МЕТОДОЛОГИЧЕСКИЕ ПОДХОДЫ И СТРАТЕГИИ КАРТИРОВАНИЯ ГЕНОВ, КОНТРОЛИРУЮЩИХ КОМПЛЕКСНЫЕ ПРИЗНАКИ ЧЕЛОВЕКА

Ю.С. Аульченко[1, 2], Т.И. Аксенович[1]

[1] Институт цитологии и генетики СО РАН, Новосибирск, Россия; [2] Erasmus Medical Center Rotterdam, The Netherlands, e-mail: yurii@bionet.nsc.ru; i.aoultchenko@erasmusmc.nl

Данный обзор посвящен методологии поиска генов, аллельная вариация которых связана с риском развития комплексных болезней. Обсуждаются принципы, лежащие в основе статистических методов, используемых при анализе генетико-эпидемиологических данных, и проблемы, возникающие при их использовании. Рассматриваются две основные стратегии генетического картирования – анализ генов-кандидатов и позиционное клонирование – и обсуждаются их основные ограничения. Также анализируются некоторые проблемы, связанные с генетической гетерогенностью комплексных признаков и взаимодействием генов, участвующих в их контроле.

Комплексные признаки человека

Комплексными (или сложными, мультифакторными, полигенными) называются такие признаки, которые контролируются множественными взаимодействующими факторами как генетической, так и средовой природы. С точки зрения медицинской генетики наиболее интересными комплексными признаками являются распространенные болезни (например, диабет или гипертония), а также некоторые количественные признаки, выступающие в роли эндогенных факторов риска этих болезней (например, физическая активность, недостаток которой является фактором риска сердечно-сосудистых и других патологий; индекс массы тела, увеличение которого способствует развитию диабета и коронарной болезни сердца; курение и др.).

Средовые и генетические факторы, контролирующие формирование комплексных признаков, различаются по времени действия и степени детерминированности. Каждый человек получает от своих родителей генетические факторы в форме специфичных аллелей. Уже на стадии зиготы наличие этих факторов детерминировано, они не меняются на протяжении жизни. Средовые факторы определяются состоянием окружающей среды и образом жизни. Они варьируют в разные периоды времени, и не всегда можно предсказать или проконтролировать их изменение. Для того чтобы понять механизм генетического контроля признака, необходимо не только идентифицировать гены, принимающие участие в контроле признака, описать их аллельные варианты и взаимодействие между ними, но и выяснить, существует ли взаимодействие между аллелями и средовыми факторами.

В чем заключается отличие комплексных болезней от классических моногенных? Для моногенных (таких, как фенилкетонурия или муковисцидоз) выполняется правило классической генетики «один ген – один белок – один признак». Это значит, что причиной моногенной болезни является повреждение одного определенного гена, и проявление мутации этого гена, т. е. развитие болезни, слабо модифицируется другими генами и средовыми эффектами. В этом случае говорят о полной пенетрантности мутантного генотипа. Мутации, вовлеченные в моногенные болезни, как правило, приводят к качественным изменениям – полной потере геном его функции. Это могут быть делеции и мутации, приводящие к преждевременной остановке трансляции или изменению аминокислотной последовательности. Каждая из таких мутаций редко встречается в популя-

ции, но ассоциированный с ними риск болезни превышает среднепопуляционный в десятки и сотни раз. Следует отметить, что определенные семейные формы распространенных заболеваний могут также контролироваться моногенно. Например, для болезни Альцгеймера (OMIM #104300) известны семейные формы, вызываемые редкими мутациями большого эффекта в генах *APP*, *PSEN1* и *PSEN2*.

Как правило, мутации, вовлеченные в контроль комплексных заболеваний, имеют количественный эффект, т. е. приводят не к полному отсутствию определенного белка, а к изменению его концентрации, зачастую только в определенной ткани и/или на определенной стадии развития. Часто это мутации регуляторных генов, изменяющие интенсивность транскрипции кодирующих участков генома, имеющих нормальную структуру. Мутации в регуляторных последовательностях – не единственные причины изменения трансляционных процессов, которые также зависят от многих внешних и внутренних факторов. Поэтому в случае комплексных болезней мутантный генотип обладает неполной пенетрантностью, и риск, ассоциированный с такой мутацией, превышает среднепопуляционный всего в несколько (2–3) раз.

Некоторые аллели малого эффекта достаточно часто встречаются в популяции. В этом случае говорят о модели «распространенная болезнь–распространенный генетический вариант» (Common Disease – Common Variant hypothesis) (Lander, 1996; Reich, Lander, 2001). Примером такой модели является участие аллеля *e4* гена *APOE* в контроле болезни Альцгеймера. Частота этого аллеля в европейских популяциях составляет от 5 до 20 %, а риск заболевания для его носителей повышается приблизительно в 3 раза по сравнению со среднепопуляционным. Для гомозигот по аллелю *e4* риск повышается в 15 раз. Аллель *e4* объясняет значительную долю случаев болезни Альцгеймера (около 20 %). Сходная картина обнаруживается для диабета второго типа, риск которого повышен в полтора раза для носителей аллеля *Pro12Ala* гена *PPAR-γ* – частота этого аллеля в европейских популяциях составляет примерно 80–90 %.

Кроме описанной выше модели в архитектуре комплексных болезней встречается другая модель – «распространенная болезнь – множественные редкие генетические варианты» (Common Disease – Multiple Rare Variants) (Weiss, Terwilliger, 2000; Wright, Hastie, 2001). Ее примером может служить детерминация рака молочной железы, обусловленная мутациями гена *BRCA1* (OMIM +113705), ответственного за репарацию ДНК. Наличие мутантного аллеля в генотипе повышает риск рака груди до 80 %, что примерно в 10 раз выше среднепопуляционного риска. Известно несколько сотен мутантных вариантов гена *BRCA1*. Однако суммарная частота этих аллелей мала, в Европе только около 3 % заболеваний раком груди может быть объяснено мутациями этого гена.

В настоящее время общепринятой является точка зрения, что для большинства комплексных болезней контроль осуществляется по смешанному типу, т. е. в нем принимают участие как редкие, так и относительно распространенные аллели (Reich, Lander, 2001; Wright, Hastie, 2001; Pritchard, Cox, 2002). Такая ситуация возможна даже в рамках одного гена. Например, у 60 % людей, страдающих болезнью Крона (OMIM #266600), в гене *CARD15* обнаруживается один из трех распространенных аллелей, а еще у 20 % болезнь объясняется присутствием одного из 27 редких аллелей этого же гена (Lesage *et al*., 2002).

Идентификация генов, аллельные варианты которых изменяют риск болезни, имеет большое фундаментальное значение, так как позволяет понять биологию развития заболевания, разработать новые подходы к его лечению. Кроме того, знание аллельных вариантов, повышающих риск заболевания, имеет большое практическое значение. С его помощью можно будет устанавливать генетические профили больных и проводить индивидуальное лечение, специфически компенсируя функцию, затронутую генетическим дефектом. Используя знания о взаимодействии аллельных вариантов со средовыми факторами, можно разрабатывать индивидуальные рекомендации по изменению стиля жизни, что позволит минимизировать риск заболевания.

