Что стоит за цифрами

Сергей Козлов

Несколько дней назад по русскоязычной части Интернета разошлось громкое интервью руководителя компании «Генотек» Валерия Ильинского. Оказывается, «среднестатистический житель России 16 процентов генома унаследовал от коренных жителей Центральной России, а все остальные участки являются мозаикой составленной из фрагментов геномов характерных для обитателей других регионов». Впоследствии это высказывание превратилось в «исследователи Genotek установили, что современные россияне являются коренными русскими только на 16%». Согласно опубликованной инфографике, россияне, к примеру, на 10.8% являются потомками британцев (6% генофонда составляет «английский» и 4.8% «шотландский» компонент), на 6.3% потомками венгров, на 1.3% — пакистанцев и т.д. Говорит же все это «о том, что Россия это действительно большой плавильный котел наций» Такая трактовка вызвала удивление у многих людей, интересующихся изучением генофондов народов России. Хорошо известна относительная однородность аутосомного генофонда основной части русских. Конечно, существует изменчивость в направлении Север-Юг, но это явно не то, что подразумевал Ильинский. На вопрос генетического генеалога Владимира Гурьянова о «шотландских» участках представитель компании ответил в социальной сети «Фэйсбук» следующее:

«Вы совершенно правы, речь идет об одинаковых геномных локусах, присутствующих как в генетическом материале испытуемых, так и в ДНК шотландской популяции. Далее происходит примерно следующее: выделяем некую древнюю популяцию (назовем ее популяцией X). Сравниваем количество участков, характерных для древней популяции X, в геноме современных русских и геноме современных шотландцев. Видим, что геном современных шотландцев обогащен участками, происходящими от популяции X; принимаем популяцию X за “древних шотландцев” и на основе этого делаем заключение о доле “шотландской крови” в геноме испытуемого. О времени и обстоятельствах, сопутствующих процессу формирования современного национального генетического профиля, судить, действительно, непросто, хотя и возможно в отдельных случаях.

Под “коренными русскими” понимались участки, характерные для населения именно Центральной России, не сибиряков и не кавказцев».

Таким образом, речь в буквальном смысле идет о частичном происхождении современных россиян от популяции, предковой в первую очередь для современных шотландцев.

У меня есть предположения о том, откуда же появились эти выводы. В качестве одной из услуг для своих клиентов компания предоставляет так называемую расшифровку этнического состава («мы сравнивали кусочки Вашего генома с соответствующими фрагментами геномов представителей 36 национальных групп.»). В личном кабинете клиента выводится список этих групп (в основном обозначенных по географическому принципу), и каждой из них сопоставляется своя доля наследственности. Названия групп совпадают с приведенными на инфографике для «россиян в целом». Судя по всему, в ней попросту показан результат усреднения по клиентам компании. Однако не спешите думать, что 10.8% «британской» наследственности объясняются вкладом протестировавшихся иностранцев. Для объяснения мне понадобится довольно длинное отступление.

«Счетчики этничности»

Среди любителей генетической генеалогии распространено использование популяционных калькуляторов на базе программы Admixture (нередко их называют «этнокалькуляторами», хотя в строгом смысле слова этот термин неверен – ведь этничность «находится в голове»). Автор калькулятора задает количество предковых компонентов (далее оно обозначается буквой K), выборки и режим работы. На базе этих данных программой создается файл аллельных частот для каждого компонента, при помощи которого (а также некоторых дополнительных сервисов либо утилит) любой желающий впоследствии может «спроецировать» имеющиеся у него широкогеномные данные, свои собственные или чужие, на модель и получить «раскладку» по предковым компонентам. Сравнив их пропорции с усреднениями по современным выборкам, можно судить о степени сходства или различия с ними для тестируемого генома. Кроме того, если исходные данные были заданы адекватно, сами по себе пропорции предковых компонентов будут говорить о вкладе различных древних популяций. Например, для современных европейцев должны выделяться компоненты, соответствующие европейским охотникам-собирателям мезолита и пришедшим с Ближнего Востока в неолите земледельцам. На северо-востоке проявляется «восточносибирский» компонент. Если увеличивать K, можно выделить компонент, связанный с миграциями в Европу степных групп, и так далее. Численная доля тех или иных компонентов может несколько варьировать в зависимости от калькулятора, однако в общем и целом картина получается адекватная. Название калькулятору обычно дается, исходя из того, кто был его автором, какое количество компонентов выделяется, и текущей версии – например, MDLP K23b, Eurogenes K15new и т.д.

