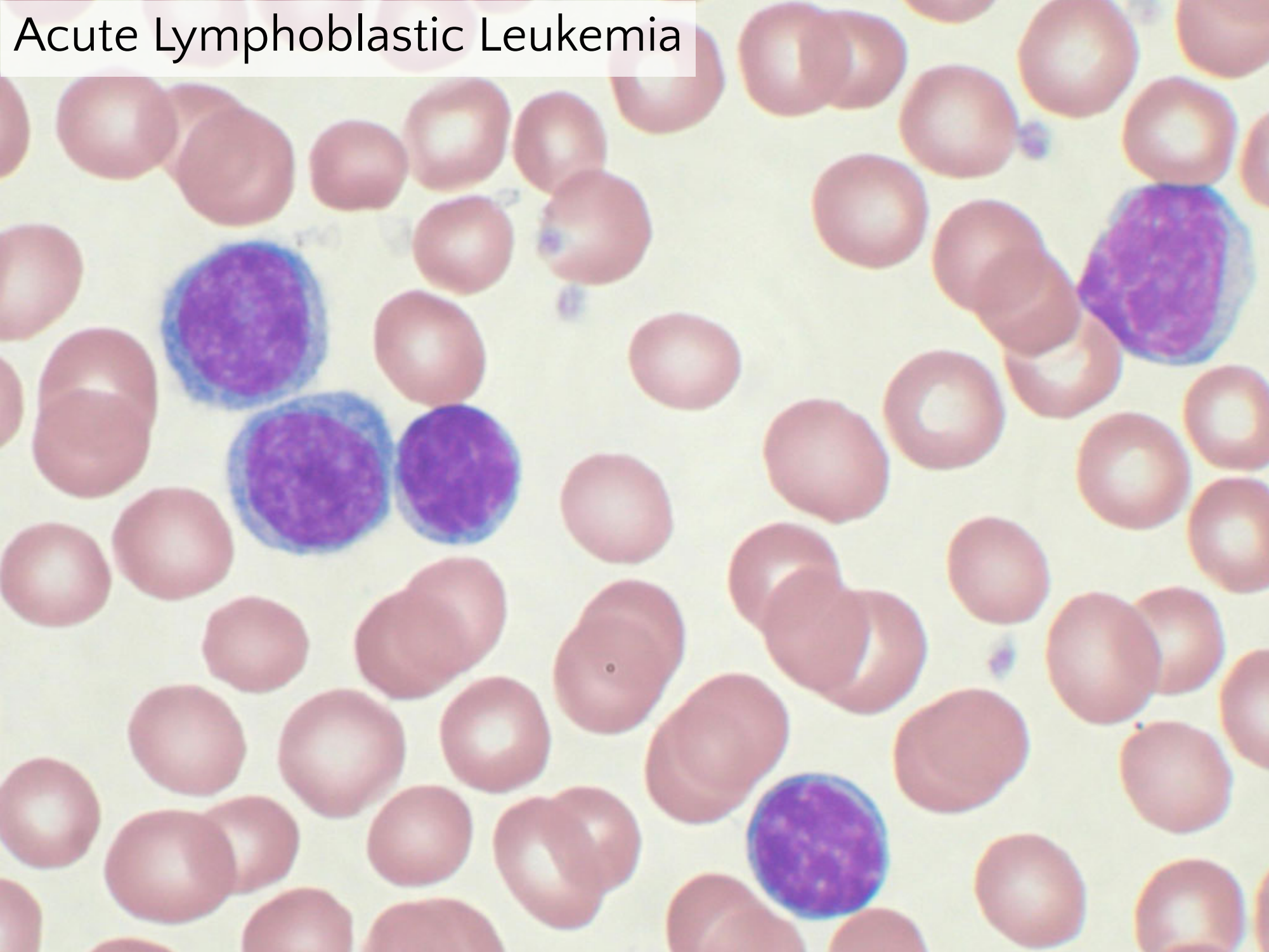
Vidjil, efficient sequencing data processing for onco-hematology

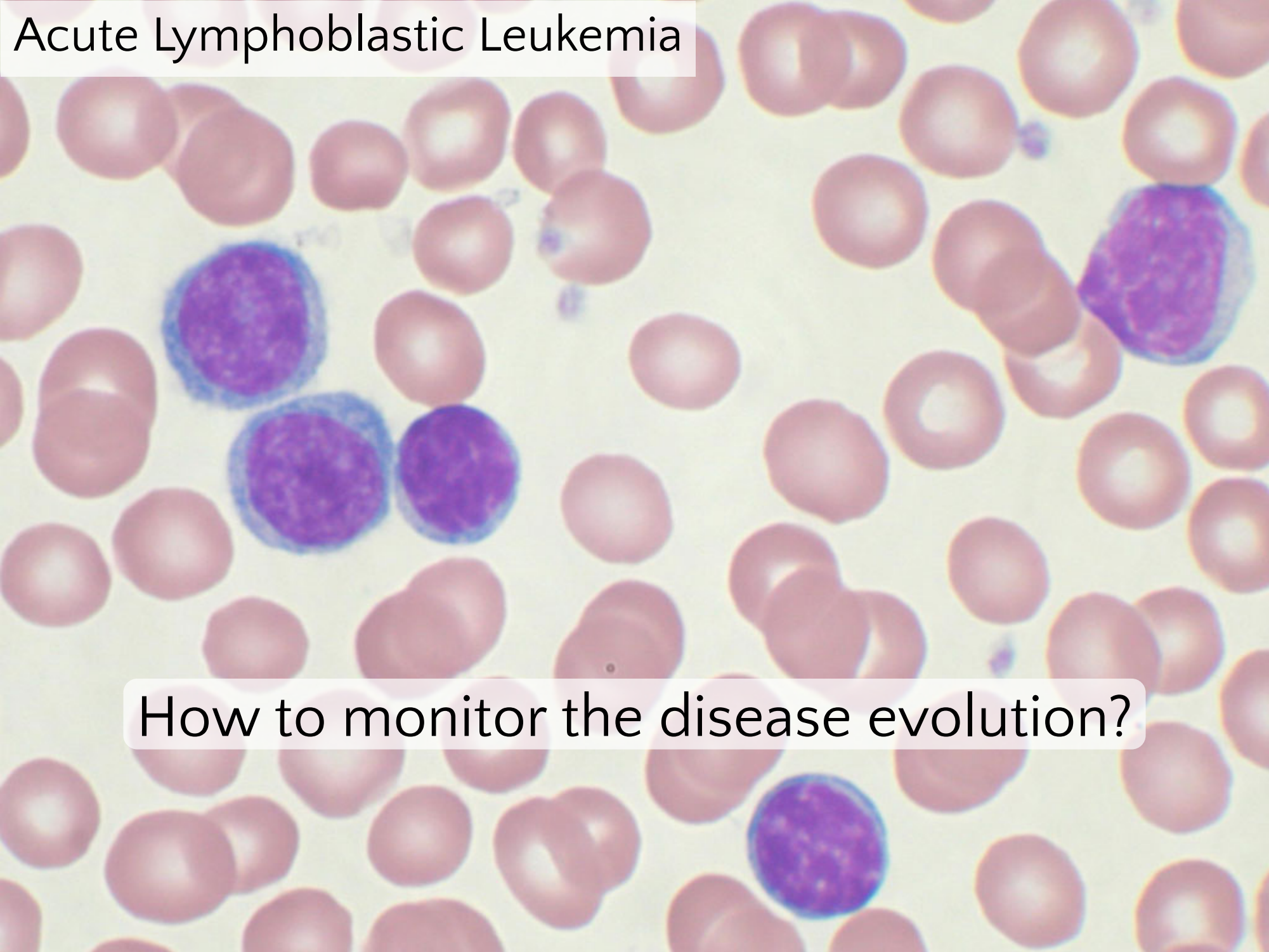
Marc Duez, Mathieu Giraud, Mikaël Salson, Florian Thonier

Équipe Bonsai, CRIStAL (U. Lille, CNRS)

Acute Lymphoblastic Leukemia



Acute Lymphoblastic Leukemia

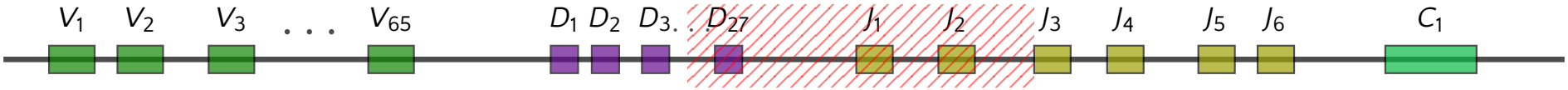


How to monitor the disease evolution?

