Immunoinformatics: application of algorithmic approaches to solving immunological problems

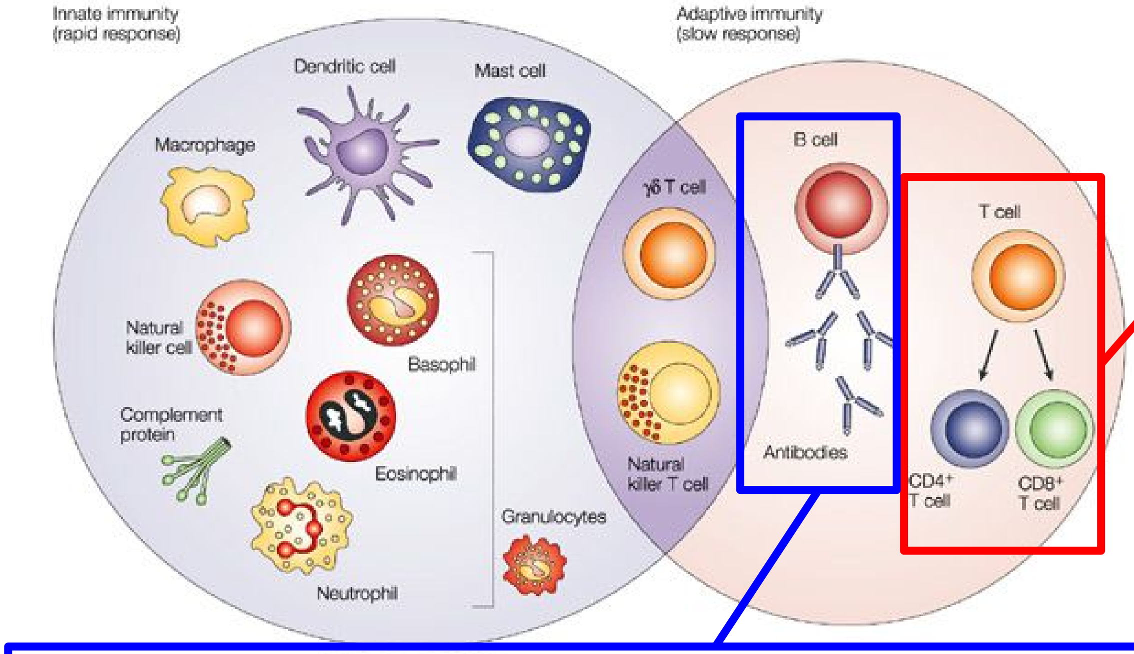
Yana Safonova

Center for Algorithmic Biotechnology
St. Petersburg State University

Outline

- **Introduction**
- Repertoire construction problem
- Evolutionary analysis of antibodies
- Analysis of immune response dynamics
- Analysis of paired antibody repertoires & new biological insights from analysis of paired repertoires

Innate & adaptive immune system



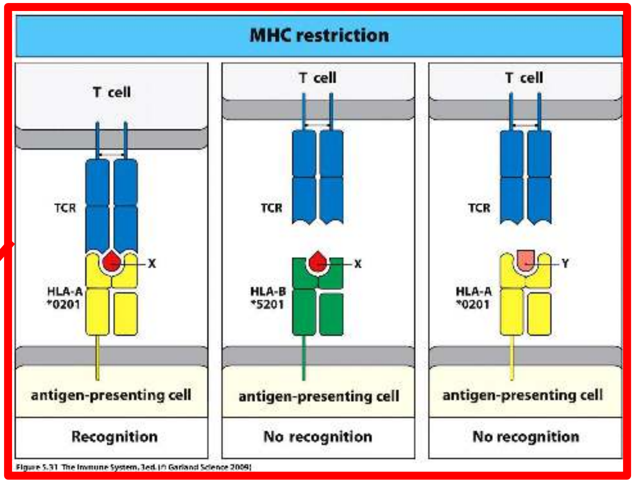


Figure 5.31 The Immune System, 3ed.(© Garland Science 2009)

cell-mediated immune response

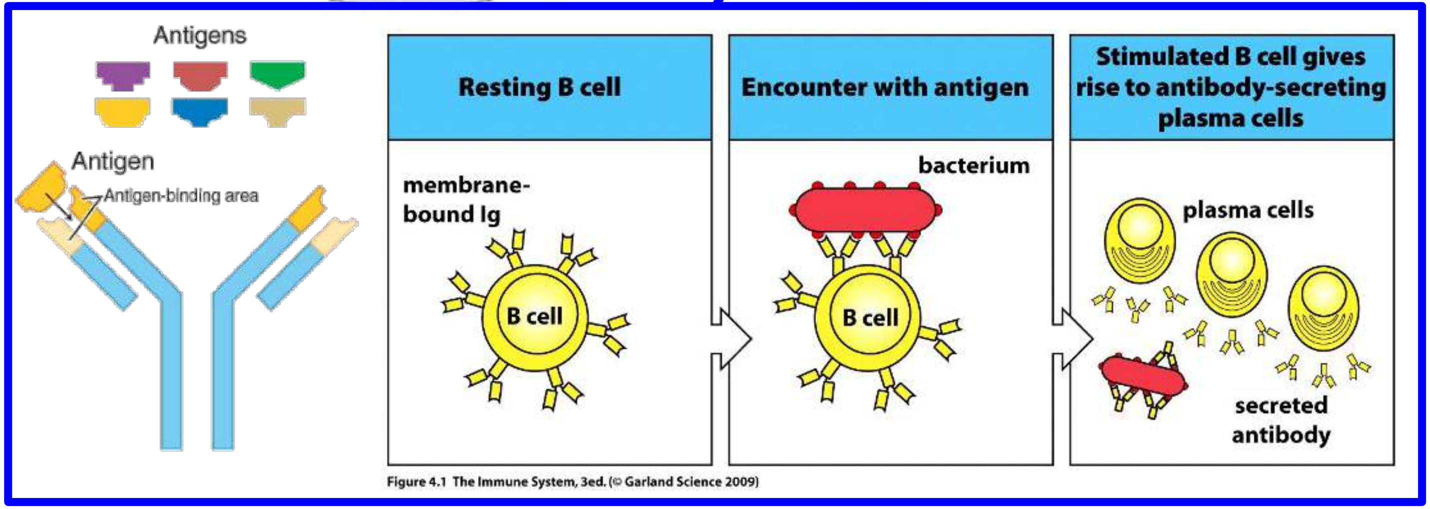


Figure 4.1 The Immune System, 3ed. (© Garland Science 2009)

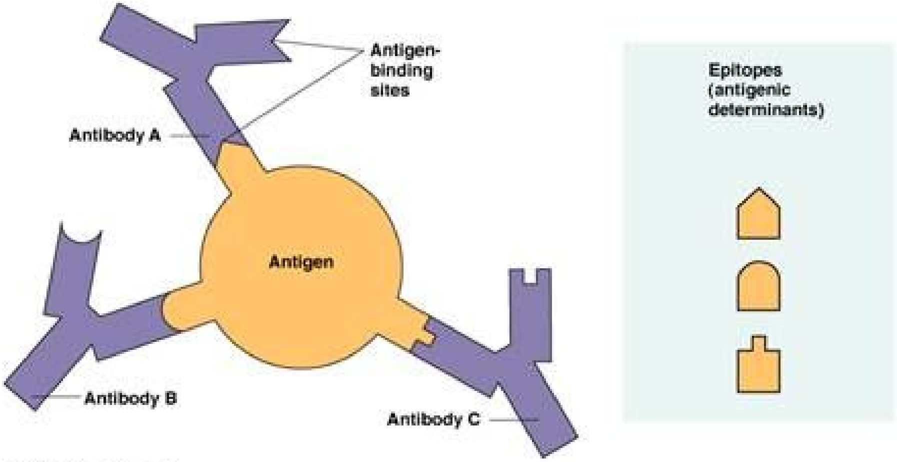
humoral immune response

Antibody & antigen

Antigen recognition

Antibody & antigen

Antigen recognition

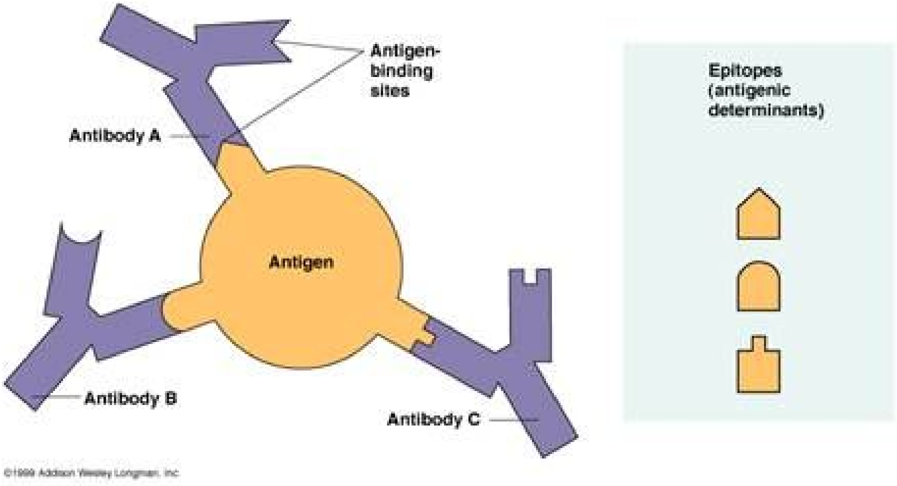


©1999 Addison Wesley Longman, Inc

Antibody - antigen binding

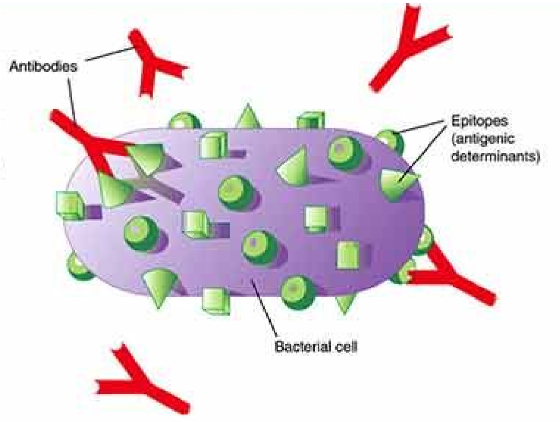
Antibody & antigen

Antigen recognition



©1999 Addison Wesley Longman, Inc

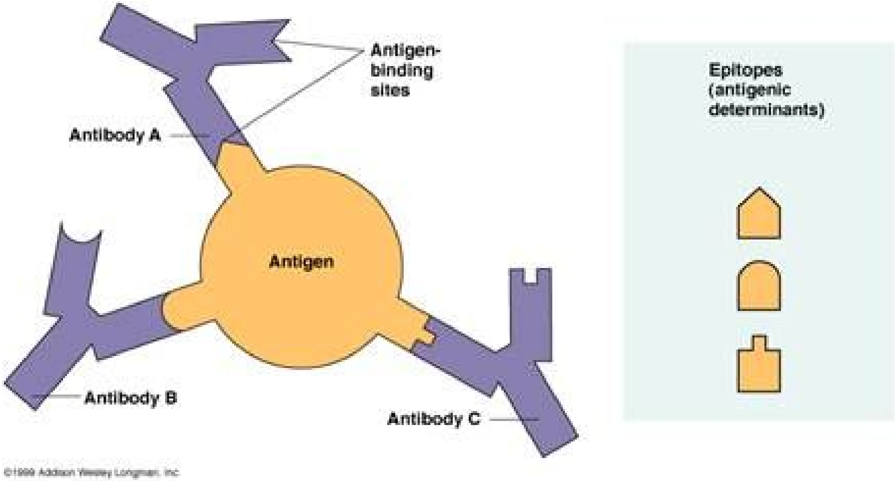
Antibody - antigen binding



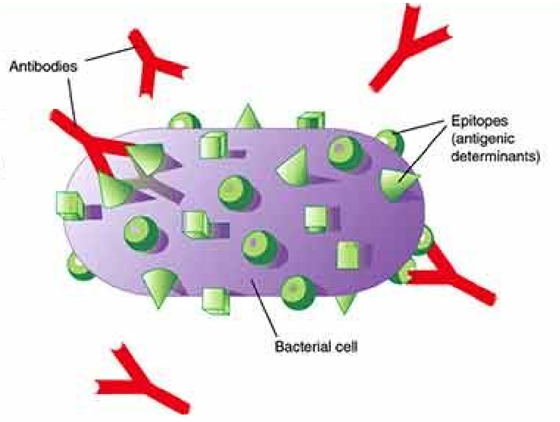
1. Antigen neutralization

Antibody & antigen

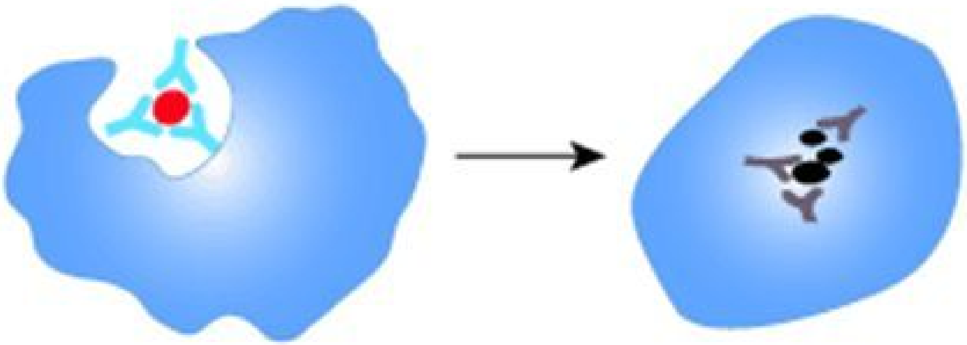
Antigen recognition


Antigen-binding sites
Antibody A
Antigen
Antibody B
Antibody C
Epitopes (antigenic determinants)
©1999 Addison Wesley Longman, Inc

Antibody - antigen binding

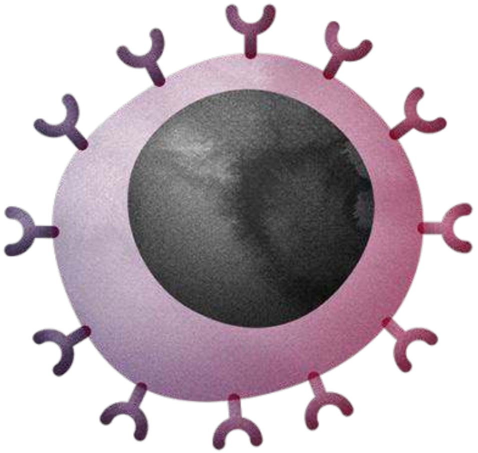


1. Antigen neutralization

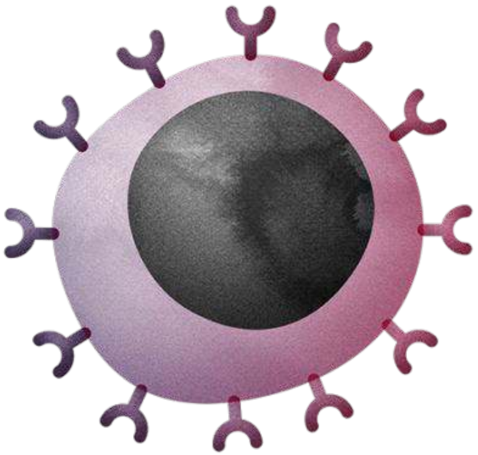


2. Destroying antigen by immune cells

Once you’ve met
an antigen,
your adaptive
immune system
never forgets it!



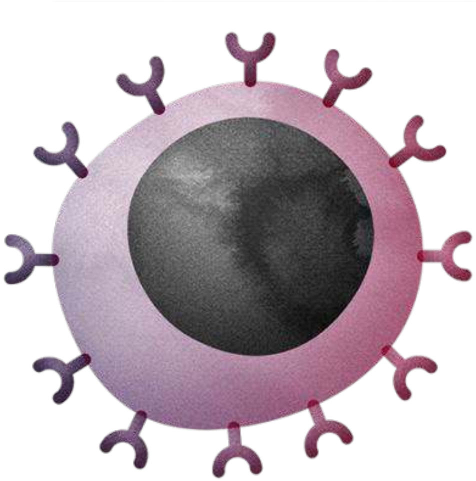
Once you've met
an antigen,
your adaptive
immune system
never forgets it!



This principle is used for vaccine design:

Real antigens

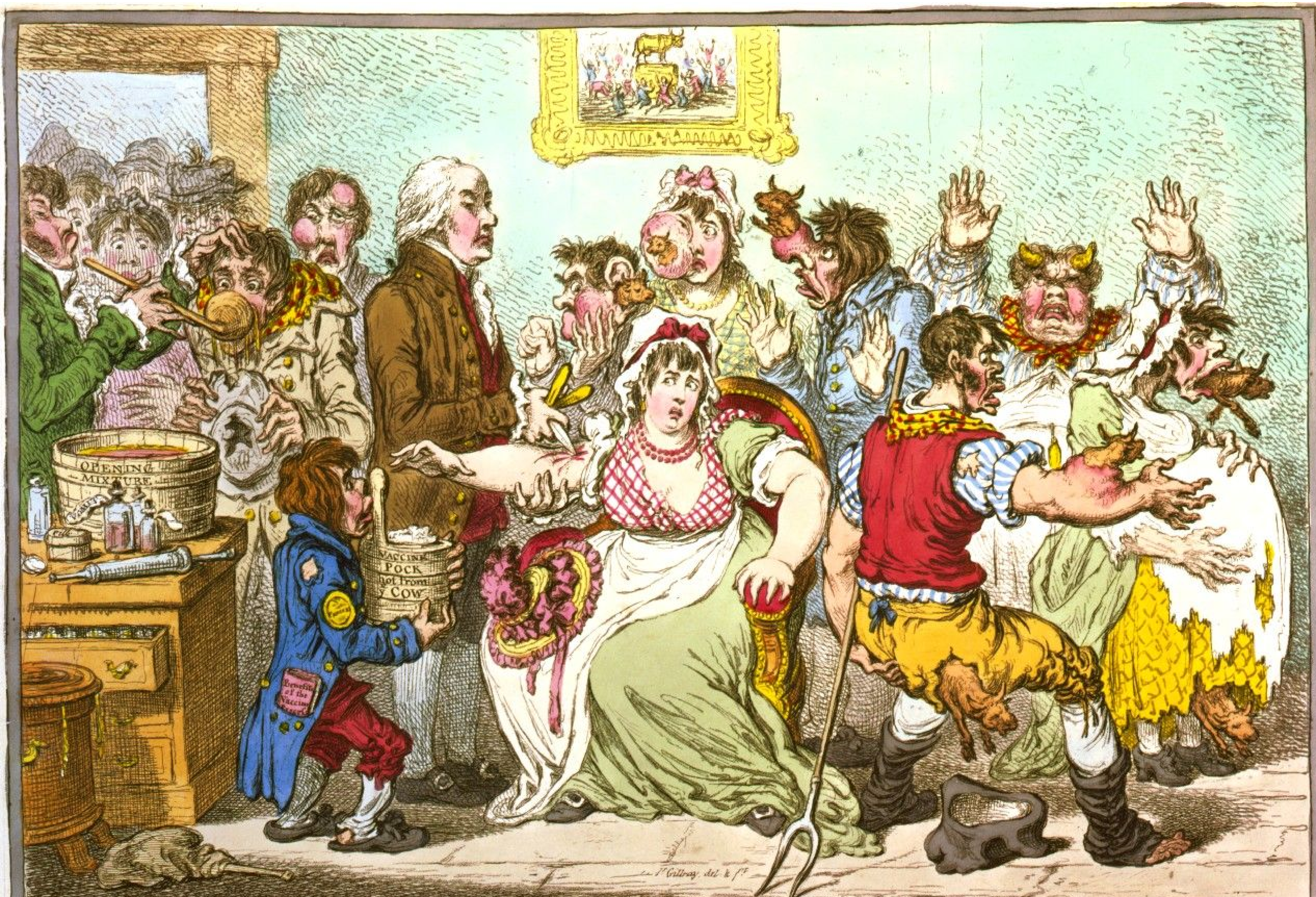
Once you've met
an antigen,
your adaptive
immune system
never forgets it!



This principle is used for vaccine design:

Real antigens

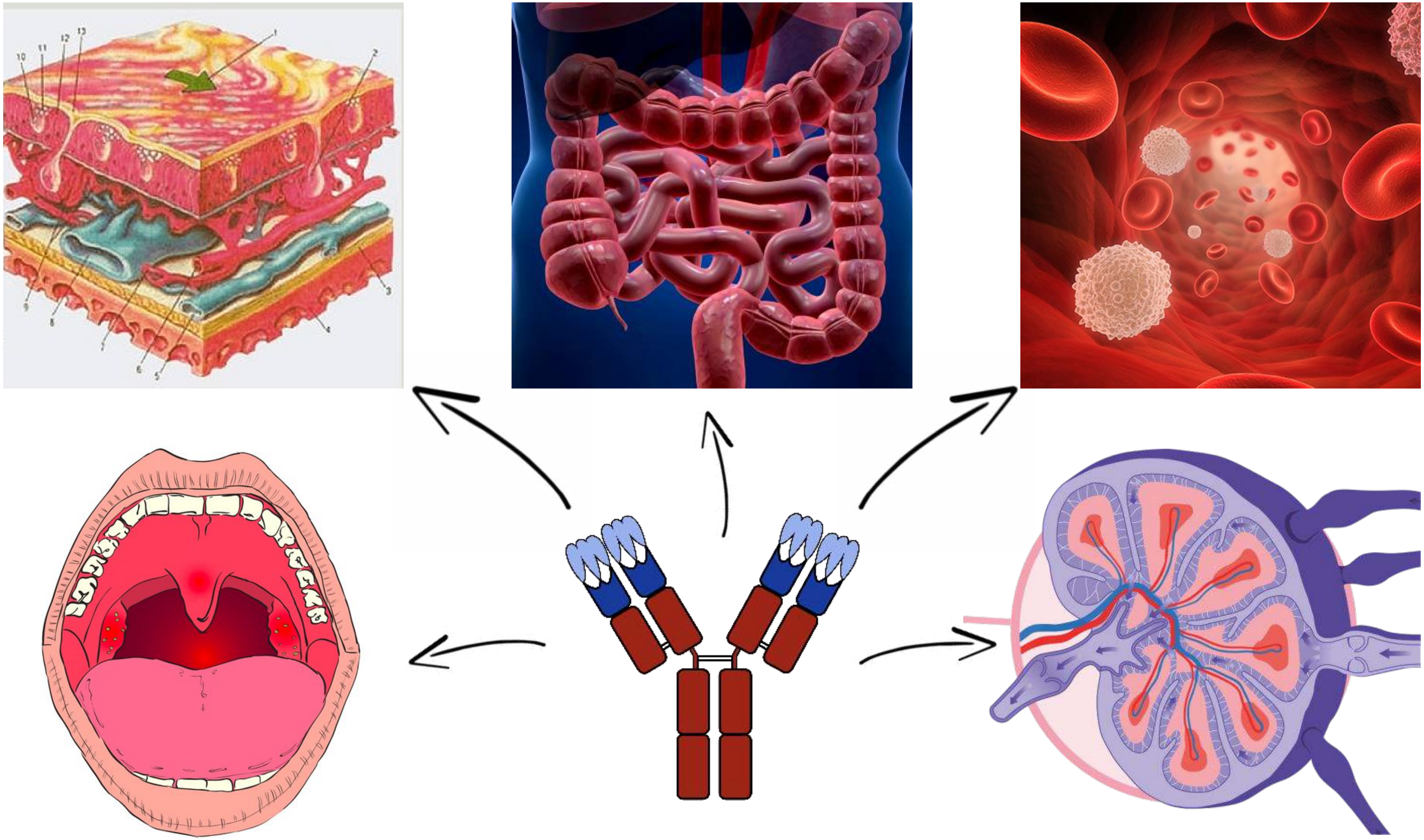
Vaccine



Pubd June 12th 1802. by H. Humphrey St James's Street.

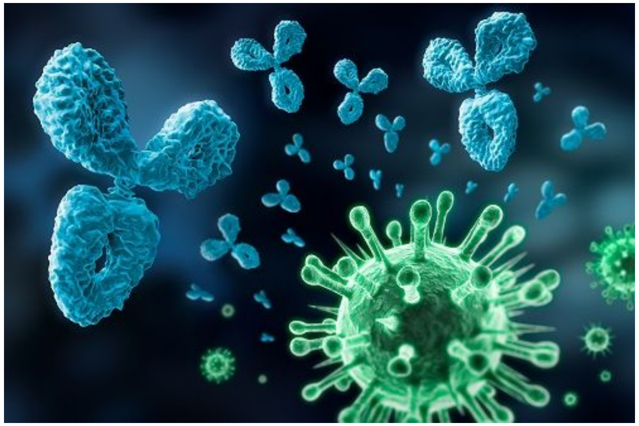
The Cow-Pock — or — the Wonderful Effects of the New Inoculation! — Vide. the Publications of ye Anti-Vaccine Society.

Where do antibody live?



Antibody repertoires

There is a **billion** of B-cells circulating in human blood at any given moment (out of 10^{18} estimated antibodies)

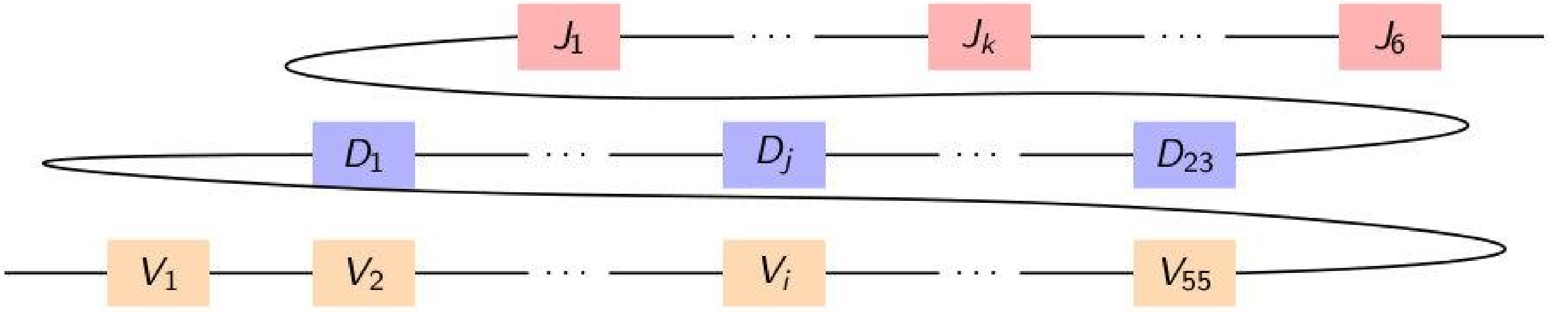


*Analysis of concentrations of all antibodies in the organism (**antibody repertoire**) is a fundamental problem in immunology*

While generation of antibody repertoires provides a new avenue for antibody drug development, it remains unclear how to construct antibody repertoires from NGS data

V(D)J recombination

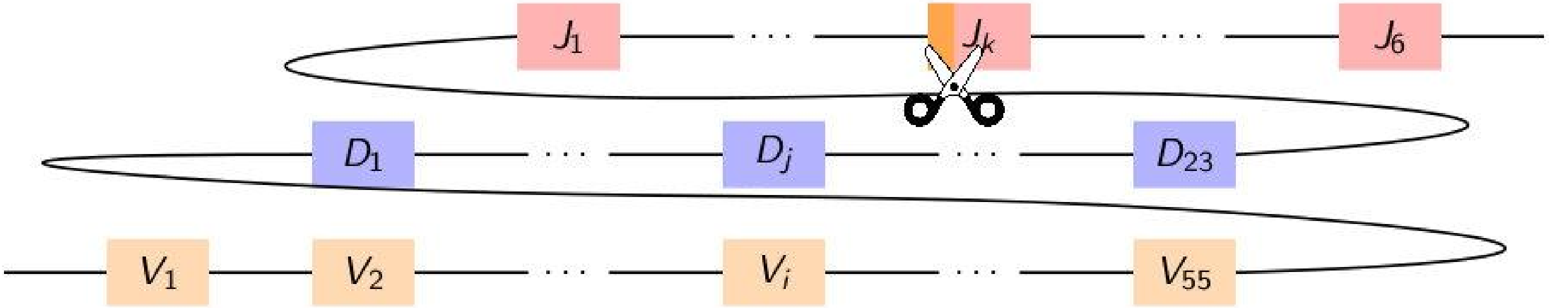
Antibodies are produced by ***B-cells***, each with unique genome:



IGH locus in human genome (1 MB length)

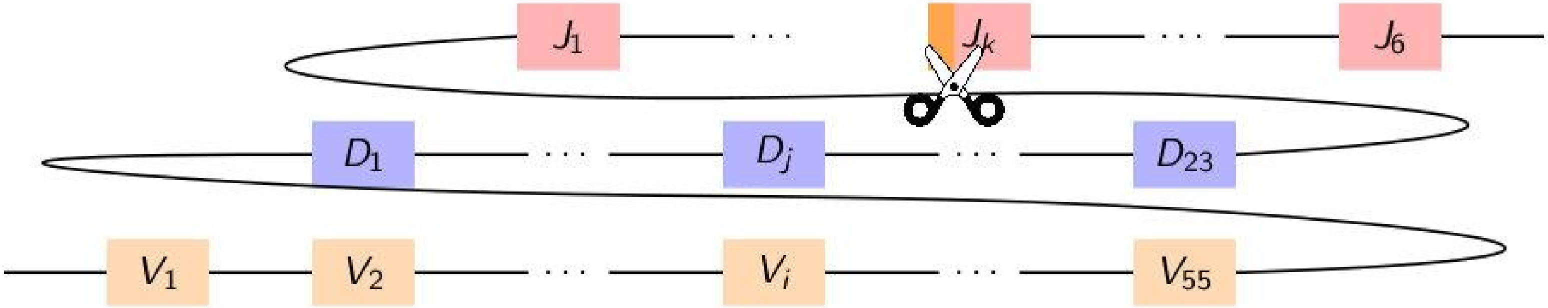
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



