

на правах рукописи

ЦЕПИЛОВ ЯКОВ АЛЕКСАНДРОВИЧ

РАЗРАБОТКА И ПРИМЕНЕНИЕ НОВЫХ МОДЕЛЕЙ В
ПОЛНОГЕНОМНОМ АНАЛИЗЕ АССОЦИАЦИЙ

03.02.07 - Генетика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Новосибирск – 2016

Работа выполнена в лаборатории рекомбинационного и сегрегационного анализа Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук» (ИЦиГ СО РАН), г. Новосибирск.

Научный руководитель: доктор биологических наук
Аульченко Юрий Сергеевич

Официальные оппоненты: доктор биологических наук, проф.
Вадим Анатольевич Степанов
ФГБНУ НИИ медицинской генетики,
г. Томск

доктор биологических наук, проф.
Мошковский Сергей Александрович
ФГБНУ НИИ биомедицинской химии имени
В.Н. Ореховича,
г. Москва

Ведущее учреждение: Федеральное государственное бюджетное
учреждение науки Институт общей генетики им.
Н.И. Вавилова Российской академии наук,
г. Москва

Защита диссертации состоится «__» _____ 2016 года на заседании диссертационного совета по защите диссертаций Д 003.011.01 на базе Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук» (ИЦиГ СО РАН), в конференц-зале Института по адресу: проспект Лаврентьева, 10, г.Новосибирск, 630090 Тел (383) 3634906, факс (383) 3331278, email: dissov@bionet.nsc.ru

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте Института: www.bionet.nsc.ru

Автореферат разослан «__» _____ 2016 г.

Ученый секретарь
диссертационного совета,
доктор биологических наук

Т.М. Хлебодарова

Общая характеристика работы

Актуальность

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 [Kochi, Suzuki, Yamamoto, 2014; Polychronakos, Alriyami, 2015; Reitz, 2014].

Несмотря на огромный прогресс, для большинства сложных признаков человека идентифицированные в рамках ПГАА локусы объясняют только часть наследуемости признака. Например, такой количественный признак как рост человека имеет наследуемость порядка 80%, однако суммарный вклад всех 180 достоверно ассоциированных локусов объясняет только 10% дисперсии признака [Lango Allen и др., 2010]. 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 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, корреляция между генотипом и фенотипом может также быть вызвана сопутствующими факторами, связанными как на фенотипом, так и с генотипами различных локусов. При ПГАА генетическая гетерогенность выборки является одним из важнейших сопутствующих факторов. Если анализ не учитывает влияние структуры популяции, тестовая статистика будет завышена [Devlin & Roeder, 1999], что затрудняет статистическую интерпретацию и может привести к ложноположительным результатам. Чтобы избежать ложноположительных выводов при интерпретации результатов ПГАА, необходимо проводить их коррекцию, учитывающую генетическую структурированность (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 [Devlin, Roeder, 1999; Yan, Hou, Yang, 2009; Zang и др., 2007; Zheng и др., 2005; Zheng, Freidlin, Gastwirth, 2006b]. 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 [Devlin, Roeder, Wasserman, 2001; Zheng и др., 2005].

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 XX века были разработаны теории и гипотезы [Fisher, 1928; Haldane, 1930; Ott, 1991; Wright, 1929], 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 метаболизм 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.

Цели и задачи

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.

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 было не известно. Исследование внесло вклад в фундаментальное знание о распространенности неаддитивных эффектов в генетическом контроле уровней метаболитов крови человека.

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.

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 научного сотрудничества. Автор разработал методы коррекции статистики, реализовал эти методы в виде программного продукта и провел анализ неаддитивных эффектов на метаболизме человека. Дизайн вычислительных экспериментов, моделирование, анализ данных и интерпретация полученных результатов были проведены автором.

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.

Публикации

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.

Апробация работы

Результаты данной работы были представлены и обсуждены: на конференции BGRS «Bioinformatics of Genome Regulation and Structure\Systems Biology - BGRS\SB-2012» (Новосибирск, 2012); на конференции AC.CES 2013 «Aachen Conference on Computational Engineering Science» (Ахен, Германия, 2013); на конференции EMGM 2013 «European Mathematical Genetics Meeting» (Лейден, Нидерланды, 2013); на конференции EMGM 2014 «European Mathematical Genetics Meeting» (Кёльн, Германия, 2014).

