На правах рукописи

Львовс Дмитрийс

СТАТИСТИЧЕСКИЙ ПОИСК И ВАЛИДАЦИЯ БИОМАРКЕРОВ, СВЯЗАННЫХ С РАССЕЯННЫМ СКЛЕРОЗОМ

03.01.09 Математическая биология, биоинформатика

А В Т О Р Е Ф Е Р А Т

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2012

Работа выполнена в Государственном научном центре Российской федерации Федеральном государственном унитарном предприятии «Государственный научно-исследовательский институт генетики и селекции промышленных микроорганизмов».

Научный руководитель:
кандидат физико-математических наук Александр Владимирович Фаворов

Научный консультант:
доктор биологических наук, профессор Ольга Олеговна Фаворова

Официальные оппоненты:
кандидат физико-математических наук, доктор биологических наук, профессор
Андрей Александрович Миронов
Факультет биоинженерии и биоинформатики
Федерального государственного бюджетного образовательного
учреждения высшего профессионального образования
Московского государственного университета им. М.В. Ломоносова,
профессор

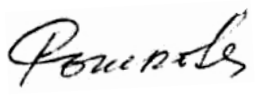
доктор физико-математических наук, профессор
Михаил Абрамович Ройтберг
Федеральное государственное бюджетное учреждение науки
Институт математических проблем биологии Российской академии наук,
заведующий лабораторией

Ведущая организация:
Федеральное государственное бюджетное учреждение науки
Институт общей генетики им. Н.И. Вавилова Российской академии наук

Защита диссертации состоится 29 октября 2012 года в 16-00 часов на заседании диссертационного совета Д002.077.04 на базе Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук по адресу: 127994, г. Москва, ГСП-4, Большой Каретный переулок, д. 19, стр.1.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук.

Автореферат разослан 28 сентября 2012 г.

Ученый секретарь диссертационного совета
доктор биологических наук, профессор  Рожкова Г.И.

Общая характеристика работы

Актуальность темы

Абсолютное большинство заболеваний человека относится к полигенным заболеваниям. Причины возникновения многих из них до конца не ясны, хотя выявлено множество факторов предрасположенности, включая как генетические, так и факторы окружающей среды. Часто комбинации из аллельных вариантов генов ассоциированы с заболеванием заметно сильнее, чем их составляющие, но поиск таких составных (композитных) генетических биомаркеров затруднён по причине комбинаторного взрыва и может быть эффективным только при наличии мощного арсенала прикладных средств математической статистики.

Рассеянный склероз (РС) часто рассматривают как классическое полигенное заболевание, которое может служить прототипом для разработки новых методов полигенного анализа. РС представляет собой тяжелое хроническое воспалительное заболевание центральной нервной системы, характеризующееся многоочаговой демиелинизацией, что может приводить к расстройству органов осязания, обоняния, зрения, нарушению опорно-двигательного аппарата. РС представляет собой серьезную медико-социальную проблему из-за его высокой распространённости среди населения (в Европе около 80 больных РС на 100 000 человек), хронического течения и частой инвалидизации, а также раннего возраста начала заболевания (в типичных случаях в возрасте от 16 до 45 лет) и высокой стоимости терапии. Недавние исследования показали, что расходы, связанные с РС в Европе, составляют более 12 миллиардов Евро в год.

Существует несколько препаратов иммуномодулирующего действия, которые используются в терапевтических целях при РС. Из них препаратами первой линии являются глатирамира ацетат и интерферон бета. Долгосрочное использование этих препаратов уменьшает количество рецидивов, задерживает формирование новых поражений и прогрессирование инвалидности, однако часто оказывается неэффективным у значительной части пациентов (от 30 до 50%). Существование проблемы гетерогенности ответа на лечение стало стимулом для развития персонализированной медицины.

Одним их основных направлений современной биоинформатики, находящейся на стыке медицины, биологии и вычислительных технологий, является разработка методов, которые позволили бы применить достижения нового поколения биотехнологий, создающих всё

большие и большие объемы биологических данных, для решения задач персонализированной медицины.

Цели и задачи исследования

Цель настоящей работы состояла в решении применительно к РС некоторых биоинформатических задач, направленных на поиск составных (композитных) генетических биомаркеров, предсказывающих индивидуальную предрасположенность к гетерогенному полигенному заболеванию и характеру его течения, а также индивидуальный ответ на то или иное лечение.

В задачи работы входило:

- Создание алгоритма валидации генетических биомаркеров и его реализация в виде программного обеспечения (ПО) на основе алгоритма APSampler, ранее разработанного для поиска составных генетических биомаркеров методом Монте-Карло Марковскими цепями (MCMC).
- Разработка языка передачи вероятностной информации о свойствах генов на основе стандарта XML и его интеграция с системой APSampler.
- Разработка методов определения характера кумулятивного эффекта между компонентами составных генетических биомаркеров.
- Разработка методов представления и визуализации результатов поиска составных биомаркеров и оценки характера кумулятивного эффекта в них.
- Выбор, в зависимости от поставленных задач, оптимального метода поиска составных генетических биомаркеров из имеющегося арсенала методов.
- Использование разработанных методов для выявления составных генетических маркеров предрасположенности к РС, характера течения заболевания и ответа на лечение по экспериментальным данным, полученным методом случай/контроль.
- Поиск составных генетических биомаркеров предрасположенности к рассеянному склерозу другим методом полигенного анализа ассоциаций – на основании семейных данных – при помощи ПО Famhap, основанного на тесте неравновесной передачи аллелей (TDT).

Научная новизна и практическая значимость

Разработан и внедрен метод валидации найденных программой APSampler составных и одиночных генетических биомаркеров, основывающийся на критерии гамма Гудмана-

Крускалла и позволяющий в качестве фенотипического признака использовать упорядоченный ряд, например, шкалу инвалидизации EDSS в случае РС.

Впервые проведено разностороннее сравнение наиболее известных методов полигенного анализа, включая такие средства как MDR, PLINK, BEAM, LogReg и APSampler, облегчающее пользователю выбор инструмента, адекватного его задачам.

Впервые разработан набор вычислительных методов, проверяющих наличие эпистаза, то есть нелинейного взаимодействия между частями составного биомаркера, по критериям значимости, сходным с применяемыми для оценки ассоциации.
Разработана программа для графического отображения до пяти взаимодействующих параметров в виде диаграммы Венна.

На основе популярного языка XML разработан новый формат передачи генетической информации OnionTree XML.

В процессе работы созданы следующие web-ресурсы: code.google.com/p/vienna5/, onion-xml.sourceforge.net, code.google.com/p/apsampler/.

Апробация работы

Результаты работы регулярно представлялись на конференциях сети UEPHA*MS, созданной под эгидой седьмой рамочной программы Евросоюза: Joint EUROKUP and UEPHA*MS workshop, Rotterdam 2010; UEPHA*MS Summer School, Berlin 2011; UEPHA*MS Autumn School, Barcelona 2011; Multiple Sclerosis and the Omics Spring Conference, Bilbao 2012. Они были также доложены на следующих международных конференциях: Moscow Conference on Computational Molecular Biology, Moscow, 2011; 1st International SystemsX.ch Conference on Systems Biology, Basel 2011; BIO IT World Asia, Singapore 2012.

Структура и объем работы

Работа выполнена на 107 страницах машинописного текста, состоит из введения, четырёх разделов, выводов и списка цитируемой литературы, содержит 12 иллюстраций и 11 таблиц. Список цитируемой литературы содержит 151 наименование.

Во «**Введении»** дано краткое обоснование важности исследований генетической природы РС, кратко описаны существующие в наше время проблемы, связанные с прогнозированием

и терапией полигенных заболеваний, а также факторы, способствующие решению этих проблем.

Раздел **«Обзор литературы»** посвящен детальному рассмотрению достижений современной науки в разработке подходов к исследованию генетической природы полигенных заболеваний, к которым относится РС. Рассмотрены некоторые подходы к лечению РС, описаны достижения персонализированной медицины и некоторые факторы, ограничивающие ее клиническое применение.

Также описаны исследования последних лет по поиску генетических биомаркеров предрасположенности к РС и биомаркеров эффективности ответа на лечение РС препаратами первой линии – интерфероном-бета и глатирамера ацетатом (Копаксоном); обоснована необходимость поиска именно составных генетических биомаркеров. Наконец, рассмотрены типы исследований, подходящих для решения этой задачи, используемые при этом математические методы, а также известные компьютерные программы, в которых реализованы описанные подходы; дана оценка функций этих программ с точки зрения пользователя.