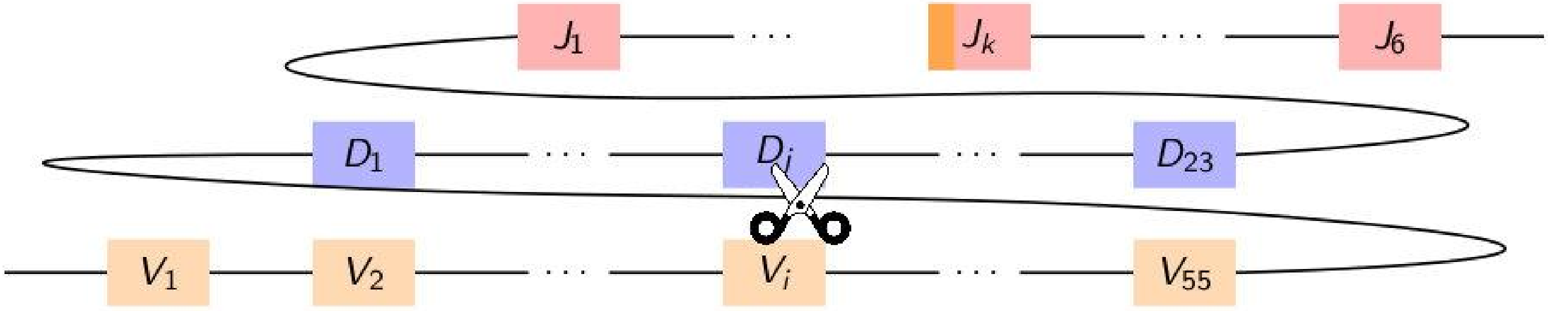
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



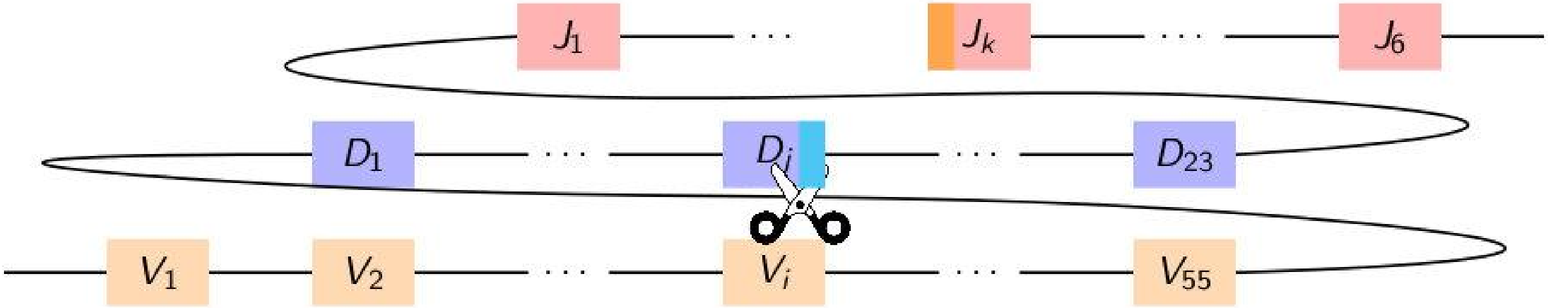
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



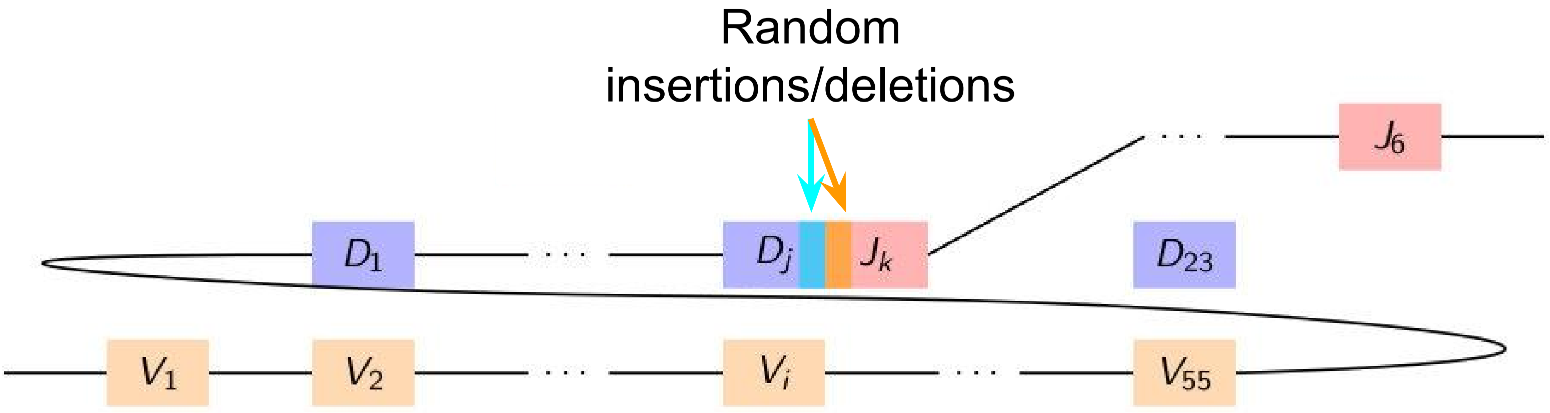
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



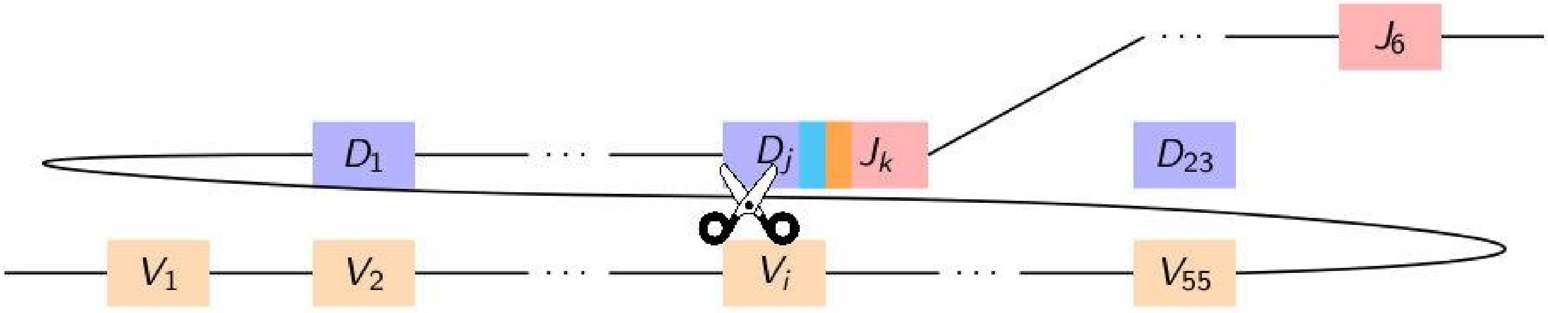
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



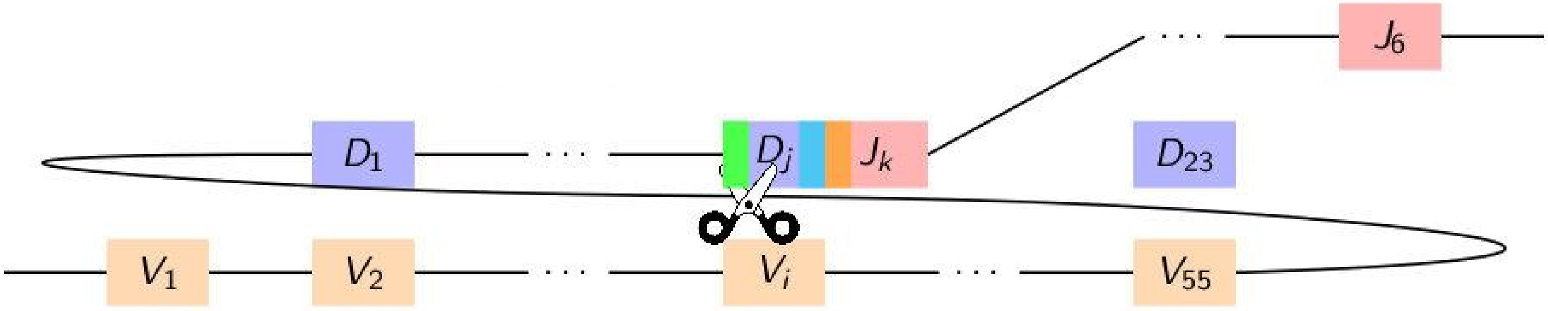
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



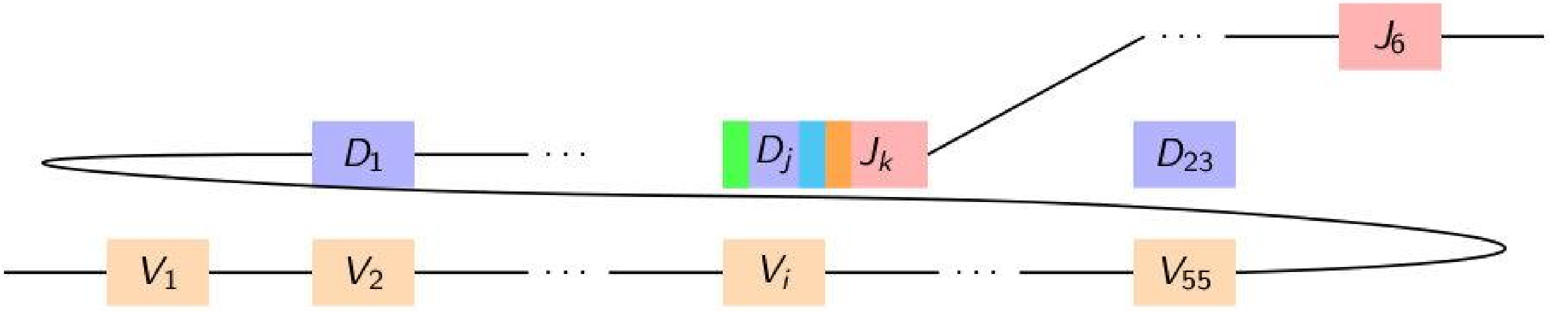
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



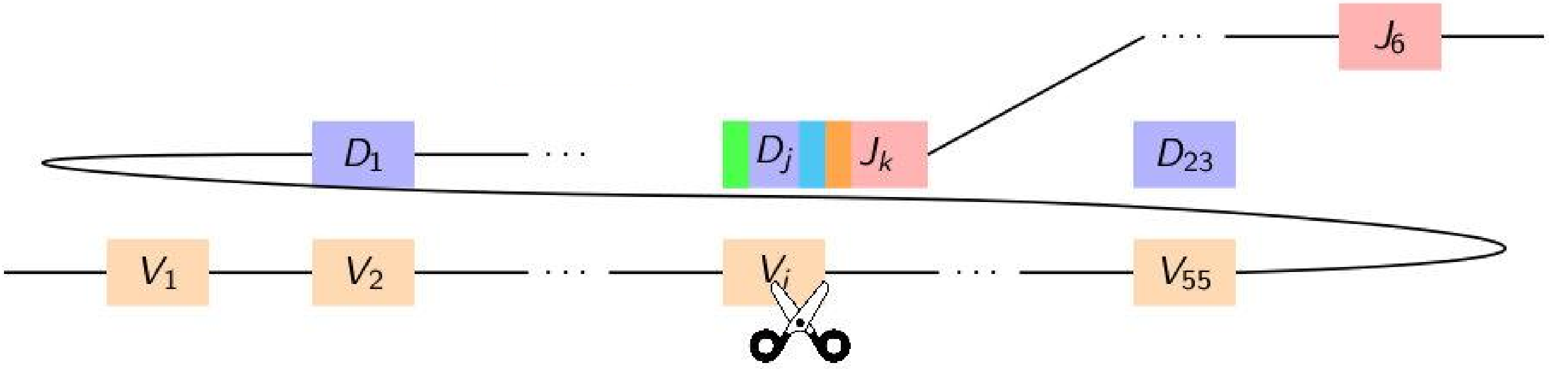
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



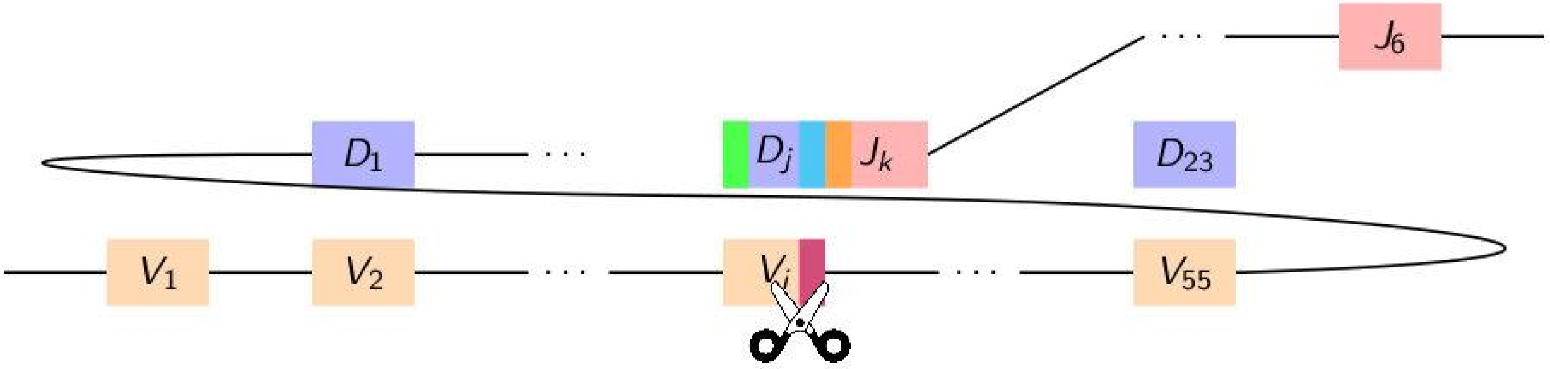
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



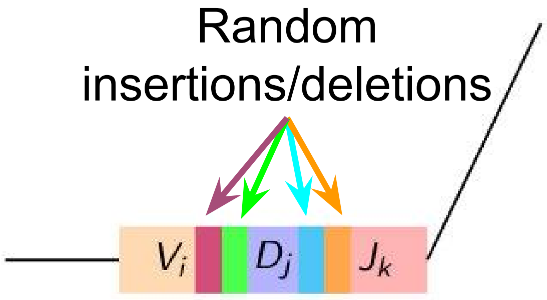
Antibody somatic recombination

Antibodies are produced by ***B-cells***, each with unique genome:



Antibody somatic recombination

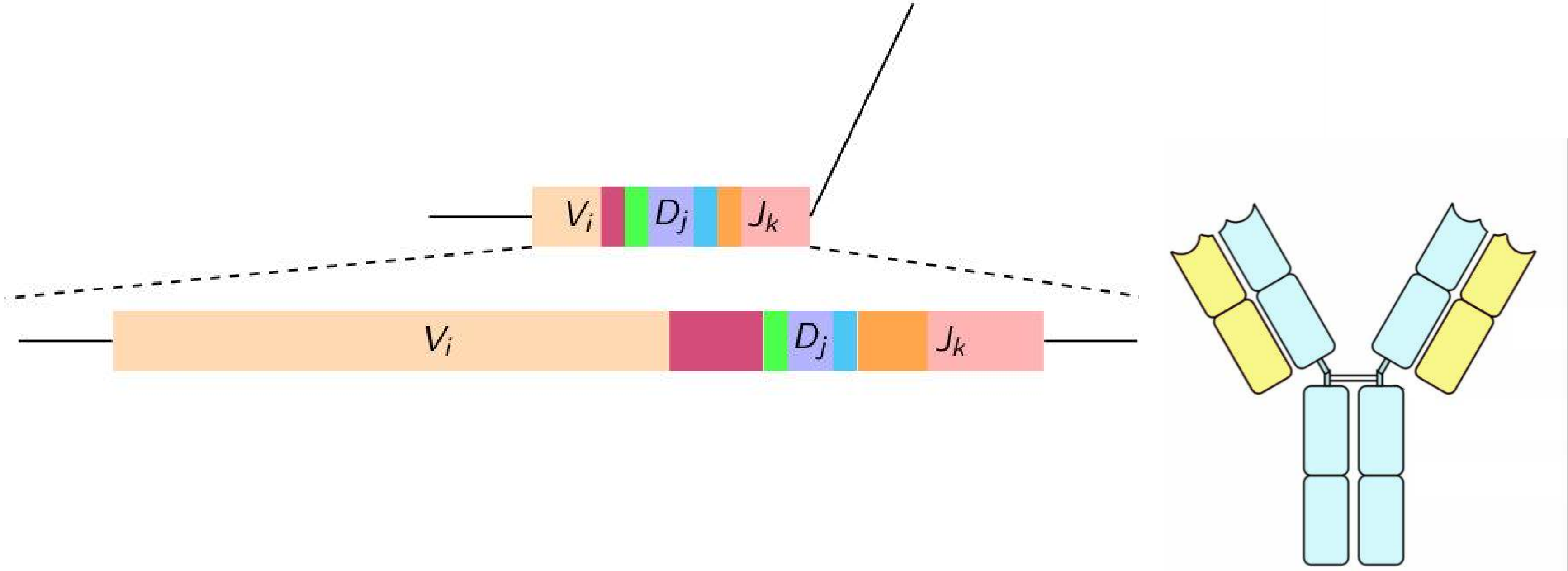
Antibodies are produced by ***B-cells***, each with unique genome:



Random insertions/deletions

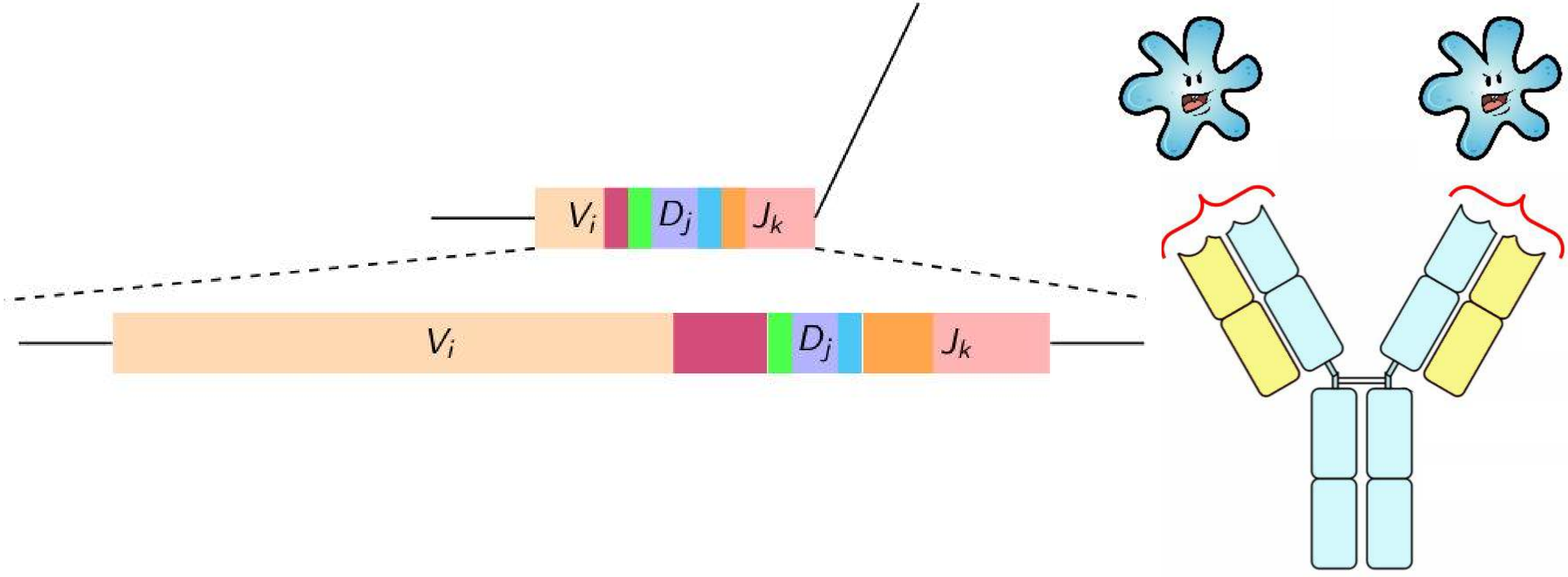
Antibody somatic recombination

Somatic recombination results in unique immunoglobulins genes encoding amino acid sequence of ***antibodies***



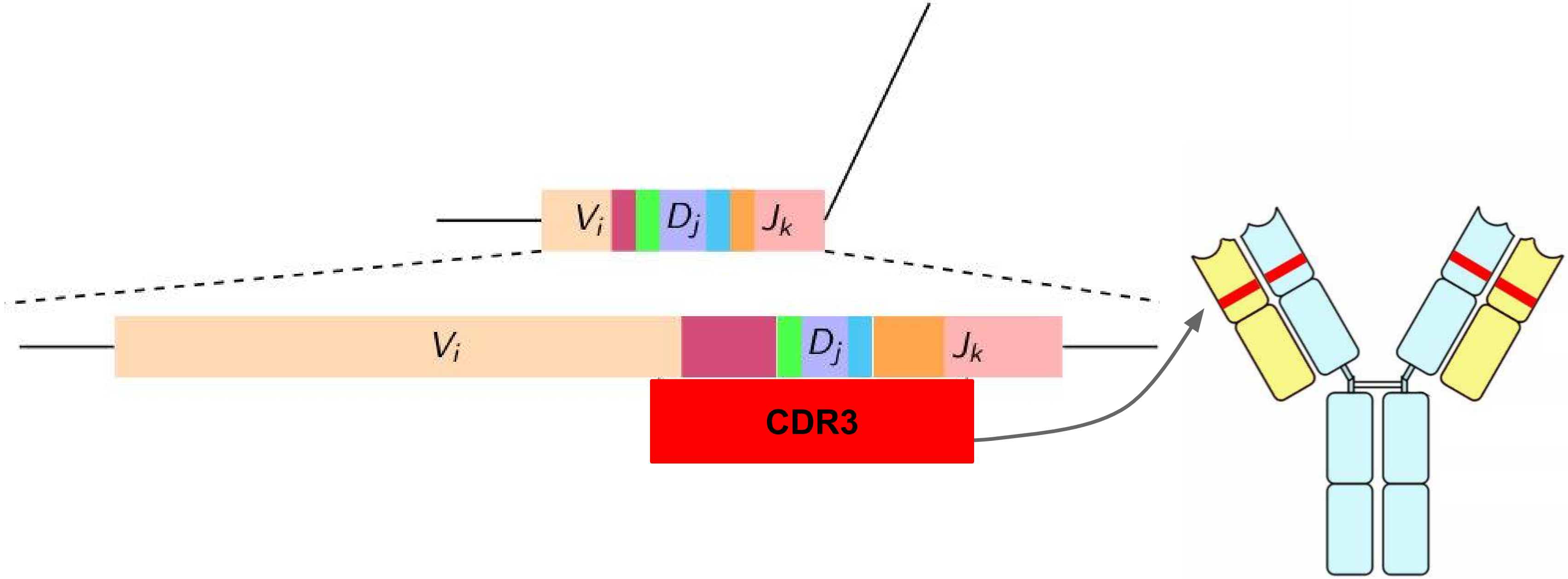
Antibody versus antigen

An antibody recognizes a foreign agent (***antigen***) using its ***antigen-binding site***



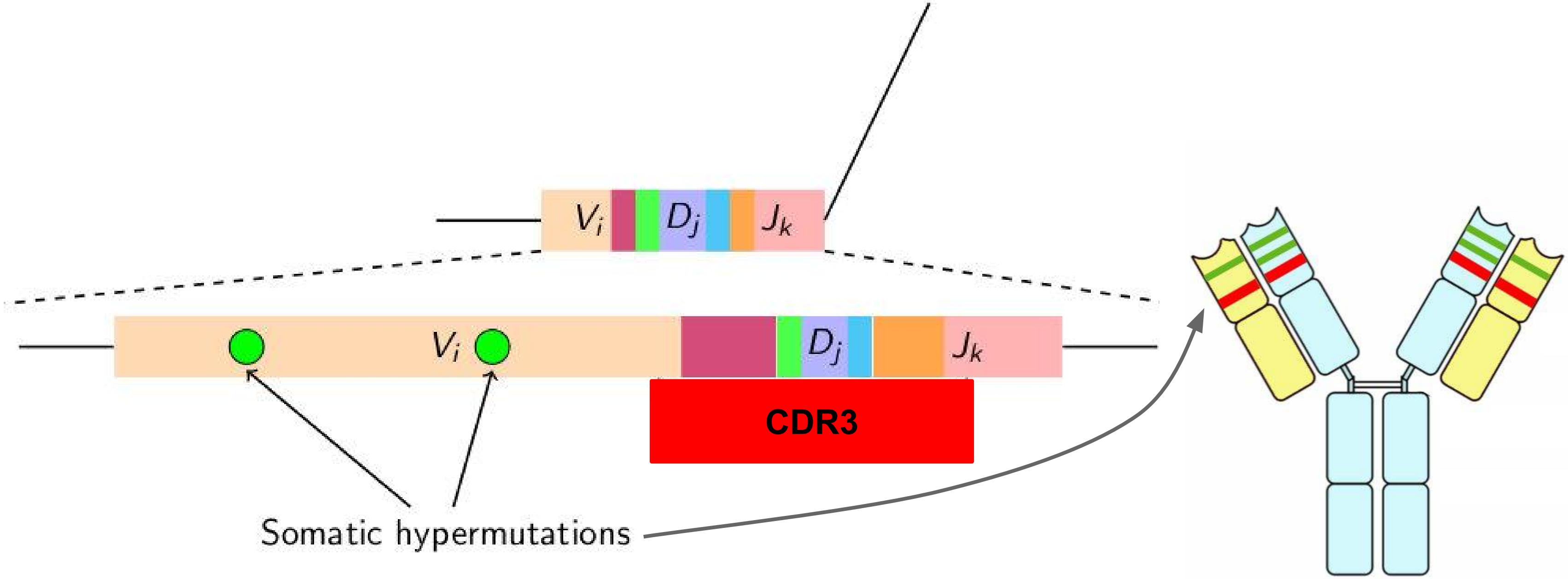
Antigen binding site in antibody

The most diverged part of antigen-binding site is **complementarity determining region 3 (CDR3)**

