Crash-course in genomics

Molecular biology : How does the genome code for function?

Genetics: How is the genome passed on from parent to child ?

Genetic variation: How does the genome change when it is passed on ?

Population and evolutionary genetics: How does the genome vary across populations and species?

Genome sequencing: How do we read the genome ?

Outline

Molecular biology : How does the genome code for function?

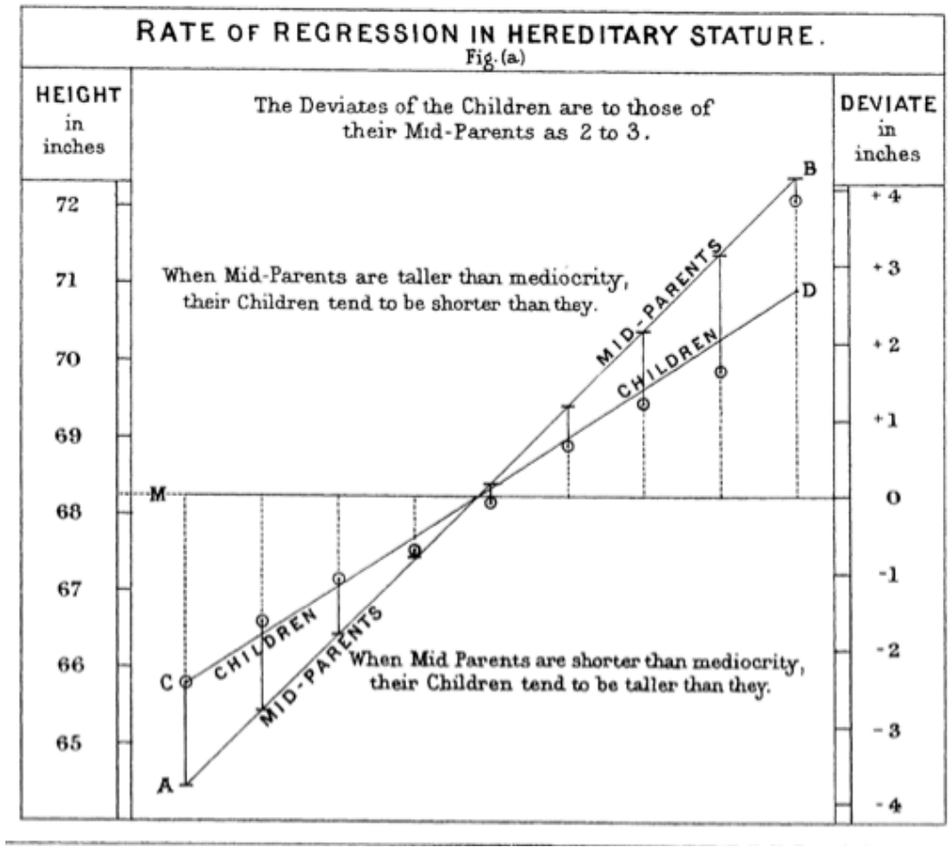
Genetics: How is the genome passed on from parent to child ?

Genetic variation: How does the genome change when it is passed on ?

What can we learn from genetic variation ?

Genome sequencing: How do we read the genome ?

Traits/Phenotype



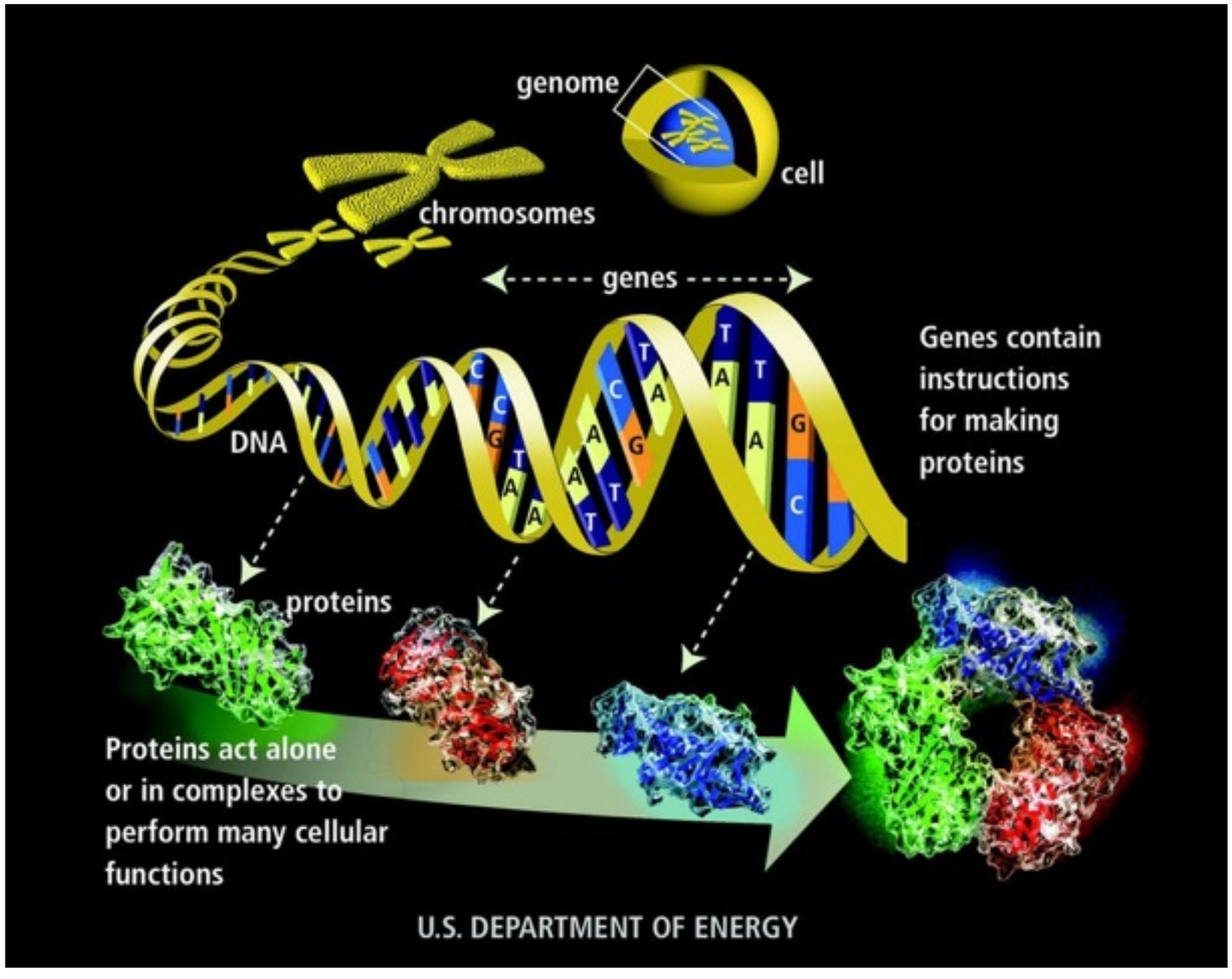
Galton et al. 1877

Trait/phenotype: Any observable that is inherited

Height, eye color, disease status, cellular measurements, IQ

Instructions that modulate traits found in the genome

Cells and DNA



Outline

Molecular biology : How does the genome code for function?

Genetics: How is the genome passed on from parent to child ?

Genetic variation: How does the genome change when it is passed on ?

What can we learn from genetic variation ?

