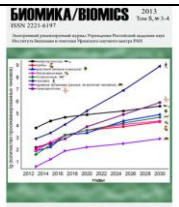
Биомика, 2013, Том 5, № 3-4, 75-125

БИОМИКА/BIOMICS

http://biomics.ru



НЕКОТОРОЕ ТЕХНОЛОГИЧЕСКОЕ ПРОШЛОЕ, НАСТОЯЩЕЕ, А ТАКЖЕ БУДУЩЕЕ СОВРЕМЕННОЙ БИОЛОГИИ К 2030 ГОДУ *(ЧАСТЬ ВТОРАЯ)*[34]

Чемерис А.В., Магданов Э.Г., Гарафутдинов Р.Р., Матниязов Р.Т., Баймиев Ал.Х., Баймиев Ан.Х., Бикбулатова С.М., Гималов Ф.Р., Вахитов В.А.

Федеральное государственное бюджетное учреждение науки
Институт биохимии и генетики Уфимского научного центра Российской академии наук
Россия, Республика Башкортостан, 450054, г. Уфа, 450054, пр. Октября, 71
тел./факс: (347) 235-60-88, e-mail: chemeris@anrb.ru

Резюме

Сделаны предположения о тенденциях развития полногеномного и полнотранскриптомного секвенирования будущего и местах его проведения. Описано потенциальное влияние методов секвенирования нуклеиновых кислот (геномов и транскриптомов) 4-го и 5-го поколений на выявление полиморфных состояний геномов организмов разных уровней генетической сложности, включая человека. Дана характеристика различных типов полиморфизма ДНК человека и эволюция их исследований. Уделено некоторое внимание размерам геномов и так называемому парадоксу C-value. Сделан прогноз о количестве полностью секвенированных геномов и транскриптомов к 2030 году с распределением по годам. Изложен один из способов генотипирования штаммов бактерий с присвоением последним уникальных генетических штрих-кодов на основе метода КМРФ (концевого мечения рестриктазных фрагментов ДНК). Рассмотрен метод присвоения различным организмам ДНК-штрих-кодов на основе полиморфизма гена цитохром С оксидазы. Цитированная литература охватывает почти столетний период.

Ключевые слова: ДНК, РНК, белок, геном, секвенирование, ПЦР, гель-электрофорез, КМРФ, ДНК-чипы, олигонуклеотиды, геномика, транскриптомика, метиломика, GWAS, CNV, VNTR, STR, SNP, снипы

Если не грешить против разума
нельзя вообще ни к чему прийти

Альберт Эйнштейн

Оглавление

[34] Первая часть статьи опубликована в предыдущем номере – Биомика. 2013. Т.5, №1-2. С. 10-43.

Предисловие ко второй части статьи

В первой части данной статьи рассматривались преимущественно прошлое и настоящее физико-химической биологии, изложенные в перечисленных ниже главах и разделах – «Молекулярная биология и связанные с ней биологические дисциплины», «Прогнозы 1960, 1970, 2000 гг.», «1983 - . . . . - 2012 гг. (*1983 г., 1970-ые, 1960-ые гг.; От 1983 г. и до 2012 г.*)»; «2013 г. (*Секвенирование ДНК (полногеномное) новых поколений; Полные геномы свободноживущих организмов; Прочие технологии исследования нуклеиновых кислот*)». В послесловии к ней мы известили читателя, что готовящаяся вторая часть[35] будет посвящена, главным образом, технологическому будущему современной биологии. Вынуждены признаться, что не вполне справились со взятыми на себя обязательствами, в том смысле, что одной второй части оказалось маловато, чтобы изложить в ней все то, что планировалось. Или она оказалась бы чрезмерно большой. Поэтому за второй частью данной статьи в следующем номере данного журнала последует и третья, которая, надеемся, завершит описание нашего видения как близкого, так и относительно далекого будущего современной биологии. Причем для лучшего понимания тенденций развития физико-химической биологии и прогнозирования ее будущего неизбежно придется возвращаться в прошлое различных методических приемов и подходов. Хотим еще раз напомнить, что и дальше будем придерживаться взятого на себя обязательства не упоминать никаких производителей каких-либо приборов и оборудования.

2014 - . . . . - 2030 гг.

Итак, вперед, в будущее. Но прежде чем перейти к изложению нашего видения будущих технологических возможностей молекулярной биологии хотим заметить, что за очень редким исключением не хотим брать на себя смелость буквально по годам предположить, когда что произойдет, и поэтому здесь в заголовке указываем весь диапазон, в котором можно ждать в области молекулярной биологии чего угодно.

Полногеномное секвенирование будущего

Как читатель, вероятно, уже догадался, главный предмет нашего интереса в этой статье - это современные технологии обращения с нуклеиновыми кислотами и, в первую очередь, высокопроизводительное секвенирование геномной ДНК. На основании фрагментарного изложения в предыдущих разделах некоторых новых методов секвенирования ДНК второго и третьего поколений[36], надеемся, что создалось впечатление, что полногеномное секвенирование ДНК - это область поистине гигантских возможностей, раскрытых пока далеко не полностью. Несмотря на произошедший огромный прогресс в этой сфере за последнее десятилетие, все равно полногеномное секвенирование по нынешним воззрениям осуществляется довольно дорогостоящими, весьма трудоемкими и недостаточно производительными методами. Исходя из всего вышеизложенного, мы даже не будем пытаться прогнозировать, какой же конкретно метод секвенирования полных геномов четвертого, пятого поколения победит (будет лидировать) сначала в ближайшее, затем в отдаленное время, поскольку таких временных победителей наверняка будет немало (что уже наблюдалось в прошедшее десятилетие). Однако все же возьмем на себя смелость утверждать, что на будущее (к 2030 г. уж точно) более перспективными выглядят пока уступающие технологиям, рассчитанным на массивный параллелизм, методы мономолекулярного секвенирования. Причин тому несколько. К ним ниже и перейдем.

Так, методы массивного параллелизма требуют очень непростой пробоподготовки, занимающей довольно много времени и состоящей из многих стадий (фрагментация, фракционирование, пришивка адапторов, амплификация, обогащение и др.), к тому же есть риск сделать что-то не так на каждой из них. По сравнению с мономолекулярными, требуется большее количество исходной ДНК, что, например, при выполнении проектов/задач (число которых, вне всякого сомнения, будет расти) по секвенированию геномов древних вымерших организмов, а также геномов в разные времена почивших людей и их далеких предков, доступно бывает не всегда. Пожалуй, еще более важным моментом является то, что методы массивного параллелизма рассчитаны на амплификацию матриц для секвенирования с помощью какого-либо варианта ПЦР, что на самом деле не есть хорошо. Так, амплификация на стадии пробоподготовки может приводить к ошибкам при прочтении последовательности нуклеотидов из-за ошибок ДНК-полимеразы при синтезе. С другой стороны этап амплификации вообще может не привести к наработке нужной матрицы в силу того,

[35] Думалось, что будет завершающей.

[36] С нумерацией поколений методов секвенирования ДНК не все так просто, поскольку разные авторы относятся к этому вопросу по-разному и некоторые даже выделяют (под)поколения, например 2.5.

что в цепочке поврежденной временем (либо недавними физическими или химическими воздействиями) ДНК может быть место, лишенное азотистого основания, на котором ДНК-полимераза не просто «споткнется», а остановится, хотя в таких случаях уже предлагается использовать спектр репарирующих ферментов. Что касается секвенирования таких поврежденных молекул ДНК мономолекулярным методом даже с использованием для построения новой цепи ДНК-полимеразы, то данный фермент, пусть и не пройдет это место (без репарации), но даст возможность прочитать последовательность хотя бы до него. Причем, нельзя исключать, что в будущем может появиться производительный метод мономолекулярного секвенирования, в котором вовсе не будет этапа ферментативного построения новой цепи ДНК и уже есть статьи, посвященные использованию, например, разных микроскопов для определения последовательности оснований в единичных молекулах ДНК [Thomas, Glover, 2008; Tanaka, Kawai, 2009; Bell et al., 2012], не говоря уже о множестве работ по нанопорному секвенированию (имеются в виду варианты секвенирования цельных молекул ДНК), где ДНК не нужно подвергать никаким ферментативным воздействиям. При использовании таких методов, даже при наличии отсутствия отдельных азотистых оснований будут читаться участки, как до такого места, так и после.

Потенциально мономолекулярные методы могут позволить читать более протяженные участки до нескольких тысяч, десятков тысяч и даже сотен тысяч оснований, чего принципиально невозможно достичь методами массивного параллелизма. Еще одним важным моментом является относительно легкая масштабируемость таких методов, которая позволит тратить расходных материалов каждый раз столько, сколько нужно даже для выполнения небольшого проекта по секвенированию, даже, если сам прибор принципиально будет позволять «читать» многие терабазы.

Еще одним важным преимуществом некоторых мономолекулярных методов является возможность непосредственно выявлять модифицированные основания (например, метилцитозин), без каких-либо предварительных химических обработок молекул ДНК. Тогда как использование метабисульфитного полногеномного секвенирования с целью выявления метилированных цитозинов методами массивного параллелизма обладает рядом ограничений и требует специального программного обеспечения для сборки контигов [Kreck et al., 2012]. Так, после метабисульфитной обработки ДНК образуются последовательности фрагментов ДНК, принадлежащие как бы сразу двум геномам[37], причем с уменьшенным числом комбинаций нуклеотидов в них, и их сборка с помощью биоинформатики (о которой еще поговорим отдельно) представляется довольно проблематичной. Пояснить эту мысль можно следующим образом. Поскольку неметилированные цитозины в результате дезаминирования превращаются в урацилы, а в ходе ПЦР - в тимины, то заведомо произойдет уменьшение числа комбинаций перемежающихся нуклеотидов, которых в большей части ДНК (ввиду того, что метилированных цитозинов все же не так много в геноме) останется всего по три для каждой из комплементарных цепей (A, G и T / T, C и A), причем их соотношения будут соответственно близки к 1:1:2, что еще дополнительно снизит разнообразие и как следствие последующую сборку.

Выше мы уже отмечали возможность детекции метилированных производных азотистых оснований при секвенировании ДНК в реальном времени с «пришитой» ДНК-полимеразой [Flusberg et al., 2010; Fang et al., 2012]. При этом мы уверены, что в будущем появится возможность сразу безо всяких нежелательных химических обработок (которые всегда будут заставлять думать об их полноте, точнее сомневаться в ней) фактически «видеть» модифицированные азотистые основания. Например, в какой-либо микроскоп. Или каким иным способом. При наступлении такой возможности можно будет серьезно говорить об анализе метиломов, что сейчас носит лишь некий допустимый и фрагментарный характер, поскольку с помощью антител к метилированному цитозину возможно только направленное обогащение подлежащего секвенированию материала, содержащего так называемые «метилированные островки».

Мономолекулярные методы секвенирования можно еще подразделить на динамические и стационарные. С помощью первых информация о нуклеотидных последовательностях считывается по мере ферментативного роста цепи ДНК или при прохождении цельной (в том числе одноцепочечной) ДНК либо отдельных дМНФ через нанопоры и не может быть считана повторно для одного и того же образца. Стационарные же методы

[37] После метабисульфитной обработки молекул ДНК и превращения цитозинов в урацилы образуются некомплементарные друг другу цепи, фактически представляющие собой уже неродственные последовательности, образующие при амплификации с помощью ПЦР в общей сложности целых четыре цепи.

секвенирования (которые еще только разрабатываются, например, с помощью подходящих микроскопов), будут позволять в случае необходимости провести считывание последовательностей повторно с меньшей скоростью и с соответственно более высоким разрешением, повышая тем самым достоверность результатов.

Таким образом, широко используемыми в ближайшее десятилетие методами полногеномного мономолекулярного секвенирования ДНК станут те, что позволят без особых трудностей с пробоподготовкой секвенировать *de novo* геномы равные по размеру человеческому (и даже более крупные) в виде протяженных фрагментов ДНК (до 10 и более тысяч нуклеотидов, заметно снизив необходимое многократное покрытие геномов) с минимальными (для полногеномного секвенирования) ошибками при чтении каждой матрицы за относительно короткое время (например, за 8-ми часовой рабочий день), недорого (порядка 150-300 долларов на геном), обеспечивая не менее терабазы последовательностей, причем легко позволяя масштабировать процесс, вплоть до секвенирования всего тысяч нуклеотидов и не неся при этом лишних затрат. Сам прибор также должен быть не дороже 10, а лучше 5 млн. руб. (в России).

Другим очень важным моментом для массового использования нового удобного метода секвенирования ДНК четвертого/пятого поколений станет легкая возможность аналогичного секвенирования молекул РНК без их перевода в кДНК. В идеале практически не должна требоваться какая-либо специальная пробоподготовка, которая после выделения нуклеиновых кислот может заключаться лишь в определении тем или иным способом нужного количества РНК, которое необходимо брать в работу. Фактически должен секвенироваться АБСОЛЮТНО весь пул молекул РНК, присутствующих в клетке, точнее в группе одинаковых клеток одного типа ткани, поскольку все же для исключения ошибок (потерь некоторых последовательностей) будет необходимо 10-20-ти кратное покрытие транскриптома, учитывая, что РНК - более ломкие молекулы. Поэтому необходим будет мягкий лизис группы клеток одного типа и выделение всей РНК, включая рибосомную, матричную, транспортную и иные. Скорее всего, для этой цели для некоторых типов клеток будет требоваться использование лазерного микродиссектора с захватом.

Патенты (которые кем-то когда-нибудь будут получены) на такие удобные во всех отношениях методы секвенирования ДНК и РНК могут стоить миллиард долларов или даже больше, что будет зависеть от того, сколько сходных по производительности технологий нового секвенирования геномов и транскриптомов, способных секвенировать аналогично недорого и без чрезмерных усилий, к тому времени будут разработаны. При этом это вовсе не будет означать, что разработка все новых технологий полногеномного секвенирования тотчас же прекратится, а посему в один прекрасный день может появиться, если не более производительная, то, по крайней мере, более удобная и дешевая технология.

Что касается методов полногеномного секвенирования, рассчитанных на массивный параллелизм вкупе с амплификацией и доминирующих сейчас, нам представляется, что их время скоро уйдет. По крайней мере, чтобы им (нынешним) задержаться и составить конкуренцию мономолекулярным методам будущего, им «самим» нужно будет резко повысить свою чувствительность при детекции тех или иных сигналов, возникающих при секвенировании, приближаясь по этому показателю к мономолекулярным. При этом все же нельзя исключать появление некоего абсолютно нового, крайне удобного метода секвенирования ДНК с помощью массивного параллелизма, характеризующегося улучшенной и упрощенной пробоподготовкой, который опередит все прочие, включая мономолекулярные.

Хотим признаться, что на протяжении ряда лет нами, насколько это позволяют финансы и людские ресурсы, ведутся работы по разработке сразу двух новых оригинальных методов мономолекулярного секвенирования ДНК/РНК стационарного плана четвертого и пятого поколений, в которых не предусматривается использование ферментов. Разве что перед определением концентрации нуклеиновых кислот с помощью спектрофотометра будут задействованы ДНКаза или РНКаза, поскольку будет требоваться очистка препарата ДНК от РНК и наоборот. Собственно, это и будет почти вся пробоподготовка для второго подхода, который мы условно относим к пятому поколению методов секвенирования геномной ДНК. Для метода четвертого поколения, по которому у нас уже есть определенный прогресс, будут дополнительно требоваться также некие этапы сорбции секвенируемых нуклеиновых кислот на подходящем сорбенте и последующей элюции. Причем за последние годы появились статьи зарубежных авторов, посвященные некоторым другим вопросам анализа ДНК, но описанные в них методические приемы косвенно подтверждают, что обе наши технологии могут быть вполне работоспособны, высокопроизводительны и

недороги. Важной чертой данных технологий является то, что они легко масштабируемы и способны вести секвенирование (причем молекул и ДНК, и РНК) как сотен миллиардов нуклеотидов, так и просто всего сотен, неся пропорциональные затраты расходных материалов, что коренным образом отличает их от всех ныне существующих подходов. Длина чтения составит до десятков тысяч нуклеотидов, что в первую очередь будет зависеть от качества (размера) изначально выделенной ДНК и в настоящее время абсолютно недостижимо никакими ныне используемыми методами. Разрабатываемый нами метод секвенирования ДНК пятого поколения, помимо само собой разумеющейся возможности секвенирования РНК, после соответствующих адаптаций теоретически будет позволять определять последовательности аминокислот в белках и даже иных мономерных звеньев в прочих биополимерах.

В завершении рассмотрения возможностей будущих технологий секвенирования нуклеиновых кислот следует, пожалуй, немного порассуждать о некоем «потолке» производительности ДНК-секвенаторов, достижимом к 2030 г. или даже раньше. Сейчас за один запуск самого производительного ныне ДНК-секвенатора читается около 600 гигабаз оснований (6×10^{11}), что округленно можно принять как 10^{12} нуклеотидов, тем более, что скоро столько, наверное, уже и будет читаться зараз. Прогнозируемое нами в будущем получение информации о последовательностях 10^{15} нуклеотидов[38] будет означать, что в некоем устройстве должно проанализироваться около 1 микрограмма ДНК – а это довольно много. А для чтения 10^{18} азотистых оснований нужно прочитать уже целый миллиграмм ДНК. Очень много! Конечно, из живущих организмов можно выделить и большее количество ДНК - вопрос не в этом! Нужно ли столько секвенировать в одном эксперименте?! Ведь это количество ДНК сравнимо с весом геномов всех живущих ныне людей, если взять у каждого по одной копии. Скорее всего, технологии секвенирования ДНК, достигнув производительности порядка 10^{13} - 10^{14} нуклеотидов за один запуск прибора, далее наращивать свои мощности не будут, поскольку это и так будет означать тысячекратные покрытия генома, равного по размеру человеческому. Конечно, нам могут возразить, что фрагменты ДНК каждого генома можно будет по концам пометить некими тэгами[39] и потом при анализе всего прочтенного массива нуклеотидных последовательностей распределить их по их законным владельцам. Да, теоретически можно. Тогда, за один запуск ДНК-секвенатора, выдающего на-гора 10^{18} нуклеотидов, учитывая, что даже при ресеквенировании будет необходимо, например, 10-50-кратное покрытие генома, теоретически удастся прочитать геномы сразу нескольких миллионов человек. Только надо ли столько? Имеется в виду - в одном эксперименте, ведь такой вариант из-за многочисленных тэгов может заметно повысить стоимость секвенирования, сведя на нет весь эффект от столь увеличенной производительности.

Теперь о времени. Самое быстрое, но так пока и нереализованное нанопорное секвенирование ДНК обеспечивает прохождение через нанопору 1 млрд. оснований за 100 сек. Вряд ли какая новая технология секвенирования будет быстрее, тем более, что и эта-то скорость пока не подвластна для регистрации, поскольку на одно событие понуклеотидного прохождения нанопоры приходится всего 100 пикосекунд. Впрочем, нельзя исключать, что в будущем за этот промежуток времени соответствующими приборами будет успевать вестись регистрация возникающих в тот момент изменений ионных токов. Хотя сомнительно. Но, тем не менее, подсчитаем. Так, для прохождения через одну нанопору 10^{18} азотистых оснований потребуется 10^{11} сек или более трех тысяч лет. Конечно, нанопор может быть не одна, а 10^6, например. И тогда уже потребуется 10^5 сек или всего-то чуть больше суток. А для некоего кластера из 10^8 нанопор потребуется всего-навсего 1000 сек. Опять-таки требуется массивный параллелизм, только другого толка. Но как только долго потом компьютер будущего и какой всю эту прочитанную информацию будет анализировать и выстраивать правильную последовательность нуклеотидов!? Также надо еще учесть, что нет в Природе такой непрерывной ДНК из миллиардов нуклеотидов, не говоря уже о неизбежно разрушаемых при выделении образцах ДНК, и это тоже надо принимать во внимание и понимать, что промежуток времени между прохождением разноразмерных обрывков ДНК может быть куда продолжительнее самих моментов секвенирования.

Исходя из всего вышесказанного, можно прийти к заключению, что более важным является разработка секвенирования длинных фрагментов

[38] См. рис. 1 в первой части статьи.

[39] Такими тэгами могут быть синтетические олигонуклеотиды длиной, например, 30 звеньев, что обеспечит их абсолютную уникальность за счет более 10^{18} комбинаций, при этом должны быть исключены те варианты последовательностей, что встречаются в геноме человека или вообще в любых известных к тому времени геномах.

ДНК или РНК с очень высокой точностью, обеспечивающего определение 10^{12} - 10^{15} нуклеотидов за запуск прибора. И весьма важно, чтобы технология была недорогой, быстрой, легко масштабируемой, и тогда метод Сэнгера из арсенала молекулярно-биологических методов может уйти совсем, как это произошло с его серьезным конкурентом и даже лидировавшим в первые годы тогдашнего секвенирования - методом химической деградации ДНК по Максаму и Гильберту.

Перспективы и последствия развития полногеномного секвенирования

К чему же приведут в будущем новые высокопроизводительные технологии секвенирования ДНК четвертого и пятого поколений? Изменений на поле биологических наук произойдет немало. Первым делом перестанут использоваться разнообразные нынешние технологии выявления полиморфизма ДНК, основанные на амплификации отдельных участков генома с помощью ПЦР и иных способов, поскольку определять нуклеотидные последовательности сразу всего генома любых организмов даже *de novo* станет куда дешевле, проще, удобнее и быстрее. Не говоря уже о ресеквенировании. Полиморфизм ДНК будет выявляться с помощью биоинформатики путем сравнения геномов целиком или соответствующих участков геномных последовательностей, где с точностью до нуклеотида могут/будут обнаруживаться различия между особями, штаммами, видами, родами, представителями других таксономических единиц. При этом исследования полиморфизма ДНК будут фактически заменены на сравнительную геномику исследуемых организмов, которая оперирует последовательностями не отдельных генов и/или прочих участков ДНК, а сопоставляет нуклеотидные последовательности всего генома, либо его фрагментов, включая в анализ также различные геномные перестройки и синтению генов в виде структурного сходства групп сцепления генов у организмов разных биологических видов.

Прежде чем продолжить рассмотрение перспектив полногеномного секвенирования необходимо немного задержаться на собственно полиморфизме ДНК, тем более, что дальше хронологическое описание его выявления будет связано с разными технологиями. Так, под полиморфизмом ДНК понимают некоторые отличия нуклеотидных последовательностей в сходных участках геномов у родственных организмов[40]. Практически весь полиморфизм ДНК или геномный полиморфизм за небольшими исключениями можно уложить в два основных типа: замены одиночных нуклеотидов и инсерции/делеции (или так называемые «инделы») или также одиночных нуклеотидов или их блоков весьма разной протяженности. К первому[41] типу относятся однонуклеотидные замены, или снипы (от Single-Nucleotide Polymorphism - SNP), число которых, например, в геномах людей может достигать нескольких миллионов, приходясь в среднем по одному SNP на каждые 300 пар нуклеотидов. Таким образом, на долю однонуклеотидных замен приходится около 0,3% (всего-то) геномных различий у человека.[42] К тому же снипы в большинстве своем представляют собой самую малозначимую часть генома[43], поскольку относительно легко подвержены мутациям, что свидетельствует о том, что они не находятся под строгим давлением отбора, как бы это имело место, будь эти места генома крайне важны для нормального функционирования. В статье с броским названием «The birth and death of human single-nucleotide polymorphisms: …» известных американских специалистов по снипам [Miller, Kwok, 2001] отмечается, что жизненный цикл снипов у человека делится на 4 фазы: (i) появление новых аллельных вариантов за счет мутаций; (ii) выживание образовавшихся вариантов аллелей за счет ранних генераций; (iii) значительное увеличение частот встречаемости с учетом популяционных флуктуаций; (iv) фиксация снипов. В своей предыдущей статье эти же авторы сравнили спектр снипов у человека и орангутанга и пришли к выводу, что имеющиеся у этих видов однонуклеотидные замены имеют независимое происхождение, свидетельствующее, что время жизни снипов, определенное ими как 284 тысячи лет, короче, чем время дивергенции данных приматов [Miller et al., 2001].

Подавляющее количество снипов биаллельны, но есть и триаллельные, и тетрааллельные снипы. Под биаллельностью снипов понимается, что в одних и тех же характеризующихся некоторым полиморфизмом

---

[40] Причины и механизмы возникновения различий между родственными геномами могут быть довольно различными и выходят за рамки рассматриваемого нами здесь материала.

[41] Первому не по хронологии обнаружения и не по значимости, а просто в порядке упоминания в предыдущем предложении.

[42] Ниже к этой цифре мы еще не раз вернемся.

[43] Здесь мы не принимаем во внимание различные типы сателлитных повторов, функция которых пока все же доподлинно не известна.

местах геномов людей разным индивидам присущи только два каких-либо нуклеотида из четырех возможных, для триаллельных снипов таковых будет уже три, но при этом у конкретного человека их будет также всего два из трех возможных[44]. Также не более двух у одного человека, но у разных людей все четыре нуклеотида могут встречаться в тетрааллельных снипах. Благодаря довольно большой вариабельности снипов, их этнические особенности выражены не столь ярко, и потому они могут служить удобными маркерными характеристиками отдельных особей у человека, формируя некий персональный набор (код), способный служить для ДНК-идентификации личности, что будет изложено в третьей части данной статьи.

Касательно преимущественной биаллельности снипов надо заметить следующее. Так, в первых экспериментах на примере нескольких десятков или даже сотен человек для какого-нибудь снипа выявляется его биаллельность и обнаруживается, что в нем в исследованных образцах встречаются только два нуклеотида, например G и T. Информация об этом помещается в базу данных по снипам и далее все остальные, кто продолжает исследовать другие популяции на предмет анализа статуса этого снипа в них, используют соответствующие праймеры или применяют иные методические подходы, либо пользуются коммерческими наборами для анализа данного снипа (если таковые выпускаются), дискриминирующими именно эти нуклеотиды (в нашем примере G и T). По крайней мере, ныне так поступает большинство экспериментаторов. При этом, если в данном снипе, которому, как правило, присущи нуклеотиды G и T, у какого-либо индивида будут находиться нуклеотиды G и A, то последний нуклеотид у него просто не будет обнаруживаться, и человек может быть признан гомозиготным по данному снипу, поскольку будет считаться, что у него в обоих аллельных вариантах присутствует гуанин[45], что на самом деле не будет соответствовать действительности. Вероятность таких событий исключать нельзя, и уже появились статьи [Huebner et al., 2007; Morita et al., 2007; Westen et al., 2009], посвященные этому вопросу, где как раз продемонстрирована подобная ситуация с получением ложных сведений о представленности нуклеотидов в некоторых снипах. Таким образом, информация о «нынешней» биаллельности или даже триаллельности любых снипов не должна восприниматься как абсолютно верная и потому каждый снип для любого генетического анализа должен считаться потенциально тетрааллельным тетрааллельным, и соответственно используемые подходы должны обеспечивать выявление всей четверки нуклеотидов. Что, однако, сразу удорожает такой анализ и является серьезным стимулирующим моментом для выявления именно биаллельности снипов. Собственно, после полного перехода на новое полногеномное секвенирование данная проблема отпадет сама собой, поскольку при изучении структурной сравнительной геномики с помощью биоинформатики будут однозначно выявляться конкретные нуклеотиды сразу во всех снипах любых индивидов, равно как и все другие полиморфные участки их геномов, к рассмотрению которых переходим.

* * *

Полиморфизм ДНК в виде инсерций/делеций проявляется в различиях между геномами по числу базовых элементов повторов в микро- и минисателлитных последовательностях (STR - Short Tandem Repeats и VNTR - Variable Number of Tandem Repeats, соответственно)[46]; по варьирующему числу копий каких-либо протяженных последовательностей (CNV - Copy Number Variation) или непосредственно самих коротких делеций и инсерций, не несущих в себе повторяющихся участков и могущих быть представленными в виде единичных нуклеотидов. Надо сказать, что границы принадлежности того или иного варианта последовательности к соответствующему типу инделов довольно условны.

---

[44] При этом подразумевается, что данный конкретный снип встречается в геноме единожды. Т.е. не сами вариабельные нуклеотиды A/C, G/A, C/T или другие их комбинации, а фланкирующие данный снип последовательности, протяженность которых берется таковой, что должна обеспечивать уникальность такого участка в геноме, но при этом быть относительно небольшой (порядка 25 нуклеотидов в обе стороны).

