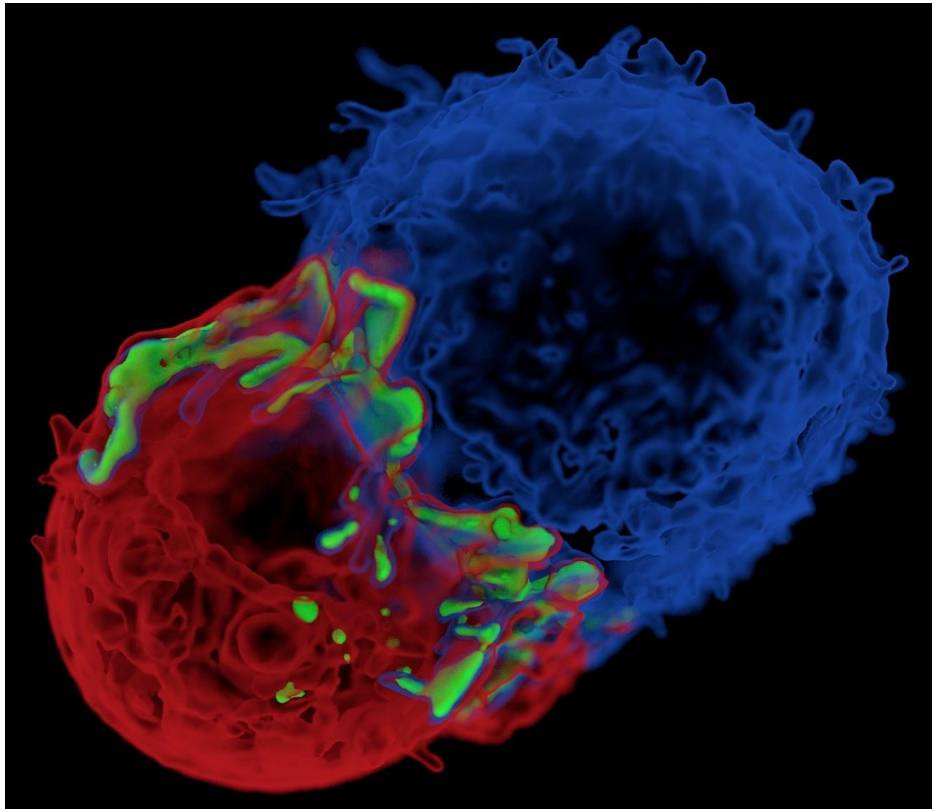
TOWARDS T-CELL RECEPTOR SPECIFICITY PREDICTION

MIKHAIL SHUGAY, SKOLTECH/IBCH RAS/RNRMU

BIOHACK 2 – 4 March 2018 | Saint-Petersbourg

INTRODUCTION: T-CELLS



From James and Vale, Nature 2012,
https://valelab.ucsf.edu/images/

TCR-expressing cell (red), antigen-presenting cell (blue) and activated TCR complex in green

T-cells form the cellular branch of the adaptive immune system. They can be classified into two large subpopulations:

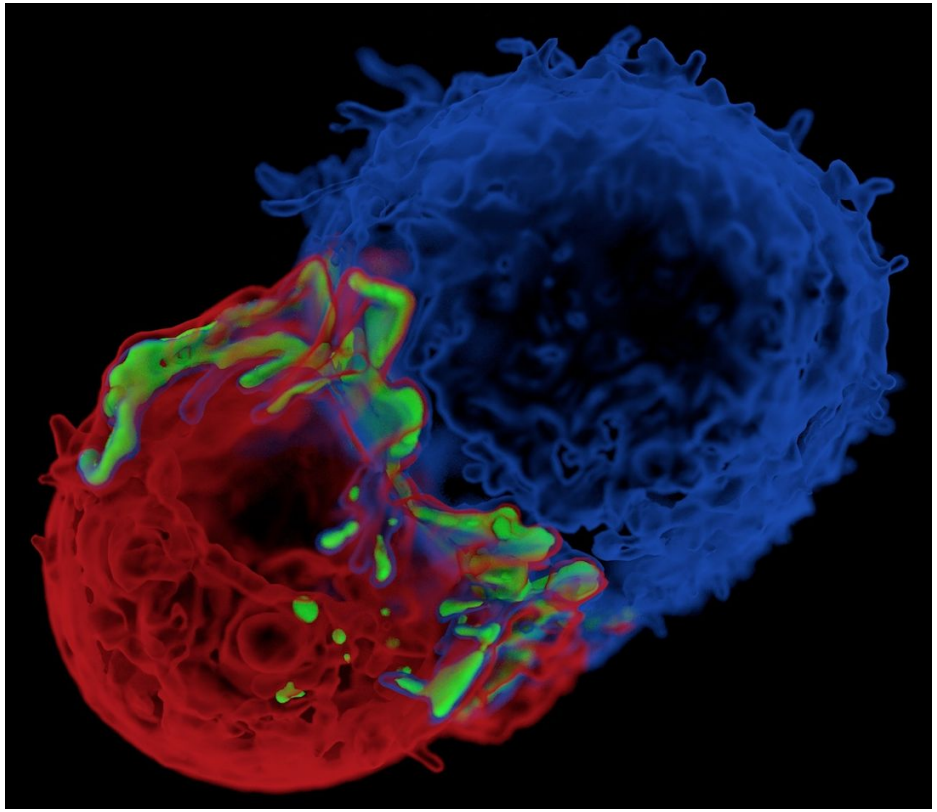
- CD8+ T-cells have the unique ability to recognize and kill infected and malignant cells
- CD4+ T-cells are able to interact with antigen presenting cells and guide immune response

In both cases these interactions are mediated by a molecule at the surface of a T-cell called the T-cell receptor (TCR).

TCR is a heterodimer build from two chains: TCRalpha and TCRbeta.

TCR recognizes antigenic peptides present by major histocompatibility complex (MHC) molecules (also called HLA in human).

