Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
к магистерской диссертации

«Предсказание взаимодействий между раковой опухолью и организмом по данным метилирования ДНК, экзомного и РНК-секвенирования»

Автор: Эсаулова Екатерина Николаевна _____
Направление подготовки (специальность): 01.04.02 Прикладная математика и информатика
Квалификация: Магистр
Руководитель: Артемов М.Н., PhD _____

К защите допустить
Зав. кафедрой Васильев В.Н., докт. техн. наук, проф. _____

«___» _____ 20___ г.

Санкт-Петербург, 2017 г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Недавний прогресс в развитии поддерживающего лечения рака в виде иммунотерапии и вакцин показывает, что возможно изготовление эффективных нетоксичиных противо-опухолевых лекарств. Лекарства прошлого поколения — вакцины, нацеленные на сверхэкспрессированные (присутствующие в большом количестве в клетке) или частично экспрессированные нативные белки, ассоциированные с опухолью, и лекарства, блокирующие подавление иммунной системы, работали, но их результатом было частое появление аутоиммунности (иммунный ответ на здоровые ткани организма) и малый процент пациентов, отвечающих на лечение. Со временем удалось понять, как снизить побочные эффекты, но предопределение наличия ответа на лекарство остается нерешенной задачей.

Отличительным признаком развития раковой опухоли является накопление мутаций в ее клетках. Эти мутации дают возможность делать клетку раковой опухоли «целью» вакцины и иммунитета в целом благодаря присутствию в ней уникального мутированного белка. Мутированный белок присутствует в клетках опухоли и отсутствует в нормальной ткани. Ученые стали связывать количество мутаций в раковых клетках и их характеристики с эффективностью иммунных лекарств.

ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

В данной главе будут рассмотрена предметная область системной биологии, введено понятие севенирования нового поколения, рассмотрены существующие объяснения присутствия и отсутствия ответа на лечение у пациента, и перечислены технологии и программы, использующиеся для анализа геномных раковых данных.

1.1. Системная биология

Системная биология — это область науки, образованной на стыке биологии и теории сложных систем, нацеленная на изучение взаимодействий между составляющими частями биологических систем и на исследование механизмов формирования функций и системных свойств в результате этих взаимодействий. Для проверки созданных моделей и гипотез системная биология работает с различными типами экспериментальных данных, такими как высокопроизводительное секвенирование ДНК и РНК, эпигенетика (изучение факторов транскрипции, не кодируемых в ДНК).

1.2. Высокопроизводительные методы секвенирования

Самые популярные методы:

а) Секвенирование ДНК — определение нуклеотидной последовательности генома. В результате становится известна похромосомная последовательность ДНК.

б) Секвенирование экзома — определение нуклеотидной последовательности кодирующей части ДНК. При сравнении последовательностей здоровых и раковых тканей помогает определить мутации, накопленные в генах клеток опухоли.

в) Секвенирование РНК, или анализ экспрессии генов — определение нуклеотидной последовательности РНК, присутствующей в клетке. Анализ экспрессии помогает определить, как ведут себя гены в различных тканях и организмах при различных условиях.

г) Анализ метилирования — определение изменений молекулы ДНК, не меняющих ее последовательность, но влияющих на экспрессию генов в клетке.

Результаты секвенирования ДНК и РНК представляются в двух форматах:

а) fastq файл, необработанные данные. Это текстовый формат, где на каждую запись приходится четыре строки: имя записи, прочитанная последовательность нуклеотидов, разделитель, строка качества, состоящая из ASCII-символов.

```
@D00124:417:H5NY7BCXY:1:1104:1739:2189 1:N:0:1 NACTTGACTTT
ACTGAGGCCATGATTAAGAGGGACGGCCGGGGGCATTCGTA
+
DDDDDHIHHIIIIIIIIIIGIIIIIIHIHIIIIIIIIIIH
@D00124:417:H5NY7BCXY:1:1104:1682:2207 1:N:0:1 NACTTGACTTT
GGAGCTGGATGACCTGGAGAAGGATGAGTTGTGAGCCACGG
+
DDDDDHHIEHHFHIIIIHIHIIGHHIHHHHIIIIIHIIIII
```

Рисунок 1 – Пример fastq файла

б) SAM/BAM файл. Этот файл получается в результате выравнивания fastq файла на геном. В нем одна запись соответствует одной строке и несет информацию о том, с какого места пришла мутация, и набор других статистик. SAM-файл является текстовым, BAM — бинарным представлением SAM-файла.

```
1 D00124:417:H5NY7BCXY:1:2110:1586:90150  0   chr1    3029352 255 41M *   0   0
  GACAAAGAGAAAGGAGAGAGGAGACAAGACAAAGAGAGAGG   DDCDDHHEH<<GHHHHDG<GHHIHIIIHHHIHIHHGGHGHH
     NH:i:1  HI:i:1  AS:i:40 nM:i:0
2 D00124:417:H5NY7BCXY:1:2116:1414:23472  0   chr1    3029352 255 41M *   0   0
  GACAAAGAGAAAGGAGAGAGGAGACAAGACAAAGAGAGAGG   DDDDDIHHHGGGEHFHHHGEGCHHHIGIIIIIHIHHDG@EE
     NH:i:1  HI:i:1  AS:i:40 nM:i:0
3 D00124:417:H5NY7BCXY:1:1210:5360:90297  16  chr1    3059116 0   41M *   0   0
  GTCTGATTGCATAAAGGTTGAGTGCCCTAGGGTCCTTTCCC   IHEHIIIIIHGIIGIIIIIIIIHIIGIIIHIHIHHHDDDDD
     NH:i:7  HI:i:1  AS:i:40 nM:i:0
4 D00124:417:H5NY7BCXY:1:2207:20814:27669 16  chr1    3059116 0   41M *   0   0
  GTCTGATTGCATAAAGGTTGAGTGCCCTAGGGTCCTTTCCC   EHGC<<1EHEHHE@<1D1@C<0C<1?FC<<1<HF<<0?@@B
     NH:i:7  HI:i:1  AS:i:40 nM:i:0
5 D00124:417:H5NY7BCXY:1:2215:10576:41984 16  chr1    3059119 0   41M *   0   0
  TGATTGCATAAAGGTTGAGTGCCCTAGGGTCCTTTCCCCAG   IIHIIIIIIIHIIEIHIHEC1IHHIIHHEGHFHHEDDD?DD
     NH:i:6  HI:i:1  AS:i:40 nM:i:0
6 D00124:417:H5NY7BCXY:1:1102:7011:97243  16  chr1    3063352 0   41M *   0   0
  ATTTATTTATTTATTTATTATACGTAAGTACACTGTAGCTG   CIIIIIIIHIIHIIIIHIIIHHIIIHIHIHHHFHFFBDADA
     NH:i:6  HI:i:1  AS:i:40 nM:i:0
```

Рисунок 2 – Пример SAM файла

Файлы в формате fastq могут быть получены из BAM/SAM файлов. Для таких случаев используется программа SamToFastq из набора Picard tools [1].

1.3. Роль мутаций в развитии и прогрессии рака

Присутствие мутаций в кодирующих регионах ДНК клетки ведет к тому, что эта клетка может быть распознана имунной системой и ликвидирована. Мутации бывают:

а) Одноточечные, приводящие к замене одного нуклеотида на другой. Если мутация попадает в кодирующий регион белка, она может вызвать изменение одной его аминокислоты, либо привести к обрыву белка, либо ничего не изменить.

б) Короткие вставки и удаления – инделы (indel, от английского INsertion-DELetion).

Клетки опухолей накапливают много мутаций (иногда сотни и тысячи), и их анализ может помочь в понимании развития опухолей.

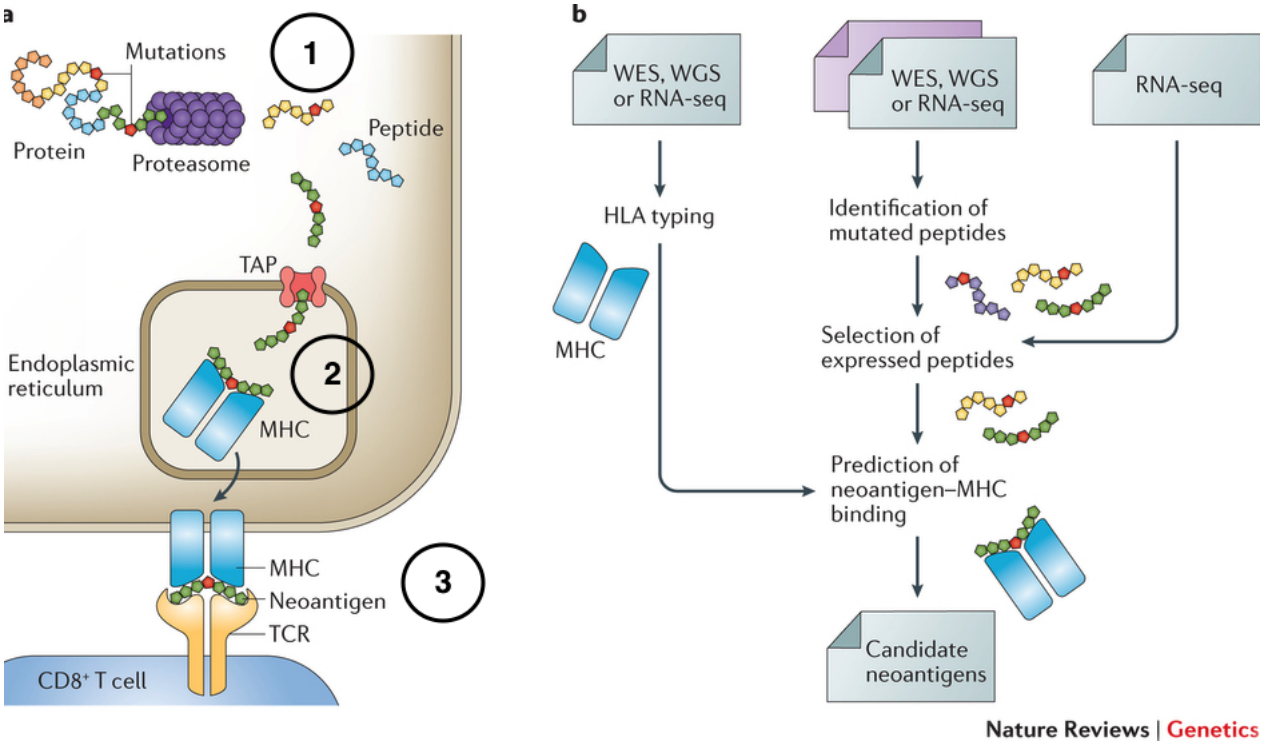


Рисунок 3 – а) Отражение мутаций на внутренней жизни клетки, б) Способ изучения мутаций и их влияния на развитие клетки методами секвенирования РНК и ДНК [2]

а) В случае, когда мутация меняет аминокислотную последовательность белка, в клетке появляется белок с измененной последовательностью (номер 1 на рисунке 3).

б) Клеточный белковый комплекс — протеосома разрезает белок на маленькие белки (пептиды), длиной 8-11 аминокислот.

в) Затем, в эндоплазматическом ретикуле полученный маленький белок может связаться с другим белковым комплексом — главным комплексом гистосовместимости (MHC) класса 1 или 2 (номер 2 на рисунке 3). Данный комплекс кодируется генами класса HLA.

г) MHC-комплекс со связанным пептидом появляется на поверхности клетки, где он презентует пептид клеткам иммунитета. В случае, если пептид является мутированным или инородным, данная клетка получает сигнал к самоликвидации. Такой уникальный пептид называется **неоантигеном**.

На рисунке 3 б) показана последовательность изучения неоантигенов с помощью данных секвенирования.

1.4. Главные комплексы гистосовместимости

Это белковые комплексы MHC, которые присутствуют на поверхностях клеток и играют важную роль во врожденном и приобретенном иммунитете. Они кодируются высоко вариабельными генами HLA. Они играют ключевую роль в вопросе приживаемости органов после пересадки, и для них разработаны клинические методы определения высокой точности [3]. Для хранения известных последовательностей генов HLA создана база IMGT [4].

1.4.1. MHC Class I

MHC Class I присутствует на поверхности всех клеток и кодируется генами HLA-A, HLA-B и HLA-C. Каждый ген может кодировать отдельный комплекс на поверхности клетки.

1.4.2. MHC Class II

MHC Class I присутствует на поверхности антиген-презентирующих клеток (например, на дендритных клетках) и кодируется парой генов:

— HLA-DRA и HLA-DRB. HLA-DRA является консервативным геном и не нуждается в предсказании.
— HLA-DQA и HLA-DQB.
— HLA-DPA и HLA-DPB. Эта пара генов является наименее изученной и не применяется для изучения неоантигенов в контексте раковых опухолей.

1.5. IC50

IC50 — это термин, обозначающий концентрацию полумаксимального ингибирования. IC50 является количественным индикатором того,

сколько нужно пептида в клетке для того, чтобы он связался с MHC комплексом. IC50 ранжируется от 0 до 50000. Когда значение меньше 500, считается, что данный пептид имеет хорошее связывание с MHC комплексом. IC50 экспериментально измерено для небольшого числа пептидов, в других случаях используются методы предсказания, основанные на анализе последовательности MHC и пептида, или нейронные сети.

1.6. Роль количества мутаций в иммунотерапии

При изучении ответа на лечение иммунотерапией доктора смотрят в том числе на характеристики раковой опухоли, такие как экспрессия генов и количество мутаций. В статье *Rizvi et al.* [5] было показано, что пациенты, имеющие большее количество мутаций, в целом лучше отвечают на лечение (см. рисунок 4, ящик с усами. DCB — пациенты с Durable Clinical Benefit, отвечающие на лекарство, и с NDB — Non-Durable Benefit).

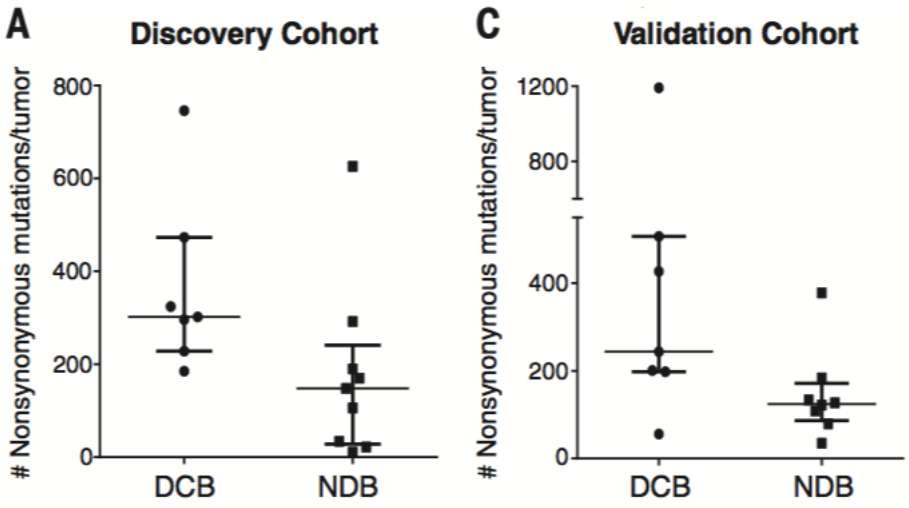


Рисунок 4 – Разница в количестве мутаций между пациентами, отвечающими и не отвечающими на лечение, среди двух когорт [5]

1.7. Цитолитическая активность как показатель иммунного ответа

На данный момент существует мало данных по пациентам, прошедшим лечение иммунотерапией, поэтому для изучения механизмов иммуного ответа используются обширные базы данных по раковым пациентам, такие как TCGA [6]. В качестве меры имунного ответа используют **цитолитическую активность** (далее - **CYT**) [7]. Она определяется через экспрессию генов GZMA и PRF1, которые экспрессируются на

клетках имуннитета, ответственных за ликвидацию других клеток. Поэтому, когда они экспрессируются в опухоли, считается, что в ней находятся иммунные клетки, ее уничтожающие. Экспрессия обоих генов представляет собой численное значение, и за CYT берется среднее геометрическое двух значений [7].

Rizvi et al. [5] указали на важность количества мутаций. *Hacohen et al* [7] исследовали пациентов из TCGA, и посмотрели на корреляцию CYT с количеством мутаций и с количеством иммуногенных мутаций: тех, которые могут произвести неоантиген, имеющий $IC50 < 500$ к MHC Class I.

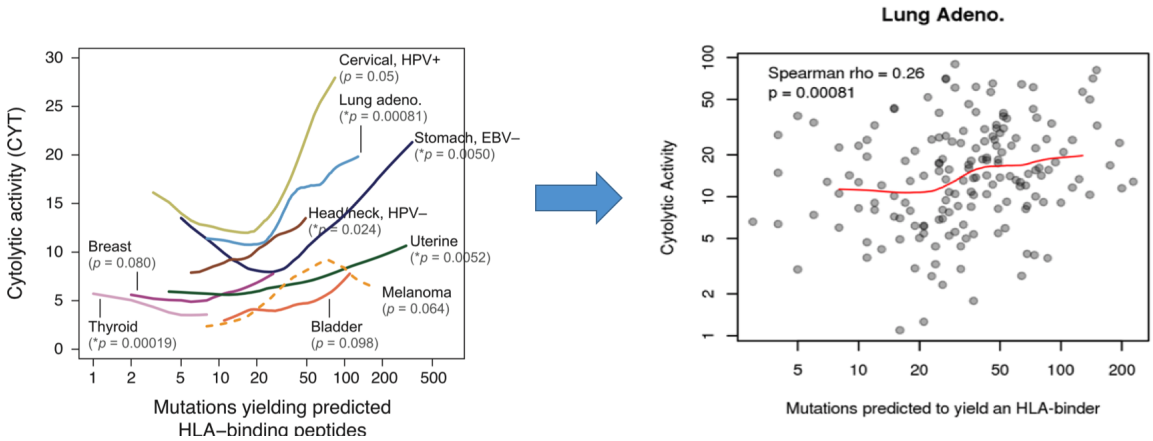


Рисунок 5 – Корреляция цитолитической активности (CYT) с количеством иммуногенных мутаций [7]

Они не рассматривали неоантигены к MHC Class II, поскольку на тот момент их роль в анти-опухолевом иммунитете не была определена. Но позже *Sahin et al.* [8] указал на их важность.

