ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК»

на правах рукописи

ЦЕПИЛОВ ЯКОВ АЛЕКСАНДРОВИЧ

РАЗРАБОТКА И ПРИМЕНЕНИЕ НОВЫХ МОДЕЛЕЙ В ПОЛНОГЕНОМНОМ АНАЛИЗЕ АССОЦИАЦИЙ

03.02.07 - Генетика

Диссертация на соискание ученой степени

кандидата биологических наук

Научный руководитель

доктор биологических наук

Аульченко Ю.С.

Новосибирск – 2016

Оглавление

Список сокращений

SNP – Single Nucleotide Polymorphism; однонуклеотидная замена

ПГАА – полногеномный анализ ассоциаций

VIF – Variance Inflation Factor, фактор инфляции дисперсии (коэффициент геномного контроля)

ВГ – взаимодействие генов

ГК – геномный контроль

QTL – Quantative Trait Locus; локус, ассоциированный с количественным признаком

пн – пар нуклеотидных оснований

тпн – тысяча пар нуклеотидных оснований

LD – linkage disequilibrium, неравновесие по сцеплению

ERF – Erasmus Rucphen Family study, семейное исследование «Эразмус-Рукфен»

KORA – Cooperative Health Research in the region of Augsburg, Кооперационное медицинское исследование в регионе Аугсбург

сМ – сантиморган

HWE – Hardy-Weinberg Equilibrium, равновесие Харди-Вайнберга

1 Введение

1.1 Актуальность

Идентификация генов и аллелей, контролирующих разнообразие сложных признаков, является важной теоретической и прикладной задачей генетики и генетической эпидемиологии. Информация об этих генах позволяет получить новые знания о биологических системах, участвующих в формировании таких признаков. Кроме того, знание конкретных аллельных вариантов, контролирующих сложные признаки человека, может найти применение в медицине, например, для определения риска заболевания, или для выявления его молекулярного подтипа. У сельскохозяйственных и домашних животных идентификация аллельных вариантов позволяет вести направленную эффективную селекцию.

Полногеномный анализ ассоциации (ПГАА) является одним из основных методов идентификации аллелей, влияющих на риск возникновения распространенных болезней человека. В рамках этого метода большие популяционные выборки, включающие тысячи особей, используются для исследования ассоциаций между картируемым признаком и большим числом (как минимум несколько сотен тысяч) маркерных локусов, равномерно распределенных по геному. За последнее десятилетие с помощью ПГАА были идентифицированы тысячи локусов, связанных со сложными признаками, что внесло фундаментальный вклад в развитие биологии и генетики [1–3].

Несмотря на огромный прогресс, для большинства сложных признаков человека идентифицированные в рамках ПГАА объясняют только часть наследуемости признака. Например, такой классический количественный признак как рост человека имеет наследуемость порядка 80%, однако суммарный аддитивный вклад всех 180 достоверно ассоциированных

локусов объясняет только 10% дисперсии признака [4]. Феномен «потерянной наследуемости» – невозможность на данном этапе полностью объяснить наследственную компоненту многих признаков и непонимание того, какие механизмы могут отвечать за эту наследственность – свидетельствует о неполноте наших знаний о генетическом контроле наследственных заболеваний и сложных признаков человека.

Одним из аспектов генетического контроля сложных признаков человека, которые до настоящего времени не получили достаточного внимания, являются модели неаддитивного контроля. В большинстве полногеномных исследований ассоциаций используется аддитивная модель наследования признака, в рамках которой предполагается, что вклад каждого аллеля является независимым от вклада других аллелей и прочих факторов. Другие, неаддитивные модели наследования, такие как рецессивная, кодоминантная, доминантная, сверхдоминантная, в контексте ПГАА как правило не рассматриваются. Понятно, что ПГАА с использованием аддитивной модели помогают нам понять основы наследуемости в узком смысле, то есть аддитивной ее компоненты. В тоже время молекулярно-генетические основы наследуемости в широком смысле (т.е. доли фенотипической изменчивости в популяции, обусловленной её генетической изменчивостью) в настоящий момент изучены мало, так как неаддитивные эффекты, как правило, игнорируются в рамках современных ПГАА. Это связано как с недостаточно проработанной методологической базой, так и с практическими трудностями применения новых моделей для анализа реальных данных.

Одной из методологических проблем, затрудняющих проведение полногеномных исследований с применением неаддитивных моделей, является отсутствие для них методов геномного контроля (ГК). Стандартные статистические методы, используемые для ПГАА, такие как линейная регрессия, предполагают, что корреляции между фенотипом и маркером

существуют либо благодаря тому, что аллели маркера оказывают непосредственное влияние на признак (являются функциональными), либо благодаря их неравновесию по сцеплению с функциональными аллелями. Это предположение, как правило, верно, если выборка состоит из представителей одной панмиксной популяции, которые находятся между собой в дальнем родстве. Однако, возможны другие корреляционные взаимосвязи, вызванные сопутствующими факторами, влияющими как на фенотип, так и на генотип различных локусов. При ПГАА генетическая гетерогенность выборки является одним из важнейших сопутствующих факторов. Если анализ не учитывает влияние структуры популяции, тестовая статистика будет завышена [5], что затрудняет статистическую интерпретацию и может привести к ложноположительным результатам (ложное утверждение о наличии «статистически значимой ассоциации» и, как следствие, «идентификации локуса»). Чтобы избежать ложноположительных выводов при интерпретации результатов ПГАА, необходимо проводить их коррекцию, учитывающую генетическую структурированность (генетическую гетерогенность) выборки. Одним из статистических методов, позволяющих проводить коррекцию результатов ПГАА, является ГК, который основывается на использовании информации о несвязанных с признаком маркерах. При нулевой гипотезе об отсутствии ассоциации распределение стандартных тестовых статистик может быть аппроксимировано распределением хи-квадрат с одной степенью свободы. Было показано, что структурированность выборки приводит к увеличению ожидаемого значения статистики на определенную константу, λ, которую называют «коэффициент геномного контроля» или «фактор инфляции» тестовой статистики [5–9]. Если этот коэффициент известен, коррекцию результатов тестирования можно провести, разделив значение каждого полученного теста на эту константу. Было показано, что при предположении об аддитивном вкладе, фактор инфляции λ не зависит от частот аллелей

маркерного локуса. Однако для других моделей наследования (рецессивная, доминантная, сверхдоминантная, кодоминантная) это не так. Для таких моделей фактор инфляции λ является неизвестной функцией от частот аллелей, что затрудняет использование метода геномного контроля, и, как следствие, интерпретацию результатов ПГАА при использовании неаддитивных моделей [6,10].

Однако, проблема ГК не является единственной проблемой, которая затрудняет исследования неаддитивных моделей в рамках ПГАА. Исследователь неаддитивных эффектов столкнется как с проблемой выбора метода для полногеномного скрининга потенциально неаддитивных локусов, так и с последующей проблемой выбора конкретной модели наследования для идентифицированных локусов.

Таким образом, отсутствие проработанной методологической базы и сопутствующего программного обеспечения приводит к тому, что в контексте ПГАА неаддитивные эффекты, как правило, не изучаются, что приводит к неполноте наших знаний о возможных молекулярно-генетических основах наследуемости в широком смысле.

Разработка методов ПГАА с использованием неаддитивных моделей откроет широкие возможности для исследования этого типа генетического контроля сложных признаков человека. Наследуемость в широком смысле, в частности, доминантность, может играть большую роль в контроле некоторых классов функционально-геномных признаков. Ещё в 30-е годы 20-го века были разработаны теории и гипотезы [11–14], которые подчеркивали значимость доминантных эффектов для признаков, зависящих от биохимических механизмов. На основании этих теорий можно ожидать, что доминантные эффекты могут быть особенно распространены при генетическом контроле метаболитов, так как их концентрации напрямую определяются последовательностями биохимических реакций. Однако,

систематического анализа неаддитивных эффектов генов на метаболом человека ранее проведено не было. Поэтому для апробации новых методов ПГАА с учетом неаддитивных эффектов представляется как методологически целесообразным, так и биологически интересным исследовать генетический контроль уровней метаболитов.

1.2 Цели и задачи

Таким образом, принимая во внимание недостаточное методологическое обеспечение и ограниченное число работ, посвященных полногеномным исследованиям неаддитивных эффектов генов, разработка новых полногеномных методов для анализа неаддитивных эффектов является актуальной проблемой современной статистической геномики. Целью данной работы является разработка и апробация методов полногеномного анализа ассоциаций с использованием неаддитивных моделей наследования (рецессивные, кодоминантные, доминантные и сверхдоминантные); применение разработанных методов для анализа генетического контроля уровней метаболитов крови человека. Для достижения цели были поставлены следующие задачи:

1. Получить аналитические выражения для фактора инфляции тестовой статистики для неаддитивных моделей наследования в условиях генетической гетерогенности выборки.
2. На основе полученных аналитических выражений разработать программное обеспечение, реализующее методы геномного контроля неаддитивных моделей.
3. Оценить статистические свойства разработанных методов геномного контроля и протестировать программное обеспечение с использованием модельных и реальных данных.

4. Разработать методику проведения ПГАА с использованием неаддитивных моделей наследования, позволяющую оптимизировать анализ многих признаков.
5. Использовать разработанные методы и программное обеспечение для исследования роли доминантности в контроле сложных признаков человека на примере уровней метаболитов сыворотки крови.

1.3 Научная новизна

Нами были разработаны методы ГК для широкого спектра моделей неаддитивных аллельных взаимодействий (кодоминантной, доминантной, рецессивной, сверхдоминантной). Была предложена и отработана новая методология двухшагового поиска и анализа неаддитивных эффектов. Методология предполагает ПГАА с использованием общей кодоминантной модели для идентификации локусов, потенциально обладающих неаддитивными эффектами. Далее, для исследования модели наследования достоверно идентифицированных локусов, нами предложен набор статистических тестов, которые позволяют установить наиболее парсимонную модель наследования.

Апробация разработанных методов осуществлялась на материале концентраций большой панели метаболитов сыворотки крови человека (22,801 признаков) в крупном популяционном исследовании KORA. В рамках апробации впервые в мире осуществлен неаддитивный ПГАА концентраций метаболитов сыворотки крови человека. Были идентифицированы четыре локуса, обладающих значимыми неаддитивными эффектами. Отклонение от аддитивности для этих локусов ранее не было известно.

1.4 Научно-практическая ценность

Разработанные методы геномного контроля можно использовать для коррекции статистических результатов, полученных для неаддитивных моделей наследования. Эти методы особенно востребованы при наличии остаточной инфляции при мета-анализе результатов ПГАА. Предложенные в диссертации подходы по поиску неаддитивных эффектов могут быть использованы при полногеномном анализе широкого спектра признаков; применение этих подходов будет особенно актуально в исследованиях с более полным геномным покрытием.

1.5 Личный вклад автора

Цели и задачи исследования были сформулированы автором в сотрудничестве с коллегами. Реальные данные для анализа были любезно предоставлены немецкими (KORA) и голландскими (ERF) коллегами в рамках научных коллабораций. Автор разработал методы коррекции статистики, реализовал эти методы в виде программного продукта и провел анализ неаддитивных эффектов на метаболоме человека. Дизайн вычислительных экспериментов, моделирование, анализ данных и интерпретация полученных результатов проведены автором.

1.6 Основные положения, выносимые на защиту

1. Разработанные методы геномного контроля позволяют проводить коррекцию статистических результатов, полученных при ПГАА с применением неаддитивных моделей наследования.

2. Идентификация локусов с неаддитивными эффектами, и определение их генетической модели на данных ПГАА может быть эффективно осуществлена с использованием предложенного нами двухшагового подхода.

3. Генетический контроль уровней метаболитов сыворотки крови человека осуществляется с помощью как аддитивных, так и значимых и реплицируемых неаддитивных внутрилокусных эффектов.

1.7 Публикации по теме диссертации

Материал диссертации представлен в шести работах, из которых две являются публикациями в зарубежных журналах, реферируемых в ISI Web of Science, и четыре являются тезисами конференций.

1.8 Структура и объем диссертации

Объем диссертации составляет 101 страница. Диссертация включает 14 таблиц и 8 иллюстраций, 4 приложения.

2 Обзор литературы

2.1 Генетическая эпидемиология – наука на стыке генетики и клинической эпидемиологии

Эпидемиология в широком смысле – наука, изучающая закономерности возникновения и распространения заболеваний различной этиологии с целью разработки профилактических мероприятий [15]. Предметом изучения эпидемиологии является заболеваемость - совокупность случаев болезни на определённой территории в определённое время среди определённой группы населения. История развития эпидемиологии берет свое начало со времен Гиппократа (460-370 гг. до н.э.) и его работ «Семь книг об эпидемиях», «О воздухе, водах и местностях». Одним из основных вопросов того времени являлся вопрос о причинах возникновения заболеваний. В разные эпохи вплоть до 17 века превалирующей была контагионистская гипотеза, предполагающая, что причиной развития эпидемий является распространение среди людей живого болезнетворного агента. Эту точку зрения впервые высказал древнегреческий философ Аристотель (IV в. до н. э.). В эпоху возрождения контагионистская теория получила множество подтверждений и была окончательно подтверждена работами А. Левенгука (1632–1723 гг.), Л. Пастера (1822–1895 гг.) и Р. Коха (1843–1910 гг.).

Современный вид эпидемиология, как наука, приобрела уже в XIX–XX веках, когда были сформированы ее основные положения. Цель классической эпидемиологии, как уже говорилось, заключается в выявлении закономерностей возникновения, распространения и прекращения болезней человека, борьбы с ними и разработке мер профилактики. Сам термин классической эпидемиологии определить довольно сложно, принято считать, что речь идет о клинической или инфекционной эпидемиологии. Объектом

эпидемиологии инфекционных болезней является эпидемический процесс, закономерности его развития и формы проявления. На данный момент эпидемиологическое учение включает в себя множество областей, отличающихся по специфике изучаемых болезней, специфике факторов риска и распространения, а также - различным методам, применяемым в той или иной подобласти эпидемиологии.

Генетическая эпидемиология выделилась в отдельную область во второй половине прошлого столетия, когда стало возможным измерять и проверять вклад генетических факторов в развитие заболеваний [16]. К тому времени уже было показано, что многие распространённые заболевания человека имеют тенденцию передаваться из поколения в поколение, что было продемонстрировано для таких заболеваний как болезнь Альцгеймера, некоторые формы диабета, рак молочной железы у женщин и т.д. [16].

Современная генетическая эпидемиология базируется одновременно на принципах генетики и эпидемиологии. Однако, специфика генетико-эпидемиологических исследований заключается в том, что основные исследуемые факторы риска – это генетические факторы [15].

В генетической эпидемиологии возможные способы формирования выборки (дизайна исследования) зачастую совпадают с таковыми в классической эпидемиологии. Эпидемиологические исследования отличаются по временному интервалу, в котором проводится исследование, и по способу выбора группы обследуемых. Различают одномоментные (cross-sectional) и многомоментные (longitudinal, или проспективные, prospective) исследования. При одномоментных исследованиях характеристики исследуемых особей определяются только в один момент времени, тогда как в многомоментных исследованиях собирается информация о динамике признака или болезни в определенном временном интервале. По способу выбора можно выделить исследования со случайным выбором (randomly

ascertained), в которых выборка формируется случайным образом относительно исследуемого признака, и исследования, в которых выборка формируется на основе исследуемого признака. По наличию в выборке родственников, различают семейные (family-based) и популяционные (population-based) исследования; промежуточное положение занимают исследования генетически изолированных популяций человека. Особое положение в генетической эпидемиологии занимают семейные исследования близнецов.

Примеры типов исследований, используемые в генетической эпидемиологии, вместе с описанием целей подобных исследований и некоторых подходов, показаны в Таблице 1.

Таблица 1. Типы исследований, которые используются для оценки генетических воздействий на риск возникновения заболевания в популяции людей. Адаптировано из [16].

Тип исследования	Цель	Пример
Описательные исследования	Изучение внешних факторов, например, пола, возраста, расовой принадлежности	Более высокая частота гемофилии у мужчин
Изучение распространения заболевания в семьях	Изучение риска повторного возникновения заболевания в семьях и влияние близкородственных браков	Более высокий риск рака молочной железы у дочерей женщин с раком молочной железы
Исследование специфических генетических факторов	Определение рисков, связанных со специфическими генетическими факторами (аллелями)	Более высокий риск инсулинозависимого сахарного диабета у людей с определенными антигенами главного комплекса гистосовместимости (HLA)

2.2 Методы генетического картирования признаков человека

Одной из основных целей эпидемиологического исследования является оценка эффекта факторов риска на изучаемые болезни (признаки). Факторы

риска могут быть средовыми (курение, профессия), эндогенными (например, избыточная масса тела) и генетическими (генотип определенного локуса). Поиск и определение возможных факторов, влияющих на проявление признака или болезни является важной задачей эпидемиологии. В случае генетической эпидемиологии эта задача включает в себя задачу идентификации генов и регуляторных районов, так или иначе вовлеченных в формирование генетической архитектуры признака - картирование генов.

Существует три основных подхода к картированию генов, отвечающих за развитие болезней: функциональное картирование, тестирование генов-кандидатов и позиционное картирование. При функциональном картировании сначала устанавливается взаимосвязь болезни с дефектом определенного белка, затем идентифицируется ген, кодирующий этот белок, и определяется его локализация в геноме (болезнь - функция - ген - карта). Однако, этот подход оказывается полезным только при картировании менделевских болезней, для которых хорошо изучены биологические причины [17].

Подход, основанный на тестировании генов-кандидатов, требует определенных знаний о биологической природе болезни, позволяющих предположить причастность некоторых генов, так называемых генов-кандидатов, к формированию патологии. Для проверки этих предположений изучается, связан ли полиморфизм признака с полиморфизмом аллелей гена-кандидата. Примером использования этого подхода является анализ наследования диабета второго типа, показавший, что гены, кодирующие инсулин и рецептор инсулина, причастны к детерминации болезни [18].

Позиционное картирование используется в случаях, когда биохимическая природа болезни неизвестна, и не высказывается никаких предположений о генах-кандидатах. Здесь стартовым этапом является анализ совместной сегрегации болезни с маркерными генами, расположение

которых в геноме заранее определено. Результатом такого анализа является локализация предполагаемого гена на генетической карте, что служит основой для дальнейшей идентификации этого гена и выяснения его роли в формировании болезни (болезнь -карта - ген - функция).

В основе методов, осуществляющих такое картирование, лежат хорошо известные биологические явления: сцепление генов, их рекомбинация во время мейоза и полиморфность генома. Благодаря сцеплению, мутация, детерминирующая болезнь, передается потомкам вместе с блоком окружающих ее аллелей соседних локусов. Рекомбинация в ряду поколений уменьшает размер этих блоков. Чем ближе расположены два локуса, тем дольше их аллели сохраняются в одном блоке. Идентификация блоков, полученных от различных родителей, обеспечивается полиморфностью генома, многие локусы которого имеют не один, а несколько вариантов нуклеотидных последовательностей. Такие локусы служат генетическими маркерами. Для того, чтобы картировать ген, вызывающий болезнь, достаточно доказать совместную сегрегацию болезни и блока маркерных аллелей.

Существует два методических подхода, позволяющих выявить те блоки маркерных аллелей, которые сегрегируют вместе с комплексной болезнью: анализ сцепления и анализ ассоциаций [19].

2.2.1 Анализ сцепления

Основной идеей анализа сцепления, или рекомбинационного анализа, является поиск блоков маркеров, которые передаются от больного родителя преимущественно больным потомкам и не передаются здоровым. В разных семьях аллельный состав таких блоков может отличаться, но их позиция в геноме должна быть одинакова. Для анализа сцепления информативными являются только гетерозиготные маркерные локусы. Поэтому предпочтительными для анализа являются мультиаллельные маркеры.

Материалом для анализа сцепления всегда служат выборки с родственной структурой: это могут быть пары родственных больных или расширенные родословные. Анализ сцепления позволяет локализовать ген на участке в 5-50 сМ, т.к. для генотипирования доступны представители не более 2-4 поколений, а размеры семей, как правило, не превышают несколько десятков человек. В таких родословных происходит не так много рекомбинационных событий, и блоки передаваемых генов велики.

Идентификация блоков маркеров, косегрегирующих с болезнью, осуществляется с помощью различных методов статистического анализа. Наиболее эффективными считаются методы, основывающиеся на известной модели наследования признака, которая включает оценку популяционной частоты мутантного аллеля и пенетрантности генотипов [20]. Однако, установление точной модели наследования – достаточно сложная задача. Из-за этого, а также, поскольку искажение модели наследования приводит к потере мощности, популярными являются статистические методы, свободные от модели наследования. В их основе лежит анализ идентичности по происхождению маркерных аллелей у пар больных родственников [21].

2.2.2 Полногеномный анализ ассоциаций

Второй метод картирования является анализ ассоциаций, основанный на феномене неравновесия по сцеплению. Неравновесие по сцеплению между двумя аллелями разных локусов выражается в том, что частота их совместной встречаемости в популяции отличается от ожидаемой при случайной независимой встрече. Одной из основных, хотя и не единственной, причиной существования неравновесия по сцеплению в популяции является совместная передача в ряду поколений (тесное сцепление). Например, если в момент возникновения мутации, вызывающей болезнь, рядом находился определенный аллель, то в течение многих поколений этот аллель будет передаваться вместе с мутацией. Рекомбинация

постепенно разрушает ассоциацию и это происходит тем быстрее, чем дальше друг от друга расположены локусы. Для сильно сцепленных (1-2 сМ) локусов неравновесие по сцеплению сохраняется десятки поколений [22]. Основная идея картирования с помощью анализа ассоциаций заключается в следующем: если у большинства больных в популяции мутантный аллель имеет общее происхождение, окружающие маркеры находятся с ним в неравновесии по сцеплению и наследуются совместно. Для картирования гена, контролирующего болезнь, требуется найти такой маркер, один из аллелей которого находится в неравновесии по сцеплению с функциональным мутантным аллелем, который определяет повышенный риск болезни. В отличие от анализа сцепления, здесь предполагается, что у больных из разных семей этот маркер имеет не только одинаковую локализацию в геноме, но также что один и тот же маркерный аллель, находится в (одинаковом) неравновесии по сцеплению с мутацией. Если это предположение об отсутствии аллельной гетерогенности верно, при анализе ассоциаций не надо исследовать родословные, материалом для этого анализа могут служить независимые группы больных и здоровых людей. Тем не менее, предположение об общности мутации у большинства больных означает наличие общего предка, существовавшего много поколений назад. В течении времени, необходимого для распространения болезни в популяции, произошло много рекомбинационных событий, и неравновесие по сцеплению могло сохраниться только между мутацией и аллелем тесно сцепленного маркера. Поэтому с помощью анализа неравновесия по сцеплению удается локализовать ген на участке менее 1 сМ. Маркеры должны плотно располагаться на генетической карте, однако число аллелей не обязательно должно быть большим. Идеальными маркерами для анализа неравновесия по сцеплению являются SNP (single nucleotide polymorphism) маркеры (однонуклеотидные полиморфизмы).

Как видно, анализ ассоциаций обладает рядом преимуществ по сравнению с анализом сцепления, а именно, он может осуществляться на популяционных данных, имеет более высокую мощность идентификации распространенных функциональных аллелей, и обладает более высокой разрешающей способностью. Поэтому в настоящее время ПГАА является наиболее используемым методом картирования сложных признаков человека. Его применение основывается, как правило, на технологии ДНК-микрочипов, которые позволяют генотипировать сотни тысяч маркеров в геномах тысяч особей. Наличие таких данных позволяет тестировать практически весь геном на присутствие ассоциаций с исследуемым признаком, что позволяет устанавливать ассоциацию с новыми локусами, вовлеченность которых в контроль признака не могла быть предположена на основании уже имеющихся биологических и генетических знаний.

Хотя основные идеи генотипирования с помощью ДНК-микрочипов были сформулированы в конце 1980-х годов, коммерческие чипы для ПГАА стали доступны для рядового исследователя начиная с 2005 года. В том же году вышла первая публикация об успешном применении метода ПГАА для исследования сложного признака человека – возрастной дегенерации жёлтого пятна [23]. Начиная с 2005 года наблюдается быстрый рост количества публикаций упоминающих ПГАА (Рисунок 1). Вследствие большого числа проведенных исследований было найдено множество локусов, ассоциированных с различными количественными признаками и распространенными болезнями человека. Если в 2005 году локусы, вовлеченность которых в контроль сложных признаков человека была установлена, можно было пересчитать по пальцам, то на 2015 год количество найденных ассоциаций превысило 15000 (информация из базы данных “ NHGRI GWAS Catalog” [24] по состоянию на август 2015 года).

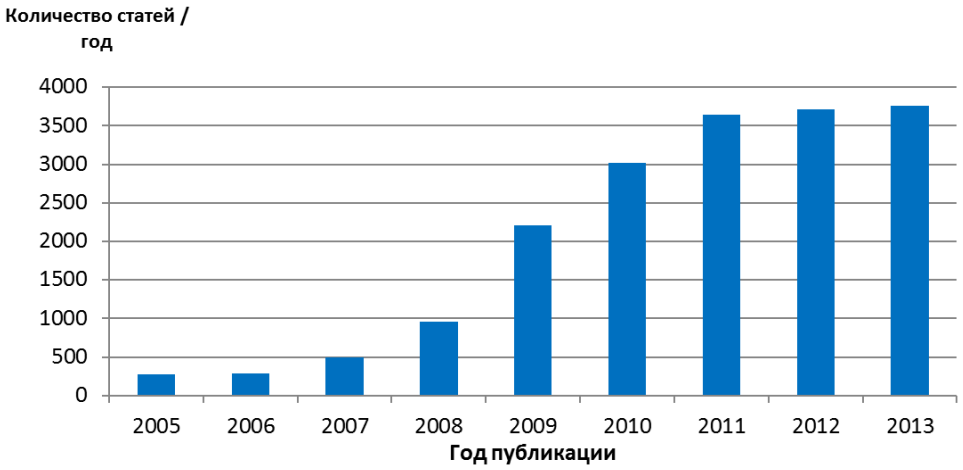


Рисунок 1. Упоминание «Genome wide association» в статьях вплоть до 2014 года (по материалам базы данных PubMed).

Результаты ПГАА могут быть представлены с помощью так называемого «Manhattan Plot» (Рисунок 2) [25]. Каждая точка на таком графике соответствует одному из тестировавшихся генетических маркеров. Ось абсцисс показывает положение маркера на карте генома, а ось ординат – десятичный логарифм от p-value, взятый с обратным знаком. То есть, чем точка выше, тем ассоциация статистически сильнее (достовернее).

Ввиду того, что при ПГАА проводится множественное тестирование, делать выводы об ассоциации с использованием номинального p-value ≤ 5% как границы значимости нельзя. Например, при исследовании 1 млн маркеров даже при отсутствии ассоциации ожидается, что около 50,000 (5% от 1,000,000) маркеров будут иметь p-value ≤ 5%. Из классической статистики известно, что если эксперименты независимы, в случае множественного тестирования можно использовать поправку Бонферрони [26]. Если проводится N независимых экспериментов, то для того, чтобы общая экспериментальная ошибка первого рода составляла α (как правило, 5%), заключение о значимом отклонении от нулевой гипотезы об отсутствии ассоциации необходимо принимать, когда номинальное p-value < $p_{гран}$, где $p_{гран} = \alpha/N$. Но, если эксперименты зависимы, использование такой поправки

приводит к статистической консервативности теста, повышая ошибку второго рода и увеличивая частоту ложноотрицательных результатов. В ПГАА тесты не являются независимыми, так как геномные маркеры находятся в неравновесии по сцеплению.

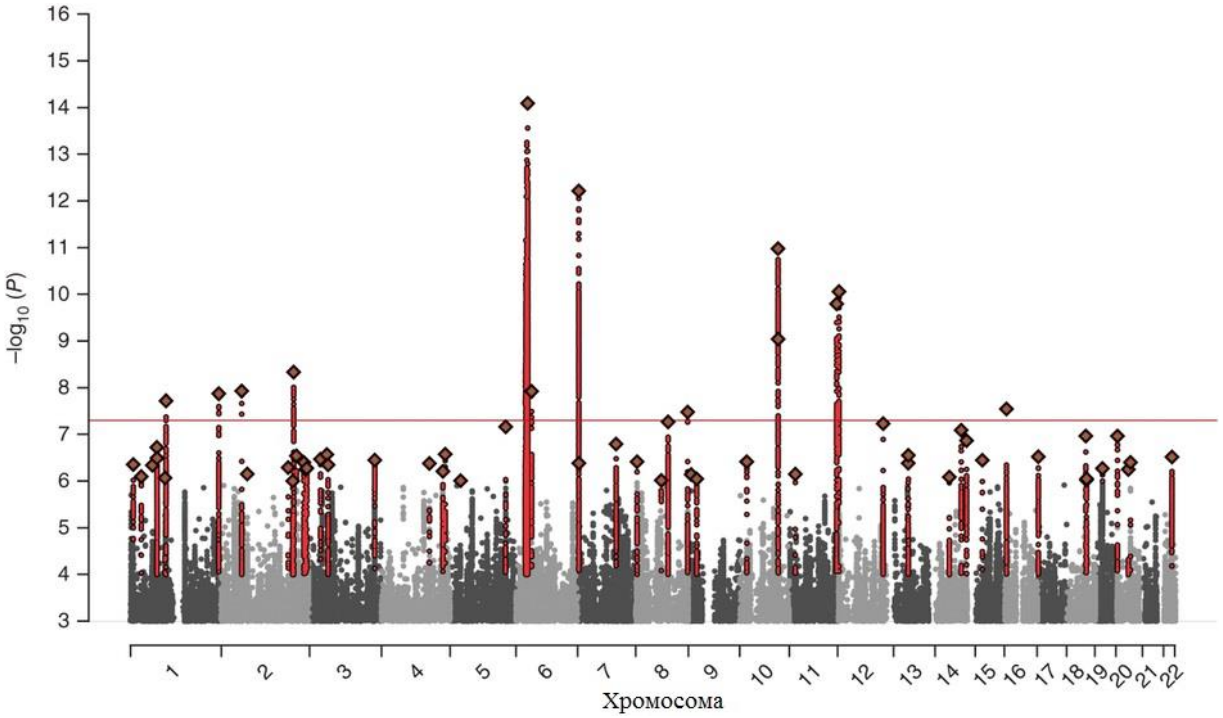


Рисунок 2. Пример представления результатов ПГАА в виде «Manhattan Plot». Каждая точка на таком графике соответствует одному из тестировавшихся генетических маркеров. Ось абсцисс показывает положение маркера на карте генома, а ось ординат – десятичный логарифм от p-value, взятый с обратным знаком.

