Балановская Е.В.

ФГБНУ Медико-генетический научный центр имени академика Н.П. Бочкова, лаборатория популяционной генетики человека, Москворечье, д. 1, Москва, 115478, Россия

СОЮЗ АНТРОПОЛОГИИ И ПОПУЛЯЦИОННОЙ ГЕНЕТИКИ

Введение. *Российская популяционная генетика человека выросла в недрах антропологии. Но постепенно бурно развивающиеся технологии генетики стали создавать проблемы во взаимопонимании этих наук. В надежде, что это поможет укрепить давний союз антропологии и генетики, в работе предпринята попытка конкретизировать: особенности признаков, с которыми работает генетика; вопросы репрезентативности выборок для столь разных генетических признаков; специфику применения в генетике тех методов, с которыми работают обе науки. Но главное внимание уделено методу предковых компонент ADMIXTURE, хорошо известному палеоантропологам, привлекающим данные палеогенетики.*

Результаты и обсуждение. *В работе показано, насколько эти методы полезны в этнической антропологии современного населения. Приведены примеры анализа и главных компонент, и предковых компонент для разных регионов (Русский Север, Дальний Восток, Северная Евразия) и для решения разных задач. Метод ADMIXTURE может дать варианты количественной оценки для вклада расово-антропологических комплексов разного иерархического уровня, причем количественная оценка основана на данных об огромном массиве независимых генетических маркеров.*

Заключение. *Конечно же, рассмотрена лишь малая часть обширной области взаимодействия антропологии и генетики. Но если эта попытка поможет взаимной заинтересованности генетиков и антропологов в совместных исследованиях, то задачу данной работы можно считать решенной.*

Ключевые слова: этническая антропология; популяционная генетика человека; метод предковых компонент ADMIXTURE; метод главных компонент

DOI: 10.55959/MSU2074-8132-24-4-4

Введение

Союз или противостояние?

«Союз или противостояние: антропология и молекулярная генетика» – серия семинаров под таким общим названием в 2000-х годах собирала в Институте антропологии МГУ большую аудиторию антропологов и генетиков (рис. 1). Экспансия бурно развивавшейся молекулярной генетики вызывала у антропологов интерес, смешанный с опасениями. И опасения зачастую оправдывались, когда молекулярные генетики с энтузиазмом неофитов вторгались в давно обсуждаемые проблемы антропологии, не обладая необходимым научным багажом и тактом. Поэтому вместе с Е.З. Годиной мы и создали площадку для обсуждения той проблематики, где интересы обеих наук пересекались, но технологии и стиль мышления были различны.

За прошедшие четверть века проблема взаимодействия антропологии и генетики приобрела новые черты. Стремительный прогресс в изучении древней ДНК создал в России обширное научное пространство для обоюдного интереса и естественного объединения достижений палеоантропологии и палеогенетики. Робкие шаги использования данных генетики сделаны в российской спортивной антропологии. Но в области этнической антропологии, которая ранее столь органично включала в себя популяционную генетику, теперь антропологи и генетики ограничиваются вежливыми поклонами, но оба потока исследований протекают независимо друг от друга.

Это вызывает недоумение, поскольку именно этническая антропология проявляла самый активный интерес уже к первым шагам популяционной генетики человека. Все три «кита» российской антропологии – Яков Яковлевич Рогинский, Виктор Валерьянович Бунак, Георгий Францевич Дебец – включали в свои глубокие исследования данные по этническим вариациям первых генетических систем AB0 и Rhesus. Их работы до сих пор могут служить образцом элегантного объединения мощного антропологического фундамента с воздушными строениями популяционной генетики в первой половине XX века. Конечно, данные популяционной генетики тогда были слабо информативны – технологии позволяли изучать только малый набор генетических маркеров, причем находящихся под давлением естественного отбора. Но насколько осторожно, корректно и креативно они использовались в классических работах Я.Я. Рогинского, В.В. Бунака и в обширных планах Г.Ф. Дебеца.

Поэтому закономерно, что российская школа популяционной генетики человека выросла именно из антропологии. Ее основатель – Юрий Григорьевич Рычков – на кафедре антропологии МГУ сумел совместить обе системы мышления: популяционный стиль интерпретации данных органично вошел в антропологию, а огромный массив знаний, накопленный этнической антропологией, и ее традиция глубоких обобщений определили сочетание глубины, междисциплинарности и креативности, ставших определяющей чертой российской популяционной генетики человека. Однако органичный синтез этих двух наук, состоявшийся у истоков популяционной генетики в России, со временем распался. Их резкий разлом произошел при взрыве возможностей молекулярной генетики.

EAA & МОГиС

ЕВРОПЕЙСКАЯ АНТРОПОЛОГИЧЕСКАЯ АССОЦИАЦИЯ (Российское отделение)
МОСКОВСКОЕ ОБЩЕСТВО ГЕНЕТИКОВ и СЕЛЕКЦИОНЕРОВ (МО ВОГиС)

семинар

и круглый стол на тему

«Антропология и молекулярная генетика: союз или противостояние?»

Ведущие: акад. РАН Т.И. АЛЕКСЕЕВА
член-корр. РАМН Е.К. ГИНТЕР

Друзья! Первая встреча на нашем семинаре (15 декабря 2000) выявила две стороны медали.

Во-первых, противостояния нет – мы все стремимся к союзу и сотрудничеству.

Во-вторых, увы, союза тоже нет – мы говорим на разных языках, хотя очень хотим понять друг друга.

Поэтому, приглашая Вас на вторую встречу (7 февраля 2001), мы попытаемся учесть все высказанные пожелания. Основной доклад – о молекулярной генетике коренного населения Америки - будет сделан антропогенетиком. Надеемся, что это позволит соединить в одном докладе три языка – антропологии, популяционной генетики и молекулярной генетики. Но чтобы упростить столь сложную задачу, сначала будут прочитаны четыре мини-лекции о тех первооснова, которые необходимы для полноценного и корректного обсуждения основного доклада самыми разными специалистами. В них будут даны необходимые для обсуждения понятия и факты: по молекулярной и популяционной генетике, антропологии и палеогеографии. А Круглый стол откроется докладом о возможных погрешностях при оценках времени происхождения рас человека по ДНК-маркерам.

Мы надеемся, что так постепенно найдем общий язык, и встреча наших наук состоится!

В программе:

ЧАСТЬ ПЕРВАЯ: **ПЕРВООСНОВЫ**

1. ***МОЛЕКУЛЯРНАЯ ГЕНЕТИКА*: О.Л. КУРБАТОВА. ДНК маркеры – это так просто.**
2. ***ПОПУЛЯЦИОННАЯ ГЕНЕТИКА:* Е.В. БАЛАНОВСКАЯ. Генетическая память и структура популяции.**
3. ***ЭТНИЧЕСКАЯ АНТРОПОЛОГИЯ*: И.В. ПЕРЕВОЗЧИКОВ. Лики коренных народов Сибири и Америки.**
4. ***ПАЛЕОГЕОГРАФИЯ:* А.А. ВЕЛИЧКО. Древний климат Сибири и Америки.**

ЧАСТЬ ВТОРАЯ: **ТЕМА**

5. **В.А СПИЦЫН. Заселение Америки по данным молекулярной генетики.**

14.00-15.00: чай, кофе;

ПРОСМОТР И ОБСУЖДЕНИЕ СТЕНДОВЫХ ДОКЛАДОВ.

ЧАСТЬ ТРЕТЬЯ: **КРУГЛЫЙ СТОЛ**

6. **Л.А.ЖИВОТОВСКИЙ. Оценка времени дивергенции рас человека по ДНК-маркерам**
7. **КРУГЛЫЙ СТОЛ. «Антропология и молекулярная генетика: союз или противостояние?»**

Оргкомитет семинара

Рисунок 1. Программа одного из семинаров «Союз или противостояние: антропология и молекулярная генетика»

Figure 1. Program of one of the seminars «Union or Opposition: Anthropology and Molecular Genetics»

Тогда в популяционную генетику, ставшую очень модной наукой, пришли специалисты из других областей знания – биохимии, молекулярной генетики, медицины и т.д., не имеющие опыта работы с традициями этнической антропологии. Быстрое развитие генетических технологий приводило к быстрой смене «моды» на генетические маркеры, к беспрестанной погоне за новыми методами биоинформатики. Антропологам было сложно их отслеживать, а генетики, утратив органичную связь с этнической антропологией (и комплексом всех тесно взаимосвязанных с ней гуманитарных наук – этнологией, археологией, историей), постепенно утрачивали и традицию междисциплинарности, в том числе синтеза данных антропологии и генетики.

