*Математическая биология и биоинформатика*
*2022. Т. 17. № 2. С. 325–337. doi: 10.17537/2022.17.325*

========== БИОИНФОРМАТИКА ==========

# Учет неравновесного сцепления при подборе панелей таргетного секвенирования

**Романов Д.Е.***[1,2], **Скобликов Н.Э.**[1,3,4]

[1]*Медицинская лаборатория CL, Краснодар, Россия*
[2]*Южный федеральный университет, Ростов-на-Дону, Россия*
[3]*Кубанский государственный медицинский университет, Краснодар, Россия*
[4]*Краснодарский научный центр по зоотехнии и ветеринарии, Краснодар, Россия*

***Аннотация.*** В работе предложен способ оптимизации разработки генодиагностических панелей на основе построения карт неравновесного сцепления. Подбор генов осуществляется на основании результатов полногеномного анализа ассоциаций (GWAS). Полногеномный анализ ассоциаций позволяет выявлять связь геномных вариантов с изучаемым фенотипом. Однако, нуклеотидные варианты, продемонстрировавшие наибольшую степень ассоциации, могут быть связаны с фенотипом лишь статистически, не являясь истинной причиной проявления фенотипа. При этом они могут оказаться в блоке сцепленного наследования с нуклеотидными вариантами, реально влияющими на проявление фенотипа. Построение карт неравновесного сцепления нуклеотидов позволяет оптимальным способом определить границы блоков сцепления, в который попадают искомые варианты. Целью данного исследования была оптимизация определения границ геномных локусов для создания таргетных панелей, направленных на предикцию восприимчивости к SARS-CoV-2 и тяжести течения COVID-19. Предложенная в данной работе методика подбора локусов для таргетной панели с учётом неравновесного сцепления позволяет использовать явление неравновесного сцепления с целью максимально охватить участки, задействованные в развитии фенотипа, с одновременной минимизацией длины этих участков, а вместе с тем и затрат на секвенирование.

***Ключевые слова:*** *COVID-19, таргетное секвенирование, неравновесное сцепление.*

## ВВЕДЕНИЕ

Пандемия COVID-19, вызванная коронавирусом SARS-CoV-2, и в 2022 остаётся одной из самых актуальных тем исследования современной биомедицины. Чрезвычайно широкий спектр клинических форм инфекции – от бессимптомных форм до тяжелейших полиорганных поражений с летальными исходами [1] – послужил основанием для предположения о том, что такое разнообразие обеспечивается высокой мутационной активностью вируса [2]. В числе факторов, предопределяющих течение и исход инфекции, ожидалась повышенная вероятность возникновения тяжёлых форм у пациентов с ослабленной или иным образом нарушенной регуляцией защитных систем организма, в первую очередь – иммунной системы и системы гемостаза [3, 4]. Однако, уже в первый период пандемии отмечалось, что значительная часть (по разным данным – от 25 до 60 %) пациентов (в том числе – входящих в «группу риска») с ПЦР-детектированным в носоглотке присутствием генных маркеров SARS-CoV-2 не

*rdme@yandex.ru

демонстрирует никаких симптомов инфекции. Позднее было высказано предположение, что такой феномен обусловлен генетическими особенностями самих пациентов [5]. Генетическими же особенностями самих пациентов стала предположительно объясняться и непредсказуемость развивающейся формы инфекции [6].

Как правило, для определения каких-либо генетических факторов предрасположенности к заболеваниям применяют метод полногеномного анализа ассоциаций. Полногеномный анализ ассоциаций (англ. genome-wide association studies, GWAS) – совокупность статистических методов, направленных на выявление связи между геномными вариантами и фенотипическими признаками.

Часто под полногеномным поиском ассоциаций подразумевают только поиск связей между однонуклеотидными полиморфизмами (англ. single-nucleotide polymorphism, SNP) в рамках всего генома и проявлением какого-либо фенотипа, в т.ч. восприимчивости или тяжести развития определённой патологии.

Появление методов генотипирования таких, как ДНК-микрочипирование (англ. DNA microarray, гибридизационные панели), позволяющий на основе явления гибридизации олигонуклеотидов выявить полиморфизмы в наиболее вариабельных позициях в геноме или секвенирование нового поколения (англ. next generation sequencing, NGS), который дает возможность получить последовательность полного генома, сделало возможным проведение статистического анализа частот встречаемости нуклеотидов в каждой полученной позиции среди индивидуумов в различных группах, например, в группе носителей патологии (кейсов) по сравнению с группой контроля. Статистический анализ заключается в построении моделей логистической регрессии.

Так, в исследовании [7] на основании полногеномного анализа ассоциаций на выборке из 1980 пациентов с подтвержденным SARS-CoV-2 и диагнозом COVID-19 и 2381 контроля, тяжесть течения COVID-19 в основном ассоциирована с двумя геномными локусами: 3p21.31 и 9q34.2 с наиболее ассоциированными полиморфизмами rs11385942 и rs657152. Интересно, что эволюционная история выявленных хромосомных локусов указывает на их возможное происхождение от неандертальских геномов [8, 9].