Раздел **«Методы»** содержит описание основных методов и программ, как известных, так и оригинальных, разработанных непосредственно для решения поставленных задач. Первая глава посвящена разработке программного обеспечения APSampler. Вторая глава описывает язык ONION-TREE XML, предназначенный для передачи вероятностной информации о свойствах генов. Две последующие главы описывают методы, связанные с оценкой кумулятивного эффекта отдельных компонентов составных генетических биомаркеров. Эти методы основываются на анализе таблиц сопряженности. Проведена проверка правильности оценки доверительного интервала (CI) для критерия SF (фактор синергии) (Cortina-Borja, 2009), и подтверждена обоснованность предложения при расчете не учитывать корреляции переменных, на основе которых оценивают дисперсию величины SF. Предложен метод оценки взаимодействия ORR, имеющий ряд преимуществ перед критерием SF. Наконец, предложена адаптация уже существующего (White et al, 1973)., но еще не применявшегося в биологии точного метода оценки нелинейного взаимодействия частей составных генетических биомаркеров. В последней главе предложен инструмент для создания векторных диаграмм Венна для 5 и менее пересекающихся множеств, удобный для представления результатов тестов на эпистаз.

Раздел **«Результаты»** открывается описанием ПО APSampler. Затем компьютерные программы, описанные в обзоре литературы, анализируются с точки зрения возможностей

найти, исходя из одинаковых данных, составные генетические маркеры, и делается выбор в пользу наиболее чувствительной из программ. Далее описаны найденные при анализе методом случай/контроль составные генетических биомаркеры, ассоциированные с развитием РС, а также с характером его течения и эффективностью лечения иммуномодулирующими препаратами. Эти результаты получены с помощью представленных в разделе «Методы» оригинальных приёмов анализа и визуализации данных. Также в разделе представлены результаты анализа предрасположенности на семейном материале, которые получены с помощью известного, описанного в главе «Обзор литературы», метода Famhap (Becker and Knapp, 2009).

Методы

Ниже без ссылок описаны методы и программы, разработанные в ходе представленной работы. Если представление требует описания известного метода, на последний даётся отсылка.

Алгоритм валидации генетических биомаркеров и его реализация в виде программного обеспечения на основе алгоритма APSampler

Степень тяжести заболевания РС часто описывают, используя многоуровневые международные шкалы заболевания EDSS и MSSS. Одним из достоинств алгоритма APSampler (Favorov, 2005) является использование рангового критерия Вилкоксона, что даёт возможность для поиска ассоциаций использовать ранжируемые фенотипы и полиаллельные гены. Недостатком, ограничивающим применение APSampler при ранговом многоуровневом фенотипе, являлось отсутствие методов валидации результатов, полученных на таких данных. Для валидации ассоциации найденных APSampler составных генетических биомаркеров с многоуровневым фенотипом нами был использован критерий гамма Гудмана-Крускалла.

Был создан документированный пользовательский комплект ПО и web-сайт http://code.google.com/a/apsampler. Разработана система параллельного запуска программы валидации, адаптируемая под различные типы многозадачных ОС. В качестве рабочих примеров реализованы и отлажены версии для стандартной UNIX-подобной многозадачности (nohup &) и для кластеров SGE. Кроме того, созданы скрипты для конвертации данных из/в другие популярные форматы, принятые при решении подобных задач (lgen, ped, vcf). Введены другие улучшения алгоритма и самого ПО.

При анализе ассоциации может оказаться полезным дополнение исходных данных какими-либо сведениями об исследуемых генах. Такими сведениями могут быть, например, списки генов, в которых находятся анализируемые однонуклеотидные полиморфизмы (SNP), либо метаболические пути, в которых участвуют рассматриваемые гены. Такая информация может быть применена на любых этапах исследования, например, для аннотации полученных в ходе анализа результатов или в самой процедуре поиска в виде априорных вероятностей.

Для осуществления возможности передавать между различными приложениями информацию о генах, их аллелях или сходных объектах в условно-вероятностной, вероятностной или текстовой формах на основе синтаксиса популярного языка описания данных XML, был создан язык ONION-TREE XML, схема которого отображена на Рис. 1.

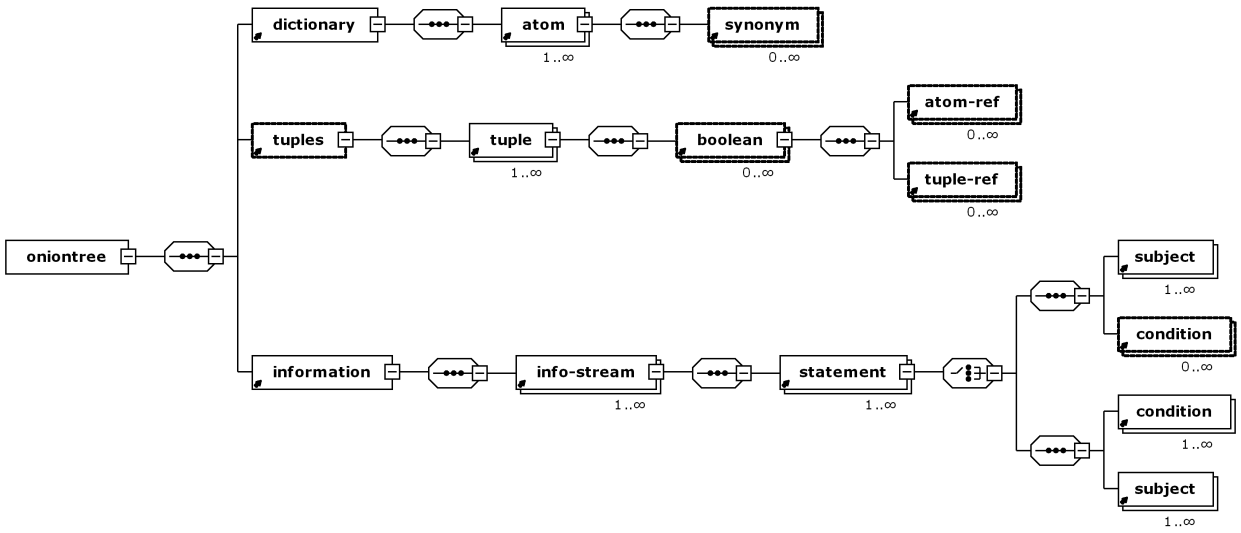


Рис. 1. Схема (XML schema) языка для передачи структурированной информации ONION-TREE XML, позволяющая передавать информацию как иерархического, так и вероятностного характера.

Главным контейнером ONION-TREE XML является объект <oniontree>, атрибуты которого позволяют хранить описание данных, которые можно ожидать в данном файле. Объект <dictionary> позволяет хранить информацию о возможных синонимах главного элемента информации, и сам <atom>. Каждый <atom> имеет свой уникальный идентификатор, который используется далее в описании данных. <tuple> представляют собой

сгруппированные посредством <boolean> различные <atom>, при этом <tuple> может также содержать и другие <tuple>, а может и состоять только из комбинаций <atom>. Наряду с обычной группировкой данных, для реализации которой хватает <atom>, <boolean> и <tuple>, ONION–TREE может также применяться для передачи вероятностной информации, например о сцеплении, или априорных вероятностях, связанных с генетическими данными. В этом случае связанные с <atom> какие-либо утверждения, а так же идентификаторы <atom> являются частью <infostream>, при чем каждое из утверждений представляется как отдельный объект <statement>, который содержит <subject>, соответствующую этому <subject> вероятность целевого высказывания (например, ассоциации) и, возможно, <condition>.

Интерпретация сообщения в формате ONION-TREE XML зависит от принимающей стороны, сам язык не описывает природы целевого высказывания.

Язык ONION-TREE использовался в ходе работы для организации данных по включениям SNP в гены и генов в метаболические пути, а также для последующей аннотации результатов работы программы APSampler исходя из переданной информации.

Методы определения характера кумулятивного эффекта компонентов составного генетического биомаркера

При полигенных фенотипических признаках эффект одного гена часто не проявляется. Комбинации из эффектов нескольких генов может иметь больший наблюдаемый эффект, чем ожидается в предположении о независимости эффектов. Напротив, эффект одного гена может нивелироваться эффектом другого. Степень отличия наблюдаемого эффекта от ожидаемого можно оценить при помощи критериев взаимодействия. Наилучшим с точки зрения авторов существующим критерием оценки взаимодействия является критерий SF (Cortina-Borja, 2009). Остальные критерии, такие как S, AP, RERI, основанные на аддитивной модели, имеют ограничения в применении. Недостатком SF является рассмотрение только полных таблиц сопряженности, по причине чего при анализе используется меньшее количество данных, чем доступно.

В настоящей работе предложено два критерия, ORR и FLINT. Оба они основываются на полных таблицах сопряженности (как в Таблице 1) и маргинальных (краевых) таблицах сопряженности, получаемых путем сложения значений по одному из свойств i,j или k числа n, записываемого в клетки таблицы (Таблица 2).

