Contents

[ВВЕДЕНИЕ 2](#_Toc105197276)

[ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ 5](#_Toc105197277)

[**1.1 Позитивный и негативный отбор Т-клеток** 7](#_Toc105197278)

[1.2 T-клеточный иммунный ответ 9](#_Toc105197279)

[**1.3 Локализация Т-клеток и их роль в аспекте разных периодов жизни** 11](#_Toc105197280)

[**1.3 Белковая структура Т-клеточного рецептора и основные молекулы корецептора** 12](#_Toc105197281)

[**1.4 Т-клеточное распознавание антигена** 13](#_Toc105197282)

[**1. 5 V(D)J – реаранжировка** 16](#_Toc105197283)

[**1.6 Механическая модель сборки ТКР** 19](#_Toc105197284)

[**1.7 Изучение репертуара Т-клеток методами высокопроизводительного секвенирования** 20](#_Toc105197285)

[**1. 8 Различные метрики для изучения распределения клонов** 23](#_Toc105197286)

[**1.9** **Публичные клоны** 25](#_Toc105197287)

[**1.10 Общая характеристика SARS-CoV-2** 27](#_Toc105197288)

[**1.11 Иммунный ответ против SARS-CoV-2** 30](#_Toc105197289)

[*1.11.1 Врожденный иммунный ответ против SARS-CoV-2* 30](#_Toc105197290)

[*1.11.2 Адаптивный иммунный ответ против SARS-CoV-2* 31](#_Toc105197291)

[**ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ** 32](#_Toc105197292)

[ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ 32](#_Toc105197293)

# 

APC – антигенпрезентирующая клетка (англ. antigen-presenting cell)

CDR – гипервариабельный регион (англ. complementarity determining region)

DN – “двойные негативные” тимоциты (англ. “double-negative” thymocytes)

DP – “двойные позитивные” тимоциты (англ. “double-positive” thymocytes)

EM – Т-клетки эффекторной памяти (англ. effector memory)

HLA – человеческий лейкоцитарный антиген (англ. human leukocyte antigen)

MHC – главный комплекс гистосовместимости (англ. major histocompatibility complex

RACE – метод синтеза кДНК на базе РНК (англ. rapid amplification of cDNA ends)

RAG1 и RAG2 – гены, запускающие рекомбинацию (англ. recombination activating genes)

RSSs – сигнальные последовательности рекомбинации (англ. recombination signal sequences)

TCR – Т-клеточный рецептор (англ. T-cell receptor)

TRA – α цепь Т-клеточного рецептора

TRB – α цепь Т-клеточного рецептора

C – константный сегмент гена Т-клеточного рецептора

J – соединительный сегмент гена Т-клеточного рецептора

V – вариабельный сегмент гена Т-клеточного рецептора

Treg – регуляторные Т-клетки (англ. regulatory T-cells)

UMI – уникальный молекулярный идентификатор (англ. unique molecular identifier)

ДНК – дезоксирибонуклеиновая кислота

ПЦР – полимеразная цепная реакция

РНК – рибонуклеиновая кислота

# **ВВЕДЕНИЕ**

Каждый человек обладает уникальным набором Т-клеточных рецепторов, именуемым ТКР репертуаром, который содержит потенциально читаемую информацию о перенесенных инфекционных заболеваниях. Трудности, связанные с аннотацией антиген-специфических вариантов ТКР и интерпретацией информации отдельно взятого репертуара препятствуют проведению диагностике инфекционных заболеваний на основании анализа репертуара ТКР (\*). Несмотря на это, применение методов машинного обучения позволяет при наличии достаточного количества размеченных данных репертуаров с известим инфекционным статусом выявить варианты и паттерны последовательностей ТКР характерные для данного заболевания.  
В настоящей работе рассмотрено несколько подходов генерации данных для машинного обучения и оценена экстраполируемость полученных результатов на независимой выборке данных. Нам удалось добиться высокой точности предсказания статуса для данных репертуаров, получаемых в одинаковых экспериментальных условиях. Однако наш подход оказался чувствительным к групповым эффектам (batch effects), что, представляет серьезное препятствие для развития универсального метода диагностики, основанного на данных репертуаров ТКР. Стала очевидна необходимость предварительной нормировки данных секвенироания направленной на устранение групповых эффектов.

При встрече с новым патогеном или вакциной организм запускает механизм адаптивнго иммунитета, который отвечает на данное воздействие продукцией клонов Т-клеток, специфически распознающих чужеродные антигены (\*). Отобранные специфическим антигеном клетки, многократно делятся в ходе эффекторного иммунного ответа, образуя многочисленную популяцию Т-клеток памяти. Высоко антигенспецифическая природа ответа Т-клеток предпологает, что репертуар ТКР хранит в себе инфорацию, расшифровка которой позволит диагностировать целый ряд заболеваний, в которых задействован адаптивный иммунный ответ таких как рак, аутоимунные и инфекционных заболевания. В то же время развитие диагностики, основанной на анализе репертуара ТКР затрудняется высокой комплексностью адаптивного иммунного ответа. Более того, каждая Т-клетка распознает антиген, в контексте индивидуального антиген-презентирующего HLA комплекса (\*). Таким образом, диагностические паттерны ТКР будут разными у каждого человека в зависимости от HLA контекста, который довольно вариабелен внутри человеческой популяции (\*). Разнообразие методов используемых для подготовки ТКР библиотек и высокопроизводительного секвенирования (\*), порождающее значительные групповые эффекты усложняет поставленную задачу.  
В тоже время, при наличии достаточного количества данных образцов репертуаров с известным статусом, можно выявить характерные для данного заболевания паттерны ТКР, не используя специальные знания об антиген специфичности отдельных вариантов ТКР (\*).  
Тяжелый острый респираторный синдром коронавируса (SARS-CoV-2) — заболевание, вызвавшее мировую пандемию коронавирусной инфекции (COVID-19) в 2020-2021 годах. Понимание того, как меняется структура репертуара ТКР в результате адаптивного иммунного ответа на SARS-CoV-2 имеет решающее значение для разработки вакцины, клинической оценки состояния пациентов и оценки вероятности повторного заражения. Появляется все больше данных, свидетельствующих о важной роли Т-клеточного иммунитета в общем иммунном ответе на SARS-CoV-2. Идентификация Т-клеток, специфичных к SARS-CoV-2, имеет решающее значение для выявления паттернов ТКР характерных для инфицированных SARS-CoV-2.  
Репертуар Т-клеток представляет собой совокупность всех уникальных клонов Т-клеток , а анализ репертуара представляет собой биоинформационный подход к количественному и качественному описанию Т-клеточного иммунитета. Многие исследования подтверждают, что репертуар Т-клеток полученный методом высокопроизводительного секвенирования, может использоваться в качестве биомаркера иммунного ответа (\*).  
Таким образом, применение методов машинного обучения к данным репертуаров ТКР здоровых доноров и пациентов, перенесших COVID-19, может дать представление о свойствах иммунного ответа на данный патоген и точно предсказать иммунный статус по COVID-19.

**Цели и задачи**

Основной целью нашего исследования является анализ структуры Т-клеточного репертуара здоровых и перенесших COVID индивидов и использование данных их репертуаров для предсказания статуса COVID с помощью методов машинного обучения.

Мы стремились найти паттерны ТКР - биомаркеры, характерные для пациентов, инфицированных SARS-CoV-2 в текущий момент или перенесших заболевание в прошлом. Для классификации образцов были использованы модели Random Forest и SVM.

Перед нами были поставлены следующие задачи:

* Провести нормализацию данных секвенирования репертуаров ТКР по распределению частот генных сегментов V (V-usage) и глубине секвенирования.
* Провести анализ индивидуальных репертуаров периферических Т-лимфоцитов для выявления групп клонов, ассоциированных с SARS-CoV-2.
* Провести трансформацию данных индивидуальных репертуаров в векторы признаков (извлечение признаков) и сформировать размеченные выборки данных для машинного обучения
* Провести оптимизацию параметров алгоритмов классификации и выявить наиболее эффективный способ предварительной обработки данных для машинного обучения.

# **ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ**

Множество последовательностей репертуара обеспечивает специфическую реакцию на широкий спектр антигенов. В основе наблюдаемого разнообразия репертуаров Т-клеточных рецепторов (ТКР) лежит механизм соматической рекомбинации ДНК, называемый V(D)J-рекомбинацией [1] - стохастической комбинации генных сегментов V, D (в случае формирования TCRβ) и J, а также удаления или вставки случайного числа нуклеотидов с концов сегментов.

Цепь TCRα образуется в аналогичном процессе, объединяется с цепью TCRβ с образованием ТКР, который связывает специфический ему антиген, представленный на поверхности белков главного комплекса гистосовместимости (MHC) класса I, поэтому антигенная специфичность ТКР дополнительно модулируется ко+нтекстом MHC. После распознавания антигена активированные Т-клетки пролиферируют путем клональной экспансии, и некоторые из них становятся частью компартмента памяти, где они могут находиться в течение многих лет в виде клональных популяций клеток с идентичными ТКР благодаря их общему происхождению от единичной наивной Т-клетки.

В крови здорового взрослого человека циркулирует ~10^12 Т-клеток, экспрессирующих на своей поверхности примерно 10^7 уникальных вариантов Т-клеточных рецепторов. При этом общее число возможных перестроек генных сегментов VDJ значительно превышает наблюдаемое число вариантов ТКР. Также отличительно то, что в человеческой популяции, одни и те же цепи TCRβ встречается в тысячи раз чаще, чем можно было бы ожидать, если бы все перестройки были равновероятны. Совокупность всех возможный вариантов последовательностей ТКРβ присутствуют в репертуарах наивных Т-клеток у большинства людей, но клональной экспансии подвергаются лишь те варианты, которые вступили во взаимодействие со своей антигенной мишенью. Такие подвергавшиеся клональной экспансии клетки, несущие ТКР, представленные у большинства людей, называются публичные клонотипы. Их наличие отражает с одной стороны, общность индивидуального опыта взаимодействия с патогенами, которые наиболее распространены в соответствующей человеческой популяции с другой, наблюдаемые популяционные частоты ТКР однозначно определяются вероятностной моделью VDJ –реаранжировки.

Иммунный репертуар является динамической, подвижной структурой, состав и разнообразие репертуара TCR изменяется в процессе онтогенеза, зависит от окружающих факторов, генетических и инфекционных условий. [3, 4].

## **1.1 Позитивный и негативный отбор Т-клеток**

Предшественники Т-клеток это плюрипотентные, кроветворные, стволовые клетки, которые мигрируют из костного мозга в вилочковую железу – место, их дальнейшего созревания. К тому времени, как Т-клетки попадают в вилочковую железу, они не имеют поверхностных маркёров зрелой клетки и их гены ТКР еще не подверглись рекомбинации. Такие тимоциты называются двойными отрицательными (DN, double-negative), так как не экспрессируют корецепторы CD4 и CD8, определяющие характер взаимодейсвия с другими клетками. Развитие двойных отрицательных Т-клеток состоит из 4 стадий дифференцировки (DN1 – DN4), которые различаются типом эксрпессируемого поверхностного маркера. В течение этапа дифференцировки DN3 происходит процесс VDJ - рекомбинация бета цепи. После стадии дифференцировки DN4, тимоциты становятся двойными положительными (DP) клетками, несущими маркеры CD4 и CD8. В процессе этого этапа происходит перестроение альфа цепи. Дальнейший отбор клеток протекает в тимусе.

V(D)J рекомбинация является только первым шагом на пути к формированию Т-клеточного репертуара. На данной стадии развития тимоцитов на их клеточной поверхности могут быть выявлены почти все возможные варианты ТКР, но в дальнейшем их разнообразие значительно сокращается в связи с отбором клеток в тимусе и переферических органах иммунной системы. Основная причина для такого отбора заключается в том, что рекомбинация является случайным процессом, который может привести к появлению, с одной стороны аутореактивных Т-клеток, которые распознают собственные антигены и способных запустить аутоиммунные процессы, с другой ТРК, не способных связывать MHC-антиген комплекс.

Таким образом, функциональный отбор препятствует появлению потенциально опасных аутореактивных клеток и элиминирует не обладающие активностью T-клетки.

Предшественники Т-клеток образуются в костном мозге, однако их созревание происходит в тимусе (вилочковой железе). Ведущую роль в отборе играют тимусные эпителиальные клетки (ТЭК). Эти клетки имеют многочисленные мембранные отростки, формируют специальное микроокружение тимоцитов, выполняют функцию тканевого каркаса и являются источниками сигналов, поддерживающих жизнеспособность Т-клеток. ТЭК презентируют собственные антигены в составе MHC I класса и MHC II класса двойными положительным CD4+CD8+ Т-клеткам.  
В ходе положительного отбора лимфоциты, которые не способны связывать МНС и, следовательно, распознавать чужеродные антигены, уничтожаются апоптозом. Приблизительно 10-30% всех Т-клеток проходят положительный отбор. После отбора, двойные положительные Т-клетки экспрессируют один из двух корецепторов – в зависимости от класса МНС, взаимодействие с которым было наиболее сильным. Т-клетки, обладающие слишком высокой аффинностью к комплексу МНС-антиген также подвергаются элиминации или положительному отбору. Таким образом, индивидуальный набор аллелей HLA во многом предопределяет качественный состав репертуара ТКР.  
<https://www.mediasphera.ru/issues/arkhiv-patologii/2020/4/downloads/ru/1000419552013041009>

Для того чтобы как можно больше аутореактивных Т-клеток подверглись апоптозу, медуллярные эпителиальные клетки тимуса (mTEC) должны представлять тимоцитам широкий спектр аутоантигенов. Эти клетки экспрессируют специальный фактор транскрипции AIRE (аутоиммунный регулятор), который обеспечивает транскрипцию набора тканеспецифических белков [44]. Более того, недавние работы указывают на важную роль в этом процессе транскрипционного фактора FEZF2 [45]. Однако определенная часть популяции Т-клеток, распознающих аутоантигены, не подвергается апоптозу, вместо этого они становятся регуляторными Т-клетками [46] (см. рис. 6). Механизм, который определяет судьбу клеток: подвергнется ли она апоптозу или дифференцировке в Т-регуляторные клетки, остается неясным.

После отбора в тимусе созревшие T-клетки выходят в кровоток и циркулируют по организму в виде наивных Т-клеток. При встрече с антигеном они активируются, получая сигнал к дифференцирокве и многократно делятся, увеличивая свою долю в репретуаре Т-клеток. Было показано, что у лиц страдающих хроническими вирусными заболеваниями (например, вирус Эбштейна-Барр (ВЭБ), цитомегаловирус (ЦМВ)) в репертуаре Т-клеток содержится значительная фракция вирус-специфических ТКР[47].

## **1.2 T-клеточный иммунный ответ**

Тригером для активации T-клеток являются два сигнала: (i) взаимодействие ТКР со специфичным пептидом, экспонированным на поверхности антиген презентирующей клетки (АПК), и (ii) костимулирующие сигналы, обеспечиваемые АПК. Примером костимулирующего сигнала может служить взаимодействием белка CD28 на поверхности Т-лимфоцита с костимулирующей молекулой В7 на АПК. Кроме того, клетка как правило получает множество сигналов, модулирующих ее активность, такие как провоспалительные или противовоспалительные цитокины.

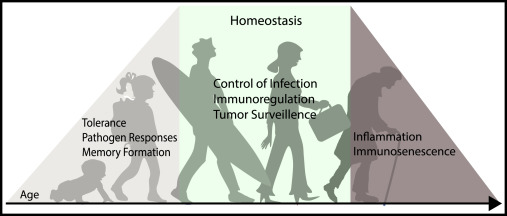
Перед активацией, ТКР группируются в небольшие островки - нанокластеры по 35-70 нм каждый из которых состоит из 7-30 ТКР. После события двойного распознавания образовавшиеся кластеры перемещаются в центр клетки, вытесняя молекулы адгезии и формируя иммуннологический синапс или супрамолекулярный активационный кластер (SMAC — Supramolecular activation cluster). Таким образом, ТКР и CD28 локализовываются в центре кластера (cSMAC), ядро которого представлено ТКР, окруженных молекулами CD28. Периферию кластера (pSMAC) занимают молекулы адгезии, такие как интегрин αLβ2 (LFA-1). Структурирование зоны контакта между клетками устраняет стерические препятствия для взаимодействия клеток, обеспечивает мобилизацию молекул адгезии, а также взаимную локализацию ТКР и CD28, обеспечивающих активацию Т-клеток.

При активации CD4+ Т-клетки дифференцируются в различные клеточные типы и могут выполнять множество функций, от активации В-лимфоцитов до подавления иммунной реакции. Основные типы клеток включают Th1 и Th2, Т-хелперы 1 и 2 типа, клетки Th17, участвующие в аллергических и аутоиммуных реакциях, клетки Treg, которые играют важную роль в регуляции иммунного ответа. Цитотоксические CD8+ Т-клетки секретируют цитокины, обладающие собственными эффекторными функциями. Данный тип клеток обеспечивает выборочную элиминацию инфицированных вирусом клеток. После активации Т-клетки многократно делятся, секретируют цитокины и образуют многочисленную популяцию клеток памяти, которая обеспечивает долговременный иммунитет.

## **1.3 Локализация Т-клеток и их роль в аспекте разных периодов жизни**

В раннем возрасте (младенчество и раннее детство) большинство Т-клеток представляют собой наивные Т-клетки, и в меньшей степени представлены Treg-клетками. На данном этапе развития, организм сталкивается с огромным количеством новых антигенов, поэтому наивные Т-клетки играют существенную роль в защите против патогенов. В свою очередь, Treg-клетки имеют решающее значение для развития толерантности к безвредным и распространенным антигенам. Также в этот период активно формируется резерв Т-клеток памяти. Процесс накопления Т-клеток памяти стабилизируется во взрослом возрасте и сохраняется в течение десятилетий [5] Разное соотношение наивных T-клеток и Т-клеток памяти в детском и взрослом возрасте, и относительная стабильность иммунитета в течение десятилетий у взрослых людей обусловлено разной ролью Т-клеток на разных этапах индивидуального развития (рис. 1). Во взрослом возрасте организм сталкивается с меньшем количеством новых антигенов и толерантность сформирована к большинству антигенов, с которыми регулярно контактирует организм, поэтому основная функция Т-клеток смещается в направлении поддержания гомеостаза и иммунорегуляции в контексте повторяющихся и часто встречающихся антигенов. У пожилых людей часто наблюдается иммунодефициты, связанные с процессами старения [6], включая и обеднение Т-клеточного репертуара.

Т-клетки населяют практически каждый орган и ткань, включая первичную и вторичную лимфоидную ткань, слизистые оболочки и барьерные участки, экзокринные органы, жировую ткань и центральную нервную систему. Что касается их численности, то большинство Т-клеток в организме человека находятся в лимфоидных тканях (BM, селезенка, миндалины и примерно 500–700 лимфатических узлов). Значительное количество клеток также присутствует в участках слизистой оболочки (легкие, тонкий и толстый кишечник) и коже, и 2-3% от общего количества Т-клеток находится в периферической крови (Clark, 2010, Ганусов и Де Бур, 2007). В раннем возрасте вновь появившиеся наивные Т-клетки и Treg-клетки заселяют основные участки лимфоидной ткани и слизистой оболочки в организме, а Т-клетки памяти локализуются преимущественно в участках слизистой оболочки, таких как тонкий кишечник и легкие (Thome et al., 2016a). По мере взросления человека Т-клетки памяти становятся преобладающим популяцией во всех структурах организма. (Thome et al., 2014). Эти данные предполагают различные роли Т-клеток в иммунитете не только на разных этапах жизни, но и в определенных анатомических компартментах.



