ГЕНЕТИКА, 2016, том 52, № 7, с. 810–830

УДК 575.17

ХРОМОСОМА-ЛЕТОПИСЕЦ: ДАТИРОВКИ ГЕНЕТИКИ, СОБЫТИЯ ИСТОРИИ, СОБЛАЗН ДНК-ГЕНЕАЛОГИИ

© 2016 г. О. П. Балановский[1, 2], В. В. Запорожченко[2, 1]

[1]*Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва 119991*

[2]*Медико-генетический научный центр, Москва 115478*

e-mail: balanovsky@inbox.ru

Поступила в редакцию 16.02.2016 г.

Нерекомбинирующие части генома – Y-хромосома и митохондриальная ДНК – широко используются для изучения генофондов популяций человека и реконструкции их истории. Эти системы дают возможность генетического датирования кластеров возникающих гаплотипов. Основным методом расчета возраста является ρ-статистика – среднее число мутаций от гаплотипа-основателя до всех современных гаплотипов. Умножая это число на скорость мутирования, исследователь получает оценку возраста кластера. Для STR-гаплотипов Y-хромосомы используется также второй метод расчета – ASD, основанный на среднеквадратичном различии в числе повторов. Кроме методов расчета, все большее значение приобретают методы байесовского моделирования. Вычислительно они значительно сложнее, но позволяют получить апостериорное распределение интересующих исследователя величин, в наибольшей степени согласующееся с экспериментальными данными. И для методов расчета, и для методов моделирования необходимо знать скорость мутирования. Она определяется либо при анализе родословных, либо путем калибровок на популяциях с известным временем формирования. Для STR-гаплотипов Y-хромосомы эти два подхода дали трехкратно различающиеся скорости. Это противоречие удалось снять только недавно благодаря использованию данных полного секвенирования Y-хромосомы: "полногеномные" скорости однонуклеотидных мутаций совпадают при использовании обоих подходов и очерчивают границы применимости разных STR-скоростей. Еще более важной, чем проблема скоростей, является проблема соотнесения реконструированной истории гаплогруппы (кластера гаплотипов) с историей популяции. Хотя необходимость различения "истории линий" и "истории популяций" была отмечена еще на заре филогеографических исследований, имеется ряд приемов и условий реконструкции популяционной истории с помощью генетических датировок. Известно, что существуют лишь некоторые демографические ситуации, в которых события истории популяции (народа) оставляют четкие следы в истории гаплогрупп. Прямое же отождествление истории народа с историей встреченных у него гаплогрупп неправомочно и избегается в популяционно-генетических работах, хотя из-за своей простоты и притягательности является постоянным соблазном для исследователей. Пример ДНК-генеалогии – любительской области, вышедшей за пределы даже гражданской науки и последовательно применяющей принцип приравнивания гаплогруппы, рода и популяции, приводящий к абсурдным результатам (например, Евразии как прародины человечества), – может послужить предостережением от упрощенного подхода к интерпретации генетических датировок.

Ключевые слова: датировка, гаплотип, кластер, Y-хромосома, митохондриальная ДНК, полногеномное секвенирование, мутация.

DOI: 10.7868/S0016675816070043

Анализ гаплоидных хромосом с однородительским типом наследования – Y-хромосомы и митохондриальной ДНК – за последние 20 лет стал одним из основных инструментов изучения популяций человека. Даже появление полногеномных данных и методов их анализа значительно дополнило, но не заменило собой использование гаплоидных систем, публикации по которым до сих пор составляют основную часть всех появляющихся статей, посвященных изучению генофондов популяций человека (библиометрический анализ проведен в [1]).

Одним из важных преимуществ "однородительских" (нерекомбинирующих) генетических систем перед аутосомными (рекомбинирующими) генетическими маркерами является возможность датировок по мутациям. Отсутствие рекомбинации (вследствие гаплоидности этих систем) приводит к тому, что возникающие мутации наслаиваются друг на друга практически бесконечно, и формирующийся гаплотип не разбивается

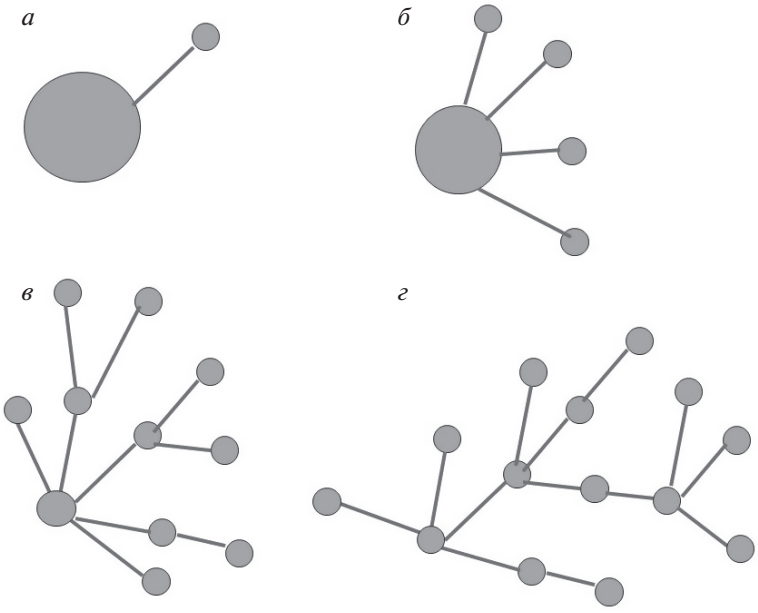


Рис. 1. Нарастание дерева гаплотипов, лежащее в основе метода молекулярных часов. Показан постепенный рост дерева гаплотипов: возникновение новых гаплотипов в результате мутаций. *а* – один исходный гаплотип (встречен у многих образцов, поэтому большой кружок) и один производный (только у одного образца, маленький кружок); *б* – возник уже целый ряд производных гаплотипов; *в* – из "дочерних" гаплотипов начали возникать "внучатые", а частота исходного гаплотипа уже почти не отличается от любого из производных; *г* – дерево приобрело сложную структуру, и уже трудно однозначно решить, какой именно гаплотип был исходным.

рекомбинацией. Это позволяет реконструировать филогенетическое дерево родства всех обнаруженных гаплотипов и, зная скорость мутирования, рассчитать возраст появления каждой ветви дерева – каждого кластера гаплотипов. А поскольку разные кластеры зачастую приурочены к разным популяциям, то датировки времени возникновения кластеров (филогенетика) в сочетании с данными по географическому распространению кластеров (филогеография) позволяют, при определенных условиях, реконструировать историю популяций.

К настоящему времени разработан ряд методов датирования кластеров гаплотипов, исследованы границы применимости этих методов, рассмотрены вопросы скоростей мутирования, накоплен опыт применения генетических датировок к реконструкции истории популяций. Сами методы, по крайней мере в своих стандартных вариантах, широко используются во множестве отечественных и зарубежных исследований. Но попытки систематического изложения основ этих методов, особенностей их применения, существующих проблем, в том числе при интерпретации результатов, в русскоязычной литературе практически отсутствуют да и в англоязычной они немногочисленны. Целью настоящего обзора является хотя бы частичное заполнение этого пробела.

Изложение строится главным образом на примере Y-хромосомы, но большинство методов применимы также к митохондриальной ДНК, как упоминается по ходу изложения. В принципе, те же подходы и модели могут использоваться и при анализе коротких фрагментов аутосомного генома с практически полным неравновесием по сцеплению.

Объектом датировки в исследованиях Y-хромосомы и мтДНК является монофилетичный кластер гаплотипов – группа гаплотипов, происходящих от одного исходного. Предположим, что мы знаем, каким был исходный гаплотип. Тогда легко понять, как возник монофилетичный кластер: из исходного гаплотипа в результате мутаций возникали все новые "дочерние" гаплотипы (отличающиеся на один мутационный шаг), а от них, в свою очередь, возникали "внучатые" гаплотипы (отличающиеся на два шага) и так далее. Чем больше пройдет времени, тем больше новых гаплотипов возникнет, и тем на большее число мутационных шагов они будут отличаться от исходного, т.е. тем более разнообразные гаплотипы мы обнаружим. Верно и обратное: чем большее разнообразие гаплотипов мы видим в пределах кластера, тем дольше кластер существует, тем он старше. Эта логика проиллюстрирована на рис. 1. По сути, это обычный принцип молекулярных часов, где время отсчитывается происходящими мутациями. Если скорость мутаций постоянна, то часы будут работать. В этих общих чертах задача расчета возраста проста (рис. 2,*а*), а оговорки, ко-

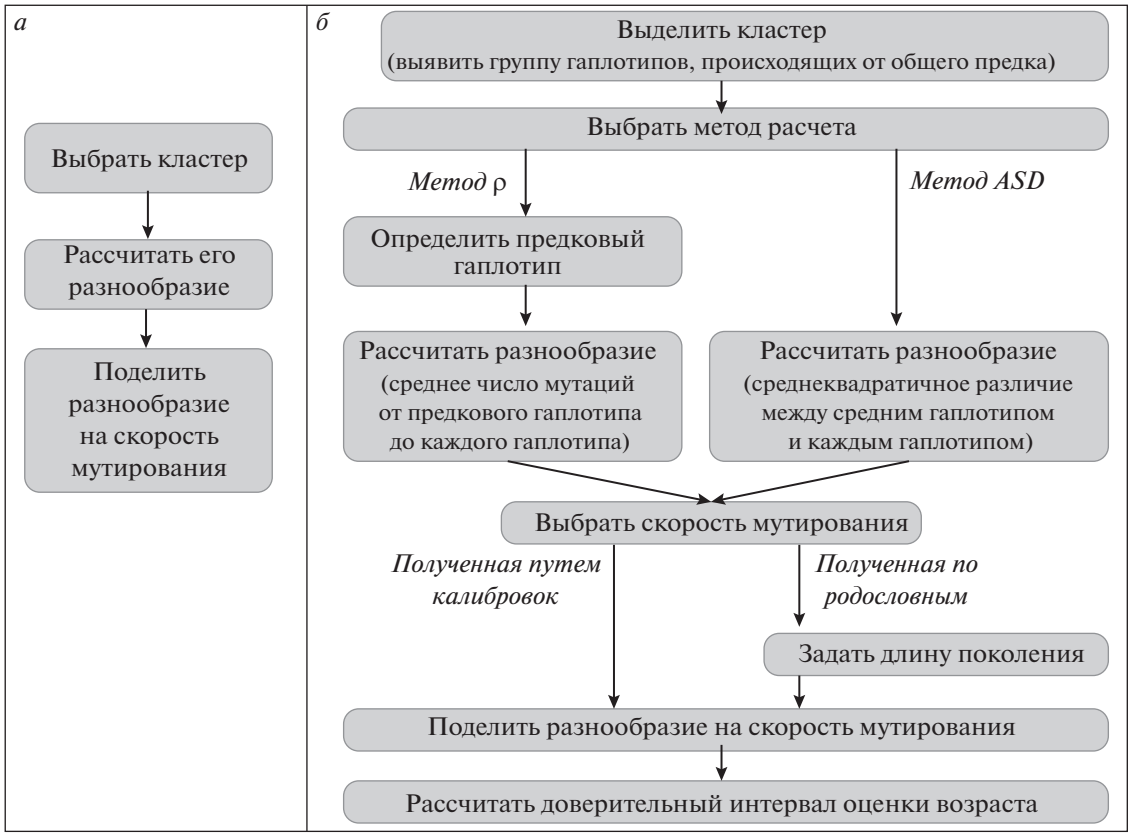


Рис. 2. Упрощенная (слева) и уточненная (справа) схема генетических датировок.

торые мы делаем, говоря “предположим” и “если”, разобраны в разделе “Допущения...”. Но все осложняется, когда приходится решать, как именно подсчитывать разнообразие гаплотипов и как именно переводить его в годы (рис. 2,*б*).

Все предложенные методы датировок кластеров гаплотипов можно поделить на методы расчета и методы моделирования. Они различаются с самого первого этапа: первые основаны на реконструкции наиболее *экономного* (парсимонного) филогенетического дерева и *прямом подсчете* числа накопившихся в кластере мутаций; вторые – на реконструкции наиболее *правдоподобного* дерева и *моделировании* распределения числа мутаций. Рассмотрим последовательно методы расчета, методы моделирования и применение полученных датировок к реконструкции истории популяции.

