T-cell repertoire annotation and motif discovery

A RepSeq data analysis tutorial in R

Mikhail Shugay, PhD

Skolkovo Institute of Science and Technology

29 October 2018, 2nd SCIAR meeting

Outline

Aims of this tutorial

Aim1 Learn how to infer T-cells specific to certain epitopes and extract T-cell receptor (TCR) sequence motifs from high-throughput sequencing data (RepSeq).

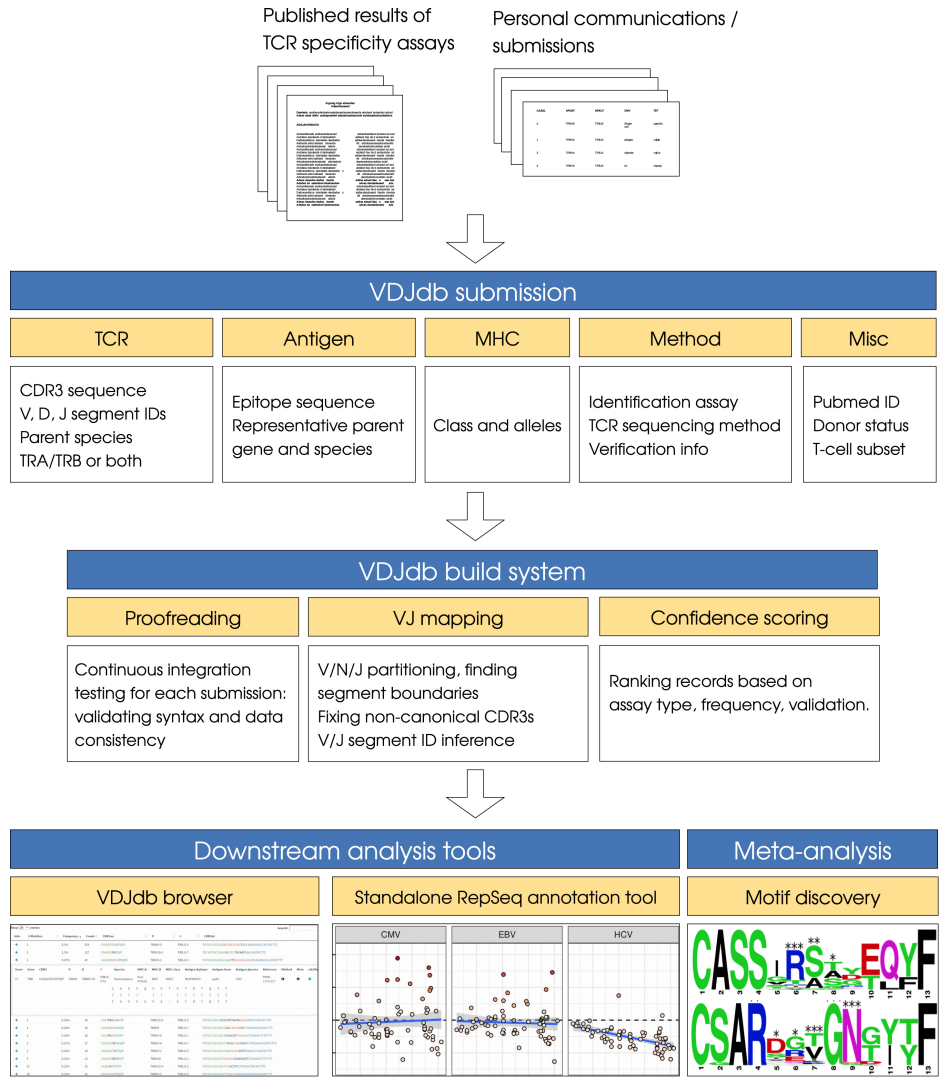
Aim2 Get familiar with `VDJdb` database, `VDJtools` software and some useful R templates for TCR sequence analysis.

Disclaimer This tutorial will not cover RepSeq data processing and basic[1] analysis.

[1] Repertoire diversity analysis, segment usage, etc. See **Examples** section of `VDJtools` docs for this.