В статье [10] были применены аналогичные методы к выборке из 2244 пациентов в критическом состоянии и 477741 контроля. Значимые сигналы ассоциации были выявлены в локусе 12q24.13, содержащем кластер генов, кодирующих рестрикционные ферменты активации антивирусной защиты *OAS1*, *OAS2* и *OAS3*, в локусе 19p13.2 в окрестности гена тирозин-киназы 2 *TYK2*, в локусе 19p13.3 внутри гена *DPP9*, кодирующего дипептидил пептидазу 9, и в локусе 21q22.1 около гена рецептора интерферона *IFNAR2*.

В работе [11] уже был исследован полный геном 7491 пациента в критическом состоянии и 48400 контролей. Секвенирование было проведено на платформах Illumina Hiseq X и NovaSeq. Анализ ассоциаций был выполнен с помощью пакета программ SAIGE посредством построения логистической смешанной регрессионной модели с поправкой на пол, возраст и 20 главных компонент. По результатам анализа было подтверждено 23 известных ранее и обнаружено 16 новых ассоциаций, включая варианты в генах *IL10RB* и *PLSCR1*, вовлеченных в сигнальный путь интерферона, в гене *BCL11A*, задействованом в дифференциации лейкоцитов и в гене *FUT2*, отвечающего за экспрессию антигенов группы крови.

В данном исследовании были использованы результаты метаанализа 46 исследований GWAS, в который вошли 49562 случая и 2 миллиона контролей и было выявлено 13 локусов, продемонстрировавших значимую ассоциацию с восприимчивостью к SARS-CoV-2 или с тяжестью течения COVID-19 [12].

Также в данное исследование были включены результаты менее масштабного

метаанализа, выполненного на более разнородной популяционной выборке жителей Объединенных Арабских Эмиратов, что позволило рассмотреть еще 8 значимых вариантов в локусах, содержащих гены, экспрессирующиеся в тканях легких [13].

Важно отметить, что большая часть наиболее ассоциированных вариантов приходилась на некодирующую часть генома, что свидетельствует в пользу мультифакторной природы заболевания [14]. Большинство этих вариантов локализуется в интронных последовательностях, но были и варианты, находившиеся в межгенном пространстве.

Следует отметить что, полиморфизмы, продемонстрировавшие по результатам GWAS наибольшую степень ассоциации (т.н. лидирующие варианты), не всегда оказываются связанными с причиной проявления фенотипа в отличие от каузальных вариантов, мутации в которых вызывают развитие исследуемого фенотипа.

Если же генотипирование проводится с помощью гибридизационных панелей, то следует учитывать, что они содержат фиксированный набор полиморфизмов и могут, таким образом, вообще не включать каузальные варианты.

Тем не менее, лидирующие и каузальные варианты могут оказаться сцепленными друг с другом и, таким образом, лидирующий вариант может выступатьпрокси-вариантом по отношению к каузальному.

В отсутствие рекомбинации все нуклеотиды, находящиеся на одной хромосоме, наследовались бы всегда одновременно. Однако, ввиду мейотического кроссинговера нуклеотиды, ранее находившиеся на одной хромосоме, могут разойтись по разным гаметам. Частота кроссинговера в различных частях генома оказывается неодинаковой — существуют т.н. «горячие точки» рекомбинации. Это приводит к тому, что весь геном разбивается на рекомбинирующие между собой блоки, причем рекомбинация внутри таких блоков маловероятна. Каждый блок образует гаплотип (сокращение от «гаплоидный генотип») и наследуется единым целым. В связи с этим, нуклеотиды внутри гаплотипа оказываются в неравновесном сцеплении друг с другом, что позволяет по одному известному нуклеотиду гаплотипа с высокой долей вероятности восстановить все остальные нуклеотиды в этом гаплотипе в отличие от равновесного сцепления нуклеотидов, когда происходит их независимое наследование.

Как правило, сцепленность нуклеотидов измеряют как квадрат коэффициента корреляции между нуклеотидами. Международный проект HapMap позволил рассчитать карты гаплотипов генома человека на основе данных с гибридизационных панелей [15]. Данные проекта 1000 Genomes, полученные по результатам секвенирования, показали высокую степень сходства с данными проекта HapMap и позволили рассчитать сцепленность нуклеотидов с учетом редких аллелей [16].

Таким образом, ввиду явления неравновесного сцепления нуклеотидов, заключающегося в высокой частоте сопоявления соответствующих аллелей, становится возможным по лидирующему варианту локализовать локус, в котором может находится каузальный вариант [17, 18].

Для построения карт неравновесного сцепления был разработан ряд программных средств [19, 20, 21]. Поскольку для построения карты неравновесного сцепления требуется произвести сравнения частот сопоявления нуклеотидов по принципу «каждый с каждым», то количество вычислительных ресурсов растет пропорционально квадрату количества рассмотренных полиморфизмов. В связи с этим, поднимается вопрос об эффективности расчетов.

