Ultima ratio

Вестник Российской Академии ДНК-генеалогии

Том 1, № 4

2008 октябрь

Российская Академия ДНК-генеалогии

ISSN 1942-7484

Вестник Российской Академии ДНК-генеалогии. Научно-публицистическое издание Российской Академии ДНК-генеалогии. Издательство Lulu inc., 2008.

Составитель
Российская Академия ДНК-генеалогии

Оформление издания
Михаил Темош

СОДЕРЖАНИЕ НОМЕРА

Загадки «западноевропейской» гаплогруппы R1b

Анатолий Клёсов

http://aklyosov.home.comcast.net

Вопреки ранним, и фактически необоснованным утверждениям, что R1b является «западноевропейской» гаплогруппой, предки которой жили в Европе 30 тысяч лет назад, и определенно были кроманьонцами, на самом деле европейский вариант R1b является относительно молодой гаплогруппой (в основном R1b1b2/M269), предок которой пришел в Европу из Азии не более 4500-5000 лет назад. Азиатский вариант этой гаплогруппы (в основном R1b1b1/M73) настолько отличается по мутациям от европейского, что это показывает, что их общий предок жил – скорее всего, в Азии – примерно 16 тысяч лет назад. Общие предки наиболее популярных в Европе субкладов – R1b1b2/R1b1c/M269, R1b1b2a1/R1b1c9/U106/S21/M405, R1b1b2a2g/R1b1c10/U152/S28 – жили примерно в одно и то же время, 4200 – 4500 лет назад. Даже баски, которых (без особых обоснований в отношении гаплогруппы R1b) считали старейшими жителями Европы, имеют общего предка по гаплогруппе R1b не ранее 4000-4600 лет назад.

Гаплогруппа R1b привлекает особое внимание любителей и профессионалов ДНК-генеалогии. Причина проста – ее больше всего у тех и у других, по крайней мере по данным, доступным в настоящее время. Эта гаплогруппа доминирует в Западной и Центральной Европе и на Британских островах, и поэтому ее чаще всего тестируют у тех, кто способен оплатить коммерческое определение гаплотипа и гаплогруппы. По ряду причин такими и оказываются жители Западной Европы.

По данным на начало 2008 года в базе данных YSearch было 44093 гаплотипа, и из них 17171, то есть почти 40%, принадлежали гаплогруппе R1b с подгруппами. Это показывает желание и

возможности европейцев узнать свои корни. Для сравнения – только 31 человек гаплогруппы R1a1 из России и Украины, из нескольких десятков миллионов, проявили пока такое сочетание желания и возможности.

Как часто бывает в ДНК-генеалогии, фактически только зарождающейся области науки, гаплогруппа R1b, как самая популярная, стала быстро обрастать легендами и фантазиями. Часть из них были основаны на скоропалительных и непроверенных результатах, тем не менее опубликованных в серьезных научных журналах и книгах. Часть появились вообще без всякого обоснования.

Наиболее устойчивая легенда говорит о том, что гаплогруппу R1b имели кроманьонцы, что эта гаплогруппа была в Европе 30-35 тысяч лет назад, ее обладатели охотились бок о бок с неандертальцами, и ее же обладатели оставили пещерные рисунки на юге Европы, которые датируются вплоть до 32 тысяч лет назад. В качестве примера особо древних носителей гаплогруппы R1b часто приводят басков. Видимо, потому что у них древний язык, не относящийся к индоевропейской языковой группе.

Вот как, например, описывает гаплогруппу R1b один из пионеров ДНК-генеалогии, Спенсер Уэллс, в своей недавней книге Deep Ancestry (2006): «Примерно 30 тысяч лет назад один из потомков клана, который находился на пути в Европу, получил мутацию M343, которая и определила его в новую гаплогруппу R1b. Потомки этого человека – прямые потомки кроманьонцев, которые доминировали в освоении Европы, создали знаменитые наскальные рисунки в пещерах на юге Франции».

Свою роль в этих легендах сыграли некорректные методы расчета «вариации» мутаций в гаплотипах, использование неких «популяционных» скоростей мутаций, в которых весьма произвольно и на основании некритичных допущений резко отдалялись времена жизни «общих предков» выборок гаплотипов, делались ошибки в отнесении гаплотипов к гаплогруппам, или расчеты проводились вообще без таких отнесений.

В данную работу не входит детальное рассмотрение того, как именно возникали легенды про гаплогруппу R1b, и в чем именно были ошибки

в расчетах. Хотя следует отметить по меньшей мере три источника таких ошибок: (а) использование неверных скоростей мутаций в гаплотипах, например, «коэффициента Животовского», который примерно в три раза отдаляет «время жизни общего предка» (Zhivotovsky et al, 2004; Sengupta et al, 2006; Karlsson et al, 2006; Derenko et al, 2007; Gayden et al, 2007; Martinez et al, 2007; Contu et al, 2008), (б) смешивание гаплотипов из разных гаплогрупп, и (в) использование коротких гаплотипов, иногда состоящих всего из трех-четырех маркеров.

Генезис гаплогруппы R1b

Рассмотрим, когда же на самом деле носители гаплогруппы R1b появились в Европе, откуда они там появились или могли появиться. Оборот «на самом деле» отражает, конечно, уровень современных представлений в ДНК-генеалогии. С появлением принципиально новых данных картина может измениться. Но это – нормальный путь научного познания.

В целом генезис гаплогруппы R1b (с некоторыми подгруппами) выглядит сейчас следующим образом:

«Адам» → M168 → M89 → M9 → M207 → M173 → M343 → P25 → P297/M73 → M269 → (S116; S21/U106/M405; S28/U152)

В начале цепочки – «хромосомный Адам», который положил начало современному человечеству. К нему сходятся генеалогические линии всех людей на Земле. Каждый мужчина на Земле унаследовал его Y-хромосому и гаплотип (то есть определенную конфигурацию нуклеотидных последовательностей в определенных локусах Y-хромосомы, которые называют маркерами), и последующие поколения добавляли лишь мутации в эти маркеры. В итоге мы имеем то разнообразие гаплотипов и гаплогрупп, которое наблюдается у современных людей.

Первая мутация, M168, дала сводную гаплогруппу C-R в будущей гаплогруппе R1b. Гаплогруппы A и B остались в Африке. Мутация M89 привела к укороченной сводной гаплогруппе F-R. Дальнейшая мутация еще более укоротила сводную гаплогруппу, до K-R. Мутация M207 дала

гаплогруппу R. Мутация M271 привела к гаплогруппе R1. Наконец, мутация M343 дала мутацию R1b. Как будет показано ниже, это произошло примерно 16 тысяч лет назад, в Азии, на участке пути в Европу, о чем первые носители гаплогруппы R1b, конечно, не знали. Пару тысяч лет спустя носители гаплогруппы R1b были в Армении. 5-6 тысяч лет назад они были, по-видимому, на Средне-Русской равнине, и, возможно, это они оставили археологические культуры боевых топоров, она же культура шнуровой керамики, она же культура одиночных могил. Не исключено, что и курганную культуру, хотя здесь необоснованных гипотез предостаточно в отношении и гаплогруппы R1a1. Но данных пока нет.

Здесь приходится постоянно говорить «возможно», потому что археологи не знают, к какому ДНК-генеалогическому роду относились те (и почти все остальные) культуры, и таких определений пока не проводилось. Поэтому я здесь строю возможную модель, которую в дальнейшем нужно проверять.

4500-5000 лет назад носители R1b прибыли в Европу, примерно тогда же, когда члены рода R1a1 начали экспансию с Балкан во все стороны света – к Атлантике, к Скандинавии, на северо-восток и восток, на территории современных России, Украины, Белоруссии, Прибалтики, и на юг, на Апеннины, Пелопоннес, в Малую Азию.

Вот это – схема, которую я буду обосновывать в данной работе. Она пока представляется крайне неортодоксальной, и является по сути принципиальной ревизией легенд о R1b в Европе 30 тысяч лет назад. По моим данным, никаких R1b в Европе не было ранее 5-6 тысяч лет назад. Или, если они и были, то следов не оставили, во всяком случае в современном населении Европы.

Более того, недавно появились данные, принципиально подтверждающие, что баски гаплогруппы R1b, которых среди басков более 90%, появились в Испании и Франции никак не ранее нескольких тысяч лет назад. Оказалось, что более 90% их имеют ген лактозной толерантности (что позволяет им пить молоко без серьезного дискомфорта), который впервые появился в Азии менее 10 тысяч лет назад, и оттуда принесен в Европу. Об этом подробнее – ниже.

Субклады R1b

Но вернемся к схеме мутаций. Мы остановились на мутации M343, которая привела к образованию гаплогруппы R1b. И далее остановимся на подгруппах. Дело в том, что мутации, которые можно использовать для классификации новых гаплогрупп, происходят в нас довольно часто. В Y-хромосоме 50 миллионов нуклеотидов, и даже при скорости мутации одна на миллион в поколение в каждом поколении может появляться до 50 мутаций, которые могут быть квалифицированы как снипы (SNP = Single Nucleotide Polymorphism). Подобных мутаций в людях уже выявлено более трех миллионов (The International HapMap Consortium, 2007). Так что в принципе в каждой семье может быть выявлена мутация, которая присуща только этой семье, идущая, например, от прадедушки. Или пра-прадедушки. Это очень удобно для проведения так называемых «фамильных исследований», как серьезное подспорье в классической генеалогии. Иначе говоря, все прямые мужские потомки определенного предка, который жил, скажем, триста лет назад, несут его снип, и определенный тест это покажет. Такие линии называются «субкладами», и они нумеруются в виде продолжения индекса основной гаплогруппы.

Ясно, что для R1b таких субкладов предложено больше всего, а именно уже более тридцати. Их может быть и сотни, и тысячи, и, видимо, в будущем так и будет. Пока же создание новых субкладов – инициатива специалистов в ДНК-генеалогии, которые делают выбор мутаций и соответствующих субкладов на основании ряда причин, включая и личные предпочтения.

Так вот, следующей за R1b подгруппой, или субкладом, является R1b1, который определяется мутацией P25. Потом идет R1b1a, мутация M18. Затем идет P297/M73, субклад R1b1b, который недавно переименован в R1b1b1. Это – в основном азиатский субклад, от Турции до Сибири. Далее – R1b1b2, мутация M269 (ранее, по предыдущей номенклатуре, был R1b1c). Это – европейский вариант R1b.

И затем идет целая галерея субкладов, нисходящих от мутации M269. Они раньше именовались R1b1c1, R1b1c2, R1b1c3, и далее до R1b1c11. Теперь они именуются в развитие индекса R1b1b2, например, R1b1b2a2c, R1b1b2a2d, R1b1b2a2e, R1b1b2a2g и так далее. Каждому поставлена в соответствие мутация с определенным индексом, часто

многозначным (например, R1b1b2a2 – это rs34276300), и по параллельной системе мутации еще обозначают индексами S28, S116, а также индексами типа U106, U152, U198 и так далее, причем маркеры могут сопровождаться значками + или - . Система весьма усложнена, и вдобавок постоянно ревизуется, причем параллельно разными администраторами и разными тестирующими компаниями, которые предпочитают свои индексы.

Я вынужден это здесь описывать, поскольку несколько таких субкладов проанализировал на время жизни общего предка. А именно, следующие (идут синонимы по записи):

R1b1b2 = R1b1c = M269 = S3 = S10 = S13 = S17

R1b1b2a1 = R1b1c9 = M405 = S21 = U106+

R1b1b2a2g = R1b1c10 = S28 = U152+

Некоторые индексы дублируются, потому что номенклатура еще не устоялась, и продолжает идти путаница. В номенклатуре 2008 года (http://www.isogg.org/tree/ISOGG_HapgrpR08.html) отмечено, что субклад R1b1b2a1 (бывший R1b1b1c9) имеет маркер S21/U106. А проект http://www.familytreedna.com/public/U106, с маркером R-U106/S21+, представляет почти исключительно гаплотипы субклада R1b1b2g. В номенклатуре 2008 года субклада R1b1b2g вообще нет. И таких примеров немало.

Для нашей цели это, строго говоря, неважно. Мы рассмотрим времена общих предков гаплогруппы R1b и субкладов, как бы их не называли, и выясним, насколько это времена жизни общих предков отличаются друг от друга.

Сравнение времени жизни общих предков европейских R1a1 и R1b

Перед этим – расчет приблизительный. В базе данных YSearch имеются более 17 тысяч гаплотипов гаплогруппы R1b с подгруппами, в подавляющем большинстве европейского происхождения. Можно для начала провести расчет времени жизни общего предка для всех этих

тысяч гаплотипов, понимая, что этот расчет будет в значительной степени условным. С другой стороны, наличие такого количества гаплотипов даст хорошую статистику, обычно недостижимую другими методами. Итак, в какой степени «условным» будет расчет? И насколько гаплотипы группы R1b окажутся «моложе» или «старше» гаплотипов группы R1a1?

Гаплогруппа R1a1 – для сравнения

В предыдущей статье, посвященной гаплогруппе R1a1 (Клёсов, 2008a), мы показали, что общий предок многих гаплотипов этой гаплогруппы жил в Европе 4400-4800 лет назад, для некоторых территорий от 4200 лет назад (Чехия, Словакия, Швеция), до 5000-5200 лет назад (Венгрия, Ирландия), для некоторых регионов раньше, чем 4200 лет назад. Но в среднем по Европе этот срок составляет примерно 4500 лет назад. А что даст массивный расчет по базе данных YSearch? Это даст некоторое представление, насколько целесообразны подобные массивные расчеты.

Итак, в базе данных есть 672 шести-маркерных гаплотипа R1a1, опять в основном европейского происхождения, с базовым (предковым) гаплотипом в системе записи FTDNA (пропущенные аллели заменены символами X)

13-25-16-10-X-X-X-12-X-X-11-X

Все 672 гаплотипа имеют 1092 мутации от базового. Это дает в среднем 1092/672/6 = 0.271 мутацию на маркер, что – с поправкой на возвратные мутации – соответствует 204 поколениям до общего предка, или 5100 лет.

В той же базе данных есть 664 12-маркерных гаплотипов гаплогруппы R1a1, с базовым гаплотипом

13-25-16-10-11-14-12-12-10-13-11-30

Как видно, предковый гаплотип тот же самый, что и для 6-маркерных гаплотипов, но более полный, 12-маркерный. Все 664 гаплотипа содержат 2180 мутаций, что дает 0.274 мутации на маркер, и соответствует 179 поколениям до общего предка, то есть 4475 лет.

Среднее количество мутаций на маркер здесь больше, чем в 6-маркерных гаплотипах, но средние скорости мутации в 6-маркерных гаплотипах меньше, чем в 12-маркерных. Поэтому и расстояние до общего предка оказалось несколько меньше.

Наконец, в базе данных есть 417 25-маркерных гаплотипов R1a1, и во всех имеется 3095 мутаций от базового гаплотипа

13-25-16-10-11-14-12-12-10-13-11-30-15-9-10-11-11-24-14-20-32-12-15-15-16

Это дает 0.297 мутаций на маркер, что соответствует 198 поколениям, или 4950 годам до общего предка. Интересно, что этот предковый 25-маркерный гаплотип в точности соответствует предковому гаплотипу русских (восточных) славян, хотя их было всего 26 гаплотипов из 417 , или всего 6%, во всей базе данных. Иначе говоря, славянский предковый гаплотип в точности равен общеевропейскому, что, впрочем, показано раньше (Клёсов, 2008a) на основе другого материала, не привлекая массированные расчеты, как делаю здесь.

Видно, что все эти величины укладываются в интервал времен жизни общих предков гаплогруппы R1a1, рассчитанных по разным европейским регионам (Клёсов, 2008a). То есть подобные массированные расчеты дают разумные величины времен жизни общего предка.

Гаплогруппа R1b

Применим тот же подход для гаплогруппы R1b (с подгруппами). Для 17171 шестимаркерных гаплотипов

13-24-14-11-X-X-X-12-X-X-13-X

имеется 21027 мутаций, что дает 0.204 мутации на маркер, и это соответствует 146 поколениям, или 3650 годам до общего предка всех 17 тысяч гаплотипов R1b в базе данных. Заметим, что общий предок R1b в Европе (а именно к Европе относится подавляющее количество гаплотипов R1b в базе данных) «моложе» общего предка R1a1. Это видно и по среднему количеству мутаций на маркер (0.271 для R1a1

против 0.204 для R1b для 6-маркерных гаплотипов, 0.274 против 0.246 для 12-маркерных, и 0.297 против 0.275 для 25-маркерных, см. ниже).

Для 12-маркерных гаплотипов, которых в базе данных 16959, базовый гаплотип R1b следующий

13-24-14-11-11-14-12-12-12-13-13-29

Как видно, он опять согласуется с предковым 6-маркерным гаплотипом. Эти гаплотипы имеют суммарно 50040 мутаций, что дает 0.246 мутаций на маркер, и приводит к 158 поколениям, или 3950 годам до общего предка. Таким образом, и по 12-маркерным гаплотипам общий предок гаплогруппы R1b в Европе «моложе», чем общий предок гаплогруппы R1a1.

В базе данных имеется 12090 25-маркерных гаплотипов R1b, и их базовый (предковый) гаплотип, полученный минимизацией мутаций по каждому маркеру (Клёсов, 2008b)

13-24-14-11-11-14-12-12-12-13-13-29-17-9-10-11-11-25-15-19-29-15-15-17-17

позволил выявить 83240 мутаций по всем гаплотипам. Это дает 0.275 мутаций на маркер, и приводит к 180 поколениям, или 4500 лет до общего предка.

Обратим внимание, что для гаплотипов R1a1 6-маркерные гаплотипы дали самый высокий интервал времени до общего предка по сравнению с 12- и 25-маркерными гаплотипами, а для гаплотипов R1b – самый низкий. Это понятно, поскольку 6-маркерные гаплотипы больше других подвержены колебаниям по числу мутаций. Я и привел их здесь в основном только для того, чтобы продемонстрировать этот результат. Тем не менее, даже 6-маркерные гаплотипы дают в целом разумные результаты по оценке времен жизни общего предка. На основании более точных 12- и 25-маркерных гаплотипов можно заключить, что общий предок европейских R1b жил примерно на 500 лет позже европейских же R1a1, а именно 4200±400 лет назад.

Как видно, это совсем не 30 тысяч лет назад, как говорят нам «научные» легенды. R1b - «младший брат» R1a1 в Европе.

Но это была, как я отмечал, примерная оценка возраста общего предка европейских R1b. Рассмотрим конкретные выборки гаплотипов более внимательно.

Европейские гаплотипы R1b двух основных генеалогических линий

На рис. 1 приведено дерево 37-маркерных европейских гаплотипов гаплогруппы R1b, в количестве 104. Они были выбраны неупорядоченным способом из всего массива R1b, и все были помечены в базе данных YSearch как "R1b (tested)".

Видно, что дерево разделяется на две части, в левой, более «молодой» части 75 гаплотипов, в правой части 24 гаплотипа, и еще малая ветвь справа из пяти гаплотипов (между 241 и 267) не обрабатывалась по причине малочисленности ветви.

При рассмотрении только первых 12 маркеров все дерево имеет один предковый гаплотип:

13-24-14-11-11-14-12-12-12-13-13-29

Это именно тот гаплотип, который приведен выше и который был получен при обработке почти 17 тысяч гаплотипов R1b. Это – самый распространенный европейский гаплотип, и он называется «Атлантический модальный гаплотип», или АМГ. Все 104 12-маркерных гаплотипа имеют 268 мутаций, что соответствует 135 поколений, или 3375 лет до общего предка.

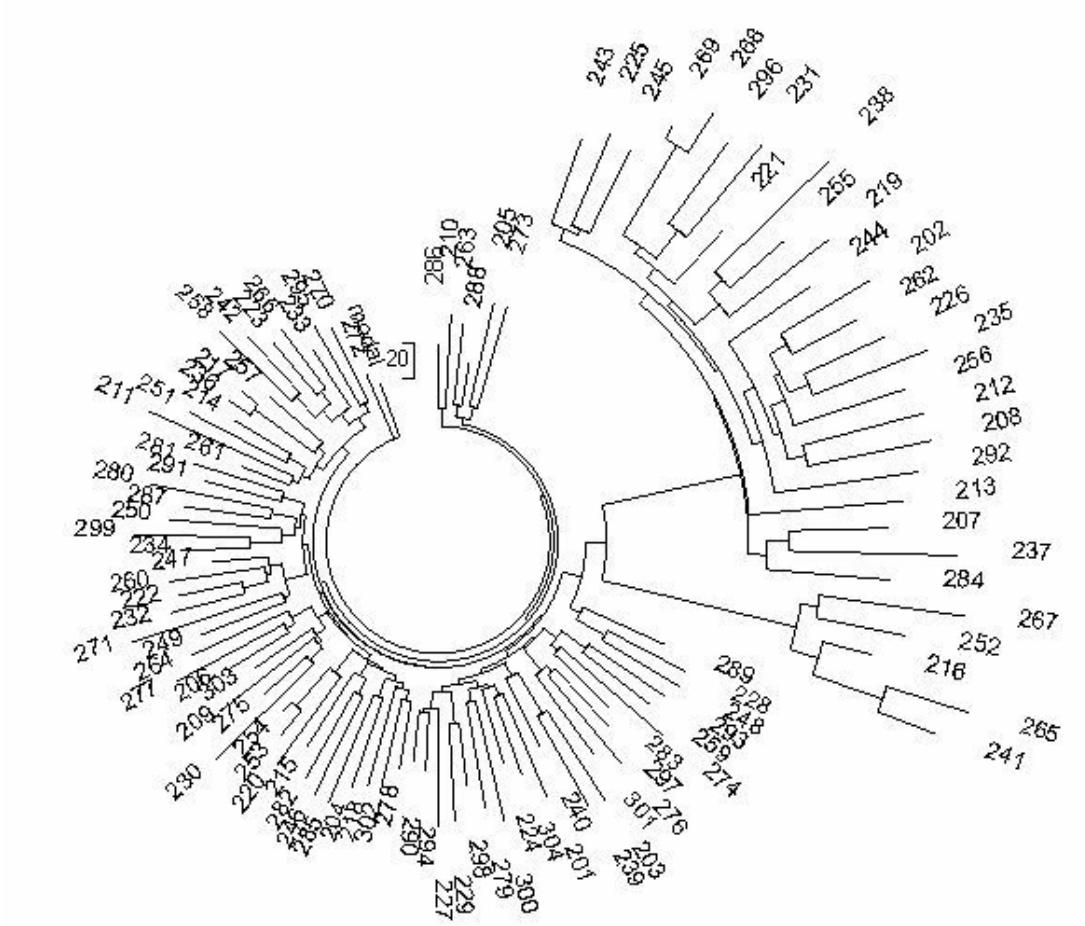


Рис. 1. Дерево 37-маркерных европейских гаплотипов R1b, неупорядоченно выбранных из базы данных YSearch (2008). В выборке – 104 гаплотипа.

При увеличении числа маркеров до 25 базовые гаплотипы начинают расходиться. Проявляются два общих предка, отстоящих по времени на 700 лет, и имевшие одинаковые 12-маркерные гаплотипы, но уже несколько разные 25-маркерные. В 25-маркерном формате их базовые гаплотипы выглядят следующим образом (мутационные различия обозначены жирным шрифтом):

13-24-14-11-11-14-12-12-12-13-13-29-17-9-**9**-11-11-25-15-19-**29**-15-15-**16**-17

13-24-14-11-11-14-12-12-12-13-13-29- 17-9-**10-**11-11-25-15-19-**30**-15-15-**17**-17

Приведенный выше 25-маркерный гаплотип, полученный из всего массива базы данных из тысяч гаплотипов – усредненный между ними.

Верхний базовый гаплотип соответствует правой верхней части дерева на рис. 1, из 24 гаплотипов, нижний базовый гаплотип – левой части, из 75 гаплотипов.

Правая часть дерева имеет 169 мутаций на 24 гаплотипа, что соответствует 185 поколениям, или 4625 годам до общего предка.

Левая часть дерева – 460 мутаций на 75 гаплотипов, что соответствует 158 поколениям, или 3950 годам до общего предка.

Как видно, эти данные близки к тем, которые были получены выше при массированном анализе более 10 тысяч гаплотипов базы данных. Как обычно и получается, результат мало зависит от величины выборки, когда выборка достигает даже небольших величин, зачастую 20-50 гаплотипов.

Разница во времени между общими предками этих двух 25-маркерных гаплотипов составляет 700 лет. Это соответствует примерно 1.2 мутациям разницы, если один общий предок является прямым потомком другого. Для показанных выше двух гаплотипов эта разница составляет 0.90+0.15+0.06 = 1.11. Совпадение прекрасное.

Более «молодые» гаплотипы (на рис. 1 слева) происходят от субклада «старшего». «Старший» - гаплотипы на рис. 1 справа - это продукт гаплогруппы R1b или R1b1, или R1b1b, или R1b1b2 (R1b1c). «Младший» субклад – это соответствующий нисходящий, как показано на схеме выше.

Итак, самый древний общий предок R1b по этим данным прибыл в Европу примерно 4600 лет назад. В это же время род R1a1 уже осваивал новые территории, тоже по всей Европе, расходясь от своей балканской «прародины», где род R1a1 жил уже более семи тысяч лет.

Так что это за молодой субклад, нисходящий от R1b, с общим предком, жившим в Европе примерно 4000 лет назад? Попытаемся его определить.

Время жизни общих предков субкладов гаплогруппы R1b1b2

Субклад R1b1b2a1/U106/S21

В сети есть собрание гаплотипов R1b/U106/S21 (http://www.familytreedna.com/public/U106), с подтвержденной мутацией U106+ и негативной U198-. Это – субклад R1b1b2a1.

