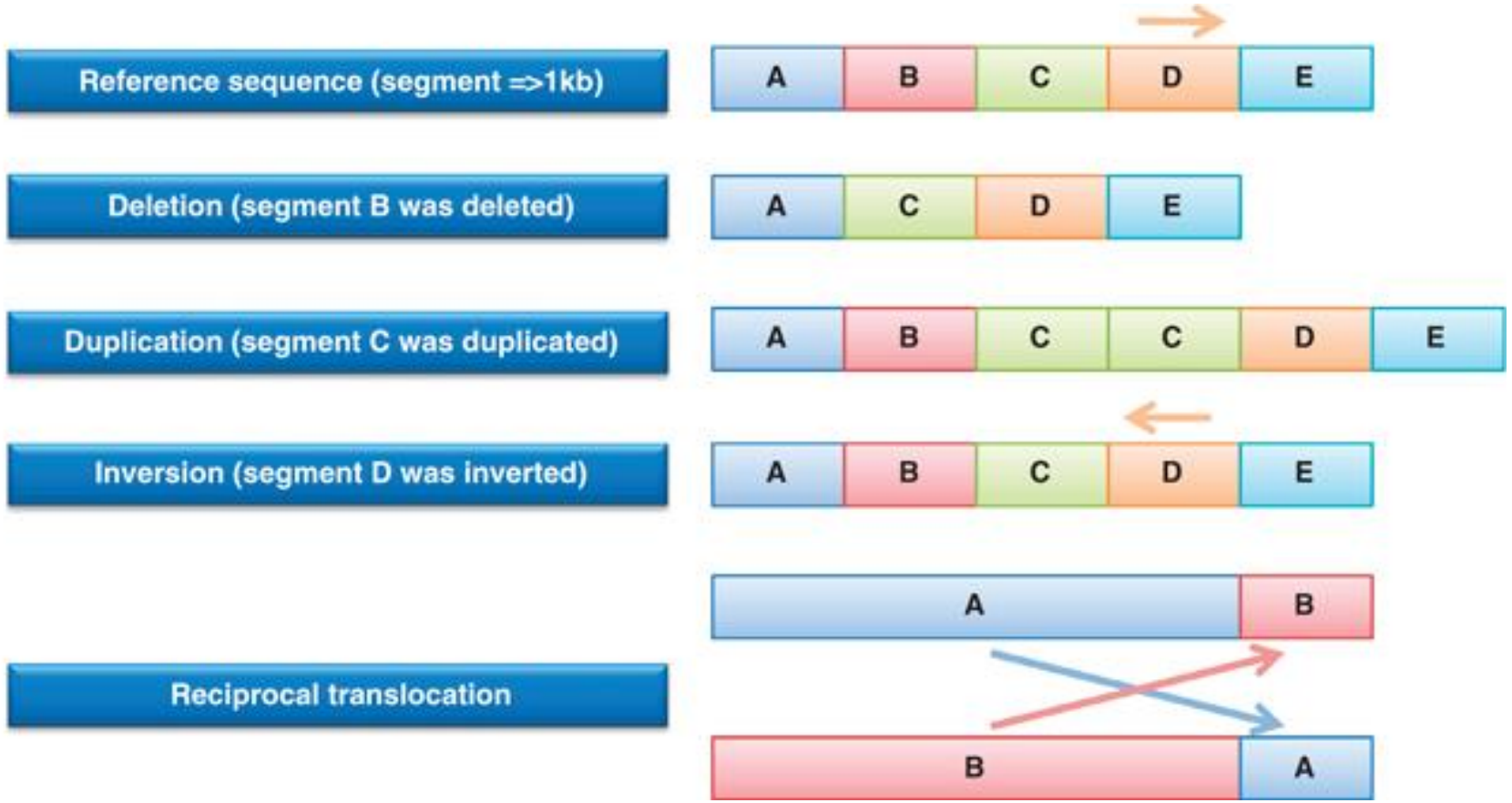


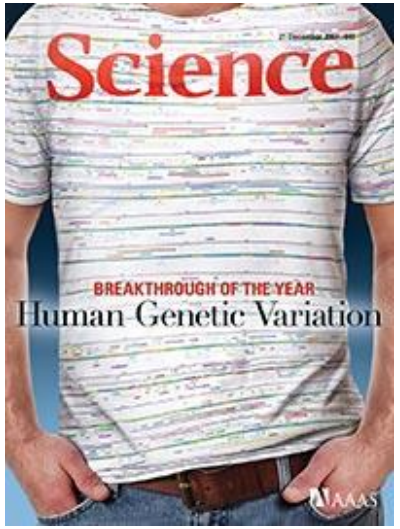
A recent study sequenced 10,545 human genomes and found more than 150 million variants



Telenti, PNAS 2016

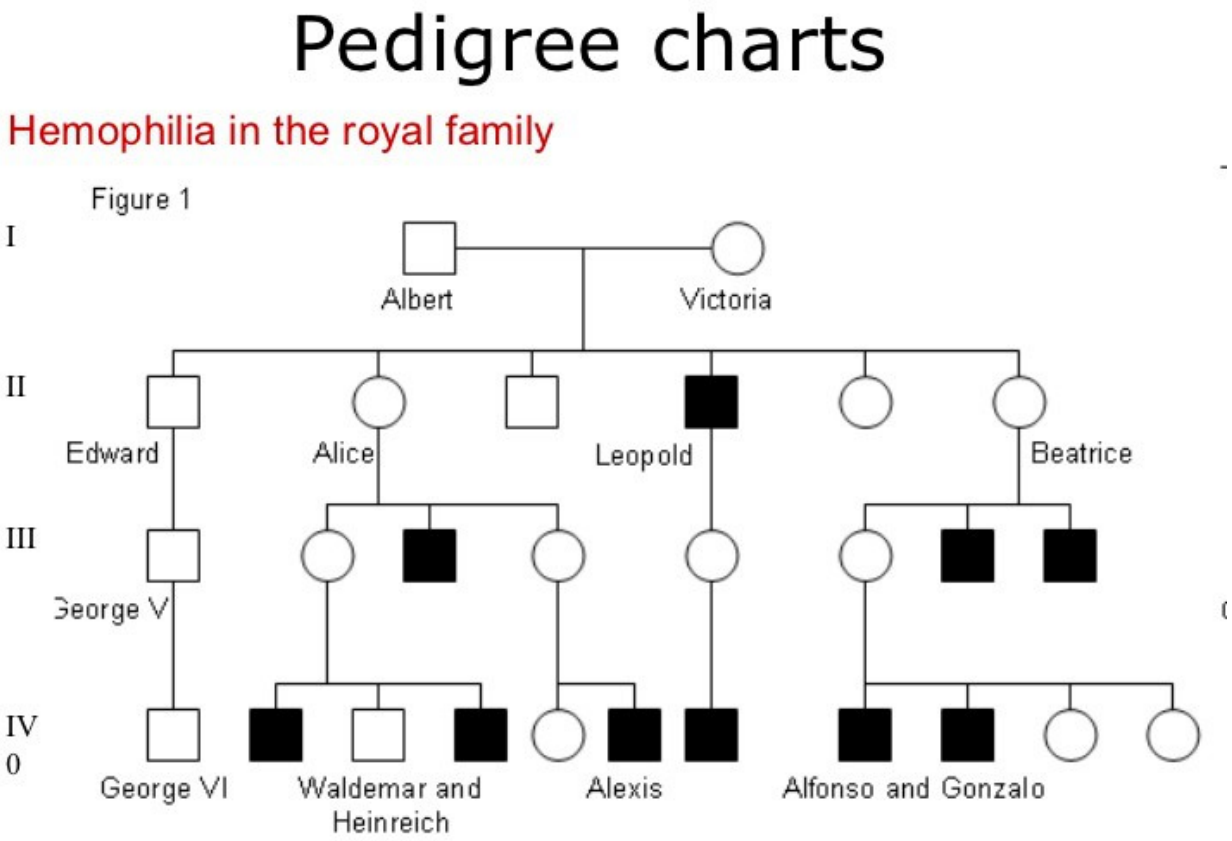
Genetic Variation – structural variation



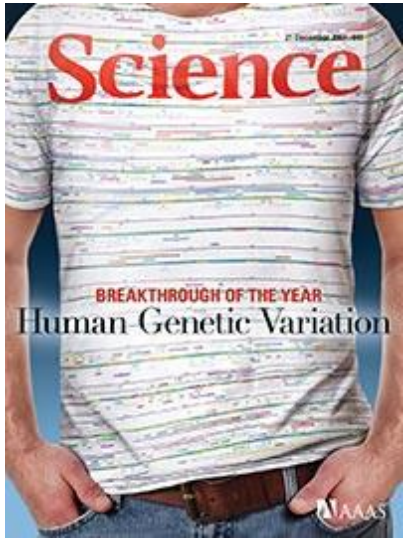


Structural variation

- Inversion in factor VIII gene causes haemophilia (clotting deficiency).