[45] Количественное определение в таких случаях одного или двух G, позволившее бы получить уточненный результат, требует использования более сложных методических приемов, которые обычно не применяются при таких исследованиях.

[46] Во избежание недоразумений необходимо указать, что различия по числу повторяющихся, например, тетрануклеотидов CTGG или аналогичных повторов в микросателлитах либо более протяженных в минисателлитах в геномах разных людей принимаются нами здесь как соответственно делеции у одного индивида или инсерции у другого, что не является общепринятым подходом к такой классификации повторяющейся ДНК, но формально имеет право на существование.

Так, принято считать, что повторяющиеся мотивы от 2 до 6 пар нуклеотидов характерны для микросателлитной ДНК, тогда как повторяющиеся единицы длиной от 7 до 100 пар нуклеотидов[47] соответствуют уже минисателлитам. Принадлежность к инделам или к CNV условно установлена на границе одной тысячи пар нуклеотидов; ниже - простые (короткие) инделы, выше - соответственно CNV. Инделы из единичных нуклеотидов формально могут быть отнесены даже к снипам, поскольку их можно отображать, например как A/0, G/0 или A/-, C/-, где «0» или «-» обозначают отсутствие соответствующего нуклеотида. При этом доля инсерционно-делеционных вариаций между геномами разных людей без учета вклада мини- и микросателлитов и снипов может достигать 12%.

Так называемые Alu-повторы, имеющие размер около 300 пар нуклеотидов, диспергированы по всему геному в виде более чем миллиона копий, что составляет свыше 10% всей последовательности ДНК человека, при этом каким-либо значительным структурным полиморфизмом они не отличаются. К тому же Alu-повторы типичны только для человека и приматов, тогда как остальные типы полиморфизмов в той или иной степени присущи всем высшим организмам.

Так называемый ПДРФ или RFLP (Полиморфизм Длины Рестрикционных Фрагментов или Restriction Fragments Length Polymorphism) в разных случаях может принадлежать как к однонуклеотидному полиморфизму, так и к инсерционно-делеционному полиморфизму в зависимости от того, что лежит в основе различий в длине выявляемых, как правило, блот-гибридизацией рестриктазных фрагментов ДНК - замена нуклеотида, нарушающая сайт узнавания фермента, или разное число копий некоего повторяющегося элемента, либо наличие инделов, локализованных внутри участка ДНК, ограниченного сайтами узнавания используемой рестрикционной эндонуклеазы.

Есть, по крайней мере, еще два типа полиморфизма геномов - инверсии и транслокации, причем если все предыдущие межгеномные различия можно обнаруживать, помимо самого секвенирования ДНК, довольно разными методами, включая высокопроизводительные ДНК-чиповые технологии, то массовое выявление инверсий и транслокаций в силу методических особенностей, возможно только с помощью полногеномного ресеквенирования.

Если проследить развитие и эволюцию исследований полиморфизма ДНК/генома человека за последние три с лишним десятилетия, то ясно видны пять этапов[48], и в каждом из них доминирующим был свой отдельный тип полиморфизма. Причем изучение того или иного (фактически любого) вида полиморфизма было напрямую связано с технологическими возможностями своего времени, сейчас кажущимися для первых типов очень скромными. Поскольку делать что-то было надо, то и делали, что получалось и давало в то время приемлемый результат[49]. Также на 5 классов основных полиморфизмов - RFLP, VNTR, STR, SNP, CNV указывает Й.Накамура в обзорной статье, посвященной своему 25-летнему опыту работы в этой области [Nakamura, 2009]. При этом он выделяет 6 поколений типов генетических вариаций с указанием годов (1980, 1985, 1989, 2000, 2005, 2012 гг.), когда каждый тип соответственно или был обнаружен или стал широко использоваться. Из них пять поколений совпадают с вышеуказанными типами полиморфизмов, а в качестве шестого поколения Й.Накамура в 2009 г. предположил (поставив знак вопроса), что с 2012 г. начнется полногеномное секвенирование, которое, надо признать, все же пока задерживается (имеется в виду, носящее действительно массовый характер и по-настоящему высокопроизводительное).

Так, начиная со второй половины 1970-х гг., в молекулярной биологии и молекулярной генетике широко использовался метод блот-гибридизации по Саузерну [Southern, 1975; 2000], с помощью которого

[47] Приводимые размеры повторяющихся единиц, разграничивающие микро- и минисателлитные повторы, довольно условны, в разных статьях и обзорах они могут несколько варьировать.

[48] Небольшое обсуждение временнЫх отрезков этих этапов, в целом совпадающих с упоминаемыми Й.Накамурой, будет проведено ниже.

[49] При этом, конечно же, нельзя «сидеть, сложа руки» и вечно ждать появления каких-то новых методов, технологий, чтобы начать с их помощью получать уже более точные и конкретные результаты, ввиду того, что и новые методы и новые технологии рождаются, как правило, на базе используемых старых, причем отталкиваясь от полученных именно этими методами результатов. Иначе движения вперед быть просто не может, а потому даже к устаревшим результатам надо относиться с уважением. Но все равно еще раз повторим мысль из первой части данной статьи, что надо стараться предвидеть в каком направлении будет идти прогресс в той или иной области науки, в том числе или даже в особенности в молекулярной биологии.

можно было определять ПДРФ разных локусов, в том числе для ДНК человека. Вполне логичным было появление работ, посвященных построению генетических карт сцепления с ПДРФ-маркерами [Botstein et al., 1980 и др.]. Надо сказать, что, несмотря на относительно небольшое число (несколько сотен) выявленных таким образом полиморфных локусов, все же для части из них, благодаря семейному анализу, удалось выявить сцепленность с десятком тяжелых генетических заболеваний, среди которых стоит упомянуть мышечную дистрофию Дюшена, болезнь Хантингтона и др. Однако в целом довольно низкая информативность ПДРФ-анализа, а также обнаружение высоковариабельных повторов ДНК - минисателлитов или VNTR [Jefreys et al., 1985] с копийностью в геноме порядка нескольких тысяч заставили исследователей переключиться на этот тип полиморфизма для поиска новых групп сцепления [Donis-Keller et al., 1987; Nakamura et al., 1987 и др.], хотя технология его выявления была не менее трудоемка, так как требовала использования той же блот-гибридизации. Обнаружение еще более полиморфной и высокоповторенной (более 100 тысяч копий в геноме) микросателлитной ДНК - STR [Weber, May, 1988; 1989; Litt, Luty, 1989; Weber, 1990], которую оказалось возможным изучать с помощью появившейся ПЦР, сразу значительно облегчило проведение экспериментов и резко повысило общую производительность исследований, направленных на выявление маркеров, сцепленных с различными заболеваниями. В результате «эксплуатации» STR-локусов оказались построенными множество «разномасштабных» карт сцепления [Weissenbach et al., 1992; Weissenbach, 1993; Dib et al., 1996 и др.], позволивших выявить определенные связи с некоторыми болезнями.

Фактически три первых этапа изучения полиморфизма ДНК пришлись на догеномную эру, когда информация о нуклеотидных последовательностях генов человека носила большей частью отрывочной характер. При этом необходимо отметить, что те исследования основывались на анализе сцепленного наследования в семьях. С завершением секвенирования генома человека стало очевидным, что имеется еще и такой тип полиморфизма, как однонуклеотидные замены[50], носящие еще более массовый характер, достигая миллионов отличий между геномами людей, не состоящих в близком кровном родстве. Причем, каждый раз переход на исследования нового типа полиморфизма происходил достаточно быстро, поэтому возможно указать довольно четкие границы начала и завершения этапов, что фактически сделал в своей статье Й.Накамура [Nakamura, 2009]. Однако ситуация с четвертым и пятым этапами несколько сложнее. Так, после обращения внимания части ученых на новый тип полиморфизма в виде вариаций числа копий, массового перехода на него не произошло, как это имело место ранее на предыдущих этапах, и снипы до сих пор остались в качестве доминирующего типа полиморфизма, отчасти по причине удобства их анализа и одновременно некоторых трудностей с выявлением новых типов полиморфизма. Скорее всего, только после широкомасштабного внедрения в повседневную практику ресеквенирования геномов людей и состоится повсеместный переход на CNV тип полиморфизма, к рассмотрению которого перейдем позже. А пока - снипы.

* * *

В настоящее время на основе снипов проводится поиск ассоциаций с поведением индивидов, отклоняющимся от нормы, и различными заболеваниями многофакторной природы с тем, чтобы выявить некие генетические факторы риска с целью дать обоснованный прогноз предрасположенности их носителей к таким состояниям. Причем, такой подход обычно не выявляет мутации, становящиеся причинами болезней и других нарушений, а лишь указывает на более/менее значимую корреляцию с ними. Допускается, что генетические отличия какой-либо особи, вызывающие болезнь или нестандартное поведение, располагается в некоем другом месте генома. Возможно, все же недалеко от найденного прогностического снипа. Правда, иногда в статьях говорится о генах-кандидатах, видимо, все же кодирующих ферменты или прочие белки, напрямую связанные с нарушенной жизнедеятельностью организма человека. В этой связи можно привести название одной статьи («Why most discovered true associations are inflated»), где автор [Ioannidis, 2008] приводит несколько причин того, что многие ассоциации снипов с чем-либо оказываются «дутыми». Так, по его мнению, часто бывает слаба теоретическая проработка; возникает большой разброс получаемых данных; имеет место селективный подход к анализу данных; наконец, возможен некий конфликт интересов.

Выявление снипов возможно с помощью довольно большого числа разнообразных методов, и как следствие стали появляться многочисленные

[50] Выявление замен нуклеотидов, расположенных в сайтах узнавания используемых при выявлении ПДРФ рестрикционных эндонуклеаз, имело по сравнению с полногеномным секвенированием столь небольшой масштаб, что их можно в расчет не принимать.

работы, посвященные анализу снипов в связи заболеваниями многофакторной природы у человека [Gray et al., 2000; Ohnishi et al., 2001; The International SNP Map Working Group, 2001; Haga et al., 2002, Ozaki et al., 2002 и др.]. Благодаря тому, что были разработаны специализированные ДНК-чипы и среди них особо выделяются высокопроизводительные микроэррейные технологии, за эти годы произошел большой прогресс в изучении однонуклеотидного полиморфизма, достигший так называемого всегеномного[51] уровня исследования ассоциаций (GWAS - Genome Wide Association Study[52]), где с помощью разных платформ сразу анализируются огромное множество снипов, в последнее время практически удваивающееся каждый год и относительно недавно достигшее миллиона. Если на предыдущих этапах изучения полиморфизма ДНК анализировали преимущественно лишь сотни и тысячи образцов ДНК людей, главным образом из-за того, что для ПДРФ, VNTR и STR анализов необходимо электрофоретическое разделение продуктов реакций, которое довольно плохо поддается масштабированию, то выявление снипов проводится на десятках и сотнях уже тысяч людей. Неудивительно, что, начиная с единичных публикаций по GWAS в 2005 - 2007 гг., их число за этот период выросло весьма значительно, превысив отметку в тысячу статей в квартал в середине 2011 г. По состоянию на начало ноября 2013 г. в PubMed хранится информация о более чем 15 тысячах статей, где в тексте встречается аббревиатура GWAS, еще около 600 работ и около 300 публикаций с аббревиатурами GWA и GWASs. Итого, 16 тысяч экспериментальных, обзорных, проблемных и редакторских статей с упоминанием о всегеномном анализе ассоциаций за 7 лет. Однако столь же резкого увеличения достоверной информации о связях различных болезней с определенными снипами, к сожалению, не произошло, и скорее всего, ввиду довольно малой прогностической значимости последних, что отмечается во множестве статей [Pearson, Manolio, 2008; Zeller et al., 2012 и др.], к вопросу о чем ниже еще вернемся. Причем, при проведении GWAS как раз вполне уместны и даже настоятельно рекомендуются так называемые репликативные исследования, выполняемые с другой популяцией и рассчитанные на подтверждение первичных результатов. Есть даже предложения принять репликативные исследования в качестве золотого стандарта для подтверждения GWAS-данных [Liu et al., 2008; Huang et al., 2011], при этом немало сообщений о неподтверждении первичных результатов [Paterson, Cardon, 2005; Morgan et al., 2007; Pearson, Manolio, 2008 и др.]. В целом подход к изучению нездоровых состояний организма человека через всегеномное изучение снипов подвергается частой критике и является предметом споров, к вопросу о чем также еще вернемся.

Такое массовое увлечение исследованием снипов объясняется наличием высокопроизводительных технологий, хотя и требующих, с одной стороны, дорогостоящего оборудования и недешевых расходных материалов, но, с другой стороны, превращающих такой анализ в довольно рутинную процедуру, выдающую в итоге огромный объем информации, которую еще надо, впрочем, правильно переработать, поскольку в литературе встречается немало предупреждений о возможных ложных выводах по результатам таких экспериментов [Sun et al., 2006; Todd, 2006; Forner et al., 2008; Ioannidis, 2008; Manolio, 2010; Hakonarson, Grant, 2011; Huang et al., 2011], а также требований усилить контроль качества проводимых процедур [Turner et al., 2011]. Китайские ученые в одной своей недавней статье честно признают, что проводить исследования по GWAS с некоторых пор стало в их стране именно модным [Du et al, 2012]. При этом они вполне самокритично замечают, что статьи по

---

[51] К сожалению, в русскоязычной литературе довольно часто встречается выражение «полногеномный анализ ассоциаций», тогда как Genome-Wide Association Study переводится точнее как «всегеномный» (также иногда использующийся в России вариант обозначения GWAS) или «изучение ассоциаций, охватывающих весь геном» (что, впрочем, слишком длинно), неся принципиальное отличие от значения-термина «полногеномный», подразумевающего анализ действительно всего полного генома, для которого в англоязычной литературе используются определения full, whole, entire, но не wide.

[52] Впервые словосочетание genome-wide association test появилось в проблемной статье американских авторов [Risch, Merikangas, 1996], где ими было предположено, что в будущем выяснение причин возникновения мультифакториальных заболеваний будет вестись не семейным анализом групп сцепления, а поиском ассоциаций диаллельных маркеров. Надо отметить, что их статья вызвала широкий резонанс и на нее последовали многочисленные отклики не во всем с ними согласных генетиков [Scott et al., 1997; Bell, Taylor, 1997 и др.]. При ответе на их отклики [Risch, Merikangas, 1997] использовали уже другой вариант описания процесса - genome-wide association study, но аббревиатуры в виде GWAS тогда не последовало.

GWAS гораздо легче опубликовать в журналах с высоким импакт-фактором, и делают вывод, что такие исследования несут, в первую очередь, публикационную направленность, а отнюдь не ориентированы на потенциальное выяснение/улучшение здоровья людей, хотя они сами же проводят подобные анализы.

Для проведения большинства GWAS исследований одним из наиболее трудоемких и продолжительных является этап создания выборки опытных и контрольных образцов, причем, если изучаются особенности некоего поведенческого характера, то при формировании групп и подгрупп, на наш взгляд, следует брать в расчет сразу очень много объективных и субъективных обстоятельств, в том числе, образ жизни испытуемых, их социальный статус, семейный уклад (при наличии), привычки и пр.. Впрочем, и при исследовании различных болезней, например, рака яичников или молочной железы для того, чтобы получить значимый результат обе группы (больных и здоровых) надо подбирать не только по возрасту и этническому признаку, а в первую очередь, учитывать многие другие более важные параметры: гормональный баланс испытуемых, употребление ими гормональных препаратов, в том числе противозачаточных, количество рожденных детей и в каком возрасте, число сделанных абортов, опять же особенности образа жизни и прочая, прочая. Иначе, достоверного результата принципиально не достичь,[53] конечно, если действительно ставить перед собой такую цель.

Возможно, не было серьезных проблем у авторов при наборе исследуемых групп при GWAS исследовании курильщиков разного пола, поскольку, к сожалению, и тех и других пока еще немало, и потому для потенциального улучшения состояния их здоровья (наверное, все же в неблизком будущем) в разных странах (кроме, к сожалению, России) уже проведено большое количество исследований [Caporaso et al., 2009; Drgon et al., 2009; 2009a; Furberg et al., 2010; Rose et al., 2010; David et al., 2012; Hubacek et al., 2012 Yoon et al., 2012; Loukola et al., 2013 и др.]. Но поскольку, в отличие от других стран, в России до сих пор производятся папиросы то, наверное, было бы весьма интересно провести комплексное генетическое исследование ассоциаций снипов с курением как обычных сигарет, так и папирос[54], чего другим странам сделать затруднительно или даже невозможно.

Надо заметить, что определенное облегчение в вопросе сбора образцов наблюдается в последние годы, благодаря появлению так называемых БиоБанков, где уже хранятся образцы ДНК большого числа людей, как здоровых, так и имеющих различные патологии, или ведущих отличный от неких условных стандартов образ жизни [Harris et al., 2012; Sandor et al., 2012; Vaught, Lockart., 2012; Vaught et al., 2012; De Souza, Greenspan, 2013].

Выше уже упоминалось, что на долю однонуклеотидных замен приходится всего около 0,3% геномных различий у человека. А если учесть, что только очень малая часть из них приходится на регуляторные или кодирующие участки[55], причем в

[53] В случае невозможности обеспечить репрезентативные выборки по опытным и контрольным группам с учетом многих важных факторов, вероятно, не следует раньше времени начинать исследования по установлению каких-либо предрасположенностей, поскольку заведомо их результаты не смогут претендовать на какую-либо полноту и тем более научную новизну и практическую значимость.

[54] В России до сих пор курят еще и самокрутки или цигарки из самосада, махорку, а также трубки и даже сигары, но, чтобы основательно охватить и любителей последних, будет требоваться уже международное исследование, а вовлечение в анализ также приверженцев нюхательного табака и электронных сигарет тянет уже на создание соответствующего Консорциума. (Хотя одно научное сообщество, несколько лет успешно проводящее такие исследования, уже есть - The Tobacco and Genetics Consortium). Еще обязательно надо учитывать что именно, сколько и как курят люди: «легкие» понемногу или крепкие пачками, без фильтра, с мундштуком, натощак... О-о-о! Тут есть огромное поле для исследований ассоциаций снипов с социально-значимым поведением, а также с вредом для здоровья в связи с курением у немалой части общества. Не забыть бы еще этническую принадлежность! И в любых таких анализах, вне всякого сомнения, будут получены интересные результаты с массой прогностических снипов в некодирующих областях генома (см. ниже сноску 55), которые проявят некую ассоциацию различных сторон курения с анонимными, но действительно ответственными за различные аспекты этого процесса нуклеотидными последовательностями (генами) в прилегающих к таким маркерным снипам сцеплено наследуемым участкам генома. Только вот последние как бы найти?! Да и сможет ли это хоть как-то повлиять на здоровье активных и пассивных курильщиков?!

[55] Проведенный американскими авторами анализ 151 GWAS исследования показал, что в этих

последних еще должна произойти замена какой-либо аминокислоты, что из-за вырожденности генетического кода имеет вероятность всего около 60%[56] [57], то реальные различия, учитывая все ограничения, которые могли бы повлиять на фенотип, скорее всего, не превысят и 0,01%, а то и вообще 0,001%. И это что называется в лучшем случае! Но эти выкладки проведены для 10 миллионов снипов[58], тогда как в настоящее время в GWAS анализ таковых берется по максимуму лишь около миллиона, а то и того меньше. Следовательно, значимые различия между геномами здоровых и больных, которые могут быть сейчас выявленными современными технологиями, составят всего 0,0001%. Очень мало. Образно говоря - «пальцем в небо». В этой связи можно или даже нужно задаться вопросом - так ли важно изучать однонуклеотидный полиморфизм человека, ожидая, что обнаруженные замены как-то связаны с его болезнями и расположены в генах-кандидатах? Тем более, что большинство обнаруживаемых в связи с разными болезнями снипов локализованы в некодирующих областях, часто вообще далеко отстоящих от каких-либо генов, которые можно было бы считать генами-кандидатами [Pearson, Manolio, 2008]. К тому же, как отмечалось выше, снипы по большей части представляют собой самую малозначимую часть генома. Особенно те, что расположены в интронах. Разве что с очень большой натяжкой можно рассматривать такие снипы как маркеры ассоциаций непонятно каких и неясно где расположенных генов, отвечающих за возникновение той или иной болезни. Но тогда, пользуясь тем, что весь геном человека уже более-менее проанализирован, необходимо, как нам представляется, взять в рассмотрение сцепленно наследуемые фланкирующие данный «прогностический» снип участки генома на предмет присутствия в них конкретных генов, хоть как-то потенциально способных быть ответственными за развитие изучаемой болезни и их детально и всесторонне исследовать, включая белковые продукты. От такого подхода действительно могла бы быть определенная польза, хотя надо признать, что данная работа весьма трудоемка, а результат тоже далеко неоднозначен. И надо ли этим сейчас заниматься? Или все же ждать появления новых технологий, включая высокоэффективное секвенирование геномов и транскриптомов? Ответ попытаемся дать в последующих абзацах.

Если пытаться рассуждать логично, учитывая низкую вероятность изменения функционального состояния гена и тем более белка/фермента при замене одного нуклеотида, то становится весьма сомнительной роль таких снипов в проявлении болезненного состояния пациентов. Если опять-таки такой снип просто маркер, то что же реально изменилось где-то там, что он маркирует? Произошла делеция, или иной тип мутации? Так надо их и искать! Причем скорее это можно будет сделать через выяснение функционального состояния молекул белковой природы. Поэтому от такого подхода в виде широкомасштабного анализа снипов в век полногеномного (ре)секвенирования генома, а затем и транскриптома с метиломом любого человека надо будет уходить, поскольку для персонифицированной медицины будущего будет стоять задача не поиска неких вероятностных ассоциаций, а необходимы будут четкие знания взаимосвязи конкретного(ых) локуса(ов), точнее даже взаимосвязи их структурно-функциональных характеристик с тем или иным заболеванием, с целью дальнейшего выявления таких НАСТОЯЩИХ генетических особенностей потенциально у каждого индивида. Причем, высказываются предположения, что медицина будущего это будет так называемая медицина P4 (predictive, preventive, personalized, participatory[59]), которая, в том числе, должна знать детальные механизмы возникновения болезней [Hood, Flores, 2012].

В качестве некоего примера обнаружения множественных ассоциаций разных снипов одного гена с весьма различными болезнями и прочими состояниями организма, вызывающими многочисленные вопросы и порождающие

работах 88% однонуклеотидных замен пришлись на некодирующие участки, из которых доля интронов составила 45%, а межгенных участков - 43% [Hindorff et al., 2009].

[56] Без учета возможного образования терминирующих кодонов.

[57] Да, чтобы еще и эффект был от такой замены - они не должны быть синонимичными а место расположения данной вариабельной аминокислоты должно располагаться в структурно-важном/активном месте/центре белка/фермента.

[58] Количество снипов в виде 10 миллионов взято как некое круглое значение, облегчающее ведение подсчетов, хотя в целом для геномов людей известно уже свыше 15 миллионов снипов.

[59] При этом «participatory» предполагает по сути лечение объединенными усилиями, где пациент также активный участник процесса - многое знает, его и дальше информируют, помогают с выбором лечения, к чему сейчас наш медперсонал абсолютно не готов, как, впрочем, и большинство больных.

множественные сомнения, можно рассмотреть так называемый ген FTO (fat mass and obesity associated), отвечающий за ожирение организма. Уже найдено немало ассоциаций разной степени достоверности снипов данного гена с тучностью хозяев - носителей определенных его вариантов, как взрослых, так и детей, в том числе, здоровых и больных астмой [Frayling et al., 2007; Frayling, Ong, 2011; Tung, Yeo, 2011; Ibba et al., 2013; Melen et al., 2013; Peters et al., 2013 и др.]. На самом деле ген FTO представляет собой 2-оксоглутарат N-метил деметилазу нуклеиновых кислот [Gerken et al., 2007], а физиологическая роль катализируемого этим белком ферментативного процесса в ожирении и прочих бедах организма до сих пор остается невыясненной [Berulava et al., 2013; Muller et al., 2013]. Тем не менее, удалось установить, что 2-оксоглутарат N-метил деметилаза нуклеиновых кислот в качестве субстрата предпочитает использовать 3-метилтимидин в одноцепочечной ДНК и 3-метилурацил в одноцепочечных участках РНК [Han et al., 2010]. На мышах и культуре клеток с увеличенной и уменьшенной экспрессией гена FTO показано, что 2-оксоглутарат N-метил деметилаза нуклеиновых кислот на транскрипционном уровне изменяет соотношение 3-метилуридин/уридин, что, собственно, и должна делать деметилаза. Недавно предпринята попытка обнаружить ингибиторы этого фермента [Aik et al., 2013], однако главного ответа на законный вопрос, как именно осуществляется обнаруженная связь 2-оксоглутарат N-метил деметилазы нуклеиновых кислот с тучностью организма человека как не было, так и нет. Однако ряд ученых из разных стран обнаружили взаимосвязь некоторых снипов FTO гена еще и с другими болезнями и прочими отклоняющимися от нормы состояниями организма. Так, обнаружены ассоциации FTO гена с риском развития меланомы [Iles et al., 2013], с риском острой коронарной недостаточности [Hubacek et al., 2010], причем позже эти же авторы с успехом решили проследить связь гена FTO с курением [Hubacek et al., 2012]. Ген FTO показал также связь с дефицитом внимания, сопровождающимся гиперактивным поведением [Choudhry et al., 2013], причем последнее для полных людей в целом надо признать нехарактерно. Потенциальным геном-кандидатом, ответственным за риск развития поликистоза яичников - комплексного заболевания эндокринной природы, китайские авторы признали тоже ген FTO [Li et al., 2013]. Удивительно ключевой ген, задействованный в стольких сторонах жизнедеятельности человеческого организма! При этом интересно и, пожалуй, даже необходимо отметить, что при поиске в PubMed статьи с одновременным наличием аббревиатур FTO и GWAS/GWA/GWASs[60] в тексте встречаются около 250 раз (из них чаще всего - более 200 раз встречается FTO и GWAS), но стоит в поисковую строку добавить еще слово «demethylase» как уже не находится ни одной работы, где бы данное слово и вышеупомянутые аббревиатуры вместе встречались. Это означает одно - тем, кто исследует всевозможные всегеномные ассоциации гена FTO, совершенно нет дела до ферментативного действия кодируемого этим геном белка, несмотря на то, что в некоторых статьях говориться о нем как о гене-кандидате.

Напомним, что делает фермент 2-оксоглутарат N-метил деметилаза нуклеиновых кислот, кодируемая геном FTO - деметилирует некоторые субстраты в одноцепочечных молекулах ДНК и РНК. Какова функциональная связь этих процессов со всеми найденными с помощью GWAS анализов с этим геном ассоциациями с очень разными болезнями и иными состояниями организма - еще выяснять и выяснять. А лучше не тратить ни времени, ни денег и не выяснять вообще, поскольку все эти обнаруженные ассоциации носят, скорее всего, случайный характер. Тем более, что значительная часть таких вариабельных снипов в гене FTO, с которыми найдены те или иные ассоциации, приходится на интронные области. И тогда надо анализировать сцепленно-наследуемые участки генома на предмет выявления в них генов, обеспечивающих возникновение истинных причин тех или иных болезненных и прочих состояний человека. Подобных удивительных примеров по результатам GWAS анализов с другими генами-кандидатами и снипами в таковых можно привести немало.

Имеется целый ряд статей, посвященных отрицательным сторонам GWAS-технологии, причем опубликованы они в высокорейтинговых генетических журналах [Pearson, Manolio, 2008; Gibson, 2012; Visscher et al., 2012; Manolio, 2013 и др.]. Предлагаем желающим самостоятельно ознакомиться с поднимаемыми в этих работах многочисленными вопросами, возникающими при всегеномных исследованиях. Здесь мы ограничимся лишь некоторыми моментами. В целом методологические подходы к проведению GWAS являются предметом споров. Так, многие такие исследования критикуются из-за пренебрежения авторами работ вопросами контроля качества. Другую серьезную проблему составляют вопросы правильного анализа и интерпретации данных, часто

[60] Имеется в виду встречаемость в статьях только одной из них: или GWAS, или GWA, либо GWASs.