В работе И. Пьер и др. (Pe'er I.) [27] было показано, что тестирование всех распространенных SNP (частота редкого аллеля ≥ 0.05) в европейских популяциях человека соответствует 1 млн гипотетических независимых тестов. Поэтому в настоящее время вывод о значимости ассоциации в полногеномных исследованиях делается, если p-value теста меньше фиксированного значения $p_{гран} = 5\times10^{-8}$. Горизонтальная линия на "Manhattan plot", показанном на Рисунке 2, соответствует этому граничному значению,

т.е. если точка, соответствующая тестируемому маркеру, выше линии - $\log_{10}(5\times10^{-8})$ на графике, то ассоциация считается статистически достоверной.

Для повышения мощности ПГАА требуются большие объемы выборок, недостижимые в рамках индивидуальных исследований. Поэтому зачастую результаты ПГАА ряда выборок объединяются. Технически обобщение полученных результатов проводится с использованием мета-анализа [28], который позволяет объединить результаты отдельных исследований. В Таблице 2 представлен пример такого анализа, проведенного с использованием метода, в котором вес исследования пропорционален обратной дисперсии оценки эффекта (“inverse variance based meta-analysis”).

Данные и результаты мета-анализа удобно отображать на так называемом Forest plot (Рисунок 3) [29], где по оси ординат расположены исследования, а по оси абсцисс – оценка силы ассоциации (например, коэффициент регрессии). Площадь прямоугольника на графике обратно пропорциональна квадрату стандартной ошибки эффекта и отражает мощность исследования. Таким образом, чем меньше площадь прямоугольника, тем шире доверительный интервал эффекта. Длина диагонали ромба, показывающего общий результат, соответствует доверительному интервалу суммарной оценки эффекта.

Таблица 2. Пример данных и результатов мета-анализа. β - оценка коэффициента регрессии; S.E.: стандартная ошибка оценки эффекта.

	β	S.E.	*P*-value
Оригинальное исследование	4.6	0.88	2×10^{-7}
Репликация 1	3.5	2.21	0.11
Репликация 2	3.6	1.59	0.02
Репликация 3	2.8	1.15	0.001
Мета-анализ	4.14	0.62	2×10^{-11}

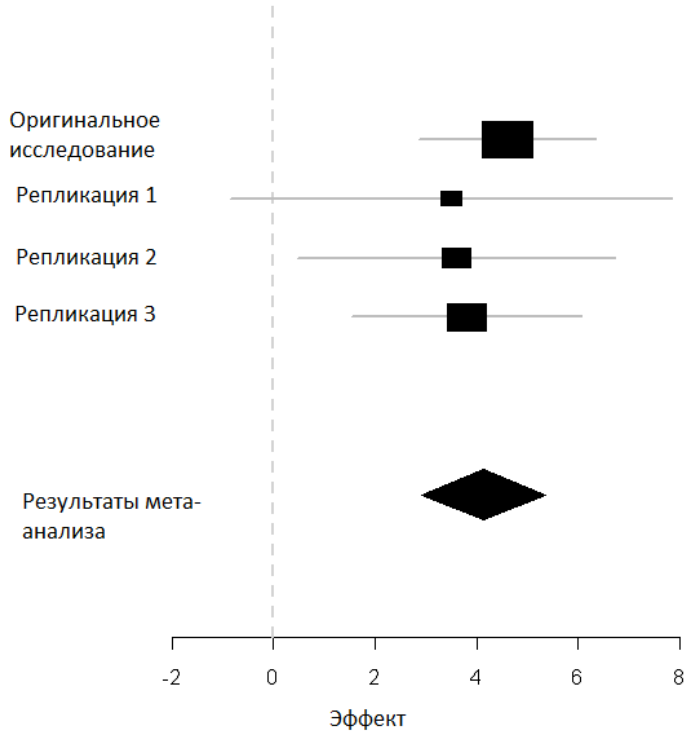


Рисунок 3. Пример представления результатов мета-анализа в виде «Forest plot».

При проведении ПГАА принято подтверждать полученные значимые ассоциации с помощью репликационного эксперимента. Для этого маркеры тестируются в дополнительной независимой выборке. Ассоциация считается реплицированной, если для репликационного эксперимента получено значение статистики, соответствующее экспериментальной ошибке первого рода $\alpha < 0.05$ [30]. Стандарт проведения репликации привел к тому, что локусы, установленные в ходе ПГАА, редко являются ложноположительными и, как правило, подтверждаются в дальнейших исследованиях.

Знание аллелей, ассоциированных с заболеванием, позволяет, по крайней мере, теоретически, предсказать риск развития этого заболевания. Однако, знание аллеля является только предпосылкой для ответа на вопрос о механизме молекулярно-генетического контроля признака. При

использовании ДНК-чипов практически всегда найденный маркер не является функциональным. Ассоциация объясняется тем, что аллели маркера находятся в неравновесии по сцеплению с функциональными аллелями, т.е. ассоциация указывает на регион, в котором находится мутация, влияющая на фенотип. Поэтому далее, как правило, проводятся ресеквенирование локуса и исследования, нацеленные на идентификацию функционального аллеля, описание механизма его действия и эффекта на определенный ген или гены. Понимание этого механизма предоставляет новую информацию о патогенезе заболевания, что, в свою очередь, позволяет разрабатывать новые биомаркеры и лекарственные средства [31–34].

2.3 Примеры использования метода ПГАА на практике

Классической статьей, в которой метод ПГАА был впервые успешно применен в том виде, в котором мы его сейчас знаем, принято считать работу Р. Дж. Кляйна и др. (R. J. Klein) [23]. В статье описываются исследования по скринингу генома для поиска генов, связанных с возрастной дегенерацией желтого пятна (age-related macular degeneration, AMD), которая является одной из причин слепоты людей старческого возраста. ПГАА проводился с использованием выборки из 96 больных и 50 здоровых людей. Генотипирование выборки проводилось с использованием коммерческого ДНК-микрочипа компании Affymetrix, содержавшего 116,204 SNP маркеров, распределенных по всем аутосомам человека. Среди исследованных SNP маркеров были найдены два маркера (см. Рисунок 4), ассоциированных с болезнью (P-value = 4.95×10^{-8}, 4.1×10^{-8}). Один из них был расположен в интроне гена фактора комплемента H (the complement factor H gene - CFH), связывающего гепарин и С-реактивный белок. У людей, гомозиготных по аллелям риска этого маркера, вероятность возникновения болезни возрастает в 7,4 раза (95% доверительный интервал 2.9-19). Секвенирование экзонов и экзон-интронных границ гена CHF показало, что аллель, повышающий риск

заболевания, находится в сильном неравновесии по сцеплению с мутацией, приводящей к замене тирозина в 402-й аминокислоте белка на гистидин. Дальнейшие независимые исследования подтвердили найденную ассоциацию на независимых выборках [35,36]. Для второго найденного маркера ассоциация объяснялась ошибкой генотипирования и в дальнейшем, как и следовало ожидать, не подтвердилась [35,36].

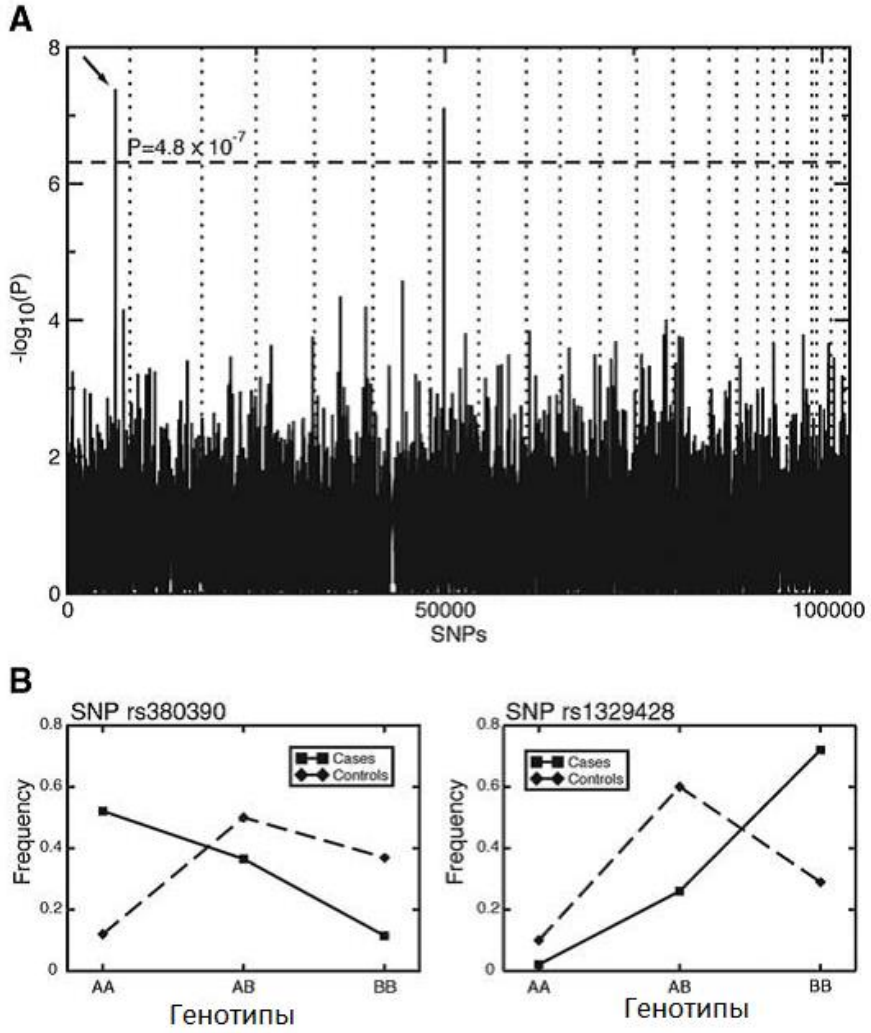


Рисунок 4. А) Результаты ПГАА возрастной дегенерации желтого пятна, представленные в виде графика “Manhattan plot”, из [23]. В) Распределение генотипов у больных и у здоровых для двух ассоциированных маркеров. Адаптировано из [23].

По сравнению с 2005 годом, в настоящее время молекулярно-генетические и вычислительные технологии сделали большой шаг вперед,

обеспечив возможность высокоточного и относительно дешевого генотипирования и анализа огромных выборок. Как пример современных возможностей ПГАА можно рассмотреть работу Т.М. Теслович (T. M. Teslovich) [37], проведенную с участием более ста соавторов. В работе описан ПГАА для поиска локусов, вовлеченных в контроль уровня липидов крови человека. В рамках исследования анализировались четыре признака: уровень липопротеинов высокой плотности, уровень липопротеинов низкой плотности, уровень общего холестерина и уровень триглицеридов в крови человека. Концентрации этих липидов и липопротеинов в сыворотке крови являются важными факторами риска развития ишемической болезни сердца. По материалу, взятому у более 100,000 человек из 46 европейских популяций, был проведен анализ ассоциации между уровнями липидов и генотипами 2.6 млн SNP-маркеров. В дальнейшем маркеры, показавшие ассоциацию, были прогенотипированы у 12,000 человек в рамках репликационного исследования. Было найдено 95 локусов, статистически значимо ассоциированных с хотя бы одним из этих 4 признаков; для 59 из них ассоциация была показана впервые. Хотя часть найденных маркеров располагалась вблизи генов, вовлеченность которых в контроль метаболизма липидов была известна (например, *CYP7A1*, *NPC1L1* и *SCARB1*), были также найдены десятки локусов, не содержащих очевидных генов-кандидатов, участвующих в метаболизме липопротеинов. Было показано, что большая часть найденных маркеров влияет на уровни липидов также в трех неевропейских популяциях (Восточной Азии, Южной Азии и Африки). Для некоторых из 95 маркеров была впервые показана ассоциация с ишемической болезнью сердца. Далее в функциональных исследованиях на модельных животных (мыши) было показано, что гены *GALNT2*, *PPP1R3B* и *TTC39B* действительно вовлечены в метаболизм липидов и связаны с риском возникновения ишемической болезни сердца [38–40].

2.4 Недостатки метода

Одной из главных проблем ПГАА является неоднозначный ответ на вопрос о конкретном полиморфизме (полиморфизмах), вовлечённом в формирование признака. Анализ сцепления маркеров позволяет определить лишь геномный регион (локус). Хотя чем выше геномное покрытие маркеров, тем точнее результаты анализа, однако точность картирования лимитирована неравновесием по сцеплению. Зачастую, локус содержит десятки генов, и определение того из них, чья функция модулируется, затруднительно и требует проведения дорогостоящего и занимающего долгое время функционального анализа. В последнее время, развитие различных молекулярно-биологических технологий сделали возможной детальную характеризацию молекулярных “омиксных” фенотипов, таких как уровни транскриптов (транскриптом), метилирования генома (эпигеноме), метаболитах (метаболоме), гликозилировании (гликоме), и так далее. Наличие таких данных позволяет в ряде случаев выдвинуть гипотезу о вовлеченности в патогенез конкретного гена в локусе, и, таким образом, проводить более направленные функциональные исследования.

Другой методологической проблемой ПГАА является возможная неоднозначность интерпретации полученных результатов: причиной ассоциации между генетическим маркером и исследуемым признаком может являться не только неравновесие по сцеплению, но и генетическая структура исследуемой популяции, миграция и дрейф генов, способ формирования выборки. Поэтому методы, позволяющие учитывать структуру выборки и проводить коррекцию полученных результатов, занимают значительное место в методологическом арсенале ПГАА.

2.4.1 Проблема «потерянной» наследуемости

Несмотря на огромные успехи метода ПГАА в идентификации новых локусов, для большинства признаков суммарный эффект найденных локусов

объясняет небольшую долю наследуемости признака. Оставшаяся необъясненной доля наследуемости признака получила название «потерянной наследуемости». Проблема потерянной наследуемости особенно очевидна на примере генетического контроля роста человека. Наследуемость этого признака составляет около 80%. Для идентификации локусов, контролирующих рост, было проведено около 50 исследований ПГАА [41–50]. В работе Ю.С. Аульченко [51], проведенной с использованием данных Роттердамского исследования [52], показано, что 54 локуса, идентифицированных на момент публикации этой работы, объясняют 4-6% дисперсии роста. В то же время метод Ф. Гальтона, описанный еще в 1886 году и основанный на корреляции между средним ростом родителей и ростом их детей, объясняет 40% дисперсии роста (показано на данных исследования ERF [53]). На Рисунке 5 представлена зависимость наблюдаемых значений роста от различных предикторов: геномный профиль по 54 локусам (Рисунок 5A), средний рост родителей (Рисунок 5B) и гипотетический случай геномного профиля по локусам, объясняющим 80% дисперсии (Рисунок 5C). Из приведенных рисунков видно, что средний рост родителей является достаточно точным предиктором по отношению к росту детей, в то время как генотипы 54 локусов объясняют лишь малую часть дисперсии признака.

В 2010 году Ланго Аллен (Lango Allen) и ряд других ученых в рамках консорциума GIANT (The Genetic Investigation of Anthropometric Traits, Генетические исследования антропометрических признаков) провели ПГАА с использованием информации о росте более чем 183,727 человек. Было найдено 180 достоверно ассоциированных локусов, суммарный аддитивный вклад которых объясняет порядка 10% дисперсии признака [4]. Аллели с максимальным вкладом объясняют каждая около 0.3%-0.7% дисперсии роста. В статье Дж. Янга и др. (J. Yang) [54] показано, что примерно 45% дисперсии роста человека может быть теоретически объяснено распространенными

аллелями (с частотой больше 0.03). Другая группа исследователей провела мета-анализ результатов ПГАА для болезни Крона (хроническое воспаление желудочно-кишечного тракта) и выявила 71 ассоциированный локус, суммарно объяснявший только 11.6% дисперсии признака [55], тогда как наследуемость признака составляет 50%.

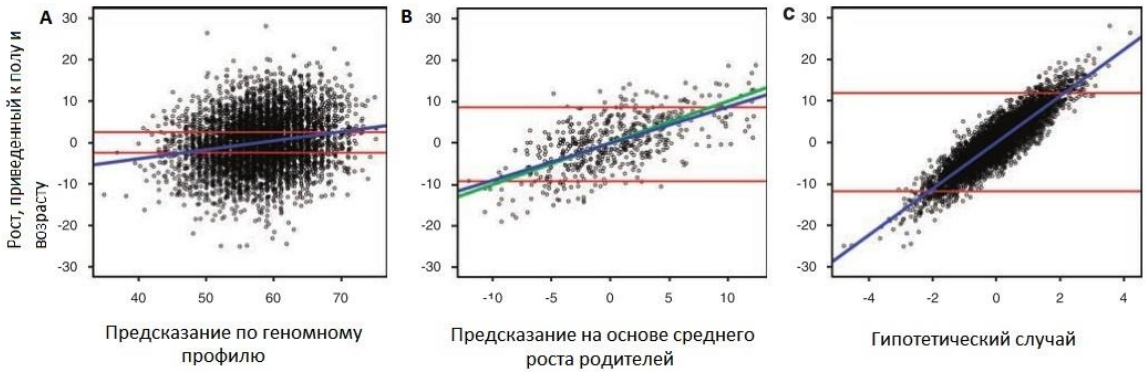


Рисунок 5. Зависимость наблюдаемого роста к предсказанному росту с учетом пола и возраста на основе различных предикторов. (A) Данные Роттердамского исследования, прогноз на основе геномного профиля по 54 локусам, (B) Данные исследования ERF, прогноз на основе среднего роста родителей, (C) Данные Роттердамского исследования, прогноз на основе гипотетического геномного профиля, объясняющего 80% дисперсии роста. Адаптировано из [51].

За проблемой потерянной наследуемости кроется неполнота наших знаний о генетической архитектуре признаков и заболеваний человека. Таким образом, объяснение феномена потерянной наследуемости является одной из важнейших задач генетики человека.

К настоящему времени предложено несколько объяснений феномена потерянной наследуемости [56]:

- наличие множества ненайденных ассоциаций с локусами, имеющими малый вклад в признак; для детекции таких локусов потребуются гигантские объемы выборок

- существование множества редких аллелей (возможно со средними эффектами); такие аллели не представлены на ДНК-чипах используемых для ПГАА;
- хромосомные перестройки (делеции, транслокации), которые могут быть ассоциированными с признаком [57,58]; такие перестройки также не представлены на ДНК-чипах, используемых для ПГАА
- неаддитивные эффекты генов;
- наличие ген-генных и ген-средовых взаимодействий [59];
- эпигенетическое влияние [59];

Все эти факторы вносят непосредственный вклад в наследуемость в широком смысле. Модели анализа, которые будут их учитывать, позволят увеличить мощность и увеличить долю объясненной наследуемости. Недавно стали появляться работы, так или иначе оценивающие вклад в генетическую дисперсию признака различных эффектов, отличных от аддитивных [60–62], хотя крупных систематических исследований в этой области по-прежнему мало. Данная ситуация обусловлена рядом причин, в том числе недостаточной проработанностью методологии анализа сложных моделей, низкой мощностью анализа, высокими вычислительными затратами и так далее. Это делает задачу разработки новых методов ПГАА актуальной в прикладном плане.

Далее более подробно будут рассмотрены модели неаддитивных эффектов генов.

2.5 Модели неаддитивных эффектов генов в статистической генетике

Для простоты предположим, что в изучаемом локусе присутствует только два аллеля, и, как следствие, в популяции возможны три генотипа.

Это предположение не накладывает ограничений на дальнейшее изложение, так как в рамках этой работы мы рассматриваем только биаллелльные модели.

Наиболее общая, *кодоминантная* (или «генотипическая») модель не накладывает ограничений на возможные значения признака при условии генотипов (см. Рисунок 6). Статистически, эта модель может быть описана тремя параметрами, которые соответствуют средним различных генотипических групп. Как правило, при описании используется терминология регрессионного анализа. Тогда ожидание признака при условии одного из генотипов (как правило, гомозиготы по более распространенному аллелю) принимается за регрессионный отступ, а для двух других генотипов оценивается отклонение ожидаемого значения признака от отступа. При определении значимости эффекта генотипа на признак кодоминантная модель сравнивается с нулевой моделью, описываемой одним параметром (отступом). Такое сравнение приводит к статистическому тесту с двумя степенями свободы.

Ряд «ограниченных» моделей получается при накладывании различных ограничений на возможное значение признака в разных генотипических группах. *Аддитивная* модель наследования количественного признака предполагает, что ожидаемое значение признака у гетерозигот является средним ожидаемых значений признака гомозигот. *Сверхдоминантная* модель предполагает, что значение признака у обоих типов гомозигот одинаково, в то время как значение признака может быть другим у гетерозиготы. Если один из аллелей, называемым доминантным, доминирует над другим (называемым рецессивным), значение признака равно для всех генотипов, где присутствует доминантный аллель, независимо от числа копий этого аллеля. Это значение отличается от того, что ожидается для генотипа, являющегося гомозиготой по альтернативному аллелю, называемому рецессивным. В том случае, если отступ приравнивается к

ожиданию признака при условии рецессивной гомозиготы, исследуется эффект доминантного аллеля, и модель называется *доминантной*. При исследовании эффекта рецессивного аллеля в рамках *рецессивной* модели за отступ принимается значение генотипов, несущих доминантный аллель.

Статистически такие ограниченные модели могут быть описаны с использованием двух параметров (отступа и генетического эффекта). Нулевая гипотеза об отсутствии генетического эффекта требует оценивания одного параметра – отступа. Таким образом, сравнение гипотезы генетического контроля против нулевой гипотезы приводит к тесту с одной степенью свободы. Ограниченные модели могут быть также сравнены с более общей кодоминантной, что также ведет к тесту с одной степенью свободы.

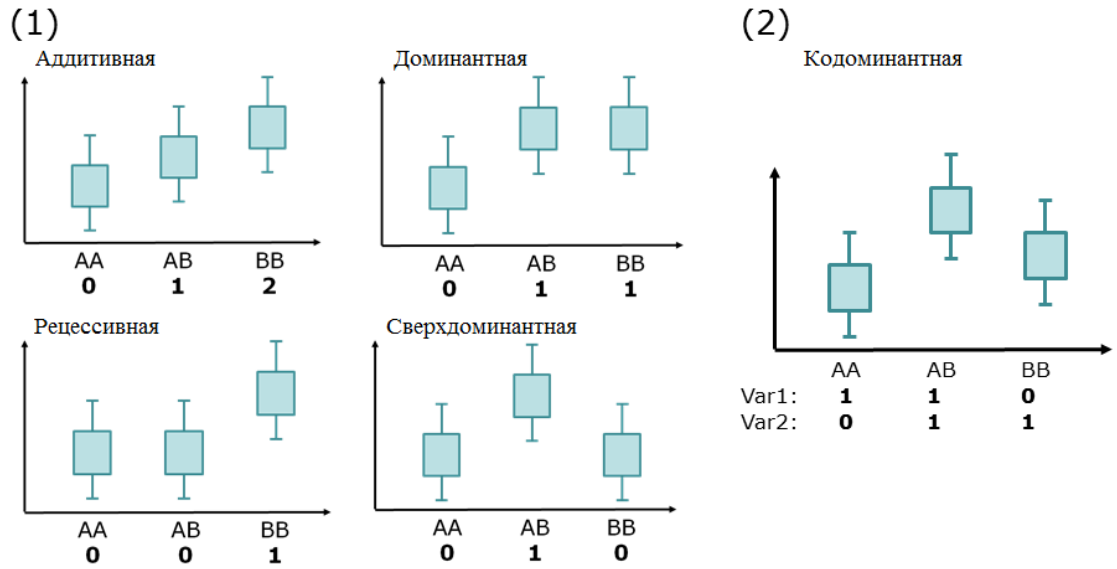


Рисунок 6. Схематический обзор эффектов генов на количественные фенотипы: (1) Аддитивная, доминантная, рецессивная и сверхдоминантная модель. Генотипический эффект, предполагаемый в этих моделях, может быть закодирован одной переменной. Значения этой переменной представлены жирным шрифтом под соответствующими генотипами. (2) Кодоминантная модель. Генотипический эффект, предполагаемый в этой модели, кодируется двумя переменными (Var1 и Var2).

2.6 Геномный контроль в ПГАА

Как уже говорилось выше, одной из основных методологических проблем при применении метода ПГАА является возможная неоднозначность интерпретации полученных результатов вследствие генетической структурированности исследуемой выборки.

Для бинарных признаков поиск ассоциации, как правило, проводится с помощью сравнения распределения частот генотипов в группах больных и здоровых особей, или с помощью логистической регрессии. Для анализа ассоциации количественных признаков можно применять линейную регрессию признака на генотип маркерного локуса [30]. Однако, такие стандартные статистические методы предполагают, что фенотипы особей коррелируют с генотипами только тех локусов, которые участвуют в контроле анализируемого признака. Это предположение не всегда выполняется для реальных данных, так как анализируемая выборка может быть генетически гетерогенной; кроме того, в выборке могут присутствовать родственники. Близкие родственники имеют сходные фенотипы, в то же время, большие доли их геномов идентичны по происхождению. Это приводит к появлению неспецифических, наведенных ассоциаций. В такой ситуации стандартные методы имеют высокую ошибку первого рода, то есть, повышена вероятность того, что анализ приведет к ложноположительному результату (ложному утверждению о наличии «статистически значимой ассоциации» и, как следствие, «идентификации локуса»).

Чтобы избежать ложноположительных выводов при интерпретации результатов ПГАА, необходимо проводить их коррекцию с учетом родственной структуры выборки.

Одним из статистических методов, позволяющих проводить коррекцию результатов ПГАА, является геномный контроль (ГК), который основывается на использовании информации о несвязанных с признаком маркерах [5]. На

сегодняшний день предложено несколько методов для ГК [5–9]. Делвин Б. и Рёдер К. (Devlin B., Roeder K.) [5] предложили использовать коэффициент коррекции, названный фактором инфляции дисперсии (variance inflation factor, VIF), для коррекции распределения тестовой статистики. Так же в литературе используется обозначение «фактор инфляции λ», являющееся синонимом фактора инфляции дисперсии VIF, однако обозначение «VIF» обычно используется в значении некоторой корректирующей функции (которая при определенных условиях является константой), а «λ» – в значении некоторого корректирующего фактора (необязательно функции), а также в качестве значения, описывающего степень инфляции тестовой статистики (см. ниже).

Было показано, что VIF является функцией частоты аллелей изучаемых маркеров и нескольких популяционных параметров [5]. Также было показано, что для аддитивных моделей VIF не зависит от частоты аллелей. Таким образом, для аддитивной модели VIF является константой, и может быть эмпирически определен для нулевых (не связанных с признаком) локусов. Стоит отметить, что для редких аллелей и меньших размеров выборок такое асимптотическое предположение не работает, а следовательно, инфляция тестовой статистики будет зависеть от частоты аллелей даже для аддитивной модели [6].

При оценке инфляции тестовой статистики, как правило, используются все маркеры, равномерно распределённые в геноме. Хотя часть маркеров может быть реально ассоциирована с признаком, считается, что доля таких маркеров мала, а следовательно, мал и их эффект на общее распределение тестовой статистики. Для оценки степени инфляции (фактора инфляции λ) могут быть использованы различные методы. Наиболее часто используемой является медианная оценка (λ_{median}), которая определяется как соотношение между медианой исследуемого распределения тестовой статистики и

медианой распределения $\chi^2_{df=1}(0.455)$ [5]. Другой оценкой является среднее распределения тестовых статистик; однако, эта оценка может быть сильно смещена при наличии сильных сигналов ассоциации. Еще одна оценка может быть определена как коэффициент регрессии изучаемой тестовой статистики на ожидаемое распределение статистики при нулевой гипотезе (регрессионная оценка - $\lambda_{regress}$). Эта оценка возникает из простого наблюдения, что ковариация между двумя упорядоченными случайными переменными, одна из которых распределена как $\chi^2_{df=1}$, а другая как $\lambda * \chi^2_{df=1}$, равна 2*λ, тогда как дисперсия ожидаемого распределения статистики теста равна 2. Все эти оценки являются константами, которые могут использоваться как индикаторы инфляции статистики или как коэффициенты, позволяющие скорректировать полученную тестовую статистику.

Общая формулировка VIF [6], в принципе, позволяет расширить применимость ГК для доминантной и рецессивной моделей наследования. Однако, для неаддитивных моделей VIF зависит как от параметров, описывающих генетическую структуру выборки, так и от частоты аллелей. Таким образом, оценить VIF эмпирически как для аддитивной модели возможно, только если частоты аллелей нулевых маркеров совпадают с таковыми для тестового маркера (специфический VIF для каждой из групп частот аллелей). Альтернативный путь предполагает оценку параметров структуры популяции. Существующие методы, учитывающие структуру популяции и кластеризующие особей [63], являются вычислительно трудоемкими.