К счастью, прогресс в изучении палеоДНК приостановил это печальное расхождение наук, поскольку без междисциплинарного синтеза палеогенетики, палеоантропологии и археологии интерпретация столь ценных данных палеоДНК невозможна. Но многие другие возможности популяционной генетики человека, важные для этнической антропологии и палеоантропологии, пока остаются за бортом интересов антропологов. Во многом это связано с некоторым сомнением в основательности методов популяционной генетики, отчасти с «гордостью и предубеждением». Поэтому, на мой взгляд, стоит рассмотреть простоту, информативность и возможности хотя бы одного из таких методов, широко используемого и в палеогенетике, и в популяционной генетике современного населения – метода предковых компонент ADMIXTURE. Что, как не празднование юбилея МГУ, может помочь восстановлению доверия между антропологией и популяционной генетикой, – доверия, родившегося именно в удивительной атмосфере Московского университета?

Результаты и обсуждение

Системы признаков: что смотрим?

Различия в системах признаков антропологии и популяционной генетики всем известны. Но когда обе науки используют аналогичные методы (главных компонент, многомерного шкалирования, дендрограмм), то при сравнении результатов различия в природе признаков и их особенностях порой упускаются.

Конечно, основное различие – в количестве генов, ответственных за признаки. Антропология обычно использует полигенные признаки, причем количество генов, отвечающих за те или иные признаки, неодинаково. К тому же в рамках конкретной антропологической системы (одонтология, дерматоглифика, морфология) ее признаки зачастую скоррелированы. Более того, такие системы признаков, как правило, дают разные количественные оценки сходства популяций. Иными словами, они предлагают разные модели микроэволюции популяций. Но синтез разных вариантов реконструкции истории популяций, полученных с разных «точек зрения» разных антропологических систем, дает более надежную картину.

Популяционная генетика использует «моногенные» признаки, причем обычно тщательно избавляется от «физиологической» (но не «исторической») скоррелированности своих признаков. Есть две основные системы генетических маркеров – аутосомного генома и однородительских генетических маркеров.

С аутосомным геномом все просто – это те же признаки, что использует и антропология, только они «моногенны» и независимы друг от друга. Этих признаков сначала было немного («классические» генетические маркеры), теперь их число выросло до полного генома, но суть не меняется. Поэтому стоит напомнить лишь термины. Аутосомные генетические маркеры, использующиеся в популяционных исследованиях, в настоящее время в основном изучаются по стандартным наборам (панелям), которые обычно называются «широкогеномными панелями» (от десятков до сотен тысяч маркеров) или «полногеномными панелями» (миллионы маркеров); всегда указывается число маркеров в каждой стандартной панели, и на сайтах есть их конкретный список. Все стандартные аутосомные панели включают только однонуклеотидные полиморфизмы (SNP, Single Nucleotide Polymorphism, в разговорном жанре – снипы; также употребляется сокращение SNV – Single Nucleotide Variant), представляющие собой отличия в последовательности ДНК на один нуклеотид (A, T, G, C). В стандартных панелях SNP подбирались таким образом, чтобы они были примерно равномерно распределены по всему аутосомному геному.

С однородительскими генетическими маркерами – все несколько своеобразнее. Признаки митохондриального (однородительского) генома утратили свое первоначальное значение: слишком мал геном митохондрий, что резко снижает его эффективность. Но другая однородительская система признаков – Y-хромосомы – не теряет своей

информативности даже на фоне полногеномных исследований аутосомного генома. Эта информативность связана с чрезвычайно полезным свойством однородительских маркеров: поскольку они не разбиваются кроссинговером, то передаются из поколения в поколения единым комплексом («паттерном») – гаплогруппами. Каждая гаплогруппа характеризуется собственным неразрывным комплексом маркеров, возникшим в ходе генетической истории. Редкие мутации создают новые гаплогруппы. Поэтому при статистическом анализе именно гаплогруппа выступает как единица анализа. Все гаплогруппы – ветви единого филогенетического древа, то есть связаны между собой тем или иным исторически сложившимся родством, прослеживаемым от Y-хромосомного Адама до современности. И крайне важно не сводить Y-хромосому просто к еще одной системе признаков, а максимально использовать ее уникальные возможности.

Конечно, систематика гаплогрупп Y-хромосомы – например, такое название как «G2-L1264(xYY9632, YY1786, YY1215)» – может навести тоску на любого. Хотя на самом деле это просто означает: что данная веточка древа Y-хромосомы относится к большой гаплогруппе G2, в пределах которой сидит на большой ветке, которая определяется важным маркером L1264, но не входит в три другие ветки, которые определяются каждая своим маркером – xYY9632, YY1786 или YY1215. Однако антропологам вовсе нет необходимости погружаться в такую неудобоваримую генетическую таксономию. Для палеоантропологии важно просто отслеживать, где и когда возникают интересующие их гаплогруппы. Для этнической антропологии – где и когда они распространились. И крайне важно использовать датировки Y-хромосомы – это уникальная возможность отделить потомков от предков, найти «прародину» и зафиксировать время миграций.

Для датировок Y-хромосомы используются обе «стрелки» молекулярных часов.

Y-SNP маркеры, по которым и определяются гаплогруппы, – это «часовая стрелка» микроэволюции. Она зависит от скорости обычных мутаций – замены одного нуклеотида на другой.

Y-STR маркеры – «минутная стрелка» молекулярных часов – работает иначе. STR (Short Tandem Repeat) – это короткий «мотив» (тандем), состоящий из нескольких нуклеотидов (как из нот). И в зависимости от того, сколько раз «мотив» повторен, получаем показатель конкретного Y-STR маркера (рис. 2). Мутации Y-STR маркеров – увеличение или уменьшение числа повторов данного «мотива» – происходят значительно чаще, чем замены одного нуклеотида в Y-SNP, поэтому Y-STR эффективнее работают для датировок относительно недавних событий генетической истории и потому именуются «минутной» стрелкой молекулярных часов. Чем больше Y-хромосома изучается полногеномными методами, чем больше мы узнаем о «мелких»

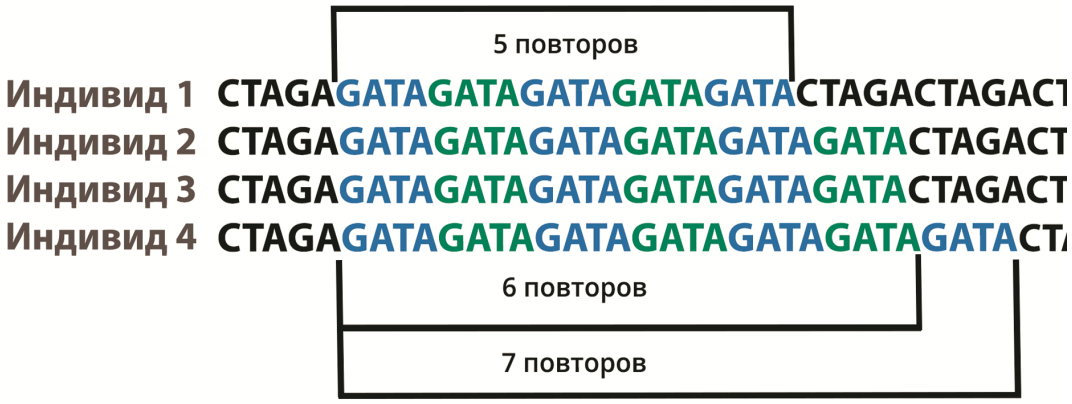

Рисунок 2. Схема определения варианта Y-STR маркера
Figure 2. Scheme for determining the variant Y-STR marker

Примечания. Короткий «мотив», состоящий всего из четырех нуклеотидов GATA, повторен у разных индивидов разное число раз. Индивиды 2 и 3 по данному Y-STR маркеру идентичны (у них по 6 повторов GATA), но отличаются от индивида 1 (5 повторов GATA) и от индивида 4 (7 раз повторен «мотив» GATA).

Notes. A short "motif" consisting of only four GATA nucleotides is repeated a different number of times in different individuals. Individuals 2 and 3 are identical according to this Y-STR marker (they have 6 GATA repeats), but differ from individual 1 (5 GATA repeats) and from individual 4 (the GATA “motif” is repeated 7 times).

веточках древа Y-хромосомы, тем больше появляется возможностей связать обе стрелки микроэволюции, тем точнее датировки и интерпретации генетической истории.

