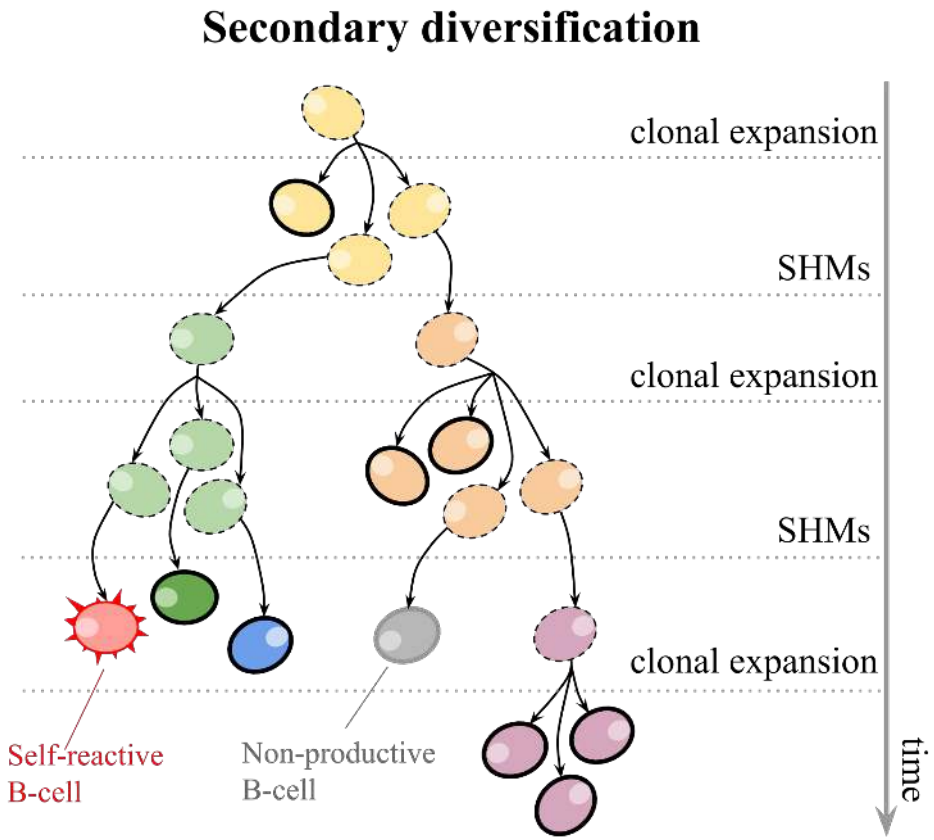
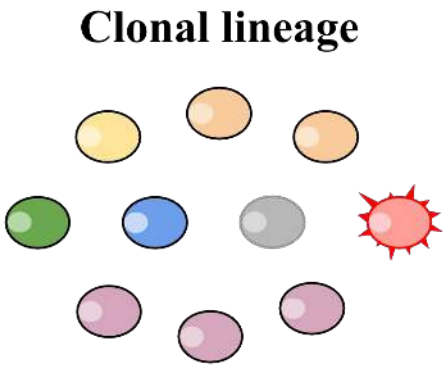
Reconstruction of clonal lineages for highly hypermutated repertoires

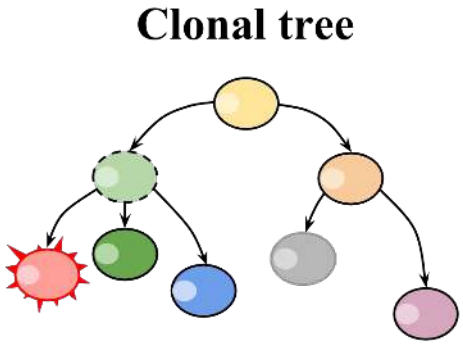
Maria Chernigovskaya

Scientific advisors:
Yana Safonova,
Andrey Slabodkin

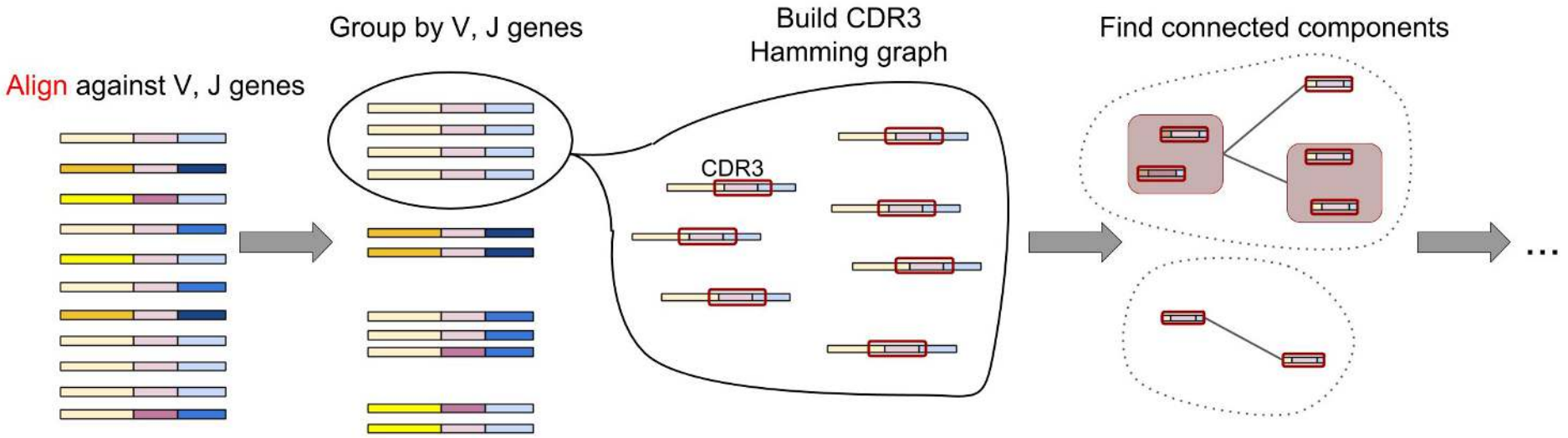
AntEvolo: an algorithm for construction of clonal trees for an antibody repertoire