1.8. Предсказание возраста при помощи данных метилирования

Horvath S. [9] использовал данные метилирования для предсказания возраста пациента. Данные метилирования представляют собой вектор из 21 тысячи числовых значений, где каждое число соответствует уровню метилирования (от 0 до 1) определенного CpG региона (участка в ДНК).

С помощью *elastic net* ему удалось построить кросс-тканевый предиктор из 353 CpG с медианной ошибкой предсказания 3.3 года.

1.9. Elastic net

Elastic net — это метод регрессии с регуляризацией, линейно объединяющий пенальти $L1$ и $L2$.

Регрессия решает задачу линейного предсказания ответа y с помощью $x_1, x_2, ..., x_p$:

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + \ldots + x_p\hat{\beta}_p$$

Где для вычисления коэффициентов β используются пенальти:

$$||\beta||_1 = \sum_{j=1}^{p} |\beta_j|, beta||^2$$

То есть решается система:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(||y - X\beta||^2 + \lambda_1||\beta||^2 + \lambda_2||\beta||_1)$$

Коэффициенты λ_1, λ_2 влияют на силу пенальти, что выражается в разном числе выбираемых предикторов.

1.10. Предсказание MHC Class I и MHC Class II-ассоциированных неоантигенов для пациента

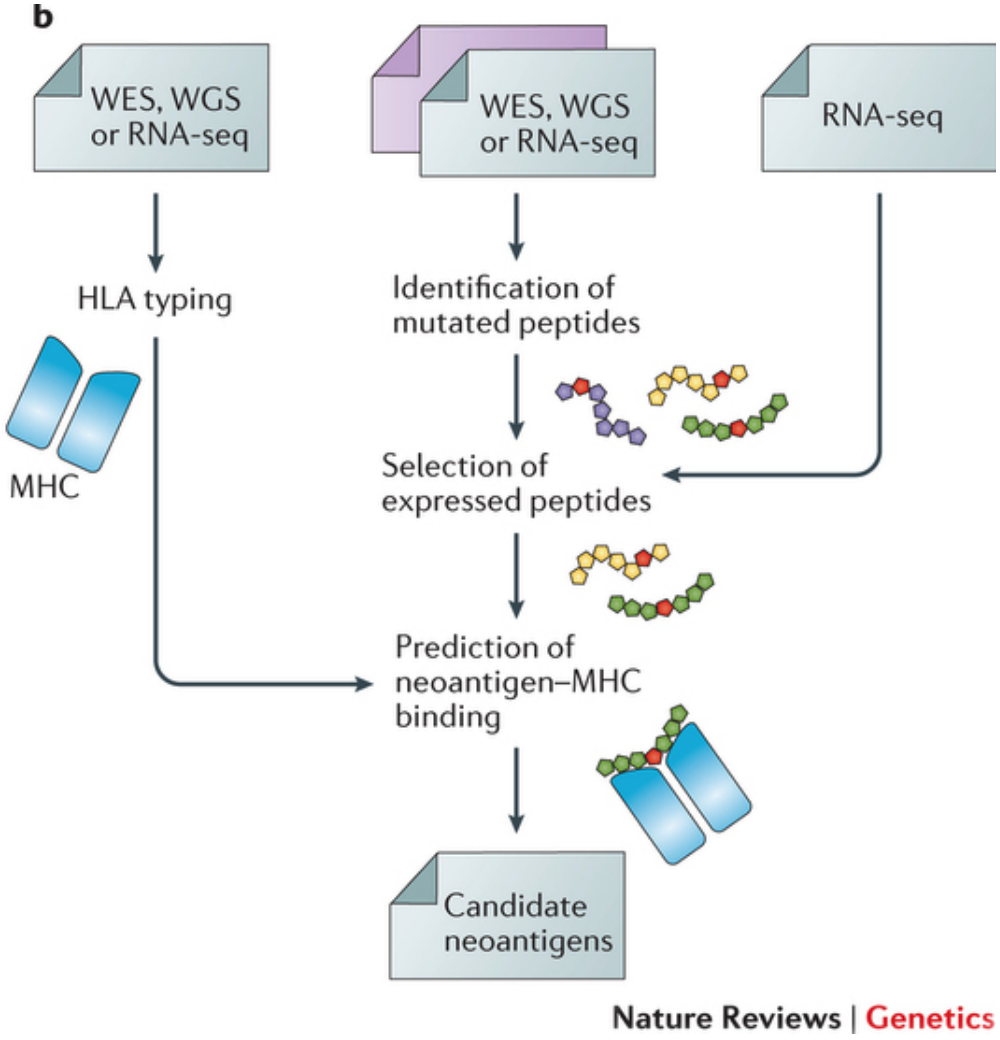


Рисунок 6 – Схема пайплайна для предсказания неоантигенов

1.10.1. Обзор существующих пайплайнов

На рисунке 7 показан список пайплайнов. Только FRED2 [10] и pVAC-seq[11] могут быть установлены локально на компьютер или кластер и обладают широкой функциональностью. Но pVAC-seq не предоставляет возможность HLA-типирования и предсказания IC50 для MHC Class II, как и FRED2. Кроме того, у них отсутствует опция работы с инделами.

Pipelines for neoantigen prediction			
FRED 2	HLA typing and T-cell epitope prediction, selection and assembly	http://github.com/FRED-2/Fred2	112
NetCTL	Prediction of immunogenic peptides by integration of proteasomal cleavage, TAP transport and pMHC-I affinity	http://www.cbs.dtu.dk/services/NetCTL	110
NetCTLpan	Pan-specific version of NetCTLpan	http://www.cbs.dtu.dk/services/NetCTLpan	149
EpiToolKit	Web-based, flexible workbench for integration of class-I HLA typing and T-cell epitope prediction and selection	http://www.epitoolkit.de	111
NetTepi	Identification of antigenic peptides based by integrating prediction of pMHC-I binding affinity and stability and T-cell propensity	http://www.cbs.dtu.dk/services/NetTepi-1.0	90
pVAC-Seq	Identification and prioritization of personalized neaontigens from mutation and expression data	http://github.com/griffithlab/pVAC-Seq	113
WAPP	Integrated prediction of pMHC-I processing and presentation: proteasomal cleavage, TAP transport and pMHC-I binding affinity	http://abi.inf.uni-tuebingen.de/Services/WAPP	87

Рисунок 7 – Существующие пайплайны для предсказания неоантигенов [2]

Таким образом, на начало 2016 года пайплайнов, осуществляющих одновременно аннотацию мутаций, типирование HLA и предсказание IC50 для MHC Class I и MHC Class II, не было.

1.10.2. Аннотация одноточечных мутаций

Для работы с мутациями необходимо знать, принадлежат ли они какому-нибудь гену, и если принадлежат, то как меняют аминокислотную последовательность. Это называется аннотацией мутаций. В случае одноточечных мутаций используется программа snpEff [12].

Для ее запуска необходимо знать версию генома, использованную для создания мутаций, и представить их в стандартном текстовом формате VCF (Variant Calling Format). Минимальные поля представлены в таблице: хромосома, позиции начала и конца мутации, нуклеотид референсного генома, нуклеотид из раковой опухоли, отличный от референсного.

Chromosome	Start	End	Reference	Mutation
1	40555169	40555169	G	T
1	149882216	149882216	C	T
3	132068776	132068776	A	G

Таблица 1 – Пример минимального VCF файла

Пример аннотированного VCF-файла:

##SnpEffVersion="3.6c (build 2014-05-20), by Pablo Cingolani"							
##SnpEffCmd="SnpEff GRCh37.74 test.vcf "							
##INFO=<ID=EFF,Number=.,Type=String,Description="Predicted effects for this variant.Format: 'Effect (Effect_Impact | Functional_Class | Codon_Change | Amino_Acid_Change | Amino_Acid_length | Gene_Name | Transcript_BioType | Gene_Coding | Transcript_ID | Exon_Rank | Genotype_Number[| RRORS | WARNINGS])'">							
1	1447408	.	G	A	.	.	EFF=INTERGENIC(MODIFIER|||||||||1),UPSTREAM(MODIFIER||123||586|ATAD3A|protein_coding|CODING|ENST00000378756|1),UPSTREAM(MODIFIER||147||634|ATAD3A|protein_coding|CODING|ENST00000378755|1),UPSTREAM(MODIFIER||502||507|ATAD3A|protein_coding|CODING|ENST00000536055|1),UPSTREAM(MODIFIER||502||571|ATAD3A|protein_coding|CODING|ENST00000339113|1|WARNING_TRANSCRIPT_NO_START_CODON)
1	5948313	.	G	T	.	.	EFF=INTRON(MODIFIER|||1426|NPHP4|protein_coding|CODING|ENST00000378156|17|1),INTRON(MODIFIER|||||NPHP4|nonsense_mediated_decay|CODING|ENST00000378169|14|1),INTRON(MODIFIER|||||NPHP4|nonsense_mediated_decay|CODING|ENST00000466897|4|1),INTRON(MODIFIER|||||NPHP4|nonsense_mediated_decay|CODING|ENST00000489180|17|1),INTRON(MODIFIER|||||NPHP4|processed_transcript|CODING|ENST00000478423|13|1)

Таблица 2 – Пример результата работы snpEff

1.10.3. Аннотация коротких вставок и удалений

Представление инделов в формате VCF по структуре совпадет с представлением одноточечных мутаций.

Chr	Start	End	Reference	Mutation
1	2090371	2090371	T	-
1	13047176	13047177	CC	-
1	21806526	21806526	-	T
1	42010489	42010491	TCT	-
1	52403072	52403072	G	-
…	…	…	…	…

Таблица 3 – Пример VCFфайла, содержащего информацию об инделах

Программа snpEff не подходит для аннотирования инделов, для них используется программа ANNOVAR [13].

1.10.4. Предсказание генов HLA

Как было сказано, HLA-типирование широко используется для пересадки органов, и для их определения разработаны лабораторные методы высокой точности. При исследовании неоантигенов клиническое типирование доступно не всегда. С 2010 года появились программы, осуществляющие типирование с помощью данных секвенирования и известных аллелей из базы данных IMGT [4]:

а) PHLAT [14],

б) seq2HLA [15],

в) ATHLATES [16],
г) HLAminer [17],
д) OptiType [18].

На данный момент, основные различия между программами составляют:

— Метод — основанный на схожести последовательностей или на нейронных сетях;
— Доступная для предсказаний база данных аллелей IMGT, использованная в продукте [4];
— Класс аллелей, доступный для предсказания: MHC Class I или Class I и Class II.

Сравнение программ показано в таблице 5.

	Последнее изменение	Данные на вход	Версия IMGT для MHC Class 1	Версия IMGT для MHC Class II
seq2HLA	2016	RNA-seq	IMGT version 3.6.0	IMGT version 3.12.0
HLAminer	2016	RNA-seq Exome-seq WGS	IMGT version 3.4.0	IMGT version 3.4.0
ATHLATES	2013	RNA-seq Exome-seq WGS	Можно добавить любую версию	Можно добавить любую версию
PHLAT	2013	RNA-seq Exome-seq WGS	Можно добавить любую версию	Можно добавить любую версию
OptiType	2014	RNA-seq Exome-seq WGS	IMGT Release 3.14.0	-

Таблица 4 – Сравнение программ для предсказания HLA аллелей

1.10.5. Пресдказание IC50

Задача предсказания состоит в том, чтобы для пептида длиной 8—11 (для MHC Class I) или 15 (для MHC Class II) аминокислот и для данной последовательности HLA-генов, составляющих комплекс, предсказать значение IC50.

Для этого разработаны программы, доступные на Immune Epitope Database Analysis Resource [19].

Для MHC Class I:

— netChop [20];
— Pickpocket [21];

— ANN [22];
— netMHCpan [23];
— SMM [24];
— Smmpmbec [25].

Для MHC Class II:
— NetMHCIIpan [26];
— SMM align [27];
— NN align [28];
— Comblib [29];
— Consensus [30].

1.11. Постановка цели и задач диссертации

Цель моей работы — использовать данные секвенирования нового поколения для исследования CYT (цитолитической активности).

Задачи:

а) Исследовать корреляцию CYT с количеством MHC Class II-иммуногенных мутаций. Для этого разработать и реализовать пайплайн, который оперирует данными ДНК и РНК секвенирования для систематического определения и фильтрации неоантигенов из репертуара мутаций раковой опухоли.
б) Применить *elastic net* для предсказания CYT на основании данных метилирования.

1.12. Выводы по главе

Для решения вопросов, связанных с восприимчивостью иммунологических лекарств, можно использовать данные секвенирования нового поколения. Анализ этих данных предполагает использование математических моделей и построения пайплайнов. CYT, мера, определяющая [7] иммунный ответ, была исследована с точки зрения MHC Class I-ассоциированных эпитопов. Исследования, рассматривающие ее в контексте MHC Class II-ассоциированных эпитопов и данных метилирования, отсутствуют.

ГЛАВА 2. ИССЛЕДОВАНИЕ ЦИТОЛОГИЧЕСКОЙ АКТИВНОСТИ В КОНТЕКСТЕ MHC CLASS II-АССОЦИРОВАННЫХ НЕОАНТИГЕНОВ

Поскольку на данный момент нет готовых пайплайнов, объединяющих в себя все этапы предсказания неоантигенов по репертуару мутаций пациента относительно его HLA, был разработан пайплайн. Были использованы языки *R*, *python 2.7* и *bash*. Пайплайн реализован на кластере WashU [31] и доступен для использования всем сотрудникам лаборатории Максима Артемова.

2.1. Построение пайплайна для предсказания иммуногенности неоантигенов

2.1.1. Входные данные

Для корректного предсказания неоантигенов необходимы следующие данные:

а) Репертуар мутаций в формате VCF.
б) Секвенирование экзома нормальной ткани и раковой опухоли.
в) (Опционально) Секвенирование РНК.
г) (Опционально) Типированные HLA-гены.

2.1.2. Требования к пайплайну

Функциональность:

а) Типирование HLA-генов.
б) Предсказание IC50 для MHC Class I и MHC Class II.
в) В случае одноточечных мутаций, для каждого неоантигена указание белка, из которого он был получен, и предсказание для последнего тех же характеристик, что и для неоантигенов, для последующего сравнения.

Схема пайплайна изображена на рисунке 8.

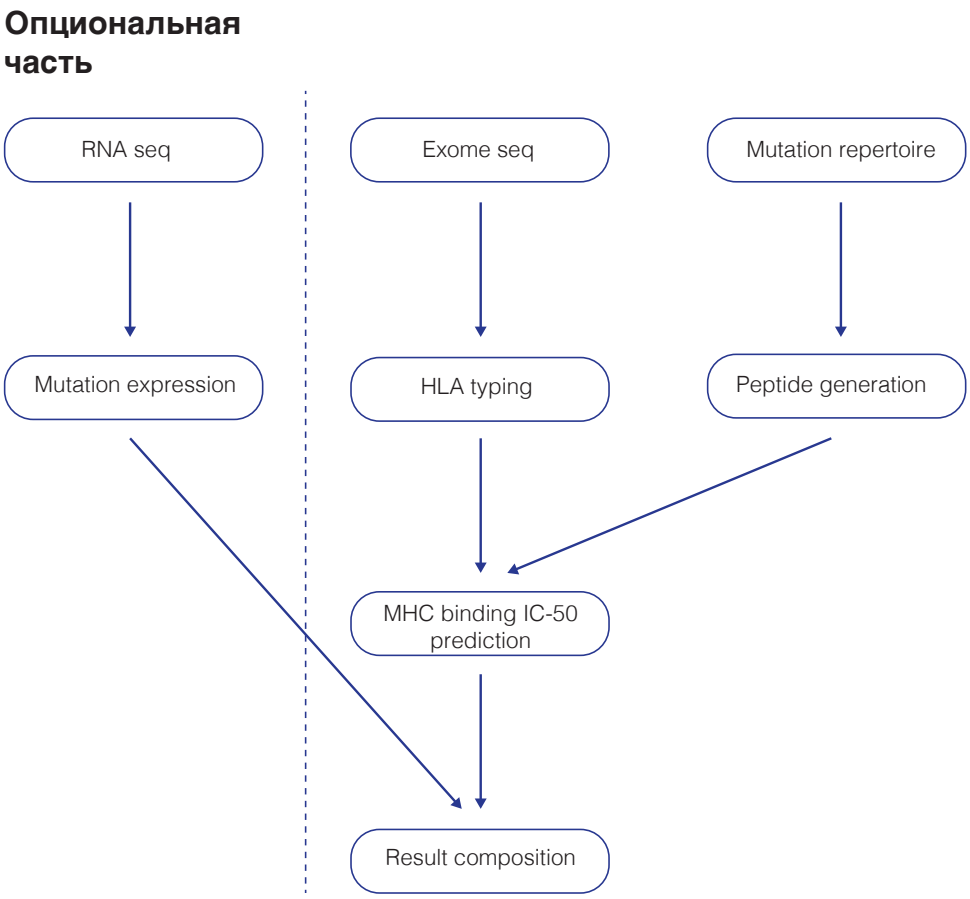


Рисунок 8 – Схема пайплайна

2.1.3. Типирование HLA

Как было указано в обзоре, существуют следующие программы для типирования:

а) PHLAT [14],
б) seq2HLA [15],
в) ATHLATES [16],
г) HLAminer [17],
д) OptiType [18].