V(D)J recombinations used as disease marker

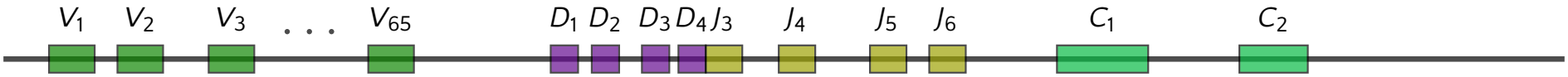
V(D)J recombinations used as disease marker

On a lymphocyte genome…



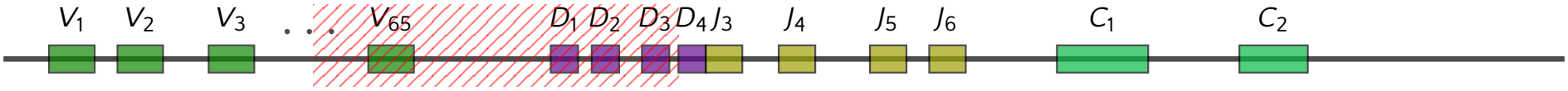
V(D)J recombinations used as disease marker

On a lymphocyte genome…



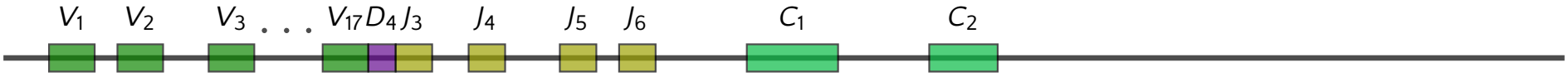
V(D)J recombinations used as disease marker

On a lymphocyte genome…



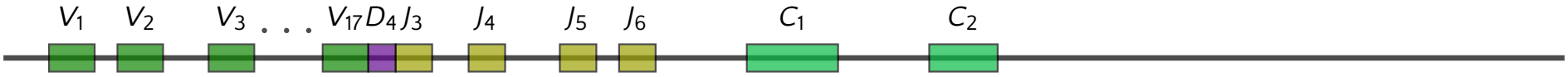
V(D)J recombinations used as disease marker

On a lymphocyte genome...

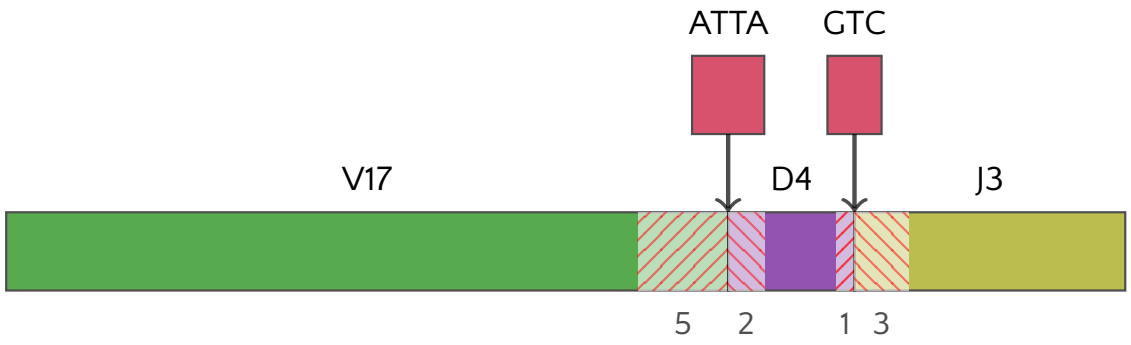


V(D)J recombinations used as disease marker

On a lymphocyte genome…

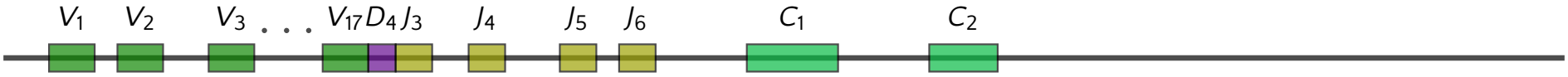


150 – 500nt

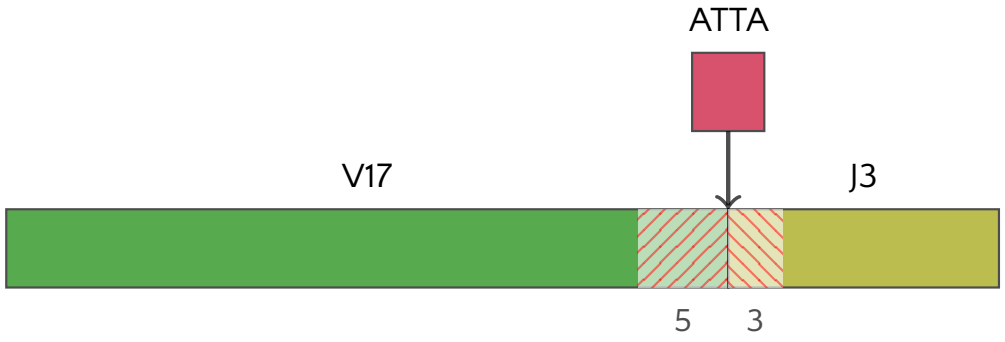


V(D)J recombinations used as disease marker

On a lymphocyte genome...

