DNA analysis in clinical genetics. A role of bioinformatics.

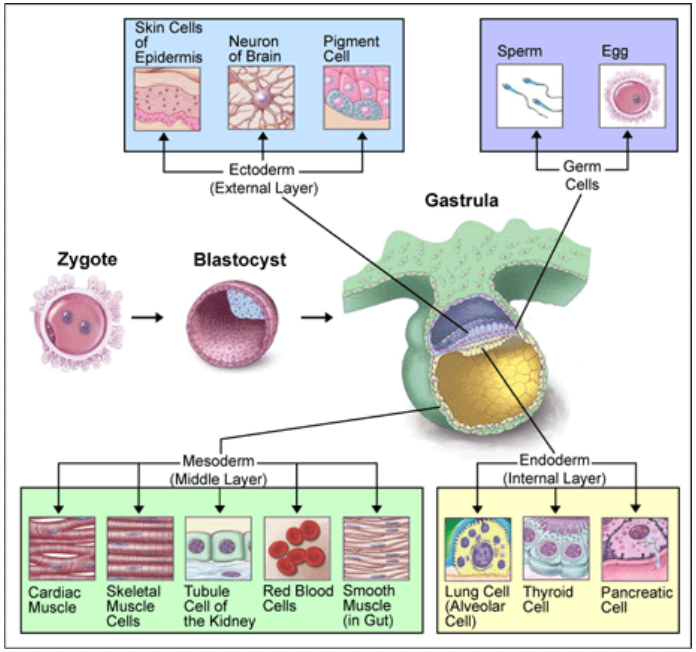
Monika Gos

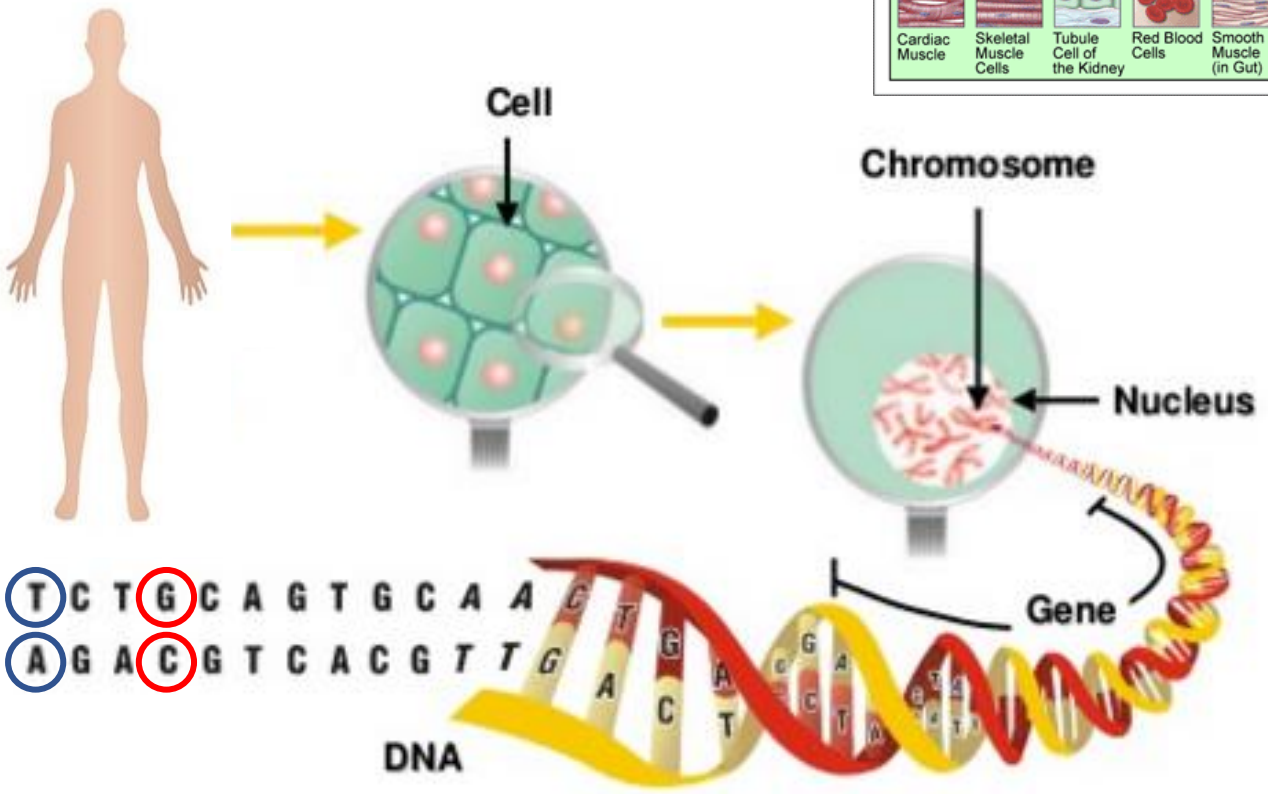
Department of Medical Genetics, Institute of Mother and Child, Warsaw

Head: prof. dr hab. n. med. Jerzy Bal

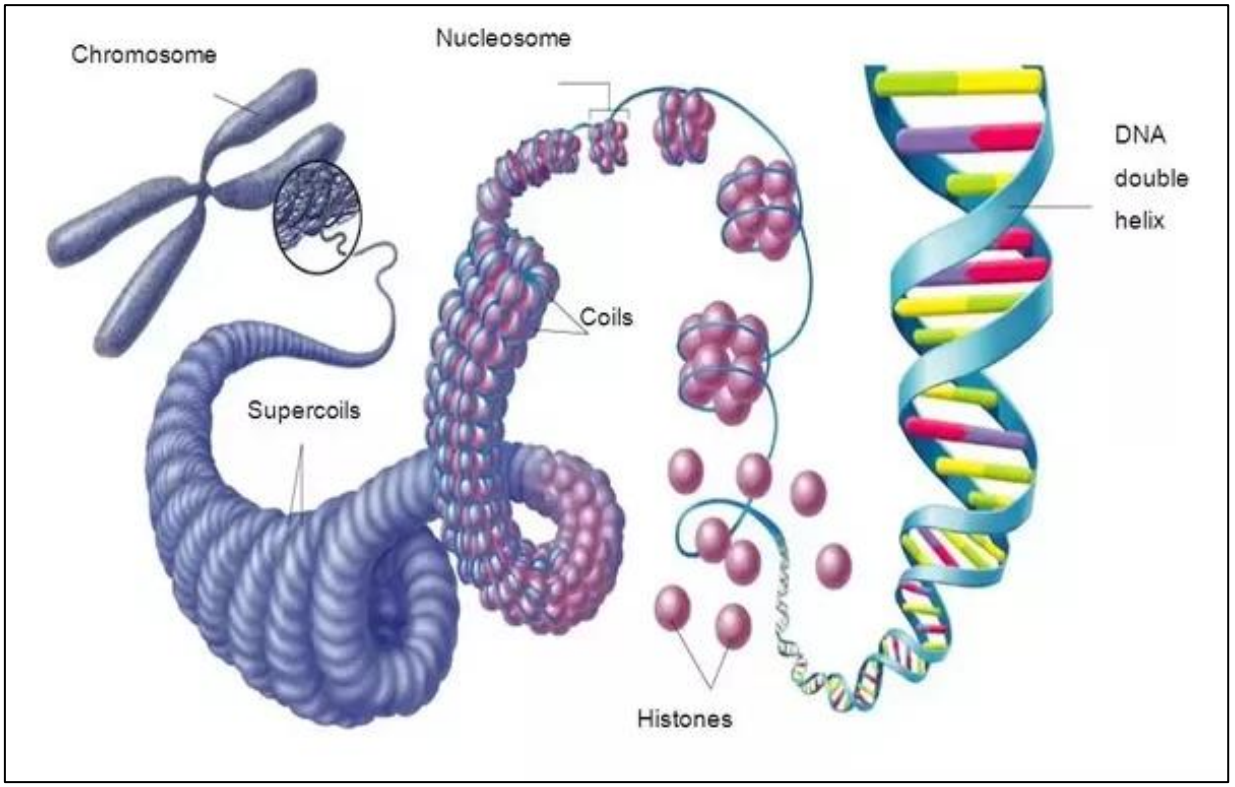
07/03/2018

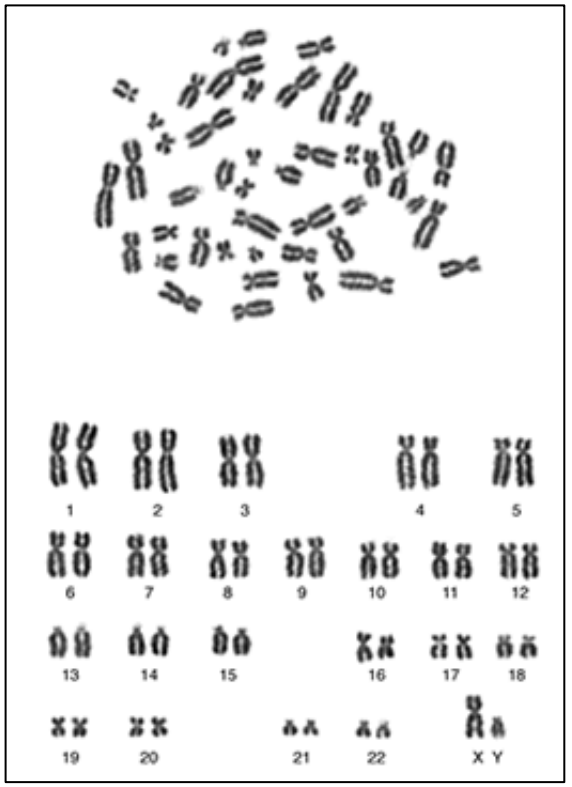
Human body consists of 5×10^{12} to 7×10^{16} cells
(mean: 3.72×10^{13})
≈200 types of cells