AntEvolo first steps



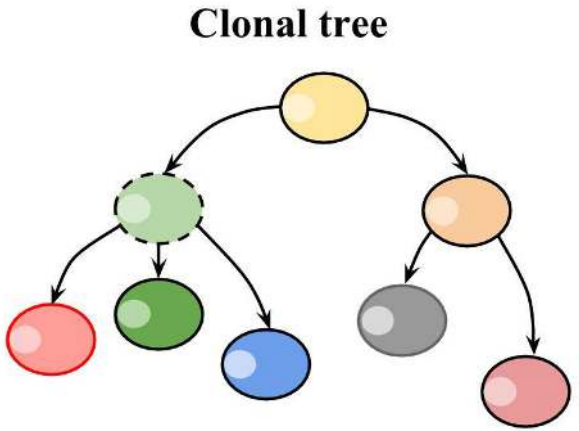
Clonal lineage = sequences from a single connected component of Hamming graph

Alignment problem

Sometimes alignment is inaccurate and sequences from the same lineage are assigned to a wrong VJ class

Why does VJ classification fail?

1. V and J genes are similar
2. Complicated data:
    a) deep clonal trees
    b) highly hypermutated data

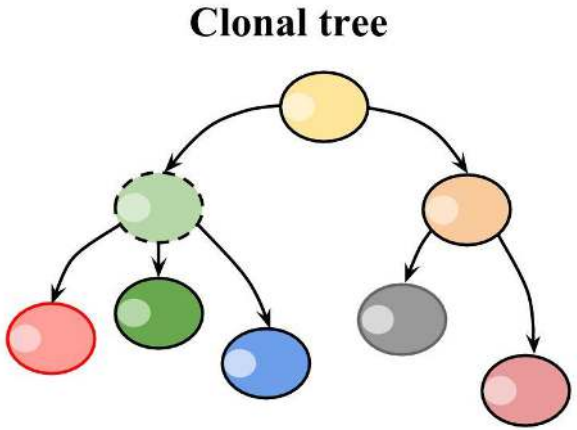


Alignment problem

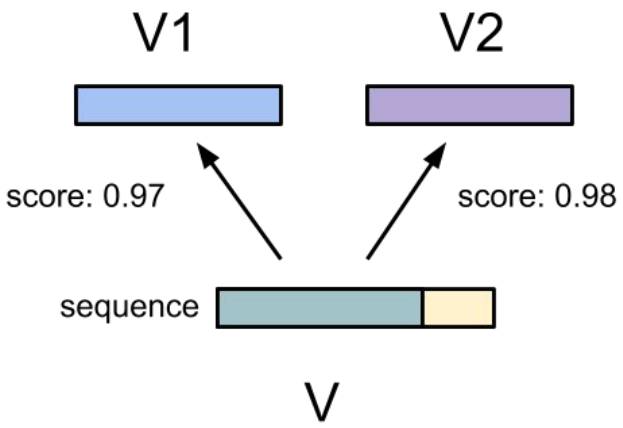
Sometimes alignment is inaccurate and sequences from the same lineage are assigned to a wrong VJ class

Why does VJ classification fail?

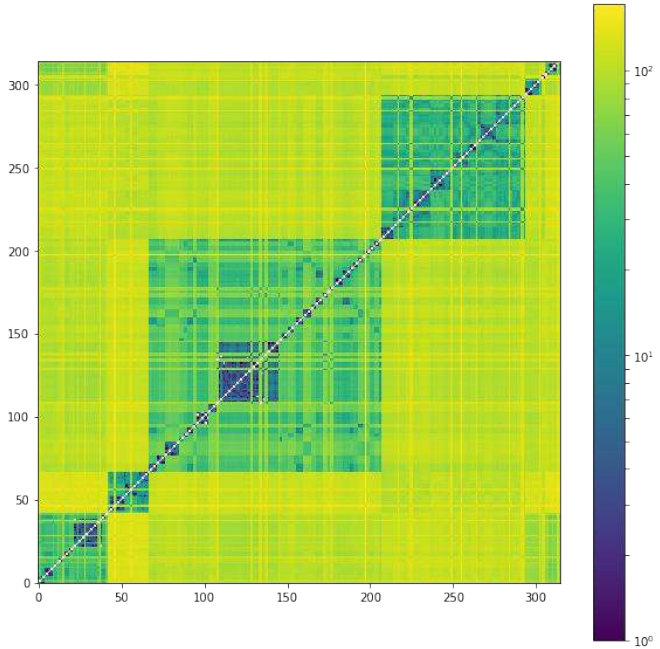
1. V and J genes are similar
2. Complicated data:
    a) deep clonal trees
    b) ~~highly hypermutated data~~