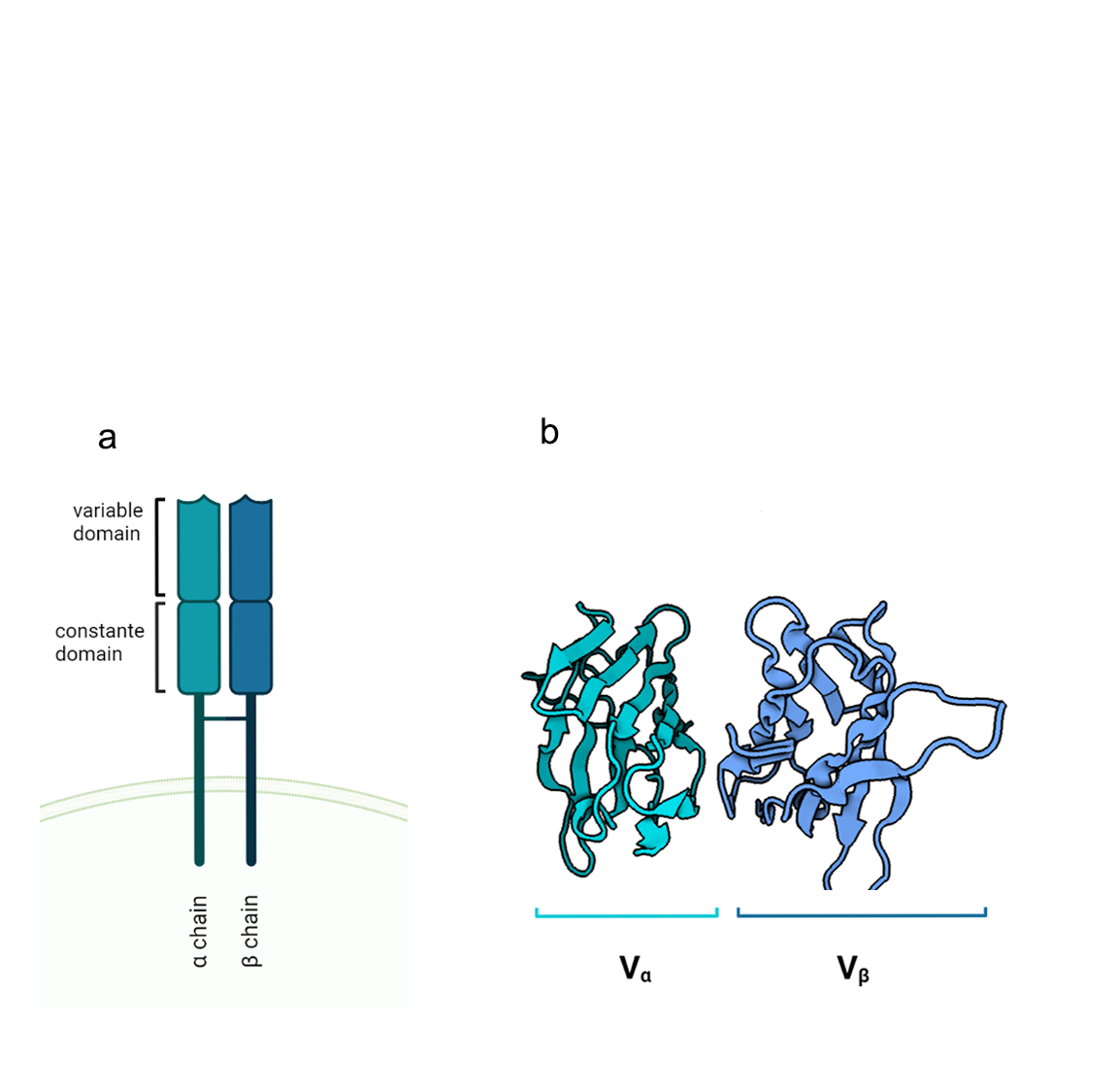
*Рисунок 1. Роль Т-клеток на разных этапах жизни [\*]*

Kumar, Brahma V.; Connors, Thomas J.; Farber, Donna L. (2018). *Human T Cell Development, Localization, and Function throughout Life. Immunity, 48(2), 202–213.*doi:10.1016/j.immuni.2018.01.007

## **1.3 Белковая структура Т-клеточного рецептора и основные молекулы корецептора**

Т-клеточный рецептор (ТКР) представляет собой заякоренный на мембране гетеродимер, состоящий из 2 цепей. Большинство Т-лимфоцитов ТКР, состоят из α- и β-цепей, соединенных дисульфидной связью [2]. Помимо αβ T-лимфоцитов, имеется также небольшое количество **γ**δ T-лимфоцитов. Функциональные свойства и распознавание антигена сильно

различаются у αβ- и **γ**δ-лимфоцитов и являются предметом исследования [3-5]. В текущей работе мы сосредоточимся на классических αβ-TCR.  
Каждая из цепей TCR состоит из 2 доменов иммуноглобулина (Ig): N-концевого вариабельного (V) домена и С-концевого константного (C) домена. α- и β-цепей также имеют трансмембранный (ТМ) домен и небольшую С-концевую цитоплазматическую область. TCR имеет 1 антигенсвязывающую область, расположенную на N-конце. Обе цепи TCR участвуют в распознавании антигена [6-8].

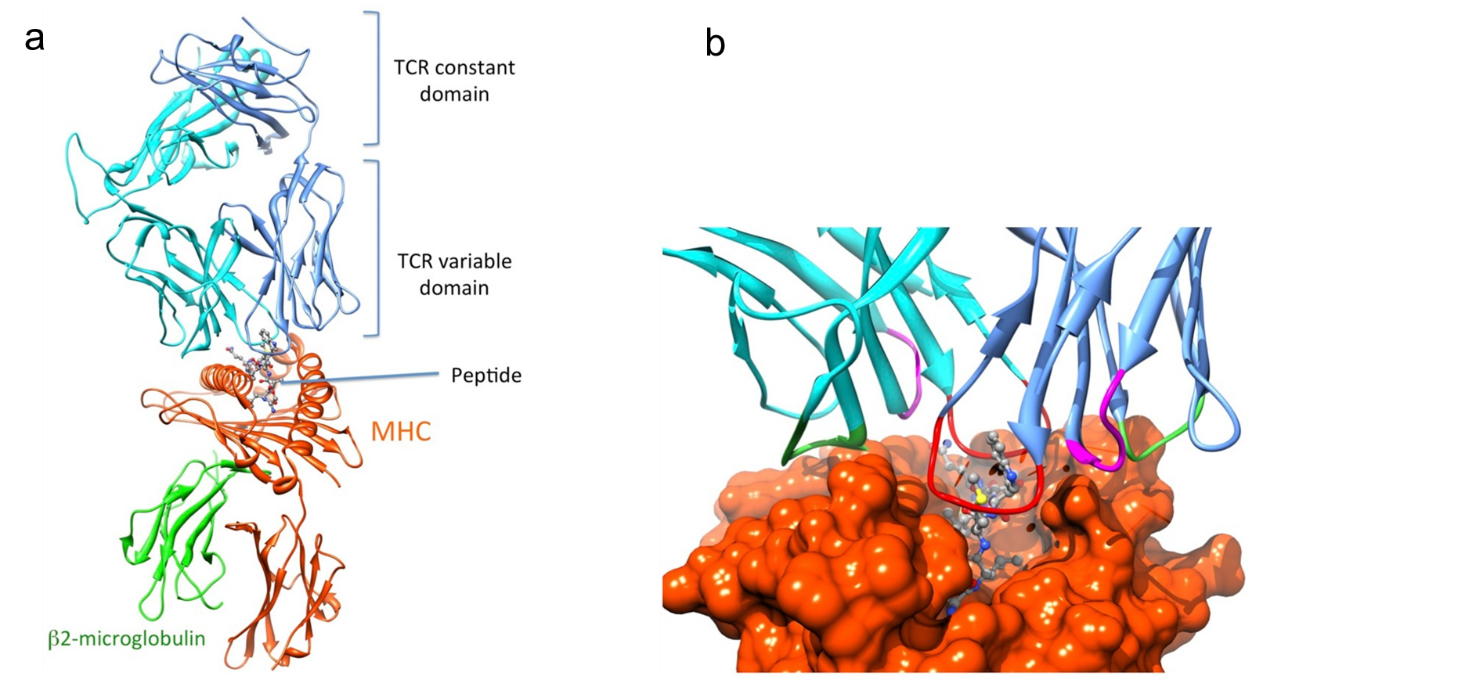
****

***Рисунок 1****. a – ТКР представляет собой гетеродимер, состоящий из α- и β-цепей, каждая из которых состоит из вариабельной (V) и константной (C) областей. b - 3D структура вариабельного домена α- и β-цепей ТКР.*

Помимо TCR, в связывании лиганда участвуют также корецепторы Т-клеток – они обозначаются как CD4 и CD8 [9-11]. CD4 представляет собой одиночную полипептидную цепь с 4 доменами Ig. CD8 представляет собой гетеродимер α- и β-цепей, соединенных дисульфидной связью. Каждая из цепей CD8 включает один домен Ig. Эти корецепторные молекулы экспрессируются на функционально разных клетках: CD8 является маркером цитотоксических Т-киллеров, а CD4 — корецептором Т-хелперов [12, 13]. CD4 и CD8 не взаимодействуют напрямую с антигеном, но обеспечивают взаимодействие между TCR и MHC.

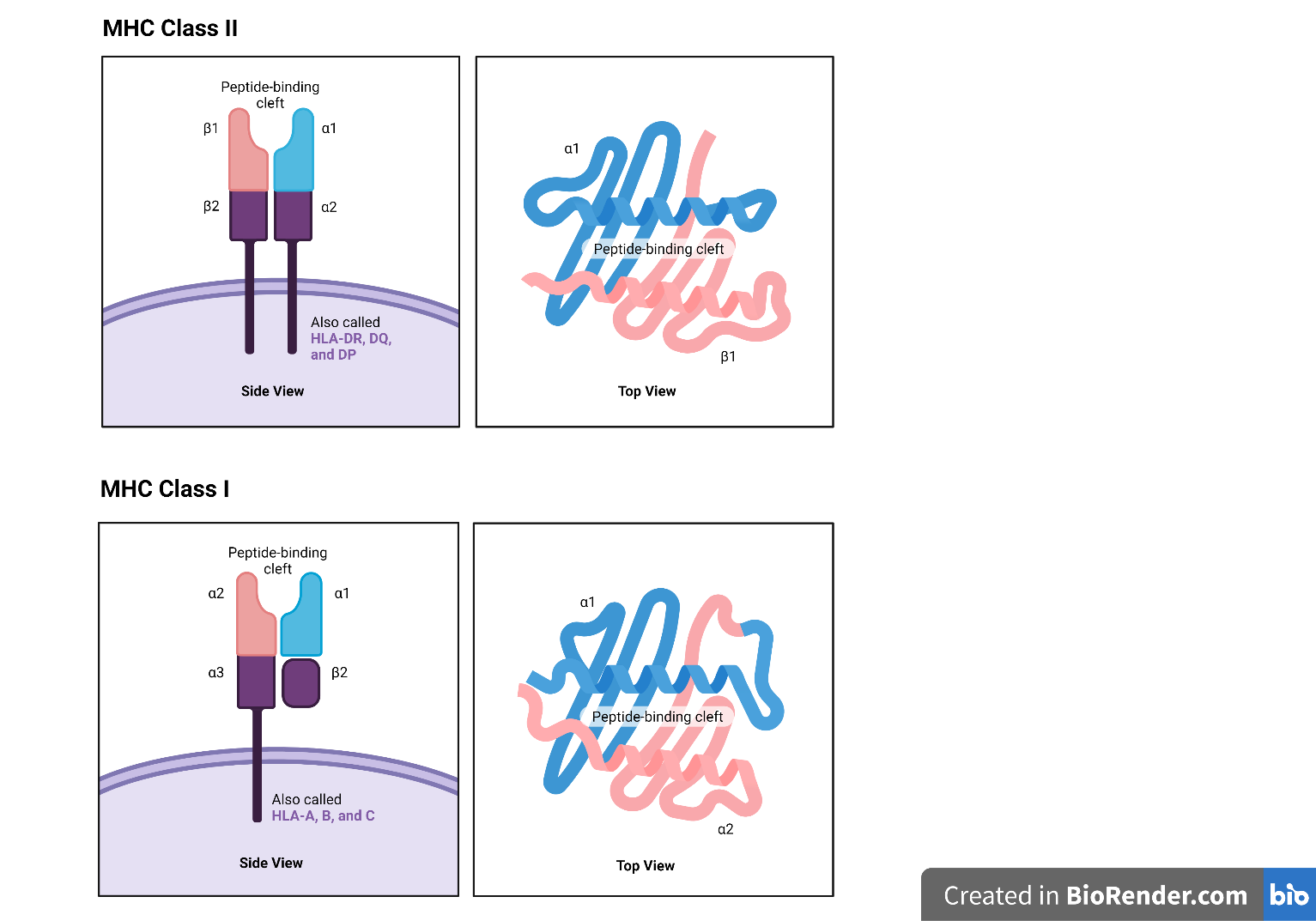
## **1.4 Т-клеточное распознавание антигена**

Способы распознавания антигена ТКР, БКР и антителами существенно различаются. БКР способны распознавать нативный антиген, то есть, они связывают доступные поверхностные эпитопы интактного белка. В то время как, ТКР распознают лишь фрагменты молекул антигенна, предварительно протестированные в макрофагах. Более того, ТКР могут распознавать эпитопы только в комплексе с главным комплексом гистосовместимости (MHC) [14, 15].

******

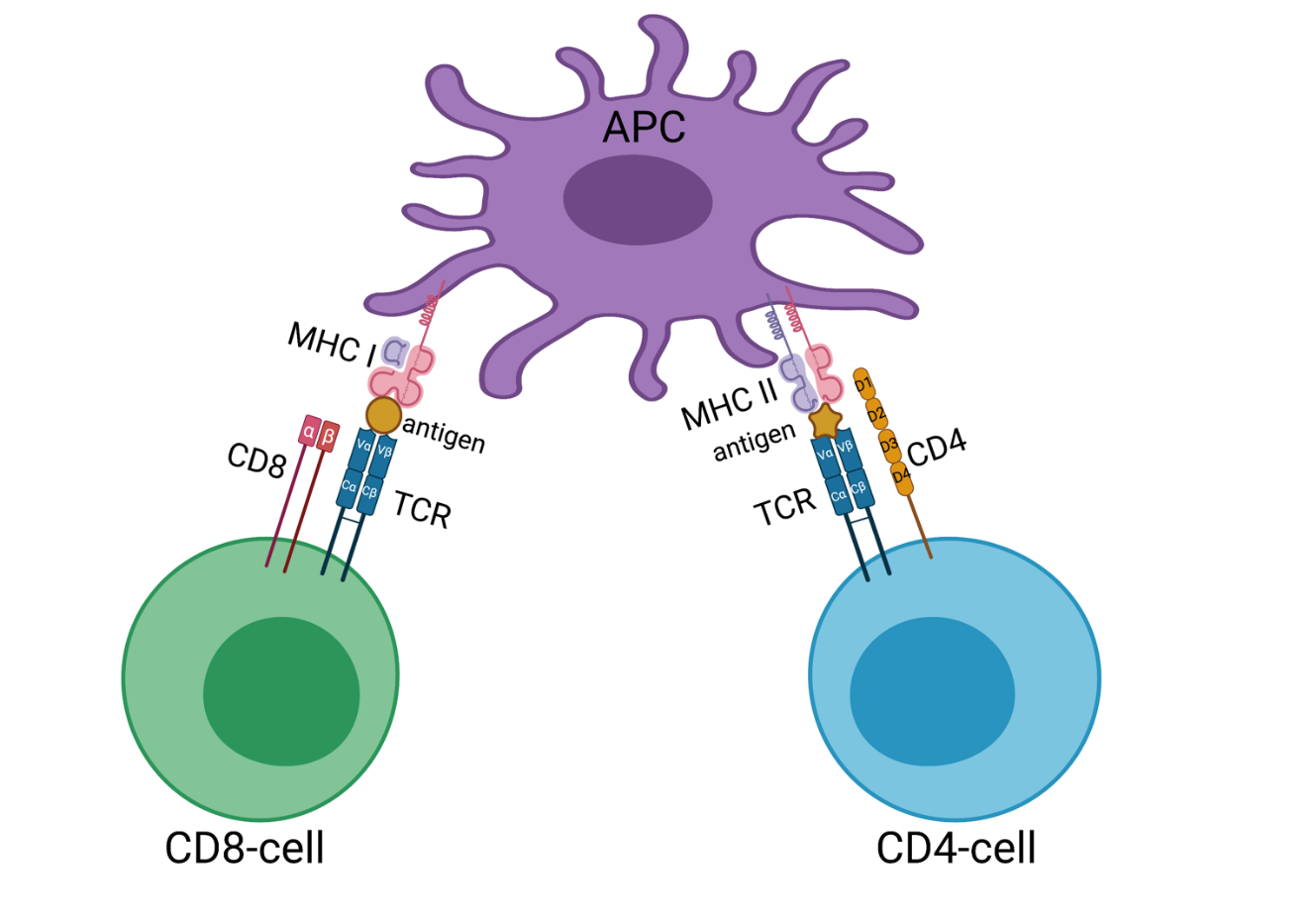
***Рисунок 2****. a – 3D структура комплекса TCR-pMHC. b – cхема взаимодействия вариабельных участков TRA и TRB с пептидом в комплексе с MHC. Участки CDR α- и β-цепей отмечены цветами: CDR1 – фиолетовым, CDR2 – зеленым, CDR3 – красным.*

MHC разделяют на 2 класса, которые имеют как структурные, так и функциональные различия. Молекулы I класса MHC состоит из 2 полипептидных цепей: α-цепь и β2- микроглобулин, которые связаны не ковалентно и формируют 4 домена. Дистальные домены α1 иα2 формируют антигенсвязывающую бороздку. Молекулы MHC II класса представлены гетеродимером 2 трансмембранных цепей: α и β. Так же, как и I класс MHC, II класс состоит из 4 доменов, 2 из которых - α1 and β1 формируют антигенсвязывающую область. Антигенсвязывающую бороздка обоих MHC имеет схожую структуру, где пептиды заключены между 2 α-спиралями [16]. Однако, бороздка в MHC II длиннее и шире, что позволяет связывать более длинные пептиды. Строго говоря, MHC класс I способен связывать пептиды длинной 8-10 аа [17], в то время как MHC класса II взаимодействует с пептидами длинной 13-25 аа [18]. Связывание пептидов это критическое событие, которое стабилизирует структуру ГКГС структуру[19].



***Рисунок 3****. Структура MHC II и MHC I. МНС класса I состоит из 4 доменов и имеет антигенсвязывающую бороздку, способную связывать пептиды длиной 8-10 аминокислот. МНС класса II также состоит из 4 доменов, но они расположены на двух полиморфных субъединицах, a его антигенсвязывающая бороздка способна связывать более длинные (18-25 а.о.) пептиды.*

Для того чтобы обеспечить презентирование всего спектра возможных антигенов Т-клеткам, MHC молекулы должны быть способны связывать огромное множество пептидов. Это реализуется благодаря полиморфизму MHC в популяции и наличию у каждого человека 6 разных MHC аллелей. У человека MHC кодируется группой плотно локализованных HLA генов, обладающих самым высоким уровнем полиморфизма человеческом организме. На сегодняшний день описаны более чем 6000 аллелей MHC I класса и 2000 аллелей ГКГС II класса [21]. Локус MHC I класса представлен тремя генами: *HLA-A, HLA-B, HLA-C*, в то время как альфа и бета цепи MHC II класса кодируются генами *HLA-DR, HLA-DQ, HLA-DP.* Таким образом, MHC присуще два свойства обеспечивающих их разнообразие: с одной стороны полигенность – наличие в организме несколько генов МНС, с другой полиморфность – сущестование в популяции множество вариантов гена МНС. Например, для HLA-А на 2020 известно 5907 аллелей. Наиболее вариабельный участок в МНС генах кодирует антигенсвязывающую область [22]. Таким образом, благодаря вариативности структур МНС, как у отдельно взятого человека, так и в популяции людей становится возможным распознавание любого существующего в окружающем мире антигена Т-клетками. Более того, каждый вариант МНС отвечает за презентирование его собственного набора пептидов [23]. По настоящее время описаны многие алгоритмы предсказания взаимодействия МНС I класса с пептидом [24-26]. Моделирование становиться более сложным, когда речь идёт о МНС II класса, так как он связывает пептиды переменной длинны Усугубляет ситуацию и то, что наличие высокой аффинность между МНС и пептидом не является достаточным для презентации, так как на этот процесс влияют множество факторов, такие как уровень экспрессии данного пептида, вероятность появления данного фрагмента пептида в результате процессинга, стабильность МНС-пептидного комплекса. Несмотря на это, разработка алгоритмов предсказания взаимодействия МНС с пептидом является широким полем для исследований. [27].



***Рисунок 4****. Взаимодействия между АПК и CD8/CD4 T-клетками. Т-клетки CD8 распознают антиген в комплексе с MHC класса I, и это распознавание опосредовано ТКР и корецептором CD8. Т-клетки CD4 распознают антиген в комплексе с MHC класса II, и распознавание опосредовано TCR и корецептором CD4. Молекула CD8 состоит из α- и β-цепей с одним доменом Ig в каждой. Молекула CD4 представляет собой одиночный полипротеин, состоящий из 4 доменов Ig.*

МНС I и II также различаются функционально. МНС I класса презентирует внутриклеточные пептиды CD8+ Т-клеткам. После распознавания Т-клеткой такого пептида в комплексе с МНС I класса, она активируется и осуществляет лизис инфицированной клетки. Таким образом, CD8+ Т-клетки играют главную роль в T-клеточном иммунитете против вирусных инфекций.

Так как вирус способен реплицироваться в любой ядросодержащей клетке, все ядерные клетки организма содержат на поверхности МНС I класса, который обеспечивает передачу информации о наличие в них инфекции T-клеткам. В отличие от МНС I класса, МНС II класса экспрессируется только в ограниченной субпопуляции клеток: дендритных клеток, макрофагов и В-клетках. Такие клетки называются антиген-презентирующие клетки (АПК), так все они способны презентировать внеклеточные антигены неактивированным Т-клеткам. Антиген-презентирующие клетки фагоцитируют антиген, процессируют его, перерабатывая макромолекулярный антиген в пептидные фрагменты и экспонируют их на поверхность в комплексе с МНС II класса. Когда CD4+ Т-хелперы распознают такой комплекс, это не приводит к гибели АПК, напротив, Т-клетки начинают секретировать различные цитокины и активировать другие иммунные клетки.

Во взаимодействие с МНС-пептидным комплексом вовлечены два Т-клеточных рецептора – ТКР и корецепторы CD4 или CD8. Корецепторы CD4 и CD8 являются посредниками взаимодействия с МНС класса II и I соответсвенно и увеличивают аффинность ТКР-антигенного взаимодействия. Сам ТКР имеет три специальных области для связывания с МНС - гипервариабельные участки (CDRs). CDR1 и CDR2 кодируются V-сегментом и обеспечивают взаимодействие с аминокислотными остатками молекул МНС. В свою очередь гипервариабельный участок CDR3 кодируется V, D (только β TRB) и J регионами. Он образует петлю антигенсвязывающей области рецептора, локализирован по центру контакта ТКР и МНС и взааимодействуют с центральной частью пептида.

## **1. 5 V(D)J – реаранжировка**

Локусы генов Т-клеточных альфа рецепторов (TRA) и бета рецептора (TRВ) расположены на хромосомах 7 и 14 соответственно. Оба эти локуса состоят из повторяющихся сегментов, принадлежащих к трём классам: V*(variable*), J (*joining*), D (*diversity*). Генные сегменты и их число записаны в Таблице 1.

