香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional · Creative
For The World

IEEE BigData 2020 Tutorial 1

Big Data Applications for High-Impact/Emerging Topics:

Case Studies on Molecular Biology and Its Applications to COVID-19

Ka-Chun Wong, Jiecong Lin, Shixiong Zhang, Xiangtao Li

Outline

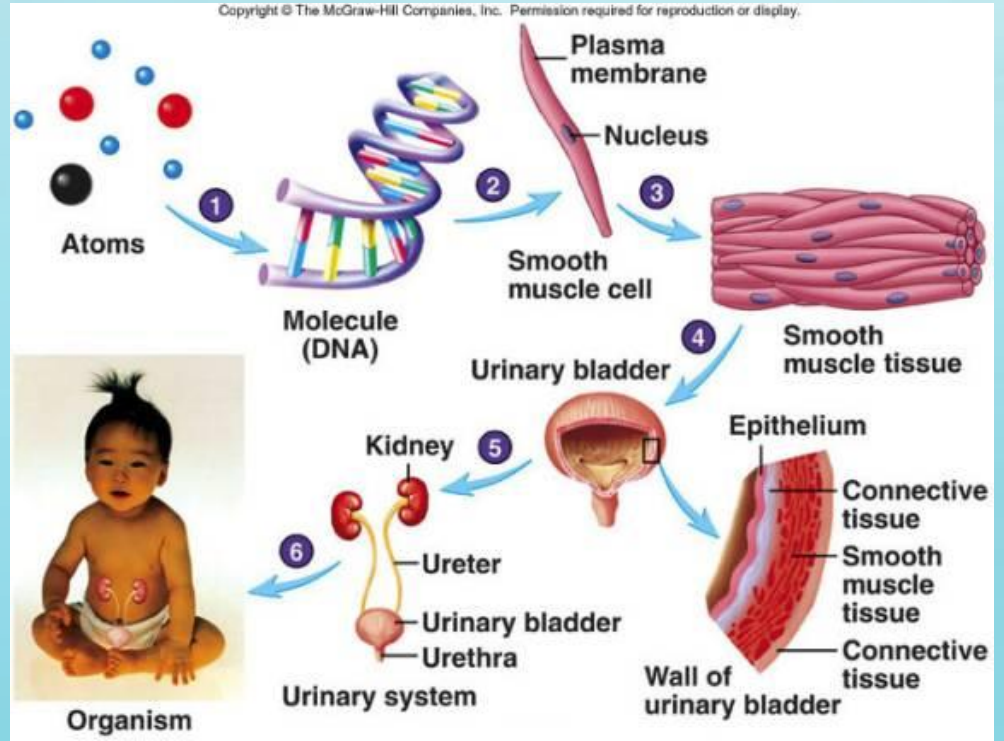
- (30 minutes) Basic Concepts in Molecular Biology
- (60 minutes) Data Application Case Studies for Molecular Biology
- (30 minutes) Data Application Case Studies for COVID-19

Basic Concepts in Molecular Biology

(30 minutes)

Our body

- Our body consists of a number of organs
- Each organ composes of a number of tissues
- Each tissue composes of cells



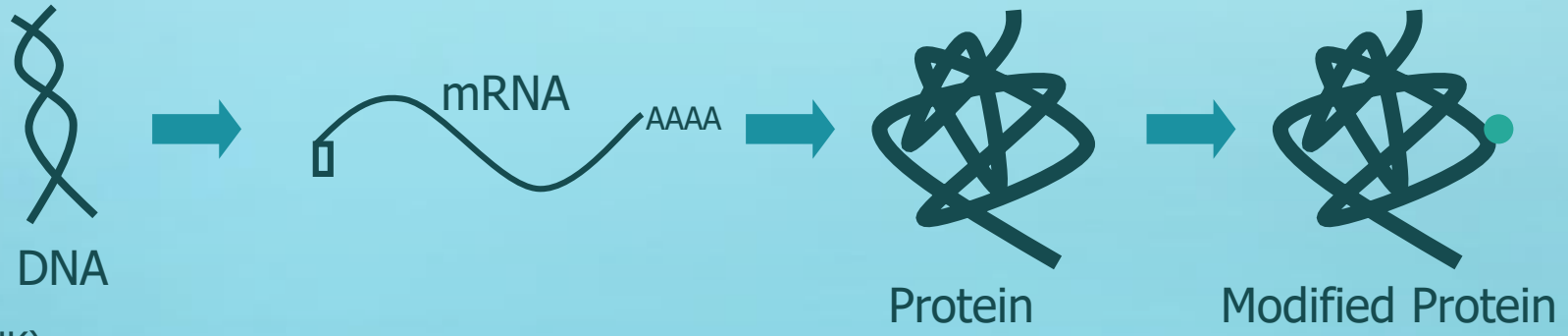
Cell

- Cell performs two type of functions:
  - Perform chemical reactions necessary to maintain our life
  - Pass the information for maintaining life to the next generation
- Actors:
  - DNA stores and passes information
  - RNA is the intermediate between DNA and proteins
  - Protein performs chemical reactions

® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Central Dogma

- Central Dogma tells us how we get the protein from a gene. This process is called gene expression.
- The expression of gene consists of steps
  - Transcription: DNA → mRNA
  - Translation: mRNA → Protein
  - Post-translation Modification: Protein → Modified protein



® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

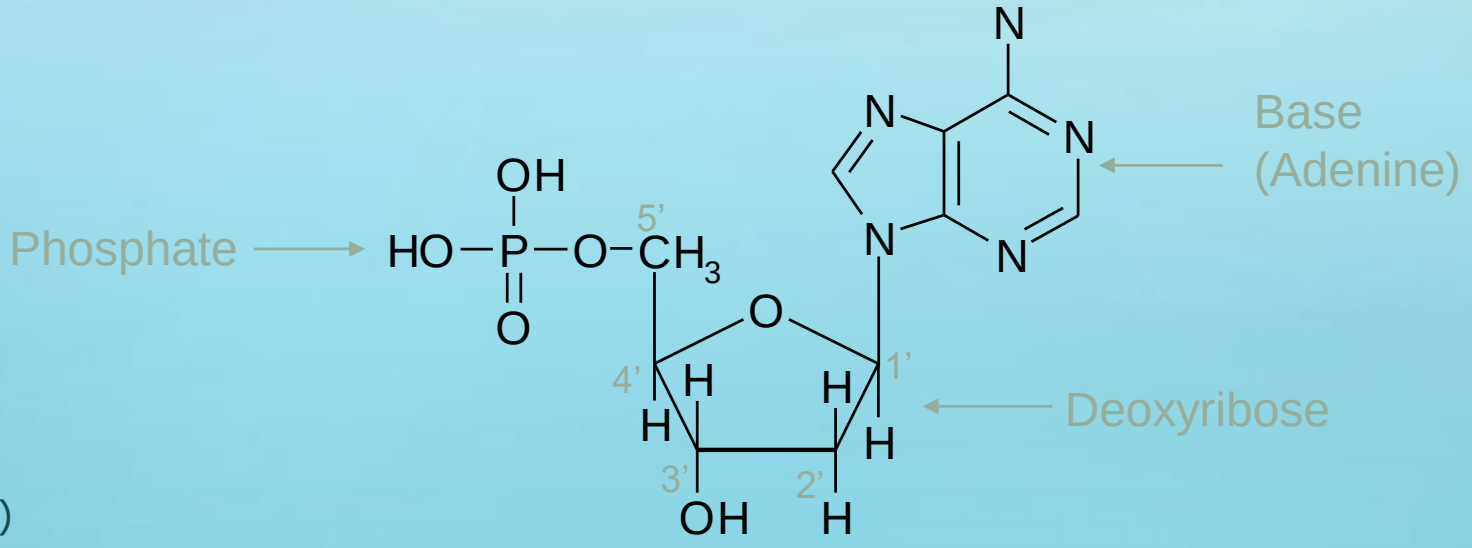
DNA

- DNA stores the instructions needed by the cell to perform daily life function.
- It consists of two strands which are interwoven together and forms a double helix.
- Each strand is a chain of some small molecules called nucleotides.
- There are 4 different nucleotides of DNA:

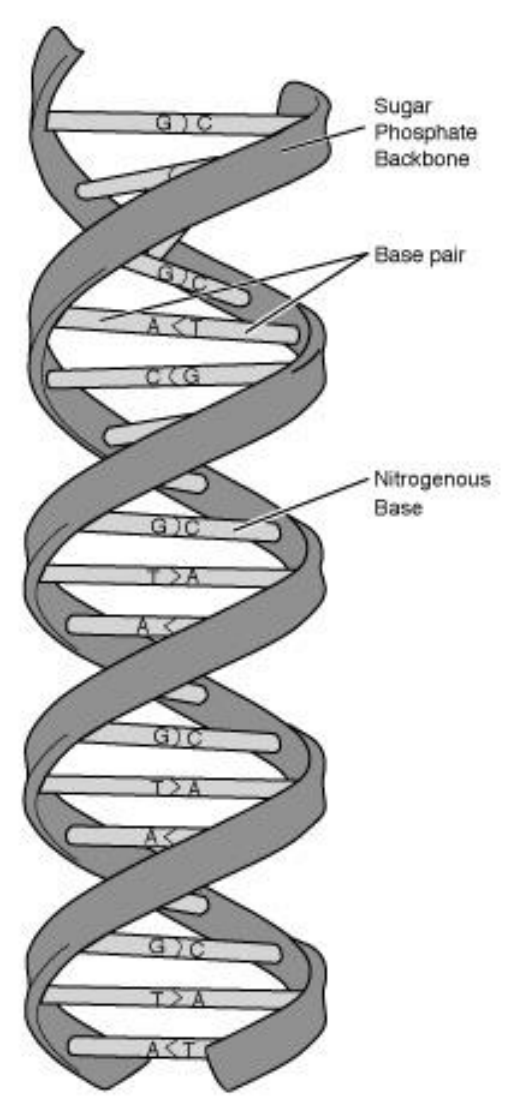
  adenine(**A**), cytosine(**C**), guanine(**G**), thymine(**T**).

Nucleotide for DNA

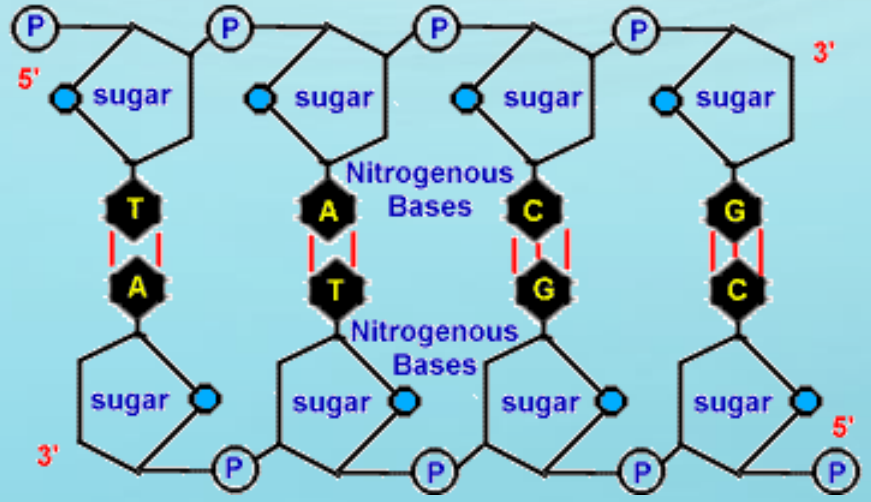
- DNA nucleotide consists of three parts:
  - Deoxyribose
  - Phosphate (bound to the 5’ carbon)
  - Base (bound to the 1’ carbon)



Double stranded DNA



- Normally, DNA is double-stranded within a cell. The two strands are antiparallel. One strand is the reverse complement of another one.
- The double strands are interwoven together and form a double helix.
- One conjectured reason for the double-stranded structure is that it eases DNA replications.



® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Chromosome

- Usually, DNA is tightly wound around histone proteins and forms a chromosome.
- The total information stored in all chromosomes constitute a genome.
- Example:
  - Human Genome: has 3G base pairs, organized in 23 pairs of chromosomes

® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Gene

- A gene is a sequence of DNA that encodes for a protein.
- In human genome, it is expected there are 20,000 – 30,000 genes.
- For the gene that encodes for a protein,
  - In Prokaryotic genome, one gene corresponds to one protein.
  - In Eukaryotic genome, one gene can correspond to more than one protein because of the process “alternative splicing”.

® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Gene

MENU

nature
International journal of science

Search E-alert Submit

NEWS · 19 JUNE 2018

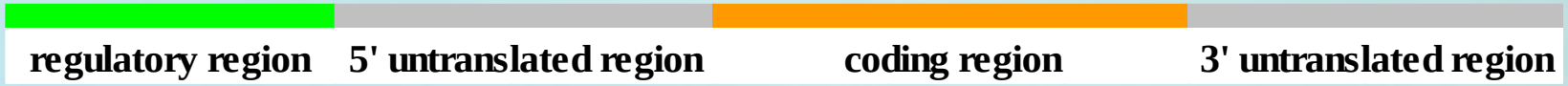
New human gene tally reignites debate

Some fifteen years after the human genome was sequenced, researchers still can't agree on how many genes it contains.

Cassandra Willyard

https://www.nature.com/articles/d41586-018-05462-w

More on Gene Structure



- Each Gene has 4 regions
  - Regulatory region contains a promoter (and enhancers) which regulate gene transcription.
  - Coding region contains the codons for protein. It is also called open reading frame. Its length is a multiple of 3. It must begin with start codon, end with end codon, and the rest of its codons are not end codon.
  - mRNA transcript contains 5’ untranslated region + coding region + 3’ untranslated region

® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Organism Complexity v.s. Genome Size

- E. coli Genome: 5M base pairs

- Human Genome: 3G base pairs

- Amoeba dubia (a single cell organism): 670G base pairs

- Genome size appears not to have any relationship with the complexity of the organism

https://www.waterfall-d-mannose.com/ecoli-uropathogen.html

Number of Genes vs. Genome Size

- Prokaryotic genome: e.g. E. coli
  - Number of base pairs: 5M
  - Number of genes: 4k
  - Average length of a gene: 1000 bp
- Eukaryotic genome: e.g. Human
  - Number of base pairs: 3G
  - Estimated number of genes: 20k – 30k
  - Estimated average length of a gene: 1000-2000 bp
- 90% of the E. coli genome consists of coding regions.
- Less than 3% of the human genome is believed to be coding regions. The rest is called non-coding regions (junk DNA?).
- Genome size appears not to have any relationship with the number of genes!

Note that, before 2001, scientists thought we had 100k genes.

	Organism	# of protein-coding genes	# of genes naïve estimate: (genome size /1000)	BNID
viruses	HIV 1	9	10	105769
	Influenza A virus	10-11	14	105767
	Bacteriophage λ	66	49	105770
	Epstein Barr virus	80	170	103246
prokaryotes	*Buchnera sp.*	610	640	105757
	T. maritima	1,900	1,900	105766
	S. aureus	2,700	2,900	105500
	V. cholerae	3,900	4,000	105760
	B. subtilis	4,400	4,200	111448
	E. coli	4,300	4,600	105443
eukaryotes	*S. cerevisiae*	6,600	12,000	105444
	C. elegans	20,000	100,000	101364
	A. thaliana	27,000	140,000	111380
	D. melanogaster	14,000	140,000	111379
	F. rubripes	19,000	400,000	111375
	Z. mays	33,000	2,300,000	110565
	M. musculus	20,000	2,800,000	100308
	H. sapiens	21,000	3,200,000	100399, 111378
	T. aestivum (hexaploid)	95,000	16,800,000	105448, 102713

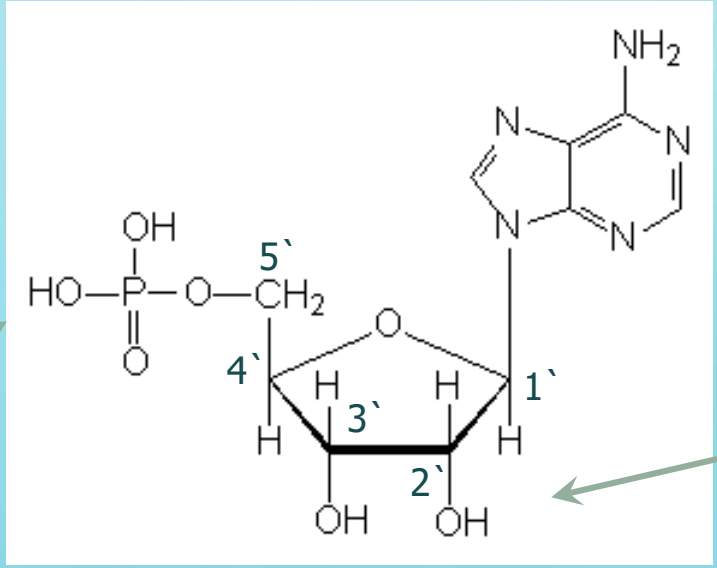
http://book.bionumbers.org/how-many-genes-are-in-a-genome/

RNA

- RNA has both the properties of DNA and protein; for instance,
  - Similar to DNA, it can store and transfer information.
  - Similar to protein, it can form complex 3-dimensional structure and perform some functions.
  - There are 4 different nucleotides for RNA:
    - adenine(A), cytosine(C), guanine(G), uracil(U)

Nucleotide for RNA

- RNA nucleotide consists of three parts:
  - Ribose Sugar (has an extra OH group at 2’)
  - Phosphate (bound to the 5’ carbon)
  - Base (bound to the 1’ carbon)



Base (Adenine)

Ribose Sugar

Phosphate

® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

RNA vs DNA

- RNA is single-stranded.
- The nucleotides of RNA are quite similar to that of DNA, except that it has an extra OH at position 2'. (see the previous slide)
  - Due to this extra OH, it can form more hydrogen bonds than DNA. Thus, RNA can form a complex 3-dimensional structure.
- RNA uses the base U instead of T.
  - U is chemically similar to T.
  - U is also complementary to A.

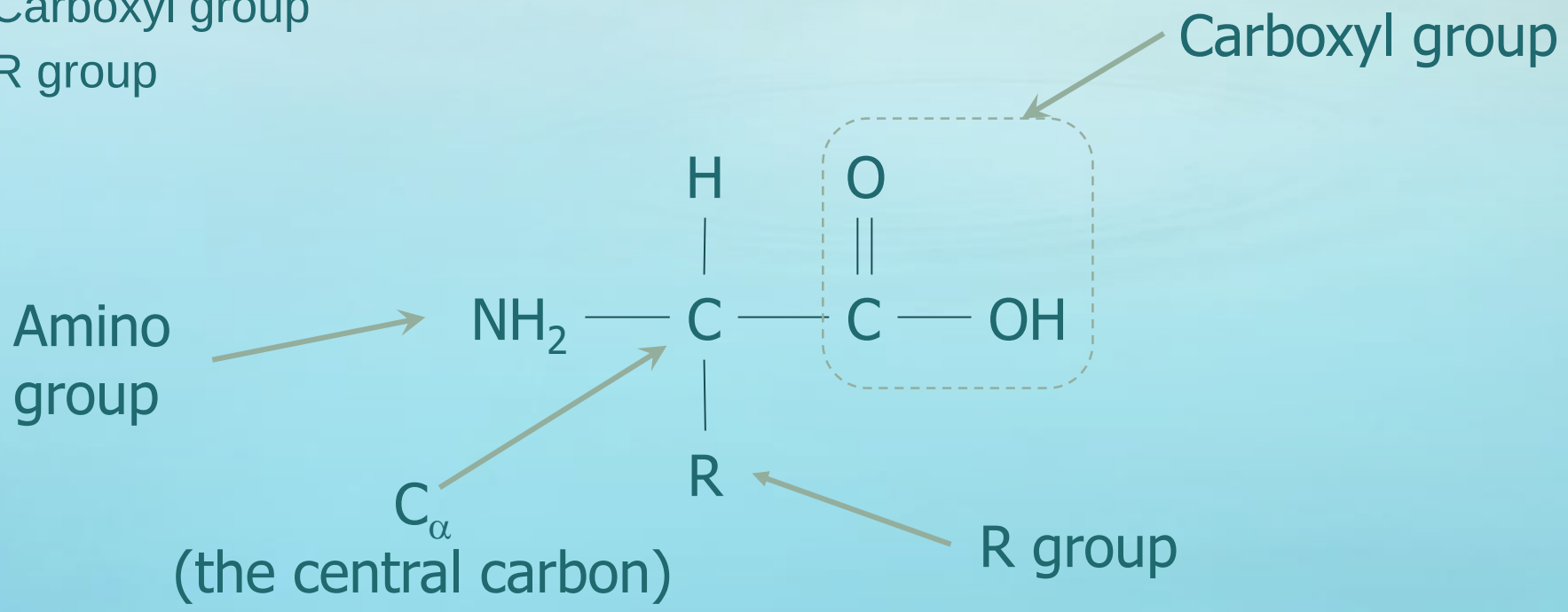
® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Protein

- Protein is a molecular sequence with an alphabet of 20 amino acids.
  - The typical length is from 20 to 5000 amino acids.
  - Average protein contains around 350 amino acids.
- Protein folds into three-dimensional shapes, which form the building blocks and perform most of the chemical reactions within a cell.

Amino acid

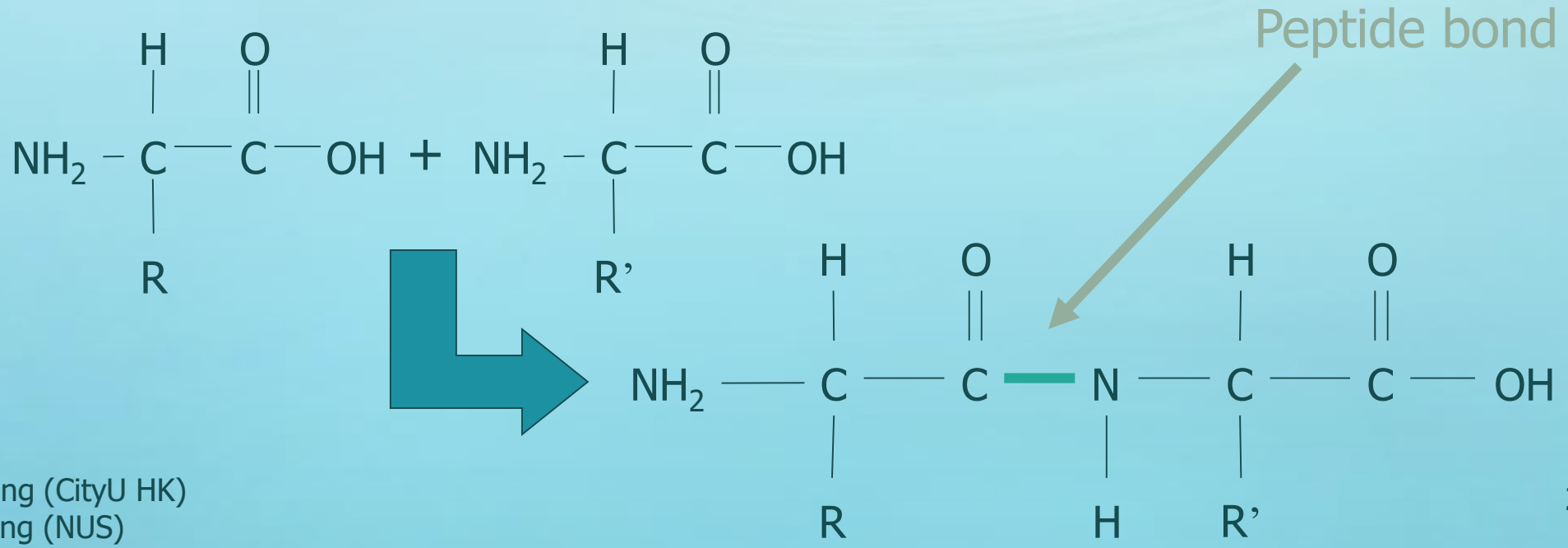
- Each amino acid consists of
  - Amino group
  - Carboxyl group
  - R group



® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Polypeptide

- Protein or polypeptide chain is formed by joining the amino acids together via a peptide bond.
- One end of the polypeptide is the amino group, which is called N-terminus. The other end of the polypeptide is the carboxyl group, which is called C-terminus.



® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Classification of amino acids (I)

- 20 common amino acids can be classified into 4 types.

- Positively charged (basic) amino acids:
  - Arginine (Arg, R)
  - Histidine (His, H)
  - Lysine (Lys, K)
- Negatively charged (acidic) amino acids:
  - Aspartic acid (Asp, D)
  - Glutamic acid (Glu, E)

® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Classification of amino acids (II)

- Polar amino acids:
  - Overall uncharged, but uneven charge distribution. Can form hydrogen bonds with water. They are called hydrophilic. Often found on the outer surface of a folded protein.
    - Asparagine (Asn, N)
    - Cysteine (Cys, C)
    - Glutamine (Gln, Q)
    - Glycine (Gly, G)
    - Serine (Ser, S)
    - Threonine (Thr, T)
    - Tyrosine (Tyr, Y)

® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Classification of amino acids (III)

- Non-polar amino acids:
  - Overall uncharged and uniform charge distribution. Cannot form hydrogen bonds with water. They are called hydrophobic. Tend to appear on the inside surface of a folded protein.
    - Alanine (Ala, A)
    - Isoleucine (Ile, I)
    - Leucine (Leu, L)
    - Methionine (Met, M)
    - Phenylalanine (Phe, F)
    - Proline (Pro, P)
    - Tryptophan (Trp, W)
    - Valine (Val, V)

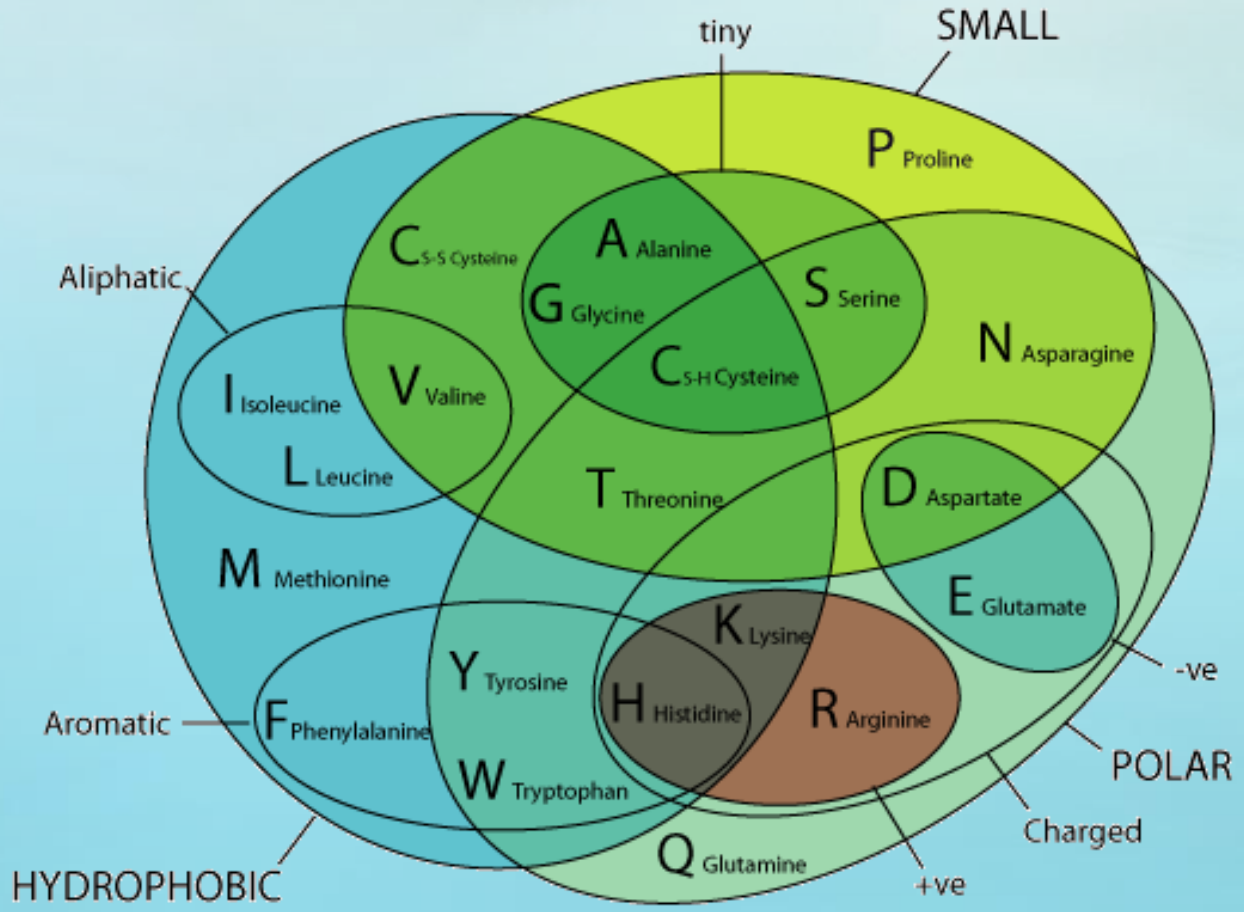
® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Amino acid properties

Amino Acid	3-Letter[133]	1-Letter[133]	Side-chain class	Side-chain polarity[133]	Side-chain charge (pH 7.4)[133]	Hydropathy index[134]	Absorbance λ_{max}(nm)[135]	ε atλ_{max} (m M^{-1} cm^{-1})[135]	MW(Weight)	Occurrence in proteins (%)[136]
Alanine	Ala	A	aliphatic	nonpolar	neutral	1.8			89.094	8.76
Arginine	Arg	R	basic	basic polar	positive	−4.5			174.203	5.78
Asparagine	Asn	N	amide	polar	neutral	−3.5			132.119	3.93
Aspartic acid	Asp	D	acid	acidic polar	negative	−3.5			133.104	5.49
Cysteine	Cys	C	sulfur-containing	nonpolar	neutral	2.5	250	0.3	121.154	1.38
Glutamic acid	Glu	E	acid	acidic polar	negative	−3.5			147.131	6.32
Glutamine	Gln	Q	amide	polar	neutral	−3.5			146.146	3.9
Glycine	Gly	G	aliphatic	nonpolar	neutral	−0.4			75.067	7.03
Histidine	His	H	basic aromatic	basic polar	positive(10%) neutral(90%)	−3.2	211	5.9	155.156	2.26
Isoleucine	Ile	I	aliphatic	nonpolar	neutral	4.5			131.175	5.49
Leucine	Leu	L	aliphatic	nonpolar	neutral	3.8			131.175	9.68
Lysine	Lys	K	basic	basic polar	positive	−3.9			146.189	5.19
Methionine	Met	M	sulfur-containing	nonpolar	neutral	1.9			149.208	2.32
Phenylalanine	Phe	F	aromatic	nonpolar	neutral	2.8	257, 206, 188	0.2, 9.3, 60.0	165.192	3.87
Proline	Pro	P	cyclic	nonpolar	neutral	−1.6			115.132	5.02
Serine	Ser	S	hydroxyl-containing	polar	neutral	−0.8			105.093	7.14
Threonine	Thr	T	hydroxyl-containing	polar	neutral	−0.7			119.119	5.53
Tryptophan	Trp	W	aromatic	nonpolar	neutral	−0.9	280, 219	5.6, 47.0	204.228	1.25
Tyrosine	Tyr	Y	aromatic	polar	neutral	−1.3	274, 222, 193	1.4, 8.0, 48.0	181.191	2.91
Valine	Val	V	aliphatic	nonpolar	neutral	4.2			117.148	6.73

https://en.wikipedia.org/wiki/Amino_acid

Summary of the amino acid properties



http://www.jalview.org/help/html/misc/properties.gif

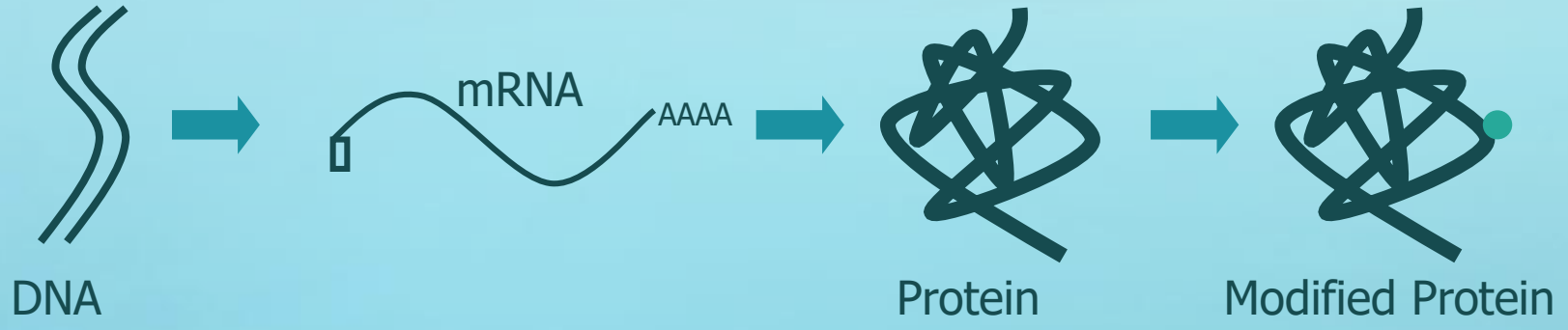
Protein Structure Levels

- Primary structures
  - Amino acid sequences.
- Secondary structures
  - Local 3D structures formed by hydrogen bonding (α-helices and β-sheets) or others (random coils).
- Tertiary structures
  - 3D structures of single protein based on the interactions of local 3D structures (i.e. secondary structures) due to the hydrophobic effect.
- Quaternary structures
  - 3D structures of multiple proteins, forming a protein complex based on different forces.

® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Central Dogma

- Central Dogma tells us how we get proteins from the genes. This process is called gene expression.
- The expression of each typical gene consists of steps
  - Transcription: DNA → mRNA
  - Translation: mRNA → Protein
  - Post-translation Modification: Protein → Modified protein



® Dr. Ka-Chun Wong (CityU HK)
® Dr. Wing-Kin Sung (NUS)

Transcription

- Synthesize a piece of RNA (messenger RNA, mRNA) from one strand of the DNA gene.
  1. An enzyme RNA polymerase temporarily separates the double-stranded DNA
  2. It begins the transcription at the transcription start site.
  3. A → A, C→C, G→G, and T→U
  4. Once the RNA polymerase reaches the transcription terminator, transcription stop.

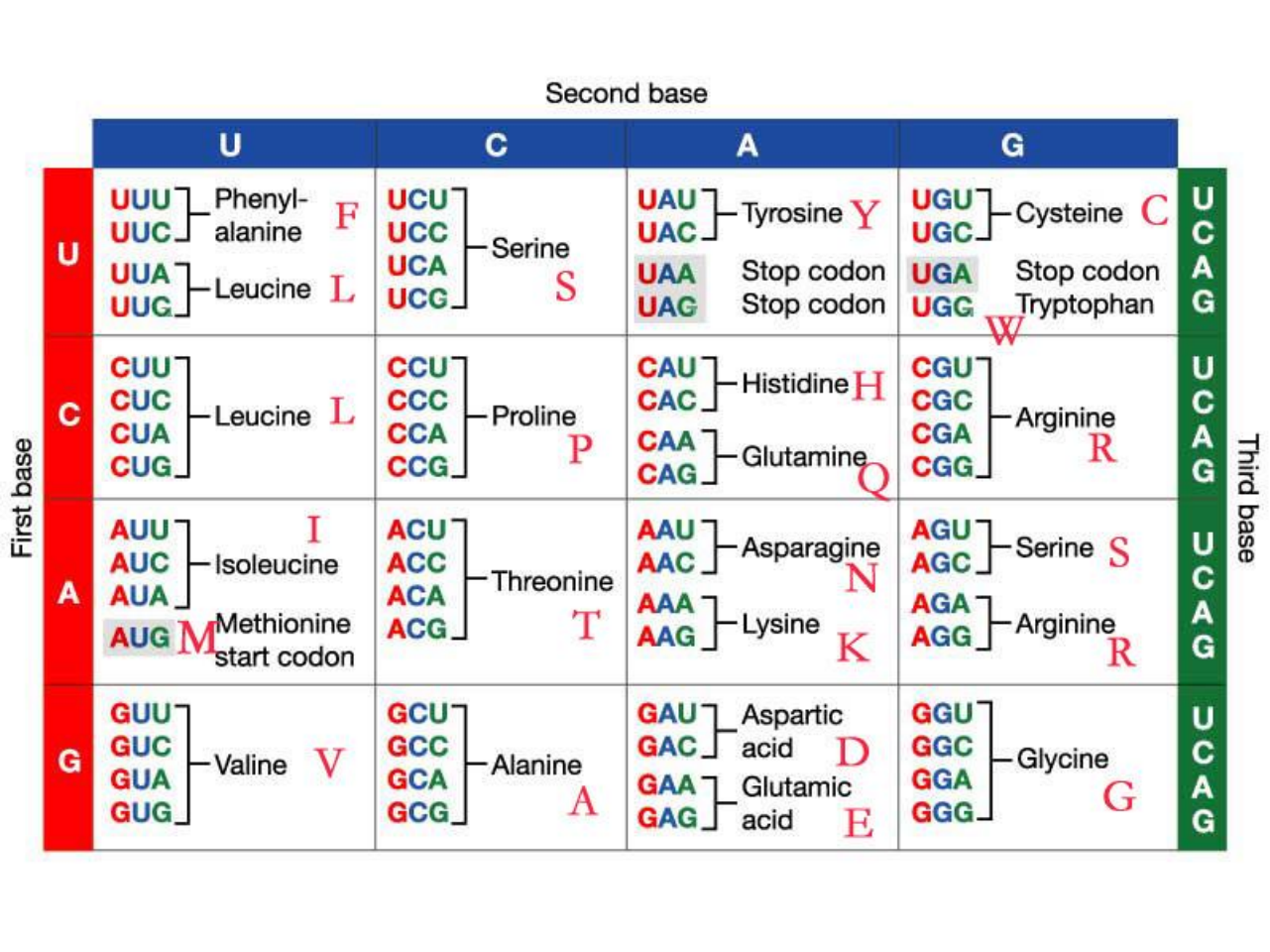
https://en.wikipedia.org/wiki/Terminator_(genetics)

Translation

- Translation synthesizes a protein from a mRNA.
- In fact, amino acids are encoded by consecutive sequences of 3 nucleotides, called codon.
- The decoding table from codon to amino acid is called Codon Table.
- Note:
  - There are 4^3=64 different codons. Thus, the codons are not one-to-one correspondence to the 20 amino acids.
  - All organisms use the same decoding table!
  - The codons that encode the same amino acid tend to have the same first and second nucleotide.
  - Recall that amino acids can be classified into 4 groups. A single base change in a codon is usually not sufficient to cause a codon to code for an amino acid in different group.