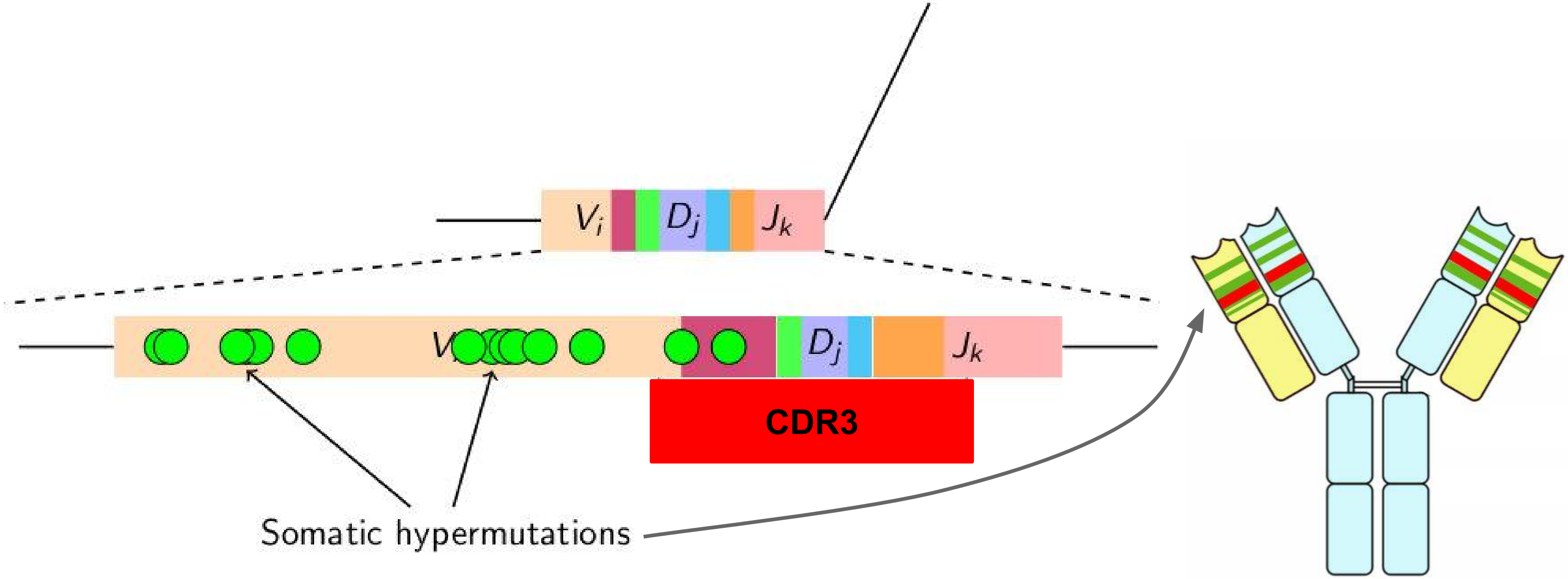


Somatic hypermutations

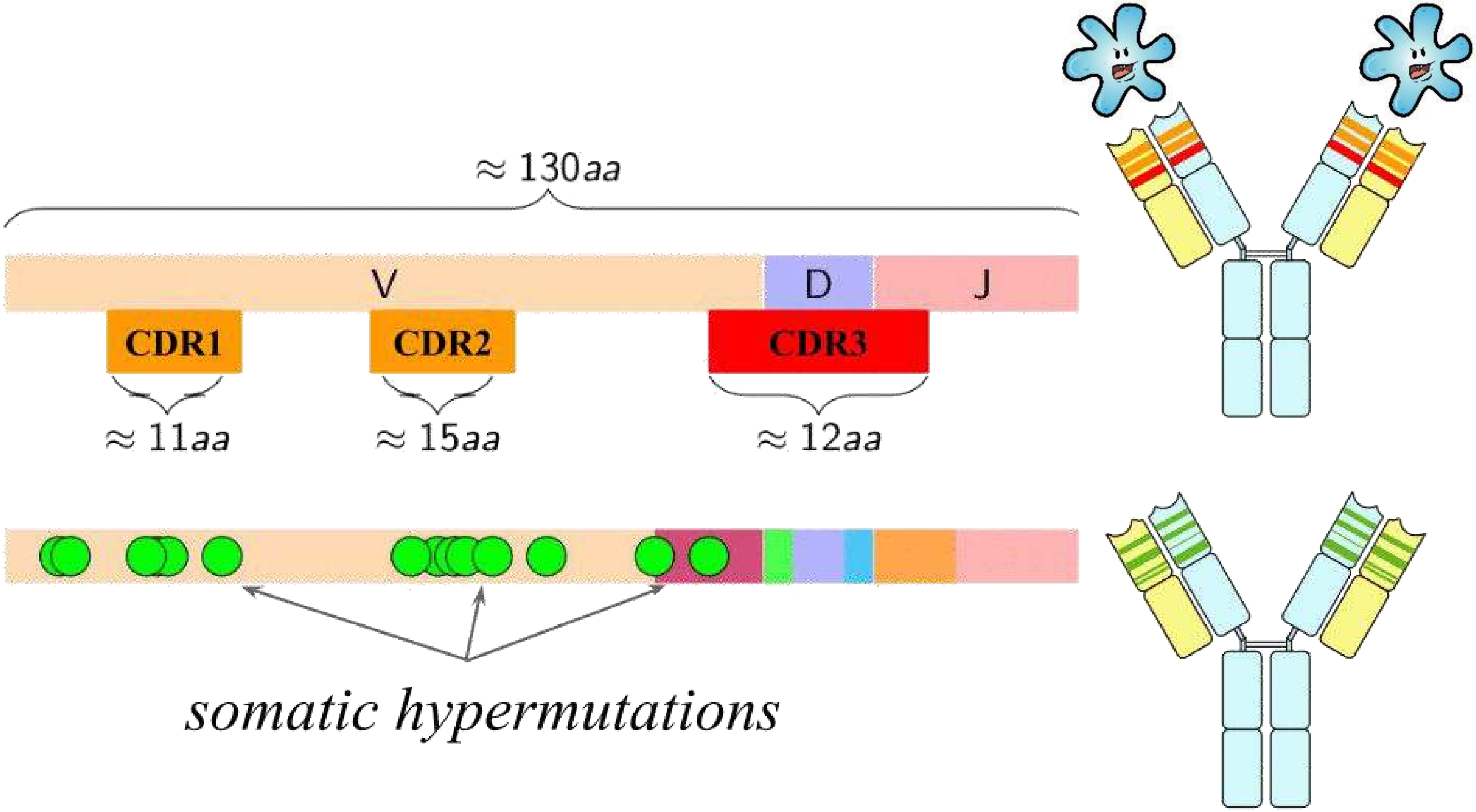
Further optimization of antibody affinity is achieved through **somatic hypermutations**



...many somatic hypermutations

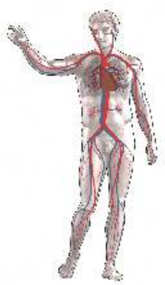


Architecture of antibodies



From biological problems to computational challenges

VDJ classification problem. Given an antibody generated from a *known set* of V, D, and J segments, identify what specific V, D, and J segments generated this antibody



From biological problems to computational challenges

VDJ classification problem. Given an antibody generated from a *known set* of V, D, and J segments, identify what specific V, D, and J segments generated this antibody

From biological problems to computational challenges

VDJ classification problem. Given an antibody generated from a *known set* of V, D, and J segments, identify what specific V, D, and J segments generated this antibody

Important model organisms in immunology with still unknown sets of V, D, and J segments

From biological problems to computational challenges

VDJ classification problem. Given an antibody generated from a *known set* of V, D, and J segments, identify what specific V, D, and J segments generated this antibody

VDJ reconstruction problem. Given a set (millions) of antibodies generated from an *unknown set* of V, D, and J segments, reconstruct these sets

Outline

- Introduction
- **Repertoire construction problem**
- Evolutionary analysis of antibodies
- Analysis of immune response dynamics
- Analysis of paired antibody repertoires & new biological insights from analysis of paired repertoires

Sequencing of antibody repertoire

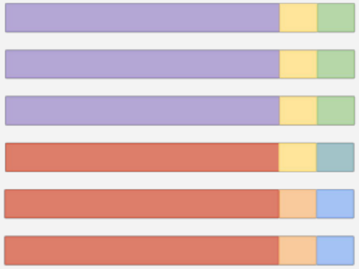
Roche
454
(2005)

low coverage

low accuracy

long reads

VDJ classification



Sequencing of antibody repertoire

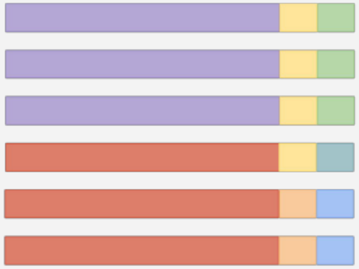
Roche
454
(2005)

low coverage

low accuracy

long reads

VDJ classification



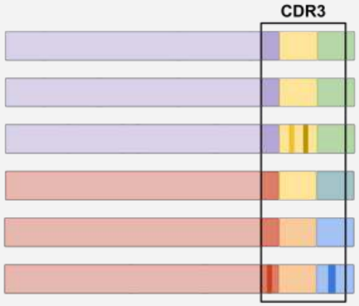
Illumina
HiSeq 2000
(2001)

high coverage

high accuracy

short reads

CDR3 classification



Sequencing of antibody repertoire

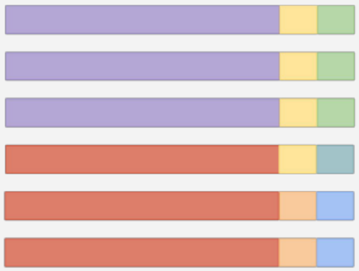
Roche 454
(2005)

low coverage

low accuracy

long reads

VDJ classification



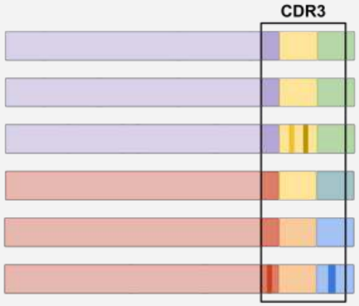
Illumina HiSeq 2000
(2001)

high coverage

high accuracy

short reads

CDR3 classification



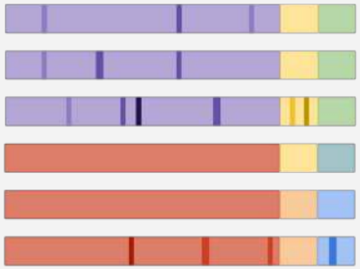
Illumina MiSeq
(2013)

med. coverage

high accuracy

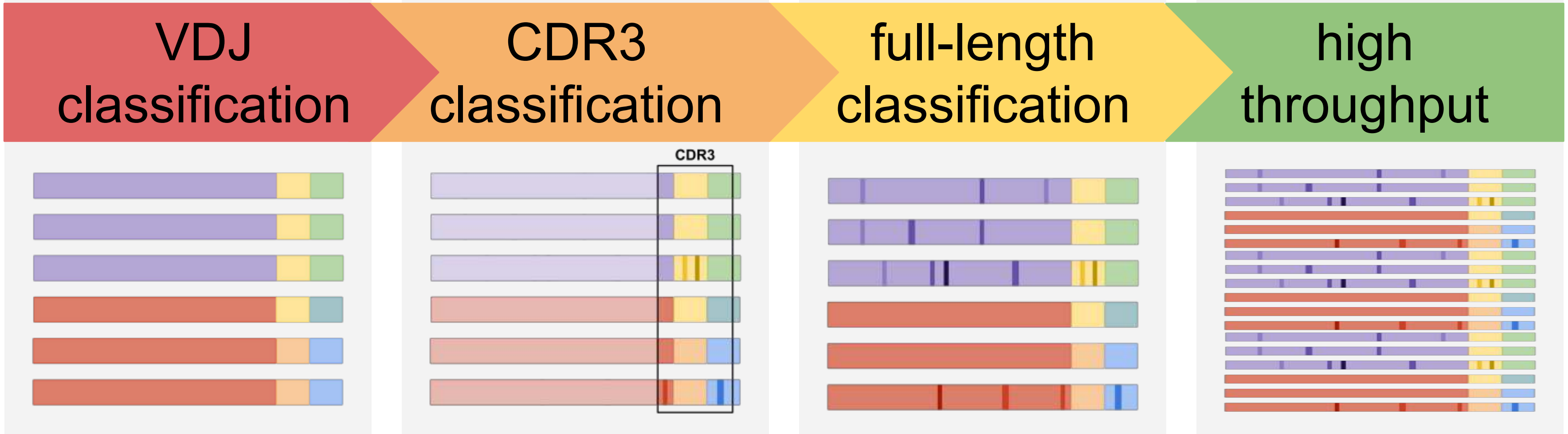
long reads

full-length classification



Sequencing of antibody repertoire

Roche 454 (2005)	Illumina HiSeq 2000 (2001)	Illumina MiSeq (2013)	HiSeq Rapid SBS Kit v2 (2015)
low coverage	high coverage	med. coverage	high coverage
low accuracy	high accuracy	high accuracy	high accuracy
long reads	short reads	long reads	long reads



VDJ classification

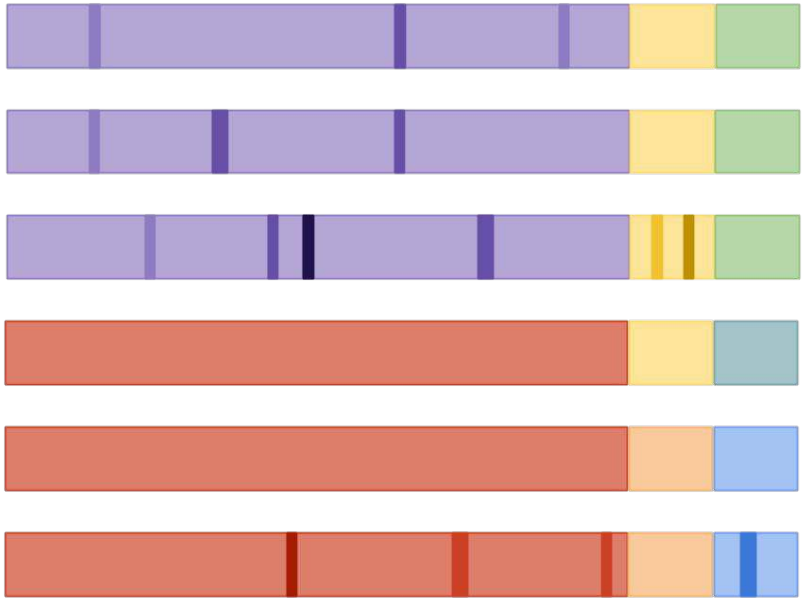
CDR3 classification

full-length classification

high throughput

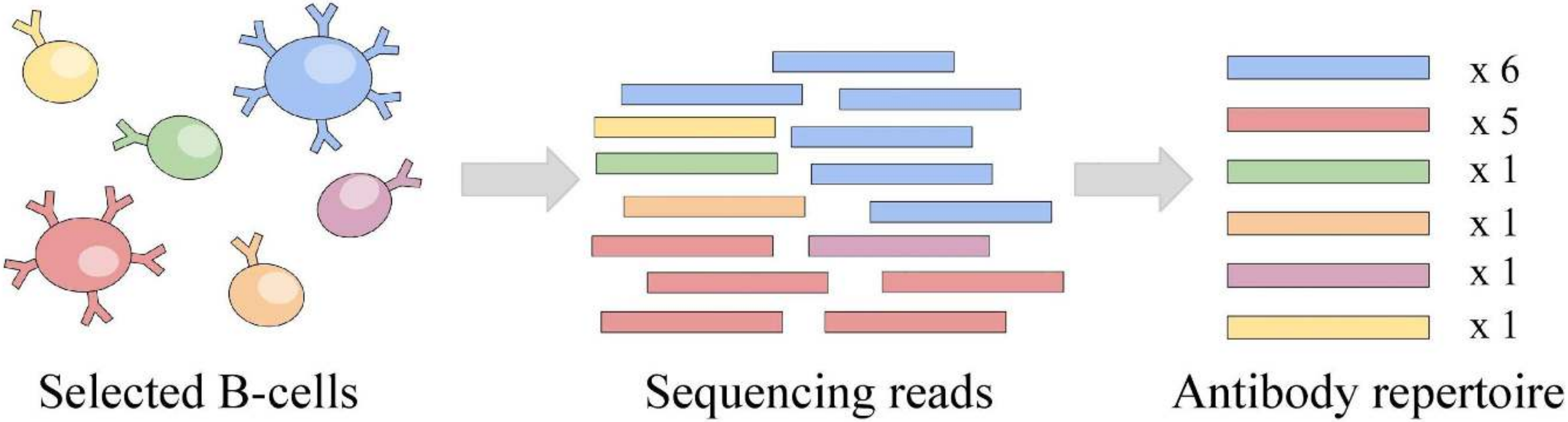
Full-length antibody classification (repertoire construction)

In contrast to well-studied **VDJ** and **CDR3 classification**, **full-length antibody classification** takes into account the entire variable region of antibody



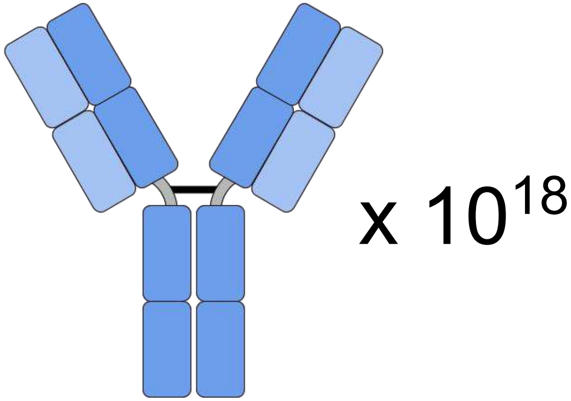
MiGEC: Shugay et al., *Nat Methods*, 2014
MiXCR: Bolotin et al., *Nat Methods*, 2015
IMSEQ: Kuchenbecker et al., *Bioinformatics*, 2015
IgRepertoireConstructor: Safonova et al., *Bioinformatics*, 2015

Repertoire construction problem

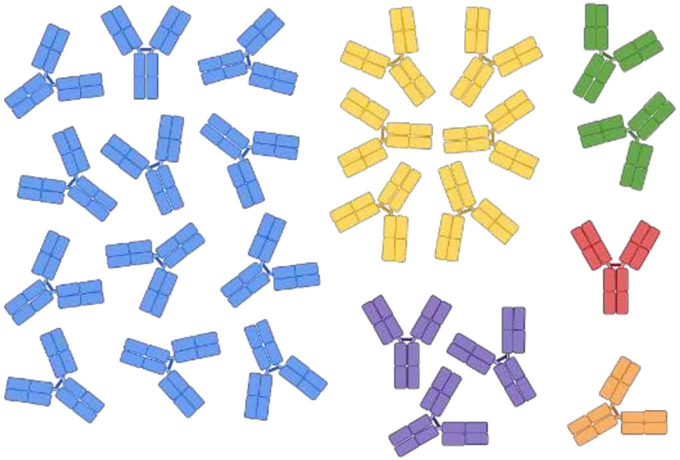


- Giant read clustering problem
- Giant error correction problem

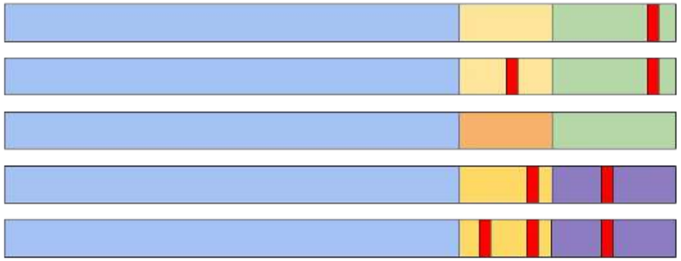
What makes this clustering problem difficult?



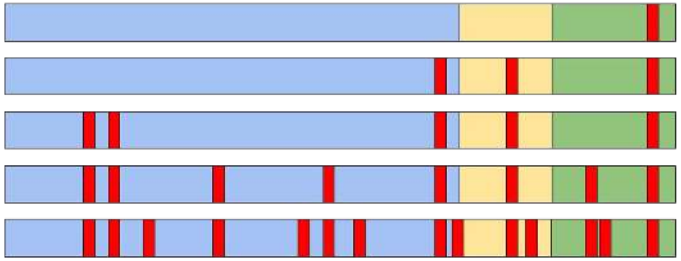
Huge repertoire size



Uneven distribution of abundances



High repetitiveness



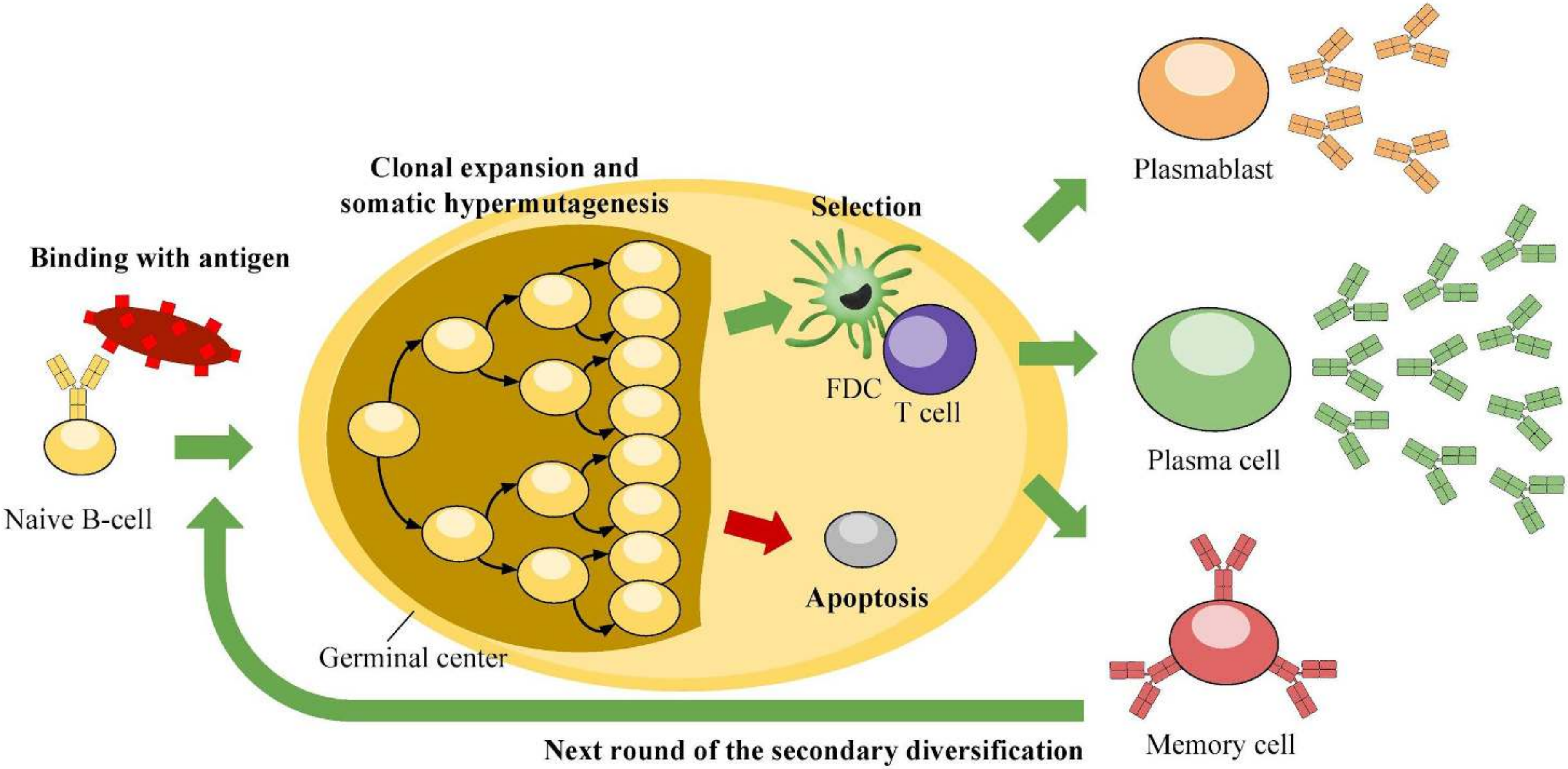
High mutation rate

- Global coverage threshold cannot be used for error correction
- Sequencing errors often look like natural variations

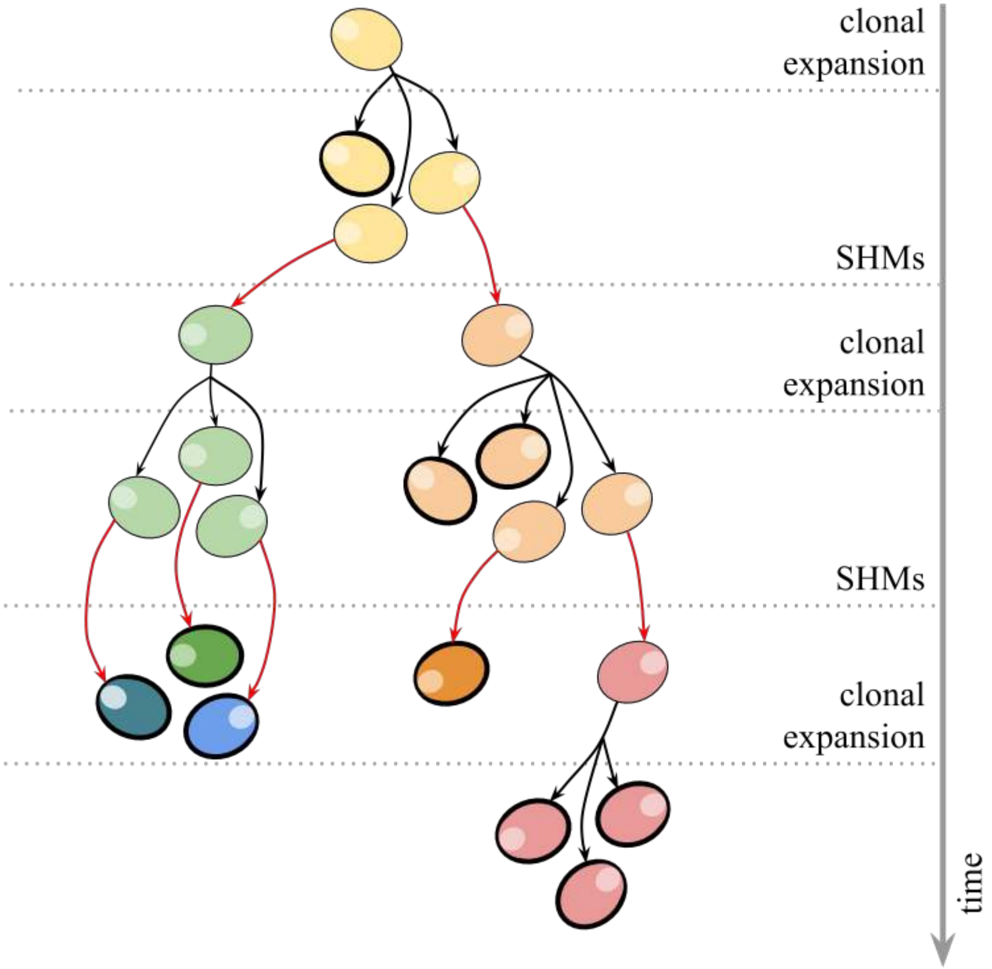
Outline

- Introduction
- Repertoire construction problem
- **Evolutionary analysis of antibodies**
- Analysis of immune response dynamics
- Analysis of paired antibody repertoires & new biological insights from analysis of paired repertoires

Secondary diversification of antibodies

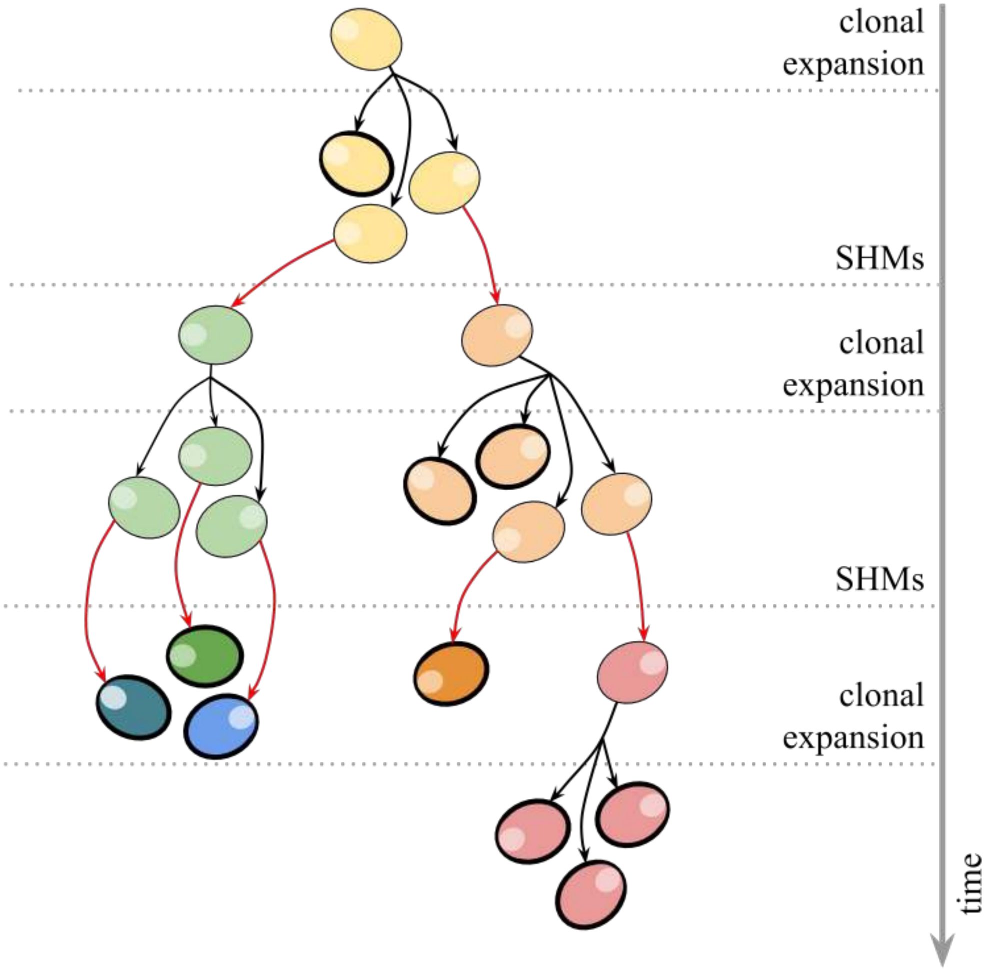


Clonal analysis of antibody repertoire

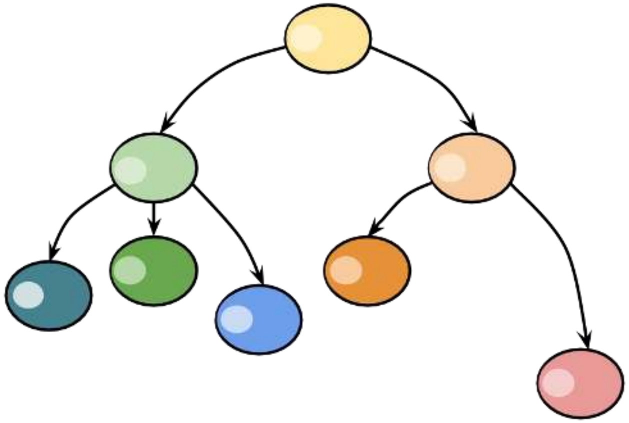


- **B-cell lineages** reflect evolutionary development of antibodies

Clonal analysis of antibody repertoire

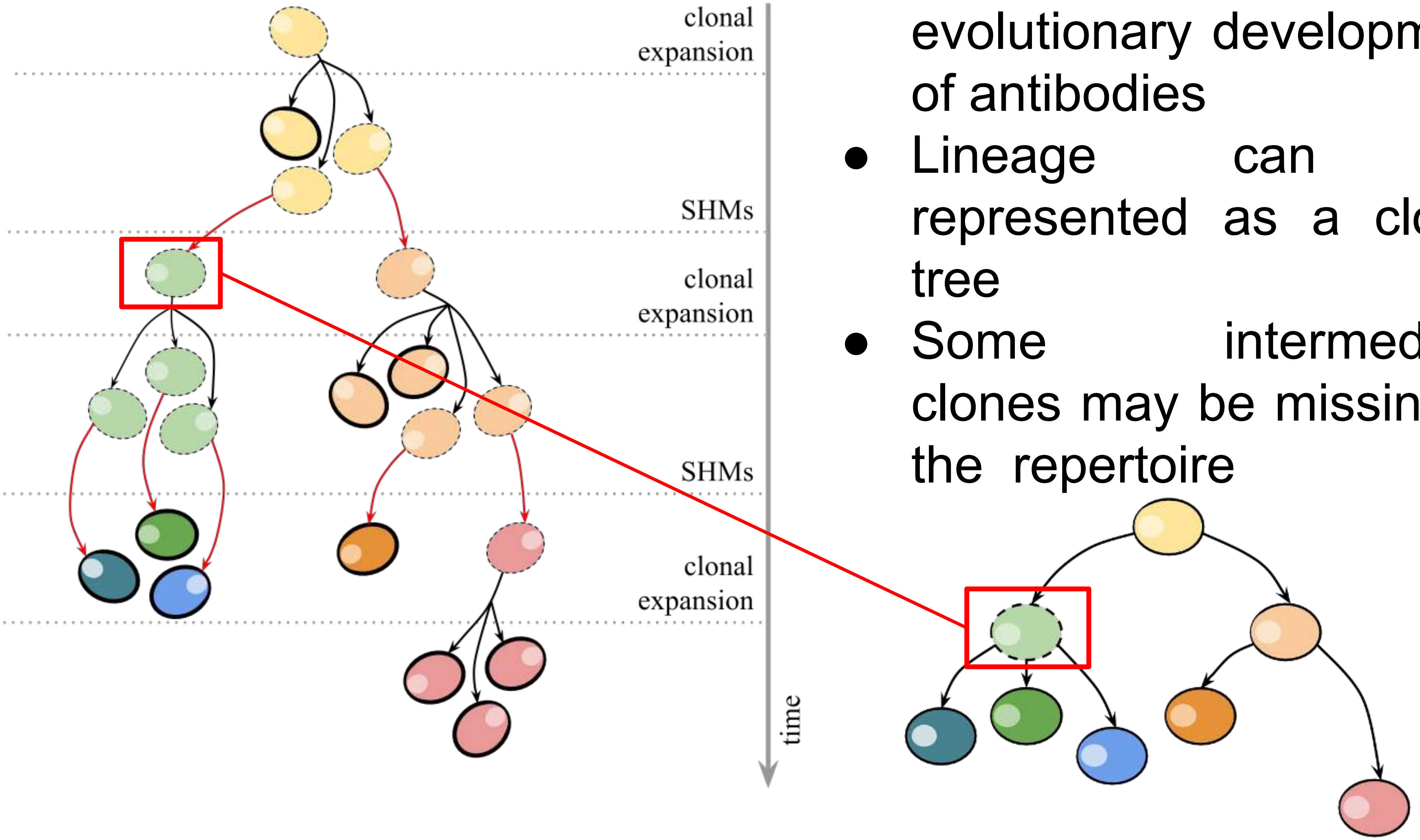


- B-cell lineages reflect evolutionary development of antibodies
- Lineage can be represented as a **clonal tree**