***Таблица 1.*** *Генные сегменты и их число (IMGT)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TCR цепь | Число сегментов | | | | Позиция на хромосоме |
| V | D | J | C |
| **α** | 49 (55) | - | 61 | 1 | 14q11.2 |
| **β** | 42 (77) | 2 | 14 | 2 | 7q34 |
| **γ** | 6 (15) | - | 5 | 2 | 7p14 |
| **δ** | 8 | 3 | 4 | 1 | 14q11.2 |

TRA локус состоит из 55 V-сегментов, 61 J-сегмента и только 1 С-сегмента. Локус Т-клеточного дельта рецептора (TRD) который экспрессирует δ-цепь γδ T-клеток расположен внутри TRA локуса [2]. Такая локализация определяет направление развития клетки: если TRA локус подвергнется рекомбинации, то такая клетка станет αβ-, а не γδ-лимфоцитом.

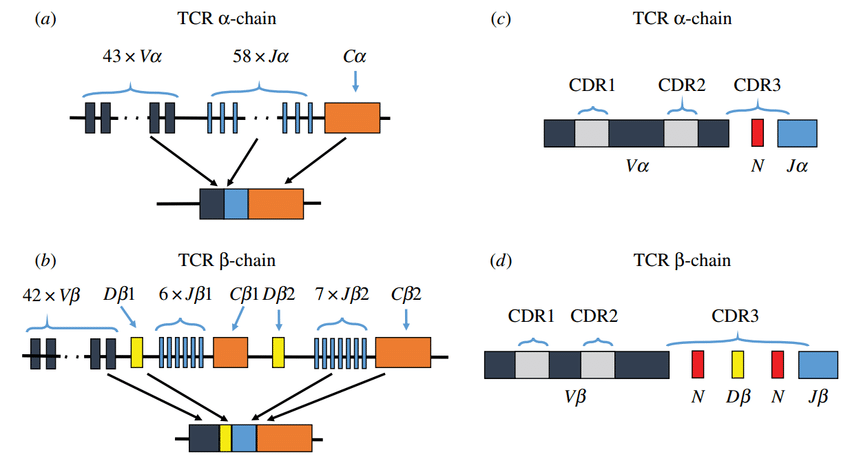
TRB локус отличается от TRA тем, что имеет 2 D-сегмента (отсутствующих в локусе TRA) и 2 С-сегмента (1 в TRA). Таким образом, геномные локусы зрелых Т-лимфоцитов состоят из V- и J-сегментов для α-цепи и V, D, J сегментов для β-цепи.

V(D)J – рекомбинация генов рецепторов лимфоцитов начинается с экспрессии белков V(D)J-рекомбинационного комплекса - димер рекомбиназ (экзонуклеаз) RAG-1/RAG-2, ДНК-зависимой протеинкиназы, ДНК-лигазы IV, терминальной дезоксинуклеотидилтрансферазы (TdT), кодирующей нематричный синтез олигодезоксинуклеотидов, гетеродимера HMG ½, гетеродимера Ku70/Ku80 [30]. Реаранжировка происходит по сигнальным последовательностям (RSS), которые фланкируют все генные сегменты. RSS представлены консервативными последовательностями: полимером из семи звеньев CACAGTG, полимером из девяти звеньев ACAAAAACC и 12 или 23 спейсерами между ними. Рекомбинация возможна только между RSS с разными спейсерами (“правило 12/23”). RSS с 23 спенсерами расположены на 3’- конце V-сегментов, в то время как 12-RSS – на 5’-конце J-сегментов. Таким образом, “правило 12/23” обеспечивает правильную ориентацию рекомбинации, позволяющую только V-J, но не V-V или J-J соединения.

При рекомбинации V(D)J также происходит делеция и вставка нуклеотидов в J-регионе. По разным оценкам, около 60% разнообразия Т-клеточного репертуара опосредовано данными событиями [32]. Именно в J-сегменте расположены антигенсвязывающие последовательности CDR3 α- и β-цепей. С одной стороны, такая стохастичность в перестройке ТКР лежит в основе широкого спектра антигенраспознающих рецепторов, с другой не все не получающиеся в результате V(D)J – рекомбинации последовательности кодируют функциональный рецептор. Делеции и вставки приводят к сдвигу рамки считывания в двух из трех случаев, увеличивая вероятность появления стоп-кодона внутри CDR3. Более того, часть генных сегментов в TRA и TRB являются псевдогенами, которые участвуют в рекомбинации, но не приводят к экспрессии ТКР.

Генерация TCR начинается с рекомбинации TRB. Сначала происходит рекомбинация сегментов D и J, затем сегмент V присоединяется к DJ. Если рекомбинация прошла успешно, синтезируется β-цепь TCR, которая образует комплекс с суррогатной α-цепью и экспонируется на поверхности клетки в виде пре-ТКР. После этого события дальнейшие перестройки TRB прекращаются на обеих хромосомах и тимоциты переходят к следующей стадии дифференцировки. В случае, если такой вариант реаранжировки сегментов кодирует нефункциональную β-цепь, то происходит второй раунд V(D)J-рекомбинации с использованием второго кластера сегментов D2-J2 (если они не участвовали в первом раунде реаранжировки) или задействуется TRB локус гомологичной хромосомы. Если клетка не может продуцировать функциональный пре-TCR, она подвергается апоптозу.

Прежде чем произойдет рекомбинация локуса TRA, Т-клеточный предшественник претерпевает несколько раундов деления. Следовательно, существуют ТКР с одинаковой β-цепью и с разной α-цепью, но противоположный случай невозможен [33-35].



***Рисунок 5.*** *Организация локусов TRB и TRA (цит. по [Laydon, Bangham, Asquith, 2015]). a, b – Схема процесса V(D)J – рекомбинации локусов генов a- и b-цепей TCR. Количество сегментов определенного типа в геноме показано рядом с названием сегмента. c, d – структура CDR1, CDR2 и CDR3 в перестроенном локусе. Сегменты на рисунке обозначены цветами: V-синий, D-желтый, J-голубой, C-оранжевый. Красным цветом помечены случайные вставочные нуклеотиды (N).*

VJ-рекомбинация генного локуса TRA происходит одновременно на обеих хромосомах. Структура локуса организована так, что в случае непродуктивной рекомбинации происходит серия последовательных рекомбинаций с задействованием все большего количества дистальных сегментов одной и той же хромосомы. В результате функциональная α-цепь имеет больше шансов сформироваться, чем β-цепь. Реаранжировки α-цепи прекращаются только после того, как произошел позитивный отбор. Так как в отличие от β-цепи, α-цепь не подвергается аллельному исключению, многие Т-клетки могут оказаться продуктивными в реаранжировках двух α-цепей [36]. В то же время, значительное большинство таких клеток будут иметь только один функциональный рецептор, потому что только небольшая группа рецепторов проходит позитивный отбор.

## **1.6 Механическая модель сборки ТКР**

Сборка каждого ТКР может быть представлена как совокупность раундов рекомбинации и может быть описана математической моделью. Механическая модель реаранжировки α и β цепи позволяет оценить вероятность сборки данного ТКР и имеет широкое практическое применение.

Рассмотрим эту модель на примере сборки TRB [32]. Чтобы сформировать данную нуклеотидную последовательность CDR3, должен произойти определенный сценарий рекомбинации. Этот сценарий представлен как набор независимых событий (Е) и имеет вероятность *Precomb.* Эта вероятность рассчитывается как произведение независимых событий рекомбинации (Формула 1):

1. Выбор определенного V, D и J сегмента
2. Число делеций на 3’ конце V-гена (*delV*)
3. Число делеций на 5’ конце J-гена (*delJ*)
4. Число делеций на концах D-гена (del5’*D*, del3’*D*)
5. Инсерция специфической нуклеотидной последовательности (x1…xins*DJ*) длинной ins*DJ* на месте соединения между D и J сегментами и последовательности (x1…xinsDJ) длиной insDJ в месте соединения сегментов D и J.

Изображение выглядит как текст

Автоматически созданное описание

Репертуар ТКР формируется не только в результате V(D)J-рекомбинации, но и под влиянием позитивного и негативного отборов. Для получения параметров данной модели, последовательности ТКР анализируются без учета событий селекции. Вероятностная модель сборки была построена на основе данных репертуаров нефункциональных вариантов TCR - последовательностей CDR3 со сдвигом рамки считывания или содержащие преждевременный стоп-кодон. Такие последовательности формируются как результат непродуктивной V(D)J рекомбинации на одной из гомологичных хромосом Т-клетки.

Существует множество способов комбинации сегментов V, D, J которыми можно получить одну и ту же нуклеотидная последовательность CDR3 (σ). Поэтому, чтобы оценить вероятность генерации данного CDR3 (*Pgen*) необходимо суммировать все вероятности событий рекомбинации, которыми может быть получена данная нуклеотидная последовательность CDR3 (Формула 2).

Изображение выглядит как текст

Автоматически созданное описание

Такая модель позволяет оценить вероятность генерации любого CDR3, которая определяется механизмом рекомбинации, но не зависит от событий отбора.

*Pgen* широко используется в современных методах исследования репертуара Т-клеток. В частности, позволяет выявить ТКР, частота встречаемости которых в определенной группе образцов (разные группы людей или разные Т-клеточные фракции) значимо отличается от ожидаемой вероятности (*Pgen*) их образования в результате V(D)J – рекомбинации.

## **1.7 Изучение репертуара Т-клеток методами высокопроизводительного секвенирования**

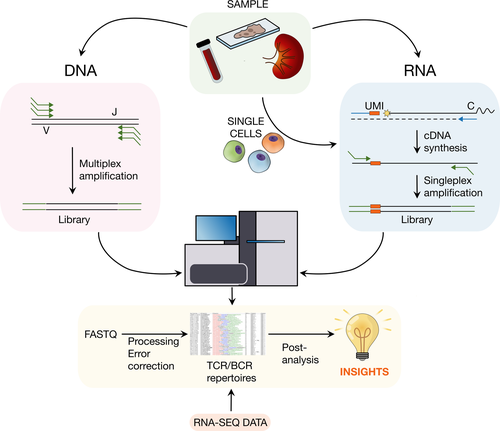
Первые попытки изучения первичной структуры TCR были основаны на спектротипировании с дальнейшим секвенированием по Сенгеру. Однако такой подход позволяет идентифицировать только перепредставленные в репертуаре клоны Т-клеток.

С появлений технологий высокопроизводительного секвенирования появилась возможность более глубокого анализа репертуаров Т-клеточных рецепторов.

В настоящее время существует несколько методов получения библиотеки TCR. Среди них, стандартные методы полнотранскриптомного секвенирования (RNAseq), в том числе секвенирования РНК одиночных клеток (single-cell RNA-Seq). Однако в библиотеке тотальной мРНК последовательности TCR представляют лишь небольшую фракцию. Принципиально другой подход заключается в использовании библиотек, обогащенных последовательностями V-сегментов TCR. Существует несколько стратегий получения библиотеки антиген-распознающих рецепторов [48–50], которые можно разделить на 2 группы в зависимости от исходного субстрата для постановки реакции - ДНК [51–53] или РНК [54–59] (см. 7)

Основанные на РНК методы используют 5′-RACE амплификацию [61] или лигирование одноцепочечных фрагментов РНК [62, 63] для введения в ПЦР универсальных праймеров. Благодаря использованию уникальных молекулярных идентификаторов (UMI) стала возможна коррекция ошибок секвенирования и доступна информация о количестве различных TCR [64]. Однако этот метод не учитывает различия в уровне экспрессии РНК в разных типах клеток, что необходимо при изучении репертуара BCR.

Основной недостаток ДНК методов – необходимость использования фиксированного набора праймеров к V и J сегментам для амплификаци рецептора. С одной стороны, это накладывает определенные ограничения на метод, так как с помощью него нельзя обнаружить новые варианты аллелей V. С другой стороны, может привести к преобладающей амплификации одних аллелей над другими и, как следствие, смещению данных.



***Рисунок 6****. Методы подготовки ТКР библиотеки. Включают в себя два направления методов в* зависимости от *исходного субстрата для постановки реакции - ДНК или РНК. Т-клеточные репертуары могут быть также получены из тотального RNAseq с последующей обработкой специализированными алгоритмами.*

Недавно разработанные методы позволяют получать парные альфа- и бета-репертуары TCR для одиночных клеток [49, 59, 67, 68]. В настоящее время применяют эмульсионные системы для получения парных ТКР с использованием до 10000 клеток с выходом более 50% [67]. Такие системы позволяют одновременно получать библиотеки TCR и тотальную мРНК. Клетки также могут быть помечены антителами или пептид-MHC-мультимерами, которые несут нуклеотидный баркод для получения дополнительной фенотипической информации о клетке.

Для обработки данных секвенирования TCR и BCR разработано различное программное обеспечение [69-73]. Главная особенность обработки таких сырых данных заключается в необходимости использования специальных алгоритмов сборки ТКР. На выходе такие алгоритмы генерируют таблицы содержащие последовательности CDR3, их представленность и соответствующие V- и J-сегменты. Строки таких таблиц называются «клонотипами» или «клонами». Некоторое программное обеспечение также может извлекать клонотипы TCR и BCR из данных тотального секвенирования мРНК — как массового, так и секвенирования одиночных клеток.

Таким образом, результатом секвенирования репертуаров Т-клеточных рецепторов, является таблица, в которой можно найти информацию о количестве клонов, аминокислотной и нуклеотидной последовательности. Как правило, существует несколько подходов для дальнейшего анализа репертуаров (см. рис. 8):

1) Изучить распределения клонов в репертуаре, т.е. оценить разнообразие и клональность

2) Исследование состава репертуара CDR3 в динамике (например, до и после вакцинации)

3) Исследование состава репертуара CDR3 в зависимости от локализации (например, между кровью и тканями)

4) Исследование состава репертуара CDR3 в разных группах (например, среди лиц с одним и тем же аутоиммунным заболеванием).

## **1. 8 Различные метрики для изучения распределения клонов**

Теоретически оцененное многообразие бета ТКР превышает экспериментально наблюдаемое на много порядков (1014 и 108  соответственно) [32]. Таким образом, существующие разрешение секвенирования позволяет идентифицировать только часть реально существующего репертуара, состоящей из фракции наиболее представленных в образце ТКР.

Для оценки разнообразия Т-клеточных репертуаров используются выборочные статистики, характеризующие распределение частот встречаемости секвенированных клонов [83-86]. Выборочная статистика позволяет описать множество компонентов структуры репертуара одним значением. Это значительно упрощает сравнение репертуаров, принадлежащих к разным биологическим группам. Наиболее широко применяемая метрика — это индекс энтропии Шеннона и его модификации [83]. Шеннон определил энтропию опыта Н, как усредненное значение неопределенности отдельных исходов. В таком случае, энтропия равна нулю, если вероятность одного из событий равна 1, и принимает максимальное значение, если все исходы равновероятны. В частном случае анализа репертуара ТКР, индекс Шеннона принимает значение 1, если все клонотипы встречаются с одинаковой частотой. Примером такого равномерного распределения в репертуаре является распределение наивных Т-клеток. Когда человек сталкивается с патогеном, клоны антиген-специфичных наивных клеток претерпевают экспансию, что приводит к изменению плотности распределения ТКР. Другие индексы разнообразия, например индекс Симпсона или индекс Чао также могут быть использованы для оценки подобных изменений плотности распределения ТКР. Важным клиническим признаком иммунного ответа является клональность лимфоцитов – наличие в репертуаре крупных популяций Т-клеток памяти, происходящих от единичных клеток. Таким образом, в общем случае, чем меньше разнообразие репертуара, тем выше его клональность. Например, высокая клональность репертуаров ТКР может наблюдаться у пациентов с аутоиммунными заболеваниями: в пробах крови пациентов с системной красной волчанкой (СКВ) [87], в образцах псориатически пораженной кожи [88], у пациентов с болезнью Крона [89]. Также метрики разнообразия репертуаров могут использоваться, в качестве маркеров успешной иммунотерапии. Высокая клональность репертуаров Т-клетки в опухолях наблюдается у пациентов, хорошо реагирующих на анти-PD1 терапию, но не у пациентов, которые получали анти-CTLA-4 терапию [90-92]. С другой стороны, высокое разнообразие репертуаров Т-клеток в периферической крови после анти-CTLA-4 терапии связано с серьезными побочными эффектами у пациентов с раком простаты [93, 94] и меланомой [95].



В общем случае произвольного опыта с ***k*** исходами, имеющими вероятности ***P1***, ***P2***, …, ***Pk***

Несмотря на широкое применение стандартных метрик разнообразия, они имеют целый ряд недостатков. Они не дают представления о природе происхождения различий между изучаемыми группами: различные биологические эффекты могут существенно влиять на разнообразие ТКР репертуаров. Например, разнообразие репертуара значительно уменьшается с возрастом [96] и при наличии хронических заболеваний [97].

## **Публичные клоны**

Ранее считалось, что вероятность встретить одинаковые последовательности ТКР в крови разных доноров крайне мала. Для их появления должны быть соблюдены следующие условия:

1. Должна произойти серия одинаковых событиях соматической рекомбинации. При этом каждое такое событие выбирается из очень большого набора вариантов.
2. Полученные клоны должны пройти тимусную селекцию
3. Они должны сохранить жизнеспособность в периферических органах иммунной системы.
4. Соответствующие наивные Т-клетки должны иметь достаточно многочисленную популяцию, чтобы выдержать конкуренцию с другими клетками, несущие специфичный к тому же самому антигену ТКР.

Если бы наступление всех сценариев реаранжировки имело равную вероятность, то появление одинаковых последовательностей в репертуарах разных доноров было бы очень редким событием и только при ответе на идентичный антиген. Тем не менее, общие ТКР могут быть найдены в разных группах доноров при иммунном ответе разной специфичности [98-101]. Более того, пересечение репертуаров наивных Т-клеток наблюдается как у людей [102], так и у мышей [103].

Для объяснения существования публичных был выдвинут ряд гипотез.

Первая группа гипотез основана на 3D структуре ТКР [104]. Основная идея состоит в том, что селекция ТКР опосредуется структурой комплекса МНС-пептид, который создает специальное окружение для экспансии публичных клонов [104]. Альтернативная теория состоит в том, что сама структура публичных ТКР способствует клональной экспансиии [105]

Другая группа гипотез основана на предположение, что наступление одних рекомбинационных сценариев более вероятно, чем других. Вследствие чего, они реализуются у большинства лиц, в популяции. Некоторые нуклеотидные последовательности могут быть получены с большей вероятностью потому, что имеют простой сценарий рекомбинации[102], например меньшее число инсерций или делеций в местах соединения сегментов . Тем не менее, это не объясняет, почему появляются ТКР с одинаковой аминокислотной последовательностью, но разными нуклеотидными последовательностями в периферической крови у отдельно взятых людей [101, 106].

Также существование публичных TCR можно объяснить механизмом конвергентной рекомбинации - продукцией вариантов TCR с разными нуклеотидными последовательностями, но с одинаковой последовательностью аминокислот. Была обнаружена прямая зависимость между частотой появления TCR и количеством вариантов нуклеотидных последовательностей, которыми кодируется соответствующая аминокислотная последовательность [107]. Таким образом, согласно этой модели, межиндивидуальное пересечение репертуаров напрямую связано с частотой генерации данной аминокислотной последовательности TCR, что в свою очередь предопределяется конвергенцией различных рекомбинационных событий, которые приводят к появлению идентичных нуклеотидных последовательностей. Также было показано, что конвергентная рекомбинация влияет как на межиндивидуальный, так и на индивидуальный репертуар – конвергентные β-цепи TCR наивных CD8+ T-клеток преобладают в индивидуальных репертуарах у мышей [108].

Данные показывают, что наиболее важными факторами участвующими в продукции общих TCR являются конвергентная рекомбинация [107-109] и повторение одних и тех же простых события во время рекомбинации, характеризуются минимальным количеством случайно удаленных и добавленных нуклеотидов. [102, 109]. Наблюдаемое фактически пересечение аминокислотных репертуаров β-цепей TCR между двумя случайно выбраним индивидуумами превышает 105 вариантов [\*\*]

\*\*Shugay M., Bolotin D.A., Putintseva E.V., Pogorelyy M.V., Mamedov I.Z., Chudakov D.M. et al. (2013). [Huge overlap of individual TCR beta repertoires](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3872297/). Front. Immunol. **4**, 466;

Несмотря на то, что тимусный отбор существенно сужает многообразие TCR, как было показано, не участвует в формировании общих репертуаров[114, 115].

## **1.10 Общая характеристика SARS-CoV-2**

Первый случай COVID-19 в Китае был официально выявлен в декабре 2019 года. Новый, ранее не идентифицированный вирус, получивший название SARS-CoV-2, вызвал вспышку тяжелой вирусной пневмонии и к Апрелю 2021 унес жизни почти трех миллионов людей по всему миру.

