

DOI 10.18699/vjgb-24-70

Импутация генотипов в геномных исследованиях человека

А.А. Бердникова ^{1, 2}, И.В. Зоркольева ¹, Я.А. Цепилов ¹, Е.Е. Елгаева ^{1, 2} 

¹ Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия

² Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

 elizabeth.elgaeva@gmail.com

Аннотация. Импутация – это метод, позволяющий восстанавливать недостающую информацию о генетических вариантах, которые не удалось генотипировать напрямую с помощью ДНК-микрочипов или секвенирования с низким покрытием. Импутация играет важнейшую роль в полногеномном анализе ассоциаций (genome wide associations study, GWAS). Она приводит к существенному увеличению количества изучаемых вариантов, что повышает разрешающую способность метода и увеличивает сопоставимость данных, полученных в разных когортах и/или с помощью разных технологий, что важно при проведении метаанализов. При ее выполнении информацию о генотипах в исследуемой выборке, у которой известна только часть генетических вариантов, дополняют за счет эталонной (референсной) выборки, имеющей более полные данные о генотипах (чаще всего это результаты полногеномного секвенирования). Импутация стала неотъемлемой частью геномных исследований человека благодаря преимуществам, которые она дает, а также увеличению доступности инструментов для импутации и данных референсных выборок. Обзор посвящен импутации в геномных исследованиях человека. В первом разделе приводятся описание технологий получения информации о генотипах человека и характеристика получаемых типов данных. Во втором разделе представлена методология импутации, перечисляются этапы ее проведения и соответствующие программы, дается описание наиболее популярных референсных панелей и способов оценки качества импутации. В заключении представлены примеры использования импутации в геномных исследованиях выборок из России. Настоящий обзор показывает важность проведения импутации, дает информацию о том, как ее выполнять, и систематизирует результаты ее применения на примере российских выборок.

Ключевые слова: импутация; генотипирование; секвенирование; полногеномный анализ ассоциаций; человек; ДНК-микрочип.

Для цитирования: Бердникова А.А., Зоркольева И.В., Цепилов Я.А., Елгаева Е.Е. Импутация генотипов в геномных исследованиях человека. *Вавиловский журнал генетики и селекции*. 2024;28(6):628-639. DOI 10.18699/vjgb-24-70

Финансирование. Работа БАА и ЕЕЕ поддержана грантом Российского научного фонда № 22-15-20037 и Правительством Новосибирской области. Работа ЗИВ и ЦЯА финансирована за счет бюджетного проекта № FWNR-2022–0020.

Благодарности. Авторы выражают благодарность В.С. Фишману за рекомендации по улучшению текста статьи.

Genotype imputation in human genomic studies

A.A. Berdnikova ^{1, 2}, I.V. Zorkoltseva ¹, Y.A. Tsepilov ¹, E.E. Elgaeva ^{1, 2} 

¹ Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

² Novosibirsk State University, Novosibirsk, Russia

 elizabeth.elgaeva@gmail.com

Abstract. Imputation is a method that supplies missing information about genetic variants that could not be directly genotyped with DNA microarrays or low-coverage sequencing. Imputation plays a critical role in genome-wide association studies (GWAS). It leads to a significant increase in the number of studied variants, which improves the resolution of the method and enhances the comparability of data obtained in different cohorts and/or by using different technologies, which is important for conducting meta-analyses. When performing imputation, genotype information from the study sample, in which only part of the genetic variants are known, is complemented using the standard (reference) sample, which has more complete genotype data (most often the results of whole-genome sequencing). Imputation has become an integral part of human genomic research due to the benefits it provides and the increasing availability of imputation tools and reference sample data. This review focuses on imputation in human genomic research. The first section of the review provides a description of technologies for obtaining information about human genotypes and characteristics of these types of data. The second section describes the imputation methodology, lists the stages of its implementation and the corresponding programs, provides a description of the most popular reference panels and methods for assessing the quality of imputation. The review

concludes with examples of the use of imputation in genomic studies of samples from Russia. This review shows the importance of imputation, provides information on how to carry it out, and systematizes the results of its application using Russian samples.

Key words: imputation; genotyping; sequencing; genome-wide association study; human; DNA-microarray.

For citation: Berdnikova A.A., Zorkoltseva I.V., Tsepilov Y.A., Elgaeva E.E. Genotype imputation in human genomic studies. *Vavilovskii Zhurnal Genetiki i Seleksii* = *Vavilov Journal of Genetics and Breeding*. 2024;28(6):628-639. DOI 10.18699/vjgb-24-70

Технологии получения данных о генотипе человека и их особенности

Данные генотипа человека являются ключевым аспектом для многих генетических исследований. Существует несколько технологий, разработанных для прочтения, анализа и интерпретации генетической информации. Наиболее часто используемые методы включают в себя секвенирование по Сэнгеру, секвенирование следующего поколения (next generation sequencing, NGS) и ДНК-микрочипы.

Генотипирование с помощью ДНК-микрочипов

ДНК-микрочип (или просто «микрочип», «чип»), однако не следует путать с РНК-микрочипом, что является другой технологией), представляет собой небольшую стеклянную или кремниевую подложку, к которой в заданном порядке с большой плотностью прикреплены десятки тысяч зондов (коротких фрагментов одноцепочечной ДНК, комплементарных определенным нуклеотидным последовательностям). Эти зонды расположены на чипе так, чтобы каждый фрагмент можно было идентифицировать по его местоположению (рис. 1).

В ходе анализа к исследуемым молекулам ДНК, разрезанным на фрагменты эндонуклеазами рестрикции, присоединяют флуоресцентные маркеры и помещают их на чип. Целевые фрагменты ДНК связываются с комплементарными ДНК-зондами, а все оставшиеся удаляют с чипа. Для детекции флуоресценции фрагментов, записи картины эмиссии (излучения) и последующей идентификации последовательностей используют лазерные лучи и компьютерную обработку. Это очень быстрый метод, позволяющий одновременно определять нуклеотидную последовательность сразу нескольких фрагментов ДНК (Govindarajan et al., 2012).

Альтернативный подход к решению задачи генотипирования был реализован академиком А.Д. Мирзабеков-

вым в отечественных разработках по созданию гелевых микрочипов (Мирзабеков, 2003). Они представляют собой подложку из стекла, пластика или силикона с фиксированными на ее поверхности полусферическими каплями гидрогеля. Отличием этого метода является то, что фрагменты ДНК оказываются иммобилизованными в трехмерном пространстве, что обеспечивает большую чувствительность и емкость микрочипа. Данная технология также нашла свое применение в анализе РНК, белковых и клеточных биочипах.

Существует несколько стратегий идентификации однонуклеотидного полиморфизма (single nucleotide polymorphism, SNP) для микрочипов (рис. 2).

Аллель-специфичная гибридизация (allele discrimination by hybridization) (рис. 2, а). Меченая целевая ДНК гибридизуется с зондами, содержащими полиморфный сайт в центре. Правильно спаренные олигонуклеотиды получают стабильнее (имеют большую температуру плавления), по сравнению с дуплексами с некомплементарным основанием. Поэтому после промывки чипа в жестких температурных условиях на нем остаются лишь верно спаренные цепочки. Принято использовать несколько фрагментов для каждого аллеля, чтобы улучшить качество сигнала по отношению к шуму (Wang D.G. et al., 1998).

GoldenGate-анализ компании Illumina (рис. 2, б) – два аллель-специфических олигонуклеотида, каждый из которых имеет 5'-конец с разными универсальными праймерами, Р1 и Р2 (праймеры метят уникальным флуорофором для последующего различия сайтов), гибридизуются в растворе с геномной ДНК. Третий олигонуклеотид, помимо универсального праймера, Р3, имеет хвостовую часть с последовательностью «штрих-код», комплементарную фрагменту на чипе. Удлиненные полимеразой аллель-специфичные праймеры лигируются с третьим олигонуклеотидом, после чего полученные фрагменты

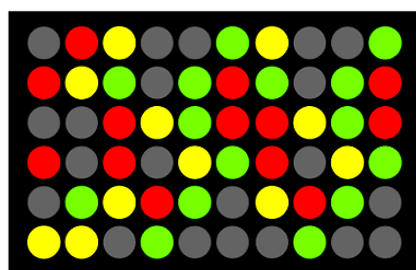
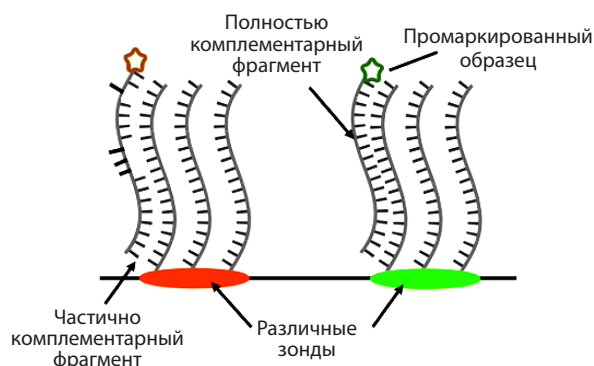


Рис. 1. ДНК-микрочип.

