Структура генофонда Европы в зеркале полногеномных маркеров

Олег Балановский

Фрагмент из книги "Генофонд Европы"

Следующий фрагмент книги О.П.Балановского «Генофонд Европы» посвящен полногеномным и широкогеномным маркерам ДНК. Это самые современные и наиболее информативные методы анализа генофонда. В первой части главы показано, как выявляемая с их помощью генетическая карта Европы соотносится с географической картой.

4.1. Особенности широкогеномного анализа

Специалист подобен флюсу – полнота его односторонняя.

Козьма Прутков

«ПОЛНОГЕНОМНЫЕ» и «ШИРОКОГЕНОМНЫЕ».

Быстрый прогресс технологий генетического анализа обеспечил возможность подробной генетической характеристики каждого образца — по всей длине его генома. Термин «полногеномные» используется в двух смыслах, различающихся степенью охвата генома.

В первом случае под «полногеномными» данными понимаются чипы высокой плотности покрытия генома, состоящие из множества (от нескольких сот тысяч до нескольких миллионов) SNP-маркеров. Эти панели маркеров точнее было бы называть широкогеномными (калька с англоязычного термина genome-wide), и именно так я буду их называть в этой книге, хотя в русскоязычной литературе их обычно называют «полногеномные».

Это нередко вызывает путаницу, потому что во втором случае под тем же термином «полногеномный анализ» имеется в виду именно расшифровка полной последовательности всего генома. Правда, и эта полнота не абсолютна – в геноме есть регионы, состоящие из часто повторяющихся последовательностей, секвенирование которых из-за этого крайне затруднено. Этих последовательностей особенно много на Y-хромосоме – они составляют более половины ее длины. Поэтому «полное» секвенирование Y-хромосомы чаще всего включает 10-15 млн.п.н., то есть примерно четверть ее длины, но и этого оказывается вполне достаточно для точных филогенетических реконструкций.

В связи с очень высокой стоимостью полного секвенирования геномов большинство «полногеномных» популяционных исследований (в том числе представленных в нашем исследовании) выполнены по «широкогеномным» данным. Но и действительно полногеномный анализ, пусть и в меньшей степени, представлен в мировой литературе и в этой книге — например, секвенирование полных древних геномов (глава 9) и результаты секвенирования Y-хромосомы (глава 8).

ПОЛНОГЕНОМНЫЕ = АУТОСОМНЫЕ?

В связи с особой важностью Y-хромосомы для популяционных работ надо пояснить, что и широкогеномные панели, и полногеномные данные, конечно, включают данные и по этой хромосоме. Но поскольку методы анализа рекомбинирующих и нерекомбинирующих систем различны, то в популяционных исследованиях широкогеномных данных информация по Y-хромосоме, а также по мтДНК и X-хромосоме, исключается из анализа. Анализируются, таким образом, лишь аутосомные маркеры, поэтому анализ широкогеномных данных является фактически синонимом анализа подробных аутосомных данных.

Нужно лишь сделать важную оговорку. Большинство широкогеномных данных разрабатывались для целей медицинской генетики и поэтому включили мало маркеров по Y-хромосоме, да и те попались в основном филогенетически мало информативные. Но две широкогеномные панели — панель Human Origin (разработанная под руководством David Reich из Гарвардского университета) и панель GenoChip (разработанная в проекте Genographic 2.0) — специально предназначались

для популяционных исследований и включили множество филогенетически информативных маркеров Y-хромосомы. Поэтому при использовании этих широкогеномных панелей данные по Y-хромосоме не игнорируются. Но и в этом случае они рассматриваются отдельно, а в основной анализ включаются лишь аутосомные маркеры.

Эта «аутосомность» широкогеномного анализа касается и полногеномного. Хотя при полном секвенировании добываются также данные по Y-хромосоме и мтДНК, они так же рассматриваются отдельно, а в основные виды анализа включаются только маркеры аутосомных хромосом.

МНОГО МАРКЕРОВ, НО МАЛО ОБРАЗЦОВ.

