УДК 575.112; 577.21

Полигенный подход к исследованиям полигенных заболеваний

D. Lvovs[1*], О. О. Фаворова[2,3], А. В. Фаворов[1,4,5]

[1] Государственный научно-исследовательский институт генетики и селекции промышленных микроорганизмов, 113545, Москва, 1-й Дорожный пр., 1

[2] Российский национальный исследовательский медицинский университет им. Н.И. Пирогова Минздравсоцразвития Российской Федерации, 117997, Москва, ул. Островитянова, 1

[3] Российский кардиологический научно-производственный комплекс Минздравсоцразвития Российской Федерации, 121552, Москва, 3-я Черепковская ул., 15а

[4] Институт общей генетики им. Н.И. Вавилова РАН, 119991, Москва, ул. Губкина, 3

[5] Johns Hopkins University School of Medicine, 550 North Broadway, Baltimore, MD 21205, US

*E-mail: dmitrijs.lvovs@gmail.com

Поступила в редакцию 02.05.2012 г.

РЕФЕРАТ **Полигенные заболевания возникают в результате совместного вклада множества независимо действующих или взаимодействующих полиморфных генов, причем вклад каждого из них по отдельности может быть невелик или вовсе не проявляться. Носительство тех или иных сочетаний генов может определять возникновение клинически гетерогенных форм заболевания и эффективность лечения. Описаны подходы к полигенному анализу данных медицинской геномики, в частности фармакогеномики, направленные на выявление кумулятивного эффекта нескольких генов. Этот эффект может возникать как в результате суммирования независимых вкладов аллелей отдельных генов, так и вследствие эпистатического взаимодействия между ними. Оба эти случая представляют несомненный интерес для исследования природы полигенных заболеваний. Рассмотрены способы, позволяющие дискриминировать эти две возможности. Представлены описание и сравнение биоинформатических методов поиска сочетаний аллелей различных генов, ассоциированных с фенотипическими признаками полигенного заболевания, а также способы отображения и валидации полученных результатов. Предпринята попытка оценить применимость существующих методов к анализу эпистаза. Описаны и обобщены результаты, полученные авторами с применением программного обеспечения APSampler.**

КЛЮЧЕВЫЕ СЛОВА **медицинская геномика, фармакогеномика, полигенные заболевания, полигенный анализ, эпистаз.**

СПИСОК СОКРАЩЕНИЙ **ПО – программное обеспечение; РС – рассеянный склероз; ИИ – ишемический инсульт; CI – доверительный интервал (Confidence Interval); RR – относительный риск (Relative Risk); OR – отношение шансов (Odds Ratio); ORR – отношение наблюдаемого отношения шансов к ожидаемому (Odds Ratios Ratio); SF – фактор синергии (Synergy Factor); FDR – частота ложных открытий (False Discovery Rate); GWAS – полногеномное исследование ассоциаций (Genome-Wide Association Study); CDRV – распространенное заболевание, редкий аллель (Common Disease, Rare Variant); CDCV – распространенное заболевание, распространенный аллель (Common Disease, Common Variant); CMC – сочетание анализа многих переменных и свертки (Combined Multivariate and Collapsing); MCMC – метод Монте-Карло Марковскими цепями (Markov Chain Monte Carlo); MDR – метод снижения размерности (Multifactor Dimensionality Reduction); TDT – тест неравновесной передачи аллелей (Transmission Disequilibrium Test).**

ВВЕДЕНИЕ

Согласно представлениям классической генетики, наследственные болезни делятся на менделирующие и комплексные (многофакторные). Первые определяются носительством мутантного варианта единственного гена, тогда как вторые зависят как от генетической компоненты, определяющейся совместным вкладом множества независимо действующих или взаимодействующих полиморфных генов, так и от других факторов. При этом вклад каждого из генов в отдельности в развитие полигенного заболевания может быть небольшим. Носительство тех или иных аллельных сочетаний генов может определять также возникновение клинически гетерогенных форм заболевания и терапевтическую эффективность тех или иных лекарственных средств.

Полигенные заболевания человека намного более распространены, чем моногенные, они имеют огром-

ное социоэкономическое значение. Однако выяснение их молекулярно-генетической природы остается по сей день нерешенной задачей. Поиск генов, вовлеченных в развитие полигенных заболеваний, проводят, используя две основные стратегии: выяснение роли того или иного гена-кандидата, выбранного исходя из возможной роли его белкового продукта в этиопатогенезе заболевания, и полное геномное сканирование с использованием панели генетических маркеров, более или менее равномерно распределенных по геному. При этом экспериментальные подходы к установлению роли отдельных генов или значения отдельных областей генома состоят в анализе их сцепления или ассоциации с заболеванием.

Анализ сцепления проводят в семьях с несколькими больными; при выявлении у них общих аллельных вариантов роль гена в формировании предрасположенности к заболеванию можно считать доказанной. Однако недостатком этого метода является его низкая чувствительность; поэтому в последнее время предпочтение отдается обладающим большей статистической мощностью методам, основанным на анализе ассоциации.

Ассоциативное исследование – это попытка найти новые или проверить известные статистические взаимосвязи различных событий. Настоящие причины, порождающие такие взаимосвязи, часто находятся за пределами понимания или экспериментальных возможностей исследователя, однако, собрав статистику встречаемости сочетаний разных исходов наблюдений, можно сделать вывод о значимости (оценивается по вероятности получить наблюдаемый результат случайно) и интенсивности этих взаимосвязей. Ассоциацию того или иного полиморфного участка генома с фенотипическим признаком анализируют путем сравнения распределений его аллелей и генотипов в репрезентативных выборках индивидов, сформированных по наличию/отсутствию этого признака, которые должны соответствовать друг другу по распределению полов, возрасту и этнической принадлежности. Анализируемые аллельные варианты могут быть локализованы в любом участке ДНК, включая кодирующие последовательности (экзоны), интроны и промоторные области генов, где очень часто находятся участки регуляции транскрипции, а также другие области ДНК. При анализе экзонов представляют интерес не только несинонимические замены, которые определяют изменения в аминокислотной последовательности кодируемой белковой молекулы, но и замены синонимические, поскольку они могут влиять на структуру и стабильность мРНК и на кинетику ее трансляции за счет использования разных изоакцепторных тРНК. Однако не следует забывать, что, помимо прямой связи между исследованным локусом и наследуемым признаком, в основе ассоциации может лежать неравновесие по сцеплению между маркерным локусом и истинным локусом заболевания, если эти локусы расположены достаточно близко друг от друга.

Ассоциативные исследования призваны связать значимые для медицины фенотипические признаки с такими характеристиками, как аллельные вариации генома, эпигенетические модификации, воздействие окружающей среды, образа жизни и т.д. В качестве фенотипических признаков, представляющих ценность для персонализированной медицины, как правило, рассматривают возникновение заболевания, характер его течения (клиническая картина, степень поражения систем организма и пр.) или эффективность лечения тем или иным препаратом (область интересов науки фармакогеномики). В этом обзоре мы предполагаем сфокусировать внимание на ассоциации индивидуальных признаков с носительством аллельных вариантов генома. Выявление таких ассоциаций позволяет оценивать риск развития заболевания (предрасположенность к нему), предсказывать характер его течения и отдавать предпочтение тем или иным методам профилактики, диагностики и лечения исходя из особенностей генома индивида.

Анализ ассоциаций полигенных заболеваний с сочетанной встречаемостью аллелей различных генов остается относительно неразработанным направлением исследований. Во многом это связано с тем, что увеличение числа рассматриваемых генов ведет к экспоненциальному росту числа сочетаний их аллельных вариантов, что делает анализ стандартными методами перебора практически невозможным.

Настоящий обзор посвящен биоинформатическим методам поиска сочетаний аллелей различных генов, ассоциированных с фенотипическими признаками полигенного заболевания, а также способам отображения и валидации полученных результатов. Эти методы, которые мы будем далее для краткости называть методами полигенного анализа, направлены на выявление кумулятивного эффекта генов и его природы. Ассоциация с сочетанием может быть обусловлена взаимозависимостью влияний входящих в сочетание аллелей на фенотип, т.е. нелинейным (эпистатическим) взаимодействием между генами. Альтернативно, аллельное сочетание, значимо влияющее на развитие признака, может возникать в результате суммирования малых независимых подпороговых вкладов аллелей, входящих в сочетание. Оба эти случая будут рассмотрены в обзоре.

