

# Genome Informatics 2022

Lesson 1 - Introduction

# Communication

[github.com/vladimirkovacevic/gi-2022-etf](https://github.com/vladimirkovacevic/gi-2022-etf)

(All info about the course. Create an issue!)

[vladimir.kovacevic-ctr@sbgenomics.com](mailto:vladimir.kovacevic-ctr@sbgenomics.com)

General info (not relevant to all)

Questions about lessons 1-6

[vojislav.varjacic@sbgenomics.com](mailto:vojislav.varjacic@sbgenomics.com)

Questions about lessons 7-10

[marko.zecevic@sbgenomics.com](mailto:marko.zecevic@sbgenomics.com)

Questions about lessons 11-12

[marko.misic@etf.bg.ac.rs](mailto:marko.misic@etf.bg.ac.rs)

Course professor

[pediao@etf.bg.ac.rs](mailto:pediao@etf.bg.ac.rs)

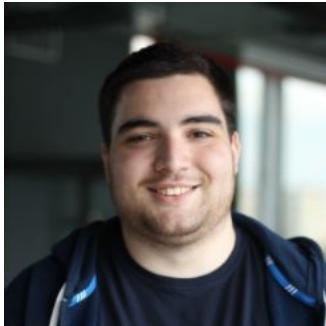
Teaching assistant

Questions about technical details and exercises

# Teachers



Vladimir Kovacevic finished computer science in 2007 at the School of Electrical Engineering where he also obtained his PhD in 2016. He worked for 5 years for Intel as algorithm developer and 5 years he's been with Seven Bridges bioinformatics company and teaching Genome informatics course at School of Electrical Engineering in Belgrade. His current interest includes machine learning applied in genomics, precision medicine, neoantigen discovery and bioinformatics in general.



Vojislav Varjacic has a masters degree from School of Electrical Engineering, department of Physical Electronics. Has been working at Seven Bridges for the last 5 years as a bioinformatics analyst; specializing in tumor heterogeneity and copy number variation analysis. His interests include automating large scale analyses and machine learning medical imaging projects.



Marko Zecevic has obtained his master's degree from School of Electrical Engineering, department of Signals and Systems. Spent several years designing relay protection and control systems for high voltage substations before transitioning to bioinformatics in 2015. His interests include transcriptomics, populational genomics and R programming. PhD student of genomics at Faculty of Biology.

# Teachers - course support



Marko Mišić is an assistant professor at the University of Belgrade, School of Electrical Engineering, Republic of Serbia. He completed B.Sc. M.Sc. and Ph.D. in the field of Computer Engineering and Informatics, at the same university. His main areas of interest are parallel programming with an emphasis on GPU programming, algorithms, data structures, and complex network analysis.



Predrag Obradović is a senior teaching assistant at the University of Belgrade, School of Electrical Engineering, Republic of Serbia. He completed B.Sc. M.Sc. in the field of Software Engineering, at the same university, where he pursues Ph.D. in Computer Engineering and Informatics. His research interests include social and complex network analysis, bioinformatics, and machine learning.

# Course info

- 13 classes (lecture + exercise) - ~2.5h
- Exam will have both theoretical and practical part
  - 40% on the exam
  - 60% during the semester
    - 40% project assignment (with presentation)
    - 20% - exercises (homework)
- Last class - presentation of student assignments
- Exercise will follow lectures - examples in python [Jupyter notebook](#)

# Course info - syllabus

1	Course info. Bioinformatics and genomics definitions. Molecular biology basics. Genome sequencing technologies. Exercise: Introduction to python and Jupiter environment.
2	Portable and reproducible bioinformatic analysis. Describing bioinformatic tools in Common Workflow Language. Exercise: Python structures. Pandas library. Writing tests.
3	Executing bioinformatic analysis locally and on the cloud. Variant calling. Cancer analysis.
4	Prof. Aleksandar Milosavljevic, Baylor College of Medicine, Houston: Introduction to metabolomics. Pattern discovery and data compression.
5	Exact string matching: Boyer-Moore, Indexing structures, Hash tables, Suffix trie/tree
6	Burrows-Wheeler Transform and FM Index
7	Approximate string matching, Edit distance, Dynamic programming, Global alignment
8	Variation on global alignment (end-space-free variant, longest common substring) , local alignment, gaps. Practical: BLAST, Bowtie
9	Shortest common superstring, Overlap graph
10	De-Bruijn graph, scaffolding, error correction
11	The central dogma of molecular biology, RNA-Seq motivation and technologies for gene expression measurement; RNA-Seq alignment.
12	Normalization procedures; Differential expression: statistical inference, multiple testing corrections; Biological pathways.
13	Presentation of the student projects.

# Literature

- Vince Buffalo: **Bioinformatics Data Skills**
- Dan Gusfield: **Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology**, Cambridge University Press
- Pavel Pevzner, Neils Jones: **An Introduction to Bioinformatics Algorithms (Computational Molecular Biology)**, MIT Press
- R. Durbin, S. Eddy, A. Krogh, G. Mitchinson: **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids** , Cambridge University Press
- Veli Mäkinen, Djamal Belazzougui, Fabio Cunial, Alexandru I. Tomescu: **Genome-Scale Algorithm Design: Biological Sequence Analysis in the Era of High-Throughput Sequencing**, Cambridge University press

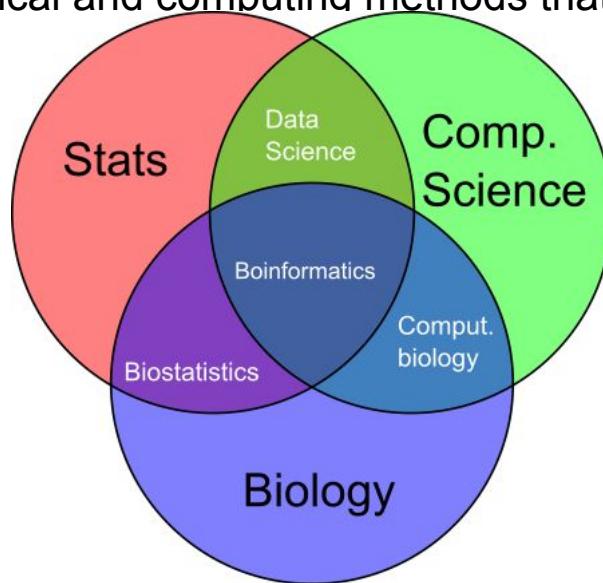
# Lesson overview

- Bioinformatics and genomics definitions
- Molecular biology basics
- Genome sequencing technologies
- We use git
- Introduction to Python

# What is bioinformatics

**Bioinformatics**, n. The science of information and information flow in biological systems, esp. of the use of computational methods in genetics and genomics. (*Oxford English Dictionary*)

**Bioinformatics** - using statistical and computing methods that aim to solve biological problems.



# What is bioinformatics

"I do not think all biological computing is bioinformatics, e.g. mathematical modelling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular genetic information."

-- Richard Durbin

**Bioinformatics in practice:** Develops methods and software tools for storing, retrieving, organizing and analyzing biological data.

# Bioinformatics principles

- Golden rule of bioinformatics: **Never ever trust your tools (or data)**
- Adapt robust and reproducible practices
  - Document each step
  - Write down key facts
  - Conclude the work done to easy understand it again in few months (make figures, reports)
  - Automate manual tasks with script
  - Always ask yourself: How much time would it take me to do it one more time?
  - Release or publish your code and tools
- Write code for humans, Write (meta)data for computers
  - Make assertions (be loud)
  - Let the code test the code
  - Use existing libraries whenever possible
  - Frequently used scripts -> tools
  - Use code versioning (git)

# Genomics 101

**Genome:** “The complete set of genes or genetic material present in a cell or organism.” (*Oxford English Dictionary*)

- “Blueprint” or “recipe” of life
- Human genome - 6 billions of base-pairs (A, C, T, G) letters
  - Can be imagined as a string 6 billion letters long

ACTGTGTCACATCGAGAGAGATCACAAACACATAGATTACGATCGAACGTAACGTAACACACC  
AAATATACTGAGTGAGGGGTGGGGACCCCCCCCCCCCCACACACATTACAAACCTAGAT  
CACCATACTACAGATATAAGAGAGAGANACGTACGTACACAATTACAAATTACAACACAAAGTA  
CTTATACATACACATGGGACCCATAGCACACACAGATATTATAATATAGAGAGACAATGT  
CGTGCAGTAA...

# Genomics: contrast with biology & genetics\*

\* Everything on this slide is  
a gross generalization

## Biology & Genetics

Targeted studies of one or a few genes

Targeted, low-throughput experiments

Clever experimental design, painstaking experimentation

## Genomics

Studies considering all genes in a genome

Global, high-throughput experiments

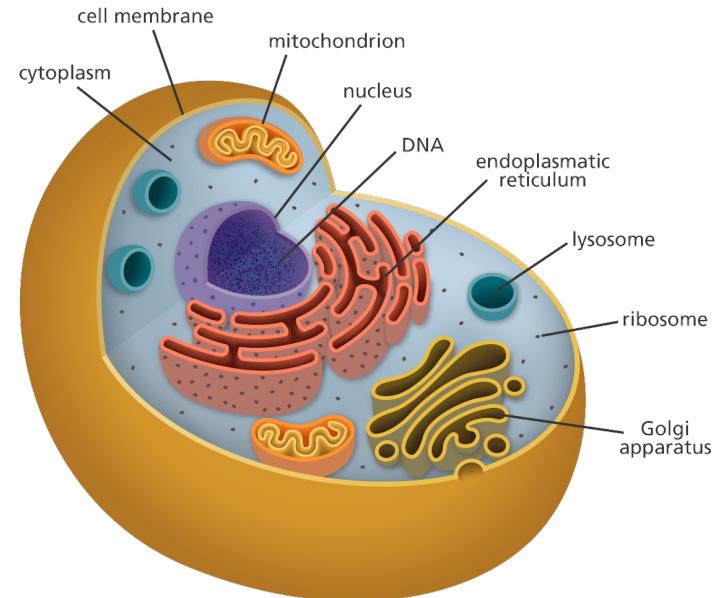
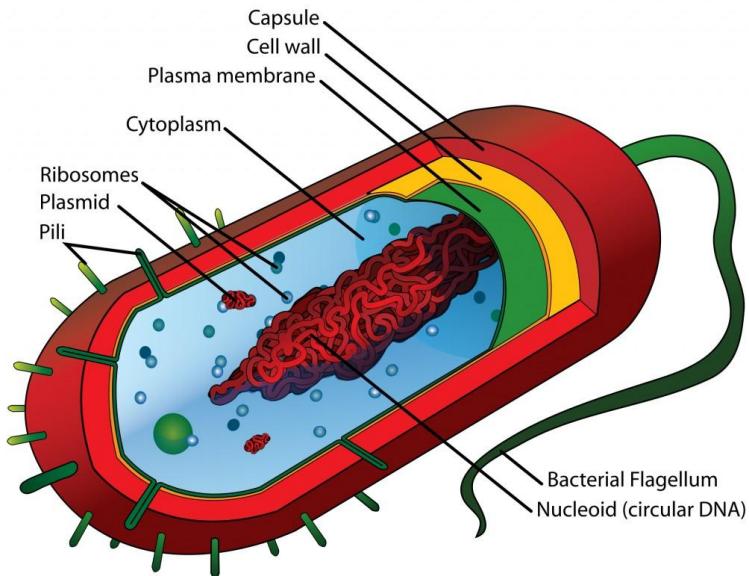
Tons of data, uncertainty, computation



