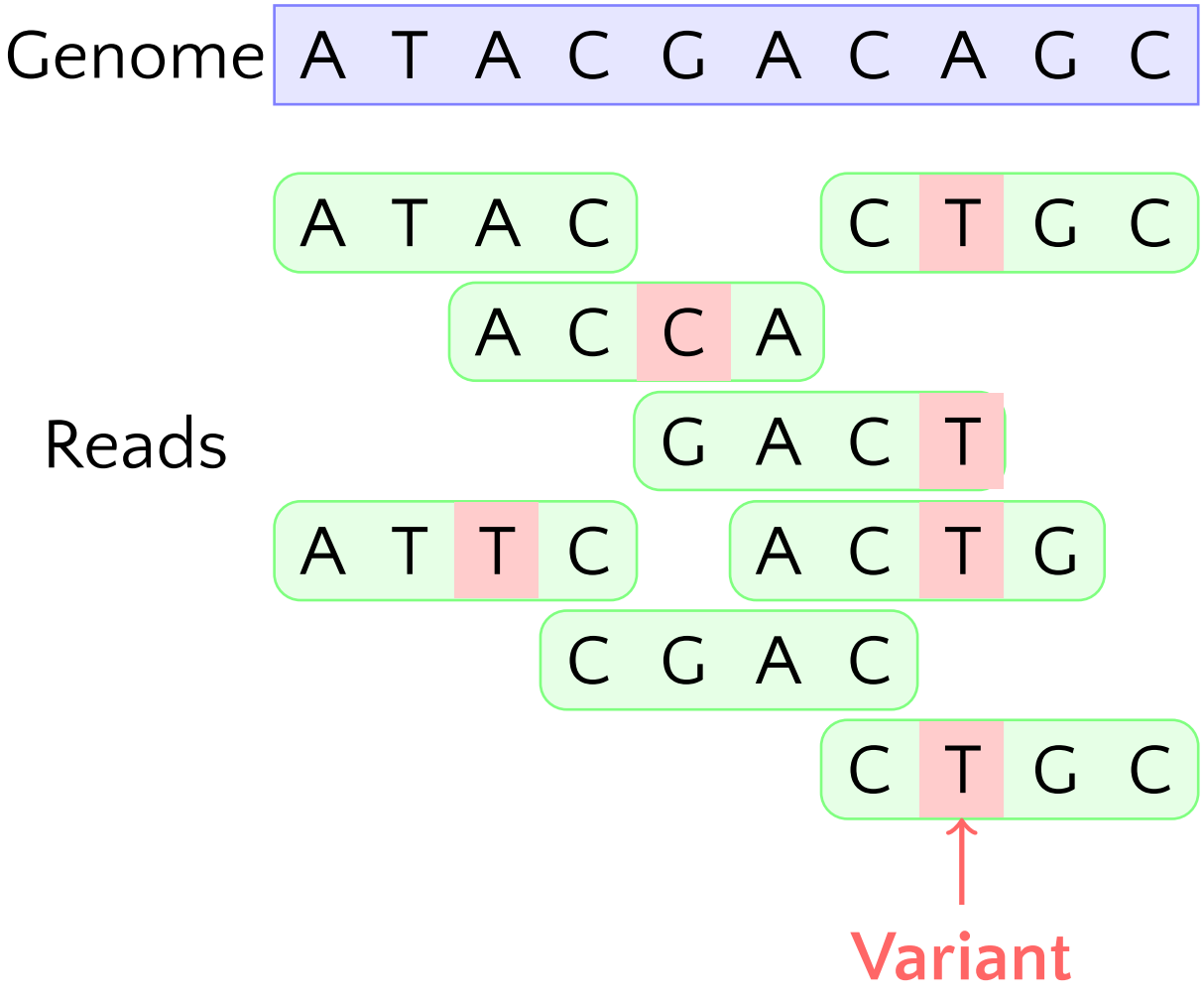
How string algorithms process high-throughput sequencing data in clinical practice

Mikaël Salson

CRIStAL (CNRS, U. Lille)