АССОЦИАТИВНЫЕ ИССЛЕДОВАНИЯ

Два основных вида ассоциативных исследований, а именно когортные и выполненные методом

«случай–контроль», различаются между собой по временно́й последовательности сбора информации и, как следствие, по параметрам, которые можно оценить исходя из наблюдений. В когортных исследованиях отобранную группу индивидов делят на две подгруппы – обладающих и не обладающих определенным индикаторным признаком (например, подгруппы носителей и неносителей определенного генотипа, подгруппы курящих и некурящих). Эти подгруппы наблюдают в течение некоторого временно́го интервала на предмет развития признака, представляющего интерес с точки зрения его предсказания (целевого признака), например, заболевания. Такой подход дает возможность численно выразить интенсивность вклада индикаторного признака в развитие целевого признака через отношение вероятностей заболевания у носителей и неносителей индикаторного признака. Оценкой этой величины является относительный риск (Relative Risk, RR).

Более распространенный вид ассоциативных исследований – исследование методом «случай–контроль». Выборку при этом делят на две группы – обладающих и не обладающих на момент исследования целевым признаком, например, больные и здоровые. В каждой из групп учитывается наличие индикаторных признаков, которые могли бы повлиять на возникновение болезни. При этом за кадром остаются люди, которые умерли до начала исследования, и чем выше уровень летальности заболевания, тем менее точна оценка уровня ассоциации по величине RR. В качестве критерия степени различия между носителями и неносителями индикаторного признака в исследованиях «случай–контроль» принято использовать величину отношения шансов (Odds Ratio, OR) [1]. Если абсолютный риск заболевания у неносителей мал, то величины OR и RR близки; чем он больше, тем сильнее OR превосходит RR. OR всегда больше, чем RR.

Результаты, полученные методом «случай–контроль», могут искажаться вследствие этнической гетерогенности сравниваемых групп, а также под влиянием неучтенных факторов окружающей среды [2]. Методы анализа на семейном материале (например, сравнение больных и здоровых братьев и сестер [3]) менее подвержены этим искажениям, но их требования ко входным данным (нужны пары больных и здоровых близких родственников, лучше всего сибсов) делают их малоприменимыми для получения достоверных зависимостей. Несколько менее жесткие требования к входной выборке предъявляет метод оценки неравновесной передачи аллелей (Transmission Disequilibrium Test, TDT) [4], в основе которого лежит анализ передачи маркерного аллеля от гетерозиготных здоровых родителей больному ребенку. Полученные данные сравнивают с ожидаемыми при менделевском наследовании и в случае неравновесного переноса аллеля делают вывод о связи аллеля с заболеванием. Еще один метод анализа ассоциации на семейном материале – метод AFBAC (Affected Family-Based Control), в котором группа сравнения составлена из совокупности тех аллелей здоровых родителей, которые не переданы больным потомкам (по одному аллелю от каждого из родителей) [5].

При анализе ассоциации часто и предсказываемые (зависимые), и предсказывающие (независимые) признаки – это категории, разделяющие выборку на два класса (например, «больной» и «здоровый» или «носитель» и «неноситель»). Такие данные удобно представлять в виде таблицы 2 × 2 (таблица сопряженности, contigency table), значения из которой используются для подсчетов величин, характеризующих силу ассоциации (OR) и ее значимость (величина p) по точному критерию Фишера, предложенному в 1922 г. и сохранившему свое значение до наших дней [6].

Если признак представляют более чем двумя классами, которые можно упорядочить (например, используя принятые медицинским сообществом шкалы для описания степени или тяжести заболевания), составляют $2n$-польные (здесь n – число градаций признака) таблицы, и для оценки силы и уровня значимости ассоциации используют критерий гамма Гудмана–Крускала [7]. Если упорядочивание не имеет смысла, то может использоваться или тест Фримена–Гальтона, дополняющий тест Фишера для более чем двух категорий [8], или критерий хи-квадрат [9].

МЕТОДЫ ПОЛИГЕННОГО АНАЛИЗА

Все подходы к многофакторному анализу и к полигенным исследованиям ассоциаций, как к его частному случаю, можно разделить на два принципиально различающихся типа: 1) использование сокращенного количества входных переменных на основе каких-либо априорных знаний и 2) полный анализ всех доступных переменных. Сокращение количества возможных переменных в полигенных исследованиях – это априорный отбор нескольких генов-кандидатов, применительно к которым проводят анализ ассоциации [10]. Этот подход позволяет существенно уменьшить расходы на генотипирование и уменьшить пространство анализа, тем самым снижая его сложность и сокращая время, необходимое на вычисления. С другой стороны, если эффект гена проявляется только во взаимодействии с другими генами и вследствие этого не наблюдается при отдельном рассмотрении (другими словами, отсутствует маргинальный эффект [11, 12]), такой ген вряд ли попадет в список

генов-кандидатов, хотя в действительности его роль может быть значимой. Сейчас, благодаря постоянному развитию как вычислительных мощностей, так и технологий генотипирования, популярность набирают методы полногеномного поиска ассоциаций по всему геному (Genome-Wide Association Study, сокращенно GWAS) [13–16], представляющие собой второй тип полигенного анализа, т.е. анализ всех доступных переменных.

При анализе полногеномных данных неизбежно наблюдаются крайне редкие аллели. Рассмотрение таких аллелей по отдельности не позволяет сделать заключение о влиянии каждого из них на заболевание, но если рассматривать общий эффект нескольких аллелей, то наблюдений может оказаться достаточно для подтверждения предположения об их общем влиянии. Иными словами, информация о каждом из редких аллелей недостаточна, но ею не следует пренебрегать: при аккумулировании информации о нескольких редких аллелях ассоциация может оказаться достоверной. Такой эффект называется аддитивным, он может наблюдаться не только на редких аллелях, однако в случае редких аллелей это основной способ ассоциативного исследования. Сейчас становится общепринятой теория, объясняющая возникновение многих распространенных болезней носительством редких аллелей, получившая название CDRV (Common Disease, Rare Variant) [17, 18] («распространенное заболевание, редкий аллель»), которая представляет собой альтернативу теории CDCV (Common Disease, Common Variant, «распространенное заболевание, распространенный аллель»). Активно развиваются методы, специально предназначенные для учета аддитивного вклада редких аллелей, такие, как метод свертки (CMC, Combined Multivariate and Collapsing) [19], статистика взвешенных сумм [20], тест нагрузки на ген (burden test) [21].

При полигенном анализе особенно актуальной становится проблема учета множественности гипотез. Кратко ее можно сформулировать так: при увеличении числа тестируемых гипотез растет вероятность случайно получить любой, в том числе и маловероятный, результат, что уменьшает значимость утверждения, согласно которому наблюдаемые статистические взаимосвязи действительно отражают какие-то закономерности, а не являются случайными.

При малом, но не равном одному, числе сравнений, используемых при исследовании ассоциации признака с несколькими аллелями одного высокополиморфного гена или при одновременной оценке роли нескольких биаллельных генов-кандидатов, такое уменьшение значимости учитывается поправкой Бонферрони [22], просто умножающей соответствующие величины p на число проведенных тестов, однако поправка Бонферрони оказывается слишком консервативной из-за лежащего в ее основе допущения о независимости тестов. Более точная поправка может быть получена методом Вестфолла–Янга [23], который не вводит этого допущения, а сравнивает лучшее из наблюдений с лучшими же наблюдениями в перемешанных выборках. Другой подход к этой проблеме заключается в том, чтобы оценивать не уровень значимости того, что среди результатов после их отбора по порогу достоверности нет ни одного ложного (случайного) (Family-Wise Error Rate, FWER), а частоту ложных открытий (False Discovery. Rate, FDR) [24, 25].

В последнее время широко обсуждается возможность взаимодействия генов, или эпистаза. В значительной степени интерес к этой теме связан с плохой воспроизводимостью результатов по оценке роли отдельных генов в формировании предрасположенности к комплексным заболеваниям, особенно при полногеномных исследованиях. Существует некоторая неоднозначность понятий «эпистаз» и «эпистатическое взаимодействие»: исходно под ними подразумевали полное маскирование эффекта полиморфизма одного локуса полиморфизмом другого локуса, а позднее – и любые другие типы влияния одних полиморфизмов на проявление других в фенотипе. Различия в интерпретациях термина «эпистаз», а также проблемы, связанные с такими разночтениями, хорошо описаны в [26, 27].

