На правах рукописи

Аульченко Юрий Сергеевич

Разработка и применение методов полногеномного анализа генетических ассоциаций сложных признаков

03.02.07 – генетика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
доктора биологических наук

Новосибирск
2010

Работа выполнена в лаборатории рекомбинационного и сегрегационного анализа Учреждения Российской академии наук Институт цитологии и генетики Сибирского отделения РАН, г. Новосибирск, Россия

Официальные оппоненты:	доктор биологических наук, профессор Маркель А. Л. Институт цитологии и генетики СО РАН, г. Новосибирск
	доктор биологических наук, профессор Поляков А. В. Медико-генетический научный центр РАМН, г. Москва
	доктор биологических наук, профессор Гуляева Л. Ф. Научно-исследовательский институт молекулярной биологии и биофизики СО РАМН, г. Новосибирск
Ведущее учреждение:	Учреждение Российской академии Медицинских наук НИИ медицинской генетики СО РАМН, г. Томск

Защита диссертации состоится "24" ноября 2010 г. на утреннем заседании диссертационного совета Д 003.011.01 при Институте цитологии и генетики СО РАН по адресу: 630090, Новосибирск, Россия, пр. ак. Лаврентьева, 10, тел/факс (383) 333-12-78, e-mail: dissov@bionet.nsc.ru

С диссертацией можно ознакомиться в библиотеке Института цитологии и генетики СО РАН

Автореферат разослан "12" октября 2010 г.

Ученый секретарь
диссертационного совета,
доктор биологических наук Т. М. Хлебодарова

Общая характеристика работы

Актуальность

Идентификация генов и аллелей, контролирующих разнообразие сложных признаков, является важной теоретической и прикладной задачей генетики. 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, находит применение в таких областях как криминалистика и медицина. Для сельскохозяйственных и домашних животных знание аллельных вариантов позволяет вести направленную эффективную селекцию.

Исходная популяция, из которой формируется выборка для изучения генетического контроля признаков, может быть инбредной (модельные объекты), либо аутбредной. По структуре, выборки подразделяют на фиксированные (направленные скрещивания инбредных линий, некоторые скрещивания сельскохозяйственных или домашних животных) и произвольные, т.е. такие, структура которых не находится под контролем исследователя. В данной работе в качестве материала для генетического анализа мы рассматриваем выборки произвольной структуры, полученные из аутбредных популяций человека, сельскохозяйственных и домашних животных. При этом предполагается, что выборка формируется из одной исходной популяции.

Существует несколько подходов к идентификации генов и аллелей в выборках произвольной структуры, полученных из аутбредных популяций. Один из подходов предусматривает тестирование генов-кандидатов, т.е. исследование ассоциации признака с аллелями гена, участие которого в формировании признака можно предположить на основании накопленных данных. Ясно, что основным недостатком этого метода является невозможность получения принципиально новой информации о биологии признака, так как метод существенно ограничен уже имеющимися знаниями.

Этот недостаток подхода, основанного на поиске генов-кандидатов, устраняется при проведении полногеномного картирования. При этом подходе для поиска локусов, контролирующих признак, используется большое количество маркеров, покрывающих весь геном. Исторически, первым широко применяющимся методом полногеномного анализа стал анализ сцепления.

При анализе сцепления выборка, состоящая из близких родственников с известными значениями исследуемого признака, генотипируется с применением панели из 200–10,000 полиморфных маркеров, покрывающих весь геном. Далее, анализируется совместное наследование (сцепление, или косегрегация) признака и маркерных генотипов. Значимое сцепление указывает регион (как правило, довольно большой – от двух до пятидесяти миллионов пар оснований), содержащий ген, высокопенетрантные аллели которого оказывают влияние на изучаемый признак. Метод анализа сцепления хорошо зарекомендовал себя при идентификации генов простых менделевских признаков. Хотя с начала 1990-х годов этот метод широко применялся для изучения сложных признаков человека, он дал удручающе мало результатов. Идентифицированные гены, как правило, объясняли малую долю случаев болезни, так как метод позволяет картировать в основном относительно редкие «менделевские» семейные формы сложных заболеваний.

Одним из наиболее перспективных современных методов, применяемых для идентификации локусов, контролирующих сложные признаки, является метод полногеномного анализа ассоциаций (Genome-Wide Association, GWA). При проведении этого анализа сотни тысяч однонуклеотидных полиморфизмов (SNP), распределенных по всему геному, типируются в группах людей с известным значением изучаемого признака. Анализ ассоциации между распределением генотипов и фенотипов позволяет установить связь между аллельной вариацией в некотором регионе генома и исследуемым признаком.

В момент начала работы по теме данной диссертации метод полногеномного анализа ассоциаций ещё не являлся общепринятым методом исследования генетики сложных признаков человека и аутбредных животных. Необходимо было оценить теоретическую возможность таких исследований, рассмотреть вопросы наиболее эффективного формирования выборки, разработать методы статистического анализа полногеномных данных и создать пакеты прикладных программ, позволяющих осуществлять такой анализ. Именно этим теоретическим вопросам, а также апробации полученных методических разработок на реальных данных посвящена данная диссертация.

Цели и задачи исследования

Целью исследования является разработка методов полногеномного анализа ассоциаций в выборках произвольной структуры из аутбредных популяций, апробация этих методов на реальных данных и

идентификация новых локусов, контролирующих сложные, в том числе патологические, признаки человека. Для достижения цели были поставлены следующие задачи:

1. Исследовать возможные стратегии формирования выборки для картирования локусов, контролирующих сложные признаки человека методом полногеномного анализа ассоциаций. В частности, сравнить преимущества и недостатки формирования выборки из представителей молодых генетически изолированных и открытых популяций человека.
2. Разработать методы и программное обеспечение для проведения полногеномного анализа ассоциаций на материале выборок произвольной структуры из аутбредных популяций человека, сельскохозяйственных и домашних животных.
3. Провести апробацию разработанных методов и программного обеспечения на примере полногеномного анализа уровней липидов крови и роста человека; определить новые локусы, участвующие в контроле этих признаков.
4. Оценить прогностический потенциал геномных данных для предсказания значений количественных признаков (на примере уровней липидов крови и роста человека) и вероятности развития определенного фенотипа (на примере гиперхолестеринемии и крайних вариантов роста).

Научная новизна

Показано, что в молодых генетически изолированных популяциях эффект дрейфа генов, выражающийся в потере или существенном увеличении частоты некоторых аллелей, выражен для аллелей с начальной частотой <1% и мало заметен для аллелей с начальной частотой ≥5%.

Показано, что длины карт неравновесия по сцеплению для молодых генетически изолированных популяций на ~30% меньше, чем для открытой популяции человека, что увеличивает мощность идентификации генетических локусов, ассоциированных с изучаемыми признаками.

Разработаны новые методы анализа ассоциации в рамках модели «измеренных генотипов», позволяющие анализировать выборки произвольной структуры из аутбредных популяций человека, сельскохозяйственных и домашних животных. Эти методы являются статистически мощными и вычислительно эффективными.

Создан пакет эффективных компьютерных программ для полногеномного анализа ассоциаций количественных и бинарных

признаков в популяционных и семейных выборках человека и аутбредных животных.

В ходе апробации разработанных методов и пакетов программ, проведен полногеномный анализ ассоциаций уровней липидов в крови человека. Впервые подобный анализ проведен на популяционных выборках, а в набор картируемых характеристик липидного обмена введен уровень общего холестерина. Идентифицировано шесть новых локусов, контролирующих уровни липидов в крови человека. Также проведен полногеномный анализ роста человека и идентифицирован новый локус, *JAZF1*, контролирующий этот признак.

Показано, что геномный профиль роста объясняет 4–6% дисперсии этого признака. Геномные профили липидов объясняют существенную долю (1–7%) дисперсии этого признака; эта доля сравнима с таковой, объясняемой индексом массы тела.

Впервые показано, что геномный профиль общего холестерина является предиктором дислипидемии и статистически значимым, независимым от уровня циркулирующего холестерина, предиктором толщины комплекса интима-медиа сосудистой стенки. Из всех геномных профилей классических липидов, геномный профиль холестерина является наилучшим предиктором дислипидемии, ишемической болезни сердца и атеросклероза.

Теоретическая и практическая ценность

Полученные нами оценки вероятности потери аллелей, коэффициентов вариации частот аллелей, а также длины карты неравновесия по сцеплению позволили сделать важный теоретический вывод о том, что мощность метода полногеномного анализа ассоциаций в среднем выше, если используются выборки из молодых генетически изолированных, а не из открытых популяций. Далее, мы показали, что результаты анализа ассоциаций, проведенного на выборках из молодых генетически изолированных популяций в целом должны совпадать результатами, полученными в открытых популяциях человека. Эти выводы предоставили теоретическое обоснование для дальнейших практических полногеномных исследований сложных признаков с использованием генетически изолированных популяций человека (например, консорциумы EUROSPAN, ENGAGE, CHARGE, другие исследования). В настоящее время использование материала из молодых генетически изолированных популяций для верификации результатов, полученных на материале открытых популяций (и наоборот) является общепринятым.

Разработанные нами новые методы и пакеты программ широко применяются в исследованиях генетики сложных признаков человека, сельскохозяйственных и домашних животных. В частности, список зарегистрированных пользователей нашего пакета GenABEL составляет более 250 человек; статьи, представляющие результаты наших методических разработок, цитировались более 70 раз.

Наши полногеномные исследования контроля уровня холестерина в крови человека доказали важность этого признака и частично сместили акценты в исследованиях геномного контроля липидов; так, на основании этого результата консорциум GLGC (Global Lipids Genetics Consortium) включил уровень общего холестерина в список исследуемых характеристик.