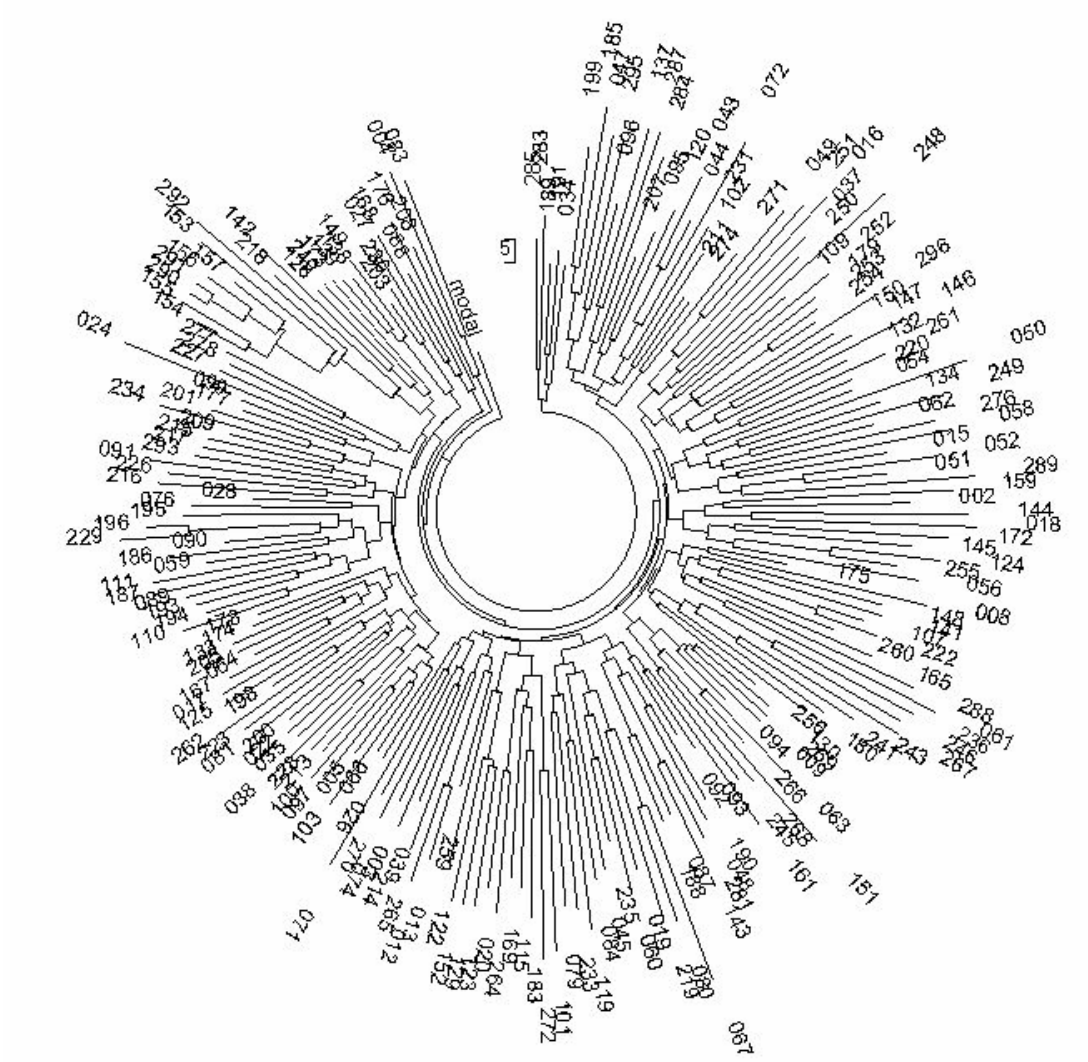


Рис. 2. Дерево 67-маркерных европейских гаплотипов R1b-U106/S21 (R1b1b2a1, ранее R1b1c9), построенное по данным проекта (http://www.familytreedna.com/public/U106 **(2008). В выборке – 207 гаплотипов.**

В этом собрании 284 полных 25-маркерных гаплотипа (и выше по числу маркеров), из них подавляющая часть гаплотипов с британских

островов и из США, потомков жителей британских островов. Их в собрании три четверти всех гаплотипов. Остальные, примерно по десятку–полтора – из разных европейских регионов – Центральная Европа (9 гаплотипов), северо-восток Европы (12 гаплотипов), северо-западная Европа (22 гаплотипа), Скандинавия (14 гаплотипов), юго-восточная и юго-западная Европа (по три гаплотипа), и ряд гаплотипов без конкретного отнесения. Дерево 67-маркерных гаплотипов, построенное по этим данных, приведено на рис. 2.

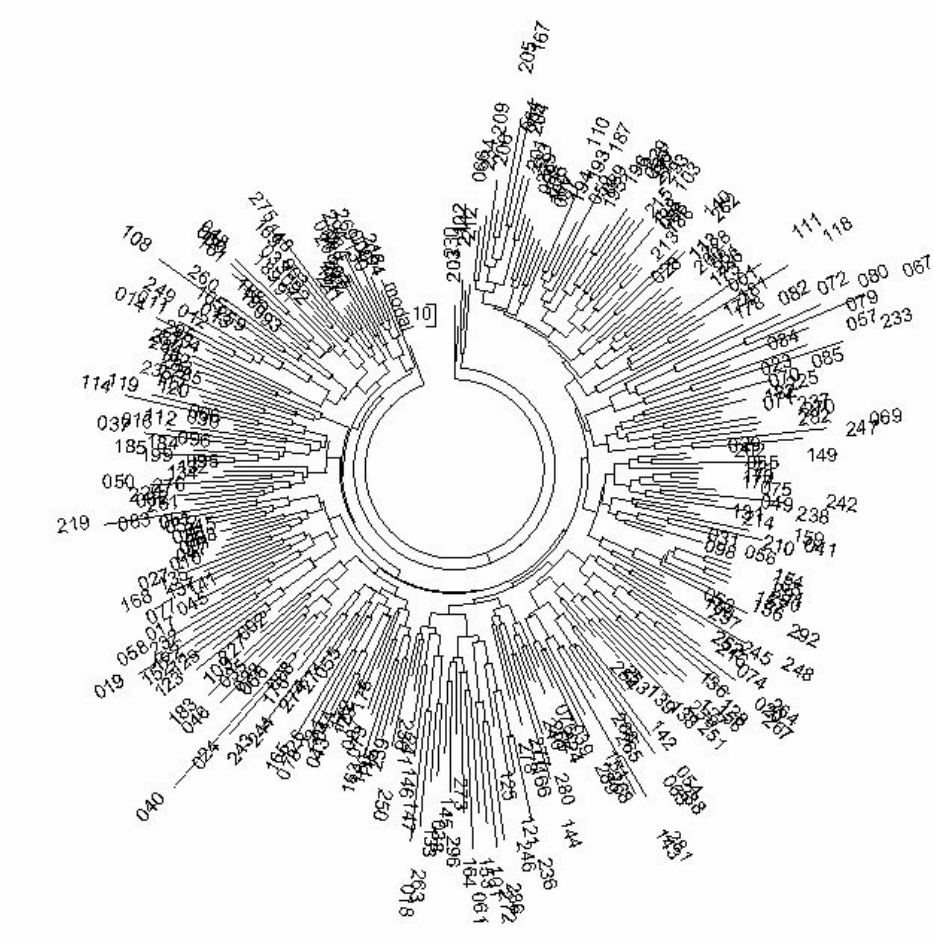


Рис. 3. Дерево 25-маркерных европейских гаплотипов R1b-U106/S21 (R1b1b2a1, ранее R1b1c9), построенное по данным проекта (http://www.familytreedna.com/public/U106 **(2008). В выборке – 284 гаплотипа.**

Видно, что для всей выборки имеется всего один общий предок. Дерево не расходится по выраженным ветвям, как это было на рис. 1, и является практически симметричным.

Естественно, что дерево с меньшим разрешением, а именно дерево 25-маркерных гаплотипов, тоже указывает на одного общего предка (рис. 3).

По данным дерева на рис. 3 было рассчитано, что базовый (предковый) гаплотип субклада R1b-U106/S21 – следующий:

13-**23**-14-11-11-14-12-12-12-13-13-29- 17-9-10-11-11-25-15-19-29-15-15-17-17

Он всего лишь на одну мутацию (отмечено жирным шрифтом) отличается от обобщенного предкового гаплотипа, полученного выше при массивной обработке базы данных YSearch, на три мутации от предкового гаплотипа материнской гаплогруппы (рис. 1 справа), и на две мутации от предкового гаплотипа «молодого» субклада на рис. 1 (слева).

Все 284 25-маркерных гаплотипа содержат 1853 мутации от этого предкового гаплотипа, что соответствует 0.261 мутации на маркер, и транслируется в 4225 лет до общего предка всех 284 носителей этих гаплотипов.

Те же 12-маркерные гаплотипы содержат 729 мутаций от базового гаплотипа, что соответствует 0.214 мутациям на маркер, что дает 3375 лет до общего предка. Это – с поправкой на возвратные мутации. Без поправки «возраст» общего предка составил бы 729/284/0.022 = 117 поколений, или 2925 лет. Поправка увеличивает этот возраст на 450 лет, до 3375 лет.

900 лет – довольно большая разница в «возрасте» предков, полученным по 12- и 25-маркерным гаплотипам. Дополнительная проверка основывалась на числе идентичных гаплотипов во всей выборке, которые и соответствуют предковым (базовым) гаплотипам. Таких среди 284 гаплотипов оказалось 21. Общий предок жил ln(284/21)/0.022 = 118 поколений назад (без поправки на возвратные мутации), то есть практически идентично величине 117 поколений, полученной по мутациям. С поправкой это даст те же 3375 лет.

Действительно, расхождение результатов расчетов между 12- и 25-маркерными гаплотипами наблюдалось нами и ранее. Но часто

расхождения нет, и данные практически идентичны для двух серий. Это означает, что проблема не в величинах используемых констант скоростей мутаций, но в особенностях конкретных серий гаплотипов – либо 12-, либо 25-маркерных. Эта проблема еще не решена в ДНК-генеалогии. Можно, конечно, начинать отбрасывать некоторые маркеры, но этим будет внесен произвол в расчеты. В данном случае проблемы особой нет, поскольку мы решаем принципиальный вопрос – когда R1b пришли в Европу – в каменном веке, 30 тысяч лет назад, или в бронзовом, 3-5 тысяч лет назад. Пока данные говорят о втором варианте, причем с большой степенью достоверности.

Субклад R1b1b2a1/U106/S21, с нулевым значением DYS425

Подборка гаплотипов субклада R1b-U106/S21 позволила проанализировать интересное явление. 22 гаплотипа из 306 имеют нулевое значение аллели в DYS425 (48-й по счету маркер в 67-маркерном гаплотипе). Это – результат так называемого эффекта RecLOH (Recombinational Loss of Heterozygosity), когда информация в локусе Y-хромосомы (в данном случае) при переписывании теряется. Поскольку это нулевое значение, раз приобретенное, передается по наследству, то интересно было узнать время жизни общего предка данных 22 человек, а значит – время, когда это обнуливание аллели произошло.

На рис. 4 показано дерево гаплотипов этих 22 человек. Дерево симметричное, и соответствует одному общему предку. Время жизни предка рассчитывалось по серии 25-маркерных гаплотипов носителей гаплотипов с нулевым значением. Предковый гаплотип их на первых 25 маркерах отличался на одну мутацию (выделено жирным шрифтом) от предкового гаплотипа для остальных объектов выборки:

13-23-14-11-11-14-12-12-12-13-13-29- 17-9-10**-**11-11-**24**-15-19-29-15-15-17-17

Иначе говоря, нулевая мутация в DYS425 пошла от человека, у которого уже случилась мутация в маркере DYS447 от обычной аллели 25 к 24.

Все 22 гаплотипа имели 54 мутации на 12-маркерных гаплотипах, и 118 мутаций на 25-маркерных гаплотипах. Это дает соответственно 128 и 135 поколений до общего предка, или 3200 и 3375 лет, соответственно.

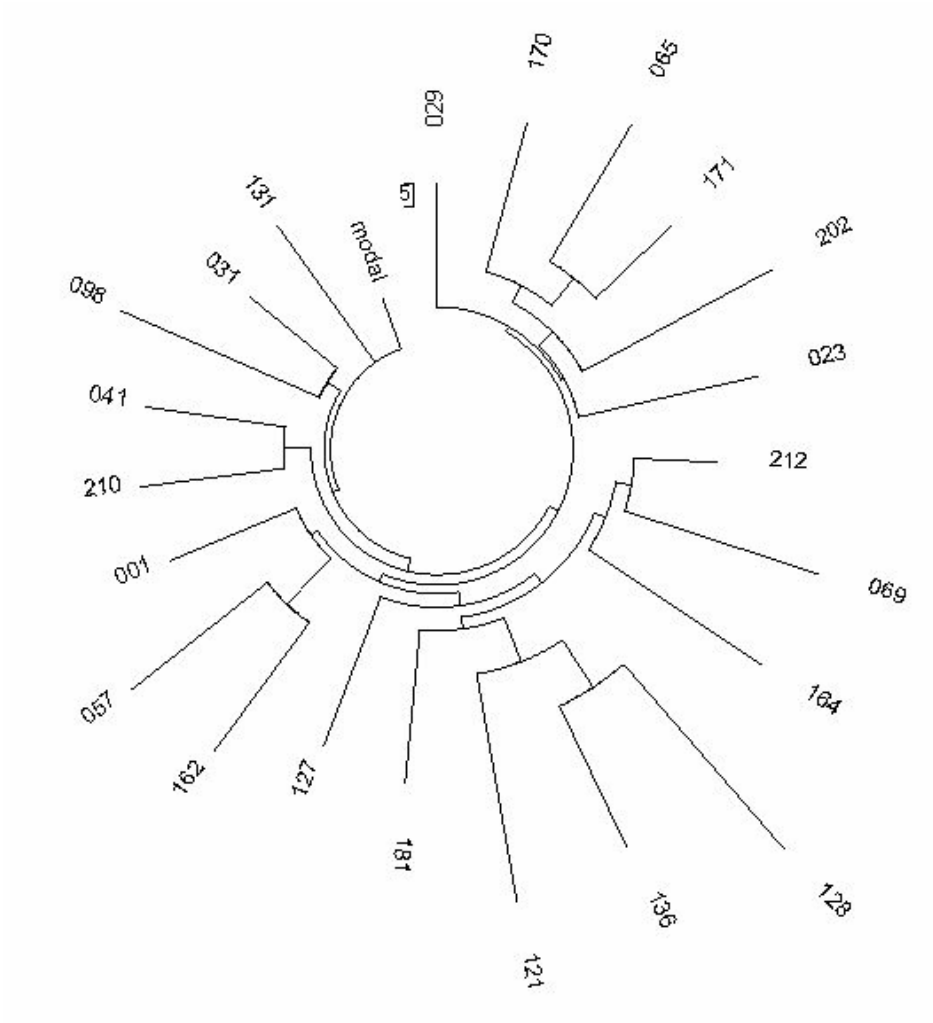


Рис. 4. Дерево 67-маркерных европейских гаплотипов R1b-U106/S21 (R1b1b2a1, ранее R1b1c9), построенное по данным проекта (http://www.familytreedna.com/public/U106 **(2008). В выборке – 22 гаплотипа, каждый из которых имеет нулевое значение аллели в локусе DYS425.**

Здесь уже разницы между двумя периодами времени по данным 12- и 25-маркерных гаплотипов почти нет, всего 7 поколений. Как и ожидалось, мутация произошла после истечения определенного времени от общего предка субклада R1b-U106/S21. Это время составляет 175 лет спустя по данным 12-маркерных гаплотипов, и 850 лет спустя по данным 25-маркерных гаплотипов. Выбор между этими

временами на данном этапе наших познаний сделать трудно. Видимо, это вообще первый расчет такого рода.

Субклад R1b1b2a2g/U152/S28

Переходим еще к одному субкладу, R1b1c10/R1b1b2h/R1b1b2a2g (http://www.davidkfaux.org/R1b1c10_Data.htm), он же субклад S28/U152. Это – субклад, параллельный тому, что представлен на рис. 3. Они исходят из одного «материнского» субклада R1b1c, но на рис. 3 представлен c9, а на рис. 5 – c10.

Среди 184 гаплотипов субклада, представленных на рис. 5, есть восемь идентичных 12-маркерных гаплотипов. Согласно принципам ДНК-генеалогии, их рассматривают как базовые, или предковые гаплотипы, наиболее сохранившиеся, в которых со времени жизни общего предка ни произошло ни одной мутации. Они уже могут быть использованы для расчета времени до общего предка (Клёсов, 2008b) по формуле ln(184/8)/0.022 = 143 поколения (без поправки на возвратные мутации). В тех же 184 12-маркерных гаплотипов имеется 568 мутаций, что дает 568/184/0.022 = 140 поколений (также без поправки). То, что число поколений практически одинаково по двум совершенно разным способам расчета показывает, что данный субклад действительно имеет одного общего предка. Об этом же свидетельствует и вид дерева гаплотипов на рис. 5, которое не содержит выраженных ветвей.

С поправкой на возвратные мутации для 12-маркерных гаплотипов число поколения до общего предка отодвигается от 140-143 до 167 поколений, что соответствует 4175 годам. 184 25-маркерных гаплотипов содержат 1236 мутаций, что соответствует 175 поколениям, то есть 4375 лет до общего предка.

Как видно, в этом случае время жизни общего предка, рассчитанное по 12- и 25-маркерным гаплотипам почти одинаково, а именно 4175 и 4375 лет.

Из этого можно сделать вывод, что по 25-маркерным гаплотипам субклады R1b1c9 (U106) и R1b1c10 (U152) образовались практически в одно и то же время, а именно 4225 и 4375 лет назад, соответственно.

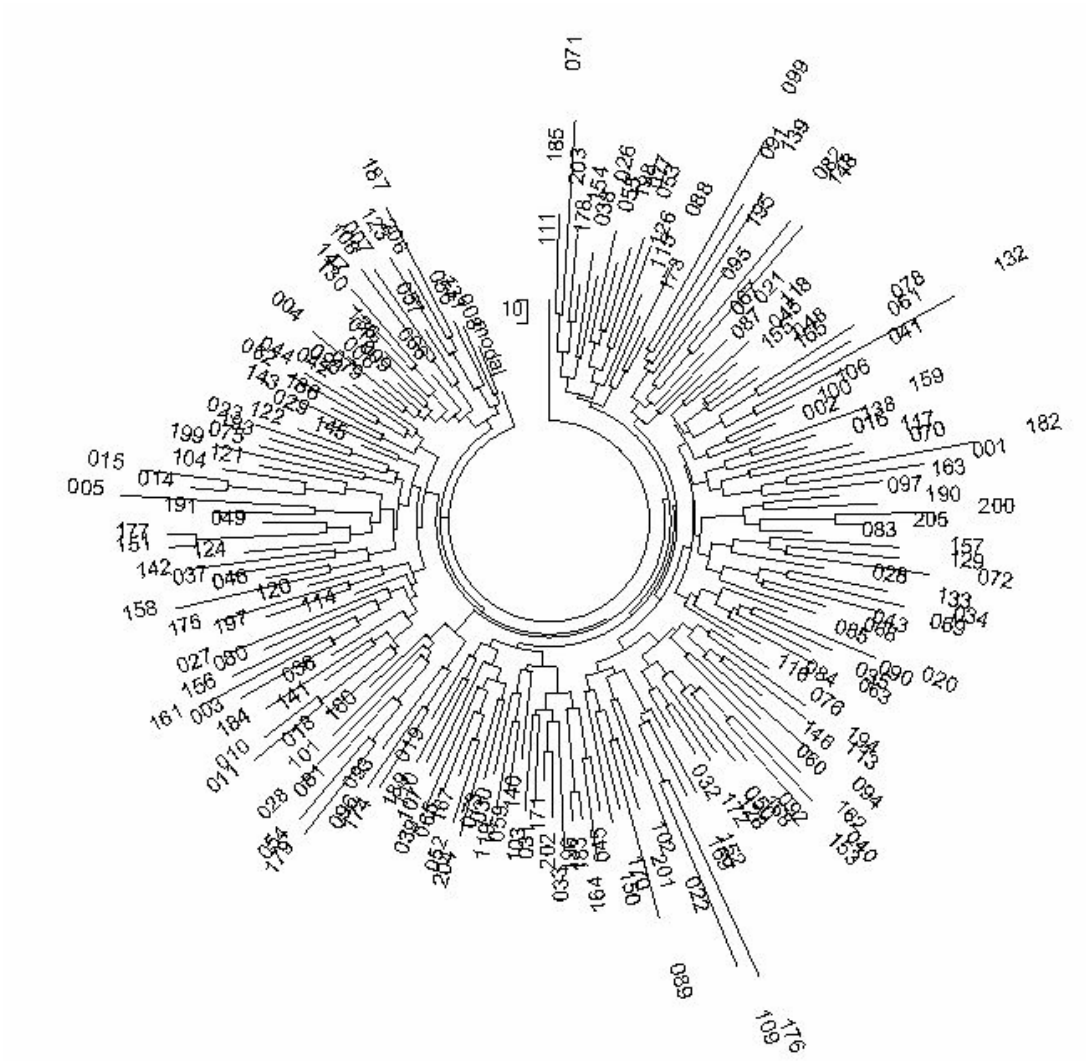


Рис. 5. Дерево 25-маркерных европейских гаплотипов R1b1b2a2g/S28/U152 (ранее R1b1c10), построенное по данным проекта (http://www.davidkfaux.org/R1b1c10_Data.htm**) (2008). В выборке – 184 гаплотипа.**

25-маркерные базовые (предковые) гаплотипы этих субкладов выглядят следующим образом:

13-**23**-14-11-11-14-12-12-12-13-13-29- 17-9-10**-**11-11-**24**-15-19-29-15-15-17-17

13-**24**-14-11-11-14-12-12-12-13-13-29- 17-9-10**-**11-11-**25**-15-19-29-15-15-17-17

Они оба отличаются от предковых гаплотипов для «материнской» и «дочерней» гаплогрупп на рис. 1:

13-**24**-14-11-11-14-12-12-12-13-13-29- 17-9-**10-**11-11-**25-**15-19-**30**-15-15-**17**-17

13-**24**-14-11-11-14-12-12-12-13-13-29-17 -9- **9**-11-11-**25-**15-19-**29**-15-15-**16**-17

Эти отличия невелики, и не превышают тысячи лет между временами жизни общих предков.

Субклад R1b1b2/M269

Еще один относительно изученный субклад – это R1b1b2 (ранее R1b1c, M269) (http://www.familytreedna.com/public/atlantic-r1b1c&fixed_columns=on). На самом деле к этому субкладу относится целый ряд производных субкладов, например, R1b1b2a2c (M153), R1b1b2a2d (M167), R1b1b2a2e (M222), R1b1b2a2g (U152), и остальная часть группы R1b1b2a2, но, как я отмечал выше, каждый субклад состоит из других субкладов, от древности до настоящего времени. И у каждого есть свой общий предок.

В данной выборке подавляющая часть гаплотипов, 83%, относится к субкладу R1b1b2 (M269), остальные – R1b1b2d, R1b1b2e и R1b1b2h, они же R1b1b2a2d, R1b1b2a2e и R1b1b2a2g.

Дерево 67-маркерных гаплотипов этого набора гаплотипов приведено на рис. 6. Видно, что у этой выборки гаплотипов – один общий предок.

Расчет времени жизни общего предка проводился по 25-маркерным гаплотипам (рис. 7).

На все 197 гаплотипов приходится 545 и 1340 мутаций, для 12- и 25-маркерных гаплотипов, соответственно. Это дает 147 и 178 поколений, или 3675 и 4450 лет, соответственно.