Codon Table

- Start codon: ATG (also code for M)
- Stop codon: TAA, TAG, TGA



http://bioinfo.bisr.res.in/project/crat/pictures/codon.jpg

Brief History of Genomics (Bioinformatics)

- 1985: Complete sequencing of genomes of various organisms
- 1990: Launch of Human Genome Project (HGP)
- 2000: By shotgun sequencing, Craig Venter and Francis Collins jointly announced the publication of the first draft of the human genome.
- 2010:
  - ENCODE Project:
    - Annotation of the whole genome including non-coding regions
  - 1000 Human Genomes Project
    - Study of the human population genomes across different regions
- Now:
  - Genotype-Tissue Expression Project (GTEx)
    - To study the relationship between gene expression and genetic variation in human tissue
  - Personal Medicine based on Human Genome
    - To develop customized medical solutions based on genome data
  - The GenomeAsia 100K Project

Data Application Case Studies for Molecular Biology

(60 minutes)

Case Studies

- Case for synthesizing Heterodimeric DNA Motifs [1]
- Case for predicting CRISPR-Cas9 Off-targets [2,3]
- Case for screening Early Cancers from Blood [4]

References

[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic Acids Research, 47(4), 1628-1636. (2019 Impact Factor: 11.501)

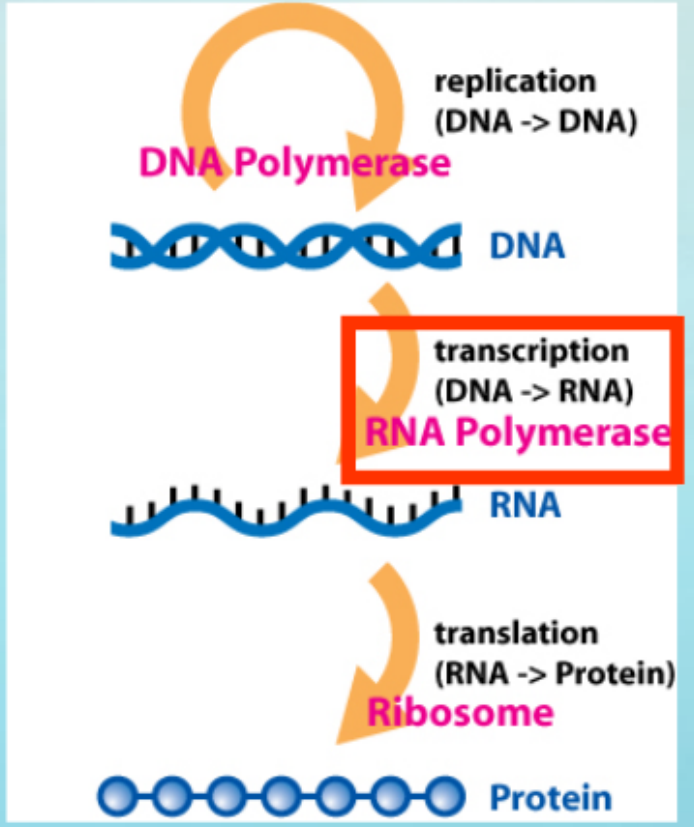
[2] Lin, J., & Wong, K. C. (2018). Off-target predictions in CRISPR-Cas9 gene editing using deep learning. Bioinformatics, 34(17), i656-i663.

[3] Lin, J., Zhang, Z., Zhang, S., Chen, J., & Wong, K. C. (2020). CRISPR-Net: A Recurrent Convolutional Network Quantifies CRISPR Off-Target Activities with Mismatches and Indels. Advanced Science, 1903562. (2019 Impact Factor =15.4)

[4] Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Background

Central Dogma of Molecular Biology:
"DNA makes RNA, RNA makes Protein"

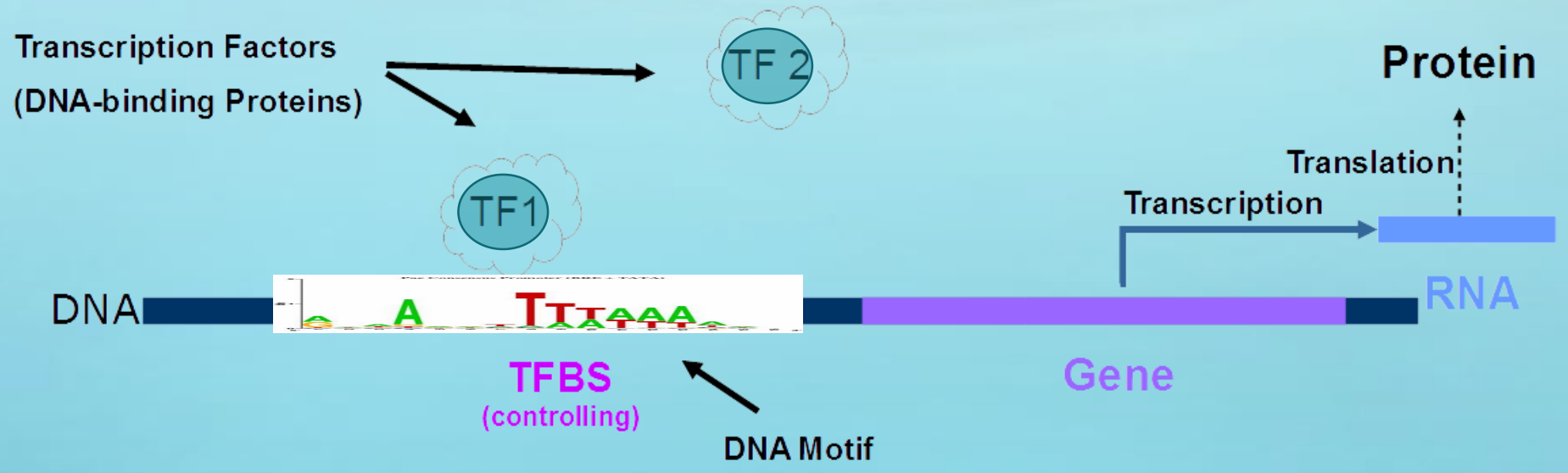


http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

Background

Protein-DNA Binding

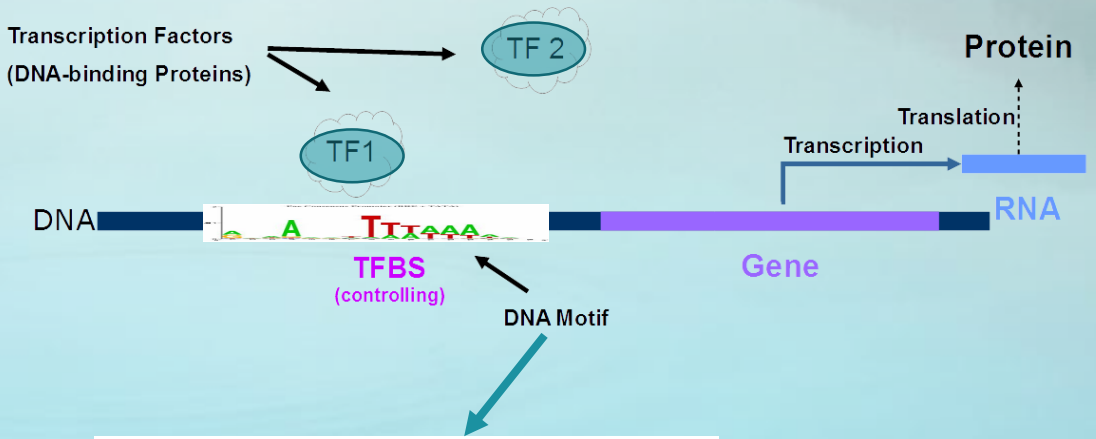
- The binding between proteins (e.g. Transcription Factors,TFs) and DNA (e.g. Transcription Factor Binding Sites, TFBSs) play an important role
- TFs bind in a sequence-specific manner to TFBSs to regulate gene transcription

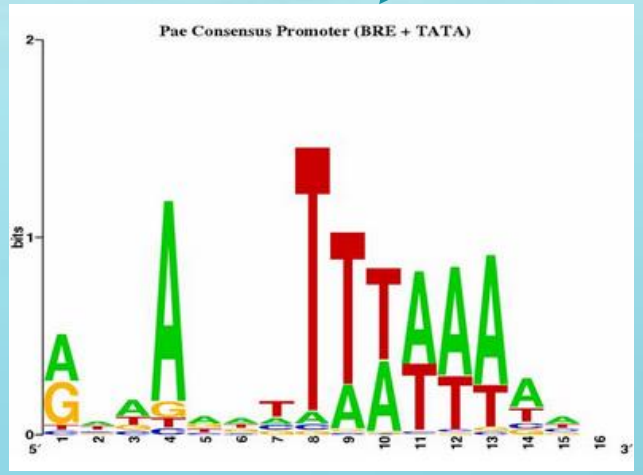


Background

- To fully understand a gene's function, it is essential to identify the TFs that regulate the gene and the corresponding TF-binding sites (also known as DNA Motifs).

- DNA motifs re relatively short (10–20 bp) and highly degenerate sequence motifs, which make their effective identification a challenging task.





A motif logo example

DNA Motif Example

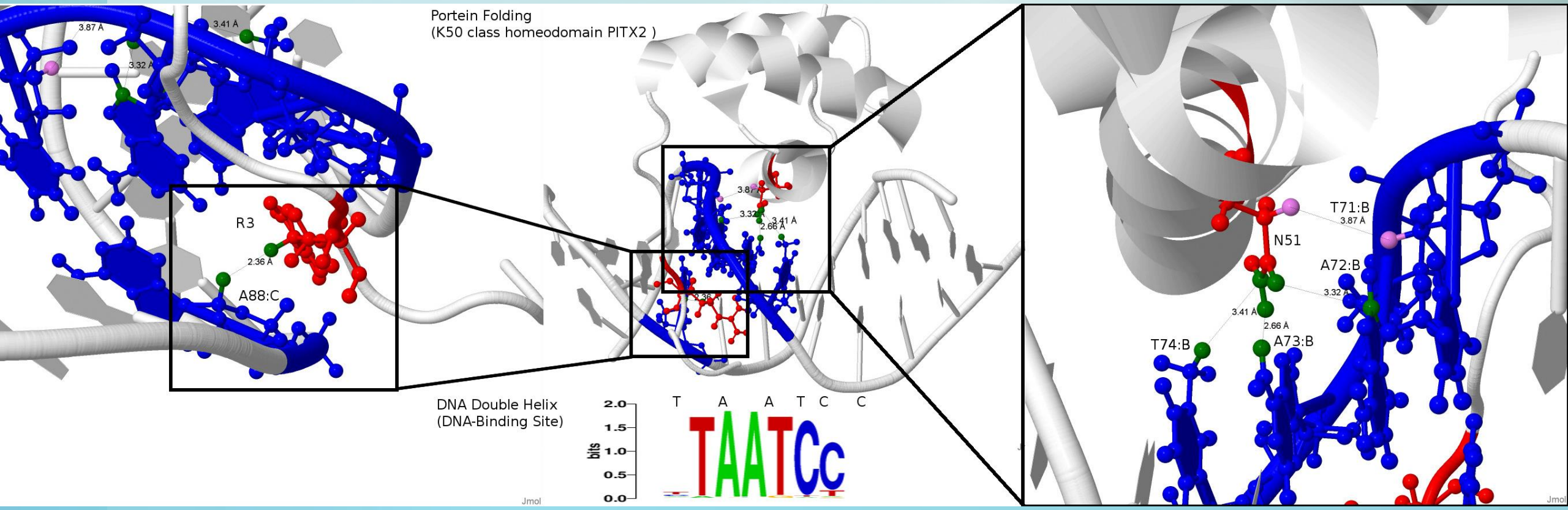
For example, given the following DNA sequences:

```
GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT
```

the PWM using relative frequencies is:

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

DNA Motif Example at 3D Level



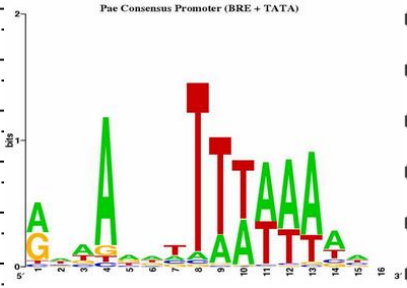
Traditional Research

Finding DNA Motifs on DNA Sequences

> 3000 sequences
……
……
……
……
……
……
……
……
……
……
……
……

Protein Binding Microarray

Sequence (with primer)	Normalized Signal Intensity
AAAAAAAAAACGGGTCTGCTGTGTTTCCATGGGCAGTCTGTGTTCCGTTGTCCGTGCTGT	1421.285654
CCATGGGCAGCGATTCCATTGAATCAACCATCCGCGTCTGTGTTCCGTTGTCCGTGCTGT	719.1041392
ACCATCCGCGTGGTGTCATACGCCCAATATCGAGCGTCTGTGTTCCGTTGTCCGTGCTGT	1762.812037
ATATCGAGCAGTAAAAGCGAAAAACAATCTTGGAAGTCTGTGTTCCGTTGTCCGTGCTGT	465.8865101
ATCTTGGAATTAATGGCCAAGGATTAGAATACCCTGTCTGTGTTCCGTTGTCCGTGCTGT	588.0584476
GAATACCCTACCAAATAGTGTAAAGACTTGTTGTAGTCTGTGTTCCGTTGTCCGTGCTGT	582.1760333
CTTGTTGTACAAAGTATCCCGTACTACGCGTTGGGGTCTGTGTTCCGTTGTCCGTGCTGT	710.6338075
CGCGTTGGGTAACTCGGGTTCTCTTGGCAGATTCCGTCTGTGTTCCGTTGTCCGTGCTGT	934.8380349
GCAGATTCCTACGGCCATCGTCAGACACGACACGGGTCTGTGTTCCGTTGTCCGTGCTGT	741.7662929
ACGACACGGGACCGTAGCAGGGACGTACATACTGTGTCTG GT	18507.18435
ACATACTGTTCCAGGTAGGCGTACGGATCTTCCAGGTCTG GT	759.8673106
ATCTTCCAGTGACTCGGTGGGTATGATGCTTTGCAGTCTG GT	868.2056512
TGCTTTGCAATTACGAACTAAAGTTGTCCCTGTTCGTCTG GT	597.3772212
TCCCTGTTCCCGCCTCGCCAACTGAACATAGCATTGTCTG GT	1156.546346
CATAGCATTATCCGATCGTCGCAGGAGGCTTATTGGTCTG GT	982.9229197
GGCTTATTGCGATTCCAGCAGTGGGCGTTCGGCATGTCTGTGTTCCGTTGTCCGTGCTGT	491.8523141
GTTCGGCATTCACAGCGGGCCGAAATAGTCCCGCCGTCTGTGTTCCGTTGTCCGTGCTGT	673.6856787



Find DNA Motifs?

…… …… ……

Segments of a DeBruijn Sequence | Primer sequence | Normalized Signal Intensity

Traditional Research

Finding DNA Motifs on DNA Sequences

- Motif finding problem is NP-hard!
- Motif discovery by over-representation
  - MEME (Expectation Maximization)
  - **Gibbs Sampling**
  - SP-STAR
  - Winnower
  - Random projection
  - Mitra
  - Weeder
  - SPACE
  - GALF

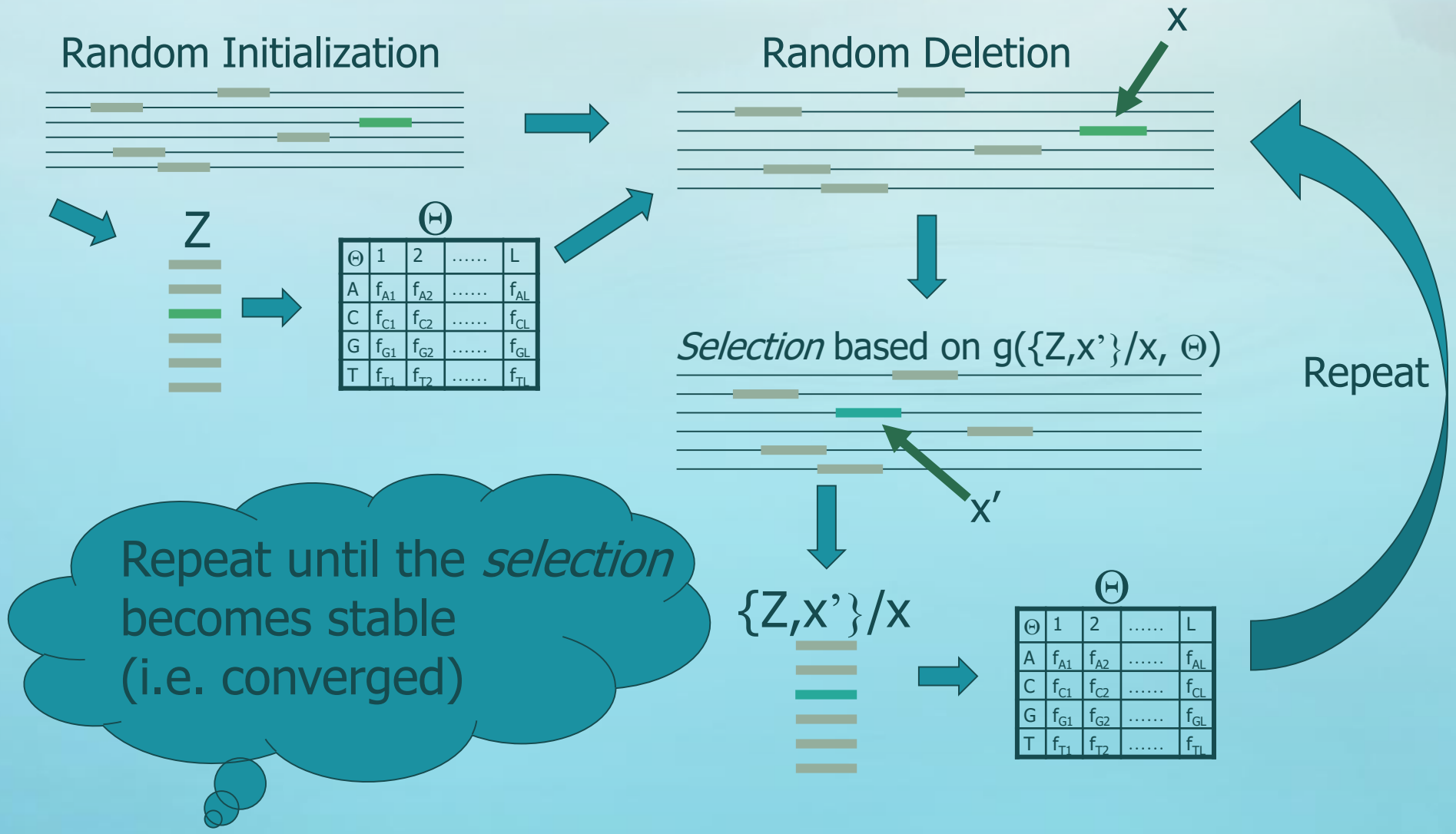
Given a set of sequences S, our problem is to find **a motif model Θ** and **the set of corresponding motif instances Z** which maximizes the log likelihood ratio score g(Z, Θ)

$$g(Z, \Theta) = \log \Pr(Z|\Theta)/\Pr(Z|\Theta_0)$$

where Θ_0 is a given background model.

Gibbs Sampling:

a technique to do random sampling



nature

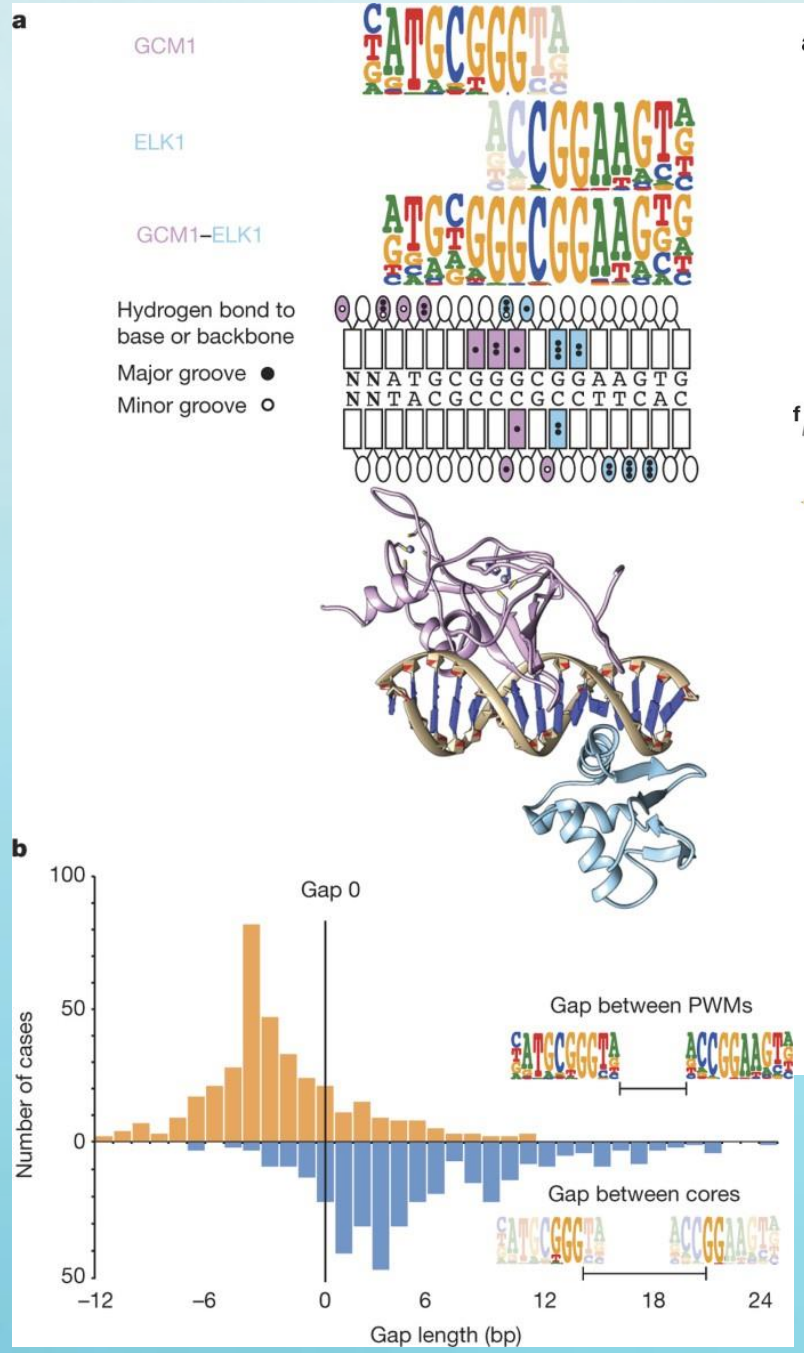
Published: 09 November 2015

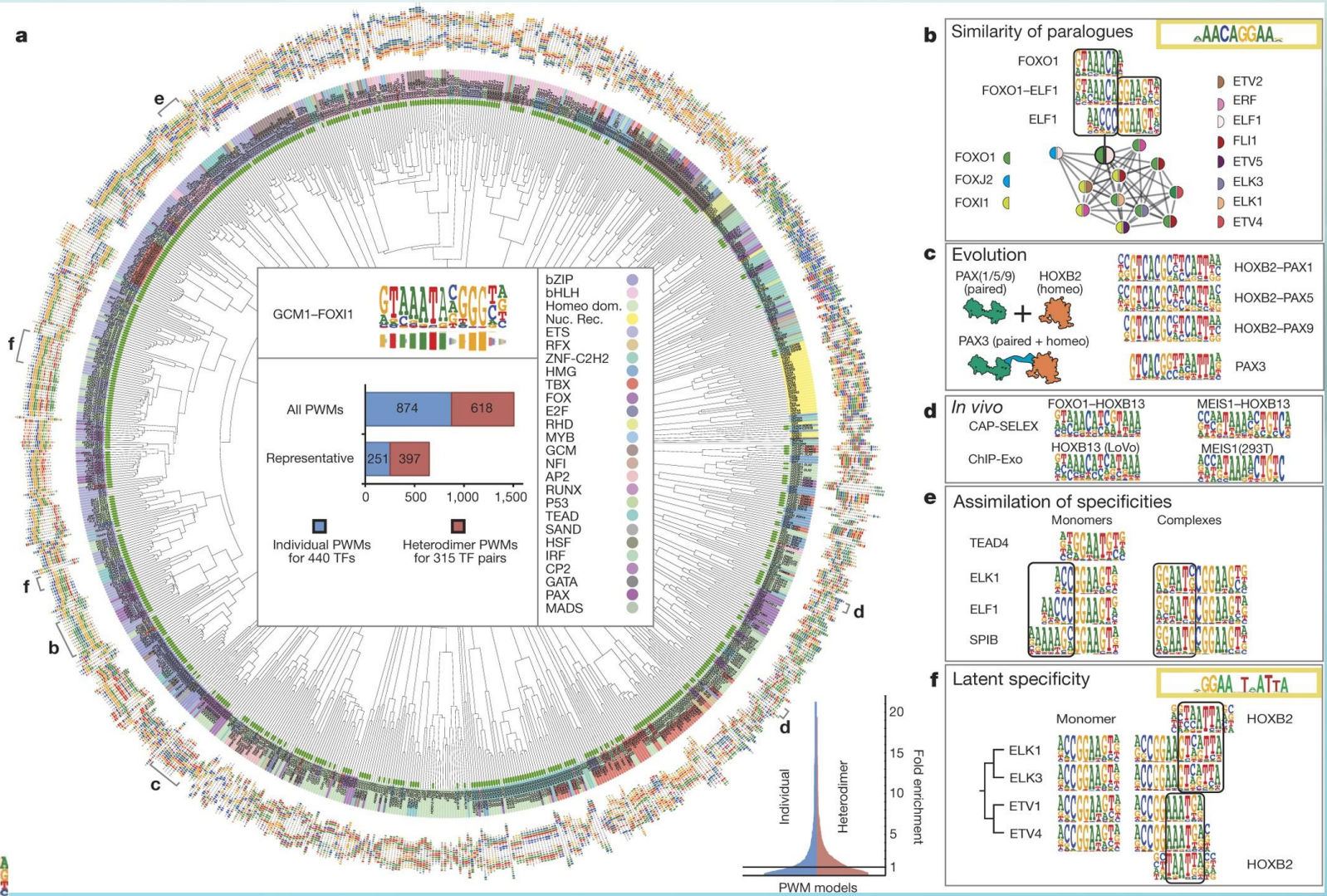
DNA-dependent formation of transcription factor pairs alters their binding specificity

Arttu Jolma, Yimeng Yin, Kazuhiro R. Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova & Jussi Taipale

Nature **527**, 384–388(2015) | Cite this article

5406 Accesses | **220** Citations | **130** Altmetric | Metrics





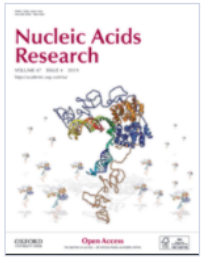
Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., ... & Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, *527*(7578), 384-388.

Most Updated Research on DNA Motif [1]

OXFORD ACADEMIC

Nucleic Acids Research

Issues | Section browse ▾ | Advance articles | Submit ▾ | Purchase | About ▾ | All Nucleic Acids Research ▾



Volume 47, Issue 4
28 February 2019

Article Contents

Abstract
INTRODUCTION
MATERIALS AND METHODS
RESULTS

Heterodimeric DNA motif synthesis and validations

Ka-Chun Wong ✉, Jiecong Lin, Xiangtao Li, Qiuzhen Lin, Cheng Liang, You-Qiang Song

Nucleic Acids Research, Volume 47, Issue 4, 28 February 2019, Pages 1628–1636,
https://doi.org/10.1093/nar/gky1297
Published: 27 December 2018 **Article history** ▾

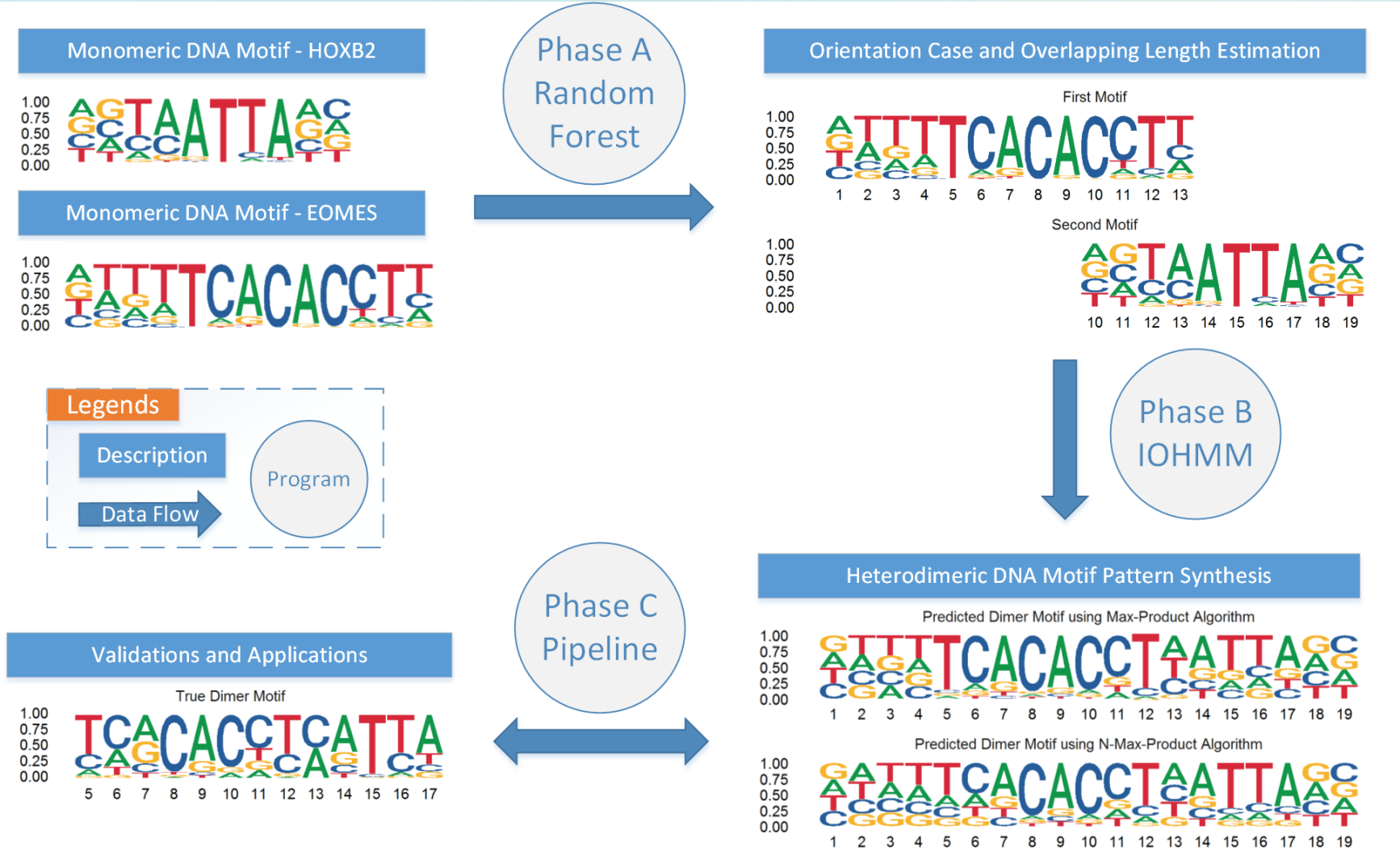
Split View | PDF | Cite | Permissions | Share ▾

Abstract

Bound by transcription factors, DNA motifs (i.e. transcription factor binding sites) are prevalent and important for gene regulation in different tissues at different developmental stages of eukaryotes. Although considerable efforts have been made on elucidating monomeric DNA motif patterns, our knowledge on heterodimeric DNA motifs are still far from complete. Therefore, we propose

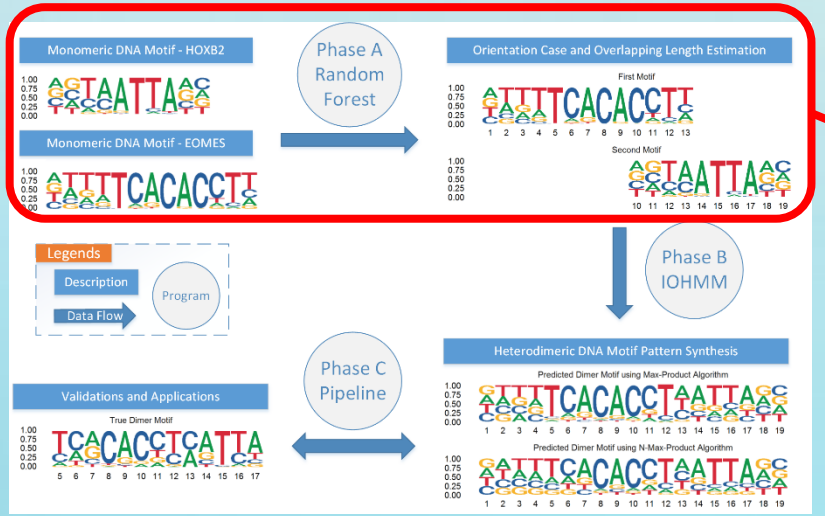
[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

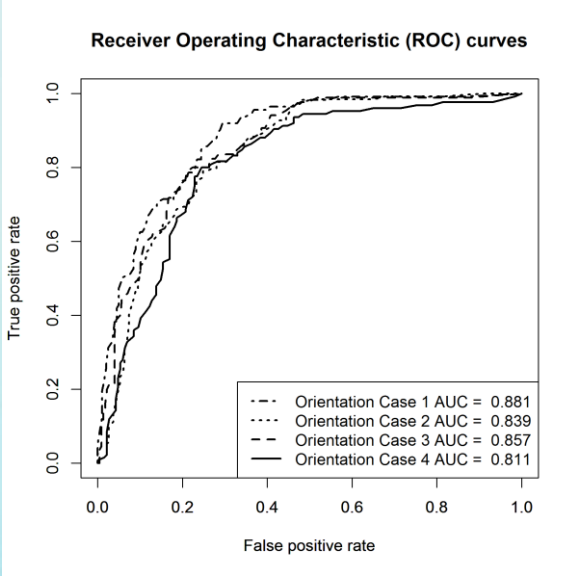
Case for synthesizing Heterodimeric DNA Motifs [1]

