UC San Diego

PDMI COMPUTER SCIENCE CLUB

UNIVERSITY OF LOUISVILLE

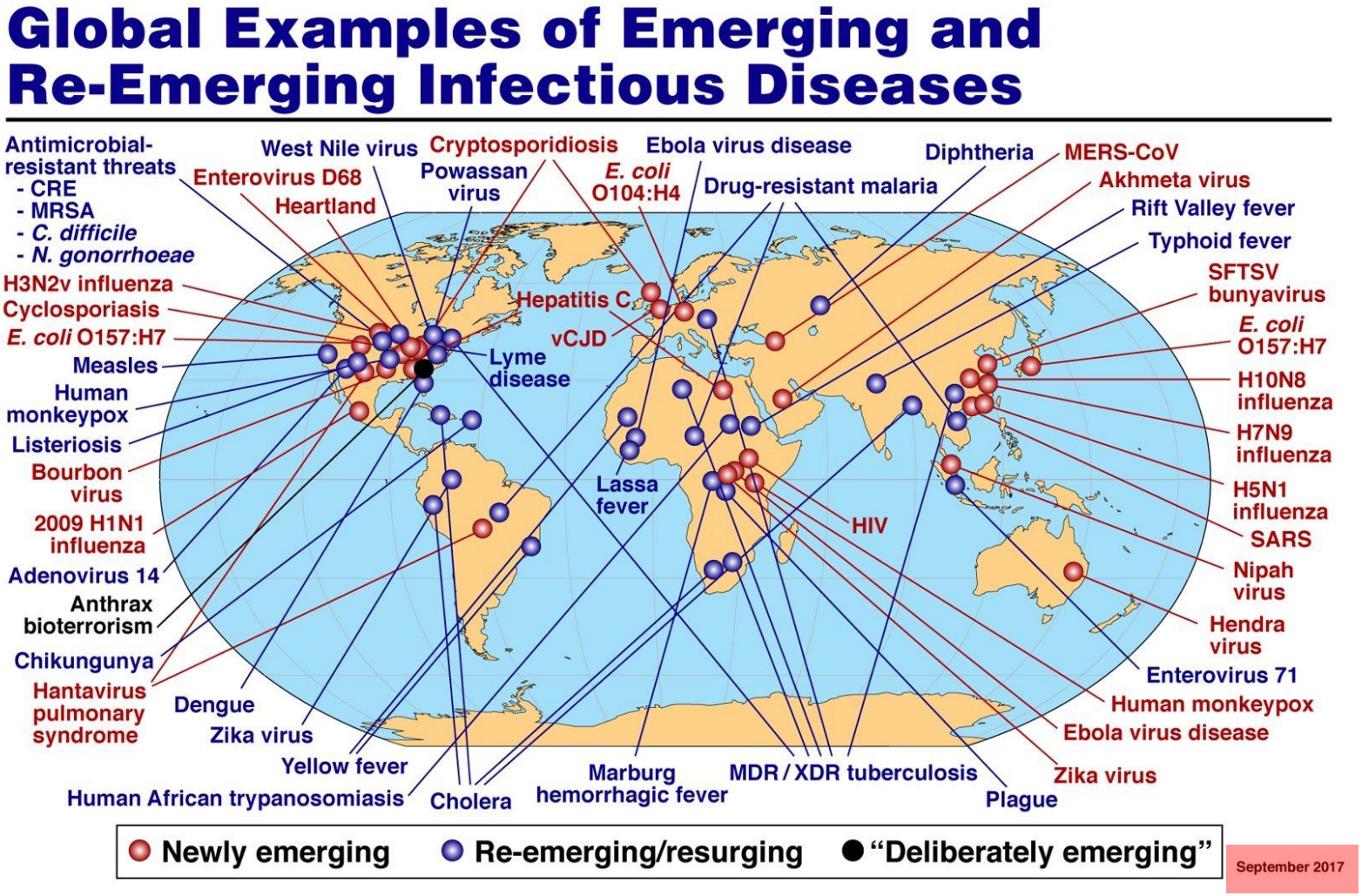
Computer Science meets Immunology:
how computational analyses help to study diseases

Yana Safonova
postdoctoral researcher, PhD

University of California San Diego
University of Louisville School of Medicine

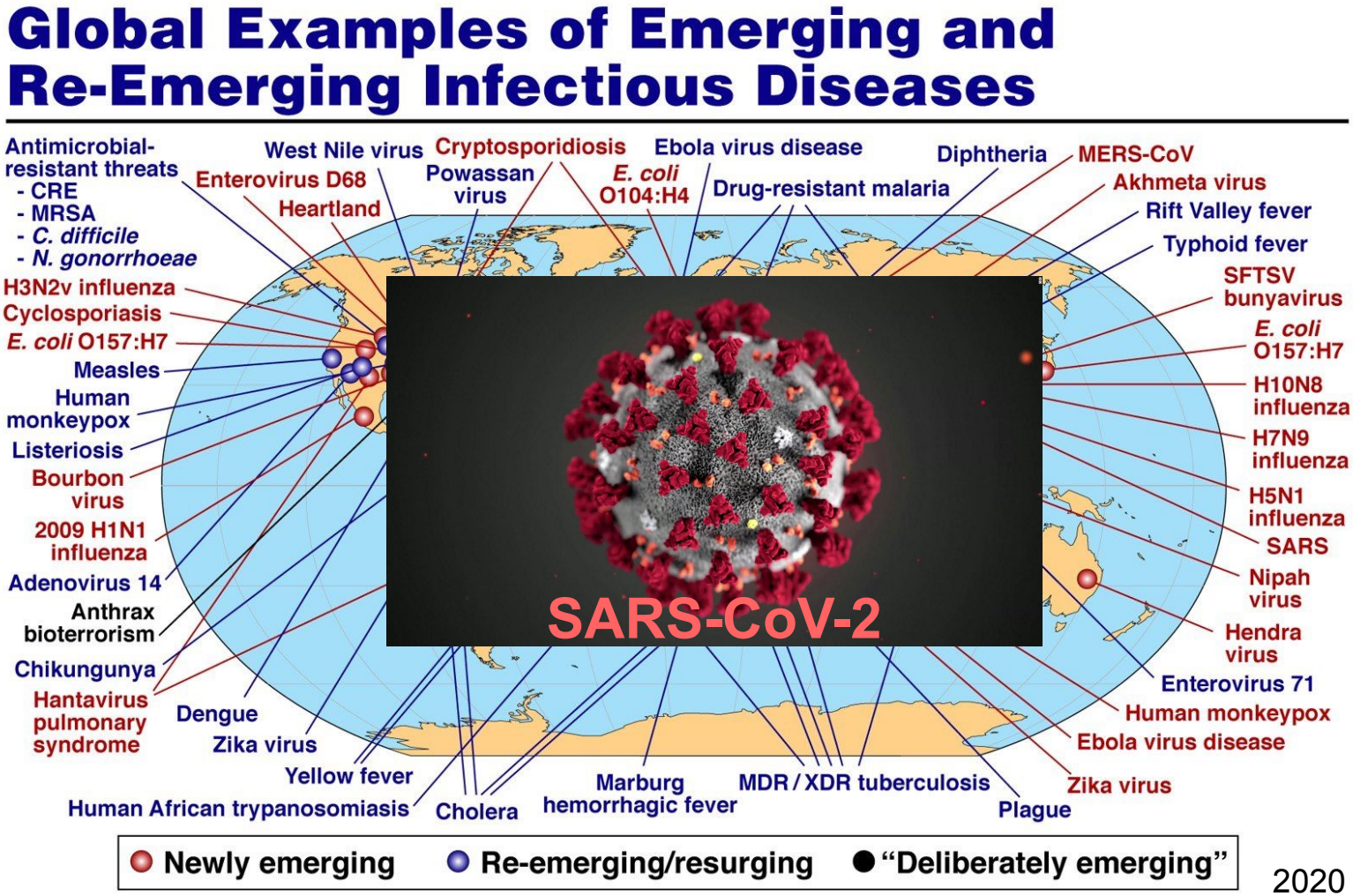
@yana_safonova_

Newly emerging and re-emerging diseases

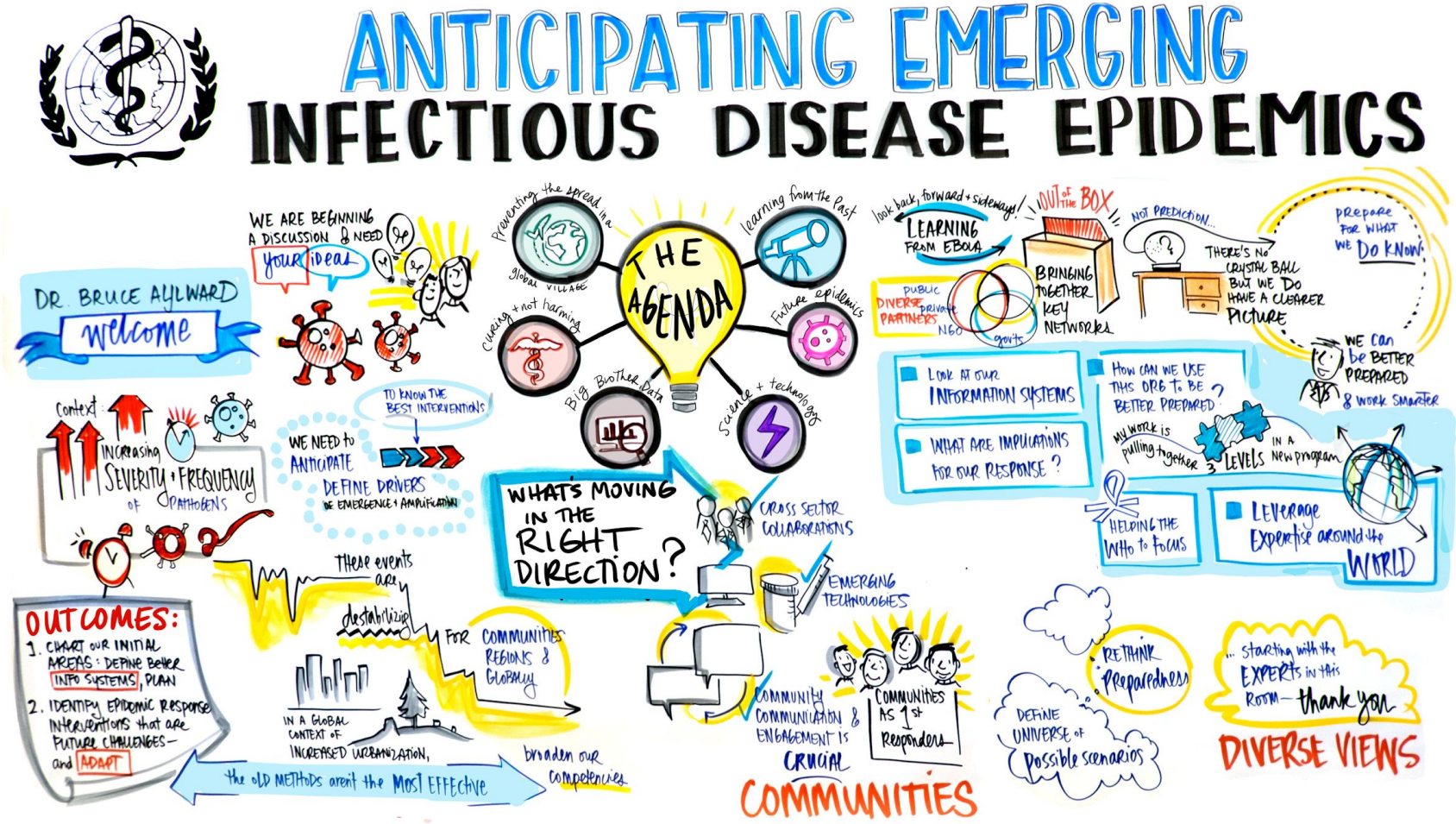


https://en.wikipedia.org/wiki/Emerging_infectious_disease

Newly emerging and re-emerging diseases

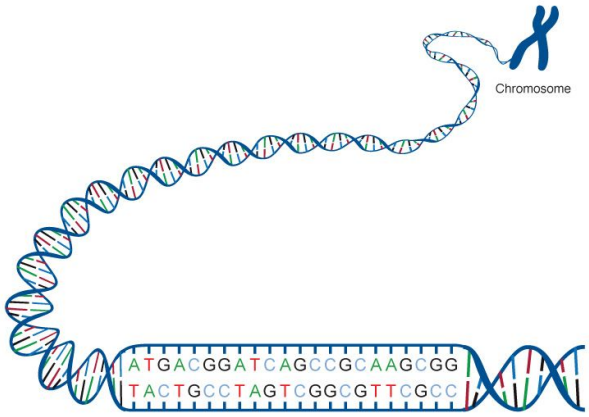


Response to (re)emerging diseases

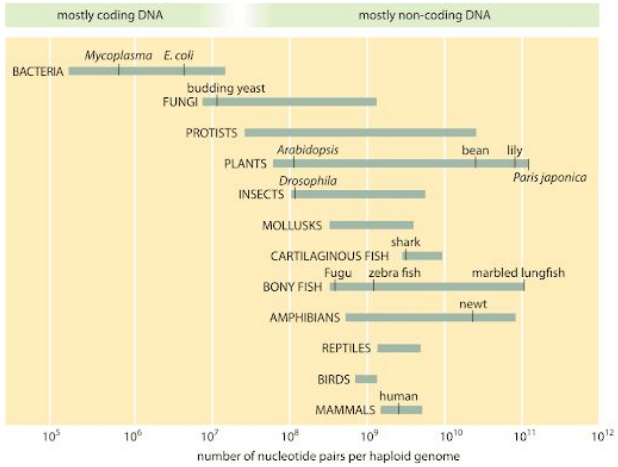


Response to (re)emerging diseases

Introduction to molecular biology

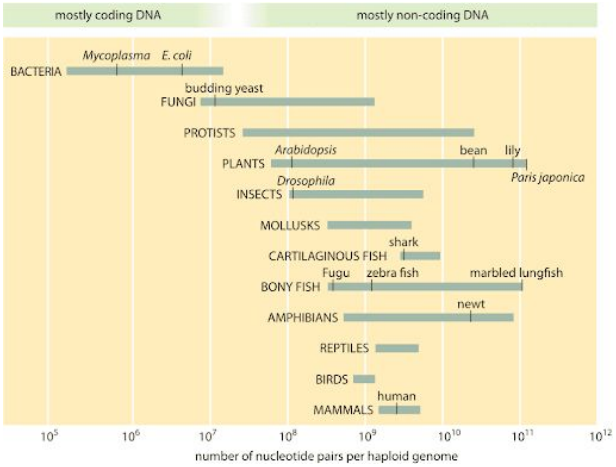


Introduction to molecular biology

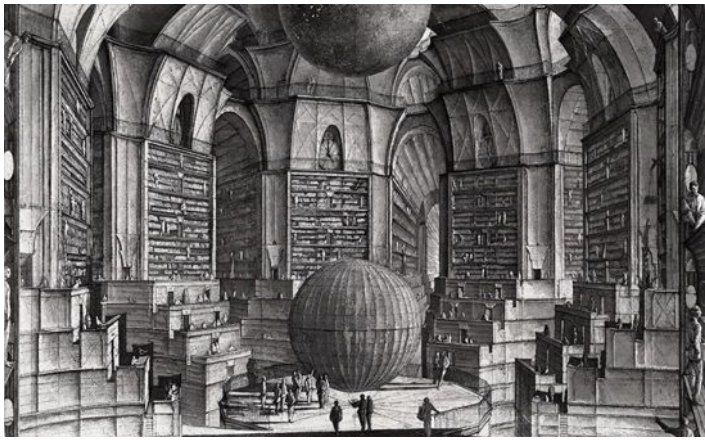


Human genome ~ 3Gbp

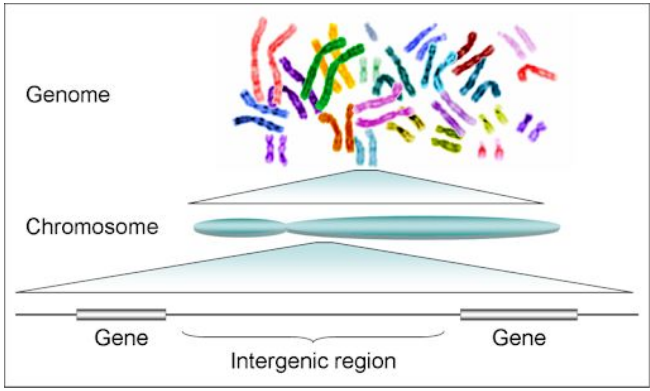
Introduction to molecular biology



Human genome ~ 3Gbp

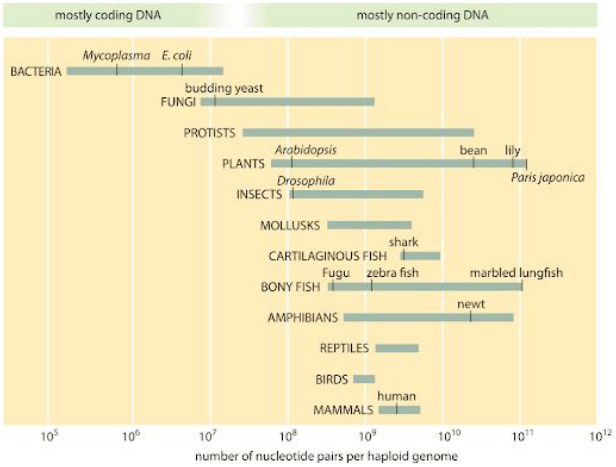


The Library of Babel (H. L. Borges) contains all possible combinations of symbols, mostly meaningless ones



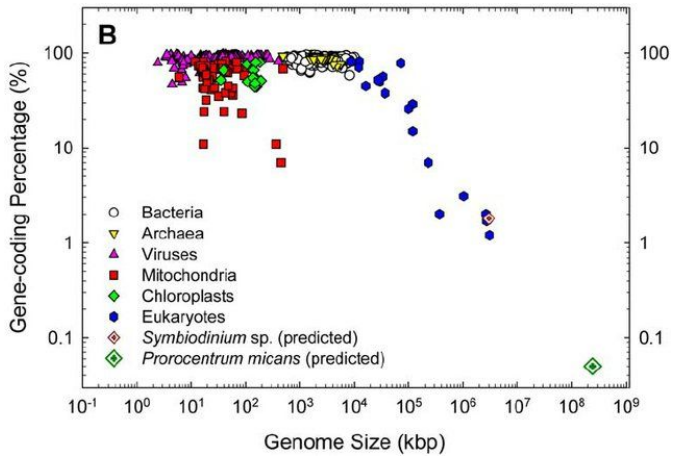
Genes in human genome take 1-2% of its length

Introduction to molecular biology

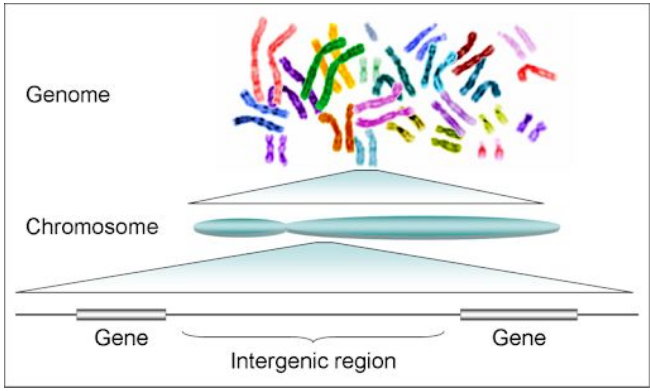


Human genome ~ 3Gbp

The Library of Babel (H. L. Borges) contains all possible combinations of symbols, mostly meaningless ones

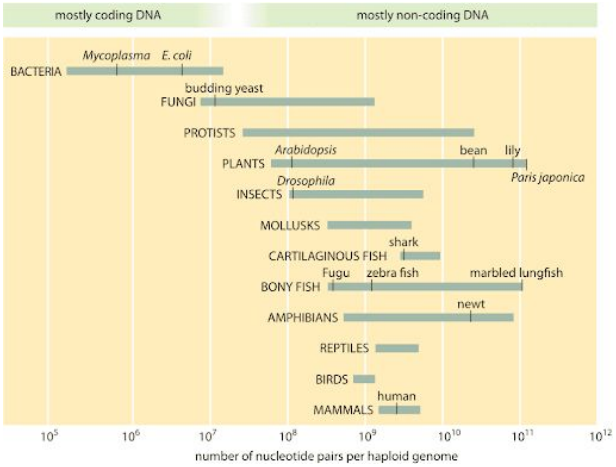


Hou and Lin, PLoS ONE, 2009



Genes in human genome take 1-2% of its length

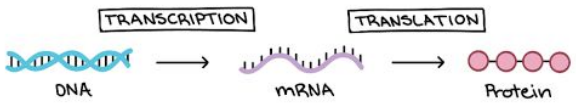
Introduction to molecular biology

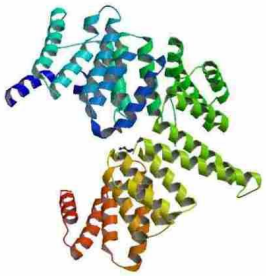


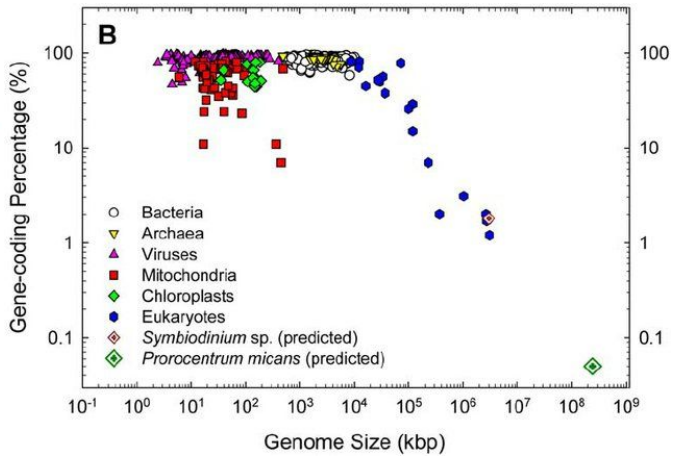
Human genome ~ 3Gbp

The Library of Babel (H. L. Borges) contains all possible combinations of symbols, mostly meaningless ones