Идентифицированные нами новые локусы, участвующие в контроле уровня липидов в крови и роста человека, расширили представления о механизмах контроля нормальной вариации этих признаков в популяциях человека. Кроме того, мы идентифицировали новый локус роста тела, *JAZF1*, обладающий плейотропным действием, что расширяет имеющиеся представления о механизмах генетического контроля роста и связи между раком, аутоиммунными заболеваниями и ростом тела.

Полученные результаты используются в нескольких курсах, преподаваемых в НГУ и медицинском центре «Эразмус» (GE03, "Advances in population-based analysis"; GE05, "Family-based genetic analysis"), а также использовались в лекциях на школах молодых ученых, проходивших в Москве, Уфе и Томске.

Положения, выносимые на защиту

1. Полногеномный анализ ассоциаций, проводимый с использованием разработанных нами алгоритмов и пакетов программ, является мощным и воспроизводимым методом идентификации эффектов распространенных аллелей.
2. Молодые генетически изолированные популяции являются лучшим ресурсом для выявления и изучения как распространенных, так и редких аллелей, влияющих на изучаемые сложные признаки.
3. Геномный профиль холестерина является наилучшим геномным предиктором дислипидемии, ишемической болезни сердца и атеросклероза.
4. Разработанный нами метод GRAMMAR-GC является статистически мощным и вычислительно эффективным методом

полногеномного анализа ассоциации в выборках особей, связанных родством.

5. Локус *JAZF1* помимо участия в контроле диабета второго типа, рака простаты и системной красной волчанки также принимает участие в детерминации роста.

Личный вклад автора

В диссертации представлены результаты, полученные автором в сотрудничестве с учеными из России и многих стран мира.

Все экспериментальные данные (выборки, генотипирование, фенотипирование) получены в рамках сотрудничества. Автор принимал активное участие в планировании выборки ERF, являвшейся одним из основных ресурсов при проведении данной работы. Во всех исследованиях, представленных в данной работе, автор выступал в качестве руководителя и/или основного исполнителя.

Апробация работы

Результаты работы, изложенной в данной диссертации, были представлены на следующих конференциях и симпозиумах:

- WEON (Werkgroep Epidemiologisch Onderzoek Nederland), (Rotterdam, The Netherlands, 2003). Presentation "Linkage disequilibrium in recently isolated Dutch population"
- 6th International Conference "Health Insurance in Transition" (Dubrovnik, Croatia, 2003). Invited talk "ERF study: Erasmus family research in isolated population"
- 9th Quantitative Trait Loci / Marker Assisted Selection Workshop (Rostock, Germany, 2004). Invited talk "Preliminary analysis of the Erasmus Rucphen Family Study"
- Haplotype Sharing Workshop, (Heidelberg, Germany, 2006). Invited talk "Haplotype sharing, linkage disequilibrium and complex genealogies"
- 9я школа-семинар по популяционной генетике (Уфа, 2006). Доклад "Методы генетической эпидемиологии сложных признаков человека"
- VIII научная конференция «Генетика человека и патология» (Томск, 2007). Доклад «Методы картирования комплексных признаков человека»
- 58th Annual Meeting of American Society of Human Genetics (Philadelphia, USA, 2008). Platform presentation "First neuronally expressed gene associated with multiple sclerosis."
- European Mathematical Genetics Meeting (Munich, Germany, 2009). Invited talk «Predicting human height by Victorian and post-genomic methods»

- Dutch human genetics society meeting (Veldhoven, The Netherlands, 2009). Presentation «Genome-wide association analysis of 16 European populations identifies novel loci influencing lipid levels»
- Genetics of complex diseases in isolated populations (Trieste, Italy, 2009). Invited talk «Meta-analysis of genome-wide association scans»
- V Съезд Вавиловского общества генетиков и селекционеров (Москва, 2009). Доклад «Количественная интегративная геномика сложных признаков человека»
- European Mathematical Genetics Meeting (Oxford, UK, 2010). Invited talk «Challenges in statistical genomics of complex human traits»

Публикации по теме диссертации

Результаты работы, изложенной в данной диссертации, были опубликованы в виде 37 статей в рецензируемых научных журналах, в том числе в «New England Journal of Medicine», «Nature Genetics», «American Journal of Human Genetics», «PLoS Genetics», «Human Molecular Genetics».

Объём и структура диссертации

Диссертация состоит из пяти глав. Объем диссертации составляет 290 страниц, диссертация включает 34 таблицы и 25 иллюстраций.

Результаты исследований и их обсуждение

Аллельный спектр и структура неравновесия по сцеплению в популяциях человека

Мощность картирования с помощью анализа ассоциаций в большой мере зависит от частот аллелей, контролирующих болезнь, и от степени неравновесия по сцеплению (linkage disequilibrium, LD) между ними и аллелями маркерных локусов (MULLER-MYHSOK and ABEL 1997). Последнее в значительной степени определяется возрастом мутаций, историей, размером и структурой исследуемой популяции.

Для будущих проектов картирования с помощью LD важно знать ожидаемые частоты маркерных аллелей, а также величину и геномный паттерн LD в различных популяциях. Распределение неравновесия по сцеплению является предметом активных дебатов и широко изучается в различных популяциях человека (EAVES *et al.* 2000; SERVICE *et al.* 2001; LONJOU *et al.* 2003; VARILO *et al.* 2003). В эмпирических исследованиях было показано, что характер снижения LD при увеличении генетического расстояния не всегда в точности соответствует

ожидаемому на основе стандартных моделей популяционной генетики. Описаны примеры слишком низкого, по сравнению с ожидаемым, LD на расстоянии нескольких тысяч пар оснований и очень высокого LD на значительно больших расстояниях (WEISS and CLARK 2002). Другие исследования показали, что LD варьирует между популяциями и что распределение LD нерегулярно в пределах генома (COLLINS *et al.* 1999; ABECASIS *et al.* 2001). Таким образом, прежде чем приступать к картированию генов методом полногеномного анализа ассоциаций, необходимо описать и сравнить LD в разных популяциях.

В открытых популяциях человека велика генетическая и средовая гетерогенность, и поэтому необходимо включать в анализ очень большие выборки (HEUTINK and OOSTRA 2002). Размер выборки можно уменьшить, если анализировать материал из генетически изолированной популяции, где средовое разнообразие меньше, а генетический фон более гомогенный (SHEFFIELD *et al.* 1998; CHAKRABORTY and DEKA 2002). Дрейф генов и эффект основателя в целом снижают генетическое разнообразие в изолированных популяциях. Однако некоторые мутации, редко встречающиеся в других популяциях, в генетических изолятах могут стать довольно частыми. Например, в популяции Финляндии с высокой частотой выявляются наследственные расстройства и аллельные варианты, которые больше практически нигде не встречаются (NORIO *et al.* 1973; PIHLAJA *et al.* 2003). С одной стороны, частота этих аллелей в популяции Финляндии относительно высока, и это позволяет изучать генетическую детерминацию таких признаков с высокой статистической мощностью. С другой стороны, перечисленные особенности популяции являются её недостатком, так как обнаруженные аллели не могут быть использованы для предсказания риска болезни в других популяциях.

Другим преимуществом изучения генетически изолированных популяций является то, что неравновесие по сцеплению может быть обнаружено на больших расстояниях. Однако степень неравновесия по сцеплению и генетическое разнообразие варьируют в разных изолированных популяциях. В результате некоторые из популяций больше подходят для полногеномного анализа ассоциаций, чем другие (WRIGHT *et al.* 1999).

В Европе существует большое число молодых генетически изолированных популяций, изоляция которых обусловлена религиозными причинами, а период основания совпадает с периодом реформации (XVIII век). Как правило, эти популяции характеризуются высокой степенью изоляции и экспоненциальным ростом в течение последних 150–200 лет. Такие популяции могут характеризоваться

измененным аллельным спектром и повышенным LD, и, таким образом, представлять ценный ресурс для картирования генов комплексных признаков. Однако генетическим исследованиям таких популяций уделялось до недавнего времени мало внимания.

Мы рассмотрели вопрос, насколько аллельный спектр и структура неравновесия по сцеплению молодых генетически изолированных популяций человека отличаются от таковых в открытых популяциях. Сравнение аллельного спектра позволяет определить насколько генетические результаты, полученные в изолированных популяциях, экстраполируемы на открытые популяции и обратно. Сравнение структуры неравновесия по сцеплению позволяет ответить на вопрос об относительной эффективности использования различных популяций человека для картирования генов с помощью полногеномного анализа ассоциаций.

При изучении эффекта дрейфа генов в качестве примера молодой генетически изолированной европейской популяции нами была использована популяция из Нидерландов, изучаемая в рамках программы GRIP (Genetic Research in Isolated Populations). Однако полученные результаты применимы к большому числу изолятов со сходной популяционной историей. Используя компьютерное моделирование и косвенные эмпирические данные, мы показали, что в популяции GRIP снижено генетическое разнообразие (Таб. 1). Это повышает мощность генетического анализа. Кроме того, мы показали, что для определенной доли аллелей, которые редки в открытой популяции, в генетически изолированных популяциях частота может быть резко повышена за счет стохастических причин. Если такие аллели обладают функциональным эффектом на фенотип или находятся в LD с функциональными вариантами, повышенная частота будет транслироваться в увеличение мощности их идентификации.
Мы показали, что аллели, распространенные (частота $\geq$5%) в исходной популяции сохраняют высокую частоту как в молодых генетических изолятах, так и в открытых популяциях. Применяющиеся в настоящее время ДНК-чипы содержат именно распространенные полиморфизмы. Поэтому следует ожидать, что при использовании таких чипов большинство найденных ассоциаций будут сходны между молодыми генетически изолированными популяциями и большими открытыми популяциями того же происхождения. Следовательно, результаты полногеномного анализа ассоциаций, проведенного в молодых генетически изолированных популяциях, могут быть обобщены на открытую популяцию, и наоборот.