Каким же образом находят гены, аллель-

ная вариация которых связана с риском комплексных болезней? Этот обзор посвящен методологии этого поиска. Мы обсудим принципы, лежащие в основе статистических методов, используемых при анализе генетико-эпидемиологических данных, и проблемы, возникающие при их использовании; опишем две основные стратегии генетического картирования – анализ генов-кандидатов и позиционное клонирование – и обсудим их основные ограничения, а также рассмотрим некоторые проблемы, связанные с генетической гетерогенностью комплексных признаков и взаимодействием генов, участвующих в их контроле.

Методы генетического картирования

В основе картирования генов лежат хорошо известные биологические явления: сцепление генов, их рекомбинация во время мейоза и полиморфность генома. Благодаря сцеплению, мутация, детерминирующая болезнь, передается потомкам вместе с блоком окружающих ее аллелей соседних локусов. Рекомбинация в ряду поколений уменьшает размер этих блоков. Чем ближе расположены два локуса, тем дольше их аллели сохраняются в одном блоке (рис. 1). Идентификация блоков, полученных от различных родителей, обеспечивается полиморфностью генома, многие локусы которого имеют не один, а несколько вариантов нуклеотидных последовательностей. Такие локусы служат генетическими маркерами. Для того чтобы картировать ген, вызывающий болезнь, достаточно доказать совместную сегрегацию болезни и блока маркерных аллелей.

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

Анализ сцепления

Основная идея анализа сцепления, или рекомбинационного анализа, заключается в поиске блока маркеров, которые передаются от больного родителя преимущественно больным потомкам и не передаются здоровым. В разных семьях аллельный состав таких блоков может различаться, но их позиция в геноме должна быть одинакова. Информативными для анализа сцепления являются только гетерозиготные маркерные локусы. Поэтому предпочтительными для анализа являются полиморфные маркеры, имеющие много аллелей. Материалом для анализа сцепления всегда служат родственники: это могут быть пары больных сибсов или расширенные родословные. Анализ сцепления позволяет локализовать ген на участке в 5–50 сМ. Это происходит потому, что

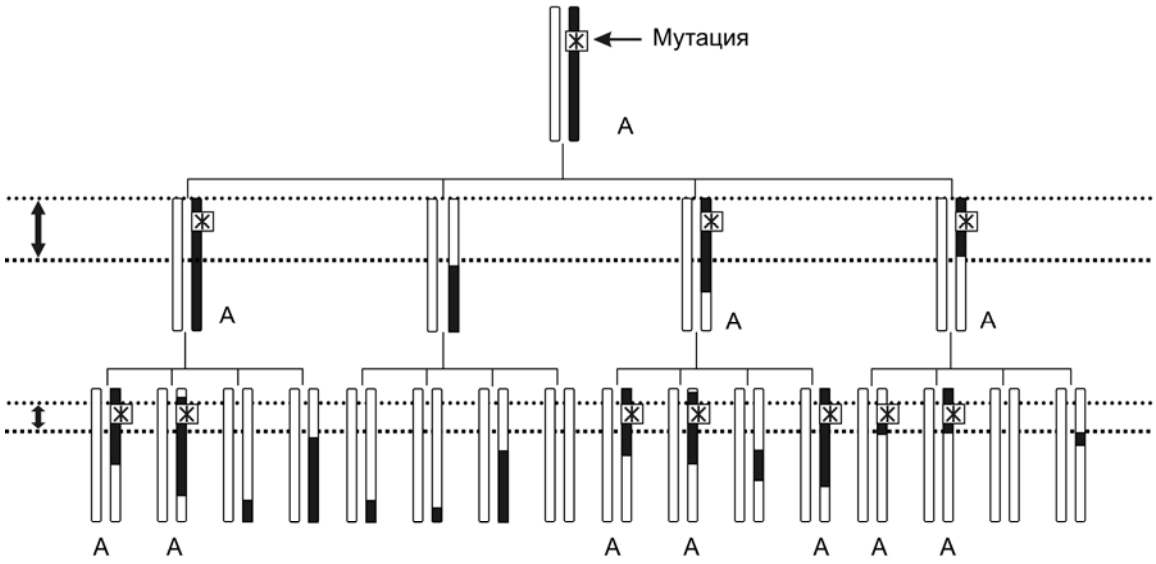

Рис. 1. Совместная сегрегация мутации (отмечено квадратом) и болезни (А) в ряду поколений. Видно, что размер блока, передающегося больным, уменьшается в ряду поколений за счет рекомбинации.

доступными для генотипирования являются представители не более 3–5 поколений, а размеры семей, как правило, не превышают несколько десятков человек. В таких родословных реализуется не так много рекомбинационных событий и блоки передаваемых генов велики. Идентификация косегрегирующих с болезнью блоков осуществляется с помощью различных методов статистического анализа. Самыми мощными считаются методы, базирующиеся на известной модели наследования признака, которая включает оценку популяционной частоты мутантного аллеля и пенетрантности генотипов (Thompson, 2001). Однако установление модели наследования для комплексных болезней – достаточно сложная задача. Из-за этих трудностей, а также поскольку искажение модели наследования приводит к потере мощности, более популярными являются статистические методы, свободные от модели наследования. В их основе лежит анализ идентичности по происхождению маркерных аллелей у пар больных родственников (Holmans, 2001).

Анализ ассоциаций

Второй метод картирования – анализ ассоциаций или неравновесия по сцеплению. Неравновесие по сцеплению между двумя аллелями разных локусов выражается в том, что частота их совместной встречи в популяции отличается от ожидаемой при случайной независимой встрече. Одной из основных, хотя и не единственной причиной существования неравновесия по сцеплению в популяции является тесное сцепление. Например, если в момент возникновения мутации, вызывающей болезнь, рядом находился определенный маркерный аллель, то в течение многих поколений этот аллель будет передаваться вместе с мутацией. Рекомбинация постепенно разрушает ассоциацию и происходит это тем быстрее, чем дальше друг от друга расположены локусы. Для тесно сцепленных (1–2 сМ) локусов неравновесие по сцеплению сохраняется десятки поколений (Ott, 1999).

Основная идея картирования с помощью анализа ассоциаций заключается в следующем. Если у большинства больных в популяции мутантный аллель имеет общее происхождение, окружающие маркеры находятся с ним в неравновесии по сцеплению. Для локализации гена, контролирующего болезнь, надо найти такой маркер, один из аллелей которого преобладает у больных. В отличие от анализа сцепления здесь предполагается, что у больных из разных семей этот маркер не только имеет одинаковую локализацию в геноме, но и содержит один и тот же аллель. Поэтому при анализе ассоциаций не надо исследовать родословные, материалом для этого анализа могут служить независимые группы больных и здоровых людей. Тем не менее предположение об общности мутации у большинства больных означает наличие общего предка, существовавшего много поколений назад. За время, необходимое для распространения болезни в популяции, произошло много рекомбинационных событий, и неравновесие по сцеплению могло сохраниться только между мутацией и аллелем тесно сцепленного маркера. Поэтому с помощью анализа неравновесия по сцеплению удается локализовать ген на участке менее 1 сМ. Маркеры должны плотно покрывать генетическую карту, и число аллелей не должно быть слишком большим. Идеальными маркерами для анализа неравновесия по сцеплению являются SNP-маркеры, характеризующиеся полиморфизмом единичных нуклеотидов.