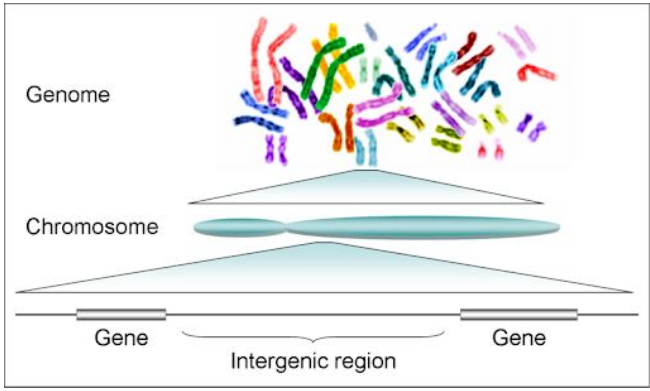
The central dogma of biology





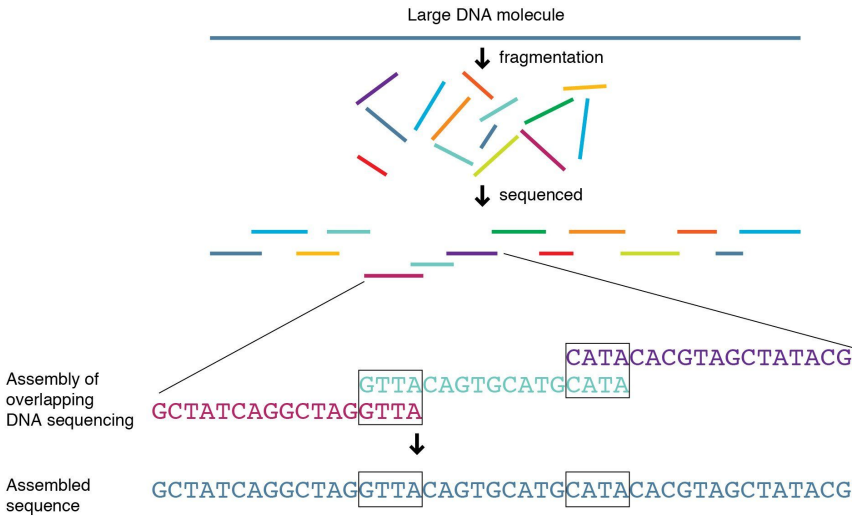


Hou and Lin, PLoS ONE, 2009

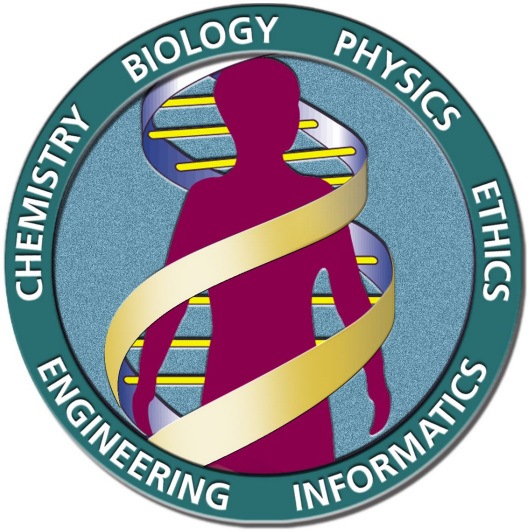


Genes in human genome take 1-2% of its length

The birth of bioinformatics

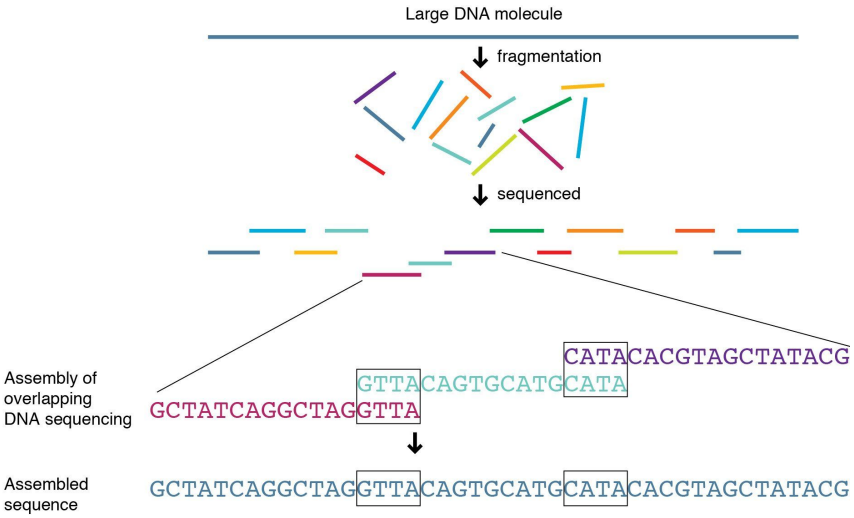


Genome assembly

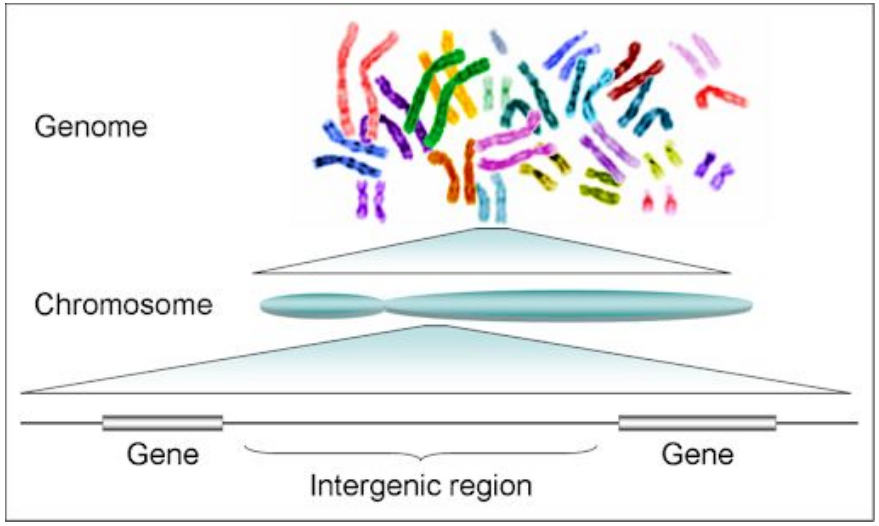


Human Genome Project
1990 – 2003

The birth of bioinformatics

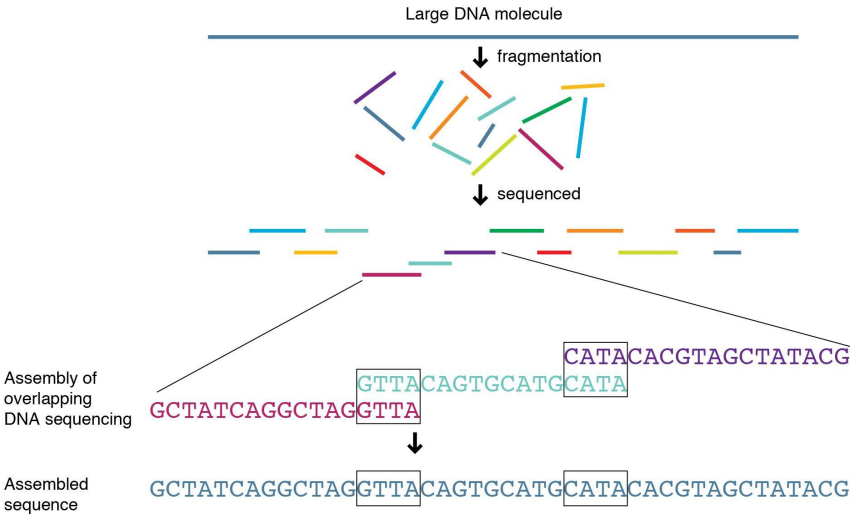


Genome assembly

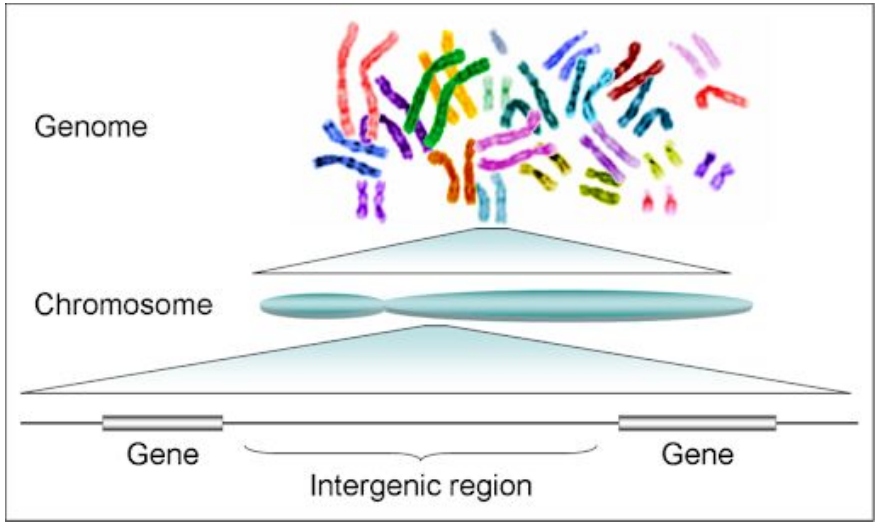


Gene prediction

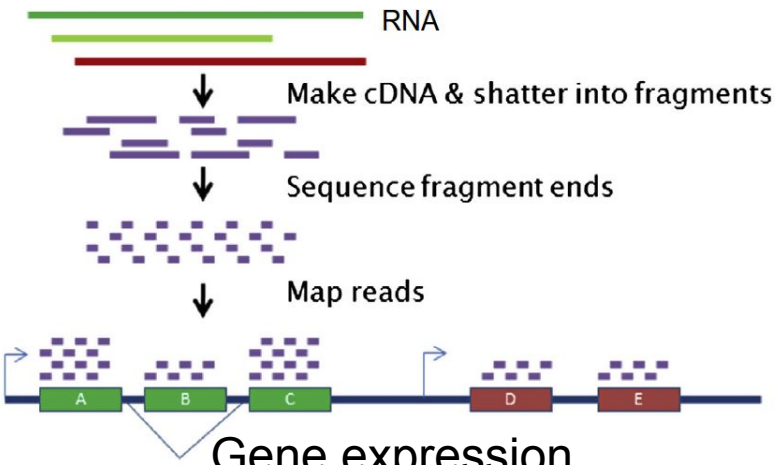
The birth of bioinformatics



Genome assembly

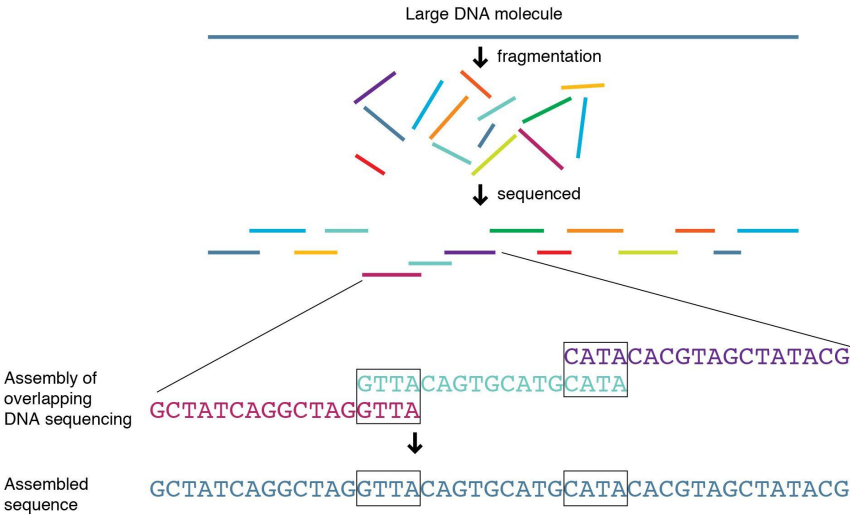


Gene prediction

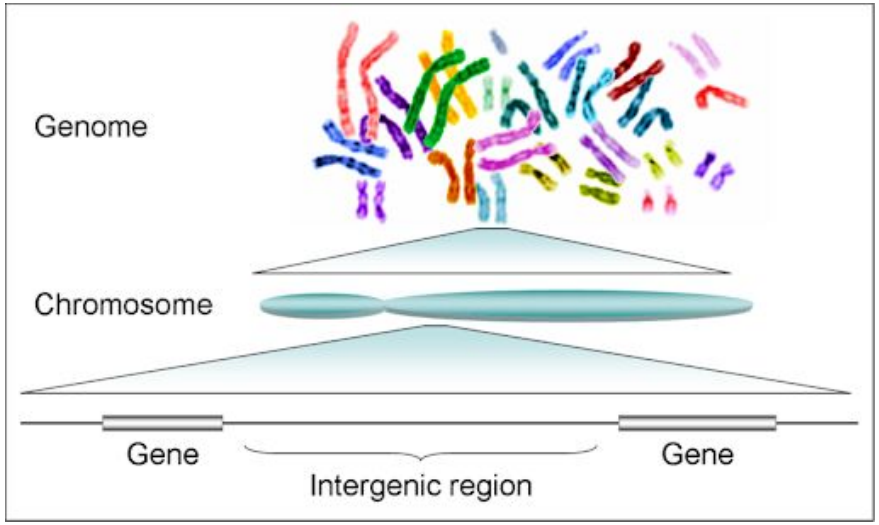


Gene expression

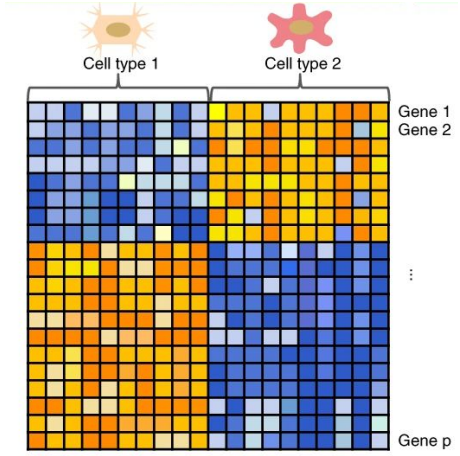
The birth of bioinformatics



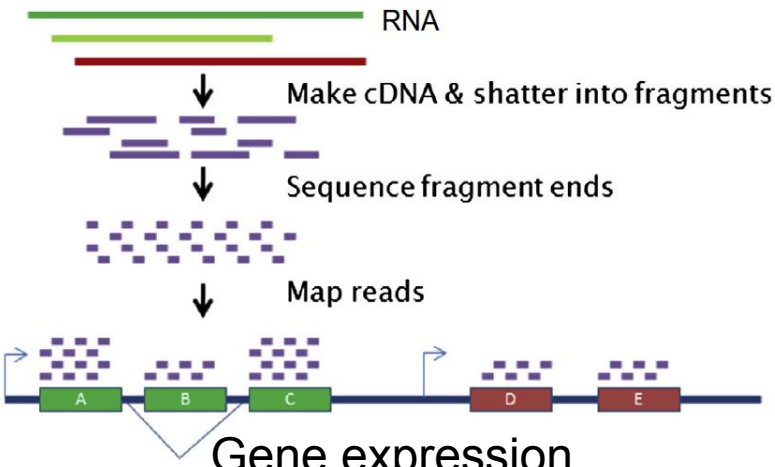
Genome assembly



Gene prediction

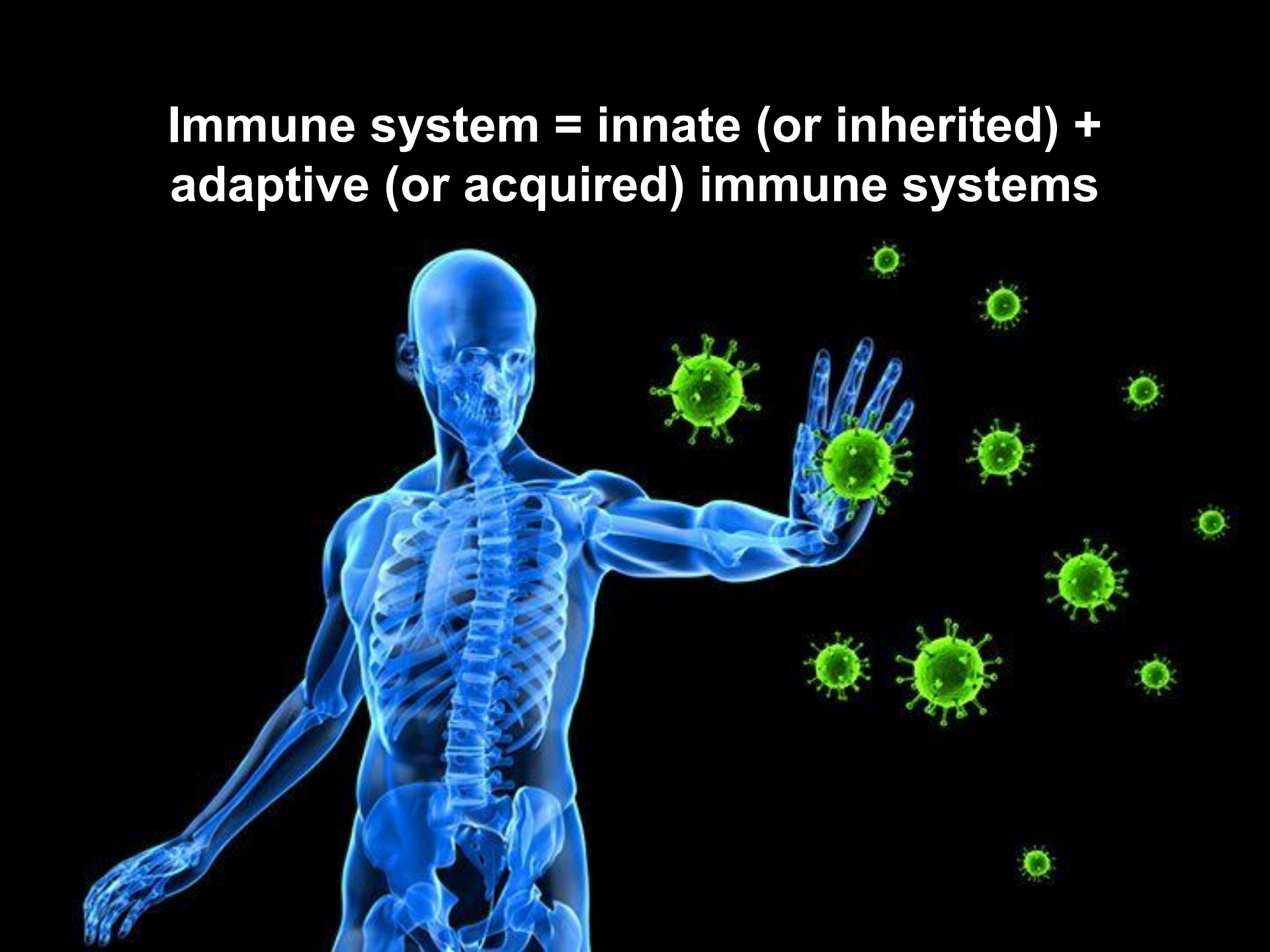


Differential gene expression



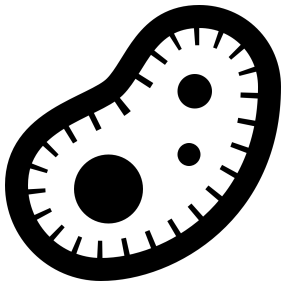
Gene expression

Immune system = innate (or inherited) + adaptive (or acquired) immune systems

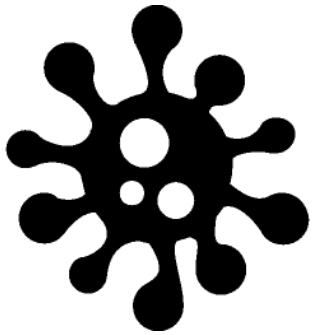


Adaptive immune system

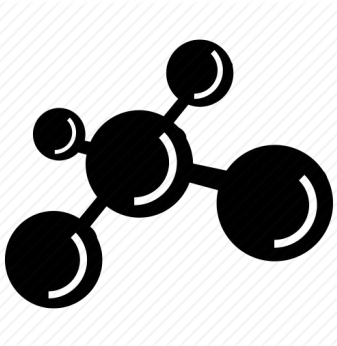
- Variety of threats to human body is huge and unpredictable



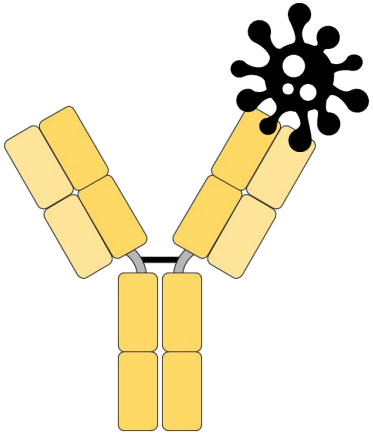
- Genome is too small to encode defences against all these threats



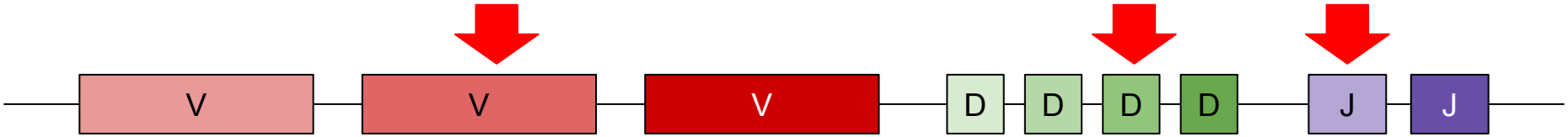
- Immune system has an ability to adapt to various threats using agents (e.g., antibodies) that **are not encoded** in the genome.



Antibodies are agents of the adaptive immune system

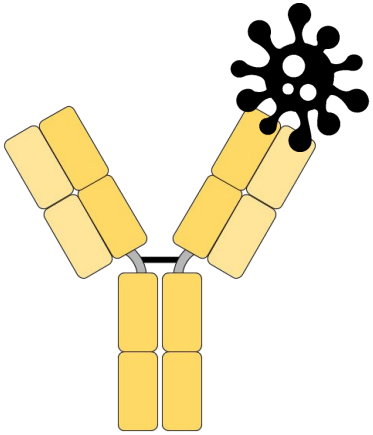


- Antibodies are proteins that bind to an antigen and cause its neutralization
- Antibodies are not encoded in the genome directly, but present a result of somatic genomic recombination

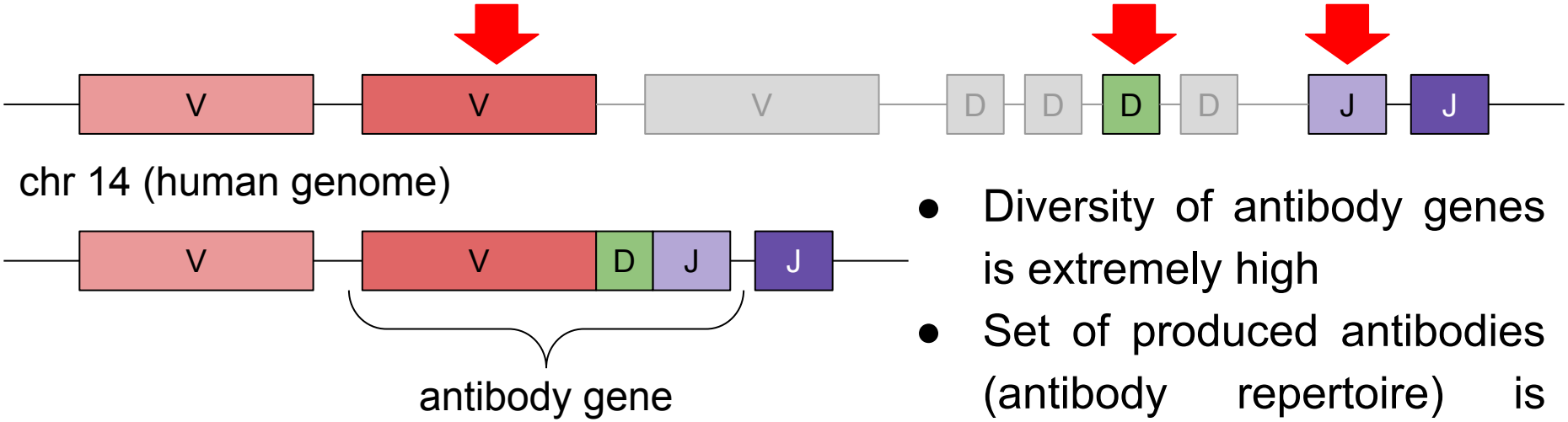


chr 14 (human genome)

Antibodies are agents of the adaptive immune system

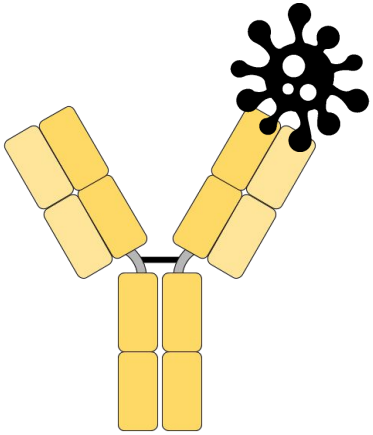


- Antibodies are proteins that bind to an antigen and cause its neutralization
- Antibodies are not encoded in the genome directly, but present a result of somatic genomic recombination



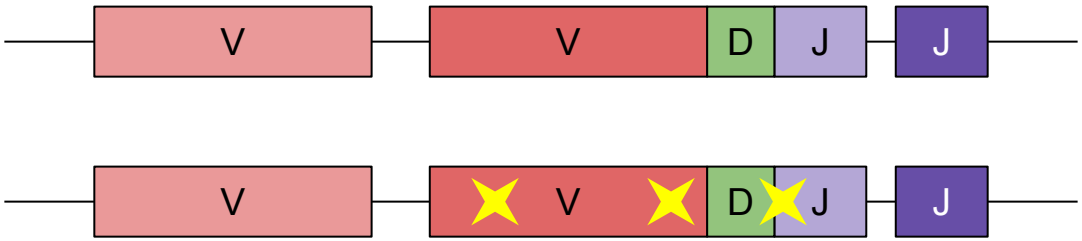
- Diversity of antibody genes is extremely high
- Set of produced antibodies (antibody repertoire) is unique for an individual

Antibodies are agents of the adaptive immune system



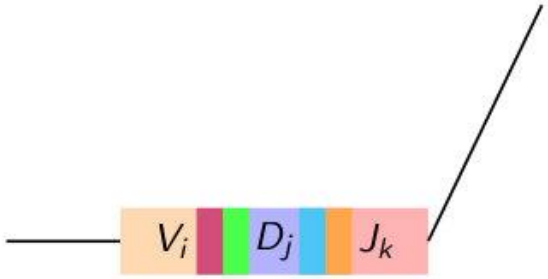
- Antibodies are proteins that bind to an antigen and cause its neutralization
- Antibodies are not encoded in the genome directly, but present a result of somatic genomic recombination

chr 14 (human genome)

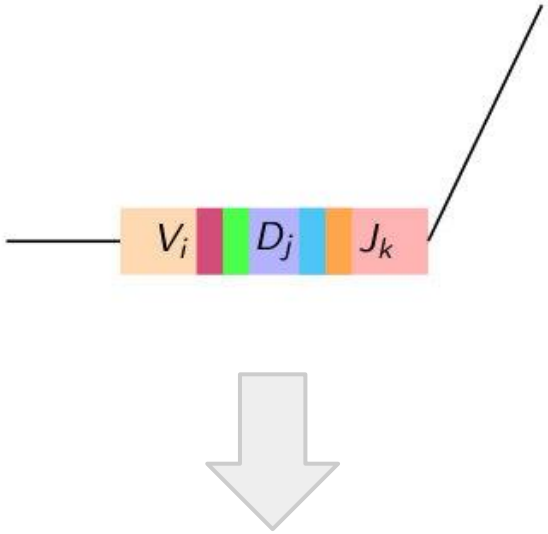


- Diversity of antibody genes is extremely high
- Set of produced antibodies (antibody repertoire) is unique for an individual

Why are antibodies so versatile if there are only 55×23×6 VDJ recombinations?

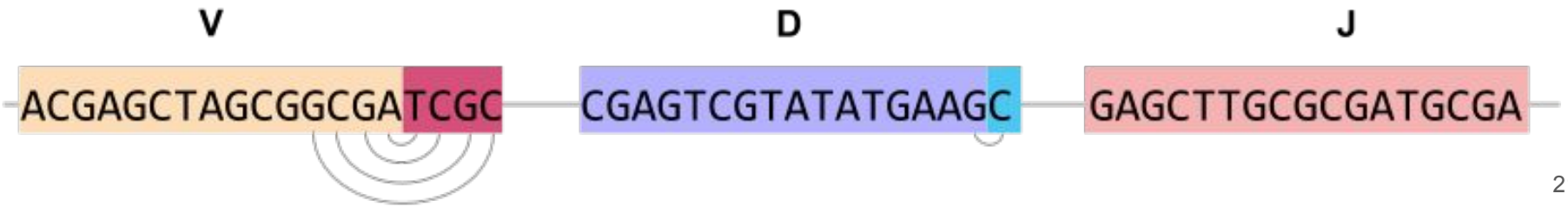


Why are antibodies so versatile if there are only 55×23×6 VDJ recombinations?

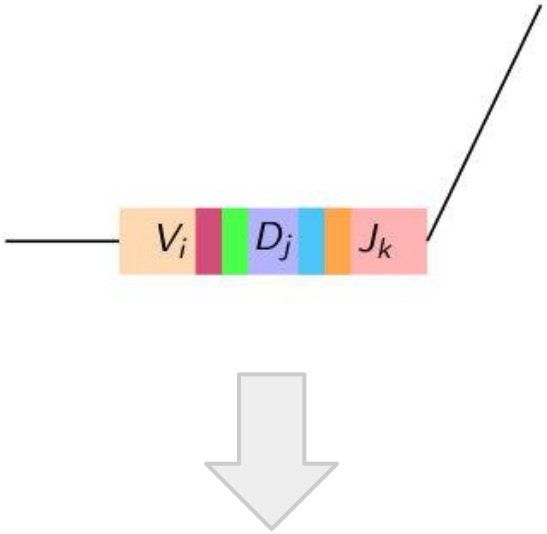


Recombination process is imperfect and includes many random processes:

- **Palindromic insertions**

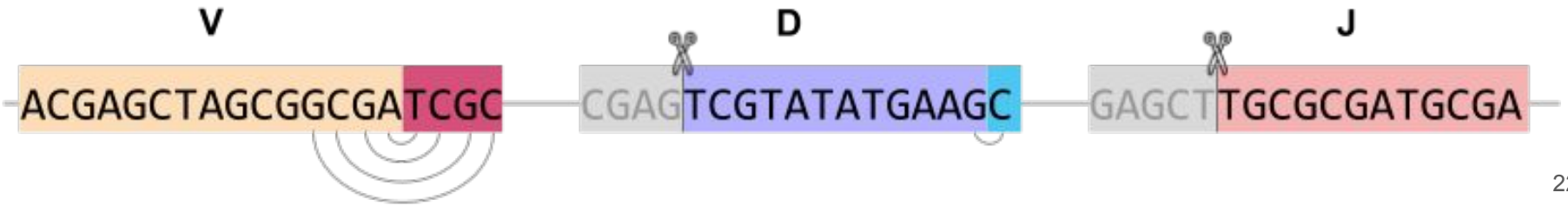


Why are antibodies so versatile if there are only 55×23×6 VDJ recombinations?