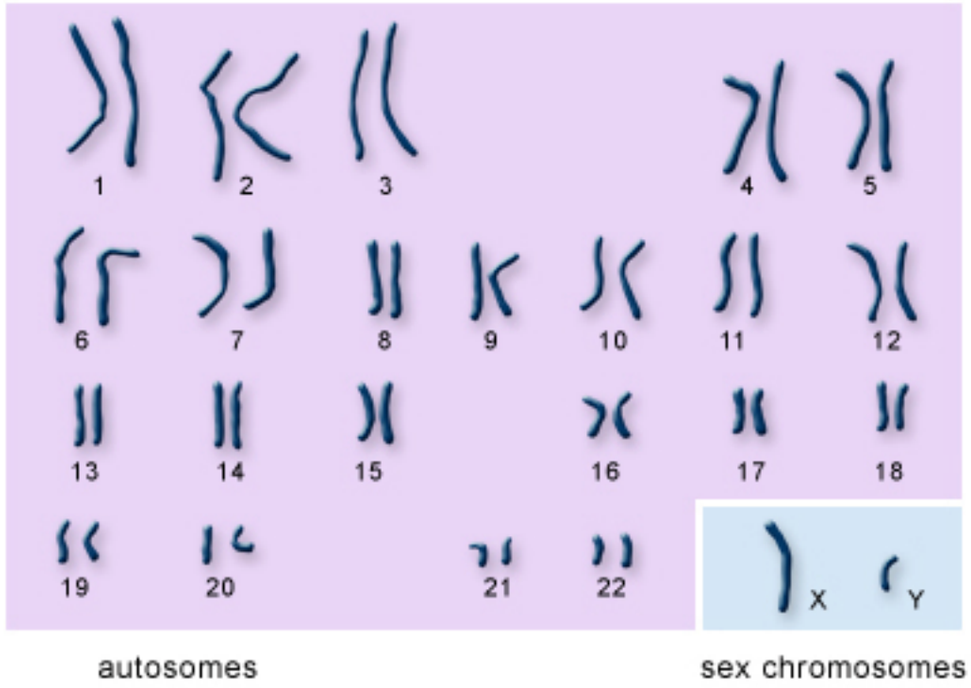
Genome sequencing: How do we read the genome ?

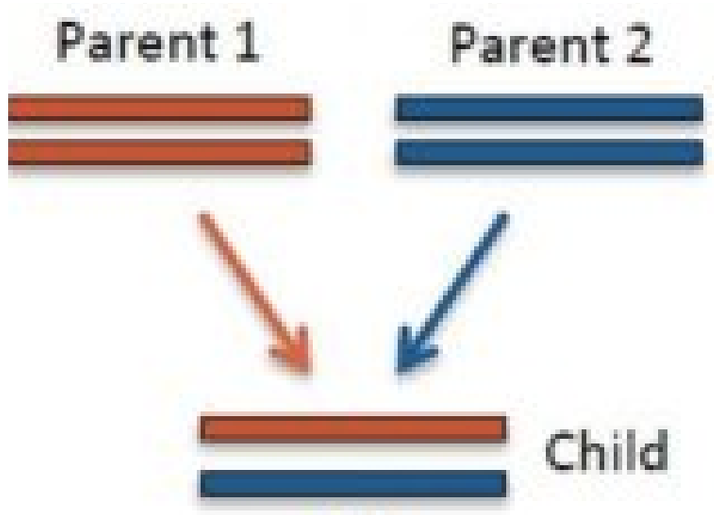
Genetics and inheritance

Typical human cell has 46 chromosomes

22 pairs of homologous chromosomes (autosomes)

1 pair of sex chromosomes

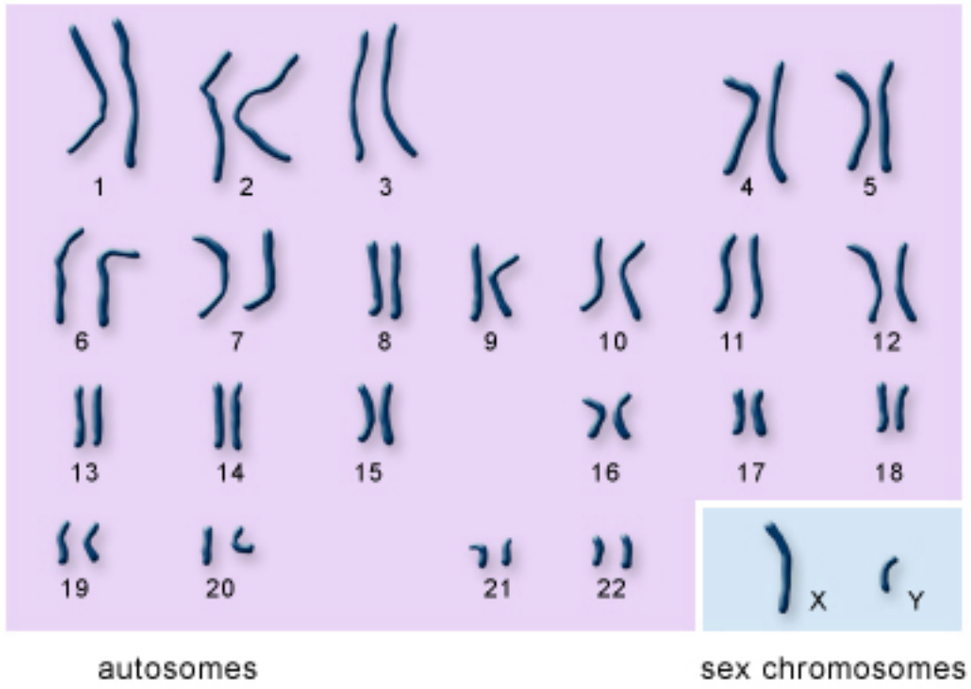




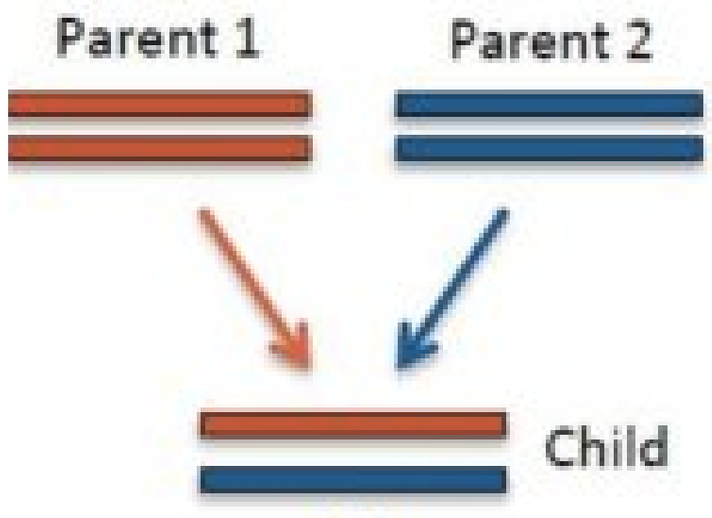
The chromosome painting collective

Genetics and inheritance

One member of each pair of homologous chromosomes comes from the father (paternal) and the other from the mother (maternal)



In males, Y from father and X from mother



The chromosome painting collective

Outline

Molecular biology : How does the genome code for function?

Genetics: How is the genome passed on from parent to child ?

Genetic variation: How does the genome change when it is passed on ?

What can we learn from genetic variation ?

Genome sequencing: How do we read the genome ?

