Домашнее задание по имууногеномике. Basic analysis of antibody sequences

ИБ. осень 2020

Чангалиди Антон

Использовал IgBlast + Python для обработки данных

Task 1.

Analyze the joint usage of V and J genes: for each sequence find the closest germline V and J genes, list of all VJ pairs occurring in the sample, and create a plot (e.g., heatmap) showing the number of sequences for each VJ pair.

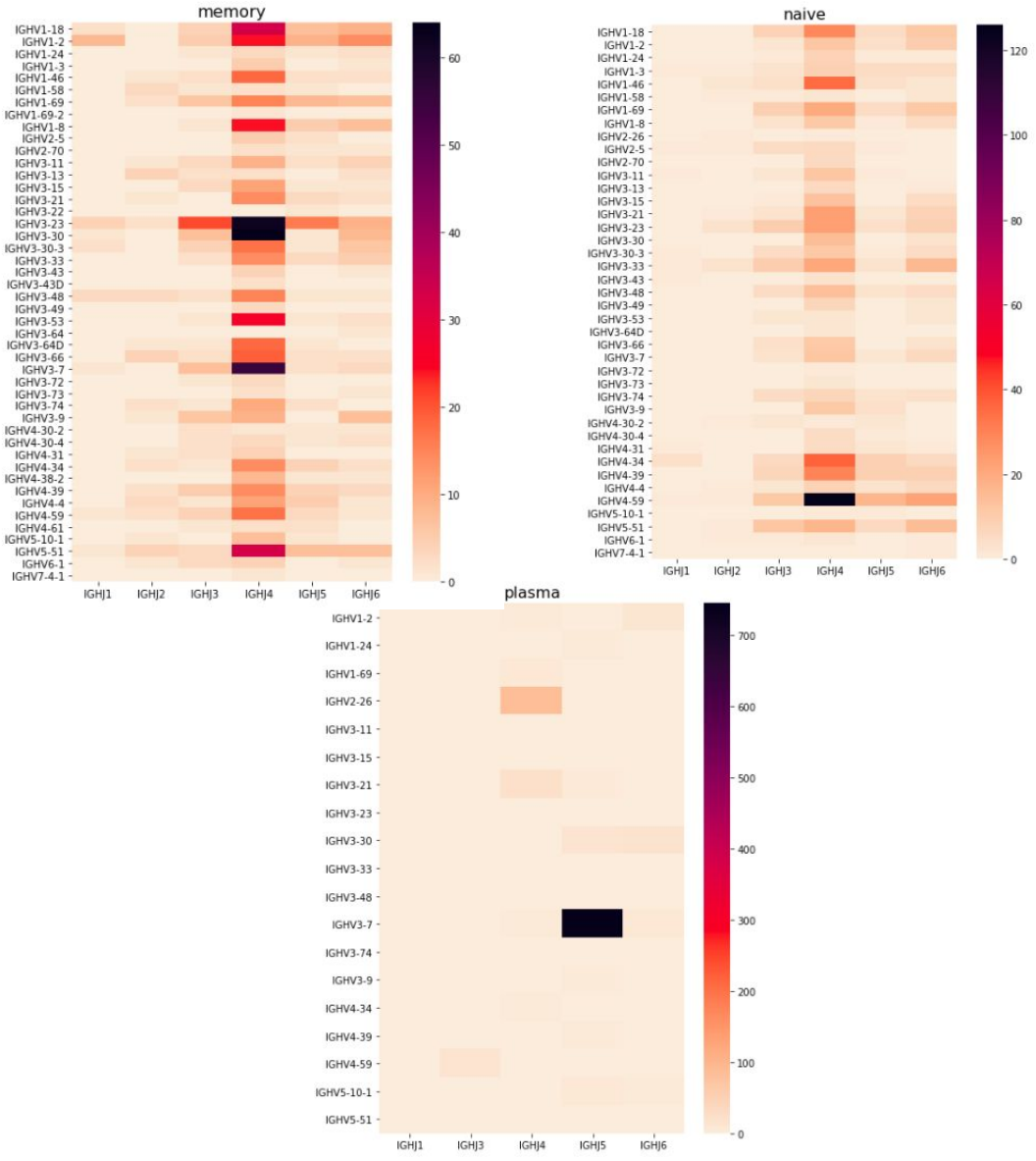


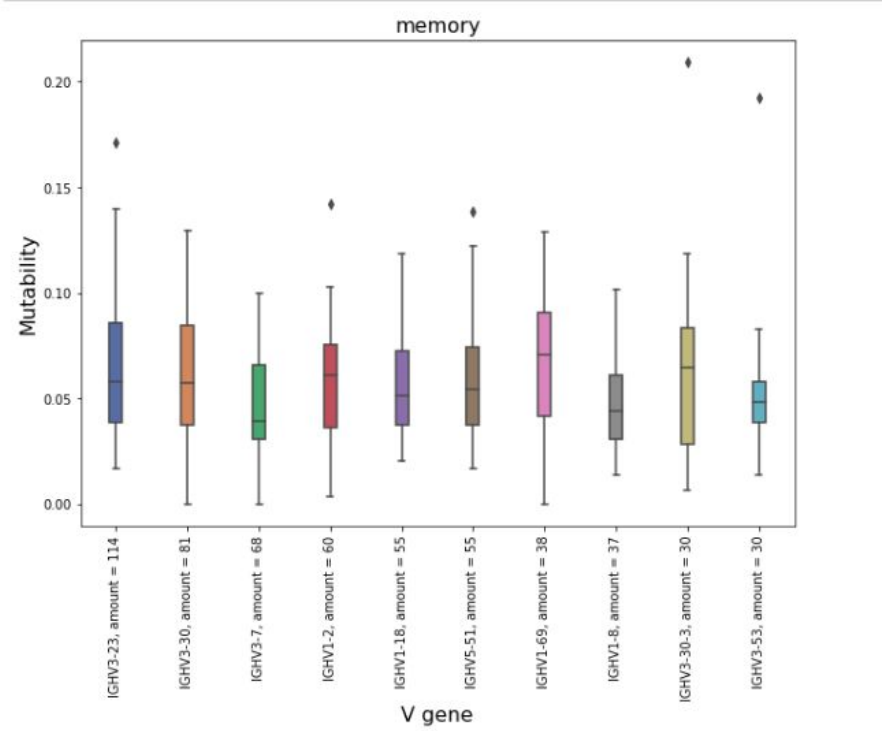
Рис.1. Хитмап представленности VJ комбинаций у разных типов имунных клеток.

