Application 1: Bioinformatics & Genomics

Applications of Machine Learning in Biology & Medicine

Raunak Shrestha, PhD

Post-doctoral Research Fellow, Bioinformatics, University of British Columbia

Laboratory for Advanced Genome Analysis, Vancouver Prostate Centre, Canada

Laboratory for Bioinformatics and Computational Genomics, Indiana University, USA

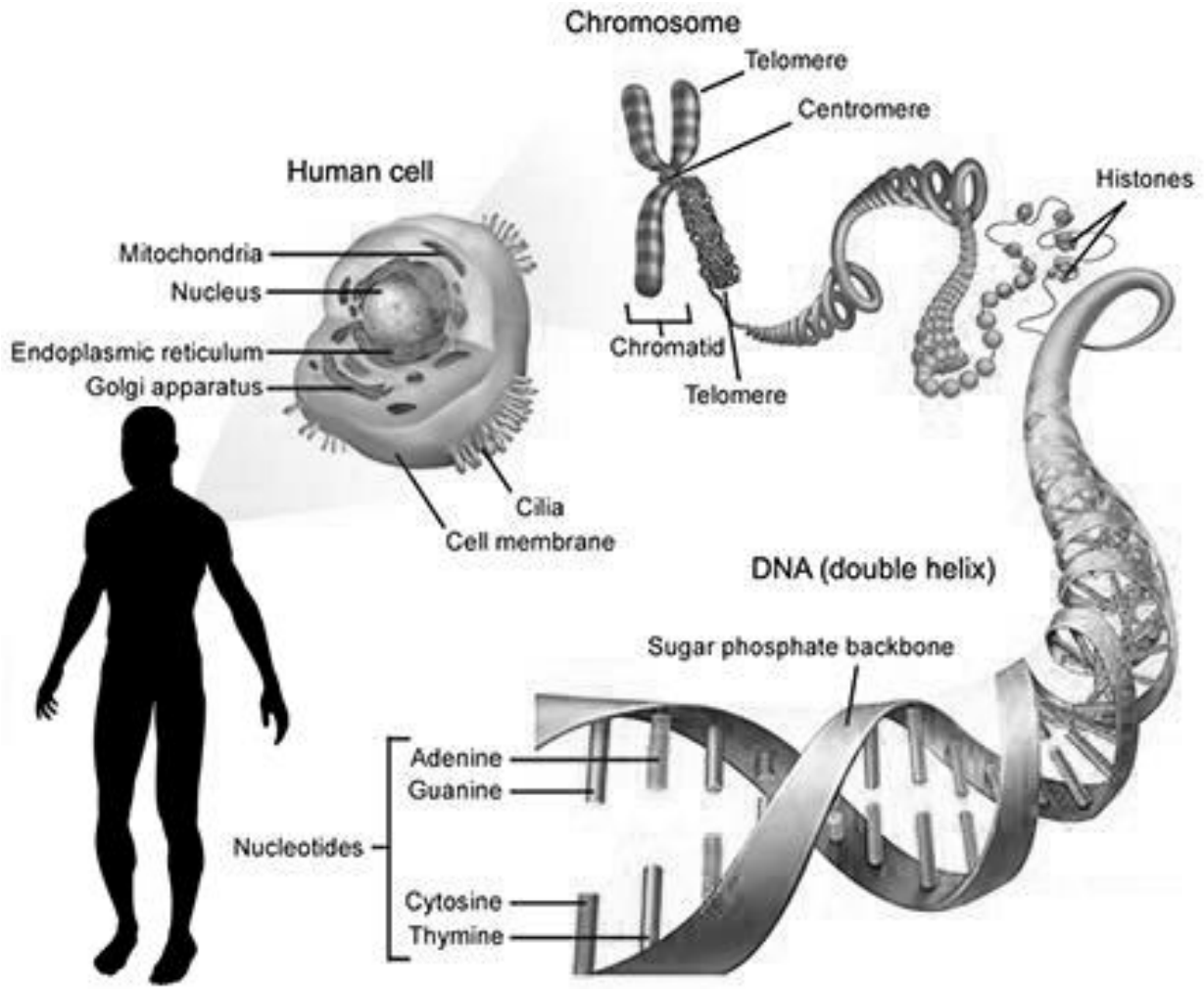
http://raunakms.github.io

@raunakms

PART - I

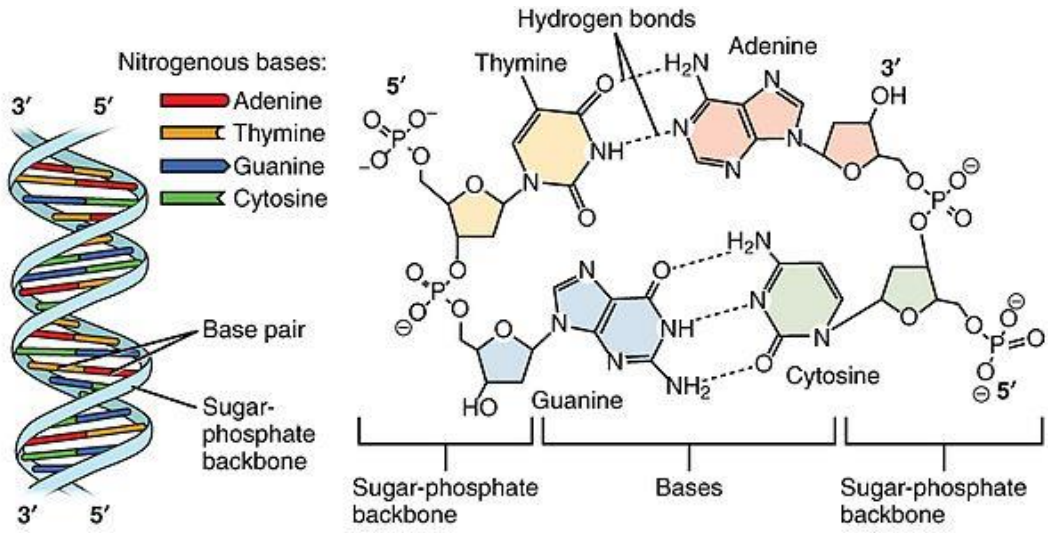
Introduction to Molecular Biology/Genomics & Bioinformatics

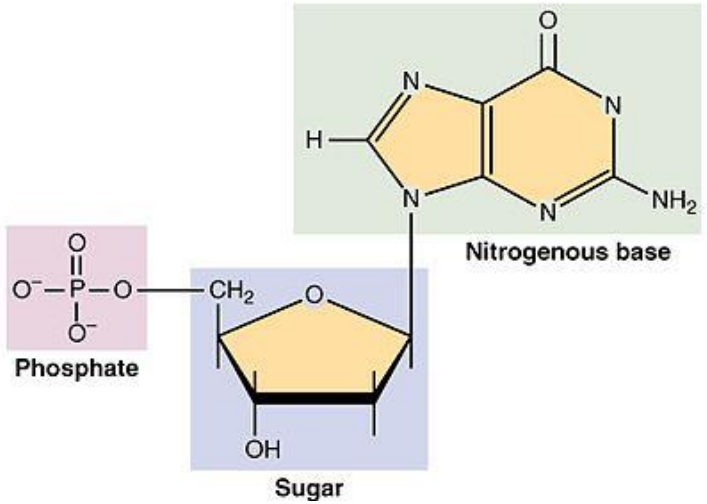
DNA: Deoxyribonucleic acid



- Carrier of genetic information
- Transmits genetic information from parents to the offspring
- Often termed as “The Molecule of Life” or the “The Software of Life”

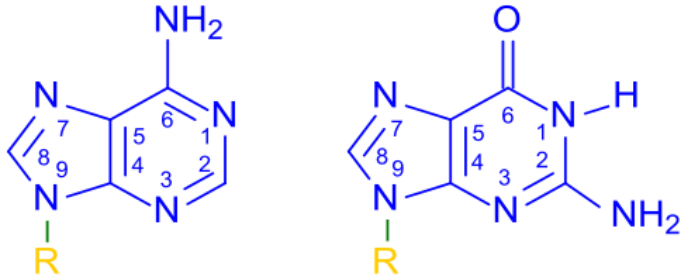
DNA: Deoxyribonucleic acid





Nitrogenous base (Nucleotides)

Purines



Adenine Guanine

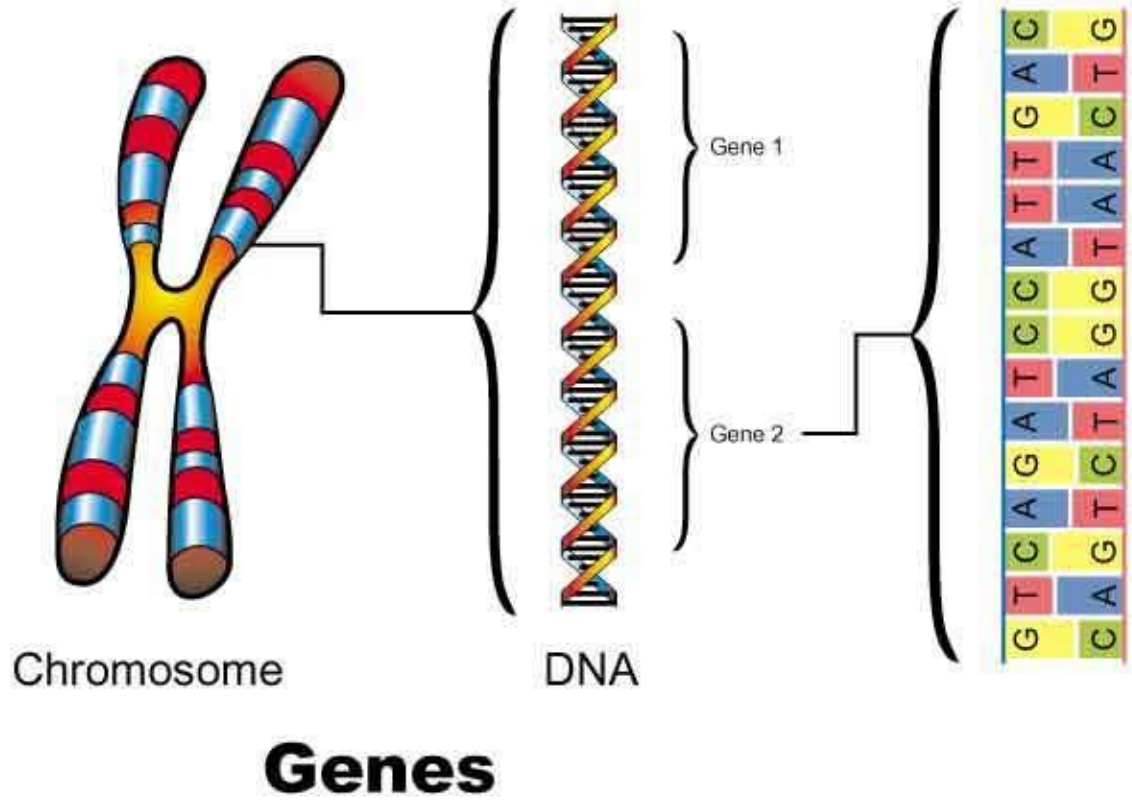
Pyrimidines

Cytosine Thymine

Size of Genome (Total Genetic Material)

Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	 Virus	 Bacteria	 Fruit fly	 Human	 Canopy Plant

Gene: The Functional Unit of the Genome



Each **Gene codes for a particular Protein**
that carries out a particular task/function

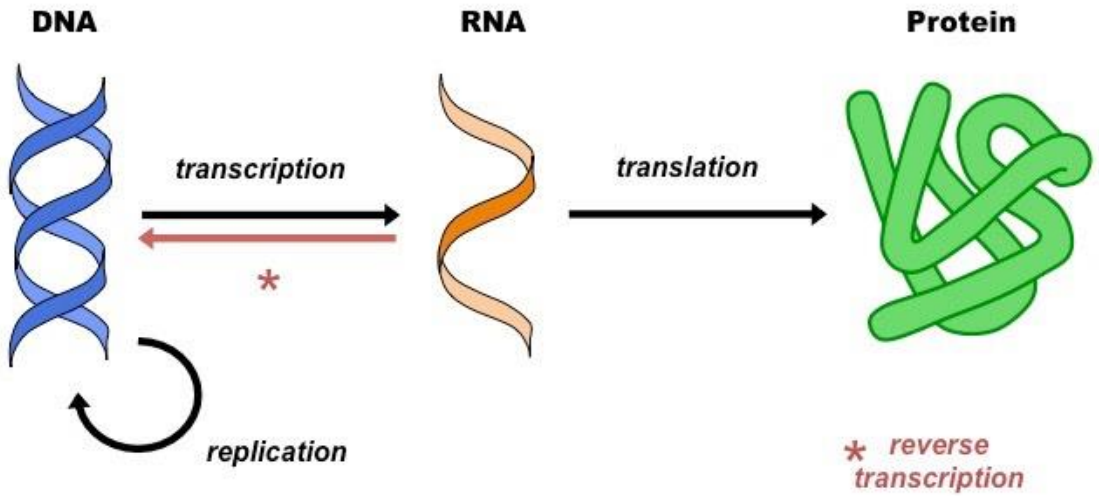
Nucleotide sequence of **BR**east **C**ancer **A**ssociated (**BRCA1**) Gene

```
   1 cagcaaggct cacctcctct cagcacctca gtctcttctt ctataaagtg aagcggtgat
  61 ttgctctcta aggtcccttc tcattgtgtg agaacaagag tcccagtggc ctgggagcac
 121 aggggagggg aggcggctgc ccacccttca acactattac ccatcggtca cttgtgctct
 181 gtctcctctc tcccttggta aatggatctg ctgcgccagc aaacaactgt ccattgtgtc
 241 ctgatgaggg acgagcaggg aaaaccaatc acttatgatg caaatgagcc gacggctgca
 301 gcctcctcac attcactgag tgcttctctt ctgcccagag cccggcgccg ccccagttgg
 361 gcctttctcc ctagccccag ttcccatctg gtcctggagg aggagttggg aggcccagac
 421 tggcctcccc attgtcgttc atttcactcc ctgcaaggag tttcaaccct caaatcctca
 481 gagcctggcc cagctcctct ccccacccca aaggtggctc agttttattt tttttaaaaa
 541 tcgattgaaa ttcacataac atacaattag ccattttaaa atgtacaatt caggggcatt
 601 tagtgcattc acaatgttgt gcaaccacca cctctggact caggttttat gccctctgcc
 661 cattctttct ttttttttt ttttttgaga cggagttttg ctctgtcgcc caggctggag
 721 tgcagtggca ggatcttggc tcactgcaag ctccgcctcc cgggttcatg ccattctcct
 781 gcctcagcct accaagtagc tgggactaca ggcacctgcc accacgcctg gctaattttt
 841 tgtattttta gtagagatgg ggtttcaccg tgttagccag gatggtctcg atctcctgac
 901 ctcgtgatct gcccgcctcg gcctcccaaa gtgctgggat tacaggcttg agccaccacg
 961 cccggcctgc cctctgccca ttctatgccc cacagccttc agtgggactg ggaagctgct
1021 gaggcctcag cagagctgag gactaatgga ggctgatgaa taggctgaga aaagcccaga
1081 gctggcctga ggtgaagagg tgattcccac actttaggat gatgttagaa ggtttgggct
1141 ttggtgggag ggggacatca ctgccttagc cagaacggac ctgtggccac gttgtagaac
1201 ttttttttt ttttttgaga tggagtttg ctcttgttgc ctgggctgga gtgcaatggc
1261 atgatctcag ctcgctgcaa cctctggctc cagggttcaa gtgattctct ggcctcagcc
1321 tcctgagtag ctgggattac aggtgcacac caccacgccc agctaatttt ctatttttaa
1381 tagagacggg gtttcaccat gttgaccagg ctggtcttga acttctgatc ttaggtgatc
1441 ggccttccaa aagtgctggg attacaggca tgagccaccg tgcctggccc cagaacattc
1501 ttgcattctt gaatttccat gggaattttt ttttccattt taatctctca ggcattagaa
1561 accagaagtg gctcctgtac agatcttggt ttccatagct ttcagagctt ctgtgggctt
1621 ggggatgatc ccgtcggcag gcttctcact ttagcagctg cagaaggccc tgagacagcc
1681 ccacagtcct ttgcctccag gaaagacctc ccgccaccag gctctaggaa caatcctaga
```

- DNA/Gene is a long stretch of nucleotide sequence
- You can think of this as a long string of text
- BRCA1 gene necessary contains information (or instructions) to code for BRCA1 protein

To view entire sequence of BRCA1 gene: http://asia.ensembl.org/Homo_sapiens/Gene/Sequence?g=ENSG00000012048;r=17:43044295-43170245

The central dogma of molecular biology

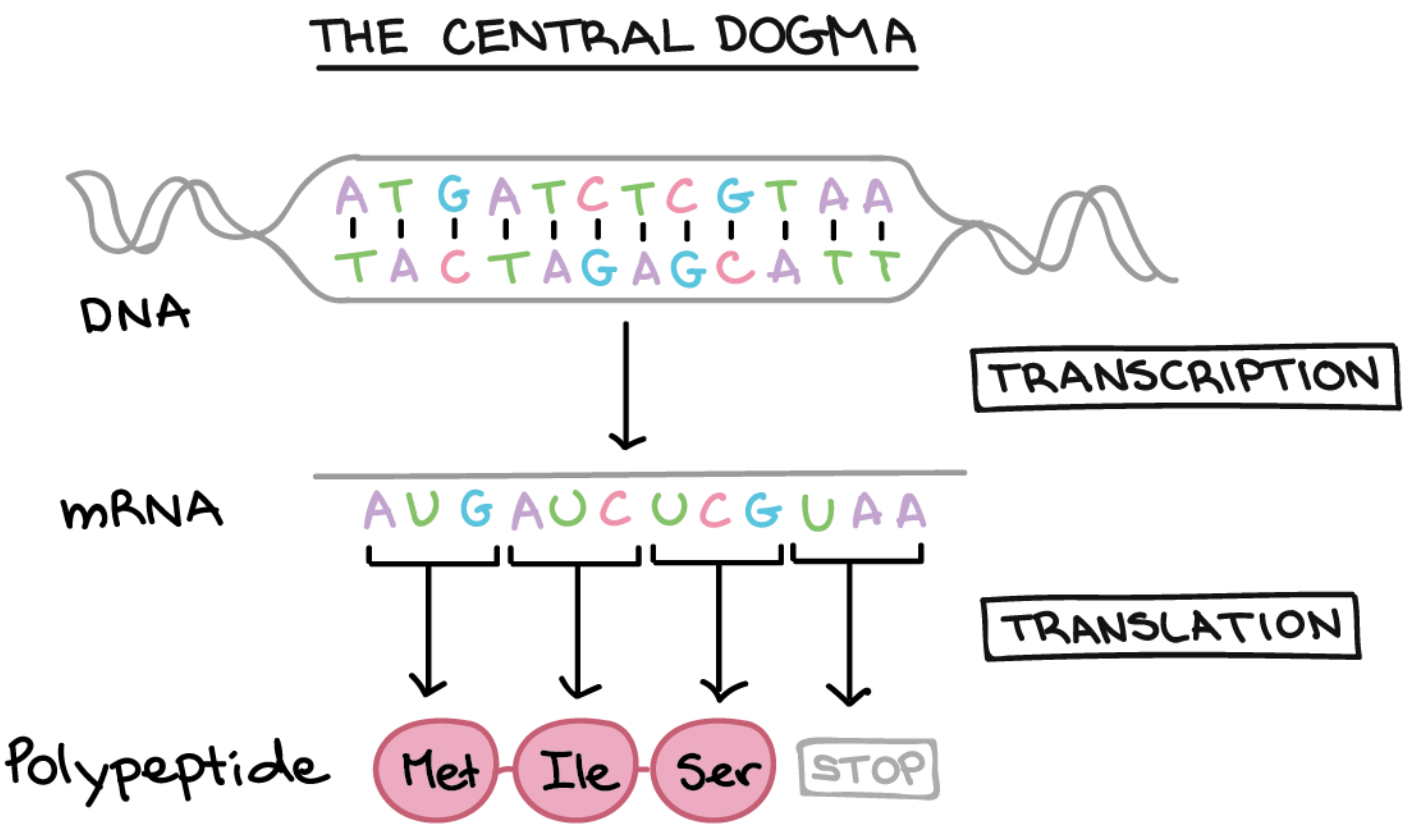


- explains the flow of genetic information within a cell
  - **DNA codes for RNA** via the process of transcription.
  - **RNA codes for protein** via the process of translation.
- This information flow was always considered to be uni-directional:

  DNA → RNA → Protein

http://ib.bioninja.com.au/_Media/central-dogma_med.jpeg

The central dogma of molecular biology



Information

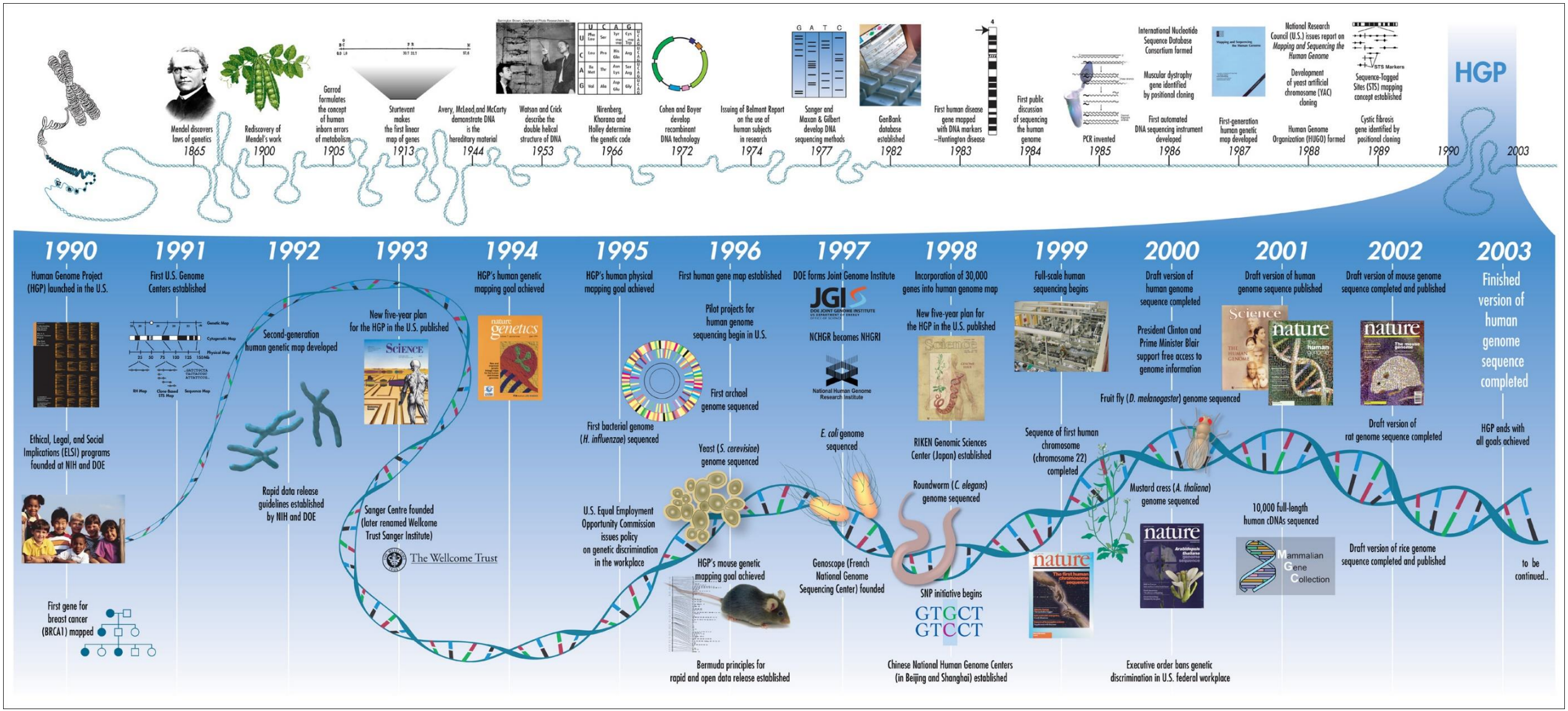
↓

Message

↓

Function

Human Genome Project



Collins et al. 24 April 2003. Nature

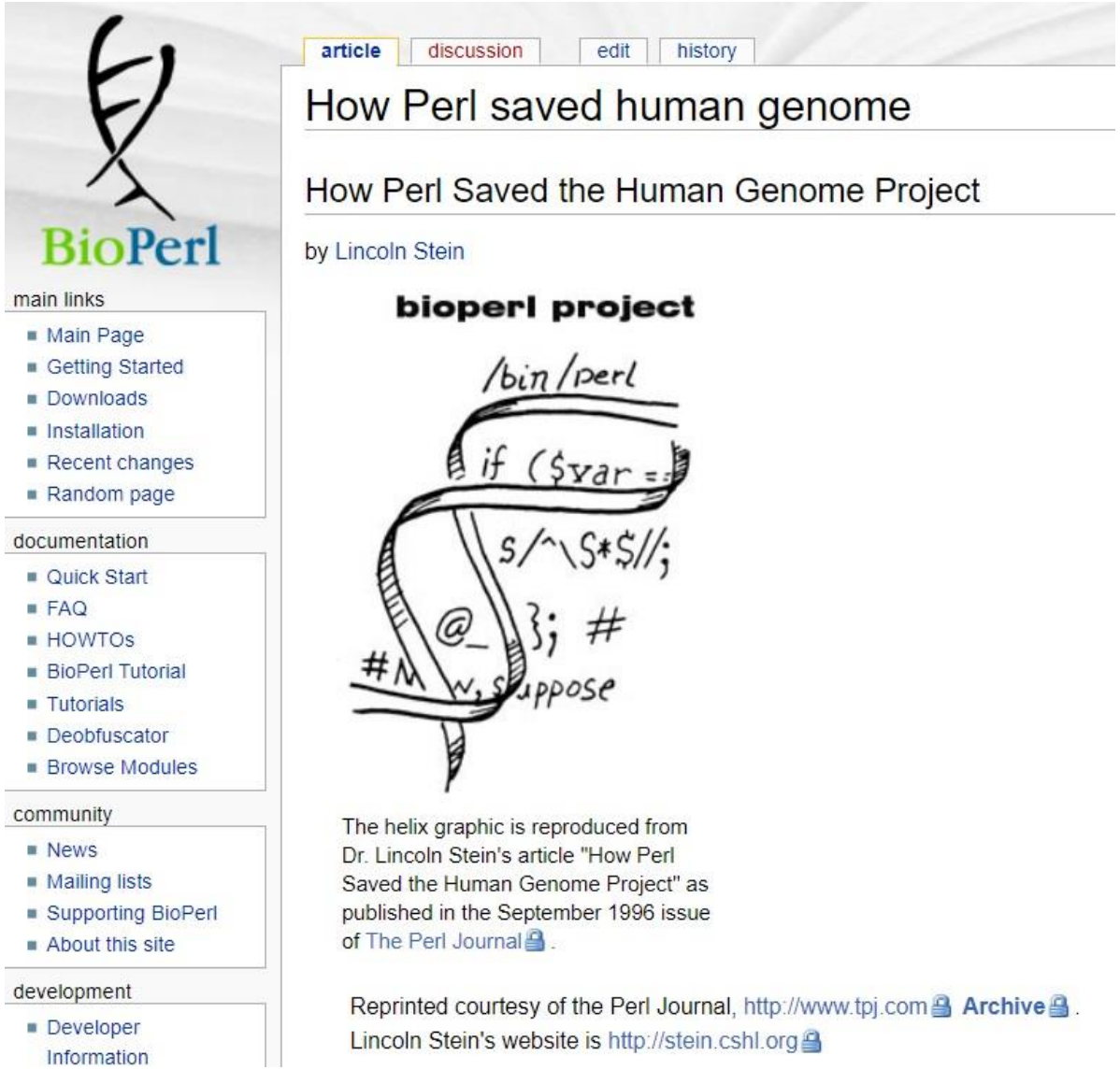
The announcement of the completion of the first draft of the human genome on 26 June 2000

Documentary on The Discovery of the DNA structure

https://youtu.be/yf4_D7HF5kU

Documentary on The Human Genome Project

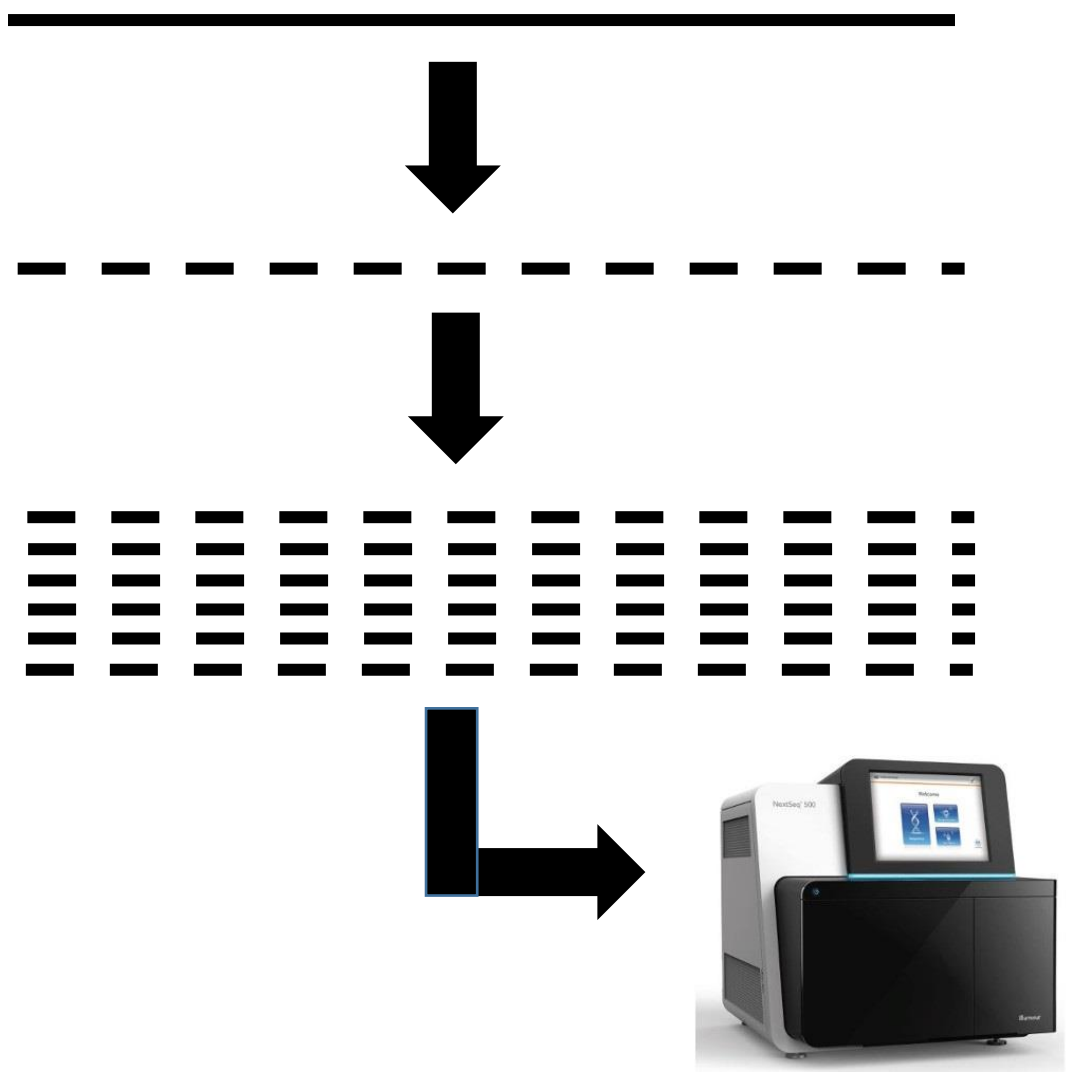
https://youtu.be/GyBued188Lk



Recommended Reading

https://web.archive.org/web/20070202101624/http://www.bioperl.org/wiki/How_Perl_saved_human_genome

DNA sequencing (Simplified version)



DNA molecule (~3 billion bp)

Fragment the entire DNA into thousands of short fragments

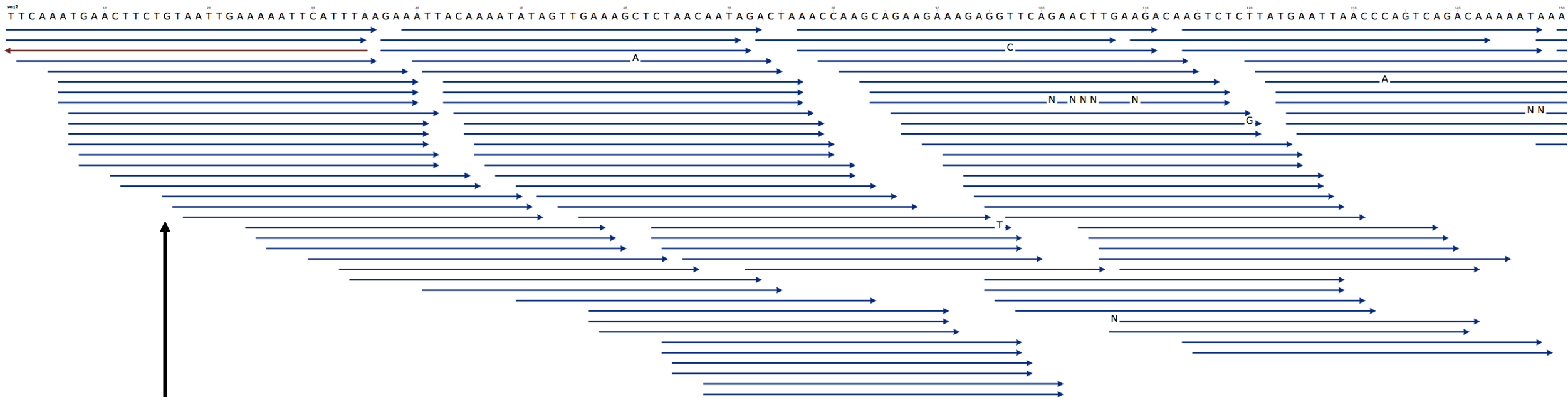
Make several copies of each fragment

DNA sequencer

ATGCCAGAACTAATCTATA
CTTATTTTTAAACATCATCA
CCAAAGAGAGAGTTCCAC

Sequence alignment of sequencing reads to Human Reference DNA sequence

Human Reference DNA sequence (from the Human Genome Project)