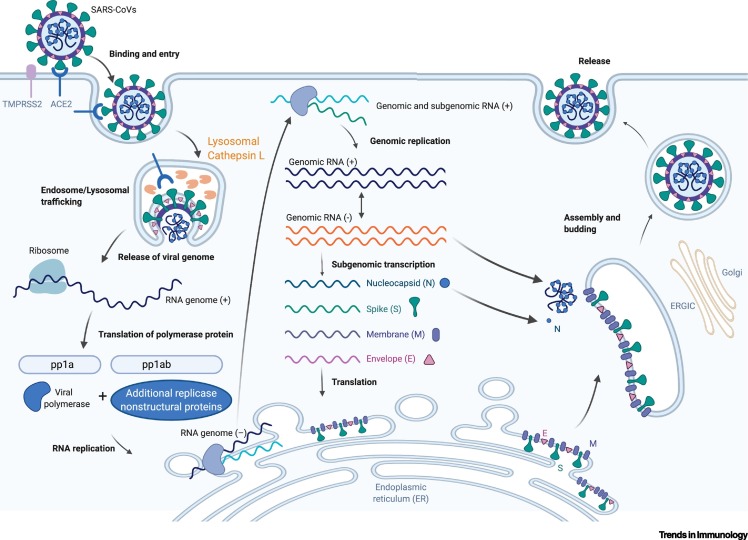
CoV семейства Coronaviridae представляют собой оболочечные одноцепочные (+) РНК-вирус [5]. Все высокопатогенные CoV, включая SARS-CoV-2, относятся к роду Betacoronavirus [5]. Последовательность генома SARS-CoV-2 на ~80% идентична последовательности SARS-CoV и на ~50% MERS-CoV [1,6]. Его геном содержит 14 открытых рамок считывания (ОРС), две трети которых кодируют 16 неструктурных белков (nsp 1–16), составляющих репликазный комплекс [6,7]. Оставшаяся треть кодирует девять дополнительных белков (ORF) и четыре структурных белка: спайковый белок — шип, представленный гликопротеином (S), оболочечный белок (E), мембранный белок (M) и нуклеокапсид (N), при этом спайковый белок обеспечивает проникновение SARS-CoV в клетку-мишень. 8]. Однако ген S SARS-CoV-2 сильно отличается от SARS-CoV, имея идентичность нуклеотидов <75% [1,6,9]. S-белок состоит из двух функциональных субъединиц: S1 (содержащей RBD-домен) обеспечивает связывание с рецептором на поверхности клет­ки хозяина - ангиотензинпревращающим ферментом 2 (ACE2) и S2 способствует слияние мембран. Этим двум событиям предшествует сайт-специфичный гидролиз по S1/S2, осуществляемого локализованным в инфицируемой клетке трансмембранной сериновой протеазой TMPRSS2 и (или) эндолизосомальным катепсином. [1,9,10]. (рис. 1)



***Рисунок 7****. Схема строения* SARS-CoV-2. *Spike (S) — шип представленный гликопротеином, Nucleocapsid (N) — нуклеокапсидный белок, Membrane (M) — мембранный белок, Envelope(E) — оболочечный белок, RNA — одноцепочечная РНК*

[10]. Как только геном высвобождается в цитозоль хозяина, начинается трансляция ORF1a и ORF1b в белки вирусной репликазы, которые затем посредством гидролиза фрагментируются на отдельные nsps (посредством протеаз хозяина и вируса: PLpro). Они образуют РНК-зависимую РНК-полимеразу (nsp12 происходит от ORF1b) [8]. Здесь компоненты репликазы перестраивают эндоплазматический ретикулум (ER) в двухмембранные везикулы (DMV), которые облегчают вирусную репликацию геномных и субгеномных РНК (sgRNA). Последние транслируются в дополнительные и вирусные структурные белки, которые облегчают образование вирусных частиц (рис. 1) [11,12].

Взаимодействие между спайковым белком и ACE2 приводит к лизосомной деградации последнего, что считается критическим этапом патогенеза COVID-19. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4231883/] ACE2 является отрицательным регулятором брадикинина, и своим отсутствием активирует выработку брадикинина, который активирует кининовый каскад, что приводит к воспалительному процессу и нарушению свертываемости крови. Брадикинин влияет на увеличение сосудистой проницаемости и стимуляцию вазодилатации. Все это провоцируют локальный отек Квинке в пораженных тканях дыхательного аппарата.

****

***Рисунок 8.*** *Проникновение SARS-CoV-2 в клетку включает следующие этапы: (1) гидролиз сайта S1/S2 поверхностной трансмембранной трансмембранной сериновой протеазой TMPRSS2; и/или (2) эндолизосомальным катепсином L, который опосредует слияние вируса с клеточной мембраной на клеточной поверхности и в эндосомальных компартментах, соответственно. (3) Геном РНК (+) высвобождается в цитозоль, где он транслируется в белки-репликазы (рамки считывания ORF1a и ORF1b). (4) Полипротеины (pp1a и pp1b) расщепляются кодируемой вирусом протеазой на отдельные неструктурные белки репликазного комплекса (nsps) (включая РНК-зависимую РНК-полимеразу: RdRp). (5) Репликация в двойных мембранных везикулах (DMV) эндоплазматического ретикулума (ER). (6) Геномная репликация в (+) полярную РНК (7) Трансляция структурных белков (N, S, M и E); (8) Структурные белки комбинируются с нуклеокапсидом в промежуточный компартмент (ERGIC); (9) Формирование зрелого вириона; (10) Экскреция вируса посредством экзоцитоза*

**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7556779/**

## **1.11 Иммунный ответ против** **SARS-CoV-2**

### *1.11.1*

Система врожденного иммунитета, является одним из первых барьеров на пути вторжения в организм патогена. CoV проникает в организм через респираторную систему, где вирус вступает во взаимодействие с респираторным эпителием.

Связавшись с рецептором клетки – хозяина SARS-CoV-2 транспортируется в эндосому, где РНК вируса выходит из каписда и распознается эндосомальными рецепторами TLR3 и TLR7. Затем сигнал передается TRIF и TRAF6 в направлении комплекса I-κB, в результате диссоциации которого образуется NF-κB. Это транскрипционный фактор отвечает за синтез широкого спектра провоспалительных цитокинов, в частности, таких как proIL-1, IL-6, proIL-18, TNFα [9,16] Другой сигнал передается через адаптерный белок TRAF3 к белкам семейства IRF, включающее, по меньшей мере, IRF3 и IRF7. Они подвергаются фосфорилированию, мигрируют в ядро клетки и инициируют синтез IFN I типа [16, 46]. Если распознавание сигнала РНК вируса происходит TLR7 он трансдуцируется по классическому пути [10]. Другой путь внедрения CoV в клетку —непосредственное слияние с клеточной мембраной. В таком случае вирион высвобождает одноцепочечную РНК, содержащей две открытых рамки считывания ORF1a и ORF1ab, на которых транслируются полипротеины pp1a и pp1ab. В результате посттрансляционных изменений из них образуются 12 неструктурных белков, расположенных в перинуклеарном пространстве клетки. Неструктурные белки образуют комплекс РНК-репликазатранскриптаза, который обеспечивает синтез отрицательной цепи РНК [59, 64]. Кроме процессов репликации, РНК распознается цитоплазматическими рецепторами RLR и MDA 5, трансдуцирующими нисходящий сигнал через MAVS на протеиновый ансамбль TRAF-TANKIKKε-TBK 1, далее на IRF3, IRF7 и NF-κB с последующим импортом в ядро и синтезом IFN I типа и провоспалительных цитокинов [16, 44]. Таким образом, в клетке существует разветвленная сеть передачи сигналов, позволяющая секретировать широкий набор цитокинов, которые в зависимости от вирулентности возбудителя и функционального состояния инфицированных клеток хозяина, могут выполнять либо защитную функцию, либо приводить к нарастанию неконтролируемого воспаления с последующим развитием тяжелых патологических реакций вплоть до гибели инфицированного организма [7, 11, 13, 14, 21, 34, 38].

file:///C:/Users/%D0%94%D0%B0%D1%88%D0%B0/Downloads/1440-8812-1-PB%20(1).pdf

### *1.11.2 Адаптивный иммунный ответ против SARS-CoV-2*

Адаптивный иммунный ответ, включающий В- и Т-лимфоциты, формируется в течение нескольких дней после контакта с патогеном. Его главные особенности — это высокая специфичность к патогену и длительная иммунная память. Первичным событием для такого ответа является так называемая презентация антигена, экспонирвоание на поверхности антиген презентирующей клетки (например макрофаги, дендритные клетки) молекулы антигена связанного с MHC I класса и его распознавание Т-клеточным рецептором. Инфицированные SARS-CoV-2 АПК, протеолитически расщепляют интактные вирусные белки внутри протеосомы. Затем пептидные фрагменты этих белков транспортируются главным комплексом тканевой совместимости (MHC класса I) в составе экзосом и реэкспрессируются на поверхность клетки. Эти пептидные фрагменты называются антигенные детерминанты, или эпитопы, которые распознаются Т- или В-клетками. После распознавания и активации, Т-клетки подвергаются клональной экспансии и дифференциируются в эффекторные клетки: CD8+ Т-клетки киллеры или CD4+ Т-клетки хелперы. В то время, как CD8+ T-клетки отвечают за уничтожение инфицированных клеток, CD4+ T-клетки выступают в качестве регуляторов иммунного ответа: они активируют макрофаги и В-клетки, определяют лимит иммунных реакций, вырабатывают множество цитокинов, чтобы рекрутировать другие клетки. В-клетки могут быть активированы таким же образом, но этот процесс часто запускается Т-клетками. В-клетки так же подвергаются клональной экспансии и дифференциируются в плазматические клетки, способные к продукции антител, или в В-клетки памяти, которые ответственны за иммуный ответ в случае повторного столкновения патогеном. Что примечательно, В-клетки после рапознавания антигена могут подвергаться процессу соматического гипермутирования, чтобы повысить свою специфичность к антигену.

Следы Т-клеточного ответа обнаруживаются в большинстве случаев после инфицирования SARS-CoV-2. Исследования показывают, при первичном инфицирование ключевую роль в иммунном ответе играют CD4+ клетки.

В настоящее время наиболее детально изучен Т-клеточный иммунный ответ на спайковый белок (S). Все экспериментальные вакцины основаны на формирование искусственного активного иммунитета против S-белка. С помощью биоинформатического анализа было показано, что другие белки SARS-CoV-2, в частности нуклеокапсидный (N) белок, мембранный белок (М) и nsp3 (вирусная хеликаза) могут стать подходящей мишенью для вакцины [122].

Анализ CD4+ T-клеточный ответа у пациентов, перенесших COVID показал, что он формируется против 21 из 25 белков SARS-CoV-2 [123]. Во время вирусной инфекции Т- хелперы преимущественно дифференциируются в Th1 и фолликулярные хелперы Т-клетки (Tfh). Th1 клетки продуцируют ИФН-гамма чтобы активировать макрофаги, в то время как фолликулярные Т-клетки специализирующейся на оказании помощи В-клетки и индуцируют продукцию антител. Интересно, что именно увеличение фракции циркулирующих Tfh связывают с наличием инфекции [125]. Что касается CD8+ T-клеток, то их присутствие связано с благоприятным клиническим исходом. Основная фракция CD8+ T-клеток специфична для одного из структурных протеинов SARS-CoV-2: S, N или M.

Формирование Т-клеточного иммунного ответа происходит в большинстве случаев инфицирования COVID-19 и

Данные показывают (\*), что 90% людей, инфицированных вирусом SARS-CoV-2, формируют устойчивый Т-клеточный иммунный ответ, несмотря на то что антитела были выявлены только у 60% из них. Анализ на антитела не может дать полноценной картины иммунного статуса человека. С целью комплексной оценки иммунного ответа, совместно с тестами на оценку уровня антител, следует идентифицировать Т-клетки, специфично отвечающие на антигены вируса SARS-CoV-2. Понимание механизмов Т-клеточного ответа, является особо важным, поэтому данное исследование посвящено Т-клеточным рецепторам и их участию в иммунном ответе против SARS-CoV-2.

\*. Sekine T et al. Robust T Cell Immunity in Convalescent Individuals with Asymptomatic or Mild COVID-19. Cell. 2020 Oct 1;183(1):158-168.e14. doi: 10.1016/j.cell.2020.08.017.

# 

# **ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ**

## **2.1 Доступные данные**

В текущем исследовании были использованы данные иммуносеквенирования гипервариабельной области локуса TCRβ (TCRβ RepSeq). Доступные данные включают в себя три набора данных (Таблица):

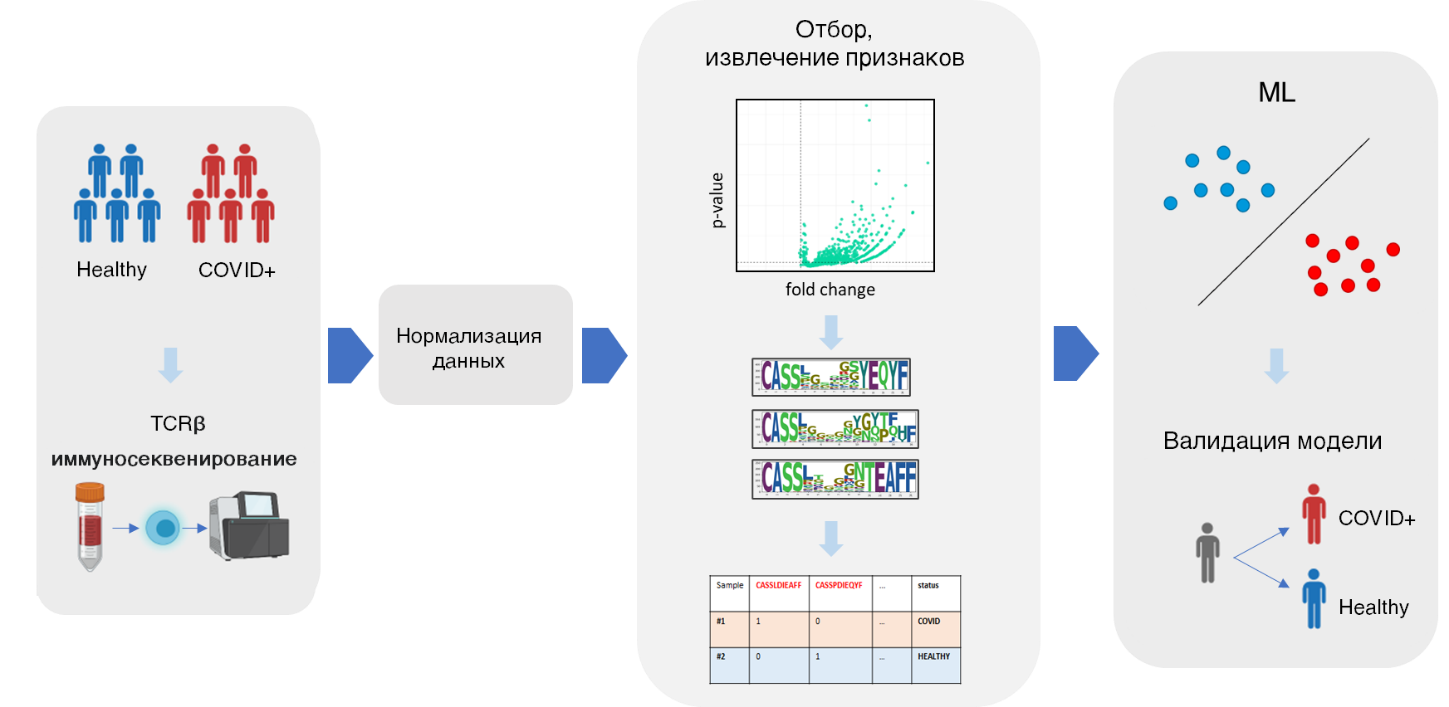
(1) Данные репертуаров Т-клеточных рецепторов здоровых доноров (AB\_2017) [\*] (доступны по ссылке [https://clients.adaptivebiotech.com/immuneaccess]), полученные до января 2020 года. Следовательно инфицирование таких субъектов вирусом SARS-CoV-2 исключено. Этот набор данных включают в себя две группы образцов HIP и KECK, состоящих из 666 и 120 проб соответственно.

(2) Данные AB\_2020 **[\*]** включают в себя 1414 образцов, полученных от доноров, которые к моменту взятия проб либо имели контакт с инфицированными, либо были инфицированы, либо перенесли COVID-19 ([https://clients.adaptivebiotech.com/pub/covid-2020]).

(3) Данные Rus\_2020, включающее в себя 505 образцов COVID+ пациентов и 446 образцов COVID- доноров.

Наборы данных содержат информацию о последовательностях CDR3 с соответствующим числом ридов (горизонтальной глубиной клона, числом копий), а также информацию о V-, D- и J-сегментах, встречающихся у данного клона.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Источник | Батч секвенирования | Число доноров COVID | Число здоровых доноров | Общее число доноров | Число TCRb CDR3 клонов | Число TCRb CDR3 прочтений | Обучающие и тестовые выборки |
| Rus\_2020 | Rus\_batch1 | 110 | 0 | 110 | 9099±4198 | 4359760±3525257 | Rus1 |
| Rus\_2020 | Rus\_batch2 | 33 | 48 | 81 | 12985±7863 | 2514178±1063350 | Rus1 |
| Rus\_2020 | Rus\_batch3 | 143 | 295 | 438 | 12697±7512 | 4076747±2404111 | Rus1 |
| Rus\_2020 | Rus\_batch4 | 120 | 0 | 120 | 9174±5285 | 5156480±4849578 | Rus2 |
| Rus\_2020 | Rus\_batch5 | 99 | 0 | 99 | 9289±3718 | 3831749±3025834 | Rus2 |
| Rus\_2020 | Rus\_batch6 | 0 | 103 | 103 | 32223±16142 | 5069993±4194279 | Rus2 |
| AB\_2020 | NIH/NIAI | 357 | 0 | 357 | 73453±58103 | 123921±101375 | adaptive1 |
| AB\_2020 | HUniv12Oct | 177 | 0 | 177 | 160588±112465 | 328975±198249 | adaptive1 |
| AB\_2020 | COVID19-Adaptive | 37 | 0 | 37 | 331591±127963 | 518031±183323 | adaptive1 |
| AB\_2020 | DLS | 433 | 0 | 433 | 161947±114241 | 292948±183785 | adaptive2 |
| AB\_2020 | ISB | 68 | 0 | 68 | 240094±172417 | 390878±249319 | adaptive2 |
| AB\_2020 | IRST/AUSL | 64 | 0 | 64 | 153487±135642 | 364108±270338 | adaptive2 |
| AB\_2017 | HIP | 0 | 665 | 665 | 190059±79128 | 3803929±2779138 | adaptive2 |
| AB\_2017 | KECK | 0 | 120 | 120 | 203475±110084 | 299828±185736 | adaptive1 |



***Рисунок 9****. Схема исследования. Забор образцов крови у людей с положительным и отрицательным статусом COVID; Иммуносеквенирование индивидуальных репертуаров ТКРβ; Контроль качества и нормализация данных; Отбор ассоциированных с SARS-CoV-2 клонов и извлечение признаков из данных – представление в виде вектора значений; машинное обучение и оценка качества предсказания полученных моделей.*

## **2.2 Нормализация образцов**

Нормализация является важным шагом предварительной обработки данных путем корректировки нежелательных смещений, вызванных преобладающей амплификацией определённых CDR3 на этапе приготовления библиотеки ТКР и различной глубиной секвенирования образцов.

## **2.3 Нормализация по V-usage**

Посредством поиска наиболее распространенные комбинации генов Variable-Joining можно выявить широко представленные в анализируемой группе клонотипы. Зачастую смещение частот вариантов V и J обусловлено не биологическими факторами, а условиями проведения эксперимента. В частности, использование разного микса праймеров в различных сериях постановки секвенирования может привести к значительному сдвига V, J –usage в соответствующих партиях образцов.

Нормализация по V-usage включает следующие шаги:

Чтобы исключить возможные ошибки идентификации генов TRBV, связанные с качеством секвенирования, (1) гены TRBV были сгруппированы в семейства генов (например, TRBV6-1 и TRBV6-4 объединялись в TRBV6).

(2) Была рассчитана частота всех встречающихся комбинаций V сегментов (V-usage) в каждом батче (Rus2020, AB2020, HIP, KECK) и затем медианное значение V-usage между батчами:

median (TRBV6) = (median (TRBV6\_in\_ Rus2020, TRBV6\_in\_ AB2020, TRBV6\_in\_ HIP, TRBV6\_in\_KECK))

(3) Рассчитывался шкалирующий коэффициент для каждого V сегмента в каждом батче. Таким образом после домножения частот всех клонотипов с данным V сегментом в данном батче на шкалирующий коэффициент, итоговый V-usage становился равным медианному.

(4) Сумма частот V-сегментов в каждом образце приводилась к единице (Сумма частот делилась на сумму изменённых частот).

## **2.4 Нормализация по глубине секвенирования**

На этапе, предшествующем нормализации был проведен контроль качества образцов на основе оценки количества прочтений клонов (глубина прочтений) и количества клонов в образце. Из дальнейшего анализа исключались образцы, показатели качества которых не удовлетворяли пороговым значениям: менее 6000 клонотипов в образце и суммарное число ридов менее 105.

Далее, суммировались частоты и количество ридов клонов, имеющих идентичные аминокислотными последовательностями. Клонотипы представленные единичным ридом исключались из анализа в целях фильтрации очень редких вариантов.

Для нормализации образцов использовали два подхода:

1. Из каждого образца отбирались 5000 клонов, имеющих наибольшее число ридов. Частоты соответствующих клонотипов пересчитывались.
2. Для того чтобы учесть относительную представленность публичных вариантов TCR, использовался альтернативный способ нормализаци. Из каждого репертуара предварительно отбирались только публичные варианты TCR. Для этого, из пула уникальных клонов Rus2020, AB2020, AB2017 были отобраны наиболее распространенные клонотипы, c частотой встречаемости более 0.02 (количество образцов, в которых была обнаружена данная аминокислотная последовательность CDR3, деленное на общее количество образцов). Затем из каждого образца отбирались только те клонотипы, которые встречались в полученном списке публичных клонов. После этого из нового публичного репертуара каждого образца отбирались 2000 клонотипов с наибольшим размеров. Частоты клонов пересчитывались.

## **2.5 Отбор признаков для машинного обучения**

Важным этапом машинного обучения являлся отбор признаков – последовательностей CDR3 и эффективного представления их в виде вектора значений. Строго говоря, перед нами стояла биологическая задача поиска клонотипов, а также паттернов объединяющих эти последователь6ности, с помощью которых можно было бы различать образцы здоровых доноров и переболевших COVID. Мы использовали две стратегии поиска значимых для модели CDR3.

(1) Первый подход основан на использование публичных клонов в качестве признаков для обучения. Он обоснован предположением, что репертуары TКР у пациентов перенесших короновирусную инфекцию близки по составу и по представленности публичных клонов. С другой стороны, образцы здоровых доноров также вероятно содержат общие специфичные клонотипы. Таким образом на вход классификатора подавалось множество (~10000) наиболее представленных среди анализируемых образцов публичных клонов. За меру встречаемости принимали количество образцов, в которых была обнаружена данная аминокислотная последовательность CDR3.