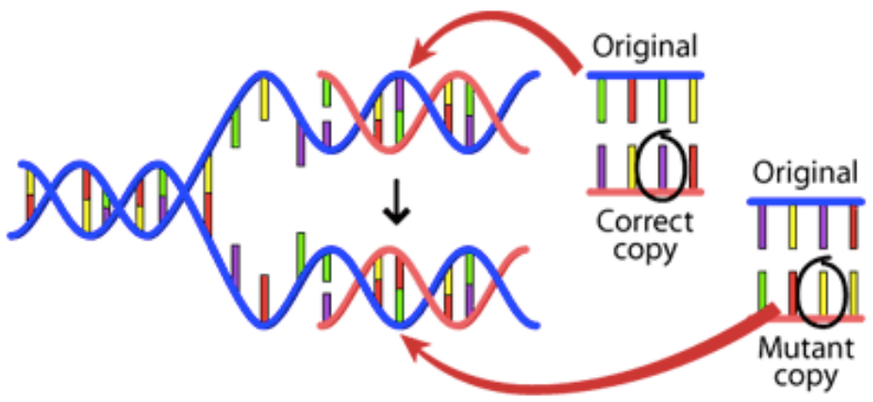
Causes of genetic variation

DNA not always inherited accurately

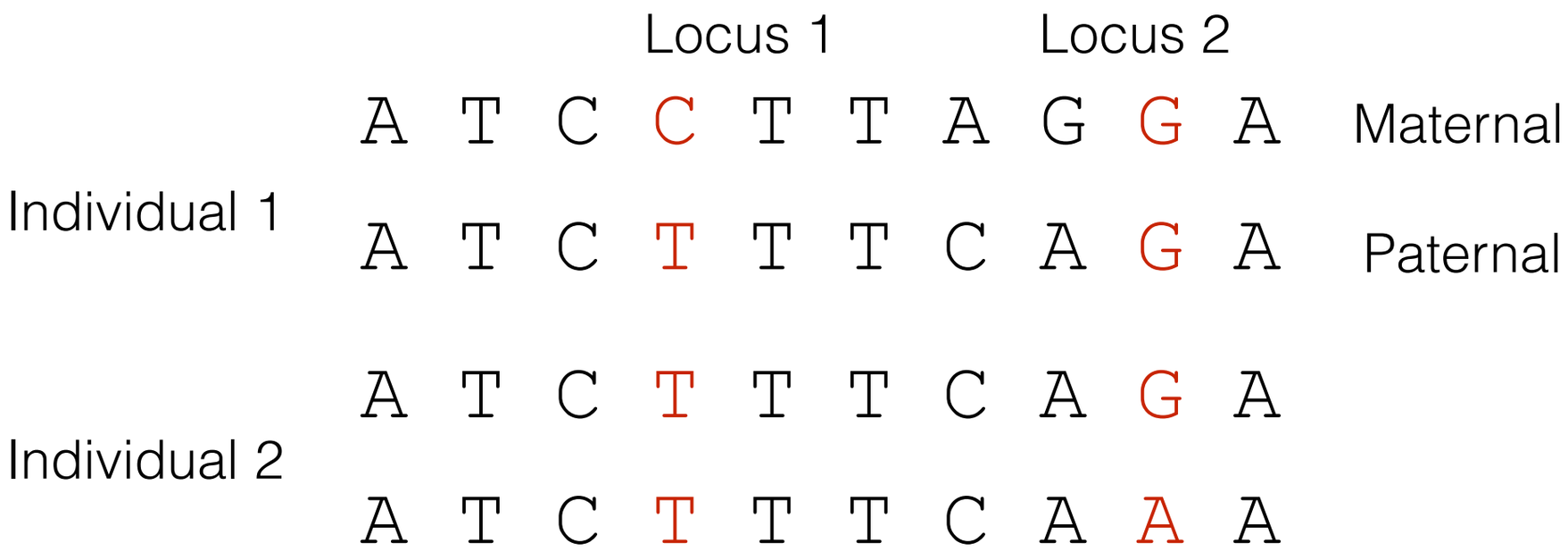
Mutations: changes in DNA

Changes at a single base (single nucleotide)

Can have more complex changes



More definitions



Locus: position along the chromosome (could be a single base or longer).

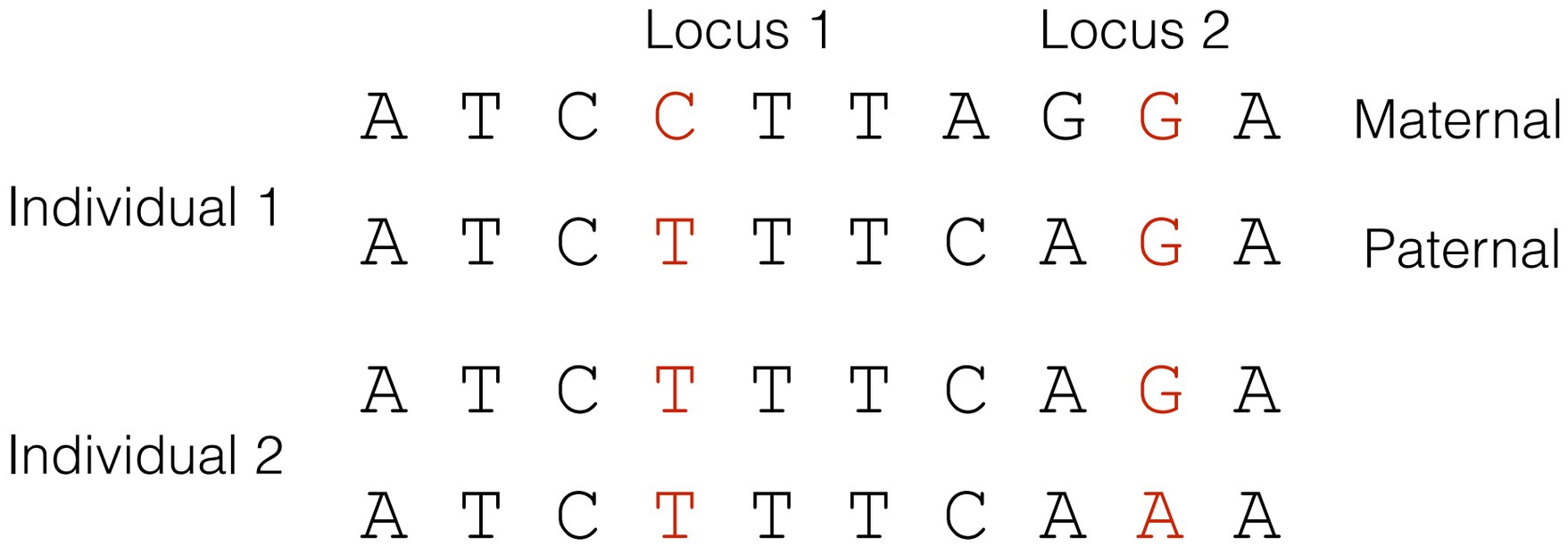
Allele: set of variants at a locus

Genotype: sequence of alleles along the loci of an individual

Individual 1: (1,CT),(2,GG)

Individual 2: (1,TT), (2,GA)

Single Nucleotide Polymorphism (SNP)



Form the basis of most genetic analyses

Easy to study in high-throughput (million at a time)

Common (80 million SNPs discovered in 2500 individuals)

Two human chromosomes have a SNP every ~1000 bases

Single Nucleotide Polymorphism (SNP)

	Locus 1						Locus 2			
Individual 1	A	T	C	C	T	T	A	G	G	A	Maternal
	A	T	C	T	T	T	C	A	G	A	Paternal
Individual 2	A	T	C	T	T	T	C	A	G	A	
	A	T	C	T	T	T	C	A	A	A	

Most SNPs are biallelic.