Проблемы выборки: какие критерии?

Размер выборки. Проблема определения разумного и достаточного размера выборки овеяна легендами и предрассудками, которые мы часто не осознаем. Но причина этого, на мой взгляд, лишь в наших привычках и стереотипах. Когда мы долго работаем с одними и теми же признаками, когда становятся столь привычными требования к размеру выборки именно по «нашим» признакам, то они по забывчивости переносятся на иные признаки. В популяционной генетике быстрая смена систем генетических признаков привела к тому, что таких «привычных» требований к размеру выборки установилось несколько: N=50 для «классических» генетических маркеров (группы крови, иммунно-биохимические маркеры и т.д.); N=70 для однородительских маркеров (Y-хромосома, мтДНК); N=10 для относительно небольших наборов аутосомных маркеров (в диапазоне 200-400 тысяч SNP); N=5 для полногеномных исследований.

Откуда такое разнообразие? Ответ был дан математиками еще в середине XX века как бы в предвидении быстрого роста возможностей генетики. Например, в [Nei, Roychoudhury, 1974] показано, что оптимальный размер выборки и число изученных маркеров связаны обратно пропорциональной связью: чем больше маркеров включено в исследование, тем меньше размер репрезентативной выборки. Поэтому даже для классических маркеров необходимый размер выборки значительно колебался: от 30 до 100 индивидов в зависимости от числа анализируемых генетических маркеров. Когда же генетики перешли к полногеномным исследованиям и стали использовать стандартные панели аутосомных маркеров от 100 тысяч до многих миллионов маркеров, то размер репрезентативной выборки стал стремиться к единицам индивидов. Но при этом требования к выборке уже стали определяться и иными критериями. Рассмотрим их.

Подразделенность. Если чисто статистически для характеристики популяции по полным геномам достаточно единичных образцов, то встает сразу вопрос, насколько достойно эти единичные образцы представляют свою популяцию. Если речь идет не о скромной локальной популяции, а об этносе или субэтносе, то возникает необходимость учесть, что они обычно являются подразделенными популяциями, обладают внутренней структурой. В этом случае возникает обязательное требование к выборке – включить в нее представителей основных субпопуляций этноса, генофонды которых могут различаться. К счастью, выборку по аутосомным ДНК-маркерам при полногеномных исследованиях легко контролировать с помощью PCA (рис. 3): поскольку на графике каждый индивид занимает свое положение, мы своими глазами видим, насколько компактна выборка. Более того, мы видим на графике даже степень «метисированности» отдельных геномов – например, все алеуты в настоящее время в разной степени метисированы с европейскими индивидами, и поэтому мы видим на графике цепь геномов, протянувшуюся от популяций Дальнего Востока в Европу, а вот геномы эвенов (которые метисированы с представителями коренного населения Дальнего Востока) тянутся от популяций нанайцев и нивхов к популяциям чукчей и коряков (рис. 3).

Важно учитывать, что подразделенность популяции выражается не только в явном виде как ряд локальных географических субпопуляций в пределах подразделенной популяции, но и в неявном виде. Например, даже городское население (Махачкалы [Балановская с соавт., 2024] или Москвы [Курбатова с соавт., 2021]) структурировано: в разных секторах города преобладают те или иные этнические группы, причем велика изменчивость такой структуры городского населения. Поэтому исследователи пришли к выводу, что при формировании генетических баз данных для мегаполиса должны создаваться отдельные «референтные популяции» для каждой этнической группы [Курбатова с соавт., 2013]. Причем их динамика столь высока, что эти данные необходимо постоянно обновлять: «Различия параметров миграции в двух возрастных группах указывают на возможность динамики этнорегионального состава населения Москвы в последующих поколениях, что непременно вызовет изменение частот многих генетических маркеров… …особенности миграционных процессов в Москве указывают на необходимость своевременного обновления и актуализации генетических баз данных для целей ДНК-идентификации в мегаполисе» [Удина с соавт., 2022, с. 1331-1332].

Критерий трех поколений. Именно для того, чтобы отобранные для анализа образцы ДНК репрезентативно представляли изучаемую популяцию, в популяционной генетике принято

правило трех поколений: в выборку включаются только индивиды, в генеалогии которых на протяжении трех поколений все индивиды принадлежали к данной популяции и данному этносу (субэтносу). Правило основано на том, что если «чужаки-мигранты» уже оставили в популяции внуков и правнуков, то «чуждые» геномы уже прочно вошли в популяцию и включены в ее современный генофонд. Поскольку у мигрантов динамика миграций намного интенсивней, чем для «укорененного» населения, это правило позволяет избежать включения в выборку тех индивидов, которые в скором времени могут поменять данную популяцию на иную и не оставят значимого следа в генофонде.

Неродственность. Еще один критерий отбора индивидов для включения в выборку – отсутствие их родства как минимум на уровне трех поколений. По этому критерию требования генетики и антропологии серьезно различаются. Наличие родственников в выборке смещает ее объективные характеристики. Поэтому генетики даже в небольшой локальной популяции стремятся включить в выборку индивидов с разных концов даже одного села, поскольку при традиции «хоть за курицу, но на соседнюю улицу» разные концы села могут генетически несколько отличаться. Конечно, даже тщательное отслеживание генеалогий не позволяет учесть все. К счастью, при использовании современных полногеномных панелей маркеров Y-хромосомы и аутосомного генома методы биоинформатики позволяют устанавливать фильтры, отсеивающие образцы близких родственников, и анализ проводится для неродственных индивидов.

Изоляты. Исключение из всех правил составляют малые изолированные популяции – в них все индивиды связаны сложными родственными связями. Для таких популяций собираются субтотальные выборки, включающие основную часть их населения с исключением ближайшего родства. И крайне важно подчеркнуть, что репрезентативный размер такой выборки не подчиняется правилам статистики. Эти правила по умолчанию предполагают, что популяция обладает неограниченным размером. Поэтому к малым популяциям нельзя применять и статистические критерии достоверности различий: если выборки субтотальны, то они уже по определению являются репрезентативными, в чем многие годы убеждал антропологов профессор Юрий Григорьевич Рычков.

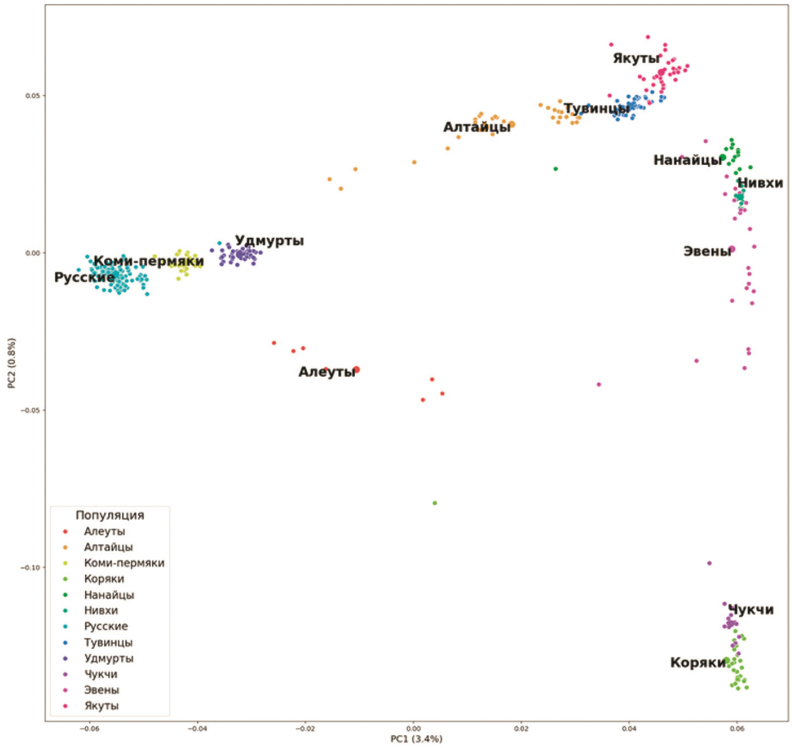

Рисунок 3. Размещение в пространстве 1 и 2 главных компонент (PC) индивидуальных аутосомных геномов и их популяционных центроидов.

Figure 3. Spatial placement of the first and second principal components (PC) of individual autosomal genomes and their population centroids

Методы статистики те же: но что-то не так?

Этот вынужденно краткий обзор признаков и выборок, с которыми работает современная популяционная генетика, важен для рассмотрения расхождений в оценке генетиками и антропологами результатов одних и тех же статистических методов.