Псевдоцвет (красный, желтый или зеленый) определяется количеством связавшихся с зондом молекул, помеченных разными красителями. Дальнейшие пояснения к рисунку см. в тексте.



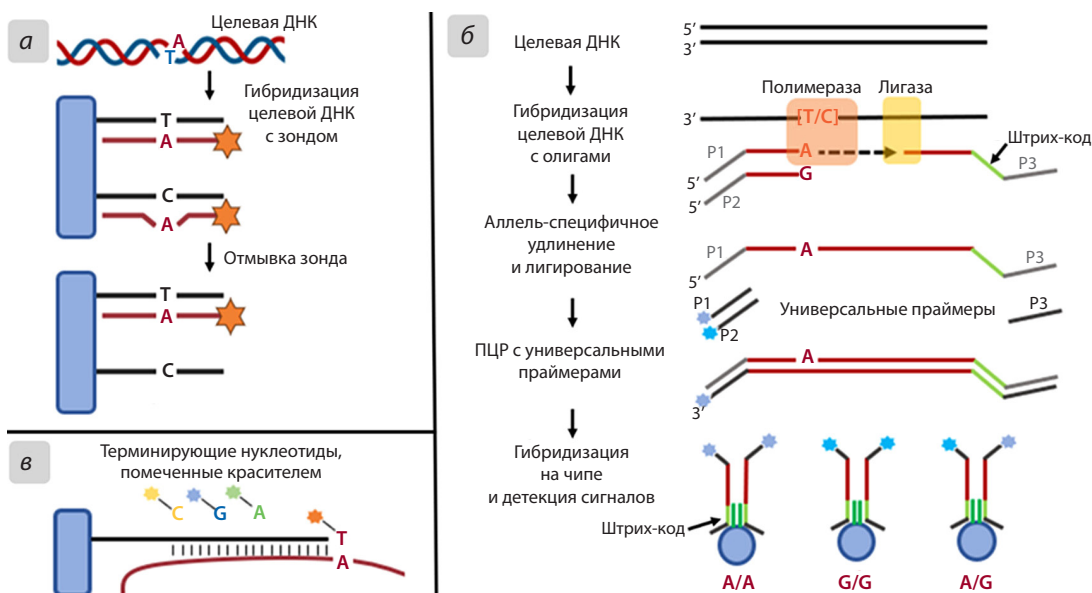


Рис. 2. Стратегии обнаружения SNP для ДНК-микрочипов.

а – аллель-специфическая гибридизация; б – GoldenGate-анализ компании Illumina; в – матричное удлинение праймера.

амплифицируют с помощью полимеразной цепной реакции и гибридизуют на чип. Использование нескольких штрих-кодов (по одному для каждого интересующего локуса) позволяет проводить анализ сразу для нескольких участков генома (Fan et al., 2003).

Матричное удлинение праймера (arrayed primer extension, APEx) (рис. 2, в). Здесь чип содержит фрагмент ДНК, 5'-конец которого закреплен на подложке, а 3'-конец заканчивается нуклеотидом, предшествующим определяемому SNP. Фрагменты геномной ДНК гибридизуются с чипом, при этом искомым SNP остается неспаренным. В ходе реакции секвенирования прикрепленная к подложке последовательность нуклеотидов удлиняется на один терминирующий нуклеотид, помеченный красителем (Kurg et al., 2000). Этот нуклеотид предотвращает дальнейший рост цепи ДНК, а цвет его красителя позволяет определить, какой из нуклеотидов (А, Т, Г или Ц) находится в данной позиции.

Одно из основных преимуществ ДНК-микрочипов – высокая пропускная способность (Nayat, 2002; Brown et al., 2024). Микрочип обеспечивает основу для одновременного генотипирования тысяч различных локусов и выявления однонуклеотидных замен. Так, микрочипы используются для анализа выборок больших объемов с целью генотипировать часто встречаемые генетические варианты (с частотой минорного аллеля в популяции >0.01).

Однако существуют и некоторые ограничения в интерпретации результатов. Данные микрочипов обычно являются бинарными (указывают на наличие или отсутствие определенного аллеля), высокопроизводительными (позволяют проанализировать тысячи и миллионы SNP) и требующими специальных методов анализа для извлечения значимой информации. Речь в данном случае идет о программном обеспечении (например, GenomeStudio (Illumina Inc., San Diego, CA, США)), которое включает в себя инструменты для контроля качества, определения

генотипа, визуализации и анализа данных. Кроме того, микрочипы могут давать как ложноположительные, так и ложноотрицательные результаты. Эти проблемы подчеркивают важность тщательной интерпретации данных и необходимость использования соответствующих статистических методов для контроля качества и подтверждения результатов.

Секвенирование генома

В этом подразделе описаны различные технологии секвенирования. Примерно в 1976 г. было разработано два метода, которые могли декодировать сотни оснований за полдня – обрыв цепи Сэнгера и Коулсона и химическое расщепление Максама и Гилберта (Maxam, Gilbert, 1977; Sanger et al., 1977). В обоих методах исследуемую ДНК распределяют по четырем пробиркам с разным составом реакционной смеси для конкретного типа азотистого основания (А, Т, Г или Ц). В методе Гилберта используются ДНК, радиоактивно меченная с одного конца, и смесь ферментов, специфично ее разрезающих перед нуклеотидом определенного вида. Секвенирование по Сэнгеру, напротив, содержит праймеры и дидезоксинуклеотиды, приводящие к прекращению синтеза цепи при включении радиоактивно-меченого дидезоксинуклеозидтрифосфата (ddNTP; в каждой пробирке свой). Таким образом, в результате применения обоих методов в каждой пробирке образуются меченые фрагменты ДНК различной длины, заканчивающиеся одинаковым основанием. Последовательности разделяют по длине с помощью электрофореза на полиакриламидных пластинчатых гелях (по одной дорожке на каждый тип основания) с разрешением в один нуклеотид. Изображение, полученное на рентгеновской пленке после электрофореза, позволяет восстановить исходную последовательность. Описанные методы сразу же вошли в обиход, а к 1987 г. автоматизированные флуоресцентные секвенаторы Сэнгера могли прочитывать уже

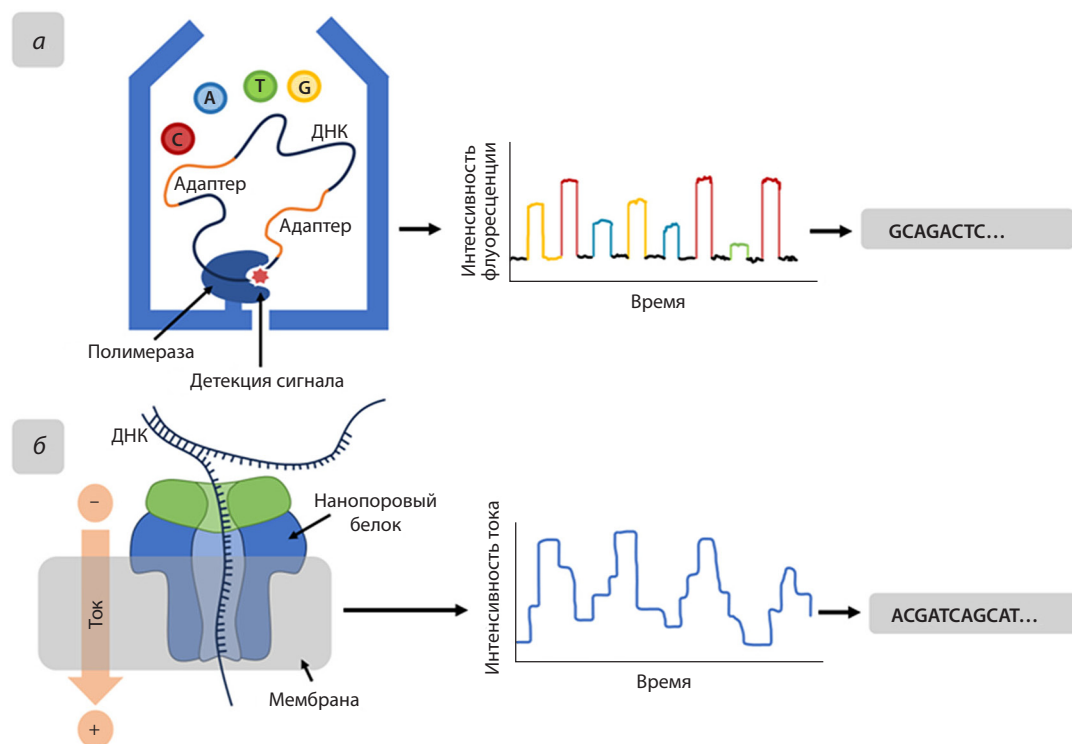


Рис. 3. Секвенирование третьего поколения.