Структура и объем диссертации

Диссертация состоит из введения, обзора литературы, описания материалов и методов исследования, результатов исследований, обсуждения, выводов и списка цитированной литературы, содержащего 110 источников (в том числе на иностранном языке 107). Объем диссертации составляет 101 страница. Диссертация включает 14 таблиц и 8 иллюстраций, 4 приложения.

Материалы и методы

Настоящая работа была выполнена с использованием симуляционных данных и четырех наборов реальных данных из трех различных исследований.

Реальные данные

Все реальные данные были предоставлены в рамках сотрудничества с другими лабораториями. Для оценки разработанных методов ГК использовались генетические данные, полученные в результате кросс-секционного исследования ERF (Erasmus Rucphen Family study), проводившегося в генетически изолированной популяции юго-западной части Нидерландов (2,699 человек, 1,093,818 SNP, уровень липопротеидов высокой плотности - ЛПВП). Данные исследования KORA (Cooperative Health Research in the Region of Augsburg) использовались в работе дважды: для валидации методов ГК и для поиска неаддитивных эффектов генов (1,785 человек, 1,717,498 SNP, уровень мочевой кислоты и концентрации 151 метаболита в сыворотке крови натошак, а также все возможные отношения их концентраций: всего 22,801 признак). Британское близнецовое исследование TwinsUK (843 человека) использовались в работе для репликации результатов анализа неаддитивных эффектов генов.

Все признаки предварительно были скорректированы на пол, возраст и номер эксперимента, а далее нормализованы с использованием обратно-нормальной трансформации (inverse-normal transformation [Beasley, Erickson, Allison, 2009]).

Моделированные данные

При исследовании статистических свойств методов геномного контроля мы использовали реальные генотипы исследования ERF и моделировали фенотипы. При анализе ошибки первого рода предрасположенность (liability) моделировалась как сумма полигенных эффектов независимых локусов и случайного, нормально распределенного, «средового» эффекта. Для оценки мощности дополнительно к полигенной и средовой компонентам моделировался локус с основным эффектом, объясняющий существенную долю дисперсии признака.

Полногеномный анализ ассоциаций

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 [Aulchenko и др., 2007] для статистической геномики. Результаты ПГАА корректировались с использованием различных методов ГК: стандартного метода, который корректирует статистику делением на коэффициент λ , а также разработанных методов (описанными далее в результатах). Для репликации мы использовали данные исследования TwinsUK.

Результаты и обсуждение

Модели неаддитивных эффектов генов в статистической генетике

Для простоты предположим, что в изучаемом локусе присутствует только два аллеля, и, как следствие, в популяции возможно три генотипа. Наиболее общая, *кододоминантная* (или «генотипическая») модель не накладывает ограничений на возможные значения

признака при условии генотипов. 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 оценивается эффект доминантного аллеля, модель называется *доминантной*, и, наоборот, эффект рецессивного аллеля – *рецессивной*.

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.

ГК для произвольной модели наследования

Мы вывели значение фактора инфляции дисперсии (VIF, variance inflation factor) как функцию, зависящую от частоты аллеля (p), модели наследования (x) и популяционных параметров (N – общий объем, F – коэффициент инбридинга Райта, K – величина, характеризующая общую структурированность выборки больных и здоровых; чем она выше, тем выше структурированность популяции).

Пример того, как выглядит VIF при заданных параметрах популяции $F=0.05$, $N=1000$, $K=10000$ приведен на Рис. 1. На этом рисунке ось x показывает исследуемую модель наследования, которая выражена как степень отклонения ожидаемого значения признака у гетерозигот по сравнению с гомозиготами. Ожидаемое значение признака у гомозигот принимается равным 0 и 1.