Например, при анализе главных компонент (PCA) вклад 1 главной компоненты (PC – Principal Component) для антропологических признаков может составлять десятки процентов, а для признаков генетики – обычно единицы или доли процентов. Но такие различия вовсе не говорят, что результат, полученный по антропологическим признакам, надежнее. Доля (%) изменчивости, объясняемой новым признаком – главной компонентой, зависит от двух параметров.

Во-первых, от степени скоррелированности признаков. Признаки антропологии зачастую «физиологически» скоррелированы. Но при расчете по геномным данным проводится процедура, стремящаяся избавиться от той корреляции, которую условно можно назвать «биологической»: корреляции R > 0,2 не включаются в расчет PCA.Такая процедура обеспечивает то, что распределение плотности покрытия генома при генотипировании не влияет на результат PCA. При этом учитывается взаиморасположение генмаркеров на хромосоме, позволяющее снять скоррелированность тесно сцепленных генмаркеров: исключаются корреляции, вызванные близким расположением маркеров друг к другу (например, при расчете графика PCA, представленного на рис. 3, стандартная процедура биоинформатики для фильтрации сцепленных маркеров использовала «окно» размером 1500 маркеров: в пределах этой группы маркеров удалялись тесно сцепленные маркеры, после чего «окно плывёт» – сдвигается еще на 150 маркеров по положению в геноме, и процедура повторяется).

Поэтому остальные корреляции уже можно считать не связанными с положением на хромосоме и с биологическими функциями генмаркеров: остается корреляция только маркеров, расположенных далеко друг от друга, то есть связанных условно «историческими» корреляциями, возникшими в ходе генетической истории популяции. Процедура избавления от корреляций R > 0,2 по признакам генетики может приводить к уменьшению доли (%) изменчивости, объясняемой главными компонентами, но вселяет надежду, что описываемая изменчивость задана историей популяции, а не функциональными особенностями использованных генмаркеров. Антропологи не избавляются от «биологической» корреляции и, как правило, не используют метод PCA для анализа межпопуляционной изменчивости – для этого в антропологии применяется канонический дискриминантный анализ, выявляющий межгрупповые («исторические») корреляции на фоне внутригрупповых («физиологических»).

Второй параметр – число исходных признаков. В антропологии обычно используется не более нескольких десятков исходных признаков, зачастую связанных физиологически или морфологически. В популяционной генетике при анализе геномных данных в анализ PCA включаются уже миллионы исходных признаков, причем в основном независимых. Такое различие в числе «степеней свободы» и может приводить к различиям в % вкладе главных компонент: в антропологии объекты исследования могут быть уже описаны в значительной мере, тогда как в генетике остаётся ещё много новых источников изменчивости. Поскольку число главных компонент равно числу исходных признаков, и каждая компонента включает какую-то долю общей изменчивости всех признаков, это создает устойчивую тенденцию к тому, что на первые главные компоненты генома приходятся не десятки, а единицы или доли % общей дисперсии. Поэтому можно только радоваться, когда в PCA по генетическим данным первая главная компонента вбирает в себя только доли процента от общей изменчивости – это значит, что признаков очень много (сотни тысяч и миллионы), они «биологически» независимы и отражают историю популяций.

Для PCA также нет требований нормальности распределения [Айвазян с соавт., 1989], требований «число признаков меньше числа наблюдений» и ряда других. Но доля объяснённой дисперсии первых компонент может указывать на важные «исторические» корреляции. Например, на приводимом графике PCA (рис. 3): 1PC (3,4% общей дисперсии) отражает изменчивость между европеоидами и монголоидами; 2PC (0,8%) демонстрирует различия между народами Дальнего Востока; 3PC (0.6%) нацелена на различия между европейскими популяциями; начиная с 4PC и 5PC график долей PC выходит на плато. Поскольку график PCA (рис. 3) построен по данным о 320 170 генетических маркерах, то при равной значимости всех ком-

понент (взаимной некоррелированности всех признаков) вклад каждой из них (PCequal) в общую изменчивость (в %) составлял бы PC_{equal} = 1 : 320170 x 100 = 0,0003%. И мы видим, что вклад 1PC> PC_{equal} в 11 000 раз, а 2PC > PC_{equal} в 2700 раз. Если бы мы использовали всего 100 исходных признаков, то PC_{equal} = 1%, и тогда даже при максимально возможном вкладе 1-й компоненты (100%): 1PC> PC_{equal} всего в 100 раз (а не в тысячи, как на рис. 3). Такой нестрогий расчет показывает, что в общем случае (особенно при очень большом количестве признаков) при оценке целесообразности применения PCA и интерпретации полученных результатов следует ориентироваться не только на величину вклада каждой компоненты (%), но и на соотношение вкладов главных компонент.

Еще одно кажущееся отличие использования метода PCA – в оценке нагрузок на главные компоненты. В антропологии этот параметр играет важную роль, но в генетике он обычно используется лишь при анализе небольшого числа исходных признаков, когда в задачи исследования входит изучение роли отдельных генетических маркеров. При большом числе исходных независимых признаков (например, в геномных исследованиях) для количественной оценки влияния отдельных генетических маркеров на исследуемый параметр (например, заболеваемость) используются иные методы биоинформатики (например, полногеномный поиск ассоциаций GWAS – Genome-Wide Association Study).

Более сложным для антропологии – но лишь психологически – могут быть графики PCA, полученные для полногеномных данных (рис. 3), где единицей наблюдения выступает индивидуальный геном, а не группа населения. При этом расчет PCA не использует информацию о том, какой образец к какой популяции принадлежит, – каждый геном занимает свое место в пространстве PCA независимо от популяционной принадлежности. Но на графике PCA указан также и центроид (рис. 3) – средний показатель для группы геномов (например, геномов одного этноса или субэтноса), который и служит привычной характеристикой популяции об ее положении в пространстве главных компонент. И тогда графики главных компонент несут намного больше информации: мы видим, насколько индивидуальные геномы отклоняются от центроида своей популяции. Это помогает нам оценить, насколько генетически гомогенна данная популяция (например, как удмурты или коряки на рисунке 3) или же гетерогенна (например, как эвены и алеуты на рисунке 3). Поэтому график PCA по геномным данным становится максимально информативным, позволяя одновременно видеть и межпопуляционное, и внутрипопуляционное разнообразие генофондов. Если учесть, что, например, данный график PCA (рис. 3) основан на информации о 443 геномах, изученных по 4 559 465 генетическим маркерам (после всех фильтров в анализ включены 320 170 SNP), то такую информативность можно считать убедительной.

Конечно, все методы имеют недостатки, и у столь любимого генетиками метода PCA их тоже немало. Самым неприятным, на мой взгляд, является субъективность выбора обсуждаемых PC и графика для интерпретации: 1PC-2PC, или 1PC-3PC, или 2PC-3PC и т.д. Для такого выбора можно использовать оценку числа главных компонент по числу обусловленности [Dormann et al., 2013; Mirkes et al., 2020]: отношение долей 1PC и последующих, когда для анализа оставляются только те компоненты, для которых это отношение не превосходит некоторого критического значения. На основе численных экспериментов предложено значение 10, но при анализе аутосомного генома (где используются мощные фильтры, в том числе и для корреляций), это критическое значение должно быть ниже. Например, на графике PCA (рис. 3) критерию 10 соответствуют все первые 10 главных компонент (1PC = 3,350; 2PC = 0,771; 3PC = 0,602; 4PC = 0,447; 5PC =0,392; 6PC = 0,385; 7PC = 0,364; 8PC = 0,358; 9PC = 0,354; 10PC = 0,343). Поэтому его надо дополнить методом с анализом излома кумулятивной кривой, которая показывает, что после 3PC график выходит на плато (PC ≈ 0,4).

Поэтому много более объективным зачастую является график многомерного шкалирования (MDS). Но и у него есть недостаток – чем больше популяций включены в анализ, тем сложнее может быть переход от многомерного пространства к двумерному представлению расположения популяций на графике. Однако этот недостаток легко компенсировать, если при интерпретации MDS постоянно ориентироваться на матрицу реальных генетических расстояний, лежащую в основе анализа MDS. И, конечно, наиболее объективную картину дают публикации, в которых приведены все основные графики главных компонент, график многомерного шкалирования и матрица генетических расстояний. Такой комплекс позволяет читателю получить наиболее полную и надежную картину сходства и различий генофондов.

Предковые компоненты: магия или расчет?

