High-throughput sequencing in immune repertoire analysis

Nika Abdollahi

Supervised by :
Prof. Martin Weigt
Prof. Frédéric Davi
Dr. Juliana Bernardes

13 Mars 2020

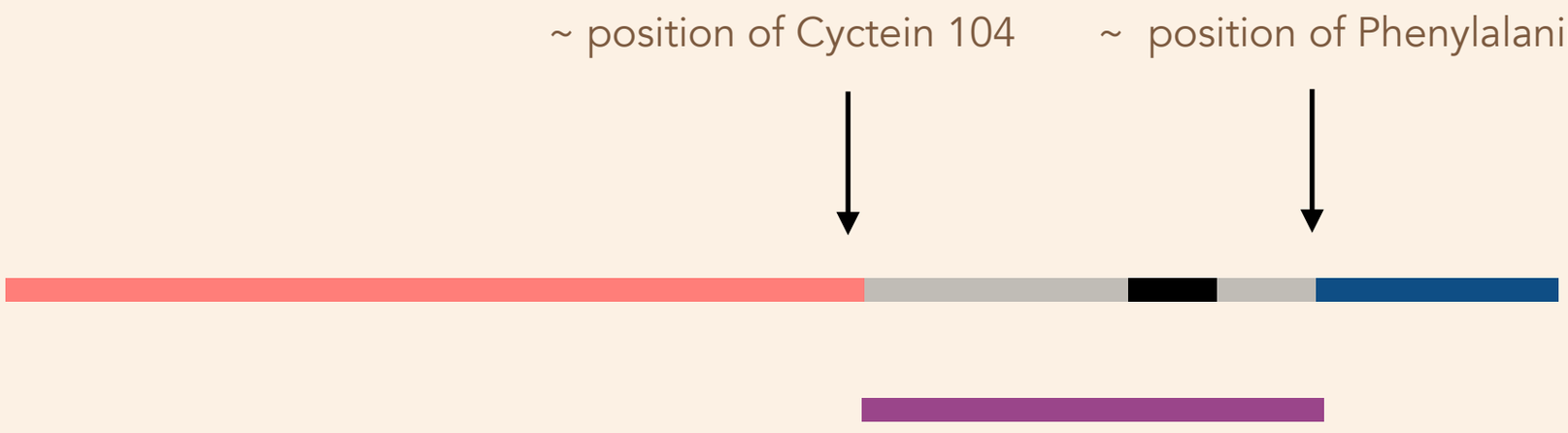
Plan

- Identifying clonal-related sequences in BCR repertoires without VDJ annotations
- Clustering T-cell beta-chain sequence receptor according to it's epitope preference
- The BCR intra-clonal diversity
- Benchmarking RepSeq analysis tools

Identifying clonal-related sequences in BCR repertoires without VDJ annotations

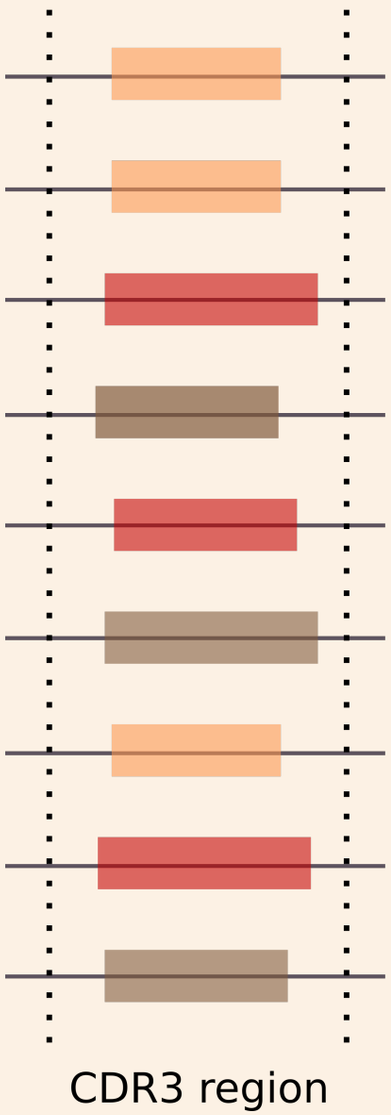
Steps of the algorithm

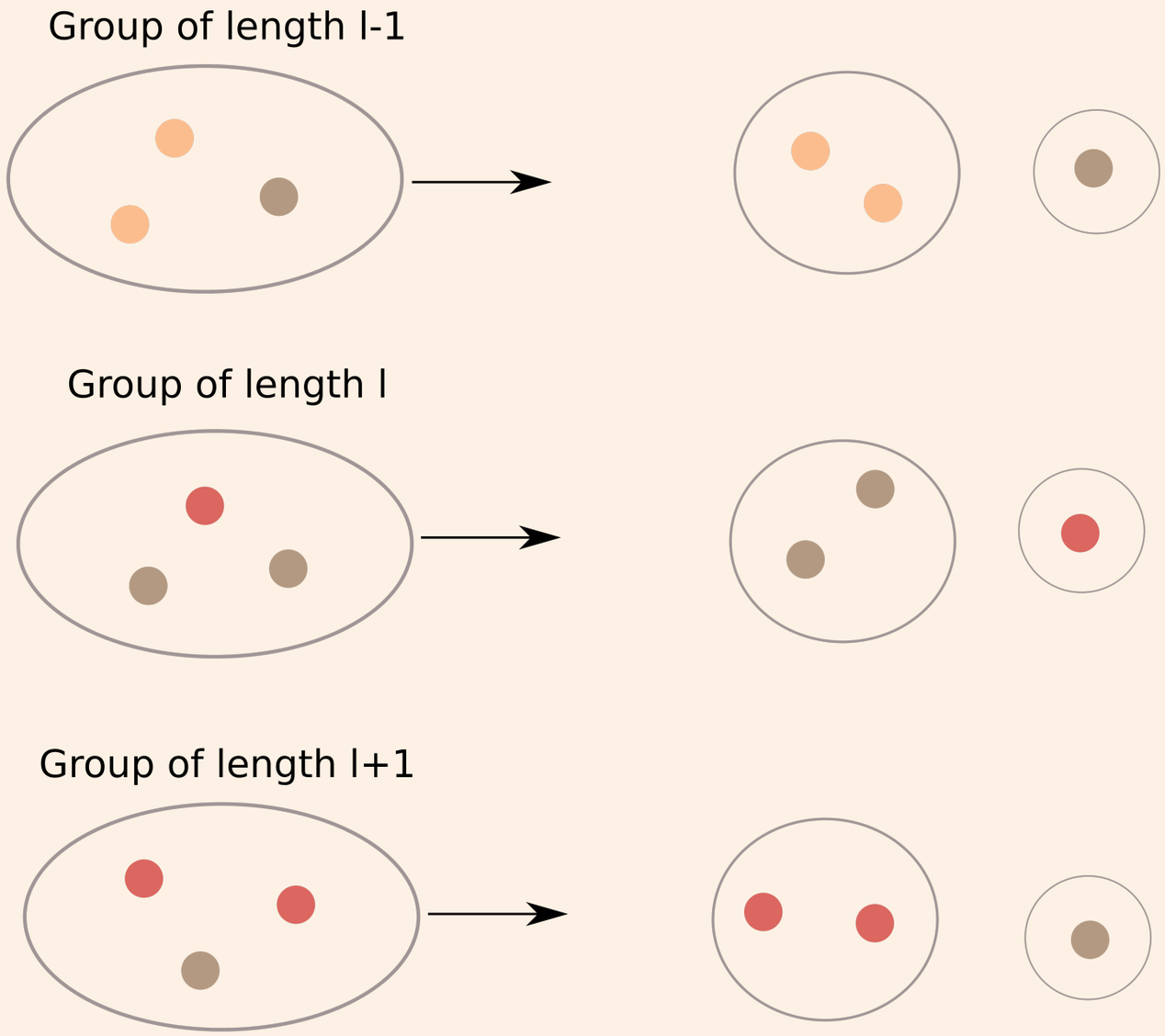
1- Junction extraction (optional)



Steps of the algorithm

2- Greedy clustering





Identifying clonal-related sequences in BCR repertoires without VDJ annotations

Improvements

Change the method's name from Mixclus to FaIR

Adapte the junction extraction part of the vidjil algorithm.

Design different methods for evaluating the performance of FaIR and its competitors.

Add few simple visualization tools for clinical use.

Extensively tested FaIR on simulated and experimental B-cell repertoires of both healthy individuals and patients with chronic lymphocytic leukemia.

Identifying clonal-related sequences in BCR repertoires without VDJ annotations

Improvements

Change the method's name from Mixclus to FaIR

Adapte the junction extraction part of the vidjil algorithm.

Design different methods for evaluating the performance of FaIR and its competitors.

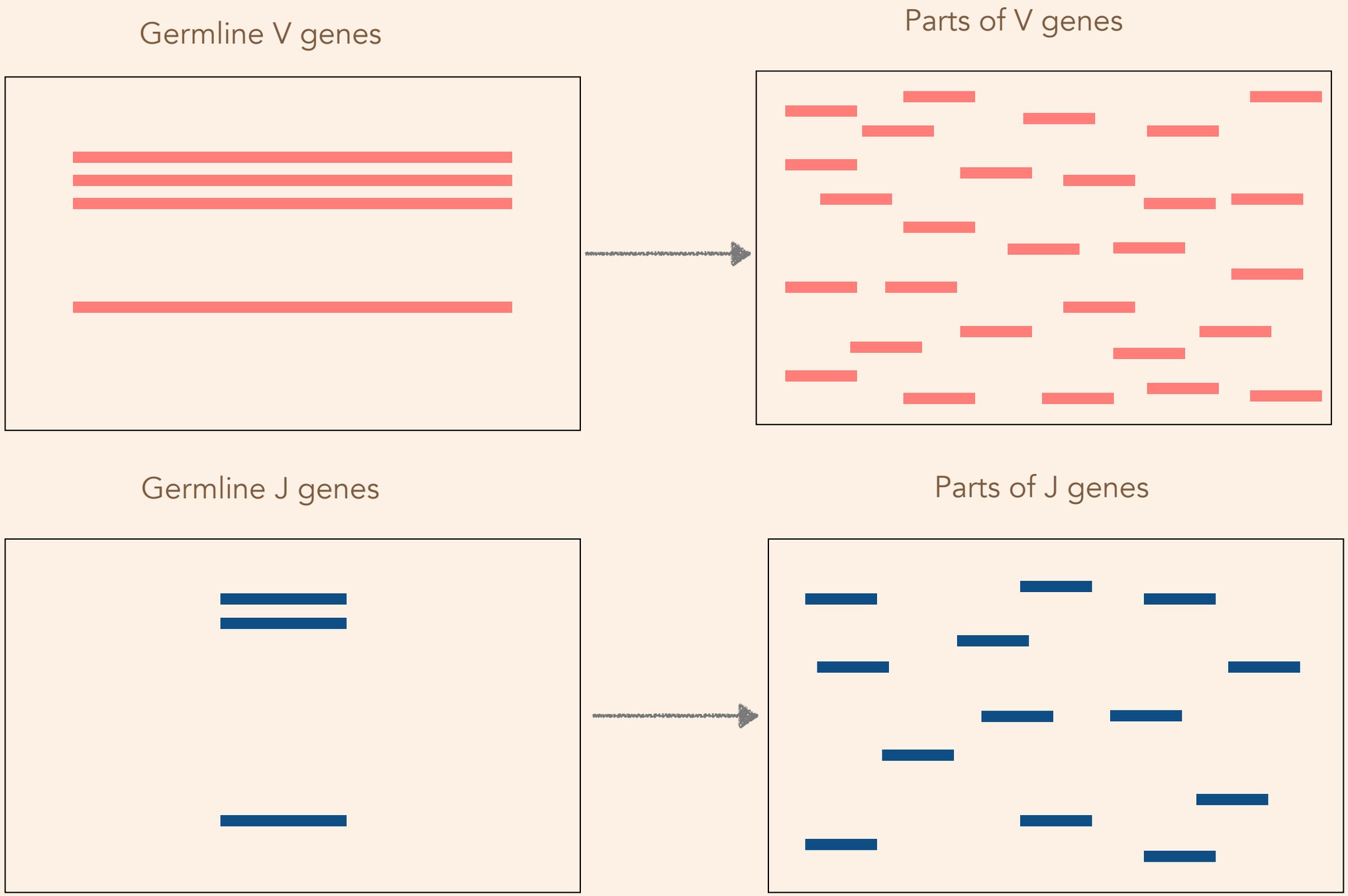
Add few simple visualization tools for clinical use.

Extensively tested FaIR on simulated and experimental B-cell repertoires of both healthy individuals and patients with chronic lymphocytic leukemia.

Adapte the junction extraction part of the vidjil algorithm

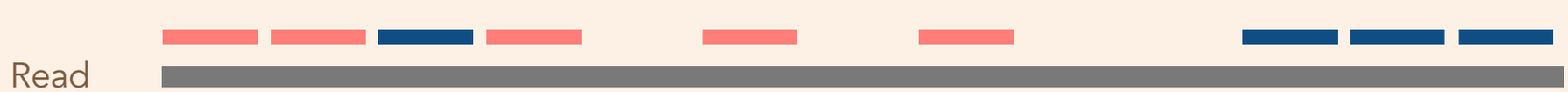
Vidjil_algo first phase

Indexing the germlines data bases

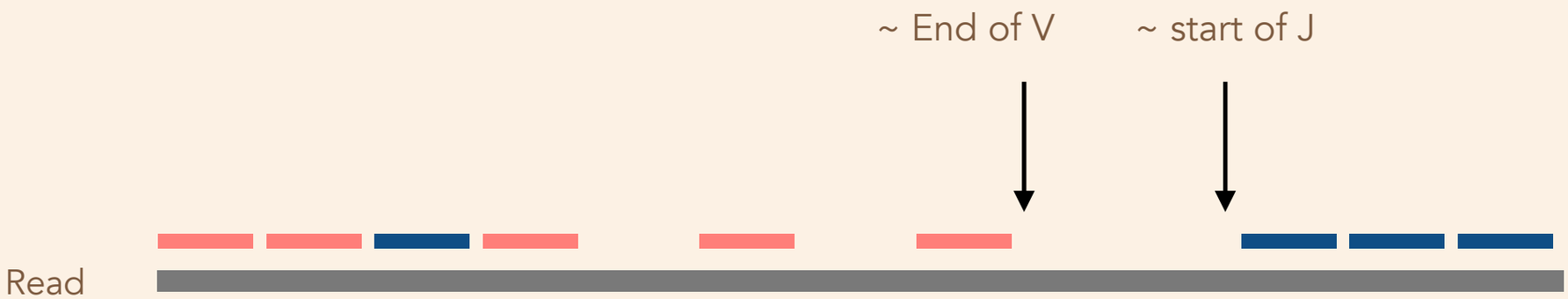


Vidjil_algo first phase

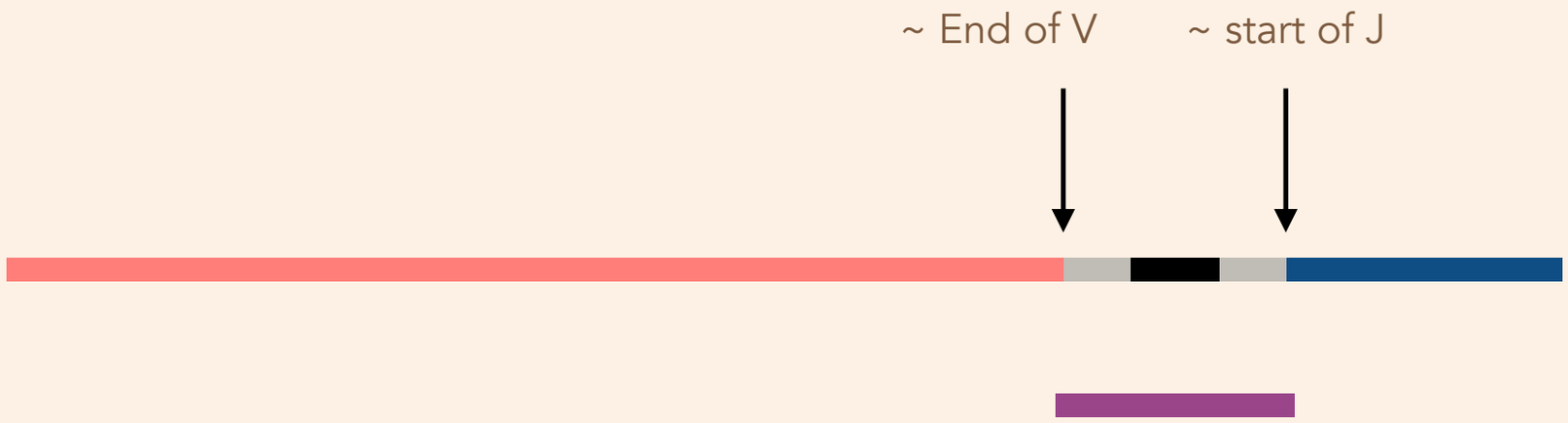
Matching the « k mers » on reads



Vidjil_algo first phase



Extracting the junction sequence



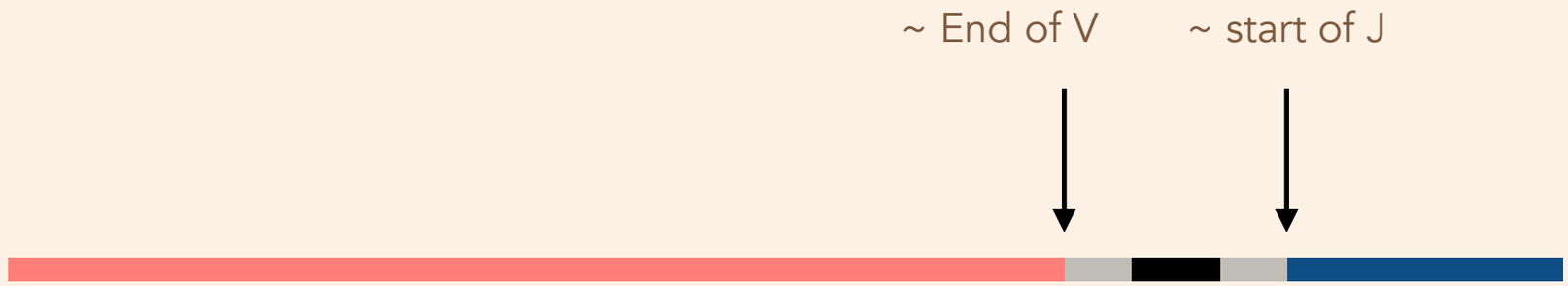
In general, It covers the two N regions and the D gene

It’s shorter than CDR3

Not enough information for clustering

Extracting the junction sequence

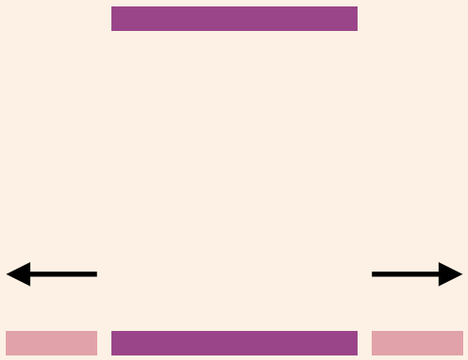
Tradeoff between the sequence length (‘effective information for clustering’) and the computational time