РАСЧЕТ ВОЗРАСТА КЛАСТЕРА ГАПЛОТИПОВ

Выявление кластеров гаплотипов

Прежде всего необходимо выделить, какие из множества экспериментально обнаруженных гаплотипов формируют монофилетичные кластеры (т.е. происходящие от единого предкового гаплотипа). Родство рассматриваемых гаплотипов принято отображать в виде филогенетического дерева. А когда однозначное дерево реконструировать не удается и остаются равновероятные возможности, что превращение одного гаплотипа в другой шло тем или иным путем, это отображается как слияние нескольких ветвей, т.е. дерево превращается в сеть. Методология построения и анализа филогенетических сетей была хорошо разработана в популяционной генетике в 90-е годы прошлого века [2] и написаны соответствующие программы, самой популярной из которых является Network 4.1.1.2 (www.fluxus-engineering.com).

На большинстве филогенетических сетей обнаруживается ряд более или менее компактных кластеров гаплотипов. Выявление и датировка их происхождения стали важным аспектом многих исследований Y-хромосомы [3–5 и мн. др.]. При этом нет алгоритма, автоматически выявляющего кластеры гаплотипов на сети, поэтому определение числа и границ кластеров оставляется на усмотрение исследователя, а значит от его произвола зависят датировки возраста кластера. Чтобы минимизировать эту произвольность, в практике наших исследований мы руководствуемся следующими правилами [6] (похожие правила применяются и другими исследователями):

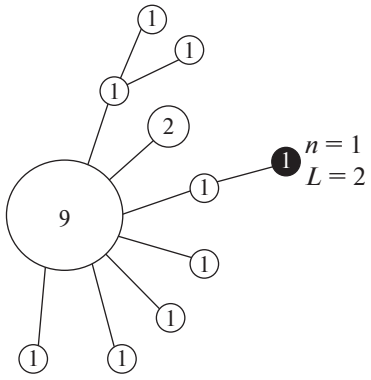
$\rho = (\Sigma nL)/N$

ρ – среднее число мутаций от исходного гаплотипа до всех встреченных гаплотипов

n – число образцов с данным гаплотипом

N – общее число образцов

L – длина ветки, т.е. число мутационных шагов от данного гаплотипа до исходного



$\rho = (6 \times 1 + 1 \times 2 + 3 \times 2 + 9 \times 0)/20 = 0.7$

Рис. 3. Расчет возраста кластера гаплотипов с помощью показателя ρ. Показано дерево родства гаплотипов: один исходный (большой кружок, встречен у 9 образцов) и 10 производных гаплотипов, из них девять встречены только у одного образца каждый, а десятый гаплотип встречен у двух образцов. Для одного из гаплотипов (выделен черным), подписаны показатели n (встречен у одного образца) и L (отстоит от исходного гаплотипа на два мутационных шага). Проведен расчет показателя ρ – суммировано число образцов в гаплотипах, умноженное на число шагов от данного гаплотипа до исходного: шесть гаплотипов отстоят на один шаг и представлены одним образцом каждый (6×1), один гаплотип отстоит на один шаг и представлен двумя образцами (1×2), три гаплотипа представлены одним образцом и отстоят на два шага (3×2), исходный гаплотип представлен 9 образцами и отстоит сам от себя на ноль шагов. Величина ρ в данном примере составляет 0.7.

1. Поскольку большинство филогенетических сетей имеют четко выделяющуюся центральную зону (вероятный корень, от которого происходит гаплогруппа), считаем кластерами только те группы гаплотипов, каждый из которых связан с этим корнем через один и тот же узловой гаплотип. Иными словами, к кластерам относим только ветви, строго монофилетичные согласно реконструированной сети.

2. Именно этот узловой гаплотип рассматривается в качестве предкового гаплотипа-основателя (founder) для своего кластера, даже если какой-то из производных гаплотипов встречен у большего числа образцов.

3. Рассматриваются только кластеры, содержащие 10 или более образцов (не гаплотипов), чтобы избежать ошибок в расчете возраста из-за малых объемов выборок.

Метод датировки с помощью показателя ρ

Этот классический метод [7] основан на подсчете среднего числа мутаций, накопившихся в пределах кластера: это число и обозначается как ρ. Для этого метода необходимо иметь филогенетическое дерево, показывающее происхождение всех гаплотипов кластера друг от друга. Также необходимо знать, какой из гаплотипов является исходным (гаплотип-основатель – founder haplotype), т.е. первый "предковый" гаплотип, из которого произошли все остальные. Расчет прост: поочередно рассматривается каждый гаплотип-потомок, определяется число мутационных шагов, отделяющих его от исходного гаплотипа-основателя, а если такой гаплотип встречен более чем у одного образца, это число умножается на число образцов. Когда такой расчет проведен для каждого гаплотипа, полученные величины суммируются и делятся на общее число образцов. Можно видеть, что по сути подсчитывается число мутационных шагов, на которое изученные образцы в среднем отстоят от исходного гаплотипа. Пример использования этого метода показан на рис. 3. Для простых филогенетических деревьев расчет легко проводится вручную, но он также автоматизирован в программе Network. В работе [8] описана возможность расчета статистической ошибки показателя ρ, а возможные погрешности обсуждаются в [9].

Метод датировки с помощью показателя ASD

Для использования этого второго метода не требуется знать ни гаплотип-основатель, ни схему возникновения из него остальных гаплотипов. Метод состоит в определении средних квадратичных различий (ASD – average squared difference) между STR-гаплотипами [10]. То есть вновь как бы рассчитывается среднее расстояние от исходного гаплотипа до каждого гаплотипа в выборке, но за исходный гаплотип по каждому STR-маркеру принимается не предполагаемый гаплотип-основатель (founder), а средневзвешенное значение этого STR-маркера во всех изученных образцах кластера. Накопленное разнообразие оценивается также иначе – по дисперсии значений в разных образцах вокруг этого среднего. Проводится рас-

чет отдельно для каждого STR-маркера, а затем результаты по всем маркерам усредняются. Этот расчет легко проводится в Excel или любой статистической программе.

Сравнение методов расчета

О двух достоинствах метода ASD уже сказано – не требуется знать ни гаплотип-основатель, ни схему возникновения из него остальных гаплотипов. Недостатком его является чрезмерная упрощенность модели: ведь "среднестатистический" гаплотип вовсе не обязательно является предковым, и, главное, усреднение значений по всем локусам как бы игнорирует, что в действительности они не являются независимыми и существуют лишь в полном сцеплении друг с другом в виде конкретных гаплотипов.

Более обоснованным является показатель ρ, но и он не свободен от недостатков. Ведь для его использования требуется – не всегда известное – древо родства гаплотипов.

Скорость мутирования

Скорость мутирования является одним из ключевых параметров при генетических датировках (рис. 2,*б*). Для ее определения возможны два подхода: прямой подсчет по родословным и калибровка.

Первый подход – подсчет по родословным – состоит в сравнении генотипов родителей и их потомства. Хотя мутации случаются редко, но при больших выборках можно обнаружить достаточное их количество. Частота встречаемости мутантных аллелей в поколении потомков по сравнению с поколением родителей и будет скоростью возникновения мутаций. Такие исследования для STR-маркеров Y-хромосомы были проведены неоднократно (в основном благодаря тому, что эти маркеры часто генотипируются при определении биологического родства), каждый раз на все больших объемах выборок [11–13], и результаты совпали. Средняя скорость мутирования изученных STR-маркеров (обычно использовался 17-маркерный набор Y-filer, широко применяемый в криминалистике) составляет 2.1×10^{-3} на локус за поколение. То есть для среднестатистического STR-маркера вероятность мутировать при передаче от отца к сыну составляет 0.0021. Поскольку эта скорость мутирования определена при прямом подсчете числа мутаций в известных родословных, она получила название "генеалогической" скорости. Аналогичный подход, примененный к скорости мутирования протяженных участков, выявляемых при полном секвенировании Y-хромосомы, рассматривается ниже.

Второй подход – калибровка – состоит в определении разнообразия гаплотипов, накопленного популяцией за время ее существования. Для этого нужно изучить популяцию, для которой есть историческая датировка ее "основания", и найти в ней кластеры гаплотипов, возникшие за все время существования популяции и роста ее численности от небольшого числа основателей с данным предковым гаплотипом до современной их численности. Понятно, что этот подход намного сложнее и у него больше потенциальных источников погрешности. В то же время подход калибровок является общепринятым в науке – достаточно привести пример калибровок в радиоуглеродном методе датирования. При этих калибровках определяется доля распавшихся атомов для образца с известным возрастом, отсюда вычисляется, с какой скоростью они распадаются, и эта скорость потом применяется для расчета возраста других образцов, для которых тоже определена доля распавшихся атомов, но возраст неизвестен.

Скорость мутирования STR-маркеров Y-хромосомы была калибрована в работе [14] на двух примерах. Ими послужили кластеры гаплотипов, накопленные в популяции маори, сформировавшейся в результате миграции (не позднее 800 лет назад) полинезийцев на дотоле необитаемую Новую Зеландию, и в популяции цыган Болгарии, сформировавшейся в результате исторически датированной (около 900–1000 лет назад) миграции предков цыган из Индии и их разделения на эндогамные группы внутри Европы. К сожалению, других работ по калибровке скорости мутирования Y-STR маркеров не было проведено, поэтому отсутствуют независимые подтверждения этой скорости. При оценке этим методом для среднестатистического STR-маркера вероятность мутировать при передаче от отца к сыну составляет 0.0007. Эта скорость получила название "эволюционной", поскольку определена для эволюционирующей популяции.

Как видим, "эволюционная" и "генеалогическая" скорости различаются в 3 раза (!) – 7 шансов или 21 шанс из 10 тысяч. Поэтому и возраст гаплогрупп, рассчитанный при использовании той или иной скорости, будет различаться в 3 раза.

Дискуссии о скоростях

Трехкратные различия между "эволюционной" и "генеалогической" скоростями мутирования представляют собой научную проблему, ставшую предметом оживленных дискуссий [14–17]. Последняя из перечисленных работ подводит итоги теоретической дискуссии и предлагает объяснение различий между скоростями.

Объяснение [17] заключается в том, что новые гаплотипы возникают в популяции в том количе-

стве, как следует из "генеалогической" скорости, но в ходе эволюции популяции часть из этих гаплотипов элиминируется дрейфом генов. Поэтому при рассмотрении сохранившихся гаплотипов (а только их, естественно, и может изучить исследователь в реальной популяции) обнаруживается меньшее разнообразие, чем разнообразие, возникавшее по ходу истории популяции, а потому и рассчитываемая по этим данным скорость мутирования оказывается меньше. Это предположение было подтверждено компьютерным моделированием и был сделан важный вывод, что соотношение "генеалогической" и "эволюционной" скоростей может быть не только трехкратным, а практически любым, в зависимости от конкретных демографических параметров популяции, а также в зависимости от – во многом стохастической – демографической истории конкретной гаплогруппы. Иного объяснения трехкратным различиям скоростей в научной литературе предложено не было, и объяснение элиминацией части гаплотипов вошло в обиход.

В то же время оставалось непонятным, какой все же скоростью пользоваться на практике. Многие исследователи полагали, что поскольку эволюционная скорость была получена не для семей, а для популяций, то именно ее логично использовать для датировок истории гаплогрупп в популяциях. Другие авторы предпочитали генеалогическую скорость, а во многих работах (например, во всех работах нашего коллектива) приводились расчеты по обеим скоростям. Последний вариант особенно наглядно показывал реально существующую неопределенность в оценках возраста большинства гаплогрупп. Поэтому большинство специалистов предпочитали не строить свои выводы на столь шатком фундаменте, как генетические датировки, и предпочитали другие основания, такие как географические закономерности в распространении гаплогрупп, градиенты их разнообразия, сравнение с лингвистическими и археологическими данными и другие подходы.

Отметим, что указанием в пользу "генеалогической" скорости являются (рассматриваемые ниже) данные полного секвенирования Y-хромосомы: в них "эволюционная" скорость не отличается существенно от "генеалогической", а значит, и для STR-гаплотипов "эволюционная" должна бы быть близка к (независимо показанной во многих исследованиях) "генеалогической" скорости.

