Bioinformatics

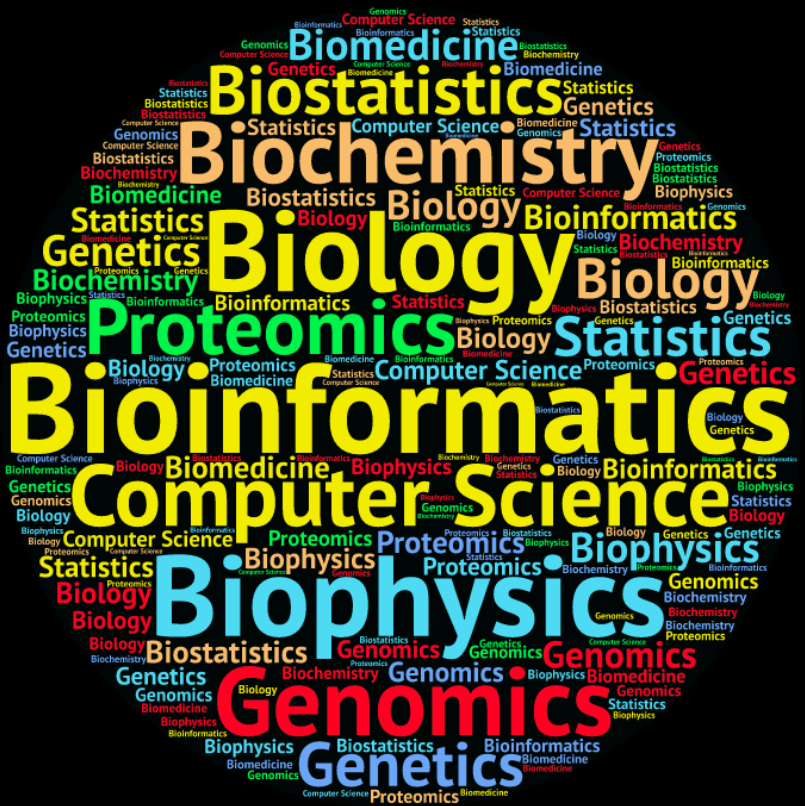
Husen Muhammad Umer, PhD

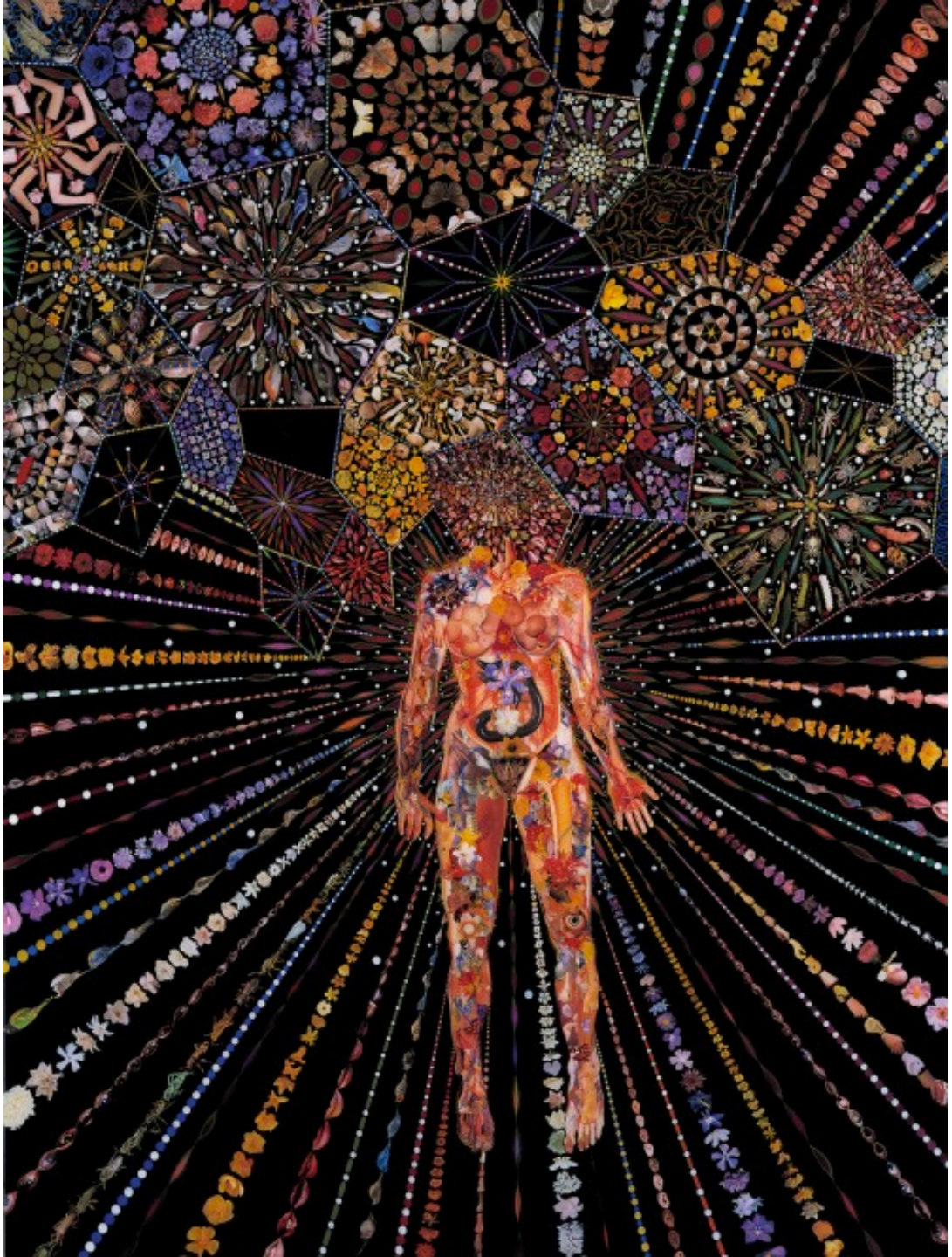
Postdoctoral research & Software engineer

Science For Life Laboratory

Karolinska Institute

15 Nov 2019

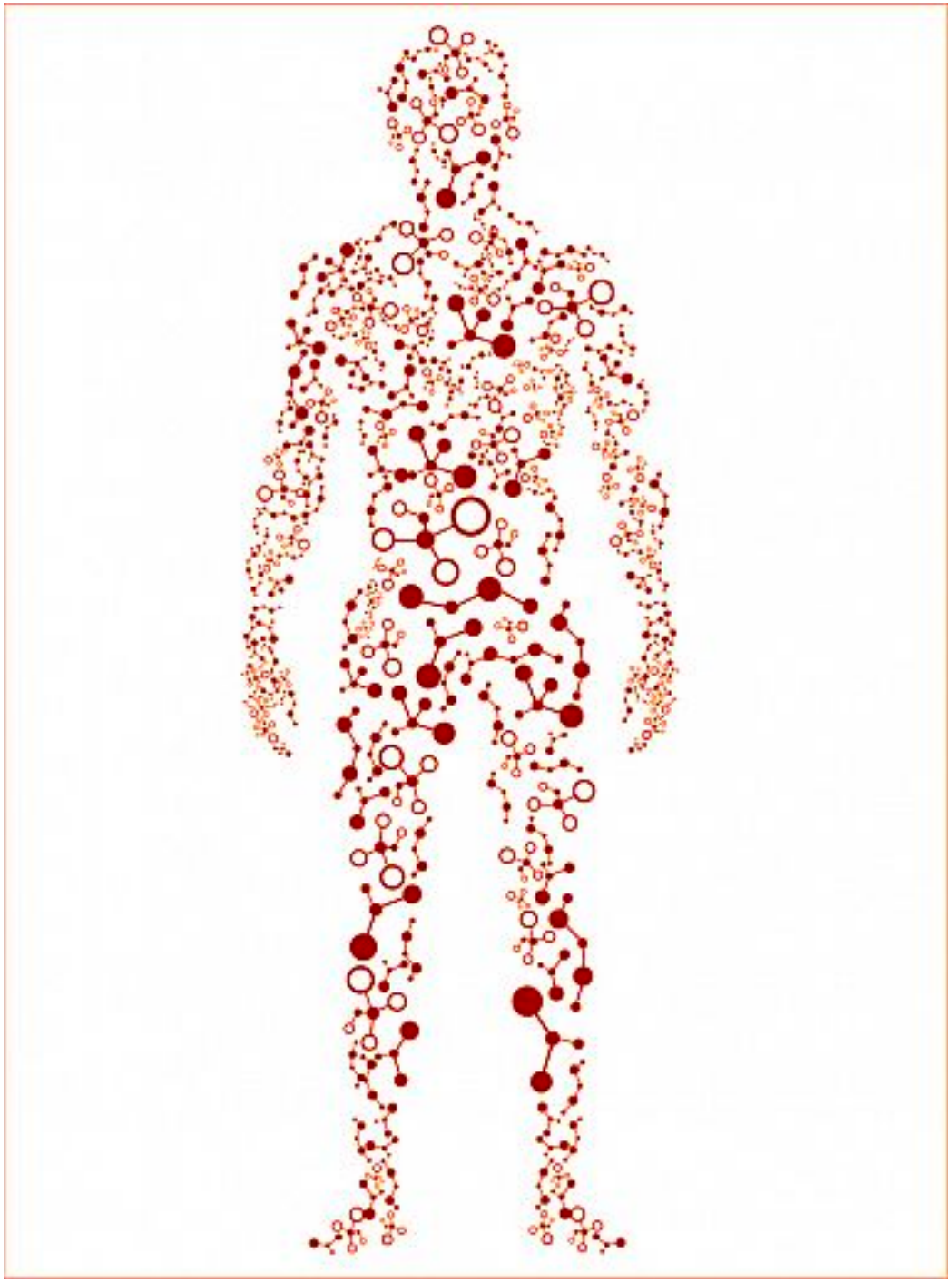

Bioinformatics
Computer Science
Biophysics
Genomics
Biology
Proteomics
Biochemistry
Biostatistics
Genetics
Statistics
Biomedicine