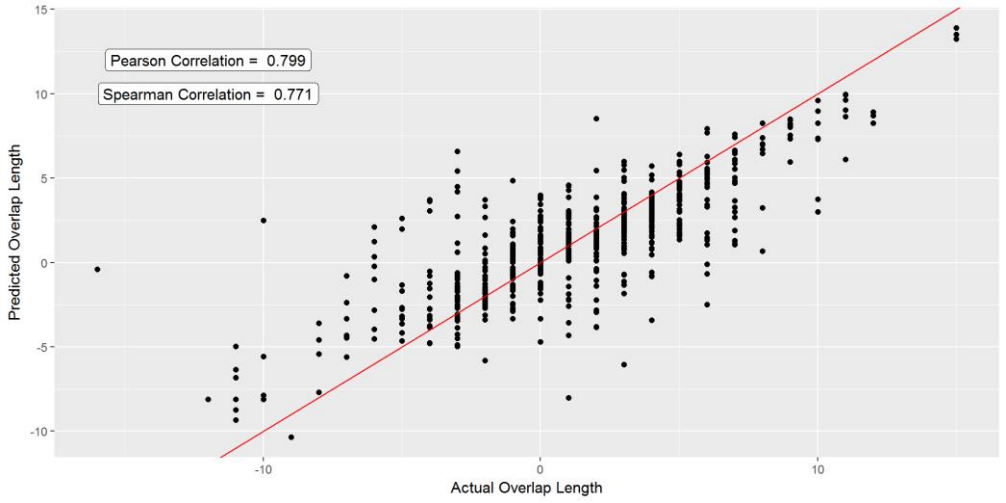


[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

AI for synthesizing Heterodimeric DNA Motifs [1]

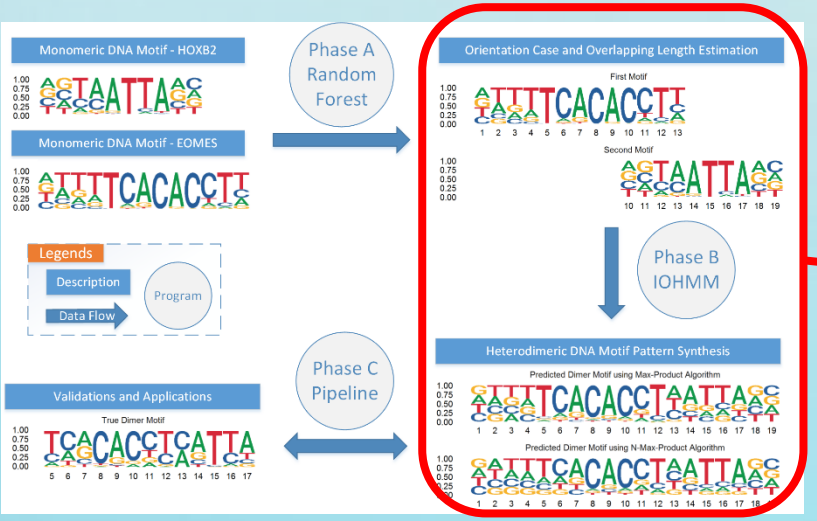


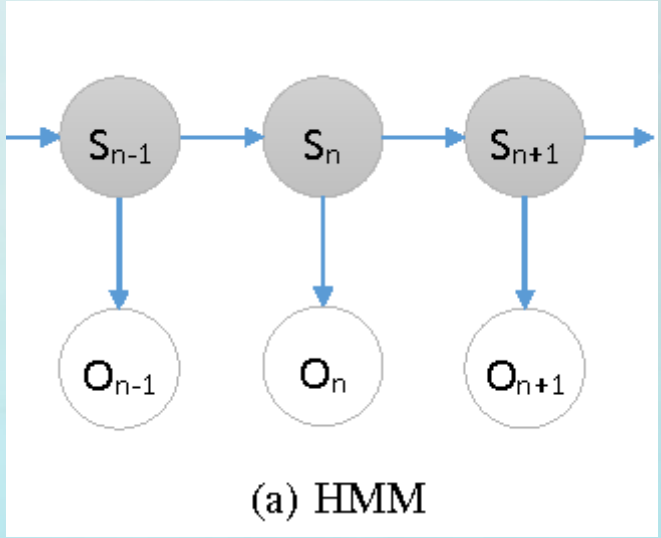




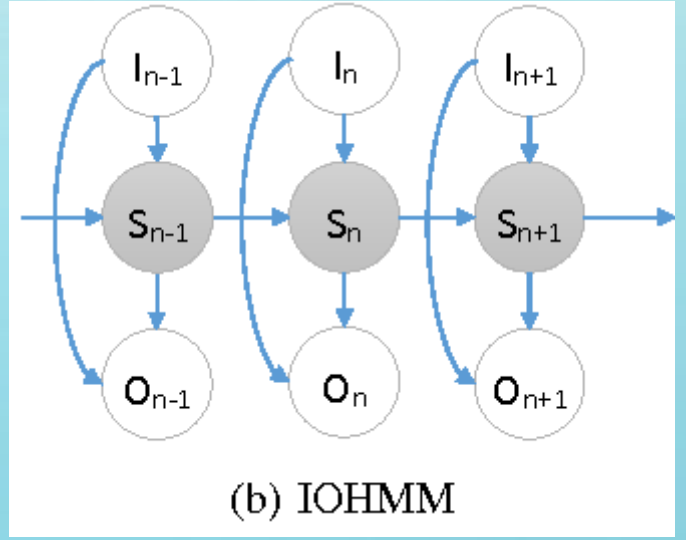
[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

AI for synthesizing Heterodimeric DNA Motifs [1]





(a) HMM

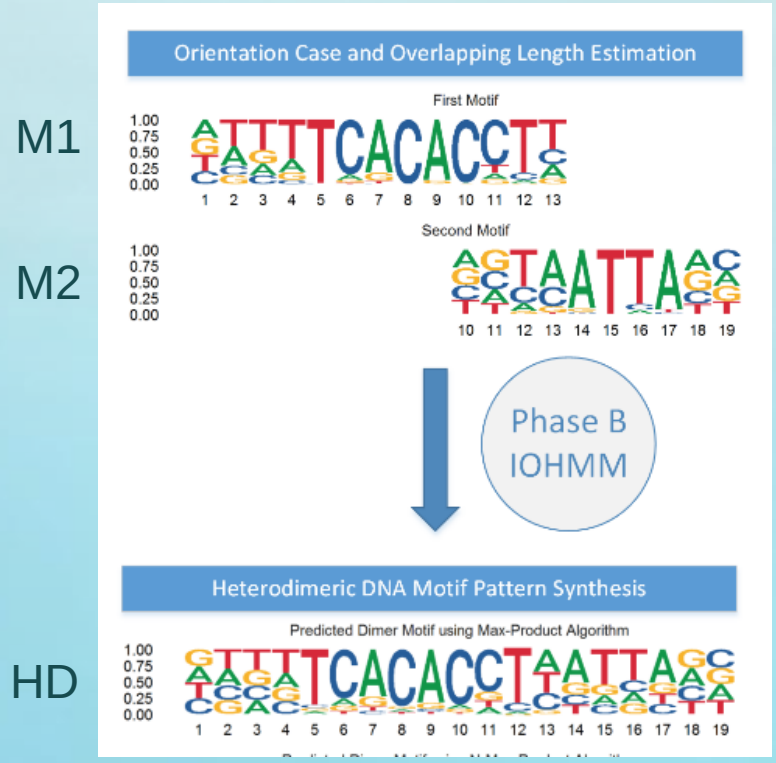


(b) IOHMM

[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

https://doi.org/10.1109/DSAA.2015.7344851

Mathematical Details of IOHMM



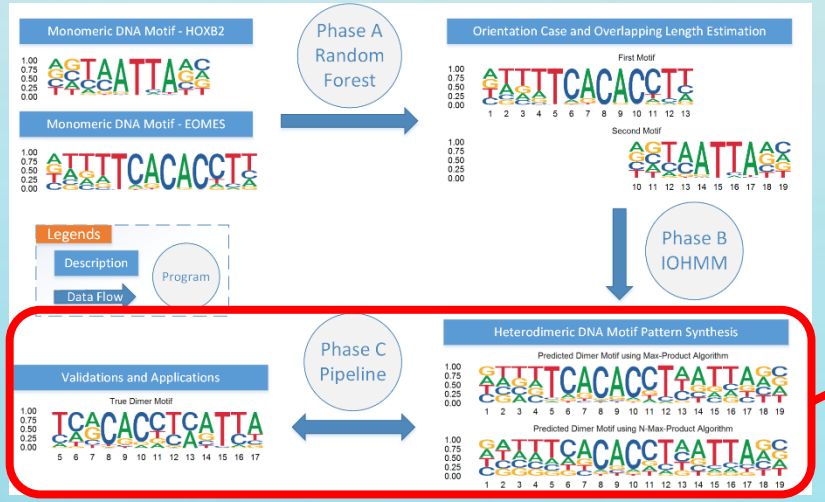
We propose to exploit the sequence information and make the Markov assumptions to infer *HD* from (*M*1, *M*2) using input-output hidden Markov models (IOHMMs) (28). Briefly, IOMMM is the generalized variant of generic hidden Markov model (HMM). The difference between IOHMM and HMM is that IOHMM takes into account inputs and outputs with hidden states while HMM only considers outputs with hidden states. Therefore, IOHMMs are chosen in this study since we have to consider the contributions from the input DNA motif model matrices (*M*1, *M*2). The IOHMM mathematical modeling $\Theta_{IO} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, f, g)$ with hidden states $k \in \{1, 2, \ldots, K\}$ is defined as follows:

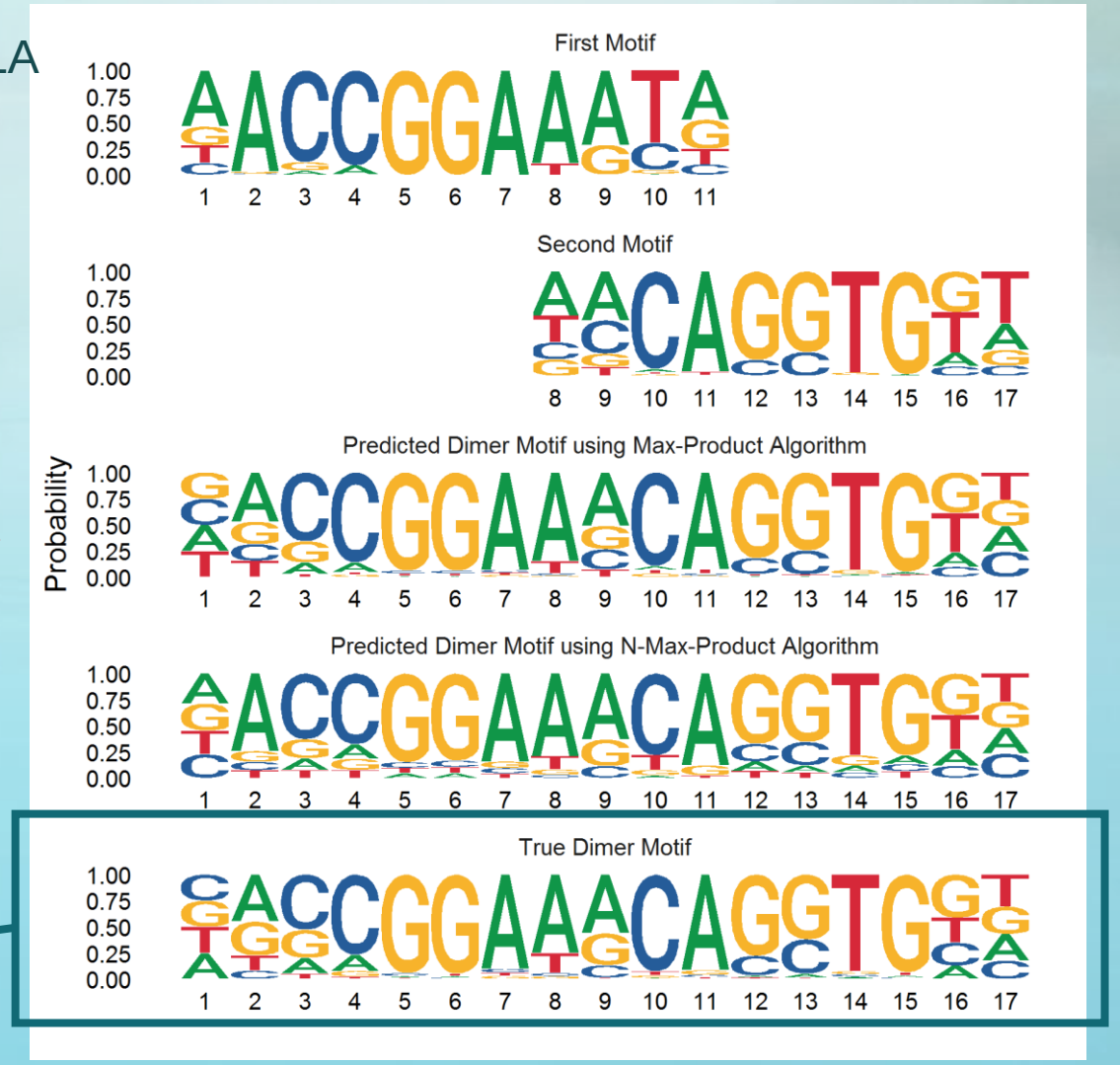
$$
\begin{aligned}
&HD[i,j] = A_{s_j}[i,j]M1[i,j] + B_{s_j}[i,j]M2[i,j] + C_{s_j}[i,j] \\
&P(s_j = k) = f(s_j, s_{j-1}, M1[:, j-1], M2[:, j-1], j-1) \\
&P(s_1 = k) = g(s_1, M1[:,1], M2[:,1]) \\
&s.t. \sum_{i=1}^{4} HD[i,j] = 1 \\
&\forall i \le 4, j \le N, k \le K \in \mathbb{N}
\end{aligned}
$$

where s_j denotes the hidden state variable at the j-th position; A_k is the regression coefficient matrix for *M*1 at state k in the set $\mathbf{A}$; B_k is the regression coefficient matrix for *M*2 at state k in the set $\mathbf{B}$; C_k is the bias coefficient matrix at state k in the set $\mathbf{C}$; the functions f and g denote the hidden state transition function and initial state estimation function respectively. In this study, given the data availability, we have adopted the classic logistic regression functions as f and g for computational efficiency (28).

AI for synthesizing Heterodimeric DNA Motifs [1]

ETS_bHLH_ETV2_FIGLA





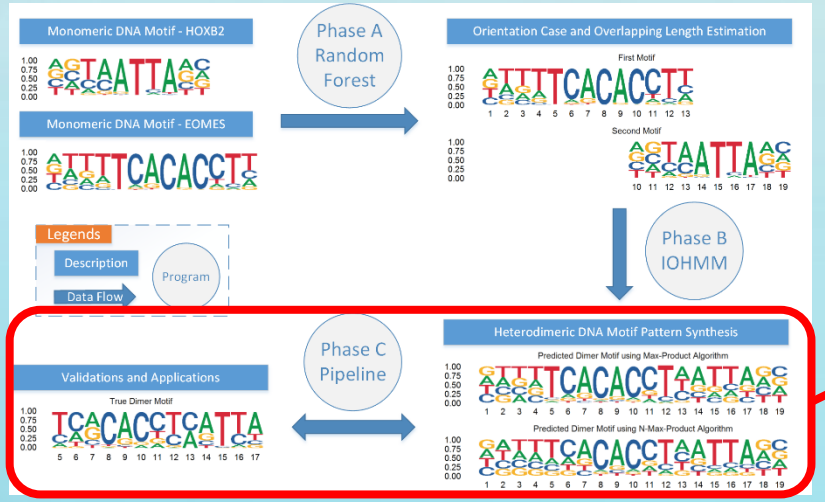
Ground Truth

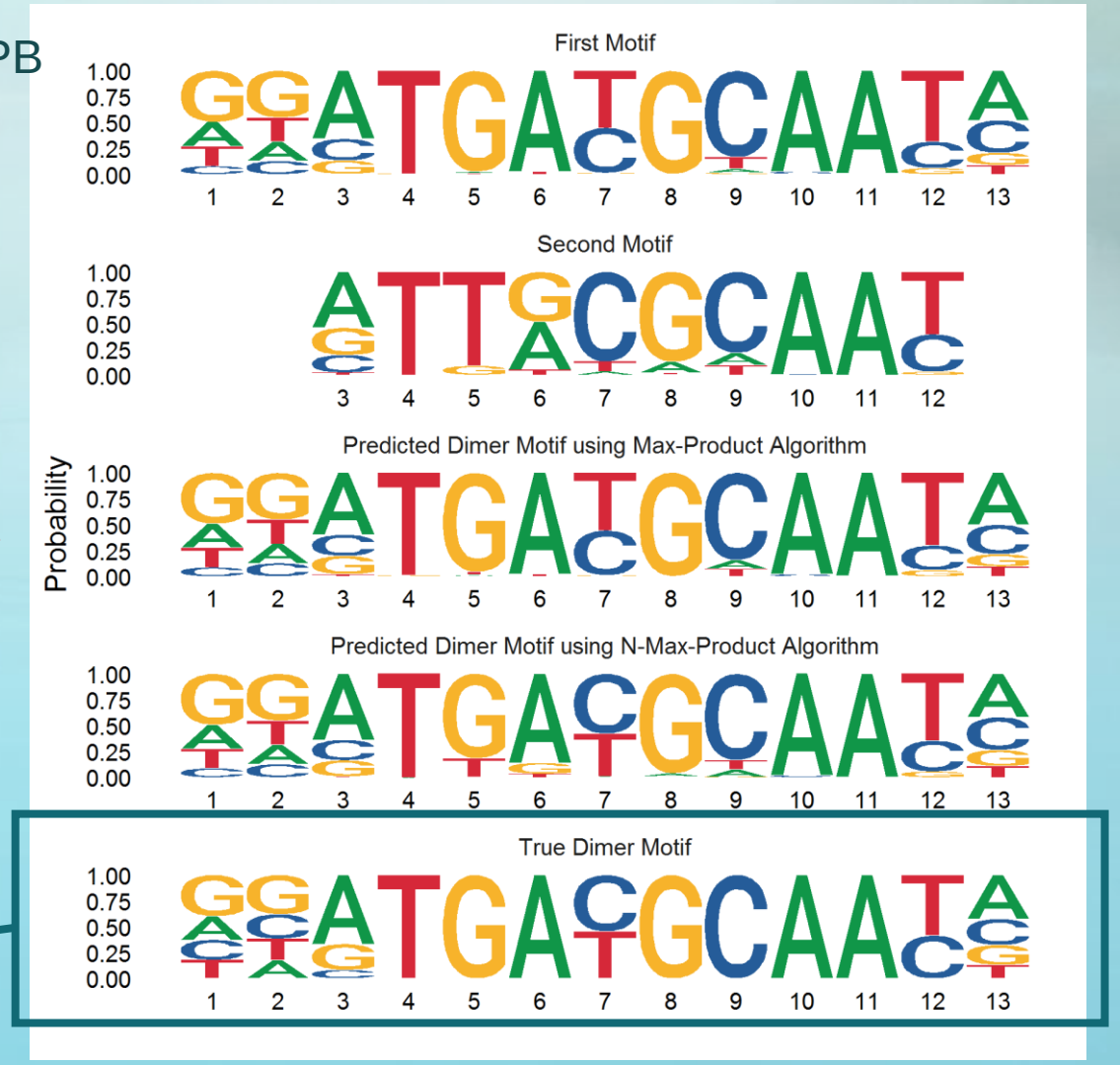
[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

https://doi.org/10.1109/DSAA.2015.7344851

AI for synthesizing Heterodimeric DNA Motifs [1]

bZIP_bZIP_ATF4_CEBPB





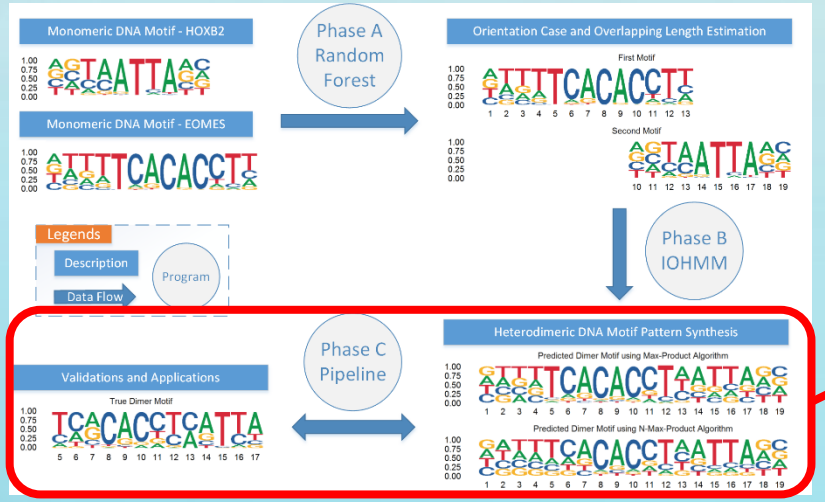
Ground Truth

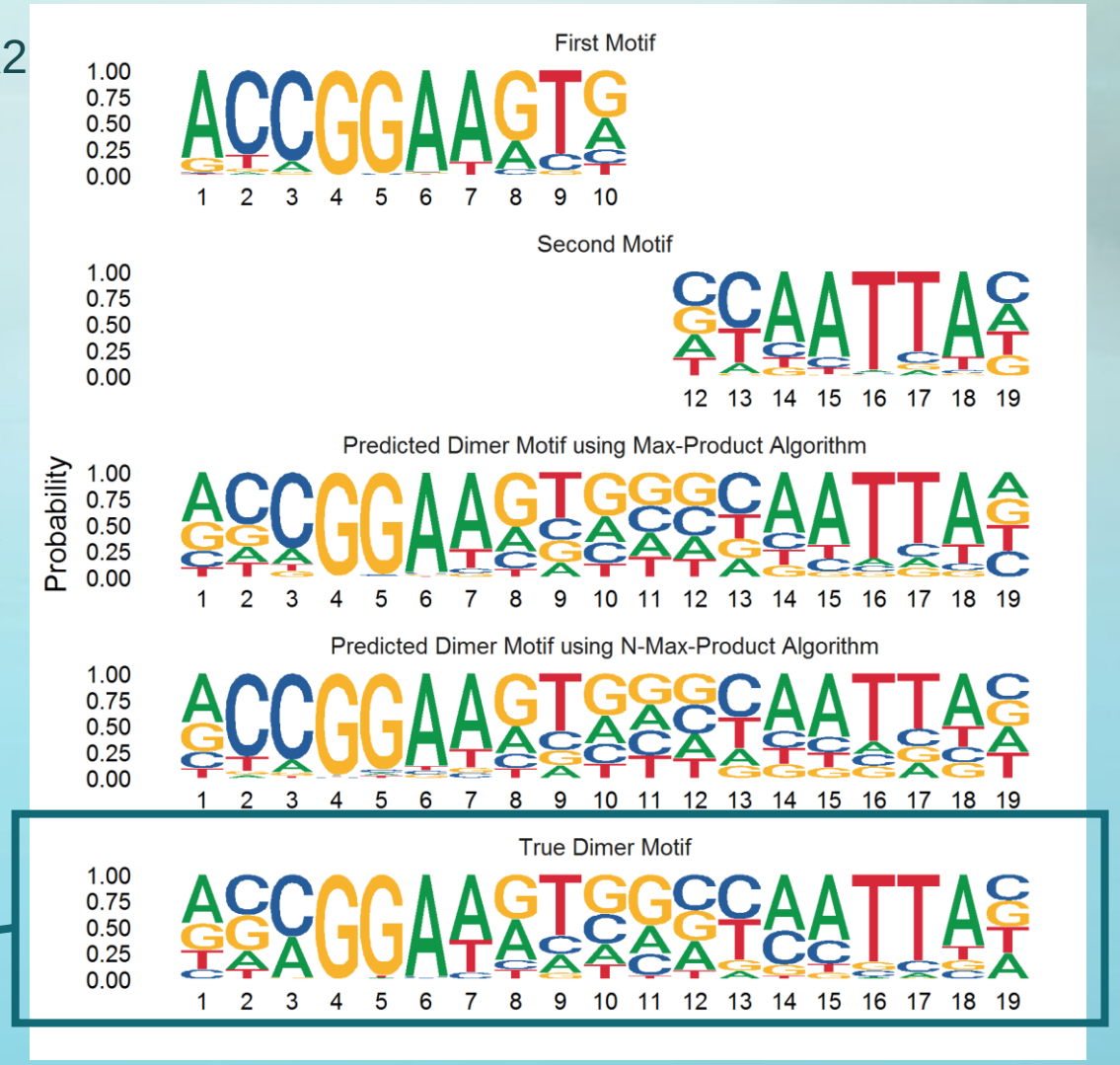
[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

https://doi.org/10.1109/DSAA.2015.7344851

AI for synthesizing Heterodimeric DNA Motifs [1]

ETS_Homeo_ERF_DLX2





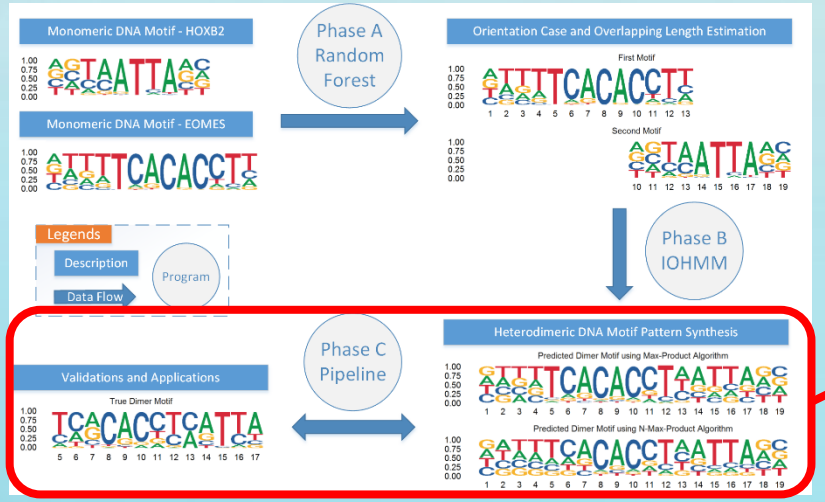
Ground Truth

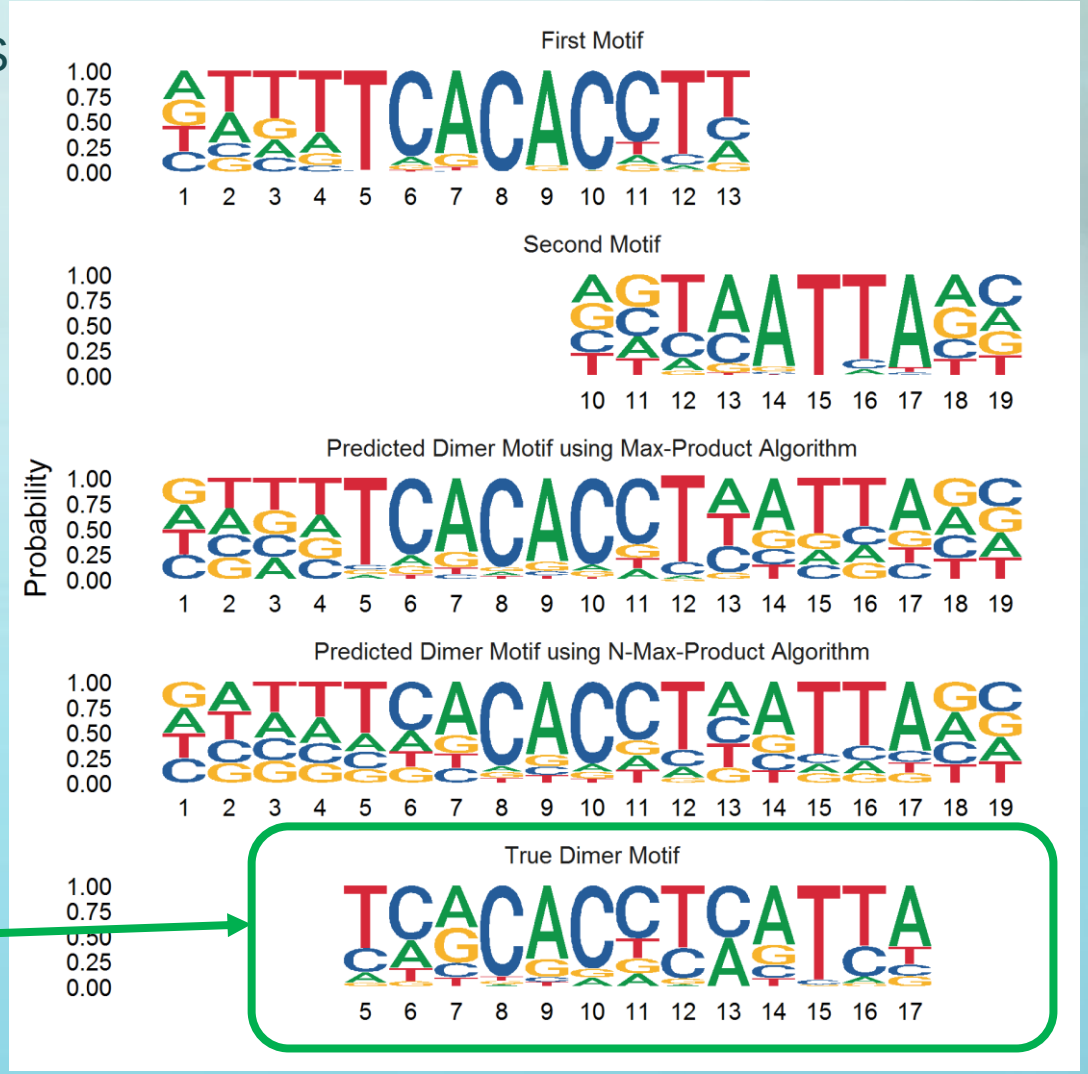
[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

https://doi.org/10.1109/DSAA.2015.7344851

AI for synthesizing Heterodimeric DNA Motifs [1]

Homeo_Tbox_HOXB2_EOMES



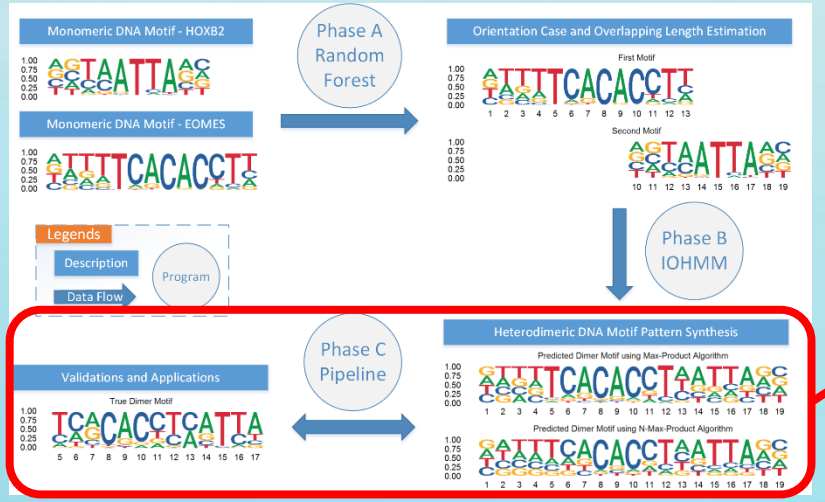


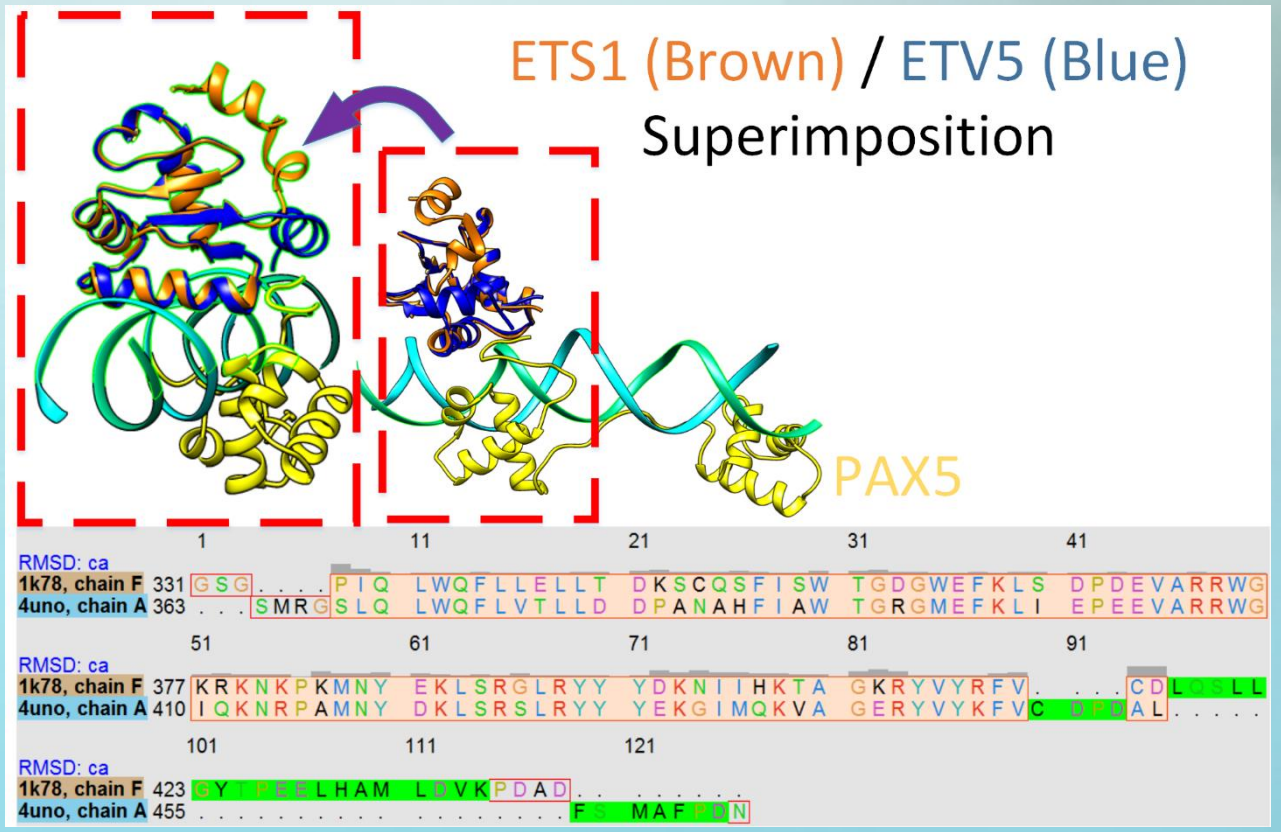
Published on Nature

[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

https://doi.org/10.1109/DSAA.2015.7344851

AI for synthesizing Heterodimeric DNA Motifs [1]



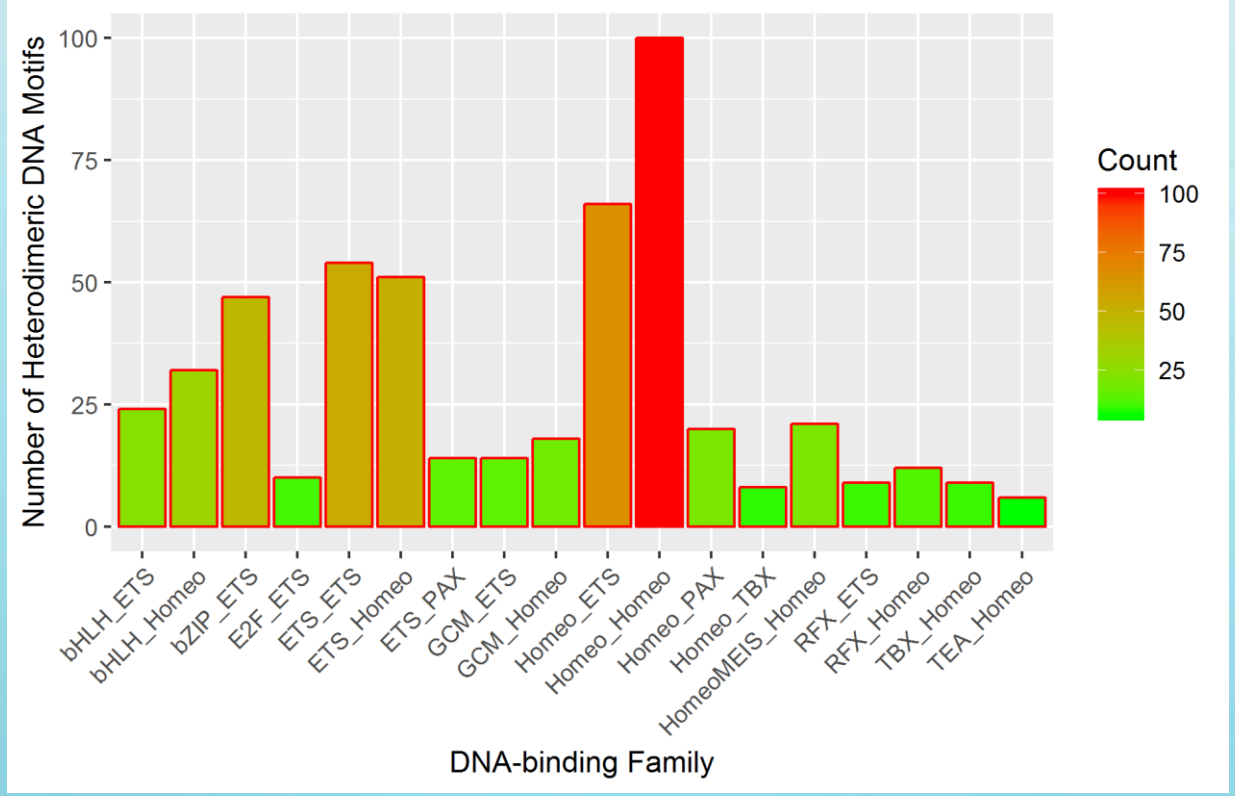


[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

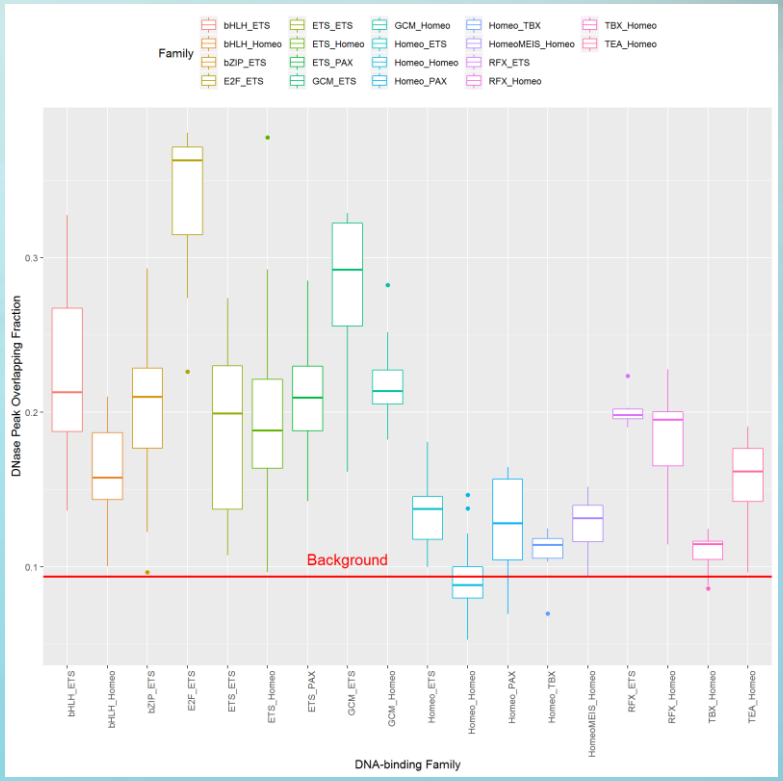
https://doi.org/10.1109/DSAA.2015.7344851

AI for synthesizing Heterodimeric DNA Motifs [1]

Lastly, we apply the proposed approach to infer previously uncharacterized heterodimeric motifs. Their motif instances are supported by DNase accessibility, gene ontology, protein-protein interactions, ChIP-seq peaks, and even structural data from PDB.



