How to start with genomics

Agnieszka Szmurło

WHOAMI

Integration / Solution / Enterprise Architect @ Insurance Companies

PhD student @ Politechnika Warszawska

http://biodatageeks.org/

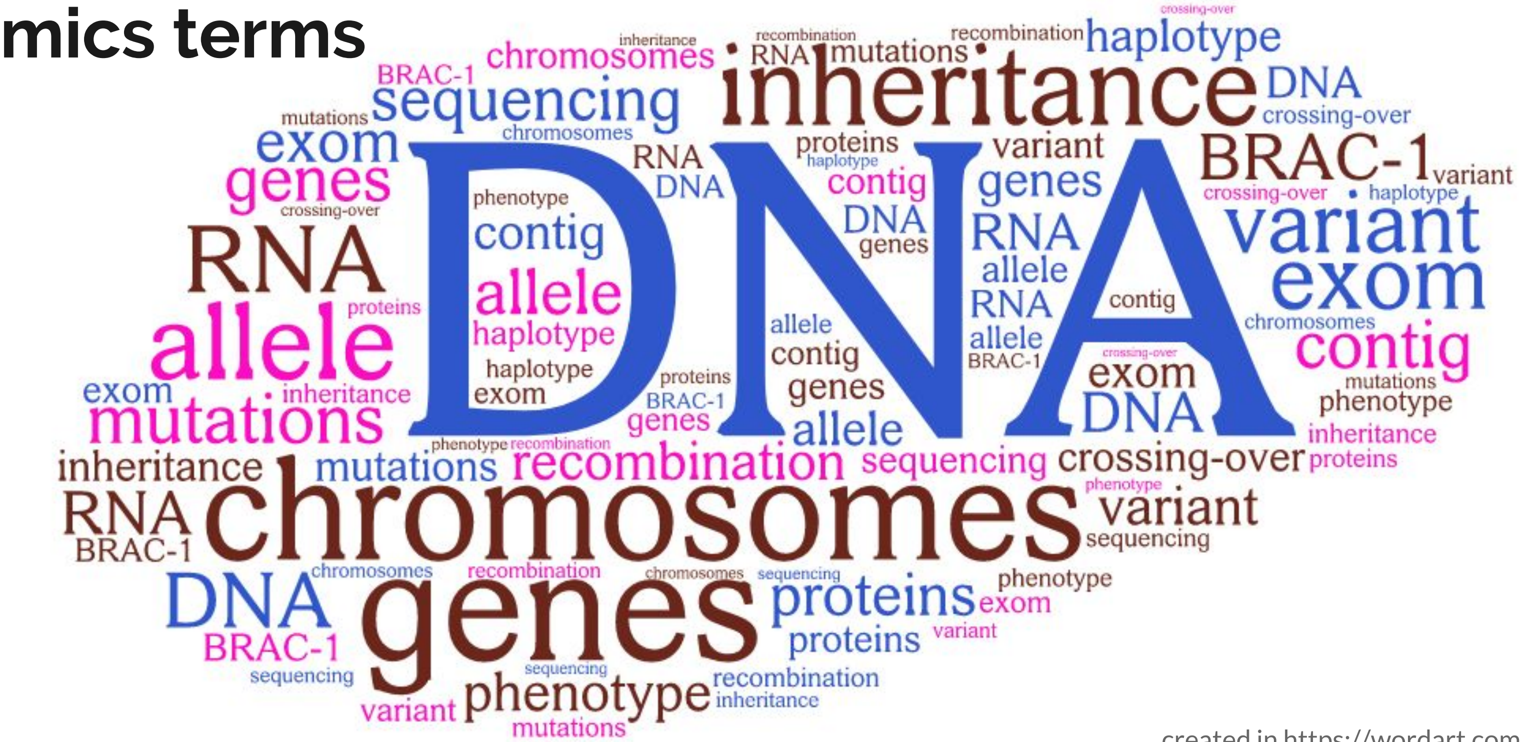
How to start with genomics?

Before that...

Why bother?

And actually what is genomics?

Genomics terms



created in https://wordart.com

So what is... ?

Genomics?

Field of science focusing on the structure, function, evolution, mapping, and editing of genomes

A genome?

An organism's complete set of DNA.

DNA?