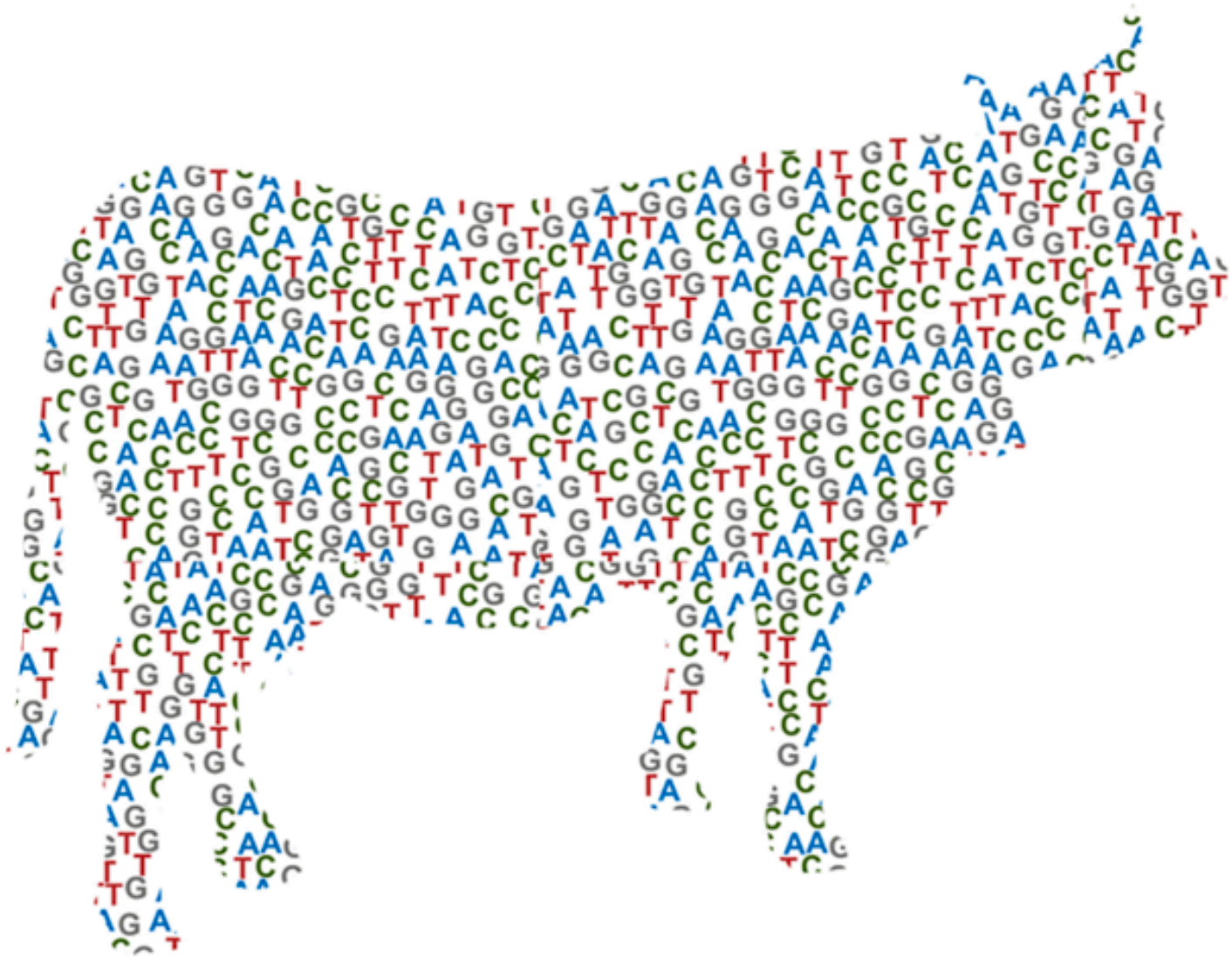


> 30 trillion cells

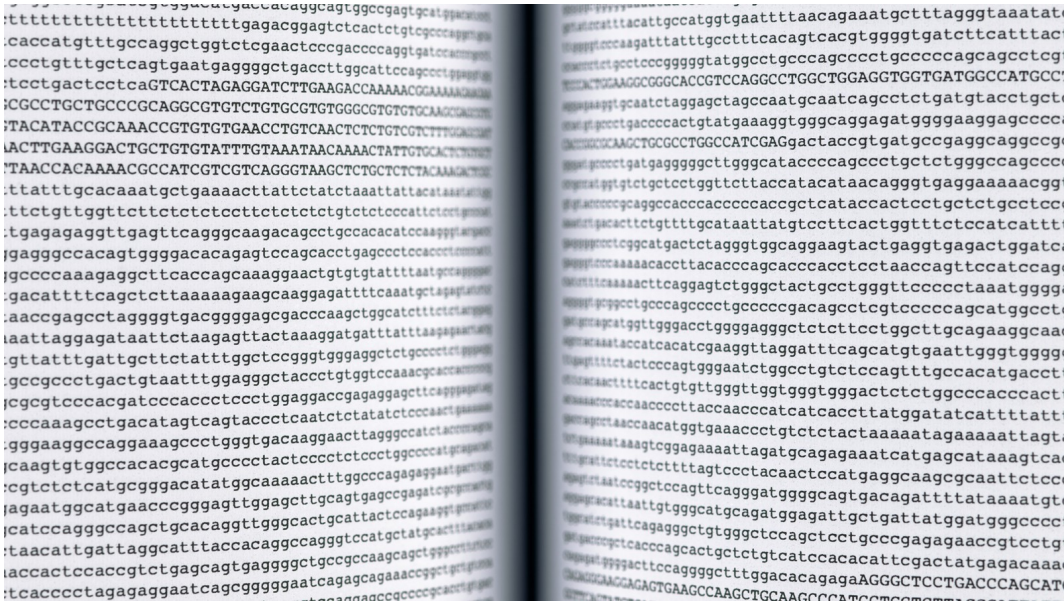
> 200 cell types

By Fred Tomaselli



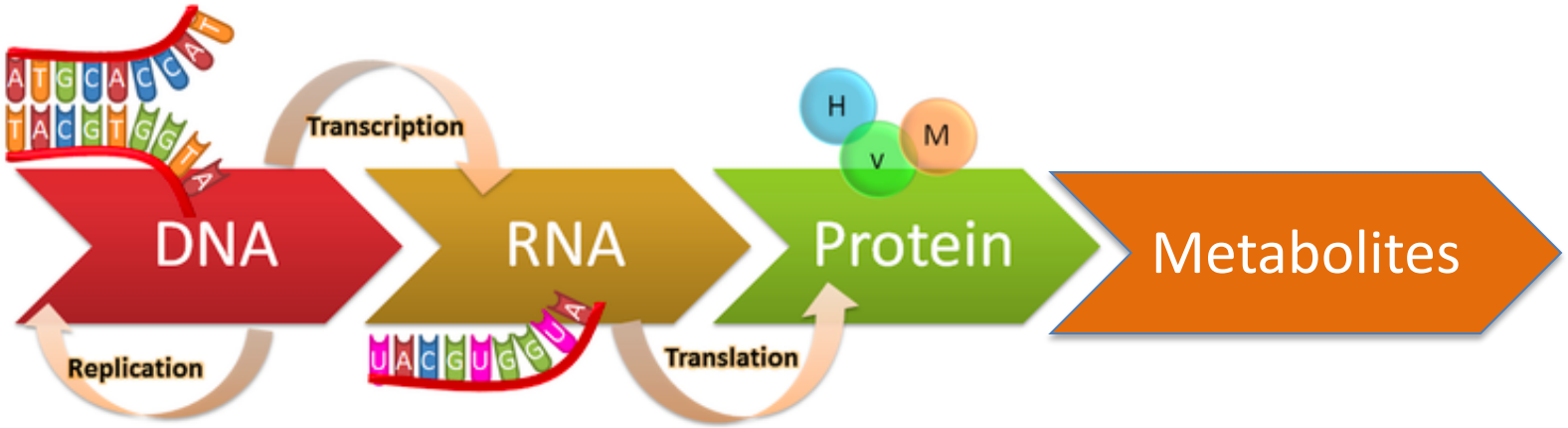


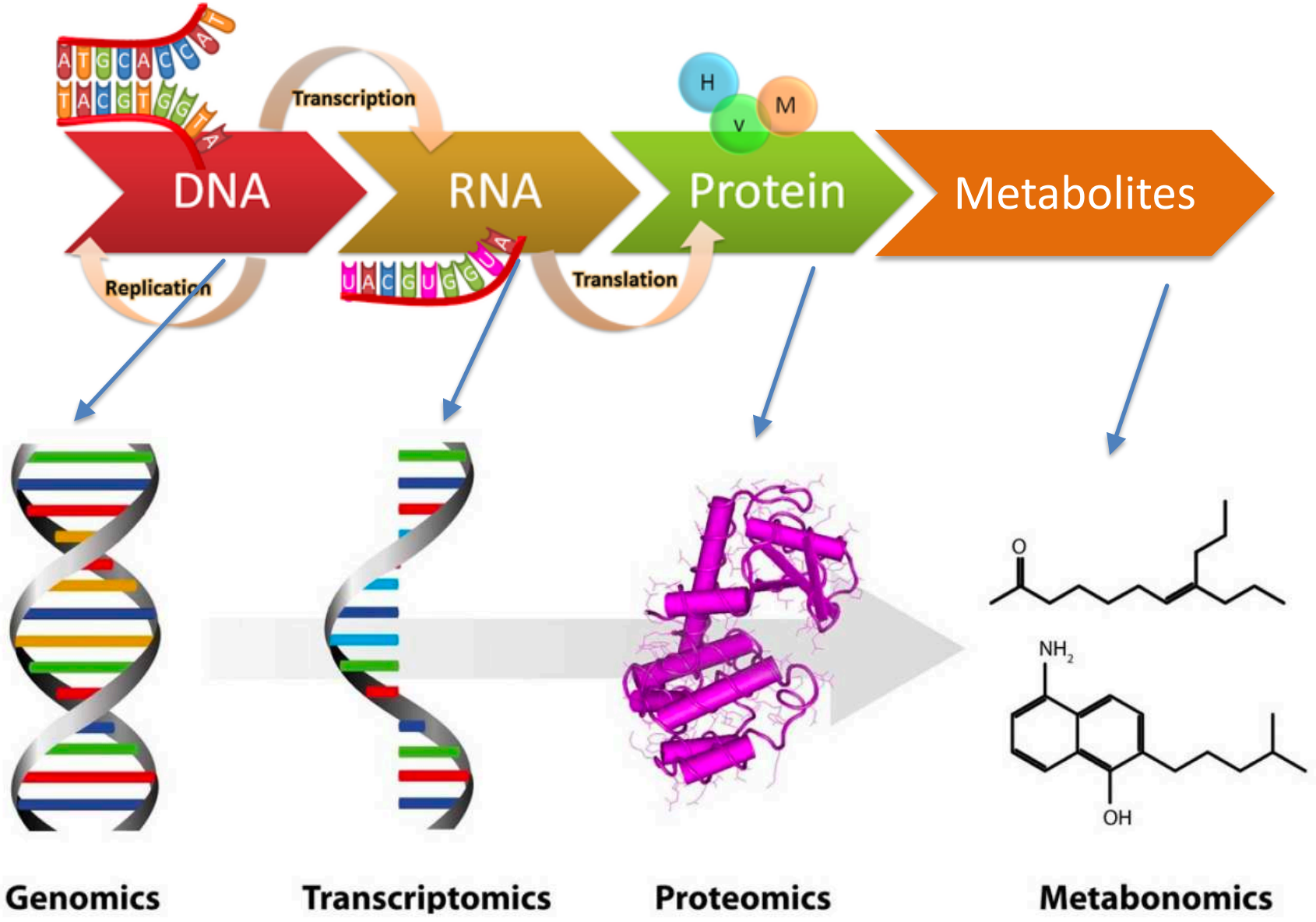
A human genome - printed



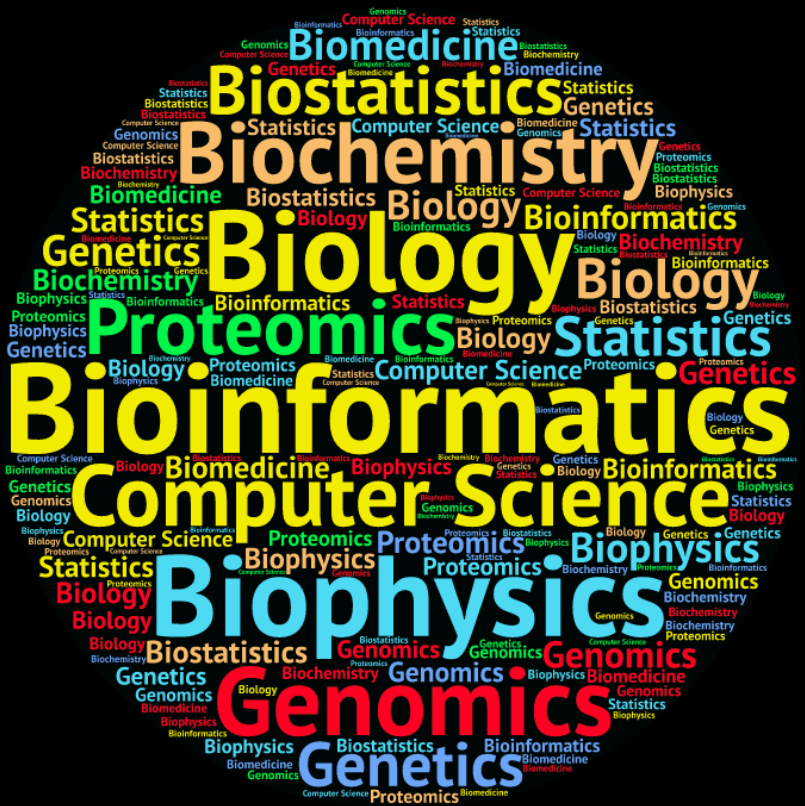
- Human DNA consists of more than 3 billion base pairs (109 books)