На *рис. 1* в форме визуализированных четырехпольных таблиц сопряженности представлены результаты анализа вклада носительства аллеля *04 гена *DRB1* HLA класса II (аллель *DRB1**04) (*А*), аллеля с делецией 32 нуклеотидов гена хемокинового рецептора *CCR5* (*CCR5**d32) (*Б*) и их сочетания (*В*) в развитие одного из типичных полигенных заболеваний – рассеянного склероза (РС), основанные на экспериментальных данных из статьи [28]. Во всех случаях больных РС и здоровых разделяли на два класса по принципу носительства/неносительства аллеля (не различая гомо- и гетерозиготы по этому аллелю). При этом полиморфизм гена *CCR5* действительно является биаллельным (аллель с делецией и аллель дикого типа), тогда как в случае гена *DRB1* анализировали 18 групп аллелей этого высокополиморфного гена, и в группу неносителей *DRB1**04 попали носители всех остальных аллелей этого гена. Как видно *на рис. 1В*, носительство сочетания *DRB1**04 и *CCR5**d32 ассоциировано с заболеванием гораздо сильнее, чем ожидается исходя из аддитивного вклада составляющих его аллелей, что можно интерпретировать как следствие эпистатического взаимодействия между рассматриваемыми генами. Этот пример демонстрирует простейший тип поли-

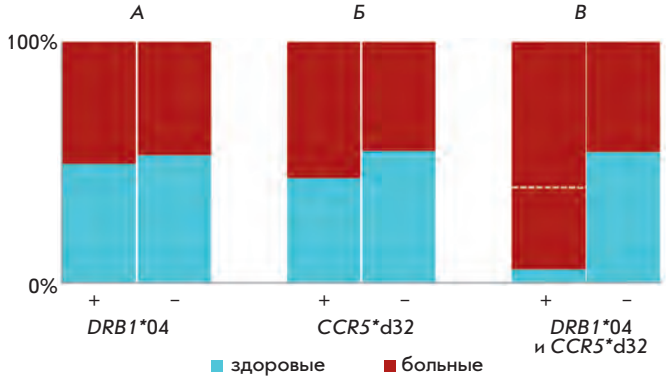

Рис. 1. Визуализация четырехпольных таблиц носительства больными РС и индивидами контрольной группы аллелей генов главного комплекса гистосовместимости *HLA-DRB1* (*А*), хемокинового рецептора *CCR5* (*Б*) и их сочетаний (*В*) (на основании данных [28] для этнических русских). Бордовым показаны поля для больных, голубым – для контрольной группы, соотношение полей по вертикали отражает распределение среди них носителей (+) и неносителей (–) аллелей *DRB1**04 (*А*), *CCR5**d32 (*Б*) и сочетания *DRB1**04 и *CCR5**d32 (*В*). Горизонтальная пунктирная линия (*В*) соответствует ожидаемому отношению числа больных и контролей среди носителей сочетания, вычисленному в предположении о независимом влиянии аллелей.

генного анализа, когда ограничиваются рассмотрением совместного вклада двух аллелей в формирование фенотипа.

В работе [29] мы предложили использовать в качестве численной меры величины эпистаза отношение наблюдаемого для сочетания аллелей OR к ожидаемому OR (далее мы будем называть ее ORR, Odds Ratios Ratio). Она основана на представлении, что если два или более аллелей в составе сочетания не взаимодействуют, то величина OR при носительстве этого сочетания будет примерно равна произведению OR отдельных аллелей, входящих в сочетание. Это произведение мы рассматривали как ожидаемое OR и сравнивали его с наблюдаемым OR. Чем больше их отношение отличается от 1, тем более сильного эпистатического взаимодействия между генами можно ожидать.

Величина ORR [29] применима для анализа взаимодействия как двух, так и многих аллелей, но существенным ее недостатком является отсутствие метода оценки доверительного интервала (Confidence Interval, CI). Мера эпистаза SF, описанная в работе [11], обладает «зеркальными» достоинствами и недостатками: для нее приведен способ расчета CI, но она применима для анализа взаимодействия только двух аллелей (или других бинарных индикаторных признаков). Обе величины представляют собой отношения наблюдаемого для сочетания аллелей OR к произведению OR, наблюдаемых порознь для его составляющих, однако значения OR вычисляются по-разному. Если ORR сравнивает у больных и контролей число носителей и неносителей индикаторного признака (будь то сочетание аллелей или отдельный аллель), как это показано на *рис. 1*, то в случае SF носители пары, а также носители каждого из аллелей, входящих в сочетание, сопоставляются с неносителями ни одного из аллелей, при этом все четыре группы не пересекаются. Как и в случае ORR, SF > 1 говорит о положительном (взаимоусиливающем) взаимодействии, а SF < 1 – об отрицательном (компенсаторном). В принципе, величину SF можно рассчитать более чем для двух аллелей, однако результат зависит от порядка их объединения в сложные признаки. Таким образом, целесообразно использовать обе эти оценки.

Существующие средства анализа кумулятивного эффекта нескольких генетических переменных используют различные алгоритмы интеллектуального анализа данных (data mining).

Среди этих средств наиболее популярна классическая логистическая регрессия, в которой взаимодействию соответствуют коэффициенты при членах модели порядка 2 и выше [30]. Для поиска таким методом наиболее сильно взаимодействующей комбинации аллелей придется проводить моделирование многократно, из-за чего статистическая мощность метода теряется. Двухступенчатый вариант, реализованный в GenABEL [31, 32], предоставляет решение проблемы множественного тестирования, используя информацию о дисперсии в индивидуальных локусах для отбора тех из них, в которых вероятность взаимодействия больше. Используются также различные эвристические методы: генетическое программирование [33], нейронные сети [34], поиск шаблонов (pattern mining) [35], методы, основанные на уменьшении размерности [36], и методы Монте-Карло Марковскими цепями (Markov Chain Monte Carlo, MCMC), к которым относятся APSampler [37, 38], BEAM [39, 40] и логическая регрессия LogicReg [41–43].

Ассоциацию носительства любого определенного сочетания аллелей (или другого индикаторного признака) с фенотипом можно оценить так же, как это делается в случае одного аллеля (признака). Иными словами, каждое сочетание может рассматриваться как единый (составной) признак и характеризоваться уровнем значимости ассоциации и величинами RR или OR. Число возможных сочетаний очень велико, и на первый план выходит задача поиска тех из них, для которых ассоциация наиболее значима.

Проведение многофакторного анализа ассоциации возможно также на семейных данных. Существуют мультиаллельные и мультилокусные версии TDT [4], который основан на статистике МакНемара и изначально был разработан для одиночных биаллельных локусов. Способы расширения TDT для нескольких аллельных вариантов предложены рядом авторов. Они включают вычисление маргинальной равномерности [44]; поочередную группировку аллелей в две группы: «исследуемый аллель» и «остальные аллели», с последующим проведением теста МакНемара и коррекцией на множественное тестирование [45]; и, наконец, вычисление неравномерного переноса аллелей с использованием логистической регрессии [46], наиболее подходящее для высокополиморфных локусов. При проведении анализа одновременно на нескольких локусах применяют методы, в которых генотип ребенка сопоставляется со всеми возможными для его родителей генотипами потомства [45, 47, 48]. Неравновесие по сцеплению между анализируемыми локусами или рассчитывается из анализируемой выборки, или берется из известных данных, например, из HAPMAP [49], как в программном обеспечении (ПО) FAMHAP [48, 50].

Далее мы подробнее рассмотрим и сравним некоторые широко распространенные инструменты для полигенного анализа ассоциаций.

PLINK

В Гарвардском университете создано и бесплатно распространяется ПО PLINK [30, 51], которое представляет собой большую взаимосвязанную коллекцию различных алгоритмов анализа генетических и фенотипических данных, в том числе и методов полигенного анализа. PLINK используется во многих исследованиях генетического взаимодействия (например, [52–55]).

Один из методов анализа взаимодействия генов в PLINK основан на рассмотрении регрессионных моделей [56]. При бивариантном исходе (например, «больной–здоровый») используют модель логистической регрессии, которая предполагает, что вероятность события (в данном случае заболевания) описывается как логистическая функция от линейной комбинации независимых переменных (предикторов) [57]. Для количественных фенотипов (таких, например, как три степени артериальной гипертонии) используют обычную линейную регрессию от тех же предикторов. Независимыми переменными при этом служат индикаторные функции, которые принимают числовые значения 1 или 0 в зависимости от того, представлен или нет в геноме определенный аллель или генотип (или от наличия другого индикаторного признака). Результатом анализа являются набор регрессионных коэффициентов при индикаторных функциях аллелей и их сочетаний и уровни статистической значимости отличий этих коэффициентов от нуля. Высокая достоверность отличия от нуля коэффициента, соответствующего определенному сочетанию аллелей, свидетельствует о взаимодействии последних. Так работает, например, тест «PLINK – epistasis».

Более простой тест на взаимодействие – «PLINK – case-only», проверяет корреляцию между носительством больными нескольких генотипов. Если корреляция пары генотипов высока, а их сцепление можно исключить из рассмотрения, то делается вывод о взаимодействии. Этот тест основан на априорном предположении, что выявленная корреляция характерна только для больных. Двухступенчатая процедура, проверяющая наличие корреляции в общей выборке, свободна от этого предположения, однако и она может давать смещенные результаты [58].