В пайплайне необходимо знать последовательности генов из MHC Class II, поэтому программа OptiType не подходит, так как предсказывает только MHC Class I.

Для определения наилучшего инструмента были запущены оставшиеся программы.

2.1.3.1. HLAminer

После запуска HLAminer оказалось, что в выходных данных возможен случай предоставления более двух аллелей на ген. Поскольку боль-

ше двух аллелей быть не может, для последующего анализа необходимо выбрать, какие аллели использовать. Статистики, предоставляемые HLAminer'ом, не всегда позволяют явно это сделать. Поэтому этот продукт не использовался для анализа.

2.1.3.2. ATHLATES

При запуске ATHLATES оказалось, что в случае малого покрытия экзома ATHLATES может выдать пустой файл. Многие пациенты отсеквенированы с малыми покрытиями, поэтому ATHLATES не использовался для анализа.

2.1.3.3. Сравнение seq2HLA и PHLAT

Программы seq2HLA и PHLAT работают и с данными, имеющими малое покрытие, и представляют данные на выходе в виде двух аллелей на ген.

Чтобы выбрать между ними, было проведено сравнение на данных, для которых было осуществлено клиническое типирование. Полноэкзомное секвенирование было получено с проекта HapMap [32]. Типирование для пациентов было проведено в нескольких лабораториях с доступными для скачивания результатами. Я запустила seq2HLA и PHLAT на 26 пациентах для трех генов MHC Class I (HLA-A, HLA-B и HLA-C) и трех генов MHC Class II (HLA-DRB1, HLA-DQA1, HLA-DRB1): всего 12 предсказываемых значений. Точность типирования для каждого гена определялась как количество правильно определенных аллелей ко всем аллелям.

	PHLAT accuracy	seq2HLA accuracy
HLA-A	0.95	0.96
HLA-B	0.96	0.92
HLA-C	0.95	0.84
HLA-DRB1	0.85	0.84
HLA-DQA1	0.78	0.64
HLA-DQB1	0.94	0.84

Рисунок 9 – Сравнение исполнения PHLAT и seq2HLA на клинических данных

Для пайплайна был выбран PHLAT, поскольку для пяти генов из шести он показал большую точность, чем seq2HLA. PHLAT работает с fastq файлами. В случае, если предоставлен BAM файл, fastq извлекаются из него программой SamToFastq из набора Picard tools.

PHLAT предсказывает одну (гомозиготный случай) или две (гетерозиготный случай) аллели на каждый ген и выдает p-value для результата.

Locus	Allele1	Allele2	LLtot	pval1	pval2
HLA_A	A*24:02:01	A*26:01:01	-48986.52	1.5E-02	4.0E-03
HLA_B	B*38:01:01	B*40:02:01	-41052.91	5.7E-04	7.8E-03
HLA_C	C*03:05	C*12:03:01	-30099.56	4.9E-04	5.8E-03
HLA_DQA1	DQA1*03:01:01	DQA1*03:01:01	-93.18	1E-01	1E-01
HLA_DQB1	DQB1*03:02:01	DQB1*03:02:01	-231.77	4.6E-03	4.6E-03
HLA_DRB1	DRB1*04:02:01	DRB1*04:07:01	-12710.26	6.8E-06	1.1E-09

Рисунок 10 – Пример результата работы PHLAT на данных полноэкзомного секвенирования

2.1.4. Реализация типирования на сервере

Типирование является ресурсоемкой задачей, так как выравнивает предоставленные файлы на геном, поэтому реализация осуществлена на сервере. Для типирования созданы следующие скрипты:

а) **type_HLA_from_fastq.sh** для случая, когда предоставлены *fastq* файлы. Должно быть представлено 2 файла с именами *fastq_prefix.1.fastq* и *fastq_prefix.2.fastq* в одной папке (в примере путь представлен как *path_to_wdir*). Запуск скрипта выглядит следующим образом:

Listing 1 – Запуск PHLAT для fastq на кластере

```
type_HLA_from_fastq.sh  fastq_prefix  path_to_wdir
```

Результат типирования будет лежать в *path_to_wdir*, с названием *fastq_prefix.sum*.

б) **type_HLA_from_bam.sh** для случая, когда предоставлен BAM файл. Запуск скрипта выглядит следующим образом:
Результат типирования будет лежать в *path_to_wdir*, с названием *BAM_prefix.sum*.

Listing 2 – Запуск PHLAT для BAM на кластере

```
type_HLA_from_bam.sh  path_to_wdir/bam_prefix.bam
```

в) **process_PHLAT_results.R**. Поскольку может быть предоставлено несколько разных файлов полноэкзомного секвенирования (из нормальных тканей и из опухолевых), и PHLAT умеет работать с данными РНК-секвенирования, в результате может иметься несколько фалов с HLA-аллелями. В связи с недостаточной глубиной секвенирования результаты в этих файлах могут отличаться. Для автоматического сравнения можно воспользоваться указанным скриптом:

Listing 3 – Автоматическое сравнение файлов с типированием для пациента

```
process_PHLAT_results.R path_to_wdir name
```

Где *fpath_to_results* — папка, содержащая все *sum файлы, name — желаемое имя для файла с конечным результатом. В случае, если все запуски имеют одинаковый результат, конечный файл будет создан, в противном случае сводная таблица с результатами будет сохранена в файл *fname_original_typing_unsucc.txt*. С его помощью в дальнейшем можно вручную определить аллели для пациента.

Ввиду множества факторов: покрытия и качества секвенирования, качества образца, типа секвенирования, доступных файлов — невозможно заранее предопределить, какой алгоритм выбора аллелей в случае с несовпадением является оптимальным.

По умолчанию PHLAT запускается в четыре потока на 48ГБ оперативной памяти с ограничением на 16 часов. Использование четырех потоков обусловлено тем, что PHLAT использует программу **bowtie2** для выравнивания последовательностей на базу данных аллелей. А такой объем оперативной памяти может быть необходим при конвертации BAM в fastq.

2.1.4.1. Предоставленные данные для типирования

В случае, если данные типирования присутствуют, они должны быть представлены в том же виде, как в таблице 5.

Аллели MHC Class I	Аллели MHC Class II
HLA-A*33:01	HLA-DQA1*05:05/HLA-DQB1*03:01
HLA-A*33:03	HLA-DRB1*11:04
HLA-B*14:02	HLA-DRB1*12:01
HLA-B*49:01	
HLA-C*07:01	
HLA-C*08:02	

Таблица 5 – Образец представления HLA аллелей для пайплайна

2.1.5. Аннотация VCF файла

На данный момент, созданный пайплайн работает со следующими версиями геномов:

а) Homo Sapiens:
   — сборка hg19, релизы 74 (GRCh37.74) и 75 (GRCh37.75);
   — сборка hg38, релиз 20 (GRCh38.p0) и 25 (GRCh38.p7).

б) Mus Musculus: сборка mm9, релиз 67 (NCBIM37.67)

Для аннотации VCF файла с помощью программы snpEFF был создан скрипт **annotate_mutations.sh**. Запуск:

Listing 4 – Запуск snpEFF для аннотации и обработка его результата

```
annotate_mutations.sh name path_to_wdir genome_version
```

После запуска скрипта в рабочей директории *path_to_wdir* будет создана папка misc, содержащая в себе файлы аннотации:

а) *name.snpeff.vcf* — оригинальная аннотация от snpEff;

##SnpEffVersion="3.6c (build 2014-05-20), by Pablo Cingolani"							
##SnpEffCmd="SnpEff GRCh37.74 test.vcf "							
##INFO=<ID=EFF,Number=.,Type=String,Description="Predicted effects for this variant.Format: 'Effect (Effect_Impact | Functional_Class | Codon_Change | Amino_Acid_Change | Amino_Acid_length | Gene_Name | Transcript_BioType | Gene_Coding | Transcript_ID | Exon_Rank | Genotype_Number[| RRORS | WARNINGS])'">							
1	1447408	.	G	A	.	.	EFF=INTERGENIC(MODIFIER|||||||||1),UPSTREAM(MODIFIER||123||586|ATAD3A|protein_coding|CODING|ENST00000378756|1),UPSTREAM(MODIFIER||147||634|ATAD3A|protein_coding|CODING|ENST00000378755|1),UPSTREAM(MODIFIER||502||507|ATAD3A|protein_coding|CODING|ENST00000536055|1),UPSTREAM(MODIFIER||502||571|ATAD3A|protein_coding|CODING|ENST00000339113|1|WARNING_TRANSCRIPT_NO_START_CODON)
1	5948313	.	G	T	.	.	EFF=INTRON(MODIFIER|||1426|NPHP4|protein_coding|CODING|ENST00000378156|17|1),INTRON(MODIFIER||||NPHP4|nonsense_mediated_decay|CODING|ENST00000378169|14|1),INTRON(MODIFIER||||NPHP4|nonsense_mediated_decay|CODING|ENST00000466897|4|1),INTRON(MODIFIER||||NPHP4|nonsense_mediated_decay|CODING|ENST00000489180|17|1),INTRON(MODIFIER||||NPHP4|processed_transcript|CODING|ENST00000478423|13|1)

Таблица 6 – Пример результата работы snpEff

б) *name.for_pull* — отформатированная аннотация для создания белковых последовательностей;

gene_name	protein_change	transcript_id
PRAMEF6	p.M201I	ENST00000355096
FBLIM1	p.A308V	ENST00000441801
LGR6	p.R371G	ENST00000439764
ATL2	p.A309V	ENST00000402054
…	…	…

Таблица 7 – Аннотация белковых последовательностей

в) *name.for_pull.full* — отформатированная аннотация с полным описанием мутации, которая будет использована для формирования файла с результатами.

1	13001080	C	G	PRAMEF6	ENST00000355096	p.M201I
1	16101324	C	T	FBLIM1	ENST00000441801	p.A308V
1	202279446	A	G	LGR6	ENST00000439764	p.R371G
2	38525479	G	A	ATL2	ENST00000402054	p.A309V
…	…	…	…	…	…	…

Таблица 8 – Полная аннотация белковых последовательностей

2.1.6. Создание неоантигенов для одноточечных мутаций

Аннотированный VCF будет использован для создания неоантигенов. Поскольку для одноточечных мутаций необходимо знать не только неоантиген, но и соответствующий ему не-мутированный белок, будет создано два файла.

Создание неоантигенов выполняет скрипт **mutations_to_peptides.sh**.

Механизм его работы:

а) Он достает белковые последовательности для генов, в которые попали мутации. Имена генов известны из аннотации VCF, последовательности находятся в аннотации генома, которая сохранена для геномов, с которыми работает пайплайн.

б) Он находит место, которое меняет мутация в последовательности, и сохраняет ее окрестность (по умолчанию 14 аминокислот). В один файл идет мутированная последовательность, в другой — оригинальная.
Запуск скрипта:

Listing 5 – Создание неоантигенов и соответсвующих им немутированных белков

```
mutations_to_peptides.sh name species genome_version
  path_to_wdir
```

Данный скрипт использует файл *name.for_pull*, находящейся в *misc*. После запуска будут созданы следующие файлы:

а) *name.wt.fasta*, *name.mut.fasta* — мутированные и референсные пептиды. Соответствие между ними определяется уникальным ID. Поскольку одному гену может соответствовать несколько транскриптов, одинаковых в окрестности мутации, производится поиск одинаковых пептидов, и найденные удаляются.

Original peptide	Mutated peptide
>l1_PRAMEF6 GMPFRNIRSILKMVNLDCIQEVEVN	>l1_PRAMEF6 GMPFRNIRSILKIVNLDCIQEVEVN
>l4_FBLIM1 YRYEKGLCTGWGAGTGRDPSRVKEL	>l4_FBLIM1 YRYEKGLCTGWGVGTGRDPSRVKEL
>l5_LGR6 EDLHLDDEESSKRPLGLLARQAENH	>l5_LGR6 EDLHLDDEESSKGPLGLLARQAENH
>l10_ATL2 IFYAARTPATLFAVMFAMYIISGLT	>l10_ATL2 IFYAARTPATLFVVMFAMYIISGLT
...	...

Таблица 9 – Пример мутированных и референсных белков, ассоциированных с мутацией

Эти файлы будут использоваться как входные данные для программ предсказания IC50.

б) *name.final.ann* — файл аннотации пептидов, будет необходим на этапе представления результатов для восстановления информации о том, с какой позиции какого транскрипта пришел ген.

ID	Transcript	Reference peptide	Mutated peptide	Mutation
l1_PRAMEF6	ENST00000355096	GMPFRNIRSILKMVNLDCIQEVEVN	GMPFRNIRSILKIVNLDCIQEVEVN	p.M201I
l4_FBLIM1	ENST00000441801	YRYEKGLCTGWGAGTGRDPSRVKEL	YRYEKGLCTGWGVGTGRDPSRVKEL	p.A308V
l5_LGR6	ENST00000439764	EDLHLDDEESSKRPLGLLARQAENH	EDLHLDDEESSKGPLGLLARQAENH	p.R371G
l10_ATL2	ENST00000406122	IFYAARTPATLFAVMFAMYIISGLT	IFYAARTPATLFVVMFAMYIISGLT	p.A309V

Таблица 10 – Аннотация белков, ассоциированных с мутацией

2.1.7. Предсказание IC50

Имея файлы с референсными и мутированными пептидами, зная HLA-гены для пациента, теперь можно запустить предсказание IC50. MHC Class I и MHC Class II IC50 предсказываются отдельно, так как для них используются пептиды разного размера и разные программы.

2.1.7.1. Предсказание для MHC Class I

Для предсказания IC50 с MHC Class I было написано несколько скриптов. В целом, в них происходит запуск разных программ и обработка промежуточных и конечных результатов.

Запуск основного скрипта показан на примере предсказания для аллели HLA-A*01:01.

Listing 6 – Предсказание IC50 для MHC Class I, аллели HLA-A*01:01

```
echo "HLA-A*01:01" | while read line;
do
    TAG=name_${line//[\*\:-\/]/}; TAG=${TAG//_/\.};
    qsub mhc_i.sh -N $TAG -l
    nodes=1:ppn=1,walltime=4:00:00,vmem=16gb
    -o log.$TAG -j oe -F "name path_to_wdir $line 1";
done
```

После этого в рабочей директории для каждой аллели будет создана папка, в которой будут находиться результаты.

а) *name.wt.table.[8, 9, 10, 11].HLA-A01:01* и *name.mut.table.[8, 9, 10, 11].HLA-A01:01*:

id	wt_peptide	mut_peptide	neoepitope_ ratio	wt_mean_ IC50	mut_mean_ IC50	wt_median_ IC50	mut_median_ IC50	wt_netchop_ left_score	wt_netchop_ right_score	wt_pickpocket_ IC50	wt_ann_ IC50	...
l10_ATL2	AARTPATLFA	AARTPATLFV	1.09	29961.35	31797.41	36426.50	33524.10	0.88	0.14	50000.00	42965.38	...
l10_ATL2	ARTPATLFAV	ARTPATLFVV	0.83	47479.89	82988.59	41352.33	50000.00	0.67	0.41	50000.00	36473.40	...
l10_ATL2	RTPATLFAVM	RTPATLFVVM	0.98	29283.20	31946.47	26702.20	27368.20	0.61	0.71	13797.25	35698.03	...
l10_ATL2	TPATLFAVMF	TPATLFVVMF	1.24	52271.91	52413.48	40708.92	32787.64	0.10	0.77	40708.92	27516.11	...

Рисунок 11 – Пример предсказания IC50 для MHC Class I

Каждый результат представляет сводную таблицу со следующими полями:

1) id неоантигена;
2) оригинальный белок;
3) мутированный белок;
4) среднее предсказанное IC50 для оригинального и мутированного белка;
5) медианное предсказанное IC50;
6) Столбцы с IC50, предсказанные каждой программой по отдельности.

б) *name.HLA-A01:01.out*: объединение всех файлов из предыдущего пункта.

2.1.7.2. Предсказание для MHC Class II

Предсказание IC50 для MHC Class II происходит сходно с MHC Class I.