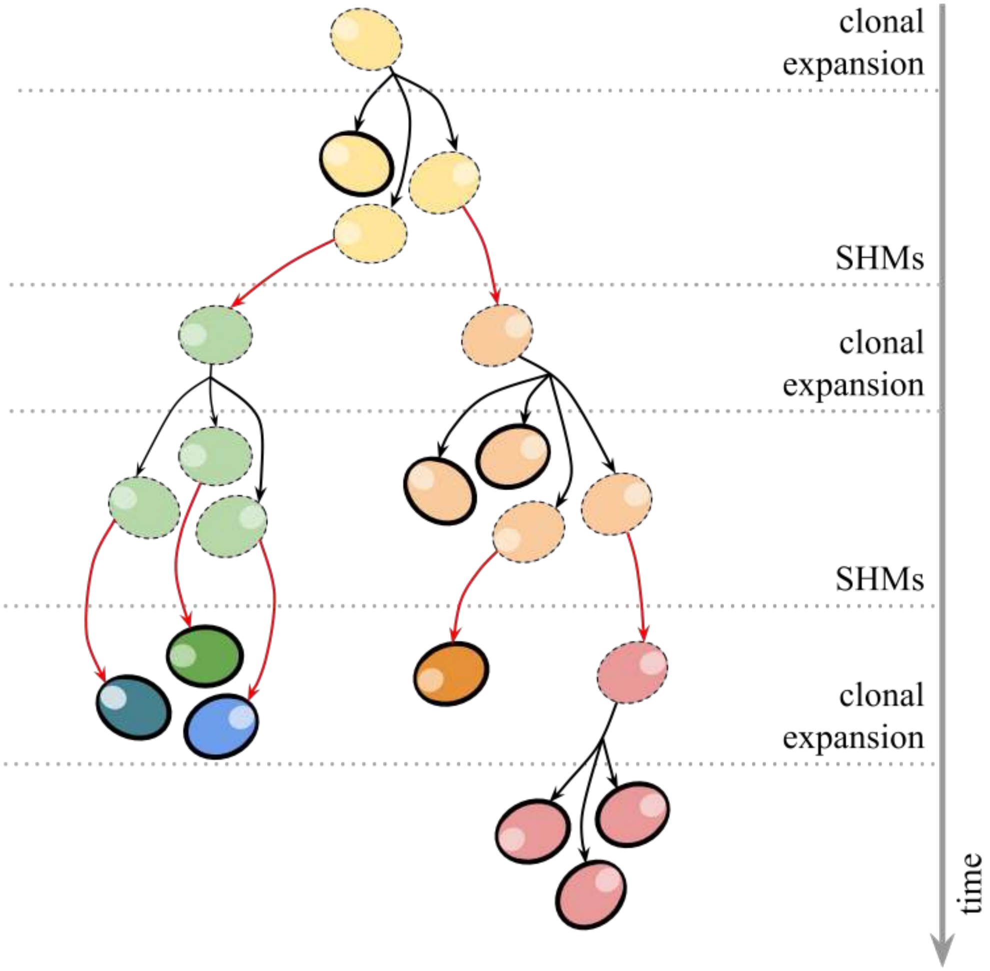


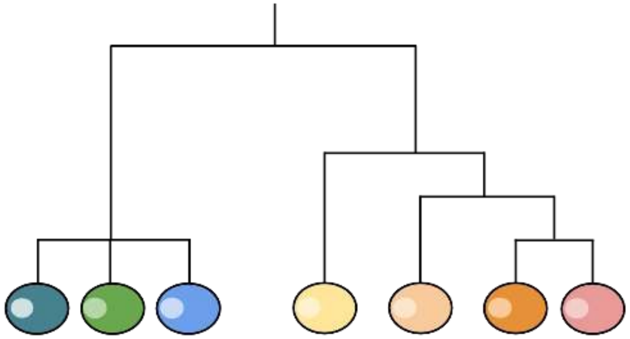
Clonal analysis of antibody repertoire

- B-cell lineages reflect evolutionary development of antibodies
- Lineage can be represented as a clonal tree
- Some intermediate clones may be missing in the repertoire

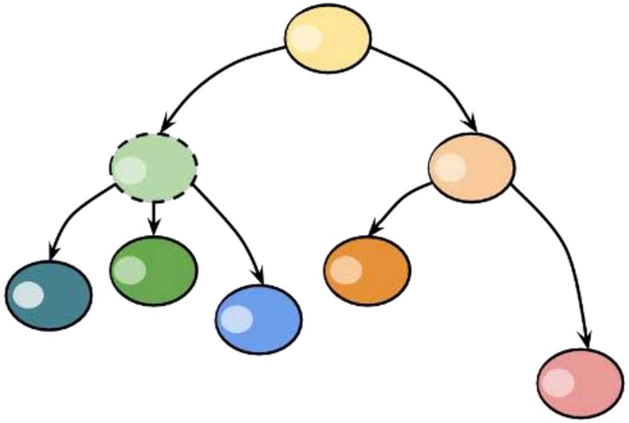


Clonal analysis of antibody repertoire

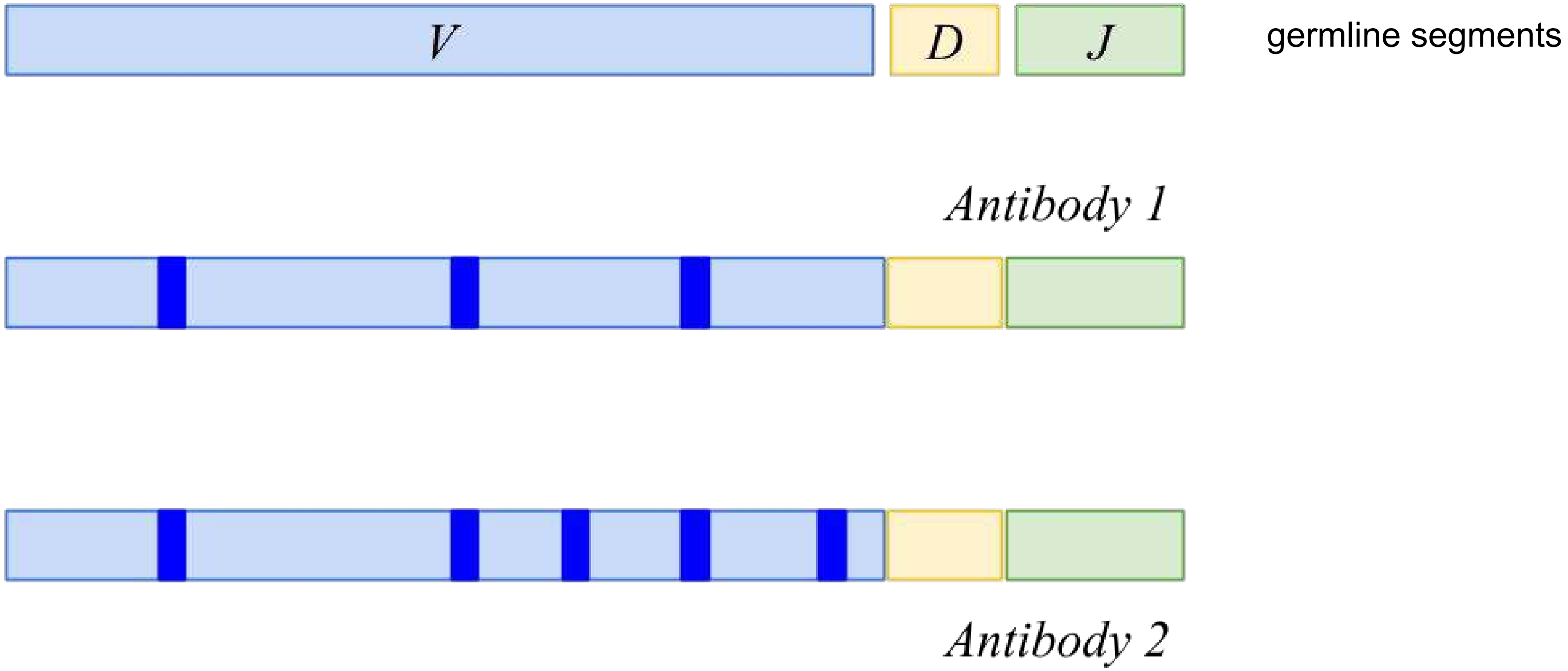




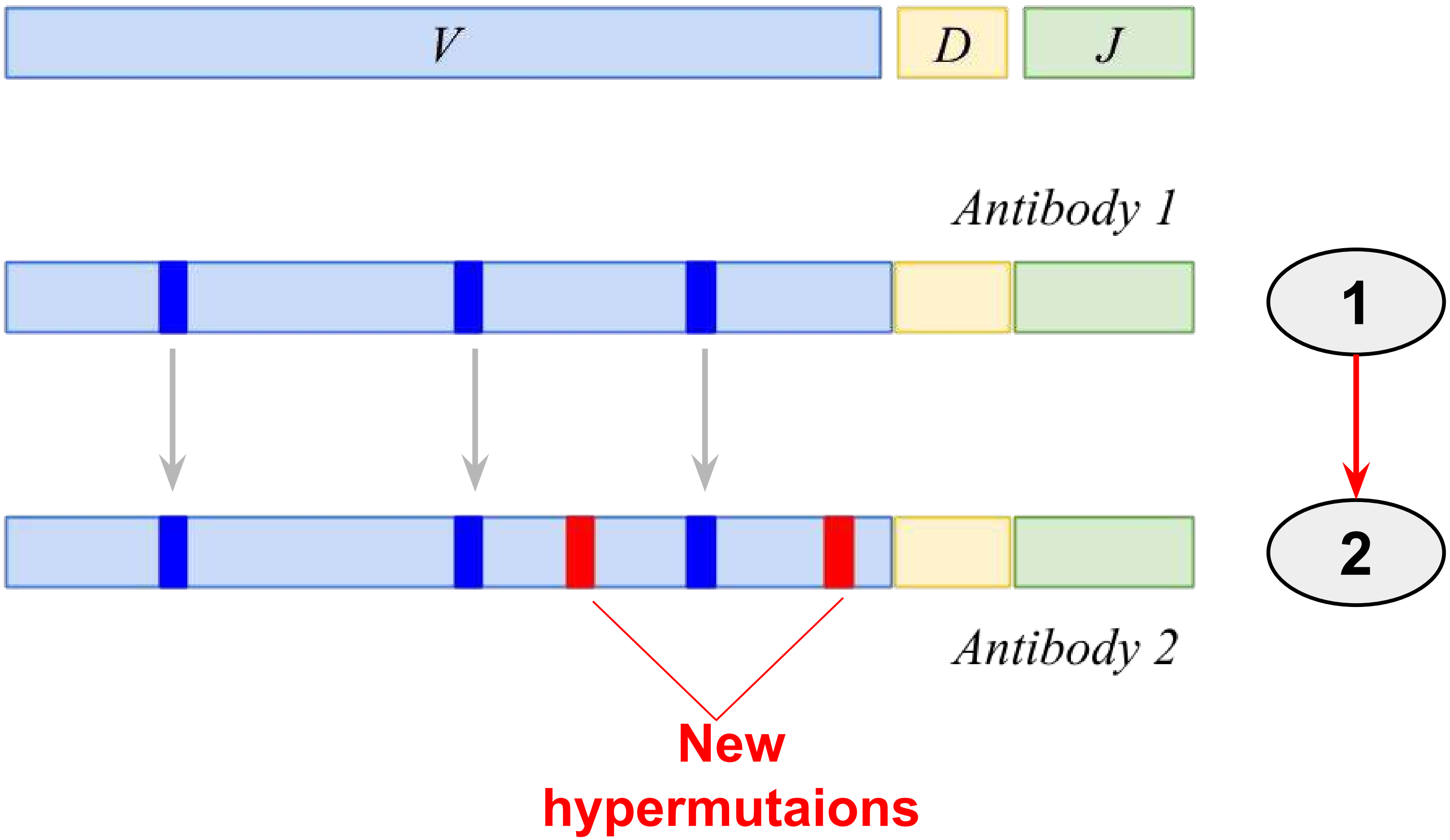
Standard phylogenetic algorithms assume that all species are represented by leaves and should be adapted for clonal trees



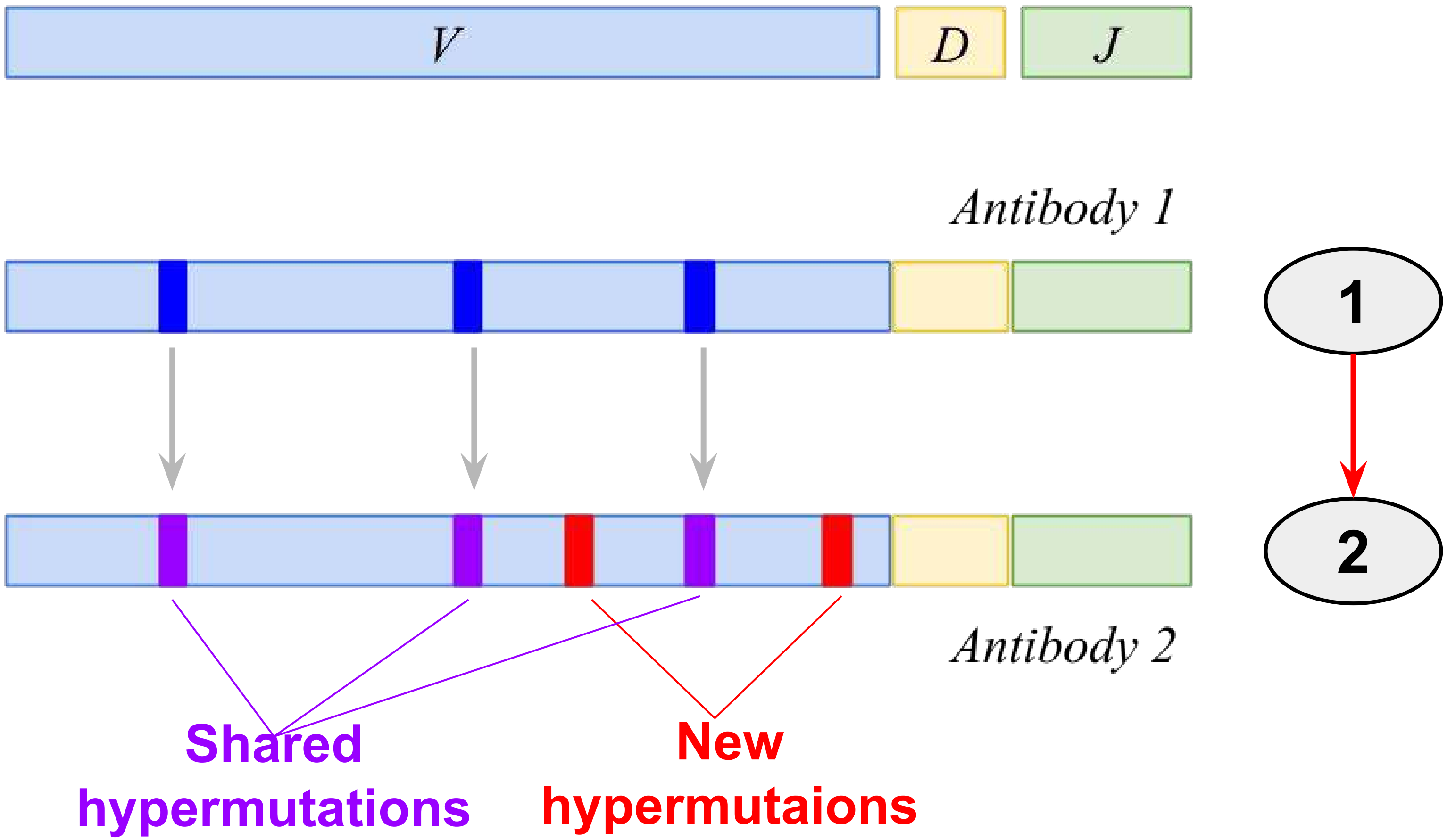
Who is the ancestor here?



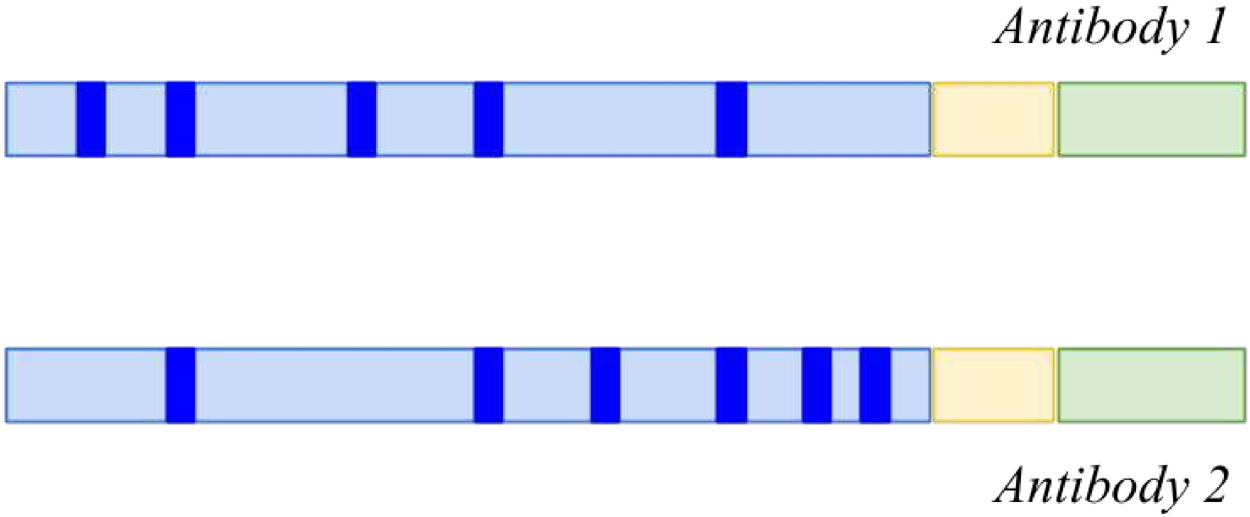
Who is the ancestor here?



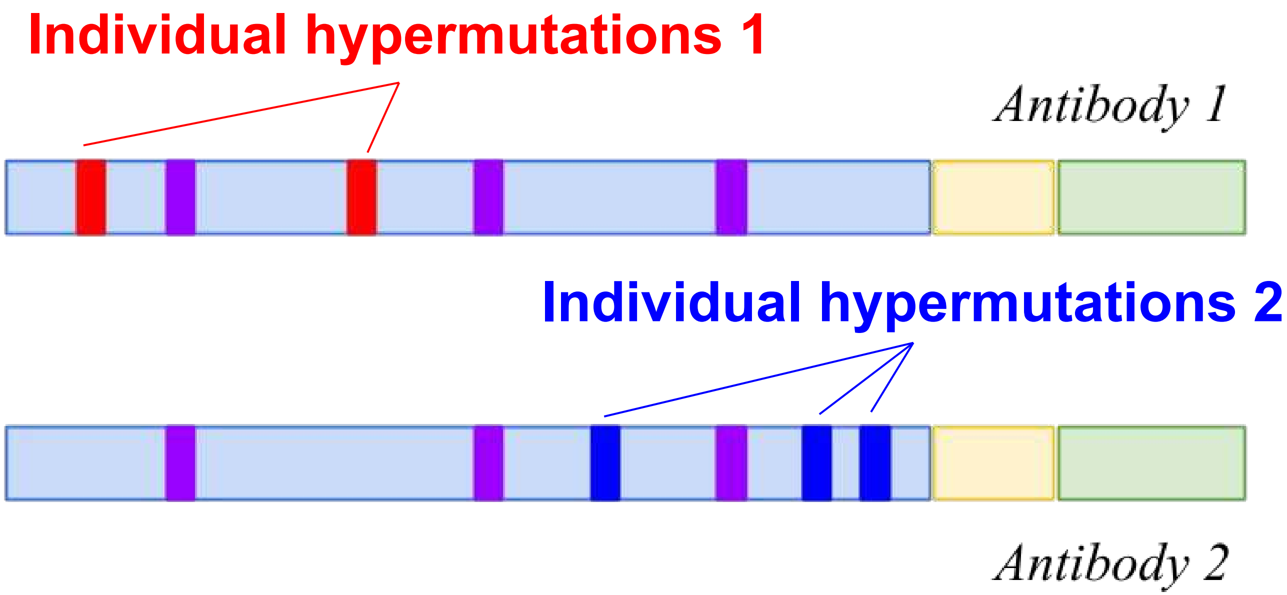
Who is the ancestor here?



Another example: who is the ancestor here?



Another example: who is the ancestor here?