# The cell

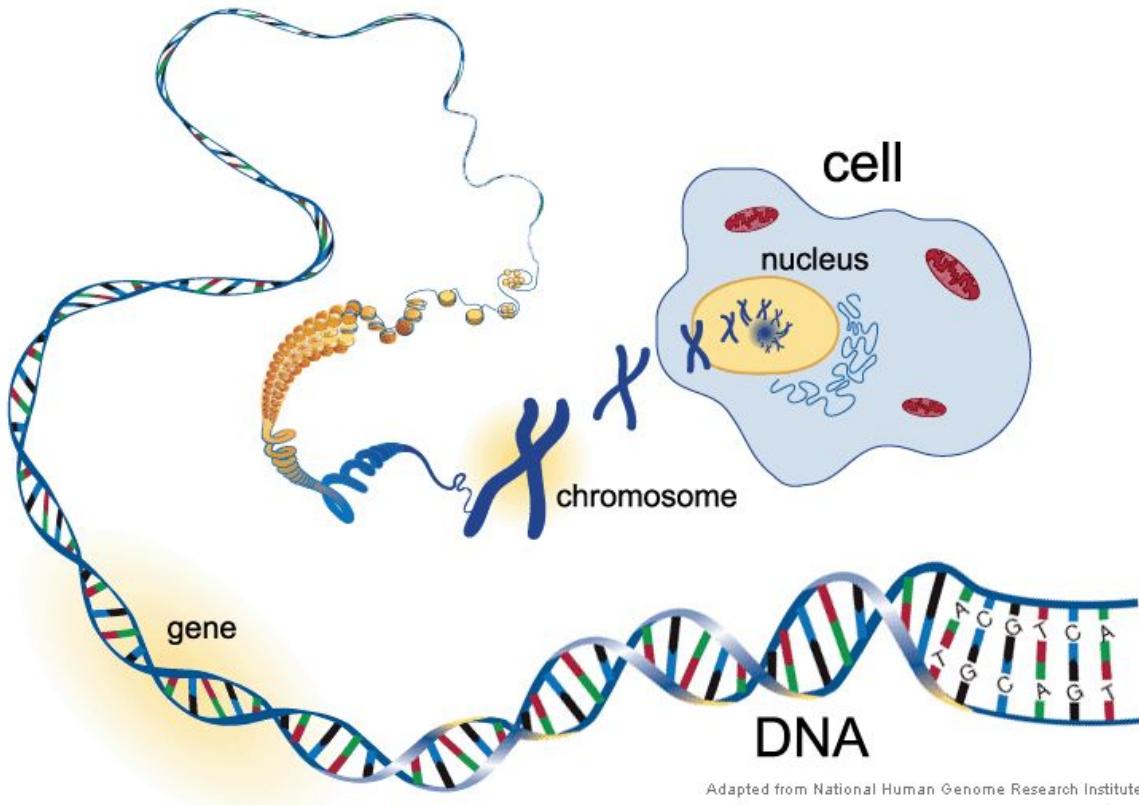
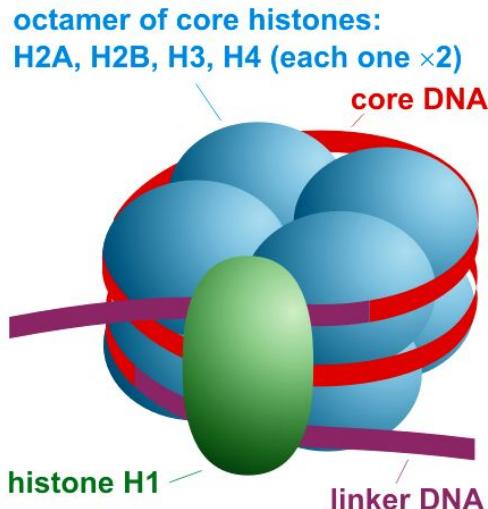
Fundamental working units of every living system.

- Prokaryotic (bacteria)
- Eukaryotic (higher organisms - animals, plants)



# “And inside the nucleus thou lays the mighty DNA”

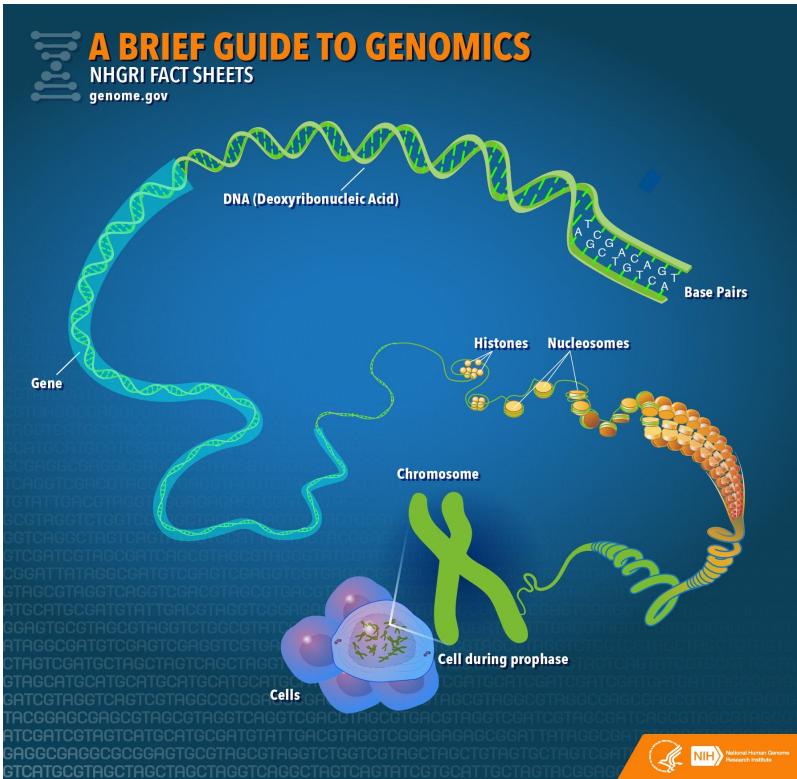
- Chromatin - tightly packed DNA
- A **nucleosome** is a basic unit of DNA packaging in eukaryotes, consisting of a segment of DNA wound in sequence around eight histone protein cores.
- Current model



Adapted from National Human Genome Research Institute

# DNA - the code of life

- DNA (deoxyribonucleic acid) - double stranded molecule
- Same in every cell - DNA replication during cell division
- More stable, redundant information - complementary double helix chain
- ~99.6% same between 2 individuals
- Base (nucleotide) pairs (complementary bases)
  - A - T (adenine and thymine)
  - C - G (cytosine and guanine)

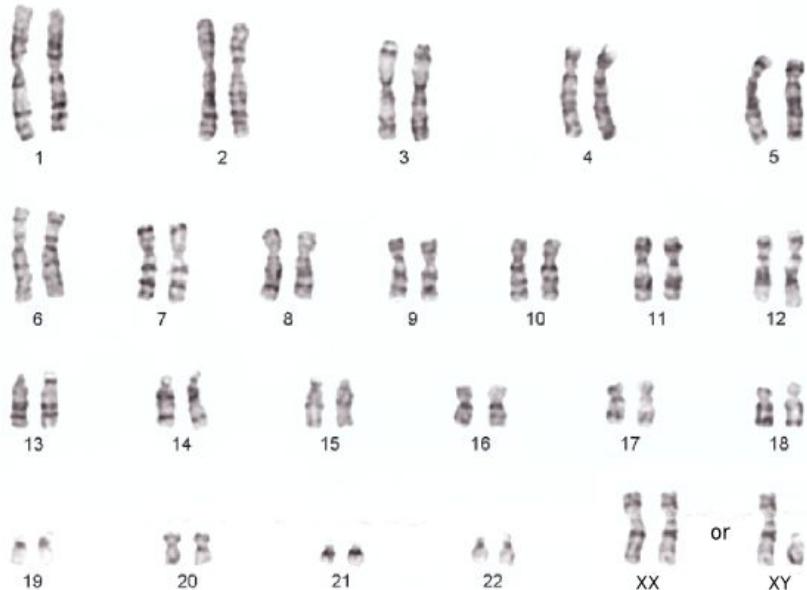


# Genome

- Set of all pairs of chromosomes

- Human genome:

- 23 pair of chromosomes (diploid)
- 22 autosomes
- 1 sex chromosome (X and/or Y)
- 3 billion base-pairs x 2
- CTGGATTATATCGAAGGGACTAT... etc
- Intron and exon (2%)

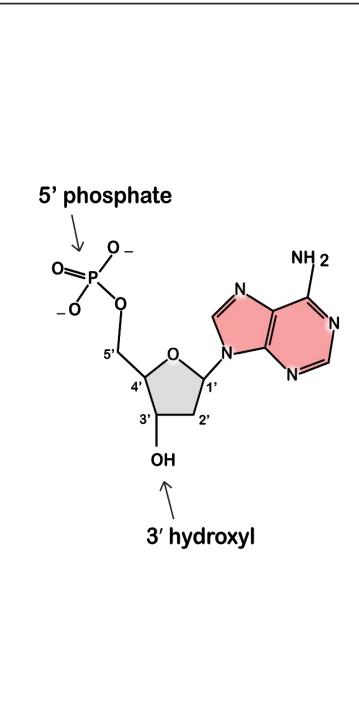
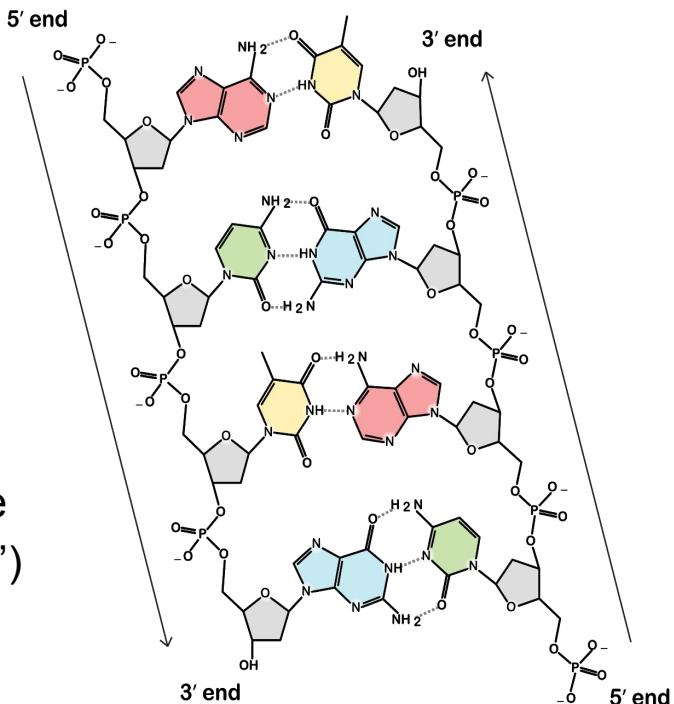


Karyogram

# DNA - structure

- Consists of:
  - Phosphate group
  - Sugar (deoxyribose)
  - Nitrogen base
- Hydrogen bonds
- Forward and reverse strand
- DNA direction:
  - 5' head and 3' tail
  - Transcribed from 5' to 3' end
- In bioinformatics we write just one strand (by convention from 5' to 3')

5' ACTG 3'  
↓  
3' TGAC 5'  
(reverse complement)

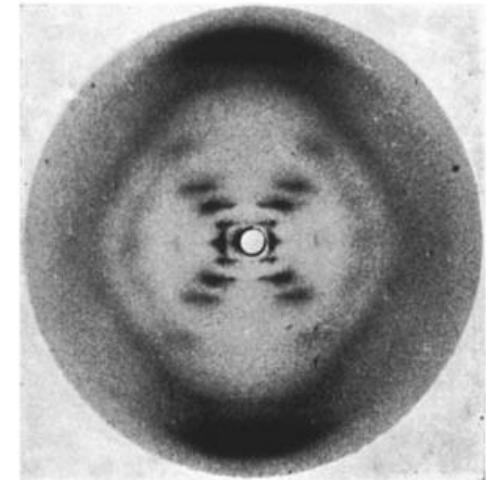


# DNA - discovery

**1952-1953** James D. Watson and Francis H. C. Crick deduced the double helical structure of DNA from X-ray diffraction images by Rosalind Franklin (provided by M. Wilkins) and data on amounts of nucleotides in DNA.

"Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid"

scientist: "does everyone here know what Watson and Crick discovered?"  
me from back of room: "Rosalind Franklin's notes"



*Photo 51*

# Central dogma of molecular biology



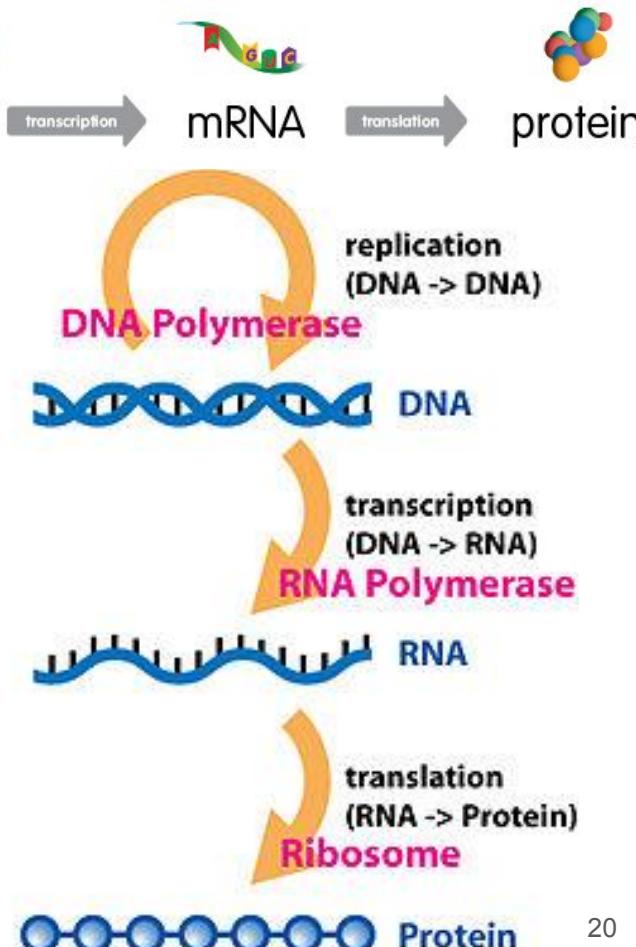
DNA ----> RNA -----> Protein

Transcription: DNA ->RNA

- particular segment of DNA is copied into RNA (especially mRNA) by the enzyme RNA polymerase.

Translation: RNA -> Protein

- process in which ribosomes synthesize proteins after the process transcription of DNA to RNA in the cell's nucleus.