Еще один метод эмпирической оценки VIF был предложен Г. Женгом и др. (Zheng G.) [7] для кодоминантной модели с двумя степенями свободы (2df), которая не накладывает ограничения на модель наследования. Этот

метод основывается на комбинации скорректированных тестовых статистик доминантной и рецессивной моделей [7] (см. Материалы и методы).

Метод ГК применим только в том случае, если ПГАА был проведен при предположении об аддитивном вкладе аллелей анализируемого локуса. В настоящее время не существует метода, позволяющего оценивать VIF для произвольной модели наследования. Кроме того, предложенные ранее поправки для рецессивной, доминантной и генотипической моделей, не тестировались на реальных данных и не были реализованы в программных пакетах для обработки генетических данных. Таким образом, разработка метода, позволяющего проводить оценку инфляции тестовой статистики произвольной модели наследования, является одной из насущных задач современной статистической геномики.

Отсутствие методов ГК для неаддитивных моделей наследования является одной из причин малого числа практических исследований неаддитивных эффектов генов на реальных данных.

2.7 Неаддитивные эффекты генов, контролирующих метаболом человека

Существующие теории доминантности подчеркивают актуальность анализа неаддитивных генетических эффектов при исследовании высокоразмерных данных функциональной геномики. В настоящий момент проблема неаддитивных эффектов в генетическом контроле признаков человека рассмотрена лишь в небольшом числе работ. В нескольких исследованиях, где проводилось сравнение аддитивных и неаддитивных эффектов полиморфизмов, поддерживается широко распространенная презумпция аддитивных эффектов. В. Г. Хилл (Hill W. G.) с группой соавторов (2008) рассмотрели ряд эмпирических исследований и показали, что сложные признаки, как правило, контролируются аддитивно.

Неаддитивные эффекты в большинстве случаев незначительно влияют на дисперсию признака [60]. Другие исследования показали, что экспрессия генов контролируется, главным образом, аддитивно [61].

На этом фоне долгое время продолжается дискуссия о встречаемости и значимости неаддитивных эффектов [11–14,64–66]. Теория эволюции доминантности Р. Фишера предполагает, что доминирование - результат селекции, которая приводит эффект гетерозиготного генотипа в соответствие с «нормальным» гомозиготным генотипом [14,64]. С. Райт (Wright S.) (1929) и Дж. Би. С. Холдейн (Haldane J. B. S.) (1930) утверждали, что доминантность является следствием физиологических факторов, и что отбор не является основной причиной доминантности [12,13]. Последующие исследователи, как правило, принимали точку зрения Райта, но не полностью исключали и аргументы Фишера, утверждая, что он может быть прав при определенных обстоятельствах [11,65]. Х. Касисер и Дж. Бернс (Kacser H., Burns J. A.) (1981) предположили, что доминантность является «следствием кинетических свойств ферментативного пути». Их современный биохимический взгляд на доминантные/рецессивные эффекты основывался на наблюдении, что увеличение скорости реакции нелинейно зависит от активности ферментов или концентрации, и замедляется к верхнему пределу активности. Они утверждают, что ни один фермент не может считаться независимым от других, а включен в целую систему «потоков» (fluxes), которые взаимосвязано снижают ответную реакцию исследуемого признака (концентрация вещества) на увеличение активности ферментов [65].

Эта теория доминантности имеет особое значение в отношении метаболитов, т.к. метаболиты являются прямыми продуктами ферментативных реакций, а следовательно, находятся ближе к действию генов и генетических сетей, чем клинические фенотипы. Если ген (SNP) имеет неаддитивный эффект на уровень метаболитов, предположение аддитивного эффекта может снизить мощность анализа. Таким образом,

предыдущие ПГАА могли упустить ассоциации с локусами, имеющими неаддитивные эффекты.

Развитие высокопроизводительных молекулярно-биологических технологий в настоящее время позволяет измерять в одном образце сотни метаболитов (как правило, это небольшие молекулы, являющиеся промежуточными продуктами метаболизма). В 2008 году К. Гигер (Gieger C.) с соавторами [67] опубликовали первый ПГАА концентраций метаболитов сыворотки крови человека. Последовали и другие успешные исследования генетического контроля метаболитов [68–77], что позволило сформировать целое новое направление – генетическую метаболомику. Многие локусы, определенные в исследованиях генетической метаболомики, позволили лучше разобраться в механизмах, лежащих в основе липидного, углеводного и аминокислотного обмена. Так как метаболиты являются биомаркерами различных заболеваний обмена веществ, энергетического обмена, заболеваний сердечно-сосудистой системы [78], поиск конкретных аллелей, вовлеченных в генетический контроль метаболитов, является актуальной задачей для медицины.

Большинство ПГАА на метаболомных данных анализируют аддитивные эффекты генов на уровне метаболитов и пренебрегают другими возможными генетическими эффектами. Одной из причин этого, помимо вычислительной сложности неаддитивных моделей, является вопрос мощности: если параллельно исследуется несколько генетических моделей, необходимо проводить коррекцию уровня значимости на множественное тестирование, что может снизить мощность анализа.

2.8 Краткое заключение

Из вышесказанного следует, что существующая теоретическая и практическая база ПГАА для анализа неаддитивных эффектов требует

дальнейшего развития и разработки. Используемая в большинстве случаев аддитивная модель не является единственно возможной, и ее огульное применение может уменьшать мощность анализа. Поскольку методологическая база ПГАА с использованием неаддитивных моделей не полна, роль неаддитивных эффектов в контроле различных признаков человека изучена недостаточно. Целью данной работы является разработка, апробация и применение методов полногеномного анализа ассоциаций с использованием неаддитивных моделей наследования (рецессивные, кодоминантные, доминантные и сверхдоминантные).

3 Материалы и методы

Настоящая работа была выполнена с использованием симуляционных данных и четырех наборов реальных данных из трех различных исследований. Все реальные данные были предоставлены в рамках сотрудничества с другими лабораториями. Далее, в подразделе «Материалы» мы описываем выборки, из которых были получены реальные данные. Методы описаны в следующем разделе, который разбит на два подраздела, описывающих методы, применявшиеся для валидации разработанных методов ГК, и методы, применявшиеся для анализа неаддитивных эффектов генов. Задачей настоящей работы являлась также разработка методов; эти новые методы описаны в разделе «Результаты».

3.1 Материалы

3.1.1 Данные исследования ERF

Для валидации разработанных методов ГК использовались реальные данные, полученные в результате кросс-секционного исследования ERF (Erasmus Rucphen family), проводящегося в генетически изолированной популяции юго-западной части Нидерландов [79]. Все протоколы исследования были одобрены комитетом по медицинской этике университета Эразмуса, а все участники дали письменное информированное согласие в соответствии с Хельсинской Декларацией. Участники исследования являются членами одного большого генеалогического древа, которое может быть прослежено в 23 поколениях и которое содержит тысячи петель [80]. Выборка, использовавшаяся для моделирования, включала 3,235 человек, по которым была доступна информация о 54,000 генотипированных SNP маркерах. Все включенные SNP имели частоту кодирующего аллеля (CAF)

$0.05 \leq CAF \leq 0.95$ и долю прогенотипированных особей по каждому SNP (call rate) ≥ 0.95.

Мы анализировали уровни липопротеидов высокой плотности (ЛПВП) на импутированных (данных, в которых вероятность неизвестных генотипов восстановлена с использованием гаплотипов из референсной выборки [81]) генотипических данных. Этот набор данных включал 2,699 человек, для которых были известны генотипы 1,093,818 SNP. Все SNP в выборке имели $0.05 \leq CAF \leq 0.95$ и call rate ≥ 0.95. Более детальное описание выборки можно найти в публикации Ю.С. Аульченко [82].

3.1.2 Данные исследования KORA

Данные исследования KORA использовалось в работе дважды: для валидации методов ГК и для поиска неаддитивных эффектов генов.

KORA (Cooperative Health Research in the Region of Augsburg) – популяционное исследование в регионе Аугсбург в южной Германии, проводимое с 1999 года по настоящее время [83]. В настоящем исследовании использовались данные подгруппы этого исследования - KORA F4, проводимое с 2006 по 2008 год. Все протоколы исследований были одобрены этическим комитетом медицинской палаты Баварии (Bayerische Landesärztekammer), а все участники дали письменное информированное согласие. Дизайн исследования подразумевал отсутствие популяционной стратификации выборки, поэтому в данной выборке близкие родственники не представлены. Возможная популяционная структурированность выборки изучалась с помощью программы EIGENSOFT [84]. Генотипирование проводилось с использованием платформы Affymetrix 6.0 (534,174 SNP маркеров после контроля качества) с последующим импутированием с использованием HapMap2 (панель 22) в качестве референсной панели с общим числом SNP 1,717,498 (детали представлены в работе [69]). Все SNP в исследовании имели $0.05 \leq CAF \leq 0.95$ и call rate ≥ 0.95.

Для валидации методов ГК мы анализировали уровни мочевой кислоты в наборе данных, включавшем 1,785 человек. Более подробное описание дизайна исследования, генотипирования и фенотипирования представлено в [69].

Для апробации предложенных методов анализа неаддитивных эффектов мы использовали данные 1,785 участников KORA F4, для которых в сыворотке крови натощак измерялись концентрации 151 метаболитов. Концентрации определялись с помощью ионизирующей электроспрейной тандемной масс-спектрометрии и набора Absolute*IDQ*™ p150 (BIOCRATES Life Sciences AG, Инсбрук, Австрия). После контроля качества была получена информация о концентрациях 151 метаболитов, из них: 20 – карнитины и ацилкарнитины, 12 – гидро- и карбоксиацилкарнитины, 14 – сфингомиелины и гидроксисфингомиелины, 36 – диацилфосфатидилхолины, 38 – ацил-алкилфосфатидилхолины, 13 - лизофосфатидилхолины, 14 – аминокислоты, 4 - углеводы. Методы измерения, описание фенотипов и контроль качества данных описаны ранее [71,85,86].

3.1.3 Исследование TwinsUK

Для репликации результатов анализа неаддитивных эффектов мы использовали данные крупного Британского близнецового исследования TwinsUK [87]. Этическое одобрение было получено от этического комитета госпиталя Гая и св. Томаса, все участники исследования дали письменное информированное согласие. Всего было генотипировано 2,277 человек европейского происхождения с использованием чипа Illumina Hap317K. Измерения метаболитов проводились с помощью той же таргетной метаболомной платформы и по такому же протоколу, как и в исследовании KORA в Центре Геномного Анализа имени Гельмгольца в Мюнхене. Для более подробного описания см. [86,88].

3.2 Валидация методов геномного контроля

3.2.1 Моделирование и симуляции

Фенотипы для оценки ошибки первого рода и мощности моделировались с использованием реальных генетических данных из исследования ERF по схеме, описанной ниже. В настоящей работе, как и в работе Г. Чжэн и др. [6], VIF был определён для бинарных признаков. Это, однако, не накладывает ограничений на применимость нашего подхода для количественных признаков (см. работу [89]).

Предрасположенность (liability scores) моделировалась как сумма эффектов независимых локусов и случайного, нормально распределенного, «средового» эффекта. Коэффициент наследуемости предрасположенности был установлен равным случайному числу из равномерного распределения, ограниченного 0.5 и 0.8, эффекты генов и средовой эффект суммировались таким образом, чтобы доля дисперсии предрасположенности, объясненная эффектами генов, была равна наследуемости. Для оценки мощности, моделировался локус с основным эффектом. Для этого, основываясь на частоте минорного аллеля (MAF), случайным образом был выбран SNP, эффект которого определялся следующим образом: для оценки ошибки первого рода SNP должен был объяснять 0% дисперсии предрасположенности, и 0.35% для оценки мощности. Для того, чтобы смоделировать полигенный эффект, были случайным образом выбраны 500 маркеров (исключая хромосому, в которой располагался SNP основного эффекта). Основываясь на частоте их аллелей, эффекты были определены так, что каждый SNP объяснял одну и ту же долю дисперсии предрасположенности. Количественный фенотип был преобразован в бинарный признак в соответствии с пороговой моделью (0, если предрасположенность была ниже порога, соответствующего 1/3 распределения, и 1, иначе). Таким образом, результирующая выборка

содержала 1/3 случаев и 2/3 контролей. Для каждого сценария для изучения ошибки первого рода было выполнено 1,000 симуляционных циклов; для изучения мощности было выполнено 100 симуляционных циклов.

3.2.2 Анализ ассоциации

Для анализа симулированных и реальных данных мы использовали стандартные тесты, реализованные функцией GWFGLS (genome-wide feasible generalized least squares) пакета MixABEL, который является частью набора программ GenABEL [90] для статистической геномики. Мы использовали опцию “score” для того, чтобы результаты GWFGLS для бинарных признаков были в точности такими же, как и для тренд-теста Кохрана-Армитажа. ПГАА был проведен для 5 различных (аддитивная, доминантная, рецессивная, сверхдоминантная, кодоминантная) моделей.

Для анализа количественных признаков мы использовали регрессию и тест множителей Лагранжа (score test), как это реализовано в MixABEL. Для анализа импутированных данных выполнялась регрессия на вероятности генотипов.

Результаты ПГАА корректировались с использованием различных методов ГК: стандартный метод, который корректирует статистику делением на коэффициент λ, а также разработанные методы (описанные далее в результатах).

3.2.3 Тест кодоминантной модели, основанный на комбинации скорректированных тестов для рецессивной и доминантной моделей

Г. Женг и соавторы (Zheng G.) [7] предложили для кодоминантной модели наследования следующий тест, имеющий распределение хи-квадрат с двумя степенями свободы:

$$X^2 = \frac{Z_0^2 + Z_1^2 - 2\rho' Z_0 Z_1}{1 - \rho'^2} \quad ,$$

где $\rho' = \sqrt{\frac{m_0 m_2}{(m_0+m_1)(m_1+m_2)}}$; *m0, m1 и m2* – количество людей с генотипом *aa, Aa* и *AA* соответственно, а Z_0^2 и Z_1^2 - значение скорректированных одностепенных тестов множителей Лагранжа (score test) для рецессивной и доминантной модели наследования, соответственно.

3.3 Методы, применявшиеся при поиске неаддитивных эффектов генов

3.3.1 Полногеномный анализ ассоциаций

Нами были проанализированы концентрации метаболитов (151) и все возможные отношения концентраций между парами (22,650). Все признаки предварительно были скорректированы на пол, возраст и номер пробы, а далее нормализированы с использованием обратно-нормальной трансформации (inversed-normal transformation [91]).

Для ПГАА использовались стандартные тесты, включенные в функцию GWFGLS (genome-wide feasible generalized least squares) с приближением Вальда для пакета MixABEL, который является частью набора программ GenABEL [90] для статистической геномики. Для анализа импутированных данных использовалась регрессия на генотипические вероятности. При анализе применялись как общая кодоминантная, так и ограниченные (аддитивная, рецессивная, доминантная и сверхдоминантная) модели.

Для коррекции возможной инфляции тестовой статистики применялись методы геномного контроля, в том числе разработанные в рамках данной диссертации.

Учитывались только SNP с долей неизмеренных генотипов менее 0.05, качеством импутации $\geq$ 0.3, значением отклонения от закона Харди-Вейнберга (HWE) p-value $\geq 10^{-6}$ и MAF $\geq$ 0.1. Более того, мы исключили все

SNP, для которых число особей с наименее вероятным генотипом было менее 30.

С целью уменьшения вычислительной трудоемкости анализа, мы провели ПГАА с использованием только прогенотипированных (неимпутированных) SNP (482,616 SNP), далее отобрали все SNP с либеральным порогом p-value $\leq 5\times10^{-7}$ и проанализировали все импутированные SNP в локусе размером 500 тпн (±250 тпн от найденных SNP). Далее, мы проверили, находятся ли значимо ассоциированные SNP на одной и той же хромосоме на расстоянии менее 250 тпн друг от друга - такие SNP мы считали находящимися в одном локусе. Для каждого локуса, была выбрана пара SNP-признак, показывающая самую сильную ассоциацию (наименьшее p-value). На основе отобранных данных были сделаны региональные графики ассоциаций для проверки наличия других сигналов ассоциаций в регионе 1,000 тпн.

3.3.2 Репликация

Для репликации мы использовали данные исследования TwinsUK. К фенотипам была применена та же трансформация, что и в исследовании KORA. Порог репликации был выбран при значении p-value = 0.05/20=0.0025 с коррекцией Бонферрони для кодоминантной модели и значении p-value = 0.05/22≈0.0023 для аддитивной модели. Мы использовали те же SNP и те же метаболиты или их отношение для репликации результатов, которые мы определили на этапе поиска.

4 Результаты

4.1 Геномный контроль при неаддитивных моделях наследования

Мы расширили возможности метода ГК для неаддитивных моделей, что позволило нам использовать для ГК маркеры с произвольной частотой аллелей. Аналитические выражения для фактора инфляции тестовой статистики, описывающей зависимость от частоты аллелей и нескольких популяционно-генетических параметров, были получены для рецессивных, доминантных и сверхдоминантных моделей наследования. Мы предложили метод для оценки требуемых параметров. Более того, мы предложили метод ГК, основанный на приближении коэффициента коррекции полиномиальной функцией частоты аллелей, и описали процедуру коррекции кодоминантного (две степени свободы) теста для случаев, когда модель наследования неизвестна. Статистические характеристики описанных методов были исследованы с использованием моделированных и реальных данных. Мы продемонстрировали, что все рассмотренные методы были эффективны для контроля ошибки первого рода в присутствии генетической гетерогенности выборки. Все методы, разработанные и протестированные в данной работе, были воплощены с использованием языка R как часть пакета GenABEL.

4.1.1 Тест множителей Лагранжа (score test) для анализа ассоциаций

Для ПГАА бинарных признаков обычно используют дизайн эксперимента типа «случай-контроль». В таких исследованиях формируются две группы индивидуумов: «случаи», которые обладают исследуемой характеристикой, например, болезнью, и «контроли», не обладающие этой характеристикой, например, здоровые люди. Методологически анализ

ассоциаций бинарных признаков при таком дизайне эксперимента проводится следующим образом.

Для каждого маркерного локуса в каждой группе подсчитывается число людей с каждым из возможных генотипов (Таблица 3).

Таблица 3. Распределение генотипов диаллельного маркера в выборке “случай-контроль”.

	aa	*Aa*	*AA*	Всего
Случай	r_0	r_1	r_2	R
Контроль	s_0	s_1	s_2	S
Всего	m_0	m_1	m_2	N

Для формализации ограниченных моделей наследования вводится понятие эффекта генотипа – t_i, где i - индекс генотипа (0,1,2 для *aa*, *Aa*, *AA*). Для аддитивной, рецессивной и доминантной моделей принято, что для генотипа *aa* эффект равен 0, для генотипа *AA* – единице, а эффект генотипа *Aa* различен при разных моделях наследования (Таблица 4).

Таблица 4. Формализация ограниченных моделей наследования с помощью различной кодировки генотипа t_i.

	Модель		
Генотип	Рецессивная	Аддитивная	Доминантная
aa	0	0	0
Aa	0	½	1
AA	1	1	1

Для тестирования таких моделей используется статистический тест Кохрана-Армитажа, принимающий во внимание силу воздействия фактора (эффект генотипа t_i) в двух группах. Этот тест является особым случаем теста множителей Лагранжа и проверяет гипотезу о равной частоте генотипов в группах больных и здоровых с учетом модели наследования (чем больше эффект генотипа, тем с большим весом разница частот генотипов у больных и здоровых входит в статистику). Статистика Кохрана-Армитажа

вычисляется как $Z^2 = T^2/Var(T)$ и при больших объемах выборки ее распределение аппроксимируется распределением хи-квадрат с одной степенью свободы. Величины T и $Var(T)$ для данных Таблица 3 определены формулами:

$$T = \sum_{i=0}^{2} t_i(r_i S - s_i R)$$

$$Var(T) = \frac{SR}{N}\left(\sum_{i=0}^{2} t_i^2 m_i(N - m_i) - 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} t_i t_j m_i m_j\right)$$

Принимая $t_0=0$, $t_1=1/2$ и $t_2=1$, тест множителей Лагранжа (score-test) для кодоминантной модели выглядит следующим образом:

$$Z_x^2 = \frac{N[N(xr_1 + r_2) - R(xm_1 + m_2)]}{R(N - R)[(x^2 m_1 + m_2) - (xm_1 + m_2)^2]}$$

В данной статистике присутствует только один параметр – x - эффект гетерозиготного генотипа, который задается выбранной моделью наследования.

4.1.2 ГК для произвольной модели наследования

Как было сказано ранее, в реальной ситуации распределение статистики, как правило, не описывается центральным распределением Хи-квадрат вследствие генетической структурированности выборки. Ввиду этого, нами был предложен следующий алгоритм для подсчета корректирующего коэффициента.

Обозначим через $G_i \in \{0,1,2\}$, i = 1,..,R маркерный генотип у i-го больного, а через H_j, j = 1,..,S – то же у j-го здорового члена выборки. Значение, полученное при проведении теста Кохрана-Армитажа Z^2 пропорционально квадрату статистики T, определенной как

$$T = \sum_{i=1}^{R} G_i - \sum_{j=1}^{S} H_j$$

В общем виде, дисперсия *T* может быть записана как:

$$Var(T) = \sum_{i=1}^{R} Var(G_j) + \sum_{j=1}^{S} Var(H_j) + 2\sum_{i<l} cov(G_i, G_l) + 2\sum_{j<l} cov\left(H_j, H_l\right) - 2\sum_{i}\sum_{j} cov\left(G_i, H_j\right)$$

При справедливости нулевой гипотезы об отсутствии связи между маркером и болезнью, математическое ожидание *E(T) = 0*, а дисперсия

$Var(G_j) = Var(H_j)$ и $cov(G_i, G_l) = cov\left(H_j, H_l\right) = cov\left(G_i, H_j\right)$, так что

$$Var_0(T) = NVar(G_j) - Ncov\left(G_i, G_j\right) \quad (1)$$

В случае, когда изучаемые особи выбраны из одной гомогенной и равновесной популяции, при нулевой гипотезе:

$$Var_0(T) = NVar(G_j) \quad (2)$$

При равновесии Харди-Вайнберга дисперсия *G* принимает вид:

$$Var(G_j) = [2pqx^2 + p^2] - [2pqx + p^2]^2 \quad (3)$$

где *p* – частота аллеля *A* в популяции, а *q = 1-p*.

Если равновесие Харди-Вайнберга нарушено, например, из-за родственной структуры или подразделенности популяции, то известно, что частоты гомозиготных генотипов увеличиваются, а гетерозиготного - уменьшается [92]:

$$\Pr(AA) = Fp + (1 - F)p^2$$

$$\Pr(Aa) = 2(1 - F)pq$$

$$\Pr(aa) = Fq + (1 - F)q^2$$

Здесь F – коэффициент инбридинга Райта. При этом $Var(G_i)$ принимает вид

$$Var(G_j) = [2pq(1-F)x^2 + Fp + (1-F)p^2] - [2pq(1-F)x + Fp + (1-F)p^2]^2$$

Или

$$Var(G_i) = p(2(F-1)(p-1)x^2 - p(2(F-1)(p-1)x + F(-p) + F + p)^2 - Fp + F + p) \quad (4)$$

Кроме того, $F \neq 0$ подразумевает наличие ковариаций между членами популяции. Чтобы оценить их, прежде всего, нужно оценить совместное распределение генотипов у пары особей, что было сделано в работе Г. Чжэн и др. (Zheng G.) [6].

Таблица 5. Вероятности совместного распределения частот генотипов.

G_1,G_2	N	Pr(G_1,G_2)*(1+F)(1+2F)
AA,AA	1	$6F^3p + 11F^2(1-F)p^2 + 6F(1-F)^2p^3 + (1-F)^3p^4$
AA,Aa	4	$2F^2(1-F)pq + 3F(1-F)^2p^2q + (1-F)^3p^3q$
AA,aa	2	$F(1-F)pq + (1-F)^3p^2q^2$
Aa,Aa	4	$F(1-F)pq + (1-F)^3p^2q^2$
aa,Aa	4	$2F^2(1-F)pq + 3F(1-F)^2pq^2 + (1-F)^3pq^3$
aa,aa	1	$6F^3q + 11F^2(1-F)q^2 + 6F(1-F)^2q^3 + (1-F)^3q^4$

Если частоты генотипов у пар особей известны (см. Таблицу 5), ковариация между эффектами генотипов у этих особей равна

$$cov(G_i, G_j) = [Pr(AA, AA) - (Pr(AA))^2] + 2x[Pr(AA, Aa) - 2\,Pr(AA)\,Pr(Aa)] + x^2[\mathrm{Pr}(Aa, Aa) - (Pr(Aa))^2] \quad (5)$$

или

$$cov(G_i, G_j) = \Pr(AA, AA) - (Fp + (1-F)p^2)^2 + 2[xPr(AA, Aa) - 4xpq(Fp + (1-F)p^2)(1-F)] + x^2[\Pr(Aa, Aa) - (2(1-F)pq)^2] \quad (6)$$

Подставив все значения из таблицы 5, можно получить

$$cov(G_i, G_j) = -\frac{2F(p-1)p(F^3(p-1)p(1-2x)^2+F^2(p(8x-4)-4x+3)}{(F+1)(2F+1)} - \frac{F(-3p^2(1-2x)^2+p(2x-1)(6x-5)-2(x-2)x)+2(-2px+p+x)^2)}{(F+1)(2F+1)} \quad (7)$$

Допустим, выборка состоит из представителей m субпопуляций. Количество представителей каждой субпопуляции в группе больных обозначим как a_1, a_2,..,a_m, а в группе здоровых как b_1, b_2,..,b_m. Можно предположить, что ковариация будет наблюдаться только внутри субпопуляций, но не между представителями разных субпопуляций. Было показано, что:

$$Var(T) = NVar(G_i) + cov(G_i, G_j)\sum_k\{a_k(a_k-1) + b_k(b_k-1) - 2a_kb_k\} \quad (8)$$

Или

$$Var(T) = N\left(Var(G_i) - cov(G_i, G_j)\right) + cov(G_i, G_j)\sum(a_k - b_k)^2 \quad (9)$$

здесь $k = 1$,..,m.

Показатель инфляции дисперсии VIF введем как

$$VIF = \frac{Var(T)}{Var_0(T)} \quad (10)$$

где $Var(T)$ и $Var_0(T)$ определены формулами (8) и (2), соответственно.

При подстановке вместо x значений 0, 0.5, 1 были получены выражения, выведенные ранее [9].

Для сверхдоминирования значения были получены аналогично, за исключением весов генотипов, принятых 0, 1, 0 для *AA*, *AB*, *BB* соответственно:

$$Var(G_j)_{свдом} = 2(1-p)p(1-F-2(-1+F)^2p+2(-1+F)^2p^2)$$

$$cov(G_i, G_j)_{свдом}$$

$$= -\frac{4F(-1+p)p((1-2p)^2+2F^3(-1+p)p+F(-1-6(-1+p)p))}{(1+F)(1+2F)}$$

Выражения для $Var(T)$ и $Var_0(T)$ также определяются формулами (8) и (2), соответственно.

Используя выражение (10), можно получить значения для фактора инфляции дисперсии. В дальнейшем для простоты записи примем:

$$K = \sum_k \{a_k(a_k-1) + b_k(b_k-1) - 2a_kb_k\} \qquad (11)$$

Таким образом, VIF является функцией частоты аллеля (p), модели наследования (x) и популяционных параметров (N – общий объем, F - коэффициент инбридинга Райта, K - величина, определяющаяся по формуле (11), характеризующая общую структурированность выборки больных и здоровых; чем она выше, тем выше популяционная структурированность). Приведем пример того, как выглядит VIF при заданных параметрах популяции *F=0.05, N=1000, K=10000* (Рисунок 7А).

Как видно из графика, VIF не зависит от частоты аллеля только в случае, когда x=1/2 (Рисунок 7Б), то есть при аддитивной модели, что уже было показано ранее [6]. Стоит отметить, что при x, стремящемуся к бесконечности, VIF все больше приближается к сверхдоминантной модели наследования. Также стоит отметить, что график VIF зеркально симметричен относительно x=1/2, т.е. при x, стремящемуся к минус бесконечности, VIF также стремится к сверхдоминантной модели наследования.