Таблица 1. Полная таблица сопряженности, используемая для оценки критериев ORR и FLINT. *A* и *B* отражают носительство генотипического признака.

A	B	случай (1)	контроль (0)
-	-	n_{100}	n_{000}
+	-	n_{110}	n_{010}
-	+	n_{101}	n_{001}
+	+	n_{111}	n_{011}

Таблица 2. Маргинальные таблицы сопряженности, получаемые путем коллапсирования n_{ijk} по j (A) и по k (B).

Аллель A	случай (1)	контроль (0)
носитель	n_{11+}	n_{01+}
неноситель	n_{10+}	n_{00+}

A

Аллель B	случай (1)	контроль (0)
носитель	n_{1+1}	n_{0+1}
неноситель	n_{1+0}	n_{0+0}

B

Дополняя друг друга, ORR и FLINT могут стать таким же стандартом для проведения анализа нелинейности взаимодействия (эпистаза), как родственные им OR и точный тест Фишера стали для ассоциативных исследований случай/контроль.

Отношение отношений шансов как критерий наличия эпистатического взаимодействия

ORR определяется, как и ранее известный критерий SF (в наших обозначениях он представлен формулой 1), как отношение отношения шансов (OR) совместного носительства к произведению двух OR носительства одного из индикаторных признаков.

Формула 1

$$SF = \frac{OR_{AB}}{OR_A \cdot OR_B} = \frac{n_{111} \cdot n_{000}}{n_{100} \cdot n_{011}} \cdot \frac{n_{100} \cdot n_{010}}{n_{110} \cdot n_{000}} \cdot \frac{n_{100} \cdot n_{001}}{n_{101} \cdot n_{000}}$$

В критерии ORR, отношения OR_A и OR_B рассчитываются как отношения шансов для всех носителей каждого из аллелей ко всем не-носителям этого аллеля (Формула 2).

Формула 2

$$ORR = \frac{OR_{AB}}{OR_A \cdot OR_B} = \frac{n_{111} \cdot n_{000}}{n_{100} \cdot n_{011}} \cdot \frac{n_{10+} \cdot n_{01+}}{n_{11+} \cdot n_{00+}} \cdot \frac{n_{1+0} \cdot n_{0+1}}{n_{1+1} \cdot n_{0+0}}$$

Такое представление OR_A и OR_B совпадает с обычным определением OR для каждого из этих признаков. При построении SF использовалась модель логистической регрессии, которая диктовала иное определение OR_A и OR_B (см. формулу 1). Таким образом, главное отличие критерия ORR от критерия SF состоит в том, что ORR использует большее количество входных данных. Доверительные интервалы для ORR, CI_{ORR} определяются способом, аналогичным оценке Вульфа (Woolf, 1955) для OR:

Формула 3

$$CI_{ORR} = e^{\ln(ORR) \pm 1.96\sigma(\ln(ORR))}$$

где $\ln(ORR) = \phi_{AB} - \phi_A - \phi_B$, $\phi_i = \ln(OR_i)$ и $\sigma_{\ln(ORR)} = \sqrt{\text{var}(\ln(ORR))}$.

Из свойств ковариации известно, что $\text{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{cov}(X_i, X_j)$ и $\text{cov}(X+Z, Y) = \text{cov}(X,Y) + \text{cov}(Z,Y)$. Тогда var(ln(ORR)) оценивается как:

Формула 4

$$\text{var}(\phi_{AB} - \phi_A - \phi_B) = \text{var}(\phi_{AB}) + \text{var}(\phi_A) + \text{var}(\phi_B) + 2\,\text{cov}(\phi_A, \phi_B) - 2\,\text{cov}(\phi_A, \phi_{AB}) - 2\,\text{cov}(\phi_B, \phi_{AB}).$$

Рассматривая все величины в клетках Таблицы 2 независимыми Пуассоновскими процессами, можно воспользоваться рядом Тэйлора и оценить неизвестные ковариации (дисперсии) как произведение производных функций и ковариаций аргументов этих функций:

Формула 5

$$\text{cov}(f(x), g(y)) = f'(x)g'(y)\text{cov}(x,y)$$

Дисперсии ϕ_{AB}, ϕ_A и ϕ_B нам известны из оценок Вульфа для соответствующих OR: $\text{var}(\phi_{AB}) = \frac{1}{n_{111}} + \frac{1}{n_{000}} + \frac{1}{n_{100}} + \frac{1}{n_{011}}$, $\text{var}(\varphi_A) = \frac{1}{n_{101}+n_{111}} + \frac{1}{n_{010}+n_{000}} + \frac{1}{n_{110}+n_{100}} + \frac{1}{n_{001}+n_{011}}$, и $\text{var}(\phi_B) = \frac{1}{n_{110}+n_{111}} + \frac{1}{n_{001}+n_{000}} + \frac{1}{n_{101}+n_{100}} + \frac{1}{n_{010}+n_{011}}$. Оценив ковариации $cov(\phi_{AB},\phi_A)$ (формула 6), $cov(\phi_{AB},\phi_B)$ по (формула 7), и $cov(\phi_A,\phi_B)$ по (формуле 8), мы сможем оценить дисперсию ORR, а так же его доверительные интервалы, подставив соответствующие результаты в Формулу 4.

Формула 6

$$\begin{aligned}&\operatorname{cov}(\phi_{AB},\phi_A)=\operatorname{cov}(\ln(n_{111})+\ln(n_{000})-\ln(n_{011})-\ln(n_{100}),\\&\ln(n_{101}+n_{111})+\ln(n_{000}+n_{010})-\ln(n_{001}+n_{011})-\ln(n_{110}+n_{100}))\\&=\operatorname{cov}(\ln(n_{111}),\ln(n_{101}+n_{111}))+\operatorname{cov}(\ln(n_{000}),\ln(n_{000}+n_{010}))+\operatorname{cov}(\ln(n_{011}),\ln(n_{001}+n_{011}))\\&+\operatorname{cov}(\ln(n_{100}),\ln(n_{110}+n_{100}))=\frac{1}{n_{101}+n_{111}}+\frac{1}{n_{000}+n_{010}}+\frac{1}{n_{001}+n_{011}}+\frac{1}{n_{110}+n_{100}}.\end{aligned}$$

Формула 7

$$\begin{aligned}&\operatorname{cov}(\phi_{AB},\phi_B)=\operatorname{cov}(\ln(n_{111})+\ln(n_{000})-\ln(n_{011})-\ln(n_{100}),\\&\ln(n_{110}+n_{111})+\ln(n_{000}+n_{001})-\ln(n_{010}+n_{011})-\ln(n_{101}+n_{100}))\\&=\operatorname{cov}(\ln(n_{111}),\ln(n_{110}+n_{111}))+\operatorname{cov}(\ln(n_{000}),\ln(n_{000}+n_{001}))+\operatorname{cov}(\ln(n_{011}),\ln(n_{010}+n_{011}))\\&+\operatorname{cov}(\ln(n_{100}),\ln(n_{101}+n_{100}))=\frac{1}{n_{110}+n_{111}}+\frac{1}{n_{000}+n_{001}}+\frac{1}{n_{010}+n_{011}}+\frac{1}{n_{101}+n_{100}}.\end{aligned}$$

Формула 8

$$\begin{aligned}&\operatorname{cov}(\phi_A,\phi_B)=\operatorname{cov}(\\&\ln(n_{101}+n_{111})+\ln(n_{000}+n_{010})-\ln(n_{001}+n_{011})-\ln(n_{100}+n_{110}),\\&\ln(n_{110}+n_{111})+\ln(n_{00}+n_{001})-\ln(n_{010}+n_{011})-\ln(n_{10}+n_{101})\\&)=\operatorname{cov}(\ln(n_{101}+n_{111}),\ln(n_{110}+n_{111}))-\operatorname{cov}(\ln(n_{101}+n_{111}),\ln(n_{100}+n_{101}))\\&+\operatorname{cov}(\ln(n_{000}+n_{010}),\ln(n_{000}+n_{001}))-\operatorname{cov}(\ln(n_{000}+n_{010}),\ln(n_{010}+n_{011}))\\&-\operatorname{cov}(\ln(n_{001}+n_{011}),\ln(n_{000}+n_{001}))+\operatorname{cov}(\ln(n_{001}+n_{011}),\ln(n_{010}+n_{011}))\\&-\operatorname{cov}(\ln(n_{100}+n_{110}),\ln(n_{110}+n_{111}))+\operatorname{cov}(\ln(n_{100}+n_{110}),\ln(n_{100}+n_{101}))=\\&\frac{n_{100}}{(n_{100}+n_{110})(n_{100}+n_{101})}+\frac{n_{000}}{(n_{000}+n_{010})(n_{000}+n_{001})}-\frac{n_{110}}{(n_{100}+n_{000})(n_{110}+n_{111})}-\frac{n_{010}}{(n_{000}+n_{010})(n_{010}+n_{011})}\\&-\frac{n_{101}}{(n_{101}+n_{111})(n_{100}+n_{101})}-\frac{n_{001}}{(n_{001}+n_{011})(n_{000}+n_{011})}+\frac{n_{111}}{(n_{101}+n_{111})(n_{111}+n_{110})}+\frac{n_{011}}{(n_{011}+n_{010})(n_{011}+n_{001})}.\end{aligned}$$