Один из наиболее популярных приемов анализа аутосомного генома – метод предковых компонент ADMIXTURE – крайне важен не только для палеоантропологии, но и для этнической антропологии. Наряду с анализом главных компонент он давно стал общепризнанным и базовым методом описания генетической структуры популяций по данным об аутосомном геноме. Метод ADMIXTURE создает гипотезу (модель), отвечая на вопрос: каков вклад разных предковых источников в данный индивидуальный геном. И подчеркнем: каждая модель ADMIXTURE для данного числа предковых компонент рассчитывается независимо для той же самой совокупности индивидуальных геномов. Меняется единственный параметр – число предковых компонент k, задаваемых исследователем для данной модели, то есть гипотезы, что в формировании каждого индивидуального генома приняло участие такое-то число (k) гипотетических предковых генофондов.

На входе расчета – только k, формальный номер образца индивида и характеристика его генома, полученная по данной геномной панели – большому набору однонуклеоидных замен (SNP). На выходе работы программы – для каждой модели (т.е. для каждого числа k предковых компонент) выдается визуальное представление вкладов всех предковых компонент всех моделей в виде разноцветного графика (рис. 4А) и таблица с указанием, какую долю составляет каждая предковая компонента (АС – Ancestral Component) в геноме каждого индивида (вклад каждой АС может варьировать от 0 до 1, суммарный вклад всех АС равен 1). Совокупность всех моделей (k = 2, k = 3, k = 4 и т.д.) показывает постепенное (двигаясь от k = 2 ко все большим значениям k) выделение все более дробных вариантов классификации геномов по их происхождению. Эта процедура аналогична выделению все более дробных таксономических единиц, двигаясь последовательно от ствола всего с двумя крупными ветвями ко все более мелким таксономическим подразделениям (например, как в антропологической классификации – от деления на крупные расовые стволы к локальным антропологическим типам).

Количественная оценка, минимум субъективности. Единственный параметр, который задает исследователь для данного набора геномов – число предковых компонент k. При этом он не ограничивается «предпочтительной» гипотезой, а последовательно проводит моделирование для максимального спектра моделей-гипотез и видит все переходы реализации гипотез: от самых простых до самых детализированных. Поэтому результаты метода можно рассматривать как максимально объективные, поскольку в своей совокупности они не зависят от гипотезы исследователя: мы получаем для целой серии моделей (последовательной серии гипотез о числе предковых компонент) количественную оценку вклада каждой предковой компоненты в каждый индивидуальный геном. При этом важно, что генетический состав каждой компоненты в каждой модели определяется программой, исходя всегда из одних и тех же первичных данных (изученных геномов).

На практике анализ проводится последовательно для набора разных (2 = k ≥ 20), на графике представляется вся совокупность результатов, где каждая строка – один набор значений предковых компонент при данном k, а каждый столбец – один и тот же индивидуальный геном при разных гипотезах относительно числа предковых компонент, участвовавших в его генетической истории.

Для удобства читателя эти индивидуальные геномы обычно группируются по этнической или региональной принадлежности (рис. 4А) – от перемены мест столбцов (индивидуальных геномов) ничего не меняется, кроме удобства восприятия. С той же целью – удобства восприятия и описания – каждой предковой компоненте АС дается условное название: обычно его дают по той популяции, где вклад данной АС максимален. Например (рис. 4Б), при гипотезе k = 8 предковых компонент, одна из них составляет в среднем 91% новгородского генофонда и потому, следуя этой традиции, названа «Новгородской», хотя в ярославском генофонде она чрезвычайно велика (составляет 90%) и потому ей можно дать и иное, столь же условное названия – «Новгородско-ярославская». Но доля вклада АС не обязательно должна достигать столь больших значений. Например, «Саамская» АС (рис. 4Б) при k = 8 составляет всего четверть их генофонда (24%), но поскольку в других генофондах она практически не встречается (0–1%), она справедливо носит условное название «Саамской». Названия чисто условные, и поэтому их можно менять так, чтобы при описании они точнее отражали суть результата. Например, при k = 8 одна из АС составляет 88% генофонда геномов води и ижоры, а также 79% генофонда карел. Поэтому для удобства описания ей дано условное название «Западно-финская» (рис. 4Б).

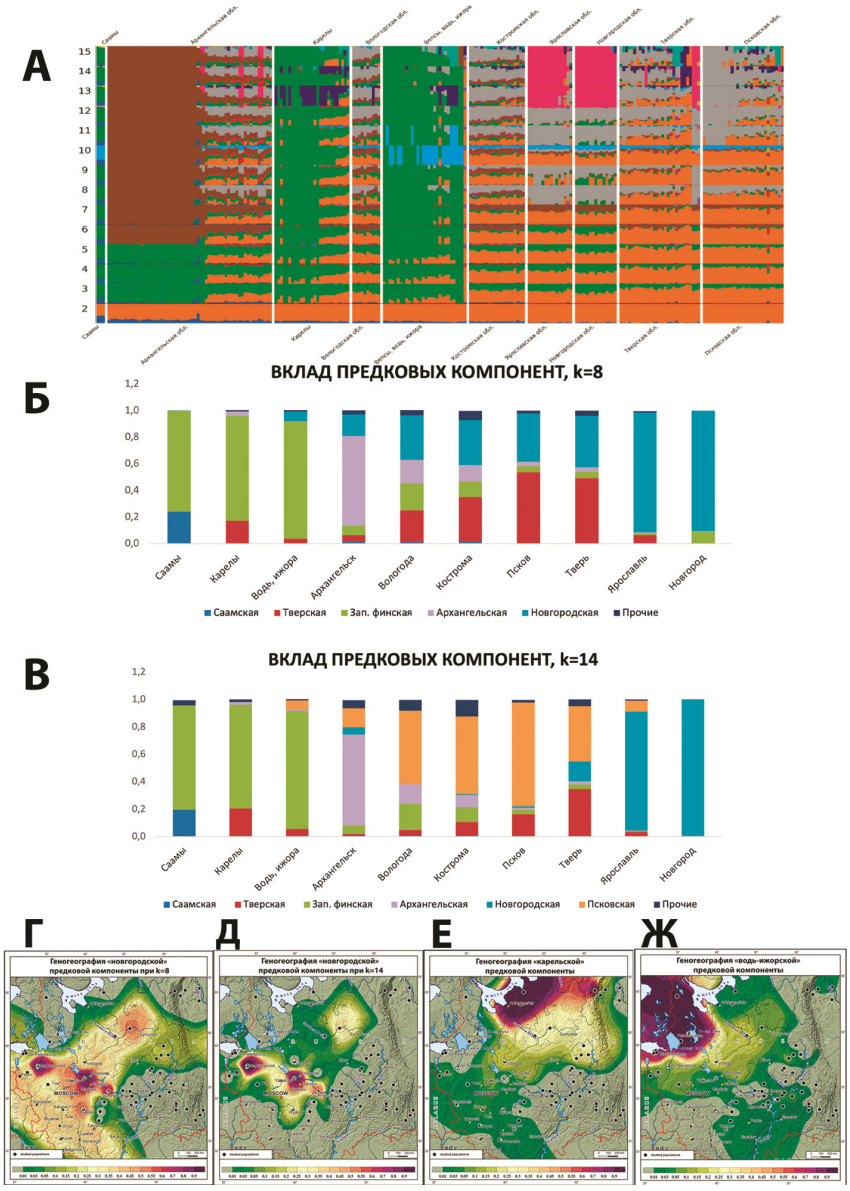

Рисунок 4. Разные способы представления результатов метода предковых компонент ADMIXTURE (приведен фрагмент для северо-восточной Европы из результатов анализа обширного спектра популяций Северной Евразии)

Figure 4. Various representations of the results of the ADMIXTURE ancestral components method (a fragment of results for northeastern Europe from an analysis of a wide range of Northern Eurasian populations is shown)

Примечания. А – стандартный график при переменном значении k от 2 до 15 предковых компонент (14 строк графика); индивидуальные геномы размещены по одной и той же вертикальной линии, независимо от значений k; геномы каждой популяции размещены рядом в столбец и отделены от других популяций белой вертикалью; Б – представление в привычном виде столбчатых диаграмм вклада предковых компонент при k= 8; В – представление в привычном виде столбчатых диаграмм вклада предковых компонент при k = 14; Г, Д, Е, Ж – картографическое представление при разных значениях k распространения предковых компонент: «Новгородской» (Г, Д), «Карельской» (Е), «Водь-ижорской» (Ж)