A classical task: variant calling



Efficient string algorithms matter

Let's compare two sequences

A T A C T G A

T A C G A C

Efficient string algorithms matter

Let's compare two sequences

A T A C T G A

T A C G A C

The optimal* solution is

A T A C T G A
  T A C – G A C

* Optimal in terms of minimising the number of differences

Efficient string algorithms matter

Let's compare two sequences

A T A C T G A

T A C G A C

The optimal* solution is

A T A C T G A
T A C – G A C

To find it we need to compute all possibilities

* Optimal in terms of minimising the number of differences

Efficient string algorithms matter

Let's compare two sequences

A T A C T G A

T A C G A C

The optimal* solution is

A T A C T G A
T A C - G A C

To find it we need to compute all possibilities

This takes $\geq$**30 microseconds** for two 300nt sequences

* Optimal in terms of minimising the number of differences

Sequencing data is produced at a (very) high throughput

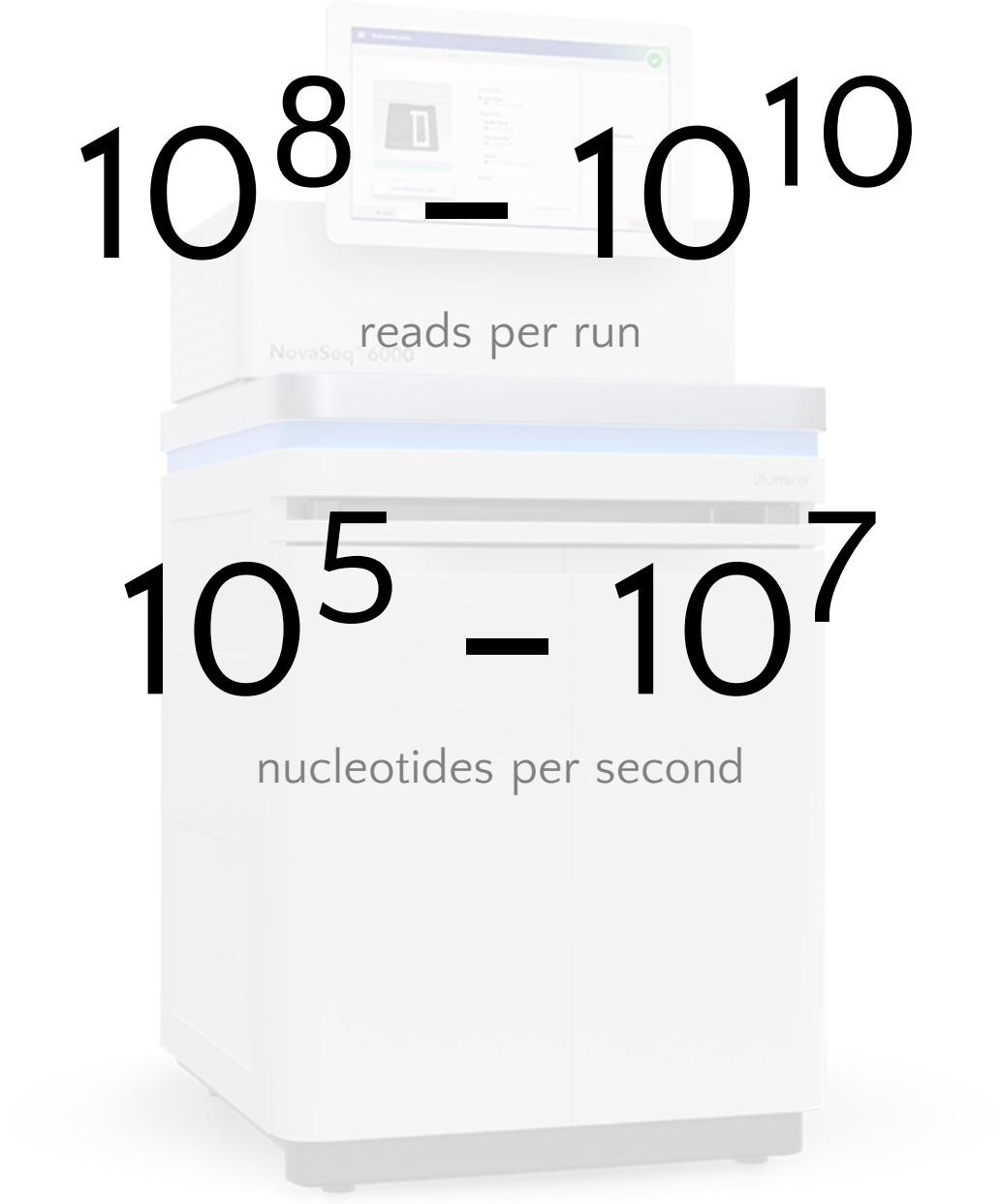
$10^8 - 10^{10}$

reads per run

Sequencing data is produced at a (very) high throughput

$10^8 - 10^{10}$

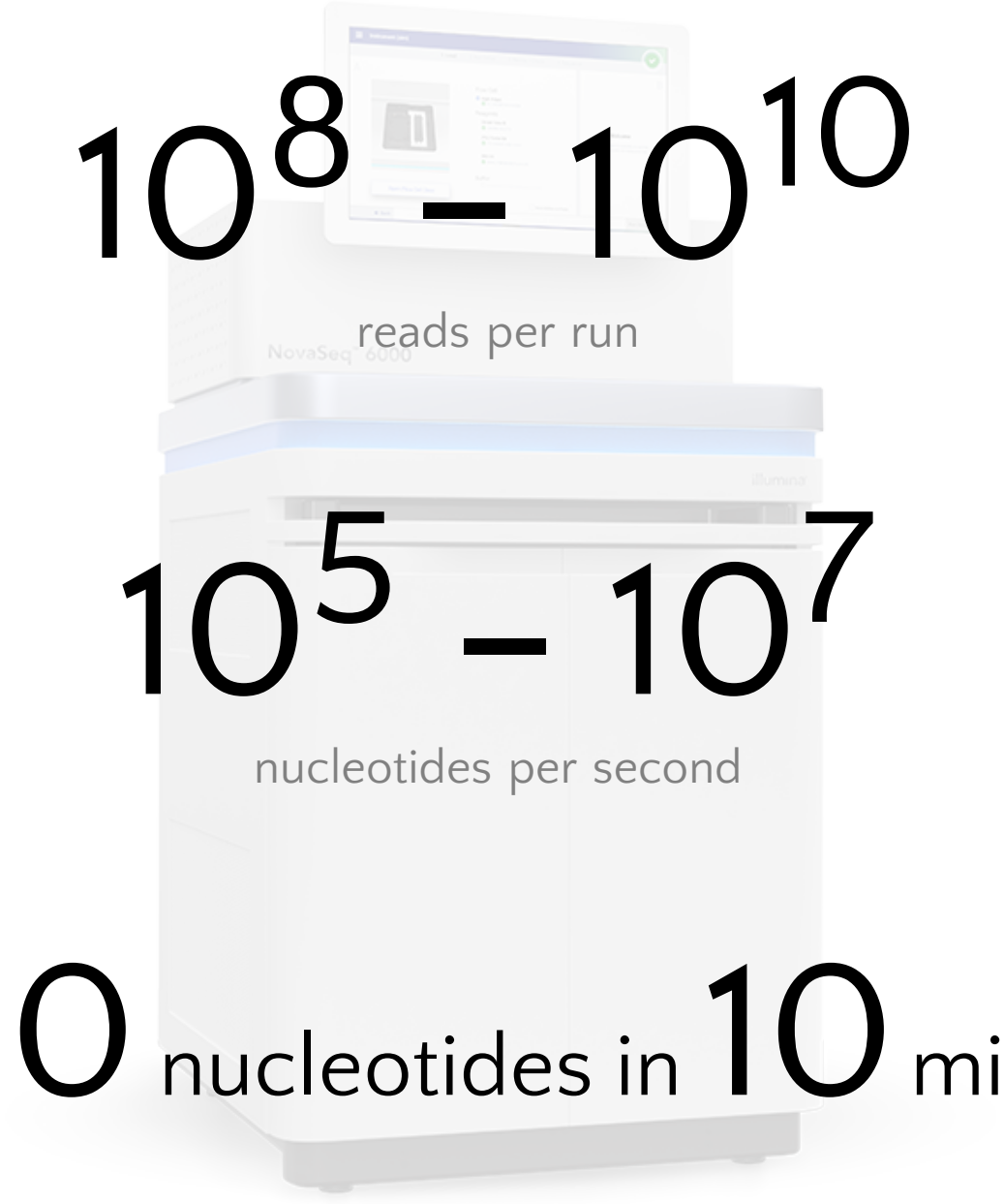
reads per run



$10^5 - 10^7$

nucleotides per second

Sequencing data is produced at a (very) high throughput



$10^8 - 10^{10}$

reads per run

$10^5 - 10^7$

nucleotides per second

up to 300 nucleotides in 10 microseconds

Comparing sequences could be longer than sequencing

Processing the data could be
10 times longer than sequencing

Comparing sequences could be longer than sequencing

Processing the data could be
10 times longer than sequencing

It's not. How?

How to process 10^5 reads per second?

How to process 10^5 reads per second?

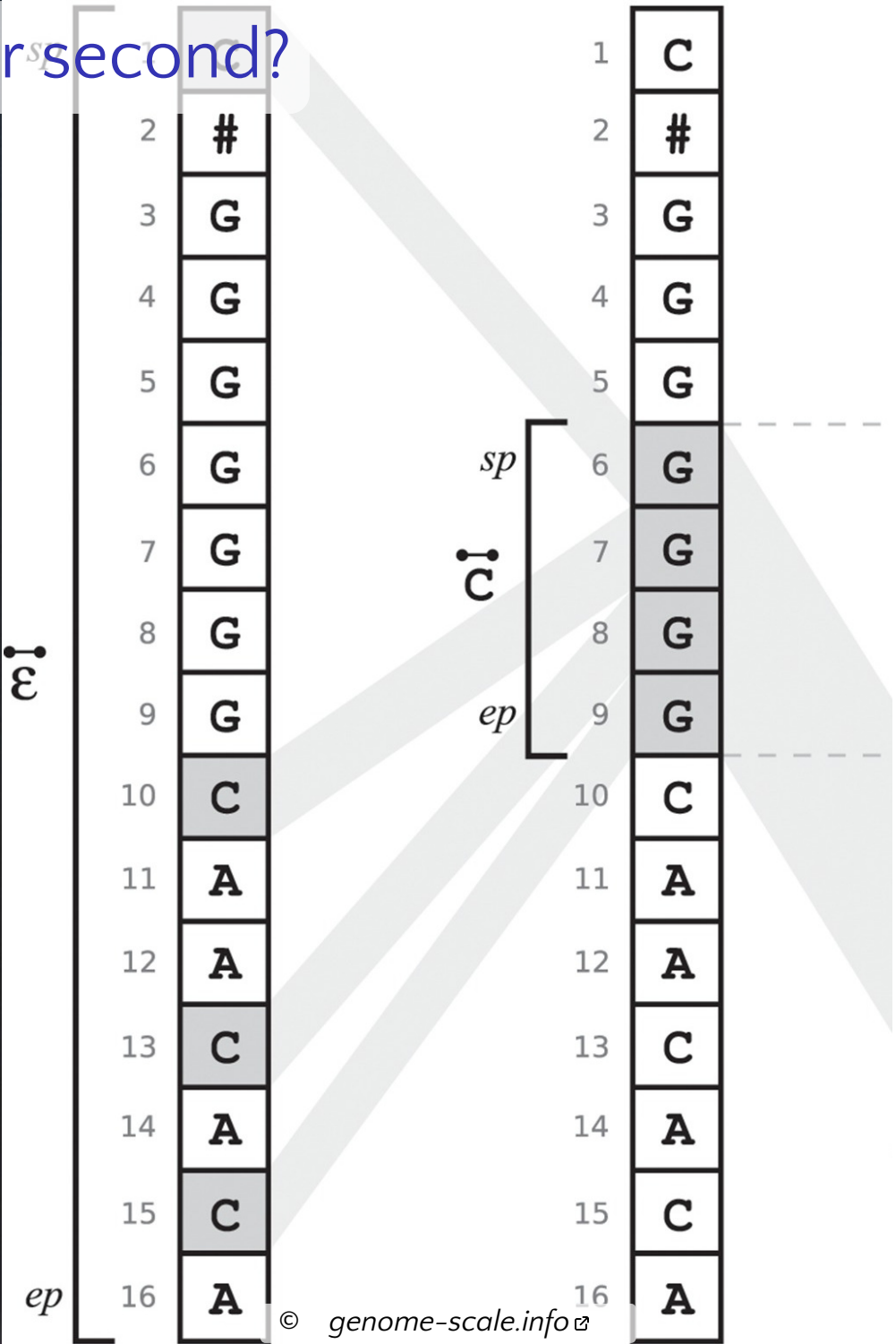
How to process 10^5 reads per second?

Costly
Wasteful
Not very challenging

How to process 10^5 reads per second?

Costly
Wasteful
Not very challenging

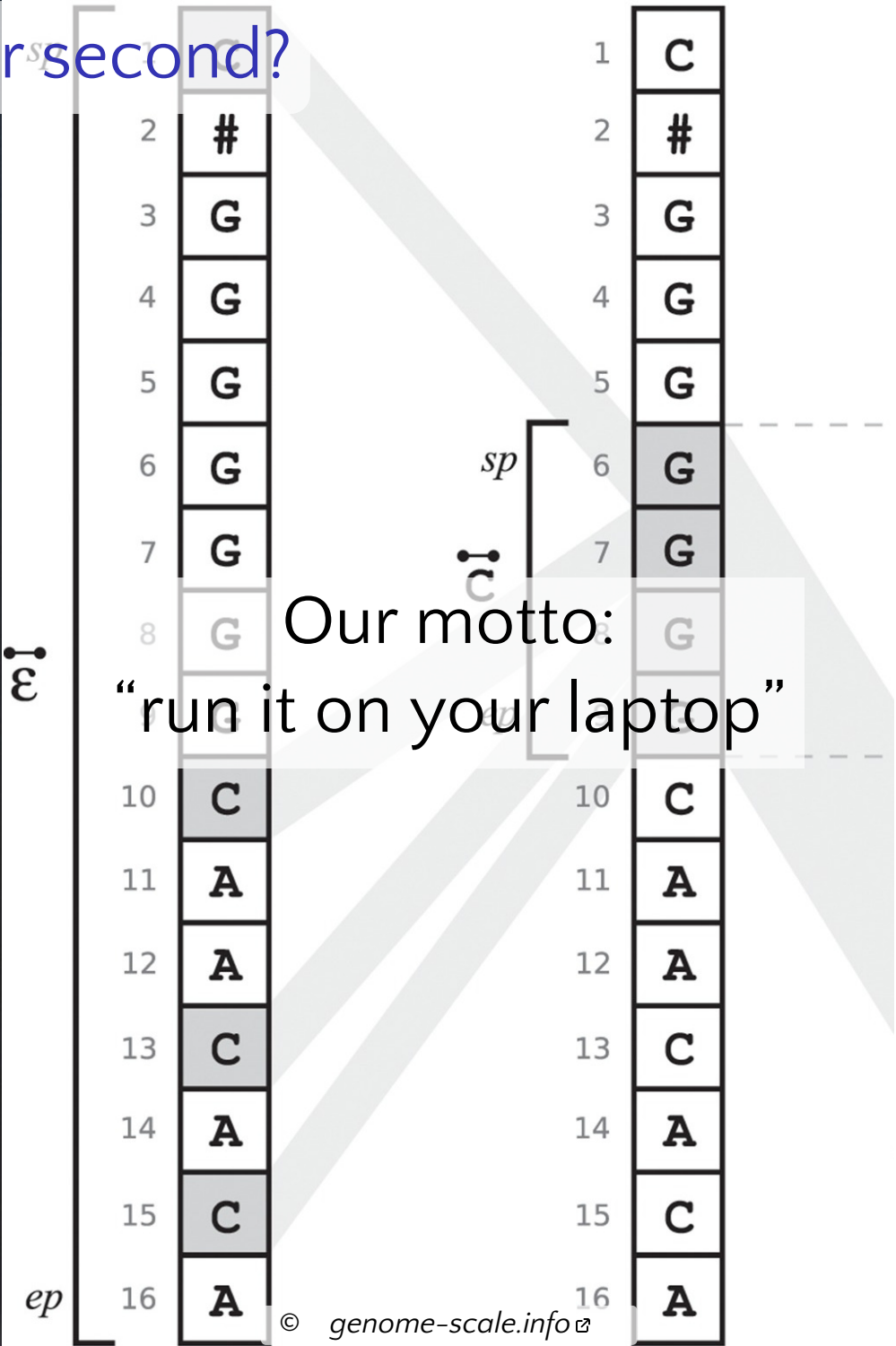
CC BY SA *Victorgrigas* Wikimedia Commons



How to process 10^5 reads per second?