Высокая стоимость широкогеномного и особенно полногеномного генотипирования резко ограничивает число изучаемых образцов. Если классическими для популяционной генетики являются выборки в 70-100 образцов, а при анализе Y-хромосомы и мтДНК их иногда старались даже увеличивать (что для гаплоидных систем действительно необходимо), то выборки при широкогеномном анализе составляют обычно 10-20 образцов, а полные геномы и вовсе чаще всего одиночны. Считается, что малый объем выборки отчасти компенсируется большой подробностью генотипирования каждого образца. Но, во-первых, только отчасти, а во-вторых, это только так считается. Что действительно помогает получать надежные результаты при столь малых выборках, это то, что единицей анализа при широкогеномных исследованиях выступает не популяция, а отдельный образец. Соответственно, если популяция представлена десятком-другим образцов, и почти все они ведут себя одинаково (например, входят в один кластер на графике главных компонент и характеризуются почти одинаковым соотношением предковых компонентов на графике ADMIXTURE) — это является сильным аргументом воспроизводимости и надежности полученного результата даже при малом объеме выборки.

Но малый объем выборок представляет еще одну опасность. Ведь если в выборке в 15 образцов из популяции А многие образцы случайно окажутся от людей, в своих родословных имеющих предков из популяции Б, то для всей популяции А будет сделан вывод о ее большом генетическом сходстве с популяцией Б – и этот вывод будет неверным. Иными словами, поскольку по данным о малой выборке – 10-15 человек – зачастую судят о генофонде многомиллионного народа, решающее значение имеет, в какой степени эти 10 человек репрезентативны для всего генофонда. Это накладывает огромную ответственность на тех, кто собирает популяционные выборки и отбирает из них образцы для широкогеномного анализа. Конечно, в популяционной генетике всегда анализируются выборки, а выводы делаются о популяциях. Но при широкогеномных исследованиях контраст между выборкой и всей популяцией может быть разительным. Итак, при широкогеномных исследованиях принципиальное значение имеет качество формирования выборок.

РАЗНЫЕ ПАНЕЛИ.

Широкогеномное генотипирование технически проводится на двух платформах, предлагаемых компаниями Illumina и Affimetrix. Для генотипирования нужно иметь и оборудование от одной из этих компаний, и наборы для генотипирования, соответствующие оборудованию, и поэтому разрабатываемые теми же производителями.

Illumina за несколько лет своей работы последовательно предлагала около десятка основных наборов – для панелей маркеров, включающих от 300 тысяч до 2 миллионов SNP-маркеров, равномерно покрывающих весь геном. Кроме того, компания предлагает разработку наборов на заказ, включающих любые маркеры, нужные пользователю, но объем заказа должен быть достаточно большим (1000 и более образцов). Основные панели были разработаны для нужд медицинской генетики, ведь для картирования генов болезней как раз нужно иметь равномерное покрытие генома. Но эти же панели оказались высокоинформативны и для популяционной генетики. Из панелей, разработанных к настоящему времени на заказ, для целей популяционной генетики предназначена лишь одна – панель GenoChip, разработанная биоинформатиком Eran Elhaik для проекта Genographic 2.0. Автор этой книги и сам участвует в проекте Genographic 2.0.. И потому, зная принцип тщательного отбора маркеров – чтобы в их число были включены характеристические маркеры для каждой популяции мира, изученной ко времени создания набора, — может рекомендовать читателю не судить о качестве набора лишь по числу маркеров (которое кажется не слишком большим по современным меркам: около 130 тысяч аутосомных и около 13 тысяч Y-хромосомных маркеров).

Affimetrix также предлагает ряд панелей, но в целом они использовались в популяционной генетике меньше, чем Illumina. Однако самое первое, ключевое широкогеномное исследование Европы (Novembre et al., 2008) выполнено на одной из первых панелей Affimetrix. И набор Human Origin, широко используемый в статьях команды David Reich, занявшей в последние пару лет лидирующие позиции, тоже генотипируется на платформе Affimetrix.

АНАЛИЗ РОССИЙСКИХ ПОПУЛЯЦИЙ.