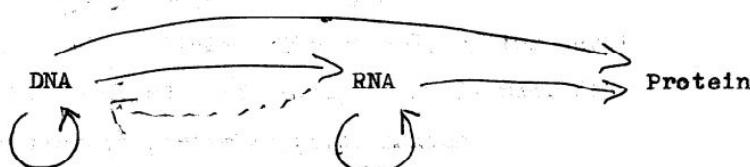
# Central dogma of molecular biology

## Ideas on Protein Synthesis (Oct. 1956)

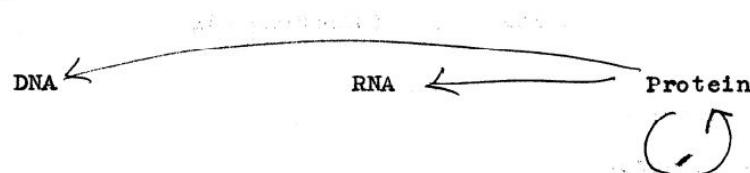
### The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have



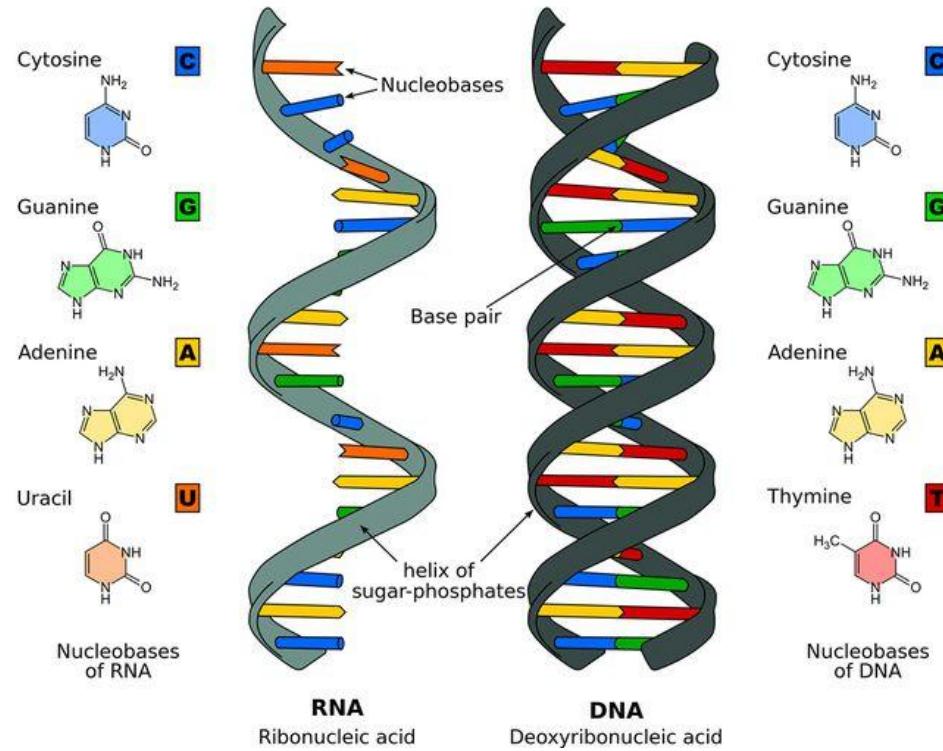
but never



where the arrows show the transfer of information.

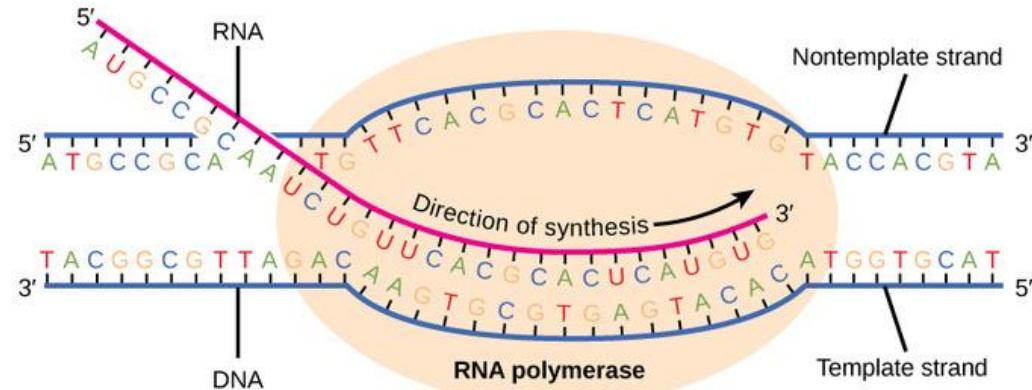
# RNA

- Single stranded
- Sugar:
  - ribose (instead of deoxyribose)
- Uracil instead of Thymine

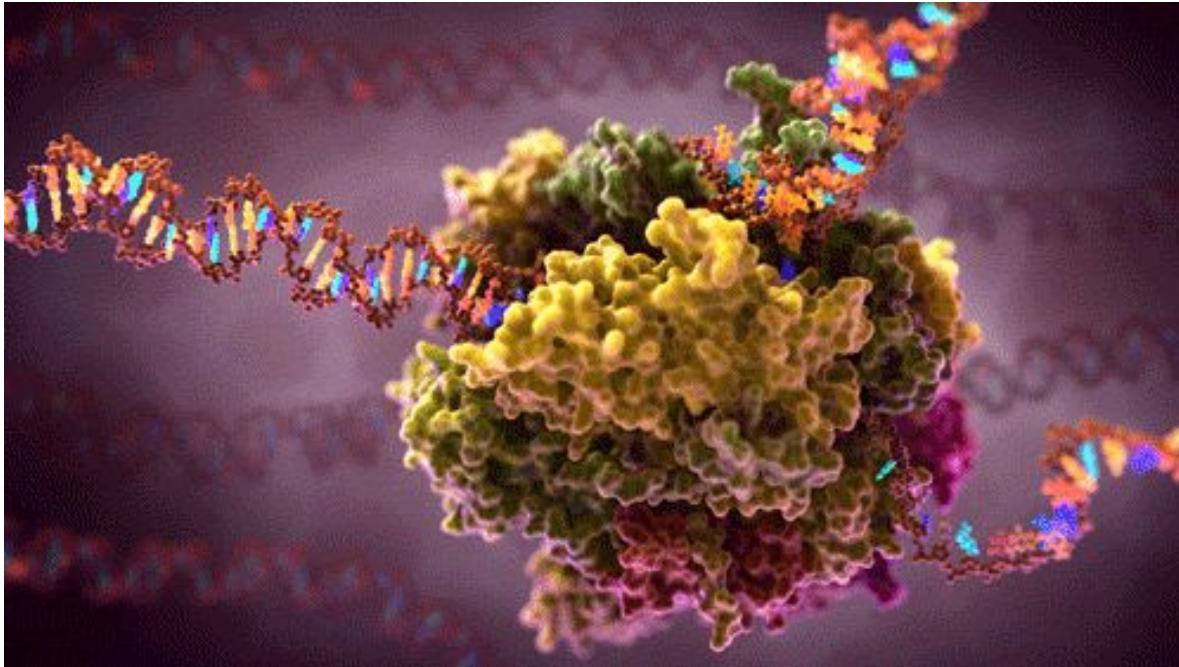


# Transcription

- Template (noncoding) strand
  - One which is transcribed by RNAP (RNA polymerase)

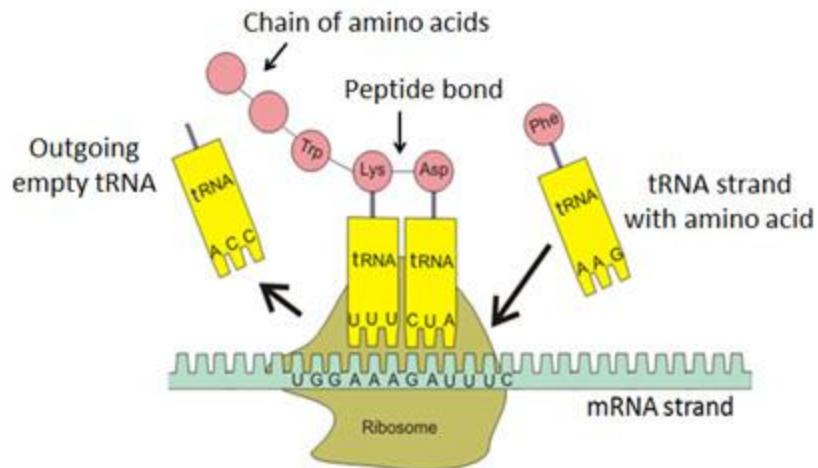


# Transcription



# Translation

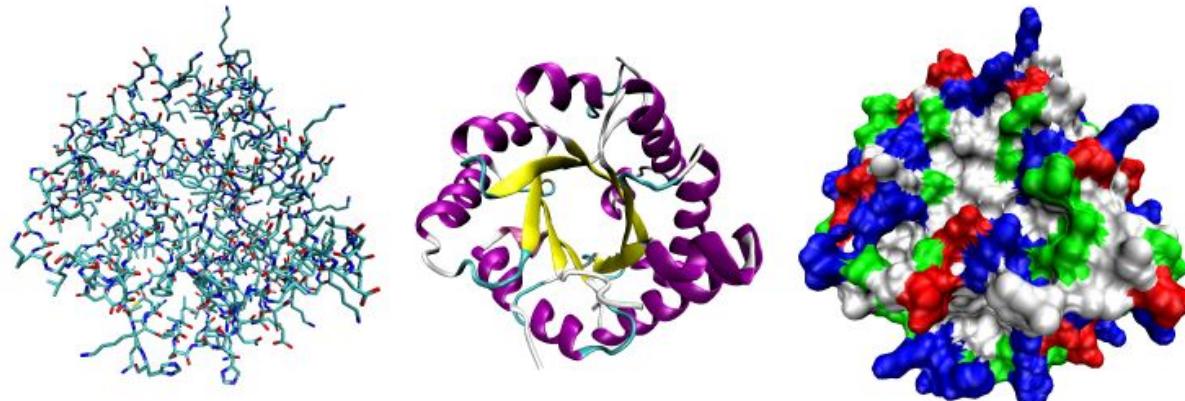
- Occurs in ribosome
- Each triplet of nucleotides (codon) codes for specific amino-acid
  - “Letters of protein code”
  - 20 amino-acid (some redundancy)



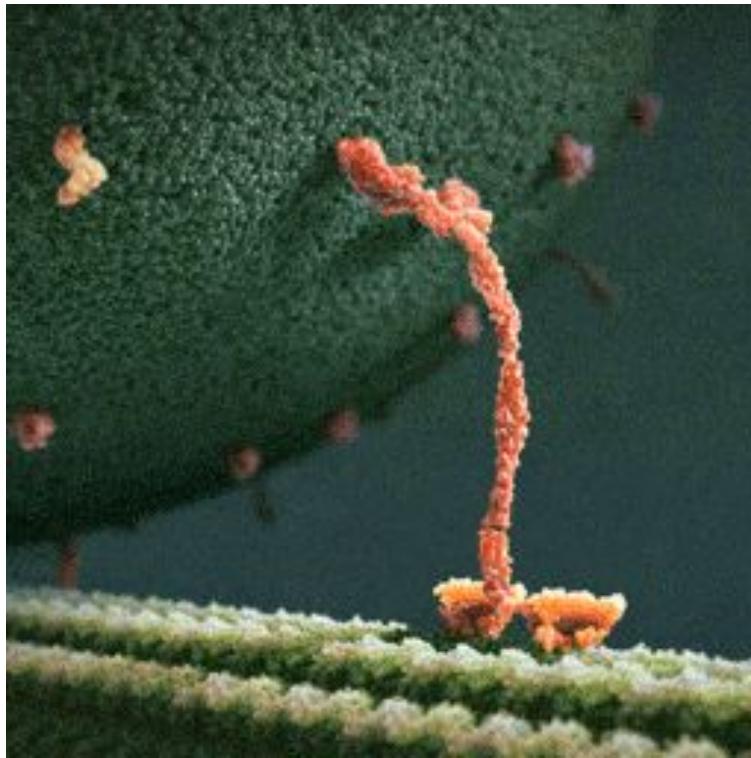
		Second letter				Third letter
		U	C	A	G	
First letter	U	UUU } Phe UUC UUA UUG	UCU } Ser UCC UCA UCG	UAU } Tyr UAC UAA Stop UAG Stop	UGU } Cys UGC UGA Stop UGG Trp	U C A G
	C	CUU } Leu CUC CUA CUG	CCU } Pro CCC CCA CCG	CAU } His CAC CAA } Gln CAG	CGU } Arg CGC CGA CGG	U C A G
	A	AUU } Ile AUC AUA AUG Met	ACU } Thr ACC ACA ACG	AAU } Asn AAC AAA } Lys AAG	AGU } Ser AGC AGA AGG	U C A G
	G	GUU } Val GUC GUA GUG	GCU } Ala GCC GCA GCG	GAU } Asp GAC GAA } Glu GAG	GGU } Gly GGC GGA GGG	U C A G

