

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 SCAIR meeting

Outline

Introduction

Setting

Interactive part

Concluding remarks

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).

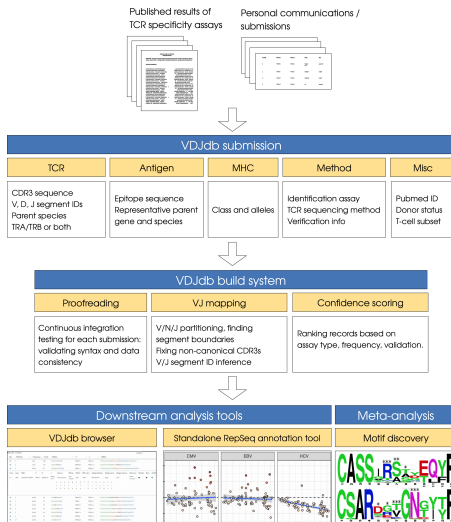
Aim2 Get familiar with VDJdb database, VDJtools software and some useful R templates for TCR sequence analysis.

Disclaimer This tutorial will not cover RepSeq data processing and basic¹ analysis.

¹Repertoire 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	CSAGGLGSTDQYF	TRBV20-1	TRBD1	TRBJ2-3	TGCAGTGTGGGGGGCTCGGTAGCACAGATACGCAGTATTTT
2	0.90%	3440	CASNSGSSYNEQFF	TRBV5-1	TRBD2	TRBJ2-1	TGCGCCAGCAATAGCGGGAGCTCTACAATGAGCAGTCTTC
3	0.79%	3021	CSARQGNQPQH	TRBV20-1	TRBD1	TRBJ1-5	TGCAGTGCGCACAGGGGAATCAGCCCCAGCATTTT
4	0.65%	2490	CASSQEPGGEQFF	TRBV4-1	TRBD2	TRBJ2-1	TGCGCCAGCAGCCAAGAGCCGGGCGGGGAGCAGTCTTC
5	0.61%	2336	CASSYGMNTEAFF	TRBV6-6	TRBD2	TRBJ1-1	TGTGCCAGCAGTTACGGGATGAACACTGAAGCTTTCTTT
6	0.52%	1992	CASSQGGRAPHTQYF	TRBV4-3	TRBD2	TRBJ2-3	TGCGCCAGCAGCCAAGGGGGAGGGCCCCCATACGCAGTATTTT
7	0.49%	1871	CASSQSGGSYEQYF	TRBV5-1	TRBD1	TRBJ2-7	TGCGCCAGCAGCAAAGTCAAGGGGGGTCTCTACGAGCAGTACTTC
8	0.48%	1847	CASSRPKSGRSGELFF	TRBV11-2	TRBD2	TRBJ2-2	TGTGCCAGCAGCCGACCCAAGAGCGGGAGAAAGTGGGGAGCTGTTTTTT

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 from a sample of interest and detects nodes having a degree higher than would be expected from placing a given node in the control sample.

Test for TCR neighbor enrichment

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

$$n_i^s \stackrel{H_0}{\sim} \text{Poisson} \left(N_i^s \frac{n_i^c}{N_i^c} \right) \quad (1)$$

where c is the control sample.

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

Outline

Introduction

Setting

Interactive part

Concluding remarks

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

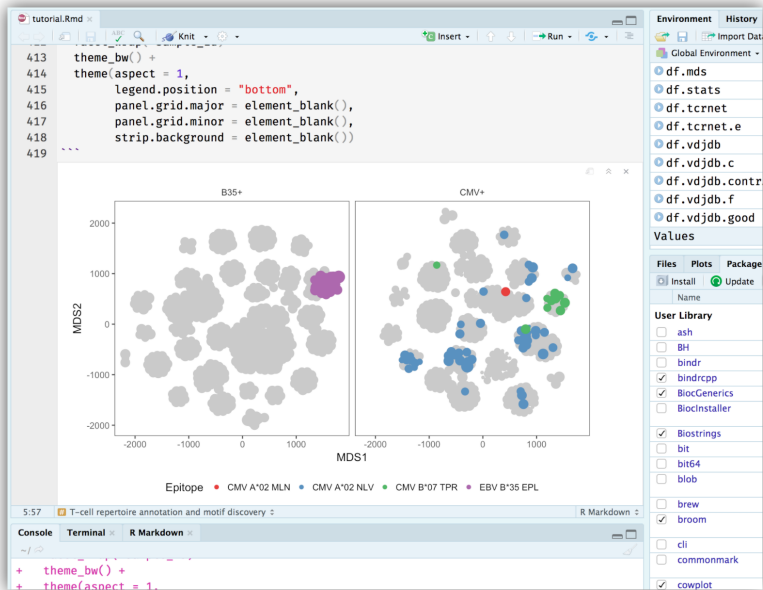
Introduction

Setting

Interactive part

Concluding remarks

Interactive part



Outline

Introduction

Setting

Interactive part

Concluding remarks

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
- ▶ Overlapped 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 specificity³ one will observe many false-positives.

Use proper controls, e.g. naive T-cells

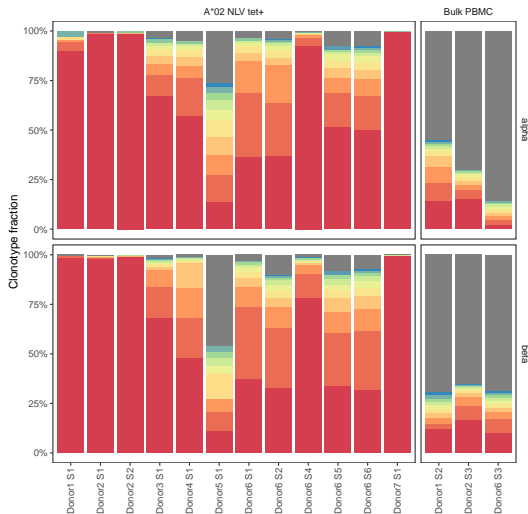
- ▶ TCRNET will fail in some cases simply because the repertoire of T-cells specific to a given antigen is dominated by a single hyperexpanded clonotype.

Always annotate and check large clonotypes

³According to VDJdb benchmark odds of matching the same epitope given 1 substitution in CDR3 β 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