Wellcome Collection


DNA
Transcription
RNA
Protein
Metabolites
Replication
Translation
H
M
V

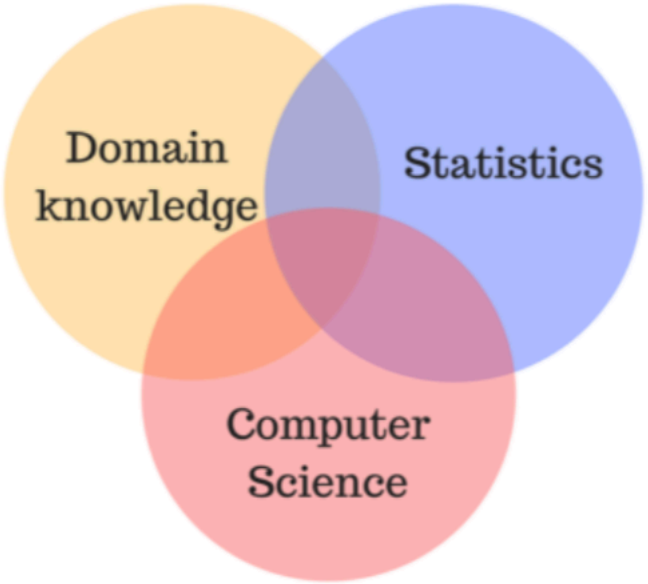


Source: https://www.imperial.ac.uk/clinical-phenotyping-centre


Bioinformatics
Biophysics
Computer Science
Biochemistry
Biology
Proteomics
Biostatistics
Genomics
Genetics
Biomedicine
Statistics

Data Science

Data Science: developing/applying computer programs and statistical methods to analyze and interpret data



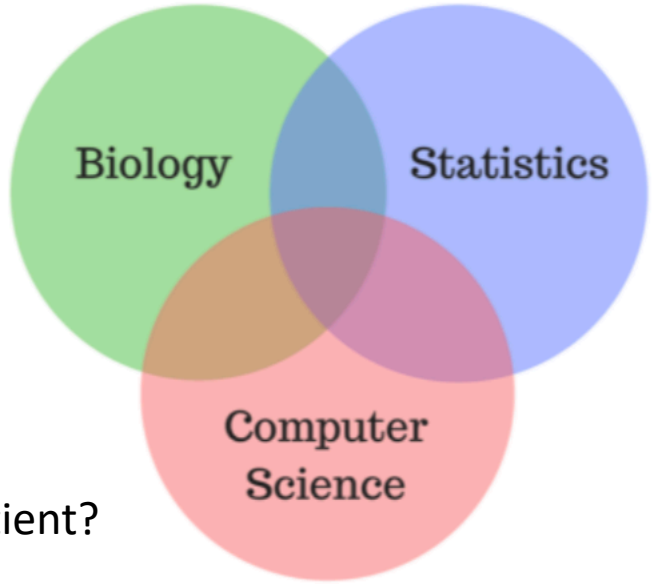
Examples?

Netflix: what movie to make?
Booking.com: where would you go next?
H&M: how much discount to make you buy?

https://www.kaggle.com/

Bioinformatics

Bioinformatics: developing/applying computer programs and statistical methods to analyze and interpret **biological data**



Examples?

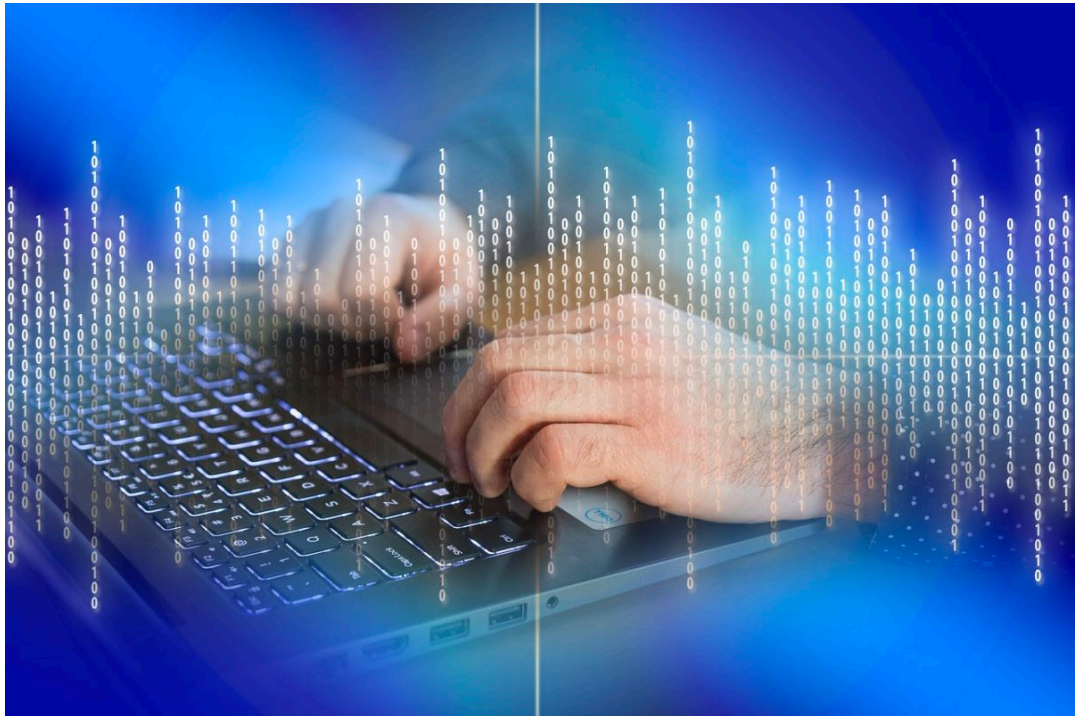
What genetic variant causes disease?

Which protein can be blocked to cure a cancer patient?

Bioinformatics

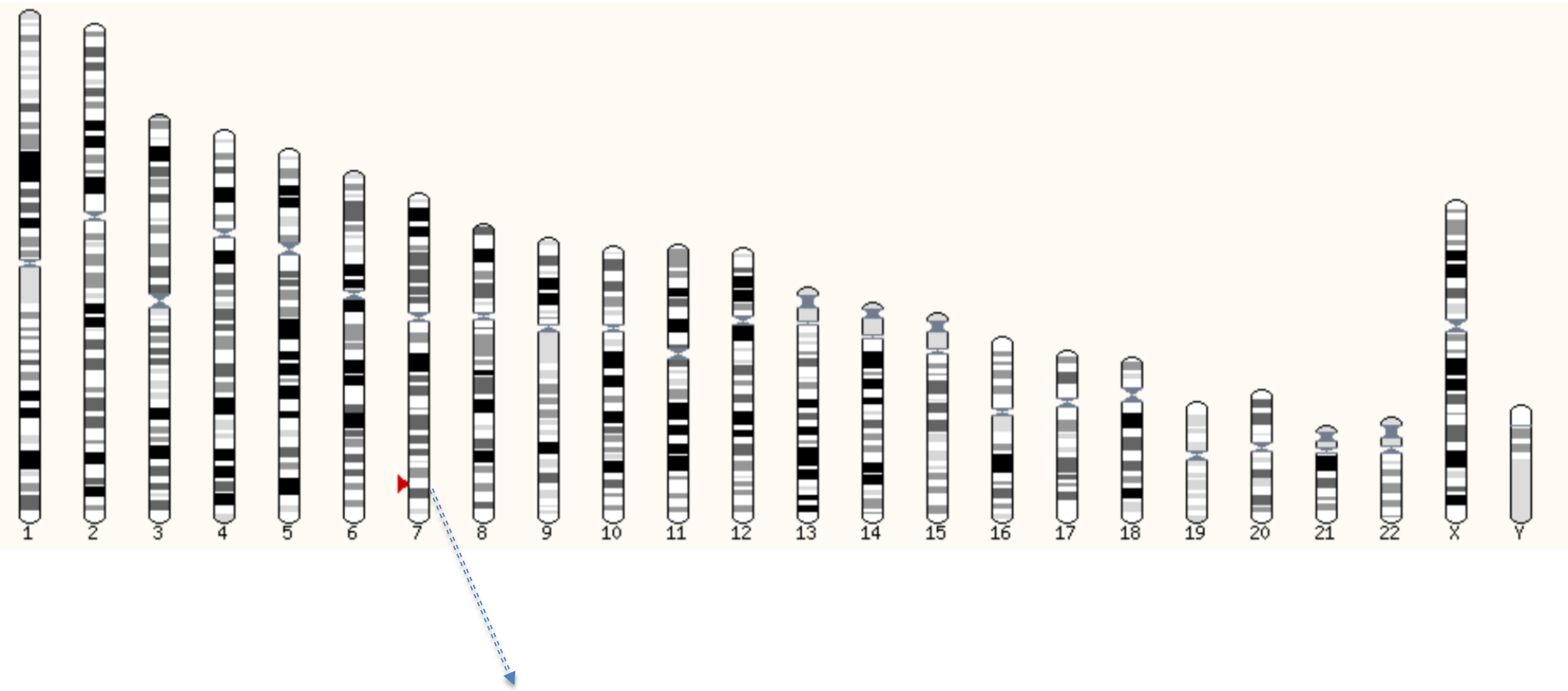
Data

- Metadata
- Sequence
- Structure
- Image
- Spectra



Bioinformatics Applications?