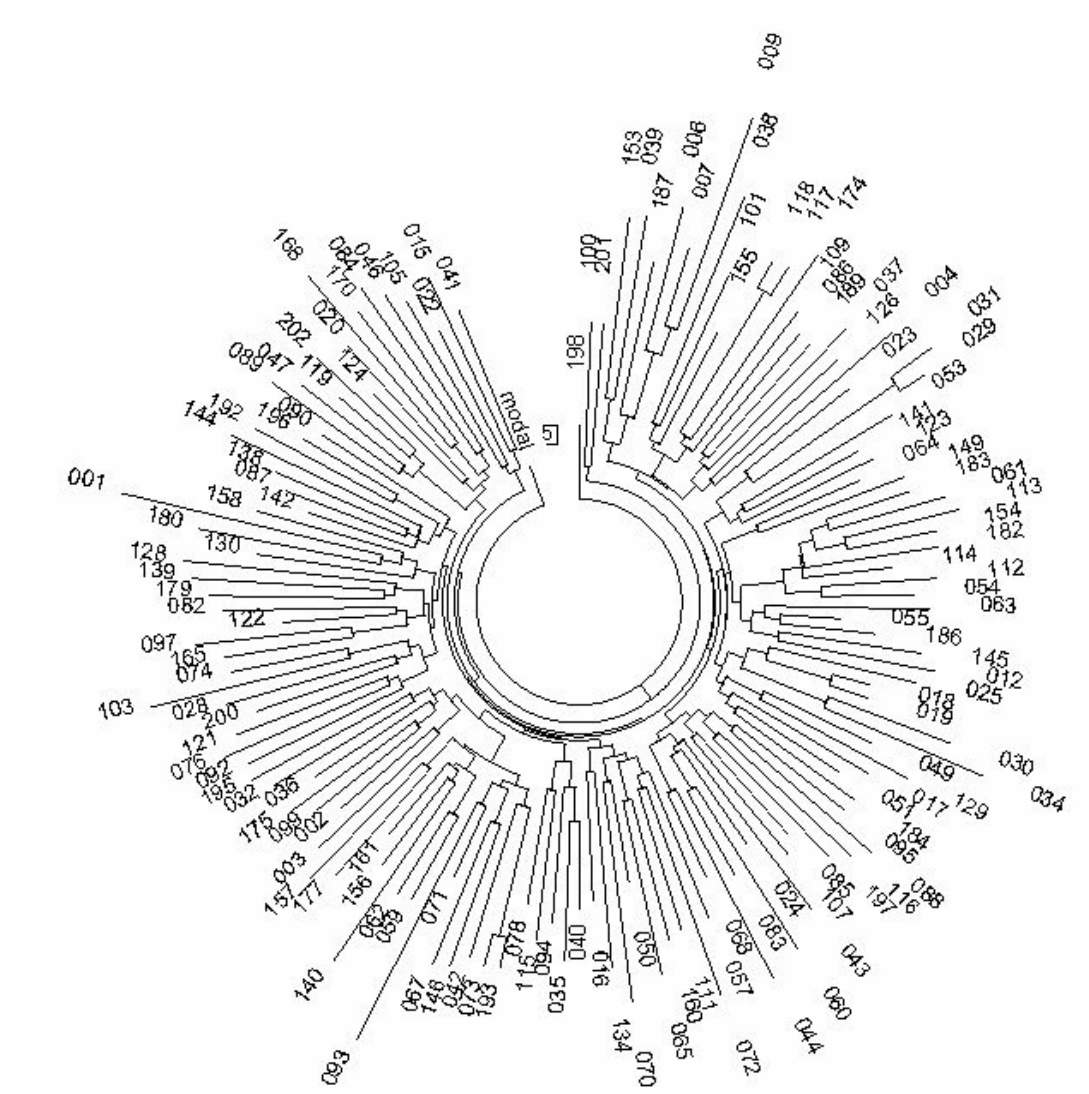


Рис. 6. Дерево 67-маркерных европейских гаплотипов R1b1b2 (ранее R1b1c, построенное по данным проекта (http://www.familytreedna.com/public/atlantic-r1b1c&fixed_columns=on **(2008). В выборке – 138 гаплотипов.**

То, что у выборки на рис. 7 имеется только один общий предок, можно показать и следующим образом. Среди 197 12-маркерных гаплотипов имеется 11 идентичных, что и соответствует базовому (предковому) гаплотипу. Число поколений до общего предка (без поправки на возвратные мутации) в этом случае равно ln(197/11)/0.022 = 131±4, где погрешность величины определяется возможным случайным отклонением числа базовых гаплотипов до 10 или 12.

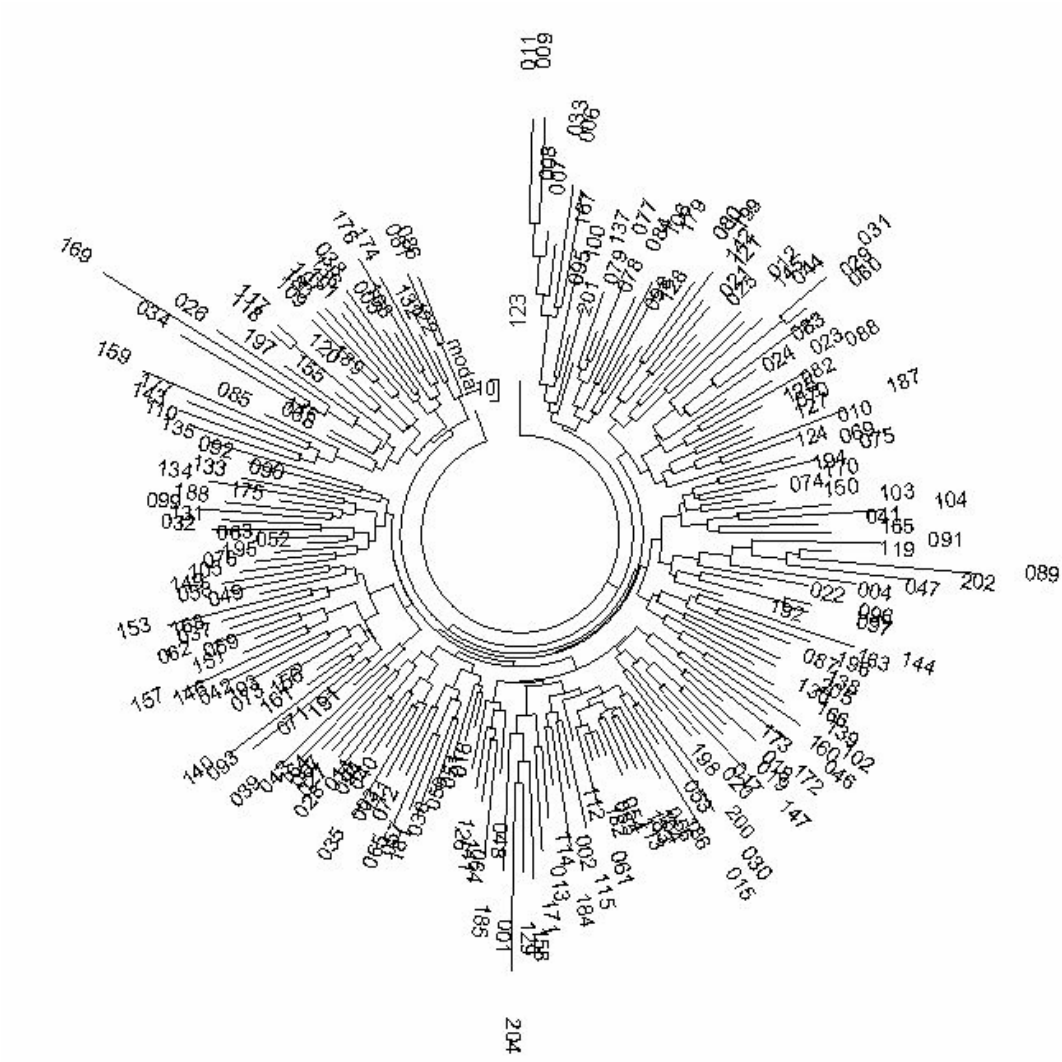


Рис. 7. Дерево 25-маркерных европейских гаплотипов R1b1b2 (ранее R1b1c), построенное по данным проекта (http://www.familytreedna.com/public/atlantic-r1b1c&fixed_columns=on **(2008). В выборке – 197 гаплотипов.**

В то же время расчет по мутациям (без поправки на возвратные мутации) дает 545/197/0.022 = 126 поколений до общего предка. Видно, что эти величины, полученные двумя принципиально разными способами (по базовым гаплотипам - логарифмам и по мутациям) практически одинаковы в пределах погрешностей расчета. Значит, у выборки на рис. 7 имеется только один общий предок (Клёсов, 2008b).

Базовый (предковый) гаплотип субклада R1b1b2 (R1b1c) (рис. 7) следующий:

13-24-14-11-11-14-12-12-12-13-13-29- 17-9-10-11-11-25-15-19-29-15-15-17-17

Он в точности равен предковому гаплотипу субклада R1b1b1ag2/S28/U152/R1b1c10 (рис. 5), да и времена жизни общих предков у них почти одинаковы – 4450 и 4375 лет назад, соответственно. От предкового гаплотипа «материнской» гаплогруппы R1b (рис. 1, ветвь справа) их отличает только одна мутация (выделено жирным шрифтом), а на самом деле доля единицы:

13-24-14-11-11-14-12-12-12-13-13-29- 17-9-10-11-11-25-15-19-**30**-15-15-17-17

Иначе говоря, все эти предковые гаплотипы являются близкородственными, и общие предки жили в пределах нескольких столетий, или даже нескольких поколений, примерно 4200-4600 лет назад. Ни о каких 30 тысяч лет назад, ни о каких кроманьоньцах не может быть и речи.

Поскольку баски, гаплогруппа подавляющего большинства которых R1b, часто приводятся в качестве древних представителей R1b, рассмотрим их гаплотипы.

Гаплотипы R1b у басков

В сети есть «Проект басков» (Basque DNA Project – http://www.familitreedna.com/public/BasqueDNA), в котором приведены 12-маркерные гаплотипы R1b (44 гаплотипа), 25-маркерные (17 гаплотипов) и меньшее количество 37- и 67-маркерных гаплотипов. Среди 12-маркерных гаплотипов нет базовых, и считать мутации по отношению к расчетному базовому гаплотипу нецелесообразно, не зная, происходит ли вся выборка от одного общего предка, или от нескольких. В любом случае, поскольку все 44 12-маркерных гаплотипа содержат 140 мутаций, предварительная оценка дает 173 поколения, или 4325 лет до общего предка. Но эту величину надо проверять.

Поэтому было построено дерево 25-маркерных гаплотипов (рис. 8). Видно, что дерево разделяется на две основные ветви, и та, что внизу – более старая ветвь, она отходит от ствола дерева дальше, чем верхняя, более молодая ветвь.

Базовые гаплотипы этих двух основных ветвей в 12-маркерном формате следующие (старая и молодая ветви, соответственно):

13-24-14-11-11-14-12-12-12-13-13-29

13-24-14-11-11-14-12-12-12-**14**-13-**30**

В 25-маркерном формате более старая баскская ветвь имеет следующий предковый гаплотип:

13-24-14-11-11-14-12-12-12-13-13-29-18-9-10-11-11-25-15-19-28-15-15-17-17

Этот гаплотип на две мутации отличается от предковых гаплотипов субкладов R1b1b2/R1b1c/M269 (рис. 7) и R1b1b2a2g/S28/U152/R1b1c10 (рис. 5) [отмечено жирным шрифтом]

13-24-14-11-11-14-12-12-12-13-13-29- **17**-9-10-11-11-25-15-19-**29**-15-15-17-17

и на три и четыре мутации, соответственно, от предковых гаплотипов «материнской» и «дочерней» ветви европейских гаплотипов на рис. 1.

13-24-14-11-11-14-12-12-12-13-13-29- **17**-9-10-11-11-25-15-19-**30**-15-15-17-17

13-24-14-11-11-14-12-12-12-13-13-29-**17** -9- **9-**11-11-25-15-19-**29**-15-15-**16**-17

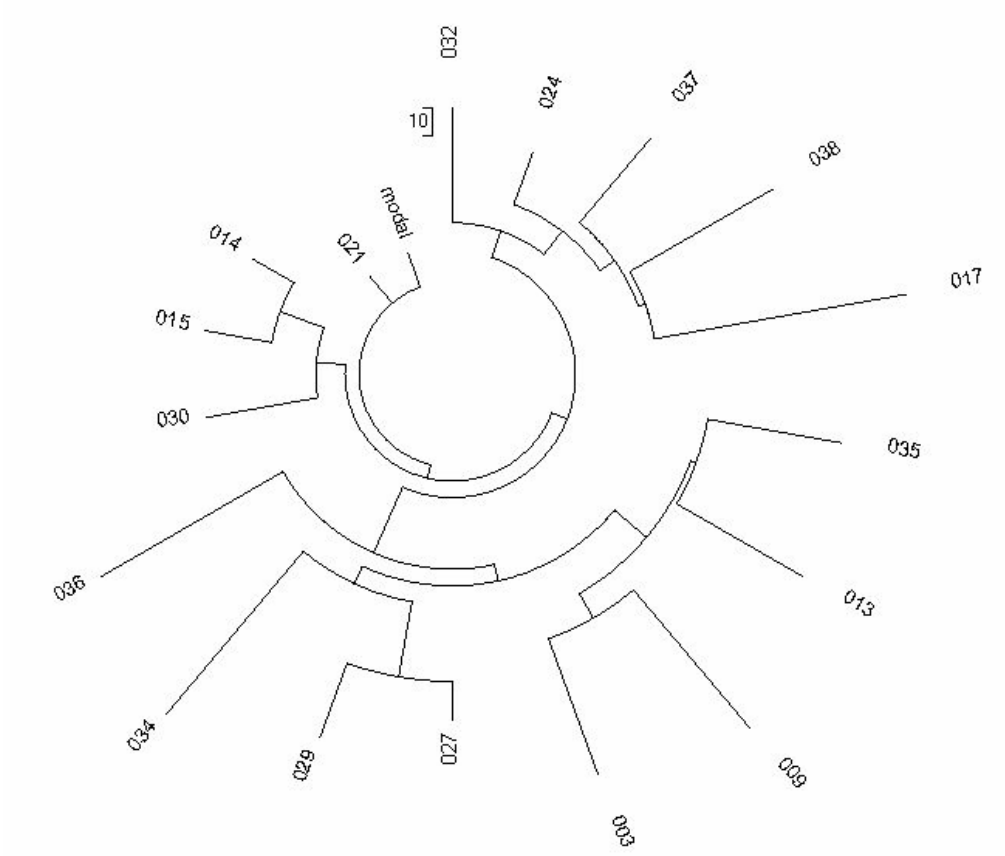


Рис. 8. Дерево 25-маркерных гаплотипов R1b басков, по данным «баскского проекта». В выборке – 17 гаплотипов.

Восемь гаплотипов более старой (нижней) ветви содержат 26 мутаций в первых 12 маркерах, что дает 177 поколений до общего предка. Все 25 маркеров этой ветви содержат 50 мутаций, что дает 161 поколений. Совпадение относительно неплохое. Таким образом, общий предок басков гаплогруппы R1b жил 4425 или 4025 лет назад, по расчетам по 12- и 25-маркерным гаплотипам.

Вторая, более молодая ветвь, содержит только пять гаплотипов, и ее базовый гаплотип следующий (отмечены мутации по сравнению со старой ветвью):

13-24-14-11-11-14-12-12-12-**14**-13-**30**-**17**-9-10-11-11-25-**14-18-29**-15-15-**16**-17

Первые 12 маркеров включают только 7 мутаций, что соответствует примерно 1725 годам до общего предка. Полная последовательность из 25 маркеров имеет 17 мутаций, что соответствует 1675 лет до общего предка. Поэтому можно заключить, что общий предок для более молодой ветви жил примерно 1700±50 лет назад, в 4-м веке нашей эры.

Заметим, что эти два 25-маркерных гаплотипа различаются на 7 мутаций (фактически, без округления, на 5.65 мутаций), что соответствует 143 поколениям, или 3575 лет между двумя общими предками. Это означает, что они не могли произойти друг от друга. Они – две различные линии. ИХ общий предок жил примерно 4600 лет назад. Подобную же цифру мы нашли раньше, анализируя европейские гаплотипы гаплогруппы R1b, и заключив, что общий предок их жил на европейском континенте 4,600±600 лет назад. Эту же цифру дает и анализ общего предка гаплогруппы R1b в Европе (рис. 1).

Таким образом, баски гаплогруппы R1b происходят от того же европейского предка, что и практически все европейские генеалогические линии R1b.

Ген лактозной толерантности у басков

В связи с датировкой возникновения гаплогруппы R1b среди басков, представляет интерес обнаружения у басков гена лактозной толерантности (или гена лактазной персистентности). Эти оба названия означают по сути одно и то же, и относятся к гену T-13910. Лактоза – это дисахарид, основной углевод в молоке, который метаболизируется на первой стадии под действием фермента лактазы. В биохимии окончание «-оза» соответствует сахарам, углеводам (что суть синонимы), а «-аза» - ферментам, биологическим катализаторам. Лактаза гидролизует, расщепляет лактозу на глюкозу и галактозу, которые легко усваиваются организмом. Если гена T-13910 в организме нет, или он подавлен, фермент лактаза в организме не образуется, углевод лактоза не усваивается, а вместо этого приводит к резкому дискомфорту у тех, кто пьет молоко. Лактоза тогда разрушается не ферментом, а микрофлорой кишечника с образованием большого количества газов (поскольку лактозы в молоке много), что приводит к тому, что «живот пучит». Природа сделала так, что лактаза образуется только при рождении младенцев, и этот фермент активен только до

достижения младенцем года-двух, после чего ген дезактивируется, перестает работать.

Но в ходе развитии молочного животноводства у древних народов, у кого-то в генах произошла такая случайная мутация, которая привела к появлению активного гена Т-13910, причем активного всю жизнь его носителя. Естественно, это дало преимущества мутированной особи в отношении естественного отбора, его носители пили молоко без вреда для организма, и мутация закрепилась в последующих поколениях. Исследования этого гена привели к региону его первоначального появления – либо в западном Предуралье, 6600-4800 лет назад (Anthony, 2007), либо на Кавказе или к северу от Кавказа, между 5 и 12 тысяч лет назад, где наблюдается наиболее высокое разнообразие мутаций соответствующего сегмента ДНК. Значительное разнообразие мутаций этого гена было выявлено также в Азии, на восточной стороне Уральских гор, а также в Средней Азии.

Недавние исследования показали, что этот ген лактозной устойчивости находится у 92% басков (Enattah et al, 2007). Более того, этот же ген распространен в Европе и западной Азии, он обусловлен единичной мутацией, и эта мутация является уникальной и окружена в соответствующем фрагменте ДНК характерной «подписью» нуклеотидов. Так что ген один и тот же и в Европе, и в Азии.

Эти данные согласуются с теми, что современные баски являются потомками мигрантов в составе рода R1b в Испанию и Францию 4300-4600 лет назад, или в любом случае не раньше 5-12 тысяч лет назад, из Азии или с Кавказа. Более того, изучение ДНК останков древних европейцев показало, что ген Т-13910 не наблюдался в них ранее, чем 7000-7800 лет назад. Это все показывает, что современные баски отнюдь не происходят от народов, населявших Европу 30 тысяч лет назад. Во всяком случае, не от кроманьонцев.

Поскольку этот ген расположен не в Y-хромосоме, то он передавался через женщин носителям других мужских гаплогрупп, и таким образом распространился по разным гаплогруппам. В отношении исходного региона возникновения гена Т-13910 исследователи склоняются к двум регионам – Средней Азии, и территории к северу от Кавказа и западу от Урала, то есть южным степям России. Оба региона могут иметь

отношение к возникновению гаплогруппы (мутации) R1b или быть на путях миграции рода R1b1 на Кавказ и в Европу.

Гаплотипы R1b1b2 у жителей Буковины, северные Карпаты

Буковина – старое название местности на северо-востоке Карпат, на стыке Украины и Румынии, со стороны Украины – Черновицкая область. Город Черновцы и есть исторический центр Буковины. В рамках археологии – часть территории трипольской культуры. В недавней работе (Клёсов, 2008а) показано, что носители гаплогруппы R1a1 появились там около 6000 лет назад, причем те носители, потомки которых живут в наше время. Точнее, по 25-маркерным гаплотипам время жизни общего предка R1a1 было 5975 лет назад, по менее точным 12-маркерным гаплотипам – 4675 лет назад.

Посмотрим на гаплотипы Буковины гаплогруппы R1b1b2. Дерево 12-маркерных гаплотипов показано на рис. 9.

Во всех 19 гаплотипах в 12-маркерном формате содержится 72 мутации, что дает 213 поколений, то есть 5325 лет до общего предка. В девяти 25-маркерных гаплотипах содержится 64 мутации, что дает до общего предка 187 поколений, или 4675 лет. Таким образом, по сравнению с гаплотипами R1a1 из того же региона 25-маркерные гаплотипы R1b «моложе» на 1300 лет, и 12-маркерные гаплотипы «старше» на 650 лет. В целом R1b гаплотипы Буковины старше центрально- и западно-европейских, что согласуется с их миграцией в Европу с востока.

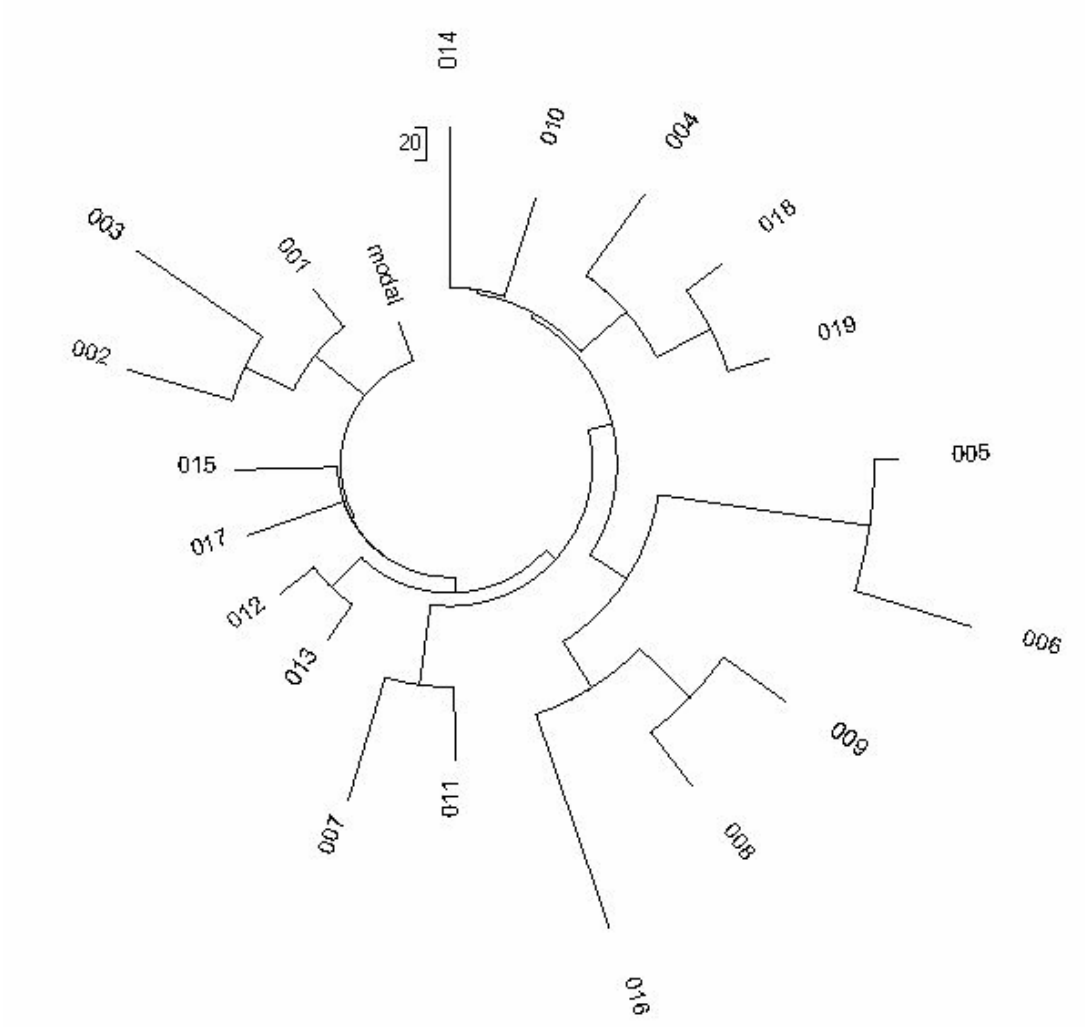


Рис. 9. Дерево 12-маркерных гаплотипов R1b1b2, по данным проекта «Буковина»». В выборке – 19 гаплотипов.

Гаплотипы R1b1b2-M269 на Сардинии

8-маркерные гаплотипы жителей этого средиземноморского острова недавно были опубликованы (Contu et al, 2008). В статье отмечалось, что остров был населен уже в неолитические времена, и во времена бронзового века численность населения стала возрастать, достигнув 300 тысяч человек 3600-2200 лет назад. Естественно, состав того населения по гаплогруппам неизвестен, но может быть хотя бы примерно реконструирован по гаплогруппам современников. Из 376

тестированных человек 64 (17%) оказались принадлежащими к гаплогруппе R1b1b2 (снипы M89, M9, M173, M269). В другой выборке тестировали 930 человек, и 158 (те же 17%) из них оказались с гаплогруппой R1b1b2. Авторы применили некритические методы расчетов (с использованием «коэффициента Животовского», упомянутого выше) и получили, что общий предок носителей этой гаплогруппы на Сардинии жил 22800 лет назад. Это, конечно, фантазийная цифра, но какая ближе к действительности?

Для ответа на этот вопрос были построены деревья 8-маркерных гаплотипов (рис. 10 и 11). Дерево на рис. 10 более древнее, в него входят гаплотипы «родительских» гаплогрупп K-R и R1, а также гаплогруппы R1b1a.

10 гаплотипов группы K-R имеют 50 мутаций от расчетного предкового гаплотипа

13-23-15-10-14-16-X-Y-Z-13-W-30

что соответствует 424 поколениям, или 10600 лет до общего предка.

Четыре гаплотипа гаплогруппы R1 принадлежали, видимо, довольно близким родственникам, так как они практически не отличались друг от друга

13-23-15-11-14-14-X-Y-Z-14-W-31

На все 32 маркера была всего лишь одна мутация, что помещает их общего предка в пределах 15 поколений, или около 400 лет от настоящего времени. Но разница в предковых гаплотипах K-R и R1, приведенных выше, помещает ИХ общего предка на 10800 лет назад, то есть почти совпадает с общим предком для сводной гаплогруппы K-R на Сардинии.