# Proteins

- Building blocks of life
  - Various functions in the organism (transportation, regulation, metabolism, DNA replication)
- Long chains of amino-acids, that also fold into complicated 3D structures
  - We often distinguish protein primary, secondary, tertiary and quaternary structure



# Proteins



# Proteins



[www.dnalc.org](http://www.dnalc.org)

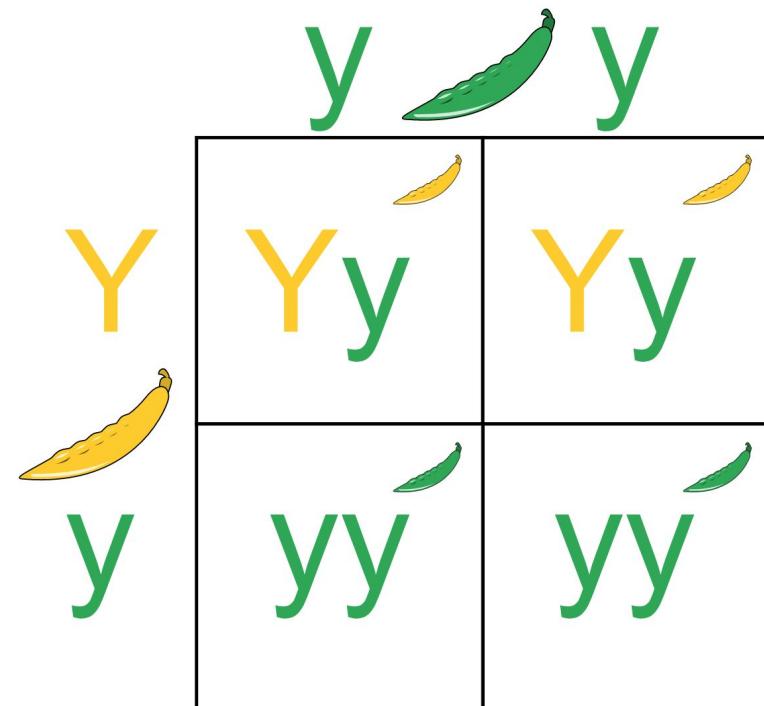
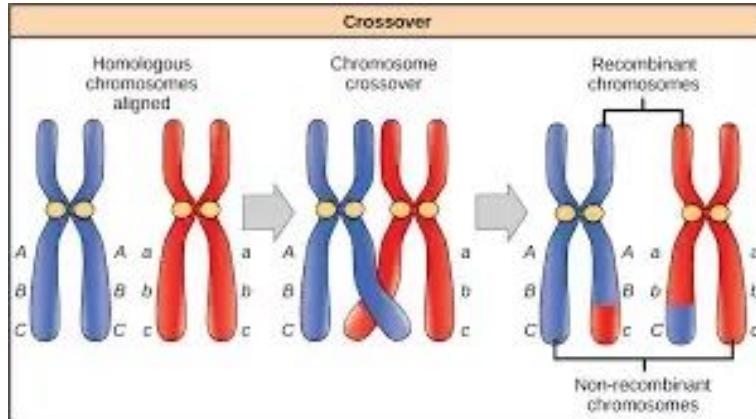
# Biology 101 - genotype vs phenotype

The **genotype** is the part of the genetic makeup of a cell, and therefore of an organism or individual, which determines one of its characteristics (phenotype).

A **phenotype** (from Greek *phainein* , meaning 'to show ', and *typos* , meaning 'type') is the composite of an organism's **observable characteristics** or traits, such as its morphology, development, biochemical or physiological properties, behavior, and products of behavior (such as a bird's nest).

# Rules of inheritance

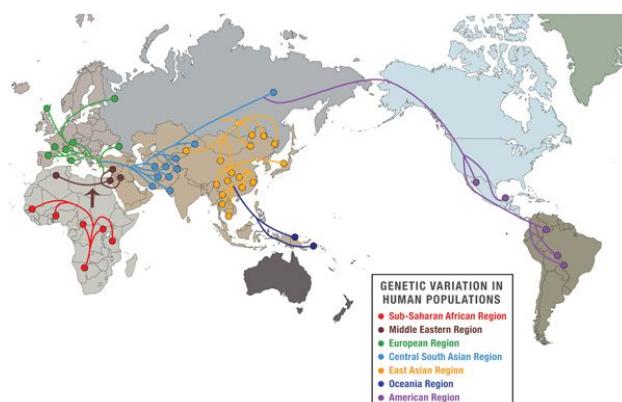
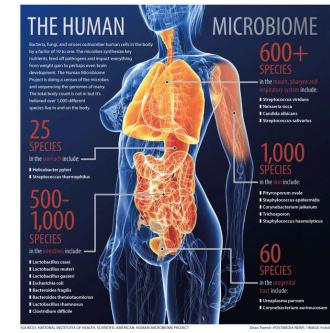
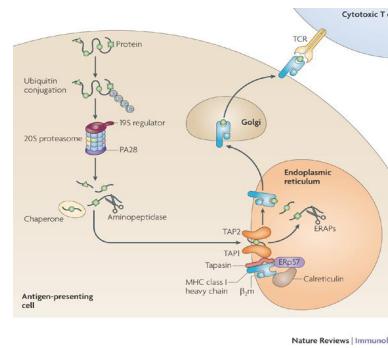
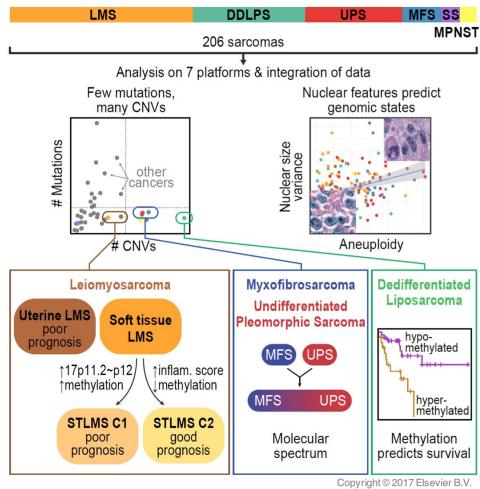
- Mendelian inheritance
- Multiple alleles of the same gene
  - One allele per chromosome
  - Dominant/recessive allele



Example of inheritance combinations

# Why perform DNA sequencing?

- Rare genetic diseases
- Origins of humans
- Precision medicine- Cancer treatment (immunotherapy)
- Microbes that live inside us (microbiome)
- Study ways that genomes work
- Gene editing
- Forensics



MEDICAL DISPATCH JULY 21, 2014 ISSUE

## ONE OF A KIND

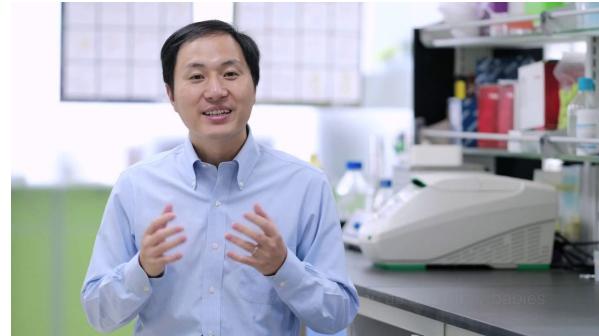
*What do you do if your child has a condition that is new to science?*

By Seth Mnookin

Stanford University

# The world's first germline genetically edited babies

- Clinical project: standard in vitro fertilization
- + CRISPR-Cas9 (technology that can modify DNA)
- Mother was HIV positive
- Modify the CCR5 gene on single egg cell before fertilization to confer genetic resistance to the HIV virus
- CCR5 codes for a protein that HIV uses to enter cells
- Clinical project was conducted secretly until November 2018
- Lulu and Nana are born healthy crying babies



He Jiankui

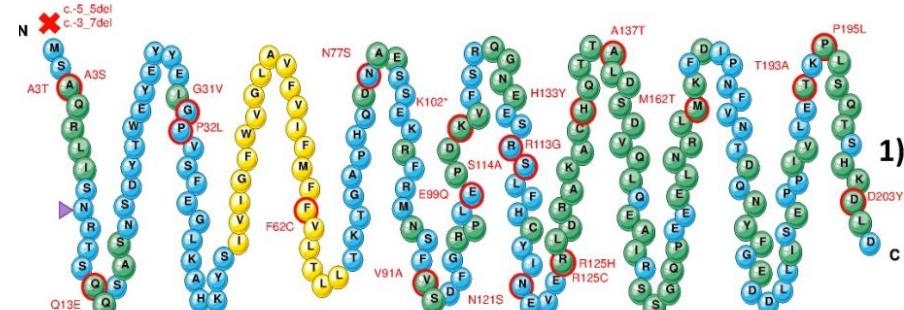
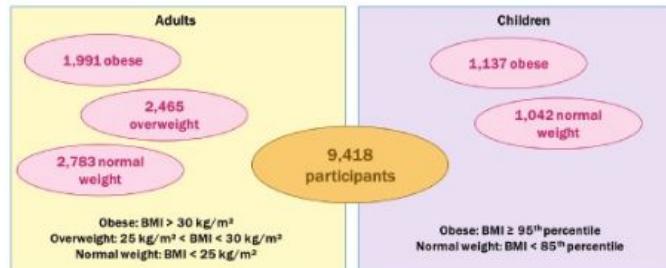
# Exome sequencing of 25000 schizophrenia cases and 100 000 controls

Tarjinder Singh

- The Schizophrenia Exome Sequencing Meta-analysis (SCHEMA) consortium (2017) - aggregating and analyzing high-throughput sequencing data
- Understand the [genetic causes of schizophrenia](#)
- Motivate the development of new therapeutics
- Life expectancy - 12-15 years shorter life expectancy
- **10 genes** that when disrupted, dramatically increase risk for schizophrenia
- Odds ratios 4 - 50,  $P < 2e-6$
- 2 genes code glutamate receptors - crucial in brain cells communication
- 10 genes have no protein-truncating variant signal

# Pathogenic loss-of-function in MRAP2 cause metabolic syndrome - A. Bonnefond

- Melanocortin receptor accessory protein 2 (MRAP2) is a transmembrane accessory protein predominantly expressed in the brain
- Deletion of Mrap2 results obesity in both mice and human
- 23 rare mutations in MRAP2
  - 2 frameshift
  - 1 non-sense
  - 20 missense



# Insight into genetic architecture of autism

Adam Rocke

- Autism sequencing consortium (SPARK)
- 18,381 autism spectrum disorder (ASD) cases and 27,969 controls
- ASD affects 1–1.5% of individuals and is highly heritable
- identifies 5 risk loci

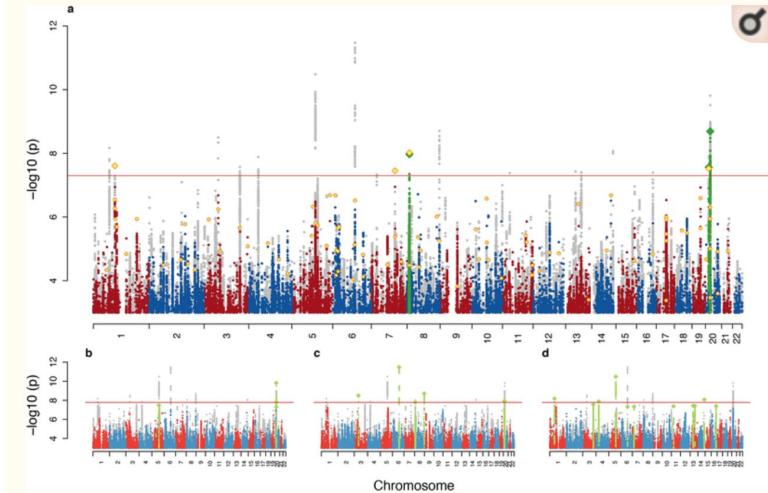


Figure 1.

Manhattans plots:

with the x axis showing genomic position (chromosomes 1–22) and the y axis showing statistical significance as  $-\log_{10}(P)$  of z statistics. a: The main ASD scan (18,381 cases and 27,969 controls) with the results of the combined analysis with the follow-up sample (2,119 cases and 142,379 controls) in yellow in the foreground. Genome-wide significant clumps are painted green with index SNPs as diamonds. b-d: Manhattan plots for three MTAG scans of ASD together with, respectively, schizophrenia<sup>14</sup> (34,129 cases and 45,512 controls), educational attainment<sup>14</sup> ( $N = 328,917$ ) and major depression<sup>14</sup> (111,902 case and 312,113 controls). See Supplementary Figures 45–48 for full size plots. In all panels the results of the composite of the five analyses (consisting for each marker of the minimal p-value of the five) is shown in grey in the background.

# Insight into genetic architecture of autism

Adam Rocke

**Supplementary Table 4:** Review of the top loci identified in the single SNP analysis with and without follow-up sample, the three MTAG analyses, and the gene-wise MAGMA analysis (see also main Table 1, Table 3, 10 and Box 1 for further details and explanations of significance tests). Loci are  $r^2 = 0.1$  clumps and the genes listed are the protein coding genes in the locus except marked with \* (see below) or on chromosome 8 where the locus around rs10099100 was restricted to a  $r^2 = 0.6$  clump plus 50kb flanks. See section 2.1.4 for details.