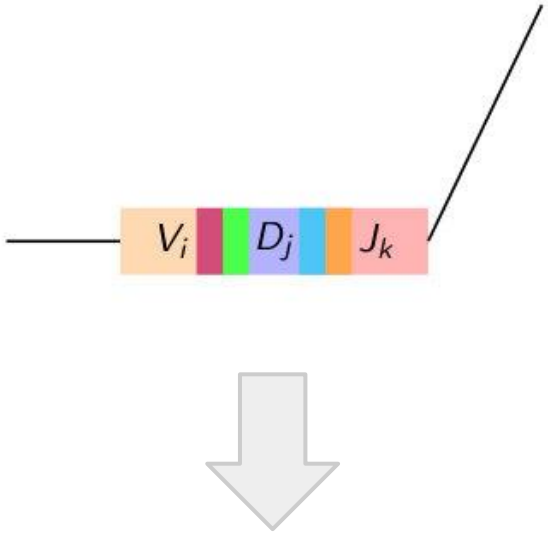


Recombination process is imperfect and includes many random processes:

- Palindromic insertions
- **Segment cleavage**

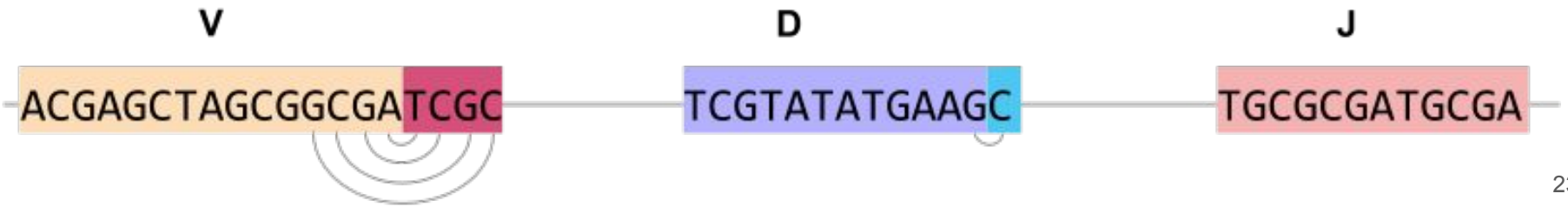


Why are antibodies so versatile if there are only 55×23×6 VDJ recombinations?

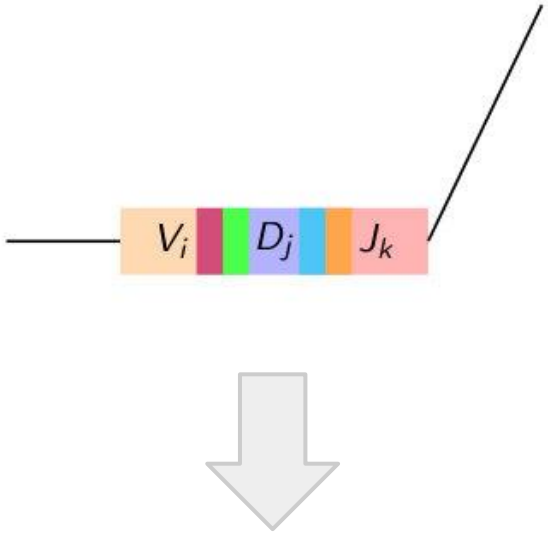


Recombination process is imperfect and includes many random processes:

- Palindromic insertions
- **Segment cleavage**

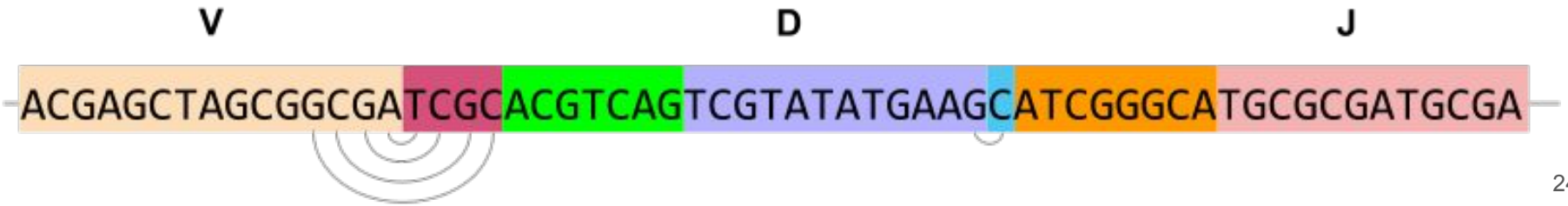


Why are antibodies so versatile if there are only 55×23×6 VDJ recombinations?



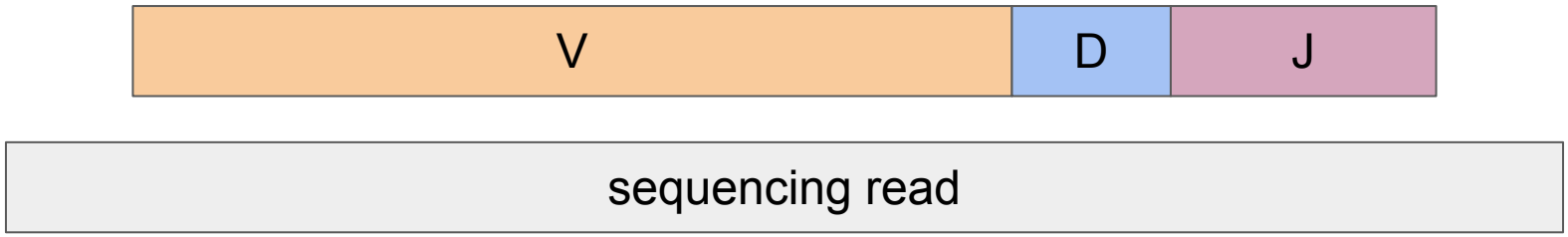
Recombination process is imperfect and includes many random processes:

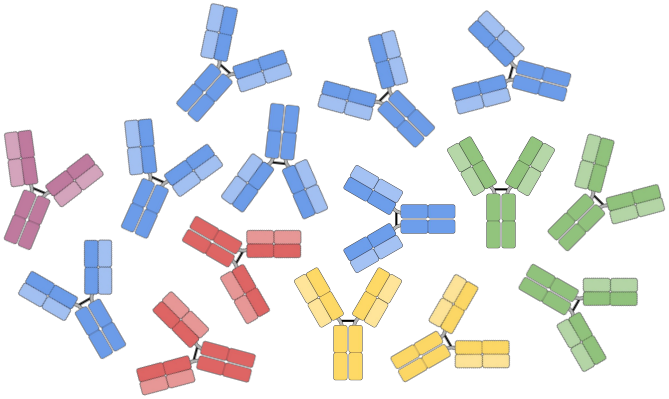
- Palindromic insertions
- Segment cleavage
- **Non-genomic insertions**



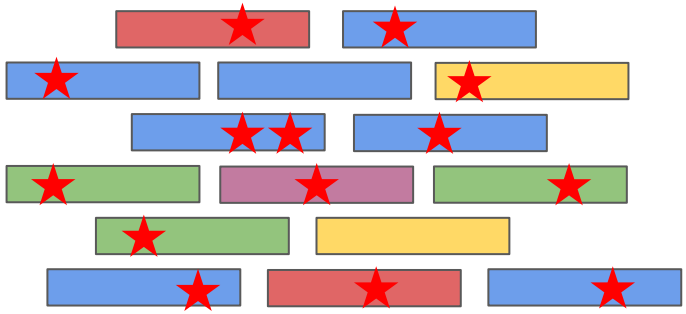
Antibody repertoire sequencing (Rep-seq)

Length: ~360 nt



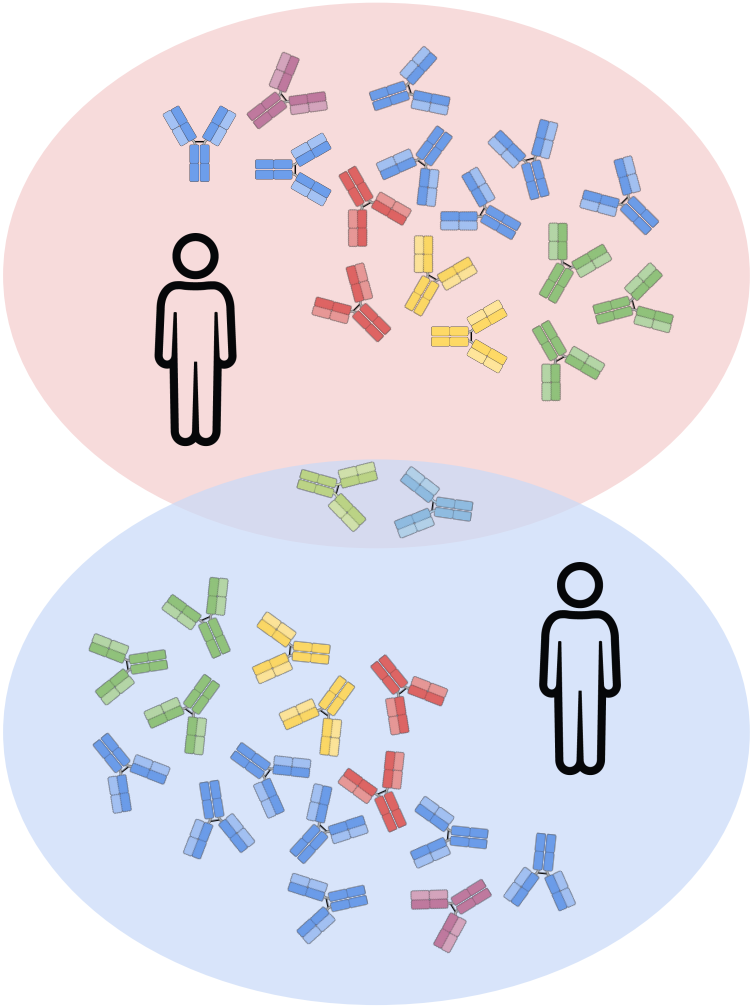


VDJ from DNA or RNA

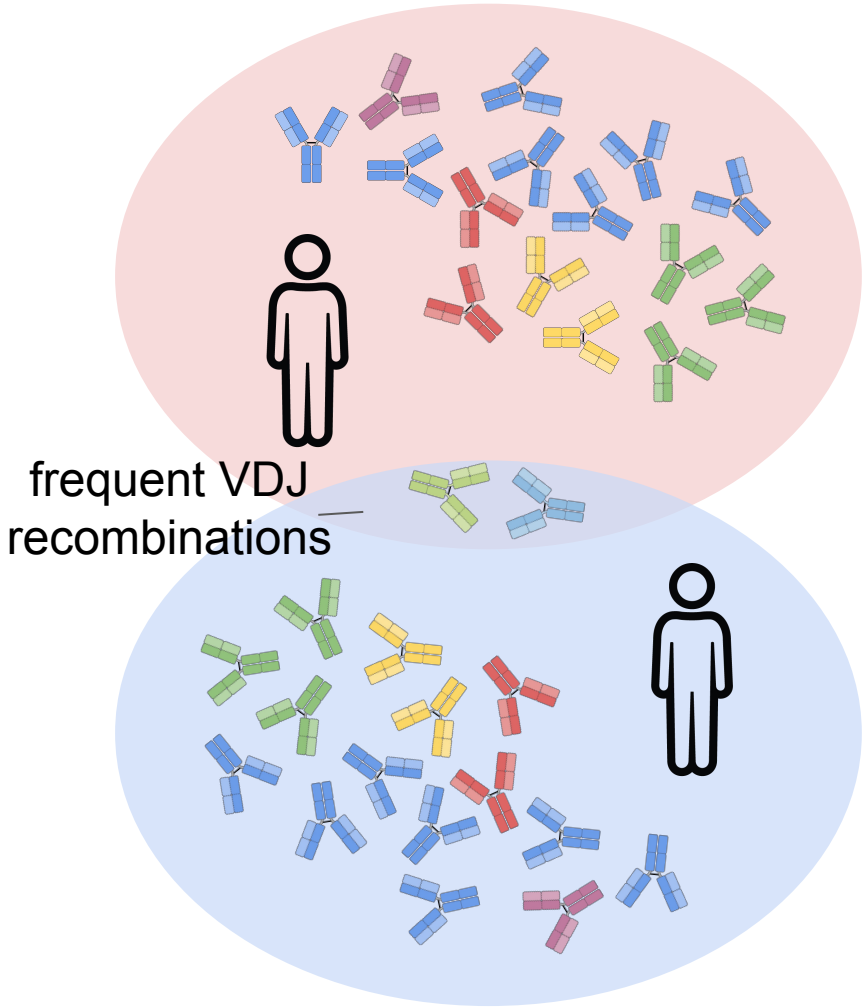


Error-prone immunosequencing reads

Antibody repertoire is unique for an individual

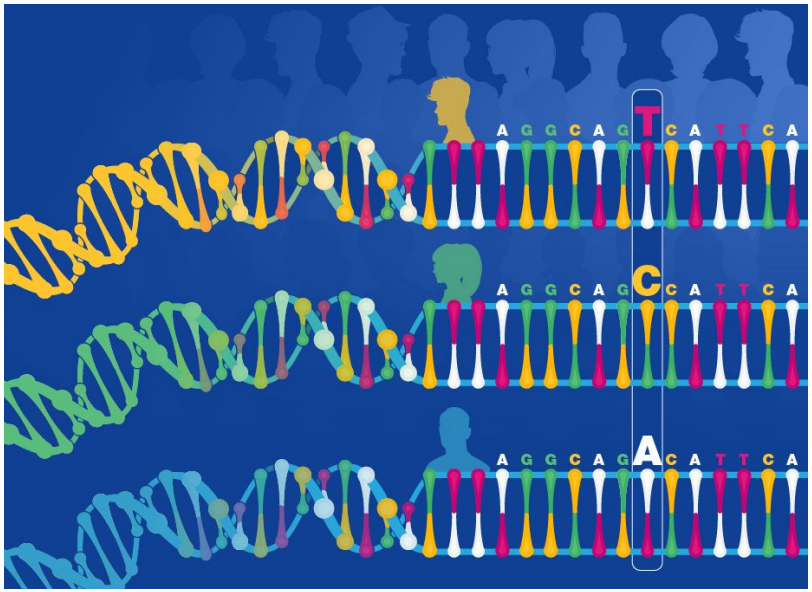


Antibody repertoire is unique for an individual

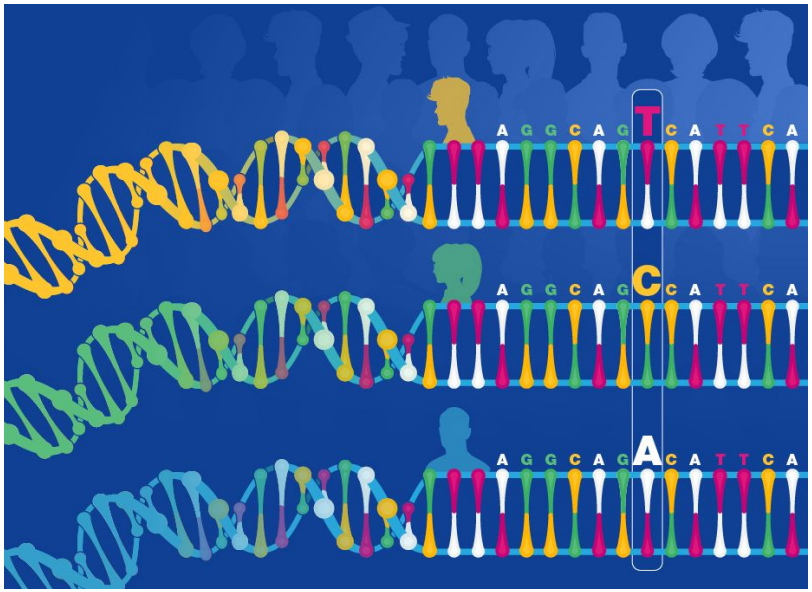


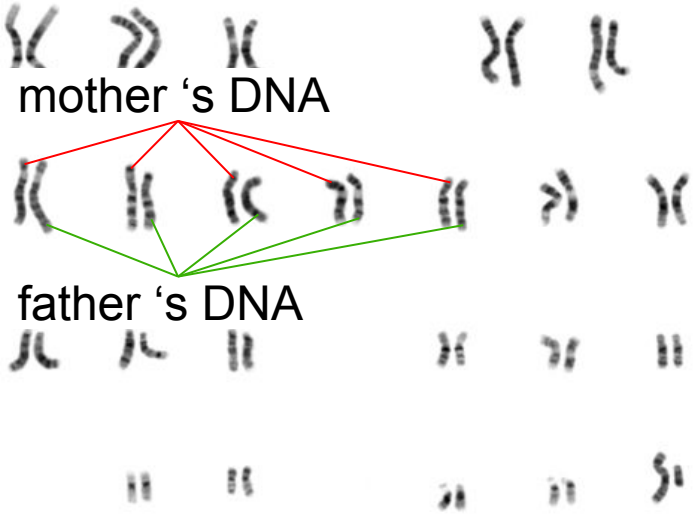
- VDJ sequences are extremely diverse
- If a VDJ sequence is shared between two individuals, it is a likely a frequent recombination rather a functionally important sequence
- **We cannot study antibody responses just comparing VDJ sequences**

Genomic variations

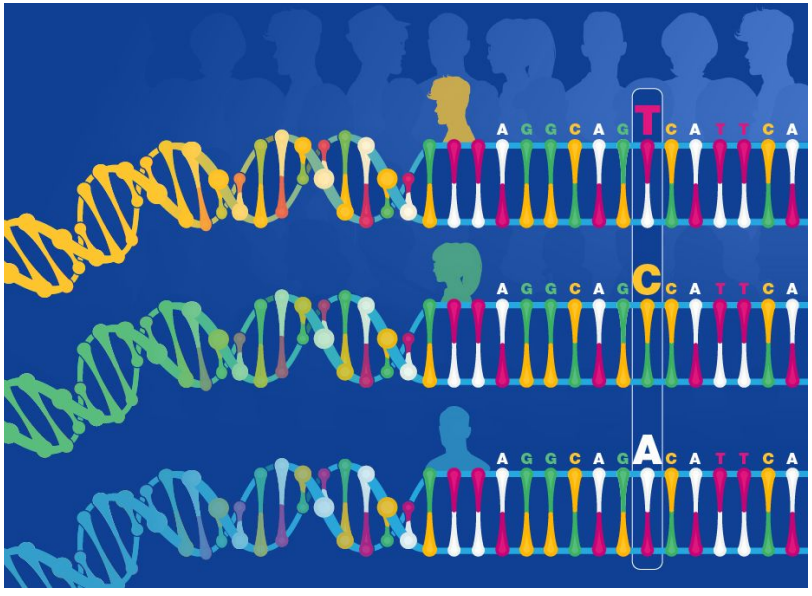


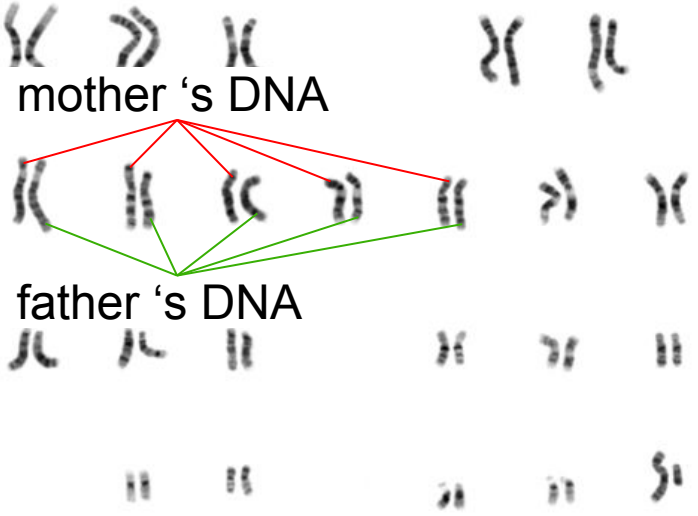
Genomic variations





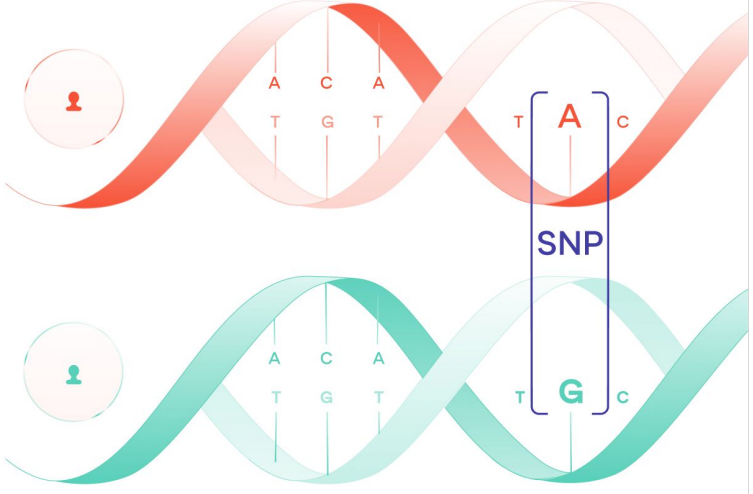
Genomic variations



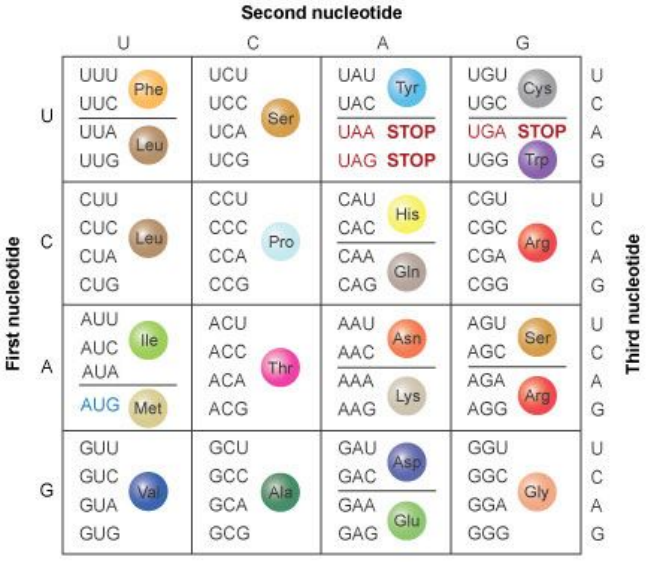


Single Nucleotide Polymorphism (SNP) is associated with a position in the genome:

- A – both chromosomes have A
- A / G – one chromosome has A, another one has G
- G – both chromosomes have G

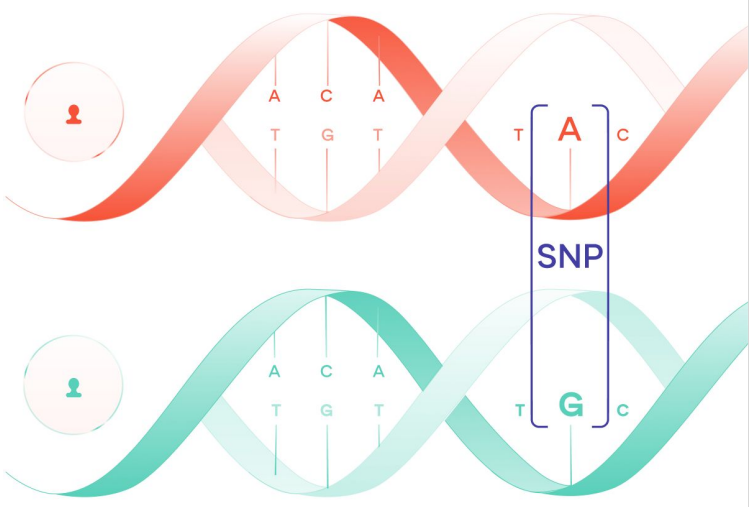


Altered genes produce altered proteins

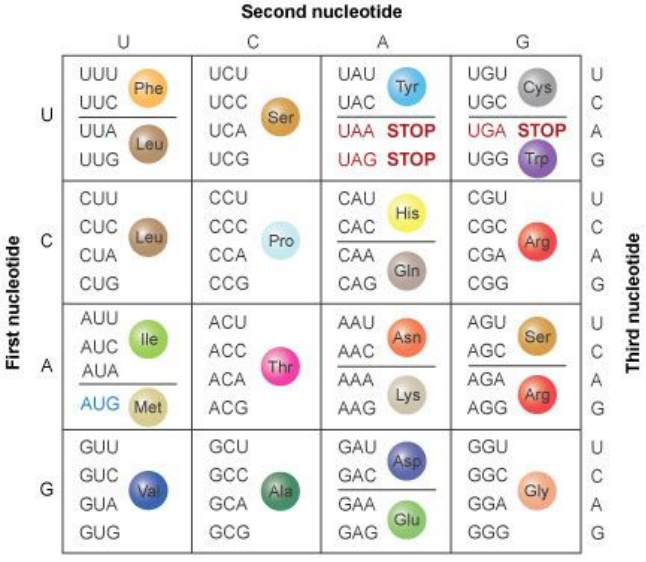


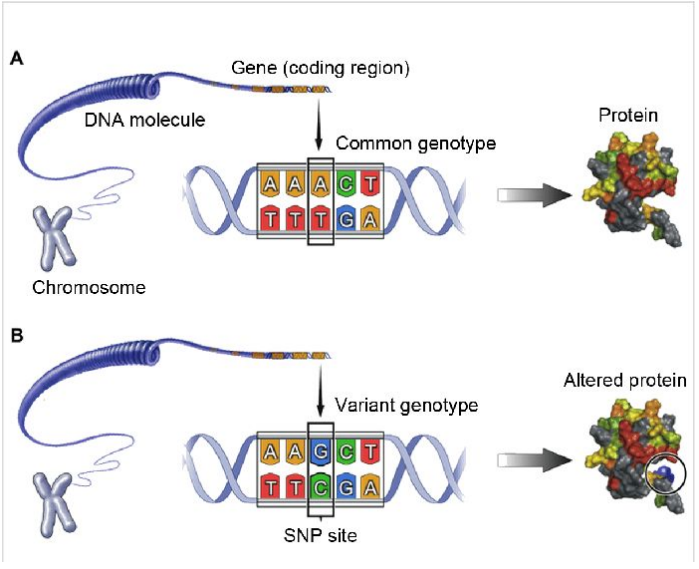
Single Nucleotide Polymorphism (SNP) is associated with a position in the genome:

- A – both chromosomes have A
- A / G – one chromosome has A, another one has G
- G – both chromosomes have G