а – Pacific Biosciences; *б* – Oxford Nanopore Technology. Пояснения см. в тексте ниже.

около 1000 оснований в день (Smith et al., 1986; Connell et al., 1987).

В 2005 г. впервые были представлены технологии секвенирования второго поколения (next generation sequencing, NGS), в основе которых лежат два подхода. Первый из них – секвенирование путем гибридизации (sequencing by hybridization, SBH). Суть метода состоит в следующем: сначала короткие участки ДНК фиксируют на стеклянной подложке (ДНК-чипе). Затем фрагменты, подлежащие идентификации, метят флуорофором и наносят на чип для гибридизации с зафиксированными участками. Одноцепочечную ДНК смывают, и картина гибридизации считывается по цветным меткам и их яркости. Альтернативным подходом в NGS является секвенирование путем синтеза (sequencing by synthesis, SBS) (Shendure et al., 2017).

В технологиях, использующих методику SBS, как правило, заранее фрагментированные последовательности закрепляют в проточной ячейке, где происходит циклический синтез новой цепи. Путем последовательного добавления одного из четырех дезоксирибонуклеотидов в смесь, при удалении заранее из нее предыдущих, можно считать сигналы из тех ячеек, где реакция синтеза прошла успешно. Тем самым на выходе получают информацию о том, где и какой нуклеотид находится.

Технологии секвенирования с подходом, отличным от NGS, были впервые описаны в 2008–2009 гг. и получили название «секвенирование третьего поколения» (Check Hayden, 2009). Они включают два основных подхода (рис. 3).

Первая технология, Pacific Biosciences (PacBio) (Rhoads, Au, 2015), заключается в оптическом наблюдении синтеза ДНК с помощью полимеразы в режиме реального време-

ни. В конструкции имеется отверстие меньше половины длины световой волны, ограничивающее флуоресцентное возбуждение небольшим объемом, в котором находятся только полимеразы и их матрица (см. рис. 3, *а*). При таком устройстве только флуоресцентно-меченые нуклеотиды, включенные в растущую цепь ДНК, излучают сигналы достаточной длительности, чтобы быть считанными. Частота ошибок при этом методе секвенирования очень высока (около 10 %), но они распределены случайным образом. Благодаря длинным прочтениям и толерантности к высокому содержанию GC и случайным ошибкам, технология PacBio позволяет получить сборки *de novo* беспрецедентного качества в отношении точности и непрерывности.

Вторая из основных технологий секвенирования третьего поколения – Oxford Nanopore (ONT) (Deamer et al., 2016). Данная методика впервые была предложена в 1980-х гг. Специальная камера, где происходит процесс секвенирования, заполнена электролитическим раствором и поделена двухслойной мембраной с нанопорой (ее размеры находятся в нанодиапазоне). После подключения напряжения ионы электролита и молекула ДНК начинают двигаться через пору. Нуклеиновая кислота физически мешает миграции ионов, что приводит к колебаниям силы тока, по которым можно определять нуклеотидную последовательность (см. рис. 3, *б*). Основным отличием от других технологий секвенирования является исключительная портативность устройств с нанопорами, которые могут быть размером с карту памяти (USB), поскольку они основаны на обнаружении электронных, а не оптических сигналов.

Сравнение технологий и их применения для решения разных задач

Наиболее часто для крупномасштабных проектов (полногеномного секвенирования, анализа транскриптома и эпигенетического профилирования) используют NGS технологии Illumina, однако для проведения сборки *de novo* больше соответствует PacBio, а для портативного секвенирования – ONT. Метод Сэнгера подходит для секвенирования коротких фрагментов ДНК, таких как отдельные гены, плазмиды или вирусные геномы.

Следует также упомянуть конкурирующий с Illumina секвенатор, разработанный компаниями Complete Genomics и MGI Tech, DNBSEQ-T7 (ранее известный как MGISEQ-T7). У DNBSEQ-T7 процесс клональной амплификации происходит по типу катящегося кольца, т. е. всегда с исходной матрицы, что позволяет исключить накопление ошибок ДНК-полимеразы (Drmanac et al., 2010). Основные преимущества MGI включают меньшую в сравнении с Illumina стоимость и возможность обработки большего объема образцов за короткое время. Как показывают последние исследования, новый секвенатор MGISEQ-2000 может быть использован в качестве полноценной альтернативы секвенаторам Illumina при проведении полногеномных исследований (поиск вариантов, выявление инделов); различия между двумя платформами незначительны (Korostin et al., 2020; Jeon et al., 2021; Feng et al., 2024).

Недавно была показана эффективность применения полногеномного секвенирования (whole-genome sequencing, WGS) для проведения GWAS (DePristo et al., 2011; Chat et al., 2022). Этот подход – многообещающая альтернатива для генотипирования с помощью ДНК-микрочипов, поскольку дает возможность получить информацию о большей доле генетических вариаций, повышая мощность тестов ассоциации и последующего анализа тонкого картирования (fine-mapping) (Wang Q.S., Huang, 2022). Однако, несмотря на снижение стоимости технологий на основе NGS, для GWAS используются главным образом высокопроизводительные и относительно дешевые ДНК-микрочипы, содержащие от сотен тысяч до миллионов общих генетических маркеров, которые дают возможность протестировать практически весь геном на наличие ассоциаций с изучаемым признаком.

Генотипирование SNP с помощью ДНК-микрочипов может содержать до 5 % ошибок в зависимости от производителя (Lamy et al., 2006; Yang et al., 2011; Guo et al., 2014). Однако существующие протоколы по контролю качества полученных данных позволяют заметно снизить количество ошибок (в среднем на 1.7 %) (Zhao et al., 2018). Таким образом, микрочипы дают возможность достаточно точно генотипировать образцы даже для видов с высокой гетерозиготностью (т. е. с большей генетической изменчивостью по сравнению с ожидаемой при равновесии Харди–Вайнберга) (Bourke et al., 2018). Более того, на конец 2023 г. цена генотипирования образца на микрочипе была на порядок меньше стоимости секвенирования NGS, что позволяет при одном и том же бюджете проекта охватить больший размер выборки. Основной их недостаток при проведении GWAS – они не дают обнару-

жить ассоциацию между SNP и признаком в том случае, если данный генетический вариант не представлен на микрочипе.

Дополнительные трудности в использовании ДНК-микрочипов могут быть вызваны тем, что информация (например, расположение SNP на хромосоме), использованная для дизайна чипа, устарела или отличается у разных производителей. Вышеперечисленные проблемы могут быть решены импутацией данных генотипирования (Pasanici et al., 2012). Этот подход позволяет увеличить плотность охвата изучаемых генетических вариантов (общее число маркеров) и долю общих вариантов при проведении метаанализа (объединения данных разных исследований и/или платформ генотипирования) (Li Y. et al., 2009).

Заменой ДНК-микрочипов может стать WGS с низким покрытием (low-coverage, lcWGS), в котором секвенируются случайные области генома (Chat et al., 2022). Исследования показывают, что lcWGS значительно превосходит микрочипы по плотности распределения маркеров, что также дает возможность более тщательно оценивать ассоциации с менее распространенными вариантами. Такие данные также требуют импутации с использованием гаплотипов (например, из проекта «1000 геномов») (Auton et al., 2015). Затраты WGS со сверхнизким покрытием (глубина секвенирования $\leq 0.5\times$) могут быть сравнимы или ниже, чем при использовании ДНК-микрочипов, однако его потенциал в качестве альтернативы еще тщательно не оценен (Martin et al., 2021).

Секвенирование ДНК и генотипирование решают задачу анализа генетической информации по-разному. Так, секвенирование позволяет читать фрагменты ДНК целиком и потому применяется для выявления редких (частота минорного аллеля $< 0.01\%$) и *de novo* мутаций, широко используется для изучения структуры отдельных генов или участков генома. Генотипирование, с другой стороны, – более быстрый и экономичный метод анализа генетической изменчивости, что особенно полезно для крупномасштабных геномных исследований, включающих тысячи или даже миллионы образцов. Таким образом, если цель исследования состоит в том, чтобы всесторонне исследовать генетическую архитектуру признака или заболевания, секвенирование, вероятно, является лучшим подходом. Однако если фокус исследования сосредоточен на частых генетических вариантах или анализе популяционной или родственной структуры выборки, то генотипирование часто бывает достаточным и более эффективным (Gresham et al., 2008).

Импутация данных генотипирования

Хотя секвенирование всего генома у сотен тысяч людей пока трудноосуществимо, можно добиться значительного прогресса, определяя лишь относительно небольшое количество генетических вариантов у каждого человека. Такой тип «неполной» информации все равно полезен, потому что данные о любом наборе SNP в группе людей дают возможность делать выводы о многих других вариантах, не наблюдаемых у тех же людей. Подход, это осуществляющий, называется импутацией.