Еще с конца 90х годов ряд российских коллективов в тесном сотрудничестве друг с другом выполнял работы по изучению восточно-европейских популяций по панелям из 3-10 аутосомных маркеров. Были охвачены почти все народы Восточной Европы и Урала, а число маркеров было для того времени достаточно большим: использовались в основном микросателлиты и минисаттелиты (Popova et al., 2001; Султанаева и др., 2001; Хуснутдинова и др., 2003; Verbenko et al., 2003ab; Stepanov et al., 2011; Ахметова и др., 2006), хотя применялись и маркеры других типов (Shabrova et al., 2004; Yunusbayev et al., 2006; Соловьева и др., 2010). Основная часть этих исследований была обобщена в монографии (Лимборская и др., 2002) и моей диссертации (Балановский, 2002).

Накопленный опыт исследований отдельных аутосомных маркеров позволил российским коллективам (в сотрудничестве с рядом зарубежных, предоставлявших технические возможности генотипирования) включиться и в исследования полногеномных маркеров, когда они вошли в арсенал популяционной генетики. К настоящему времени народонаселение России и сопредельных стран, в основном благодаря работам Эстонского биоцентра и его многочисленных российских коллег, а также ряда других российских и зарубежных лабораторий, изучено на уровне, не уступающем большинству других регионов мира. По крайней мере, большинство народов России представлены хотя бы одной выборкой, генотипированной по широкогеномным маркерам.

4.2. Генетические взаимоотношения популяций:

три вариации на европейскую тему

Все то же солнце ходит надо мной,

Но и оно не блещет новизной!

Шекспир. Сонет 76.

ВАРИАЦИЯ ПЕРВАЯ: «ГЕОГРАФИЧЕСКАЯ» КЛАСТЕРИЗАЦИЯ.

Первой статьей, подробно охарактеризовавшей генофонд Европы по широкогеномным маркерам, стала работа международного коллектива, опубликованная в 2008 году в Nature [Novembre et al., 2008]. В этой статье популяции, представляющие большинство европейских стран, были охарактеризованы по панели Affimetrix из 500 тысяч SNP-маркеров. Основной результат показан на графике главных компонент (рис. 4.1).

Как говорилось, большое число SNP-маркеров позволяет надежно определять генетическое положение не только популяционной выборки, но и каждого индивидуального образца. Тем удивительнее, что даже индивиды, несмотря на заведомо большой размах межиндивидуальных различий внутри популяции, четко кластеризуются по стране происхождения. Более того, получившаяся «генетическая карта» в значительной степени воспроизводит географическую. Этот результат был вынесен в название статьи – «Гены отражают географию». Но, положа руку на сердце, этот результат не нов. Он ранее неоднократно намечался в популяционных исследованиях и зарубежной школы геногеографии [Cavallli-Sforza et al., 1994] и отечественной школы.

Другое дело, что географическая кластеризация никогда еще не проявлялась с такой рельефностью и убедительностью. Действительно, почти все ирландские образцы образуют «этническое облако», примыкающее к облаку уроженцев Великобритании – точно так же как примыкают друг к другу два этих острова. Сразу за английским облаком – опять следуя географии – располагается облако французских образцов. А от французского облака отделено небольшим промежутком (хочется сказать – отделено Пиренеями) облако испанцев, почти слившихся – на одном полуострове – с португальцами. Полноту сходства с географической картой дополняет то, что если провести прямую линию от англичан к испанцам, то французы окажутся справа от нее и на географической, и на генетической карте, а само это пространство на генетической карте пустует – как и географической оно занято Бискайским заливом.

Генетическая карта отлично воспроизводит географическую и в Средиземноморье – видны генетические облака, соответствующие Аппенинскому и Балканскому полуостровам, и промежутки между ними, соответствующие Балеарскому и Адриатическому морям. Правда, в Восточной Европе такого полного отражения географии уже не наблюдается. Там положение этнических облаков друг относительно друга хотя обычно и соответствует географическому положению народов, но пропорции генетической карты уже явно отличаются от географической.