Анализ результатов применения Формулы 4 на данных, симулированных при помощи метода Монте Карло (10^6 симуляций), показал, что дисперсия, определённая по описанному выше методу, соответствует свойствам дисперсии, то есть никогда не принимает отрицательное значение. Однако, оценка дисперсии обращалась в ноль при всех одинаковых n_{ijk}. Были проведены две дополнительные проверки. Во-первых, определены частные производные $\frac{\delta \operatorname{var}(ORR)}{\delta n_{ijk}}$ и проверено, что все они обращаются в ноль при одинаковых n_{ijk}. Во-вторых, произведено детальное (сеткой с шагом 1) исследование поведения оценки около точки обращения (n_{ijk}=7,8,9) в ноль. Обе проверки подтвердили, что оценка никогда не становится отрицательной, а точки, в которых n_{ijk} равны между собой – это минимумы оценки. Учитывая, что с точки зрения исследований эпистатического взаимодействия такие точки не представляют никакого интереса (в них все OR=1, независимо от способа оценки), тот факт, что оценка дисперсии обращается в этих точках в ноль, не уменьшает её применимости для этих исследований.

В ассоциативных исследованиях широко используется точный тест Фишера, который рассматривает классические 4-польные таблицы сопряженности, описывающие населённости классов при классификации набора объектов (например, индивидов, наблюдаемых в исследовании случай/контроль) одновременно по двум дихотомическим признакам. Нулевой гипотезой теста является отсутствие связи между этими признаками. В этом тесте множеством всех возможных событий считается множество всех классификаций по этим двум признакам с такими же краевыми суммами, как в исходной классификации. Иными словами, суммы в строках и в столбцах таблицы сопряжённости остаются теми же. Все возможные классификации параметризуются одной из ячеек таблицы, например, верхнего левого угла. Сила ассоциации, соответствующая таблице сопряжённости, монотонно меняется с изменением этого параметра, а число возможных классификаций, описываемых этой таблицей, вычисляется из комбинаторных соображений. Таким образом, предполагая все классификации равновероятными, для каждой исходной таблицы сопряжённости можно определить вероятность классификации, описываемой такой или более «экстремальной» таблицей, в предположении нулевой гипотезы.

Чтобы построить аналог точного критерия Фишера для трёх признаков, мы рассматриваем таблицу сопряжённости в трёх измерениях (см. Рис. 2). Нулевой гипотезой в данном случае является то, что B не влияет на связь между статусом случай/контроль и A (см. Таблицы 2 и 3 с обозначениями) (или, наоборот, B не влияет на связь A и исхода). Все маргинальные суммы фиксированы.

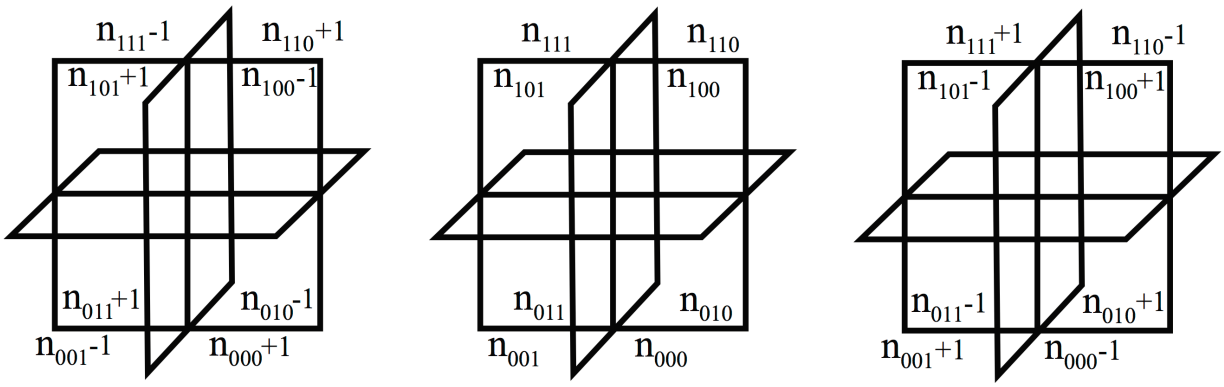


Рис. 2. Визуализация таблиц сопряженности 2x2x2 для точного теста на эпистаз. В исследовании типа случай/контроль есть n_{0++} контролей и n_{1++} случаев, а также два дихотомических фактора A и B для каждого объекта в исследовании. В центре находится таблица с наблюдаемыми данными, по аналогии с точным критерием Фишера, величина *p* определяется путем локализации данной таблицы во всей популяции таблиц.

Все возможные классификации равновероятны. Число классификаций, описанных каждой из возможных таблиц, определяется комбинаторно (формула 9):

Формула 9

$$P(n_{111}) = \binom{n_{+11}}{n_{111}}\binom{n_{+01}}{n_{101}}\binom{n_{+10}}{n_{110}}\binom{n_{+01}}{n_{101}}.$$

Все поля таблицы зависят только от одного параметра, например, n_{111}. Заметим, что степень взаимосвязанности действия факторов A и B на исход монотонно меняется с этим параметром, от наибольшего положительного влияния при максимальном значении до наибольшей компенсации при минимальном. Значение *p* рассчитывается как сумма вероятностей всех возможных (то есть с теми же маргинальными суммами) таблиц, более «экстремальных», чем наблюдаемая, например, для неожиданно маленького наблюдаемого значения n_{111}:

Формула 10

$$p = \sum_{n'_{111} \le n_{111}} P(n'_{111}).$$

Для двусторонней версии теста к результату формулы 10 необходимо добавить противоположный «хвост» распределения из класса с тем же населением, как и рассматриваемый.

Разработка методов представления и визуализации составных биомаркеров и характера кумулятивного эффекта

Одной из отличительных особенностей качественно выполненных научных работ безусловно является использование наглядных иллюстраций для упрощения анализа данных. В качестве наиболее подходящего средства визуализации данных по характеру взаимодействия между компонентами составных биомаркеров нами была выбрана диаграмма Венна. Этот вид представления результатов достаточно нагляден при количестве переменных до пяти. В настоящее время вычислительная мощность и текущие алгоритмы, используемые в методах, которые мы применяли для анализа данных, таковы, что никогда не выявляли составных маркеров из более чем пяти компонентов. Для отображения взаимодействия между составными частями биомаркера мы пробовали различные варианты диаграмм Венна, сделанные вручную с применением графических редакторов (см раздел «Результаты»), однако производство таких картинок требует больших временных затрат. Поэтому нами была создана программа, использующая к качестве образца картинку, находящуюся в свободном доступе в Интернете, соответствующие области которой

закрашиваются программой в цвет разной насыщенности, в зависимости от наблюдаемой величины эпистаза между составными частями одного биомаркера. Созданная программа доступна для скачивания в Интернете по адресу http://code.google.com/p/vienna5/.

Основные результаты работы

Программное обеспечение APSampler

Для поиска составных генетических биомаркеров в ассоциативных исследованиях в работе создана многоступенчатая параллелизуемая процедура обработки и анализа данных (Рис. 3).

