

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 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. Наиболее известный — портал от National Cancer Institute (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 уровне экспрессии различных 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.

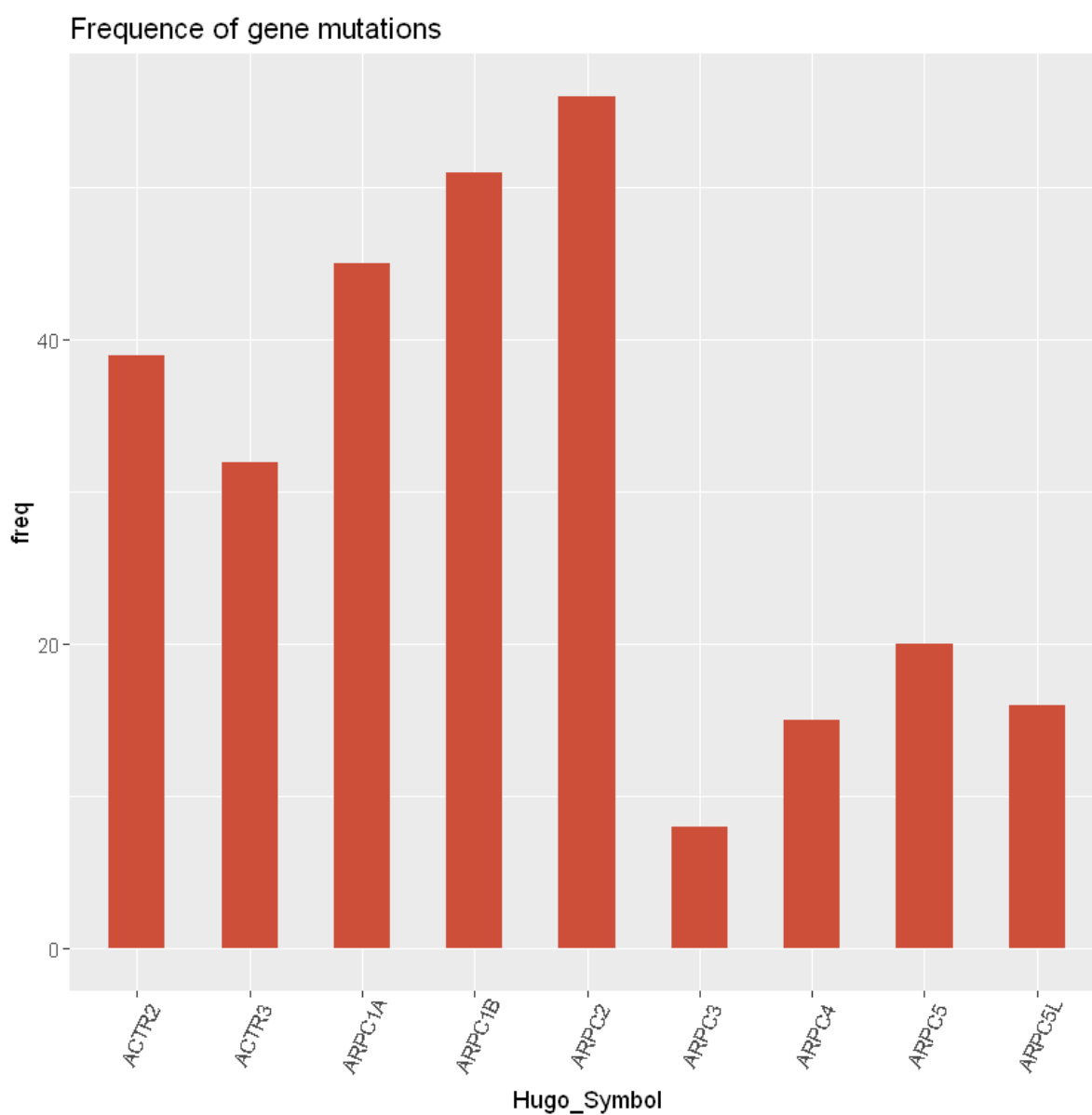
Кроме того, существует множество уже готовых интерфейсов, позволяющих определить, например, зависимость выживаемости от уровня экспрессии того или иного белка. Например, <http://kmplot.com/analysis/>. Создатели этого сайта, использовали библиотеки языка R для получения данных, нормировки и анализа данных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