приводящие к появлению ложно-позитивных результатов. Так, отмечено, что с 2005 г. идентифицировано более 2000 ассоциаций снипов с различными состояниями организма человека, однако их клиническая применимость вызывает заметный скептицизм [Manolio, 2013]. Первоначально имевшая место эйфория в отношении снипов уже прошла. Весьма показательна обзорная статья, посвященная 5-летию проведения GWAS [Visscher et al., 2012], в которой авторы привели, в том числе высказывания известных специалистов-генетиков, среди которых сэр Alec Jeffreys, профессор M.C.King и другие, отмечающие, что возлагаемые на исследования снипов определенные надежды в плане более точного установления причин патологических состояний человека не оправдались.

В литературе уже довольно много статей, посвященных так называемому post-GWAS периоду [Gershon et al., 2011; Juran et al., 2011; Du et al., 2012; Maouche, Schunkert, 2012; Vrieze et al., 2012; Dube et al., 2013; Monteiro, Freedman, 2013; Wjst et al., 2013 и др.], где авторы в большинстве своем сходятся во мнении, что будущее массовое полногеномное секвенирование и анализ его результатов с акцентом не на однонуклеотидных заменах, действительно позволят разобраться с причинами болезней человека. Некоторые исследователи, проводя аналогию с next generation sequencing, говорят о новом поколении GWAS, которое будет требовать дополнительных усилий по связыванию результатов GWAS с биологическими функциями, что является наиболее критичным моментом [Wang et al., 2011]. При этом ряд авторов упоминают такой важный аспект как эпигеномная регуляция, совершенно не учитываемую сейчас при проведении GWAS [Rakyan et al., 2011; Ben-Avraham et al., 2012; Huynh, Casaccia, 2013; Michels et al., 2013 и др.]; другие предлагают проверять высказываемые предположения на модельных животных [Queitsch et al., 2012]. Поскольку функциональные особенности генов только определяют возможность болезненного состояния организма, но проявляется сама болезнь уже на уровне белков или даже вторичных метаболитов, то правы те исследователи, которые пытаются увязать результаты экспериментов по GWAS с протеомикой и метаболомикой [Kullo, Cooper, 2010; Adamski, 2012; Homuth et al., 2012; Robinette et al., 2012; Suhre, Gieger, 2012; Adamski, Suhre, 2013; Montoliu et al., 2013 и др.]. Предложено при поиске ассоциаций учитывать и фенотипические проявления, что нашло отражение в новой аббревиатуре - PheWAS (Phenome-Wide Association Study) [Denny et al., 2013], а также семейный анализ [Ott et al., 2011]. В литературе имеются высказывания, что будущее картирования геномов будет основано на всегеномном анализе вкупе с ресеквенированием всех вариантов внутри групп сцепления [Rowe, Tenesa, 2012]. И за такими работами, по-видимому, будущее.

Помимо поиска ассоциаций снипов с болезнями, GWAS находит применение и в фармакогеномике при выявлении отличий действия различных лекарственных средств на разных людей в зависимости от присутствия у них тех или иных снипов [Riancho, Hernandez, 2012; Sadee, 2012 и др.], но все же, принимая во внимание те изменения отдельных нуклеотидов, которые могут привести лишь к незначительным изменениям функциональных свойств каких-либо ферментов или прочих белков, то возможно оно (применение GWAS) для этой цели также не совсем оправданно. Однако и в эту область приходит понимание, что гораздо больший эффект может быть вызван CNV отличиями геномов людей [Dhawan, Padh, 2009; Gamazon et al., 2011; He et al., 2011], а не снипами. Но и здесь мы не сомневаемся, что ресеквенирование полных геномов человека «скажет» вскоре свое веское «слово», а к вопросу о вкладе в функциональное состояние человека CNV-полиморфизма вернемся чуть ниже.

Пожалуй, единственной областью применения GWAS, где оно вполне оправдано, является эволюционная геногеография популяций человека, называемая еще этногеномикой, или генетической генеалогией народов мира, хотя это понятие несколько шире. Так, проведено уже довольно много исследований различных популяций на разных материках, в том числе, нашими коллегами по Институту совместно с их зарубежными партнерами [Price et al., 2010; Metspalu et al., 2011; Reich et al., 2009; 2012 и др.]. Тем не менее, после того, как станет доступным новое дешевое секвенирование ДНК, будут в целом завершены основные эксперименты по структурной и сравнительной геномике человека (ориентировочно к 2020 г.), это, несомненно, заставит пересмотреть некоторые сегодняшние стратиграфические выкладки, часть других - уточнить. Если провести некую аналогию, например, с ботаникой или микробиологией, ранее использующих для ДНК-систематики сравнение размеров рестриктазных фрагментов, а теперь для этой цели секвенирующих гены рибосомных РНК растений и микроорганизмов, то приблизительно так по оценочной силе и точности будут соотноситься нынешние GWAS исследования популяций человека и будущее ресеквенирование полных геномов их представителей. При этом

общие тенденции, скорее вспего, сохранятся неизменными, но некоторые уточнения происхождения народов и народностей обязательно потребуются. Ведь GWAS даже с миллионом снипов, пусть и взятых в анализ из более-менее равномерно распределенных по всему геному, фактически позволяет оценивать лишь 0,03% всей последовательности[61]. Вряд ли достаточно для однозначных выводов. Конечно, можно подождать и фактически не выполнять по сути двойную работу, но кто его знает, когда ЭТО НОВОЕ секвенирование появится?... А эту информацию хочется получать уже сейчас, тем более, что есть понятие приоритета и всегда должно наличествовать вполне законное желание опередить коллег. Безусловно, можно и даже, наверное, нужно делать такую работу, но при этом еще и нужно отдавать себе отчет, что неизбежно придется к этим вопросам обязательно вернуться на новом витке будущих технологических возможностей современной биологии. Здесь можно также заметить, что, не дожидаясь нового секвенирования, венгерские ученые решили привлечь для филогенетического изучения популяций совсем другие технические возможности и проанализировали особенности народных мелодий, найдя по ним довольно четкие закономерности этнического родства [Pamajav et al., 2012].

В связи с публикационной активностью, опять-таки проводя параллели между анализом пресловутых снипов и прежними полиморфизмами (STR, VNTR, ПДРФ), четко видно, что время последних давно ушло и сейчас мало какой журнал решится принять к публикации статью, посвященную «старым» полиморфизмам в связи со здоровьем человека, да, наверное, никто уже и не проводит таких работ. Из этого следует, что пройдет какое-то время[62] и статьи по всегеномному анализу ассоциаций перестанут быть востребованы. Если, хотите - выйдут из моды. И к этому надо быть готовыми, тем более, что уже и сейчас возможны исследования таких мест генома человека, которые с гораздо большей вероятностью ответственны за его здоровье/нездоровье, к рассмотрению которых ниже перейдем.

* * *

Более масштабное деление изучения геномного полиморфизма у человека на периоды, чем это сделал Й.Накамура [Nakamura, 2009], приведено в работе шведско-сингапурского коллектива авторов [Ku et al., 2010]. Так, они указали всего три временных отрезка (прошлое, настоящее, будущее) и для каждого привели свои типы ДНК-полиморфизмов. В прошлом, по их мнению, остались ПДРФ и различные тандемные повторы. К настоящему они отнесли снипы, а для будущего спрогнозировали изучение CNV, инделов, инверсий, транслокаций и некоторых других типов вариаций нуклеотидных последовательностей и их блоков в человеческих геномах. Но некоторые взыскательные исследователи уже давно обратили внимание на такой тип полиморфизма генома человека как CNV [Iafrate et al., 2004; Sebat et al., 2004; Freeman et al., 2006; Redon et al., 2006 и др.], поскольку из-за варьирования числа копий геномы людей могут отличаться между собой на 10-12%. При этом у разных людей варьирует копийность тысяч генов. Во многих случаях выявленные отличия в числе копий тех или иных последовательностей влияют на транскрипционную активность и соответственно на функциональное состояние организма. В последние годы появляется все больше работ, где обнаруживается связь между имеющими место определенными CNV и конкретными болезнями [Усманова, 2009; Голимбет, Корень, 2010; Hollox et al., 2008; Henrichsen et al., 2009; Ionita-Laza et al., 2009; Wain et al., 2009; Zhang et al., 2009; Lee et al., 2010; Stankiewicz, Lupski, 2010; Pankratz et al., 2011; Xu et al., 2011; Almal, Padh, 2012; Spielmann, Klopocki, 2013; Priebe et al., 2013; Vandeweyer, Kooy, 2013; Zhao et al., 2013 и др.]. Публикуются и статьи, в которых вариации числа копий связывают с вопросами фармакогеномики [Gamazon et al., 2011; He et al., 2011].

---

[61] 0,03% всей последовательности генома составляют собственно сами вариабельные места в виде однонуклеотидных замен без учета прилегающих областей, которые, впрочем, от одного снипа к соседнему можно принять гомологичными на 100%. При этом не взятые в рассмотрение другие снипы, теоретически находящиеся между сравниваемыми с помощью GWAS снипами, как раз могут нести несколько иную информацию о близости анализируемых популяций. Более того, между сравниваемыми снипами могут иметь место и другие структурные перестройки генома, также не попадающие в анализ и тем самым искажающие истинную ситуацию с генетическими родством или удаленностью популяций. Опять-таки исследуется только биаллельное состояние снипов.

[62] На самом деле это время известно - оно наступит после появления нового дешевого и удобного метода ресеквенирования полных человеческих геномов, когда оно начнет носит поистине массовый характер.

Почему же в 2010 г. [Ku et al., 2010] изучению связи полиморфизма Copy Number Variation с болезнями отводилось только будущее, и наступило ли оно сейчас, спустя почти 4 года? Так, за последнее десятилетие с момента, как стали известны черновые варианты генома человека, к части исследователей пришло понимание того, что наиболее серьезные отличия между людьми кроются не в однонуклеотидных заменах, а в более сложных перестройках генома и, в частности, в вариациях числа копий конкретных генов. На последний же вопрос, видимо, следует ответить отрицательно - пока еще не наступило! Одна из причин заключается в том, что снипы никак не хотят «сдавать» свои позиции, в том числе, по причине сложившегося крупномасштабного бизнеса при GWAS-исследованиях. Несколько большая сложность детекции CNV, по сравнению со снипами, представляет собой другую причину.

Самым достоверным способом обнаружить различия в числе копий тех или иных участков генома человека является полногеномное ресеквенирование [Handsaker et al., 2011; Mills et al., 2011] или секвенирование экзомов [Saithirpongsasuti et al., 2011], но и то и другое пока все же дороговато, чтобы стать массовыми настолько, чтобы превратиться в диагностические процедуры. Другими способами подсчитать вариации числа копий конкретных фрагментов генома служит ПЦР в реальном времени [Rose-Zerilli et al., 2009], но разрешающая способность данного метода не вполне достаточна для таких оценок, и ей на помощь/смену приходит цифровая ПЦР [Dube et al., 2008; Weaver et al., 2010; Whale et al., 2012]. Довольно широко используемым подходом к выявлению CNV служит поиск последних с помощью разнообразных платформ по выявлению однонуклеотидных замен, для чего разработаны соответствующие компьютерные программы анализа получаемых первичных данных [Wang et al., 2007; Cooper et al., 2008; McCarroll et al., 2008; Dellinger et al., 2010 и др.], однако достоверность получаемых результатов оставляет желать лучшего. Еще одним способом обнаружения вариаций в числе копий нуклеотидных последовательностей служит метод эррейной сравнительной гибридизации [Buffart et al., 2008; McDonnell et al., 2013], для валидации которого предлагается использовать амплификацию, основанную на мультиплексном лигировании пробы [Shen, Wu, 2009; Stuppia et al., 2012]. Существуют и другие методы выявления CNV, сравнительный анализ которых приведен в ряде обзоров [Carter, 2007; Aten et al., 2008; Winchester et al., 2009; Ceulemans et al., 2012; Li, Olivier, 2013].

* * *

Возвращаясь к вопросу о резких изменениях на поле изучения полиморфизма ДНК, нужно заметить, что в результате появления нового высокопроизводительного секвенирования полных геномов также однозначно прекратится использование разных генетических маркеров, например, применяемых сейчас для маркирования у сельскохозяйственных растений и животных так называемых хозяйственно-полезных признаков (QTL - Quantitative Trait Locus) с целью их последующего выявления при ведении селекционной работы [Korol et al., 2012; Wurschum, 2012]. Основной причиной этого станет то, что такие маркеры носят практически случайный характер и их применение мало что дает. В свою очередь, больше вопросов, чем ответов возникает при использовании GWAS при изучении генетики количественных признаков [Korte, Farlow, 2013]. Тем более, что уже есть доказательства того, что наибольший вклад в фенотипическое разнообразие пород и сортов сельскохозяйственных животных и растений вносят не однонуклеотидные замены и прочие незначительные полиморфизмы, а вариации числа копий CNV тех или иных участков генома [Alvarez, Akey, 2012; Bickhart et al., 2012; Clop et al., 2012; Doan et al., 2012; Liu, Bickhart, 2012 и др.]. Фактически в этих случаях может работать так называемая «доза гена» и тогда обычный анализ на предмет простого выявления генетических маркеров уже не сработает, и необходимо проводить их количественную оценку, например с помощью цифровой ПЦР на уровне ДНК или РНК либо опять же путем секвенирования всего генома и/или транскриптома. Или количественно детектировать конкретные белковые продукты.

Перестанут также применяться неохарактеризованные (недостаточно охарактеризованные) генетические маркеры как на ДНК-, так и на РНК-уровнях для идентификации у людей различных заболеваний многофакторной природы, либо только предрасположенности к таковым. Таким биомаркерам на смену придут конкретные точные ДНК/РНК-«свидетельства» (не ассоциации!), сопровождаемые полноценными характеристиками, досконально изученными на геномном, транскриптомном и прочих необходимых уровнях. При этом нельзя исключать, что в будущем будет проще выявлять с высокой чувствительностью сразу конкретные белки (группу белков), ответственные(х) за развитие того или иного признака, либо за болезненные состояния. Либо непосредственно самих белковых молекул, или опосредованно, например, с помощью иммуно-ПЦР или аналогичными методами с использованием

вместо белковых антител аптамерных молекул. То есть не исключен (и даже более реален) некий гибрид биохимической/генетической диагностики.

Что касается большинства нынешних ДНК-чипов, производимых многими фирмами и рассчитанными на исследование структурной геномики, в том числе для всегеномных анализов с огромным множеством исследуемых точек, то они потеряют свою актуальность и их выпуск прекратится весьма быстро после разработки нового более производительного секвенирования ДНК. Для некоторых задач, например, при выявлении инфекционных агентов будут продолжать использовать методы амплификации специфичных фрагментов ДНК, однако нельзя исключать, что такие анализы, возможно, будут делаться и с помощью ДНК-чипов будущего (отличающихся от нынешних, если не кардинально, то, по крайней мере, весьма заметно), чувствительность которых позволит вести детекцию искомых последовательностей в исследуемых образцах, не прибегая к амплификации конкретных участков. При этом такие ДНК-чипы будут представлять собой относительно небольшую подборку конкретных участков генома, ограниченную двумя-тремя сотнями локусов. Еще более недолгим окажется век цифровой ПЦР, которая сейчас позволяет устанавливать более-менее точное число копий отдельных элементов генома, варьирующих у разных организмов, поскольку много точнее и проще эта информация будет становиться доступной после секвенирования полных геномов любых индивидов. Впрочем, цифровая ПЦР будет использоваться при контроле излечения больных от некоторых инфекций путем регистрации присутствующих в организмах людей точного числа возбудителей, вплоть до полного их исчезновения, означающего полное выздоровление.

Нам представляется, что для главных объектов, к каковым следует отнести человека, а также имеющих сельскохозяйственное или иное промышленное значение животных и растения, сравнительная геномика будет практически завершена уже к 2020 г. Это будет означать, что ОБЩАЯ КАРТИНА полиморфных областей, участков генома, вплоть до однонуклеотидных замен, инделов, инверсий и транслокаций станет абсолютно ясна, например, для рас, этнических групп, и дальнейшее секвенирование геномов людей будет представлять интерес лишь в плане персонифицированной медицины, для которой, впрочем, абсолютно недостаточно знаний о полиморфизме ДНК без связи с функцией.

Так, гораздо больший интерес для научного познания и выяснения детальных механизмов функционирования живых организмов будет представлять функциональная геномика. И здесь на смену ныне широко применяемому поиску ассоциаций вариаций нуклеотидных последовательностей с теми или иными заболеваниями мультифакториальной природы, ведущемуся сейчас даже на так называемом всегеномном уровне практически вслепую, должны прийти новые технологии, основанные на метиломике, транскриптомике, протеомике, пептидомике, метаболомике. Возможно, основанные и на других «омиках». Которые позволят уверенно устанавливать причинно-следственные связи между сравнительной геномикой, функциональной геномикой и здоровьем человека. И это будет абсолютно новый уровень знаний, которые не будут нуждаться в сопоставлении с ранее полученными результатами с помощью прежних (имеется в виду - нынешних) подходов. В качестве подтверждающего эту мысль примера можно привести уже сегодняшнюю ситуацию с секвенированием полных геномов. Так, например, решив секвенировать полный геном того или иного организма, авторы такого проекта не роются по ГенБанку на предмет того, а что уже было секвенировано для этого объекта до них, как-то пытаясь уменьшить (облегчить) себе работу, а целенаправленно определяют всю последовательность генома от начала до конца, доверяя, в первую очередь, своим данным. Да и не технологично это - исключать из фронтального секвенирования какие-то уже секвенированные гены, их фрагменты или прочие участки генома. Такой принцип будет действовать и для любых новых высокопроизводительных технологий секвенирования генома/метилома/транскриптома, равно как и для новых технологий сопоставления особенностей конкретных геномов с биологическим здоровьем их носителей и никто не будет искать и тем более цитировать работы, выполненные в предыдущую (нынешнюю) эпоху.

Перспективы и последствия развития полнотранскриптомного секвенирования

Не менее важен вопрос - а к чему в плане использования тех или иных методов приведет новая высокопроизводительная технология секвенирования РНК? Точнее, полного транскриптома любой клетки. Действительно, для понимания функционального состояния любого организма важно знание не столько или точнее не только последовательности всего генома, но и транскриптомов, которых в реальности неисчислимое множество. Как уже отмечалось выше, новые технологии секвенирования РНК должны быть все же мономолекулярными и обходиться без каких-либо (пред)этапов (пред)амплификаций, которые неизбежно искажают

реальную ситуацию с копийностью тех или иных молекул. Причем крайне необходимо секвенировать весь пул молекул РНК в клетке / типе ткани еще и в динамике с тем чтобы получить полную информацию как о транскрипционной активности всех РНК-кодирующих фрагментов ДНК, так и об уровне дифференциальной экспрессии, выражающейся уже в точном числе молекул разных типов[63]. Благодаря такой возможности некоторые современные и ныне весьма передовые методы исследования транскриптомов канут в лету. Так, для определения транскрипционной активности отдельных генов потеряет свою актуальность как цифровая ПЦР, так и ПЦР в реальном времени. Метод вычитающей гибридизации, направленный на установление транскрипционной активности генов в различных клетках/тканях/условиях также перестанет использоваться. Станут абсолютно ненужными транскрипционные ДНК-чипы. Ну, разве, что они будут применяться при массовых анализах функционирования каких-либо немногих конкретных генов, с которыми будет все ясно, и можно будет однозначно ставить по ним правильный диагноз. Таким образом, полнотранскриптомное секвенирование даст возможность экспериментаторам перейти на совершенно новый уровень исследований, позволяющих оценивать функциональное состояние организма и его отдельных органов (не только у человека).

Эта самая короткая главка во всей статье на самом деле фактически рассматривает (получилось, что очень кратко) перспективы и последствия технологии, которая через десяток лет станет по сути главенствующей, поскольку в целом информация о большинстве геномов уже будет известна и структурная геномика останется только для того, чтобы продолжать получать данные, которые будут лишь способствовать расширению наших познаний о Живом в плане эволюции, филогении, систематики, без дальнейших серьезных прорывов. Функциональная же геномика будет крайне востребована, так как будет добывать сведения о различных сторонах функционирования/жизнедеятельности отдельной клетки, ткани, органа и, наконец, целого организма. Безусловно, это должно происходить в тесном единении с метиломикой, метаболомикой, а также с протеомикой, которой мы уделим определенное внимание в третьей части данной статьи.

[63] В третьей части статьи к вопросу о количестве молекул мы еще вернемся.

Размеры геномов и C-value парадокс

Прежде чем пытаться строить прогнозы относительного того, сколько же полных геномов различных свободноживущих организмов будет секвенировано до 2030 года, видимо, следует немного задержать внимание читателя на их размерах, поскольку этот параметр, вне всякого сомнения, будет сказываться на массовости таких процессов.

Что же такое C-value и в чем заключается парадокс? Впервые термин C-value, обозначающий содержание ДНК на ядро в пикограммах, был предложен в 1950 г. при описании постоянства («C» – constant) количеств ДНК вне зависимости от типа ткани в находящихся на соответствующих стадиях жизненного цикла клетках, на примере традесканции и кукурузы [Swift, 1950]. В то время подобные измерения осуществлялись с помощью метода окрашивания по Фельгену, разработанному для окраски ядер еще в 1924 г. [Feulgen, Rossenbeck, 1924]. В различных вариациях такой подход просуществовал до начала 1970-х гг., когда ему на смену пришел способ определения количеств ДНК в ядре с помощью проточной цитофлуориметрии. Первой такой работой стало исследование количества ДНК в ядрах конских бобов *Vicia faba*, окрашенных бромистым этидием после пектиназной и протеазной обработок [Heller, 1973]. Эта технология заметно улучшила процедуру количественного измерения ДНК в ядрах и, несколько упростившись сама, поставила определение C-value у различных организмов чуть ли не на поток. К тому времени, благодаря накопившимся за предшествующие годы данным по содержанию ДНК в ядрах организмов разных уровней генетической сложности, стало ясно, что нет какой-либо корреляции между количеством ДНК в ядре и продвинутостью организма по эволюционной лестнице, что привело к введению в обиход понятий «загадка C-value» или «C-value парадокс» [Thomas, 1971]. Необходимо заметить, что в специализированной литературе термин C-value часто используется с различными префиксами – 1C, 2C, 3C и т.д., но не желая углубляться в такие тонкости и возможные причины парадокса C-value, можем посоветовать интересующемуся читателю обратиться к многочисленным обзорным и проблемным статьям на эту тему, в том числе, отечественных авторов [Патрушев, Минкевич, 2007; Шереметьев и др., 2011; Gregory, 2001; 2005; Greilhuber et al., 2005; Greilhuber, 2008; Patrushev, Minkevich, 2008; Greilhuber, Dolezel, 2009; Kraaijeveld, 2010; Bennett, Leitch, 2011 и др.].

Раз уж мы здесь коснулись истории, то, видимо, стоит вспомнить, когда впервые был

предложен и термин «геном», тем более, что к обсуждению размеров этих крайне важных частей организмов скоро перейдем. Произошло это почти столетие назад еще в 1920 г. в монографии, посвященной вопросам распространения и причинам партеногенеза в растительном и животном царствах [Winkler, 1920]. В главе про связь хромосом с партеногенезом Г.Винклером на стр. 165 была высказана следующая мысль - «Ich schlage vor, für den haploiden Chromosomensatz, der im Verein mit dem zugehörigen Protoplasma die materielle Grundlage der systematischen Einheit darstellt, den Ausdruck: das **Genom**[64]…». В переводе на русский она могла бы звучать приблизительно так: «Я предлагаю использовать для гаплоидного набора хромосом, который вместе с прилежащей цитоплазмой составляет основу таксономической единицы, термин геном…». Далее в этом же предложении он рассуждает о том, что если геном содержит более чем одну такую же единицу, то его следует называть гомогеноматическим, а если разные единицы, то тогда - гетерогеноматическим. Сейчас такие организмы принято обозначать как автополиплоиды и аллополиплоиды соответственно. В этой связи можем добавить, что в некоторых случаях правильнее сложносоставные геномы подразделять на субгеномы, что для хлопчатника было предложено еще в 1998 г. [Jiang et al., 1998]. При характеристике полиплоидных пшениц в наших исследованиях мы также использовали термин «субгеном», поскольку, пройдя через стадии интеграции и функциональной «диплоидизации», геномы исходных родительских форм за счет внутри- и межгеномных перестроек (транслокаций) в составе аллополиплоидных пшениц заметно изменяются и представляют собой, по существу, уже элементы интегрального генома, и в этой связи использование для последних термина «субгеном» вполне логично так как призвано показать их отличия от донорных геномов диплоидных эгилопсов и пшениц и облегчить дифференциацию, например, самостоятельного генома **D** эгилопса *Aegilops tauschii* от его производного - субгенома **D**, являющегося частью составного гексаплоидного **BAD** генома мягкой пшеницы [Вахитов и др., 2003]. В литературе встречается еще несколько работ, где применительно к полиплоидным пшеницам авторы оперируют понятиями субгеномов [Gill et al., 2004; Goyal et al., 2005; Gupta et al., 2008; Pont et al., 2013]. Для представителей полиплоидного ряда из рода *Brassica* семейства капустных для трех основных геномов **A**, **B**, **C**, ведущих свое происхождение от разных диплоидных видов – доноров этих геномов, также предложена концепция субгеномов [Zou et al., 2010]. В последние годы на субгеномы стали подразделяться геномы и некоторых других полиплоидных растений. Представляется, что в будущем по мере накопления сведений о полных геномах полиплоидных организмов для демонстрации принадлежности различных нуклеотидных последовательностей к тем или иным исходным формам, термин «субгеном» будет применяться все более широко.

Переходя к рассмотрению значений C-value и размерам геномов, хотим заметить, что в литературе до сих пор имеются расхождения при сопоставлении этих величин. Мы здесь будем использовать C-value, опуская префикс «1», вне зависимости от уровня плоидности, и тогда термины C-value и размер генома будут являться до некоторой степени синонимами. Отличия заключаются в том, что C-value выражается в пикограммах, а размеры геномов принято выражать в парах нуклеотидах. Другое отличие заключается в том, что C-value представляет собой довольно приблизительный показатель, тогда как размер генома (если таковой полностью[65]) секвенирован, является весьма точной величиной. Тем не менее, пересчет между этими характеристиками ядерной ДНК конечно же, возможен. Так, усредненный вес одной пары нуклеотидов составляет $1{,}023 \times 10^{-9}$ пг и 1 млрд. пар нуклеотидов соответственно будет «весить» 1,023 пг, а в 1 пг «уложится» 978 млн. пар нуклеотидов. Можно заметить, что некоторое время назад под размером генома понималась только его кодирующая часть без повторяющейся ДНК [Mitra, Bhatia, 1973], но такой подход в корне неверен.

Здесь при сравнении значений C-value разных организмов оперировать будем для каждой из групп только известными на настоящий момент величинами, что не исключает, что в будущем, когда таких сведений, благодаря, в том числе, новому полногеномному секвенированию, станет гораздо больше, то нижние и верхние размерные границы геномов для некоторых групп могут раздвинуться, тогда как для некоторых, напротив, сузиться. К тому же, вне всякого сомнения, что информация о полной нуклеотидной последовательности всего генома какого-либо организма дает куда более точные сведения о его

[64] Выделено нами.

[65] Полностью секвенированными можно считать, пожалуй, лишь геномы архей и прокариот, поскольку геномы высших организмов в силу особенностей их устройства в виде разнообразных повторяющихся участков ДНК из-за несовершенства прежних и нынешних технологий секвенирования так до конца и не секвенированы.

размере, нежели определение C-value любым способом. Справедливости ради, следует заметить, что проточная цитофлуорометрия обеспечивает довольно точные значения размеров геномов, что показывает сопоставление таких данных с полногеномными последовательностями [Bennett et al., 2003]. Но уже сейчас ясно, что для одних групп организмов свойственен довольно узкий диапазон геномных размеров, тогда как внутри других геномы могут отличаться на порядки (рис. 3). В целом для всех свободноживущих организмов количества ДНК на (ядерный) геном сейчас укладываются в диапазон от 500 аттограмм (бактерии) до 700 пикограмм (простейшие), что составляет более чем 6 порядков разницы и это огромный диапазон. При этом, чем организмы более эволюционно продвинуты, тем в грубом приближении можно принять, что разброс их C-value меньше.