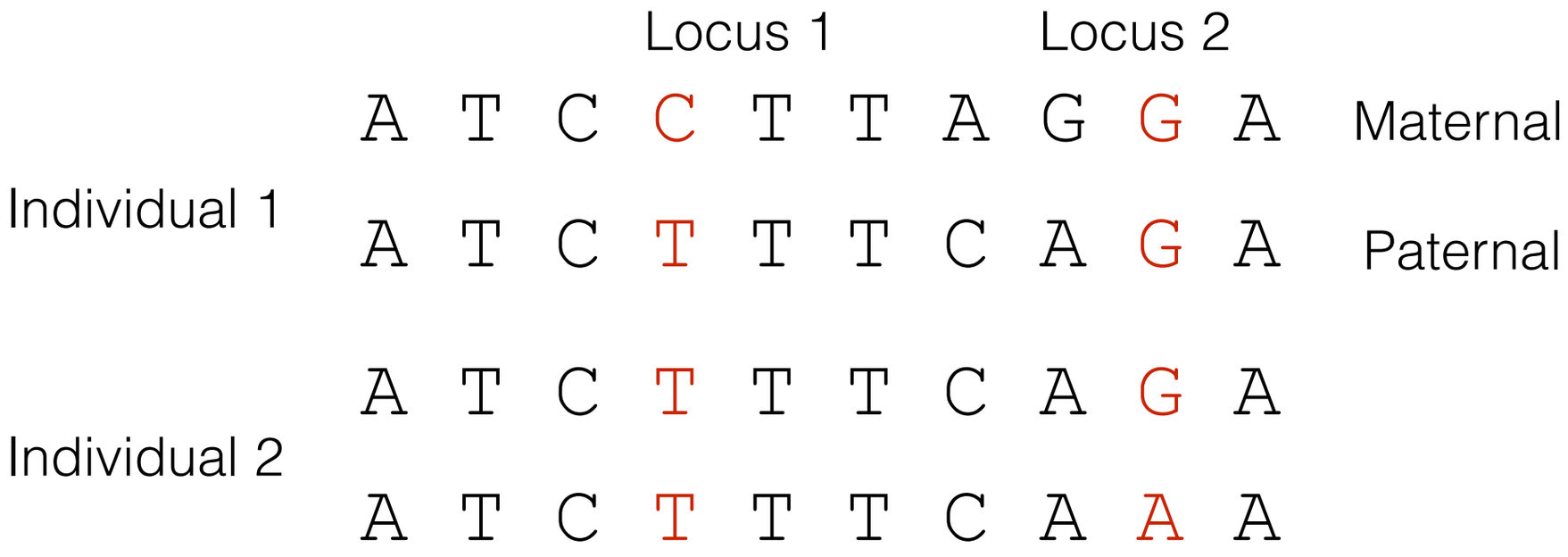
Pick one allele as the reference allele.

Can represent a genotype as the number of copies of the reference allele.

Each genotype at a single base can be 0/1/2

Locus 1:C is reference
Individual 1 has genotype 1
Individual 2 has genotype 0

Single Nucleotide Polymorphism (SNP)



Form the basis of most genetic analyses

Easy to study in high-throughput

SNP arrays have millions of common SNPs

Common (80 million SNPs discovered in 2500 individuals)

Genotype and phenotype

Phenotype = function(Genotype, Environment)

Twins have similar phenotype

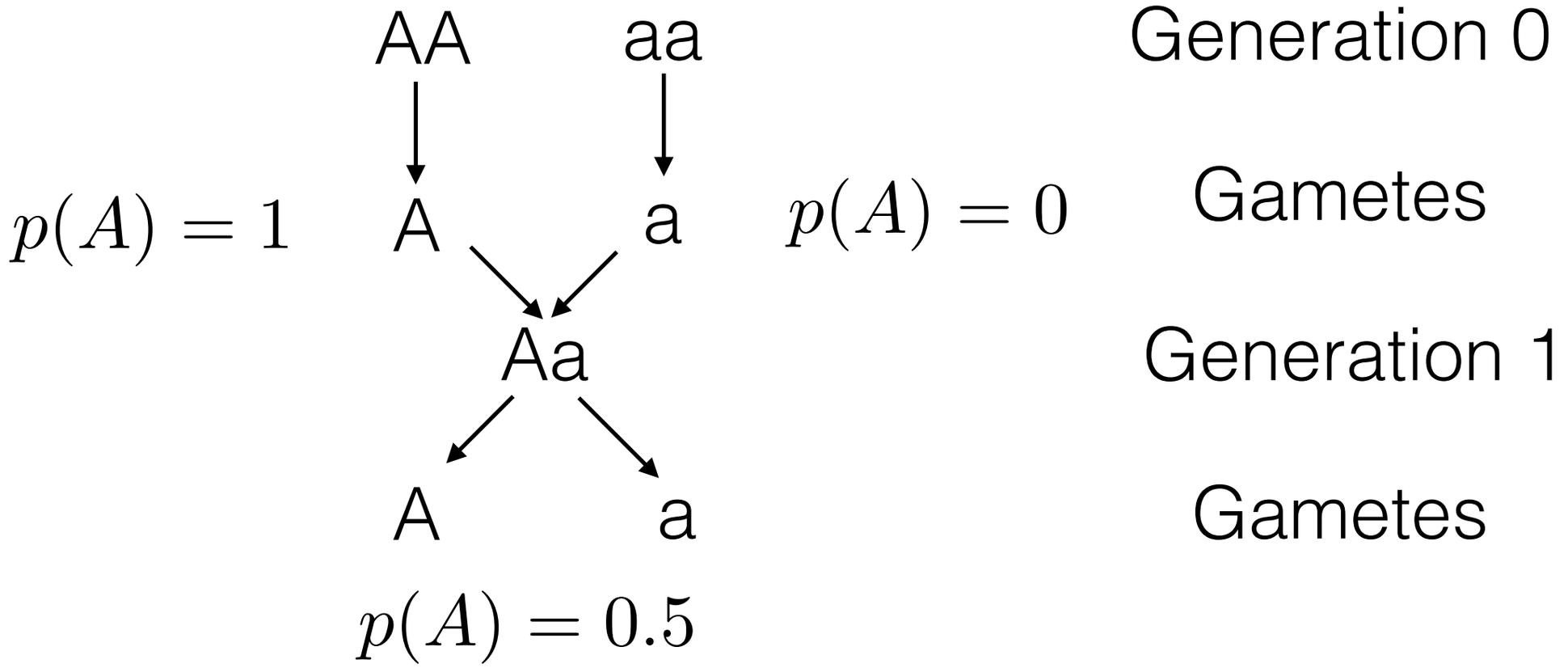
Identical twins (same genotype) can have different phenotypes

~30% are concordant for asthma, depression

Huang et al. Genetics in Medicine 2000

Back to genetic inheritance

Segregation (Mendel's first law)



Back to genetic inheritance

Segregation (Mendel's first law)

AA X aa

Aa

1.0

AA X Aa Generation 0

AA Aa Generation 1

0.5 0.5

Aa X Aa Generation 0

AA Aa aa Generation 1

0.25 0.50 0.25

Back to genetic inheritance

Assortment (Mendel's second law)

Locus 1	Locus 2			
AaBb				Generation 0
AB	Ab	aB	ab	Gametes
0.25	0.25	0.25	0.25	

Back to genetic inheritance

Assortment (Mendel's second law)