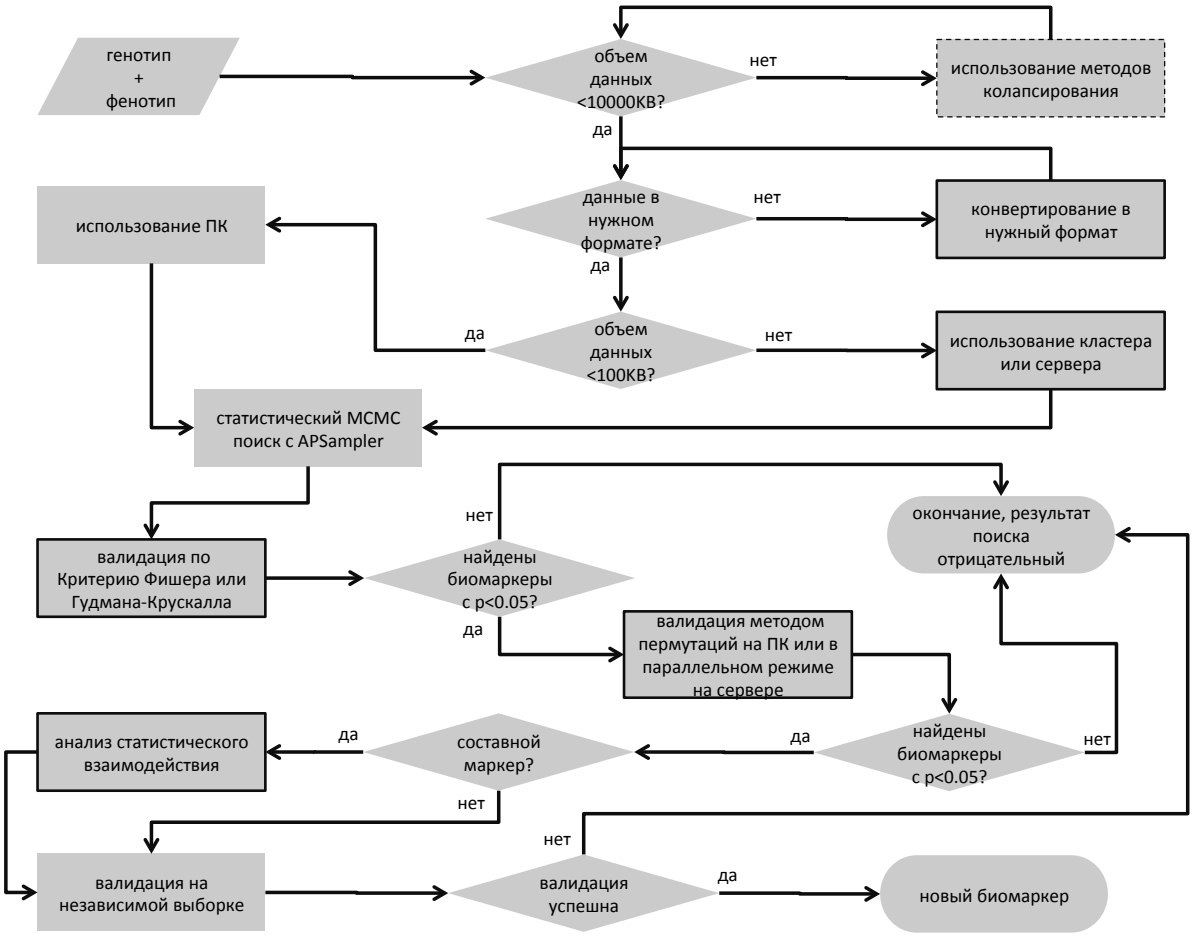


Рис. 3. Диаграмма, отражающая многоступенчатую процедуру анализа данных, предложенную в данной работе и реализованную в ПО APSampler. Ромбами обозначены решения, прямоугольниками – процессы, овалами – конец процедуры. Пунктиром отмечен предложенный теоретический процесс, не применявшийся в работе. Темным контуром выделены процессы, разработанные в данной работе.

Эта процедура реализована в ПО APSampler, которое находится в открытом доступе по адресу http://code.google.com/a/apsampler.

Выбор оптимального метода поиска составных генетических биомаркеров из имеющихся методов

С целью выбора наиболее подходящего метода поиска ассоциаций произведено сравнение BEAM, LogReg, MDR, plink и APSampler. Основные параметры сравниваемых программ приведены в таблице 3.

Таблица 3. Краткое сравнение возможностей различных ПО для полигенного анализа ассоциаций.

	APSampler	BEAM	LogicReg	MDR	plink
Графический пользовательский интерфейс	-	-[1]	-[2]	+	+
Дихотомический фенотипический признак	+	+	+	+	+
Ранговый фенотипический признак	+	-[3]	-	-	-
Работа с пропущенными данными	+	+	+	-[4]	+
Статистический поиск комбинаций конкретных аллелей локусов, ассоциированных с фенотипом	+	+	+	-[5]	+[6]
Точный критерий Фишера	+	+	-	-	-
Процедура валидации	+	+	+	+	-
Полиаллельные локусы	+	-[7]	-	+	-
Поиск эпистаза	+[8]	+	+	+	+
Графическое отображение эпистаза	-	-	-	+	-
Возможность проведения анализа ассоциации для комбинации аллелей, указанной пользователем	+	-	-	+	-[9]
Полногеномный анализ	-	+	-	-[10]	+
Возможность запуска из командной строки	+	+	+	+	+
Параллельные вычисления	+	-[11]	-	-[10]	-

[1] Существует версия BEAM, интегрированная в серверное приложение GALAXY. [2] Алгоритм реализован в пакете для статистических вычислений и графики R. [3] ПО автоматически разделяет данные на две категории, используя для этого среднее значение. [4] Авторами предлагается специальное ПО – MDR Data Tool для заполнения пустых значений. [5] Программа находит взаимодействующие и ассоциированные с фенотипом локусы, а не их аллели. [6] Предлагается только попарный поиск. [7] Количество аллелей в каждом локусе должно быть одинаковым. [8] Несмотря на то, что поиск эпистатически взаимодействующих аллелей не объявлен конкретной функцией программы APSampler, опыт практического применения ПО указывает на возможность применения данной программы для поиска эпистаза. [9] Предлагается анализ ассоциации гаплотипа. [10] Для этой цели предусмотрено специальное ПО. [11] Отдельное ПО PBEAM для параллельных исчислений.

Для сравнения алгоритмов поиска на практике, включенные в таблицу 3 программы запускали в пользовательском режиме на одних и тех же данных. Вывод LogrReg требовал дополнительной обработки, а вывод BEAM не содержал значимых результатов при заданном пороге значимости. Результаты MDR, plink и APsampler во многом совпадали (Рис. 4). При этом APSampler находил больше комбинаций, чем остальные программы, и каждая из его находок, прошедшая валидацию пермутациями, была подтверждена хотя бы одной из программ. При такой чувствительности, вывод APSampler часто содержал слишком большое количество результатов, которые должны были быть отфильтрованы.

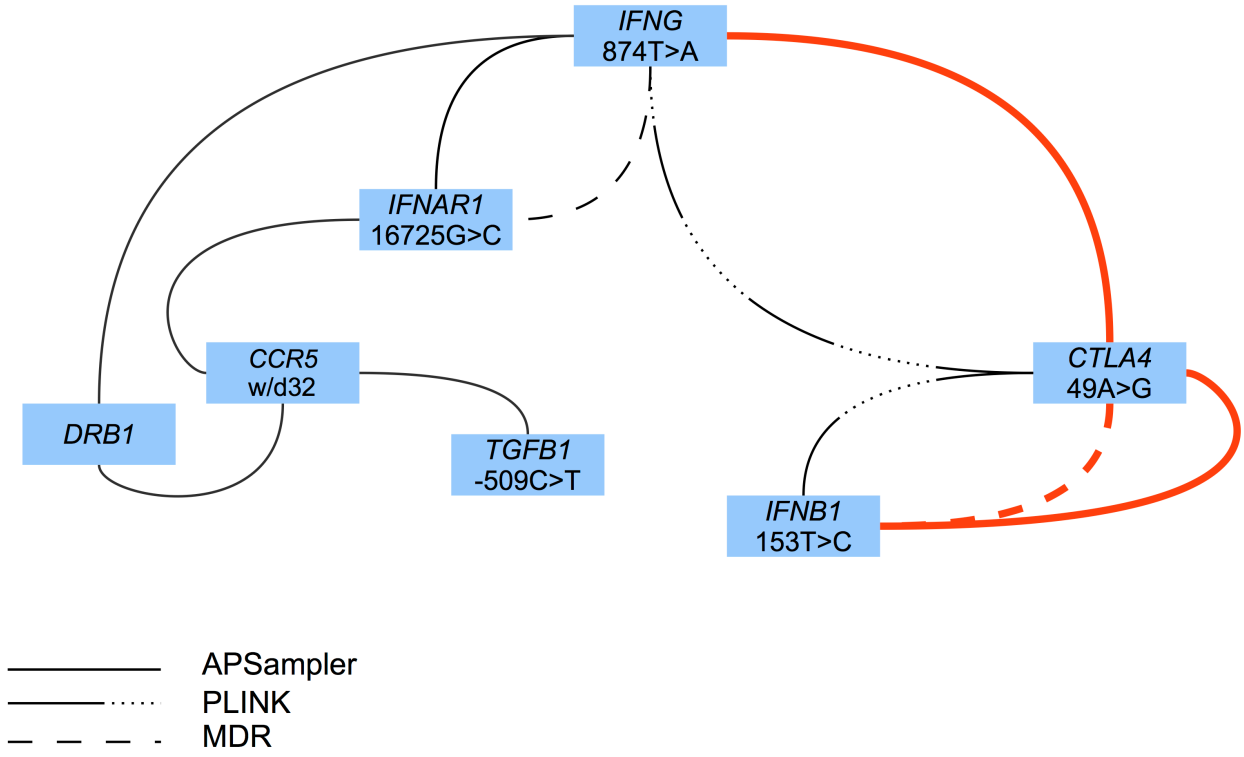


Рис. 4. Поиск программами APSampler, MDR и PLINK биаллельных сочетании генов иммунного ответа, ассоциированных с эффективностью лечения РС препаратом глатирамера ацетатом. Программа APSampler находит все биаллельные маркеры, определенные другими программами, а также идентифицирует другие сочетания. Жирными линиями отмечены сочетания, ассоциация которых с эффективностью лечения прошла валидацию пермутациями в программе APSampler ($p < 0.1$) или кросс-валидацию программой MDR (CVC > 8/10).