На данный момент наибольшую эффективность с точки зрения расхода памяти и процессорного времени демонстрирует программа LDBlockShow, позволяющая в течение нескольких секунд с расходом памяти около 1 гигабайта получить на современной ЭВМ

карту неравновесного сцепления 60000 полиморфизмов [21].

Однако, практически все имеющиеся инструменты для построения карт неравновесного сцепления имеют ограниченные возможности по отображению геномных аннотаций в изучаемой окрестности, в частности, информации о расположении генов, регуляторных элементов и проч.

С другой стороны, такие инструменты, как геномный браузер Ensembl [22] или геномный браузер UCSC [23], позволяют подробно отобразить геномный контекст изучаемого локуса, однако имеют ряд ограничений по отображению информации о неравновесном сцеплении. В частности, геномный браузер Ensembl позволяет строить карты неравновесного сцепления для окрестности длиной не более 75 тыс. н.п., причем выбор данных генотипирования ограничен только данными фазы 3 проекта 1000 Genomes.

Если локализация лидирующего варианта неизвестна в пределах гена, то в целях поиска каузального варианта прибегают к таргетному ресеквенированию последовательности всего гена, что дает возможность проанализировать вообще все полиморфизмы в пределах данного гена [24]. Для этого исследуемые регионы предварительно амплифицируют (т.е., получают большое число копий ДНК региона) методом ПЦР протяжённых участков ДНК (10 тысяч и более оснований), что обеспечивает высокое покрытие при низкой стоимости и высокой скорости секвенирования. Если исследуемых регионов несколько, то указанная последовательность действий — амплификация и секвенирование — проделывается для каждого из них. В результате вся совокупность исследуемых участков образует т.н. таргетную панель.

Таргетные панели используются там, где требуется детальное изучение небольшого числа регионов, которые могут нести клинически значимые мутации, например, при исследовании раковых заболеваний [25] и идентификации наследственных заболеваний [26]. Кроме того, это позволяет быстро и дешево выявлять популяционные особенности исследуемых регионов [27]. В случае, если лидирующий полиморфизм локализуется в межгенном пространстве, имеется вариативность в выборе участка ДНК для более детального анализа.

Анализ карты неравновесного сцепления дает возможность по лидирующему полиморфизму определить блок неравновесного сцепления. В свою очередь, таргетное ресеквенирование блока неравновесного сцепления позволяет получить все варианты в исследуемом участке, среди которых с высокой долей вероятности может находится каузальный вариант [24, 28, 29].

В связи с тем, что большинство полиморфизмов, ассоциированных с восприимчивостью к SARS-CoV-2 или с тяжестью течения COVID-19, приходятся на некодирующий геном [12, 13], в данном исследовании был осуществлен подбор участков для создания таргетной панели к COVID-19 с учетом неравновесного сцепления.

Создание такой панели позволит быстро провести скрининг жителей южных регионов России на наличие как известных вариантов, ассоциированных с COVID-19, так и более детально исследовать указанные локусы с учетом популяционных особенностей.

## МАТЕРИАЛЫ И МЕТОДЫ

В исследование вошли 13 полиморфизмов (табл. 1, верхняя часть), показавших значимую ассоциацию с тяжестью течения COVID-19 или восприимчивостью к SARS-CoV-2 [12]. Помимо этого, были рассмотрены 8 полиморфизмов (табл. 1, нижняя часть), продемонстрировавших высокую степень ассоциации с тяжестью течения COVID-19 в популяции жителей Объединенных Арабских Эмиратов [13]. Были изучены

паттерны неравновесного сцепления в геномных окрестностях этих полиморфизмов. Под окрестностью полиморфизма понимался участок с фланками длиной 100000 н.п. от позиции полиморфизма.

**Таблица 1.** Полиморфизмы, показавшие ассоциацию с тяжестью течения или восприимчивостью к COVID-19 [12, 13]. Для каждого полиморфизма приведено клиническое состояние (фенотип), продемонстрировавшее наиболее значимую ассоциацию. Везде приведены нескорректированные P-значения. Геномные координаты даны для версии генома GRCh38