Напомним, что очень близкий результат получен в нашем исследовании и по данным об Y-хромосоме (рис. 2.40): точно так же вместе группировались популяции Британских островов, близко друг к другу были популяции Пиренейского полуострова (аналогично Аппенинского и Балканского), соответствие географической и генетической карт прослеживалось и в Восточной Европе. Обнаружив один и тот же результат в двух обобщающих исследованиях, разных и по использованным системам, и по методам, и по изученным популяциям, нужно признать, что генофонд Европы действительно структурирован прежде всего по географическому принципу!

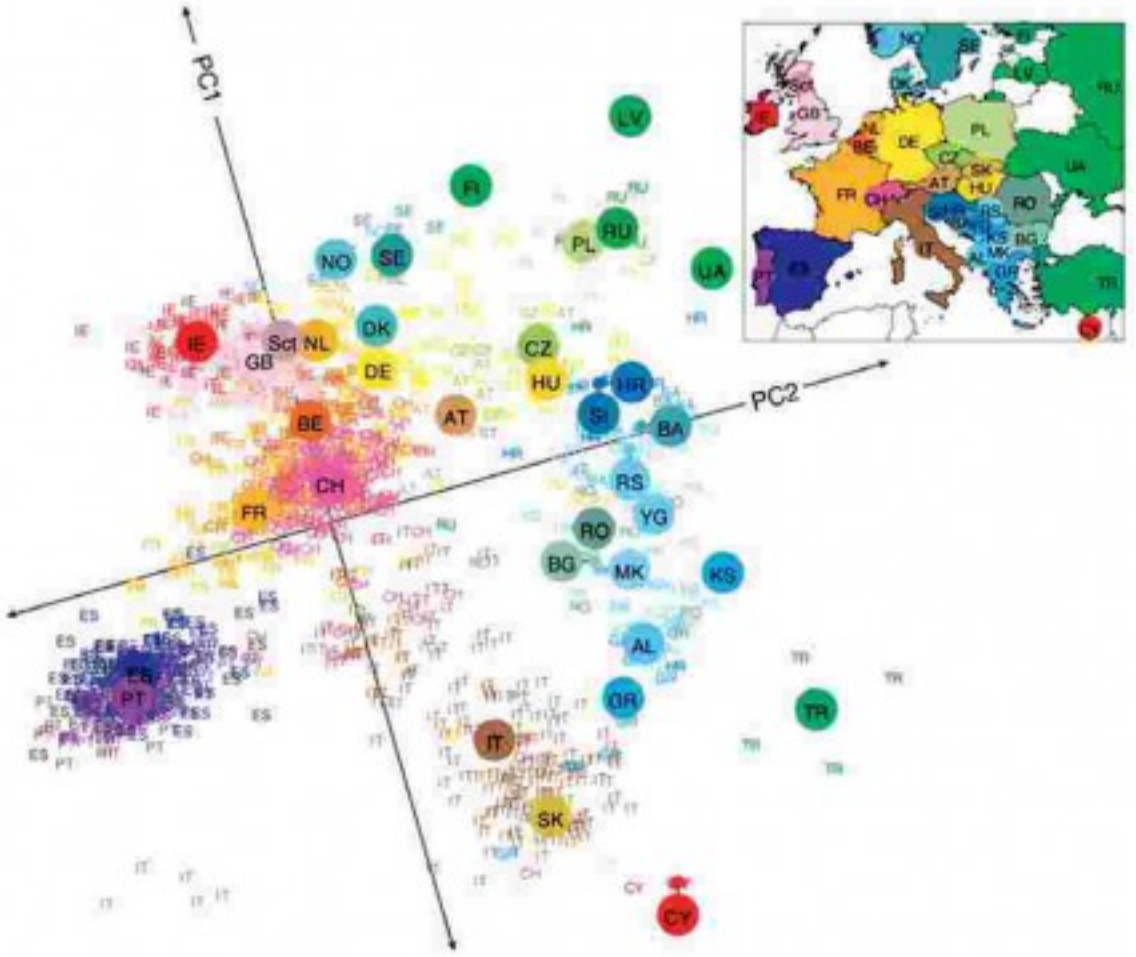

Рис. 4.1. Генетические взаимоотношения популяций Европы по данным о широкогеномных маркерах панели Affimetrix. График главных компонент, приводится по [Novembre et al., 2008].