Vidjil_algo first phase

Adjustment

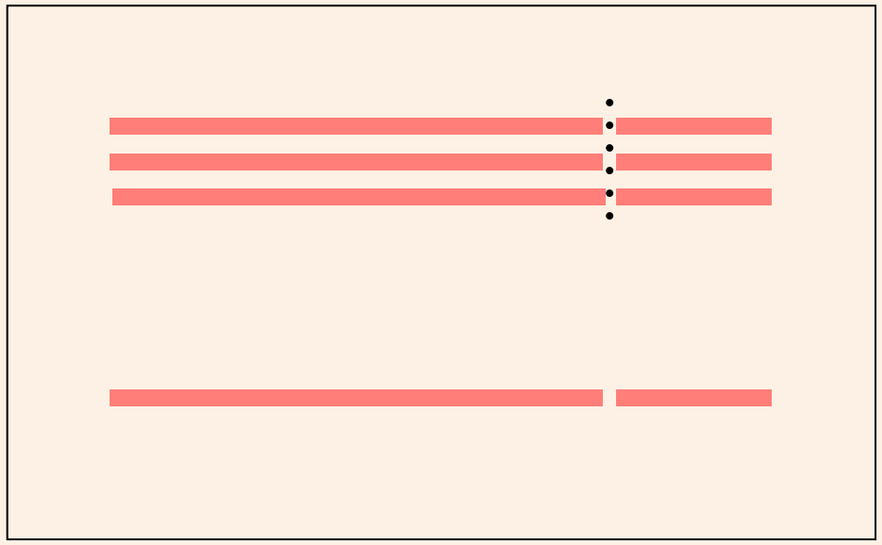
(to increase the chance of covering CDR3 region)