Similarity of V and J genes

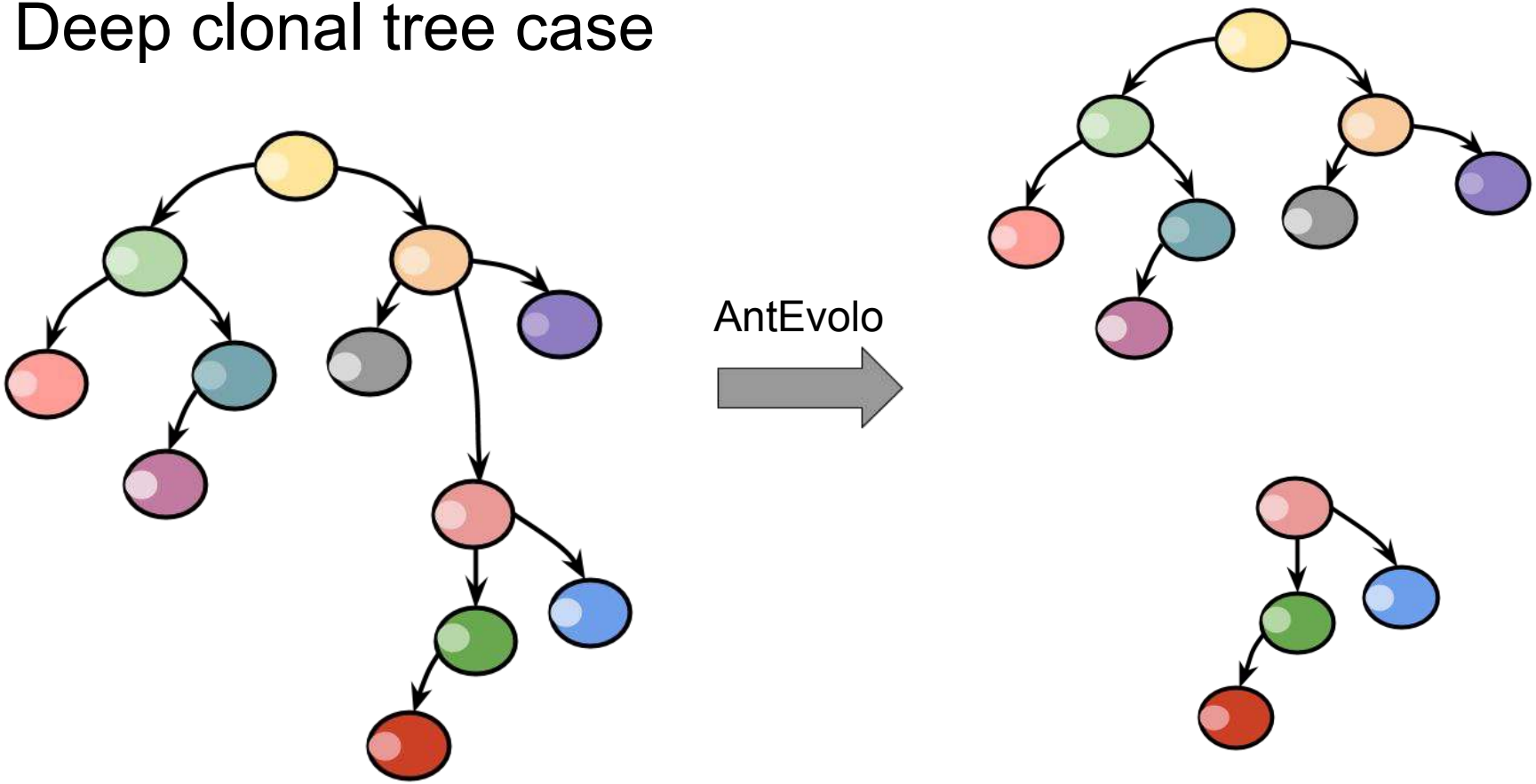


We chose V2
We also may choose V1

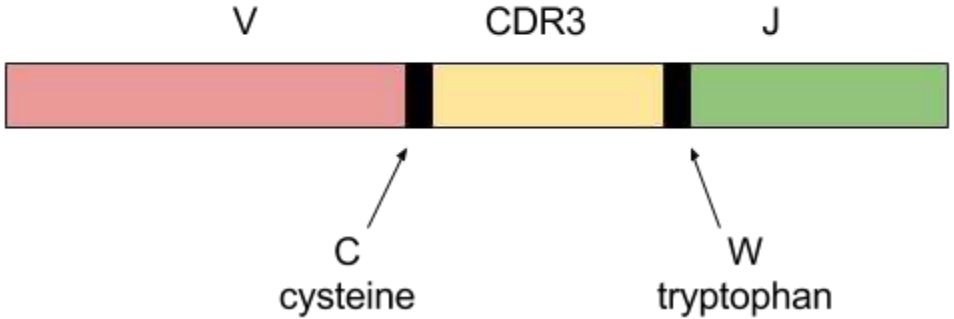


Distance matrix for V genes

Deep clonal tree case



Cropping idea

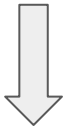


We take account of mutations in V/J genes in CDR3 region during construction of a tree. **That's wrong!**

We should crop V by the left bound of CDR3 (cysteine position) and J by the right bound of CDR3 (tryptophan position)