1.1

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

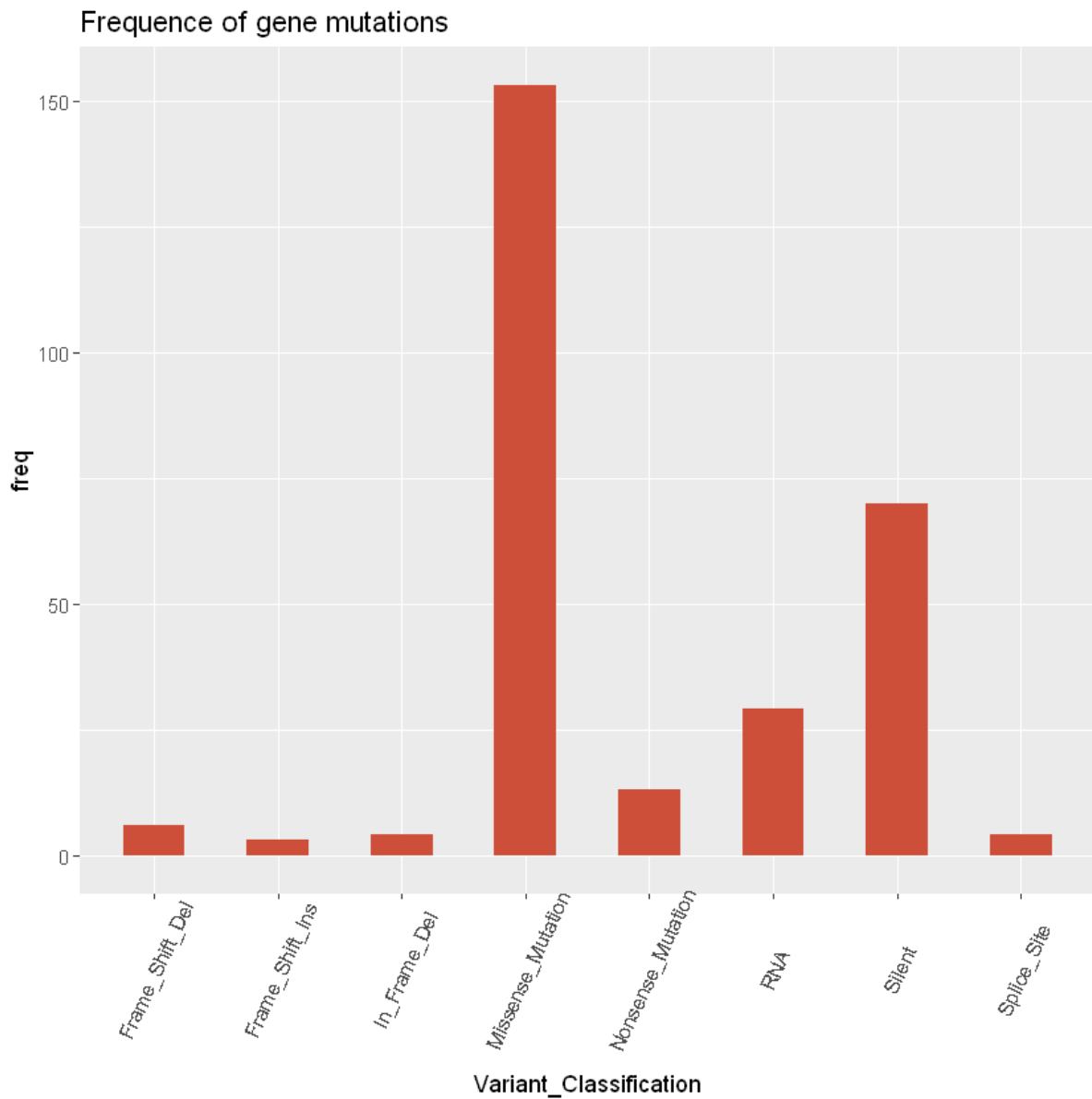
Посмотрим частоту встречаемости мутаций в отдельных белках комплекса:

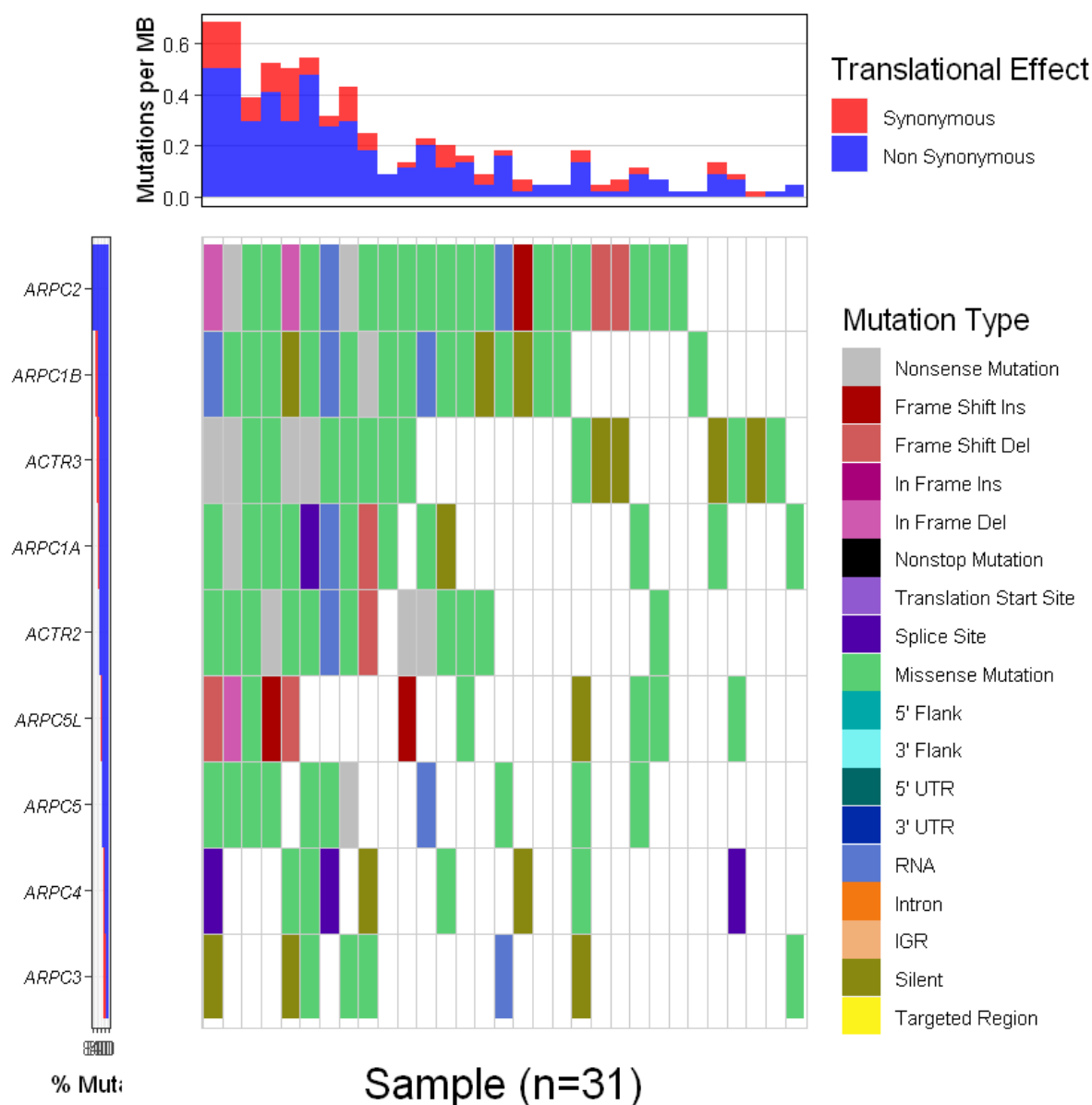
| Hugo Symbol | freq |
|-------------|------|
| ACTR2 | 39 |
| ACTR3 | 32 |
| ARPC1A | 45 |
| ARPC1B | 51 |
| ARPC2 | 56 |
| ARPC3 | 8 |
| ARPC4 | 15 |
| ARPC5 | 20 |
| ARPC5L | 16 |



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 |





Waterfall диаграмма мутаций в генах комплекса Arp2/3

В результате, максимальное количество мутаций происходит в гене ARPC2. Большая часть мутаций — точечные missense мутации.

- Отобразить места, где эти мутации встречаются чаще всего, на структуре комплекса Arp2/3
- Проверить - могут ли те или иные места, в которых бывают мутации, быть задействованы взаимодействия комплекса с лигандами или другими белками