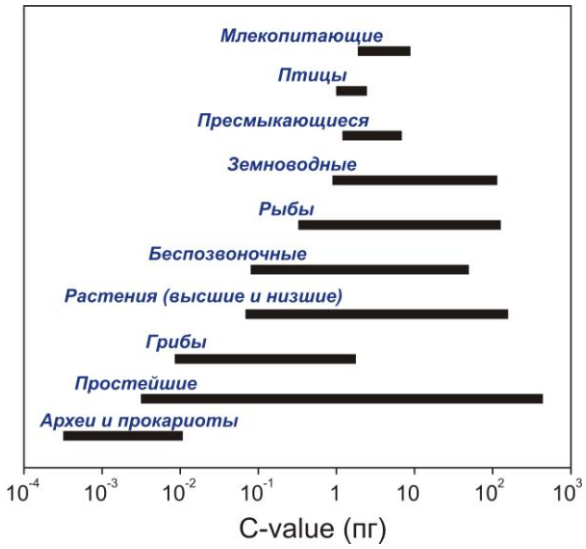


Рис. 3. «Разброс» значений C-value среди различных групп организмов

Кратко рассмотрим значения C-value и размеры геномов некоторых групп организмов. Так, следуя вверх по эволюционной лестнице, начнем с бактерий[66]. Округленно можно посчитать, что размеры геномов архей и прокариот отличаются приблизительно в 20 раз. Если считать в нуклеотидных парах, то для полностью секвенированных геномов – округленно от 500 тысяч пар нуклеотидов до 10 млн. Вполне вероятно, что среди секвенированных в будущем геномов тех архей и прокариот, которые пока остались неисследованными, могут найтись объекты и с меньшими и с большими размерами, однако ждать, что они будут отличаться между собой более чем в 25-30 раз, видимо, не следует. Таким образом, наиболее вероятен диапазон от 400 тысяч до 12 млн. пар нуклеотидов. Геномы таких размеров с помощью технологий новых поколений должны будут секвенироваться совсем легко и просто.

Для эукариотических же организмов уже сейчас известно, что размеры их геномов отличаются более существенно - на 4 – 5 порядков. Первым эукариотическим организмом, полная последовательность генома которого была определена, стали дрожжи *Saccharomyces cerevisiae* [Goffeau et al., 1996]. В результате выполнения этого проекта было определено чуть более 12 млн. пар нуклеотидов, хотя весь полный геном дрожжей, сосредоточенный по 16 хромосомам, оценивается в более чем 13 млн. C-value одной из нематод *Pratylenchus coffeae* – относительно крупного организма по сравнению с дрожжами - оценивается всего в 0,02 пг [Leroy et al., 2007]. Его еще предстоит секвенировать, но сильного расхождения C-value с истинным размером генома здесь явно не будет. Наименьший известный размер генома в 2,3 млн. пар нуклеотидов имеет паразитический организм микроспоридия *Encephalitozoon intestinalis* и он уже полностью секвенирован [Corradi et al., 2010]. Ранее с помощью пульс-гель-электрофореза было подсчитано, что геном другого вида микроспоридии *E. cuniculi* имеет размер около 2,9 млн. пар нуклеотидов [Biderre et al., 1995] и надо отметить, что этот метод также вполне пригоден для приблизительной оценки геномов размером до 10 – 15 млн. Хочется верить, что при полногеномном секвенировании эукариотических организмов со столь малыми размерами геномов не потребуется в будущем особых усилий. Что касается геномов весьма крупного размера, к секвенированию которых необходимо будет приступать более продуманно, то здесь ситуация иная, и вероятность получения в будущем заметно отличающихся цифр в размерах геномов от нынешних оценок C-value весьма высока. Пока считается, что самые крупные геномы размерами в сотни пикограмм принадлежат некоторым амебам *Amoeba dubia* (700 пг), *Chaos chaos* (1400 пг), но эти сведения были получены достаточно давно с помощью окрашивания ДНК дифениламином [Fritz, 1968], что позволяет допустить и вероятность ошибки. Хотя про амеб известно, что они могут иметь до 1000 мелких хромосом, вероятно, будучи полиплоидами в крайне

[66] В третьей части статьи при описании системной и синтетической биологии вопросу о размерах геномов микроорганизмов в связи с так называемым минимальным геномом свободноживущего организма будет уделено отдельное внимание.

высокой степени. Зачем им такие геномы, видимо, станет ясно, и то в первом приближении, только когда их полностью секвенируют. Хотя вполне возможно, что после выполнения в будущем таких проектов методами полногеномого секвенирования четвертого или пятого поколений приписываемые амебам размеры геномов не подтвердятся. Но пока «в защиту» достоверности обладания амебами геномами таких огромных размеров можно здесь привести сведения о том, что именно у них найдены так называемые мегавирусы Pandoravirus, обладающие гигантскими размерами (более 700 нм) и такими же чудовищными (по вирусным меркам) геномами (2,5 млн. пар нуклеотидов), кодирующими свыше 1000 генов, и эти показатели вполне сравнимы с таковыми у бактерий, ведущих паразитический образ жизни [Phillippe et al., 2013], что, возможно, не случайно, опять-таки учитывая количество ДНК в амебах.

Исходя из вызывающей в большей степени доверие информации, считается, что наиболее крупный геном в царстве животных принадлежит мраморной двоякодышащей рыбе, оцениваемый сейчас почти в 133 пг[67]. Хотя несколько ранее его принимали равным 142 пг [Pedersen, 1971], но используемый тогда автором метод определения содержания ДНК у этого вида основывался на грешащим неточностью окрашивании по Фельгену. Относительно недавно был найден организм, имеющий геном еще большего размера – 152 пг. Почти в 50 раз больше человеческого! Им оказалось растение из класса однодольных - вороний глаз *Paris japonica* [Pellicer et al., 2010]. Другая крайность для представителей растительного царства – это насекомоядные растения из семейства *Lentibulariaceae* – *Genlisea margaretae* (0,0648 пг), *G.aurea* (0,065 пг) и *Utricularia gibba* (0,0903 пг) [Greilhuber et al., 2006]. Причем два первых растения имеют геномы в два раза меньшие, чем арабидопсис, известный всем своими карликовыми размерами генома. По сравнению с тем же вороньим глазом генлисея имеет геном в 2330 раз меньшего размера. Казалось бы, разброс для высших растений действительно огромный. Однако проведенный отечественными авторами подробный анализ известных C-value растений позволил им прийти к заключению, что сложившееся мнение о чрезмерной изменчивости содержания ядерной ДНК несколько преувеличено, поскольку 80% геномов всех растений укладываются в относительно узкий диапазон от 0,6 до 16,7 пг, а по 10% растений с геномами меньшего и большего размеров представляют собой некие крайности [Шереметьев и др., 2011]. Можно предположить, что сходная тенденция характерна и для некоторых других групп живых организмов. Но как бы то ни было, секвенировать геномы всех этих организмов (или, по крайней мере – многих из них) когда-нибудь да придется. В этой связи неизбежным этапом перед стартом любого проекта по секвенированию *de novo* еще неизвестного генома любого организма (кроме, пожалуй, бактерий, ввиду малого размера их геномов) будет определение C-value, которое будет служить ориентиром как при выборе стратегии секвенирования, так и осуществлении сборки полных геномов.

Поскольку в последующие годы непременно будет расти интерес к секвенированию полных транскриптомов, то представляется важным знать их ориентировочные «размеры» для разных групп организмов, хотя этот показатель может меняться как между типами клеток, так и в зависимости от многих причин для одного типа клеток. Тем не менее, определенные видовые границы для размеров транскриптомов безусловно существуют. Несмотря на то, что у высших организмов ядерные геномы, как уже было сказано выше, различаются между собой на 4 - 5 порядков, считается, что транскриптомы у них варьируют по размеру всего 17-ти кратно[68] [Cavalier-Smith, 2005]. Подсчитано, что транскриптом одной клетки человека в весовом выражении составляет около 11 пикограмм, из которых на долю мРНК приходится от 1 до 3% или усреднено около 250 фемтограмм [Nygaard et al., 2005]. Важную информацию при проектировании полнотранскриптомного секвенирования будут представлять предположительные количества генов, предсказанные для полностью секвенированного генома какого-либо изучаемого организма, которые, впрочем, колеблются для свободноживущих организмов не в очень широких пределах – от 500 до 30 тысяч, если не принимать во внимание аллополиплоидные формы. Что касается копийности тех или иных типов молекул РНК, образующихся при транскрипции соответствующих генов, то эти цифры как раз станут точно известны, только когда начнется полнотранскриптомное секвенирование, не требующее перевода РНК в ДНК и обходящееся без каких-либо стадий амплификации. Хотя уже сейчас для некоторых организмов произведены довольно грубые (с помощью ПЦР в реальном времени) подсчеты, показывающие, что некоторых типов мРНК может содержаться в конкретной клетке до 10^3 молекул, а рРНК – до 10^6 молекул, при этом другие

[67] www.genomesize.com

[68] Хотя к такой оценке варьирования транскриптомов по размерам надо относиться довольно настороженно.

молекулы РНК находятся в виде единичных копий [Shimada et al., 2010].

Количества секвенированных полных геномов и транскриптомов к 2030 году

Так сколько все же полных геномов и транскриптомов будет определено к 2030 г.? Во всем мире уже довольно давно функционируют международные консорциумы, ставящие целью секвенирование, например, 5 тысяч геномов насекомых и членистоногих, другие намерены секвенировать более тысячи разных растений арабидопсиса из разных мест произрастания, есть проекты по секвенированию 10 тысяч видов позвоночных и т.д. В рамках одного такого консорциума после завершения секвенирования 1092 геномов человека [The 1000 Genome Project Consortium, 2012] в дальнейшем предстоит секвенировать уже 2500 геномов людей из 26 популяций. Однако мы в своем прогнозе опираемся больше не на эти заявленные цифры, а исходим из неких общих тенденций, также принимая во внимание число всех видов разных групп живых организмов на Земле.

Безусловно, дать точный прогноз - невозможно, но все же попытаемся представить, каковы будут тенденции этого процесса и попробуем объяснить получившиеся цифры. Конечно, необходимо опираться на уже известные и предположительные количества видов организмов разных групп. К сожалению, цифры и прогностические оценки числа видов довольно сильно разнятся [May, 1988; 1990; Mora et al., 2011]. Так, по большинству оценок, всех видов различных организмов на Земле насчитывается примерно 5±3 миллионов, т.е. от 2 до 8 миллионов [May, 1988; Costello et al., 2013]. И если принять во внимание, что к настоящему времени уже каталогизировано около 1,5 млн. видов, то существует вероятность, что остались пока без внимания всего-то полмиллиона видов. Или целых 7 миллионов при другом допущении. По мнению других исследователей, число всех видов на Планете оценивается в 8,7±1,3 миллиона [Mora et al., 2011, что по максимуму составит все 10 миллионов, из которых, как подсчитали авторы, около 2 миллионов приходится на морских обитателей. Но, как бы то ни было, цифры по количеству всех видов разнятся все же не на порядки и могут отличаться от истинных значений (к знанию которых приблизимся в будущем) всего-то в 2 - 5 раз, что не должно очень сильно повлиять на нашу весьма приблизительную оценку числа секвенированных геномов в будущем, выраженную в очень сильно округленных числах. Также довольно округленно указываем здесь число полностью секвенированных геномов к концу 2013 г.

Таблица

Прогноз количества полностью секвенированных геномов и транскриптомов для разных групп организмов

Группы организмов	**Число видов**	**Г о д ы**					
		2013	**2015**	**2017**	**2020**	**2025**	**2030**
Микроорганизмы *(прокариоты, археи)*	11000	8000	30000	60000	10^5	2×10^5	5×10^5
Простейшие	10000	100	200	300	500	2000	4000
Грибы	40000	150	500	2000	5000	10000	25000
Растения *(низшие и высшие)*	2×10^5	100	200	500	3000	30000	10^5
Животные *(беспозвоночные)*	10^6	100	500	2000	5000	15000	30000
Животные *(позвоночные)*	20000	50	300	2000	5000	20000	60000
Человек	1	1000	3000	15000	2×10^5	10^7	10^9
Древние организмы *(разные, не включая человека)*	-	5	20	100	200	400	1000
Транскриптомы *(все организмы)*	-	200	400	1000	10000	10^5	10^6

В первой части данной статьи мы уже упоминали чрезмерно завышенные оценки числа бактериальных видов, которые им одно время приписывали. На самом деле, скорее всего, видов прокариот около 10 тысяч, а архей - не более тысячи. Однако виды этих групп организмов характеризуются довольно большим разнообразием штаммов, которых может быть огромное множество, принимая к тому же во внимание происходящий обмен генетическим материалом между ними. Тем более, что уже и искусственно стали создавать новые микроорганизмы, помимо многочисленных рекомбинантных штаммов. Учитывая относительно небольшие размеры их геномов и соответственно легкость секвенирования, а также практическую значимость, к 2030 г., вероятно, будут установлены полногеномные последовательности не менее чем для полумиллиона штаммов, охватывающих все известные к тому времени виды, включая некультивируемые. А возможно, что и несколько миллионов геномов микроорганизмов будут секвенированы.

Считается, что Простейших на планете обитает около 10 тысяч видов (допускается, что может быть и 40 тысяч), но поскольку они не столь важны для хозяйственной деятельности человека, то, вероятно, серьезного внимания секвенированию полных геномов этих организмов уделяться не будет, и к 2030 г. может стать известными всего около 4 тысяч их геномов.

Грибы представляют собой более важную группу организмов, среди которых немало как весьма полезных, так и крайне вредных для человека видов. Число их видов пока оценивается в 40 тысяч, хотя допускается цифра и в 600 тысяч. При этом, многие виды включают в себя отличающиеся расы, что должно повлиять на число секвенированных геномов, которых, как можно допустить, станет известно к 2030 г. не менее 25 тысяч. А то и все 50 тысяч.

Низших и высших растений уже известно более 200 тысяч видов, и считается, что данная цифра может достичь 300 тысяч[69]. Принимая во внимание, что к этой обширной группе живых организмов относится немало культурных (сельскохозяйственных и декоративных) и прочих промышленно значимых растений, то интерес к их геномам может быть довольно высок. При этом будут секвенироваться не только виды этих растений, но и их различные сорта. Также весьма вероятно, что создаваемые трансгенные растения будут в обязательном порядке подвергаться полногеномному секвенированию, для того, чтобы обнаружить место встраивания трансгена для прогнозирования каких-либо последствий этого события на геном и транскриптом и все растение. Исторически сложилось так, что филогенетические связи в растительном мире всегда представляли заметный интерес, в связи с чем для лучшего понимания таксономических отношений могут быть секвенированы основные представители всех крупных и важных таксонов. Так что цифра в 100 тысяч геномов, возможно, даже сильно занижена.

Что касается беспозвоночных животных организмов, то их видов без насекомых известно всего 100 тысяч, с насекомыми - около миллиона. Казалось бы, здесь есть большой простор для секвенирования полных геномов, однако, учитывая значимость в хозяйственной деятельности человека довольно небольшого числа их представителей, нам представляется, что охвачены полногеномным секвенированием будут, естественно, далеко не все виды. Хотя можно предположить, что, например, к 2050 г. геносистематики определят последовательности геномов почти всех видов этой группы.

Видов позвоночных животных, можно сказать, существует довольно мало, но многие из них представляют для человека и хозяйственный, и познавательный интерес. Поэтому помимо самих видов (геномы большинства которых будут просеквенированы), немало геномов будет принадлежать породам домашних и одомашненных животных, включая трансгенные организмы. Так, только одних пород собак насчитывается уже под тысячу и, вне всякого сомнения, что для большинства их геномы будут полностью секвенированы. Поэтому, несмотря на относительно малое число видов этой группы и довольно крупные размеры геномов, цифра в 60 тысяч геномов, превышающая в два-три раз число самих видов, кажется обоснованной.

Наконец, человек. Как уже отмечалось, сравнительная геномика для человека закончится к году так к 2020, когда будет получена достаточная информация о геномах подавляющего большинства народов и этнических групп, что позволит установить довольно точную картину перемещения людских масс по континентам в прошлом и возникновения различных рас и народностей. После чего геномы людей будут преимущественно секвенироваться для целей персонифицированной медицины, причем это должна быть (будет!) медицина, основанная, как уже отмечалось, не на вероятностных ассоциациях, а на точных знаниях. Можно предположить, что к 2030 г. будут секвенированы геномы только у одного миллиарда

[69] По некоторым другим оценкам Царство растений может включать в себя и до полумиллиона видов.

человек, что составит около 10% всего населения того времени. Это пессимистический прогноз. В оптимистическом варианте можно допустить секвенирование геномов 3 - 4 миллиардов человек. Отчасти это будет зависеть от точности знаний о связи особенностей генома человека с его болезнями.

Что касается секвенирования полных геномов древних организмов. Безусловно, их количество будет оставаться небольшим, и даже дело не в новых технологиях секвенирования, а в ограниченной доступности материала. Тем не менее, если не включать в эту группу останки людей, живших в последние 3 - 4 тысячелетия, то цифра в 1000 древних геномов кажется довольно правдоподобной. В эту категорию также не попали гербарные и аналогичные музейные образцы, возраст которых исчисляется десятками и сотнями лет, секвенирование полных геномов которых, впрочем, уже началось [Staats et al., 2013].

Про секвенирование полных транскриптомов надо сказать следующее. Во-первых, их число для всего живого просто не поддается подсчету, поскольку разные типы клеток многоклеточных организмов могут иметь каждая свои транскриптомы. Во-вторых, даже в одном и том же типе клеток транскриптом может меняться как с возрастом, в связи с болезнями, так и после различных химических (лекарства, пища и др.) и физических (погодные условия, радиационный фон, магнитные бури и пр.) воздействий. При этом секвенирование полных транскриптомов у организмов из дикой флоры и фауны вряд ли будет представлять какой-либо заметный интерес, если только это не будут модельные объекты, используемые в научных исследованиях (дрозофила, мышь, крыса, арабидопсис и др.). Но число таковых немногочисленно. Поэтому основная масса секвенированных транскриптомов будет принадлежать видам (сортам, породам и др.), возделываемым и выращиваемым в сельском хозяйстве, в лесном деле, в звероводческих фермах, будут они и среди оранжерейных культур и т.д. Транскриптомы человека имеют ограниченную доступность и потому общее количество секвенированных транскриптомов для всех групп организмов, скорее всего, не превысит одного миллиона. Хотя это число может подобраться и к миллиарду, что будет зависеть, в первую очередь, от удобства применяемого(ых) метода(ов).

При анализе геномов и транскриптомов некоторых видов особый интерес представляет секвенирование их метиломов [Baranzini et al., 2010; Furukawa et al., 2013 и др.], причем число метиломов также непостоянно и может меняться под воздействием фактически тех же факторов, что действуют и на транскриптомы. Но поскольку в этом направлении пока грандиозных успехов нет, то мы решили не включать этот прогноз в таблицу. Хотя, например, уже есть сообщения, о метиломах клеточных линий человека [Lister et al., 2009], моноядерных клеток периферической крови человека [Li et al., 2013], о метиломе арабидопсиса [Cokus et al., 2008] и др. Количества выявленных метиломов для важных в хозяйственном плане видов и разных модельных организмов, а также человека, несомненно, будут расти, но темпы этого процесса будут зависеть, в первую очередь, от удобства применяемых технологий и, главным образом, достоверности получаемых с их помощью результатов. Все же, учитывая несколько повышенную трудоемкость такого секвенирования (которое и в будущем, наверняка останется более сложным, чем определение обычных нуклеотидов), можно представить, что к 2030 г. метиломов полных геномов будет определено не более чем 10^5.

Также в данную таблицу оказались не включенными сведения о метагеномах, представляющих собой некую сборную информацию о ДНК разнообразных микроорганизмов, населяющих разнообразные ниши, включая и организм (тело) человека. Нам, однако, представляется, что к 2030 г. технологии секвенирования геномов, возможно, позволят секвенировать все же большинство геномов даже некультивируемых микроорганизмов по отдельности. Тем более, что уже есть примеры этому [Raghunathan et al., 2005; Lasken, 2012; Marshall et al., 2012]. При этом нельзя исключать, что со временем будут секвенироваться внеземные метагеномные последовательности (например, марсианские, причем секвенироваться именно там, на красной Планете, с помощью специализированного робота, оснащенного компактным ДНК-секвенатором, чтобы максимально исключить контаминацию земными микроорганизмами), тем более, что определенные намерения на этот счет уже есть у К.Вентера и некоторых других. При этом все же нет уверенности, есть ли Жизнь на Марсе, и если даже есть - не факт, что она совпадает по химическим особенностям (тем же нуклеотидам[70], да и по всему

[70] Считается, что «выбор» Природы на использование в нуклеотидах рибоз и дезоксирибоз пал достаточно случайно; вместо них вполне могли быть другие сахара, например, треозы. Да и сами азотистые основания могли быть иными, не говоря уже о совсем другой, также возможной организации жизненных форм.

остальному) с Земной и, следовательно, планируемые к использованию на Марсе технологии секвенирования ДНК могут и не дать нужной информации. Однако нельзя исключать, что в будущем (возможно, что только после 2030 г.) появятся такие мономолекулярные методы секвенирования (ДНК/РНК или иного генетического материала), которые в принципе смогут распознавать последовательность из чуть ли не любых мономеров в практически любых полимерных молекулах, и тогда основными задачами станут выделение и очистка должным образом таких секвенируемых соединений. Но пока нет достаточных оснований предполагать, сколько же геномов (их фрагментов) внеземных организмов будет прочитано к 2030 г. Еще одно применение метагеномика может найти в криминалистике. Так, было показано, что оставленные отпечатки на клавиатуре компьютера несут специфичный для каждого пользователя набор бактерий [Fierer et al., 2010], к разговору о чем мы еще вернемся ниже.

Сознательно не вспоминаем здесь геномы органелл (митохондрий, хлоропластов), геномы симбиотических бактерий насекомых, вирусные геномы, поскольку они не относятся к свободноживущим организмам и при этом характеризуются более мелкими размерами геномов, хотя надо признать, что многие из них секвенировать ранее методом Сэнгера было далеко не так просто. Методы нынешнего полногеномного секвенирования существенно увеличили число прочитанных геномов органелл, крупных вирусов, симбиотических бактерий насекомых, а методы секвенирования четвертого и пятого поколений, безусловно, окажут еще более заметное влияние на процесс массового секвенирование таких геномов, которых в общем и целом будет секвенировано, наверное, не менее одного миллиона, а то и десяти.

Из приведенной таблицы следует, что для разных групп организмов в 2015 - 2030 г. интенсивность секвенирования полных геномов будет меняться по своим законам, а не только зависеть от используемых технологий. Так, для некоторых групп организмов будет наблюдаться некое насыщение полногеномными данными, для других - рост будет постоянно линейный, а для третьих - даже прогрессивный. Для большей наглядности такой динамики данные таблицы мы решили привести и в виде графика (рис. 4), из которого четко прослеживаются упомянутые особенности.

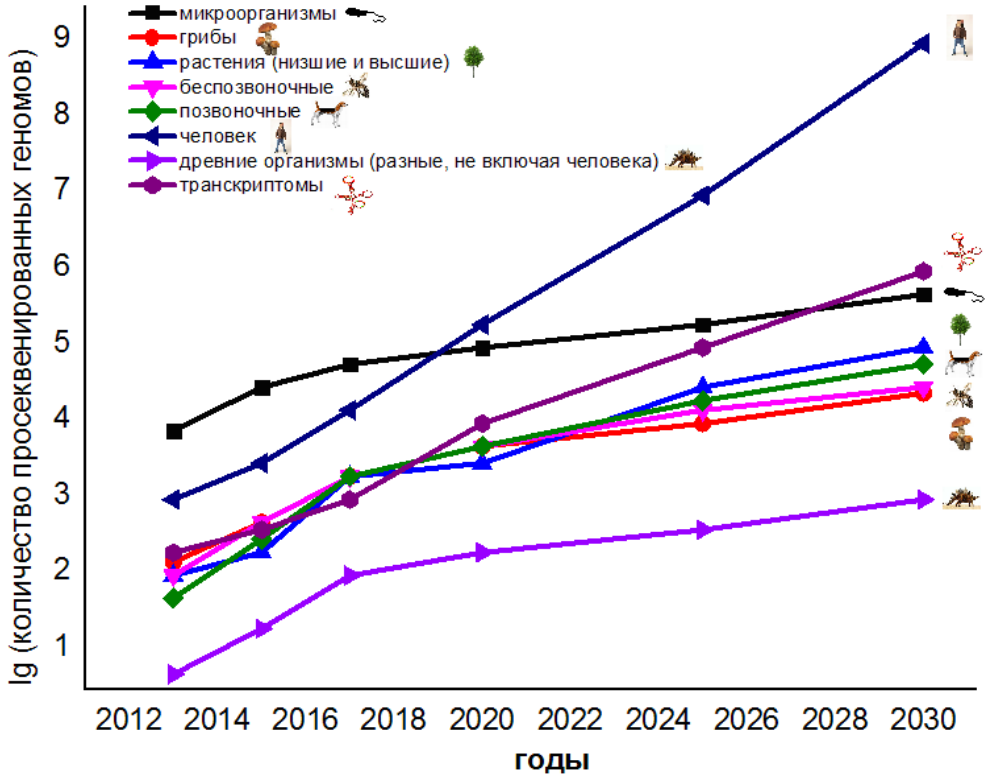


Рис. 4. Динамика предполагаемого роста числа секвенированных полных геномов и транскриптомов по разным группам организмов на период до 2030 г.

В первые годы секвенирования ДНК методами Максама-Гильберта и Сэнгера практически любая становящаяся известной последовательность нуклеотидов (даже довольно короткая) помещалась в базу данных. В настоящее время думается, что дело обстоит несколько иначе - далеко не все секвенированные последовательности нуклеиновых кислот любых организмов в силу разных причин заносятся в ГенБанк[71]. Отчасти потому, что определение нуклеиновых последовательностей уже носит иногда исключительно технический характер, например, для проверки правильности лигирования фрагментов ДНК или для проверки точной работы ДНК-полимеразы при амплификации некоего фрагмента ДНК до его использования в конструировании рекомбинантных молекул. Аналогично бесчисленным электрофоретическим разделениям продуктов рестрикции, амплификации и т.д., из которых только немногие картины гель-электрофореза могут попасть, например, в статью. Экстраполируя эту ситуацию на будущее, можно представить, что далеко не все полные геномы (для видов с уже известными полногеномными последовательностями) будут со временем регистрироваться в банках данных, так как будут носить такой же технический характер. Мы уже упоминали, что будут секвенироваться геномы трансгенных организмов на предмет выяснения мест встройки трансгенов, при этом такая процедура может носить довольно массовый характер на стадии отбора лучших форм, которые лишь и будут удостоены того, чтобы их последовательность стала известна не только самим экспериментаторам. Это только один такой пример, а ситуации оставления информации о полных геномах каких-либо организмов фактически для внутреннего пользования могут быть довольно различными. Что касается персональных геномов людей, то информация о них возможно будет храниться в неких электронных медицинских карточках пациентов. Таким образом, к 2030 г. (и даже раньше) будет абсолютно невозможно подсчитать, сколько же полных геномов всего в действительности секвенировано. И это нормально.

[71] Под ГенБанком здесь подразумевается объединенная база данных DDBJ/EMBL/GenBank, которой будет уделено соответствующее внимание в следующей части статьи в разделе про биоинформатику.

Места проведения секвенирования нуклеиновых кислот новых поколений

Теперь, похоже, настало время поразмышлять, где же будет вестись полногеномное секвенирование в будущем. Про Марс выше упомянули, но массового секвенирования там, надо думать, все же не будет, да и под «где» мы здесь подразумеваем, в какого типа организациях будут проводиться такие работы. Когда-то давно[72] секвенирование ДНК велось в ручном варианте с помощью радиоактивной метки и при обязательном наличии рекомбинантных молекул ДНК, что позволяло заниматься этим процессом только молекулярно-биологическим лабораториям, располагающим при этом соответствующим оборудованием, которое, впрочем, можно было изготовить тогда даже кустарным способом. При этом производительность секвенирования была довольно низкой. С появлением флуоресцентной метки и автоматизации определения нуклеотидных последовательностей производительность всего процесса выросла весьма значительно. Как возросла и стоимость уже довольно сложного оборудования в виде автоматических ДНК-секвенаторов первого и второго типов (в пластинах геля и в капиллярах соответственно). Рядовое секвенирование по методу Сэнгера (т.е. относительно небольших фрагментов ДНК) в таких приборах могли вести обзаведшиеся подобным оборудованием практически любые лаборатории, поскольку уже появились специальные наборы для проведения терминирующих реакций, и стало возможным нарабатывать матрицы для секвенирования без молекулярного клонирования, а путем амплификации фрагментов ДНК с помощью ПЦР. Однако выполнение крупных проектов в виде секвенирования полных геномов каких-либо свободноживущих организмов с помощью метода Сэнгера уже потребовали не одного-двух, а десятков и даже сотен таких приборов, что стало под силу только крупным специализированным центрам, лабораторные помещения которых напоминали больше заводские цехи со стоящими в ряд многочисленными ДНК-секвенаторами [Cyranoski, 2010]. В результате круглосуточной работы таких «лабораторий-цехов» до нескольких миллиардов нуклеотидных последовательностей становились известными за неделю. И потребовалось соответствующее компьютерное и программное обеспечение вместе с подготовленными специалистами-биоинформатиками. Появились также новые услуги в виде заказного секвенирования разного масштаба - от нескольких

[72] В первое десятилетие развития методов секвенирования ДНК с 1977 по 1986 гг.