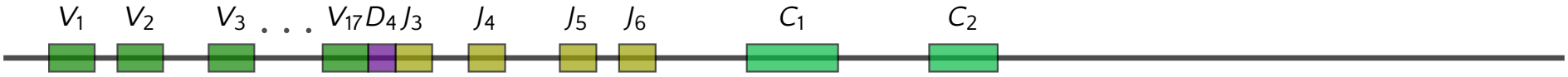


150 – 500nt

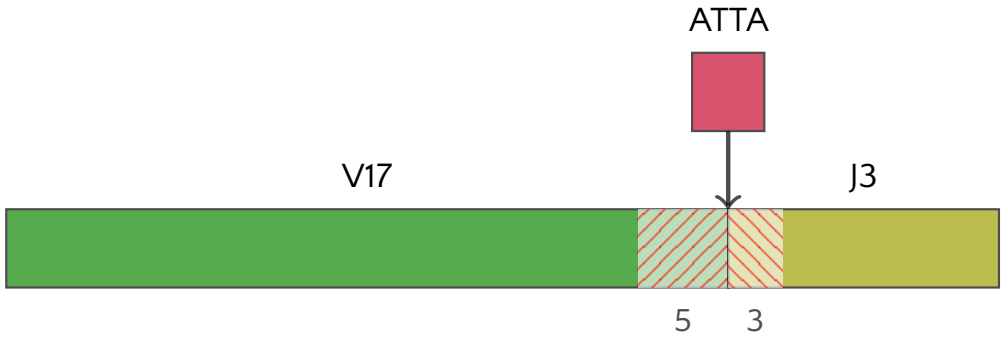


V(D)J recombinations used as disease marker

On a lymphocyte genome…

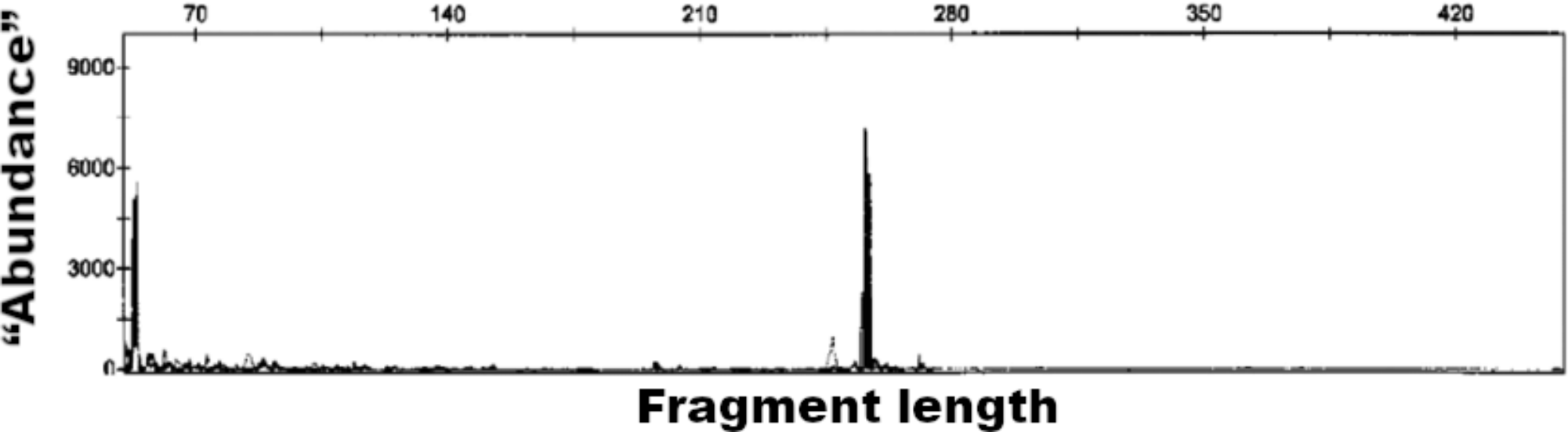


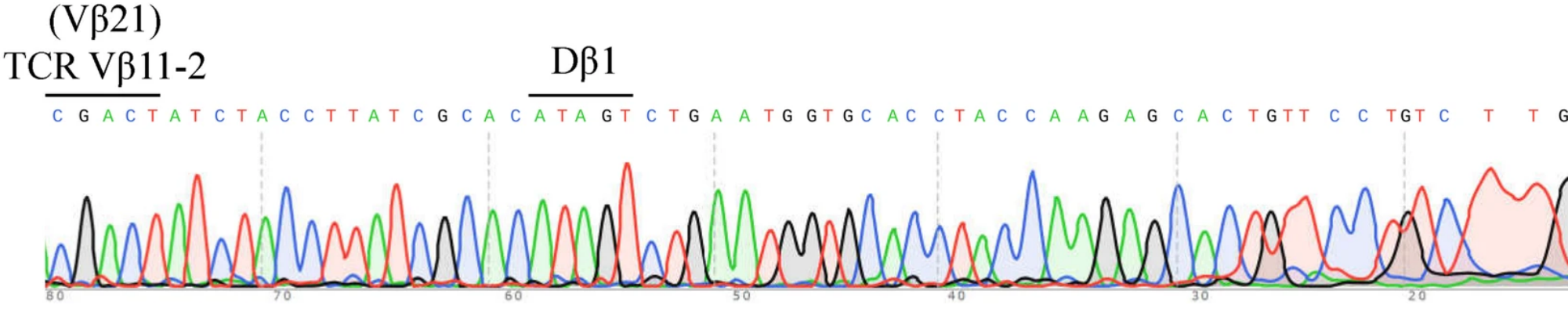
150 – 500nt



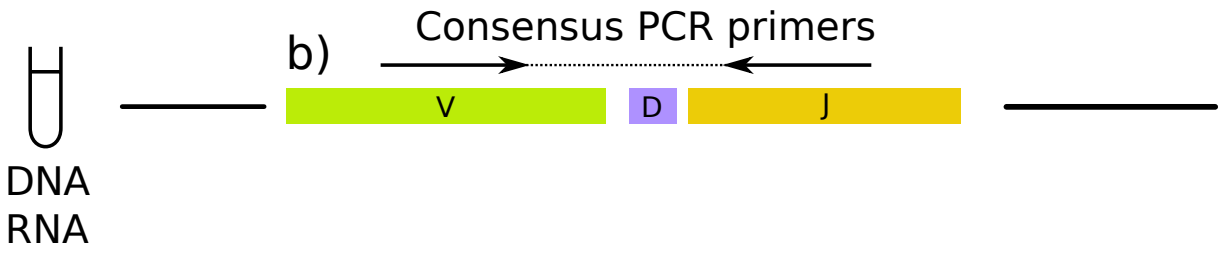
V17 -5/4/-3 J3