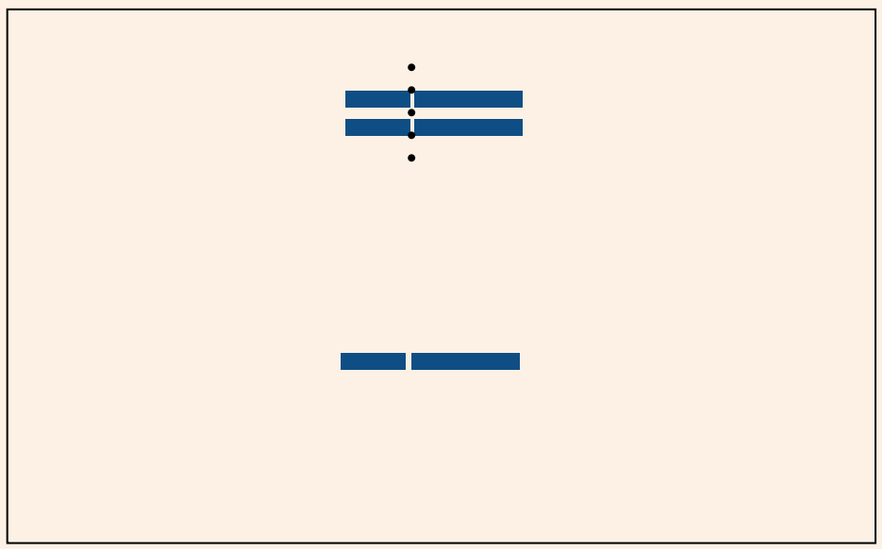
Vidjil_algo first phase

Trim the sequences based on the CDR3 anchors

Trim germline V genes after Cyctein 104

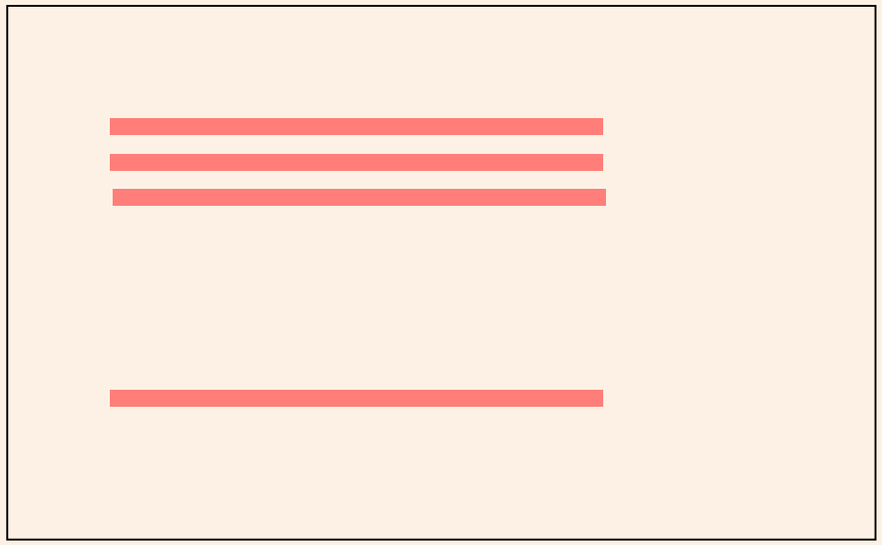


Trim germline J genes before Phenylalanine or Tryptophane 118

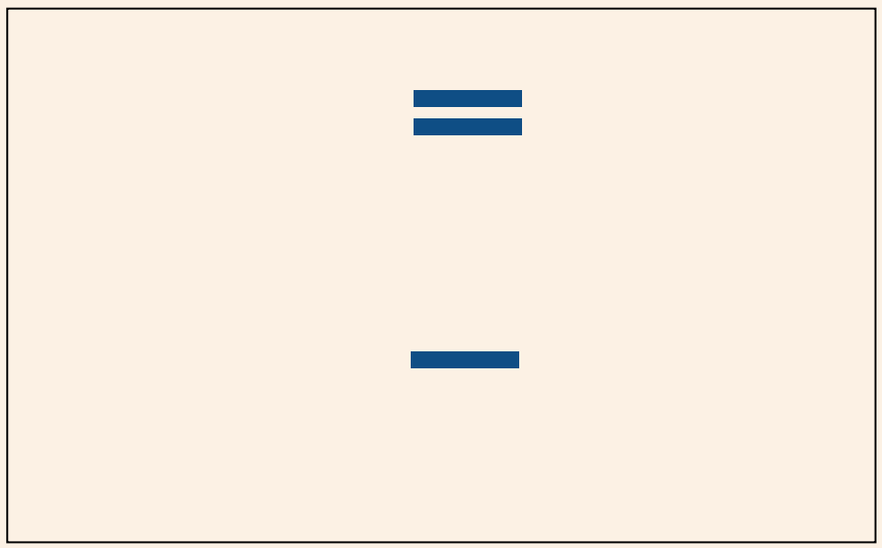


Vidjil_algo first phase

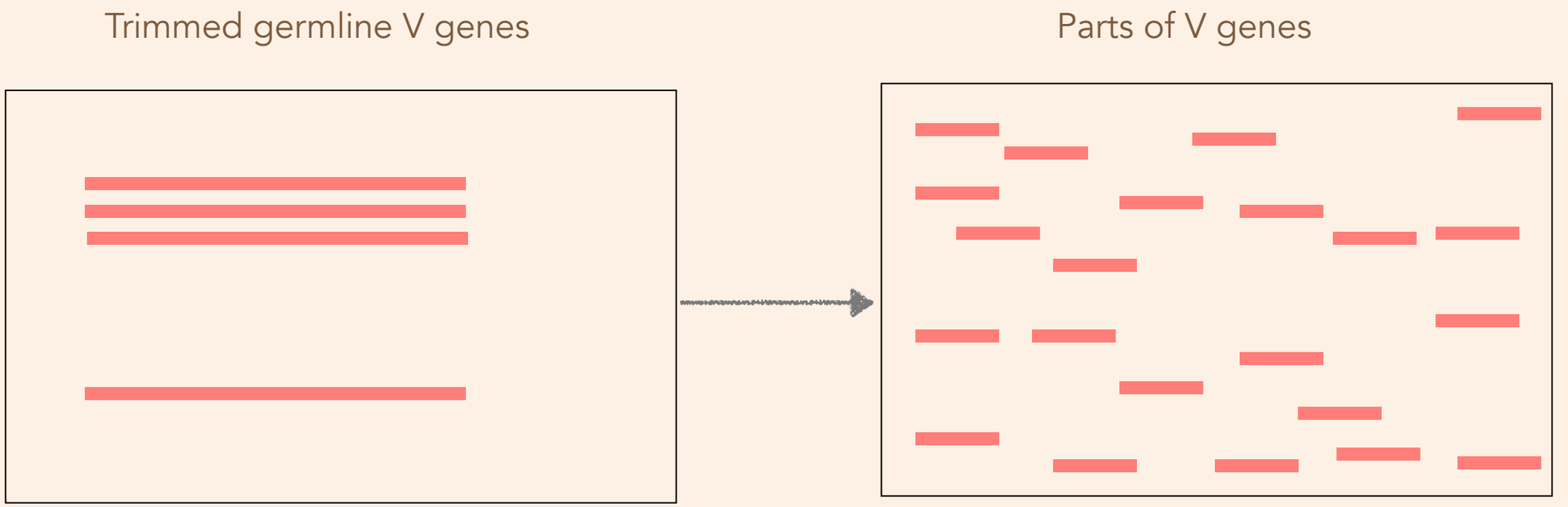
Trimmed germline V genes

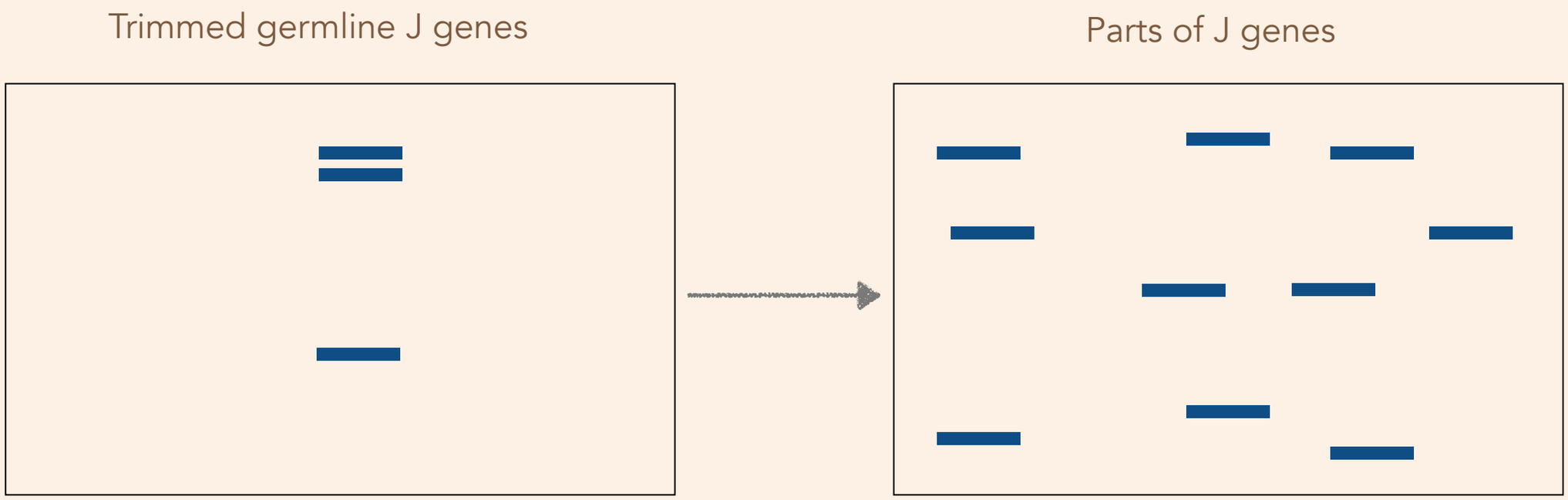


Trimmed germline J genes

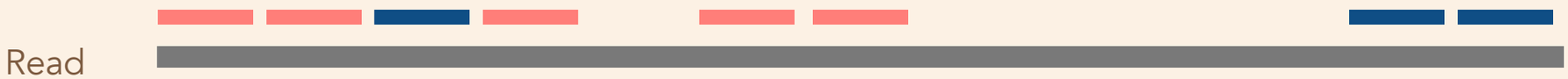


Vidjil_algo first phase

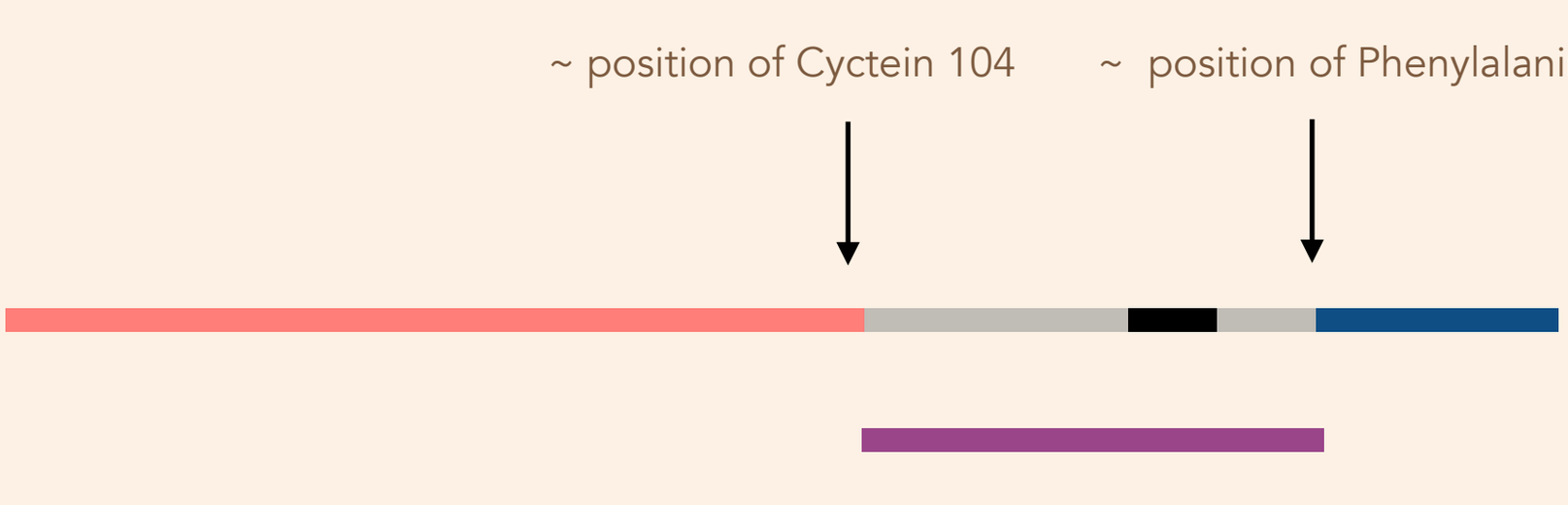




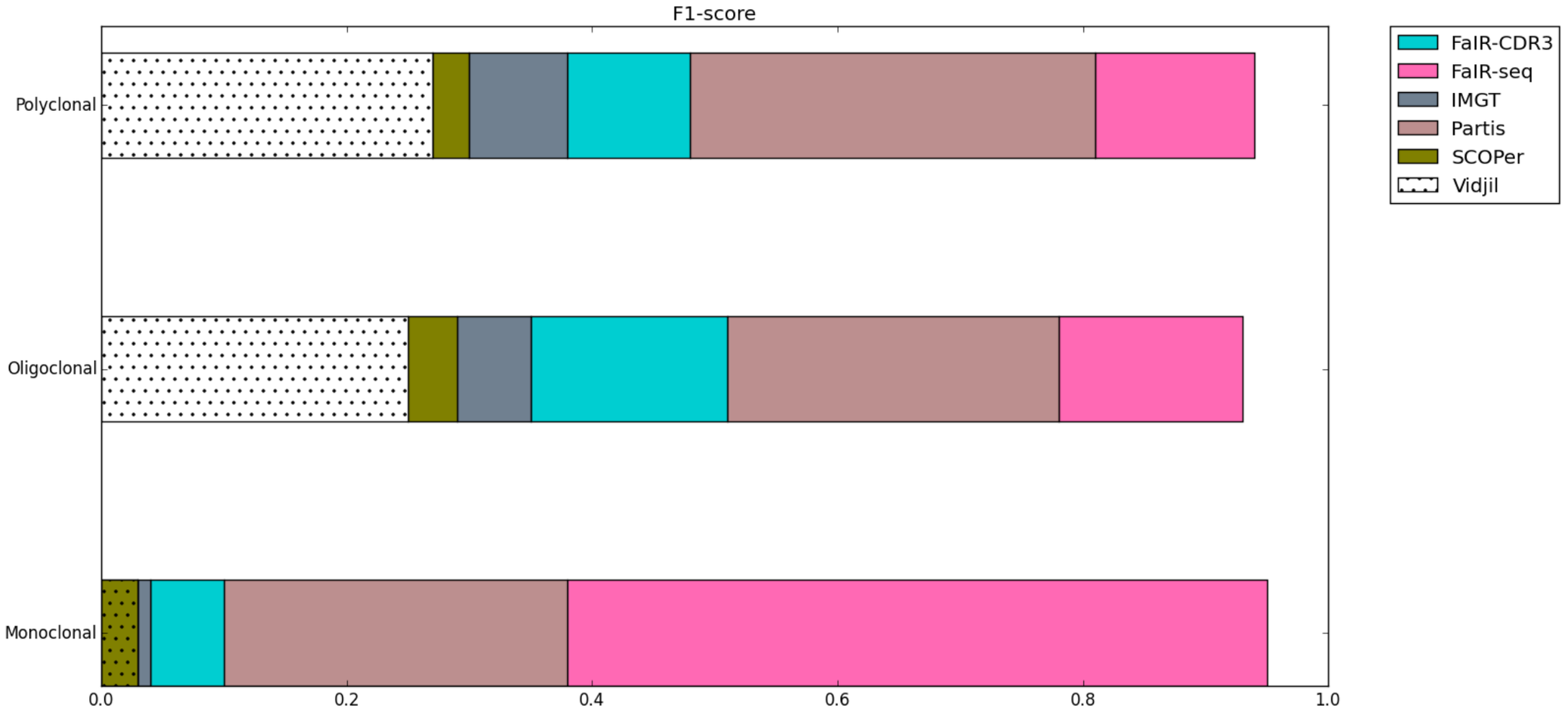
Vidjil_algo first phase

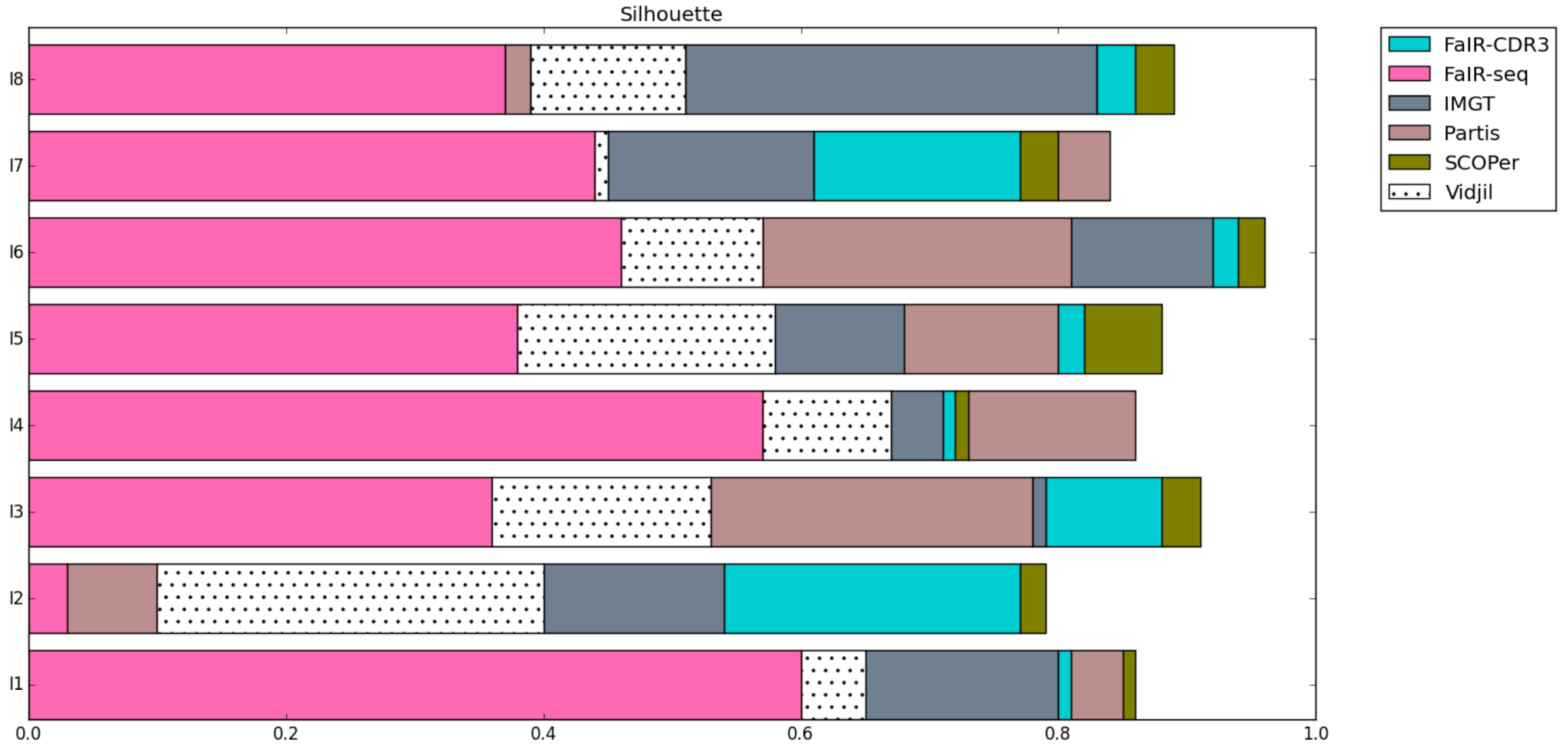


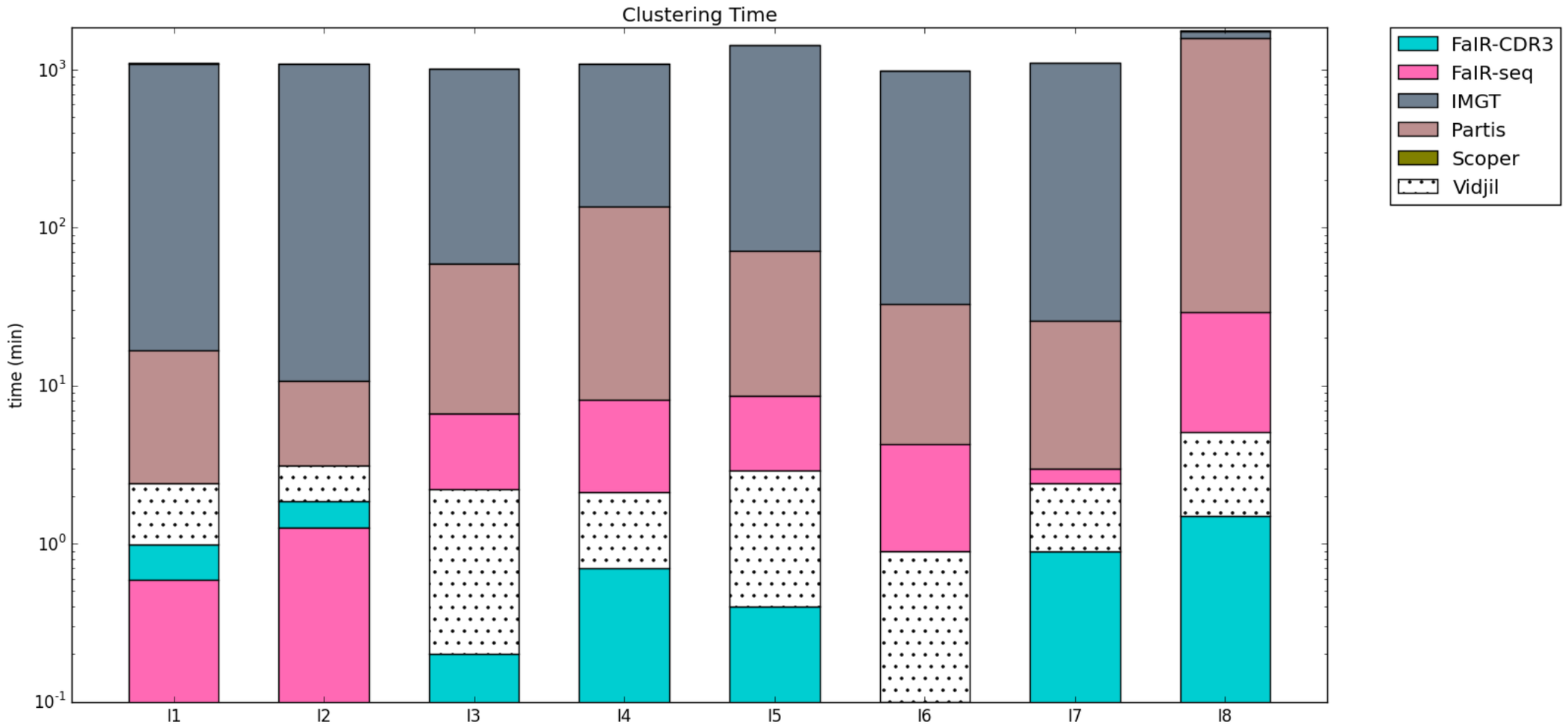
Vidjil_algo first phase



Design of different methods for evaluating the performance of FaIR and its competitors on artificial and experimental repertoires


F1-score
Polyclonal
Oligoclonal
Monoclonal
0.0
0.2
0.4
0.6
0.8
1.0
FaIR-CDR3
FaIR-seq
IMGT
Partis
SCOPer
Vidjil


Silhouette
I8
I7
I6
I5
I4
I3
I2
I1
0.0
0.2
0.4
0.6
0.8
1.0
FaIR-CDR3
FaIR-seq
IMGT
Partis
SCOPer
Vidjil


Clustering Time
time (min)
10^3
10^2
10^1
10^0
10^{-1}
I1
I2
I3
I4
I5
I6
I7
I8
FaIR-CDR3
FaIR-seq
IMGT
Partis
Scoper
Vidjil

Silhouette

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

i : a datum

a(i) : the average distance between i and all other data within the same cluster

b(i) : the smallest average distance of to all points in any other cluster, of which i is not a member

Silhouette

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

i : a datum

a(i) : the average distance between i and all other data within the same cluster

b(i) : the smallest average distance of to all points in any other cluster, of which i is not a member

Distance 1 : Levenstein distance between 2 nucleotide sequences

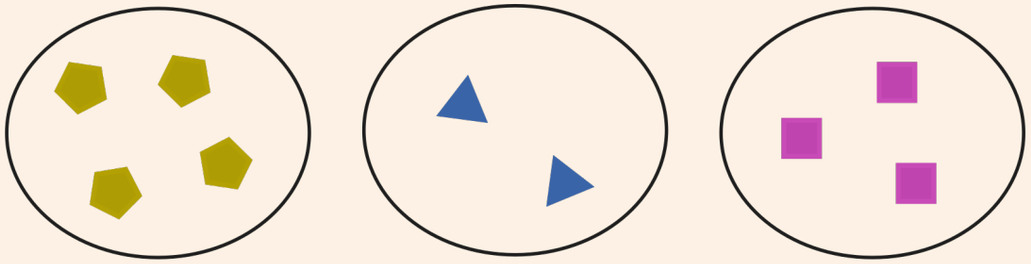
Distance 1 : Levenstein distance between 2 CDR3 sequences + V gene identity + J gene identity

Silhouette

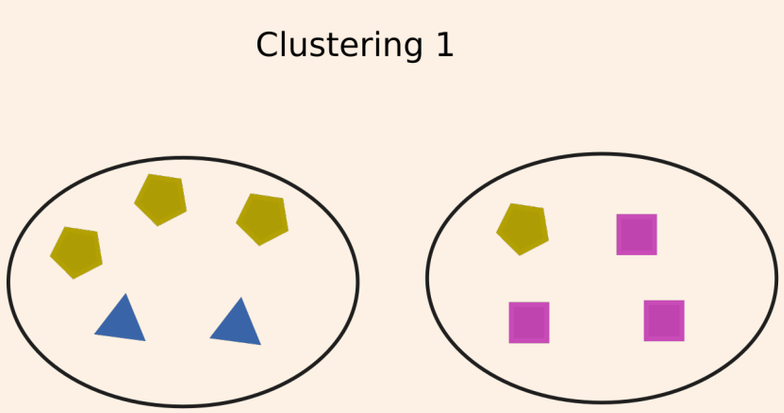
Add The results

F-score

Sequence based

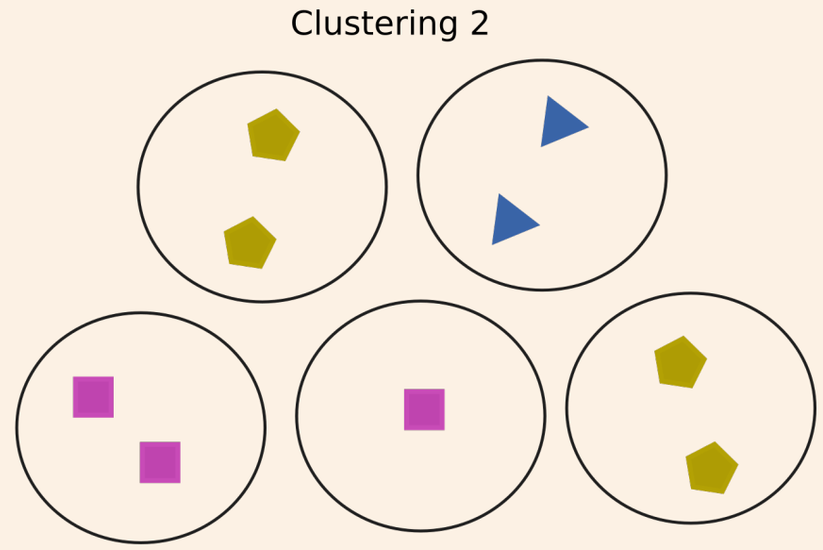


Clustering 1



TP = 14 Precision = 0.43
FN = 6 Recall = 0.70
FP = 18 F-score = 0.53

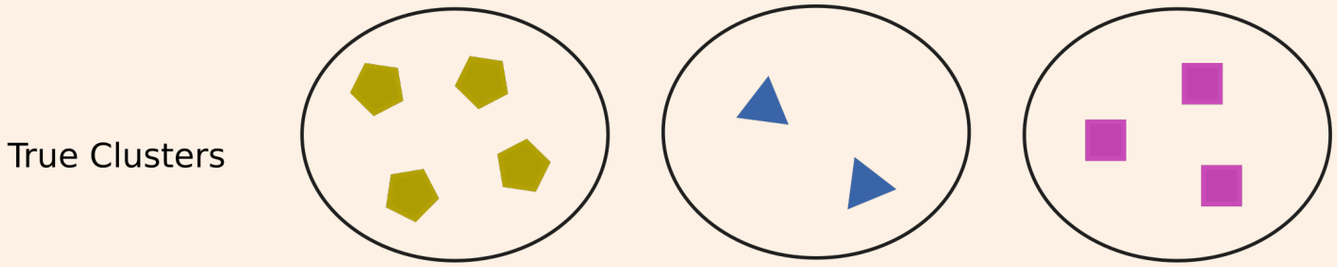
Clustering 2



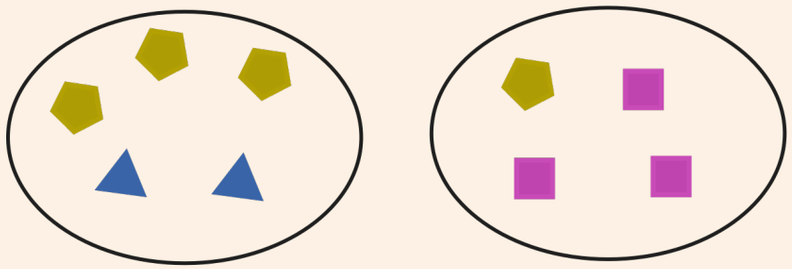
TP = 8 Precision = 1
FN =12 Recall = 0.4
FP = 0 F-score = 0.57

F-score

Cluster based

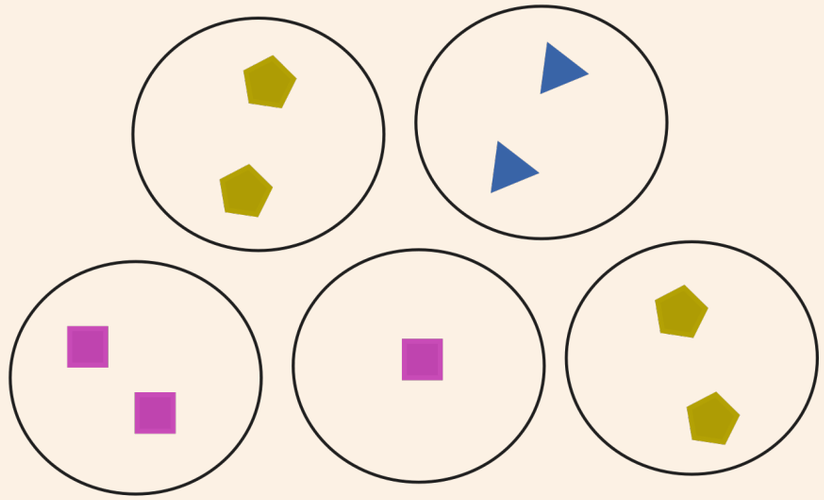


Clustering 1



TP = 6 Precision = 0.66
FN = 1 Recall = 0.85
FP = 3 F-score = 0.74

Clustering 2



TP = 9 Precision = 1
FN = 7 Recall = 0.56
FP = 0 F-score = 0.71

F-score

Add The results

Clustering T-cell beta-chain sequence receptor according to it's epitope preference

VDJdb

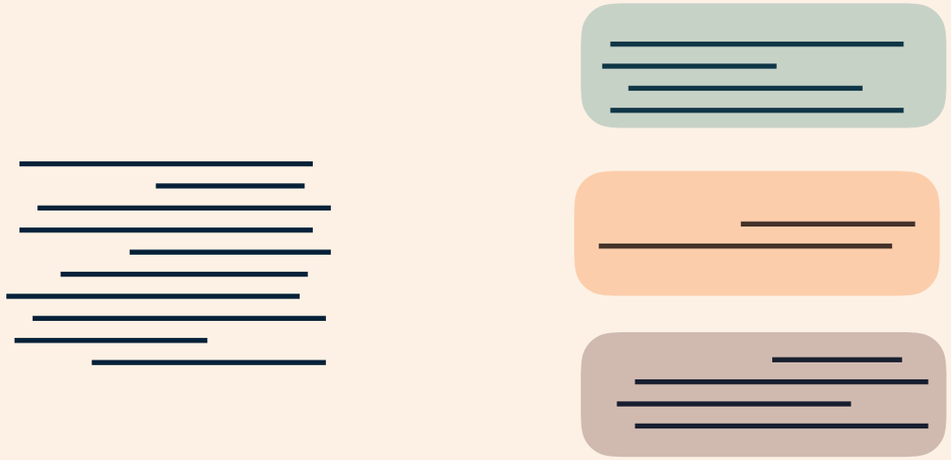
On the viability of unsupervised T-cell receptor sequence clustering for epitope preference

Pieter Meysman, Nicolas De Neuter, Sofie Gielis, Danh Bui Thi, Benson Ogunjimi, Kris Laukens

Structural distinct TCR groups ===> target identical epitopes

VDJdb

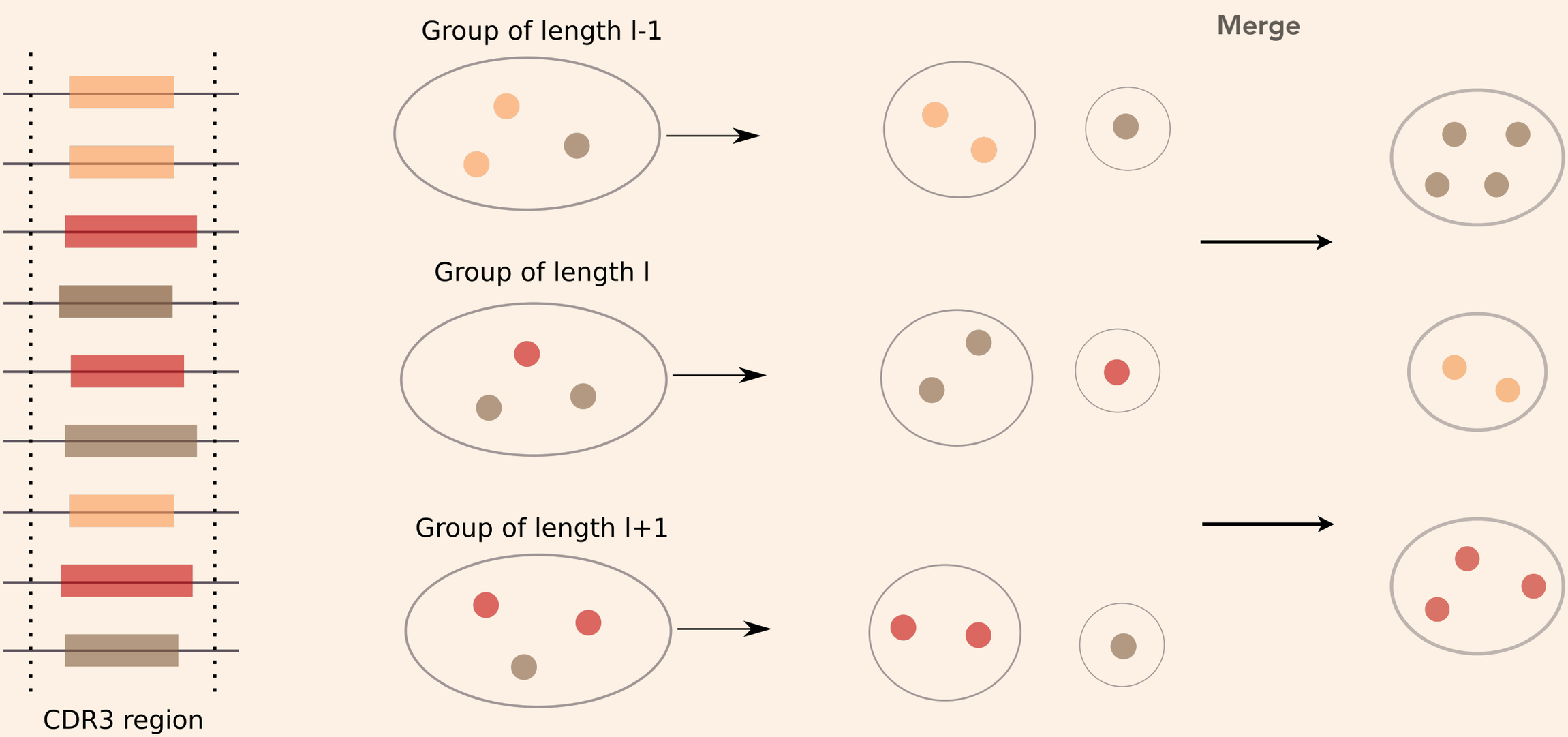
Cluster : CDR3 sequences associated to the same epitope-MHC
(#seq =1021, # true cluster= 78, #singleton = 11)



