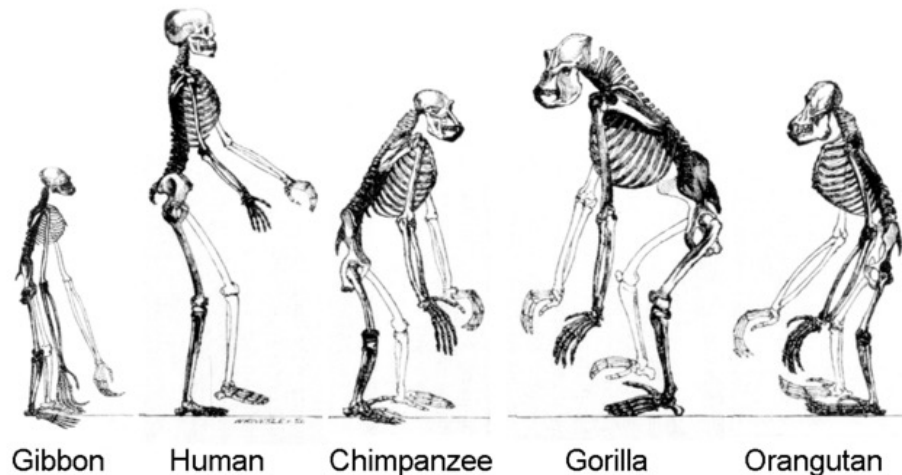


# What makes us Human?

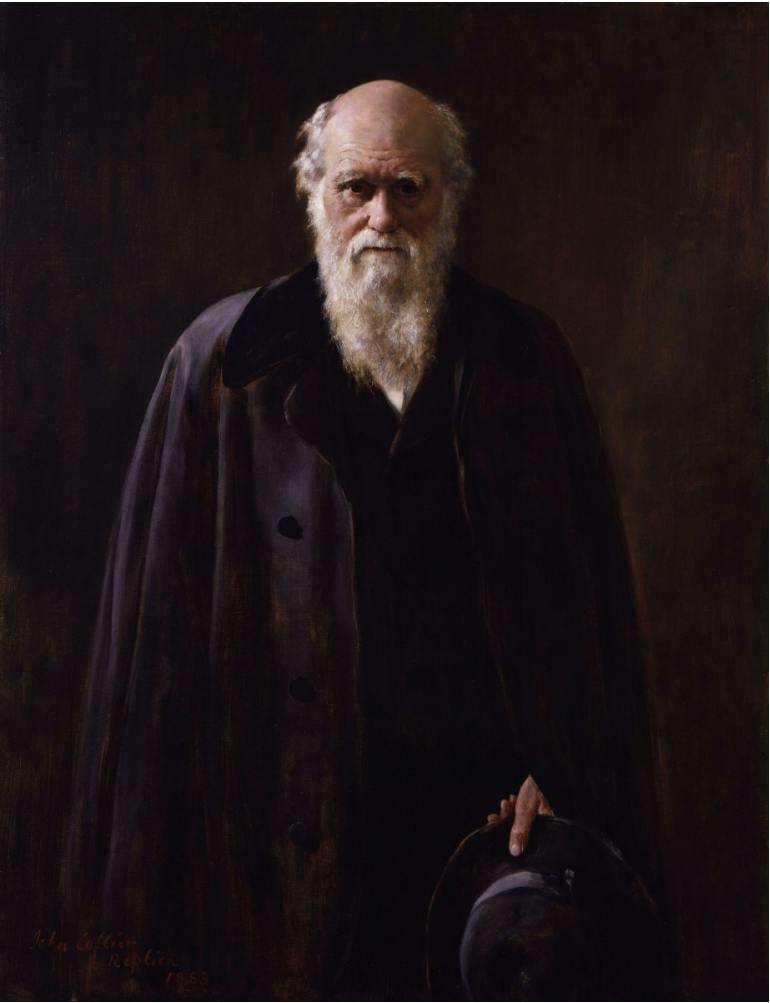
## Data Science and Genomics

---