INTRODUCTION: T-CELLS



From James and Vale, Nature 2012, https://valelab.ucsf.edu/images/

TCR-expressing cell (red), antigen-presenting cell (blue) and activated TCR complex in green

T-cells form the cellular branch of the adaptive immune system. They can be classified into two large subpopulations:

- CD8+ T-cells have the unique ability to recognize and kill infected and malignant cells
- CD4+ T-cells are able to interact with antigen presenting cells and guide immune response

In both cases these interactions are mediated by a molecule at the surface of a T-cell called the T-cell receptor (TCR).

TCR is a heterodimer build from two chains: TCRalpha and TCRbeta.

TCR recognizes antigenic peptides present by major histocompatibility complex (MHC) molecules (also called HLA in human).

INTRODUCTION: V(D)J REARRANGEMENT OF T-CELL RECEPTORS

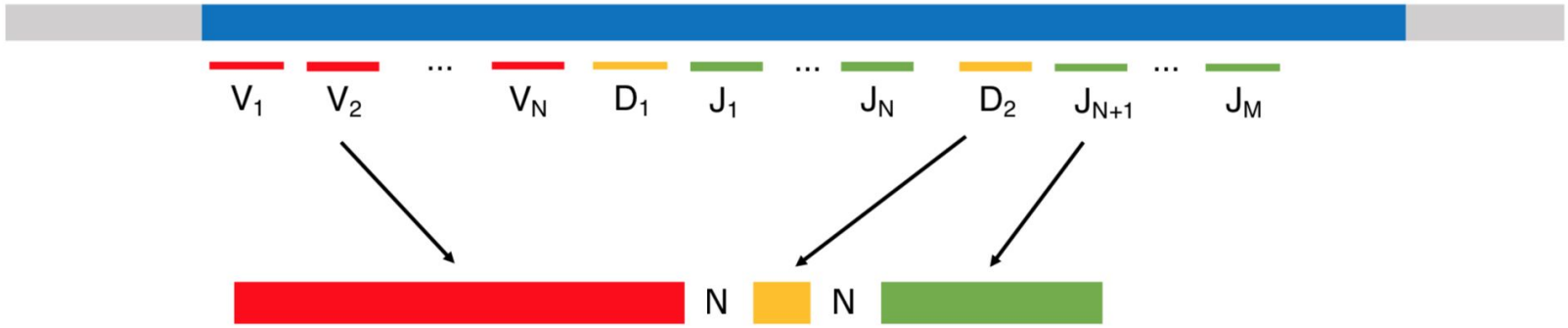
During T-cell development, each cell undergoes a process called V(D)J rearrangement producing TCRalpha and TCRbeta genes by joining genomic segments in corresponding loci.

This process generates $>10^{10}$ unique variants for a single TCR chain. Such immense diversity is required to cover the landscape of potential antigens presented by MHC molecules.

INTRODUCTION: V(D)J REARRANGEMENT OF T-CELL RECEPTORS

During T-cell development, each cell undergoes a process called V(D)J rearrangement producing TCRalpha and TCRbeta genes by joining genomic segments in corresponding loci.

This process generates >10^{10} unique variants for a single TCR chain. Such immense diversity is required to cover the landscape of potential antigens presented by MHC molecules.



V(D)J rearrangement
Variable, Diversity and Joining are chosen at random from the genome, V-D and D-J junctions are filled with non-template N nucleotides. The image is for TCRbeta, TCRalpha lacks D segments.

INTRODUCTION: T-CELL RECEPTOR STRUCTURAL DOMAINS

TCR chains consists of the following regions:

In total there are four framework (FRs) and three complementarity determining regions/loops (CDRs).

INTRODUCTION: T-CELL RECEPTOR STRUCTURAL DOMAINS

TCR chains consists of the following regions:

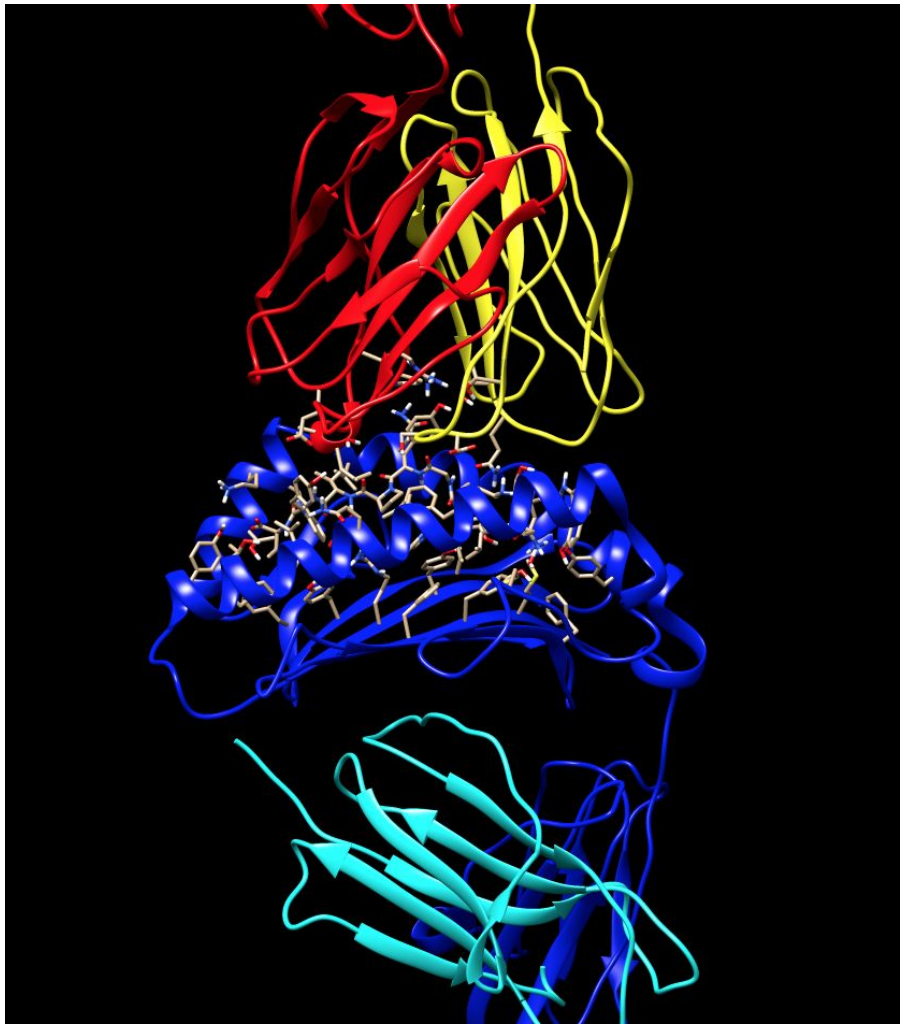
In total there are four framework (FRs) and three complementarity determining regions/loops (CDRs).