Similarity of V genes

315 genes

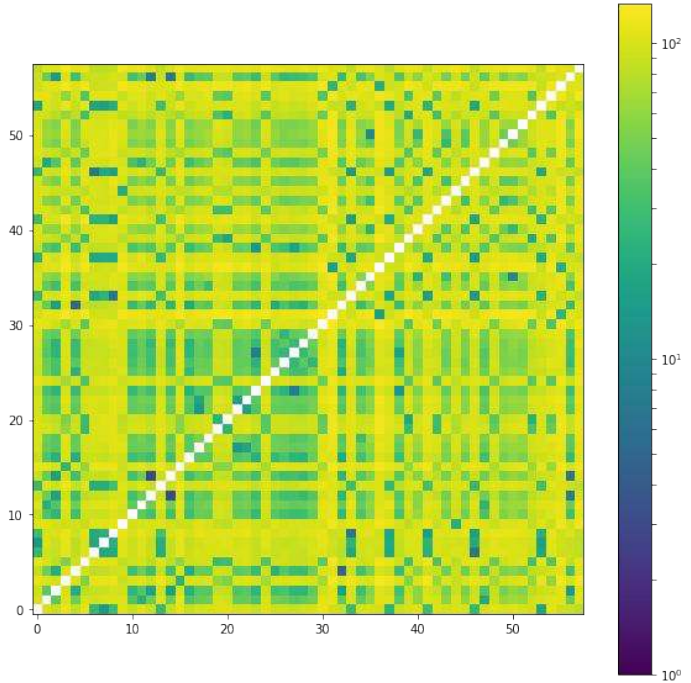


58 genes

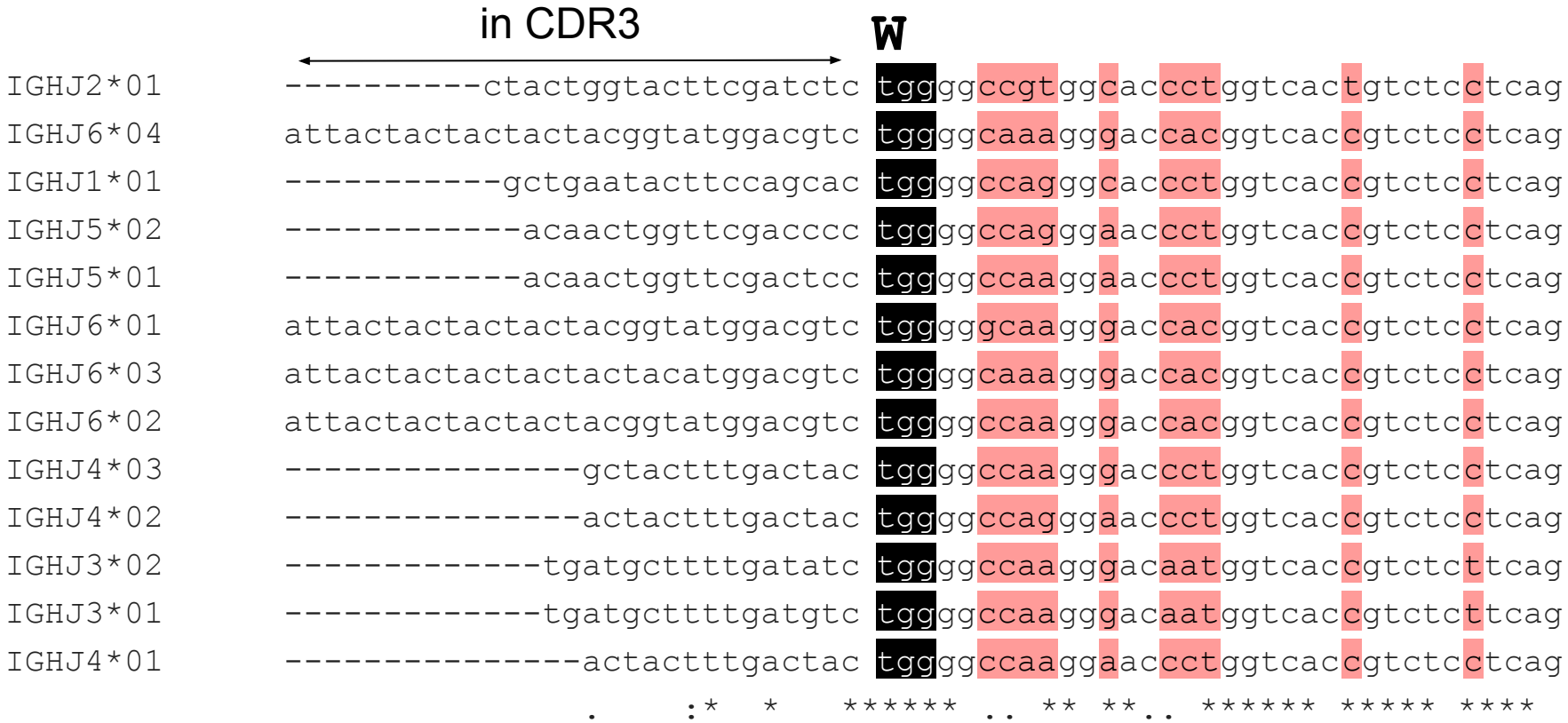
cropping + gluing allelic variations

min_distance = 5

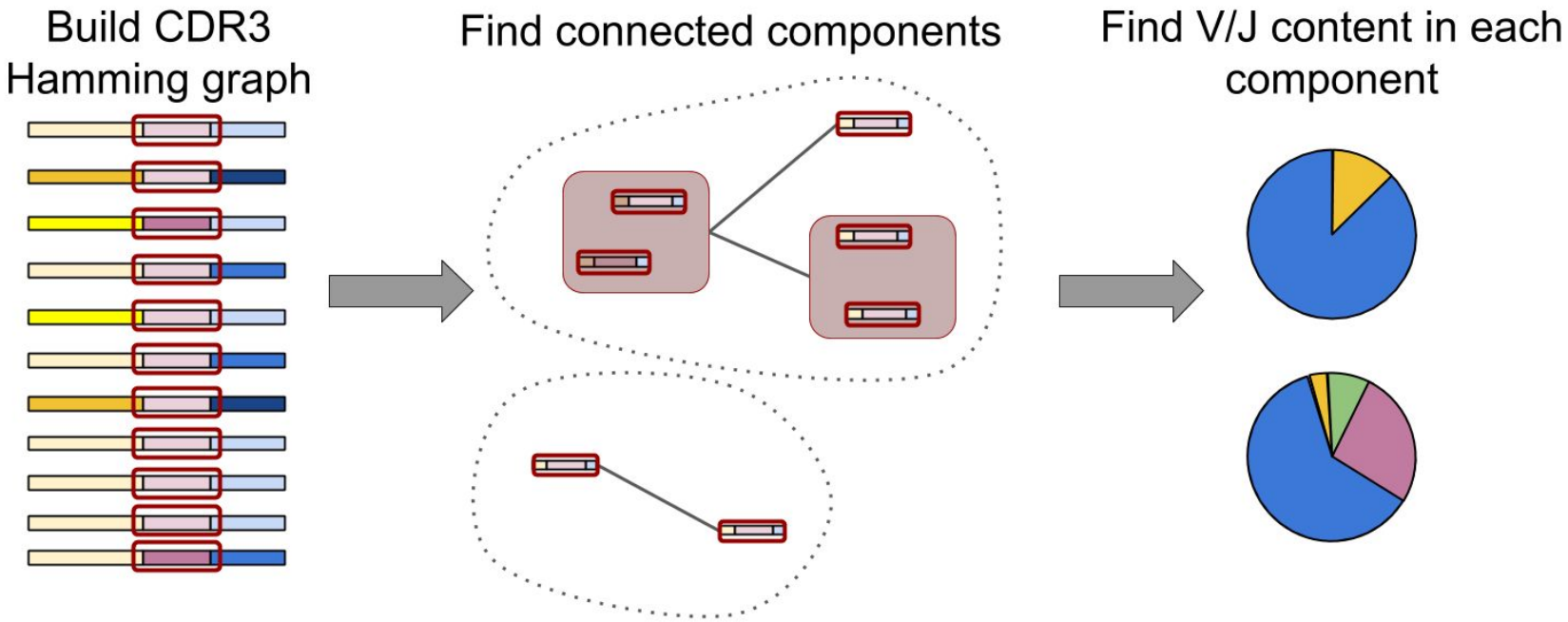
(≥ 10 for most of the new V genes)