Use case: BRAF mutation in cancer



BRAF on Chromosome 7: 140,719,327-140,924,928

BRAF – cDNA sequence

ATGGCGGCGCTGAGCGGTGGCGGTGGTGGCGGCGCGGAGCCGGGCCAGGCTCTGTTCAACGGGGACATGGAGCCCGAGGCCGGCGCC
GGCGCCGGCGCCGCGGCCTCTTCGGCTGCGGACCCTGCCATTCCGGAGGAGGTGTGGAATATCAAACAAATGATTAAGTTGACACAGGAA
CATATAGAGGCCCTATTGGACAAATTTGGTGGGGAGCATAATCCACCATCAATATATCTGGAGGCCTATGAAGAATACACCAGCAAGCTAGAT
GCACTCCAACAAAGAGAACAACAGTTATTGGAATCTCTGGGGAACGGAACTGATTTTTCTGTTTCTAGCTCTGCATCAATGGATACCGTTACA
TCTTCTTCCTCTTCTAGCCTTTCAGTGCTACCTTCATCTCTTTCAGTTTTTCAAAATCCCACAGATGTGGCACGGAGCAACCCCAAGTCACCAC
AAAAACCTATCGTTAGAGTCTTCCTGCCCAACAAACAGAGGACAGTGGTACCTGCAAGGTGTGGAGTTACAGTCCGAGACAGTCTAAAGAA
AGCACTGATGATGAGAGGTCTAATCCCAGAGTGCTGTGCTGTTTACAGAATTCAGGATGGAGAGAAGAAACCAATTGGTTGGGACACTGAT
ATTTCCTGGCTTACTGGAGAAGAATTGCATGTGGAAGTGTTGGAGAATGTTCCACTTACAACACACAACTTTGTACGAAAAACGTTTTTCAC
CTTAGCATTTTGTGACTTTTGTCGAAAGCTGCTTTTCCAGGGTTTCCGCTGTCAAACATGTGGTTATAAATTTCACCAGCGTTGTAGTACAGA
AGTTCCACTGATGTGTGTTAATTATGACCAACTTGATTTGCTGTTTGTCTCCAAGTTCTTTGAACACCACCCAATACCACAGGAAGAGGCGTC
CTTAGCAGAGACTGCCCTAACATCTGGATCATCCCCTTCCGCACCCGCCTCGGACTCTATTGGGCCCCAAATTCTCACCAGTCCGTCTCCTTCA
AAATCCATTCCAATTCCACAGCCCTTCCGACCAGCAGATGAAGATCATCGAAATCAATTTGGGCAACGAGACCGATCCTCATCAGCTCCCAAT
GTGCATATAAACACAATAGAACCTGTCAATATTGATGACTTGATTAGAGACCAAGGATTTCGTGGTGATGGAGGATCAACCACAGGTTTGTCT
GCTACCCCCCCTGCCTCATTACCTGGCTCACTAACTAACGTGAAAGCCTTACAGAAATCTCCAGGACCTCAGCGAGAAAGGAAGTCATCTTC
ATCCTCAGAAGACAGGAATCGAATGAAAACACTTGGTAGACGGGACTCGAGTGATGATTGGGAGATTCCTGATGGGCAGATTACAGTGGG
ACAAAGAATTGGATCTGGATCATTTGGAACAGTCTACAAGGGAAAGTGGCATGGTGATGTGGCAGTGAAAATGTTGAATGTGACAGCACCT
ACACCTCAGCAGTTACAAGCCTTCAAAAATGAAGTAGGAGTACTCAGGAAAACACGACATGTGAATATCCTACTCTTCATGGGCTATTCCACA
AAGCCACAACTGGCTATTGTTACCCAGTGGTGTGAGGGCTCCAGCTTGTATCACCATCTCCATATCATTGAGACCAAATTTGAGATGATCAAA
CTTATAGATATTGCACGACAGACTGCACAGGGCATGGATTACTTACACGCCAAGTCAATCATCCACAGAGACCTCAAGAGTAATAATATATTTC
TTCATGAAGACCTCACAGTAAAAATAGGTGATTTTGGTCTAGCTACAGTGAAATCTCGATGGAGTGGGTCCCATCAGTTTGAACAGTTGTCT
GGATCCATTTTGTGGATGGCACCAGAAGTCATCAGAATGCAAGATAAAAATCCATACAGCTTTCAGTCAGATGTATATGCATTTGGAATTGTT
CTGTATGAATTGATGACTGGACAGTTACCTTATTCAAACATCAACAACAGGGACCAGATAATTTTTATGGTGGGACGAGGATACCTGTCTCCA
GATCTCAGTAAGGTACGGAGTAACTGTCCAAAAGCCATGAAGAGATTAATGGCAGAGTGCCTCAAAAAGAAAAGAGATGAGAGACCACTC
TTTCCCCAAATTCTCGCCTCTATTGAGCTGCTGGCCCGCTCATTGCCAAAAATTCACCGCAGTGCATCAGAACCCTCCTTGAATCGGGCTGGT
TTCCAAACAGAGGATTTTAGTCTATATGCTTGTGCTTCTCCAAAAACACCCATCCAGGCAGGGGGATATGGTGCGTTTCCTGTCCACTGA

= 2301 nucleotides

BRAF – Protein Sequence

MAALSGGGGGGAEPGQALFNGDMEPEAGAGAGAAASSAADPAIPEEVWNIKQMIKLTQEHIEALLDKFGGEHNPPSIYL
EAYEEYTSKLDALQQREQQLLESLGNGTDFSVSSSASMDTVTSSSSSSLSVLPSSLSVFQNPTDVARSNPKSPQKPIVRVFLPN
KQRTVVPARCGVTVRDSLKKALMMRGLIPECCAVYRIQDGEKKPIGWDTDISWLTGEELHVEVLENVPLTTHNFVRKTFFT
LAFCDFCRKLLFQGFRCQTCGYKFHQRCSTEVPLMCVNYDQLDLLFVSKFFEHHPIPQEEASLAETALTSGSSPSAPASDSIGP
QILTSPSPSKSIPIPQPFRPADEDHRNQFGQRDRSSSAPNVHINTIEPVNIDDLIRDQGFRGDGGSTTGLSATPPASLPGSLTN
VKALQKSPGPQRERKSSSSSEDRNRMKTLGRRDSSDDWEIPDGQITVGQRIGSGSFGTVYKGKWHGDVAVKMLNVTAPT
PQQLQAFKNEVGVLRKTRHVNILLFMGYSTKPQLAIVTQWCEGSSLYHHLHIIETKFEMIKLIDIARQTAQGMDYLHAKSIIH
RDLKSNNIFLHEDLTVKIGDFGLATVKSRWSGSHQFEQLSGSILWMAPEVIRMQDKNPYSFQSDVYAFGIVLYELMTGQLPY
SNINNRDQIIFMVGRGYLSPDLSKVRSNCPKAMKRLMAECLKKKRDERPLFPQILASIELLARSLPKIHRSASEPSLNRAGFQT
EDFSLYACASPKTPIQAGGYGAFPVH*

= 767 nucleotides

BRAF – DNA-Protein Sequence

DNA sequence: ATG GCG GCG CTG AGC GGT GGC GGT GGT GGC GGC GCG GAG CCG GGC CAG GCT CTG TTC AAC
Protein sequence: M A A L S G G G G G G A E P G Q A L F N

GGG GAC ATG GAG CCC GAG GCC GGC GCC GGC GCC GGC GCC GCG GCC TCT TCG GCT GCG GAC
G D M E P E A G A G A G A A A S S A A D

CCT GCC ATT CCG GAG GAG GTG TGG AAT ATC AAA CAA ATG ATT AAG TTG ACA CAG GAA CAT
P A I P E E V W N I K Q M I K L T Q E H

ATA GAG GCC CTA TTG GAC AAA TTT GGT GGG GAG CAT AAT CCA CCA TCA ATA TAT CTG GAG
I E A L L D K F G G E H N P P S I Y L E

GCC TAT GAA GAA TAC ACC AGC AAG CTA GAT GCA CTC CAA CAA AGA GAA CAA CAG TTA TTG
A Y E E Y T S K L D A L Q Q R E Q Q L L

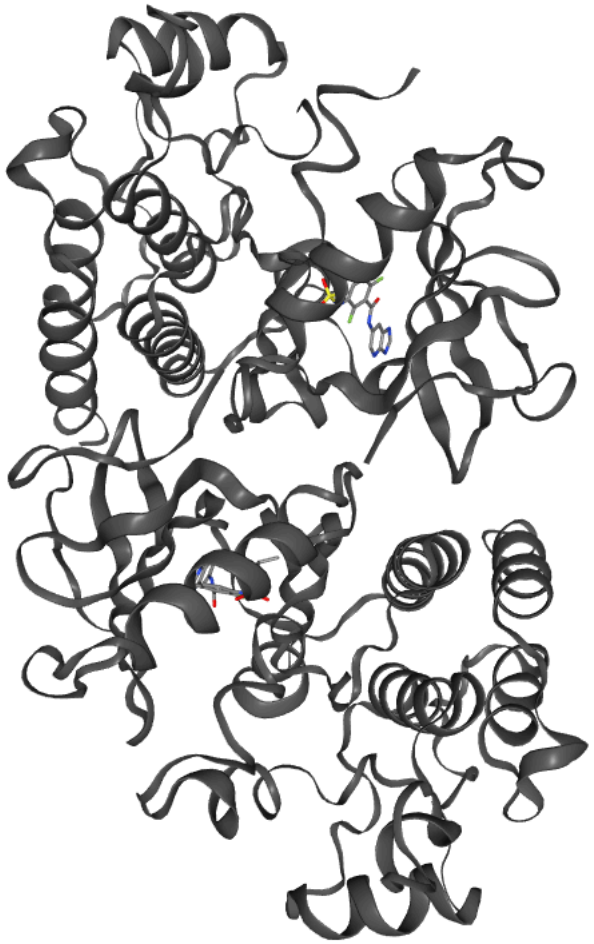
GAA TCT CTG GGG AAC GGA ACT GAT TTT TCT GTT TCT AGC TCT GCA TCA ATG GAT ACC GTT
E S L G N G T D F S V S S S A S M D T V