DNase Accessibility



[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic acids research, 47(4), 1628-1636. (2017 Impact Factor: 11.561)

Case for predicting CRISPR-Cas9 Off-targets [2,3]

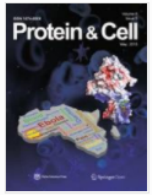
[2] Lin, J., & Wong, K. C. (2018). Off-target predictions in CRISPR-Cas9 gene editing using deep learning. Bioinformatics, 34(17), i656-i663.

[3] Lin, J., Zhang, Z., Zhang, S., Chen, J., & Wong, K. C. (2020). CRISPR-Net: A Recurrent Convolutional Network Quantifies CRISPR Off-Target Activities with Mismatches and Indels. Advanced Science, 1903562. (2019 Impact Factor =15.4)

Pattern Recognition on CRISPR-Cas9 Off-targets [3]

Springer Link

Search



Protein & Cell
May 2015, Volume 6, Issue 5, pp 363–372 | Cite as

CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes

Authors | Authors and affiliations

Puping Liang, Yanwen Xu, Xiya Zhang, Chenhui Ding, Rui Huang, Zhen Zhang, Jie Lv, Xiaowei Xie, Yuxi Chen, Yujing Li, Ying Sun, Yaofu Bai, Zhou Songyang, Wenbin Ma, Canquan Zhou, Junjiu Huang

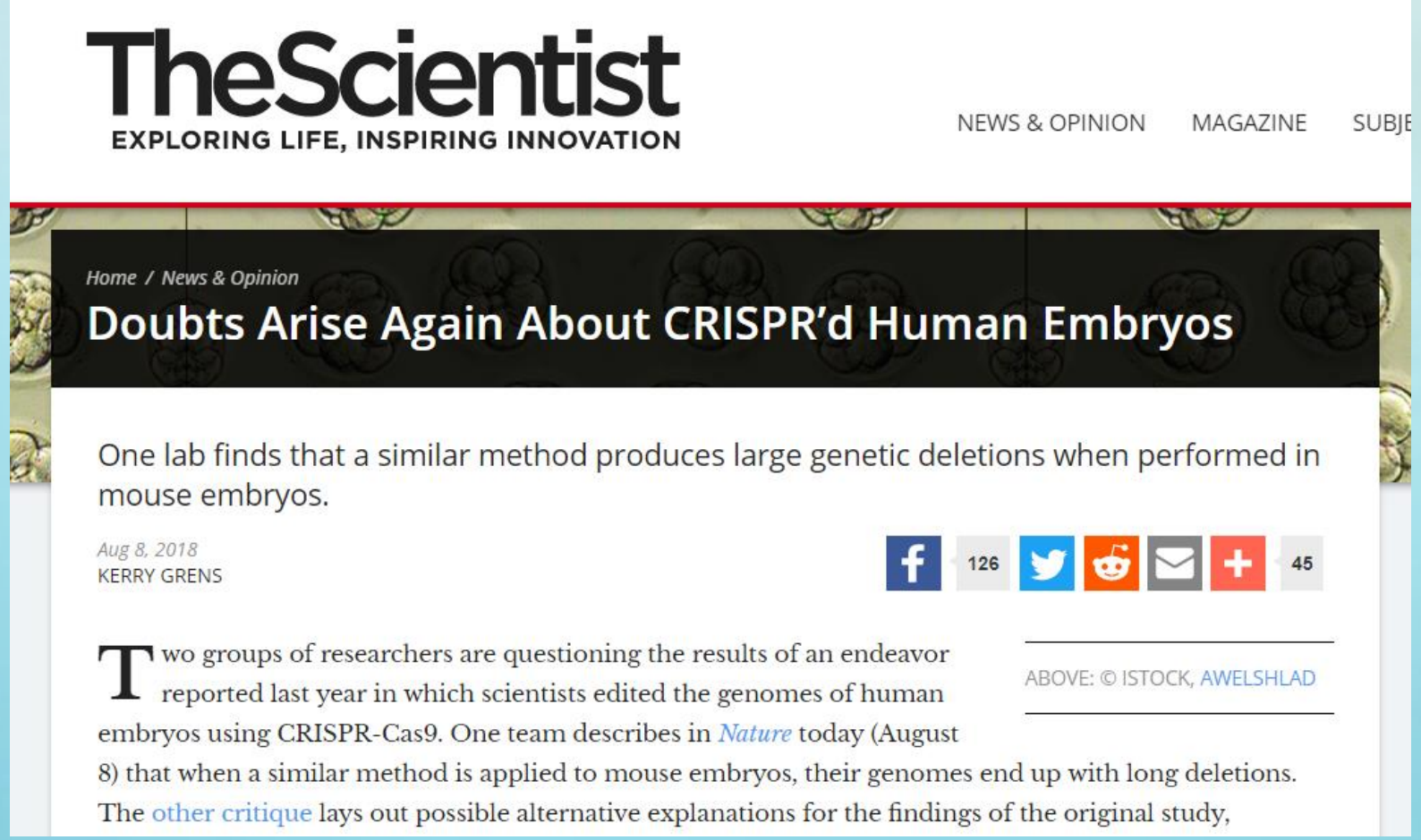
Open Access | Research article
First Online: 18 April 2015

909 Shares
209k Downloads
340 Citations

ABSTRACT

Genome editing tools such as the clustered regularly interspaced short palindromic repeat (CRISPR)-associated system (Cas) have been widely used to modify genes in model systems including animal zygotes and human cells, and hold tremendous promise for both basic research and clinical applications. To date, a serious knowledge gap remains in our

Pattern Recognition on CRISPR-Cas9 Off-targets [3]



TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

NEWS & OPINION MAGAZINE SUBJE

Home / News & Opinion

Doubts Arise Again About CRISPR'd Human Embryos

One lab finds that a similar method produces large genetic deletions when performed in mouse embryos.

Aug 8, 2018
KERRY GRENS

126 45

ABOVE: © ISTOCK, AWELSHLAD

Two groups of researchers are questioning the results of an endeavor reported last year in which scientists edited the genomes of human embryos using CRISPR-Cas9. One team describes in *Nature* today (August 8) that when a similar method is applied to mouse embryos, their genomes end up with long deletions. The other critique lays out possible alternative explanations for the findings of the original study,

https://www.the-scientist.com/news-opinion/doubts-arise-again-about-crisprd-human-embryos-64625

SECOND INTERNATIONAL SUMMIT ON HUMAN GENOME EDITING

Nobel Price in Chemistry 2020

Heroes of Chemistry and Nobel Laureates | First published: **9 October 2020** | Last updated: **12 October 2020**

The 2020 Nobel Prize in Chemistry was awarded jointly to Emmanuelle Charpentier and Jennifer A. Doudna for the development of a method for genome editing: the CRISPR/Cas9 system. This technological revolution propelled advances not only in basic research in the life sciences, but also in other fields such as medicine, materials science, and biotechnology. To honor the achievements of the laureates, and to show how this discovery has affected progress in various fields, *ChemBioChem*, *Angewandte Chemie*, *Small*, *Small*

Pattern Recognition on CRISPR-Cas9 Off-targets [3]

Nucleic Acids Research, 2014 1
doi: 10.1093/nar/gku402

CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences

Yanni Lin[1], Thomas J. Cradick[1], Matthew T. Brown[1], Harshavardhan Deshmukh[1], Piyush Ranjan[2], Neha Sarode[2], Brian M. Wile[1], Paula M. Vertino[3], Frank J. Stewart[2] and Gang Bao[1,*]

[1]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA, [2]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA and [3]Department of Radiation Oncology, Emory University School of Medicine, Atlanta, GA 30322, USA

CRISPR-Cas9 Off-targets

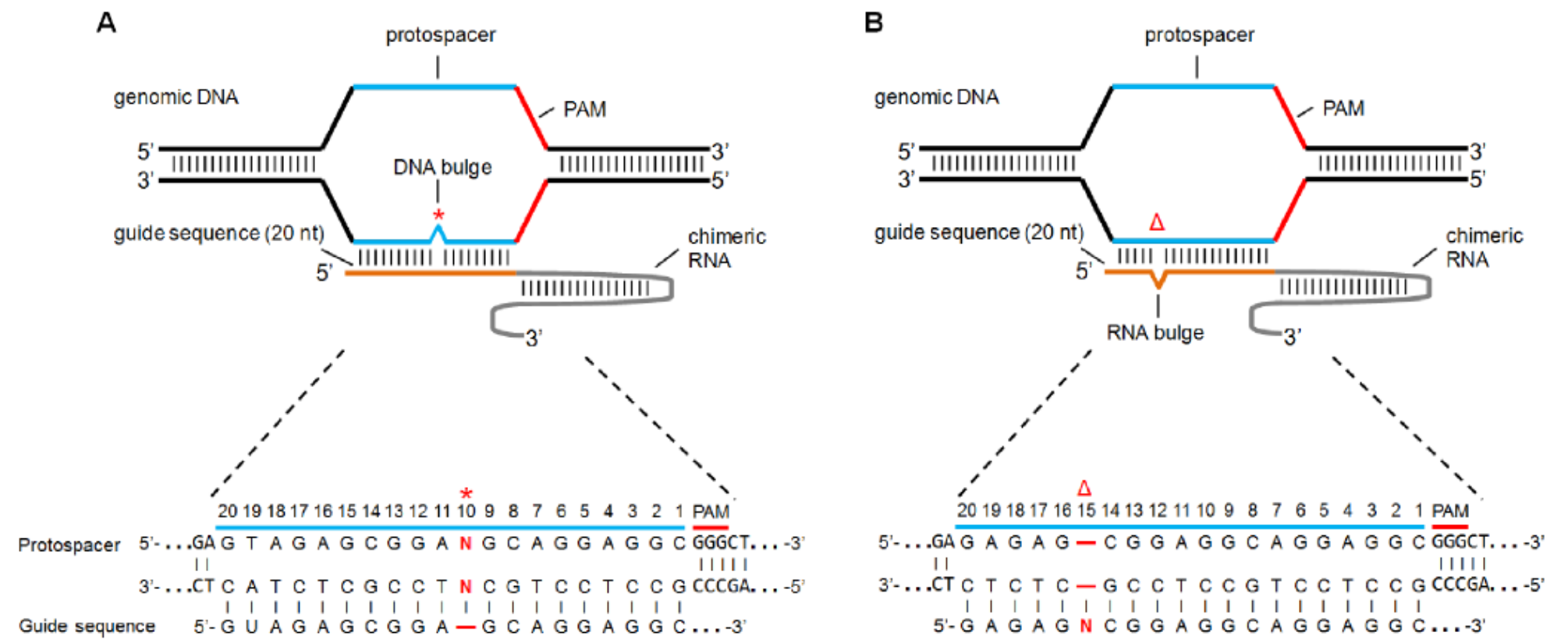


Figure 1. Schematic of CRISPR/Cas9 off-target sites with (**A**) 1-bp insertion (DNA bulge) or (**B**) 1-bp deletion (RNA bulge). The 20-nt guide sequence (orange line) in the sgRNA is shown with genomic target sequence (protospacer) containing single-base DNA bulge (red asterisk) or single-base sgRNA bulge (red Δ). The zoom-in sequences of protospacer and PAM are shown above the sgRNA guide sequence. Positions of nucleotides in the target are numbered 3′ to 5′ starting from the nucleotide next to PAM.

Lin, Yanni, et al. "CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences." *Nucleic acids research* 42.11 (2014): 7473-7485.

Motivation

Although specific fragments of DNA are aimed, sgRNA can sometimes influence other regions and incur off-target mutations (Chen et al. 2017). CRISPR-Cas9 can tolerate mismatches in sgRNA-DNA at different positions in a sequence-dependent manner, sensitive to the number, position, and distribution of mismatch (Hsu et al. 2013 and Zhang et al. 2015). Since sgRNA is able to endure some mismatches across several nucleotide positions, many off-target sites could be found on the target genome (Kim et al. 2015).

Off-targets can lead to genomic instability and disturb the normal gene function, which is still a major problem when applying CRISPR-Cas9 gene editing to clinical applications Consequently, we still need accurate off-target prediction methods for complementary purposes.

Motivation

Most of the existing off-target prediction methods just calculate scores based on the positions of the mismatches to the guide sequence The score of each base pair in sgRNA-DNA is derived using classic bioinformatics statistics on the mismatch effects based on previous gene editing experiments. For example, CFD (Cutting Frequently Determination) score is derived by infecting a large number of sgRNAs with single-bp replacement, deletion or insertion corresponding to the validated sgRNAs in MOLM13 cells; it calculates the percentage activity rates of different mutation sites based on LFC (Log Fold Change) value (Doench et al. 2016).

In light of the above, their performance are vulnerable to experimental variation. Most importantly, the existing methods can not take advantage of the growing CRISPR-Cas9 data for constant self-learning. In addition, most of the existing methods do not consider the potential relationships between mismatched and matched sites, which may be relevant to the off-target activity in CRISPR-Cas9 gene editing (Xu et al. 2017).

Proposed Approach (Deep Neural Network)

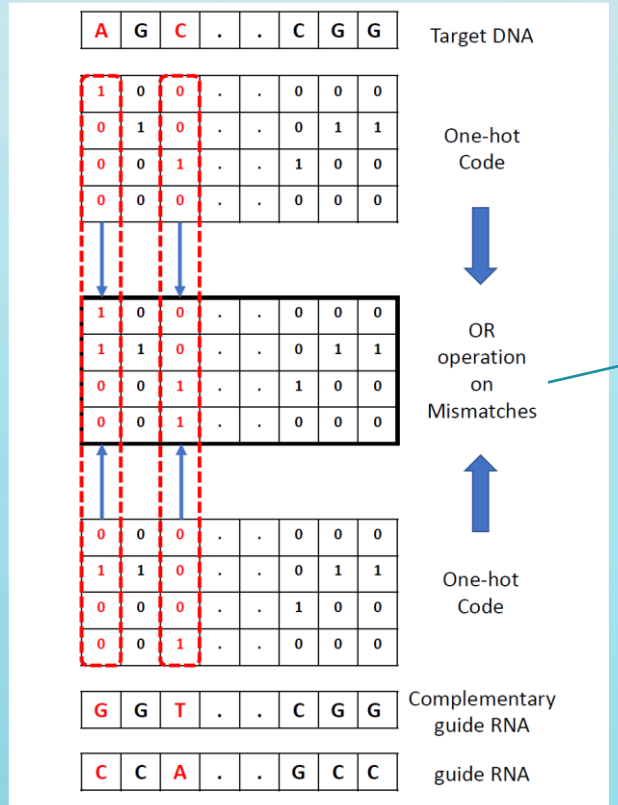


Fig. 1. An example on how to encode a sgRNA-DNA sequence pair. The table with thick borders in the middle of the figure shows the final matrix code of a sgRNA-DNA sequence pair, which can be used as the input for convolutional neural network modelling.

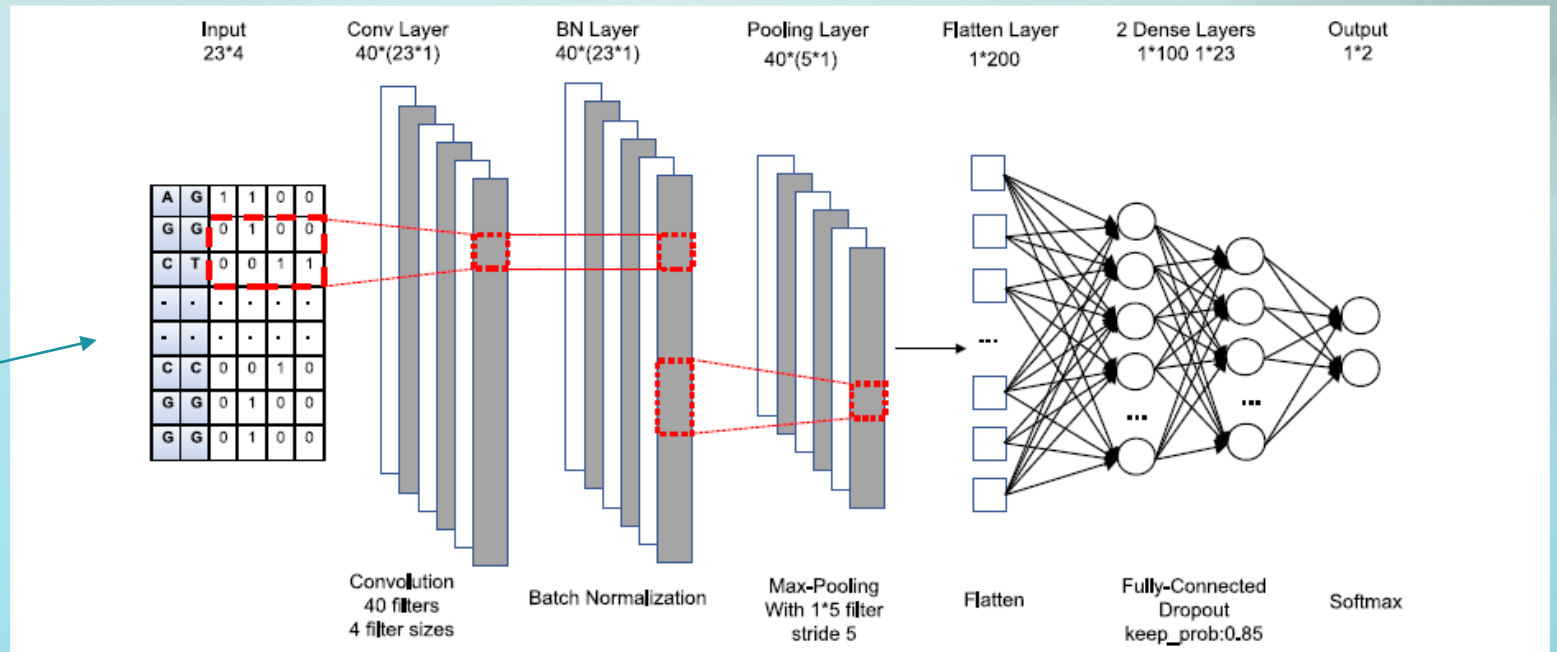


Fig. 2. The architecture of standard deep convolutional neural network (CNN_std) for off-target prediction. The input of this deep neural network is the encoded sgRNA-DNA sequence with length 23, which are converted into hidden features by one convolutional layer. The convolutional layer consists of 40 filers including 10 for each of the sizes 4×1, 4×2, 4×3 and 4×5. The BN layer is used to normalise the output of the convolutional layer in order to make the training process faster and avoid over-fitting. The global max-pooling layer applies a filter of window size 5 to the previous layers. The output of max-pooling layer are joined together into one vector by flattening. Each neurons in the flatten layer is fully connected to the first dense layer. Two dense layer consists of 100 and 23 neurons respectively. The second dense layer with a drop-out layer is fully connected to two output neurons to predict the probability of binary class. The neurons in output layer and dense layers use softmax function as the activation function, while all the neurons in other layers use ReLU as the activation function.

Performance Comparison on CRISPOR data

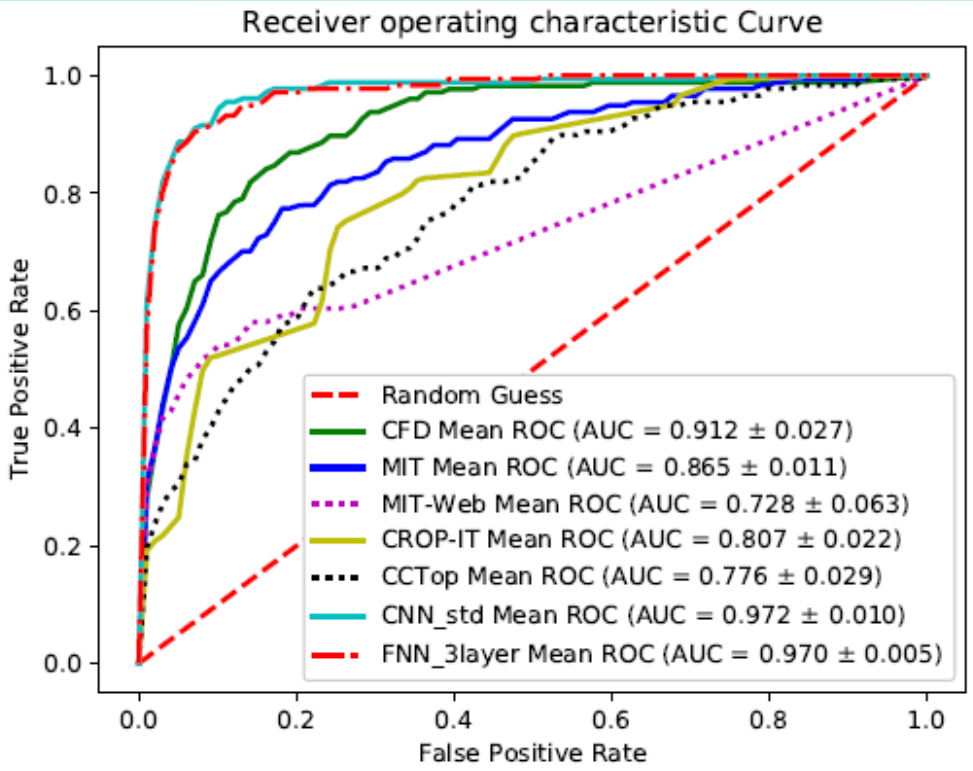


Fig. 3. ROC curves of two deep learning models and five current prediction methods under stratified 5-fold cross-validation on CRISPOR dataset.

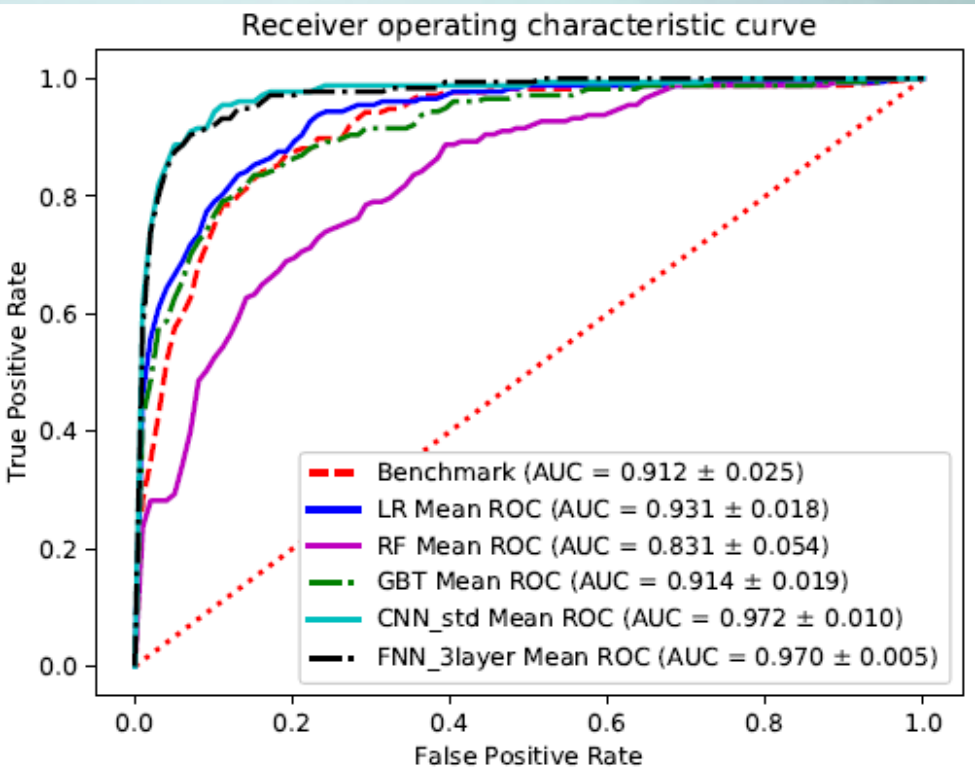


Fig. 4. ROC curves of two deep learning models (i.e. FNN_3layer and DNN_std) and three traditional machine learning models including logistic regression (LR), random forest (RF) and gradient boosting trees (GBT). The ROC curve and AUC value of CFD score were regarded as the state-of-arts benchmark.

Performance Comparison on GUIDE-seq data

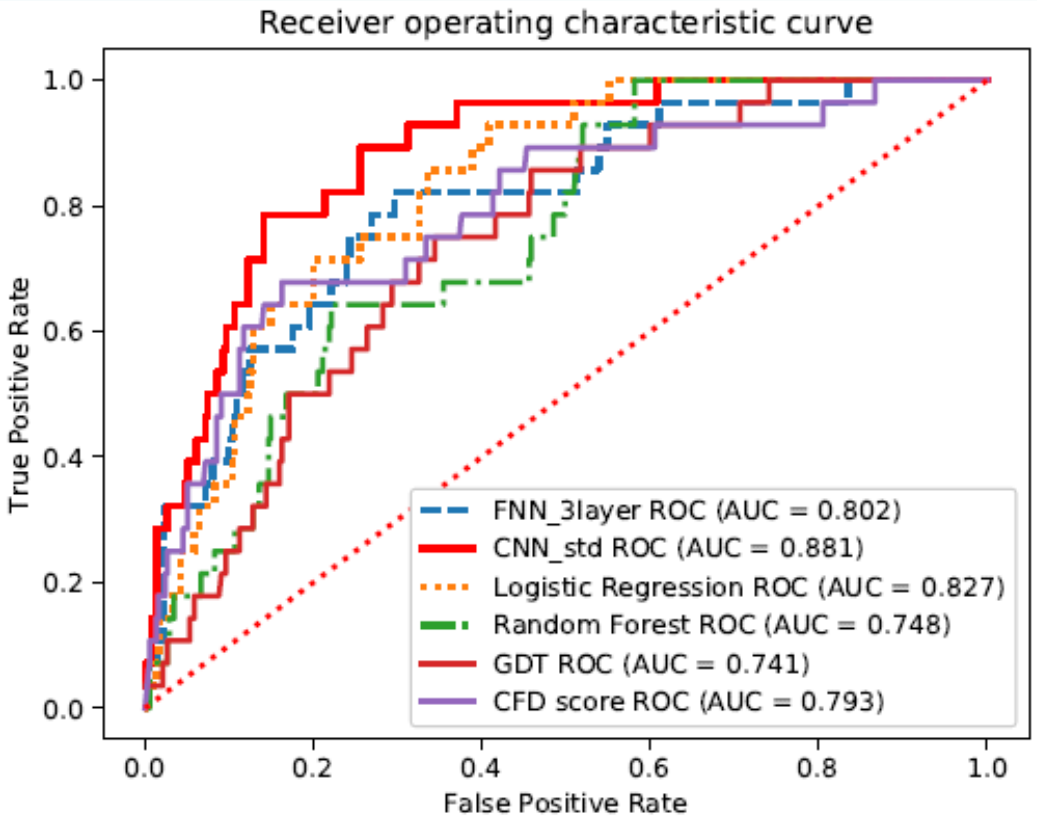


Fig. 5. ROC curves of deep learning models, CFD score and three traditional machine learning models on GUIDE-seq dataset.

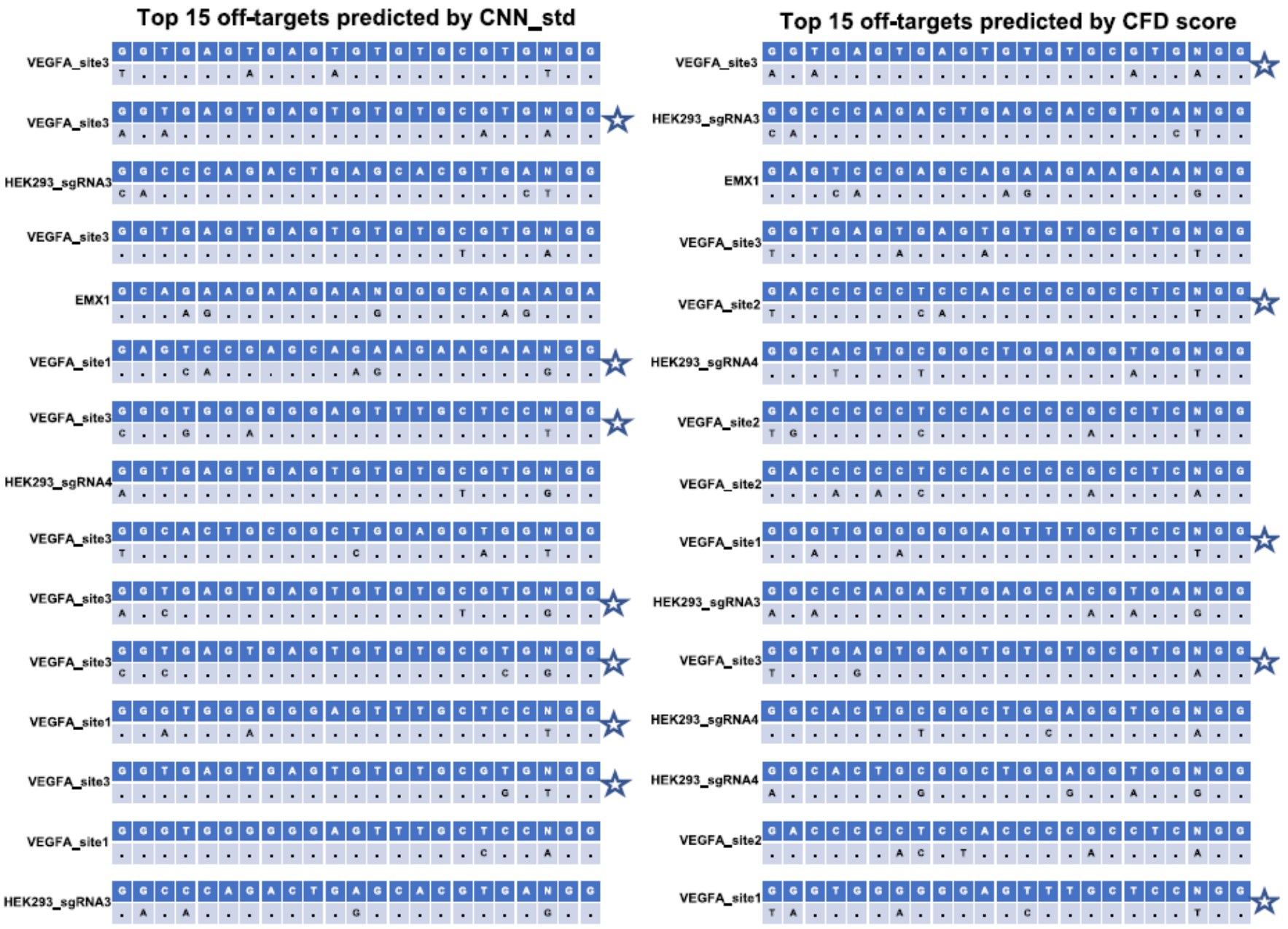


Fig. 6. 15 off-targets with the highest score predicted by final convolutional neural network and CFD score respectively on GUIDE-seq dataset. The sgRNA-DNA sequence marked with star is the true off-target.

Pattern Recognition on CRISPR-Cas9 Off-targets [2]

- We presented that deep convolutional neural networks are able to accurately predict the off-targets of CRISPR-Cas9 gene editing. To our knowledge, this is the first time that deep neural networks are designed and implemented for off-target predictions.
- Our final convolutional neural network obtained the best performance on both CRISPOR dataset and GUIDE-seq dataset, outperforming the current state-of-art off-target prediction methods and three traditional machine learning algorithms including logistic regression, random forest, and gradient boosting trees. We discussed and attributed its performance successes to the neural network layer designs which are general enough to self-learn and capture sequence features by itself.
- We believe that such intelligent approaches can contribute to CRISPR-Cas9 off-target predictions or other similar problems in a rigorous manner.

ADVANCED SCIENCE
Open Access

Full Paper | Open Access

2019 IF: 15.84

CRISPR-Net: A Recurrent Convolutional Network Quantifies CRISPR Off-Target Activities with Mismatches and Indels

Jiecong Lin, Zhaolei Zhang, Shixiong Zhang, Junyi Chen, Ka-Chun Wong

First published:20 May 2020 | https://doi.org/10.1002/advs.201903562

SECTIONS PDF TOOLS SHARE

Abstract

The off-target effects induced by guide RNAs in the CRISPR/Cas9 gene-editing system have raised substantial concerns in recent years. Many in silico predictive models have been developed for predicting the off-target activities; however, few are capable of predicting the off-target activities with insertions or deletions between guide RNA and target DNA sequence pair. In order to fill this gap, a recurrent convolutional network named CRISPR-Net is developed for scoring the gRNA-target pairs with mismatches and indels; and a machine-learning based model named CRISPR-Net-Aggregate is also developed for aggregating the scores as the consensus off-target score for each potential guide RNA. It is demonstrated that CRISPR-Net achieves competitive performance on CIRCLE-Seq and GUIDE-seq datasets with indels and mismatches, outperforming the state-of-the-art off-target prediction methods on two independent mismatch-only datasets. The CRISPR-Net-Aggregate also surpasses a competing method on the aggregation task. Moreover, a two-stage sensitivity analysis is introduced to visualize the CRISPR-Net prediction on the gRNA-target pair of interest, demonstrating how implicit knowledge encoded in CRISPR-Net contributes to the accurate off-target activity quantification. Finally, the source code is made available at the Code Ocean repository (https://codeocean.com/capsule/9553651/tree/v1).