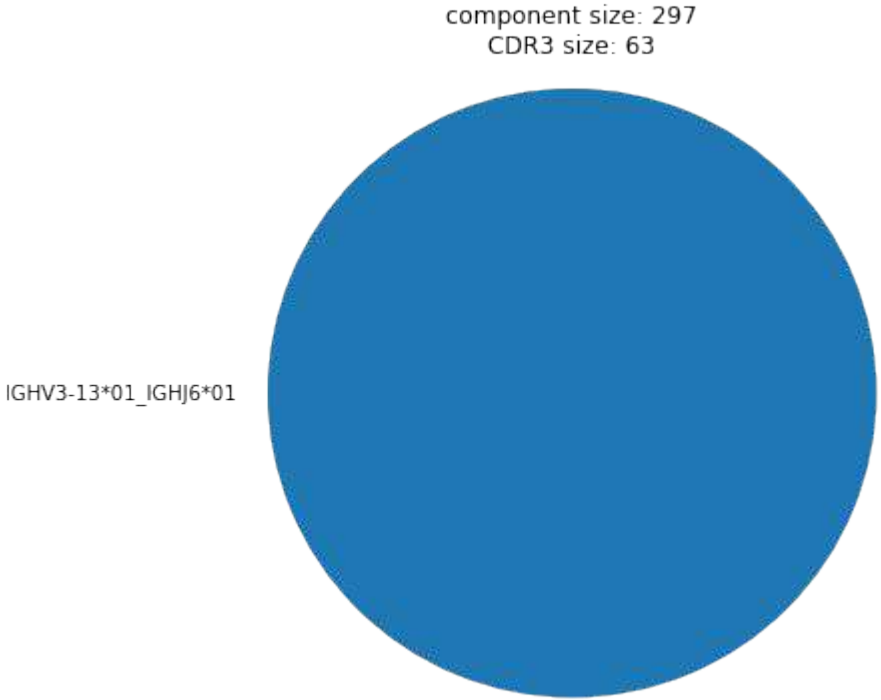
Similarity of J genes



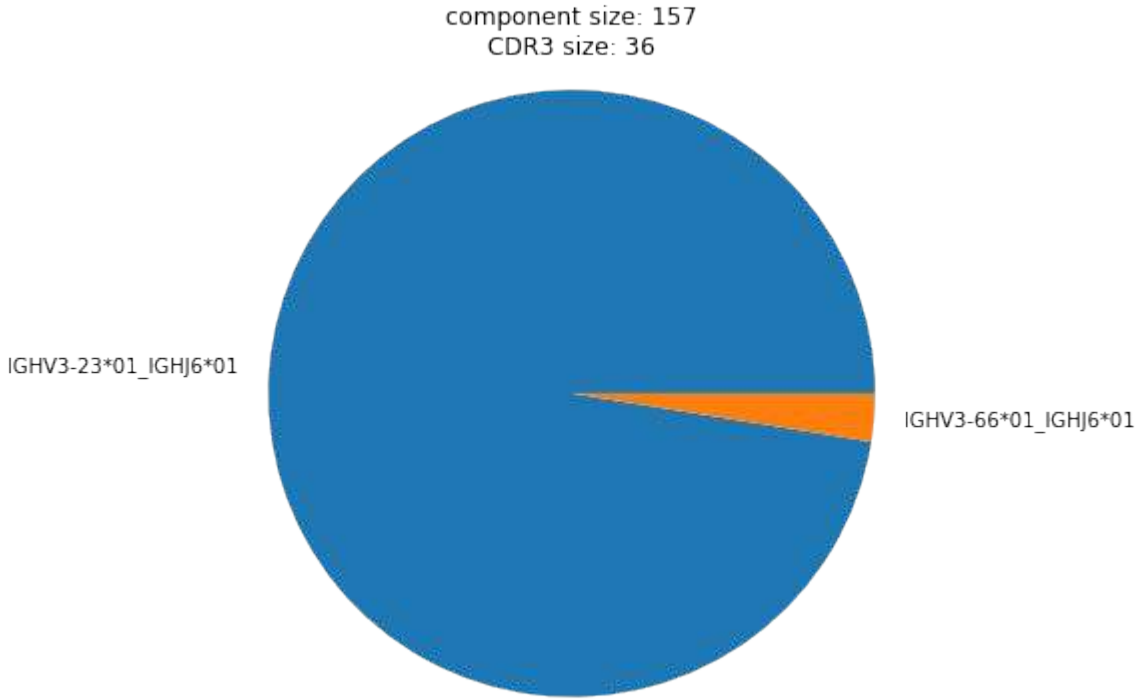
How can we verify alignment quality?