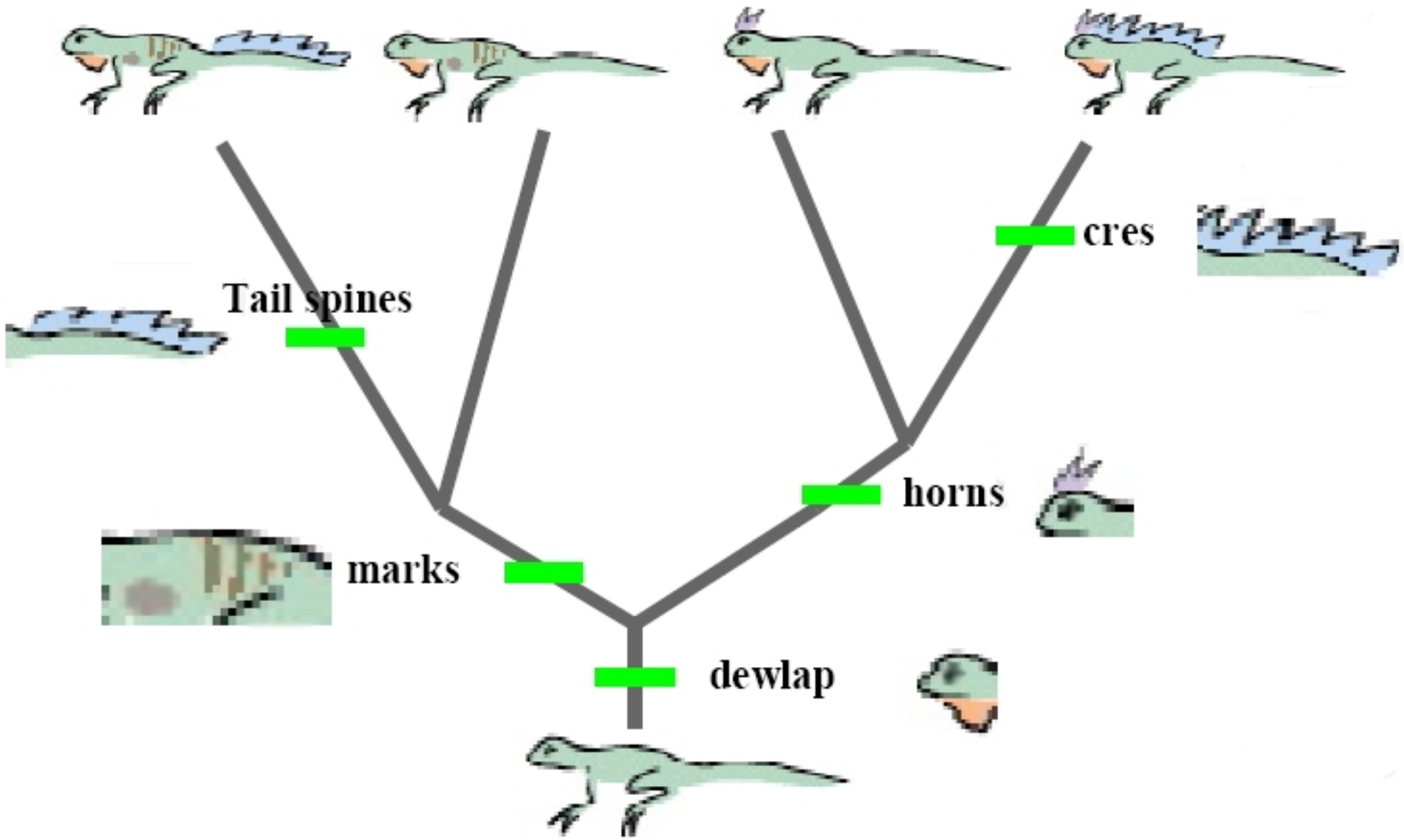
Genetic Variation



We find that a typical genome differs from the reference human genome at 4.1 million to 5.0 million sites. Although >99.9% of variants consist of SNPs and short indels, structural variants affect more bases: the typical genome contains an estimated 2,100 to 2,500 structural variants, affecting ~20 million bases of sequence.

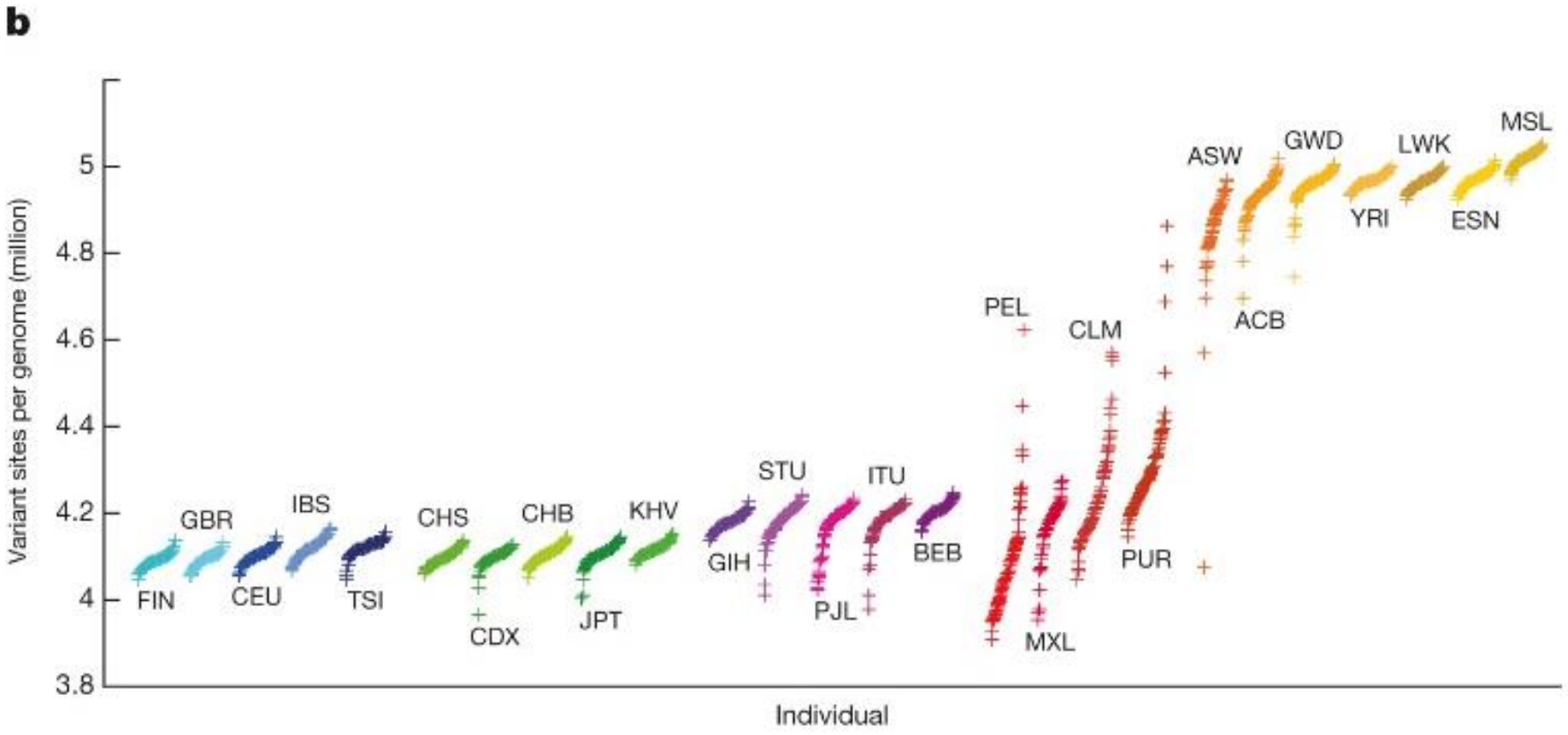
1000 Genomes project, Nature 2015

Accumulation of variants over generations



Baum 2008

Genetic diversity is greatest in Africans

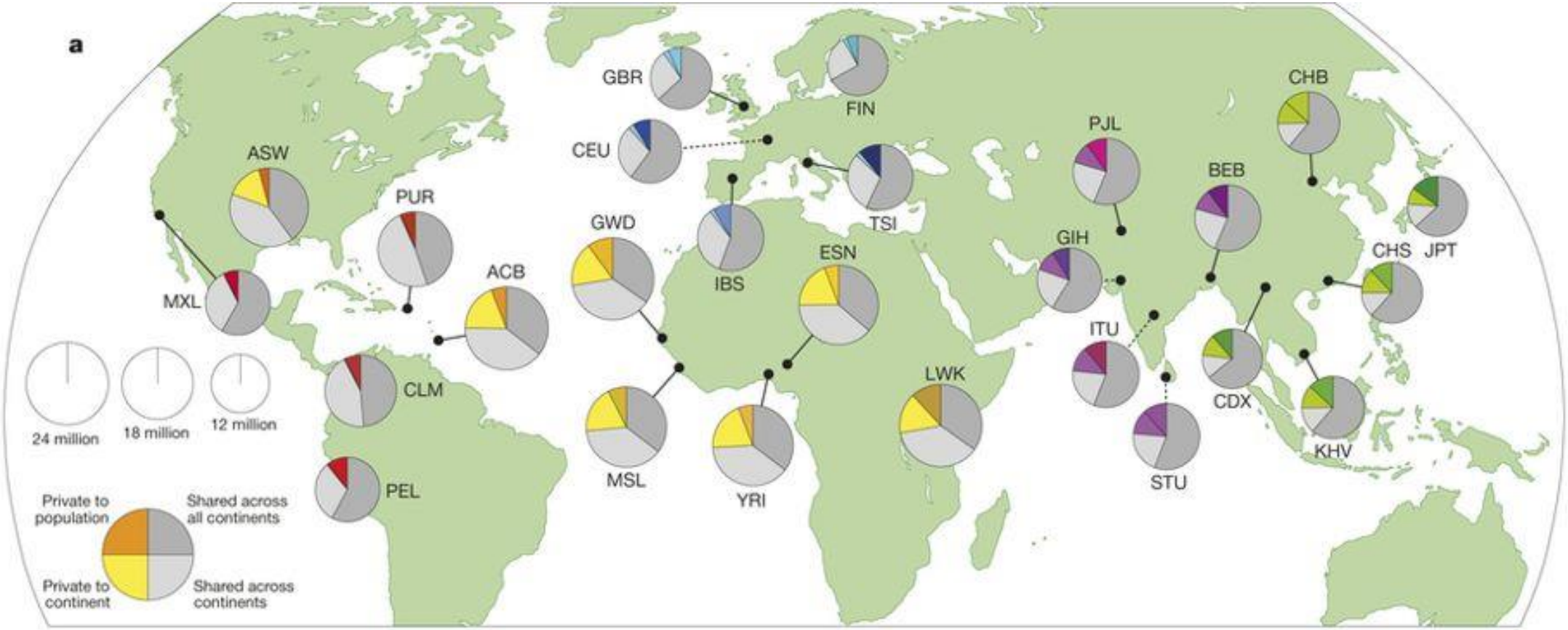


	Population Color	Continental Group Color	Analysis Panel
Africa			AFR
			AFR
			AFR
			AFR
			AFR
			AFR/AMR
			AFR/AMR
Americas			AMR
			AMR
			AMR
			AMR
East Asian			EAS
			EAS
			EAS
			EAS
			EAS
European			EUR
			EUR
			EUR
			EUR
			EUR
South Asian			SAS
			SAS
			SAS
			SAS
			SAS