Как следует из результатов сравнения, выбор программы APSampler для полигенного анализа данных полностью обоснован. Ниже описаны проведенные применительно к РС исследования, в которых для анализа результатов в основном использовалось ПО APSampler.

Поиск биомаркеров эффективности лечения РС иммуномодулирующими препаратами

В этих исследованиях впервые введен критерий ORR как мера эпистатического взаимодействия между компонентами составных биомаркеров.

Они выполнены в сотрудничестве с Московским городским центром рассеянного склероза и кафедрой молекулярной биологии и медицинской биотехнологии Государственного бюджетного образовательного учреждения высшего профессионального образования «Российский национальный исследовательский медицинский университет имени Н. И. Пирогова» Минздравсоцразвития РФ. С помощью ПО APSampler проведен анализ возможной ассоциации носительства сочетаний аллелей и/или генотипов девяти генов иммунного ответа (*DRB1, TNF, IFNAR, IFNB1, IFNG, TGFB1,CCR5, CTLA4, IL7RA*) с эффективностью лечения больных РС двумя иммуномодулирующими препаратами первой линии: глатирамера ацетатом (Копаксона) и интерфероном бета (Царева и др., 2011; Tsareva et al., 2012). Особенностью дизайна этого исследования является его унифицированность. Для обоих препаратов использовались одинаковые клинические критерии оценки эффективности лечения, на основании которых формировались группы сравнения, и одна и та же панель типируемых полиморфных участков генов-кандидатов. Больные одинаковой этнической принадлежности (русские), все проживающие в Московском регионе, находились под наблюдением в одном медицинском учреждении.

Лечение больных РС глатирамера ацетатом

В исследовании участвовало 285 больных РС. Не выявлено значимых ассоциаций ни с одним из исследованных полиморфизмов по отдельности. Носительство сочетаний аллелей 4-х генов (*DRB1**15 + *TGFB1**509T + *CCR5**d + *IFNAR1**16725G) увеличивало в 14 раз риск неэффективного лечения глатирамера ацетатом (OR = 0.072 [CI = 0.02-0.28], p = 0.00018). При этом ассоциация выдерживала пермутационный тест (p_{perm} = 0.0056).

На рис. 5 в виде диаграммы Венна представлены цветами разной насыщенности величины ORR, характеризующие взаимодействие компонентов этого «неблагоприятного» аллельного сочетания при всех возможных комбинациях входящих в него аллелей.

Триаллельное сочетание (*DRB1**15 + *CCR5**d + *TGFB1**–509T) как маркер неэффективного лечения мало отличалось от 4-аллельного, тогда как ассоциация всех остальных компонентов последнего с неэффективным лечением была существенно слабее.

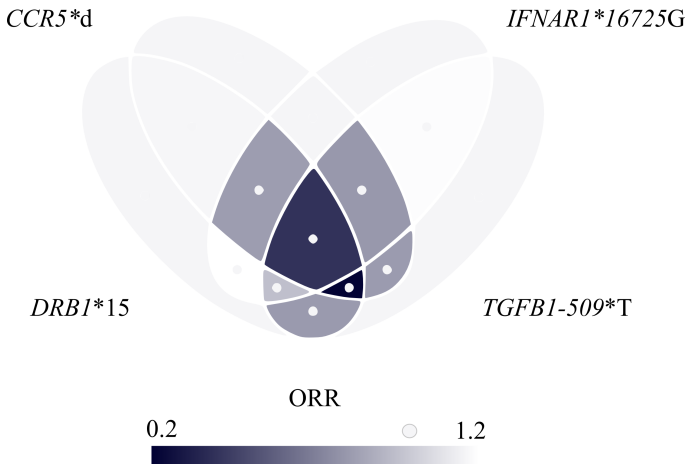


Рис. 5. Диаграмма Венна, характеризующая взаимодействия компонентов сочетания *DRB1**15 + *TGFB1**–509T + *CCR5**d + *IFNAR1**16725G, негативная ассоциация которого с эффективностью лечения РС глатирамера ацетатом выявлена с помощью ПО APSampler (Tsareva et al., 2012). Каждый из четырех эллипсов диаграммы соответствует одному из аллелей этого сочетания. Области пересечения эллипсов дают все возможные комбинации из 4 аллелей, при этом интенсивность цвета отражает значение ORR, в соответствии с градиентной шкалой, представленной внизу рисунка. Области, соответствующие отдельным аллелям, а также маленькие референтные кружки соответствуют ORR, равному 1.

В случае триаллельного сочетания (*DRB1**15+*CCR5**d+*TGFB1**–509T) ORR составляло 0.2, т.е. в 5 раз отличалось от 1, и не менялось при добавлении аллеля *IFNAR1**16725G. Мы рассматриваем эти данные как указание на эпистатическое взаимодействие аллелей генов *DRB1*, *CCR5* и *TGFB1*.

Лечение больных РС интерфероном бета

В исследовании участвовало 253 больных РС. Показано, что носительство каждого из аллелей *TGFB1**-509C и *CCR5**d по отдельности ассоциировано с благоприятным ответом на лечение интерфероном бета (IFNβ). Аллели *CCR5**d, *IFNAR1**16725G, *IFNG**874T и *IFNB1**153T/T были компонентами сочетаний, носительство которых было значительно выше в группе пациентов с оптимальным ответом на IFNβ. Триплеты (*CCR5**d + *IFNAR1**G + *IFNB1**T/T) и *(CCR5**d + *IFNAR1**G + *IFNG**T) могут рассматриваться в качестве составных маркеров, ассоциированных с оптимальным ответом на IFNβ (p_{perm} = 0.017 и 0.035, соответственно); их носительство повышало вероятность эффективного лечения в 14,3 и 2,8 раз соответственно.

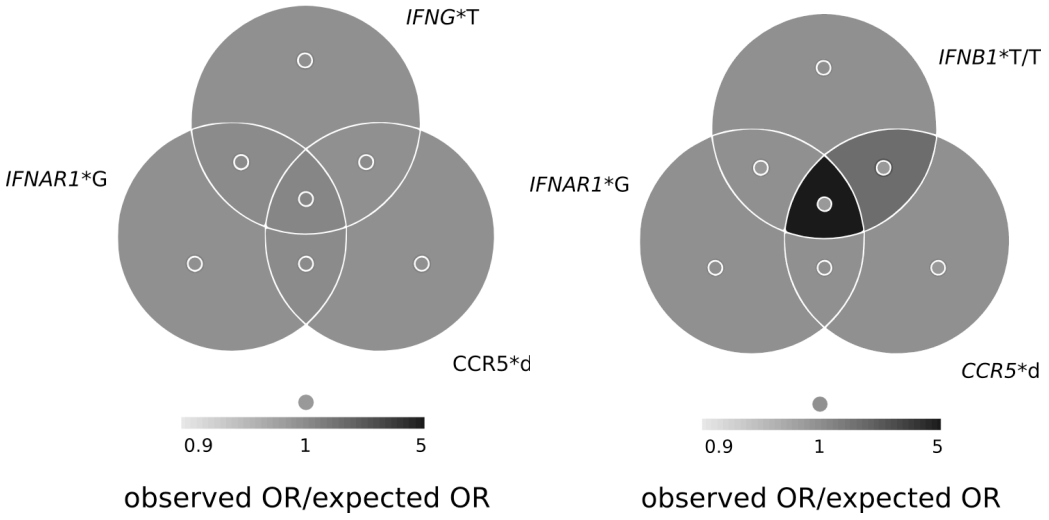


Рис. 6. Диаграмма Венна, характеризующая взаимодействия между компонентами сочетания *CCR5**d + *IFNAR*1*G + *IFNB1**T/T (справа) и сочетания *CCR5**d + *IFNAR1**G + *IFNG**T (слева). Описание графического представления дано на рис. 5.

На Рис. 6 в виде диаграммы Венна представлен анализ возможных эпистатических эффектов между компонентами названных триплетов. Виден возможный эпистаз между компонентами сочетания *CCR5**d + *IFNAR*1*G + *IFNB1**T/T и отсутствие мультипликативных эффектов между компонентами сочетания *CCR5**d + *IFNAR1**G + *IFNG**T. Полученные данные свидетельствуют о кумулятивном эффекте генов иммунного ответа на эффективность лечения РС IFNβ. Этот совместный вклад может отражать аддитивный эффект независимых аллелей отдельных генов, а также эпистатическое взаимодействие между некоторыми их них .

