Models and Inference in population genetics

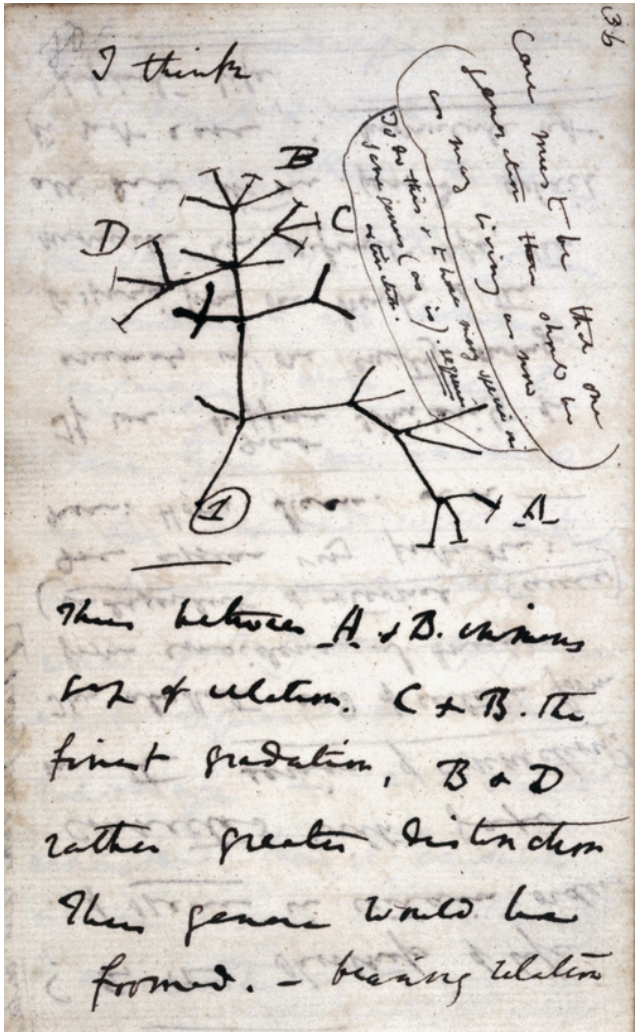
Bocconi University, 11 – 25 October 2021

Dario Spano`
D.Spano@warwick.ac.uk

1. What is population genetics about ?

An old modern ambition: understanding the mechanisms responsible for life evolution and inheritance.

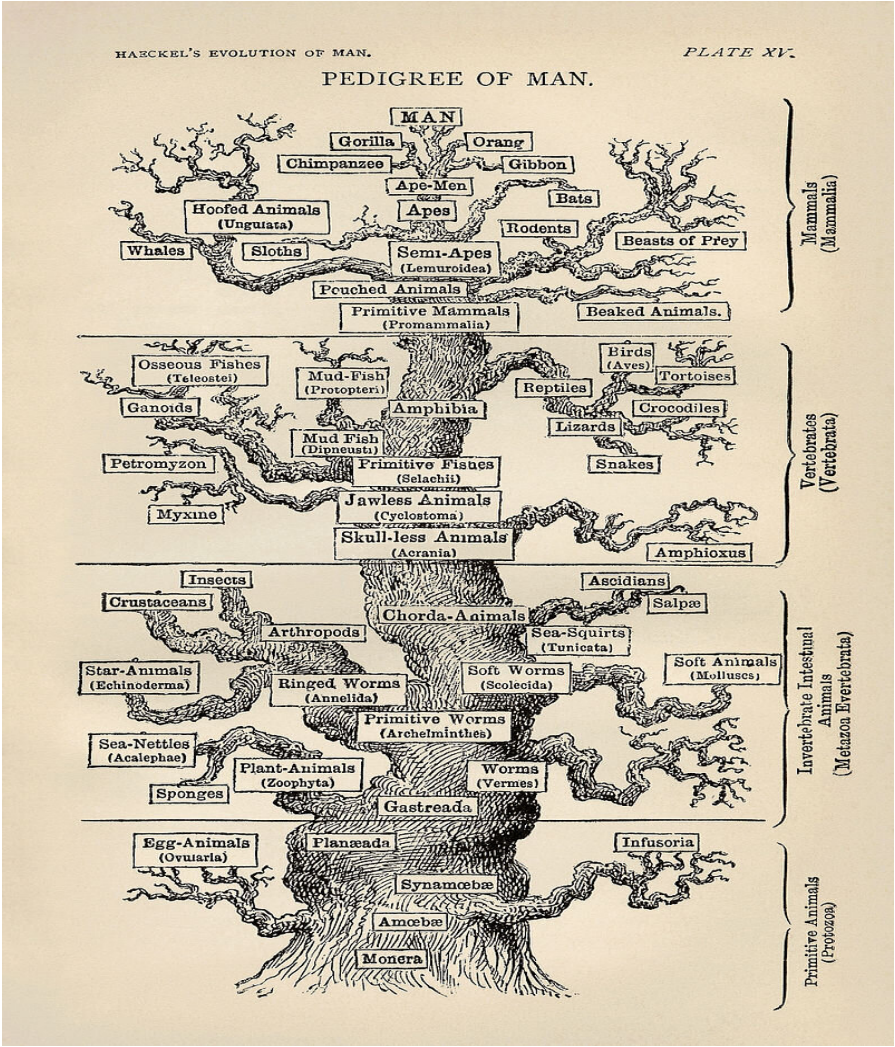
- “Will my family survive?”
- “How does a population becomes extinct?”
- “Where do we come from?”
- “What is the origin of species?”



Picture: Charles Darwin's 1837 sketch, his first diagram of an evolutionary tree from his First Notebook on Transmutation of Species (1837).