Early View
Online Version of Record before inclusion in an issue
1903562
This article also appears in:
Hot Topic: Artificial Intelligence and Machine Learning

Figures References Related Information

Metrics

Am score 2

Details

© 2020 The Authors. Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

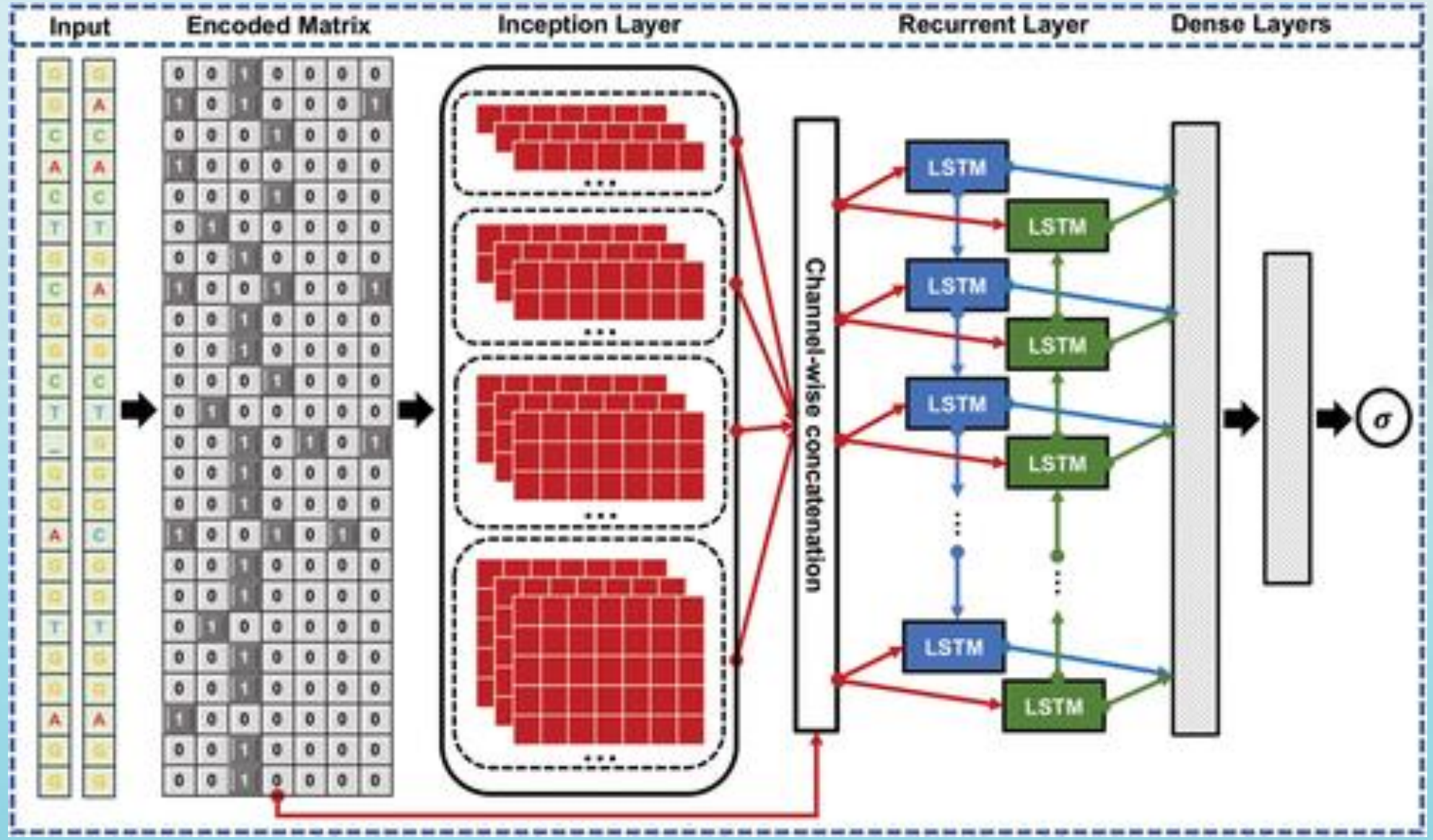
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Check for updates

Keywords

CRISPR systems deep learning off-target activities

Funding Information



Jiecong, Lin. , Zhaolei, Zhang. , Shixiong, Zhang. , Junyi, Chen. & **Ka-Chun, Wong*.** (*Corresponding Authorship 2020). CRISPR-Net: A Recurrent Convolutional Network quantifies CRISPR Off-target Activities with Mismatches and Indels. *Advanced Science*. 2019 IF: 15.84

Case for screening Early Cancers from Blood [4]

[4] Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Case for screening Early Cancers from Blood [4]

iScience

CellPress

Volume 15, 31 May 2019, Pages 332-341

Article

Early Cancer Detection from Multianalyte Blood Test Results

Ka-Chun Wong [1,7], Junyi Chen [1], Jiao Zhang [1], Jiecong Lin [1], Shankai Yan [1], Shxiong Zhang [1], Xiangtao Li [2], Cheng Liang [3], Chengbin Peng [4], Qiuzhen Lin [5], Sam Kwong [1], Jun Yu [6]

Show more

https://doi.org/10.1016/j.isci.2019.04.035

Get rights and content

Under a Creative Commons license

open access

Highlights

- We propose an approach (CancerA1DE) to detect early cancers from blood
- CancerA1DE doubles the existing sensitivity for the stage I cancer detection
- For stage II cancers, it can reach up to 90% across multiple cancer types
- The related software is opened and released for future follow-up works

[4] Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Motivation

World Health Organization

GLOBAL

Cancer Profile 2020

BURDEN OF CANCER

Total population (2019)

7,676,965,500

Total # cancer cases (2018)	Total # cancer deaths (2018)
18,078,957	9,555,027

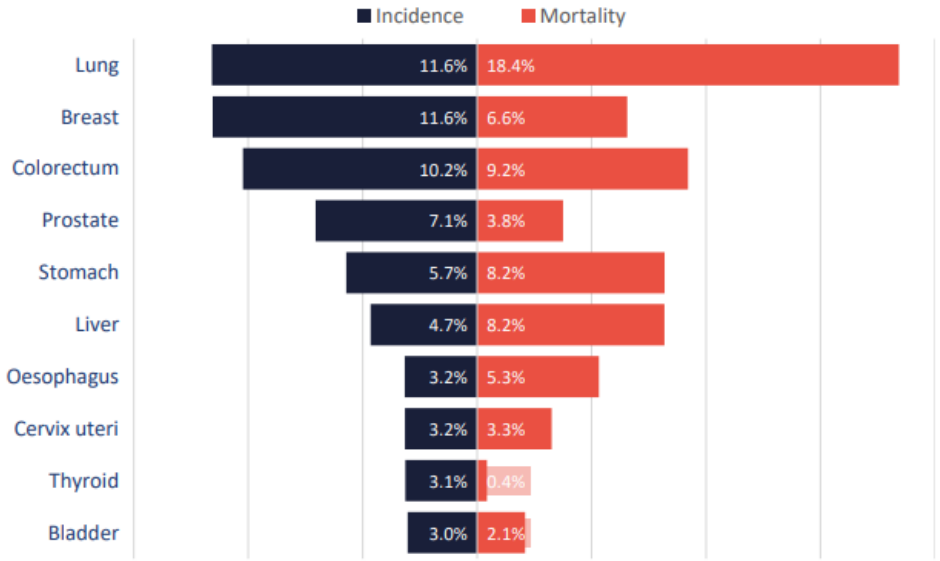
Premature deaths from NCDs (2016)

15,179,108

Cancer as % of NCD premature deaths (2016)

29.7%

Most common cancer cases (2018)

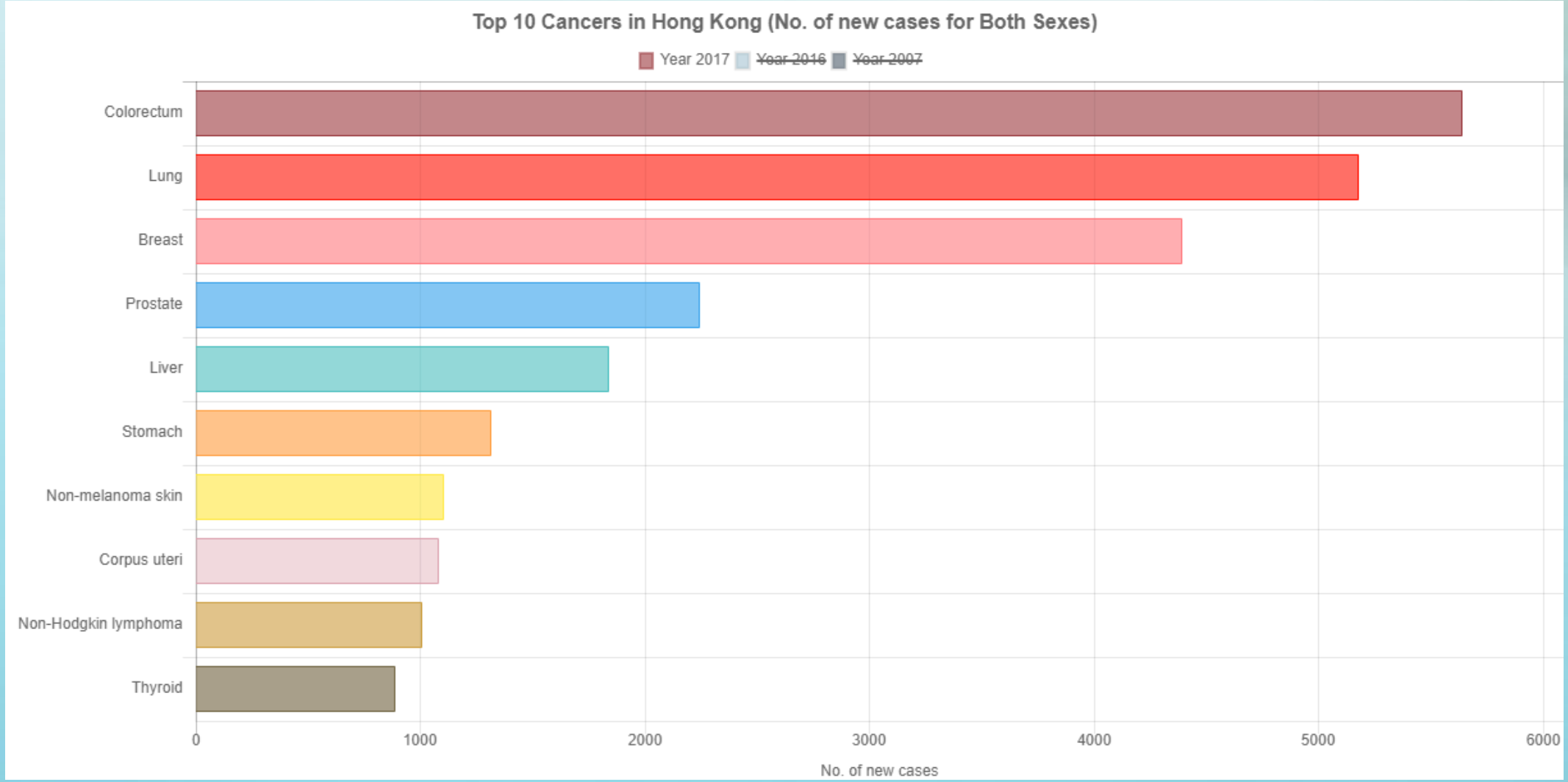


PAFs (population attributable fractions)

25.0%	4-5%	13.0%	3-4%	1.0%	2-8%
Tobacco (2017)[a]	Alcohol (2016)[a]	Infections (2012)[b]	Obesity (2012)[b]	UV (2012)[c]	Occupational risk (2017)[a]

[a] *PAF, cancer deaths* [b] *PAF, cancer cases* [c] *PAF, melanoma cases*

Motivation



https://www3.ha.org.hk/cancereg/top10incidence.html

Background

AAAS Become a Member

Science Contents News Careers Journals

Read our COVID-19 research and news.

SHARE

REPORT

Detection and localization of surgically resectable cancers with a multi-analyte blood test

Joshua D. Cohen[1,2,3,4,5], Lu Li[6], Yuxuan Wang[1,2,3,4], Christopher Thoburn[3], Bahman Afsari[7], Ludmila Danilova...

+ See all authors and affiliations

Science 23 Feb 2018:
Vol. 359, Issue 6378, pp. 926-930
DOI: 10.1126/science.aar3247

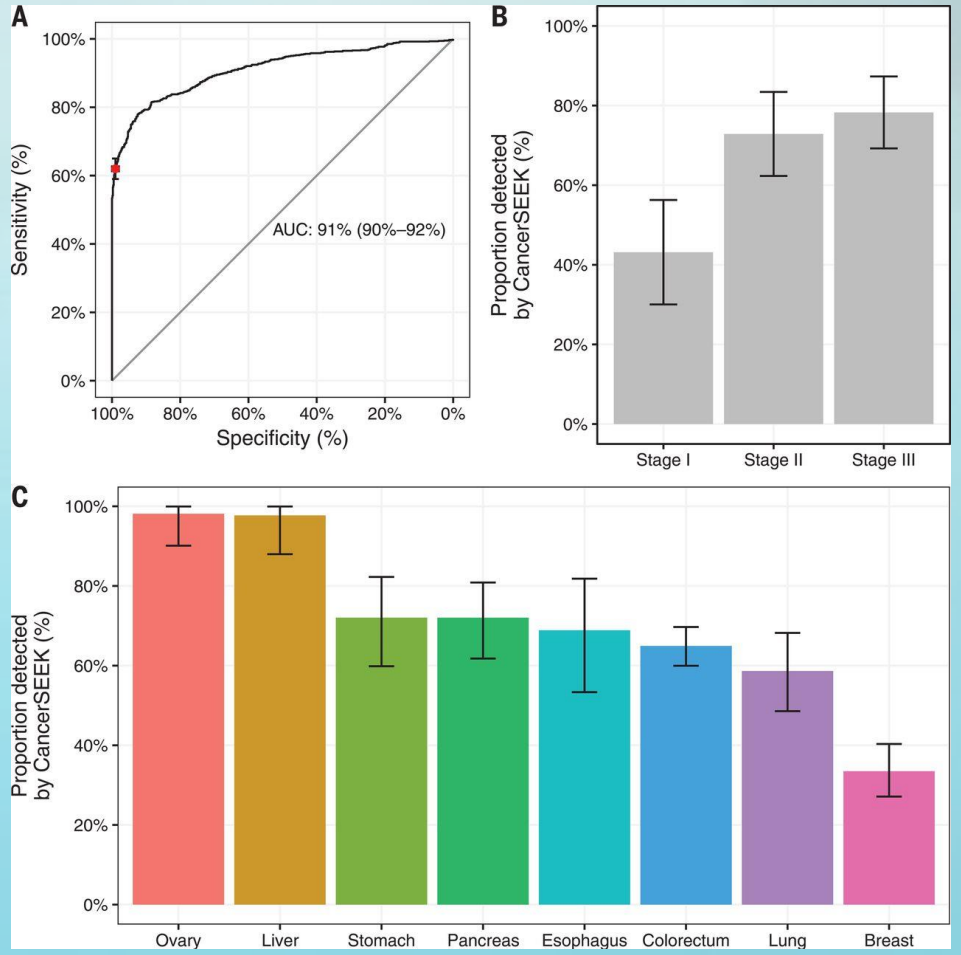
Background

Performance of CancerSEEK [1]

(A) ROC curve for CancerSEEK. The red point on the curve indicates the test's average performance (62%) at >99% specificity. Error bars represent 95% confidence intervals for sensitivity and specificity at this particular point. The median performance among the eight cancer types assessed was 70%.

(B) Sensitivity of CancerSEEK by stage. Bars represent the median sensitivity of the eight cancer types, and error bars represent standard errors of the median.

(C) Sensitivity of CancerSEEK by tumor type. Error bars represent 95% confidence intervals.



[1] Cohen, Joshua D., et al. "Detection and localization of surgically resectable cancers with a multi-analyte blood test." *Science* 359.6378 (2018): 926-930.

Background

However, we note three limitations of CancerSEEK [1]:

1. Its front-line cancer detection component is based on logistic regression, whereby linear assumption on different markers is hardly realistic.
2. Its second-line cancer type localization component is based on random forest, a modeling known to be difficult for interpretations.
3. From the user perspective, its lack of public Web service also limits its potential impacts.

[1] Cohen, Joshua D., et al. "Detection and localization of surgically resectable cancers with a multi-analyte blood test." *Science* 359.6378 (2018): 926-930.

Data Collection

- We have collected the multianalyte blood test data from Cohen et al. (2018). Those data have been processed according to the supplementary guideline provided, resulting in two datasets.
- The first dataset has 1,817 patient blood test records, which are designed and adopted to build models to detect cancers as the front-line detector in a binary manner (i.e., cancer or normal). Therefore, to be scalable and economical, it has the minimal number of input feature information involving eight circulating protein marker concentrations and one cell-free DNA mutation score (OmegaScore) as listed in the following table.
- The second dataset has 626 patient blood test records, which are designed and adopted to build models to localize cancer types as the second-line diagnosis (i.e., Breast, Colorectum, Upper GI, Liver, Lung, Ovary, or Pancreas). Therefore, its input feature set covers the previous nine features and includes additional 31 protein markers and patient gender as listed in the later slides.

First Dataset

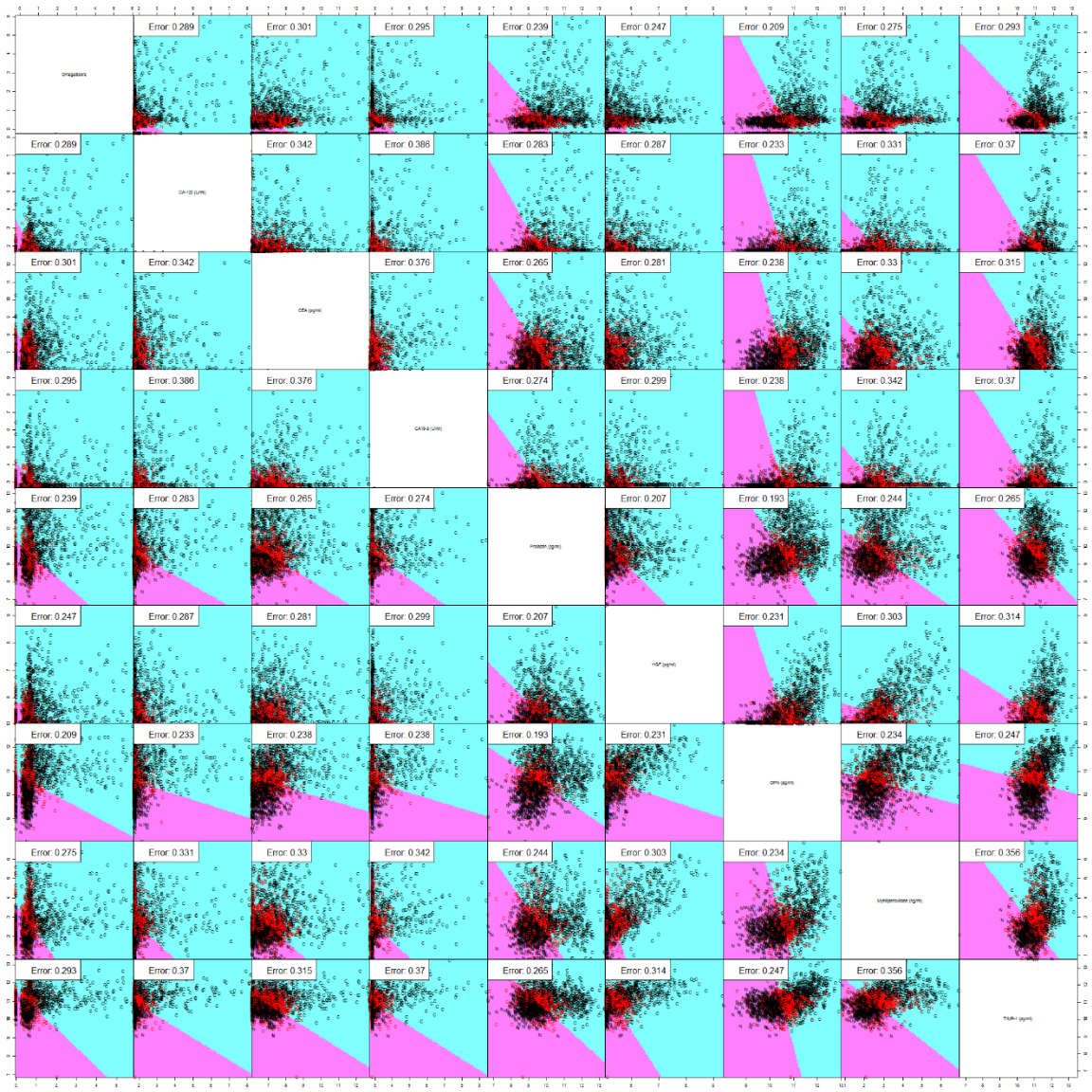
Feature Ranking for Binary Cancer Detection

InfoG	Input Features	Feature Description
0.6897	CA19-9 (U/ml)	Circulating Cancer Antigen 19-9 Concentration in U/ml
0.5119	CA-125 (U/ml)	Circulating Cancer Antigen 125 Concentration in U/ml
0.5001	HGF (pg/ml)	Circulating Hepatocryte Growth Factor Concentration in pg/ml
0.2779	OPN (pg/ml)	Circulating Osteopontin Concentration in pg/ml
0.2208	OmegaScore	Omega Score for Mutations in Circulating Cell-Free DNA
0.1826	Prolactin (pg/ml)	Circulating Prolactin Concentration in pg/ml
0.1518	CEA (pg/ml)	Circulating CarcinoEmbryonic Antigen Concentration in pg/ml
0.0989	Myeloperoxidase (ng/ml)	Circulating Myeloperoxidase Concentration in ng/ml
0.0916	TIMP-1 (pg/ml)	Circulating Tissue Inhibitor of MetalloProteinases 1 Concentration in pg/ml

Table S1: **Feature List for Cancer Detection ranked by Information Gain (InfoG), related to Figure 1**

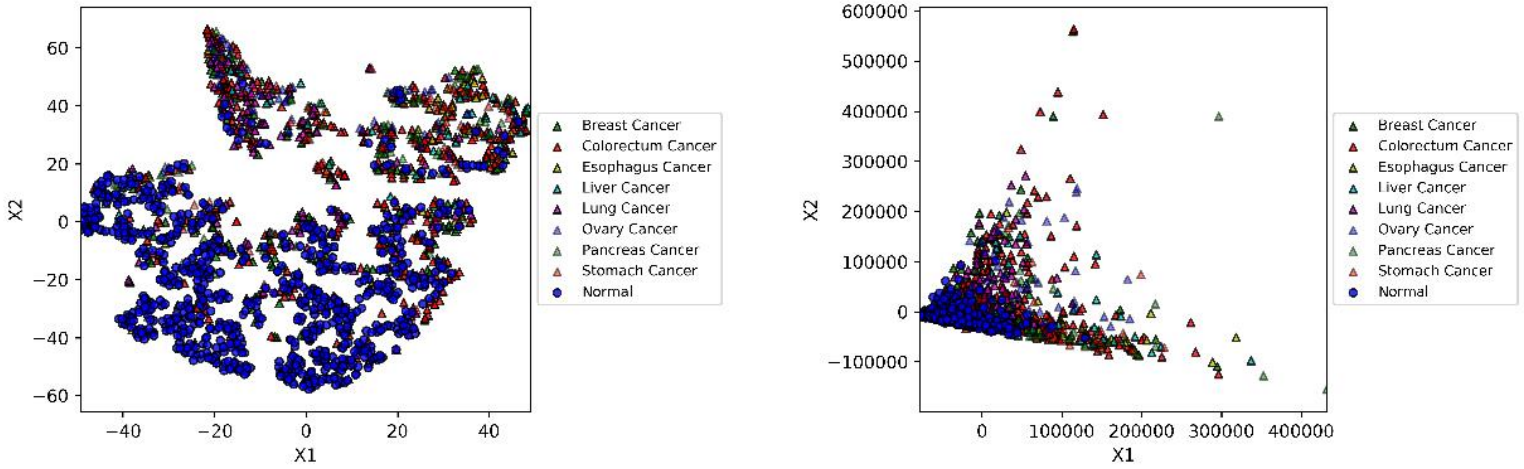
Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Linear Discriminant Analysis



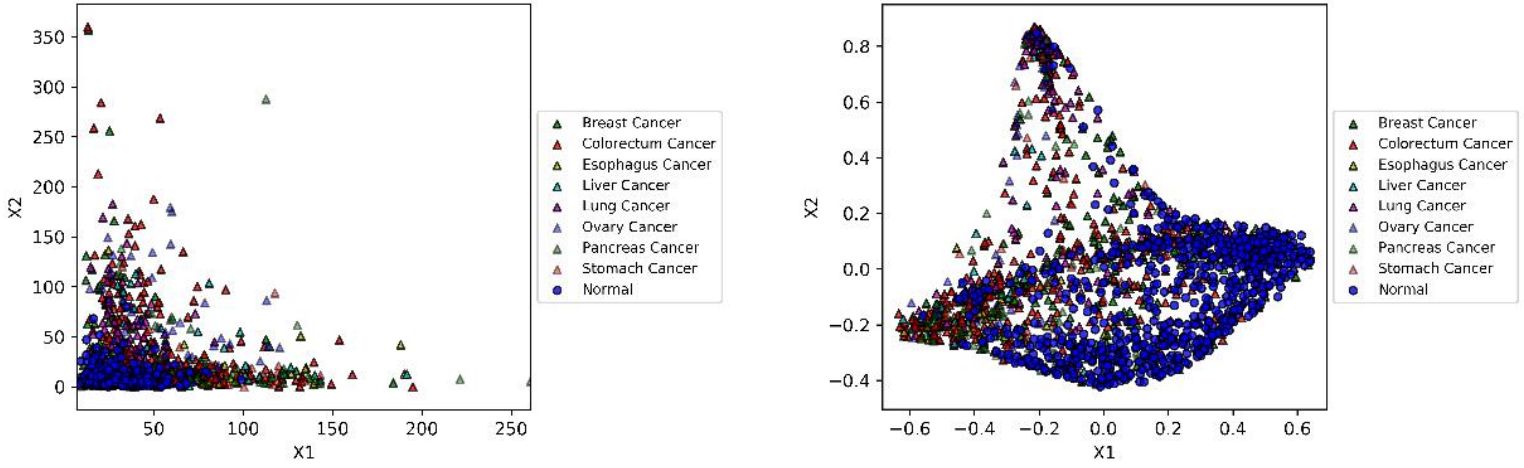
First Dataset

Dimensional Reduction



(a) t-distributed Stochastic Neighbor Embedding

(b) Principal Component Analysis



(c) Nonnegative Matrix Factorization

(d) Spectral Embedding

Proposed Approach – A1DE

WEKA
The University of Waikato

pentaho
open source business intelligence

WEKA Packages

IMPORTANT: make sure there are no old versions of Weka (<3.7.2) in your CLASSPATH before starting Weka

Installation of Packages

A GUI package manager is available from the "Tools" menu of the GUIChooser

```
java -jar weka.jar
```

For a command line package manager type:
java weka.core.WekaPackageManager -h

Running packaged algorithms from the command line

```
java weka.Run [algorithm name]
```

Substring matching is also supported. E.g. try:

```
java weka.Run Bayes
```

Available Packages (206)

AffectiveTweets	Text classification	Text Filters for Analyzing Sentiment and Emotions of Tweets
AnDE	Classification	Averaged N-Dependence Estimators (includes A1DE and A2DE)

We Have Tried All 206 Packages

It is the best for this task.

Proposed Approach – A1DE

Hence, given that each blood marker sample can be represented by a vector $x = \langle x_1, x_2, ..., x_n \rangle$ where x_i is a marker attribute value, an A1DE model can be trained and assigned cancer detection label y based on its posterior probability:

$$P(y|x) = \frac{P(y,x)}{P(x)} \propto P(y,x) \tag{1}$$

By aggregating all possible 1-dependence classifiers, $P(y,x)$ can be written as:

$$P(y,x) = \frac{\sum_{1 \le i \le n \wedge F(x_i) \ge m} P(y,x_i) P(x|y,x_i)}{|\{1 \le i \le n \wedge F(x_i) \ge m\}|} \tag{2}$$

Therefore, the label assignment (cancer detection label y) can be derived as follows:

$$\arg\max_{y} \sum_{1 \le i \le n \wedge F(x_i) \ge m} \hat{P}(y,x_i) \prod_{j=1}^{n} \hat{P}(x_j|y,x_i) \tag{3}$$

where $\hat{P}$ denotes the probability estimate. From the above, we can see that, if none of the parent attributes x_i have its $F(x_i)$ count greater than m, the A1DE is identical to a traditional NB classifier. On the other hand, the posterior of classes can be derived as follows:

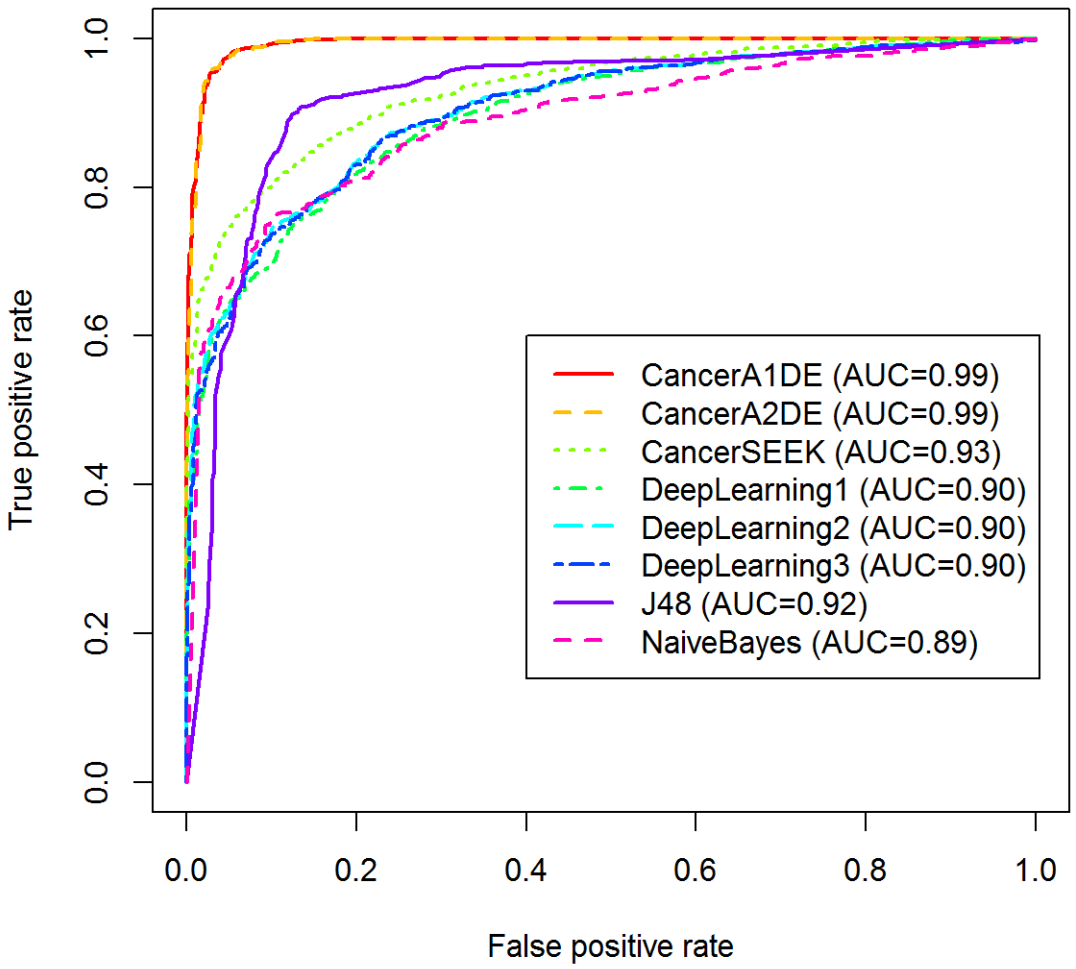
$$\hat{P}(y|x) = \frac{\sum_{1 \le n \wedge F(x_i) \ge m} \hat{P}(y,x_i) \prod_{j=1}^{n} \hat{P}(x_j|y,x_i)}{\sum_{y' \in Y} \sum_{1 \le n \wedge F(x_i) \ge m} \hat{P}(y',x_i) \prod_{j=1}^{n} \hat{P}(x_j|y',x_i)} \tag{4}$$

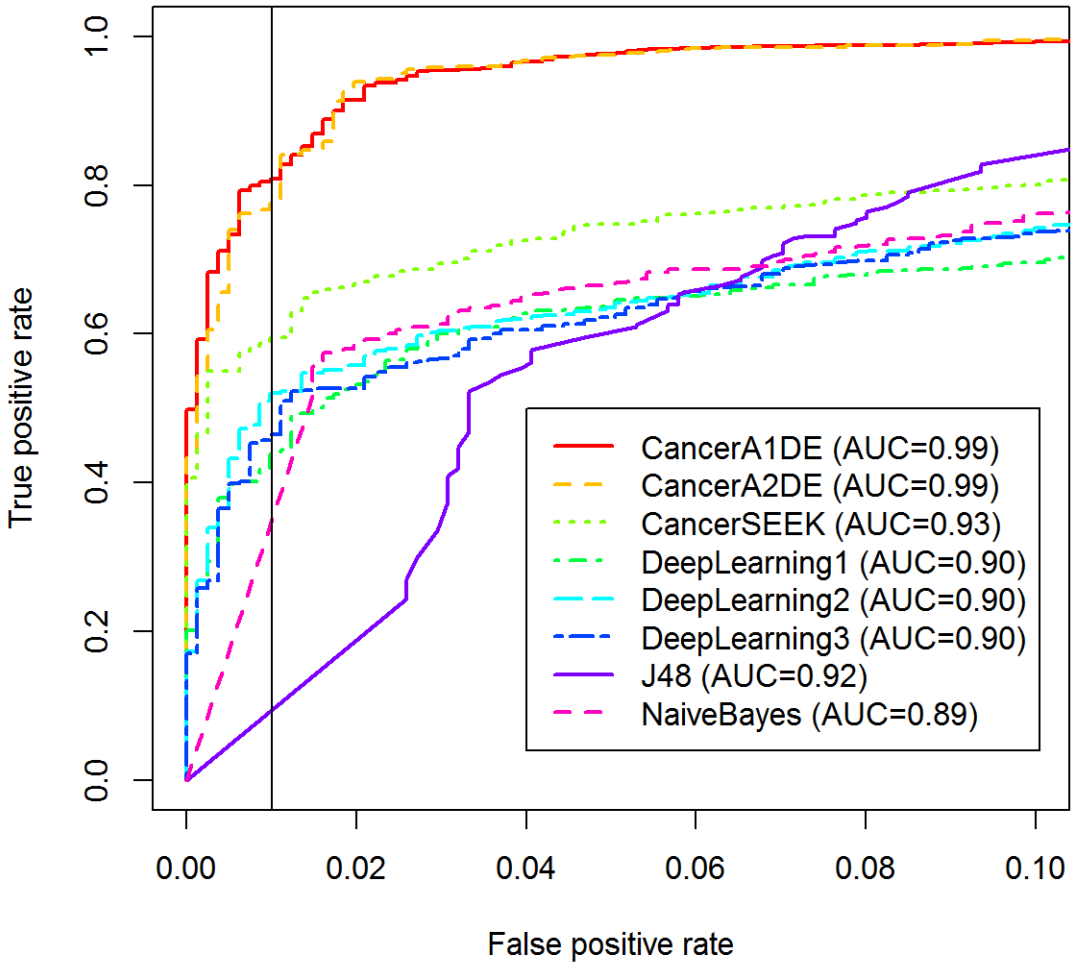
where the above formula is derived from the Bayes rule $P(y|x) = P(y,x)/P(x)$. The

*with Minimum Description Length (MDL) feature discretization

First Dataset

ROC Curves for Binary Cancer Detection





Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

First Dataset

Sensitivities (Recalls) for Binary Cancer Detection

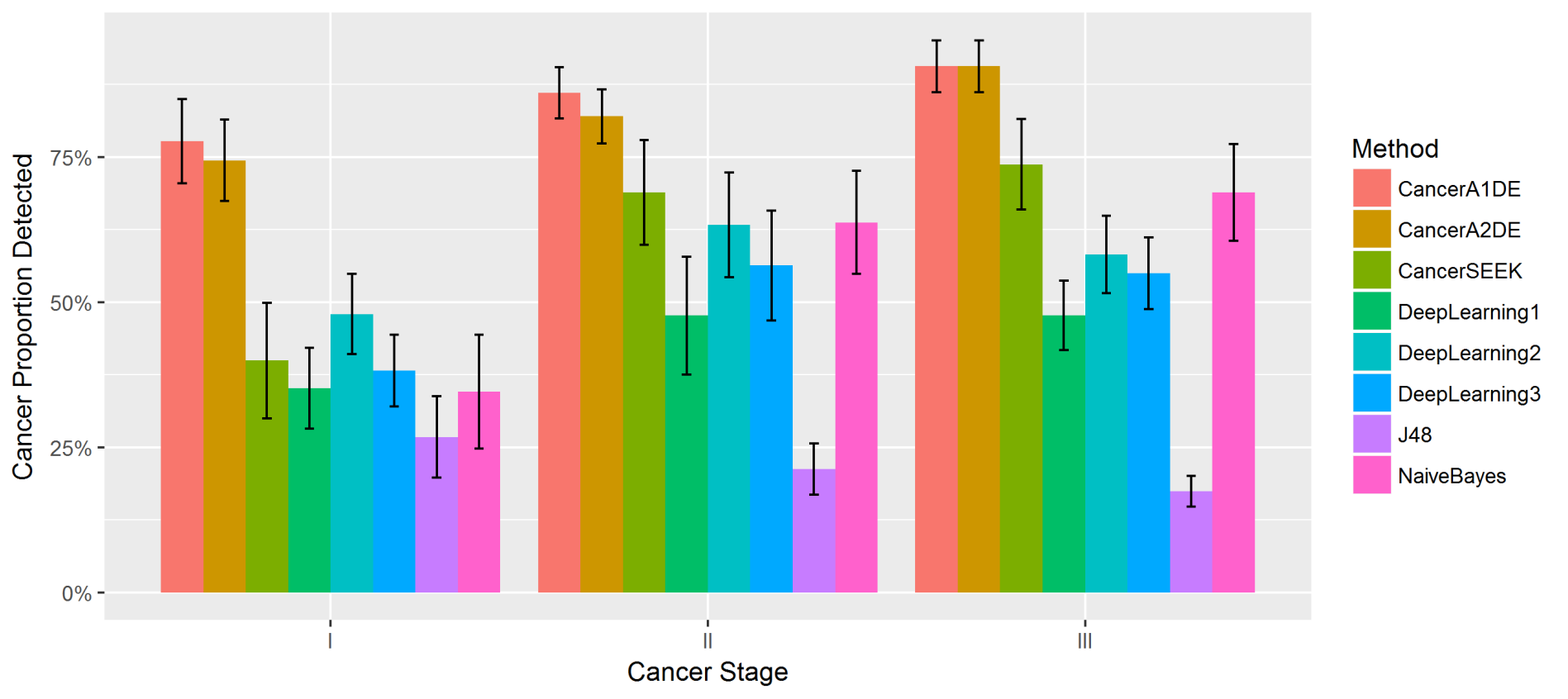


Figure 2. Proportion of Detected Cancers with Different Stages at the 99% Specificity Level. Each color represents a method, and the horizontal axis has been ordered by cancer stages. Each bar represents the median sensitivity of each method on each cancer stage with standard errors.