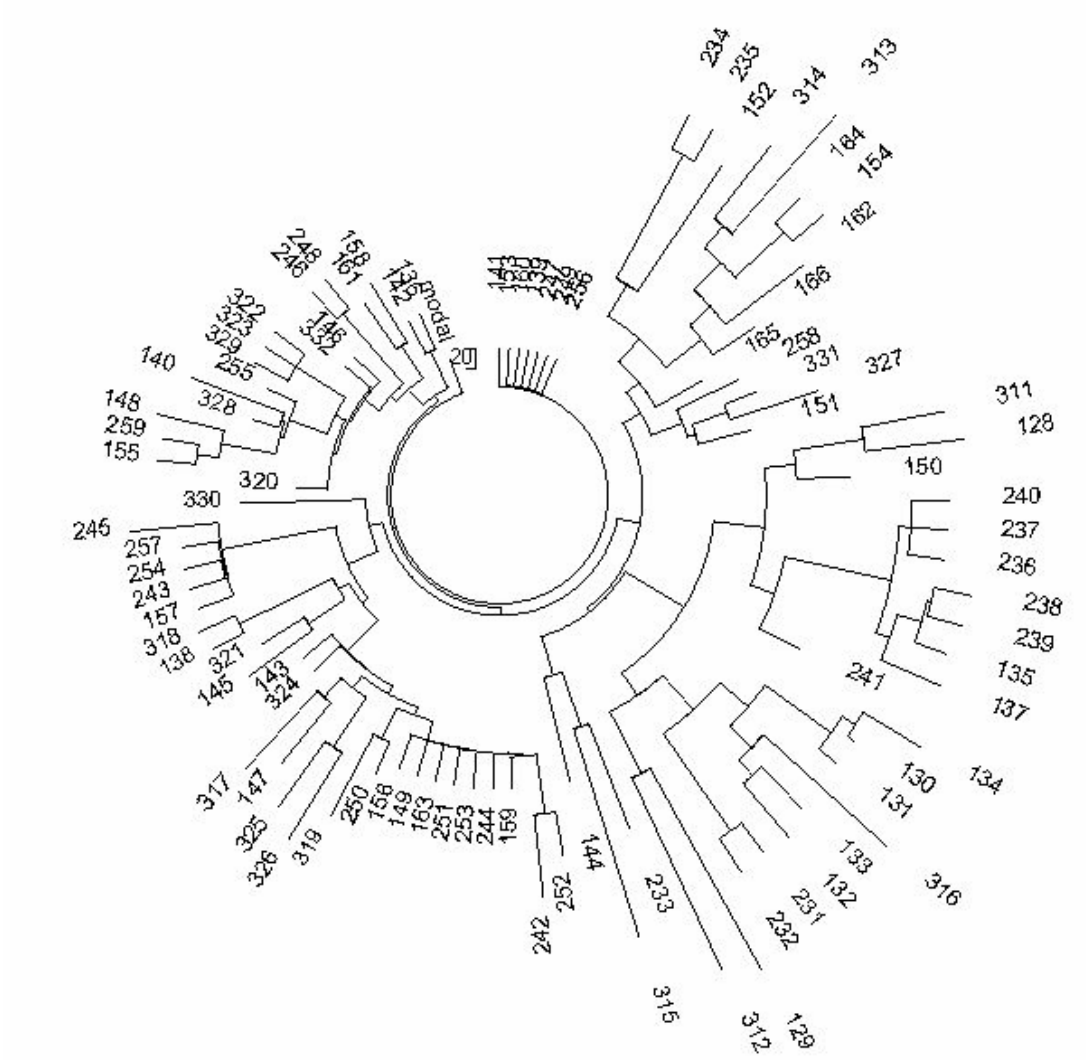


Рис. 10. Дерево 8-маркерных гаплотипов гаплогрупп K-R, R1b1a и R1b1c, построенное по данным работы (Contu et al, 2008). В выборке 90 гаплотипов. Гаплотипы гаплогрупп K-R под номерами 128, 129, 231, 232, 311-316 (все справа); гаплотипы гаплогруппы R1 под номерами 130-133 (ветвь справа внизу); гаплотипы гаплогруппы R1b1a под номерами 135-137, 236-240 (ветвь справа).

Общий предок гаплотипов гаплогруппы R1b1a с базовым гаплотипом

13-24-15-10-11-12-X-Y-Z-13-W-28

имел всего 10 мутаций на восьми гаплотипах, что дает 2000 лет до общего предка. Как будет показано ниже, этот предковый гаплотип отличается от предкового гаплотипа гаплогруппы R1b1c на Сардинии на 8 мутаций на 8 маркерах, что помещает ИХ общего предка на 15 тысяч лет назад. Кстати, наличие нескольких гаплотипов гаплогруппы R1a1 на Сардинии позволяет оценить время жизни их общего предка, что составляет 4400 лет, как и по всей Европе, за исключением Балкан (Клёсов, 2008a).

Для рассмотрения дерева гаплотипов R1b1b2 гаплотипы K-R, R1 и R1b1a были удалены, и оставшиеся гаплотипы (снип M269) приведены на рис. 11.

Видно, что полученное дерево состоит из нескольких ветвей гаплотипов: серия из семи идентичных гаплотипов на «стволе» дерева, отдельная ветвь из 12 гаплотипов справа вверху, ветвь из 16 гаплотипов слева вверху, и остальная часть дерева.

Разберем эти ветви по частям.

Семь одинаковых гаплотипов, образующих короткую «гребенку» на вершине дерева, имеют следующий состав:

13-24-14-11-11-14-X-Y-Z-13-W-29

Это – классический «Атлантический модальный гаплотип», который в полной 12-маркерной записи выглядит так:

13-24-14-11-11-14-12-12-12-13-13-29

Ясно, что эти гаплотипы – от недавнего предка, удаленного по времени от остальных гаплотипов на дереве гаплотипов на рис. 11.

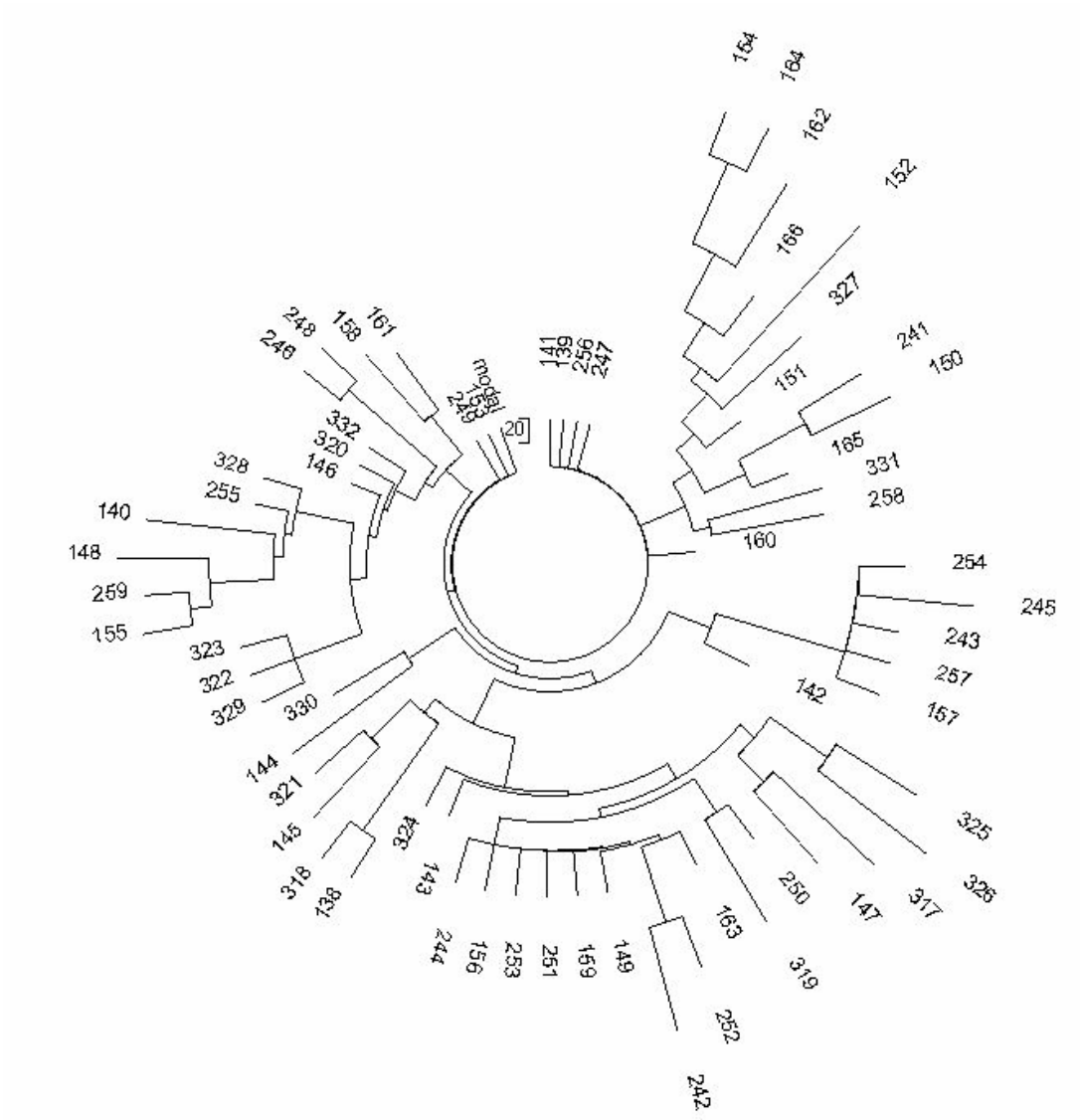


Рис. 11. Дерево 8-маркерных гаплотипов гаплогрупп R1b1b2, построенное по данным работы (Contu et al, 2008). В выборке 64 гаплотипа.

12 гаплотипов правой верхней ветви имеют 25 мутаций от базового гаплотипа

13-24-14-10-11-14-X-Y-Z-13-W-29

который отличается всего на одну мутацию от АМГ на том же дереве. Эти 25 мутаций соответствуют 141 поколению, или 3525 лет до общего предка.

16 гаплотипов левой верхней ветви имеют 28 мутаций от базового гаплотипа

13-24-14-10-11-15-X-Y-Z-13-W-29

который отличается на две мутации от АМГ на том же дереве, и всего на одну мутацию от предыдущего базового гаплотипа. Эти 28 мутаций соответствуют 116 поколениям или 2900 лет до общего предка.

Наконец, если объединить все ветви (что, строго говоря, не является правильным, поскольку ветви разного размера), то окажется, что все 64 гаплотипа содержат 180 мутаций, приводя к 201 поколению, или 5025 лет до общего предка всех носителей этих гаплотипов. Эта величина имеет условную надежность, но показывает, что она совершенно далека от величины 22800 лет, «определенную» авторами статьи (Contu et al, 2008) с использованием скорости мутации 0.00069 на маркер, то есть в 3.1 раза ниже, чем принятые в научном сообществе, причем с отбрасыванием двух маркеров из восьми.

Гаплотипы R1b в Сицилии

В работе (Di Gaetano et al, 2008) приведены всего девять 10-маркерных гаплотипов группы R1b с Сицилии. Этого количества недостаточно, чтобы строить дерево гаплотипов, но достаточно для проведения предварительной оценки времени жизни общего предка гаплогруппы на острове. Эти гаплотипы имеют следующий базовый гаплотип

13-24-14-11-11-14-X-Y-12-13-13-29

который идентичен гаплотипу относительно недавнего предка на Сардинии

13-24-14-11-11-14-X-Y-Z-13-W-29

и идентичен Атлантическому модальному гаплотипу

13-24-14-11-11-14-12-12-12-13-13-29

в пределах определенных аллелей. Все девять гаплотипов имеют 25 мутаций, что соответствует 143 поколениям, или 3600 лет до общего предка гаплогруппы R1b на Сицилии. Это – обычные европейские величины для гаплогруппы R1b.

Гаплотипы R1b1b2 в Ливане

Недавно были опубликованы 44 гаплотипа гаплогруппы R1b1b2 (бывшая R1b1c) из Ливана, наряду с десятью R* гаплотипами и тремя гаплотипами R1b1a из того же региона (Zalloua et al, 2008). Общее дерево этих гаплотипов дано на рис. 12.

Видно, что только три гаплотипа группы R* (под номерами 09, 10, 11) и все три гаплотипа R1b1a (14, 15, 16) выбиваются из общей структуры дерева. Остальные семь гаплотипов R* встроены в общую структуру, и даже являются частью относительно «молодых» ветвей.

При последующих расчетах я использовал только девять маркеров из 11, приведенных авторами, так как маркеры DYS437 и DYS438 принадлежат другим панелям, и могут внести неопределенность в расчеты.

Все десять гаплотипов R* сходятся к их предковому (базовому) гаплотипу, который идентичен, как будет показано ниже, базовому гаплотипу R1b1b2

12-24-14-10-X-Y-Z-12-12-13-13-29

Этот гаплотип отклоняется на две мутации от Атлантического модального гаплотипа:

13-24-14-**11-**11-14-12-12-12-13-13-29

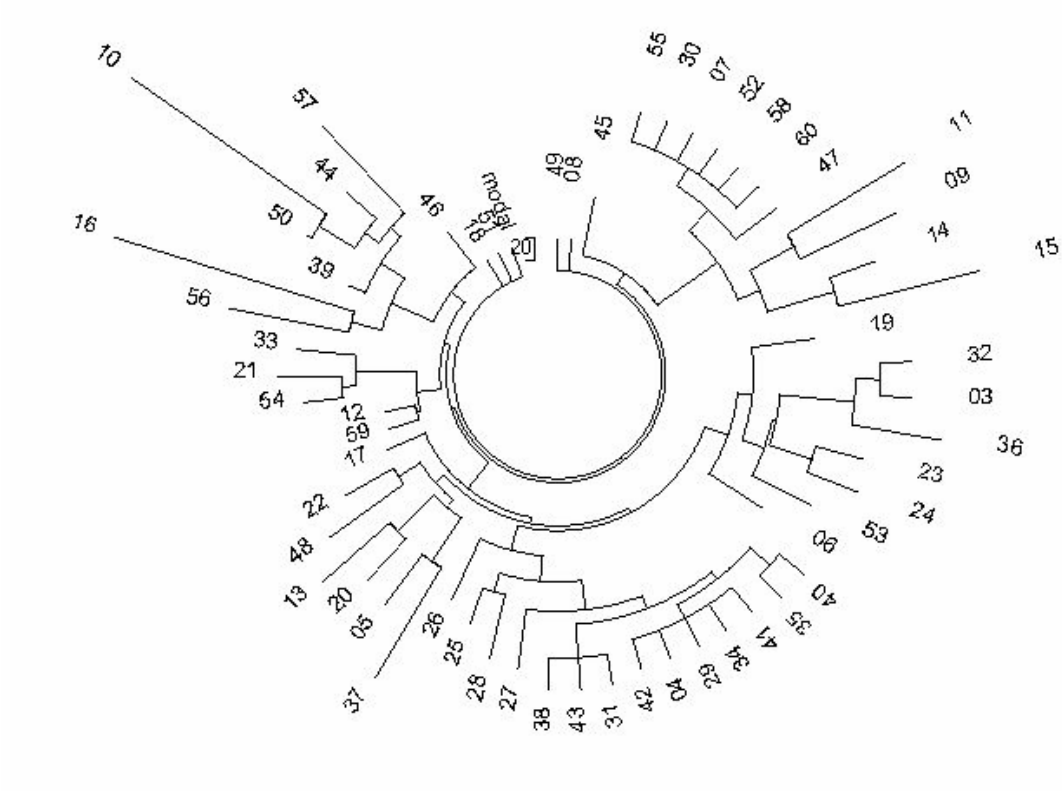


Рис. 12. Дерево 11-маркерных гаплотипов гаплогрупп R*, R1b*, R1b1a и R1b1b2, построенное по данным работы (Zalloua et al, 2008). В выборке 58 гаплотипов. R1b* - гаплотипы номер 03-12, R1b* - 13, R1b1a – 14-16, R1b1b2 – 17-60.

Все 10 гаплотипов группы R* содержат 37 мутаций от базового, что соответствует 262 поколениям, или 6550 годам до общего предка этих гаплотипов.

Три гаплотипа группы R1b1a, конечно, недостаточны для расчетов, поэтому была проведены просто приблизительная оценка их «возраста» относительно общего предка с гаплотипом (ориентировочно)

13-24-15-10-X-Y-Z-13-12-13-13-29

На три гаплотипа приходится 8 мутаций, что примерно соответствует 4400 лет до общего предка.

Перейдем к гаплотипам группы R1b1b2 (рис. 13).

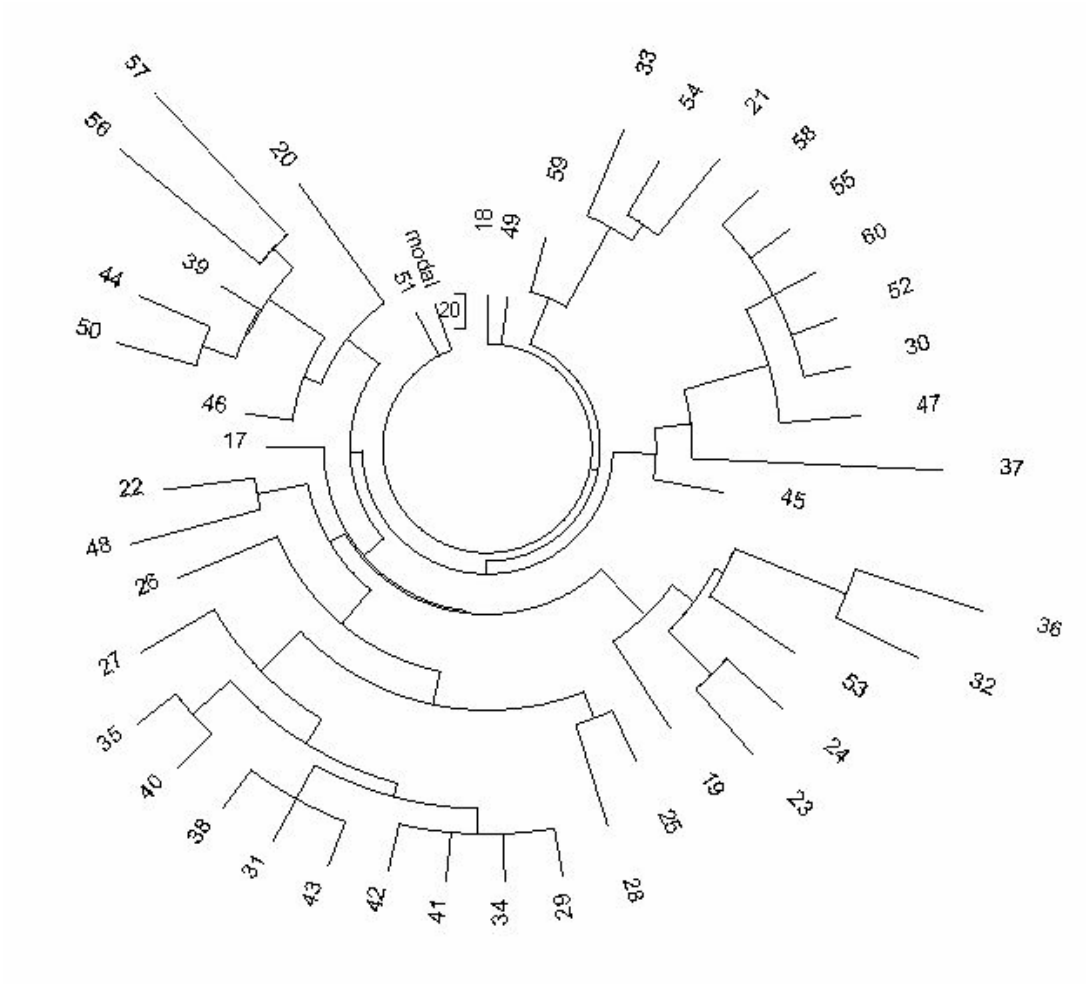


Рис. 13. Дерево 11-маркерных гаплотипов гаплогруппы R1b1b2, построенное по данным работы (Zalloua et al, 2008). В выборке 44 гаплотипа.

Видно, что выборка содержит несколько линий недавних предков (плоские ветви на рис. 13), но в целом не содержит сепаратных ветвей. Это означает, что для дерева можно принять одного общего предка. Его гаплотип – тот самый предковый гаплотип R*, приведенный выше

12-24-14-10-X-Y-Z-12-12-13-13-29

и который отклоняется на две мутации от АМГ. Все 44 гаплотипа содержат 137 мутаций, что дает 212 поколений, или 5300 лет до общего предка гаплогруппы R1b1b2 в Ливане. Это – несколько выше, хотя и не намного, возраста общих предков этой гаплогруппы в Европе.

Чтобы определить, действительно ли гаплогруппа R1b1b2 в Ливане старше европейских, применим два критерия – сравним с возрастом гаплогруппы R1a1 в том же регионе, и сравним с возрастом европейских гаплотипов R1b (M343), представленных в той же статье (Zalloua et al, 2008), по их 12- и 9-маркерным гаплотипам.

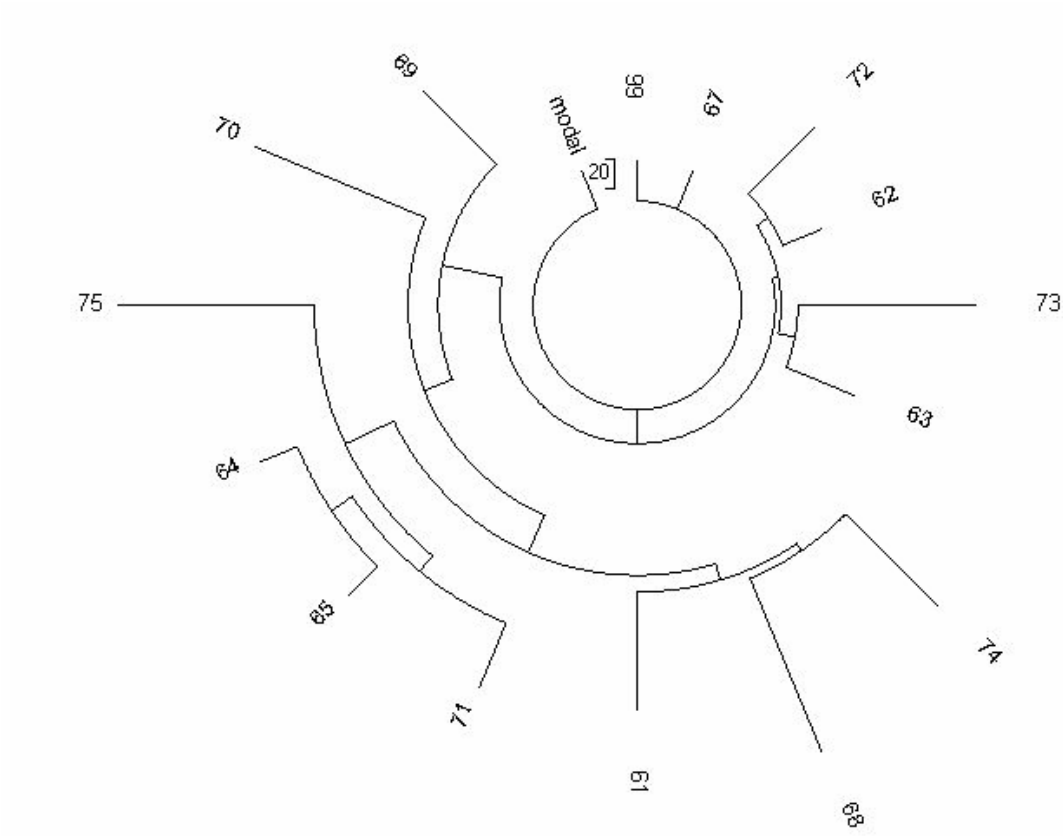


Рис. 14. Дерево 11-маркерных гаплотипов гаплогруппы R1a1, построенное по данным работы (Zalloua et al, 2008). В выборке 15 гаплотипов.

15 гаплотипов R1a1 из Ливана расходятся по двум совершенно разным ветвям (рис. 14).

Справа – относительно молодая ветвь, в ней всего четыре гаплотипа, содержащих пять мутаций от ориентировочного предкового гаплотипа R1a1 в Ливане:

13-24-15-11-X-Y-Z-12-10-14-11-31

Это дает «возраст» общего предка данных четырех гаплотипов всего лишь 75 поколений назад, то есть 1875 лет. Наша эра, 2-й век.

Более старая, левая ветвь на рис. 14, имеет базовый гаплотип

13-25-16-10-X-Y-Z-12-10-13-11-30

и содержит 29 мутаций на 9 гаплотипов. Это соответствует 221 поколению, или 5525 лет до общего предка. Сам гаплотип идентичен на данных маркерах типичному европейскому предковому гаплотипу R1a1

13-25-16-10-11-14-12-12-10-13-11-30

который в свою очередь в точности соответствует восточно-славянскому базовому (предковому) гаплотипу.

В любом случае, выборки ливанских гаплотипов R1b1b2 и R1a1 имеют общих предков практически одинакового возраста, 5300 и 5525 лет, соответственно.

Проверим это на других системах.

На рис. 15 приведено дерево 12-маркерных европейских гаплотипов, построенное по данным, приведенным в той же работе (Zalloua et al, 2008). Видно, что дерево практически симметричное, следовательно, происходит от одного общего предка. Все 112 гаплотипов содержат 334 мутации, что соответствует 160 поколениям, или 4000 лет до общего предка. Те же 112 гаплотипов, но 9-маркерные, содержат 258 мутаций, что соответствует 148 поколениями, или 3700 лет до общего предка.