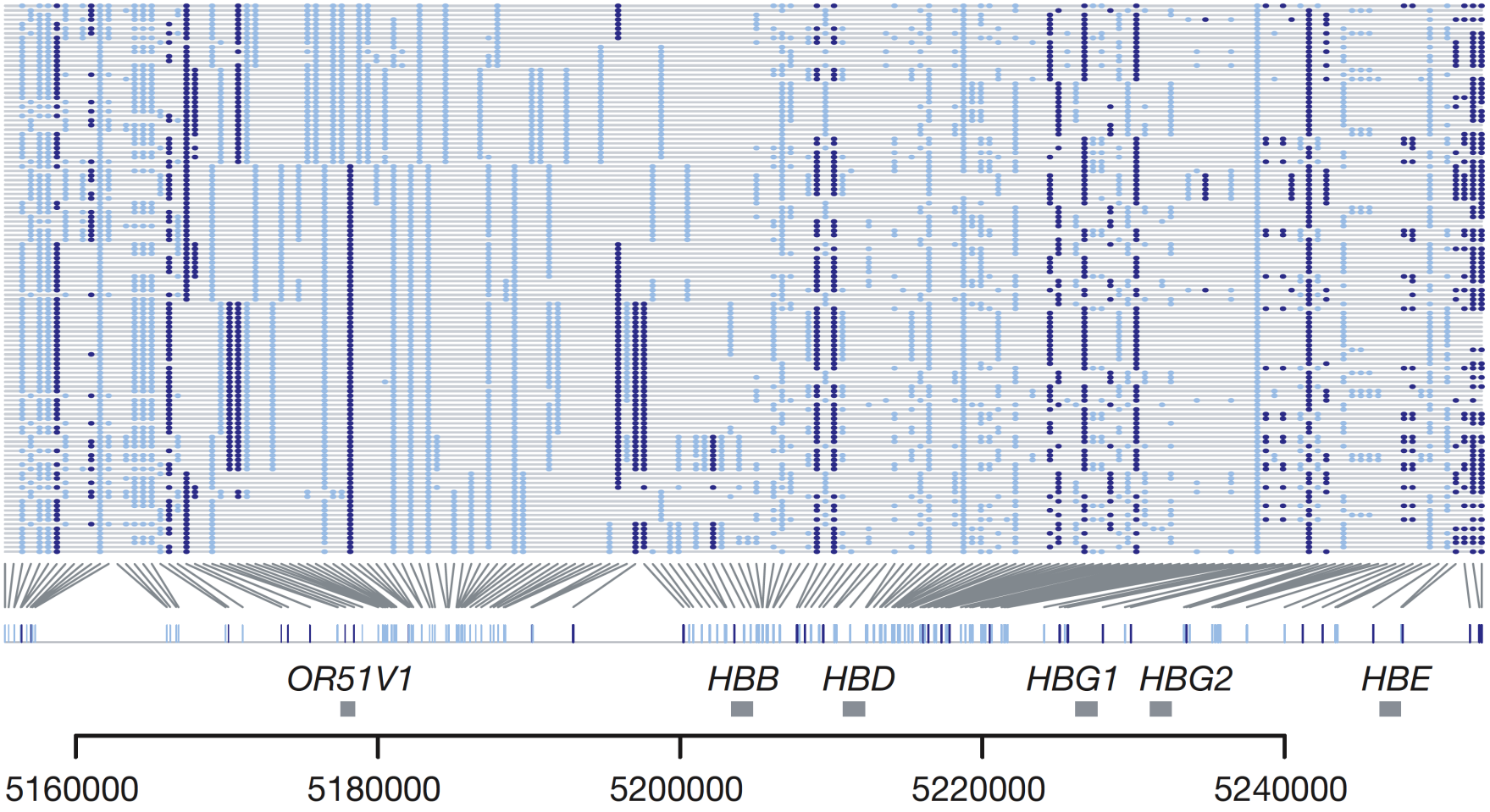
1. What is population genetics about ?

- Accepted idea of relatedness of species
- How to define species ?
- How to measure evolutionary distance?
- Are humans really pinnacle of evolution?
- How to measure genetic time?



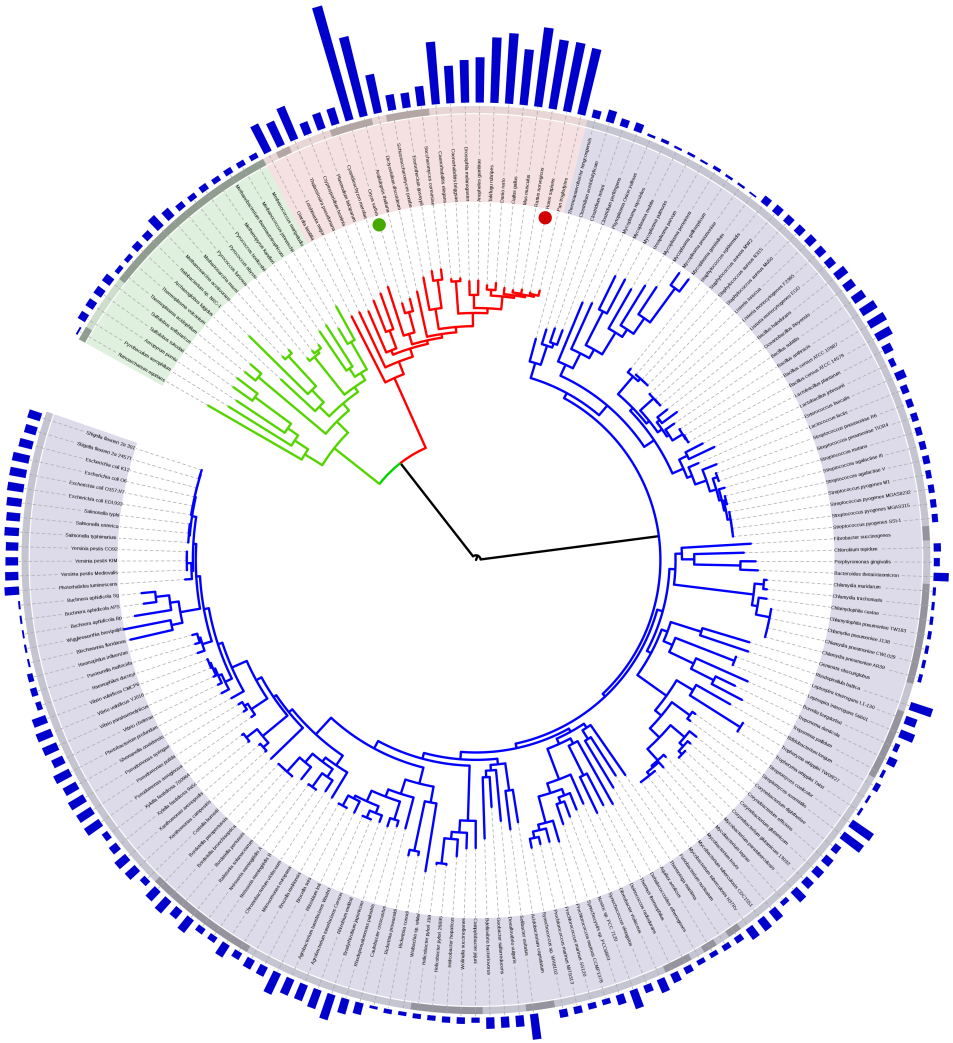
Picture: Ernst Haeckel's tree from the The Evolution of Man (Published 1879)