ВАРИАЦИЯ ВТОРАЯ: ОБОСОБЛЕННОСТЬ СЕВЕРО-ВОСТОКА.

Последующие широкогеномные исследования генофонда Европы основывались, главным образом, на различных панелях с разным, все увеличивающимся, числом SNP-маркеров компании Illumina [Nelis et al., 2009; Yunusbaev et al., 2012], хотя использовались и панели Affimetrix [Xing et al., 2009]. Эти работы подтвердили вывод о примате географического фактора в структурировании аутосомного генофонда Европы по широкогеномным данным. Так, вышедшая уже в следующем году статья [Nelis et al., 2009] тоже констатировала кластеризацию по географическому принципу. Более того, роль географии была показана не только на уровне крупных популяций (уровня стран), но и внутри столь небольшой по площади страны, как Эстония (группы индивидов, выделенные на графике по критерию их генетического сходства, соответствуют их происхождению из различных районов Эстонии).

Другой важный вывод этой работы состоит в том, что популяции северо-востока Европы (представленные в этой статье главным образом южными и северными финнами, а также эстонцами) оказываются одним из трех генетических «полюсов» Европы. Степень выраженности этих полюсов такова, что все европейские популяции вместе формируют как бы трехлучевую звезду, с лучами, вытянутыми от общеевропейского центра к трем «полюсам» (рис. 4.2). В работе [Novembre et al., 2008] северо-восток был почти не представлен, — видимо, поэтому в ней не выявилась «трехлучевая» структура. А когда в работе [Nelis et al., 2009] добавились генетические контрастные финны (куда более контрастные, чем можно было бы прогнозировать из их географического положения), все остальное генетическое разнообразие Европы сплющилось в одну линию. Правда, и в

статье [Nelis et al., 2009] очень многие народы Европы отсутствовали, поэтому и трехлучевая картина не является окончательной.

Отметим, что вывод [Nelis et al., 2009] о резко выраженном своеобразии генофонда финнов по большому счету тоже не нов – не новее вывода о географической кластеризации [Novembre et al., 2008]. Это своеобразие финнов хорошо известно в популяционной генетике и объясняется сильным действием дрейфа генов на редко заселенных территориях северо-востока Европы. У медицинских генетиков есть даже термин «финские болезни» — те редкие наследственные варианты, которые за счет дрейфа генов достигли заметной частоты у финнов, а в других популяциях Европы практически отсутствуют. Финны во всех этих случаях являются наиболее подробно изученными представителями всего населения северо-востока Европы (от Балтики до Урала), и выводы о финнах можно, хотя и с определенной осторожностью, относить ко всему этому региону.

В статье [Nelis et al., 2009] проведен также анализ генетических границ. Для этого использован не геногеографический подход через локальное межпопуляционное разнообразие (описанный в главе 8), а традиционный подход совмещения данных о генетических расстояниях между популяциями и их географических координатах, реализованный в программе Barrier (использованной и нами в главе 7). Любопытно, что генетические границы (барьеры для потока генов) выявлены как между этническими облаками, так и внутри некоторых народов (например, между южными и северными финнами, между южными и северными итальянцами).

Вообще эта работа ничем не уступает по своему уровню исследованию [Novembre et al., 2008], но, появившись чуть позже, она была опубликована и в менее престижном журнале (PLoS One), и реже цитируется.