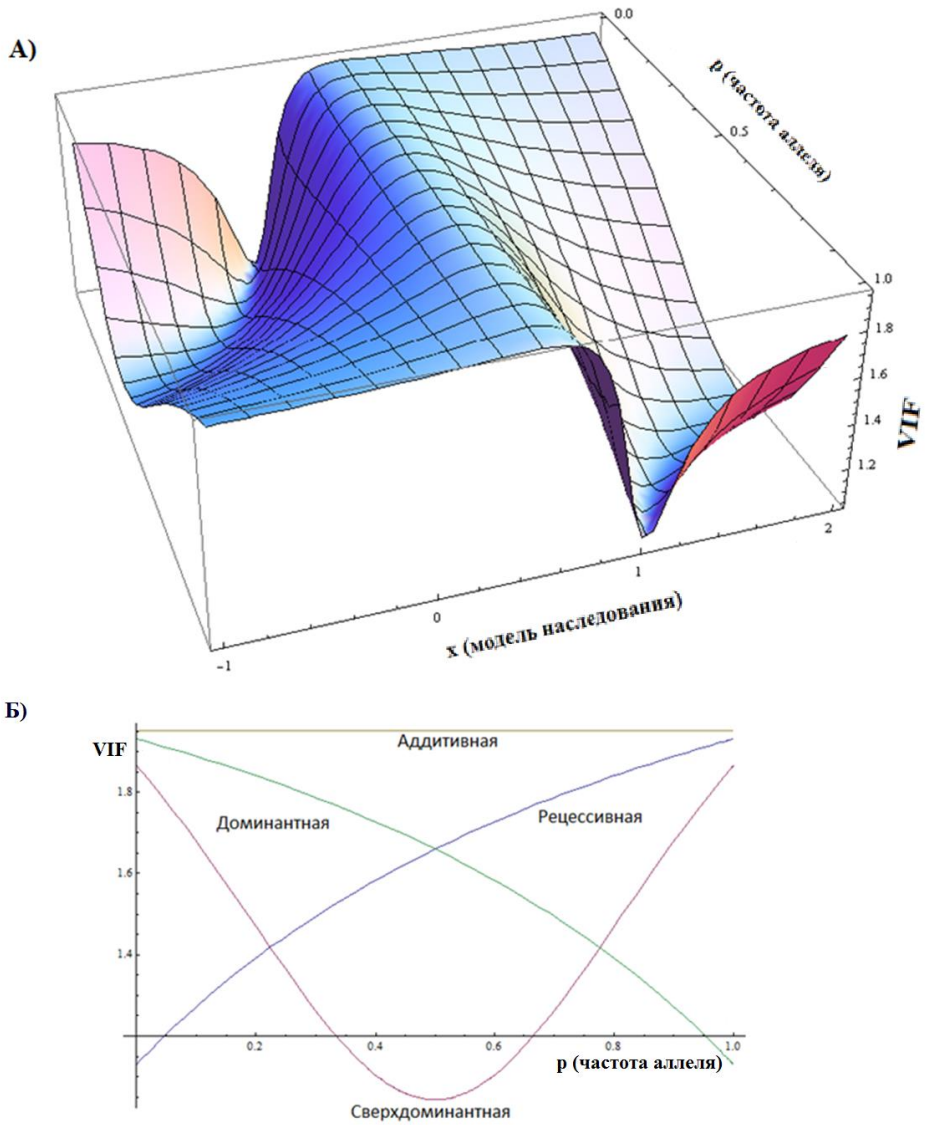


Рисунок 7. А) График VIF при значениях *F=0.05, N=1000 K=10000, x:{-1,2}, p{0,1}.* Б). Крайние случаи VIF при доминантной, рецессивной, аддитивной и сверхдоминантной моделях.

Важным замечанием является то, что для данного метода коррекции непринципиально использование статического теста Кохрана-Армитажа. Метод полностью применим и для анализа количественных признаков, для которого используют линейную регрессию признака на генотип [89].

Суммируя вышесказанное, мы определили VIF как функцию частоты аллеля (*p*), типа наследования (*x* указывает на эффект гетерозиготного генотипа; для рецессивной, аддитивной и доминантной модели *x* равно 0, 1/2

и 1 соответственно), размера выборки (N) и популяционных параметров, таких, как коэффициент инбридинга Райта F (в диапазоне от 0 до 1) и коэффициент K, описывающий субструктуру популяции. Сверхдоминантная модель (эффект генотипа равен 0 для гомозигот и 1 для гетерозиготы) описывалась отдельно.

В реальности, для большинства популяций человека коэффициент инбридинга F принимает значения < 0.01, однако может достигать значения 0.04 для популяций, где распространены близкородственные браки [93]. Значение K/N^2 стремится к 0, когда дизайн исследования сбалансирован (например, соотношение «случай:контроль» примерно 1:1 в каждой субпопуляции) и стремится к своему максимуму ½, когда субпопуляции представлены только случаями или только контролями.

VIF выражается как $\lambda = \frac{N\left(Var\left(G_i\right) - cov\left(G_i, G_j\right)\right) + K * cov\left(G_i, G_j\right)}{NVar(G_j)}$, где G_i – маркер генотипа i случая ($G_i \in \{0,1,2\}$). $Var(G_j)$ и $cov(G_i, G_j)$ определяются как:

$$Var(G_j) = \left[2p(1-p)(1-F)x^2 + Fp + (1-F)p^2\right] - [2p(1-p)(1-F)x + Fp + (1-F)p^2]^2$$ и

$$cov\left(G_i, G_j\right) = -\frac{2F(p-1)p(F^3(p-1)p\left(1-2x\right)^2 + F^2(p(8x-4)-4x+3)}{(F+1)(2F+1)} -$$

$$-\frac{F(-3p^2\left(1-2x\right)^2 + p(2x-1)(6x-5) - 2(x-2)x) + 2\left(-2px + p + x\right)^2)}{\left(F+1\right)\left(2F+1\right)}$$

соответственно.

Из анализа графика функции VIF можно сделать несколько выводов (Рисунок 7). Во-первых, VIF для аддитивной модели всегда больше, чем для неаддитивной, что также подтверждается при анализе симулированных данных нескорректированных тестов (Таблица 6). Во-вторых, применение простой коррекции с помощью константы к результатам, полученным для

неаддитивной модели ПГАА, способно контролировать «среднюю» ошибку первого рода на номинальном уровне; однако, для некоторых частотных групп маркеров тест будет консервативным, тогда как для других – либеральным. Например, для доминантной модели такая коррекция приведет к либерализации теста для SNP, для которых доминантно-кодированный аллель имел высокую частоту, и к консервативным результатам теста для SNP для которых доминантно-кодированный аллель имел низкую частоту (Рисунок 7Б). Эти результаты подтверждаются результатами константной коррекции в приложении к симулированным данным (Таблица 6). Коррекция при помощи константы, как правило, сохраняет уровень ошибки первого рода «в среднем» для всех маркеров, однако для конкретных групп частот маркеров это не соблюдается.

4.1.3 Оценка параметров VIF

Методы оценки VIF требуют знания параметров, описывающих генетическую структуру популяции. Если эти параметры (в нашем случае *F* и *K)* неизвестны, то можно использовать несколько подходов для их оценки. Наша оценка основывается на идее, что распределение тестовой статистики должно соответствовать $\chi^2_{df=1}$ после коррекции. Таким образом, оценка неизвестных функциональных параметров возможна за счет минимизации функции ошибки, описывающей отклонение наблюдаемого распределения от предполагаемого. Как функцию ошибки мы использовали сумму квадратов отклонений упорядоченной скорректированной статистики (Z^2) от теоретически ожидаемого распределения:

$$f_{err} = \sum_{i=1}^{M} \left(\frac{Z_i^2(x)}{VIF_i(p, x, N, F, K)} \right) - \chi_1^2)^2$$

Необходимо отметить, что только параметры популяции *F* и *K* должны подлежать оценке, тогда как *N* (объем выборки), *M* (число SNP), *p*

(частота аллеля) и *x* заданы данными и аналитической моделью. Этот метод был обозначен как VIFGC.

4.1.4 Полиномиальный ГК

Нами был предложен полиномиальный ГК (Polynomial Genomic Control, PGC) для неаддитивных моделей. PGC аппроксимирует функцию коррекции λ полиномом *l*-степени от частоты аллеля *p:*

$$\lambda(p) = \sum_{i=0}^{l} a_i * p^i$$

Для оценки коэффициентов a_i, мы использовали ту же идею, что и для оценки параметров *F* и *K* в методе VIFGC, а именно, что скорректированная статистика $Z^{2*} = Z^2 / \lambda(p)$ должна быть распределена как χ^2. Решение использовать полиномы третьей степени в процессе оптимизации принято эмпирическим путем.

Следует отметить, что для метода PGC мы можем, в принципе, использовать экспоненциальную функцию вместо полиномиальной, но использование экспоненциальной функции ограничивает доступные модели только рецессивной и доминантной. Использование полиномиальной функции снимает ограничения на использование PGC для других моделей, таких как сверх- и кодоминантная.

Предложенные методы коррекции, VIFGC и PGC, сравнивались со стандартным методом ГК – коррекцией тестовой статистики на константу λ. Для оценки ошибки первого рода и мощности методов использовались смоделированные и реальные данные. Ошибка первого рода характеризовалась тремя параметрами: λ_{median}, отношением медианы наблюдаемого распределения к ожидаемой медиане (0.455); $\lambda_{regress}$, коэффициентом регрессии между наблюдаемым распределением статистики и теоретически ожидаемым χ^2; и *E* - пропорцией тестов со значением p-value

≤ 0.05. Также нами проведено оценивание ошибки первого рода для пяти частотных групп аллелей: [0.05,0.25), [0.25,0.4), [0.4,0.6), [0.6,0.75), и [0.75,0.95].

4.1.5 Результаты моделирования

Процедура, использовавшаяся для моделирования, представлена в разделе «Материалы и методы». Результаты моделирования для ошибки первого рода ограниченных тестов представлены в Таблице 6. Как и ожидалось в соответствии с полученными нами теоретическими результатами, тесты, использующие коррекцию на константу, имеют значительное отклонение от ожидаемых значений ошибки первого рода для определенных групп частот аллелей для неаддитивных моделей. В отличие от метода коррекции при помощи константы, методы PGC и VIFGC имеют ошибку первого рода, близкую к ожидаемой как для всех SNP вместе, так и для всего спектра частотных групп.

Результаты симуляций для ошибки первого рода кодоминантной модели представлены в Таблице 7. Из таблицы видно, что коррекция при помощи константы приводит к либеральному тесту. Метод, основанный на ограниченных VIFGC-скорректированных тестах, является консервативным (ошибка первого рода ниже номинального уровня). В тоже время, PGC-коррекция приводит к тесту, имеющему ошибку первого рода, близкую к номинальной.

Мощность различных методов представлена в Таблице 8. Показано, что все методы для коррекции, включая VIFGC и PGC, дают оптимальные результаты, когда коррекционная модель (используемая для симуляции) используется также и для анализа. Как и ожидалось, кодоминантный тест имеет меньшую мощность, чем ограниченный тест использующий корректную моделью, но оказывается более устойчивым при неправильном выборе ограниченной модели наследования.

4.1.6 Апробация на реальных данных

Использование реальных данных по двум независимым когортам, KORA и ERF, предоставило возможность тестирования наших методов для ситуаций, возможно, не отраженных в наших модельных исследованиях. В обоих исследованиях мы анализировали импутированные генотипы (выраженные через оцененные вероятности) и количественные признаки с использованием методов линейной регрессии.

В исследовании KORA нами был проведен анализ уровня мочевой кислоты, а в исследовании ERF - уровней липопротеидов высокой плотности (ЛПВП) (см. раздел «Материалы и методы»). Необходимо отметить, что для полногеномного анализа в ERF, как правило используются методы, основанные на смешанных моделях [94]; здесь мы использовали ERF в качестве примера генетически высокоструктурированной популяции, и анализировали ее с использованием моделей с фиксированным эффектом.

Таблица 9 показывает результаты анализа ошибки первого рода в ERF, где влияние генетической структуры на результаты анализа ассоциаций велико. Если применялась аддитивная модель без использования коррекции генетической структуры смешанными моделями, λ для ЛПВП составляла 1.2. Для неаддитивных моделей мы воспроизвели те же основные результаты, которые были получены для моделированных данных: коррекция с помощью константы привела к консервативному тесту для некоторых частотных групп и либеральному для других, тогда как коррекции VIFGC и PGC давали ошибки первого рода, близкие к номинальной, независимо от частоты аллелей маркеров.

Результаты ошибки первого рода для KORA, популяционного исследования, где стратификация минимальна, представлены в Таблица 10. При использовании аддитивной модели для анализа уровня мочевой кислоты мы наблюдали значения λ, равные 1.03. Для неаддитивных моделей, выводы

совпадают с теми, что были получены в симуляциях и при анализе данных ERF.

4.1.7 Краткое заключение

В рамках этой работы мы расширили возможности метода ГК для неаддитивных моделей, что позволило нам использовать для ГК маркеры с произвольной частотой аллелей. Для рецессивных, доминантных и сверхдоминантных моделей наследования были получены аналитические выражения для фактора инфляции тестовой статистики. Инфляция зависит от частоты аллелей и нескольких популяционно-генетических параметров. Мы предложили метод для оценки требуемых параметров. Более того, нами предложен метод ГК, основанный на приближении коэффициента коррекции полиномиальной функцией частоты аллелей, и описана процедура коррекции кодоминантного теста для случаев, когда модель наследования неизвестна. Статистические характеристики описанных методов были исследованы с использованием моделированных и реальных данных. Мы продемонстрировали, что предложенные нами методы эффективны для контроля ошибки первого рода в присутствии генетической гетерогенности выборки. Предложенные методы ГК могут быть применены к статистическим тестам для ПГАА с различными моделями наследования. Все методы, разработанные и протестированные в данной работе, были воплощены с использованием языка R как часть пакета GenABEL.

Таблица 6. Ошибка первого рода для тестов с одной степенью свободы. Ошибка первого рода была оценена тремя способами: λ_{median} – отношение медианы полученного распределения статистики к медиане ожидаемого распределения; $\lambda_{regress}$ – коэффициент регрессии между полученным распределением и теоретически ожидаемым; E – доля тестов с p-value ≤ 0.05. Значения даны как для всех SNP, так и для определённых частотных групп.

		Некорр.			**Константная корр.**			**VIFGC корр.**			**PGC корр.**		
Модель	**Частоты**	λ_{median}	$\lambda_{regress}$	*E*	λ_{median}	$\lambda_{regress}$	*E*	λ_{median}	$\lambda_{regress}$	*E*	λ_{median}	$\lambda_{regress}$	*E*
Рецессивная	все	1.301	1.305	0.086	1.000	1.003	0.051	1.000	1.000	0.050	0.999	0.999	0.050
	[0.05,0.25)	1.175	1.170	0.069	0.905	0.900	0.038	0.990	0.983	0.048	1.004	0.998	0.049
	[0.25,0.4)	1.245	1.245	0.079	0.957	0.957	0.045	0.995	0.995	0.049	1.000	1.000	0.050
	[0.4,0.6)	1.320	1.322	0.088	1.014	1.015	0.052	1.002	1.004	0.051	0.998	0.999	0.050
	[0.6,0.75)	1.377	1.381	0.095	1.057	1.060	0.057	1.006	1.009	0.051	0.996	0.999	0.050
	[0.75,0.95]	1.412	1.416	0.100	1.084	1.087	0.060	1.007	1.010	0.051	0.997	1.000	0.050
Аддитивная	все	1.453	1.458	0.104	1.000	1.003	0.051	0.997	1.000	0.050	0.991	1.034	0.050
	[0.05,0.25)	1.451	1.455	0.104	0.998	1.001	0.050	0.995	0.998	0.050	0.991	1.033	0.050
	[0.25,0.4)	1.455	1.460	0.105	1.001	1.005	0.051	0.998	1.002	0.050	0.991	1.035	0.050
	[0.4,0.6)	1.456	1.461	0.105	1.002	1.006	0.051	0.999	1.002	0.051	0.990	1.035	0.050
	[0.6,0.75)	1.454	1.458	0.104	1.000	1.003	0.051	0.997	1.000	0.050	0.990	1.034	0.050
	[0.75,0.95]	1.452	1.456	0.104	0.999	1.002	0.051	0.996	0.998	0.050	0.992	1.036	0.050
Доминантная	все	1.302	1.306	0.086	1.000	1.003	0.051	0.999	1.000	0.050	0.999	1.000	0.050
	[0.05,0.25)	1.413	1.416	0.099	1.084	1.086	0.060	1.007	1.009	0.051	0.997	0.999	0.050
	[0.25,0.4)	1.379	1.383	0.095	1.058	1.061	0.057	1.007	1.010	0.051	0.997	1.000	0.050
	[0.4,0.6)	1.320	1.323	0.088	1.013	1.016	0.052	1.002	1.004	0.051	0.998	1.001	0.050
	[0.6,0.75)	1.244	1.245	0.079	0.956	0.956	0.045	0.993	0.993	0.049	1.000	1.000	0.050
	[0.75,0.95]	1.174	1.171	0.070	0.903	0.900	0.039	0.988	0.984	0.048	1.003	0.999	0.050
Сверхдоминантная	все	1.176	1.181	0.072	1.000	1.004	0.051	0.999	1.000	0.050	0.999	1.000	0.050
	[0.05,0.25)	1.281	1.282	0.083	1.088	1.089	0.061	1.007	1.008	0.051	0.996	0.997	0.050
	[0.25,0.4)	1.143	1.146	0.067	0.972	0.974	0.047	0.998	1.000	0.050	1.006	1.008	0.051
	[0.4,0.6)	1.060	1.058	0.057	0.902	0.901	0.039	0.987	0.985	0.048	0.991	0.990	0.049
	[0.6,0.75)	1.142	1.143	0.067	0.971	0.972	0.047	0.998	0.999	0.050	1.006	1.007	0.051
	[0.75,0.95]	1.279	1.282	0.083	1.086	1.089	0.060	1.007	1.008	0.051	0.996	0.997	0.050

Таблица 7. Ошибка первого рода для тестов с двумя степенями свободы. Обозначения по аналогии Таблице 6. Значения даны как для всех SNP, так и для определённых частотных групп.

	Некорр.			**Константная корр.**			**Основанный на корр. df1 (VIFGC)***			**PGC корр.**		
Частоты	λ_{median}	$\lambda_{regress}$	E	λ_{median}	$\lambda_{regress}$	E	λ_{median}	$\lambda_{regress}$	E	λ_{median}	$\lambda_{regress}$	E
все	1.239	1.250	0.092	1.000	1.009	0.053	0.951	0.957	0.045	0.991	1.000	0.051
[0.05,0.25)	1.239	1.248	0.091	1.000	1.007	0.052	0.959	0.962	0.045	0.992	1.000	0.051
[0.25,0.4)	1.241	1.252	0.092	1.001	1.010	0.053	0.948	0.955	0.045	0.991	1.000	0.051
[0.4,0.6)	1.240	1.252	0.092	1.001	1.010	0.053	0.942	0.951	0.044	0.990	1.000	0.051
[0.6,0.75)	1.239	1.251	0.092	1.000	1.009	0.053	0.946	0.955	0.045	0.990	1.000	0.052
[0.75,0.95]	1.239	1.249	0.092	1.000	1.008	0.053	0.959	0.963	0.045	0.992	1.000	0.051

*кодоминантный тест, основанный на скорректированных ограниченных тестах (здесь ограниченные тесты были скорректированы методом VIFGC)[7].

Таблица 8. Мощность (% тестов с p-value ≤ 0.05) для различных тестов. r, a, d, o и g – рецессивная, аддитивная, доминантная, сверхдоминантная и кодоминантная (генотипическая) модели, соответственно.

Симулированная модель	**Рецессивная**					**Аддитивная**					**Доминантная**					**Сверхдоминантная**				
Анализируемая модель	r	a	d	o	g	r	a	d	o	g	r	a	d	o	g	r	a	d	o	g
Некорр.	0.87	0.71	0.26	0.44	0.78	0.60	0.78	0.64	0.42	0.67	0.20	0.74	0.84	0.39	0.78	0.37	0.32	0.40	0.83	0.76
Константная корр.	0.79	0.59	0.15	0.43	0.64	0.48	0.67	0.58	0.38	0.62	0.15	0.63	0.80	0.35	0.72	0.31	0.26	0.33	0.78	0.60
VIFGC корр.*	0.80	0.59	0.16	0.41	0.62	0.50	0.66	0.55	0.38	0.58	0.15	0.63	0.80	0.35	0.68	0.30	0.26	0.32	0.77	0.57
PGC корр.	0.81	0.58	0.16	0.41	0.63	0.50	0.67	0.56	0.38	0.62	0.15	0.64	0.80	0.36	0.72	0.30	0.26	0.32	0.77	0.59

*Генотипическая модель для VIFGC скорректированных тестов – тест с двумя степенями свободы, основанный на рецессивном и доминантном тестах, скорректированных VIFGC [7].

Таблица 9. Ошибка первого рода для результатов анализа ЛПВП в данных ERF. Обозначения по аналогии Таблице 6. Значения даны как для всех SNP, так и для определённых частотных групп.

		Некорр.			**Константная корр.**			**VIFGC корр.**			**PGC корр.**		
Модель	**Частоты**	λ_{median}	$\lambda_{regress}$	E	λ_{median}	$\lambda_{regress}$	E	λ_{median}	$\lambda_{regress}$	E	λ_{median}	$\lambda_{regress}$	E
Рецессивная	все	1.201	1.200	0.074	1.000	1.000	0.050	1.001	1.000	0.050	1.000	1.000	0.050
	[0.05,0.25)	1.105	1.109	0.063	0.921	0.924	0.042	0.993	0.997	0.050	0.998	1.002	0.051
	[0.25,0.5)	1.188	1.180	0.072	0.989	0.983	0.048	1.005	0.999	0.050	0.998	0.992	0.050
	[0.5,0.75)	1.240	1.252	0.079	1.033	1.043	0.055	1.004	1.013	0.051	0.996	1.006	0.051
	[0.75,0.95]	1.273	1.261	0.081	1.061	1.051	0.056	1.001	0.991	0.049	1.008	0.999	0.050
Аддитивная	all	1.298	1.302	0.086	1.000	1.003	0.051	0.997	1.000	0.050	0.996	1.000	0.050
	[0.05,0.25)	1.299	1.290	0.085	1.001	0.994	0.050	0.998	0.991	0.049	1.008	1.000	0.050
	[0.25,0.5)	1.298	1.313	0.088	1.001	1.012	0.052	0.997	1.008	0.052	0.986	0.997	0.050
	[0.5,0.75)	1.301	1.320	0.088	1.003	1.017	0.053	1.000	1.014	0.052	0.989	1.003	0.051
	[0.75,0.95]	1.292	1.286	0.084	0.995	0.991	0.049	0.992	0.988	0.049	1.003	0.999	0.050
Доминантная	all	1.202	1.203	0.074	1.000	1.001	0.050	1.000	1.000	0.050	1.000	1.000	0.050
	[0.05,0.25)	1.277	1.265	0.081	1.063	1.053	0.056	1.001	0.991	0.049	1.008	0.999	0.050
	[0.25,0.5)	1.243	1.252	0.079	1.035	1.042	0.054	1.003	1.011	0.051	0.997	1.005	0.051
	[0.5,0.75)	1.190	1.183	0.072	0.990	0.985	0.049	1.005	1.000	0.050	0.999	0.994	0.050
	[0.75,0.95]	1.104	1.112	0.064	0.919	0.925	0.042	0.991	0.998	0.050	0.995	1.002	0.051
Сверхдоминантная	all	1.133	1.123	0.064	1.000	0.991	0.049	1.011	1.000	0.050	1.011	1.000	0.050
	[0.05,0.25)	1.203	1.193	0.072	1.061	1.053	0.056	1.017	1.010	0.051	1.011	1.003	0.050
	[0.25,0.5)	1.076	1.060	0.057	0.950	0.935	0.043	1.011	0.995	0.049	1.013	0.998	0.049
	[0.5,0.75)	1.064	1.053	0.056	0.939	0.929	0.042	1.000	0.989	0.049	1.004	0.993	0.049
	[0.75,0.95]	1.204	1.190	0.072	1.062	1.050	0.056	1.017	1.007	0.051	1.014	1.004	0.051
Кодоминантная (df=2)	all	1.157	1.162	0.077	1.000	1.004	0.052	0.964	0.966	0.046	0.996	1.000	0.051
	[0.05,0.25)	1.163	1.159	0.077	1.005	1.002	0.052	0.973	0.969	0.047	1.004	1.001	0.051
	[0.25,0.5)	1.152	1.165	0.077	0.996	1.006	0.052	0.954	0.964	0.045	0.988	0.999	0.051
	[0.5,0.75)	1.151	1.164	0.077	0.995	1.006	0.052	0.953	0.963	0.046	0.989	1.000	0.051
	[0.75,0.95]	1.163	1.159	0.077	1.005	1.001	0.052	0.973	0.970	0.047	1.004	1.000	0.052

*Кодоминантная модель для VIFGC скорректированных тестов – тест с двумя степенями свободы, основанный на рецессивном и доминантном тестах, скорректированных VIFGC [7].

Таблица 10. Ошибка первого рода для результатов анализа уровня мочевой кислоты в данных KORA. Обозначения по аналогии Таблице 6. Значения даны как для всех SNP, так и для определённых частотных групп.

		Некорр.			Константная корр.			VIFGC корр.			PGC корр.		
Модель	**Частоты**	λ_{median}	$\lambda_{regress}$	***E***	λ_{median}	$\lambda_{regress}$	***E***	λ_{median}	$\lambda_{regress}$	***E***	λ_{median}	$\lambda_{regress}$	***E***
Рецессивная	все	1.016	1.020	0.053	1.000	1.004	0.051	0.996	1.000	0.050	0.996	1.000	0.050
	[0.05,0.25)	1.015	1.016	0.052	0.998	1.000	0.050	1.004	1.005	0.051	1.000	1.001	0.050
	[0.25,0.5)	1.014	1.023	0.053	0.998	1.006	0.051	0.996	1.005	0.051	0.992	1.001	0.050
	[0.5,0.75)	1.017	1.021	0.053	1.001	1.005	0.051	0.993	0.997	0.050	0.993	0.997	0.050
	[0.75,0.95]	1.020	1.020	0.053	1.004	1.004	0.051	0.993	0.993	0.050	1.000	1.001	0.051
Аддитивная	all	1.019	1.024	0.053	1.000	1.005	0.051	0.995	1.000	0.050	0.995	1.000	0.050
	[0.05,0.25)	1.024	1.030	0.053	1.005	1.011	0.051	1.000	1.006	0.051	0.997	1.003	0.050
	[0.25,0.5)	1.020	1.024	0.053	1.001	1.004	0.051	0.996	0.999	0.050	0.994	0.998	0.050
	[0.5,0.75)	1.017	1.019	0.052	0.998	1.000	0.050	0.993	0.995	0.050	0.994	0.996	0.050
	[0.75,0.95]	1.016	1.024	0.053	0.997	1.005	0.051	0.992	1.000	0.050	0.995	1.003	0.051
Доминантная	all	1.021	1.025	0.053	1.000	1.004	0.051	0.996	1.000	0.050	0.996	1.000	0.050
	[0.05,0.25)	1.024	1.026	0.053	1.002	1.005	0.051	0.989	0.992	0.049	0.999	1.002	0.050
	[0.25,0.5)	1.025	1.028	0.054	1.004	1.007	0.051	0.995	0.999	0.050	0.994	0.998	0.050
	[0.5,0.75)	1.015	1.026	0.053	0.994	1.005	0.051	0.992	1.003	0.050	0.987	0.997	0.050
	[0.75,0.95]	1.022	1.021	0.053	1.000	1.000	0.050	1.007	1.007	0.051	1.003	1.002	0.050
Сверхдоминантная	all	1.027	1.027	0.053	1.000	1.000	0.050	1.000	1.000	0.050	1.000	1.000	0.050
	[0.05,0.25)	1.027	1.027	0.054	1.000	1.001	0.051	0.987	0.988	0.049	0.999	1.000	0.050
	[0.25,0.5)	1.029	1.027	0.053	1.003	1.000	0.050	1.016	1.013	0.051	1.003	1.001	0.050
	[0.5,0.75)	1.028	1.025	0.053	1.002	0.998	0.050	1.014	1.011	0.051	1.002	0.999	0.050
	[0.75,0.95]	1.021	1.028	0.054	0.995	1.001	0.051	0.982	0.988	0.049	0.994	1.000	0.051
Кодоминантная (df=2)	all	1.021	1.024	0.054	1.000	1.003	0.051	0.999	1.002	0.051	0.997	1.000	0.051
	[0.05,0.25)	1.021	1.022	0.054	1.000	1.001	0.051	0.997	0.998	0.050	1.000	1.001	0.051
	[0.25,0.5)	1.022	1.026	0.055	1.001	1.005	0.051	0.997	1.001	0.051	0.996	1.000	0.051
	[0.5,0.75)	1.020	1.024	0.054	0.999	1.003	0.051	0.996	0.999	0.050	0.994	0.998	0.050
	[0.75,0.95]	1.021	1.025	0.055	1.000	1.004	0.052	1.006	1.009	0.052	0.997	1.001	0.051

*Кодоминантная модель для VIFGC скорректированных тестов – тест с двумя степенями свободы, основанный на рецессивном и доминантном тестах, скорректированных VIFGC [7].

4.2 Неаддитивные эффекты генов на метаболоме человека

Развитие высокопроизводительных технологий измерения метаболитов (высокопроизводительная метаболомика, highthroughput metabolomics) обусловило возможность проведение ПГАА, который успешно определил локусы, влияющие на концентрации метаболитов, и биохимические пути, лежащие в их основе. В большинстве ПГАА предполагается, что эффект каждого локуса на фенотип является аддитивным. Другие генетические модели, такие как рецессивная, доминантная или сверхдоминантная, рассматривались лишь в небольшом числе исследований. В тоже время, существуют теории, которые постулируют неаддитивные эффекты как следствие фундаментальных свойств цепей биохимических реакций. Эти теории могут быть особенно релевантны для метаболитов, так как их концентрации напрямую контролируются биохимическими реакциями. Поэтому в данном исследовании мы систематически проанализировали неаддитивные эффекты генов на большой панели метаболитов с использованием всего комплекса разработанных методов. Исследовались концентрации 151 метаболита в сыворотке крови человека, а также все возможные отношения их концентраций (всего 22,801 признак). Исследование проводилось на материале крупного популяционного исследования (KORA F4, N=1,785).