Costly
Wasteful
Not very challenging

CC BY SA *Victorgrigas* Wikimedia Commons



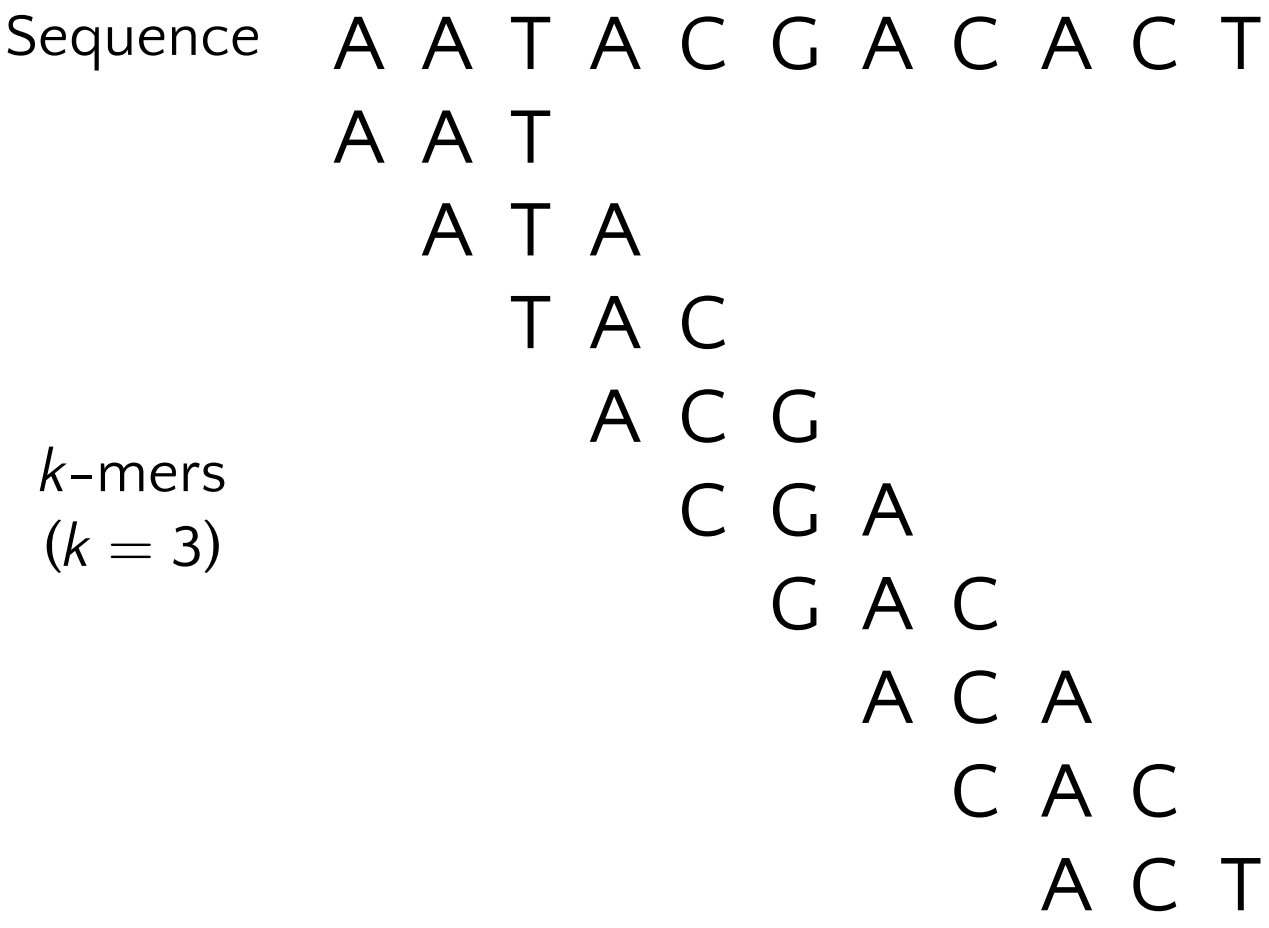
Our motto:
"run it on your laptop"

© *genome-scale.info*

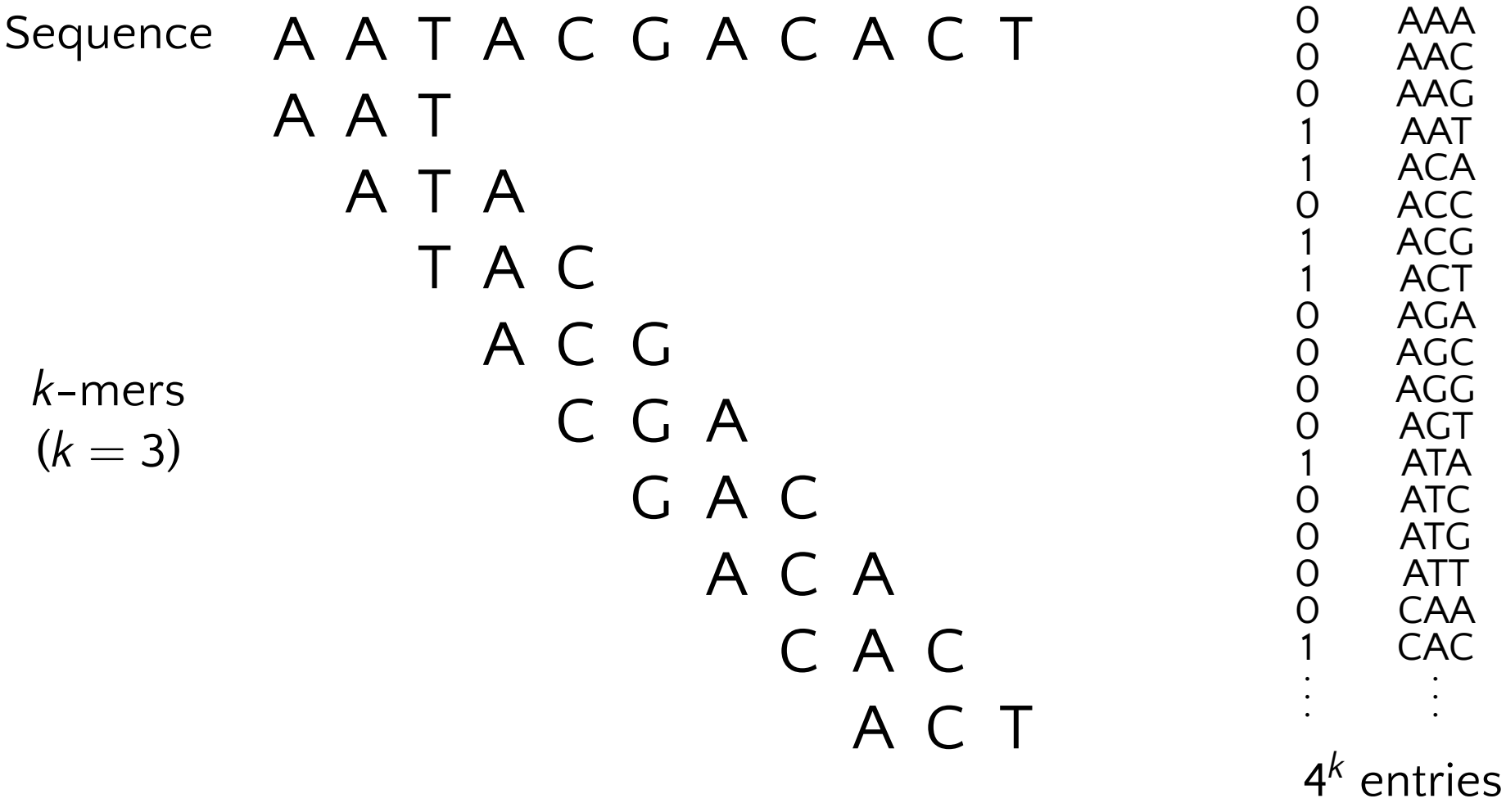
One solution: splitting sequences in k-mers

Sequence A A T A C G A C A C T

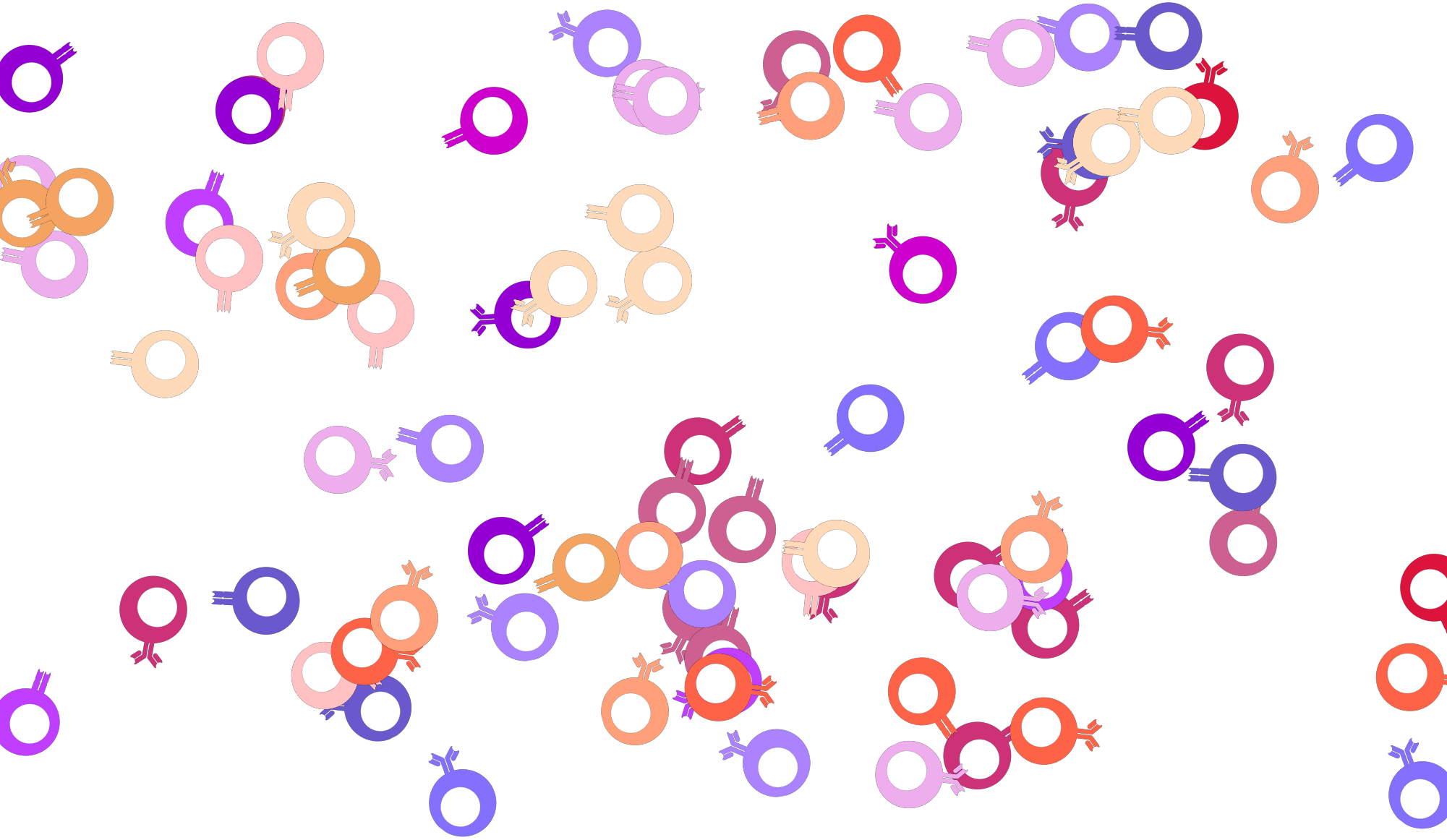
One solution: splitting sequences in k-mers



