TCGA Mutations MIPT

Виктория Гросс

March 2019

1 Введение в использование базы данных TCGA

The Cancer Genome Atlas (TCGA) — это база данных, содержащая и способная анализировать данные о человеческих раковых опухолях, созданная как инструмент для исследования молекулярных нарушений в организме, приводящих к раку. Существует множество интерфейсов, используемых для получения данных с портала, а также много уже написанных анализаторов.

Один из наиболее известных — портал от National Cancer Institute (https://portal.gdc.cancer.gov/). В разделе Exploration можно подобрать случаи необходимые для исследования: местоположение опухоли, база данных (TCGA или другие подобные), тип рака. Данные можно отфильтровать по характеристикам пациентов (пол, возраст, раса, информация о выживании). Для выбранных случаев можно скачать файлы, содержащие данные об уровни экспрессии разничных генов, информацию о мутациях, клинические данные пациента. Затем эти данные в формате csv или json могут быть обработаны стандартными методами обработки данных python или R.

Данные могут экспортированы при помощи специальных библиотек: xenaPython для Python или RTCGA для R.

Кроме того, существует множество уже готовых интерфейсов, позволяющих определить, например, зависимость выживаемости от уровня экспрессии того или иного белка. Например, http://kmplot.com/analysis/ . Создатели этого сайта, использовали библиотеки языка R для получения данных, нормировки и анализа данных с нескольких баз данных: TCGA и Affymetrix. Аналогичными методами, описанными в их статье, можно повторить их результаты для выбранной базы данных.

Кроме того, специально для анализа и визуализации данных TCGA был создан XenaBrowser (https://xenabrowser.net/heatmap/). С его помощью можно отобразить мутации определенного гена при определенном типе рака. Далее эти мутации можно отобразить на стурктуре белка, кодируемого этим геном (http://hg19.cravat.us/MuPIT_Interactive/). При этом можно наблюдать какие мутации характерны для определенной локализации рака.

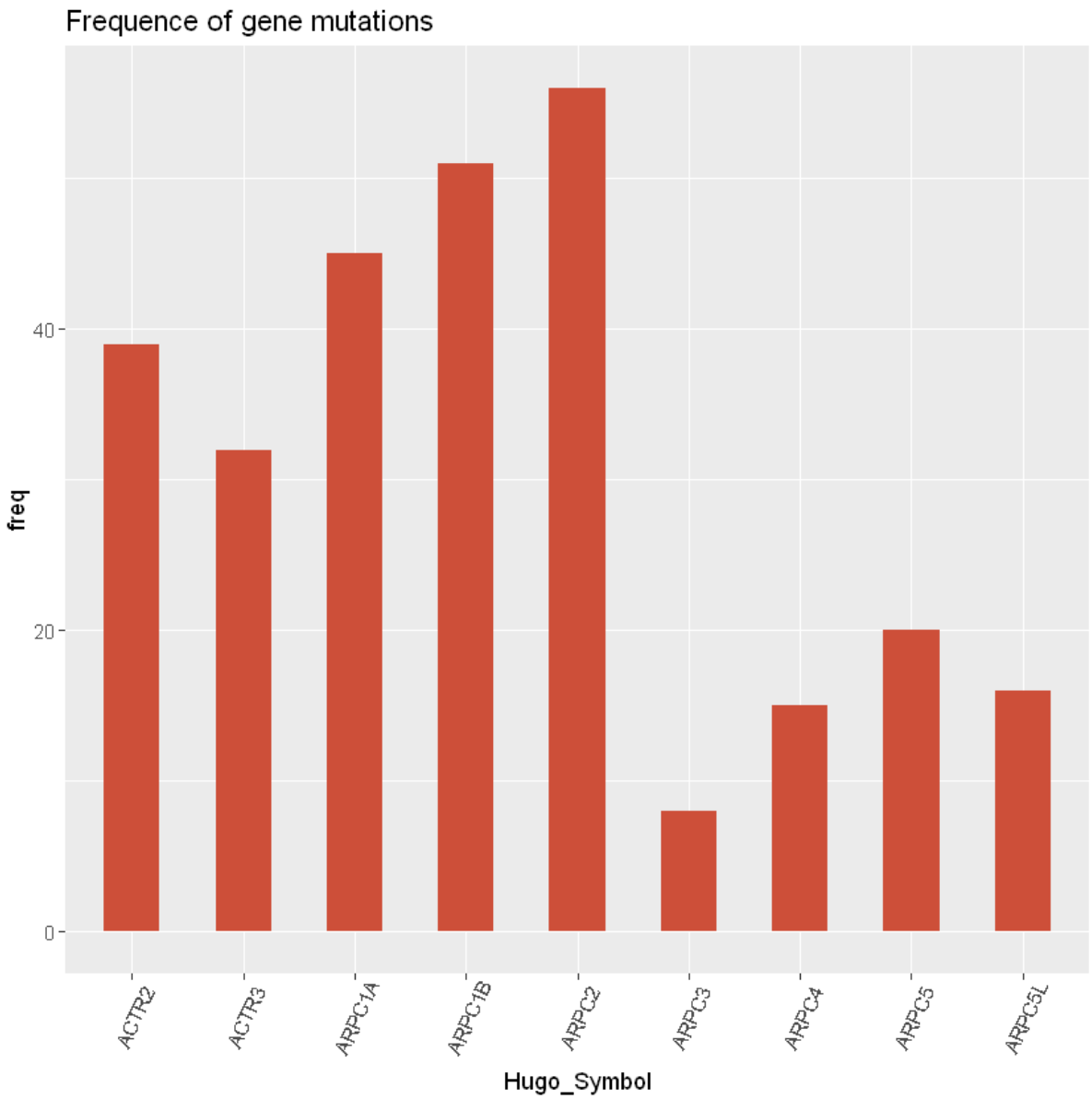
2 С помощью базы данных узнать какие мутации чаще всего бывают в комплексе Arp2/3 с помощью базы данных