Not quite

AaBb

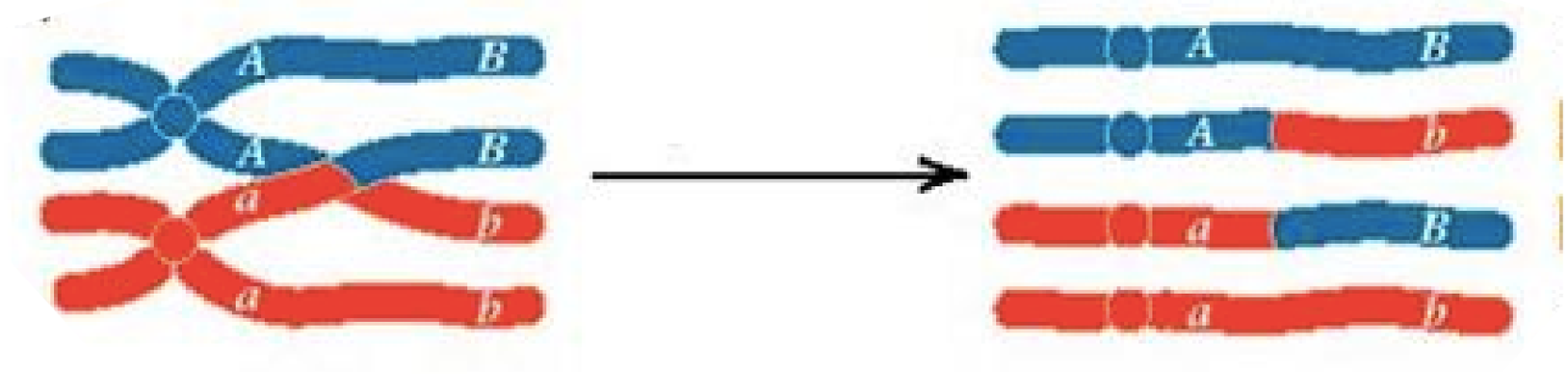
Generation 0

AB	Ab	aB	ab	Gametes
0.25	0.25	0.25	0.25	

Back to genetic inheritance

Assortment (Mendel's second law)

Not quite. Crossover recombination



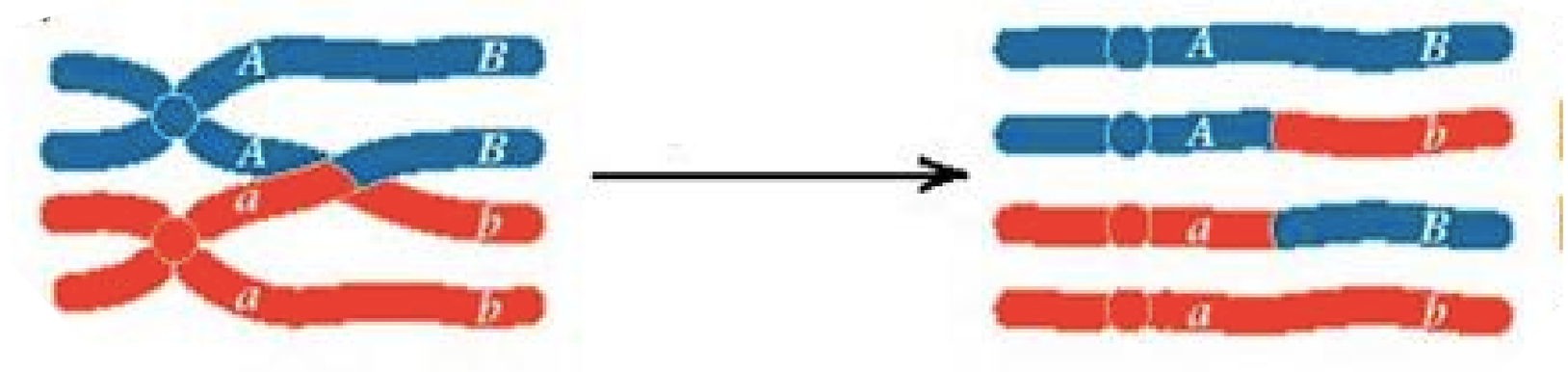
AB ab Ab aB

Parental Recombinant

Back to genetic inheritance

Assortment (Mendel's second law)

Not quite. Crossover recombination



AB ab Ab aB

(1-r)/2 (1-r)/2 r/2 r/2

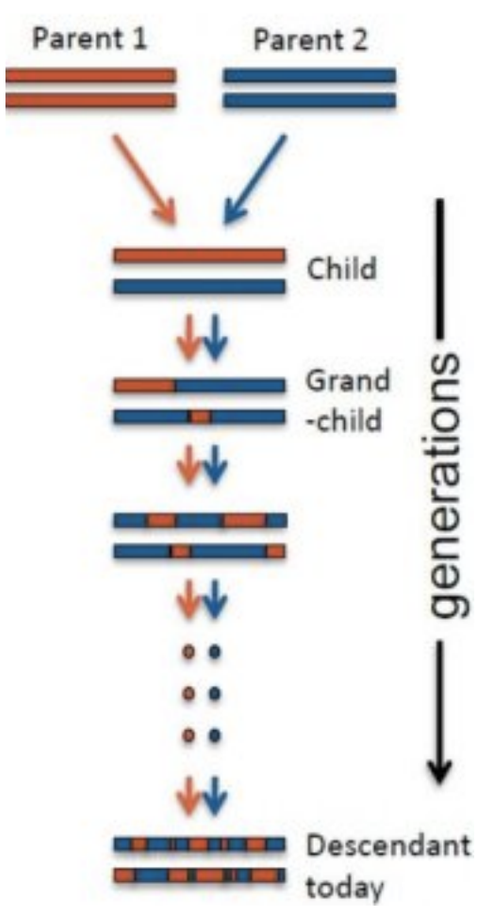
r: recombination fraction (0<=r<=1/2)

Back to genetic inheritance

Assortment (Mendel's second law)

Linkage: Positions nearby inherited together.

Important idea for mapping disease genes.



The chromosome painting collective

Back to genetic inheritance

Mutation and recombination (among other forces that we will learn about later) produce genetic variation

Mutation produces differences

Recombination shuffles these differences

Outline

Molecular biology : How does the genome code for function?

Genetics: How is the genome passed on from parent to child ?

Genetic variation: How does the genome change when it is passed on ?

What can we learn from genetic variation ?

Genome sequencing: How do we read the genome ?

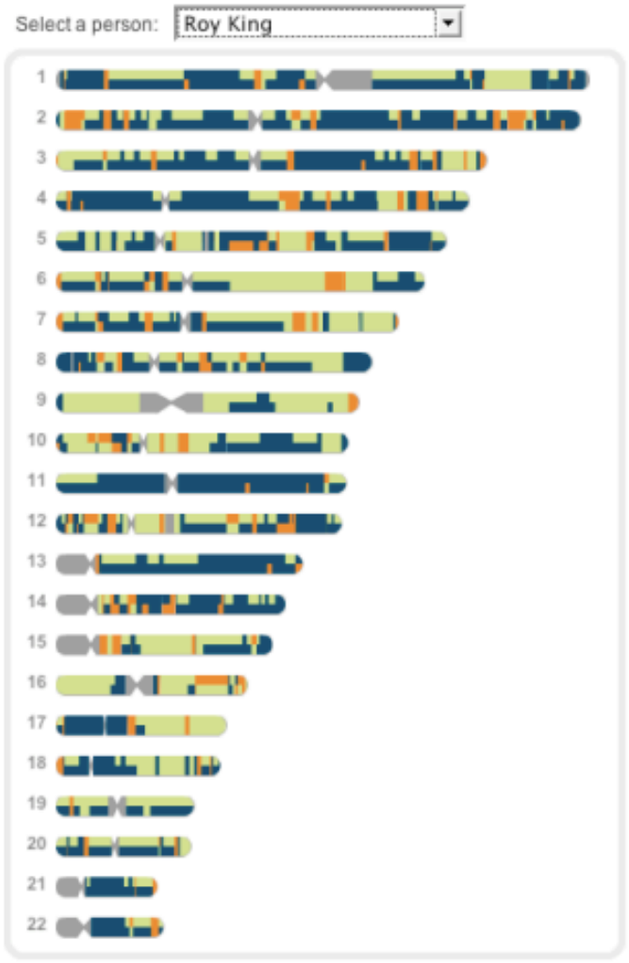
What can we learn from genetic variation ?