The “old-fashioned” strategy



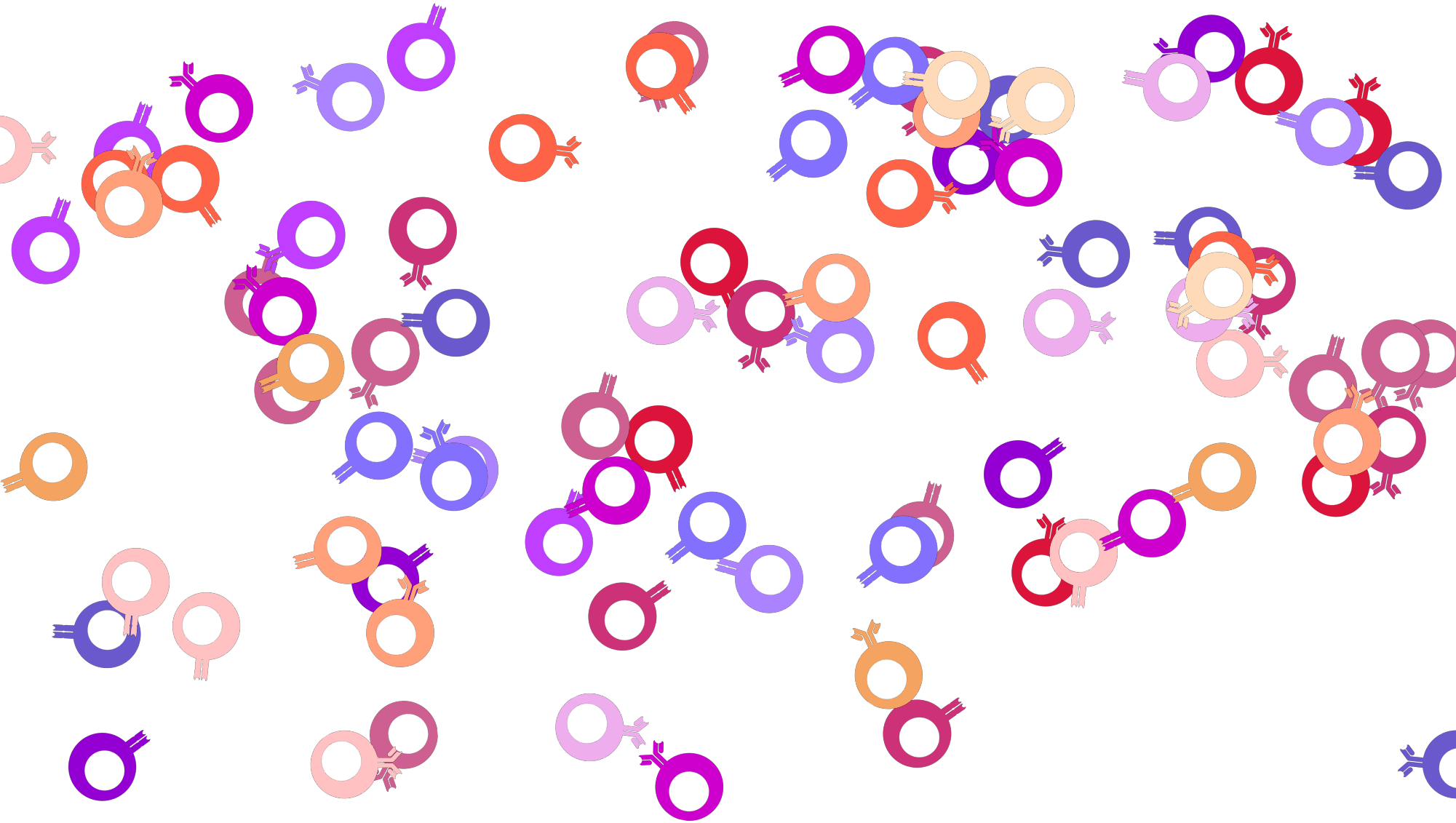


The high-throughput sequencing strategy

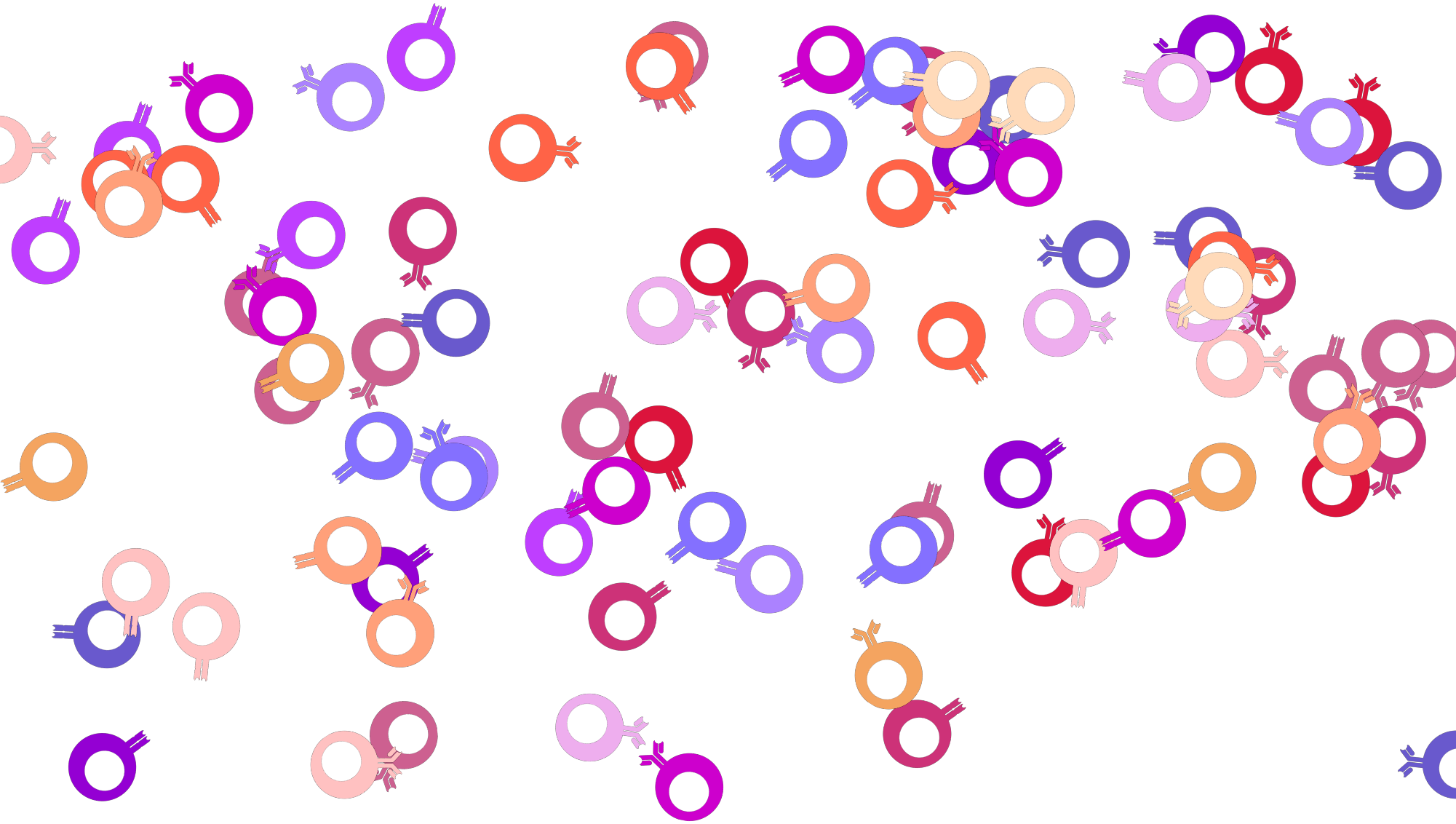


High-throughput sequencing

Studying immune repertoires with *k*-mers



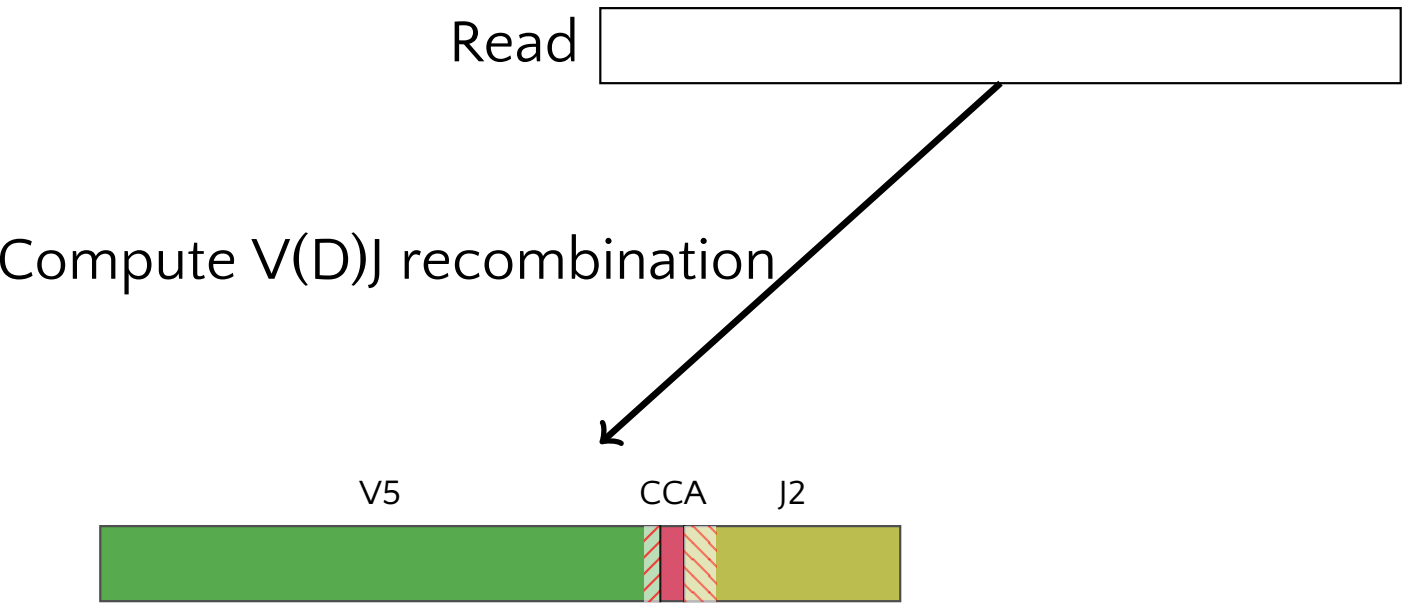
Studying immune repertoires with *k*-mers