фрагментов ДНК до полных геномов.

Дальнейшее развитие техники определения нуклеотидных последовательностей в виде разработки методов полногеномного секвенирования нескольких новых поколений повлекло за собой необходимость производства еще более дорогостоящих приборов, приобрести которые оказалось не всем под силу. К тому же экспериментаторам пришлось учитывать масштаб стоящих перед ними задач, поскольку дешевле и проще стало секвенировать отдельные геномы разово «под заказ», нежели иметь у себя часто или подолгу простаивающее дорогостоящее оборудование. Более того, секвенирование на приборах новых поколений потребовало новых знаний и навыков, позволяющих эксплуатировать такое оборудование только высококлассным специалистам. Отдельной задачей стала сборка геномов секвенируемых организмов из огромного множества так называемых ридов (коротких черновых последовательностей), объединяемых в более протяженные контиги в виде перекрывающихся последовательностей. Несмотря на то, что эту работу выполняет компьютер, очень важное значение имеет, какого типа секвенирование было проведено - ресеквенирование или секвенирование *de novo*. В последнем случае задача усложняется даже не на порядок, равно как и стоимость такого проекта.

Таким образом, оборудование для «нового» секвенирования стало концентрироваться в крупных специализированных центрах, как в ранее действующих при эксплуатации метода Сэнгера, так и в новых. Ситуация несколько поменялась после появления на рынке относительно недорогих полногеномных секвенаторов полупроводникового типа. Что касается будущего полногеномного секвенирования, то при сохранении и, особенно при развитии нынешней тенденции, некоторого упрощения методов и технологий, приборами четвертого и пятого поколений смогут владеть профессионалы уже не столь высокого уровня, который требуется сейчас. Более того, как упоминалось выше, существует огромная разница между ресеквенированием и секвенированием *de novo*, однако со временем будет все больше ресеквенирования, поскольку многие геномы часто исследуемых видов, а также близкородственные им будут уже многократно секвенированы, и фактически, имея базовые референсные последовательности геномов многих организмов, в них необходимо будет лишь определять полиморфные участки. Таким образом, и биоинформатическая составляющая процесса полногеномного секвенирования будет несколько упрощаться, сопровождаемая к тому же усилением мощи компьютерной техники. Все это может привести к тому, что секвенировать геномы свободноживущих организмов (по крайней мере, ресеквенировать) смогут опять-таки практически любые лаборатории, которые будут в состоянии позволить себе приобрести полногеномные секвенаторы будущих поколений и покупать расходные материалы к ним. Есть даже прогнозы, что выполнять различные молекулярно-биологические эксперименты и секвенировать геномы со временем смогут неспециалисты чуть ли не на кухне или в гараже, поскольку во всем мире и, в том числе, в биологии ширится такое движение как «DIY - Do It Yourself» [Ledford, 2010; Biba, 2011; Stevens, 2011]. Сомнительно, конечно, что дело дойдет до этого, но то, что во всем мире прослеживается тенденция проводить все больше разнообразных анализов по так называемому месту лечения (Point-Of-Care-Testing - POCT) неоспоримо, и кто знает - может быть, к 2030 г. действительно каждый, кто захочет, сможет секвенировать геномы чуть ли не дома. Пока к таким POCT-тестам можно отнести лишь установление беременности и определение сахара в крови, но развитие микрофлюидных технологий может сделать возможным и домашний анализ нуклеиновых кислот, а там, глядишь, и полногеномное секвенирование станет доступным.

Archea* и *Procarya

Учитывая относительно небольшие геномы этих ветвей Жизни, «полногеномное будущее» собственно для них уже давно началось. Однако, несмотря на это, в микробиологии до сих пор сохранились некоторые архаичные[73] подходы при описании новых видов и родов микроорганизмов. Так, ранее для принятия видов и родов за новые следовало обязательно определять GC-состав их ДНК, служивший тогда довольно важным таксономическим критерием, учитывая, что с морфологическими особенностями у этой группы живых организмов дело обстоит небогато. Здесь необходимо остановиться на GC-составах и точности их определения чуть подробнее. Так, например, в конце 60-х гг. при описании нового вида термофильной бактерии *Thermus aquaticus* GC-состав ее ДНК с помощью аналитического

[73] Здесь под определением «архаичные» подразумевается «сильно устаревшие», а никак не связанные с почти одноименной ветвью Жизни, тем более, что очередность появления на Планете двух первых ветвей Жизни остается под вопросом (см. ниже).

ультрацентрифугирования в градиенте плотности хлористого цезия был определен в пределах от 65,4 до 67,4% [Brock, Freeze, 1969]. Позднее отечественными авторами было показано, что GC-состав отдельных штаммов рода *Thermus* варьирует от 65,3 до 70,8% [Александрушкина, Егорова, 1978], что на самом деле следует отнести к погрешностям измерения ввиду того, что, как сейчас известно из результатов полногеномного секвенирования представителей рода *Thermus*, GC-состав их ДНК, который в этом случае рассчитывается арифметическим путем, даже от вида к виду меняется весьма незначительно. Так, например, по 69% GC-пар характерно для целого ряда штаммов *T.thermophilus*, такое же содержание азотистых оснований присуще ДНК *T.igniterrae* и *T.oshimai*, у *T.aquaticus* оно чуть ниже - 68 %. Приведенные примеры свидетельствуют о довольно большой неточности определения содержания GC-пар косвенными и даже прямыми методами, к которым относятся уже упоминавшееся аналитическое ультрацентрифугирование, плавление ДНК, проточная цитофлуорометрия, бумажная, тонкослойная или иная хроматография.

С развитием молекулярно-биологических методов, в частности ДНК/ДНК-гибридизации на твердой фазе, появилось требование, согласно которому необходимо определять уровни гомологии с другими уже описанными видами. Все бы ничего, но этот метод довольно капризен (что уже давно отмечается многими авторами [Konstantinidis, Tiedje, 2005; Goris et al., 2007]) в том смысле, что в разных руках (у разных экспериментаторов) одни и те же сравниваемые образцы могут показать весьма отличающиеся уровни сходства нуклеотидных последовательностей из-за практически невозможной полной стандартизации условий гибридизации, главным образом стадий удаления присутствующих в растворе несвязавшихся меченых проб, в качестве которых выступают разноразмерные фрагменты ДНК одного из сравниваемых видов, тогда как ДНК другого - фиксирована на мембранном фильтре.

Казалось бы, давно нужно отказаться от этих критериев и ограничиться секвенированием генов рРНК, дающих куда более полную информацию о родстве сравниваемых микроорганизмов. Справедливости ради следует отметить, что для разных групп этих организмов установлены несколько отличающиеся требования (ныне действующие) при описании новых видов и родов. Так, например, для аэробных спорообразующих бактерий секвенирование гена 16S рРНК и ДНК/ДНК-гибридизация входят в группу основных требуемых характеристик, тогда как GC-состав рекомендован как дополнительный показатель [Logan et al., 2009]. При описании новых видов, относящихся к бактериям семейства галомонад - *Halomonadaceae*, наряду с секвенированием гена 16S рРНК и ДНК/ДНК-гибридизацией, GC-состав также включен в группу основных характеристик [Arahal et al., 2007].

При этом мы абсолютно уверены, что в связи с бурным развитием полногеномного секвенирования не за горами то время, когда новый вид бактерий будет приниматься за новый только после установления всей последовательности его генома, что сразу снимет вопросы о GC-составе (поскольку он неизбежно станет известным абсолютно точно) и прочих характеристиках на основе ДНК. Более того, можно не сомневаться, что весьма скоро при работе со штаммами микроорганизмов как из групп прокариот, так и архей вместо секвенирования генов 16S рРНК абсолютной нормой станет установление полной нуклеотидной последовательности их геномов, причем это будет касаться не только вновь описываемых видов и родов бактерий.

* * *

Прежде чем перейти в следующей подглавке к рассмотрению биоразнообразия высших жизненных форм, возможно, следует здесь несколько задержать внимание читателей на вопросе, откуда те взялись. Сейчас общепринятой считается точка зрения, что на Земле существует три ветви Жизни: прокариоты, эукариоты и археи, причем последние, будучи фактически микроорганизмами, по некоторым чертам своей генетической и биохимической организации напоминают эукариотические формы (или, точнее, наоборот). Так, в частности, можно сказать, что нынешняя основа основ всего живого в виде механизма репликации ДНК довольно четко делится на прокариотический тип и на архейно-эукариотический [Forterre, 2013]. Причем в этой же работе высказывается предположение, что их общий предок имел РНК-геном. Ранее этим же автором было высказано предположение, что пре-кариотические организмы могли возникнуть в результате поглощения некими бактериями-археями из группы *thaumarcheon* каких-либо прокариотических бактерий, сопровождаемого инвазией ретровирусов [Forterre, 2011]. Другие авторы, придерживаясь в целом сходной точки зрения, дополнительно считают, что в качестве предков будущих митохондрий выступили некие α-протеобактерии, напоминающие тогда неких паразитов периплазмы пре-кариотической клетки, при этом, предполагая, что именно прокариоты, а не архебактерии были первыми организмами на нашей

Планете [Vesteg, Krajcovic, 2011]. Собственно и ранее высказывались различные соображения относительно взаимосвязей этих трех ветвей Жизни, как совпадающие с вышеупомянутыми, так и совсем иные [Martin, Muller, 1998; Poole, Penny, 2007; Cox et al., 2008; Lake et al., 2009 и др.].

Большой вопрос, насколько серьезно можно рассуждать, какие из этих организмов были первыми, если даже нет единого мнения относительно климата добиотической Земли, который большей частью ученых принимается за весьма жаркий с кипящим океаном воды, тогда как встречаются и совсем противоположные мнения. Так, выдвигаются предположения, что молодое Солнце было не таким ярким, как сейчас, и вода на нашей планете в самом начале ее существования была замерзшей и локально подтаивала лишь в ходе массовых «бомбардировок» Земли космическими болидами, особенно интенсивными в период между 4 и 3,6 миллиардами лет тому назад, благодаря чему в результате повторяющихся циклов оттаивания/замораживания Жизнь на Земле и появилась [Bada et al., 1994]. Но при этом даже, если это было и так, все же не надо забывать, что ядро Земли из-за высокого давления было и тогда, скорее всего, горячим и даже раскаленным, и от него по трещинам земной коры и через кратеры поднималась и выходила на поверхность лава, также способствующая растоплению первольда, принесенного на Землю кометами водного типа. Причем процессы извержения сопровождались выходом расплавов разнообразных химических соединений, а также газов, способных создать со временем первые органические молекулы. Одно можно только сказать с уверенностью, что у самых первых протоорганизмов[74], ДНК (или, вероятнее, РНК) должна была включать некий случайный (до некоторой степени, конечно же) набор всех четырех нуклеотидов, но представленных, скорее всего, не в равных пропорциях, а несколько обогащенных более прочными GC-парами, поскольку Жизнь все же зарождалась в относительно высокотемпературных условиях, где разнообразные каталитические реакции протекали быстрее. Так, предполагается, что благодаря температурной конвекции в гидротермальных порах, например, арагонита происходило в некоторых местах резкое повышение концентрации тех же нуклеотидов, что в итоге создавало нужные условия формирования полимерных молекул, ставших основой современной Жизни [Baaske et al., 2007].

С изменением (смягчением) климата появилась возможность заселения новых мест обитания, в большей степени благоприятствующих существованию и размножению микроорганизмов, что не могло не сказаться на их преобразовании (отчасти упрощении, отчасти усложнении) и увеличившимся биоразнообразии. В какие-то моменты древней истории Земли действительно могли происходить события захвата архебактериями эубактерий с образованием пре-кариотической клетки, причем доминирующим оказывался (чаще) геном архебактерий, который обособился с помощью подходящих эубактериальных компонентов, сформировавших некое подобие ядерной мембраны. Впоследствии геном архебактерии превратился в ядерный с соответствующей оболочкой, тогда как геном эубактерии-донора мог со временем фактически редуцироваться, передав часть своей генетической информации в новообразование в виде ядерного генома. Такой сценарий, по крайней мере, объясняет, почему организация ДНК и ее функционирование у эукариот больше напоминает архейный тип, тогда как с устройством прокариотического генома общего имеется гораздо меньше. При этом считается, что хлоропластный геном происходит от цианобактерий, некогда находящихся в симбиотических отношениях с пре-кариотической клеткой. С предками митохондрий дело обстоит несколько хуже, хотя в литературе высказываются предположения, что митохондриальный геном ведет начало от внутриклеточных паразитов типа риккетсий [Emelyanov, 2003; 2003a]. Но пока даже полногеномное секвенирование вкупе с биоинформатикой со всеми этими вопросами до конца разобраться не позволяет, хотя дало возможность заметно приблизиться к пониманию того, как развивалась Жизнь на нашей Планете и дальнейшего прогресса в этих вопросах вполне можно ожидать.

Пожалуй, здесь нельзя обойти вниманием довольно уникальный микроорганизм *Thermotoga maritima*, геном которого был секвенирован еще в 1999 г. [Nelson et al., 1999]. Проведенный в этой работе анализ нуклеотидных последовательностей показал, что около четверти всех генов данной бактерии имеют, скорее всего, архейное происхождение, причем эти гены (в количестве 81) располагаются в геноме в виде 15 обособленных кластеров, имеющих размеры от 4 до 20 т.п.н., что подтверждает горизонтальный перенос данных генов между представителями разных ветвей Жизни. Фактически микроорганизм *Th. maritima* можно рассматривать как имеющий смешанный геном доминирующего эубактериального типа. Еще задолго до завершения секвенирования геномной

[74] Вопрос только - кто же ими был?

ДНК *Th. maritima* указывалось на своеобразие этого вида и делалось предположение, что все эубактерии возникли из их термофильных предшественников [Achenbach-Richter et al., 1987], но только определение нуклеотидных последовательностей полных геномов этого и некоторых других видов микроорганизмов в действительности подтвердило их своеобразное генетическое родство.

Как бы то ни было, все равно можно считать, что и сроки, и даже само происхождение как эубактерий, так архебактерий покрыто глубокой тайной, разгадать которую, скорее всего, до конца (точнее от начала) никогда не удастся, ввиду отсутствия нужных окаменелостей и их слишком большого возраста, даже если бы таковые нашлись. Впрочем, не исключено, что помочь здесь сможет астробиология и секвенирование форм Жизни на том же Марсе (если они все же там есть и при этом похожи на те, что имеются тут у нас на Земле), чтобы можно было провести некую аналогию. Однако предпринять попытки для подтверждения гипотезы возникновения именно таким описанным выше образом первоядерных организмов можно уже сейчас, фактически заново конструируя искусственную пре-кариотическую клетку из соответствующих архе- и эубактерий, что в настоящее время технически вполне осуществимо, как это станет видно при дальнейшем изложении в третьей части данной статьи в разделе про системную и синтетическую биологии.

* * *

Помимо занимательных вопросов происхождения и эволюции бактерий, в современной микробиологии и биотехнологии важное место отводится также генетической паспортизации штаммов микроорганизмов, поскольку немалое их число являются продуцентами каких-либо ценных продуктов и коммерчески используются. Можно не сомневаться, что ДНК-паспортизация бактериальных штаммов продолжится и в будущем. В основе такого генотипирования лежит все тот же полиморфизм молекул ДНК, который у бактерий не столь велик по сравнению с высшими организмами и соответствует их небольшим геномам. Один из подходов к генотипированию бактерий заключается в специфичном рестриктазном расщеплении их ДНК редкощепящими ферментами и последующим разделением получающихся фрагментов довольно крупного размера с помощью гель-электрофореза в пульсирующем поле [Goering, 2010]. Была предложена специальная система улучшения интерпретации таких ДНК-паттернов, служащая, в том числе, для определения степени филогенетического родства микроорганизмов [Duck et al., 2003]. Однако данный метод требует для анализа весьма высокополимерной ДНК, что сразу приводит к резкому усложнению процедур по ее выделению из бактерий и по дальнейшему обращению с ней. Более того, ввиду разделения этим методом очень крупных фрагментов ДНК, совершенно невозможно точно оценивать их размеры и по существу сравнение штаммов ведется по принципу «похожи / не похожи». Да, и количество требуемой для такого анализа ДНК довольно велико.

Появление и бурное развитие такого мощного метода, как ПЦР заставило микробиологов, в том числе, занимающихся систематикой микроорганизмов обратить на него свое внимание. Так, в настоящее время существует достаточно много вариантов выявления полиморфизма бактериальной ДНК с помощью ПЦР. Одним из наиболее простых и широко используемых можно считать RAPD-метод случайно амплифицируемой полиморфной ДНК (Random Amplified Polymorphic DNA). Немало и других методов, схожих с ним, которые приводят к образованию множественных полос ДНК, позволяющих детектировать различия подчас даже между близкородственными штаммами. Но главным недостатком таких способов выявления полиморфизма ДНК является их довольно плохая отмечаемая многими авторами даже внутрилабораторная воспроизводимость, не говоря уже о межлабораторной. Что касается упомянутого выше способа идентификации бактерий путем секвенирования генов 16S или 23S рРНК, то ввиду их слишком высокой эволюционной консервативности этот метод не позволяет обнаруживать различия между штаммами одного вида бактерий. Поэтому сравнение нуклеотидных последовательностей генов рРНК микроорганизмов имеет серьезное значение, главным образом, при построении филогенетических древ и установлении таксономического статуса того или иного вида бактерий и не может служить в качестве паспортизирующего для отдельных штаммов.

Таким образом, следует признать, что удобного и универсального метода генотипирования штаммов бактерий пока не предложено. При этом секвенирование полных геномов бактерий и сопоставление их нуклеотидных последовательностей, безусловно, позволит однозначно идентифицировать любые штаммы, но эта процедура даже тогда, когда станет очень быстрой и дешевой, вряд ли будет пригодна для ДНК-паспортизации всех (очень многих) штаммов в силу большого объема хранимой информации (в том числе ненужной для

ДНК-паспортизации), среди которой все равно придется выделять наиболее значимую. Так что в вопросе будущего ДНК-паспортизации штаммов бактерий многое пока неясно, и нет уверенности в том, как она будет производиться в будущем, хотя нельзя исключать, что будут целиком сравниваться полные геномы микроорганизмов. Тем не менее, позволим себе остановиться несколько подробнее на одном довольно перспективном подходе к генотипированию микроорганизмов.

Так, предложенный уже достаточно давно [van Steenbergen et al., 1995] метод идентификации микроорганизмов путем концевого мечения рестриктазных фрагментов ДНК (КМРФ) довольно прост по исполнению и при этом весьма информативен (рис. 5).

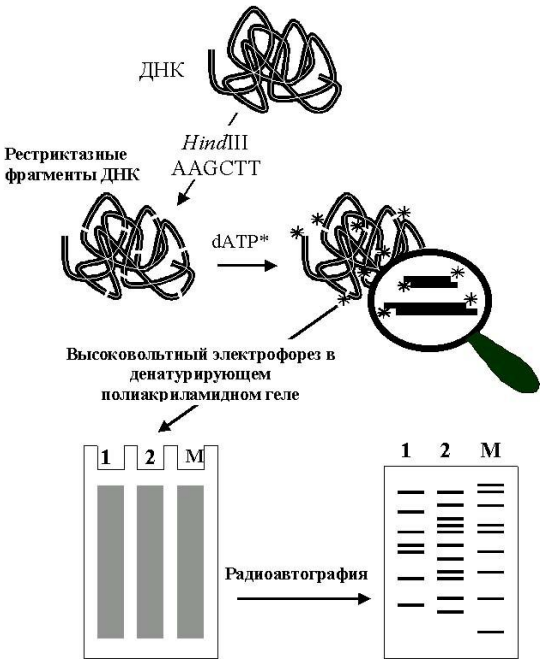


Рис. 5. Схема проведения КМРФ бактериальной ДНК

Главное отличие метода КМРФ от других подходов по электрофоретическому разделению рестриктазных фрагментов ДНК каких-либо организмов для целей геносистематики заключается в использовании полиакриламидного геля с денатурирующим агентом в виде мочевины, что позволяет разделять одноцепочечные фрагменты ДНК и определять их истинные размеры с точностью до нуклеотида. Кроме высокоточного определения размеров рестриктазных фрагментов ДНК, данный подход направлен и на уменьшение числа анализируемых фрагментов, что также очень важно. Именно это сочетание используемой для расщепления ДНК рестрикционной эндонуклеазы с гексануклеотидным сайтом узнавания/расщепления, а также полиакриламидного геля с денатурирующим агентом для электрофоретического разделения одноцепочечных рестриктазных фрагментов и определяет уникальные возможности данного метода.

Некоторое время спустя в части интерпретации данных этот метод был нами значительно усовершенствован [Баймиев и др., 1999; Baymiev et al., 2007], что позволило присваивать штаммам микроорганизмов уникальные генетические штрих-коды, формируемые на основе электрофоретического разделения рестриктазных фрагментов ДНК, как это показано на рис. 6 и 7.

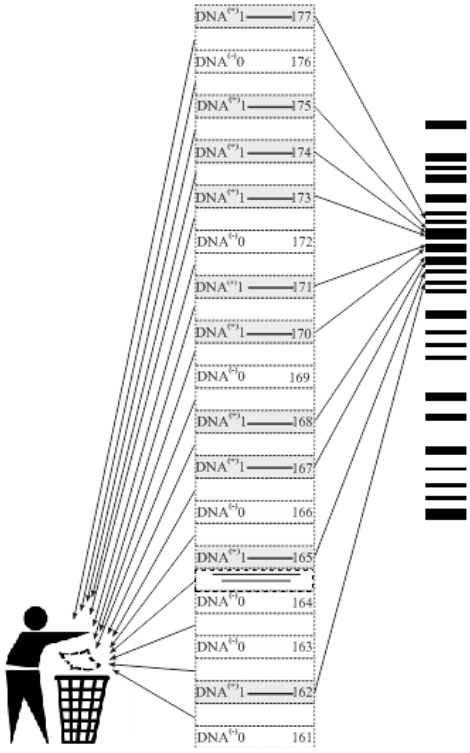


Рис. 6. Принцип формирования генетического штрих-кода штаммов микроорганизмов на основе гель-электрофоретического разделения подвергнутых температурной или щелочной денатурации и потому одноцепочечных рестриктазных фрагментов ДНК

Для большей наглядности как происходит создание генетического штрих-кода на рис. 5 приведен некий условный участок радиоавтографа геля, однако формирование полос штрих-кода сходным образом может быть осуществлено и при разделении флуоресцентно-меченных по концам фрагментов ДНК с помощью автоматического

секвенатора ДНК. Здесь радиоавтограф геля был нами условно подразделен на «ДНК-ячейки» и на промежутки между ними, которые не несут никакой информации и потому показаны, как отправляемые в мусорную корзину. ДНК-ячейки в свою очередь подразделены на содержащие рестриктазный фрагмент ДНК (ДНК$^+$-ячейки) и на те, где такой фрагмент отсутствует (ДНК$^-$-ячейки). Поскольку штрих-код несет в себе бинарную информацию в виде перемежающихся «единиц» (черных зон) и «нулей» (белых зон), то ДНК$^+$- и ДНК$^-$-ячейки как раз соответствуют компьютерным «1» и «0», обеспечивая привычную исчерченность штрих-кода, показанного на рисунке справа. При этом вероятностное нахождение в какой-либо ячейке двух одинаковых по размеру рестриктазных фрагментов ДНК (или даже большего их числа), но ведущих свое происхождение из разных мест генома, рассматривается как обычная ДНК$^+$-ячейка, поскольку для формирования данного генетического штрих-кода важны качественные, а не количественные показатели.

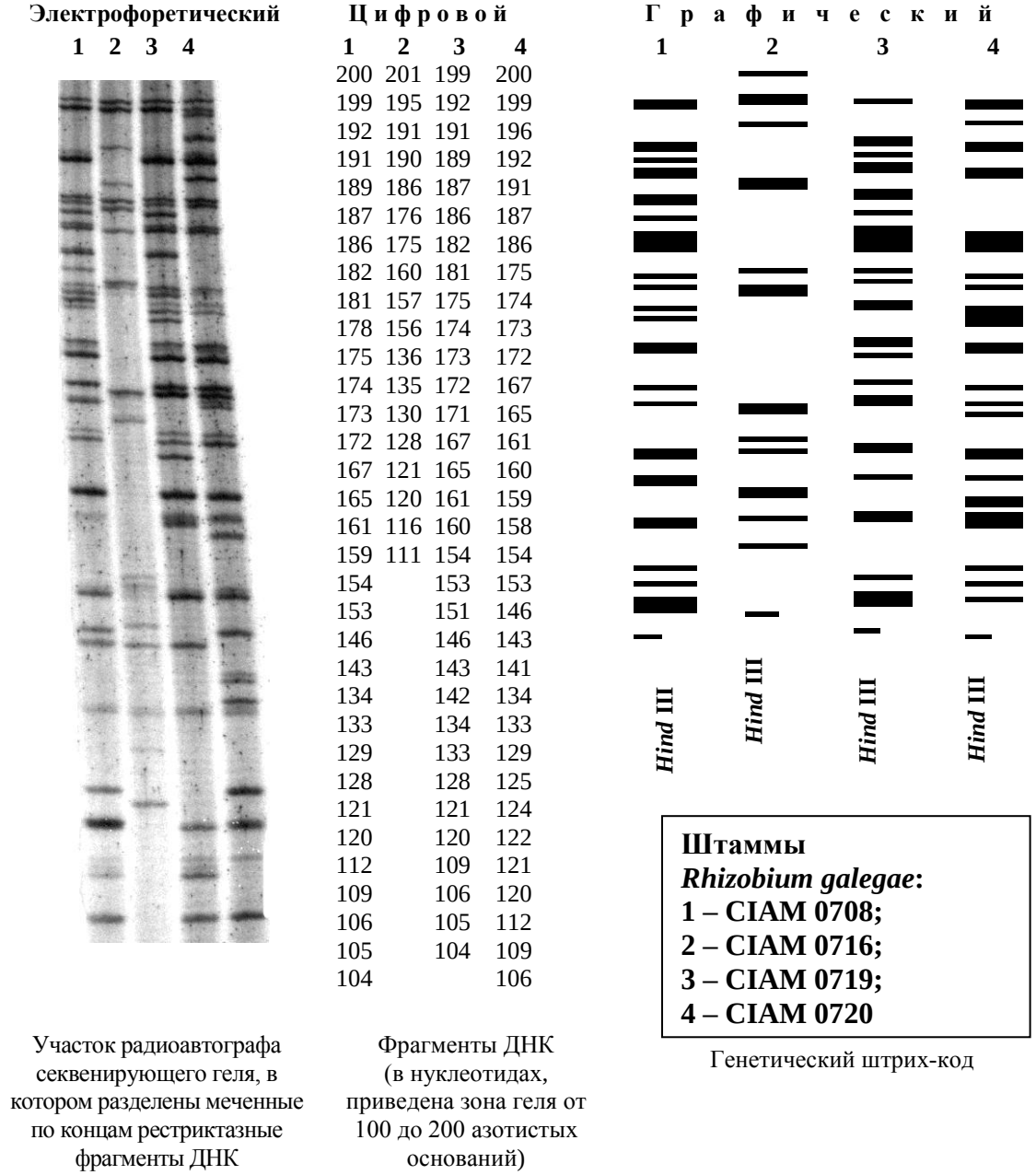


Рис. 7. Генетические портреты некоторых штаммов азотфиксирующих бактерий *Rhizobium galegae*, созданные путем расщепления их ДНК рестрикционной эндонуклеазой *Hind*III

Теоретическое число комбинаций рестриктазных фрагментов ДНК разного размера, на основе которых могут создаваться генетические «портреты» бактерий, рассчитывается по формуле Бернулли -

$$C_m^n = \frac{m!}{n!(m-n)!},$$

где в нашем случае C - общее число вероятностных комбинаций рестриктазных фрагментов бактериальной ДНК определенного размерного диапазона, m - число всех анализируемых в выбранном диапазоне ДНК-ячеек, n - число ДНК$^+$-ячеек, (m-n) - число ДНК$^-$-ячеек.