What we want to see



What we actually see: good case



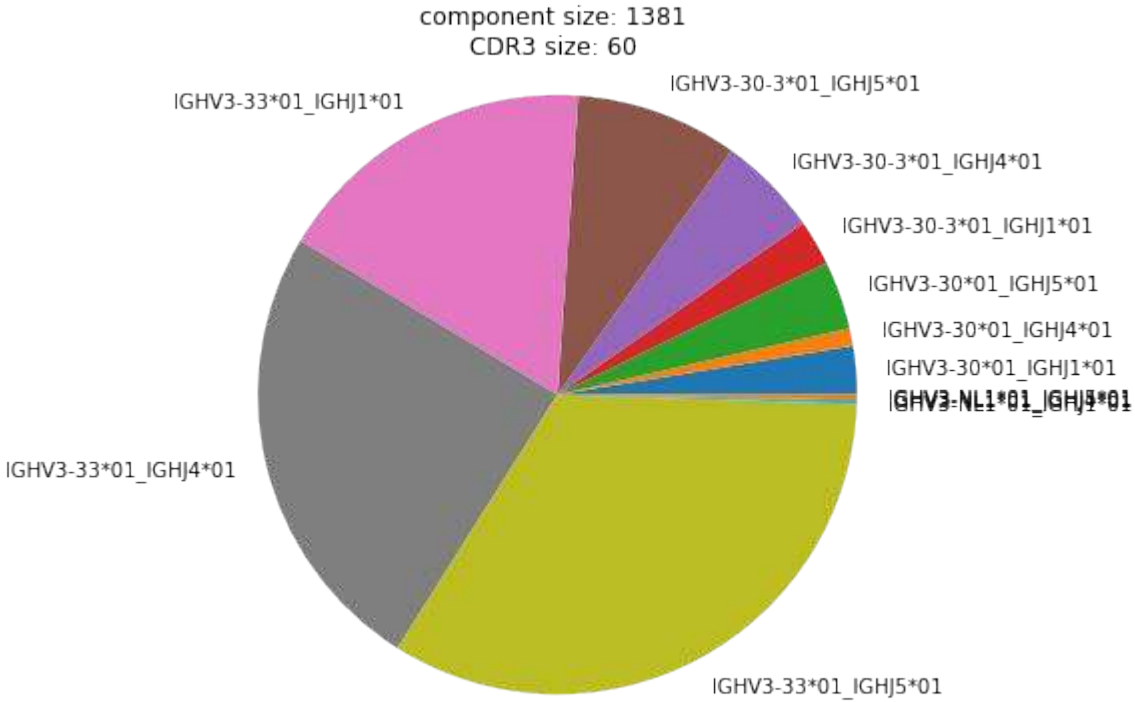
What we actually see: bad case

V genes:

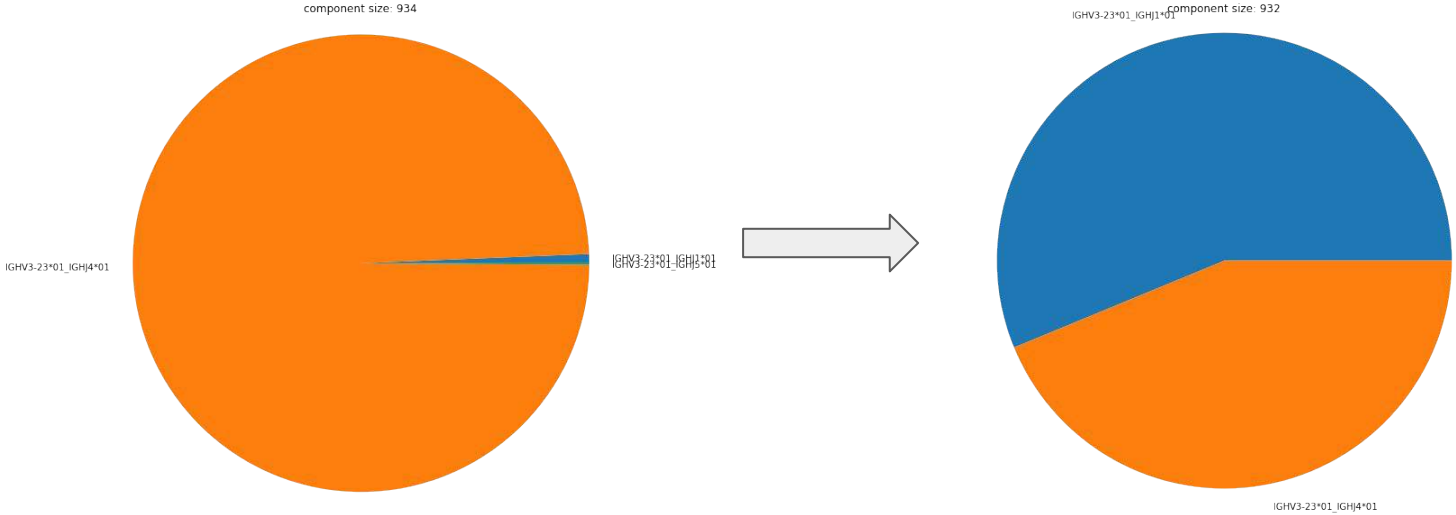
IGHV3-33
IGHV3-30
IGHV3-30-3
IGHV3-NL1

J genes:

IGHJ1
IGHJ4
IGHJ5



Problem with J genes alignment after cropping




```
IGHJ1*01      gctgaatacttccagcactgggccagggcaccctggtcaccgtctcctcag
IGHJ4*01      ----actactttgactactgggccaaggaaccctggtcaccgtctcctcag
```