Listing 7 – Предсказание IC50 для MHC Class II, аллели HLA-DRB1*01:01

```
echo "HLA-DRB1*01:01" | while read line;
do
TAG=name_${line//[\*\:-\/]/}; TAG=${TAG//_/\.};
qsub  mhc_ii.sh -N $TAG -l
nodes=1:ppn=1,walltime=4:00:00,vmem=16gb -o log.$TAG
-j oe  -F "name path_to_wdir 1"; done
```

После этого в рабочей директории для каждой аллели будет создана папка, в которой будет находиться следущие файлы:

а) *name.HLA-DRB101:01.pep.ann*:

Peptide for prediction	Original peptide
AAAEIREQGDGAEDE	SKGAAAEIREQGDGAEDEEWDD
AAAEIREQGDGAEDE	SKGAAAEIREQGDGAEDEEWPEEIQ
AAAEIREQGDGAEDE	SKGAAAEIREQGDGAEDEEWVFVET
AADTAAQISKRKCEA	NEDLRSWTAADTAAQISKRKCEAAN
AAEIREQGDGAEDEE	SKGAAAEIREQGDGAEDEEWDD
AANAVKAGGTDHRKP	EAPAANAVKAGGTDHRKPLISPQTS
AARTPATLFAVMFAM	IFYAARTPATLFAVMFAMYIISGLT
…	…

Рисунок 12 – Пример аннотированных пептидов для предсказания

б) *name.HLA-DRB101:01.out* — сводная таблица с результатами, аналогичными результатам для MHC Class I.

wt_peptide	mut_peptide	wt_median_aff	mut_median_aff	neoepitope_ratio	wt_NetMHCIIpan_IC50	wt_smm_align_IC50	wt_nn_align_IC50	wt_comblib_IC50	wt_consensus3_rank	mut_NetMHCIIpan_IC50	mut_smm_align_IC50	mut_nn_align_IC50	mut_comblib_IC50	mut_consensus3_rank
IFYAARTPATLFAVM	IFYAARTPATLFVVM	0.0276	0.0295	1.0714	18.26	2951.0	NA	NA	31.43	17.05	2580.0	NA	NA	28.46
FYAARTPATLFAVMF	FYAARTPATLFVVMF	0.0039	0.0045	1.1523	131.51	3765.0	NA	NA	37.16	113.78	3578.0	NA	NA	35.92
YAARTPATLFAVMFA	YAARTPATLFVVMFA	0.0016	0.0017	1.0903	336.96	4597.0	NA	NA	42.02	306.95	4649.0	NA	NA	42.29
AARTPATLFAVMFAM	AARTPATLFVVMFAM	0.0013	0.0012	0.9015	410.19	4937.0	NA	NA	43.77	451.32	6070.0	NA	NA	48.9
ARTPATLFAVMFAMY	ARTPATLFVVMFAMY	0.0036	0.0026	0.7197	184.05	571.0	NA	NA	6.03	244.39	927.0	NA	NA	10.9

Рисунок 13 – Пример результата предсказания IC50 для MHC Class II

2.1.8. Формирование результата

В результате нужно получить единую таблицу для MHC Class I и для MHC Class II. Для оперирования результатами был написан скрипт **process_snv_prediction.sh**. Они использует файлы с аннотациями, созданные на прошлых шагах, такие как:

а) *name.for_pull.full*;
б) *name.final.ann*;
в) *name.allele.pep.ann* (для MHC Class II).

Запуск скрипта:

Listing 8 – Пример запуска скрипта для форматирования результатов

```
process_snv_prediction.sh name path_to_wdir
```

После запуска в рабочей директории будут созданы следующие файлы:

а) *name_snv_class1.txt* и *name_snv_class2.txt*. Это наиболее полные таблицы, содержащие следующие информацию:
   1) Каждую мутацию, меняющая последовательность белка;
   2) Для каждой вышеуказанной мутации — все возможные неоантигены, имеющие потенциал связаться с MHC Class I или MHC Class II;
   3) Для каждого неоантигена указан соответствующий ему референсный белок;
   4) Для пары неоэпитоп-референсный белок, для каждой предсказанной аллели HLA указаны:
      — IC50, предсказанные выбранными программами;
      — Посчитано среднее и медианное IC-50;

— Подсчитано отношение между средними IC-50, что позволяет оценить разницу в эффективности взаимодействия между мутированнным и референсным белком.

б) *name_snv_class1_by_mutation.txt* и *name_snv_class2_by_mutation.txt*. Это наиболее сокращенные таблицы, показывающие лучший неоантиген для каждой мутации и его характеристики. Такая таблица дает понять, есть ли потенциальные неоантигены для вакцин, каково их количество и качество.

chr	pos	ref	alt	gene symbol	allele	mut_peptide	mut_median_IC50	wt_peptide	wt_median_IC50	neoepitope_ratio	transcript	mut	mut statistics...
1	16101324	C	T	FBLIM1	HLA-A33:01	TGWGVGTGR	543.50	TGWGAGTGR	1331.07	2.45	ENST00000441801	p.A308V	...
1	202279446	A	G	LGR6	HLA-C08:02	ESSKGPLGL	1062.03	ESSKRPLGL	1006.10	0.95	ENST00000439764	p.R371G	...
1	13001080	C	G	PRAMEF6	HLA-C07:01	IRSILKIVN	1437.83	IRSILKMVN	1276.50	0.89	ENST00000355096	p.M201I	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Рисунок 14 – Пример сокращенной таблицы с лучшим неоантигеном для каждой мутации

в) *name_snv_class1_full_table.txt* и *name_snv_class2_full_table.txt*. Это таблицы такого же формата, как в предыдущем пункте, но содержащие все комбинации неоантиген-аллель.

2.1.9. Доработка пайплайна для поиска неоантигенов, образованных инделами

Для учета инделов в пайплайне нужно принять во внимание отличие их от точечных мутаций. Особенности работы с ними заключаются в:

а) Другой механизм создания мутированных последовательностей;

б) Отсутствие референсного белка для сравнения эффективности присоединения к MHC комплексу. Поскольку начиная с позиции мутации аминокислотная последовательность может быть полностью изменена, то референсного белка, отличающегося от потенциального неоантигена и присутствующего в здоровых тканях, нет.

Для возможности работы с инделами пайплайн был модернизирован, добавлена возможность извлекать неоантигены для инделов, и изменена обработка данных предсказания для работы без референсных белков.

2.1.9.1. Извлечение мутированных белковых последовательностей

Формат представления инделов в vcf по структуре совпадает с представлением одноточечных мутаций.

Chr	Start	End	Reference	Mutation
1	2090371	2090371	T	-
1	13047176	13047177	CC	-
1	21806526	21806526	-	T
1	42010489	42010491	TCT	-
1	52403072	52403072	G	-
...	...	...	...	...

Рисунок 15 – Пример vcf файла, содержащего информацию об инделах

Потенциально разница между точечными мутациями и инделами состоит в том, что с индела можно получить больше неоантигенов. Например, при точечной мутации и пептиде длины N можно получить максимум N различных пептидов. При инделах количество новых неоантигенов определяется длиной измененной последовательности.

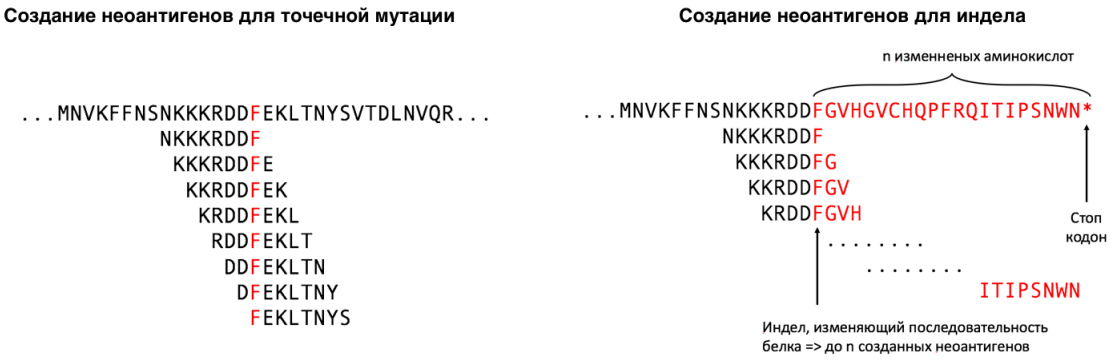


Рисунок 16 – Разница в генерации неоантигенов для точечной мутации и для инделов

В качестве программы для аннотации был использован ANNOVAR [13]. Был написан скрипт, который запускает программу и форматирует результат ее выхода: **annotate_indels.sh**

Запуск скрипта для аннотации:

Listing 9 – Пример запуска скрипта для аннотации инделов

```
annotate_indels.sh path_to_wdir/indel.vcf genome_version
```

2.1.9.2. Предсказание IC50 в случае инделов

Аналогично случаю с одноточечными мутациями, был написан скрипт, выполняющий запуск программ и обработку результатов. Разница состоит в том, что в данном случае предсказание идет только для мутированных пептидов.

Listing 10 – Предсказание IC50 для инделов, MHC Class I

```
main_indel_class1.sh name_indel path_to_wdir typing.txt
```

Вид результата для MHC Class I:

peptide	pickpocket_ IC50	ann_IC50	netmhcpan_ IC50	smm_IC50	smmpmbec_ IC50
GRSGTGPSQ	8758.62	31100.00	45745.08	409.41	444.65
PSQQPITVG	10992.95	34464.00	41896.67	937.91	816.62
SGTGPSQQP	6330.92	38812.00	44529.11	1607.53	837.57
TGPSQQPIT	16404.99	19423.00	42390.48	3918.86	7925.38
QIWDTAGRS	18882.64	31095.00	43145.19	2421.92	5321.33
RSGTGPSQQ	3343.70	34914.00	39330.80	3918.86	1986.19
…	…	…	…	…	…

Рисунок 17 – Результат предсказания для MHC Class I

Listing 11 – Предсказание IC50 для инделов, MHC Class II

```
main_indel_class2.sh name_indel path_to_wdir typing.txt
```

Вид результата для MHC Class II:

peptide	NetMHCIIpan_ IC50	smm_align_ IC50	nn_align_ IC50	comblib_ IC50	consensus3_ rank
LQIWDTAGRSGTGPS	1301.26	NA	NA	NA	13.32
RSGTGPSQQPITVGP	7087.08	NA	NA	NA	34.99
GTGPSQQPITVGPWA	4262.76	NA	NA	NA	34.99
KLQIWDTAGRSGTGP	490.17	NA	NA	NA	13.32
SGTGPSQQPITVGPW	5772.59	NA	NA	NA	34.99
…	…	…	…	…	…

Рисунок 18 – Результат предсказания для MHC Class II

2.1.9.3. Форматирование результата для инделов

Как и в случае с одноточеными мутациями, все предсказания объединяются в один файл с сохранением информации об аллелях, и ге-

нерирую файл с лучшим неоантигеном на индел. Для форматирования написан скрипт **main_indel_post_processing.sh**.

Запуск:

Listing 12 – Форматирование результатов предсказания для инделов

```
main_indel_post_processing.sh  name  path_to_wdir
```

Пример полной итоговой таблицы для MHC Class I:

transcript_name	chr	start	end	ref	alt	gene_symbol	mut_exon	change	peptide	allele	median_IC50
NM_080863	17	42248186	42248186	C	-	ASB16	exon1	c.29delC	AGARQVPG	HLA-A32:01	37473.45
NM_001267716	19	23158289	23158289	C	-	ZNF728	exon4	c.1850delG	THKRIHTGEK	HLA-A32:01	37543
NM_080863	17	42248186	42248186	C	-	ASB16	exon1	c.29delC	GWNGRTGGGRL	HLA-A32:01	37625
NM_002867	1	52403072	52403072	G	-	RAB3B	exon3	c.241delC	TGPSQQPITV	HLA-A32:01	37749.5
NM_080863	17	42248186	42248186	C	-	ASB16	exon1	c.29delC	RQVPGPDS	HLA-A32:01	37797
…	…	…	…	…	…	…	…	…	…	…	…

Таблица 11 – Итоговая таблица MHC Class I

2.2. Применение пайплайна для характеризации CYT

Для исследования были рассмотрены 308 пациентов с аденокарциномой легких (LUAD) из базы данных TCGA [6]. Для каждого пациента присутствовали данные:

а) Полноэкзомное секвенирование нормальной ткани.
б) Полноэкзомное секвенирование раковой опухоли.
в) Репертуар мутаций раковой опухоли.
г) РНК-секвенирование раковой опухоли.

Пайплайн был запущен для предсказания неоантигенов из точечных мутаций и инделов.

В статье *Hacohen et al.* [7] на рассматриваемых данных было показано, что мутации, порождающие неоантигены MHC Class I положительно коррелируют с CYT.

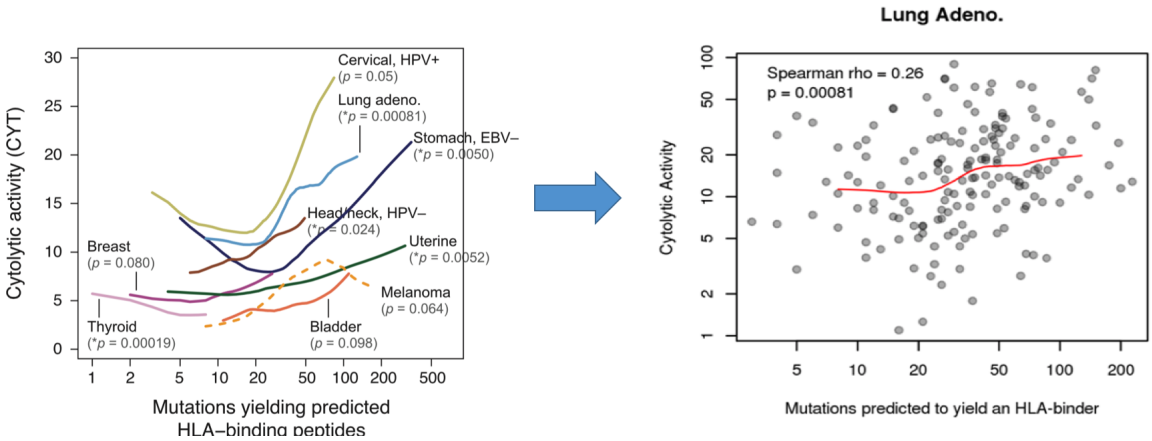


Рисунок 19 – Пропорция мутаций, порождающих эффективный антиген для MHC Class II

2.2.1. Сравнение Class I и Class II иммуногенности мутаций

Было исследована разница в вероятности порождения мутацией неоантигена с IC50 < 500 для MHC Class I и MHC Class II. Для исследованых 308 пациентов было подсчитано, какая пропорция мутаций несет в себе сильные неоантигены.

Class I:

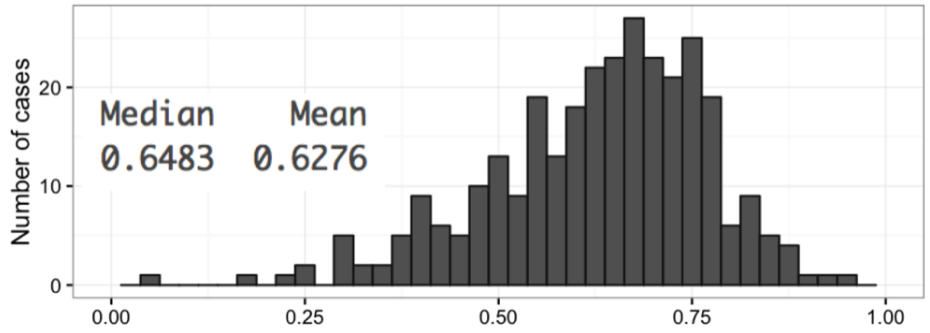


Рисунок 20 – Пропорция мутаций, порждающий эффективный антиген для MHC Class I

Class II:

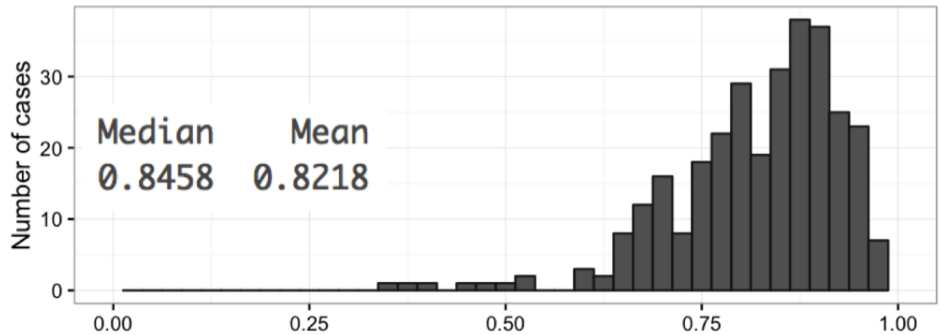


Рисунок 21 – Пропорция мутаций, порждающий эффективный антиген для MHC Class II

В соответствии с данными, каждая мутация порождает сильный эпитоп для Class I в среднем в 60% случаев, для Class II — в 80% случаев.

На самом деле, наиболее интересным является эффективный неоантиген, представленный в клетке. Поскольку для пациентов были доступны данные РНК-секвенирования, были подсчитаны отношения экспрессирующихся мутаций с эффективным неоантигеном по отношению ко всем мутациям.