1000 Genomes project, Nature 2015

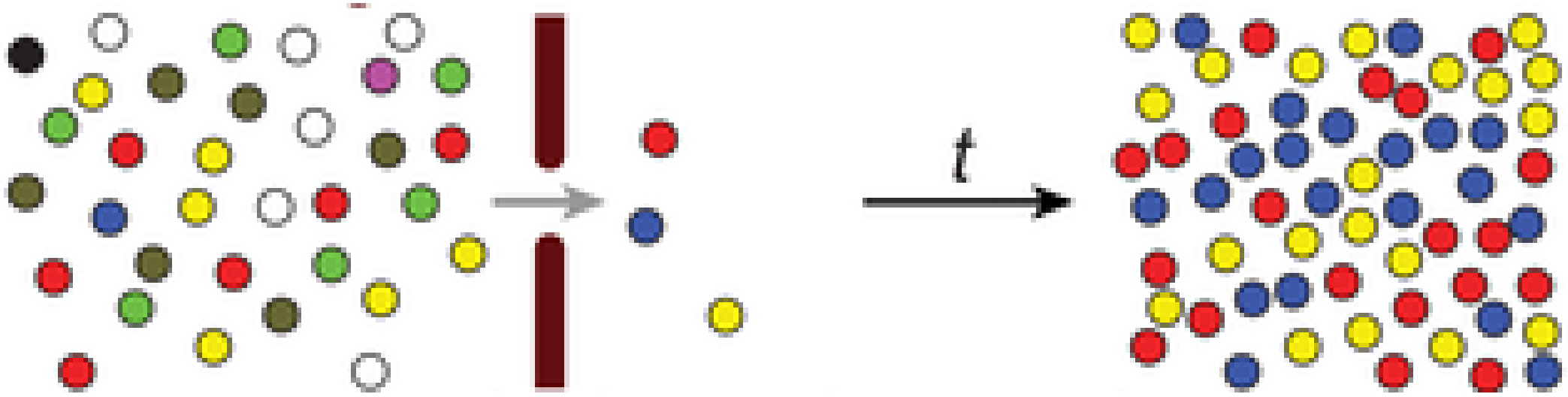
deMenocal & Stringer, Nature 2016

Variation in global populations



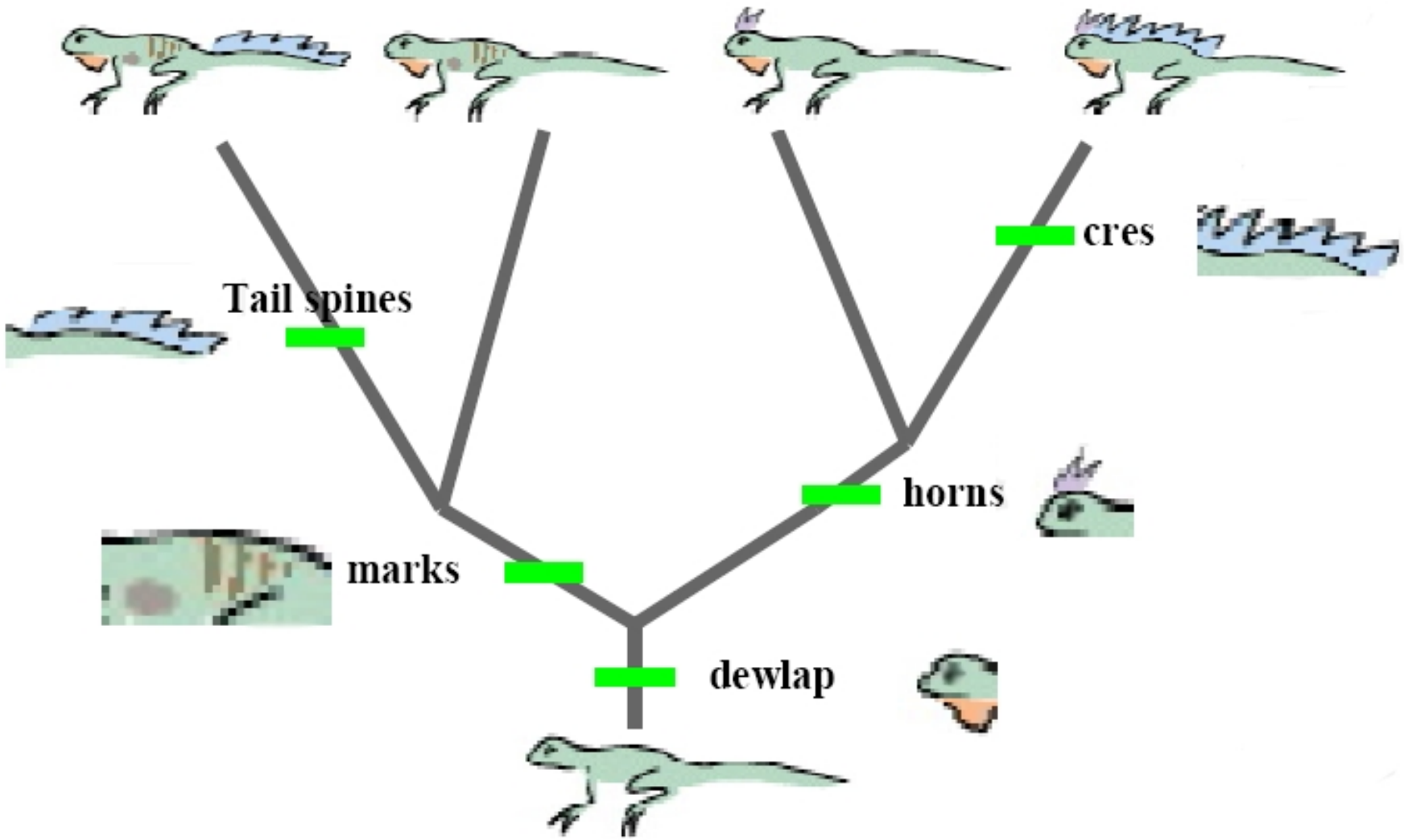
1000 Genomes project, Nature 2015

The “Out-of-Africa” migration is an example of a Population Bottleneck



Abel, PLoS Pathogenics 2015

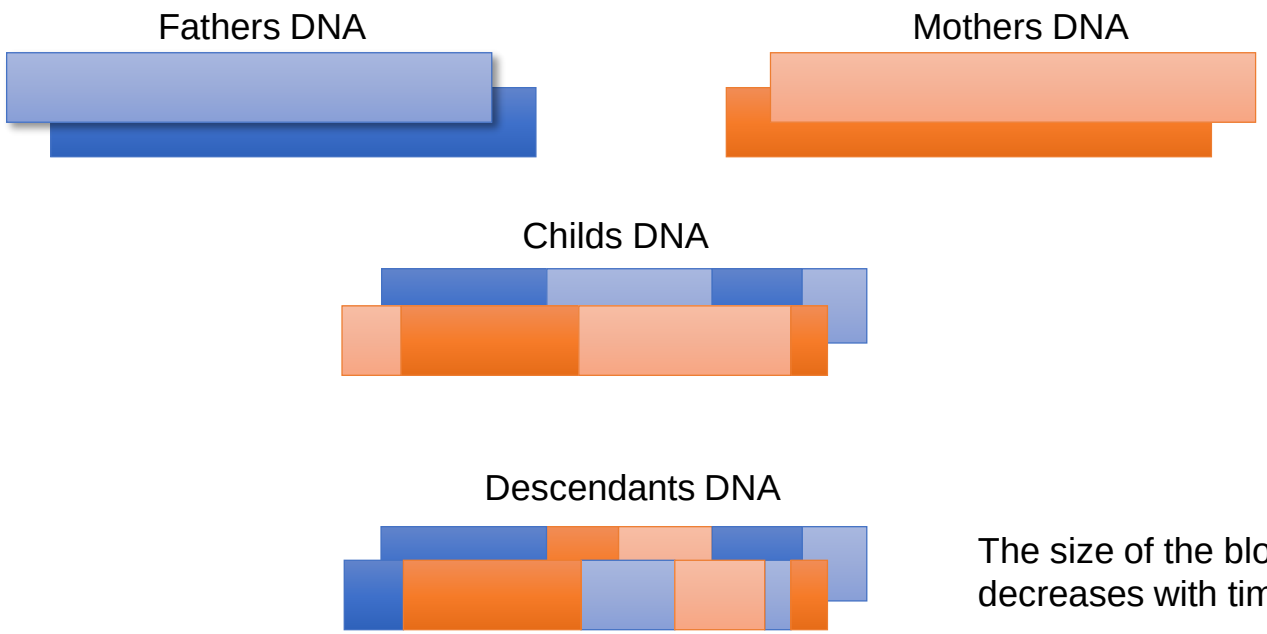
Accumulation of variants over generations