Challenge

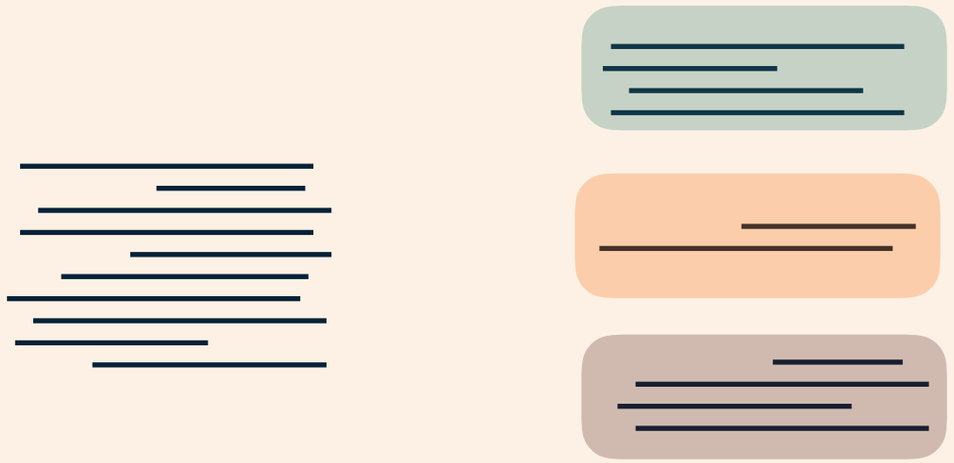

Highly similar CDR3
associated with different
epitope - MHC
Different CDR3 associated
with same epitope - MHC

Adapted algorithm



Performance

Cluster : CDR3 sequences associated with the same epitope-MHC
(#seq =1021, # true cluster= 78, #singleton = 11)

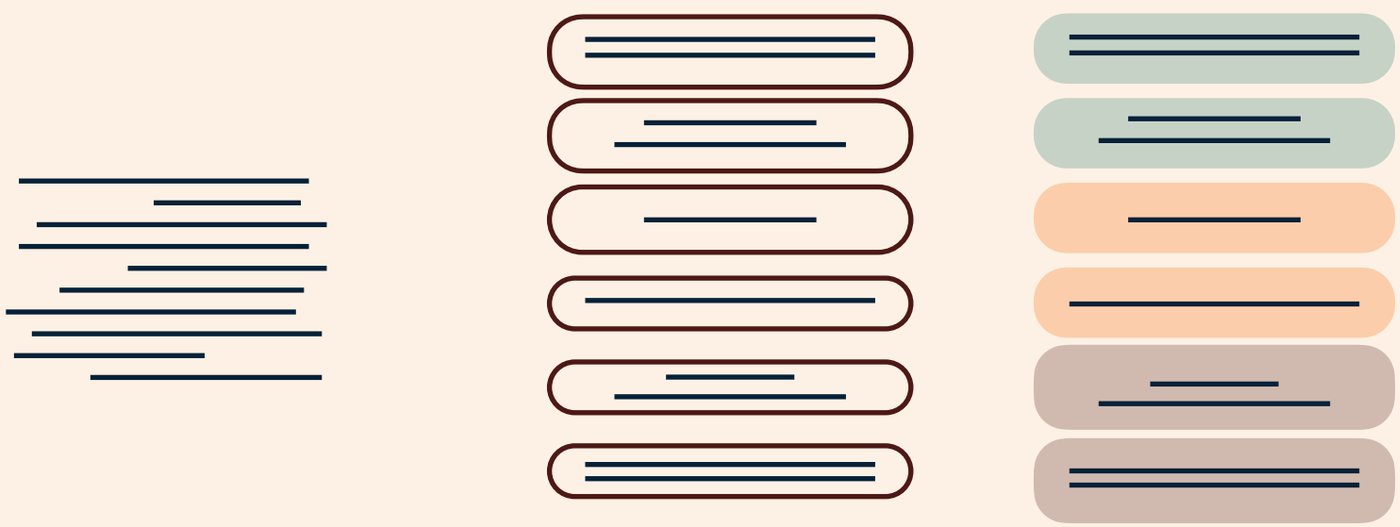


				Nearest cluster				Pairwise sequences		
	#cluster	#singleton	Non assigned	Recall	precision	F-Score	# non singleton cluster	Retention	Purity (Precision)	Consistency (Recall)
DBSCAN	594	441	441	0,025	0,88	0,049	153	0,568	0,868	0,204
Mixclus	93	9	0	0,06	0,78	0,11	84	0.91	0,76	0,48

VDJdb

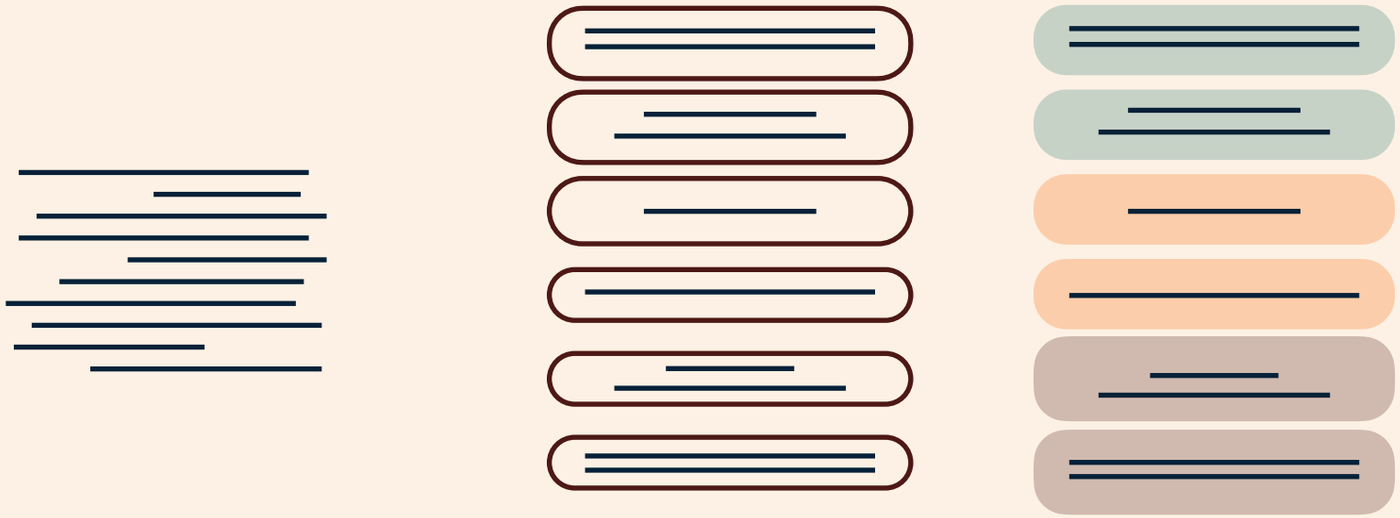
Cluster : CDR3 sequences associated to the same epitope-MHC
(#seq =1021, # true cluster= 78, #singleton = 11)

Cluster : CDR3 sequences with the same V and J genes associated to the same epitope-MHC
(#seq =1021, # true cluster= 544, #singleton =243)



Performance

Cluster : CDR3 sequences with the same V and J genes associated to the same epitope-MHC
(#seq =1021, # true cluster= 544, #singleton = 367)

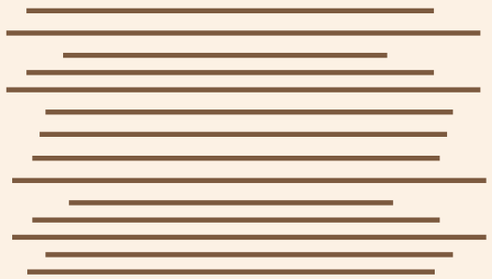


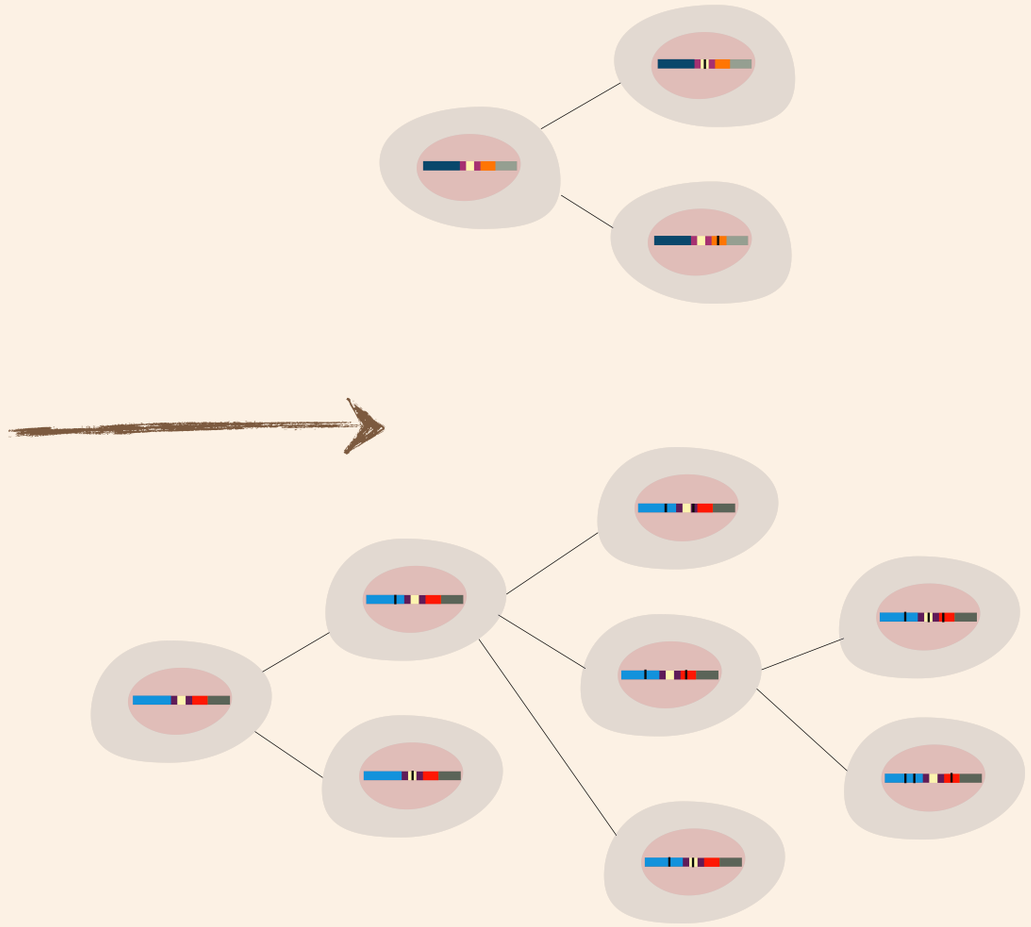
				Nearest cluster				Pairwise sequences		
	#cluster	#singleton	Non assigned	Recall	precision	F-Score	# non singleton cluster	Retention	Purity (Precision)	Consistency (Recall)
DBSCAN	632	470	470	0.43	0,90	0,55	162	0,539	0,94	0,55
Mixclus	112	33	0	0.78	0.66	0.71	112	0,95	0.60	0,7

The BCR intra-clonal diversity

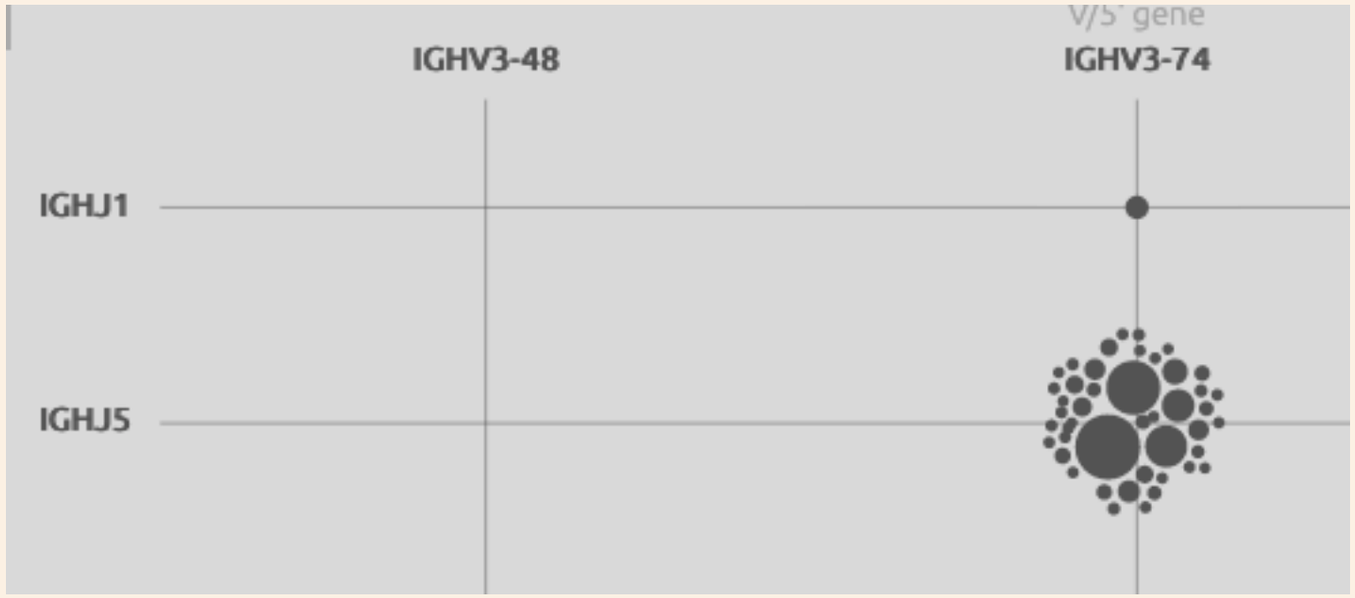
Objectif

Sequences/Reads





Intra-clonal diversity example

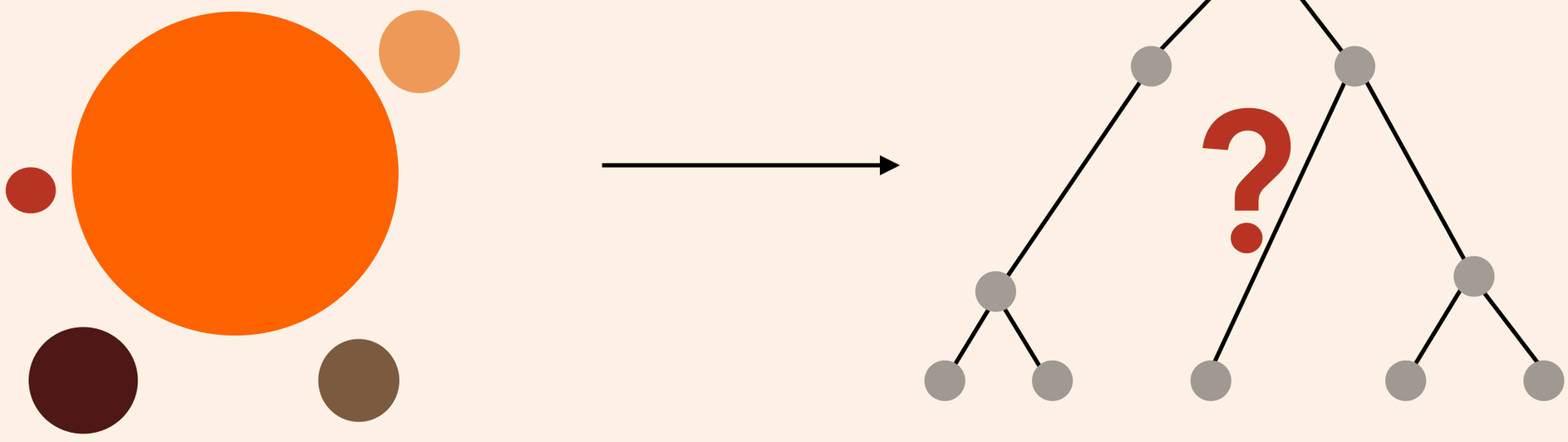


Intra-clonal diversity example

One clone :

2317 sequences

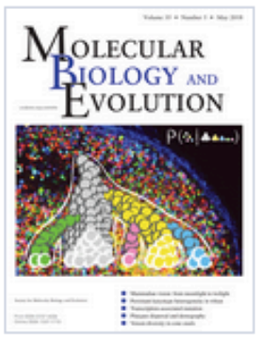
554 unique sequences



GC tree

Phylogebetic inference of genotype-collapsed tree

MOLECULAR BIOLOGY AND EVOLUTION



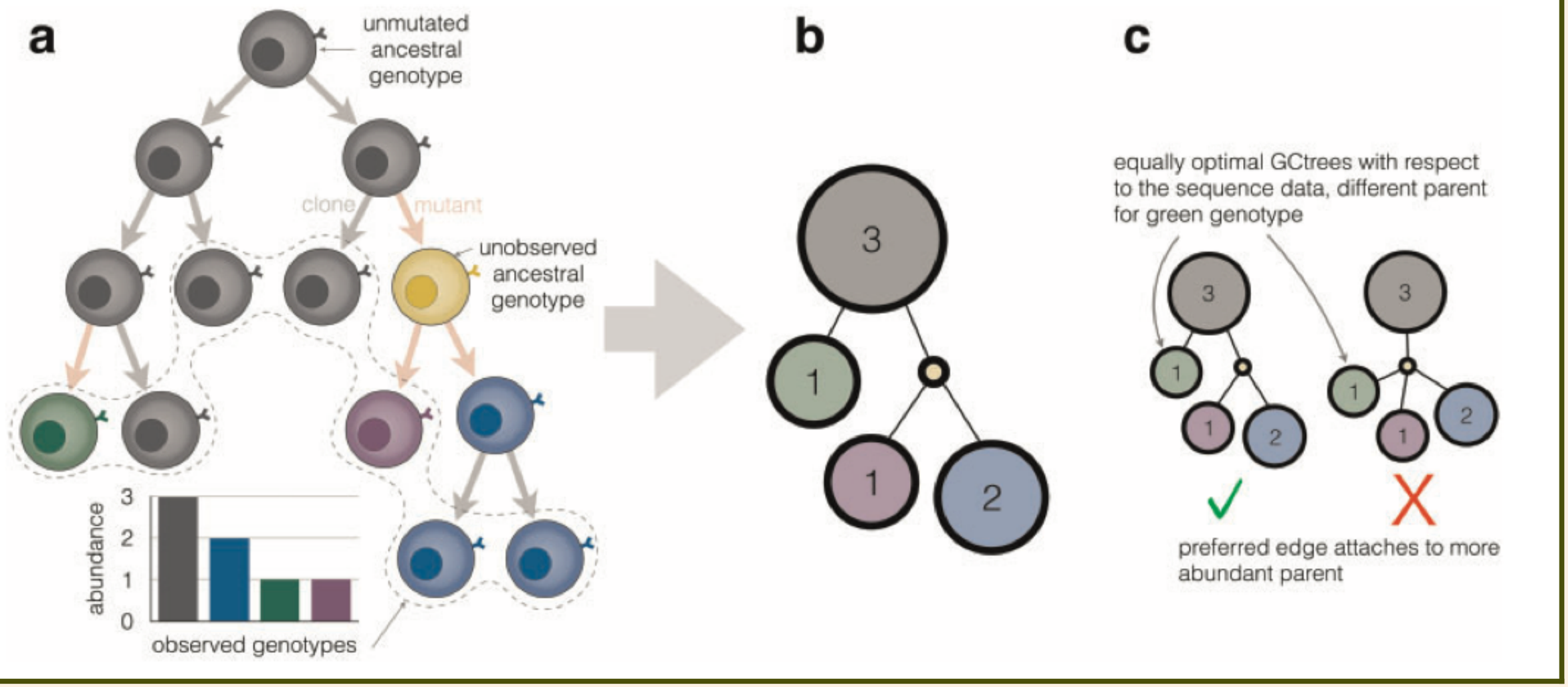
Volume 35, Issue 5
May 2018

Using Genotype Abundance to Improve Phylogenetic Inference

William S DeWitt, III, Luka Mesin, Gabriel D Victora, Vladimir N Minin ✉, Frederick A Matsen, IV ✉