Genetic data increasingly more refined



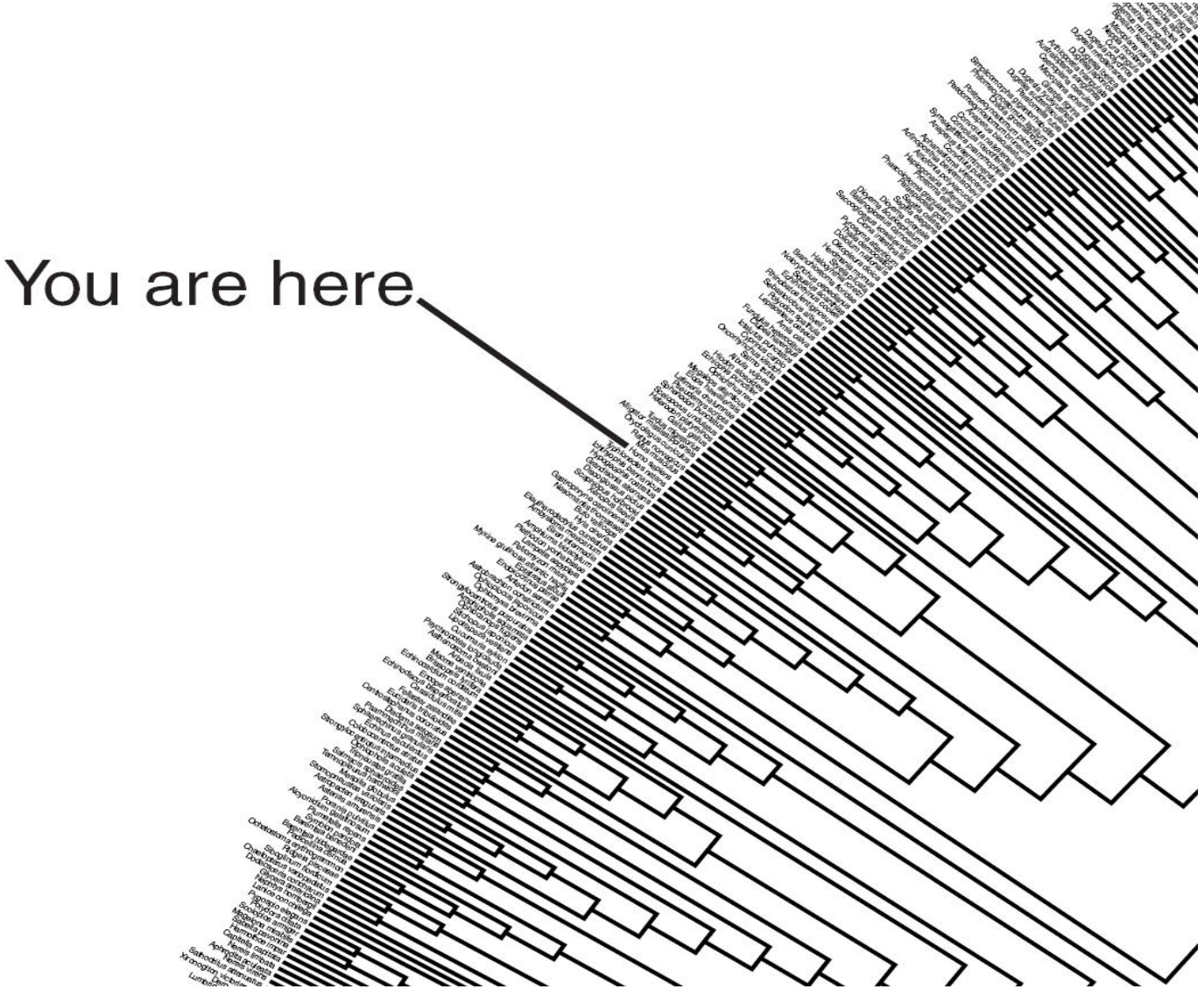
International HapMap Consortium (2007)

1. A huge classification task.



Picture: HD Tree of life from genome sequencing
https://www.pngitem.com/middle/hommwi_tree-of-life-genome-sequencing-hd-png-download/

Tree of life



Aim of this mini-course

- To understand some fundamental models of evolution:
  - Allele-frequencies: forward in time
  - Genealogy: backward in time
- To provide an answer to some key questions:
  - What is the probability that a genetic type will eventually disappear, or that a new mutation will prevail ?
  - What is the time of the most recent common ancestor of a collection of “genes”?
- To understand how to estimate relevant quantities such as rate of mutation, age of an allele, etc. NOT JUST CLASSIFICATION.

Two important objects

To understand basic models of evolution:

- Allele-frequencies: forward in time diffusion process.
- Genealogy: backward in time coalescent trees.

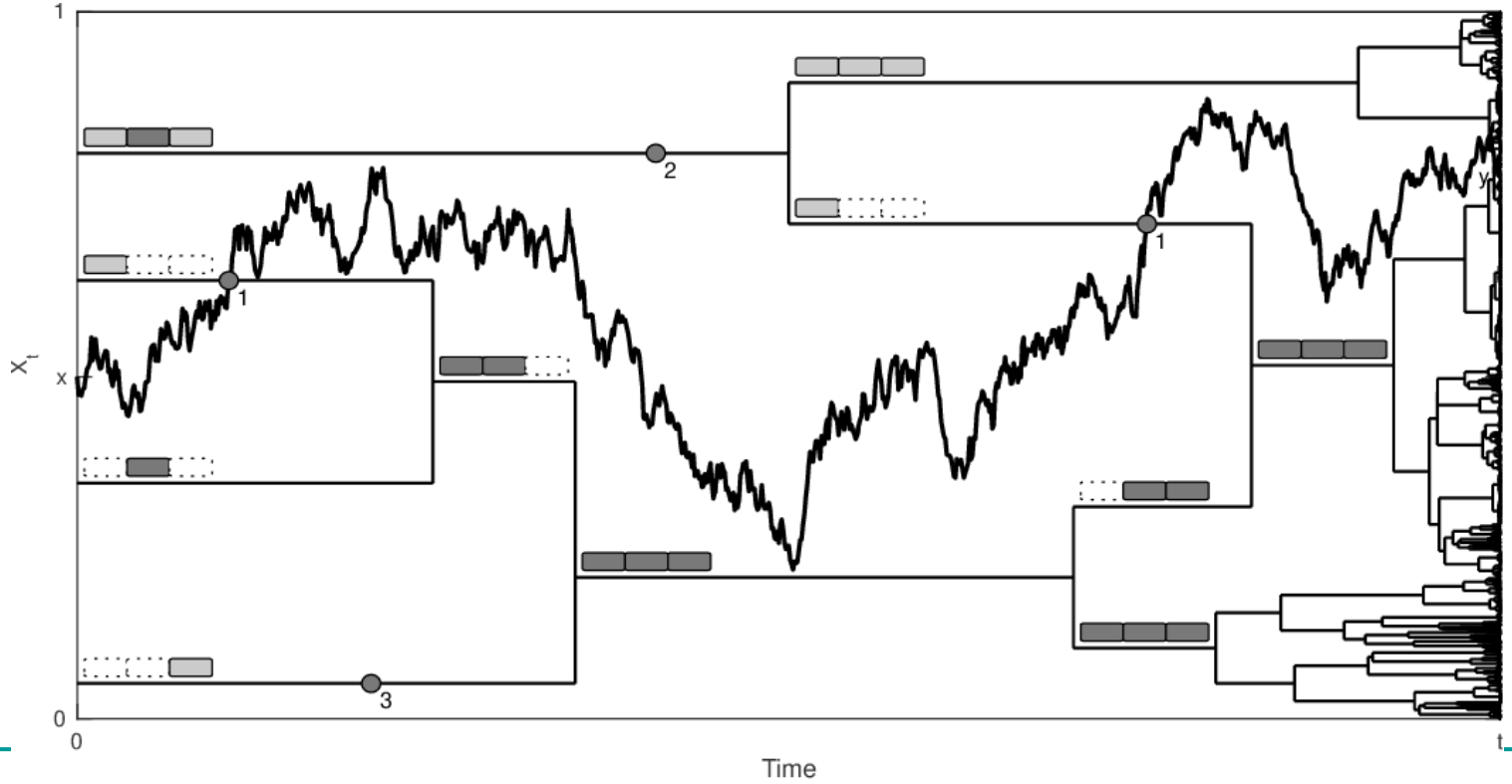


Image: Griffiths et al. (2016), Theoretical Population Biology, 112 (2016), 126–138

1. Introduction to genetic data

- This minicourse does not assume any knowledge of biology, but does contain a lot of biological jargon.
- We will first familiarise with some key terminology and notions from genetics.
- Further reference:
  - Molecular Biology of the Cell, 4th edition. Free online at http://www.ncbi.nlm.nih.gov/books/NBK21054/

The human genome

- 23 pairs of chromosomes:
  - Compose the human ***genome***
  - Found in the nucleus of each cell
  - Pairs show high similarity; they are ***homologous***
  - One pair of sex chromosomes and 22 pairs of ***autosomes***
  - In humans, an XY sex-determination system:
    - Females are XX
    - Males are XY.