Baum 2008

DNA inheritance

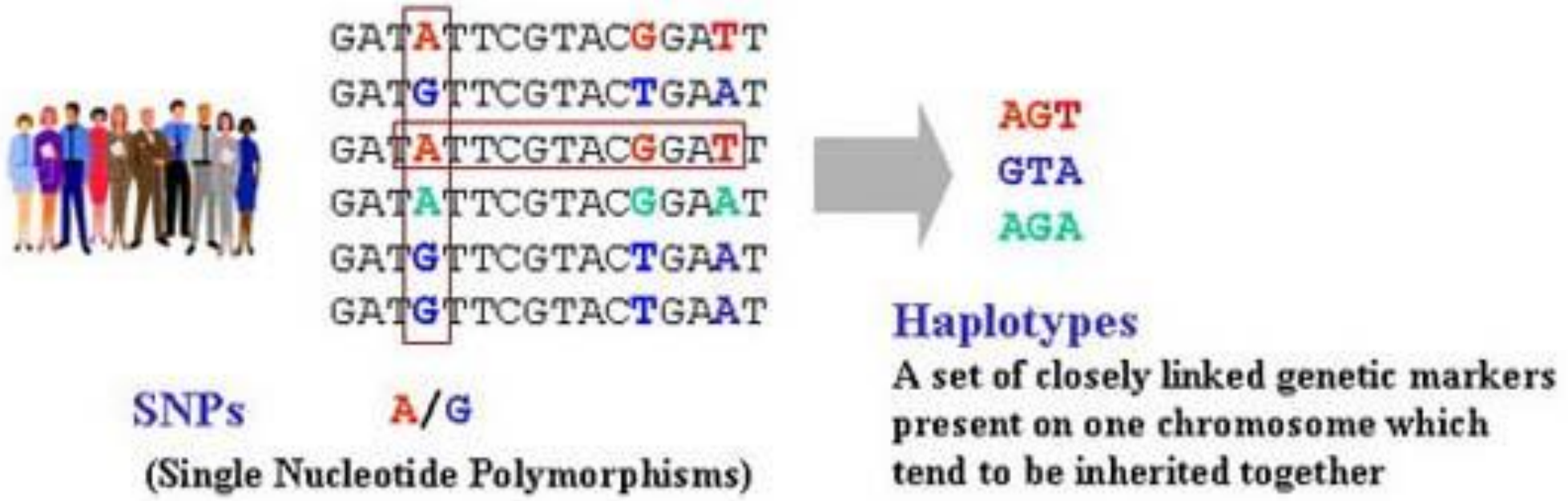
We inherit “blocks” of the genome from our parents (and not independent base-pairs)



The size of the blocks decreases with time

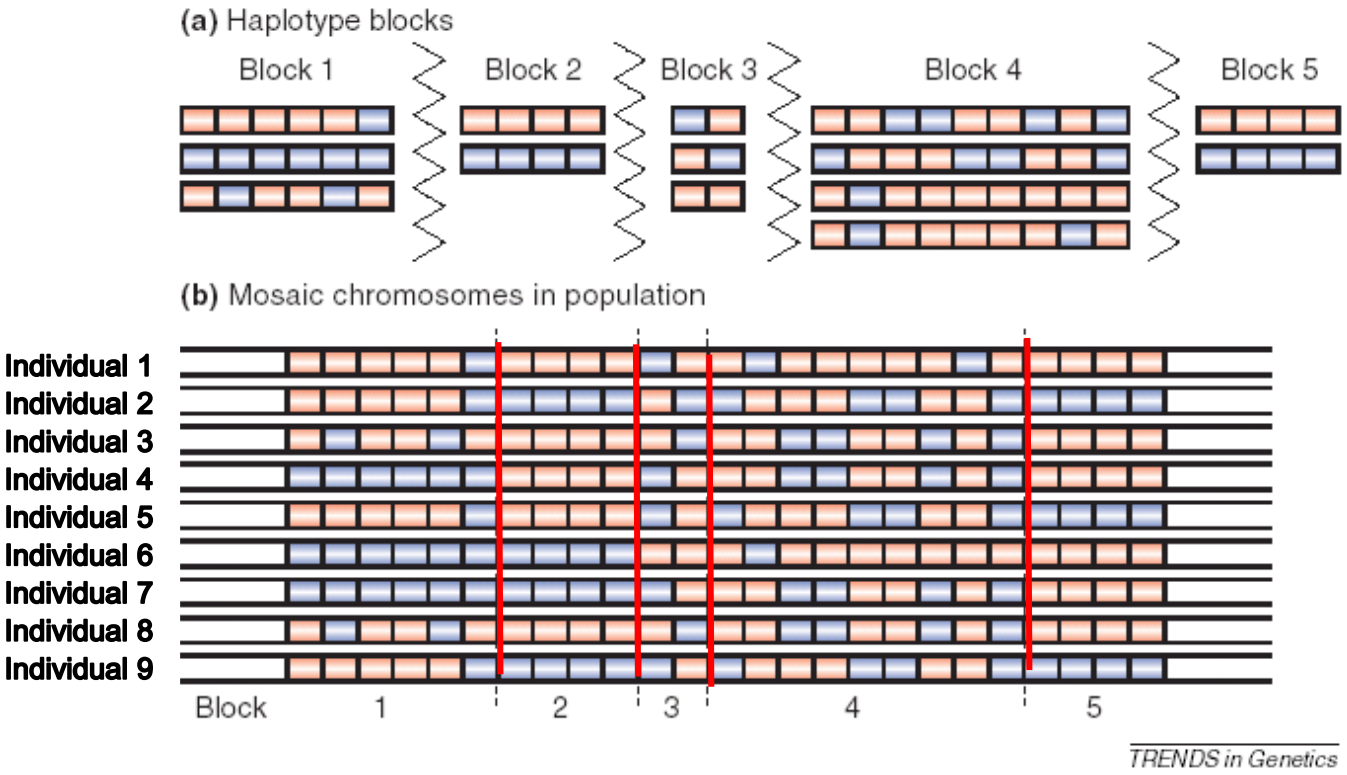
Haplotypes

Specific combination of SNPs occurring on the same segment of chromosome.



Haplotype blocks

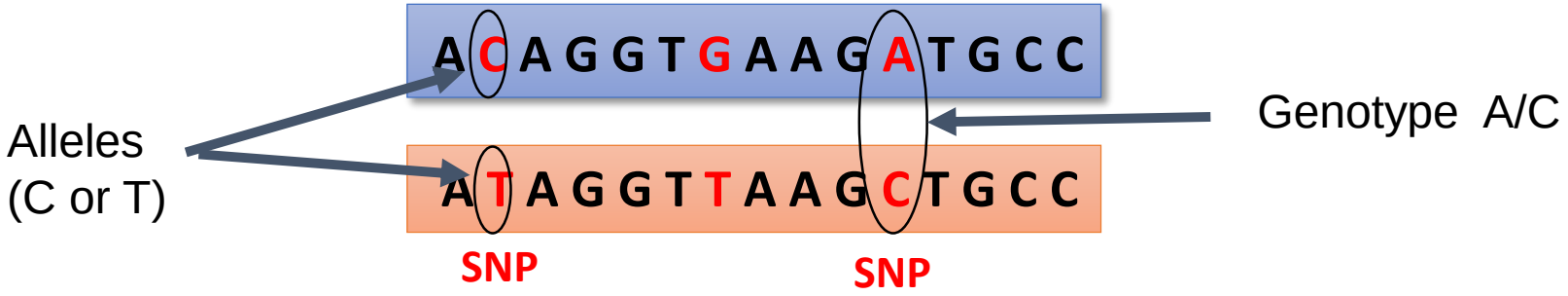
For N SNPs, there are 2^N possible haplotypes



Cardon & Abecasis, Trends In Genetics 2003

Allele vs. genotype

We inherit two copies of each chromosome



Genotypes
(A/A) – homozygous
(A/C) – heterozygous
(C/C) - homozygous

Haplotype phasing

When we genotype SNPs, we only see the genotype, and not the chromosome

For an individual who is **C/T,G/G, A/C**:

Genotype	SNP 1 C/T	SNP 2 G/G	SNP 3 A/C
Option 1	C	G	A
	T	G	C
Option 2	C	G	C
	T	G	A

Which one is true?

Determine haplotype phase

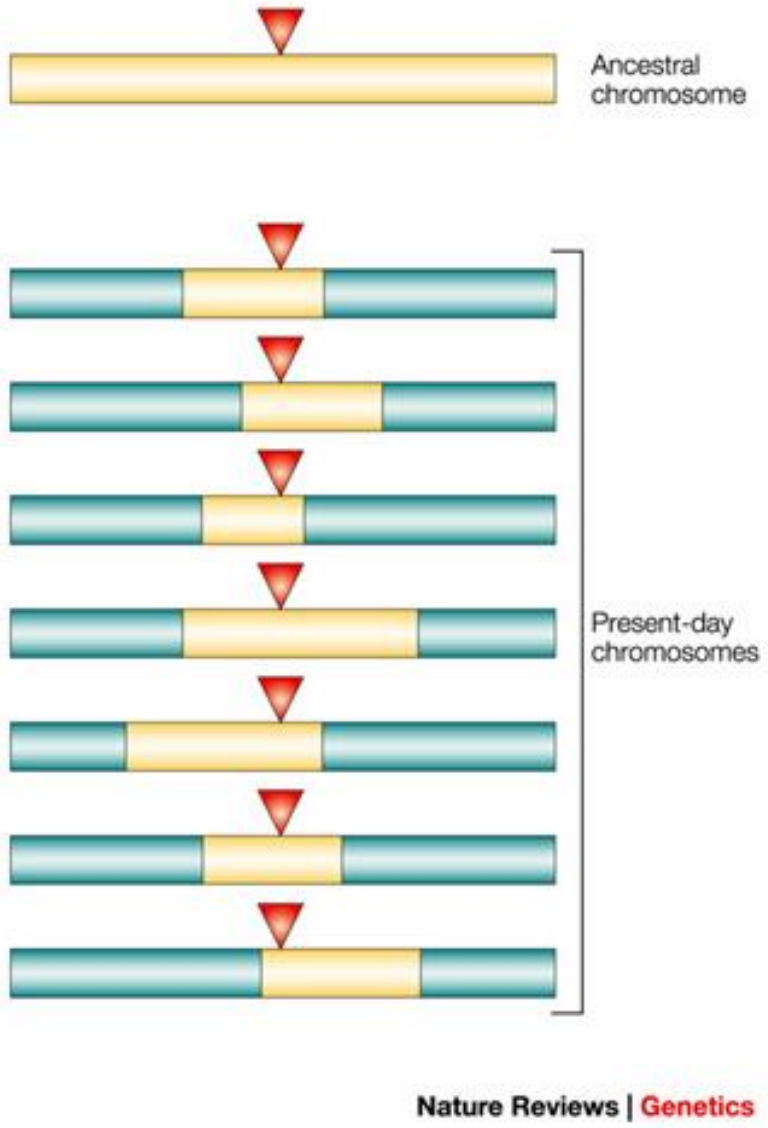
1) Look at family data
   - We seldom have this information