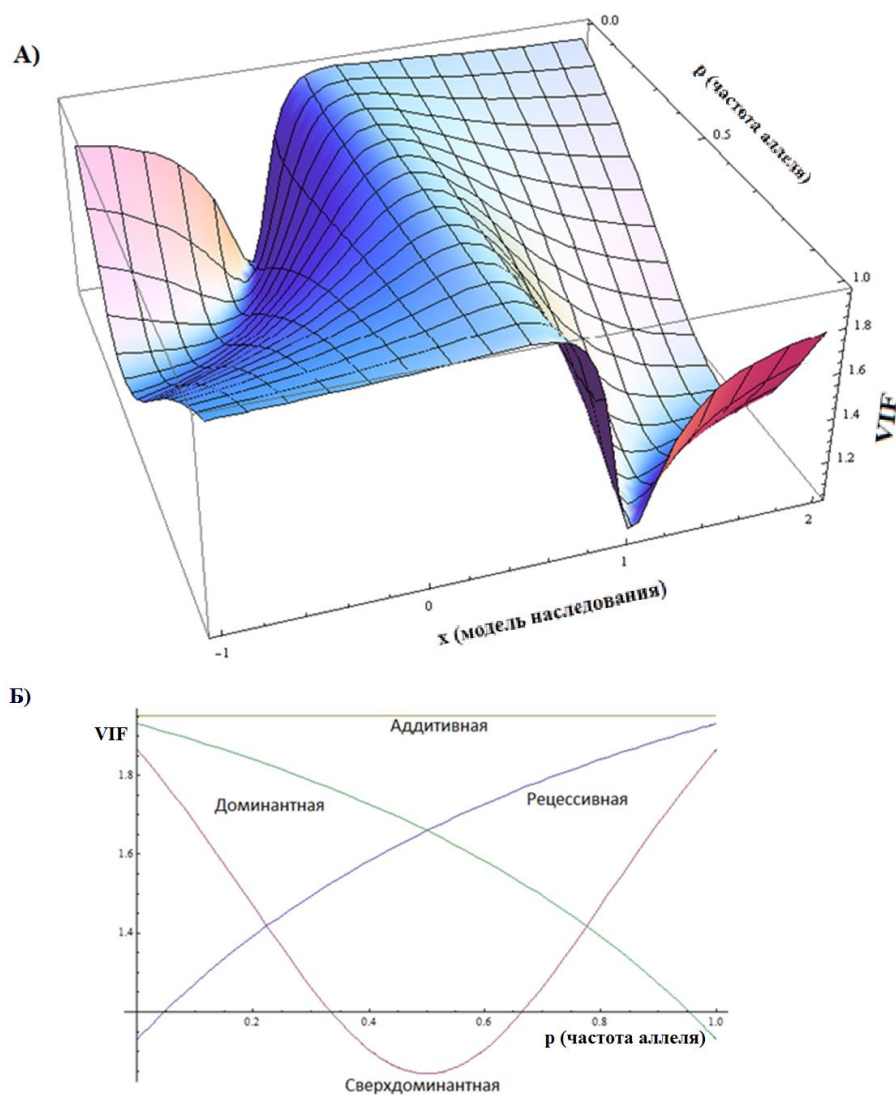


Рисунок 1. А) График VIF при значениях $F=0.05$, $N=1000$, $K=10000$, $x:\{-1,2\}$, $p\{0,1\}$. Б). Крайние случаи VIF при доминантной, рецессивной, аддитивной и сверхдоминантной моделях ($x = 1, 0, \frac{1}{2}, +\infty$ соответственно).

Как видно из графика, VIF не зависит от частоты аллеля только в случае, когда $x=1/2$ (Рисунок 1Б), то есть при аддитивной модели, что уже было показано ранее [Zheng и др., 2005]. Стоит отметить, что при x , стремящемся 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 , стремящемся к минус бесконечности, VIF также стремится к сверхдоминантной модели наследования. Более того, VIF для аддитивной модели всегда больше, чем для неаддитивной, что также подтверждается нами при анализе моделированных данных нескорректированных тестов (Таблица 1). На основании полученных результатов можно сделать вывод, что при использовании неаддитивных моделей коррекция с помощью деления на константу способна контролировать только 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 имеет высокую частоту, и к консервативным результатам теста для SNP для которых доминантно-кодированный аллель имеет низкую частоту (Рисунок 1Б). Этот вывод подтверждается результатами анализа модельных данных (Таблица 1). Коррекция при помощи константы, как правило, сохраняет уровень ошибки первого рода для всех маркеров, однако для конкретных частотных групп это не соблюдается.