Chemical acid structured like twisted ladder. The steps of the ladder are complementary bases A-T, G-C

source: wikipedia

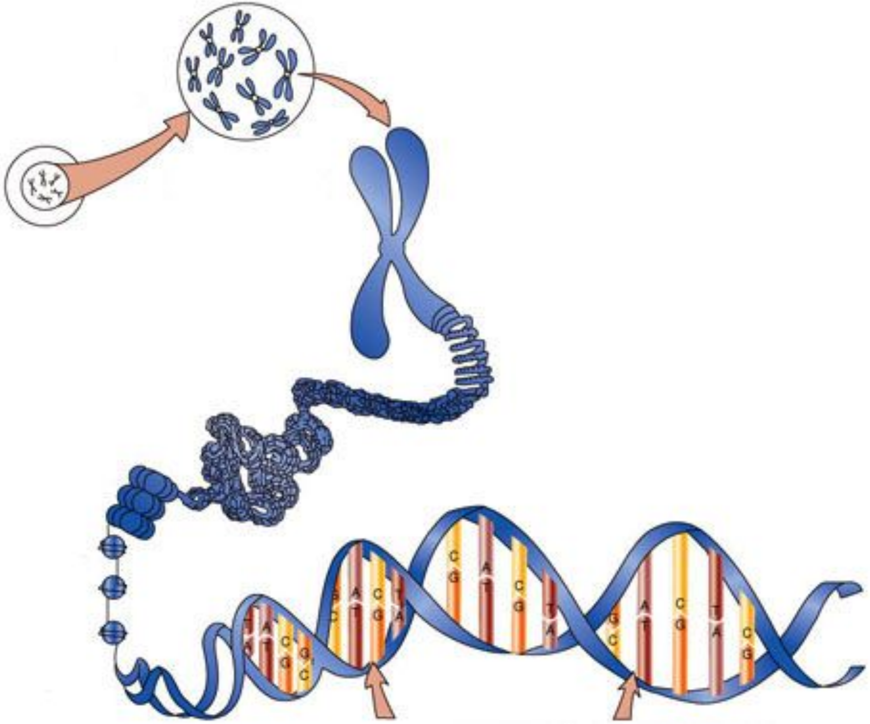
Genome. Some facts

Genome appears in each cell

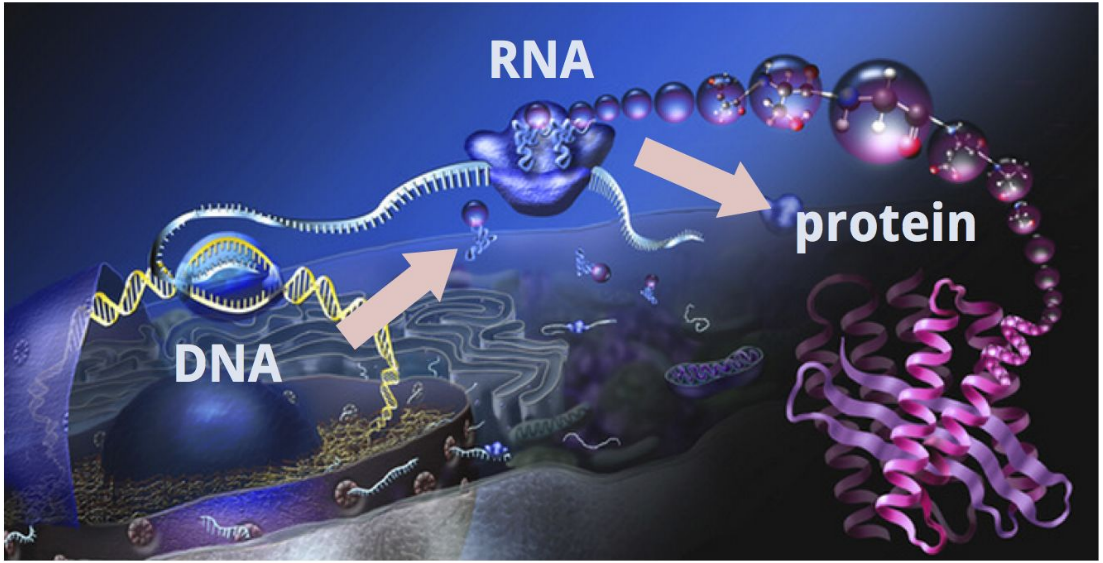
Genome is packaged in chromosomes

Genome is **VERY BIG** and very small at the same time (human cell’s DNA length is 3 m, while cell’s nucleus is 6 microns)

Genes are “fragments” of DNA



1958. Central Dogma



DNA gets transcribed into RNA.