2) “Genotype” each chromosome
   - Very laboratory intensive and low-throughput

3) Infer the haplotype phase from the genotype data
   - Clark’s algorithm (*Clark, Mol Biol Evol, 1990)*
   - Expectation-Maximization algorithm *(Excoffier, Mol Biol Evol, 1995)*
   - Coalescent-based methods and hidden Markov models *(Li Genetics, 2003)*

Linkage Disequilibrium (LD)

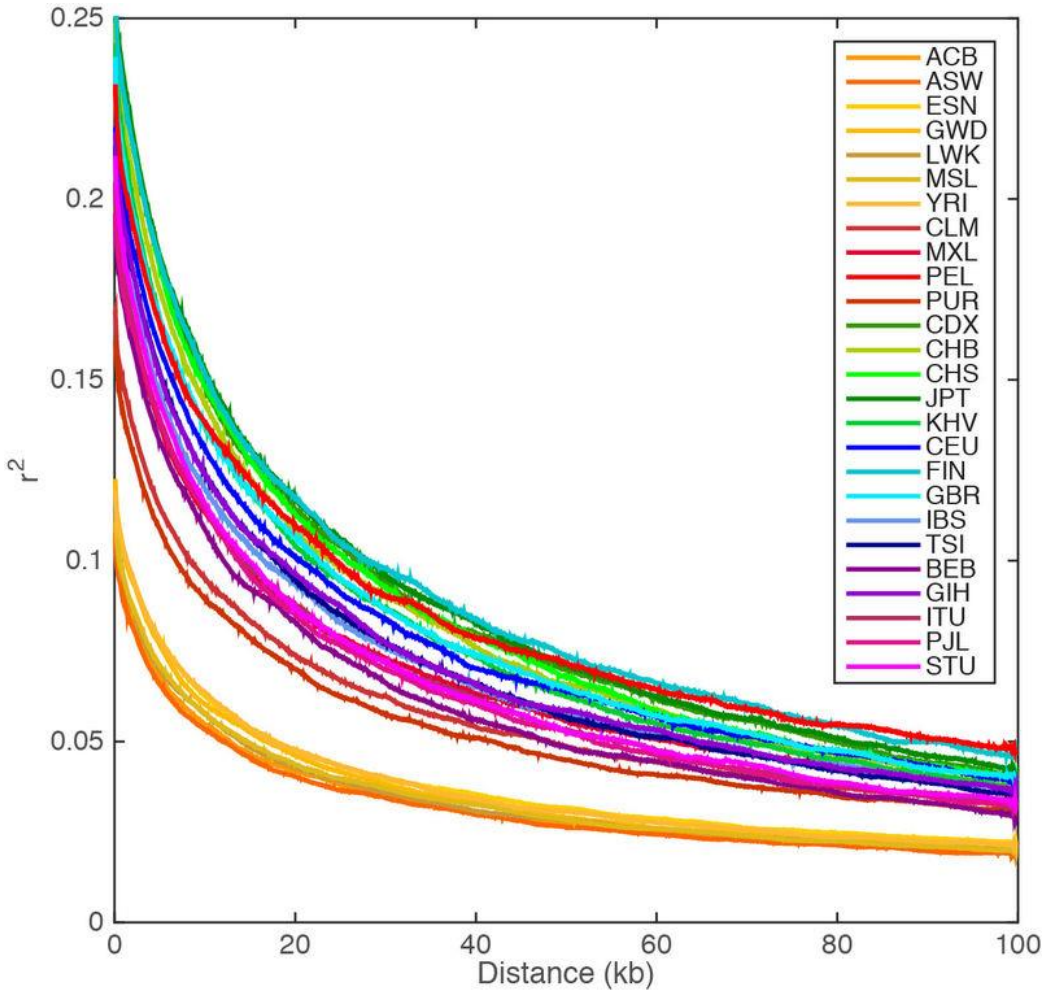
Linkage Disequilibrium (LD) is the non-random association between alleles at two or more loci



Chromosomal stretches derived from the common ancestor of all chromosomes are shown in yellow, and new stretches introduced by recombination are shown in blue. Markers that are physically close (that is, in the yellow regions of present-day chromosomes) tend to remain associated with the ancestral mutation (red arrow) even as recombination limits the extent of the region of association over time.

Ardlie, Nat Rev Genetics 2002

SNPs physically closer to each other tend to be in stronger LD



1000 Genomes Project, Nature 2015

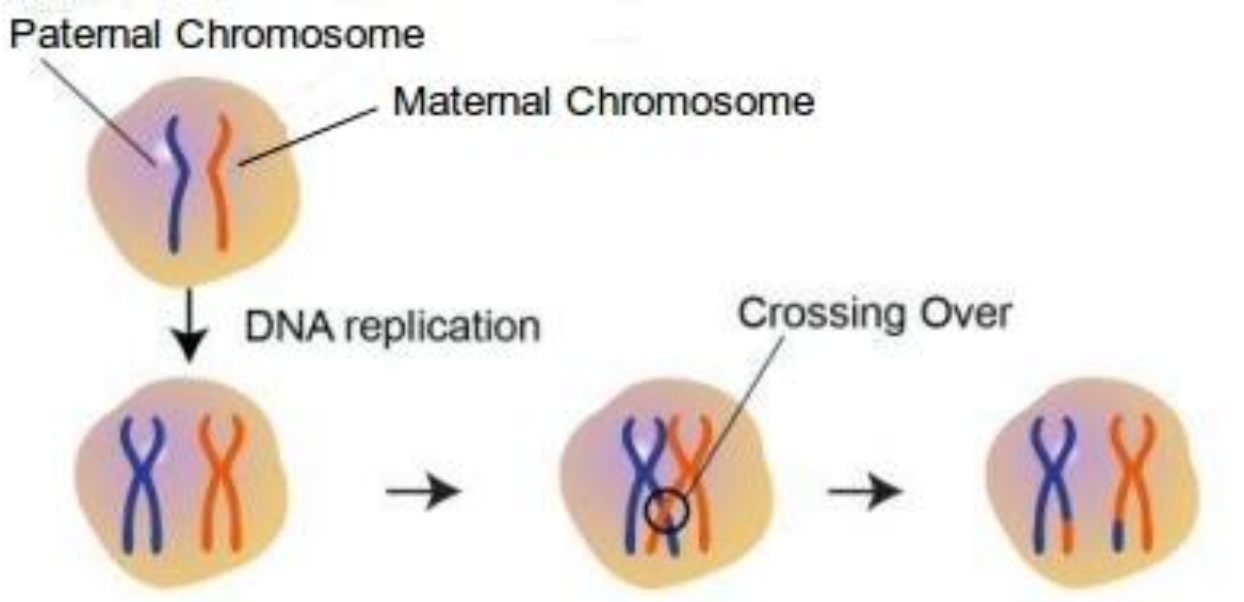
Factors that influence LD

- Recombination

Recombination

- Alleles on the same chromosome are inherited together unless *recombination (crossing over)* occurs

- The probability of recombination between two alleles increases with the distance between them



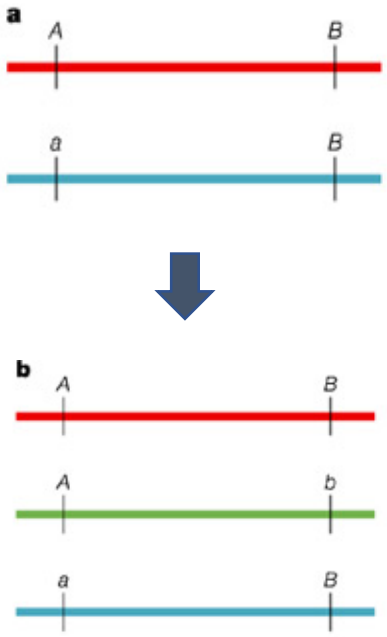
Recombination between 2 homologous chromosomes

Recombination breaks up LD

Start with a polymorphic locus with alleles *A* and *a*.

Ardlie, Nat Rev Genetics 2002

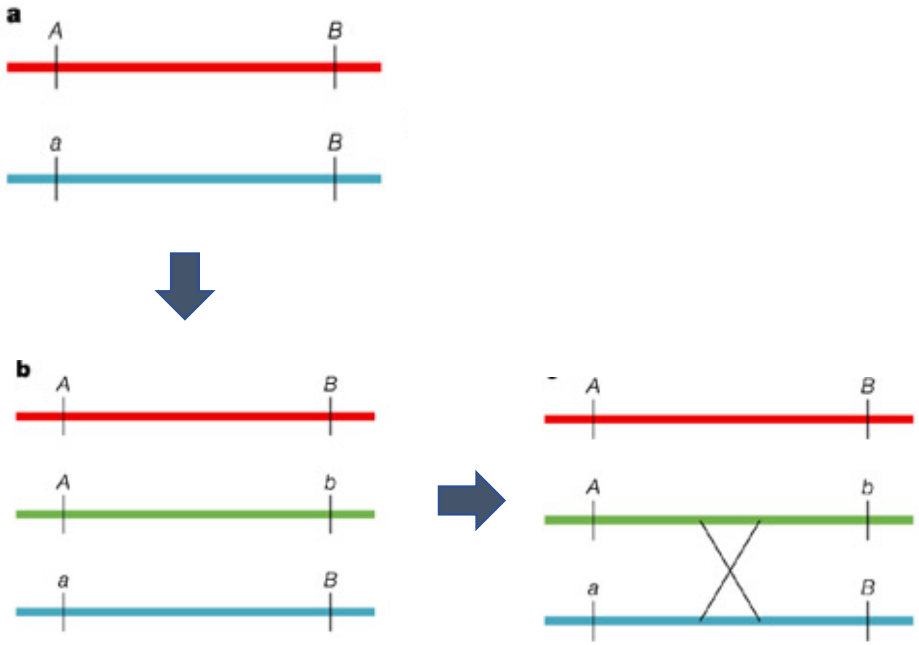
Recombination breaks up LD



When a mutation occurs at a nearby locus (B->b), this occurs on a single chromosome bearing either allele *A* or *a* at the first locus (*A* in this example). So, early in the lifetime of the mutation, only three out of the four possible haplotypes will be observed in the population. The *b* allele will always be found on a chromosome with the *A* allele.