Таб. 1. Распределение частот аллелей в последних поколениях родословной ERF при различных начальных частотах p_0.

Начальная частота (P_0)	Средняя	Медиана	S.D	S.D ***	Минимум	Максимум	95% C.I.	Коэфф. вариации**	Потеря***	Потеря/возрастание частоты аллеля в* Потеря	> 2 раза	> 5 раз	> 10 раз
0.0001	0.0001	0	0.0012	0.001	0	0.0626	0.000–0.001	12	0.973	0.954	0.046	0.046	0.026
0.001	0.001	0	0.0031	0.002	0	0.0631	0.000–0.009	3.1	0.759	0.615	0.115	0.0483	0.02
0.01	0.0099	0.0068	0.0095	0.007	0	0.0883	0.001–0.037	0.9	0.064	0.0083	0.1178	0.0068	0
0.025	0.0304	0.0268	0.0163	0.011	0.0021	0.1104	0.008–0.072	0.5	0.001	0	0.11	0	0
0.05	0.0499	0.0468	0.0207	0.015	0.0089	0.1729	0.019–0.098	0.42	0	0	0	0	0
0.1	0.1002	0.0967	0.0292	0.021	0.0278	0.2634	0.053–0.167	0.29	0	0	0	0	0
0.25	0.2496	0.2482	0.0413	0.030	0.1215	0.4306	0.175–0.335	0.16	0	0	0	0	0
0.5	0.5003	0.5	0.0468	0.035	0.3333	0.6666	0.407–0.591	0.09	0	0	0	0	0

* – оцененная в численном эксперименте частота потери аллеля / возрастания его начальной частоты в определенное число раз; ** – коэффициент вариации, оцененный как стандартное отклонение (SD), деленное на среднее значение; *** – SD и вероятности потери, оцененные аналитически на основе популяционно-генетической теории.

Однако разница в структуре LD может привести к различию в мощности анализа в этих двух типах популяций. Поэтому далее мы анализировали эмпирические данные по генотипам полиморфных маркеров для характеризации LD в ряде генетически изолированных популяций человека.

Мы изучали LD в популяции GRIP с помощью высокополиморфных микросателлитных маркеров и провели сравнение с молодыми изолированными популяциями Палау, Микронезии (DEVLIN *et al.* 2001) и Центральной Долины Коста-Рики (SERVICE *et al.* 2001). В этих популяциях, а также в более старых популяциях, подверженных сильному генетическому дрейфу (саамы и гавои, (VARILO *et al.* 2000; ZAVATTARI *et al.* 2000)) распределение LD было сходным. Для синтенных локусов, неравновесие по сцеплению было найдено на больших расстояниях, что подчеркивает ценность молодых генетически изолированных популяций для картирования генов. Неравновесие по сцеплению было меньше и убывало с расстоянием быстрее в открытой популяции Великобритании и в более старых изолятах большого размера, претерпевших экспоненциальное расширение (Сардиния, Финляндия) (VARILO *et al.* 2000; ZAVATTARI *et al.* 2000).

В принципе, смешение с другими популяциями и дрейф могут приводить к «ложному» LD между несцепленными локусами, затрудняя полногеномный анализ ассоциаций. Однако для популяции GRIP нами было показано отсутствие статистически значимого LD между несцепленными локусами.

Далее, мы сконструировали метрические карты неравновесия по сцеплению для одиннадцати молодых и старых генетических изолятов различного размера, а также для открытой популяции (Таб. 2). В целом, сравнение двенадцати популяций демонстрирует, что изолированные популяции, недавно пережившие период быстрого роста и берущие начало от небольшого числа основателей, имеют более высокий общий уровень LD, чем открытые популяции, а также имеют гораздо меньше районов очень низкого LD. Было показано, что в таких популяциях карта LD на ~20–45% короче, чем в открытых популяциях. Таким образом, следует ожидать, что при использовании одной и той же панели маркеров геномное покрытие в генетически изолированных популяциях будет лучше, чем в открытой популяции, приводя к аналогичному (~20–45%) повышению ожидаемой мощности полногеномного анализа ассоциаций. Принимая во внимание большой масштаб полногеномных исследований (тысячи образцов, генотипирование каждого из которых может быть довольно дорого),

Таб. 2. Карта LD хромосомы 22 для двенадцати популяций.

Популяция	Длина карты LD в LDU*	Отношение LDU/Мб	Число пробелов LD	Общий размер пробелов LD
Антиокия, Колумбия	581.9	17.01	31	1,092
Ашкенази	656.5	19.19	26	975
Азоры	864.5	25.27	84	2,709
Открытая популяция	845.1	24.70	84	2,574
Центральная Долина Коста-Рики	572.1	16.72	23	821
Юго-восток Нидерландов	620.8	18.15	29	1,166
Северная Финляндия	523.9	15.31	21	821
Финский изолят Куусамо	368.3	10.77	5	146
Население Финляндии	606.5	17.73	26	731
Ньюфаундленд	790.4	23.10	69	2,014
Африканеры	794.1	23.21	76	2,633
Сардиния, провинция Нуоро	681.4	19.92	38	1,241

* LDU – LD units; расстояние, на котором LD падает в два раза.

исследование генетически изолированных популяций следует признать экономически выгодным.

В некоторых районах генома, которые мы назвали «пробелами», неравновесие по сцеплению падает чрезвычайно быстро, и, таким образом, в этих районах картирование с помощью анализа ассоциаций может быть затруднено. Пробелы LD были определены как промежутки размером ≥ 2.5 LDU (расстояние, на котором LD падает в два раза) между соседними SNP на карте LD. В целом, как и ожидалось, число пробелов LD было связано обратной зависимостью с длиной карты LD (Таб. 2). Представляется вероятным, что некоторые из таких областей, обладающих, судя по всему, чрезвычайно высокой рекомбинационной частотой, не могут быть исследованы в рамках полногеномного анализа ассоциаций и будут исследованы с помощью технологий нового поколения, позволяющих ресеквенировать индивидуальные геномы.

Разработка методов генетического картирования с помощью неравновесия по сцеплению

Мы показали, что использование генетически изолированных популяций человека позволяет повысить мощность картирования. Однако полногеномный анализ ассоциаций в таких популяциях, выборки из которых представляют собой, по существу, большие родословные, требует разработки специальных методов. Сходная структура выборок – большие родословные – характерна для популяций аутбредных домашних и сельскохозяйственных животных.

Ожидается, что генетический контроль сложных признаков осуществляется множественными генами, при этом вклад каждого отдельного гена может быть невелик. Например, один из наиболее изученных локусов количественного признака (quantitative trait locus, QTL) – *APOE*, – достоверно и устойчиво ассоциирован с повышенным уровнем общего холестерина. Все же он объясняет лишь около 2–5% дисперсии данного признака (SING and DAVIGNON 1985; ISAACS *et al.* 2007).

При идентификации аллелей малого эффекта анализ ассоциаций потенциально имеет более высокую мощность и более высокую разрешающую способность по сравнению с классическим анализом сцепления (RISCH and MERIKANGAS 1996). В последние годы был достигнут значительный методологический и технический прогресс в области анализа ассоциаций. Большой упор был сделан на анализ выборок неродственных пациентов и здоровых людей (выборка типа «случай-контроль»), взятых из открытой популяции, а также на картирование бинарных и количественных признаков с использованием семейных данных (см. обзор (FORABOSCO *et al.* 2005)). Для картирования QTL в родословных был разработан ряд методов анализа ассоциаций и программных пакетов, использующих информацию о передаче аллелей. Эти методы включают, например, ортогональный тест значимости внутрисемейной дисперсии (quantitative trait transmission disequilibrium test, QTDT) (ABECASIS *et al.* 2000) и метод тестирования ассоциаций на семейных данных (family-based association test, FBAT) (LANGE *et al.* 2002; HORVATH *et al.* 2004). Поскольку эти методы анализируют ассоциацию между признаком и передачей аллелей маркера, т. е. используют только внутрисемейную дисперсию, их результаты являются несмещенными даже в присутствии подразделенности (этнической гетерогенности) исследуемой популяции. Однако эти методы игнорируют большой объем информации, заключенный в межсемейной дисперсии, оставляя

пространство для дальнейшего совершенствования методов картирования.

При анализе открытых и недавно смешанных популяций можно ожидать, что в выборку могут попасть этнически разные особи. В то же время, в популяциях, которые тщательно отбирались для анализа с использованием строгих критериев этнического происхождения, а также в генетически изолированных популяциях, риск подразделенности минимален. Более того, генетическая подразделенность может быть обнаружена с помощью генетических маркеров (PRITCHARD *et al.* 2000; FALUSH *et al.* 2003), и особи, значительно отличающиеся от остальной выборки, могут быть исключены из дальнейшего анализа, либо анализ может быть скорректирован на подразделенность.