Для практического завершения дискуссии решающее значение имеет работа [18], в которой для нескольких десятков гаплогрупп были сопоставлены датировки, полученные по STR-маркерам и по данным полного секвенирования. Оказалось, что для молодых гаплогрупп (в пределах 5–10 тыс. лет) "эволюционная" скорость сильно занижает возраст, и для соответствия полногеномным результатам скорость мутирования STR-маркеров должна быть близка к генеалогической или даже превосходить ее. Зато для более древних гаплогрупп "эволюционная" скорость подходит лучше, особенно для возраста порядка 30 тыс. лет [18].

Итак, хотя окончательной ясности в этом вопросе нет, большинство аргументов указывают на обоснованность применения в большинстве случаев "генеалогической" скорости мутирования при обработке данных по Y-STR маркерам. Впрочем, этот вопрос уже потерял свою остроту, поскольку возраст гаплогрупп теперь может быть надежно определен по результатам полного секвенирования Y-хромосомы. А для датировок тех совсем молодых кластеров в пределах этих гаплогрупп, где требуется использование быстромутирующих STR-маркеров, сама их молодость (приближающая их временной диапазон к диапазону известных генеалогий) обосновывает применение к ним "генеалогической" скорости.

Длина поколения

Если скорость мутирования выражена в числе мутаций на поколение, то рассчитанное разнообразие (в числе мутаций) сначала пересчитывается в число поколений, и чтобы затем пересчитать его в годы, требуется задать длину поколения. Длина поколения определяется в демографических исследованиях популяций, и наиболее обоснованной для мужских поколений (Y-хромосомы) является оценка около 30 лет, а для женских (мтДНК) – 25–28 лет [19, 20]. Подробнее исследования длины поколения в популяциях человека описаны в [21].

Отметим, что пользоваться реальной длиной поколения необходимо только при использовании скорости мутирования, полученной на парах близких родственников (поскольку в этом случае число мутаций рассчитано для определенного числа поколений, а не числа лет). Но при использовании скорости мутирования, основанной на калибровках, скорость первоначально выражается в числе мутаций за определенное число лет (а не за число поколений). Для удобства сравнения такие скорости принято тоже переводить в число мутаций на поколение. И та длина поколения, которая была использована автором калибровки для этого пересчета, должна применяться всеми, кто пользуется этой скоростью. Например, в работе [14] длина поколения принималась равной 25 годам, и при использовании "эволюционной скорости" следует брать именно такую длину поколения.

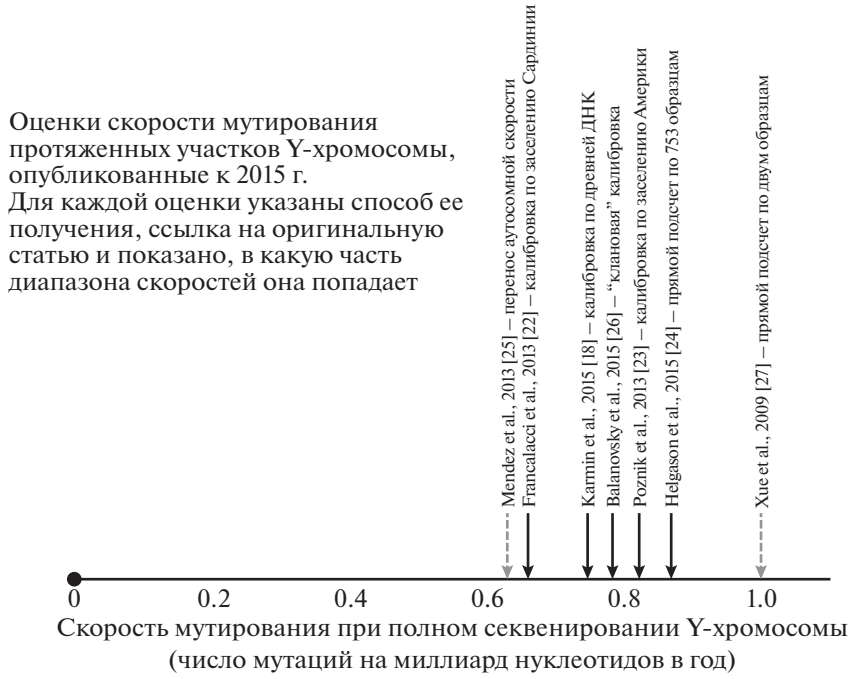


Рис. 4. Оценки "полногеномной" скорости мутирования на Y-хромосоме. Пунктирными стрелками показаны две первые опубликованные оценки, а темными – результаты последних исследований. [27] – скорость 1×10^{-9} мутаций на нуклеотид в год. Проведен прямой подсчет мутаций по 1 родословной между 2 образцами, разделенными 13 поколениями. [25] – скорость 0.62×10^{-9} мутаций на нуклеотид в год. На Y-хромосому просто перенесена скорость, принятая для аутосом. [23] – скорость 0.82×10^{-9} мутаций на нуклеотид в год. Калибровка по заселению Америки, принимая дату заселения равной 15 тыс. лет. [22] – скорость 0.65×10^{-9} мутаций на нуклеотид в год. Калибровка по археологическим датам заселения Сардинии. [26] – скорость 0.78×10^{-9} мутаций на нуклеотид в год. Калибровка по времени жизни генеалогического предка изученных представителей рода казахов-аргынов (справедливость в данном случае генеалогических преданий подтверждается совпадением с реконструированным генетическим деревом). [18] – скорость 0.74×10^{-9} мутаций на нуклеотид в год. Калибровка по древней ДНК, датированной радиоуглеродным методом, и современным образцам – ее потомкам. [24] – скорость 0.87×10^{-9} мутаций на нуклеотид в год. Проведен прямой подсчет мутаций по 274 родословным между 753 образцами, разделенными 2449 поколениями.

ДАТИРОВКИ ПО ДАННЫМ ПОЛНОГО СЕКВЕНИРОВАНИЯ

В последние годы в изучении мтДНК, а затем и Y-хромосомы произошла "полногеномная революция" – стали стремительно накапливаться данные не только по гипервариабельному сегменту мтДНК и по нескольким десяткам STR-маркеров Y-хромосомы, но и по всей последовательности мтДНК и нескольким миллионам позиций на Y-хромосоме, каждая из которых представляет собой потенциальный SNP-маркер. И хотя скорость мутирования SNP-маркеров намного меньше, чем STR, но за счет их огромного числа теперь каждый секвенированный образец характеризуется своим собственным набором встреченных только у него SNP-маркеров, а каждая ветвь на дереве тоже несет ряд специфичных для нее маркеров. Тем самым решается проблема неоднозначной реконструкции дерева – дерево теперь обычно реконструируется однозначно. Предковый гаплотип тоже четко определяется по дереву. То есть снимаются основные проблемы, свойственные применению показателя ρ. Метод расчета возраста гаплогрупп по полногеномным данным столь очевиден, что обычно даже не обсуждается, но по сути это именно ρ – среднее число мутаций от предкового узла-гаплотипа до гаплотипов-потомков служит мерой возраста кластера, происходящего от этого предкового узла. Мера возраста выражена в числе мутаций – поэтому нужно знать скорость мутирования.

Скорость мутирования для полного секвенирования Y-хромосомы уже определена во многих исследованиях – и генеалогических, и по калибровкам. В отличие от скорости мутирования STR-маркеров, калибровки получены не в одной, в нескольких работах, причем для калибровок использованы самые разные подходы. На рис. 4 показаны величины скорости, полученные во всех этих исследованиях. Видно, что в первых двух работах [25, 27] разброс был почти двукратный, но все последующие исследования дали близкие оценки, и центр тяжести определений скорости находится около величины 0.8×10^{-9} мутаций на нуклеотид в год. Иначе говоря, для среднестатистического нуклеотида на Y-хромосоме вероят-

ность мутировать за один год составляет примерно восемь шансов из 10 миллиардов. Разумеется, для одного отдельно взятого нуклеотида это весьма редкое событие, но если секвенируются 10 миллионов нуклеотидов, то вероятность мутации хотя бы одного за один год составляет восемь шансов из 100. А за 100 лет – уже 80 шансов из 100, т.е. одна мутация произойдет в среднем за 120 лет. Значит, за тысячу лет (это небольшой срок для объектов популяционных исследований) произойдет уже целых восемь мутаций, и по числу мутаций можно оценить возраст гаплогруппы с приемлемой статистической погрешностью.

Важно, что полногеномная скорость, полученная по калибровкам заселения Сардинии [22] и Америки [23], совпала с полногеномной "генеалогической" скоростью, полученной по многочисленным исландским родословным [24]. Расхождение между этой [24] и более ранней генеалогической оценкой [27] составляет лишь 13% и, несомненно, что предпочтение следует отдать более поздней и основанной на несравненно большем объеме данных "исландской" скорости. Конечно, любая из этих оценок не бесспорна: например, использованная дата заселения Америки (15 тыс. лет) весьма условна, поскольку, например, при калибровке скорости мутирования митохондриальной ДНК [7] та же дата принималась равной 25 тыс. лет.

Подход к оценке скорости мутирования, использованный в нашей собственной работе [26], можно назвать "клановым", потому что он является промежуточным между калибровкой и подсчетом по родословной: скорость определена по разнообразию, накопленному в популяции (что роднит его с калибровкой), но изученная казахская популяция аргынов происходит преимущественно от одного основателя-родоначальника (что роднит его с подсчетом по родословной).

Наконец, в работе [18] использован еще один подход: подсчет числа мутаций, по которым современные гаплотипы отличаются от секвенированного древнего образца ДНК, который филогенетически может считаться их предком (точнее, расположен на филогенетическом дереве очень близко к своему общему предку с современными гаплотипами).

Итак, в отличие от скоростей мутирования Y-STR-маркеров, в которых конкурируют две оценки, различающиеся в 3 раза, скорости мутирования при датировках по полным сиквенсам Y-хромосомы определены с приемлемой точностью, существенных разногласий в этом вопросе нет.

Для полного митохондиального генома также получены надежные оценки скоростей мутирования: для всего генома – одна мутация на 3600 лет [28], а для быстромутирующего кодирующего региона – одна мутация на 4600 лет [29].

ДОПУЩЕНИЯ ПО УМОЛЧАНИЮ

При расчете возраста кластеров делается ряд допущений или встречаются затруднения, которые обычно не упоминаются в текстах статей, но могут иметь значение для оценки надежности результатов. Рассмотрим их в этом разделе.

Селективная нейтральность и постоянство скорости

Важнейшим допущением является принцип нейтральности Кимуры, согласно которому наблюдаемые мутации не подвержены (или слабо подвержены) отбору и происходят независимо друг от друга. Также считается, что скорость возникновения мутаций не меняется ни во времени, ни в пространстве, так что вероятность возникновения мутации в данном гаплотипе не зависит от того, принадлежит ли этот гаплотип палеолитическому сибиряку или же современному африканцу. Эти допущения хорошо обоснованы, и нет экспериментальных данных, которые заставляли бы в них усомниться, указывая на роль окружающей среды. Другое дело, что скорость мутаций может несколько различаться в зависимости от генетического окружения (были работы по несколько различающимся скоростям мутирования для разных гаплогрупп мтДНК и Y-хромосомы). Достоверно показано и то, что частота мутаций на Y-хромосоме увеличивается с возрастом отца. Поэтому в принципе возможно, что в популяциях с различной длиной мужских поколений скорость накопления мутаций будет несколько различаться, хотя на обычно исследуемых промежутках времени длина поколений относительно постоянна для всего человечества. В целом на практике в постоянстве и селективной нейтральности мутаций можно не сомневаться, чего нельзя сказать о других рассматриваемых допущениях.