Как видно, анализ ассоциаций обладает рядом преимуществ, а именно: он может осуществляться на популяционных данных и обладает высокой разрешающей способностью. Вместе с тем анализ ассоциаций имеет ряд недостатков.

Как показывает опыт, воспроизводимость результатов, полученных этим методом, может быть низка (Cardon, Bell, 2001; Hirschhorn *et al*., 2002). Так, Хиршхорн с соавторами обнаружили, что из 166 повторно тестируемых локусов только 6 продемонстрировали ассоциацию во всех повторах (Hirschhorn *et al*., 2002). В чем здесь причина? Почему результаты, полученные с помощью анализа ассоциаций, часто не подтверждаются на других выборках или в других популяциях? Почему в локусах, указанных этим методом, часто не находят генов, контролирующих болезнь?

Одной из причин получения ложноположительных результатов является то, что тесное сцепление генов – не единственная причина возникновения неравновесия по сцеплению. Оно может появиться в выборке из-за кластеризации данных или подразделенности популяции (Abecasis *et al*., 2005). Кластеризация выражается в том, что при наследственной патологии в группу больных часто попадают близкие родственники. Если в семье таких людей с большой частотой встречается какой-то аллель любого локуса, то его частота будет повышенной в группе больных, независимо от локализации маркера. В этом случае на основании анализа ассоциаций будет указан локус, не содержащий искомого гена, и дальнейшие исследования будут направлены по ложному пути. Проблема кластеризации данных является достаточно серьезной, и сейчас разработан ряд пакетов программ, позволяющих установить степень родства двух людей на основании информации об их маркерных генотипах (Abecasis *et al*., 2001). Без такой предварительной проверки положительный результат анализа ассоциаций нельзя интерпретировать как указание на сцепление маркера с искомым геном.

Вторая причина, когда может быть получен ложно-положительный результат, – подразделенность популяций. Популяции человека, строго говоря, не являются панмиксными. В них всегда присутствует подразделенность, основанная на этнических, религиозных, социальных, культурных особенностях (Freedman *et al*., 2004). Если в одной субпопуляции одновременно наблюдается повышенная частота болезни и повышенная частота какого-то аллеля-маркера, то ассоциация болезни с этим аллелем будет установлена независимо от его положения. В настоящее время существуют пакеты программ, которые позволяют тестировать подразделенность популяций на основе геномных данных (Pritchard, Rosenberg, 1999; Pritchard *et al*., 2000).

Для того чтобы решить проблему подразделенности, было предложено использовать родительский контроль (Spielman, Ewens, 1996). В этом случае вместо группы больных и здоровых людей набирается только группа больных, но у каждого из них определяются генотипы родителей (рис. 2). Из четырех родительских аллелей маркерного гена только два передаются больному потомку. Два других составляют контрольную группу. Если частота определенного аллеля в группе переданных оказывается выше, чем в группе непереданных аллелей, то ассоциация считается установленной. Эта ассоциация уже однозначно интерпретируется как сцепление и указывает на расположение искомого гена в геноме. С помощью родительского контроля решается также и проблема кластеризации данных.

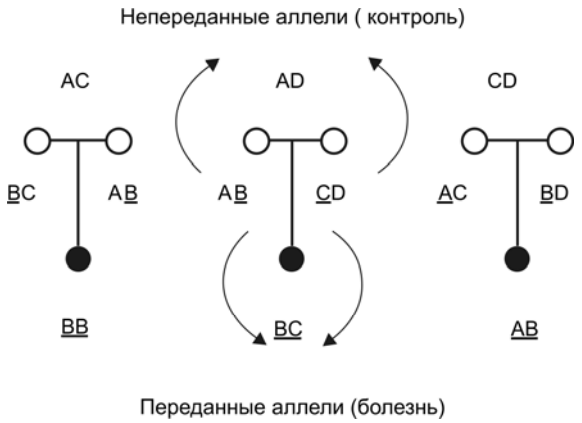

Рис. 2. Схема формирования данных с использованием родительского контроля.

Другое ограничение метода анализа ассоциаций связано с тем, что с его помощью не всегда можно обнаружить сцепление, устанавливаемое с помощью рекомбинационного анализа (низкая мощность). Одной из причин этого служит постепенное разрушение неравновесия по сцеплению между мутацией и аллелем соседнего локуса, возникшее в результате тесного сцепления. Чаще всего это происходит из-за рекомбинаций между геном, контролирующим болезнь, и маркерным локусом. В результате этого у разных людей рядом с мутировавшим геном оказываются разные маркерные аллели. Поэтому при картировании с помощью анализа ассоциаций следует использовать очень плотные генетические карты, чтобы для каждой возможной позиции искомого гена существовал близко расположенный маркер. Другая причина разрушения неравновесия по сцеплению – мутации в маркерных локусах. К счастью, они возникают достаточно редко, особенно, если в качестве маркеров

используются SNP (Johnson, Todd, 2000). Хотя анализ ассоциаций часто обладает меньшей мощностью, чем анализ сцепления, в ряде случаев наблюдается обратная ситуация. Например, было показано, что анализ ассоциаций демонстрирует большую мощность, чем свободный от модели наследования анализ сцепления, если контроль признака осуществляется сравнительно распространенным аллелем малого эффекта (Risch, Merikangas, 1996).

Помимо разрушения неравновесия по сцеплению, к снижению мощности приводит разное происхождение мутации в гене, контролирующем болезнь. Для распространенных болезней человека предположение об общности происхождения мутации в большой открытой популяции, на котором основано использование неравновесия по сцеплению для картирования генов, кажется весьма спорным. Разное происхождение мутаций в одном и том же гене подтверждается тем, что для многих генов, участвующих в детерминации болезни, обнаружены различные патогенные мутации, приводящие к нарушению одного и того же метаболического процесса и имеющие одинаковое фенотипическое проявление (см., например, Gent, Braakman, 2004). Очевидно, что такие мутации впервые возникали у разных людей и что у каждого из них рядом с мутировавшим аллелем могли находиться разные маркерные аллели. В настоящее время разработан ряд методов, которые позволяют учесть аллельную гетерогенность (например, CLUMP), однако статистические свойства этих методов изучены плохо.

Ошибки генотипирования также влияют на результаты анализа ассоциаций. В том случае, если эти ошибки возникают независимо от генотипа и не элиминируются в процессе контроля качества (именно так, как правило, происходит при анализе выборок больных и здоровых), они снижают общую мощность анализа и могут приводить к ложно-негативным результатам. Однако если выборка сформирована из семейных данных с использованием родительского контроля, ошибки генотипирования могут быть выявлены при анализе передачи аллелей от родителей к потомкам. При этом ошибки элиминируются неслучайным образом: например, чаще обнаруживаются и элиминируются ошибки в генотипах, содержащих редкие аллели. Теоретически это может привести к ложно-положительной ассоциации, при которой редкий аллель будет интерпретироваться как «протективный» – предотвращающий развитие болезни.