Существенные достоинства ПО PLINK – его применимость для GWAS и большой набор доступных средств анализа, а недостаток – ограничения по формату данных: программа работает только с биаллельными маркерами.

MDR

Для поиска полигенных ассоциаций методом «случай–контроль» сейчас широко применяется алгоритм снижения размерности MDR (Multifactor Dimensionality Reduction) [59–62].

На первом этапе все данные случайным образом делят на две выборки: обучающую (например, 9/10 данных) и тестовую (например, 1/10 данных). Далее, для каждой комбинации аллелей и генотипов, присутствующей в обучающей выборке, рассчитывается параметр, характеризующий соотношение количества больных и здоровых, несущих эту комбинацию, и в зависимости от величины этого параметра комбинации классифицируются на категории, например, высокого и низкого риска. Таким образом осуществляется переход от n-мерного пространства всех единичных полиморфных участков и фенотипа к двумерному пространству, где одно измерение – это уровень риска, а второе – носительство данной комбинации аллелей. Среди всех комбинаций будет существовать такая, которая имеет наименьшую ошибку классификации в обучающей (Training Accuracy) и тестовой (Testing Accuracy) выборках. При этом разбиение на группы производят 10 раз, изменяя каждый раз параметры генератора случайных чисел. Параметр согласованности модели (Cross Validation Consistency) показывает, сколько раз из этих 10 она была найдена как лучшая. Модель считается валидированной, если ее согласованность не меньше 9/10.

Пакет ПО MDR включает в себя, кроме текстового отображения результатов, также графики-дендрограммы, отображающие попарный анализ взаимодействия, где цветом показан тип взаимодействия локусов – от эпистаза до независимости, а длиной связи – его сила.

МЕТОДЫ, ОСНОВАННЫЕ НА MCMC

Полный перебор всех комбинаций (например, применяемый в MDR) становится неэффективным при росте числа анализируемых аллелей из-за огромного количества возможных сочетаний: происходит так называемый комбинаторный взрыв. Кроме того, при таком переборе статистическая значимость найденных сочетаний становится неочевидной из-за проблемы множественного тестирования. С другой стороны, простые градиентные («жадные», greedy) методы, пошагово улучшающие промежуточный результат, часто могут не привести вообще ни к какому разумному итогу, поскольку находят лучшие варианты локально, а не глобально.

Существуют различные эвристические методы, позволяющие искать глобальный оптимум, не прибегая к полному перебору. Один их них – это метод Монте-Карло Марковскими цепями (Markov Chain Monte Carlo, MCMC), использованный разными авторами [38, 40, 63–65]. Главная идея метода состоит в том, что он, как и градиентный поиск, стремится перейти к решению, лучшему, чем имеющееся на данный момент, но, в отличие от градиентного поиска, с некоторой вероятностью может перейти и к худшему, причем вероятность эта уменьшается со степенью ухудшения решения.

BEAM

Алгоритм BEAM (Bayesian Epistatis Association Mapping) [40, 66] при поиске ассоциаций опирается на то, что у ассоциированных с заболеванием локусов распределение генотипов у больных будет обязательно отличаться от распределения генотипов в контрольной группе. Цель алгоритма – классификация всех локусов на не ассоциированные с заболеванием, ассоциированные с ним поодиночке, ассоциированные и при этом эпистатически взаимодействующие. Программа находит наиболее вероятное при данных генотипах и уровнях заболевания разбиение локусов на эти три класса с помощью метода MCMC. Локусы считаются эпистатически взаимодействующими, если совместное распределение их аллелей/генотипов лучше соответствует данным, чем распределение, следующее из независимой модели (произведение распределений аллелей/генотипов). BEAM может учитывать информацию о гаплотипах, чтобы отличать их от эпистатической взаимозависимости.

Логическая регрессия

Существует также метод, использующий MCMC для оптимизации моделей регрессионного поиска полигенных ассоциаций, – это MCMC-версия алгоритма логической регрессии (Logic Regression) [43, 65]. Название метода напоминает более известную логистическую регрессию, которая решает сходную задачу другим способом. В качестве предикторов логической регрессии используются индикаторные функции логических комбинаций (логические функции) присутствия различных аллелей, при этом набор оптимальных функций определяется с помощью MCMC. Найденные логические функции явно показывают тип взаимодействия аллелей.

APSampler

Логика анализа полигенных данных программой APSampler [38] значительно отличается от представленных ранее программных пакетов, в которых предсказываемый фенотипический признак может принимать только два значения, например «больной» и «здоровый». Использование в этом ПО непараметрического критерия Вилкоксона создает возможность проводить не только категориальное сравнение, но и анализировать данные с более чем двумя значениями целевого признака, в случае если эти значения можно ранжировать. Например, в случае инсульта для формирования сравниваемых групп можно использовать ряд международных шкал, по которым оценивают степень угнетения сознания, исходную тяжесть заболевания, клиническое состояние больного в остром периоде, исход инсульта (т.е. степень восстановления утраченных функций за определенный период от начала заболевания) и др., причем каждая из шкал имеет свой диапазон балльных оценок, но не менее трех. Основной объект, с которым работает программа APSampler в поисках предсказывающего индикаторного признака – это генетический паттерн, т.е. сочетание аллелей или генотипов различных локусов, ассоциированное с фенотипическим признаком. Поиск паттернов осуществляется с помощью MCMC, при этом на каждом шаге рассматривается сразу несколько паттернов, и их набор оптимизируется от шага к шагу с точки зрения вероятности того, что все паттерны из набора независимо друг от друга и одновременно ассоциированы с признаком. Для оценки вероятности ассоциации каждого из паттернов применяется непараметрический критерий Вилкоксона, при этом сравниваемые подвыборки устроены так, что они отличаются носительством только одного паттерна из набора. Результатом работы первого этапа программы является список паттернов, которые встретились при работе ПО. Затем происходит валидация этих результатов.

Краткое сравнение возможностей различных ПО для полигенного анализа ассоциаций

	APSampler [38]	BEAM [40]	LogicReg [43]	MDR [60]	PLINK [30]
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]
Оценка ассоциации для найденных сочетаний точным критерием Фишера	+	+	-	-	-
Процедура валидации	+	+	+	+	-
Полиаллельные локусы	+	-[7]	-	+	-
Поиск эпистаза	+[8]	+	+	+	+
Графическое отображение эпистаза	-[9]	-	-	+	-
Возможность проведения анализа ассоциации для комбинации аллелей, указанной пользователем	+	-	-	+	-[10]
Полногеномный анализ	-	+	-	-[11]	+
Возможность запуска из командной строки (например, на сервере)	+	+	+	+	+
Работа в среде UNIX	+	+	+	+	+
Работа в среде Windows	+	+	+	+	+
Параллельные вычисления	+	-[12]	-	-[11]	-

[1] Существует версия BEAM, интегрированная в серверное приложение GALAXY [83].
[2] Алгоритм реализован в пакете для статистических вычислений и графики R [84].
[3] ПО автоматически разделяет данные на две категории, используя для этого среднее значение.
[4] Авторами предлагается специальное ПО – MDR Data Tool [85] 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] Создана программа на языке Perl для графического отображения эпистаза [37].
[10] Предлагается анализ ассоциации гаплотипа.
[11] Для этой цели предусмотрено специальное ПО [86].
[12] Отдельное ПО PBEAM для параллельных исчислений [87].

Для каждого паттерна из списка вычисляется значимость ассоциации по Фишеру в случае двоичного исхода или по Крускалу–Гудману [7] в случае более чем двух категорий. Затем программа несколько раз перемешивает метку фенотипического признака и еще раз запускает поиск ассоциированных паттернов. Достоверности ассоциации по результатам запусков с перемешанным фенотипом дают распределение достоверностей находок при условии нулевой гипотезы, утверждающей отсутствие действительных ассоциаций в начальных данных. Это нулевое распределение используется для валидации сочетаний, найденных на первом этапе.