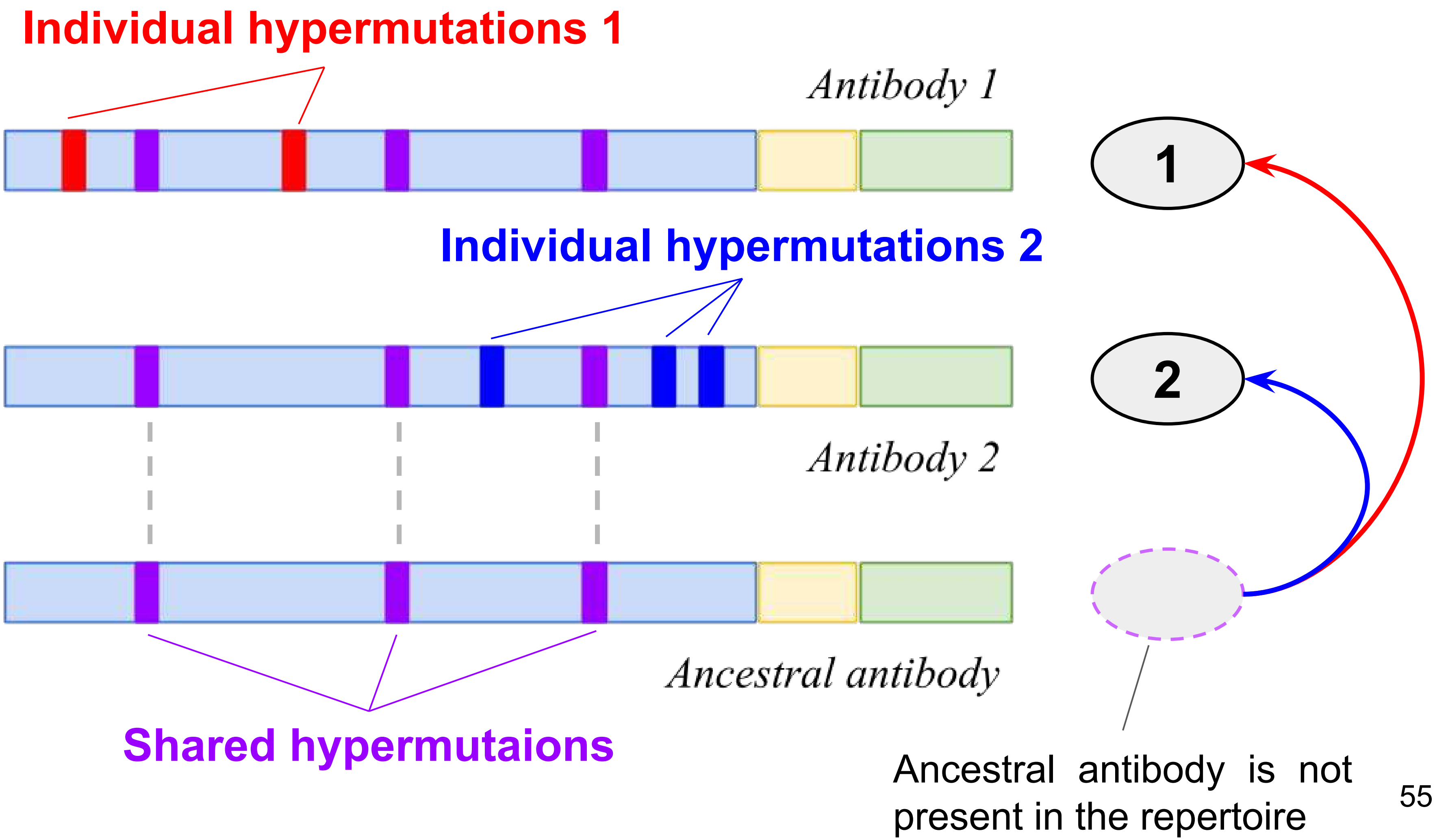
Individual hypermutations 1

Antibody 1

Individual hypermutations 2

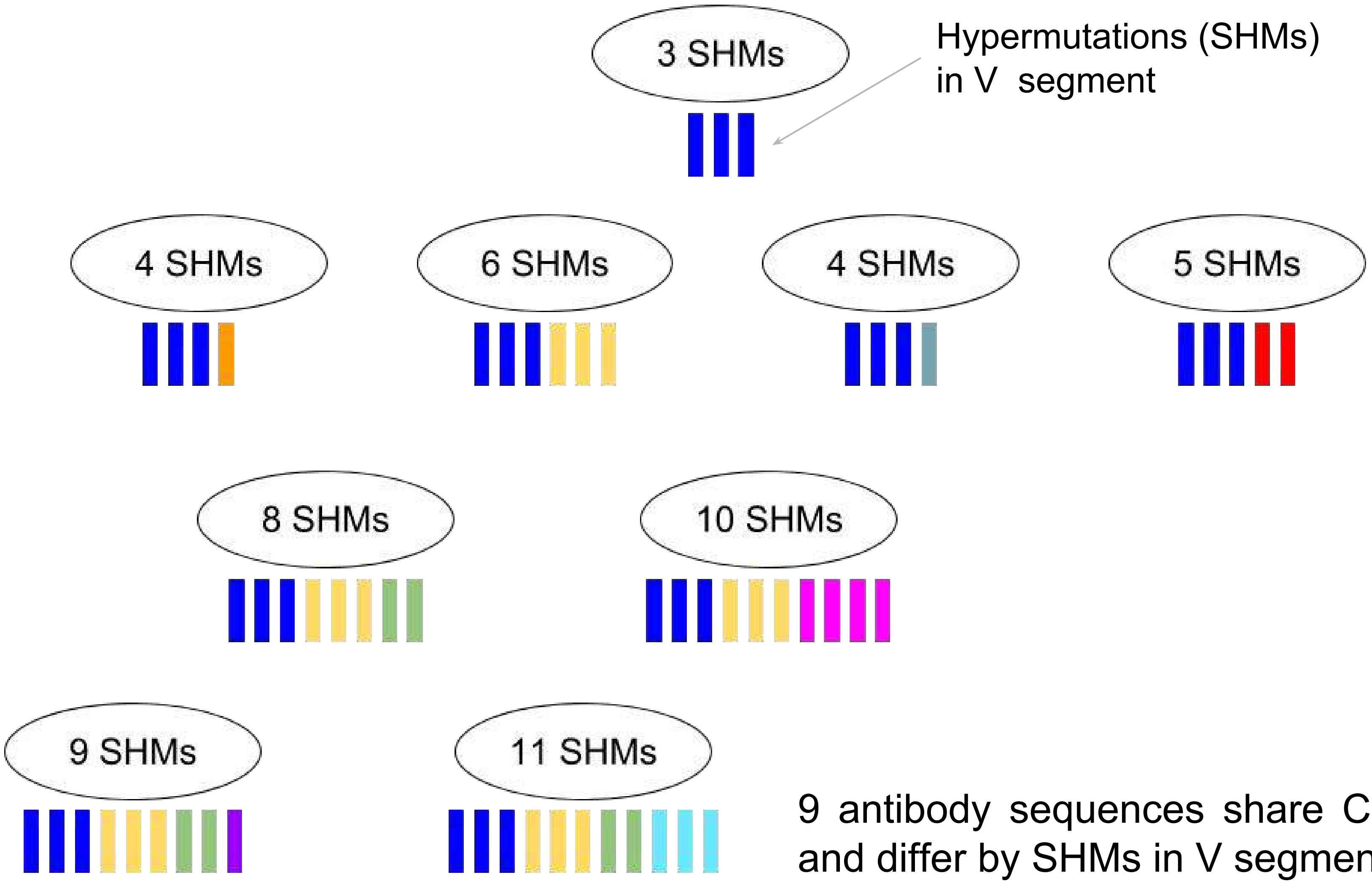
Antibody 2

Ancestral antibody may be missing…



Ancestral antibody is not present in the repertoire

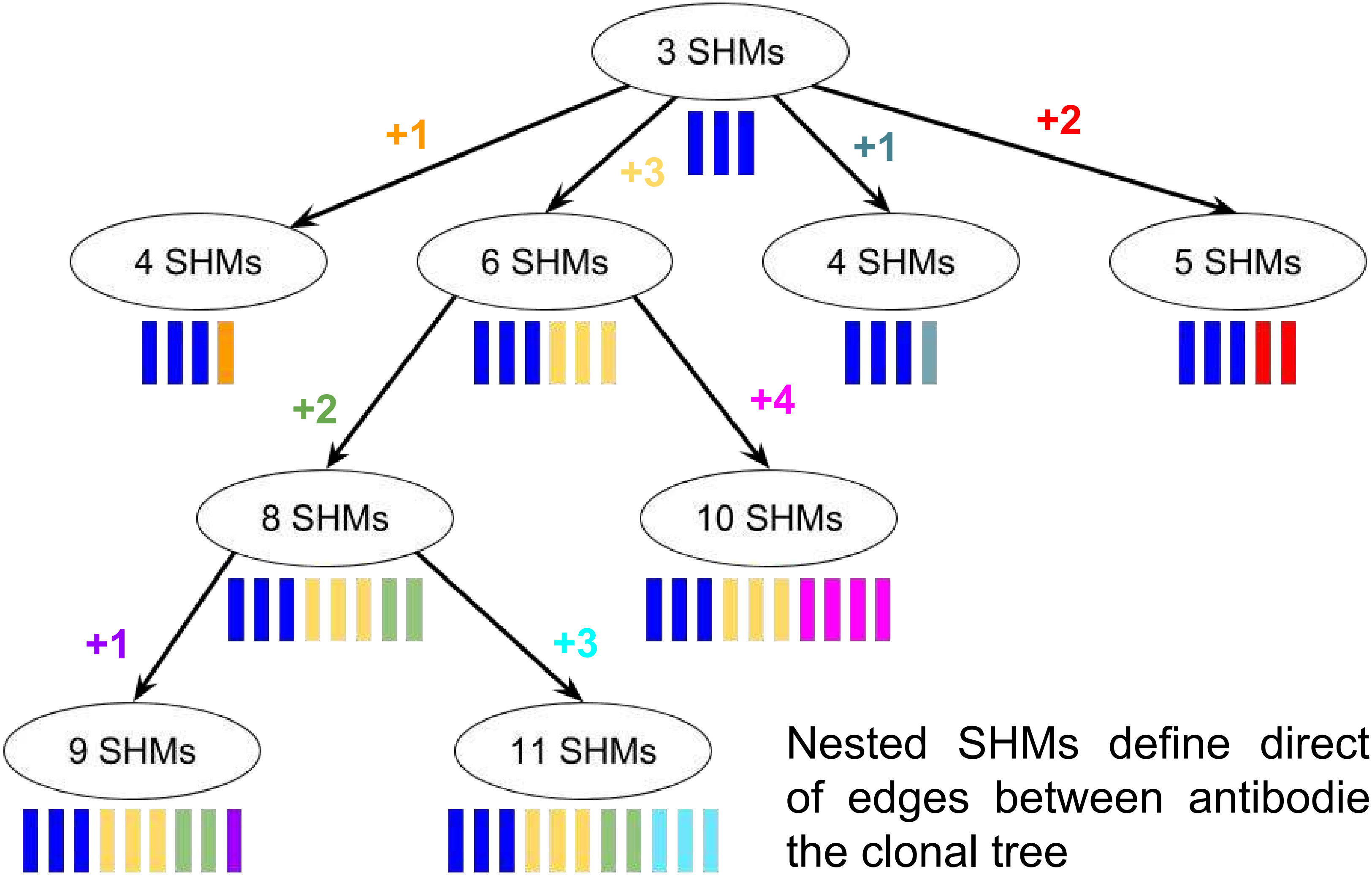
What is the evolutionary tree?



Hypermutations (SHMs) in V segment

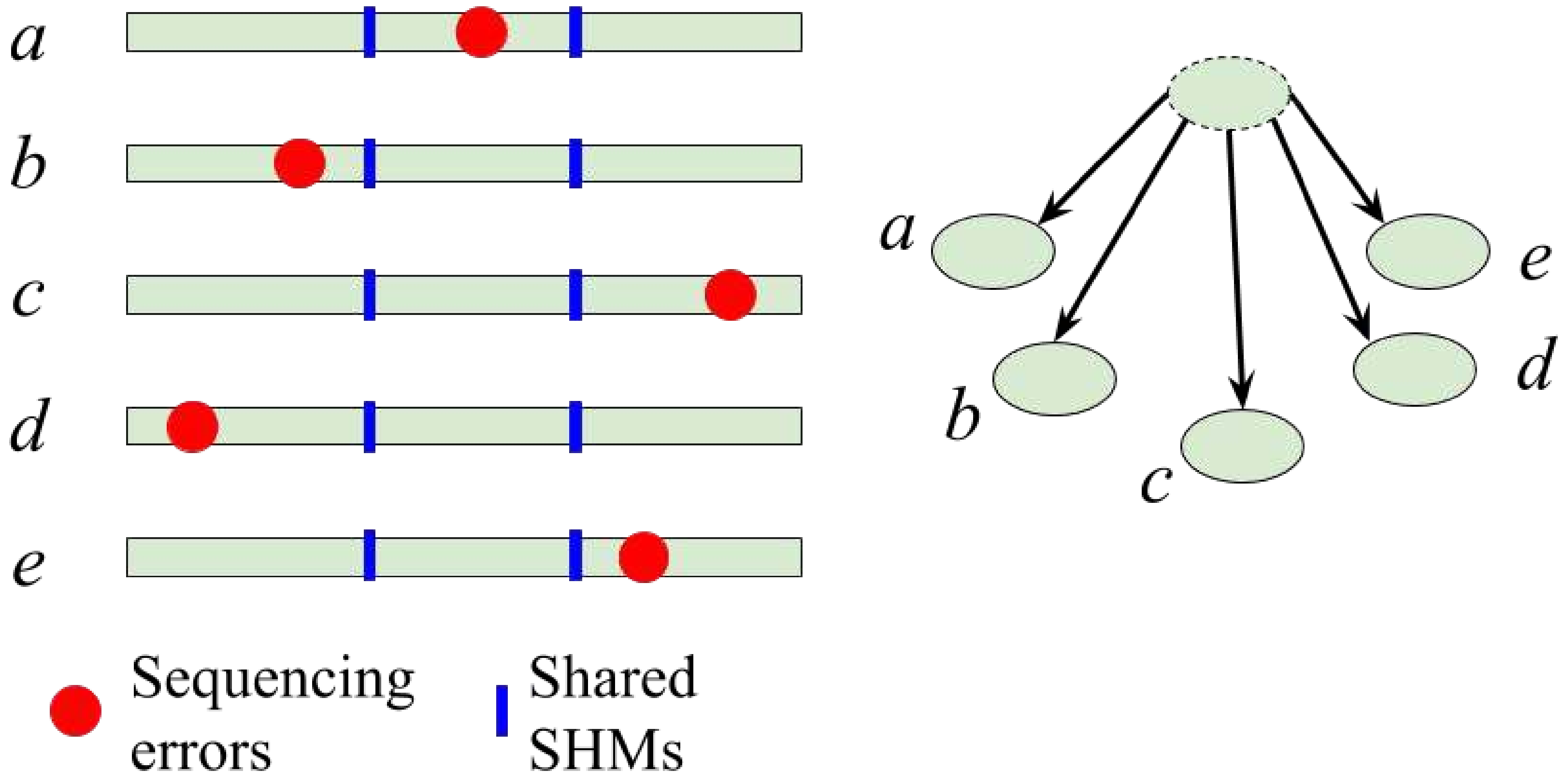
9 antibody sequences share CDR3 and differ by SHMs in V segments

Any tree reconstruction approach will work

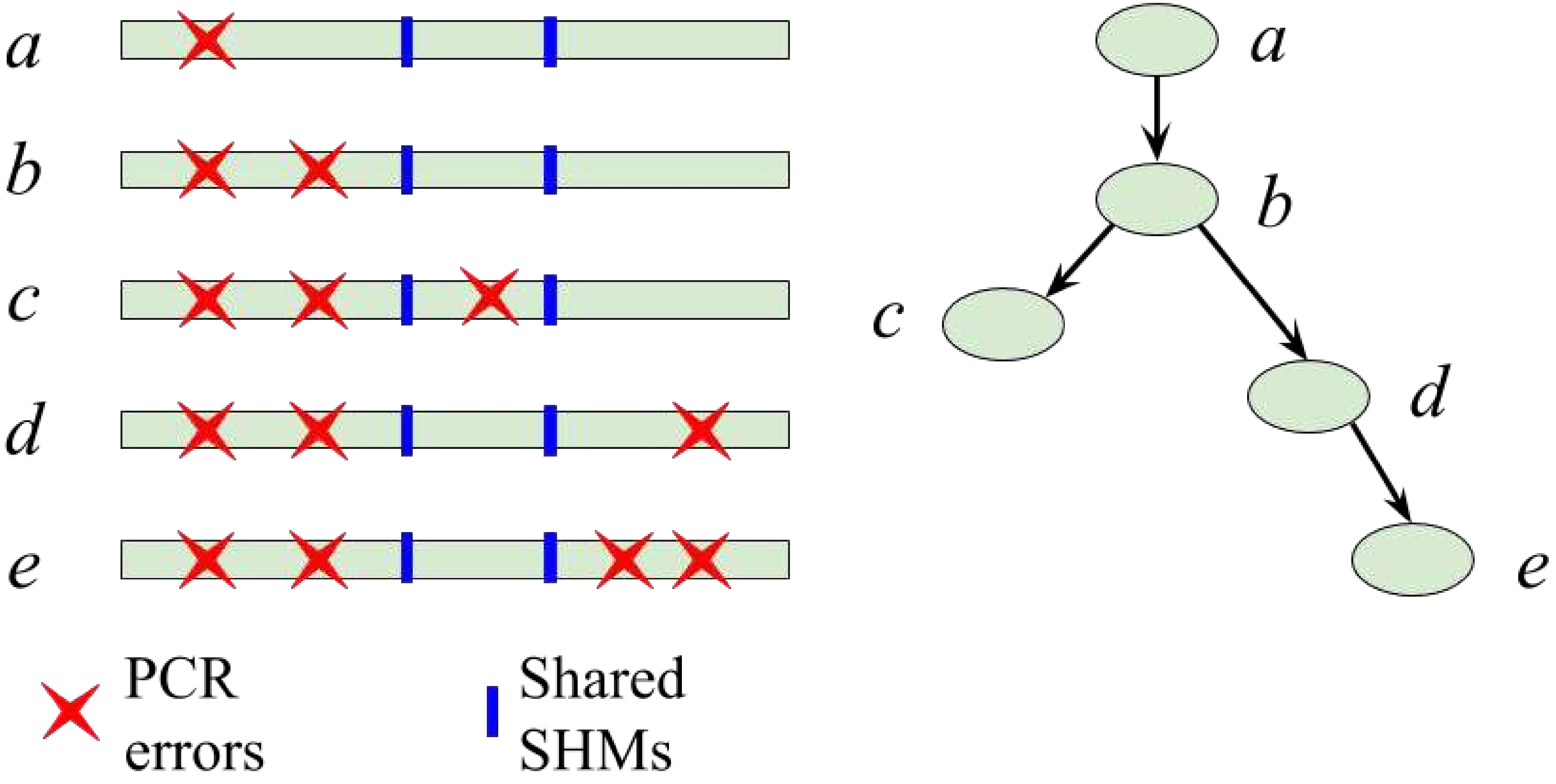


Nested SHMs define directions of edges between antibodies in the clonal tree

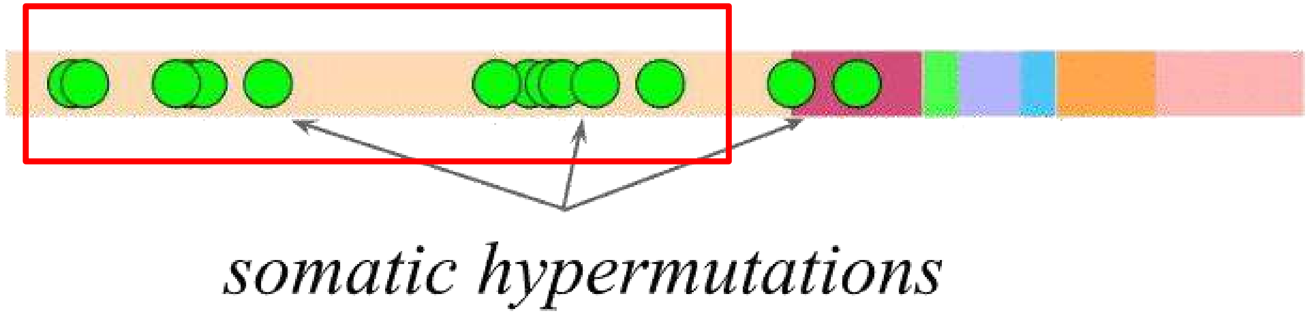
Repertoire construction step is very important for clonal analysis!



Repertoire construction step is very important for clonal analysis!

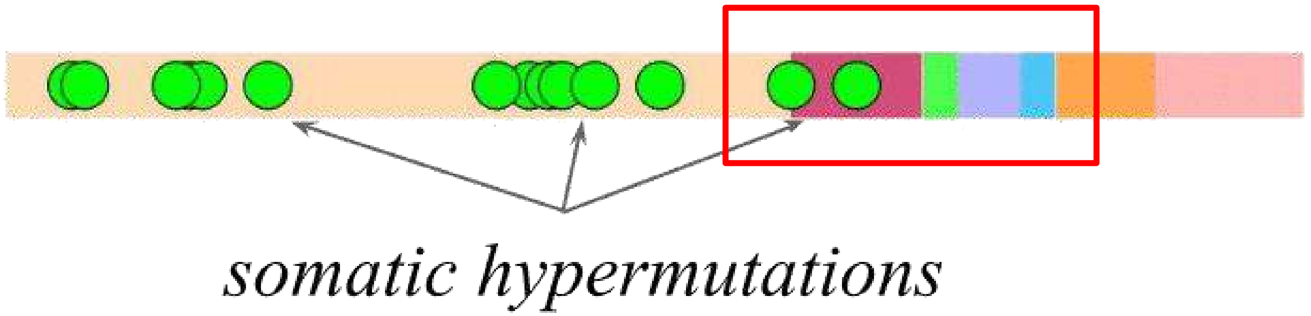


SHMs in V segments are easy to find



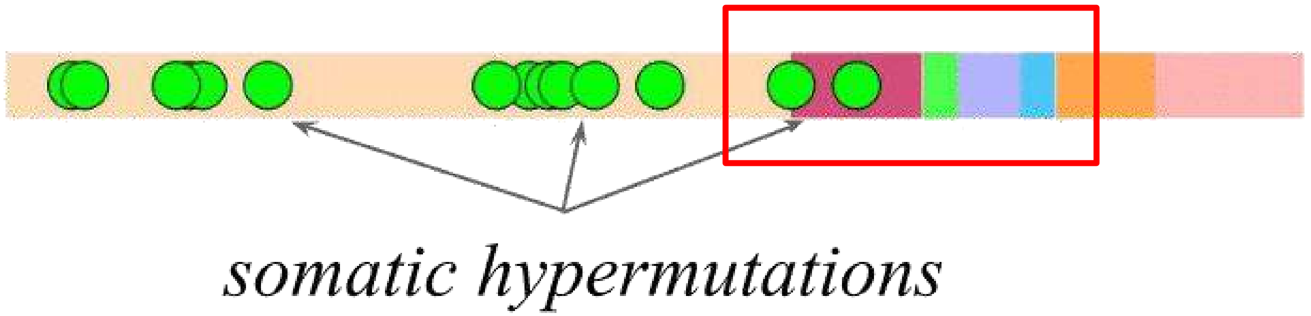
- One can easily identify mutations in the V segment using alignment against the **template** (germline V segment)

SHMs in CDR3 are difficult to identify



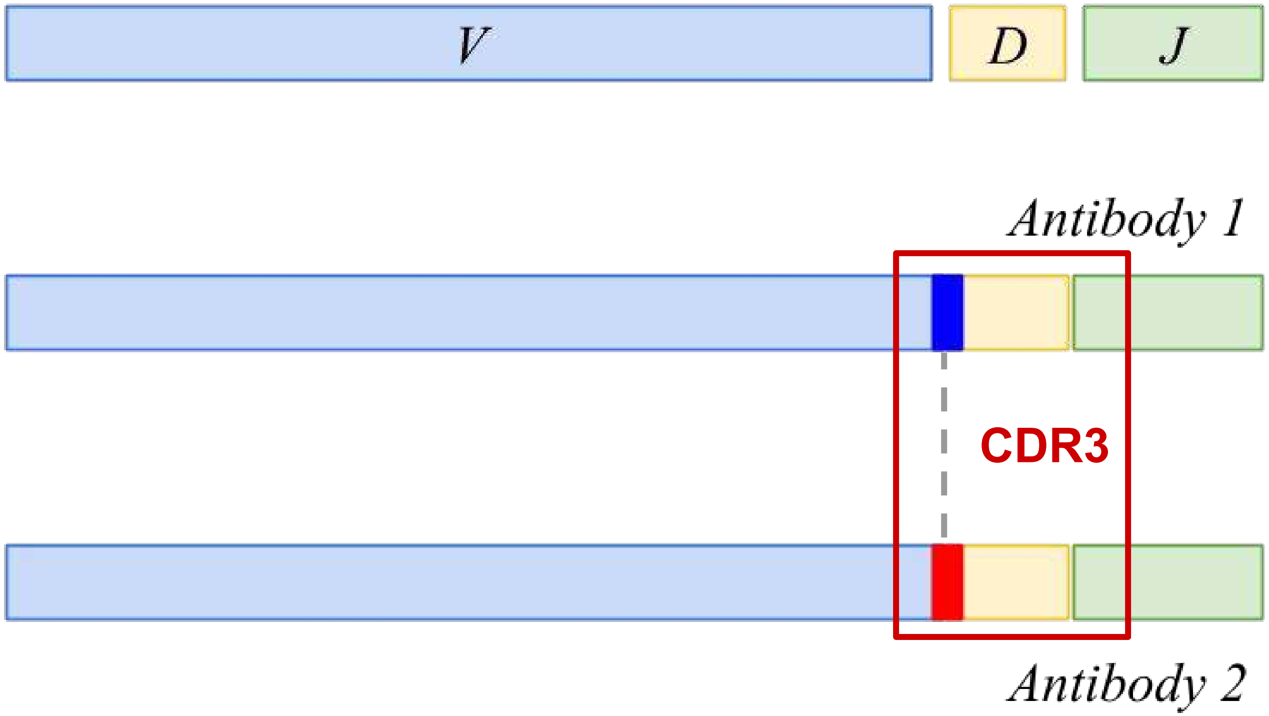
- One can easily identify mutations in the V segment using alignment against the **template** (germline V segment)
- **But there is no template for CDR3!**

SHMs in CDR3 are difficult to identify

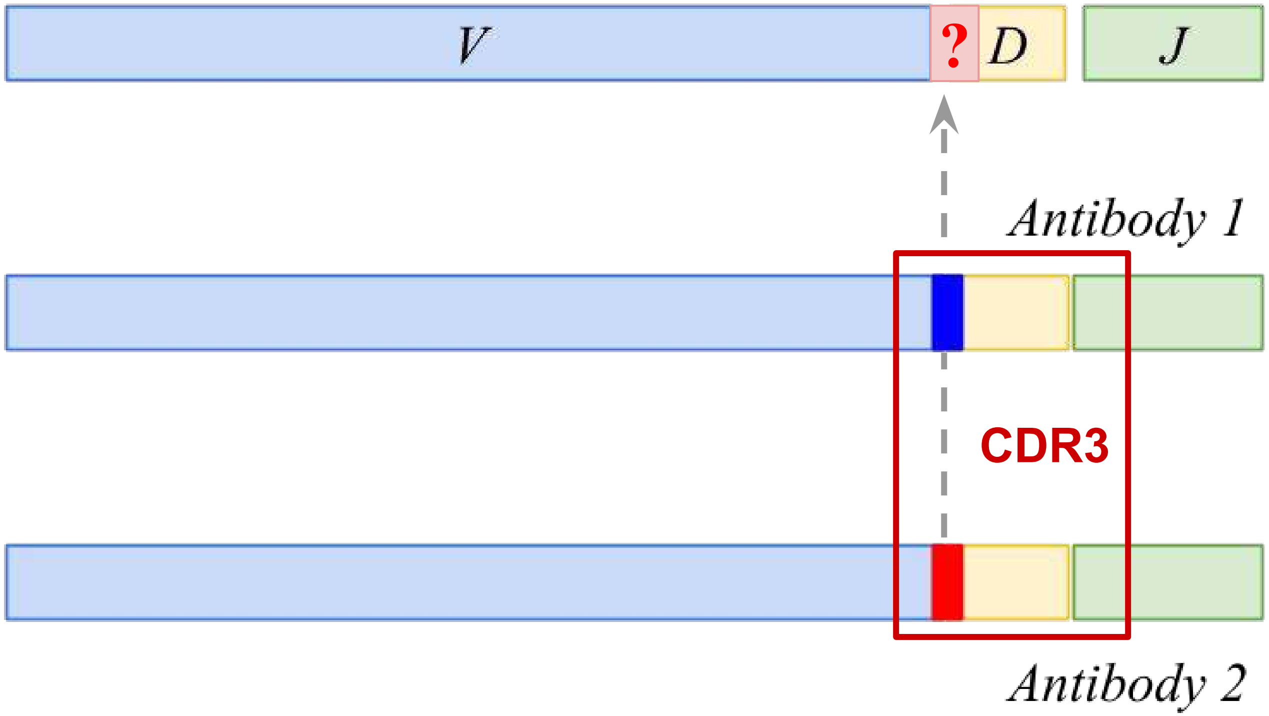


- One can easily identify mutations in the V segment using alignment against the **template** (germline V segment)
- **But there is no template for CDR3!**
    - **deletions** in gene segments
    - non-genomic VD and DJ **insertions**
    - addition of **palindromes**

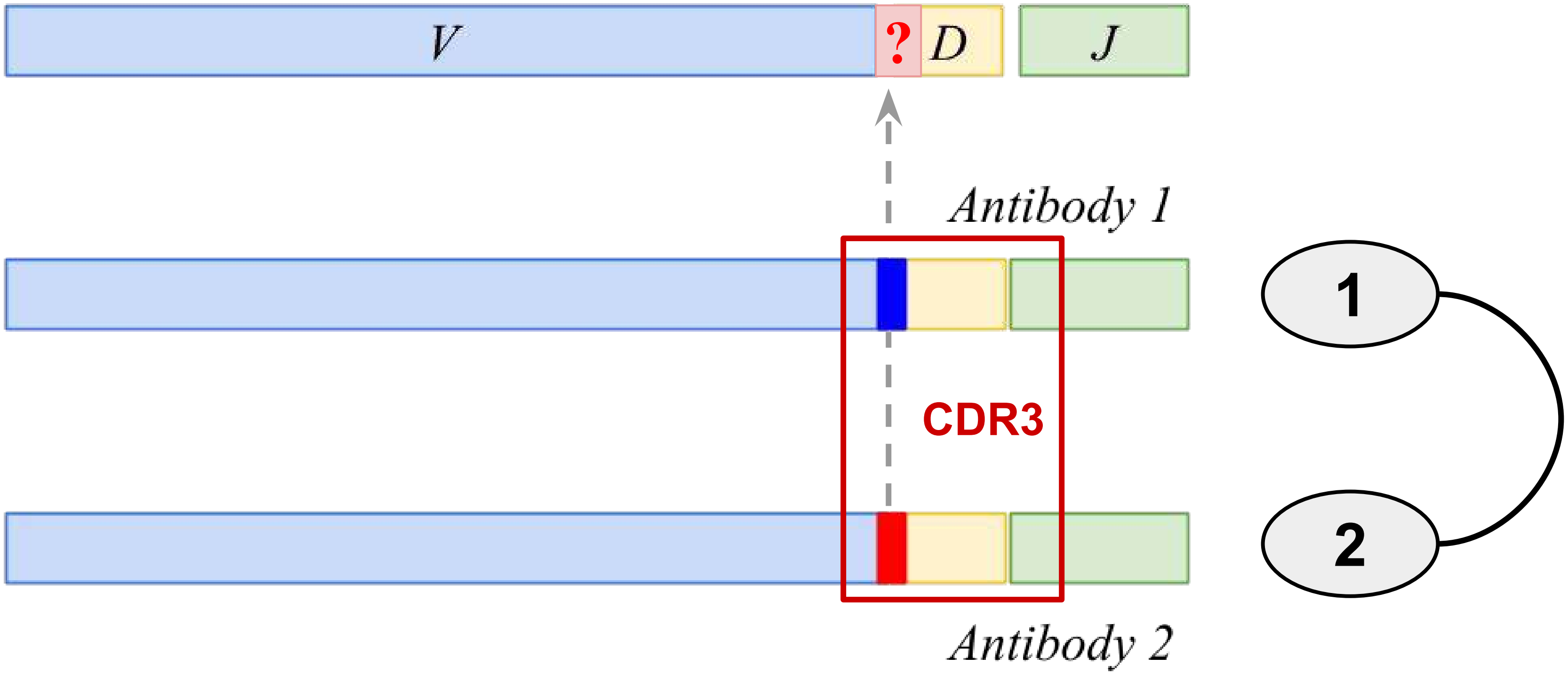
A more complex case: who is the ancestor?



A more complex case: who is the ancestor?



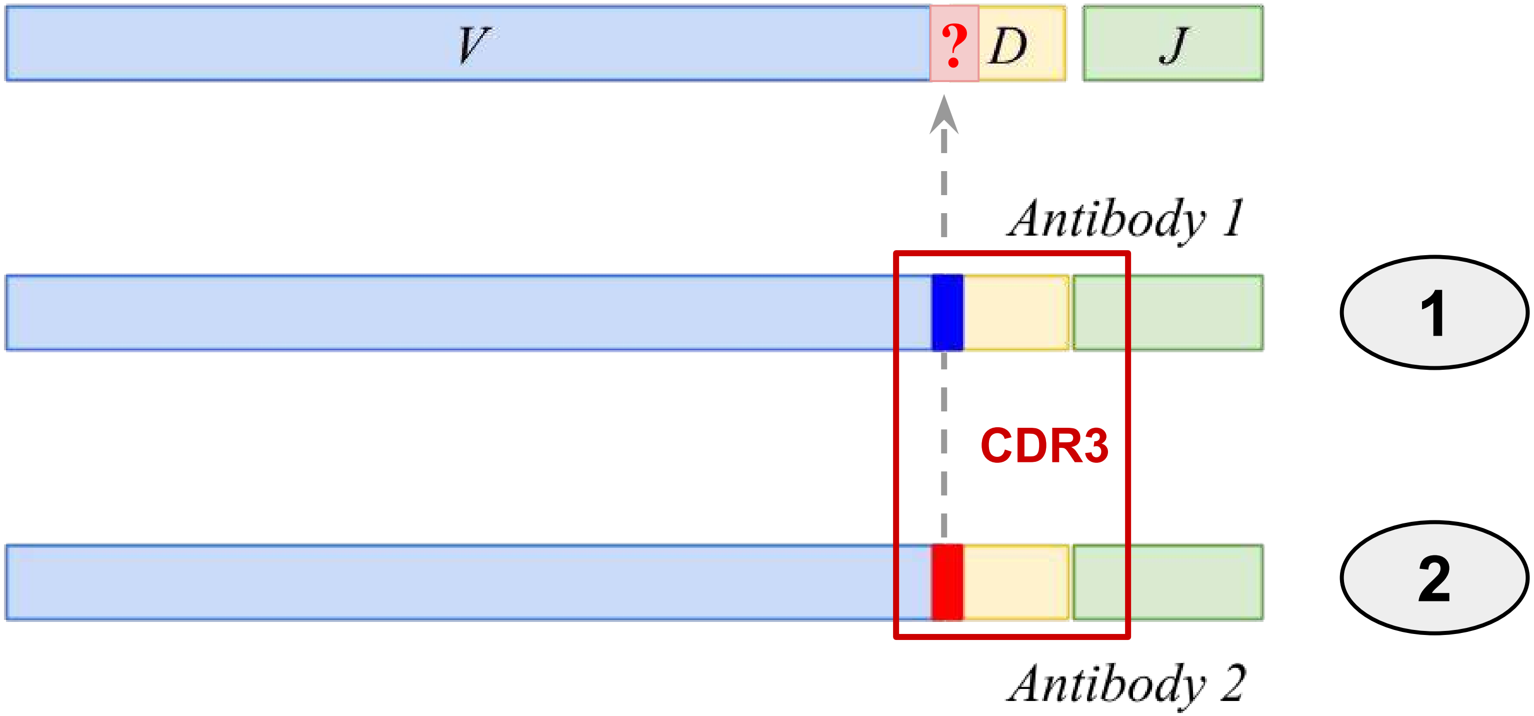
A more complex case: who is the ancestor?