Все перечисленные факторы снижают мощность анализа ассоциаций, но не анализа сцепления, поскольку последний не привязан к определенному аллелю маркера. Оптимальной является стратегия, когда сначала с помощью анализа сцепления выявляется крупный блок, содержащий картируемый ген, а затем с помощью анализа ассоциаций этот блок сужается (Abecasis *et al*., 2005).

Таким образом, анализ основных принципов картирования генов, контролирующих распространенные болезни человека, выявляет ряд ограничений методов, основанных на тестировании неравновесия по сцеплению. Многих проблем удается избежать при правильном формировании выборки.

Так, если выборка формируется из групп больных и здоровых людей, необходимо следить, чтобы группы были как можно более однородны. Прежде всего, нужно контролировать этническую принадлежность членов выборки, используя специальные опросники или интервью. Желательно выровнять выборки здоровых и больных людей по таким факторам, как пол, возраст, место рождения, социальный статус и т. п. Различие выборок по этим параметрам может привести к ложноположительным результатам.

Если получение выровненных групп больных и здоровых невозможно, вместо группы здоровых следует использовать родительский контроль. К сожалению, для болезней с поздним возрастом проявления генотипы родителей, как правило, недоступны, в этом случае можно использовать родственный контроль – например, сибсов.

Стратегии генетического картирования

Существует два основных подхода к картированию генов комплексных признаков: изучение генов-кандидатов и позиционное клонирование, также называемое сканированием генома (рис. 3). Эти подходы принципиально отличаются между собой в способе

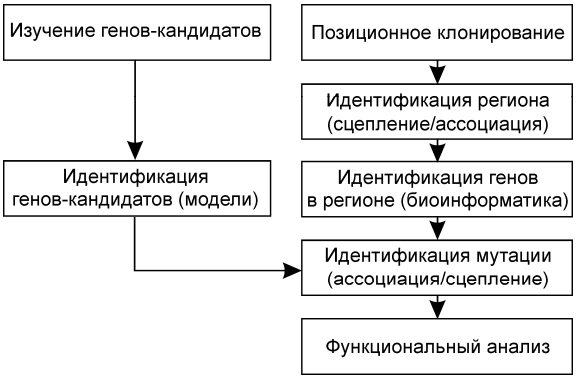

Рис. 3. Основные стратегии генетического картирования.

идентификации гена, потенциально вовлеченного в контроль признака. Однако после идентификации гена оба подхода используют одинаковый инструментарий для выявления аллельной вариации и доказательства вовлеченности гена в контроль признака.

Изучение генов-кандидатов

Ключевым моментом при применении стратегии изучения генов-кандидатов является их выбор, который осуществляется на основе знаний о биологии признака. Как правило, такие знания получаются при изучении модельных объектов. Идеальными свойствами модельного объекта являются короткий жизненный цикл, возможность проведения направленных скрещиваний и других биологических манипуляций, а также адекватность моделируемой системе. Действительно, червь *C. elegans* является идеальным объектом с точки зрения двух первых свойств, однако его адекватность при моделировании многих аспектов сложных признаков человека не очевидна. В то же время модельные объекты, которые максимально близки человеку, например, человекообразные обезьяны, обладают длительным жизненным циклом и сравнительно мало изучены генетически. В этом смысле самыми «сбалансированными» модельными объектами являются, по-видимому, лабораторная мышь и крыса, в экспериментах над которыми получена большая часть наших знаний о генетике сложных признаков.

Одним из наиболее известных успехов применения стратегии генов-кандидатов в популяциях человека было изучение гена, кодирующего ангиотензиноген (*AGT*, OMIM *106150). Роль ренин-ангиотензиновой системы в контроле давления крови была установлена уже в 1970-е гг. при изучении модельных объектов. В 1980-е гг. была охарактеризована структура человеческого гена *AGT* и установлена его позиция в геноме. В 1992 г. Jeunemaitre *et al.* (1992) на примере 379 пар сибсов с повышенным давлением показали, что регион, содержащий *AGT*, сцеплен с этим признаком. В процессе секвенирования *AGT* они нашли 15 полиморфизмов и обнаружили, что два из них ассоциированы с повышенным давлением. В последующие годы было проведено несколько десятков исследований ассоциации и сцепления *AGT* с повышенным давлением крови и другими сердечно-сосудистыми патологиями. В настоящее время может считаться установленным, что некоторые полиморфизмы в гене *AGT* связаны с различными заболеваниями сердца и сосудов.

Для других признаков, таких, как например, ожирение, применение стратегии изучения генов-кандидатов было менее успешным. Ожирение можно охарактеризовать с помощью индекса массы тела (ИМТ), который вычисляется как вес в килограммах, деленный на квадрат роста в метрах. В соответствии с классификацией Всемирной организации здравоохранения ожирением страдают люди с ИМТ более 30, в то время как ИМТ от 25 до 30 классифицируется как избыточная масса тела. Ожирение является фактором риска и осложняет течение диабета второго типа, болезней сердечно-сосудистой системы и опорно-двигательного аппарата. Частота ожирения повышена в индустриально развитых странах, принявших западный стиль жизни. Например, в США частота ожирения выросла в несколько раз за предыдущие несколько десятков лет и составляет сейчас около 30 %; в Москве и Новосибирске эта частота составляет около 10 %, в то время как в Китае – менее 5 %. Более того, в рамках одной страны частота ожирения может быть выше среди городского населения по сравнению с сельским. Эти факты указывают на ключевую роль средовых факторов в контроле ожирения. В то же время анализ наследуемости ИМТ указывает на то, что ожирение являет-

ся генетическим признаком, и около 30 % вариации ИМТ может быть объяснено генами.

До сравнительно недавнего времени о генетическом контроле ожирения было известно мало. Ситуация изменилась в 1994 г., когда позиционно клонировали ген *Ob* (Zhang *et al*., 1994), отсутствие продукта которого (гормона лептина) вызывает повышение в несколько раз веса тела у мышей мутантной линии (рис. 4). Было показано, что гормон лептин вырабатывается клетками жировой ткани, поступает в кровоток и связывается рецептором лептина (*Ob-R*) в гипоталамусе, регулируя аппетит по принципу обратной связи: чем больше жировой ткани, тем больше уровень лептина в крови и тем меньше аппетит, и наоборот.

В том же году была инициирована серия работ, нацеленных на подтверждение роли лептинной системы регуляции веса тела у человека. Однако многочисленные исследования, проводившиеся с помощью сравнения выборок людей с ожирением и без него, не показали наличия аллельных вариантов, ассоциированных с ожирением, ни в гене, кодирующем лептин, ни в его рецепторе.

Только спустя три года, в 1997 г., Фаруки с коллегами (Farooqi *et al*., 2002) удалось идентифицировать троих детей с экстремальной степенью ожирения, вызванной генетически обусловленным отсутствием лептина. Знание генетического и физиологического механизмов контроля признака позволило провести эффективную заместительную терапию, в результате которой понизился аппетит, восстановились относительно нормальные пропорции тела, а также уровень инсулина и липидов крови.