В *таблице* сведена воедино информация о функциональных возможностях описанных выше программ для полигенного анализа ассоциаций. Данные таблицы свидетельствуют о том, что предлагаемые программы для полигенного анализа существенно различны по своим функциям. Например, MDR очень удобен благодаря наличию пользовательского

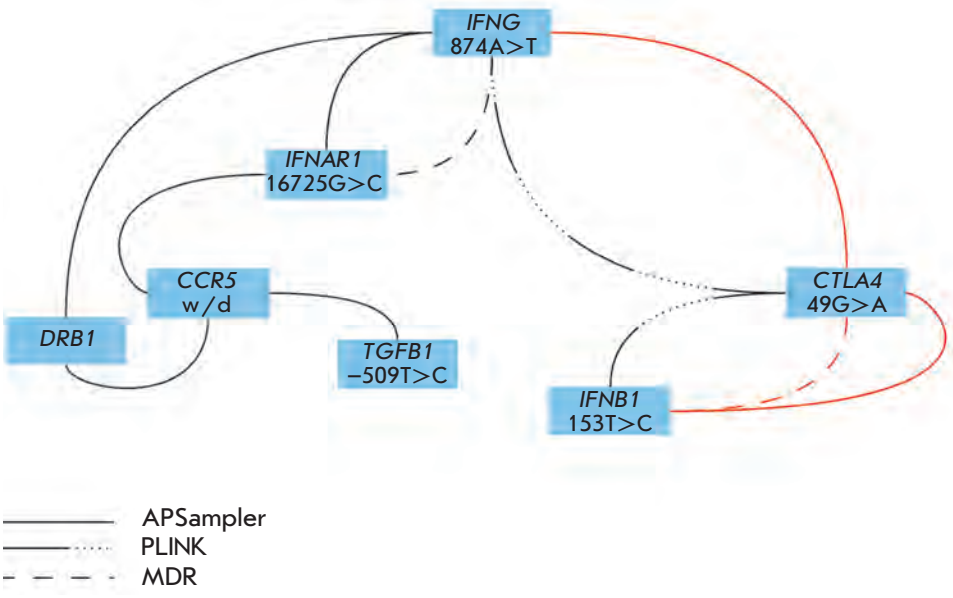

Рис. 2. Поиск программами APSampler, MDR и PLINK биаллельных сочетаний генов иммунного ответа, ассоциированных с эффективностью лечения РС препаратом глатирамера ацетатом (на основании данных [29] для этнических русских). Программа APSampler [38] находит все биаллельные маркеры, определенные другими программами, а также идентифицирует другие сочетания. Красным отмечены сочетания, ассоциация которых с эффективностью лечения прошла валидацию пермутациями в программе APSampler ($p < 0.1$) или кроссвалидацию программой MDR (CVC > 8/10).

интерфейса и графическому отображению результатов, в том числе эпистаза. Следовательно, сравниваемые ПО применимы в разных случаях в зависимости от имеющихся генетических и фенотипических данных, содержания и формата желаемого результата, а также возможности пользователя оперировать ПО на уровне командной строки. Нужно также учитывать, что сам искомый результат сильно различается для разных программ. Например, MDR выдает найденные ассоциированные с фенотипом локусы и их комбинации, тогда как APSampler учитывает направление ассоциации, определяемое носительством аллелей локусов и их комбинации. И APSampler, и MDR работают с поливариантными входными признаками, в то время как остальные – только с бивариантными. Эти две программы схожи также в том, что позволяют провести анализ эпистатического взаимодействия уже после выявления ассоциации, тогда как ПО BEAM заранее разделяет все аллели на три группы: с краевым эффектом, эпистазом и отсутствием эффекта. Характеристики комбинаций локусов, которые приводит MDR, статистически обоснованы, но неочевидно соотносятся с силой ассоциации. LogicReg не приводит классически интерпретируемых величин ассоциации вообще. APSampler и BEAM решают эту проблему, приводя точный тест Фишера на ассоциацию найденных индикаторных признаков с фенотипом. В целом, BEAM, PLINK, MDR и LogicReg хорошо применимы к фундаментальным исследованиям, в том числе к изучению взаимодействия генов, либо для работы в составе большей интегрированной программной среды, но по умолчанию не обладают нужным набором функций для таких прикладных медико-генетических задач, как поиск маркеров предрасположенности или поиск фармакогенетических маркеров, для которых приспособлено ПО APSampler.

Мы применили эти пять программ в пользовательском режиме (т.е. со всеми настройками по умолчанию) к данным из [29]. BEAM не нашел ни одного сочетания с $p < 0.05$, а выдача LogicReg требовала дополнительной обработки. Результаты применения APSampler, MDR и PLINK представлены на *рис. 2*, из которого видно, что ПО APSampler находит как те сочетания, которые нашел MDR, так и те, которые нашел PLINK, при этом все находки APSampler'а, прошедшие валидацию, подтверждены хотя бы одной из этих программ.

ИССЛЕДОВАНИЯ, ПРОВЕДЕННЫЕ С ПОМОЩЬЮ APSampler

С момента первой публикации [38] было проведено довольно большое количество исследований с использованием ПО APSampler, причем в большинстве из них участвовали авторы этой программы по причине относительной сложности ее применения на начальных этапах разработки. Это позволило по мере использования алгоритма и с учетом пожеланий пользователей совершенствовать ПО, поэтапно добавляя к нему новые части, расширяющие возможности валидации [67], управления данными, представления результатов и получения справки

по использованию и устройству APSampler. В настоящее время созданы условия для свободного использования программы [37].

Используя ПО APSampler, авторы в рамках различных проектов анализировали кумулятивный эффект аллелей ряда генов-кандидатов с развитием рассеянного склероза (РС) [68], различных форм артериальной гипертонии [69–71], инфаркта миокарда [72], ишемического инсульта (ИИ) [73, 74] и геморрагического инсульта [75]. Исследования проводили, следуя принципу этнической гомогенности групп, у русских или у якутов. Популяция якутов представляет особый интерес с точки зрения этногеномики, поскольку в ней наблюдается эффект основателя, а также определенная географическая и культурная изолированность [76]. APSampler применяли также при фармакогенетических исследованиях РС, изучая связь генетического статуса пациентов с эффективностью лечения иммуномодулирующими препаратами – интерфероном бета ([67] у ирландцев) и глатирамера ацетатом ([29, 77] у русских).

В большинстве перечисленных работ сравнивали попарно группу неродственных больных с контрольной группой неродственных индивидов без изучаемого заболевания, сходной с выборкой больных по этнической принадлежности, соотношению полов и среднему возрасту. В некоторых случаях сравнивали две группы больных с клинически гетерогенными формами одного и того же заболевания (например, артериальная гипертензия, протекающая с гиперальдостеронизмом и без гиперальдостеронизма [69]). При исследовании генетической предрасположенности к артериальной гипертонии, предшествующей развитию ИИ, сначала разбивали больных на две подгруппы по уровню гипертонии, а затем с помощью таблицы сопряженности 2 × 4 искали среди выявленных алгоритмом APSampler аллельных сочетаний такие, носительство которых характеризуется изменением в ряду от нормотоников до гипертоников 3-й степени [71]. При фармакогенетических исследованиях сравнивали попарно больных, отвечающих и не отвечающих на лечение, используя также подход «сравнение крайних».

Гены-кандидаты выбирали исходя из представлений об участии их белковых продуктов в процессах, вовлеченных в патогенез заболевания. При анализе генетической предрасположенности к сердечно-сосудистым заболеваниям выбирали гены, белковые продукты которых участвуют в воспалении, гены систем гемостаза, транспорта и метаболизма липидов, гены ренин-ангиотензин-альдостероновой системы и некоторые другие. В случае РС продукты генов-кандидатов вовлечены в развитие иммунного ответа и хронического воспалительного процесса.

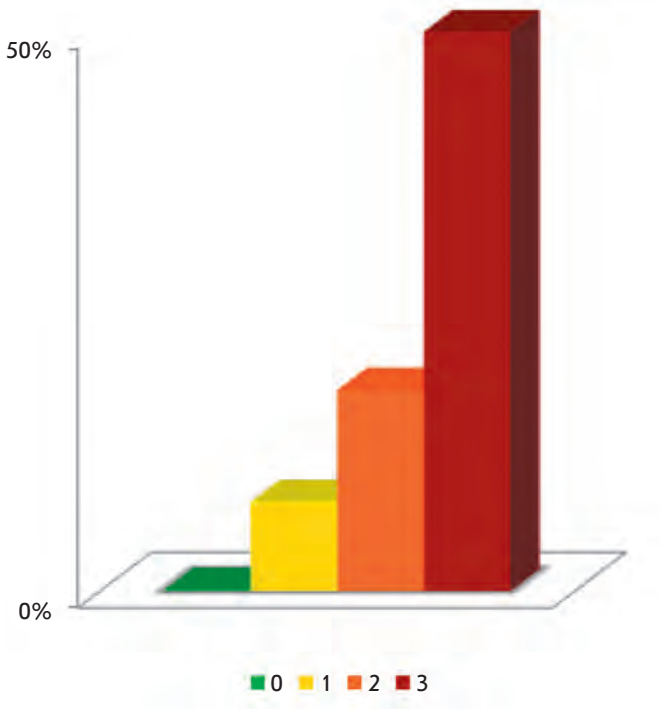

Рис. 3. Носительство выявленного с помощью ПО APSampler трехаллельного сочетания *FGB**–249C + *APOE**ε4 + *CMA**–1903A у якутских больных, перенесших ИИ и различающихся по уровню предшествующего артериального давления [71]. 0 – нормотоники, 1–3 – гипертоники 1-й, 2-й и 3-й степени соответственно, согласно критериям 2003 ESH/ESC [82]. Носительство представлено в процентах от общей численности каждой подгруппы.