\* Nearest protein coding gene within  $\pm 1$  Mb from locus.

\*\* Results based on samples overlapping the present ASD sample, hence not independent.

† Previous GWS hits for brain disorders in locus.

Chr	Gene	Index SNP (BP)	Analysis P-value	Function	Tissue specificity	Disease	Previous GWS hits in locus <sup>†</sup>
1	NEGR1	rs1620977 (72729142)	MTAG MD: $6.7 \cdot 10^{-9}$	NEGR1 encodes the Neuronal growth regulator 1, a GPI-anchored IgLON protein belonging to the immunoglobulin superfamily of cell adhesion molecules. NEGR1 has been identified as a raft-associated component of the brain that involved in neurite outgrowth[72–74] and neurodevelopmental determination of synapse number in the hippocampus[75]. NEGR1 expression levels are affected by nutritional state in brain areas relevant to feeding[76] and NEGR1 may also serve a role in intracellular cholesterol trafficking[77].	NEGR1 is localized at postsynaptic sites of dendritic and somatic synapses[78] and is expressed at high levels during postnatal development in cerebral cortex, hippocampus, cerebellum, and hypothalamus[78–80].	A 1p31.1 deletion including a part of <i>NEGR1</i> (and no other genes) has been identified in two siblings with a history of neuropsychiatric and behavioral problems, learning difficulties, hypotonia, mild aortic root dilatation, hypermobility and scoliosis[81].	Meta-analysis of autism spectrum disorder and schizophrenia[24] <sup>**</sup> , Schizophrenia[82], Educational attainment[50], Depressive symptoms[62], Major depression[48], Intelligence[83]
1	PTBP2*	rs201910565 (96561801)	Comb: $2.5 \cdot 10^{-8}$	PTBP2 encodes Polypyrimidine tract-binding protein 2. PTBP2 is a PTBP1 (Polypyrimidine tract-binding protein 1) paralog and is also known as nPTB (neuronal PTB) or brPTB (brain PTB). PTBP1 and PTBP2 binds to intronic polypyrimidine clusters in pre-mRNA molecules and each target large sets of exons to coordinate programs of splicing events during development[84]. During neuronal development and differentiation, several switches in the expression of <i>PTBP1</i> and <i>PTBP2</i> activates networks of new spliced isoforms[85–88].	PTBP2 is expressed at high levels in adult brain, testis, myoblasts and lymphocytes[89, 90]. Different isoforms generated by alternative splicing of <i>PTBP2</i> are expressed in a tissue specific manner[90, 91]. In neuronal cells, <i>PTBP2</i> acts by autoregulating its own exon 10 inclusion, leading to an increased expression level of the PTBP2 protein[92]. PTBP1 promotes the exon 10 exclusion from <i>PTBP2</i> transcripts, leading to NMD of <i>PTBP2</i> transcripts[93].	Hemizygous deletion of varying genomic regions containing <i>PTBP2</i> , <i>DPYD</i> and <i>MIR137</i> have been reported in cases with ASD, severe speech delay and intellectual disability[94, 95]. Two brothers with ASD and intellectual disability born to consanguineous parents were found to be homozygous for a novel 5 bp indel variant located in a human accelerated region (HAR) between <i>DPYD</i> and <i>PTBP2</i> [96]. This intergenic indel variant was suggested to affect the <i>PTBP2</i> -directed enhancer activity of the region and that this effect was relatively specific to neurons[96]. Further evidence pointing to <i>PTBP2</i> as a ASD risk gene include identification of a <i>de novo</i> <i>PTBP2</i> potentially damaging missense variant in an ASD proband[97] as well as identification of a <i>de novo</i> <i>PTBP2</i> intronic insertion in a ASD proband from a simplex family[98].	Educational attainment[50]

# Golden state killer

**SEEKING  
INFORMATION**

East Area Rapist/Golden State Killer  
California  
1976 to 1986

**UNKNOWN SUSPECT**



# Golden state killer

- <https://www.gedmatch.com>, tutorial
- applications for comparing your DNA test results with other people
- Genealogical Data Communication
- Software developed by the Church of the Latter Day Saints 
- [Site found 10-20 distant relatives](#) of the killer, roughly, equivalent of third cousins
- “When you go that far back in time, you have trees that grow huge,” Holes said.
- Census data, old newspaper and a gravesite locator relatives, websites such as LexisNexis.

## DNA Applications:

- One-To-Many Beta - give it a try
- One-To-Many DNA Comparison Result
- One-to-One Autosomal DNA Comparison
- One-to-One X-DNA Comparison
- Admixture (heritage)
- Admixture / Oracle Population Search
- People who match me or 1 of 2 kits
- DNA File Diagnosis Analyze DNA file uploaded
- Are you related?
- 3-D Chromosome
- Archaic DNA



# Golden state killer



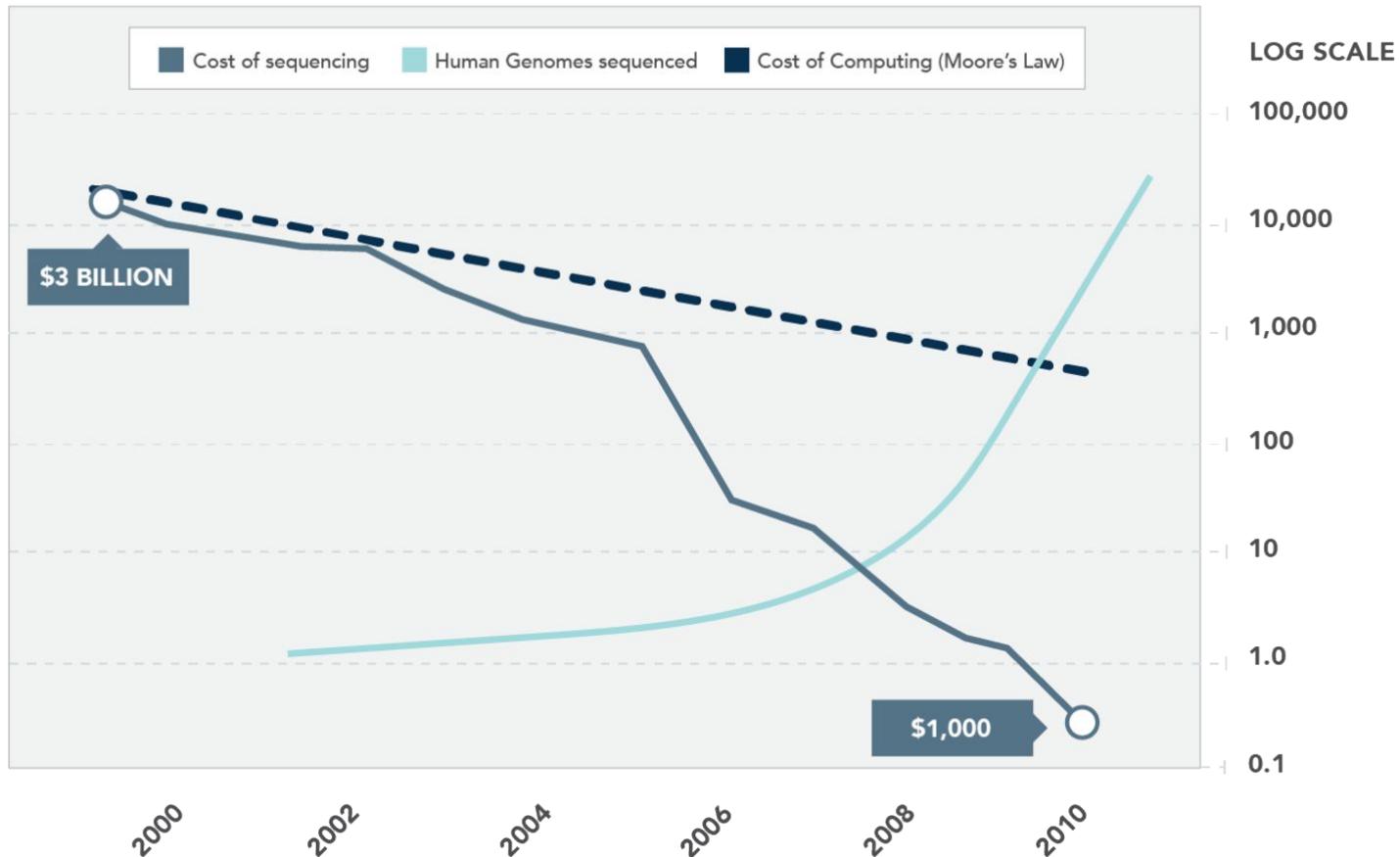
Joseph James DeAngelo

# Genome sequencing

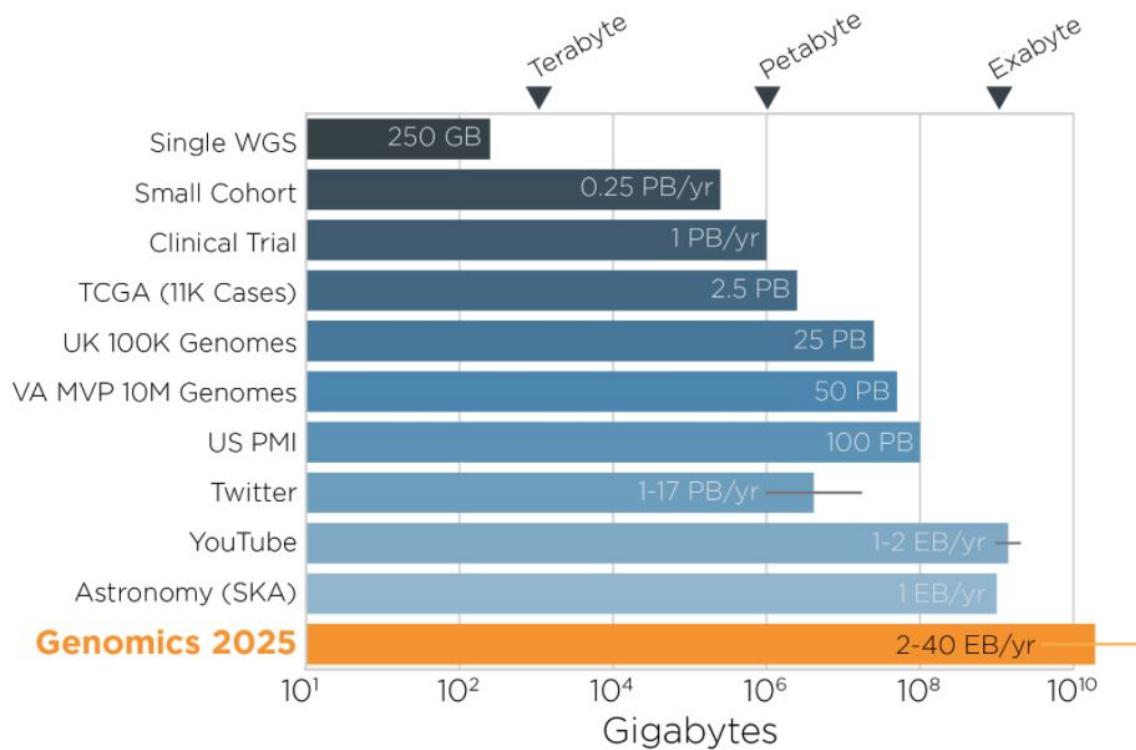
- Digitalization of genome
- **Human Genome Project** (1990-2003), 3B \$
- Birth of bioinformatics
- Sanger sequencing (First generation sequencing)
  - Long (took 13 years)
  - Costly (3B\$ for one human genome)
- Currently NGS (next generation sequencing)
  - Illumina
  - Around 200\$ and 1 day needed to sequence the genome
- Also third generation sequencing in use
  - Longer read-length (up to 50k base)
  - Oxford nanopore, PacBio
  - Higher error rate
  - Smaller in size
  - Sequencing in space



## GROWTH OF DNA SEQUENCING



# Genomics is Big Data



Source: "Big Data: Astronomical or Genomical?" *PLoS Biology* (2015).