В настоящее время исследования на модельных объектах позволили идентифицировать множество генов, вовлеченных в регуляцию веса тела. Однако изучение этих генов у человека, как правило, повторяет историю гена *Ob* – для многих из них не найдено функциональных аллельных вариантов, если же их находят, они объясняют лишь редкие семейные случаи. Для некоторых генов, таких, как *Ghrelin* (OMIM *605353) и *MC4R* (OMIM *155541), мутантные формы относительно часто встречаются в выборках детей с ожирением. Таким образом, несмотря на огромные успехи генетики ожирения модельных объектов, о генетическом контроле ожирения в популяциях человека мы знаем мало.

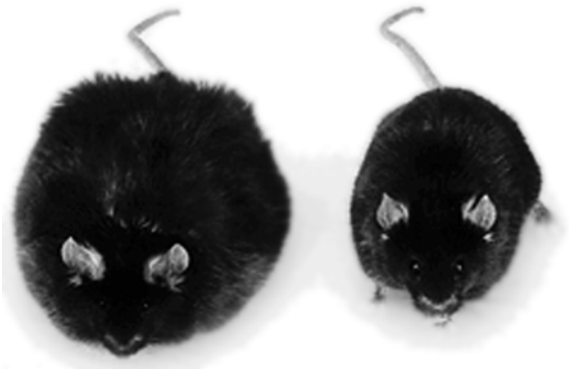

Рис. 4. Лабораторная мышь, гомозиготная по мутации гена *Ob* (слева) и мышь контрольной линии (справа).

Чем вызван относительно малый успех стратегии изучения генов-кандидатов ожирения человека? Очевидно, немаловажную роль играет то, что существующие модели не вполне адекватны. Ожирение, эпидемию которого мы наблюдаем в сообществах, принявших западный стиль жизни, развивается на фоне избытка (зачастую высококалорийной и низкокачественной) еды и отсутствия физической активности. Необходимо разрабатывать линии мышей, моделирующие именно этот тип ожирения. В настоящее время подобные работы активно ведутся.

При изучении генов-кандидатов следует понимать, что не всякий ген, потенциально вовлеченный в контроль признака, имеет аллельные варианты, изменяющие его функцию. Теоретические исследования показывают, что только малая доля таких генов (около 5–10 %) полиморфна и вносит вклад в генетический контроль признака (Pritchard, 2001; Pritchard, Cox, 2002). Таким образом, если список генов-кандидатов составлен на основе знаний о модельных объектах, то шанс найти функциональный вариант относительно мал, и при изучении контроля признаков человека необходимо исследовать десятки таких генов. В идеальном случае ген-кандидат следует выбирать на основании не только биологических знаний, но также и предварительно проведенного геномного сканирования.

Позиционное клонирование

В отличие от анализа генов-кандидатов позиционное клонирование не предполагает использования априорных знаний о генетике признака: каждый участок генома последовательно тестируется на участие в контроле признака. Этот процесс называют сканированием генома. Он позволяет идентифицировать новые гены и расширять наши знания о биологии признака. Сканирование генома можно проводить как с помощью анализа сцепления, так и изучения ассоциаций.

Рассмотрим применение стратегии сканирования генома на примере картирования гена *DJ-1*, вовлеченного в болезнь Паркинсона (OMIM #168600), которое проводилось в генетически изолированной популяции, расположенной на юго-западе Нидерландов. В этой популяции было идентифицировано около 70 больных. Детальное исследование показало, что четверо больных имели раннюю форму болезни и являлись родственниками (рис. 5). Как видно из рисунка 5, родословная является инбредной, что может указывать на рецессивную модель контроля.

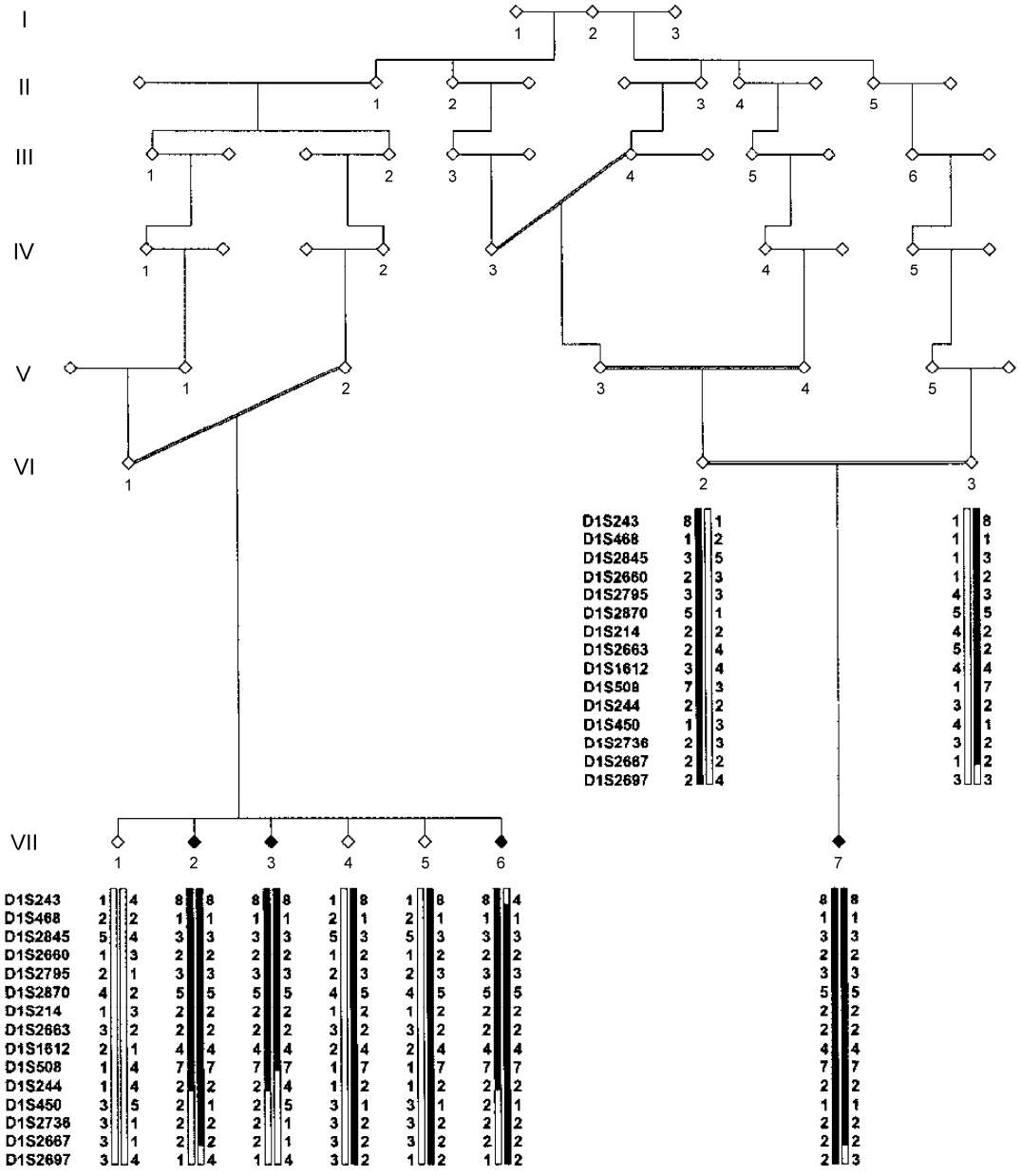

Рис. 5. Родословная пациентов с болезнью Паркинсона (van Duijn *et al*., 2001).