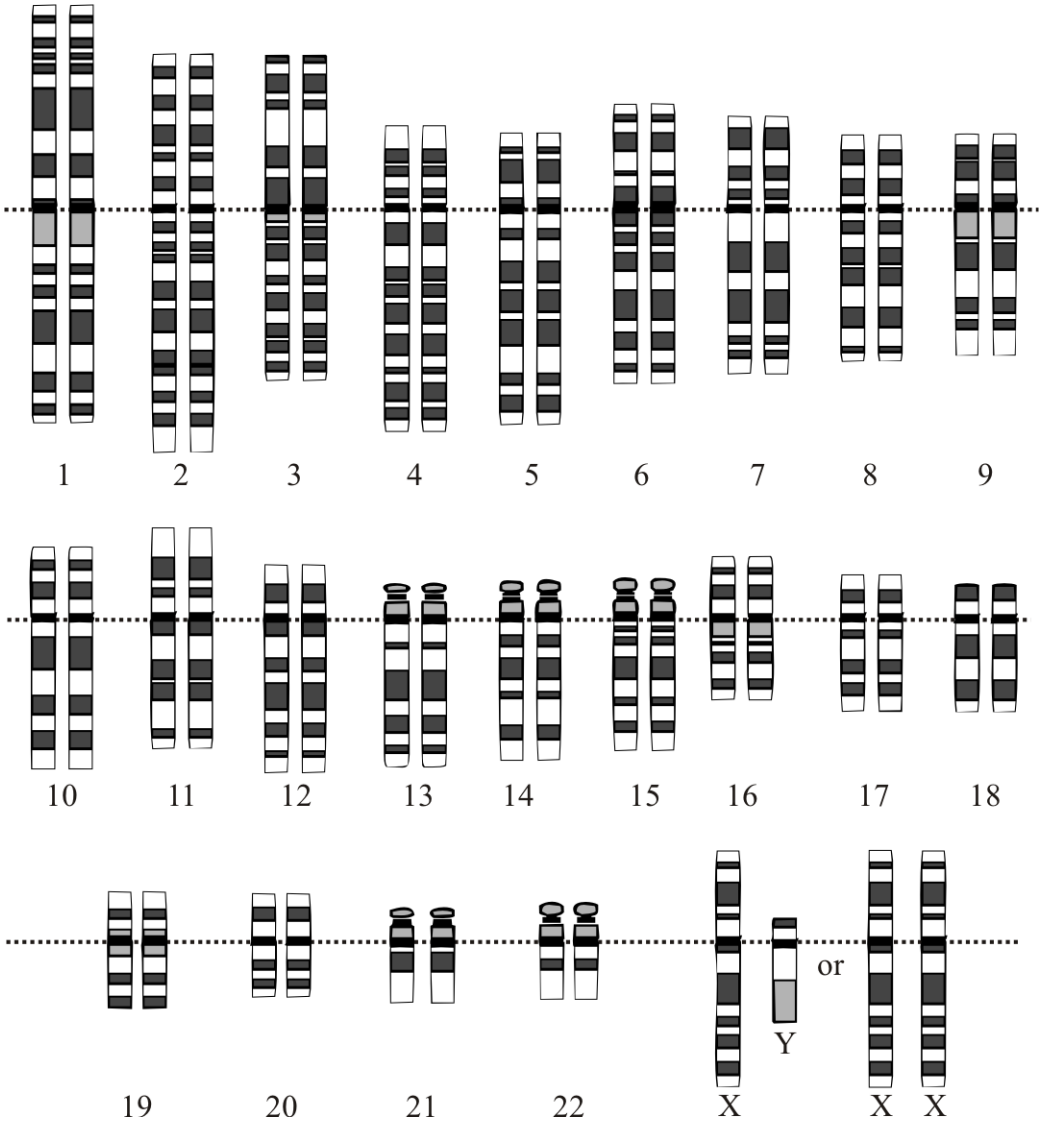


Image: Molecular biology of the cell, 4th edition.

Ploidy

- Humans (and many plants and animals) are ***diploid*** – cells contain two copies of each chromosome.
  - But sex cells, or gametes (sperm and ovum) are ***haploid*** – carrying one unpaired copy.
  - In sexual reproduction gametes unite randomly in ***fertilization*** to produce a ***zygote***.
  - The sex of the zygote is determined by the sex chromosome carried by the sperm.

- We are generally mostly be interested in ***dioecious*** organisms – those with two sexes, each contributing one gamete to offspring, but some models are simpler if we assume a ***monoecious*** organism.

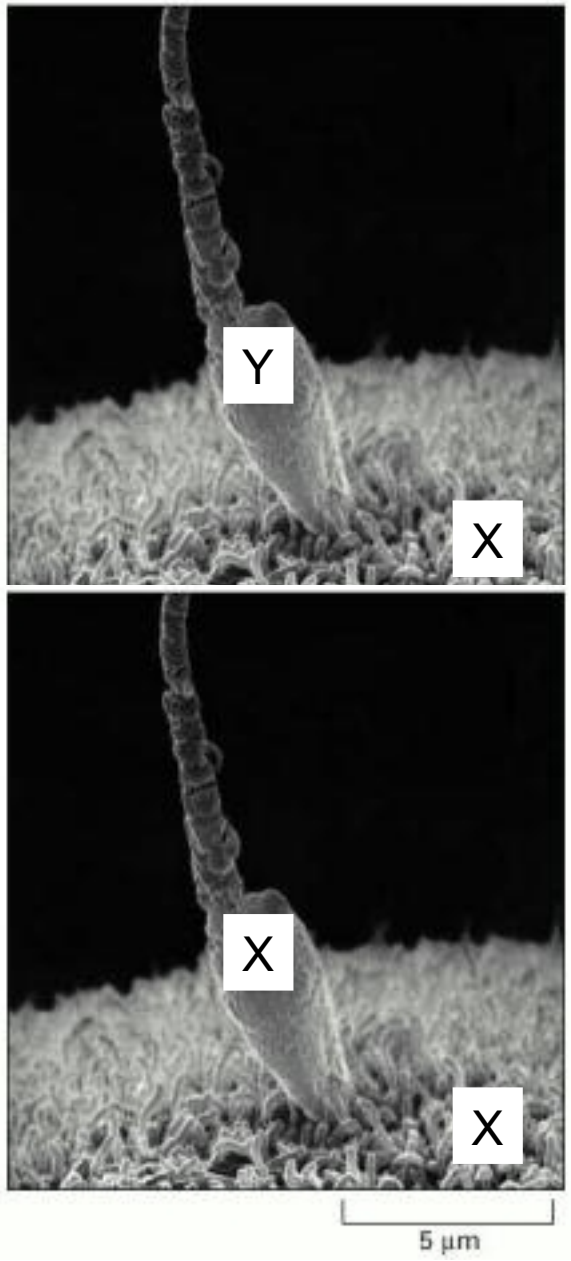


Image: Molecular biology of the cell, 4th edition.