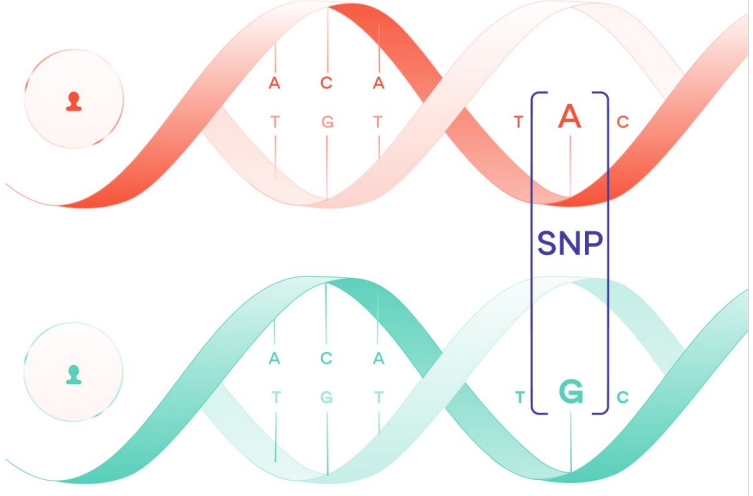
Altered genes produce altered proteins



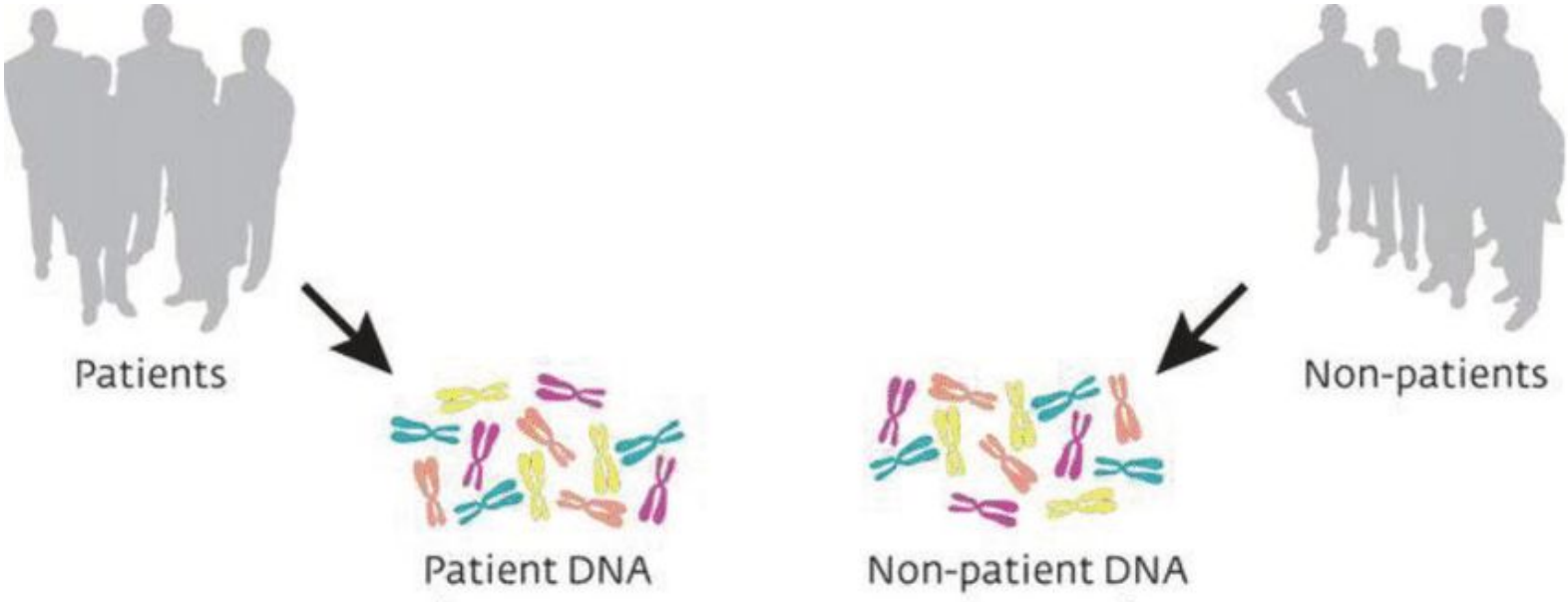


Single Nucleotide Polymorphism (SNP) is associated with a position in the genome:

- A – both chromosomes have A
- A / G – one chromosome has A, another one has G
- G – both chromosomes have G



Genome-wide association studies (GWAS)

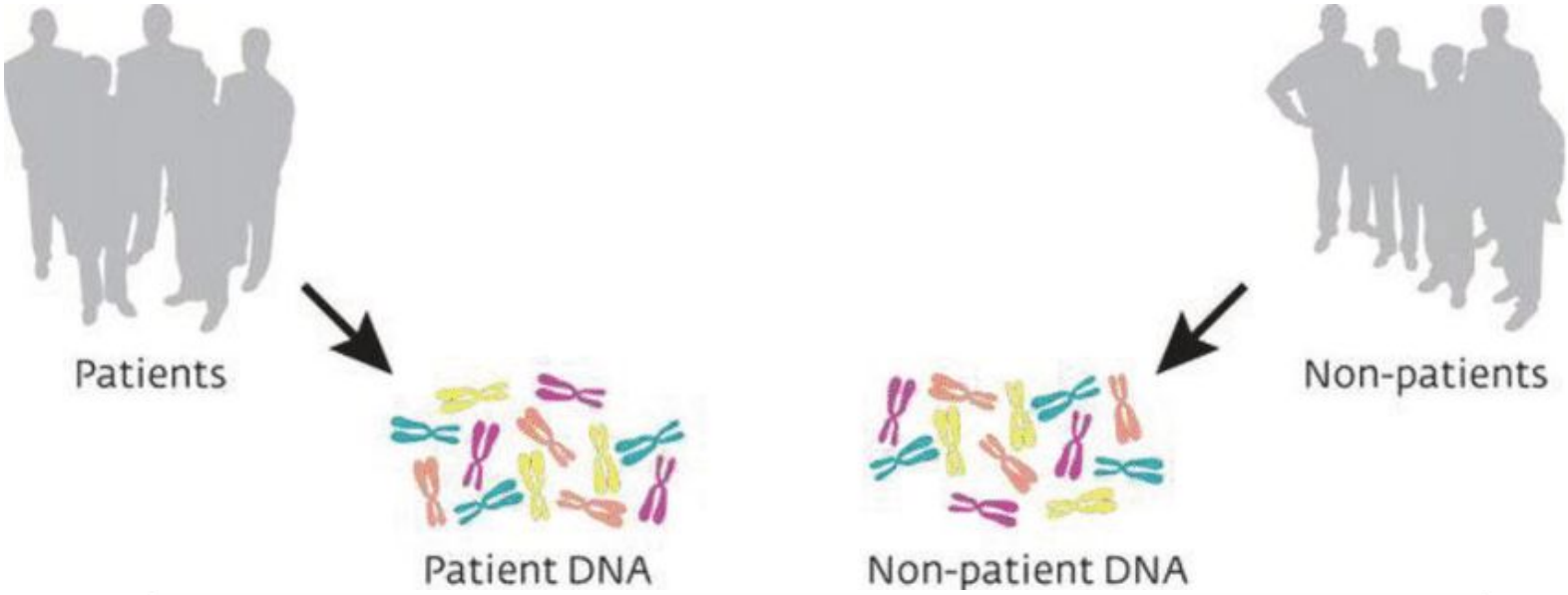


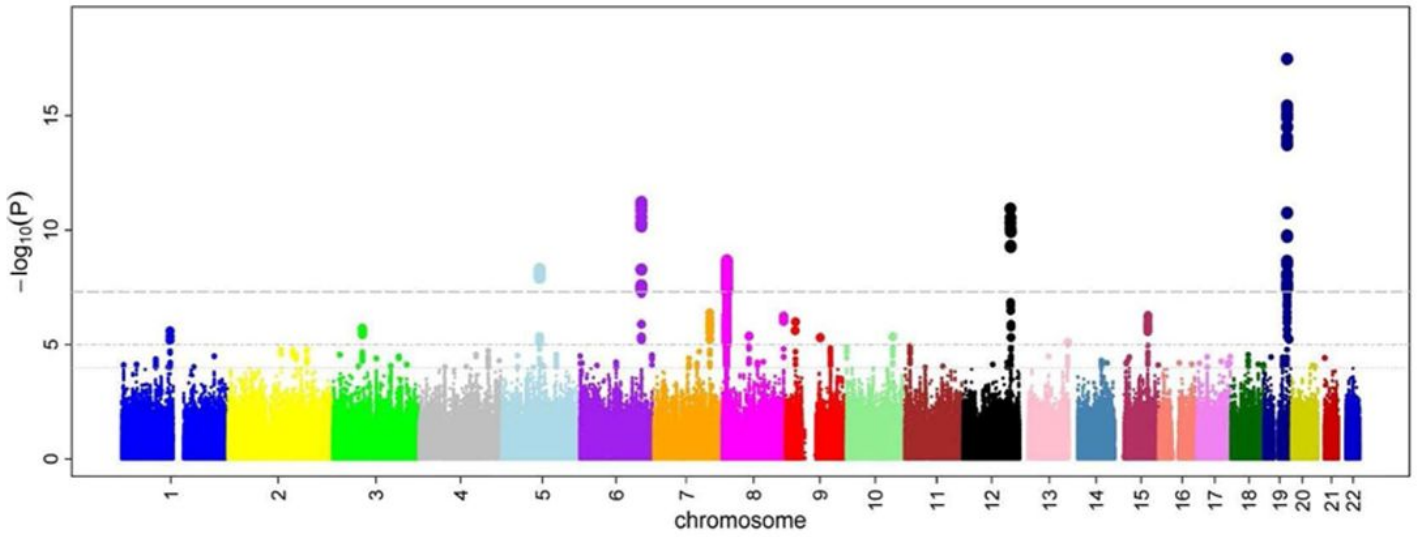
	A	A/G	G
Patients	15	10	8
Controls	2	6	25

a single SNP

P-value (Fisher exact probability test) = 0.000026

Genome-wide association studies (GWAS)





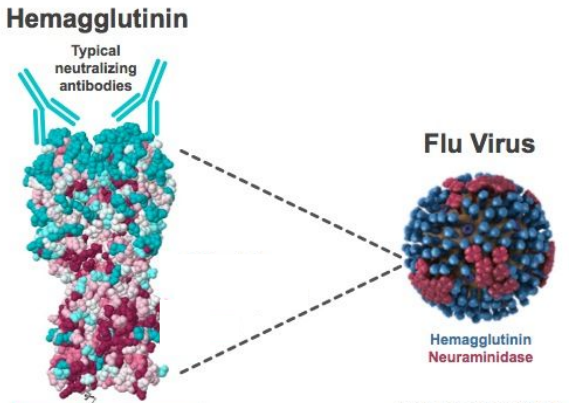
Variants of IGHV1-69 shape Ab response to flu

V	D	J

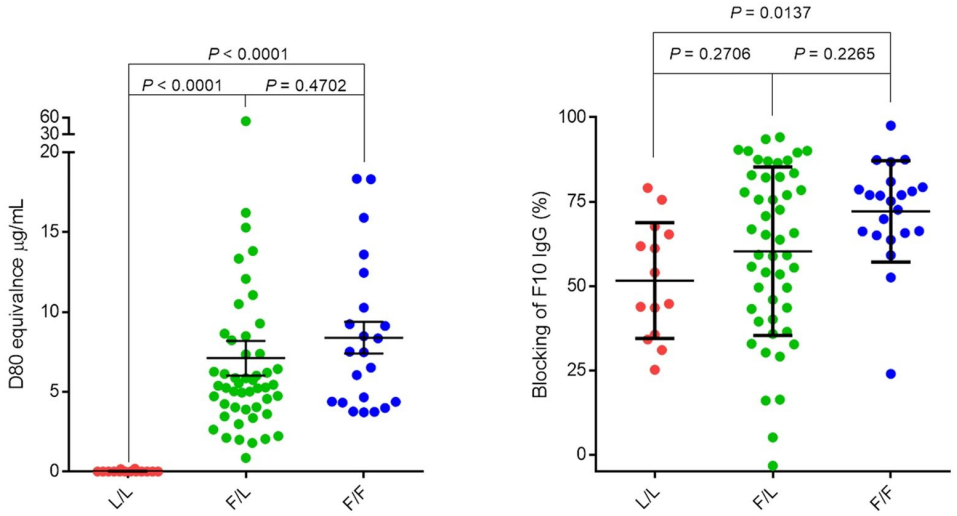
IGHV1-69

IGHV1-69*01	54: F
IGHV1-69*02	54: L
IGHV1-69*03	54: F
IGHV1-69*04	54: L
...	
IGHV1-69*14	54: F

Successful binding to hemagglutinin
Loss of binding properties

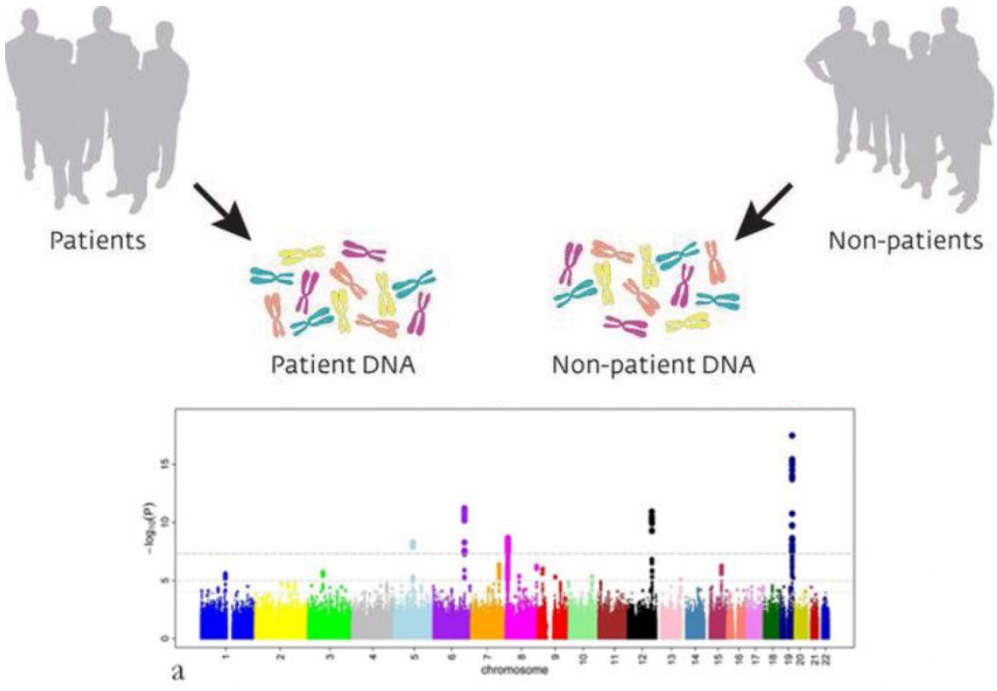


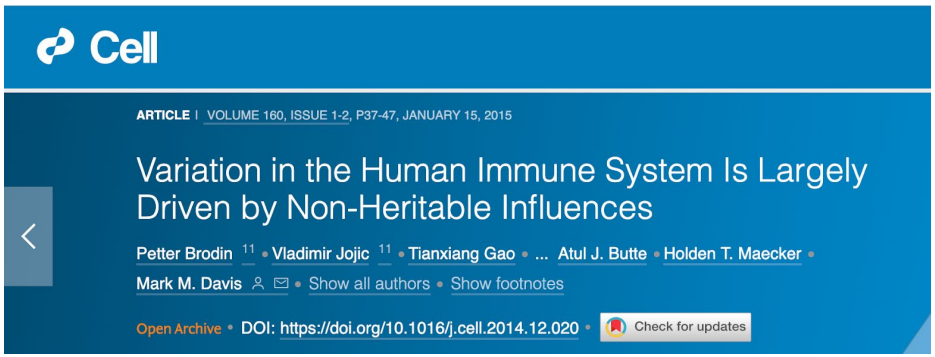
Titers (= antibody counts) before and after immunization



Avnir et al., *Sci Rep*, 2016

GWAS of antibody responses

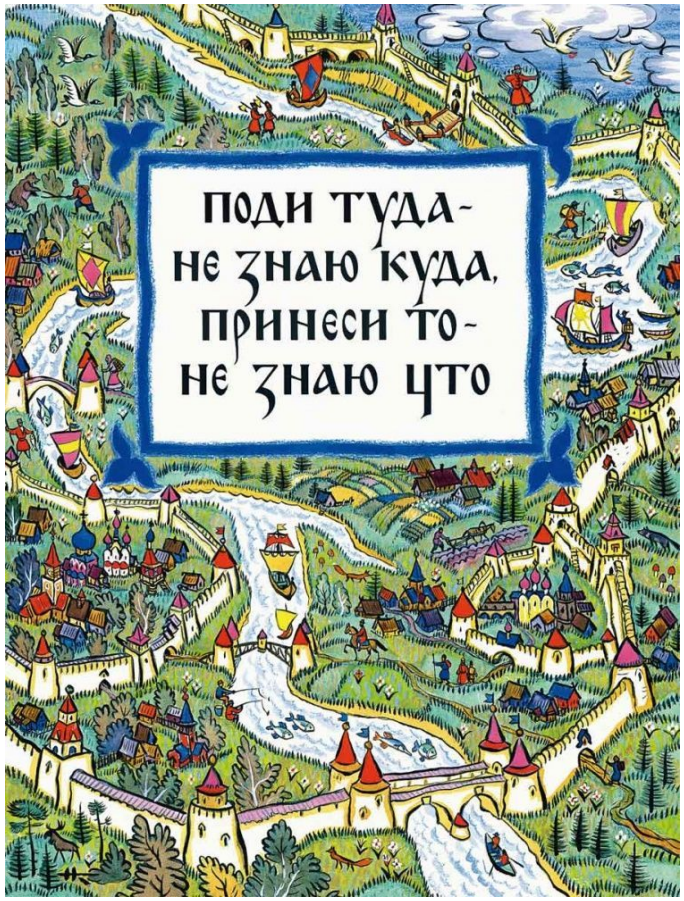




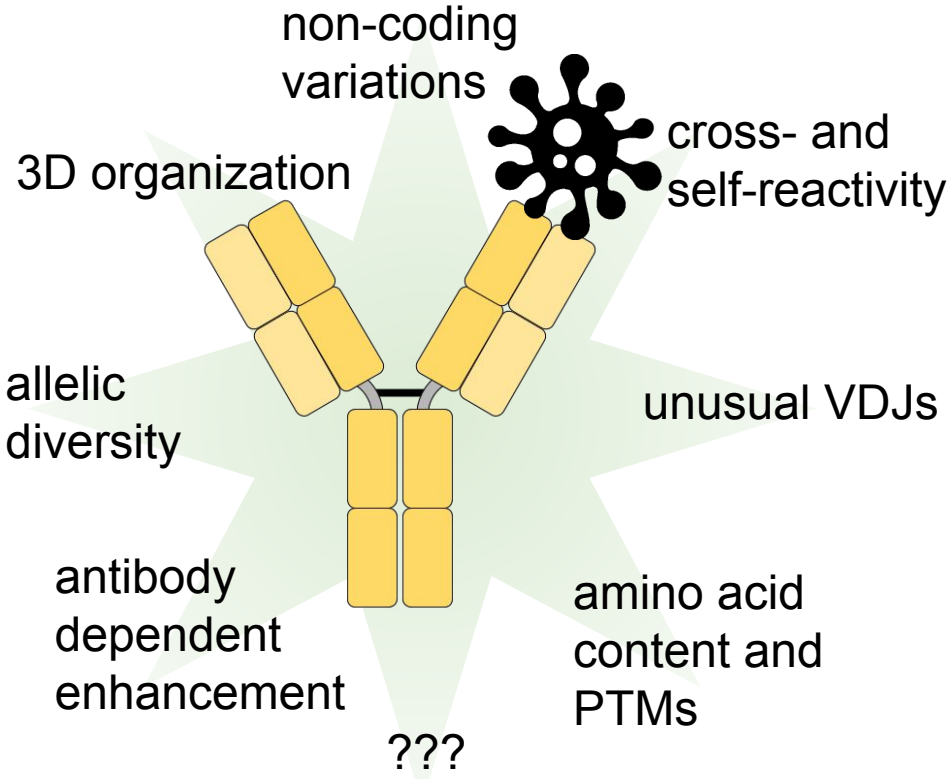
Challenges

- A single IG locus includes many V, D, and J genes. **If a single gene loses functionality, in most cases others can replace it**
- Recent studies report a lack of associations between genomic variations outside IG loci and adaptive immune responses
- Many other factors (age, diet, environment) influence adaptive immune responses
- **Features of antibody repertoires go far beyond genomic variations**

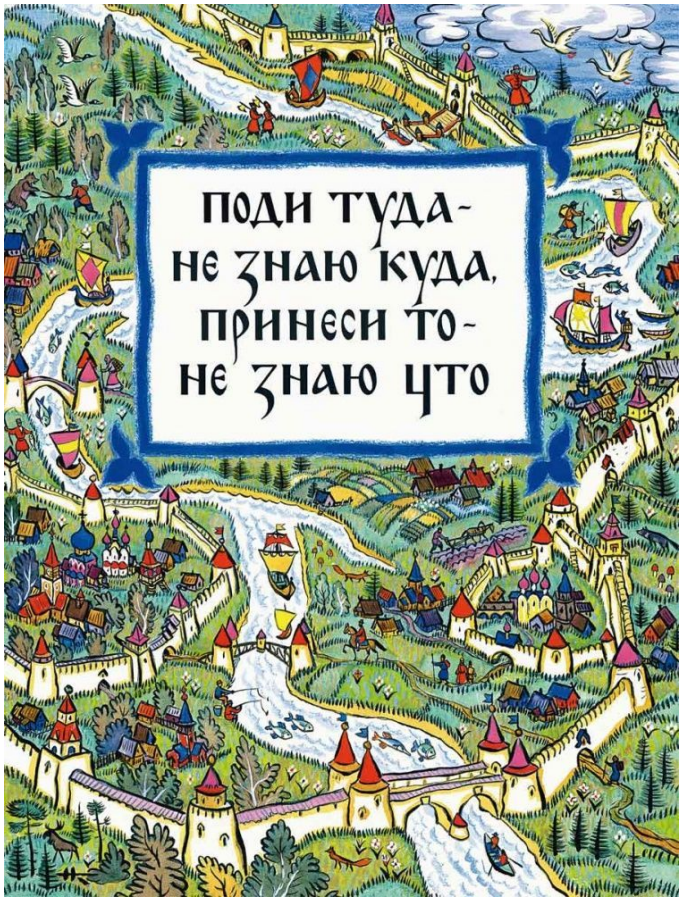
Bioinformatics + immunology = immunoinformatics



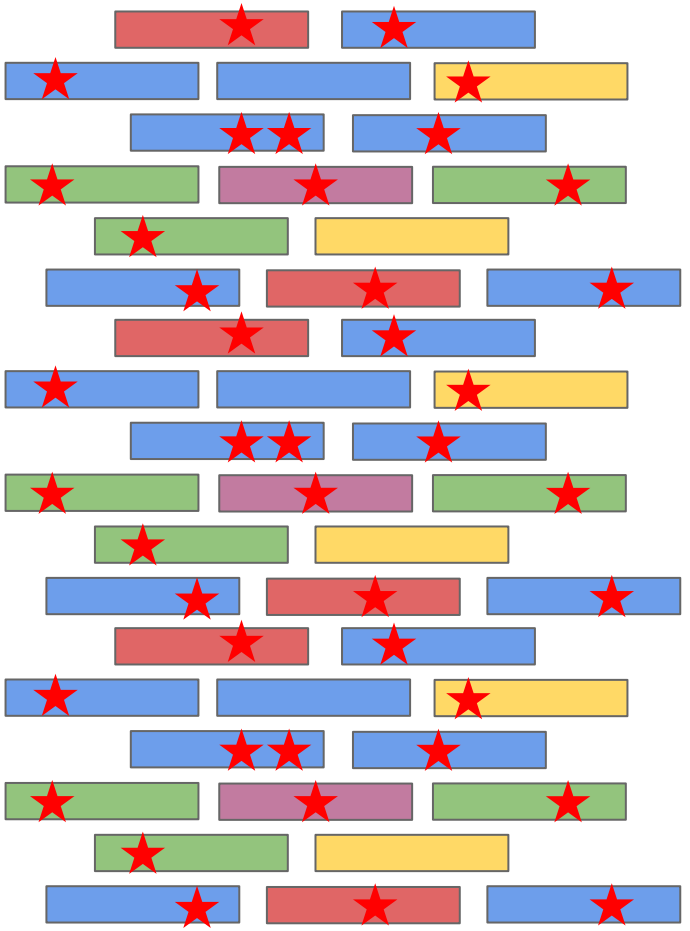
Go I Know Not Whither and
Fetch I Know Not What