Методология

Процедура импутации предполагает следующие этапы: контроль качества данных генотипирования, фазирование, собственно импутация и на завершающем этапе – контроль качества импутированных генотипов (рис. 4).

Генетические варианты, расположенные рядом на хромосоме, имеют большую вероятность наследоваться вместе, что происходит из-за того, что на каждую хромосому приходится всего несколько рекомбинаций. Данный принцип получил название «неравновесие по сцеплению» (linkage disequilibrium, LD). Благодаря этому принципу мы наблюдаем блоки гаплогрупп (гапоблоки) – наборы близко расположенных генетических вариантов, которые в ходе эволюции наследовались вместе.

В импутации гапоблоки используются для выявления общих коротких участков ДНК на хромосомах, которые индивиды в случайно выбранной популяции могли унаследовать от общего предка. Путем сравнения гаплогрупп в двух выборках (исследуемой и референсной) по набору общих генетических вариантов алгоритмы импутации предоставляют вывод о генотипах исследуемых индивидов. Обе эти выборки должны относиться к одной этнической группе, чтобы импутация давала точные результаты (Mills et al., 2020).

Хотя данные генотипирования не содержат информации о гаплогруппах, их можно вывести и реконструировать с помощью поэтапного анализа. Фазирование – это процесс статистической оценки гаплогрупп. Импутация может быть реализована как на необработанных данных нефазированного генотипирования, так и на реконструированных смешанных гаплогруппах, хотя известно, что фазирование приводит к повышению точности импутации (Anderson et al., 2010). Кроме того, фазирование зачастую необходимо ввиду того, что стандартные алгоритмы для импутации (подробнее ниже) работают именно с гаплогруппами.

Контроль качества данных генотипирования

Важным шагом любого геномного исследования является проведение контроля качества данных. Значимость этого этапа иллюстрирует пример со статьей, опубликованной в “Science”, которая была отозвана из-за недостаточного учета технических ошибок при генотипировании на чипе Illumina (Marees et al., 2018).

Контроль качества данных ДНК-микрочипового генотипирования подразделяется на два основных шага: конт-

роль на уровне индивидов и контроль на уровне маркеров. Контроль на уровне индивидов заключается в удалении из выборки образца в следующих случаях (Anderson et al., 2010):

- наблюдается несоответствие между фенотипом и генотипом (в частности, фенотипический пол отличается от генетического);
- количество гетерозиготных локусов в геноме отклоняется от ожидаемого значения (завышение или занижение этого показателя может свидетельствовать о контаминации образца или об инбридинге соответственно);
- в выборке присутствуют дубликаты, родственники первой или второй степени родства (схожие генотипы будут представлены избыточно, вследствие чего частоты аллелей в популяции могут быть отображены недостоверно);
- имеет другое этническое происхождение, т.е. наблюдается стратификация популяции (наиболее распространенный подход для выявления таких индивидов – анализ главных компонент (principal component analysis, PCA) на матрице родства).

Контроль качества данных на уровне отдельных маркеров также состоит из нескольких пунктов, которые предполагают удаление SNP в случае, если:

- частота минорного аллеля (minor allele frequency, MAF) < 0.01 ;
- они отсутствуют у большей части индивидов в выборке;
- значительно отклоняются от равновесия Харди–Вайнберга.

Для проведения контроля качества используют ряд готовых программ, находящихся в открытом доступе: PLINK 1.9/PLINK 2 (Purcell et al., 2007; Chang et al., 2015), RCOLI (Lam et al., 2020), SMARTPCA (Price et al., 2006) и FlashPCA (Abraham et al., 2017).

Инструменты для импутации

За последние 20 лет несколько различных исследовательских групп разработали и опубликовали ряд инструментов для фазирования и последующей импутации, большинство из которых основано на скрытой марковской модели (Hidden Markov Model, HMM) Ли и Стивенса (LS) (Li N., Stephens, 2003). Эта статистическая модель, впервые описанная в 2003 г., предполагает, что гаплогруппы наследуются в виде гаплогрупп и события рекомбинации происходят на их границах. Модель вероятностно реконструирует ис-



Рис. 4. Импутация данных генотипирования.

1 – фазирование; 2 – собственно импутация. Пояснения см. в тексте.



Рис. 5. Визуализация работы алгоритмов, основанных на HMM, для четырех индивидов из референсной выборки.

Каждая колонка – отдельный SNP с двумя аллелями (пустые и зачеркнутые квадраты обозначают разные аллели одного SNP), а каждая пара строк представляет собой две копии ДНК (от каждого из родителей). Тесно связанные SNP объединены в группы по цветам; каждый гаплотип моделируется как мозаика цветных комбинаций (Scheet, Stephens, 2006).

следующие гаплотипы в виде мозаики, составленной из гаплотипов небольшой референсной выборки (рис. 5). Было показано, что методы, основанные на HMM Ли и Стивенса, точнее и эффективнее (Weale, 2004) таких подходов, как, например, алгоритм Кларка (Clark, 1990) или ЕМ-алгоритм (expectation-maximization – максимизация ожидания) (Dempster et al., 1977) (Browning S.R., Browning B.L., 2011). В настоящее время наиболее часто применяемыми программами, использующими HMM Ли и Стивенса, являются Beagle 5 (Browning et al., 2021), Eagle 2 (Loh et al., 2016) и ShapeIT (Delaneau et al., 2012) для фазирования, а также Beagle 5 (Browning et al., 2018), Impute 5 (Rubinacci et al., 2020) и Minimac 4 (Das et al., 2016) – для импутации. Beagle 5 и ShapeIT 2 при этом позволяют проводить обе эти процедуры.

Сравнительный анализ текущего программного обеспечения (ПО) для фазирования и импутации показал, что в целом Beagle5.4 несколько лучше, чем Impute5 и Minimac4. Эта программа отличается более высоким коэффициентом конкордации и высокой производительностью даже на больших наборах данных (De Marino et al., 2022). Однако Minimac4 и Impute5, как правило, лучше работают с редкими вариантами, потому что, в отличие от Beagle5.4, которая вычисляет кластеры гаплотипов и выполняет вычисления на их основе, эти программы выполняют поиск во всем пространстве гаплотипов. Программа Minimac4 требует наименьшего объема памяти, но вычисления занимают больше времени. Если использование памяти ограничено, а потеря точности приемлема, тогда Minimac4 может быть оптимальным выбором программного обеспечения для импутации.

Вышеперечисленные программы предполагают запуск с локального сервера и требуют наличия референсных гаплотипов. Но большинства таких крупномасштабных наборов данных нет в открытом доступе. По этой причине наиболее часто для импутации данных человека используются специальные серверы, содержащие информацию о разных референсных панелях, такие как Michigan Imputation Server¹ (Das et al., 2016) и TOPMed Imputation Server²

(Das et al., 2016). Исследователи могут загружать туда свои наборы данных, настраивать параметры через пользовательский веб-интерфейс (выбирать инструменты, референсные панели и т.д.), выполнять фазирование и импутацию генотипов на сервере и загружать выходные файлы.

В качестве недостатков данного подхода нужно отметить необходимость отправки своих данных за пределы локального сервера (хотя и с использованием безопасных протоколов соединения) и возможные очереди. Кроме того, пользователи часто ограничены выбором программ или референсных панелей, не могут объединять несколько панелей или интегрировать свои. Однако есть возможность обойти данные ограничения, например с помощью ПО Docker (Das et al., 2016), и запускать импутацию на своем сервере. Проблема автономного запуска заключается в чуть большей сложности из-за настроек вручную, где пользователю нужно устанавливать дополнительные программы для конвейера и учитывать конфликты библиотек. В Приложении 1³ можно сравнить инструментарий, доступный на двух описанных выше серверах.

Референсные панели для импутации данных генотипирования человека

Один из важных вопросов при импутации генотипов заключается в том, как выбрать референсную панель, обеспечивающую высокую точность импутации в интересующей популяции. Так, было показано (Huang, Tseng, 2014), что на качество импутации влияет не только размер панели, но и этнический состав референсной выборки. Наиболее часто используемыми панелями для европейских популяций на сегодняшний день являются: 1000 Genomes (Sudmant et al., 2015), Haplotype Reference Consortium (HRC) (Haplotype Reference Consortium, 2016) и Trans-Omics for Precision Medicine (TOPMed) (Taliun et al., 2021).

Референсная панель 1000 Genomes Phase 3 Version 5 была подготовлена в рамках проекта «1000 геномов» в 2008 г. (Auton et al., 2015). Всего в этом проекте путем использования комбинации полногеномного секвенирования с низким покрытием, экзомного секвенирования с высоким покрытием и генотипирования на микрочипах удалось охарактеризовать 88 млн генетических вариантов (84.7 млн SNP, 3.6 млн коротких вставок/делеций и 60000 структурных вариантов). Данная версия референсной панели включает в себя 49 млн маркеров от 2504 индивидов из смешанной популяции.