Sequencing DNA reads

Human Reference DNA sequence

Target Genome ATTTGCGCAGAGACCTAAGGCATTAGCTTGGCCCTAAAG

Reads ATTTGC AGAGACCTAAG TTAGCTTGGC AAG

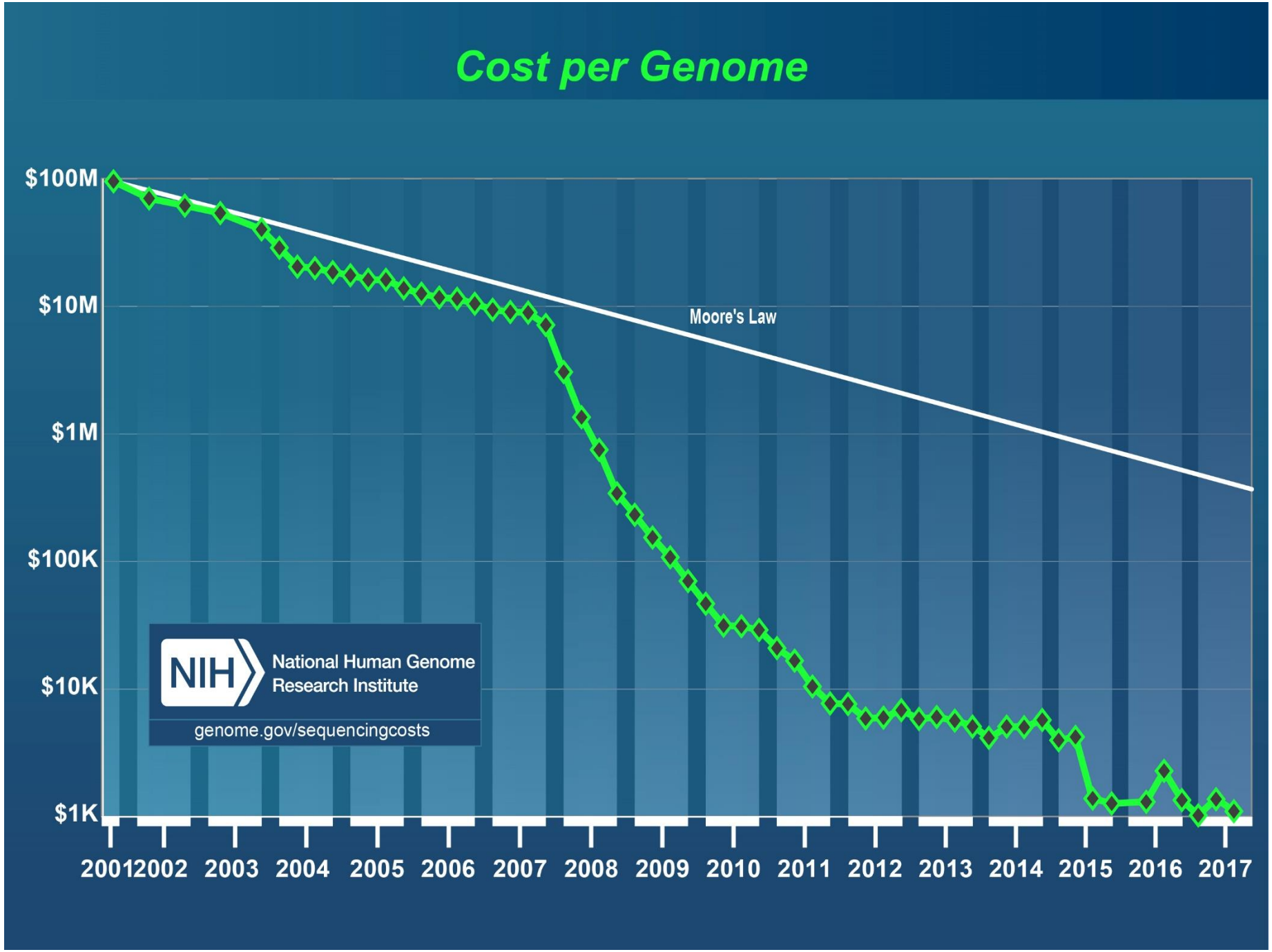
TGCGCAGA TGGCCCTAA

Overlapping ATTTGC AGAGACCTAAG TTAGCTTGGC AAG

TGCGCAGA TGGCCCTAA

Contigs ATTTGCGCAGAGACCTAAG TTAGCTTGGCCCTAAAG

https://towardsdatascience.com


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

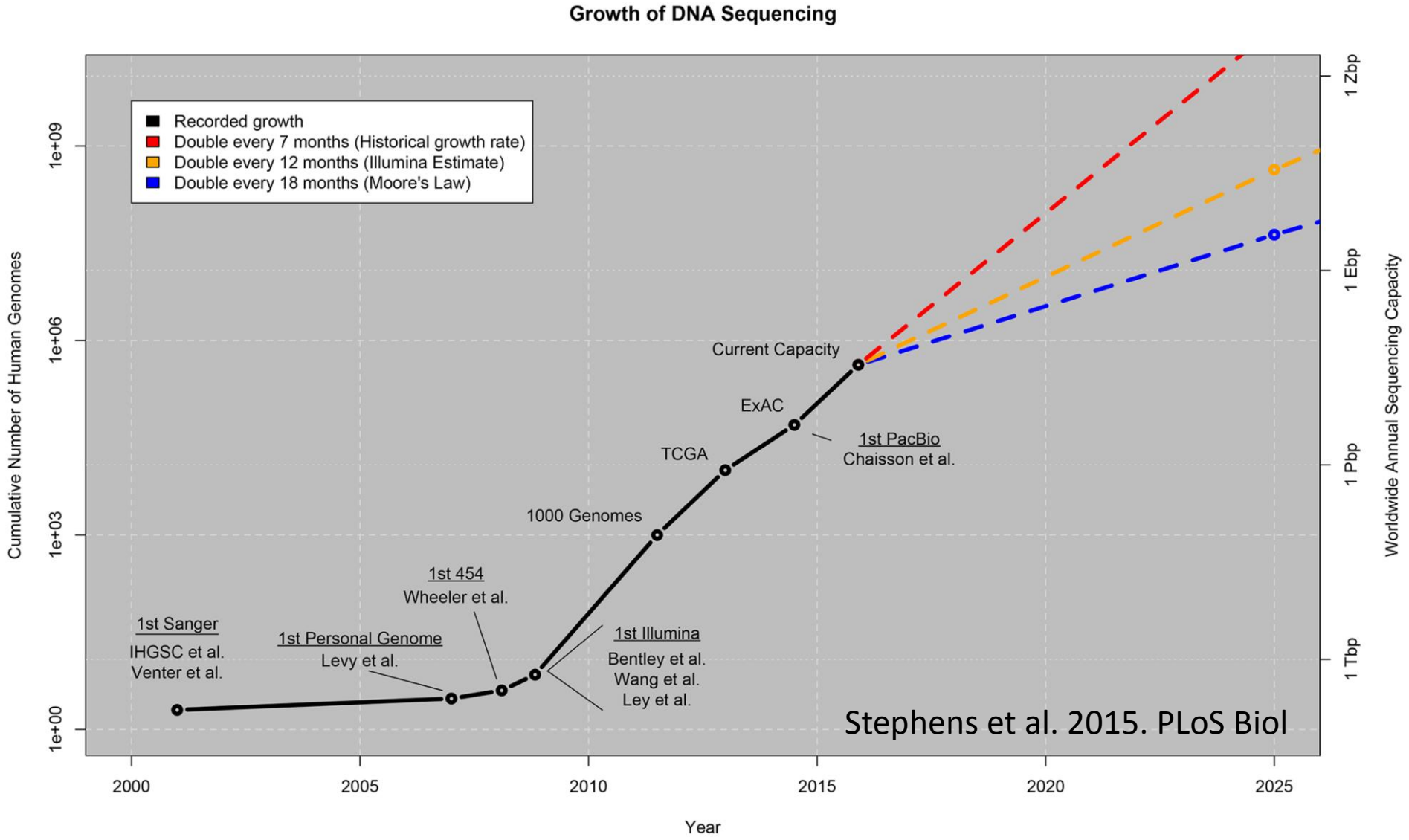
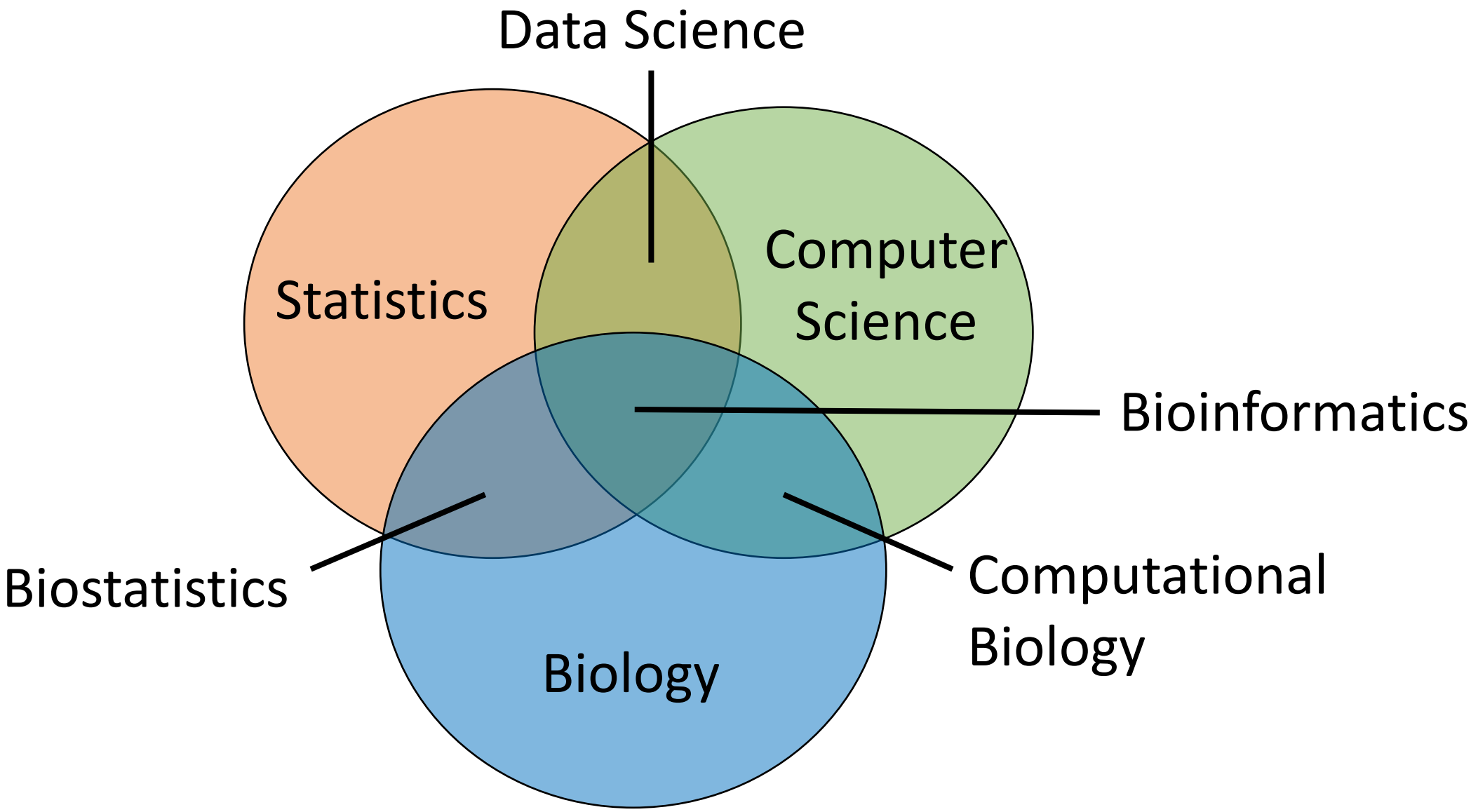


Fig 1. Growth of DNA sequencing. The plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepairs (Ebp), Zetta-basepairs (Zbps)). The values through 2015 are based on the historical publication record, with selected milestones in sequencing (first Sanger through first PacBio human genome published) as well as three exemplar projects using large-scale sequencing: the 1000 Genomes Project, aggregating hundreds of human genomes by 2012 [3]; The Cancer Genome Atlas (TCGA), aggregating over several thousand tumor/normal genome pairs [4]; and the Exome Aggregation Consortium (ExAC), aggregating over 60,000 human exomes [5]. Many of the genomes sequenced to date have been whole exome rather than whole genome, but we expect the ratio to be increasingly favored towards whole genome in the future. The values beyond 2015 represent our projection under three possible growth curves as described in the main text.

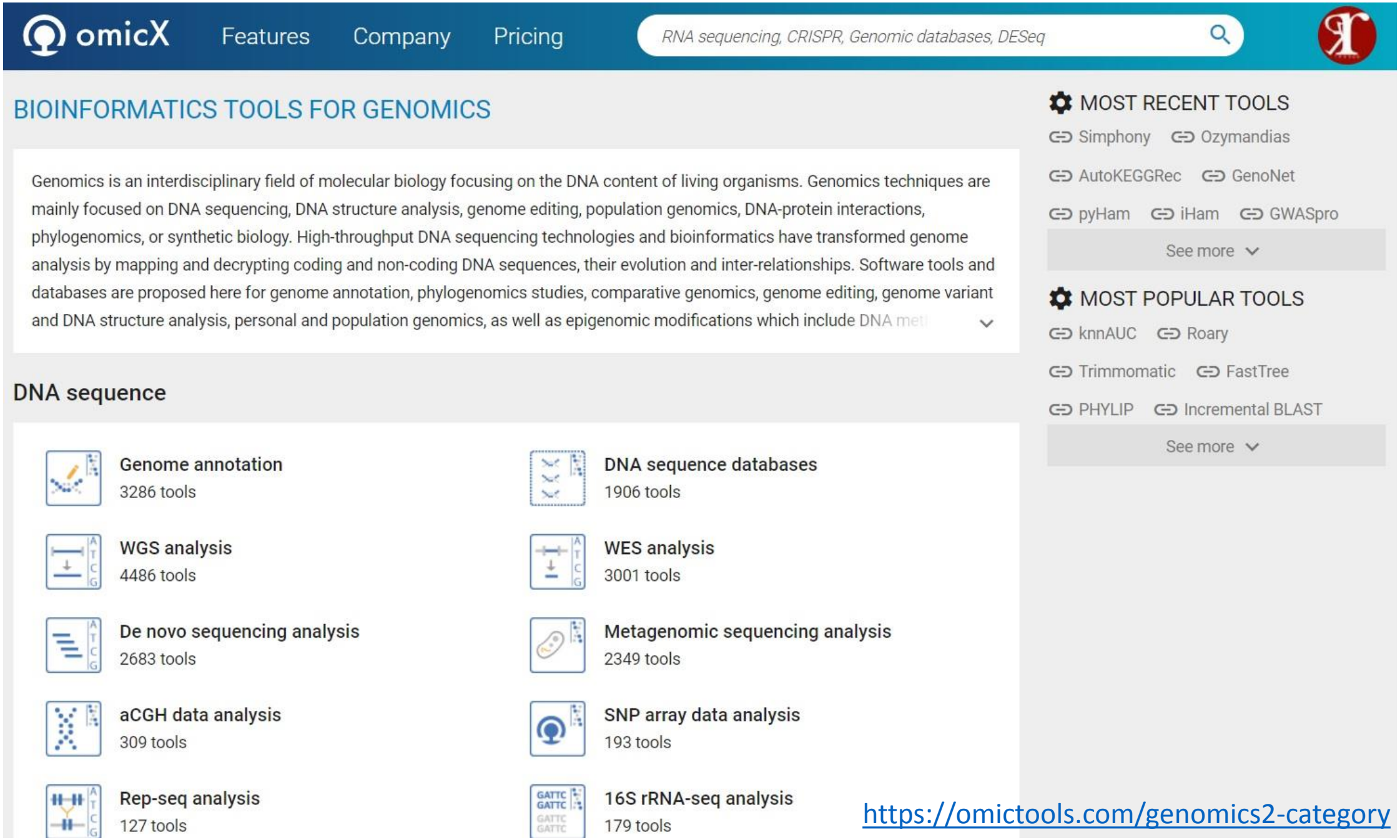


you can think of **bioinformatics/computational biology** as **data science in biology**

Different Roles of Bioinformatician / Computational Biologist

- **Algorithm Development**
  - Skills in Maths + Stats. + Comp. Sci. essential. Knowledge in biology required.
- **Software Development**
  - Skills in coding & documentations

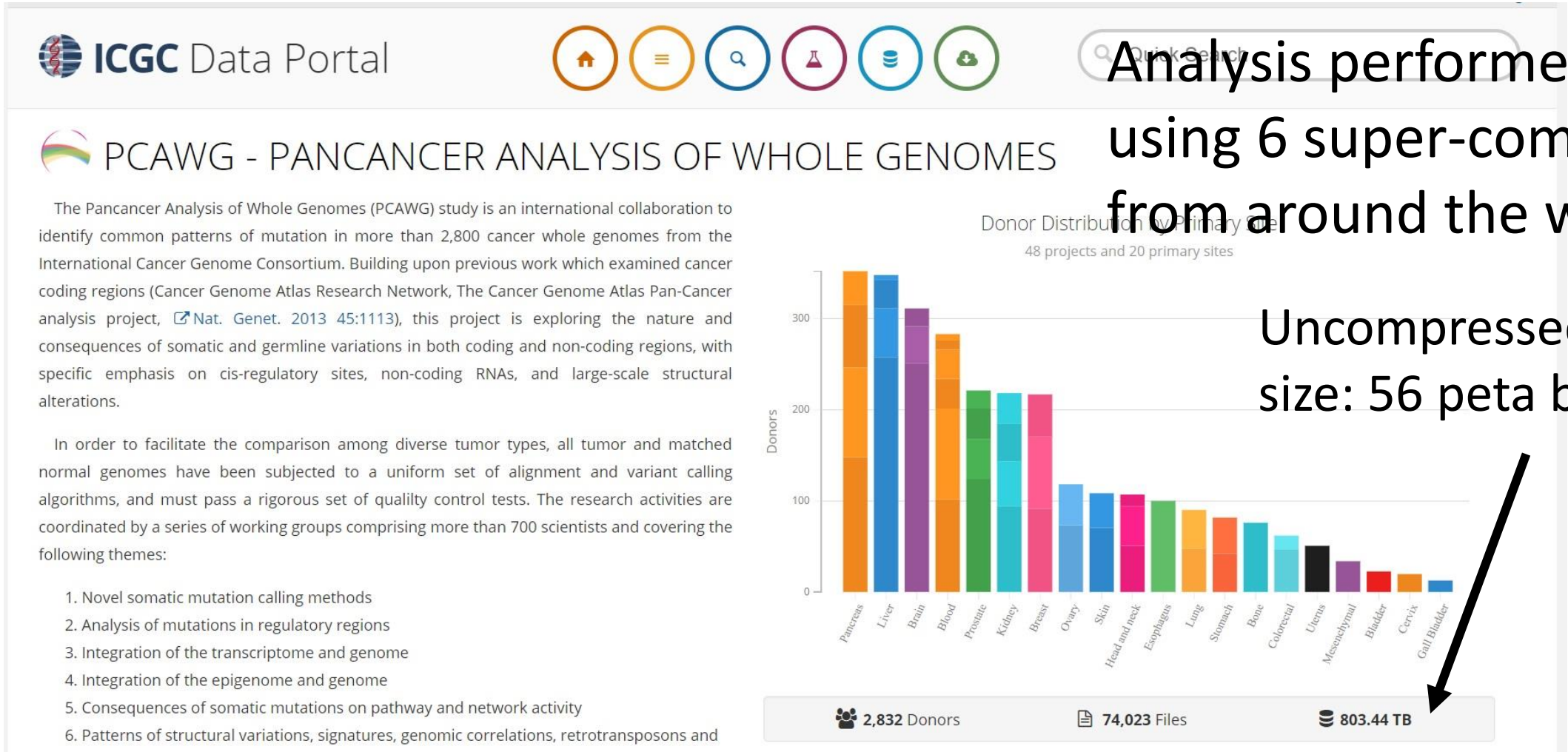
- **Data Processing**
  - implementation of bioinformatics pipelines, i.e. processing data generated by high-throughput analytical instruments ready to be analyzed by biologist
- **Data Analysis**
  - Knowledge in biology essential. Skills in Maths + Stats. + Comp. Sci. required.



https://omictools.com/genomics2-category

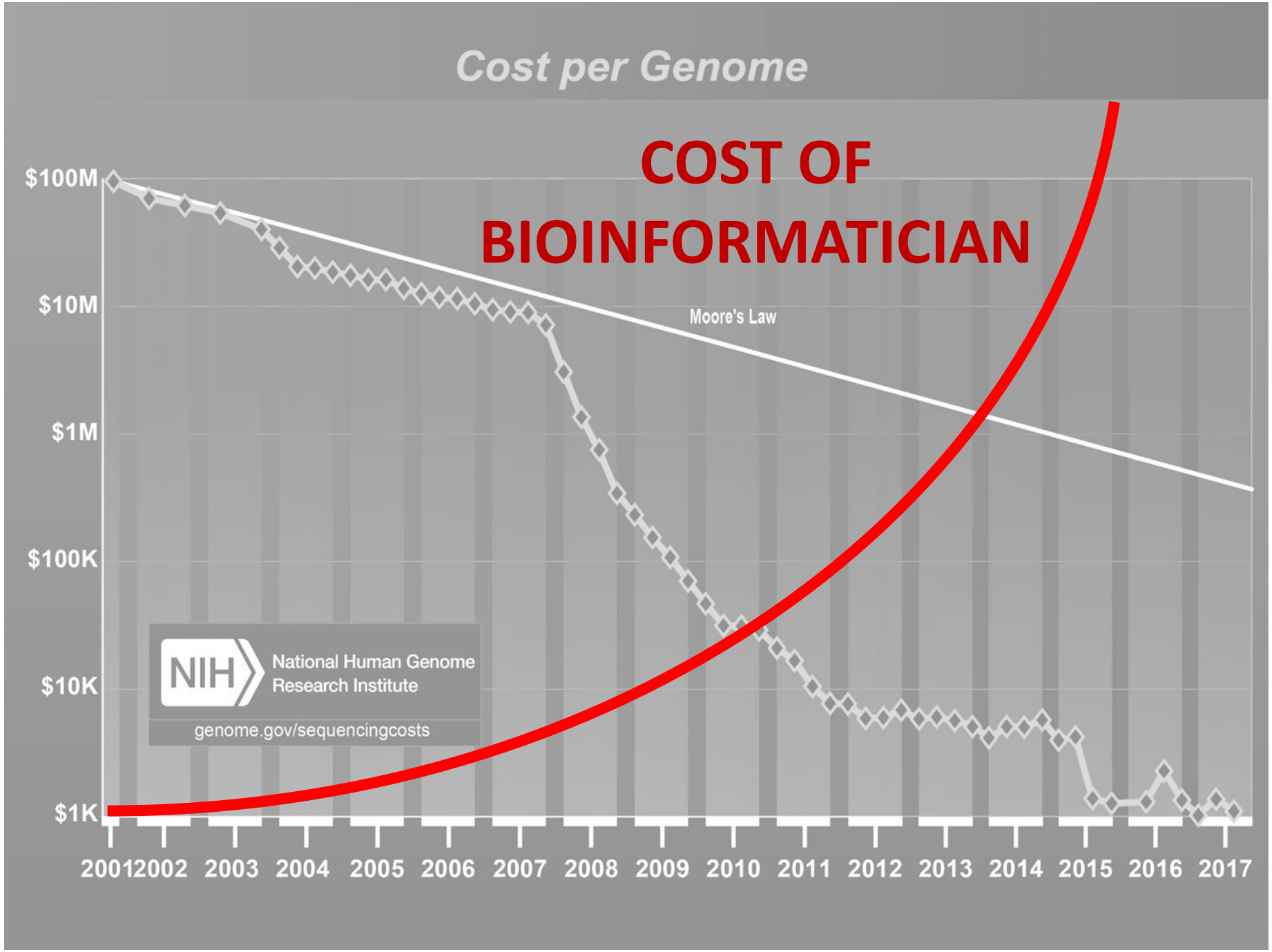
Bioinformatics is a Big Data Problem

https://dcc.icgc.org/pcawg

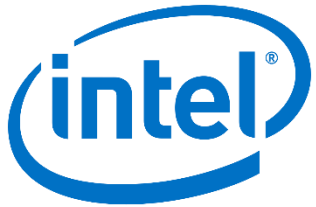


Analysis performed using 6 super-computers from around the world

Uncompressed data size: 56 peta bytes


Cost per Genome
COST OF BIOINFORMATICIAN
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

Major IT companies heavily investing in Bioinformatics



TEA BREAK

After the break …..