The structural functions of these regions are:

- FR regions maintain TCR secondary structure and (possibly) play role in MHC binding
- CDR1,2 are germline encoded and play role in antigen recognition, as well as (possibly) MHC binding
- CDR3 plays a major role in antigen recognition and is extremely variable

INTRODUCTION: TCR-peptide-MHC interaction

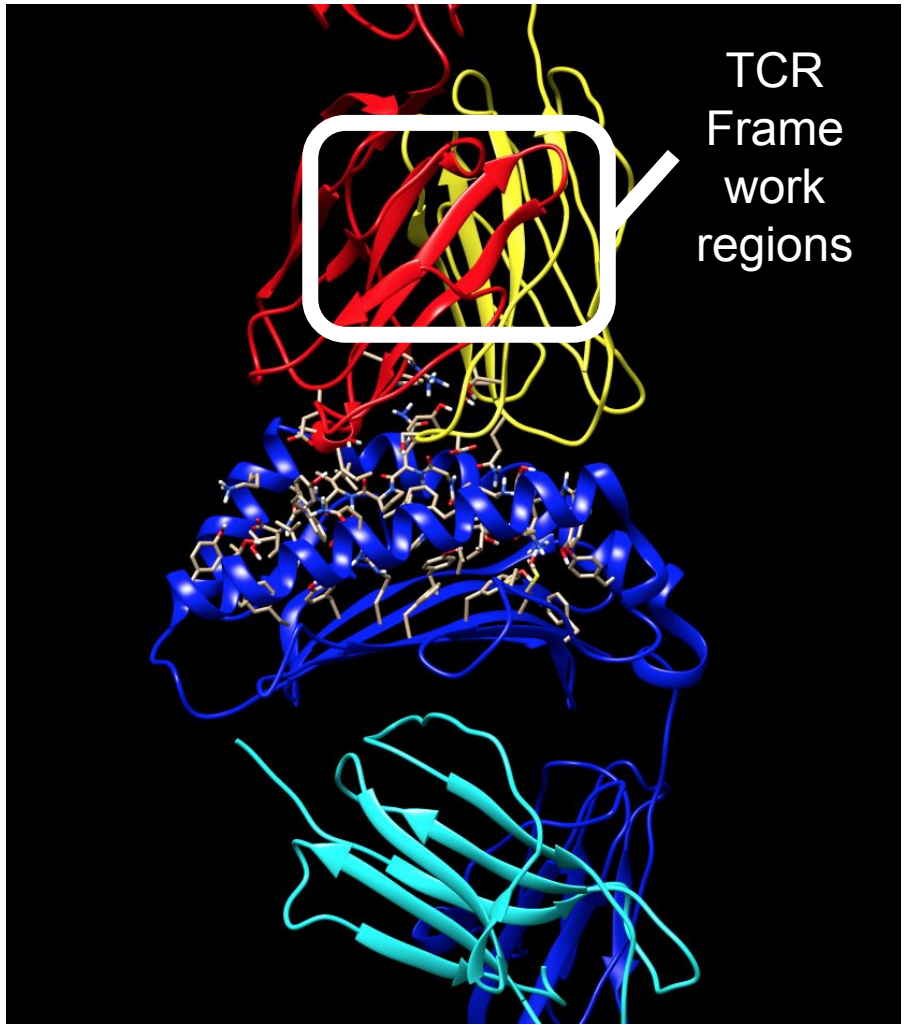
TCRalpha+beta



MHC class I