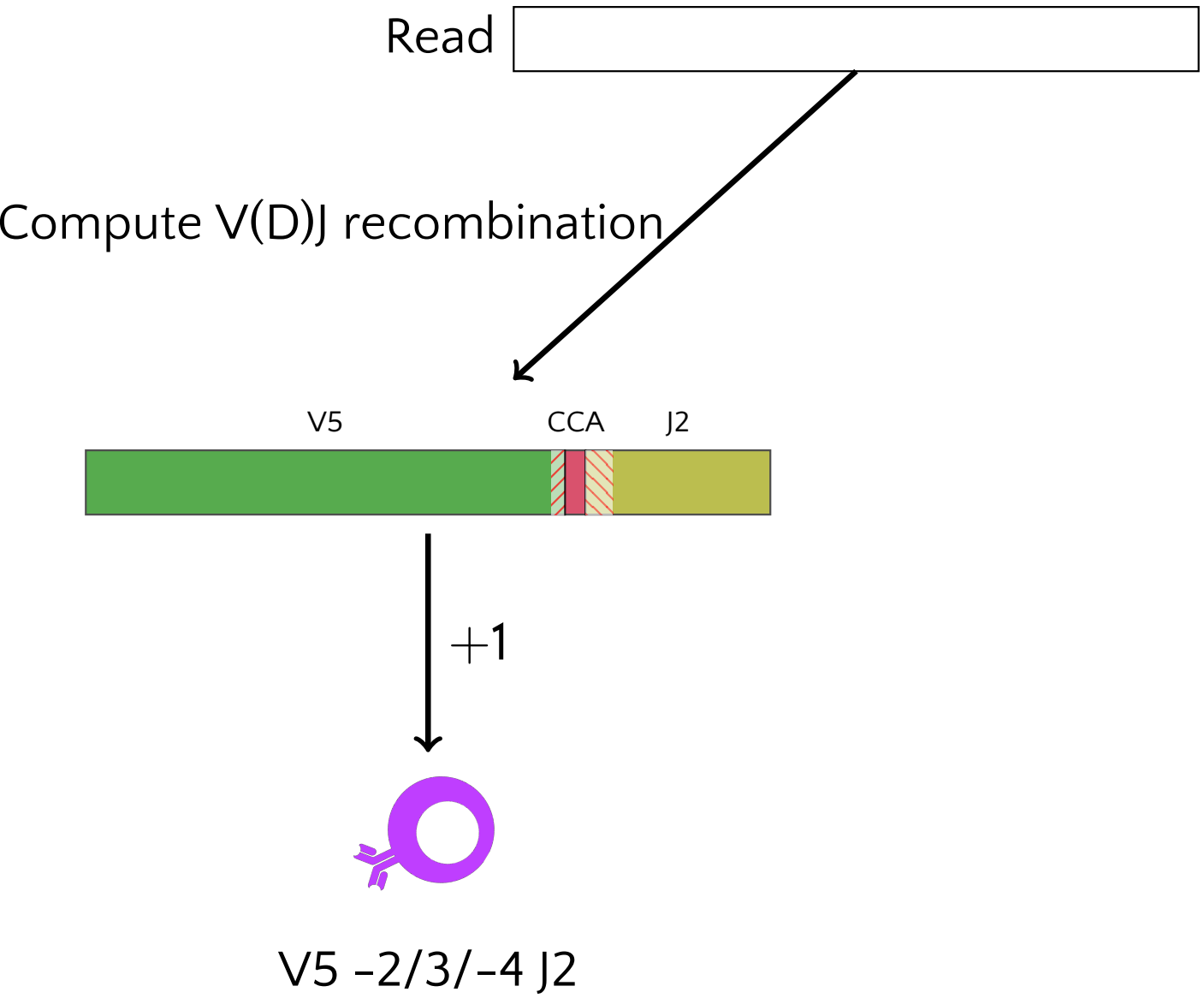
What are the most abundant lymphocytes?

Two solutions to detect V(D)J recombinations

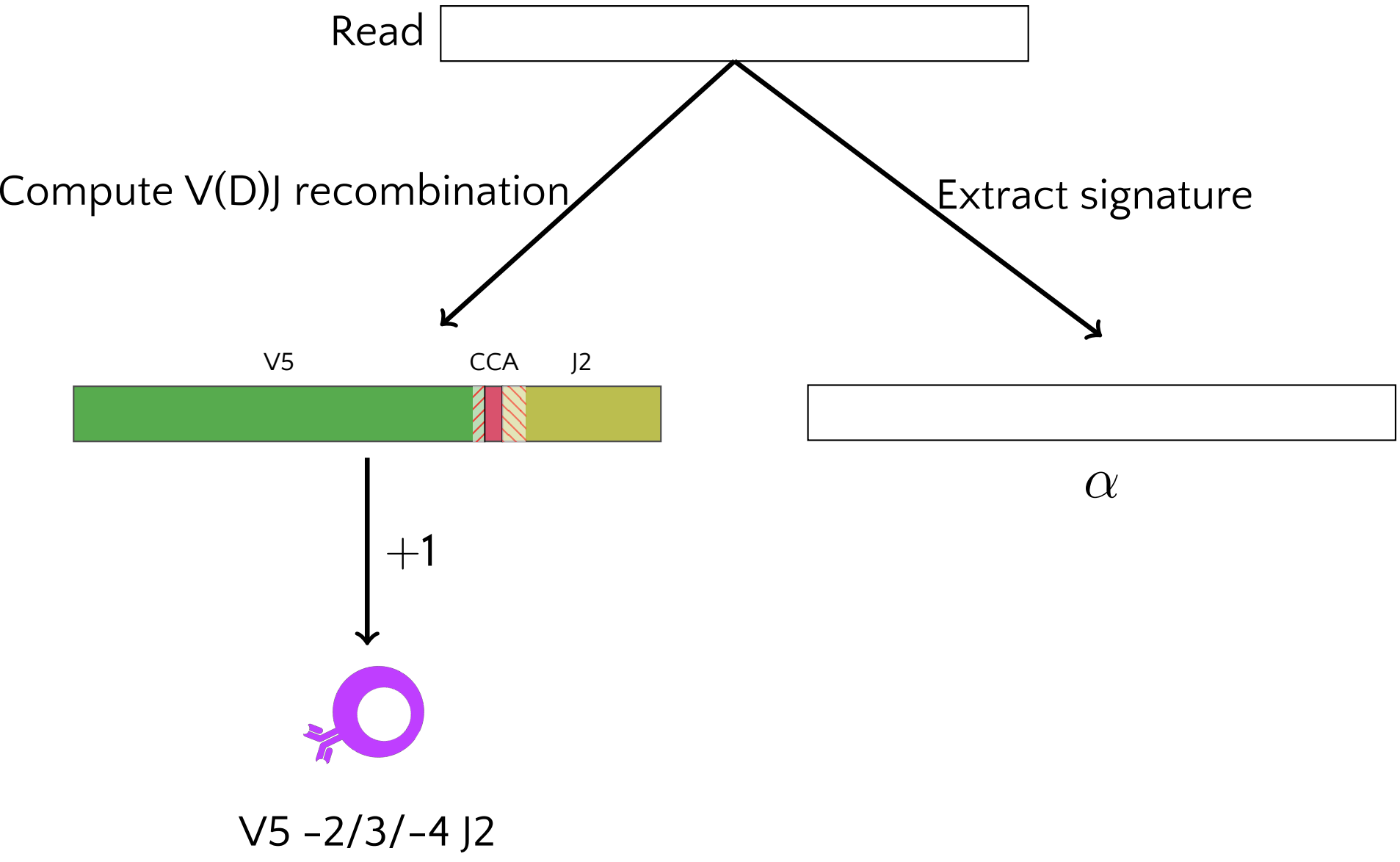
Two solutions to detect V(D)J recombinations