ACA TCT TCT TCC TCT TCT AGC CTT TCA GTG CTA CCT TCA TCT CTT TCA GTT TTT CAA AAT
T S S S S S S L S V L P S S L S V F Q N

CCC ACA GAT GTG GCA CGG AGC AAC CCC AAG TCA CCA CAA AAA CCT ATC GTT AGA GTC TTC
P T D V A R S N P K S P Q K P I V R V F

CTG CCC AAC AAA CAG AGG ACA GTG GTA CCT GCA AGG TGT GGA GTT ACA GTC CGA GAC AGT
L P N K Q R T V V P A R C G V T V R D S

BRAF - Protein



B-Raf proto-oncogene, serine/threonine kinase

Function:
Sends signals inside the cell to control cell growth

BRAF – cDNA sequence in cancer patients

ATGGCGGCGCTGAGCGGTGGCGGTGGTGGCGGCGCGGAGCCGGGCCAGGCTCTGTTCAACGGGGACATGGAGCCCGAGGCCGGCGCC
GGCGCCGGCGCCGCGGCCTCTTCGGCTGCGGACCCTGCCATTCCGGAGGAGGTGTGGAATATCAAACAAATGATTAAGTTGACACAGGAA
CATATAGAGGCCCTATTGGACAAATTTGGTGGGGAGCATAATCCACCATCAATATATCTGGAGGCCTATGAAGAATACACCAGCAAGCTAGAT
GCACTCCAACAAAGAGAACAACAGTTATTGGAATCTCTGGGGAACGGAACTGATTTTTCTGTTTCTAGCTCTGCATCAATGGATACCGTTACA
TCTTCTTCCTCTTCTAGCCTTTCAGTGCTACCTTCATCTCTTTCAGTTTTTCAAAATCCCACAGATGTGGCACGGAGCAACCCCAAGTCACCAC
AAAAACCTATCGTTAGAGTCTTCCTGCCCAACAAACAGAGGACAGTGGTACCTGCAAGGTGTGGAGTTACAGTCCGAGACAGTCTAAAGAA
AGCACTGATGATGAGAGGTCTAATCCCAGAGTGCTGTGCTGTTTACAGAATTCAGGATGGAGAGAAGAAACCAATTGGTTGGGACACTGAT
ATTTCCTGGCTTACTGGAGAAGAATTGCATGTGGAAGTGTTGGAGAATGTTCCACTTACAACACACAACTTTGTACGAAAAACGTTTTTCAC
CTTAGCATTTTGTGACTTTTGTCGAAAGCTGCTTTTCCAGGGTTTCCGCTGTCAAACATGTGGTTATAAATTTCACCAGCGTTGTAGTACAGA
AGTTCCACTGATGTGTGTTAATTATGACCAACTTGATTTGCTGTTTGTCTCCAAGTTCTTTGAACACCACCCAATACCACAGGAAGAGGCGTC
CTTAGCAGAGACTGCCCTAACATCTGGATCATCCCCTTCCGCACCCGCCTCGGACTCTATTGGGCCCCAAATTCTCACCAGTCCGTCTCCTTCA
AAATCCATTCCAATTCCACAGCCCTTCCGACCAGCAGATGAAGATCATCGAAATCAATTTGGGCAACGAGACCGATCCTCATCAGCTCCCAAT
GTGCATATAAACACAATAGAACCTGTCAATATTGATGACTTGATTAGAGACCAAGGATTTCGTGGTGATGGAGGATCAACCACAGGTTTGTCT
GCTACCCCCCCTGCCTCATTACCTGGCTCACTAACTAACGTGAAAGCCTTACAGAAATCTCCAGGACCTCAGCGAGAAAGGAAGTCATCTTC
ATCCTCAGAAGACAGGAATCGAATGAAAACACTTGGTAGACGGGACTCGAGTGATGATTGGGAGATTCCTGATGGGCAGATTACAGTGGG
ACAAAGAATTGGATCTGGATCATTTGGAACAGTCTACAAGGGAAAGTGGCATGGTGATGTGGCAGTGAAAATGTTGAATGTGACAGCACCT
ACACCTCAGCAGTTACAAGCCTTCAAAAATGAAGTAGGAGTACTCAGGAAAACACGACATGTGAATATCCTACTCTTCATGGGCTATTCCACA
AAGCCACAACTGGCTATTGTTACCCAGTGGTGTGAGGGCTCCAGCTTGTATCACCATCTCCATATCATTGAGACCAAATTTGAGATGATCAAA
CTTATAGATATTGCACGACAGACTGCACAGGGCATGGATTACTTACACGCCAAGTCAATCATCCACAGAGACCTCAAGAGTAATAATATATTTC
TTCATGAAGACCTCACAGTAAAAATAGGTGATTTTGGTCTAGCTACAGAGAAATCTCGATGGAGTGGGTCCCATCAGTTTGAACAGTTGTCT
GGATCCATTTTGTGGATGGCACCAGAAGTCATCAGAATGCAAGATAAAAATCCATACAGCTTTCAGTCAGATGTATATGCATTTGGAATTGTT
CTGTATGAATTGATGACTGGACAGTTACCTTATTCAAACATCAACAACAGGGACCAGATAATTTTTATGGTGGGACGAGGATACCTGTCTCCA
GATCTCAGTAAGGTACGGAGTAACTGTCCAAAAGCCATGAAGAGATTAATGGCAGAGTGCCTCAAAAAGAAAAGAGATGAGAGACCACTC
TTTCCCCAAATTCTCGCCTCTATTGAGCTGCTGGCCCGCTCATTGCCAAAAATTCACCGCAGTGCATCAGAACCCTCCTTGAATCGGGCTGGT
TTCCAAACAGAGGATTTTAGTCTATATGCTTGTGCTTCTCCAAAAACACCCATCCAGGCAGGGGGATATGGTGCGTTTCCTGTCCACTGA

= 2301 nucleotides

BRAF – Protein Sequence in cancer patients

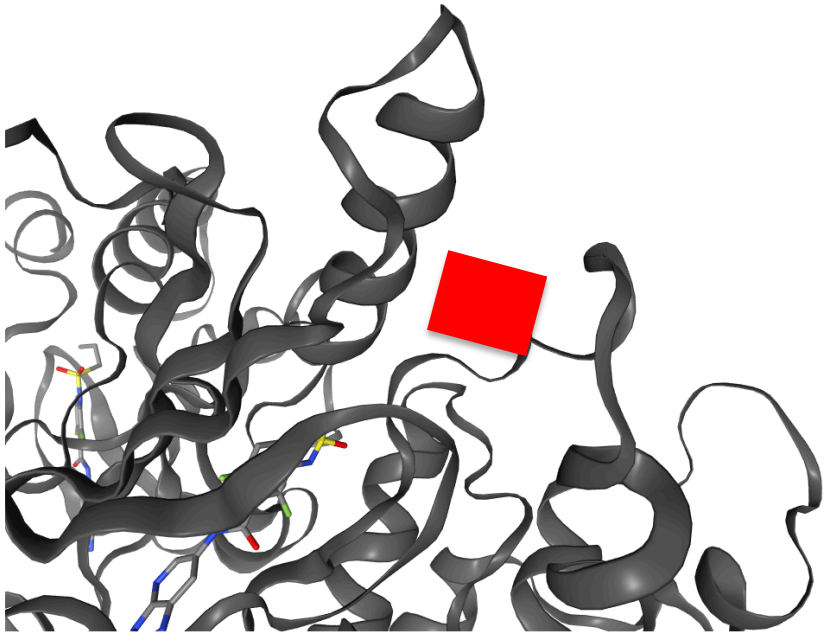
MAALSGGGGGGAEPGQALFNGDMEPEAGAGAGAAASSAADPAIPEEVWNIKQMIKLTQEHIEALLDKFGGEHNPPSIYL
EAYEEYTSKLDALQQREQQLLESLGNGTDFSVSSSASMDTVTSSSSSSLSVLPSSLSVFQNPTDVARSNPKSPQKPIVRVFLPN
KQRTVVPARCGVTVRDSLKKALMMRGLIPECCAVYRIQDGEKKPIGWDTDISWLTGEELHVEVLENVPLTTHNFVRKTFFT
LAFCDFCRKLLFQGFRCQTCGYKFHQRCSTEVPLMCVNYDQLDLLFVSKFFEHHPIPQEEASLAETALTSGSSPSAPASDSIGP
QILTSPSPSKSIPIPQPFRPADEDHRNQFGQRDRSSSAPNVHINTIEPVNIDDLIRDQGFRGDGGSTTGLSATPPASLPGSLTN
VKALQKSPGPQRERKSSSSSEDRNRMKTLGRRDSSDDWEIPDGQITVGQRIGSGSFGTVYKGKWHGDVAVKMLNVTAPT
PQQLQAFKNEVGVLRKTRHVNILLFMGYSTKPQLAIVTQWCEGSSLYHHLHIIETKFEMIKLIDIARQTAQGMDYLHAKSIIH
RDLKSNNIFLHEDLTVKIGDFGLATMVKSRWSGSHQFEQLSGSILWMAPEVIRMQDKNPYSFQSDVYAFGIVLYELMTGQL
PYSNINNRDQIIFMVGRGYLSPDLSKVRSNCPKAMKRLMAECLKKKRDERPLFPQILASIELLARSLPKIHRSASEPSLNRAGF
QTEDFSLYACASPKTPIQAGGYGAFPVH*

= 767 nucleotides

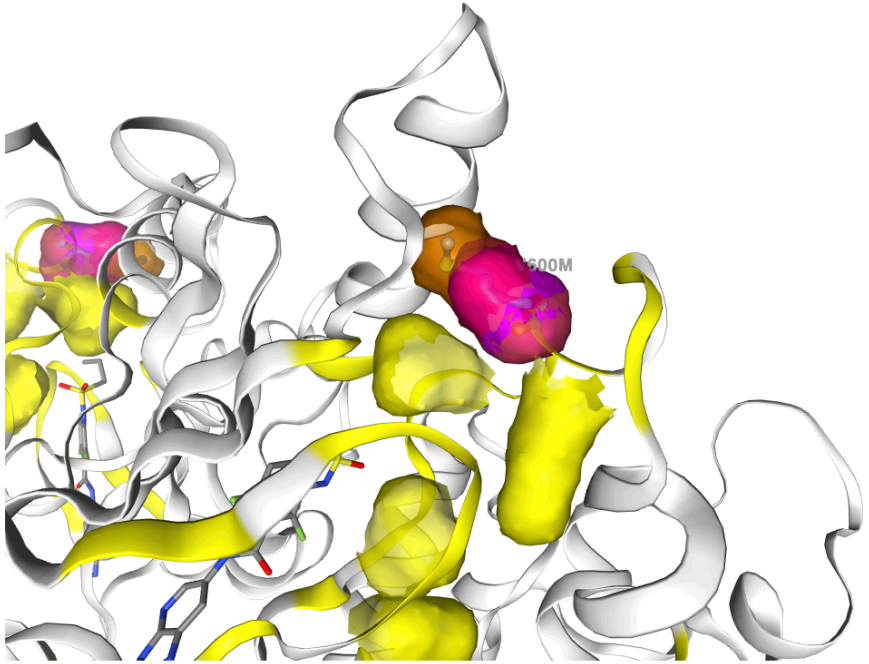
Valine (V)
Mass = 99 Da

Methionine (M)
Mass = 131 Da

BRAF - Protein in cancer patients



Valine (V)
Mass = 99 Da



Methionine (M)
Mass = 131 Da

Solution to BRAF mutations?