## Sequences:

1 zetta-bases/yr

## Storage needs:

2-40 exabytes

## Compute for Alignment:

**10,000 trillion** CPU hrs  
= 83x time since Big Bang

## Variant Calling:

**~2 trillion** CPU hrs

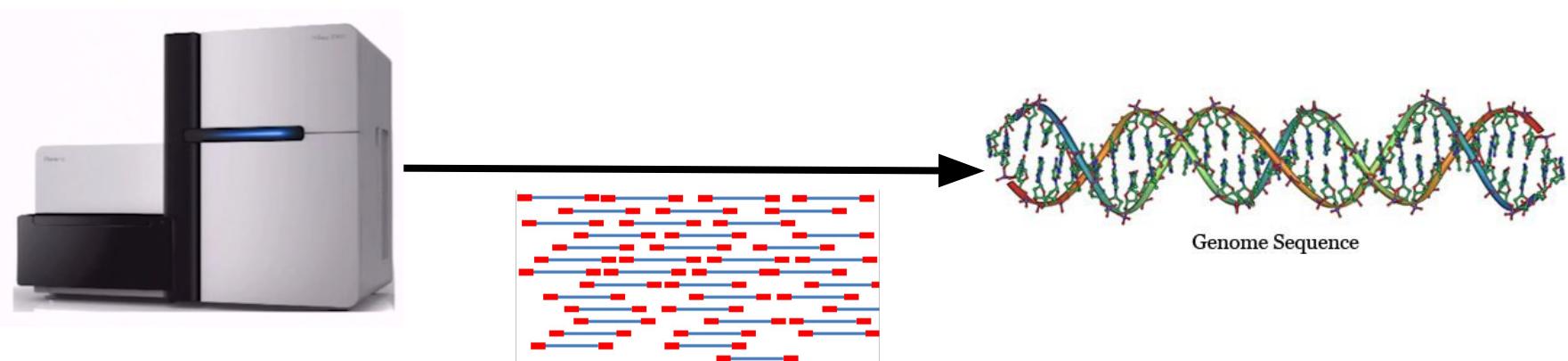
## Tertiary Analysis:

**~4 trillion** CPU hrs

= time since land-breathing mammals evolved

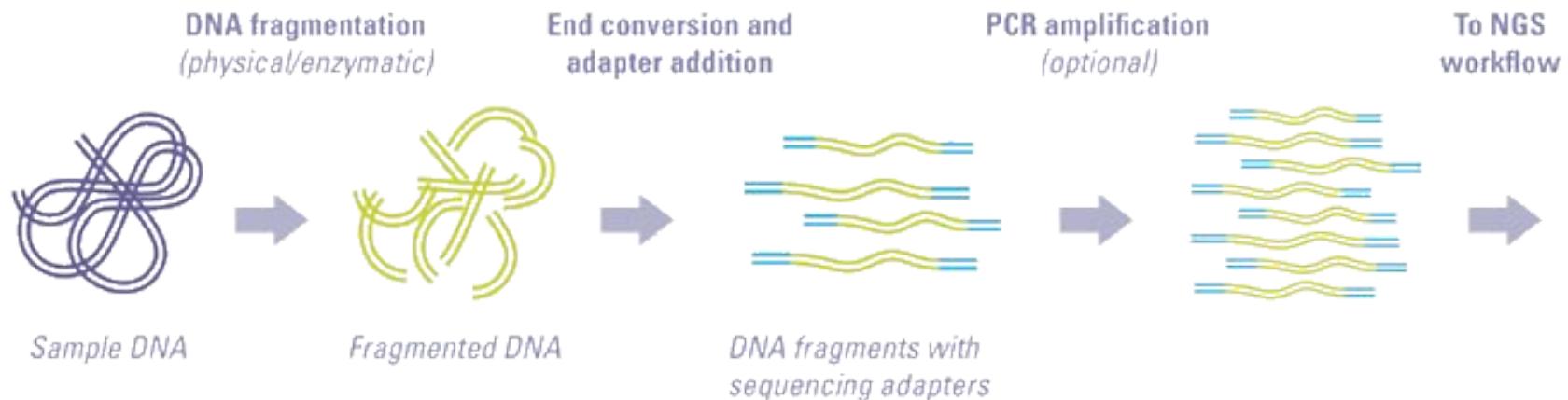
# Bioinformatics to the rescue!

- Genomes of all species are arrays of nucleotides (A, T, C, G) - strings
- The process of DNA sequencing returns only fragments of it
- Our mission: RECONSTRUCT IT!

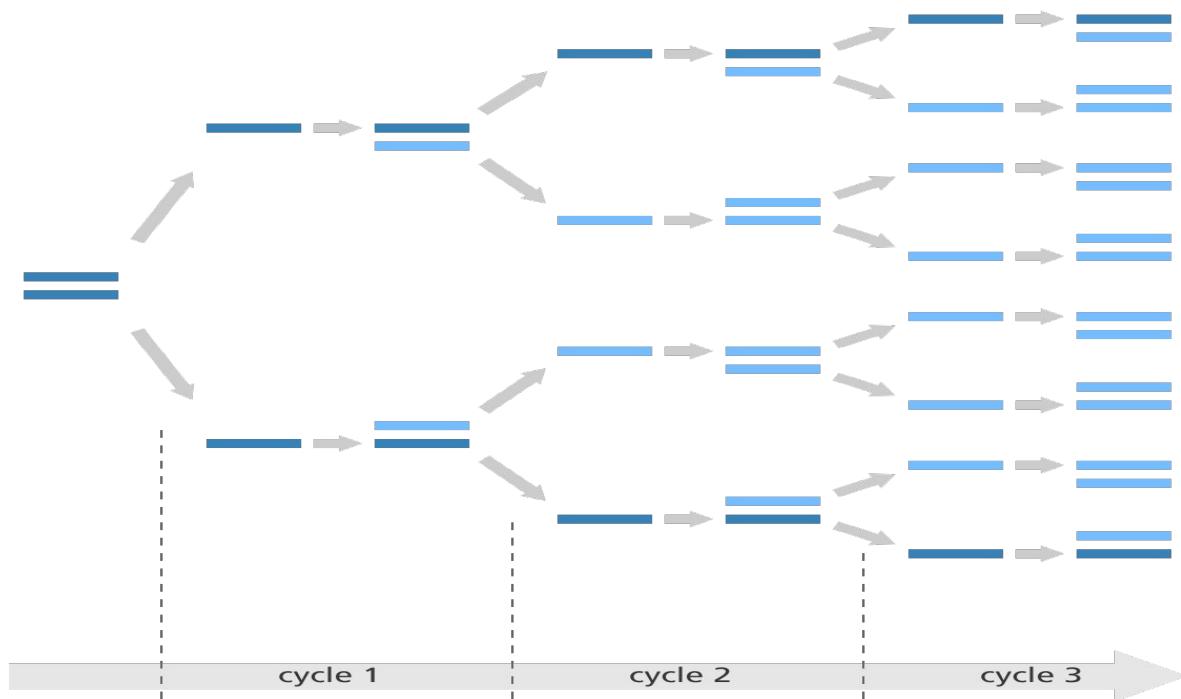


# Illumina sequencing

- Read - DNA fragment after reading it in sequencer
- Typical whole genome sequencing experiment:
  - 200-500 million reads
  - 150-250 bases (letters long)