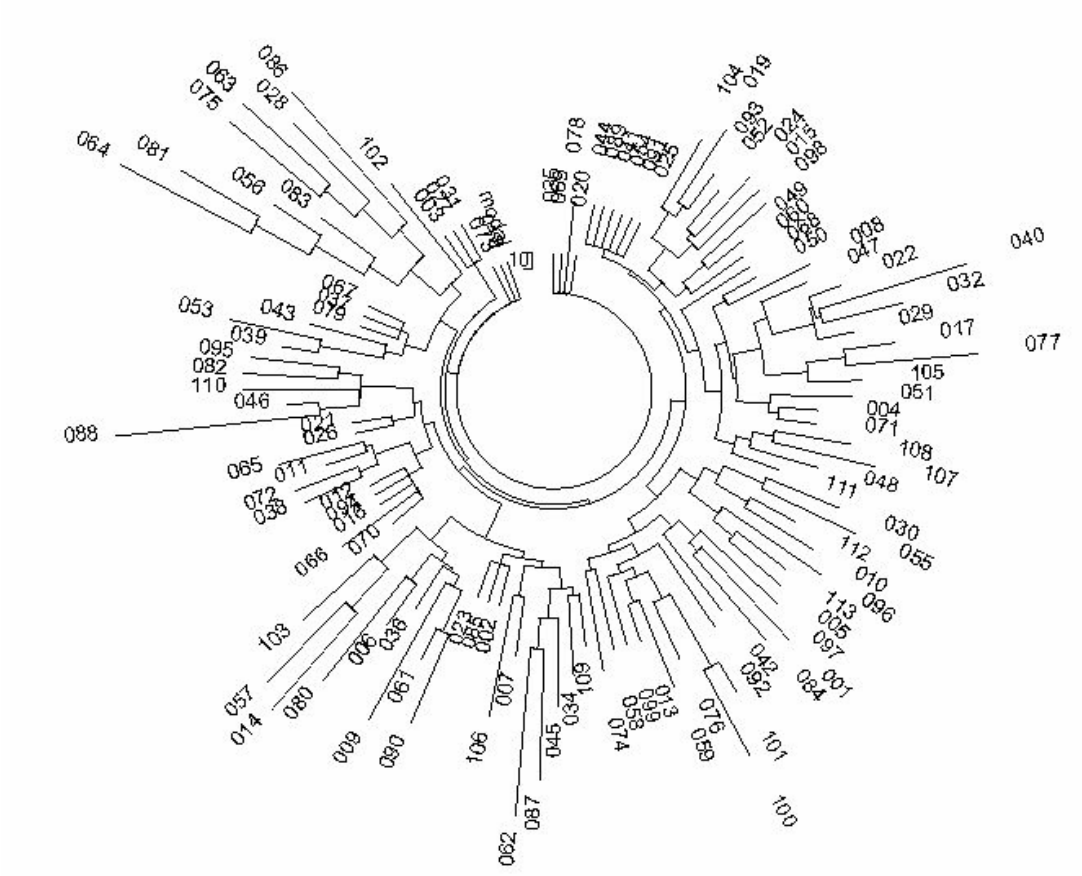


Рис. 15. Дерево 12-маркерных европейских гаплотипов гаплогруппы R1b (M343), построенное по данным работы (Zalloua et al, 2008). В выборке 112 гаплотипов. Представлены гаплотипы из Франции (29 гаплотипов), Англии (25), Германии (31), Италии (27 гаплотипов)

Иначе говоря, расчет по 9-маркерным гаплотипам в данном случае даже несколько занижает расстояние до общего предка, и указанные 5300 (R1b1b2) и 5525 (R1a1) лет в Ливане могут быть даже выше на несколько сот лет, например, 5700 и 6000 лет до общего предка.

Гаплотипы R1b1b2 в Анатолии

84 десятимаркерных гаплотипа R1b1b2 (M269) были приведены в статье (Cinnioglu et al, 2004). Среди них был маркер DYS#A7.2 (DYS461), который не числится в первых 37 маркерах в системе записи FTDNA, и поэтому был мной исключен из рассмотрения. Дерево 9-маркерных гаплотипов приведено на рис. 16.

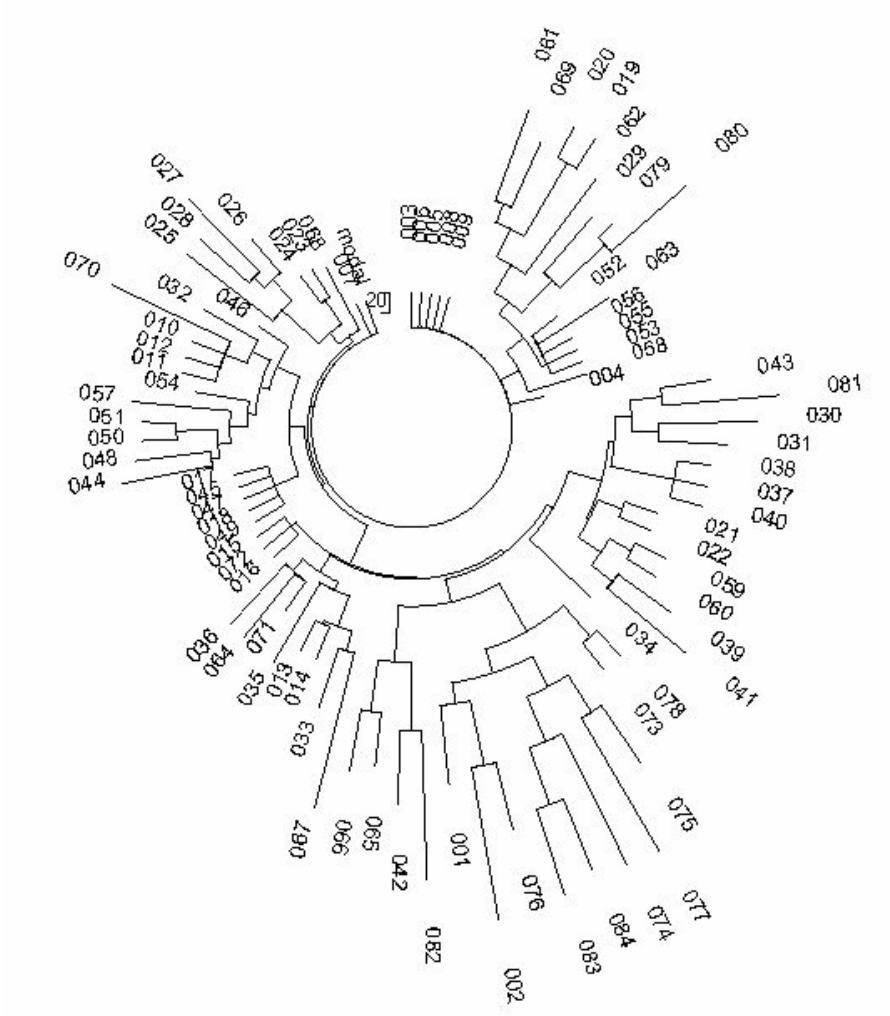


Рис. 16. Дерево 9-маркерных анатолийских гаплотипов гаплогруппы R1b1b2 (M269) и нескольких других (см. ниже), построенное по данным работы (Cinnioglu et al, 2004). В выборке 84 гаплотипа. Гаплотипы 001 и 002 принадлежат гаплогруппе R1b1 (P25), 079-082 представляют гаплогруппу R1b1b (M73), 083-084 относятся к исходной гаплогруппе R1b (M343).

Из 84 гаплотипов восемь относятся к отдельным гаплогруппам или субкладам (гаплотипы 001 и 002 принадлежат гаплогруппе R1b1/P25,

079-082 представляют гаплогруппу R1b1b/M73, 083-084 относятся к исходной гаплогруппе R1b/M343), и не были включены в расчеты.

Видно, что дерево на рис. 16 состоит из нескольких выраженных ветвей. На вершине дерева находятся семь идентичных гаплотипов

12-24-14-11-X-Y-Z-12-12-13-13-29

определенно от относительно недавнего общего предка. Этот гаплотип отклоняется всего лишь на одну мутацию от предкового ливанского гаплотипа той же гаплогруппы R1b1b2

12-24-14-**10**-X-Y-Z-12-12-13-13-29,

на одну и две мутации от предковых гаплотипов R1b1b2 Сардинии

13-24-14-11-11-14-X-Y-Z-13-W-29
13-24-14-**10**-11-15-X-Y-Z-13-W-29

на одну мутацию от предкового гаплотипа Сицилии (идентичного АМГ)

13-24-14-11-11-14-X-Y-12-13-13-29

и на одну мутацию от усеченного Атлантического модального гаплотипа (АМГ):

13-24-14-11-11-14-12-12-12-13-13-29

Ветвь справа вверху на рис. 16 состоит из 13 гаплотипов R1b1b2, с предковым гаплотипом

13-24-14-11-X-Y-Z-12-12-13-13-29

Это – опять усеченный Атлантический модальный гаплотип. Все 13 гаплотипов содержат 27 мутаций, что дает 132 поколения, или 3300 лет до общего предка.

Широкая ветвь из 33 гаплотипов R1b1b2 в нижней части дерева на рис. 16 содержит 99 мутаций, что дает 200 поколений, или 5000 лет до общего предка с базовым гаплотипом

12-24-14-10-X-Y-Z-12-12-13-13-29

Наконец, все 76 R1b1b2 гаплотипов содержат 207 мутаций, что дает 180 поколений, или 4500 лет до общего предка. Эта величина занижена, поскольку ее «тянут вниз» гаплотипы от недавних предков с малым количеством мутаций. Именно поэтому необходимо строить деревья и разделять гаплотипы на ветви.

Подводя итог данному разделу, гаплотипы R1b1b2 Анатолии имеют древнейшего общего предка, который жил 5000 лет назад, примерно тогда же, когда и в Ливане (5300 лет назад). Расчеты времени жизни общего предка гаплогруппы R1a1 на тех же территориях дали 4500 и 5525 лет, соответственно. Иначе говоря, представители родов R1a1 и R1b1b2 жили на этих территориях в одно и то же время.

Гаплотипы R1b в Армении

Выборка из 238 гаплотипов армян гаплогруппы R1b была представлена в работе (Weale et al, 2001). Она включала гаплотипы из шести регионов Армении, Карабаха, Ирана, и других районов армянской диаспоры в мире. Рассмотрение нами этих гаплотипов показало, что часть из них относятся к древним, а часть - к относительно «молодым» предкам. Иначе говоря, все эти выборки представляют неоднородную смесь популяций, и для всех расчеты по остаточным гаплотипам и по мутациям дают весьма различные результаты.

Тем не менее, во всех шести регионах базовый гаплотип был один и тот же:

12-24-14-11-X-X-X-12-X-X-13-X

Он отличается на одну мутацию от «Атлантического модального гаплотипа»

12-24-14-**10**-X-Y-Z-12-12-13-13-29,

Если не проводить разделение на отдельные ветви гаплотипов, и считать общих предков по регионам, то, например, для Иранского региона число поколений до «общего предка», рассчитанное по остаточным базовым гаплотипам (Клёсов, 2008b) и по мутациям оказалось равно ln(18/4)/0.00877 = 214 и 244, соответственно. Для Карабаха это равно 185 и 206 поколений. Для Восточной Турции – 237 и 286 поколений. Для Северной Армении – 249 и 291 поколение. Для Южной Армении – 151 и 178 поколений. Самое высокое число поколений оказалось для Араратского региона – 258 и 400 поколений, то есть, формально (и неправильно) говоря, 6,450 и 10,000 лет до общего предка. Дело в том, что при смеси общих предков в выборке первая цифра обычно оказывается завышенной, а вторая – заниженной. При разделении предков эти цифры обычно расходятся шире. Насколько шире расходятся – зависит от ряда факторов, в особенности от того, сколько потомков в выборке от каждого общего предка, и сколько до него поколений от настоящего времени.

Для того, чтобы разобраться с этой проблемой, было построено дерево для 52-х гаплотипов для объединенной выборки Араратского региона и Северной Армении (рис. 17)

Дерево подразделилось на пять ветвей. Две самые древние, но очень малочисленные ветви расположены слева внизу и справа вверху на рис. 17. Шесть гаплотипов первой содержат 18 мутаций, что формально соответствует 491 поколению, или 12300 лет до общего предка. Другая ветвь дала 13 мутаций на четыре гаплотипа, что соответствует 553 поколениям, или 13800 лет до общего предка. Третья, широкая ветвь в нижней правой части на рис. 17, состоит из 21 гаплотипа и содержит 36 мутаций, что соответствует 235 поколениям, или примерно 5900 лет до общего предка. Четвертая, плоская ветвь в левом верхнем секторе, из 10 гаплотипов, содержит 16 мутаций, что соответствует 216 поколениям, или 5400 лет до общего предка. Наконец, пятая ветвь оказалась совсем недавней, в ней были только 11 идентичных гаплотипов, поэтому возраст посчитать было нельзя. Примерно несколько сот лет до общего предка.

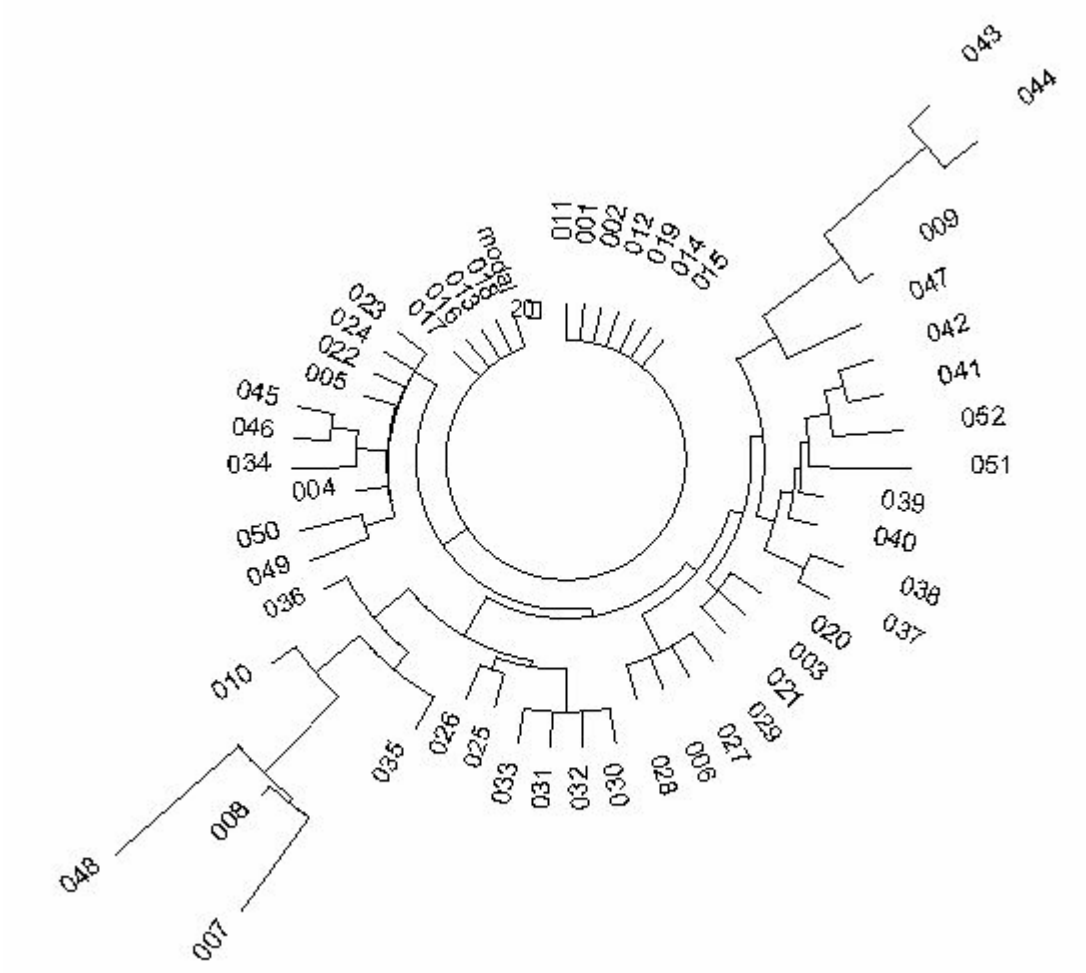


Рис. 17. Дерево 6-маркерных гаплотипов объединений выборки Араратского региона и Северной Армении, гаплогруппа R1b, построено по данным (Weale et al, 2001). В выборке – 52 гаплотипа.

Для сравнения, общий предок армянских R1a1 гаплотипов жил 4400 лет назад (Клёсов, 2008а).

Ясно, что большинство армянских общих предков гаплогруппы R1b значительно старше западно-европейских, которые жили в основном между 3400 и 4400 годами назад.

Гаплотипы R1b на Балканах

В частном порядке я получил от д-ра Marijana Pericic серию из 36 гаплотипов гаплогруппы R1b, собранных на разных территориях Балкан, побережья Адриатического моря, и островах Адриатики. Дерево гаплотипов показано на рис. 18.

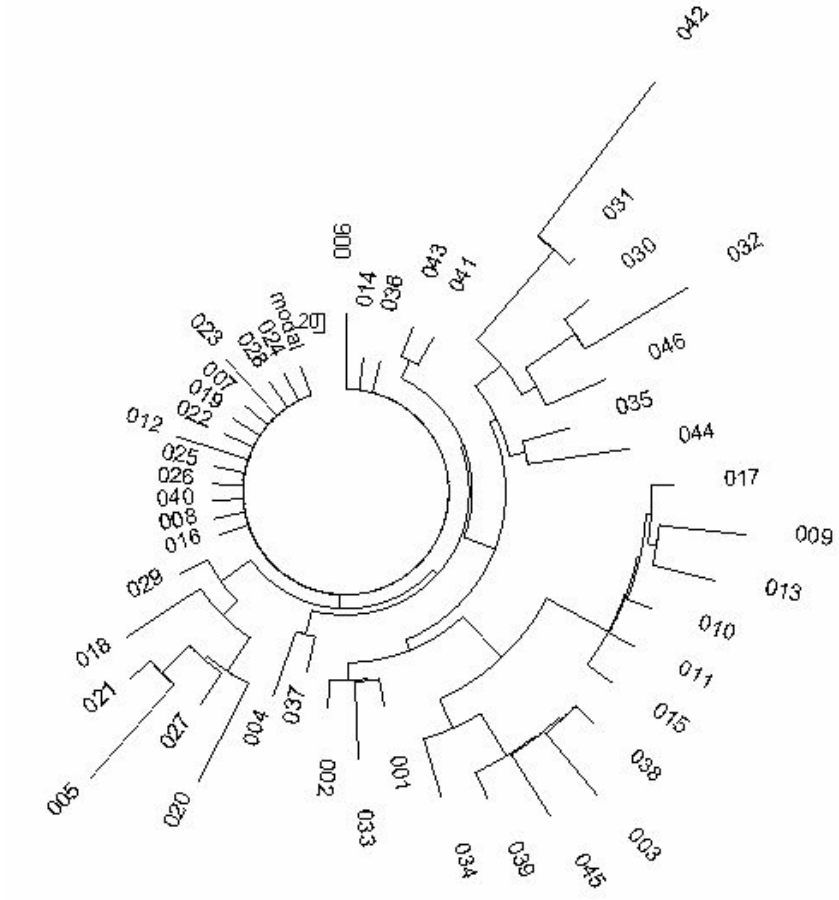


Рис.18. Дерево 9-маркерных балканских гаплотипов R1b, представленных д-ром Марьяной Pericic, Хорватия. В выборке – 46 гаплотипов.

Видно, что дерево гаплотипов содержит несколько ветвей, что соответствует нескольким общим предкам. Можно выделить четыре явно разные ветви, соответствующие четырем генеалогическим линиям. Одна из них совсем недавняя, из 15 гаплотипов, включает 12 базовых гаплотипов, и содержит всего три мутации. Это дает по остаточным базовым гаплотипам $\ln(15/12)/0.0173 = 13$ поколений, и по мутациям $3/15/0.0173 = 12$ поколений. Таким образом, молодая ветвь, в

количестве одной трети от всех R1b, имеет общего предка, который жил всего 300-325 лет назад, в конце 17-го – начале 18-го веков. Почти все они – жители острова Корсула в Адриатике. Базовый гаплотип у них следующий:

13-24-14-11-11-11-X-Y-Z-13-13-29

Он на три мутации отходит от известного Атлантического модального гаплотипа (АМГ):

13-24-14-11-11-**14**-12-12-12-13-13-29

Вторая ветвь, справа внизу на рис. 18, содержит 11 или 14 гаплотипов, в зависимости от того, как ее рассматривать. Базовый гаплотип при этом не меняется:

12-24-14-11-11-15-X-Y-Z-13-13-29

Здесь – две мутации в сторону от АМГ

В этой ветви 26 мутаций при 11 гаплотипах, или 31 мутация при 14 гаплотипах. Это дает соответственно 157 или 146 поколений, то есть 3,900 или 3,700 лет до общего предка. Можно записать 3,800±100 лет.

Третья ветвь, слева внизу, из 6 гаплотипов, содержит 17 мутаций от базового гаплотипа

13-24-14-11-11-**11**-X-Y-Z-**14**-13-29

и отстоит на четыре мутации от АМГ. Возраст до общего предка этой ветви составляет 4,900 лет.

Наконец, самая древняя ветвь, как и видно из дерева гаплотипов, состоит из 7 гаплотипов и содержит 21 мутацию. Это дает 208 поколений, или 5200 лет до общего предка, что немногим отличается от других цифр для возраста общего предка R1b в Европе. Более того, этот возраст в данном случае держится только на одном гаплотипе, номер 042 на рис. 18. Если его убрать, то останется 15 мутаций на 6 гаплотипах, и это даст вполне умеренные для Европы 168 поколений,

то есть примерно 4,200 лет до общего предка. Но такие разбросы, к сожалению, это общие недостатки малых выборок.

Шведские гаплотипы гаплогруппы R1b1b2

Набор из 383 девятимаркерных гаплотипов Швеции, включавший 76 гаплотипов R1b1b2, был опубликован в работе (Karlsson et al, 2006).

Базовый гаплотип R1b1b2 в Швеции следующий:

13-24-14-11-11-14-X-Y-Z-13-13-29

Это – классический Атлантический модальный гаплотип.

Все 76 шведских гаплотипов гаплогруппы R1b1b2 включают 7 базовых гаплотипов и содержат 187 мутаций. Без поправки на возвратные мутации это дает ln(76/7)/0.0173 = 138 поколений, и 187/76/0.0173 = 142 поколений до общего предка. Опять мы видим почти одинаковые величины, что указывает на одного общего предка в выборке. Поскольку число поколений превышает сотню, необходимо ввести поправку на возвратные мутации, которые в данном случае увеличат возраст предка примерно на 16%, а именно до 165 поколений, то есть до 4,125 лет до общего предка. Это – совершенно обычная цифра для европейских гаплотипов гаплогруппы R1b1b2.

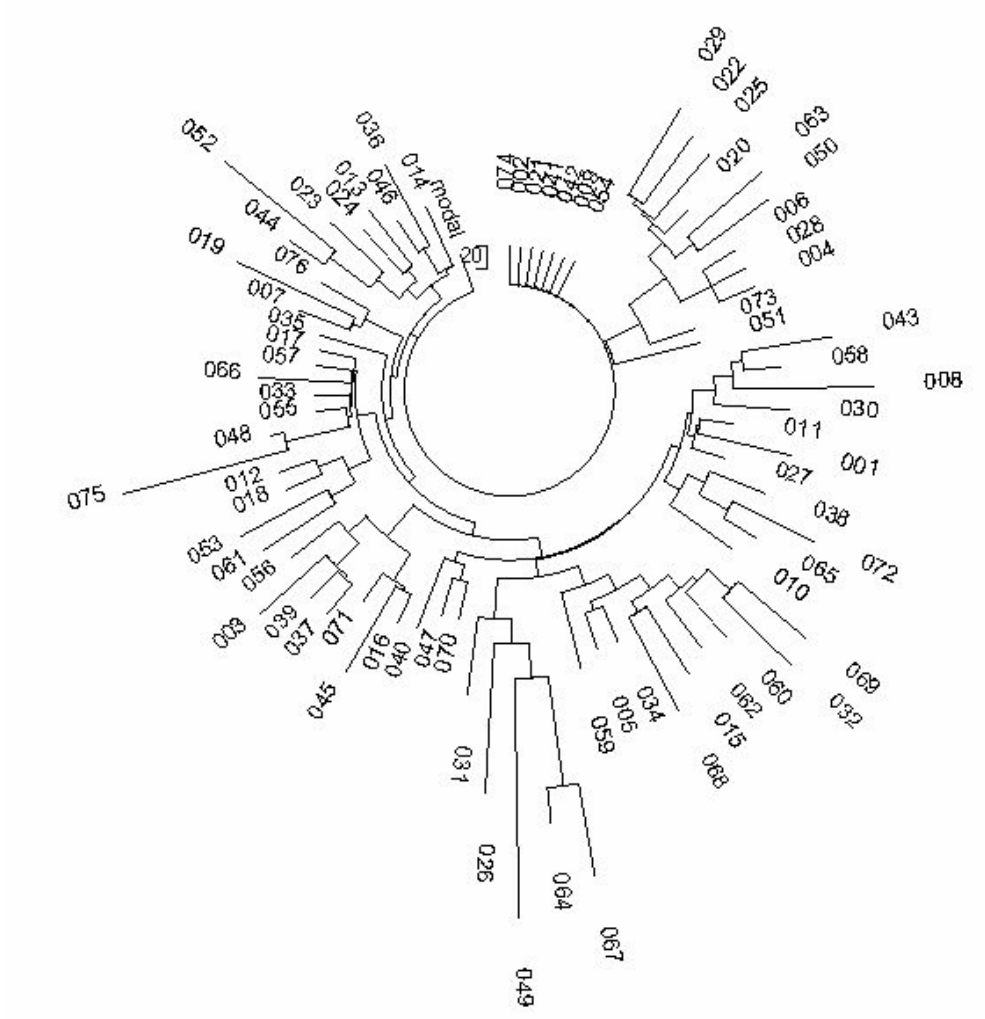


Рис.19. Дерево 9-маркерных шведских гаплотипов R1b1b2, по данным (Karlsson et al, 2006). В выборке – 76 гаплотипов.

Дерево гаплотипов (рис. 19) также показывает, что общий предок был один. Никакого расщепления дерева на несвязанные между собой ветви не отмечается. Некоторое подозрение может вызвать ветвь вверху справа, но там 11 гаплотипов, имеющих всего 16 мутаций, что дает всего 91 поколение, или 2,275 лет до общего предка. Это – просто один из более поздних вариантов генеалогической линии, в которой маркер DYS#389-1 мутировал из 13 в 14, что и сдвинуло также маркер DYS#389-2, слегка изменив базовый гаплотип:

13-24-14-11-11-14-X-Y-Z-**14**-13-**30**

Это – нормальные мутации в гаплотипах, которые скорее интересны тем, кто занимается «фамильной историей».