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

First Dataset

Sensitivities (Recalls) for Binary Cancer Detection

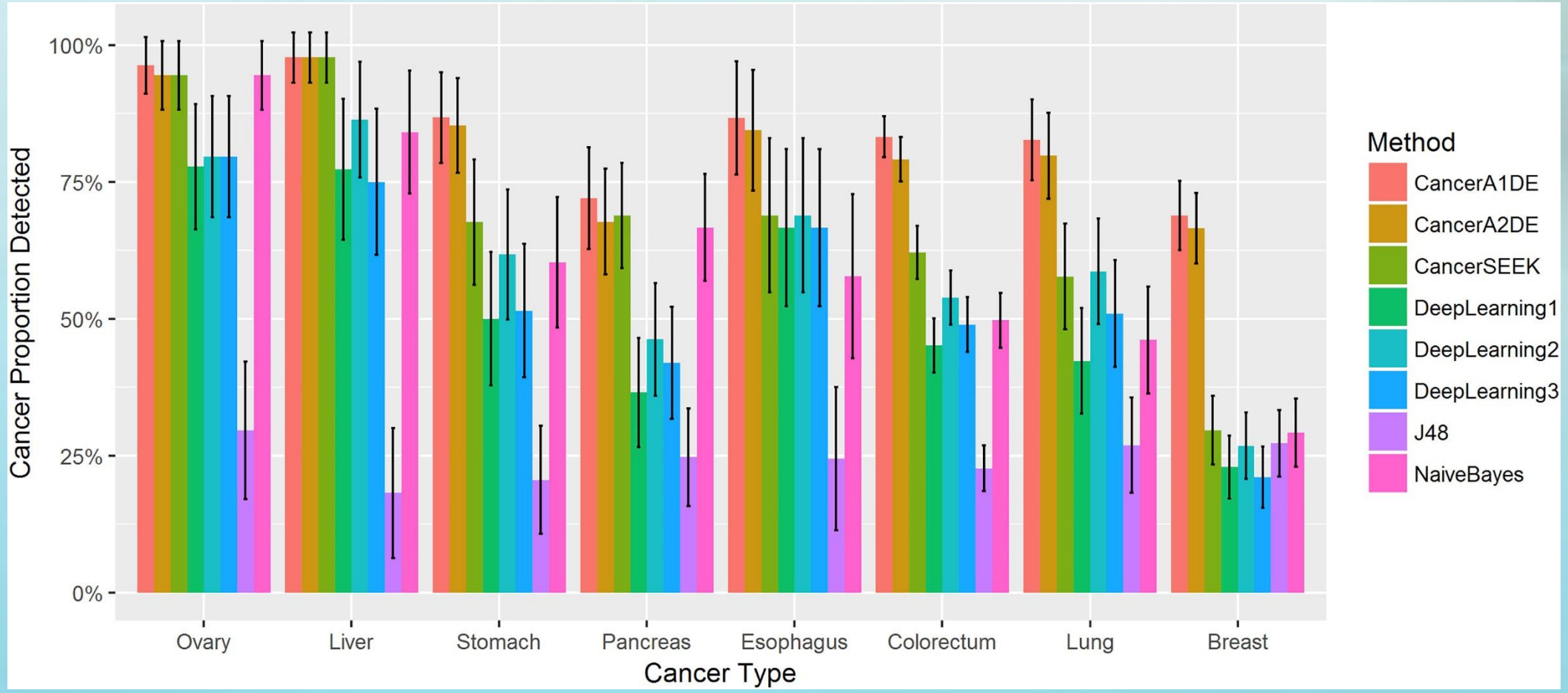


Figure 3. Detected Proportions of Different Cancer Types at the 99% Specificity Level
Different colors represent different methods. The horizontal axis is ordered by cancer types. Each bar represents the sensitivity of each method on each cancer type with 95% confidence intervals.

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Second Dataset

Feature Ranking for Multiple Cancer Classification

InfoG	Input Features	Feature Description
1.0389	TGFa (pg/ml)	Circulating Transforming Growth Factor Alpha Concentration in pg/ml
0.8301	HE4 (pg/ml)	Circulating Human Epididymis Protein 4 Concentration in pg/ml
0.6135	sFas (pg/ml)	Circulating soluble Fas Cell Surface Death Receptor Concentration in pg/ml
0.5372	Thrombospondin-2 (pg/ml)	Circulating Thrombospondin-2 Concentration in pg/ml
0.5073	AFP (pg/ml)	Circulating AlphaFetoprotein Precursor Concentration in pg/ml
0.3759	G-CSF (pg/ml)	Circulating Granulocyte-Colony Stimulating Factor Concentration in pg/ml
0.3633	IL-6 (pg/ml)	Circulating InterLeukin-6 Concentration in pg/ml
0.3597	CA-125 (U/ml)	Circulating Cancer Antigen 125 Concentration in U/ml
0.2568	Sex	Patient Gender Information (Male or Female)
0.2352	sHER2/sEGFR2/sErbB2 (pg/ml)	Circulating sHER2/sEGFR2/sErbB2 Concentration in pg/ml
0.2259	TIMP-2 (pg/ml)	Circulating Tissue Inhibitor of MetalloProteinases 2 Concentration in pg/ml
0.2231	CD44 (ng/ml)	Circulating CD44 Concentration in pg/ml
0.183	CA19-9 (U/ml)	Circulating Cancer Antigen 19-9 Concentration in U/ml
0.1805	IL-8 (pg/ml)	Circulating InterLeukin-8 Concentration in pg/ml
0.164	CA 15-3 (U/ml)	Circulating Cancer Antigen 15-3 Concentration in U/ml
0.1448	HGF (pg/ml)	Circulating Hepatocryte Growth Factor Concentration in pg/ml
0.1431	OPG (ng/ml)	Circulating Osteopontin Concentration in pg/ml
0.1414	GDF15 (ng/ml)	Circulating Growth Differentiation Factor 15 Concentration in ng/ml
0.1384	Leptin (pg/ml)	Circulating Leptin Concentration in pg/ml
0.1271	Myeloperoxidase (ng/ml)	Circulating Myeloperoxidase Concentration in ng/ml
0.125	Kallikrein-6 (pg/ml)	Circulating Kallikrein-6 Concentration in pg/ml
0.1173	TIMP-1 (pg/ml)	Circulating Tissue Inhibitor of MetalloProteinases 1 Concentration in pg/ml
0.1122	Midkine (pg/ml)	Circulating Midkine Concentration in pg/ml
0.1095	Prolactin (pg/ml)	Circulating Prolactin Concentration in pg/ml
0.1032	Mesothelin (ng/ml)	Circulating Mesothelin Concentration in ng/ml
0.103	Galectin-3 (ng/ml)	Circulating Galectin-3 Concentration in ng/ml
0.096	OPN (pg/ml)	Circulating Osteopontin Concentration in pg/ml
0.0956	NSE (ng/ml)	Circulating Neuron-Specific Enolase Concentration in ng/ml
0.0901	sEGFR (pg/ml)	Circulating soluble Epidermal Growth Factor Receptor Concentration in pg/ml
0.0901	CEA (pg/ml)	Circulating CarcinoEmbryonic Antigen Concentration in pg/ml
0.085	AXL (pg/ml)	Circulating AXL Receptor Tyrosine Kinase Concentration in pg/ml
0.0771	sPECAM-1 (pg/ml)	Circulating soluble Platelet and Endothelial Cell Adhesion Molecule 1 Concentration in pg/ml
0.0637	SHBG (nM)	Circulating Sex Hormone-Binding Globulin Concentration in nM
0.0635	OmegaScore	Omega Score for Mutations in Circulating Cell-Free DNA
0	Angiopoietin-2 (pg/ml)	Circulating Angiopoietin-2 Concentration in pg/ml
0	DKK1 (ng/ml)	Circulating Dickkopf WNT Signaling Pathway Inhibitor 1 Concentration in ng/ml
0	CYFRA 21-1 (pg/ml)	Circulating Cytokeratin-19 Fragment Concentration in pg/ml
0	PAR (pg/ml)	Circulating Protease-Activated Receptor Concentration in pg/ml
0	Endoglin (pg/ml)	Circulating Endoglin Concentration in pg/ml
0	FGF2 (pg/ml)	Circulating Fibroblast Growth Factor 2 Concentration in pg/ml
0	Follistatin (pg/ml)	Circulating Follistatin Concentration in pg/ml

Table 1: **Feature List for Cancer Type Localization ranked by Information Gain (InfoG)**

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Second Dataset

Sensitivities (Recalls) for Multiple Cancer Classification

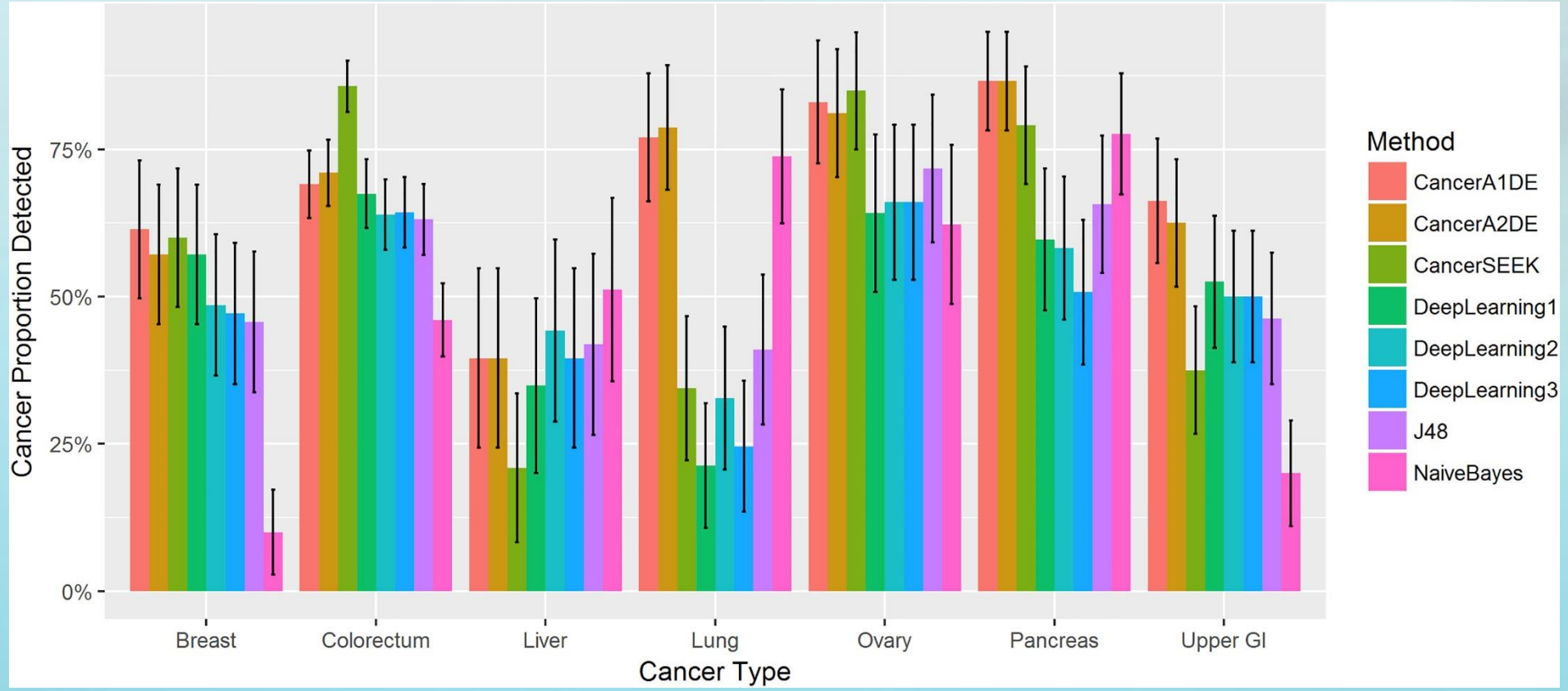


Figure 4. Localized Proportions of Different Cancer Types using the Top One Prediction Approach

Different colors represent different methods. The horizontal axis is ordered by cancer types. Each bar represents the sensitivity of each method on each cancer type with 95% confidence intervals.

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Second Dataset

Feature Importance Bi-Clustering

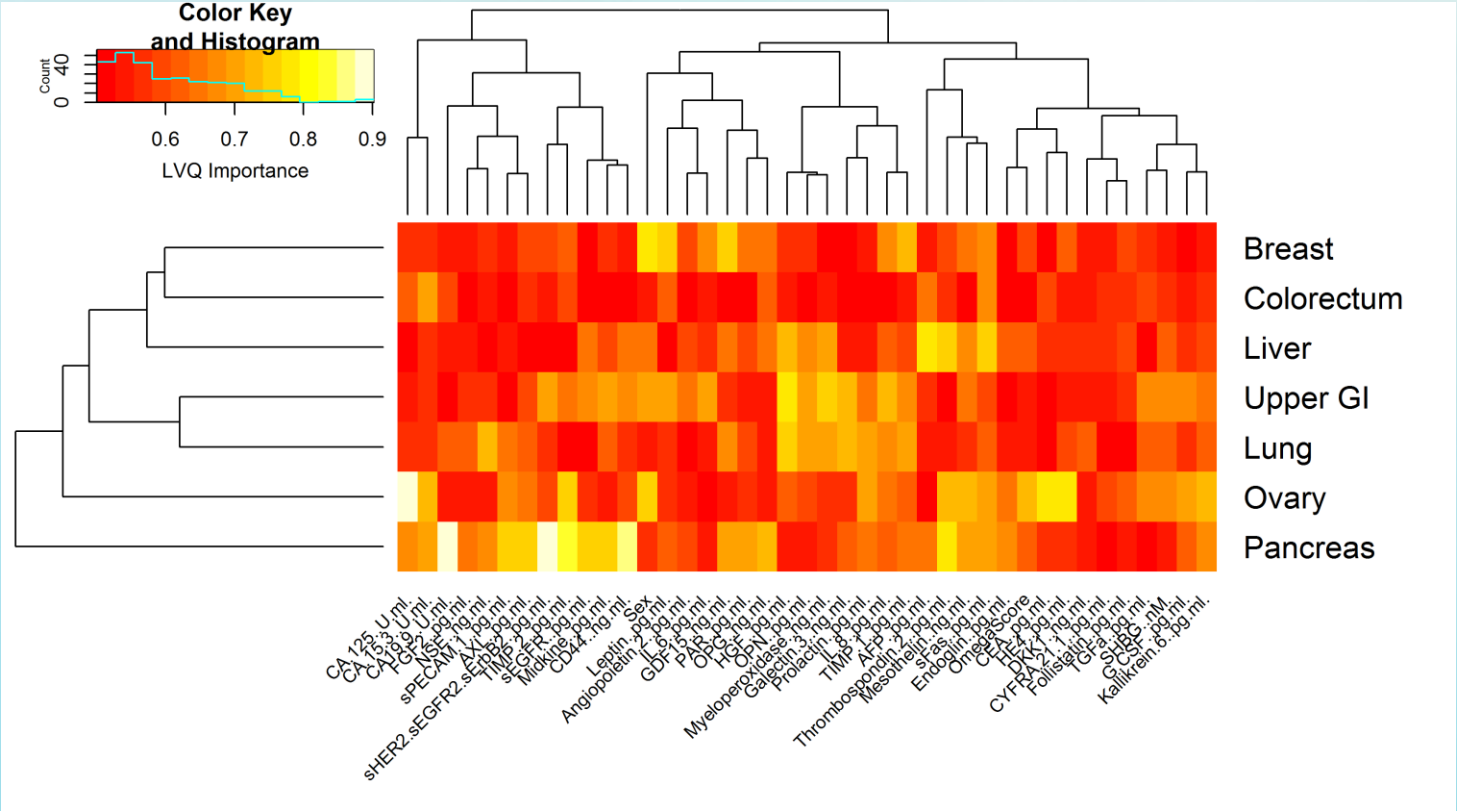


Figure 5. Feature Importance Heatmap for Cancer Type Localization under One-Class-versus-Others Setting

The feature rankings are measured based on the Learning Vector Quantization (LVQ) building under Python caret package (Bischl et al., 2016). Ten-fold cross-validations are run to compute the feature importance values. After that, the function “heatmap.2” in R language is adopted with the default setting to cluster and visualize the feature importance values. Further details can be found in Figure S14.

Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

http://cancer.cs.cityu.edu.hk/

Cancer Detection from Blood | CANCER DETECTION | CANCER TYPE LOCALIZATION | DOWNLOADS | FAQ | ABOUT | TEAM | CONTACT

CANCER DETECTION

Do You Want to Know Your Probabilty to Get a Cancer? Just Have a Try!

Step 1. Blood Test Marker Values	
OmegaScore	2.96
CA-125 (U/ml)	5.09
CEA (pg/ml)	540.1
CA19-9 (U/ml)	16.45
Prolactin (pg/ml)	11606.6
HGF (pg/ml)	377.26
OPN (pg/ml)	56516.58
Myeloperoxidase (ng/ml)	14.22
TIMP-1 (pg/ml)	56428.71
Step 2. Model Selection	
Model	CancerA1DE

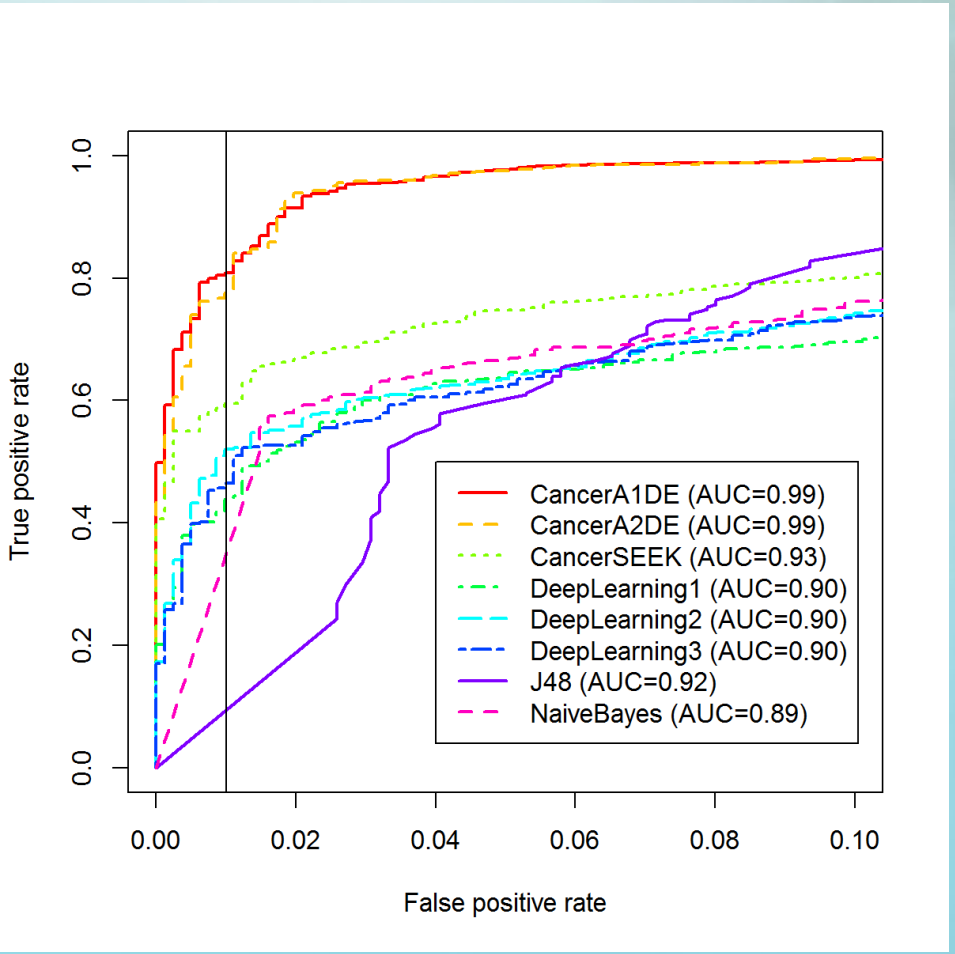
AUTOFILL RESET SUBMIT

The probability to get a cancer is:

Conclusion

Summary

We explore different supervised learning approaches for multiple cancer type detection and observe significant improvements; for instance, one of our approaches **(i.e., CancerA1DE) can double the existing sensitivity from 38% to 77% for the earliest cancer detection (i.e., Stage I) at the 99% specificity level.** For Stage II, it can even reach up to about 90% across multiple cancer types. In addition, CancerA1DE can also double the existing sensitivity from 30% to 70% for detecting breast cancers at the 99% specificity level. Data and model analysis are conducted to reveal the underlying reasons. A website is built at http://cancer.cs.cityu.edu.hk/.



Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Summary - Case Studies

- Case for synthesizing Heterodimeric DNA Motifs [1]
- Case for predicting CRISPR-Cas9 Off-targets [2,3]
- Case for screening Early Cancers from Blood [4]

References

[1] Wong, K. C., Lin, J., Li, X., Lin, Q., Liang, C., & Song, Y. Q. (2018). Heterodimeric DNA motif synthesis and validations. Nucleic Acids Research, 47(4), 1628-1636. (2019 Impact Factor: 11.501)

[2] Lin, J., & Wong, K. C. (2018). Off-target predictions in CRISPR-Cas9 gene editing using deep learning. Bioinformatics, 34(17), i656-i663.

[3] Lin, J., Zhang, Z., Zhang, S., Chen, J., & Wong, K. C. (2020). CRISPR-Net: A Recurrent Convolutional Network Quantifies CRISPR Off-Target Activities with Mismatches and Indels. Advanced Science, 1903562. (2019 Impact Factor =15.4)

[4] Wong, K. C. et al. (2019). Early Cancer Detection from Multianalyte Blood Test Results. iScience.

Data Application Case Studies for COVID-19

(30 minutes)

Outline

- Data Application Cases at Molecular Biology Level
  - 1D Sequence Data from COVID-19
  - 3D Molecular Structure Data from COVID-19
  - 3D Molecular Vaccine Design for COVID-19 Treatment
- Data Application Cases at Public Health / Clinical Level
  - Infection Case Data on COVID-19
  - Integrated Data Analysis on COVID-19
- Data Application Cases at Text Level
  - Academic Paper Analysis on COVID-19
  - Social Media Text Analysis on COVID-19

Molecular Biology Level

COVID-19 at Molecular Biology Level

https://www.nature.com/collections/hajgidghjb

nature research

Search Login

nature > collection

COLLECTION | 03 FEBRUARY 2020

Coronavirus

To support urgent research to combat the ongoing outbreak of COVID-19, caused by the novel coronavirus SARS-CoV-2, the editorial teams at Nature Research have curated a collection of relevant articles. Our collection includes research into the basic biology of... show more

Latest SARS-CoV-2/COVID-19 research

ARTICLE
OPEN ACCESS
27 OCT 2020
Nature Communications

Disease burden and clinical severity of the first pandemic wave of COVID-19 in Wuhan, China

Here the authors analyze disease burden and clinical severity of COVID-19 during the first wave in Wuhan, China in comparison to past influenza virus pandemics and COVID-19 in the US and Canada. These estimates of symptomatic cases, medical consultations,... show more

Juan Yang, Xinhua Chen ··· Prof Hongjie Yu

ARTICLE
21 OCT 2020
Nature

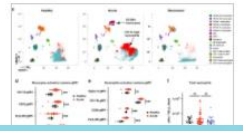
A pooled testing strategy for identifying SARS-CoV-2 at low prevalence

Leon Mutesa, Pacifique Ndishimye ··· Wilfred Ndifon

ARTICLE
OPEN ACCESS
16 OCT 2020
Nature Communications

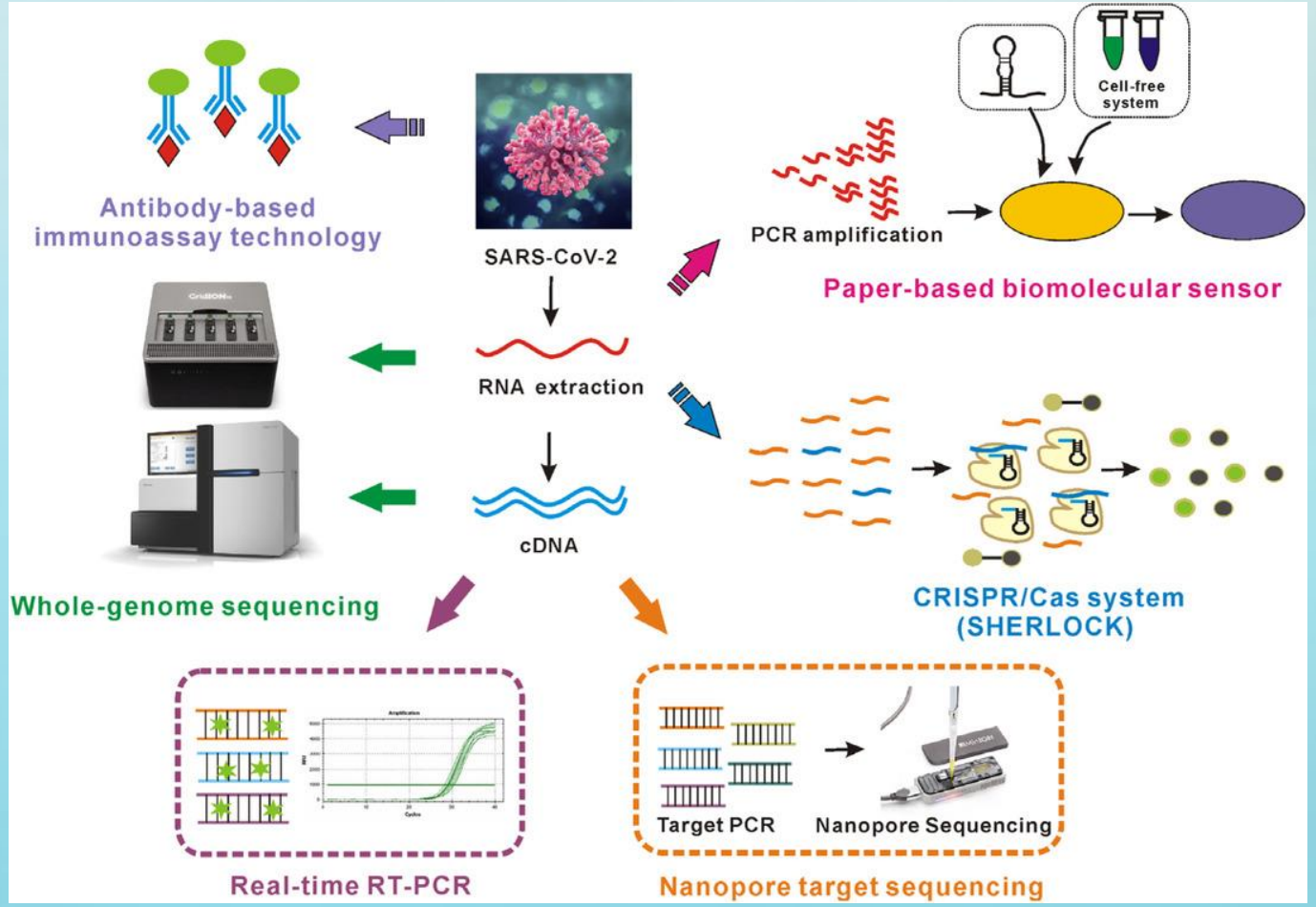
Whole blood immunophenotyping uncovers immature neutrophil-to-VD2 T-cell ratio as an early marker for severe COVID-19

COVID-19 severity is associated with cytokine levels and lymphopenia, but the role of immune cell subsets is not well understood. Here the authors



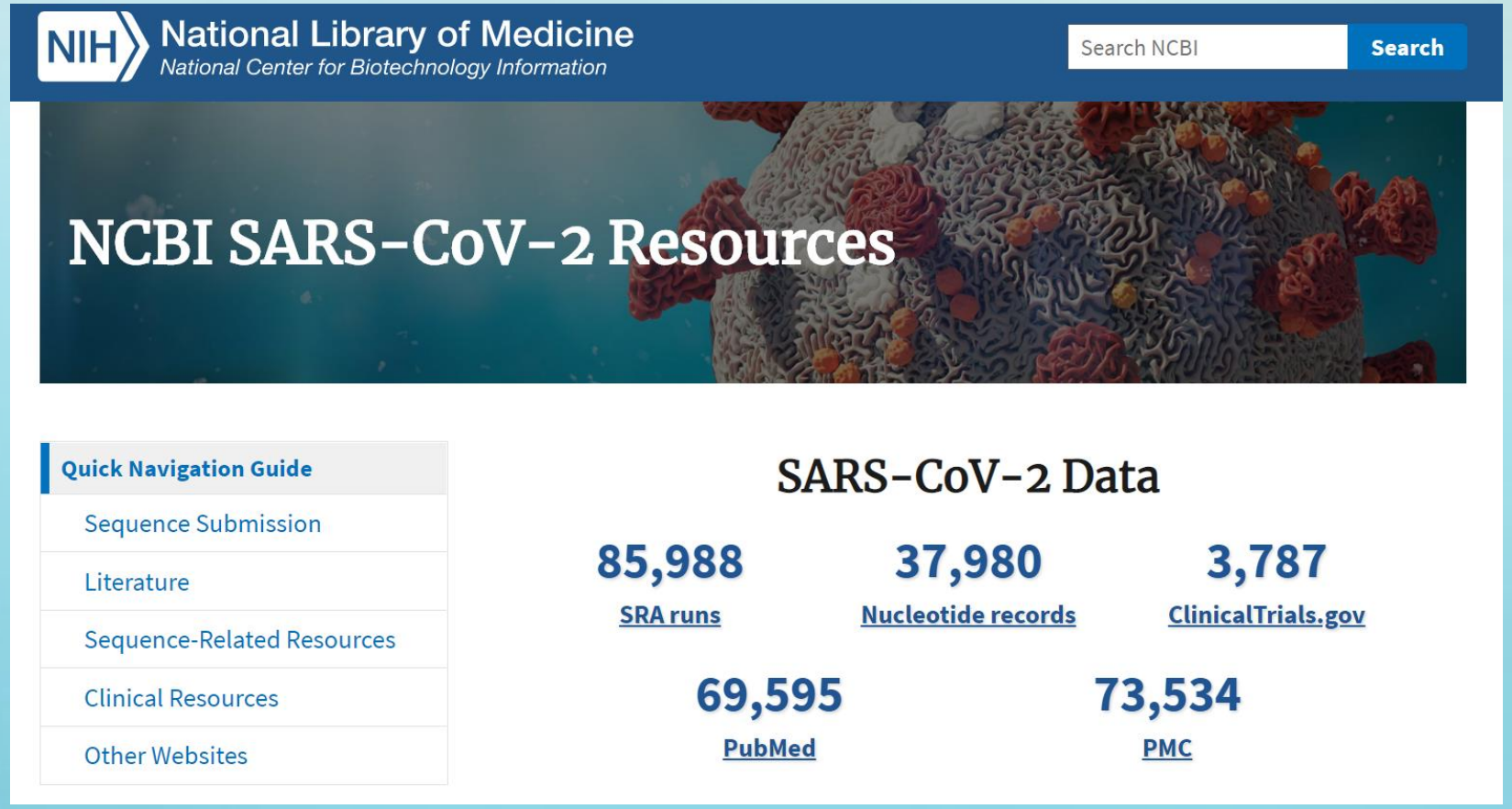
Liu, R., Fu, A., Deng, Z., Li, Y., & Liu, T. (2020). Promising methods for detection of novel coronavirus SARS-CoV-2. *View*, *1*(1), e4.