Методы оценки VIF требуют знания параметров, описывающих структуру популяции. Если эти параметры (в нашем случае F и K) неизвестны, их можно оценить, если предположить, что распределение тестовой статистики должно соответствовать $\chi^2_{df=1}$ после коррекции. Таким образом, оценка неизвестных параметров возможна за счет минимизации функции ошибки, описывающей отклонение наблюдаемого распределения от предполагаемого. Метод геномного контроля, основанный на описанной выше методологии, был назван нами VIFGC. Следует отметить, что этот метод применим только для коррекции результатов одностепенного тестирования ограниченных моделей. В тоже время тест кодоминантной модели может быть представлен как функция (скорректированных) тестов ограниченных моделей.

Также нами был предложен полиномиальный ГК (Polynomial Genomic Control, PGC) для неаддитивных моделей, который приближает функцию коррекции λ как полином l -степени от частоты аллеля p . 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 должна быть распределена как χ^2 . Этот метод был назван нами PGC. Отметим, что PGC может применяться как для коррекции результатов тестирования ограниченных моделей, так и для коррекции результатов тестирования кодоминантной модели.

Статистические характеристики предложенных методов ГК

Для оценки ошибки первого рода и мощности методов использовались моделированные 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 $p\text{-value} \leq 0.05$. Также нами проведена оценка ошибки первого рода для каждой из 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].

Результаты исследования ошибки первого рода ограниченных тестов представлены в Таблице 1. 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 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 аллельных частот.

При анализе альтернативной гипотезы, как и ожидалось, наибольшей мощностью обладали модели, предполагавшие «правильную» модель наследования. Кодоминантный тест имеет меньшую мощность, чем ограниченный тест использующий корректную

моделью, но оказывается более устойчивым при неправильном выборе ограниченной модели наследования.

Использование двух независимых когорт, KORA и ERF, позволило тестировать наши методы на реальных данных. Для обоих случаев полученные выводы совпадают с таковыми, полученными в симуляционных исследованиях.

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 1). 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.

Апробация разработанных методов: исследование неаддитивных эффектов генов на метаболом человека

Двухэтапный подход к идентификации неаддитивных эффектов

Нами предложен и апробирован двухшаговый подход к идентификации неаддитивных эффектов в рамках ПГАА. На первом шаге ПГАА проводится с помощью двухступенного теста для кодоминантной модели, которая не накладывает ограничений на модель наследования. Результаты корректируются с помощью разработанных нами методов ГК. На втором шаге, для каждого значимого локуса, определенного на первом этапе, кодоминантная модель сравнивается с ограниченными моделями (рецессивной, доминантной, аддитивной и сверхдоминантной) с использованием критерия отношения правдоподобия (likelihood ratio test, LRT). Ограниченная модель принимается в том случае, если она незначимо отличается от модели кодоминантного контроля. Если отвергаются все ограниченные модели – то есть данные описываются ограниченными моделями значимо хуже – принимается кодоминантная (общая) модель. Если не отвергается только одна ограниченная модель, такая модель принимается как наиболее парсимонная. В тоже время, четко определить наиболее парсимонную модель только с использованием LRT не всегда представляется возможным. Поэтому как дополнительный критерий, позволяющий ранжировать модели, мы предложили использовать информационный критерий Акаике (Akaike information criterion, AIC) [Akaike, 1974], который позволяет сравнивать невложенные модели. Следует отметить, что хотя отдельные этапы предложенного нами подхода и не являются принципиально новыми, однако, изложенный выше подход для систематической идентификации неаддитивных эффектов является новым.

Мы использовали предложенный нами подход для изучения неаддитивных эффектов в ПГАА концентраций метаболитов крови человека на материале исследования KORA (n=1,785 человек с данными по 151 метаболиту и генотипированных по 1,717,498 SNP).

ПГАА с использованием неаддитивных моделей

При анализе с использованием кодоминантной модели мы идентифицировали двадцать локусов, которые были достоверно ассоциированы как минимум с одним метаболитом или отношением метаболитов. При коррекции уровня значимости на

множественное тестирование применялся метод Бонферрони. Граница значимости была установлена как $5 \times 10^{-8} / 22801 \approx 2.19 \times 10^{-12}$, где 5×10^{-8} – общепринятая граница значимости ПГАА при анализе одного признака, а 22801 – число проанализированных нами признаков.