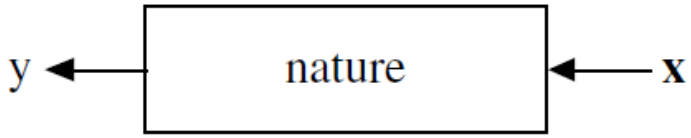
How machine learning is being used to solve biological problems?

PART - II

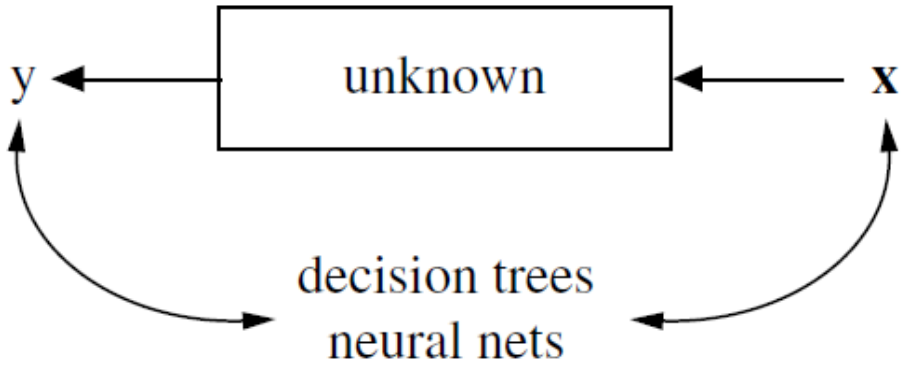
Applications of Machine Learning in Biology & Medicine

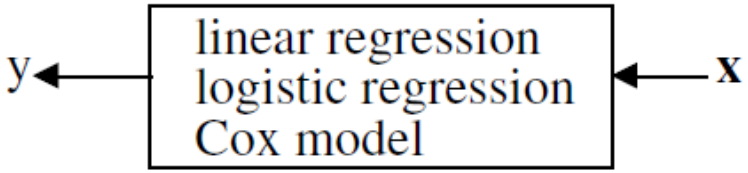
- **WHY** is machine learning a good tool for biological studies?
- **WHERE** in biology & medicine machine learning tools are being used?
- **WHEN NOT TO USE** machine learning?

Two cultures of Machine Learning



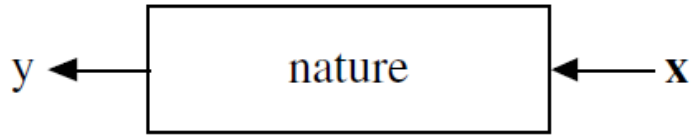
BLACK BOX APPROACH **OPEN BOX APPROACH**





Breiman L. Statistical Science. 2001

Two cultures of Machine Learning



BLACK BOX APPROACH

Classification of Cat vs Dog

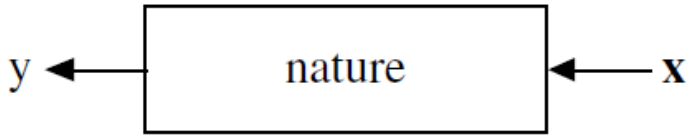
Features that separate Cat vs Dog **may not be important** for the end-user

OPEN BOX APPROACH

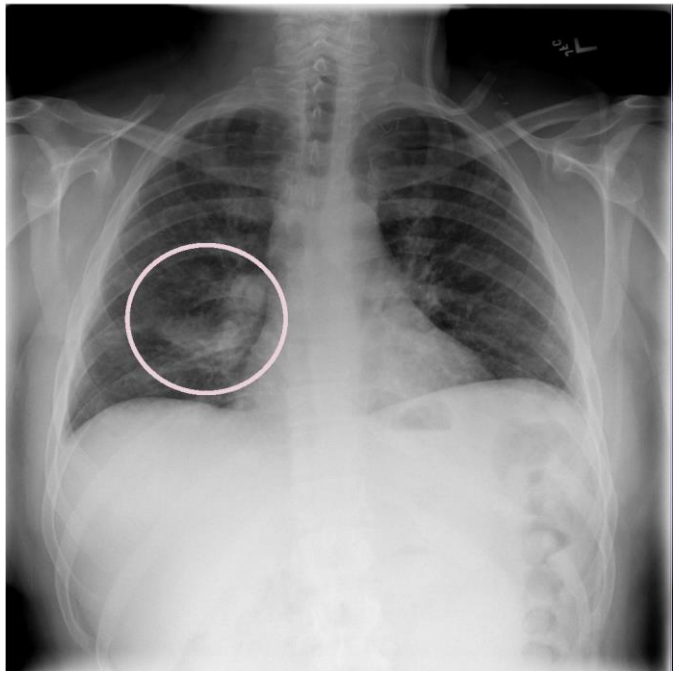
Classification of Cat vs Dog

Features that separate Cat vs Dog **is important** for the end-user

Two cultures of Machine Learning



BLACK BOX APPROACH



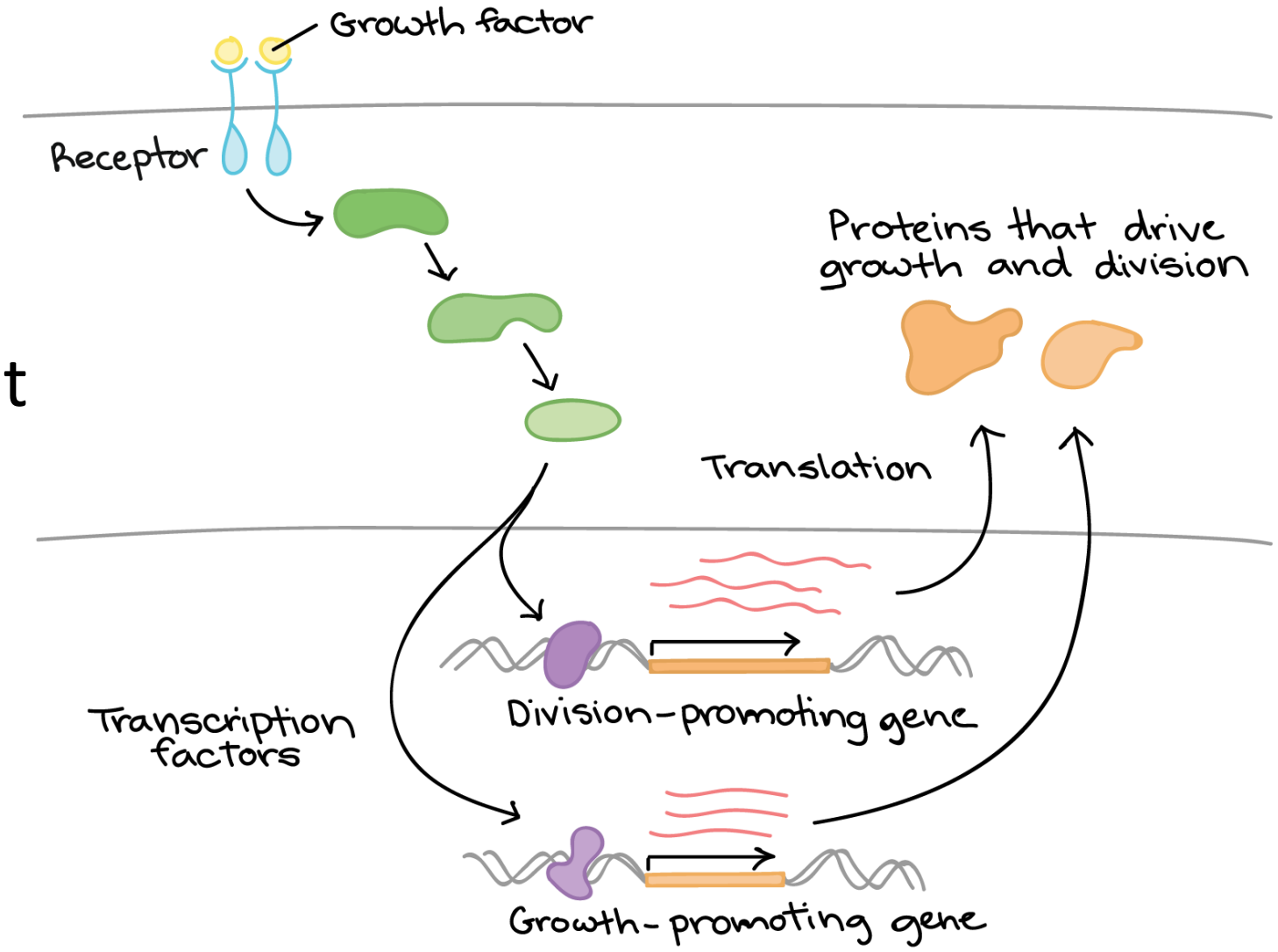
Diagnosis of Tuberculosis (TB) using X-ray images (Image classification)

OPEN BOX APPROACH

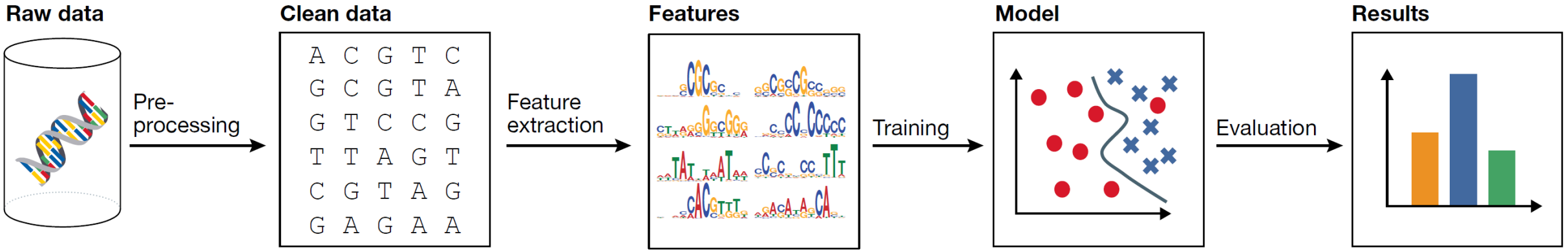
WHY does some patients responds to Drug-X while others does not respond the drug at all?

Gene Regulation

The pattern of DNA sequence is very important for attachment of key protein that regulate the function of a gene.



Machine learning workflow

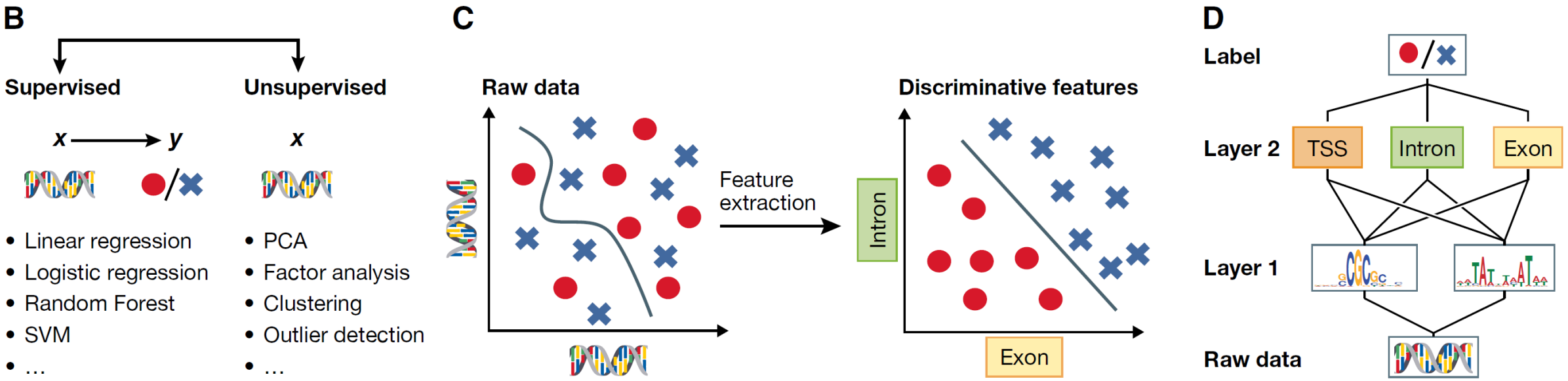


The classical machine learning workflow can be broken down into four steps:

- **data pre-processing**,
- **feature extraction**,
- **model learning** and
- **model evaluation**.

Angermueller et al. 2016. Mol. Sys. Bio.

Machine learning approaches



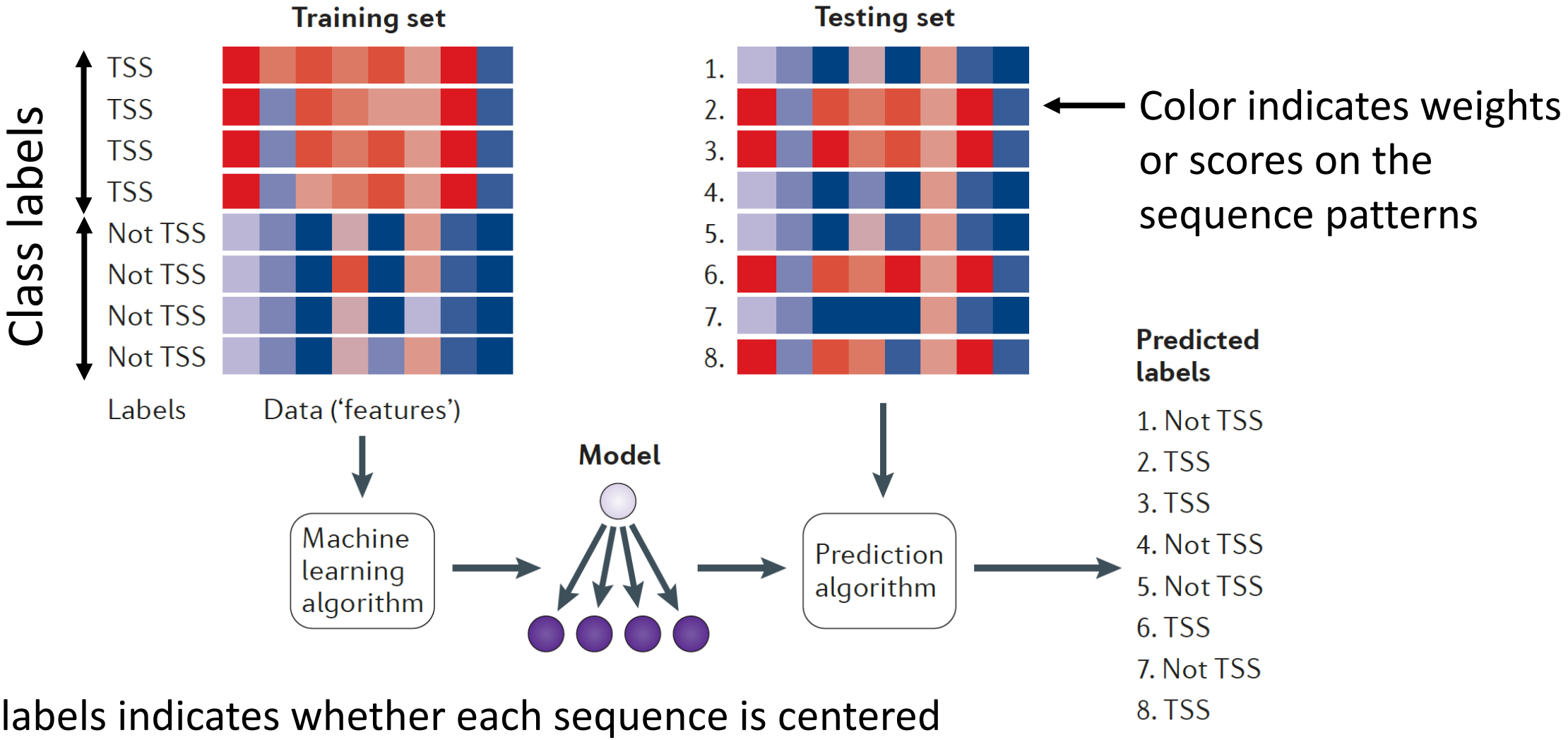
Raw input data are often high-dimensional and related to the corresponding label in a complicated way, which is challenging for many classical machine learning algorithms (left plot). Alternatively, higher-level features extracted using a deep model may be able to better discriminate between classes (right plot).

Deep networks use a hierarchical structure to learn increasingly abstract feature representations from the raw data.

Angermueller et al. 2016. Mol. Sys. Bio.