Almost each cell of teh organism contains the same genetic information (DNA)

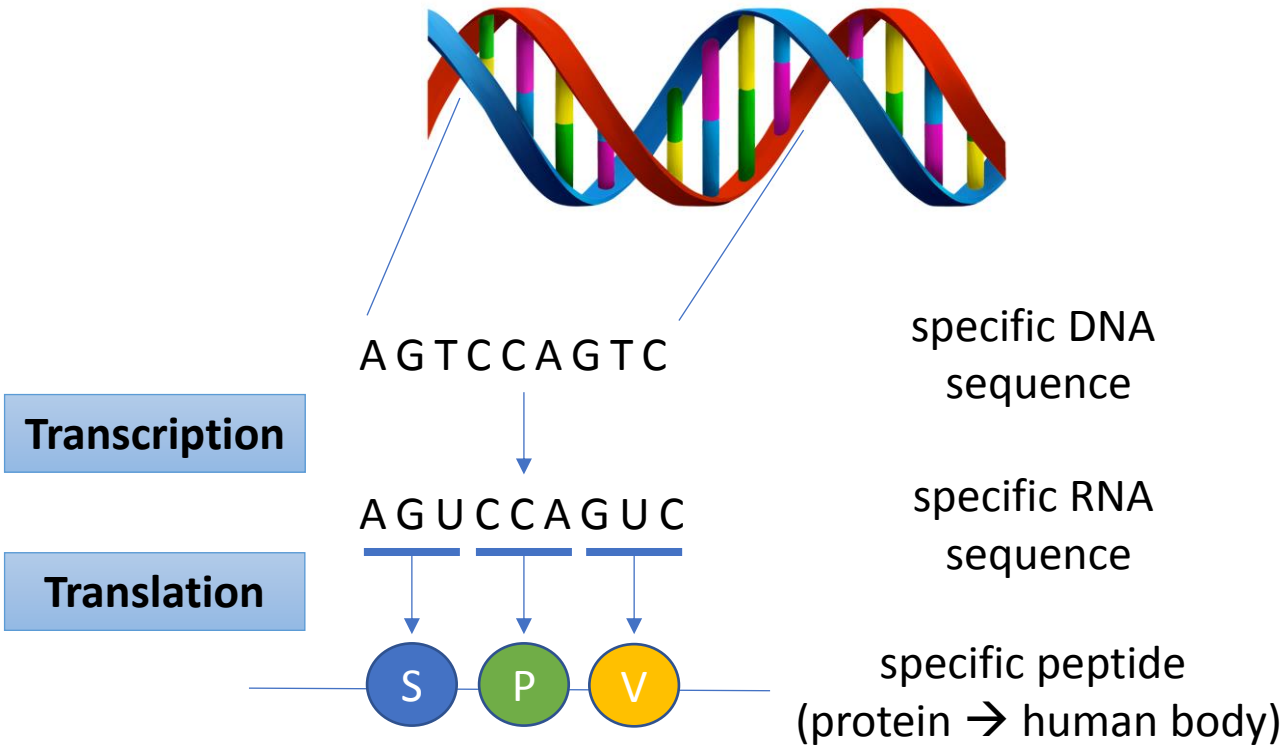




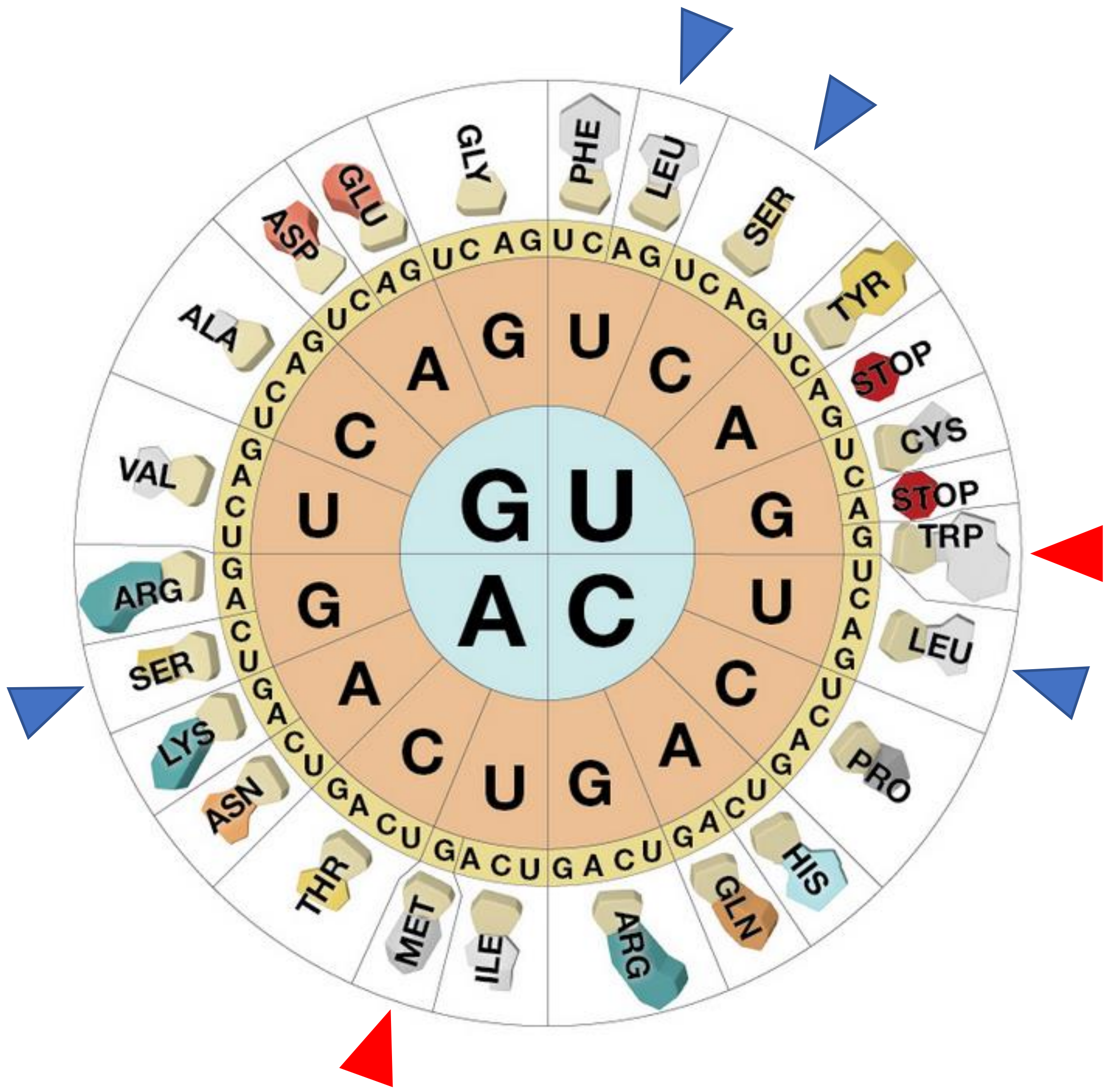
GENOME: 3 bilion base pairs (bp) - 46 chromosomes (22 autosomal + 2 sex chromosomes)
≈ 2m DNA / cell
2.0×10^{13} meters of the DNA in all human body cells (70x sun-earth-sun)
The DNA is packed – histone core – 146bp (wrap)

GENE – DNA fragment (RNA in some viruses) encoding specific protein or functional RNA.
In common sense: a DNA fragment that determines a specific feature of the organism

How 4 bases (A, T, G, C) form an organism that has 3.72×10^{13} cells?
THE GENETIC CODE – a system that is used to describe genetic information based on the combination of 3 bases from the 4 (64 potential combinations)