Гаплотипы R1b на британских островах

Выборка из 1242 гаплотипов R1b была представлена в работе (Campbell, 2007). К сожалению, автор выбрал их из общего списка в 1625 гаплотипов как наиболее «популярные», Остальные гаплотипы для счета оказались потеряны, а следовательно, фактически потеряны все, включая и те, что были представлены. На самом деле автор привел в работе только 50 десяти-маркерных «наиболее популярных» гаплотипов со всеми их повторами, а всех вариантов гаплотипов было 291. В итоге не показаны 383 гаплотипа с наибольшим количеством мутаций, которые и должны определить возраст до общего предка.

Рассмотрим, что нам даст даже неполный список гаплотипов в отношении расчетов времени до общего предка.

Все показанные 1242 гаплотипа включают 262 идентичных, базовых, предковых гаплотипа

13-24-14-11-X-Y-12-12-12-13-13-29

Это – все тот же классический Атлантический модальный гаплотип в 12-маркерном формате:

13-24-14-11-11-14-12-12-12-13-13-29

Все 1242 гаплотипа содержали 1708 мутаций от приведенного базового. Поскольку наиболее мутированные гаплотипы оказались потерянными, естественно, время до «общего предка» окажется значительно заниженным, то есть ближе к нам. Действительно, 1708/1242/0.0176 = 78 поколений, или всего 1950 лет назад, 1-й век нашей эры. Понятно, что это искаженные данные, однако, мы уже знаем их верхний потолок. А из доли базового гаплотипа в общем списке (эти данные уже предоставлены полностью) получаем ln(1625/262)/0.0176 = 104 поколений, или 2,600 лет до общего предка.

Что-то здесь не то, маловато. Возможно, это относительно недавний общий предок, наличие которого мы не можем проверить по мутациям, хотя верхний потолок этим данным не противоречит. Выше мы видели, что по Европе в целом 12-маркерные гаплотипы дают 3400-4000 лет до общего предка гаплогруппы R1b. Остается условно предположить, что R1b на Британских островах намного моложе по сравнению с остальной Европой. Мы знаем, что римские легионеры были в Британии в конце прошлой эры, но мы не знаем, какие именно гаплогруппы они принесли, и какие популяционные бутылочные горлышки имели место после того. Но данные, полученные с полутора тысячами гаплотипов, нельзя просто так отбросить. Они несут важную информацию, нуждающуюся в объяснении. Возможно, недавние генеалогические линии полностью «забивают» старые на Британских островах, в отличие от остальной Европы.

Фламандские гаплотипы R1b (северная Бельгия)

Серия из 64 двенадцати-маркерных гаплотипов была приведена в статье (Mertens, 2007). Но эти 12-маркерные включали только 10 маркеров из первой панели, а остальные два (DYS## 437 и 438) были взяты из 25- и 37-маркерной панелей. Для того, чтобы можно было напрямую сравнивать данные с другими выборками, я рассмотрел только 10-маркерные гаплотипы, в которых не было маркеров DYS## 388 и 426.

Вся серия имела следующий базовый гаплотип

13-24-14-11-11-14-X-Y-12-13-13-29

Это – опять классический Атлантический модальный гаплотип (АМГ), приведенный выше. Кстати, маркеры 437 и 438, удаленные из расчетов, имели «базовые» аллели, равные 15 и 12, соответственно, что тоже характерно для АМГ.

Во всей выборке было только два таких базовых гаплотипа. Это уже показывает, что выборка не содержит примесей АМГ от недавних общих предков, иначе этих «базовых» было бы гораздо больше. Например, среди 1242 гаплотипов с британских островов 262 были базовыми (см. выше), то есть 21%. У фламандцев их только 3%, в 7 раз

меньше. Именно поэтому британские гаплотипы в представленной выше выборке – в основном от недавних предков (или предка). У фламандцев – от предка древнего.

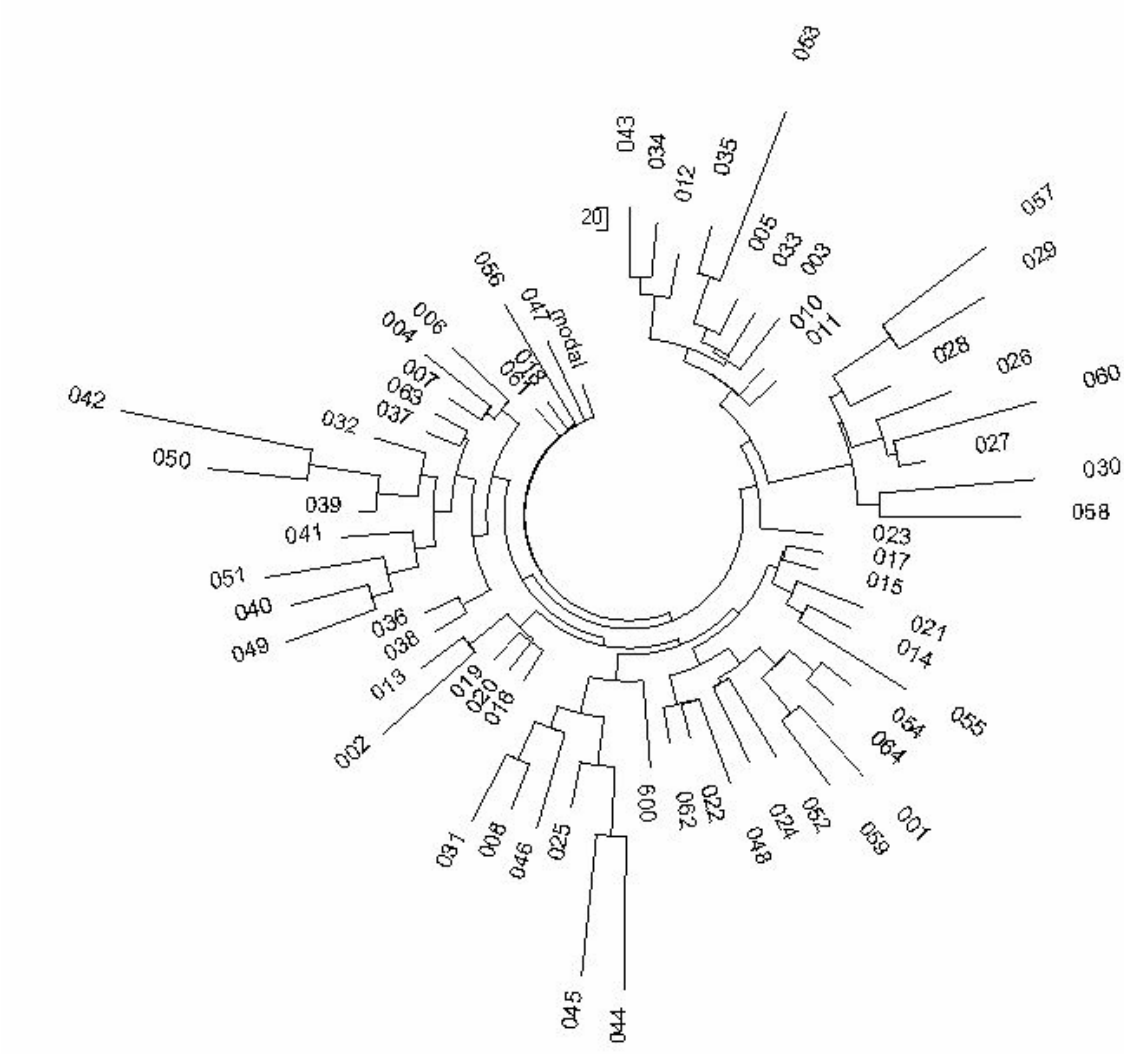


Рис.20. Дерево 10-маркерных фламандских (северная часть Бельгии) гаплотипов R1b, по данным (Mertens, 2007). В выборке – 64 гаплотипа.

Сравним два дерева гаплотипов R1b, общего по Европе (рис. 1) и фламандское (рис. 20). Разница принципиальная. У фламандского – явно один общий предок. По доле остаточных базовых гаплотипов число поколений до общего предка равно ln(64/2)/0.0218 = 159±20. Поскольку все 64 гаплотипа содержат 199 мутаций, то 199/64/0.0218 =

143 поколения. Погрешность в логарифмическом варианте рассчитывалась исходя из того, что если бы в выборке по статистической случайности оказалось не два, а три базовых гаплотипа, то ln(64/3)/0.0218 = 140 поколений. Согласование числа поколений до общего предка, рассчитанного по числу базовых гаплотипов и по числу мутаций показывает, что общий предок был один для всех 64 человек, и наши методы расчета полностью применимы.

С поправкой на возвратные мутации получим, что общий предок фламандцев гаплогруппы R1b из данной выборки жил 166 поколений, или 4,150 лет назад. Это близко к времени жизни общего предка гаплогруппы R1b1b2 в Швеции (4,125 лет назад), общим предком R1b в Европе, рассчитанным по 12-маркерным гаплотипам (3,375 лет назад), и с прочими данными для R1b на разных европейских территориях.

Гаплотипы R1b1b2 и R1b1b1/M73 из Средней Азии

Ниже приведены четырнадцать 25-маркерных гаплотипов среднеазиатского происхождения. Это гаплотипы узбеков, таджиков, тувинцев, уйгур, казахов, полученные непосредственно от работающих там исследователей.

001 13 22 14 11 13 17 12 12 12 13 13 30 15 9 9 11 11 23 15 20 31 12 14 15 17
002 13 22 14 11 13 17 12 12 12 13 13 30 15 9 9 11 11 23 15 20 31 12 15 15 17
003 13 22 14 11 13 17 12 12 13 13 13 30 15 9 9 11 11 23 15 20 29 12 15 15 17
004 13 22 14 11 13 17 12 12 13 13 13 30 15 9 9 11 11 23 15 20 29 12 15 15 17
005 14 24 14 13 13 16 12 12 12 13 13 31 16 9 9 11 11 23 15 20 29 12 15 15 17
006 14 24 14 13 13 16 12 12 12 13 13 31 17 9 9 11 11 24 15 20 29 12 15 15 17
101 13 19 14 10 13 13 12 12 13 14 13 30 18 9 9 11 11 23 15 19 33 12 15 15 16
102 13 19 14 11 13 13 12 12 13 14 13 30 17 9 9 11 11 21 14 19 33 12 15 15 16
103 13 19 14 11 13 13 12 12 13 14 13 30 17 9 9 11 11 23 15 19 33 12 14 15 15
104 13 19 14 11 13 13 12 12 13 14 13 30 17 9 9 11 11 23 15 19 33 12 15 15 16
105 13 19 14 11 13 13 12 12 14 14 13 30 17 9 9 11 11 23 14 19 33 12 15 15 16
106 13 19 14 11 13 13 12 12 14 14 13 30 17 9 9 11 11 23 15 19 31 12 12 15 15
107 13 19 14 11 13 13 12 12 14 14 13 30 17 9 9 11 11 23 15 19 33 12 15 15 16
108 13 19 14 11 13 13 12 12 14 14 13 30 17 9 9 11 11 23 15 19 33 12 15 15 16

(список гаплотипов приведен в ссылке
http://dnatree.ru/index.php?name=Forums&file=viewtopic&p=18889#18824)

Видно, что этот список гаплотипов неоднородный, и представляет две разные серии, каждая со своим общим предком. Этому соответствует нумерация гаплотипов от 001 до 006, и от 101 до 108. На дереве гаплотипов (рис. 21) они расходятся по совершенно разным ветвям.

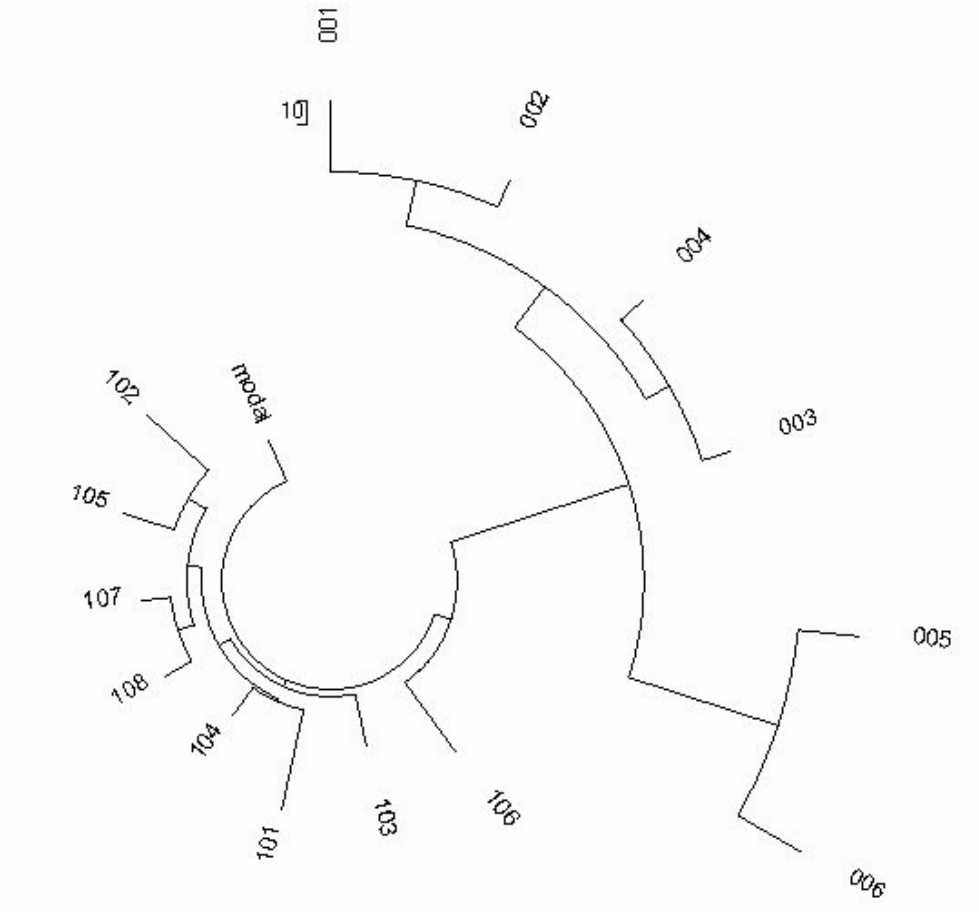


Рис. 21. Дерево из четырнадцати 25-маркерных гаплотипов среднеазиатских гаплотипов гаплогруппы R1b.
Для правой ветви базовый, или предковый гаплотип следующий:

13 22 14 11 13 17 12 12 12 13 13 30 15 9 9 11 11 23 15 20 29 12 15 15 17

Все 6 гаплотипов в сумме имеют 16 мутаций на первых 72 маркерах (то есть на первой 12-маркерной панели), у них мутированы 22% маркеров. Это – хорошая статистика, предок довольно древний. Во всех 25-маркерных панелях у них 25 мутаций.

По числу мутаций получаем – для 12-маркерных гаплотипов 141 поколение, или 3525 лет до общего предка, и для 25-маркерных 102 поколения, или 2550 лет до общего предка. В среднем 122±28 поколений, или 3100±700 лет до общего предка. Это – довольно обычные времена жизни общего предка R1b в Европе, и можно было бы полагать, что у приведенных шести гаплотипов он имеет европейское происхождение.

Однако это не так. По сравнению с «модальным атлантическим гаплотипом» (АМГ)

13 **24** 14 11 **11 14** 12 12 12 13 13 **29 17** 9 9(10) 11 11 **25** 15 **19** 29(30) **15** 15 **16(17)** 17

определенный выше базовый гаплотип имеет 8 мутаций на первых 12 маркерах, и 15-18 мутаций на всех 25 маркерах (соответствующие аллели выделены жирными цифрами). Это – очень большое расстояние во времени, и оно соответствует примерно 580 поколений на первых 12 маркерах, и 500-650 поколений на 25-маркерных гаплотипах. А поскольку мы знаем, что западноевропейский АМГ имеет общего предка 185 поколений, или 4625 лет назад (см. выше), то можно рассчитать, что общий предок западноевропейского АМГ и среднеазиатского предкового гаплотипа, приведенного выше, жил [122+185+(500-650)]/2 = от 400 до 480 поколений назад, или между 10000 и 12000 лет назад. Если усреднить, получится 11000±1400 лет назад. Он и был общим предком для европейских и приведенных выше шести азиатских представителей гаплогруппы R1b.

У левой ветви на рисунке базовый, или предковый гаплотип следующий:

13 19 14 11 13 13 12 12 14(13) 14 13 30 17 9 9 11 11 23 15 19 33 12 15 15 16

Общий предок всех восьми человек с этими гаплотипами жил совсем недавно по историческим меркам. Все восемь гаплотипов в сумме имеют всего пять мутаций на первых 96 маркерах (то есть на первой 12-маркерной панели), у них мутированы всего 5% маркеров. Во всех 25-маркерных панелях у них всего 18 мутаций.

Поскольку из восьми 12-маркерных гаплотипов четыре идентичны, то есть сохраняют предковый гаплотип, то расчеты показывают: ln(8/4)/0.022 = 31.5 поколений до общего предка. По числу мутаций в тех же восьми гаплотипах получаем 5/8/0.022 = 28.4 поколения до общего предка. Эти величины весьма близки, и в среднем дают 30±2 поколений. Такое совпадение показывает, что предок у данных восьми человек был действительно один, общий для всей серии.

Во всех 25-маркерных гаплотипах содержится 18 мутаций, поэтому 18/8/0.046 = 49 поколений, или 1225 лет до общего предка. Подобные разнобои между двумя панелями встречаются обычно тогда, когда количество мутаций в панели меньше 8-10% от числа маркеров. Поэтому примем, что общий предок для всех 8 человек жил 1225 лет назад, то есть примерно в 8-м веке нашей эры.

Самое интересное здесь то, что предковый гаплотип имеет аллель 19 во втором слева маркере (DYS390, по принятой номенклатуре), в то время как «классический» западноевропейский «атлантический модальный гаплотип» (АМГ) имеет там аллель 24. Пять мутаций – это не шутка, это многие тысячи лет разницы в эволюции гаплотипа. Вот так, с отмеченными мутациями, выглядит западноевропейский АМГ:

13 **24** 14 11 **11 14** 12 12 **12 13** 13 **29** 17 9 9(**10)** 11 11 **25** 15 19 **29**(**30) 15** 15 **16(17) 17**

Отличие на 12-маркерной панели составляет 11 или 12 мутаций, на всей 25-маркерной панели – от 19 до 25 мутаций. Это – огромное расстояние во времени, и соответствует примерно от 960 до 1100 поколений. А поскольку мы знаем, что западноевропейский АМГ имеет общего предка примерно 185 поколений, или 4625 лет назад (см. выше), то можно рассчитать, что общий предок западноевропейского АМГ и среднеазиатского предкового гаплотипа, приведенного выше, жил [49+185+(960-1100)]/2 = от 600 до 680 поколений назад. Таким образом, разумно будет принять, что общий предок R1b жил примерно

16000±1400 лет назад. Он и был предком для европейских и азиатских представителей гаплогруппы R1b.

На самом деле азиатский гаплотип с DYS390 = 19 принадлежит гаплогруппе R1b1b1 (M73). Он довольно широко распространен от Турции до Сибири. Как показано выше, он весьма далеко отстоит по мутациям (и, соответственно, по времени) от обычного европейского гаплотипа R1b1b2 (M269).

Посмотрим, наконец, а как далеко разнесены во времени два приведенных выше азиатских предковых гаплотипа R1b:

13 **22** 14 11 13 **17** 12 12 **12 13** 13 30 **15** 9 9 11 11 23 15 **20 29** 12 15 15 **17**
13 **19** 14 11 13 **13** 12 12 **14(13) 14** 13 30 **17** 9 9 11 11 23 15 **19 33** 12 15 15 **16**

Между ними – 9-10 мутаций на 12-маркерной панели, и 17-18 мутаций на 25-маркерных предковых гаплотипах. Это – 690-810 поколений на 12-маркерной панели, и 590-650 поколений на 25-маркерной панели, или 690±90 поколений. Поскольку мы знаем, что обладатели этих базовых гаплотипов жили примерно 122 и 49 поколений назад, то ИХ общий предок жил (122+49+690)/2 = 430 поколений, или 10800 лет назад. Это соответствует значению 11000±1400 лет назад, приведенному выше.

Заключение

Итак, общий предок гаплогруппы R1b в Азии жил примерно 16 тысяч лет назад. Потомки его ушли долгим путем в Европу, перед этим 12000-14000 лет назад жили в Армении, на что указывают армянские гаплотипы R1b (см. выше), и примерно 4000-4500 лет назад пришли в Европу. Тем временем оставшаяся азиатская ветвь примерно 11 тысяч лет назад в свою очередь расщепилась на две ветви, обе среднеазиатские. Они сейчас обнаружены по их относительно недавним предкам, жившим примерно 3100 и 1220 лет назад. Гаплотипы этих предков описаны выше. Возможно, обе популяции прошли бутылочное горлышко три тысячелетия и тысячелетие назад, и то, что мы видим – это две выживших генеалогических линии, довольно далеко разошедшиеся за 10-12 тысяч лет. Но они продолжают вести линию их древнего общего предка, основавшего род R1b.

Можно предположить, что с Кавказа носители гаплогруппы R1b отправились сначала на север, вслед за уходящими ледниками. Много позже они оставили следы своего пребывания в виде археологических культур – возможно, культуры шнуровой керамики, она же культура боевых топоров, она же культура одиночных могил. У археологов пока нет ответа на вопрос, к каким родам принадлежали эти культуры. Мы не можем исключить, что ямная или/курганная культуры также были оставлены родом R1b. Похоже, пути движения родов R1b и R1a1 пересеклись 6000-4000 лет назад, первые в целом передвигались на запад, вторые – на восток. Территория Восточной Европы хранит память о том времени в виде большого количества древних черепов, раздробленных каменными боевыми топорами.

Типичные представители гаплогруппы R1b – кельты, которые появились в Западной Европе 3500-4500 лет назад. Кстати, кельты – имя собирательное и впервые примененное в его современном значении не так давно, в начале 18-го века, Эдвардом Лайдом, директором Ашмолеанского музея в Оксфорде. Путешествуя по делам музея, он обратил внимание на сходство языков уэльсцев, корнишей, бретонцев, ирландцев, шотландских галлов, и древних галльских языков. Он и объединил эти языки под общим, придуманным им именем кельтских языков. Хотя само имя кельты упоминал еще Юлий Цезарь в книге «Записки о галльской войне», как синоним галлов.

Таблица ниже представляет полученные в настоящей работе данные по временам жизни общих предков.

Гаплогруппа (субклад)	**Времена жизни общего предка популяции рода R1b или его субкладов (в годах от настоящего времени)**	
	25-маркерные	**12-маркерные**

	гаплотипы	**гаплотипы**
Все R1b1 из базы данных, 16959 и 12090 гаплотипов (12 и 25-маркерных)	4500	3950
R1b (R1b1), Европа	4625	3375
Субклад R1b (неидентифицирован)	3950	3375
R1b1c/M269 (R1b1b2)	4450	3675
R1b1c9/R1b-U106/S21	4225	3375
То же, но DYS425 null	3375	3200
R1b1c10/R1b1b2a2g/S28/U152	4375	4175
R1b Баски (старая и молодая ветвь)	4025 1675	4425 1725
R1b Баски, общий предок по всем ветвям	4600	4325
R1b Буковины	4675	5325
R1b1b2 Сардиния (8-маркерные гаплотипы), отдельные ветви и вычисленный общий предок		3525 2900 5025
R1b Сицилия (10-маркерные гаплотипы)		3600
R1b1b2 Ливан (9-маркерные гаплотипы)		5300
R1b (M343) европейские гаплотипы (Англия, Франция, Италия, Германия), 12-маркерные гаплотипы		4000
То же, 9-маркерные гаплотипы		3700
R1b1b2 Анатолия (9-маркерные гаплотипы), отдельные ветви и суммарный общий предок		5000 3300 4500

R1b Армении (6-маркерные гаплотипы), отдельные ветви		13,800 12,300 5,900 5,400
R1b на Балканах (9-маркерные гаплотипы), отдельные ветви		5200-4200 4900 3800
R1b1b2 в Швеции (9-маркерные гаплотипы)		4125
R1b1 в северной Бельгии (10-маркерные гаплотипы)		4150
R1b в Британии (10-маркерные гаплотипы, рассчитано по числу идентичных гаплотипов в выборке из 1625 гаплотипов)		2600
R1b1b2 в Средней Азии	3525	2550
R1b1b1 (M73) в Средней Азии	750	1225

Литература

Клёсов, А.А. Откуда появились славяне и «индоевропейцы» и где их прародина? Ответ дает ДНК-генеалогия. Вестник Российской Академии ДНК-генеалогии. 1, 400-477 (2008a)

Клёсов, А.А. Основные положения ДНК-генеалогии (хромосома Y), скорости мутаций, их калибровка и примеры расчетов. Вестник Российской Академии ДНК-генеалогии. 1, № 2, 252-348 (2008b).

Anthony, D.W. The Horse, The Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World. Princeton University Press, 2007, p. 326.

Campbell, K.D. Geographic patterns of haplogroup R1b in the British Isles. J. Genetic Genealogy 3, 1-13 (2007)

Cinnioglu, C., King, R., Kivisild, T., Kalfoglu, E., Atasoy, S., Cavalleri, G., Lillie, A.S., Roseman, C.C., Lin, A.A., Prince, K., Oefner, P.J., Shen, P., Semino, O., Cavalli-Sforza, L.L. and Underhill, P,A. Excavating Y-chromosome haplotype strata in Anatolia. Hum. Genet. 114, 127-148 (2004)

Contu, D., Morelli, L., Santoni, F., Foster, J.W., Francalacci, P. And Cucca, F. Y-Chromosome based evidence for pre-neolithic origin of the genetically homogeneous but diverse Sardinian population: inference for association scans. PLoS ONE, 1, 1-8 (2008).