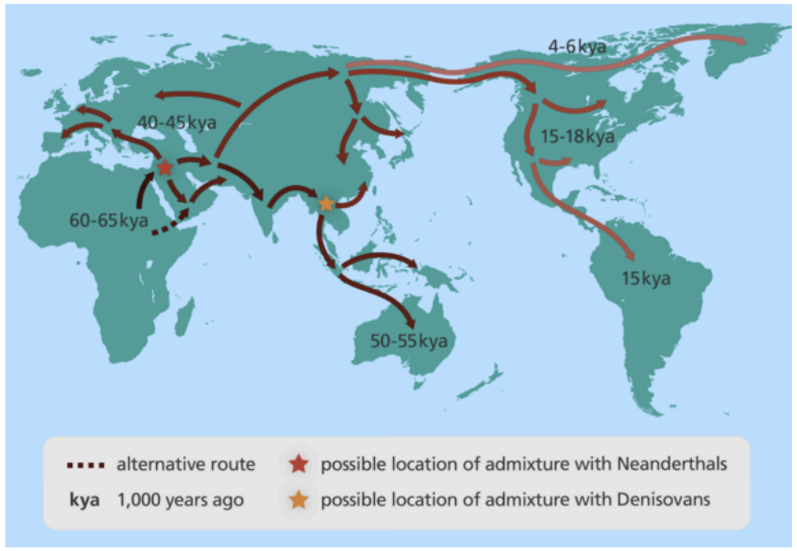
Evolution and history

Biological function and disease

What can we learn from genetic variation ?

History of human populations learned from genetics





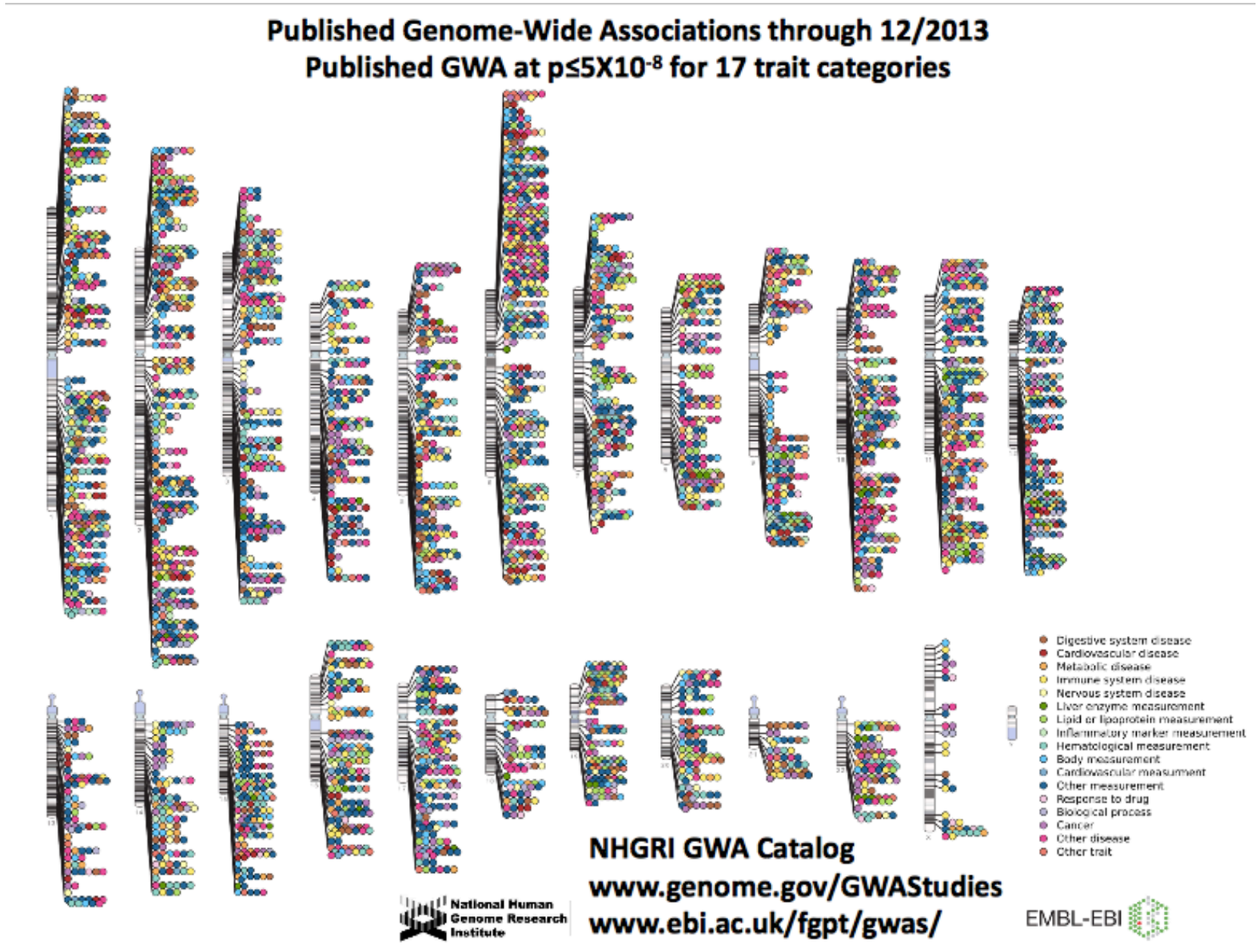
https://blog.23andme.com/23andme-and-you/genetics-101/a-beautiful-ancestry-painting/

What can we learn from genetic variation ?

Evolution and history

Biological function and disease

Genome-wide Association Studies (GWAS)



Outline

Molecular biology : How does the genome code for function?

Genetics: How is the genome passed on from parent to child ?

Genetic variation: How does the genome change when it is passed on ?

What can we learn from genetic variation ?

Genome sequencing: How do we read the genome ?

Reading the genome

Goal: Determine the sequence of bases along each chromosome

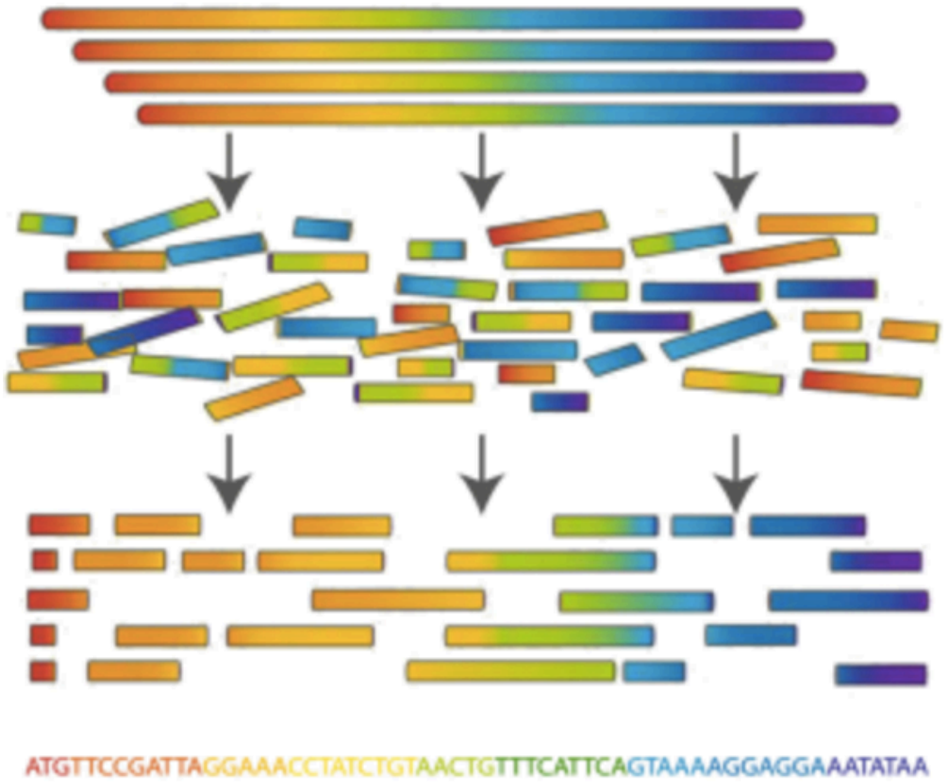
Fragment the chromosomes

Read each fragment

Assemble the fragments

Details depend on the technology

Computationally hard



The human genome project

Goals

Sequence an accurate reference human genome

Find the set of all genes

Draft published in 2001

High-quality version completed in 2003

Cost: ~$3 billion.