INTRODUCTION: TCR-peptide-MHC interaction

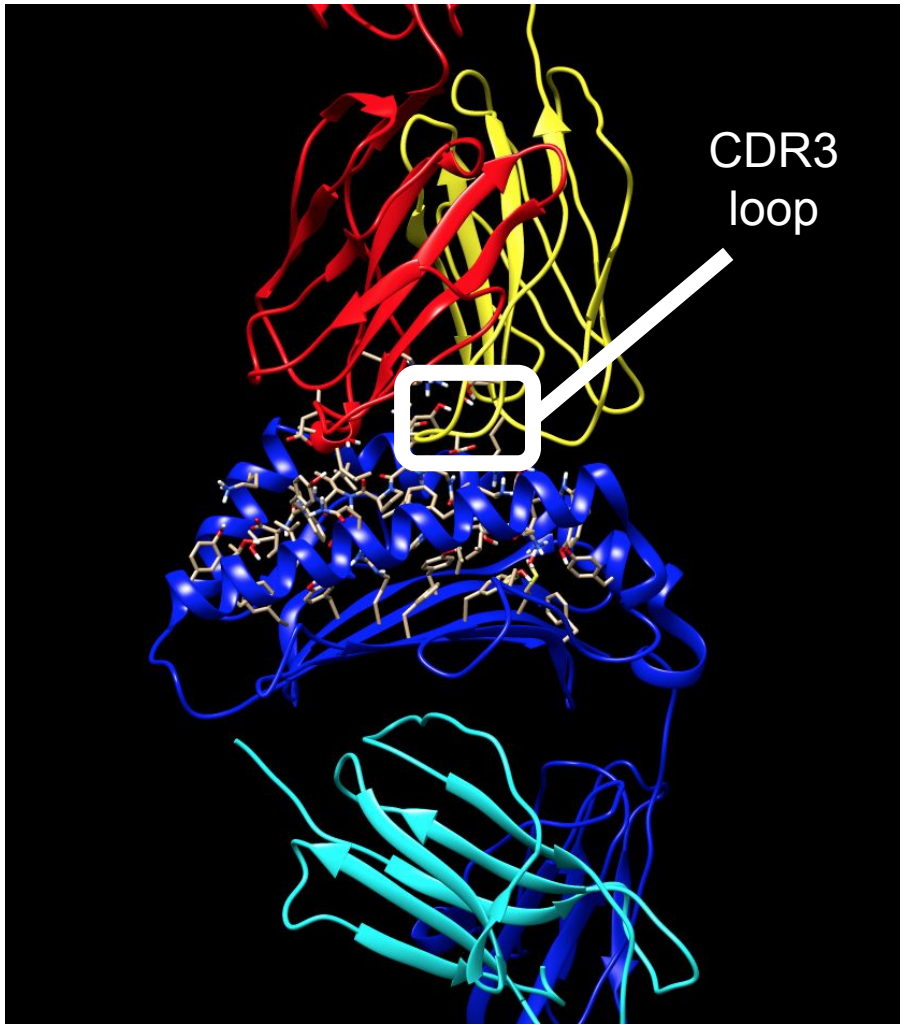
TCRalpha+beta



MHC class I

INTRODUCTION: TCR-peptide-MHC interaction

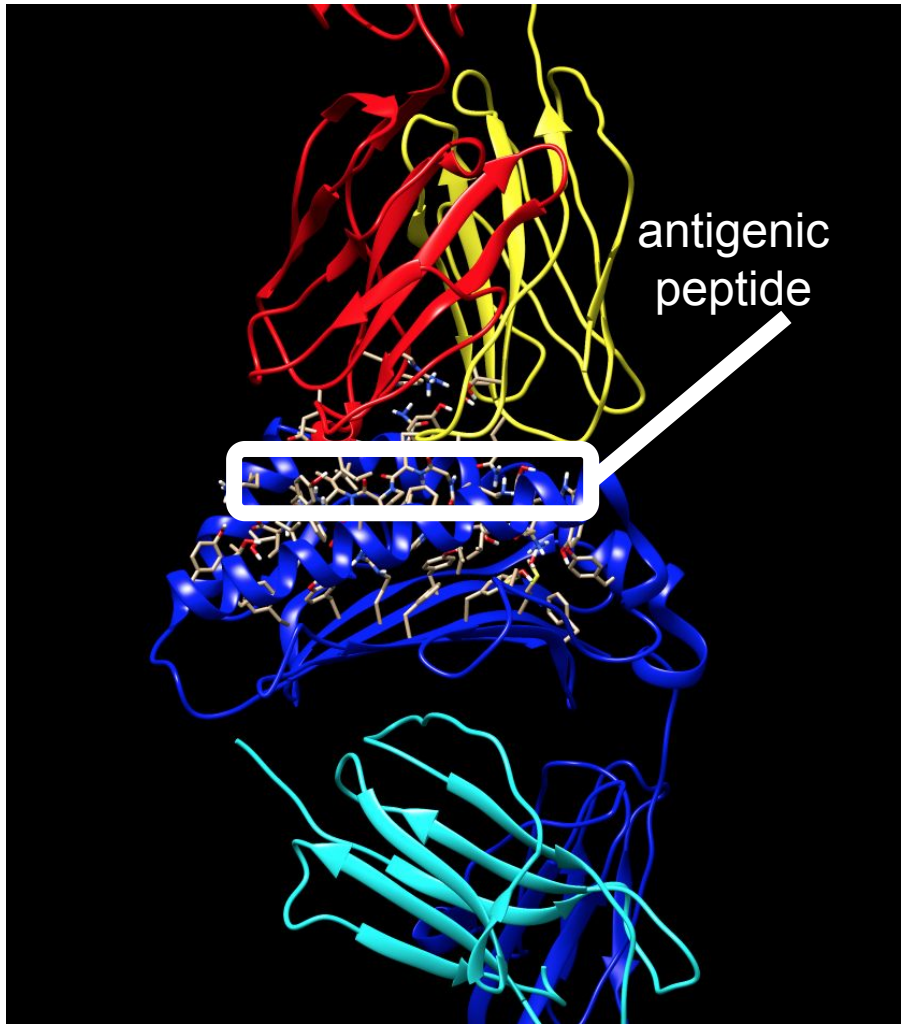
TCRalpha+beta



MHC class I

INTRODUCTION: TCR-peptide-MHC interaction

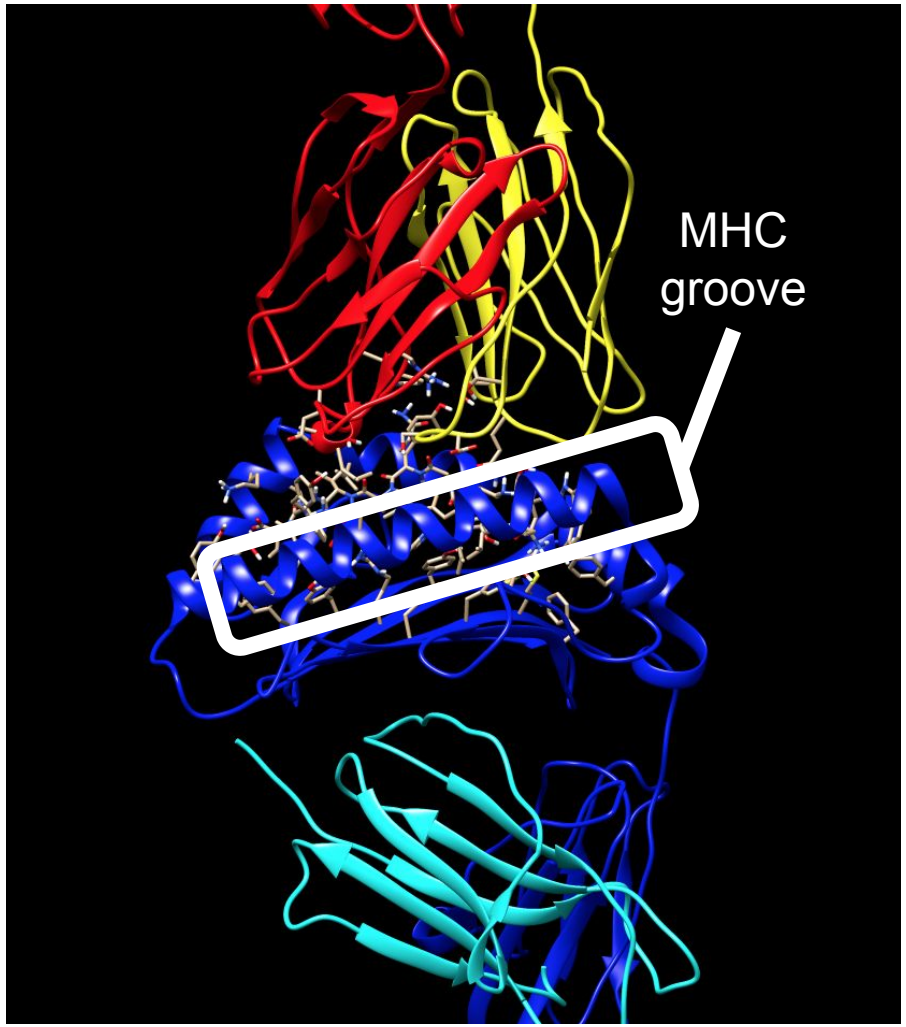
TCRalpha+beta



MHC class I

INTRODUCTION: TCR-peptide-MHC interaction

TCRalpha+beta