Ardlie, Nat Rev Genetics 2002

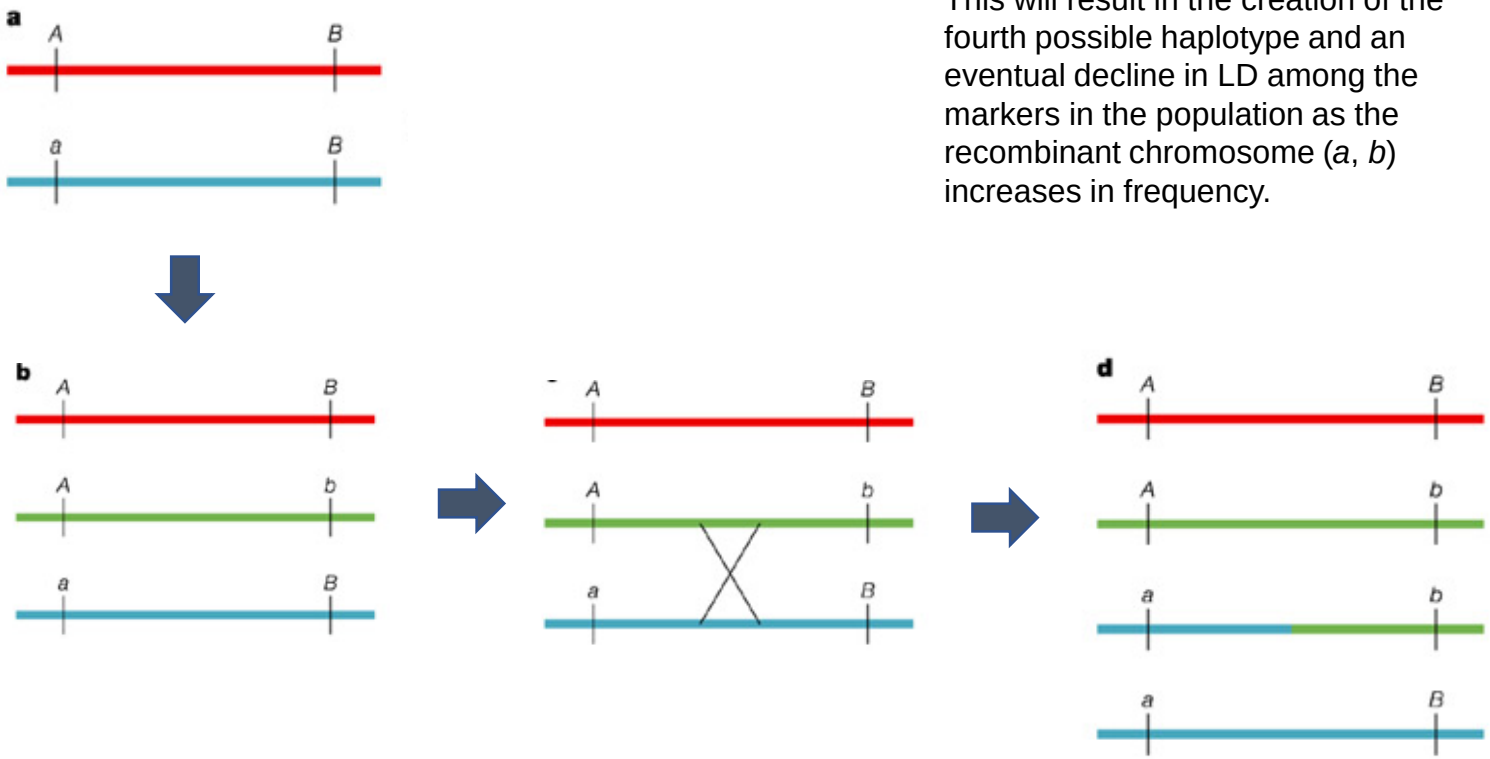
Recombination breaks up LD



With time, a recombination event will take place and the association between alleles at the two loci will gradually be disrupted

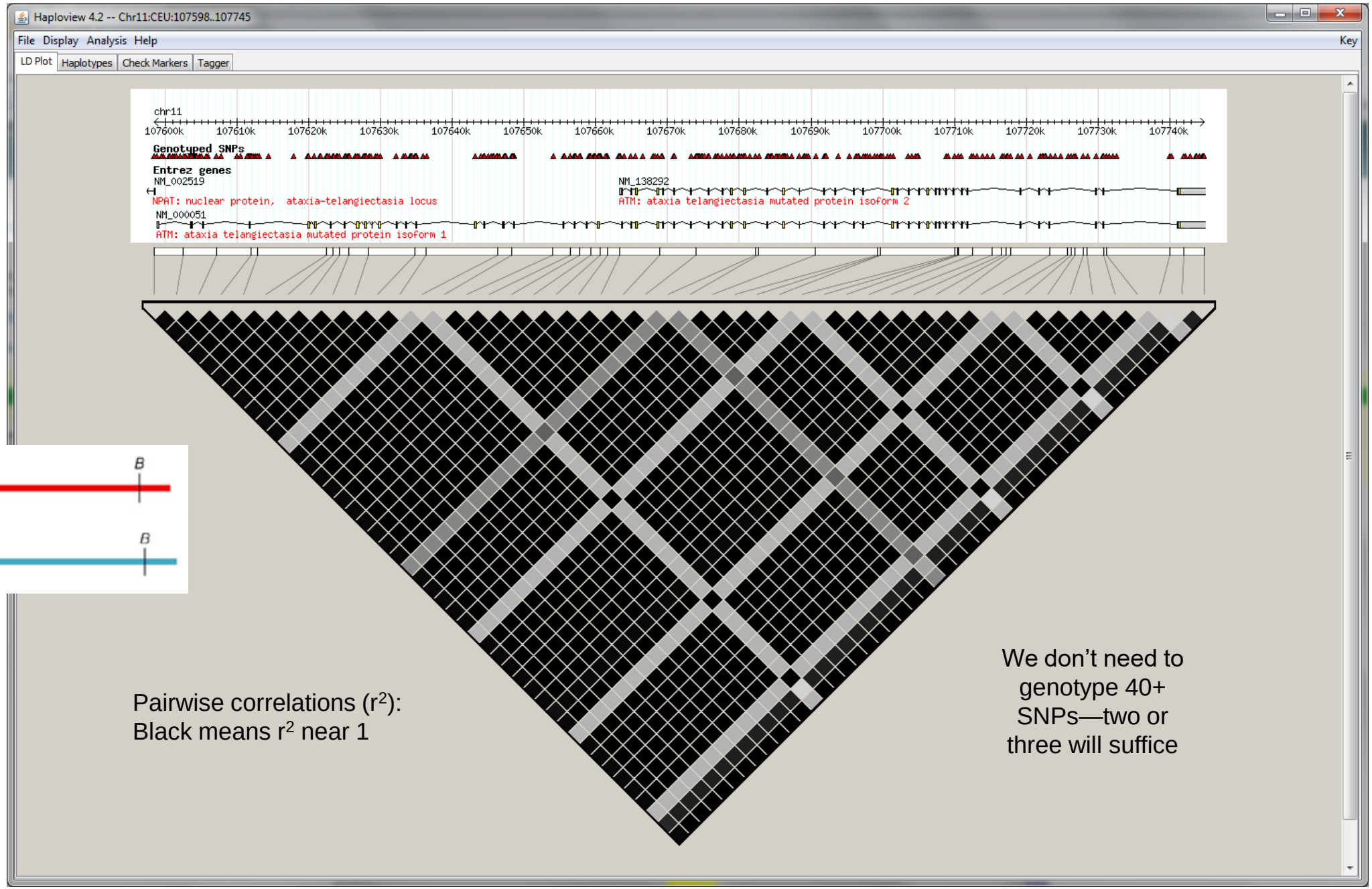
Ardlie, Nat Rev Genetics 2002

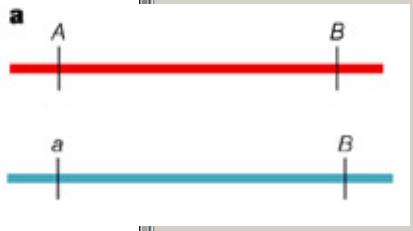
Recombination breaks up LD



This will result in the creation of the fourth possible haplotype and an eventual decline in LD among the markers in the population as the recombinant chromosome (*a*, *b*) increases in frequency.