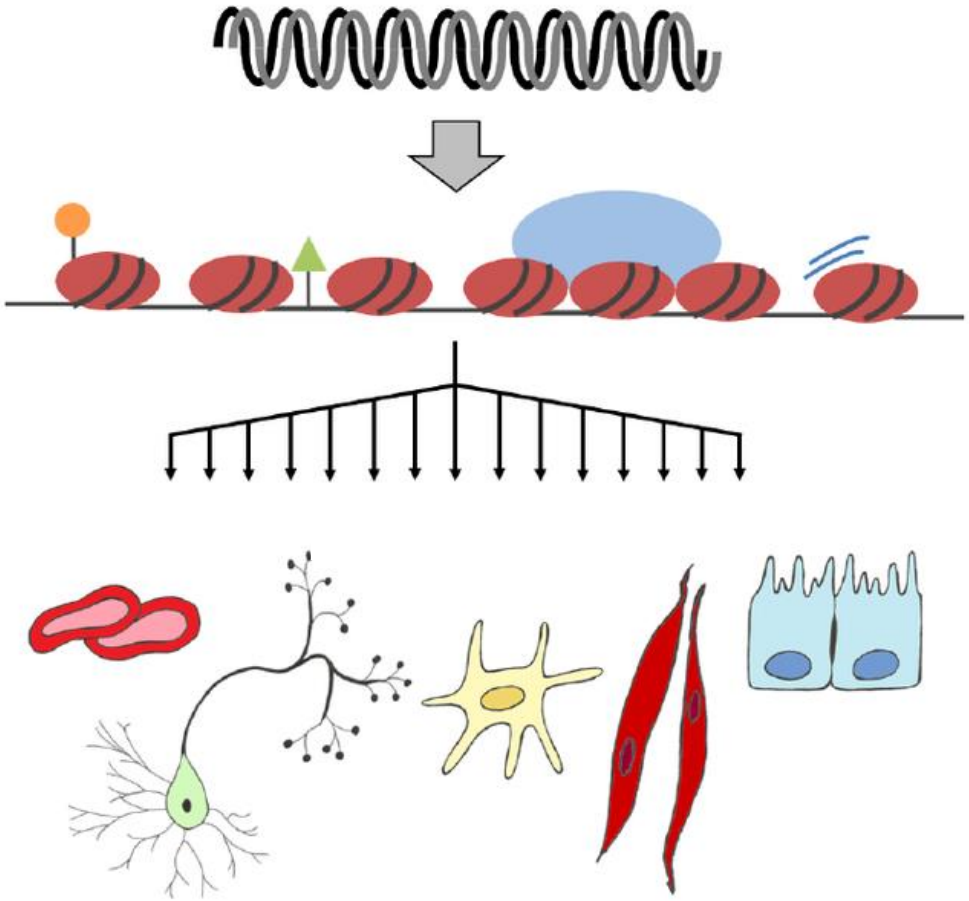


In practice: 64 3-letter combinations –> 21 amino acids + STOP codon



In each cell the genetic information is the same.
Why the cells are different?

Different gene expression (transcription) – time- and tissue- specific
epigenetic regulation – silencing of gene expression during embryonic development
EPIGENOME – DNA methylation, histone modification, chromatine structure (eu- i hetetro-)



EPIGENOME

modyfikacje epigenetyczne

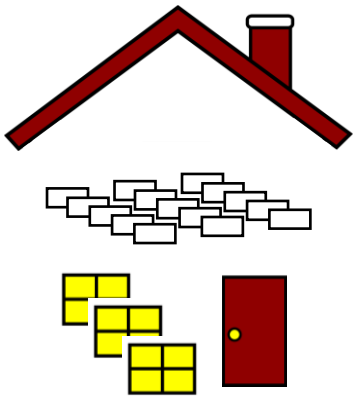
GENOME | TRANSCRIPTOME | PROTEOME | METABOLOME

DNA → **RNA** → **protein** → **organism**

PROJECT
„house”

- brick
- roof
- window
- door

ORDER



PRODUCTION

PRODUCT

Genome: 3 000 000 000 bp ($3x10^9$) – 25 000 gene

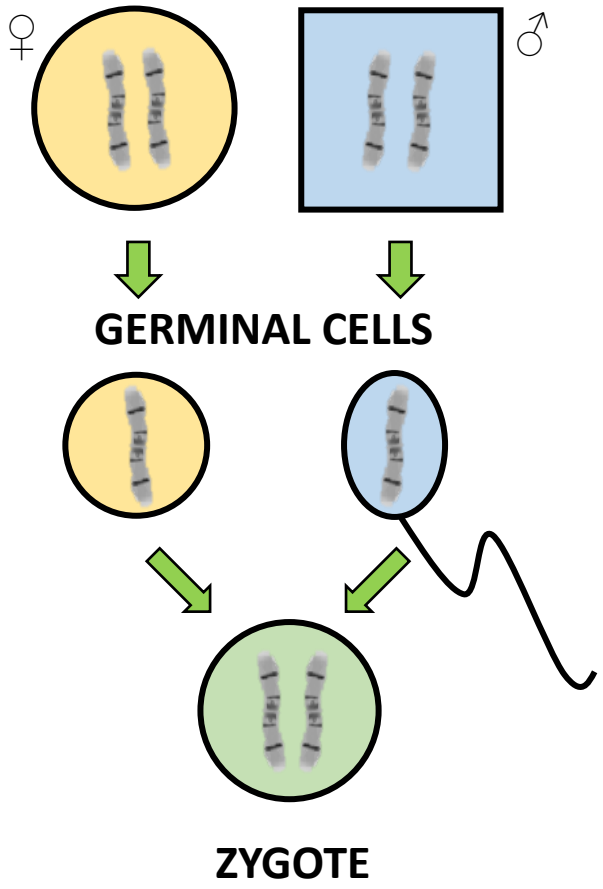
Mean gene length: 16,3 x 10^3 bp

x gene number = 4,07x10^8bp

1.5% genome – unique genes, protein/RNA coding (**EXOME**)

>95% genome – non-coding DNA (e.g. repeated sequences)

Gene inheritance



Each person has a pair of identical chromosomes in each cell and two copies of each gene
(exception: X chromosome in males)

Each parent passes
one chromosome of each pair
(one copy of the gene)
to their pedigree.

As a consequence, their progeny
has two chromosomes
– one inherited from the mother,
the other inherited from the father.

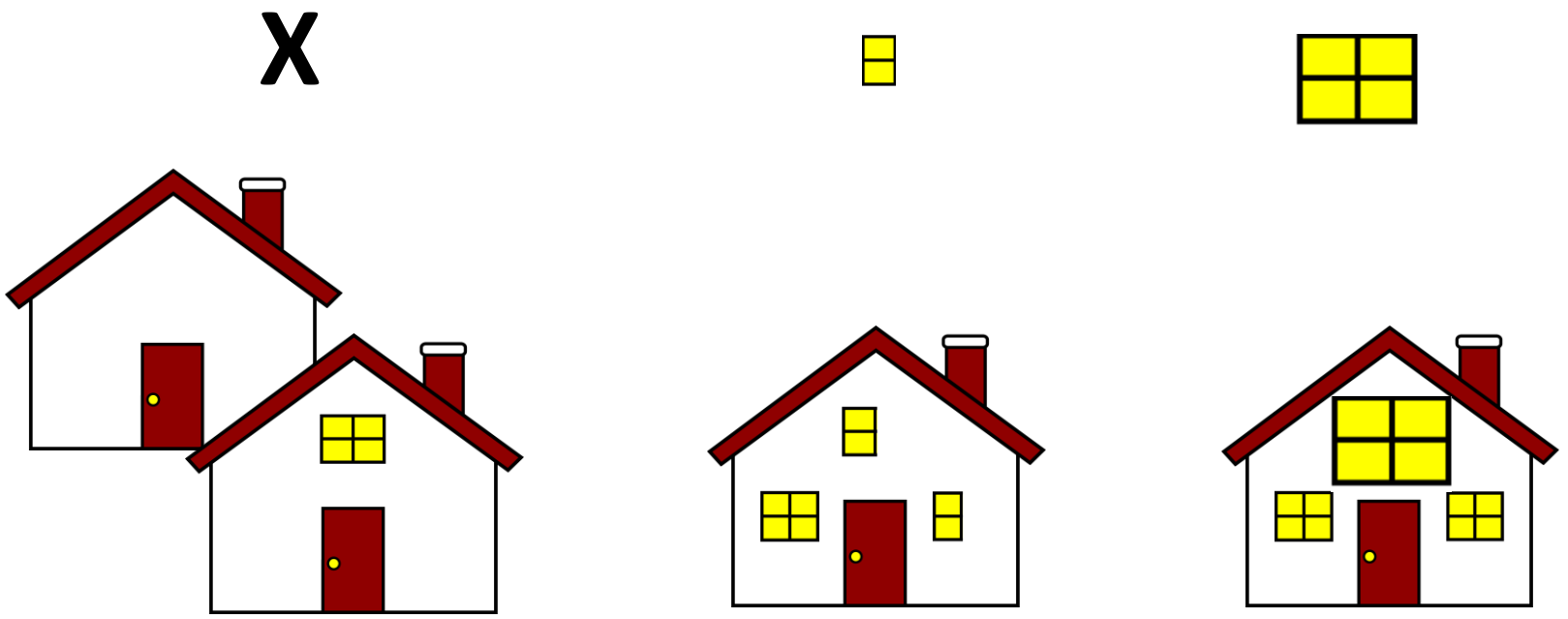
GENOTYPE vs. PHENOTYPE

MUTATION

dynamic change of genetic information within the cell (spontanic/induced)
phenotypic effect: neutral, beneficial, unfavourable (lethal, sublethal)
inheritance: somatic i germline (diveristy source)
types: chromosomal aberrations (numeric, structural) or gene alterations (m. point)

No protein

Abnormal protein



INHERITED DISORDERS

Disease caused by the mutation within the specific gene or genes (chromosmes), that has an impact on proper organism development and functioning.