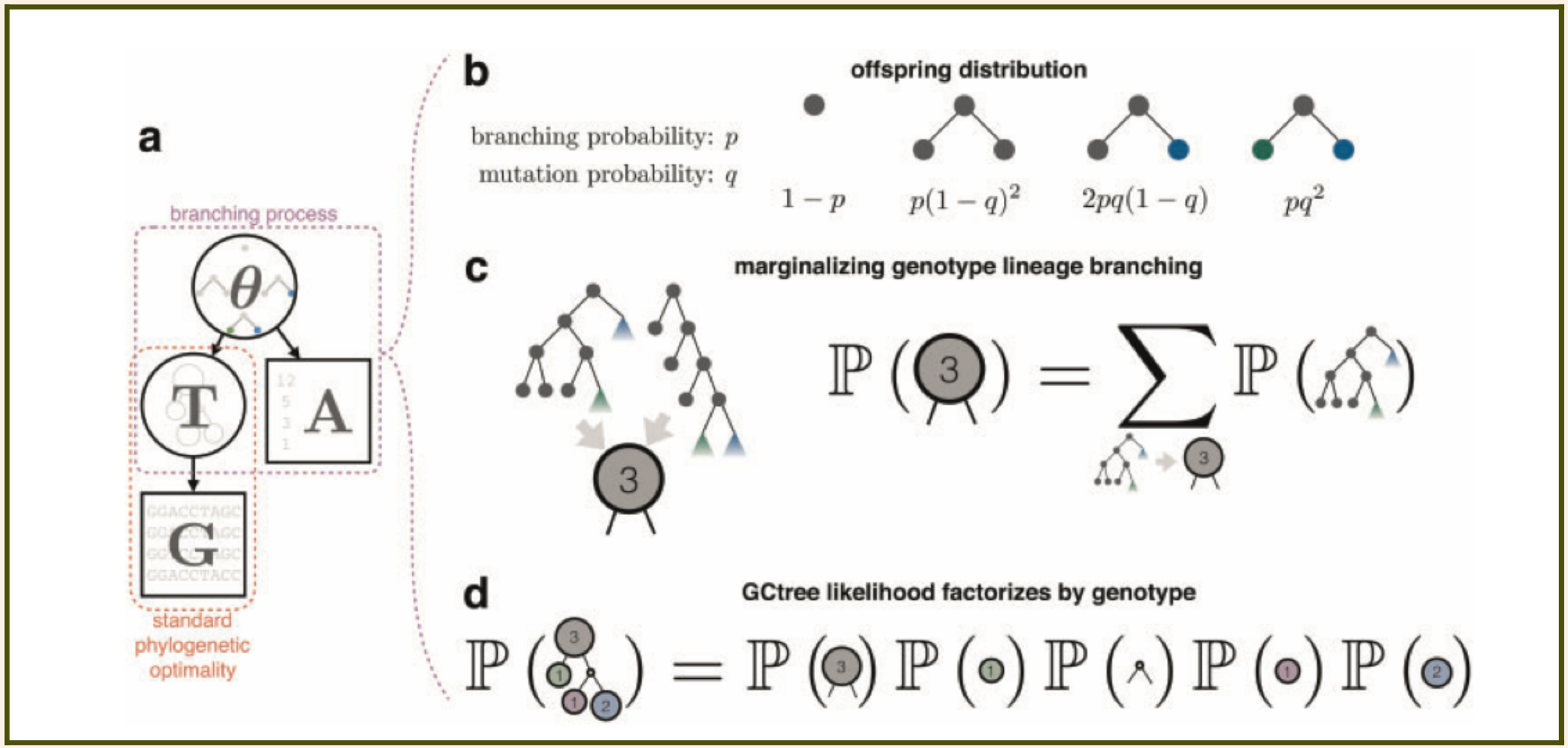
Molecular Biology and Evolution, Volume 35, Issue 5, May 2018, Pages 1253–1265,
https://doi.org/10.1093/molbev/msy020
Published: 20 February 2018

GC tree



The more abundant parent is more likely to have generated mutant descendants

GC tree


a
branching process
θ
T
A
G
standard
phylogenetic
optimality
b
branching probability: p
mutation probability: q
offspring distribution
1 − p
p(1 − q)²
2pq(1 − q)
pq²
c
marginalizing genotype lineage branching
3
ℙ(3) = ∑ ℙ(·)
d
GCtree likelihood factorizes by genotype
ℙ(3 1 2) = ℙ(3) ℙ(1) ℙ(·) ℙ(1) ℙ(2)

GC tree

simulated dataset

GCtree-simulated dataset

- containing 50 heavy chain V gene sequences
- germline : the germline sequence of the V gene used in the V(D)J rearrangement that defines this clonal family

True tree

7

9 5 3 3 2 2 1 1 1 1 1 1

3 1 1 1 1 1

4 1

GC tree rendered tree

7

naive

9 5 3 3 2 2 1 1 1 1 1 1

seq4 seq10 seq14 seq7 seq12 seq13 seq6 seq11 seq17 seq16 seq18 seq15

3 1 1 1 1 1

seq8 seq1 seq2 seq20 seq5 seq19

4 1

seq3 seq9

GC tree

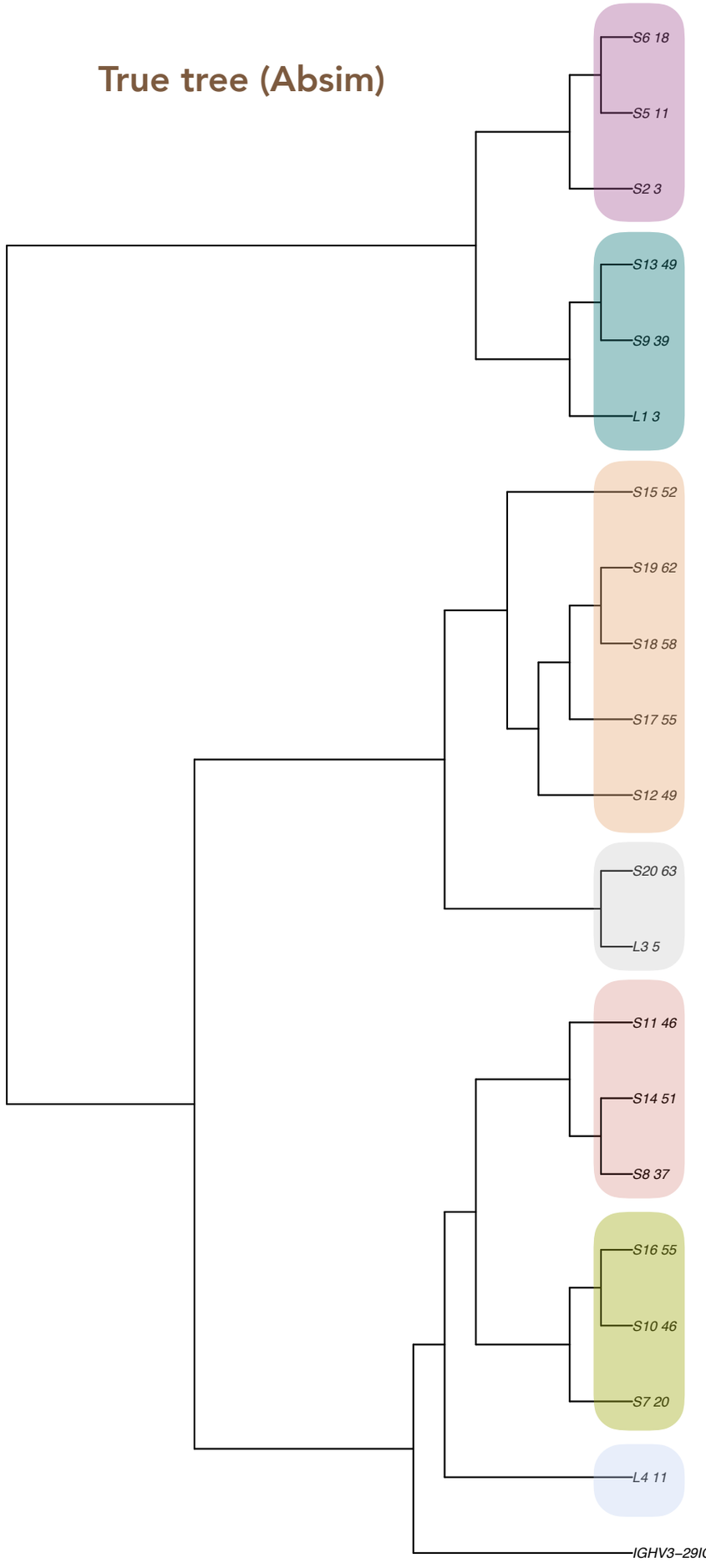
simulated dataset

Absim-simulated dataset

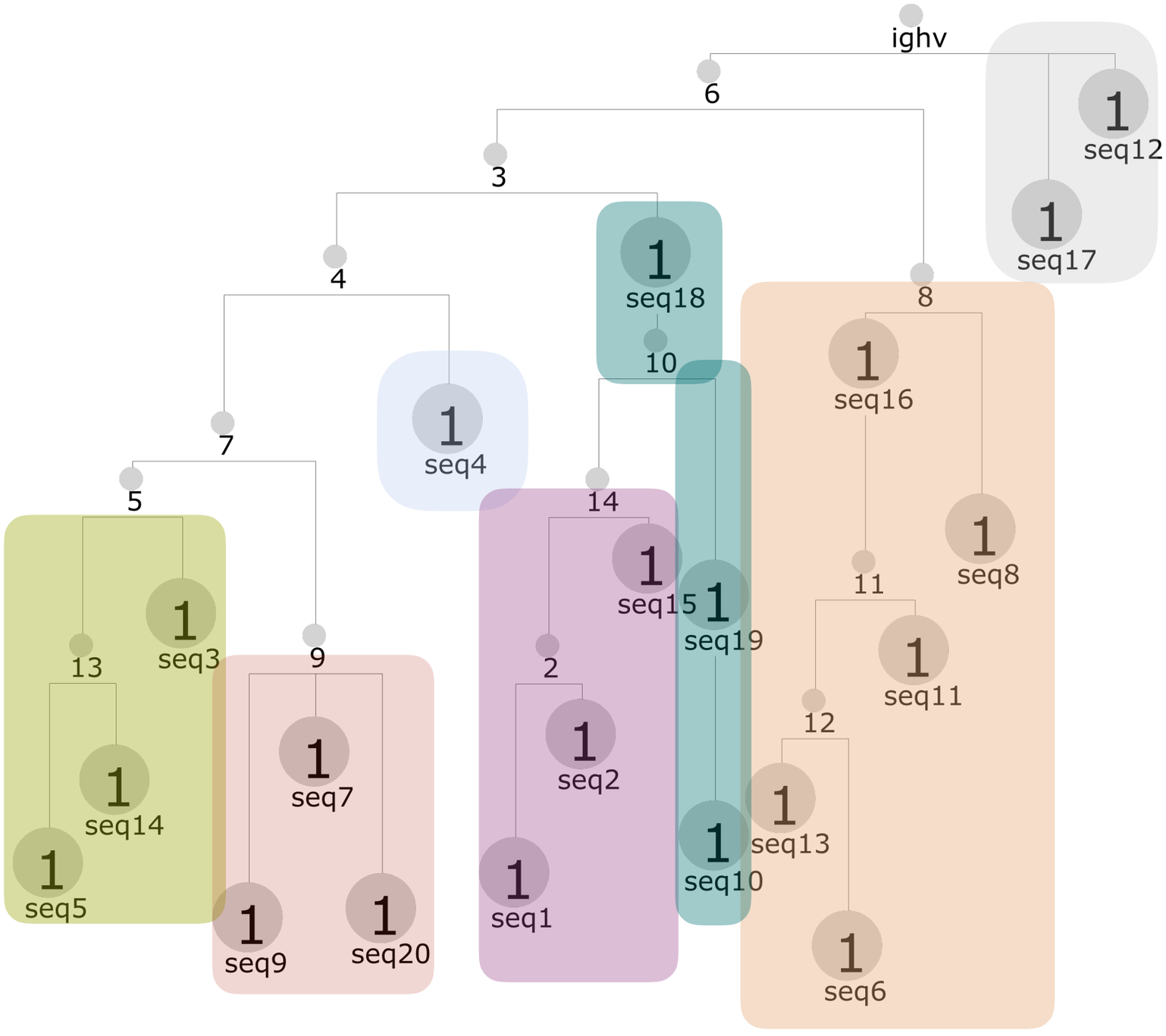
- containing 20 IGH chain sequences
- germline : the germline sequence used in the V(D)J rearrangement that defines this clonal family

Gap as un unknown nucleotide

True tree (Absim)

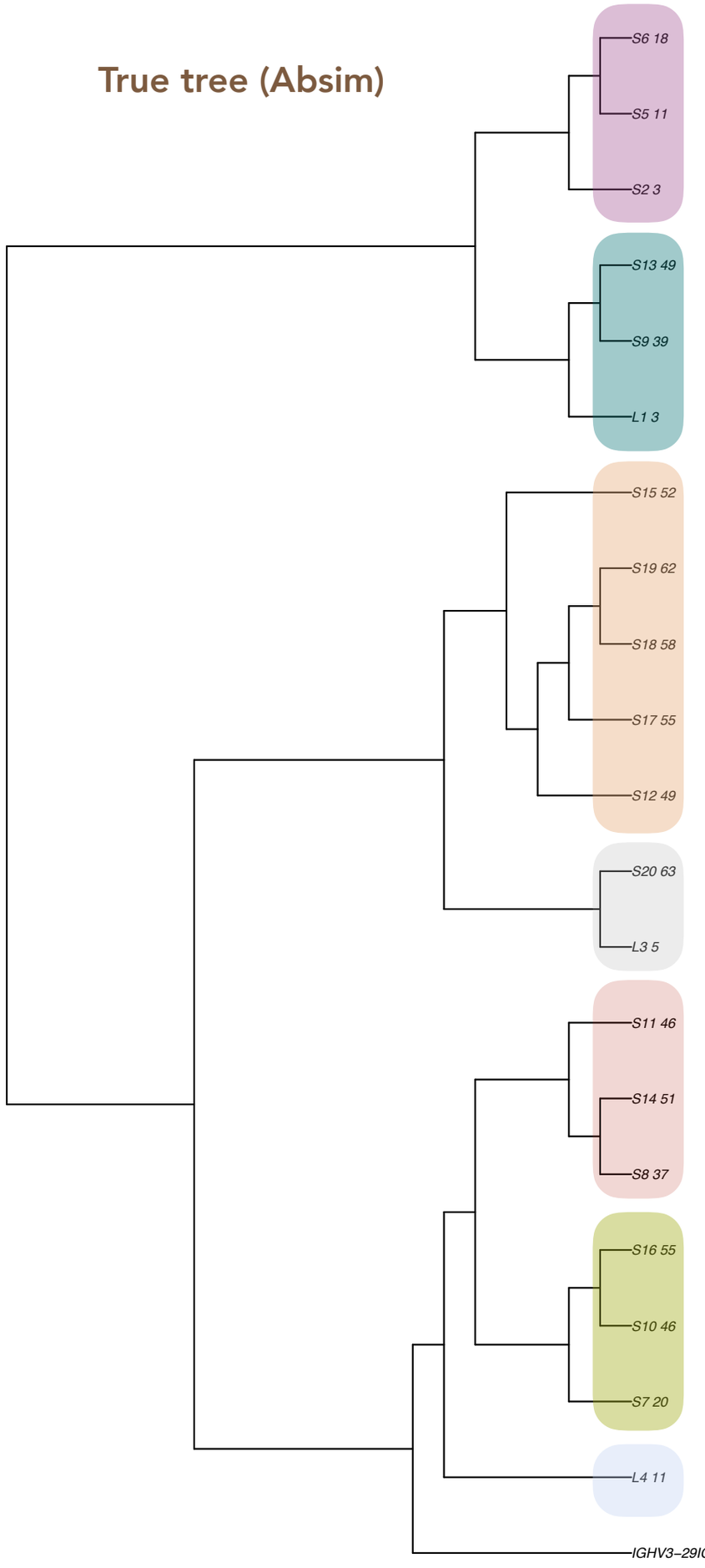


GC tree rendered tree

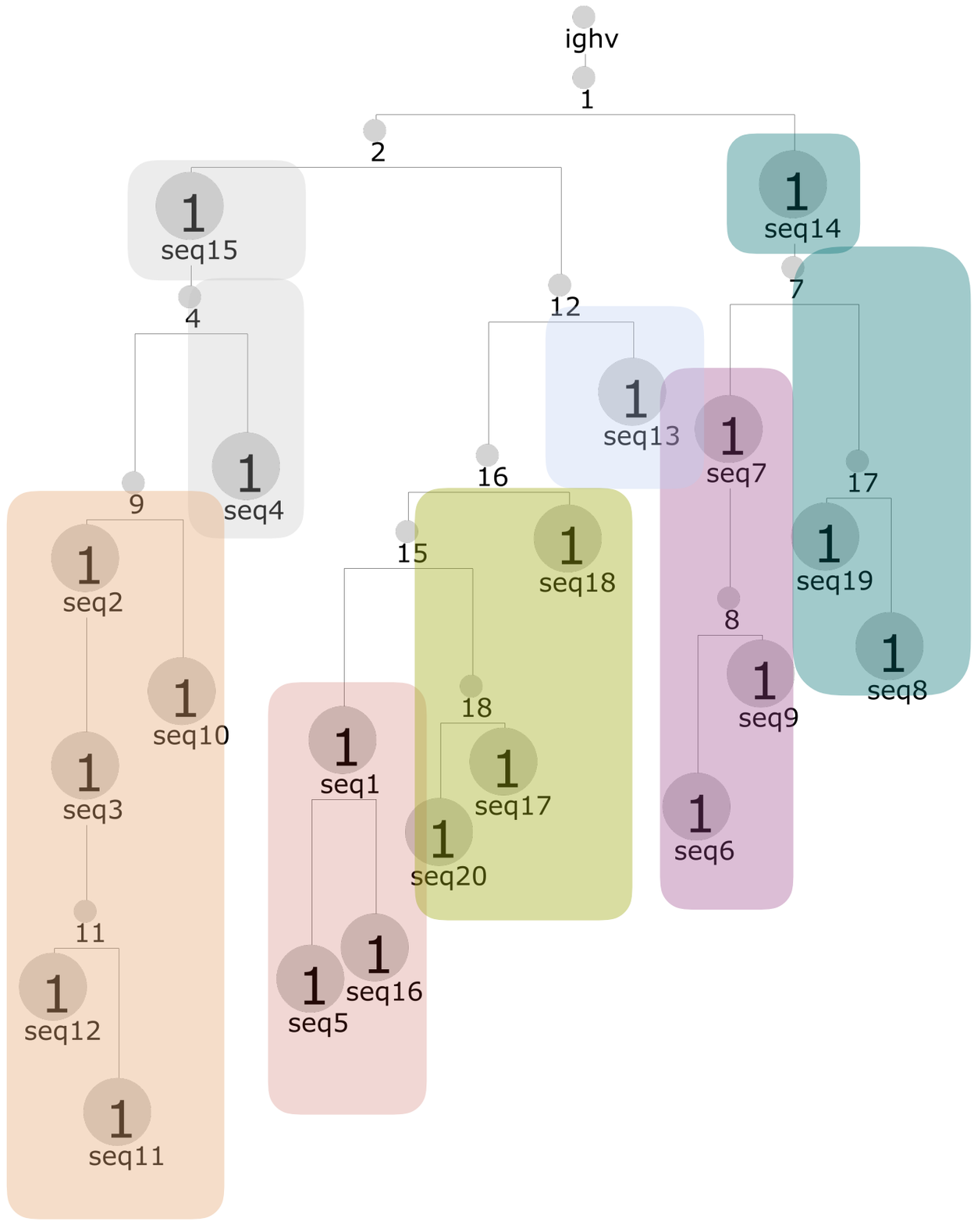


Gap as a fifth nucleotide state

True tree (Absim)