4.2.1 Двухэтапный подход к идентификации неаддитивных эффектов

Нами предложен и апробирован двухшаговый подход к идентификации неаддитивных эффектов в рамках ПГАА. На первом шаге ПГАА проводился с помощью двухстепенного теста для кодоминантной модели, который не накладывает ограничение на модель наследования. Результаты корректировались с помощью разработанных нами методов ГК. На втором шаге, для каждого значимого локуса определенного на первом этапе, кодоминантная модель сравнивалась с ограниченными моделями (рецессивной, доминантной, аддитивной и сверхдоминантной) с использованием критерия отношения

правдоподобия (likelihood ratio test, LRT [95]). Ограниченная модель принималась в том случае, если она незначимо отличалась от модели кодоминантного контроля (p-value 0.05/20=0.0025 после коррекции Бонферрони). Если отвергались все ограниченные модели – то есть данные описывались ограниченными моделями значимо хуже – принималась кодоминантная (общая) модель. Если не отвергалась только одна ограниченная модель, такая модель принималась как наиболее парсимонная. В тоже время, четко определить наиболее парсимонную модель только с использованием LRT не всегда представлялось возможным. Поэтому как дополнительный критерий, позволяющий ранжировать модели, использовался информационный критерий Акаике (Akaike information criterion, AIC) [96], который позволяет сравнивать невложенные модели:

AIC = 2k – 2ln(L), где k — число параметров в статистической модели, и L — максимизированное значение функции правдоподобия модели. В соответствии с этим критерием, наилучшей считается модель с минимальным значением теста.

Следует отметить, что хотя отдельные этапы предложенного нами подхода и были известны ранее, однако, изложенный здесь подход для систематической идентификации неаддитивных эффектов в целом является новым.

4.2.2 Результаты анализа с использованием двухэтапного подхода

Мы провели ПГАА с использованием кодоминантной модели. Коэффициент инфляции λ для всех признаков варьировал от 1.00 до 1.03. Было обнаружено двадцать локусов, достоверно ассоциированных как минимум с одним метаболитом или отношением метаболитов. При коррекции уровня значимости на множественное тестирование применялся метод Бонферрони. Граница значимости была установлена как $5\times10^{-8}/22801\approx2.19\times10^{-12}$, где 5×10^{-8} – общепринятая граница значимости ПГАА при анализе одного признака, а 22801 – число проанализированных нами признаков.

Достоверно ассоциированные локусы представлены в Таблице 11. Шестнадцать локусов были реплицированы на данных TwinsUK (p-value, с

коррекцией Бонферрони, < 0.05/20=0.0025). Необходимо отметить, что SNP rs715 находился в сильном неравновесии по сцеплению с SNP rs7422339 (R^2=0.91), который был сильнее ассоциирован на данных KORA (p-value = 5.19×10^{-74}), однако не был генотипирован в исследовании TwinsUK. Таким образом, SNP rs715 был выбран основным для этого локуса и был успешно реплицирован. SNP rs6970485 не был генотипирован в TwinsUK. Определить прокси-SNP хорошего качества для него также не удалось, поэтому мы не смогли проанализировать этот SNP на этапе репликации.

Для каждого найденного локуса мы сравнили ограниченные модели с кодоминантной моделью с помощью LRT и выбрали наиболее парсимонную модель с помощью AIC. Результаты этого анализа представлены в Таблице 12. Для шестнадцати из идентифицированных локусов кодоминантная модель не отличалась значимо от аддитивной модели (указаны в Таблица 12 с «a» в столбике LRT для KORA), что можно интерпретировать как свидетельство аддитивности. При использовании AIC, наилучшей моделью для этих локусов оказалась аддитивная (10 локусов) или кодоминантная (4 локуса, представленные rs273913, rs174547, rs1077989, rs603424). Для двух локусов (rs4902242 и rs7200543) кодоминантная модель не была достоверно лучше (LRT) двух ограниченных моделей (аддитивная/рецессивная и аддитивная/доминантная, соответственно); AIC определил наилучшую модель как кодоминантную для rs4902242 и аддитивную для rs7200543.

Оставшиеся четыре локуса показали явные признаки неаддитивных эффектов. Для локуса rs715 кодоминантная модель отличалась от рецессивной модели недостоверно (однако кодоминантная модель была наилучшей по AIC). Это наблюдение было реплицировано на данных TwinsUK. Для двух локусов (rs2066938 и rs7601356) кодоминантная модель была достоверно лучше, чем все ограниченные модели, что подтвердилось также на материале TwinsUK. Для SNP rs6970485 кодоминантная модель была недостоверно лучше, чем доминантная модель (LRT); тест AIC также указал на доминантную модель как наиболее

подходящую. Однако, эти результаты не удалось повторить в исследовании TwinsUK, так как указанный SNP не был ни генотипирован, ни импутирован.

4.2.3 Поиск локусов с использованием ограниченных моделей

Анализ с использованием аддитивной модели позволил определить двадцать локусов, которые были обнаружены кодоминантной моделью (частично представленные другими SNP и отношениями метаболитов) и два дополнительных локуса (представленных rs477992 и rs1374804). Для обоих SNP значение p-value для кодоминантной модели было чуть ниже порогового (p-value для rs477992: 6.43×10^{-12}; p-value для rs1374804: 1.43×10^{-11}). Новый локус rs1374804 не удалось реплицировать в исследовании TwinsUK.

Далее мы провели ПГАА для рецессивной и доминантной моделей (Приложение 3). Даже с использованием либерального уровня значимости ($5\times10^{-8}/22801$) вместо строгого ($5\times10^{-8}/(22801\times4)$), мы не смогли обнаружить дополнительные локусы. Из 20 локусов, определенных с помощью кодоминантной модели, четырнадцать были обнаружены с помощью рецессивной модели и восемнадцать – доминантной. Использование сверхдоминантной модели выявило десять из 20 описанных локусов и одну дополнительную ассоциацию между rs219040 на седьмой хромосоме (p-value $<3\ .94\times10^{-13}$) и отношением C5.1/C6.1. Локус располагался вблизи гена *STEAP2-AS1* (кодирующего антисмысловую РНК гена RNA1), биологическую роль которого нельзя напрямую соотнести с контролем метаболизма. Его p-value для HWE было близко к пороговому для контроля качетсва (p-value $< 1.03\times10^{-05}$), и его не удалось реплицировать на данных TwinsUK (p-value = 0.8).

4.2.4 Сравнение с предыдущими опубликованными результатами ПГАА

Мы сравнили наши результаты с предыдущим исследованием с использованием аддитивной модели на данных того же исследования [71]. Только локус rs477992 (в данном исследовании) / rs541503 (в предыдущем исследовании) не удалось обнаружить, ввиду того, что его p-value было чуть ниже порогового (p-

value=3.88×10^{-11}). Некоторые локусы, идентифицированные в данном исследовании, были представлены другими SNP в предыдущем. Мы повторили анализ для кодоминантной модели именно для тех SNP, которые были опубликованы для аддитивного ПГАА ранее [71] (Таблица 13). Только один SNP - rs11158519 – не был достоверно ассоциирован с ранее выявленными метаболитами или их отношением. Этот локус был идентифицирован в ходе анализа кодоминантной моделью с другим найденным значимым SNP на расстоянии 134 тпн от искомого и для другого отношения концентраций фосфатидилхолинов. Оставшиеся 13 SNP показали достоверное значение p-value как для кодоминантной, так и для аддитивной моделей (p-value < 2.19×10^{-12}). Возможно, полученные различия объясняются разницей в контроле качества генотипов и признаков между исследованиями, хотя в обоих анализах использовались данные одних и тех же индивидуумов. Для восемнадцати из двадцати локусов, которые были определены в обоих анализах аддитивной и кодоминантной моделью, была выявлена одна и та же пара SNP - метаболит. Необходимо отметить, что при использовании аддитивной модели мы смогли реплицировать локус rs1894832, который не удавалось реплицировать при использовании кодоминантной модели.

4.2.5 Новые локусы с аддитивными эффектами

В результате проведенных исследований мы идентифицировали и реплицировали пять новых локусов (представленных SNP rs1466448, rs7200543, rs2657879, rs5746636 и rs1894832), которые не были найдены ранее [71]. По данным LRT и AIC, наилучшей моделью для этих SNP в нашем анализе была аддитивная.

Локус, включающий SNP rs1466448, расположен в области гена *CERS4*, который кодирует фермент церамидсинтазу, вовлеченный в биосинтез церамидов (простой формы сфинголипидов, состоящих из сфингозина или некоторых его производных, и жирной кислоты). Мы обнаружили ассоциацию с отношением

концентраций двух сфингомиелинов SM.C18.1 и SM.C16.1, что, предположительно, связано с деятельностью гена *CERS4*.

Локус, включающий rs7200543, расположен в области гена *NTAN1*, кодирующего белок N-терминальную аспарагинамидазу. Этот локус был ассоциирован с отношением концентраций фосфатидилхолинов PC.aa.C36.2 и PC.aa.C38.3. На первый взгляд, прямой связи между функцией гена и ассоциированным признаком нет.

Два локуса (rs2657879 и rs5746636) расположены вблизи генов, участвующих в метаболизме аминокислот - *GLS2* (кодирующего фермент глутаминазу) и *PRODH (*кодирующего пролиндегидрогеназу), соответственно. В нашем исследовании rs2657879 был ассоциирован с отношением концентраций гистидина и глутамина, а rs5746636 ассоциирован с отношением концентраций лейцина (и изолейцина) к пролину.

Локус rs1894832, который был реплицирован только для аддитивной модели, расположен рядом с генами *PSPH* и *PHKG1* (которые кодируют фосфосеринфосфотазу и фосфорилаз киназу, соответственно). Мы не можем сказать, какой ген является функциональным без проведения дополнительного анализа, но оба гена могут быть связаны с ассоциированным признаком (отношением серина и триптофана).

4.2.6 Локусы с неаддитивными эффектами

Для визуализации генетических эффектов найденных локусов на признаки мы использовали так называемые диаграммы размаха (иногда также называемые «ящик с усами»), на которых наглядно показана разница между аддитивными и неаддитивными моделями (см. Рисунок 6 и Рисунок 8).

Если принять среднее концентрации признака для всех образцов с гомозиготным генотипом по эффекторному аллелю за 100%, а среднее для всех образцов с гомозиготным генотипом по другому аллелю за 0%; то для этой шкалы среднее признака для всех образцов с гетерозиготным генотипом следует ожидать

около 50% для аддитивной модели. Для рецессивной (доминантной) модели ожидаемые оценки будут 0% (100%).

Для двух локусов (rs2066938 и rs7601356) кодоминантная модель оказалась достоверно лучше, чем любая из ограниченных одностепенных. Диаграммы размаха для этих двух SNP (Рисунок 8А) показывают, что лежащая в основе генетическая модель не может быть отнесена ни к рецессивной, ни к доминантной, а занимает некоторое «среднее» положение между аддитивной и доминантной/рецессивной. В терминах процентной шкалы, для rs2066938 и rs7601356 эффект гетерозиготного генотипа был 67% и 22%, соответственно.

Для локуса rs715 кодоминантная модель незначимо отличалась от рецессивной, однако согласно AIC наилучшей (парсимонной) моделью являлась кодоминантная. Для этого SNP среднее признака для гетерозиготного генотипа было 82% (Рисунок 8В).

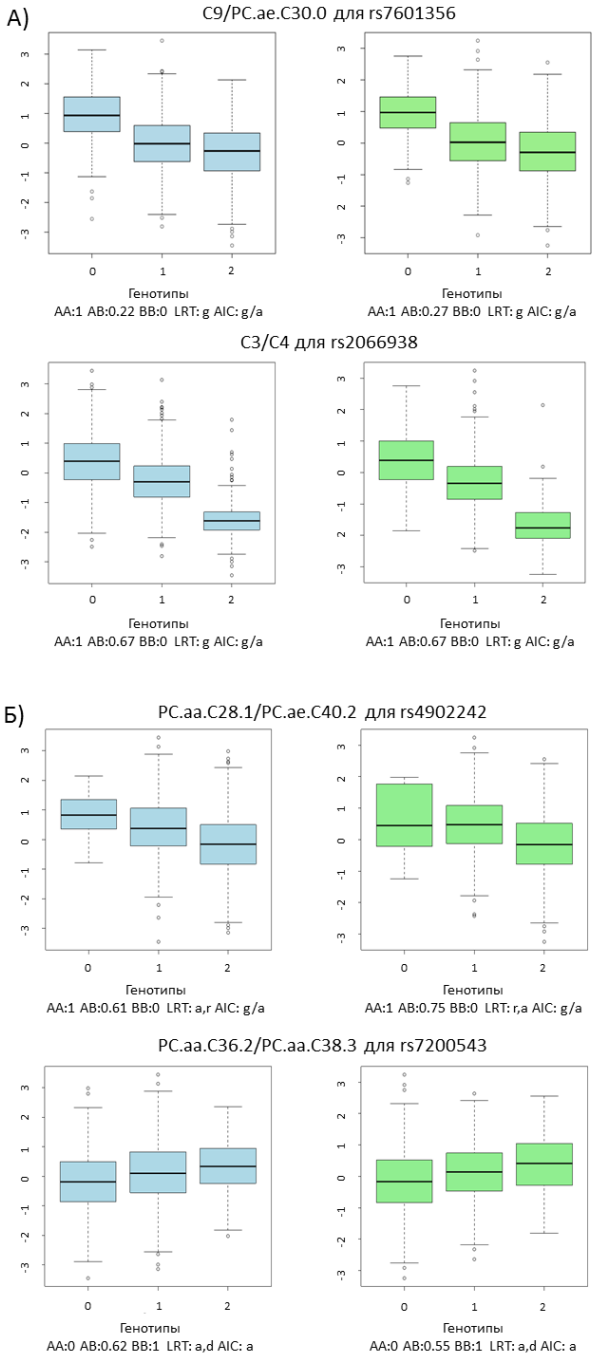
Для SNP rs6970485 доминантная модель оказалась лучшей (по данным AIC и LRT), диаграммы представлены на Рисунке 8Г. Для этого SNP среднее для гетерозиготного генотипа было 89%.

Для двух локусов (rs4902242 и rs7200543), для которых кодоминантная модель незначимо отличалась от двух ограниченных моделей, одной из которых была аддитивная, среднее значение признака у гетерозигот было 61% и 62%, соответственно (Рисунок 8Б).

4.2.7 Краткое заключение

Мы провели ПГАА уровней метаболитов крови человека с использованием кодоминантной, аддитивной, доминантной, рецессивной и сверхдоминантной моделей. Нами было выявлено 23 локуса, значимо ассоциированных (p-value с коррекцией Бонферрони $\leq 2.19\times10^{-12}$) как минимум с одним метаболитом или отношением концентраций. Для четырех из них мы показали наличие неаддитивных эффектов. Семнадцать локусов (включая 3 локуса с неаддитивными эффектами) были реплицированы на данных независимого

исследования (TwinsUK, N=846). Таким образом, мы обнаружили, что большинство генетических эффектов на концентрации метаболитов и их отношения являются аддитивными, что подтверждает практику использования аддитивной модели для анализа эффектов SNP на концентрации метаболитов.



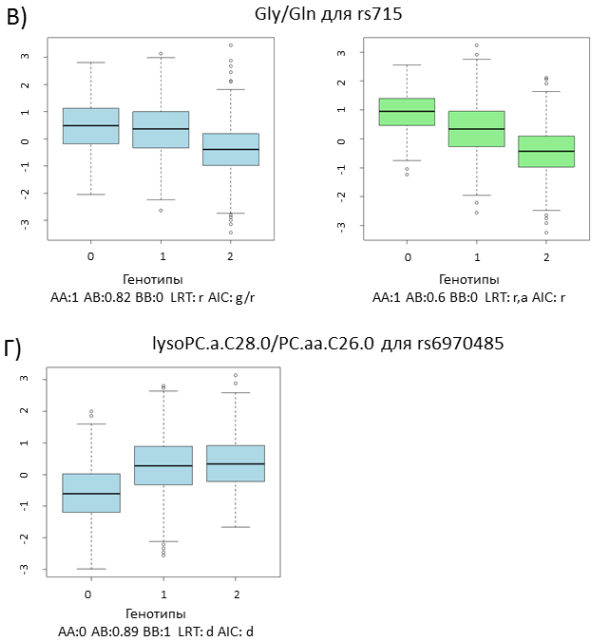


Рисунок 8. Диаграммы размаха для локусов с неаддитивными эффектами. Показано распределение фенотипов при условии различных генотипов. Жирная линия является медианой признака в группе генотипа. Столбцы данных KORA представлены синим цветом, TwinsUK – зеленым. Под каждой диаграммой приведена дополнительная информация. Во-первых, среднее значение концентрации (отношения концентраций) метаболитов у гетерозигот выражено как процент от разницы между средними значениями признака для двух гомозигот (принятых за 0% и 100%). Во-вторых, представлена информация о результатах тестов LRT и AIC. Для LRT представлены все одностепенные модели, для которых кодоминантная модель не показала значимо более хороших результатов (в порядке уменьшения значения p-value). Если кодоминантная модель оказывалась достоверно лучше, чем все ограниченные модели по результатам LRT, это отмечено как «g». Для AIC указана самая хорошая (парсимонная) модель. В случаях, когда самой хорошей является кодоминантная модель, после косой черты приведена наилучшая одностепенная модель. Обозначения для ограниченных одностепенных моделей: r, a, d, o, g – рецессивная, аддитивная, доминантная, сверхдоминантная и кодоминантная, соответственно.

А) диаграммы для rs7601356 and rs2066938. Для этих локусов наилучшей моделью была кодоминантная.

Б) Диаграммы для rs4902242 и rs7200543. Для этих локусов две ограниченные модели незначимо отличались от кодоминантной модели.

В) Диаграммы для rs715. Для этого локуса наилучшей моделью была рецессивная.

Г) Диаграммы для rs6970485. Для этого локуса наилучшей моделью была доминантная. Этот SNP был недоступен в TwinsUK.

Таблица 11. Результаты ПГАА концентраций (отношений концентраций) метаболитов крови человека с использованием кодоминантной модели. В таблице представлены результаты для двадцати локусов, найденных при анализе кодоминантной модели на данных KORA (p-value < 2.19×10^{-12}). Шестнадцать локусов были реплицированы на данных TwinsUK (p-value < 0.0025), они отмечены звездочкой в последнем столбце. Таблица разделена на 2 части: в верхней части представлены локусы, которые уже описаны в ранее опубликованных ПГАА по тем же данным (Illig et al. 2010), в нижней части приведены новые локусы.

					KORA		TwinsUK		
SNP	**Метаболит (отношение)**	**Хр.**	**Позиция**	**Ген**	**AF**	**p-value**	**AF**	**p-value**	**Репл.**
rs7552404	C12/C10	1	75,908,534	*ACADM*	0.300	1.69E-72	0.314	1.89E-29	*
rs7601356	C9/PC.ae.C30.0	2	210,764,902	*ACADL*	0.632	1.24E-70	0.649	6.86E-28	*
rs715	Gly/Gln	2	211,251,300	*CPS1*	0.687	4.28E-69	0.703	1.12E-48	*
rs8396	C7.DC/C10	4	159,850,267	*ETFDH*	0.707	5.98E-26	0.678	3.14E-17	*
rs2046813	PC.ae.C42.5/PC.ae.C44.5	4	186,006,153	*ACSL1*	0.688	6.29E-17	0.687	1.18E-03	*
rs273913	C5/PC.ae.C34.1	5	131,689,055	*SLC22A4*	0.405	1.60E-16	0.351	4.19E-02	
rs3798719	PC.aa.C42.5/PC.aa.C40.3	6	11,144,811	*ELOVL2*	0.248	5.01E-32	0.234	4.01E-04	*
rs12356193	C0	10	61,083,359	*SLC16A9*	0.166	2.18E-27	0.161	1.20E-07	*
rs603424	C16.1/C14	10	102,065,469	*PKD2L1*	0.801	3.70E-18	0.818	1.99E-02	
rs174547	PC.aa.C36.3/PC.aa.C36.4	11	61,327,359	*FADS1*	0.701	2.29E-208	0.649	2.09E-76	*
rs2066938	C3/C4	12	119,644,998	*ACADS*	0.270	1.73E-159	0.257	2.17E-67	*
rs4902242	PC.aa.C28.1/PC.ae.C40.2	14	63,299,842	*SYNE2*	0.849	2.00E-35	0.872	4.78E-15	*
rs1077989	PC.ae.C32.1/PC.ae.C34.1	14	67,045,575	*PLEKHH1*	0.463	6.80E-42	0.472	4.05E-18	*
rs4814176	SM.OH.C24.1/SM.OH.C22.1	20	12,907,398	*SPTLC3*	0.364	2.69E-31	0.416	9.69E-09	*
rs6970485	lysoPC.a.C28.0/PC.aa.C26.0	7	11,752,704	*THSD7A*	0.354	1.21E-47	-	-	
rs1894832	Ser/Trp	7	56,144,740	*LOC389493*	0.508	1.98E-12	0.511	4.02E-03	
rs2657879	His/Gln	12	55,151,605	*GLS2*	0.207	2.89E-14	0.186	1.90E-06	*
rs7200543	PC.aa.C36.2/PC.aa.C38.3	16	15,037,471	*NTAN1*	0.312	7.45E-16	0.277	1.66E-06	*
rs1466448	SM.C18.1/SM.C16.1	19	8,195,519	*CERS4*	0.222	7.01E-16	0.194	3.90E-10	*
rs5746636	xLeu/Pro	22	17,276,301	*DGCR6*	0.236	2.98E-20	0.273	2.40E-03	*

Хр.: Хромосома; AF – частота эффекторного аллеля.

Таблица 12. Результаты анализа неаддитивных генетических эффектов на концентрации (отношения концентраций) метаболитов крови человека. В столбце LRT представлены все ограниченные модели, которые недостоверно отличались от кодоминантной модели (в порядке уменьшения значения p-value). Кодоминантная модель представлена в этой колонке, если она была достоверно лучше, чем все ограниченные модели. В столбце AIC показана самая парсимонная модель. Если самая парсимонная модель – кодоминантная, то она отделена косой чертой от следующей самой хорошей ограниченной модели. Обозначения для ограниченных одностепенных моделей: r, a, d, o, g – рецессивная, аддитивная, доминантная, сверхдоминантная и кодоминантная, соответственно.

				кодоминантная модель			аддитивная модель		KORA		TwinsUK		
SNP	**метаболит/отношение**	**A/a**	**AF**	**beta(*AA*) (se)**	**beta(*Aa*) (se)**	**p-value**	**beta (se)**	**p-value**	**LRT**	**AIC**	**LRT**	**AIC**	**Репл.**
rs7552404	C12/C10	G/A	0.30/0.31	1.150(0.076)	0.642(0.046)	1.69E-72	0.600(0.033)	1.84E-73	a	a	a	a	*
rs7601356	C9/PC.ae.C30.0	C/T	0.63/0.65	-1.225(0.068)	-0.961(0.068)	1.24E-70	-0.523(0.032)	1.57E-58	g	g/d	g	g/a	*
rs715	Gly/Gln	C/T	0.69/0.70	-0.937(0.082)	-0.139(0.084)	4.28E-69	-0.590(0.036)	1.00E-61	r	g/r	r,a	r	*
rs8396	C7.DC/C10	C/T	0.71/0.68	-0.760(0.086)	-0.360(0.087)	5.98E-26	-0.388(0.036)	4.54E-27	a	a	r,a	g/r	*
rs2046813	PC.ae.C42.5/PC.ae.C44.5	C/T	0.69/0.69	0.647(0.084)	0.360(0.084)	6.29E-17	0.309(0.036)	7.11E-18	a	a	a,r,d	a	*
rs273913	C5/PC.ae.C34.1	T/C	0.41/0.35	0.606(0.071)	0.203(0.052)	1.60E-16	0.283(0.034)	1.28E-16	a	g/a	a,d,r,o	a	
rs3798719	PC.aa.C42.5/PC.aa.C40.3	T/C	0.25/0.23	-0.983(0.100)	-0.418(0.048)	5.01E-32	-0.453(0.038)	6.78E-33	a	a	a,d,r	a	*
rs12356193	C0	G/A	0.17/0.16	-1.145(0.167)	-0.483(0.053)	2.18E-27	-0.507(0.046)	2.42E-28	a	a	a,d	a	*
rs603424	C16.1/C14	A/G	0.80/0.82	0.872(0.121)	0.558(0.124)	3.70E-18	0.362(0.041)	1.42E-18	a	g/a	a,r,o,d	a	
rs174547	PC.aa.C36.3/PC.aa.C36.4	C/T	0.70/0.65	-1.872(0.068)	-1.019(0.069)	2.29E-208	-0.904(0.029)	5.98E-209	a	g/a	g	g/a	*
rs2066938	C3/C4	G/A	0.27/0.26	-1.942(0.077)	-0.649(0.043)	1.73E-159	-0.832(0.032)	7.49E-149	g	g/a	g	g/a	*
rs4902242	PC.aa.C28.1/PC.ae.C40.2	C/T	0.85/0.87	-1.019(0.181)	-0.356(0.189)	2.00E-35	-0.619(0.049)	3.77E-36	a,r	g/a	r,a	r	*
rs1077989	PC.ae.C32.1/PC.ae.C34.1	C/A	0.46/0.47	-0.888(0.066)	-0.529(0.056)	6.80E-42	-0.450(0.033)	1.99E-42	a	g/a	a,d	g/a	*
rs4814176	SM.OH.C24.1/SM.OH.C22.1	T/C	0.36/0.42	0.792(0.074)	0.422(0.050)	2.69E-31	0.403(0.034)	2.00E-32	a	a	d,a	d	*
rs6970485	lysoPC.a.C28.0/PC.aa.C26.0	C/T	0.35/-	0.980(0.107)	0.877(0.063)	1.21E-47	0.635(0.047)	4.66E-41	d	d	-	-	
rs1894832	Ser/Trp	C/T	0.51/0.51	0.491(0.067)	0.302(0.061)	1.98E-12	0.245(0.034)	4.02E-13	a	a	a,r,d	a	
rs2657879	His/Gln	G/A	0.21/0.19	0.747(0.128)	0.303(0.050)	2.89E-14	0.328(0.042)	4.44E-15	a	a	d,o,a	d	*
rs7200543	PC.aa.C36.2/PC.aa.C38.3	G/A	0.31/0.28	0.531(0.078)	0.328(0.049)	7.45E-16	0.287(0.035)	1.40E-16	a,d	a	a,d	a	*
rs1466448	SM.C18.1/SM.C16.1	C/A	0.22/0.19	-0.775(0.120)	-0.305(0.049)	7.01E-16	-0.337(0.041)	1.31E-16	a	a	a,d	a	*
rs5746636	xLeu/Pro	T/G	0.24/0.27	0.705(0.103)	0.379(0.049)	2.98E-20	0.366(0.039)	2.64E-21	a	a	a,r,d	a	*

A – эффекторный аллель; a – референсный аллель; AF – частота эффекторного аллеля в KORA / в TwinsUK; beta(se) – оценка эффекта (ее стандартная ошибка) для SNP; beta(*AA*) (se) и beta(*Aa*) (se) – оценка эффекта (ее стандартная ошибка) для генотипов *AA* и *Aa* в двухстепенной модели.

Таблица 13. Результаты анализа локусов, опубликованных ранее (Illig et al. 2010). В таблице приведено обобщение результатов пятнадцати ранее найденных SNP и соответствующих признаков (Illig et al. 2010). Результаты представлены для аддитивной и кодоминантной модели для этих локусов. Также для каждого локуса представлены найденные в этой работе наилучшие SNP и метаболит (или отношение концентраций метаболитов), и дистанция до SNP, описанного в (Illig et al. 2010). Один локус не был обнаружен в ходе анализа кодоминантной моделью.

Наилучший SNP опубликов. В (Illig et al. 2010)	Наилучший метаболит/отношение, опублик. в (Illig et al. 2010)	Хр.	Позиция	A/a	p-value генотипич.	p-value аддитивн.	AF (A)	Ген	Наилучший SNP в этой работе	Наилучший метаболит/отнош.	Дист. В пн
rs211718	C12/C10	1	75,879,263	T/C	1.9E-72	1.6E-73	0.3	*ACADM*	rs7552404	C12/C10	29,271
rs541503	Orn/Ser	1	120,009,820	C/T	3.9E-11	5.0E-11	0.63	*PHGDH*	-	-	-
rs2286963	C9/C10.2	2	210,768,295	G/T	1.3E-65	5.8E-60	0.63	*ACADL*	rs7601356	C9/PC.ae.C30.0	3,393
rs2216405	Gly/PC.ae.C38.2	2	211,325,139	G/A	5.6E-31	3.3E-31	0.19	*CPS1*	rs715	Gly/Gln	73,839
rs8396	C14.1.OH/C10	4	159,850,267	C/T	3.4E-23	3.9E-24	0.71	*ETFDH*	rs8396	C7.DC/C10	0
rs2046813	PC.ae.C44.5/PC.ae.C42.5	4	186,006,153	C/T	8.3E-17	9.8E-18	0.69	*ACSL1*	rs2046813	PC.ae.C42.5/PC.ae.C44.5	0
rs272889	Val/C5	5	131,693,277	A/G	4.4E-15	5.1E-16	0.62	*SLC22A4*	rs273913	C5/PC.ae.C34.1	4,222
rs9393903	PC.aa.C40.3/PC.aa.C42.5	6	11,150,895	A/G	6.9E-32	1.0E-32	0.75	*ELOVL2*	rs3798719	PC.aa.C42.5/PC.aa.C40.3	6,084
rs7094971	C0	10	61,119,570	G/A	3.7E-21	3.2E-22	0.15	*SLC16A9*	rs12356193	C0	36,211
rs603424	C14/C16.1	10	102,065,469	A/G	5.4E-18	1.6E-18	0.8	*SCD*	rs603424	C16.1/C14	0
rs174547	PC.aa.C36.3/PC.aa.C36.4	11	61,327,359	C/T	6.3E-209	1.5E-209	0.7	*FADS1*	rs174547	PC.aa.C36.3/PC.aa.C36.4	0
rs2014355	C3/C4	12	119,659,907	C/T	8.9E-121	1.7E-110	0.73	*ACADS*	rs2066938	C3/C4	14,909
rs11158519	PC.ae.C38.1/PC.aa.C28.1	14	63,434,338	A/G	4.4E-06	8.0E-07	0.86	*SYNE2*	rs4902242	PC.aa.C28.1/PC.ae.C40.2	134,496
rs7156144	PC.ae.C32.1/PC.ae.C34.1	14	67,049,466	A/G	2.1E-31	4.8E-32	0.59	*PLEKHH1*	rs1077989	PC.ae.C32.1/PC.ae.C34.1	3,891
rs168622	SM.OH.C24.1/SM.C16.0	20	12,914,089	T/G	3.5E-21	3.1E-22	0.36	*SPTLC3*	rs4814176	SM.OH.C24.1/SM.OH.C22.1	6,691

Хр.: хромосома; A – эффекторный аллель; a – референсный аллель; AF – частота аллеля эффекторного аллеля.