Besides the counselling and clinical examination (that includes interview about the diseases in patients family) and routine diagnostuic testing (imaging scans, biochemical testing), the genetic tests are performed that aim to identify the genetic defect responsible for the disease.

GENETIC COUNSELLING UNIT (clinical geneticist)

BUT: you should never forget that similar phenotype can be obserevd as an effect of other non-genetic conditions (environmental causes – infections, autoimmunological)

Why genetic tests shoul be performed?

confirmation of the clinical diagnosis –
the proper diagnosis

genetic counselling (family members)

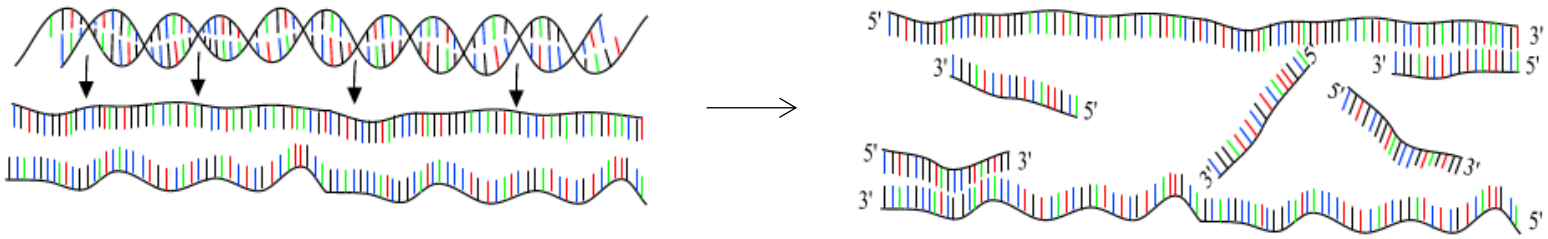
disease prognosis

therapeutic targets (personalized medicine)

With time, modern techniques are implemented into genetic testing…

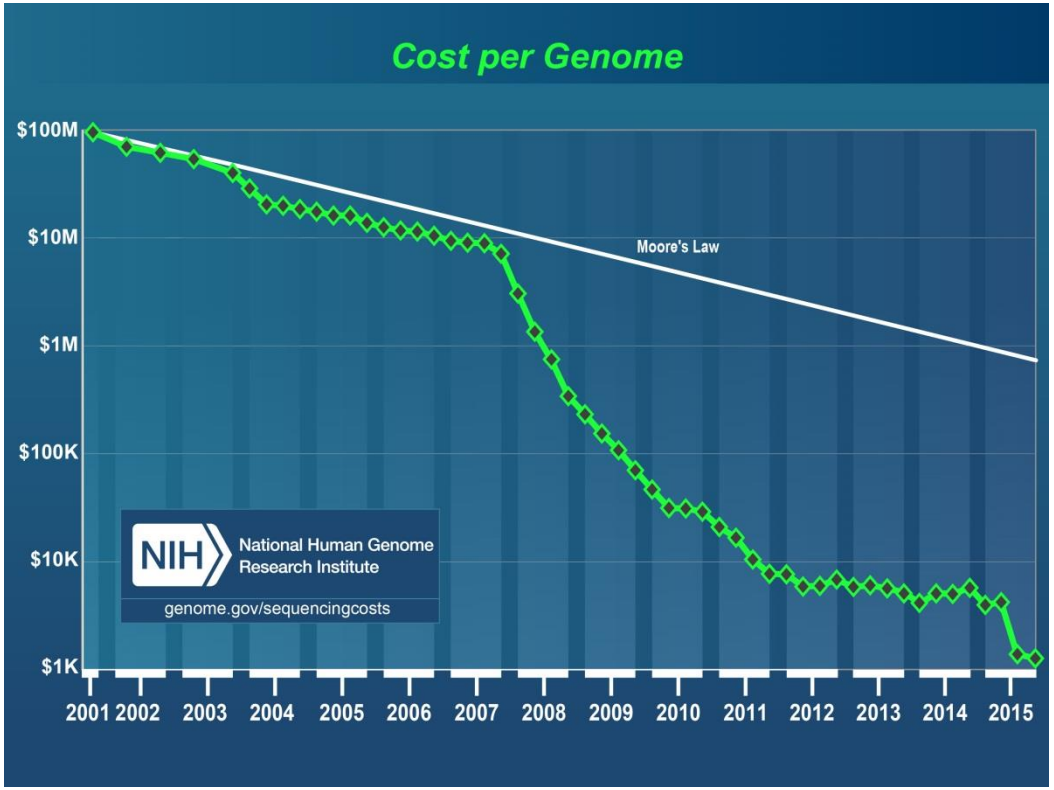
mutation → **gene** → **genome**

All techniques are based of the ability of DNA to form complementary structures…



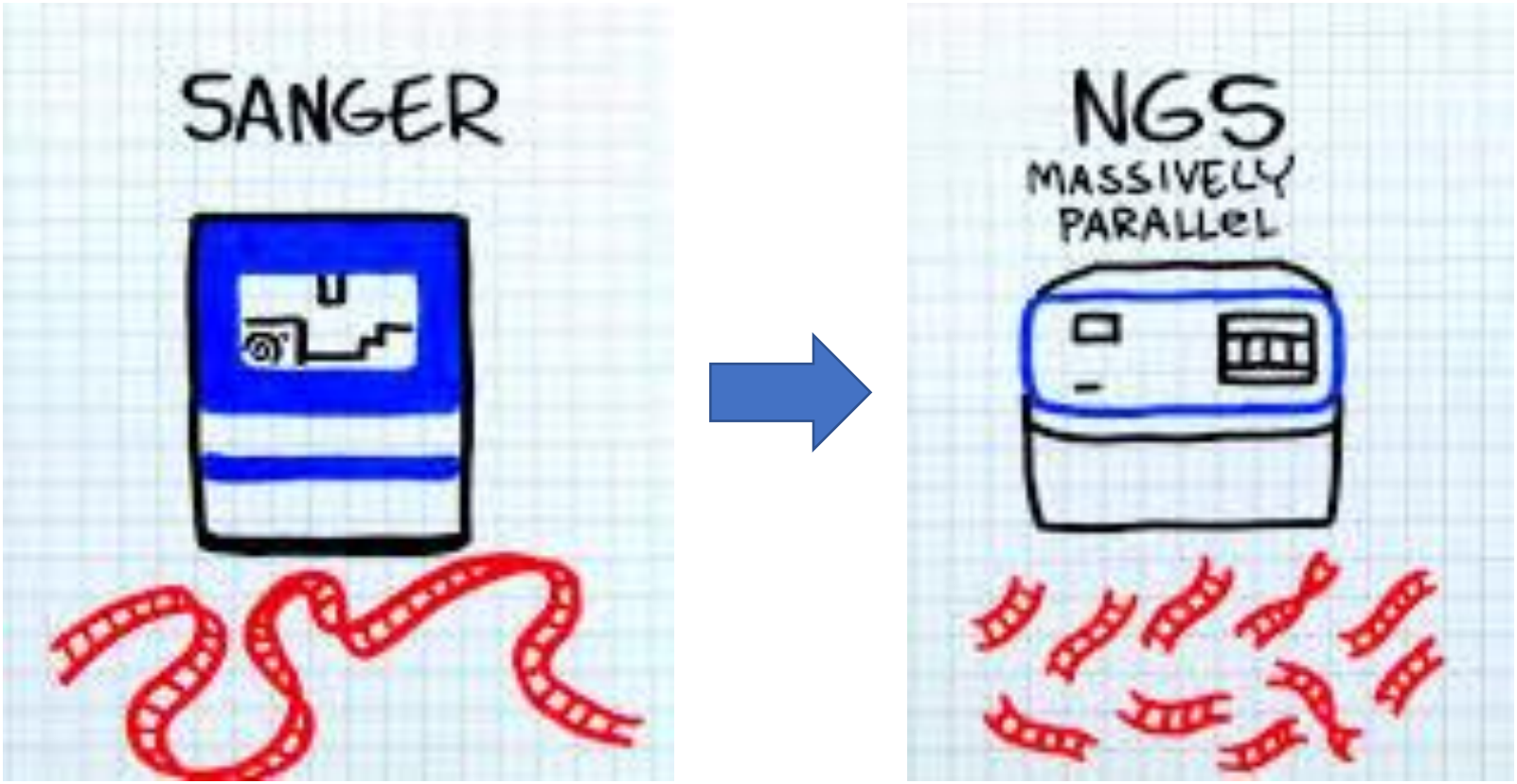
Genetic revolution: 2001 – human genome published