В отсутствии подразделенности «золотым стандартом» статистической генетики является традиционная смешанная полигенная модель наследования признака y

$$y = \mu + G + e,$$

где μ – популяционное среднее значение признака, G – вектор случайных полигенных эффектов, а e – вектор случайных остаточных эффектов. Эту модель можно расширить для исследования ассоциации, включив в нее элемент kg

$$y = \mu + G + kg + e,$$

где k – эффект маркерного генотипа, а g – вектор маркерных генотипов. Такая модель, реализующая общий тест внутри- и межсемейной дисперсии, известна под названием модель измеренных генотипов (measured genotype, MG) (HOPPER and MATHEWS 1982; BOERWINKLE *et al.* 1986; GEORGE and ELSTON 1987). Статистическая значимость эффекта полиморфизма маркерного локуса оценивается с помощью критерия отношения правдоподобия (при использовании максимума правдоподобия, maximum likelihood, ML) или теста Вальда (при использовании ограниченного максимума правдоподобия, restricted maximum likelihood, REML).

Подход MG является мощным инструментом анализа количественных признаков в ситуациях, когда эффекты подразделенности можно игнорировать (HAVILL *et al.* 2005; LANGE *et al.* 2005). К сожалению, если анализируются большие родословные, что особенно часто бывает при исследовании генетически изолированных популяций (NEWMAN *et al.* 2001; BOURGAIN and GENIN 2005; PARDO *et al.* 2005) или некоторых этнических подгрупп (CHARLESWORTH *et al.* 2005; LEHMAN *et al.* 2006), метод измеренных генотипов требует большого объема вычислений. Это обусловлено необходимостью оценки параметров сложной смешанной модели для каждого тестируемого

маркера. Проверка эффекта одного полиморфизма может занять от нескольких минут до нескольких часов и, следовательно, полногеномный анализ ассоциаций с применением этого метода потребует значительных вычислительных ресурсов; реализация такого подхода с применением одного компьютера не представляется практически возможной и анализ требует применения распределенных вычислений.

Другим существенным недостатком метода измеренных генотипов является то, что в его рамках невозможен эмпирический анализ значимости с помощью пермутаций и бутстрепа: пермутации значений признака в выборке родословных нарушают не только зависимость между маркером и признаком, но и зависимости между признаками родственников, обусловленные полигенным наследованием.

Мы исследовали альтернативные подходы к картированию QTL методом анализа ассоциаций в выборках родственников и разработали семейство новых, быстрых и простых методов полногеномного анализа ассоциаций с использованием смешанной модели и регрессии, GRAMMAR (Genomewide Rapid Association using Mixed Model And Regression).

Основная идея предложенного метода заключается в том, что анализ полигенной модели выполняется отдельно, с использованием информации о родственной структуре выборки, но без учета маркерных данных. Затем оценки средовых остатков признака, полученные в рамках этой модели и скорректированные на полигенную ковариацию и фиксированные эффекты, используются как количественный признак для анализа ассоциаций с каждым из множества маркеров. Этот анализ проводится классическими методами, применяемыми для анализа неродственных особей из популяции.

Было показано, что метод GRAMMAR является достаточно быстрым для проведения полногеномного анализа. В то же время, было показано, что GRAMMAR является консервативным тестом. Поэтому далее нами был предложен метод, позволяющий контролировать ошибку первого рода за счет использования полногеномной информации. Действительно, большинство локусов в геноме не ассоциировано с признаком и для них справедлива нулевая гипотеза об отсутствии ассоциации. По этим локусам можно оценить распределение статистики при справедливости нулевой гипотезы и скорректировать пороги значимости. Далее мы предложили использовать полногеномные данные, а не родословную, для оценки матрицы родства.

Данный метод, названный нами GRAMMAR-GC, позволяет повысить мощность метода GRAMMAR практически до уровня метода

измеренных генотипов (Рис. 1), который является теоретически наиболее мощным, но в тоже время чрезвычайно вычислительно сложным и практически не применимым при полногеномном анализе ассоциаций.

Одно из преимуществ методов GRAMMAR по сравнению с другими методами, позволяющими анализировать ассоциацию в родословных, состоит в том, что средовые остатки полигенной модели, используемые при анализе, свободны от семейных корреляций. Следовательно, структура данных становится взаимозаменяемой, и к ним можно применить технику пермутаций для получения эмпирических оценок границ значимости. Это свойство метода GRAMMAR также позволяет использовать для анализа целый ряд современных методов, разработанных для выборок «неродственных особей».

Другим преимуществом метода GRAMMAR по сравнению с существующими методами, позволяющими анализировать ассоциацию в родословных, является то, что GRAMMAR очень просто модифицировать для тестирования целого ряда моделей, например, включить дополнительные независимые переменные, учитывающие взаимодействие с полом и факторами внешней среды, взаимодействие между генами, эффект родительского происхождения аллелей и так далее. Недавно нами был реализован вариант GRAMMAR, позволяющий исследовать эффект родительского происхождения аллелей в полногеномном анализе ассоциаций (BELONOGOVA *et al.* 2009).

Нами также была предложена реализация метода измеренных генотипов с помощью скор-теста, не требующего оценки дисперсии при альтернативном значении тестируемого параметра, которая может стать мощной альтернативой метода GRAMMAR. Подобная реализация была описана в независимой работе Чена и Абекасиса (CHEN and ABECASIS 2007); эта модель была расширена нами в пакете ProbABEL (AULCHENKO and STRUCHALIN 2010).

Следует отметить, что хотя новые методы были разработаны нами для анализа количественных признаков, они также применимы для анализа бинарных признаков. При этом получаемые оценки уровня значимости ассоциаций хорошо совпадают с таковыми, полученными при использовании более корректных (и вычислительно значительно более сложных) методов, разработанных специально для анализа бинарных признаков (личное сообщение, N. Pirastu).

Таким образом, нами был сформулирован и реализован ряд новых методов, позволяющих проводить полногеномный анализ ассоциаций

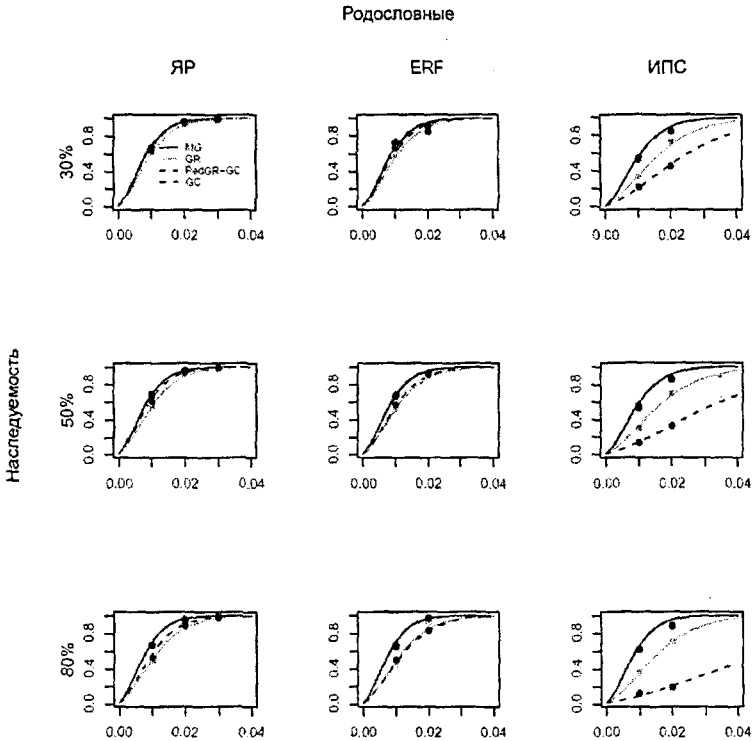

Рис. 1. Мощность методов измеренных генотипов, GRAMMAR-GC (перекрывающиеся верхние линии), GRAMMAR (серая линия) и GC (нижняя пунктирная линия) при разных значениях наследуемости и структурах родословных.
Ряды отличаются значениями наследуемости (от 30% до 80%), а колонки – структурой родословных: ядерные родословные (ЯР), родословная из молодой генетически изолированной популяции из Нидерландов (ERF) и идеализированная популяция свиней (ИПС). Ось Y каждой панели указывает мощность, тогда как ось X – долю дисперсии признака, объясненной исследуемым QTL. Кружки соответствуют эмпирическим оценкам мощности, посчитанным при $\alpha = 0.01$. Оценки мощности основаны на 1000 повторах для ЯР и ИПС и на 100 повторах для ERF.

признаков в выборках родственников. Эти методы не требуют априорного знания степени родства между исследуемыми особями (родословной), так как генетическое родство между особями

оценивается с помощью геномных данных; эти методы позволяют проводить быстрые вычисления. В то же время мощность новых методов практически не уступает мощности метода измеренных генотипов, который является «золотым стандартом» для методов исследования ассоциаций по выборкам родственников.

Разработанные методы, а также большой набор существующих методов были реализованы в пакете прикладных программ для анализа полногеномных данных, GenABEL (AULCHENKO *et al.* 2007b). Пакет распространяется свободно и доступен по адресу http://mga.bionet.nsc.ru/~yurii/ABEL.

Полногеномное исследование количественных признаков человека

Мы использовали разработанные методы и программное обеспечение для идентификации локусов, генетическая вариация которых ассоциирована с такими признаками, как уровень липидов в крови и рост человека. Кроме того, нами был исследован вопрос прогностической мощности геномного профилирования для предсказания исследованных признаков.

Генетические и физиологические основы метаболизма липидов хорошо изучены как на модельных объектах, так и на примере моногенных менделевских заболеваний. Не будет преувеличением сказать, что уровень липидов в крови человека – один из наиболее хорошо генетически изученных сложных количественных признаков человека (FRIEDLANDER *et al.* 1997; PILIA *et al.* 2006). Более того, в отличие от большинства сложных количественных признаков человека, для уровней липидов известен ряд генов, вариация которых объясняет существенную долю дисперсии этих признаков в популяции (например, аллели ε2/3/4 гена *APOE* (SING and DAVIGNON 1985)). Таким образом, в методологическом смысле, изучение уровней липидов в крови человека предоставляет прекрасную возможность для тестирования метода полногеномного анализа ассоциаций: ожидается, что метод должен подтвердить ряд ранее известных локусов (таким образом, имеется «позитивный контроль»).