| Полиморфизм | Хромосома | Позиция | Клиническое состояние / Фенотип | P-значение |
|---|---|---|---|---|
| rs2271616 | 3 | 45796521 | подтверждённое инфицирование | $1.79 \cdot 10^{-34}$ |
| rs10490770 | 3 | 45823240 | критическое состояние | $2.20 \cdot 10^{-61}$ |
| rs11919389 | 3 | 101705614 | подтверждённое инфицирование | $3.46 \cdot 10^{-15}$ |
| rs1886814 | 6 | 41534945 | госпитализация | $1.11 \cdot 10^{-9}$ |
| rs72711165 | 8 | 124324323 | госпитализация | $2.13 \cdot 10^{-9}$ |
| rs912805253 | 9 | 133274084 | подтверждённое инфицирование | $1.45 \cdot 10^{-39}$ |
| rs10774671 | 12 | 112919388 | критическое состояние | $4.08 \cdot 10^{-13}$ |
| rs1819040 | 17 | 46142465 | госпитализация | $1.83 \cdot 10^{-10}$ |
| rs77534576 | 17 | 49863303 | критическое состояние | $4.37 \cdot 10^{-9}$ |
| rs2109069 | 19 | 4719431 | критическое состояние | $9.68 \cdot 10^{-22}$ |
| rs74956615 | 19 | 10317045 | критическое состояние | $9.71 \cdot 10^{-12}$ |
| rs4801778 | 19 | 48867352 | подтверждённое инфицирование | $1.18 \cdot 10^{-8}$ |
| rs13050728 | 21 | 33242905 | госпитализация | $2.72 \cdot 10^{-20}$ |
| rs7605851 | 2 | 136948487 | госпитализация | $3.07 \cdot 10^{-6}$ |
| rs7595310 | 2 | 167953628 | госпитализация | $4.55 \cdot 10^{-6}$ |
| rs7715119 | 5 | 77379594 | госпитализация | $2.19 \cdot 10^{-6}$ |
| rs72953026 | 11 | 88436432 | госпитализация | $2.38 \cdot 10^{-6}$ |
| rs10507497 | 13 | 41746692 | госпитализация | $9.54 \cdot 10^{-7}$ |
| rs599976 | 13 | 102695867 | госпитализация | $8.95 \cdot 10^{-6}$ |
| rs10140801 | 14 | 55263687 | госпитализация | $8.26 \cdot 10^{-6}$ |
| rs11659676 | 18 | 6461232 | госпитализация | $8.88 \cdot 10^{-6}$ |

Генотипы геномных окрестностей каждого из рассмотренных полиморфизмов были извлечены из набора данных 1000 Genomes 30x on GRCh38 [30], содержащего 3202 образца из 26 популяций. Для этого по геномным координатам окрестностей с помощью утилиты tabix из набора утилит Samtools с сайта проекта были извлечены генотипы окрестностей в виде vcf-файлов.

Карты неравновесного сцепления были построены с помощью набора программ LDBlockShow [21] отдельно для всей группы популяций (3202 образца), только для европейской группы (EUR, 633 образца) и только для центрально-европейской группы (CEU, 179 образцов).

Для каждого полиморфизма с помощью реализации скриптов на языке bash были получены снимки окрестности этого полиморфизма из геномного браузера UCSC с позициями генов и других геномных элементов.


**Рис. 1.** Окрестность полиморфизма rs10490770. Снимок геномного браузера UCSC окрестности полиморфизма наложен на карту неравновесного сцепления нуклеотидов. Отобранный локус обозначен синим прямоугольником и захватывает 3′-UTR гена *LZTFL1*.

Полученные изображения в виде png-файлов были совмещены с помощью утилиты convert из набора программ ImageMagick. Относительные смещения изображений

**Таблица 2.** Геномные координаты локусов для создания таргетной панели к COVID-19, подобранных с учетом неравновесного сцепления и содержащих исследуемые полиморфизмы. Геномные координаты даны для версии генома GRCh38

| Полимор-физм | Хро-мо-сома | Позиция | Левая граница локуса | Правая граница локуса | Длина локуса | Геномное положение локуса |
|---|---|---|---|---|---|---|
| rs10490770 | 3 | 45823240 | 45793240 | 45843240 | 50000 | chr3:45793240-45843240 |
| rs11919389 | 3 | 101705614 | 101700614 | 101720614 | 20000 | chr3:101700614-101720614 |
| rs1886814 | 6 | 41534945 | 41530795 | 41539095 | 8300 | chr6:41530795-41539095 |
| rs72711165 | 8 | 124324323 | 124306823 | 124333073 | 26250 | chr8:124306823-124333073 |
| rs912805253 | 9 | 133274084 | 133266584 | 133282834 | 16250 | chr9:133266584-133282834 |
| rs10774671 | 12 | 112919388 | 112909388 | 112936888 | 27500 | chr12:112909388-112936888 |
| rs1819040 | 17 | 46142465 | 46139965 | 46144965 | 5000 | chr17:46139965-46144965 |
| rs77534576 | 17 | 49863303 | 49858303 | 49870803 | 12500 | chr17:49858303-49870803 |
| rs2109069 | 19 | 4719431 | 4714431 | 4729431 | 15000 | chr19:4714431-4729431 |
| rs74956615 | 19 | 10317045 | 10304545 | 10342045 | 37500 | chr19:10304545-10342045 |
| rs4801778 | 19 | 48867352 | 48857352 | 48873602 | 16250 | chr19:48857352-48873602 |
| rs13050728 | 21 | 33242905 | 33217905 | 33262905 | 45000 | chr21:33217905-33262905 |
| rs7605851 | 2 | 136948487 | 136942237 | 136963487 | 21250 | chr2:136942237-136963487 |
| rs7595310 | 2 | 167953628 | 167948628 | 167957378 | 8750 | chr2:167948628-167957378 |
| rs7715119 | 5 | 77379594 | 77375844 | 77394594 | 18750 | chr5:77375844-77394594 |
| rs72953026 | 11 | 88436432 | 88426432 | 88441432 | 15000 | chr11:88426432-88441432 |
| rs10507497 | 13 | 41746692 | 41741692 | 41754192 | 12500 | chr13:41741692-41754192 |
| rs599976 | 13 | 102695867 | 102690867 | 102703367 | 12500 | chr13:102690867-102703367 |
| rs10140801 | 14 | 55263687 | 55258687 | 55271187 | 12500 | chr14:55258687-55271187 |
| rs11659676 | 18 | 6461232 | 6451232 | 6463732 | 12500 | chr18:6451232-6463732 |