RNA gets translated into proteins.

Proteins build body parts

http://thisisartlab.com/tag/central-dogma/

Representing genome

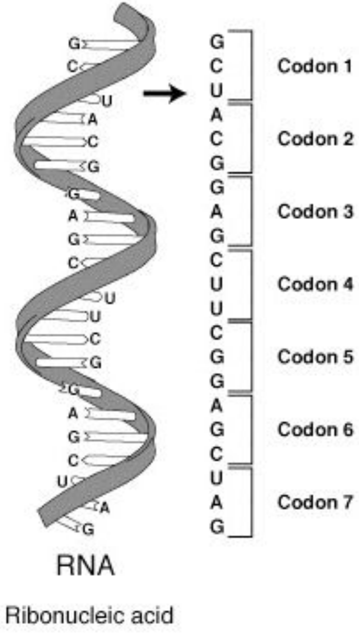
Genome => string (3 billion letters) build upon 4 letter alphabet {A, C , G, T}

Genes => substrings of genome

RNA => substring of genome, transformed , T->U

Genome is like “source-code” for humans?

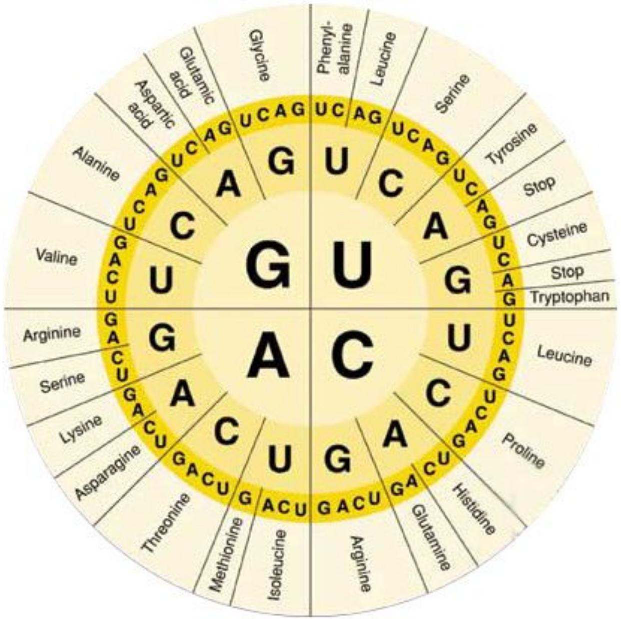
Genetic code



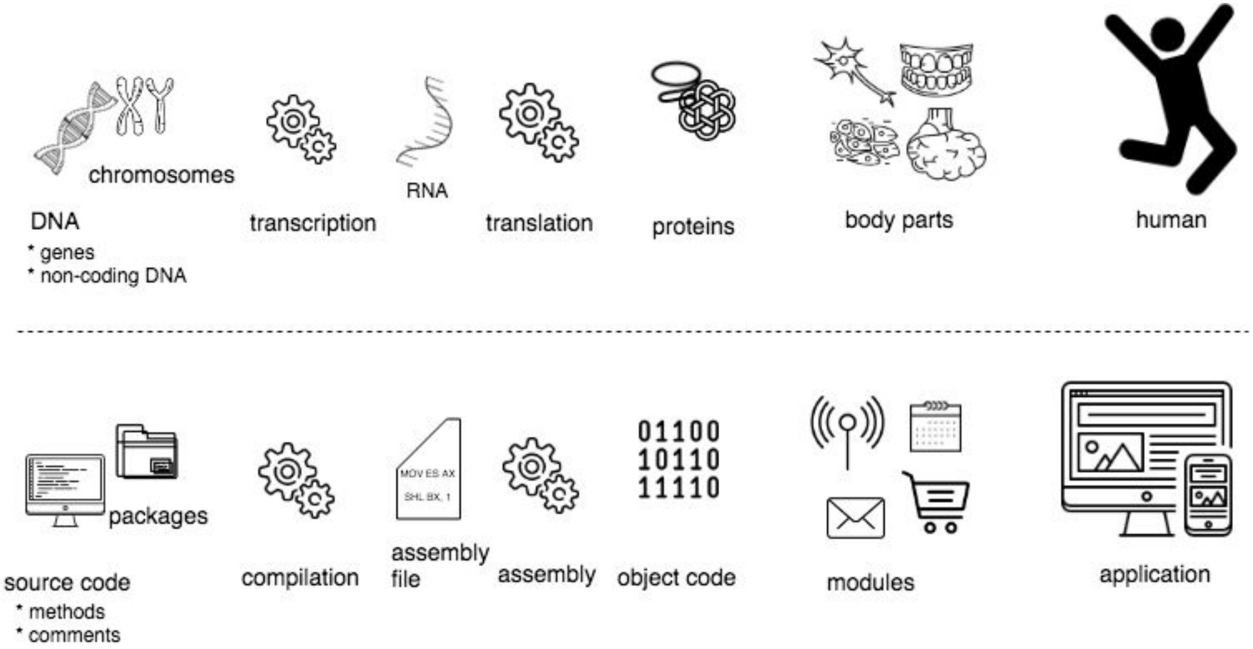
RNA gets translated into chain of amino acids using genetic code.

RNA is divided into 3 letter long "codons" which identify particular amino acid.

There is also codon for START and STOP.