Notes. A – Standard graph with variable k from 2 to 15 ancestral components (14 rows in the graph); individual genomes are placed along the same vertical line, regardless of k values; genomes of each population are placed next to each other in a column, separated from other populations by a white vertical line. Б – Representation in the usual form of bar charts showing the contribution of ancestral components at k = 8. В – Representation in the usual form of bar charts showing the contribution of ancestral components at k = 14. Г, Д, Е, Ж – Cartographic representation of the distribution of ancestral components at different k values: «Novgorodian» (Г, Д), «Karelian» (Е), and «Votian-Ingrian» (Ж)

Независимость моделей. При k = 2 для каждого проанализированного генома выявляется вклад всего двух предковых компонент. При увеличении k программа производит расчет заново для тех же геномов и рассчитывает все большее число более дробных предковых компонент: три предковых компоненты при k = 3, четыре компоненты при k = 4, а при k = 20 предполагается участие двадцати предковых компонент. Таким образом при большом числе k можно выделить все более «частные» компоненты, характерные для малых групп или даже отдельных популяций. При каждом k программа проводит новый расчет по совокупности всех включенных в анализ геномов, и каждый новый вариант расчета никак не зависит от всех других вариантов. Поэтому каждая модель (гипотеза) строится независимо от предыдущих моделей, и для каждой модели программа выдает количественную оценку вклада каждой из предковых компонент в каждый индивидуальный геном (рис. 4А).

Формы представления могут быть разными. Наиболее распространенная – стандартный график ADMIXTURE (рис. 4А), где каждая вертикальная линия – индивидуальный геном, каждая строка – модель при k предковых компонент, а разными оттенками цвета обозначены вклады предковых компонент (у каждой свой цвет). Точные доли (%) вклада каждой предковой компоненты в каждый индивидуальный геном для каждого значения k представлены в серии таблиц, сопровождающих график (для каждой модели при данном k – своя таблица).

Конечно, чтение такого графика требует некоторого опыта. Поэтому можно представлять полученные результаты в более привычном виде. Удобная форма представления – столбчатые диаграммы (рис. 4Б, В) с вкладом каждой предковой компоненты в генофонд каждой популяции («популяционный вклад» АС – это среднее значение АС по всем индивидуальным геномам членов данной популяции). При этом, когда вклад каких-то АС очень мал или они мало информативны для данной модели, мы можем для удобства сократить диаграмму, суммировав вклады некоторых предковых компонент в «прочие» (на рис. 4Б все «прочие» составили всего 2%, на рис. 4В – всего 4%).

Поскольку популяции различаются по размеру ареала, мы можем увеличить информативность результатов, если учтем положение популяций в географическом пространстве. Для этого вклад каждой предковой компоненты при каждом k можно картографировать (рис. 4Г-Ж). Например, для реконструкции генетической истории Новгородчины мы построили массив, включающий 119 карт (карты для каждой АС при каждом значении k от 2 до 15), но для публикации [Балановская с соавт., 2021] выбрали из них 4 карты для трех «предковых компонент»: две карты компоненты, условно названной «Новгородской» – при k = 8 (рис. 4Г) и при k = 14 (рис. 4Д); карту компоненты, условно названной «Карельской» (рис. 4Е); карту предковой компоненты, условно названной «Водь-ижорской» (рис. 4Ж). Пример интерпретации результатов, полученных методом ADMIXTURE, в плане этнической и генетической истории населения северо-восточной Европы в связи с новгородской экспансией, дан в этой же публикации, рассчитанной на этнографов [Балановская с соавт., 2021].

Предковые компоненты ADMIXTURE как источник информации для антропологической классификации

Предковые компоненты ADMIXTURE могут стать для этнической антропологии новым вспомогательным инструментом при уточнении расово-антропологической классификации народов. Важность этого инструмента в том, что он не зависит от набора тех или иных антропологических признаков. Каждая система антропологических признаков (морфологических, дерматоглифических, одонтологических) контролируется небольшой совокупностью генов, к тому же частично скоррелированных, и поэтому каждая система по-своему классифицирует популяции. Важное преимущество предковых компонент ADMIXTURE в том, что они опираются на очень большую совокупность независимых друг от друга ДНК-маркеров (от сотен тысяч до миллионов). Еще одно крайне важное преимущество – прямая количественная оценка вклада того или иного генетического комплекса в каждую популяцию при разных гипотезах ее таксономического положения.

Пример использования такой классификации для количественной оценки вклада «монголоидности» и «европеоидности» в генофонды популяций (табл. 1) был опубликован недавно в данном журнале [Козлов с соавт., 2023]. Это позволяет на данном примере показать, как метод ADMIXTURE может помочь детализировать антропологическую таксономию популяций, причем с количественной оценкой вклада генетических комплексов разного уровня иерархии.

На рис. 5А представлен стандартный график ADMIXTURE для тех же исходных геномов,

что и в таблице 1: информация нижней строки стандартного графика ADMIXTURE (при k = 2) представлена в таблице 1 в виде линейчатой диаграммы. Ее можно ассоциировать с двумя стволами антропологической классификации, и дать количественную оценку вклада «монголоидности-европеоидности» в генофонды популяций. Справа от стандартного графика ADMIXTURE рисунка 5 та же самая информация – вклад этих же двух предковых компонент – представлена для 7 этнических групп в виде привычных столбчатых диаграмм с указанием количественного вклада каждой компоненты в генофонд каждого этноса.

Однако мы получим больше полезной информации, если не остановимся на выделении только двух предковых компонент, а будем постепенно выделять все более дробные таксономические единицы. Для этого на стандартном графике ADMIXTURE (рис. 5) приведены четыре модели с k = 2, k = 3, k = 4, k = 5, а справа от каждой строки стандартного графика ADMIXTURE представлен вклад предковых компонент для одних и тех же 7 этносов в виде «столбчатых» диаграмм с указанием количественного вклада каждой предковой компоненты (АС) в генофонд каждого этноса. Что же мы видим?

Таблица 1. Величины вкладов двух предковых компонент (k = 2) в 18 этнических группах Северной Евразии
Table 1. The contributions of two ancestral components (k = 2) in 18 ethnic groups of Northern Eurasia

Этнос	Вклад «Европеидной» предковой компоненты, %	Вклад «Монголоидной» предковой компоненты, %
Русские	98	2
Ягнобцы	86	14
Народы Памира	81	19
Таджики	75	25
Башкиры	63	37
Туркмены	58	42
Узбеки	58	42
Татары сибирские	51	49
Казахи	37	63
Алтайцы северные	36	64
Шорцы	31	69
Киргизы	25	75
Хакасы	25	75
Алтайцы южные	20	80
Калмыки	12	88
Тофалары	9	91
Монголы не халха	8	92
Тувинцы	6	94
Буряты	6	94
Тоджинцы	5	95
Монголы халха	4	96
Якуты	3	97

При k = 3 появляется третья предковая компонента (окрашенная на графике зеленым цветом). Важно, что новая АС включает в себя часть и «Европеиодной», и «Монголоидной» компонент, которые выделились при k = 2: выведенные доли (%) вклада каждой компоненты на «столбиках» (рис. 5, правая часть) помогают оценить степень такого уточнения таксономии. Третья АС наиболее ярко представлена у народов Средней Азии: она составила 97% генофонда ягнобцев Таджикистана, 95% – народов Памира, 85% – таджиков, 64% – туркмен, 59% – узбеков (на стандартном графике рис. 6 для удобства чтения приведены более крупные объединения популяций, но таблицы для каждого значения k дают возможность оценить вклад АС отдельно для народов Таджикистана или же разных популяций алтайцев или хакасов). Важно, что при выделении новых предковых компонент (k = 4, k = 5) вклад этой АС (обозначенной зеленым цветом) в генофонды этих народов практически не меняется: мы видим стабильность вклада «зеленой» АС и на стандартном графике (от ягнобцев до туркмен и узбеков); и в столбчатой диаграмме, где для объединенной популяции с условным названием «Таджики» вклад «зеленой» АС колеблется в узких пределах от 85% до 82% при 3 ≤ k ≤ 5. Такая устойчивость вклада «зеленой» АС позволяет говорить, что это действительная важная предковая компонента для среднеазиатских генофондов. Но в Южной Сибири мы видим иную картину. При k = 3 «Среднеазиатская» АС (зеленый тон) составляет у шорцев, хакасов и алтайцев около трети генофонда. Но при переходе к более дробным таксономическим моделям (при k = 5) вклад «Среднеазиатской» АС в генофонды народов Южной Сибири почти полностью исчезает, поскольку заменяется новой собственной «Алтайской» предковой компонентой (сиреневый тон), которая при таксономической модели пяти предковых компонент составляет половину генофондов народов Южной Сибири (рис. 5).