GC tree rendered tree



Study of intraclonal-diversity in a chronic lymphocytic leukaemia case Using Vidjil and GCtree

Pipeline

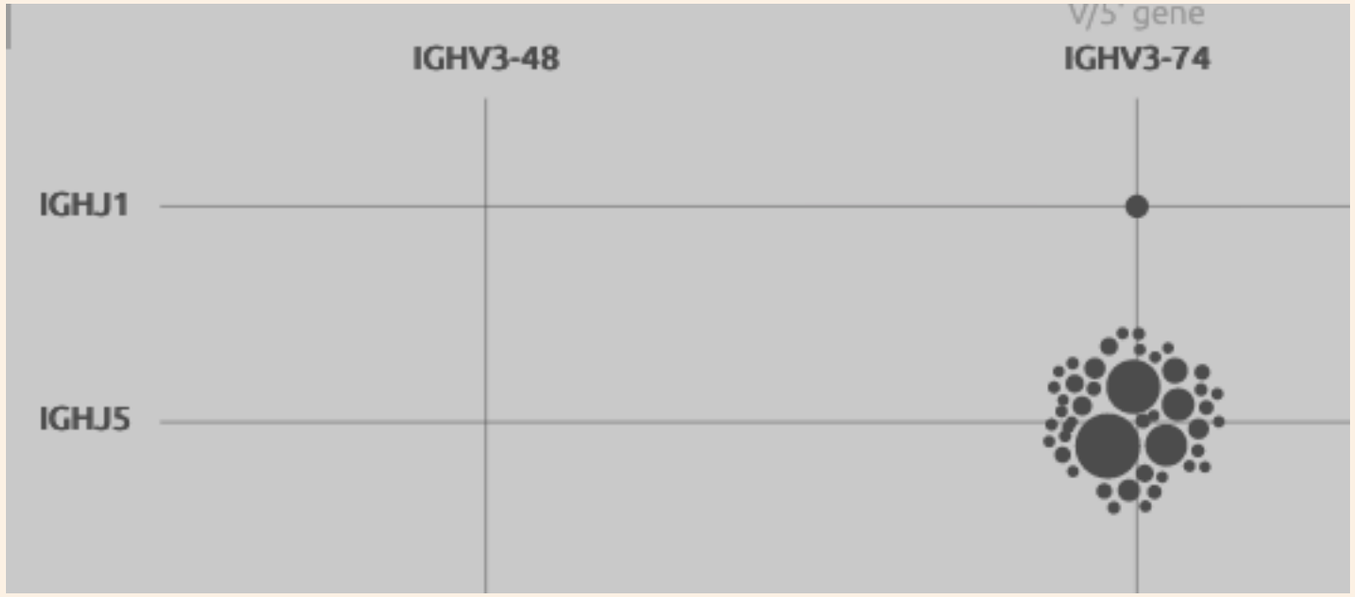
P2 dataset with 33599 seq

Run Vidjil

Select 10 most abundant clones

Run GCtree on the selection (consensus sequence)

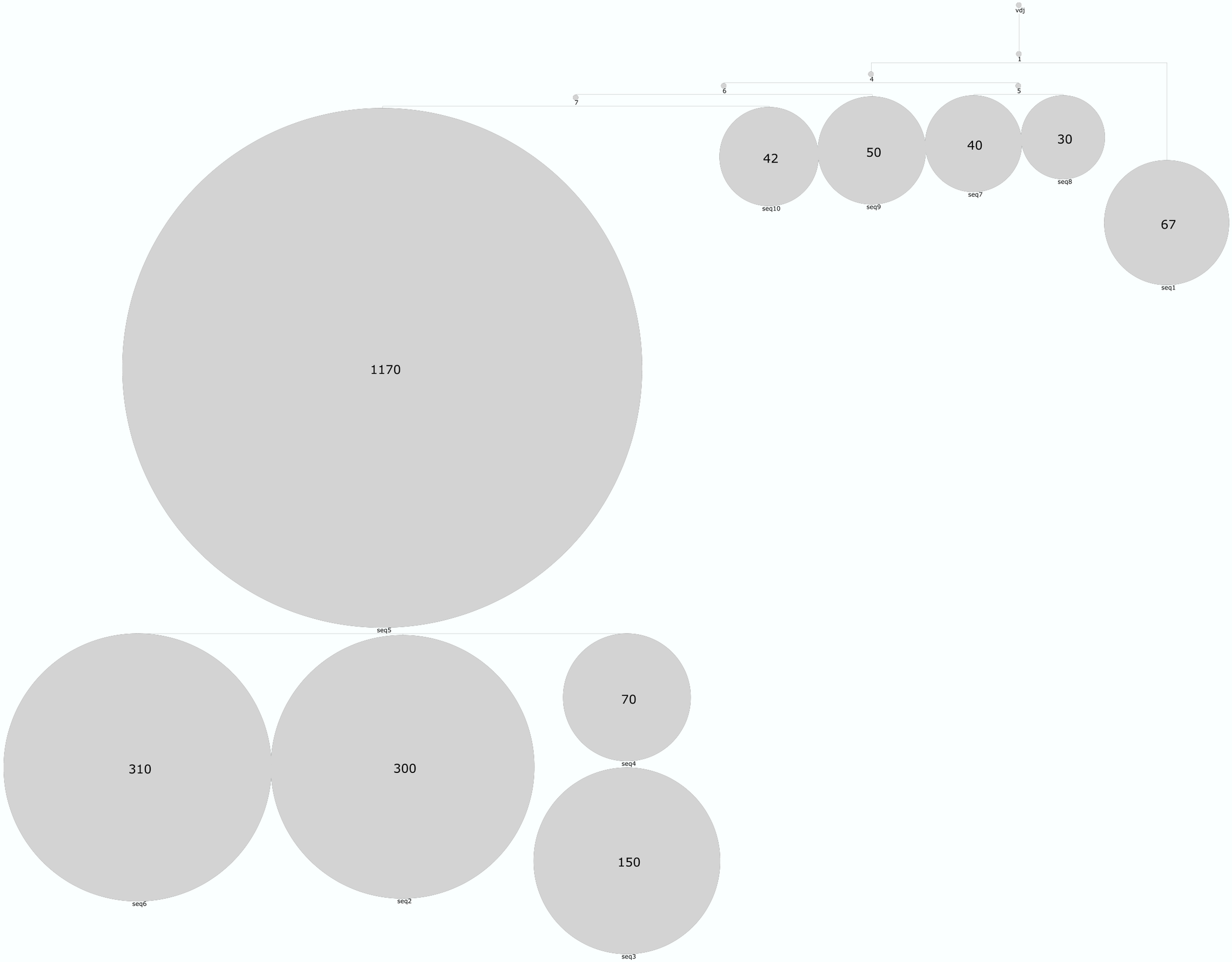
P2 major clones and satellites



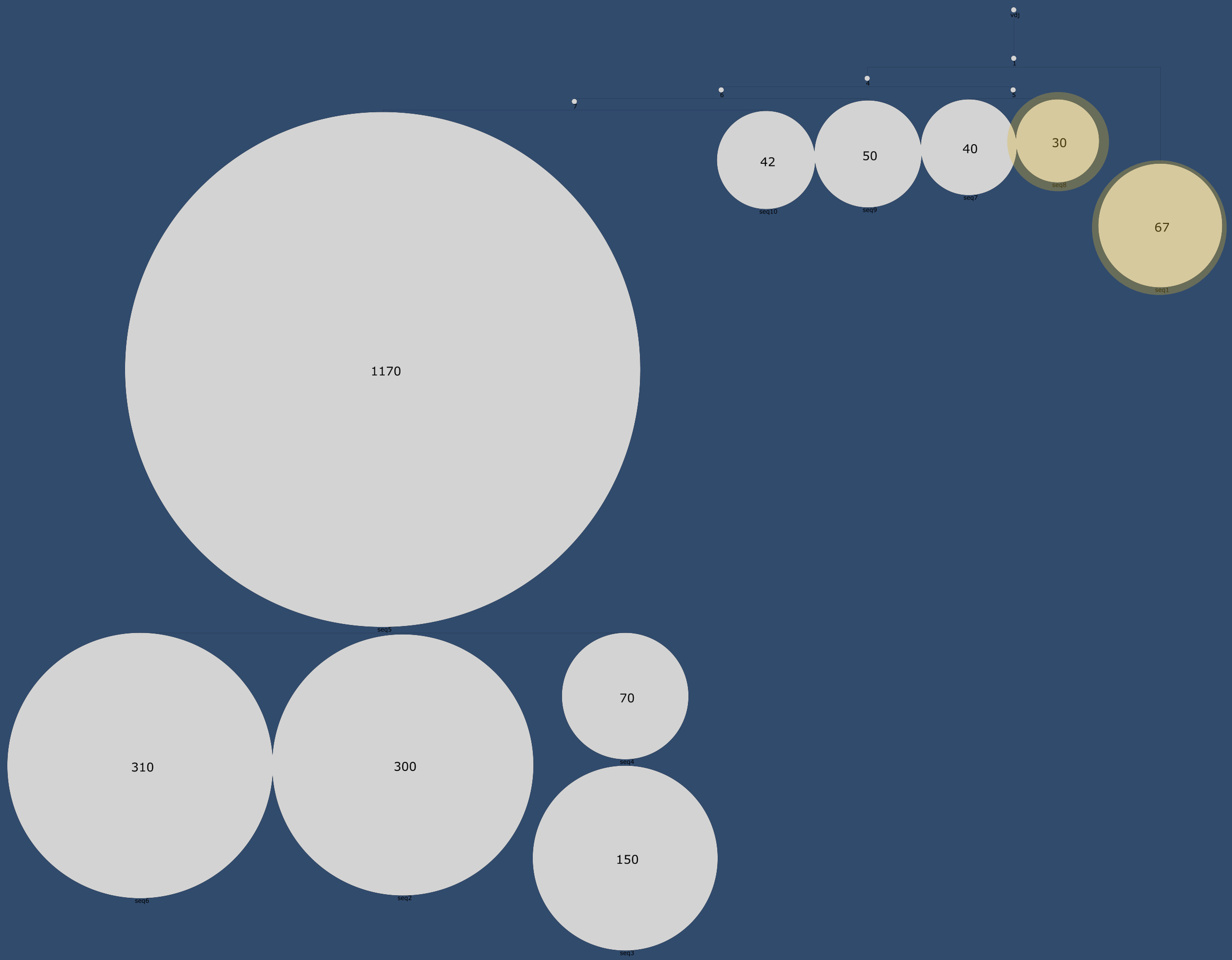
P2 major clones and satellites

Clone rank	# read	% read	IGHV	IGHJ
clone 1	11772	35 %	V3-74*01	J5*02
clone 2	6778	20 %	V3-74*02	J5*02
clone 3	3191	10 %	V3-74*01	J5*02
clone 4	1494	4 %	V3-74*01	J5*02
clone 5	699	2 %	V3-74*01	J5*02
clone 6	513	2 %	V3-74*03	J1*01
clone 7	464	1 %	V3-74*01	J5*02
clone 8	395	1 %	V3-74*01	J5*02
clone 9	369	1 %	V3-74*01	J5*02
clone 10	301	0,08 %	V3-74*01	J5*02

Intraclonal diversity via GCtree



Intraclonal diversity



Intraclonal diversity

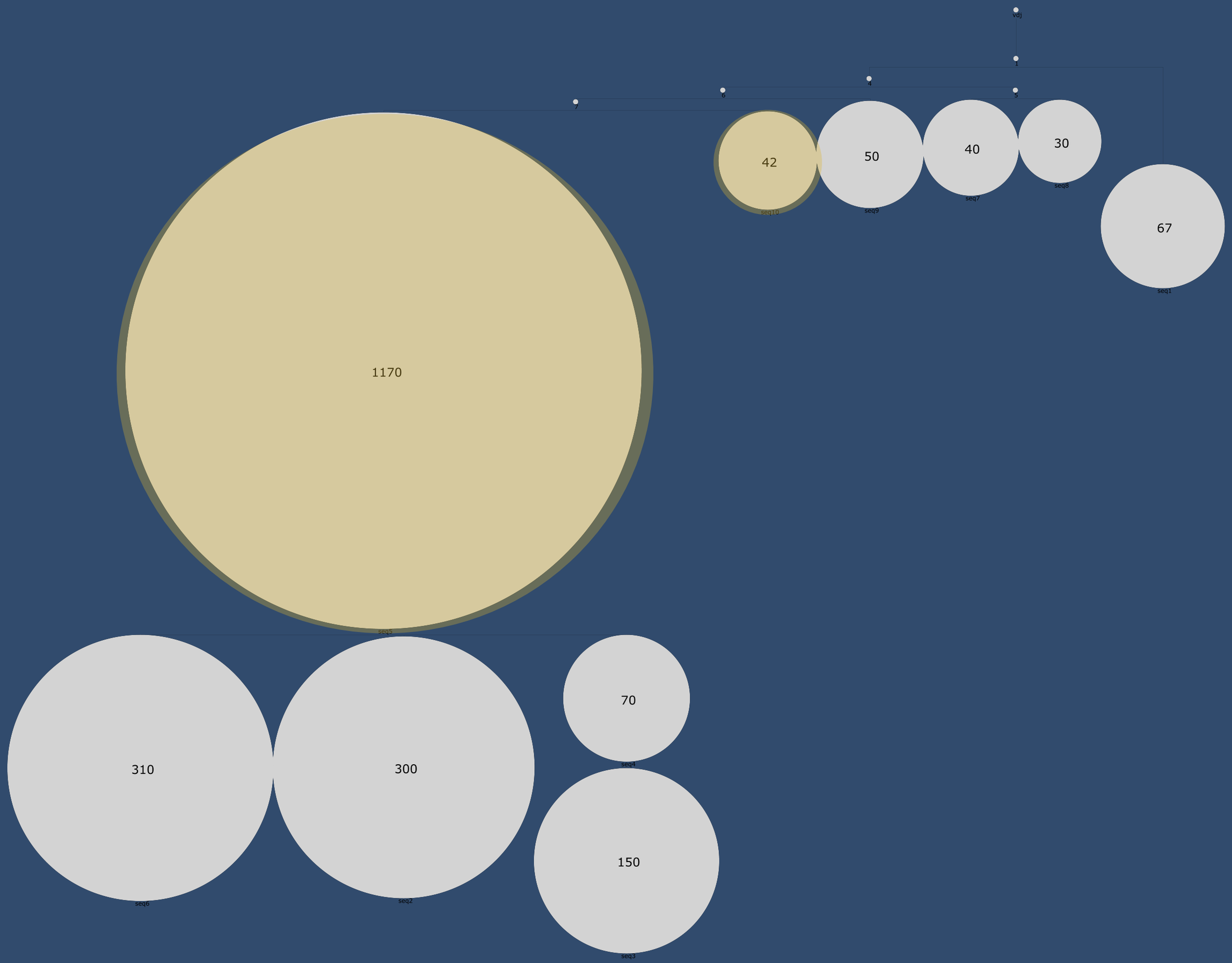
30

	104	105	106	107	108	109	110	111	111.1	111.2	111.3	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI	PhysicoChemical Descriptor (by BRFAA)
	C	A	R	D	Q	G	S	A	D	T	G	V	G	T	A	V	P	F	D	S	W					
30	tgt	gca	aga	gat	cag	ggg	tca	gcg	gat	aca	ggt	gta	gga	aca	gcc	gta	ccc	ttt	gac	tcc	tgg	+	19	2,140.28	4.15	CARDQGSADTGVGTAVPFDSW

67

	104	105	106	107	108	109	110	111	111.1	111.2	111.3	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI	PhysicoChemical Descriptor (by BRFAA)
	C	A	R	D	Q	G	S	A	D	T	G	V	G	T	A	V	P	F	D	S	W					
67	tgt	gca	aga	gat	cag	ggg	tca	gcg	gat	aca	ggt	gtg	ggt	aca	gcc	gta	ccc	ttt	gac	tcc	tgg	+	19	2,140.28	4.15	CARDQGSADTGVGTAVPFDSW