были подобраны таким образом, чтобы экранные координаты окрестностей на обоих изображениях совпадали друг с другом.

На основании визуального анализа карты неравновесного сцепления были определены границы блоков сцепления в окрестностях исследуемых полиморфизмов. Границами локусов для создания панели выступали границы блоков сцепления, содержавших соответствующий полиморфизм. Полученные границы локусов были дополнительно скорректированы с учетом взаимного расположения геномных элементов

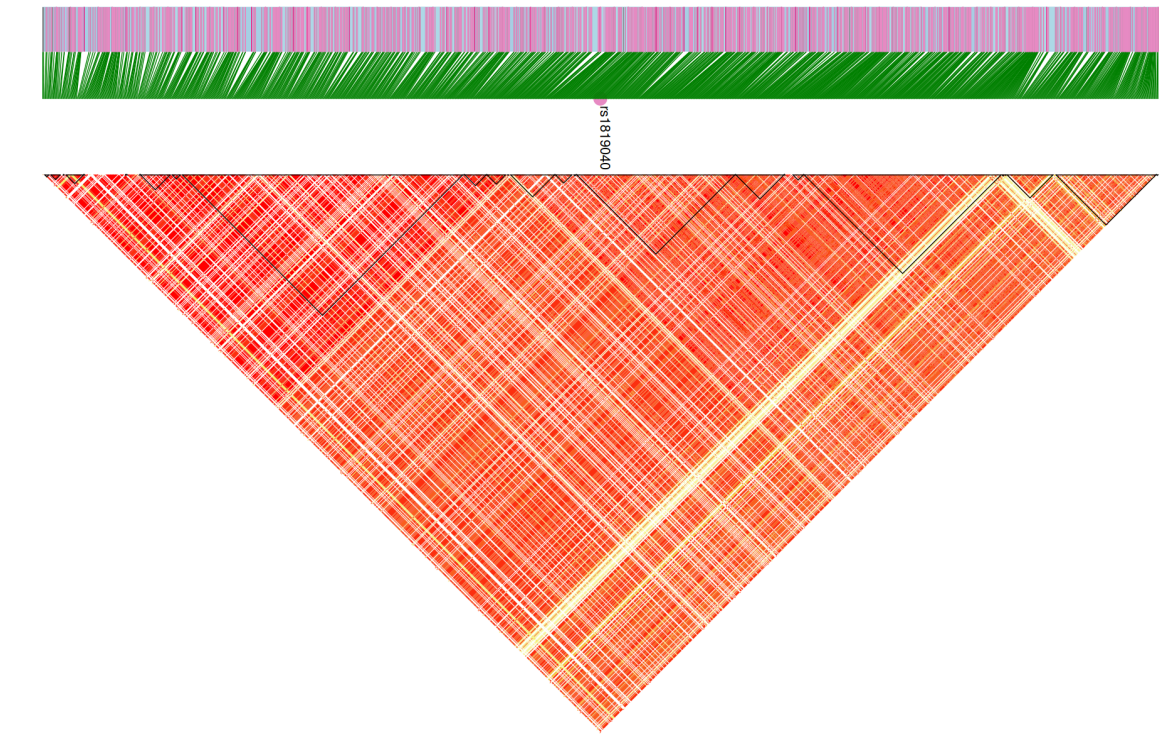
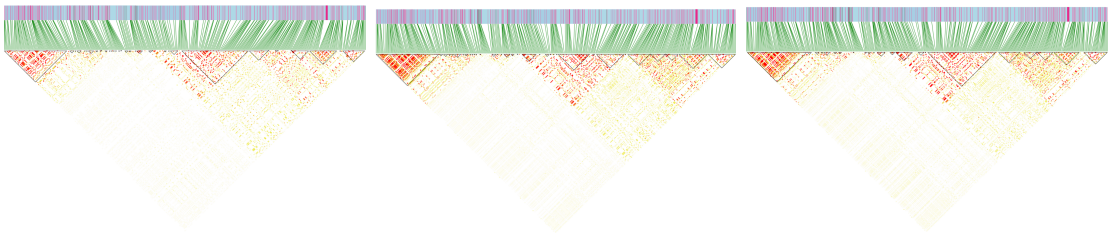



**Рис. 2.** Неравновесное сцепление нуклеотидов в окрестности полиморфизма rs1819040. Вся окрестность данного полиморфизма попадает в одну протяженную группу сцепления.