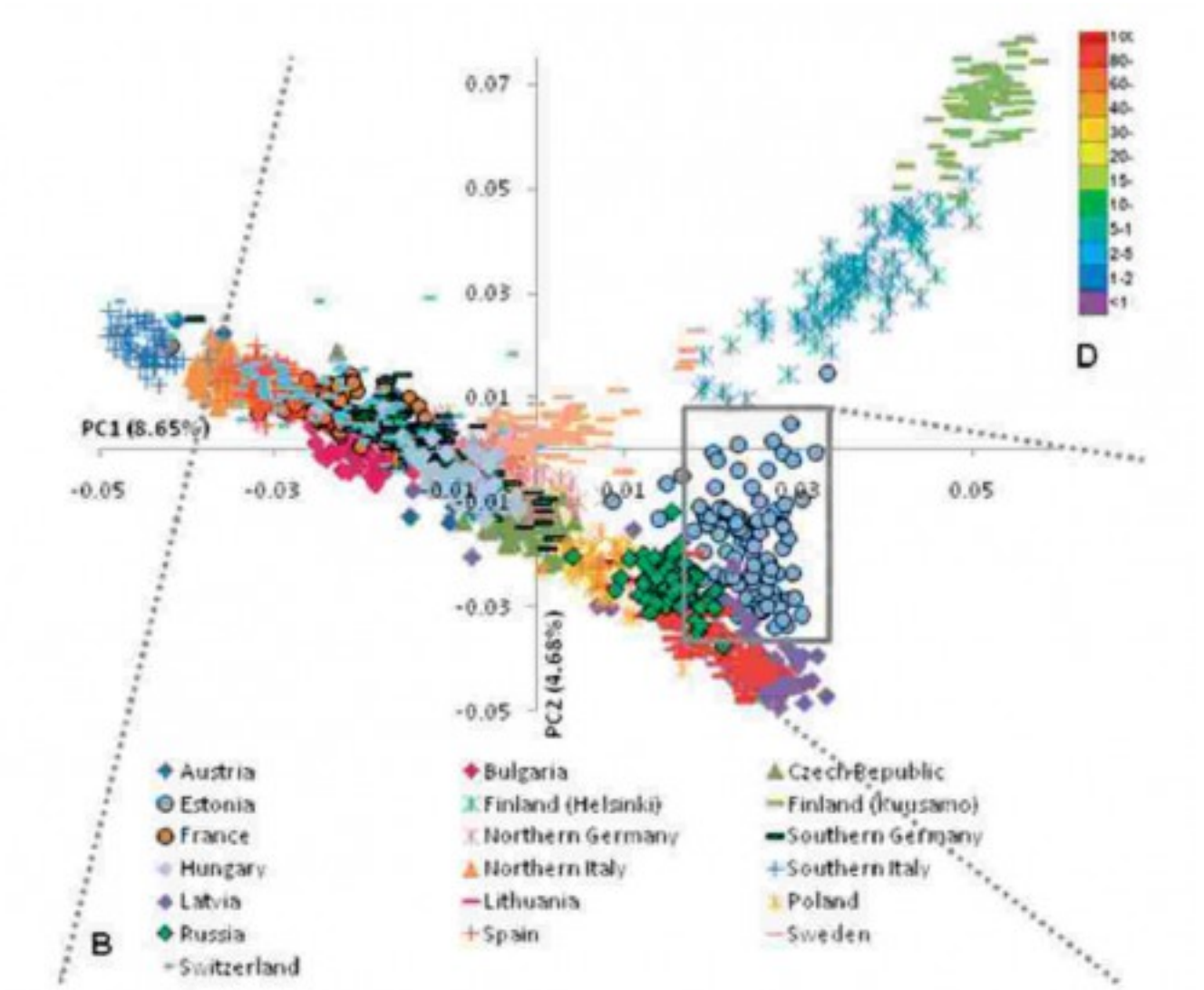

Рис. 4.2 Генетические взаимоотношения популяций Европы по данным о широкогеномных маркерах панели Illumina. График главных компонент [Nelis et al., 2009].

Последующие работы не изменили сколь-либо существенно выводов этих двух основных работ, и были посвящены отдельным регионам Европы [Filippova et al., 2012; Karafet et al., 2015; Khrunin et al., 2013 и другие]. В частности, в статье [Khrunin et al., 2013] представлены данные по 4 русским популяциям, а также финно-угорским популяциям севера Европейской части России и показано, что население крайнего северо-востока Европы (представленного в статье популяцией коми) может быть еще одним полюсом генетического разнообразия европейского генофонда.

ВАРИАЦИЯ ТРЕТЬЯ: МОСТ К БЛИЖНЕМУ ВОСТОКУ

Следующий виток исследований связан с использованием двух методических приемов: специальных «исторических» панелей маркеров и совмещением данных по современным и древним популяциям.