Genes

- ***Locus***: a region of interest on a chromosome.
- ***Gene***: (roughly) a contiguous region of DNA responsible for making a protein.
  - ~20,000 genes in humans
  - <2% of the human genome
  - Mean length 3 kilobases (kb)

- Alternative forms of genes (or loci) are known as ***alleles***.
- Loci at which more than one allele has been observed are said to be ***polymorphic***.

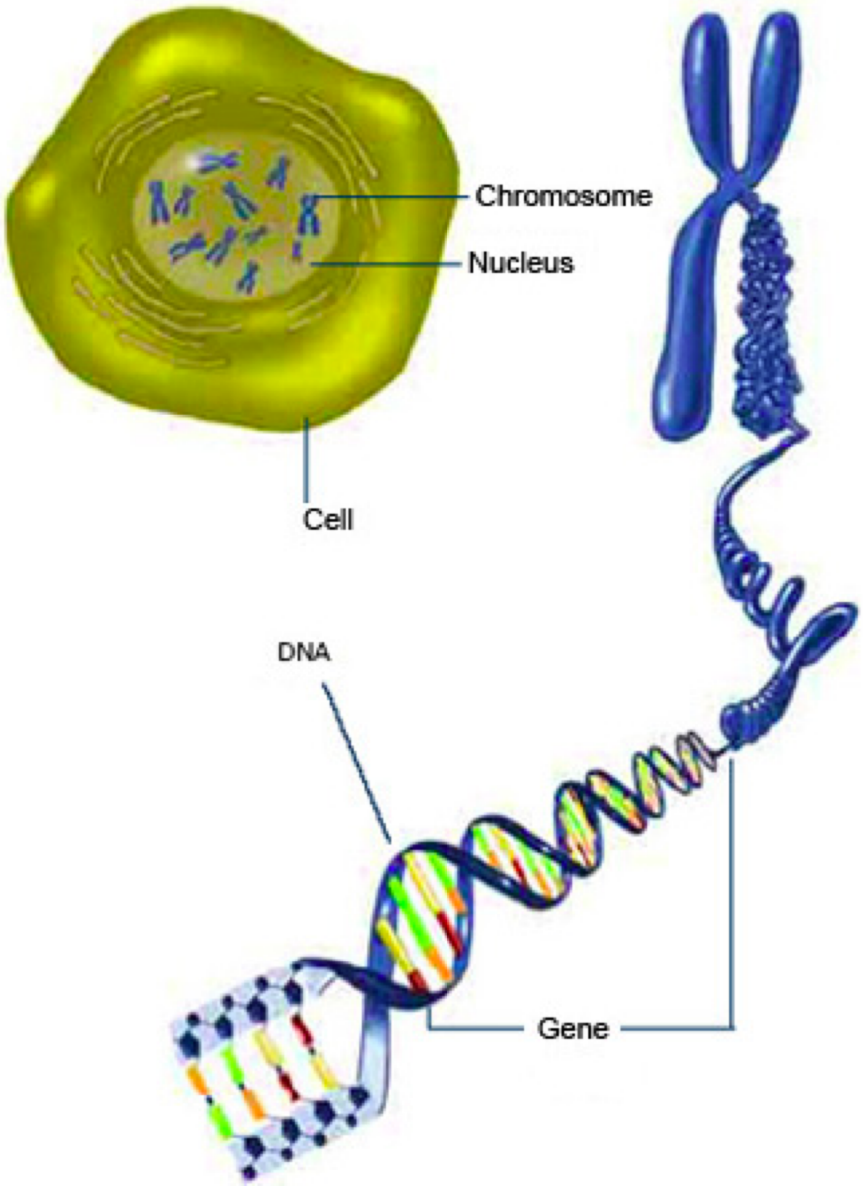


Image: NIH.

What do we mean by an alternative allele?

- ***Single nucleotide polymorphisms (SNPs)*** – variation observed at a single DNA nucleotide (i.e. a single letter at one site).
- ***Insertions/deletions (Indels)*** – variation in which base pairs are inserted or deleted at a locus in some individuals.
- ***Tandem repeats*** – a short sequence repeated a variable number of times, e.g. ATATAT versus ATATATATATAT.
- ***Structural variants*** – large chromosomal rearrangements, duplications, translocations, inversions, insertions or deletion.

What do we mean by an alternative allele?

- ***Single nucleotide polymorphisms (SNPs)*** – variation observed at a single DNA nucleotide (i.e. a single letter at one site).
- ***Insertions/deletions (Indels)*** – variation in which base pairs are inserted or deleted at a locus in some individuals.
- ***Tandem repeats*** – a short sequence repeated a variable number of times, e.g. ATATAT versus ATATATATATAT.
- ***Structural variants*** – large chromosomal rearrangements, duplications, translocations, inversions, insertions or deletion.

- For this course we'll only think about SNPs (but all are important!).