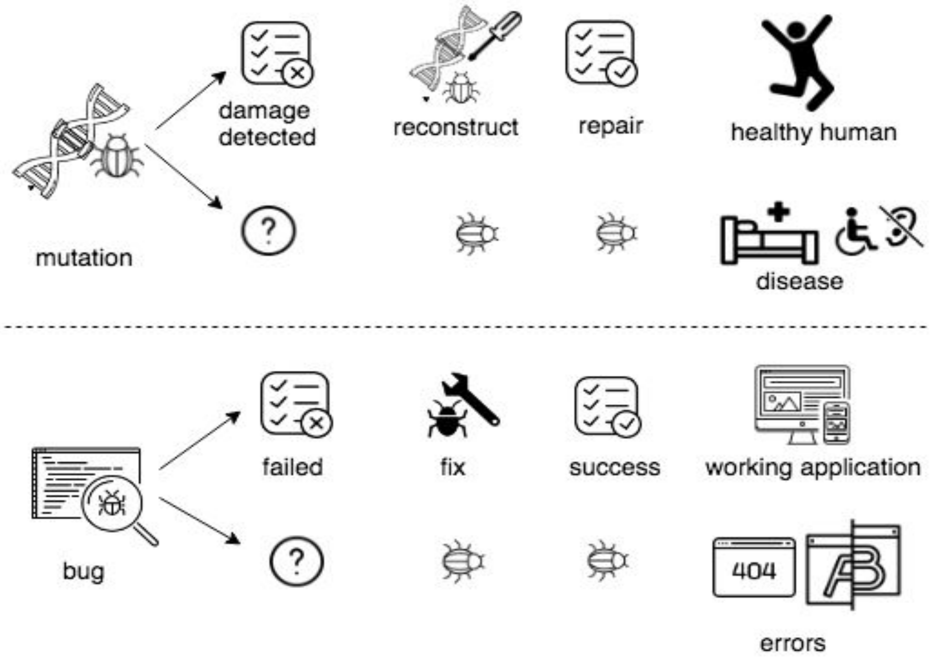


Programming human? Happy path



icons from: www.flaticon.com

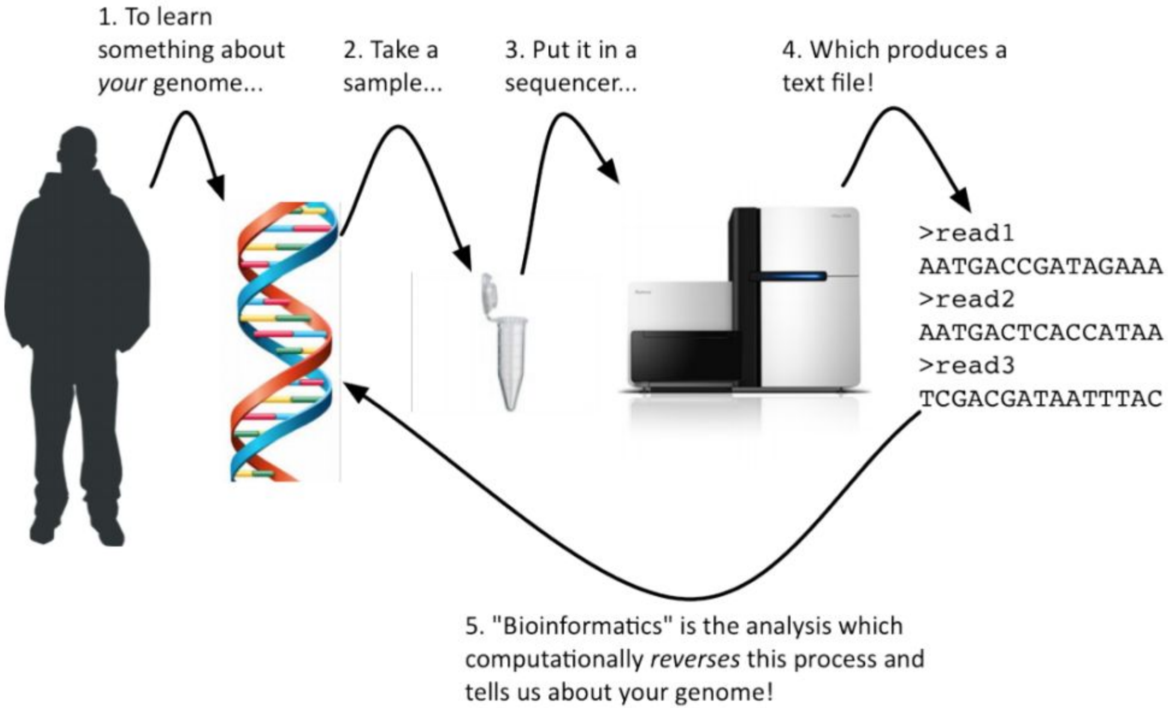
Programming human? Quality checks



how to get to know your genome?