Если первые полногеномные популяционно-генетические исследования опирались на чипы высокой плотности, разработанные для целей медицинской генетики, то в последние годы все шире используются «прицельные» панели — широкогеномные панели маркеров, специально подобранные для целей выявления истории популяций. Наиболее подробные данные о современных популяциях опубликованы к настоящему времени по набору Human Origins в серии статей под руководством David Reich, и к рассмотрению их результатов мы сейчас и переходим. Что касается совмещения на одном графике данных по современным и древним образцам, то этот прием широко используется в тех же статьях, но мы его рассмотрим в главе 9, специально посвященной исследованиям древней ДНК.

График главных компонент, полученный по набору Human Origin, представлен на рисунке 4.3. В этом исследовании были подробно изучены не только европейские, но и ближневосточные популяции. На графике прекрасно выделяются эти два региона, популяции которых протянулись двумя параллельными линиями. Генетическими полюсами Европы являются: сардинцы, а также народы Иберийского полуострова (первый полюс), и (второй полюс) народы Восточно-европейской равнины. То есть опять паттерн в точности следует географической оси, вдоль которой с запада на восток вытянута Европа. Генетические полюса Ближнего Востока, рассматриваемого в самом широком смысле, включая и смежные регионы, – это Северная Африка и Кавказ. Обращает на себя внимание, что изменчивость генофонда Ближнего Востока (и Северной Африки) выше, чем Европы, хотя по площади территории эти регионы сопоставимы.

Два больших облака на графике — европейских и ближневосточных популяций – четко отделены друг от друга, как и географически они разделены Средиземным и Черным морями. Мостом между ними служат популяции двух народов, населяющих острова Средиземноморья – греков и сицилийцев. Острова Эгейского моря и Сицилия действительно являлись путями многих исторически документированных миграций между южным и северным побережьями Средиземного моря. В этом плане скорее неожиданно, почему население вдоль третьего пути (Гибралтарского) по этим данным разнится между собой (например, популяции испанцев и запада Северной Африки генетически удалены друг от друга).

Еще одним мостом являются популяции евреев (из Европы, Ближнего Востока и Северной Африки), многие из которых выведены на этот график. Возможно, в ходе своей сложной истории миграций они включили в себя как европейские, так и ближневосточные компоненты. Лишь евреи Грузии, Ирана и Йемена генетически почти неотличимы от основного населения этих стран – вероятно, за счет преобладания в их генофонде компонента принявших иудаизм местных уроженцев над компонентом, принесенным в ходе миграции евреев из их исторического ареала.

Любопытно положение сардинцев, находящихся в генетическом пространстве близко к основной массе европейских популяций, но не присоединяющихся к ней. Действительно, население Сардинии, как считается, сохранило генофонд неолитической миграционной волны в Европу, которая во всех других местах значительно смешалась с донеолитическим населением или приняла в себя позднейшие миграционные волны. Подтверждение этому мы увидим в главе 9, где с современным генофондом Сардинии почти совпадают геномы первых неолитических популяций Европы, реконструированные по древней ДНК.