Class I:

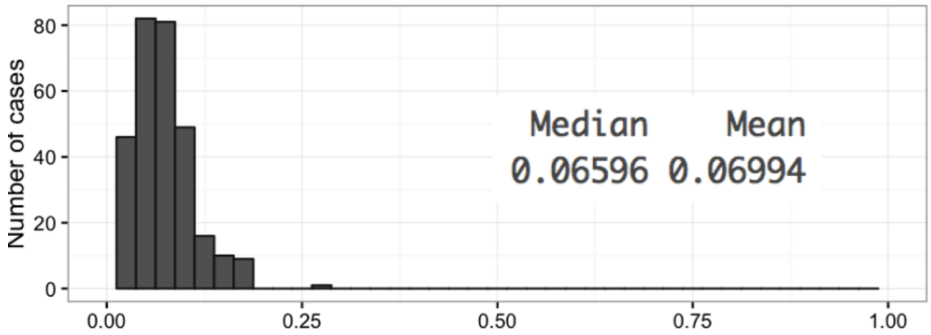


Рисунок 22 – Пропорция мутаций, порждающий эффективный антиген для MHC Class I

Class II:

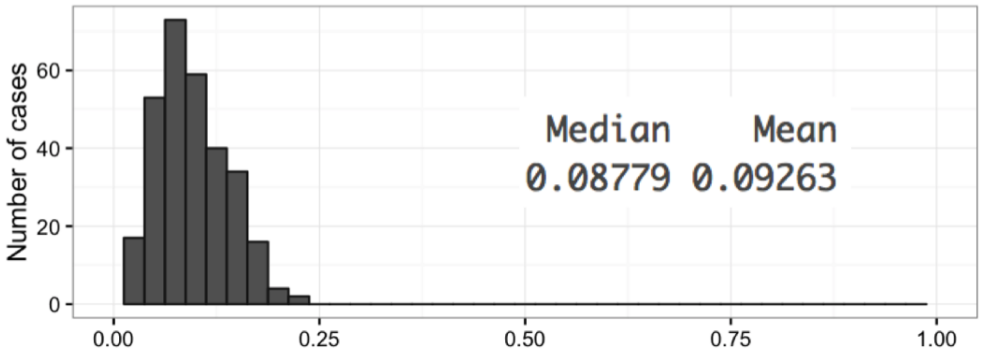


Рисунок 23 – Пропорция мутаций, порждающий эффективный антиген для MHC Class II

Как видно из графика, вероятность получить эффективный MHC Class II-ассоциированный неоантиген, присутствующий в клетках, выше, чем вероятность получить MHC Class I-ассоциированный неоантиген.

Затем, была подсчитана корреляция количества мутаций с количеством мутаций, порождающих сильные антигены. Оказалось, что в случае обоих классов она выше 0.9:

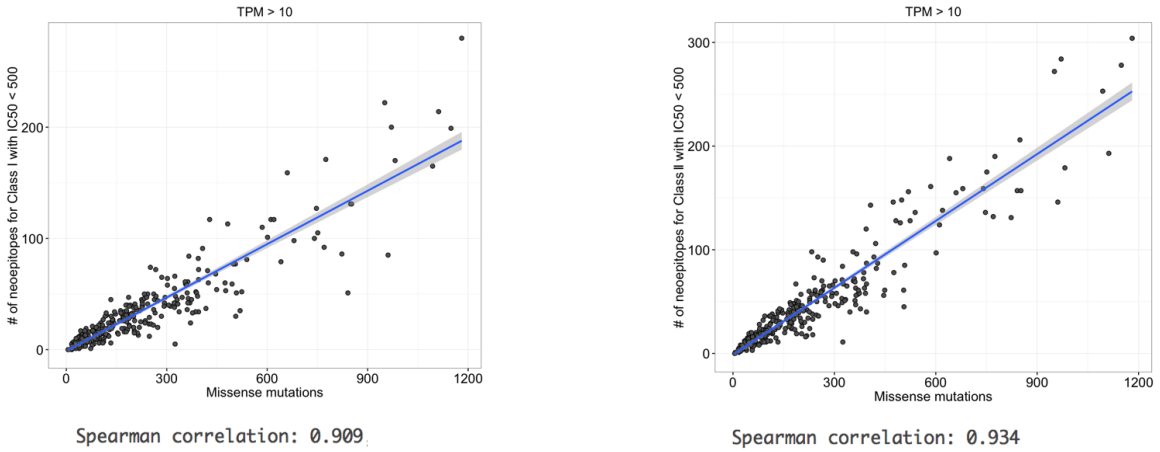


Рисунок 24 – Пропорция мутаций, порождающий эффективный антиген для MHC Class II

Здесь по оси X отложено количество мутаций, меняющих последовательность белка, по оси Y — количество мутаций, порождающих сильный неоантиген для MHC Class I (слева) и MHC Class I (справа).

2.2.2. Сравнение корреляций с количеством иммуногенных мутаций двух классов MHC

Применив свой пайплайн, стало понятно, что требование для мутации порождения неоантигена для MHC Class II не улучшает корреляцию:

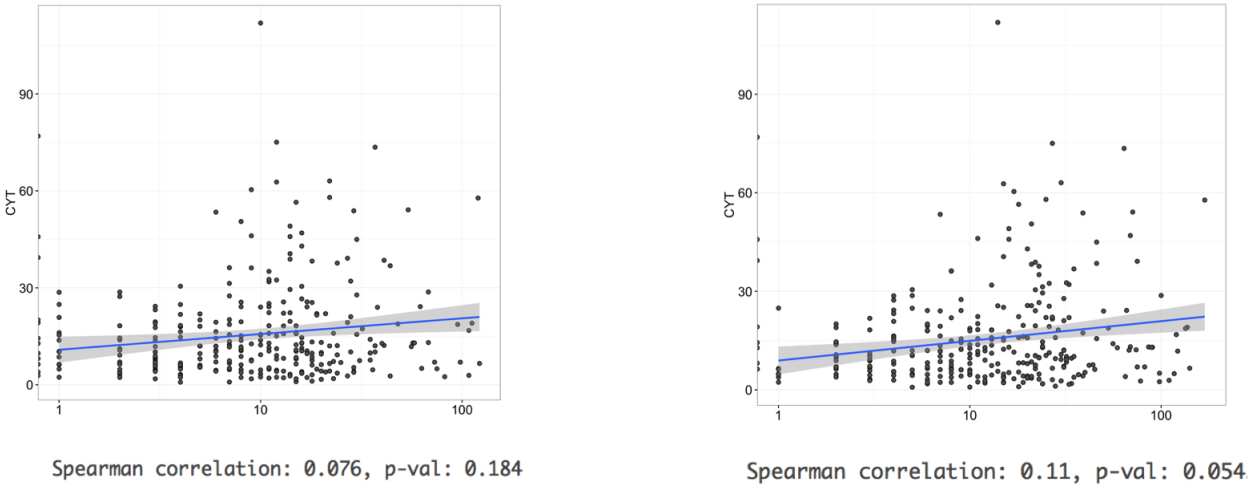


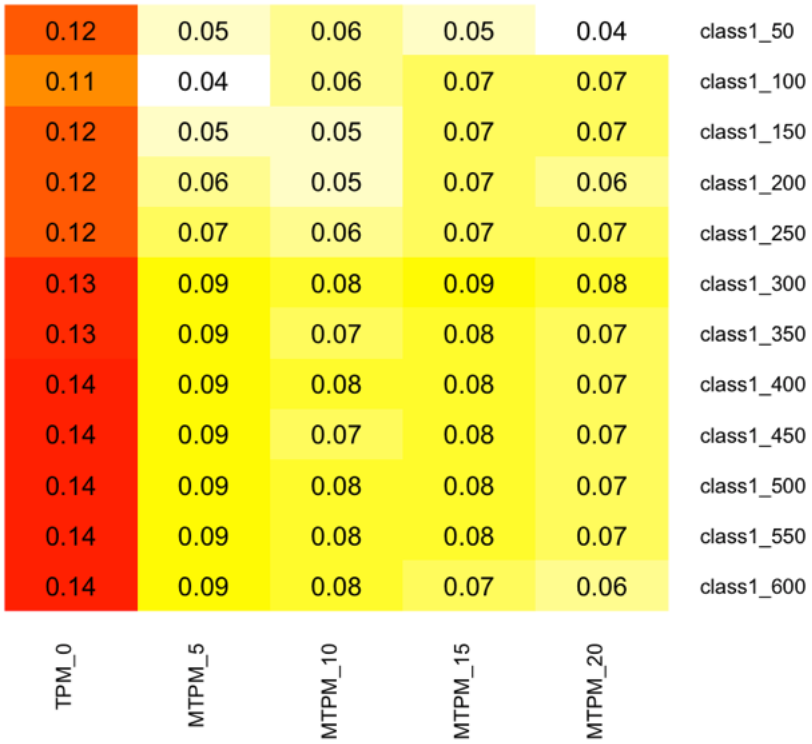
Рисунок 25 – Сравнение корреляции MHC Class I-ассоциированных эпитопов (слева) и MHC Class II-ассоциированных эпитопов (справа)

2.2.3. Сравнение разных отсечек на IC50 их влияние на корреляцию с CYT

Стандартной отсечкой, определяющей силу связывания пептида с MHC, является 500. Кроме того, естественным фильтром для мутации является экспрессия мутированных аллелей (представленность в клетке).

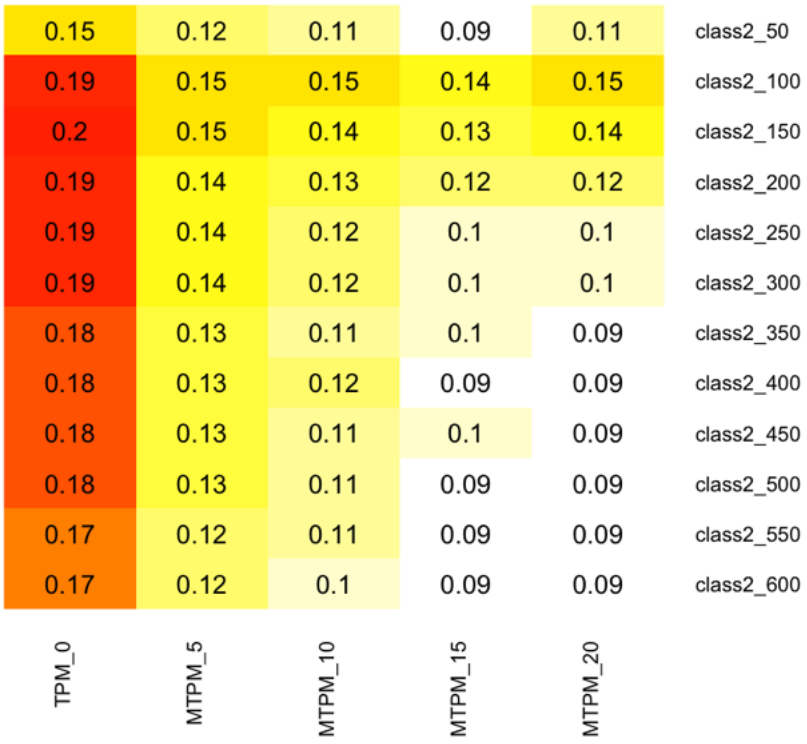
Таким образом, было проведено исследование влияния отсечки и экспрессии на корреляцию с CYT.

а) Корреляция CYT с количеством мутаций, представляющих сильный эпитоп для MHC Class I, при разных отсечках на IC50:

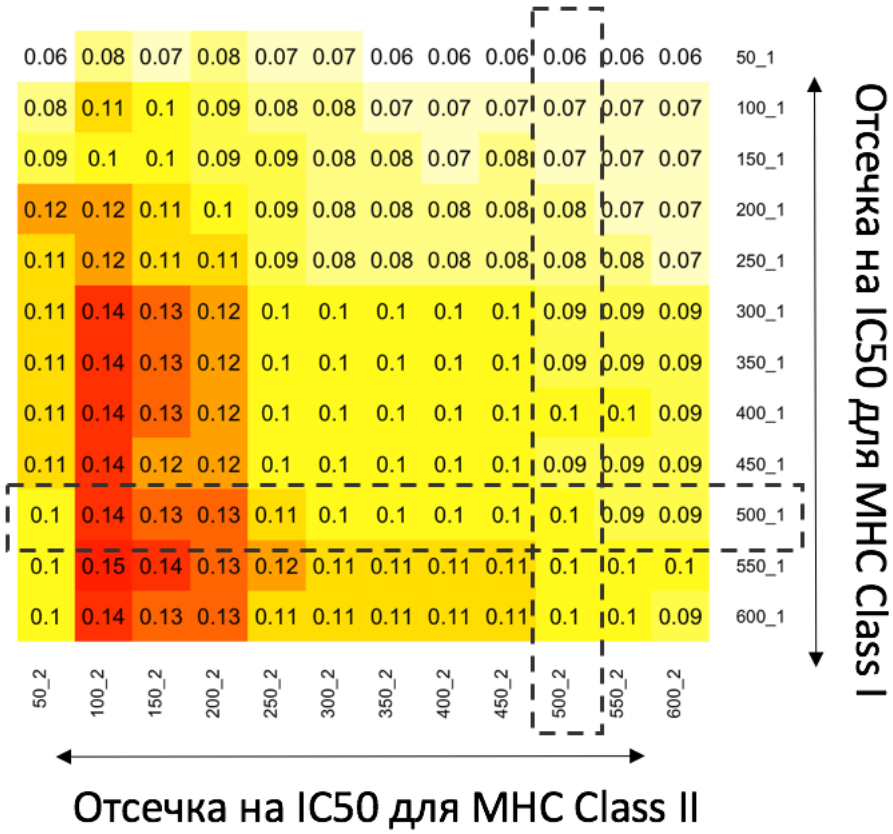


По оси X отложены разные значения на экспрессию, по оси Y - разные значения IC50. Цвет соответствует уровню корреляции, красный — самая лучшая корреляция, белый - сама низкая. По графику видно, что даже комбинация фильтров не улучшает значение корреляции.

б) Корреляция CYT с количеством мутаций, представляющих сильный эпитоп для MHC Class II, при разных отсечках на IC50:



в) Так же в качестве фильтра было решено использовать требование порождения мутацией сильных эпитопов как для MHC Class I, так и для MHC Class II.



Получается, что даже всевозможные комбинации фильтрующих факторов не ведут к увеличению корреляции с CYT, то есть мутации не могут использоваться как характеристика CYT.

2.3. Использование пайплайна для создания анти-опухолевых вакцин

При создании вакцин из репертуара мутаций пациента выбирается до 10-15 неоантигенов, имеющих IC50 < 500, создаются аминокислотные последовательности, им соответствующие, и формируется вакцина, инъекция которой будет произведена в пациента. В коллаборации с Genome Institute [33] пайплайн использовался для предсказания MHC Class II-ассоциированных неоэпитопов для пациентов с меланомой, раком мозга, раком простаты.

2.4. Выводы по главе

Был построен пайплайн для проверки гипотезы о том, что фильтрация мутаций по признаку наличия MHC Class II-ассоциированных неоантигенов улучшает корреляцию количества мутаций с CYT. Оказалось, что улучшение является незначительным. Возможными причинами может быть высокая (> 0.9) корреляция количества мутаций с количеством тех мутаций, которые порождают сильные неоантигены. В целом мутации не могут быть использованы как характеристика CYT.

ГЛАВА 3. ПРЕДСКАЗАНИЕ ЦИТОЛИТИЧЕСКОЙ АКТИВНОСТИ НА ОСНОВАНИИ ДАННЫХ МЕТИЛИРОВАНИЯ ДНК

Для исследования цитолитической активности (CYT) из данных секвенирования нового поколения использовались только данные по мутациям [7], была изучена корреляция этих двух величин.

Данные экспрессии использовались для корреляции экспрессии различных генов с CYT. Данные метилирования не использовались вообще.

3.1. Выбор модели для предсказания

В случае с метилированием, для каждому наблюдению CYT соответствует более 20000 значений от 0 до 1, и значения могут быть скоррелированны между собой. Было решено сделать линейную регрессию. Требования к модели:

а) Малое (меньше 1000) число финальных предикторов.
б) Возможность сохранения коррелированных предикторов.

По примеру *Horvath, S. et al* была использована *elastic net*, удовлетворяющая поставленным требованиям.

3.2. Формирования набора данных для тренировки

Самой обширной базой данных по пациентам раковых опухолей, содержащей качественные данные разных профилей, является TCGA. Из нее были получены данные экспресии и метилирования для пациентов с опухолями, в которых экспресси. CYT значительно отличчается от экспрессии CYT в здоровых тканях.

а) 408 пациентов с BLCA, карциномой мочевого пузыря;
б) 304 пацинта с CESC, аденокарциномой шейки матки;
в) 500 пациентов с HNSC, раком головы и шеи;
г) 527 пациентов с KIRC, почечно-клеточный раком;
д) 288 пациентов с KIRP, раком папиллярный клеток почки;
е) 440 пациентов с LUAD, аденокарциномой легких;
ж) 370 пацентов с LUSC, мелкоклеточным раком легких;
и) 466 пацентов с SKCM, меланомой;
к) 375 пацентов с STAD, аденокарциномой желудка;
л) 543 пацентов с UCEC, карциномой эндометрия сосудов матки.

Cancer	BLCA	CESC	HNSC	KIRC	KIRP	LUAD	LUSC	SKCM	STAD	UCEC
# of patients	408	304	500	527	288	450	370	466	375	543

Таблица 12 – Количество данных по каждому типу рака

В набор данных вошли пациенты, для которых присутствовали и данные метилирования, и данные экспрессии.

Раковые опухоли являются гетерогенными и редко носят схожие признаки в экспрессии и метилировании, особенно если сравнивать между собой разные болезни. Тем не менее, ввиду малого количества данных, они рассматривались в совокупности. В помощью *elastic net* нужно было изучить:

а) Предсказание CYT на всей совокупности данных с:
   — Выбором коэффициента λ с помощью кросс-валидации.
   — Варьирование коэффициента λ.

б) LOOCV (leave-one-out-cross-validation) — кросс-валидация с исключением одного типа раковых опухолей и предсказание CYT для него.

в) Предсказание экспрессии других генов, анализ возможности уловить биологический сигнал в данных метилирования.

3.3. Построение предиктора на языке R

Анализ и визуализация осуществлялись на языке **R**. Для использования *elastic net* был использован пакет *glment*, для визуализации — *ggplot2*. Кросс-валидация выполнялась методами, реализованными в *glment*. Отношение между тренировочным и тестовым набором данных было 80:20.