Анализ сцепления, проведенный на основе этой модели с использованием 420 высокополиморфных маркеров, показал, что два соседних маркера, локализованных на хромосоме 1p36, были идентичны и гомозиготны у всех больных, а у здоровых родственников эти аллели либо не были представлены, либо находились в гетерозиготе. Насыщение региона еще 13 маркерами позволило подтвердить его участие в контроле признака (рис. 5) (van Duin *et al*., 2001).

В сцепленном регионе находилось более 30 генов, ни один из которых не был явным кандидатом. Секвенирование 5 этих генов не дало результата. После этого был проведен анализ транскрипции всех генов и было обнаружено, что транскрипт одного из генов, *DJ-1*, отсутствует. Изучение этого гена показало наличие большой делеции (Bonifati *et al*., 2003). В дальнейшем участие этого гена в контроле болезни Паркинсона было показано в других популяциях, были обнаружены различные мутации этого гена, приводящие к болезни.

Идентификация гена *DJ-1* позволила подтвердить роль оксидативного стресса в контроле болезни Паркинсона и расширила наши знания о биологии этой болезни.

Однако, как было отмечено выше, в изучаемой популяции было идентифицировано около 70 пациентов с болезнью Паркинсона. Мы выяснили, что причиной болезни четырех из них являлась редкая моногенная мутация. А как же остальные? К сожалению, сканирование генома является мощным способом идентификации мутаций только большого эффекта. Сейчас геномные данные остальных пациентов изучаются другими методами, в частности, анализом неравновесия по сцеплению.

Основными проблемами стратегии геномного сканирования в настоящее время являются огромная вычислительная сложность этой задачи, связанная с этим относительно слабая разработанность методологии и отсутствие пакетов прикладных программ.

В то время как исследование генов-кандидатов, как правило, предполагает анализ эффекта нескольких полиморфизмов, анализ генома предполагает анализ сотен и тысяч маркеров. Недавно компании Illumina и Affymetrix выпустили на рынок чипы, позволяющие типировать 500 тыс. SNP-маркеров; скоро будут доступны чипы с миллионом и более маркеров. Цены такого анализа стремительно падают, и в недалеком будущем можно ожидать, что генотипирование нескольких сотен или тысяч людей с помощью этих чипов будет финансово приемлемо. Понятно, что анализ подобного объема данных представляет собой огромную задачу. Необходимо разрабатывать новые эффективные методы и пакеты программ для автоматизированного анализа. Более того, проблема геномного сканирования, особенно при рассмотрении эффектов взаимодействия генов, должна решаться с привлечением супер- и параллельных компьютеров. В настоящее время в этом направлении ведутся активные работы (см., например, Dietter *et al*., 2004; Marchini *et al*., 2005; DHPAS).

Проблема генетической гетерогенности

Одной из основных проблем генетического анализа распространенных болезней человека является их генетическая гетерогенность, выражающаяся в том, что в разных семьях и у разных людей в контроле болезни принимают участие разные наборы генов. Эта гетерогенность затрудняет картирование независимо от того, каким методом оно осуществляется, и требует для анализа выборки огромного размера. Чтобы решить эту проблему, необходимо снизить генетическую гетерогенность в анализируемой выборке. Сделать это можно различными способами.

Первый из них заключается в анализе отдельных форм болезни. Для многих болезней с поздним возрастом проявления известны случаи, когда болезнь проявляется относительно рано. Было показано, что частота больных среди родственников пробанда особенно высока в семьях с ранним проявлением болезни (Murff *et al*., 2004). Это свидетельствует о высоком вкладе генетической компоненты в контроль ранних форм болезни и позволяет думать, что в их детерминации участвует небольшое число генов. Перспективность использования этого подхода продемонстрирована на примере ряда болезней: болезни Паркинсона, рассмотренной выше в этом обзоре (van Duijn *et al*., 2001;

Bonifati *et al*., 2003), а также рака молочной железы (Spurr *et al*., 1993; Wooster *et al*., 1994), диабета второго типа (Bowden *et al*., 1992), рака толстого кишечника (Peltomaki *et al*., 1991), болезни Альцгеймера (Chartier-Harlin *et al*., 2004) и др. Однако решить все задачи картирования с помощью этого подхода вряд ли удастся – не для каждой болезни существуют рано проявляющиеся формы, и при их анализе выявляются только те гены, которые специфичны для данной формы болезни.

Второй способ снижения генетической гетерогенности заключается в изучении изолированных популяций человека (Terwilliger *et al*., 1998; Peltonen *et al*., 2000; Chapman, Thompson, 2001; Rannala, 2001). В таких популяциях велик эффект основателя, в результате чего генетическая полиморфность понижена. Как правило, такие популяции зачастую проживают компактно, что снижает вариабельность внешних факторов: климата, питания, социального статуса. Сейчас созданы проекты по изучению ряда изолятов и получены первые многообещающие результаты (Njajou *et al*., 2001; Vaessen *et al*., 2002; Aulchenko *et al*., 2003; Abecasis *et al*., 2005). Изолированные популяции дают совершенно уникальный материал для генетического анализа распространенных заболеваний, но вряд ли можно рассчитывать, что с их помощью будут выявлены все гены значимого эффекта, встречающиеся в открытых популяциях.

В этом смысле более перспективным кажется анализ расширенных родословных (Almasy, Blangero, 2000; Terwilliger, Goring, 2000; Gulcher *et al*., 2001) из открытых популяций. В каждой из таких родословных значительно сужен набор генов, участвующих в контроле болезни. Поэтому, если размер родословной велик, на ее материале можно картировать гены, оказывающие в этой семье большой вклад в предрасположенность к болезни (см., например, Martin *et al*., 2002). Наличие больших родословных позволяет осуществлять картирование, используя как анализ сцепления, так и анализ ассоциаций, а для анализа ассоциаций обеспечивается родительский контроль. Кроме того, имея набор расширенных родословных, можно картировать несколько генов, максимально повышающих риск болезни в каждой из семей. И наконец, в любой большой родословной встречаются разные болезни из списка наиболее распространенных и появляется возможность одновременно анализировать несколько патологий и изучать плейотропное действие генов.

К сожалению, в открытых популяциях генеалогическая информация доступна только в некоторых странах, где она систематически собиралась на государственном уровне. Поэтому перспективным является материал, собранный в молодых генетических изолятах с малым эффектом основателя, таких, как некоторые из религиозных изолятов Европы (Pardo *et al*., 2005).

