Immune repertoire annotation: a RepSeq data analysis tutorial

Introduction

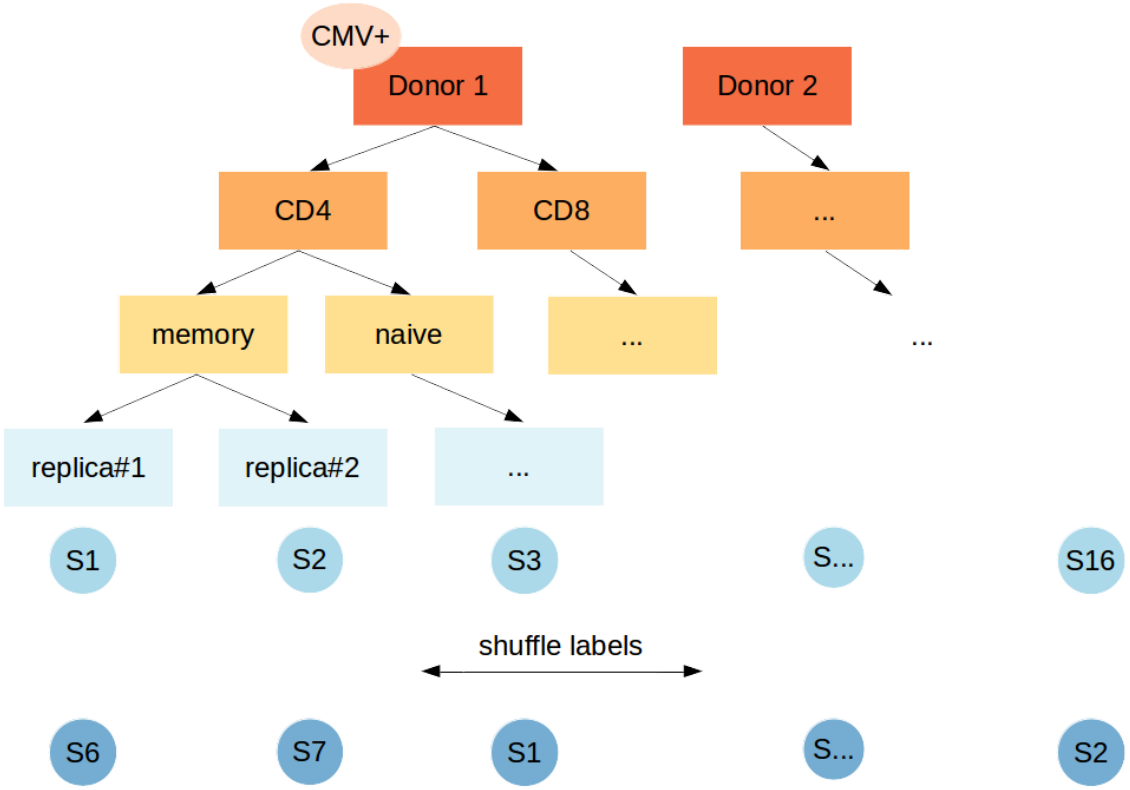
https://github.com/mikeraiko/repseq-annotation-tutorial

This tutorial covers some basic aspects of Immune Repertoire Sequencing (RepSeq) data analysis focused on T-cell receptor (TCR) repertoires:

- Repertoire diversity analysis
- Segment usage analysis
- Repertoire overlap analysis
- Annotation of antigen-specific TCR sequences

The main idea of this tutorial is to demonstrate the immense amount of information encoded in immune repertoires and the ability to decode relevant characteristics from the RepSeq data using relatively simple bioinformatic/data mining methods.

Given a set of unlabeled samples from different donors (generated as shown below), T-cell subpopulations and phenotypes, we can reliably infer the sample origin and even some properties of the immunological status of a (relatively) healthy donor.



This analysis uses 16 samples of 10,000 random reads from two donors from [Qi et al. PNAS 2014] study (sample labels and TCR nucleotide sequences are removed).

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	CD4	naive	CMV+
s3	D1	CD8	naive	CMV-
...	...	...	...	...

Table filling rules:

- Column names should match those on previous slide
- Sample id should be one of s1::s16
- 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 CMVI 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.

Additional materials

To read:

- "Анализ индивидуальных репертуаров Т-клеточных рецепторов" https://biomolecula.ru/articles/analiz-individualnykh-repertuarov-t-kletochnykh-retseptorov
- "12 методов в картинках: иммунологические технологии" https://biomolecula.ru/articles/12-metodov-v-kartinkakh-immunologicheskie-tekhnologii
- Short review: "Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges" https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6058020/

To watch:

- “Иммуноинформатика: алгоритмический подход к решению прикладных задач иммунологии | Яна Сафонова” https://www.youtube.com/watch?v=6UEHdyxRZtw
- “Молекулярное баркодирование, анализ репертуаров Т-клеточных рецепторов и антител | Дмитрий Чудаков” https://www.youtube.com/watch?v=88LH7Ge0lRo

To do:

Another (extended and advanced) RepSeq tutorial, including all analysis steps and tools:
https://github.com/mikessh/repseq-tutorial