Immune repertoire forensics

A RepSeq data analysis tutorial

Mikhail Shugay, PhD

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

December 1, 2017

Outline

Introduction

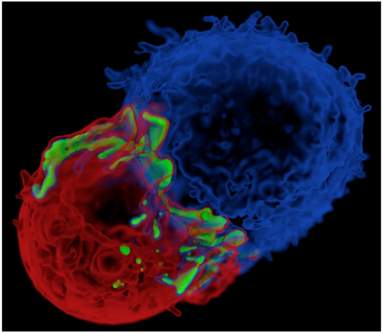
Getting started

Interactive part

The assignment

T-cell receptor

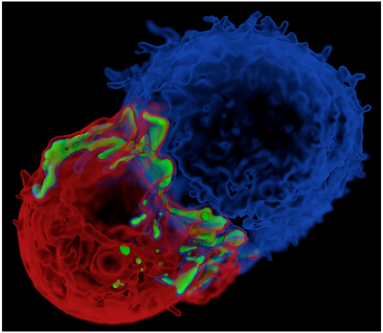
T-cell:APC contact



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

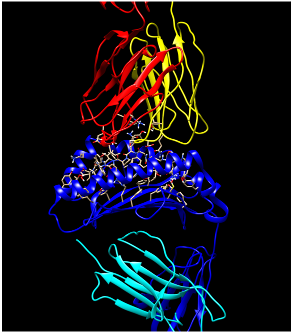
T-cell receptor

T-cell:APC contact



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

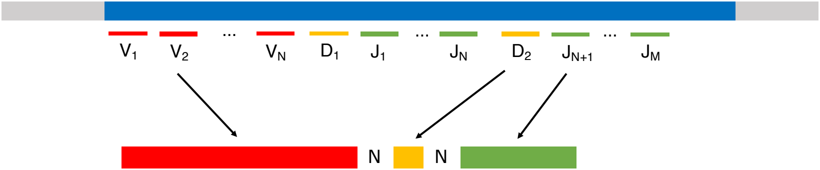
TCR:pMHC structure



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

VDJ rearrangement

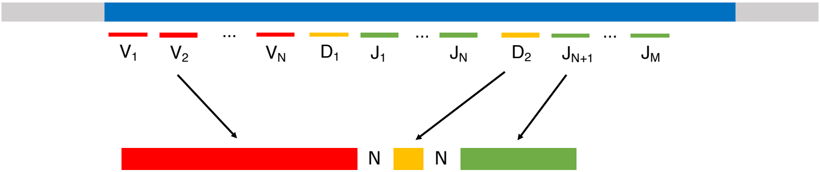
An example schema for TCRβ locus



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



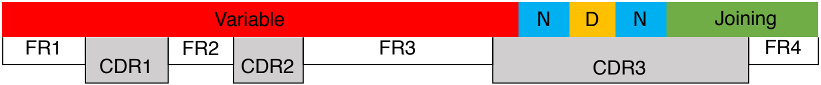
Variable, Diversity and Joining are chosen at random, V-D and D-J junctions are filled with non-template N bases.

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

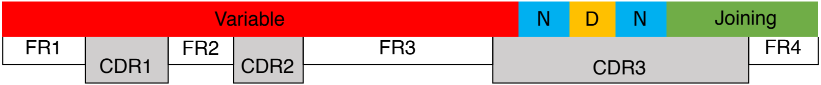
A TCR chains consists of the following regions:



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

TCR regions

A TCR chains consists of the following regions:

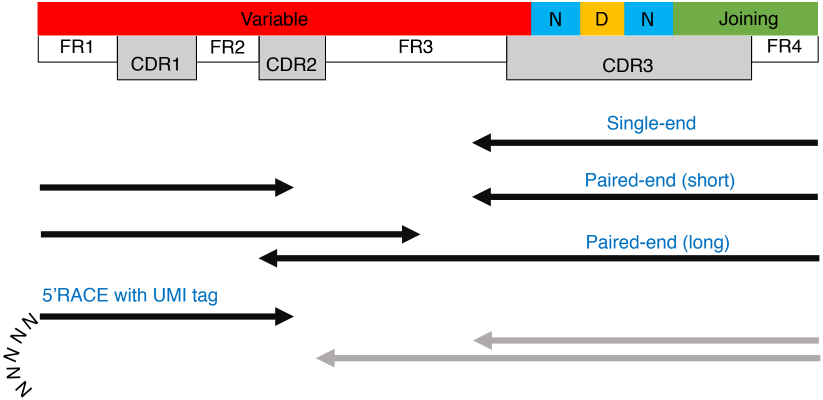


In total there are four framework (FRs) and three complementarity determining regions/loops (CDRs).

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 sequencing



An example of a RepSeq dataset

After all pre-processing steps:

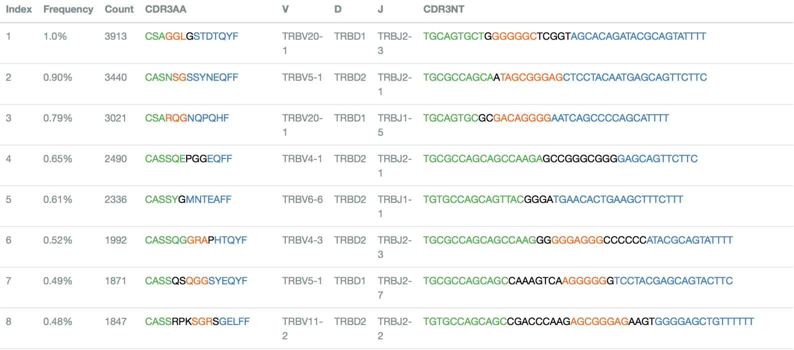
- Read grooming (filtering, etc)
- UMI-based assembly (for molecular barcoded data)
- V-D-J mapping and clonotype assembly

An example of a 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

Outline

Introduction

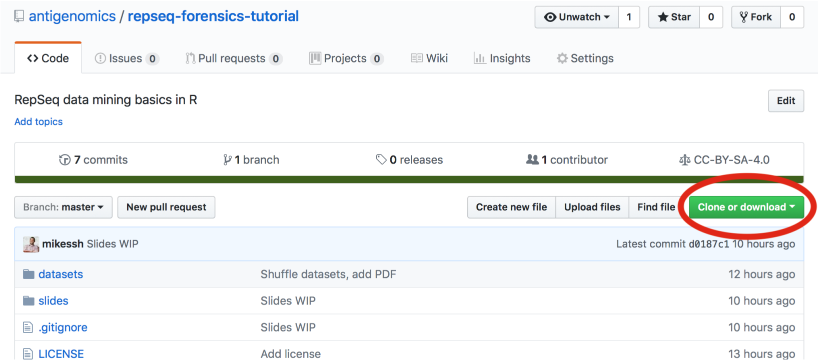
Getting started

Interactive part

The assignment

Downloading data

Navigate to `https://github.com/antigenomics/repseq-forensics-tutorial` and download the data + code bundle as zip



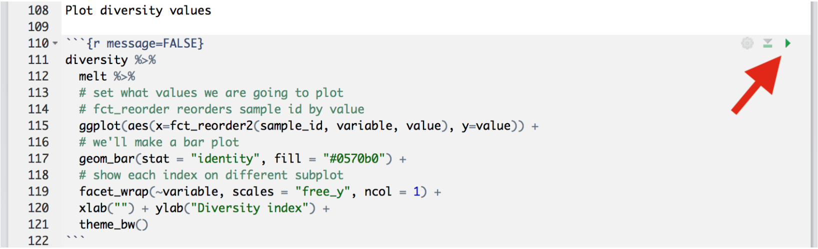
RepSeq data mining basics in R

mikessh Slides WIP — Latest commit d0187c1 10 hours ago

datasets	Shuffle datasets, add PDF	12 hours ago
slides	Slides WIP	10 hours ago
.gitignore	Slides WIP	10 hours ago
LICENSE	Add license	13 hours ago