- **Targeted thereapy**

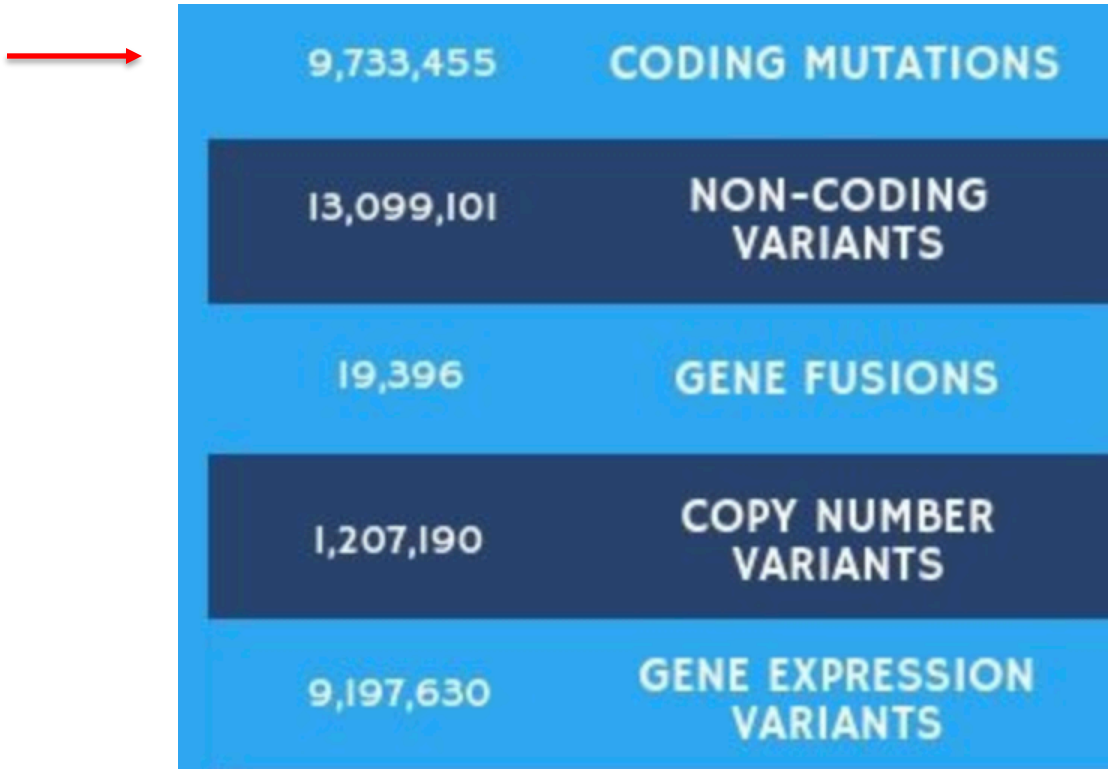
- BRAF inhibitors:
  - Vemurafenib (Zelboraf), dabrafenib (Tafinlar), and encorafenib (Braftovi) are drugs that attack the BRAF protein directly.

- Inhibitions of BRAF's partners?
  - trametinib (Mekinist), cobimetinib (Cotellic), and binimetinib (Mektovi) Inhibit MEK which is a BRAF partner

Solution to all cancers?

- BRAF is mutated in some cancer patients especially in melanoma (skin cancers)
- Other patients?

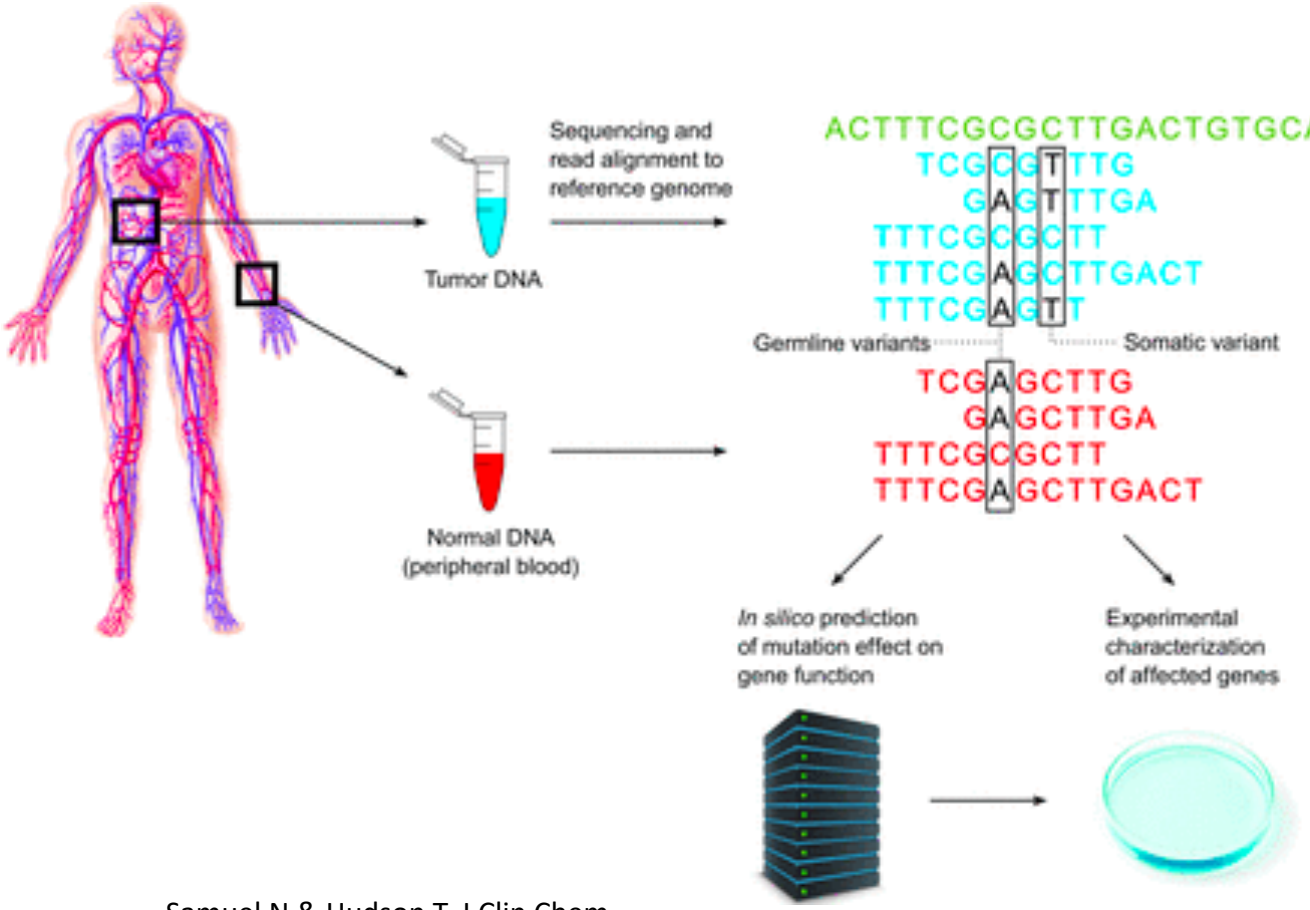
Why computer algorithms & Bioinformatics?



COSMIC release v90

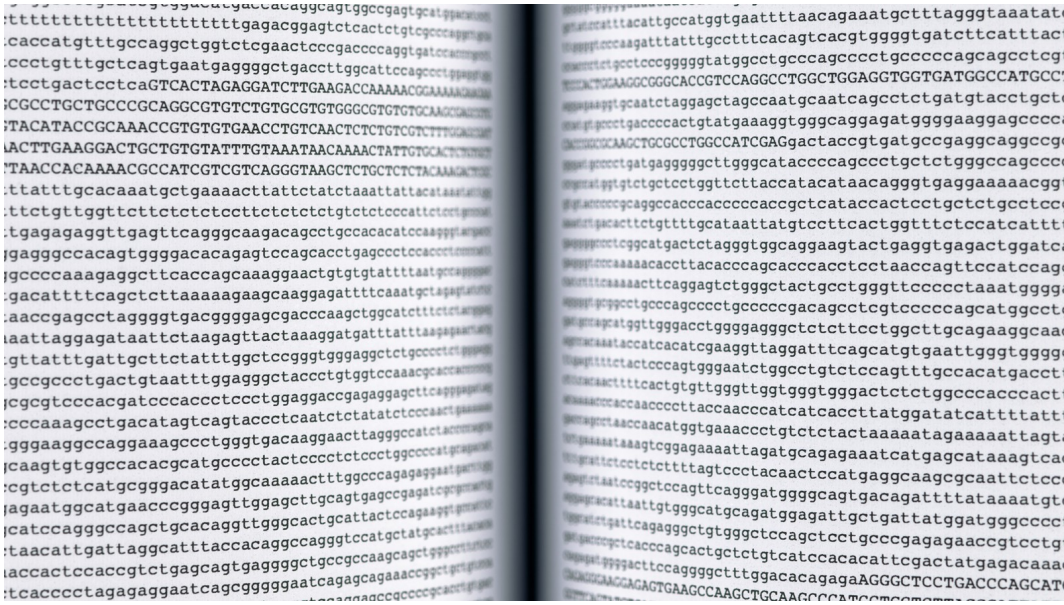
Approach 1 - Genomics

Sequence a normal cells and a cancer cells from the same patient then compare them to find the mutations.



Samuel N & Hudson T. J Clin Chem

A human genome - printed



- Human DNA consists of more than 3 billion base pairs (109 books)

Wellcome Collection

Approach 1 - Genomics

- How to compare 3 billion letters from normal cells to 3 billion letters in cancer cells?
- Which mutations will be the ones that impact the cancer?
- What drug can be made/used to handle the mutated protein?
  - Start diving into bioinformatics to find out

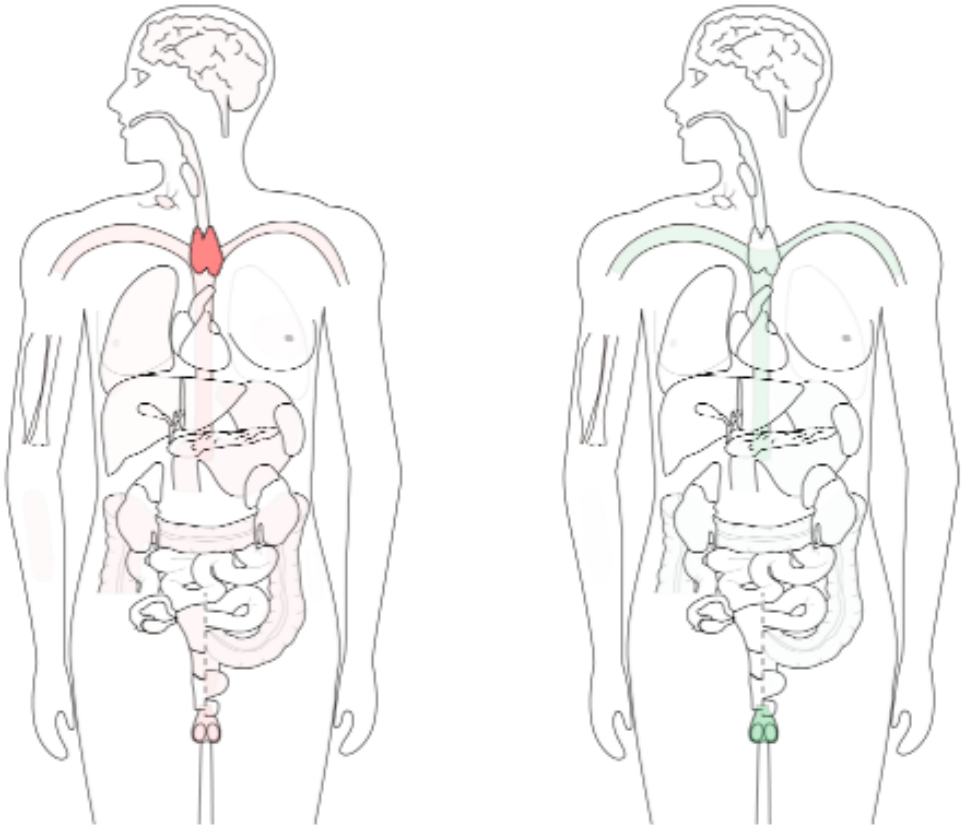
Approach 2 – Transcriptomics & Proteomics