Случайное совпадение данных характеристик микроорганизмов теоретически может происходить с частотой, по крайней мере, один случай на 10^{50} исследуемых штаммов, если генетический «портрет» бактерии будет основан на анализе участка секвенирующего геля, в котором разделены фрагменты ДНК, укладывающиеся в диапазон длин от 51 до 250 нуклеотидов при условии, что для каждого штамма таких фрагментов будет приблизительно от четверти до трех четвертей от максимально возможной представленности, под чем подразумевается занятие всех ДНК-ячеек и соответственно превращение их всех в ДНК$^+$-ячейки. То есть в данном диапазоне длин рестриктазных фрагментов ДНК из 200 ДНК-ячеек можно ожидать 50-150 ДНК$^+$-ячеек с любыми размерами как, например 53, 57, 75, 76, 84 ... 221, 224, 233 и т.д. нуклеотидов. При этом количества рестриктазных фрагментов в указанном диапазоне и сами длины фрагментов для разных бактерий действительно могут весьма заметно меняться. В первую очередь это зависит от того, каких среди значащих нуклеотидов больше - GC- или AT-пар в сайте узнавания используемой гексануклеотидной рестрикционной эндонуклеазы и от общего GC-состава ДНК исследуемого микроорганизма и, кроме того, от размера генома микроорганизма. Проведенное нами исследование *in silico* целого ряда полных геномов различных микроорганизмов показало абсолютную пригодность такого подхода и присутствие в анализируемой зоне, если не оптимального, то вполне подходящего числа фрагментов ДНК [Усанов и др., неопубл.; Usanov et al., unpublished].

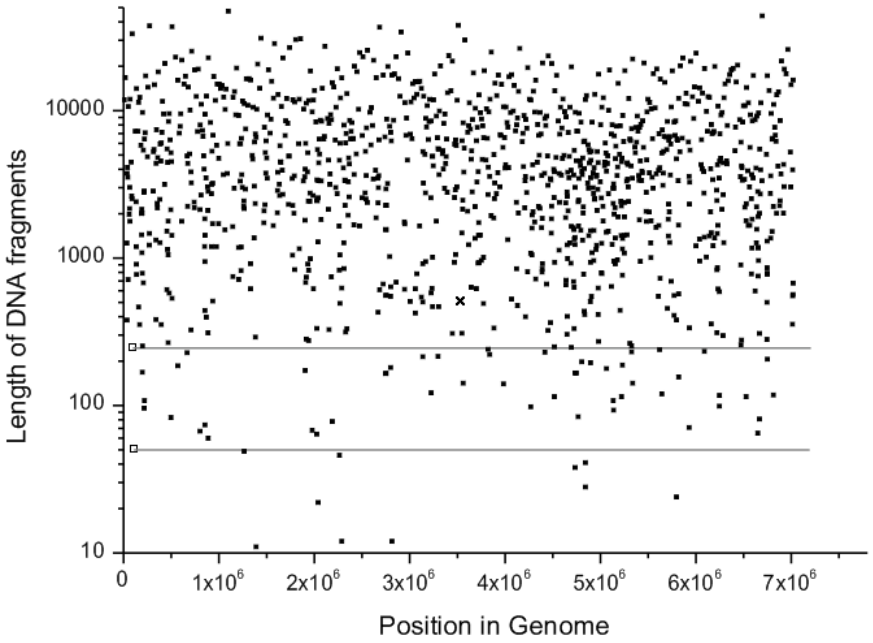


Рис.8. Распределение рестрикционных фрагментов в геноме штамма MAFF330099 *Mesorhizobium loti* ДНК которого была «расщеплена» *in silico* рестрикционной эндонуклеазой *Hind*III. По оси абсцисс отложена длина (позиции) геномной ДНК, а по оси ординат - размеры рестриктазных фрагментов. Бледными горизонтальными линиями выделена зона, соответствующая длинам фрагментов ДНК от 51 до 250 п.н.

Так, на рис.8 показано распределение в геноме штамма MAFF330099 *Mesorhizobium loti* (секвенированного ранее японскими авторами [Kaneko et al., 2000]) *Hind*III-фрагментов, при этом по оси абсцисс отложена длина всей ДНК микроорганизма и точки обозначают места генома, где обнаружены те или иные *Hind*III-фрагменты, а по оси ординат - размеры образующихся рестриктазных фрагментов. Как и следовало ожидать, значительная часть всех *Hind*III-фрагментов у данной бактерии имеют длину около 4 т.п.н., поскольку по теории вероятности сайт гексануклеотидной рестриктазы может встречаться через каждые 4096 п.н. (4^6). Можно видеть, что в представляющей наибольший интерес для генетического штрих-кодирования (выделенной) зоне от 51 до 250 п.н. *Hind*III-фрагментов обнаруживается около 50, но часть из них совпадают по размеру, что хорошо видно уже из рис. 8, где показан сгенерированный для данного штамма его генетический штрих-код.

Рис. 9. Сгенерированный *in silico* генетический штрих-код штамма MAFF330099 *Mesorhizobium loti*, включающий информацию о его *Hind*III-фрагментах ДНК как геномной, так и плазмидной природы. Носящая технический характер двояко ломаная линия (один полный элемент которой соответствует 10 нуклеотидам) показывает границы анализируемой зоны рестриктазных фрагментов от 51 до 250 нуклеотидов.

На рис. 9 в строке с одинарными ДНК$^+$-ячейками представлен сгенерированный *in silico* генетический штрих-код штамма MAFF330099 *Mesorhizobium loti*, включающий информацию о геномной ДНК (рис.8), а также данные о *Hind*III-фрагментах ДНК двух мегаплазмид, свойственных этому штамму, для которых нами проведен аналогичный анализ по распределению сайтов рестрикции данного фермента в них, свидетельствующий о наличие небольшого числа таких плазмидных фрагментов. При этом некоторые ДНК$^+$-ячейки несут в себе дополнительные «сигналы» (в виде присутствия в них двух[75] и даже трех[76] рестриктазных фрагментов), но, как отмечалось уже выше, наличие множественных фрагментов ДНК в таких ДНК$^+$-ячейках никак не сказывается на формировании генетического штрих-кода, поскольку для него важны качественные показатели типа «да/нет» и потому принято деление ДНК-ячеек всего на две разновидности - ДНК$^+$- и ДНК$^-$-ячейки с присвоением им компьютерных «единиц» и «нулей», а в графическом варианте черных и белых зон соответственно.

Следует отметить, что указанное количество возможных комбинаций (10^{50}) для оговоренного размерного диапазона рестриктазных фрагментов ДНК, которое могут обеспечить штаммы микроорганизмов, возникает при условии использования только одной рестрикционной эндонуклеазы. Однако для однозначной «ДНК-паспортизации» близкородственных штаммов этого может оказаться недостаточно, но увеличивать диапазон длин рестриктазных фрагментов для анализа менее целесообразно, нежели применять вторую рестрикционную эндонуклеазу. Тем более, что микроорганизмы-носители генов рестрикции-модификации первого фермента или его изошизомеров не будут, например, этим ферментом расщепляться вовсе и для них обязательно нужна другая рестрикционная эндонуклеаза[77].

Так, в случае, если одна из используемых рестрикционных эндонуклеаз за счет имеющейся у какого-либо исследуемого штамма соответствующей системы рестрикции-

[75] 58-я, 93-я, 105-я и 106-я.

[76] 64-я.

[77] Вероятность нахождения в каком-либо штамме сразу двух систем рестрикции-модификации, совпадающих с обеими использующимися для генетического штрих-кодирования рестрикционными эндонуклеазами, представляется практически невероятным событием.

модификации и тем самым метилирования некоторых азотистых оснований не будет способна расщеплять анализируемую ДНК, то возникающая «пустая» область генетического штрих-кода, которая должна была бы наполняться данными по этому ферменту, также будет служить дополнительной характеристической информацией, хотя и снижающей общее число комбинаций, но, тем не менее, все равно обеспечивающей уникальность такого штамма. Еще одной причиной использования второго фермента служит то, что содержание G+C/A+T-пар в ДНК у разных микроорганизмов весьма заметно варьирует, и для достижения оптимальных результатов генотипирования любых штаммов абсолютно всех бактерий, независимо от состава их геномов, желательно использовать рестрикционные эндонуклеазы с преимущественными представленностями GC-пар в сайте узнавания/расщепления для одной и AT-пар для второй, например *Nco*I (C▼CATGG) и *Hind*III (A▼AGCTT) соответственно. При этом использование пары ферментов соответственно с одними AT- и GC-нуклеотидами в сайтах узнавания, например, таких как *Ase*I (AT▼TAAT) и *Nar*I (GG▼CGCC) представляется нежелательным и даже неприемлемым. Еще одним важным моментом является использование для концевого мечения рестриктазных фрагментов, полученных под действием данных ферментов, разных дНТФ, которые могут быть мечены подходящими флуорохромами, что позволит в секвенирующем гель-электрофорезе фрагменты ДНК, образованные действиями обеих рестриктаз (по отдельности) разделять вместе.

Благодаря второму ферменту для каждого штамма будет происходить создание второго штрих-кода (точнее становящегося продолжением первого), что таким образом увеличит число возможных комбинаций рестриктазных фрагментов для подавляющего большинства микроорганизмов до гигантского числа в гугол комбинаций (10^{100}). В целях более компактного отображения такого сдвоенного генетического штрих-кода можно использовать двумерное размещение информации, получившее в последние годы заметное развитие. Также в генетических штрих-кодах, безусловно, нужна дополнительная реперная информация, носящая некий информационно-технический характер, в том числе, обозначающая стартовую точку отсчета (считывания) штрих-кода для правильного позиционирования полос.

Прежде чем перейти к рассмотрению биоразнообразия и методов его оценки на уровне ДНК у высших организмов, включая ДНК-идентификацию человеческих особей, все же необходимо уделить вопросу генотипирования бактерий еще немного внимания. Несколько забегая вперед, надо отметить, что, например, для человека ДНК-идентификация индивидов в настоящее время ведется по конкретным локусам, которые, хотя и высокополиморфны, но в том или ином виде имеются у всех людей. В будущем в вопросе ДНК-идентификации личности неизбежно должен будет произойти переход на иной тип полиморфных маркеров, но и они в любом случае будут фланкированы одинаковыми участками генома. Поэтому как для нынешних, так и для будущих подходов можно для всех людей использовать одни и те же (соответствующие) наборы олигонуклеотидных праймеров. Сходный принцип действует и для ДНК-паспортизации ценных пород животных. С ценными сортами растений пока дело обстоит сложнее. Что же касается бактерий, то они все весьма разные и выбрать для них одинаковые участки ДНК, отжиг одних и тех же наборов праймеров на которых обеспечивал бы штаммовую специфичность для всех видов (да и просто обычную амплификацию), не представляется возможным. Абсолютно невозможным! В то же время проводить ДНК-паспортизацию отдельно, например, штаммов кишечной палочки или штаммов бацилл, для чего требовались бы свои комплекты праймеров по сути на каждый род бактерий, представляется неразумным. Таким образом, использование метода КМРФ (хотя и не лишенного некоторых недостатков, о чем мы фактически пока умолчали, но они не столь критичны и могут быть преодолены различными способами) позволяет ДНК-паспортизировать любые бактерии, не думая об их нуклеотидных последовательностях.

Все же, видимо, стоит хотя бы кратко упомянуть недостатки КМРФ-генотипирования штаммов бактерий. Главным, пожалуй, может являться неполнота расщепления ДНК рестрикционной эндонуклеазой, но заведомый избыток фермента и высокая степень очистки ДНК должны преодолевать эту ситуацию. К тому же после электрофоретического разделения продуктов по распределению фрагментов ДНК почти наверняка можно будет сделать вывод об имевшей место «недорестрикции» и внести коррективы в повторную процедуру. Другим недостатком является необходимость для проведения КМРФ наличия довольно значительного количества ДНК, что, впрочем, при паспортизации культивируемых штаммов никакой проблемы не представляет. Хуже дело обстоит с принципиально некультурабельными микроорганизмами, но в

таких случаях крохотные количества выделенной из них ДНК можно амплифицировать до требуемых количеств в изотермических условиях с помощью ДНК-полимеразы фага phi29 и таких методик уже предложено немало [Silander, Saarela, 2008; Hongoh, Toyoda, 2011 и др.]. Поскольку при проведении КМРФ для рестриктазного расщепления достаточно довольно коротких фрагментов ДНК, то таковые с большей вероятностью и без каких-либо проблем будут нарабатываться при использовании phi29 ДНК-полимеразы. Еще один недостаток, свойственный не только методу КМРФ, но и всем прочим, рассчитанным на выявление только части полиморфных локусов, а не на получение информации о всем геноме, включая плазмидные ДНК, заключается в возможном пропуске различий между близкородственными штаммами, поскольку произошедшая в одном из них важная замена нуклеотида(ов) может, например, прийтись на совсем другую (неанализируемую) область(и) генома. Известно, что многие бактерии несут плазмиды и даже мегаплазмиды, которые также могут вносить свой вклад в генетический штрих-код, генерируемый путем КМРФ, но при этом плазмиды имеют свойство относительно легко передаваться/теряться, причем это может происходить с внесением или не внесением изменений в штрих-код в зависимости от нуклеотидной последовательности плазмид, хотя штамм, вероятно, приобретет другие свойства. Во всех этих случаях только полногеномное секвенирование даст возможность выявить имеющиеся между такими близкородственными штаммами различия, включая наличие плазмидной ДНК, но тогда возникают сложности иного порядка (см. выше). При этом, владея информацией обо всей ДНК какой-либо бактерии, ей «рестрикцией» *in silico* может быть присвоен аналогичный штрих-код, что выше нами уже продемонстрировано. Итак, использование метода КМРФ пока еще (до того как начнется поистине массовое секвенирование полных геномов всех штаммов) представляется весьма удобным. Да, возможно оно останется вполне приемлемым (применимым) и после того как такое секвенирование начнется. Т.е. данный метод обеспечивает единый подход для ДНК-паспортизации любых штаммов из всех групп бактерий, что принципиально невозможно для всех высших организмов, хотя для них и стоит задача паспортизации отдельных особей, пород, сортов, не столь актуально, чтобы все виды высших организмов были ДНК-паспортизированы на основе одинаковых участков их геномов, поскольку вполне достаточно внутривидовой ДНК-паспортизации/ДНК-идентификации. Тем не менее, свои ДНК-штрих-коды для некоторых эукариотических организмов уже предложены (см. следующую подглавку), но они способны разграничивать лишь таксономические отношения на уровне семейств, родов и, если повезет, то и на уровне некоторых видов, в отличие от предлагаемого нами подхода, который позволяет уверенно генотипировать конкретные штаммы микроорганизмов. Или на основе иного принципа вести ДНК-идентификацию/ДНК-паспортизацию людских особей, к вопросу о чем вернемся в третьей части статьи.

Биоразнообразие высших организмов

Что касается вопроса о биоразнообразии всего Живого на нашей Планете надо сказать следующее. Действительно, только за последние десятилетия исчезло немало видимых глазом видов живых организмов и продолжает исчезать. Причем в связи с ухудшающимися природными условиями (т.е. неизбежно удаляющимися от первозданных, точнее, от тех, что существовали до активной деятельности человека[78]) процесс исчезновения отдельных видов, скорее всего, будет только нарастать, и для максимально возможного поддержания биоразнообразия первым шагом должно стать описание всех видов, чтобы знать что необходимо сохранять. Однако применение стандартных подходов на основе морфологических признаков для классификации часто оказывается недостаточным, и в этом случае на помощь приходят молекулярно-биологические методы, главным образом, секвенирование отдельных генов или иных участков геномов разных организмов.

Ботаниками для целей систематики уже давно взято на вооружение секвенирование так называемых внутренних транскрибируемых спейсеров рДНК (ITS - Internal Transcribed Spacers), как сочетающих в себе достаточно высокую

[78] Все же надо признать, что огромное множество видов из так называемых доисторических флоры и фауны, оказались исчезнувшими (и мы даже приблизительно не знаем сколько, поскольку они исчезли для нас практически совсем бесследно) без всякого на то влияния человека (еще задолго до его появления на Планете), поскольку были вытеснены их конкурентами, гораздо лучше приспособившимися, в том числе, к новым климатическим и иным условиям мест обитания, однако сейчас главным фактором (относительно быстрого) исчезновения видов все же служит хозяйственная и прочая деятельность человека.

эволюционную консервативность, так и довольно заметную вариабельность, благодаря чему стали возможны уточнение/ревизия таксономических положений отдельных видов. И эта работа активно продолжается, в том числе, в наступившую новую эру полногеномного секвенирования, причем секвенирование геномов целиком уже принесло свои неожиданные «плоды». Так, в результате секвенирования всех копий ITS у конкретных видов обнаружилась высокая интрагеномная вариабельность, из-за которой, например, минорный вариант этой последовательности одного вида растений полностью совпадает с мажорным вариантом другого вида растений этого же семейства [Song et al., 2012]. Это сразу вызывает ряд вопросов по полученным прежде результатам секвенирования этих участков рДНК у разных растений для геносистематических построений. Однако применявшиеся технологии фактически позволяли секвенировать именно мажорные последовательности, имеющие все же несколько большую прогностическую значимость. Однако как бы то ни было, вопросы остаются, в том числе, куда и как двигаться дальше. Возможно, для целей систематики предпочтительнее сравнивать несколько однокопийных консервативно/вариабельных генов, которые с помощью биоинформатики еще предстоит выбрать после полногеномного секвенирования большого числа видов растений, хотя и уже сейчас некоторые такие гены успешно секвенируются для геносистематики. Для особенно трудных случаев при определении родства некоторых видов и других таксонов обязательно будет проводиться секвенирование полных геномов (включая геномы органелл - митохондрий и хлоропластов) отдельных представителей и соответственно сравнение их последовательностей друг с другом и с прочими близкими и дальними сородичами, что в итоге позволит к 2030 г. воссоздать довольно полную и точную картину родственных взаимоотношений в царстве растений. Равно как и в других группах живых организмов, поскольку к тому времени будут весьма широко секвенироваться полные геномы и пластомы представителей так называемой дикой природы с целью установления/подтверждения/ревизии их таксономического положения и филогенетических связей.

До некоторой степени схожая с ITS ситуация наблюдается с данными, получаемыми при реализации существующего с 2004 г. международного проекта The Consortium for the Barcode of Life, в ходе которого определяется последовательность митохондриального гена цитохром С оксидазы, и, исходя из полиморфизма 5’-участка данного гена, строится цветной ДНК-штрих-код с полосками-нуклеотидами соответствующего цвета (A – зеленый, C - синий, G - черный, T - красный). Однако будучи изначально предложенным [Hebert et al., 2003; 2003a] для разграничений видов и родов животных организмов, данный подход не вполне применим к некоторым другим высшим организмам и растениям [Cowan, Fay, 2012]. Есть свои проблемы при работе с организмами с внутриклеточными симбионтами, также имеющими свой аналогичный ген цитохром С оксидазы. Не всегда работают установленные (в %) для разграничения родов и видов критерии. Многие проблемные вопросы по штрих-кодированию Жизни отмечаются в ряде обзоров [Meyer, Paulay, 2005; Шнеер, 2009; Taylor, Harris, 2012 и др.]. Есть предложения включать в ДНК-штрих-код растений информацию о последовательности вышеупомянутого ITS или хотя бы его частей [China Plant BOL Group, 2011; Hollingsworth, 2011].

Что касается нынешнего разноцветного ДНК-штрих-кода «Жизни», то, несмотря на его некоторую зрелищность, он «проигрывает» обычному двуцветному (черно-белому) и в экономии дискового пространства при хранении и в возможности быть легко считанным любым устройством, поскольку, как известно, цветопередача на разных дисплеях и принтерах не идеальна и часто заметно отличается. В настоящее время общепринятый способ оцифровки нуклеотидов заключается в присвоении им в двухбитной кодировке компьютерных нулей и единиц в следующем порядке: A - 00, C - 01, G - 10, T - 11. Может быть, Консорциуму The Consortium for the Barcode of Life стоит перейти на этот принцип оцифровки нуклеотидов (или чуть иной) и присваивать ДНК-штрих-коды (пока они еще хоть немного востребованы) всем живым организмам в более емком и все шире использующемся последнее время двумерном формате? И генов (участков ДНК) использовать для него побольше!

Возлагаемые участниками Консорциума надежды на скорое появление мобильного секвенирующего устройства, благодаря которому будет легко и просто определить последовательность нужного гена пойманного насекомого (или другого живого объекта) и с помощью интернета прямо в поле по существующим базам данных понять - новый ли это вид или уже описанный ранее[79], не лишены

[79] Для этого, впрочем, совсем не обязательно использование ДНК-штрих-кода!

некоторых оснований, поскольку подобные технологии развиваются очень быстро. Скорее всего, это может быть некое микрожидкостное устройство, сочетающее в себе сразу экстрактор ДНК *плюс* ДНК-амплификатор (требующий смену температур, или амплифицирующий фрагменты ДНК изотермически) *плюс* прибор для капиллярного электрофореза ДНК с детекцией результатов разделения. Собственно, подобные устройства уже есть, рассчитанные на электрофоретическое разделение ампликонов и даже разделение продуктов секвенирующих реакций, подготовленных методом Сэнгера. Осталось в таком комплексном, но компактном приборе обеспечить прочтение последовательности нуклеотидов и преобразование данных в подходящий формат для поиска по базам данных. При этом секвенировать в нем можно будет лишь отдельные гены или их фрагменты, но никак не цельные геномы, да и, наверное, еще с несколько большим процентом ошибок, ввиду довольно малой длины пробега разделяемых фрагментов и иных конструктивных особенностей прибора в угоду миниатюризации (см. соответствующую подглавку в третьей части статьи). И нужен ли будет такой хотя и малогабаритный, но и маломощный (уже сейчас и тем более по меркам нового времени) прибор года через три-четыре?

И так ли уж будет нужен штрих-код на основе полиморфизма одного участка (фрагмента гена) ДНК?! К тому же из митохондриального генома! Ведь как признают сами участники данного Консорциума, часто получаемые сведения не обеспечивают однозначного результата! Тем более, надо, безусловно, иметь в виду, что эволюционирование главенствующего ядерного генома идет по другим законам, нежели происходит изменение ДНК митохондрий. Так, уже довольно давно звучали голоса, что для систематики предпочтительнее использовать однонуклеотидные замены в хлоропластном геноме и в 10-100 ядерных генах [Kane, Cronk, 2008]. Исходя из этого и из быстро растущих технологических возможностей, представляется целесообразным как можно быстрее отказаться от присвоения ДНК-штрих-кода на основе одного (того или иного - любого) гена и тем более их отдельных участков в пользу обладающей гораздо большей информативностью набора из целого ряда генов из ядерного генома, сведения о которых могут добываться пусть и не в полевых условиях (а может и в полевых, только прибор будет уже не карманного размера и питаться не от батарейки), зато обеспечат полную ясность с конкретным представителем флоры или фауны. Здесь надо иметь ввиду, что, в целом, снипы весьма вариабельны, и у особей одного вида, размножающегося половым путем, они могут заметно различаться, формируя фактически персональный набор таковых. Исходя из этого, для верификации видовой и тем более родовой принадлежности исследуемых объектов необходимо вводить в комплект анализируемых снипов только те, что локализованы в эволюционно консервативных генах. Таким образом, можно предположить, что проект The Consortium for the Barcode of Life должен претерпеть, по крайней мере, существенные изменения, поскольку в нынешнем виде он и не универсален, и не однозначен, но при этом до некоторой степени может быть актуальным для систематизации живых организмов на уровне видов и родов. Но, по всей видимости, нет необходимости присваивать всевозможным живым организмам подобный единообразный ДНК-штрих-код, поскольку это фактически является самоцелью, а решить главную задачу каталогизации видов, родов и прочих таксонов живых организмов можно иначе, где главенствующим подходом через некоторое время непременно станет полногеномное секвенирование.

Послесловие ко второй части статьи

Завершая вторую часть данной статьи, необходимо заметить, что основное внимание в ней уделено перспективам или, точнее, последствиям полногеномного и полнотранскриптомного секвенирований нуклеиновых кислот. Предсказано исчезновение ряда направлений современной физико-химической биологии. Затронуты вопросы каталогизации на основе полиморфизма ДНК микро- и макроорганизмов, кроме человека, поскольку вопросы ДНК-идентификации / ДНК-паспортизации индивидов будут рассмотрены в третьей (завершающей) части данной статьи, которая будет опубликована в очередном номере данного журнала. Из приведенного ниже оглавления третьей части можно видеть, что в ней будут затронуты весьма разные вопросы технологических возможностей современной и будущей физико-химической биологии, компьютерного обеспечения, баз данных. При этом неизбежно будет осуществляться возврат в прошлое. Причем довольно далекое, а именно в 1869 год, когда впервые была открыта сама ДНК. Придется коснуться и околонаучных дел в виде различных подходов к финансированию науки и подсчета эффективности этих процессов[80].

[80] При этом просим нас простить тех, для кого наукометрия является по-настоящему научным исследованием.

Оглавление третьей части статьи

Литература, цитированная во второй части статьи

1. Александрушкина Н.И., Егорова Л.А. Нуклеотидный состав ДНК из термофильных бактерий рода *Thermus* // Микробиология. 1978. Т.47. С.250-252.
2. Баймиев Ал.Х., Чемерис А.В., Вахитов В.А. Анализ информативности некоторых современных методов идентификации полиморфизма ДНК микроорганизмов на примере симбиотических клубеньковых бактерий Rhizobium galegae // Генетика. 1999. Т.35. С.1613-1621.
3. Вахитов В.А., Чемерис А.В., Сабиржанов Б.Е., Ахунов Э.Д., Куликов А.М., Никоноров Ю.М., Гималов Ф.Р., Бикбулатова С.М., Баймиев Ал.Х. Исследование филогении *Triticum* L. и *Aegilops* L. на основе сравнения нуклеотидных последовательностей промоторных областей рДНК // Генетика. 2003. Т.39. С.5-17.
4. Голимбет В.Е., Корень Е.В. Вариации числа копий в геноме – новая страница в генетических исследованиях в области психиатрии: международный проект PsychCNVs // Журнал неврологии и психиатрии. 2010. Т. 1. С. 107-109.
5. Патрушев Л.И., Минкевич И.Г. Проблема размера геномов эукариот // Успехи биологической химии. 2007. Т.47. С. 293–370.
6. Усманова Н.М. Вариации числа копий отдельных сегментов – новая форма изменчивости генома // Цитология. 2009. Т. 51. С. 549-550.
7. Шереметьев С.Н., Гамалей Ю.В., Слезнев Н.Н. Направления эволюции генома покрытосеменных // Цитология. 2011. Т. 53. С. 295-311.
8. Шнеер В.С. ДНК – штрихкодирование видов животных и растений – способ их молекулярной идентификации и изучения биоразнообразия // Журнал Общей Биологии. 2009. Т. 70. № 4. С. 296-315.
9. 1000 Genomes Project Consortium, Abecasis G.R., Auton A., Brooks L.D., DePristo M.A., Durbin R.M., Handsaker R.E., Kang H.M., Marth G.T., McVean G.A. An integrated map of genetic variation from 1,092 human genomes // Nature. 2012. V. 491. P. 56-65.
10. Achenbach-Richter L., Gupta R., Stetter K.O., Woese C.R. Were the original eubacteria thermophiles? // Syst. Appl. Microbiol. 1987. V. 9. P. 34-39.
11. Adamski J. Genome-wide association studies with metabolomics // Genome Med. 2012. V. 4. P. 34.
12. Adamski J., Suhre K. Metabolomics platforms for genome wide association studies--linking the genome to the metabolome // Curr Opin Biotechnol. 2013. V. 24. P. 39-47.
13. Aik W., Demetriades M., Hamdan M.K., Bagg E.A., Yeoh K.K., Lejeune C., Zhang Z., McDonough M.A., Schofield C.J. Structural basis for inhibition of the fat mass and obesity associated protein (FTO) // J. Med. Chem. 2013. V. 56. P. 3680-3688.