COVID-19 at Molecular Biology Level

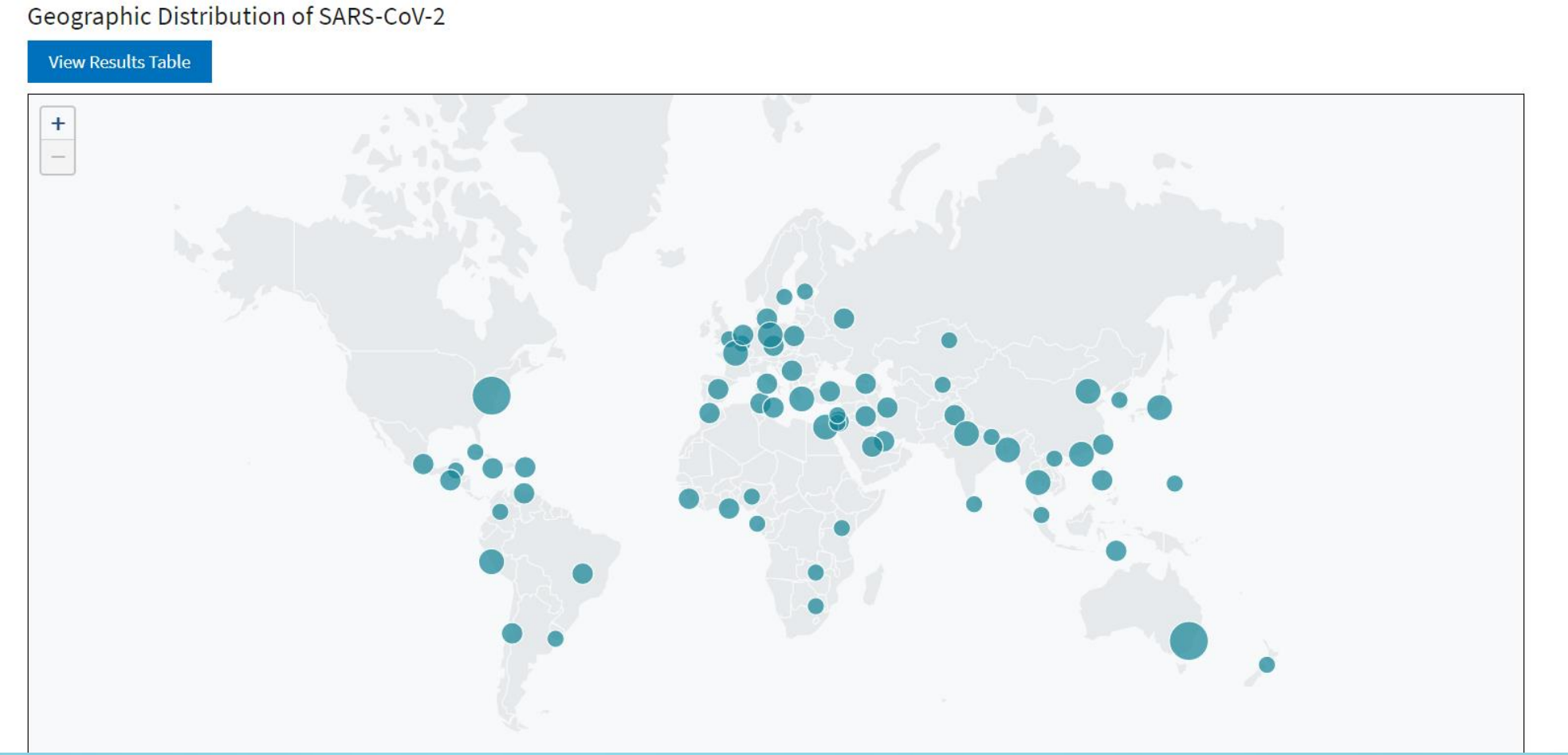


COVID-19 at Molecular Biology Level

- https://www.ncbi.nlm.nih.gov/sars-cov-2/

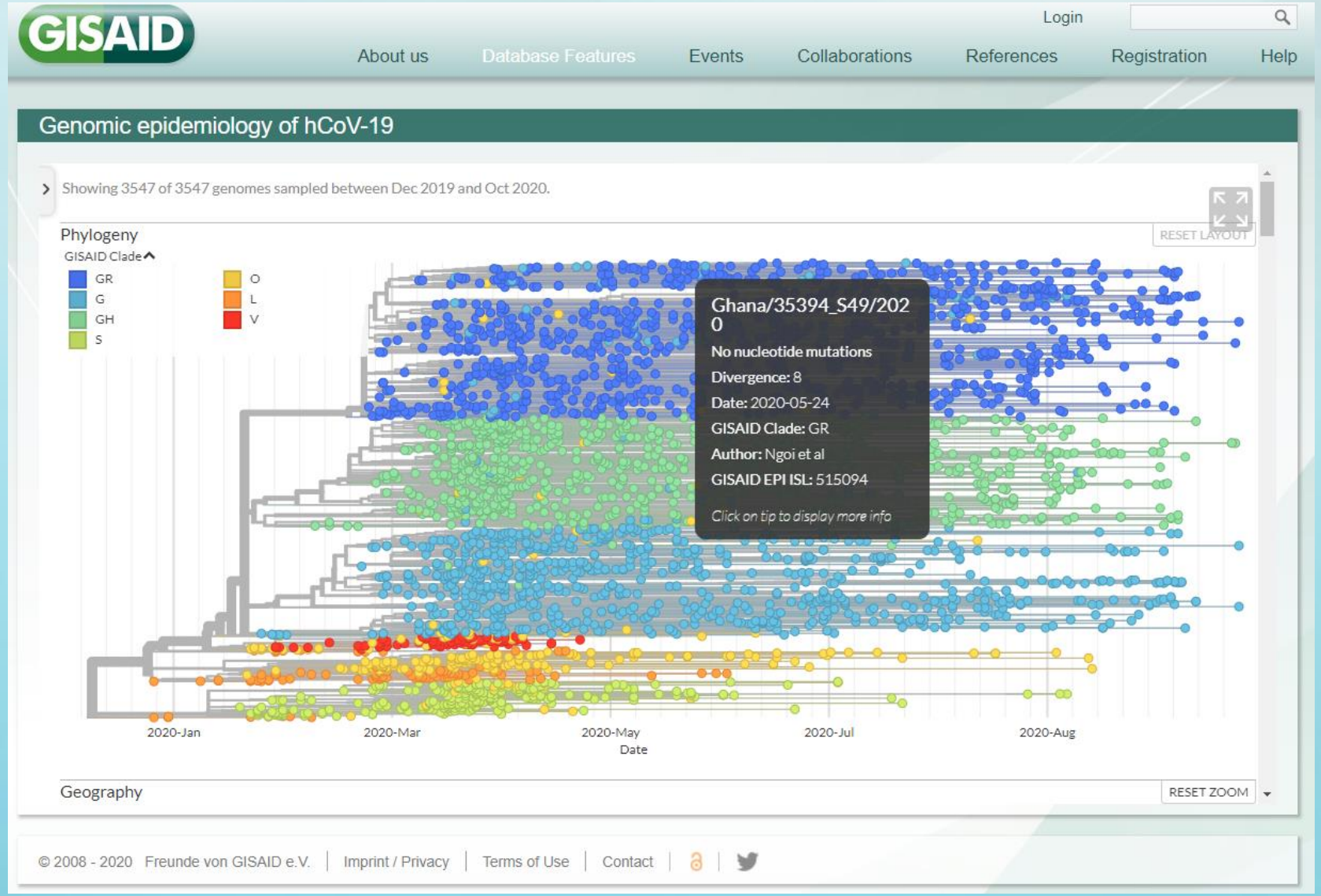


COVID-19 at Molecular Biology Level



https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2

COVID-19 at Molecular Biology Level

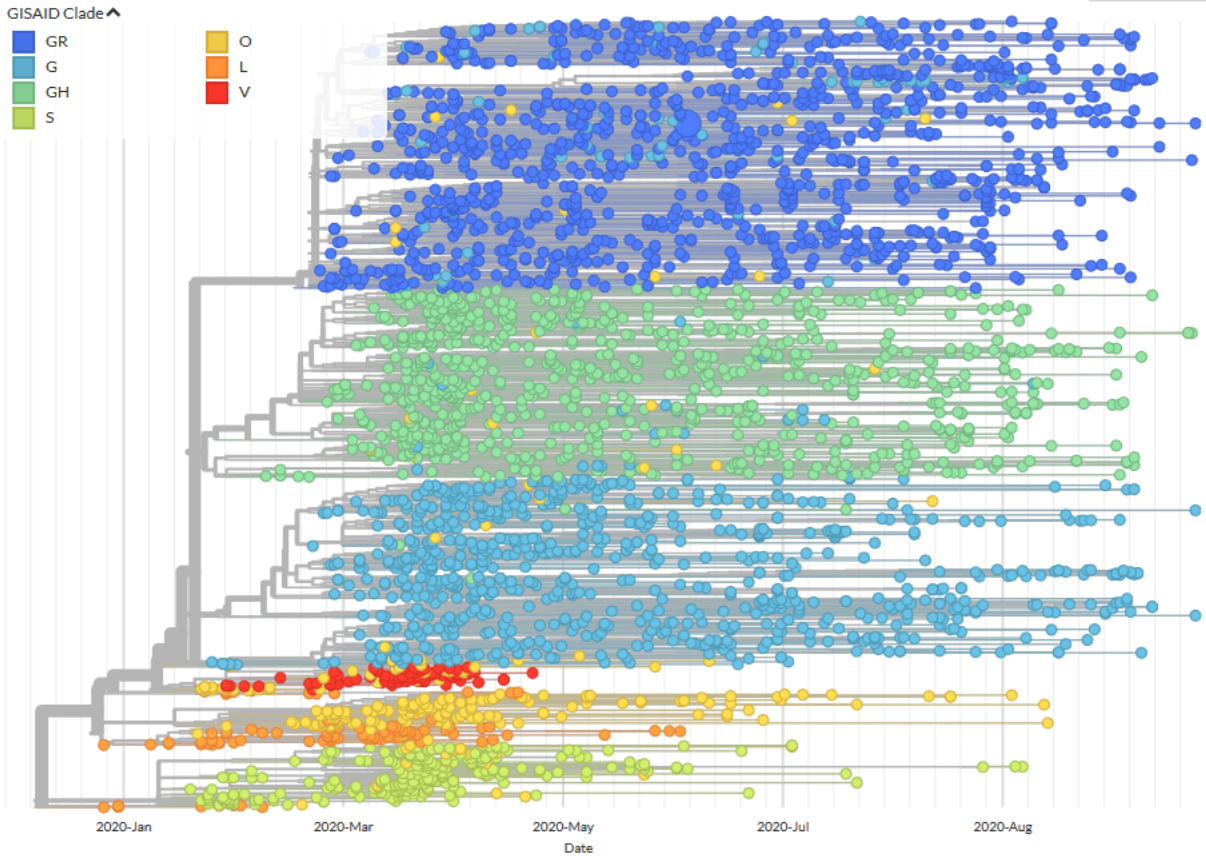


https://www.gisaid.org/epiflu-applications/phylodynamics/

Showing 3547 of 3547 genomes sampled between Dec 2019 and Oct 2020.

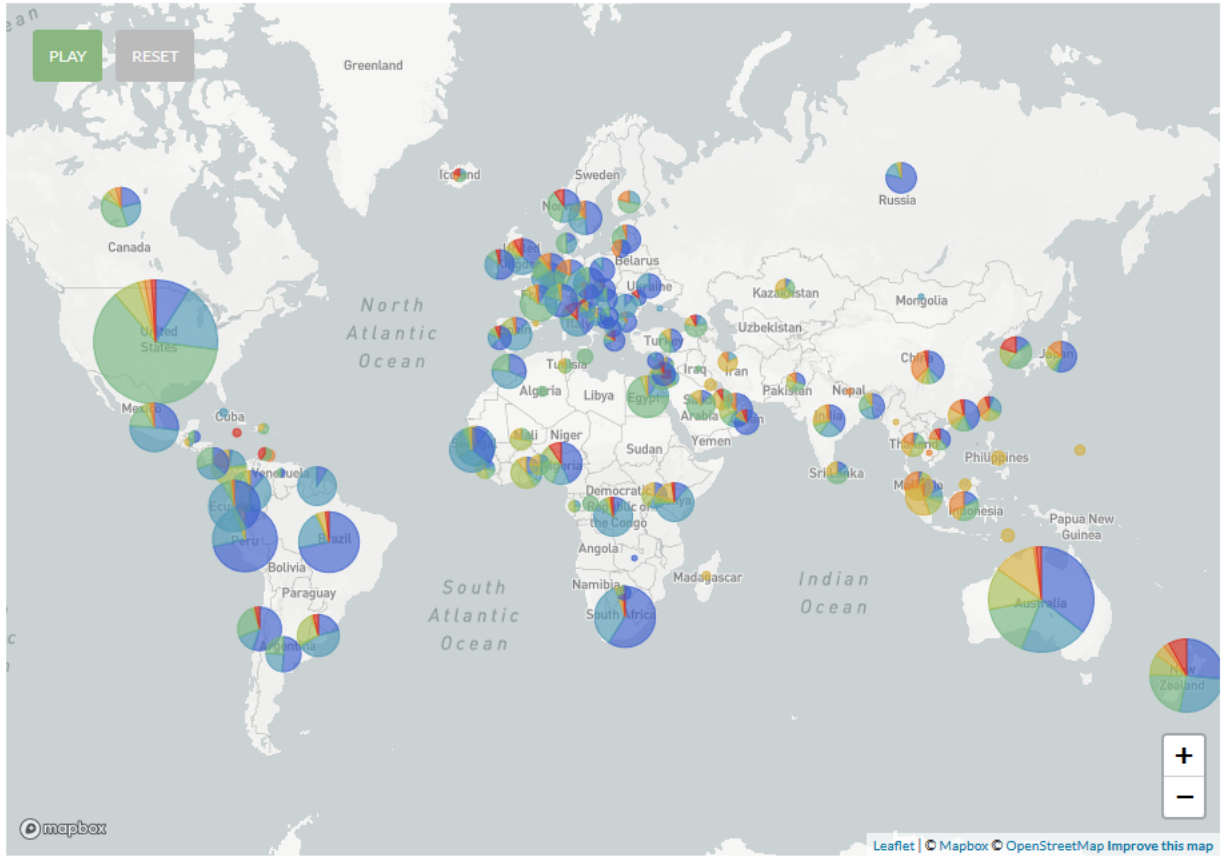
Phylogeny

RESET LAYOUT



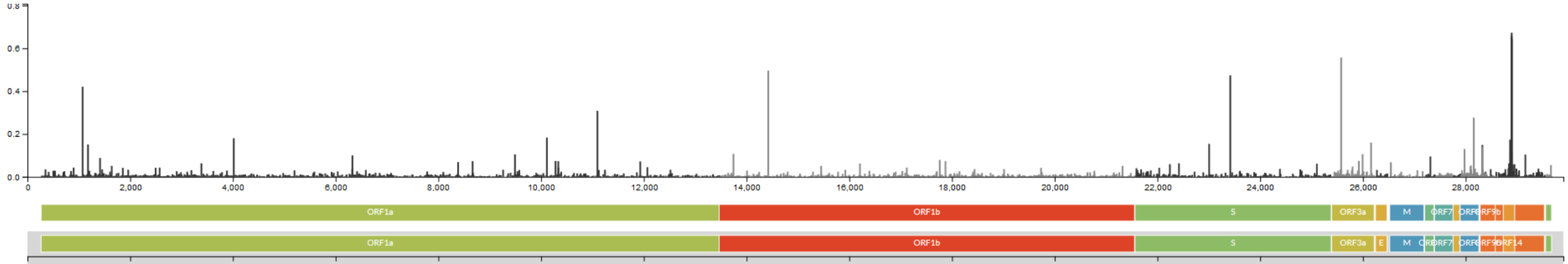
Geography

RESET ZOOM



Diversity

ENTROPY EVENTS AA NT



COVID-19 at Molecular Biology Level

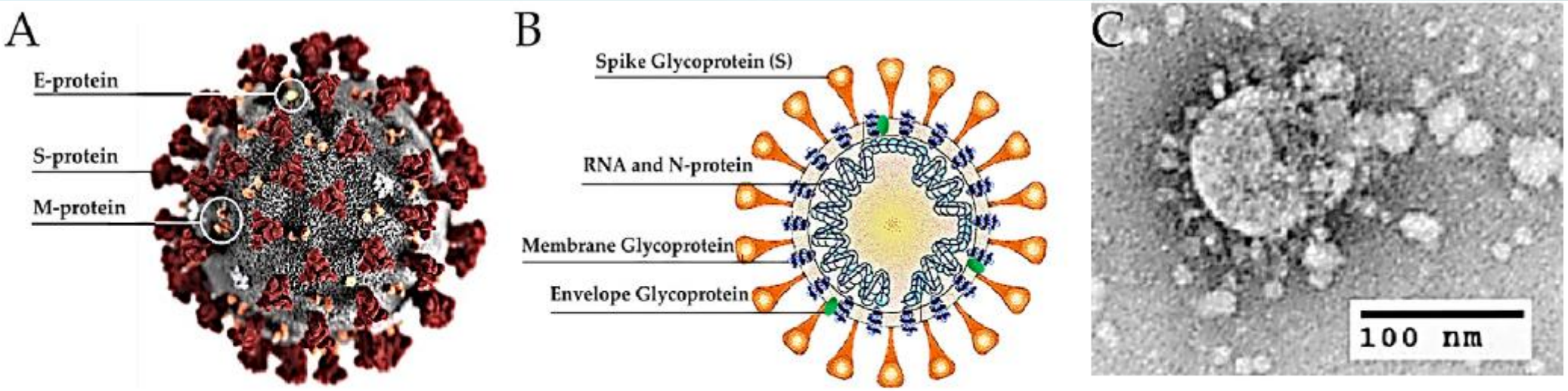


Fig. 1. Structure of SARS-CoV-2. (A) Illustration of the SARS-CoV-2 virion created at the Centers for Disease Control and Prevention (CDC). The spikes on the outer edge of the virus particles look like a crown, giving the disease its characteristic name. (B) Schematic representation of the structure of SARS-CoV-2. It has four structural proteins, S (spike), E (envelope), M (membrane), and N (nucleocapsid) proteins; the N protein holds the RNA genome, and the S, E, and M proteins together create the viral envelope. (C) An electron microscopic image of a thin section of SARS-CoV-2 within the cytoplasm of an infected cell, showing the spherical particles and cross-sections through the viral nucleocapsid (Sohrabi et al., 2020).

Abd El-Aziz, T. M., & Stockand, J. D. (2020). Recent progress and challenges in drug development against COVID-19 coronavirus (SARS-CoV-2)-an update on the status. Infection, Genetics and Evolution, 104327.

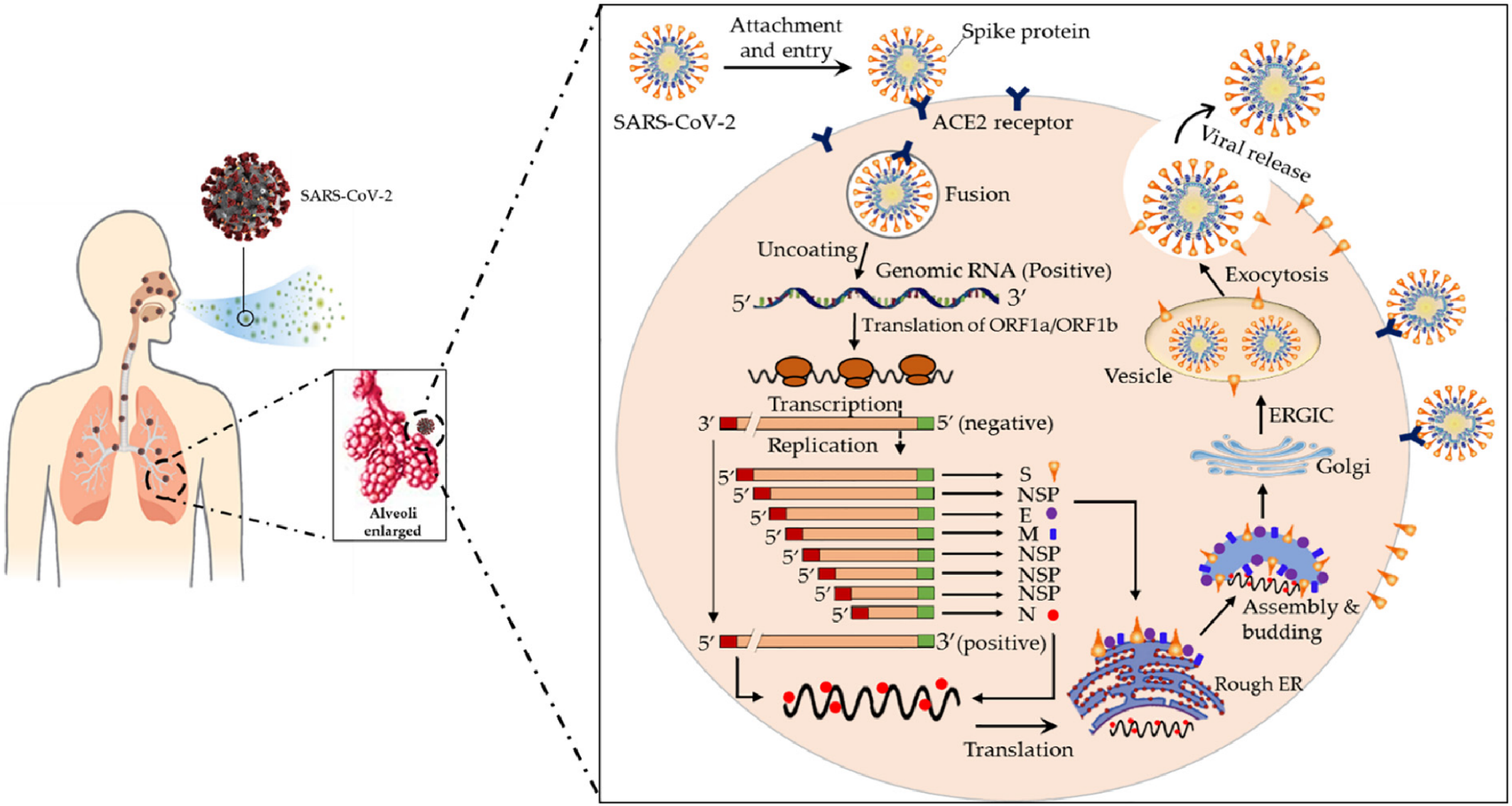
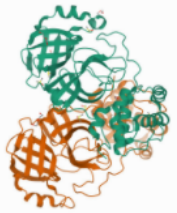


Fig. 2. The life cycle of SARS-CoV-2 in human lung cells. Coronavirus is most often transmitted by droplets while sneezing and coughing and its journey begins in the first days after infiltration from the upper respiratory tract. The spike proteins of SARS-CoV-2 binds to ACE2 receptors. The virion then releases RNA genome into the cell and translation of structural and non-structural proteins follows. ORF1a and ORF1ab are translated to produce pp1a and pp1ab polyproteins, which are cleaved by the proteases that are encoded by ORF1a to yield non-structural proteins. This is followed by assembly and budding into the lumen of the ERGIC. Virions are then released from the infected cell through exocytosis (Adnan Shereen et al., 2020). NSP, non-structural proteins; ACE2, Angiotensin-Converting Enzyme 2; Rough ER, Rough Endoplasmic Reticulum; ERGIC, Endoplasmic Reticulum Golgi Intermediate Compartment.

COVID-19 at Molecular Biology Level

https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52&feature=true

COVID-19 at Molecular Biology Level



3D View

7KFI

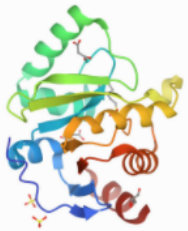
Download File | View File

SARS-CoV-2 Main protease immature form - apo structure

Noske, G.D., Nakamura, A.M., Gawriljuk, V.O., Lima, G.A.M., Zeri, A.C.M., Nascimento, A.F.Z., Oliva, G., Godoy, A.S.

To be published

Released	2020-10-28
Method	X-RAY DIFFRACTION 1.6 Å
Organisms	Severe acute respiratory syndrome coronavirus 2
Macromolecule	Main protease (protein)
Unique Ligands	DMS, PEG



3D View

7KG3

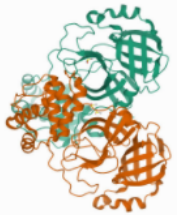
Download File | View File

Crystal structure of CoV-2 Nsp3 Macrodomain

Arvai, A., Brosey, C.A., Link, T., Jones, D.E., Ahmed, Z., Tainer, J.A.

To be published

Released	2020-10-28
Method	X-RAY DIFFRACTION 1.45 Å
Organisms	Severe acute respiratory syndrome coronavirus 2
Macromolecule	Non-structural protein 3 (protein)
Unique Ligands	GOL, MES, MLI, SO4



3D View

7KHP

Download File | View File

Acyl-enzyme intermediate structure of SARS-CoV-2 Mpro in complex with its C-terminal autoprocessing sequence.

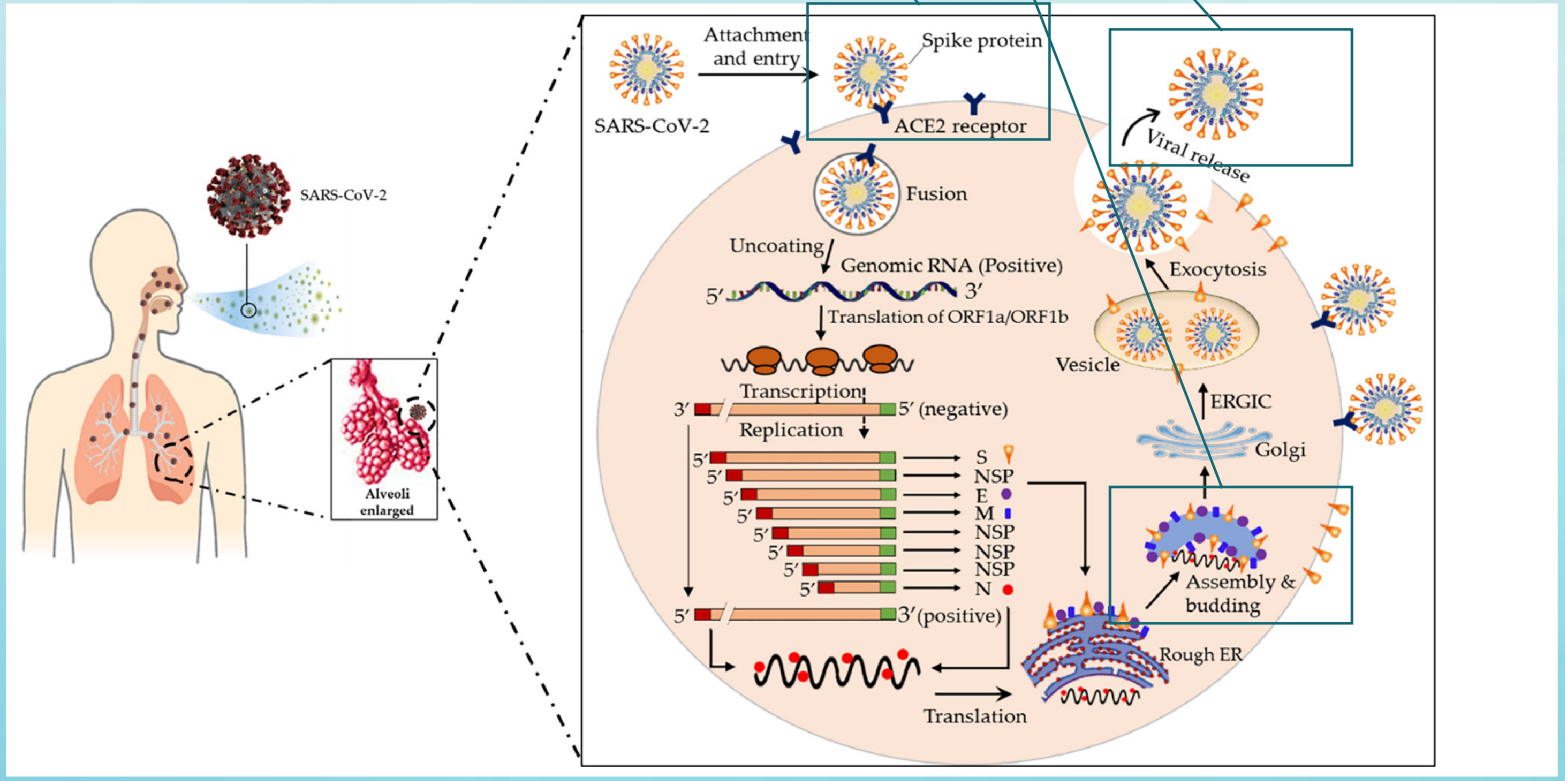
Lee, J., Worrall, L.J., Paetzel, M., Strynadka, N.C.J.

To be published

Released	2020-10-28
Method	X-RAY DIFFRACTION 1.95 Å
Organisms	Severe acute respiratory syndrome coronavirus 2
Macromolecule	3C-like proteinase (protein)

COVID-19 at Molecular Biology Level

Potential Vaccine Targets



COVID-19 at Molecular Biology Level

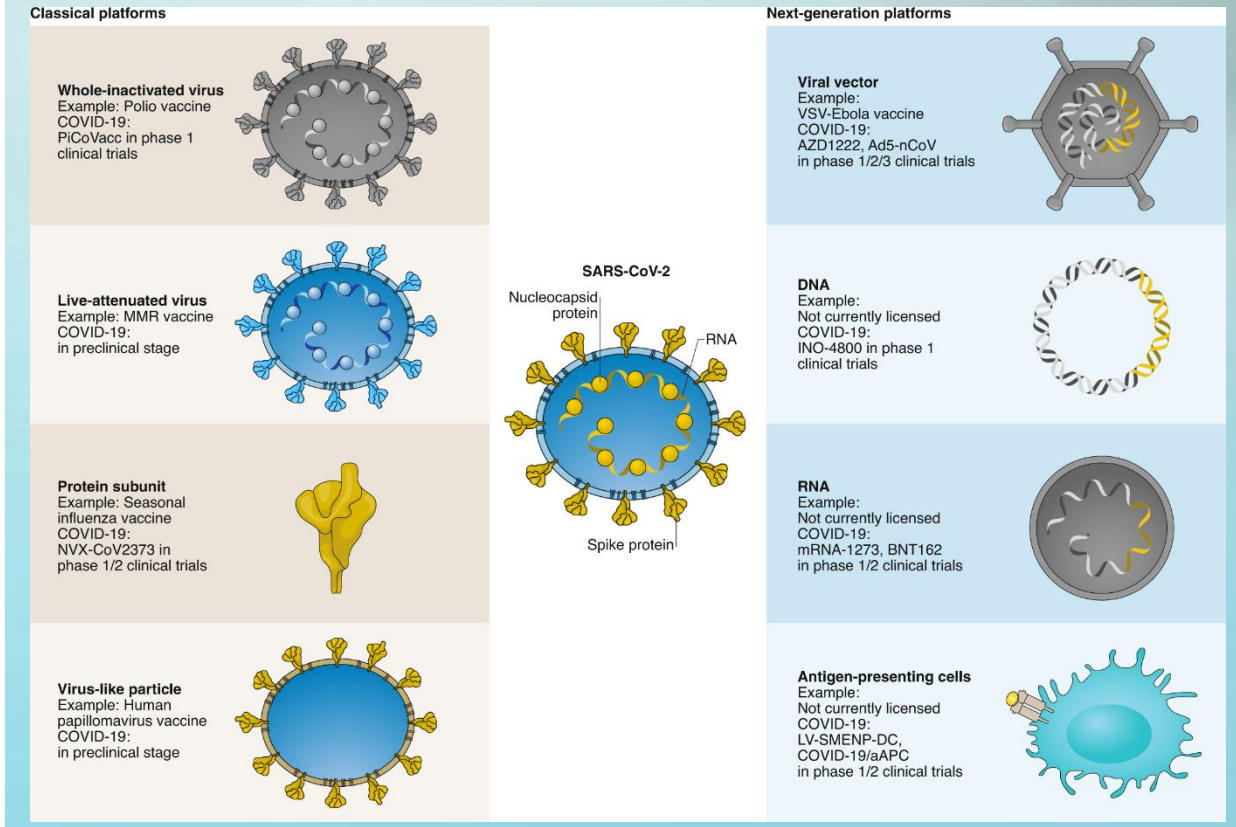
Types of coronavirus vaccine approaches

Scientists are casting a wide net to see what works best against the novel coronavirus.

Types of vaccines	DNA and RNA	Live attenuated	Inactivated	Subunit	Viral vector
How it works	This vaccine uses DNA or RNA molecules to teach the immune system to target key viral proteins.	This is a weakened version of the actual virus.	An inactivated vaccine uses the whole virus after it has been killed with heat or chemicals.	This vaccine uses a piece of a virus' surface to focus your immune system on a single target.	This approach takes a harmless virus and uses it to deliver viral genes to build immunity.
Advantages	Easy and quick to design.	Stimulates a robust immune response without causing serious disease.	Safe because the virus is already dead and is easy to make.	Focuses the immune response on the most important part of the virus for protection and cannot cause infection.	Live viruses tend to elicit stronger immune responses than dead viruses or subunit vaccines.
Disadvantages	Never been done before. There are no licensed DNA or RNA vaccines currently in use.	May not be safe for those with compromised immune systems.	Not as effective as a live virus. Some previous inactivated vaccines have made the disease worse; safety for the novel coronavirus needs to be shown in clinical trials.	May not stimulate a strong response, other chemicals may need to be added to boost long-term immunity.	Important to pick a viral vector that is truly safe. An immune response to the viral vector could make the vaccine less effective.
Existing examples	• None	• Measles, Mumps and Rubella • Chickenpox	• Polio	• Pertussis • Hepatitis B • Human papillomavirus (HPV)	• Ebola • Veterinary medicine
Group testing this approach for COVID-19	• Moderna (RNA) • Inovio (DNA)	• Codagenix • Indian Immunologicals Ltd.	• Sinovac • Sinopharm	• Novavax • AdaptVac	• University of Oxford & AstraZeneca • CanSino Biologics • Johnson & Johnson

Sources: CDC; NIAID; FDA

MICHELLE GUERRERO and JONATHAN WOSEN U-T

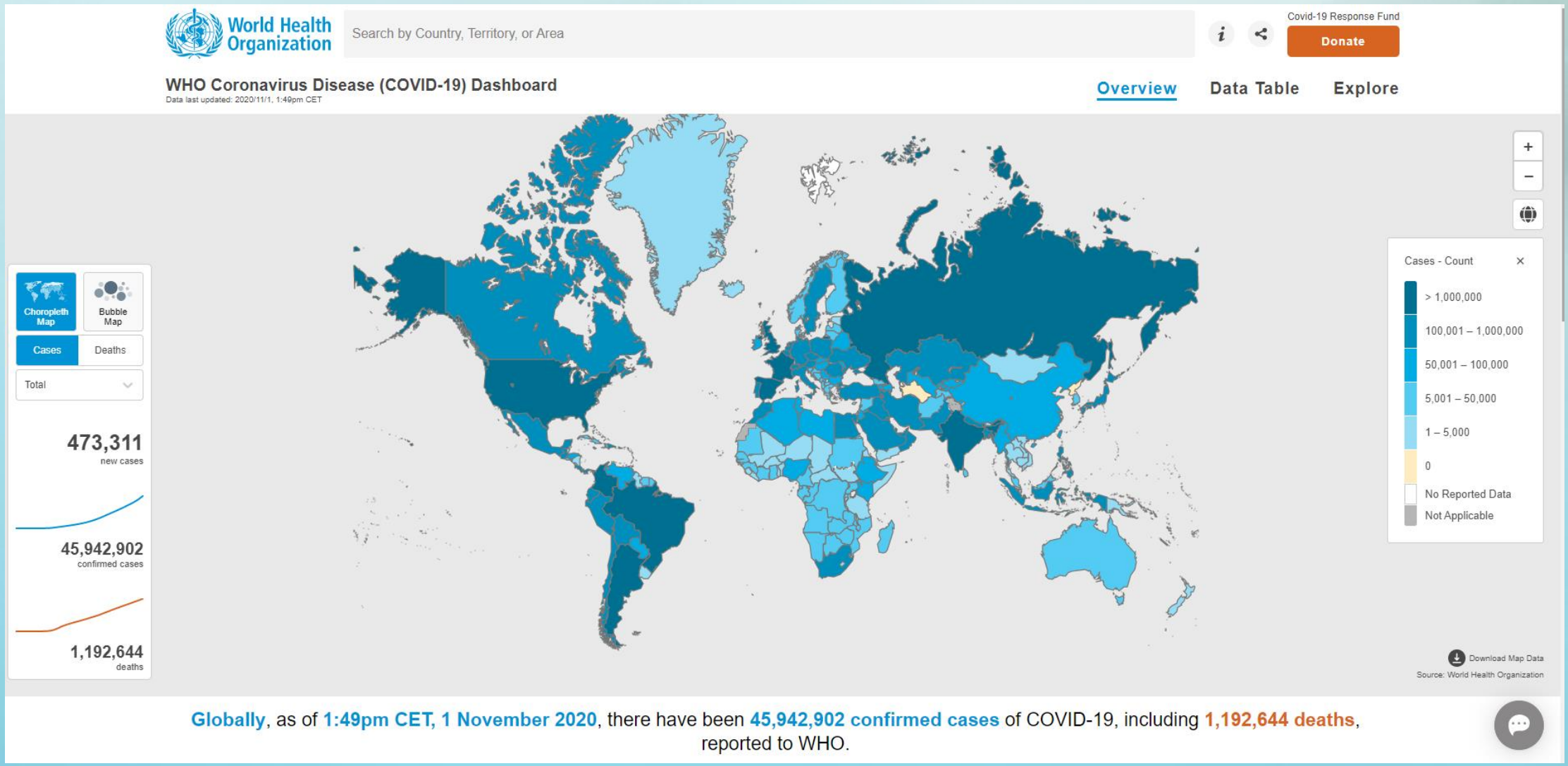


https://www.sandiegouniontribune.com/news/science/story/2020-06-06/race-for-vaccine
https://www.nature.com/articles/s41563-020-0746-0/figures/1