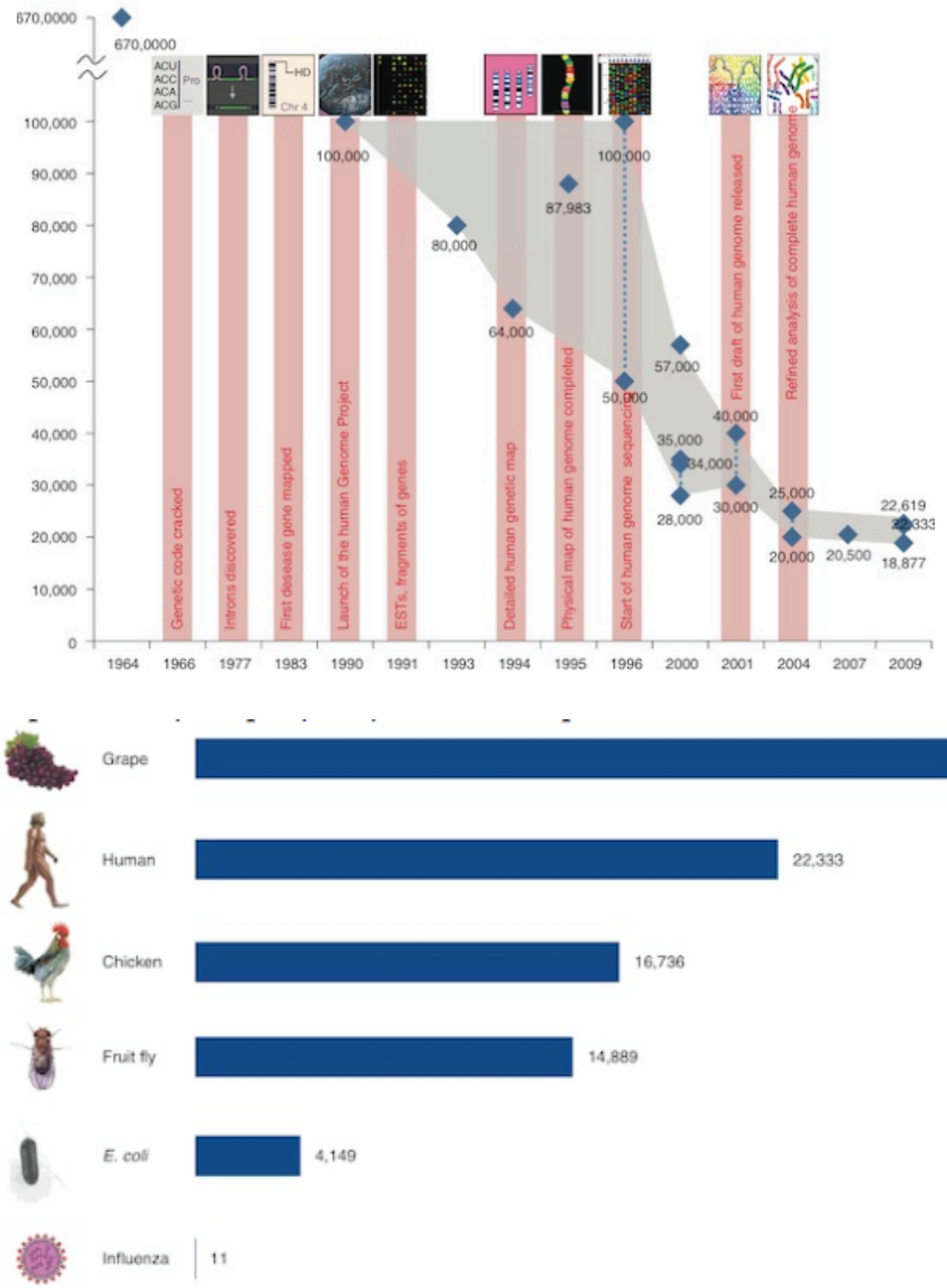
Time: ~13 years.

Two competing groups (public and private)

The human genome project

Major findings

Fewer genes than previously thought (~20K)



Pertea and Salzberg, *Genome Biology* 2010

The human genome project

Other outcomes

International collaborations

Power of computing

Reading the genome

Human genome provides a reference

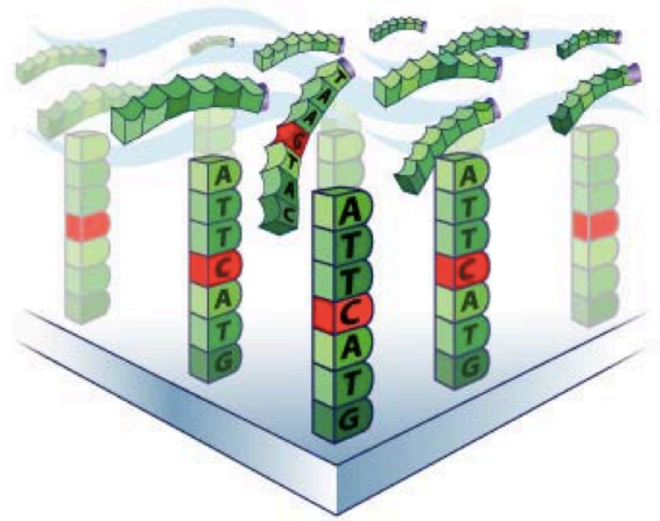
Humans share most of their genome (~99.9%)

Can focus on reading the differences

Reading the genome

A T C C T T A G G A

A T C T T T C A G A



High-throughput genotyping

Hybridization of DNA molecules

Nucleotides bind to their complementary bases

A=T, C=G

Can be used to get the genotype at a chosen set of SNPs

Maps of genetic variation

International HapMap Project

Goals: Describe common patterns of genetic variation in human populations

Phase 1: Genotyped ~1 million SNPs from 270 individuals in 4 populations.
Aims to capture all SNPs with a frequency of >5%.

Phase 3: 7 additional populations included

All data publicly available.