Кроме того, идентификация геномных полиморфизмов, ассоциированных с уровнем липидов, представляет собой практическую ценность. Изменение уровней липидов сыворотки крови относительно нормы является одним из первостепенных факторов риска сердечнососудистых заболеваний (KANNEL *et al.* 1961; MILLER and MILLER 1975; PILIA *et al.* 2006). Теоретически, на основании генетического профиля риска возможна ранняя (до появления

клинических симптомов) идентификация людей с повышенным риском гиперхолестеринемии. Это знание может быть критически важным для предотвращения как гиперхолестеринемии, так и, в конечном счете, сопутствующих сердечнососудистых заболеваний. Действительно, уровень холестерина в крови как правило может быть модифицирован с помощью изменения стиля жизни и питания, а также с помощью различных лекарственных препаратов.

Нами было проведено полногеномное исследование ассоциаций уровней липидов в сыворотке крови человека. Мы использовали данные из 16 когорт, собранных по всей Европе; общий объем выборки составлял от 17 797 до 22 562 человек; полногеномное генотипирование каждого образца проводилось с использованием более 300 тысяч SNP.

Результаты полногеномного анализа ассоциаций уровня общего холестерина – признака, не исследовавшегося ранее с помощью этого метода – представлены на Рис. 2. Одиннадцать локусов показали ассоциацию с полногеномным уровнем значимости p-value $< 5 \times 10^{-8}$. Для трех из этих локусов (*FADS1/2/3, ABCG5/8, TMEM57*) вовлеченность в контроль уровней липидов в популяциях человека была ранее не известна. Для остальных локусов ассоциация с уровнями других липидов (холестерина липопротеидов низкой плотности или триглицеридов) была ранее уже показана.

В целом, мы идентифицировали шесть новых локусов (*FADS1/2/3, ABCG5/8, TMEM57, MADD-FOLH1, CTCF-PMRT8, DNAH11*), значимо ассоциированных с уровнями липидов, и подтвердили 16 локусов, ассоциация которых с метаболизмом липидов была показана ранее в полногеномных исследованиях ассоциаций (KATHIRESAN *et al.* 2008a; KATHIRESAN *et al.* 2008b; KOONER *et al.* 2008; WILLER *et al.* 2008). Ранее мы предположили, что исследование уровня липидов в крови человека может представлять также методологический интерес за счет того, что для некоторых липидов известны локусы, объясняющие большую долю дисперсии и представляющие, таким образом, «позитивный контроль». Наше исследование подтвердило это предположение: например, вариация в локусе *LDLR* была высоко значимо (p-value = 10^{-23}) ассоциирована с уровнем общего холестерина (Рис. 2), а вариация в локусе *CETP* объясняла ~2% дисперсии уровня холестерина липопротеидов высокой плотности и была детектирована с p-value = 10^{-93}.

Рост тела является классическим примером полигенно наследуемого признака человека. Многочисленные исследования показали, что доля дисперсии роста, объясняемая семейными факторами, составляет 80–90%. Сходство роста родственников в

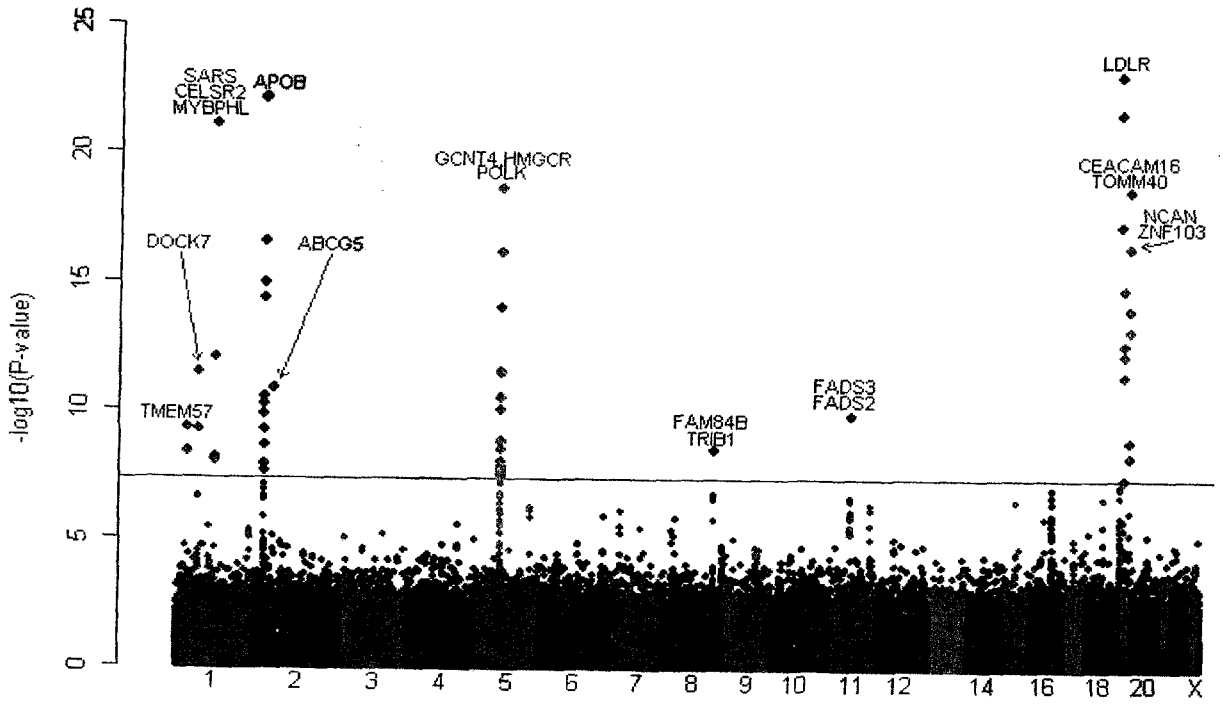

Рис. 2. Результаты полногеномного анализа ассоциаций уровня общего холестерина в крови в 16 популяционных когортах.

основном обусловлено генетическими факторами, поскольку эффекты *негенетических причин* сходства сибсов пренебрежимо малы (VISSCHER *et al.* 2006). В то же время, как до, так и после недавно проведенных полногеномных исследований ассоциаций (GUDBJARTSSON *et al.* 2008; LETTRE *et al.* 2008; WEEDON *et al.* 2008) ни одного распространенного аллеля, объясняющего существенную долю дисперсии роста в популяции человека, идентифицировано не было (локус, наиболее сильно ассоциированный с нормальной вариацией роста – *HMGA2* – объясняет только ~0.3% дисперсии).

Высокая наследуемость роста может быть объяснена как присутствием большого числа распространенных аллелей малого эффекта, так и присутствием большого числа редких аллелей с сильным эффектом на фенотип. При этом как распространенные, так и редкие аллели могут встречаться в рамках одного локуса. Например, такова аллельная архитектура локуса *LDLR*, принимающего участие в контроле уровня холестерина липопротеидов низкой плотности. Для идентификации распространенных аллелей малого эффекта наиболее эффективной стратегией является полногеномный анализ ассоциаций с использованием больших выборок. Однако этот метод неприменим для идентификации локусов, в которых встречаются редкие аллели, даже если таковые обладают большим эффектом на фенотип: распространенные полиморфизмы, используемые в ДНК-чипах, находятся в статистически слабой ассоциации с редкими полиморфизмами. Для идентификации локусов, содержащих редкие аллели с большим эффектом на фенотип, может применяться классический метод анализа сцепления. К сожалению, анализ сцепления позволяет идентифицировать только большие геномные регионы, содержащие как правило десятки или даже сотни генов. Однако если аллельная архитектура исследуемого локуса включает как редкие, так и распространенные аллели, можно ожидать, что анализ ассоциаций в регионе сцепления позволит провести точное картирование исследуемого локуса. При этом, в отличие от полногеномного анализа, можно применять более слабые критерии значимости, что позволит идентифицировать локусы, которые невозможно обнаружить только с помощью полногеномного анализа ассоциаций.

Таким образом, анализ сцепления с последующим анализом ассоциаций является стратегией, которая может позволить идентифицировать локусы со смешанной аллельной архитектурой. Мы применили эту стратегию для исследования генетики роста человека. Исследование было проведено в рамках консорциума по генетике генетически изолированных популяций (EUROSPAN). Анализ

сцепления был проведен на материале из четырех популяций. LOD score пяти хромосомных районов достиг границы возможного сцепления. Для трех из этих районов (хромосомы 2, 7 и 17) самое высокое значение LOD было получено при анализе объединенной выборки. В остальных двух районах сцепления (хромосомы 9 и 16) общее значение LOD было высоким благодаря сильному эффекту в одной из популяций при практически нулевом значении LOD в других популяциях. Следующим шагом было исследование ассоциаций между SNP и признаком в районах сцепления. Пять идентифицированных нами районов сцепления были большими, включая от 887 до 3176 SNP. В сумме было исследовано 9852 маркеров. Только в одном районе (хромосома 7) была найдена статистически значимая ассоциация с ростом при мета-анализе. В этом районе два соседних SNP (rs849140 и rs1635852) были ассоциированы с ростом ($p < 0.05$ после коррекции Бонферрони на 9852 протестированных SNP); более сильная ассоциация наблюдалась при анализе роста мужчин (Рис. 3). Оба SNP локализованы в гене *JAZF1*. Последующая проверка rs849140 с привлечением дополнительного материала показала значимость ассоциации этого SNP с ростом тела.

