Immune repertoire forensics

A RepSeq data analysis tutorial

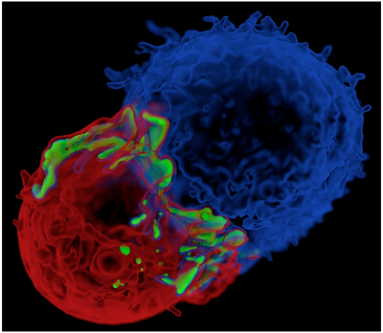
Mikhail Shugay, PhD

Skoltech, MA03172 course [Term 2, 2017-2018]

December 1, 2017

T-cell receptor

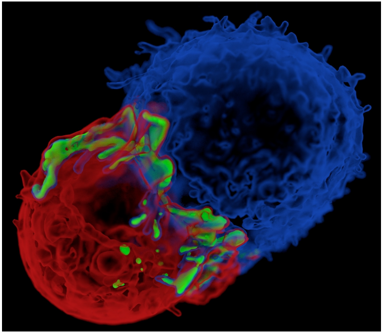
T-cell:APC contact



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

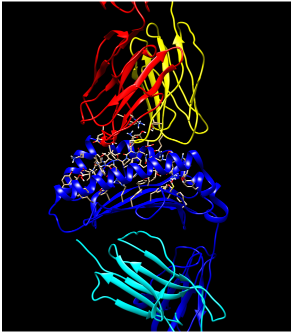
T-cell receptor

T-cell:APC contact



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

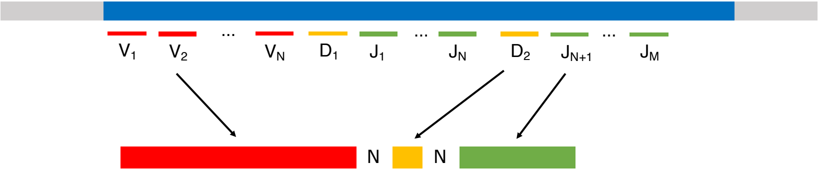
TCR:pMHC structure



PDB:1ao7, rendered using UCSF chimera, colored by chain

VDJ rearrangement

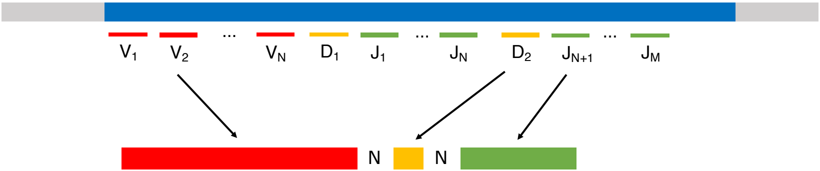
An example schema for TCRβ locus



Variable, Diversity and Joining are chosen at random, V-D and D-J junctions are filled with non-template N bases.

VDJ rearrangement

An example schema for TCRβ locus



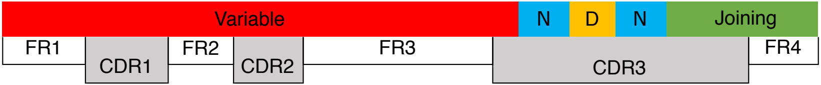
Variable, Diversity and Joining are chosen at random, V-D and D-J junctions are filled with non-template N bases.

VDJ rearrangement mechanism can be efficiently recaptured with a probabilistic model [`Murugan et al. PNAS 2012`]

$$
\begin{aligned}
P(\sigma) &= P(V)P(D,J) \\
&\times P(\#del_V|V)P(\#del_J|J)P(\#del_{D5},\#del_{D3}|D) \\
&\times P(\#ins_{VD})P(\#ins_{DJ}) \prod_{i\in ins_{VD}} P(b_i|b_{i-1}) \prod_{i\in ins_{DJ}} P(b_i|b_{i+1})
\end{aligned}
$$

TCR regions

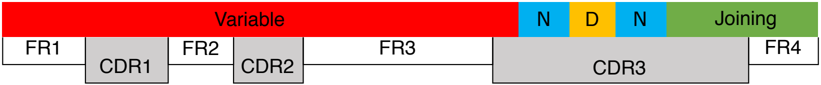
A TCR chains consists of the following regions:



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

TCR regions

A TCR chains consists of the following regions:

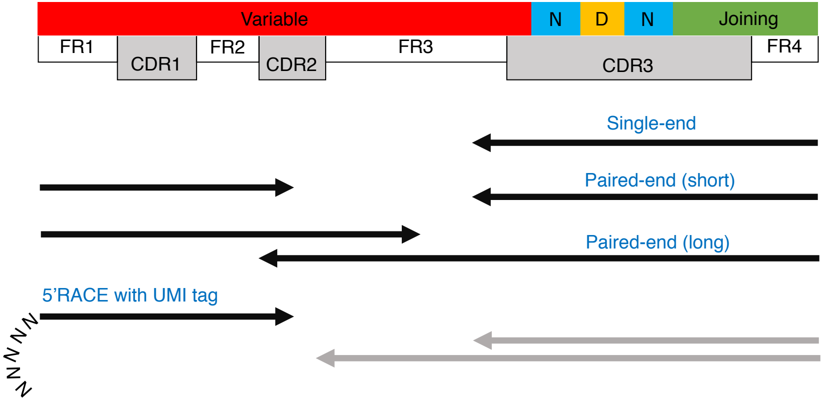


In total there are four framework (FRs) and three complementarity determining regions/loops (CDRs).

The likely 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

TCR repertoire sequecing



An example of RepSeq dataset

After all pre-processing steps:

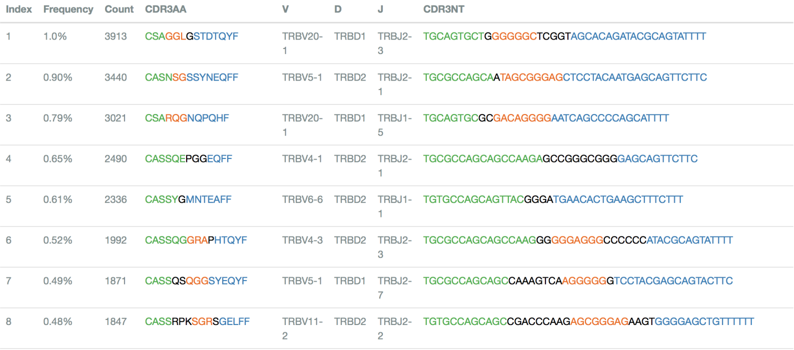
- Read grooming (filtering, etc)
- UMI-based assembly (for molecular barcoded data)
- V-D-J mapping and clonotype assembly

An example of RepSeq dataset

After all pre-processing steps:

- Read grooming (filtering, etc)
- UMI-based assembly (for molecular barcoded data)
- V-D-J mapping and clonotype assembly

We finally get clonotype frequency tables that look like



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