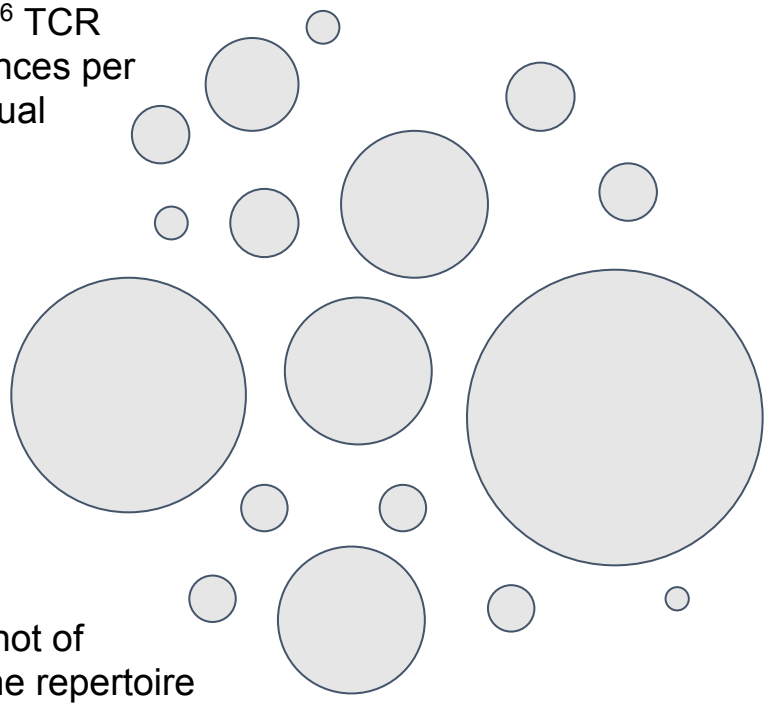
MHC class I

PROJECT SUMMARY: PROBLEM STATEMENT

With the advent of high-throughput sequencing methods (namely, immune repertoire sequencing or RepSeq) it is now possible to sequence millions of TCRs from a sample of interest.

While basic analysis of TCR sequence features can help to answer important biological questions, RepSeq data lacks one major component: annotation of antigen specificities.