Хотя окончательное доказательство того, что локус *JAZF1* является примером смешанной аллельной архитектуры, может быть предоставлено только последующими исследованиями, общую стратегию поиска таких локусов, основанную на анализе ассоциаций в регионах сцепления, можно рассматривать как многообещающую. Следует отметить, что эта стратегия представляет собой вариант классической стратегии позиционного клонирования, незаслуженно забытой в последнее время.

Биологически чрезвычайно интересным представляется тот факт, что локус *JAZF1* является примером плейотропного локуса – SNP rs849140, ассоциированный с ростом в нашем исследовании, также ассоциирован с диабетом второго типа (ZEGGINI *et al.* 2008) и системной красной волчанкой (GATEVA *et al.* 2009). Другие SNP этого же локуса ассоциированы с раком простаты (THOMAS *et al.* 2008). Как для уровней липидов, так и для роста человека, нами был исследован потенциал использования геномных данных для предсказания этих признаков. Было показано, что геномные профили объясняют 4–6% дисперсии роста и 1–7% дисперсии липидов в разных популяциях. Также показано, что геномный профиль холестерина является статистически значимым, независимым от уровня циркулирующего холестерина, предиктором дислипидемии и толщины комплекса интима-медиа стенки сосуда. Кроме того, мы показали, что на современном этапе простое предсказание на основе фенотипов родственников (метод Гальтона)

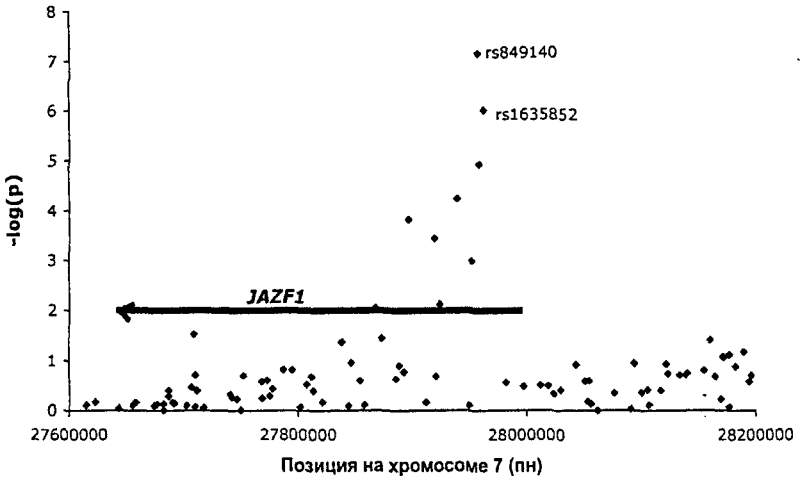

Рис. 3. Ассоциация роста мужчин с SNP, расположенными в районе гена *JAZF1* хромосомы 7.
Ось Y: $-\log_{10}$ (номинальное *p*-value); ось X: позиция (пн) на хромосоме 7.

может значительно превосходить по качеству сложные предсказания на основе геномных данных. Так, для роста тела гальтоновское среднеродительское предсказание было на порядок лучше геномного предсказания. Дополнительное включение геномного профиля в гальтоновскую модель улучшало модель не значительно (Таб. 3).

Мы рассмотрели вопрос, будет ли гальтоновское среднеродительское предсказание значительно лучше геномного предсказания для других фенотипов. Доля фенотипического разнообразия потомков, которое объясняется среднеродительским фенотипическим значением, выражается формулой $(h^2)^2/2$, где h^2 – наследуемость признака. Мы показали, что 11 SNP объясняют 3–5% дисперсии общего холестерина; сходные оценки были получены для липопротеидов высокой и низкой плотности и триглицеридов. Эти признаки обычно проявляют около 30% наследуемости. Следовательно, гальтоновское предсказание не может объяснить более 5% дисперсии признака. Таким образом, для уровней липидов предсказание на основе геномных данных уже работает столь же хорошо (или столь же плохо) как гальтоновское. Однако геномные профили, в отличие от гальтоновского, имеют потенциал к совершенствованию по мере обнаружения новых локусов, влияющих на фенотип.

Таб. 3. Доля дисперсии, объясненной различными профилями.

Профиль	Популяция	*N*	Доля объяснен-ной дисперсии, %	$\Delta_{5,95}$, см*
Профиль на основе 54 геномных локусов	Роттердамское исследование	5748	3.8	4.95
Гипотетический профиль	Роттердамское исследование	5748	80.0	23.4 ± 0.01
Среднеродительский профиль Гальтона	ERF	550	40.1	17.68
Среднеродительский профиль Гальтона	ERF	257	44.9	21.18
Профиль Гальтона + 54 локуса	ERF	257	46.2	21.28

* $\Delta_{5,95}$ – разница между средними значениями роста в верхних и нижних 5% распределения профиля. Два последних профиля тестировались на выборке участников ERF с известными фенотипами родителей (N = 257).

Генетическая архитектура признака является важным фактором, который следует рассматривать при оценке потенциала прогностического тестирования (JANSSENS *et al.* 2006). Например, популяционное разнообразие цвета радужной оболочки глаза в значительной степени объясняется генетическим разнообразием единственного локуса (*HERC2*), и высокая точность предсказания достигается при использовании генотипов всего трех SNP (KAYSER *et al.* 2008a). Однако для таких признаков как артериальное давление крови известно буквально несколько локусов, объясняющих лишь небольшую долю дисперсии; для таких признаков перспективы применения геномных профилей на данном этапе развития генетики значительно хуже.

Нами, как и другими исследователями, было показано, что при использовании строгих критериев полногеномной значимости (поправка Бонферрони или использование фиксированного граничного значения $p < 5 \times 10^{-8}$ для популяций европейского происхождения) результаты полногеномного анализа ассоциаций являются в целом хорошо воспроизводимыми при условии достаточного объема репликационной

выборки. Например, из шести новых липидных локусов, описанных нами, пять было подтверждено в последующих независимых исследованиях (KATHIRESAN *et al.* 2009; MANOLIO 2009; SABATTI *et al.* 2009). При проверке SNP, ассоциация которых с ростом считается установленной (GUDBJARTSSON *et al.* 2008; LETTRE *et al.* 2008; WEEDON *et al.* 2008), на выборке Роттердамского исследования 34 из 54 SNP были значимо ассоциированы при $\alpha = 0.05$, и только для двух SNP направление (не значимой) ассоциации с ростом не соответствовало описанному в первоначальных работах. При этом следует отметить, что при исследовании роста выборка Роттердамского исследования не являлась достаточно мощной для подтверждения всех ассоциаций, и, таким образом, отсутствие значимой ассоциации для некоторых локусов (с малым эффектом) не могло считаться доказательством отсутствия эффекта этих локусов. Сходным образом, 18 из 33-х ранее идентифицированных SNP, которые могли быть протестированы на ассоциацию с ростом в выборке EUROSPAN, показали ассоциацию с p-value < 5% – результат, показывающий действительную насыщенность этого набора SNP реальными сигналами ассоциаций.

Таким образом, на основании наших исследований можно сделать заключение, что полногеномный анализ ассоциаций является мощным и надежным методом идентификации локусов, вариация которых ассоциирована со значениями сложных признаков; результаты, полученные с использованием метода полногеномного анализа ассоциаций, в целом хорошо воспроизводимы. На основании знания этих локусов возможно конструирование генетических профилей риска, которые (уже или в перспективе) могут предсказывать долговременный риск развития заболевания.

Заключение

Основной целью исследований, описанных в данной работе, являлась разработка методов полногеномного анализа ассоциаций в выборках произвольной структуры из аутбредных популяций, апробация этих методов на реальных данных и идентификация новых локусов, контролирующих сложные признаки человека.

В момент начала работы по этой теме (2003 год), за исключением моногенных форм, объясняющих ничтожную долю случаев, было известно всего несколько локусов, генетическая вариация которых была достоверно связана с разнообразием сложных признаков человека, в частности, с риском распространенных болезней. За прошедшее время

эта ситуация изменилась кардинальным образом – в настоящий момент известны более тысячи локусов, достоверно ассоциированных с сотнями признаков человека (см. неполный каталог результатов полногеномных исследований ассоциаций на сайтах http://www.genome.gov/gwastudies/, а также «GWAS Integrator», http://hugenavigator.net/HuGENavigator/gWAHitStartPage.do). Эти успехи в большой степени обусловлены применением нового метода – полногеномного анализа ассоциаций.

Автору данной диссертации посчастливилось принимать участие в работах, приведших к современному буму полногеномных исследований ассоциаций. В настоящее время он является (со)автором работ, в которых был проведен полногеномный анализ ассоциаций 32 признаков; в результате было идентифицировано 238 значимых ассоциаций в 148-и различных регионах генома (по данным сайта «GWAS Integrator», запрос произведен 27 апреля 2010 года). Были исследованы такие сложные признаки человека, как ожирение (JOHANSSON *et al.* 2009b) и антропометрические характеристики (HEARD-COSTA *et al.* 2009; LINDGREN *et al.* 2009), остеопороз (RICHARDS *et al.* 2009; RIVADENEIRA *et al.* 2009), рассеянный склероз (AULCHENKO *et al.* 2008; HOPPENBROUWERS *et al.* 2009), уровни липидов в крови (AULCHENKO *et al.* 2009a; HICKS *et al.* 2009), уровни различных метаболитов (KOLZ *et al.* 2009; PATTARO *et al.* 2009; PROKOPENKO *et al.* 2009) и пептидов (KOLLERITS *et al.* 2009), рост (ESTRADA *et al.* 2009; JOHANSSON *et al.* 2009a), функция почки (KOTTGEN *et al.* 2009), артериальное давление крови (LEVY *et al.* 2009), инсульт (IKRAM *et al.* 2009), курение (VINK *et al.* 2009), структура и функционирование сердца (VASAN *et al.* 2009), цвет радужной оболочки глаза (KAYSER *et al.* 2008b) и так далее.