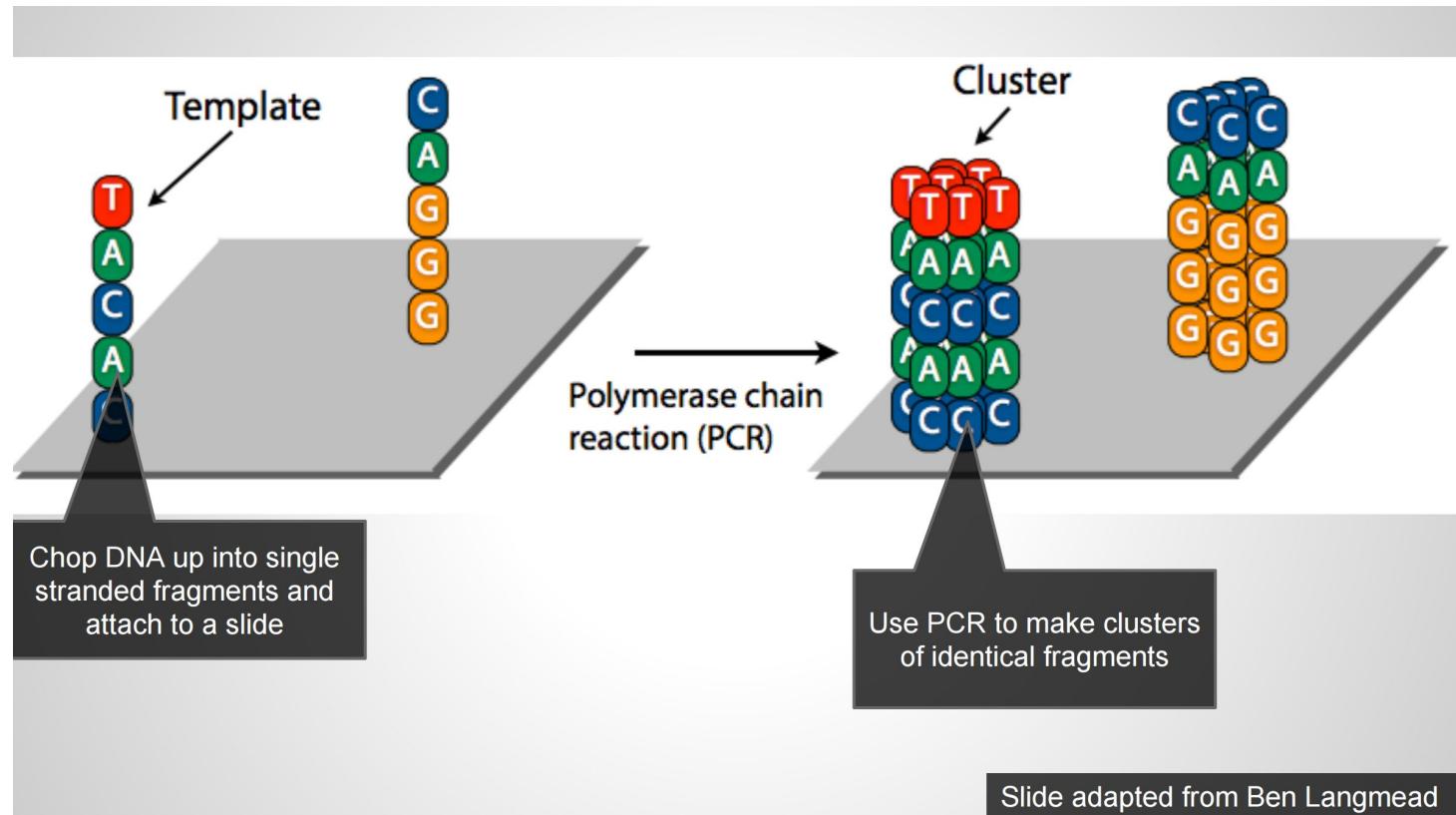


# Sequencing - PCR (polymerase chain reaction)

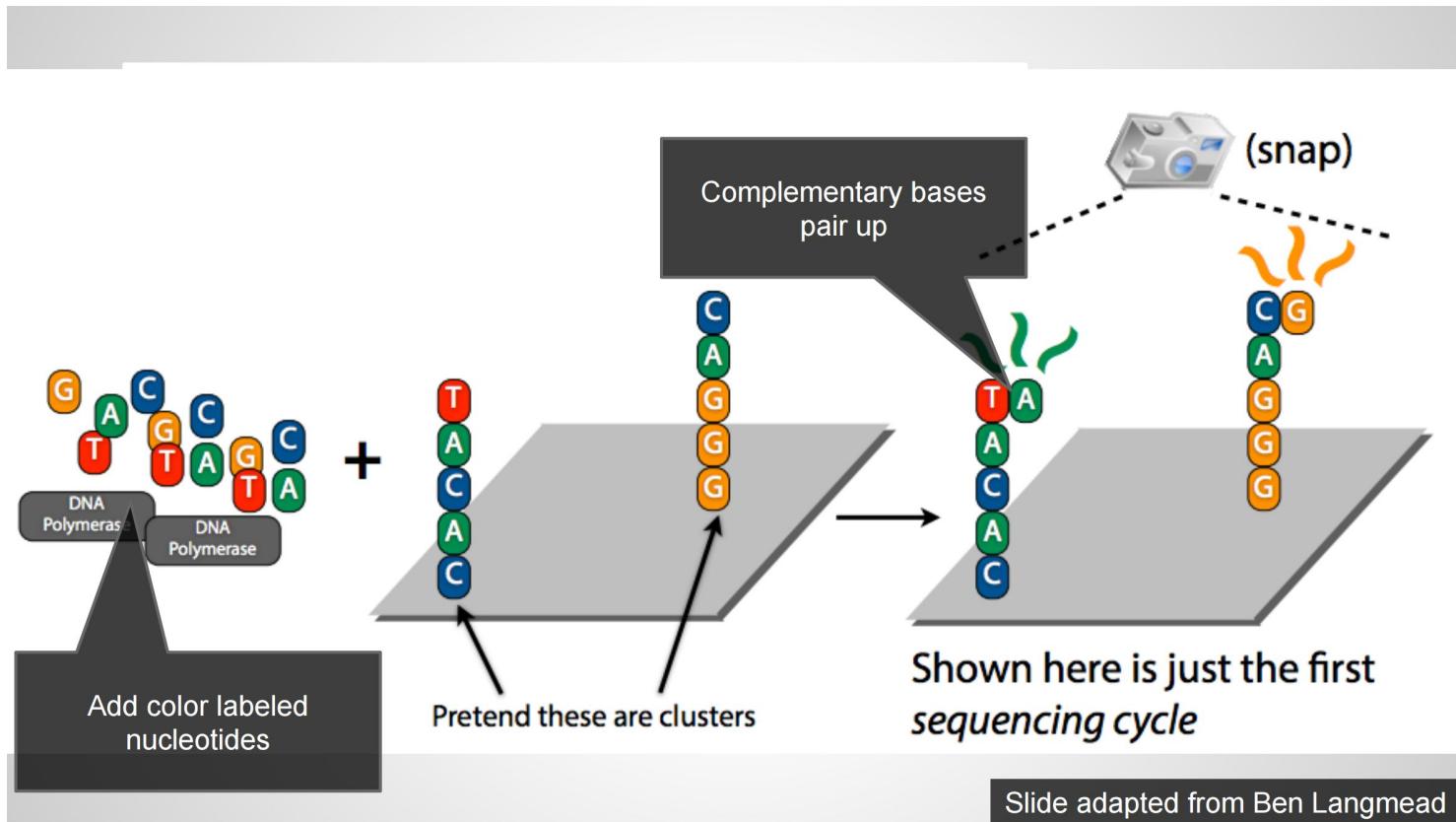


Bridge amplification

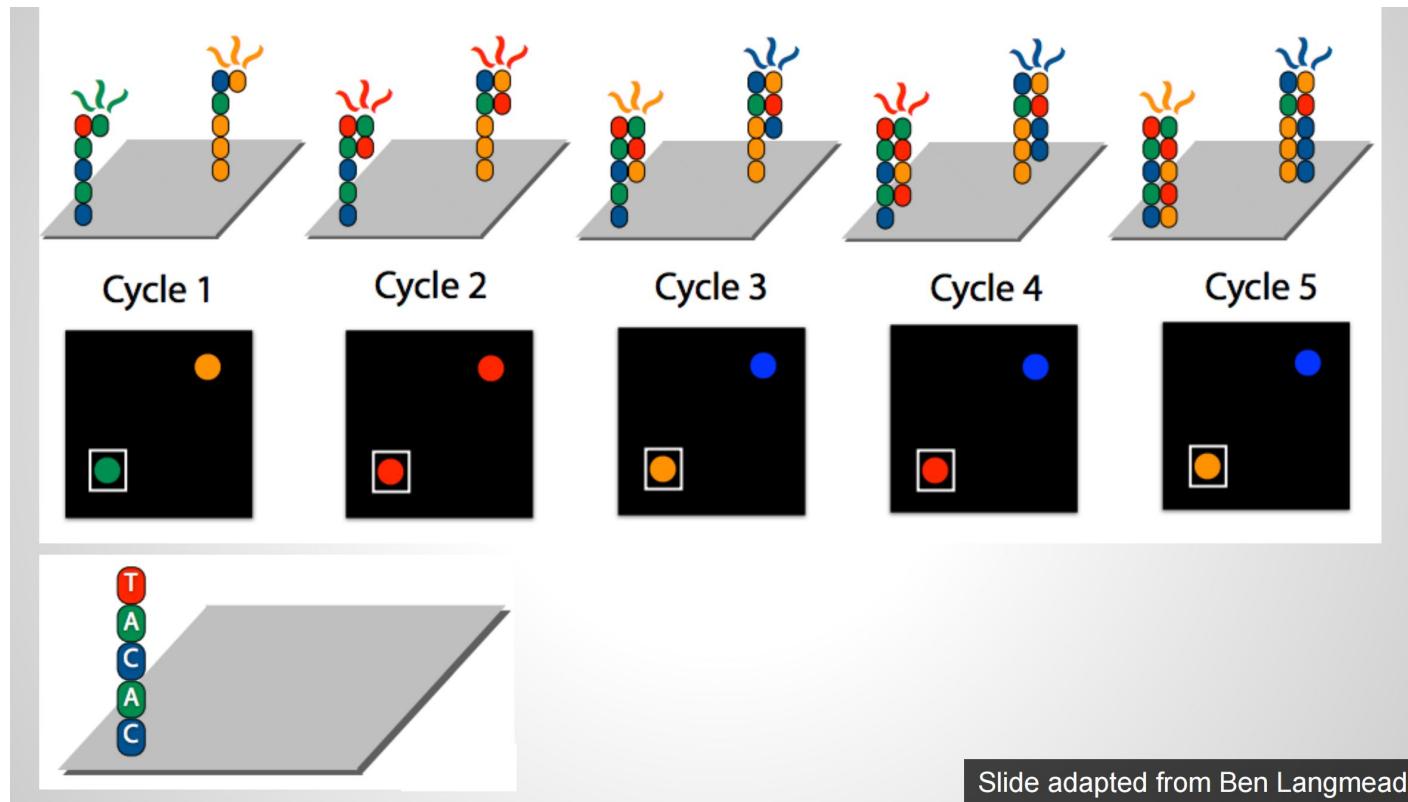
# Sequencing (Illumina)