Разные скорости для разных маркеров

При рассмотрении скоростей мутирования упоминалось, что они относятся к "среднестатистическому" маркеру. Но частота мутирования, например, разных STR-маркеров, отличается на один–два порядка, и даже крупные сегменты Y-хромосомы неодинаковы по скорости мутирования [24]. К счастью, частота мутирований конкретного локуса обычно не важна, поскольку всегда анализируются их наборы. Поэтому нужно знать скорость мутирования именно данного набора локусов (т.е. средняя вероятность того, что смутирует какой-то любой – не важно, какой именно – локус из набора). Хотя скорости мутирования разных наборов могут различаться, если в один набор включены быстромутирующие, а в другой – медленно мутирующие маркеры, но по-

скольку используемые в анализе гаплогрупп наборы, за редким исключением, не подбирались специально по скорости мутирования, то случайное включение быстрых и медленных маркеров взаимно компенсируется, и скорости мутирования разных наборов STR-маркеров мало отличаются друг от друга.

Проблема монофилетичности кластера

Все методы датировок кластеров исходят из предположения, что рассматриваемые гаплотипы реально происходят от одного предка. Но для этого нужно знать точное дерево происхождения гаплотипов друг от друга. При использовании полного секвенирования Y-хромосомы это достигается, но при использовании данных по STR-гаплотипам получаемое дерево является лишь приближенной реконструкцией происходившей эволюции гаплотипов. Конечно, чем больше STR-маркеров анализируется, тем точнее выделяются гаплотипы и лучше реконструируется их родство друг с другом. Но даже при использовании и нескольких десятков, и нескольких сотен STR-маркеров остаются неоднозначности в филогенетическом дереве их гаплотипов.

Публикуемые схемы родства гаплотипов, даже если они отображаются в виде сетей (ретикуляции на сетях как раз предназначены для отображения разных возможных путей эволюции), часто создают иллюзию почти полной однозначности дерева. В действительности же на этих схемах степень отсечения ретикуляций определяется порогами, задаваемыми автором программе при построении дерева, или же внутренними алгоритмами самой программы. При использовании наиболее популярной программы Network сеть при отображении большинства возможных путей эволюции выглядит как спутанный клубок. Но при задании жестких условий исключения всех ветвей, кроме наиболее вероятных, клубок можно превратить в красивое звездообразное дерево: ретикуляции на нем остаются лишь там, где решительно невозможно отдать предпочтение одному из двух возможных путей. Но и не отображаемые на дереве пути эволюции тоже вполне возможны. А значит та группа гаплотипов, которая на полученном дереве показана как происходящая из одного гаплотипа-основателя, на самом деле может включать гаплотипы, имевшие других предков.

Эта погрешность, неизбежная при использовании Y-STR данных, является серьезным препятствием для надежного определения возрастов кластеров – в ряде случаев не менее серьезным, чем трехкратные различия при использовании разных скоростей мутирования. Для минимизации этой погрешности необходимо сравнивать схемы, полученные при разных параметрах и порогах, тщательно рассматривать детали топологии и брать в работу лишь те кластеры, которые при разных параметрах и порогах представляются действительно монофилетическими группами.

Найти основателя

Выделение гаплотипа-основателя представляет собой отдельную проблему. В идеальном случае на сети выявляется четкая звездообразная структура с одним гаплотипом в центре, представленном у большого числа образцов, и расходящимися от него по лучам гаплотипами-потомками (причем каждый из них представлен меньшим числом образцов, чем основатель) и потомками потомков (рис. 1,*б*, 1,*в*). Такая структура действительно иногда выявляется в реальных данных. Но куда чаще приходится работать с филогенетической сетью сложной топологии, на которой просматриваются группы гаплотипов, лишь отдаленно напоминающие звездообразную структуру. В этом случае предковым стоит считать тот гаплотип, который удовлетворяет двум условиям: во-первых, представлен у большого числа образцов; а во-вторых, имеет большое число потомков. В случае противоречия между этими правилами предпочтение стоит отдавать “правилу потомков”, потому что частота предкового гаплотипа могла за время истории популяции случайно уменьшиться, а частота одного из потомков случайно вырасти. Но если у гаплотипа на сети совсем мало потомков, то и у нас совсем мало оснований считать его исходным гаплотипом кластера.

МОДЕЛИРОВАНИЕ ВОЗРАСТА КЛАСТЕРА

Стоит четырехэтажный дом, в каждом этаже по восьми окон, на крыше – два слуховых окна и две трубы, в каждом этаже по два квартиранта. А теперь скажите, господа, в каком году умерла у швейцара бабушка?

Я. Гашек

Несмотря на хрестоматийную нелепость постановки Швейком этой задачи, она может быть успешно решена моделированием, использующим теорему Байеса. А именно, можно показать, что распределение вероятностей того, что в доходном доме столько-то стен, окон и дымоходов, не влияет на распределение периода жизни родственников швейцаров – указанные события независимы. Но в науке отрицательный результат – тоже результат, и методы, которыми он получен, определенно имеют ценность.

Байесовский подход к генетическим датировкам

Байесовское моделирование оперирует понятием правдоподобия гипотезы о событиях в про-

шлом для оценки вероятности того, что наблюдаемые (сегодня) результаты эксперимента поддерживают данную гипотезу (о прошлом) более, чем какую-либо другую. Вероятность часто пытаются измерять с помощью иных понятий и приближений, но не всегда удачно. Например, филогенетическое дерево, описывающее данные максимально экономным (парсимонным) способом, иногда может обманывать исследователей, ведь эволюция не всегда шла наиболее экономным путем [30], а дерево, построенное по принципу максимального байесовского правдоподобия, будет в этих случаях ближе к реальности.

В 90-х годах быстрое развитие как аппаратной, так и алгоритмической основы вычислений позволило применить к поиску наиболее вероятных деревьев гаплотипов мощные байесовские эвристики, которые давно использовали в физике для изучения сложных распределений [31, 32]. Если обозначить *данные* через D, а *модель* (дерево, реконструкцию нуклеотидных состояний его внутренних узлов, а также их возраста) – через M, то из формулы Байеса следует, что

$$P(M|D) = P(D|M)P(M)/P(D),$$

где $P(M|D)$ – *апостериорная* вероятность корректности некоторой модели M при конкретных экспериментальных данных D;

$P(M)$ – ее *априорная* вероятность, независимая от нашего эксперимента (например, известная из литературы скорость мутирования);

$P(D|M)$ – вероятность, что наши экспериментальные данные корректны при условии истинности модели M;

и наконец,

$P(D)$ – вероятность того, что данные корректны вообще, независимо от каких-либо условий.

Априорные вероятности $P(M)$ мы задаем на входе анализа в виде распределения – например, можем предположить, что скорость мутирования сайтов ДНК равномерно распределена на некотором интервале [0, a] или что все теоретически возможные деревья равновероятны. Вычисление вероятности $P(D|M)$ достаточно трудоемко, особенно если включает в себя реконструкцию всех предковых гаплотипов. К счастью, последнее из выражений – знаменатель $P(D)$ – константно и игнорируется.

Например, на вход программы подаются экспериментальные данные (обозначим их D), а также основные параметры модели (обозначим ее M) – вероятность замены одного нуклеотида на другой, априорное распределение возраста клады (TMRCA), априорное предположение о равновероятности всех возможных деревьев и т.д. Результатом применения байесовского метода должно стать получение апостериорного *распределения* для каждого интересующего нас параметра M. Априорное распределение не обязано быть правильным – это лишь стартовая точка, с которой программа начнет свою работу. Но полученное апостериорное распределение будет оптимально соответствующим экспериментальным данным. А получив распределение интересующего нас параметра, мы легко сможем рассчитать и его наиболее вероятное значение, и доверительный интервал.

Ход моделирования

Программа получает исходные данные и генерирует начальные значения параметров M в соответствии с переданными ей априорными распределениями. А именно: строит стартовое дерево; выполняет разметку нуклеотидных состояний на его внутренних узлах; вычисляет вероятность получения данных D, т.е. известных нам гаплотипов на листьях дерева, в предположении истинности M; а затем с учетом известной ей априорной вероятности данного значения M программа получает $P(M|D)$ по приведенной выше формуле. На практике обычно вычисляется не вероятность $P(M|D)$, а некая функция $f(M|D)$, пропорциональная плотности искомого распределения, поскольку важно не точное значение вероятности модели, а сравнение значений в различных точках.

Во время первых итераций программа должна определить, нет ли противоречия между исходными данными и параметрами модели. Далее программа вызывает минорные возмущения текущей модели M_i, например одномерные, и переходит к следующей модели M_{i+1}, оценивая ее апостериорную вероятность. Некоторые программы дают пользователю возможность выбирать типы и характеристики возмущений M (“операторов”). Но в целом изменениями управляет алгоритм Метрополиса–Гастингса: его реализация должна гарантировать, что в процессе итерации возмущений M будет найдено единственное распределение вероятностей (т.е. имеет место “сходимость” к распределению), и притом за как можно меньшее число шагов. Количество итераций (“длину цепи”) пользователь определяет сам, и на выходе программа сообщает количество независимых состояний M. Их обычно во много раз меньше, чем было выполнено всего итераций – по причине того, что соседние состояния M мало отличаются друг от друга и потому находятся в зависимости.

Выходные данные байесовских программ обычно представляют собой журнал результатов, куда для каждой итерации записаны все параметры M и полученное для нее апостериорное значение вероятности. Соответственно, можно отобрать результаты с высокими вероятностями и среди этих результатов найти среднее для каждого параметра, в том числе средний возраст интересующего нас

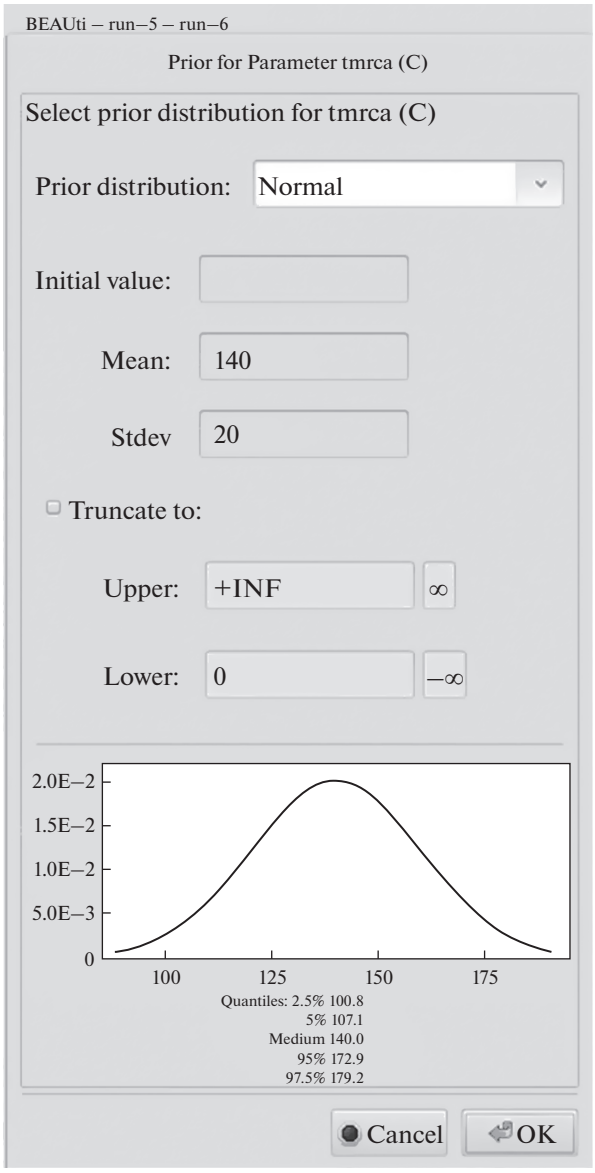


Рис. 5. Пример задания априорного распределения в программе BEAST. Возраст кластера (TMRCA – time to the most recent common ancestor, время от наиболее близкого общего предка) задан равным 140 годам. Это генеалогически документированный интервал между годами рождения общего предка семьи C и его обследованных потомков. В качестве типа распределения выбрано нормальное, а стандартное отклонение задано равным 20 годам.

кластера гаплотипов, консенсусное филогенетическое дерево и т.д. О доверительном интервале полученных значений можно судить, выбрав наименьший интервал изменения некоторого признака, представленный в заданной доле результатов, – например, промежуток возраста TMRCA от a до b лет назад, в который попадает 95% построенных деревьев. Также может быть интересным сравнение входных (априорных) и выходных (апостериорных) распределений параметров модели: они не обязаны совпадать.