Было показано, что в молодых генетически изолированных популяциях, представленных в Европе в большом количестве, частота редких (начальная частота <1%) аллелей может быть как резко (в разы) повышена, так и резко (вплоть до полного исчезновения) снижена, что приводит к повышению мощности генетического анализа в таких популяциях. Относительно распространенных аллелей, нами было показано, что генетические варианты с начальной частотой 5% или выше будут присутствовать как в молодых генетических изолятах, так и в открытых популяциях. Следовательно, результаты геномного сканирования, проведенного с использованием ДНК-чипов в молодых генетически изолированных популяциях, могут быть обобщены на открытую популяцию, и наоборот. Далее, нами было показано, что в изолированных популяциях, недавно переживших период быстрого роста и берущих начало от небольшой популяции основателей,

неравновесие по сцеплению распространяется на значительно большие дистанции по сравнению с большими открытыми популяциями; в частности, для хромосомы 22 карта неравновесия по сцеплению для генетических изолятов на ~20–45% короче, чем для открытых популяций, что приводит к аналогичному повышению ожидаемой мощности полногеномного анализа ассоциаций. Таким образом, на основании наших исследований можно сделать заключение, что молодые генетически изолированные популяции представляют ценный ресурс для картирования локусов сложных признаков методом полногеномного анализа ассоциаций.

Далее, нами был разработан и реализован ряд новых, быстрых и простых методов, позволяющих проводить полногеномный анализ ассоциаций признаков в выборках родственников. Разработанные нами методы не требуют априорного знания степени родства между исследуемыми особями (родословной), так как для оценки генетического родства используются геномные данные. Мощность новых методов практически не уступает мощности «золотого стандарта» для методов исследования ассоциаций по выборкам родственников (классический метод измеренных генотипов). Разработанные методы были реализованы в пакете прикладных программ для анализа полногеномных данных, GenABEL.

На основании результатов, полученных нами при исследовании молодых изолятов, было решено проводить исследование генетики сложных признаков человека в генетически изолированных популяциях Европы (например, консорциум EUROSPAN). Полногеномный анализ ассоциаций в этих эмпирических исследованиях проводился с использованием разработанных нами методов. В настоящее время возможность использования генетически изолированных популяций для идентификации локусов сложных признаков с использованием метода полногеномного анализа ассоциаций не вызывает сомнения, а методы, описанные и реализованные нами, вошли в стандартный арсенал полногеномных исследований ассоциаций.

Следует отдельно отметить, что применение методов, разработанных нами для анализа генетически изолированных популяций человека, не ограничено только этими популяциями. В первую очередь, наши методы применимы для анализа любых семейных выборок человека. Принимая во внимание то, что при субтотальном (>10%) обследовании любой популяции в выборке обязательно начинают встречаться родственные особи, и что многие исследования в настоящий момент выходят на субтотальный уровень, роль разработанных нами методов в дальнейшем будет повышаться. Более того,

сконструированные нами методы могут применяться при полногеномном анализе признаков сельскохозяйственных и домашних животных. В частности, нам известно, что в настоящий момент разработанные нами методы и пакеты программ применяются при исследовании генетики крупного рогатого скота и собак.

В целом, результаты работ по созданию новых методов полногеномного анализа ассоциаций следует признать одними из наиболее успешных из представленных в данной диссертации. Так, число пользователей, которые обращались с вопросами к разработчикам нашего пакета полногеномного анализа ассоциаций GenABEL составляет более двухсот пятидесяти человек, число опубликованных работ, использовавших пакет, составляет более 50; наш пакет был упомянут в статье *New York Times*, посвященной вычислительной среде R.

Нами были идентифицированы новые локусы, генетическая вариация которых ассоциирована с изменением уровня липидов в крови и ростом тела человека. Одним из наиболее интересных биологических результатов представляется то, что уровень липидов в крови человека зачастую контролируется вариацией в генах, которые представлены гомологичными кластерами (*FADS1/2/3*, *ABCG5/8*). Также интересен факт, что SNP rs849140, находящийся в локусе *JAZF1* и ассоциированный с ростом в нашем исследовании, также показал ассоциацию с диабетом второго типа (ZEGGINI *et al.* 2008) и системной красной волчанкой (GATEVA *et al.* 2009). Другие SNP этого же локуса ассоциированы с раком простаты (THOMAS *et al.* 2008).

Методологически, нами, как и другими исследователями, было показано, что полногеномный анализ ассоциаций является мощным методом идентификации распространенных аллелей, контролирующих сложные признаки. Результаты, полученные с использованием метода полногеномного анализа ассоциаций, в целом хорошо воспроизводимы при использовании строгих критериев полногеномной значимости и адекватных объемов репликационных выборок. Эти результаты оправдывают дальнейшее широкое применение метода полногеномного анализа ассоциаций – метода, который за последние несколько лет стал *de-facto* стандартом идентификации локусов сложных признаков человека.

Нами также были описаны методологические основы генетического предсказания. За несколько последних лет методология, предложенная Janssens et al. (JANSSENS *et al.* 2004) и в дальнейшем развитая нами (JANSSENS *et al.* 2006) – оценка предсказательной мощности генетического профиля площадью под кривой, показывающей

соотношение между ложно-положительными и истинно-положительными результатами теста – стала стандартной, и используется во многих работах, представляющих результаты полногеномного анализа. Мы показали, что знание локусов, идентифицированных в ходе полногеномных анализов ассоциаций, позволяет конструировать генетические профили риска, которые (уже или в перспективе) могут предсказывать значение количественных признаков и долговременный риск развития заболевания. С ростом числа известных локусов геномное профилирование может стать стандартной процедурой при предсказании некоторых признаков. Однако потенциал этого метода в значительной степени зависит от генетической архитектуры признака.

Работы, представленные в данной диссертации, получили широкий отклик в научной среде: так, работы, представленные в главе 2 диссертации (AULCHENKO *et al.* 2003; AULCHENKO *et al.* 2004; PARDO *et al.* 2005; SERVICE *et al.* 2006) были процитированы 164 раза, работы, представленные в главе 3 (AMIN *et al.* 2007; AULCHENKO *et al.* 2007a; AULCHENKO *et al.* 2007b) – 72 раза, а работы, представленные в главе 4 (JANSSENS *et al.* 2006; KAYSER *et al.* 2008a; AULCHENKO *et al.* 2009a; AULCHENKO *et al.* 2009b; JOHANSSON *et al.* 2009a) – 165 раз (ISI Web of Knowledge, запрос произведен 27 апреля 2010). В сумме работы автора данной диссертации (в том числе работы, не включенные в данную диссертацию), цитируются более тысячи раз (из них более 400 цитирований за 2009 год).

Следует отметить, что хотя идентификация локусов сложных признаков с помощью метода полногеномного анализа ассоциаций и является важным этапом генетического анализа, этот метод зачастую не дает окончательного ответа на вопрос, продукт какого гена вовлечен в контроль признака. Для ответа на этот несомненно биологически важный вопрос необходимо проведение функциональных, молекулярно-генетических и физиологических исследований. Однако рассмотрение вопроса функциональности идентифицированных полиморфизмов находится за рамками поставленной нами цели.

Разрешающая способность метода полногеномного анализа ассоциаций ограничена распространенными аллелями (с частотой редкого аллеля $\geq$5%). В то же время, в контроле многих признаков, судя по всему, велика роль множественных редких аллелей (гипотеза «распространенная болезнь – множество редких аллелей», CDMRV). Такие аллели можно детектировать с помощью современных технологий, которые уже позволяют ресеквенировать индивидуальные геномы; цена таких исследований стремительно снижается. Однако

генетический анализ редких аллелей представляет собой большую методическую проблему, так как статистическая мощность оценки эффекта редкого фактора чрезвычайно мала. Чтобы успешно решить эту проблему и определить роль редких аллелей в детерминации сложных признаков, потребуется создать принципиально новые методы анализа, которые, скорее всего, будут лишь в малой степени сходны с методами классической эпидемиологии.

Выводы

1. Исследован эффект дрейфа генов в молодых генетически изолированных популяциях человека. Показано, что в таких популяциях эффект дрейфа генов мал для распространенных (частота $\geq$5%) аллелей и выражен для аллелей, имеющих начальную частоту <1%.
2. Проведен сравнительный анализ структуры неравновесия по сцеплению в различных популяциях человека. Показано, что длины карт неравновесия по сцеплению для молодых генетически изолированных популяций на ~30% меньше, чем для открытых популяций человека.
3. Разработаны новые методы для проведения полногеномного анализа ассоциаций в выборках произвольной структуры из аутбредных популяций человека, сельскохозяйственных и домашних животных. Эти методы являются статистически мощными и вычислительно эффективными.
4. Разработано новое программное обеспечение для проведения полногеномного анализа ассоциаций. Разработанный пакет программ GenABEL реализует большое число современных методов полногеномного анализа ассоциаций и позволяет анализировать миллионы SNP, типированных в тысячах образцов, на персональных компьютерах.
5. С использованием разработанных методов и программ проведен полногеномный анализ ассоциаций уровней липидов в крови человека. Впервые в мире, такой анализ проведен на популяционных выборках. Также впервые проведен полногеномный анализ ассоциаций уровней общего холестерина. Идентифицированы шесть новых локусов, контролирующих уровни липидов.