Bioinformatics + immunology = immunoinformatics

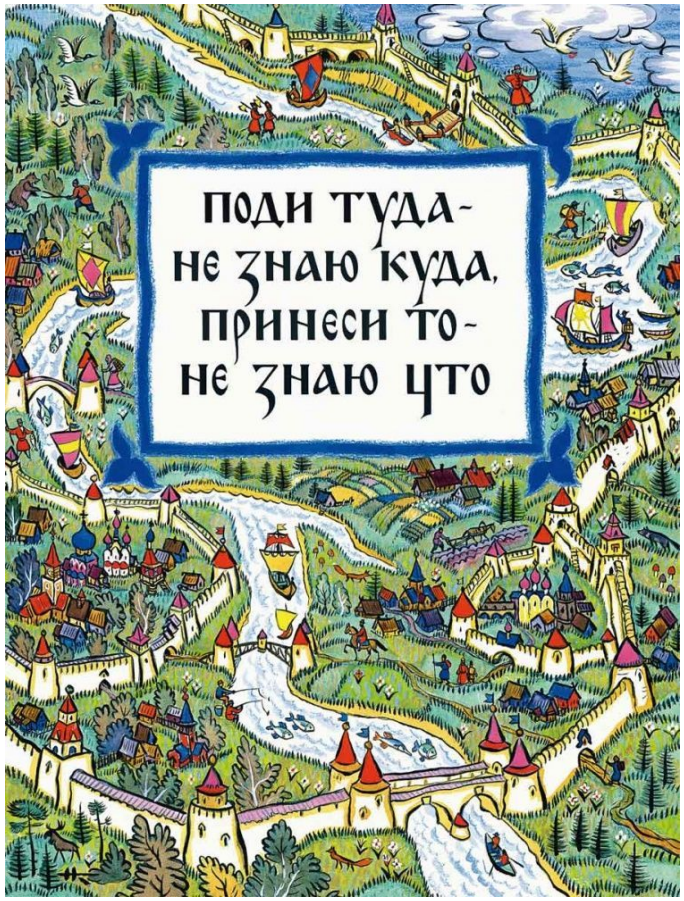


Go I Know Not Whither and
Fetch I Know Not What

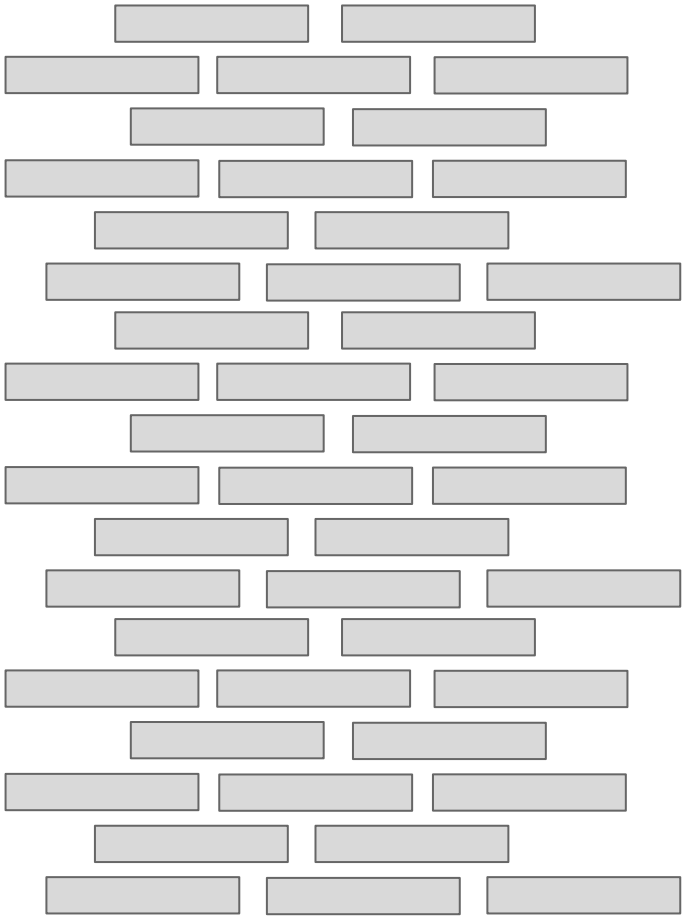


Error-prone immunosequencing reads

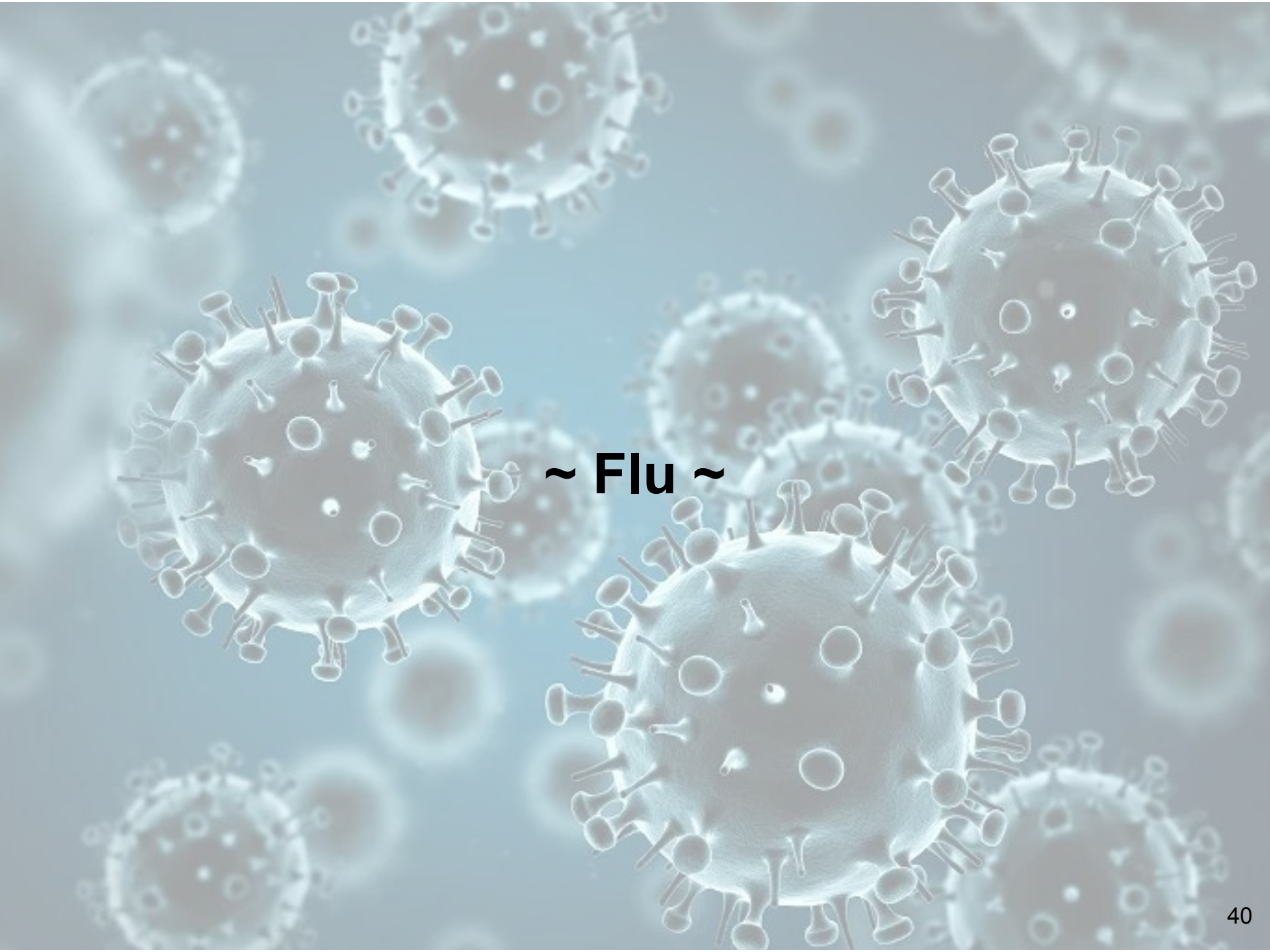
Bioinformatics + immunology = immunoinformatics



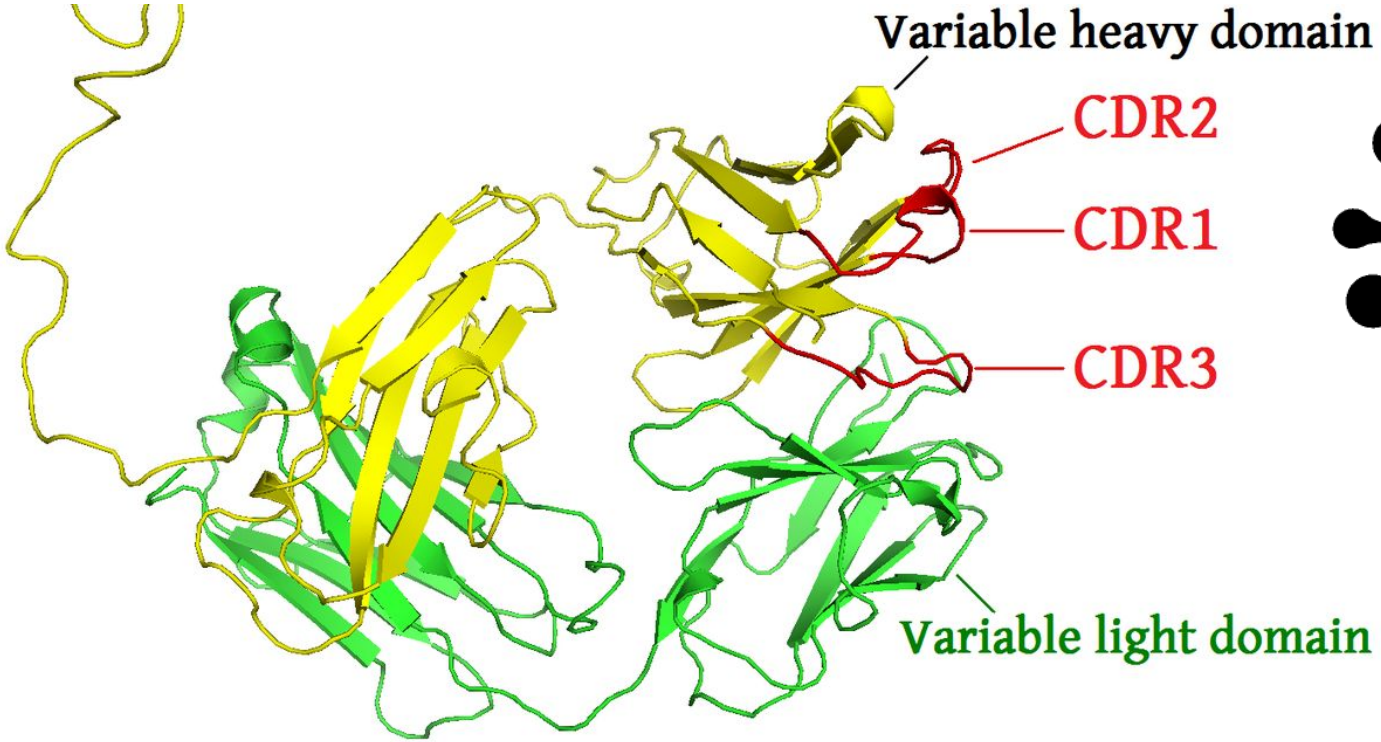
Go I Know Not Whither and
Fetch I Know Not What

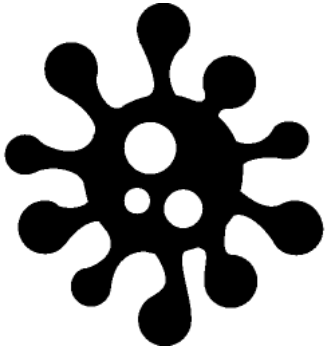


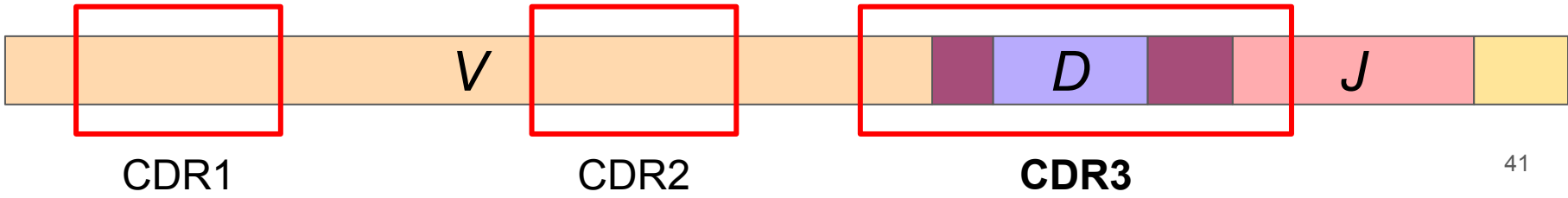
Error-prone immunosequencing reads


~ Flu ~

CDRs represent antigen-binding sites

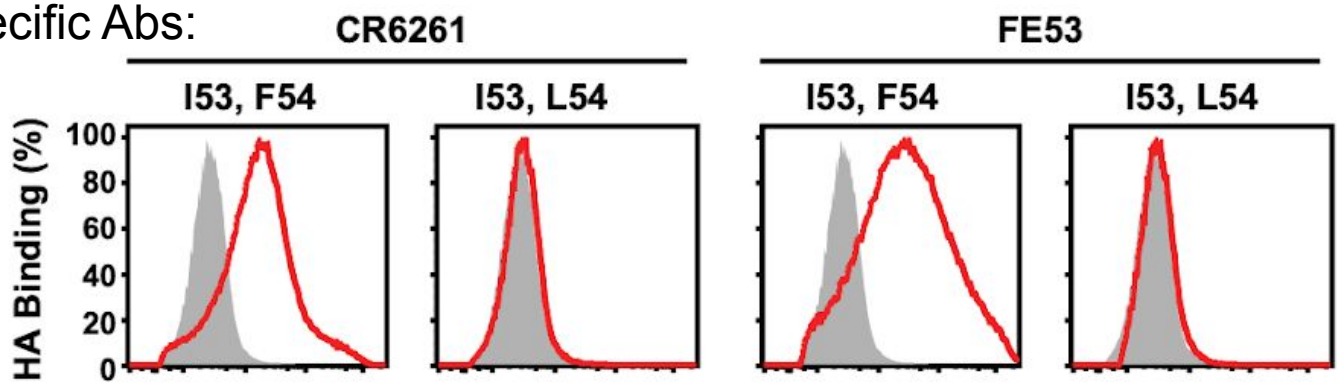






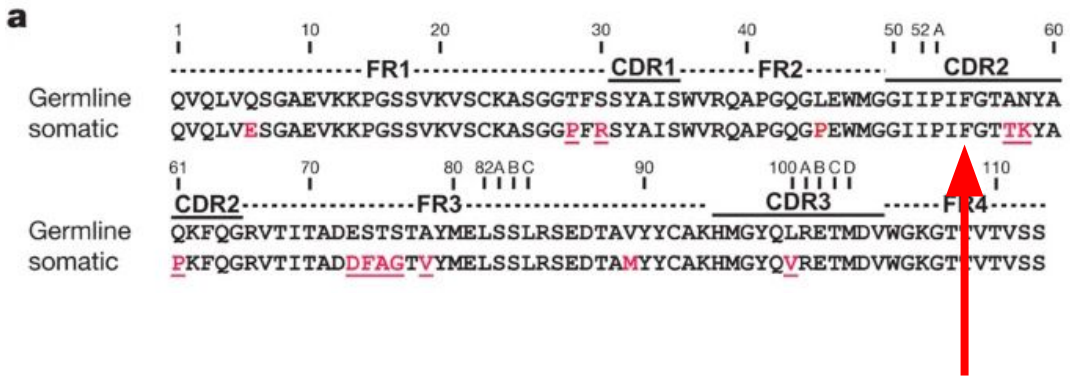
Anatomy of IGHV1-69-guided response to flu

Two flu-specific Abs:



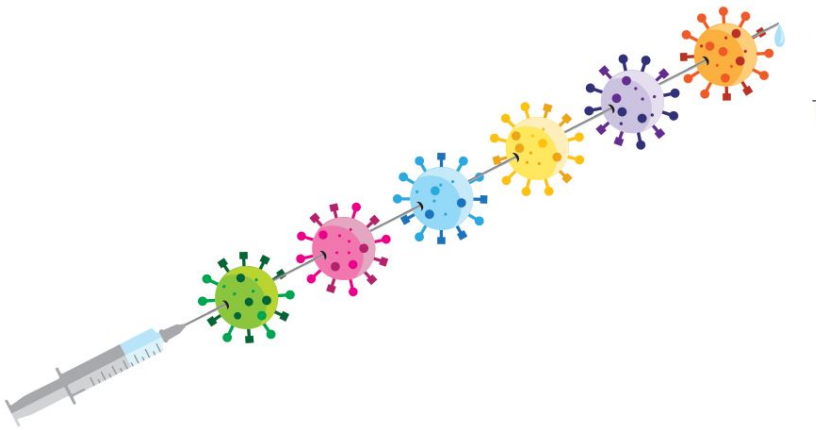
Lindwood et al., Nature, 2012

IGHV1-69*01	54: F
IGHV1-69*02	54: L
IGHV1-69*03	54: F
IGHV1-69*04	54: L
...	
IGHV1-69*14	54: F



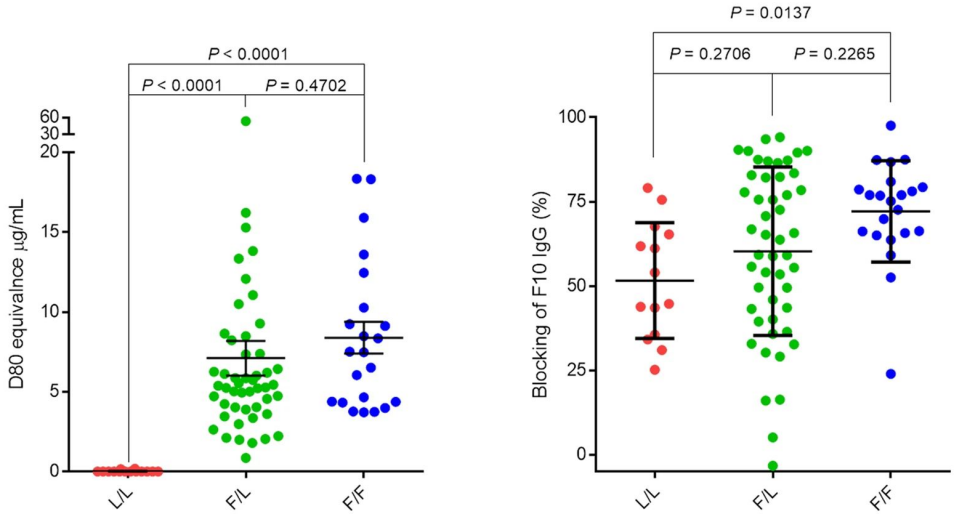
Antibody titers vs IGHV1-69 genotype

The titer data suggests that genotype of IGHV1-69 shapes the response to flu and other V genes do not fully replace the “bad” variant of IGHV1-69



~20% of people have L/L variants of IGHV1-69

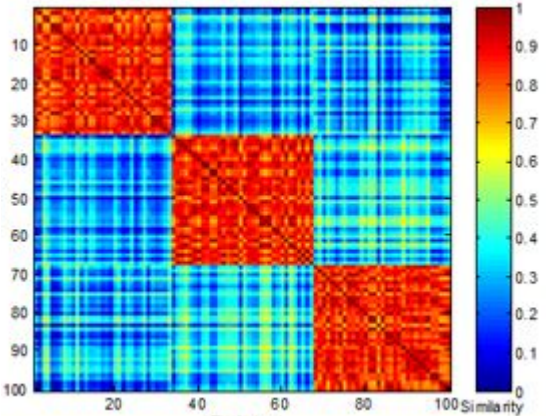
Titers (= antibody counts) before and after immunization

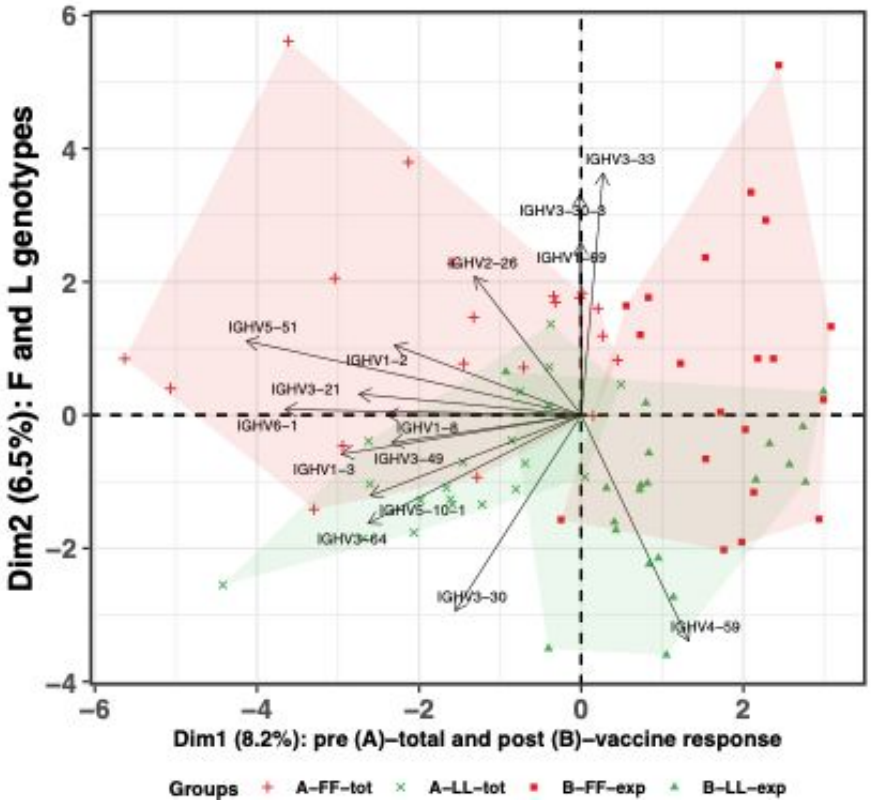


Avnir et al., *Sci Rep*, 2016

Bioinformatics analysis of flu response

- Usage of gene *G* = the fraction of VDJ sequences derived from *G*
- Individual antibody repertoire can be described as a *usage vector* for all existing genes
- We can compare usage vectors for F/F, F/L, and L/L individuals:





Ke, Nouri, Safonova, et al., in preparation

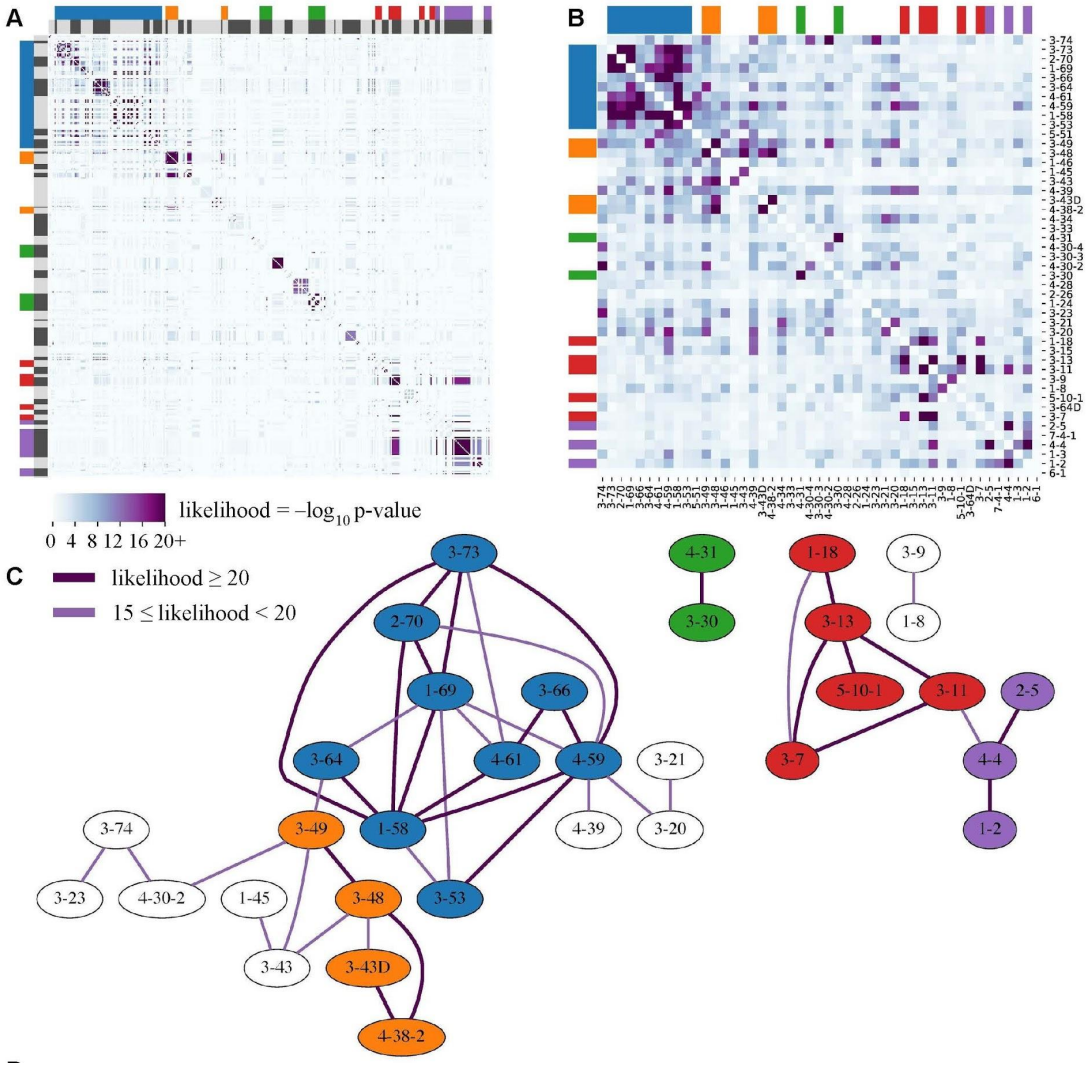
Linked variations of V genes

	A	A/G	G
C	15	10	8
A/C	6	8	3
A	2	5	17

P-value = 0.00967

Blue group is a set of V genes associated with flu response

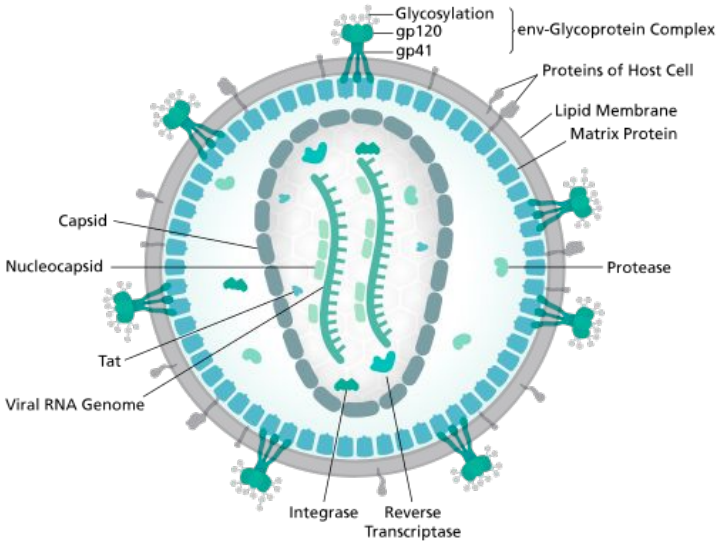
These genes have linked variations and perhaps close specificities