SNPs

1 AACG**A**GTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT

2 AACG**A**GTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT

3 AACGGGTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT

4 AACGGGTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT

5 AACGGGTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT

6 AACGGGTACTGGCTAAAGCTCGACTCGCCTACGTCAGTCTCTTT

7 AACGGGTACTGGCTAAAGCTCGACTCGCCTACGTCAGTCTCTTT

8 AACGGGTACTGGCTAAAGCTCGACTCGCCTACGTCAGTCTC**C**TT

9 AACG**A**GTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT

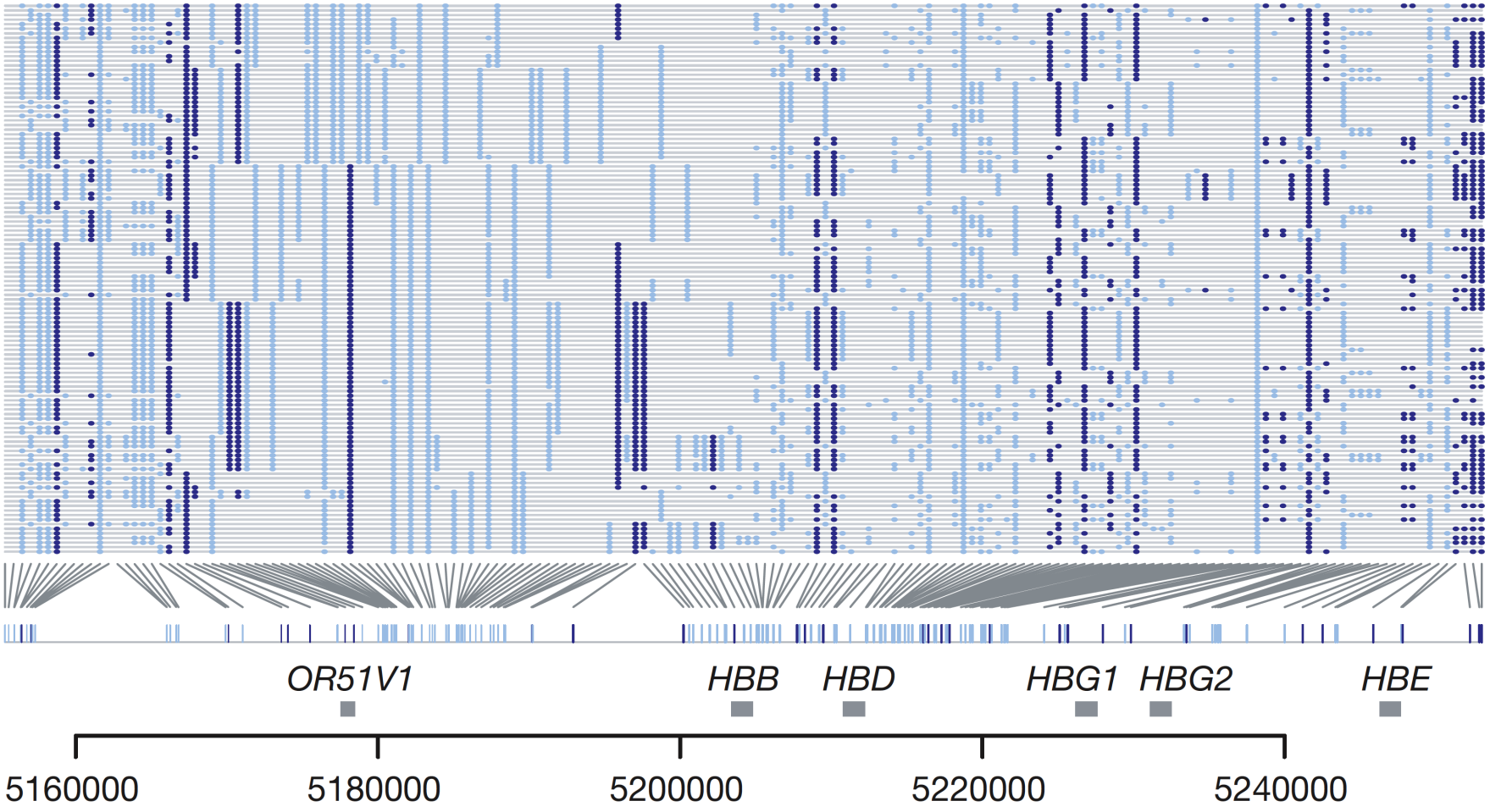
10 AACGGGTACTGGCTAAAGCTCGACTCGCCTACGTCAGTCTC**C**TT

SNPs

1 AACG**A**GTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT
2 AACG**A**GTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT
3 AACGGGTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT
4 AACGGGTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT
5 AACGGGTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT
6 AACGGGTACTGGCTAAAGCTCGACTCGCCTACGTCAGTCTCTTT
7 AACGGGTACTGGCTAAAGCTCGACTCGCCTACGTCAGTCTCTTT
8 AACGGGTACTGGCTAAAGCTCGACTCGCCTACGTCAGTCTC**C**TT
9 AACG**A**GTACTGGCTAAAGCTCGACTCGC**T**TACGTCAGTCTCTTT
10 AACGGGTACTGGCTAAAGCTCGACTCGCCTACGTCAGTCTC**C**TT

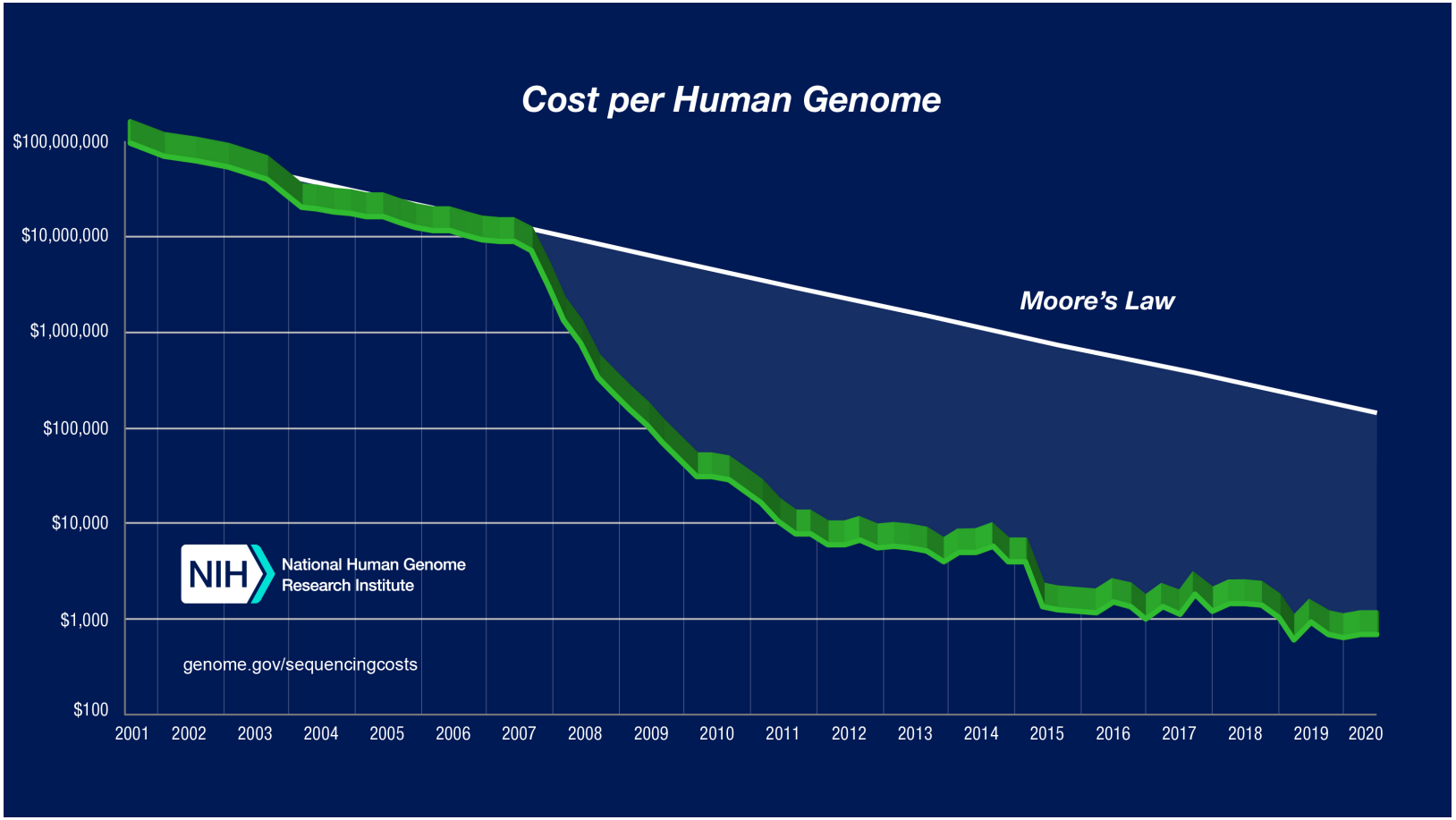
- The collection of alleles inherited together from different loci along a chromosome is called a ***haplotype***.