Пример моделирования

Приведем пример использования программы BEAST [33], которая сейчас наиболее широко используется для генетических датировок. Байесовский метод был применен к данным полного секвенирования Y-хромосомы у 23 адыгейцев.

Семья A – 4 образца; гаплогруппа G2.

Семья H – 1 образец; гаплогруппа G2; есть версия о родстве с семьей A.

Семья B – 5 образцов; гаплогруппа G2; общий предок родился 110 лет назад.

Семья C – 7 образцов; гаплогруппа G2; общий предок родился 140 лет назад.

Семья D – 2 образца; гаплогруппа G2.

Семья E – 1 образец; гаплогруппа G2.

Семья G – 3 образца; гаплогруппа J2.

Были заданы всего несколько априорных параметров модели, отличных от предлагаемых по умолчанию: распределение скорости мутирования положили равномерным вокруг литературного значения, а для более точного ее определения задали калибровку по возрасту семей B и C, время жизни основателей которых известно документально (рис. 5).

Проведя 20 миллионов итераций, получили ряд результатов, из которых для краткости приведем только два – выходное (апостериорное) распределение возраста всего дерева (рис. 6) и консенсусное дерево с возрастами кластеров (рис. 7).

Средний возраст дерева составил 52 тыс. лет (рис. 6). Все изученные образцы относятся только к гаплогруппам G и J, поэтому дерево, построенное по этим образцам, должно укореняться в точке разделения этих гаплогрупп. И действительно, литературные данные о времени разделения гаплогрупп G и J, полученные для глобального филогенетического дерева [18], дают такие же оценки (50–55 тыс. лет назад). Байесовское моделирование истории гаплогрупп дало нам не только датировку возраста, но и доверительный интервал: оказалось, возраст корня дерева для 95% состояний моделей укладывается в промежуток 26000–82000 лет назад. Данные границы шире, чем те, которые обычно получаются методом ρ-статистики: последний сильно зависит от дисперсии длины пути от корня до листьев дерева, а длина пути от глубокого узла до листьев обычно меняется слабо, поэтому величина ошибки ρ часто оказывается неправдоподобно низкой.

Согласно консенсусному дереву (рис. 7), семейное предание о родстве A и H оказалось правдоподобным, их общий предок мог жить 345 лет назад (разброс 140–580 лет). Сложнее ситуация с

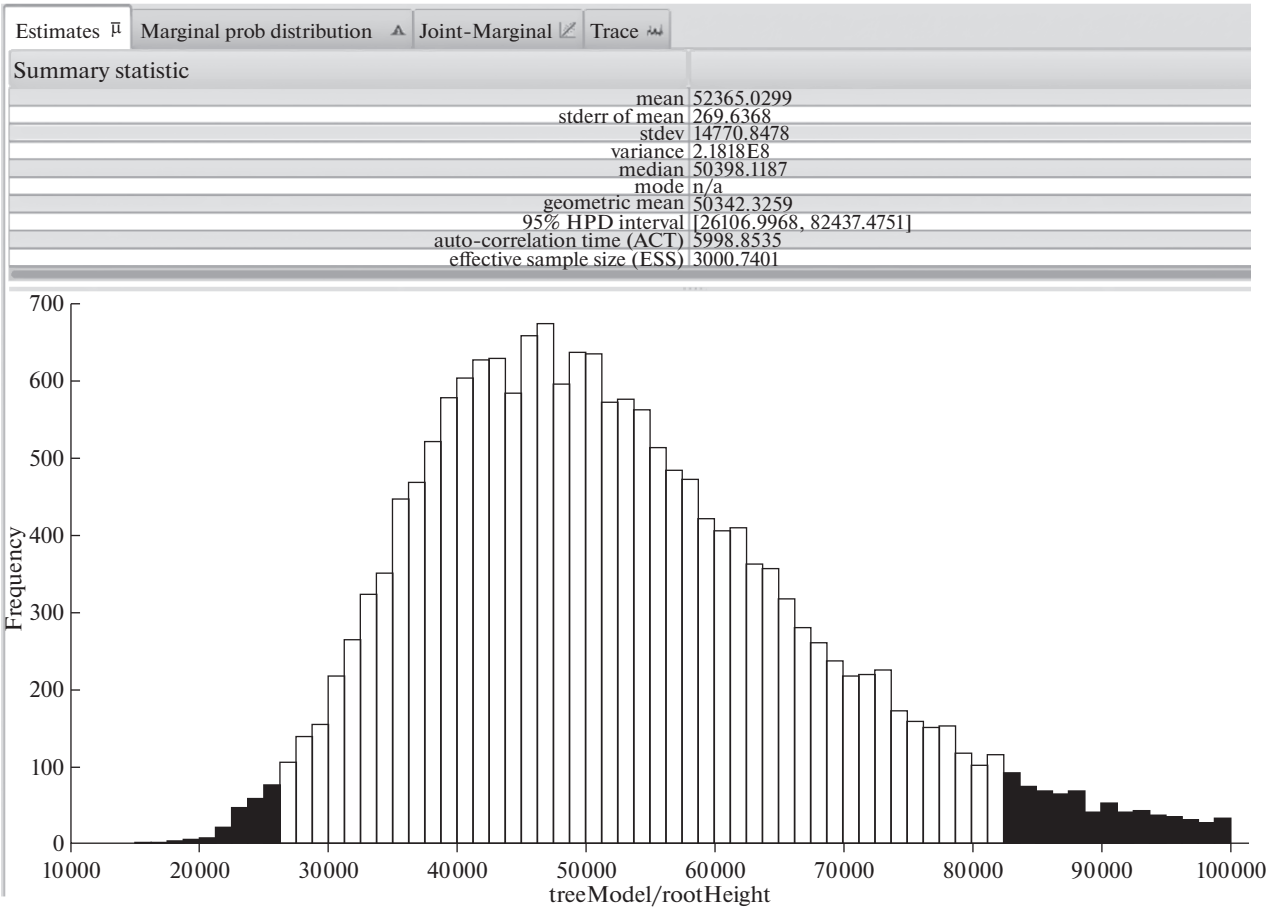


Рис. 6. Апостериорное распределение на примере возраста дерева. По оси абсцисс – возраст дерева; по оси ординат – число итераций (вариантов дерева), в которых был получен данный возраст. Черным цветом показаны значения на краях распределения, в которые суммарно попадают только 5% результатов, а белым – интервал, в который попадают 95% результатов. Видно, что вершина распределения находится между 40 и 60 тыс. лет, среднее значение (mean в таблице вверху) составляет 52000 лет.

историей семьи G – при среднем возрасте общего предка 450 лет назад доверительный интервал составляет 170–770 лет назад, что не позволяет сделать однозначного вывода о достоверности фамильного предания о родстве изученных людей.

За рамками данного обзора остается обсуждение многих интересных вопросов, например, разбиение ДНК-сайтов по группам с различной скоростью мутирования или же возможное различие скоростей мутирования одного и того же сайта в разных ветвях дерева, упомянутые выше в разделе “Допущения”. Современные программы, основанные на байесовском моделировании, позволяют включать подобные параметры и учитывать множество других фактов, накопленных при изучении мутаций.

Функциональность и степень поддержки другого байесовского пакета, BATWING, несколько ниже, чем у BEAST. Но BATWING чаще используется для работы с STR-мутациями, так как предлагает предопределенные модели эволюции микросателлитов, в то время как BEAST для работы с STR нуждается в более тонкой настройке. Интересный пример использования BATWING можно найти в работах [34, 35], где набор SNP расширен за счет использования “виртуальных мутаций”, маркирующих кластеры STR-гаплотипов.

ГЕНЕТИЧЕСКИЕ ДАТИРОВКИ ИСТОРИЧЕСКИХ СОБЫТИЙ

Под “историческим событием” в контексте рассматриваемой области исследований будем понимать только события истории населения: лишь события, приведшие к значительным изменениям в демографии, могут оставить свой след в генофонде. Поэтому открытие Америки Колумбом таким событием не является (приплыл и уплыл обратно, не оставив больших следов в генофонде), а вот походы Кортеса и Писарро – являются, поскольку они привели и к сокращению чис-

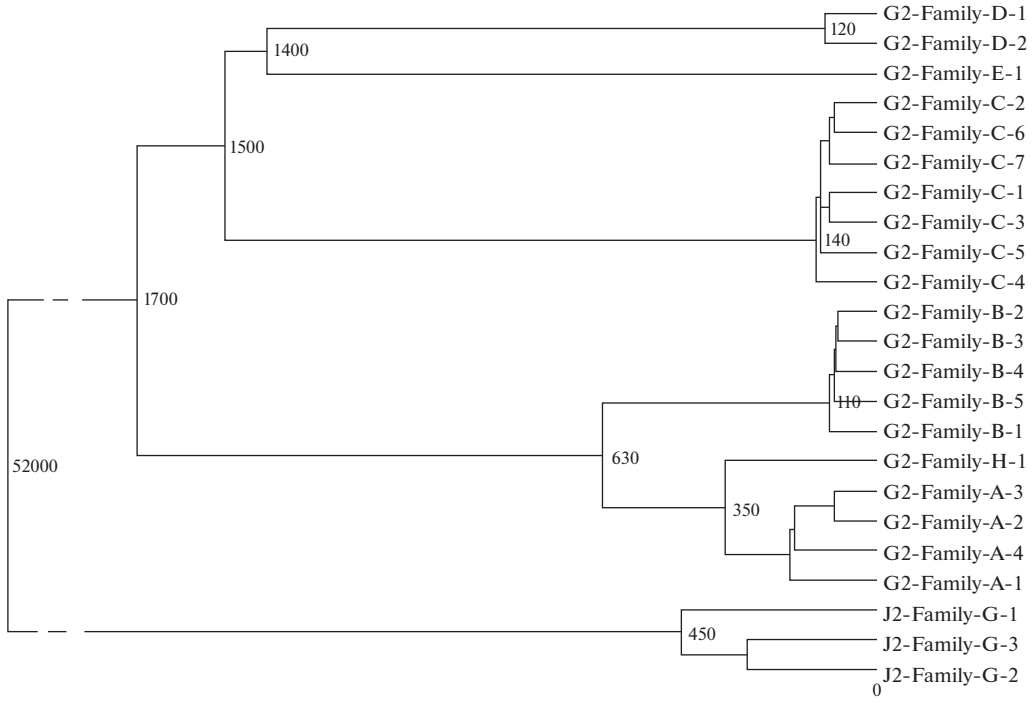


Рис. 7. Консенсусное филогенетическое дерево – результат работы программы BEAST. "Листья" дерева – непосредственно исследованные 23 Y-хромосомы. Две основные ветви соответствуют гаплогруппам G (большинство семей) и J (одна семья). Возраст основных кластеров дан у их первых развилок. Эти даты представляют собой средние значения соответствующих распределений, так же как величина 52 тыс. лет представляет собой среднее значение распределения на рис. 6.

ленности индейского населения, и к массовым смешанным бракам европейцев с индианками.

Отсутствие связи между гаплотипами и историческими событиями

Возникает вопрос, можно ли с помощью генетических датировок определять время исторических событий? И, более широко, можно ли реконструировать историю популяций исходя из филогенетических соотношений гаплотипов? Поспешность с положительным ответом на этот вопрос чревата...

Дело в том, что реконструкции филогенетических деревьев гаплотипов в целом и датировки кластеров гаплотипов в частности описывают генетические линии (хромосомы, гаплотипы, циркулирующие в популяциях), а вовсе не сами эти популяции. И, как правило, история линий (возникновение новых в результате мутаций или их исчезновение), и история популяций – это два совершенно разных процесса, идущих параллельно и зачастую независимо друг от друга. Например, популяция может целиком переселиться на новое место, а гаплотипы внутри нее продолжают возникать и исчезать точно так же, как если бы она оставалась на месте; или же популяция может поделиться на две таким образом, что гаплотипы внутри каждой из них никак не изменятся, и т.д. Поэтому в общем случае знания о гаплотипах еще ничего не говорят о популяциях.