Оба фармакогеномных исследования свидетельствуют о продуктивности подхода, основанного на изучении совместного вклада вариантов генов, который обеспечивает более точный прогноз эффективности ответа на лечение, чем отдельные гены.

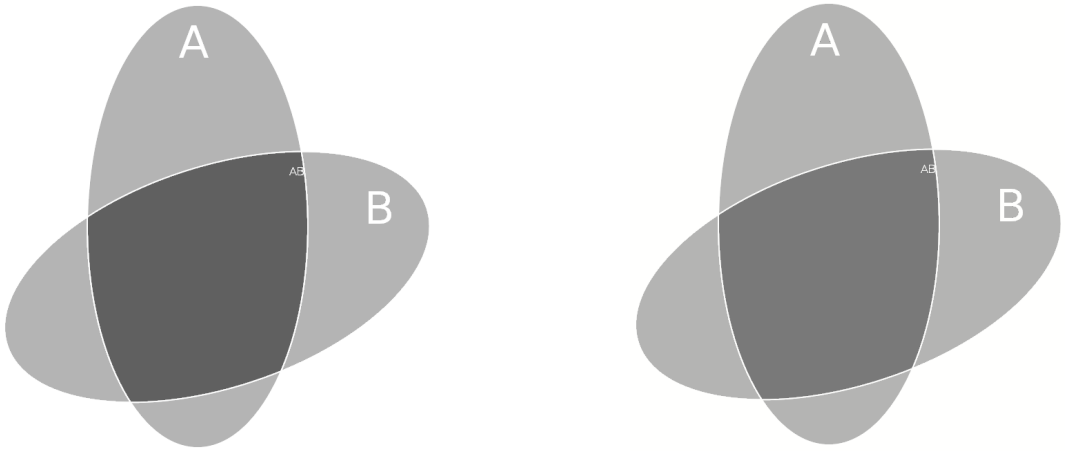
Анализ предрасположенности к РС методом случай/контроль

В этих исследованиях впервые была применена программа графического изображения характера эпистатического взаимодействия между компонентами составного биомаркера в виде диаграмм Венна (доступна в Интернете по адресу http://code.google.com/p/vienna5/).

Мы провели полигенный анализ предрасположенности к РС у европеоидов, проживающих в Германии, при помощи ПО APSampler. Данные о геномном типировании 601 больного и 600 контролей по 12 полиморфным участкам пяти известных генов-кандидатов,

локализованных в области главного комплекса гистосовместимости (HLA класс II): *DQB1*, *DRB1*, *DRB3*, *DRB4*, *DRB5*, и 7 SNP из той же области генома, были предоставлены в рамках сотрудничества в сети UEPHA*MS Робертом Гёрчесом (Robert Goertches) из Ростокского Университета. Ассоциации носительства 60 сочетаний аллелей или одиночных аллелей характеризовались уровнем значимости $p_{Wesfall\text{-}Young}$<=0.01 (на 100 перемешиваниях) (Westfall, Young, 1993). Такое большое количество статистически значимых сигналов требует дополнительного анализа результатов поиска.

С помощью известного критерия SF (Cortina-Borja, 2009) мы провели анализ характера взаимодействия между составными частями всех маркеров, проходящих установленный порог значимости. Статистически значимый эпистаз был обнаружен только у двух пар маркеров, а именно у сочетаний аллелей генов *DQB1* и *DRB3* и аллелей генов *DQB1* с *DRB5*: (SF=5.93, $CI_{99\%}$ 2.85..12.33, и SF=4.54, $CI_{99\%}$ 1.94..10.62, соответственно). Результат визуализации программой vienna5 взаимодействия компонентов этих двух комплексных маркеров предрасположенности к РС, сочетаний *DQB1**0602 + *DRB3* и *DQB1**0602 + *DRB5*, представлен на Рис 7.



1 2

Рис. 7. Анализ взаимодействия между компонентами составных маркеров предрасположенности к РС, сочетаний *DQB1**0602 + *DRB3* (панель 1) и *DQB1**0602 + *DRB5* (панель 2). Обозначения: A - *DQB1**0602 на обоих панелях, B - *DRB3* на панели 1 и *DRB5* на панели 2.

Эти результаты предполагают, что носительство *DRB5* и *DRB3,* минорных локусов, кодирующих бета-цепь молекул главного комплекса гистосовместимости класса II, влияет на предрасполагающую к развитию РС роль аллеля *DQB1**0602.

Во второй части этого исследования мы разделили выборку пациентов на две подгруппы: больные с первично-прогрессирующей формой РС, течение болезни у которых происходит без ремиссий, и все остальные больные с различными формами течения заболевания, характеризующиеся наличием ремиссий, и проверили ассоциацию генотипа с таким признаком. Из-за малого количества больных с первично-прогрессирующей формой РС (49 человек), мы, чтобы уменьшить влияние текущей выборки (эффект переобучения), не рассматривали ассоциации с составными маркерами. Выявлен один маркер – аллель *DRB5*:y, достоверно ассоциированный с наличием ремиссий (p_{perm} = 0.004, OR=2.2 CI(95%)=[1.2..3.9]).

В рамках программы UEPHA*MS в лаборатории Куна Ванденбрука (Koen Vandenbroeck) в Университете страны басков (г. Бильбао), были генотипированы все (388) SNP, отвечающие за вариацию в популяции басков более 50 генов цитокинов, включая регуляторные области. В исследование было включено 462 пациента с РС и 470 человек контрольной группы. Эта выборка уже использовалась для поиска единичных аллелей, по ней были получены значимые новые результаты (Vandenbroeck et al., 2011). С помощью ПО APSampler нами было найдено 13 составных маркеров, прошедших валидацию методом Вестфола-Янга ($p_{Wesfall\text{-}Young}$<=0.01, в таблице 4 представлены 3 лучших из них), все они содержали аллель *DRB1**1501 главного комплекса гистосовместимости – самый известный генетический фактор предрасположенности к РС. Однако, ассоциация носительства этого аллеля как такового не выдерживала валидации методом Вестфола-Янга.

Таблица 4 Три наиболее сильно ассоциированных с РС составных генетических биомаркера

Маркер	Сочетание аллелей	Параметры ассоциации
1	rs700179:A; rs9509348:A; rs1554999:A; rs3135388:A.	P_{Fisher} = 5.52e-12 OR=11.23314 CI(95%)=[4.71830..26.74342] P_{perm} = <1/200000000 $p_{Wesfall\text{-}Young}$<=0.01
2	rs7119430:A; rs1939152:C; rs875643:G; rs3135388:A.	P_{Fisher} = 1.82e-11 OR=5.21303 CI(95%)=[3.08653..8.80462] P_{perm} = 1.5e-08 $p_{Wesfall\text{-}Young}$<=0.01
3	rs1569922:A; rs875643:G; rs3135388:A.	P_{Fisher} = 3.1e-11 OR=8.73204 CI(95%)=[4.05198..18.81762] P_{perm} = 1.75e-07 $p_{Wesfall\text{-}Young}$<=0.01

Дальнейший анализ был направлен на исследование взаимодействия компонентов этих составных маркеров с использованием стандартного критерия SF и разработанного нами критерия FLINT. Рассматривали только сочетания из двух компонентов, и только те из них,

для которых значения p_{perm} были не более 0.05, а CI(SF) не пересекали 1. Выявленные три эпистатически взаимодействующих биаллельных сочетания представлены в Таблице 5.

Таблица 5. Анализ эпистатического взаимодействия в парных составных генетических биомаркерах. Показаны единственные три значимые комбинации.

Маркер	Сочетание аллелей	SF [CI(95%)]	FLINT (двусторонний *p-value*)
1	rs3135388:A; rs1569922:A	4.78 [1.13..20.28]	0.0005
2	rs3135388:A; rs875643:G	6.81 [1.21..38.33]	0.0004
3	rs3135388:A; rs8177633:G	5.15 [1.15..23.12]	0.0005

Интересно заметить, что два из них (1 и 2) в совокупности формируют сильный маркер предрасположенности (OR=8.7), найденный ранее (он представлен в Таблице 4 под номером 3). Логичным следующим шагом стала оценка взаимодействия между аллелем rs3135388:A и составным маркером rs1569922:A; rs875643:G. Фактор синергии SF для такой комбинации еще более значим: SF=13.51 [1.50..121.87] (см. Рис. 8).