SNPs



International HapMap Consortium (2007)

Big (and cheap?) data



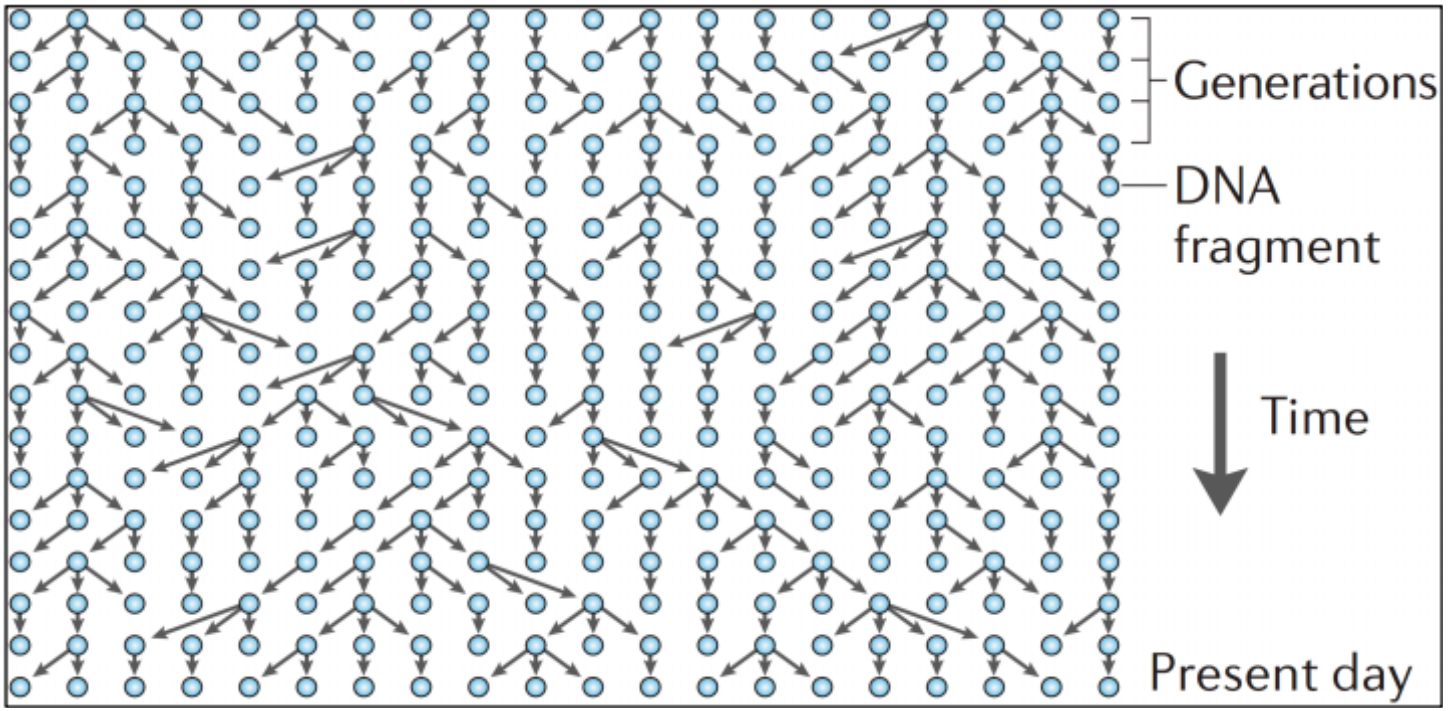
Picture: NHGRI

Genetics and data science buzzwords

- SNP or Genome arrays: BIG DATA
- Likelihood function of SNP: BIG MODELS

Key role of stochastic processes !!!

2. The Wright-Fisher model



- **[Note 2]**

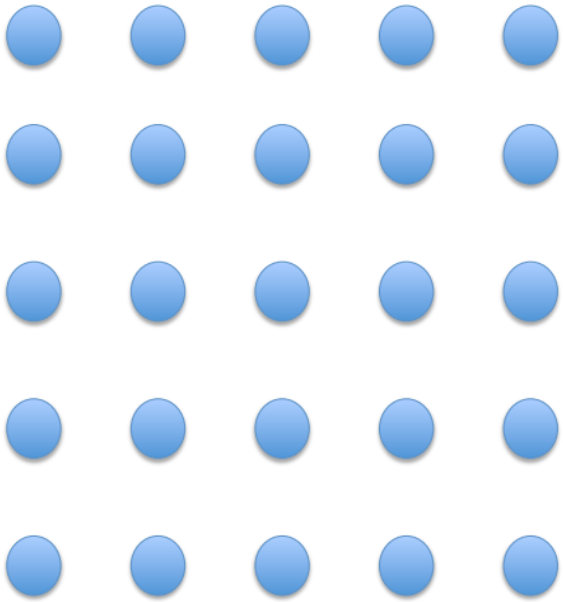
Image: Marjoram & Tavaré (2006)

The Wright-Fisher model

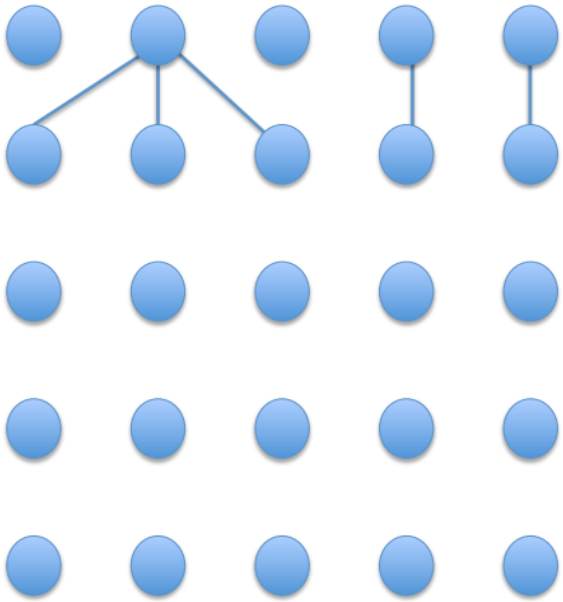
Assumptions

- Finite population size N
- At each generation, each individual chooses its parent uniformly at random from the previous generation, independently of all other individuals (no selection, pure genetic drift)
- Every individual inherits the same allelic type as the parent's (no mutation)