Special Topics in DS2000: Introduction to Programming with Data

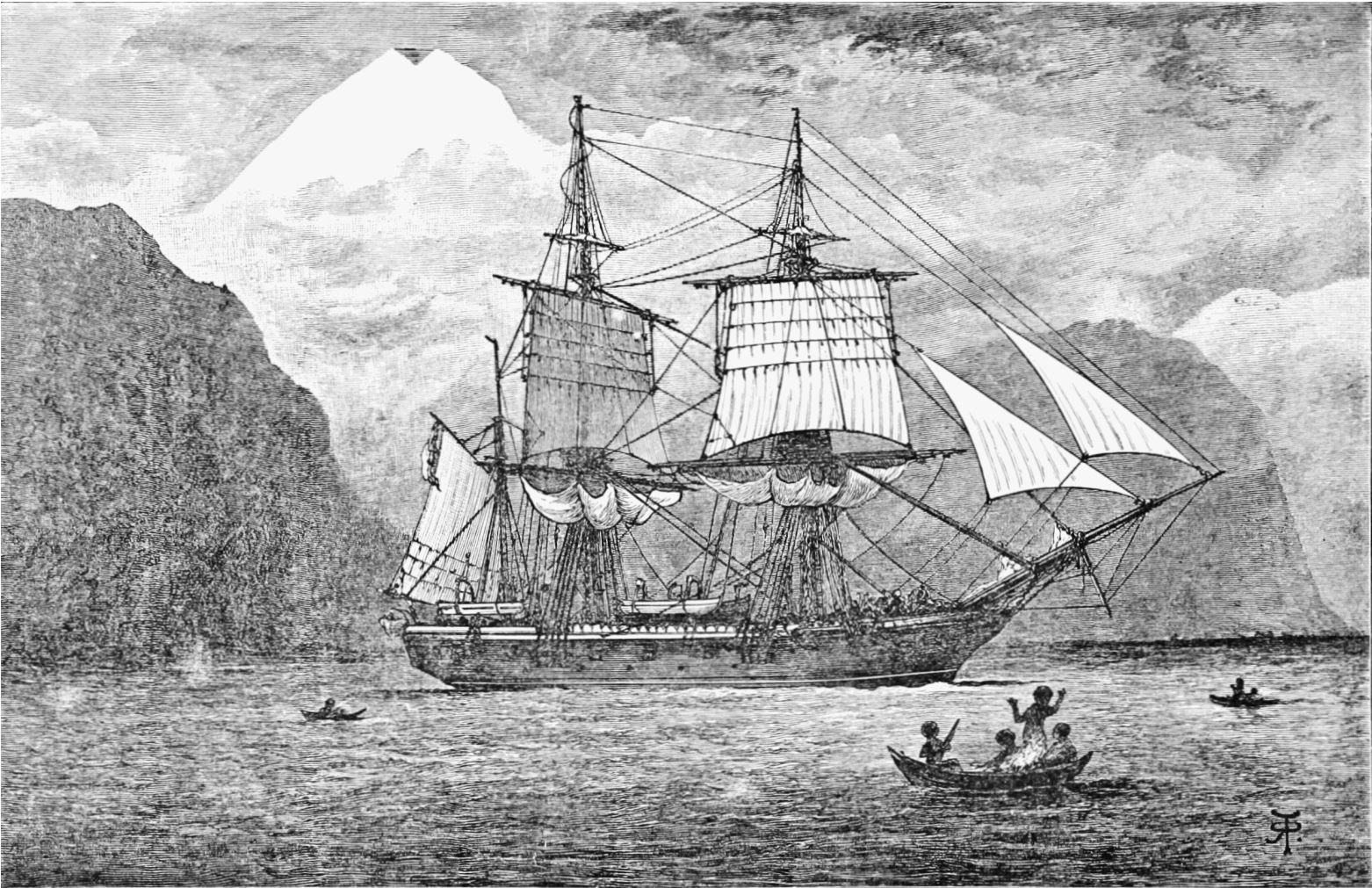


# Charles Darwin



1809 - 1882

On the Origin of Species (1859)