10^5-10^6 TCR sequences per individual



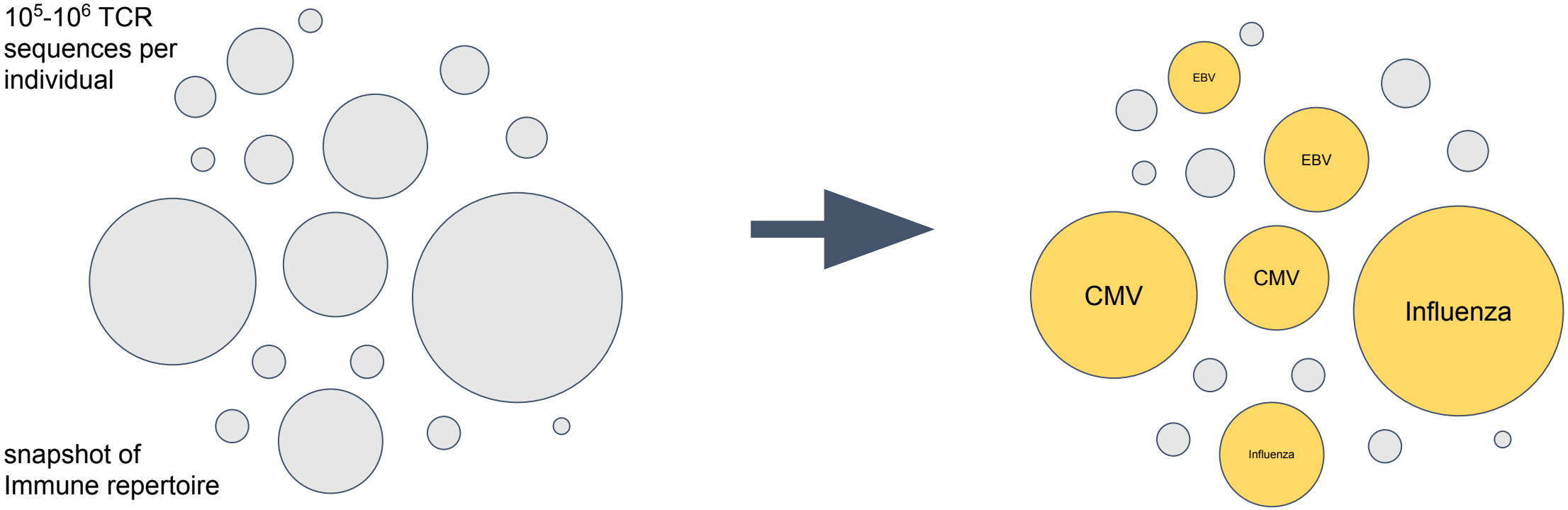
snapshot of Immune repertoire

PROJECT SUMMARY: PROBLEM STATEMENT

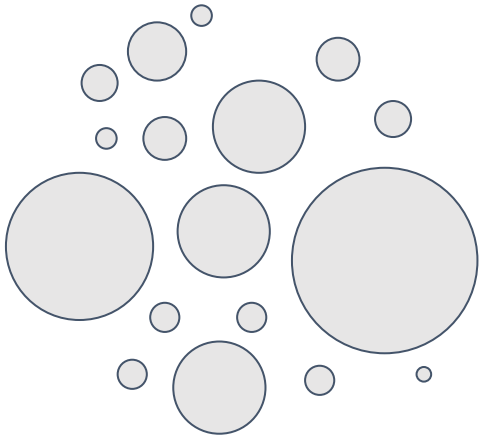
With the advent of high-throughput sequencing methods (namely, immune repertoire sequencing or RepSeq) it is now possible to sequence millions of TCRs from a sample of interest.

While basic analysis of TCR sequence features can help to answer important biological questions, RepSeq data lacks one major component: annotation of antigen specificities.

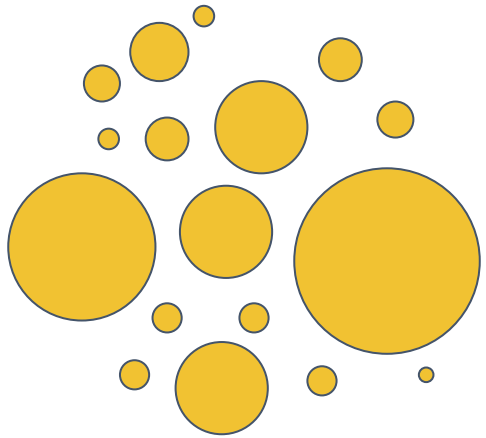
Development of the VDJdb database (**vdjdb.cdr3.net**) that compiles experimentally determined TCR sequence specificities is the first step in this direction.



PROJECT SUMMARY: PROBLEM STATEMENT



Index	Frequency	Count	CDR3AA	V	D	J	CDR3NT
1	1.0%	3913	CSAGGLGSTDTQYF	TRBV20-1	TRBD1	TRBJ2-3	TGCAGTGCTGGGGGGCTCGGTAGCACAGATACGCAGTATTTT
2	0.90%	3440	CASNSGSSYNEQFF	TRBV5-1	TRBD2	TRBJ2-1	TGCGCCAGCAATAGCGGGAGCTCCTACAATGAGCAGTTCTTC
3	0.79%	3021	CSARQGNQPQHF	TRBV20-1	TRBD1	TRBJ1-5	TGCAGTGCGCGACAGGGGAATCAGCCCCAGCATTTT
4	0.65%	2490	CASSQEPGGEQFF	TRBV4-1	TRBD2	TRBJ2-1	TGCGCCAGCAGCCAAGAGCCGGGCGGGGAGCAGTTCTTC
5	0.61%	2336	CASSYGMNTEAFF	TRBV6-6	TRBD2	TRBJ1-1	TGTGCCAGCAGTTACGGGATGAACACTGAAGCTTTCTTT



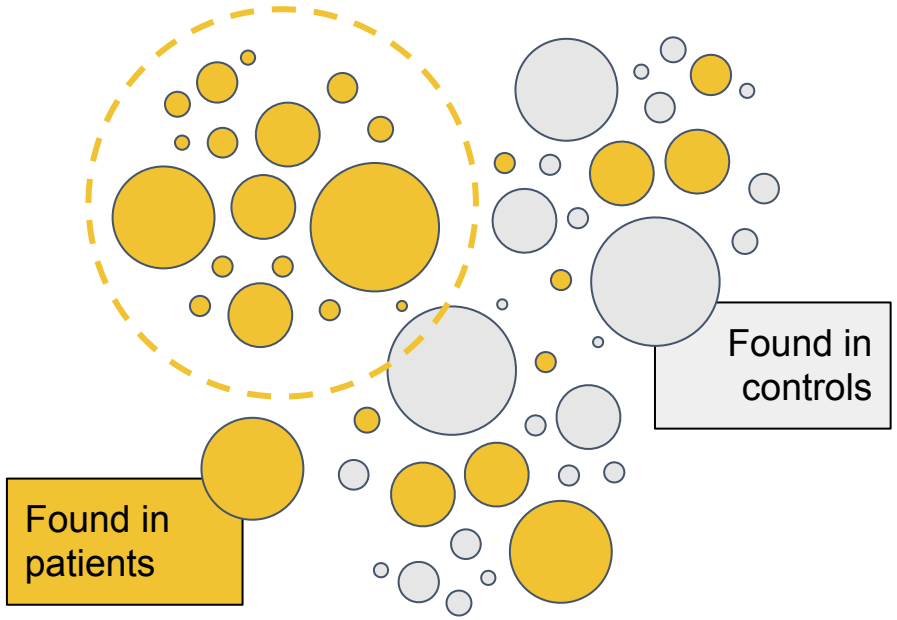
Details	# matches	Rank	Frequency	Count	CDR3aa	V	J	Tags
▾	1	3	4.63e-2	10832	CASSRGPGGTDTQYF	TRBV12-3	TRBJ2-3	GILGFVFTL, InfluenzaA, HLA-A*02, B2M
▾	2	4	3.65e-2	8532	CASSPSGGTDTQYF	TRBV18	TRBJ2-3	NLVPMVATV, CMV, HLA-A*02, B2M
▾	7	5	3.48e-2	8153	CASSLIGVSSYNEQFF	TRBV7-9	TRBJ2-1	TPRVTGGGAM, LPRRSGAAGA, NLVPMVATV, CMV, InfluenzaA, HLA-B*07:02, HLA-B*07, HLA-A*02, B2M