Information about VDJ scenarios allows us to make the a choice:

- Antibodies 1 and 2 belong to the same lineage

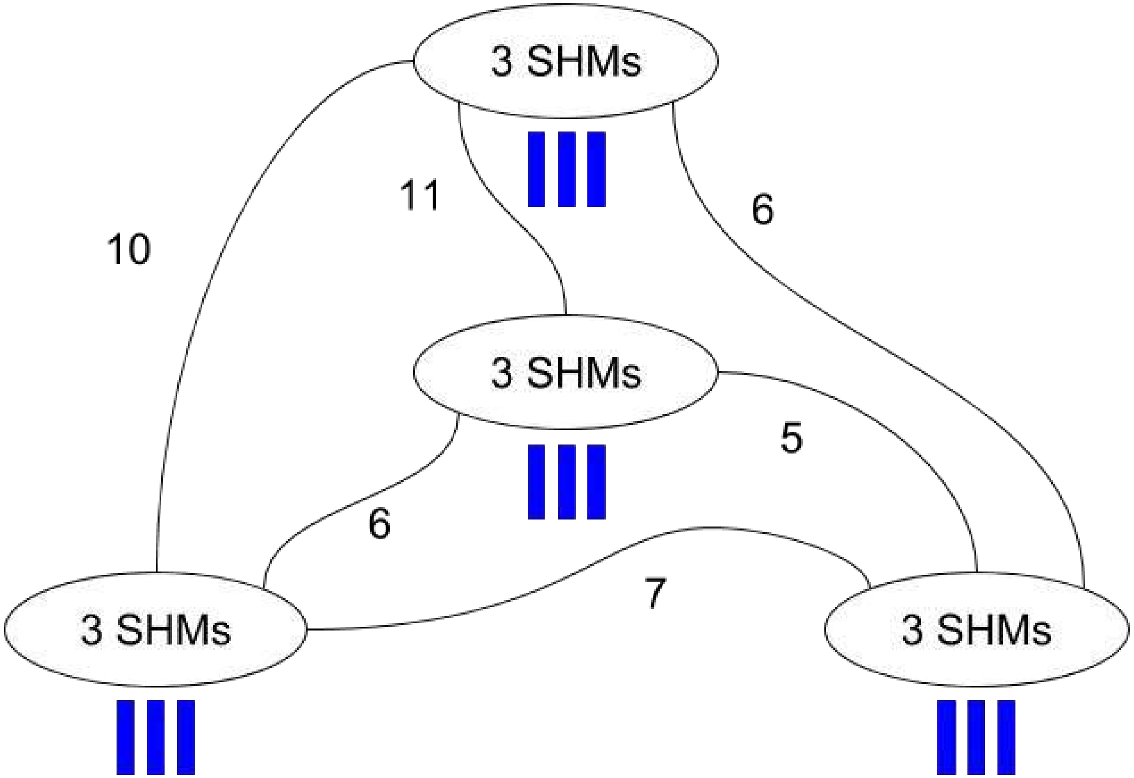
A more complex case: who is the ancestor?



Information about VDJ scenarios allows us to make the right choice:

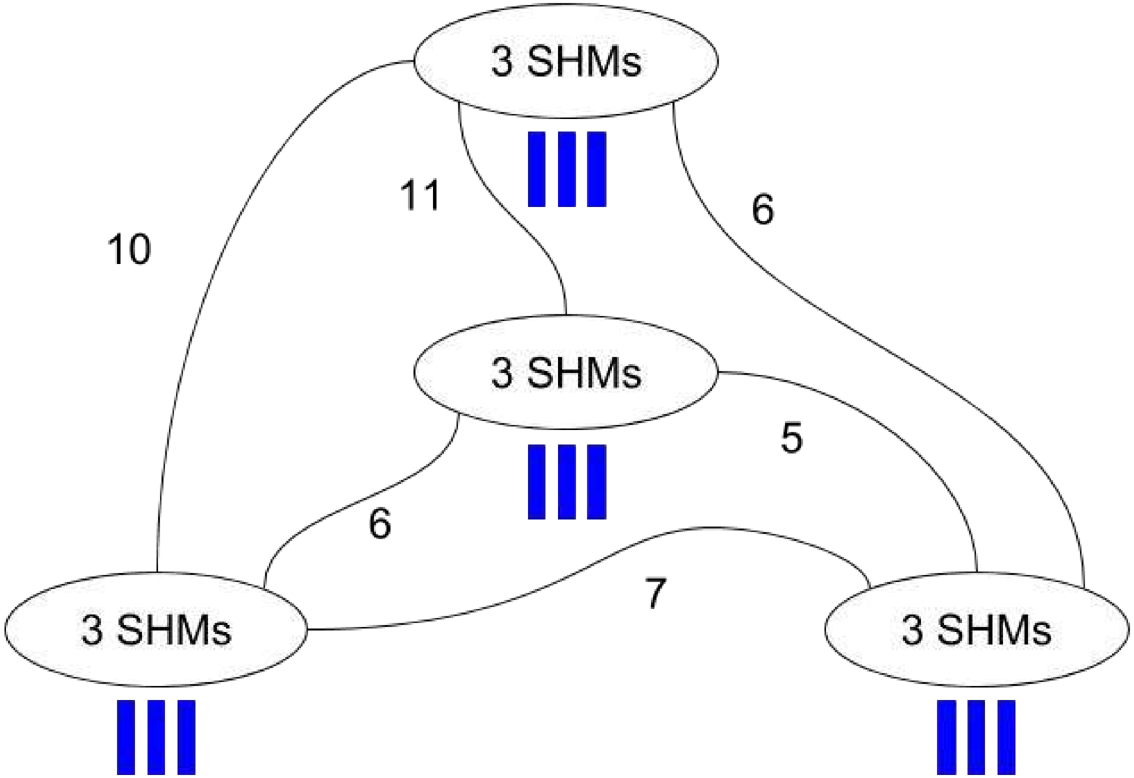
- Antibodies 1 and 2 belong to the same lineage
- Antibodies 1 and 2 are not related

Another puzzle



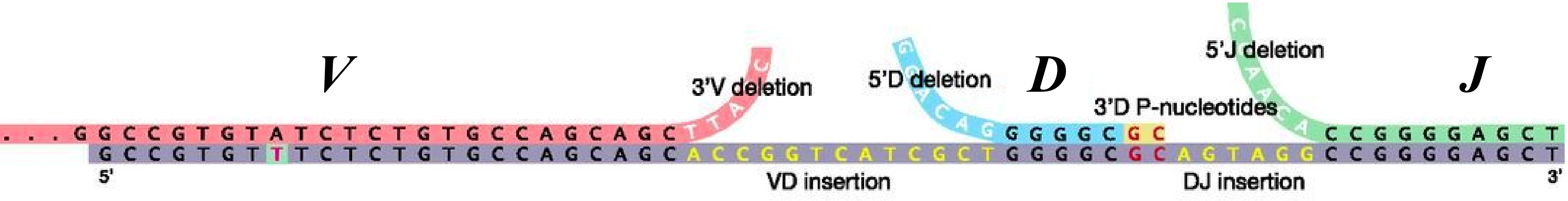
4 antibodies share SHMs in V segments but differ in CDR3s

Another puzzle



- It is unclear how to select direction between two similar CDR3s
- It is unclear whether two similar CDR3s belong to a single clonal tree or not

Why do we need a VDJ probabilistic model?

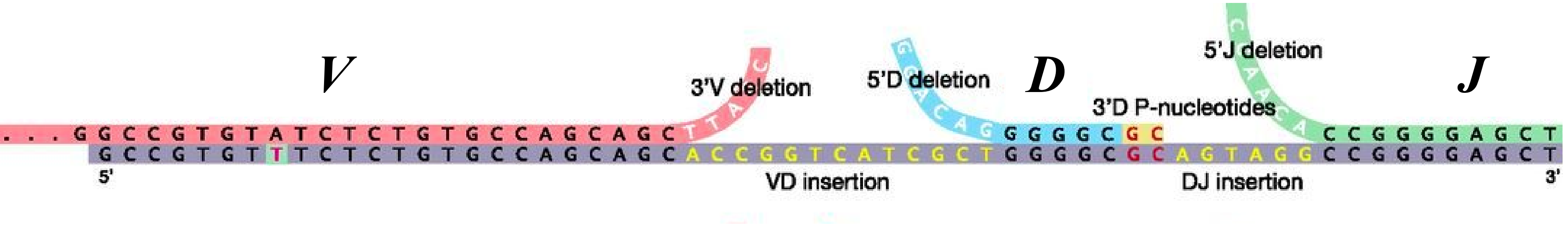

V
3'V deletion
5'D deletion
D
5'J deletion
J
3'D P-nucleotides
. . . . G G C C G T G T A T C T C T G T G C C A G C A G C T T A C
G C A C A G G G G G C G C
C A A A C A C C G G G G A G C T
G C C G T G T T T C T C T G T G C C A G C A G C A C C G G T C A T C G C T G G G G C G C A G T A G G C C G G G G A G C T
5'
3'
VD insertion
DJ insertion

To compute **VDJ scenario**, we need to:

- perform VDJ classification to find germline segments (well-studied problem)
- specify **deletions** in gene segments
- specify non-genomic **insertions**
- specify addition of **palindromes**

Murugan et al., *PNAS*, 2012

Why do we need a VDJ probabilistic model?


V
3'V deletion
5'D deletion
D
5'J deletion
J
3'D P-nucleotides
. . . . G G C C G T G T A T C T C T G T G C C A G C A G C T T A C
G C A C A G G G G G C G C
C A A C A C C G G G G A G C T
G C C G T G T T T C T C T G T G C C A G C A G C A C C G G T C A T C G C T G G G G C G C A G T A G G C C G G G G A G C T
5'
VD insertion
DJ insertion
3'

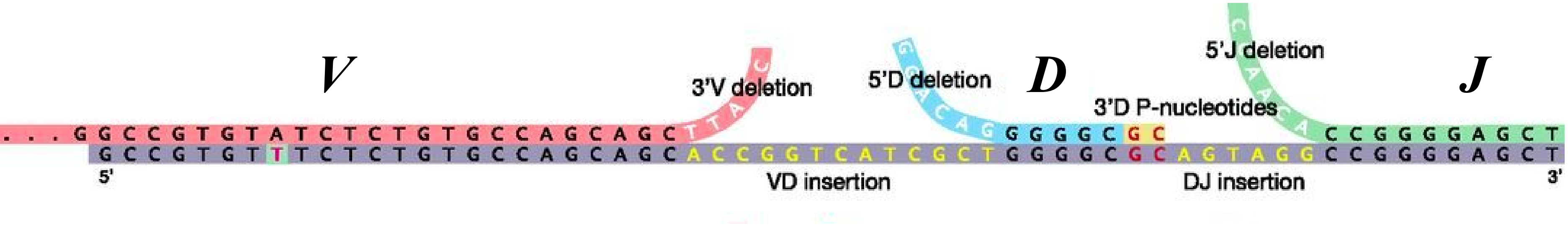
To compute VDJ scenario, we need to:

- perform VDJ classification to find germline segments (well-studied problem)
- specify deletions in gene segments
- specify non-genomic insertions
- specify addition of palindromes

Recombination events are not distributed uniformly

Murugan et al., *PNAS*, 2012

Why do we need a VDJ probabilistic model?


V
3'V deletion
5'D deletion
D
5'J deletion
3'D P-nucleotides
J
. . . G G C C G T G T A T C T C T G T G C C A G C A G C T T A C
G C A C A G G G G G C G C
C A A A C C G G G G A G C T
G C C G T G T T T C T C T G T G C C A G C A G C A C C G G T C A T C G C T G G G G C G C A G T A G G C C G G G G A G C T
5'
3'
VD insertion
DJ insertion

To compute VDJ scenario, we need to:

- perform VDJ classification to find germline segments (well-studied problem)
- specify deletions in gene segments
- specify non-genomic insertions
- specify addition of palindromes

Recombination events are not distributed uniformly

We need a probabilistic VDJ recombination model for a realistic description of these events

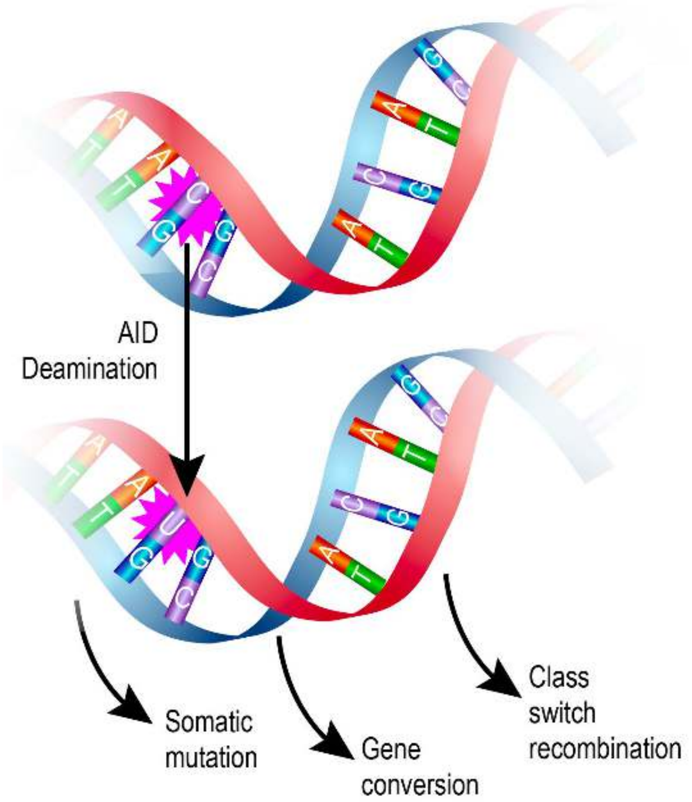
Murugan et al., *PNAS*, 2012

Why do we need an SHM probabilistic model?

SHM hotspots such as the degenerative 4-mers:

$$\begin{Bmatrix}A\\T\end{Bmatrix}\begin{Bmatrix}A\\G\end{Bmatrix}C\begin{Bmatrix}C\\T\end{Bmatrix}$$

trigger mutations in antibodies



Somatic hypermutagenesis engages AID enzyme that changes immunoglobulin genes to improve antibody affinity

Rogozin and Kolchanov, *Biochimica et Biophysica Acta*, 1992

Building probabilistic SHM model

5-mer	Freq	A	C	G	T
ACAAC	**83**	−	0.24	***0.48***	0.28
GGCGT	**1742**	0.22	−	0.12	***0.66***
CCGTC	**12**	0.35	***0.52***	−	0.13
TCTCC	**516**	0.32	***0.54***	0.14	−

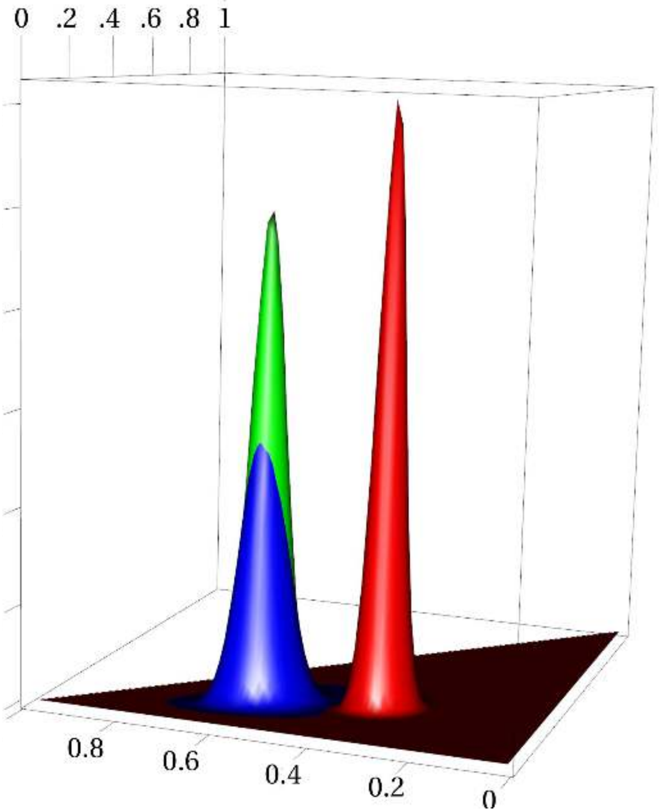
- The SHM model takes into account both the mutated nucleotide and its neighbours
- Detect new hot spots and compares SHMs in IG chains

Yaari et al., *Front Immunol,* 2013

Building probabilistic SHM model

5-mer	Freq	A	C	G	T
ACAAC	83	–	0.24	***0.48***	0.28
GGCGT	1742	0.22	–	0.12	***0.66***
CCGTC	12	0.35	***0.52***	–	0.13
TCTCC	516	0.32	***0.54***	0.14	–

- The SHM model takes into account both the mutated nucleotide and its neighbours
- Detect new hot spots and compares SHMs in IG chains



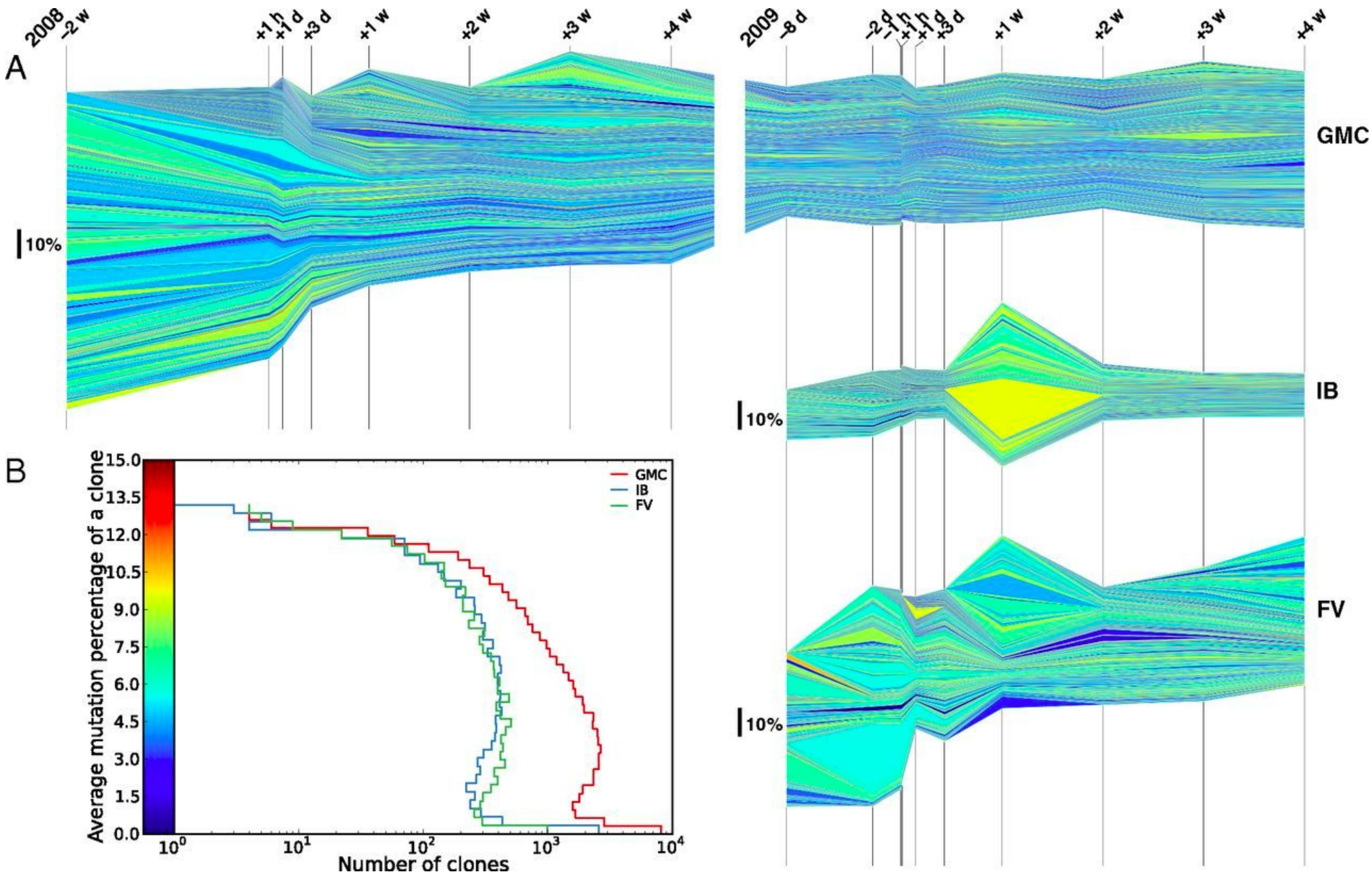
TCTCC 5-mer profiles for IGL, IGH, and IGK chains aggregated over 60 datasets

Yaari et al., *Front Immunol,* 2013

Outline

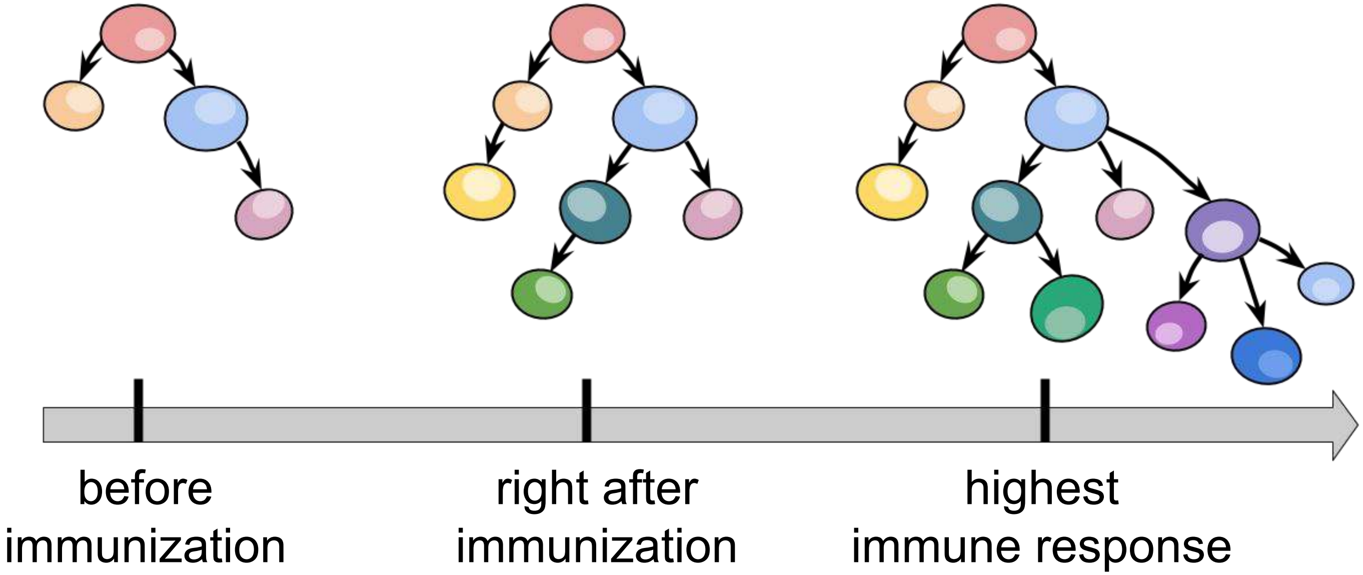
- Introduction
- Repertoire construction problem
- Evolutionary analysis of antibodies
- **Analysis of immune response dynamics**
- Analysis of paired antibody repertoires & new biological insights from analysis of paired repertoires

Time series



Laserson et al, *PNAS*, 2014

Clonal analysis in time



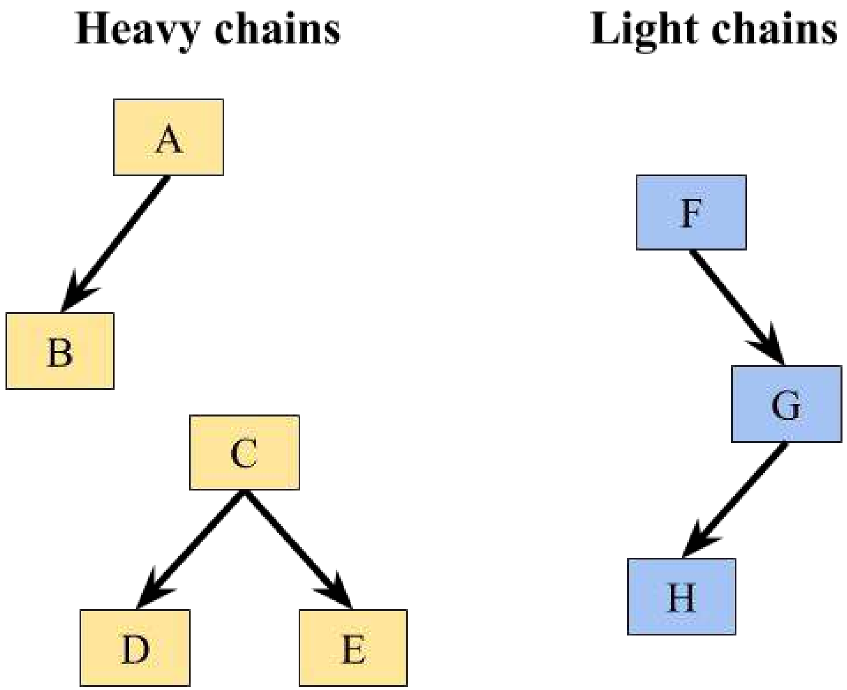
Clonal analysis of time series of antibody repertoire allows one to estimate efficiency of immune response

Sequencing data provided by AbVITRO

Outline

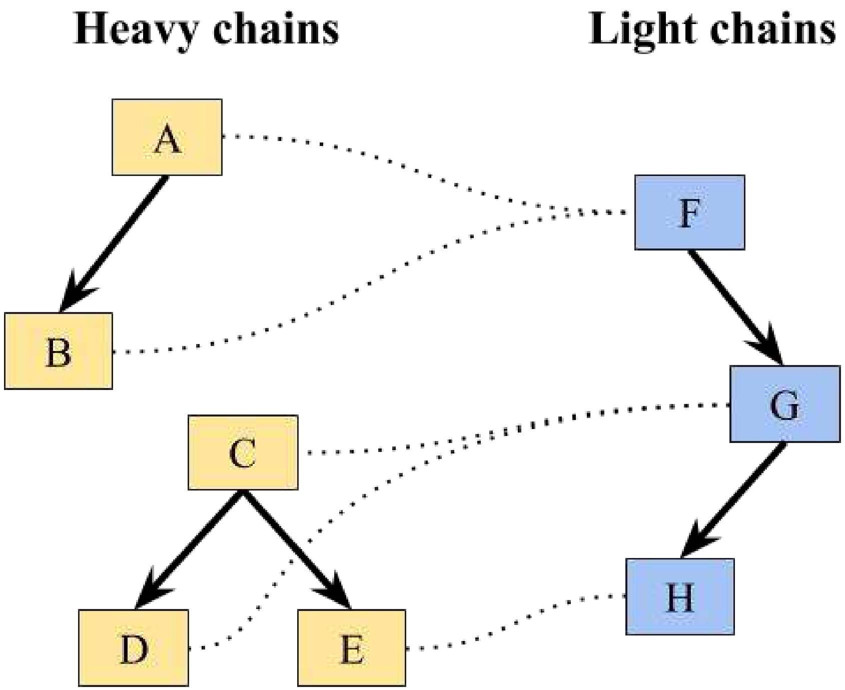
- Introduction
- Repertoire construction problem
- Evolutionary analysis of antibodies
- Analysis of immune response dynamics
- **Analysis of paired antibody repertoires & new biological insights from analysis of paired repertoires**

Clonal analysis for antibody repertoire



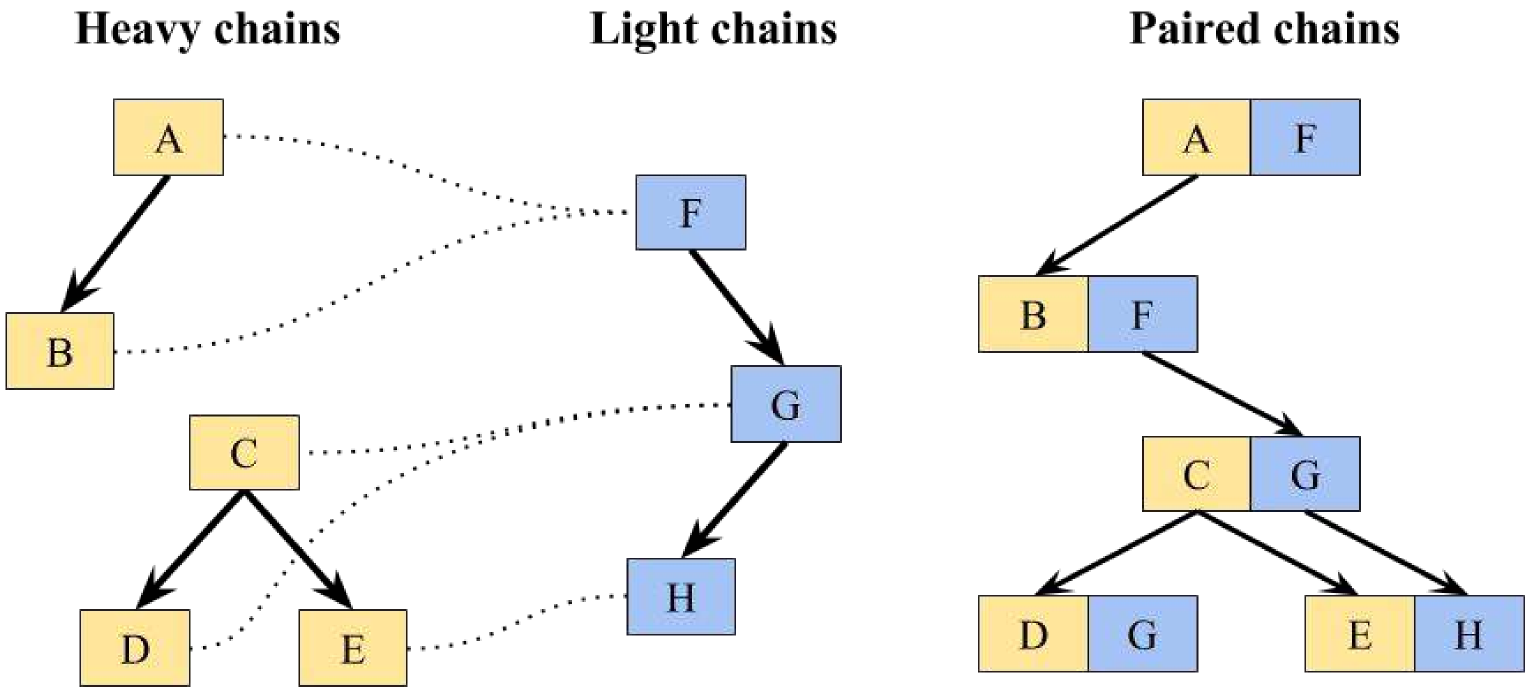
Sequencing data provided by AbVitro

Clonal analysis for paired antibody repertoire



Sequencing data provided by AbVitro

Clonal analysis for antibody repertoire

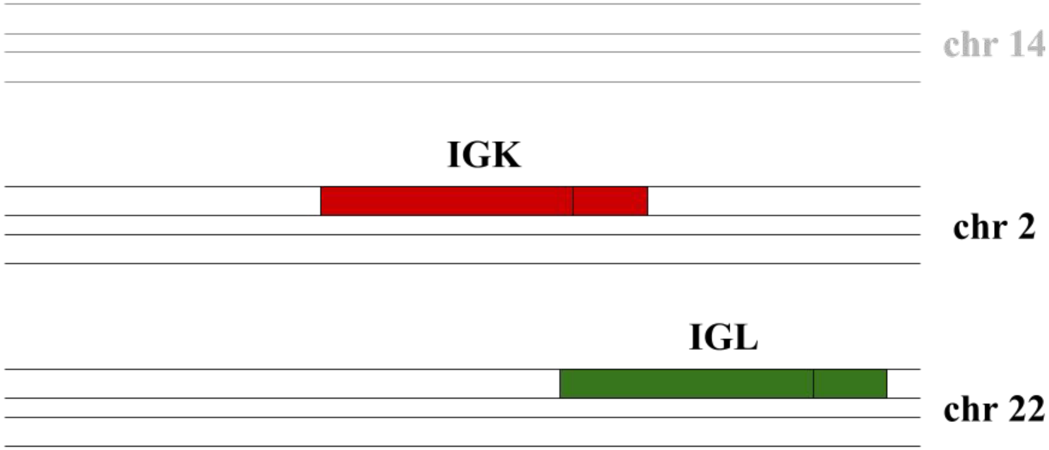


- utilizes information about chain pairing to construct **paired clonal tree**
- reveals that, contrary to previous views, B-cells **often** co-express multiple heavy and light chains.