HMS Beagle

Voyage of the Beagle: 1831 – 1836



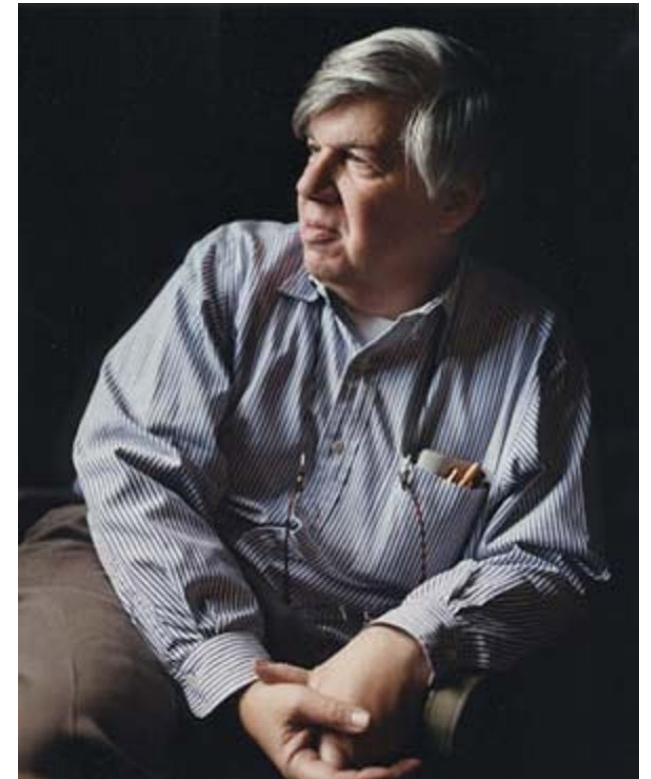
# Natural Selection

---

*[The] basis of natural selection is simplicity itself – two undeniable facts and an inescapable conclusion.*