Достоверно ассоциированные локусы представлены в Таблице 2. Шестнадцать ассоциаций были реплицированы на данных TwinsUK (p-value, с учетом коррекции Бонферрони, $< 0.05/20 = 0.0025$). Для каждого найденного SNP мы сравнили ограниченные модели с кодоминантной моделью с помощью LRT, а также провели сравнение моделей с помощью AIC. Для шестнадцати из идентифицированных локусов кодоминантная модель не отличалась значимо от аддитивной модели (указаны в Таблице 2 с «а» 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.

Для четырех оставшихся локусов аддитивная модель была значимо хуже кодоминантной. При этом для локуса заданного SNP 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, где информация о генотипах по этому SNP не была доступна.

Далее мы эмпирически сравнили предложенный нами подход, в котором поиск потенциально неаддитивных локусов осуществляется с использованием кодоминантной модели, с альтернативным подходом, предполагающим использование ограниченных неаддитивных моделей для полногеномного анализа. Мы провели ПГАА для рецессивной, доминантной и сверхдоминантной моделей. Даже с использованием либерального уровня значимости $5 \times 10^{-8} / 22801$, 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 локусов и одну дополнительную ассоциацию между rs219040 на седьмой хромосоме (p-value $< 3.94 \times 10^{-13}$) 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. Однако, нам не удалось реплицировать эту ассоциацию на данных исследования Twins UK (p-value < 0.78).

Таким образом, нам не удалось идентифицировать дополнительные неаддитивные локусы с использованием ограниченных моделей, что говорит о том, что предложенный нами метод скрининга с использованием кодоминантной модели может являться оптимальным.

Следует отметить, что ПГАА с использованием аддитивной модели позволил определить все двадцать локусов, которые были обнаружены кодоминантной моделью (частично представленные другими SNP и отношениями метаболитов). Более детальное рассмотрение полученных результатов (Таблица 2) позволяет утверждать, что для неаддитивных локусов результаты были, как правило, более значимы с использованием кодоминантной модели. Неаддитивные локусы характеризовались очень большими эффектами, высокой частотой минорного аллеля, и более или менее выраженным неполным доминированием. Именно такая модель контроля, не являясь аддитивной, по-прежнему предоставляет возможность картирования с использованием аддитивной модели. Хотя при этом и теряется мощность анализа, однако, при условии достаточно больших эффектов – что наблюдается в анализе метаболитов – значимость по-прежнему может достигать полногеномной. Однако, аддитивная модель будет терять мощность, если рецессивный аллель является редким, при условии полной доминантности, а особенно – при сверхдоминантности. Следовательно, использование аддитивной модели для скрининга потенциально неаддитивных локусов не рекомендуется.

Таким образом, нами сформулирована и апробирована стратегия идентификации неаддитивных эффектов с использованием кодоминантной модели и LRT/AIC для определения модели наследования значимо ассоциированных локусов. Предложенная нами стратегия может быть использована и в других полногеномных исследованиях.

Кодоминантная модель может быть также использована при проведении ПГАА, особенно когда известно, что значительная часть ожидаемых генетических эффектов – неаддитивна [Loley и др., 2013; Zheng, Freidlin, Gastwirth, 2006a]. 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 этот подход с анализом всех ограниченных моделей. Для фенотипических данных с высокой размерностью, таких как метаболомные, это соображение может быть важным.

Локусы с неаддитивными эффектами

Для шестнадцати локусов, найденных в нашем анализе, модель наследования не отличалась значимо от аддитивной. В тоже время четыре локуса обладали значимыми неаддитивными эффектами. Мы обнаружили два локуса (rs6970485, rs715), которые 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. 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 [Hill, Goddard, Visscher, 2008; Powell и др., 2013]. Ранее было показано, что наблюдение преимущественно аддитивных генетических эффектов может быть следствием искажения модели в сторону аддитивности в случае неполного неравновесия по сцеплению (linkage disequilibrium, LD) между функциональным и маркерным аллелями [Vukcevic и др., 2011; Zondervan, Cardon, 2004]; 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. Таким образом, в исследованиях, использующих генетические данные с более высоким геномным покрытием (например, данные геномного секвенирования), мы можем ожидать более выраженные неаддитивные эффекты.