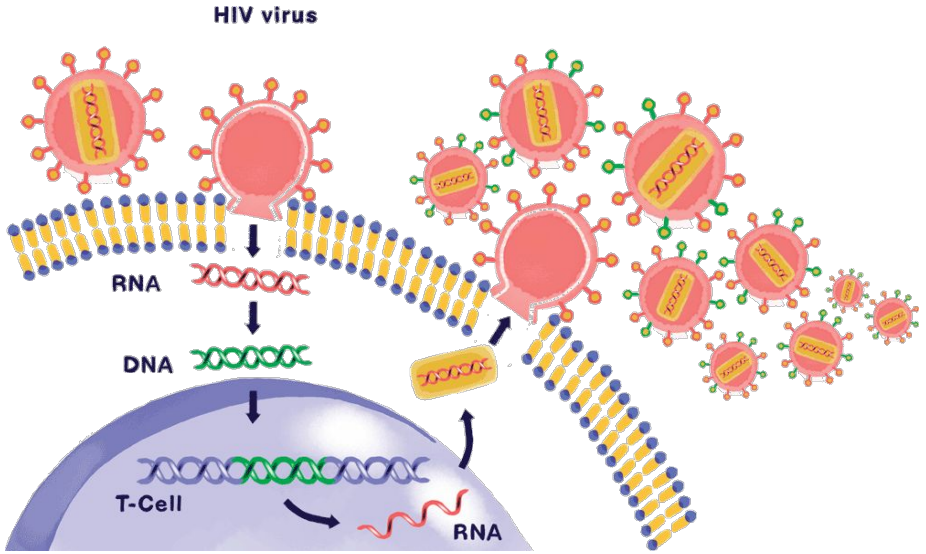


Safonova and Watson, in preparation


~ HIV ~

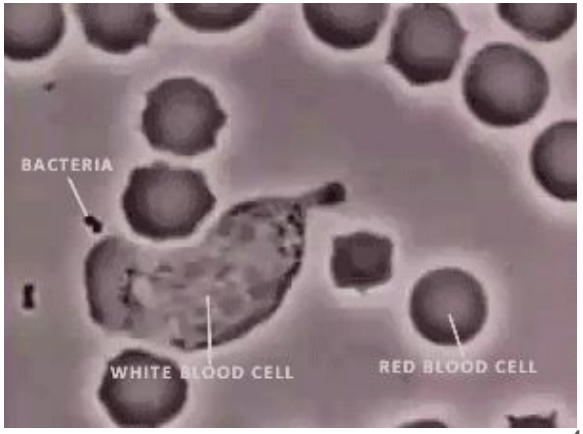
HIV and antibody response



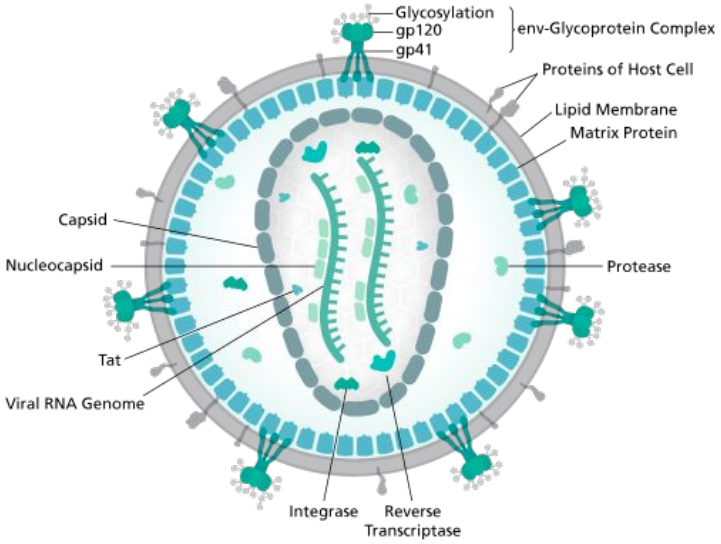


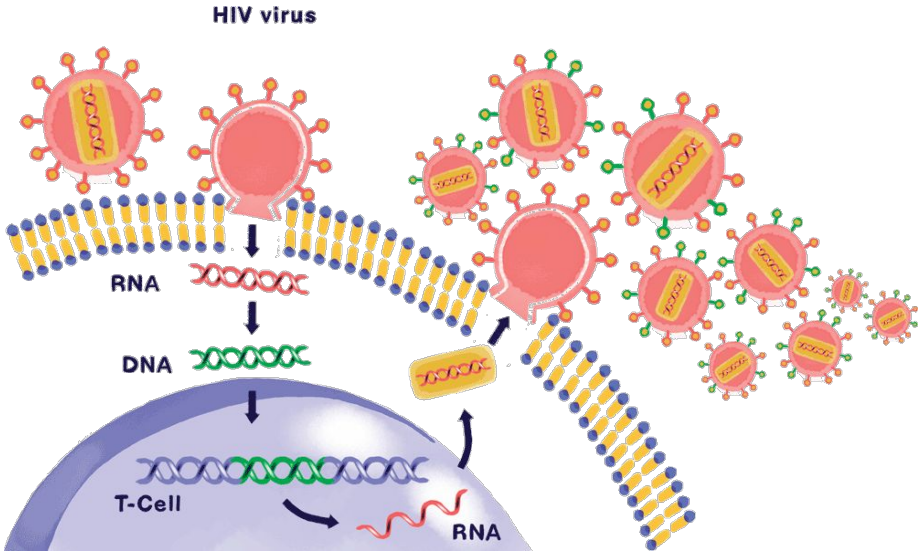
HIV infects immune cells:

- T cells
- Macrophages
- Dendritic cells



HIV and antibody response



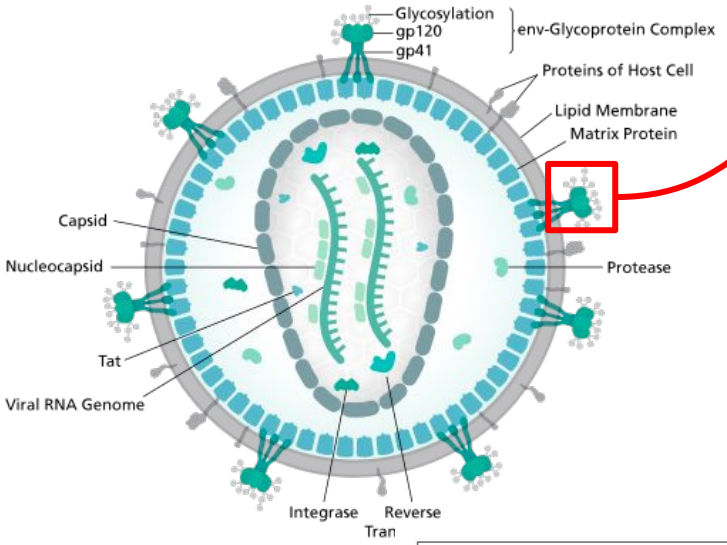


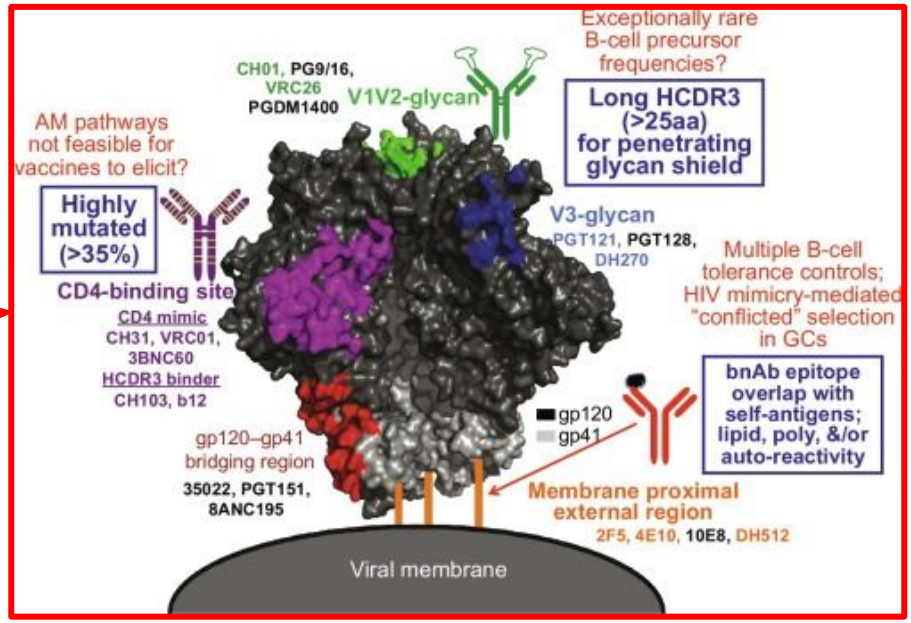
HIV infects immune cells:

- T cells
- Macrophages
- Dendritic cells

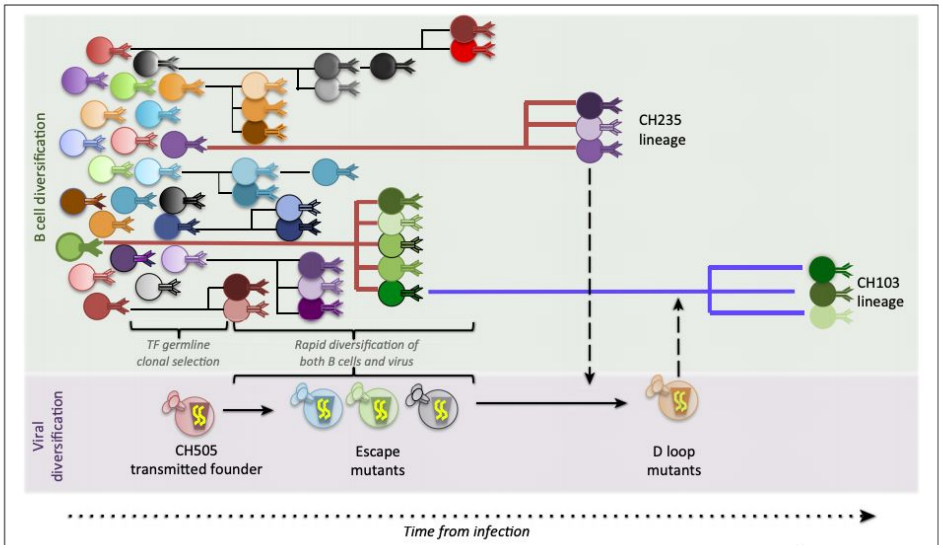
B cells producing antibodies are not affected by HIV but cannot fight it because of a high mutation rate of the HIV genome

Recognition of HIV



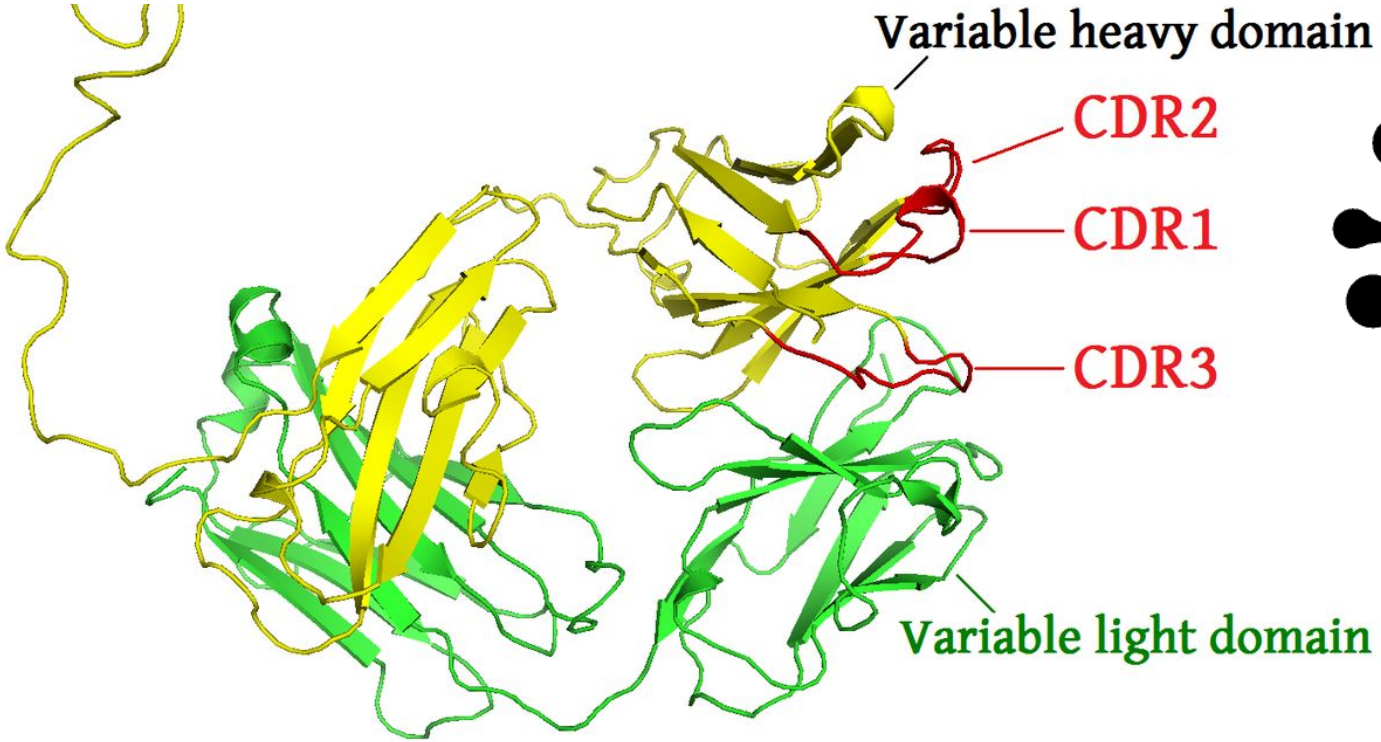


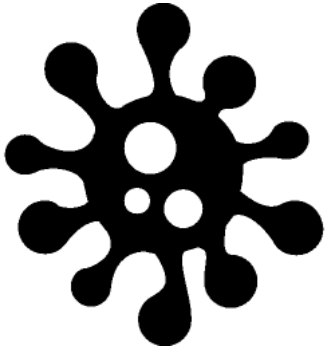
Verkoczy, Adv Immunol, 2017

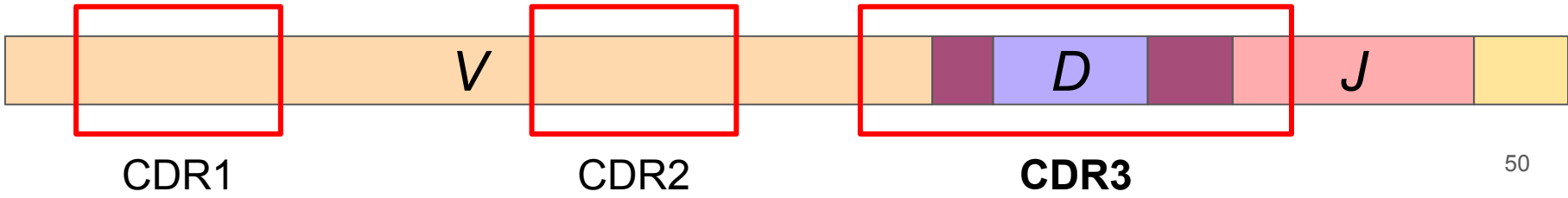


Alter & Ackerman, *Cell*, 2014

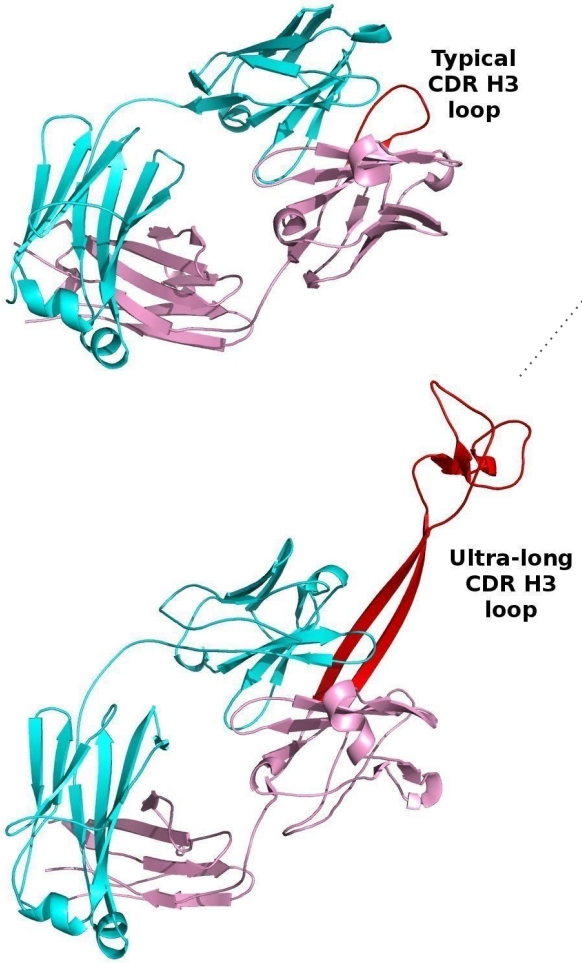
CDRs represent antigen-binding sites







Long CDR3s in response to HIV



Broadly neutralizing antibodies against HIV are characterized by extremely long CDR3s

Ultralong CDR3s in human antibodies present a result of **VDDJ** recombination:

```
D3:                                          GTATTACGATTTTTGGAGTGGTTATtatacc
D5:                                                                  GTGGATACAGCTATGGttac
CDR3:   ACCACAGAACCGCTTCAGTTTAGTCCGTATTACGATTTTTGGAGTGGTTATCAGCCAGTGGATACAGCTATGGACCCGTTGACT
span:                                        GTATTACGATTTTTGGAGTGGTTATCAGCCAGTGGATACAGCTATGG
inter-D insertion:                                                   CAGCCA
```

Sok et al., *Nature*, 2017

Safonova and Pevzner, *Front Immunol*, 2019

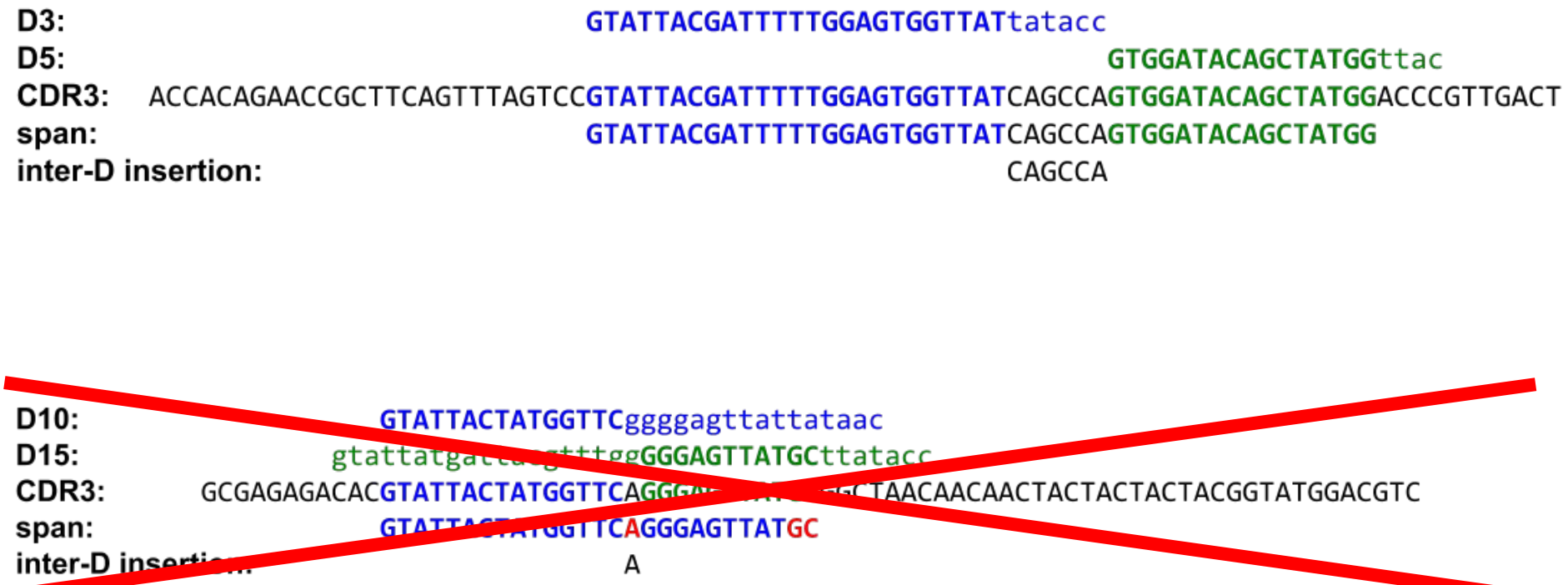
Finding VDDJ recombinations

D3: GTATTACGATTTTTGGAGTGGTTATtatacc
D5: GTGGATACAGCTATGGttac
CDR3: ACCACAGAACCGCTTCAGTTTAGTCCGTATTACGATTTTTGGAGTGGTTATCAGCCAGTGGATACAGCTATGGACCCGTTGACT
span: GTATTACGATTTTTGGAGTGGTTATCAGCCAGTGGATACAGCTATGG
inter-D insertion: CAGCCA

D10: GTATTACTATGGTTCggggagttattataac
D15: gtattatgattacgtttggGGGAGTTATGCttatacc
CDR3: GCGAGAGACACGTATTACTATGGTTCAGGGAGTTATGCGGCTAACAACAACTACTACTACTACGGTATGGACGTC
span: GTATTACTATGGTTCAGGGAGTTATGC
inter-D insertion: A

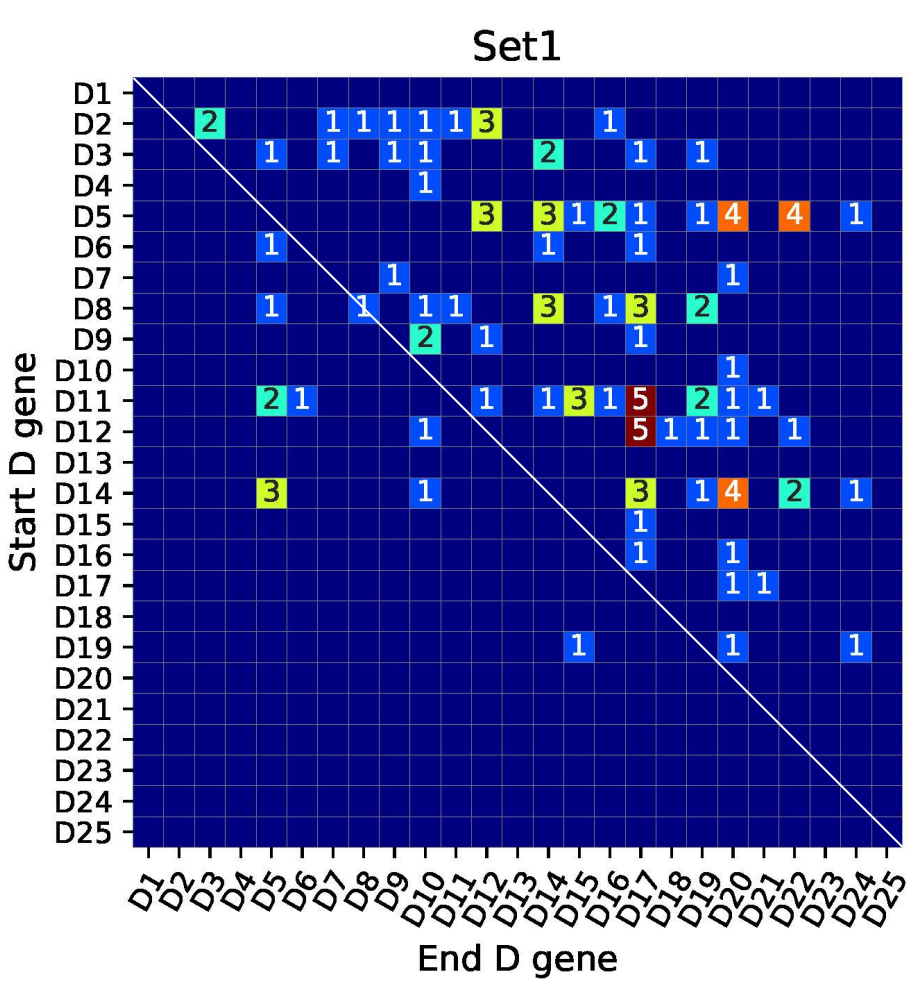
Safonova and Pevzner, *Front Immunol*, 2019

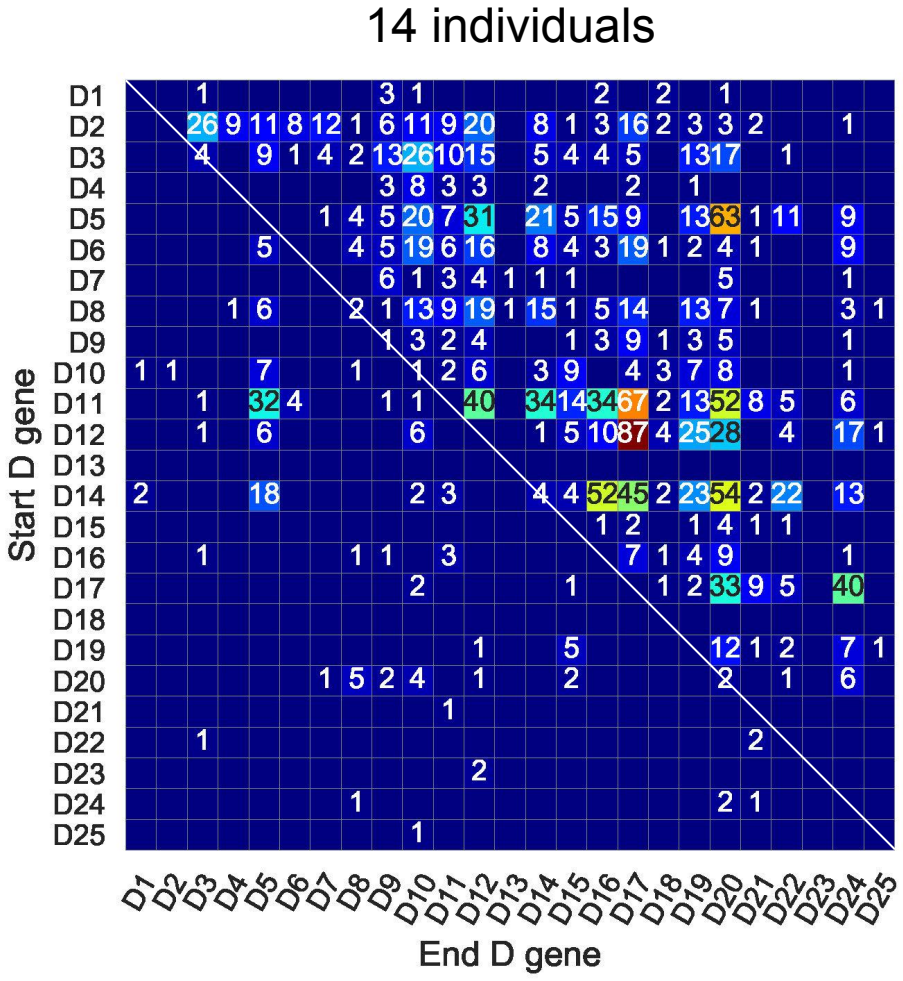
Finding VDDJ recombinations



Safonova and Pevzner, *Front Immunol*, 2019

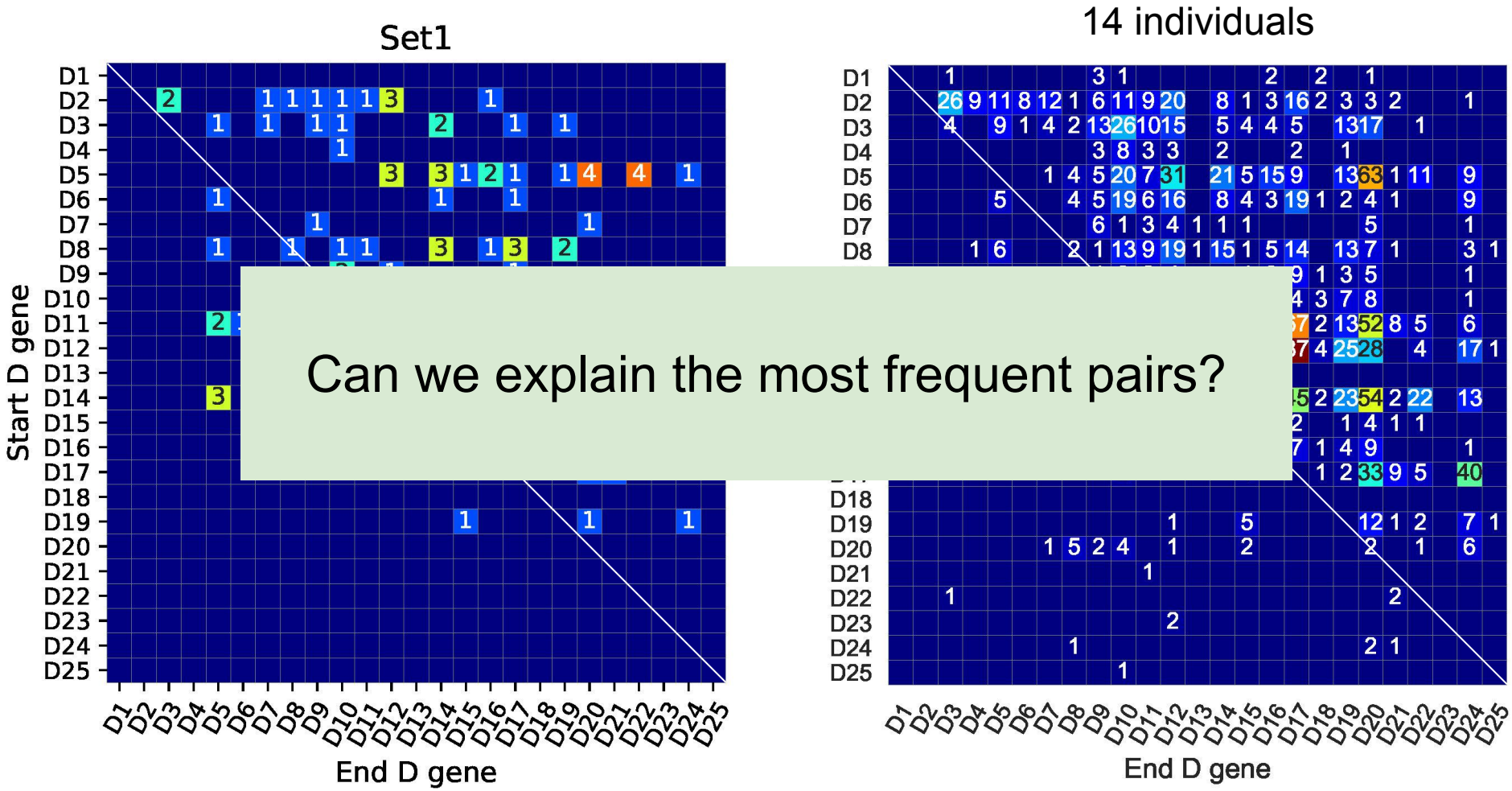
Most D-D pairs follow ordering in IGH locus