Genome	Method	Cost
Human Genome Project (13y)	Capillary sequencer	2,7 mld $
Venter Genome (9 months)	Capillary sequencer (>340tys. reactions)	70 mln $
James Watson	Roche, 454 (234 reactions)	1 mln $
James Lupski	Life/APG (3 reactions)	75 tys. $



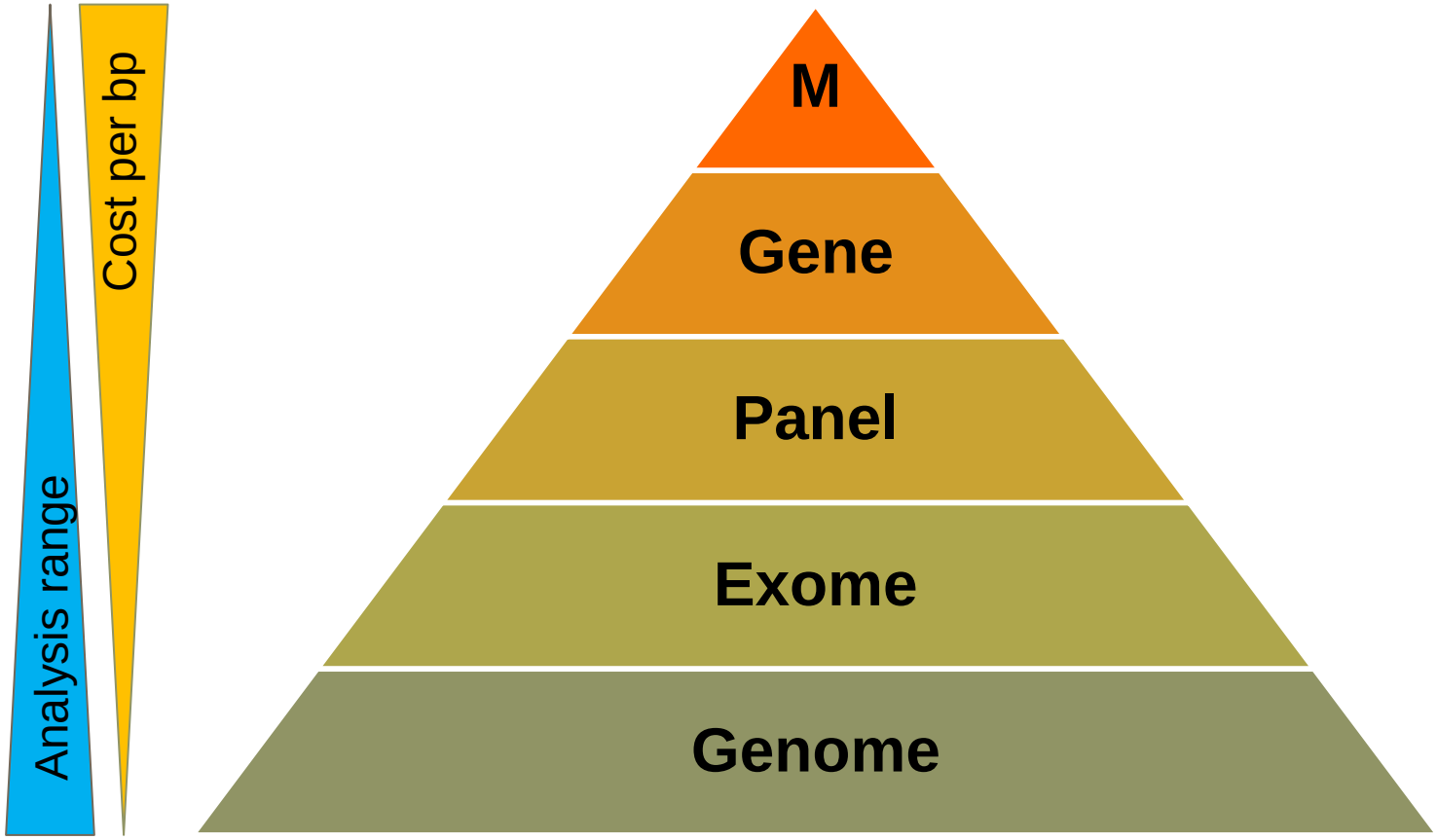
Breakthrough: High throughtput techniques implementation

Array comparative genomic hybridization (aCGH) – CNV analysis
Next generation sequencing (NGS) – analysis of SNV, CNV, structural changes

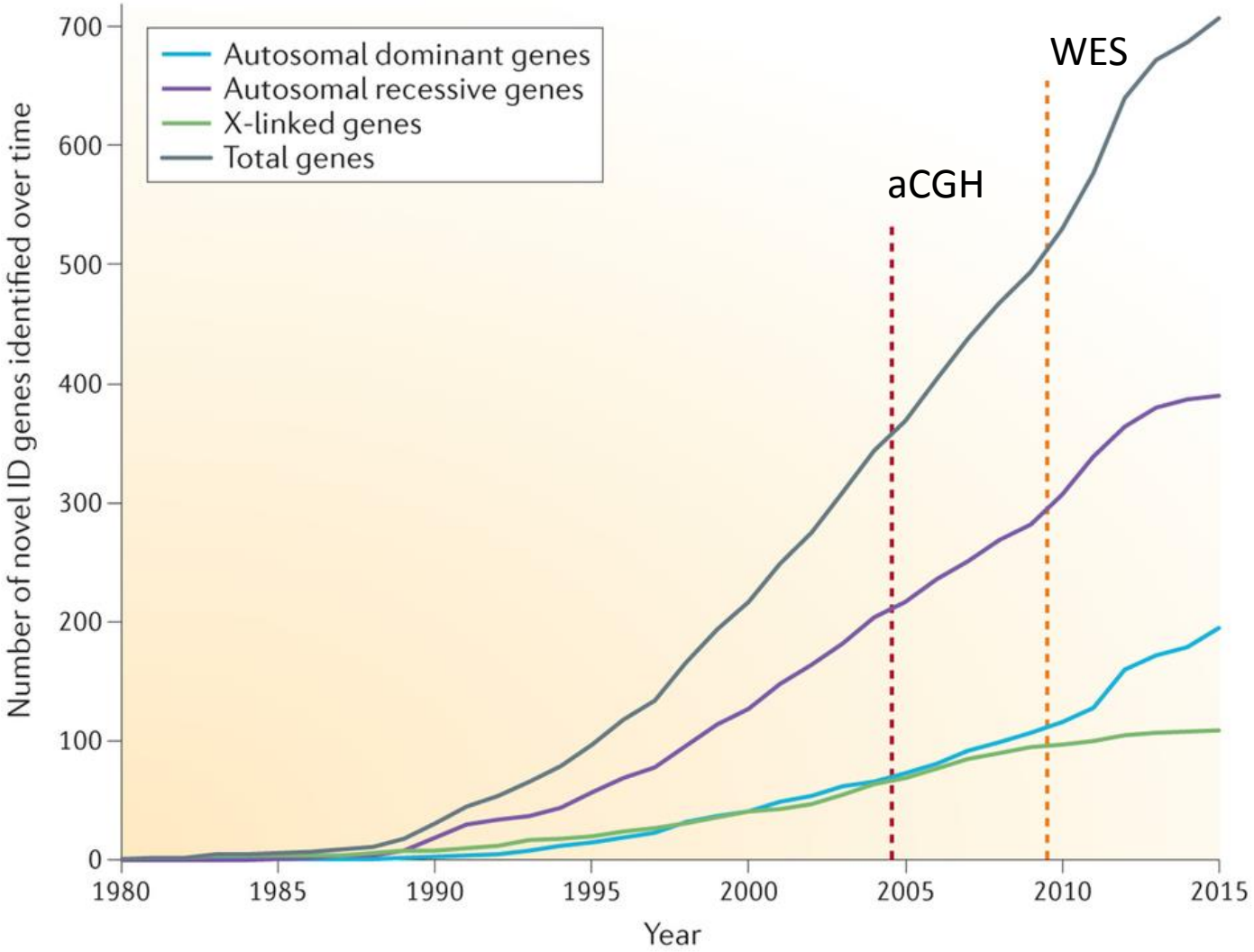


Massive parallel sequencing
Primary: genome sequencing *de novo ;* now: re-sequencing