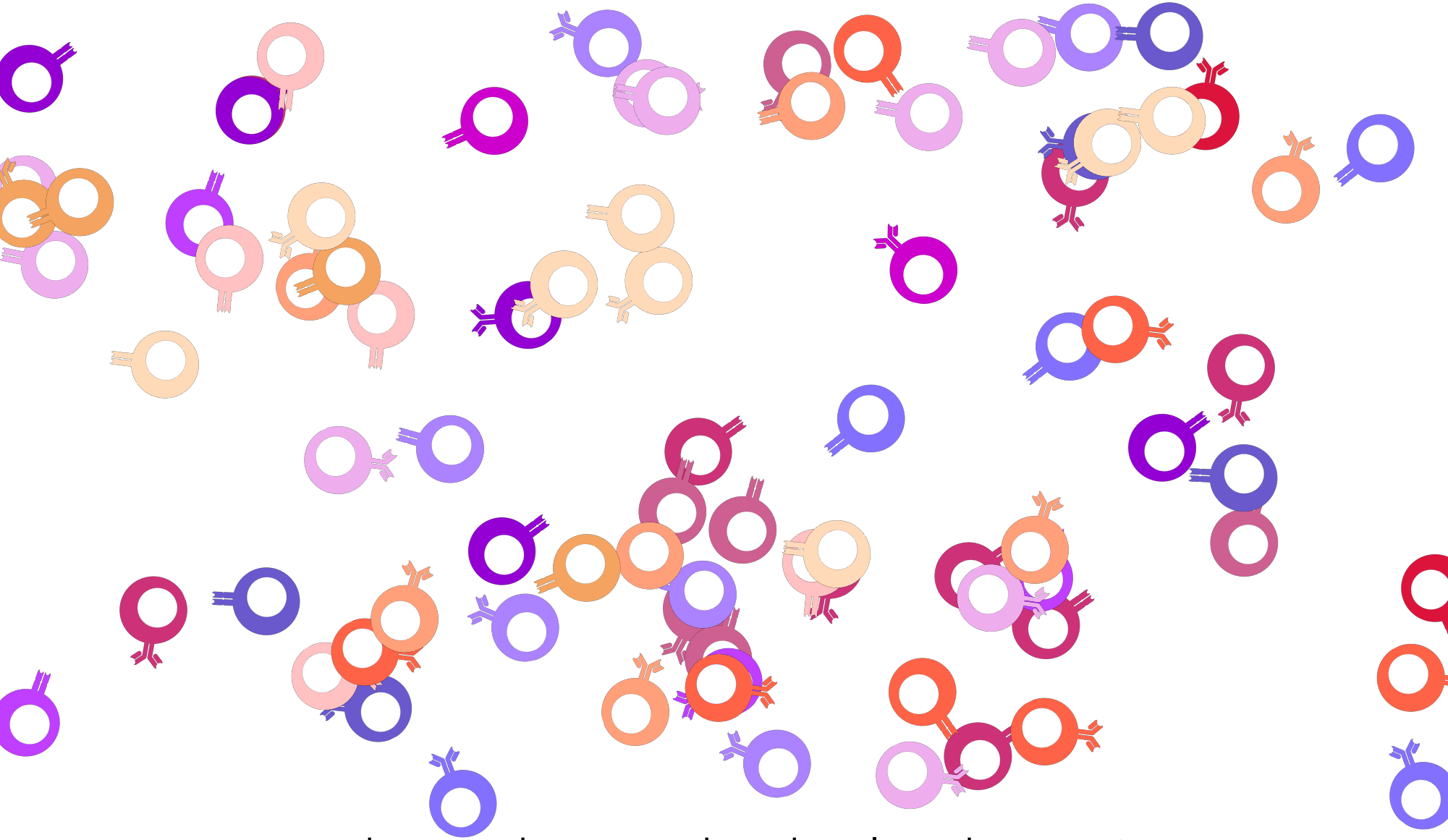
One solution: splitting sequences in k-mers



Studying immune repertoires with *k*-mers



Studying immune repertoires with *k*-mers

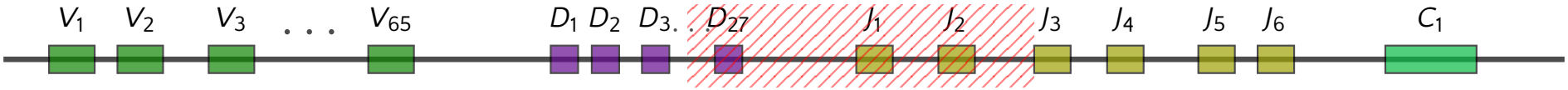


What are the most abundant lymphocytes?

Counting lymphocytes through their V(D)J recombinations

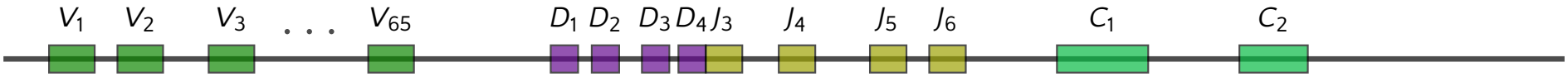
Counting lymphocytes through their V(D)J recombinations

On a lymphoblast genome…



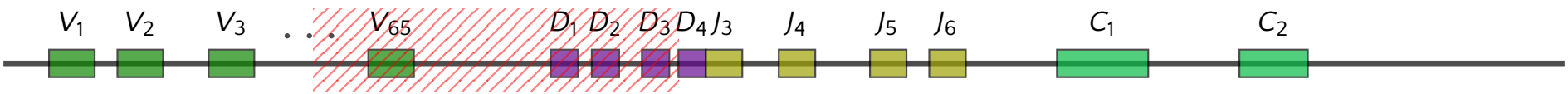
Counting lymphocytes through their V(D)J recombinations

On a lymphoblast genome…



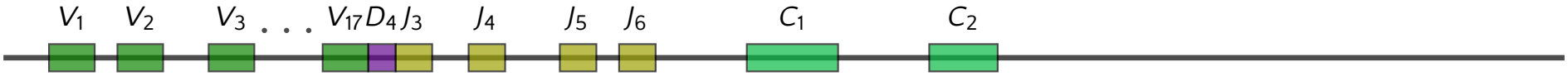
Counting lymphocytes through their V(D)J recombinations

On a lymphoblast genome…



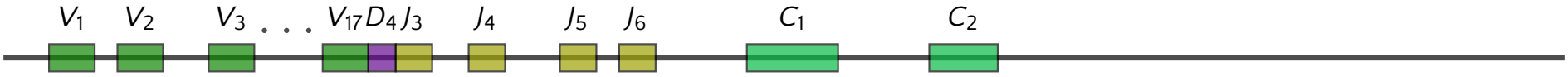
Counting lymphocytes through their V(D)J recombinations