Конечно же, пытливые умы геномных блогеров не остановились на простых вариантах и пошли далее))) Так, Дэвидом Веселовским, известным под псевдонимами Davidski и Polako, в конце 2012 — начале 2013 года был создан в экспериментальных целях довольно специфический калькулятор Eurogenes K36. Задачей было создание инструмента, позволяющего напрямую судить не только о вкладе в геном тестируемого древних популяций, но и о долях влияния популяций, гораздо более приближенных к нашим дням. Для этого выборки современных европейцев были сгруппированы в 14 кластеров (их выбор определялся исключительно волей автора), и на их основе были заданы предковые компоненты. В программе Admixture разбиение на предковые компоненты может производиться при помощи встроенного алгоритма, тогда ищутся «естественные» компоненты, а может искусственно задаваться непосредственно исследователем. Остальные 22 компонента относятся к неевропейским популяциям.

Поскольку аутосомный генофонд большинства европейцев сложился из одних и тех же источников (разница в основном в пропорциях их влияния), эти 14 искусственных компонентов оказались родственными друг другу и перекрывающимися между собой. В результате у почти любого из европейских пользователей выделялся десяток или более компонентов. Автор разместил в своем блоге следующее разъяснение (перевод с английского мой):

«Важно понимать, что результаты нельзя воспринимать буквально. Если вы, допустим, англичанин, и получили 12% компонента «Пиренейский полуостров», это не означает, что у вас есть недавние предки из Испании или Португалии. На деле это лишь показывает, что 12% ваших аллелей схожи с образцами, использованными для выведения «иберийского» компонента. Предполагать влияние пиренейцев можно лишь в случае, если ваши значения явственно превышают результаты большинства остальных англичан.»

Ссылка: http://bga101.blogspot.ru/2013/03/eurogenes-k36-at-gedmatch.html

Как эрзя и мокша оказались «коренными русскими»

Совпадение количества компонентов в калькуляторе Дэвида Веселовского и в «расшифровке этнического состава» от Genotek неслучайно. У меня есть «этнические» результаты нескольких клиентов этой компании, и они совпадают с полученными для них же при помощи Eurogenes K36 с точностью до десятых долей процента. Лишь названия компонентов переведены на русский язык и в некоторых случаях изменены. Например, компонент East Central Euro, для выведения которого Дэвидом были использованы выборки украинцев, белорусов и литовцев, назван «Белоруссия и Украина». Компонент Central European, для выведения которого были использованы выборки венгров и нескольких американцев европейского происхождения, назван «Венгрия». Поэтому я не сомневаюсь, что компания использует именно этот инструмент, некогда выложенный его автором в открытый доступ.

Что же за выборку нам предъявил «Генотек» в качестве «коренных русских»? Согласно сведениям, предоставленным автором Eurogenes K36, для выведения компонента Eastern Euro (в переводе ставшим «центральной Россией») было использовано 14 геномов представителей «мордовских» народов (эрзя и мокша), 25 геномов русских из выборки HGDP (место сбора выборки локализуется на границе Архангельской и Вологодской областей) и 3 генома других северных русских. Если быть корректным, то назвать это можно «выборка, в наиболее доступной нам сейчас (однако очень неполной) степени представляющая дославянское население территорий от Волго-Окского междуречья до Онежского озера». Видимо, в таком качестве она и задумывалась автором. Следовательно, заявление о «коренных русских» можно назвать полной дезинформацией, что же касается более осторожного варианта определения этой выборки, как «коренное население Центральной России», то частично он верен. Однако, как уже отмечалось выше, интерпретировать полученные при помощи K36 проценты, как показатель прямого и непосредственного вклада той или иной выборки в генофонд, нельзя. Например, сам Веселовский (по происхождению он поляк, живет и работает в Австралии), получается при такой интерпретации «коренным русским» на 13,04%. Значения других предковых компонентов у него следующие:

«Белоруссия и Украина» 27,06%

«Англия» 13,08%

«Финляндия» 11,67%

«Венгрия» 11,01%

«Шотландия» 6,13%

«Балканы» 5,48%

«Чувашия» 2,94%

«Франция» 2,87%

«Пиренейский полуостров» 2,65%

«Северный Кавказ» 2,27%

«Пакистан» 1,7%

Как видно, Польша «оказалась» «плавильным котлом наций» не хуже России. То же самое можно отнести и к большинству других европейских стран. На деле же это попросту неверная интерпретация результатов.

Все это делает представленные в интервью выводы не имеющими отношения к реальности. Критике можно подвергнуть и другие моменты. Например, на инфографике представлены результаты неких «россиян в целом» — большая часть из них русские, но присутствуют и представители множества других народов. Видимо, в основном это жители крупных городов. Таким образом, данные не показывают ни результаты «среднего русского», ни даже «среднего россиянина» — это просто «средний клиент». Для науки это бесполезно, для информирования общественности вредно, зато прекрасно сработало на повышение известности самой компании.

Что же касается Eurogenes K36, после 2013 года автор инструмента не стал продолжать эксперименты в этом направлении и сейчас использует другие подходы.