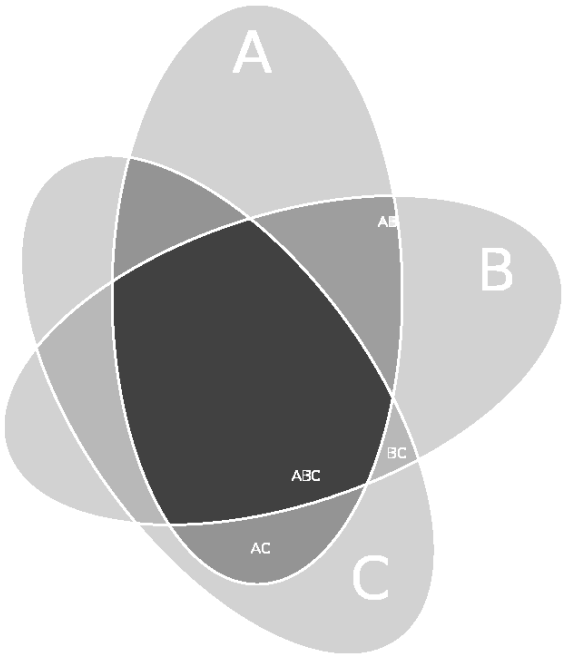


Рис. 8. Анализ взаимодействия между компонентами предрасполагающего к развитию РС сочетания аллелей rs1569922:A (A), rs3135388:A (B) и rs875643:G (C). Рисунок сделан при помощи программы vienna5.

Из этого примера видно, что одним из перспективных методов анализа эпистаза является построение сетей (графов) попарных взаимодействий маркеров на основе информации о статистической достоверности этих взаимодействий.

Поиск биомаркеров предрасположенности к РС на семейном материале

Семейный анализ, в отличие от анализа случай/контроль, не подвержен ошибкам стратификации выборки, поэтому возможность его расширения на полигенные исследования представляет большой научный и практический интерес.

Для реализации этой возможности мы применили в настоящей работе ПО Famhap (Herold and Becker, 2009) для выявления составных маркёров, ассоциированных с РС, в семейной выборке, типированной по генам *DRB1*, *CTLA4*, *TGFB1*, *IL4*, *CCR5*, *RANTES*, *MMP9*, *TIMP1*, *IL-6*, *IFNG* и *TNF*, на кафедре молекулярной биологии РНИМУ им. Н.И. Пирогова. Выборка представляла собой 104 ядерные семьи, состоящие из 106 больных РС русской этнической принадлежности и их здоровых родителей. Ассоциацию с РС перечисленных генов по отдельности проанализировали ранее методом TDT (Макарычева, 2010, 2011). В Famhap поиск реализован только методом полного перебора, что наложило ограничения на размерность составных маркёров. Поэтому сочетания с числом членов, большим, чем 2, нами не были рассмотрены. Результаты поиска парных генетических маркеров представлены в таблице 6.

Таблица 6 Анализ ассоциации с РС попарных сочетаний аллелей полиморфных участков генов *TNF, MMP9, CTLA4, IFNG, IL6, TGFB1, IL4, CCR5, RANTES* и *TIMP1* методом HAP-TDT при помощи ПО Famhap. Показаны только значимые результаты (p<0.05)

Сочетания аллелей	p (Famhap)
*IL6**(–174)G + *IFNG**(+874)A	0.043
*CCR5**w + *TNF**(–308)G	0.025

Таким образом, полигенное расширение семейного анализа представляется перспективным направлением исследований.

Заключение

В ближайшие годы будет падать цена индивидуального генотипирования, и вскоре оно превратится в рутинную процедуру. В результате произойдёт резкий скачок объёма данных по индивидуальным полиморфизмам в геноме. Полигенный анализ, оперирующий с сочетаниями генетических маркёров, скорее всего, станет главным инструментом для выявления связей генотип-фенотип. Методы полигенного анализа будут всё более и более востребованы в практике. Таким образом, представленная работа является шагом в одном из самых перспективных направлений развития современной биоинформатики.

1. Предложен алгоритм валидации ассоциаций, найденных с помощью ранее предложенного алгоритма APSampler. Это позволило создать программное обеспечение (ПО) для поиска и валидации составных генетических биомаркеров, ассоциированных с фенотипами полигенного заболевания.

2. На основе стандарта расширяемого языка разметки XML создан язык ONION TREE XML. Он предназначен для передачи информации, описывающей набор генов в виде тех или иных утверждений о свойствах этих генов и соответствующих этим утверждениям условных или безусловных вероятностей. Этот язык использован при передаче в программный комплекс APSampler данных о принадлежности генов к определенным каскадам реакций (метаболическим путям) для визуального представления результатов.

3. Разработан набор вычислительных методов, проверяющих наличие эпистаза, то есть нелинейного взаимодействия между частями составного биомаркера, по критериям значимости, сходным с применяемыми для оценки ассоциации. Предложен критерий наличия эпистаза – величина отношения отношений шансов (ORR), который включает расчёт доверительных интервалов, основанный на модели восьми независимых Пуассоновских процессов. Эта модель расширяет модель четырех независимых Пуассоновских процессов, используемую для оценки доверительного интервала отношения шансов (OR) в ассоциативных исследованиях. Предложен точный критерий анализа эпистаза FLINT, который основан на модели, аналогичной модели точного критерия Фишера, используемого в ассоциативных исследованиях.

4. Создана программа, позволяющая визуализировать характер взаимодействия между компонентами составного биомаркера с помощью построения диаграмм Венна, фрагменты которых различаются интенсивностью окраски в соответствие с величиной ORR взаимодействующих компонент.

5. Проведено сравнение широко используемых ПО поиска и валидации составных генетических биомаркеров, направленное на определение областей применения. Анализ свидетельствует, что среди сравниваемых программ ПО APSampler оптимально для решения прикладных медико-биологических задач методами исследования ассоциаций.

6. Совокупность разработанных методов успешно использована для поиска и анализа составных генетических биомаркеров, ассоциированных с развитием рассеянного склероза, с характером его течения и эффективностью лечения иммуномодулирующими препаратами, на основании экспериментальных данных, полученных методом случай/контроль. Поиск составных генетических биомаркеров предрасположенности к рассеянному склерозу проведен также методом полигенного анализа ассоциаций на семейных данных при помощи ПО Famhap, основанного на тесте неравновесной передачи сразу аллелей (TDT).

*Работа поддержана грантом European Community's Seventh Framework Programme [FP7/2007-2013] No.212877 (UEPHA*MS). Автор благодарит руководителей программы UEPHA*MS Куна Ванденброка (Koen Vandendroeck) и Ирэйд Аллоза (Iraide Alloza) за возможность участия программе.*

Автор выражает глубокую благодарность научному руководителю, Александру Владимировичу Фаворову за яркие, творческие научные дискуссии и методы работы; научному консультанту, профессору Ольге Олеговне Фаворовой за исключительные навыки и знания, полученные в ходе совместной работы; руководителю лаборатории биоинформатики ФГУП ГосНИИгенетика Всеволоду Юрьевичу Макееву за постоянную поддержку.

Список работ, опубликованных по теме диссертации

Статьи в научных журналах

Lvovs D, Фаворова ОО, Фаворов АВ. Полигенный подход к исследованиям полигенных заболеваний. Acta Naturae. 2012. т. 4 №. 3(14) с. 62-75.

Tsareva EY, Kulakova OG, Boyko AN, Shchur SG, Lvovs D, Favorov AV, Gusev EI, Vandenbroeck K, Favorova OO. Allelic combinations of immune-response genes associated with glatiramer acetate treatment response in Russian multiple sclerosis patients. Pharmacogenomics. 2012. V. 13 N. 1 pp. 43-53.

Favorov, A., D. Lvovs, W. Speier, G. Parmigiani, and M. F Ochs. OnionTree XML: A Format to Exchange Gene-Related Probabilities. Journal of Biomolecular Structure & Dynamics. 2011. V. 29, N. 2 pp. 417-423.

Царева ЕЮ, Львов ДВ, Фаворов АВ, Ochs MF, Фаворова ОО, Кулакова ОГ, Макарычева ОЮ, Бойко АН, Щур СГ, Лащ НЮ, Попова НФ, Гусев ЕИ, Башинская ВВ. Фармакогеномика рассеянного склероза: ассоциация полиморфизма генов иммунного ответа с эффективностью лечения копаксоном. Молекулярная Биология. 2011. т.45 №6 с. 963-972.

Тезисы международных конференций

Favorov, A., Lvovs, D., Sudomoina, M., Favorova, O., Parmigiani, G., Ochs, M.F. APSampler: open-source software for identifying multigene effects in genetic data. Department of Bioengineering and Bioinformatics of MV Lomonosov Moscow State University, 2011, p. 109.

Lvovs, D., Makeev, V., Fridman, M., Oparina, N. Tandem repeat polymorphisms in the human genome. Department of Bioengineering and Bioinformatics of MV Lomonosov Moscow State University, 2011, p. 208.