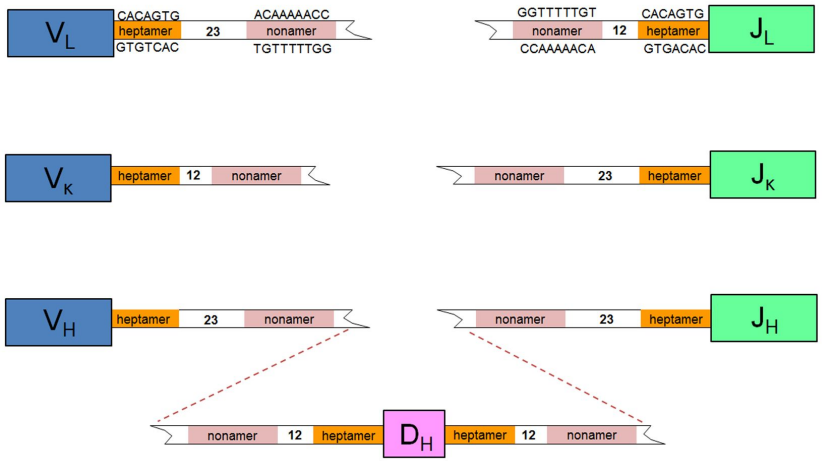
Safonova and Pevzner, *Front Immunol*, 2019

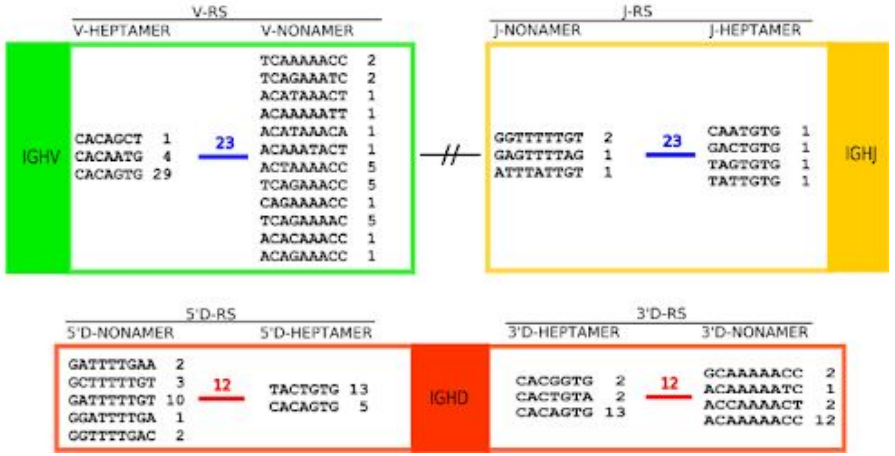
Most D-D pairs follow ordering in IGH locus

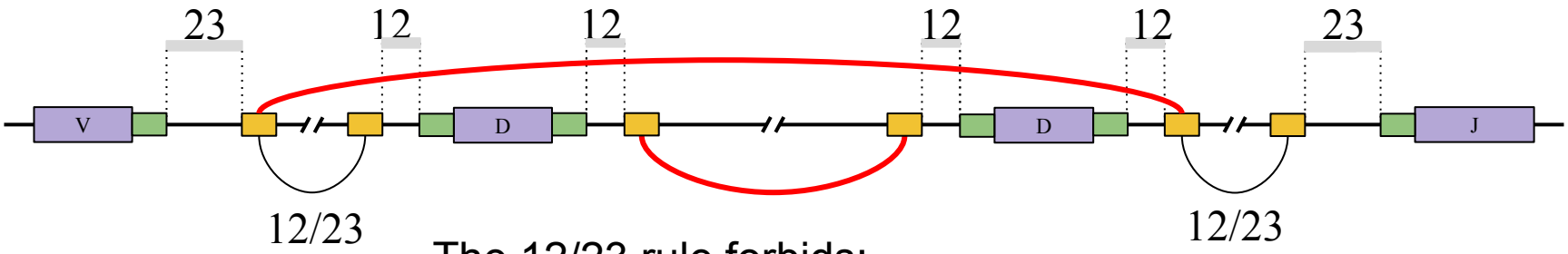


Safonova and Pevzner, *Front Immunol*, 2019

Recombination signal sequences





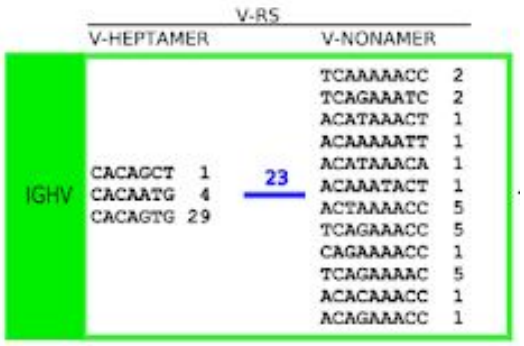


The 12/23 rule forbids:

- V-J pairs
- D-D pairs
- V-V pairs
- J-J pairs

12 nt = 1 turn of DNA
23 nt = 2 turns of DNA

Non-canonical recombination signals

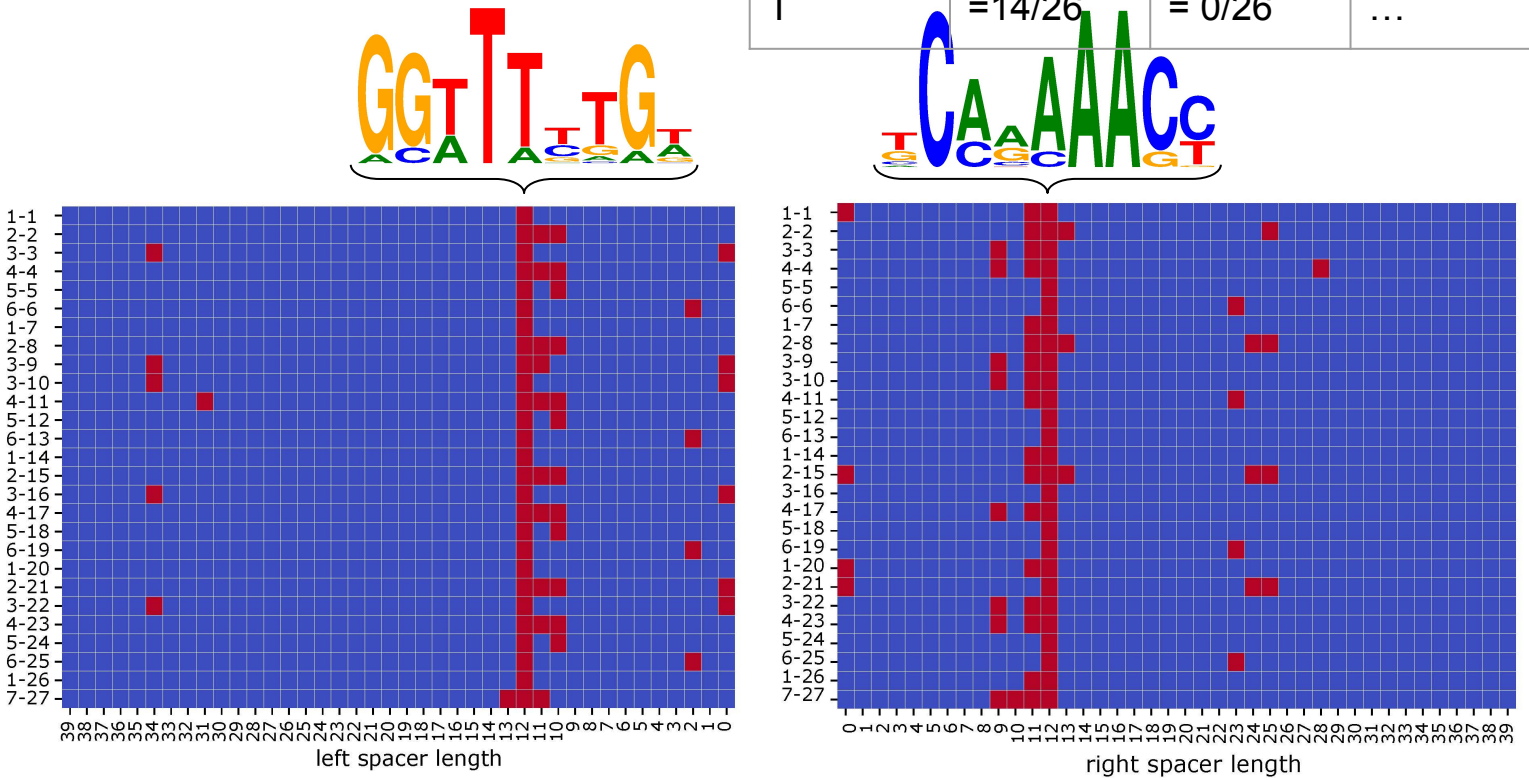


	Pos 1	Pos 2	…
A	= 11/26	= 1/26	…
C	= 1/26	= 25/26	…
G	= 0/26	= 0/26	…
T	=14/26	= 0/26	…

Non-canonical recombination signals

Prob(ACGTACGTA) =
= 11/26 * 25/26 * ...

	Pos 1	Pos 2	…
A	= 11/26	= 1/26	…
C	= 1/26	= 25/26	…
G	= 0/26	= 0/26	…
T	=14/26	= 0/26	…



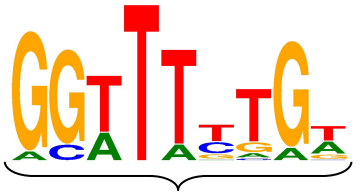
Safonova and Pevzner, *Genome Res*, 2020

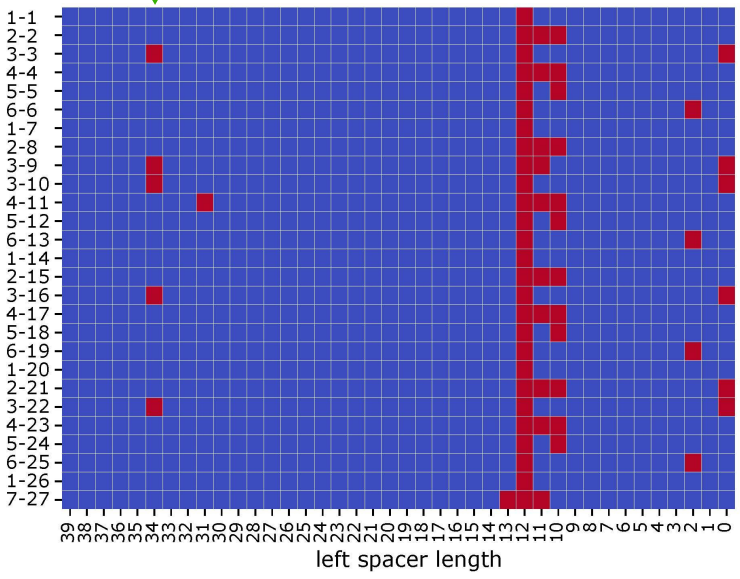
Non-canonical recombination signals

Prob(ACGTACGTA) =
= 11/26 * 25/26 * ...

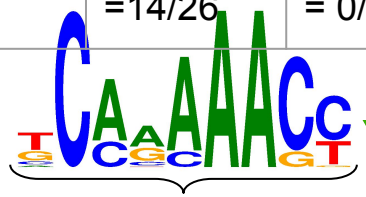
	Pos 1	Pos 2	…
A	= 11/26	= 1/26	…
C	= 1/26	= 25/26	…
G	= 0/26	= 0/26	…
T	=14/26	= 0/26	…

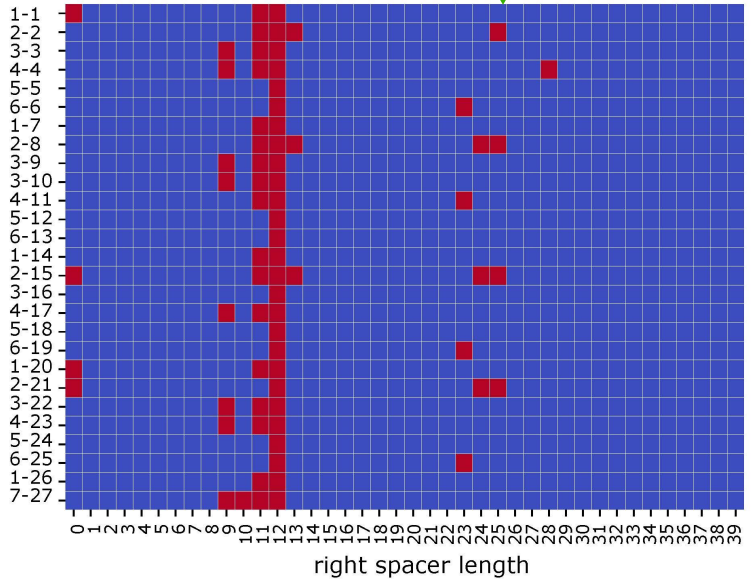
3 DNA turns





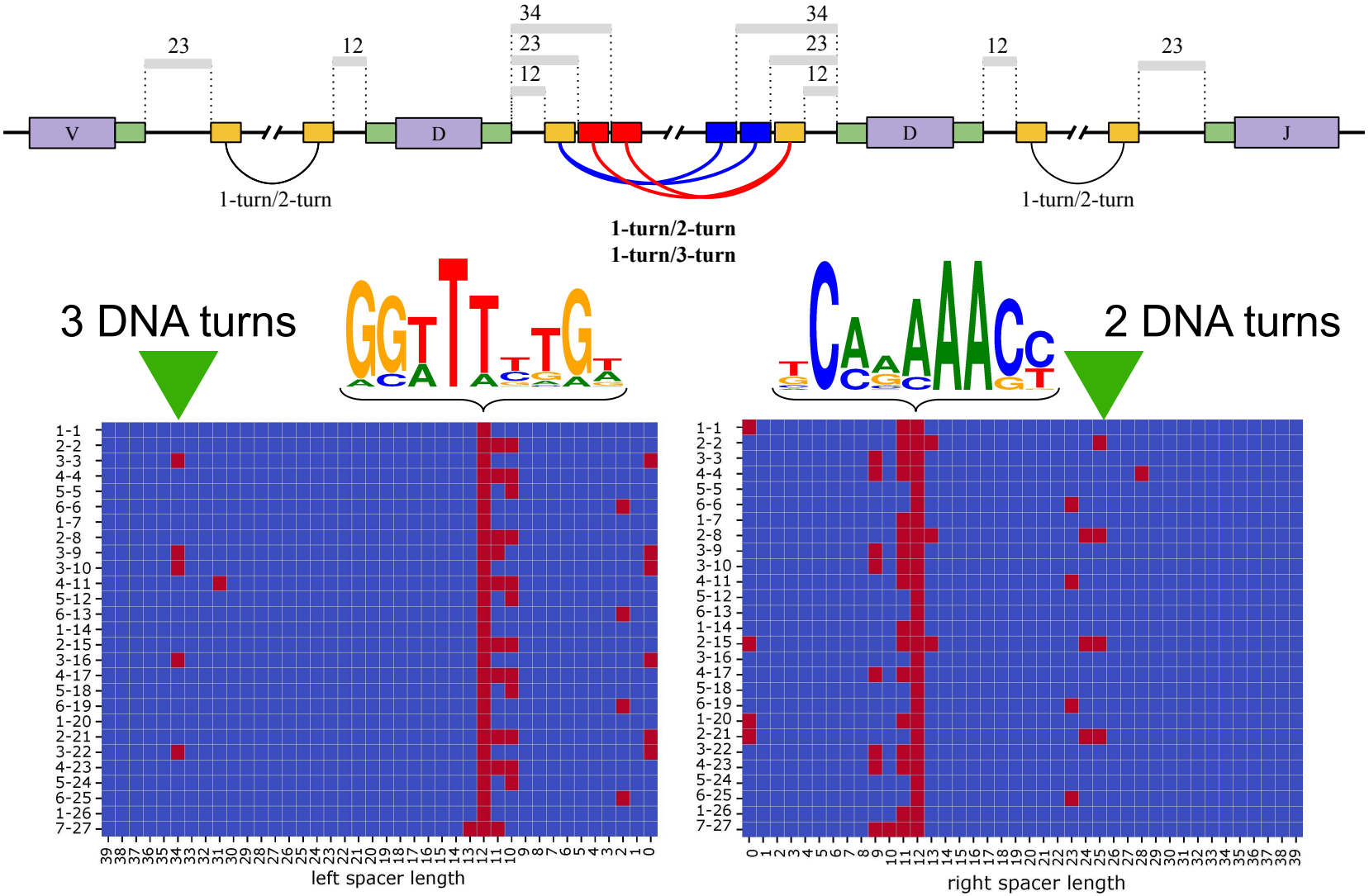
2 DNA turns





Safonova and Pevzner, *Genome Res*, 2020

Non-canonical recombination signals



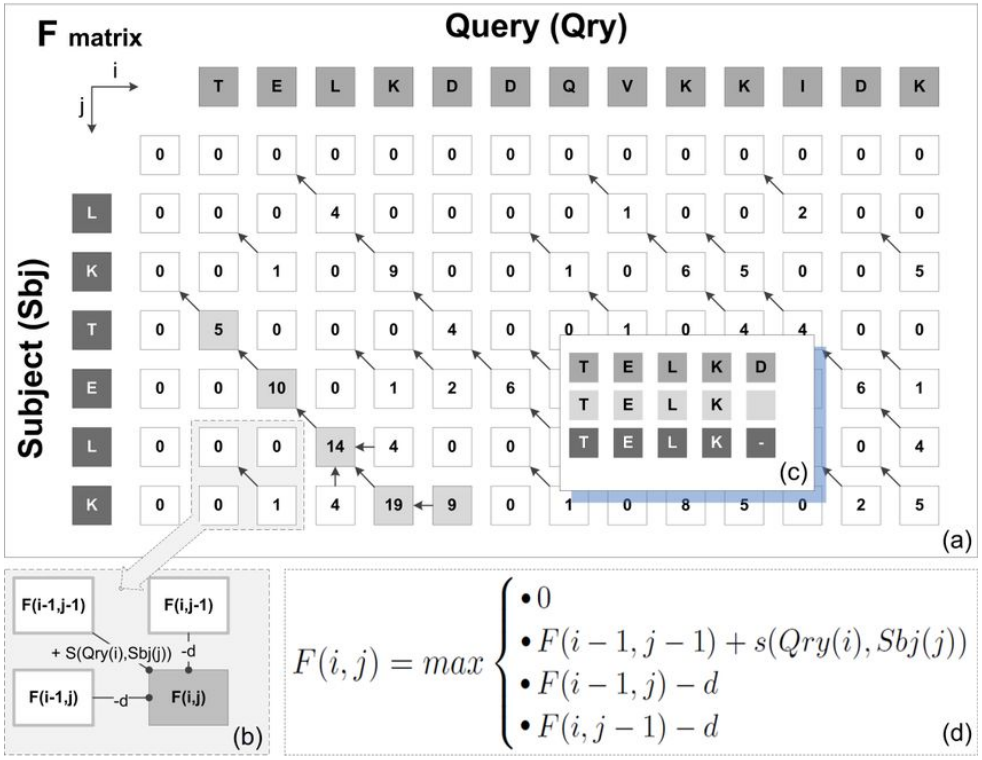
Safonova and Pevzner, *Genome Res*, 2020