Alignment	Gene	CDR3	V	J	Species	MHC A	MHC B	MHC class	Epitope	Epitope gene	Epitope species	Reference	Method	Meta	CDR3fix	Score
CASSLIGVSSYNEQFF ||||||||||||||| CASSLIGVSSYNEQFF	TRB	CASSLIGVSSYNEQFF	TRBV7-9*01	TRBJ2-1*01	HomoSapiens	HLA-B*07:02	B2M	MHCI	TPRVTGGGAM	p65	CMV	PMID:23267020	ℹ	ℹ	ℹ	1
CASSLIGVSSYNEQFF |||||||||||||||	TRB	CASSLIGVSSYNEQFF	TRBV7-9*01	TRBJ2-1*01	HomoSapiens	HLA-B*07:02	B2M	MHCI	TPRVTGGGAM	p65	CMV	PMID:28636589	ℹ	ℹ	ℹ	1

PROJECT SUMMARY: PROBLEM STATEMENT

The main idea of this project is to develop algorithms for functional annotation of TCR repertoires that can solve two major tasks:

- Finding groups of homologous TCR sequences that can potentially recognize the same antigen. This task is critical for biomarker discovery in studies involving donors with cancer/infection/autoimmune disease and healthy controls

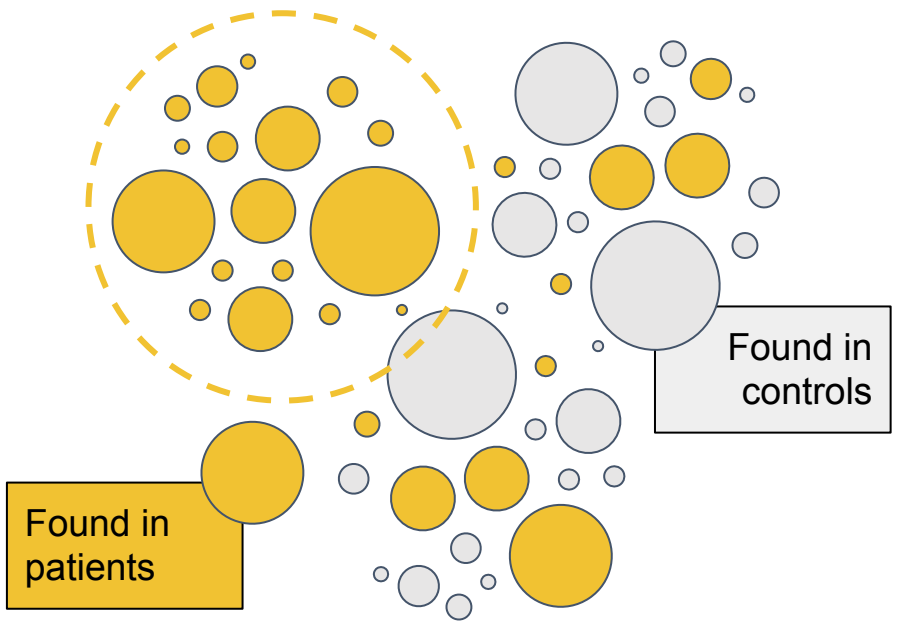


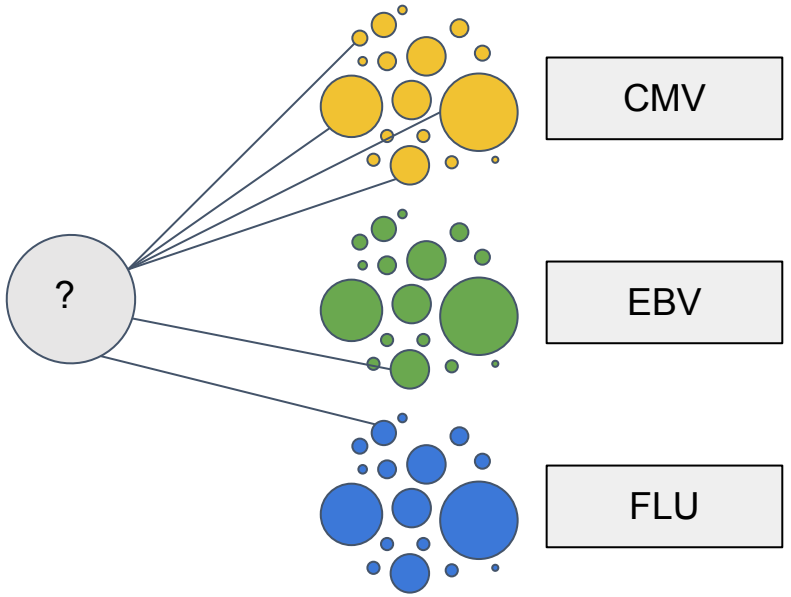
PROJECT SUMMARY: PROBLEM STATEMENT

The main idea of this project is to develop algorithms for functional annotation of TCR repertoires that can solve two major tasks:

- Finding groups of homologous TCR sequences that can potentially recognize the same antigen. This task is critical for biomarker discovery in studies involving donors with cancer/infection/autoimmune disease and healthy controls

- Querying TCR repertoire against a database of TCR sequences with known antigen specificity to predict specificities encoded in the repertoire.





PROJECT SUMMARY: POTENTIAL APPROACHES

We can highlight two reasonable approaches to this problem:

Alignment-based approach (see Dash et al. Nature 2017). In this case a distance metric between two TCR sequences is computed using sequence alignment:

- Custom alignment scoring matrices can be used, scores for CDR1-3 regions that are directly contacting the antigen are given additional weight in the aggregate score.
- Distance metric can be used for unsupervised clustering of TCR sequences, aggregate distance score to the set of antigen-specific TCRs in the database is used for annotation

Motif-based approach (see Glanville et al. Nature 2017). In this case motifs (k-mers, amino acid words of length k) are used for TCR sequence clustering

- TCR similarity is computed based on distance between k-mer profiles. K-mers can be clustered based on amino acid features/weighted based on their chance to randomly occur in the repertoire. Note that k-mers also encode position in TCR sequence.
- TCR annotation is performed by computing the similarity of TCR k-mer profile to the aggregate k-mer profile of the set of antigen-specific TCRs.