(a) (b) (c)

**Рис. 3.** Неравновесное сцепление нуклеотидов в окрестности полиморфизма rs10490770 в объединенной группе популяций (a), в европейской группе (b) и центрально-европейской группе (c). Сцепленность нуклеотидов слабо зависит от популяции.

в окрестностях исследуемых полиморфизмов.

## РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Описание подобранных локусов дано в таблице 2. Суммарная длина локусов составила 393300 н.п. На рисунке 1 приведена окрестность полиморфизма rs10490770 и отобранный для панели локус, содержащий данный полиморфизм. Следует отметить, что подобранный локус захватывает весь 3′-UTR гена фактора транскрипции лейциновой молнии *LZTFL1*, известного своей связью с COVID-19 [31, 32]. Примечательно, что полиморфизм rs2109069, находящийся в окрестности полиморфизма rs10490770 и расположенный в 5′-UTR гена *SLC6A20*, оказывается несцепленным с полиморфизмом rs10490770, и его ассоциация с тяжестью течения COVID-19 может носить опосредованный характер [33, 34]. В связи с этим данный полиморфизм был исключен из создаваемой панели.

В случае, если группа сцепления охватывала всю окрестность полиморфизма, то длина подбираемого локуса была ограничена минимальным значением 2500 н.п. в каждую сторону от позиции полиморфизма (рис. 2).

Карты неравновесного сцепления для объединенной группы популяций, европейской группы (EUR) и центрально-европейской группы (CEU) показали высокую степень сходства друг с другом (рис. 3). На основании этого следует ожидать, что выявленные границы локусов будут сохраняться и в других популяциях.

Таким образом, предложенная в данной работе методика подбора локусов для таргетной панели позволяет использовать явление неравновесного сцепления с целью максимально охватить участки, задействованные в развитии фенотипа, с одновременной минимизацией длины этих участков, а вместе с тем и затрат на секвенирование. Кроме того, наш способ подбора локусов позволяет использовать неравновесное сцепление в качестве еще одного параметра оценки корректности секвенирования отобранных локусов. Например, при наблюдении аномального сцепления нуклеотидов в панели следует уделить повышенное внимание контролю качества образцов. Для дальнейшей оптимизации предлагаемой методики может дополнительно учитываться критерий консервативности последовательности. К примеру, если отобранный локус делится на консервативную и эволюционно-вариабельную часть, то этот локус следует ограничить более консервативной частью.

Исследование выполнено при финансовой поддержке Кубанского научного фонда и ООО «СЛ МедикалГруп» в рамках научного проекта МФИ-П-20.1/10, а также при финансовой поддержке РФФИ и Краснодарского края в рамках научного проекта №19-44-230040-р_а.