J genes alignment

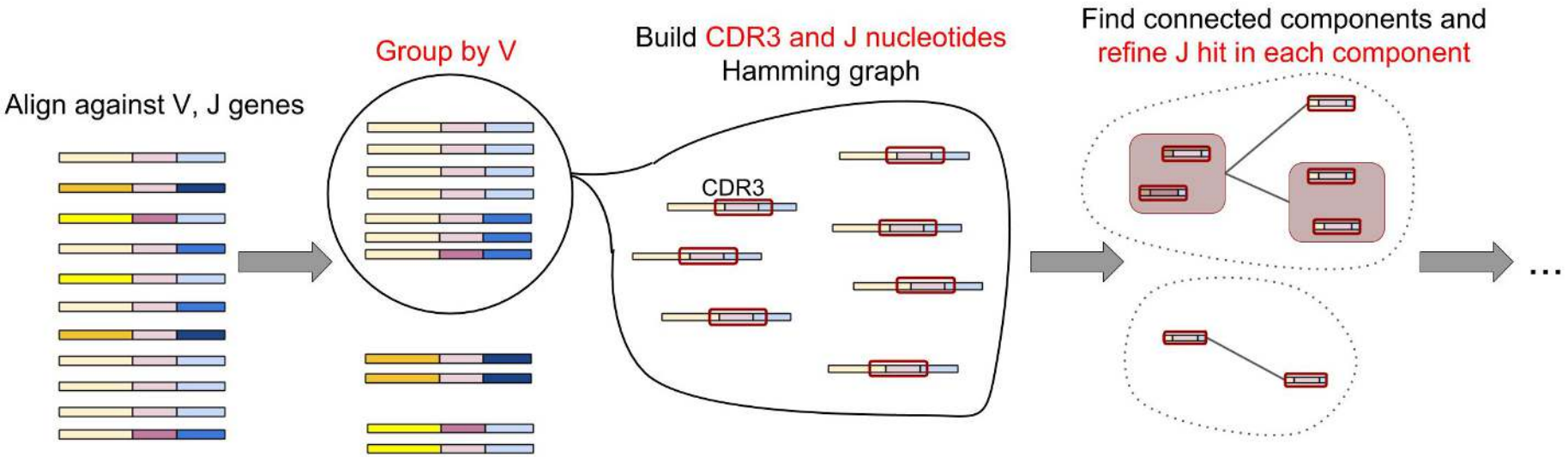
We should align J genes to the “universal” J gene

tgggg????gg?ac???ggtcac?gtctc?tcag

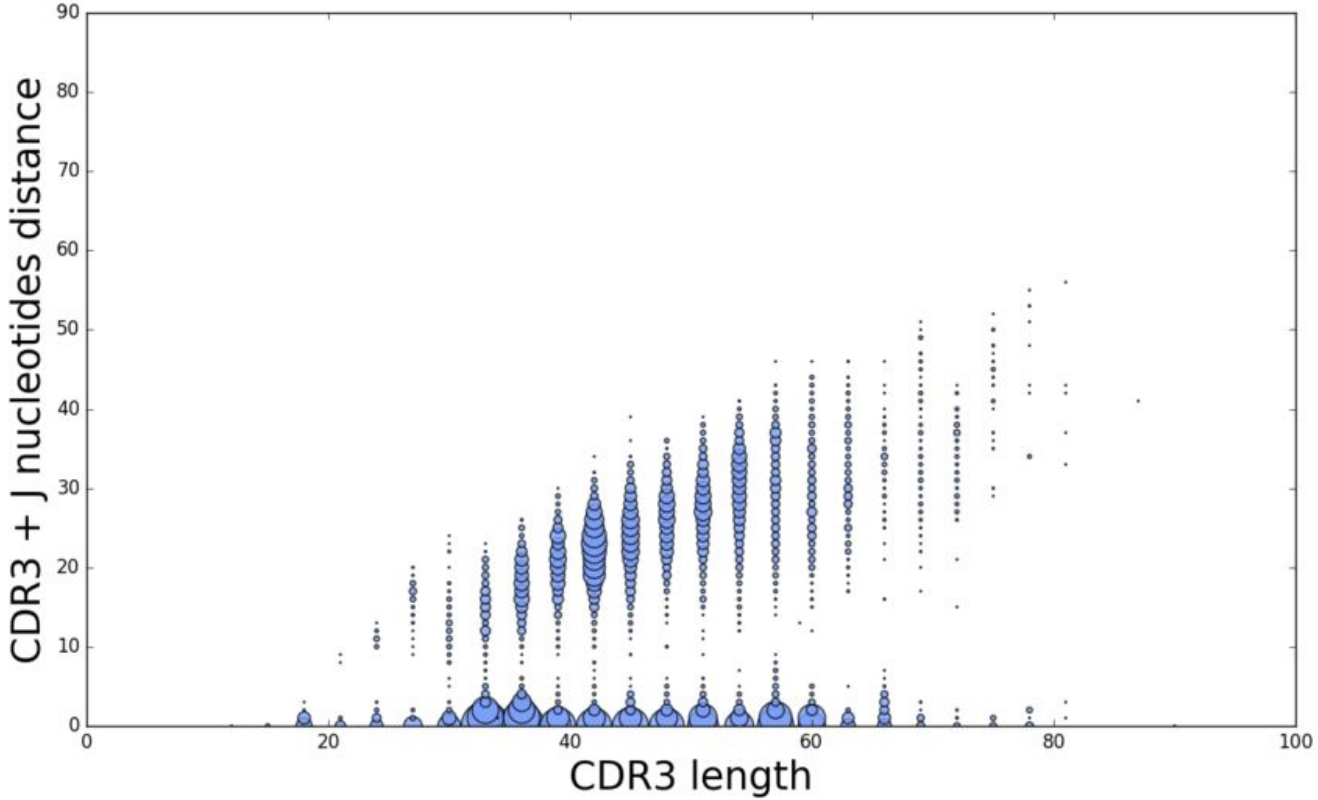
and take into account only ? nucleotides (the first 8)