Взаимодействие генов

Сегодня ни у кого не вызывает сомнения, что идентификация генов, контролирующих распространенные болезни человека, является чрезвычайно сложной задачей. Для ее решения необходимо рассматривать полиморфизм многих сайтов, расположенных по всему геному. Именно поэтому современные методы картирования базируются на сканировании генома с помощью анализа сцепления или анализа ассоциаций. На первый взгляд, кажется, что такое рассмотрение учитывает многие особенности комплексных болезней: участие большого числа генов, их взаимодействие. Однако, несмотря на вовлечение в анализ всего генома, на каждом отдельном этапе анализа рассматривается только один локус и тестируется его причастность к контролю болезни. Тестируя шаг за шагом все участки генома, мы можем идентифицировать те его сайты, полиморфизм которых ассоциирован с болезнью. Хотя в настоящее время существуют методы и даже программное обеспечение, позволяющее одновременно тестировать эффекты нескольких локусов и учитывать их взаимодействие (Dietter *et al.*, 2004; Marchini *et al*., 2005), но они требуют огромных вычислительных ресурсов. Более того, методология анализа взаимодействий является слабо разработанной. Как правило, существующие методы проверяют совместные эффекты только тех локусов, значимость вклада которых доказана при их индивидуальном тести-

ровании (Hoh *et al*., 2000; Nelson *et al*., 2001; Hoh, Ott, 2003). Таким образом, получается, что общая идеология анализа комплексных болезней мало чем отличается от идеологии моногенных болезней – в поле зрения исследователей оказываются лишь те гены, индивидуальный вклад которых в контроль признака значим (Terwilliger, Goring, 2000; Weiss, Terwilliger, 2000; Terwilliger, 2001). Конечно, такой подход существенно ограничивает наши возможности, и сейчас ведется интенсивная работа по созданию новых методов и подходов, учитывающих специфику комплексных болезней (Nelson *et al*., 2001; Hirschhorn, Daly, 2005; Wang *et al*., 2005). Тем не менее существует много ситуаций, где традиционный подход оказывается эффективным. Известно, что для многих болезней вклады детерминирующих их генов различаются, т. е. существует иерархия вкладов генов. Наиболее выраженный случай такой иерархии представляют майоргенные болезни, где ведущую роль в развитии патологии играют аллели одного гена, а все остальные лишь усиливают или ослабляют его эффект. Часто в больших семьях, где снижена генетическая гетерогенность признака, болезнь ассоциирована с аллелями одного локуса. В этих случаях гены, контролирующие комплексные болезни, успешно картируются с помощью традиционного подхода. Несмотря на недостатки и ограничения, именно этот подход позволил нам получить те знания о генетической природе комплексных болезней, которыми мы уже располагаем, именно с его помощью выполняются все работы по картированию генов, публикуемые в каждом номере ведущих генетических журналов. Это позволяет надеяться, что возможности существующих методов еще не исчерпаны, с их помощью можно продолжать исследования и получать новую информацию о генетике распространенных болезней в то время, пока ведется интенсивная работа по созданию новых методов и подходов, учитывающих специфику этих болезней.

Работа выполнена при поддержке гранта Российского фонда фундаментальных исследований (04-04-48074), программы РАН «Динамика генофондов растений, животных и человека» и гранта NWO-RFBR (047-016-009).

Литература

Abecasis G.R., Cherny S.S., Cookson W.O., Cardon L.R. GRR: graphical representation of relationship errors // Bioinformatics. 2001. V. 17. № 8. P. 742–743.

Abecasis G.R., Ghosh D., Nichols T.E. Linkage disequilibrium: ancient history drives the new genetics // Hum. Hered. 2005. V. 59. № 2. P. 118–124.

Almasy L., Blangero J. Challenges for genetic analysis in the 21st century: localizing and characterizing genes for common complex diseases and their quantitative risk factors // Gene Screen. 2000. V. 1. P. 113–116.

Aulchenko Y.S., Vaessen N., Heutink P. *et al.* A genome-wide search for genes involved in type 2 diabetes in a recently genetically isolated population from the Netherlands // Diabetes. 2003. V. 52. № 12. P. 3001–3004.

Bonifati V., Rizzu P., van Baren M.J. *et al.* Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism // Science. 2003. V. 299 (5604). № 10. P. 256–259.

Bowden D.W., Akots G., Rothschild C.B. *et al.* Linkage analysis of maturity-onset diabetes of the young (MODY): genetic heterogeneity and nonpenetrance // Am. J. Hum. Genet. 1992. V. 50. № 3. P. 607–618.

Cardon L.R., Bell J.I. Association study designs for complex diseases // Nat. Rev. Genet. 2001. V. 2. № 2. P. 91–99.

Chapman N.H., Thompson E.A. Linkage disequilibrium mapping: the role of population history, size, and structure // Adv. Genet. 2001. V. 42. P. 413–437.

Chartier-Harlin M.C., Araria-Goumidi L., Lambert J.C. Genetic complexity of Alzheimer's disease // Rev. Neurol. (Paris). 2004. V. 160. № 2. P. 251–255.

CLUMP, Monte Carlo method for assessing significance of case-control association studies with multi-allelic markers. Available at http://www.smd.qmul.ac.uk/statgen/dcurtis/software.html

DHPAS. Development of high-performance angorithms and software for genetic epidemiology of complex traits. Available at http://mga.bionet.nsc.ru/NLRU/

Dietter J., Spiegel A., an Mey D. *et al.* Efficient two-trait-locus linkage analysis through program optimization and parallelization: application to hypercholesterolemia // Eur. J Hum. Genet. 2004. V. 12. № 7. P. 542–550.

Duijn C.M. van, Dekker M.C., Bonifati V. *et al.* Park7, a novel locus for autosomal recessive early-onset parkinsonism, on chromosome 1p36 // Am. J. Hum. Genet. 2001. V. 69. № 3. P. 629–634.

Farooqi I.S., Matarese G., Lord G.M. *et al.* Beneficial effects of leptin on obesity, T cell hyporesponsiveness, and neuroendocrine/metabolic dysfunction of human congenital leptin deficiency // J. Clin. Invest. 2002. V. 110. № 8. P. 1093–1103.

Freedman M.L., Reich D., Penney K.L. *et al.* Assessing the impact of population stratification on genetic association studies // Nat. Genet. 2004. V. 36. № 4. P. 388–393.

Gent J., Braakman I. Low-density lipoprotein receptor structure and folding // Cell Mol. Life Sci. 2004. V. 61. № 19/20. P. 2461–2470.

Gulcher J.R., Kong A., Stefansson K. The role of linkage studies for common diseases // Curr. Opin. Genet. Dev. 2001. V. 11. № 3. P. 264–267.

Hirschhorn J.N., Daly M.J. Genome-wide association studies for common diseases and complex traits // Nat. Rev. Genet. 2005. V. 6. № 2. P. 95–108.

Hirschhorn J.N., Lohmueller K., Byrne E., Hirschhorn K. A comprehensive review of genetic association studies // Genet. Med. 2002. V. 4. № 2. P. 45–61.

Hoh J., Ott J. Mathematical multi-locus approaches to localizing complex human trait genes // Nat. Rev. Genet. 2003. V. 4. № 9. P. 701–709.

Hoh J., Wille A., Zee R. *et al.* Selecting SNPs in two-stage analysis of disease association data: a model-free approach // Ann. Hum. Genet. 2000. V. 64. № 5. P. 413–417.