*-- Ever Since Darwin (1977)*

Stephen Jay Gould  
Harvard Paleontologist  
1941 - 2002



# Variation and Inheritance

***1. Organisms vary, and these variations are inherited (at least in part) by their offspring.***





# Population explosion

***2. Organisms produce more offspring than can possibly survive.***



# Survival of the Fittest

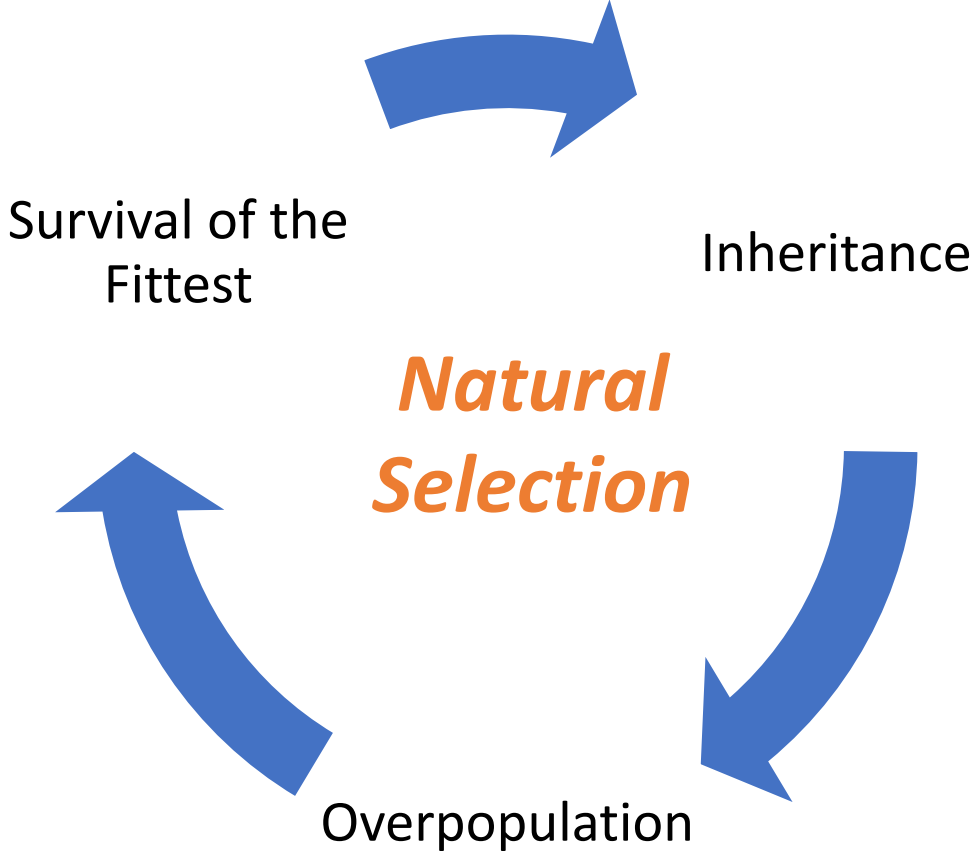
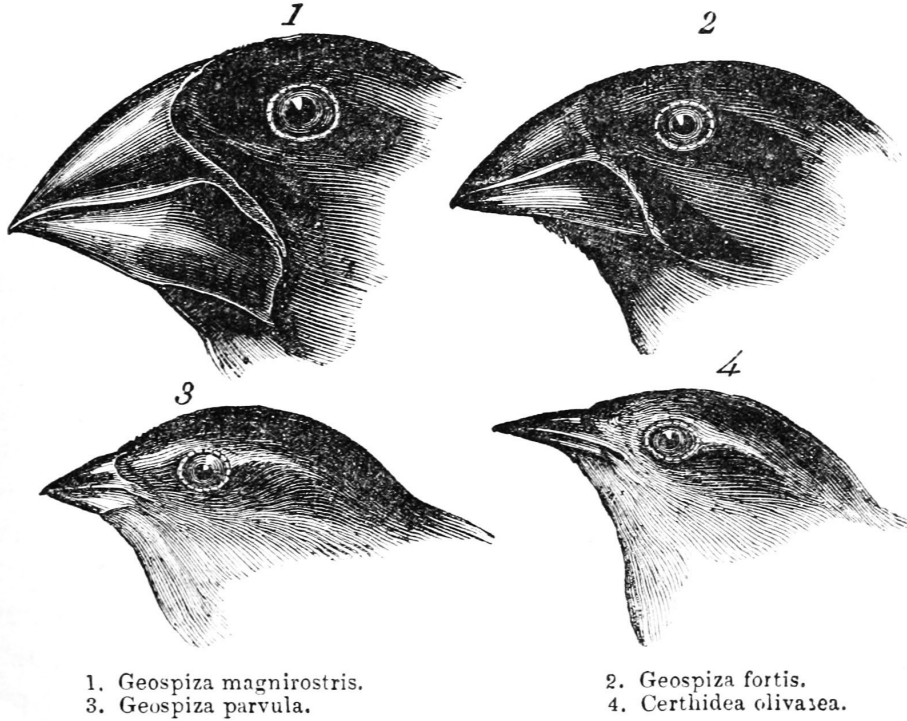
---

***3. On average, offspring that vary most strongly in directions favored by the environment will survive and propagate.***





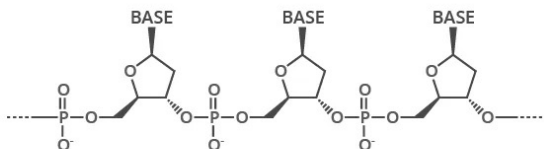
# Evolution by Natural Selection



# DNA Structure

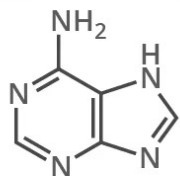
## THE CHEMICAL STRUCTURE OF DNA

### THE SUGAR PHOSPHATE 'BACKBONE'

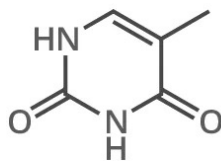


DNA is a polymer made up of units called nucleotides. The nucleotides are made of three different components: a sugar group, a phosphate group, and a base. There are four different bases: adenine, thymine, guanine and cytosine.