(2) Вторая стратегия отбора включает в себя предварительный этап поиска ассоциированных с COVID последовательных среди пула идентифицированных образцов и выявления наиболее распространенных мотивов последовательностей с помощью методов кластеризации. На вход модели подавалось относительно небольшое число (~2000) клонов с заранее известной специфичностью к COVID.

Для поиска последовательности TCRβ, ассоциированных с COVID, мы использовали U-критерий Манна-Уитни. Для каждого клонотипа, взятого из пула уникальных публичных клонов Rus\_2020, AB\_2020 и AB\_2017 рассчитывалась его частота в каждом из образцов. Затем с помощью U-критерия Манна-Уитни были выявлены ​​варианты, вес которых статистически значимо выше (порог p-значения <0,01) в выборках образцов со статусом COVID+.

## **2.6 Кластеризация признаков и поиск паттернов последовательностей**

Одной из задач нашего исследования был поиск групп клонов, ассоциированных с SARS-CoV-2, имеющих общие паттерны аминокислотных последовательностей. Для этого были взяты клонопипы полученные в пункте 2 раздела Материалов и методов «Отбор признаков для машинного обучения». После чего, было найдено расстояние (расстояние Хэмминга, см. «Извлечение признаков», глава «Материалы и методы», Пункт 2) между всеми парами элементов и построена матрица расстояний. Затем полученная матрица подавалась на вход алгоритма DBSCAN cо следующими параметрами: eps = 40, min\_samples = 15 и остальными параметрами по умолчанию.

## **2.7 Извлечение признаков**

Мы использовали два способа представления фичей в образце. В первом случае учитывалась частота встретившегося клона в образце, так называемое, взвешенное представление. В противоположном подходе вес клона не учитывается, оценивается только факт вхождения клонотипа в данный репертуар (0, если клонотип не найден, 1, если найден) – невзвешенное представление.

Кроме того, было необходимо учитывать тот факт, что последовательности, обладающие высокой гомологией, в большинстве случаев, имеют общую антиген-специфичность. Поэтому один лишь поиск клонов с одинаковой последовательностью привел бы к потере большой части информации о структуре репертуара. Ввиду этого, для выявления вариантов, содержащих те же паттерны, что и интересующий нас клон были использованы метрики различия последовательностей одинаковой размерности:

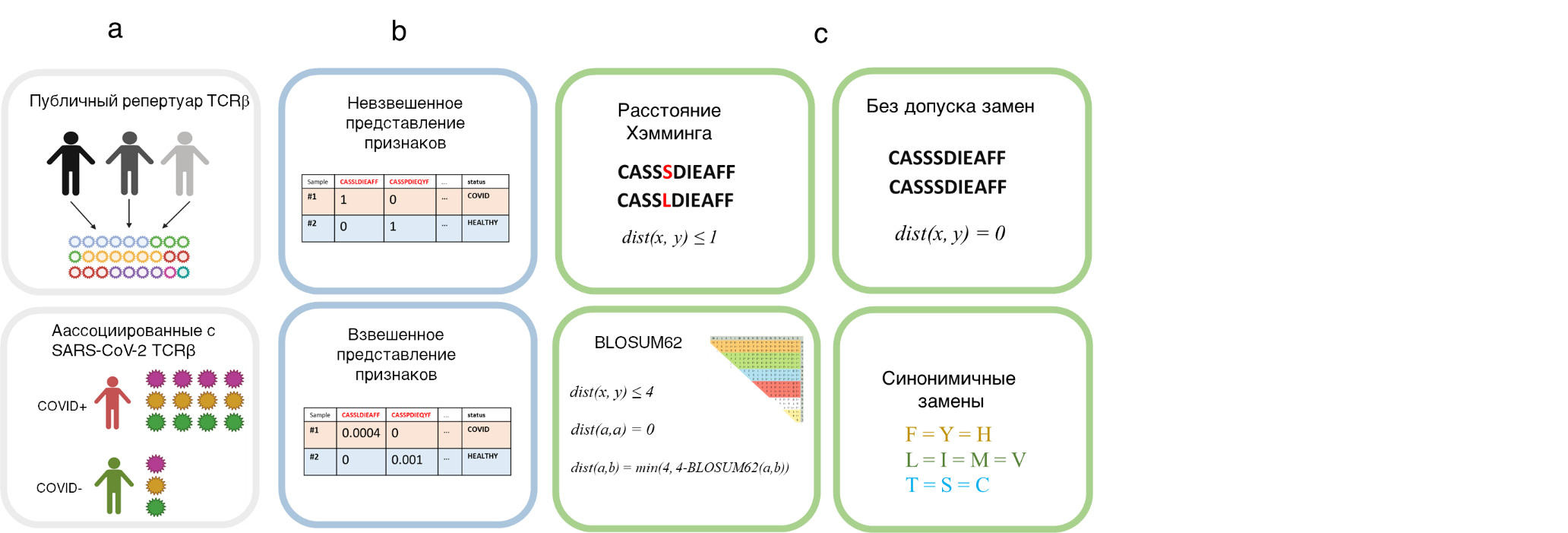
(1) Число позиций, в которых различаются соответствующие аминокислоты двух последовательностей одинаковой длины (расстояние Хэмминга, *d(x, y)*) должно быть меньше или равно 1. Таким образом, при расчете признаков допускалась единичная аминокислотная замена.

*d(x, y) ≤ 1*

(2) Расстояние между последовательностями CDR3 одинаковой длины (число различающихся позиций) рассчитывалось на основе матрицы аминокислотных замен BLOSUM62 следующим образом: dist(a,a) = 0; dist(a,b) = min(4, 4-BLOSUM62(a,b)), где 4 на 1 единицу больше, чем максимальная оценка по матрице BLOSUM62 для аминокислотных замен. Это приводит к уменьшению штрафа за замену для пар аминокислот с положительными значениями BLOSUM62 (например, dist(I,V)=1; dist(D,E)=2; dist(Q,K)= 3). Пороговым расстоянием считалось значение меньшее или равное 4.

*dist(x, y) ≤ 4*

1. Объединение аминокислот в группы с общими физико-химическими свойствами. Данный подход предполагает, что замена аминокислоты на аминокислоту с близкими физико-химическими свойствами не влияет на функциональную активность клона.  
   Таким образом, аминокислоты в составе следующих групп считались идентичными: FYH, LIMV, TSC, a соответствующие замены синонимичными.



***Рисунок 10****. Различные подходы отбора и извлечения признаков из данных.* ***a*** *– предварительный отбор признаков;* ***b*** *– численно представление признака, может быть, взвешенных или невзвешенным. c – различные параметры поиска признака в пределах репертуара (детали в тексте).*

## **2.8 Задача бинарной классификации**

Пусть X — множество описаний объектов, Y —множество меток классов. Существует неизвестная целевая зависимость  y^{*}:\; X\to Y, значения которой известны только для объектов [обучающей выборки](http://www.machinelearning.ru/wiki/index.php?title=%D0%92%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0)  X^m = \{(x_1,y_1),\dots,(x_m,y_m)\}.

Необходимо построить [алгоритм](http://www.machinelearning.ru/wiki/index.php?title=%D0%90%D0%BB%D0%B3%D0%BE%D1%80%D0%B8%D1%82%D0%BC) a:\; X\to Y, способный классифицировать произвольный объект x \in X.

В данном частном случае, каждый образец описывается вектором значений X, каждое из которых является отражением встречаемости данного клона в репертуаре. Тогда, решением задачи классификации является алгоритм, получающий на входе множество X и способный классифицировать данный произвольный образец (присвоить статус COVID или healthy). Функция Изображение выглядит как текст

Автоматически созданное описание определяется как функция потерь, которая оптимизируется для решения данной задачи, где y — истинное значение выхода модели и y’  
— фактическое значение.

Изображение выглядит как текст

Автоматически созданное описание  
**2.9 Применение моделей для предсказания статуса COVID**

Для прогнозирования статуса пациентов по набору клонотипов использовались модели машинного обучения Random Forest и SVM. Оптимизация гиперпараметров моделей осуществлялась методом перекрёстной проверки (кросс-валидацией). Для наших целей использовалась кросс-валидация по 3 блокам с поиском оптимальных гиперпараметров среди следующего списка параметров.

Список параметров RF: {'criterion': ['gini'], 'max\_depth' : [3, 5, 7], 'min\_samples\_leaf': [4, 10, 20], 'min\_samples\_split': [10, 20, 40], 'max\_features': [15, 20, 50]}

Список параметров SVM: {"C": [0.01, 0.5, 1], "kernel": ["rbf"], "grade": [1, 2], "gamma": ["scale"]}

Поиск гиперпараметров выполнялся для наброра обучающих выборок, после чего найденные гиперпараметры применялись к тестовой выборке.  
При разделении пациентов на группы учитывались партии секвенирования, то есть не допускалось попадания в разные группы образцов, принадлежащих одной партии секвенирования. Также мы стремились добиться количественного баланса между классами COVID+ и COVID- в обучающих и тестовых выборках. Таким образом были сформированы следующие группы пациентов (Таблица).

**Rus1**: rus\_batch1, rus\_batch2, rus\_batch3

**Rus2**: rus\_batch4, rus\_batch5, rus\_batch6

**Adaptive1**: COVID-19-NIH/NIAID, COVID-19-HUniv12Oct, COVID-19- COVID19-Adaptive, KECK

**Adaptive2**: COVID-19-DLS, COVID-19-ISB, COVID-19-IRST/AUSL, HIP

Затем были получены комбинации (размещение) упомянутых групп и для каждой комбинации рассчитывались оценки точности моделей. Результат обобщен на рисунке ...

Мы использовали две модели машинного обучения: Random Forest и SVM.

Random Forest (случайный лес) строит несколько деревьев решений и объединяет их вместе для получения более точного и стабильного прогноза. Random Forest работает в два этапа: создается случайный лес путем объединения N деревьев решений, затем рассчитывается прогноз для каждого дерева, созданного на первом этапе.

Алгоритм можно описать с помощью следующих шагов:

1. Выбираем случайные точки данных K из тренировочного набора.  
   Строим деревья решений выбранных точек данных.
2. Комбинируем N деревьев решений.
3. Повторяем шаги 1 и 2
4. ля новых точек данных получаем прогнозы каждого дерева решений и присваиваем точкам данных классы, которые набирают большинство голосов.

Цель алгоритма SVM (метода опорных векторов) — найти гиперплоскость в N-мерном пространстве (N — количество признаков), которая наиболее эффективно разделяет точки данных.  
Чтобы разделить два класса точек данных, можно выбрать множество возможных гиперплоскостей. Задача алгоритма найти плоскость с максимальным зазором - максимальным расстоянием между точками данных обоих классов. Опорные векторы — это точки данных, расположенные наиболее близко к гиперплоскости и влияющие на положение и ориентацию гиперплоскости. С помощью опорных векторов, максимизируется зазор классификатора. Удаление опорных векторов изменяет положение гиперплоскости.

**2.10 Обучающие, контрольные и валидационные наборы данных**

Чтобы обучить нашу модель и оценить ее качество, необходимо разделить совокупность данных на 3 части:

(1) Для непосредственного обучения модели обучающая выборка (training sample), по которой производится настройка (оптимизация гиперпараметров) алгоритма  
(2) Для оценки качества модели используется тестовая (контрольная) выборка (test sample), которая, не должна пересекаться с обучающей выборкой.;  
(3) Для выбора наилучшей модели машинного обучения понадобится проверочная (валидационная) выборка (validation sample), которая является независимой от обучающей выборки.

## **2.11 Метрики качества модели** Точность (Accuracy) – доля правильных ответов. Однако этот показатель очень чувствителен к балансу размеров классов, поэтому его следует использовать с осторожностью.

  
Сбалансированная точность – среднее значение точности для каждого класса.

Эта метрика более надежна в случае несбалансированного размера классов.  
AUC-ROC – площадь под ROC-кривой, где ROC-кривая – это графическая характеристика, отражающая предсказательную способность модели бинарного классификатора, зависимость доли истинно положительных классификаций (True Positive Rate (TPR)) от доли ложноположительных классификаций (False Positive Rate (FPR)) при изменении порога решающего правила. AUC-ROC в свою очередь, зависит от точности и полноты.



  
Precision (точность) доля объектов положительного класса, которые являются истинно положительными

  
Recall (полнота) доля объектов положительного класса из всех объектов положительного класса, которая была выявлена моделью.  
<https://habr.com/ru/company/ods/blog/328372/>



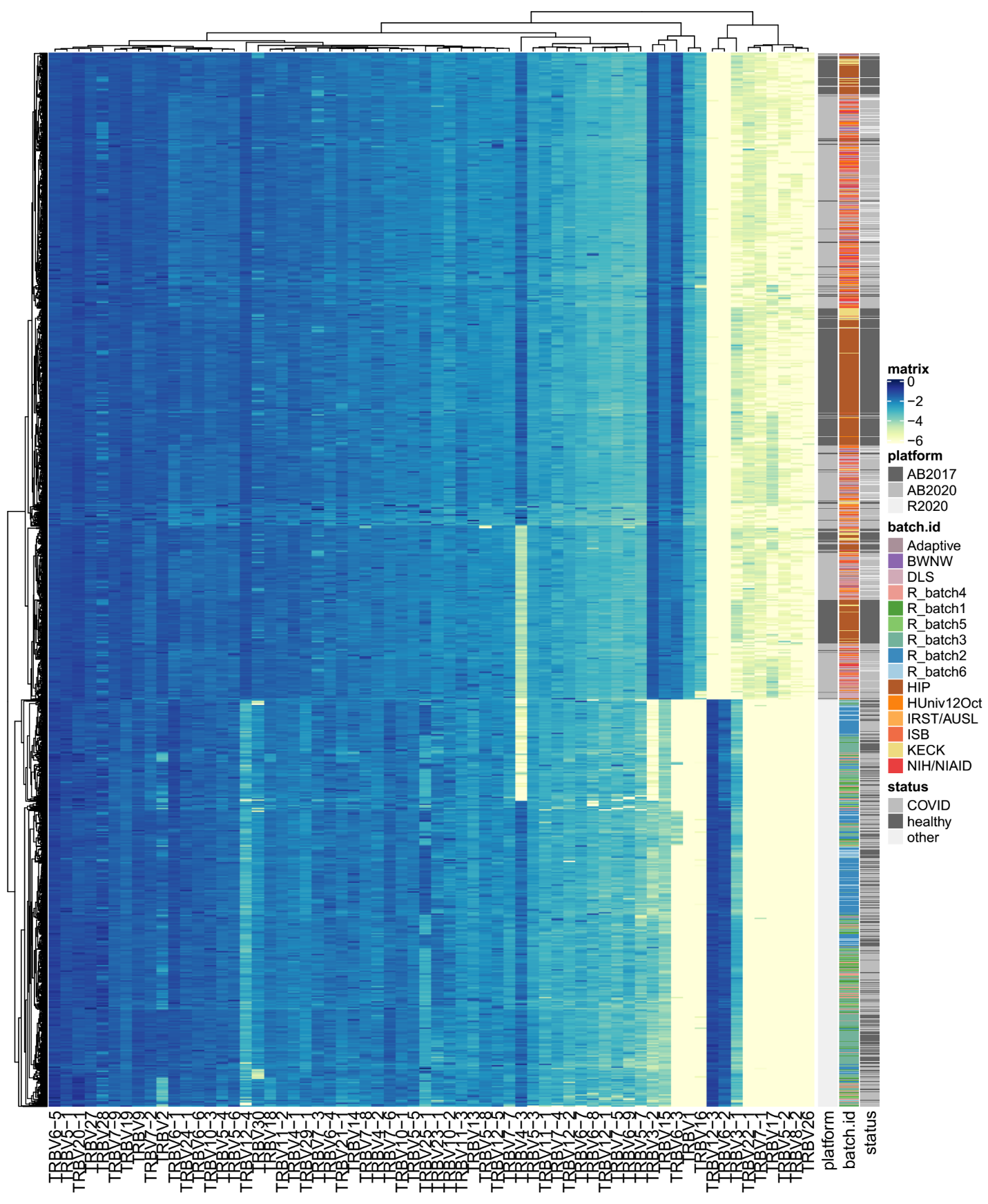
# ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Тепловая карта частот комбинаций вариабельного гена и соединительного гена (VJ-ген) индивидуальных образцов является эффективным представлением данных репертуаров Т-клеточных рецепторов.  Иерархическая организация данных позволяет выявить кластеры образцов с идентичным характером распределения VJ-usage, которое непосредственно связано с количественным и качественным CDR3 составом репертуара. Особенно информативно может быть сравнение образцов с разным статусом инфекционного заболевания или статусом вакцинирования. Рис иллюстрирует негомогенность репертуаров анализируемых выборок по VJ-usage, связанную с особенностями методики подготовки библиотеки TCR. Праймеры для аллелей J или константной области α- и β-цепей TCR используют вместе со смесью праймеров для всех известных аллелей V. Это приводит к специфической амплификации транскрипта TCR в области CDR3. Таким образом, методы мультиплексной ПЦР подвержены систематической ошибке амплификации [42], то есть преобладающей амплификации одних аллелей по сравнению с другими, что приводит к искажению относительного содержания полученных продуктов.

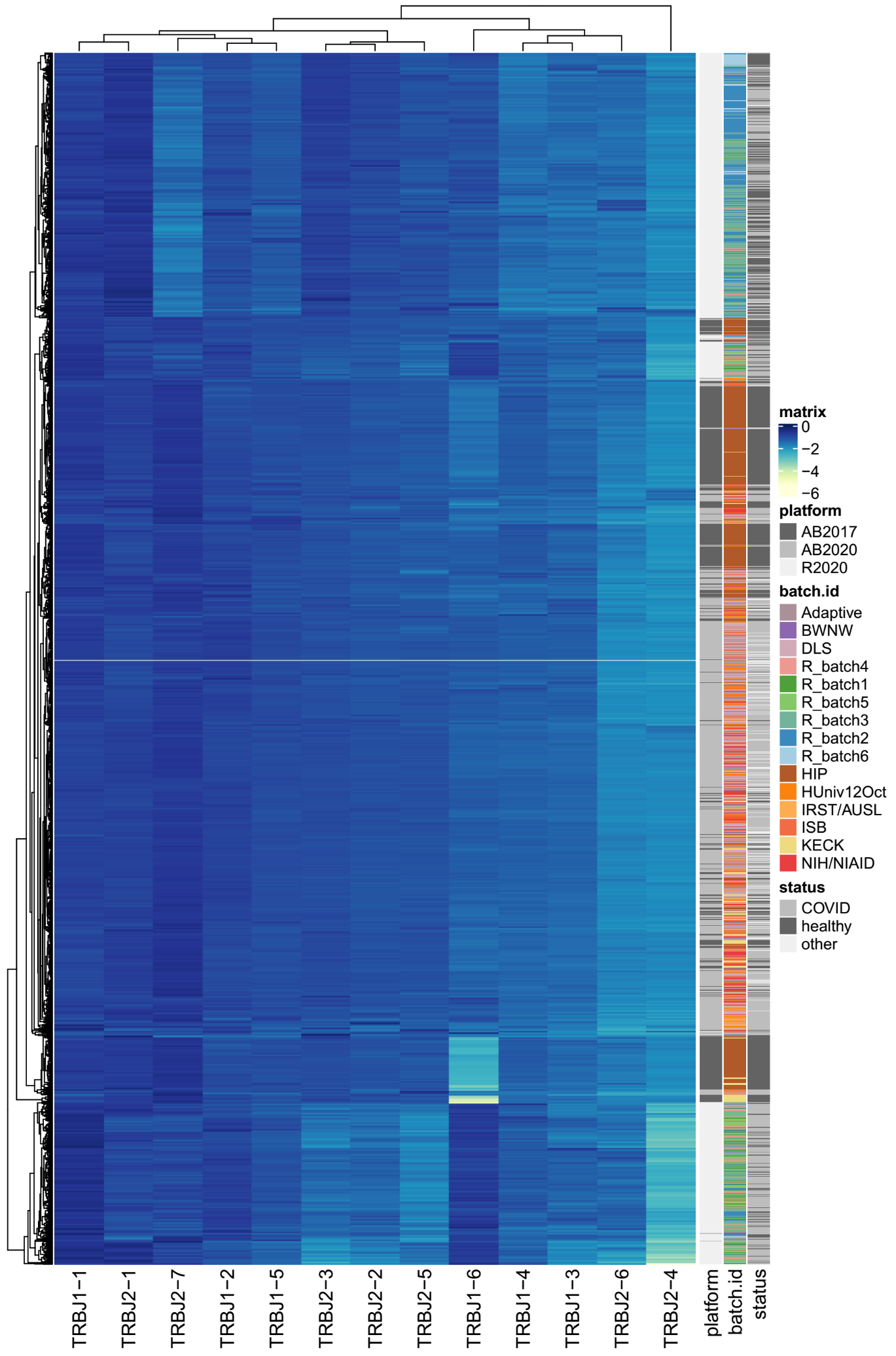
Мы можем наблюдать довольно явную кластеризацию, по выборкам данных, полученных в разное время и в разных лабораториях (РисA, platform, рисB) вероятно связанную с систематической ошибкой амплификации. Несмотря на то, что данные AB17 и AB20 получались в одной лаборатории с разницей в три года групповой эффект по этим выборкам, является сильно выраженным. Все это делает невозможным разделение образцов из различных выборок по статусу и порождает необходимость приведения наблюдаемых частот в искусственных группах данных (AB2017, AB2020, Rus2020) к общему распределению частот.