Public Health Level

COVID-19 at Public Health Level

https://covid19.who.int/



COVID-19 at Public Health Level

https://covid19.who.int/

World Health Organization

Covid-19 Response Fund

Donate

WHO Coronavirus Disease (COVID-19) Dashboard

Data last updated: 2020/11/1, 1:49pm CET

Overview | Data Table | Explore

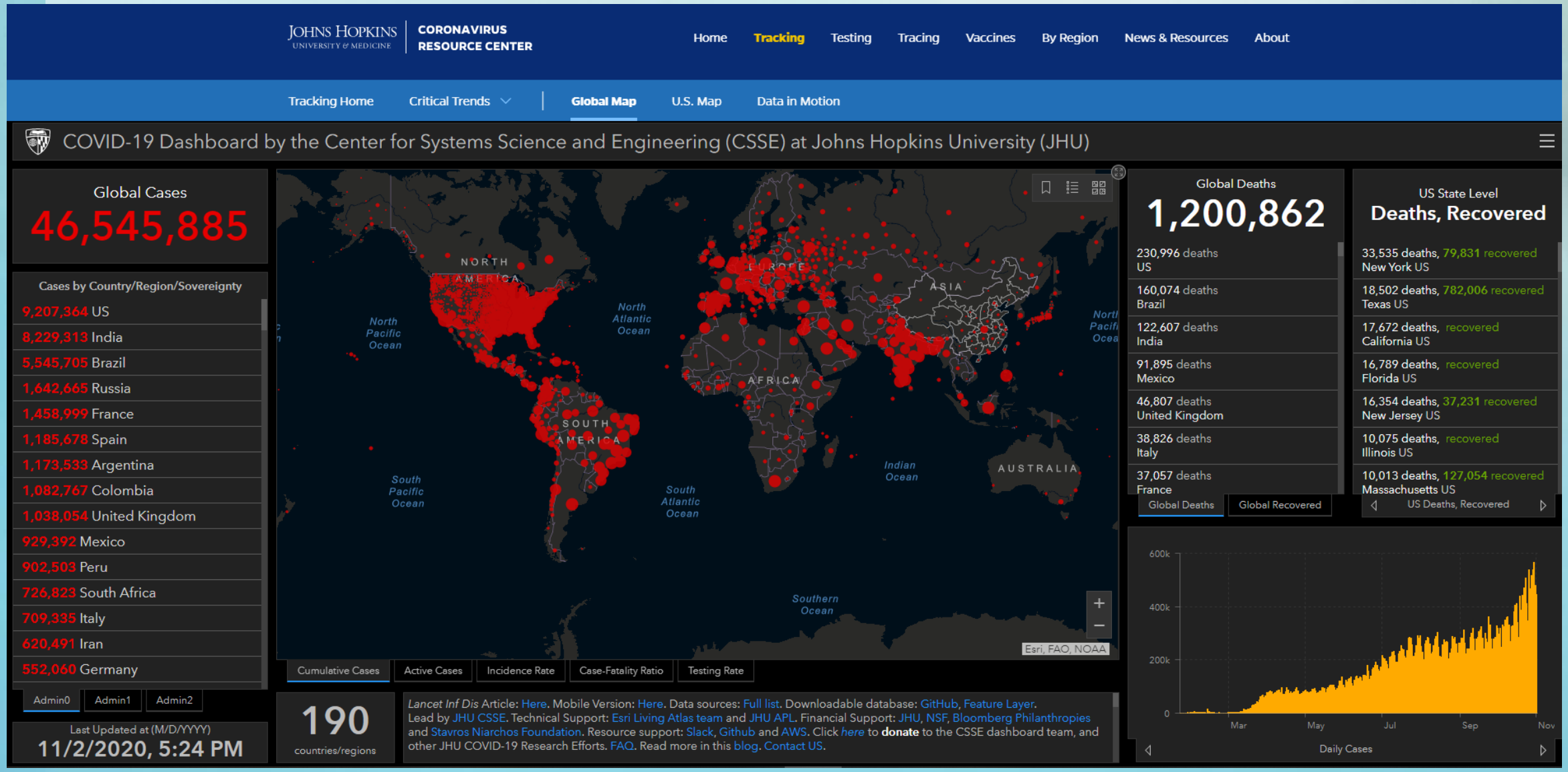
Latest | Yesterday | Search by Country, Territory, or Area

Situation by Country, Territory & Area

Name	Cases - cumulative total	Cases - newly reported in last 24 hours	Deaths - cumulative total	Deaths - newly reported in last 24 hours	Transmission Classification
Global	**45,942,902**	**473,311**	**1,192,644**	**6,578**	
United States of Ame...	8,952,086	99,356	228,185	1,007	Community transmission
India	8,184,082	46,963	122,111	470	Clusters of cases
Brazil	5,516,658	22,282	159,477	508	Community transmission
Russian Federation	1,636,781	18,665	28,235	245	Clusters of cases
France	1,331,808	32,530	36,473	223	Community transmission
Spain	1,185,678	0	35,878	0	Community transmission
Argentina	1,157,179	13,379	30,792	350	Community transmission
Colombia	1,063,151	10,029	31,135	209	Community transmission
The United Kingdom	1,011,664	21,915	46,555	326	Community transmission
Mexico	918,811	6,000	91,289	516	Community transmission

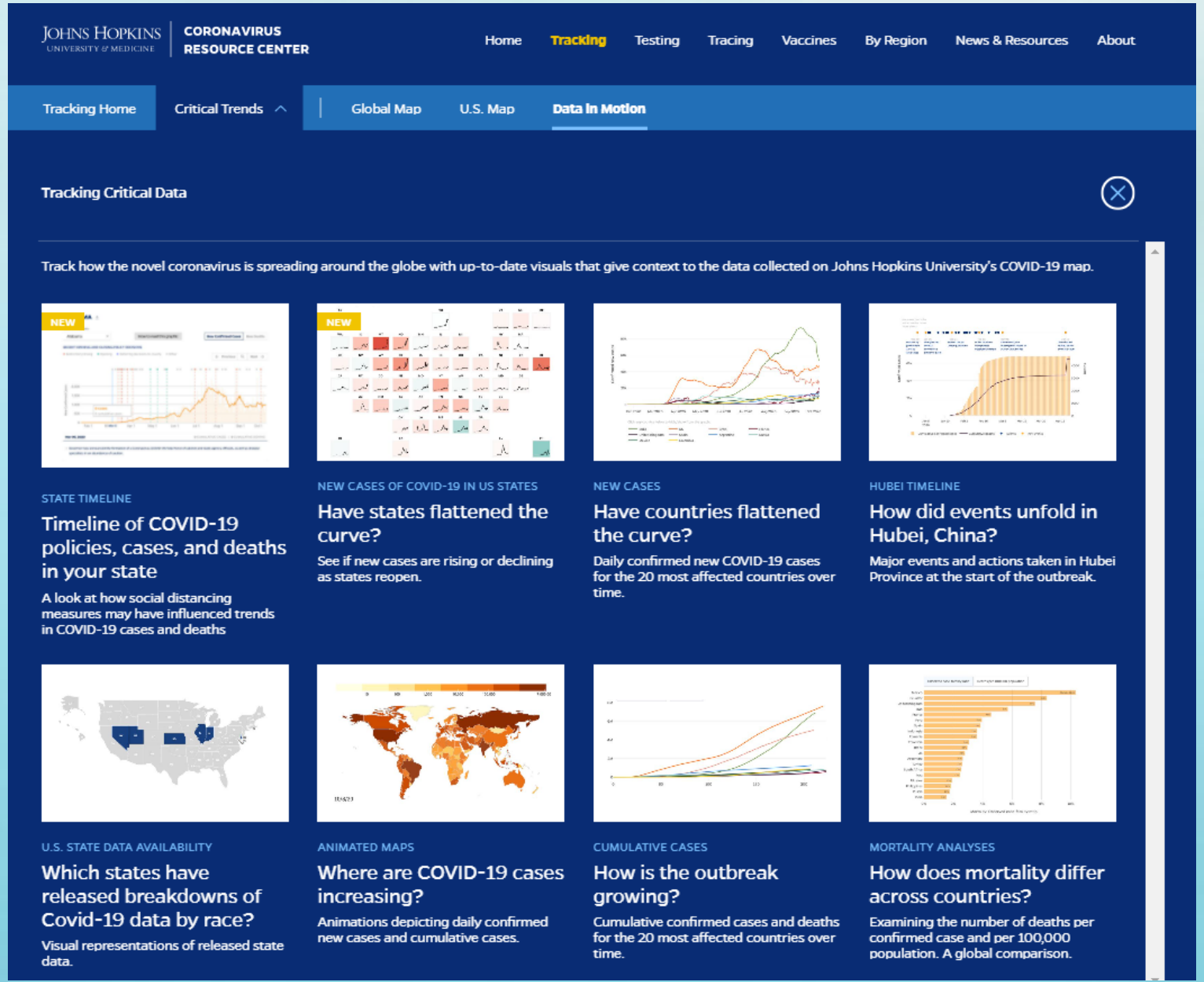
COVID-19 at Public Health Level

https://coronavirus.jhu.edu/map.html



COVID-19 at Public Health Level

https://coronavirus.jhu.edu/covid-19-daily-video



COVID-19 at Public Health Level

https://ourworldindata.org/coronavirus-data

Our World in Data | Articles by topic | Search... | Latest | About | Donate

All charts | Sustainable Development Goals Tracker

Coronavirus Pandemic (COVID-19) – the data

Research and data: Hannah Ritchie, Esteban Ortiz-Ospina, Diana Beltekian, Edouard Mathieu, Joe Hasell, Bobbie Macdonald, Charlie Giattino, and Max Roser
Web development: Breck Yunits, Ernst van Woerden, Daniel Gavrilov, Matthieu Bergel, Shahid Ahmad, and Jason Crawford

The data on the coronavirus pandemic is updated daily. Last update: **November 1, 2020** (10:00, London time).

Coronavirus > **By country** **Data explorer** Deaths Cases Tests Hospitalizations Mortality risk Excess mortality Policy responses Exemplars All charts

Our work belongs to everyone

Download the complete *Our World in Data* COVID-19 dataset
.xslx .csv .json (daily updated)

All our code is open-source
All our research and visualizations are free for everyone to use for all purposes

The purpose of this page here is simply to *lists all our visualizations* on the pandemic.

COVID-19 at Public Health Level

https://github.com/owid/covid-19-data/tree/master/public/data

Data on COVID-19 (coronavirus) by *Our World in Data*

Our complete COVID-19 dataset is a collection of the COVID-19 data maintained by *Our World in Data*. It is updated daily and includes data on confirmed cases, deaths, hospitalizations, and testing, as well as other variables of potential interest.

📁 Download our complete COVID-19 dataset : CSV | XLSX | JSON

We will continue to publish up-to-date data on confirmed cases, deaths, hospitalizations, and testing, throughout the duration of the COVID-19 pandemic.

Our data sources

- **Confirmed cases and deaths:** our data comes from the European Centre for Disease Prevention and Control (ECDC). We discuss how and when the ECDC collects and publishes this data here. The cases & deaths dataset is updated daily. *Note: the number of cases or deaths reported by any institution—including the ECDC, the WHO, Johns Hopkins and others—on a given day does not necessarily represent the actual number on that date. This is because of the long reporting chain that exists between a new case/death and its inclusion in statistics.* ***This also means that negative values in cases and deaths can sometimes appear when a country sends a correction to the ECDC, because it had previously overestimated the number of cases/deaths. Alternatively, large changes can sometimes (although rarely) be made to a country's entire time series if the ECDC decides (and has access to the necessary data) to correct values retrospectively.***
- **Hospitalizations and intensive care unit (ICU) admissions:** our data comes from the European Centre for Disease Prevention and Control (ECDC), who provide these statistics only for a select number of European countries, and update it on a weekly basis. Unfortunately, we are unable to provide data on hospitalizations for other countries: there is currently no global, aggregated database on COVID-19 hospitalization, and our team at *Our World in Data* does not have the capacity to build such a dataset.
- **Testing for COVID-19:** this data is collected by the *Our World in Data* team from official reports; you can find further details in our post on COVID-19 testing, including our checklist of questions to understand testing data, information on geographical and temporal coverage, and detailed country-by-country source information. The testing dataset is updated around twice a week.
- **Other variables:** this data is collected from a variety of sources (United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government, etc.). More information is available in our codebook.

The complete *Our World in Data* COVID-19 dataset

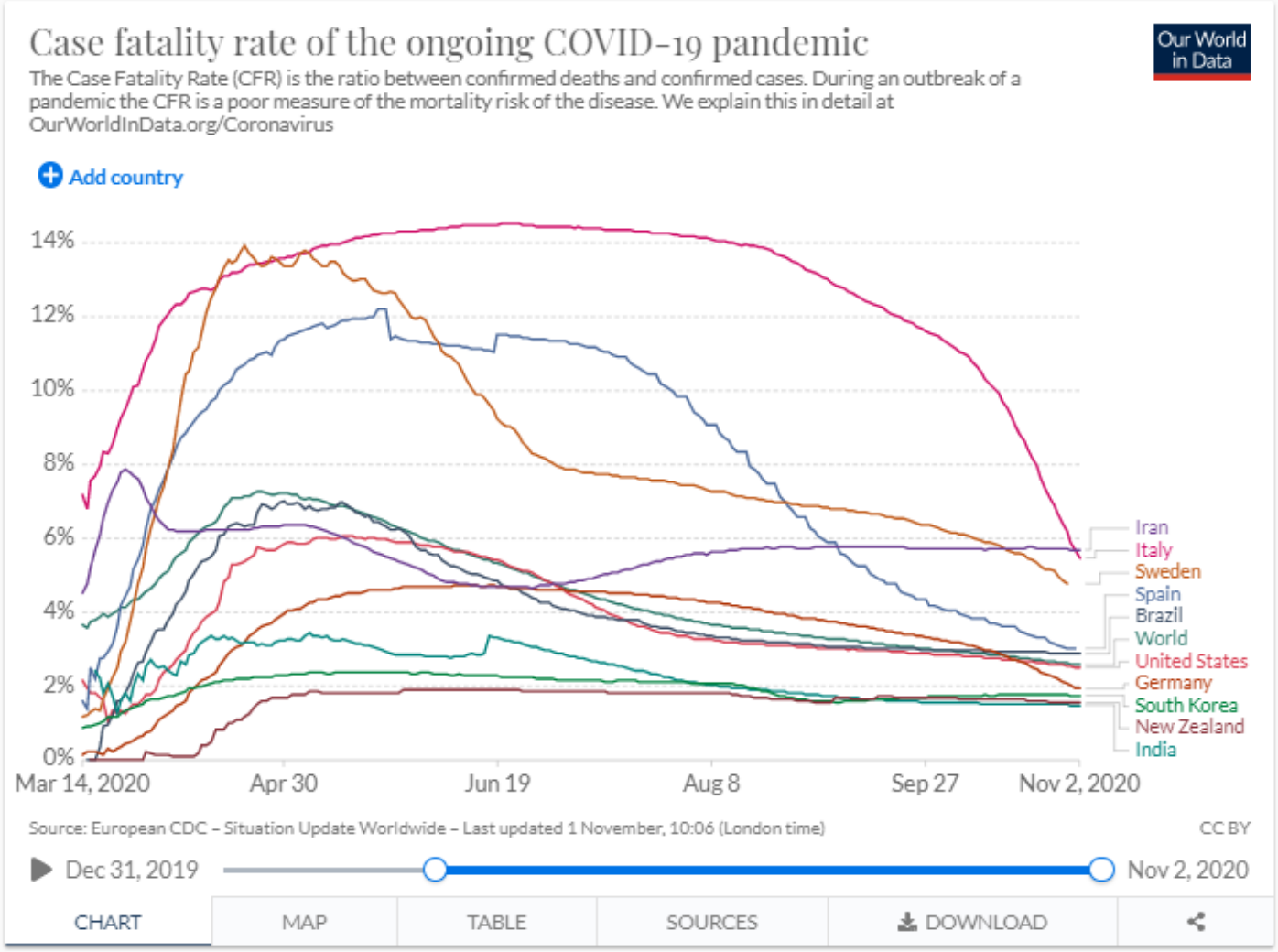
Our complete COVID-19 dataset is available in CSV, XLSX, and JSON formats, and includes all of our historical data on the pandemic up to the date of publication.

The CSV and XLSX files follow a format of 1 row per location and date. The JSON version is split by country ISO code, with static variables and an array of daily records.

COVID-19 at Public Health Level

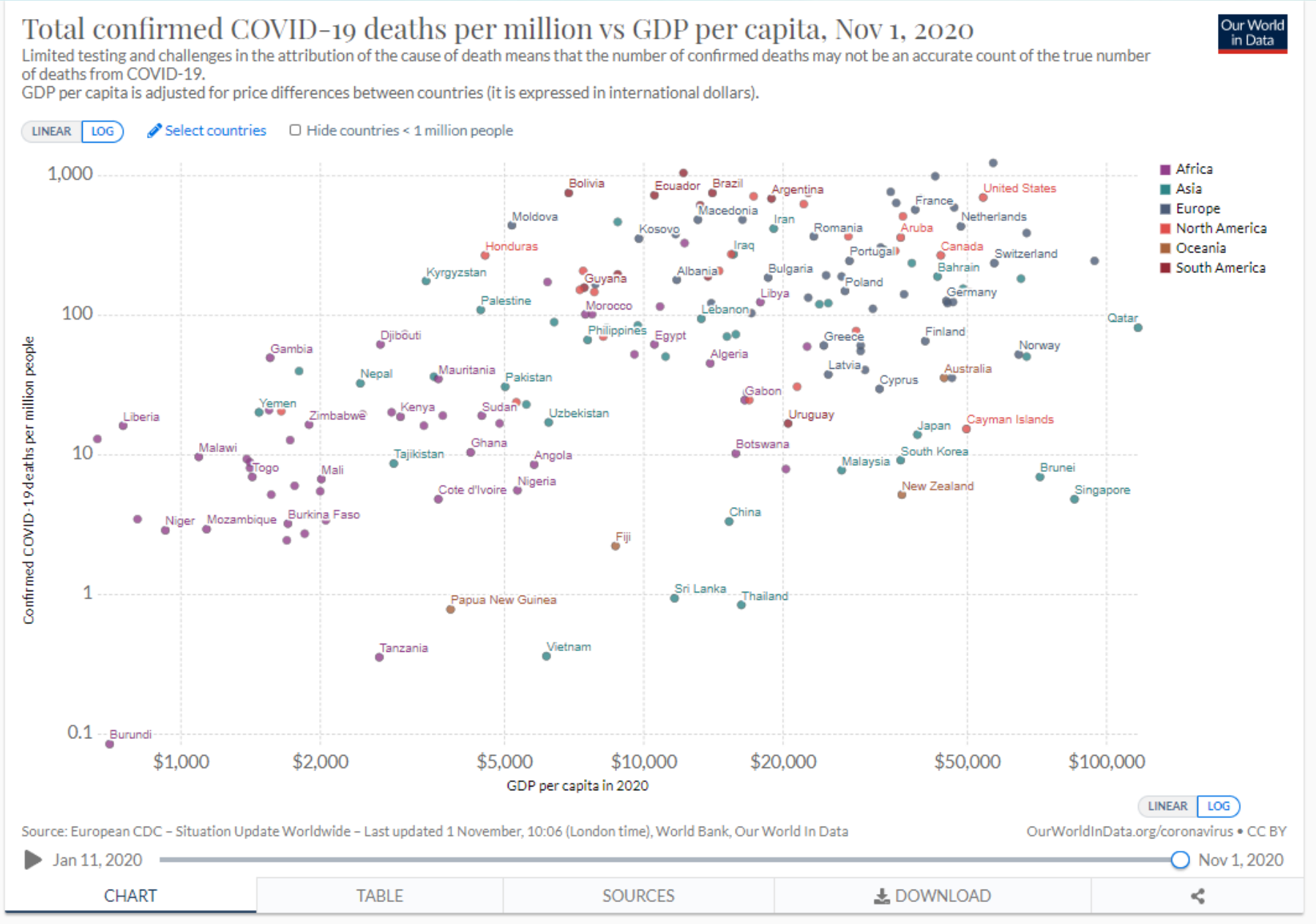
https://ourworldindata.org/mortality-risk-covid

The case fatality rate



COVID-19 at Public Health Level

https://ourworldindata.org/grapher/total-confirmed-deaths-of-covid-19-per-million-people-vs-gdp-per-capita?time=2020-11-01..latest



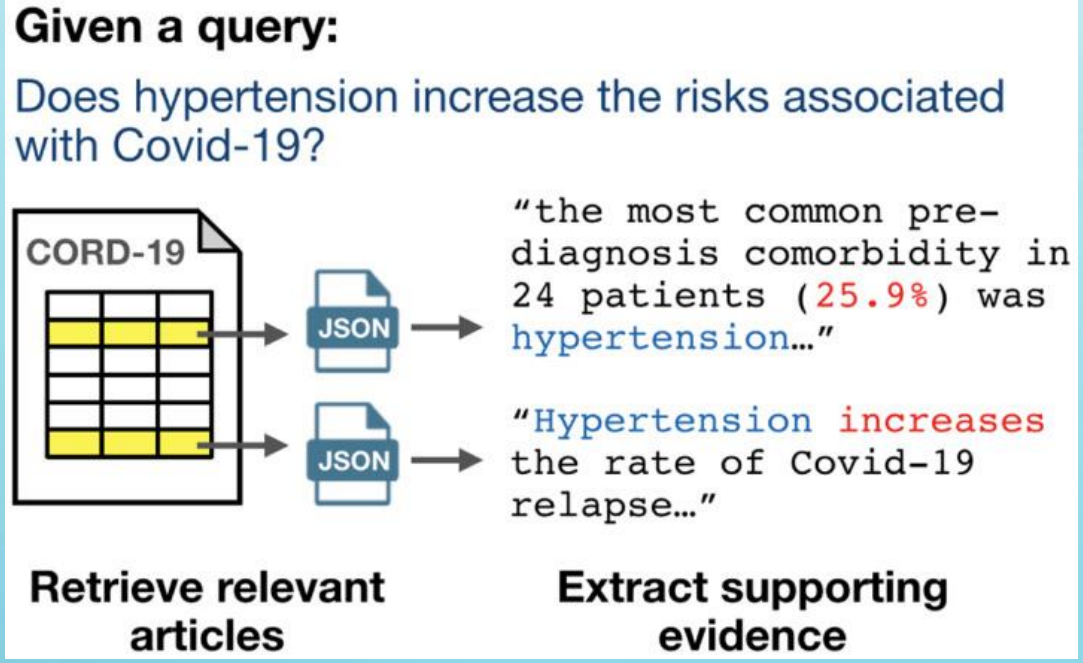
Text Level

COVID-19 at Text Level

- Text Mining on Academic Papers
- https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
- Over 200,000 academic papers have been collected.
- Total Data Size: 22 GB (as of 2nd Nov 2020)

COVID-19 at Text Level

- Text Mining on Academic Papers
- https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955/

COVID-19 at Text Level

- Text Mining on Tweets

https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset

COVID-19 at Text Level

- — Number of tweets : 694,080,653 tweets
- — Coverage : Global
- — Language : English (EN)
- — Geo-tagged Version: Coronavirus (COVID-19) Geo-tagged Tweets Dataset (GeoCOV19Tweets Dataset)
- — Dataset updates : Everyday
- — Usage policy : As per Twitter's Developer Policy
- — Active keywords and hashtags (archive: keywords.tsv) : "corona", "#corona", "coronavirus", "#coronavirus", "covid", "#covid", "covid19", "#covid19", "covid-19", "#covid-19", "sarscov2", "#sarscov2", "sars cov2", "sars cov 2", "covid_19", "#covid_19", "#ncov", "ncov", "#ncov2019", "ncov2019", "2019-ncov", "#2019-ncov", "pandemic", "#pandemic" "#2019ncov", "2019ncov", "quarantine", "#quarantine", "flatten the curve", "flattening the curve", "#flatteningthecurve", "#flattenthecurve", "hand sanitizer", "#handsanitizer", "#lockdown", "lockdown", "social distancing", "#socialdistancing", "work from home", "#workfromhome", "working from home", "#workingfromhome", "ppe", "n95", "#ppe", "#n95", "#covidiots", "covidiots", "herd immunity", "#herdimmunity", "pneumonia", "#pneumonia", "chinese virus", "#chinesevirus", "wuhan virus", "#wuhanvirus", "kung flu", "#kungflu", "wearamask", "#wearamask", "wear a mask", "vaccine", "vaccines", "#vaccine", "#vaccines", "corona vaccine", "corona vaccines", "#coronavaccine", "#coronavaccines", "face shield", "#faceshield", "face shields", "#faceshields", "health worker", "#healthworker", "health workers", "#healthworkers", "#stayhomestaysafe", "#coronaupdate", "#frontlineheroes", "#coronawarriors", "#homeschool", "#homeschooling", "#hometasking", "#masks4all", "#wfh", "wash ur hands", "wash your hands", "#washurhands", "#washyourhands", "#stayathome", "#stayhome", "#selfisolating", "self isolating"

https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset

COVID-19 at Text Level

- Aguilar-Gallegos, N., Romero-García, L. E., Martínez-González, E. G., García-Sánchez, E. I., & Aguilar-Ávila, J. (2020). Dataset on dynamics of Coronavirus on Twitter. *Data in Brief*, 105684.

COVID-19 at Text Level

- Lopez, C. E., Vasu, M., & Gallemore, C. (2020). Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. arXiv preprint arXiv:2003.10359.

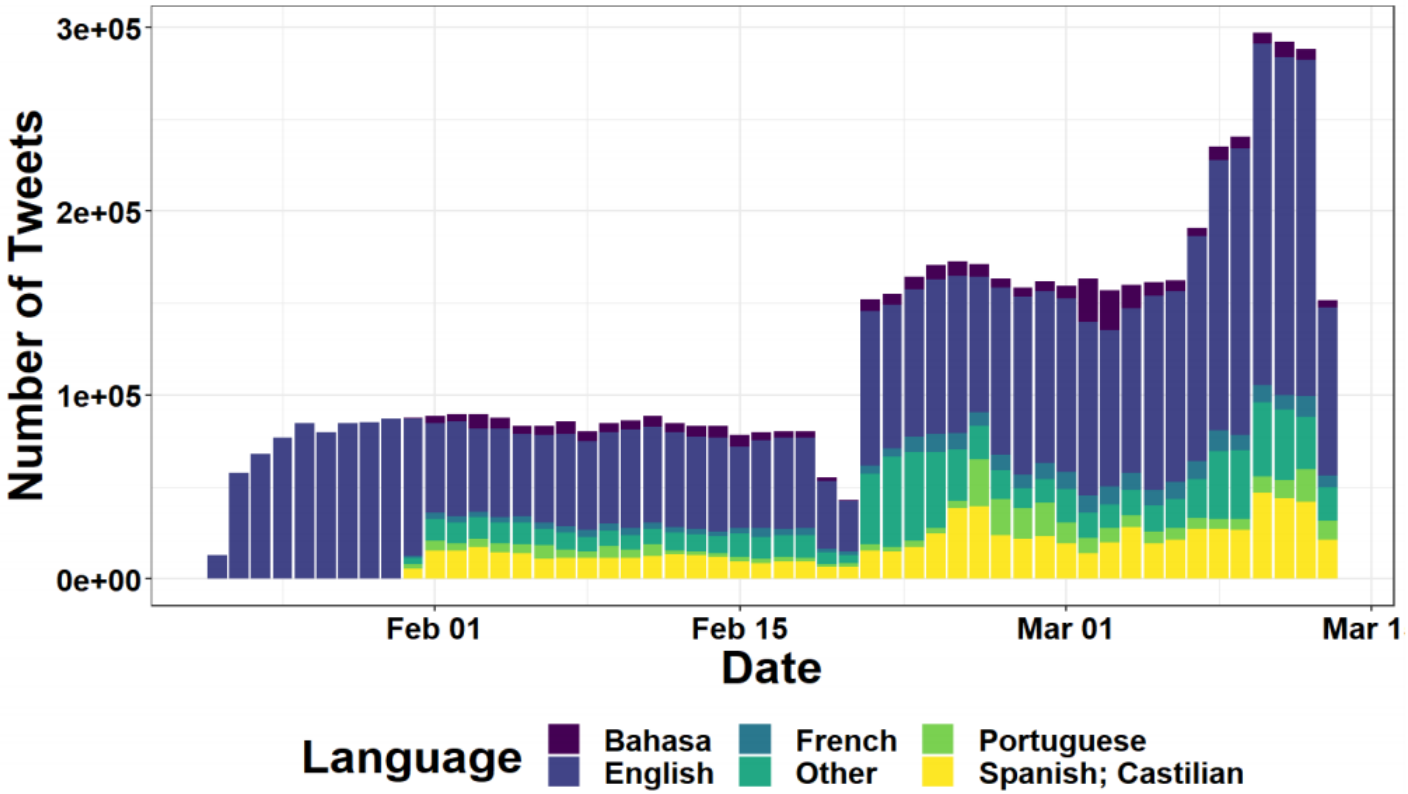
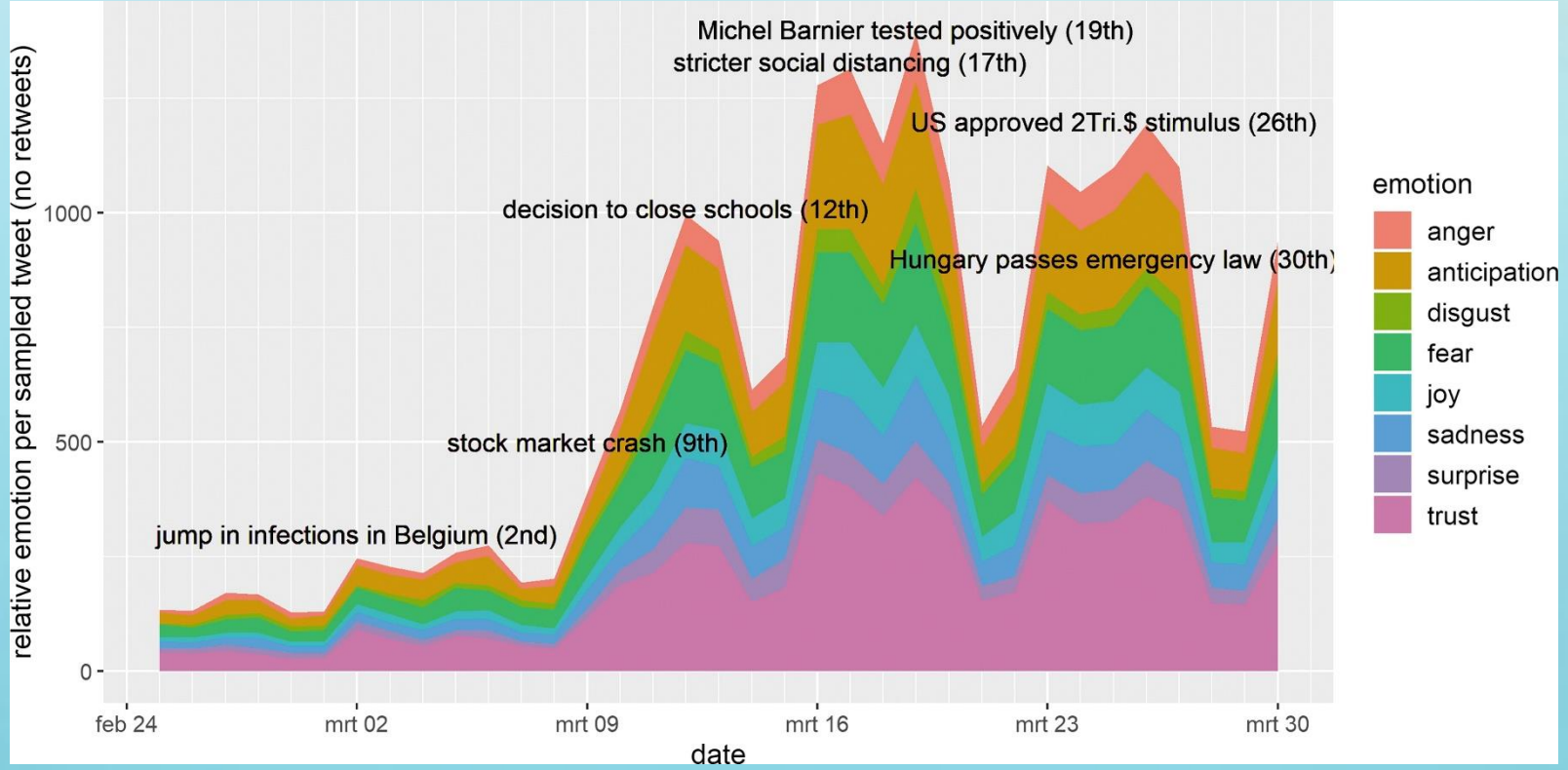


Figure 1. Collected tweets, by language, as of 13 March 2020.

COVID-19 at Text Level

- Kurten, S., & Beullens, K. (2020). # Coronavirus: Monitoring the Belgian Twitter Discourse on the Severe Acute Respiratory Syndrome Coronavirus 2 Pandemic. Cyberpsychology, Behavior, and Social Networking.

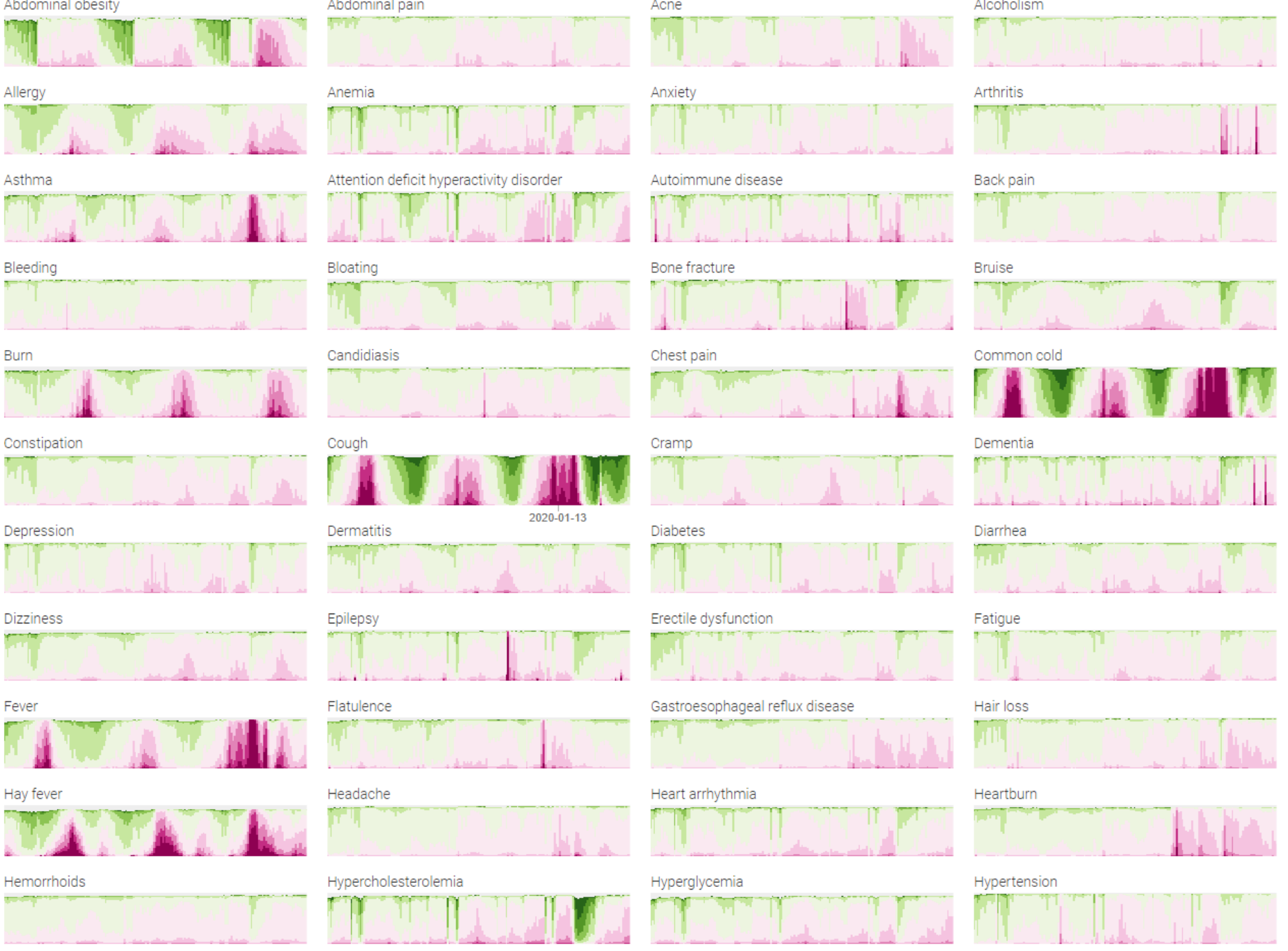


COVID-19 at Text Level

- Google Search Keyword Analysis
- https://trends.google.com/trends/story/US_cu_4Rjdh3ABAABMHM_en

The COVID-19 Search Trends symptoms dataset shows aggregated, anonymized trends in Google searches for more than 400 health symptoms, signs, and conditions, such as cough, fever and difficulty breathing. The dataset provides a time series for each region showing the relative volume of searches for each symptom.

Google Search Trend from 2017 to 2020



END