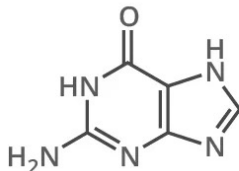
### A ADENINE



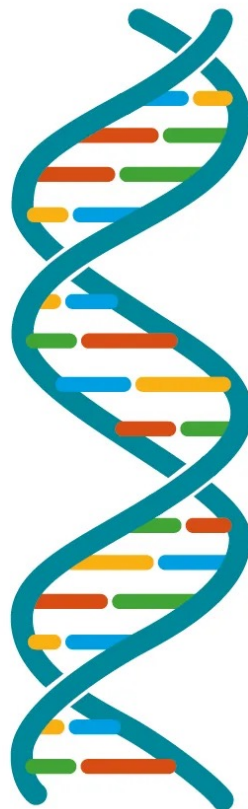
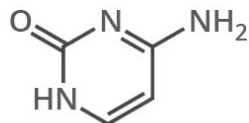
### T THYMINE



### G GUANINE

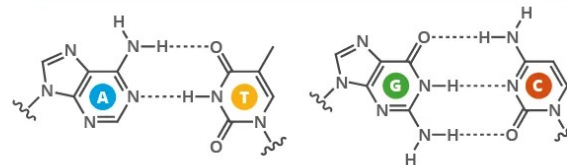


### C CYTOSINE



### WHAT HOLDS DNA STRANDS TOGETHER?

DNA strands are held together by hydrogen bonds between bases on adjacent strands. Adenine (A) always pairs with thymine (T), while guanine (G) always pairs with cytosine (C). Adenine pairs with uracil (U) in RNA.

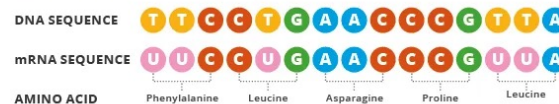


### FROM DNA TO PROTEINS

The bases on a single strand of DNA act as a code. The letters form three letter codons, which code for amino acids - the building blocks of proteins.



An enzyme, RNA polymerase, transcribes DNA into mRNA (messenger ribonucleic acid). It splits apart the two strands that form the double helix, then reads a strand and copies the sequence of nucleotides. The only difference between the RNA and the original DNA is that in the place of thymine (T), another base with a similar structure is used: uracil (U).



In multicellular organisms, the mRNA carries genetic code out of the cell nucleus, to the cytoplasm. Here, protein synthesis takes place. 'Translation' is the process of turning the mRNA's 'code' into proteins. Molecules called ribosomes carry out this process, building up proteins from the amino acids coded for.



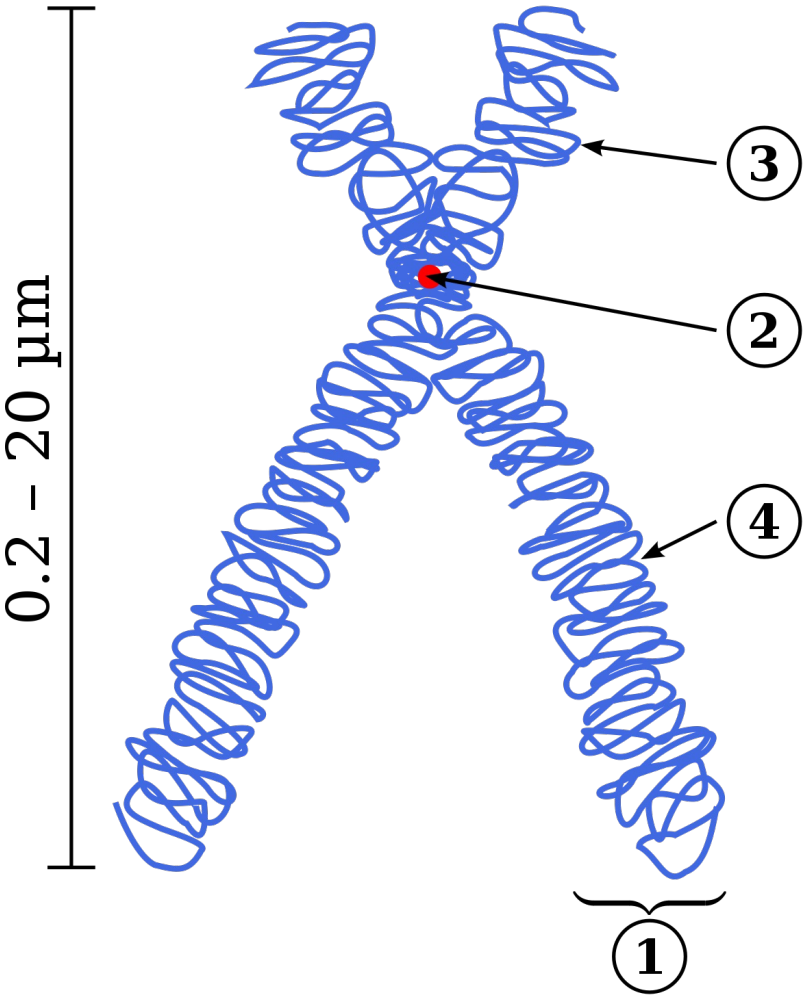
© Andy Brunning/Compound Interest 2018 - [www.compoundchem.com](http://www.compoundchem.com) | Twitter: @compoundchem | FB: [www.facebook.com/compoundchem](https://www.facebook.com/compoundchem)  
This graphic is shared under a Creative Commons Attribution-NonCommercial-NoDerivatives licence.