Together with new techniques, the throughput of the analysis increased while the cost of the anaysis dropped down

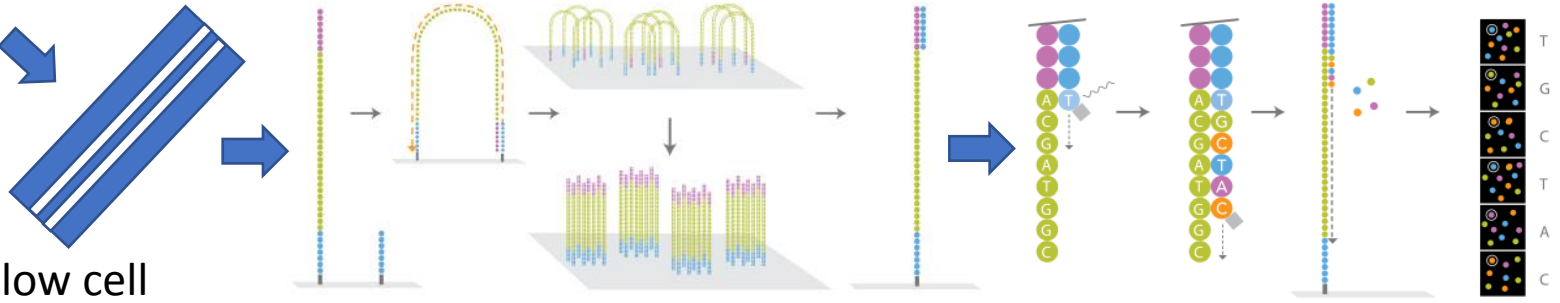


The implementation of new molecular techniques – an increase of the clinical entities with known genetic background



Sequencing by synthesis – Illumina technology

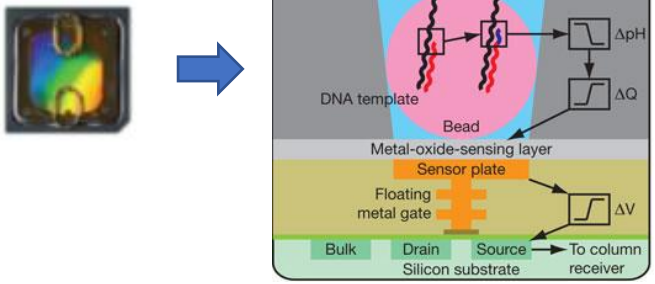
Library
(DNA)

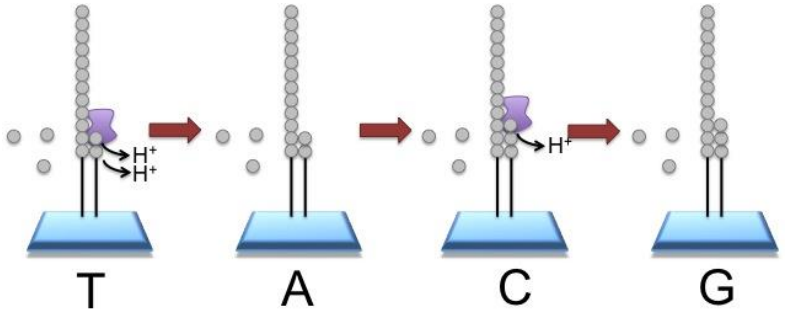


Flow cell

Sequencing by synthesis - Ion Torrent technology

Library
(DNA)





Single sequencing reaction – **read**
Number of reads per nucleotide – **coverage**
Sequence length - **read length**

Bioinformatic analysis – pipeline

*.bcl file → *.fastq file (CASAVA)

assembly with the reference genome (hg19)

Data annotation and pipeline can be „automated" (bioinformatics), but the final analysis and data interpretation towards specific phenotype is made by a human (diagnostician together with clinical specialist)

*.vcf file (e.g. GATK, SAMTools, Annovar)

functional analysis – population databases (1000 genomes, EVS/NHLBI, ExAC, in-house), clinical databases (OMIM, ClinVar, HGMD) + prediction tests

Excel file (*.tsv, *.txt, *.csv)

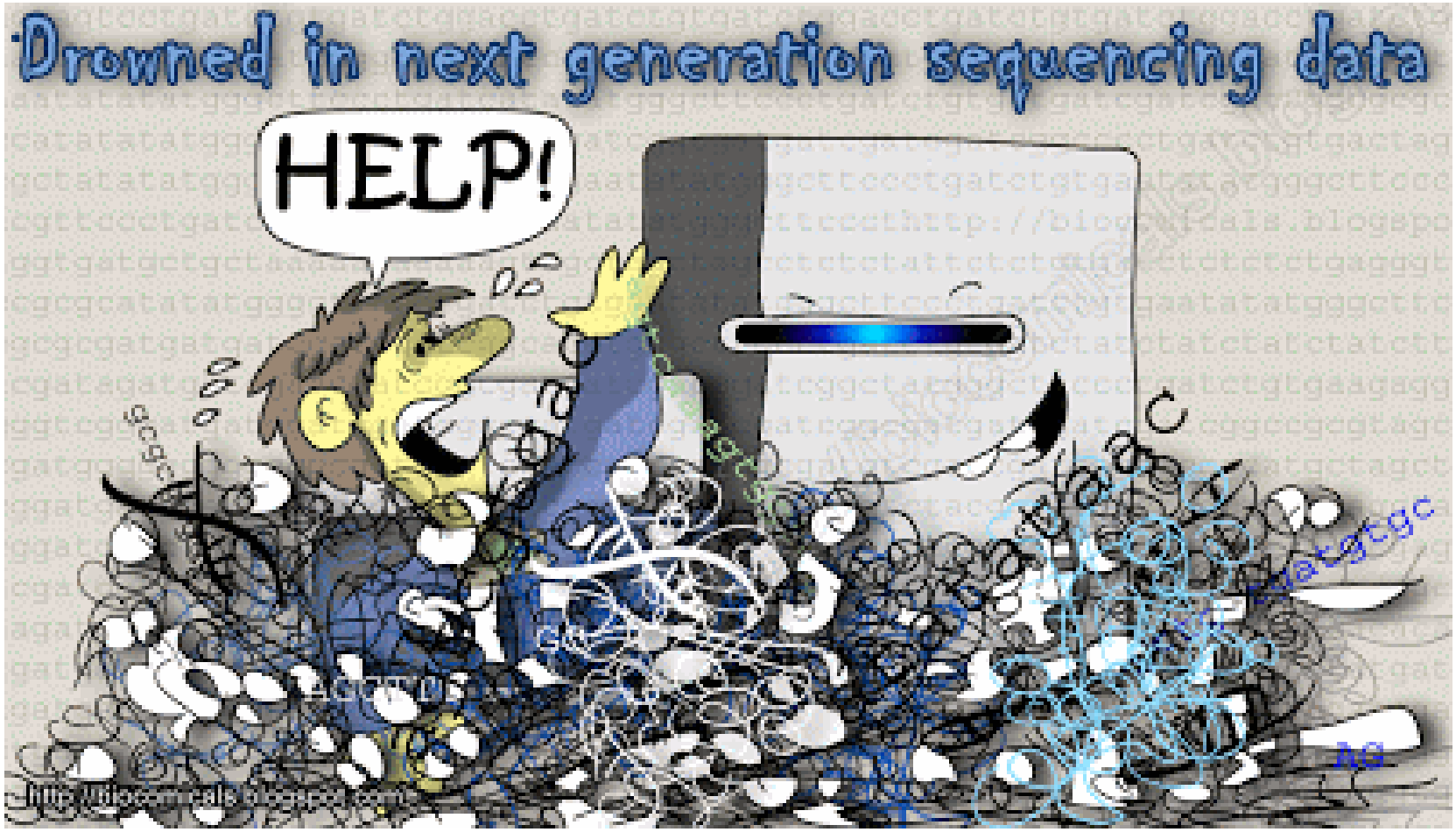
selection of the variants related to the disease pathogenesis