Наличие связи между гаплотипами и историческими событиями

К счастью, из всякого правила есть исключения, и бывают такие события в истории популяций, которые приводят к характерным изменениям в составе и соотношении гаплотипов. К числу таких событий относятся: 1) резкое сокращение численности популяции; 2) резкое ее увеличение; 3) смешение с другими популяциями; 4) разделение популяций, сопровождающееся дальней миграцией одной из них. Резкое сокращение численности (случай 1) приводит к падению разнообразия гаплотипов; резкое увеличение (случай 2) – к одновременному порождению несколькими одиночными гаплотипами крупных кластеров; смешение (случай 3) тоже увеличивает разнообразие, но, в отличие от случая 2, гаплотипы будут не родственны друг другу; наконец, отпочкование с дальней миграцией (4) приводит к тому, что гаплотипы дочерней популяции представляют собой кластеры гаплотипов, происходящие от гаплотипов материнской популяции.

Вариант (4) случается всего реже, но выявляется всего достовернее и часто связан с важнейшими событиями. Например, реконструкция глобального филогенетического дерева человечества показывает, что гаплотипы коренного населения Америки представляют собой кластеры, сформировавшиеся из гаплотипов, распространенных в Евразии; а гаплотипы Евразии, в свою очередь, являются побегами одной–двух из множества африканских ветвей. Этого достаточно, чтобы реконструировать ход заселения планеты человечеством: первоначальный ареал – Африка, оттуда миграция в Евразию, а оттуда – в Америку. В этом случае, как и во многих подобных, работает вавиловский принцип: наибольшее генетическое разнообразие сохраняется на прародине (прародиной для населения Америки является Евразия, а для населения Евразии – Африка).

Примером реконструкции случаев (1) и (2) служит исследование глобального филогенетического дерева Y-хромосомы человечества [18]. Выявив все основные кластеры и датировав их, авторы определили, в какие периоды кластеров возникало много (следовательно, в эти периоды население быстро росло), а в какие численность была стабильной или сокращалась.

Когда возникновение какого-то кластера гаплотипов удается обоснованно связать с конкретным историческим событием, генетическая датировка кластера может помочь датировать само событие. Например, когда в коренном населении Америки обнаруживается кластер, свойственный только ему, то возникновение этого кластера может указывать на дату первоначального заселения Америки. Но при этом важно всегда помнить, что он может указывать на иные события. Например, если этот кластер возник у предков американских индейцев еще на их сибирской прародине, то генетическая датировка кластера будет раньше реального времени заселения Америки, поскольку ко времени прибытия в Америку молекулярные часы уже какое-то время шли. Ведь в Америку прибыла популяция не с нулевым, а уже неким накопленным генетическим разнообразием. Если же этот кластер возник во время последующей истории населения, то дата его возникновения будет позже реального времени заселения Америки. Примером такого хода событий являются кластеры, специфичные для отдельных народов Кавказа [6]: у части кластеров генетическая дата совпадает с лингвистическими датировками формирования народа, а другие кластеры значительно моложе этой даты, поскольку могли возникнуть у народа в любой период его последующей истории.

Типичные ошибки и как их избежать

При детальной изученности населения и тщательном учете этих источников погрешностей риск ошибок в связывании даты возникновения кластера и даты исторического события удается если не исключить, то снизить до минимума. Но если таких специальных усилий не прилагать, то ошибки имеют свойство возрастать до максимума.

Примером может служить обсуждение, состоявшееся на заре филогеографических исследований в России, когда опыта корректных интерпретаций таких данных было еще мало, а энтузиазма и веры в возможности молекулярных методов – еще много. Были исследованы гаплотипы русского населения Верхней Волги; рассматривая все гаплотипы как относящиеся к одному большому кластеру, была получена датировка его возникновения 20 тыс. лет. И обсуждался вывод, что эта русская популяция сформировалась на Волге 20 тыс. лет назад. Неправомочность самого вывода очевидна: 20 тыс. лет назад не только русских не было, но и сама эта территория была покрыта ледником. Но поскольку возраст кластера рассчитан правильно, то где же тут ошибка в логике? Их несколько.

Во-первых, не рассматривалось, в каких еще популяциях распространены гаплотипы этого большого кластера. А поскольку они распространены почти по всей Евразии, то датировка кластера относится к одному из этапов истории всего населения Евразии, а не к одной малой его части – русской популяции Верхней Волги.

Во-вторых, возникновение кластера было без достаточных оснований отнесено к той территории, где он найден; но ведь кластер мог возникнуть совсем в другом месте, где эта же популяция жила раньше.

В-третьих, даже если кластер был бы специфичен только для данной популяции и не было бы оснований предполагать ее миграцию с другой территории, то кластер мог возникнуть не обязательно одновременно с формированием популяции, а в любое время позже, поэтому дата кластера дает лишь верхнюю границу даты формирования популяции.

Резюмируем: в общем случае формирование дерева гаплотипов в значительной мере независимо от исторических событий, и потому одни и те же деревья и кластеры могли сформироваться при самых разных сценариях истории популяции. Но есть ряд исключений – некоторые из значительных демографических изменений оставляют в дереве гаплотипов характерные отпечатки, и по наличию таких следов можно заключать о том, что такие события имели место, а по датировкам соответствующих кластеров датировать и сами события. Самым ярким случаем является наличие кла-

стера, специфичного только для изолированной популяции (неважно, какого масштаба – от материка до поселка): в этом случае возраст кластера дает представление о времени формирования популяции. Но и тут нужна осторожность: если у основателей популяции существовал не только исходный гаплотип, но и ряд дочерних гаплотипов, то возраст кластера будет больше, чем возраст популяции. А если кластер возник какое-то время спустя после формирования популяции, то кластер, наоборот, будет моложе ее.

ГЕНЕТИЧЕСКИЕ ДАТИРОВКИ И ГРАЖДАНСКАЯ НАУКА

Всеобщая грамотность, открывшая людям доступ к научным текстам, создает иллюзии, что чтение равнозначно пониманию.

В.А. Шнирельман

Академическая наука и гражданская наука

Под “гражданской наукой” (нечасто используемая в России калька с распространенного английского термина citizen science) понимают исследования, которые проводят не академические ученые, а люди, непрофессионально интересующиеся наукой. Являясь научными по предмету исследований, эти исследования далеко не всегда следуют строгой научной методологии. А их авторы находятся около науки и формально, поскольку не интегрированы в сложившуюся структуру науки, хотя иногда взаимодействуют с ней. Гражданская наука сопутствует многим сферам академической науки, вызывающим общественный интерес, и для популяционной генетики человека “парной” к ней гражданской наукой является генетическая генеалогия.

Настойчивость, энтузиазм и несомненная талантливость помогают многим представителям генетической генеалогии получить значимые результаты, часть из которых со временем получает признание и входит в академическую науку. Однако отсутствие навыков критического подхода, поверхностность познаний и склонность к ангажированным интерпретациям приводят других представителей генетической генеалогии к выводам, имеющим с наукой мало общего. Это свойственно многим сферам знания: так, ангажированный подход к физической антропологии порождает расизм, в случае истории он порождает фолк-хистори, в случае археологии – “черных копателей”, а в случае популяционной генетики – ДНК-генеалогию. Рассмотрим сначала положительную, научную сторону генетической генеалогии, оставив ее тень – ДНК-генеалогию – напоследок.

Генетическая генеалогия и популяционная генетика: расхождения и схождения

Генеалогия в целом – это вспомогательная историческая дисциплина, изучающая родословные. Генетическая генеалогия в первоначальном, узком смысле термина – это применение генетических, молекулярных методов определения биологического родства для более точной реконструкции родословных, а также для получения хотя бы предположительных сведений о предках, когда архивные данные отсутствуют. Эта область исследований с популяционной генетикой перекрывается мало – не более, чем, например, судебная экспертиза, где тоже устанавливается биологическое родство. Но на практике общественный интерес к происхождению народов столь велик, что многие генетические генеалоги, начав с выяснения происхождения *семьи* (что относится к генеалогии), потом постепенно переключаются на выяснение происхождения *популяций*, что относится уже к сфере популяционной генетики. При этом они продолжают называть себя генетическими генеалогами, а их научная деятельность в этой сфере характеризуется несколькими особенностями.

Во-первых, не занимаясь по понятным причинам генотипированием, они целиком сосредотачиваются на анализе данных. Результаты такого анализа зависят от того, следуют ли авторы научной методологии или нет.

Во-вторых, предметом их изучения является не столько коренное население, сколько общее население разных стран – те люди, которые прошли платное коммерческое ДНК-тестирование. Исключение составляют “этнические” ДНК проекты: в них ведется отбор представителей данного этноса, но выборки не всегда репрезентативны в отношении всего генофонда этноса. Интерес к генетическому изучению своих генеалогий велик, поэтому суммарный объем баз данных “генеалогических образцов” общего населения в несколько раз больше суммарного объема выборок образцов коренного населения, генотипированных в популяционно-генетических работах. Впрочем, и те, и другие базы данных общедоступны и могут анализироваться параллельно.

В-третьих, большинство генетических генеалогов специализируется на детальном изучении одной гаплогруппы Y-хромосомы или мтДНК (чаще всего своей собственной). Среди популяционных генетиков столь узкая специализация не встречается, а генетические генеалоги, “копая” на узком поле, имеют возможность “копать” очень глубоко. Поэтому в академической науке в последние годы сформировалось мнение, что наиболее детальными познаниями по каждой отдельной гаплогруппе обладают как раз отдельные представители генетико-генеалогического сооб-

щества. В этой связи необходимо упомянуть о важном достижении генетической генеалогии – интернет-ресурсе по структуре глобального филогенетического дерева Y-хромосомы человека (http://isogg.org/tree), которое по причине частых обновлений является более широко используемым, чем аналогичные реконструкции дерева, изредка публикуемые популяционными генетиками [18, 36, 37].

Безоблачная картина совместной работы представителей академической и гражданской науки, которую один из авторов наблюдал, например, на конференции в Вашингтоне (http://i4gg.org), в условиях российского климата нередко осложняется появлением грозовых фронтов. Из многолетнего опыта мы вынесли личное, но глубокое убеждение, что расхождения между представителями генетической генеалогии и популяционной генетики обусловлены причинами психологическими. Представителям академической науки нелегко принять форму, в которой генетическая генеалогия подает свои результаты (например, в виде серии не всегда связанных между собой кратких "постов" на интернет-форумах). Представители же генетической генеалогии склонны компенсировать естественно возникающее в их положении "любителей" ощущение дискомфорта острой критикой в адрес "профессионалов". Излюбленным предметом этой критики является скорость мутирования.

Неудивительно, что генетическая генеалогия выбрала для своих расчетов "генеалогическую" скорость мутирования, которая, действительно, подтверждается не только на парах "отец-сын", но и на более глубоких генеалогиях. Но удивительно, с каким энтузиазмом критикуется скорость "эволюционная" – интернет-форумы генетических генеалогов пестрят утверждениями о "дискредитации" результатов популяционно-генетических исследований в результате использования "эволюционной" скорости, изложенной в работе [14]. С одной стороны, действительно, последние 10 лет большинство популяционных генетиков при датировках гаплогрупп по STR-гаплотипам пользовались "эволюционной" скоростью, тогда как последние работы [18, 26] показывают, что генеалогическая скорость в большинстве случаев дает лучшие результаты. Но с другой стороны, все эти годы в рамках популяционной генетики обе скорости использовались параллельно и не прекращалось обсуждение их оптимальности [6, 38–40]. Причем недавнее решение [18], совпадающее с воззрениями генетических генеалогов, было найдено в рамках собственного развития популяционных исследований. Более того, генетики, как правило, сознают трудность надежного определения возраста гаплогруппы – ведь даже если решить проблему скорости мутирования, останутся не меньшие проблемы неполноты выборки гаплотипов, неточности реконструированного дерева, и даже точно определенный возраст гаплогруппы далеко не всегда поможет датировать события в истории популяций. Но в сообществе генетических генеалогов сформировалась некая аберрация зрения: хотя расчеты возраста по STR-гаплотипам с использованием "эволюционной" скорости являются очень небольшой частью исследований Y-хромосомы, проводимых популяционными генетиками, генеалоги искренне убеждены, что "неправильная эволюционная скорость" является для популяционной генетики краеугольным камнем. К тому же в интернет-дискуссиях на эту тему нередко смешиваются скорости мутирования и метод расчета возраста кластера – хотя очевидно, что любой метод можно скомбинировать с любой скоростью.