We suggest trying out the second approach.

DATASETS

All datasets and their description can be found at **https://github.com/antigenomics/biohack-2018**

The repository contains

- A sample from VDJdb database (Human TRB, epitopes that have at least 30 associated TCR sequences)
- Memory and Naive T-cell TCR beta repertoires of CMV+ and CMV- individuals.
- Miscellaneous information: substitution scoring matrices for CDR3 region called VDJAM, Variable segment sequences, k-mer similarity scores

DATASETS: VDJDB database layout

TCR sequence is recorded as CDR3 region amino acid sequence, Variable (v.segm) and Joining (j.segm) segment identifiers, v.end and j.start mark positions of germline subregions of CDR3.

Antigen and MHC information is stored in ‘antigen.*’ and ‘mhc.*’ fields. The ‘antigen.epitope’ field should be used as a tag/class id in all classification tasks!

```
> str(dt.vdjdb.good)
'data.frame':   13357 obs. of  11 variables:
 $ cdr3          : chr  "CASSFSGTQYF" "CASSLGTSGGNEQFF" "CASSQVEEFHGELFF" "CASRTQRWETQYF" ...
 $ antigen.epitope: chr  "NLVPMVATV" "NLVPMVATV" "GILGFVFTL" "QASQEVKNW" ...
 $ antigen.gene  : chr  "p65" "p65" "M1" "p24" ...
 $ antigen.species: chr  "CMV" "CMV" "InfluenzaA" "HIV-1" ...
 $ v.segm        : chr  "TRBV28*01" "TRBV7-6*01" "TRBV7-9*01" "TRBV27*01" ...
 $ j.segm        : chr  "TRBJ2-3*01" "TRBJ2-1*01" "TRBJ2-2*01" "TRBJ2-5*01" ...
 $ v.end         : int  4 5 4 3 4 4 4 4 4 5 ...
 $ j.start       : int  7 10 10 8 10 12 10 11 13 9 ...
 $ mhc.a         : chr  "HLA-A*02" "HLA-A*02" "HLA-A*02" "HLA-B*57" ...
 $ mhc.b         : chr  "B2M" "B2M" "B2M" "B2M" ...
 $ mhc.class     : chr  "MHCI" "MHCI" "MHCI" "MHCI" ...
```

DATASETS: RepSeq sample layout

Similar to VDJdb, yet different column naming (‘cdr3’ is renamed to ‘cdr3aa’) and V/J ids have ‘*01’ allele identifier trimmed. The quantity of TCR clonotype is reflected by the ‘count’ column, which contains the number of sequencing reads.

```
> str(dt.sample)
Classes ‘data.table’ and 'data.frame':  10136 obs. of  11 variables:
 $ count : int  16460 13951 13473 10832 8532 8153 4442 3791 3241 2355 ...
 $ freq  : num  0.0703 0.0596 0.0576 0.0463 0.0365 ...
 $ cdr3nt: chr  "TGTGCCAGCAGCTTAGTTTTTGGTAGCGGGGGCGCCTACAATGAGCAGTTCTTC" "TGCGC
GCAGGGGGTGGGGAATGAGCAGTTCTTC" "TGTGCCAGCAGCAGGGGACCAGGGGGCACGGATACGCAGTATTTT" .
 $ cdr3aa: chr  "CASSLVFGSGGAYNEQFF" "CASSGDGMNTEAFF" "CATSDLGQGVGNEQFF" "CASSR
 $ v     : chr  "TRBV11-2" "TRBV10-2" "TRBV24-1" "TRBV12-3" ...
 $ d     : chr  "TRBD2" "TRBD1" "TRBD1" "TRBD1" ...
 $ j     : chr  "TRBJ2-1" "TRBJ1-1" "TRBJ2-1" "TRBJ2-3" ...
 $ VEnd  : int  16 13 18 11 15 15 11 14 12 12 ...
 $ DStart: int  27 13 21 11 -1 17 17 13 14 20 ...
 $ DEnd  : int  32 18 28 17 -1 22 22 19 22 25 ...
 $ JStart: int  34 22 33 25 22 26 21 23 34 32 ...
```

TASKS SUMMARY & SUGGESTIONS

1. Develop a TCR sequence similarity metric that gives higher similarity score to TCR pairs specific to the same antigen than to pairs specific to different antigens.

Similarity metric can be based on V/CDR3 alignment scores or distance between K-mer profile vectors. Optimize the metric by training it on the set of TCR pairs matching the same and different antigens from VDJdb

TASKS SUMMARY & SUGGESTIONS

1. Develop a TCR sequence similarity metric that gives higher similarity score to TCR pairs specific to the same antigen than to pairs specific to different antigens.

Similarity metric can be based on V/CDR3 alignment scores or distance between K-mer profile vectors. Optimize the metric by training it on the set of TCR pairs matching the same and different antigens from VDJdb

2. Choose optimal TCR repertoire unsupervised clustering method. Given the TCR similarities the method should group VDJdb records into clusters in a way that TCRs specific to the same antigen are placed in the same cluster.

Select clustering method that relies on pairwise distances, such as hierarchical clustering or DBSCAN

TASKS SUMMARY & SUGGESTIONS

1. Develop a TCR sequence similarity metric that gives higher similarity score to TCR pairs specific to the same antigen than to pairs specific to different antigens.

Similarity metric can be based on V/CDR3 alignment scores or distance between K-mer profile vectors. Optimize the metric by training it on the set of TCR pairs matching the same and different antigens from VDJdb

2. Choose optimal TCR repertoire unsupervised clustering method. Given the TCR similarities the method should group VDJdb records into clusters in a way that TCRs specific to the same antigen are placed in the same cluster.

Select clustering method that relies on pairwise distances, such as hierarchical clustering or DBSCAN

3. Develop a scoring rule that aggregates similarity scores obtained by matching a given TCR against a set of TCRs specific to a given epitope. This method should be assessed by splitting VDJdb in test and training set. TCRs should have higher aggregate similarity score to the set specific to their cognate antigen.

VDJdb records can be weighted by their “informativeness”, i.e. the probability of random match between a given record and other records specific to a different antigen