Как правило, в этих генах типировали полиморфные участки (в основном однонуклеотидные полиморфизмы, или SNP), представляющие интерес с функциональной точки зрения, т.е. заведомо влияющие на количество или свойство кодируемого белкового продукта. Анализировали совместный вклад от одного десятка до нескольких десятков полиморфных маркеров в относительно небольших выборках, составляющих максимум 500 человек. Хотя этот типичный для российских исследований объем выборок не идет ни в какое сравнение с численностью групп, формируемых международными консорциумами, нам удалось выявить с помощью ПО APSampler высокозначимые ассоциации сочетаний аллелей/генотипов с изучаемым фенотипом. Это утверждение можно проиллюстрировать данными по ассоциации сочетания из аллелей трех генов (*FGB**–249C + *APOE**ε4 + *CMA**–1903A) с уровнем артериальной гипертензии, предшествующей развитию ИИ у якутов (*рис. 3*). В выборке всего из 115 больных наблюдали монотонное нарастание частоты носительства

названного триаллельного сочетания от 0% у нормотоников до 47% от общей численности подгруппы у гипертоников 3-й степени; величина p, оцененная по Фишеру в таблице сопряженности 2 x 4, оказалась равной 0.0003. В этом случае мы наблюдали яркий пример эффекта совместного вклада генов, кодирующих компоненты трех различных важнейших систем гомеостаза – гемостаза (*FGB*), метаболизма липидов (*APOE*) и ренин-ангиотензин-альдостероновой системы (*CMA*) в развитие полигенного заболевания – артериальной гипертензии, которая скорее всего возникает в результате суммирования независимых вкладов отдельных генов.

Причина такой высокой информативности при скромном объеме экспериментальных данных может быть связана с преимуществами, возникающими за счет этнической и клинической гомогенности используемых нами групп. Группы же из десятков тысяч пациентов из разных стран и лечебных учреждений, формируемые в рамках консорциумов, как правило, не отвечают условиям гомогенности ни по этнической принадлежности индивидов, ни по клинической картине, что может нивелировать их генетические отличия от контрольной группы. Однако основная причина высокой информативности результатов, полученных с помощью ПО APSampler, кроется, по-видимому, в высокой статистической мощности анализа. Не вдаваясь здесь в рассмотрение вопроса о том, что лежит в основе наблюдаемого феномена, можно констатировать, что выявление ассоциации аллелей/генотипов отдельных генов при анализе любого из исследуемых заболеваний было нечастым событием, тогда как обнаружить ассоциированные с фенотипом сочетания из двух–четырех аллелей нам удавалось практически во всех случаях. Здесь уместно оговориться, что наблюдаемая ассоциация могла быть как позитивной, так и негативной, причем выявлять разнонаправленный эффект альтернативных аллелей нам удавалось во многих, но не во всех случаях.

Ассоциация РС с аллелем гена *DRB1**15 главного комплекса гистосовместимости [78, 79], с микросателлитным маркером TNFa9 [80] и с биаллельным сочетанием *DRB1**04 и *CCR5**d32 [28] (см. *рис. 1*) у русских была показана ранее без применения APSampler'а и воспроизведена при анализе на независимой выборке с помощью алгоритма APSampler [68]. Репликация данных по ассоциации этих генетических факторов с развитием РС не только отвечает критериям, принятым мировым научным сообществом для признания полученных результатов, но и свидетельствует об эффективности использованного ПО.

Опираясь на описанные выше наблюдения, мы сформулировали понятие сочетания минимального множества (минимального сочетания) аллелей как фактора генетического риска, выявляемого в том или ином исследовании [68]. Под этим понимается, что любое подмножество этого множества характеризуется меньшей значимостью ассоциации. Так, нами были выявлены [68] два ассоциированных с РС триаллельных сочетания, включающие аллели полиморфных участков генов *DRB1*, *TGFB1*, *CTLA4* и *TNF*. Различия в частотах носительства входящих в состав «трио» биаллельных сочетаний и отдельных аллелей между больными и контрольной группой не достигали уровня значимости ($p < 0.01$). Важно отметить, что подгруппы индивидов, несущих предрасполагающие к РС сочетания 1 и 2, не перекрывались и составляли около 5 и 9% больных РС и не выявлялись в контрольной группе. Таким образом, как при классическом моногенном доминантном заболевании, все носители того или другого сочетания в нашей выборке оказались больными. Аналогичные результаты получены и в других наших работах. В любом случае, минимальное множество аллелей представляет собой составной генетический маркер полигенного заболевания или другого фенотипа.

Вопрос о типе взаимодействия между аллелями входящих в сочетание генов – эпистатическом или аддитивном – мы попытались решить в ходе фармакогенетического исследования, в котором анализировали ассоциацию между эффективностью лечения больных РС иммуномодулирующим препаратом глатирамера ацетатом и аллельным полиморфизмом ряда генов иммунного ответа [29]. Носительство сочетаний аллелей четырех генов (*DRB1**15+*TGFB1**–509T+*CCR5**d+*IFNAR1**16725G) увеличивало в 14 раз риск неэффективного лечения препаратом (OR = 0.072 [CI = 0.02–0.28]; p = 0.00018), причем ассоциация выдерживала пермутационный тест (p_{perm} = 0.0056), ко времени этого исследования включенный в программу. Трехаллельное сочетание (*DRB1**15+*CCR5**d+*TGFB1**–509T) как маркер неэффективности лечения мало отличалось от четырехаллельного, тогда как ассоциация всех остальных компонентов последнего с неэффективным лечением была существенно слабее. На *рис. 4* в графическом представлении (в виде диаграммы Венна) приведена оценка характера взаимодействия различных компонентов «неблагоприятного» аллельного сочетания (*DRB1**15+*TGFB1**–509T+*CCR5**d+*IFNAR1**16725G). 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*.

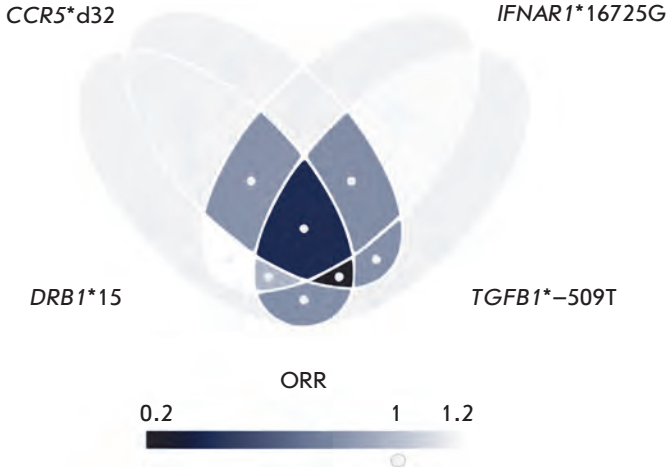

Рис. 4. Диаграмма Венна, характеризующая возможность взаимодействия компонентов сочетания *DRB1**15 + *TGFB1**–509T + *CCR5**d32 + *IFNAR1**16725G, негативная ассоциация которого с эффективностью лечения РС препаратом глатирамера ацетатом выявлена с помощью ПО APSampler [29]. Каждый из четырех эллипсов диаграммы соответствует одному из аллелей этого сочетания. Области пересечения эллипсов дают все возможные комбинации из четырех аллелей, при этом интенсивность цвета отражает отношение наблюдаемого OR к ожидаемому OR (ORR) в соответствии с градиентной шкалой, представленной ниже. Области, соответствующие отдельным аллелям, а также маленькие референтные кружки, соответствуют ORR, равному единице. Чем более цвет отличается от серого, соответствующего единице, тем интенсивнее эпистатическое взаимодействие, оцененное для этой области. Значения ожидаемого OR вычислены для каждого из сочетаний как произведение OR отдельных аллелей, соответствующих пересекающимся областям.