В то же время имеется немалое число совместных достижений представителей генетической генеалогии и российской популяционной генетики: база данных Y-base, в наполнении которой большую помощь оказал Роман Сычев; программа-предиктор, разработанная для популяционных генетиков Вадимом Урасиным; совместные интернет-публикации (http://генофонд.рф/?page_id=439, http://генофонд.рф/?page_id=524), помощь ряду генеалогов, исследующих связи гаплотипов с фамилиями и родами у народов Кавказа, Закавказья и Приуралья, совместные исследования генетиков и томских генеалогов (http://ling.tspu.edu.ru/files/ling/PDF/articles/volkov_v._g._109_122_4_10_2015.pdf), и мн. др. Остается надеяться, что эти реальные совместные достижения помогут русскоязычным генетикам и генеалогам догнать и перегнать своих англоязычных коллег по части сотрудничества друг с другом.

Гаплогруппа – род – народ

Когда слова утрачивают свой смысл – народы утрачивают свободу.

Конфуций

Выше уже обсуждалось, что связь кластеров гаплотипов (и их датировок) с историей популяций неоднозначна и прямолинейные интерпретации могут приводить к ошибкам. К сожалению, именно это произошло с маргинальной ветвью генетической генеалогии, представители которой называют свою область "ДНК-генеалогией"[1]. В этой области гаплогруппы мыслятся как популяции – они мигрируют, смешиваются, развивают

[1] В англоязычной литературе термин "ДНК-генеалогия" практически не используется, а в русскоязычных текстах встречается в двух смыслах – и как обозначение маргинальной ветви генетической генеалогии, также называемой по фамилии ее предводителя "ДНК-генеалогией Клесова", и как устаревший синоним "генетической генеалогии" в целом.

археологические культуры и даже воюют друг с другом, побеждают и вытесняют одни других, имеют разный "социальный" ранг. Применительно к фрагментам ДНК, это, конечно, бессмыслица, но в рамках ДНК-генеалогии "гаплогруппы" отождествляются с людьми – их носителями – и с популяциями. Подобная логика основывается на убеждении, что каждая популяция состоит из носителей одной гаплогруппы и "маркируется" ей, а если в популяции встречено несколько гаплогрупп, то такая популяция возникла слиянием нескольких гаплогрупп-популяций. В силу такого убеждения ДНК-генеалогия механически переносит на историю популяций и народов все обилие имеющихся данных по филогеографии преобладающей в данной популяции гаплогруппы, включая место и время ее возникновения, пути распространения, деление на дочерние субгаплогруппы и т.д. В результате ДНК-генеалогия легко "решает" многие проблемы истории популяций, включая прародину человечества (ею оказывается не Африка, а Россия), индоевропейскую проблему, происхождение славян (праславяне оказываются ариями), и т.д.

Причем лежащее в основе ненаучных реконструкций убеждение, что популяция – это гаплогруппа, и что поэтому реконструкция истории гаплогруппы рассказывает историю популяции – вовсе не является ни безоговорочно неверным, ни чуждым академической науке. Такие исключения существуют, но ДНК-генеалогией они возводятся в правило. Можно привести ряд примеров, когда популяция действительно состоит из одной гаплогруппы: все коренное население Америки относится только к гаплогруппе Q-M3; многие крупные роды и кланы тюркоязычных народов восходят каждая к своему биологическому основателю и представляют, соответственно, кластеры гаплотипов Y-хромосомы; "народ желтых листьев" Таиланда имеет не только одну гаплогруппу, но даже единственный гаплотип митохондриальной ДНК. Однако риск приравнять гаплогруппу к популяции дамокловым мечом нависает над всеми исследователями гаплоидных генетических систем в популяциях человека. Уж очень велик соблазн принять историю гаплогруппы, встреченной в популяции, за историю самой этой популяции, забыв о том, что, хотя демографическая история популяции действительно влияет на историю содержащихся в ней гаплогрупп, эти две истории редко когда могут быть приравнены друг к другу.

В действительности же популяций, генофонд которых состоит из одной гаплогруппы, ничтожное меньшинство. Причина в том, что для возникновения "моногаплогруппных" популяций (фиксации аллеля) необходим чрезвычайно сильный дрейф генов, который случается редко и длится исторически недолго. А для того, чтобы "моногаплогруппная" популяция перестала быть таковой, достаточно даже небольшого потока генов из других популяций, что происходит постоянно с любыми популяциями человека. Если же в популяции и присутствует только одна гаплогруппа (например, Q-M3 в коренном населении Америки), то филогеография именно этой гаплогруппы будет информативна только на одном отрезке времени – при изучении только этапа формирования популяции, а для изучения последующих этапов ее истории – будет непригодна. Для анализа последующих этапов истории популяции уже надо изучать не Q-M3, а те субгаплогруппы, которые возникли в ее пределах. Но в отношении этих субгаплогрупп популяции индейцев уже не являются "моногаплогруппыми", а включают сразу несколько субгаплогрупп.

Итак, полной связи гаплогрупп с популяциями нет. Возникает вопрос, есть ли связь частичная? Например, если какая-то гаплогруппа составляет заметную часть генофонда популяции, то происхождение этой гаплогруппы маркирует происхождение части генофонда популяции, т.е. какой-то субстрат или миграционный поток в нее. Это справедливо, и на этом основано использование гаплогрупп как маркеров миграций. Например, обнаружив в современных популяциях индейцев гаплогруппы, свойственные европейцам, прослеживают миграции Нового времени [41]. Но дело в том, что такая логика работает только на "один шаг", на короткий отрезок времени. Потому что, раз соединившись в одной популяции, разные гаплогруппы уже являются частями одного генофонда и далее мигрируют только совместно. Например, если эта метисированная популяция мигрирует дальше, то маркерами этой миграции являются уже в равной мере не только "европейские", но и "индейские" гаплогруппы, и история этой дочерней популяции не сводится к истории "европейских" гаплогрупп.

Остается рассмотреть еще один аргумент, которым ДНК-генеалогия обосновывает правомочность своего подхода: гаплогруппы называются "родами", и история народа описывается как история входящих в него родов-гаплогрупп. Ведь то, что многие *народы* состоят или состояли из *родов*, хорошо известно из этнографии, а для обыденной логики следует и из самого сходства этих слов. Род определяется в этнологии как социальная группа, члены которой считают, что происходят от общего предка (как правило, по мужской линии) – мифического или исторически реального. А гаплогруппа объединяет людей, которые действительно происходят по мужской линии от общего предка. Казалось бы, если люди справедливо считают себя родственниками, то понятия рода и гаплогруппы совпадают. И описано много

таких популяций-родов, действительно имеющих в своем генофонде одну основную гаплогруппу [26, 42–44 и мн. др. работы]. Но на деле ни одна из таких популяций-родов не является состоящей только из носителей одной гаплогруппы: любой крупный род всегда включает “приемных” членов, приносящих в генофонд популяции другие гаплогруппы. И наоборот, одна и та же гаплогруппа часто встречается у различных родоплеменных групп: возникнув изначально в одной из них, она затем распространяется по широкому кругу популяций. Поэтому связь рода и гаплогруппы, во-первых, может быть или же отсутствовать (и выяснение этого вопроса является важной задачей популяционных генетиков). А во-вторых, если такая связь обнаруживается, то она работает только на коротких временны́х дистанциях. Поэтому “когда слова утрачивают свой смысл” и гаплогруппу называют “родом”, приписывая ему многотысячелетнюю историю гаплогруппы, это становится искажением, смешивающим биологическое и социальное понятия.

Итак, ДНК-генеалогия описывает историю народов как историю отдельных родов-гаплогрупп, каждый из которых имеет свою историю, свои маршруты миграции по планете и которые, если иногда и соединяются в одной популяции, то затем вновь расходятся, сохраняя свое гаплогруппное “лицо”, каждый своей дорогой. Но, как показано выше, такой взгляд ненаучен: он неправомочен ни как общее правило (гаплогруппа – это не популяция), ни как исключение (история гаплогруппы – это не история популяции, даже если гаплогруппа в популяции только одна), ни даже как описание родоплеменных групп (гаплогруппа – это не род).

Датировки по доле исходного гаплотипа (логарифмический метод)

В рамках основной темы данного обзора необходимо рассмотреть и особенности генетических датировок, практикуемые в ДНК-генеалогии. Они включают не только два выше описанных метода расчета (ρ и ASD), но и “логарифмическую” формулу, которую в популяционной генетике не используют из-за отсутствия каких-либо ее преимуществ. Эта формула легко выводится из распределения Пуассона и описывает логарифмический закон убывания исходного варианта при постоянной скорости его превращения в производные варианты. Она широко используется в ядерной физике для описания радиоактивного распада, в химической кинетике для описания химической реакции, в лингвистике для описания замены слов, а ДНК-генеалогия применила ее для описания возникновения кластеров гаплотипов. И, подобно тому как радиоуглеродный метод позволяет датировать образец, а глоттохронологический метод позволяет датировать время расхождения языков, в ДНК-генеалогии датируется время возникновения кластера гаплотипов.

Этот метод требует только знания того, какой гаплотип является исходным. Определяется его частота, и из нее рассчитывается прошедшее время: чем больше прошло времени, чем меньше становится эта частота, поскольку исходный гаплотип постепенно мутирует в производные. Достоинством такого метода является только его простота для вычислений, что важно для любителей. Можно было бы сказать, что другим достоинством является то, что не требуется знать филогенетическое дерево гаплотипов, но тем же достоинством обладает и метод ASD. Недостатком же – и весьма существенным – метода доли исходного гаплотипа является то, что он применим только для самых простых филогенетических схем, а если доля исходного гаплотипа стала мала по сравнению с суммарным числом производных гаплотипов, точность и надежность метода стремительно падает.

При использовании этого метода используется поправка на обратные мутации. Действительно, могут происходить и обратные мутации производных гаплотипов в исходный, и они занизят наблюдаемую долю производных гаплотипов по сравнению с долей реально возникавших производных гаплотипов. Лежащая в основе метода теоретическая модель STR-мутаций в предположении пошагового мутирования разработана Дмитрием Адамовым; эта модель позволила вывести аппроксимационную формулу, учитывающую возвратные мутации, которая носит название формулы Адамова–Клесова (Адамов, личн. сообщ; http://dna-academy.ru/wp-content/uploads/2_1_2009.pdf). Отметим, что эта поправка обоснована при использовании внешне заданной скорости мутирования, но при использовании калиброванных скоростей мутирования такая поправка не нужна – ведь обратные мутации могли происходить и в примере, послужившем основой калибровки, а значит уже в неявном виде учтены.

Критическое обсуждение различных методов, применяемых в ДНК-генеалогии, уже имеется в нескольких интернет-публикациях генетических генеалогов Дмитрия Адамова и Сергея Каржавина, в том числе сравнение ρ-статистики, ASD и логарифмического метода, а также обсуждение границ применимости каждого из методов. Можно надеяться, что эти исследования будут вскоре опубликованы и в научной периодике.

Можно видеть, что наработки ДНК-генеалогии имеют ограниченное значение для генетической генеалогии в целом, и почти никакого – для популяционной генетики. К тому же эта составляющая, которая может иметь отношение к нау-

ке, – лишь малая часть всей “теории” ДНК-генеалогии. К основной вненаучной части относится, например, постулат, что генетика изучает только гены, а все остальные – негенные – участки ДНК являются предметом изучения ДНК-генеалогии. С помощью наслаивания одних неверных постулатов на другие и формируется комплекс воззрений ДНК-генеалогии, получивший в научных кругах статус лженаучного (“Троицкий Вариант”, № 1(170), 13 января 2015 г.).

Даже столь краткому обсуждению ДНК-генеалогии было бы не место в научной статье, если бы, во-первых, она не распространялась столь широко в русскоязычном интернете, что многие ученые, не являющиеся популяционными генетиками, поневоле знакомятся с ней и затрудняются в отделении зерен от плевел, и во-вторых, если бы идея отождествления истории гаплогрупп с историей популяций, в законченном виде воплощенная ДНК-генеалогией, не была бы постоянным соблазном и для самих популяционных генетиков, не исключая и авторов этих строк. Этот соблазн только возрастает при современном буме исследований древней ДНК, где изучаются выборки крайне малого объема и поэтому индивидуальные данные вынужденно распространяются на популяцию в целом.