Ardlie, Nat Rev Genetics 2002



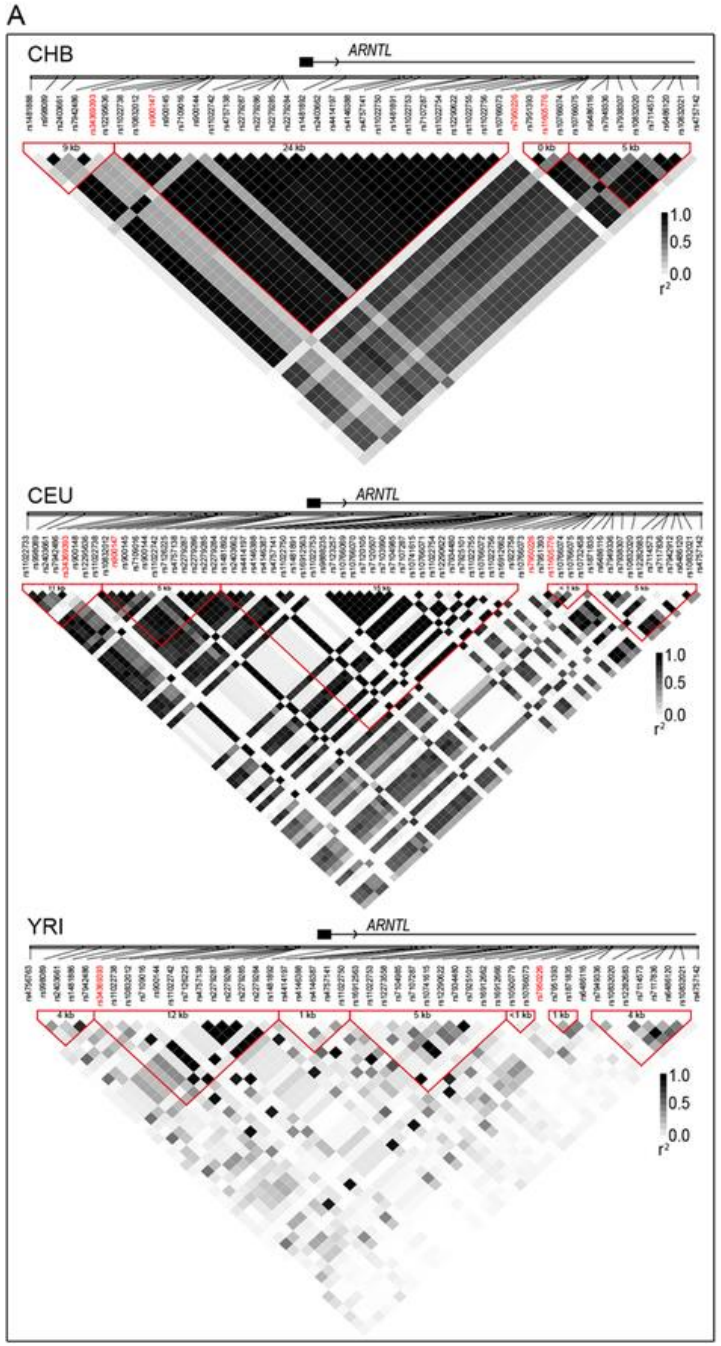


Pairwise correlations (r^2):
Black means r^2 near 1

We don't need to genotype 40+ SNPs—two or three will suffice

Sometimes just a few SNPs are enough to explain the genetic variation in a region. These SNPs are called ‘tag’ SNPs

Caveat: Tag SNPs are not particularly efficient for rare SNPs



Region associated with Parkinson’s Disease in Han Chinese

Gu, Sci Rep 2015

Linkage disequilibrium

- 20% sunny days in Seattle.
- How often do sunny days fall on the weekend?

Linkage disequilibrium

- 20% sunny days in Seattle
- How often do sunny days fall on the weekend?

Assume no linkage:

20% sunny days, 2/7 weekend days (0.29)

Likelihood have sunny day and weekend = p(sunny day)*p(weekend)

= 0.20 *(0.29) = 0.059 = 6% of days will be sunny and a weekend.

Linkage disequilibrium

- 20% sunny days
- How often do sunny days fall on the weekend?

Assume complete linkage.

29% weekends and 20% sunny days means that there has to be some weekend days that are not sunny even if all sunny days happen on a weekend.

Calculation of LD

There are 4 possible haplotypes for SNP1 (Aa) and SNP2 (Bb)

	SNP2 (Bb)		
SNP1 (Aa)	AB	Ab	p_A
	aB	ab	p_a
	p_B	p_b	1

Calculation of LD

Haplotypes frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other. (This is called linkage equilibrium)

	SNP2 (Bb)		
SNP1 (Aa)	AB $p_{AB} = p_A p_B$	Ab $p_{Ab} = p_A p_b$	p_A
	aB $p_{aB} = p_a p_B$	ab $p_{ab} = p_a p_b$	p_a
	p_B	p_b	1

Calculation of LD

Haplotypes frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other (This is called linkage equilibrium)

	SNP2 (Bb)		
SNP1 (Aa)	AB $p_{AB}= p_A p_B$	Ab $p_{Ab}= p_A p_b$	p_A
	aB $p_{aB}= p_a p_B$	ab $p_{ab}= p_a p_b$	p_a
	p_B	p_b	1

We can infer LD as the deviation of observed haplotype frequency from its corresponding allele frequencies if SNP1 and SNP2 are independent of each other

	SNP2 (Bb)		
SNP1 (Aa)	AB $p_A p_B+D$	Ab $p_A p_b-D$	p_A
	aB $p_a p_B-D$	ab $p_a p_b+D$	p_a
	p_B	p_b	1

$$D=p_{AB}p_{ab}-p_{Ab}p_{aB}$$

Calculation of LD

Haplotypes frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other (This is called linkage equilibrium)

	Sunny		
	Y	**N**	
Y **Weekend** **N**	0.06 $p_{AB} = p_A p_B$	.23 $p_{Ab} = p_A p_b$	0.29
	.14 $p_{aB} = p_a p_B$	0.57 $p_{ab} = p_a p_b$	0.71
	0.20	0.80	1

Calculation of LD – All sunny days happen on a weekend…

Haplotypes frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other (This is called linkage equilibrium)

	Sunny		
	Y	**N**	
Y Weekend N	0.06 $p_{AB}= p_A p_B$	.23 $p_{Ab}= p_A p_b$	0.29
	.14 $p_{aB}= p_a p_B$	0.57 $p_{ab}= p_a p_b$	0.71
	0.20	0.80	1

We can infer LD as the deviation of observed haplotype frequency from its corresponding allele frequencies if SNP1 and SNP2 are independent of each other

	Sunny		
	Y	**N**	
Y Weekend N	0.20 $p_A p_B+D$	0.09 $p_A p_b-D$	0.29
	0 $p_a p_B-D$	0.71 $p_a p_b+D$	0.71
	0.20	0.80	1

$$D=p_{AB}p_{ab}-p_{Ab}p_{aB}$$

Calculation of LD – All sunny days happen on a weekend…

Haplotypes frequencies if SNP1 (Aa) and SNP2 (Bb) are independent of each other (This is called linkage equilibrium)

	Sunny		
	Y	**N**	
Y Weekend N	0.06 $p_{AB}= p_A p_B$	.23 $p_{Ab}= p_A p_b$	0.29
	.14 $p_{aB}= p_a p_B$	0.57 $p_{ab}= p_a p_b$	0.71
	0.20	0.80	1

We can infer LD as the deviation of observed haplotype frequency from its corresponding allele frequencies if SNP1 and SNP2 are independent of each other

	Sunny		
	Y	**N**	
Y Weekend N	0.20 0.06+D	0.09 0.23-D	0.29
	0 0.14-D	0.71 0.57+D	0.71
	0.20	0.80	1

$$D=p_{AB}p_{ab}-p_{Ab}p_{aB}$$

Instead of D, we often express LD in terms of D’ (normalized D) or r^2 (correlation coefficient)

$$D' = \frac{D}{D_{max}},$$

$$D_{max} = \begin{cases} max\{-p_A p_B, -(1-p_A)(1-p_B)\}, when\ D < 0 \\ min\{p_A(1-p_B), (1-p_A)p_B\}, when\ D > 0 \end{cases}$$

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

Instead of D, we often express LD in terms of D’ (normalized D) or r^2 (correlation coefficient)

$$D' = \frac{0.14}{D_{max}},$$

$$D_{max} = \begin{cases} min\{0.29(1-0.20), (1-0.29)0.20\}, when\ D > 0 \end{cases}$$

Dmax = min(0.232, 0.142)
Dmax = 0.142
D’ = 0.14/0.142 = 0.99

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

Instead of D, we often express LD in terms of D’ (normalized D) or r^2 (correlation coefficient)

$$D' = \frac{0.14}{D_{max}},$$

$$D_{max} = \left\{ min\{0.29(1-0.20), (1-0.29)0.20\}, when\ D > 0 \right.$$

Dmax = min(0.232, 0.142)
Dmax = 0.142
D’ = 0.14/0.142 = 0.99

$$r^2 = \frac{(0.14)^2}{0.29*0.20*0.71*0.80}$$

Instead of D, we often express LD in terms of D’ (normalized D) or r^2 (correlation coefficient)

$$D' = \frac{0.14}{D_{max}},$$

$$D_{max} = \left\{ min\{0.29(1-0.20), (1-0.29)0.20\}, when\ D > 0 \right.$$

Dmax = min(0.232, 0.142)
Dmax = 0.142
D’ = 0.14/0.142 = 0.99

$$r^2 = \frac{(0.14)^2}{0.29*0.20*0.71*0.80}$$

$$r^2 = 0.59$$

How does LD influence our study power?

- If a SNP C and causal SNP G are in LD with $\mathbf{r^2}$, then a study with N cases and controls which measures C (but not G) will have the same power to detect an association between C and disease as a study with $\mathbf{r^2}$ **N** cases and controls that directly measured G.

- $\mathbf{r^2}$ **N** is the "effective sample size"
  - If the r^2 between your measured SNP C and causal SNP G is 0.5 you need to double your sample size to obtain the same power as if you had measured (genotyped) G directly.