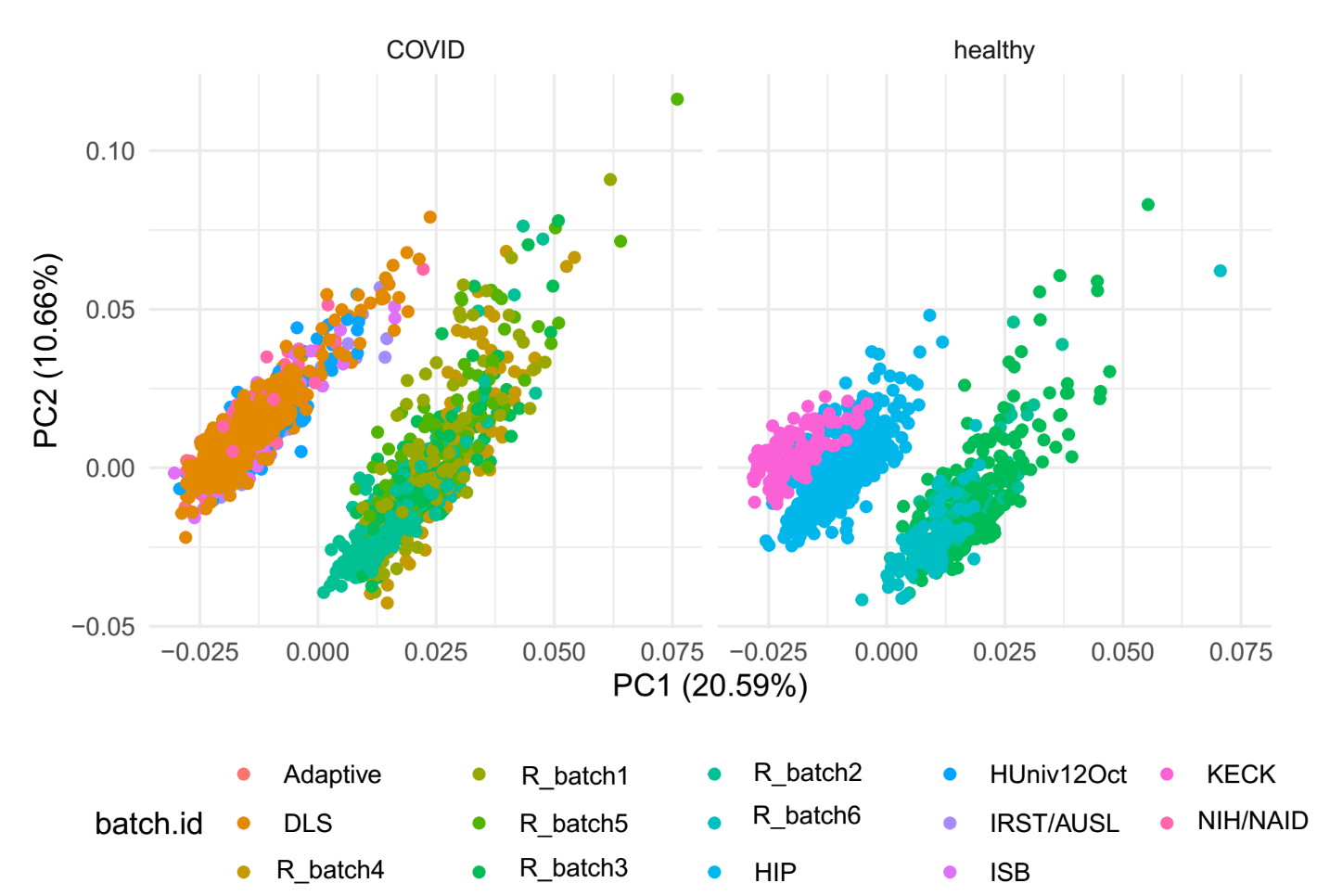
Была проведена нормализация по V-usage, описанная в соответствующем разделе главы материалов и методов.

Okino ST, Kong M, Sarras H, Wang Y. Evaluation of bias associated with high-multiplex, target-specific pre-amplification. *Biomol Detect Quantif.*2016;6:13–21. doi: 10.1016/j.bdq.2015.12.001. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4822213/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/27077043)] [[CrossRef](https://doi.org/10.1016%2Fj.bdq.2015.12.001)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=Biomol+Detect+Quantif&title=Evaluation+of+bias+associated+with+high-multiplex,+target-specific+pre-amplification&author=ST+Okino&author=M+Kong&author=H+Sarras&author=Y+Wang&volume=6&publication_year=2016&pages=13-21&pmid=27077043&doi=10.1016/j.bdq.2015.12.001&)]



a

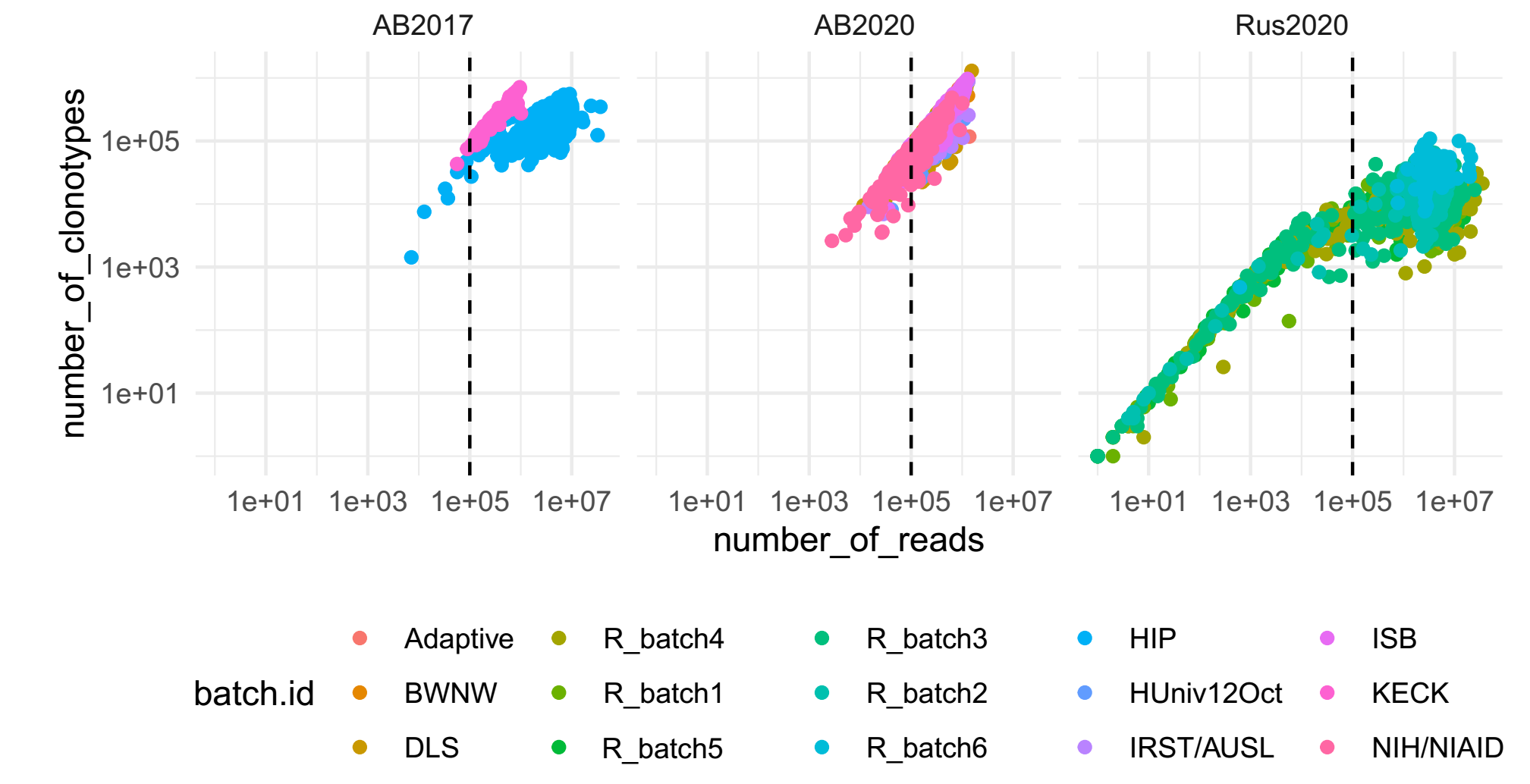
 b

 с

***Рисунок 11****.* ***a*** *и* ***b*** *- тепловые карты, отражающие кластеризацию данных по частотам TRBV сегментов.* ***с*** *– две главные компоненты данных частот комбинаций* VJ-usage. *Platform - ряд данных полученных в различное время и в разных лабораториях (AB17 – HIP и KECK, AB20 – все остальные выборки Adaptive, полученные в 2020 году, Rus2020(таблица)), batch.id – ряд данных представленный разными ранами секвенирования, status – COVID-статус образца (healthy, COVID, other)*

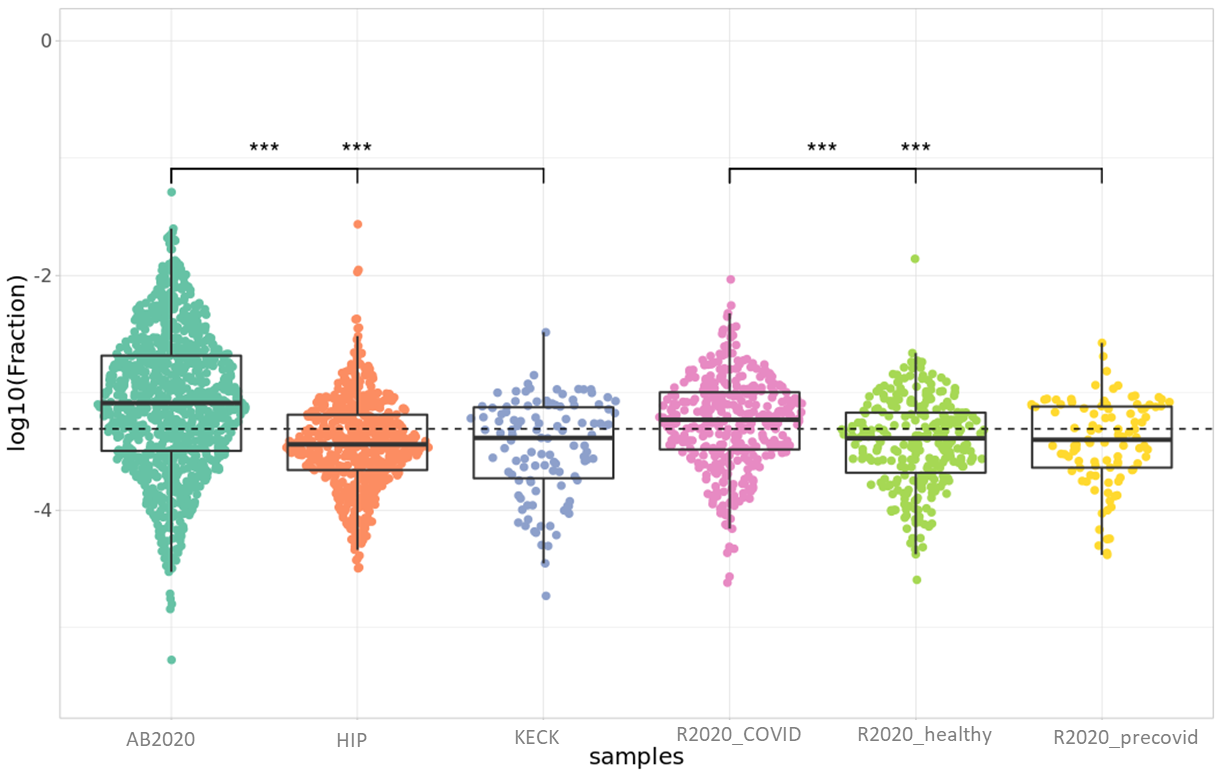
Разрешение секвенирования может также сильно отличаться для разных образцов, что отражается на общем количестве клонов и на представленности отдельного рида (клона).

На рисунке показана зависимость числа клонов от суммы числа ридов в образце для ненормированных по глубине секвенирования данных. Как можно видеть выборка данных Rus2020 содержит большое число образцов низкого качества секвенирования. В целях контроля качества из всех имеющихся данных были отфильтрованы образцы, содержащие менее 6000 клонов с суммой ридов менее 105. Также была проведена нормализация данных по глубине секвенирования, описанная в советующем разделе Материалов и методов.



***Рисунок 12****. Число ридов и количество клонов в данных секвенирования различных образцов.*

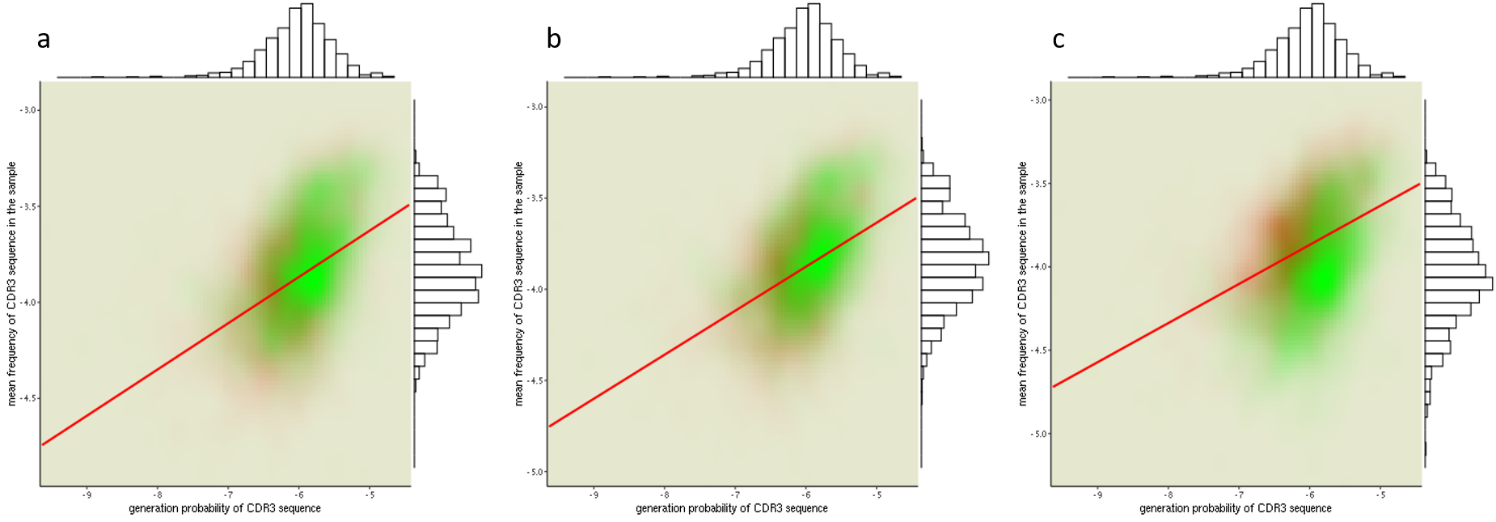
Затем, на когортах AB2020, AB2017, R2020 нормированных по V – usage и глубине секвенирования («**Нормализация по глубине секвенирования**», пункт 2) были получены признаки для обучения модели, представляющие собой набор 2000 вариантов TCRb, обогащенных у пациентов инфицированных COVID. На рисунке показана доля COVID-ассоциированных клонов в различных образцах в зависимости от выборки данных.  Этот показатель статистически значимо отличается между группами пациентов инфицированных COVID и группами здоровыми донорами (Таблица).

****

***Рисунок 13****. Представленность COVID-ассоциированных клонов в различных выборках образцов.*

|  |  |
| --- | --- |
| **Dataset** | P, Kruskal-Wallis ANOVA |
| AB2020 vs AB2017 | 4.2e-45 |
| Rus2020\_COVID vs Rus2020\_healthy, Rus2020\_precovid | 1.1e-08 |

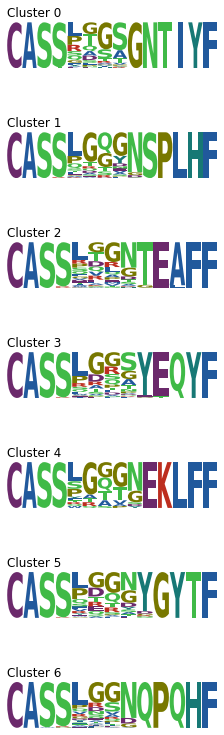
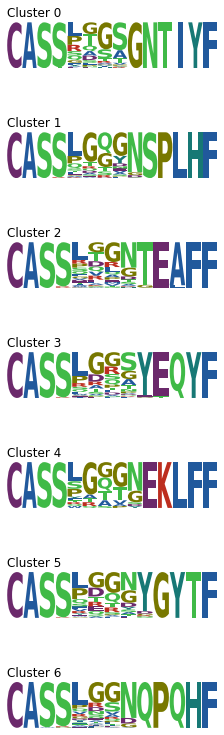
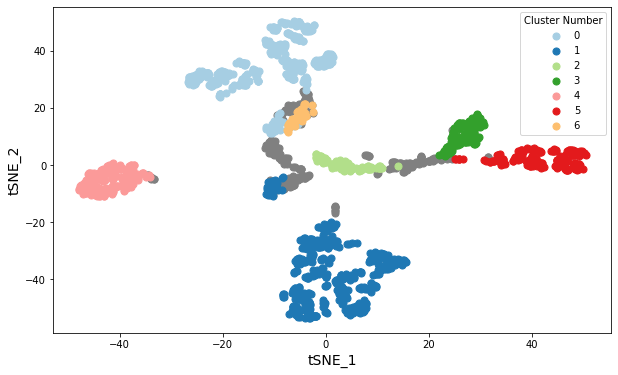
Как видно из Рисунка 4 обогащенные по COVID+ публичные клонотипы присутствуют с большей частотой в как репертуарах больных в острой фазе инфекции (*IgM +*), так и в переболевших пациентах (*IgM-IgG +,*) по сравнению с репертуаром здоровых доноров.

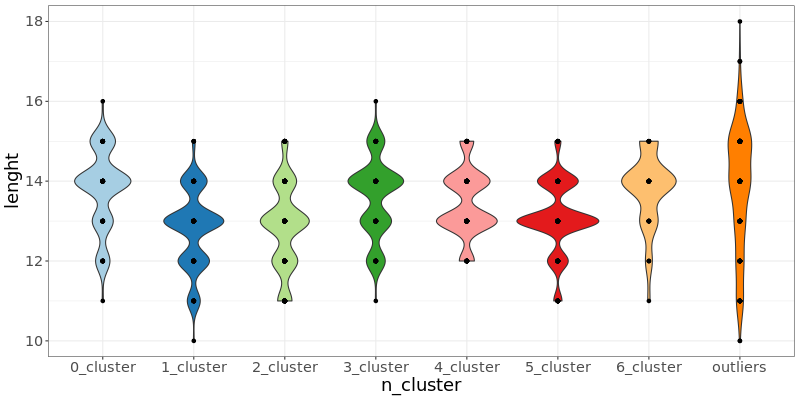


***Рисунок 14****. Наблюдаемые частоты клонотипов TRB в разных выборочных когортах (ось Y) по сравнению с теоретическими (ось X) частотами герерация (Pgen) 3000 случайно взятых клонов (красный цвет), взятых из общего пула клонов. Зеленым цветом обозначены 1000 клонов, ассоциированных с SARS-CoV-2, обнаруженных с помощью теста U-критерия Манна-Уитни. Для 3000 случайно взятых клонов постоена линейная регрессия (красная линия). а - образцы из когорты IgM +, b - IgM-IgG +, c - здоровые*.