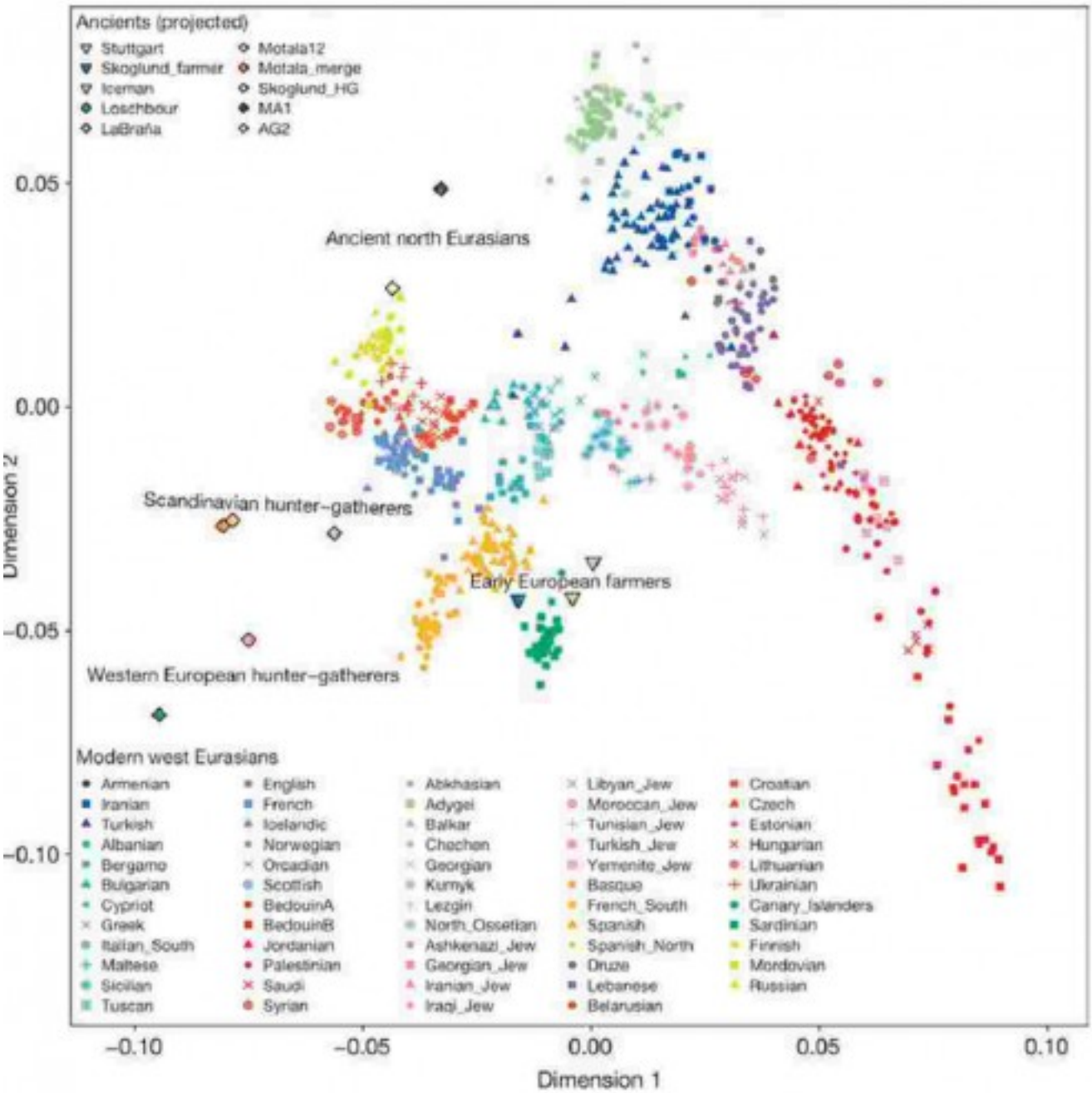

Рис. 4.3. Генетические взаимоотношения популяций Европы и Ближнего Востока по данным о широкогеномных маркерах (панель Human Origin – Affimetrix). График главных компонент [Lazaridis et al., 2014].

Итак, результаты этого третьего крупного исследования, хотя и добавляют ряд важных штрихов – в основном про взаимодействие Европы и Ближнего Востока – тоже являются вариацией на тему географической кластеризации европейских генофондов.

Дальнейшее уточнение структуры европейского генофонда с помощью широкогеномных данных может идти двумя путями. Первый путь — это увеличение числа изученных популяций, ведь в каждом из трех описанных исследований карта изученных популяций зияет многочисленными и обширными дырами. Второй путь – это расширение спектра использованных методов анализа данных, ведь до сих пор мы рассматривали лишь графики главных компонент. Оба этих пути реализованы в данной книге в региональном масштабе — для балто-славянских народов Европы (глава 6). Но прежде чем перейти к анализу славянских генофондов, рассмотрим еще одни результаты по широкогеномным маркерам в глобальном масштабе. Речь идет о применении метода ADMIXTURE, который, наряду с главными компонентами, уже стал традиционным при обработке широкогеномных данных.