Derenko, M., Malyarchuk, B., Denisova, G., Wozniak, M., Grzybowski, T., Dambueva, I. and Zakharov, I. Y-chromosome haplogroup N dispersals from south Siberia to Europe. J. Hum. Genet. 52, 763-770 (2007)

Di Gaetano, C., Cerutti, N., Crobu, F., Robino, C., Inturri, S., Gino, S., Guarrera, S., Underhill, P.A., King, R.J., Romano, V., Cali, F., Gasparini, M., Matullo, G., Salerno, A., Torre, C., and Piazza, A. Differential Greek and northern African migrations to Sicily are supported by genetic evidence from the Y chromosome. Eur. J. Human Genetics, August 2008, preprint.

Enattah, N. S., Trudeau, A., Pimenoff, V., Maiuri, L., Auricchio, S., Greco, L., Rossi, M., Lentze, M., Seo, J.K., Rahgozar, S., Khalil, I., Alifrangis, M., Natah, S., Groop, L., Shaat, N., Kozlov, A., Verschubskaya, G., Comas, D., Bulayeva, K., Mehdi, S.Q., Terwilliger, J.D., Sahi, T., Savilahti, E., Perola, M., Sajantila, A., Järvelä, I., and Peltonen L. Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. Am J Hum Genet. 81, 615-25 (2007).

Gayden, T., Cadenas, A.M., Regueiro, M., Singh, N.B., Zhivotovsky, L.A., Underhill, P.A., Cavalli-Sforza, and Herrera, R. J. The Himalayas as a directional barrier to gene flow. Amer. J. Human Genetics, 80 (2007)

Karlsson, A.O., Wallerstrom, T., Gotherstrom, A. and Holmlund, G. Y-chromosome diversity in Sweden – A long-time perspective. Europ. J. Human Genetics, 14, 963-970 (2006)

Martinez, L., Underhill, P.A., Zhivotovsly, L.A., Gayden, T., Moschonas, N.K., Chow, C.-E. T., Conti, S., Mamolini, E., Cavalli-Sforza, L.L. and Herrera, R.J. Paleolitic Y-haplogroup heritage predominates in a Cretan highland plateau. Eur. J. Human Genetics, 15, 485-493 (2007)

Mertens, G. Y-Haplogroup frequencies in the Flemish population. J. Genetic Genealogy 3, 19-25 (2007)

Sengupta, S., Zhivotovsky, L.A., King, R., Mehdi, S.Q., Edmonds, C.A., Chow, C.-E. T., Lin, A.A., Mitra, M., Sil, S.K., Ramesh, A., Rani, M.V.U., Thakur, C.M., Cavalli-Sforza, L.L., Majumder, P.P., and Underhill, P.A. 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 Pastoralis. Amer. J. Human Genet. 78, 202-221 (2006)

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-862 (2007).

Weale, M.E., Yepiskoposyan, L., Jager, R.F., Hovhannisyan, N., Khudoyan, A., Burbage-Hall, O., Bradman, N. and Thomas, M. Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group. Hum. Genet. 109, 659-674 (2001)

Zalloua, P.A., Xue, Y., Khalife, J., Makhoul, N., Debiane, L., Platt, D.E., Royyuru, A.K., Herrera, R.J., Hernanz, D.F.S., Blue-Smith, J., Wells, R.S., Comas, D., Bertranpetit, J., Tyler-Smith et al. Y-Chromosomal diversity in Lebanon is structured by recent historical events. Amer. J. Hum. Genet. 82, 973-882 (2008).

Zhivotovsky, L.A., Underhill, P.A., Cinnoglu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, F., Destro-Bisol, G., Spedini, G., Chambers, G.K., Herrera, R.J., Yong, K.K., Gresham, D., Tournev, I., Feldman, M.W., and Kalaydjieva, L. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am. J. Human Genet. 74, 50-61 (2004).

Теоретическая и практическая оценка возвратных мутаций в гаплотипах Y-хромосомы

Дмитрий С. Адамов и Анатолий А. Клёсов

nimissin@mail.ru
http://aklyosov.home.comcast.net

В рамках пуассоновского подхода выведены формулы для оценки вклада возвратных мутаций при расчетах времен жизни общих предков выборки гаплотипов. Рассчитаны таблицы поправок.

«Возвратные мутации» - это теоретическая концепция, имеющая важное практическое значение. Вряд ли кто на самом деле видел возвратные мутации при рассмотрении гаплотипов человека, но нет сомнений, что они имеют место, и тем самым меняют картину наблюдаемых мутаций в гаплотипах.

При рассмотрении мутаций в гаплотипах очевидно, что они происходят по кинетике первого порядка, то есть «сами по себе», спонтанно, и зависят только от случайности этого события, то есть происходят статистически. Это – допущение, но в целом хорошо подтверждаемое практикой. При рассмотрении массивов реальных гаплотипов видно, что отклонения от картины чисто случайных мутаций имеют место – то мутация произойдет скачком сразу на два, а то и на три шага в одну сторону в одном поколении, то происходит чаще в одну сторону, чем в другую, то один маркер мутирует значительно быстрее, чем другой. Тем не менее, общая картина мутаций легко усредняется, и в целом подчиняется, как было отмечено, простой кинетике первого порядка, как по отдельным маркерам, или локусам, так и по гаплотипам в целом.

Поэтому разумным является подход, при котором не стоит обращать особое внимание на частные отклонения от общей картины мутаций, и тем более не приходить от этого в отчаяние, а попытаться получить максимум информации при анализе этой не идеальной системы. А далее, в ходе развития науки, вносить усовершенствования и поправки. Это – нормальный ход развития науки. В физической химии, например, уравнения выведены, как правило, для идеальных растворов, далеких от реальных систем. Учебник одного из авторов

настоящей статьи начинается словами «До настоящего времени базис химической кинетики остается наименее строгим среди всех разделов физической химии» (Березин и Клёсов, 1976). Это не помешало ряду ученых получить Нобелевские премии за работы по химической кинетике.

Когда гаплотип мутирует, он скорее всего мутирует и обратно, притом, возможно, с той же частотой, или вероятностью. Если аллель в гаплотипе, скажем, из 17 мутирует в 18, то есть вероятность, что 18 в одном из следующих поколений мутирует не только далее в 19, но и обратно в 17. То есть произойдет возвратная мутация, скрывающая от наблюдателя факт двух последовательных мутаций. Ясно, что такие возвратные мутации должны приводить к кажущемуся замедлению мутаций. Процесс продолжает быть неравновесным, исходного гаплотипа в течение долгого времени больше, чем его мутированных производных, но со временем, в ходе накопления мутированных гаплотипов, количество возвратных мутаций постепенно нарастает, прогрессивно замедляя общую скорость мутаций исходного гаплотипа.

Иначе говоря, возвратные мутации «упрятаны» среди общего массива наблюдаемых мутаций, и незаметны для исследователя. Но чем старше предок, чем больше в линии потомков было возвратных мутаций, и тем выше ошибка расчета времени, отдаляющего нас от предка.

Именно согласно кинетике первого порядка, пока мутированного гаплотипа (или маркера) мало, скорость возвратной мутации пренебрежимо мала. С накоплением мутированного маркера скорость возвратной мутации становится ощутимой, и затем, как было сказано, нарастает прогрессивно. Но для этого в реальных условиях нужны тысячелетия. Для 12-маркерного гаплотипа, например, если простой расчет по мутациям дает время до общего предка, равное 320 поколений, или 8000 лет, на самом деле это уже 484 поколения, или 12100 лет. Таким образом, наблюдаемая скорость мутаций будет уже не 0.0022 мутаций на маркер на поколение, а 0.00145.

Это явление, иногда называемое гомоплазией, может быть описано в виде следующей схемы

$$D \leftrightarrow B \leftrightarrow A \leftrightarrow C \leftrightarrow E$$

где A – базовая, или предковая аллель, и остальные – ее ближайшие мутированные формы. При гомоплазии мы видим меньше мутаций, чем их на самом деле имело место, и чем глубже во времени, тем больше несоответствие. Мы видим меньше мутаций, и рассчитываем (без учета

гомоплазии), что предок жил, скажем, 10 тысяч лет назад, а он на самом деле жил более 16 тысяч лет назад.

Давайте разберемся, какими формулами это несоответствие можно описать, и составим удобные таблицы, которые может использовать любой желающий, тем более не разбирающийся в математике.

Будем исходить из того, что скорость мутаций, усредненная по всем маркерам, равна 0.002 мутаций на маркер на поколение, или одна мутация на маркер в среднем раз в 500 поколений (12500 лет, считая 25 лет на поколение). Отсюда легко скорректировать табличные данные для любой средней скорости мутации на маркер, поскольку скорость мутации различается для различных наборов маркеров в гаплотипах.

Введение поправок будем сравнивать с простейшей линейной моделью, для которой число наблюдаемых мутаций пропорционально числу поколений до общего предка для выборки гаплотипов. Эта модель применима для расчетов при относительно малом числе поколений, или (что по смыслу то же самое) малом числе мутаций в базовом (предковом) гаплотипе. В этом случае формула для расчета времени, прошедшего от общего для всей выборки предка, проста:

$t = n/N/\mu$

где t – время до общего предка, в поколениях,
n – количество мутаций во всех N гаплотипах выборки,
μ – средняя скорость (частота) мутаций, выраженная в числе мутаций на гаплотип на поколение.

Поскольку мутации в маркерах редки, примерно раз в 500 поколений, то можно принять, что основными вариантами обратной мутации являются одношаговые, как показано на схеме выше. На практике, при изучении мутаций в гаплотипах для пар «отец-сын», наблюдались и разовые двойные мутации, что позволяет оценить их вероятность. Например, в работе (Gusmão et al, 2005) было найдено, что из 137 мутаций 131 была одношаговой, 4 – двойные, 1 – тройная и еще 1 – с изменением аллели сразу на 4. Таким образом, разница в вероятности таких событий составляет (в данном примере) 22 раза в пользу разовых мутаций.

Как отмечалось выше, чем больше мутаций, тем больше в них «упрятано» возвратных мутаций. Приведем расчетный пример (Каржавин, 2008). В таблице ниже первая колонка – номер мутации от начала, то есть от предкового гаплотипа, на одном маркере. Для 20 мутаций понадобится среднее

время (усредненное по разным маркерам) 10 тысяч поколений, то есть 250 тысяч лет. Иначе говоря, пример нереальный, просто иллюстрационный. Он показывает, почему мутации не могут уходить в сторону до бесконечности, и практически доходят только до определенного предела за время жизни человечества. Иначе говоря, почему пределы мутаций ограничены. Образно говоря, мутации как бы сжимают пружину, которая далеко их не пускает. На этот счет было много дискуссий среди биологов, которые, к сожалению, не вспоминали про обратные мутации. Но это – не единственная возможная причина ограничений мутаций. Есть и другие возможные причины. Прежде всего, число STR-повторов ограничено. Нижнее ограничение – 2, так как меньше повторов не бывает. Фактически по используемым маркерам – 6. Верхняя граница для тетрануклеотидов составляет около 40. Можно предположить, что она как-то связана с длиной участка ДНК, образующего нуклеосому (147 п.о.), элементарную единицу упаковки хроматина.

Итак, первая колонка – номер мутации от предковой аллели, вторая колонка – вероятность того, что за прошедший путь было хотя бы одно одношаговое возвращение в состояние предка, то есть хотя бы одна возвратная мутация в предковую аллель, и третья колонка – вероятность того, что первое возвращение к предковому маркеру произошло на данной мутации.

2	0.500	0.500
4	0.875	0.125
6	0.938	0.062
8	0.961	0.039
10	0.973	0.027
12	0.979	0.021
14	0.984	0.016
16	0.987	0.013
18	0.989	0.011
20	0.991	0.009

Как видно, уже на четвертой мутации вероятность возвратной почти 90%, и с четвертой мутации вероятность впервые вернуться к предковому гаплотипу весьма мала.

Часть математическая. Распределения Пуассона

Вероятность мутаций мала, они происходят независимо друг от друга. Эти два положения позволяют утверждать, что распределение числа мутаций на локус (маркер) описывается распределением Пуассона. Тогда фактическое число мутаций в целом на гаплотип также подчиняется распределению Пуассона:

$$P(m) = \frac{\lambda^m}{m!} e^{-\lambda}$$

где:

P(m) – вероятность возникновения m мутаций в гаплотипе из исследуемой выборки,
m – фактическое число мутаций в индивидуальном гаплотипе, состоящий из n локусов,
λ – среднее число мутаций на гаплотип по всей выборке, то есть произведение суммарной скорости мутаций (μ) в локусах гаплотипа на число поколений (T), отделяющих гаплотип от предкового, $\lambda = \mu T$.
Ясно, что m – всегда целое положительное число, а λ – любое положительное.

Если μ_i – скорость мутаций в отдельном локусе *i,* i = 1, 2, …, *n.*

то

$\mu = \mu_1 + \mu_2 + \ldots + \mu_n$,

Теперь введем важное допущение. Поскольку количество мутаций всегда положительное число, мутационные события редки и независимы, то скорость мутаций в локусе *i* равна

$$\mu_i = \mu_i(+) + \mu_i(-),$$

где $\mu_i(+)$ - скорость мутаций, приводящая к увеличению аллели на 1,

$\mu_i(-)$ - скорость мутаций, уменьшающая аллель на 1.

В этом приближении мы пренебрегаем мутациями, изменяющими аллель более чем на единицу за одно поколение. Далее, заметим, что в случае $\mu_i = \mu_i(+)$ или $\mu_i = \mu_i(-)$ возвратных мутаций нет, поскольку изменение аллели происходит только в одну сторону: либо увеличения, либо уменьшения. Кроме того, уровень возвратных мутаций в гаплотипе зависит от разброса скоростей мутаций μ_i для разных локусов (маркеров).

Перейдем к расчету вклада возвратных мутаций в пуассоновском приближении

Обозначив $\lambda_i = \mu_i T$, получаем

$$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n, \qquad (1)$$

где λ_i - среднее число мутаций в локусе i.

Вероятность того, что в гаплотипе мутаций не произошло, равна

$$P(0) = e^{-\lambda}.$$

В реальности мы наблюдаем не только гаплотипы без мутаций, но и гаплотипы с четным числом возвратных мутаций, приводящих в итоге к $m = 0$.

Наблюдаемая вероятность $P_{набл}(0)$ завышена за счет возвратных мутаций:

$$P_{набл}(0) = P(0) + P_{возвр}(2 \to 0) + P_{возвр}(4 \to 0) + \ldots$$

где
$2 \to 0$ означает, что произошли две взаимно исключающие друг друга мутации,
$4 \to 0$ означает, что в результате четырех последовательных мутаций аллель вернулась в исходное состояние, и т.д. Очевидно, что нечетное число мутаций не может дать вклада в наблюдаемую вероятность.

Вероятность образования в гаплотипе двух мутаций:

$$P(2) = \frac{\lambda^2}{2!} e^{-\lambda}$$

Подставляя в эту формулу выражение (1), получаем:

$$P(2) = \frac{1}{2!} e^{-\lambda} (\lambda_1^2 + \lambda_2^2 + \ldots + \lambda_n^2 + 2(\lambda_1\lambda_2 + \lambda_1\lambda_3 + \ldots + \lambda_1\lambda_n + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \ldots + \lambda_2\lambda_n + \ldots + \lambda_{n-1}\lambda_n))$$

Члены $\lambda_i \lambda_k$, где $i \neq k$, не дают вклада в возвратные мутации, так как описывают две мутации на гаплотип, происшедшие в разных локусах i и k, по одной на каждый локус.

С учетом разложения $\mu_i = \mu_i(+) + \mu_i(-)$ вероятность двух мутаций запишется в следующем виде:

$$P(2) = \frac{1}{2!} e^{-\lambda} (\lambda_1^2(+) + \lambda_1^2(-) + \lambda_2^2(+) + \lambda_2^2(-) + ... + \lambda_n^2(+) + \lambda_n^2(-) +$$
$$2(\lambda_1(+)\lambda_1(-) + \lambda_2(+)\lambda_2(-) + ... + \lambda_n(+)\lambda_n(-)) +$$
члены, не дающие вклада в возвратные мутации),

где

$$\lambda_i(+) = \mu_i(+)T$$
$$\lambda_i(-) = \mu_i(-)T$$

Члены $\lambda_i^2(+)$ описывают две мутации в локусе *i,* увеличившие аллель на два значения,

члены $\lambda_i^2(-)$ описывают две мутации в локусе *i,* уменьшившие аллель на два значения.

Следовательно, они также не дают вклада в возвратные мутации.

Напротив, члены $\lambda_i(+)\lambda_i(-)$ описывают как раз те самые мутации, приводящие к кажущемуся отсутствию мутаций в гаплотипе.

В итоге, получаем следующее выражение для вероятности двух возвратных мутаций:

$$P_{возвр}(2 \to 0) = e^{-\lambda} (\lambda_1(+)\lambda_1(-) + \lambda_2(+)\lambda_2(-) + ... + \lambda_n(+)\lambda_n(-)) . \quad (2)$$

Аналогичным способом выводится выражение для вероятности четырех возвратных мутаций:

$$P_{возвр}(4 \to 0) = e^{-\lambda} (\frac{1}{4} (\lambda_1^2(+)\lambda_1^2(-) + \lambda_2^2(+)\lambda_2^2(-) + ... + \lambda_n^2(+)\lambda_n^2(-)) +$$
$$\lambda_1(+)\lambda_1(-)(\lambda_2(+)\lambda_2(-) + \lambda_3(+)\lambda_3(-) + ... + \lambda_n(+)\lambda_n(-)) +$$
$$\lambda_2(+)\lambda_2(-)(\lambda_3(+)\lambda_3(-) + \lambda_4(+)\lambda_4(-) + ... + \lambda_n(+)\lambda_n(-)) + ... + \quad (3)$$
$$\lambda_{n-1}(+)\lambda_{n-1}(-)\lambda_n(+)\lambda_n(-))$$

Вероятность образования в гаплотипе одной мутации:

$$P(1) = \lambda e^{-\lambda}$$

Как и для случая $m = 0$ наблюдаемая вероятность $P_{набл}(1)$ увеличена за счет возвратных мутаций:

$$P_{набл}(1) = P(1) + P_{возвр}(3 \to 1) + P_{возвр}(5 \to 1) + \ldots$$

Применяя вышеизложенный подход, получаем для вклада возвратных трех мутаций:

$$P_{возвр}(3 \to 1) = e^{-\lambda}(\frac{1}{2}(\lambda_1^2(+)\lambda_1(-) + \lambda_1(+)\lambda_1^2(-) + \lambda_2^2(+)\lambda_2(-) + \lambda_2(+)\lambda_2^2(-) + \ldots + \lambda_n^2(+)\lambda_n(-) + \lambda_n(+)\lambda_n^2(-)) + \lambda_1(+)\lambda_1(-)(\lambda_2 + \lambda_3 + \ldots + \lambda_n) + \lambda_2(+)\lambda_2(-)(\lambda_1 + \lambda_3 + \lambda_4 + \ldots + \lambda_n) + \ldots + \lambda_n(+)\lambda_n(-)(\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_{n-1})) \quad (4)$$

Далее, наблюдаемая вероятность двух мутаций складывается следующим образом:
1) из вероятности возникновения в гаплотипе двух мутаций исключается вероятность возвратной мутации $2 \to 0$ (она не наблюдается);
2) добавляются вероятности возвратных мутаций за счет мутаций с кратностью 4, 6 и т.д.

$$P_{набл}(2) = P(2) - P_{возвр}(2 \to 0) + P_{возвр}(4 \to 2) + P_{возвр}(6 \to 2) + \ldots$$

Аналогично, для наблюдаемой вероятности трех мутаций запишем:

$$P_{набл}(3) = P(3) - P_{возвр}(3 \to 1) + P_{возвр}(5 \to 3) + P_{возвр}(7 \to 3) + \ldots$$

Подобные формулы можно привести для любого количества мутаций в гаплотипе.

Более строгий вывод формул (2-4) производится из многомерного распределения Пуассона.

Приложение к частному случаю

Рассмотрим частный случай симметричного дерева мутаций, полезный для проведения оценки верхнего предела вклада возвратных мутаций. Симметричное дерево мутаций означает, что скорости мутаций $\mu_i(+)$ и $\mu_i(-)$ для локуса *i* равны друг другу:

$$\mu_i(+) = \mu_i(-) = \frac{\mu_i}{2}.$$

В этом случае вероятность возвратных мутаций максимальна. Поэтому этот частный случай может быть использован для оценки верхнего предела уровня возвратных мутаций.

Для упрощения формул предположим, что скорости мутаций во всех локусах одинаковы:

$$\mu_1 = \mu_2 = ... = \mu_n = \frac{\mu}{n}.$$

Из общих формул, приведенных выше, получаем:

$$P_{набл}(0) = e^{-\lambda}(1 + \frac{\lambda^2}{4n} + \frac{2n-1}{64n^3}\lambda^4 + \frac{6n^2-9n+4}{2304n^5}\lambda^6 + ...), \quad (5)$$

$$P_{набл}(1) = \lambda e^{-\lambda}(1 + \frac{2n-1}{8n^2}\lambda^2 + \frac{6n^2-9n+4}{192n^4}\lambda^4 + ...), \quad (6)$$

$$P_{набл}(2) = \frac{\lambda^2}{2}e^{-\lambda}(1 - \frac{1}{2n} + \frac{6n^2-9n+4}{24n^3}\lambda^2 + \frac{24n^3-72n^2+82n-33}{768n^5}\lambda^4 + ...), \quad (7)$$

$$P_{набл}(3) = \frac{\lambda^3}{6}e^{-\lambda}(1 - \frac{3(2n-1)}{4n^2} + \frac{16n^3-48n^2+56n-23}{64n^4}\lambda^2 + ...), \quad (8)$$

$$P_{набл}(4) = \frac{\lambda^4}{24}e^{-\lambda}(1 - \frac{24n^2-30n+13}{8n^3} + \frac{40n^4-200n^3+430n^2-435n+166}{160n^5}\lambda^2 + ...), \quad (9)$$

$$P_{набл}(5) = \frac{\lambda^5}{120}e^{-\lambda}(1 - \frac{80n^3-180n^2+190n-75}{16n^4} + ...), \quad (10)$$

$$P_{набл}(6) = \frac{\lambda^6}{720}e^{-\lambda}(1 - \frac{240n^4-840n^3+1560n^2-1470n+541}{32n^5} + ...). \quad (11)$$

Формулы (5-11) применимы для среднего числа мутаций на гаплотип до 1.4 ($\lambda \le 1.4$) с точностью не хуже 0.1 %. 1.4 мутации на 12-маркерный гаплотип при линейной модели соответствуют 64-м поколениям (1600 лет до общего предка), а при учете возвратных мутаций соответствуют 68 поколениям (1700 лет до общего предка).

Эти формулы применимы также и для случая отдельного локуса (n=1), когда удается получить точные формулы, позволяющие рассчитать вероятность возвратных мутаций для любых значений наблюдаемых мутаций (m):

$P(m \to 0) = (p(\frac{m}{2}, \frac{\lambda}{2}))^2$, $m \geq 0$ четное.

$P(m \to k) = 2p(\frac{m+k}{2}, \frac{\lambda}{2}) p(\frac{m-k}{2}, \frac{\lambda}{2})$, $m \geq k > 0$ одновременно четные или нечетные.

Здесь функция $p(m,\lambda)$ – распределение Пуассона:

$$p(m, \lambda) = \frac{\lambda^m}{m!} e^{-\lambda}.$$

Из формул (5-11) можно получить оценку влияния возвратных мутаций. Результаты расчетов приведены конце в виде таблиц соответствия наблюдаемого времени $T_{набл}$ и теоретического времени T для часто используемых гаплотипов с 6, 12, 25 и 37 маркерами, а также для отдельного локуса. Удельные скорости мутаций, использованные в расчетах, следующие (Клёсов, 2008):

6-маркерный гаплотип – 0.0016 на поколение на маркер;
12-маркерный гаплотип – 0.0018 на поколение на маркер;
25-маркерный гаплотип – 0.0018 на поколение на маркер;
37-маркерный гаплотип – 0.00243 на поколение на маркер;
отдельный маркер – 0.002 на поколение.

Среднее число наблюдаемых мутаций $\lambda_{набл} = \mu T_{набл}$ равно:

$$\lambda_{набл} = 1P_{набл}(1) + 2P_{набл}(2) + 3P_{набл}(3) + ... \qquad (12)$$

Подставляя в формулу (12) значения $P_{набл}(m)$, можно получить следующую аппроксимационную формулу:

$$\lambda_{набл} = \frac{\lambda}{2}(1 + e^{-\frac{\lambda}{n}}) \qquad (13)$$

Для гаплотипов с числом маркеров шесть и выше ($n \geq 6$) точность формулы (13) не хуже 0.2 % при $\lambda \leq 1.4$. Для одиночных локусов точность аппроксимации (13) не хуже 1 % при $\lambda \leq 0.8$.

Из формулы (9) следует, что величина поправки

$$\frac{\lambda_{набл}}{\lambda} = \frac{1}{2}(1 + e^{-\frac{\lambda}{n}})$$

зависит только от удельного количества мутаций на один маркер $\frac{\lambda}{n}$.

Отсюда можно сделать один весьма важный практический вывод: поправка на возвратные мутации на гаплотип из *n* маркеров и количеством мутаций λ такая же, как и поправка на возвратные мутации на отдельный локус с числом мутаций $\frac{\lambda}{n}$.