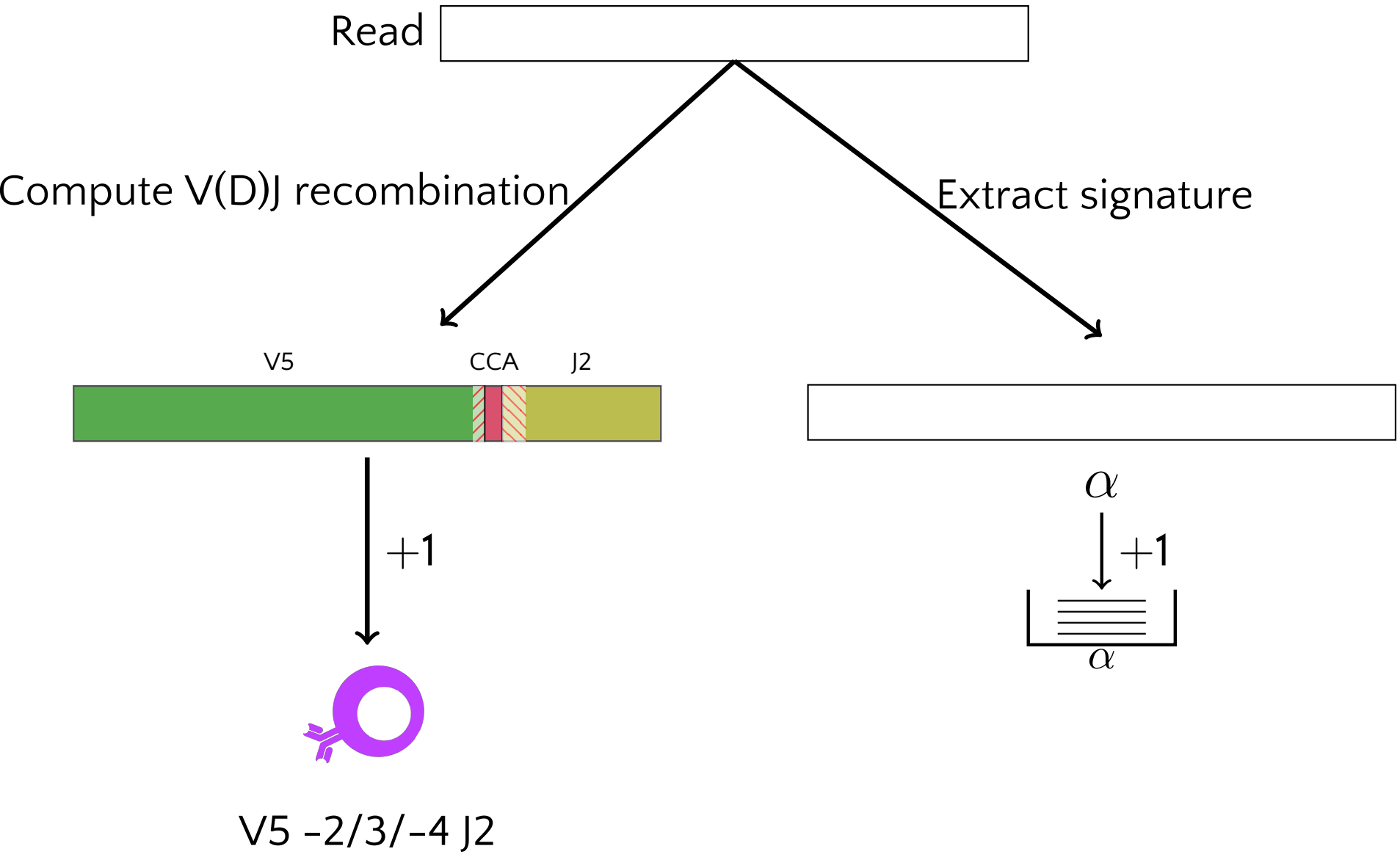
Two solutions to detect V(D)J recombinations



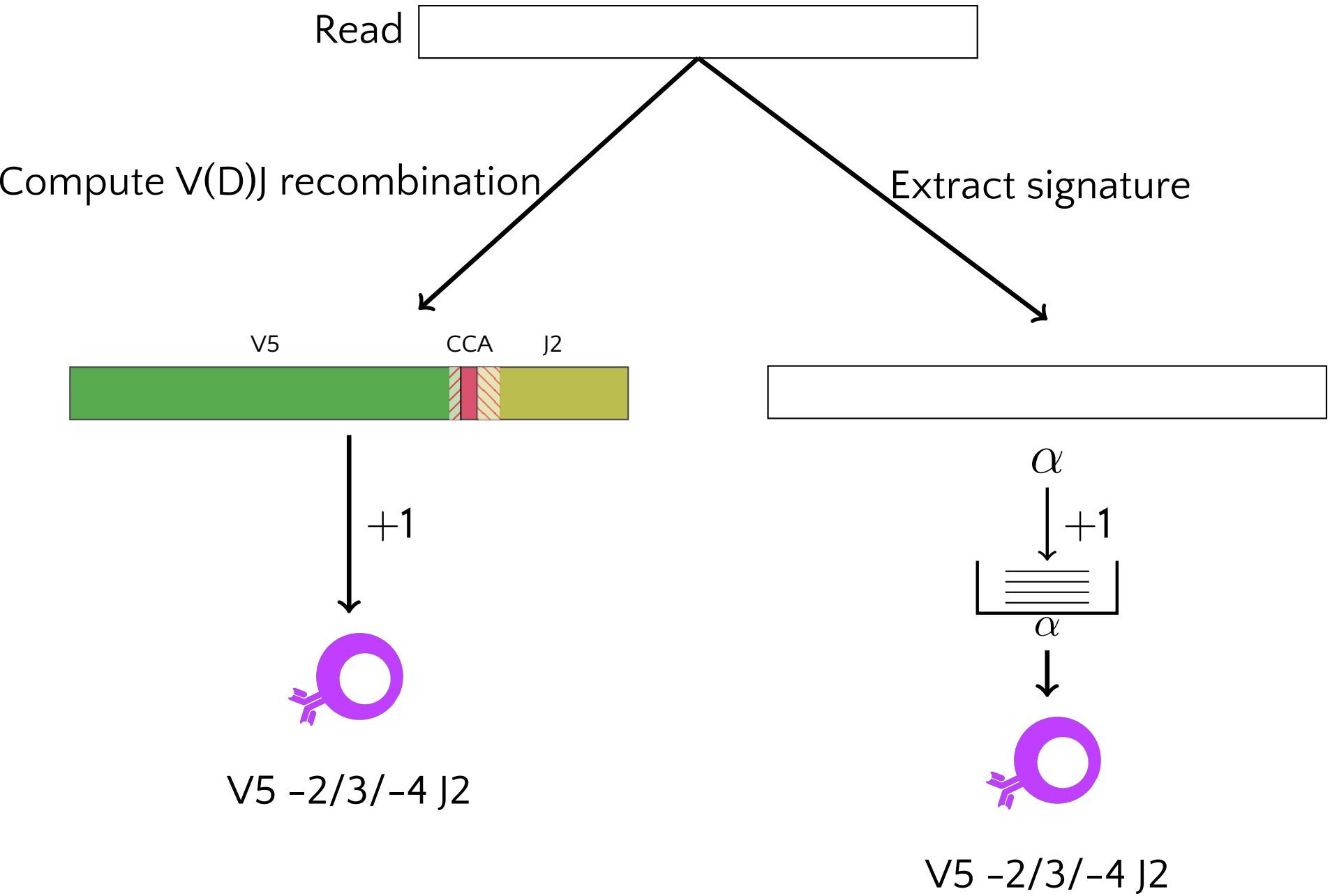
Two solutions to detect V(D)J recombinations



Two solutions to detect V(D)J recombinations



Two solutions to detect V(D)J recombinations



Extracting V(D)J recombination signatures with k-mers

ACACGGCCGTGTATTACTGTGCGAGAGAGCTGAATACTTCCAGCACTGGGGCC

Extracting V(D)J recombination signatures with *k*-mers

parts of V genes

```
ACAC CACG ACGG CGGC GGCC
GCCG TCTT CTTC TTCC TCCA
CCAA CAAC AACC ACCT CCTT
CTTG TTGG TGGA GGAC  ...
```

parts of J genes

```
ATAC TACT ACTG CCAG CAGC
AGCA GCAC TGGG GGGC GGCA
GCAA CAAG AAGA AGAG GAGT
AGTT GTTG TTGG  ...
```

```
ACACGGCCGTGTATTACTGTGCGAGAGAGCTGAATACTTCCAGCACTGGGGCC
```