14. Almal S.H., Padh H. Implications of gene copy-number variation in health and diseases // J. Hum. Genet. 2012. V. 57. P. 6-13.
15. Alvarez C.E., Akey J.M. Copy number variation in the domestic dog // Mamm. Genome. 2012. V.23. P.144-163.
16. Arahal D.R., Vreeland R.H., Litchfield C.D., Mormile M.R., Tindall B.J., Oren A, Bejar V., Quesada E., Ventosa A. Recommended minimal standards for describing new taxa of the family Halomonadaceae // Int J Syst Evol Microbiol. 2007. V. 57. P. 2436-2446.
17. Aten E., White S.J., Kalf M.E. et al. Methods to detect CNVs in the human genome // Cytogenet. Genome Res. 2008. V. 123. P. 313-321.
18. Baaske P., Weinert F.M., Duhr S., Lemke K.H., Russell M.J., Braun D. Extreme accumulation of nucleotides in simulated hydrothermal pore systems // Proc Natl Acad Sci U S A. 2007. V. 104. P. 9346-9351.
19. Bada J.L., Bigham C., Miller S.L. Impact melting of frozen oceans on the early Earth: implications for the origin of life // Proc. Natl. Acad. Sci. USA. 1994. V. 91. P. 1248-1250.
20. Baranzini S.E., Mudge J., van Velkinburgh J.C., Khankhanian P., Khrebtukova I., Miller N.A., Zhang L., Farmer A.D., Bell C.J., Kim R.W., May G.D., Woodward J.E., Caillier S.J., McElroy J.P., Gomez R., Pando M.J., Clendenen L.E., Ganusova E.E., Schilkey F.D., Ramaraj T., Khan O.A., Huntley J.J., Luo S., Kwok P.Y., Wu T.D., Schroth G.P., Oksenberg J.R., Hauser S.L., Ben-Avraham D., Muzumdar R.H., Atzmon G. Epigenetic genome-wide association methylation in aging and longevity // Epigenomics. 2012. V. 4. P. 503-509.
21. Baymiev Al.K., Chemeris A.V., Chemeris D.A.; Korpela T.K., Usanov N.G., Vakhitov V.A. Digital identification of genetic materials and methods for acquiring data for it // US Patent Application Publication 2007/0092873 A1. April 26, 2007.
22. Bell D.C., Thomas W.K., Murtagh K.M., Dionne C.A., Graham A.C., Anderson J.E., Glover W.R. DNA base identification by electron microscopy // Microsc. Microanal. 2012. V.18. P.1049-1053.
23. Bennett M.D., Leitch I.J. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow // Ann Bot. 2011. 107. P. 467-590.
24. Bennett M.D., Leitch I.J., Price H.J., Johnston J.S. Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb // Ann. Bot. 2003. V.91. P.547-557.
25. Berulava T., Ziehe M., Klein-Hitpass L., Mladenov E., Thomale J., Rüther U., Horsthemke B. FTO levels affect RNA modification and the transcriptome // Eur. J. Hum. Genet. 2013. 21. P. 317-323.
26. Biba E. Genome at home: biohackers build their own labs // Wired. 2011. 19.09.
27. Bickhart D.M., Hou Y., Schroeder S.G., Alkan C., Cardone M.F., Matukumalli L.K., Song J., Schnabel R.D., Ventura M., Taylor J.F., Garcia J.F., Van Tassell C.P., Sonstegard T.S., Eichler E.E., Liu G. E. Copy number variation of individual cattle genomes using next-generation sequencing // Genome Res. 2012. V. 22. P. 778-790.
28. Biderre C., Pages M., Méténier G., Canning E.U., Vivarès C.P. Evidence for the smallest nuclear genome (2.9 Mb) in the microsporidium Encephalitozoon cuniculi // Mol. Biochem. Parasitol. 1995. V. 74. P. 229-231.
29. Botstein D., White R.L., Skolnick M., Davis R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms // Am. J. Hum. Genet. 1980. V.32. P.14-31.
30. Brock T.D., Freeze H. *Thermus aquaticus* gen. n. and sp. n., a nonsporulating extreme thermophile // J Bacteriol. 1969. V. 98. P. 289-297.
31. Buffart T.E., Israeli D., Tijssen M., Vosse S.J., Mrsić A., Meijer G.A., Ylstra B. Across array comparative genomic hybridization: a strategy to reduce reference channel hybridizations // Genes Chromosomes Cancer. 2008. V.47. P.994-1004.
32. Caporaso N., Gu F., Chatterjee N., Sheng-Chih J., Yu K., Yeager M., Chen C., Jacobs K., Wheeler W., Landi M.T., Ziegler R.G., Hunter D.J., Chanock S., Hankinson S., Kraft P., Bergen A.W. Genome-wide and candidate gene association study of cigarette smoking behaviors // PLoS One. 2009. V. 4. e4653.
33. Carter N.P. Methods and strategies for analyzing copy number variation using DNA microarrays // Nat. Genet. 2007. V. 39. P. 16-21.
34. Cavalier-Smith T. Economy, speed and size matter: evolutionary forces driving nuclear

genome miniaturization and expansion // Ann. Bot. 2005. V. 95. P. 147-175.
35. Ceulemans S., van der Ven K., Del-Favero J. Targeted screening and validation of copy number variations // Methods Mol. Biol. 2012. V. 838. P. 311-328.
36. China Plant BOL Group, Li D.Z., Gao L.M., Li H.T., Wang H., Ge X.J., Liu J.Q., Chen Z.D., Zhou S.L., Chen S.L., Yang J.B., Fu C.X., Zeng C.X., Yan H.F., Zhu Y.J., Sun Y.S., Chen S.Y., Zhao L., Wang K., Yang T., Duan G.W. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants //Proc Natl Acad Sci U S A. 2011. V. 108. P. 19641-19646.
37. Choudhry Z., Sengupta S.M., Grizenko N., Thakur G.A., Fortier M.E., Schmitz N., Joober R. Association between obesity-related gene FTO and ADHD // Obesity (Silver Spring). 2013 Mar 20. doi: 10.1002/oby.20444. [Epub ahead of print]
38. Clop A., Vidal O., Amills M. Copy number variation in the genomes of domestic animals // Anim Genet. 2012. V. 43. P. 503-517.
39. Cokus S.J., Feng S., Zhang X., Chen Z., Merriman B., Haudenschild C.D., Pradhan S., Nelson S.F., Pellegrini M., Jacobsen S.E. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning // Nature. 2008. V. 452. P. 215-219.
40. Cooper G.M., Zerr T., Kidd J.M., Eichler E.E., Nickerson D.A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping // Nat. Genet. 2008. V. 40. P. 1199-1203.
41. Corradi N., Pombert J.F., Farinelli L., Didier E.S., Keeling P.J. The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis* // Nat. Commun. 2010. V.1. Article number: 77.
42. Cowan R.S., Fay M.F. Challenges in the DNA barcoding of plant material // Methods Mol. Biol. 2012. V. 862. P. 23-33.
43. Cox C.J., Foster P.G., Hirt R.P., Harris S.R., Embley T.M. The archaebacterial origin of eukaryotes // Proc. Natl. Acad. Sci. USA.2008. V. 105. P. 20356-20361.
44. Cyranoski D. Chinese bioscience: The sequence factory // Nature. 2010. V. 464. P. 22-24.
45. David S.P., Hamidovic A., Chen G.K., Bergen A.W., Wessel J., Kasberger J.L., Brown W.M., Petruzella S., Thacker E.L., Kim Y., Nalls M.A., Tranah G.J., Sung Y.J., Ambrosone C.B., Arnett D., Bandera E.V., Becker D.M., Becker L, Berndt S.I., Bernstein L., Blot W.J., Broeckel U., Buxbaum S.G., Caporaso N., Casey G., Chanock S.J., Deming S.L., Diver W.R., Eaton C.B., Evans D.S., Evans M.K., Fornage M., Franceschini N., Harris T.B., Henderson B.E., Hernandez D.G., Hitsman B., Hu J.J., Hunt S.C., Ingles S.A., John E.M., Kittles R., Kolb S., Kolonel L.N., Le Marchand L., Liu Y., Lohman K.K., McKnight B., Millikan R.C., Murphy A., Neslund-Dudas C., Nyante S., Press M., Psaty B.M., Rao D.C., Redline S., Rodriguez-Gil J.L., Rybicki B.A., Signorello L.B., Singleton A.B., Smoller J., Snively B., Spring B., Stanford J.L., Strom S.S., Swan G.E., Taylor K.D., Thun M.J., Wilson A.F., Witte J.S., Yamamura Y., Yanek L.R., Yu K., Zheng W., Ziegler R.G., Zonderman A.B., Jorgenson E., Haiman C.A., Furberg H. Genome-wide meta-analyses of smoking behaviors in African Americans // Transl Psychiatry. 2012. V. 2. e119.
46. Dellinger A.E., Saw S.M., Goh L.K., Seielstad M., Young T.L., Li Y.J. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays // Nucleic Acids Res. 2010. V. 38. e105.
47. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, Basford MA, Carrell DS, Peissig PL, Kho AN, Pacheco JA, Rasmussen LV, Crosslin DR, Crane PK, Pathak J, Bielinski SJ, Pendergrass SA, Xu H, Hindorff LA, Li R, Manolio TA, Chute CG, Chisholm RL, Larson EB, Jarvik GP, Brilliant MH, McCarty CA, Kullo IJ, Haines JL, Crawford DC, Masys DR, Roden DM. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data // Nat. Biotechnol. 2013. V.31. P.1102-1110.
48. De Souza Y.G., Greenspan J.S. Biobanking past, present and future: responsibilities and benefits // AIDS. 2013. 27. P. 303-312.
49. Dhawan D., Padh H. Pharmacogenetics: technologies to detect copy number variations // Curr. Opin. Mol. Ther. 2009. V. 11. P. 670-680.
50. Dib C., Fauré S., Fizames C., Samson D., Drouot N., Vignal A., Millasseau P., Marc S., Hazan J., Seboun E., Lathrop M., Gyapay G., Morissette J., Weissenbach J. A comprehensive genetic map of the human genome based on

5,264 microsatellites // Nature. 1996. V. 380. 152-154.
51. Doan R., Cohen N., Harrington J., Veazy K., Juras R., Cothran G., McCue M.E., Skow L., Dindot S. V. Identification of copy number variants in horses // Genome Res. 2012. V. 22. P. 899-907.
52. Donis-Keller H., Green P., Helms C., Cartinhour S., Weiffenbach B., Stephens K., Keith T.P., Bowden D.W., Smith D.R., Lander E.S., et al. A genetic linkage map of the human genome // Cell. 1987. V. 51. P. 319-337.
53. Drgon T., Montoya I., Johnson C., Liu Q.R., Walther D., Hamer D., Uhl G.R. Genome-wide association for nicotine dependence and smoking cessation success in NIH research volunteers // Mol Med. 2009. V. 15. P. 21-27.
54. Drgon T., Johnson C., Walther D., Albino A.P., Rose J.E., Uhl G.R. Genome-wide association for smoking cessation success: participants in a trial with adjunctive denicotinized cigarettes // Mol. Med. 2009a. V.15. P.268-274.
55. Du Y., Xie J., Chang W., Han Y., Cao G. Genome-wide association studies: inherent limitations and future challenges // Front Med. 2012. V. 6. P. 444-450.
56. Dubé J.B., Hegele R.A. Genetics 100 for cardiologists: basics of genome-wide association studies // Can. J. Cardiol. 2013. V. 29. P. 10-17.
57. Dube S., Qin J., Ramakrishnan R. Mathematical analysis of copy number variation in a DNA sample using digital PCR on a nanofluidic device // PLoS One. 2008. V. 3. e2876.
58. Duck W.M., Steward C.D., Banerjee S.N., McGowan J.E. Jr., Tenover F.C. Optimization of computer software settings improves accuracy of pulsed-field gel electrophoresis macrorestriction fragment pattern analysis // J. Clin. Microbiol. 2003. V.41. P.3035-3042.
59. Emelyanov V.V. Common evolutionary origin of mitochondrial and rickettsial respiratory chains // Arch. Biochem. Biophys. 2003. V.420. P.130-141.
60. Emelyanov V.V. Mitochondrial connection to the origin of the eukaryotic cell // Eur. J. Biochem. 2003a. V. 270. P. 1599-1618.
61. Emelyanov V.V. Rickettsiaceae, rickettsia-like endosymbionts, and the origin of mitochondria // Biosci Rep. 2001. V. 21. P. 1-17.
62. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ, Deikus G, Clark TA, Luong K, Murray IA, Davis BM, Keren-Paz A, Chess A, Roberts RJ, Korlach J, Turner SW, Kumar V, Waldor MK, Schadt EE. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing // Nat. Biotechnol. 2012. V.30. P.1232-1239. Erratum in Nat. Biotechnol. 2013. V.31. P.566.
63. Feulgen R, Rossenbeck H. Mikroskopisch-chemischer Nachweis einer Nucleinsäure vom Typus der Thymonukleinsäure und die darauf beruhende elektive Färbung von Zellkernen in mikroskopischen Präparaten // Hoppe-Seyler's Zeitschr. Physiologis. Chem. 1924. V.135. P.203–248.
64. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW. Direct detection of DNA methylation during single-molecule, real-time sequencing // Nat. Methods. 2010. V.7. P.461-465.
65. Forner K., Lamarine M., Guedj M., Dauvillier J., Wojcik J. Universal false discovery rate estimation methodology for genome-wide association studies // Hum Hered. 2008. V. 65. P. 183-194.
66. Forterre P. A new fusion hypothesis for the origin of Eukarya: better than previous ones, but probably also wrong // Res. Microbiol. 2011. V. 162. P. 77-91.
67. Forterre P. Why are there so many diverse replication machineries? // J. Mol. Biol. 2013 . V. 425. P. 4714-4726.
68. Frayling T.M., Ong K. Piecing together the FTO jigsaw // Genome Biol. 2011. V. 12. P. 104.
69. Frayling T.M., Timpson N.J., Weedon M.N., Zeggini E., Freathy R.M., Lindgren C.M., Perry J.R., Elliott K.S., Lango H., Rayner N.W., Shields B, Harries L.W., Barrett J.C., Ellard S, Groves C.J., Knight B., Patch A.M., Ness A.R., Ebrahim S., Lawlor D.A., Ring S.M., Ben-Shlomo Y., Jarvelin M.R., Sovio U., Bennett A.J., Melzer D., Ferrucci L., Loos R.J., Barroso I., Wareham N.J., Karpe F., Owen K.R., Cardon L.R., Walker M., Hitman G.A., Palmer C.N., Doney A.S., Morris A.D., Smith G.D., Hattersley A.T., McCarthy M.I. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity // Science. 2007. V. 316. P. 889-894.
70. Freeman J.L., Perry G.H., Feuk L. et al. Copy number variation: New insights in genome diversity // Genome Res. 2006. V. 16. P. 949-961.

71. Friz C.T. The biochemical composition of the free-living amoebae Chaos chaos, Amoeba dubia and Amoeba proteus // Comp. Biochem. Physiol. 1968. V. 26. P. 81-90.
72. Furberg H., Ostroff J., Lerman C., Sullivan P.F. The public health utility of genome-wide association study results for smoking behavior // Genome Med. 2010. V. 2. P. 26.
73. Furukawa H., Oka S., Matsui T., Hashimoto A., Arinuma Y., Komiya A., Fukui N., Tsuchiya N., Tohma S. Genome, epigenome and transcriptome analyses of a pair of monozygotic twins discordant for systemic lupus erythematosus // Hum Immunol. 2013. V. 74. P. 170-175.
74. Gamazon E.R., Huang R.S., Dolan M.E., Cox N.J. Copy number polymorphisms and anticancer pharmacogenomics // Genome Biol. 2011. 12. R46.
75. Gerken T., Girard C.A., Tung Y.C., Webby C.J., Saudek V., Hewitson K.S., Yeo G.S., McDonough M.A., Cunliffe S., McNeill L.A., Galvanovskis J., Rorsman P., Robins P., Prieur X., Coll A.P., Ma M., Jovanovic Z., Farooqi I.S., Sedgwick B., Barroso I., Lindahl T., Ponting C.P., Ashcroft F.M., O'Rahilly S., Schofield C.J. The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase // Science. 2007. V. 318. P. 1469-1472.
76. Gershon E.S., Alliey-Rodriguez N., Liu C. After GWAS: searching for genetic risk for schizophrenia and bipolar disorder // Am J Psychiatry. 2011. V. 168. P. 253-256.
77. Gibson G. Rare and common variants: twenty arguments // Nat. Rev. Genet. 2012. V. 13. P. 135-145.
78. Gill B.S., Appels R., Botha-Oberholster A.M., Buell C.R., Bennetzen J.L., Chalhoub B., Chumley F., Dvorák J., Iwanaga M., Keller B., Li W., McCombie W.R., Ogihara Y., Quetier F., Sasaki T. A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium // Genetics. 2004. V.168. P.1087-`1096.
79. Goering R.V. Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease // Infect. Genet. Evol. 2010. V.10. P.866-875.
80. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldman, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philipsen, P., Tettelin, H., Oliver, S.G. Life with 6000 genes // Science. 1996. V.274. P.546-567.
81. Goris J., Konstantinidis K.T., Klappenbach J.A., Coenye T., Vandamme P., Tiedje J.M. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities // Int. J. Syst. Evol. Microbiol. 2007. V. 57. P. 81-91.
82. Goyal A., Bandopadhyay R., Sourdille P., Endo T.R., Balyan H.S., Gupta P.K. Physical molecular maps of wheat chromosomes // Funct. Integr. Genomics. 2005. V. 5. P. 260-263.
83. Gray I.C., Campbell D.A., Spurr N.K. Single nucleotide polymorphisms as tools in human genetics // Hum. Mol. Genet. 2000. V.9. P.2403-2408.
84. Gregory T.R. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma // Biol. Rev. Camb. Philos. Soc. 2001. V. 76. P. 65-101.
85. Gregory T.R. Synergy between sequence and size in large-scale genomics // Nat. Rev. Genet. 2005. V. 6. P. 699-708.
86. Gregory T.R.The C-value enigma in plants and animals: a review of parallels and an appeal for partnership // Ann. Bot. 2005. V. 95. P. 133-146.
87. Greilhuber J. Cytochemistry and C-values: the less-well-known world of nuclear DNA amounts // Ann. Bot. 2008. V. 101. P. 791-804.
88. Greilhuber J., Borsch T., Müller K., Worberg A., Porembski S., Barthlott W. Smallest angiosperm genomes found in lentibulariaceae, with chromosomes of bacterial size // Plant Biol. (Stuttg). 2006. V. 8. P. 770-777.
89. Greilhuber J., Dolezel J. 2C or not 2C: a closer look at cell nuclei and their DNA content // Chromosoma. 2009. V. 118. P. 391-400.
90. Greilhuber J., Dolezel J., Lysák M.A., Bennett M.D. The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents // Ann. Bot. 2005. V. 95. 255-260.
91. Gupta P.K., Mir R.R., Mohan A., Kumar J. Wheat genomics: present status and future prospects // Int. J. Plant Genomics. 2008. Article ID 896451. 36 pp.
92. Haga H., Yamada R., Ohnishi Y., Nakamura Y., Tanaka T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism // J. Hum. Genet. 2002. V. 47. 605-610.

93. Hakonarson H., Grant S.F. Planning a genome-wide association study: points to consider // Ann Med. 2011. V. 43. P. 451-460
94. Han Z., Niu T., Chang J., Lei X., Zhao M., Wang Q., Cheng W., Wang J., Feng Y., Chai J. Crystal structure of the FTO protein reveals basis for its substrate specificity // Nature. 2010. V. 464. P. 1205-1209.
95. Handsaker R.E., Korn J.M., Nemesh J., McCarroll S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale // Nat. Genet. 2011. V. 43. P. 269-276.
96. Harris J.R., Burton P., Knoppers B.M., Lindpaintner K., Bledsoe M., Brookes A.J., Budin-Ljøsne I., Chisholm R., Cox D., Deschênes M., Fortier I., Hainaut P., Hewitt R., Kaye J., Litton J.E., Metspalu A., Ollier B., Palmer L.J., Palotie A., Pasterk M., Perola M., Riegman P.H., van Ommen G.J., Yuille M., Zatloukal K. Toward a roadmap in global biobanking for health // Eur J Hum Genet. 2012. V. 20. P. 1105-1111.
97. He Y., Hoskins J.M., McLeod H.L. Copy number variants in pharmacogenetic genes // Trends Mol Med. 2011. V. 17. P. 244-251.
98. Hebert P.D., Cywinska A., Ball S.L., deWaard J.R. Biological identifications through DNA barcodes // Proc Biol Sci. 2003. V. 270. P. 313-321.
99. Heller F.O. DNS-Bestimmung an Keimwurzeln von *Vicia faba* L. mit Hilfe der Impulscytophotometrie // Ber. Deutsch. Bot. Ges. (now – Plant Biology) 1973. V. 86. P.437–441.
100. Henrichsen C.N., Chaignat E., Reymond A. Copy number variants, diseases and gene expression // Hum. Mol. Genet. 2009. V. 18. R1-8.
101. Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S., Manolio T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits // Proc Natl Acad Sci U S A. 2009. V. 106. P. 9362-9367.
102. Hollingsworth P.M. Refining the DNA barcode for land plants // Proc Natl Acad Sci USA. 2011. V. 108. P. 19451-19452.
103. Hollox E.J., Huffmeier U., Zeeuwen P.L. et al. Psoriasis is associated with increased beta-defensin genomic copy number // Nat. Genet. 2008. V. 40. P. 23-25.
104. Homuth G., Teumer A., Völker U., Nauck M. A description of large-scale metabolomics studies: increasing value by combining metabolomics with genome-wide SNP genotyping and transcriptional profiling // J Endocrinol. 2012. V. 215. P. 17-28.
105. Hongoh Y., Toyoda A. Whole-genome sequencing of unculturable bacterium using whole-genome amplification // Methods Mol. Biol. 2011. V.733. P.25-33.
106. Hood L., Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory // N. Biotechnol. 2012. V. 29. P. 613-624.
107. Huang Z., Wang J., Wu C.C., Houlston R.S., Bondy M.L., Shete S. False-Negative-Rate Based Approach for Selecting Top Single-Nucleotide Polymorphisms in the First Stage of a Two-Stage Genome-Wide Association Study // Stat. Interface. 2011. V. 4. P. 359-371.
108. Hubacek JA, Dlouha D, Lanska V, Adamkova V. Lack of an association between three tagging SNPs within the FTO gene and smoking behavior // Nicotine Tob. Res. 2012. V.14. P.998-1002.
109. Hubacek J.A, Piper B.J., Pikhart H., Peasey A., Kubinova R., Bobak M. Lack of an association between left-handedness and APOE polymorphism in a large sample of adults: Results of the Czech HAPIEE study // Laterality. 2013. V.18. P.513-519.
110. Hubacek J.A., Stanek V., Gebauerová M., Pilipcincová A., Dlouhá D., Poledne R, Aschermann M., Skalická H, Matousková J., Kruger A., Penicka M., Hrabáková H., Veselka J., Hájek P., Lánská V., Adámková V., Pitha J. A FTO variant and risk of acute coronary syndrome // Clin Chim Acta. 2010. V. 411. P. 1069-1072.
111. Hüebner C, Petermann I, Browning BL, Shelling AN, Ferguson LR. Triallelic single nucleotide polymorphisms and genotyping error in genetic epidemiology studies: MDR1 (ABCB1) G2677/T/A as an example // Cancer Epidemiol. Biomarkers Prev. 2007. V.16. P.1185-1192.
112. Huynh J.L., Casaccia P. Epigenetic mechanisms in multiple sclerosis: implications for pathogenesis and treatment // Lancet Neurol. 2013. V. 12. P. 195-206.
113. Iafrate A.J., Feuk L., Rivera M.N., Listewnik M.L., Donahoe P.K., Qi Y., Scherer S.W., Lee C. Detection of large-scale variation in the human genome // Nat Genet. 2004. V. 36. P. 949-951.
114. Ibba A., Pilia S., Zavattari P., Loche A., Guzzetti C., Casini M.R., Minerba L., Loche S.

The role of FTO genotype on eating behavior in obese Sardinian children and adolescents // J. Pediatr Endocrinol. Metab. 2013 V. 26. P. 539-544.

115. Iles M.M., Law M.H., Stacey S.N., Han J., Fang S., Pfeiffer R., Harland M., Macgregor S., Taylor J.C., Aben K.K., Akslen L.A., Avril M.F., Azizi E., Bakker B., Benediktsdottir K.R., Bergman W, Scarrà G.B., Brown K.M., Calista D., Chaudru V., Fargnoli M.C., Cust A.E., Demenais F., de Waal A.C., Dębniak T., Elder D.E., Friedman E., Galan P., Ghiorzo P., Gillanders E.M., Goldstein A.M., Gruis N.A., Hansson J., Helsing P., Hočevar M., Höiom V., Hopper J.L., Ingvar C., Janssen M., Jenkins M.A., Kanetsky P.A., Kiemeney L.A., Lang J., Lathrop G.M., Leachman S., Lee J.E., Lubiński J., Mackie R.M., Mann G.J., Martin N.G., Mayordomo J.I., Molven A., Mulder S., Nagore E., Novaković S., Okamoto I., Olafsson J.H., Olsson H., Pehamberger H., Peris K., Grasa M.P., Planelles D., Puig S., Puig-Butille J.A., Randerson-Moor J., Requena C., Rivoltini L., Rodolfo M., Santinami M., Sigurgeirsson B., Snowden H., Song F., Sulem P., Thorisdottir K., Tuominen R., Van Belle P., van der Stoep N., van Rossum M.M., Wei Q., Wendt J., Zelenika D., Zhang M., Landi M.T., Thorleifsson G., Bishop D.T., Amos C.I., Hayward N.K., Stefansson K., Bishop J.A., Barrett J.H; GenoMEL Consortium; Q-MEGA and AMFS Investigators. A variant in FTO shows association with melanoma risk not due to BMI // Nat. Genet. 2013. V. 45. P. 428-432, 432e1.

116. International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome // Nature. 2001. V.409. P.860-921.

117. Ioannidis J.P. Why most discovered true associations are inflated // Epidemiology. 2008. V. 19. P. 640-648.

118. Ionita-Laza I., Rogers A.J., Lange C., Raby B.A., Lee C. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis // Genomics. 2009. V. 93. P. 22-26.

119. Jiang C., Wright R.J., El-Zik K.M., Paterson A.H. Polyploid formation created unique avenues for response to selection in Gossypium (cotton) // Proc. Nat. Acad. Sci. USA. 1998. V. 95. P. 4419-4424.

120. Juran B.D., Lazaridis K.N. Genomics in the post-GWAS era // Semin. Liver Dis. 2011. V. 31. P. 215-222.

121. Kane N.C., Cronk Q. Botany without borders: barcoding in focus // Mol. Ecol. 2008. V. 17. P. 5175-5176.

122. Kaneko T., Nakamura Y., Sato S., Asamizu E., Kato T., Sasamoto S., Watanabe A., Idesawa K., Ishikawa A., Kawashima K., Kimura T., Kishida Y., Kiyokawa C., Kohara M., Matsumoto M., Matsuno A., Mochizuki Y., Nakayama S., Nakazaki N., Shimpo S., Sugimoto M., Takeuchi C., Yamada M., Tabata S. Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti // DNA Res. 2000. V.7. P.331-338.

123. Konstantinidis K.T., Tiedje J.M. Genomic insights that advance the species definition for prokaryotes // Proc Natl Acad Sci U S A. 2005. V. 102. P. 2567-2572.

124. Korol A., Frenkel Z., Orion O., Ronin Y. Some ways to improve QTL mapping accuracy // Anim Genet. 2012. V. 43. P. 36-44.

125. Korte A., Farlow A. The advantages and limitations of trait analysis with GWAS: a review // Plant Methods. 2013. 9:29.