Reading the genome

Limitations of genotyping

Can only read SNPs that are on the chip

Biased by how these SNPs are chosen (e.g. common SNPs)

Reading the genome

High-throughput (or next-gen) sequencing

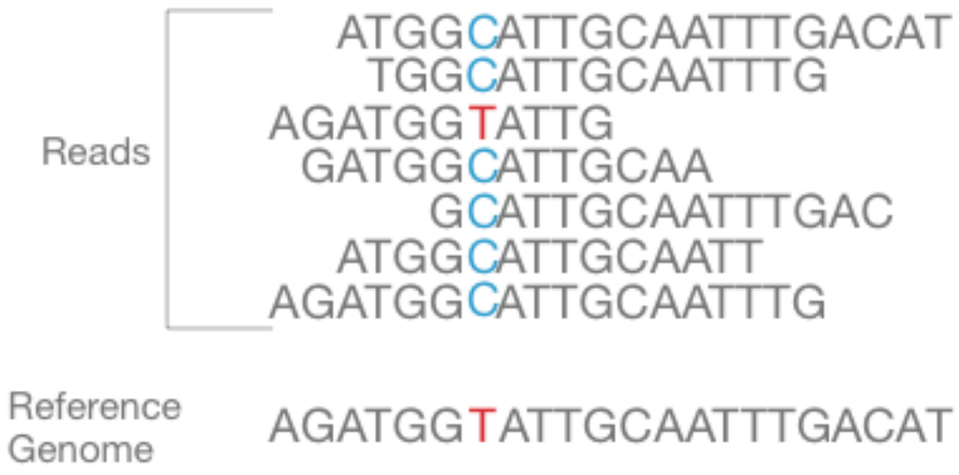
Technologies: Illumina, IonTorrent, PacBio

Can read small pieces of the genome (~100bp)

Two major differences

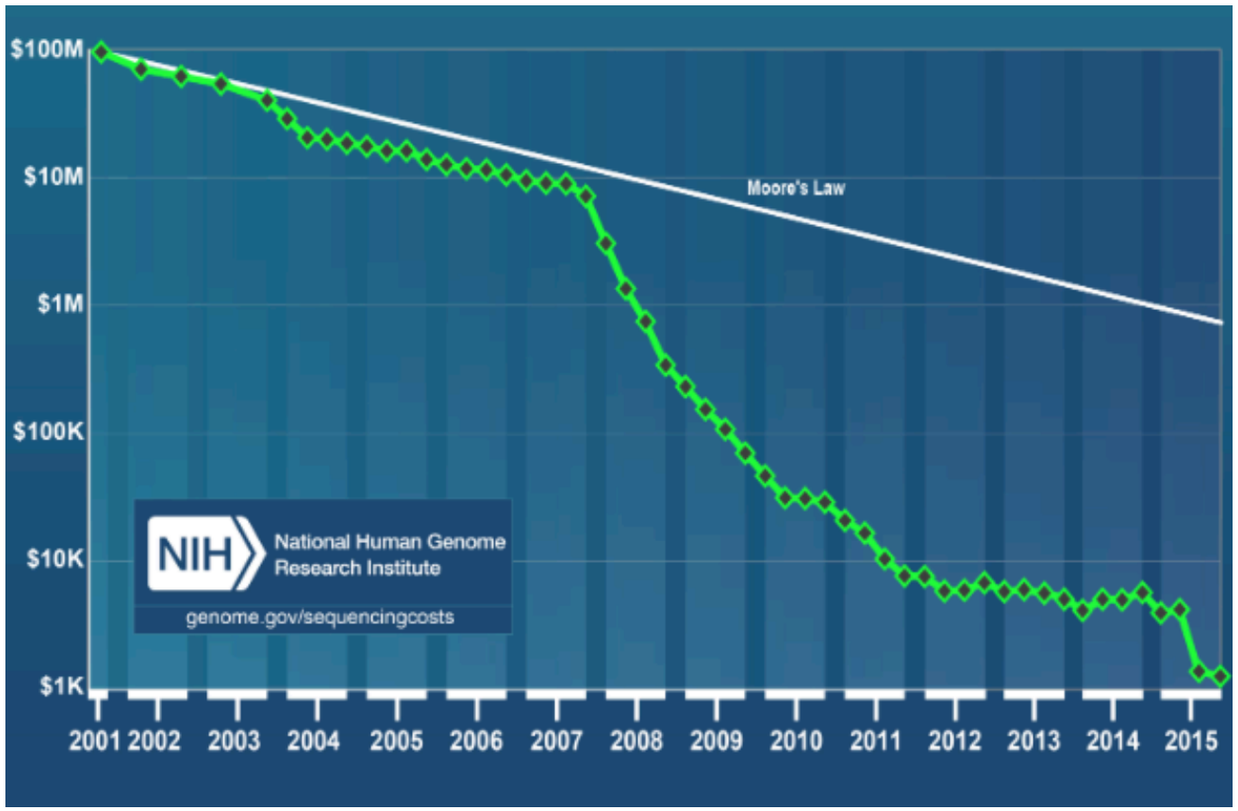
Sequence hundreds of thousands of fragments in parallel

Use the reference human genome to **find** the locations of the reads (and to infer mutations)



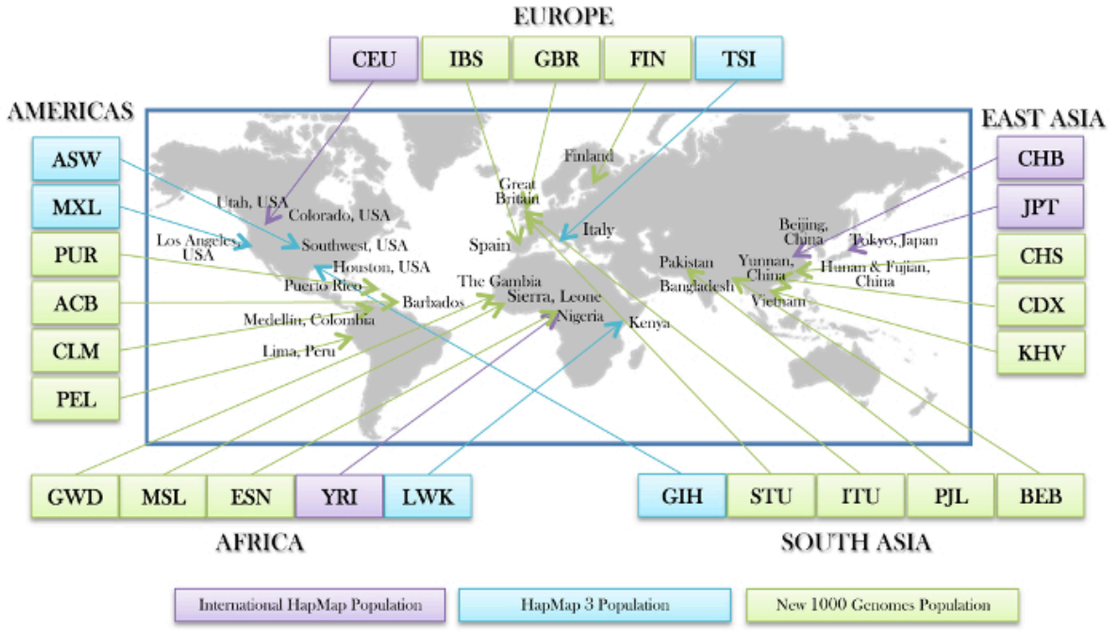
Reading the genome

Cost of genome sequencing



Maps of genetic variation

1000 Genomes Project



2500 individuals from 26 populations

Discover ~90 million SNPs

Includes >99% of SNPs with frequency >1%

All data publicly available

Maps of genetic variation

Many more such efforts underway

Example: Simons Genome Diversity Project : 260 genomes from 127 populations

Also publicly available

Other interesting data

EXAC data: Exomes from ~60,000 individuals

Also publicly available

UK Biobank: 500,000 individuals with 200 phenotypes

Not publicly available

This class

Overview of the biological questions

Basic concepts in genetics

Genomes are inherited according to well-known rules (Mendel's laws)

Genomes change

Genetic variation forms the starting point for inference.

Possible inferences: history, disease risk and many more

Advances in technology are allowing us to read many more genomes