On a lymphoblast genome…

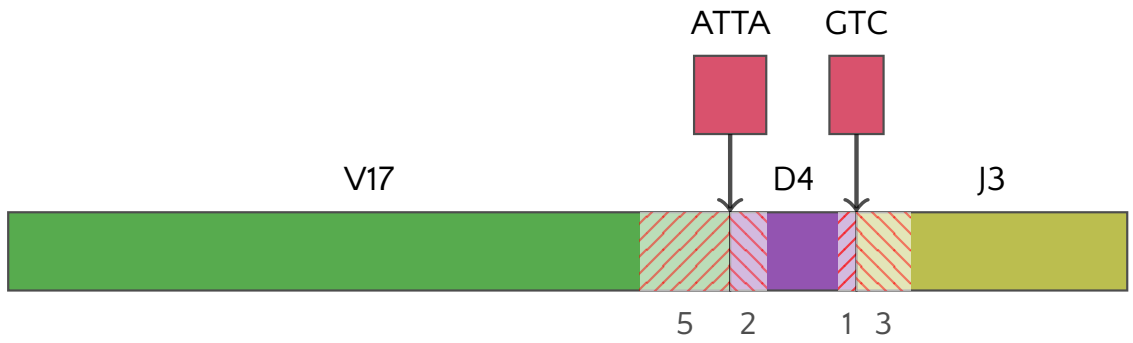


Counting lymphocytes through their V(D)J recombinations

On a lymphoblast genome…

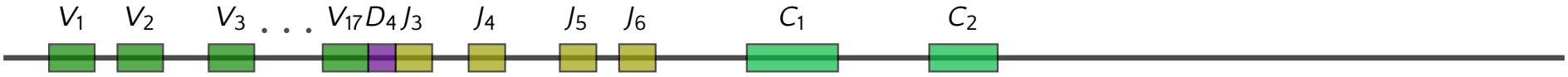


150 – 500nt

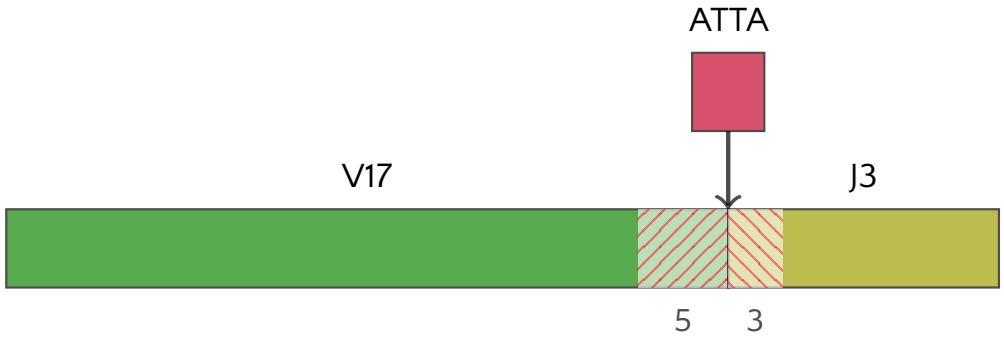


Counting lymphocytes through their V(D)J recombinations

On a lymphoblast genome…

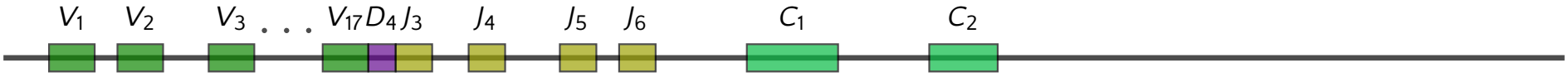


150 – 500nt

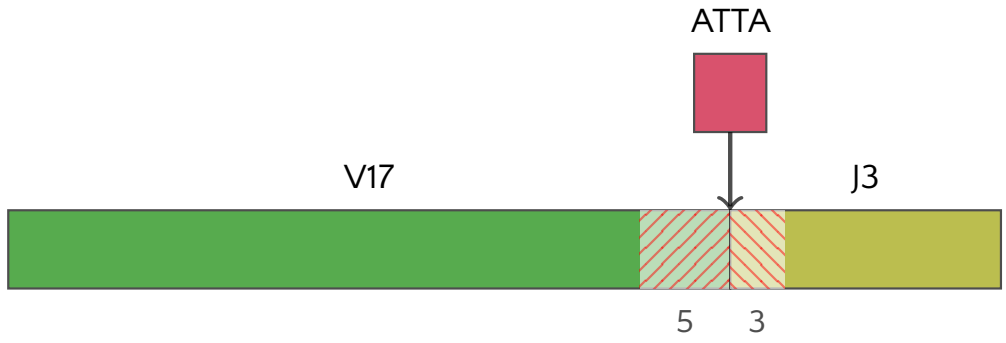


Counting lymphocytes through their V(D)J recombinations

On a lymphoblast genome…



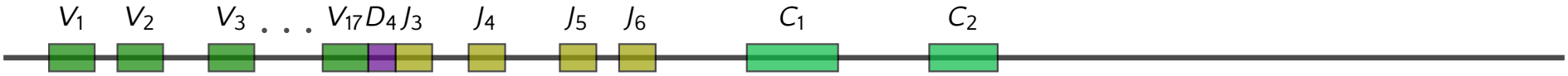
150 – 500nt

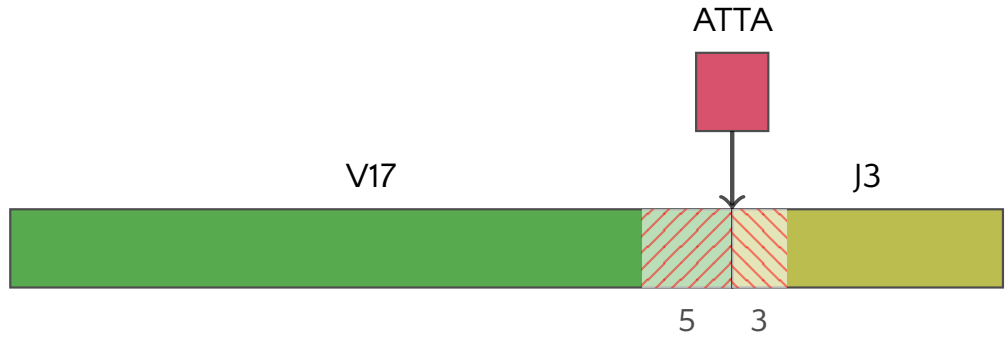


V17 -5/4/-3 J3

Counting lymphocytes through their V(D)J recombinations

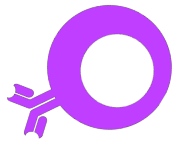
On a lymphoblast genome…



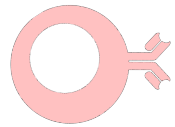


150 – 500nt

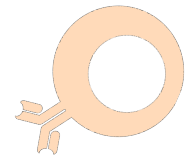
V17 -5/4/-3 J3



V5 -2/3/-4 J2
1.2 %



V3 -1/12/-1 J5
0.7 %

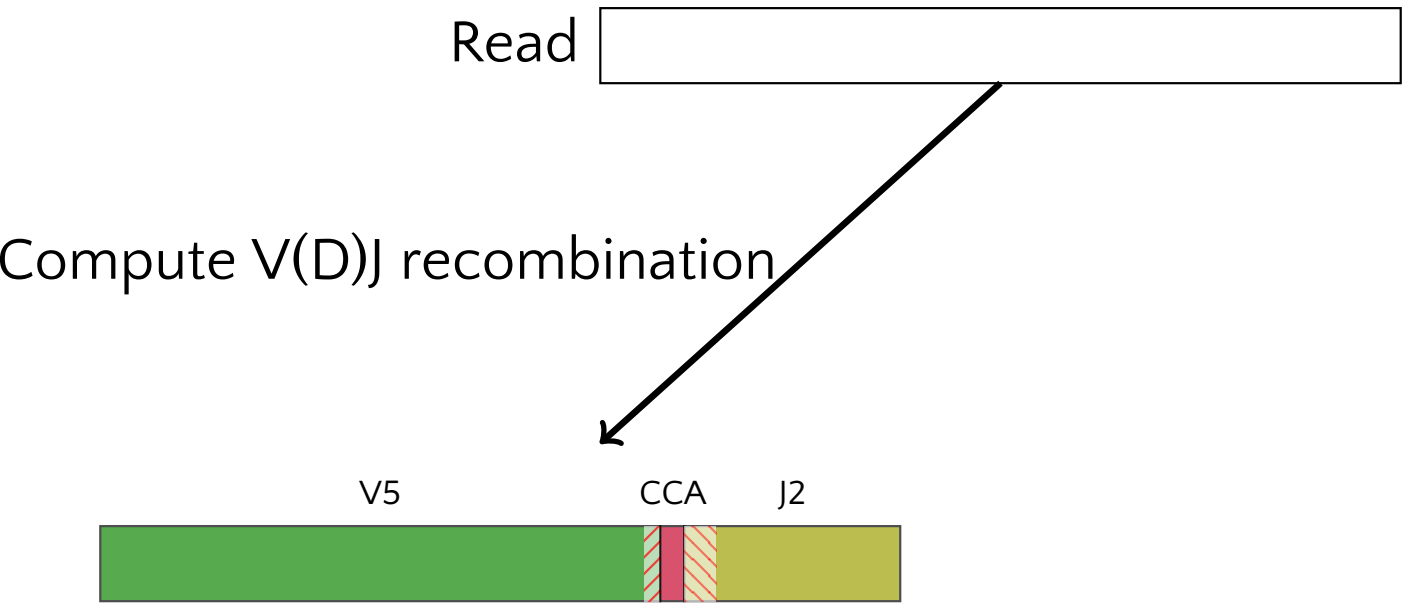


V1 -2/0/-7 J3
0.9 %

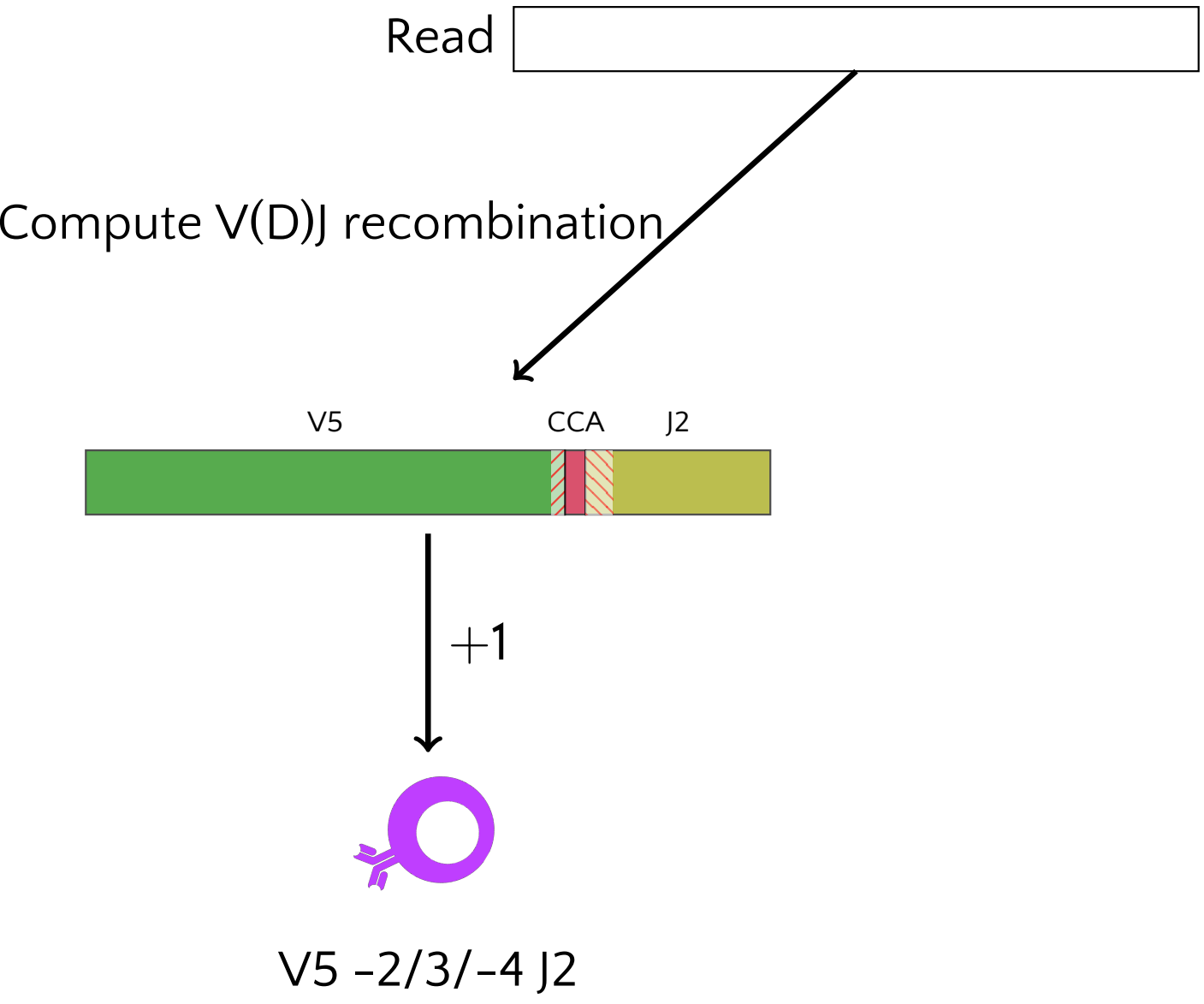
. . .