126. Kraaijeveld K. Genome Size and Species Diversification // Evol. Biol. 2010. V. 37. P. 227-233.

127. Kreck B., Marnellos G., Richter J., Krueger F., Siebert R., Franke A. B-SOLANA: an approach for the analysis of two-base encoding bisulfite sequencing data // Bioinformatics. 2012. V. 28. P. 428-429.

128. Ku C.S., Loy E.Y., Salim A., Pawitan Y., Chia K.S. The discovery of human genetic variations and their use as disease markers: past, present and future // J. Hum. Genet. 2010. V. 55. P. 403-415.

129. Kullo I.J., Cooper LT. Early identification of cardiovascular risk using genomics and proteomics // Nat. Rev .Cardiol. 2010. V. 7. P. 309-317.

130. Lake J.A., Skophammer R.G., Herbold C.W., Servin J.A. Genome beginnings: rooting the tree of life // Philos. Trans. R. Soc. Lond. B. Biol. Sci. 2009. V. 364. P. 2177-2185.

131. Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C. et al. International Human Genome Sequencing Consortium Initial sequencing and analysis of the human genome // Nature. 2001. V. 409. P. 860-921.

132. Lasken R.S. Genomic sequencing of uncultured microorganisms from single cells // Nat Rev Microbiol. 2012. V. 10. P. 631-640.

133. Ledford H. Garage biotech: Life hackers // Nature. 2010. V. 467. P. 650-652.

134. Lee C., Hyland C., Lee A.S., Hislop S., Ihm C. Copy number variation and human health / Essentials of Genomic and Personalized Medicine. Ginsburg G.S., Willard H.F. (Eds). 2010. P.46-59.
135. Leroy S., Bouamer S., Morand S., Fargette M. Genome size of plant-parasitic nematodes // Nematology. 2007. V.9. P.449-450.
136. Li T., Wu K., You L., Xing X., Wang P., Cui L., Liu H., Cui Y., Bian Y., Ning Y., Zhao H., Tang R., Chen Z.J. Common Variant rs9939609 in Gene FTO Confers Risk to Polycystic Ovary Syndrome // PLoS One. 2013. V. 8. e66250.
137. Li W., Olivier M. Current analysis platforms and methods for detecting copy number variation // Physiol Genomics. 2013. V. 45. P. 1-16.
138. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. Human DNA methylomes at base resolution show widespread epigenomic differences // Nature. 2009. V.462. P.315-322.
139. Litt M., Luty J.A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene // Am. J. Hum. Genet. 1989. V. 44. P. 397-401.
140. Liu G.E., Bickhart D.M. Copy number variation in the cattle genome // Funct. Integr. Genomics. 2012. V.12. P609-624.
141. Liu Y.J., Papasian C.J., Liu J.F., Hamilton J., Deng H.W. Is replication the gold standard for validating genome-wide association findings? // PLoS One. 2008. V. 3. e4037.
142. Logan N.A., Berge O., Bishop A.H., Busse H.J., De Vos P., Fritze D., Heyndrickx M., Kämpfer P., Rabinovitch L., Salkinoja-Salonen M.S., Seldin L., Ventosa A. Proposed minimal standards for describing new taxa of aerobic, endospore-forming bacteria // Int. J. Syst. Evol. Microbiol. 2009. V. 59. P. 2114-2121.
143. Loukola A., Wedenoja J., Keskitalo-Vuokko K., Broms U., Korhonen T., Ripatti S., Sarin A.P., Pitkäniemi J., He L., Häppölä A., Heikkilä K., Chou Y.L., Pergadia M.L., Heath A.C., Montgomery G.W., Martin N.G., Madden P.A., Kaprio J. Genome-wide association study on detailed profiles of smoking behavior and nicotine dependence in a twin sample // Mol Psychiatry. 2013. Jun 11. doi: 10.1038/mp.2013.72. [Epub ahead of print]
144. Manolio T.A. Bringing genome-wide association findings into clinical use // Nat. Rev. Genet. 2013. V. 14. P. 549-558.
145. Manolio T.A. Genomewide association studies and assessment of the risk of disease // N. Engl. J. Med. 2010. V. 363. P. 166-176.
146. Maouche S., Schunkert H. Strategies beyond genome-wide association studies for atherosclerosis // Arterioscler Thromb. Vasc. Biol. 2012. V. 32. P. 170-181.
147. Marshall I.P., Blainey P.C., Spormann A.M., Quake S.R. A Single-cell genome for Thiovulum sp. // Appl Environ Microbiol. 2012. V. 78. P. 8555-8563.
148. Martin W., Müller M. The hydrogen hypothesis for the first eukaryote // Nature. 1998. V. 392. P. 37-41.
149. May RM. How many species are there on Earth? // Science. 1988. V. 24. P. 1441-1449.
150. May R.M. How many species? // Phil. Trans. R. Soc. Lond. 1990. V. 330. P. 293-304.
151. McCarroll S.A., Kuruvilla F.G., Korn J.M. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation // Nat. Genet. 2008. V. 40. P. 1166-1174.
152. McDonnell S.K., Riska S.M., Klee E.W. et al. Experimental designs for array comparative genomic hybridization technology // Cytogenet. Genome Res. 2013. V. 139. P. 250-257.
153. Melén E., Granell R., Kogevinas M., Strachan D., Gonzalez J.R., Wjst M., Jarvis D., Ege M., Braun-Fahrländer C., Genuneit J., Horak E., Bouzigon E., Demenais F., Kauffmann F., Siroux V., Michel S., von Berg A., Heinzmann A., Kabesch M., Probst-Hensch N.M., Curjuric I., Imboden M., Rochat T., Henderson J., Sterne J.A., McArdle W.L., Hui J., James A.L., William Musk A., Palmer L.J., Becker A., Kozyrskyj A.L., Chan-Young M., Park J.E., Leung A., Daley D., Freidin M.B., Deev I.A., Ogorodova L.M., Puzyrev V.P., Celedón J.C., Brehm J.M., Cloutier M.M., Canino G., Acosta-Pérez E., Soto-Quiros M., Avila L., Bergström A., Magnusson J., Söderhäll C., Kull I., Scholtens S., Marike Boezen H., Koppelman G.H., Wijga A.H., Marenholz I., Esparza-Gordillo J., Lau S., Lee Y.A., Standl M., Tiesler C.M., Flexeder C., Heinrich J., Myers R.A., Ober C., Nicolae D.L., Farrall M., Kumar A., Moffatt M.F., Cookson W.O., Lasky-Su J. Genome-wide association study of body mass index in 23 000 individuals with

and without asthma // Clin. Exp. Allergy. 2013. V. 43. P. 463-474.
154. Metspalu M., Romero I.G., Yunusbayev B., Chaubey G., Mallick C.B., Hudjashov G., Nelis M., Mägi R., Metspalu E., Remm M., Pitchappan R., Singh L., Thangaraj K., Villems R., Kivisild T. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia // Am. J. Hum. Genet. 2011. V. 89. P. 731-744.
155. Meyer C.P., Paulay G. DNA barcoding: error rates based on comprehensive sampling PLoS Biol. 2005. V.3. e422.
156. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, Houseman EA, Izzi B, Kelsey KT, Meissner A, Milosavljevic A, Siegmund KD, Bock C, Irizarry RA. Recommendations for the design and analysis of epigenome-wide association studies // Nat. Methods. 2013. V.10. P.949-955.
157. Miller R.D., Kwok P.Y. The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine // Hum. Mol. Genet. 2001. V. 10. P. 2195-2198.
158. Miller R.D., Taillon-Miller P., Kwok P.Y. Regions of low single-nucleotide polymorphism incidence in human and orangutan xq: deserts and recent coalescences // Genomics. 2001. V. 71. P. 78-88.
159. Mills R.E., Walter K., Stewart C. et al. 1000 Genomes Project Mapping copy number variation by population-scale genome sequencing // Nature. 2011. V. 470. P. 59-65.
160. Mitra R., Bhatia C.R. Repeated and non-repeated nucleotide sequences in diploid and polyploid wheat species // Heredity. 1973. V. 31. P. 251-262.
161. Monteiro A.N., Freedman M.L. Lessons from postgenome-wide association studies: functional analysis of cancer predisposition loci // J. Intern. Med. 2013. V.274. P.414-424.
162. Montoliu I., Genick U., Ledda M., Collino S., Martin F.P., le Coutre J., Rezzi S. Current status on genome-metabolome-wide associations: an opportunity in nutrition research // Genes Nutr. 2013. V. 8. P. 19-27.
163. Mora C., Tittensor D.P., Adl S., Simpson A.G., Worm B. How many species are there on Earth and in the ocean? // PLoS Biol. 2011. V. 9. e1001127.
164. Morgan T.M., Krumholz H.M., Lifton R.P., Spertus J.A. Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study // JAMA. 2007. V.297. P.1551-61. Erratum in: JAMA. 2007. V.298. P.973.
165. Morita A., Nakayama T., Doba N., Hinohara Sh., Mizutani T., Soma M. Genotyping of triallelic SNPs using TaqMan PCR // Mol. Cell. Probes. 2007. V.21. P.171-176.
166. Müller T.D., Tschöp M.H., Hofmann S. Emerging function of fat mass and obesity-associated protein (fto) // PLoS Genet. 2013. V. 9. e1003223.
167. Nakamura Y. DNA variations in human and medical genetics: 25 years of my experience // J. Hum. Genet. 2009. V. 54. P. 1-8.
168. Nakamura Y., Leppert M., O'Connell P., Wolff R., Holm T., Culver M., Martin C., Fujimoto E., Hoff M, Kumlin E., et al. Variable number of tandem repeat (VNTR) markers for human gene mapping // Science. 1987. V. 235. P. 1616-1622.
169. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritime* // Nature. 1999. V.399. P.323-329.
170. Nygaard V, Holden M, Løland A, Langaas M, Myklebost O, Hovig E. Limitations of mRNA amplification from small-size cell samples // BMC Genomics. 2005. V.27. P.147.
171. Ohnishi Y., Tanaka T., Ozaki K., Yamada R., Suzuki H., Nakamura Y. A high-throughput SNP typing system for genome-wide association studies // J. Hum. Genet. 2001. V.46. P.471-477.
172. Ott J., Kamatani Y., Lathrop M. Family-based designs for genome-wide association studies // Nat. Rev. Genet. 2011. V. 12. P. 465-474.
173. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction // Nat. Genet. 2002. V.32. P.650-654. - Erratum in: Nat. Genet. 2003. V.33. P.107.
174. Pamjav H., Juhász Z., Zalán A., Németh E., Damdin B. A comparative phylogenetic study of genetics and folk music // Mol. Genet. Genomics. 2012. V. 287. P. 337-349.

175. Pankratz N., Dumitriu A., Hetrick K.N. et al. PSG-PROGENI and GenePD Investigators, Coordinators and Molecular Genetic Laboratories Copy number variation in familial Parkinson disease // PLoS One. 2011. V. 6. e20988.
176. Patrushev L.I., Minkevich I.G. The problem of the eukaryotic genome size // Biochemistry (Mosc). 2008. V.73. P.1519-1552.
177. Patterson M., Cardon L. Replication publication // PLoS Biol. 2005. V. 3. e327.
178. Pearson T.A., Manolio T.A. How to interpret a genom-wide association study // JAMA. 2008. V. 299. P. 1335-1344.
179. Pearson T.A., Manolio T.A. How to interpret a genome-wide association study // JAMA. 2008. V.299. P.1335-44. Erratum in: JAMA. 2008. V.299. P.2150.
180. Pedersen R.A. DNA content, ribosomal gene multiplicity, and cell size in fish // J. Experiment. Zool. 1971. V.177. P. 65-79.
181. Pellicer J., Garcia S., Canela M.A. et al. Genome size dynamics in Artemisia L. (Asteraceae): following the track of polyploidy // Plant Biol. (Stuttg). 2010. V. 12. P. 820-830.
182. Peters U., North K.E., Sethupathy P., Buyske S., Haessler J., Jiao S., Fesinmeyer M.D., Jackson R.D., Kuller L.H., Rajkovic A., Lim U., Cheng I., Schumacher F.,Wilkens L., Li R., Monda K., Ehret G., Nguyen K.D., Cooper R., Lewis C.E., Leppert M., Irvin M.R., Gu C.C., Houston D., Buzkova P., Ritchie M., Matise T.C., Le Marchand L., Hindorff L.A., Crawford D.C., Haiman C.A., Kooperberg C. A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: results from the Population Architecture using Genomics and Epidemiology (PAGE) study // PLoS Genet. 2013. V. 9. e1003171.
183. Philippe N., Legendre M., Doutre G. et al. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. // Science. 2013. V. 341. P. 281-286.
184. Pont C., Murat F., Guizard S. et al. Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes // Plant J. 2013. V. 76. P. 1030-1044.
185. Poole A.M., Penny D. Evaluating hypotheses for the origin of eukaryotes // Bioessays. 2007. V. 29. P. 74-84.
186. Price A.L., Zaitlen N.A., Reich D., Patterson N. New approaches to population stratification in genome-wide association studies // Nat Rev Genet. 2010. V. 11. P. 459-463.
187. Priebe L, Degenhardt F, Strohmaier J, Breuer R, Herms S, Witt SH, Hoffmann P, Kulbida R, Mattheisen M, Moebus S, Meyer-Lindenberg A, Walter H, Mössner R, Nenadic I, Sauer H, Rujescu D, Maier W, Rietschel M, Nöthen MM, Cichon S. Copy number variants in German patients with schizophrenia // PLoS One. 2013. V.8. e64035.
188. Queitsch C., Carlson K.D., Girirajan S. Lessons from model organisms: phenotypic robustness and missing heritability in complex disease // PLoS Genet. 2012. V. 8. e1003041.
189. Raghunathan A., Ferguson H.R.Jr., Bornarth C.J., Song W., Driscoll M., Lasken R.S. Genomic DNA amplification from a single bacterium // Appl Environ Microbiol. 2005. V. 71. P. 3342-3347.
190. Rakyan V.K., Down T.A., Balding D.J., Beck S. Epigenome-wide association studies for common human diseases // Nat Rev Genet. 2011. V.12. P. 529-541.
191. Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., Fiegler H., Shapero M.H., Carson A.R., Chen W., Cho E.K., Dallaire S., Freeman J.L., González J.R., Gratacòs M., Huang J., Kalaitzopoulos D., Komura D., MacDonald J.R., Marshall C.R., Mei R., Montgomery L., Nishimura K., Okamura K., Shen F., Somerville M.J., Tchinda J., Valsesia A., Woodwark C., Yang F., Zhang J., Zerjal T., Zhang J., Armengol L., Conrad D.F., Estivill X., Tyler-Smith C., Carter N.P., Aburatani H., Lee C., Jones K.W., Scherer S.W., Hurles M.E. Global variation in copy number in the human genome // Nature. 2006. V. 444. P. 444-454.
192. Reich D., Patterson N., Campbell D., Tandon A., Mazieres S., Ray N., Parra M.V., Rojas W. , Duque C., Mesa N., García L.F., Triana O., Blair S., Maestre A., Dib J.C.,Bravi C.M., Bailliet G., Corach D., Hünemeier T., Bortolini M.C., Salzano F.M., Petzl-Erler M.L., Acuña-Alonzo V., Aguilar-Salinas C., Canizales-Quinteros S., Tusié-Luna T., Riba L., Rodríguez-Cruz M., Lopez-Alarcón M., Coral-Vazquez R., Canto-Cetina T., Silva-Zolezzi I., Fernandez-Lopez J.C., Contreras A.V., Jimenez-Sanchez G., Gómez-Vázquez M.J., Molina J., Carracedo A., Salas A., Gallo C., Poletti G., Witonsky D.B., Alkorta-Aranburu G., Sukernik R.I., Osipova L., Fedorova S.A.,

Vasquez R., Villena M., Moreau C., Barrantes R., Pauls D., Excoffier L., Bedoya G., Rothhammer F., Dugoujon J.M., Larrouy G., Klitz W., Labuda D., Kidd J., Kidd K., Di Rienzo A., Freimer N.B., Price A.L., Ruiz-Linares A. Reconstructing Native American population history // Nature. 2012. V. 488. P. 370-374.

193. Riancho JA, Hernández JL. Pharmacogenomics of osteoporosis: a pathway approach // Pharmacogenomics. 2012. V.13. P.815-829.
194. Risch N., Merikangas K. The future of genetic studies of complex human diseases // Science. 1996. V.273. P.1516-1517. Comments - Genetic analysis of complex diseases // Science. 1997. V.275. P. 1327-1328. Scott W.K., Pericak-Vance M.A., Haines J.L. / Bell D.A., Taylor J.A. / Long A.D., Grote M.N., Langley C.H. author reply 1329-1330.
195. Robinette S.L., Holmes E., Nicholson J.K., Dumas M.E. Genetic determinants of metabolism in health and disease: from biochemical genetics to genome-wide associations // Genome Med. 2012. V. 4. P. 30.
196. Rose J.E., Behm F.M., Drgon T., Johnson C, Uhl G.R. Personalized smoking cessation: interactions between nicotine dose, dependence and quit-success genotype score // Mol Med. 2010. V. 16. P. 247-253.
197. Rose-Zerilli M.J., Barton S.J., Henderson A.J., Shaheen S.O., Holloway J.W. Copy-number variation genotyping of GSTT1 and GSTM1 gene deletions by real-time PCR // Clin. Chem. 2009. V. 55. P. 1680-1685.
198. Rowe S.J., Tenesa A. Human complex trait genetics: lifting the lid of the genomics toolbox - from pathways to prediction // Curr. Genomics. 2012. V. 13. P. 213-224.
199. Sadee W. The relevance of "missing heritability " in pharmacogenomics // Clin. Pharmacol. Ther. 2012. V.92. P.428-423.
200. Sándor J., Bárd P., Tamburrini C., Tännsjö T. The case of biobank with the law: between a legal and scientific fiction // J Med Ethics. 2012. V. 38. P. 347-350.
201. Sathirapongsasuti J.F., Lee H., Horst B.A. et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: Exome CNV // Bioinformatics. 2011. V. 27. P. 2648-2654.
202. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome // Science. 2004. V.305. P.525-528.
203. Shen Y., Wu B.L. Designing a simple multiplex ligation-dependent probe amplification (MLPA) assay for rapid detection of copy number variants in the genome // J. Genet. Genomics. 2009. V. 36. P. 257-265.
204. Shimada H., Obayashi T., Takahashi N., Matsui M., Sakamoto A. Normalization using ploidy and genomic DNA copy number allows absolute quantification of transcripts, proteins and metabolites in cells // Plant Methods. 2010. V. 6. P. 29.
205. Silander K., Saarela J. Whole genome amplification with Phi29 DNA polymerase to enable genetic or genomic analysis of samples of low DNA yield // Methods Mol. Biol. 2008. V.439. P.1-18.
206. Southern E.M. Detection of specific sequences among DNA fragments separated by gel electrophoresis // J. Mol. Biol. 1975. V.98. P.503-517.
207. Southern E.M. Blotting at 25 // Trends Biochem. Sci. 2000. V.25. P.585-588.
208. Spielmann M., Klopocki E. CNVs of noncoding cis-regulatory elements in human disease // Curr. Opin. Genet. Dev. 2013. V.23. P.249-256.
209. Song J., Shi L., Li D., Sun Y., Niu Y., Chen Z., Luo H., Pang X., Sun Z., Liu C., Lv A., Deng Y., Larson-Rabin Z., Wilkinson M., Chen S. Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA // PLoS One. 2012. V. 7. e43971.
210. Staats M., Erkens R.H., van de Vossenberg B., Wieringa J.J., Kraaijeveld K., Stielow B., Geml J., Richardson JE, Bakker F.T. Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens //PLoS One. 2013. V. 8. e69189.
211. Stankiewicz P., Lupski J.R. Structural variation in the human genome and its role in disease // Ann. Rev. Med. 2010. V. 61. P. 437-455.
212. Stevens H. Dr. Sanger, meet Mr. Moore: next-generation sequencing is driving new questions and new modes of research // Bioessays. 2012. V. 34. P. 103-105.
213. Stuppia L., Antonucci I., Palka G., Gatta V. Use of the MLPA Assay in the Molecular Diagnosis of Gene Copy Number Alterations in Human Genetic Diseases // Int. J. Mol. Sci. 2012. V. 13. P. 3245-3276.

214. Suhre K., Gieger C. Genetic variation in metabolic phenotypes: study designs and applications // Nat. Rev. Genet. 2012. V. 13. P. 759-769
215. Sun L., Craiu R.V., Paterson A.D., Bull S.B. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies // Genet Epidemiol. 2006. V. 30. P. 519-530.
216. Swift H. The constancy of desoxyribose nucleic acid in plant nuclei // Proc. Natl. Acad. Sci. USA. 1950. V. 36. P.643–654.
217. Tanaka H., Kawai T. Partial sequencing of a single DNA molecule with a scanning tunnelling microscope // Nature Nanotechnol. 2009. V.4. P.518-522.
218. Taylor H.R., Harris W.E. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding // Mol Ecol Resour. 2012. V. 12. P. 377-388.
219. Thomas C.A, Jr. The genetic organization of chromosomes // Annual Rev. Genet. 1971. V. 5. P. 237-256.
220. Thomas W.K, Glover W. Direct sequencing by TEM of Z-substituted DNA molecules / Next Generation Genome Sequencing: Towards Personalized Medicine. Ed.: M.Janitz. 2008. P.103-116.
221. Tobacco and Genetics Consortium.Collaborators (116) Genome-wide meta-analyses identify multiple loci associated with smoking behavior // Nat. Genet. 2010. V. 42. P. 441-447.
222. Todd JA. Statistical false positive or true disease pathway? // Nat Genet. 2006. V. 38. P. 731-733.
223. Tung Y.C., Yeo G.S. From GWAS to biology: lessons from FTO // Ann. N. Y. Acad Sci. 2011. V. 1220. P. 162-171.
224. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, Jarvik G, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto M, McCarty CA, McDavid AN, Mirel DB, Paschall JE, Pugh EW, Rasmussen LV, Wilke RA, Zuvich RL, Ritchie MD. Quality control procedures for genome-wide association studies // Curr. Protoc. Hum. Genet. 2011. Chapter 1. Unit1.19.
225. van Steenbergen T.J., Colloms S.D., Hermans P.W., de Graaff J., Plasterk R.H. Genomic DNA fingerprinting by restriction fragment end labeling // Proc. Natl. Acad. Sci. USA. 1995. V.92. P.5572-5576.
226. Vandeweyer G., Kooy R.F. Detection and interpretation of genomic structural variation in health and disease // Expert. Rev. Mol. Diagn. 2013. V. 13. P. 61-82.
227. Vaught J., Lockhart N.C. The evolution of biobanking best practices // Clin Chim Acta. 2012. V. 413. P. 1569-1575.
228. Vaught J.B., Henderson M.K., Compton C.C. Biospecimens and biorepositories: from afterthought to science // Cancer Epidemiol Biomarkers Prev. 2012. V. 21. P. 253-255.
229. Vesteg M., Krajčovič J. The falsifiability of the models for the origin of eukaryotes // Curr Genet. 2011. V. 57. P. 367-390.
230. Visscher P.M., Brown M.A., McCarthy M.I., Yang J. Five years of GWAS discovery // Am. J. Hum. Genet. 2012. V. 90. P. 7-24
231. Vrieze S.I., Iacono W.G., McGue M. Confluence of genes, environment, development, and behavior in a post Genome-Wide Association Study world // Dev. Psychopathol. 2012. 24. P. 1195-1214.
232. Wain L.V., Armour J.A., Tobin M.D. Genomic copy number variation, human health, and disease // Lancet. 2009. V. 374. P. 340-350.
233. Wang K., Li M., Hadley D. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data // Genome Res. 2007. V. 17. P. 1665-1674.
234. Wang X., Prins B.P., Sõber S., Laan M., Snieder H. Beyond genome-wide association studies: new strategies for identifying genetic determinants of hypertension // Curr Hypertens Rep. 2011. V.13. P. 442-451.
235. Weaver S., Dube S., Mir A. et al. Taking qPCR to a higher level: Analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution // Methods. 2010. V. 50. P. 271-276.
236. Weber J.L. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms // Genomics. 1990. V. 7. P. 524-530.
237. Weber J.L., May P.E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction // Am J Hum Genet. 1989. V. 44. P. 388-396.
238. Weissenbach J. Microsatellite polymorphisms and the genetic linkage map of the human genome // Curr Opin Genet Dev. 1993. V. 3. P. 414-417.
239. Weissenbach J., Gyapay G., Dib C., Vignal A., Morissette J., Millasseau P., Vaysseix G., Lathrop M. A second-generation linkage map

of the human genome // Nature. 1992. V. 359. P. 794-801.
240. Westen A.A., Matai A.S., Laros J.F., Meiland H.C., Jasper M., de Leeuw W.J., de Knijff P., Sijen T. Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples // Forensic Sci. Int. Genet. 2009. V. 3. P. 233-241.
241. Whale A.S., Huggett J.F., Cowen S., Speirs V., Shaw J., Ellison S., Foy C.A., Scott D.J. Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation // Nucleic Acids Res. 2012. V.40. e82.
242. Winchester L., Yau C., Ragoussis J. Comparing CNV detection methods for SNP arrays // Brief. Funct. Genomic. Proteomic. 2009. V. 8. P. 353-366.
243. Wjst M., Sargurupremraj M., Arnold M. Genome-wide association studies in asthma: what they really told us about pathogenesis // Curr Opin Allergy Clin Immunol. 2013. V. 13. P. 112-118.
244. Winkler H. Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche. 1920. Jena: Gustav Fischer Verlag.
245. Würschum T. Mapping QTL for agronomic traits in breeding populations // Theor Appl Genet. 2012. V. 125. P. 201-210.
246. Xu Y., Peng B., Fu Y., Amos C.I. Genome-wide algorithm for detecting CNV associations with diseases // BMC Bioinformatics. 2011. V.12. P. 331.
247. Yoon D., Kim Y.J., Park T. Phenotype prediction from genome-wide association studies: application to smoking behaviors // BMC Syst Biol. 2012. V. 6. S11.
248. Zeller T., Blankenberg S., Diemert P. Genomewide association studies in cardiovascular disease--an update 2011 // Clin Chem. 2012. V. 58. P. 92-103.
249. Zhang F., Gu W., Hurles M.E., Lupski J.R. Copy Number Variation in Human Health, Disease, and Evolution // Annual Review of Genomics and Human Genetics. 2009. V. 10. P.451-481.
250. Zhao W, Niu G, Shen B, Zheng Y, Gong F, Wang X, Lee J, Mulvihill JJ, Chen X, Li S. High-resolution analysis of copy number variants in adults with simple-to-moderate congenital heart disease // Am. J. Med. Genet. A. 2013. V.161A. P.3087-3094.
251. Zou J., Zhu J., Huang S. et al. Broadening the avenue of intersubgenomic heterosis in oilseed Brassica // Theor. Appl. Genet. 2010. V. 120. P. 283-290.

SOME TECHNOLOGICAL PAST, PRESENT AND ALSO FUTURE OF MODERN BIOLOGY UNTIL THE YEAR 2030 *(PART TWO)*

Chemeris A.V., Magdanov E.G., Garafutdinov R.R., Matniyazov R.T., Baymiev Al.K., Baymiev An.Kh., Bikbulatova S.M., Gimalov F.R., Vakhitov V.A.

Institute of Biochemistry and Genetics of Ufa Science Centre of Russian Academy of Sciences, Ufa, chemeris@anrb.ru

Abstract

Suggestions about the future tendencies of the genome and transcriptome sequencing and places of their realization have being made. The potential impact methods sequencing of nucleic acids (genomes and transcriptomes) belonging to the 4th and 5th generations for the detection of polymorphic states of the genomes of organisms of different levels of genetic complexity, including humans is described. The characteristic of different types of polymorphism of human DNA and evolution of their research are shown. Some attention was made to sizes of genomes and the so-called paradox C value. The forecast has been made about the number of fully sequenced genomes and transcriptomes by 2030, with breakdown by years. One approach for genotyping strains of bacteria with the assignment of unique genetic barcodes to microorganisms based on the method of RFEL (Restriction Fragment End Labelling) is described. The method of assigning various organisms DNA barcode-based polymorphism of cytochrome C oxidase is viewed. Cited literature covers nearly one hundred years.

Keywords: DNA, RNA, protein, genome, sequencing, PCR, gel electrophoresis, RFEL, DNA-chips, oligonucleotide, genomics, transcriptomics, methylomics, GWAS, CNV, VNTR, STR, SNP