Несмотря на то, Т-клеточный ответ на конкретный патоген специфичен для каждого человека, ТКР, получаемые в ходе экспансии клонов клеток, обладают определённой степенью гомологии у разных людей. Следующим шагом после идентификации ассоциированных с ковид клонов был поиск общих мотивов или паттернов среди пула этих последовательностей. Для этого были рассчитаны попарные расстояния (расстояния Хэмминга) между выявленными клонами, построена матрица попарных растояний и с помощью алгоритма DBSCAN найдены группы клонов, обладающих гомологией. Получившиеся кластеры и соответствующие им мотивы представлены на рисунке (A). Как можно видеть на рисунке (B), обособленные группы клонов имеют отличающиеся распределения длин последовательностей с выраженными модальными значениями. Были предприняты попытки поиска найденных вариантов в базе данных VDJdb (URL: <https://vdjdb.cdr3.net/>) c разрешением одной замены, совпадений найдено не было.

a

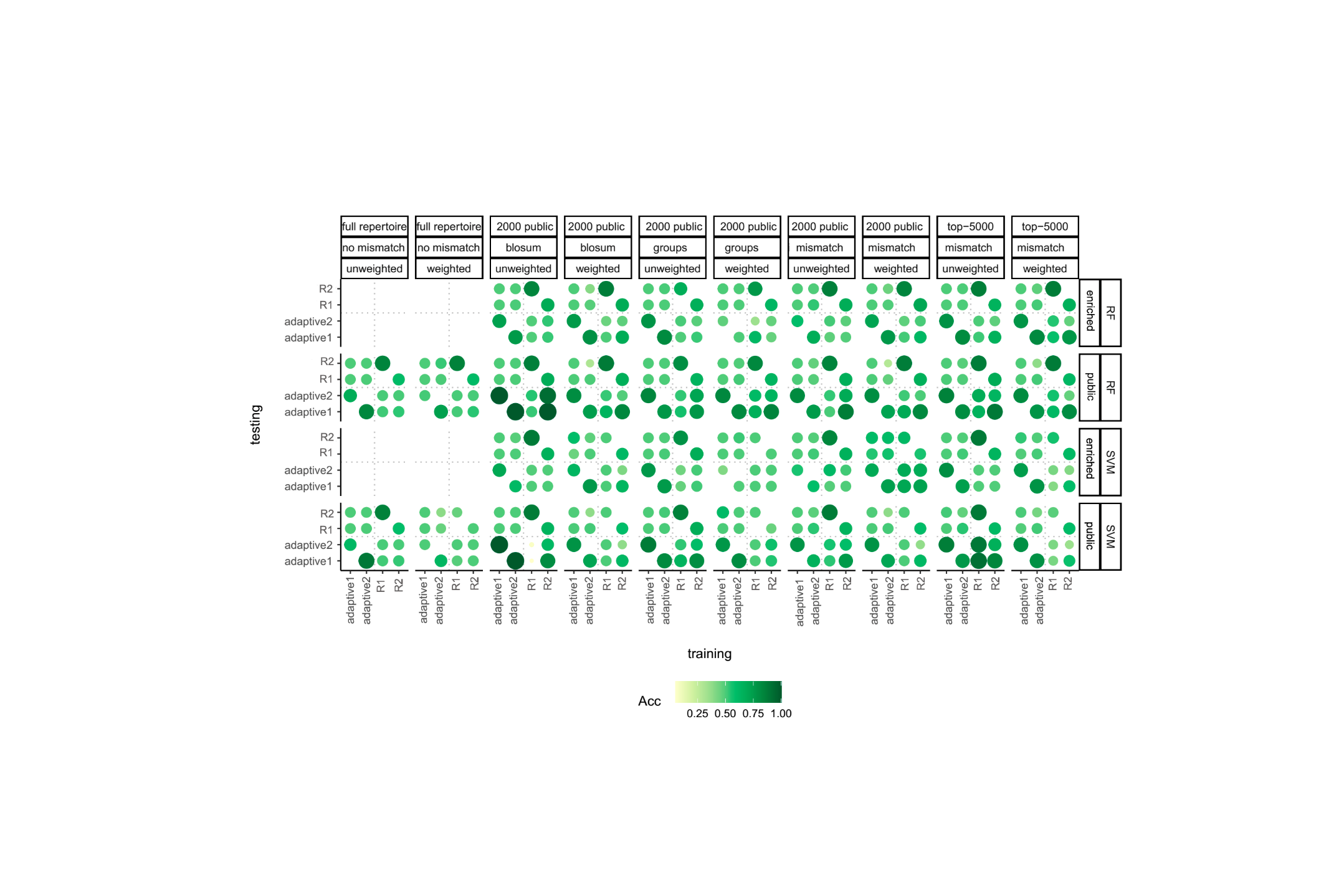


**b

***Рисунок 15****. А - Кластеры COVID-ассоциированных клонов, полученные на основе матрицы попарных расстояний и соответствующие мотивы групп клонов, составляющих кластер. B – Распределение клонов, принадлежащих в выявленных кластерам по длине. Outliers – точки- выбросы, расположенные в области низкой плотности точек.*

Также в качестве набора признаков для обучения модели использовался набор наиболее распространенных публичных вариантов TCR, присутствующих как минимум у 60 из 2872 пациентов, независимо от их связи с COVID-статусом. Кроме того, размер соответствующих клонотипов в пределах репертуара (т. е. частота клонотипов) мог учитываться (взвешенное представление) или нет. В последнем случае учитывалось только наличие или отсутствие клонотипа (встречаемость). Также поиск паттернов исходного признака в пределах репертуара осуществлялся с помощью трех подходов, использующих разные метрики различия последовательностей (детали в методах).

Затем мы использовали метод перекрестной проверки для обучения моделей Random Forest и SVM, на отобранных для этих целей выборках (Таблица) с последующей валидацией моделей. Таким образом, мы рассмотрели все случаи, в которых каждая из сформированных нами выборок выступала как в качестве обучающего, так и тестового набора данных. При этом мы использовали все комбинации параметров предварительной обработки данных: описанные способы нормализации по глубине секвенирования, отбора и извлечения признаков. В качестве контрольной выборки использовали данные ненормированные по глубине секвенирования, с параметром поиска клонов - без допуска аминокислотных замен. Оценкой производительности модели служит сбалансированная точность. Полученные значения точности для всех раундов валидации моделей показаны на рисунке.



*Рисунок 16. Зависимость точности предсказания модели от способов нормализации, отбора и извлечения признаков.*

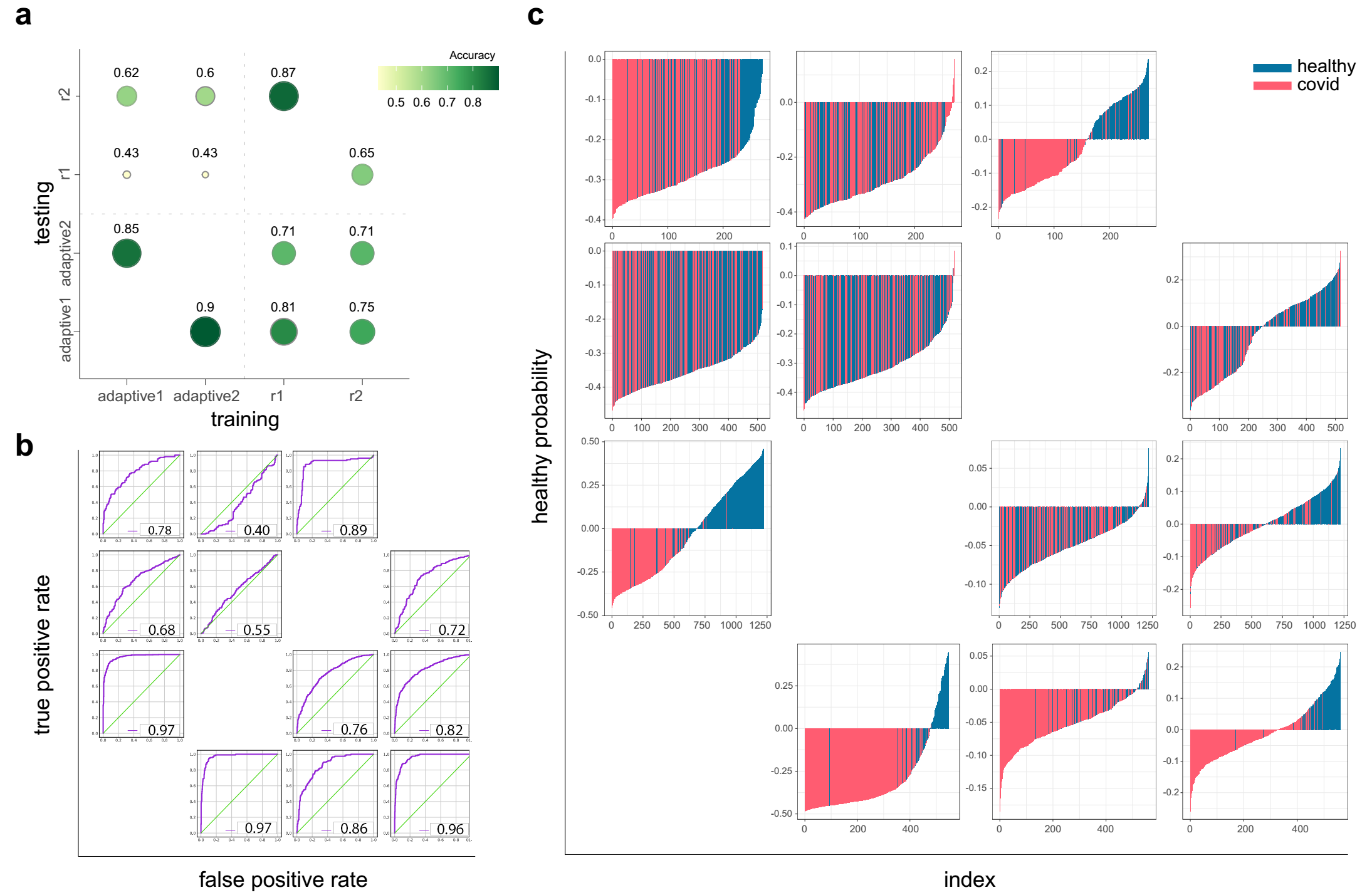
Как можно видеть, производительность прогнозирования модели сильно зависит от выбора обучающей и тестовой выборок и в меньшей степени определяется набором параметров предварительной обработки данных. Обучение модели на выборках данных R2020 приводит к хорошей производительности модели на выборках AB2020 и AB2017, в то время как наблюдается переобучение классификатора на данных Adaptive. Это указывает на наличие батч эффекта и его влияние на показатель точности и, следовательно недостаточную эффективность методов нормализации данных.

Использование публичных клонов в качестве признаков для машинного привело к более высоким показателям точности. Это можно объяснить наличием общих закономерностей в структуре репертуаров неинфицированных доноров и существованием клонов, специфичных для людей с отрицательным статусом COVID. На это, косвенно, указывает, тот факт, что часть признаков, внесших наибольший вклад в итоговое предсказание модели (выявленных с помощью встроенного алгоритма Random Forest) не связаны статистически значимо с целевой переменной COVID.

Невзвешенный способ представления признаков оказался более эффективным. С одной стороны, этот подход позволяет скорректировать батч эффект, обусловленный разной глубиной прочтения данных (в дополнение к существующим способам нормализации), так как значения признаков, в таком случае, зависят от числа ридов косвенным образом (учитывается лишь факт наличия клона в репертуаре, но не его количество). Несмотря на то, что при таком подходе, утрачивается часть информации о количественном составе репертуара, описанное преимущество имеет больший вес.

Также стоит отметить, что поиск клона по гомологии приводит к большей производительности модели, чем поиск полностью идентичных клонов в пределах репертуара. Очевидно, что такой подход расширяет поиск одного клона до нескольких вариантов последовательностей (в зависимости от параметров поиска) с схожими структурно-функциональными свойствами и идентичной антиген-специфичностью. Однако при сравнении трех подходов поиска не удалось выявить предпочтительный.

Результирующая точность прогноза, графики кривых ROC-AUC и водопады показаны на рисунке для наиболее производительной модели.

**

***Рисунок 17****. a – значения сбалансированной точности предсказания для COVID-статусов. b. ROC-кривые. с. Графики водопады. Для обучения и тестирования моделей использовались четыре набора данных, каждый из которых включал здоровых доноров и доноров инфицированных COVID. Используемые параметры: предварительная обработка наборов данных: datasets preprocessing: top5000 clonotypes; features: public clonotypes; mode: single mismatch, unweighted.*

*В текущем исследовании мы стремились обучить классификатор, который будет различать здоровых людей и людей, перенесших COVID, на основе их репертуара Т-клеток. Мы исследовали различные наборы функций и способы их кодирования, чтобы создать надежную и точную модель. Мы использовали наборы данных, которые различались по глубине секвенирования и имели множество предубеждений, связанных со сбором данных, что, с одной стороны, было сложным, но, с другой стороны, способствовало надежности полученной модели. На основании полученных результатов можно сделать вывод, что можно создать такой классификатор, используя общедоступные клонотипы, закодированные как соответствующие признаки. В будущем потенциал такой модели может быть расширен, если будет предоставлена информация о V-использовании и HLA-статусе.*

***Выводы***

1. Разработаны алгоритмы для нормализации данных секвенирования репертуаров ТКР по распределению частот генных сегментов V (V-usage) и глубине секвенирования.
2. Разработан алгоритм для идентификации в репертуарах Т-клеточных рецепторов групп клонов, обладающих идентичной структурой и ассоциированных с SARS-CoV-2.
3. Разработано несколько подходов извлечения признаков из данных индивидуальных репертуаров TCR.
4. С использованием разработанных алгоритмов была проведена подготовка данных для машинного обучения и на основе этих данных построены модели, способные предсказывать инфекционный статус COVID.
5. Проведен анализ эффективности подходов предварительной обработки данных для машинного обучения

Список литературы

[1] S. Tonegawa, "Somatic generation of antibody diversity," *Nature,* vol. 302, no. 5909, pp. 575-81, Apr 14 1983, doi: 10.1038/302575a0.

[2] M. M. Davis and P. J. Bjorkman, "T-cell antigen receptor genes and T-cell recognition," *Nature,* vol. 334, no. 6181, pp. 395-402, Aug 4 1988, doi: 10.1038/334395a0.

[3] E. J. Adams, S. Gu, and A. M. Luoma, "Human gamma delta T cells: Evolution and ligand recognition," *Cell Immunol,* vol. 296, no. 1, pp. 31-40, Jul 2015, doi: 10.1016/j.cellimm.2015.04.008.

[4] D. Kabelitz and J. Dechanet-Merville, "Editorial: "Recent Advances in Gamma/Delta T Cell Biology: New Ligands, New Functions, and New Translational Perspectives"," *Front Immunol,* vol. 6, p. 371, 2015, doi: 10.3389/fimmu.2015.00371.

[5] B. Silva-Santos and J. Strid, "gammadelta T cells get adaptive," *Nat Immunol,* vol. 18, no. 4, pp. 370-372, Mar 22 2017, doi: 10.1038/ni.3705.

[6] P. Dash *et al.*, "Quantifiable predictive features define epitope-specific T cell receptor repertoires," *Nature,* vol. 547, no. 7661, pp. 89-93, Jul 6 2017, doi: 10.1038/nature22383.

[7] J. Glanville *et al.*, "Identifying specificity groups in the T cell receptor repertoire," *Nature,* vol. 547, no. 7661, pp. 94-98, Jul 6 2017, doi: 10.1038/nature22976.

[8] I. Song, A. Gil, R. Mishra, D. Ghersi, L. K. Selin, and L. J. Stern, "Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant CD8(+) T cell epitope," *Nat Struct Mol Biol,* vol. 24, no. 4, pp. 395-406, Apr 2017, doi: 10.1038/nsmb.3383.

[9] R. Bosselut, "T cell antigen recognition: Evolution-driven affinities," *Proc Natl Acad Sci U S A,* vol. 116, no. 44, pp. 21969-21971, Oct 29 2019, doi: 10.1073/pnas.1916129116.

[10] M. N. Artyomov, M. Lis, S. Devadas, M. M. Davis, and A. K. Chakraborty, "CD4 and CD8 binding to MHC molecules primarily acts to enhance Lck delivery," *Proc Natl Acad Sci U S A,* vol. 107, no. 39, pp. 16916-21, Sep 28 2010, doi: 10.1073/pnas.1010568107.

[11] R. Zamoyska, "CD4 and CD8: modulators of T-cell receptor recognition of antigen and of immune responses?," *Curr Opin Immunol,* vol. 10, no. 1, pp. 82-7, Feb 1998, doi: 10.1016/s0952-7915(98)80036-8.

[12] H. Cantor and E. A. Boyse, "Development and function of subclasses of T cells," *J Reticuloendothel Soc,* vol. 17, no. 2, pp. 115-8, Feb 1975. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/47392>.

[13] P. Kisielow *et al.*, "Ly antigens as markers for functionally distinct subpopulations of thymus-derived lymphocytes of the mouse," *Nature,* vol. 253, no. 5488, pp. 219-20, Jan 17 1975, doi: 10.1038/253219a0.

[14] A. R. Townsend, J. Rothbard, F. M. Gotch, G. Bahadur, D. Wraith, and A. J. McMichael, "The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides," *Cell,* vol. 44, no. 6, pp. 959-68, Mar 28 1986, doi: 10.1016/0092-8674(86)90019-x.

[15] R. M. Zinkernagel and P. C. Doherty, "Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system," *Nature,* vol. 248, no. 5450, pp. 701-2, Apr 19 1974, doi: 10.1038/248701a0.

[16] J. M. Khan and S. Ranganathan, "Understanding TR binding to pMHC complexes: how does a TR scan many pMHC complexes yet preferentially bind to one," *PLoS One,* vol. 6, no. 2, p. e17194, Feb 22 2011, doi: 10.1371/journal.pone.0017194.

[17] D. Gfeller *et al.*, "The Length Distribution and Multiple Specificity of Naturally Presented HLA-I Ligands," *J Immunol,* vol. 201, no. 12, pp. 3705-3716, Dec 15 2018, doi: 10.4049/jimmunol.1800914.

[18] P. Wang, J. Sidney, C. Dow, B. Mothe, A. Sette, and B. Peters, "A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach," *PLoS Comput Biol,* vol. 4, no. 4, p. e1000048, Apr 4 2008, doi: 10.1371/journal.pcbi.1000048.

[19] A. Simon, Z. Dosztanyi, E. Rajnavolgyi, and I. Simon, "Function-related regulation of the stability of MHC proteins," *Biophys J,* vol. 79, no. 5, pp. 2305-13, Nov 2000, doi: 10.1016/S0006-3495(00)76476-9.

[20] B. Rodenko *et al.*, "Generation of peptide-MHC class I complexes through UV-mediated ligand exchange," *Nat Protoc,* vol. 1, no. 3, pp. 1120-32, 2006, doi: 10.1038/nprot.2006.121.

[21] J. Robinson, J. A. Halliwell, H. McWilliam, R. Lopez, P. Parham, and S. G. Marsh, "The IMGT/HLA database," *Nucleic Acids Res,* vol. 41, no. Database issue, pp. D1222-7, Jan 2013, doi: 10.1093/nar/gks949.

[22] J. Robinson *et al.*, "Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles," *PLoS Genet,* vol. 13, no. 6, p. e1006862, Jun 2017, doi: 10.1371/journal.pgen.1006862.

[23] J. Sidney *et al.*, "Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries," *Immunome Res,* vol. 4, p. 2, Jan 25 2008, doi: 10.1186/1745-7580-4-2.

[24] M. Andreatta and M. Nielsen, "Gapped sequence alignment using artificial neural networks: application to the MHC class I system," *Bioinformatics,* vol. 32, no. 4, pp. 511-7, Feb 15 2016, doi: 10.1093/bioinformatics/btv639.

[25] M. Nielsen, C. Lundegaard, O. Lund, and C. Kesmir, "The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage," *Immunogenetics,* vol. 57, no. 1-2, pp. 33-41, Apr 2005, doi: 10.1007/s00251-005-0781-7.

[26] M. Nielsen and M. Andreatta, "NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions," *Nucleic Acids Res,* vol. 45, no. W1, pp. W344-W349, Jul 3 2017, doi: 10.1093/nar/gkx276.

[27] A. Mosch, S. Raffegerst, M. Weis, D. J. Schendel, and D. Frishman, "Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors," *Front Genet,* vol. 10, p. 1141, 2019, doi: 10.3389/fgene.2019.01141.

[28] M. S. Krangel, "Mechanics of T cell receptor gene rearrangement," *Curr Opin Immunol,* vol. 21, no. 2, pp. 133-9, Apr 2009, doi: 10.1016/j.coi.2009.03.009.

[29] D. J. Laydon, C. R. Bangham, and B. Asquith, "Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach," *Philos Trans R Soc Lond B Biol Sci,* vol. 370, no. 1675, Aug 19 2015, doi: 10.1098/rstb.2014.0291.

[30] C. H. Bassing, W. Swat, and F. W. Alt, "The mechanism and regulation of chromosomal V(D)J recombination," *Cell,* vol. 109 Suppl, pp. S45-55, Apr 2002, doi: 10.1016/s0092-8674(02)00675-x.

[31] M. Nishana and S. C. Raghavan, "Role of recombination activating genes in the generation of antigen receptor diversity and beyond," *Immunology,* vol. 137, no. 4, pp. 271-81, Dec 2012, doi: 10.1111/imm.12009.

[32] A. Murugan, T. Mora, A. M. Walczak, and C. G. Callan, Jr., "Statistical inference of the generation probability of T-cell receptors from sequence repertoires," *Proc Natl Acad Sci U S A,* vol. 109, no. 40, pp. 16161-6, Oct 2 2012, doi: 10.1073/pnas.1212755109.

[33] T. Dupic, Q. Marcou, A. M. Walczak, and T. Mora, "Genesis of the alphabeta T-cell receptor," *PLoS Comput Biol,* vol. 15, no. 3, p. e1006874, Mar 2019, doi: 10.1371/journal.pcbi.1006874.

[34] B. Howie *et al.*, "High-throughput pairing of T cell receptor alpha and beta sequences," *Sci Transl Med,* vol. 7, no. 301, p. 301ra131, Aug 19 2015, doi: 10.1126/scitranslmed.aac5624.

[35] E. S. Lee, P. G. Thomas, J. E. Mold, and A. J. Yates, "Identifying T Cell Receptors from High-Throughput Sequencing: Dealing with Promiscuity in TCRalpha and TCRbeta Pairing," *PLoS Comput Biol,* vol. 13, no. 1, p. e1005313, Jan 2017, doi: 10.1371/journal.pcbi.1005313.

[36] A. Balakrishnan and G. P. Morris, "The highly alloreactive nature of dual TCR T cells," *Curr Opin Organ Transplant,* vol. 21, no. 1, pp. 22-8, Feb 2016, doi: 10.1097/MOT.0000000000000261.

[37] R. Ahmed *et al.*, "A Public BCR Present in a Unique Dual-Receptor-Expressing Lymphocyte from Type 1 Diabetes Patients Encodes a Potent T Cell Autoantigen," *Cell,* vol. 177, no. 6, pp. 1583-1599 e16, May 30 2019, doi: 10.1016/j.cell.2019.05.007.

[38] N. J. Schuldt and B. A. Binstadt, "Dual TCR T Cells: Identity Crisis or Multitaskers?," *J Immunol,* vol. 202, no. 3, pp. 637-644, Feb 1 2019, doi: 10.4049/jimmunol.1800904.

[39] R. Ceredig and T. Rolink, "A positive look at double-negative thymocytes," *Nat Rev Immunol,* vol. 2, no. 11, pp. 888-97, Nov 2002, doi: 10.1038/nri937.

[40] H. Takaba and H. Takayanagi, "The Mechanisms of T Cell Selection in the Thymus," *Trends Immunol,* vol. 38, no. 11, pp. 805-816, Nov 2017, doi: 10.1016/j.it.2017.07.010.

[41] K. Takada and Y. Takahama, "Positive-selection-inducing self-peptides displayed by cortical thymic epithelial cells," *Adv Immunol,* vol. 125, pp. 87-110, 2015, doi: 10.1016/bs.ai.2014.09.003.

[42] E. V. Putintseva *et al.*, "Mother and child T cell receptor repertoires: deep profiling study," *Front Immunol,* vol. 4, p. 463, 2013, doi: 10.3389/fimmu.2013.00463.