3.3.1. Методы оценки построенной модели

Для оценки построенной модели применяются стандартные критерии регрессии:

а) Корреляция Пирсона между реальными и предсказанными значениями.

б) Гомоскедастичность — однородная вариативность значений.

в) Медиана ошибки между реальными и предсказанными значениями.

3.4. Предсказание CYT на всей совокупности данных

На рисунке 26 показан запуск *elastic net* на всей совокупности данный. С выбранным с помощью кросс-валидации λ предикторами стало 553 CpG. Левый столбец рисунка соответствует тестовому набору данных, правый — тренировочному. Графики в первом ряду показывают, как предсказанные значения лежат относительно реальных (значения упорядочены по величине для удобства). Во втором ряду показана корреляция между реальными и предсказанными значениями. В третьем ряду график показывает гомоскедастичность — вариативность ошибки в зависимости от величины предсказываемого значения.

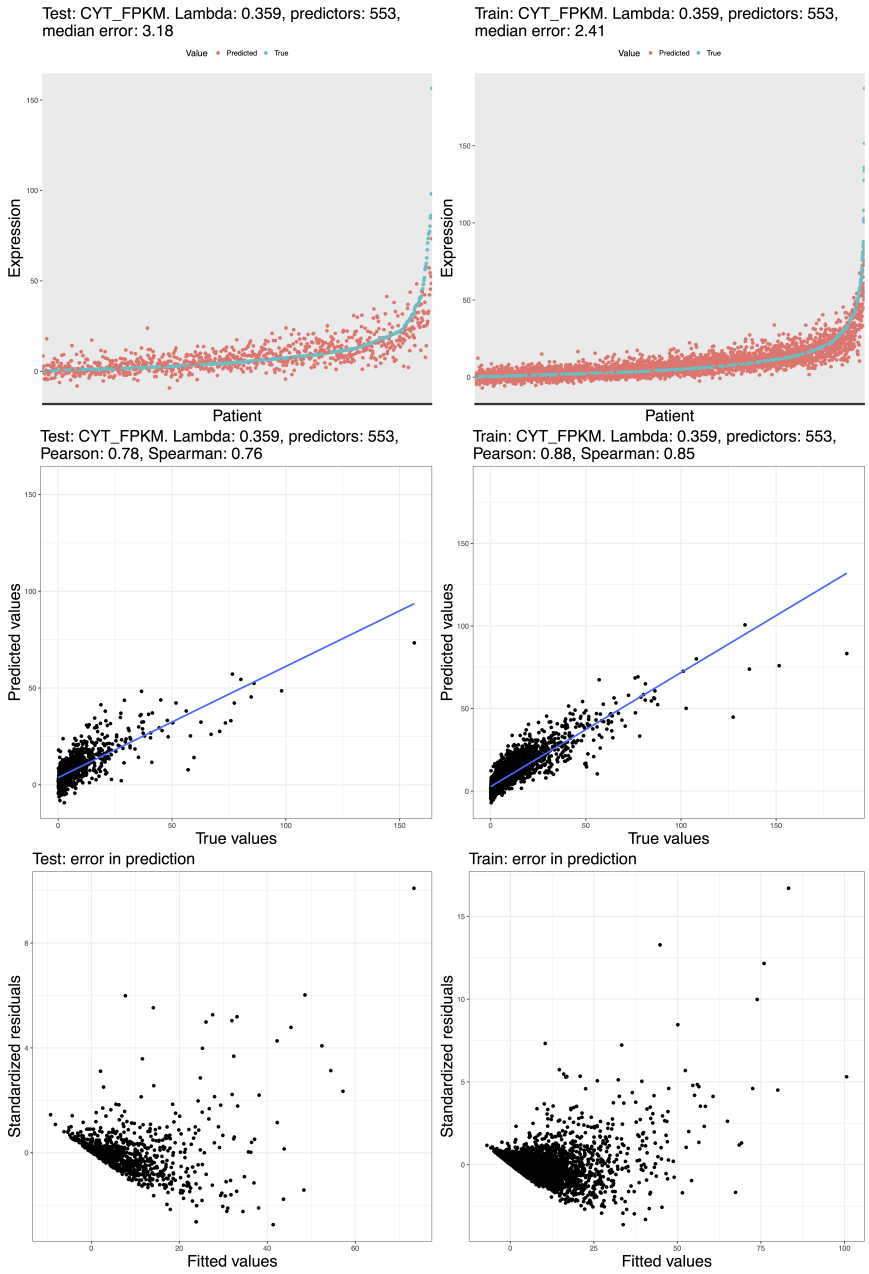


Рисунок 26 – Предсказание CYT, выбор λ с помощью кросс-валидации

На графике, показывающем гомоскедастичность, значения по оси Y должны иметь одинаковую вариацию вдоль оси X, однако видно, что вариация растет с ростом X. Получается, что экспрессия в текущем представлении не может быть представлена линейной комбинацией значений метилирования. Это может быть исправлено преобразованием CYT.

3.4.1. Преобразование CYT

В качестве преобразований будут рассмотрены:

а) логарифмическое преобразование;
б) преобразование Бокса-Кокса;
в) ранговое представление;
г) ранговое персентильное представление.

Критерием выбора является удовлетворение моделью оценок регрессионной модели.

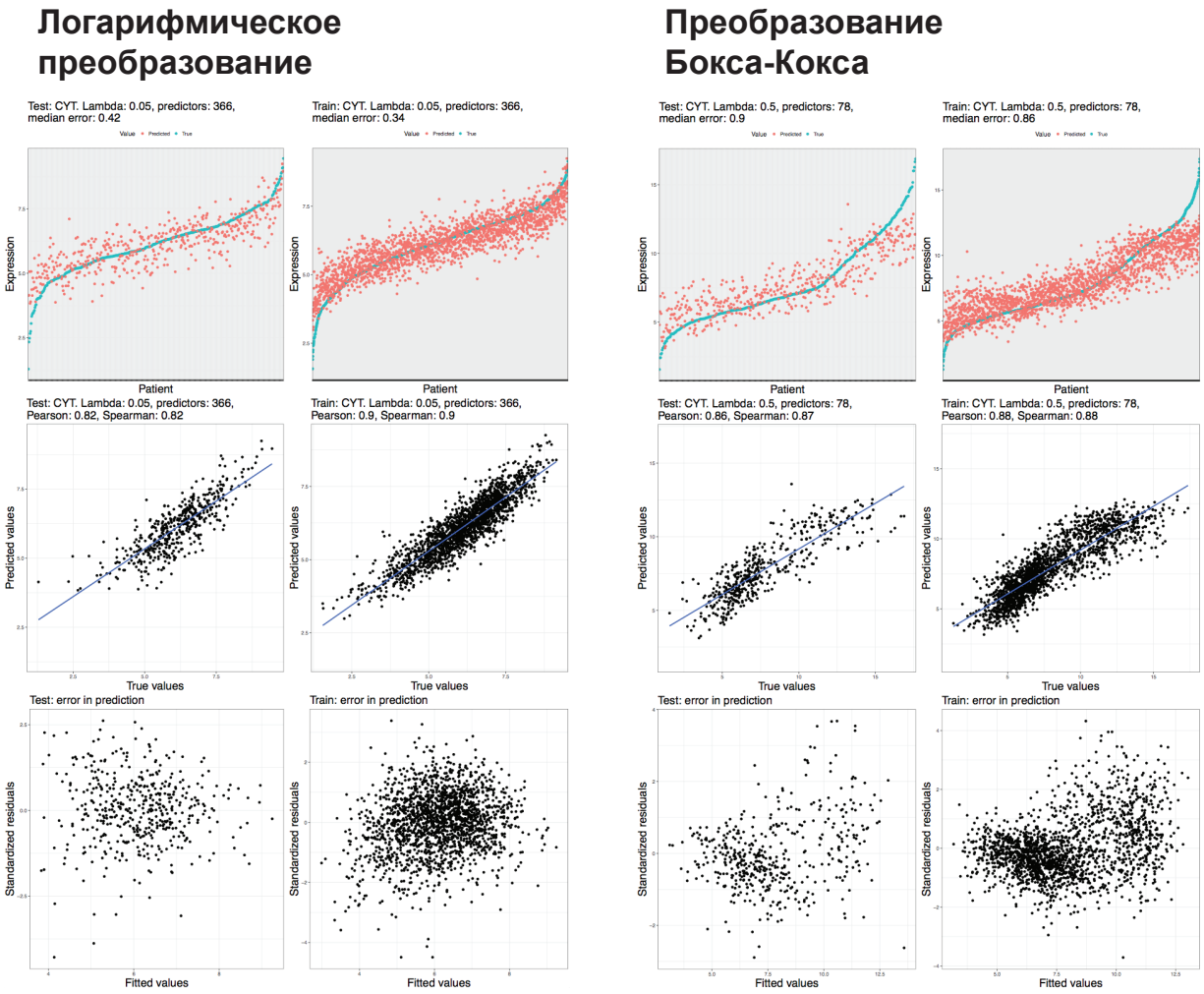


Рисунок 27 – Логарифмическое преобразование и преобразование Бокса-Кокса для CYT

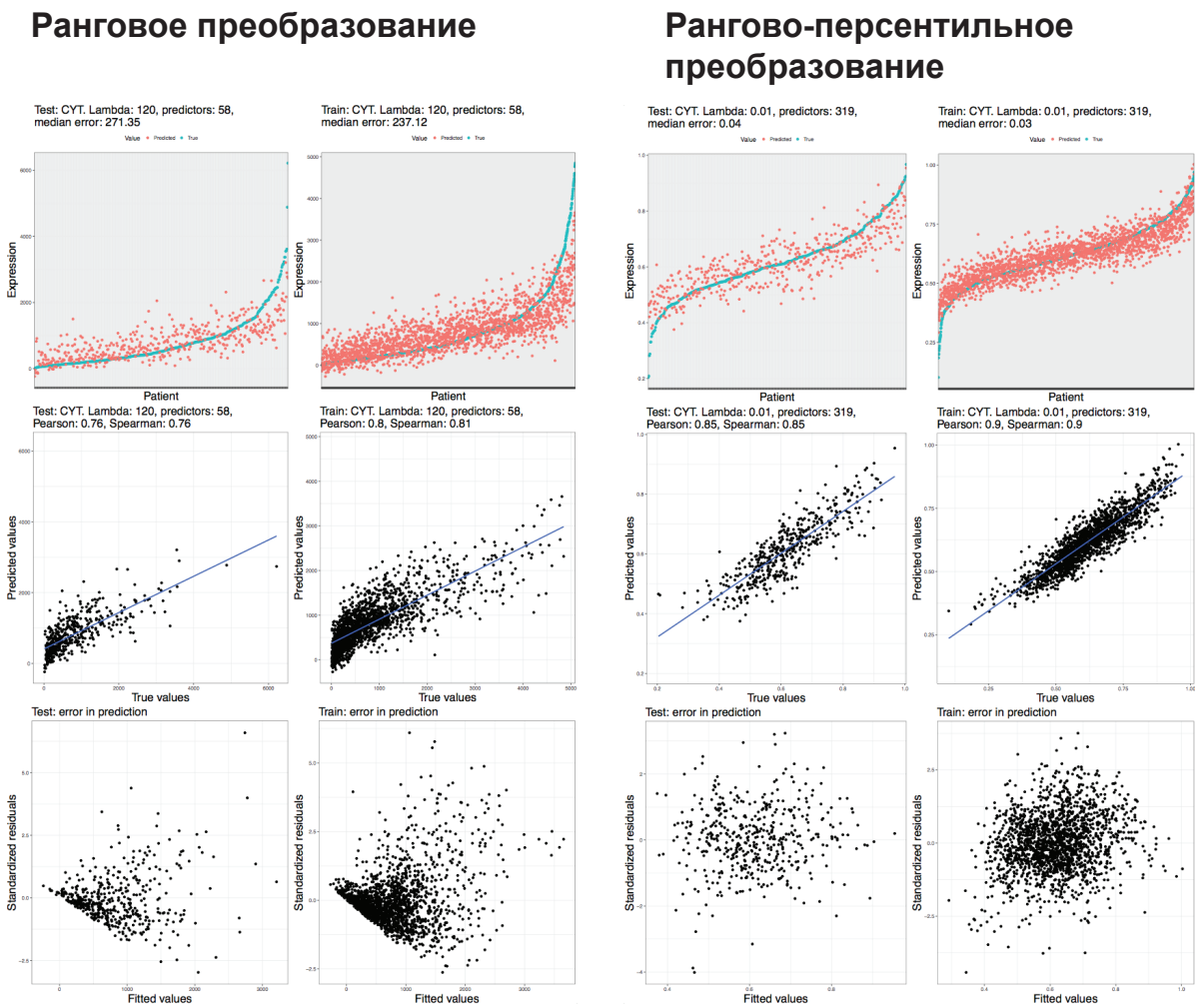


Рисунок 28 – Ранговые преобразования CYT

Преобразование Бокса-Кокса и обычное ранговое не являются гомоскедастичными. Логарифмическое и рангово-персентильное похожи между собой по уровню корреляции между реальными и предсказываемыми значениями, и имеют схожий разброс предсказываемых значений. Ввиду простоты использования, было выбрано логарифмическое преобразование.

3.5. LOOCV для разных типов раковых опухолей

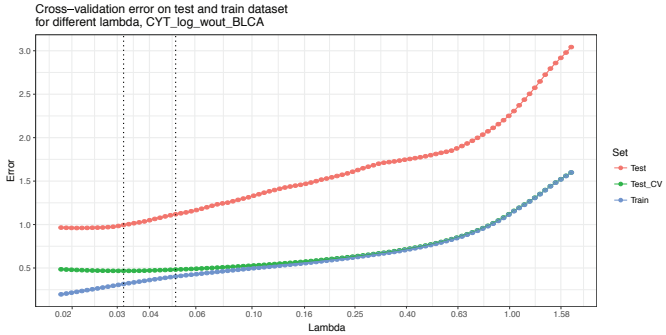
Раковые опухоли отличаются друг от друга тканями, из которых они появляются и процессами, которые в них протекают. В общем случае нет ожидания того, что модель, натренированная на одних типах опухолей, будет работать на других. Чтобы проверить это и посмотреть, на каких опухолях в целом модель работает лучше, будет выполнен LOOCV анализ для каждого вида опухоли из набора следующим образом:

а) Будет сгенерирован предиктор, не имеющий в своих тренировочных данных пациентов с определенной опухолью.

б) Для каждой модели (сгенерированной и обычной), будет построен график, где по оси X — величина λ, по оси Y — величина ошибки при данной λ. Будет рассмотрено три вида ошибок:

1) ошибка кросс-валидации;
2) ошибка на тренировочном наборе данных;
3) ошибка на тестовом наборе данных, состоящих из рассматриваемого типа опухоли.