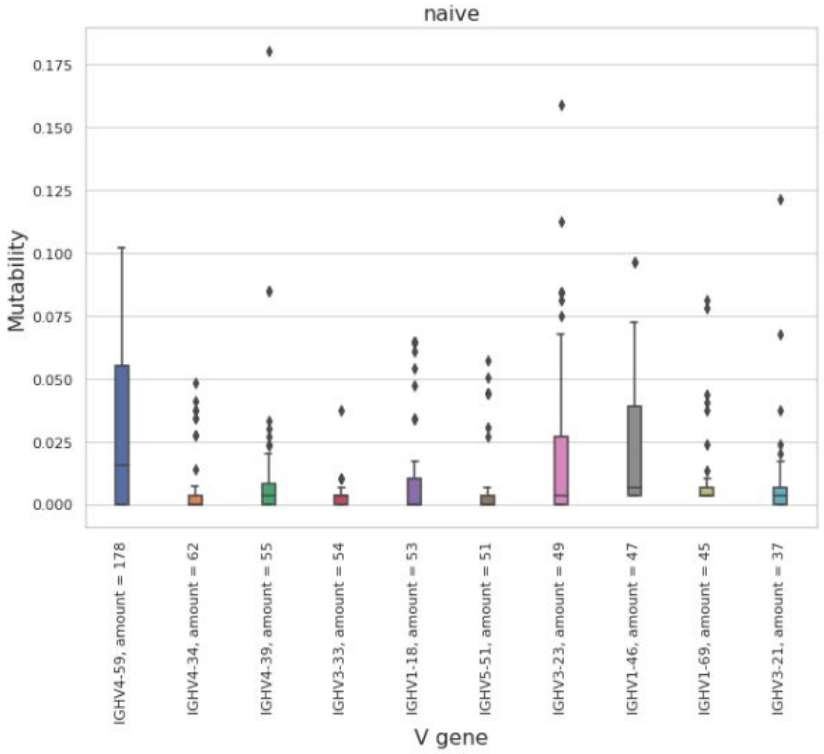
На рис.1 выше представлен хитмап представленности V-J комбинаций у разных типов иммунных клеток (всего 144 VJ-комбинации в naive-клетках, 171 комбинация в memory-клетках, и 32 - в plasma-клетках). У плазматических клеток наблюдается малое число комбинаций (самая представленная комбинация: IGHV3-7 и IGHJ5) - малое число может быть связано с тем, что эти клетки образовались от одной B-клетки с одной V+J комбинацией (направленной на борьбу с одним антигеном).