VDJdb database

VDJdb is a curated database of T-cell receptor sequences of known antigen specificity that can be accessed at `http://vdjdb.cdr3.net`



Clonotype tables

We define a clonotype as a unique combination of Variable (V) and Joining (J) segment and CDR3nt sequence observed in our sequencing data.

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
6	0.52%	1992	CASSQGGRAPHTQYF	TRBV4-3	TRBD2	TRBJ2-3	TGCGCCAGCAGCCAAGGGGGGAGGGCCCCCCATACGCAGTATTTT
7	0.49%	1871	CASSQSQGGSYEQYF	TRBV5-1	TRBD1	TRBJ2-7	TGCGCCAGCAGCCAAAGTCAAGGGGGGTCCTACGAGCAGTACTTC
8	0.48%	1847	CASSRPKSGRSGELFF	TRBV11-2	TRBD2	TRBJ2-2	TGTGCCAGCAGCCGACCCAAGAGCGGGAGAAGTGGGGAGCTGTTTTTT

TCR motif inference

Motif inference in present tutorial is based on the TCR neighbor enrichment test (`TCRNET`) implemented in `VDJtools`.

`TCRNET` scans TCR sequence graph for nodes having a degree higher than expected by chance.

TCRNET

Let n_i^s be the number of clonotypes in sample s that differ from i^{th} clonotype by no more than $d = 1$ substitutions in the CDR3aa sequence.[2]

Let N_i^s be the total number of clonotypes having the same V/J segments as i^{th} clonotype in sample s.

Select clonotypes with more neighbors than expected by chance by assuming that

$$n_i^s \overset{H_0}{\sim} Poisson\left(N_i^s \frac{n_i^c}{N_i^c}\right) \tag{1}$$

where c is the control sample.

[2] They don't have to have the same V/J segments

Outline

Dataset

We'll use RepSeq data from Emerson et al. Nat Genet 2017

Sample	**ID**	**a1**	**a2**	**b1**	**b2**	**status**
B35+	HIP02877	A*26	A*33	B*14	B*35	CMV-
CMV+	HIP13994	A*02	A*02	B*07	B*44	CMV+
Control-1	HIP03484	A*02	A*02	B*07	B*58	CMV-
Control-2	HIP03592	A*02	A*32	B*07	B*39	CMV-
Control-3	HIP04532	A*02	A*24	B*07	B*51	CMV-
Control-4	HIP04576	A*02	A*30	B*07	B*18	CMV-

Experiment - 1

Comparing B35+ sample versus samples without this allele

Sample	**ID**	**a1**	**a2**	**b1**	**b2**	**status**
B35+	HIP02877	A*26	A*33	B*14	B*35	CMV-
CMV+	HIP13994	A*02	A*02	B*07	B*44	CMV+
Control-1	HIP03484	A*02	A*02	B*07	B*58	CMV-
Control-2	HIP03592	A*02	A*32	B*07	B*39	CMV-
Control-3	HIP04532	A*02	A*24	B*07	B*51	CMV-
Control-4	HIP04576	A*02	A*30	B*07	B*18	CMV-