Intraclonal diversity



Intraclonal diversity

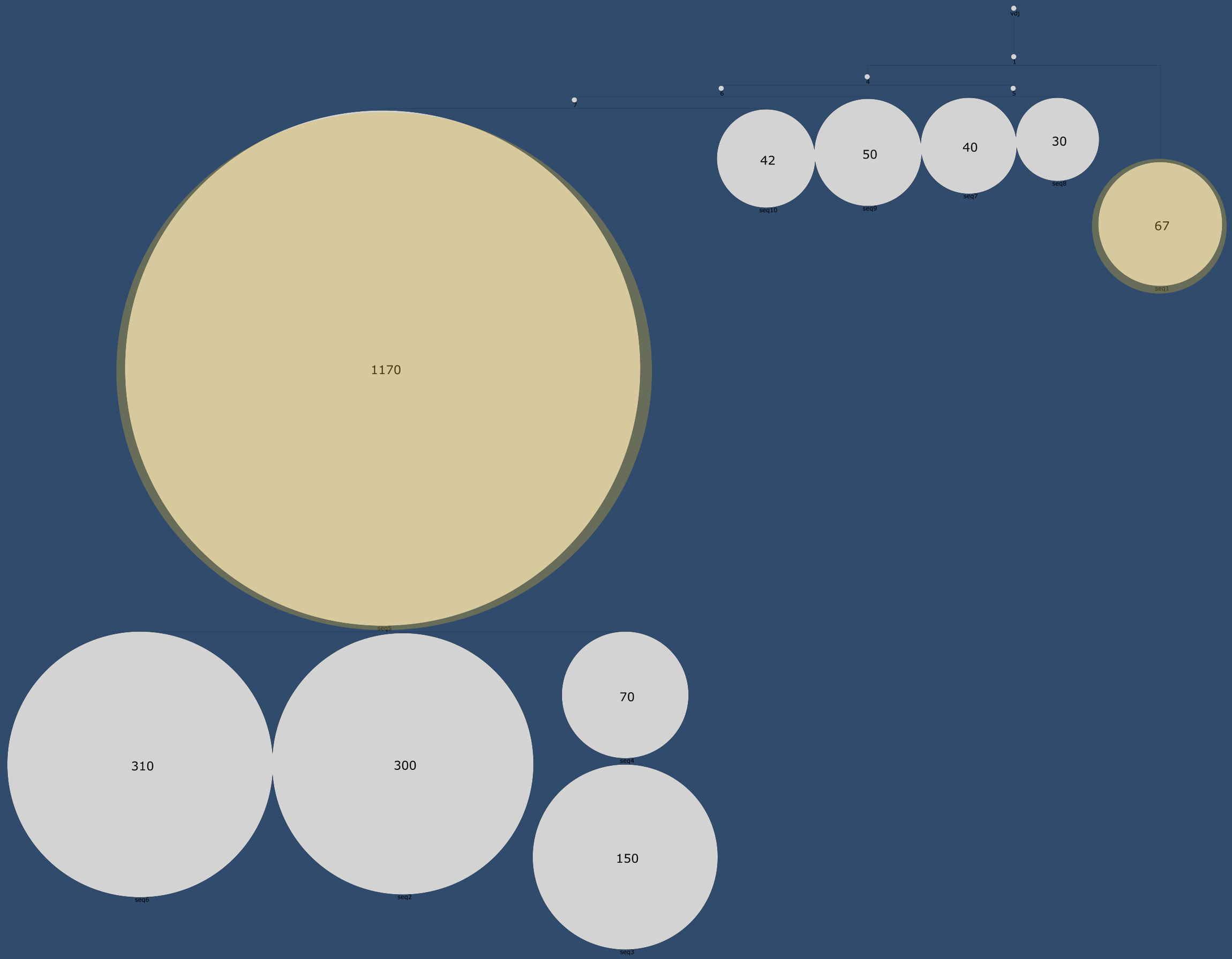
1170

	104	105	106	107	108	109	110	111	111.1	111.2	111.3	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI	PhysicoChemical Descriptor (by BRFAA)
	C	V	R	D	Q	G	S	A	D	T	G	V	G	T	A	V	P	F	D	S	W					
1170	tgt	gta	aga	gat	cag	ggg	tca	gcg	gac	aca	ggt	gtg	ggc	aca	gcc	gta	ccc	ttt	gac	tca	tgg	+	19	2,168.33	4.15	CVRDQGSADTGVGTAVPFDSW

42

	104	105	106	107	108	109	110	111	111.1	111.2	111.3	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI	PhysicoChemical Descriptor (by BRFAA)
	C	A	R	D	Q	G	S	A	D	T	G	V	G	T	A	V	P	F	D	S	W					
42	tgt	gca	aga	gat	cag	ggg	tca	gcg	gac	aca	ggt	gtg	ggc	aca	gcc	gtc	ccc	ttt	gac	tca	tgg	+	19	2,140.28	4.15	CARDQGSADTGVGTAVPFDSW

Intraclonal diversity



Intraclonal diversity

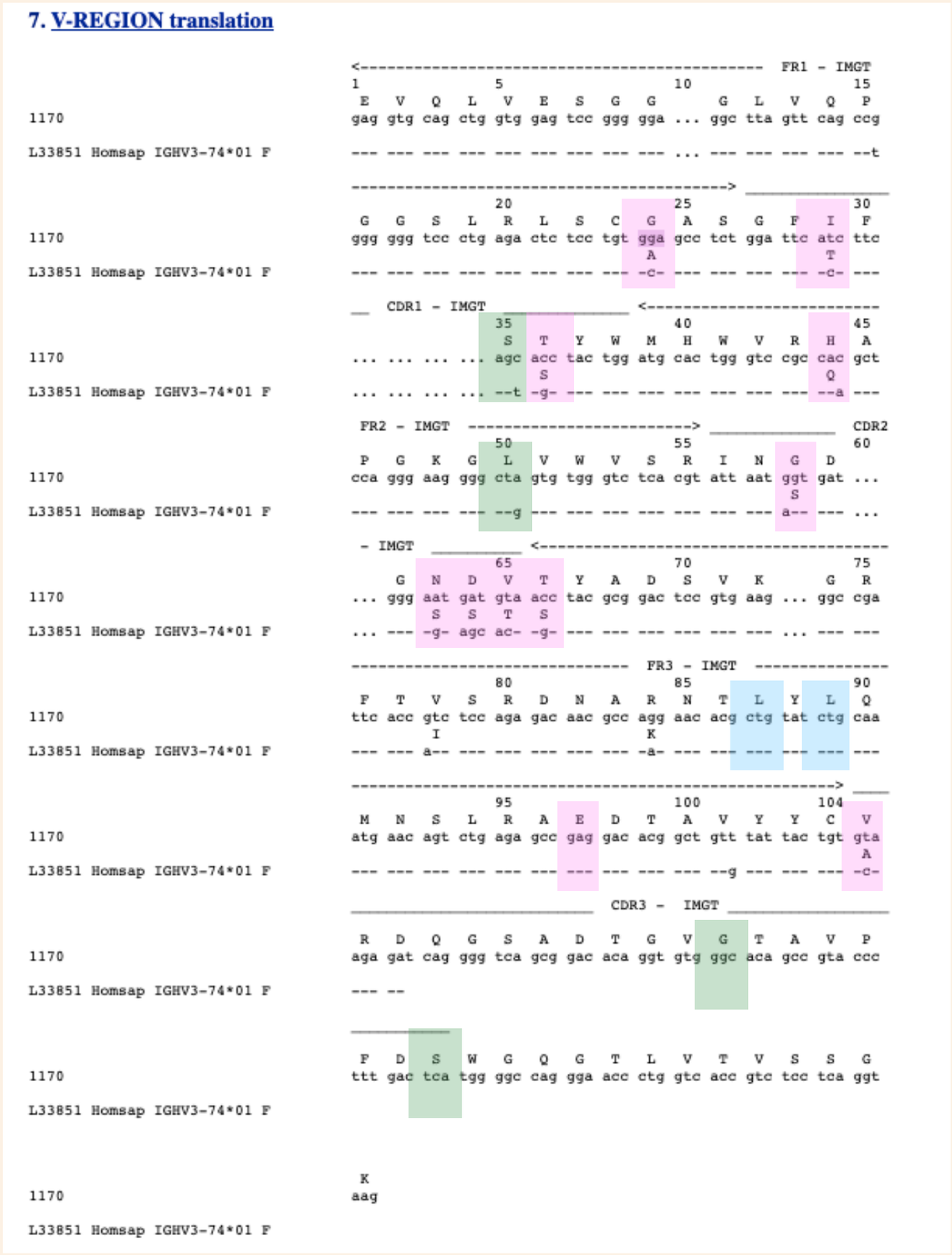
1170

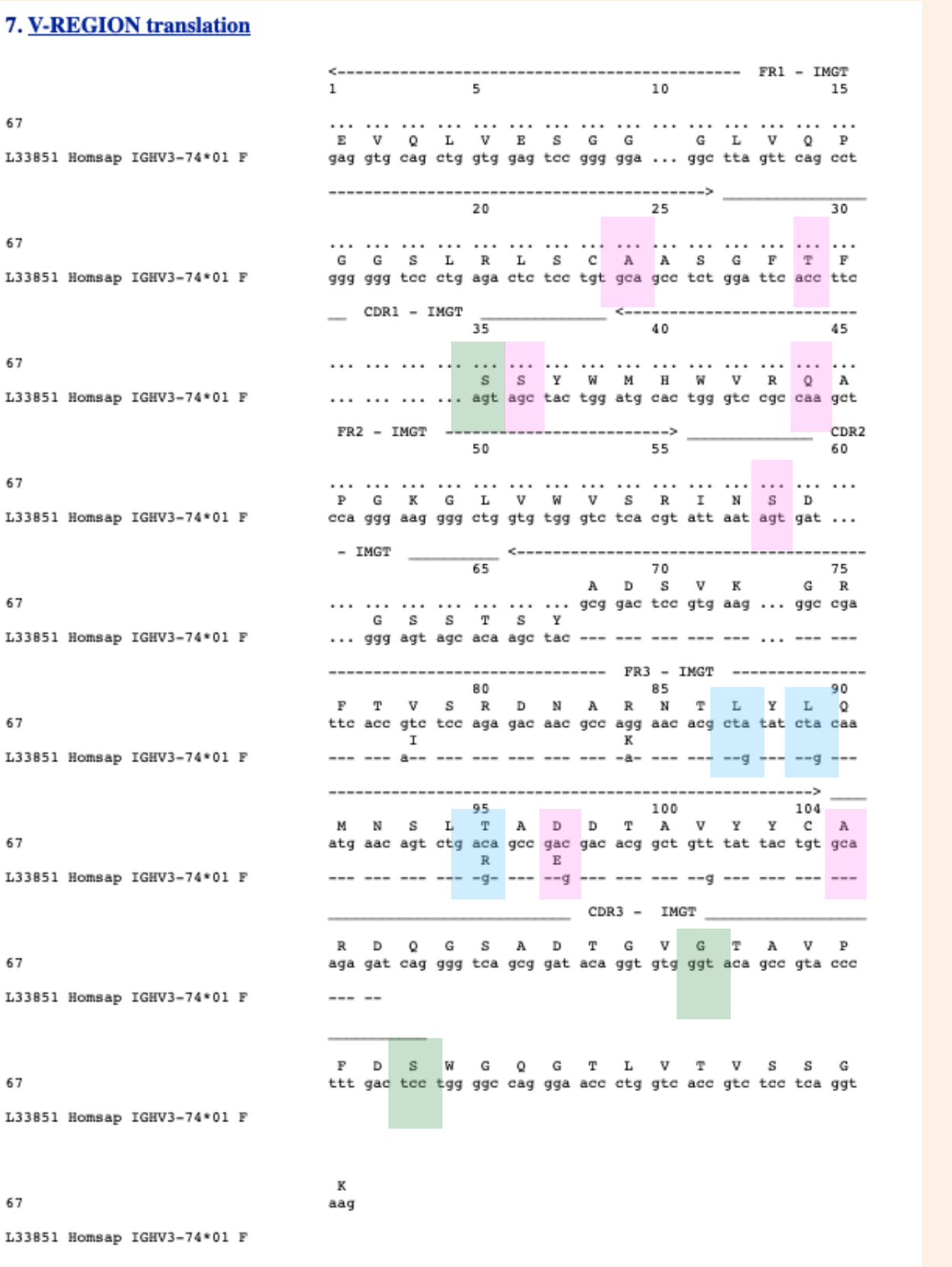
	104	105	106	107	108	109	110	111	111.1	111.2	111.3	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI	PhysicoChemical Descriptor (by BRFAA)
	C	V	R	D	Q	G	S	A	D	T	G	V	G	T	A	V	P	F	D	S	W					
1170	tgt	gta	aga	gat	cag	ggg	tca	gcg	gac	aca	ggt	gtg	ggc	aca	gcc	gta	ccc	ttt	gac	tca	tgg	+	19	2,168.33	4.15	CVRDQGSADTGVGTAVPFDSW

67

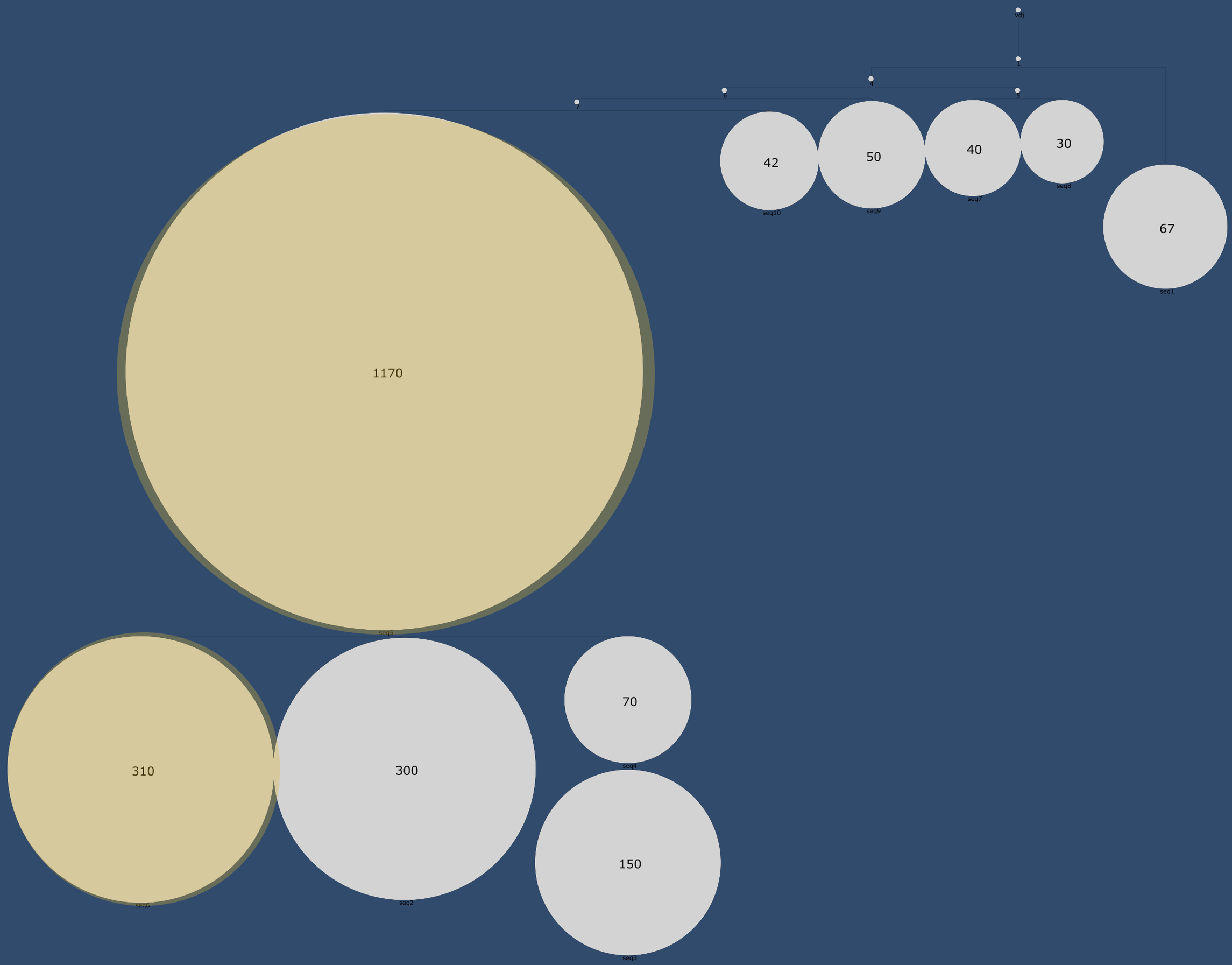
	104	105	106	107	108	109	110	111	111.1	111.2	111.3	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI	PhysicoChemical Descriptor (by BRFAA)
	C	A	R	D	Q	G	S	A	D	T	G	V	G	T	A	V	P	F	D	S	W					
67	tgt	gca	aga	gat	cag	ggg	tca	gcg	gat	aca	ggt	gtg	ggt	aca	gcc	gta	ccc	ttt	gac	tcc	tgg	+	19	2,140.28	4.15	CARDQGSADTGVGTAVPFDSW