Референсная панель HRC r1.1 2016 набрана консорциумом HRC (The Haplotype Reference Consortium) с целью создания крупной справочной панели гаплотипов. Панель HRC объединяет наборы данных из 20 различных исследований, большинство из которых получены с помощью полногеномного секвенирования с низким покрытием (4–8x) и состоят из выборок индивидов преимущественно европейского происхождения. Референсная панель состоит из 64976 гаплотипов 32 тыс. индивидов с 39235157 SNP, делеций и вставок не содержит.

Проект TOPMed (The Trans-Omics for Precision Medicine) был инициирован в 2010 г. с целью сбора и анализа

¹ <https://imputationserver.sph.umich.edu/index.html#!pages/home>

² <https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/home>

³ Приложения 1–3 см. по адресу:

<https://vavilovj-icg.ru/download/pict-2024-28/appx22.pdf>

данных полногеномного секвенирования. По состоянию на сентябрь 2021 г. TOPMed насчитывает около 180 тыс. участников, преимущественно неевропейского происхождения из более 85 различных исследований. На базе данных TOPMed была создана референсная панель, которая включает 286068980 SNP; 5815513 вставок и 16222592 делеций в генотипах 97256 индивидов. Эти генетические варианты распределены по 22 аутосомам и X-хромосоме. TOPMed (Version r3) – первая панель, которая основана исключительно на данных глубокого полногеномного секвенирования и при этом значительно превосходит ранее опубликованные альтернативы.

Несмотря на то что большинство генетических исследований и референсных панелей ориентированы на выборки индивидов европейского происхождения, нужно отметить существование различных проектов, направленных на изучение генетического разнообразия других популяций. В их число входят ChinaMAP (10588 образцов и 136.7 млн SNP) (Li L. et al., 2021), NARD (1779 индивидов, 40.6 млн SNP) (Yoo et al., 2019), GAsP (1739 образцов, 1 млн аутосомных SNP) (Wall et al., 2019), SG10K (4810 образцов, 89.1 млн SNP) (Wu et al., 2019) для выборки людей азиатского происхождения; AFAM (2269 образцов, 45 млн SNP) (O’Connell et al., 2021) и UGR (4778 образцов, 2.2 млн маркеров) (Fatumo et al., 2022) для афроамериканцев. Панель TOPMed также может быть использована для импутации неевропейских выборок индивидов как африканского, так и азиатского происхождения.

Идеальным решением при выборе панели для импутации стало объединение данных от нескольких референсных выборок для построения комбинированной эталонной панели. Однако в различных исследованиях, как правило, используются разные стратегии контроля качества и фильтрации вариантов, что может затруднить объединение результатов.

Еще одна важная проблема – ограничения на совместное использование данных. Например, информация о генотипе на индивидуальном уровне во многих референсных панелях не является общедоступной; поэтому может быть невозможно напрямую объединить ее с результатами секвенирования других выборок. В связи с этим был предложен метод метаимпутации (Yu et al., 2022). Вместо того чтобы объединять эталонные панели, сначала импугируют генотипы с использованием нескольких эталонных панелей по отдельности, а затем объединяют импугированные результаты в согласованный набор данных.

Оценка качества импутации

Качество импутации данных генотипирования можно оценить: 1) с помощью стандартных метрик качества импутации; 2) опытным путем (например, провести GWAS на исследуемом признаке и проверить воспроизводимость известных из литературы сигналов ассоциации или вычислить полигенную оценку признака и сравнить ее с реальными фенотипами).

Метрики качества импутации также можно разделить на две большие группы (Stahl et al., 2021): 1) те, что оценивают качество импутации без использования генотипиро-

ванных напрямую SNP и вычисляются автоматически при запуске соответствующего ПО для импутации, и 2) те, что позволяют сравнить импугированные SNP с генотипами и вычисляются вручную.

Метрики качества из первой группы специфичны для каждой отдельной программы. Так, например, для программ Minimac4 и Beagle 5 оценивается показатель R^2 (Marchini, Howie, 2010), который при этом считается поразному для каждой из программ, в то время как Impute 5 вычисляет параметр Info (Marchini, Howie, 2010). Из-за своей специфичности они не подходят для сравнения качества данных, импугированных разными методами. С этой задачей успешно справляются метрики из второй группы, к которым относятся: коэффициент конкордации (concordance rate, CR), показатель качества импутации IQS (imputation quality score) (Lin et al., 2010), показатель Хеллингера (Hellinger score) (Roshyara et al., 2014), квадрат евклидовой нормы (SEN, squared Euclidean norm score) (Roshyara et al., 2014) и другие. На практике чаще всего используют стандартные метрики первой группы.

При импутации оцениваются апостериорные вероятности генотипа. Так, для биаллельных SNP в аддитивной модели (где генотип закодирован как 0, 1 и 2, а референсный и альтернативный аллель – 0 и 1 соответственно) оцененная вероятность индивиду i иметь генотип j в конкретном локусе обозначается G_j^i ($j = 0, 1, 2$). Этот показатель рассчитывается соответствующим ПО для импутации по данным референсной и целевой выборок с помощью встроенных алгоритмов (например, скрытой марковской модели, как описано выше). Доза альтернативного аллеля рассчитывается как $D_i = G_1^i + 2G_2^i$.

Метрика R^2 представляет собой аппроксимацию квадрата корреляции между дозой импугированного аллеля и ожидаемым генотипом и считается как отношение дисперсии дозы аллеля и ожидаемой дисперсии при равновесии Харди–Вайнберга

$$\hat{R}_d^2 = \frac{\frac{1}{N} \sum_{i=1}^N (D_i - 2\hat{p})^2}{2\hat{p}(1-\hat{p})}, \quad (1)$$

$$\hat{p} = \sum_{i=1}^N \frac{D_i}{2N},$$

где N – количество индивидов в выборке; D_i – доза импугированного аллеля для i -го индивида; $\hat{p}$ – оценка частоты аллеля.

Многие современные алгоритмы (как, например, Minimac), проводят импутацию на предварительно фазированных генотипах, т. е. работают с гаплотипами. Формула претерпевает небольшие изменения, так как теперь множество генотипов описывается как пул из $2N$ бинарно закодированных аллелей

$$\hat{R}_h^2 = \frac{\frac{1}{2N} \sum_{i=1}^{2N} (H_i - \hat{p})^2}{\hat{p}(1-\hat{p})}, \quad (2)$$

$$\hat{p} = \sum_{i=1}^{2N} \frac{H_i}{2N},$$

где H_i – вероятность импугированного аллеля в i -м гаплотипе (варьируется от 0 до 1 и оценивается встроенными

алгоритмами скрытых марковских моделей); N – размер выборки; $\hat{p}$ – оценка частоты аллеля (вывод формул представлен в Приложениях 2 и 3).

При расчете метрик второго типа часть информации о генотипах в исследуемой выборке искусственно «маскируется» (удаляется из общего набора данных с сохранением информации об этих SNP). После чего полученные пропуски импутируют и сравнивают с реальными генотипами. Например, CR представляет собой долю правильно рассчитанных SNP ко всем SNP. Показатель Хеллингера является мерой расстояния между двумя распределениями вероятностей генотипа, в его основе лежит коэффициент Бхаттачарии (Bhattacharyya, 1946), который измеряет степень перекрытия между двумя распределениями. Показатель SEN представляет собой масштабированное евклидово расстояние между распределением истинной дозы и импутированной. Как показатель Хеллингера, так и показатель SEN рассчитываются для отдельных SNP каждого индивида. Метрика IQS основана на каппа-статистике Козна и позволяет учитывать случайные совпадения импутированных и реальных SNP.

Как было сказано выше, помимо перечисленных метрик, для контроля качества импутации можно использовать полигенную оценку признака (polygenic score, PGS) (Choi et al., 2020). Она представляет собой показатель индивидуального генетического риска в отношении некоторого признака, полученный путем суммирования количественной оценки эффекта многих распространенных вариантов (обычно с частотой минорных аллелей $\geq 1\%$) в геноме, каждый из которых может делать небольшой вклад в генетический риск человека для данного признака или заболевания. Показатель PGS обычно вычисляется как взвешенная сумма набора генетических вариантов, как правило, SNP, определяемых как вариации одной пары оснований из референсного генома. Полученная оценка имеет распределение, близкое к нормальному в общей популяции, причем высокие оценки указывают на более высокий риск.

В общем виде уравнение для расчета взвешенной оценки полигенного риска для отдельного индивида выглядит следующим образом (Collister et al., 2022):

$$PGS_i = \sum_j M_j \hat{\beta}_j \text{dosage}_{ij},$$

где M – количество SNP в модели; $\hat{\beta}_j$ – оценка размера эффекта j -го варианта; dosage_{ij} – генотип в кодировке 0, 1, 2 для j -го варианта в генотипе i -го индивида. Величину эффекта SNP (β) часто получают из результатов GWAS.