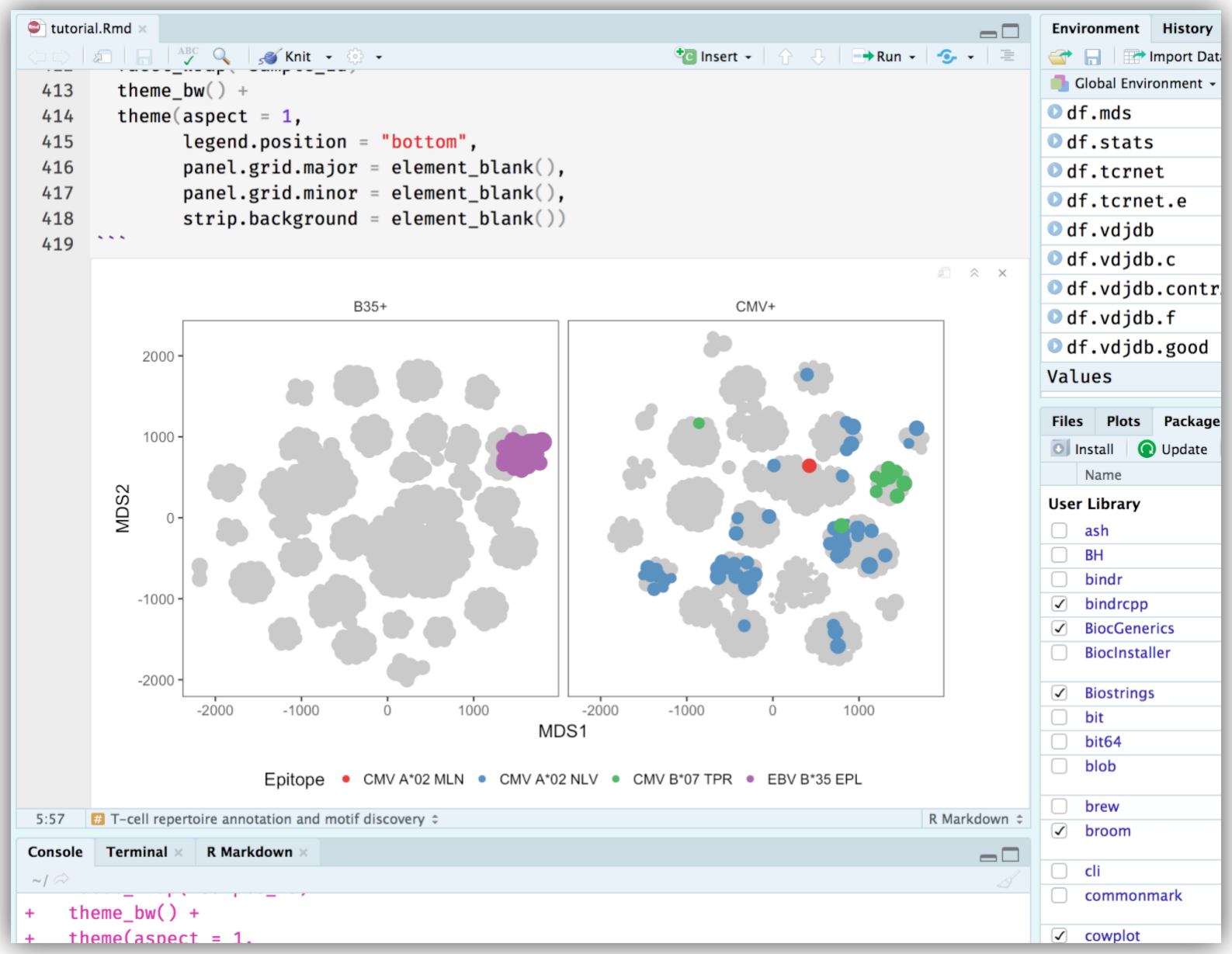
Experiment - 2

Comparing CMV+ and CMV- samples for A*02 and B*07

Sample	**ID**	**a1**	**a2**	**b1**	**b2**	**status**
B35+	HIP02877	A*26	A*33	B*14	B*35	CMV-
CMV+	HIP13994	A*02	A*02	B*07	B*44	CMV+
Control-1	HIP03484	A*02	A*02	B*07	B*58	CMV-
Control-2	HIP03592	A*02	A*32	B*07	B*39	CMV-
Control-3	HIP04532	A*02	A*24	B*07	B*51	CMV-
Control-4	HIP04576	A*02	A*30	B*07	B*18	CMV-

Outline

Interactive part



Outline

Overview

- Annotated our samples with `VDJdb` and quality-filtered results
- Filtered annotation results based on our allele of interest (HLA-B35) or donor status (CMV+)
- Inferred antigen-specific clonotype groups using `TCRNET` algorithm in `VDJtools`
- Overlaped `VDJdb` annotations and `TCRNET` results and extract CDR3 sequence motifs