Таблица 14. Результаты ПГАА (отношений) концентраций метаболитов крови человека с использованием аддитивной модели. В таблице представлены результаты по двадцати двум локусам, достоверно ассоциированным в исследовании KORA (p-value < 2.19×10^{-12}) для аддитивной модели. Восемнадцать локусов были реплицированы (p-value < 0.0023). Для каждого локуса приведены SNP и метаболит (отношение метаболитов) с наименьшим p-value (в исследовании KORA). Таблица разделена на 2 части: в верхней части представлены локусы, которые были описаны ранее (Illig et al. 2010), в нижней части приведены новые локусы. Реплицированные SNP отмечены звездочкой.

					KORA		TwinsUK		информ. о гене		
SNP	**метаболит (отнош.)**	**Хр.**	**Позиция**	**A/a**	**AF**	**p-value аддитивное**	**AF**	**p-value аддитивное**	**Ген**	**Позиция отн. гена**	
rs7552404	C12/C10	1	75,908,534	G/A	0.300	1.84E-73	0.314	1.25E-30	*ACADM*	Ниже	*
rs477992	Ser/Orn	1	120,059,099	A/G	0.702	5.75E-13	0.658	8.11E-09	*PHGDH*	В гене	*
rs2286963	C9/PC.ae.C34.0	2	210,768,295	G/T	0.629	1.68E-60	0.651	2.14E-31	*ACADL*	В гене	*
rs715	Gly/Gln	2	211,251,300	C/T	0.687	1.00E-61	0.703	4.37E-49	*CPS1*	Выше	*
rs8396	C7.DC/C10	4	159,850,267	C/T	0.707	4.54E-27	0.678	2.72E-17	*ETFDH*	В гене	*
rs2046813	PC.ae.C42.5/PC.ae.C44.5	4	186,006,153	C/T	0.688	7.11E-18	0.687	2.71E-04	*ACSL1*	Выше	*
rs270613	C5/Val	5	131,668,482	A/G	0.606	8.03E-17	0.647	2.57E-03	*SLC22A4*	В гене	
rs3798719	PC.aa.C42.5/PC.aa.C40.3	6	11,144,811	T/C	0.248	6.78E-33	0.234	1.32E-04	*ELOVL2*	В гене	*
rs12356193	C0	10	61,083,359	G/A	0.166	2.42E-28	0.161	1.70E-08	*SLC16A9*	В гене	*
rs603424	C16.1/C14	10	102,065,469	A/G	0.801	1.42E-18	0.818	5.36E-03	*PKD2L1*	В гене	
rs174547	PC.aa.C36.3/PC.aa.C36.4	11	61,327,359	C/T	0.701	5.98E-209	0.649	1.00E-74	*FADS1*	В гене	*
rs2066938	C3/C4	12	119,644,998	G/A	0.270	7.49E-149	0.257	1.59E-63	*ACADS*	Ниже	*
rs4902242	PC.aa.C28.1/PC.ae.C40.2	14	63,299,842	C/T	0.849	3.77E-36	0.872	1.30E-15	*SYNE2*	Выше	*
rs1077989	PC.ae.C32.1/PC.ae.C34.1	14	67,045,575	C/A	0.463	1.99E-42	0.472	5.18E-18	*PLEKHH1*	Ниже	*
rs4814176	SM.OH.C24.1/SM.OH.C22.1	20	12,907,398	T/C	0.364	2.00E-32	0.416	7.67E-08	*SPTLC3*	Ниже	*
rs1374804	Ser/Gly	3	127,391,188	A/G	0.638	1.70E-12	0.619	1.03E-01	*ALDH1L1*	Выше	
rs6970485	lysoPC.a.C28.0/PC.aa.C26.0	7	11,752,704	C/T	0.354	4.66E-41	-	-	*THSD7A*	В гене	
rs1894832	Ser/Trp	7	56,144,740	C/T	0.508	4.02E-13	0.511	8.92E-04	*LOC389493*	Ниже	*
rs2657879	His/Gln	12	55,151,605	G/A	0.207	4.44E-15	0.186	1.57E-06	*GLS2*	В гене	*
rs7200543	PC.aa.C36.2/PC.aa.C38.3	16	15,037,471	G/A	0.312	1.40E-16	0.277	2.55E-07	*NTAN1*	Ниже	*
rs1466448	SM.C18.1/SM.C16.1	19	8,195,519	C/A	0.222	1.31E-16	0.194	5.12E-11	*CERS4*	В гене	*
rs5746636	xLeu/Pro	22	17,276,301	T/G	0.236	2.64E-21	0.273	8.94E-04	*DGCR6*	В гене	*

Хр.: хромосома; A – эффекторный аллель; AF – частота аллеля эффективного аллеля

5 Обсуждение

5.1 Методы геномного контроля для неаддитивных моделей наследования

Геномный контроль (ГК) занимает важное место в методологии полногеномного анализа ассоциаций. Во-первых, коэффициент геномного контроля λ может использоваться для коррекции завышенной тестовой статистики, что позволяет контролировать ошибку первого рода исследования. Во-вторых, λ выступает в роли важного индикатора адекватности используемой модели для анализируемых данных. Если инфляция тестовой статистики относительно большая, то, как правило, это свидетельствует о том, что конкретный метод, примененный для ПГАА, не может достаточно полно моделировать структуру изучаемой выборки. В таком случае модель анализа должна быть пересмотрена; например, вместо линейной регрессии с использованием фиксированных эффектов, необходимо использование таких методов, как стратифицированный анализ [97], анализ с поправкой на основные компоненты геномного родства EIGENSTRAT [98] или анализ с использованием смешанных моделей (mixed models) [99–101]. Стоит отметить, что даже после применения адекватных и сложных моделей анализа, небольшая остаточная инфляция тестовой статистики может по-прежнему присутствовать. Эта остаточная инфляция обычно и корректируется ГК, т.к. даже незначительная инфляция может привести к значительному повышению доли ложноположительных результатов в ПГАА. Так, например, при использовании стандартного порога p-value $< 5\times10^{-8}$ (что соответствует экспериментальной ошибке первого рода $\approx 5\%$), в случае если инфляция $\lambda=1.05$ и ГК не проводится, ошибка первого рода повышается приблизительно в два раза. В случае проведения мета-анализа результатов ПГАА, что является стандартным методом при

проведении больших полногеномных исследований, инфляция тестовой статистики может амплифицироваться [102].

В этой работе мы предложили новые и изучили существующие методы ГК неаддитивных моделей наследования. При помощи моделирования и на реальных данных мы показали, что VIFGC и PGC могут использоваться для коррекции результатов ПГАА, полученных при использовании неаддитивных моделей. Рассмотренные модели включали в себя ограниченные (доминантная, рецессивная, сверхдоминантная) и кодоминантную. В наших примерах, включавших анализ реальных фенотипов, ГК в высокоструктурированной выборке ERF использовался, в основном, в качестве индикатора. Несмотря на то, что номинальная ошибка первого рода может быть достигнута при использовании предложенных методов ГК, ПГАА для таких популяций все равно должен проводиться при помощи методов смешанных моделей, а ГК должен использоваться только для коррекции остаточной инфляции. Анализ данных KORA, которые являются тщательно спланированным популяционным исследованием с незначительной стратификацией, предоставляет более реалистичный пример случая, когда метод ГК может использоваться и как индикатор, и как единственный метод коррекции.

Мы показали, что в целом, для ограниченных моделей VIFGC и PGC действуют одинаково хорошо, тогда как для кодоминантной модели статистически более оптимальным является PGC, в то время как тест, основанный на использовании VIFGC-скорректированных ограниченных статистик, консервативен. Для всех частотных групп генетических маркеров предложенные нами методы имеют меньший разброс ошибки первого рода по сравнению с ГК с использованием константы (p-value < 10^{-20}, см. Приложение 1, Табл. S1).

Высокое качество коррекции результатов неаддитивных ПГАА сопряжено с повышенной сложностью моделей и временем вычислений. В то время как коррекция результатов аддитивной модели очень проста с вычислительной точки зрения, коррекции VIFGC и PGC включают шаги по оптимизации параметров функции. Однако даже оптимизация четырех параметров PGC по 1.5 миллионам маркеров требует всего нескольких минут вычислений на обычном персональном компьютере. Поскольку шаг коррекции проводится всего один раз для одного раунда ПГАА и само вычисление статистики ПГАА зачастую значительно более трудоемко, можно сказать, что дополнительная вычислительная трудоемкость, налагаемая предложенными нами методами коррекции, мала.

В наших методах ГК мы оцениваем параметры модели, минимизируя регрессионную функцию потери. Эта функция потери чувствительна к наличию больших значений статистики, которые могут присутствовать, например, в случае сильных сигналов ассоциаций. Мы минимизируем негативные эффекты больших значений статистик за счет использования нижних 95% распределения тестов, что похоже на метод, предложенный ранее [103]. Еще одним решением может быть использование других функций потерь, которые менее чувствительны к большим отклонениям наблюдаемой статистики от ожидаемой. Например, можно сформулировать функцию потерь, определенную как сумма абсолютных отклонений полученного распределения от ожидаемого, или – что обусловит еще меньшую чувствительность к большим отклонениям – суммой квадратных корней абсолютных отклонений.

Как уже говорилось выше, для аддитивных моделей фактор инфляции λ является важным показателем адекватности модели ПГАА, использованной при анализе конкретной выборки. Для рецессивной, доминантной и сверхдоминантной моделей мы показали, что для одних и тех же исходных данных инфляция тестовой статистики всегда меньше, чем инфляция для

аддитивной модели (Рисунок 6). Основываясь на этом результате, мы предлагаем использование коэффициента ГК λ аддитивной модели как показатель адекватности использованного метода ПГАА, в том числе и для неаддитивных моделей. Этот результат является теоретически и практически важным.

Мы проводили аналитическое исследование метода ГК предполагая, что исследуемый признак – бинарный, а используемая тестовая статистика является тестом множителей Лагранжа (score test). Однако ранее было показано, что полученные результаты можно обобщить также для случая исследования количественного признака с использованием линейной регрессии [89]. Это положение подтверждено нами эмпирически в исследовании статистических свойств методов ГК на реальных данных, представленных рядом количественных признаков в двух различных популяциях человека.

Все методы, рассмотренные в этой работе, были реализованы с использованием языка R в пакете GenABEL [90], который является частью проекта GenABEL для статистической геномики (http://www.genabel.org).

5.2 ПГАА с использованием неаддитивных моделей

Разработка и реализация методов ГК неаддитивных моделей ПГАА являлась необходимым условием для проведения систематического исследования неаддитивных генетических эффектов на концентрации метаболитов сыворотки крови человека. Однако, помимо этого, необходимо было выбрать стратегию ПГАА.

Для ПГАА без ограничения на модель наследования помимо анализа кодоминантной модели, существуют такие методы, как MERT (Maximin Efficiency Robust Tests) [104] и MAX (max test) [105], основанные на комбинации тестов для ограниченных одностепенных моделей с выведением их эмпирического распределения тестовой статистики. Ранее было проведено

сравнение этих методов между собой и с кодоминантным тестом [106,107]. Все три стратегии выигрывают в эффективности при разных условиях. Мы решили использовать кодоминантный тест по нескольким причинам: наличие программного обеспечения, быстрота вычислений, отсутствие ограничений на частоты аллелей [107], применимость к количественным признакам и импутированным данным. Кодоминантный тест также эффективен, если лежащая в основе модель – сверхдоминантная [107,108].

Предложенная нами стратегия идентификации неаддитивных эффектов с использованием кодоминантной модели может быть использована и в других анализах, особенно когда известно, что значительная часть ожидаемых генетических эффектов – неаддитивна [106,107]. По сравнению с наилучшей ограниченной моделью, кодоминантная модель может иметь меньшую мощность выявления ассоциированных локусов из-за большего числа степеней свободы. Тем не менее, она может использоваться как поисковый инструмент, который снижает число множественных тестирований и является более эффективной в плане времени, затрачиваемого на вычисления, если сравнивать с анализом каждой из ограниченных моделей по отдельности (см. Приложение 3). Для фенотипических данных с высокой размерностью, таких как метаболомные, это особенно важно.

5.3 Поиск неаддитивных эффектов генов на концентрации метаболитов сыворотки крови человека

Разработка и воплощение методов ГК неаддитивных моделей ПГАА, а формирование стратегии проведения такого ПГАА, позволило нам систематически исследовать неаддитивные генетические эффекты на концентрации метаболитов сыворотки крови человека. В контексте исследования доминантности, концентрации метаболитов занимают особое место. Физиологическая теория доминантности С. Райта (Wright S., 1929), в

дальнейшем развитая в работах Х. Касисера и Дж. Бернса (Kacser H., Burns J. A., 1981), постулирует, что неаддитивнось генетических эффектов может являться следствием фундаментальных свойств цепей биохимических реакций. Эта теория может быть особенно релевантна для метаболитов, так как их концентрации напрямую контролируются биохимическими реакциями.

При проведении неаддитивного ПГАА нами использовались данные двух больших независимых исследований: KORA F4 и TwinsUK. Следует отметить, что использование большого числа образцов в выборке является еще более критичным при исследовании неаддитивных эффектов, чем при анализе аддитивной модели. Статистическая мощность такого исследования для конкретного SNP зависит от наличия в выборке образцов с каждым из трех возможных генотипов. Малое число образцов с одним конкретным генотипом снизит возможность выявления потенциального неаддитивного эффекта. В нашем анализе мы исключили SNP, для которых число образцов с редким генотипом было менее 30.

Для шестнадцати локусов, найденных в нашем анализе, была принята аддитивная модель эффекта на ассоциированные метаболиты. Четыре локуса обладали значимыми неаддитивными эффектами. Из них два (rs6970485, rs715) следовали доминантной генетической модели. Тот факт, что для двух других локусов (rs2066938, rs7601356) наилучшей моделью была кодоминантная, не доказывает, что возможный функциональный вариант не является доминантным – смещение модели в сторону аддитивности может быть обусловлено такими причинами как слабое неравновесие по сцеплению между маркерным и функциональным аллелями. Однако кодоминантная модель сама по себе является неаддитивной, и кодоминантный механизм контроля укладывается в рамки физиологической теории доминирования Райта. Мы продемонстрировали устойчивость полученной нами неаддитивной модели к различным трансформациям фенотипа – таких как

логарифмическая и обратно-нормальная трансформация рангов (см. Приложение 2, Таблица S1). Полученные нами результаты проливают свет на общую картину генетического контроля метаболитов, а так же уточняют характер наследования определенных локусов.

Применимость аддитивных моделей для ПГАА «омиксных» признаков

Наблюдение, что большинство обнаруженных нами локусов незначимо отклонялось от аддитивной модели подтверждает постулат о том, что большая часть генетически обусловленной вариативности признаков контролируется аддитивно [60,61]. Ранее было показано, что наблюдение преимущественно аддитивных генетических эффектов может быть следствием сильного искажения модели в сторону аддитивности в случае неполного неравновесия по сцеплению (LD) между функциональным и маркерным аллелями [109,110]; при этом, чем ниже LD, тем больше будет приближение к аддитивности. Мы провели дополнительные модельные эксперименты для того, чтобы оценить влияние ошибки измерения генотипа (LD) или фенотипа на генетическую модель, и получили результаты, согласующиеся с предыдущими исследованиями (см. Приложение 4). Таким образом, в исследованиях, использующих генетические данные с более высоким геномным покрытием (например, данные геномного секвенирования), мы можем ожидать более выраженные неаддитивные эффекты. Необходимо отметить, что мы предполагали наличие в исследуемом локусе только одного биаллельного функционального варианта, аллели которого идентичны по происхождению. Ситуации, при которых в локусе может наблюдаться аллельная гетерогенность или множественный аллелизм, здесь не рассматривались.

Мы наблюдали, что даже SNP с неаддитивными эффектами могут быть обнаружены при использовании аддитивной модели. Аддитивная модель может рассматриваться как приближение к рецессивной и доминантной

моделям, которая обладает адекватной мощностью в случае превалирования рецессивного аллеля. Более того, не все локусы, которые были определены с помощью аддитивной модели, могли быть определены с помощью кодоминантной модели. Это еще больше подчеркивает, что предположение об аддитивности генетических эффектов при ПГАА концентраций метаболитов и их отношений приемлемо.

Мы можем предположить, что и для других «омиксных» фенотипов (транскриптомных, гликомных, протеомных и т.д.) генетический контроль, в основном, тоже осуществляется аддитивно. Логично предположить, что если теории С. Райта и Касисера и Бернса [13,65] верны, то «омики», наиболее близкие к биохимическим системам – такие как метаболомика – будут иметь высокую степень неаддитивности. Но, как было показано в этой работе, для метаболомики (и ранее для транскриптомики [61]), это не так. Поэтому для других «омик», скорее всего, стоит ожидать еще меньшую степень неаддитивности. Однако, этот вопрос требует дополнительных исследований.

6 Заключение

В настоящий момент в большинстве исследований по полногеномному анализу ассоциаций (ПГАА) используется аддитивная модель наследования признака, в рамках которой предполагается, что вклад каждого аллеля независим от вклада других аллелей и прочих факторов. Отсутствие различных неаддитивных эффектов в используемых моделях может быть одной из причин, по которым нам пока не удается объяснить наследуемость в широком смысле. Малое число работ, использующих модели, отличные от аддитивной, связано в том числе и с недостаточно проработанной методологией ПГАА с использованием таких моделей. Таким образом, разработка новых полногеномных методов для анализа неаддитивных эффектов и применение этих методов для анализа реальных данных являются актуальной проблемой современной статистической генетики и геномики.

Данная работа посвящена разработке, апробации и применению методов ПГАА с использованием неаддитивных моделей наследования. Были разработаны методы, повышающие статистическую корректность ПГАА с использованием неаддитивных моделей, а также стратегии полногеномного анализа неаддитивных эффектов. С использованием разработанных методов и стратегий был проведен систематический поиск неаддитивных эффектов генов на концентрации метаболитов сыворотки крови человека. Полученные новые результаты проливают свет на общую картину генетического контроля метаболитов, а также уточняют характер наследования определенных локусов.

В соответствии с поставленными задачами, были достигнуты следующие результаты:

1. Получены аналитические выражения для коэффициента коррекции тестовой статистики (геномного контроля) как функции частоты аллеля, модели наследования и популяционно-генетических параметров.
2. Разработаны алгоритмы и методы оценки параметров коэффициента коррекции (VIFGC и PGC) геномного контроля.
3. Предложенные подходы протестированы на моделированных и реальных данных и реализованы в виде программного продукта.
4. Предложена и отработана методика двухшагового поиска неаддитивных эффектов на основе кодоминантного теста.
5. Проведен полногеномный анализ ассоциаций концентраций метаболитов сыворотки крови человека для поиска неаддитивных эффектов генов.

7 Выводы

На основе проделанной работы можно сделать следующие выводы:

1. Показано, что при использовании неаддитивных моделей наследования для ПГАА коррекцию полученных статистических результатов можно проводить с использованием разработанных нами методов геномного контроля.
2. Продемонстрирована эффективность предложенного двухшагового подхода идентификации локусов с неаддитивными эффектами.
3. Выявлено, что эффективность поиска неаддитивных локусов зависит от плотности генетических маркеров.
4. Показано, что генетический контроль уровней метаболитов сыворотки крови человека осуществляется с помощью как аддитивных, так и неаддитивных внутрилокусных эффектов.

8 Список литературы

1. Polychronakos C., Alriyami M. Diabetes in the post-GWAS era. // Nat. Genet. 2015. Vol. 47, № 12. P. 1373–1374.
2. Reitz C. Genetic loci associated with Alzheimer's disease. // Future Neurol. 2014. Vol. 9, № 2. P. 119–122.
3. Kochi Y., Suzuki A., Yamamoto K. Genetic basis of rheumatoid arthritis: a current review. // Biochem. Biophys. Res. Commun. 2014. Vol. 452, № 2. P. 254–262.
4. Lango Allen H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. // Nature. Nature Publishing Group, 2010. Vol. 467, № 7317. P. 832–838.
5. Devlin B., Roeder K. Genomic control for association studies. // Biometrics. 1999. Vol. 55, № 4. P. 997–1004.
6. Zheng G. et al. Genomic control for association studies under various genetic models. // Biometrics. 2005. Vol. 61, № 1. P. 186–192.
7. Zheng G., Freidlin B., Gastwirth J.L. Robust genomic control for association studies. // Am. J. Hum. Genet. 2006. Vol. 78, № 2. P. 350–356.
8. Zang Y. et al. Robust genomic control and robust delta centralization tests for case-control association studies. // Hum. Hered. 2007. Vol. 63, № 3-4. P. 187–195.
9. Yan T., Hou B., Yang Y. Correcting for cryptic relatedness by a regression-based genomic control method. // BMC Genet. 2009. Vol. 10. P. 78.
10. Devlin B., Roeder K., Wasserman L. Genomic control, a new approach to genetic-based association studies. // Theor. Popul. Biol. 2001. Vol. 60, № 3. P. 155–166.

11. Orr H.A. A test of Fisher’s theory of dominance. // Proc. Natl. Acad. Sci. U. S. A. 1991. Vol. 88, № 24. P. 11413–11415.

12. Haldane J.B.S. A note on Fisher’s theory of the origin of dominance and a correlation between dominance and linkage. // Am. Nat. 1930. Vol. 64. P. 87–90.

13. Wright S. Fisher’s theory of dominance. // Am. Nat. 1929. Vol. 63. P. 274–279.

14. Fisher R. The possible modification of the response of the wild type to recurrent mutations. // Am. Nat. 1928. Vol. 62. P. 115–126.

15. Покровский et al. Инфекционные болезни и эпидемиология. 3rd ed. Москва, 2013. 1008 p.

16. Morton N.E., Chung C.S. Genetic Epidemiology. New York: Academic Press, 1978.

17. Аксенович Т.И. Статистические методы гентического анализа признаков человека. Новосибирск: Новосибирский Государственный Университет, 2003. 160 p.

18. Olefsky J.M., Kolterman O.G. Mechanisms of insulin resistance in obesity and noninsulin-dependent (type II) diabetes. // Am. J. Med. 1981. Vol. 70, № 1. P. 151–168.

19. Аульченко Ю.С., Аксенович Т.И. МЕТОДОЛОГИЧЕСКИЕ ПОДХОДЫ И СТРАТЕГИИ КАРТИРОВАНИЯ ГЕНОВ, КОНТРОЛИРУЮЩИХ КОМПЛЕКСНЫЕ ПРИЗНАКИ ЧЕЛОВЕКА // Вестник ВОГиС. 2006. Vol. 10, № 1.

20. Thompson E.A. Linkage analysis // Handb. Stat. Genet. / ed. D.J., Al B. et. John Wiley, Sons, Ltd., 2001. P. 541–563.

21. Holmans P. Nonparametric Linkage // Handb. Stat. Genet. / ed. Al B. et. John Wiley, Sons, Ltd, 2001. P. 487–505.

22. WATERWORTH D. Analysis of Human Genetic Linkage . By J. Ott. Baltimore, London: Johns Hopkins University Press. 1999 (3rd edition). Pp. 382. £38.00. // Ann. Hum. Genet. 2000. Vol. 64, № 1. P. 89–92.

23. Klein R.J. et al. Complement factor H polymorphism in age-related macular degeneration. // Science. 2005. Vol. 308, № 5720. P. 385–389.

24. Welter D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations // Nucleic Acids Res. 2014. Vol. 42, № D1.

25. Gibson G. Hints of hidden heritability in GWAS. // Nat. Genet. 2010. Vol. 42, № 7. P. 558–560.

26. HOMMEL G. A stagewise rejective multiple test procedure based on a modified Bonferroni test // Biometrika. 1988. Vol. 75, № 2. P. 383–386.

27. Pe'er I. et al. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants // Genet. Epidemiol. 2008. Vol. 32, № 4. P. 381–385.

28. Egger M., Smith G.D. Meta-Analysis. Potentials and promise. // BMJ. 1997. Vol. 315, № 7119. P. 1371–1374.

29. Maher B. Personal genomes: The case of the missing heritability. // Nature. 2008. Vol. 456, № 7218. P. 18–21.

30. Clarke G.M. et al. Basic statistical analysis in genetic case-control studies. // Nat. Protoc. 2011. Vol. 6, № 2. P. 121–133.

31. Ferrari R. et al. A genome-wide screening and SNPs-to-genes approach to identify novel genetic risk factors associated with frontotemporal dementia. // Neurobiol. Aging. 2015.

32. Chen P.-L. et al. Genetic determinants of antithyroid drug-induced agranulocytosis by human leukocyte antigen genotyping and genome-wide association study. // Nat. Commun. 2015. Vol. 6. P. 7633.

33. Zheng W., Rao S. Knowledge-based analysis of genetic associations of rheumatoid arthritis to inform studies searching for pleiotropic genes: a literature review and network analysis. // Arthritis Res. Ther. 2015. Vol. 17. P. 202.

34. Hu Y. et al. A Pooling Genome-Wide Association Study Combining a Pathway Analysis for Typical Sporadic Parkinson's Disease in the Han Population of Chinese Mainland. // Mol. Neurobiol. 2015.

35. Edwards A.O. et al. Complement factor H polymorphism and age-related macular degeneration. // Science. 2005. Vol. 308, № 5720. P. 421–424.

36. Haines J.L. et al. Complement factor H variant increases the risk of age-related macular degeneration. // Science. 2005. Vol. 308, № 5720. P. 419–421.

37. Teslovich T.M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. // Nature. Nature Publishing Group, 2010. Vol. 466, № 7307. P. 707–713.

38. Marucci A. et al. GALNT2 Expression Is Reduced in Patients with Type 2 Diabetes: Possible Role of Hyperglycemia // PLoS One. 2013. Vol. 8, № 7.

39. Dunn J.S. et al. Examination of PPP1R3B as a candidate gene for the type 2 diabetes and MODY loci on chromosome 8p23 // Ann. Hum. Genet. 2006. Vol. 70, № 5. P. 587–593.

40. Waterworth D.M. et al. Genetic variants influencing circulating lipid levels and risk of coronary artery disease // Arterioscler. Thromb. Vasc. Biol. 2010. Vol. 30, № 11. P. 2264–2276.

41. Sanna S. et al. Common variants in the GDF5-UQCC region are associated with variation in human height. // Nat. Genet. Nature Publishing Group, 2008. Vol. 40, № 2. P. 198–203.

42. Weedon M.N. et al. A common variant of HMGA2 is associated with adult

and childhood height in the general population. // Nat. Genet. 2007. Vol. 39, № 10. P. 1245–1250.

43. Weedon M.N. et al. Genome-wide association analysis identifies 20 loci that influence adult height. // Nat. Genet. 2008. Vol. 40, № 5. P. 575–583.

44. Carty C.L. et al. Genome-wide association study of body height in African Americans: the Women's Health Initiative SNP Health Association Resource (SHARe). // Hum. Mol. Genet. Oxford University Press, 2012. Vol. 21, № 3. P. 711–720.

45. Estrada K. et al. A genome-wide association study of northwestern Europeans involves the C-type natriuretic peptide signaling pathway in the etiology of human height variation. // Hum. Mol. Genet. 2009. Vol. 18, № 18. P. 3516–3524.

46. Lettre G. et al. Identification of ten loci associated with height highlights new biological pathways in human growth. // Nat. Genet. 2008. Vol. 40, № 5. P. 584–591.

47. Liu J.Z. et al. Genome-wide association study of height and body mass index in Australian twin families. // Twin Res. Hum. Genet. 2010. Vol. 13, № 2. P. 179–193.

48. Yang J. et al. FTO genotype is associated with phenotypic variability of body mass index. // Nature. 2012. Vol. 490, № 7419. P. 267–272.

49. Struchalin M. V et al. An R package "VariABEL" for genome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity. // BMC Genet. BioMed Central Ltd, 2012. Vol. 13, № 1. P. 4.

50. Tönjes A. et al. Genetic variation in GPR133 is associated with height: genome wide association study in the self-contained population of Sorbs. // Hum. Mol. Genet. 2009. Vol. 18, № 23. P. 4662–4668.

51. Aulchenko Y.S. et al. Predicting human height by Victorian and genomic

methods. // Eur. J. Hum. Genet. 2009. Vol. 17, № 8. P. 1070–1075.

52. Hofman A. et al. The Rotterdam Study: objectives and design update. // Eur. J. Epidemiol. 2007. Vol. 22, № 11. P. 819–829.