Extracting V(D)J recombination signatures with *k*-mers

parts of V genes

```
ACAC CACG ACGG CGGC GGCC
GCCG TCTT CTTC TTCC TCCA
CCAA CAAC AACC ACCT CCTT
CTTG TTGG TGGA GGAC  ...
```

parts of J genes

```
ATAC TACT ACTG CCAG CAGC
AGCA GCAC TGGG GGGC GGCA
GCAA CAAG AAGA AGAG GAGT
AGTT GTTG TTGG  ...
```

Extracting V(D)J recombination signatures with *k*-mers

parts of V genes

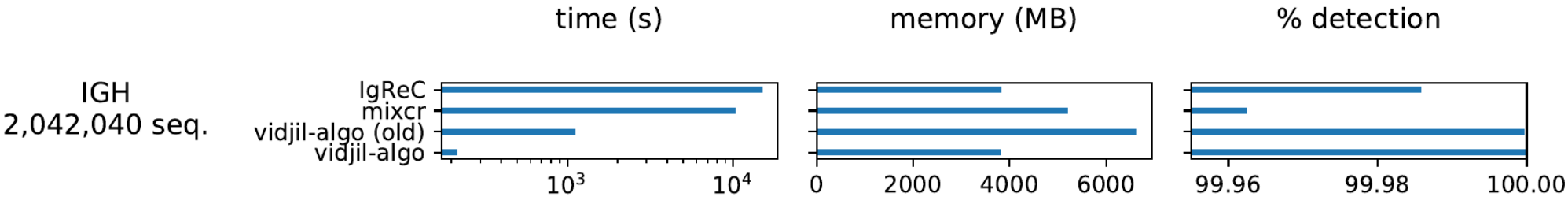
```
ACAC CACG ACGG CGGC GGCC
GCCG TCTT CTTC TTCC TCCA
CCAA CAAC AACC ACCT CCTT
CTTG TTGG TGGA GGAC  ...
```

parts of J genes

```
ATAC TACT ACTG CCAG CAGC
AGCA GCAC TGGG GGGC GGCA
GCAA CAAG AAGA AGAG GAGT
AGTT GTTG TTGG  ...
```

Signature

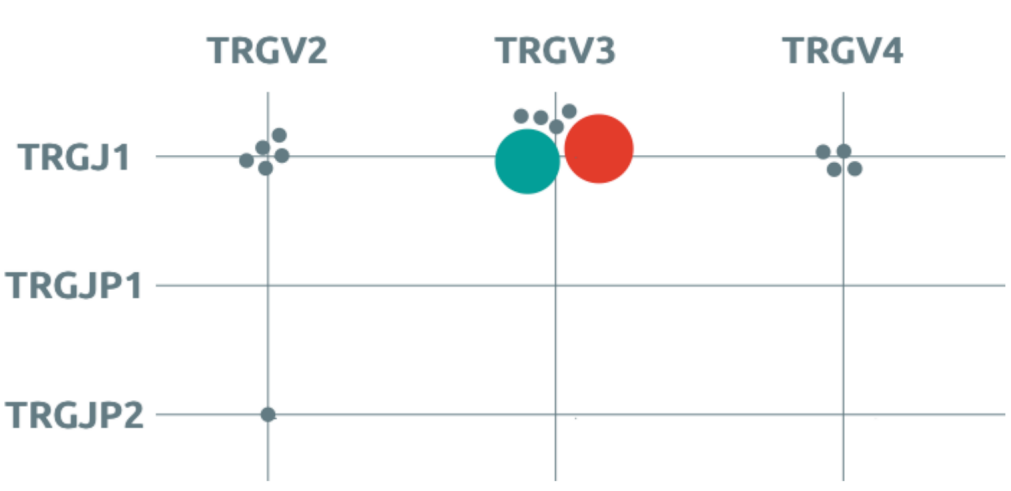
A continuously improving algorithm

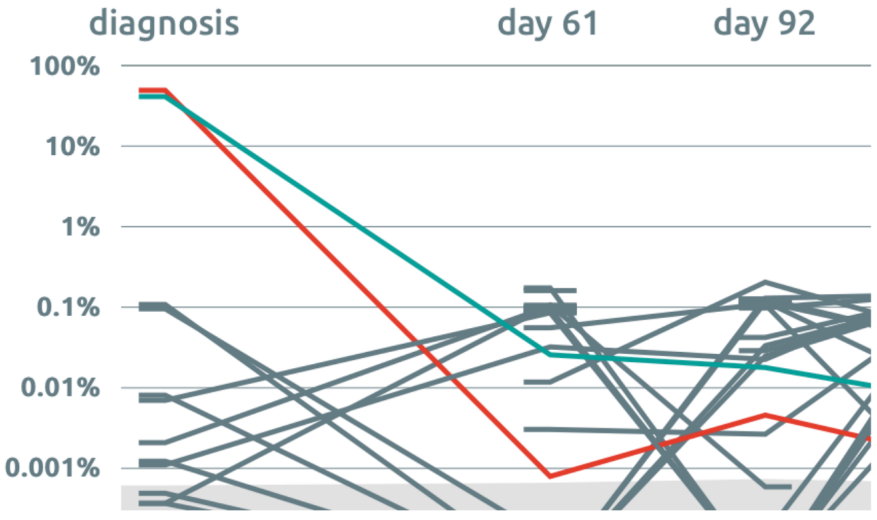


Vidjil – Ultrafast V(D)J recombination detection

A story started in 2011 as a collaboration with Lille hospital

Vidjil – Also a ready-to-use web application





merge align › to IMGT/V-QUEST › to IgBlast › to Blast

4 clones, 738 494 reads (90.53%)

✖ TRGV3 1/11/2 J1	49.30%	★	i	GCCACCTGGGACAGCTCCC-TT-GTTC--ATTATAAGAAACTCTTTGGCAGTG
✖ TRGV3 4/1/2 J1	41.23%	★	i	GCCACCTGGG--A--T--A--T---T--ATTATAAGAAACTCTTTGGCAGTG
✖ TRGV3 3/16/3 J1	0.0021%	★	i	GCCG-CTTGGA-ACCCCAATTTGGTACGGGTTATAAGAAACTCTTTGGCAGTG
✖ TRGV3 5/4/2 J1	+	★	i	GCCACCTGGG---GC--CA-AT---T--A-TA--AGAAACTCTTTGGCAGTG