- Transcriptomics: measure transcripts for each gene (RNA-sequencing or Mass-spectrometry Proteomics).
- Check which genes differ between normal cells and cancer cells

- More than 20,000 genes! How to compare?
  - Start diving into bioinformatics to find out

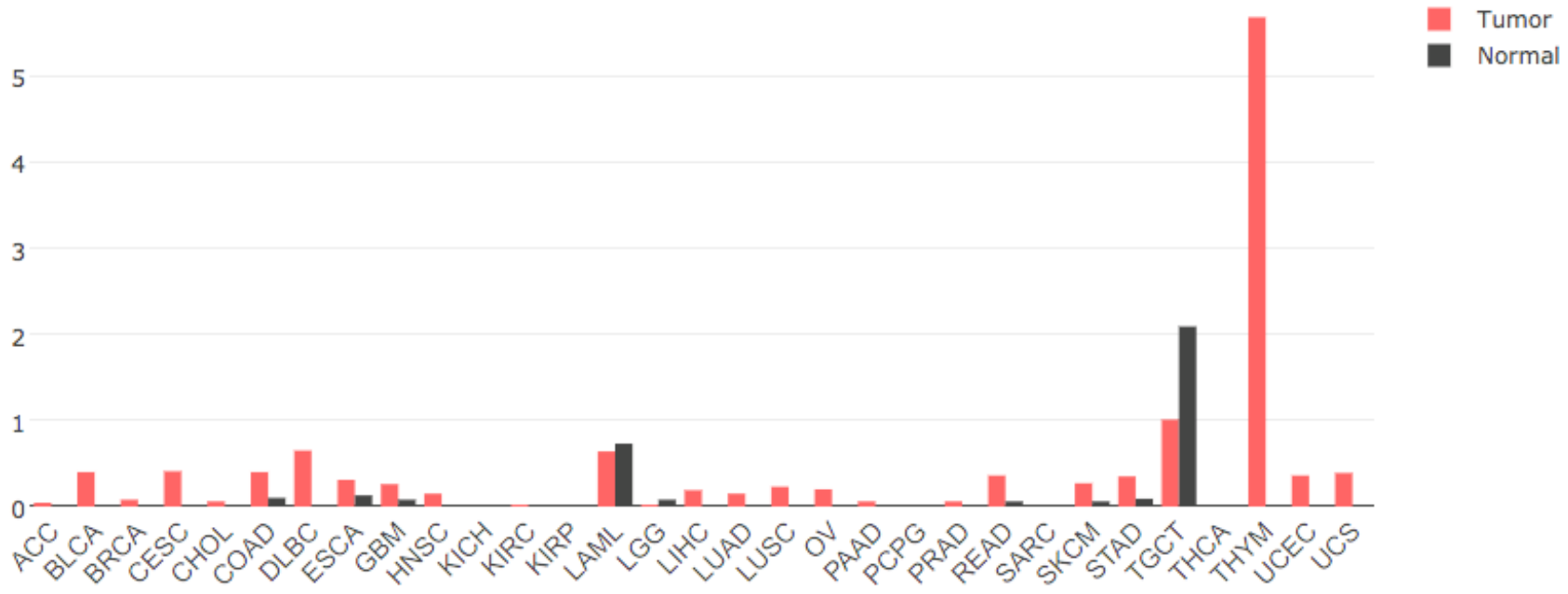
Approaches

TERT over-expression
in Thymoma cancers

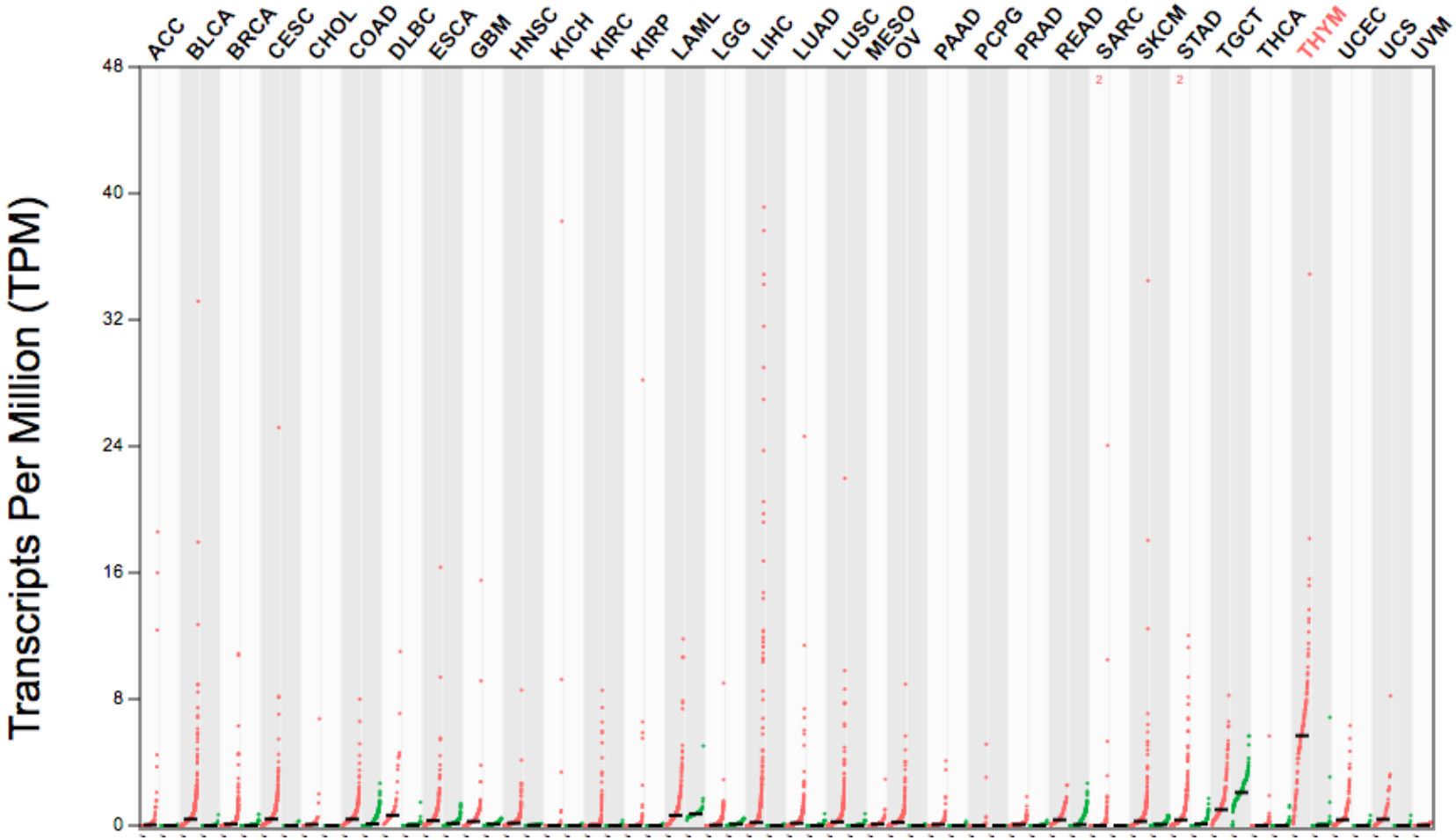


http://gepia.cancer-pku.cn/detail.php?gene=TERT

Approach 2


Tumor
Normal
5
4
3
2
1
0
ACC
BLCA
BRCA
CESC
CHOL
COAD
DLBC
ESCA
GBM
HNSC
KICH
KIRC
KIRP
LAML
LGG
LIHC
LUAD
LUSC
OV
PAAD
PCPG
PRAD
READ
SARC
SKCM
STAD
TGCT
THCA
THYM
UCEC
UCS