# 24 Chromosomes, 3 Billion Base Pairs

Chromosome	Length (bp)	Length (cm)	Weight (pg)	Weight (fg)	GC%
1	248,956,422	8.14 ± 0.08	0.25	254.57	41.72
2	242,193,529	7.92 ± 0.08	0.25	247.65	40.23
3	198,295,559	6.48 ± 0.06	0.20	202.76	39.67
4	190,214,555	6.22 ± 0.06	0.19	194.49	38.24
5	181,538,259	5.93 ± 0.06	0.19	185.63	39.51
6	170,805,979	5.58 ± 0.05	0.17	174.65	39.61
7	159,345,973	5.21 ± 0.05	0.16	162.94	40.70
8	145,138,636	4.74 ± 0.05	0.15	148.41	40.16
9	138,394,717	4.52 ± 0.04	0.14	141.51	41.28
10	133,797,422	4.37 ± 0.04	0.14	136.81	41.54
11	135,086,622	4.42 ± 0.04	0.14	138.13	41.54
12	133,275,309	4.36 ± 0.04	0.14	136.28	40.77
13	114,364,328	3.74 ± 0.04	0.12	116.94	38.55
14	107,043,718	3.50 ± 0.03	0.11	109.46	40.83
15	101,991,189	3.33 ± 0.03	0.10	104.29	42.03
16	90,338,345	2.95 ± 0.03	0.09	92.38	44.58
17	83,257,441	2.72 ± 0.03	0.09	85.14	45.32
18	80,373,285	2.63 ± 0.03	0.08	82.18	39.78
19	58,617,616	1.92 ± 0.02	0.06	59.95	47.94
20	64,444,167	2.11 ± 0.02	0.07	65.90	43.80
21	46,709,983	1.53 ± 0.01	0.05	47.76	40.94
22	50,818,468	1.66 ± 0.02	0.05	51.97	47.00
X	156,040,895	5.10 ± 0.05	0.16	159.55	39.53
Y	57,227,415	1.87 ± 0.02	0.06	58.52	40.03
Total (1–22, X, Y) <sup>a</sup>	3,088,269,832	100.96 ± 0.97	3.16	3157.87	40.87