icons from: www.flaticon.com

A One-Slide Introduction to Genomics



1. To learn something about *your* genome...

2. Take a sample...

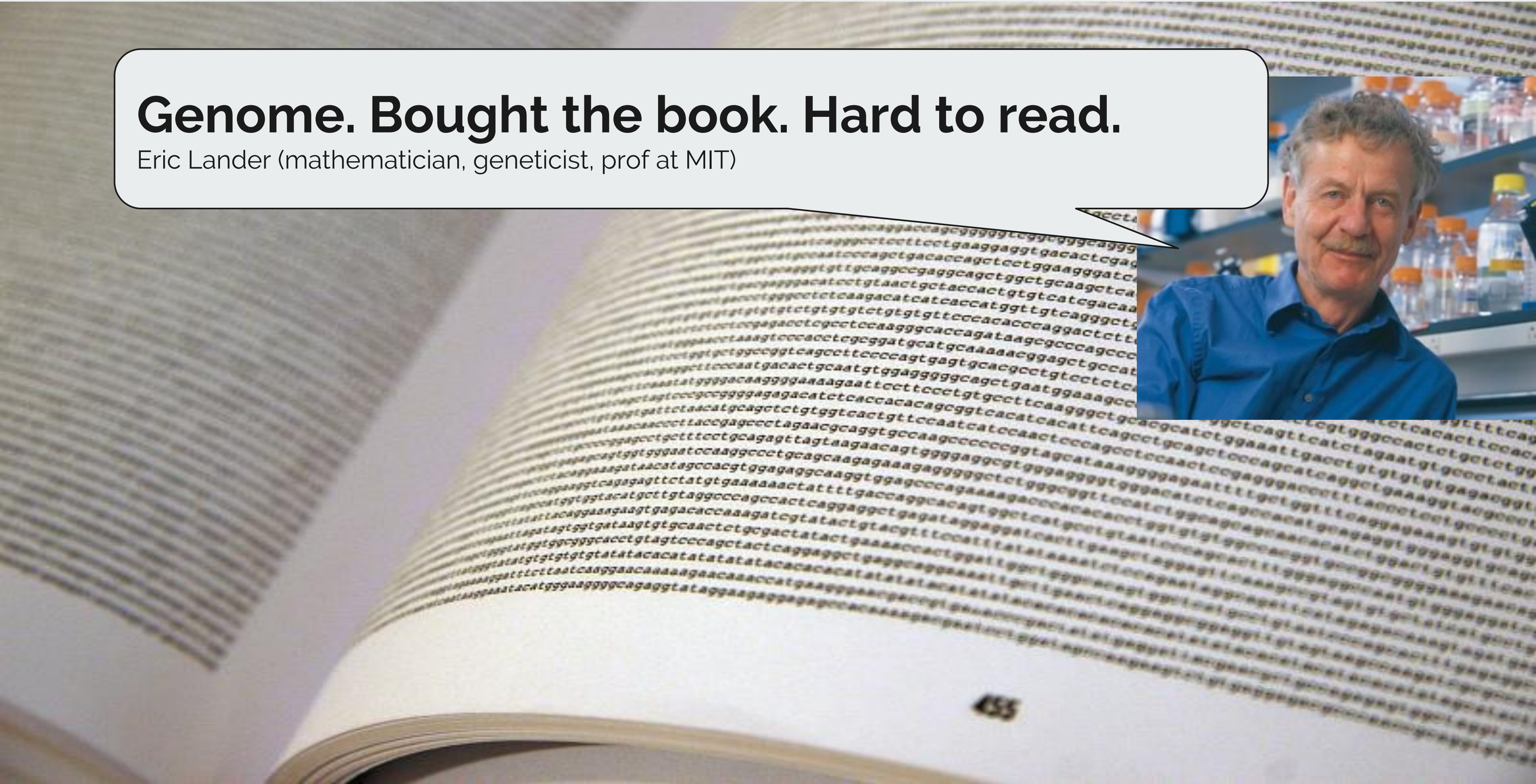
3. Put it in a sequencer...

4. Which produces a text file!

```
>read1
AATGACCGATAGAAA
>read2
AATGACTCACCATAA
>read3
TCGACGATAATTTAC
```

5. "Bioinformatics" is the analysis which computationally *reverses* this process and tells us about your genome!

Figure: Źródło:http://www.slideshare.net/TimothyDanford/tdanford-spark



Genome. Bought the book. Hard to read.

Eric Lander (mathematician, geneticist, prof at MIT)

Hard to read? Why?

- very long . 3 billion letters. When stored in file: 150 GB
- fragmented (genes are split by exons)
- many repetitive fragments
- mutation or sequencing error?
- meaningless or meaningful? (rubbish DNA or regulatory sequences)

Main problems targeted by genomics

- DIAGNOSTICS
  - finding root cause of genetic disease
  - cancer (and other diseases) prevention
- THERAPY
  - deciding on the the best treatment for patient
  - developing gene therapy - repairing DNA fragments
  - predicting organism’s response to drug
- RESEARCH
  - discovering knowledge (genotype-phenotype relationship, exposom-phenotype relationship)

PERSONALIZED MEDICINE

IT challenges in genomics

- classic algorithms optimizations for data analysis (scaling and distributed computing)
- data science - discovering knowledge - genotype-phenotype relationship, gene-gene relationship, gene annotations

But first, researchers need a lot of genomes. The bigger the database, the better.

- embracing genomics Big Data

How to share safely your sensitive data?