Problems: indels, N

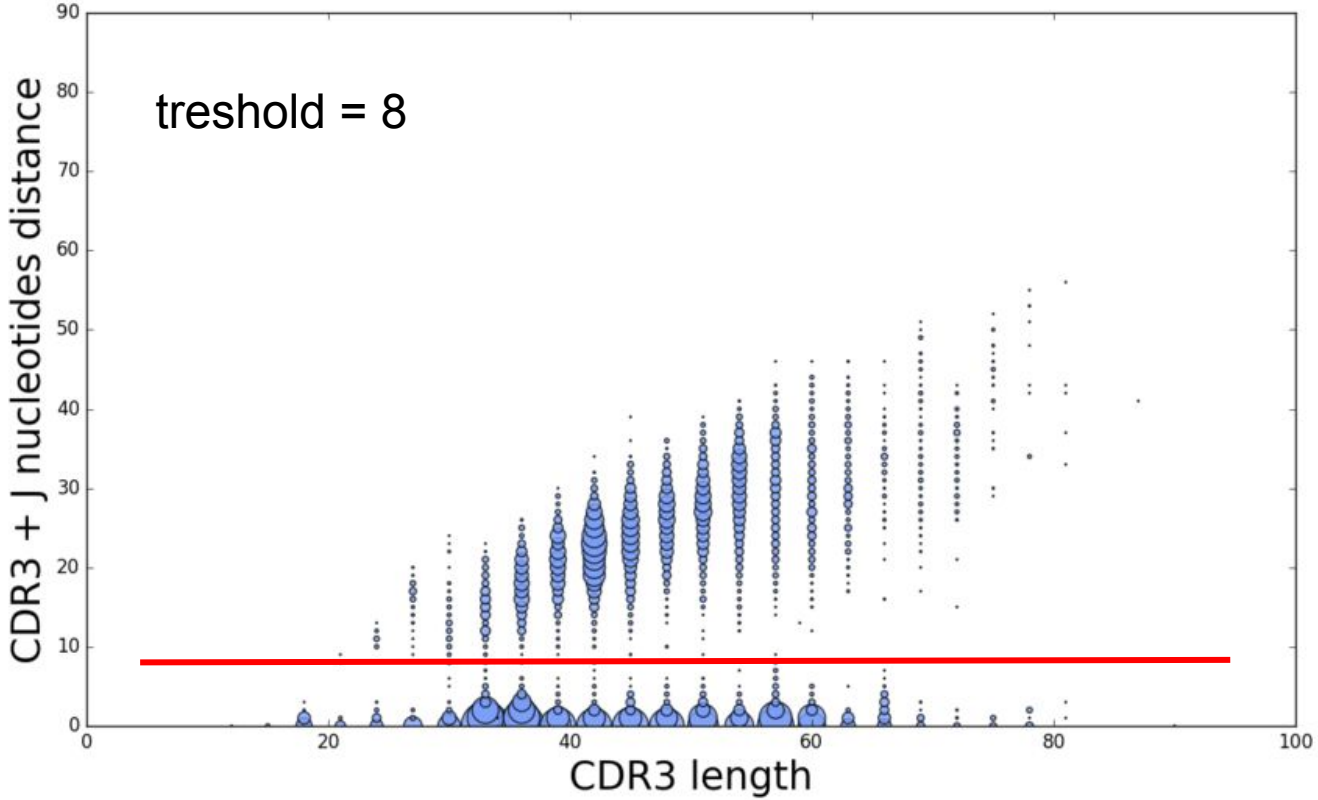
AntEvolo modification



CDR3 and J nucleotides Hamming Graph: treshold



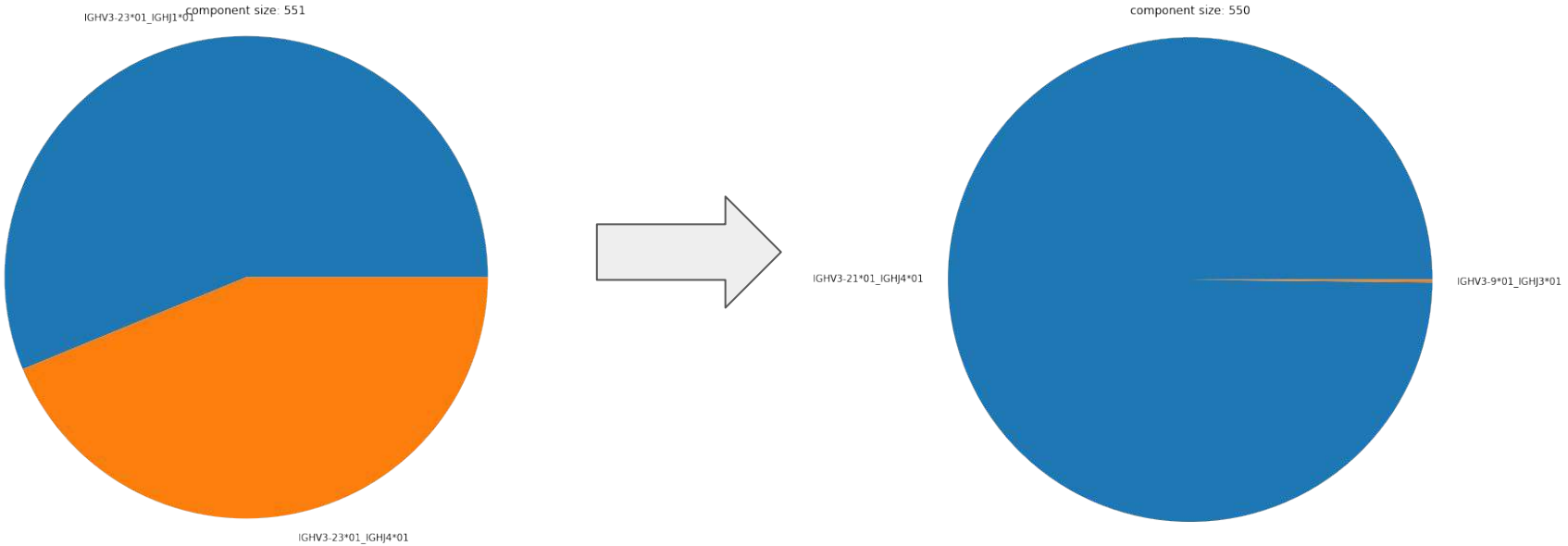
CDR3 and J nucleotides Hamming Graph: treshold



Results: FV dataset

	#cc: size > 100	#cc: 100%	#cc: singletone	#cc: trash
FV1: AntEvolo	115	65	29	1
FV1: modification	114	104	4	0
FV2: AntEvolo	150	77	39	1
FV2: modification	147	117	13	1
FV3: AntEvolo	239	141	40	8
FV3: modification	238	189	16	2
FV4: AntEvolo	352	222	63	8
FV4: modification	344	296	22	0

Results: a single lineage of lymphoma B cells



Questions?