# Sequencing (Illumina)

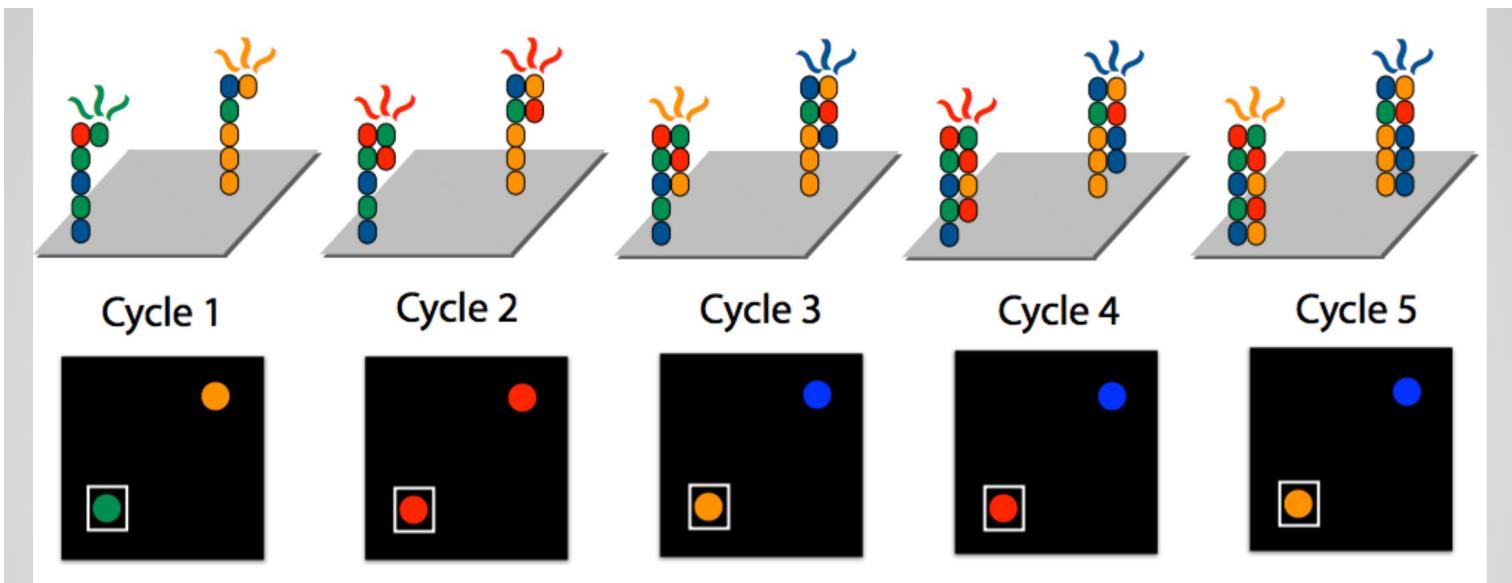


# Sequencing (Illumina)

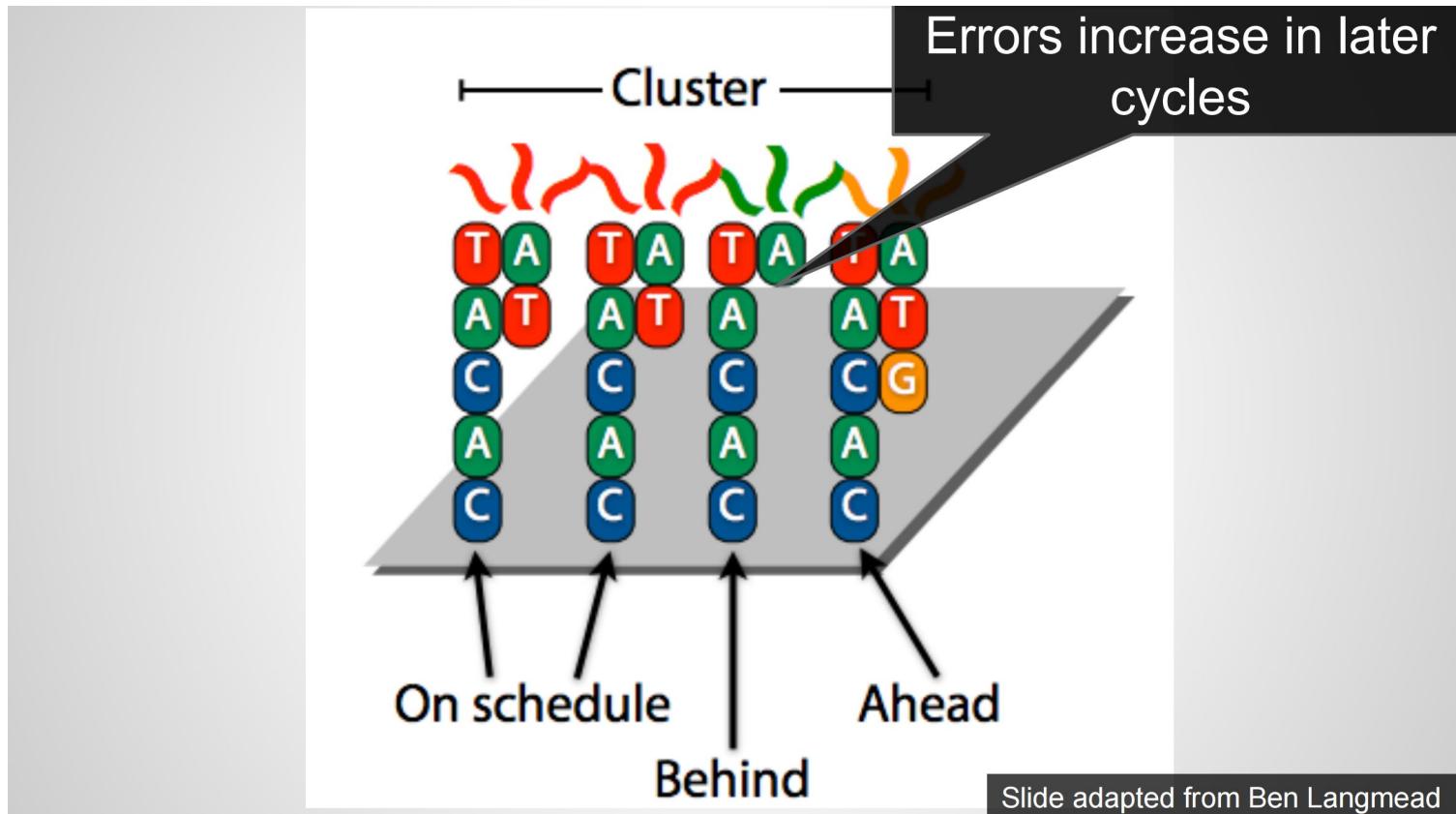


Slide adapted from Ben Langmead

# Sequencing (Illumina)

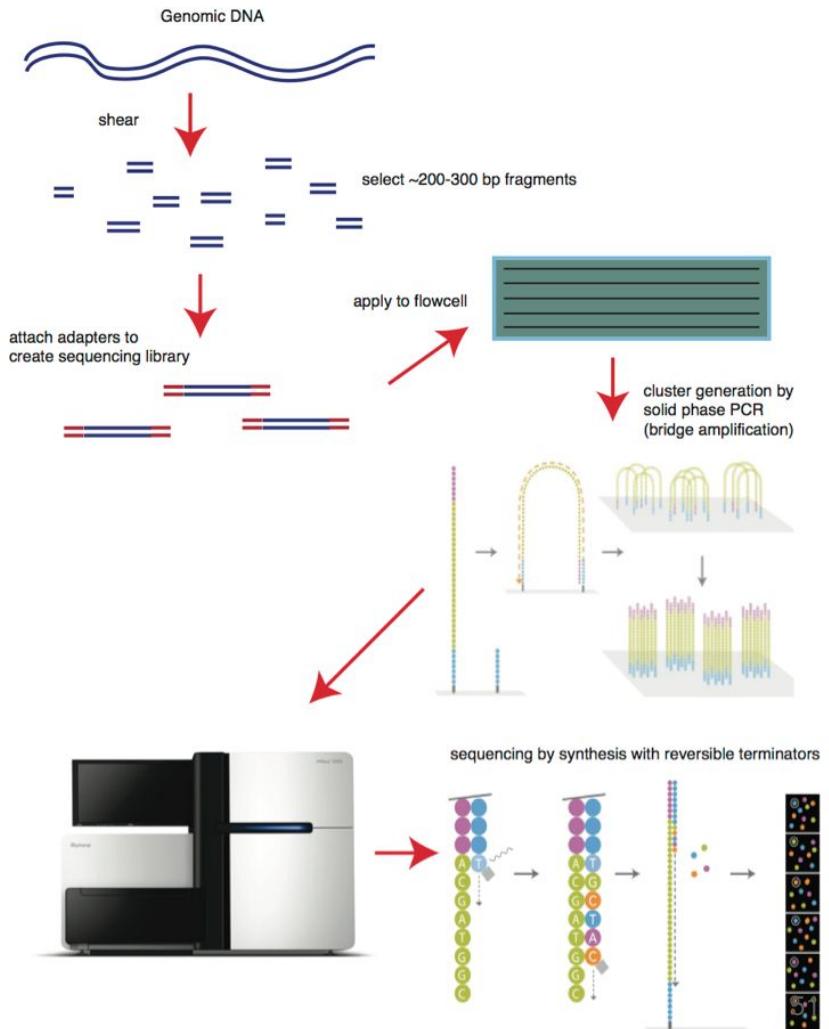


# Sequencing error

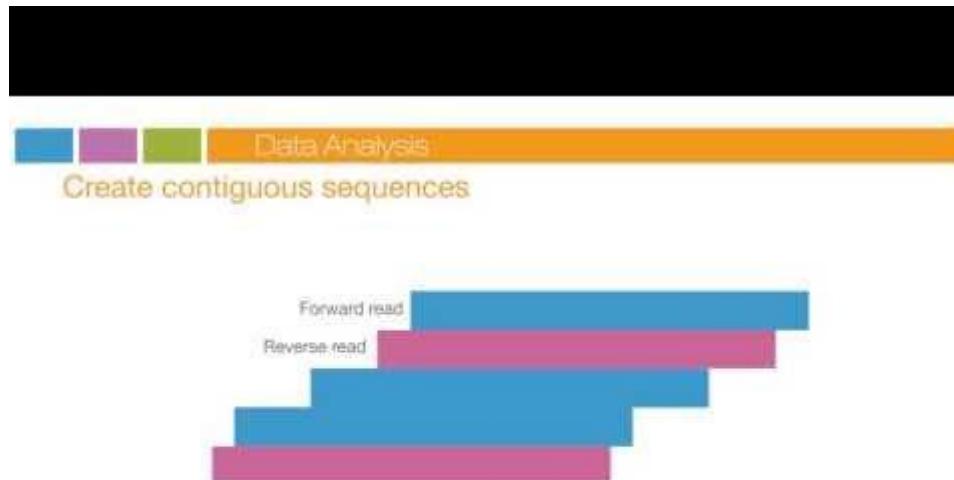


# Sequencing (sum up)

1. Shearing (fragmentation of the genome)
2. Attaching adapters
3. PCR amplification (optional)
4. Attaching template to surface/flowcel
5. PCR/bridge amplification (cluster creation)
6. Adding fluorescent bases and taking a picture after each cycle (repeat this many times)
7. Stack up images and read the sequence



# Illumina sequencing

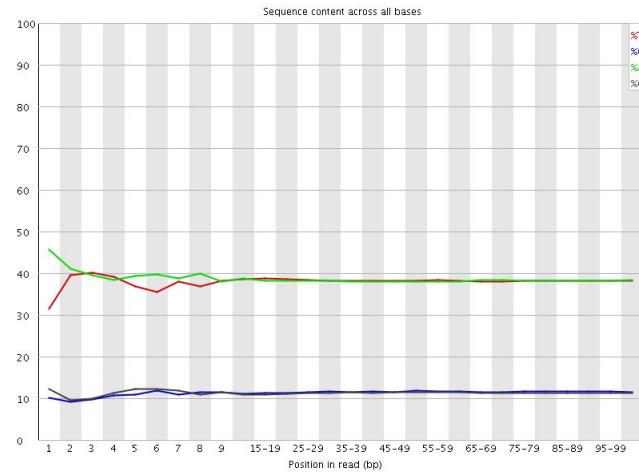
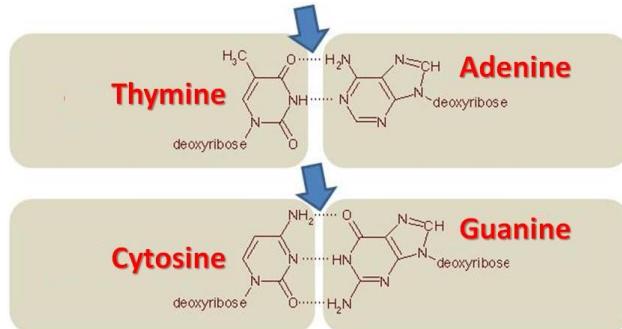


# Sequencing errors

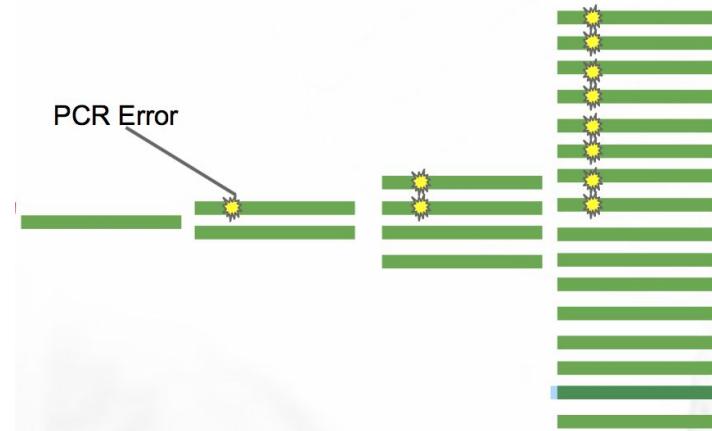
## 1. GC bias

35% to 60% - human

~20% - Plasmodium falciparum

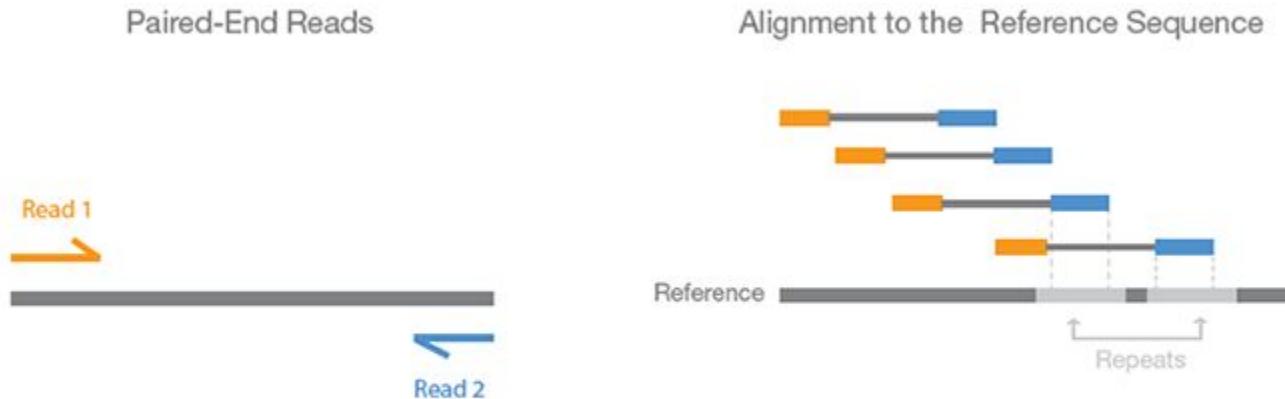


## 2. Error propagation (1 in 10.000 error rate)



# Paired-end sequencing

Figure 4. Paired-End Sequencing and Alignment

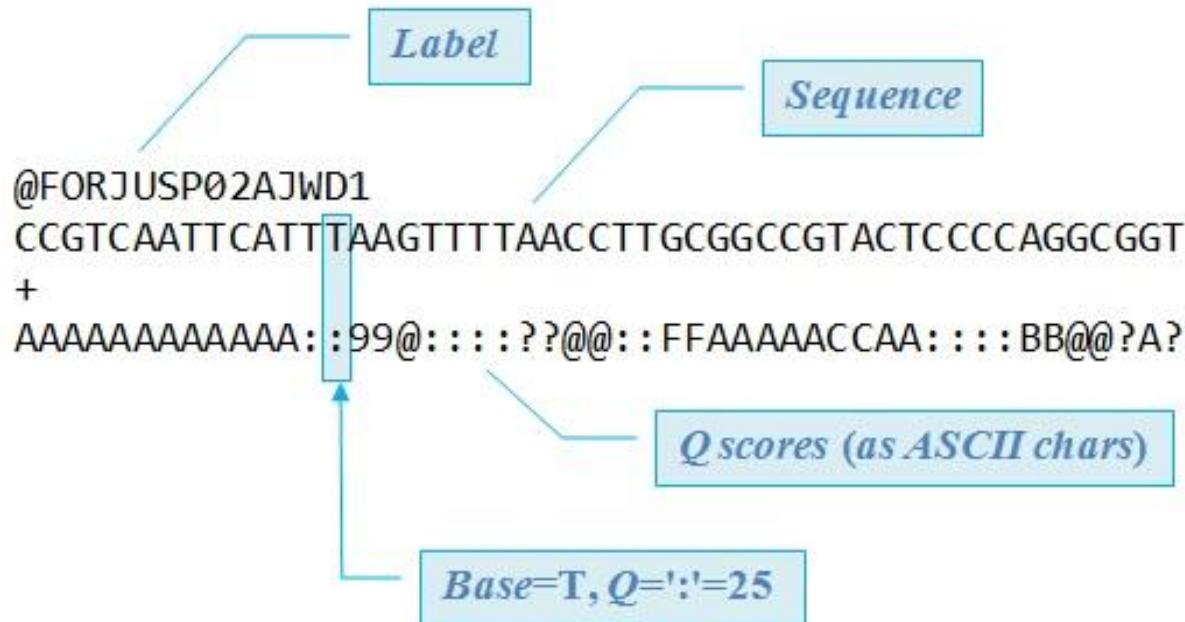


Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

# Sequencing data - FASTQ file

4 lines for each read

- Read id
- Read sequence
- + sign
- ASCII encoded quality



## Sequencing data - FASTQ file

# Genome reconstruction

Result of sequencing experiment

- FASTQ file
- 100-500 GB
- Each read(line) containing a genome sequence 50-250 bp long



# Genome reconstruction

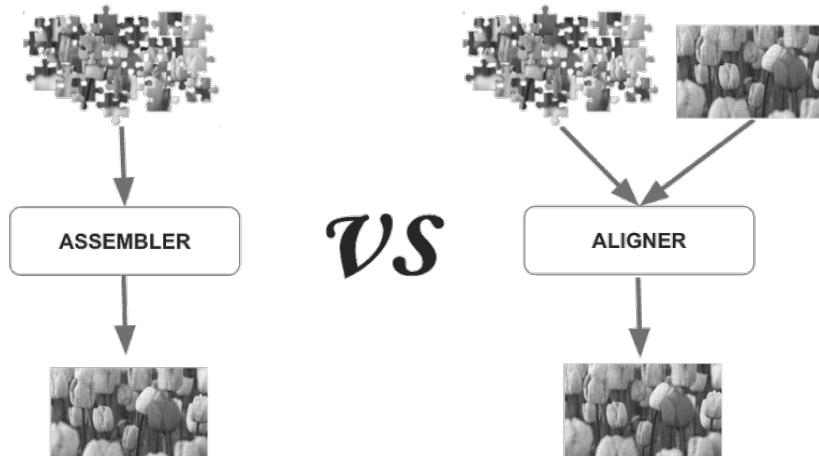
How do we reconstruct genome from reads?

## 1. Alignment

- Using reference genome to map the position of the reads

## 2. Assembly

- Reconstructing the genome by finding the links between the reads



# Alignment

AAGGACAAGA	TCTTTTATG	
ATGA <del>CCAC</del>	<del>GA</del> ATGC <del>AAGG</del>	<del>CCAC</del> <del>A</del> TCTTT
ATGATTAGA		

# Assembly

AAGGACAAGA TCTTTTATG  
ATGA~~CCAC~~ GAATGC~~AAGG~~ CCAC~~A~~TCTTT  
ATGATTAGA

# Resources and additional reads

Presentation available at: [github.com/vladimirkovacevic/gi-2020-etf](https://github.com/vladimirkovacevic/gi-2020-etf)

- [A Computer Scientist's Guide to Cell Biology, A Travelogue from a Stranger in a Strange Land](#)
- [Genomics 101, Edition 2016](#)
- [Bioinformatics at COMAV - SNP Calling](#)
- [Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM](#)
- [High-Throughput Sequencing Technologies - Review paper](#)
- Vince Buffalo: Bioinformatics Data Skills
- Dan Gusfield: Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge
- Pavel Pevzner, Neils Jones: An Introduction to Bioinformatics Algorithms (Computational Molecular Biology), MIT
- R. Durbin, S. Eddy, A. Krogh, G. Mitchinson: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press
- Veli Mäkinen, Djamal Belazzougui, Fabio Cunial, Alexandru I. Tomescu: Genome-Scale Algorithm Design: Biological Sequence Analysis in the Era of High-Throughput Sequencing, Cambridge University press