## СПИСОК ЛИТЕРАТУРЫ


1. Li Q., Guan X., Wu P., Wang X., Zhou L., Tong Y., Ren R., Leung K.S.M., Lau E.H.Y., Wong J.Y. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.* 2020. V. 382. No. 13. P. 1199–1207. doi: 10.1056/nejmoa2001316
2. Rahimi A., Mirzazadeh A., Tavakolpour S. Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. *Genomics*. 2021. V. 113. No. 1. P. 1221–1232. doi: 10.1016/j.ygeno.2020.09.059
3. Мойсова Д.Л., Городин В.Н., Скобликов Н.Э., Зотов С.В., Тихоненко Ю.В. Особенности полиморфизма некоторых генов системы гемостаза у больных с COVID-19. *Медицинский вестник Башкортостана*. 2021. Т. 16. № 6. С. 35–40.
4. Liao M., Liu Y., Yuan J., Wen Y., Xu G., Zhao J., Chen L., Li J., Wang X., Wang F. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature Medicine*. 2020. V. 26. No. 6. P. 842–844 doi: 10.1038/s41591-020-0901-9
5. Villapol S. Gastrointestinal symptoms associated with COVID-19: impact on the gut microbiome. *Transl. Res.* 2020. V. 226. P. 57–69. doi: 10.1016/j.trsl.2020.08.004
6. Matzaraki V., Kumar V., Wijmenga C., Zhernakova A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biology*. 2021. V. 18. No. 1. P. 76. doi: 10.1186/s13059-017-1207-1
7. Ellinghaus D., Degenhardt F., Bujanda L., Buti M., Albillos A., Invernizzi P., Fernández J., Prati D., Baselli G., Asselta R. et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N. Engl. J. Med.* 2020. V. 383. No. 16. P. 1522–1534. doi: 10.1056/nejmoa2020283
8. Zeberg H., Pääbo S. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proc. Natl. Acad. Sci. USA*. 2021. V. 118. No. 9. P. e2026309118. doi: 10.1073/pnas.2026309118
9. Zeberg H., Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*. 2020. V. 587. No. 7835. P. 610–612. doi: 10.1038/s41586-020-2818-3
10. Pairo-Castineira E., Clohisey S., Klaric L., Bretherick A.D., Rawlik K., Pasko D., Walker S., Parkinson N., Fourman M.H., Russell C.D. et al. Genetic mechanisms of critical illness in COVID-19. *Nature*. 2021. V. 591. No. 7848. P. 92–98. doi: 10.1038/s41586-020-03065-y
11. Kousathanas A., Pairo-Castineira E., Rawlik K., Stuckey A., Odhams C.A., Walker S., Russell C.D., Malinauskas T., Wu Y., Millar J. et al. Whole genome sequencing reveals host factors underlying critical COVID-19. *Nature*. 2022. V. 607. P. 97–103. doi: 10.1038/s41586-022-04576-6
12. Niemi M.E.K., Karjalainen J., Liao R.G., Neale B.M., Daly M., Ganna A., Pathak G.A., Andrews S.J., Kanai M., Veerapen K. et al. Mapping the human genetic architecture of COVID-19. *Nature*. 2021. V. 600. No. 7889. P. 472–477. doi: 10.1038/s41586-021-03767-x
13. Mousa M., Vurivi H., Kannout H., Uddin M., Alkaabi N., Mahboub B., Tay G.K., Alsafar H.S. Genome-wide association study of hospitalized COVID-19 patients in the United Arab Emirates. *EBioMedicine*. 2021. V. 74. P. 103695. doi: 10.1016/j.ebiom.2021.103695
14. Secolin R., de Araujo T.K., Gonsales M.C., Rocha C.S., Naslavsky M., Marco L., Bicalho M.A.C., Vazquez V.L., Zatz M., Silva W.A., Lopes-Cendes I. Genetic variability in COVID-19-related genes in the Brazilian population. *Hum. Genome. Var.* 2021. V. 8. P. 15. doi: 10.1038/s41439-021-00146-w
15. Consortium International HapMap. A haplotype map of the human genome. *Nature*. 2005. V. 437. No. 7063. P. 1299–1320. doi: 10.1038/nature04226
16. Buchanan C.C., Torstenson E.S., Bush W.S., Ritchie M.D. A comparison of cataloged

variation between International HapMap Consortium and 1000 Genomes Project data. *J. Am. Med. Inform. Assoc.* 2012. V. 19. No. 2. P. 289–294. doi: 10.1136/amiajnl-2011-000652
17. Justice C.M., Musolf A.M., Cuellar A., Lattanzi W., Simeonov E., Kaneva R., Paschall J., Cunningham M., Wilkie A.O.M., Wilson A.F. et al. Targeted Sequencing of Candidate Regions Associated with Sagittal and Metopic Nonsyndromic Craniosynostosis. *Genes*. 2022. V. 13. No. 5. P. 816. doi: 10.3390/genes13050816
18. Schaid D.J., Chen W., Larson N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* 2018. V. 19. No. 8. P. 491–504. doi: 10.1038/s41576-018-0016-z
19. Barrett J.C., Fry B., Maller J., Daly M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005. V. 21. No. 2. P. 263–265. doi: 10.1093/bioinformatics/bth457
20. Shin J.-H., Blay S., McNeney B., Graham J. LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software*. 2006. V. 16. P. 1–9. doi: 10.18637/jss.v016.c03
21. Dong S.S., He W.M., Ji J.J., Zhang C., Guo Y., Yang T.L. LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief. Bioinform.* 2021. V. 22. No. 4. P. bbaa227. doi: 10.1093/bib/bbaa227
22. Birney E., Andrews T.D., Bevan P., Caccamo M., Chen Y., Clarke L., Coates G., Cuff J., Curwen V., Cutts T. et al. An overview of Ensembl. *Genome Research*. 2004. V. 14. No. 5. P. 925–928. doi: 10.1101/gr.1860604
23. Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. The human genome browser at UCSC. *Genome Research*. 2002. V. 12. No. 6. P. 996–1006. doi: 10.1101/gr.229102
24. Kitamoto T., Kitamoto A., Yoneda M., Hyogo H., Ochi H., Mizusawa S., Ueno T., Nakao K., Sekine A., Chayama K. et al. Targeted next-generation sequencing and fine linkage disequilibrium mapping reveals association of PNPLA3 and PARVB with the severity of nonalcoholic fatty liver disease. *J. Hum. Genet.* 2014. V. 59. No. 5. P. 241–246. doi: 10.1038/jhg.2014.17
25. Bewicke-Copley F., Arjun Kumar E., Palladino G., Korfi K., Wang J. Applications and analysis of targeted genomic sequencing in cancer studies. *Comput. Struct. Biotechnol. J.* 2019. V. 17. P. 1348–1359. doi: 10.1016/j.csbj.2019.10.004
26. Qin D. Next-generation sequencing and its clinical application. *Cancer Biol. Med.* 2019. V. 16. No. 1. P. 4–10. doi: 10.20892/j.issn.2095-3941.2018.0055
27. Luo Y., Kanai M., Choi W., Li X., Sakaue S., Yamamoto K., Ogawa K., Gutierrez-Arcelus M., Gregersen P.K., Stuart P.E. et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nature Genetics*. 2021. V. 53. No. 10. P. 1504–1516. doi: 10.1038/s41588-021-00935-7
28. Justice C.M., Kim J., Kim S.D., Kim K., Yagnik G., Cuellar A., Carrington B., Lu C.L., Sood R., Boyadjiev S.A., Wilson A.F. A variant associated with sagittal nonsyndromic craniosynostosis alters the regulatory function of a non-coding element. *Am. J. Med. Genet. A*. 2017. V. 173. No. 11. P. 2893–2897. doi: 10.1002/ajmg.a.38392
29. Justice C.M., Yagnik G., Kim Y., Peter I., Jabs E.W., Erazo M., Ye X., Ainehsazan E., Shi L., Cunningham M.L. et al. A genome-wide association study identifies susceptibility loci for nonsyndromic sagittal craniosynostosis near BMP2 and within BBS9. *Nat. Genet.* 2012. V. 44. No. 12. P. 1360–1364. doi: 10.1038/ng.2463
30. Byrska-Bishop M., Evani U.S., Zhao X., Basile A.O., Abel H.J., Regier A.A., Corvelo A., Clarke W.E., Musunuri R., Nagulapalli K. et al. High coverage whole genome sequencing of