Библиотека RTCGA языка R дает возможность скачать и объединить данные из TCGA, используя кодовый номер пациента, облегчающий хранение данных. Далее при помощи этой библиотеки можно преобразить данные TCGA в удобный для анализа формат.

Все мутации для всех типов опухолей, присутствующих в базе данных, были получены при помощи функции mutationsTCGA(...). Для каждого пациента были получены локализация опухоли, ген, тип мутации (Frame Shift Del, Frame Shift Ins, In Frame Del, Missense Mutation, Nonsense Mutation, RNA, Silent, Splice Site). Затем из них были отобраны только те, что относятся к генам, кодирующим комплекс Arp2/3 (ACTR2, ACTR3, ARPC1A, ARPC1B, ARPC2, ARPC3, ARPC4, ARPC5, ARPC5L).

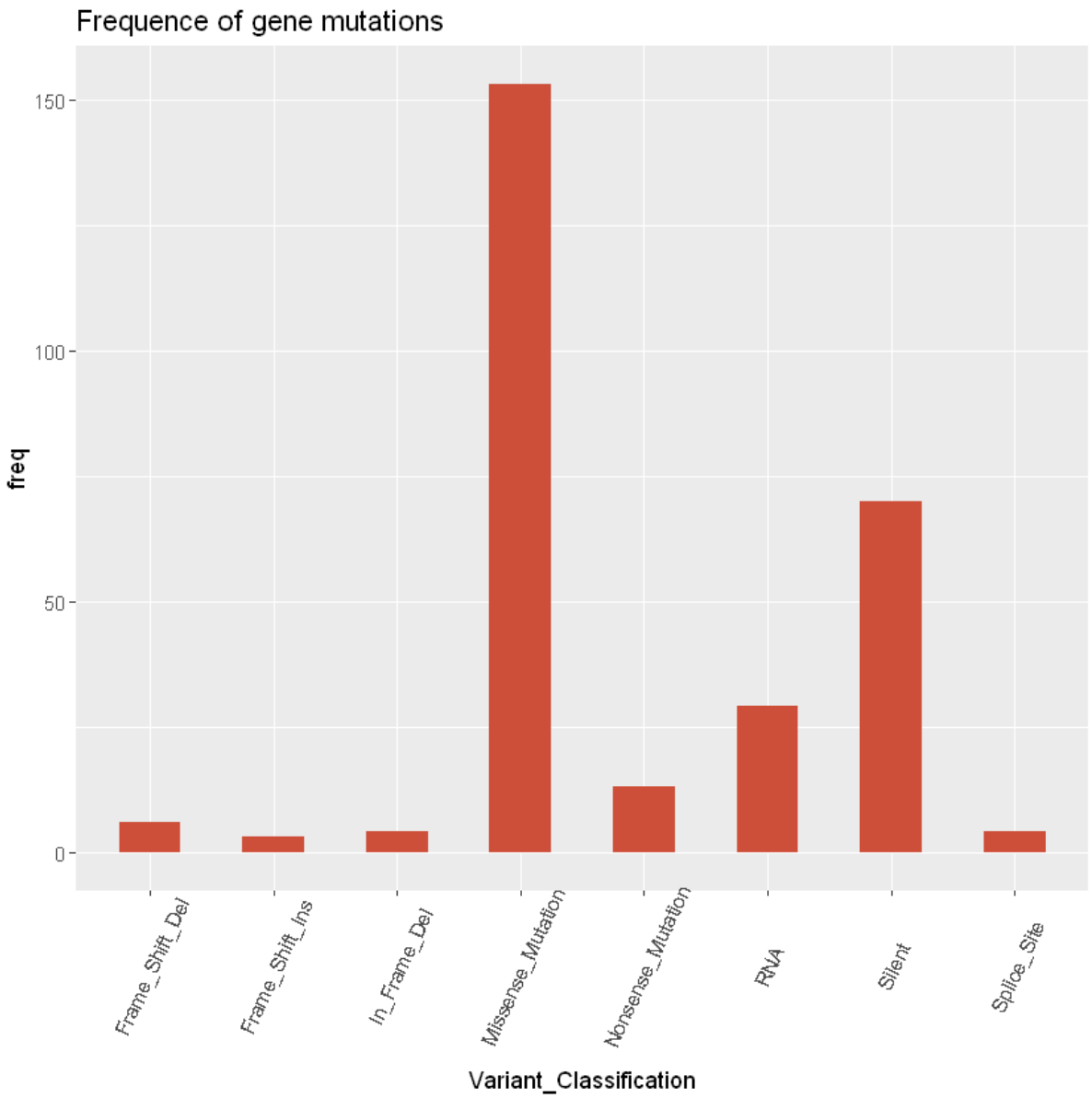
Посмотрим частоту встречаемости мутаций в отдельных белках комплекса:

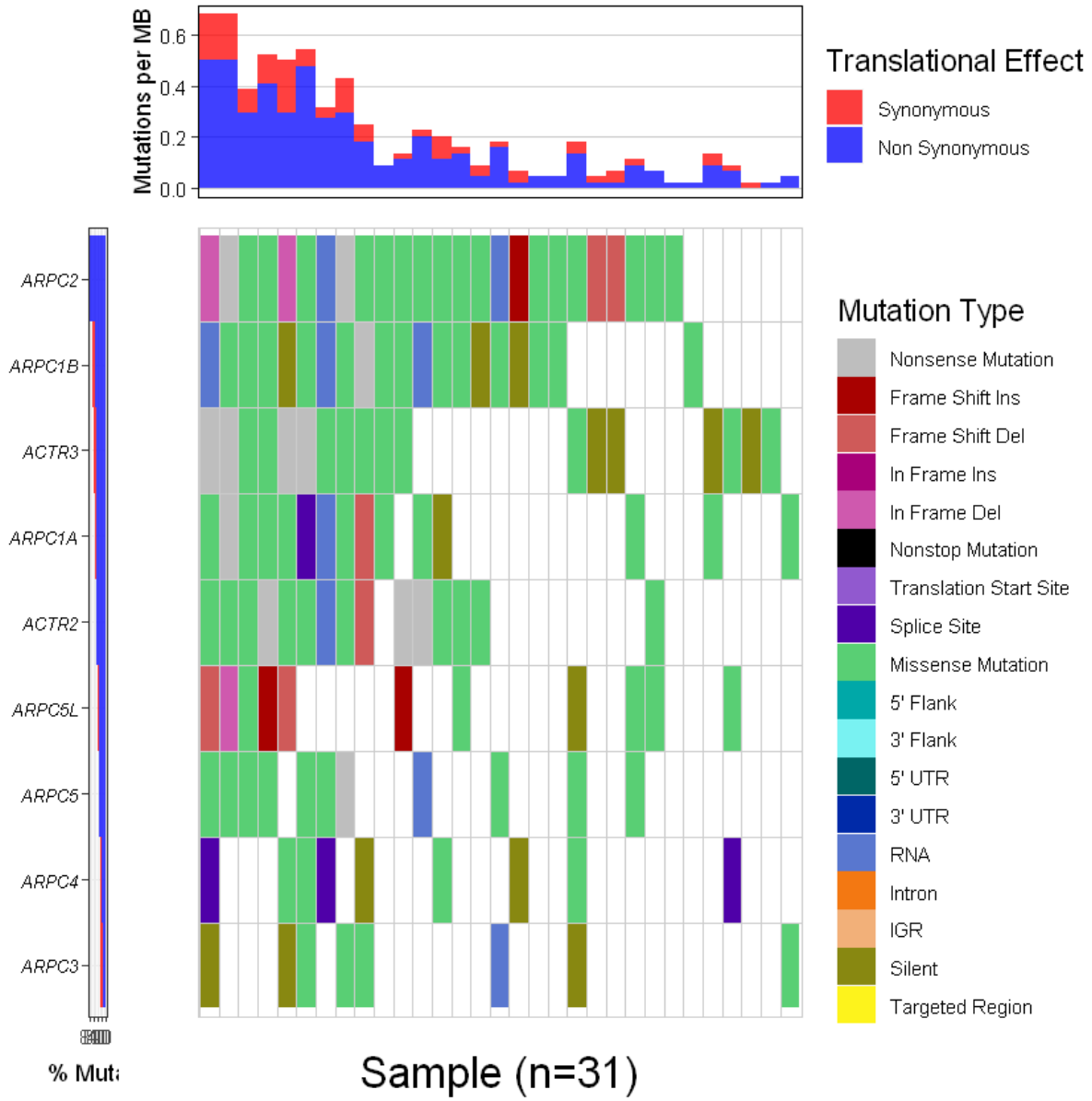
Hugo Symbol	freq
ACTR2	39
ACTR3	32
ARPC1A	45
ARPC1B	51
ARPC2	56
ARPC3	8
ARPC4	15
ARPC5	20
ARPC5L	16



И на частоту встречаемости мутаций различного типа:

Variant Classification	freq
Frame Shift Del	6
Frame Shift Ins	3
In Frame Del	4
Missense Mutation	153
Nonsense Mutation	13
RNA	29
Silent	70
Splice Site	4


Frequence of gene mutations
150
100
50
0
freq
Frame_Shift_Del
Frame_Shift_Ins
In_Frame_Del
Missense_Mutation
Nonsense_Mutation
RNA
Silent
Splice_Site
Variant_Classification



Waterfall диаграмма мутаций в генах комплекса Arp2/3

В результате, макисмальное количество мутаций происходит в гене ARPC2. Большая часть мутаций — точечные missence мутации.

3 Отобразить места, где эти мутации встречаются чаще всего, на структуре комплекса Arp2/3

4 Проверить - могут ли те или иные места, в которых бывают мутации, быть задействованы взаимодействии комплекса с лигандами или другими белками