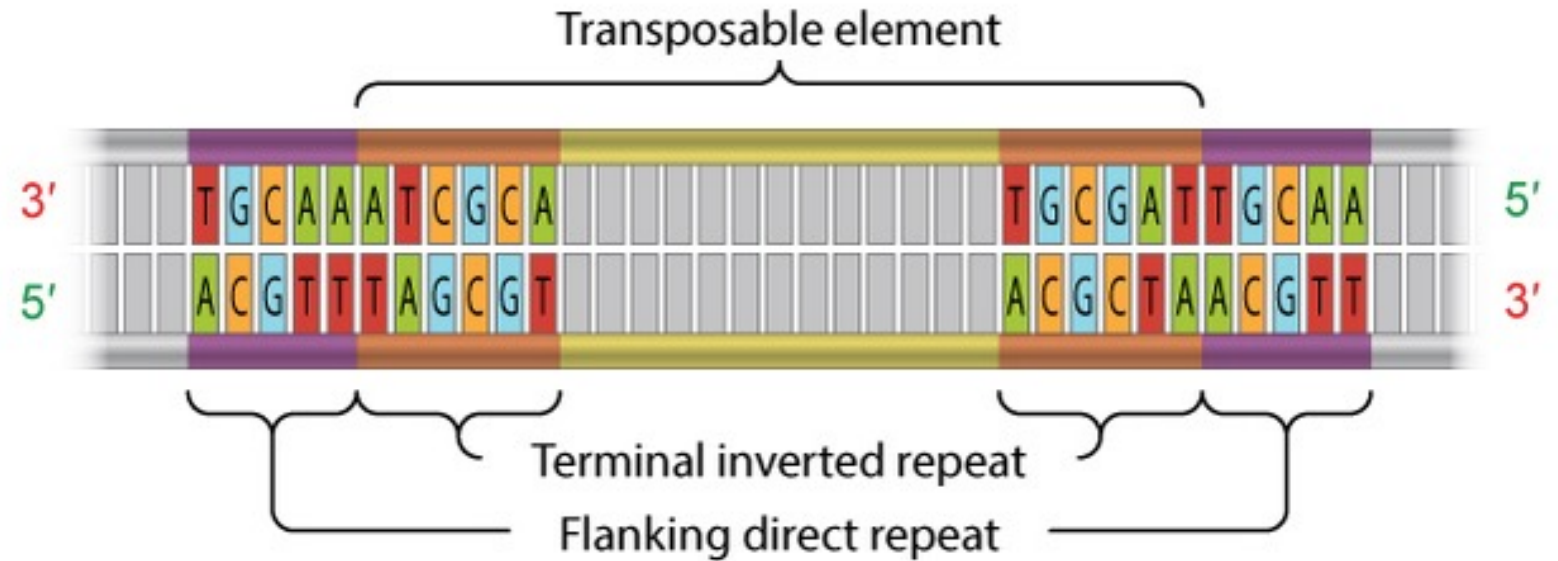


- 1. Chromatid
- 2. Centromere
- 3. Short arm
- 4. Long arm





# Barbara McClintock and Jumping Genes



1902 – 1992

Winner of the 1983 Nobel Prize in Physiology or Medicine

“for her discovery of mobile genetic elements”



# Genomics Resources (and many more)

<https://www.ncbi.nlm.nih.gov>

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

### Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

#### Submit

Deposit data or manuscripts into NCBI databases



#### Download

Transfer NCBI data to your computer



#### Learn

Find help documents, attend a class or watch a tutorial



#### Develop

Use NCBI APIs and code libraries to build applications



#### Analyze

Identify an NCBI tool for your data analysis task



#### Research

Explore NCBI research and collaborative projects



[https://uswest.ensembl.org/Homo\\_sapiens/Info/Index](https://uswest.ensembl.org/Homo_sapiens/Info/Index)

The screenshot shows the Ensembl genome browser interface for Homo sapiens. At the top, there is a navigation bar with the Ensembl logo, a search bar containing "Search Human...", and links for "Login/Register", "BLAST/BLAT", "VEP", "Tools", "BioMart", "Downloads", "Help & Docs", and "Blog". Below the navigation bar, the current species is set to "Human (GRCh38.p13)". A search box is titled "Search Human (Homo sapiens)" and contains the text "Search all categories Search... Go". Below the search box, there are example search terms: "e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis". The main content area is divided into several sections: "Genome assembly: GRCh38.p13 (GCA\_000001405.28)" with links for "More information and statistics", "Download DNA sequence (FASTA)", "Convert your data to GRCh38 coordinates", and "Display your data in Ensembl"; "Gene annotation" with a "What can I find?" section listing protein-coding and non-coding genes, splice variants, cDNA, and protein sequences, and links for "More about this genebuild", "Download FASTA files for genes, cDNAs, ncRNA, proteins", "Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins", and "Update your old Ensembl IDs"; and "Other assemblies" with a dropdown menu set to "GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart" and a "Go" button. On the right side, there is a "View karyotype" link and an "Example region" section showing a genomic track. At the bottom right, there is an "Example gene" section showing a gene model for Pax6, with subunits FOXP2, BRCA2, and DMU, and an "Example transcript" section showing a transcript model.