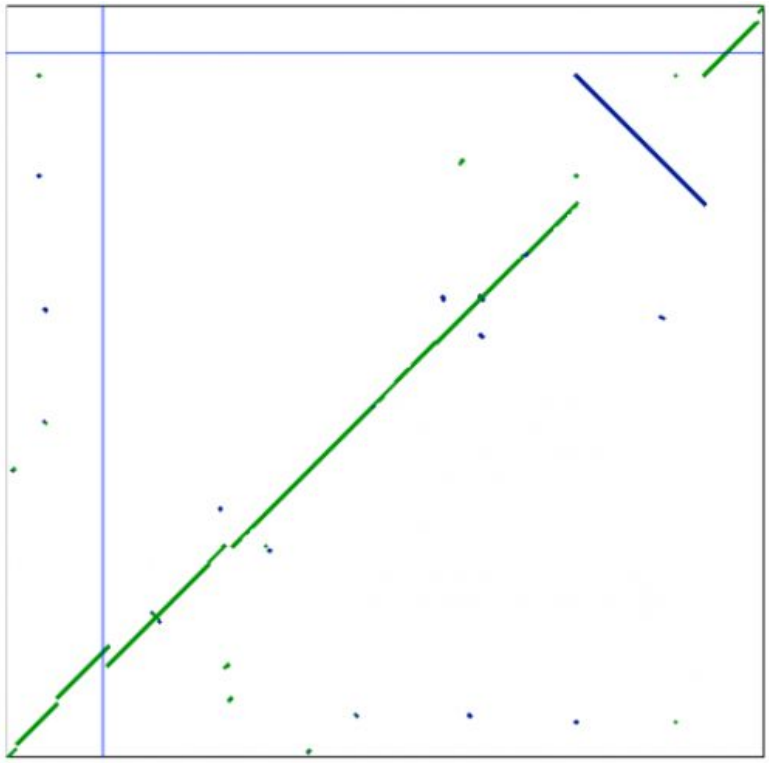
Sequence alignment

Target Sequence

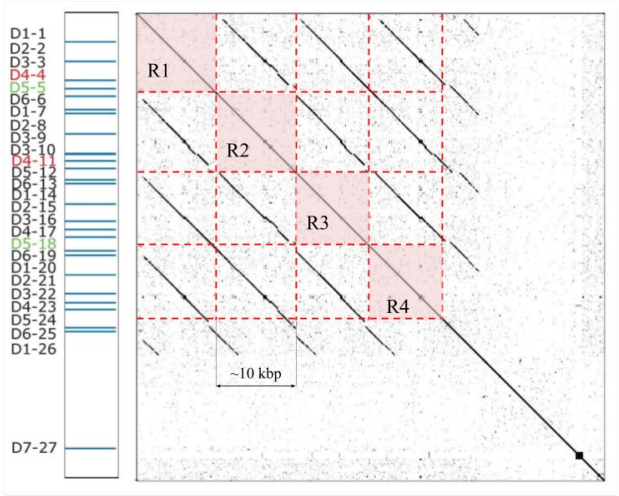
```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||||    |||||||  ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence



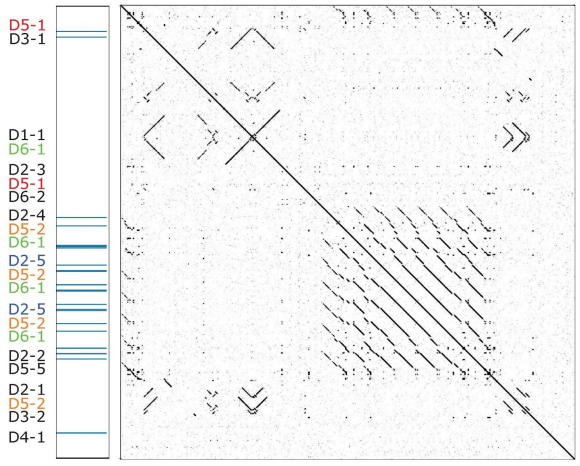


Tandem repeats correlate with D-D fusions

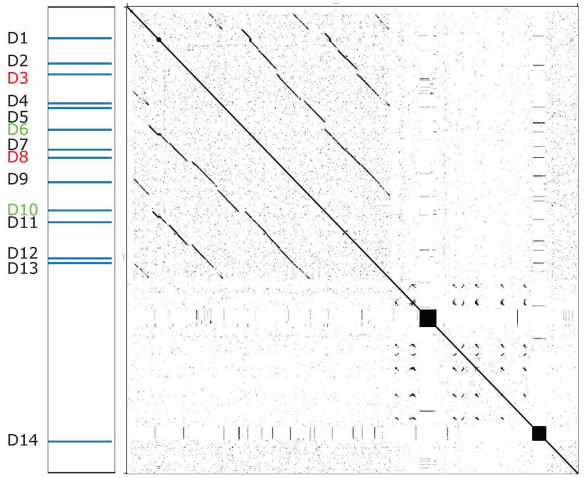


R1	R2	R3	R4
–	D6-6	D6-13	D6-19
D1-1	D1-7	D1-14	D1-20
D2-2	D2-8	D2-15	D2-21
D3-3	D3-9 D3-10	D3-16	D3-22
D4-4	D4-11	D4-17	D4-23
D5-5	D5-12	D5-18	D5-24

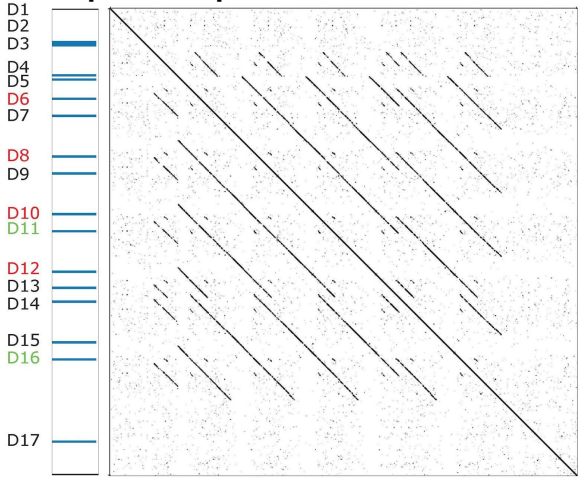
mouse



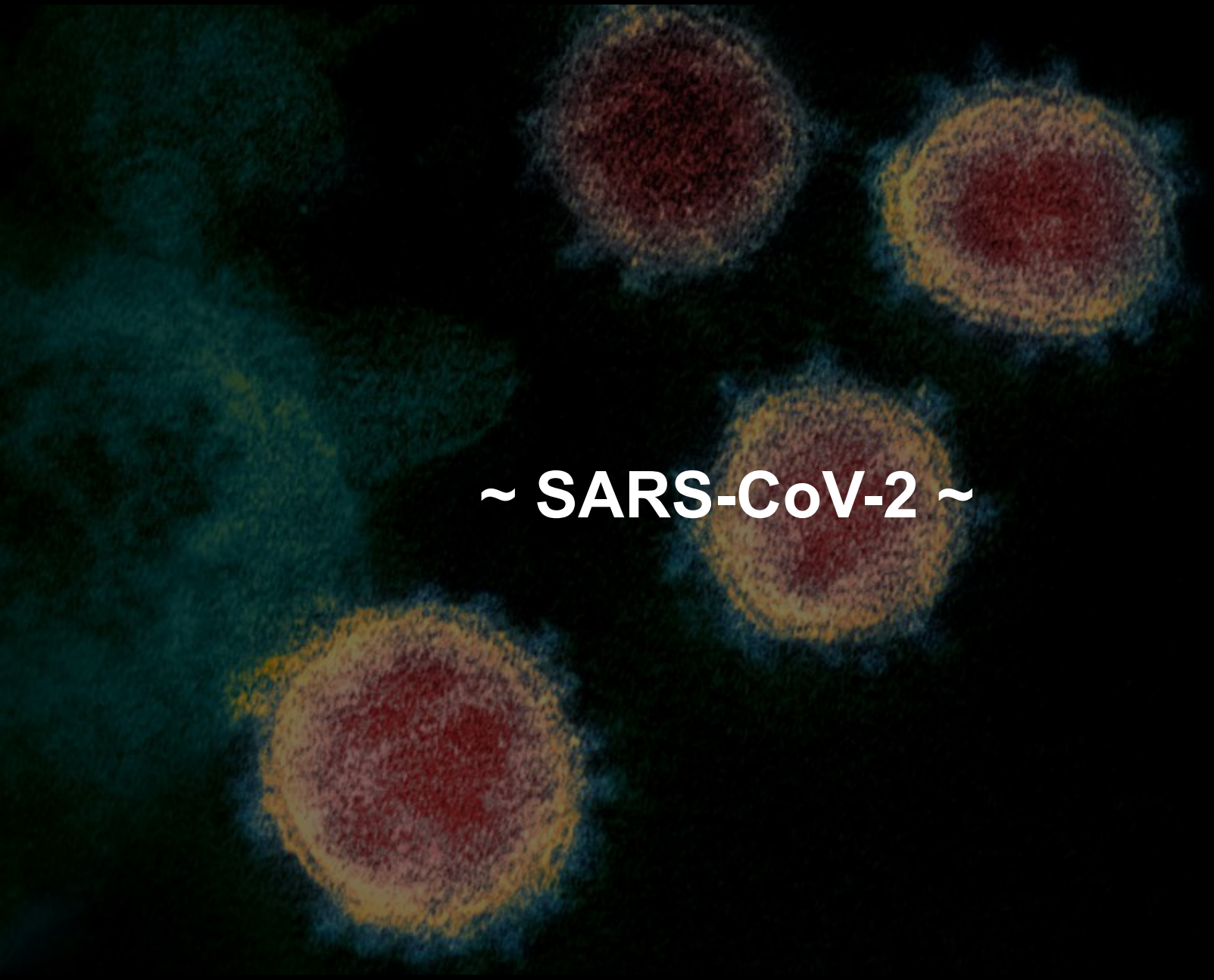
Common marmoset



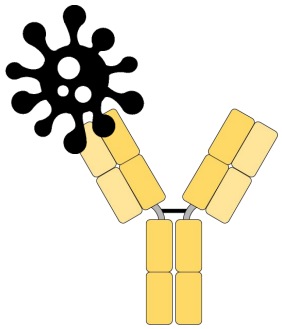
pale spear-nosed bat



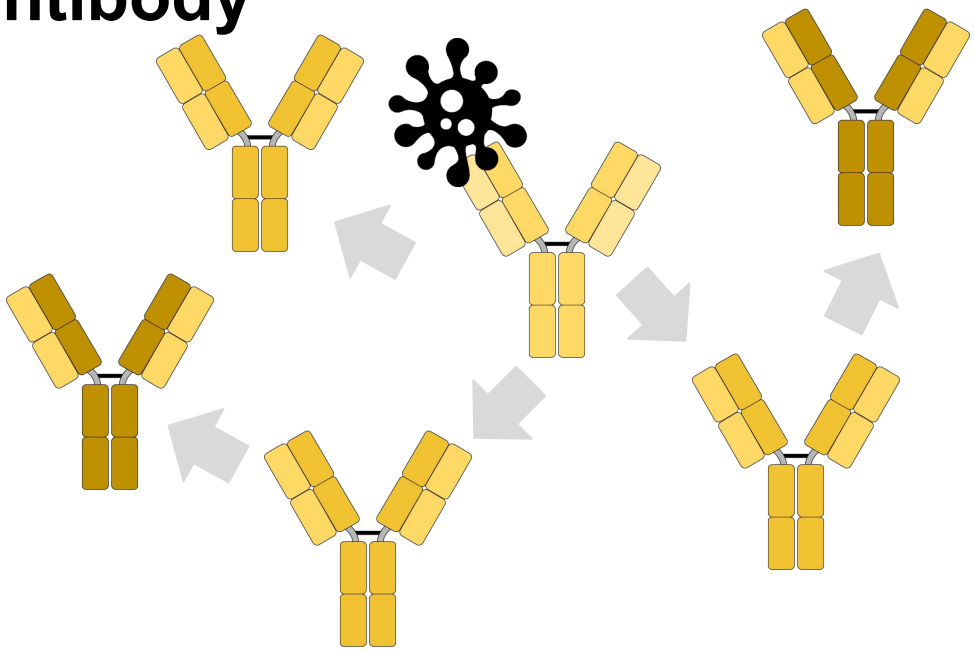
Safonova and Pevzner, *Genome Res*, 2020


~ SARS-CoV-2 ~

Antibodies are subjects of fast evolution

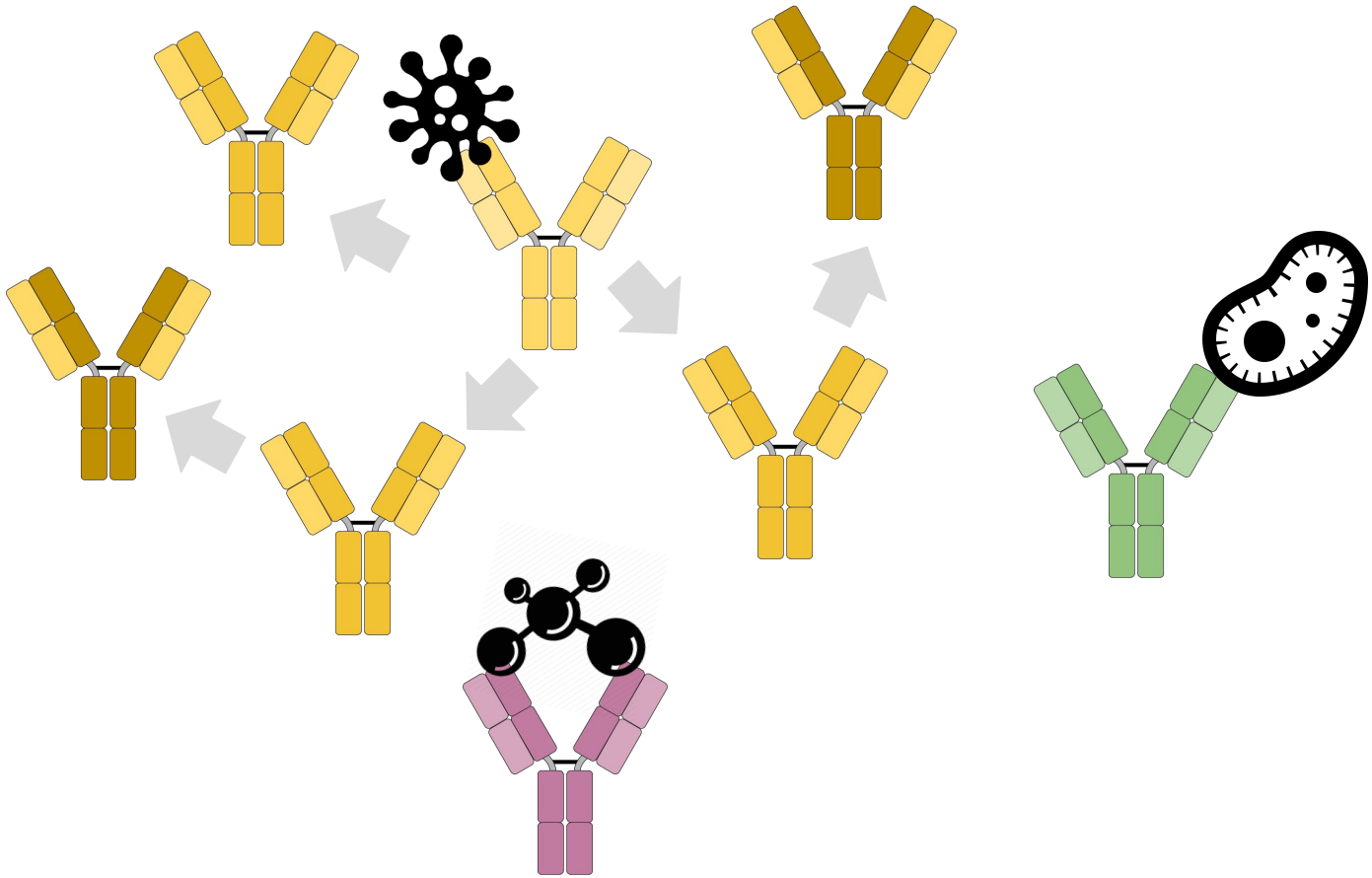


Immune system mutates and amplifies a binding antibody

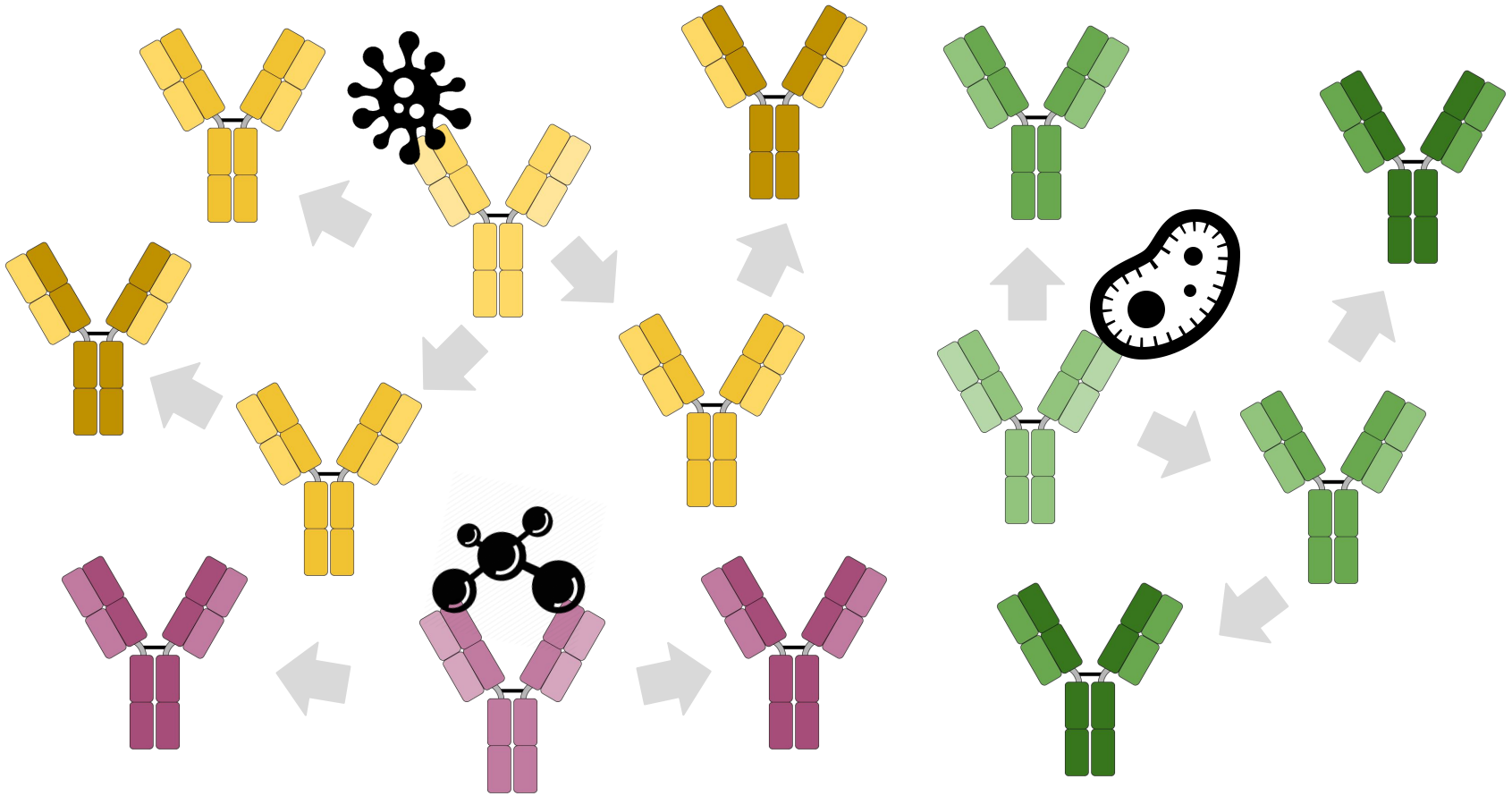


Mutation rate in antibody genes is 3-4 order of magnitude higher than in other genome

One antibody = one antigen

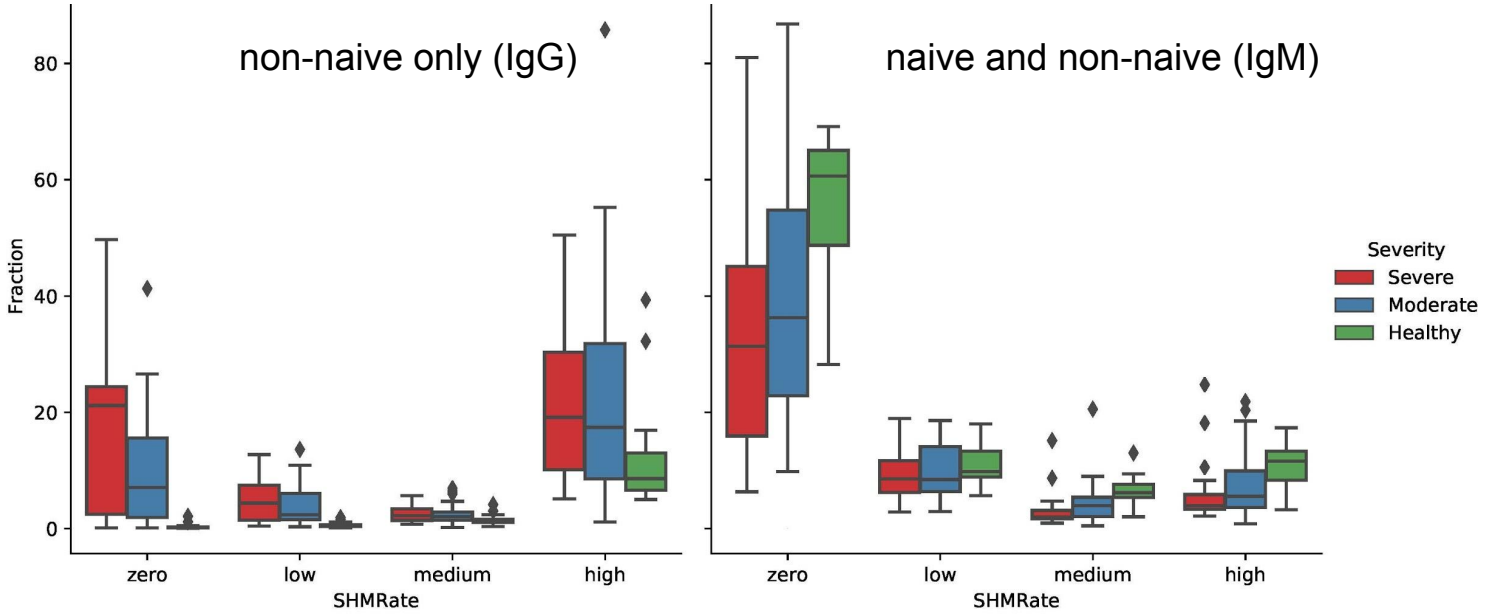


Antibody repertoire is a set of clonal lineages



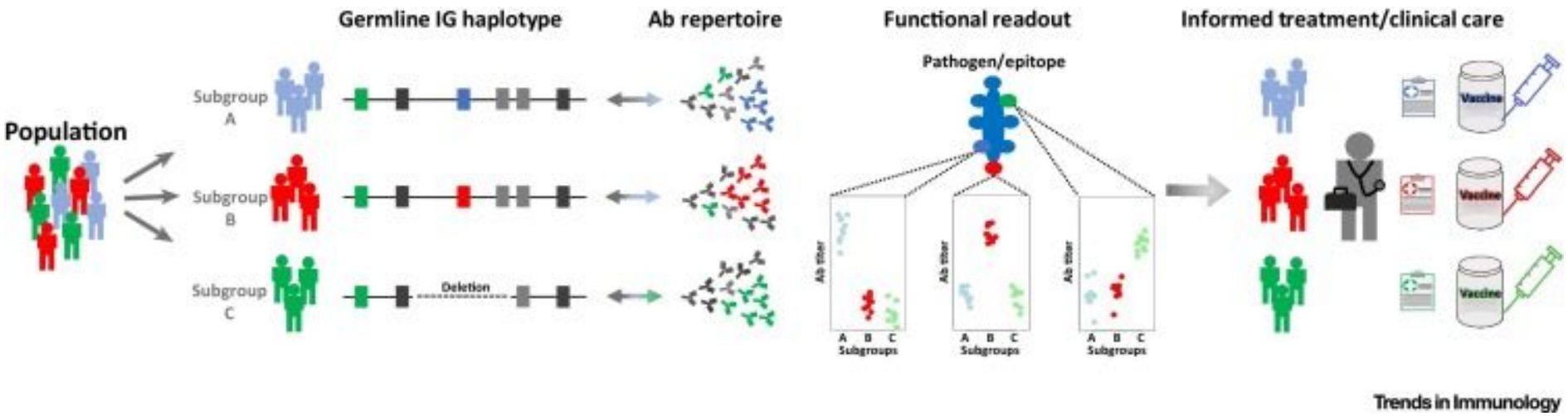
Severity of COVID-19 and Ab response

```
                                                <-------------------------------FR1-IMGT-------------------------------><-------CDR1-IMG
                                                   V  Q  L  V  Q  S  G  G  G  L  V  Q  P  G  G  S  L  R  L  S  C  V  A  S  G  F  T  F  S
                                 Query_1      2    AGGTGCAGCTGGTGCAGTCTGGAGGAGGCTTGGTCCAGCCTGGGGGGTCCCTGAGACTCTCCTGTGTAGCGTCTGGCTTCACCTTCAGTA  91
V  90.2% (266/295)               IGHV3-30*02  2    .............G.......G......G.....................................C.........A.............  91
                                                   V  Q  L  V  E  S  G  G  G  V  V  Q  P  G  G  S  L  R  L  S  C  A  A  S  G  F  T  F  S
V  90.2% (266/295)               IGHV3-30-5*02 2    .............G.......G......G.....................................C.........A.............  91
V  89.5% (264/295)               IGHV3-30*18  2    .............G.......G......G..............A......................C...C.....A.............  91

                                                T------><--------------------FR2-IMGT--------------------><-------CDR2-IMGT------><------
                                                N  F  G  M  H  W  V  R  Q  A  P  G  K  G  L  E  W  V  A  M  T  Q  P  D  E  G  T  E  Y  Y
                                 Query_1      92   ATTTTGGCATGCACTGGGTCCGCCAGGCTCCAGGCAAGGGACTGGAGTGGGTGGCAATGACACAACCTGATGAAGGTACTGAATACTATG  181
V  90.2% (266/295)               IGHV3-30*02  92   GC.A.....................................G..............T.T.T..GGTA.....G.A...A.A.........  181
                                                S  Y  G  M  H  W  V  R  Q  A  P  G  K  G  L  E  W  V  A  F  I  R  Y  D  G  S  N  K  Y  Y
V  90.2% (266/295)               IGHV3-30-5*02 92   GC.A.....................................G..............T.T.T..GGTA.....G.A...A.A.........  181
V  89.5% (264/295)               IGHV3-30*18  92   GC.A.....................................G..............G.T.T.TC.TA.....G.A...A.A.........  181

                                                ---------------------------------------------FR3-IMGT-------------------------------------
                                                G  D  S  V  K  G  R  F  T  I  S  R  D  N  S  R  N  T  L  Y  L  Q  M  N  S  L  R  T  E  D
                                 Query_1      182  GAGACTCCGTGAAGGGCCGATTTACCATCTCCAGAGACAATTCCAGGAACACACTGTATCTGCAAATGAACAGCCTGAGAACTGAGGACA  271
V  90.2% (266/295)               IGHV3-30*02  182  C.....................C.....................A......G.....................G.........  271
                                                A  D  S  V  K  G  R  F  T  I  S  R  D  N  S  K  N  T  L  Y  L  Q  M  N  S  L  R  A  E  D
V  90.2% (266/295)               IGHV3-30-5*02 182  C.....................C.....................A......G.....................G.........  271
V  89.5% (264/295)               IGHV3-30*18  182  C.....................C.....................A......G.....................G.........  271
```



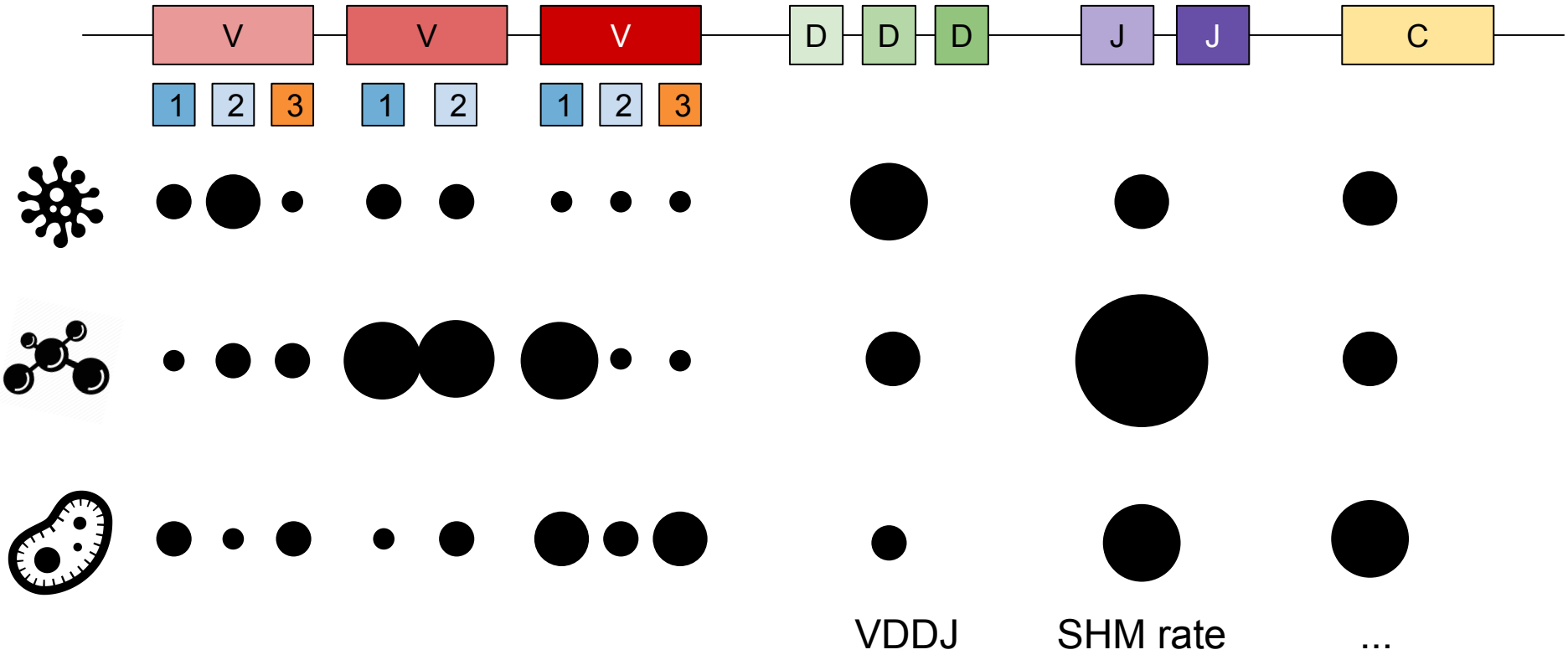
Safonova, Tieri, et al., in preparation

Immunogenomics approach to vaccine design



Watson, Glanville, Marasco, *Trends in Immunol*, 2017

Data science approach to predicting the efficiency of antibody response



Acknowledges

Pavel Pevzner
UCSD

Corey Watson
U of Louisville

Harvard Medical School
Wayne Marasco
Hanzhong Ke

Yale University
Steven Kleinstein
Nima Nouri

UC San Diego
Data Science postdoctoral fellowship

THE AMERICAN ASSOCIATION OF IMMUNOLOGISTS
Intersect fellowship for computational immunologists

U of Louisville
William Gibson
Justin Kos
Oscar Rodriguez
David Tieri
Jun Yan

U of New South Wales
Andrew Collins
Katherine Jackson

UCSD
Massimo Franceschetti
Siavash Mirarab
Ramesh Rao
Andrey Bzikadze
Vinnu Bhardwaj
Chao Zhang

USDA
Tim Smith
Sung Bong Shin

Iowa State U
James Reecy
Luke Kramer

Scripps Institute
Raiees Andrabi
Vaughn Smider

Smithsonian Conservation Biology Institute
Klaus-Peter Koepfli

Institute for Bioorganic Chemistry
Ivan Zvyagin
Artem Mikelov
Mikhail Shugay