Выражаю признательность Дмитрию Адамову за конструктивные комментарии к ранней версии этого обзора.

Работа выполнена при финансовой поддержке Российского научного фонда, проект 14-04-00827. Данные секвенирования, использованные в примере байесовского моделирования, получены в рамках проекта лаборатории исторической генетики МФТИ.

СПИСОК ЛИТЕРАТУРЫ

1. *Балановский О.П.* Генофонд Европы. М.: Тов-во науч. изд. КМК, 2015. 356 с.
2. *Bandelt H.J., Macaulay V., Richards M.* Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA // Mol. Phylogenet. Evol. 2000. V. 16. № 1. P. 8–28.
3. *Rootsi S., Zhivotovsky L.A., Baldovic M. et al.* A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe // Eur. J. Hum. Genet. 2007. V. 15. № 2. P. 204–211. doi 10.1038/sj.ejhg.5201748
4. *Derenko M., Malyarchuk B., Denisova G. et al.* Y-chromosome haplogroup N dispersals from south Siberia to Europe // J. Hum. Genet. 2007. V. 52. № 9. P. 763–770. doi 10.1007/s10038-007-0179-5
5. *Haber M., Platt D.E., Badro D.A. et al.* Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon // Eur. J. Hum. Genet. 2011. V. 19. № 3. P. 334–340. doi 10.1038/ejhg.2010.177
6. *Balanovsky O., Dibirova Kh., Dybo A. et al.* Parallel Evolution of Genes and Languages in the Caucasus Region // Mol. Biol. Evol. 2011. V. 28. № 10. P. 2905–2920. doi 10.1093/molbev/msr126
7. *Forster P., Harding R., Torroni A., Bandelt H.J.* Origin and evolution of Native American mtDNA variation: a reappraisal // Am. J. Hum. Genet. 1996. V. 59. № 4. P. 935–945.
8. *Saillard J., Forster P., Lynnerup N. et al.* mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion // Am. J. Hum. Genet. 2000. V. 67. № 3. P. 718–726.
9. *Cox M.P.* Accuracy of molecular dating with the rho statistic: deviations from coalescent expectations under a range of demographic models // Hum. Biol. 2008. V. 80. P. 335–357. doi 10.3378/1534-6617-80.4.335
10. *Sengupta S., Zhivotovsky L.A., King R. et al.* Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists // Am. J. Hum. Genet. 2006. V. 78. № 2. P. 202–221.
11. *Gusmão L., Sánchez-Diz P., Calafell F. et al.* Mutation rates at Y chromosome specific microsatellites // Hum. Mutat. 2005. V. 26. № 6. P. 520–528.
12. *Sánchez-Diz P., Alves C., Carvalho E. et al.* Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study // Int. J. Legal Med. 2008. V. 122. № 6. P. 529–533. doi 10.1007/s00414-008-0265-z
13. *Ge J., Budowle B., Aranda X.G. et al.* Mutation rates at Y chromosome short tandem repeats in Texas populations // Forensic Sci. Int. Genet. 2009. V. 3. № 3. P. 179–184. doi 10.1016/j.fsigen.2009.01.007
14. *Zhivotovsky L.A., Underhill P.A., Cinniođlu C. et al.* The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time // Am. J. Hum. Genet. 2004. V. 37. № 1. P. 50–61. doi 10.1086/380911
15. *Di Giacomo F., Luca F., Popa L.O. et al.* Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe // Hum. Genet. 2004. V. 115. № 5. P. 357–371.
16. *Zhivotovsky L.A., Underhill P.A.* On the evolutionary mutation rate at Y-chromosome STRs: comments on paper by Di Giacomo et al. (2004) // Hum. Genet. 2005. V. 116. № 6. P. 529–532.
17. *Zhivotovsky L.A., Underhill P.A., Feldman M.W.* Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size // Mol. Biol. Evol. 2006. V. 23. № 12. P. 2268–2270. doi 10.1093/molbev/msl105
18. *Karmin M., Saag L., Vicente M. et al.* A recent bottleneck of Y chromosome diversity coincides with a global change in culture // Genome Res. 2015. V. 25. № 4. P. 459–466. doi 10.1101/gr.186684.114
19. *Fenner J.N.* Cross-cultural estimation of the human generation interval for use in genetics-based population

divergence studies // Am. J. Phys. Anthropol. 2005. V. 128. P. 415–423.

20. *Почешхова Э.А.* Геногеографическое изучение народов Западного Кавказа: Дис. д-ра мед. наук. М., 2008. 298 с.
21. *Тетушкин Е.Я.* Генетическая генеалогия: история и методология // Генетика. 2011. Т. 47. № 5. С. 581–596.
22. *Francalacci P., Morelli L., Angius A. et al.* Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny // Science. 2013. V. 341. № 6145. P. 565–569. doi 10.1126/science.1237947
23. *Poznik G.D., Henn B.M., Yee M.C. et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females // Science. 2013. V. 341. № 6145. P. 562–565. doi 10.1126/science.1237619
24. *Helgason A., Einarsson A.W., Guđmundsdóttir V.B. et al.* The Y-chromosome point mutation rate in humans // Nat. Genet. 2015. V. 47. № 5. P. 453–457. doi 10.1038/ng.3171
25. *Mendez F.L., Krahn T., Schrack B. et al.* An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree // Am. J. Hum. Genet. 2013. V. 92. № 3. P. 454–459. doi 10.1016/j.ajhg.2013.02.002
26. *Balanovsky O., Zhabagin M., Agdzhoyan A. et al.* Deep phylogenetic analysis of haplogroup G1 provides estimates of SNP and STR mutation rates on the human Y-chromosome and reveals migrations of Iranic speakers // PLoS One. 2015. V. 10. № 4. doi 10.1371/journal.pone.0122968
27. *Xue Y., Wang Q., Long Q. et al.* Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree // Curr. Biol. 2009. V. 19. № 17. P. 1453–1457. doi 10.1016/j.cub.2009.07.032
28. *Soares P., Ermini L., Thomson N. et al.* Correcting for purifying selection: an improved human mitochondrial molecular clock // Am. J. Hum. Genet. 2009. V. 84. P. 740–759.
29. *Perego U.A., Achilli A., Angerhofer N. et al.* Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups // Curr. Biol. 2009. V. 19. P. 1–8. doi 10.1016/j.cub.2008.11.058
30. *Felsenstein J.* Evolutionary trees from DNA sequences: a maximum likelihood approach // J. Mol. Evol. 1981. V. 17. P. 368–376.
31. *Li S., Pearl D.K., Doss H.* Phylogenetic three construction using Markov chain Monte Carlo // J. Amer. Stat. Assoc. 2000. V. 95. P. 493–508.
32. *Larget B., Simon D.L.* Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees // Mol. Biol. Evol. 1999. V. 16. № 6. P. 750–759.
33. *Drummond A.J., Suchard M.A., Xie D., Rambaut A.* Bayesian phylogenetics with BEAUti and the BEAST 1.7 // Mol. Biol. Evol. 2012. V. 29. P. 1969–1973.
34. *Cruciani F., La Fratta R., Santolamazza P. et al.* Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa // Am. J. Hum. Genet. 2004. V. 74. № 5. P. 1014–1022. doi 10.1086/386294
35. *Cruciani F., La Fratta R., Torroni A. et al.* Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers // Hum. Mutat. 2006. V. 27. № 8. P. 831–832.
36. Y Chromosome Consortium. A nomenclature system for the three of human Y-chromosomal binary haplogroups // Genome Res. 2002. V. 12. № 2. P. 339–348.
37. *Karafet T.M., Mendez F.L., Meilerman M.B. et al.* New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree // Genome Res. 2008. V. 18. № 5. P. 830–838. doi 10.1101/gr.7172008
38. *Pichler I., Fuchsberger C., Platzer C. et al.* Drawing the history of the Hutterite population on a genetic landscape: inference from Y-chromosome and mtDNA genotypes // Eur. J. Hum. Genet. 2010. V. 18. P. 463–470.
39. *Dulik M.C., Zhadanov S.I., Osipova L.P. et al.* Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between native Americans and indigenous Altaians // Am. J. Hum. Genet. 2012. V. 90. P. 573.
40. *Rootsi S., Myres N.M., Lin A.A. et al.* Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus // Europ. J. Hum. Genet. 2012. V. 20. № 12. P. 1275–1282. doi 10.1038/ejhg.2012.86
41. *Vilar M.G., Melendez C., Sanders A.B.* Genetic diversity in Puerto Rico and its implications for the peopling of the Island and the West Indies // Am. J. Phys. Anthropol. 2014. V. 155. № 3. P. 352–368. doi 10.1002/ajpa.22569
42. *Abilev S., Malyarchuk B., Derenko M. et al.* The Y-chromosome C3* star-cluster attributed to Genghis Khan's descendants is present at high frequency in the Kerey clan from Kazakhstan // Human Biol. 2012. V. 84. № 1. Article 12.
43. *Жабагин М.К., Дибирова Х.Д., Фролова С.А. и др.* Связь изменчивости Y-хромосомы и родовой структуры: генофонд степной аристократии и духовенства казахов // Вестн. Московск. ун-та. Сер. XXIII "Антропология". 2014. № 1. С. 96–101.
44. *Богунов Ю.В., Мальцева О.В., Богунова А.А., Балановская Е.В.* Нанайский род самар: структура генофонда по данным маркеров Y-хромосомы // Археология, этнография и антропология Евразии. 2015. Т. 43. № 2. С. 146–152. doi 10.17746/1563-0102.2015.43.2.146-152

Chromosome as a Chronicler: Genetic Dating, Historical Events, and DNA-Genealogic Temptation

O. P. Balanovsky[a,b] and V. V. Zaporozhchenko[b,a]

[a] *Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, 119991 Russia*
e-mail: balanovsky@inbox.ru
[b] *Research Centre of Medical Genetics, Moscow, 115478 Russia*

Received February 16, 2016

Abstract—Nonrecombinant portions of the genome, Y chromosome and mitochondrial DNA, are widely used for research on human population gene pools and reconstruction of their history. These systems allow the genetic dating of clusters of emerging haplotypes. The main method for age estimations is ρ statistics, which is an average number of mutations from founder haplotype to all modern-day haplotypes. A researcher can estimate the age of the cluster by multiplying this number by the mutation rate. The second method of estimation, ASD, is used for STR haplotypes of the Y chromosome and is based on the squared difference in the number of repeats. In addition to the methods of calculation, methods of Bayesian modeling assume a new significance. They have greater computational cost and complexity, but they allow obtaining an a posteriori distribution of the value of interest that is the most consistent with experimental data. The mutation rate must be known for both calculation methods and modeling methods. It can be determined either during the analysis of lineages or by providing calibration points based on populations with known formation time. These two approaches resulted in rate estimations for Y-chromosomal STR haplotypes with threefold difference. This contradiction was only recently refuted through the use of sequence data for the complete Y chromosome; "whole-genomic" rates of single nucleotide mutations obtained by both methods are mutually consistent and mark the area of application for different rates of STR markers. An issue even more crucial than that of the rates is correlation of the reconstructed history of the haplogroup (a cluster of haplotypes) and the history of the population. Although the need for distinguishing "lineage history" and "population history" arose in the earliest days of phylogeographic research, reconstructing the population history using genetic dating requires a number of methods and conditions. It is known that population history events leave distinct traces in the history of haplogroups only under certain demographic conditions. Direct identification of national history with the history of its occurring haplogroups is inappropriate and is avoided in population genetic studies, although because of its simplicity and attractiveness it is a constant temptation for researchers. An example of DNA genealogy, an amateur field that went beyond the borders of even citizen science and is consistently using the principle of equating haplogroup with lineage and population, which leads to absurd results (e.g., Eurasia as an origin of humankind), can serve as a warning against a simplified approach for interpretation of genetic dating results. English translation of the paper published in Russian Journal of Genetics, 2016, Vol. 52, No. 7, is available ONLINE by subscription from: http://www.springer.com/, http://link.springer.com

Keywords: dating, haplotype, cluster, Y chromosome, mitochondrial DNA, whole genome sequencing, mutation