An example of a machine learning application

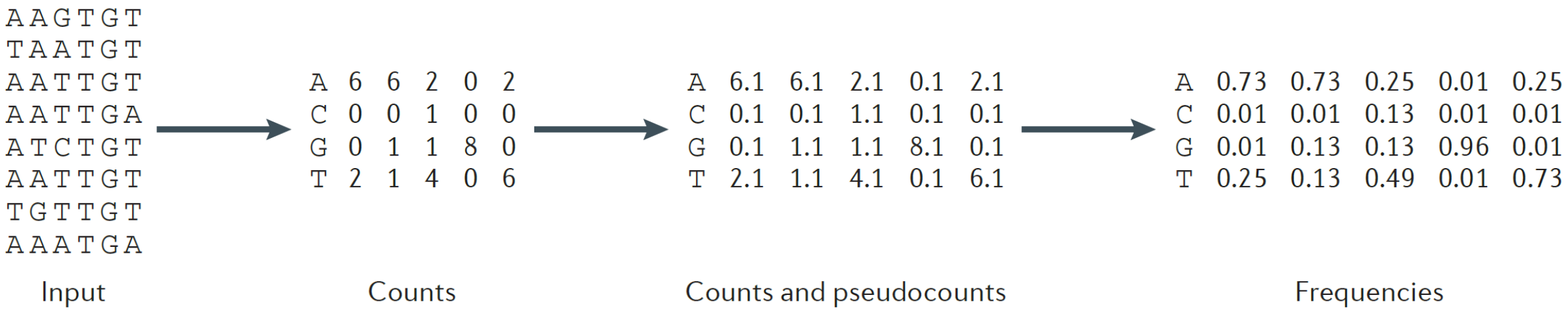
Training set of DNA sequences



binary labels indicates whether each sequence is centered on a transcription start site (TSS) or not.

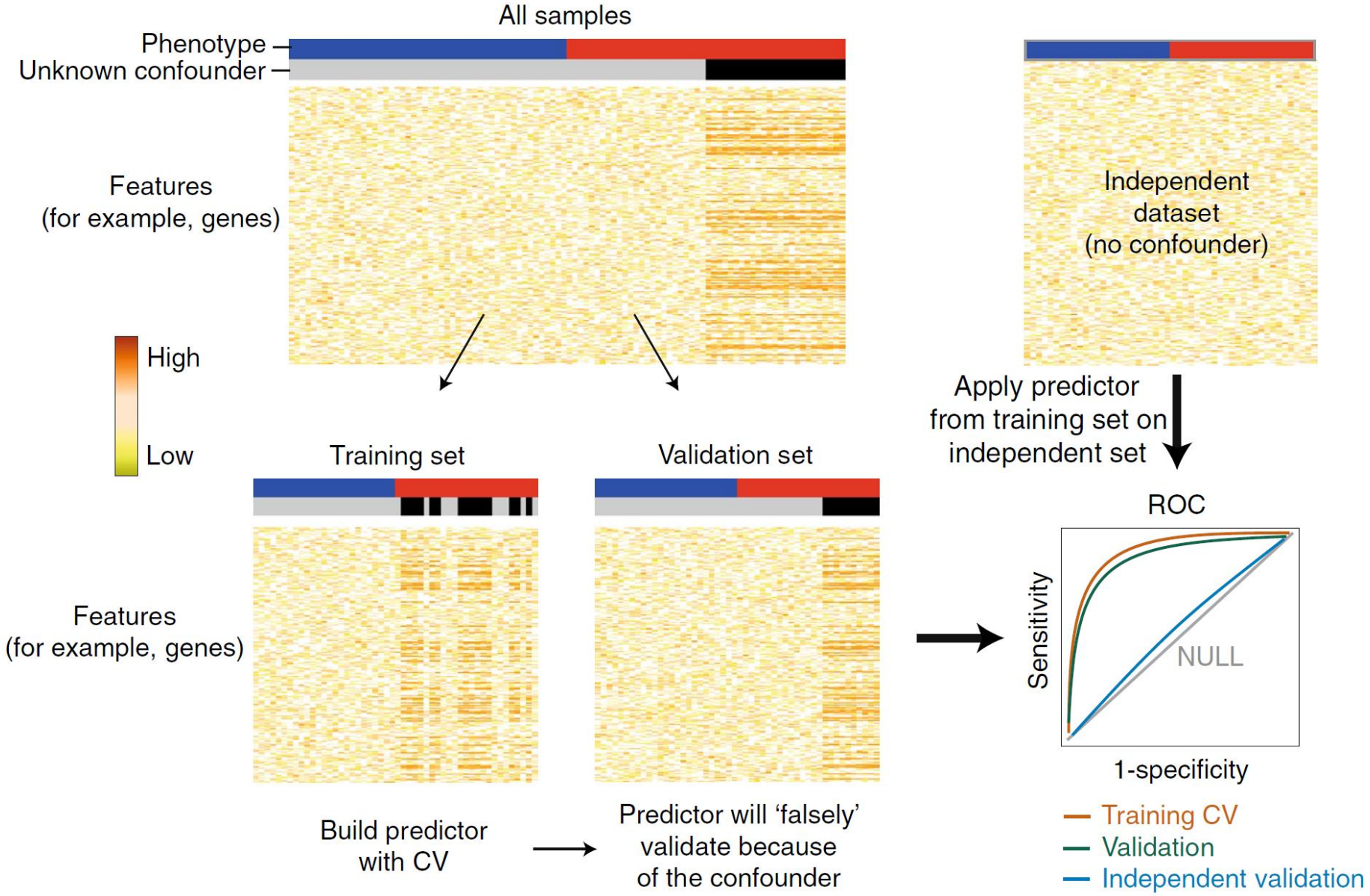
Libbrecht and Nobel. 2015. Nature Review Genetics

Sequence to Scores



Libbrecht and Nobel. 2015. Nature Review Genetics

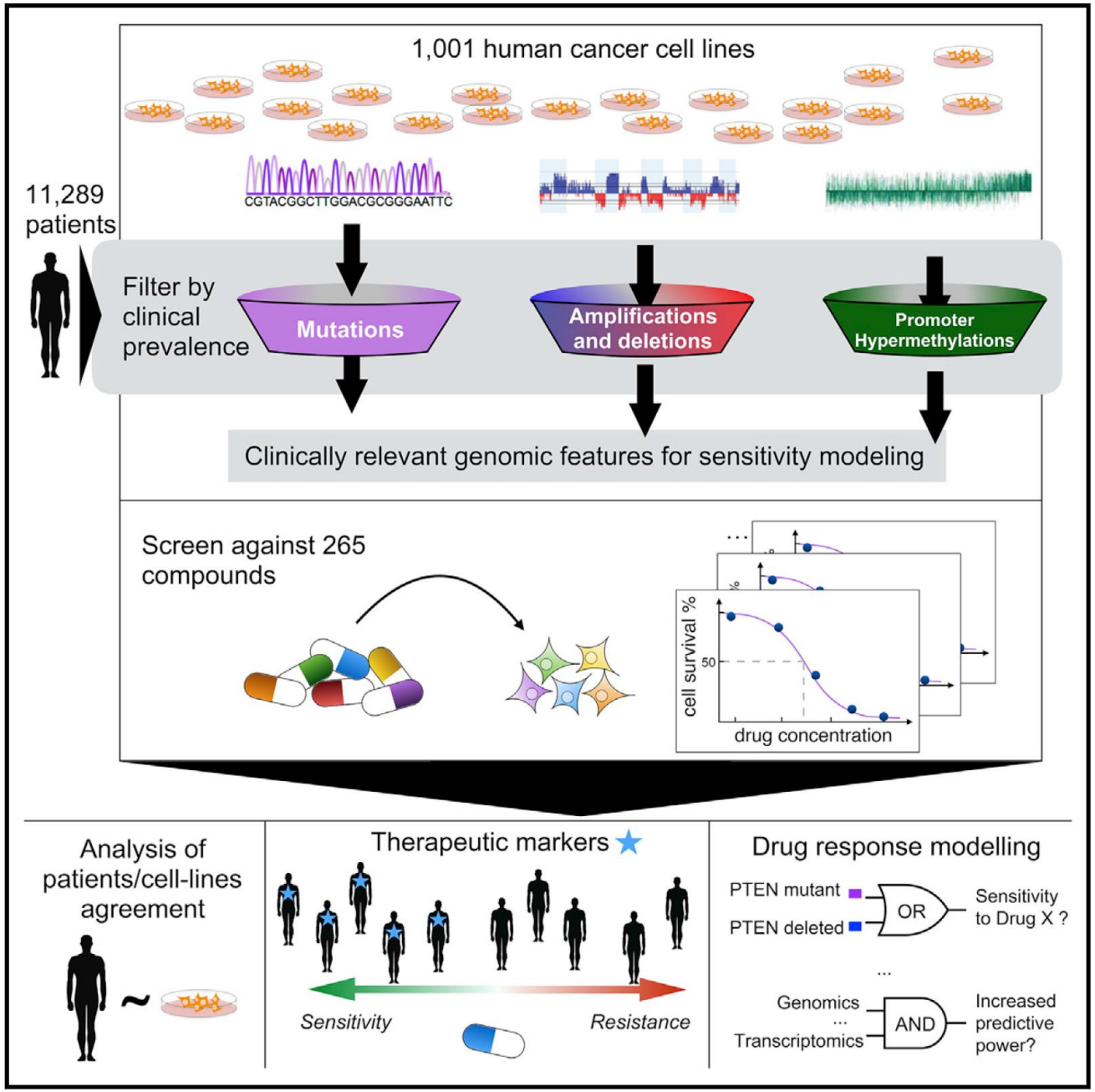
Using Gene Expression Data

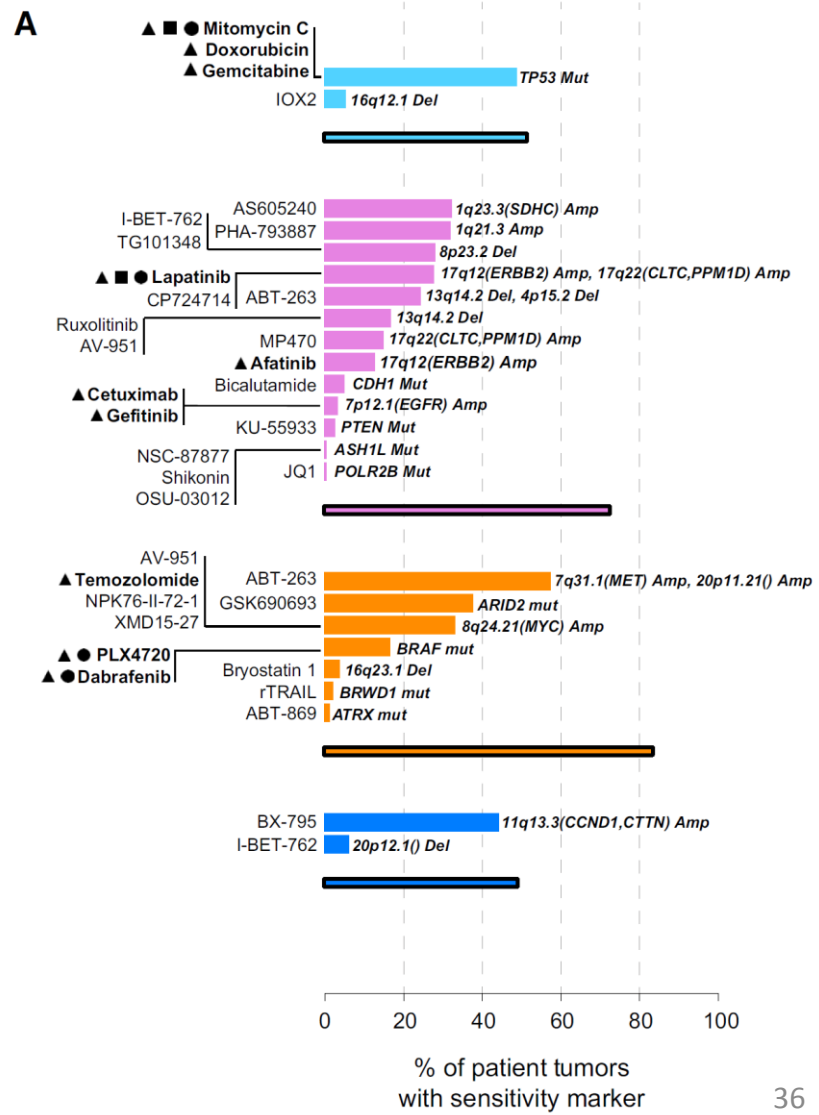


Teschendorff, 2018, Nature Materials

Cell

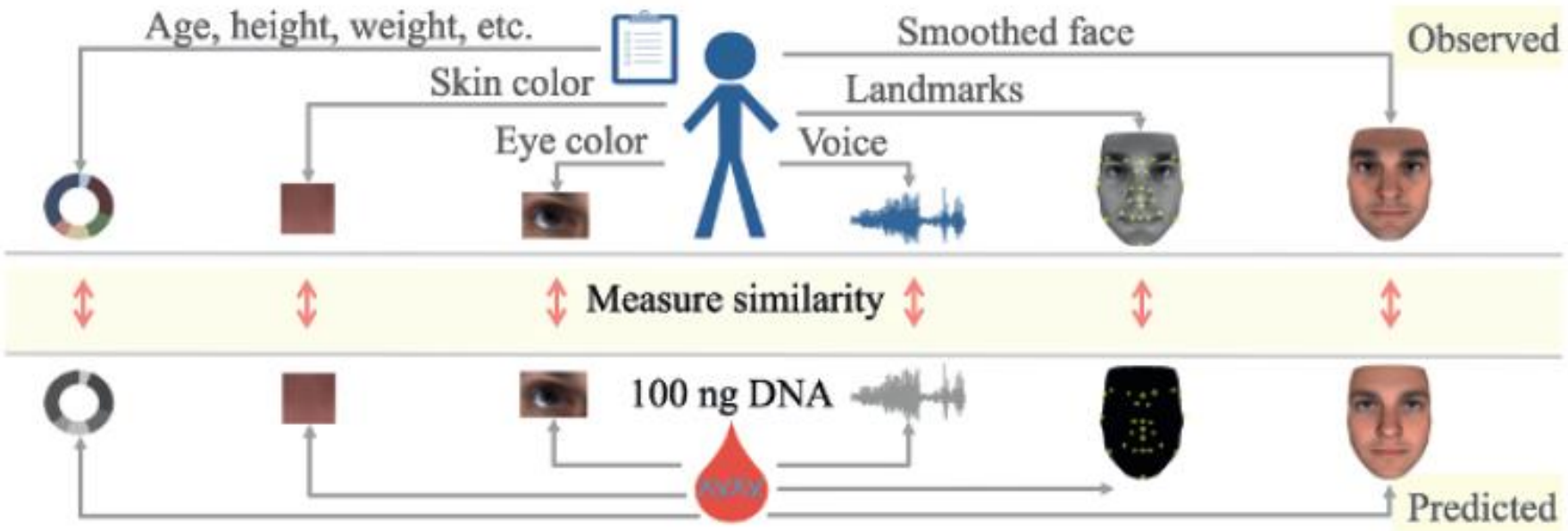
A Landscape of Pharmacogenomic Interactions in Cancer





Identification of individuals by trait prediction using whole-genome sequencing data

Christoph Lippert[a,1], Riccardo Sabatini[a], M. Cyrus Maher[a], Eun Yong Kang[a], Seunghak Lee[a], Okan Arikan[a], Alena Harley[a], Axel Bernal[a], Peter Garst[a], Victor Lavrenko[a], Ken Yocum[a], Theodore Wong[a], Mingfu Zhu[a], Wen-Yun Yang[a], Chris Chang[a], Tim Lu[b], Charlie W. H. Lee[b], Barry Hicks[a], Smriti Ramakrishnan[a], Haibao Tang[a], Chao Xie[c], Jason Piper[c], Suzanne Brewerton[c], Yaron Turpaz[b,c], Amalio Telenti[b], Rhonda K. Roby[b,d,2], Franz J. Och[a], and J. Craig Venter[b,d,1]



A universal SNP and small-indel variant caller using deep neural networks

nature biotechnology

Ryan Poplin[1,2], Pi-Chuan Chang[2], David Alexander[2], Scott Schwartz[2], Thomas Colthurst[2], Alexander Ku[2], Dan Newburger[1], Jojo Dijamco[1], Nam Nguyen[1], Pegah T Afshar[1], Sam S Gross[1], Lizzie Dorfman[1,2], Cory Y McLean[1,2] & Mark A DePristo[1,2]

Pileup: Standard format for mapped data, position summaries

```
seq1 272 T 24  ,.$.....,.,.,...,,,.,..^+.   <<<+;<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,.,.,...,,,.,..A       <<<;<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$....,.,.,...,,,.,...      7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$....,.,.,...,,,.,...^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,...,,,.,....           33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ....,.,.,.C.,,,.,..G.            +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  ....,.,.,...,,,.,....^k.  %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,...,,,.,.....      ;75&<<<<<<<<<=<<<9<<:<<
```

Seq.

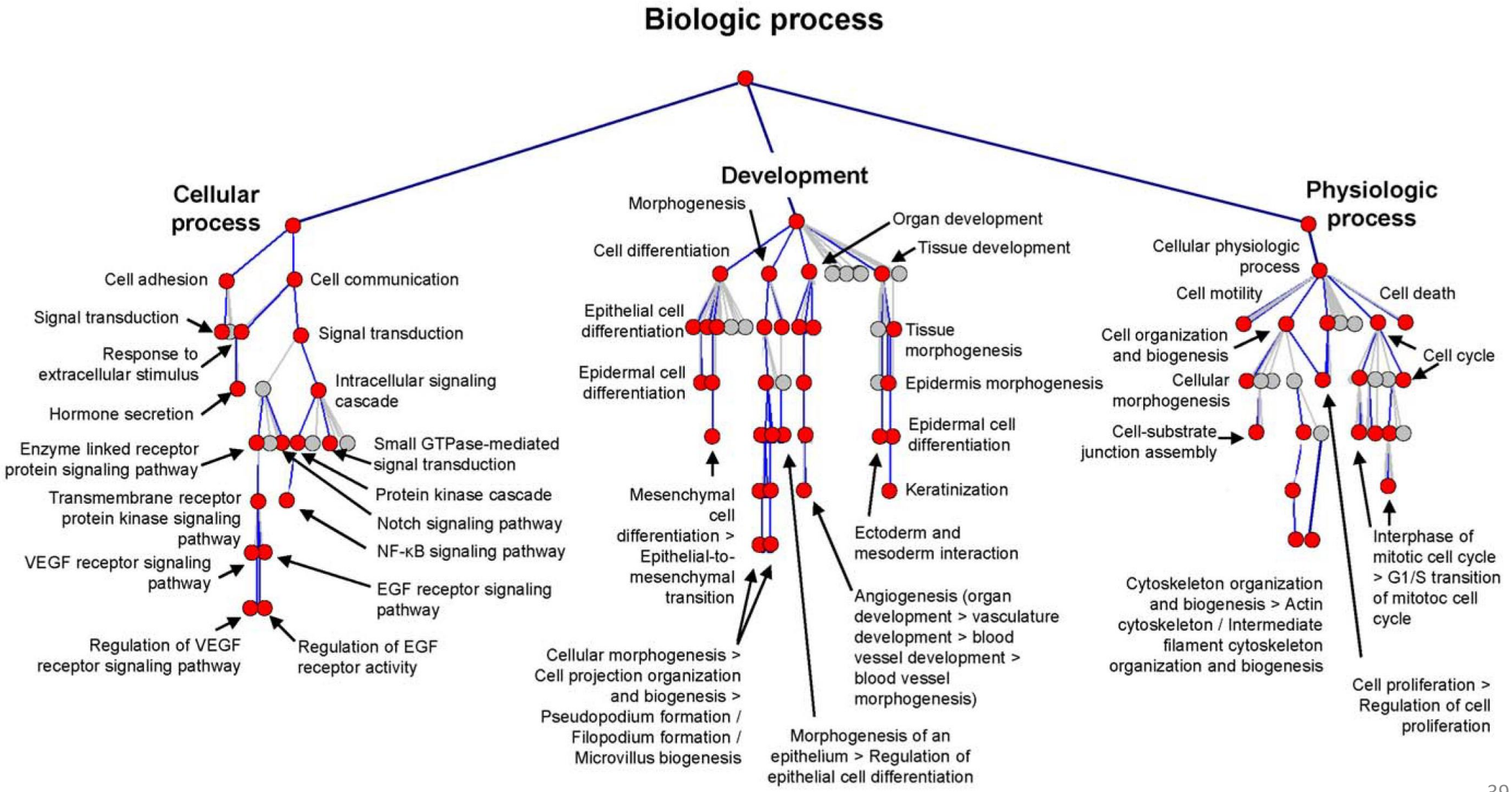
Pos. Len. Alignment Quality

Ref.