Vidjil is used throughout the world

A public web server accessible to anyone

`app.vidjil.org`

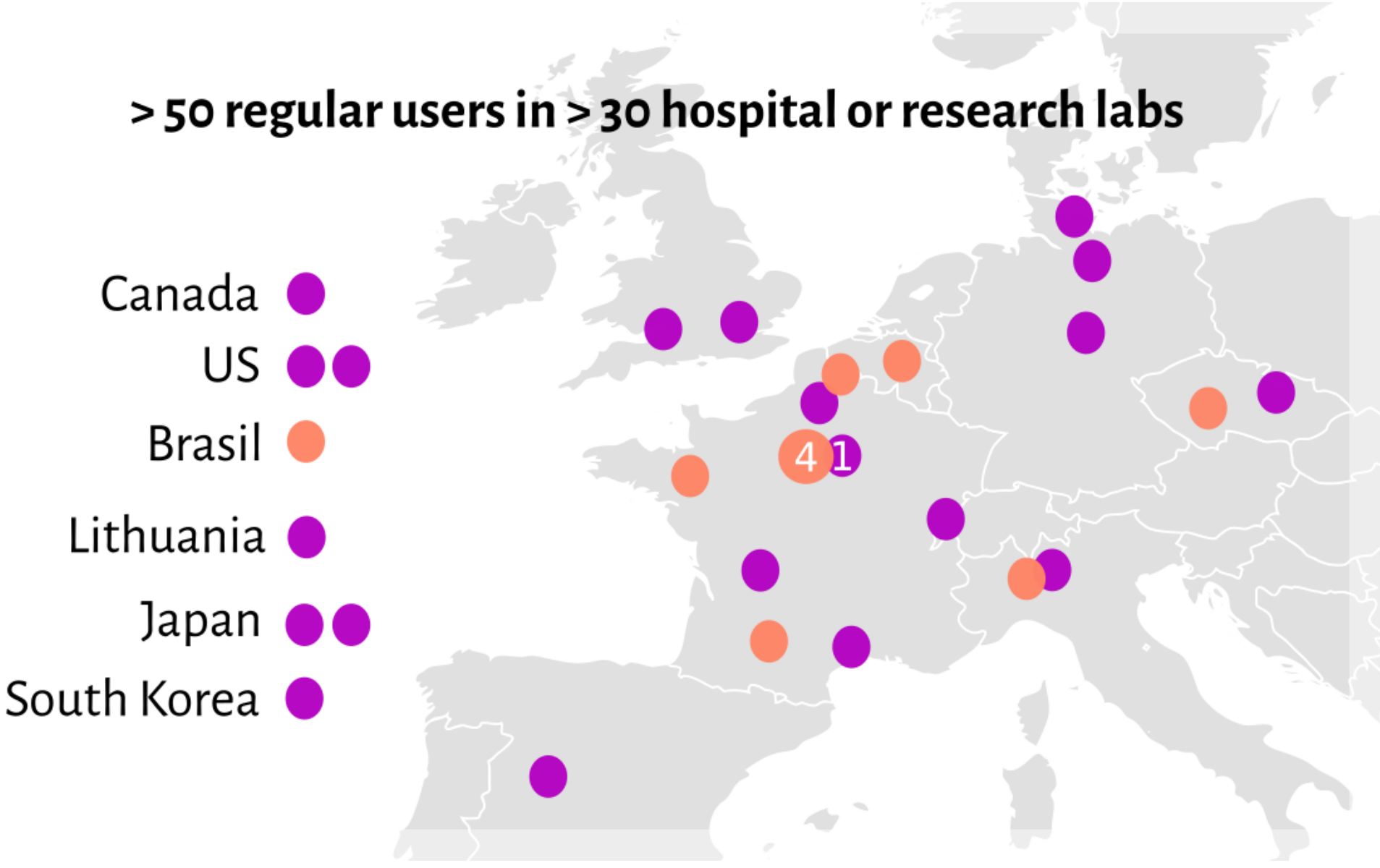
An open-source software

`gitlab.vidjil.org`

A nonprofit Inria consortium to support and enhance the software

`vidjil.net`

Vidjil is used throughout the world



30+ publications using Vidjil

Abdo et al., *Caution encouraged in next-generation sequencing immunogenetic analyses in acute lymphoblastic leukemia*, Blood, 2020

Allain et al., *IGHV segment utilization in immunoglobulin gene rearrangement differentiates patients with anti-myelin-associated glycoprotein neuropathy from others immunoglobulin M-gammopathies*, Haematologica, 2018

Bartram et al., *High throughput sequencing in acute lymphoblastic leukemia reveals clonal architecture of central nervous system and bone marrow compartments*, Haematologica, 2018

Bender et al., *Immunoglobulin variable domain high-throughput sequencing reveals specific novel mutational patterns in POEMS syndrome*, Blood, 2020

Brüggemann et al., *Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study*, Leukemia, 2019

Cavagna et al., *Capture-based Next-Generation Sequencing Improves the Identification of Immunoglobulin/T-Cell Receptor Clonal Markers and Gene Mutations in Adult Acute Lymphoblastic Leukemia Patients Lacking Molecular Probes*, Cancers, 2020

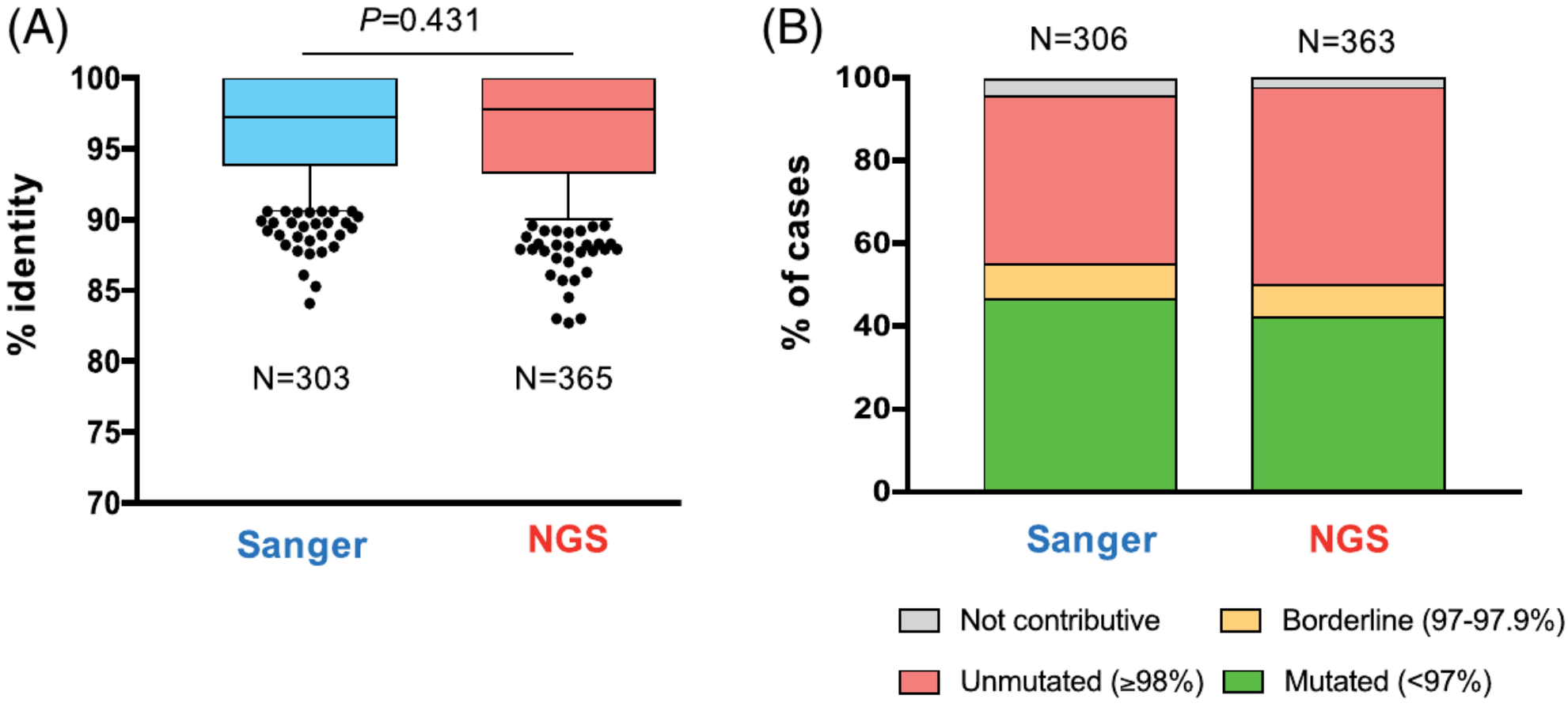
Correia et al., *High-throughput sequencing of immunoglobulin heavy chain for minimal residual disease detection in B-lymphoblastic leukemia*, Int. Journal of Laboratory Hematology, 2021

Davi et al., on behalf of ERIC, and the EuroClonality-NGS Working Group, *Immunoglobulin gene analysis in chronic lymphocytic leukemia in the era of next generation sequencing*, Leukemia, 2020

...

(full list on `vidjil.org`)

An example: Sanger/Vidjil comparison for CLL



Does high-throughput sequencing (+ Vidjil) provides similar results to Sanger?

Vidjil – from string algorithmics to clinical practice

Efficient string algorithms (really) matter (as well as strong interdisciplinary collaborations)

We need good theoretical research to conceive practical tools

Ongoing: detecting any type of unusual V(D)J recombinations (eg. $V_H V_\alpha J_H$)

Perspective: failure-aware heuristics