После расчета PGS оценки признака ее значения сравнивают со значениями реальных фенотипов. В случае если наблюдается достоверная корреляция между этими двумя наборами данных, мы можем сделать вывод о высоком качестве данных после импутации.

Примеры импутации в геномных исследованиях на российских выборках

Несмотря на описанные выше преимущества импутации и фазирования, существует крайне мало упоминаний их использования в исследованиях российских выборок. Так, в работе (Pinakhina et al., 2022), посвященной изучению депрессии в выборке из 4520 индивидов из различных

регионов России, проведена импутация с помощью референсных панелей HRC и 1000G с использованием программы Beagle 5.1. Аналогично в исследовании генетической структуры западнорусской популяции (выборка из 4145 индивидов) в качестве панели для импутации была выбрана панель HRC, сама процедура проводилась с помощью Beagle 4.0 и позволила рассматривать в анализе, помимо 623 249 генотипированных вариантов, еще 10454514 импутированных (Usoltsev et al., 2023). А в работе (Moreland et al., 2022) по изучению маркеров, ассоциированных с мышечной силой и мощностью у 292 россиян (из них 83 профессиональные атлеты), проведены не только импутация на панели 1000G, но и фазирование с помощью SHAPEIT.

Один из важнейших факторов для осуществления качественной импутации – правильный выбор референсной панели. N. Kolosov с коллегами (2022) оценивали надежность импутации генотипов выборки из 230 пожилых людей из Санкт-Петербурга (501 100 SNP) такими панелями как HRC, 1000G, HGDP (Human Genome Diversity Project (Cann et al., 2002) – референсная панель на основе 929 человек различной этнической принадлежности). Им удалось увеличить с помощью Beagle 5.1 (данные были заранее фазированы) общее количество исследуемых вариантов до 37.6, 37.5 и 26.6 млн SNP для каждой из панелей соответственно. Помимо этого, HRC, по сравнению с двумя другими панелями, показала наибольшую точность импутации (метрики IQS и CR).

Все эти работы в качестве референсных панелей используют HRC или 1000G, однако данный подход несколько устарел и подлежит пересмотру в связи с появлением более крупного набора данных TOPMed, применение которой служит своего рода золотым стандартом в зарубежных исследованиях на сегодняшний день. Из программного обеспечения в рассмотренных работах применяют различные версии Beagle.

В упомянутых исследованиях на российских выборках не проводились метаанализы или тонкое картирование генов, однако, как показывают примеры других работ (Barton et al., 2021), именно благодаря импутации и фазированию такие анализы могут быть выполнены существенно более качественно.

Заключение

В настоящее время проведение импутации данных генотипирования – неотъемлемая часть многих геномных исследований человека, в частности GWAS. Она обеспечивает увеличение числа анализируемых SNP и дает возможность объединять результаты разных исследований. Импутация также существенно улучшает результаты тонкого картирования, позволяя наиболее точно определять конкретные генетические варианты и гены, обуславливающие ассоциацию всего участка генома с изучаемым признаком (Chundru et al., 2019).

Следует отметить, что для крупномасштабных исследований, где важны размер выборки и покрытие генотипирования, комбинация ДНК-микрочипов/секвенирования с низким покрытием и дальнейшей импутацией является наиболее оптимальной и дешевой стратегией получения данных, подходящих для большинства дизайнов ге-

номных исследований. Эта комбинация используется во всех крупных национальных биобанках, как, например, UK Biobank (Sudlow et al., 2015), AlloFUs (Ramirez et al., 2022) и другие.

Наряду с перечисленными достоинствами метод импутации имеет ряд недостатков и ограничений. В частности, ошибки чтения при низком покрытии, а также некорректный выбор параметров для импутации наравне с неподходящей референсной панелью часто приводят к низкой точности импутированных данных, что может отрицательно влиять на результаты дальнейших этапов анализа. Необходимо также помнить, что импутация использует информацию о гаплотипах из референсной выборки, поэтому при ее устаревании генетические варианты, которые стали частыми в популяции сравнительно недавно, могут импутироваться хуже (Ali et al., 2022). Помимо этого, высокий уровень рекомбинации снижает точность фазирования и последующей импутации генотипов, в связи с чем в некоторых случаях необходим дополнительный анализ рекомбинаций (Weng et al., 2014).

Кроме того, импутация может сгладить генетические различия между индивидами в выборках типа «случай–контроль» (Lau et al., 2024): в импутированных данных могут возникнуть неточные генотипы в регионах, где ожидаются различия между случаем и контролем. При этом такой эффект проявляется независимо от того, насколько велика и разнообразна референсная панель. Наконец, при использовании метода важно помнить, что то, что верно для популяции в целом, не всегда может быть верно для конкретного индивида.

Существует большое разнообразие программ и референсных панелей для проведения импутации геномных данных человека, и, как следствие, множество их комбинаций. Благодаря этому у исследователей есть возможность подобрать оптимальный набор инструментов для импутации, наиболее подходящий под особенности выборки и задачи конкретного исследования. Обзор работ на российских выборках показал, что наиболее популярное ПО для импутации – Beagle различных версий, а среди референсных панелей чаще всего используют HRC и 1000G, что несколько отличается от зарубежной практики, где лидером среди референсных панелей является TOPMed.

Большая осведомленность о тонкостях проведения импутации и осознанный подход к подбору инструментов позволят улучшить качество геномных данных без увеличения стоимости их получения, облегчат их интеграцию с результатами других исследований и дадут более точную информацию о генетическом контроле признаков человека.

Список литературы / References

- Мирзабеков А.Д. Биочипы в биологии и медицине XXI века. *Вестн. Рос. акад. наук.* 2003;73(5):412.
[Mirzabekov A.D. Biochips in the biology and medicine of the XXI century. *Vestnik Rossiyskoj Akademii Nauk = Herald of the Russian Academy of Sciences.* 2003;73(5):412 (in Russian)]
Abraham G., Qiu Y., Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics.* 2017; 33(17):2776-2778. DOI 10.1093/bioinformatics/btx299

- Ali A.T., Liebert A., Lau W., Maniatis N., Swallow D.M. The hazards of genotype imputation in chromosomal regions under selection: A case study using the lactase gene region. *Ann. Hum. Genet.* 2022; 86(1):24-33. DOI 10.1111/ahg.12444
Anderson C.A., Pettersson F.H., Clarke G.M., Cardon L.R., Morris A.P., Zondervan K.T. Data quality control in genetic case-control association studies. *Nat. Protoc.* 2010;5(9):1564-1573. DOI 10.1038/nprot.2010.116
Auton A., Abecasis G.R., Altshuler D.M., Durbin R.M., Abecasis G.R., Bentley D.R., ... Min Kang H., Korbel J.O., Marchini J.L., McCarthy S., McVean G.A., Abecasis G.R. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74. DOI 10.1038/nature15393
Barton A.R., Sherman M.A., Mukamel R.E., Loh P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* 2021;53(8):1260-1269. DOI 10.1038/s41588-021-00892-1
Bhattacharyya A. On a measure of divergence between two multinomial populations. *Sankhyā: Ind. J. Stat.* 1946;7(4):401-406
Bourke P.M., Voorrips R.E., Visser R.G.F., Maliepaard C. Tools for genetic studies in experimental populations of polyploids. *Front. Plant. Sci.* 2018;9:513. DOI 10.3389/fpls.2018.00513
Brown A., Ampratwum P.O., Ray S.D. Microarray analysis. In: Encyclopedia of Toxicology. 4 ed. 2024;6:385-392. DOI 10.1016/B978-0-12-824315-2.00210-4
Browning B.L., Zhou Y., Browning S.R. A One-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 2018;103(3):338-348. DOI 10.1016/j.ajhg.2018.07.015
Browning B.L., Tian X., Zhou Y., Browning S.R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* 2021;108(10):1880-1890. DOI 10.1016/j.ajhg.2021.08.005
Browning S.R., Browning B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 2011;12(10):703-714. DOI 10.1038/nrg3054
Cann H.M., de Toma C., Cazes L., Legrand M.F., Morel V., Piouffre L., Bodmer J., ... Zhu S., Weber J.L., Greely H.T., Feldman M.W., Thomas G., Dausset J., Cavalli-Sforza L.L. A human genome diversity cell line panel. *Science.* 2002;296(5566):261-262. DOI 10.1126/science.296.5566.261b
Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell S.M., Lee J.J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience.* 2015;4(1):7. DOI 10.1186/s13742-015-0047-8
Chat V., Ferguson R., Morales L., Kirchhoff T. Ultra low-coverage whole-genome sequencing as an alternative to genotyping arrays in genome-wide association studies. *Front. Genet.* 2022;12:790445. DOI 10.3389/fgene.2021.790445
Check Hayden E. Genome sequencing: the third generation. *Nature.* 2009;457(7231):768-769. DOI 10.1038/news.2009.86
Choi S.W., Mak T.S.-H., O'Reilly P.F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* 2020;15(9):2759-2772. DOI 10.1038/s41596-020-0353-1
Chundru V.K., Marioni R.E., Prendergast J.G.D., Vallerger C.L., Lin T., Beveridge A.J., Gratten J., Hume D.A., Deary I.J., Wray N.R., Visscher P.M., McRae A.F. Examining the impact of imputation errors on fine-mapping using DNA methylation QTL as a model trait. *Genetics.* 2019;212(3):577-586. DOI 10.1534/genetics.118.301861
Clark A.G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* 1990;7(2):111-122. DOI 10.1093/oxfordjournals.molbev.a040591
Collister J.A., Liu X., Clifton L. Calculating polygenic risk scores (PRS) in UK biobank: A practical guide for epidemiologists. *Front. Genet.* 2022;13:818574. DOI 10.3389/fgene.2022.818574
Connell C., Fung S., Heiner C., Bridgman J., Chakerian V., Heron E., Jones B., Menchen S., Mordan W., Raff M., Recknor M., Smith L.M., Springer J., Woo S., Hunkapiller M. Automated DNA-sequence analysis. *Biotechniques.* 1987;5:342-348