Executing R code

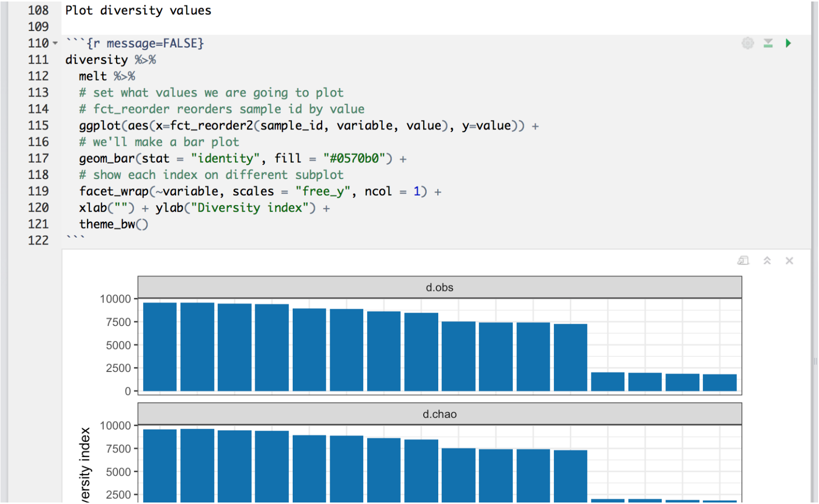
Open the `tutorial.Rmd` in RStudio, it can be found in the root folder of the bundle.



~~~
Plot diversity values

```{r message=FALSE}
diversity %>%
  melt %>%
  # set what values we are going to plot
  # fct_reorder reorders sample id by value
  ggplot(aes(x=fct_reorder2(sample_id, variable, value), y=value)) +
  # we'll make a bar plot
  geom_bar(stat = "identity", fill = "#0570b0") +
  # show each index on different subplot
  facet_wrap(~variable, scales = "free_y", ncol = 1) +
  xlab("") + ylab("Diversity index") +
  theme_bw()
```
~~~

Executing R code



~~~
Plot diversity values

```{r message=FALSE}
diversity %>%
  melt %>%
  # set what values we are going to plot
  # fct_reorder reorders sample id by value
  ggplot(aes(x=fct_reorder2(sample_id, variable, value), y=value)) +
  # we'll make a bar plot
  geom_bar(stat = "identity", fill = "#0570b0") +
  # show each index on different subplot
  facet_wrap(~variable, scales = "free_y", ncol = 1) +
  xlab("") + ylab("Diversity index") +
  theme_bw()
```
~~~

Outline

Introduction

Getting started

Interactive part

The assignment

Basic RepSeq analysis approaches

We're going to perform 4 types of TCR RepSeq data analysis:

- **Diversity analysis** is useful to tell naive T-cells having a rather uniform clonal frequency distribution from antigen-experienced T-cells which display a number of highly expanded (high frequency) clones

Basic RepSeq analysis approaches

We're going to perform 4 types of TCR RepSeq data analysis:

- **Diversity analysis** is useful to tell naive T-cells having a rather uniform clonal frequency distribution from antigen-experienced T-cells which display a number of highly expanded (high frequency) clones
- **Variable segment usage** profile can be useful for distinguishing different subsets of T-cells: different subsets typically have a characteristic pattern of preferred V segments.