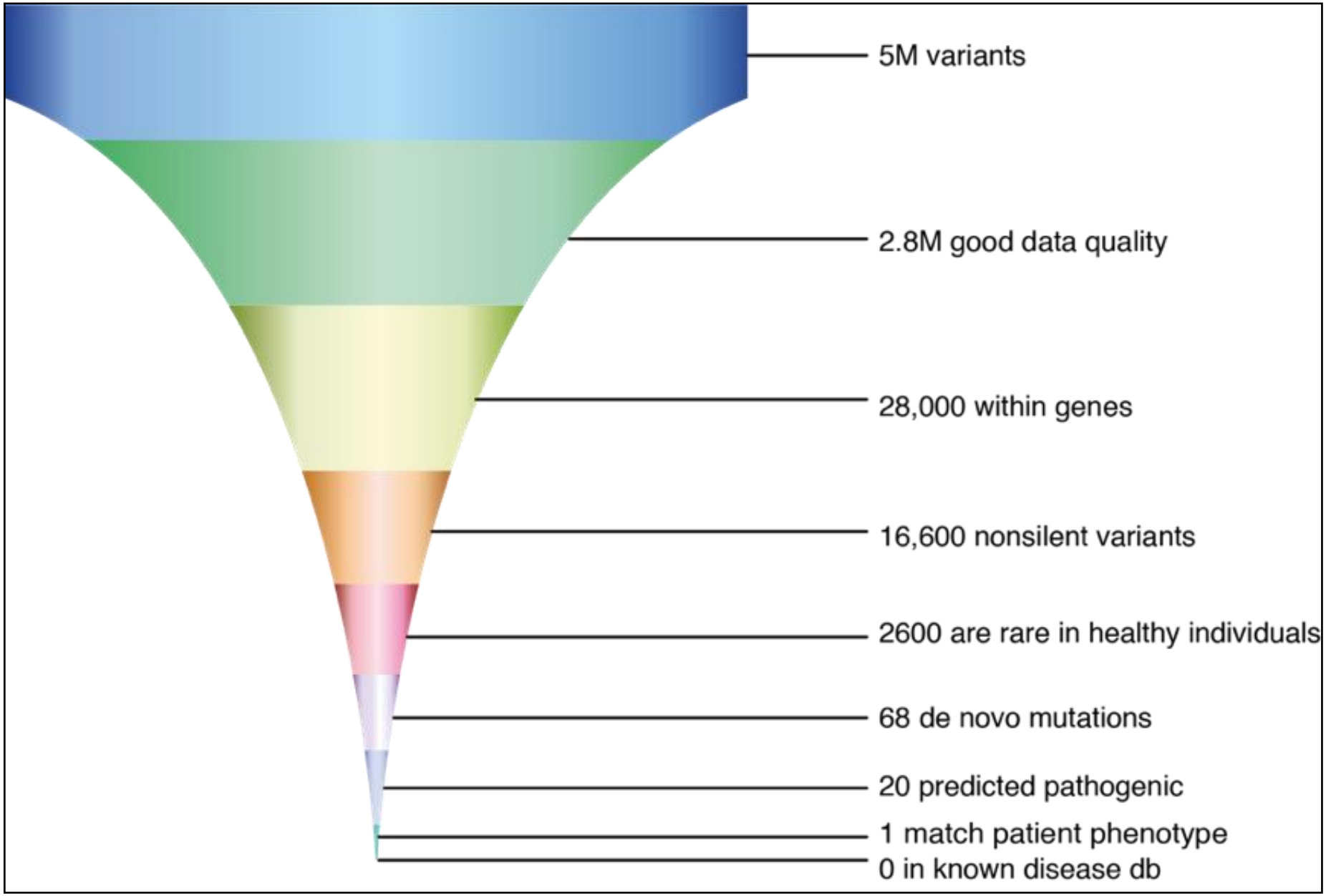
What we are going to sequence depend on the sequencer we have in our lab

	MiniSeq	MiSeq	NextSeq	HiSeq	NovaSeq
Run Time	4–24 hours	4–55 hours	12–30 hours	< 1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	16–36 hours (Dual S2 flow cells) 44 hours (Dual S2 flow cells)
Maximum Output	7.5 Gb (5Gb)	15 Gb (5Gb)	120 Gb (80Gb)	1500 Gb	6000 Gb
Maximum Reads Per Run	25 million	25 million*	400 million	5 billion	20 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp

Genome – exome – clinome – panel – comparison

Parameter	Genome	Exome	Clinome	Trageted NGS
cost	++	++	++	+
coverage	+	++	+++	++++
enrichment bias	-	+	+	+
Wet-lab time	1 library/ many diseases	1 library/ many diseases	1 library/ many diseases	1 library/ 1 or several iseases
Data amount	++++	+++	++	+
CNV	+ (structural)	+/-	+/-	+/-
New genes ?	+	+	-	-


Drowned in next generation sequencing data
HELP!


5M variants
2.8M good data quality
28,000 within genes
16,600 nonsilent variants
2600 are rare in healthy individuals
68 de novo mutations
20 predicted pathogenic
1 match patient phenotype
0 in known disease db

Databases used in genetic analyses

- Population databases: in-house, 1000Genomes, NHLBI, ExAC

  http://www.1000genomes.org/1000-genomes-browsers

  http://evs.gs.washington.edu/EVS/

  http://exac.broadinstitute.org/

  http://gnomad.broadinstitute.org/

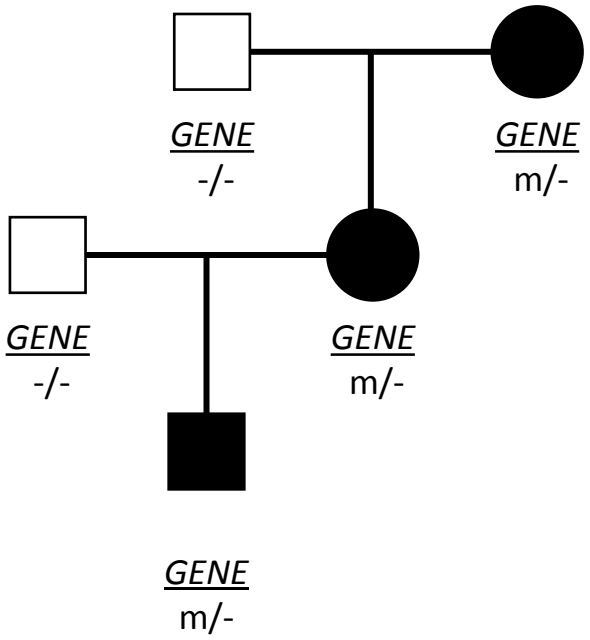
  The data set provided on this website spans 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current

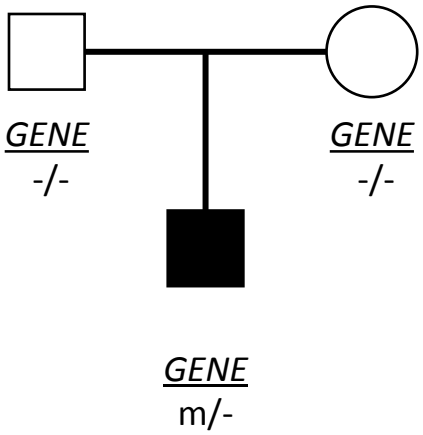
Databases used for NGS analyses

- Clinical databases
  - OMIM - http://www.omim.org/
  - HGMD - http://www.hgmd.cf.ac.uk/ac/index.php
  - LOVD - http://www.lovd.nl/3.0/home
  - ClinVar - http://www.ncbi.nlm.nih.gov/clinvar/
- Supporting algorithms
  - Phenomizer - http://compbio.charite.de/phenomizer/
- *In silico* analysis – predictive, but not as good as functional analysis or cosegregation analysis
- Google – GeneCards, Orphanet, GeneReview, Pubmed, UCSC

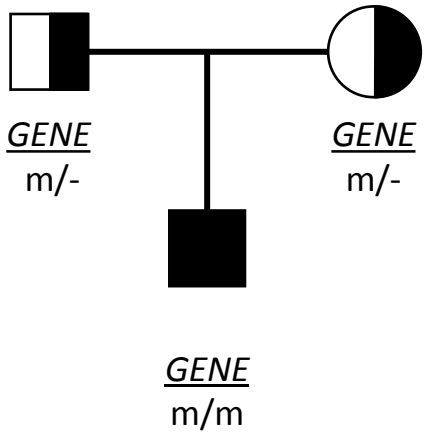
How to check if a variant is probably pathogenic?

Autosomal dominant inheritance





Autosomal recessive inheritance



Integrative Genomic Viewer (IGV)

One, two genes, but exome???