- Das S., Forer L., Schönherr S., Sidore C., Locke A.E., Kwong A., Vrieze S.I., Chew E.Y., Levy S., McGue M., Schlessinger D., Stambolian D., Loh P.-R., Iacono W.G., Swaroop A., Scott L.J., Cucca F., Kronenberg F., Boehnke M., Abecasis G.R., Fuchsberger C. Next-generation genotype imputation service and methods. *Nat. Genet.* 2016;48(10):1284-1287. DOI 10.1038/ng.3656
- De Marino A., Mahmoud A.A., Bose M., Bircan K.O., Terpolovsky A., Bamunusinghe V., Bohn S., Khan U., Novković B., Yazdi P.G. A comparative analysis of current phasing and imputation software. *PLoS One.* 2022;17(10):e0260177. DOI 10.1371/journal.pone.0260177
- Deamer D., Akeson M., Branton D. Three decades of nanopore sequencing. *Nat. Biotechnol.* 2016;34(5):518-524. DOI 10.1038/nbt.3423
- Delaneau O., Marchini J., Zagury J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods.* 2012;9(2):179-181. DOI 10.1038/nmeth.1785
- Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Society.* 1977;39(1):1-38. DOI 10.1111/j.2517-6161.1977.tb01600.x
- DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernysky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D., Daly M.J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 2011;43(5):491-498. DOI 10.1038/ng.806
- Drmanac R., Sparks A.B., Callow M.J., Halpern A.L., Burns N.L., Kermani B.G., Carnevali P., ... Drmanac S., Oliphant A.R., Banyai W.C., Martin B., Ballinger D.G., Church G.M., Reid C.A. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science.* 2010;327(5961):78-81. DOI 10.1126/science.1181498
- Fan J.B., Oliphant A., Shen R., Kermani B.G., Garcia F., Gunderon K.L., Hansen M., ... Kruglyak S., Bentley D., Haas J., Rigault P., Zhou L., Stuelcpnagel J., Chee M.S. Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol.* 2003;68:69-78. DOI 10.1101/sqb.2003.68.69
- Fatumo S., Mugisha J., Soremekun O.S., Kalungi A., Mayanja R., Kintu C., Makanga R., Kakande A., Abaasa A., Asiki G., Kalyesubula R., Newton R., Nyirenda M., Sandhu M.S., Kaleebu P. Uganda genome resource: A rich research database for genomic studies of communicable and non-communicable diseases in Africa. *Cell Genom.* 2022;2(11):100209. DOI 10.1016/j.xgen.2022.100209
- Feng Z., Peng F., Xie F., Liu Y., Zhang H., Ma J., Xing J., Guo X. Comparison of capture-based mtDNA sequencing performance between MGI and illumina sequencing platforms in various sample types. *BMC Genomics.* 2024;25(1):41. DOI 10.1186/s12864-023-09938-6
- Govindarajan R., Duraiyan J., Kaliyappan K., Palanisamy M. Microarray and its applications. *J. Pharm. Bioallied Sci.* 2012;4(6):310. DOI 10.4103/0975-7406.100283
- Gresham D., Dunham M.J., Botstein D. Comparing whole genomes using DNA microarrays. *Nat. Rev. Genet.* 2008;9(4):291-302. DOI 10.1038/nrg2335
- Guo Y., He J., Zhao S., Wu H., Zhong X., Sheng Q., Samuels D.C., Shyr Y., Long J. Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* 2014;9(11):2643-2662. DOI 10.1038/nprot.2014.174
- Hayat M.A. DNA microarrays technology. In: Handbook of Immunohistochemistry and *in situ* Hybridization of Human Carcinomas. 2002;49-55. DOI 10.1016/S1874-5784(04)80015-1
- Huang G.-H., Tseng Y.-C. Genotype imputation accuracy with different reference panels in admixed populations. *BMC Proc.* 2014;8(S1):S64. DOI 10.1186/1753-6561-8-S1-S64
- Jeon S.A., Park J.L., Park S.-J., Kim J.H., Goh S.-H., Han J.-Y., Kim S.-Y. Comparison between MGI and illumina sequencing platforms for whole genome sequencing. *Genes Genom.* 2021;43(7):713-724. DOI 10.1007/s13258-021-01096-x
- Kolosov N., Rezapova V., Rotar O., Loboda A., Freylikhman O., Melnik O., Sergushichev A., Stevens C., Voortman T., Kostareva A., Konradi A., Daly M.J., Artomov M. Genotype imputation and polygenic score estimation in northwestern Russian population. *PLoS One.* 2022;17(6):e0269434. DOI 10.1371/journal.pone.0269434
- Korostin D., Kulemin N., Naumov V., Belova V., Kwon D., Gorbachev A. Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *PLoS One.* 2020;15(3):e0230301. DOI 10.1371/journal.pone.0230301
- Kurg A., Töniss N., Georgiou I., Shumaker J., Tollett J., Metspalu A. Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. *Genet. Test.* 2000;4(1):1-7. DOI 10.1089/109065700316408
- Lam M., Awasthi S., Watson H.J., Goldstein J., Panagiotaropoulou G., Trubetskoy V., Karlsson R., Frei O., Fan C.-C., De Witte W., Mota N.R., Mullins N., Brügger K., Lee S.H., Wray N.R., Skarabis N., Huang H., Neale B., Daly M.J., Mattheisen M., Walters R., Ripke S. RICOPII: rapid imputation for CONsortias PIPELine. *Bioinformatics.* 2020;36(3):930-933. DOI 10.1093/bioinformatics/btz633
- Lamy P., Andersen C.L., Wikman F.P., Wiuf C. Genotyping and annotation of Affymetrix SNP arrays. *Nucleic Acids Res.* 2006;34(14):e100. DOI 10.1093/nar/gkl475
- Lau W., Ali A., Maude H., Andrew T., Swallow D.M., Maniatis N. The hazards of genotype imputation when mapping disease susceptibility variants. *Genome Biol.* 2024;25(1):7. DOI 10.1186/s13059-023-03140-3
- Li L., Huang P., Sun X., Wang S., Xu M., Liu S., Feng Z., Zhang Q., Wang X., Zheng X., Dai M., Bi Y., Ning G., Cao Y., Wang W. The ChinaMAP reference panel for the accurate genotype imputation in Chinese populations. *Cell Res.* 2021;31(12):1308-1310. DOI 10.1038/s41422-021-00564-z
- Li N., Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics.* 2003;165(4):2213-2233. DOI 10.1093/genetics/165.4.2213
- Li Y., Willer C., Sanna S., Abecasis G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* 2009;10(1):387-406. DOI 10.1146/annurev.genom.9.081307.164242
- Lin P., Hartz S.M., Zhang Z., Saccone S.F., Wang J., Tischfield J.A., Edenberg H.J., Kramer J.R., Goate A.M., Bierut L.J., Rice J.P. A new statistic to evaluate imputation reliability. *PLoS One.* 2010;5(3):e9697. DOI 10.1371/journal.pone.0009697
- Loh P.-R., Danecek P., Palamara P.F., Fuchsberger C., Reshef Y.A., Finucane H.K., Schoenherr S., Forer L., McCarthy S., Abecasis G.R., Durbin R., L Price A. Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* 2016;48(11):1443-1448. DOI 10.1038/ng.3679
- Marchini J., Howie B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 2010;11(7):499-511. DOI 10.1038/nrg2796
- Marees A.T., de Kluiver H., Stringer S., Vorspan F., Curis E., Marie Claire C., Derks E.M. A tutorial on conducting genome wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* 2018;27(2). DOI 10.1002/mpr.1608
- Martin A.R., Atkinson E.G., Chapman S.B., Stevenson A., Stroud R.E., Abebe T., Akena D., ... Ramesar R., Shiferaw W., Stein D.J., Teffera S., van der Merwe C., Zingela Z. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am. J. Hum. Genet.* 2021;108(4):656-668. DOI 10.1016/j.ajhg.2021.03.012
- Maxam A.M., Gilbert W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA.* 1977;74(2):560-564. DOI 10.1073/pnas.74.2.560
- Mills M.C., Barban N., Tropf F.C. An Introduction to Statistical Genetic Data Analysis. Cambridge, MA: MIT Press, 2020
- Moreland E., Borisov O.V., Semenova E.A., Larin A.K., Andryushchenko O.N., Andryushchenko L.B., Generozov E.V., Wil-