Holmans P. Nonparametric linkage // Handbook of Statistical Genetics / Ed. D.J. Balding *et al*. N.Y.: John Wiley, Sons, Ltd, 2001. P. 487–505.

IUNS (International Union of Nutritional Sciences). The Global Challenge of Obesity and the International Obesity Task Force. Available at http://www.iuns.org/features/obesity/tabfig.htm

Jeunemaitre X., Soubrier F., Kotelevtsev Y.V. *et al.* Molecular basis of human hypertension: role of angiotensinogen // Cell. 1992. V. 71. P. 7–20.

Johnson G.C., Todd J.A. Strategies in complex disease mapping // Curr. Opin. Genet. Dev. 2000. V. 10. № 3. P. 330–334.

Lander E.S. The new genomics: global views of biology // Science. 1996. V. 274 (5287). P. 536–539.

Lesage S., Zouali H., Cezard J.P. *et al.* CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease // Am. J. Hum. Genet. 2002. V. 70. № 4. P. 845–857.

Marchini J., Donnelly P., Cardon L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases // Nat. Genet. 2005. V. 37. № 4. P. 413–417.

Martin L.J., Comuzzie A.G., Dupont S. *et al.* A quantitative trait locus influencing type 2 diabetes susceptibility maps to a region on 5q in an extended French family // Diabetes. 2002. V. 51. P. 3568–3572.

Murff H.J., Byrne D., Syngal S. Cancer risk assessment: quality and impact of the family history interview // Am. J. Prev. Med. 2004. V. 27. № 3. P. 239–245.

Nelson M.R., Kardia S.L., Ferrell R.E., Sing C.F. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation // Genome Res. 2001. V. 11. № 3. P. 458–470.

Njajou O.T., Vaessen N., Joosse M. *et al.* A mutation in SLC11A3 is associated with autosomal dominant hemochromatosis // Nat. Genet. 2001. V. 28. № 3. P. 213–214.

OMIM (Online Mendelian Inheritance in Men). Available at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Ott J. Analysis of Human Genetic Linkage. Baltimore; London: The Johns Hopkins Univ. Press, 1999. 382 p.

Pardo L.M., MacKay I., Oostra B. *et al.* The effect of genetic drift in a young genetically isolated population // Ann. Hum. Genet. 2005. V. 69. № 3. P. 288–295.

Peltomaki P., Sistonen P., Mecklin J.P. *et al.* Evidence supporting exclusion of the DCC gene and a portion of chromosome 18q as the locus for susceptibility to hereditary nonpolyposis colorectal carcinoma in five kindreds // Cancer Res. 1991. V. 51. № 16. P. 4135–4140.

Peltonen L., Palotie A., Lange K. Use of population isolates for mapping complex traits // Nat. Rev. Genet. 2000. V. 1. № 3. P. 182–190.

Pritchard J.K. Are rare variants responsible for susceptibility to complex disease? // Am. J. Hum. Genet. 2001. V. 69. P. 124–137.

Pritchard J.K., Cox N.J. The allelic architecture of human disease genes: common disease-common variant ... or not? // Hum. Mol. Genet. 2002. V. 11. № 20. P. 2417–2423.

Pritchard J.K., Rosenberg N.A. Use of unlinked genetic markers to detect population stratification in association studies // Am. J. Hum. Genet. 1999. V. 65. P. 220–228.

Pritchard J.K., Stephens M., Donnelly P.J. Inference of population structure using multilocus genotype data // Genetics. 2000. V. 155. P. 945–959.

Rannala B. Finding genes influencing susceptibility to complex diseases in the post-genome era // Am. J. Pharmacogenomics. 2001. V. 1. № 3. P. 203–221.

Reich D.E., Lander E.S. On the allelic spectrum of human disease // Trends Genet. 2001. V. 17. № 9. P. 502–510.

Risch N., Merikangas K. The future of genetic studies of complex human diseases // Science.

1996. V. 273(5281). № 13. P. 1516–1517.
Spielman R.S., Ewens W.J. The TDT and other family-based tests for linkage disequilibrium and association // Am. J. Hum. Genet. 1996. V. 59. № 5. P. 983–989.
Spurr N.K., Kelsell D.P., Black D.M. *et al.* Linkage analysis of early-onset breast and ovarian cancer families, with markers on the long arm of chromosome 17 // Am. J. Hum. Genet. 1993. V. 52. № 4. P. 777–785.
Terwilliger J.D. On the resolution and feasibility of genome scanning approaches // Adv. Genet. 2001. V. 42. P. 351–391.
Terwilliger J.D., Goring H.H. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design // Hum. Biol. 2000. V. 72. № 1. P. 63–132.
Terwilliger J.D., Zollner S., Laan M., Paabo S. Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion // Hum. Hered. 1998. V. 48. № 3. P. 138–154.
Thompson E.A. Linkage analysis // Handbook of Statistical Genetics / Ed. D.J. Balding *et al*. N.Y.: John Wiley, Sons, Ltd, 2001. P. 541–563.
Vaessen N., Heutink P., Houwing-Duistermaat J.J. *et al.* A genome-wide search for linkage-disequilibrium with type 1 diabetes in a recent genetically isolated population from the Netherlands // Diabetes. 2002. V. 51. № 3. P. 856–859.
Wang W.Y., Barratt B.J., Clayton D.G., Todd J.A. Genome-wide association studies: theoretical and practical concerns // Nat. Rev. Genet. 2005. V. 6. № 2. P. 109–118.
Weiss K.M., Terwilliger J.D. How many diseases does it take to map a gene with SNPs? // Nat. Genet. 2000. V. 26. № 2. P. 151–157.
WHO (World Health Organization). Obesity and Overweight: facts sheet. Available at http:// www.who.int/dietphysicalactivity/media/en/ gsfs_obesity.pdf
Wooster R., Neuhausen S.L., Mangion J. *et al.* Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13 // Science. 1994. V. 265. № 5181. P. 2088–2090.
Wright A.F., Hastie N.D. Complex genetic diseases: controversy over the Croesus code // Genome Biol. 2001. V. 8. № 2. P. COMMENT2007.
Zhang Y., Proenca R., Maffei M. *et al.* Positional cloning of the mouse obese gene and its human homologue // Nature. 1994. V. 372. P. 425–432.

METHODOLOGICAL APPROACHES AND STRATEGIES FOR MAPPING GENES CONTROLLING COMPLEX HUMAN TRAITS

Yu.S. Aulchenko[1, 2], T.I. Axenovich[1]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Erasmus Medical Center Rotterdam, The Netherlands, e-mail: yurii@bionet.nsc.ru; i.aoultchenko@erasmusmc.nl

Summary

This review is dedicated to the methodology which is used for identification of genes, whose allelic variation changes susceptibility to complex diseases. Basic principles which underlie statistical methods for analysis of genetic-epidemiological data are described and common problems appearing during such analysis are discussed. We describe two basic strategies of genetic mapping (analysis of candidate genes and positional cloning) and major limitations of these strategies. We also address the issue of genetic heterogeneity and interaction between genes involved into control of complex diseases.