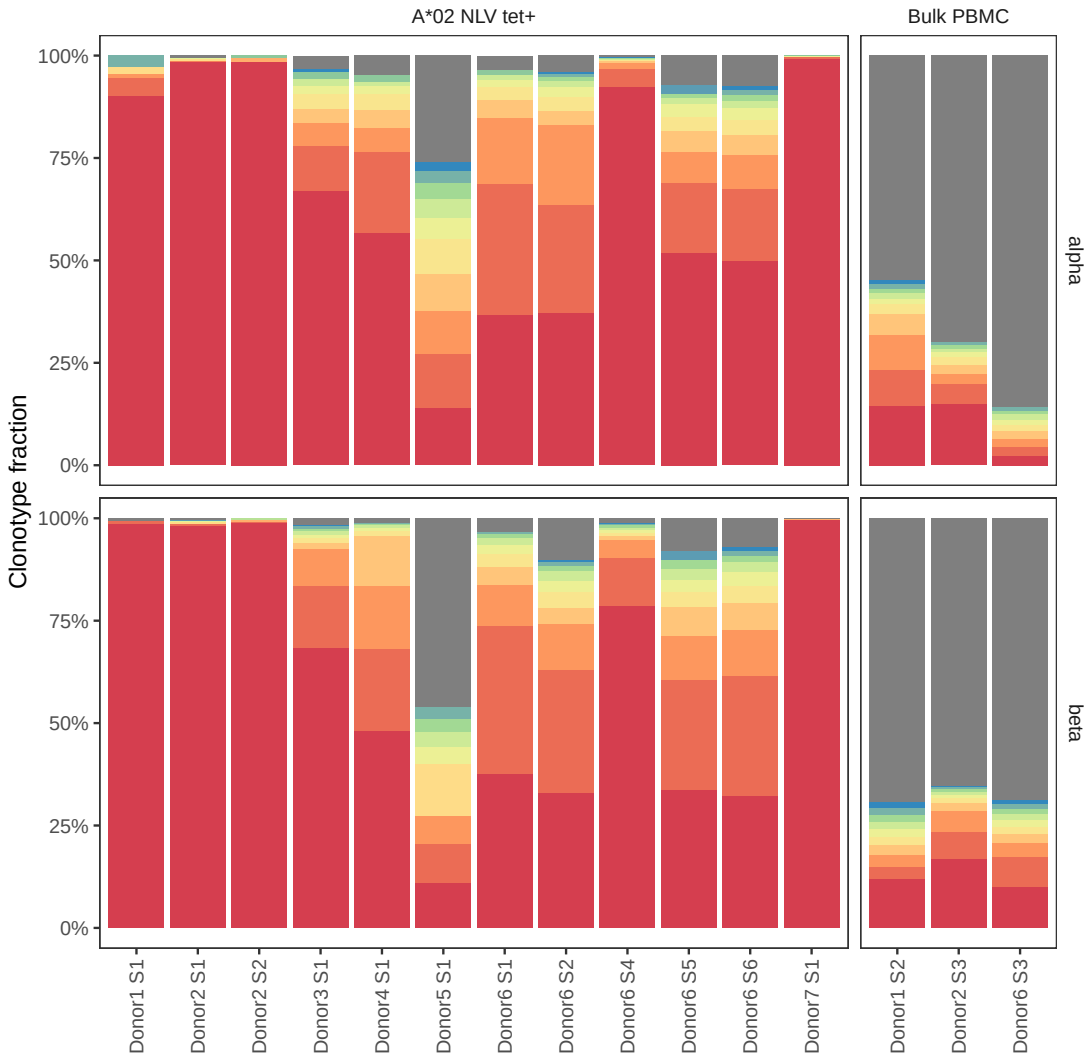
Potential pitfalls

- Repertoires are extremely diverse making TCR annotation an imbalanced classification problem. Even when matching against `VDJdb` with high specificity[3] one will observe many false-positives.
  Use proper controls, e.g. naive T-cells
- `TCRNET` will fail in some cases simply because the repertoire of T-cells specific to a given antigen is dominated by a single hyperexpanded clonotype.
  Always annotate and check large clonotypes

[3] According to VDJdb benchmark odds of matching the same epitope given 1 substitution in CDR3β are around 86 to 1

Note on CMV-specific clonotypes

Diverse repertoire of dissimilar TCRs, individuals are likely to carry a single hyperexpanded clonotype with no subvariants



Thank you for listening!

Contacts


antigenomics


mikessh, antigenomics


mikhail.shugay@gmail.com