Two solutions to detect V(D)J recombinations

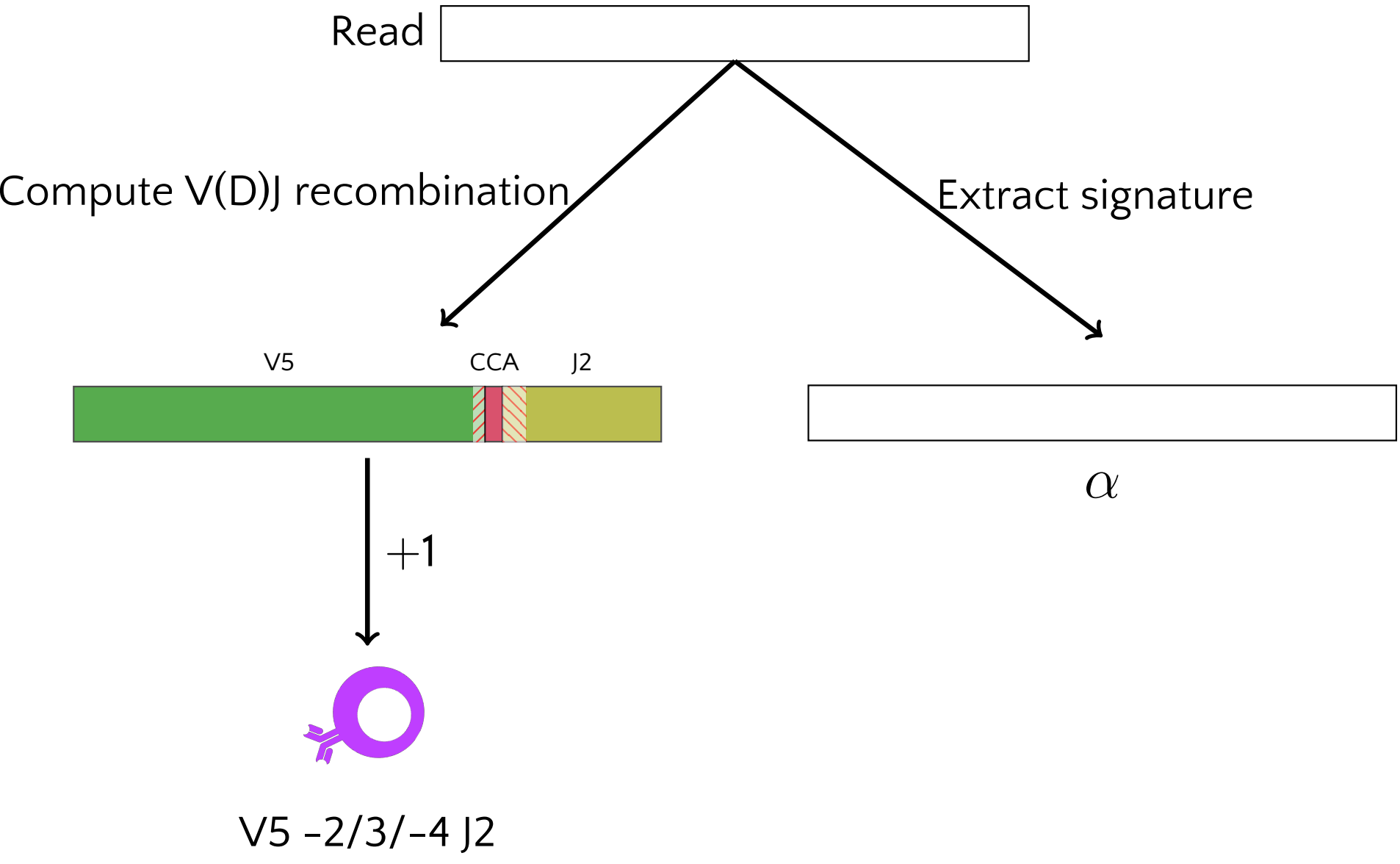
Two solutions to detect V(D)J recombinations



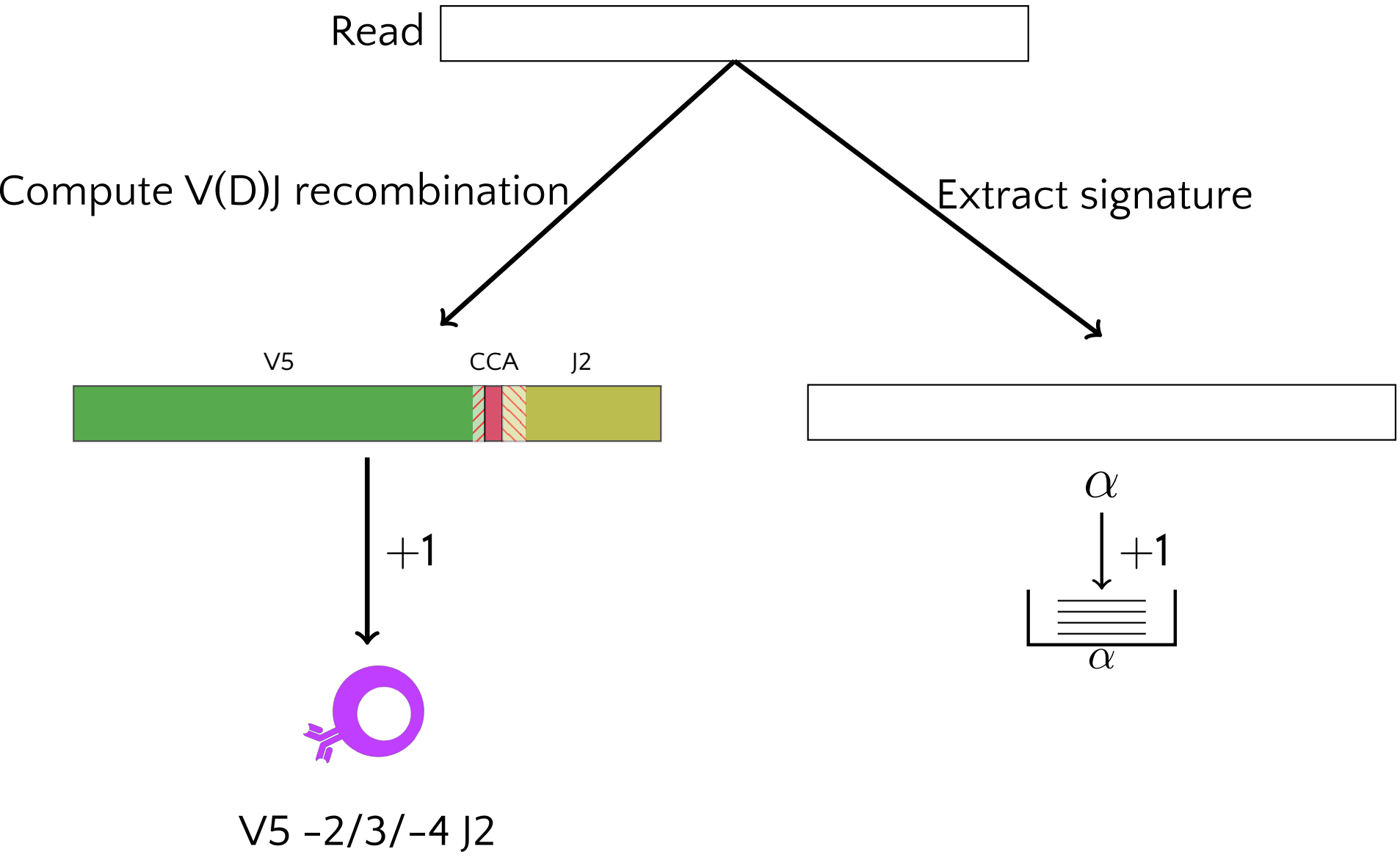
Two solutions to detect V(D)J recombinations



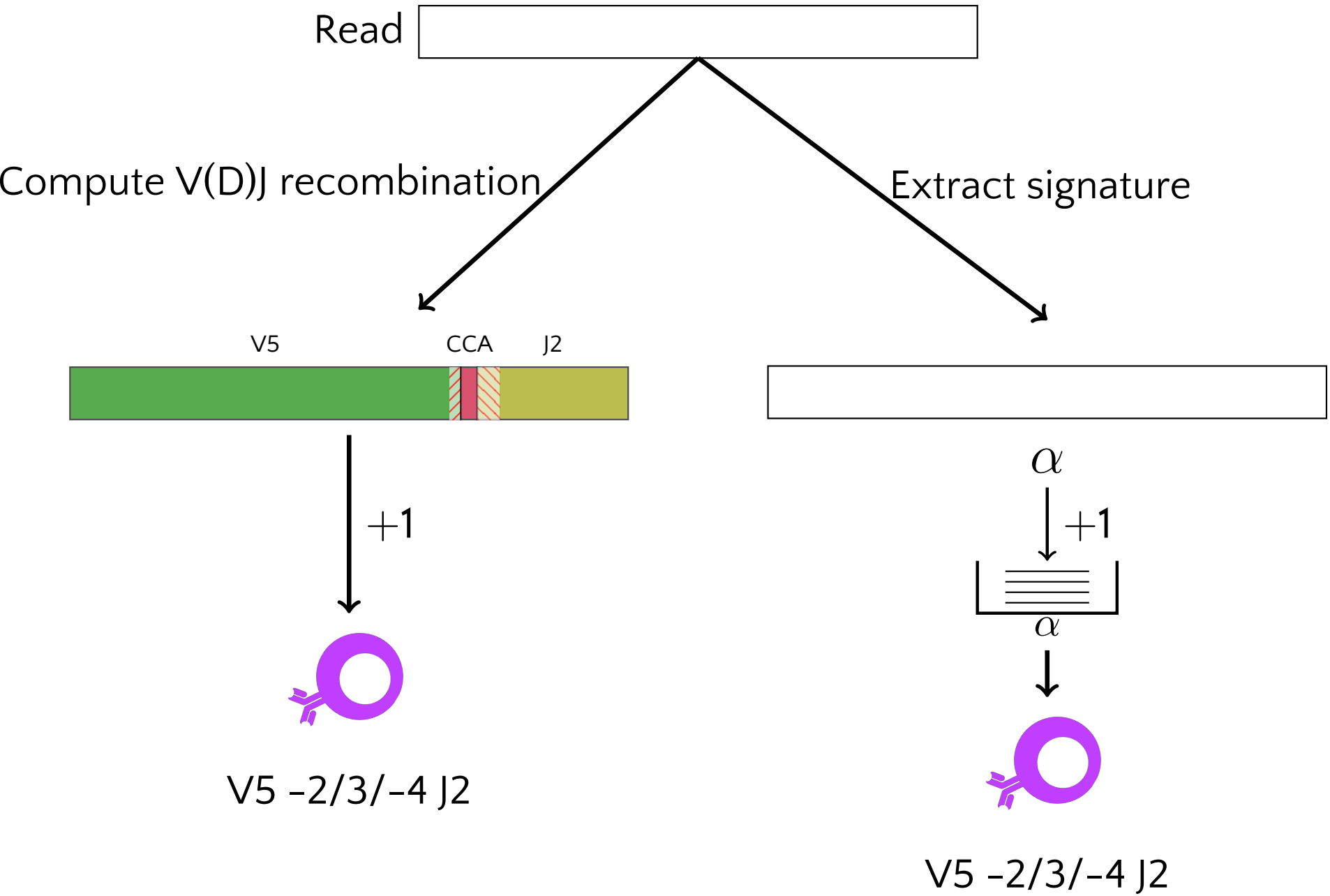
Two solutions to detect V(D)J recombinations