Approach 2

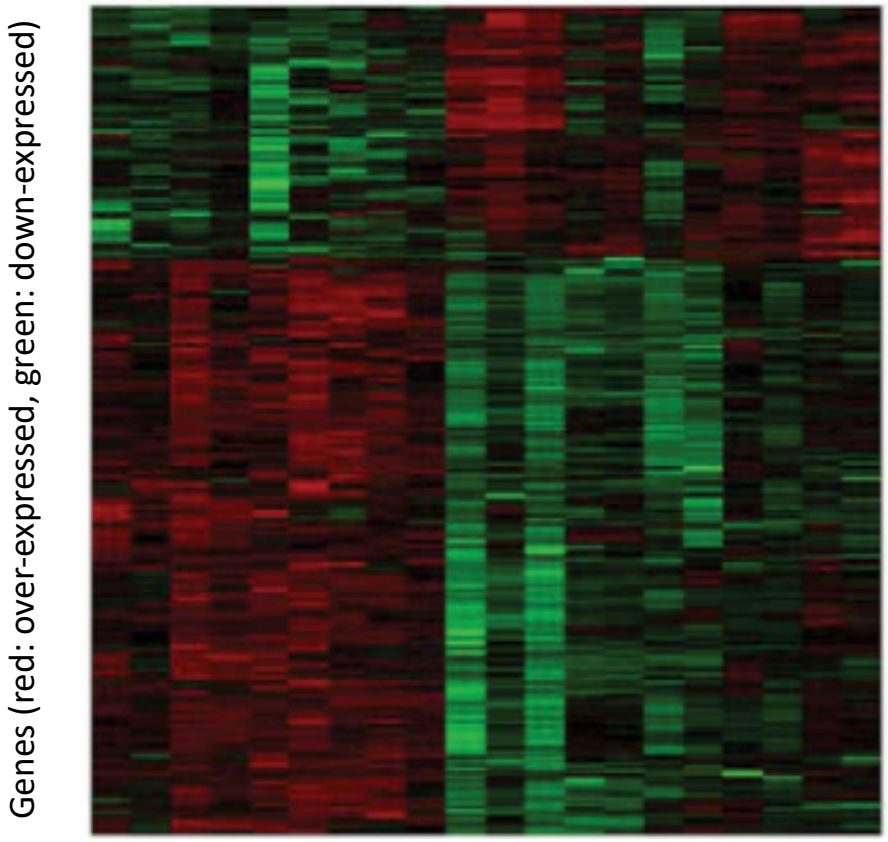


Over-expression of TERT

http://gepia.cancer-pku.cn/detail.php?gene=TERT

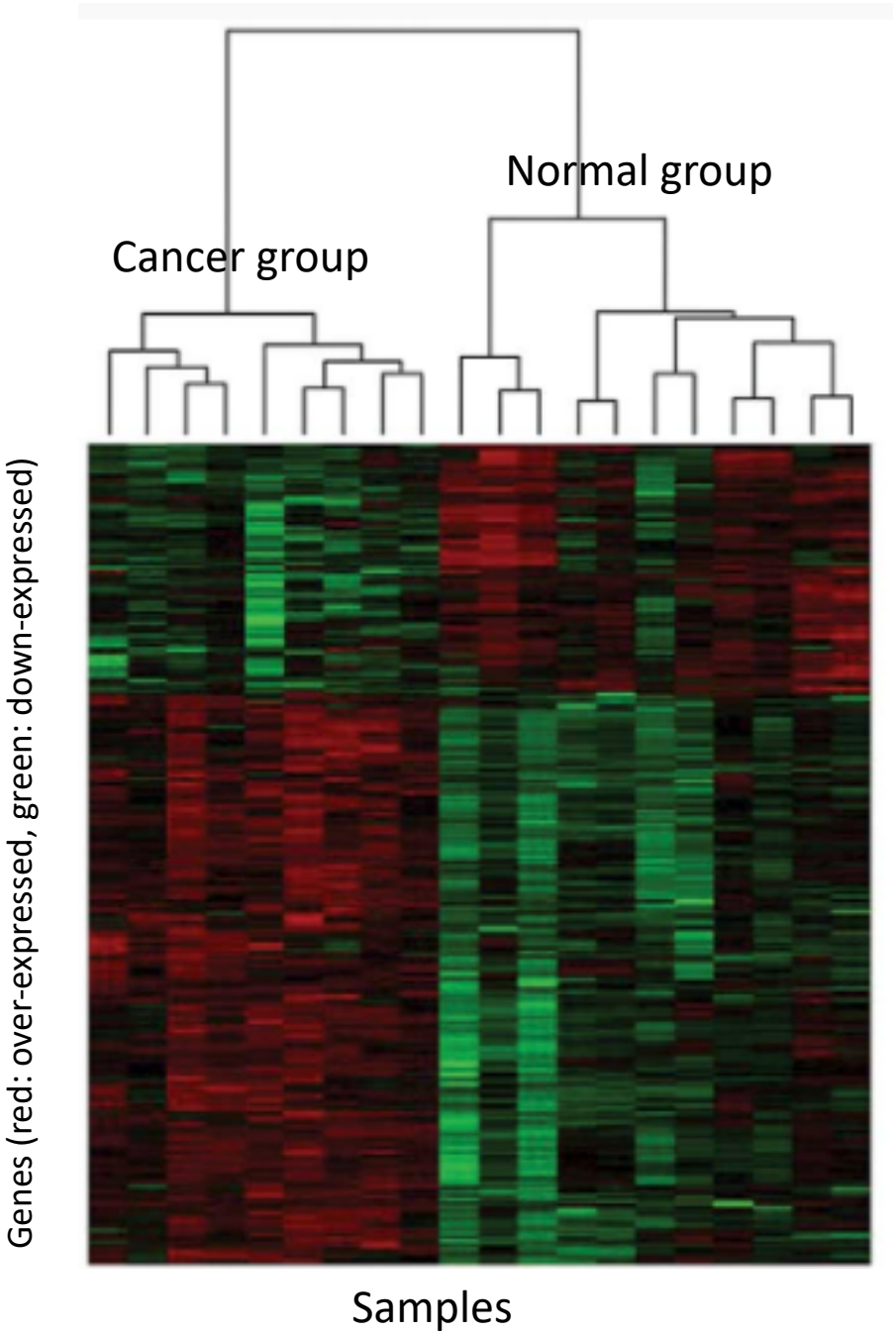
Approach 2

Group samples based on their gene expression



Approach 2

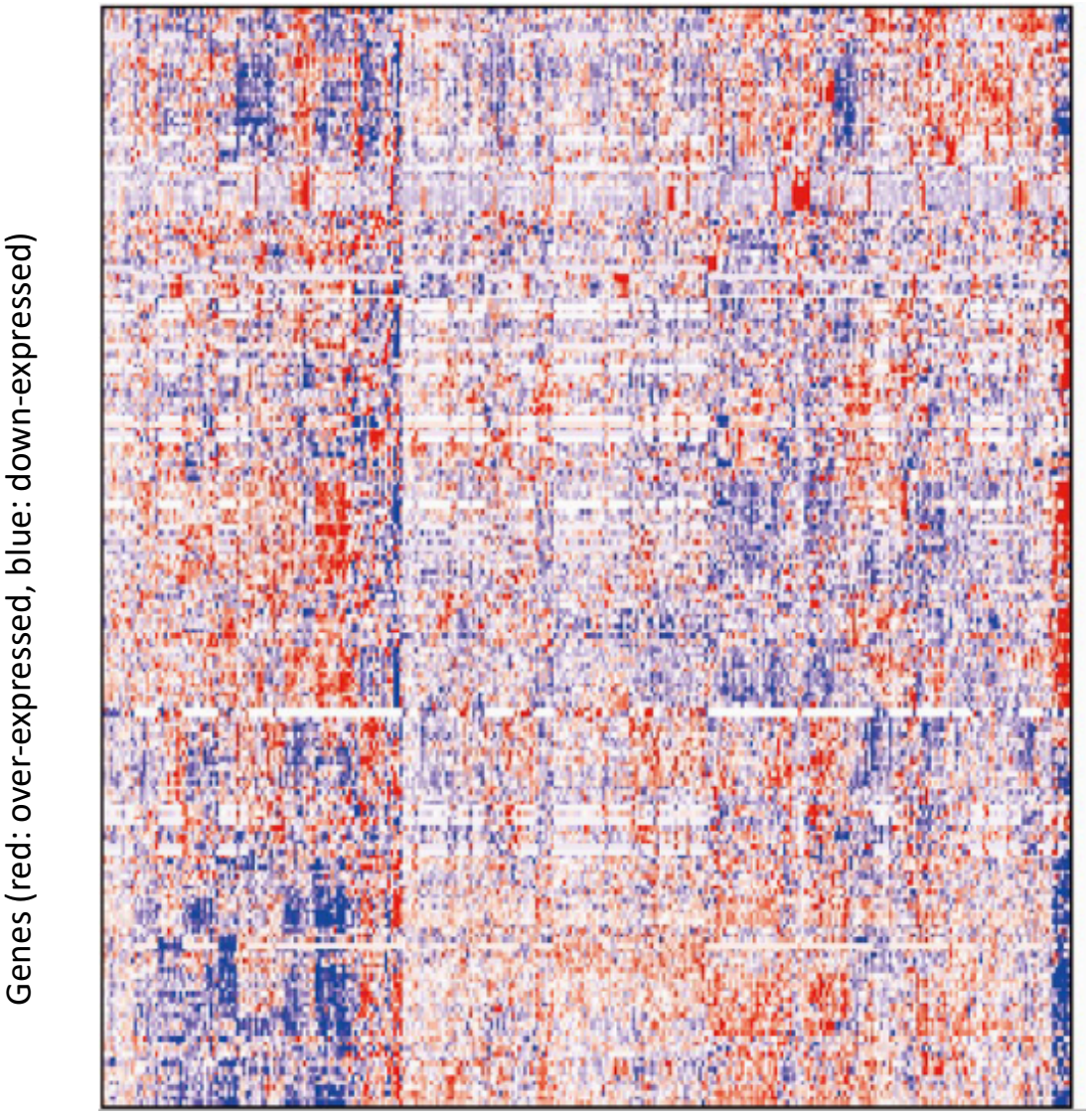
Group samples based on their gene expression



Approach 2

Group samples based on their gene expression?

Start diving into Bioinformatics



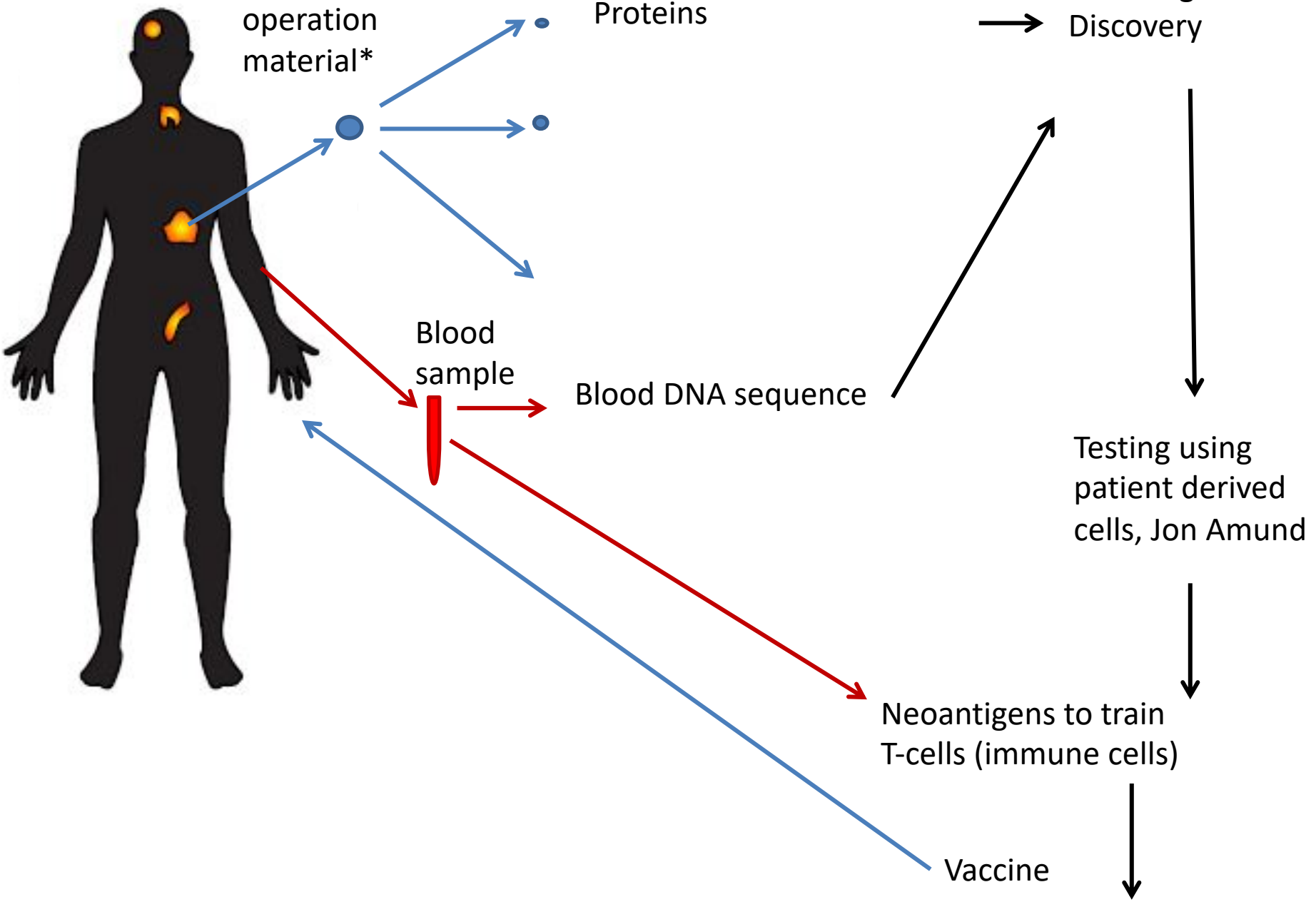
Peng et al. J TCR 2019

My research:
MultiOmics Approach

Integrate:

- GenOmics
- TranscriptOmics
- ProteOmics

Goal: Cancer immunotherapy


Tumor piece, operation material*
Cancer DBA sequence and Proteins
Neoantigen Discovery
Blood sample
Blood DNA sequence
Testing using patient derived cells, Jon Amund
Neoantigens to train T-cells (immune cells)
Vaccine

Useful links

- https://www.nature.com/scitable/topics/
- http://rosalind.info/problems/list-view/

Cancer data resources:

- https://dcc.icgc.org/
- https://cancer.sanger.ac.uk/

Questions?