Sequencing data provided by AbVitro

Light chain duality

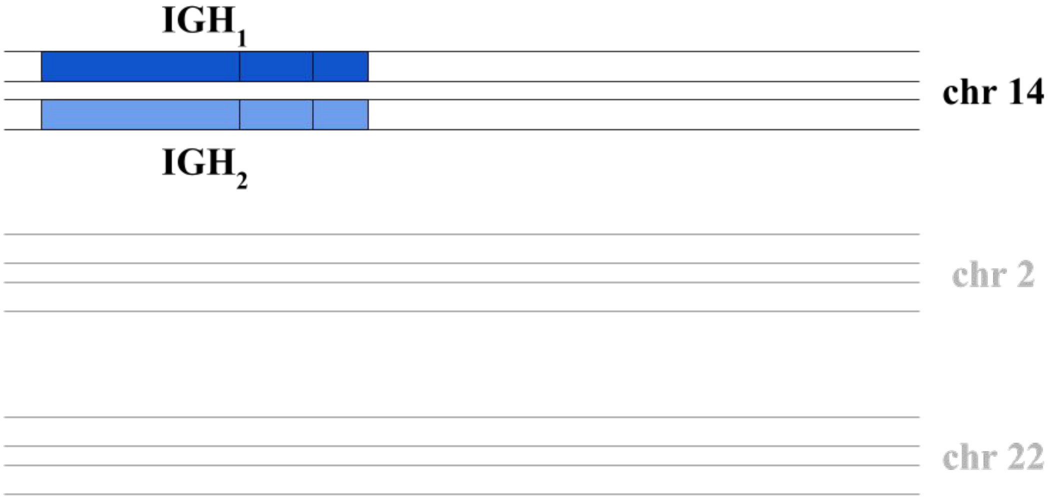
co-expression of both kappa and lambda chains by a single B-cell



Pelanda et al., *Cur Opin Immunol*, 2014
Giachino et al., *J Exp Med*, 1995

Allelic inclusion

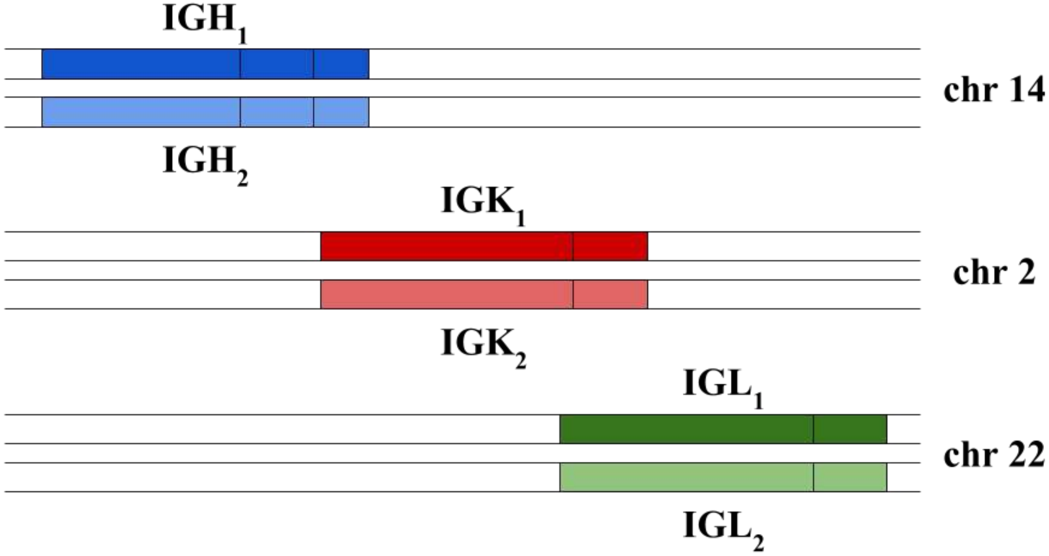
production of chains from both haplomes by B-cells



Casellas et al., *J Exp Med*, 2007
Beck-Engeser et al., *PNAS,* 1987

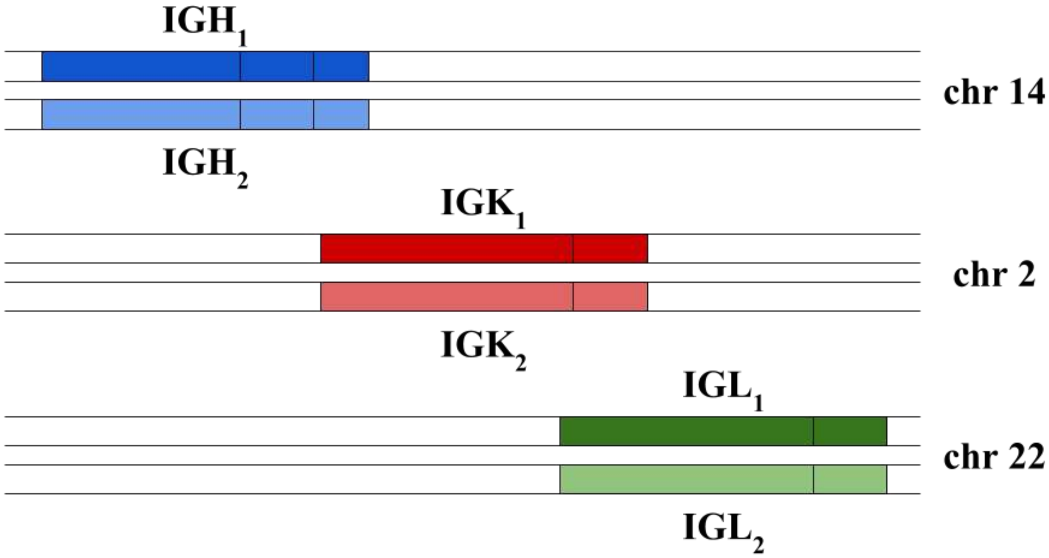
Duality + allelic inclusion

A single B-cell may express multiple chains due to allelic inclusions and/or light chain duality

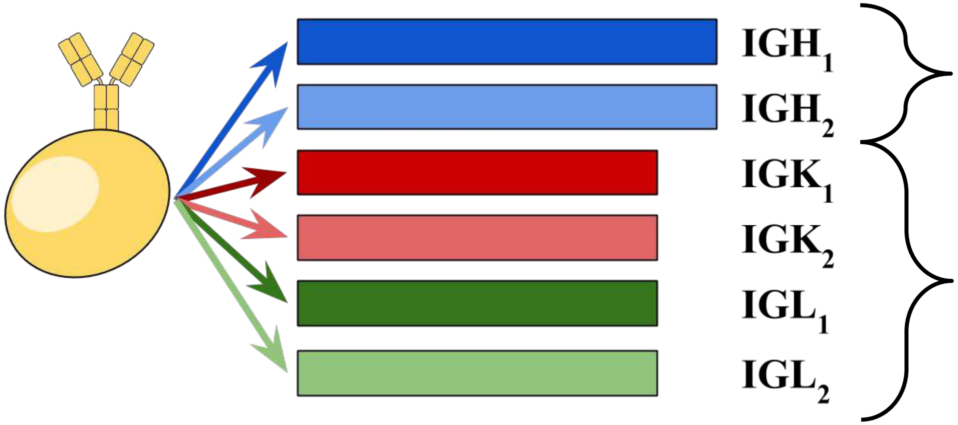


Multi-chain effect

A single B-cell may express multiple chains due to allelic inclusions and/or light chain duality

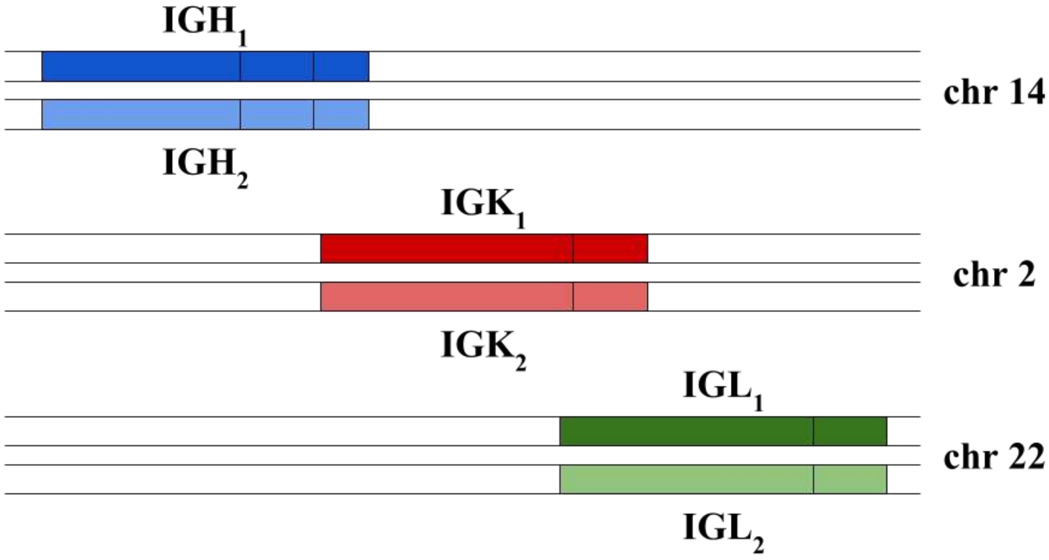


Multi-chain effect: B-cell can express up to **6 different chains:**

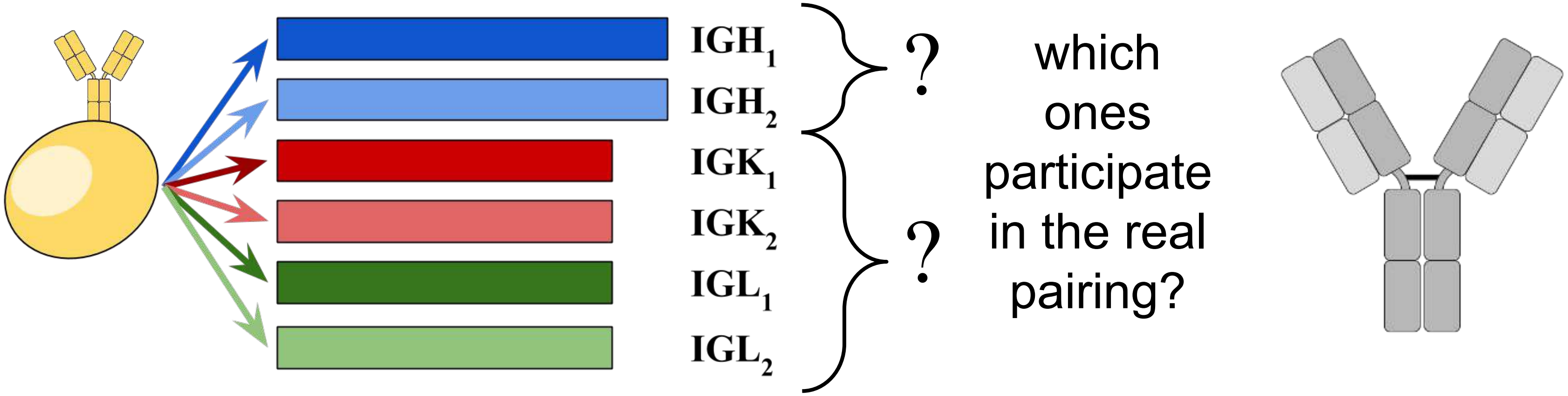


Multi-chain effect

A single B-cell may express multiple chains due to allelic inclusions and/or light chain duality

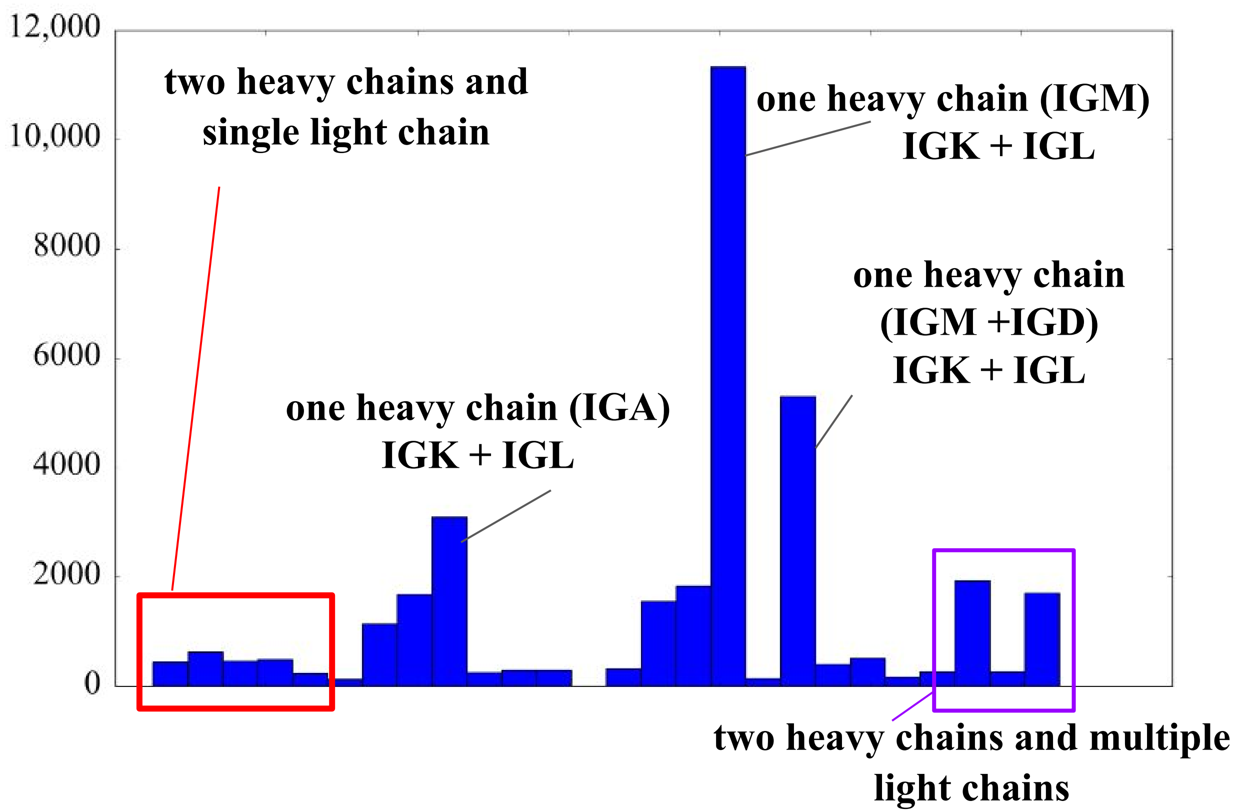


Multi-chain effect: B-cell can express up to **6 different chains:**



Multi-chain effect is common in healthy B-cells!

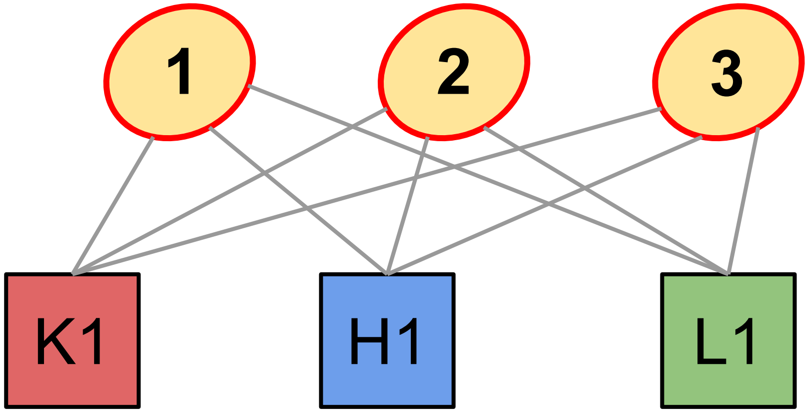
25% (!) of B-cells with known pairing have allelic inclusions and/or light chain duality



Clonal analysis reveals true chain pairing

Cells 1, 2, and 3 express identical **heavy**, **kappa** and **lambda** chains. Thus, 1, 2, and 3 are clones of the same B-cell

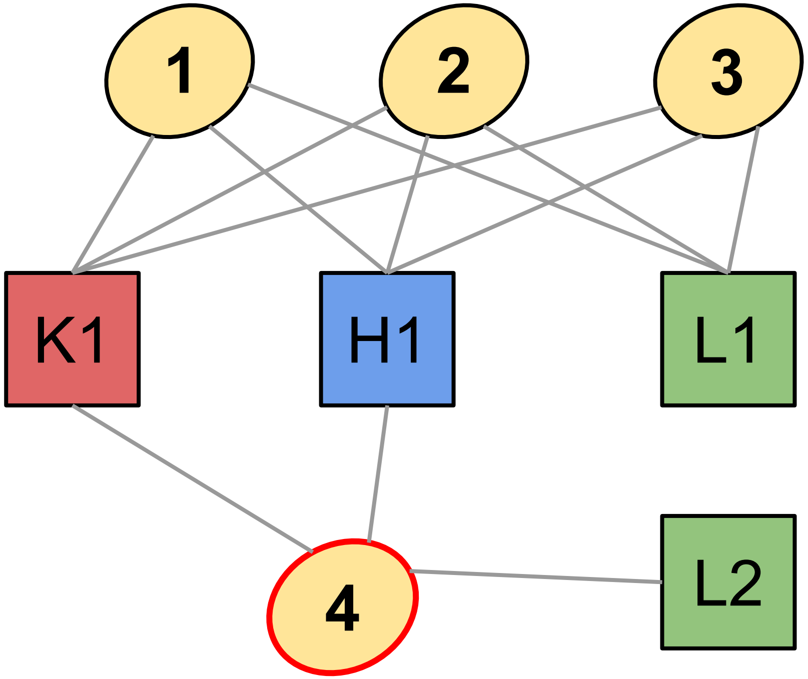
Which light chain contributes to the antibody:
kappa or **lambda?**



Example from AbVitro sequencing data

Clonal analysis reveals true chain pairing

Cell 4 shares **heavy** and **kappa** chains with cells 1, 2 and 3, but has different **lambda** chain (L2)

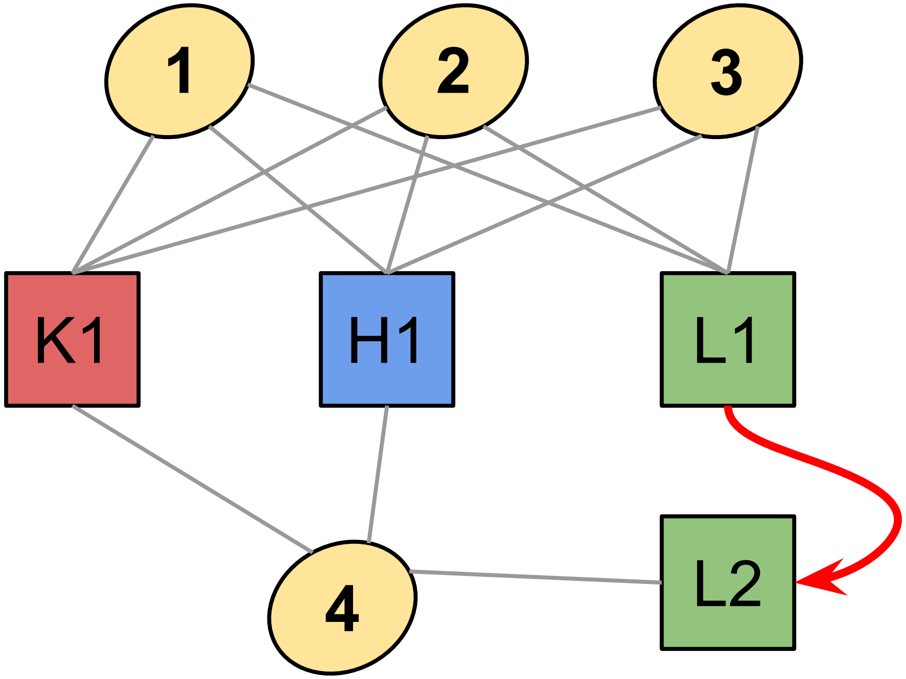


Clonal analysis reveals true chain pairing

Alignment of L1 and L2 reveals that L1 is an ancestor of L2

Thus, cell 4 is a descendant of cells 1, 2, and 3

Cell 4 shares **heavy** and **kappa** chains with cells 1, 2 and 3, but has different **lambda** chain (L2)

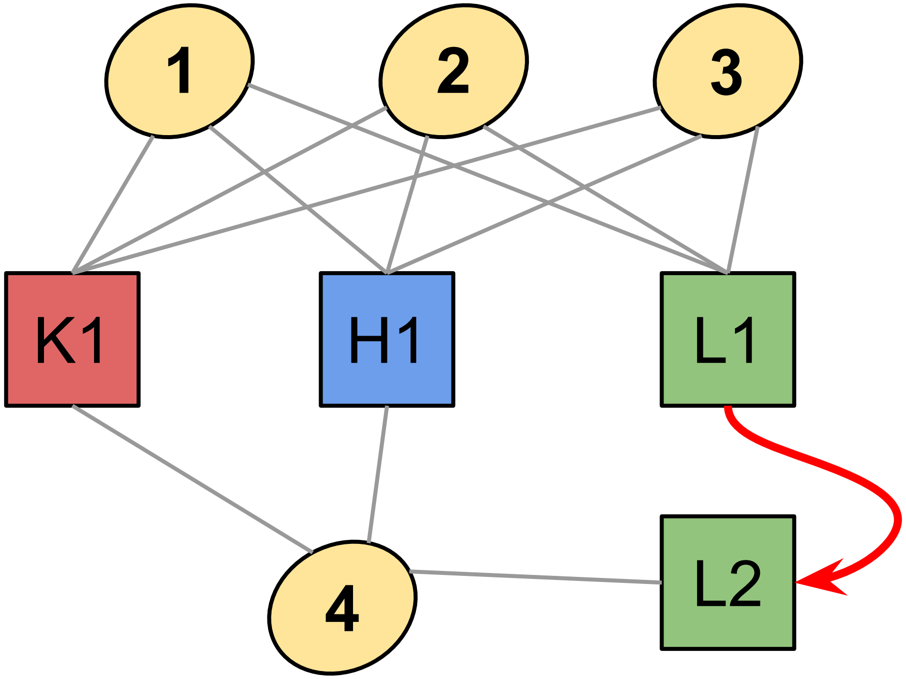


Clonal analysis reveals true chain pairing

Alignment of L1 and L2 reveals that L1 is an ancestor of L2

Thus, cell 4 is a descendant of cells 1, 2, and 3

Evolution of L1 into L2 provides evidence that cells 1, 2, 3, and 4 generate **functional antibodies**



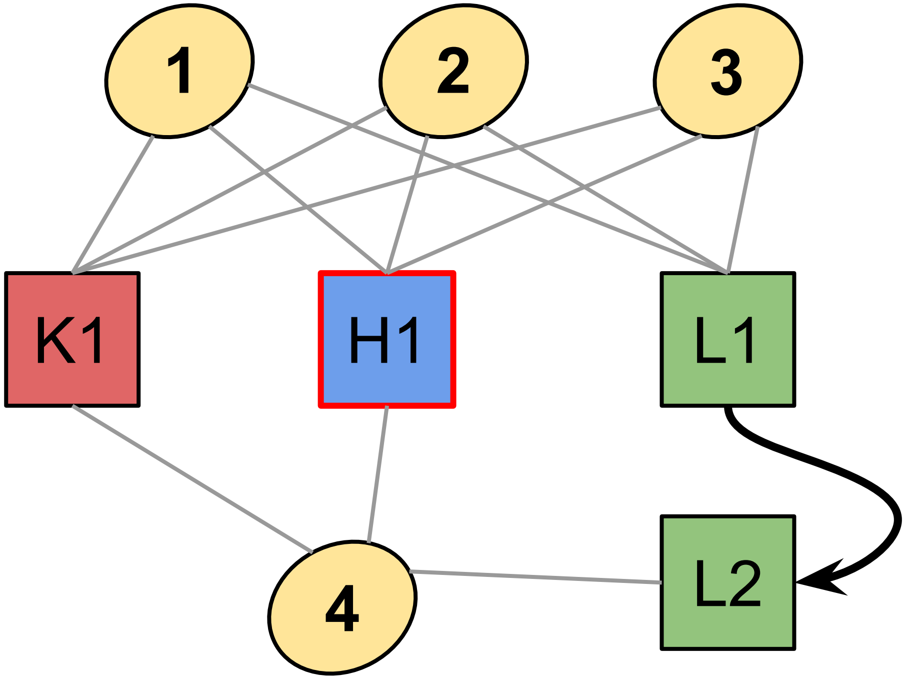
Clonal analysis reveals true chain pairing

Alignment of L1 and L2 reveals that L1 is an ancestor of L2

Thus, cell 4 is a descendant of cells 1, 2, and 3

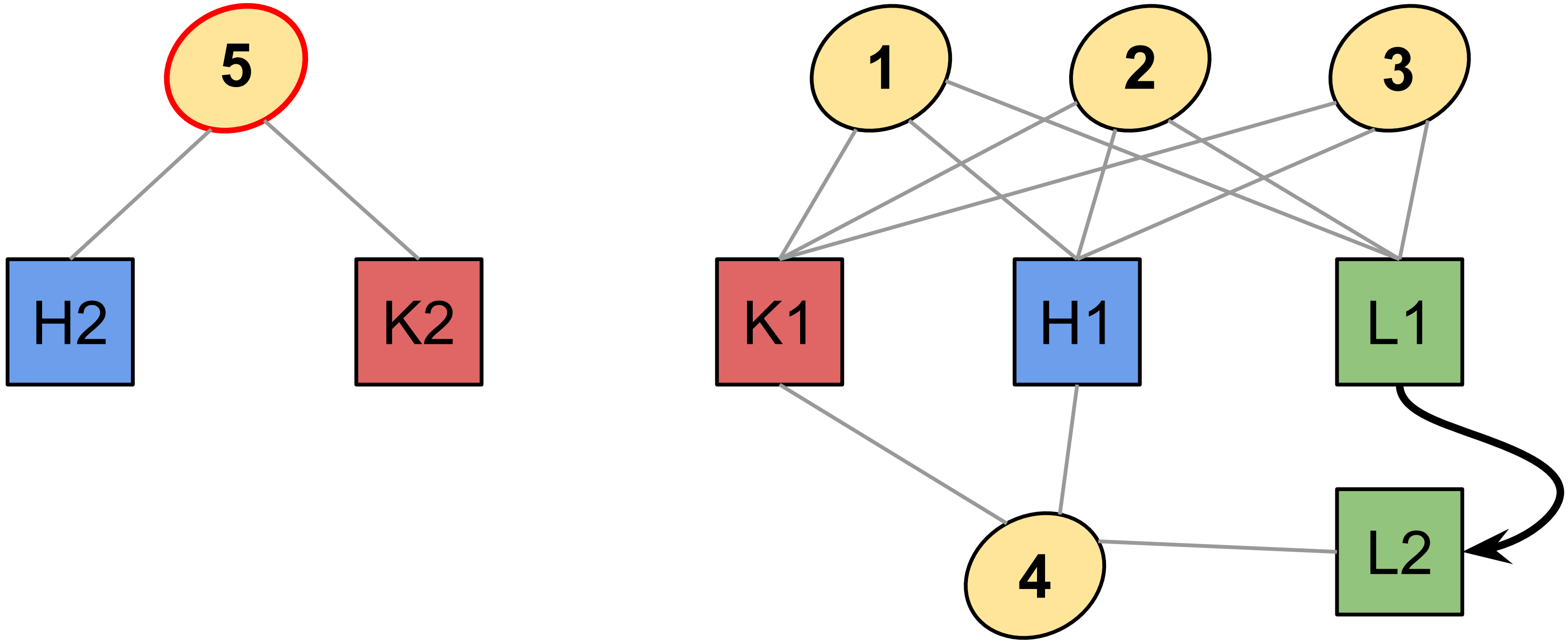
Evolution of L1 into L2 provides evidence that cells 1, 2, 3, and 4 generate **functional antibodies**

But it contradicts with a fact that H1 is non-productive



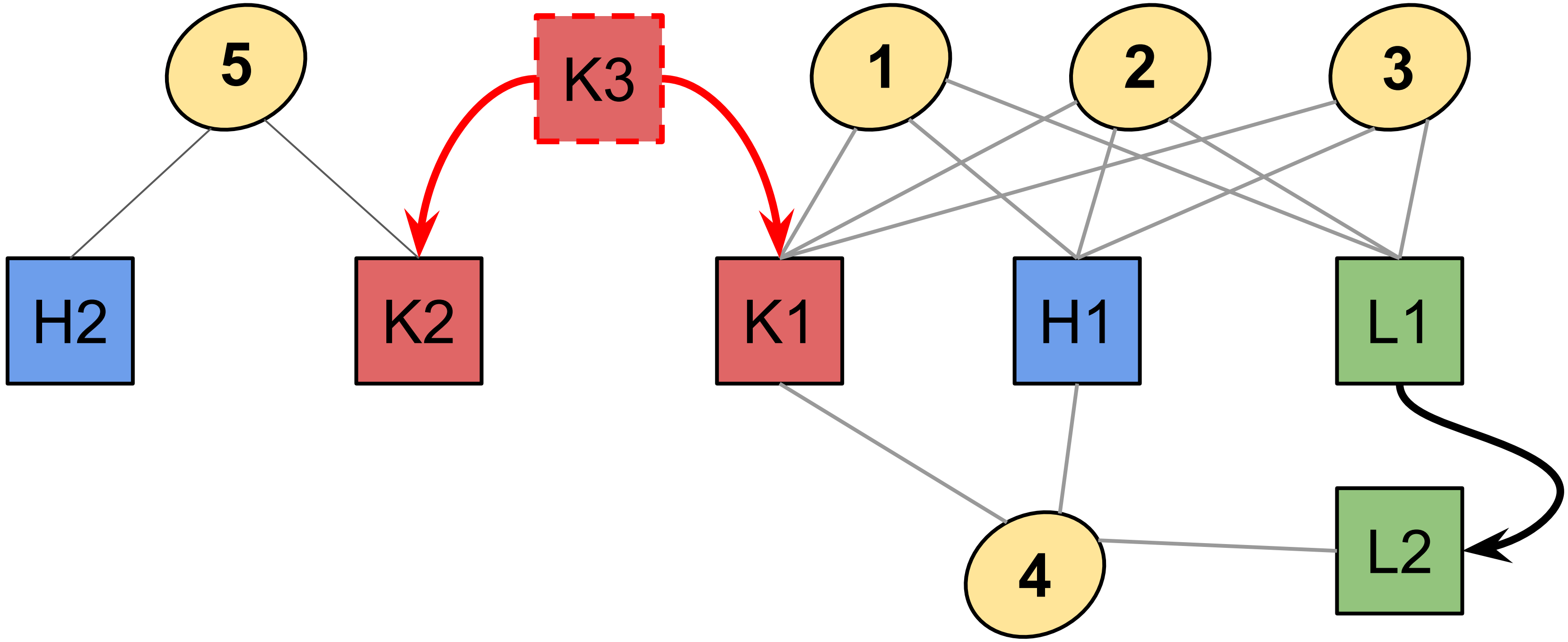
There are more B-cells to analyze!