Intraclonal diversity





Intraclonal diversity



Intraclonal diversity

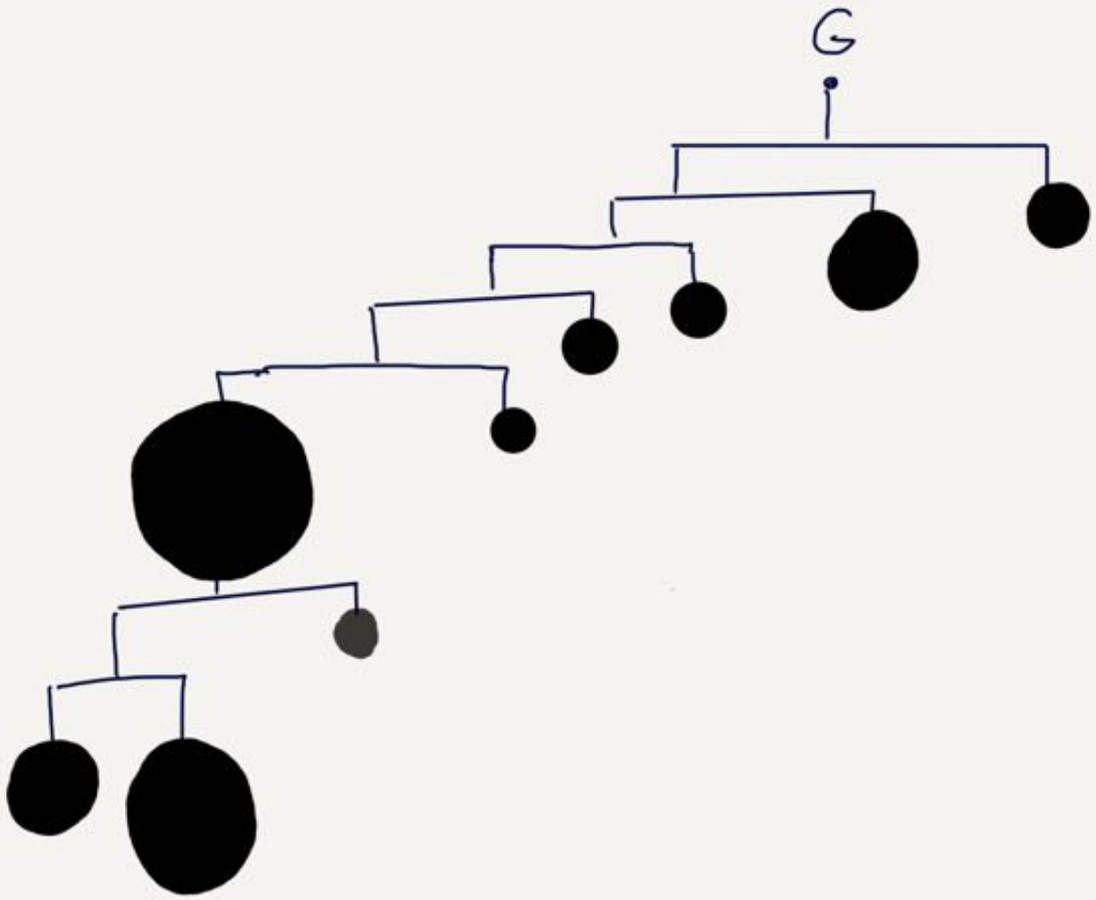
1170

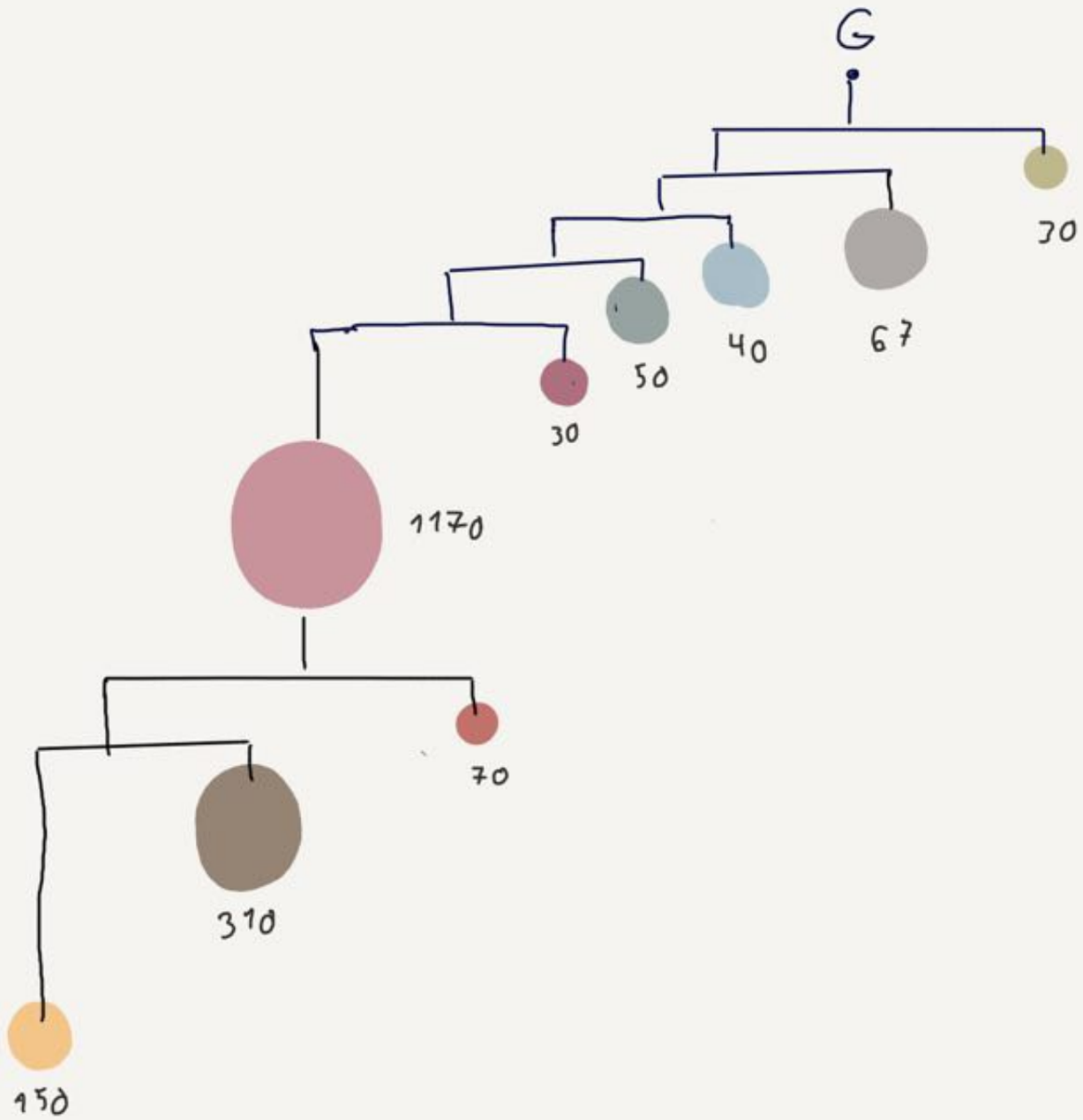
	104	105	106	107	108	109	110	111	111.1	111.2	111.3	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI	PhysicoChemical Descriptor (by BRFAA)
	C	V	R	D	Q	G	S	A	D	T	G	V	G	T	A	V	P	F	D	S	W					
1170	tgt	gta	aga	gat	cag	ggg	tca	gcg	gac	aca	ggt	gtg	ggc	aca	gcc	gta	ccc	ttt	gac	tca	tgg	+	19	2,168.33	4.15	CVRDQGSADTGVGTAVPFDSW

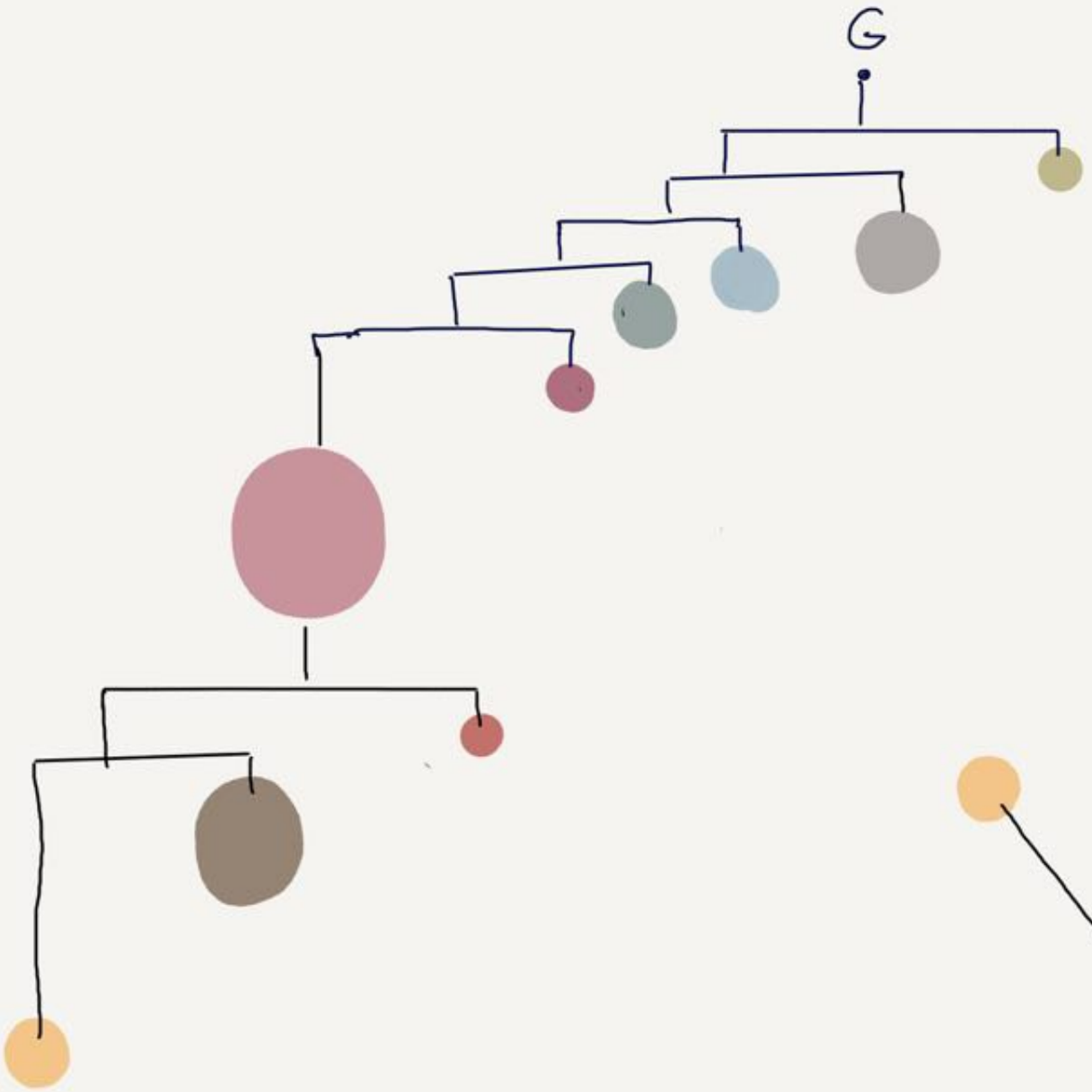
310

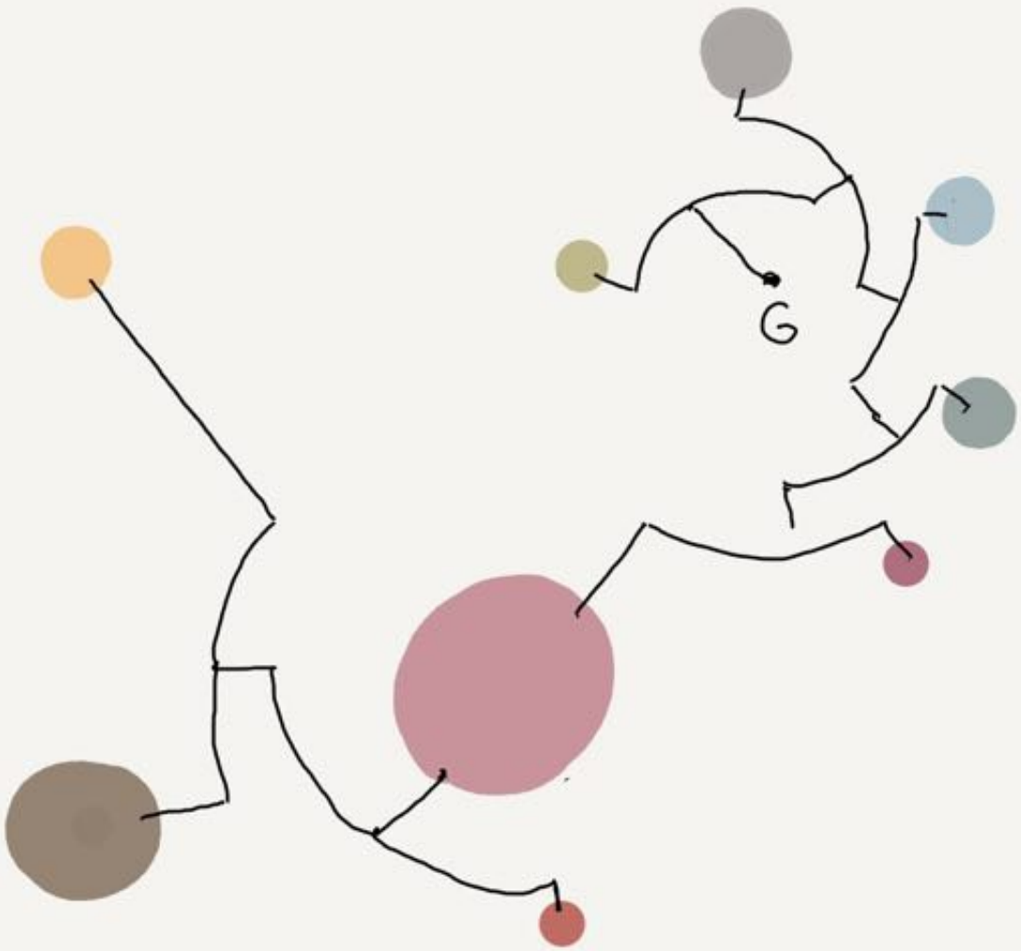
	104	105	106	107	108	109	110	111	111.1	111.2	111.3	112.3	112.2	112.1	112	113	114	115	116	117	118	Frame	CDR3-IMGT length	Molecular mass	pI	PhysicoChemical Descriptor (by BRFAA)
	C	V	R	D	Q	G	S	A	D	T	G	V	G	T	A	V	P	F	D	S	W					
310	tgt	gta	aga	gat	cag	ggg	tca	gcg	gac	aca	ggt	gtg	ggc	aca	gcc	gtg	ccc	ttt	gac	tca	tgg	+	19	2,168.33	4.15	CVRDQGSADTGVGTAVPFDSW

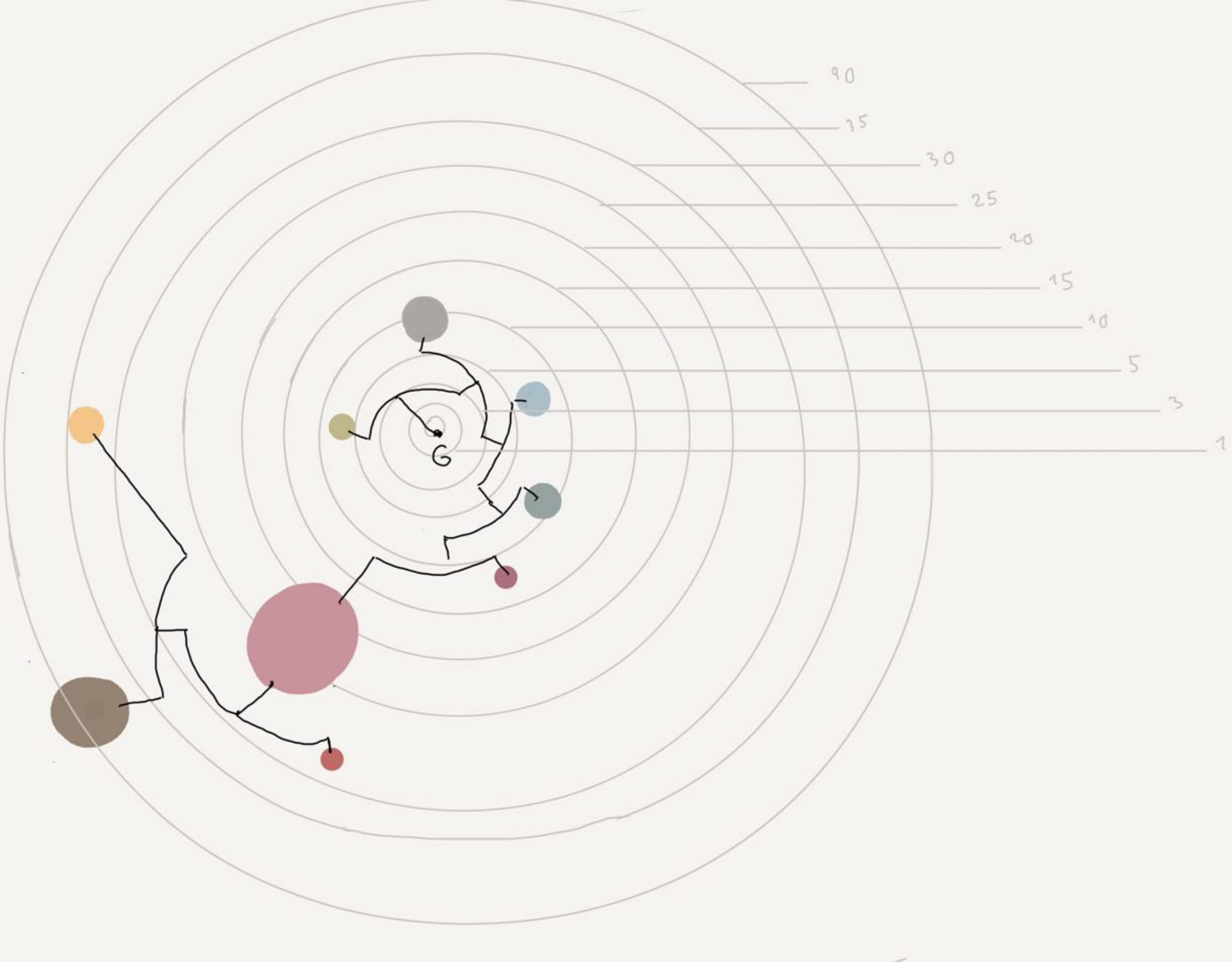
Visualisation prototype

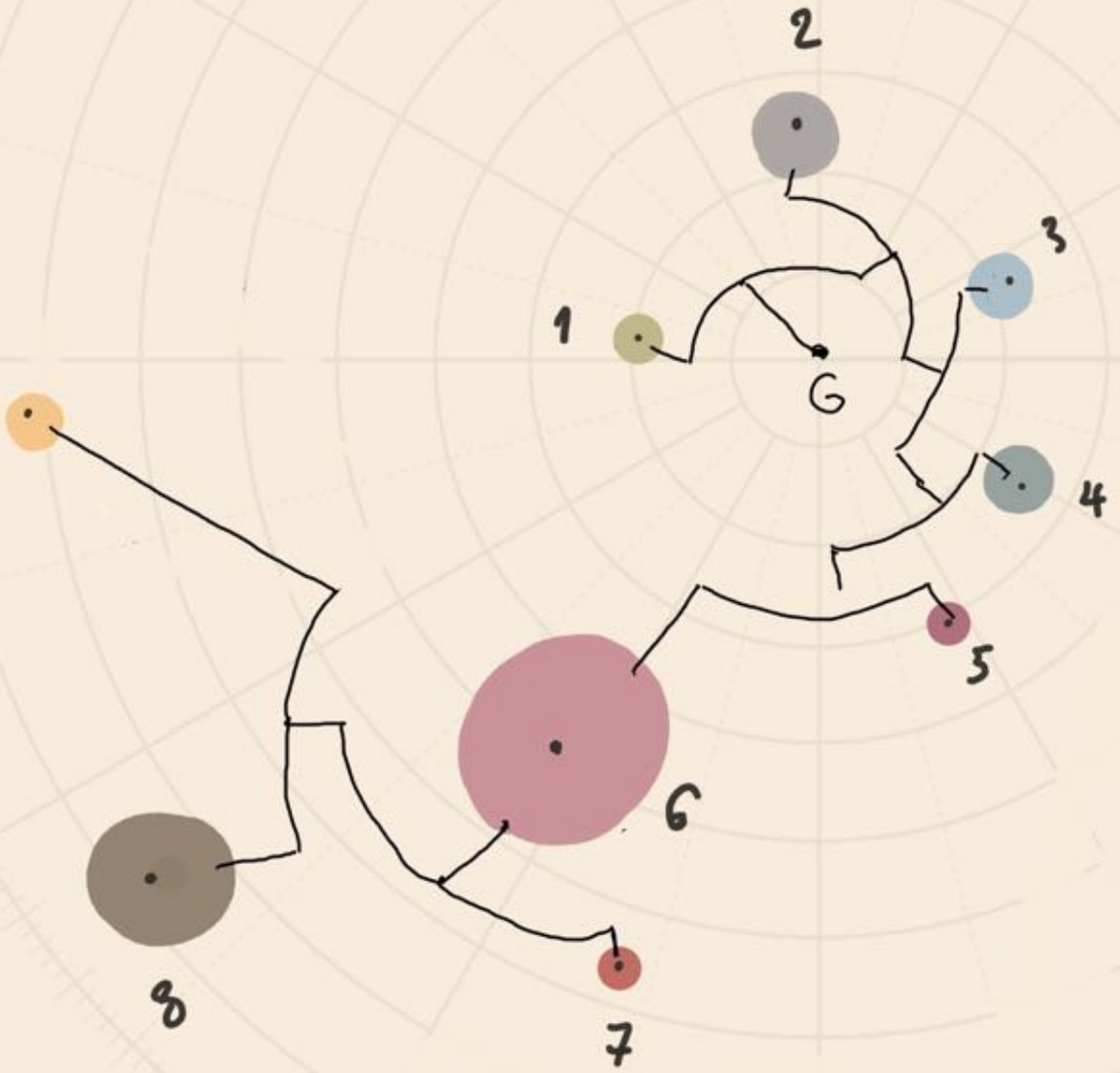



G
30
67
40
50
30
1170
70
310
150


G


G


90
75
30
25
20
15
10
5
3
1
G


2
3
1
G
4
5
6
8
7

Benchmarking RepSeq analysis tools

Benchmarking RepSeq analysis tools

Improvements

Studied **eight** tools

Analyzed different repertoire simulators

Studied different performance evaluation measures and found out that for the same input data, using different tools results in having

- Variable number of clusters
- Variable cluster sizes (especially for the major clones of each repertoire)
- Various standards to accept a sequence as an input.

Prepared an *in vitro* benchmark to evaluate the clone detection caliber of each tool. Data set is constructed by a serial dilution of a known clone (1\%, 0.1\%,and 0.01\%) in a polyclonal background

Thank you