При k =4 новая (четвертая) предковая компонента (красный тон) может быть названа «Южносибирской». Она вносит наибольший вклад в генофонды тофаларов (92%) и тувинцев (49%), но с заметной частотой присутствует у многих народов Сибири: шорцев (50%), хакасов (31%, с максимумом 38% у сагайцев), алтайцев (30%, с максимумом 40% у челканцев), якутов (18%), сибирских татар (15%, с максимумом 25% у заболотных татар).

При k = 5 «Южносибирская» компонента (красный тон) сохраняет свое значение у тофаларов, тувинцев и якутов. Но новая более дробная «Алтайская» АС (сиреневый тон) составляет значительную часть генофондов шорцев (95%), алтайцев (45% с максимумом 83% у челканцев) и хакасов (39%, с максимумом 59% у сагайцев и минимумом 16% у качинцев).

Так работает метод предковых компонент ADMIXTURE. Позволяя выделять все большее число предковых компонент и все более дробные таксоны, программа количественно оценивает их вклад в индивидуальные геномы, что дает возможность последовательного количественного анализа. Сравнивая динамику этих количественных оценок (полученных по очень большей совокупности генов) для одной и той же популяции, или для разных популяций при разных моделях (гипотезах их происхождения), мы получаем инструмент для проверки различных версий антропологических классификаций и формулировки новых гипотез. Если мы сможем сопоставить выявленные предковые компоненты с теми или иными группами антропологической классификации, то получим уникальную возможность количественно оценивать вклад расово-антропологических различий в заболеваемость, спортивные достижения, ростовые процессы и многие другие параметры.

Каким особенностям метода ADMIXTURE надо уделять внимание? Отмечу три наиболее важных.

Во-первых, охват популяций. Чем шире представлены популяции региона и чем разнообразнее их генофонды, тем более корректные результаты мы получим. Поэтому наиболее перспективно включение в анализ ADMIXTURE максимального спектра самых разных популяций максимально обширной территории. Предел нашим желаниям ставит подробность данных об аутосомных геномах и мощность сервера (сейчас даже на мощных серверах расчет при больших значениях k может занимать несколько суток). Но при решении региональных задач (как в случаях работы по реконструкции истории Новгородчины, рис. 4) мы можем из всего спектра популяций, использованных при расчете ADMIXTURE, рассматривать только популяции конкретного региона.

Во-вторых, определенное влияние оказывает число изученных геномов в популяции. Если для какой-то популяции число изученных геномов резко превышает число геномов в других популяциях, то наиболее полно представленная популяция может «перетягивать одеяло на себя», повышая значимость своих предковых компонент.

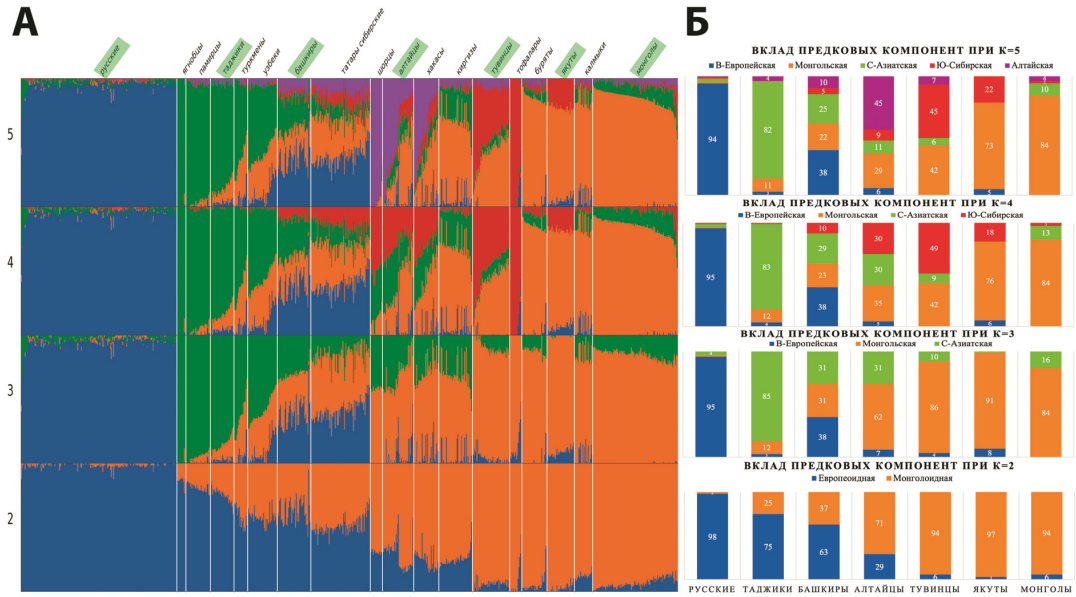

Рисунок 5. Представление результатов метода предковых компонент ADMIXTURE [Козлов с соавт., 2023] для k = 2, 3, 4, 5 в виде стандартного графика (5А, слева) и столбчатых диаграмм (5Б, справа)

Figure 5. Representation of the results of the ADMIXTURE ancestral components method [Kozlov et al., 2023] for k = 2, 3, 4, 5 in the form of a standard graph (5A, left) and bar charts (5Б, right)

В-третьих, моделирование при каждом значении k происходит настолько независимо, что, многократно повторяя расчет ADMIXTURE при одном и том же значении k, эти модели могут в некоторых деталях различаться. Например, могут появляться «случайные» предковые компоненты, которые не подтверждаются другими моделями и в других повторениях расчета при том же значении k. Этот случай продемонстрирован на рисунке 4 для геномов карел: при k=13 появляется предковая компонента (окрашенная темно-синим тоном), которая не подтверждается другими моделями. Поэтому каждая модель при том же значении k строится несколько раз, что позволяет статистически оценить ее надежность с помощью двух подходов – оценки ошибки кросс-валидации (аналогично анализу главных компонент и факторному анализу) и расчета логарифма функции правдоподобия (более высокие значения логарифма функции правдоподобия указывают на более вероятную модель). Хотя в целом рекомендуется для каждого значения k повторить расчет 5–10 раз [Alexander et al., 2015, 2020], однако в работах по популяционной генетике при наличии мощных серверов для каждого k строится даже по 100 моделей [Rasmussen et al., 2010; Yunusbaev et al., 2015; Tambets et al., 2018], что обеспечивает надежность анализа.

Заключение

Конечно же, метод предковых компонент ADMIXTURE – лишь одно из направлений, где союз антропологии и генетики перспективен. У каждой из наук есть свой собственный центральный «ареал», но есть и пограничные области, где взаимодействие с другой наукой дает полезные плоды. Однако пока все больше генетики пасутся на плохо охраняемых территориях антропологии. Конечно, антропология может усилить охрану и заключить свои нивы в крепостные стены, но при этом есть серьезный риск превращения в резервацию. Намного полезнее ответить смелыми вылазками на приветливую территорию генетики: такие рейды могут принести антропологии богатые трофеи. А совместное изучение приграничных территорий, несомненно, даст синэргетический эффект.

Благодарности

Автор благодарен всем коллегам – генетикам и антропологам, которые помогли получить данные и внесли уточнения в текст: С.М. Кошелю, Н.Н. Гончаровой, И.О. Горину, И.В. Евсюкову, А.Ю. Потаниной, А.А. Агджоян. Исследование выполнено в рамках государственного задания Минобрнауки России для ФГБНУ «МГНЦ».

Обследование проведено на добровольной основе с письменным информированным согласием, одобренным Этическим комитетом ФГБНУ «МГНЦ» (Заключение от 29.06.2020 г).

Библиография

Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности: Справочное издание. М.: Финансы и статистика. 1989. 606 с. ISBN: 527900054X.

Балановская Е.В., Потанина А.Ю., Кошель С.М., Адамов Д.С., Борисова А.Л., с соавт. Технология оценки частот ДНК-маркеров в многонациональных административных единицах по данным о коренном народонаселении в связи с заболеваемостью (на примере кардиоваскулярных заболеваний) // Кардиоваскулярная терапия и профилактика, 2024. (В печати).

Балановская Е.В., Черневский Д.К., Балановский О.П. Своеобразие Новгородского генофонда в контексте народонаселения европейской части России // Вестник Новгородского государственного университета. Сер.: Медицинские науки, 2021. № 3. С. 51–57. DOI: 10.34680/2076-8052.2021.3(124).51-57.