Cell 5 expresses **heavy** and **kappa** chains

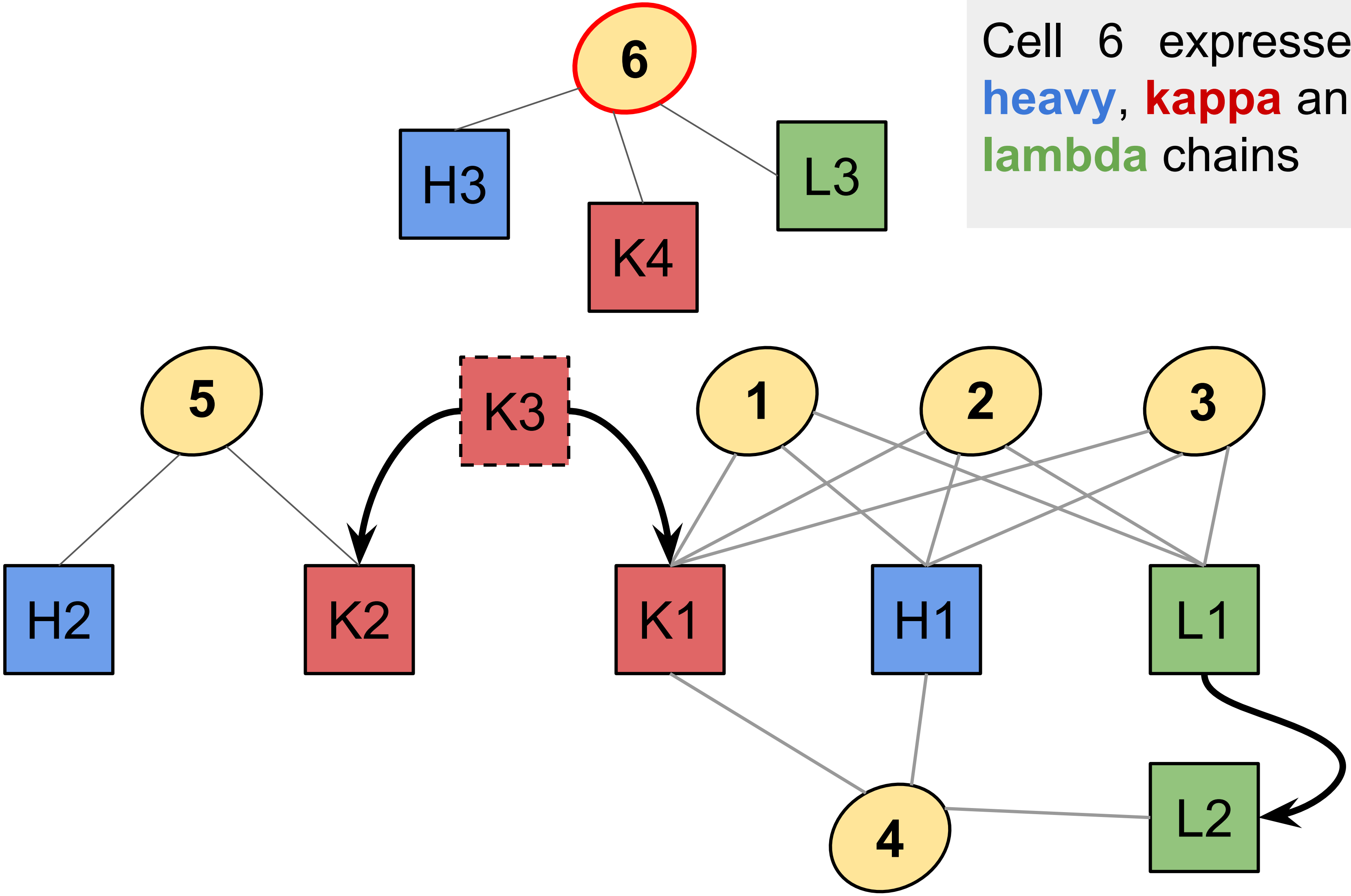


There are more B-cells to analyze!

K2 and K1 have originated from a an unknown kappa chain K3 that is missing in the repertoire

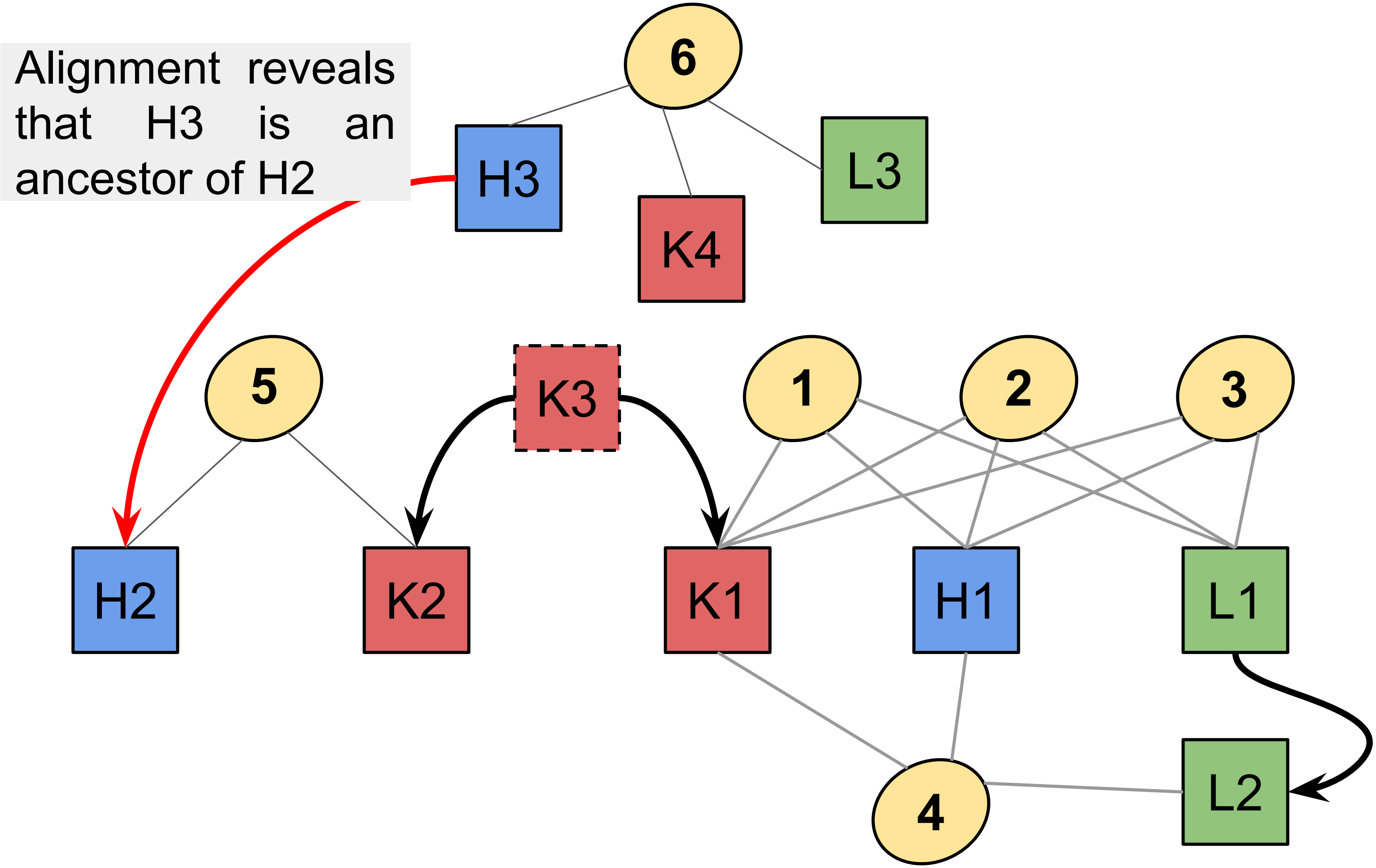


We are not done yet…



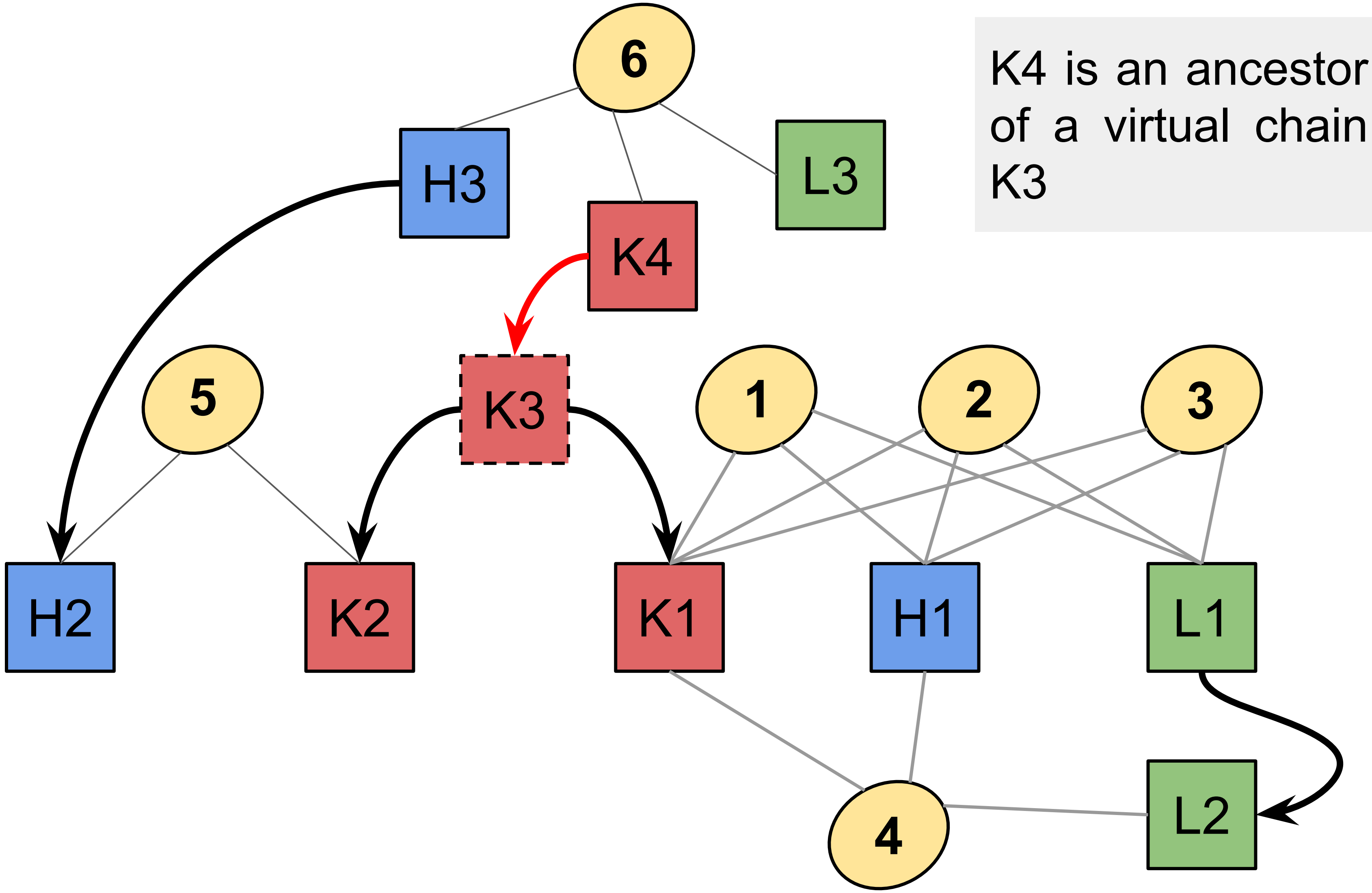
Cell 6 expresses **heavy**, **kappa** and **lambda** chains

We are not done yet…



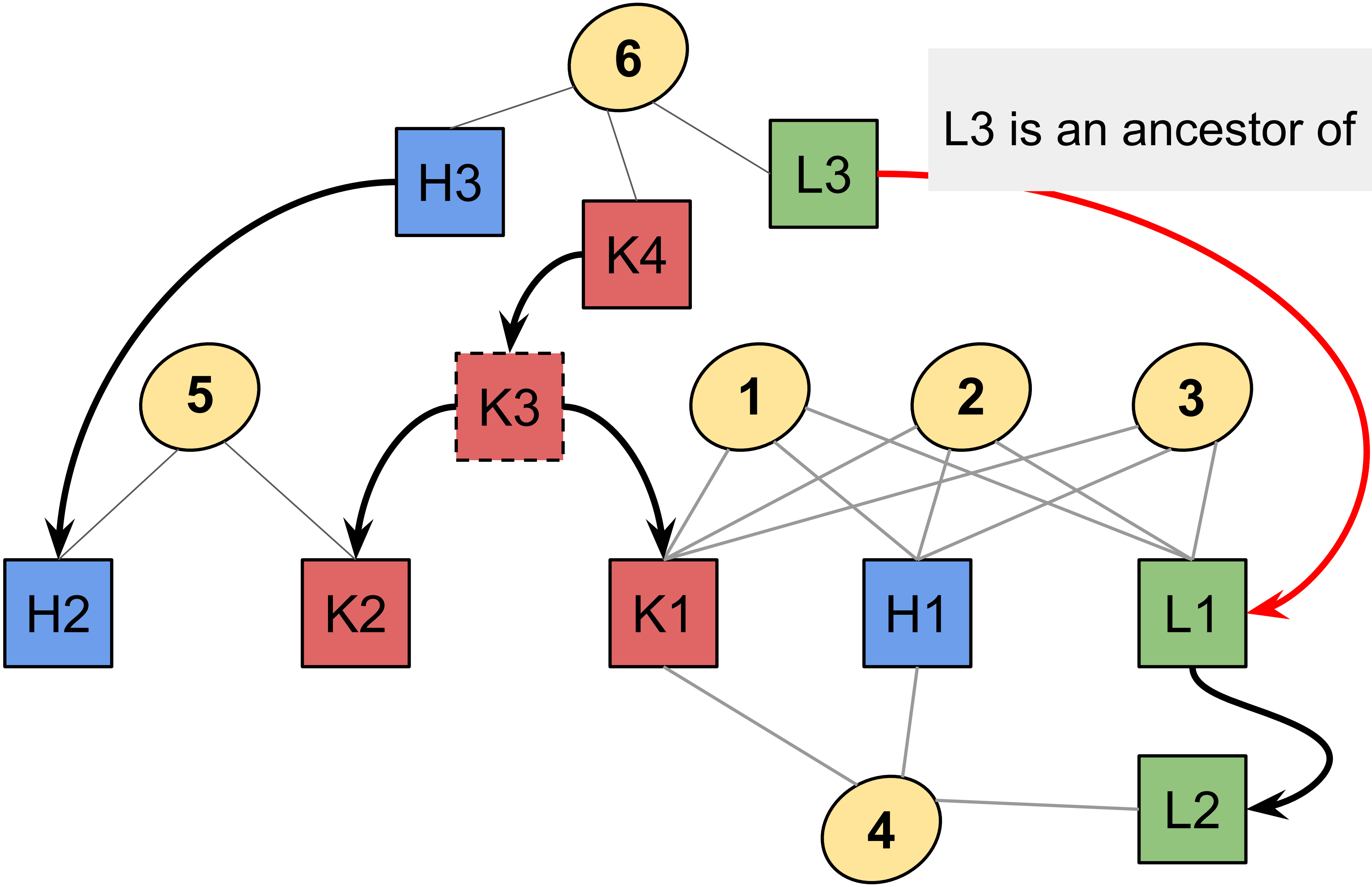
Alignment reveals that H3 is an ancestor of H2

We are not done yet…



K4 is an ancestor of a virtual chain K3

We are not done yet…

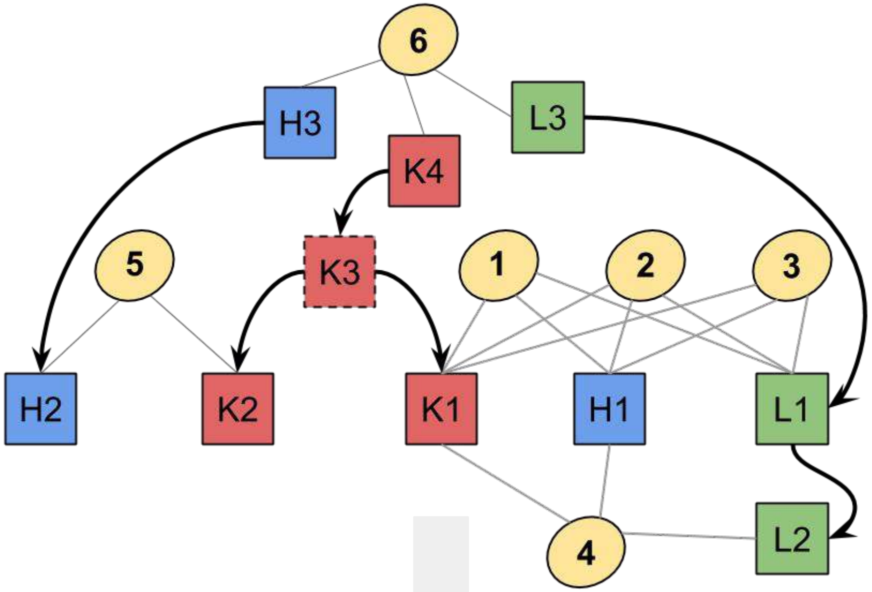


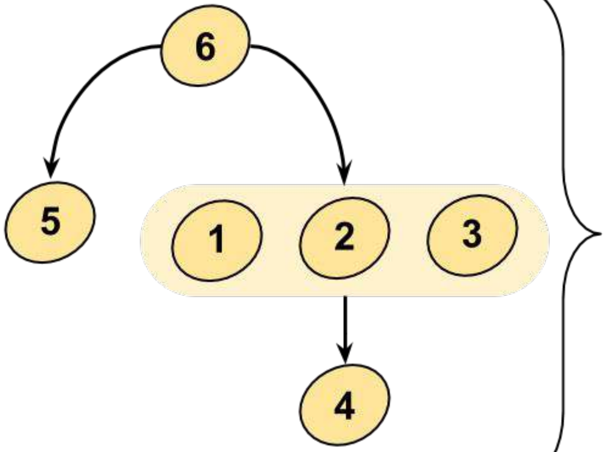
L3 is an ancestor of L1

Evolutionary analysis helps to understand true chain pairing

H1 lineage is non-productive, so it does not participate in pairing

Lineage H3 → H2 is more likely to participate in chain pairing





H: H3 → H2 👍

H: H1 ✗

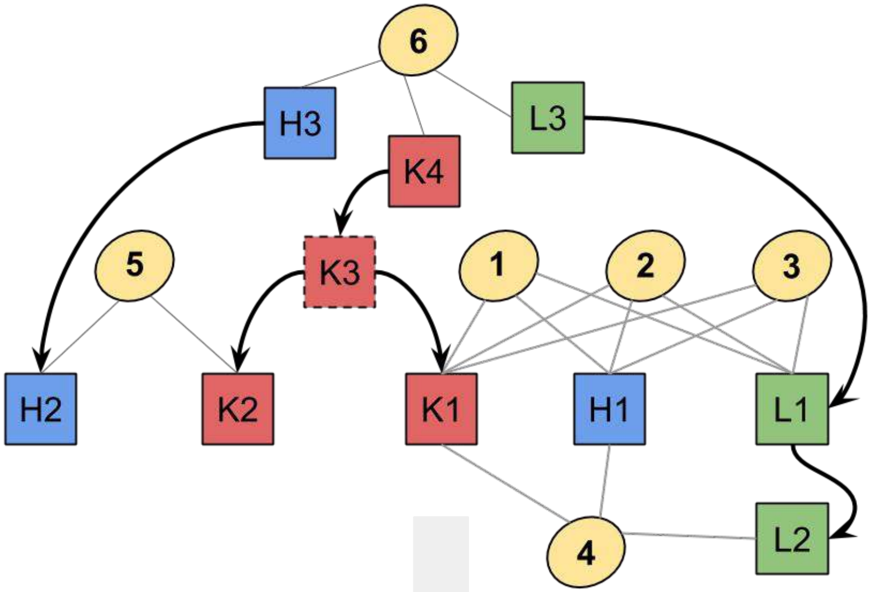
L: L3 → L1 → L2

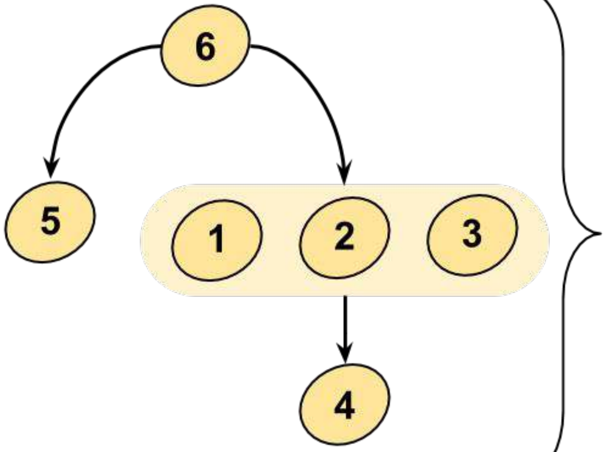
K: K4 → K3,
K3 → K1,
K3 → K2

Evolutionary analysis helps to understand true chain pairing

- Lambda lineage contain synonymous mutations
- Mutations in lambda lineage are grouped into CDRs
- Mutations in kappa chain are distributed randomly along variable region

Lambda lineage undergoes selection, thus it more likely participates in chain pairing





H: H3 → H2

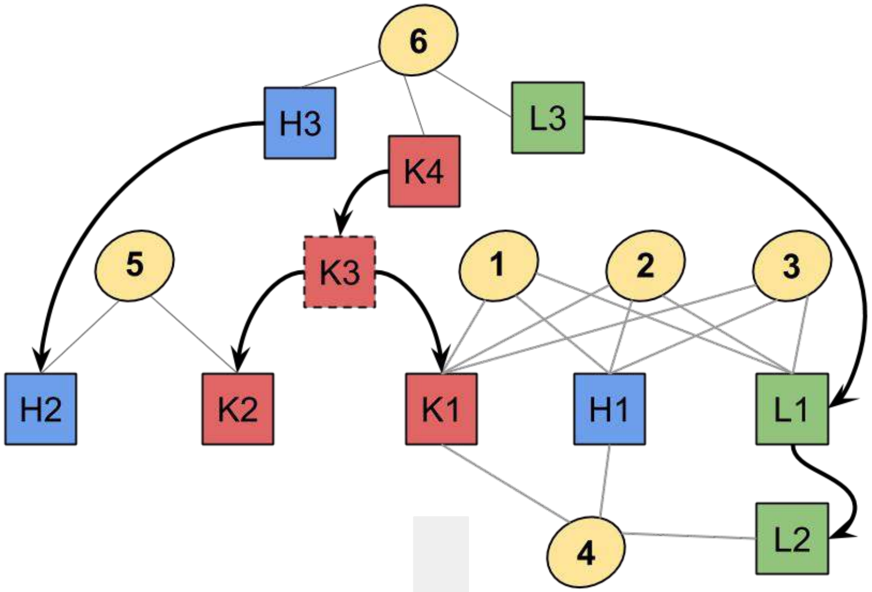
H: H1

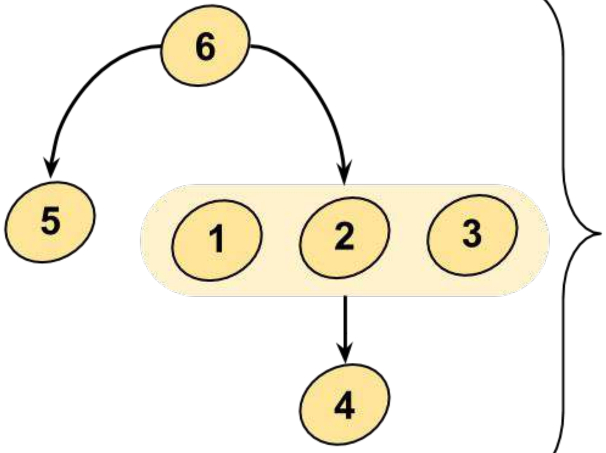
L: L3 → L1 → L2

K: K4 → K3,
K3 → K1,
K3 → K2

Evolutionary analysis helps to understand true chain pairing

Using information about clonal lineages for H, K and L chains and the SHM model, we can select the most likely chain pairing



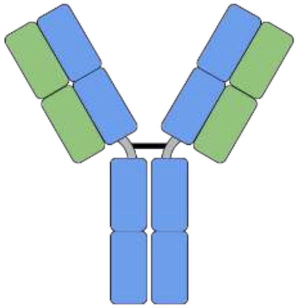


H: H3 → H2

H: H1

L: L3 → L1 → L2

K: K4 → K3,
K3 → K1,
K3 → K2



H3 → H2

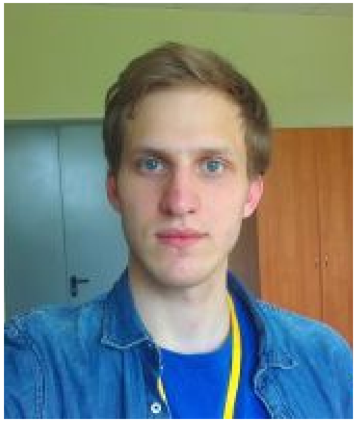
L3 → L1 → L2

Yana
Safonova

Alexander
Shlemov

Andrey
Bzikadze

Sergey
Bankevich



Timofey
Prodanov

Andrey
Slabodkin

Alla
Lapidus

Pavel A.
Pevzner

Thank you!