Неожиданные данные об эпистатических взаимодействиях при формировании генетической предрасположенности к ИИ у русских получены в работе [73]. Анализ с помощью алгоритма APSampler выявил биаллельные протективные сочетания (*IL6**–174C/C + *FGA**4266A) и (*IL6**–174C/C + *FGB**–249C), которые несколько более значимо, чем один входящий в каждый из них генотип *IL6**–174C/C, ассоциированы с ИИ и имеют практически такую же величину OR (0.32–0.35). Одновременно аллели, входящие в эти сочетания, *FGA**4266A или *FGB**–249C, при совместном носительстве с альтернативным генотипу *IL6**–174C/C аллелем G гена *IL6*, «нейтрализовали» его значение как аллеля риска, снижая как уровни значимости, так и величины OR (с 2.9 до 1.9–2.1). Иными словами, мы наблюдали ассоциацию с ИИ сочетаний аллелей/генотипов *IL6*, *FGA* и *FGB*, в которых гену *IL6* принадлежит ведущая роль, а генам *FGA* и *FGB* – модулирующая. Это наблюдение, возможно, отражает присутствие в генах *FGA* и *FGB* элементов, чувствительных к интерлейкину-6, связывающих STAT3 – основной транскрипционный фактор, передающий сигнал от рецептора интерлейкина-6 к ядру [81].

ЗАКЛЮЧЕНИЕ

Анализ, ставящий своей целью поиск полигенных сочетаний, ассоциированных с фенотипическим признаком, т.е. составных генетических маркеров, является адекватным инструментом для исследования полигенных заболеваний. Сейчас статистические методы, предоставляющие возможности такого анализа, переживают период быстрого роста.

В соответствии со всем сказанным выше, составные генетические маркеры могут возникать вследствие эпистатического взаимодействия между компонентами или же иметь аддитивную природу. Принимая во внимание сложность и разнонаправленность различных кумулятивных эффектов, можно утверждать, что обнаружение достоверного составного маркера, пусть несущего даже небольшое число компонентов, является важным шагом в понимании этиопатогенеза заболевания. Действительно, такой маркер может указывать на важный узел в сложной регуляторной сети взаимодействия биологических макромолекул. ●

Авторы выражают благодарность О.Г. Кулаковой и Е.Ю. Царёвой (РНИМУ им. Н.И. Пирогова, Москва), а также И. Русинскому (I. Ruczinski, Johns Hopkins University, Baltimore, MD) за полезные замечания и советы.

*Работа выполнена при поддержке РФФИ (проекты № 11-04-01644а и 11-04-02016а), научно-технической программы Правительства г. Москвы (№ 8/3-280н-10), гранта Johns Hopkins University Framework for the Future, гранта Commonwealth Foundation and the SKCCC Center for Personalized Cancer Medicine, а также программы European Community's Seventh Framework Programme [FP7/2007-2013] № 212877 (UEPHA*MS).*

СПИСОК ЛИТЕРАТУРЫ

1. Bland J.M., Altman D.G. // BMJ. 2000. V. 320. № 7247. P. 1468.
2. Hattersley A.T., McCarthy M.I. // Lancet. 2005. V. 366. № 9493. P. 1315–1323.
3. Laird N.M., Lange C. // Nat. Rev. Genet. 2006. V. 7. № 5. P. 385–394.
4. Spielman R.S., McGinnis R.E., Ewens W.J. // Am. J. Hum. Gen. 1993. V. 52. № 3. P. 506–516.
5. Thomson G. // Am. J. Hum. Genet. 1995. V. 57. № 2. P. 474–486.
6. Fisher R.A. // J. Roy. Statistical Society. 1922. V. 85. № 1. P. 87–94.
7. Sheskin D. Handbook of parametric and nonparametric statistical procedures. Boca Raton, Florida: CRC Press, 2004. 752 p.
8. Freeman G.H., Halton J.H. // Biometrika. 1951. V. 38. № 1–2. P. 141–149.
9. Mantel N. // JASA. 1963. V. 58. № 303. P. 690–700.
10. Kwon J.M., Goate A.M. // Alcohol Res. Health. 2000. V. 24. № 3. P. 164–168.
11. Cortina-Borja M., Smith A.D., Combarros O., Lehmann D.J. // BMC Res. Notes. 2009. V. 2. № 1. P. 105.
12. Cordell H.J. // Nat. Rev. Genet. 2009. V. 10. № 6. P. 392–404.
13. Ahn J., Yu K., Stolzenberg-Solomon R., Simon K.C., McCullough M.L., Gallicchio L., Jacobs E.J., Ascherio A., Helzlsouer K., Jacobs K.B., et al. // Hum. Mol. Genet. 2010. V. 19. № 13. P. 2739–2745.
14. Jakkula E., Leppä V., Sulonen A.-M., Varilo T., Kallio S., Kemppinen A., Purcell S., Koivisto K., Tienari P., Sumelahti M.-L., et al. // Am. J. Hum. Genet. 2010. V. 86. № 2. P. 285–291.
15. Kemppinen A., Sawcer S., Compston A. // Brief Funct Genomics. 2011. V. 10. № 2. P. 61–70.
16. Wang J.H., Pappas D., Jager P.L.D., Pelletier D., de Bakker P.I., Kappos L., Polman C.H., Australian and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene), Chibnik L.B., Hafler D.A., et al. // Genome Med. 2011. V. 3. № 1. P. 3.
17. Hirschhorn J.N., Daly M.J. // Nat. Rev. Genet. 2005. V. 6. № 2. P. 95–108.
18. Schork N.J., Murray S.S., Frazer K.A., Topol E.J. // Curr. Opin. Genet. Dev. 2009. V. 19. № 3. P. 212–219.
19. Li B., Leal S.M. // Am. J. Human Genet. 2008. V. 83. № 3. P. 311–321.
20. Madsen B.E., Browning S.R. // PLoS Genet. 2009. V. 5. № 2. P. e1000384.
21. Neale B.M., Rivas M.A., Voight B.F., Altshuler D., Devlin B., Orho-Melander M., Kathiresan S., Purcell S.M., Roeder K., Daly M.J. // PLoS Genet. 2011. V. 7. № 3. P. e1001322.
22. Bland J.M., Altman D.G. // BMJ. 1995. V. 310. № 6973. P. 170.
23. Westfall P.H., Young S.S. Resampling-based multiple testing: Examples and methods for p-value adjustment. N. Y.: Wiley-Intersci, 1993. 316 p.
24. Benjamini Y., Hochberg Y. // J. Roy. Statistical. Soc. Ser. B Stat. Methodol. 1995. P. 289–300.
25. Storey J.D., Tibshirani R. // Proc. Natl. Acad. Sci. USA. 2003. V. 100. № 16. P. 9440–9445.
26. Cordell H.J. // Hum. Mol. Genet. 2002. V. 11. № 20. P. 2463–2468.
27. Phillips P.C. // Nat. Rev. Genet. 2008. V. 9. № 11. P. 855–867.
28. Favorova O.O., Andreewski T.V., Boiko A.N., Sudomoina M.A., Alekseenkov A.D., Kulakova O.G., Slanova A.V., Gusev E.I. // Neurology. 2002. V. 59. № 10. P. 1652.
29. Tsareva E.Y., Kulakova O.G., Boyko A.N., Shchur S.G., Lvovs D., Favorov A.V., Gusev E.I., Vandenbroeck K., Favorova O.O. // Pharmacogenomics. 2012. V. 13. № 1. P. 43–53.
30. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J., et al. // Am. J. Hum. Genet. 2007. V. 81. № 3. P. 559–575.
31. http://www.genabel.org/
32. Aulchenko Y.S., Ripke S., Isaacs A., van Duijn C.M. // Bioinformatics. 2007. V. 23. № 10. P. 1294–1296.
33. Nunkesser R., Bernholt T., Schwender H., Ickstadt K., Wegener I. // Bioinformatics. 2007. V. 23. № 24. P. 3280–3288.
34. Motsinger-Reif A.A., Dudek S.M., Hahn L.W., Ritchie M.D. // Genet. Epidemiol. 2008. V. 32. № 4. P. 325–340.
35. Long Q., Zhang Q., Ott J. // BMC Bioinformatics. 2009. V. 10. № Suppl 1. P. S75.
36. Ritchie M.D., Hahn L.W., Roodi N., Bailey L.R., Dupont W.D., Parl F.F., Moore J.H. // Am. J. Hum. Genet. 2001. V. 69. № 1. P. 138–147.
37. http://code.google.com/p/apsampler/
38. Favorov A.V., Andreewski T.V., Sudomoina M.A., Favorova O.O., Parmigiani G., Ochs M.F. // Genetics. 2005. V. 171. № 4. P. 2113–2121.
39. http://sites.stat.psu.edu/~yuzhang/
40. Zhang Y., Liu J.S. // Nat. Genet. 2007. V. 39. № 9. P. 1167–1173.
41. http://cran.r-project.org/web/packages/LogicReg/index.html
42. Ruczinski C.K.I., LeBlanc M.L., Hsu L. // Genet. Epidemiol. 2001. V. 21. № 1. P. S626–S631.
43. Kooperberg C., Ruczinski I. // Genet. Epidemiol. 2005. V. 28. № 2. P. 157–170.
44. Cleves M.A., Olson J.M., Jacobs K.B. // Genet. Epidemiol. 1997. V. 14. № 4. P. 337–347.
45. Becker T., Knapp M. // Genet. Epidemiol. 2004. V. 27. № 1. P. 21–32.
46. Sham P.C., Curtis D. // Ann. Hum. Genet. 1995. V. 59. Pt 3. P. 323–336.
47. Cordell H.J., Barratt B.J., Clayton D.G. // Genet. Epidemiol. 2004. V. 26. № 3. P. 167–185.
48. Herold C., Becker T. // Bioinformatics. 2009. V. 25. № 1. P. 134–136.
49. Gibbs R.A., Belmont J.W., Hardenbol P., Willis T.D., Yu F., Yang H., Ch'ang L.Y., Huang W., Liu B., Shen Y., et al. // Nature. 2003. V. 426. № 6968. P. 789–796.
50. http://famhap.meb.uni-bonn.de/
51. http://pngu.mgh.harvard.edu/~purcell/plink/
52. Julià A., Ballina J., Cañete J.D., Balsa A., Tornero-Molina J., Naranjo A., Alperi-López M., Erra A., Pascual-Salcedo D., Barceló P., et al. // Arthritis Rheum. 2008. V. 58. № 8. P. 2275–2286.
53. Shen L., Kim S., Risacher S.L., Nho K., Swaminathan S., West J.D., Foroud T., Pankratz N., Moore J.H., Sloan C.D., et al. // Neuroimage. 2010. V. 53. № 3. P. 1051–1063.
54. Turton J.C., Bullock J., Medway C., Shi H., Brown K., Belbin O., Kalsheker N., Carrasquillo M.M., Dickson D.W., Graff-Radford N.R., et al. // J. Alzheimers Dis. 2011. V. 25. № 4. P. 635–644.
55. Orton S.M., Ramagopalan S.V., Para A.E., Lincoln M.R., Handunnetthi L., Chao M.J., Morahan J., Morrison K.M., Sadovnick A.D., Ebers G.C. // J. Neurol. Sci. 2011. V. 305. № 1–2. P. 116–120.
56. Schüpbach T., Xenarios I., Bergmann S., Kapur K. // Bioinformatics. 2010. V. 26. № 11. P. 1468–1469.
57. Hosmer D.W., Lemeshow S. Applied logistic regression. N. Y.: John Wiley & Sons, 2000. 373 p.
58. Mukherjee B., Ahn J., Gruber S.B., Rennert G., Moreno V., Chatterjee N. // Genet. Epidemiol. 2008. V. 32. № 7. P. 615–626.