- liams A.G., Ahmetov I.I. Polygenic profile of elite strength athletes. *J. Strength. Cond. Res.* 2022;36(9):2509-2514. DOI 10.1519/JSC.0000000000003901
- O'Connell J., Yun T., Moreno M., Li H., Litterman N., Kolesnikov A., Noblin E., ... Wang W., Weldon C.H., Wilton P., Wong C., Auton A., Carroll A., McLean C.Y. A population-specific reference panel for improved genotype imputation in African Americans. *Commun. Biol.* 2021;4(1):1269. DOI 10.1038/s42003-021-02777-9
- Pasaniuc B., Rohland N., McLaren P.J., Garimella K., Zaitlen N., Li H., Gupta N., ... Haas D.W., Liang L., Sunyaev S., Patterson N., de Bakker P.I.W., Reich D., Price A.L. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* 2012;44(6):631-635. DOI 10.1038/ng.2283
- Pinakhina D., Yermakovich D., Vergasova E., Kasyanov E., Rukavishnikov G., Rezapova V., Kolosov, ... Plotnikov N., Ilinsky V., Neznanov N., Mazo G., Kibitov A., Rakitko A., Artomov M. GWAS of depression in 4,520 individuals from the Russian population highlights the role of MAGI2 (S-SCAM) in the gut-brain axis. *Front. Genet.* 2022;13:972196. DOI 10.3389/fgene.2022.972196
- Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A., Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 2006;38(8):904-909. DOI 10.1038/ng1847
- Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A.R., Bender D., Maller J., Sklar P., de Bakker P.I.W., Daly M.J., Sham P.C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 2007;81(3):559-575. DOI 10.1086/519795
- Ramirez A.H., Sulieman L., Schlueter D.J., Halvorson A., Qian J., Ratsimbazafy F., Loperena R., ... Denny J.C., Carroll R.J., Glazer D., Harris P.A., Hripsak G., Philippakis A., Roden D.M.; All of Us research program. The All of Us research program: Data quality, utility, and diversity. *Patterns (N Y)*. 2022;3(8):100570. DOI 10.1016/j.patter.2022.100570
- Rhoads A., Au K.F. PacBio Sequencing and its applications. *Genomics Proteomics Bioinformatics.* 2015;13(5):278-289. DOI 10.1016/j.gpb.2015.08.002
- Roshyara N.R., Kirsten H., Horn K., Ahnert P., Scholz M. Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genet.* 2014;15(1):88. DOI 10.1186/s12863-014-0088-5
- Rubinacci S., Delaneau O., Marchini J. Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS Genet.* 2020;16(11):e1009049. DOI 10.1371/journal.pgen.1009049
- Sanger F., Nicklen S., Coulson A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA.* 1977;74(12):5463-5467. DOI 10.1073/pnas.74.12.5463
- Scheet P., Stephens M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 2006;78(4):629-644. DOI 10.1086/502802
- Shendure J., Balasubramanian S., Church G.M., Gilbert W., Rogers J., Schloss J.A., Waterston R.H. DNA sequencing at 40: past, present and future. *Nature.* 2017;550(7676):345-353. DOI 10.1038/nature24286
- Smith L.M., Sanders J.Z., Kaiser R.J., Hughes P., Dodd C., Connell C.R., Heiner C., Kent S.B.H., Hood L.E. Fluorescence detection in automated DNA sequence analysis. *Nature.* 1986;321(6071):674-679. DOI 10.1038/321674a0
- Stahl K., Gola D., König I.R. Assessment of imputation quality: comparison of phasing and imputation algorithms in real data. *Front. Genet.* 2021;12:724037. DOI 10.3389/fgene.2021.724037
- Sudlow C., Gallacher J., Allen N., Beral V., Burton P., Danesh J., Downey P., Elliott P., Green J., Landray M., Liu B., Matthews P., Ong G., Pell J., Silman A., Young A., Sprosen T., Peakman T., Collins R. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12(3):e1001779. DOI 10.1371/journal.pmed.1001779
- Sudmant P.H., Rausch T., Gardner E.J., Handsaker R.E., Abyzov A., Huddleston J., Zhang Y., ... Gerstein M.B., Bashir A., Stegle O., Devine S.E., Lee C., Eichler E.E., Korbel J.O. An integrated map of structural variation in 2,504 human genomes. *Nature.* 2015;526(7571):75-81. DOI 10.1038/nature15394
- Taliun D., Harris D.N., Kessler M.D., Carlson J., Szpiech Z.A., Torres R., ... Cupples L.A., Laurie C.C., Jaquish C.E., Hernandez R.D., O'Connor T.D., Abecasis G.R. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature.* 2021;590(7845):290-299. DOI 10.1038/s41586-021-03205-y
- The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 2016;48:1279-1283. DOI 10.1038/ng.3643
- Usoltsev D., Kolosov N., Rotar O., Loboda A., Boyarinova M., Moguchaya E., Kolesova E., ... Laiho P., Kostareva A., Konradi A., Shlyakhto E., Palotie A., Daly M.J., Artomov M. Understanding complex trait susceptibilities and ethnical diversity in a sample of 4,145 Russians through analysis of clinical and genetic data. *bioRxiv.* 2023. DOI 10.1101/2023.03.23.534000
- Wall J.D., Stawiski E.W., Ratan A., Kim H.L., Kim C., Gupta R., Suryamohan K., ... Radha V., Mohan V., Majumder P.P., Seshagiri S., Seo J.-S., Schuster S.C., Peterson A.S. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature.* 2019;576(7785):106-111. DOI 10.1038/s41586-019-1793-z
- Wang D.G., Fan J.-B., Siao C.-J., Berno A., Young P., Sapolsky R., Ghandour G., Perkins N., Winchester E., Spencer J., Kruglyak L., Stein L., Hsie L., Topaloglou T., Hubbell E., Robinson E., Mittmann M., Morris M.S., Shen N., Kilburn D., Rioux J., Nusbaum C., Rozen S., Hudson T.J., Lipshutz R., Chee M., Lander E.S. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science.* 1998;280(5366):1077-1082. DOI 10.1126/science.280.5366.1077
- Wang Q.S., Huang H. Methods for statistical fine-mapping and their applications to auto-immune diseases. *Semin. Immunopathol.* 2022;44(1):101-113. DOI 10.1007/s00281-021-00902-8
- Weale M.E. A survey of current software for haplotype phase inference. *Hum. Genomics.* 2004;1(2):141. DOI 10.1186/1479-7364-1-2-141
- Weng Z.-Q., Saatchi M., Schnabel R.D., Taylor J.F., Garrick D.J. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Gen. Select. Evol.* 2014;46(1):34. DOI 10.1186/1297-9686-46-34
- Wu D., Dou J., Chai X., Bellis C., Wilm A., Shih C.C., ... Wong W.-C., Xie Z., Yeo K.K., Zhang L., Zhai W., Zhao Y. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell.* 2019;179(3):736-749.e15. DOI 10.1016/j.cell.2019.09.019
- Yang H.-C., Lin H.-C., Kang M., Chen C.-H., Lin C.-W., Li L.-H., Wu J.-Y., Chen Y.-T., Pan W.-H. SAQC: SNP array quality control. *BMC Bioinformatics.* 2011;12(1):100. DOI 10.1186/1471-2105-12-100
- Yoo S.-K., Kim C.-U., Kim H.L., Kim S., Shin J.-Y., Kim N., Yang J.S.W., Lo K.-W., Cho B., Matsuda F., Schuster S.C., Kim C., Kim J.-I., Seo J.-S. NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med.* 2019;11(1):64. DOI 10.1186/s13073-019-0677-z
- Yu K., Das S., LeFaive J., Kwong A., Pleinness J., Forer L., Schönherr S., Fuchsberger C., Smith A.V., Abecasis G.R. Meta-imputation: An efficient method to combine genotype data after imputation with multiple reference panels. *Am. J. Hum. Genet.* 2022;109(6):1007-1015. DOI 10.1016/j.ajhg.2022.04.002
- Zhao S., Jing W., Samuels D.C., Sheng Q., Shyr Y., Guo Y. Strategies for processing and quality control of Illumina genotyping arrays. *Brief. Bioinform.* 2018;19(5):765-775. DOI 10.1093/bib/bbx012

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 12.12.2023. После доработки 23.04.2024. Принята к публикации 04.07.2024.