Pritchard and Przeworski (2001)

LD calculation exercise

SNPs rs6025 and rs4524 are both associated with venous thromboembolism (blood clot in a vein). The number of alleles for each SNP based on 503 individuals are displayed in the table below. Based on these numbers, calculate

a) **Frequencies of the four alleles (rs6025-C, rs6025-T, rs4524-G, rs4524-A)**
b) **Frequencies for the four haplotypes (C-G, C-A, T-G and T-A)**
c) **D' and r^2 between the two SNPs.**

Distribution of alleles for rs6025 and rs4524 across 503 individuals.

rs6025/rs4524	rs4524-G	rs4524-A	Total
rs6025-C	255	739	994
rs6025-T	0	12	12
Total	255	751	1006

LD calculation exercise

a) Frequencies of the four alleles (rs6025-C, rs6025-T, rs4524-G, rs4524-A)

rs6025/rs4524	rs4524-G	rs4524-A	Total
rs6025-C	255	739	994
rs6025-T	0	12	12
Total	255	751	1006

rs6025/ rs4524	**G**	**A**		
C	**255**	**739**	**C=994**	**pC=0.988**
T	**0**	**12**	**T=12**	**pT=0.012**
	G=255	**A=751**	**1006**	**1**
	pG=0.253	**pA=0.747**	**1**	

LD calculation exercise

b) Frequencies for the four haplotypes (C-G, C-A, T-G and T-A)

rs6025/rs4524	rs4524-G	rs4524-A	Total
rs6025-C	255	739	994
rs6025-T	0	12	12
Total	255	751	1006

rs6025/rs4524	**G**	**A**
C	**pCG=255/1006=0.253**	**pCA=739/1006=0.735**
T	**pTG=0**	**pTA=12/1006=0.0119**

LD calculation exercise

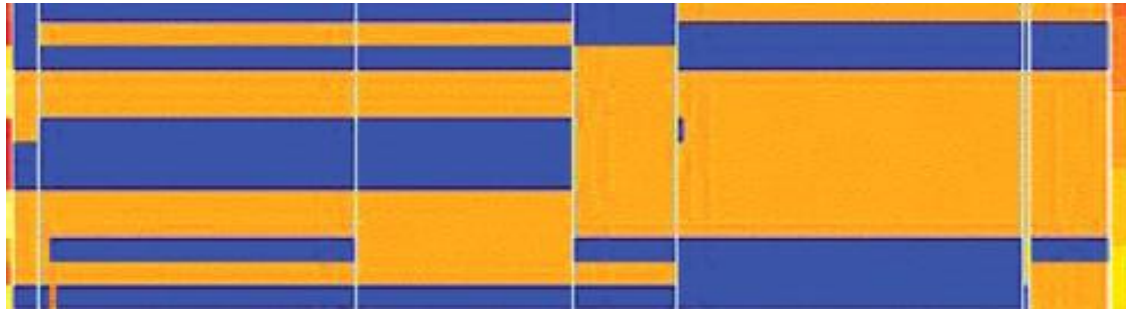
c) D’ and r^2 between the two SNPs.

D=pCG*pTA-pCA*pTG
=0.253*0.0119-0.735*0
=0.0030

D’=D/Dmax
=0.003/min{0.253*(1-0.988), (1-0.253)*0.998)
=0.003/min{0.003, 0.746}
=0.003/0.003
=1

$r^2=D^2/(p_{rs6025\text{-}C}*p_{rs6025\text{-}T}*p_{rs4524\text{-}G}*p_{rs4524\text{-}A})$
$=0.003^2/(0.988*0.012*0.253*0.747)$
$=9.217\text{x}10^{-6}/0.0022$
=0.0041

Linkage patterns and ancestry



Population 1

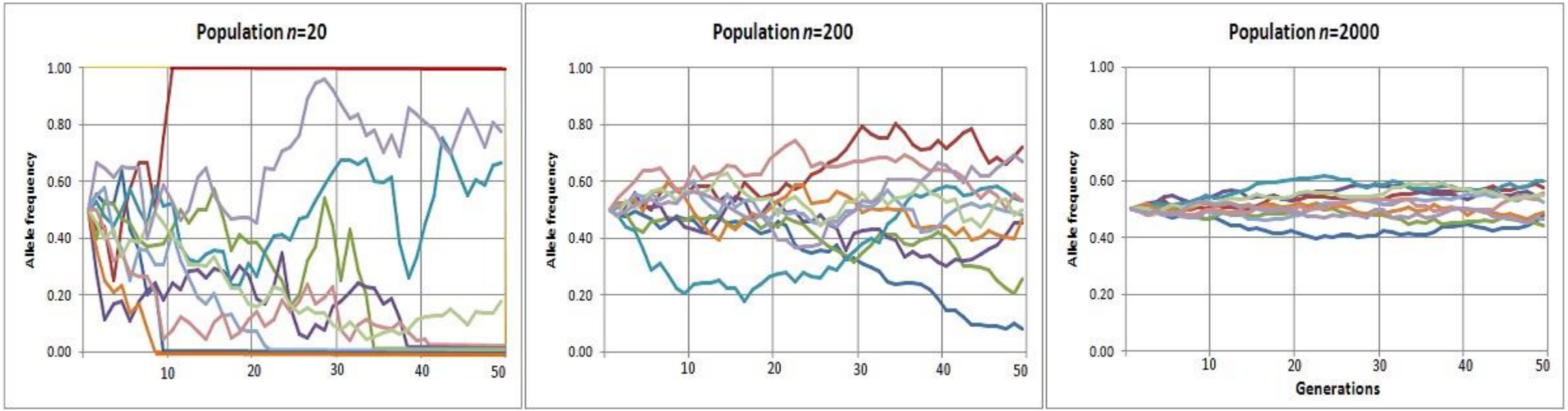
Population 2

Factors that influence LD

- New mutations

Factors that influence LD

- New mutations
- Genetic drift



Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth

Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations

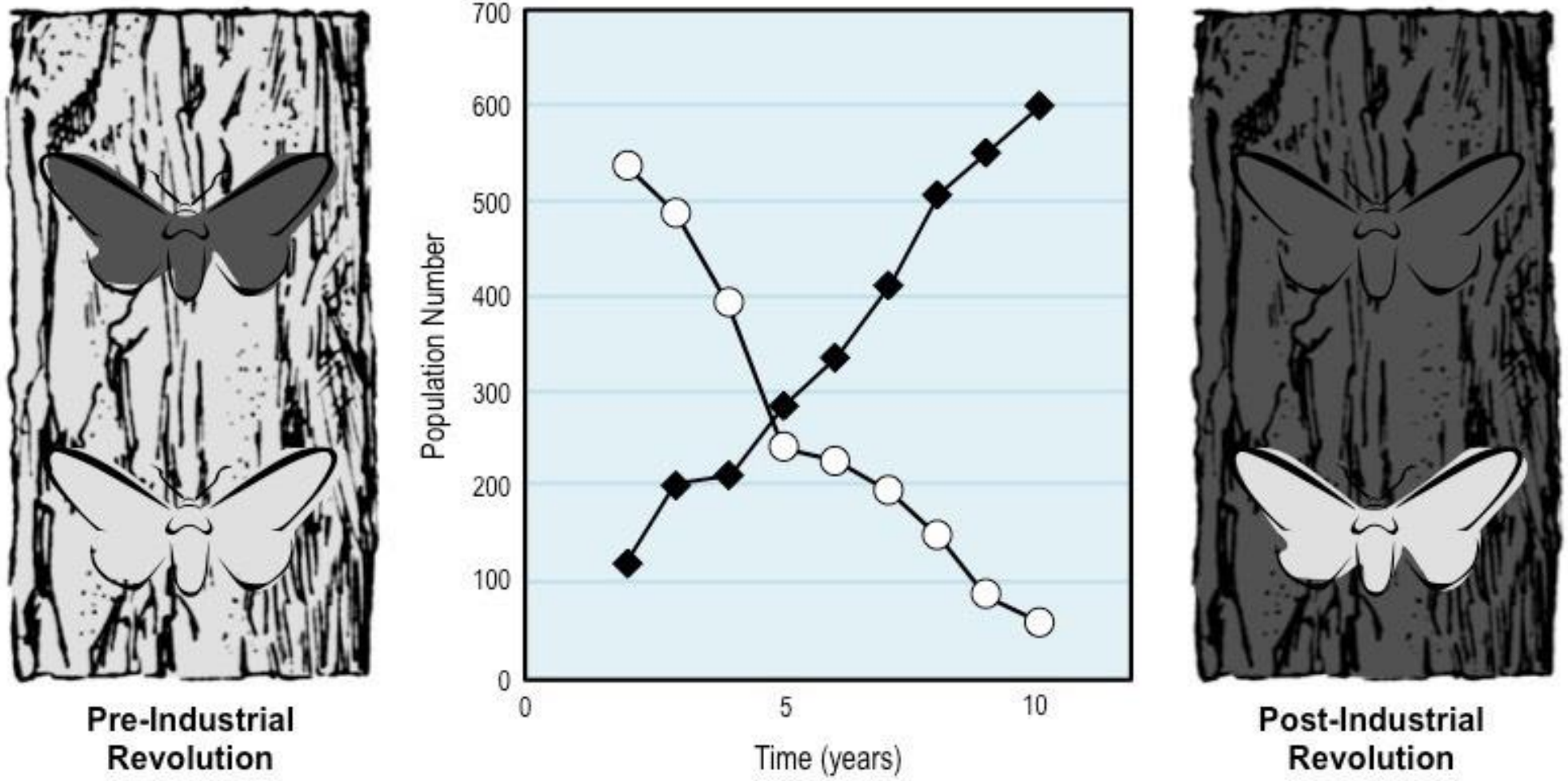
Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations
- Population structure – inbreeding

Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations
- Population structure – inbreeding
- Natural selection
  - Haplotypes that carry favorable mutations increase in frequency

Natural selection



Factors that influence LD

- New mutations
- Genetic drift
- Rapid population growth
- Admixture between populations
- Population structure – inbreeding
- Natural selection
  - Haplotypes that carry favorable mutations increase in frequency
- Recombination (recombination hotspots)
- Gene conversion (one-side recombination)

Summary

- Genetic variation can affect single nucleotides or longer segments through structural changes.
- Chunks of DNA are inherited together, allowing imputation and tagging SNPs for capturing genetic diversity (resulting from linkage disequilibrium).