53. Pardo L.M. et al. The effect of genetic drift in a young genetically isolated population. // Ann. Hum. Genet. 2005. Vol. 69, № Pt 3. P. 288–295.

54. Yang J. et al. Common SNPs explain a large proportion of the heritability for human height. // Nat. Genet. 2010. Vol. 42, № 7. P. 565–569.

55. Franke A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. // Nat. Genet. 2010. Vol. 42, № 12. P. 1118–1125.

56. Manolio T. a et al. Finding the missing heritability of complex diseases. // Nature. 2009. Vol. 461, № 7265. P. 747–753.

57. Weiss L.A. et al. Association between Microdeletion and Microduplication at 16p11.2 and Autism // N. Engl. J. Med. 2008. Vol. 358, № 7. P. 667–675.

58. Stefansson H. et al. Large recurrent microdeletions associated with schizophrenia. // Nature. 2008. Vol. 455, № 7210. P. 232–236.

59. Zuk O. et al. The mystery of missing heritability: Genetic interactions create phantom heritability. // Proc. Natl. Acad. Sci. U. S. A. 2012. Vol. 109, № 4. P. 1193–1198.

60. Hill W.G., Goddard M.E., Visscher P.M. Data and theory point to mainly additive genetic variance for complex traits. // PLoS Genet. 2008. Vol. 4, № 2. P. e1000008.

61. Powell J.E. et al. Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. // PLoS Genet. 2013. Vol. 9, № 5. P. e1003502.

62. Shen X. et al. Inheritance beyond plain heritability: variance-controlling

genes in Arabidopsis thaliana. // PLoS Genet. 2012. Vol. 8, № 8. P. e1002839.

63. Pritchard J.K., Stephens M., Donnelly P. Inference of population structure using multilocus genotype data. // Genetics. Genetics Soc America, 2000. Vol. 155, № 2. P. 945–959.

64. Fisher R. Two further notes on the origin of dominance. // Am. Nat. 1928. Vol. 62. P. 571–574.

65. Kacser H., Burns J.A. The molecular basis of dominance. // Genetics. 1981. № 97. P. 639–666.

66. Porteous J.W. Dominance--one hundred and fifteen years after Mendel's paper. // J. Theor. Biol. 1996. Vol. 182, № 3. P. 223–232.

67. Gieger C. et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. // PLoS Genet. 2008. Vol. 4, № 11. P. e1000282.

68. Tanaka T. et al. Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. // PLoS Genet. 2009. Vol. 5, № 1. P. e1000338.

69. Kolz M. et al. Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. // PLoS Genet. 2009. Vol. 5, № 6. P. e1000504.

70. Hicks A.A. et al. Genetic determinants of circulating sphingolipid concentrations in European populations. // PLoS Genet. 2009. Vol. 5, № 10. P. e1000672.

71. Illig T. et al. A genome-wide perspective of genetic variation in human metabolism. // Nat. Genet. Nature Publishing Group, 2010. Vol. 42, № 2. P. 137–141.

72. Nicholson G. et al. A genome-wide metabolic QTL analysis in Europeans

implicates two loci shaped by recent positive selection. // PLoS Genet. 2011. Vol. 7, № 9. P. e1002270.

73. Suhre K. et al. A genome-wide association study of metabolic traits in human urine. // Nat. Genet. 2011. Vol. 43, № 6. P. 565–569.

74. Demirkan A. et al. Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. // PLoS Genet. 2012. Vol. 8, № 2. P. e1002490.

75. Kettunen J. et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. // Nat. Genet. 2012. Vol. 44, № 3. P. 269–276.

76. Suhre K., Gieger C. Genetic variation in metabolic phenotypes: study designs and applications. // Nat. Rev. Genet. 2012. Vol. 13, № 11. P. 759–769.

77. Shin S.-Y. et al. An atlas of genetic influences on human blood metabolites // Nat. Genet. 2014. Vol. 46, № April. P. 543–550.

78. Kastenmüller G. et al. Genetics of human metabolism: an update. // Hum. Mol. Genet. 2015.

79. Pardo B., Marcand S. Rap1 prevents telomere fusions by nonhomologous end joining // EMBO J. 2005. Vol. 24, № 17. P. 3117–3127.

80. Liu F. et al. A study of the SORL1 gene in Alzheimer's disease and cognitive function. // J. Alzheimers. Dis. 2009. Vol. 18, № 1. P. 51–64.

81. 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. Vol. 78, № 4. P. 629–644.

82. Aulchenko Y.S. et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. // Nat. Genet. 2009. Vol. 41, № 1. P. 47–55.

83. Wichmann H.-E., Gieger C., Illig T. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. // Gesundheitswesen. 2005. Vol. 67 Suppl 1. P. S26–S30.

84. Steffens M. et al. SNP-based analysis of genetic substructure in the German population. // Hum. Hered. 2006. Vol. 62, № 1. P. 20–29.

85. Ried J.S. et al. PSEA: Phenotype Set Enrichment Analysis--a new method for analysis of multiple phenotypes. // Genet. Epidemiol. 2012. Vol. 36, № 3. P. 244–252.

86. Römisch-Margl W. et al. Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics // Metabolomics. 2012. Vol. 8, № 1. P. 133–142.

87. Moayyeri A. et al. The UK Adult Twin Registry (TwinsUK Resource) // Twin Res. Hum. Genet. 2012. Vol. 16, № 01. P. 144–149.

88. Menni C. et al. Targeted metabolomics profiles are strongly correlated with nutritional patterns in women // Metabolomics. 2013. Vol. 9, № 2. P. 506–514.

89. Bacanu S.-A., Devlin B., Roeder K. Association studies for quantitative traits in structured populations. // Genet. Epidemiol. 2002. Vol. 22, № 1. P. 78–93.

90. Aulchenko Y.S. et al. GenABEL: an R library for genome-wide association analysis. // Bioinformatics. 2007. Vol. 23, № 10. P. 1294–1296.

91. Beasley T.M., Erickson S., Allison D.B. Rank-based inverse normal transformations are increasingly used, but are they merited? // Behav. Genet. 2009. Vol. 39, № 5. P. 580–595.

92. Hardy G.H. Mendelian proportions in a mixed population. 1908. // Yale J. Biol. Med. 2003. Vol. 76, № 2. P. 79–80.

93. Bittles A. Consanguinity and its relevance to clinical genetics. // Clin. Genet. 2001. Vol. 60, № 2. P. 89–98.

94. Aulchenko Y.S., Struchalin M. V, van Duijn C.M. ProbABEL package for genome-wide association analysis of imputed data. // BMC Bioinformatics. 2010. Vol. 11. P. 134.

95. Neyman J., Pearson E.S. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I // Biometrika. 1928. Vol. 20A, № 1/2. P. 175.

96. Akaike H. A new look at the statistical model identification // IEEE Trans. Automat. Contr. 1974. Vol. 19, № 6. P. 716–723.

97. Gorroochurn P. et al. Centralizing the non-central chi-square: A new method to correct for population stratification in genetic case-control association studies. // Genet. Epidemiol. 2006. Vol. 30, № 4. P. 277–289.

98. Price A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. // Nat. Genet. 2006. Vol. 38, № 8. P. 904–909.

99. Yu J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. // Nat. Genet. 2006. Vol. 38, № 2. P. 203–208.

100. Chen W.-M., Abecasis G.R. Family-based association tests for genomewide association scans. // Am. J. Hum. Genet. 2007. Vol. 81, № 5. P. 913–926.

101. Dupuis J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. // Nat. Genet. 2010. Vol. 42, № 2. P. 105–116.

102. de Bakker P.I.W. et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. // Hum. Mol. Genet. 2008. Vol. 17, № R2. P. R122–R128.

103. de Bakker P.I.W., Raychaudhuri S. Interrogating the major histocompatibility complex with high-throughput genomics. // Hum. Mol.

Genet. 2012. Vol. 21, № R1. P. R29–R36.

104. Gastwirth J.L.J.L.J.L. The Use of Maximin Efficiency Robust Tests in Combining Contingency Tables and Survival Analysis // J. Am. Stat. Assoc. 1985. Vol. 80. P. 380.

105. Davies R.B. Hypothesis testing when a nuisance parameter is present only under the alternative // Biometrika. 1987. Vol. 74. P. 33–43.

106. Zheng G., Freidlin B., Gastwirth J.L. Comparison of robust tests for genetic association using case-control studies. 2006. Vol. 49. P. 253–265.

107. Loley C. et al. A unifying framework for robust association testing, estimation, and genetic model selection using the generalized linear model. // Eur. J. Hum. Genet. 2013. Vol. 21. P. 1442–1448.

108. Tsepilov Y.A. et al. Development and application of genomic control methods for genome-wide association studies using non-additive models. // PLoS One. 2013. Vol. 8, № 12. P. e81431.

109. Zondervan K.T., Cardon L.R. The complex interplay among factors that influence allelic association. // Nat. Rev. Genet. 2004. Vol. 5. P. 89–100.

110. Vukcevic D. et al. Disease model distortion in association studies. // Genet. Epidemiol. 2011. Vol. 290. P. 278–290.

Приложение 1

Геномный контроль для аддитивной модели наследования

Рассмотрим биаллельный маркер с аллелями A и B. Пусть маркерный генотип может принимать значение 0, 1 или 2 для генотипов AA, AB и BB, соответственно.

Обозначим через $G_i \in \{0,1,2\}$, i = 1,...,R маркерный генотип у i-го больного, а через H_j, j = 1,...,S – то же у j-го здорового члена выборки. Учитывая, что наш тест определяется разницей частот аллелей у больных и здоровых, при R=S статистика Кохран-Армитажа Z^2 при аддитивной модели наследования пропорциональна квадрату статистики T, определенной как

$$T^{add} = \sum_i G_i^{add} - \sum_j H_j^{add}$$

Предположим, что выборка сформирована из m генетически различных субпопуляций. Пусть число больных, происходящих из каждой субпопуляции равно $a_1,\ldots,a_m$, а число здоровых – $b_1,\ldots,b_m$. Здесь $R = \sum_k a_k$ и $S = \sum_k b_k$, где R и S - количество больных и здоровых соответственно.

В общем виде, дисперсия T может быть записана как

$$Var(T) = \sum_{i=1}^{R} Var(G_j) + \sum_{j=1}^{S} Var(H_j) + 2\sum_{i<l} cov(G_i, G_l) + 2\sum_{j<l} cov(H_j, H_l) - 2\sum_i \sum_j cov(G_i, H_j)$$

Предположив, что при нулевой гипотезе дисперсии и ковариации генотипов равны между больными и здоровыми ($var(G_i^{add}) = var(H_i^{add})$, $cov(G_i^{add}, G_l^{add}) = cov(G_i^{add}, H_j^{add}) = cov(H_j^{add}, H_l^{add})$, $i \neq l, j \neq l$), при R = S получаем:

$$Var(T^{add}) = 2RVar\left(G_1^{add}\right) + \sum_k \{a_k(a_k-1) + b_k(b_k-1) - 2a_k b_k\} cov(G_1^{add}, G_2^{add})$$

Учитывая, что:

$$Var\left(G_1^{add}\right) = 2pq(1+F)$$

$$cov\left(G_1^{add}, G_2^{add}\right) = 4Fpq$$

(где p - частота аллеля, q=1-p, F- коэффициент инбридинга Райта),

можно показать что VIF, определенный как отношение дисперсии T^{add} к дисперсии статистики Кохран-Армитажа ($\lambda_{add} = \frac{Var_{H_0}(T^{add})}{4pq(1+F)}$) равен:

$$\lambda_{add} = 1 + \frac{F\sum_k \{a_k(a_k-1) + b_k(b_k-1) - 2a_k b_k\} cov(G_1^{add}, G_2^{add})}{R(1+3F)}$$

Как видно из формулы, фактор инфляции для аддитивной модели не зависит от частоты аллеля p.

Геномный контроль в случае неаддитивной модели

Выше было показано, что:

$$Var(T) = NVar(G_i) + cov(G_i, G_j)\sum_k \{a_k(a_k-1) + b_k(b_k-1) - 2a_k b_k\} \quad (1)$$

Для рецессивной модели было показано, что:

$$Var(G_1^{rec}) = p(F + p - Fp - p(F + p - Fp)^2)$$

Значение ковариации получено через знание информации о вероятности совместного распределения частот генотипов:

$$cov\left(G_i, G_j\right) = \Pr(AA, AA) - (Fp + (1-F)p^2)^2 + 2[xPr(AA, Aa) - 4xpq(Fp + (1-F)p^2)(1-F)] + x^2[\Pr(Aa, Aa) - (2(1-F)pq)^2] \quad (2)$$

Подставив $\Pr(AA) = \Pr(AA, AA) + 2\Pr(AA, Aa) + \Pr(AA, aa)$ в формулу (1):

$$Var(T^{rec}) = 2RVar(G_1^{rec}) + \sum_k \{a_k(a_k - 1) + b_k(b_k - 1) - 2a_k b_k\} cov(G_1^{rec}, G_2^{rec})$$

Имеем:

$$cov(G_1^{rec}, G_2^{rec}) = -\frac{2F(F(-1+p)-p)(-1+p)p(-3F+(-2+F+F^2)p)}{(1+F)(1+2F)}$$

$$Var(G_1^{rec}) = p(F + p - Fp - p(F + p - Fp)^2)$$

Аналогично получены значения для доминантной модели:

$$cov\left(G_1^{dom}, G_2^{dom}\right) = -\frac{2F(-1+p)p(1+(-1+F)p)(2-2p+F(2+F(-1+p)+p))}{(1+F)(1+2F)}$$

$$Var\left(G_1^{dom}\right) = -(2 + F(-1+p) - p)(-1+p)p(1+(-1+F)p)$$

Использую для рецессивной модели можно записать:

$$\lambda_{rec} = \frac{var(T^{rec})}{2Rpq(p+qF)\{1+p(1-F)\}}$$

Как видно из формулы, для неаддитивной модели VIF не является константой, которая может быть относительно просто оценена из эмпирических данных, а зависит от частоты аллеля.

Таблица S1. Результаты Левине теста для гомогенности дисперсий между значениями *E* (доля тестов с p-value ≤ 0.05) для двух корректирующих тестов в симуляционых исследованиях для ошибки 1 рода. Var1 и Var2 – общая дисперсия E* во всех частотных группах для первого и второго метода соответственно. Ratio –отношение Var1 и Var2.

Сравниваемые методы	Модель	Var1	Var2	Ratio	F-statistic	P-value
Коррекция на константу и VIF	Рецессивная	7.34E-05	6.69E-06	10.97	5518.13	0.00E+00
	Аддитивная	6.76E-06	4.78E-06	1.41	131.05	1.95E-29
	Доминантная	7.13E-05	6.45E-06	11.06	5670.45	0.00E+00
	Сверхдоминантная	8.26E-05	5.94E-06	13.91	5398.39	0.00E+00
	Генотипическая (df=2)	7.72E-06	5.49E-06	1.41	105.57	3.62E-24
Коррекция на константу и PGC	Рецессивная	7.34E-05	3.59E-06	20.45	7106.93	0.00E+00
	Аддитивная	6.76E-06	1.85E-05	0.36	23.51	1.34E-06
	Доминантная	7.13E-05	3.54E-06	20.13	7244.72	0.00E+00
	Сверхдоминантная	8.26E-05	3.91E-06	21.12	6257.51	0.00E+00
	Генотипическая (df=2)	7.72E-06	3.49E-06	2.21	598.67	7.28E-116
PGC и VIF	Рецессивная	6.69E-06	3.59E-06	1.86	430.68	8.64E-87
	Аддитивная	4.78E-06	1.85E-05	0.26	0.56	4.54E-01
	Доминантная	6.45E-06	3.54E-06	1.82	415.04	5.56E-84
	Сверхдоминантная	5.94E-06	3.91E-06	1.52	178.38	4.91E-39
	Генотипическая (df=2)	5.49E-06	3.49E-06	1.57	229.22	4.19E-49

Приложение 2

Таблица S1. Результаты анализа неаддитивных эффектов для различных трансформаций признаков. Представлены результаты ПГАА для кодоминантной и аддитивной моедли для 20 наденных на KORA локусов. Четрые типа трансформации были использвоаны. Каждая трансоффрмация обозначенна в буквенной форме: каждая буква соответвует трансформации в порядке следования букв: **G** – гауссинизация, **R** – коррекция на пол, возраст и номер пробы, **L** – логарифмизация (так, например, **LR** – последовательная трансоффрмация с использвоаниием логарифмизации, а затем с коррекцией на ковариаты; в нашем исследовании **RG** трансформация была использована). Для каждой тарнсформаии представлены результаты для тестов LRT и AIC. В столбце LRT представлены все ограниченные модели, которые недостоверно отличались от кодоминантной модели (в порядке уменьшения значения p-value). Кодоминантнпя модель представлена в этой колонке, если она была достоверно лучше, чем все ограниченные модели. В столбце AIC показана самая парсимонная модель. Если самая парсимонная модель – кодоминантная, то она отделена через слэш от следующей самой хорошей ограниченной модели. Обозначения для ограниченных одностепенных моделей: r, a, d, o, g – рецессивная, аддитивная, доминантная, сверхдоминантная и кодоминантная, соответственно. Таблица разделена на 2 части: в верхней части представлены локусы, которые уже описаны в ранее опубликованных ПГАА по тем же данным (Illig et al. 2010), в нижней части приведены новые локусы.

				GR		LR		LRG		R		RG	
SNP	**Метаболит (отношение)**	**Хр.**	**Позиция**	**LRT**	**AIC**	**LRT**	**AIC**	**LRT**	**AIC**	**LRT**	**AIC**	**LRT**	**AIC**
rs7552404	C12/C10	1	75,908,534	a	a	a	g/a	a	a	a	a	a	a
rs7601356	C9/PC.ae.C30.0	2	210,764,902	g	g/d	g	g/d	g	g/d	g	g/d	g	g/d
rs715	Gly/Gln	2	211,251,300	r	g/r	r	g/r	r	g/r	r	g/r	r	g/r
rs8396	C7.DC/C10	4	159,850,267	a	a	a	a	a	a	a	a	a	a
rs2046813	PC.ae.C42.5/PC.ae.C44.5	4	186,006,153	a	a	a	a	a	a	a	a	a	a
rs273913	C5/PC.ae.C34.1	5	131,689,055	a	g/a	a	g/a	a	g/a	a	g/a	a	g/a
rs3798719	PC.aa.C42.5/PC.aa.C40.3	6	11,144,811	a	a	a	a	a	a	a	a	a	a
rs12356193	C0	10	61,083,359	a	a	a	a	a	a	a	a	a	a
rs603424	C16.1/C14	10	102,065,469	a	g/a	a	g/a	a	g/a	a	g/a	a	g/a
rs174547	PC.aa.C36.3/PC.aa.C36.4	11	61,327,359	a	g/a	a	g/a	a	g/a	g	g/a	a	g/a
rs2066938	C3/C4	12	119,644,998	g	g/a	g	g/a	g	g/a	g	g/a	g	g/a
rs4902242	PC.aa.C28.1/PC.ae.C40.2	14	63,299,842	a,r	g/a	a,r	g/a	a,r	g/a	a,r	a	a,r	g/a
rs1077989	PC.ae.C32.1/PC.ae.C34.1	14	67,045,575	a	g/a	a	g/a	a	g/a	a	g/a	a	g/a
rs4814176	SM..OH..C24.1/SM..OH..C22.1	20	12,907,398	a	a	a	a	a	a	a	a	a	a
rs6970485	lysoPC.a.C28.0/PC.aa.C26.0	7	11,752,704	d	d	d	d	d	d	d	g/d	d	d

rs1894832	Ser/Trp	7	56,144,740	a	a	a	a	a	a	a	a	a	a
rs2657879	His/Gln	12	55,151,605	a	a	a	a	a	a	a	a	a	a
rs7200543	PC.aa.C36.2/PC.aa.C38.3	16	15,037,471	a,d	a	a,d	a	a,d	a	a,d	g/a	a,d	a
rs1466448	SM.C18.1/SM.C16.1	19	8,195,519	a	a	a	a	a	a	a	a	a	a
rs5746636	xLeu/Pro	22	17,276,301	a,d	a	a,d	a	a,d	a	a	a	a	a

Приложение 3

Результаты тестирования рецессивной, доминантной, сверхдоминантной моделями и MAX-тестом

Мы провели ПГАА для рецессивной и доминантной моделей. Даже с использованием либерального уровня значимости ($5\times10^{-8}/22801$) вместо строгого ($5\times10^{-8}/(22801\times4)$), мы не смогли обнаружить дополнительные локусы. Из 20 локусов, определенных с помощью кодоминантной модели, четырнадцать были обнаружены с помощью рецессивной модели и восемнадцать – доминантной. Использование сверхдоминантной модели выявило десять из 20 описанных локусов и одну дополнительную ассоциацию между rs219040 на седьмой хромосоме (p-value $<3\ .94\times10^{-13}$) и отношением C5.1/C6.1. Локус располагался вблизи гена *STEAP2-AS1* (кодирующего антисмысловую РНК гена RNA1), биологическую роль которого нельзя напрямую соотнести с контролем метаболизма. Его p-value для HWE было близко к пороговому для контроля качетсва (p-value $< 1.03\times10^{-05}$), и его не удалось отреплицировать на данных TwinsUK (p-value = 0.8).

Таблица S1. Результаты ПГАА для рецессивной модели. В таблице представлены результаты для 14 значимых локусов для рецессивной модели (P-value $<2.19\times10^{-12}$). chr: Хромосома; AF – частота эффекторного аллеля.

				KORA sample		TwinsUK sample		
SNP	**metabolite (ratio)**	**chr**	**position**	**AF**	**p-value recessive**	**AF**	**p-value recessive**	**gene**
rs11161521	C8/C12	1	75,988,918	0.70	2.74E-62	0.69	3.98E-28	*ACADM*
rs7558218	C9/PC.ae.C30.0	2	210,811,690	0.36	3.25E-61	0.35	2.94E-23	*ACADL*
rs7422339	Gly/Gln	2	211,248,752	0.69	7.37E-75	-	-	*CPS1*
rs8396	C7.DC/C10	4	159,850,267	0.71	3.75E-23	0.68	1.92E-17	*PPID*
rs2046813	PC.ae.C42.5/PC.ae.C44.5	4	186,006,153	0.69	8.14E-14	0.69	1.88E-03	*SLED1*
rs273913	C5/PC.ae.C34.1	5	131,689,055	0.41	4.00E-14	0.35	8.12E-02	*SLC22A4*
rs3798723	PC.aa.C42.5/PC.aa.C40.3	6	11,149,706	0.75	4.76E-26	-	-	*ELOVL2*
rs603424	C14/C16.1	10	102,065,469	0.80	9.86E-15	0.82	1.53E-02	*PKD2L1*
rs174547	PC.aa.C36.3/PC.aa.C36.4	11	61,327,359	0.70	1.02E-145	0.65	8.09E-44	*FADS1*
rs2066938	C4/C3	12	119,644,998	0.27	5.22E-99	0.26	7.29E-39	*ACADS*
rs7156144	PC.ae.C32.1/PC.ae.C34.1	14	67,049,466	0.59	1.14E-27	0.57	6.80E-14	*SGPP1*
rs1741	PC.aa.C38.3/PC.aa.C36.2	16	15,037,852	0.69	1.98E-13	0.72	2.07E-06	*NTAN1*
rs364585	SM..OH..C24.1/SM.C24.0	20	12,910,718	0.64	9.12E-28	0.59	1.02E-12	*SPTLC3*
rs5747922	xLeu/Pro	22	17,269,755	0.77	6.63E-19	0.73	6.18E-03	*DGCR6*

Таблица S2. Результаты ПГАА для доминантной модели. В таблице представлены результаты для 18 значимых локусов для доминантной модели (P-value $<2.19\times10^{-12}$). chr: Хромосома; AF – частота эффекторного аллеля.

				KORA sample		TwinsUK sample		
SNP	**metabolite (ratio)**	**chr**	**position**	**AF**	**p-value dominant**	**AF**	**p-value dominant**	**gene**
rs7552404	C8/C12	1	75,908,534	0.30	5.48E-64	0.31	4.02E-28	*ACADM*
rs7601356	C9/PC.ae.C30.0	2	210,764,902	0.63	2.78E-64	0.65	1.43E-23	*ACADL*
rs2216405	Gly/Gln	2	211,325,139	0.19	8.84E-40	0.16	1.15E-19	*CPS1*
rs12505475	C7.DC/C10	4	159,854,694	0.29	4.37E-23	0.33	1.51E-17	*PPID*
rs4862429	PC.ae.C42.5/PC.ae.C44.5	4	186,006,834	0.31	1.25E-13	0.31	1.65E-03	*SLED1*
rs270605	C5/PC.ae.C34.1	5	131,679,710	0.60	4.71E-14	0.65	8.17E-02	*SLC22A4*
rs3798719	PC.aa.C42.5/PC.aa.C40.3	6	11,144,811	0.25	3.73E-26	0.23	1.24E-03	*ELOVL2*
rs6970485	PC.aa.C26.0/PC.ae.C38.1	7	11,752,704	0.35	2.33E-17	-	-	*THSD7A*
rs12356193	C0	10	61,083,359	0.17	4.51E-25	0.16	4.25E-08	*SLC16A9*
rs174556	PC.aa.C36.3/PC.aa.C36.4	11	61,337,211	0.27	2.78E-144	0.32	4.65E-46	*FADS1*
rs1043011	Gln/Met	12	55,151,307	0.21	4.03E-13	0.19	4.67E-04	*GLS2*
rs3916	C3/C4	12	119,661,655	0.73	2.46E-97	0.75	4.07E-36	*ACADS*
rs4902243	PC.aa.C28.1/PC.ae.C40.2	14	63,303,996	0.17	3.66E-36	0.14	4.24E-17	*SGPP1*
rs1077989	PC.ae.C32.1/PC.ae.C34.1	14	67,045,575	0.46	3.60E-35	0.47	3.99E-17	*PLEKHH1*
rs7200543	PC.aa.C36.2/PC.aa.C38.3	16	15,037,471	0.31	2.14E-15	0.28	1.47E-06	*NTAN1*
rs1466448	SM.C16.1/SM.C18.1	19	8,195,519	0.22	1.45E-13	0.19	1.75E-10	*CERS4*
rs4814176	SM..OH..C24.1/SM.C24.0	20	12,907,398	0.36	6.70E-28	0.42	3.69E-13	*SPTLC3*
rs5746636	xLeu/Pro	22	17,276,301	0.24	3.80E-19	0.27	4.62E-03	*DGCR6*

Таблица S3. Результаты ПГАА для сверхдоминантной модели. В таблице представлены результаты для 11 значимых локусов для сверхдоминантной модели (P-value $<2.19\times10^{-12}$). chr: Хромосома; AF – частота эффекторного аллеля.

				KORA sample		TwinsUK sample		
SNP	**metabolite (ratio)**	**chr**	**position**	**AF**	**p-value**	**AF**	**p-value**	**gene**
rs7365179	C10/C12	1	76,096,212	0.22	1.74E-30	0.24	5.01E-15	*ACADM*
rs12468576	C5.M.DC/C9	2	210,662,236	0.20	2.07E-14	0.22	1.65E-04	*ACADL*
rs7422339	Gly/Gln	2	211,248,752	0.69	3.26E-42	-	-	*CPS1*
rs3756963	PC.aa.C42.6/PC.aa.C38.5	6	11,130,140	0.76	1.13E-13	-	-	*ELOVL2*
rs6970485	lysoPC.a.C28.0/PC.aa.C26.0	7	11,752,704	0.35	3.47E-29	-	-	*THSD7A*
rs2190401	C5.1/C6.1	7	89,504,946	0.76	3.94E-13	0.78	8.37E-01	*STEAP2-AS1*
rs12356193	C0	10	61,083,359	0.17	9.27E-18	0.16	3.15E-06	*SLC16A9*
rs968567	PC.aa.C36.3/PC.aa.C36.4	11	61,352,140	0.15	6.05E-58	0.19	9.20E-13	*FADS1*
rs12310160	C3/C4	12	119,584,265	0.86	2.02E-26	0.85	2.76E-10	*ACADS*
rs7157785	PC.ae.C40.2/PC.aa.C28.1	14	63,305,309	0.17	2.08E-30	0.17	3.59E-12	*SGPP1*
rs4508668	SM.C24.0/SM..OH..C24.1	20	12,903,601	0.32	2.64E-13	0.37	2.06E-04	*SPTLC3*

Таблица S4. Результаты MAX-теста для 20 локусов, найденных двухстепенным тестом. chr: Хромосома; AF – частота эффекторного аллеля. MAX_KORA и MAX_TUK – p-values для **MAX-теста** для KORA и TwinsUK соответственно.