Google's DeepVariant: Uses Pileup images as input to identify changes in DNA sequences utilizing CNN.

Treats pattern of strings as "images"

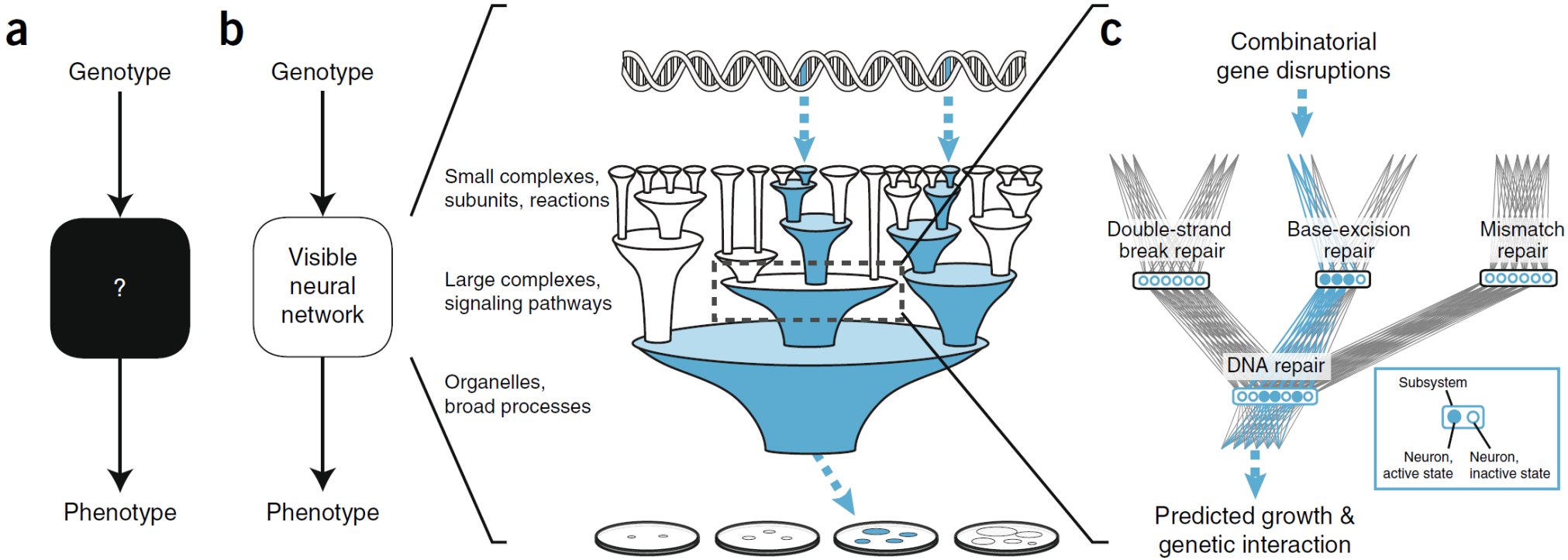
Hierarchical mapping of Gene Ontology (GO) categories



NATURE METHODS

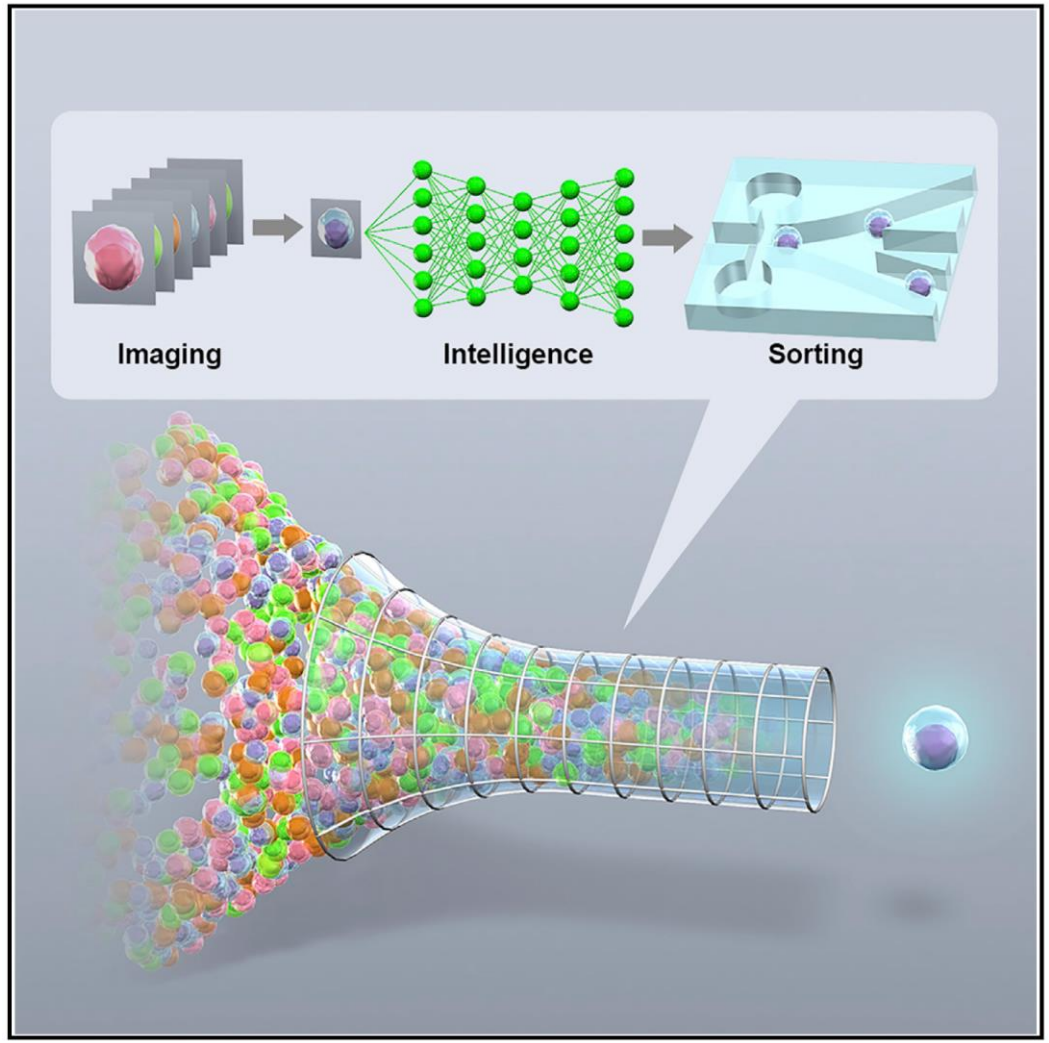
Using deep learning to model the hierarchical structure and function of a cell

Jianzhu Ma[1,5], Michael Ku Yu[1,2,5], Samson Fong[1,3,5], Keiichiro Ono[1], Eric Sage[1], Barry Demchak[1], Roded Sharan[4] & Trey Ideker[1–3]



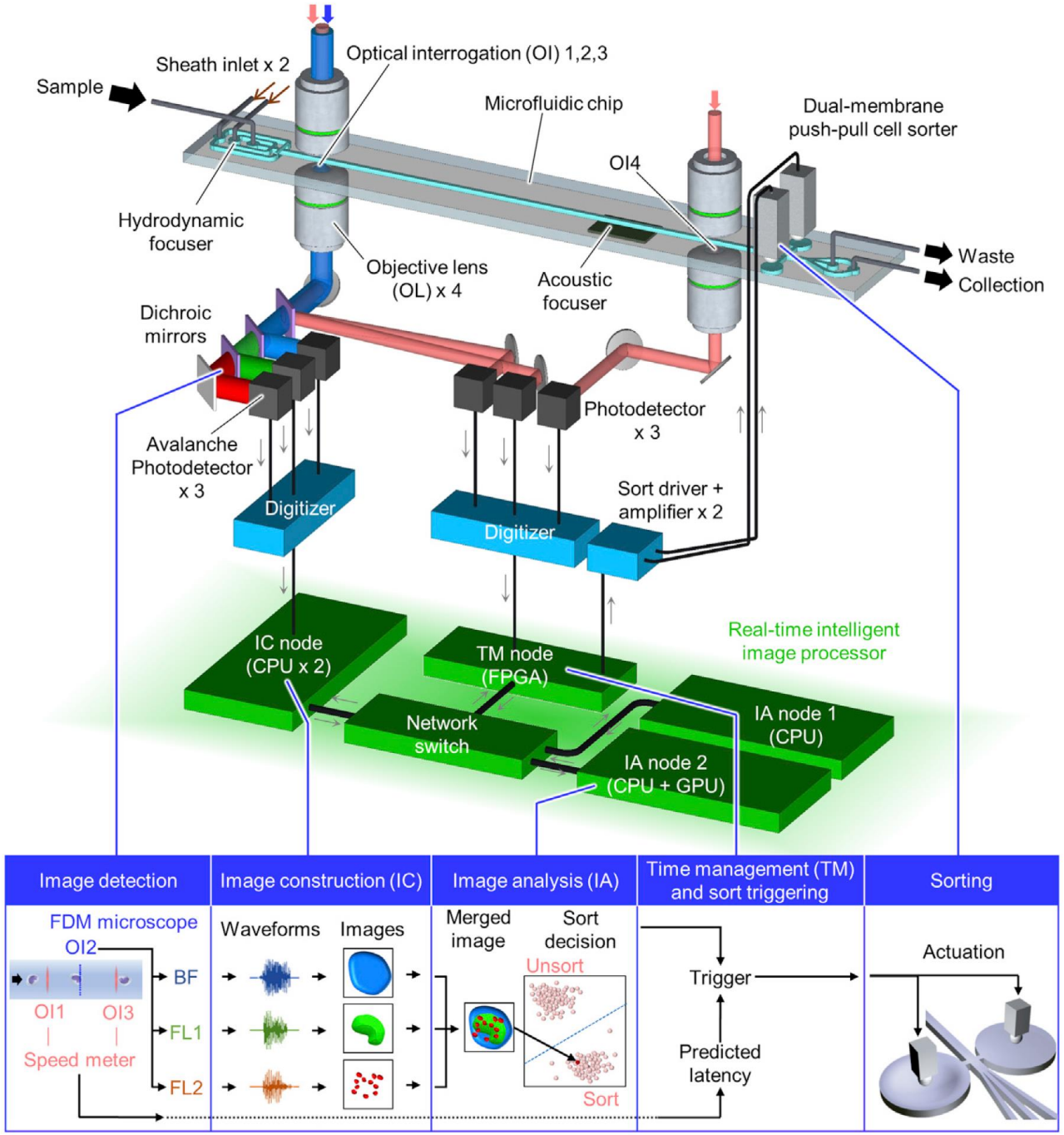
Cell

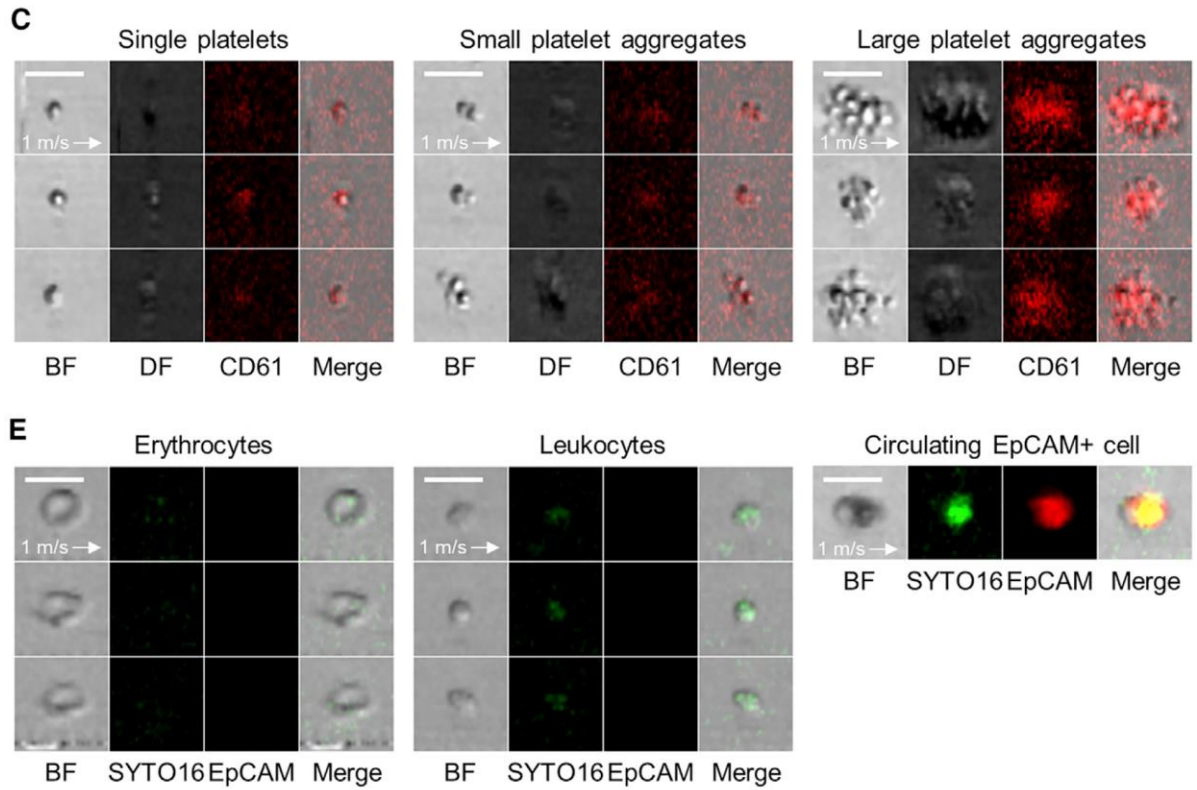
Intelligent Image-Activated Cell Sorting



In Brief

Artificial-intelligence-assisted, image-based flow cytometry in real-time enables rapid cell sorting based on unique chemical and morphological features.

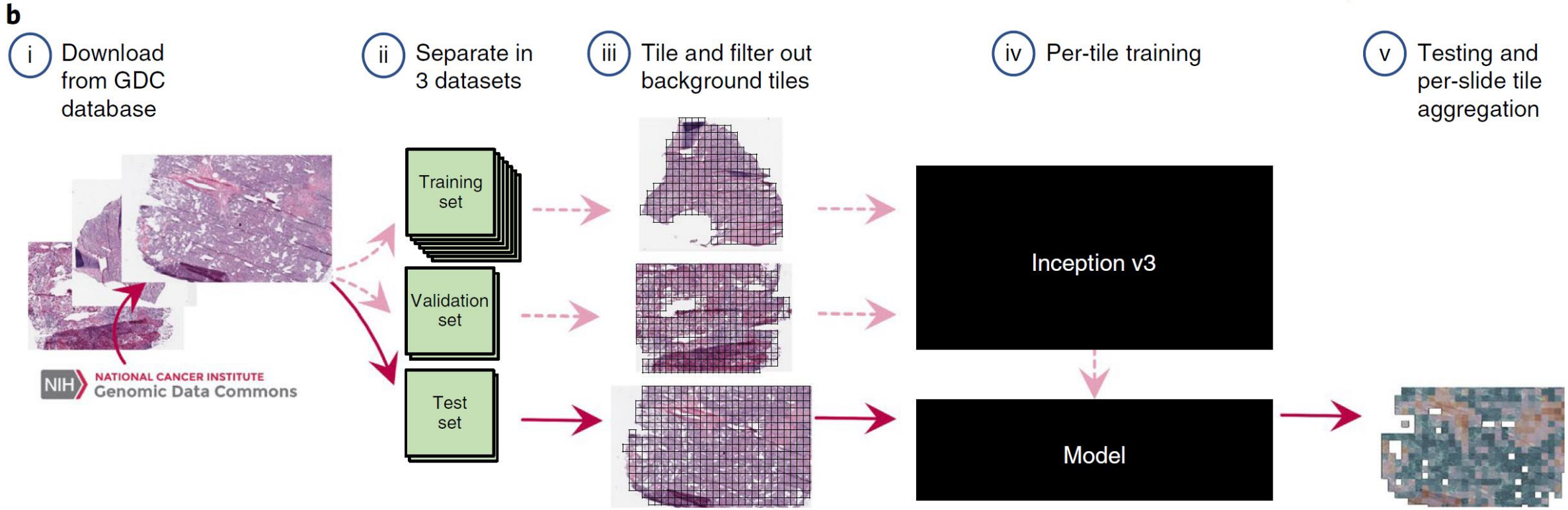

Sample
Sheath inlet x 2
Optical interrogation (OI) 1,2,3
Microfluidic chip
Dual-membrane push-pull cell sorter
OI4
Hydrodynamic focuser
Objective lens (OL) x 4
Acoustic focuser
Waste
Collection
Dichroic mirrors
Photodetector x 3
Avalanche Photodetector x 3
Digitizer
Digitizer
Sort driver + amplifier x 2
Real-time intelligent image processor
IC node (CPU x 2)
TM node (FPGA)
Network switch
IA node 1 (CPU)
IA node 2 (CPU + GPU)
Image detection
Image construction (IC)
Image analysis (IA)
Time management (TM) and sort triggering
Sorting
FDM microscope
OI2
OI1
OI3
Speed meter
BF
FL1
FL2
Waveforms
Images
Merged image
Sort decision
Unsort
Sort
Trigger
Predicted latency
Actuation


C
Single platelets
Small platelet aggregates
Large platelet aggregates
1 m/s
BF
DF
CD61
Merge
E
Erythrocytes
Leukocytes
Circulating EpCAM+ cell
SYTO16
EpCAM

Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning

nature medicine

Nicolas Coudray[1,2,9], Paolo Santiago Ocampo[3,9], Theodore Sakellaropoulos[4], Navneet Narula[3], Matija Snuderl[3], David Fenyö[5,6], Andre L. Moreira[3,7], Narges Razavian[8*] and Aristotelis Tsirigos[1,3*]



Deep Genomic Signature for early metastasis prediction in prostate cancer

Hossein Sharifi-Noghabi [1,3], Yang Liu [2], Nicholas Erho [2], Raunak Shrestha [3,4], Mohammed Alshalalfa [2], Elai Davicioni [2], Colin C. Collins [1,3,4], and Martin Ester [1,3,*]

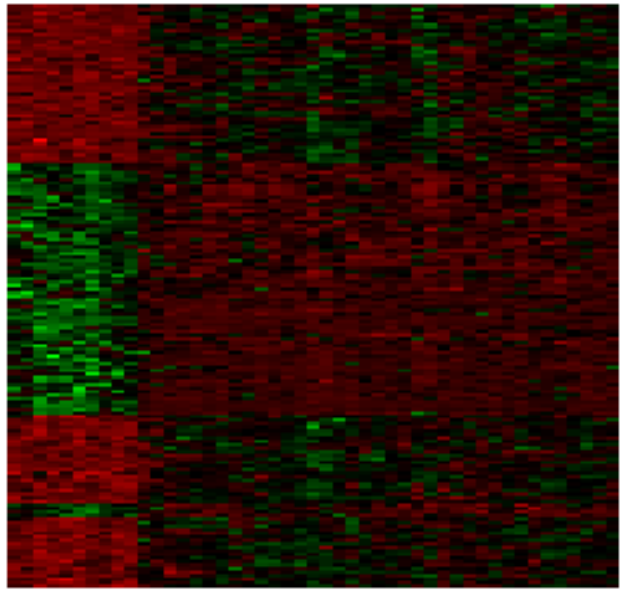
[1]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada.
[2]GenomeDx Biosciences, Vancouver, BC, Canada.
[3]Vancouver Prostate Centre, Vancouver, BC, Canada.
[4]Department of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada

Sharifi-Noghabi et al. (in submission)
https://doi.org/10.1101/276055

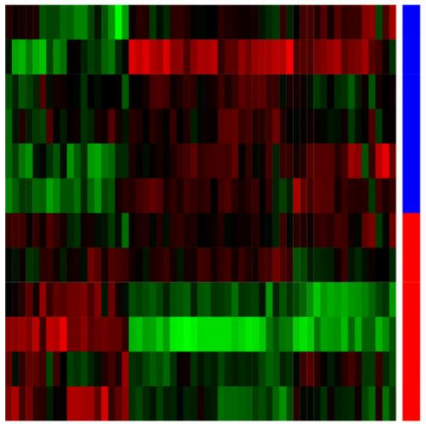
- Metastasis is the medical term for cancer that spreads to a different part of the body from where it started.
- Metastasis are mostly lethal
- So **Early Prediction** of which patients are highly likely to develop metastasis is very important

Can we improve the performance of genomic classifiers by using poorly- and well-annotated cohorts together?

Poorly-annotated cohort
(Without long-term clinical outcomes)



Well-annotated cohort
(With long-term clinical outcomes)

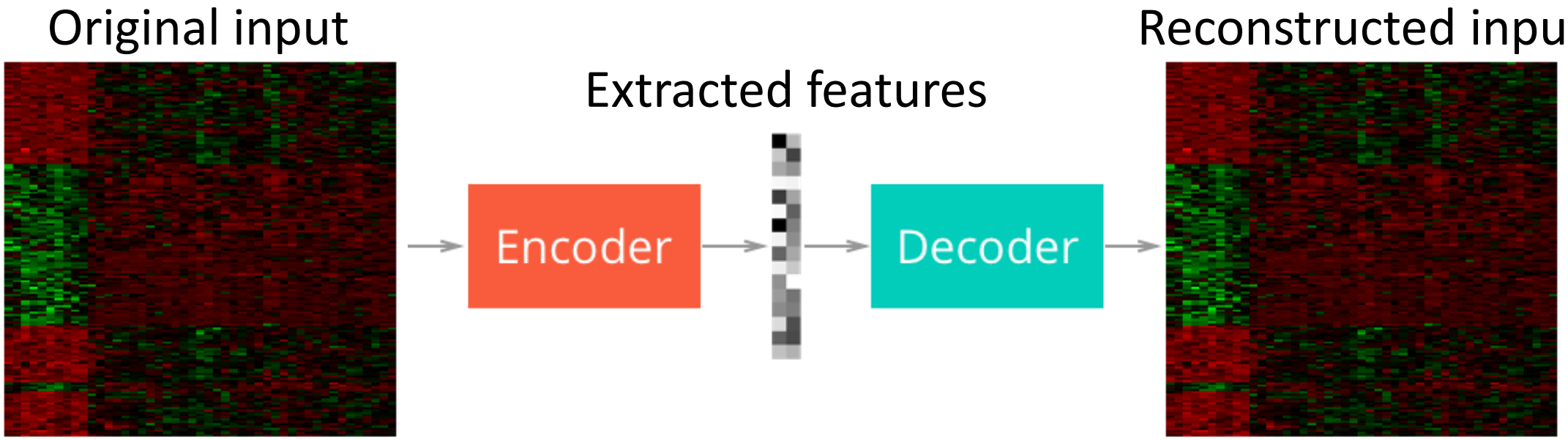


?

Classifier

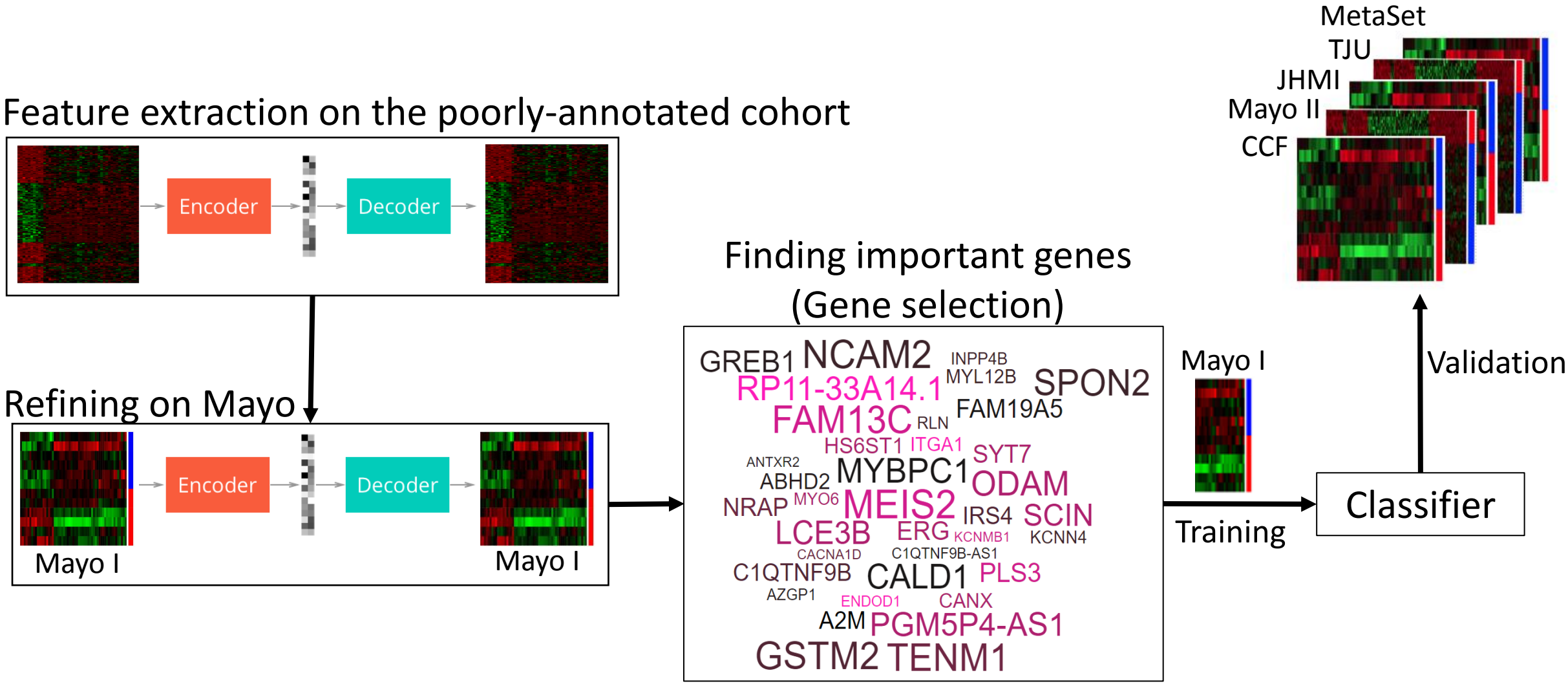
Sharifi-Noghabi et al. (in submission)

Feature extraction by Autoencoders



Sharifi-Noghabi et al. (in submission)

Deep Genomic Signature (DGS)



Sharifi-Noghabi et al. (in submission)

HIT'nDRIVE: Multi-driver Gene Prioritization Based on Hitting Time

Raunak Shrestha[1,2,⋆], Ermin Hodzic[3,⋆], Jake Yeung[2,4,⋆], Kendric Wang[2], Thomas Sauerwald[5], Phuong Dao[6], Shawn Anderson[2], Himisha Beltran[7], Mark A. Rubin[7], Colin C. Collins[2,8], Gholamreza Haffari[9], and S. Cenk Sahinalp[3,10]

Shrestha et al. 2014 (RECOMB-2014)

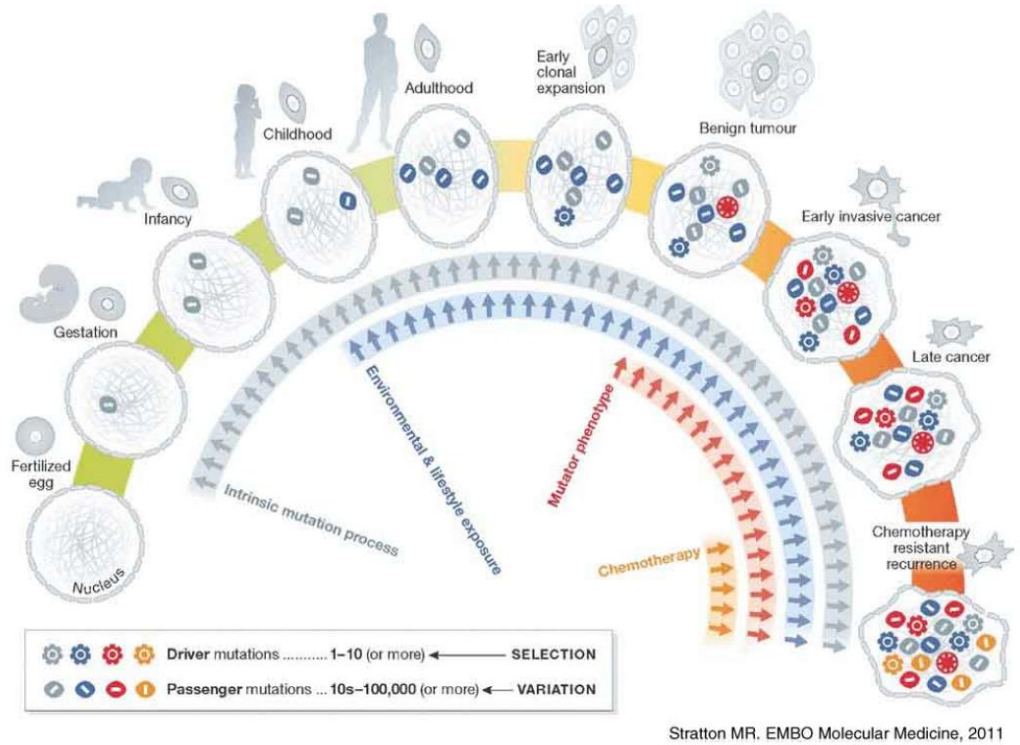
Method

HIT'nDRIVE: patient-specific multidriver gene prioritization for precision oncology

Raunak Shrestha,[1,2,10] Ermin Hodzic,[3,10] Thomas Sauerwald,[4] Phuong Dao,[5] Kendric Wang,[2] Jake Yeung,[2] Shawn Anderson,[2] Fabio Vandin,[6] Gholamreza Haffari,[7] Colin C. Collins,[2,8] and S. Cenk Sahinalp[2,3,9]

Shrestha et al. 2017 (Genome Research)

Genes that Drive Cancer



Stratton MR. EMBO Molecular Medicine, 2011

- Cancer is mediated by somatic evolution of various alterations in genome

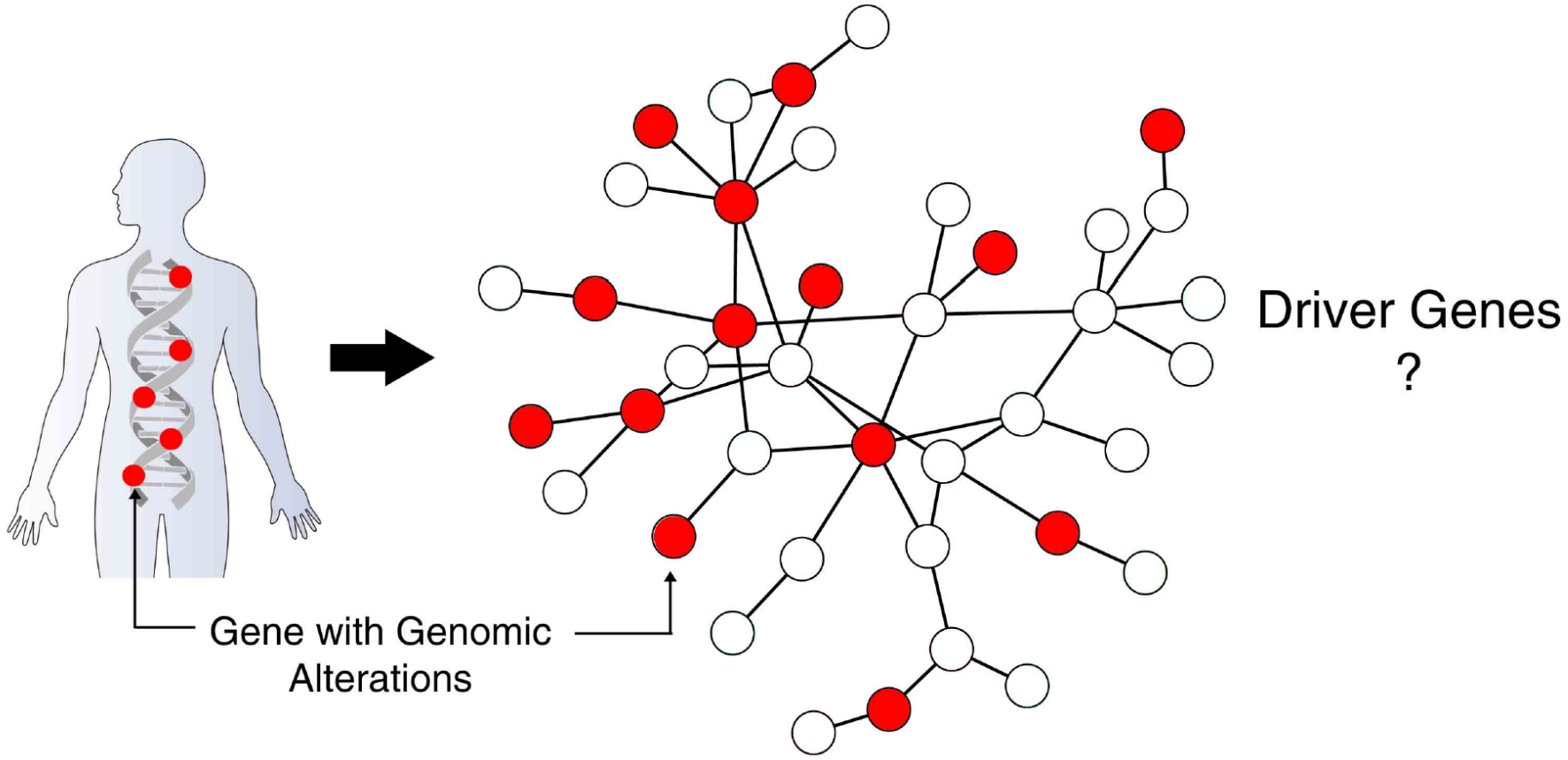
- Most of the alterations are neutral and provide no growth advantage to the tumor: known as "passenger" alterations
- Very few alterations provide evolutionary advantage and are positively selected for during tumorigenesis: known as "driver" alterations

Challenge

- Driver alterations are diluted and are outnumbered by the passenger alterations
- This makes identification of driver alterations more complicated

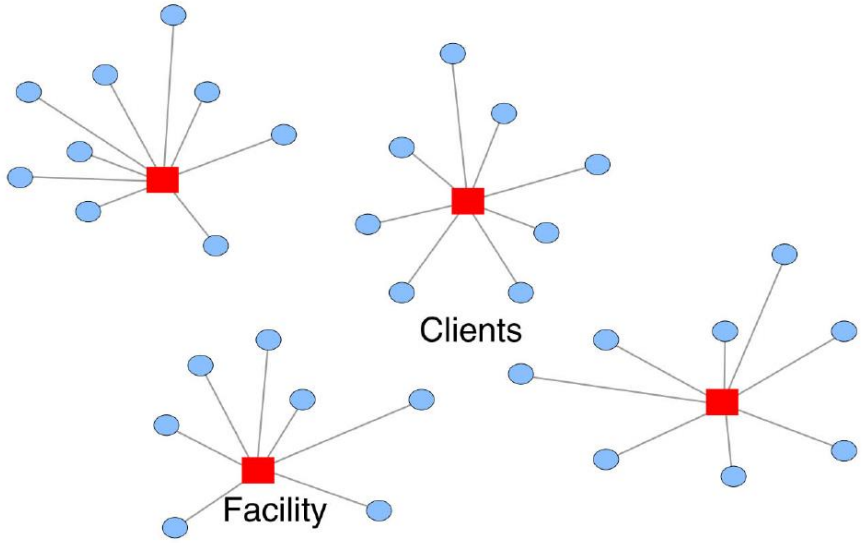
Problem Formulation:

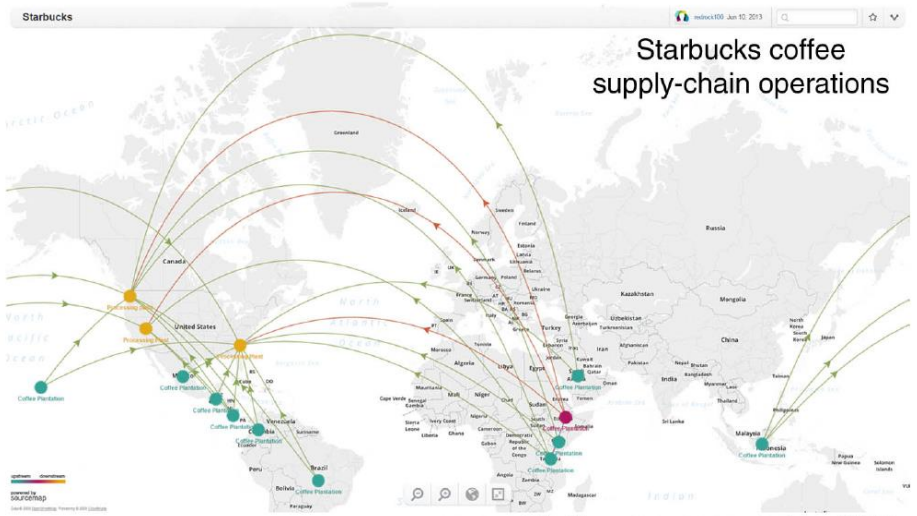
- Given a set of genes with genomic alterations in a tumor, identify the location of the gene(s) in the network that are (most likely) responsible for driving the tumor



Facility location Problem

- Concerned with the **optimal placement of facilities** to **minimize associated costs**

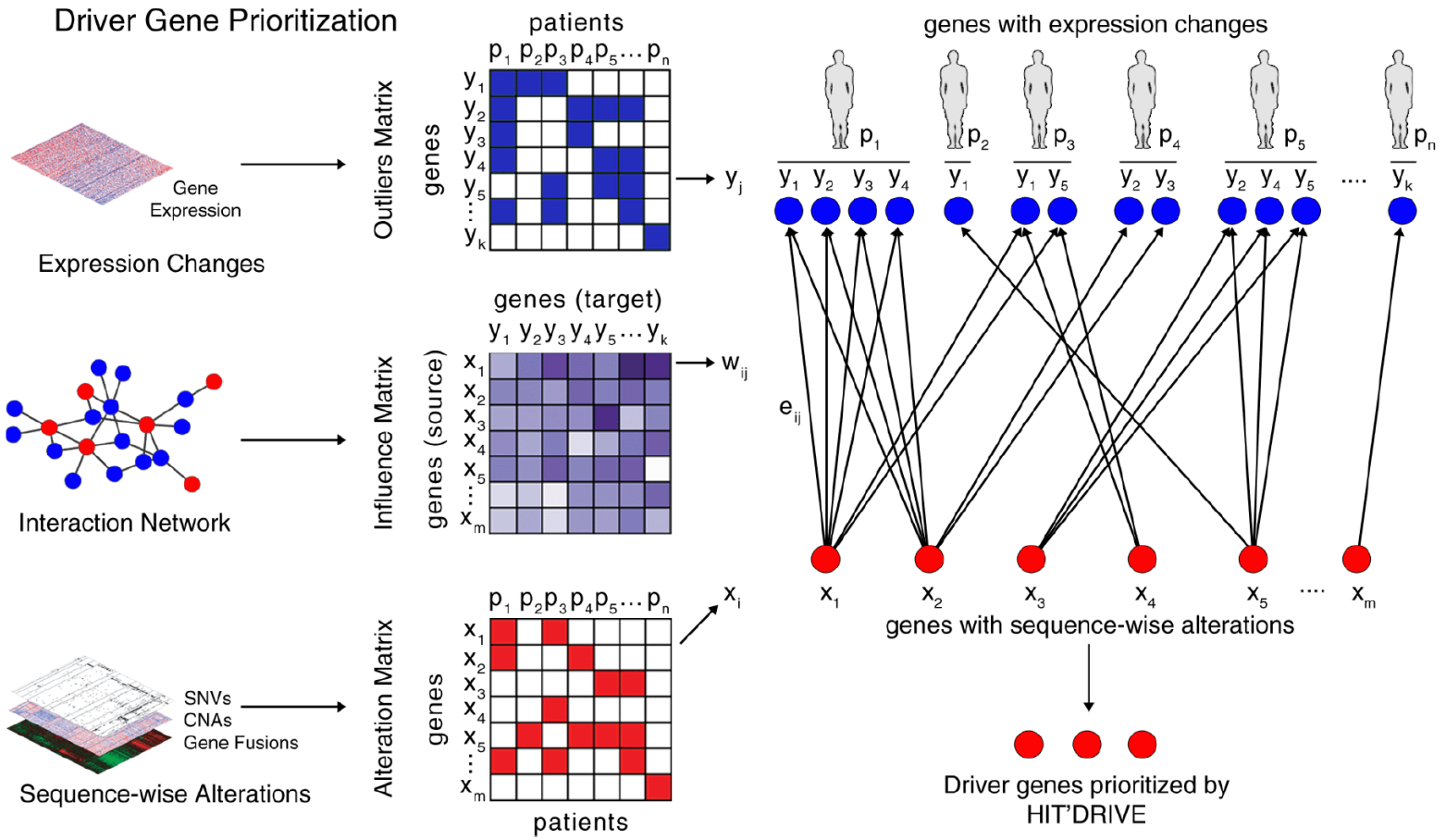




source: http://free.sourcemap.com/view/6039

- This formulation is used to decide
  - Hospital placement
  - Airline hub airport selection
  - Chain restaurant/coffee shop location identification
  - Identification of top influencers in a social network: fake news placement

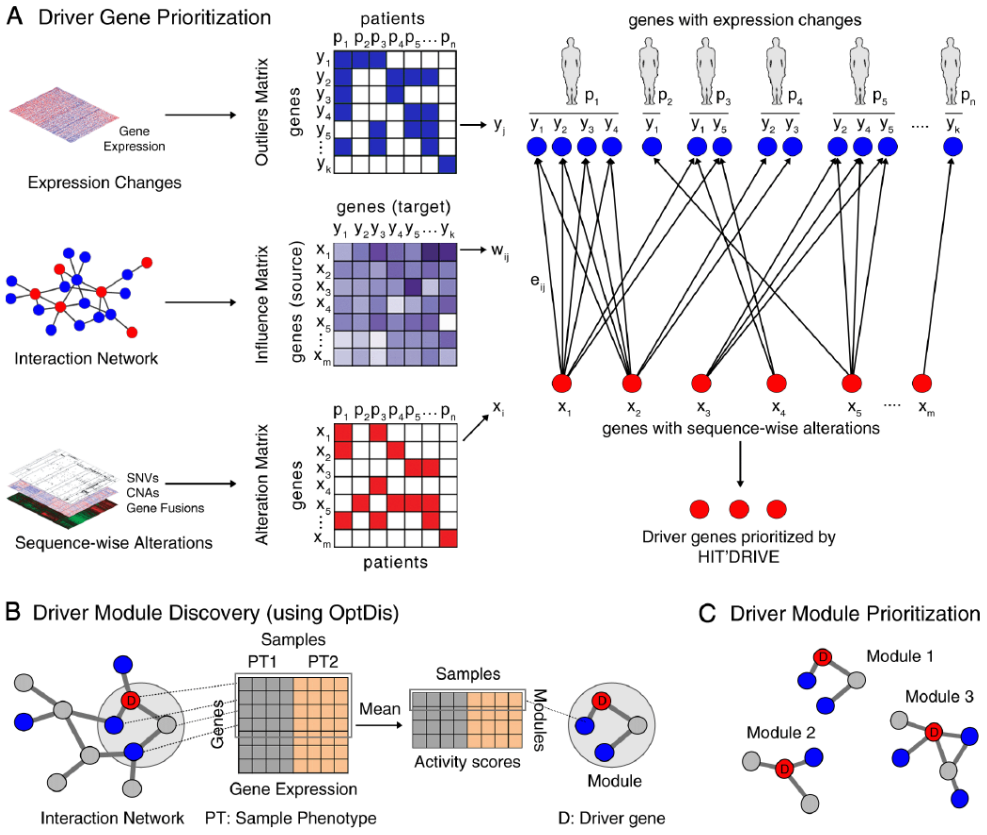
Find the most influential set of genes that potentially driver the cancer



- Find the smallest set of genes with sequence level changes that can explain the changes in gene expression of large fraction of genes
- We solve this using Integer Linear Programming (ILP)

Shrestha et al. 2017 (Genome Research)

Cancer sub-type classification

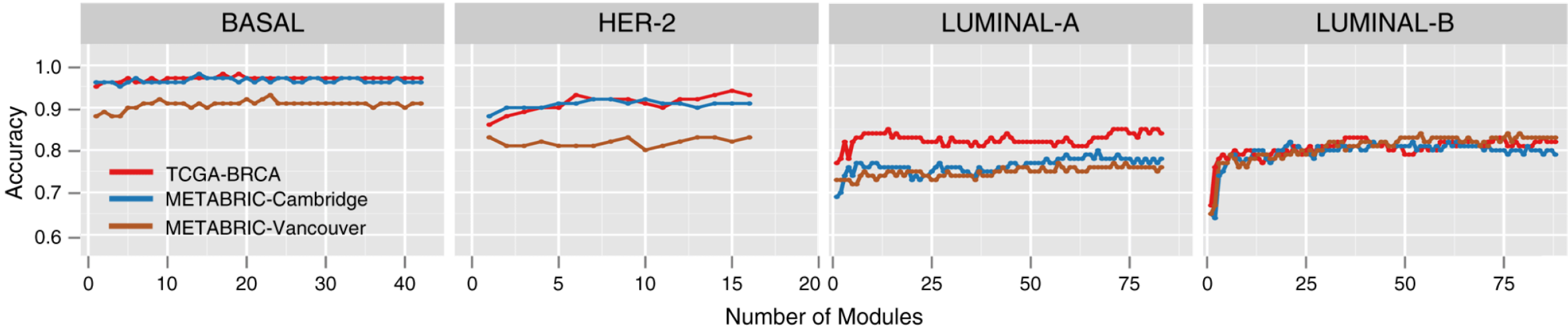


- We used (and improved) OptDis[1] for *de novo* identification of modules inside the interaction network which are seeded by at least one predicted driver
- modules are chosen so that their discriminative power (for phenotype classification) is greatest among connected subnetworks of similar size that contain the individual predicted drivers.

[1] Dao *et al.* Bioinformatics, 2011 (ISMB 2011)

Shrestha et al. 2017 (Genome Research)

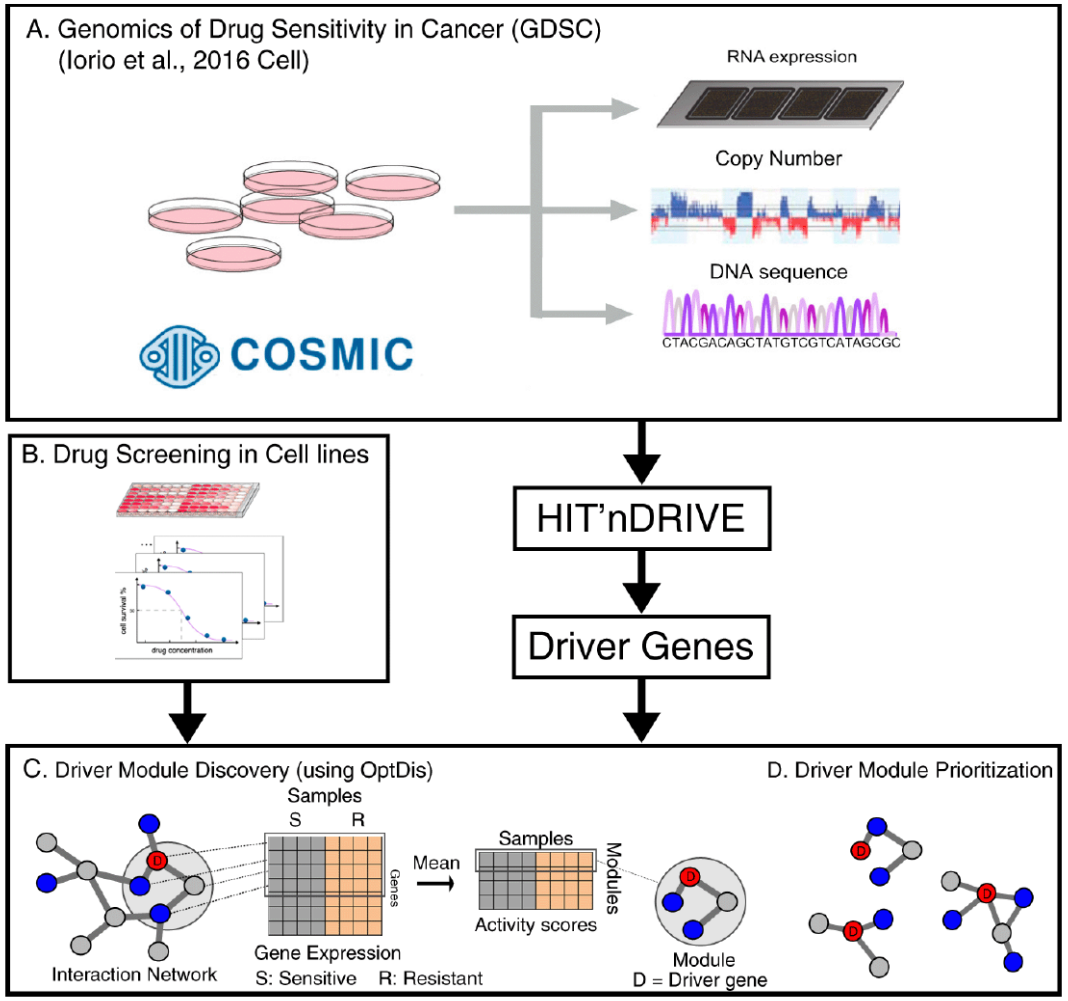
Breast Cancer Subtype classification



- Classified four major subtypes of breast cancer - Basal, HER2, Luminal-A and Luminal-B
- We respectively obtained 37, 16, 43 and 39 subtype specific driver modules for Basal, HER2, Luminal-A and Luminal-B subtypes
- Classified Basal-like tumors with much higher accuracy (98%) as compared to other BRCA-subtypes - HER2 (94%), Luminal-A (85%) and Luminal-B (83%)

Shrestha et al. 2017 (Genome Research)

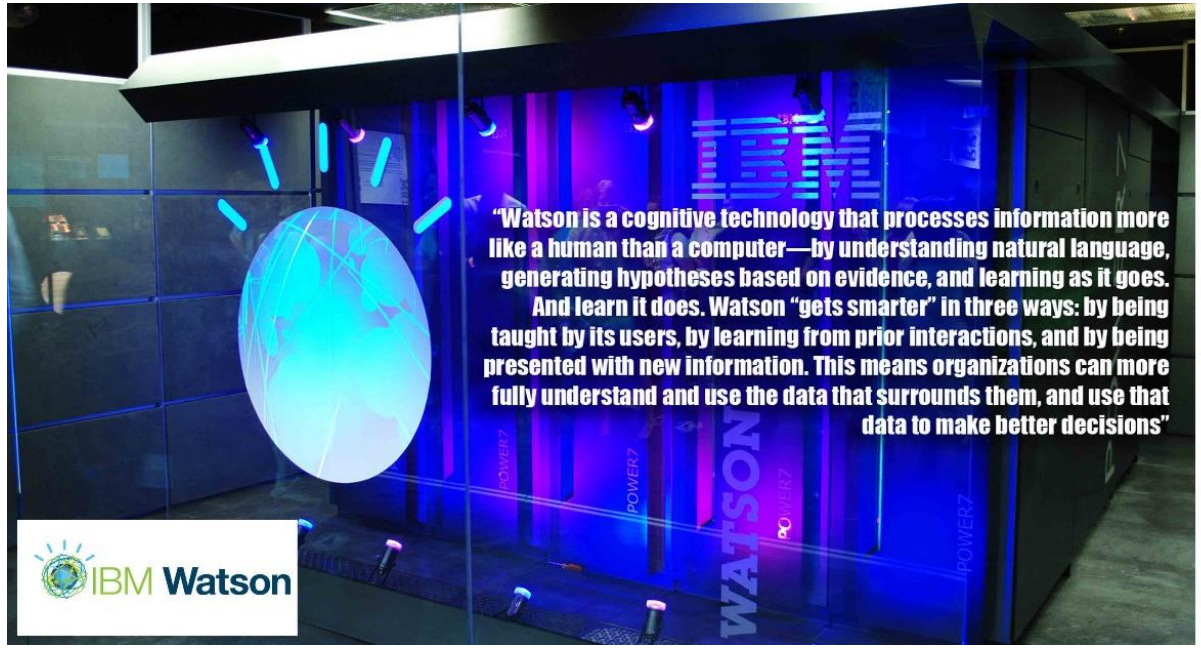
Predict Drug Efficacy

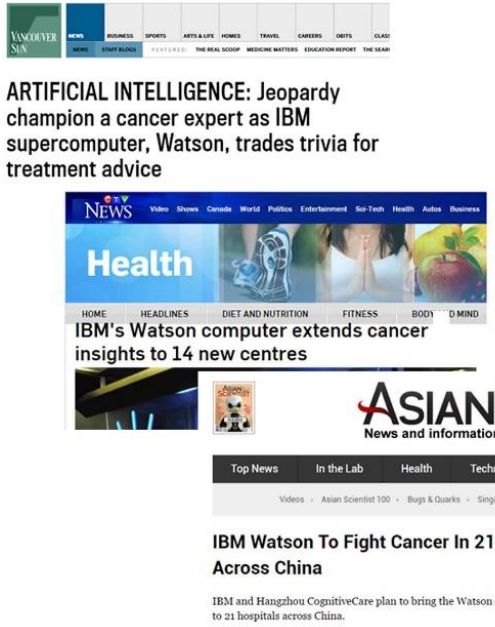


- Pan-Cancer Cell Lines (1001 cell lines, 30 cancer types)
- Treatment with 265 drugs in Pan-Cancer Cell Lines
  - Targeted & cytotoxic drugs
  - FDA approved drugs, drugs in clinical trials & pre-clinical ligands

Shrestha et al. 2017 (Genome Research)

AI Expert systems are beginning to be tested in cancer research centres across the world





Google DeepMind

NEWS ANALYSIS

Did IBM overhype Watson Health's AI promise?

IBM's Watson Health division has been under fire for not delivering on its promise to use AI to enable smarter, more personalized medicine. But IBM officials maintain that hospitals are seeing benefits.

SCIENCE | TECH | HEALTH

IBM's Watson gave unsafe recommendations for treating cancer

Doctors fed it hypothetical scenarios, not real patient data

By Angela Chen | @chengela | Jul 26, 2018, 4:29pm EDT

Emergent Tech ▸ Artificial Intelligence

IBM's Watson Health wing left looking poorly after 'massive' layoffs

Up to 70% of staff shown the door this week, insiders claim

By Iain Thomson in San Francisco 25 May 2018 at 20:08 40 SHARE ▼

IBM has laid off approximately 50 and 70 per cent of staff this week in its Watson Health division, according to inside sources.

Take away message

- Machine Learning is a great approach to tackle a number of problems of biology
- Rise in biological (or genomic) data will certainly require machine learning models to hunt for hidden patterns of disease
- **However, all problems in biology CANNOT be solved using machine learning**
- Try to explore other approaches to model a problem beyond machine learning!