Тренировочный набор данных, сформированный без BLCA



Тренировочный набор данных, включающий BLCA

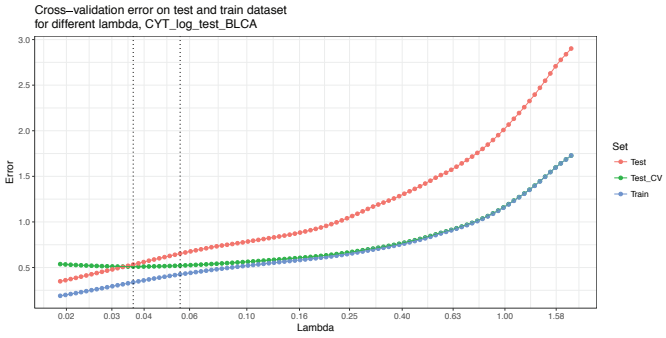
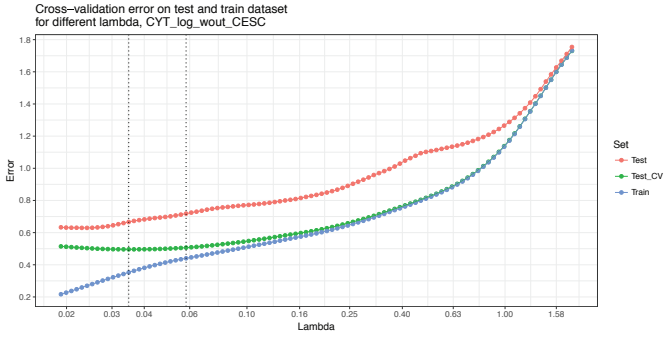


Рисунок 29 – Предсказание BLCA

Тренировочный набор данных, сформированный без CESC



Тренировочный набор данных, включающий CESC

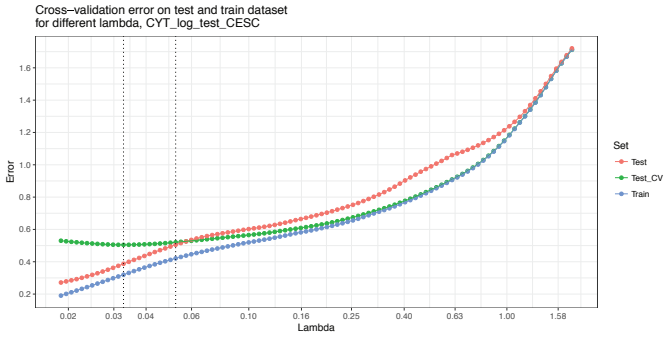
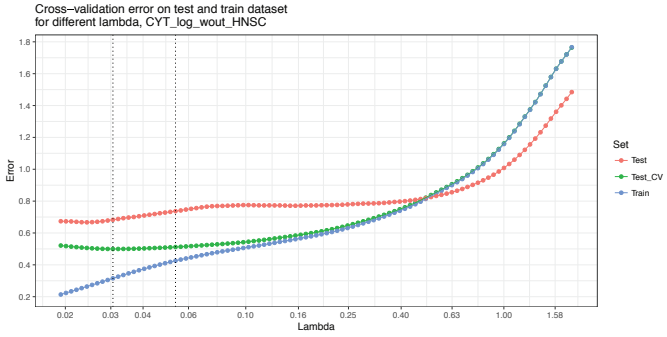


Рисунок 30 – Предсказание CESC

Тренировочный набор данных, сформированный без HNSC



Тренировочный набор данных, включающий HNSC

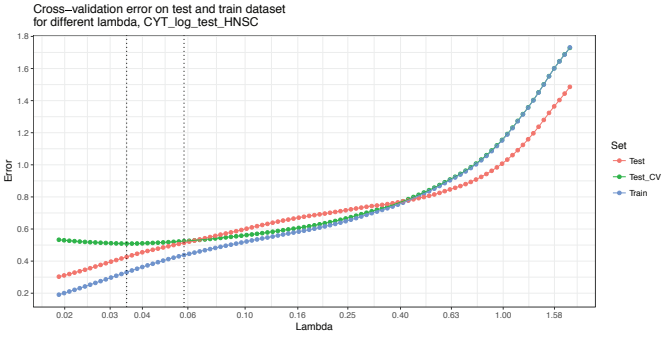
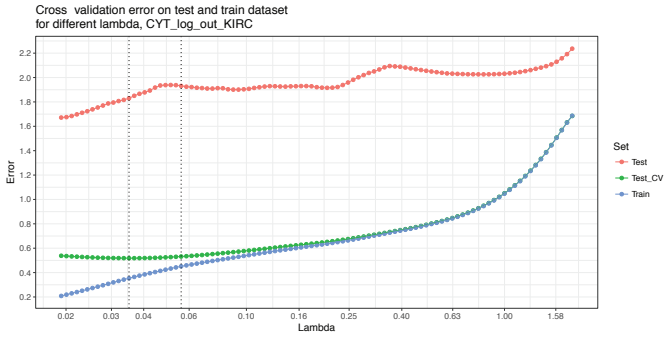


Рисунок 31 – Предсказание HNSC

Тренировочный набор данных, сформированный без KIRC



Тренировочный набор данных, включающий KIRC

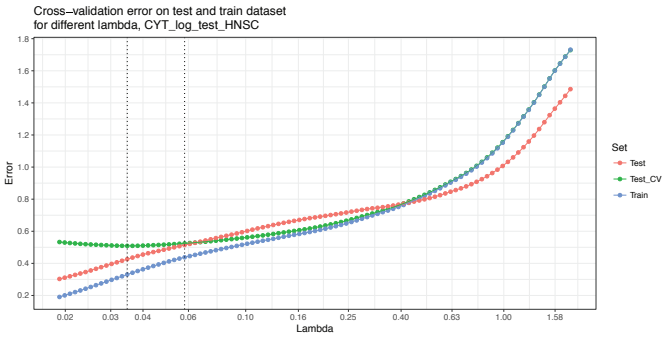
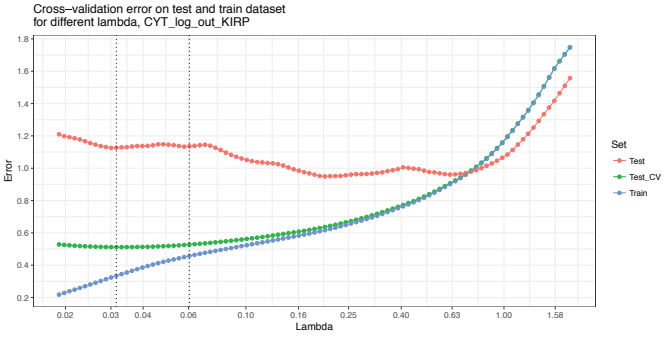


Рисунок 32 – Предсказание KIRC

Тренировочный набор данных, сформированный без KIRP



Тренировочный набор данных, включающий KIRP

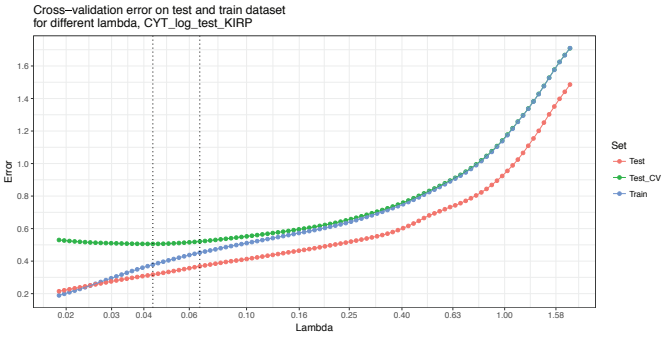
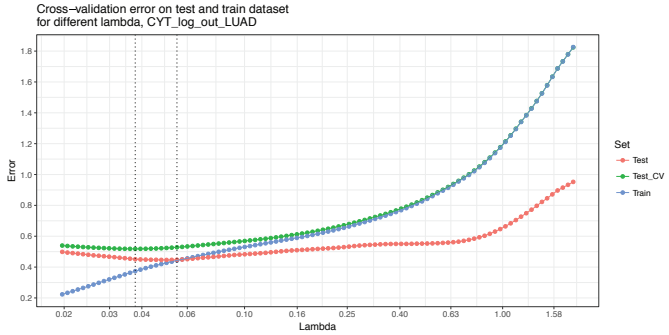


Рисунок 33 – Предсказание KIRP

Тренировочный набор данных, сформированный без LUAD



Тренировочный набор данных, включающий LUAD

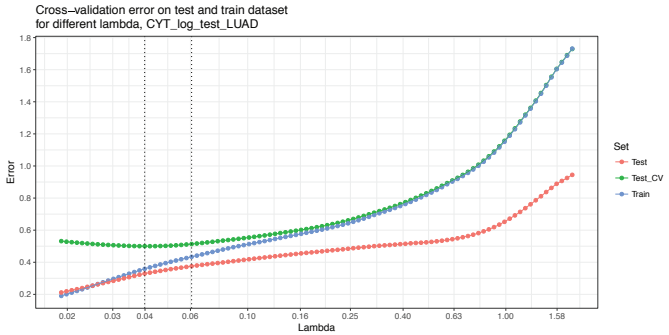
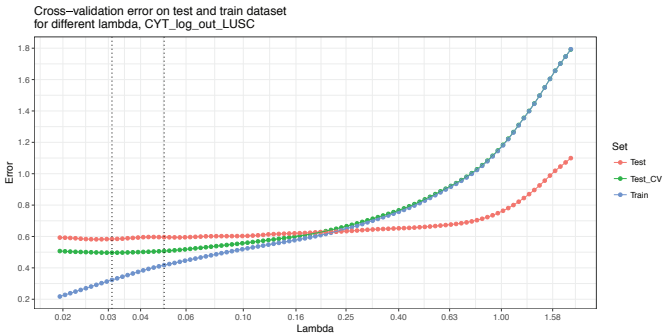


Рисунок 34 – Предсказание LUAD

Тренировочный набор данных, сформированный без LUSC



Тренировочный набор данных, включающий LUSC

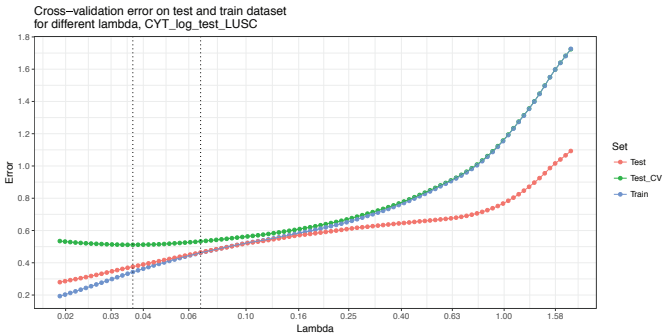
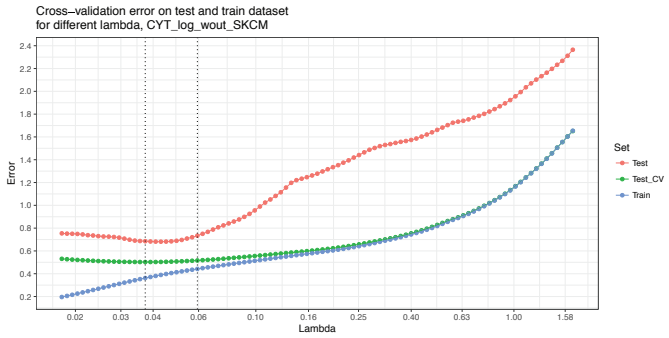


Рисунок 35 – Предсказание LUSC

Тренировочный набор данных, сформированный без SKCM



Тренировочный набор данных, включающий SKCM

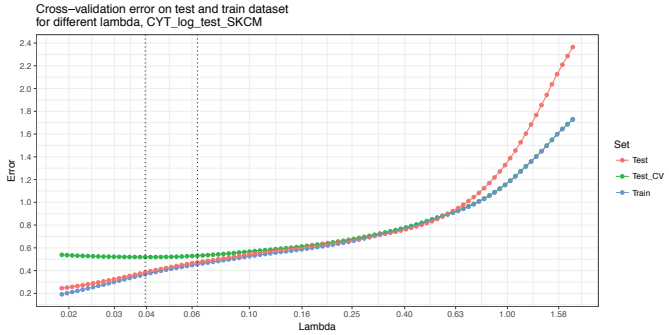
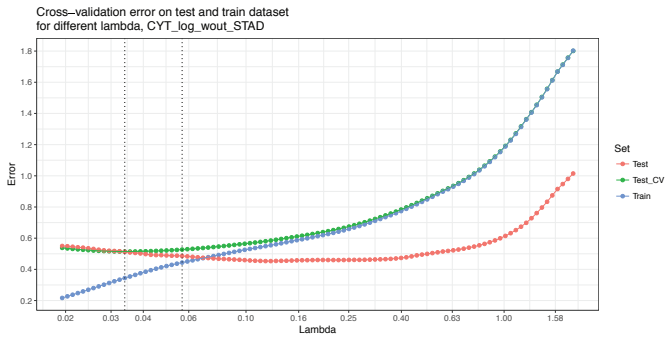


Рисунок 36 – Предсказание SKCM

Тренировочный набор данных, сформированный без STAD



Тренировочный набор данных, включающий STAD

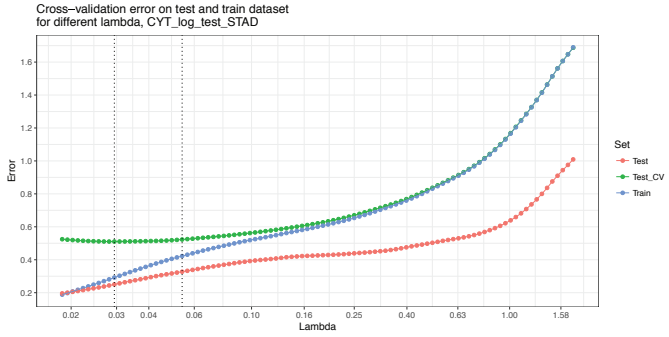
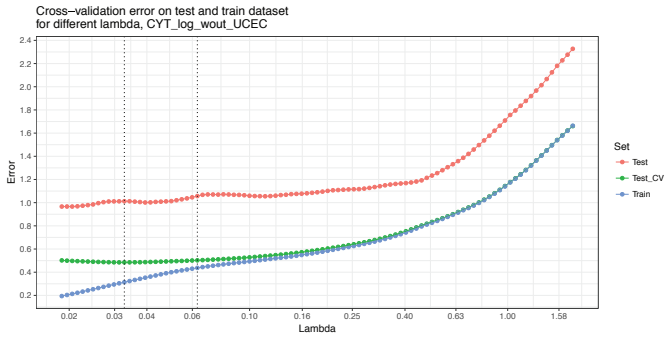


Рисунок 37 – Предсказание STAD

Тренировочный набор данных, сформированный без UCEC



Тренировочный набор данных, включающий UCEC

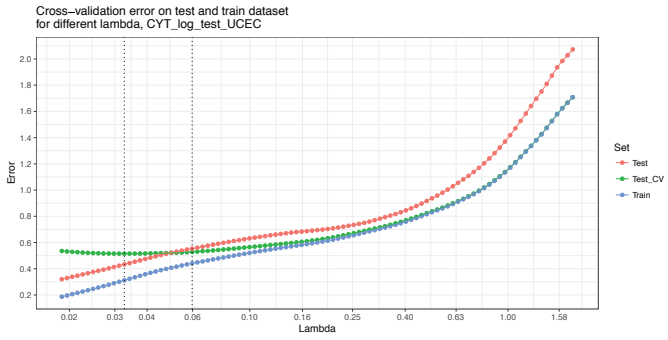


Рисунок 38 – Предсказание UCEC

Из LOOCV видно, что отсутствие данных для большинства раковых опухолей сказывается на ошибке, увеличивая ее до нескольких раз, как в BLCA (рисунок 29), KIRC (рисунок 32), KIRP (рисунок 33), UCEC (рисунок 38). Интересно, что в случае нескольких видов опухолей ошибка на тестовом наборе изначально меньше, чем на тренировочном. В случае с SKCM (рисунок 36) это может быть вызвано либо фактом, что предикторы хорошо предсказывают CYT, либо что экспрессия в SKCM меньше, чем в других опухолях.

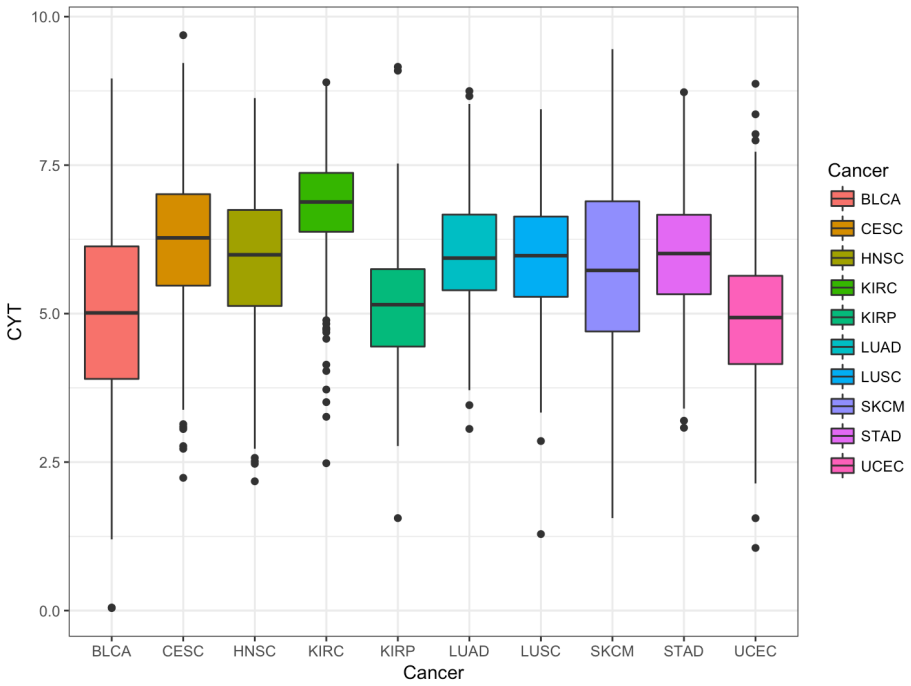


Рисунок 39 – Экспрессия CYT среди разных раковых опухолей

Как видно из рисунка 39, уровень экспрессии CYT в SKCM сравним с уровнем в других опухолях, поэтому наиболее вероятным является объяснение лучшей предсказуемости. В случае с LUAD (рисунок 34) и LUSC (рисунок 35) причиной их маленькой ошибки может быть то, что обе опухоли расположены в легких, и могут иметь схожие профили метилирования. Для того, чтобы проверить это, из тренировочной выборки были исключены оба типа.