Basic RepSeq analysis approaches

We're going to perform 4 types of TCR RepSeq data analysis:

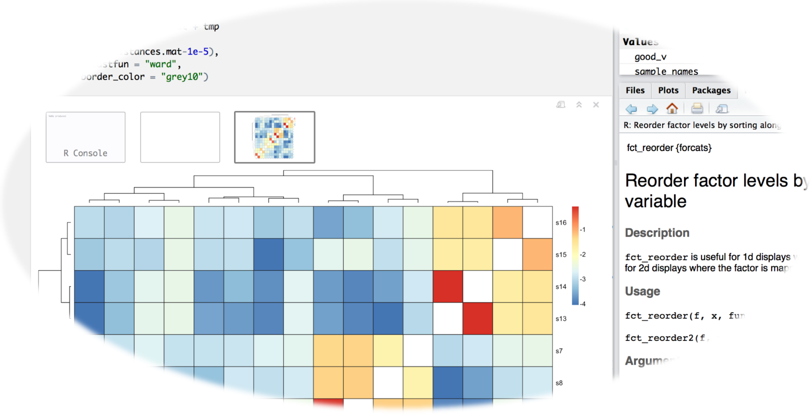
- **Diversity analysis** is useful to tell naive T-cells having a rather uniform clonal frequency distribution from antigen-experienced T-cells which display a number of highly expanded (high frequency) clones
- **Variable segment usage** profile can be useful for distinguishing different subsets of T-cells: different subsets typically have a characteristic pattern of preferred V segments.
- **Clonotype sharing** - the overlap of hypervariable CDR3 region sequences, generally useful for determining sample origin.

Basic RepSeq analysis approaches

We're going to perform 4 types of TCR RepSeq data analysis:

- **Diversity analysis** is useful to tell naive T-cells having a rather uniform clonal frequency distribution from antigen-experienced T-cells which display a number of highly expanded (high frequency) clones
- **Variable segment usage** profile can be useful for distinguishing different subsets of T-cells: different subsets typically have a characteristic pattern of preferred V segments.
- **Clonotype sharing** - the overlap of hypervariable CDR3 region sequences, generally useful for determining sample origin.
- **TCR sequence annotation** using a curated database of TCRs with known antigen specificity.

Interactive part



```
tmp

tances.mat-1e-5),
stfun = "ward",
order_color = "grey10")
```

R: Reorder factor levels by sorting alon

fct_reorder {forcats}

Reorder factor levels by variable

Description

`fct_reorder` is useful for 1d displays for 2d displays where the factor is map

Usage

```
fct_reorder(f, x, fu
fct_reorder2(f,
```

Argume

Outline

Introduction

Getting started

Interactive part

The assignment

The assignment

Using the analysis results we've obtained we need to assign feature labels to each sample. Namely, you need to fill the table with the following structure:

sample	**donor**	**subset**	**phenotype**	**CMVstatus**
s1	D1	CD4	memory	CMV-
s2	D2		naive	CMV+
s3	D1	CD8	naive	CMV-
...	...	...	...	...

Details

Table filling rules:

- Column names should match those on previous slide
- Sample id should be one of $s_1..s_{16}$
- Two distinct donor IDs should be used, naming doesn't matter
- Subset should be either **CD4** or **CD8**
- Phenotype should be either **memory** or **naive**
- CMV status should be either **CMV+** or **CMV-**
- Unknown/ambiguous fields should be left blank

A hint

While you can unambiguously assign CD4/8 and memory/naive labels, as well as point out biological replicates of the same sample, assigning donor labels is tricky.

First, it is impossible to link CD4-CD8 cells of the same donor. Same for CMV status, that is unambiguous only for CD8+ memory T-cells. Therefore I expect that you mark donors in the way they will distinguish samples/replicas coming from the same and different donors.

I.e. there is no problem if donor labels are swapped between CD4 and CD8 T-cells as far as they point to distinct donors for CD4 or CD8 T-cells coming from different donor and the same donor for replicas.

Feedback

Send me filled tables as plain text tab-delimited files, the file name should be in format `your-name.assignment.txt`

Final remarks

Thanks for your attention!

These slides and a PDF file containing compiled analysis results can be found in `slides/` and root folders of the data and code bundle.