Козлов А.И., Пылев В.Ю., Вершубская Г.Г., Балановская Е.В. Клинальная изменчивость генетических детерминант трегалазной недостаточности в популяциях Южной Сибири, Казахстана, Центральной Азии и Монголии // Вестник Московского университета. Серия 23. Антропология, 2023. № 3. С. 63–71. DOI: 10.32521/2074-8132.2023.3.063-071.

Курбатова О.Л., Грачева А.С., Победоносцева Е.Ю., Удина И.Г. Генетико-демографические параметры населения г. Москвы. Миграционные процессы // Генетика, 2021. Вып. 56. № 12. С. 1438–1449. DOI: 10.31857/S0016675821120080

Курбатова О.Л., Победоносцева Е.Ю., Веремейчик В.М., Прудникова А.С., Атраментова Л.А., с соавт. Особенности генетико-демографических процессов в населении трех мегаполисов в связи с проблемой создания генетических баз данных // Генетика, 2013. Вып. 49. № 4. С. 513. DOI: 10.7868/S0016675813040085

Удина И. Г., Грачева А.С., Курбатова О.Л. Частоты гаплогрупп Y-хромосомы и процессы миграции в трех поколениях жителей Москвы // Генетика, 2022. Вып. 58. № 11. С. 1325–1333. DOI: 10.31857/S001667582110121

Информация об авторе

Балановская Елена Владимировна, проф., д.б.н.;
ORCID ID: 0000-0002-3882-8300; balanovska@mail.ru.

Поступила в редакцию 07.10.2024,
принята к публикации 20.10.2024

Balanovska E.V.

Research Centre of Medical Genetics, Laboratory of human population genetics, Moskvorechye, 1, Moscow, 115478, Russia

THE ALLIANCE OF ANTHROPOLOGY AND POPULATIONS GENETICS

Introduction. *Russian population genetics arose in the depths of anthropology. Over time, the rapid development of genetic technologies created a tension between these two fields of science. In hope to strengthen the long-standing alliance between anthropology and genetics, this work attempts to describe some aspects of genetic characteristics that genetics deals with, discuss the problem of sample representativeness for so diverse genetic features, and explain how methods harnessed by both sciences are used in genetics. The main focus of the article is on ADMIXTURE, a method of ancestry estimation which makes use of paleogenetic data and is well known to paleoanthropologists.*

Results and discussion. *Our study shows how this method can benefit the ethnic anthropology of modern populations. We provide examples of PCA and ancestral component analysis for different regions (the Russian North, the Far East, Northern Eurasia) and for different tasks. ADMIXTURE can quantitatively estimate the contributions of racial and anthropologic components on different hierarchical levels; its estimates are based on huge arrays of independent genetic markers.*

Conclusion. *Only a small part of the extensive research field that both anthropology and genetics deal with is discussed in this paper. But if our attempt boosts collaboration between geneticists and anthropologists, the mission of this paper can be considered accomplished.*

Keywords: ethnic anthropology; human population genetics; ancestor component method ADMIXTURE; principal component method

DOI: 10.55959/MSU2074-8132-24-4-4

References

Aivazyan S. A., Bukhshtaber V. M., Enyukov I. S., Meshalkin L. D. *Prikladnaya statistika. Klassifikatsiya i snizhenie razmernosti: Spravochnoe izdanie* [Applied statistics. Classification and dimensionality reduction: A reference edition]. M.: Finansy i statistika, 1989. p. 606. ISBN: 527900054X. (In Russ.).

Balanovskaya E.V., Potanina A.Yu., Koshel' S.M., Adamov D.S., Borisova A.L., et al. Tekhnologiya otsenki chastot DNK-markerov v mnogonatsional'nykh administrativnykh edinitsakh po dannym o korennom narodonaselenii v svyazi s zabolevaemost'yu (na primere kardiovaskulyarnykh zabolevanii) [Technology for estimating DNA marker frequencies in multinational administrative units from indigenous population data in relation to morbidity (cardiovascular diseases as an example)]. *Kardiovaskulyarnaya terapiya i profilaktika* [Cardiovascular Therapy and Prevention], 2024. (In print). (In Russ.).

Balanovskaya E.V., Chernevskii D.K., Balanovskii O.P. Svoeobrazie Novgorodskogo genofonda v kontekste narodonaseleniya evropeiskoi chasti Rossii [The peculiarity of the Novgorod gene pool in the context of the population of the European part of Russia]. *Vestnik Novgorodskogo gosudarstvennogo universiteta. Ser.: Meditsinskie nauki* [Bulletin of Novgorod State University. Ser.: Medical Sciences], 2021, 3, pp. 51–57. (In Russ.). DOI: 10.34680/2076-8052.2021.3(124).51-57

Kozlov A.I., Pylev V.Yu., Vershubskaya G.G., Balanovskaya E.V. Clinal variability of genetic determinants of trehalase deficiency in populations of South Siberia, Kazakhstan, Central Asia and Mongolia. *Moscow University Anthropology Bulletin* [Vestnik Moskovskogo Universiteta. Seriya XXIII. Antropologia], 2023, 3, pp. 63–71. (In Russ.). DOI: 10.32521/2074-8132.2023.3.063-071

Kurbatova O.L., Gracheva A.S., Pobedonostseva E.Y., Udina I.G. Genetiko-demograficheskie parametry naseleniya g. Moskvy. Migratsionnye protsessy [Genetic and demographic parameters of the population of Moscow. Migration processes.]. *Genetika* [Russian J Genetics], 2021, 56 (12), pp. 1438–1449. (In Russ.). DOI: 10.31857/S0016675821120080

Kurbatova O.L., Pobedonostseva E.Yu., Veremeichik V.M., Prudnikova A.S., Atramentova L.A., et al. Osobennosti genetiko-demograficheskikh protsessov v naselenii trekh megapolisov v svyazi s problemoi sozdaniya geneticheskikh baz dannykh [Peculiarities of genetic and demographic processes in the population of three megacities in connection with the problem of creating genetic databases]. *Genetika* [Russian J Genetics], 2013, 49 (4), pp. 513. (In Russ.). DOI:10.7868/S0016675813040085

Udina I.G., Gracheva A.S., Kurbatova O.L. Chastoty gaplogrupp Y-khromosomy i protsessy migratsii v trekh pokoleniyakh zhitelei Moskvy [Frequencies of Y-chromosome haplogroups and migration processes in three generations of Moscow residents]. *Genetika* [Russian J Genetics], 2022, 58 (11), pp. 1325–1333. (In Russ.). DOI: 10.31857/S001667582110121

Alexander D.H., Shringarpure S.S., John Novembre J., Lange K. *Admixture 1.3 Software Manual*. 2015. https://vcru.wisc.edu/simonlab/bioinformatics/programs/admixture/admixture-manual.pdf (Accessed 20.09.2024).

Alexander D.H., Shringarpure S.S., John Novembre J., Lange K. *Admixture 1.3 Software Manual*. 2020. Available at: http://dalexander.github.io/admixture/admixture-manual.pdf (Accessed 20.09.2024).

Dormann C.F., Elith J., Bacher S., Buchmann C., Carl G., et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 2013, 36 (1), pp. 27–46. DOI: 10.1111/j.1600-0587.2012.07348.

Mirkes E.M., Allohibi J., Gorban A. Fractional Norms and Quasinorms Do Not Help to Overcome the Curse of Dimensionality. *Entropy*, 2020, 22 (10), pp.1105. DOI: 10.3390/e22101105.

Nei M., Roychoudhury A.K. Sampling variances of heterozygosity and genetic distance. *Genetics*, 1974, 76 (2), pp. 379–390. DOI: 10.1093/genetics/76.2.379.

Rasmussen M., Li Y., Lindgreen S., Pedersen J., Albrechtsen A., et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 2010, 463, pp. 757–762. DOI: 10.1038/nature08835.

Tambets K., Yunusbayev B., Hudjashov G., Ilumäe A.M., Rootsi S., et al. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biol.*, 2018, 19 (1), pp. 139. DOI: 10.1186/s13059-018-1522-1.

Yunusbayev B., Metspalu M., Metspalu E., Valeev A., Litvinov S., et al. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *PLoS Genet.*, 2015, 11 (4), pp. e1005068. DOI: 10.1371/journal.pgen.1005068.

Information about the author

Balanovska Elena V., professor, DSc. in Biology; ORCID ID: 0000-0002-3882-8300; balanovska@mail.ru.