В memory и naive клетках самый представленный J-ген - IGHJ4, причем, в memory-клетках имеется бОльшее количество комбинаций, чем в других клетках (наверное, из-за того, что эти клетки содержат рецепторы, восприимчивые к различным антигенам).

Task 2

Find 10 most used V genes in the sample and analyze their mutability . For each gene, analyze sequences aligned to it and compute the number of differences in each alignment. Mutability is the distribution of the number of differences. Visualize mutability of 10 most used V genes in any convenient form (e.g., using boxplot).





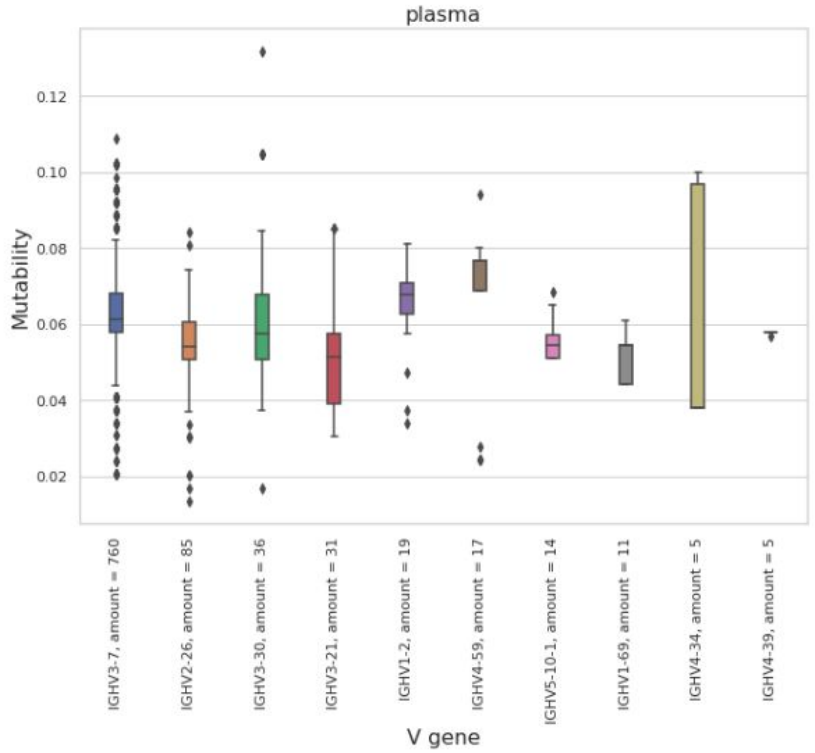


Рис.2 Мутабельность самых частых генов разных клетках

Самый низкий процент мутаций - у наивных клеток. Это логично, так как они еще не прошли профилирование и не имеют "специализации". Мутации возникают в клетках плазмы и памяти - после встречи с антигенами, поэтому эти клетки более изменчивы. При этом, клетки плазмы имеют больший разброс (возможно, из-за того, что им необходимо иметь иммунный респонс к большему числу антигенов).

Task 3

Visualize distributions of CDR3 lengths.