6. Проведен полногеномный анализ сцепления с последующим анализом генетических ассоциаций с ростом человека. Идентифицирован новый локус, *JAZF1*, контролирующий рост тела, и имеющий плейотропное влияние на ряд других признаков, в том числе патологических.

7. Оценен потенциал метода предсказания значения сложного признака на основе генотипических данных и проведено практическое исследование возможности использования геномных данных для предсказания таких признаков человека, как роста тела, уровень липидов в крови и риск дислипидемии. Показано, что геномные профили объясняют 4–6% дисперсии роста и 1–7% дисперсии липидов. Также показано, что геномный профиль холестерина является статистически значимым, независимым от уровня циркулирующего холестерина, предиктором толщины интима-медиа и дислипидемии.

Список публикаций по теме диссертации

1. АКСЕНОВИЧ, Т. И., Г. Р. СВИЩЕВА и Ю. С. АУЛЬЧЕНКО, 2000 Картирование генов, детерминирующих количественные признаки животных: метод разложения дисперсий. Генетика 36: 986–993.
2. АУЛЬЧЕНКО, Ю. С. и Т. И. АКСЕНОВИЧ, 2006 Методологические подходы и стратегии картирования генов, контролирующих комплексные признаки человека. Вестник ВОГиС 10: 189-202.
3. ТИМОФЕЕВА, О. А., М. Л. ФИЛИПЕНКО, Ю. С., АУЛЬЧЕНКО, Е. А. ВОРОНИНА, А. Б., МАСЛЕННИКОВ и Н. П. МЕРТВЕЦОВ, 1999 Анализ распределения аллелей тетрануклеотидного повтора в интроне 6 гена липопротеинлипазы среди населения г. Новосибирска. Генетика 35: 862-864.
4. AMIN, N., C. M. VAN DUIJN and Y. S. AULCHENKO, 2007 A genomic background based method for association analysis in related individuals. PLoS ONE 2: e1274.
5. AULCHENKO, Y. S., T. I. AXENOVICH, I. MACKAY and C. M. VAN DUIJN, 2003 miLD and booLD programs for calculation and analysis of corrected linkage disequilibrium. Ann Hum Genet 67: 372-375.
6. AULCHENKO, Y. S., D. J. DE KONING and C. HALEY, 2007a Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177: 577-585.
7. AULCHENKO, Y. S., P. HEUTINK, I. MACKAY, A. M. BERTOLI-AVELLA, J. PULLEN et al., 2004 Linkage disequilibrium in young genetically isolated Dutch population. Eur J Hum Genet 12: 527-534.
8. AULCHENKO, Y. S., I. A. HOPPENBROUWERS, S. V. RAMAGOPALAN, L. BROER, N. JAFARI et al., 2008 Genetic variation in the KIF1B locus influences susceptibility to multiple sclerosis. Nat Genet 40: 1402-1403.

9. AULCHENKO, Y. S., S. RIPATTI, I. LINDQVIST, D. BOOMSMA, I. M. HEID et al., 2009a Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41: 47-55.
10. AULCHENKO, Y. S., S. RIPKE, A. ISAACS and C. M. VAN DUIJN, 2007b GenABEL: an R library for genome-wide association analysis. Bioinformatics 23: 1294-1296.
11. AULCHENKO, Y. S., M. V. STRUCHALIN, N. M. BELONOGOVA, T. I. AXENOVICH, M. N. WEEDON et al., 2009b Predicting human height by Victorian and genomic methods. Eur J Hum Genet 17: 1070-1075.
12. AXENOVICH, T. I., I. V. ZORKOLTSEVA, N. M. BELONOGOVA, M. V. STRUCHALIN, A. V. KIRICHENKO et al., 2009 Linkage analysis of adult height in a large pedigree from a Dutch genetically isolated population. Hum Genet 126: 457-471.
13. BELONOGOVA, N. M., T. I. AXENOVICH and Y. S. AULCHENKO, 2009 A powerful genome-wide feasible approach to detect parent-of-origin effects in studies of quantitative traits. Eur J Hum Genet.
14. ESTRADA, K., M. KRAWCZAK, S. SCHREIBER, K. VAN DUIJN, L. STOLK et al., 2009 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 18: 3516-3524.
15. HEARD-COSTA, N. L., M. C. ZILLIKENS, K. L. MONDA, A. JOHANSSON, T. B. HARRIS et al., 2009 NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. PLoS Genet 5: e1000539.
16. HICKS, A. A., P. P. PRAMSTALLER, A. JOHANSSON, V. VITART, I. RUDAN et al., 2009 Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS Genet 5: e1000672.
17. HOPPENBROUWERS, I. A., Y. S. AULCHENKO, A. C. JANSSENS, S. V. RAMAGOPALAN, L. BROER et al., 2009 Replication of CD58 and CLEC16A as genome-wide significant risk genes for multiple sclerosis. J Hum Genet 54: 676-680.
18. IKRAM, M. A., S. SESHADRI, J. C. BIS, M. FORNAGE, A. L. DESTEFANO et al., 2009 Genomewide association studies of stroke. N Engl J Med 360: 1718-1728.
19. JANSSENS, A. C., Y. S. AULCHENKO, S. ELEFANTE, G. J. BORSBOOM, E. W. STEYERBERG et al., 2006 Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med 8: 395-400.
20. JOHANSSON, A., F. MARRONI, C. HAYWARD, C. S. FRANKLIN, A. V. KIRICHENKO et al., 2009a Common variants in the JAZF1 gene associated with height identified by linkage and genome-wide association analysis. Hum Mol Genet 18: 373-380.
21. JOHANSSON, A., F. MARRONI, C. HAYWARD, C. S. FRANKLIN, A. V. KIRICHENKO et al., 2009b Linkage and Genome-wide Association Analysis of Obesity-related Phenotypes: Association of Weight With the MGAT1 Gene. Obesity (Silver Spring).
22. KAYSER, M., F. LIU, A. C. JANSSENS, F. RIVADENEIRA, O. LAO et al., 2008 Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. Am J Hum Genet 82: 411-423.
23. KOLLERITS, B., S. COASSIN, N. D. BECKMANN, A. TEUMER, S. KIECHL et al., 2009 Genetic evidence for a role of adiponutrin in the metabolism of apolipoprotein B-containing lipoproteins. Hum Mol Genet 18: 4669-4676.

24. KOLZ, M., T. JOHNSON, S. SANNA, A. TEUMER, V. VITART et al., 2009 Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. PLoS Genet 5: e1000504.
25. KOTTGEN, A., N. L. GLAZER, A. DEHGHAN, S. J. HWANG, R. KATZ et al., 2009 Multiple loci associated with indices of renal function and chronic kidney disease. Nat Genet.
26. LAM, A. C., M. SCHOUTEN, Y. S. AULCHENKO, C. S. HALEY and D. J. DE KONING, 2007 Rapid and robust association mapping of expression quantitative trait loci. BMC Proc 1 Suppl 1: S144.
27. LEVY, D., G. B. EHRET, K. RICE, G. C. VERWOERT, L. J. LAUNER et al., 2009 Genome-wide association study of blood pressure and hypertension. Nat Genet.
28. LINDGREN, C. M., I. M. HEID, J. C. RANDALL, C. LAMINA, V. STEINTHORSDOTTIR et al., 2009 Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. PLoS Genet 5: e1000508.
29. PARDO, L. M., I. MACKAY, B. OOSTRA, C. M. VAN DUIJN and Y. S. AULCHENKO, 2005 The effect of genetic drift in a young genetically isolated population. Ann Hum Genet 69: 288-295.
30. PATTARO, C., Y. S. AULCHENKO, A. ISAACS, V. VITART, C. HAYWARD et al., 2009 Genome-wide linkage analysis of serum creatinine in three isolated European populations. Kidney Int 76: 297-306.
31. PROKOPENKO, I., C. LANGENBERG, J. C. FLOREZ, R. SAXENA, N. SORANZO et al., 2009 Variants in MTNR1B influence fasting glucose levels. Nat Genet 41: 77-81.
32. RICHARDS, J. B., F. K. KAVVOURA, F. RIVADENEIRA, U. STYRKARSDOTTIR, K. ESTRADA et al., 2009 Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. Ann Intern Med 151: 528-537.
33. RIVADENEIRA, F., U. STYRKARSDOTTIR, K. ESTRADA, B. V. HALLDORSSON, Y. H. HSU et al., 2009 Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. Nat Genet 41: 1199-1206.
34. SERVICE, S., J. DEYOUNG, M. KARAYIORGOU, J. L. ROOS, H. PRETORIOUS et al., 2006 Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 38: 556-560.
35. VAN DIEMEN, C. C., D. S. POSTMA, Y. S. AULCHENKO, P. J. SNIJDERS, B. A. OOSTRA et al., 2009 Novel strategy to identify genetic risk factors for COPD severity: a genetic isolate. Eur Respir J.
36. VASAN, R. S., N. L. GLAZER, J. F. FELIX, W. LIEB, P. S. WILD et al., 2009 Genetic variants associated with cardiac structure and function: a meta-analysis and replication of genome-wide association data. JAMA 302: 168-178.
37. VINK, J. M., A. B. SMIT, E. J. DE GEUS, P. SULLIVAN, G. WILLEMSEN et al., 2009 Genome-wide association study of smoking initiation and current smoking. Am J Hum Genet 84: 367-379.

Подписано к печати 24.06.2010 г.
Формат бумаги 60 х 90 1/16, печ. л. 2, уч. изд. л.1,4
Тираж 110 Заказ № 63

Ротапринт Института цитологии и генетики СО РАН
630090, Новосибирск, пр. акад. Лаврентьева, 10