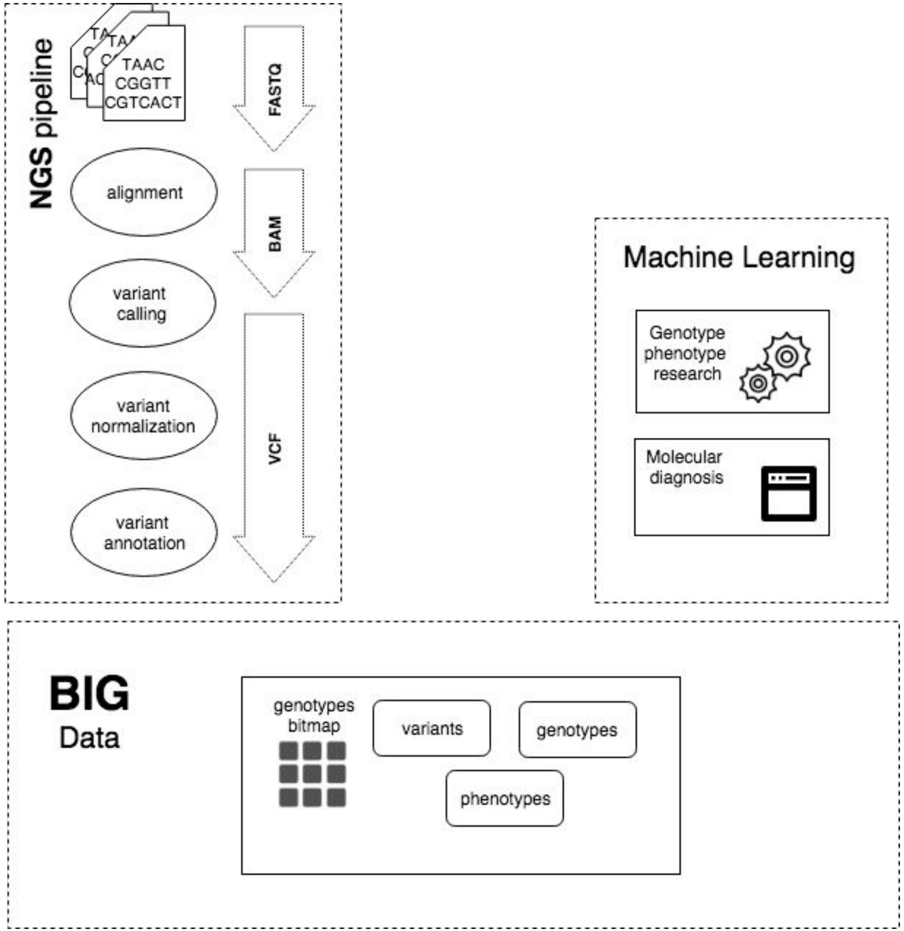
- genomics and blockchain

iGAP

NGS pipeline - data transformations, normalization and preprocessing

Data model - fit for random access and analytical access patterns

Machine Learning - tools to support molecular diagnosis and general research



Scalable range joins

In bioinformatics we frequently perform joins as keys between datasets. Our datasets are billions of rows. They have to be efficient.

Adding custom strategy into SparkSQL to efficiently join interval ranges, based on Interval Tree algorithms

CNV detection

CNVs may have severe impact on human health

Current solutions provide results with many False Positives and that complicates analysis.

We are working on classifier detecting rare CNVs (Bayes, neural networks).

Some latest genomics highlights:

Editing your genes at home using CRISPR technology -> BioHackers, [2017]
http://mysteriousuniverse.org/2017/11/biohackers-are-using-crispr-to-hack-their-own-dna/

Luxterna, Gene Therapy for blindness approved by FDA: [2017]
https://edition.cnn.com/2017/12/20/health/fda-gene-therapy-blindness-bn/index.html

Google Deep Variant [2017]

https://www.face2gene.com/ [2011 +]

How to start?

Quick biology recap (PL/EN)

KHANACADEMY https://pl.khanacademy.org/science/biology

YouTube https://www.youtube.com/playlist?list=PLInNVsmlBUIQT_peuWctrmGMiLngK-6fb

Comprehensive course:

coursera https://www.coursera.org/specializations/genomic-data-science

How to start?

Some recommended readings:

“Algorithms For Next Generation Sequencing”, 2017, Wing-Kin Sung

“Genetyka Medyczna i Molekularna”, 2017, Jerzy Bal

Events:

NGSchool - Summer School this year in Lublin http://ngschool.eu/

BioHack - Hackathon in Lodz, http://www.biohack.linuxpl.eu/

How to start?

Reach out:

Biodatageeks weekly meetings: www.biodatageeks.org

RNA CLUB: https://www.facebook.com/RNAClubWarsaw/

MIMUW http://bioputer.mimuw.edu.pl/

MINI, PW: http://mi2.mini.pw.edu.pl/

PTBI : https://www.ptbi.org.pl/website/home/

Thank you!