59. http://www.multifactordimensionalityreduction.org/
60. Ritchie M.D., Motsinger A.A. // Pharmacogenomics. 2005. V. 6. № 8. P. 823–834.
61. Brassat D., Motsinger A.A., Caillier S.J., Erlich H.A., Walker K., Steiner L.L., Cree B.A.C., Barcellos L.F., Pericak-Vance M.A., Schmidt S., et al. // Genes Immun. 2006. V. 7. № 4. P. 310–315.
62. Greene C.S., Sinnott-Armstrong N.A., Himmelstein D.S., Park P.J., Moore J.H., Harris B.T. // Bioinformatics. 2010. V. 26. № 5. P. 694–695.
63. Ma J., Amos C.I., Warwick Daw E. // Genet. Epidemiol. 2007. V. 31. № 6. P. 594–604.
64. Albrechtsen A., Castella S., Andersen G., Hansen T., Pedersen O., Nielsen R. // Genetics. 2007. V. 176. № 2. P. 1197–1208.
65. Schwender H., Ruczinski I. // Adv. Genet. 2010. V. 72. P. 25–45.
66. Zhang Y., Jiang B., Zhu J., Liu J.S. // Ann. Hum. Genet. 2011. V. 75. № 1. P. 183–193.
67. O'Doherty C., Favorov A., Heggarty S., Graham C., Favorova O., Ochs M., Hawkins S., Hutchinson M., O'Rourke K., Vandenbroeck K. // Pharmacogenomics. 2009. V. 10. № 7. P. 1177–1186.
68. Favorova O.O., Favorov A.V., Boiko A.N., Andreewski T.V., Sudomoina M.A., Alekseenkov A.D., Kulakova O.G., Gusev E.I., Parmigiani G., Ochs M.F. // BMC Med. Genet. 2006. V. 7. P. 63–72.
69. Чихладзе Н.М., Самедова Х.Ф., Судомоина М.А., Thant M., Htut Z.M., Литонова Г.Н., Фаворов А.В., Чазова И.Е., Фаворова О.О. // Кардиология. 2008. Т. 48. № 1. С. 37–42.
70. Судомоина М.А., Николаева Т.Я., Парфенов М.Г., Алексеенков А.Д., Фаворов А.В., Гехт А.Б., Гусев Е.И., Фаворова О.О. // Кардиологический вестник. 2007. Т. 2. № 1. С. 22–25.
71. Парфенов М.Г., Чугунова С.А., Николаева Т. Я., Кобылина О. В., Судомоина М.А., Колядина Ю.А., Гехт А.Б., Гусев Е.И., Фаворова О.О. // Молекуляр. медицина. 2008. № 2. С. 55–59.
72. Судомоина М.А., Сухинина Т.С., Барсова Р.М., Фаворов А.В., Шахнович Р.М., Титов Б.В., Матвеева Н.А., Рыбалкин И.Н., Власик Т.Н., Руда М.Я. и др. // Молекуляр. биология. 2010. Т. 44. № 3. С. 463–471.
73. Титов Б.В., Барсова Р.М., Мартынов М.Ю., Никонова А.А., Фаворов А.В., Гусев Е.И., Фаворова О.О. // Молекуляр. биология. 2012. Т. 46. № 1. С. 93–102.
74. Парфенов М.Г., Титов Б.В., Судомоина М.А., Мартынов М.Ю., Фаворов А.В., Ochs M.F., Гусев Е.И., Фаворова О.О. // Молекуляр. биология. 2009. V. 43. № 5. P. 937–945.
75. Чугунова С.А., Судомоина М.А., Николаева Т.Я., Парфенов М.Г., Макарычева О.Ю., Гехт А.Б., Фаворова О.О. // Якутский мед. журн. 2009. Т. 2. № 26. С. 105–107.
76. Федорова С.А., Бермишева М.А., Виллемс Р., Максимова Н.Р., Кононова С.К., Степанова С.К., Куличкин С.С., Хуснутдинова Э.К. // Якутский мед. журн. 2003. № 1. С. 16–21.
77. Царёва Е.Ю., Кулакова О.Г., Макарычева О.Ю., Бойко А.Н., Щур С.Г., Лащ Н.Ю., Попова Н.Ф., Гусев Е.И., Башинская В.В., Львов Д.В., и др. // Молекуляр. биология. 2011. Т. 45. № 6. С. 963–972.
78. Судомоина М.А., Бойко А.Н., Демина Т.Л., Гусев Е.И., Болдырева М.Н., Трофимов Д.Ю., Алексеев А.Л., Фаворова О.О. // Молекуляр. биология. 1998. Т. 32. № 2. С. 291–296.
79. Boiko A.N., Gusev E.I., Sudomoina M.A., Alekseenkov A.D., Kulakova O.G., Bikova O.V., Maslova O.I., Guseva M.R., Boiko S.Y., Guseva M.E., et al. // Neurology. 2002. V. 58. № 4. P. 658.
80. Gusev E., Sudomoina M., Boiko A., Deomina T., Favorova O. Frontiers in multiple sclerosis. //Eds Abramsky O., Ovadia H. London: Martin Dunitz Publishers, 1997. P. 35–41.
81. Fuller G.M., Zhang Z. // Ann. N.Y. Acad. Sci. 2001. V. 936. № 1. P. 469–479.
82. O'Brien E., Asmar R., Beilin L., Imai Y., Mallion J.M., Mancia G., Mengden T., Myers M., Padfield P., Palatini P., et al. // J. Hypertens. 2003. V. 21. № 5. P. 821–848.
83. Goecks J., Nekrutenko A., Taylor J., Team T.G. // Genome Biol. 2010. V. 11. № 8. P. R86.
84. Ihaka R., Gentleman R. // J. Comput. Graph. Statist. 1996. P. 299–314.
85. Moore J.H. http://compgen.blogspot.com/2005/05/mdr-data-tool.html
86. Bush W.S., Dudek S.M., Ritchie M.D. // Bioinformatics. 2006. V. 22. № 17. P. 2173–2174.
87. Peng T., Du P., Li Y. // Bioinformation. 2009. V. 3. № 8. P. 349–351.