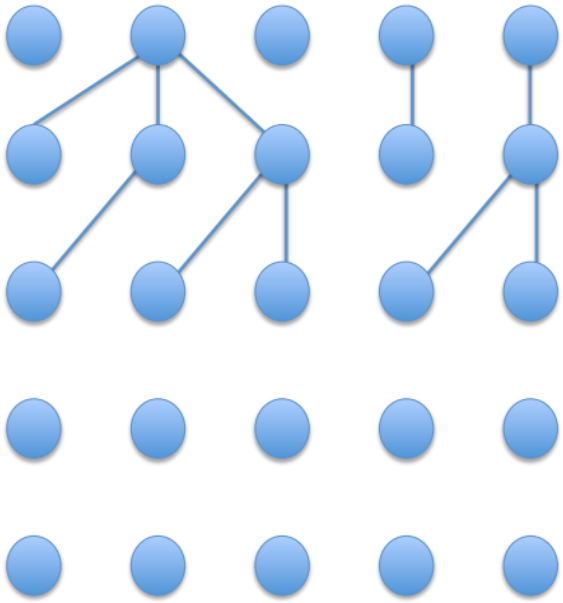
Wright-Fisher graph



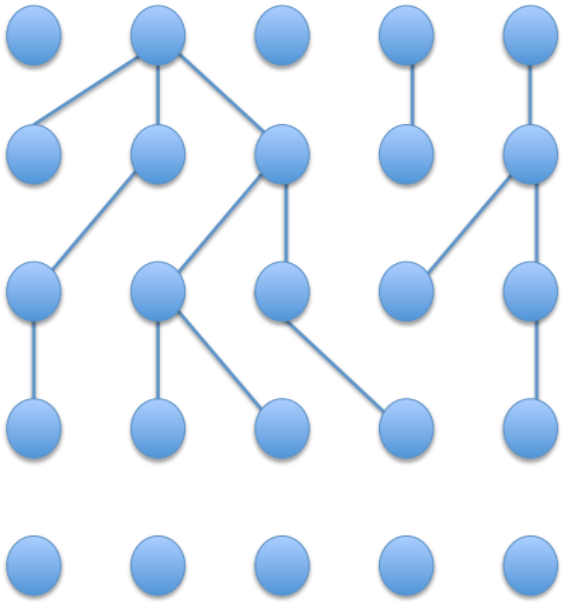
Wright-Fisher graph



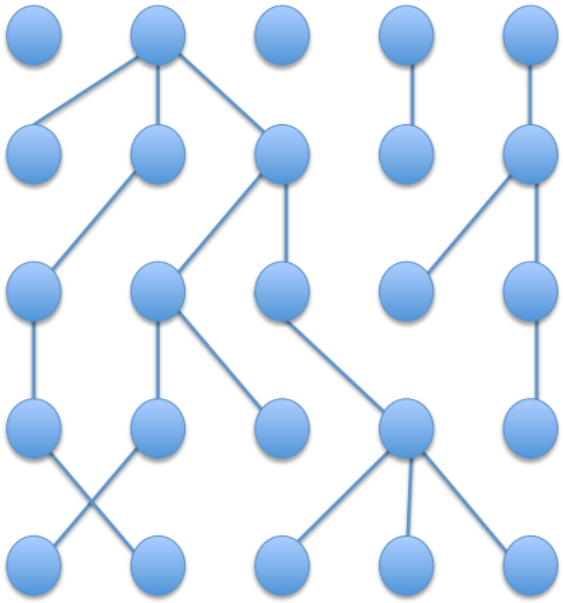
Wright-Fisher graph



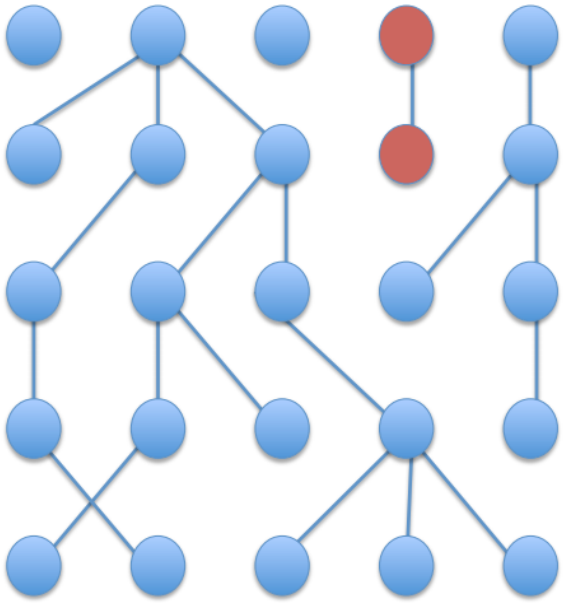
Wright-Fisher graph



Wright-Fisher graph



Wright-Fisher graph: allele frequencies



Y(0)=1/5

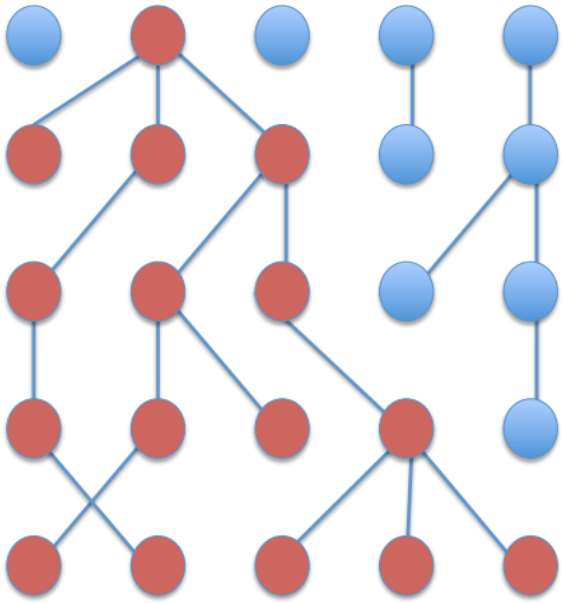
Y(1)=1/5

Y(2)=0

Y(3)=0

Y(4)=0

Wright-Fisher graph: allele frequencies



Y(0)=1/5

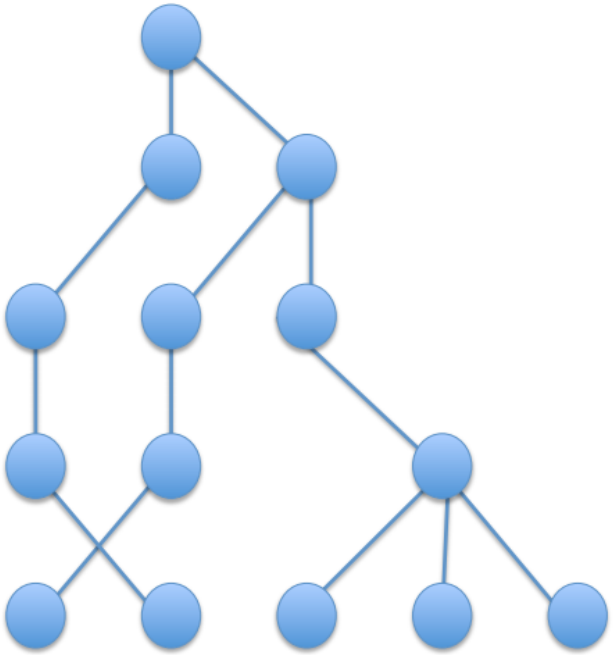
Y(1)=3/5

Y(2)=3/5

Y(3)=4/5

Y(4)=1

Wright-Fisher graph: genealogies



{1,2,3,4,5}

{1,3,4,5} , {2}

{1} , {2} , {3,4,5}

{1} , {2} , {3,4,5}

{1} , {2} , {3} , {4} , {5}

Questions (exercises):

Assume at generation 0 only one individual out of N=6 is of the “red” type

1. What is the probability that all the 6 individuals in generation 1 are of red type?

2. What is the probability that all in generation k are all of the red type?

3. Suppose we have observed that the history of red allele counts up to generation 5 is {1,3,2,5,2}. Given this information, what is the conditional probability that at generation 6 the number of red individuals is j, for any j=0,…,6 ?

- 4. Suppose that, out of N individuals, a fraction (frequency) x is of red type at generation 0. What is the expected value of the frequency of reds at time k (k=1,2,…)?