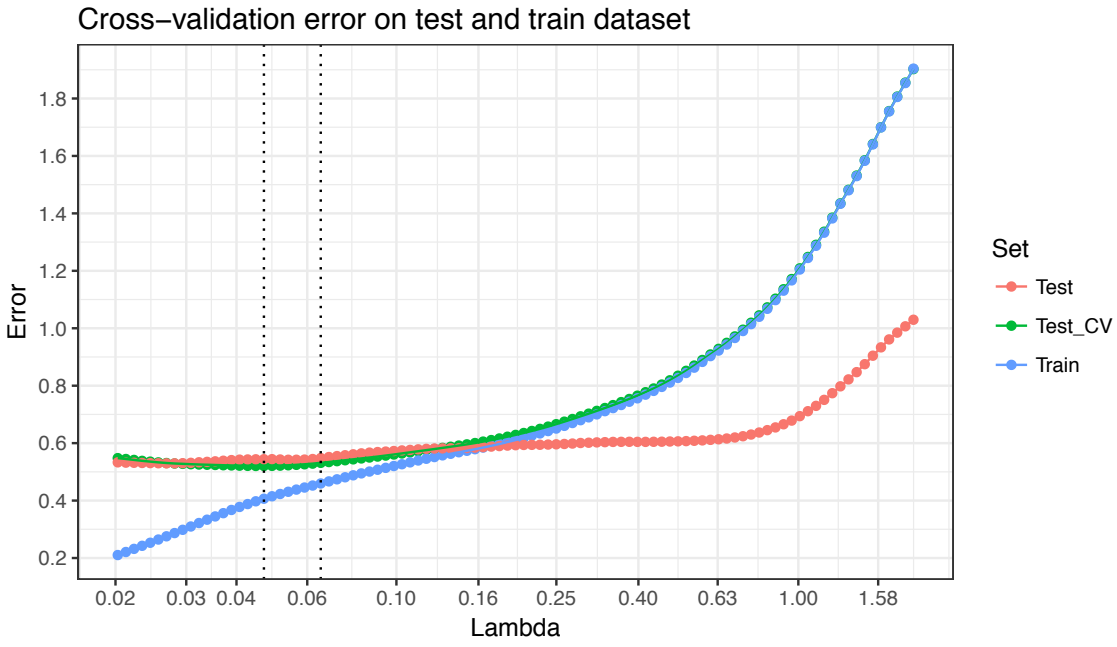


Рисунок 40 – Предсказание CYT в LUAD и LUSC, когда пациенты с такими типами опухолей отсутствуют в тренировочном наборе данных

На рисунке 40 показано, что даже в случае, когда в тестовом наборе данных не присутствуют образы из опухолей легких, ошибка на LUAD и LUSC меньше, чем ошибка на тренировочном наборе данных.

3.6. Предсказание экспрессии других генов данной моделью

В организме человека в разных тканях экспрессируется до 20 тысяч различных генов. В описанном материале было описано предсказание двух ко-экспрессирующихся генов. *elastic net* был применен для предсказания остальных генов. Целью являлось измерение корреляции как меры качества предсказания и изучения распределение корреляции в зависимости от экспрессии генов. В предположении, что изменения в экспрессии связаны с изменениями в метилировании, множество генов, не играющих роли в развитии рака и в иммунном ответе, меняющих свою экспрессию вне зависимости от процессов, происходящих опухоли, должны были предсказываться плохо.

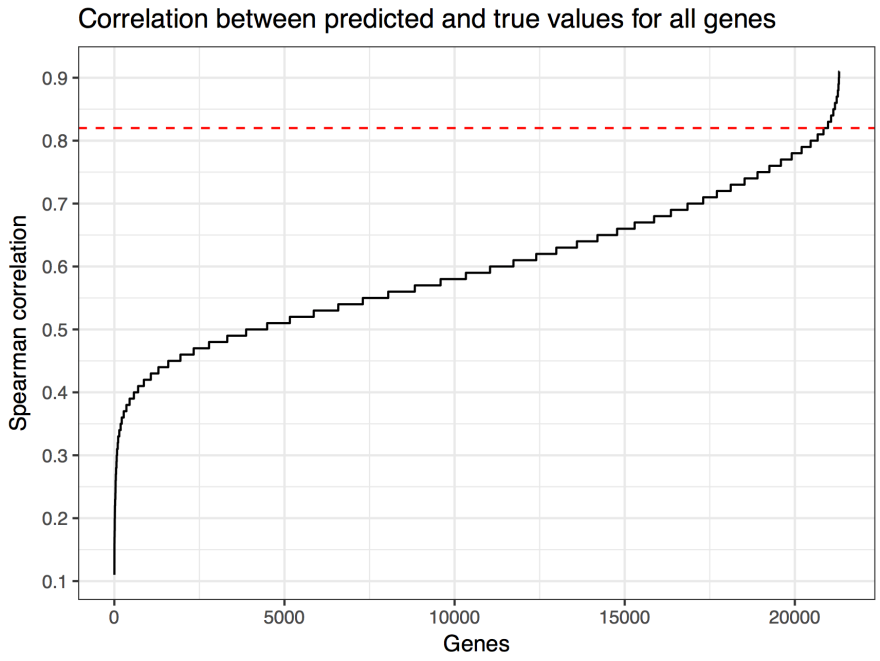


Рисунок 41 – Распределение корреляций между предсказанными и реальными значениями для всех генов. Красным указан уровень корреляции генов GZMA и PRF1, составляющих CYT

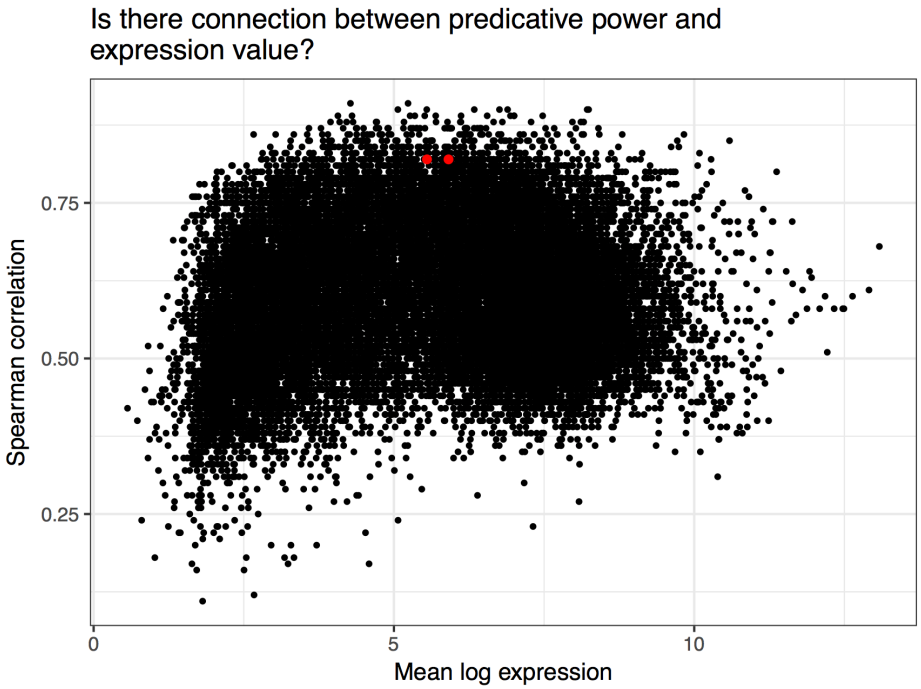


Рисунок 42 – Распределение корреляций в зависимости от значений экспрессии. Красным указан уровень корреляции генов GZMA и PRF1, составляющих CYT

На рисунках 27 и 28 видно, что большинство генов имеет корреляцию менее 0.75, а экспрессия генов с высокой корреляцией ранжируется на всем множестве значений экспрессии.

3.7. Выводы по главе

В этой главе была построена линейная регрессия на данных метилирования для предсказания цитологической активности. Для удовлетворения критериям линейной модели данные экспрессии были прологарифмированы. Количество предикторов, в зависимости от настроек составляло 300-500. Было показано, как предсказания отличается для разных типов опухолей, и что при отсутствии определенных типов раковых опухолей в тренировочном наборе данных, ошибка на пациентов с этой болезнью будет выше, чем на тестовых данных. Медианная ошибка на тестовом наборе составила 0.42 при диапазоне значений от 0 до 10. LUAD, LUSC и SKCM оказались опухолями, наилучшим образом поддающимися предсказанию.

Таким образом, данные метилирования можно использовать для предсказания экспрессии генов, релевантных развитию раковых опухолей.

ЗАКЛЮЧЕНИЕ

В данной магистерской работе был разработан и выполнен пайплайн для предсказания неоантигенов из данных ДНК и РНК секвенирования на основании точечных мутаций и инделов.

В данной магистерской работе было исследовано предсказание и объяснение цитологической активности (CYT), выраженной экспрессией генов GZMA и PRF1, с помощью MHC-ассоциированых антигенов и профилей метилирования. Для этого был разработан и выполнен пайплайн для предсказания неоантигенов из данных ДНК и РНК секвенирования на основании точечных мутаций и инделов, была использована *elastic net* для предсказания CYT с помощью данных метилирования и исследованы свойства полученного предиктора.

В том числе, созданный пайплайн использовался для создания анти-опухолевых вакцин и результаты его работы использованы в публикации:

1 Genomic landscape of high-grade meningiomas / W. L. Bi [et al.] // npj Genomic Medicine. — 2017. — Vol. 2, no. 15. — ISSN 1756-994X. — DOI: 0.1038/s41525-017-0014-7.

При написании магистерской так же было участие в следующих публикациях:

1 *Chatterjee S.* [et al.] Structural basis for human respiratory syncytial virus NS1-mediated modulation of host responses // Nature Microbiology. — 2017.

2 *Steed A.* [et al.] The Microbial Metabolite Desaminotyrosine Protects from Influenza through Type I Interferon // Science. — -.

3 *Nair S.* [et al.] Immune-Responsive Gene 1 prevents neutrophil mediated immunopathology during Mycobacterium tuberculosis infection // Immunology. — -.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Picard tools. — URL: http://broadinstitute.github.io/picard (visited on 05/08/2017).

2 Computational genomics tools for dissecting tumour-immune cell interactions / H. Hackl [et al.] // Nature reviews. Genetics. — 2016. — Vol. 17, no. 8.

3 *Mahdi B. M.* A glow of HLA typing in organ transplantation // Clinical and Translational Medicine. — 2013. — Vol. 2, no. 1. — P. 6. — ISSN 2001-1326. — DOI: 10.1186/2001-1326-2-6.

4 The IPD and IMGT/HLA database: allele variant databases / J. Robinson [et al.] // Nucleic Acids Research. — 2014. — Vol. 43, Database issue. — P. D423–D431. — DOI: 10.1093/nar/gku1161.

5 *Rizvi N.* [et al.] Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer // Cancer Immunology. — 2015.

6 Genomic Data Commons. — URL: https://gdc.cancer.gov (visited on 05/08/2017).

7 *Rooney M.* [et al.] Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity // Cell. — 2015.

8 *Kreiter S.* [et al.] Mutant MHC class II epitopes drive therapeutic immune responses to cancer // Nature. — 2015.

9 *Horvath S.* DNA methylation age of human tissues and cell types // Genome Biology. — 2015.

10 FRED 2: an immunoinformatics framework for Python / B. Schubert [et al.] // Bioinformatics. — 2016. — Vol. 32, no. 13. — P. 2044. — DOI: 10.1093/bioinformatics/btw113.

11 pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens / J. Hundal [et al.] // Genome Medicine. — 2016. — Vol. 8, no. 1. — P. 11. — ISSN 1756-994X. — DOI: 10.1186/s13073-016-0264-5.

12 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3 / P. Cingolani [et al.] // Fly. — 2012. — Vol. 6, no. 2. — P. 80–92.

13 *Wang K.*, *Li M.*, *Hakonarson H.* ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data // Nucleic Acids Research. — 2010. — Vol. 38, no. 16. — e164. — DOI: 10.1093/nar/gkq603.

14 Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads / Y. Bai [et al.] // BMC Genomics. — 2014. — Vol. 15, no. 1. — P. 325. — ISSN 1471-2164. — DOI: 10.1186/1471-2164-15-325. — URL: http://dx.doi.org/10.1186/1471-2164-15-325.

15 HLA typing from RNA-Seq sequence reads / S. Boegel [et al.] // Genome Medicine. — 2012. — Vol. 4, no. 12. — P. 102. — ISSN 1756-994X. — DOI: 10.1186/gm403.

16 ATHLATES: accurate typing of human leukocyte antigen through exome sequencing / C. Liu [et al.] // Nucleic Acids Research. — 2013. — Vol. 41, no. 14. — e142. — DOI: 10.1093/nar/gkt481.

17 Derivation of HLA types from shotgun sequence datasets / R. L. Warren [et al.] // Genome Medicine. — 2012. — Vol. 4, no. 12. — P. 95. — ISSN 1756-994X. — DOI: 10.1186/gm396.

18 OptiType: precision HLA typing from next-generation sequencing data / A. Szolek [et al.] // Bioinformatics. — 2014. — Vol. 30, no. 23. — P. 3310. — DOI: 10.1093/bioinformatics/btu548.

19 Immune Epitope Database Analysis Resource. — URL: http://tools.immuneepitope.org/main/ (visited on 05/08/2017).

20 The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage / M. Nielsen [et al.] // Immunogenetics. — 2005. — Vol. 57, no. 1. — P. 33–41. — DOI: 10.1007/s00251-005-0781-7.

21 *Zhang H.*, *Lund O.*, *Nielsen M.* The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding // Bioinformatics. — 2009. — Vol. 25, no. 10. — P. 1293. — DOI: 10.1093/bioinformatics/btp137.

22 *Andreatta M.*, *Nielsen M.* Gapped sequence alignment using artificial neural networks: application to the MHC class I system // Bioinformatics. — 2016. — Vol. 32, no. 4. — P. 511. — DOI: 10.1093/bioinformatics/btv639.

23 *Nielsen M.*, *Andreatta M.* NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets // Genome Medicine. — 2016. — Vol. 8, no. 1. — P. 33. — ISSN 1756-994X. — DOI: 10.1186/s13073-016-0288-x.

24 *Peters B.*, *Sette A.* Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method // BMC Bioinformatics. — 2005. — Vol. 6, no. 1. — P. 132. — ISSN 1471-2105. — DOI: 10.1186/1471-2105-6-132.

25 Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a Bayesian prior / Y. Kim [et al.] // BMC Bioinformatics. — 2009. — Vol. 10, no. 1. — P. 394. — ISSN 1471-2105. — DOI: 10.1186/1471-2105-10-394.

26 Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification / M. Andreatta [et al.] // Immunogenetics. — 2015. — Vol. 67, no. 11. — P. 641–650. — DOI: 10.1007/s00251-015-0873-y.

27 *Nielsen M.*, *Lundegaard C.*, *Lund O.* Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method // BMC Bioinformatics. — 2007. — Vol. 8, no. 1. — P. 238. — ISSN 1471-2105. — DOI: 10.1186/1471-2105-8-238.

28 *Nielsen M.*, *Lund O.* NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction // BMC Bioinformatics. — 2009. — Vol. 10, no. 1. — P. 296. — ISSN 1471-2105. — DOI: 10.1186/1471-2105-10-296.

29 Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries / J. Sidney [et al.] // Immunome Research. — 2008. — Vol. 4, no. 2. — ISSN 1471-2105. — DOI: 10.1186/1745-7580-4-2.

30 *Wang P.* [et al.] A Systematic Assessment of MHC Class II Peptide Binding Predictions and Evaluation of a Consensus Approach // PLOS Computational Biology. — 2008. — Apr. — Vol. 4, no. 4. — P. 1–10. — DOI: 10.1371/journal.pcbi.1000048. — URL: https://doi.org/10.1371/journal.pcbi.1000048.

31 WashU CHPC Cluster. — URL: http://mgt2.chpc.wustl.edu/wiki119/index.php/Main_Page (visited on 05/08/2017).

32 HapMap database. — URL: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-197/ (visited on 05/08/2017).

33 McDonnell Genome institute. — URL: http://genome.wustl.edu (visited on 05/08/2017).

34 *Steed A.* [et al.] The Microbial Metabolite Desaminotyrosine Protects from Influenza through Type I Interferon // Science. — -.

35 *Chatterjee S.* [et al.] Structural basis for human respiratory syncytial virus NS1-mediated modulation of host responses // Nature Microbiology. — 2017.

36 *Nair S.* [et al.] Immune-Responsive Gene 1 prevents neutrophil mediated immunopathology during Mycobacterium tuberculosis infection // Immunology. — -.

37 *Zou H.*, *Hastie T.* Regularization and variable selection via the elastic net // Journal of the Royal Statistical Society: Series B (Statistical Methodology). — 2005.

38 Genomic landscape of high-grade meningiomas / W. L. Bi [et al.] // npj Genomic Medicine. — 2017. — Vol. 2, no. 15. — ISSN 1756-994X. — DOI: 0.1038/s41525-017-0014-7.

39 Improved Survival with Ipilimumab in Patients with Metastatic Melanoma / F. S. Hodi [et al.] // New England Journal of Medicine. — 2010. — Vol. 363, no. 8. — P. 711–723.

40 *Sensi M.*, *Anichini A.* Unique Tumor Antigens: Evidence for Immune Control of Genome Integrity and Immunogenic Targets for T Cell–Mediated Patient-Specific Immunotherapy // Clinical Cancer Research. — 2006. — Vol. 12, no. 17. — P. 5023–5032. — ISSN 1078-0432. — DOI: 10.1158/1078-0432.CCR-05-2682.

41 The response of autologous T cells to a human melanoma is dominated by mutated neoantigens / V. Lennerz [et al.] // Proceedings of the National Academy of Sciences of the United States of America. — 2005. — Vol. 102, no. 44. — P. 16013–16018. — DOI: 10.1073/pnas.0500090102.

42 *Lander E.*, *Birren B.* Initial sequencing and analysis of the human genome // Nature. — 2001. — Vol. 409, no. 6822. — P. 860–921. — ISSN 0028-0836. — DOI: 10.1038/35057062.

43 EpiToolKit—a web-based workbench for vaccine design / B. Schubert [et al.] // Bioinformatics. — 2015. — Vol. 31, no. 13. — P. 2211. — DOI: 10.1093/bioinformatics/btv116.

44 Neoantigen prediction tools. — URL: https://www.nature.com/nrg/journal/v17/n8/fig_tab/nrg.2016.67_T1.html (visited on 05/08/2017).

45 TESLA, Tumor neoantigEn SeLection Alliance. — URL: http://www.parkerici.org/media/2016/parker-institute-for-cancer-immunotherapy-cri-neoantigen-alliance (visited on 05/08/2017).