SNP	Trait	chr	Pos	Freq_KORA	g_pval_KORA	Freq_TUK	g_pval_TUK	MAX_KORA*	MAX_TUK*
rs7552404	C12/C10	1	75,908,534	0.300	1.69E-72	0.314	1.89E-29	0.00E+00	0.00E+00
rs7601356	C9/PC.ae.C30.0	2	210,764,902	0.632	1.24E-70	0.649	6.86E-28	0.00E+00	0.00E+00
rs715	Gly/Gln	2	211,251,300	0.687	4.28E-69	0.703	1.12E-48	0.00E+00	0.00E+00
rs8396	C7.DC/C10	4	159,850,267	0.707	5.98E-26	0.678	3.14E-17	0.00E+00	0.00E+00
rs2046813	PC.ae.C42.5/PC.ae.C44.5	4	186,006,153	0.688	6.29E-17	0.687	1.18E-03	0.00E+00	4.46E-04
rs273913	C5/PC.ae.C34.1	5	131,689,055	0.405	1.60E-16	0.351	4.19E-02	0.00E+00	2.21E-02
rs3798719	PC.aa.C42.5/PC.aa.C40.3	6	11,144,811	0.248	5.01E-32	0.234	4.01E-04	0.00E+00	1.11E-04
rs12356193	C0	10	61,083,359	0.166	2.18E-27	0.161	1.20E-07	0.00E+00	7.82E-06
rs603424	C16.1/C14	10	102,065,469	0.801	3.70E-18	0.818	1.99E-02	0.00E+00	1.47E-02
rs174547	PC.aa.C36.3/PC.aa.C36.4	11	61,327,359	0.701	2.29E-208	0.649	2.09E-76	0.00E+00	0.00E+00
rs2066938	C3/C4	12	119,644,998	0.270	1.73E-159	0.257	2.17E-67	0.00E+00	0.00E+00
rs4902242	PC.aa.C28.1/PC.ae.C40.2	14	63,299,842	0.849	2.00E-35	0.872	4.78E-15	0.00E+00	2.46E-07
rs1077989	PC.ae.C32.1/PC.ae.C34.1	14	67,045,575	0.463	6.80E-42	0.472	4.05E-18	0.00E+00	0.00E+00
rs4814176	SM..OH..C24.1/SM..OH..C22.1	20	12,907,398	0.364	2.69E-31	0.416	9.69E-09	0.00E+00	5.96E-09
rs6970485	lysoPC.a.C28.0/PC.aa.C26.0	7	11,752,704	0.354	1.21E-47	-	-	0.00E+00	9.82E-01
rs1894832	Ser/Trp	7	56,144,740	0.508	1.98E-12	0.511	4.02E-03	3.32E-13	1.97E-03
rs2657879	His/Gln	12	55,151,605	0.207	2.89E-14	0.186	1.90E-06	2.58E-13	7.33E-03
rs7200543	PC.aa.C36.2/PC.aa.C38.3	16	15,037,471	0.312	7.45E-16	0.277	1.66E-06	0.00E+00	5.65E-07
rs1466448	SM.C18.1/SM.C16.1	19	8,195,519	0.222	7.01E-16	0.194	3.90E-10	1.78E-15	1.88E-08
rs5746636	xLeu/Pro	22	17,276,301	0.236	2.98E-20	0.273	2.40E-03	0.00E+00	9.81E-04

*Значение 0.00E+00 значит, что p-value<1e-15

Приложение 4

Моделирование

Мы провели исследование на моделированных данных для определения влияния шума в фенотипических и генотипических данных на модель наследования генетических маркеров в локусе в случае неаддитивной модели эффекта функционального варианта. Данный параграф разбит на две части – моделирование фенотипического шума и моделирование генетических корреляций.

Фенотипический шум.

Мы использовали предположение о том, что измеряемые признаки (в данном случае концентрации метаболитов, измеренные технологией масс-спектрометрии) высоко скоррелированы с реальными биологическими фенотипами, которые могут контролироваться неаддитивными эффектами генов. Другими словами, в процессе измерения всегда появляются шумы. В данном исследовании изучаемые фенотипы могли быть высоко скоррелированными (но быть эквивалентными) с фенотипами, контролирующимися неаддитивными эффектами генов. Таким образом, мы проверили как образом данные шумы могут влиять на модель наследования ассоциируемого варианта.

В ходе каждой итерации мы моделировали генотипы, распределенные биномиально с фиксированной частотой (были использованы частоты: 0.25, 0.50 и 0.75), которые были ассоциированы с признаком. Моделирование данных проводилось для выборки объемом 2 000 людей.

Уровень предрасположенности для изначальных фенотипов моделировался как сумма независимых эффектов ассоциированного локуса, полигенетических эффектов и средовой компоненты. Коэффициент наследуемости был равен 0.7. Эффект ассоциированного локуса подбирался в зависимости от частоты

минорного аллеля таким образом, чтобы эффект локуса объяснял 5% общей дисперсии признака, после чего мы кодировали генотипы в соответствии с моделью и умножали их на полученное значение эффекта. Для моделирования полигенетического эффекта мы случайным образом сгенерировали 50 генетических маркеров, которым были приписаны эффекты на фенотип в зависимости от их частоты минорного аллеля таким образом, чтобы каждый маркер объяснял одинаковую долю наследуемости, оставшейся после вычета эффекта ассоциированного локуса. Средовая компонента была смоделирована нормально распределенной со средним 0 и стандартным отклонением равным 0.3.

Мы смоделировали фенотипы для рецессивной, доминантной и сверхдоминантной моделей. Скоррелированные признаки были равны сумме изначального фенотипа и вектора шума, распределенного нормально с фиксированным стандартным отклонением (0, 0.5, 1, 1.5, 2). Таким образом мы получили 4 скоррелированных признака и один оригинальный. В ходе исследования мы провели 1,000 итераций описанного выше моделирования.

Результаты представлены в таблице S1, а также на рисунке S1. В таблице представлены результаты только для фенотипов, корреляция которых была не ниже 0.5.

Генетические корреляции

В случае использования технологии геномных микрочипов, SNP, находящийся в неравновесии по сцеплению с функциональным вариантом, будет найден в ПГАА в большинстве случаев (а не см функциональный вариант). Мы проверили, каким образом меняется неаддитивная модель эффекта в случае, если вместо таргетного SNP мы будем анализировать нетаргетный. Схема моделирования данных была сходна с описанной выше схемой моделирования для исследования фенотипических корреляций. Вместо генерирования

скоррелированных фенотипов, мы симулировали скоррелированные генотипы с одинаковой частотой минорного аллеля. Сперва мы смоделировали функциональный вариант, распределенный биномиально с фиксированный частотой аллеля. Затем мы случайным образом изменяли генотипы для получения вектора генотипов с желаемым уровнем корреляции между генотипами с сохранением частоты аллеля и выполнением равновесия Харди-Вайеньберга. В итоге мы получили 3 скоррелированных генотипа (квадрат корреляции был равен 0.7, 0.8 и 0.9) и один оригинальный функциональный генотип. Результаты представлены в таблице S2 и на рисунке S2.

Также мы использовали реальные данные генотипов, для которых моделировались искусственные фенотипы. Мы выбрали локус, представленный SNP rs419291, расположенный на 5ой хромосоме. Согласно графику региональной ассоциации, данный локус содержит большое число генотипов, высоко скоррелированных с наиболее ассоциированным (в данном исследовании) SNP - rs419291 В районе 1Мегабазы от rs419291 были выбраны 8 SNP, находящиеся в неравновесии по сцеплению с rs419291 (квадрат корреляции был в пределах от 0.39 до 1). Мы смоделировали фенотипы с функциональным вариантом для рецессивной, доминантной и сверхдоминантной модели по схеме, описанной ранее, а затем мы вычислили тестовую статистику для каждого выбранного SNP. Выбор модели проводился по минимальному p-value среди тестировавшихся моделей. Для каждой симулированной модели мы проверили различия между ее –log10(p-value) с оставшимися моделями, в том числе, с генотипической (кодоминантной) моделью. Для каждого SNP был оценен его эффект и его стандартная ошибка в каждой из моделей. Результаты представлены в таблице S3 и на рисунке S3. Число итераций составило 100.

Результаты

По результатам исследования различных сценариев моделирования генотипов и фенотипов, мы можем сделалать вывод, что различия между симулированными неаддитивными моделями и аддитивной моделью уменьшаются в случае наличия сильного шума у признаков или уменьшения неравновесия по сцеплению. Для выбранных нами параметров симуляции, согласно тесту LRT, разница между генотипической (кодоминантной) и аддтивными моделями станвоиться все меньше с уменьшением значимости ассоциации для всех моделей. Во всех случаях лучшей моделью эффекта SNP была симулированная модель. Ожидаемо, двух-степенной генотипический (кодоминантный) тест имел меньшую мощность по сравнению с моделью, использовавшейся для симуляции, однако он был устойчив к симулирвоанной модели по сравнению с остальными моделями, особенно при моделировании сверхдоминантной моедли, которую очень трудно детектировать аддитивным тестом.

Таблица S1. Результаты моделирования неаддитивных эффектов для признаков, коррелированных с оригинальным фенотипом. В таблице представлены средние значения и стандартные отклонения признаков. В строках “–log(BM)”, “–log(A)”, “–log(G)” представлены логарифмы p-value для симмулированной, аддитивной и генотипической моделей соответственно. Строка “LRT(A,G)” содержит значения p-value теста отношения правдоподобия между аддитивной и генотипической моделями.

	AF		Original	0.89+-0	0.71+-0.01	0.55+-0.02
Recessive	0.25	-log(BM)	23.4+-4.61	18.78+-4.08	12.15+-3.23	7.72+-2.49
		-log(A)	9.9+-3.09	8.05+-2.7	5.39+-2.17	3.49+-1.65
		-log(G)	22.52+-4.58	17.94+-4.05	11.41+-3.19	7.09+-2.43
		LRT(A, G)	14.53+-3.39	11.69+-3.07	7.63+-2.43	5+-1.96
	0.5	-log(BM)	23.34+-4.49	18.82+-4.02	11.97+-3.17	7.8+-2.59
		-log(A)	15.93+-3.69	12.91+-3.29	8.33+-2.57	5.47+-2.11
		-log(G)	22.46+-4.47	18+-3.99	11.24+-3.13	7.18+-2.54
		LRT(A, G)	8.41+-2.65	6.87+-2.43	4.49+-1.9	3.1+-1.62
	0.75	-log(BM)	23.48+-4.32	18.93+-3.86	12.13+-3.19	7.8+-2.48
		-log(A)	20.15+-4	16.28+-3.59	10.49+-2.94	6.76+-2.31
		-log(G)	22.59+-4.27	18.09+-3.83	11.38+-3.15	7.15+-2.42
		LRT(A, G)	4.21+-1.79	3.49+-1.64	2.38+-1.33	1.68+-1.08
Dominant	0.25	-log(BM)	23.71+-4.41	19.08+-3.98	12.24+-3.17	7.8+-2.46
		-log(A)	20.31+-4.06	16.41+-3.63	10.56+-2.9	6.8+-2.26
		-log(G)	22.82+-4.37	18.23+-3.95	11.51+-3.12	7.16+-2.41
		LRT(A, G)	4.28+-1.86	3.5+-1.68	2.43+-1.43	1.65+-1.15
	0.5	-log(BM)	23.69+-4.28	19.06+-3.95	12.13+-3.11	7.9+-2.46
		-log(A)	15.91+-3.49	12.86+-3.2	8.38+-2.57	5.56+-2.06
		-log(G)	22.81+-4.24	18.23+-3.9	11.41+-3.06	7.27+-2.4
		LRT(A, G)	8.8+-2.72	7.17+-2.44	4.62+-1.96	3.1+-1.58
	0.75	-log(BM)	23.63+-4.85	18.96+-4.23	12.12+-3.4	7.78+-2.63
		-log(A)	9.85+-3.13	8.04+-2.76	5.27+-2.21	3.51+-1.69
		-log(G)	22.75+-4.82	18.12+-4.2	11.4+-3.36	7.13+-2.58
		LRT(A, G)	14.8+-3.56	11.89+-3.18	7.73+-2.62	5.02+-2.03
Over-dominant	0.25	-log(BM)	23.66+-4.34	19.04+-3.92	12.14+-3.22	7.92+-2.58
		-log(A)	10+-2.99	8.12+-2.69	5.36+-2.15	3.65+-1.74
		-log(G)	22.78+-4.31	18.2+-3.89	11.42+-3.16	7.29+-2.51
		LRT(A, G)	14.69+-3.36	11.89+-3.06	7.66+-2.49	5.06+-2
	0.5	-log(BM)	23.58+-4.42	19.03+-3.98	11.98+-3.1	7.74+-2.45
		-log(A)	0.43+-0.43	0.45+-0.43	0.46+-0.47	0.43+-0.42
		-log(G)	22.69+-4.39	18.2+-3.94	11.26+-3.06	7.1+-2.39
		LRT(A, G)	23.58+-4.42	19.03+-3.98	11.98+-3.1	7.74+-2.45
	0.75	-log(BM)	23.5+-4.35	18.92+-3.94	12.13+-3.15	7.84+-2.54
		-log(A)	9.82+-2.89	8.02+-2.6	5.28+-2.03	3.55+-1.69

	-log(G)	22.62+-4.32	18.09+-3.91	11.4+-3.1	7.2+-2.48
	LRT(A, G)	14.7+-3.51	11.88+-3.16	7.73+-2.56	5.05+-1.98

Рисунок S1. Графики p-value для различных моделей в зависимости от корреляции признаков. На каждом графки представлено 3 линии: красная – симмулированная модель, синяя – генотипическая, зеленая – аддитивная. Горизонтальная линия – уровень значимости 5y-8. Соркращения r, a, d, o, g являются рецессивной, аддитивной, доминантной и сверхдоминатной моделями соотвественно.

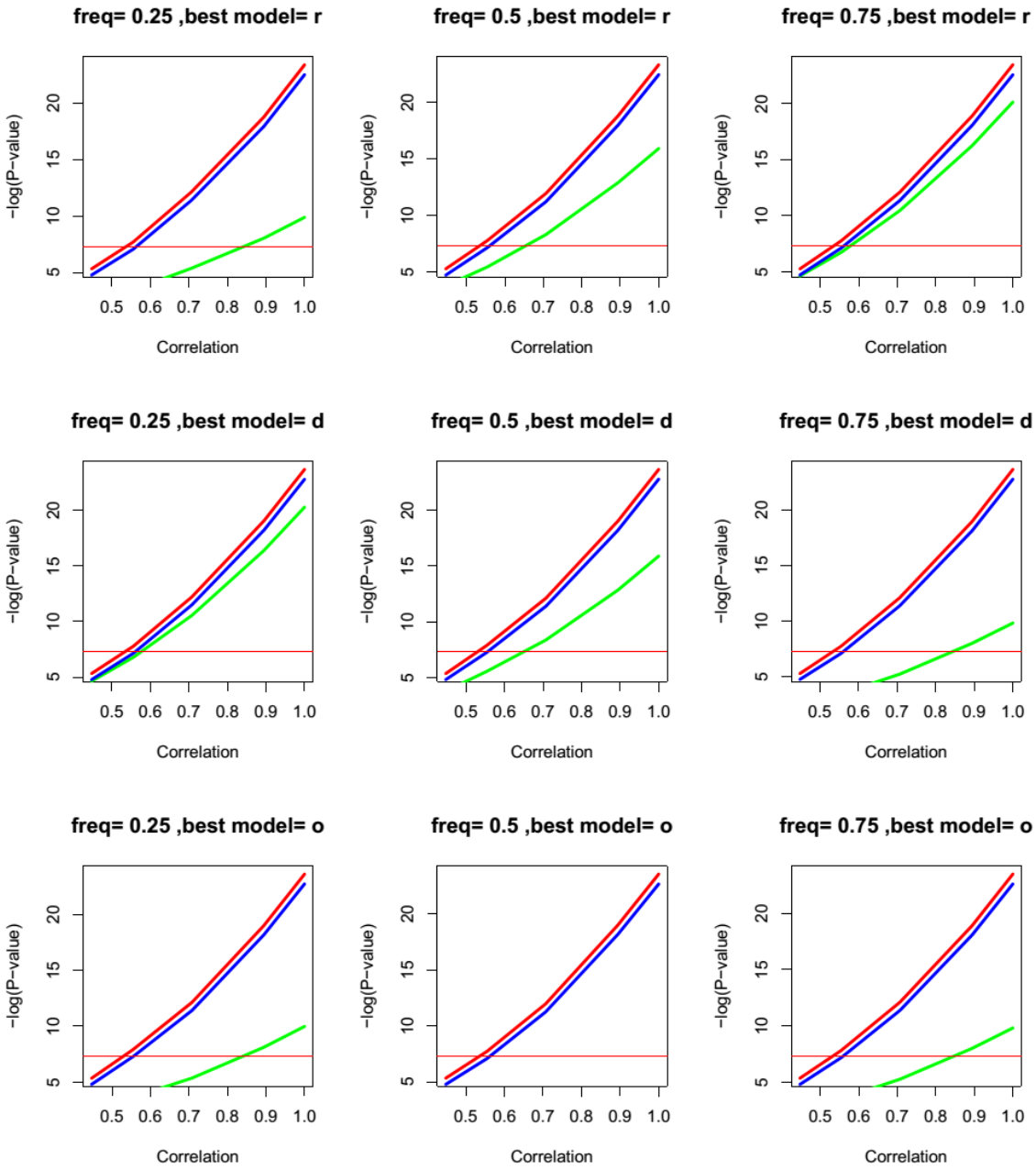


Таблица S2. Результаты моделирования неаддитивных эффектов SNP, скореллированных с функциональным вариантом. В таблице представлены средние значения и стандартные отклонения признаков. В строках **“–log(BM)”, “–log(A)”, “–log(G)”** представлены логарифмы p-value для симмулированной, аддитивной и генотипической моделей соответственно. Строка **“LRT(A,G)”** содержит значения p-value теста отношения правдоподобия между аддитивной и генотипической моделями.

	AF		Target SNP	0.9	0.8	0.7
Recessive	0.25	-log(BM)	23.9+-4.63	19.78+-4.22	15.82+-3.73	12.36+-3.33
		-log(A)	9.94+-2.95	8.3+-2.69	6.77+-2.38	5.39+-2.14
		-log(G)	23+-4.59	18.92+-4.2	15.01+-3.68	11.62+-3.27
		LRT(A,G)	14.97+-3.44	12.44+-3.19	9.97+-2.77	7.84+-2.5
	0.5	-log(BM)	23.57+-4.34	19.53+-3.86	15.64+-3.49	12.18+-3.14
		-log(A)	15.95+-3.51	13.28+-3.13	10.73+-2.85	8.4+-2.49
		-log(G)	22.67+-4.3	18.67+-3.82	14.83+-3.46	11.45+-3.1
		LRT(A,G)	8.61+-2.59	7.2+-2.34	5.81+-2.16	4.64+-1.97
	0.75	-log(BM)	23.44+-4.27	19.36+-3.9	15.64+-3.68	12.1+-3.23
		-log(A)	20.16+-3.93	16.65+-3.55	13.53+-3.4	10.5+-2.97
		-log(G)	22.57+-4.22	18.53+-3.85	14.85+-3.64	11.36+-3.17
		LRT(A,G)	4.18+-1.83	3.56+-1.72	2.91+-1.49	2.34+-1.3
Dominant	0.25	-log(BM)	23.6+-4.35	19.56+-3.88	15.66+-3.62	12.31+-3.11
		-log(A)	20.37+-4.04	16.92+-3.59	13.57+-3.4	10.72+-2.91
		-log(G)	22.7+-4.31	18.7+-3.85	14.87+-3.59	11.58+-3.07
		LRT(A,G)	4.1+-1.76	3.47+-1.62	2.89+-1.49	2.34+-1.33
	0.5	-log(BM)	23.57+-4.33	19.51+-3.89	15.63+-3.7	12.18+-3.08
		-log(A)	15.98+-3.6	13.32+-3.27	10.68+-3.05	8.38+-2.52
		-log(G)	22.69+-4.3	18.68+-3.86	14.84+-3.66	11.45+-3.04
		LRT(A,G)	8.6+-2.69	7.16+-2.42	5.87+-2.18	4.66+-1.98
	0.75	-log(BM)	23.51+-4.56	19.41+-4.16	15.68+-3.8	12.26+-3.34
		-log(A)	9.92+-3.05	8.34+-2.81	6.77+-2.45	5.43+-2.13
		-log(G)	22.64+-4.54	18.58+-4.14	14.89+-3.77	11.5+-3.3
		LRT(A,G)	14.62+-3.44	12.06+-3.17	9.84+-2.91	7.68+-2.52
Over-dominant	0.25	-log(BM)	23.61+-4.07	19.61+-3.88	15.66+-3.53	12.22+-3.11
		-log(A)	9.84+-2.78	8.31+-2.63	6.75+-2.29	5.35+-2.04
		-log(G)	22.72+-4.05	18.76+-3.86	14.86+-3.49	11.48+-3.06
		LRT(A,G)	14.78+-3.35	12.28+-3.13	9.82+-2.84	7.74+-2.52
	0.5	-log(BM)	23.47+-4.35	19.46+-4.06	15.62+-3.57	12.18+-3.15
		-log(A)	0.43+-0.44	0.43+-0.44	0.44+-0.45	0.43+-0.43
		-log(G)	22.58+-4.33	18.62+-4.03	14.83+-3.53	11.44+-3.11
		LRT(A,G)	23.47+-4.35	19.47+-4.06	15.62+-3.56	12.18+-3.15

	0.75	-log(BM)	23.47+-4.23	19.48+-3.83	15.68+-3.63	12.14+-3.17
		-log(A)	9.86+-2.76	8.28+-2.53	6.72+-2.27	5.31+-2.09
		-log(G)	22.58+-4.19	18.63+-3.79	14.88+-3.58	11.4+-3.12
		LRT(A,G)	14.62+-3.53	12.16+-3.16	9.88+-2.94	7.7+-2.48

Рисунок S2. Графики p-value для различных моделей в зависимости от корреляции генотипов. На каждом графки представлено 3 линии: красная – симмулированная модель, синяя – генотипическая, зеленая – аддитивная. Горизонтальная линия – уровень значимости 5e-8. Соркращения r, a, d, o, g являются рецессивной, аддитивной, доминантной и сверхдоминатной моделями соответственно.

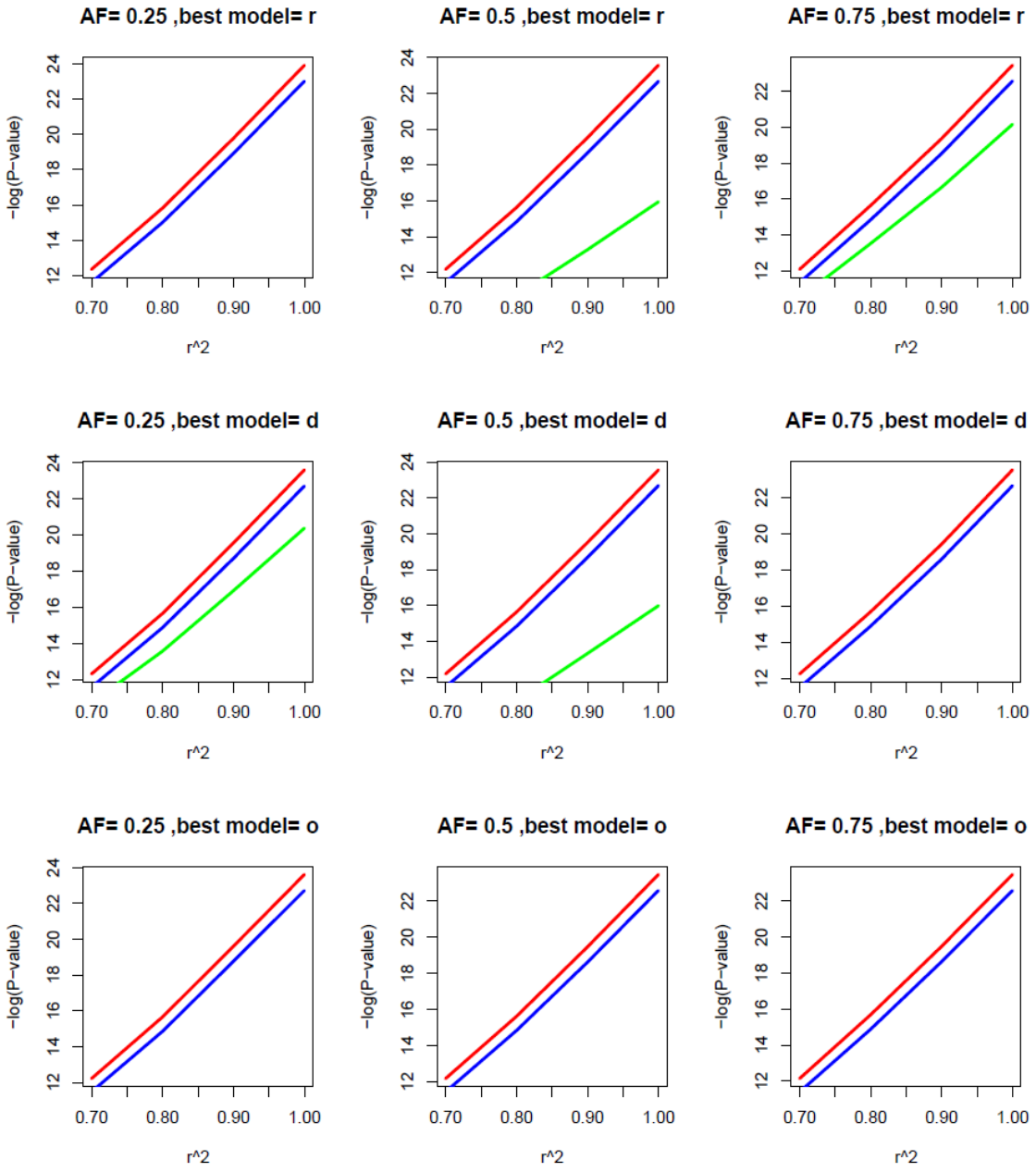


Таблица S3. Результаты моделирования неаддитивных эффектов SNP, взятых из реальных данных, скореллированных с функциональным вариантом. В таблице представлены различия между –log10(p-value) симулированной модели и других одно и двух степенных тестов. Из реальных генетических данных были выбраны 8 SNP согласно их высокому значению квадратов корреляции с функциональным вариантом. Для каждой из симулированных моделей (сокращенно **R, A, D, O, G -** рецессивной, аддитивной, доминантной и сверхдоминатной) представлены –log10(pvalue) теста ассоциации соответствующей модели (r, a, d, o) и различия с другими моделями (da, dr, do, dg - рецессивной, аддитивной, доминантной и сверх-доминатной соответственно)

		R^2							
Simulated model		1	0.98	0.92	0.82	0.76	0.6	0.51	0.39
R	r	20.8+-4.16	20.23+-4.09	11.76+-3.13	19.04+-4.22	17.23+-3.89	6.71+-2.17	6.02+-1.98	5.82+-2.03
	da	8.25+-2.58	8.02+-2.54	3.12+-2.18	7.49+-2.71	6.55+-2.73	0.83+-1.59	0.59+-1.45	0.57+-1.42
	dd	17.37+-3.85	17.01+-3.76	8.05+-2.92	16.21+-4.02	14.46+-3.9	3.81+-2.12	3.29+-1.89	3.29+-1.85
	do	17.28+-3.94	16.81+-3.9	11.28+-3.22	15.22+-3.81	14.1+-3.46	6.3+-2.23	5.64+-2.04	5.43+-2.04
	dg	0.88+-0.28	0.86+-0.28	0.21+-0.69	0.83+-0.29	0.76+-0.36	0.06+-0.71	0.03+-0.71	0.05+-0.64
A	a	6.03+-2.27	5.84+-2.19	4.4+-1.85	5.36+-2.07	5.25+-2.08	3.09+-1.58	2.85+-1.45	2.8+-1.39
	dr	2.17+-1.4	2.14+-1.37	1.98+-1.37	1.89+-1.31	1.97+-1.3	1.32+-1.21	1.23+-1.04	1.22+-1.03
	dd	1.25+-0.96	1.3+-0.94	0.69+-0.8	1.3+-0.89	1.22+-0.93	0.45+-0.73	0.43+-0.64	0.43+-0.67
	do	5.03+-2.14	4.91+-2.07	2.85+-1.61	4.64+-1.97	4.43+-2.02	2.01+-1.4	1.88+-1.3	1.91+-1.27
	dg	0.6+-0.27	0.6+-0.27	0.5+-0.25	0.59+-0.25	0.58+-0.27	0.38+-0.3	0.41+-0.25	0.41+-0.26
D	d	19.47+-3.61	18.87+-3.62	12.7+-2.89	16.69+-3.41	16.37+-3.31	8.48+-2.57	7.6+-2.33	7.15+-2.23
	da	4.21+-1.87	3.96+-1.85	1.86+-1.31	3.36+-1.82	3.07+-1.78	0.88+-1.24	0.67+-1.12	0.48+-1.18
	dr	16.19+-3.33	15.83+-3.37	10.65+-2.72	13.82+-3.25	13.55+-3.13	6.49+-2.56	5.72+-2.19	5.21+-2.18
	do	8.96+-2.74	8.79+-2.65	4.13+-1.77	8.5+-2.45	8.23+-2.45	3.6+-1.54	3.28+-1.53	3.45+-1.52
	dg	0.85+-0.24	0.84+-0.25	0.76+-0.28	0.79+-0.3	0.76+-0.34	0.57+-0.4	0.53+-0.41	0.44+-0.49
O	o	18.01+-3.74	17.19+-3.65	8.7+-2.57	15.25+-3.36	14.29+-3.07	4.46+-1.58	3.78+-1.37	3.51+-1.35
	dr	14.84+-3.69	14.01+-3.62	6.69+-2.58	12.2+-3.19	11.6+-2.9	3.61+-1.6	3.04+-1.38	2.82+-1.29
	dd	8.13+-2.55	7.44+-2.41	3.46+-1.53	6.54+-2.42	5.63+-2.36	1.25+-1.15	0.91+-1.06	0.76+-1.07
	da	15.66+-3.44	14.83+-3.29	6.81+-2.24	13.29+-3.2	12.09+-3.05	3.02+-1.51	2.4+-1.41	2.2+-1.44
	dg	0.81+-0.33	0.8+-0.34	0.5+-0.48	0.76+-0.39	0.72+-0.42	0.47+-0.37	0.43+-0.35	0.4+-0.36