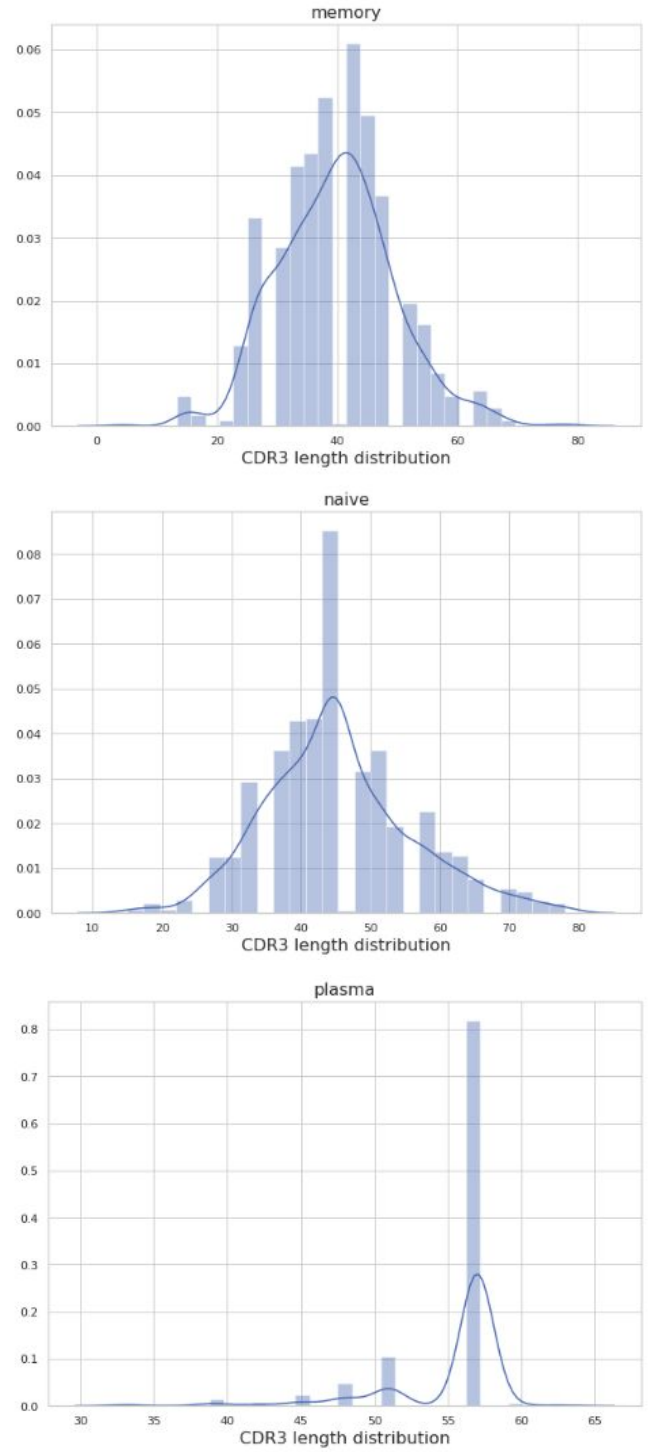


Рис.3. Распределение длины CDR3 в разных клетках.

У наивных клеток и клеток памяти схожий пик распределения CDR3 - у 40-45. Распределение CDR3 клеток плазмы отличается от предыдущих (выделяется 1 пик и почти только он). Скорее всего, это из-за небольшой вариабельности пар V+J генов в этой клетке, что может означать высокую специфичность к определенному антигену.

Task 4

Compute the fraction of non-productive sequences in the sample. Both IgBlast and DiversityAnalyzer report productiveness of input sequences as a part of the output.

Часть непродуктивных последовательностей в клетках памяти: 7.9%
Часть непродуктивных последовательностей в наивных клетках: 3.1%
Часть непродуктивных последовательностей в клетках плазмы: 3.2%

Процент непродуктивных клеток во всех 3 типах клеток достаточно мал. В клетках памяти мы видим высокий процент непродуктивных последовательностей (относительно других). Это может быть из-за более низкого отбора и более высокой мутабельности в этих клетках.