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

Vidjil – Ultrafast V(D)J recombination detection

A story started in 2011 as a collaboration with Lille hospital

Mathieu Giraud
CRIStAL

Vidjil – Ultrafast V(D)J recombination detection

A story started in 2011 as a collaboration with Lille hospital

Mathieu Giraud
CRIStAL

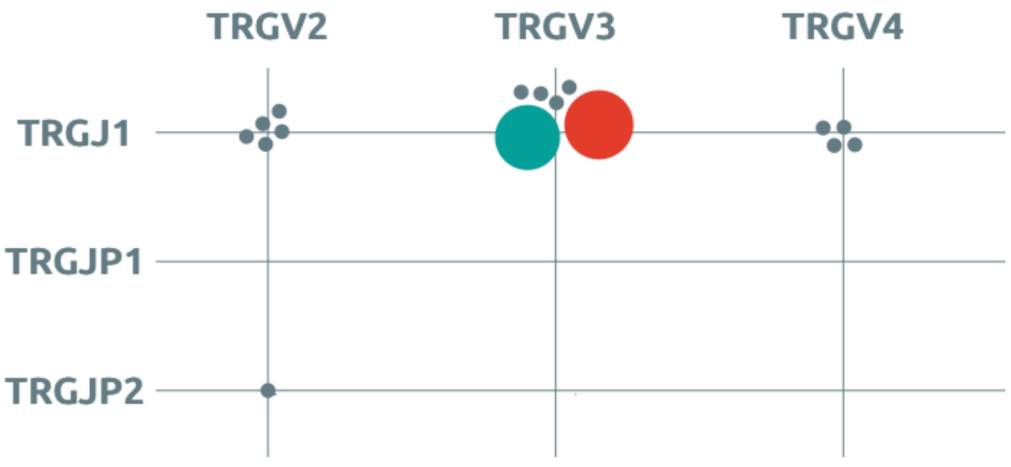
Marc Duez

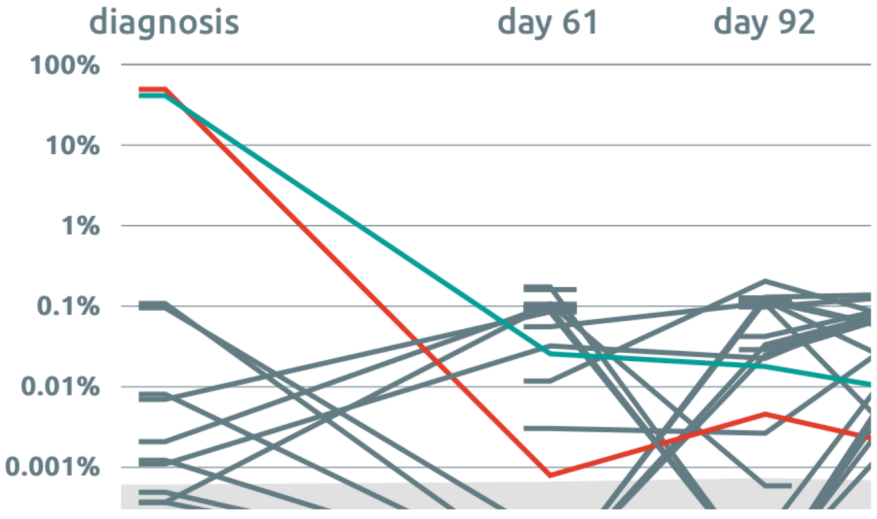
Florian Thonier

Tatiana Rocher

Ryan Herbert

Vidjil – Also a ready-to-use web application



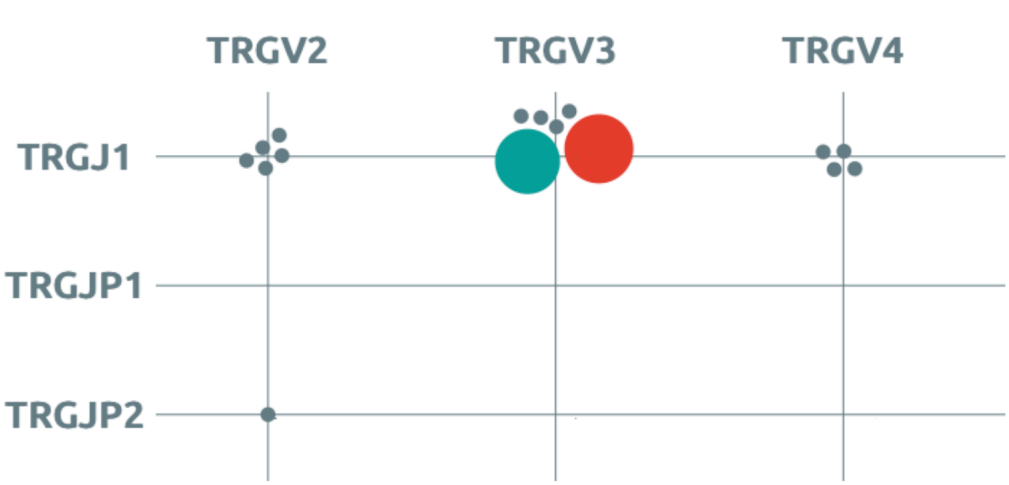


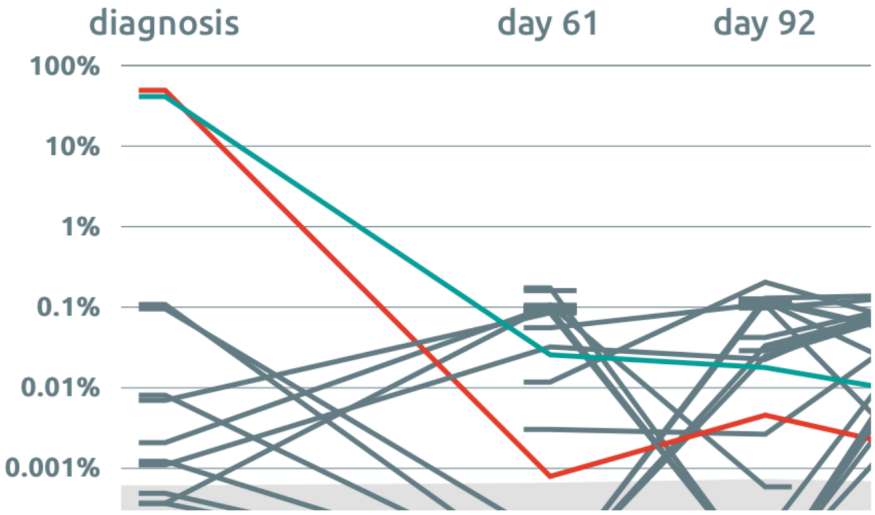
merge align › to IMGT/V-QUEST › to IgBlast › to Blast

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 – 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 3 min

MiXCR 50 min

On 2M reads

Vidjil is used throughout the world

Vidjil is used throughout the world

A public web server accessible to anyone

`app.vidjil.org`

Vidjil is used throughout the world

A public web server accessible to anyone

`app.vidjil.org`

An open-source software

`gitlab.vidjil.org`

Vidjil is used throughout the world

A public web server accessible to anyone

`app.vidjil.org`

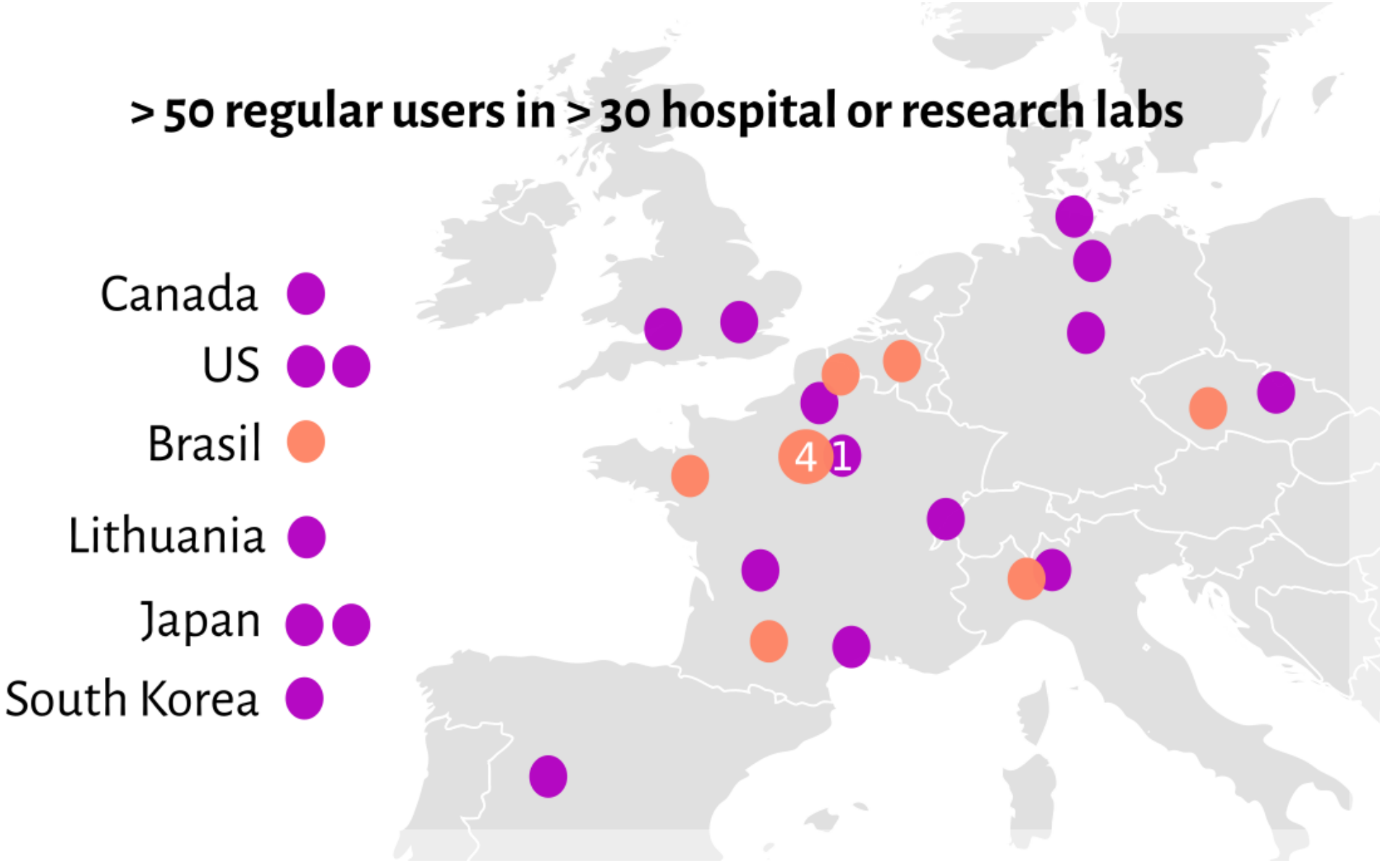
An open-source software

`gitlab.vidjil.org`

A nonprofit consortium to support and enhance the software

`vidjil.net`

Vidjil is used throughout the world



Why do the algorithms matter?