the expanded 1000 Genomes Project cohort including 602 trios. *Cell*. 2022. V. 185. No. 18. P. 3426–3440. doi: 10.1016/j.cell.2022.08.004

31. Downes D.J., Cross A.R., Hua P., Roberts N., Schwessinger R., Cutler A.J., Munis A.M., Brown J., Mielczarek O., de Andrea C.E. et al. Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. *Nat. Genet.* 2021. V. 53. No. 11. P. 1606–1615. doi: 10.1038/s41588-021-00955-3
32. Fink-Baldauf I.M., Stuart W.D., Brewington J.J., Guo M., Maeda Y. CRISPRi links COVID-19 GWAS loci to LZTFL1 and RAVER1. *EBioMedicine*. 2022. V. 75. P. 103806. doi: 10.1016/j.ebiom.2021.103806
33. Kasela S., Daniloski Z., Bollepalli S., Jordan T.X., tenOever B.R., Sanjana N.E., Lappalainen T. Integrative approach identifies SLC6A20 and CXCR6 as putative causal genes for the COVID-19 GWAS signal in the 3p21.31 locus. *Genome Biol.* 2021. V. 22. No. 1. P. 242. doi: 10.1186/s13059-021-02454-4
34. Semiz S. SIT1 transporter as a potential novel target in treatment of COVID-19. *Biomol. Concepts*. 2021. V. 12. No. 1. P. 156–163. doi: 10.1515/bmc-2021-0017

Рукопись поступила в редакцию 31.08.2022.

Переработанный вариант поступил 10.10.2022.

Дата опубликования 22.11.2022.

*Mathematical Biology and Bioinformatics*
*doi: 10.17537/2022.17.325*

========== BIOINFORMATICS ==========

# Linkage Disequilibrium in Targeted Sequencing

**Dmitriy Romanov*[1,2], Nikolai Skoblikow[1,3,4]**

[1]*Medical laboratory CL, Krasnodar, Russia*
[2]*Southern Federal University, Rostov-on-Don, Russia*
[3]*Kuban State Medical University, Krasnodar, Russia*
[4]*Krasnodar Research Center for Animal Husbandry and Veterinary Medicine, Krasnodar, Russia*

***Abstract.*** We propose an approach for optimizing the development of gene diagnostic panels, which is based on the construction of non-equilibrium linkage maps. In the process of gene selection we essentially use genome-wide association analysis (GWAS). Whole-genome analysis of associations makes it possible to reveal the relationship of genomic variants with the studied phenotype. However, the nucleotide variants that showed the highest degree of association can only be statistically associated with the phenotype, not being the true cause of the phenotype. In this case, they may be in the block of linked inheritance with nucleotide variants that really affect the manifestation of the phenotype. The construction of maps of non-equilibrium linkage of nucleotides makes it possible to optimally determine the boundaries of linkage blocks, in which the desired variants fall. The aim of this study was to optimize the demarcation of genomic loci to create targeted panels aimed at predicting susceptibility to SARS-CoV-2 and the severity of COVID-19. The proposed method for selecting loci for a target panel, taking into account nonequilibrium linkage, makes it possible to use the phenomenon of nonequilibrium linkage in order to maximally cover the regions involved in the development of the phenotype, while simultaneously minimizing the length of these regions, and, at the same time, the cost of sequencing.

***Key words:*** *COVID-19, linkage disequilibrium, targeted sequencing.*

*rdme@yandex.ru