Variant annotation – Annovar / VEP

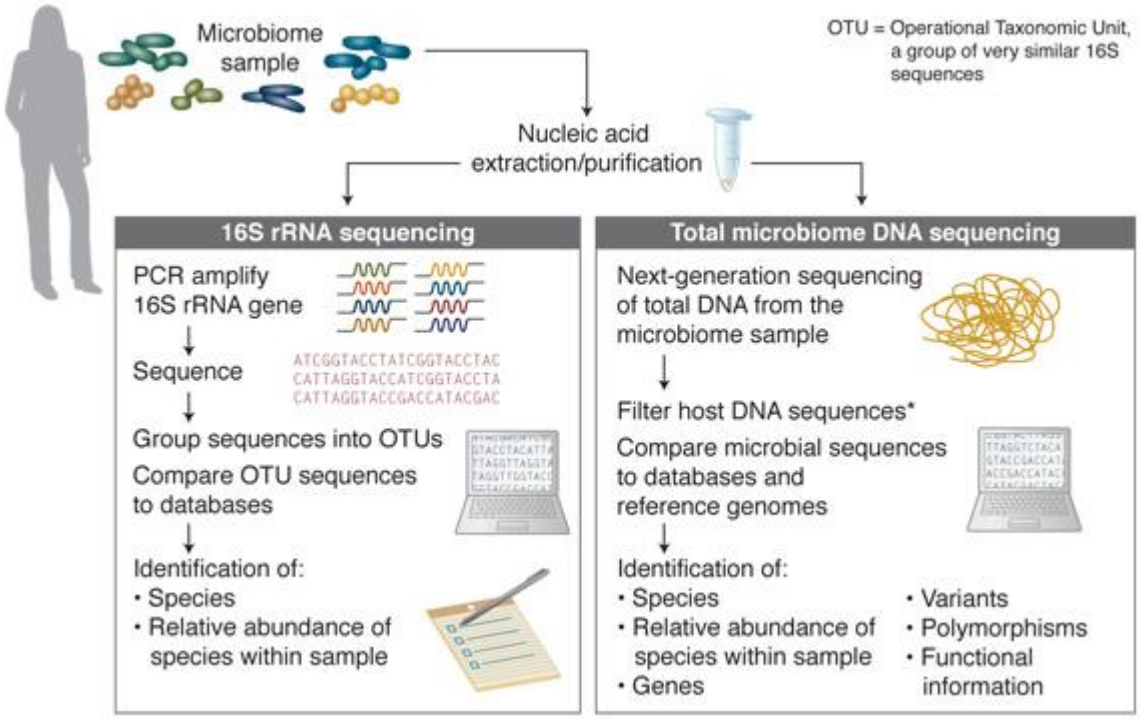
Reference genome: hg19 (panele) / hg38 (eksomy)
Population databases: 1000genomes, EVS/NHLBI, ExAC, gnomAD, in-house
Clinical databases: OMIM, ClinVar
Predictive algorithms: PolyPhen, SIFT, MutationTaster, FATHMM, RadialSVM, LRT, CADD, etc.
CNV analysis: Conifer
Cooperation: BIO-IT Tomasz Gambin

Not only SNV – CNV from panel testing

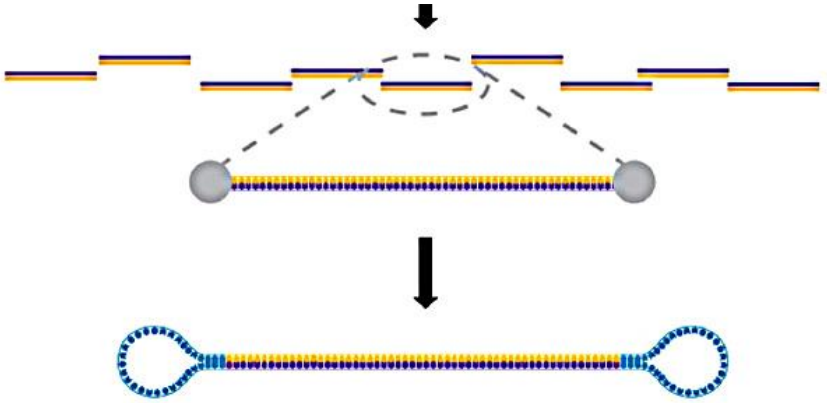
DNA sequencing to analyze the genetic disorders is not the only NGS use

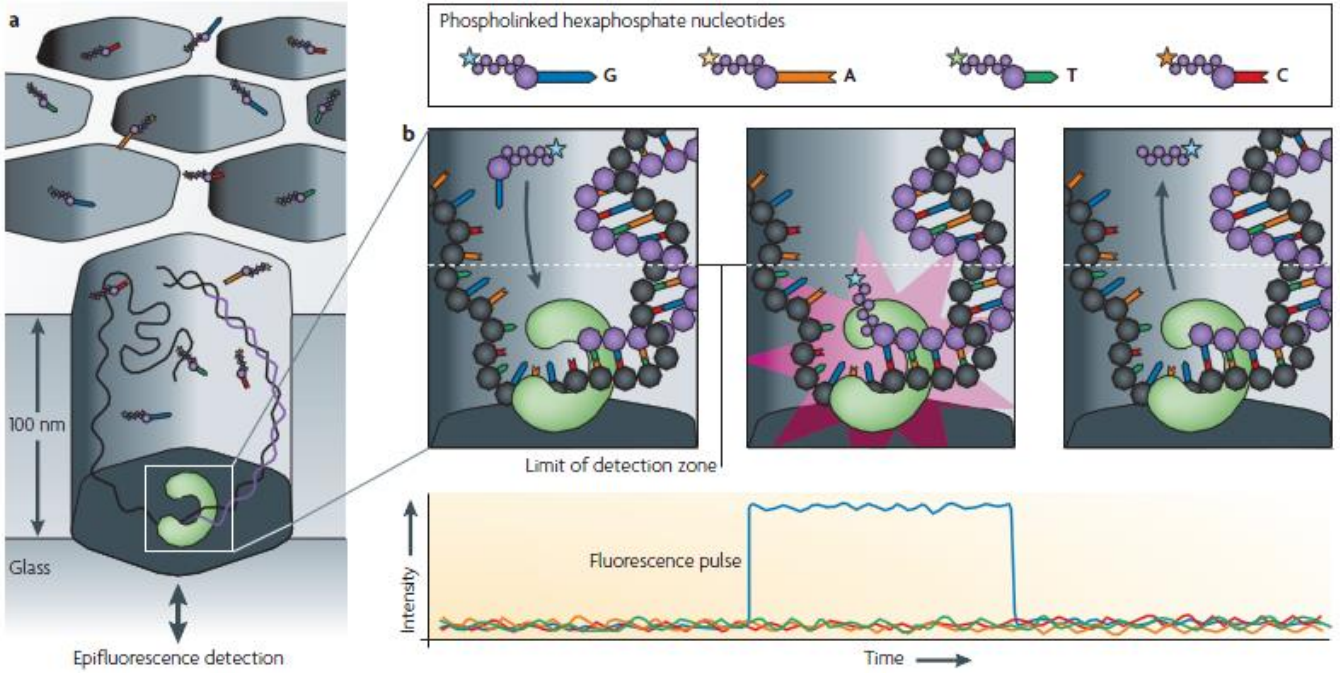
DNA sequencing in cancer (somatic variants)
Cell free DNA analysis (NIPT, cancer)

transcriptome sequencing (gene expression analysis) – RNA-Seq
epigenome sequencing (DNA methylation analysis) – Epi-Seq
analysis of DNA-protein interaction – ChiP-seq
microbiome sequencing



PacBio





https://www.youtube.com/watch?v=WMZmG00uhwU

Oxford Nanopore

https://www.youtube.com/watch?v=E9-Rm5AoZGw

	PacBio[1]		Oxford Nanopore[2]	
Instrument Specifications	RS II (P6-C4)	Sequel	MinION	PromethION
Average read length	10 – 15 kb	10 – 15 kb	Variable (up to 900 kb)[3,4]	*
Error rate	10 – 15 %	10 – 15 %	5 – 15 %[4,5]	*
Output	500 Mb – 1 Gb	5 Gb – 10 Gb	~5 Gb[4]	*
# of reads	~50k	~500k	Variable (up to 1M)[6,7]	*
Instrument price/Access fee[a]	$700k	$350k	$1000[8]	$135k bundle[9]
Run price	~$400	~$850	$500-$900[7]	*

https://blog.genohub.com/2017/06/16/pacbio-vs-oxford-nanopore-sequencing/

Department of Medical Genetics
Institute of Mother and Child
Head: prof. dr hab. med. Jerzy Bal
RASopathies / floppy child syndrome: Monika Gos
Genodermatoses: Katarzyna Wertheim-Tysarowska, Dominika Śniegórska, Sylwia Radomska
Epileptic encephalopathies: Dorota Hoffman-Zacharska, Paulina Górka-Skoczylas, Karolina Kanabus
Hearing loss: Katarzyna Niepokój
Intellectual disability: Agnieszka Charzewska, Sylwia Rzońca
Microcephaly: Paweł Gawliński, Mateusz Dawidziuk
Chronic panceatitis: Agnieszka Rygiel, Aleksandra Kujko
DiGeorge syndrome: Beata Nowakowska

Bioinformatic team: Tomasz Gambin, Justyna Sawicka and others from PW

Studies supported by National Science Centre
and IMID intramural grants