[43] W. S. DeWitt, 3rd, A. Smith, G. Schoch, J. A. Hansen, F. A. t. Matsen, and P. Bradley, "Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity," *Elife,* vol. 7, Aug 28 2018, doi: 10.7554/eLife.38358.

[44] K. Bansal, H. Yoshida, C. Benoist, and D. Mathis, "The transcriptional regulator Aire binds to and activates super-enhancers," *Nat Immunol,* vol. 18, no. 3, pp. 263-273, Mar 2017, doi: 10.1038/ni.3675.

[45] H. Takaba *et al.*, "Fezf2 Orchestrates a Thymic Program of Self-Antigen Expression for Immune Tolerance," *Cell,* vol. 163, no. 4, pp. 975-87, Nov 5 2015, doi: 10.1016/j.cell.2015.10.013.

[46] P. Kraj and L. Ignatowicz, "The mechanisms shaping the repertoire of CD4(+) Foxp3(+) regulatory T cells," *Immunology,* vol. 153, no. 3, pp. 290-296, Mar 2018, doi: 10.1111/imm.12859.

[47] R. O. Emerson *et al.*, "Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire," *Nat Genet,* vol. 49, no. 5, pp. 659-665, May 2017, doi: 10.1038/ng.3822.

[48] S. D. Boyd and J. E. Crowe, Jr., "Deep sequencing and human antibody repertoire analysis," *Curr Opin Immunol,* vol. 40, pp. 103-9, Jun 2016, doi: 10.1016/j.coi.2016.03.008.

[49] S. Friedensohn, T. A. Khan, and S. T. Reddy, "Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires," *Trends Biotechnol,* vol. 35, no. 3, pp. 203-214, Mar 2017, doi: 10.1016/j.tibtech.2016.09.010.

[50] E. Rosati, C. M. Dowds, E. Liaskou, E. K. K. Henriksen, T. H. Karlsen, and A. Franke, "Overview of methodologies for T-cell receptor repertoire analysis," *BMC Biotechnol,* vol. 17, no. 1, p. 61, Jul 10 2017, doi: 10.1186/s12896-017-0379-9.

[51] K. Kitaura, T. Shini, T. Matsutani, and R. Suzuki, "A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) alpha and beta repertoires and identifying potential new invariant TCR alpha chains," *BMC Immunol,* vol. 17, no. 1, p. 38, Oct 11 2016, doi: 10.1186/s12865-016-0177-5.

[52] S. G. Lin, Z. Ba, Z. Du, Y. Zhang, J. Hu, and F. W. Alt, "Highly sensitive and unbiased approach for elucidating antibody repertoires," *Proc Natl Acad Sci U S A,* vol. 113, no. 28, pp. 7846-51, Jul 12 2016, doi: 10.1073/pnas.1608649113.

[53] H. S. Robins *et al.*, "Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells," *Blood,* vol. 114, no. 19, pp. 4099-107, Nov 5 2009, doi: 10.1182/blood-2009-04-217604.

[54] C. Cole, R. Volden, S. Dharmadhikari, C. Scelfo-Dalbey, and C. Vollmers, "Highly Accurate Sequencing of Full-Length Immune Repertoire Amplicons Using Tn5-Enabled and Molecular Identifier-Guided Amplicon Assembly," *J Immunol,* vol. 196, no. 6, pp. 2902-7, Mar 15 2016, doi: 10.4049/jimmunol.1502563.

[55] J. D. Freeman, R. L. Warren, J. R. Webb, B. H. Nelson, and R. A. Holt, "Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing," *Genome Res,* vol. 19, no. 10, pp. 1817-24, Oct 2009, doi: 10.1101/gr.092924.109.

[56] I. Z. Mamedov *et al.*, "Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling," *Front Immunol,* vol. 4, p. 456, 2013, doi: 10.3389/fimmu.2013.00456.

[57] M. V. Pogorelyy *et al.*, "Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires," *PLoS Comput Biol,* vol. 13, no. 7, p. e1005572, Jul 2017, doi: 10.1371/journal.pcbi.1005572.

[58] M. F. Quigley, J. R. Almeida, D. A. Price, and D. C. Douek, "Unbiased molecular analysis of T cell receptor expression using template-switch anchored RT-PCR," *Curr Protoc Immunol,* vol. Chapter 10, p. Unit10 33, Aug 2011, doi: 10.1002/0471142735.im1033s94.

[59] M. A. Turchaninova *et al.*, "High-quality full-length immunoglobulin profiling with unique molecular barcoding," *Nat Protoc,* vol. 11, no. 9, pp. 1599-616, Sep 2016, doi: 10.1038/nprot.2016.093.

[60] A. Minervina, M. Pogorelyy, and I. Mamedov, "T-cell receptor and B-cell receptor repertoire profiling in adaptive immunity," *Transpl Int,* vol. 32, no. 11, pp. 1111-1123, Nov 2019, doi: 10.1111/tri.13475.

[61] M. Matz *et al.*, "Amplification of cDNA ends based on template-switching effect and step-out PCR," *Nucleic Acids Res,* vol. 27, no. 6, pp. 1558-60, Mar 15 1999, doi: 10.1093/nar/27.6.1558.

[62] K. Best, T. Oakes, J. M. Heather, J. Shawe-Taylor, and B. Chain, "Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding," *Sci Rep,* vol. 5, p. 14629, Oct 13 2015, doi: 10.1038/srep14629.

[63] T. Oakes *et al.*, "Quantitative Characterization of the T Cell Receptor Repertoire of Naive and Memory Subsets Using an Integrated Experimental and Computational Pipeline Which Is Robust, Economical, and Versatile," *Front Immunol,* vol. 8, p. 1267, 2017, doi: 10.3389/fimmu.2017.01267.

[64] T. Kivioja *et al.*, "Counting absolute numbers of molecules using unique molecular identifiers," *Nat Methods,* vol. 9, no. 1, pp. 72-4, Nov 20 2011, doi: 10.1038/nmeth.1778.

[65] C. Linnemann *et al.*, "High-throughput identification of antigen-specific TCRs by TCR gene capture," *Nat Med,* vol. 19, no. 11, pp. 1534-41, Nov 2013, doi: 10.1038/nm.3359.

[66] E. Ruggiero *et al.*, "High-resolution analysis of the human T-cell receptor repertoire," *Nat Commun,* vol. 6, p. 8081, Sep 1 2015, doi: 10.1038/ncomms9081.

[67] M. De Simone, G. Rossetti, and M. Pagani, "Single Cell T Cell Receptor Sequencing: Techniques and Future Challenges," *Front Immunol,* vol. 9, p. 1638, 2018, doi: 10.3389/fimmu.2018.01638.

[68] A. Han, J. Glanville, L. Hansmann, and M. M. Davis, "Linking T-cell receptor sequence to functional phenotype at the single-cell level," *Nat Biotechnol,* vol. 32, no. 7, pp. 684-92, Jul 2014, doi: 10.1038/nbt.2938.

[69] D. A. Bolotin *et al.*, "MiXCR: software for comprehensive adaptive immunity profiling," *Nat Methods,* vol. 12, no. 5, pp. 380-1, May 2015, doi: 10.1038/nmeth.3364.

[70] M. Duez, M. Giraud, R. Herbert, T. Rocher, M. Salson, and F. Thonier, "Vidjil: A Web Platform for Analysis of High-Throughput Repertoire Sequencing," *PLoS One,* vol. 11, no. 11, p. e0166126, 2016, doi: 10.1371/journal.pone.0166126.

[71] L. Kuchenbecker *et al.*, "IMSEQ--a fast and error aware approach to immunogenetic sequence analysis," *Bioinformatics,* vol. 31, no. 18, pp. 2963-71, Sep 15 2015, doi: 10.1093/bioinformatics/btv309.

[72] B. Li *et al.*, "Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA-seq data," *Nat Genet,* vol. 49, no. 4, pp. 482-483, Mar 30 2017, doi: 10.1038/ng.3820.

[73] E. Miho, A. Yermanos, C. R. Weber, C. T. Berger, S. T. Reddy, and V. Greiff, "Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires," *Front Immunol,* vol. 9, p. 224, 2018, doi: 10.3389/fimmu.2018.00224.

[74] M. Shugay *et al.*, "Towards error-free profiling of immune repertoires," *Nat Methods,* vol. 11, no. 6, pp. 653-5, Jun 2014, doi: 10.1038/nmeth.2960.

[75] V. I. Nazarov *et al.*, "Reliability of immune receptor rearrangements as genetic markers for minimal residual disease monitoring," *Bone Marrow Transplant,* vol. 51, no. 10, pp. 1408-1410, Oct 2016, doi: 10.1038/bmt.2016.148.

[76] S. Schaller *et al.*, "ImmunExplorer (IMEX): a software framework for diversity and clonality analyses of immunoglobulins and T cell receptors on the basis of IMGT/HighV-QUEST preprocessed NGS data," *BMC Bioinformatics,* vol. 16, p. 252, Aug 12 2015, doi: 10.1186/s12859-015-0687-9.

[77] M. Shugay *et al.*, "VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires," *PLoS Comput Biol,* vol. 11, no. 11, p. e1004503, Nov 2015, doi: 10.1371/journal.pcbi.1004503.

[78] N. De Neuter *et al.*, "On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition," *Immunogenetics,* vol. 70, no. 3, pp. 159-168, Mar 2018, doi: 10.1007/s00251-017-1023-5.

[79] S. Gielis *et al.*, "Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires," *Front Immunol,* vol. 10, p. 2820, 2019, doi: 10.3389/fimmu.2019.02820.

[80] D. V. Bagaev *et al.*, "VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium," *Nucleic Acids Res,* vol. 48, no. D1, pp. D1057-D1062, Jan 8 2020, doi: 10.1093/nar/gkz874.

[81] M. Shugay *et al.*, "VDJdb: a curated database of T-cell receptor sequences with known antigen specificity," *Nucleic Acids Res,* vol. 46, no. D1, pp. D419-D427, Jan 4 2018, doi: 10.1093/nar/gkx760.

[82] N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, and N. Friedman, "McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences," *Bioinformatics,* vol. 33, no. 18, pp. 2924-2929, Sep 15 2017, doi: 10.1093/bioinformatics/btx286.

[83] W. Chaara, A. Gonzalez-Tort, L. M. Florez, D. Klatzmann, E. Mariotti-Ferrandiz, and A. Six, "RepSeq Data Representativeness and Robustness Assessment by Shannon Entropy," *Front Immunol,* vol. 9, p. 1038, 2018, doi: 10.3389/fimmu.2018.01038.

[84] J. Kaplinsky and R. Arnaout, "Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples," *Nat Commun,* vol. 7, p. 11881, Jun 15 2016, doi: 10.1038/ncomms11881.

[85] D. J. Laydon *et al.*, "Quantification of HTLV-1 clonality and TCR diversity," *PLoS Comput Biol,* vol. 10, no. 6, p. e1003646, Jun 2014, doi: 10.1371/journal.pcbi.1003646.

[86] S. M. Lee and A. Chao, "Estimating population size via sample coverage for closed capture-recapture models," *Biometrics,* vol. 50, no. 1, pp. 88-97, Mar 1994. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/19480084>.

[87] D. R. Thapa *et al.*, "Longitudinal analysis of peripheral blood T cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing," *Arthritis Res Ther,* vol. 17, p. 132, May 23 2015, doi: 10.1186/s13075-015-0655-9.

[88] T. R. Matos *et al.*, "Clinically resolved psoriatic lesions contain psoriasis-specific IL-17-producing alphabeta T cell clones," *J Clin Invest,* vol. 127, no. 11, pp. 4031-4041, Nov 1 2017, doi: 10.1172/JCI93396.

[89] M. Allez *et al.*, "T cell clonal expansions in ileal Crohn's disease are associated with smoking behaviour and postoperative recurrence," *Gut,* vol. 68, no. 11, pp. 1961-1970, Nov 2019, doi: 10.1136/gutjnl-2018-317878.

[90] R. N. Amaria *et al.*, "Neoadjuvant immune checkpoint blockade in high-risk resectable melanoma," *Nat Med,* vol. 24, no. 11, pp. 1649-1654, Nov 2018, doi: 10.1038/s41591-018-0197-1.

[91] P. M. Forde *et al.*, "Neoadjuvant PD-1 Blockade in Resectable Lung Cancer," *N Engl J Med,* vol. 378, no. 21, pp. 1976-1986, May 24 2018, doi: 10.1056/NEJMoa1716078.

[92] W. Roh *et al.*, "Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance," *Sci Transl Med,* vol. 9, no. 379, Mar 1 2017, doi: 10.1126/scitranslmed.aah3560.

[93] D. Y. Oh *et al.*, "Immune Toxicities Elicted by CTLA-4 Blockade in Cancer Patients Are Associated with Early Diversification of the T-cell Repertoire," *Cancer Res,* vol. 77, no. 6, pp. 1322-1330, Mar 15 2017, doi: 10.1158/0008-5472.CAN-16-2324.

[94] S. K. Subudhi *et al.*, "Clonal expansion of CD8 T cells in the systemic circulation precedes development of ipilimumab-induced toxicities," *Proc Natl Acad Sci U S A,* vol. 113, no. 42, pp. 11919-11924, Oct 18 2016, doi: 10.1073/pnas.1611421113.

[95] L. Robert *et al.*, "CTLA4 blockade broadens the peripheral T-cell receptor repertoire," *Clin Cancer Res,* vol. 20, no. 9, pp. 2424-32, May 1 2014, doi: 10.1158/1078-0432.CCR-13-2648.

[96] O. V. Britanova *et al.*, "Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians," *J Immunol,* vol. 196, no. 12, pp. 5005-13, Jun 15 2016, doi: 10.4049/jimmunol.1600005.

[97] W. Tu and S. Rao, "Mechanisms Underlying T Cell Immunosenescence: Aging and Cytomegalovirus Infection," *Front Microbiol,* vol. 7, p. 2111, 2016, doi: 10.3389/fmicb.2016.02111.

[98] V. P. Argaet *et al.*, "Dominant selection of an invariant T cell antigen receptor in response to persistent infection by Epstein-Barr virus," *J Exp Med,* vol. 180, no. 6, pp. 2335-40, Dec 1 1994, doi: 10.1084/jem.180.6.2335.

[99] R. Cibotti *et al.*, "Public and private V beta T cell receptor repertoires against hen egg white lysozyme (HEL) in nontransgenic versus HEL transgenic mice," *J Exp Med,* vol. 180, no. 3, pp. 861-72, Sep 1 1994, doi: 10.1084/jem.180.3.861.

[100] N. E. Annels, M. F. Callan, L. Tan, and A. B. Rickinson, "Changing patterns of dominant TCR usage with maturation of an EBV-specific cytotoxic T cell response," *J Immunol,* vol. 165, no. 9, pp. 4831-41, Nov 1 2000, doi: 10.4049/jimmunol.165.9.4831.

[101] D. A. Price *et al.*, "T cell receptor recognition motifs govern immune escape patterns in acute SIV infection," *Immunity,* vol. 21, no. 6, pp. 793-803, Dec 2004, doi: 10.1016/j.immuni.2004.10.010.

[102] H. S. Robins *et al.*, "Overlap and effective size of the human CD8+ T cell receptor repertoire," *Sci Transl Med,* vol. 2, no. 47, p. 47ra64, Sep 1 2010, doi: 10.1126/scitranslmed.3001442.

[103] P. Bousso *et al.*, "Individual variations in the murine T cell response to a specific peptide reflect variability in naive repertoires," *Immunity,* vol. 9, no. 2, pp. 169-78, Aug 1998, doi: 10.1016/s1074-7613(00)80599-3.

[104] S. J. Turner, P. C. Doherty, J. McCluskey, and J. Rossjohn, "Structural determinants of T-cell receptor bias in immunity," *Nat Rev Immunol,* vol. 6, no. 12, pp. 883-94, Dec 2006, doi: 10.1038/nri1977.

[105] L. Kjer-Nielsen *et al.*, "A structural basis for the selection of dominant alphabeta T cell receptors in antiviral immunity," *Immunity,* vol. 18, no. 1, pp. 53-64, Jan 2003, doi: 10.1016/s1074-7613(02)00513-7.

[106] D. A. Price *et al.*, "Avidity for antigen shapes clonal dominance in CD8+ T cell populations specific for persistent DNA viruses," *J Exp Med,* vol. 202, no. 10, pp. 1349-61, Nov 21 2005, doi: 10.1084/jem.20051357.

[107] V. Venturi *et al.*, "Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination," *Proc Natl Acad Sci U S A,* vol. 103, no. 49, pp. 18691-6, Dec 5 2006, doi: 10.1073/pnas.0608907103.

[108] V. Venturi *et al.*, "A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing," *J Immunol,* vol. 186, no. 7, pp. 4285-94, Apr 1 2011, doi: 10.4049/jimmunol.1003898.

[109] V. Venturi, D. A. Price, D. C. Douek, and M. P. Davenport, "The molecular basis for public T-cell responses?," *Nat Rev Immunol,* vol. 8, no. 3, pp. 231-8, Mar 2008, doi: 10.1038/nri2260.

[110] H. Li, C. Ye, G. Ji, and J. Han, "Determinants of public T cell responses," *Cell Res,* vol. 22, no. 1, pp. 33-42, Jan 2012, doi: 10.1038/cr.2012.1.

[111] M. F. Quigley *et al.*, "Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire," *Proc Natl Acad Sci U S A,* vol. 107, no. 45, pp. 19414-9, Nov 9 2010, doi: 10.1073/pnas.1010586107.

[112] V. Venturi, H. Y. Chin, D. A. Price, D. C. Douek, and M. P. Davenport, "The role of production frequency in the sharing of simian immunodeficiency virus-specific CD8+ TCRs between macaques," *J Immunol,* vol. 181, no. 4, pp. 2597-609, Aug 15 2008, doi: 10.4049/jimmunol.181.4.2597.

[113] V. Venturi *et al.*, "TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV," *J Immunol,* vol. 181, no. 11, pp. 7853-62, Dec 1 2008, doi: 10.4049/jimmunol.181.11.7853.

[114] H. Li *et al.*, "Recombinatorial biases and convergent recombination determine interindividual TCRbeta sharing in murine thymocytes," *J Immunol,* vol. 189, no. 5, pp. 2404-13, Sep 1 2012, doi: 10.4049/jimmunol.1102087.

[115] A. L. Furmanski *et al.*, "Public T cell receptor beta-chains are not advantaged during positive selection," *J Immunol,* vol. 180, no. 2, pp. 1029-39, Jan 15 2008, doi: 10.4049/jimmunol.180.2.1029.

[116] C. Baruah, P. Devi, and D. K. Sharma, "Sequence Analysis and Structure Prediction of SARS-CoV-2 Accessory Proteins 9b and ORF14: Evolutionary Analysis Indicates Close Relatedness to Bat Coronavirus," *Biomed Res Int,* vol. 2020, p. 7234961, 2020, doi: 10.1155/2020/7234961.

[117] D. E. Gordon *et al.*, "A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing," *bioRxiv,* Mar 22 2020, doi: 10.1101/2020.03.22.002386.

[118] L. K. Clark, T. J. Green, and C. M. Petit, "Structure of nonstructural protein 1 from SARS-CoV-2," *bioRxiv,* Nov 3 2020, doi: 10.1101/2020.11.03.366757.

[119] E. Tazikeh-Lemeski, S. Moradi, R. Raoufi, M. Shahlaei, M. A. M. Janlou, and S. Zolghadri, "Targeting SARS-COV-2 non-structural protein 16: a virtual drug repurposing study," *J Biomol Struct Dyn,* pp. 1-14, Jun 23 2020, doi: 10.1080/07391102.2020.1779133.

[120] X. Lei *et al.*, "Activation and evasion of type I interferon responses by SARS-CoV-2," *Nat Commun,* vol. 11, no. 1, p. 3810, Jul 30 2020, doi: 10.1038/s41467-020-17665-9.

[121] C. G. K. Ziegler *et al.*, "SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues," *Cell,* vol. 181, no. 5, pp. 1016-1035 e19, May 28 2020, doi: 10.1016/j.cell.2020.04.035.

[122] S. C. Oliveira, M. T. Q. de Magalhaes, and E. J. Homan, "Immunoinformatic Analysis of SARS-CoV-2 Nucleocapsid Protein and Identification of COVID-19 Vaccine Targets," *Front Immunol,* vol. 11, p. 587615, 2020, doi: 10.3389/fimmu.2020.587615.

[123] A. Grifoni *et al.*, "Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals," *Cell,* vol. 181, no. 7, pp. 1489-1501 e15, Jun 25 2020, doi: 10.1016/j.cell.2020.05.015.

[124] D. F. Robbiani *et al.*, "Convergent Antibody Responses to SARS-CoV-2 Infection in Convalescent Individuals," *bioRxiv,* May 22 2020, doi: 10.1101/2020.05.13.092619.

[125] C. Rydyznski Moderbacher *et al.*, "Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and Associations with Age and Disease Severity," *Cell,* vol. 183, no. 4, pp. 996-1012 e19, Nov 12 2020, doi: 10.1016/j.cell.2020.09.038.

[126] P. J. M. Brouwer *et al.*, "Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability," *Science,* vol. 369, no. 6504, pp. 643-650, Aug 7 2020, doi: 10.1126/science.abc5902.

[127] S. Nolan *et al.*, "A large-scale database of T-cell receptor beta (TCRbeta) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2," *Res Sq,* Aug 4 2020, doi: 10.21203/rs.3.rs-51964/v1.

[128] A. S. Shomuradova *et al.*, "SARS-CoV-2 Epitopes Are Recognized by a Public and Diverse Repertoire of Human T Cell Receptors," *Immunity,* vol. 53, no. 6, pp. 1245-1257 e5, Dec 15 2020, doi: 10.1016/j.immuni.2020.11.004.