Мы наблюдали, что SNP с неаддитивными эффектами могут быть обнаружены и при использовании аддитивной модели. Аддитивная модель может рассматриваться как приближение к рецессивной/доминантной моделям, которая обладает адекватной мощностью в случае превалирования рецессивного аллеля большого генетического эффекта. Следует отметить, что не все локусы, которые были определены с помощью аддитивной модели, могли быть определены с помощью кодоминантной модели. Это значит, что при идентификации локусов, влияющих на концентрации метаболитов, в рамках ПГАА предположение об аддитивности генетических эффектов приемлемо.

Мы можем предположить, что и для других «омиксных» фенотипов (транскриптомных, гликомных, протеомных и т.д.) генетический контроль, в основном, осуществляется аддитивно. Логично предположить, что если теории С. Райта и Касисера и Бернса [Kacser, Burns, 1981; Wright, 1929] 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, это не так. Поэтому для других «омик», скорее всего, стоит ожидать еще меньшую степень неаддитивности; например, это уже было показано для транскриптомики [Powell и др., 2013]. Однако, этот вопрос требует дополнительных исследований.

Заключение

В результате проделанной работы 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.

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.

Таблица 1. 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; λ_{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.

Модель	Частоты	Некорр.			Константная корр.			VIFGC корр.			PGC корр.		
		λ_{median}	λ_{regress}	E	λ_{median}	λ_{regress}	E	λ_{median}	λ_{regress}	E	λ_{median}	λ_{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 2. Результаты анализа неаддитивных эффектов. Результаты ПГАА для кодоминантной и аддитивной модели для двадцати найденных локусов. Локусы, которые были реплицированы на данных TwinsUK, отмечены звездочкой в последнем столбце. 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. Таблица разделена на 2 части: в верхней части представлены локусы, которые уже описаны в ранее опубликованных ПГАА на тех же данных (Illig et al. 2010), в нижней части приведены новые локусы.

SNP	метаболит/отношение	A/a	AF	кодоминантная модель			аддитивная модель		KORA		TwinsUK		Реплик.
				beta(AA) (se)	beta(Aa) (se)	p-value	beta (se)	p-value	LRT	AIC	LRT	AIC	
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.

Список публикаций по теме диссертации

Статьи в журналах из списка ВАК:

1. **Tsepilov, Y.A.**, Ried, J.S., Strauch, K., и др. Nonadditive Effects of Genes in Human Metabolomics. // *Genetics*. 2015. Т. 200. № 3. С. 707–18., IF=6.0
2. **Tsepilov, Y.A.**, Shin, S.-Y., Soranzo, N., и др. Development and application of genomic control methods for genome-wide association studies using non-additive models. // *PLoS One*. 2013. Т. 8. № 12. e81431. IF=3.5

Тезисы конференций:

1. **Tsepilov Y.A.**, Shin S., et al., Genome-wide scan of metabolomics data using non-additive intra-locus models. // EMGM 2014, European Mathematical Genetics Meeting, Cologne, Germany, 2014. - Vol. 77. - P. 94
2. **Tsepilov Y.A.**, Ried J.S., et al., Genomic Control for genome-wide association analysis under non-additive models.// EMGM 2013, European Mathematical Genetics Meeting, Leiden, The Netherlands, 2013.- Vol. 76. - P. 94.
3. **Tsepilov Y.A.**, Ried J.S., et al., Exploration of Metabolomics Space with Genome-Wide Association Analysis.// AC.CES 2013, Aachen Conference on Computational Engineering Science, Aachen, Germany, 2013. - URL: <http://www.ac-ces.rwth-aachen.de/> (14.04.2016).
4. **Tsepilov Y.A.**, Ried J.S., et al., Development and application of the Genomic Control methods for genome-wide association analysis using non-additive models.// BGRS 2012, Novosibirsk, Russia, 2012. – Vol. 1. – P. 319