52 Vidjil launched every day

Median value for 2019. D1=6, D9=166

Why do the algorithms matter?

52 Vidjil launched every day

Median value for 2019. D1=6, D9=166

33 seconds for each job

Median value for 2019. D1=1, D9=433

Why do the algorithms matter?

52 Vidjil launched every day

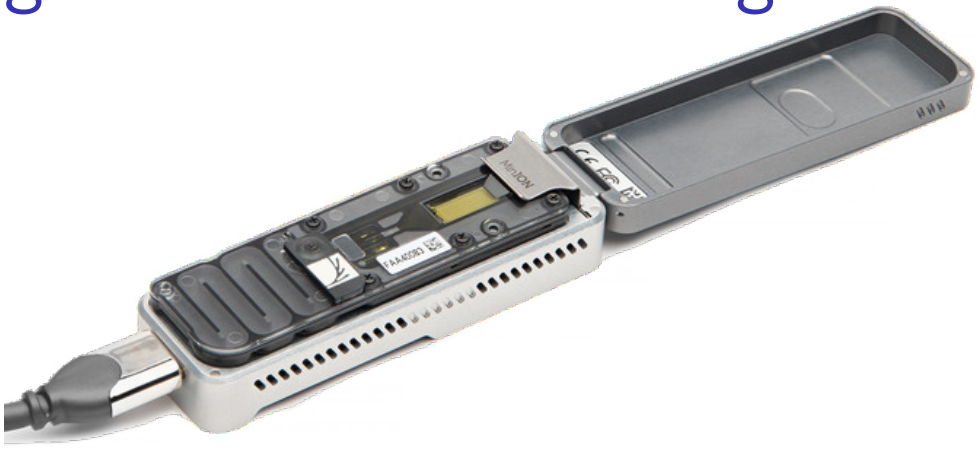
Median value for 2019. D1=6, D9=166

33 seconds for each job

Median value for 2019. D1=1, D9=433

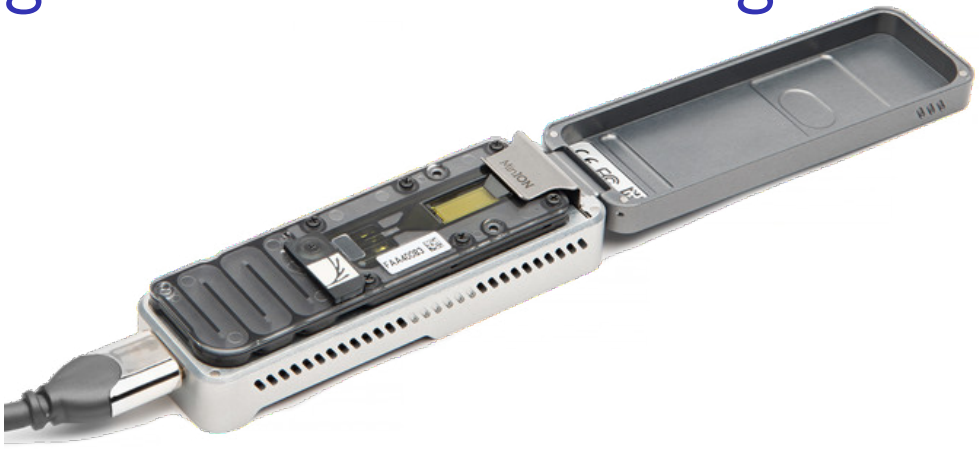
For years, the server was less powerful than my 5-year old laptop…

What string algorithms can still bring to Vidjil?



Long reads, amplification-free...but 10-15 % of errors

What string algorithms can still bring to Vidjil?

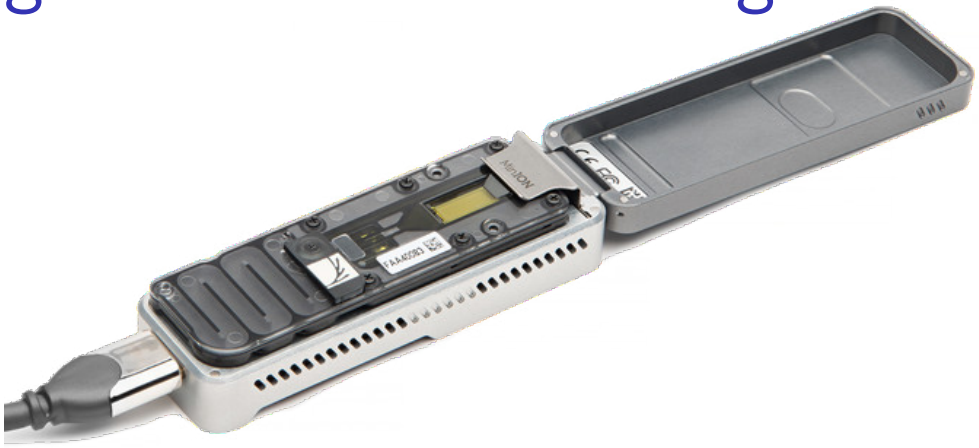


Long reads, amplification-free…but 10-15 % of errors

genome A A G A A G A C C C T

read A A T A C G A C A C T

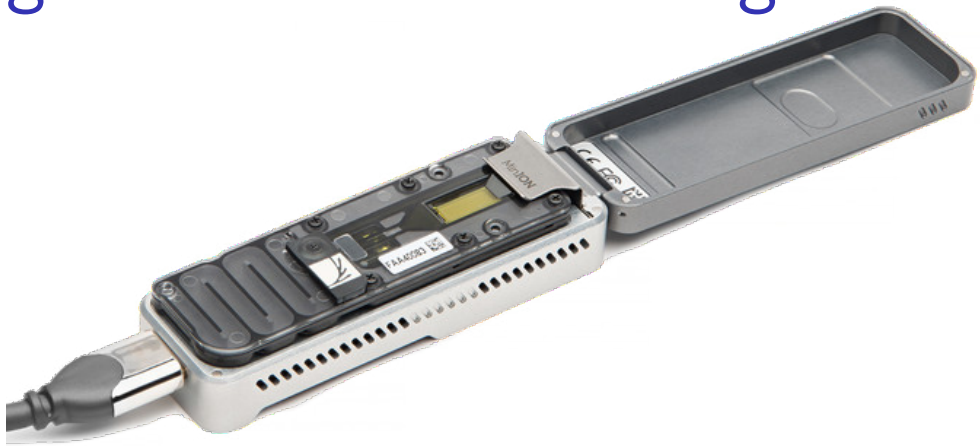
What string algorithms can still bring to Vidjil?



Long reads, amplification-free…but 10-15 % of errors

```
genome  A A G A A G A C C C T
read    A A T A C G A C A C T
        A A T A
          A T A C
            T A C G
              A C G A
                C G A C
                  G A C A
                    A C A C
                      C A C T
```

What string algorithms can still bring to Vidjil?



Long reads, amplification-free…but 10-15 % of errors

```
genome  A A G A A G A C C C T
read    A A T A C G A C A C T
        A A T A
          A T A C
            T A C G
              A C G A
                C G A C
                  G A C A
                    A C A C
                      C A C T
```

No k-mer in common between the read and the genome

01^*0 seeds: finding similarities beyond the k-mers

We want to accept a possible error in a k-mer

01^*0 seeds: finding similarities beyond the k-mers

We want to accept a possible error in a k-mer

Split the k-mer in two parts

01^*0 seeds: finding similarities beyond the k-mers

We want to accept a possible error in a k-mer

Split the k-mer in two parts

k-mer k-mer

01^*0 seeds: finding similarities beyond the k-mers

We want to accept a possible error in a k-mer

Split the k-mer in two parts

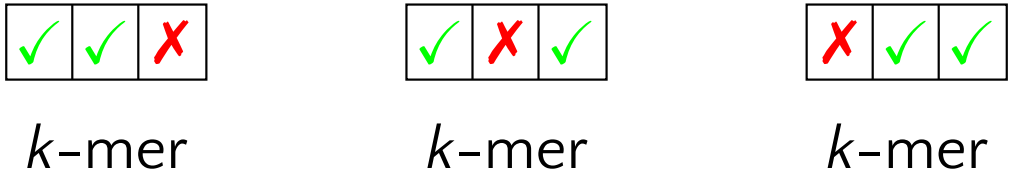
Tolerate one error, more efficiently: split in 3 parts

01^*0 seeds: finding similarities beyond the k-mers

We want to accept a possible error in a k-mer

Split the k-mer in two parts

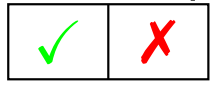
Tolerate one error, more efficiently: split in 3 parts



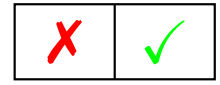
01^*0 seeds: finding similarities beyond the k-mers

We want to accept a possible error in a k-mer

Split the k-mer in two parts

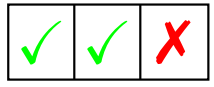


k-mer

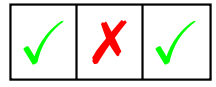


k-mer

Tolerate one error, more efficiently: split in 3 parts



k-mer



k-mer

k-mer

With e errors, split in $e + 2$ parts.

There will always be two parts with 0 error with all interleaving parts having one error

Vidjil – from string algorithmics to clinical practice

Efficient string algorithms (really) matter

We need good theoretical research to conceive practical tools

Why using a computing cluster when a laptop is enough?

Strong interdisciplinary collaborations can have great impacts

Even if not very successful when applying for grants…