Из структуры формул (5-11) можно предположить, что эта закономерность носит универсальный характер и может применяться к значениям $\frac{\lambda}{n} \leq 0.8$. С этой точки зрения для оценки верхнего предела поправки на возвратные мутации в гаплотипе достаточно одной таблицы для одиночного локуса с соответствующим значением удельной скорости мутаций на маркер.

Литература

Березин, И.В. и Клёсов, А.А. Практический курс химической и ферментативной кинетики. Изд-во Московского университета, 1976, 320 стр.

Каржавин, С. Математический анализ возможности определения исторического времени жизни основателя генеалогического древа по гаплотипам его мужских потомков», Вестник Российской Академии ДНК-генеалогии, 1, №4, сентябрь 2008.

Клёсов, А.А. Основные положения ДНК-генеалогии. Вестник Российской Академии ДНК-генеалогии. 1, № 2, 252-348 (2008).

Gusmão, L., Sanchez-Diz, P., Calafell, F., Martin, P., Alonso, C.A. et al. Mutation Rates at Y Chromosome Specific Microsatellites. Human Mutation, 26, 520-528 (2005).

ПРИЛОЖЕНИЯ

В первом столбце таблиц приводится наблюдаемый возраст популяции с интервалом в 10 поколений. Остальные столбцы соответствуют единицам поколений. Например, соответствие для возраста 26 поколений надо искать на пересечении строки 20 и столбца 6, для возраста 50 – пересечение строки 50 и столбца 0.

Теоретический возраст рассчитан для симметричного дерева мутаций и одинаковой (средней) скорости мутаций во всех локусах гаплотипа.

Таблица 1.

Соответствие возраста популяции, рассчитанного из общего числа мутаций на 6-маркерный гаплотип (DYS 19, 388, 390, 393, 392, 393), теоретическому возрасту (число поколений). Рассчитано для скорости мутаций 0.0016 на маркер на поколение (типичная средняя скорость мутаций для 6-маркерных гаплотипов).

	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
10	10	11	12	13	14	15	16	17	18	19
20	20	21	22	23	24	25	27	28	29	30
30	31	32	33	34	35	36	37	38	39	40
40	41	42	43	44	46	47	48	49	50	51
50	52	53	54	55	56	57	59	60	61	62
60	63	64	65	66	67	69	70	71	72	73
70	74	75	76	77	79	80	81	82	83	84
80	85	87	88	89	90	91	92	93	95	96
90	97	98	99	100	102	103	104	105	106	107
100	109	110	111	112	113	115	116	117	118	119
110	121	122	123	124	125	127	128	129	130	132
120	133	134	135	136	138	139	140	141	143	144
130	145									

Таблица 2.

Соответствие возраста популяции, рассчитанного из общего числа мутаций на 12-маркерный гаплотип, теоретическому возрасту

(поколений). Скорость мутаций 0.0018 на маркер на поколение (типичная средняя скорость мутаций для 12-маркерных гаплотипов в системе записи FTDNA).

	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
10	10	11	12	13	14	15	16	17	18	19
20	20	21	22	23	25	26	27	28	29	30
30	31	32	33	34	35	36	37	38	39	40
40	41	43	44	45	46	47	48	49	50	51
50	52	53	55	56	57	58	59	60	61	62
60	63	65								

Таблица 3.

Соответствие возраста популяции, рассчитанного из общего числа мутаций на 25-маркерный гаплотип, теоретическому возрасту (поколений). Скорость мутаций 0.0018 на маркер на поколение (типичная средняя скорость мутаций для 25-маркерных гаплотипов в системе записи FTDNA).

	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
10	10	11	12	13	14	15	16	17	18	19
20	20	21	22	23	25	26	27	28	29	30

Таблица 4.

Соответствие возраста популяции, рассчитанного из общего числа мутаций на 37-маркерный гаплотип, теоретическому возрасту (поколений). Скорость мутаций 0.00243 на маркер на поколение (типичная средняя скорость мутаций для 37-маркерных гаплотипов в системе записи FTDNA).

	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
10	10	11	12	13	14	15				

Таблица 5.

Первая колонка – среднее число мутаций на маркер,
вторая колонка – число поколений до общего предка выборки, рассчитанное для линейной модели ($t = n/N/\mu$, где t – время до общего предка, в поколениях, n – количество мутаций во всех N гаплотипах выборки, μ – средняя скорость (частота) мутаций, выраженная в числе мутаций на гаплотип на поколение, принятая в данном случае 0.002),
третья колонка – число поколений до общего предка выборки с учетом возвратных мутаций.
Условия расчета: один общий предок для всей выборки, дерево мутаций симметричное.

$\lambda_{набл}$	$T_{набл}$	$T_{пред}$
0.01	5	5
0.02	10	10
0.03	15	15
0.04	20	20
0.05	25	26
0.06	30	31
0.07	35	36
0.08	40	42
0.09	45	47
0.10	50	53
0.11	55	58
0.12	60	64
0.13	65	69
0.14	70	75
0.15	75	81
0.16	80	87
0.17	85	93
0.18	90	99
0.19	95	105
0.20	100	111
0.21	105	117
0.22	110	123
0.23	115	130
0.24	120	136
0.25	125	143
0.26	130	149
0.27	135	156
0.28	140	163

$\lambda_{набл}$	$T_{набл}$	$T_{пред}$
0.41	205	257
0.42	210	265
0.43	215	273
0.44	220	281
0.45	225	289
0.46	230	297
0.47	235	306
0.48	240	314
0.49	245	323
0.50	250	331
0.51	255	340
0.52	260	349
0.53	265	358
0.54	270	367
0.55	275	376
0.56	280	385
0.57	285	395
0.58	290	404
0.59	295	414
0.60	300	424
0.61	305	434
0.62	310	444
0.63	315	454
0.64	320	464
0.65	325	474
0.66	330	485
0.67	335	496
0.68	340	506

$\lambda_{набл}$	$T_{набл}$	$T_{пред}$
0.81	405	659
0.82	410	671
0.83	415	684
0.84	420	697
0.85	425	710
0.86	430	724
0.87	435	737
0.88	440	750
0.89	445	764
0.90	450	778
0.91	455	792
0.92	460	806
0.93	465	820
0.94	470	835
0.95	475	850
0.96	480	864
0.97	485	879
0.98	490	894
0.99	495	910
1.00	500	925
1.01	505	940
1.02	510	956
1.03	515	972
1.04	520	988
1.05	525	1004
1.06	530	1020
1.07	535	1037
1.08	540	1054

0.29	145	169
0.30	150	176
0.31	155	183
0.32	160	190
0.33	165	197
0.34	170	204
0.35	175	212
0.36	180	219
0.37	185	226
0.38	190	234
0.39	195	241
0.40	200	249

0.69	345	517
0.70	350	528
0.71	355	539
0.72	360	551
0.73	365	562
0.74	370	574
0.75	375	585
0.76	380	597
0.77	385	609
0.78	390	621
0.79	395	634
0.80	400	646

1.09	545	1070
1.10	550	1087
1.11	555	1104
1.12	560	1122
1.13	565	1139
1.14	570	1157
1.15	575	1174
1.16	580	1192
1.17	585	1210
1.18	590	1229
1.19	595	1247
1.20	600	1266

$\lambda_{набл}$	$T_{набл}$	$T_{пред}$
1.3	650	1460
1.4	700	1671
1.5	750	1898
1.6	800	2140
1.7	850	2399
1.8	900	2673
1.9	950	2963
2.0	1000	3269
2.1	1050	3591
2.2	1100	3929
2.3	1150	4282
2.4	1200	4651
2.5	1250	5035
2.6	1300	5436
2.7	1350	5852
2.8	1400	6284
2.9	1450	6731
3.0	1500	7195

Расчет возраста популяции якутов, принадлежащих к гаплогруппе N1c1

Д.С. Адамов

nimissin@mail.ru

Предположение о пуассоновском распределения мутаций в STR локусах Y-хромосомы позволило получить непротиворечивую оценку времени жизни предка якутов по мужской линии с гаплогруппой N1c1: 1300 ± 200 лет назад.

Введение.

Исследования Y-хромосомы якутов показали, что гаплотипы подавляющего большинства мужчин принадлежат к гаплогруппе N1c1 по новой классификации (N3a по старой). По данным Бригитты Пакендорф и др. [1] доля популяции якутов с гаплогруппой N1c1 составляет 94 %, в работе В.Н.Харькова и др. [2] получен показатель 89 %. По этой гаплогруппе наблюдается ярко выраженный эффект основателя. Вместе с тем, оценки временного интервала, отделяющего наше время от времени жизни предка, противоречивы. Б.Пакендорф и соавторы оценивают время жизни основателя в 880 ± 440 лет назад [1], томская группа [2] - в 4450 ± 1960 лет назад.

Метод.

Для расчета использовались данные по гаплотипам Y-хромосомы якутов, опубликованные в работе Б.Пакендорф и др. [1] и имеющиеся в широком доступе. Данные приведены в Приложении 1. Гаплотипы определялись по 9 маркерам STR локусов: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a и DYS385b. Гаплогруппа N1c1 была выявлена у 172 человек, распределившихся по 34 гаплотипам.

В работе томской группы [2] был определен предковый гаплотип якутов с гаплогруппой N1c1 по 7 маркерам. В приведенной выше последовательности маркеров он выглядит следующим образом: 14-14-18-23-11-16-14-X-Y. В выборке Б.Пакендорф и др. [1] у 167 человек из 172 число повторов по маркеру DYS385a равно 11, мутаций всего 5. По маркеру DYS385b у 163 человек 13 повторов, мутаций 9. Исходя из этого можно определить

предковый гаплотип по всем 9 маркерам: 14-14-32-23-11-16-14-11-13. К нему относится 74 человека, остальные гаплотипы образовались в результате мутаций в исследованных локусах.

Используемый метод впервые был описан в работах А.А.Клесова по основам ДНК-генеалогии [3]. Известно, что мутации в STR локусах происходят редко. Кроме того, можно считать, что каждая мутация появляется независимо от других мутаций. На основании этих соображений предположим, что наблюдаемое число мутаций (отклонений от предкового числа повторов) в каждом локусе описывается распределением Пуассона. Важнейшим свойством распределения Пуассона является следующее: сумма пуассоновских случайных величин также имеет распределение Пуассона. В нашем случае это означает, что суммарное число мутаций на гаплотип в целом также описывается распределением Пуассона:

$$P(m) = \frac{\lambda^m}{m!} e^{-\lambda}$$

где m – число мутаций на гаплотип, $\lambda = \mu T$, μ – суммарная скорость мутаций в локусах гаплотипа, T – число поколений с момента появления основателя рода.

Рассмотрим 2 частных случая.

1) Вероятность сохранения предкового гаплотипа:

$$P(0) = e^{-\mu T}$$

С другой стороны, $P(0) = N(0)/N_{общ}$,

где $N(0)$ – наблюдаемое число мужчин с предковым гаплотипом, $N_{общ}$ – общая численность обследованных. Тогда

$$T = \frac{\ln \frac{N_{общ}}{N(0)}}{\mu} \qquad (1)$$

Эта формула была впервые получена А.А.Клесовым из уравнения кинетики первого порядка [3].

2) Общее количество мутаций по всей выборке равно:

$$M = N_{общ} \ (1P(1) + 2P(2) + 3P(3) + \ldots) = N_{общ} (1\frac{\lambda^1}{1!}e^{-\lambda} + 2\frac{\lambda^2}{2!}e^{-\lambda} + 3\frac{\lambda^3}{3!}e^{-\lambda} + \ldots) = N_{общ}\ \mu T$$

Отсюда $T = \dfrac{M}{\mu N_{общ}}$ (2)

Формула (2) широко применяется для расчета возраста популяций [3].

Критерием того, что исследуемая выборка действительно восходит к одному предку, является соответствие между собой возрастов, вычисленных по формулам (1) и (2) (более подробно см. в работе А.А.Клесова [3]). Заметим, что предположение о пуассоновском характере наблюдаемого распределения удовлетворяет этому критерию.

Из данных Б.Пакендорф и др. [1] получается следующее распределение по числу мутаций на 9-маркерный гаплотип:

Таблица 1.

Число мутаций на гаплотип *m*	Распределение гаплотипов по мутациям *N(m)*, чел.	Нормированное распределение *P(m)*	Общее количество мутаций
0 (предковый)	74	0.430	0
1	51	0.297	51
2	33	0.192	66
3	9	0.052	27
4	2	0.012	8
5	1	0.006	5
6	2	0.012	12
Итого:	172	1	169

$N_{общ}$ = 172, *N(0)* = 74, *M* = 169.

Суммарную скорость мутаций по исследованным локусам определим из данных Джона Чандлера [4]: μ = 0.0174 на поколение в 25 лет.

Можно предположить, что наилучшая оценка возраста популяции достигается путем аппроксимации наблюдаемого распределения пуассоновским

распределением по методу наименьших квадратов, так как в оценке будет использоваться вся имеющаяся информация по мутациям.

Результаты.

Расчет по формуле (1) дает значение:

T = 48 поколений, с учетом погрешностей 1200 ± 300 лет.

Расчет по формуле (2):

T = 56 поколений, с учетом погрешностей 1400 ± 300 лет.

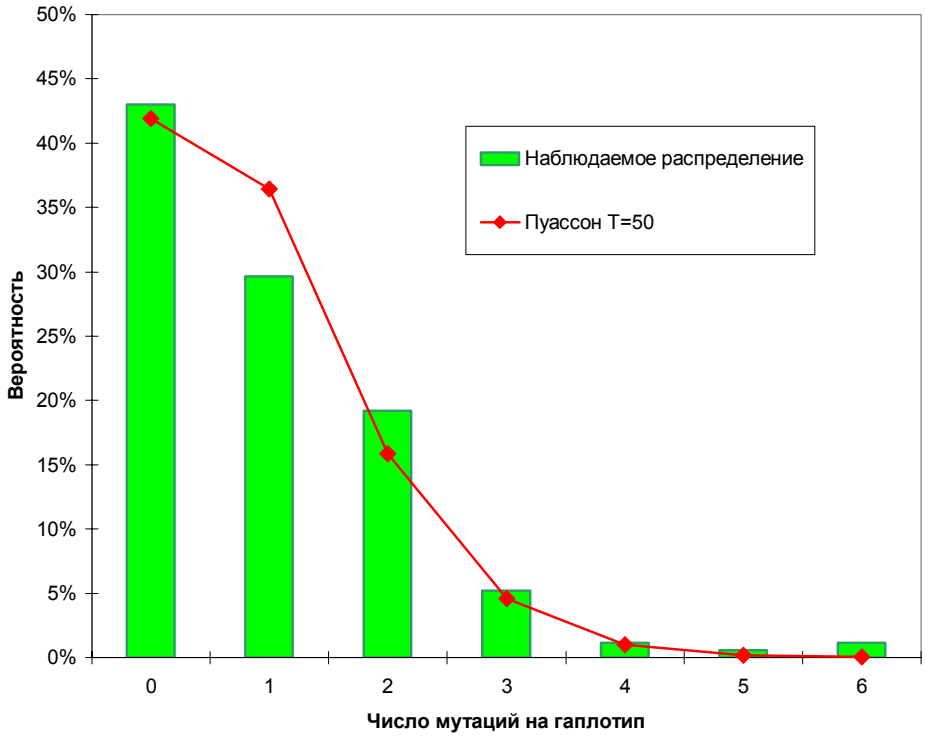
Полученные величины близки друг к другу в пределах статистической погрешности. Это позволяет сделать вывод о том, что исследуемая выборка не противоречит предположению о наличии единого предка. Некоторое сомнение вызывают гаплотипы № 32 и № 35 с шестью мутациями (см. Приложение 1), у которых DYS389II = 16 (в предковом гаплотипе 18) и DYS392 = 14 (в предковом гаплотипе 16). Без учета этих гаплотипов, дающих 12 мутаций, оценка возраста по формуле (2) снижается до 52 поколений.

Тем не менее, для дальнейшего анализа будем использовать все гаплотипы. На рисунке 1 приведено наблюдаемое распределение числа мутаций на гаплотип. Оно было аппроксимировано пуассоновским распределением с различными значениями возраста T. По методу наименьших квадратов наилучшая аппроксимация наблюдаемого распределения достигается при T = 50 поколений.

Отсюда время жизни предка якутов по мужской линии с гаплогруппой N1c1 составляет 1300 ± 200 лет назад, или 500 - 900 г. н.э.

Оценка возраста популяции Б.Пакендорф и др. [1] 880 ± 440 лет была произведена с применением скорости мутации 0.003/локус/поколение, которая в 1.6 раза выше скоростей Чандлера [4]: 0.0174/9 = 0.0019. С учетом поправки на скорость мутаций получаем 880*1.6 = 1400 лет. Погрешность примененного в работе [1] метода не обсуждаем.

Рис.1.



Аналогичным образом объясняется расхождение полученного времени жизни с оценками томской группы 4450 ± 1960 лет по методу Животовского и др. [5] и 3980 ± 2570 лет по методу Форстера [6]. В работе [2] в обоих способах расчетов была использована т.н. «скорость Животовского» 0.00069/локус за 25 лет [5], которая в 2.6 раза ниже скорости, определенной по Чандлеру [4] для 7-маркерного гаплотипа DYS19-DYS389I-DYS389II- DYS390-DYS391-DYS392-DYS393 (0.0128/7 = 0.0018). Поправка на скорость мутаций существенно снижает возраст: 4450*0.00069/0.0018 = 1700 лет и 3980*0.00069/0.0018 = 1500 лет. С учетом погрешностей эти оценки не противоречат результатам настоящей работы. Кроме того, авторы работы [2] указывают, что около 55 % выборки якутов относится к предковому гаплотипу. Применяя формулу (2), получаем оценку времени жизни предка:

$T = ln(1/0.55)/0.0128 = 47$ поколений, или 1200 ± 400 лет назад.

Эта оценка также не противоречит нашим расчетам.

Число носителей предкового гаплотипа может быть завышено из-за возвратных мутаций (гомоплазии) [3,7]. Оценка верхнего предела уровня возвратных мутаций показывает, что наблюдаемый возраст исследуемой популяции якутов может быть занижен не более, чем на 3 поколения по сравнению с истинным возрастом. Эта поправка значительно меньше погрешностей использованного метода.

Любопытно отметить, что в работе Б.Пакендорф и др. [1] был обнаружен эффект основателя и по митохондриальной гаплогруппе D5a. В настоящее время к этой субгаплогруппе относят 15.7 % якутов [1]. Возраст этой популяции авторы работы [1] определяют в 1286 ± 800 лет! Авторы работы [8] подтверждают эффект основателя по гаплогруппе D5a, их датировка 1630 ± 997 лет.

Таким образом, вся имеющаяся в настоящее время информация по якутам с гаплогруппой N1c1 не противоречит оценке возраста популяции в 1300 ± 200 лет.

Обсуждение.

Расчеты для различных европейских ветвей гаплогруппы N1c1 были проведены А.А.Клесовым [3].

1. Прибалтийские народы. Из 249 гаплотипов была выделена большая ветвь из 126 человек с возрастом 1000-1250 лет. Остальные 123 гаплотипа отнесены к 9 другим ветвям с общим предком, жившим около 2750 лет назад.

2. Русские. Был проведен анализ 37 гаплотипов из работы Деренко М. и др. [9]. Обнаружено 2 молодых ветви, 7 и 9 гаплотипов, с возрастами 225 и 500 лет. По расстоянию между ветвями определено, что общий предок жил 3000 лет назад. В самой древней ветви 8 гаплотипов, ее возраст оценивается в 2500 лет.

3. Швеция. 72 гаплотипа, относящиеся к гаплогруппе N1c, имеют разных предков. Для 18 гаплотипов базовый гаплотип идентичен прибалтийскому для гаплогруппы N1c1: 14-24-14-11-11-13-11-12-10-14-14-30 (в стандартной последовательности Family Tree DNA для первой панели). Возраст этой ветви 1000-1250 лет.

Таким образом, ко времени появления предка якутов представители гаплогруппы N1c1 расселились по всем северу Евразии со своей первоначальной родины, предположительно северного Китая или Монголии.

В хронологическом плане появление около 1300 лет назад предка якутов с гаплогруппой N1c1 относится к времени Первого и Второго тюркских каганатов (551 – 745 годы) и Уйгурского каганата (745 – 840 годы) [10]. Обзор гипотез о происхождении якутов (саха) был недавно составлен В.В.Ушницким [11]. Одной из наиболее старых и разработанных гипотез является т.н. "курыканская".

Из надписи Бильге-кагана известно, что уч-курыканы (гулигань китайских хроник) присылали своих послов на похороны Бумын-кагана в 552 году и/или его брата Истеми-кагана (ок. 576 г.). Курыканы перечисляются в числе врагов Эльтерес-кагана, отца Бильге-кагана (конец VII века). На похоронах Кюль-тегина в 731 году уч-курыканы уже не упоминаются [10]. Эльтерес-каган 47 раз ходил с войском в поход и дал 20 сражений своим врагам. Можно предположить, что именно в это время произошел отбор предков якутов через "бутылочное горлышко популяции" [3]. Заметим, что возраст популяции и время переселения предков якутов на свою современную родину не обязательно должны совпадать, хотя вероятность людских потерь при переселениях в те времена могла быть высокой.

Выводы.

Предположение о пуассоновском распределения мутаций в локусах Y-хромосомы в рамках общего метода, развитого А.А.Клесовым в работе [3], позволило получить непротиворечивую оценку возраста популяции якутов с гаплогруппой N1c1: 1300 ± 200 лет. Возраст популяции согласуется с гипотезой происхождения якутов от курыкан.

Автор выражает глубокую благодарность Анатолию Алексеевичу Клесову и Владимиру Геннадьевичу Волкову за поддержку и плодотворные обсуждения темы.

Литература

1. Pakendorf B. et al. Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. Hum.Genet., 2006, 120: 334-353.

2. Харьков В.Н. и др. Происхождение якутов: анализ гаплотипов Y-хромосомы. Молекулярная биология, 2008, том 42, № 2, с.226-237.

3. Клесов А.А. Основные положения ДНК-генеалогии. Вестник Российской Академии ДНК-генеалогии. 2008, том 1, № 2, с.252-348.

4. Chandler J.F. Estimating Per-Locus Mutation Rates. J. Genet. Genealogy, 2006, 2: 27-33.

5. Zhivotovsky L.A. et al. The Effective Mutation Rate at Y Chromosome Short Tandem Repeats, with Application to Human Population-Divergence Time. Am.J.Hum.Genet., 2004, 74: 50-61.

6. Forster P. et al. A short tandem repeat-based phylogeny for the human Y-chromosome. Am.J.Hum.Genet., 2000, 67: 182-196.

7. Адамов Д.С., Клесов А.А. Теоретическая и практическая оценка возвратных мутаций в гаплотипах Y-хромосомы. Вестник Российской Академии ДНК-генеалогии. В печати.

8. Zlojutro M. et al. The Origins of the Yakut People: Evidence from Mitochondrial DNA Diversity. Int.J.Hum.Genet., 2008, 8(1-2): 119-130.

9. Derenko M. et al. Y-chromosome haplogroup N dispersals from south Siberia to Europe. J.Hum.Genet., 2007, 52: 763-770.

10. Гумилев Л.Н. Древние тюрки. М.: Айрис-пресс, 2002, 560 с.

11. Ушницкий В.В. Историографический обзор проблемы происхождения саха. http://www.ethnonet.ru/ru/pub/11-04-2008.html.

Приложение 1

Гаплотипы Y-хромосомы якутов (гаплогруппа N1c1)
из работы Pakendorf B. et al. Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. Hum.Genet., 2006, 120: 334-353.

№ гаплотипа из работы [1]	DYS19[1)]	DYS389I[2)]	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385a	DYS385b	Кол-во людей	Мутаций на гаплотип	Число мутаций на выборку
1	14	14	18	23	11	16	14	11	13	***74***	0	**0**
2	14	14	17	23	11	16	14	11	13	***36***	1	**36**
3	14	14	17	23	11	15	14	11	13	***17***	2	**34**
5	14	14	18	23	11	16	14	11	12	***2***	1	**2**
6	14	14	18	23	10	16	14	11	13	***3***	1	**3**
7	14	15	18	23	11	16	14	11	13	***1***	1	**1**
8	15	14	18	23	11	16	14	11	13	***1***	1	**1**
9	14	14	18	23	11	17	14	11	13	***1***	1	**1**
10	14	14	18	23	11	16	13	11	13	***1***	1	**1**
11	14	14	18	22	11	16	14	11	13	***1***	1	**1**
12	14	14	18	23	11	16	14	11	14	***1***	1	**1**
13	14	14	18	23	11	15	14	11	13	***1***	1	**1**
14	14	14	19	23	11	16	14	11	13	***1***	1	**1**
16	14	14	18	23	11	16	14	12	13	***2***	1	**2**
17	14	14	17	23	11	16	14	11	12	***1***	2	**2**
18	14	14	17	23	10	16	14	11	13	***1***	2	**2**
19	14	14	17	23	11	16	13	11	13	***2***	2	**4**
20	14	14	17	23	11	16	14	11	14	***1***	2	**2**
21	14	14	16	23	11	16	14	11	13	***2***	2	**4**
22	14	14	17	24	11	16	14	11	13	***3***	2	**6**
23	14	15	17	23	11	15	14	11	13	***1***	3	**3**
24	14	14	17	23	11	15	14	12	13	***2***	3	**6**
25	14	13	17	23	11	15	14	11	13	***1***	3	**3**
26	14	14	17	23	11	15	13	11	13	***2***	3	**6**
28	14	13	18	23	11	16	14	11	14	***2***	2	**4**
29	14	14	16	23	10	15	14	11	13	***2***	4	**8**

30	14	14	17	23	10	15	14	11	13	***3***	3	**9**
31	14	14	18	23	12	16	14	11	12	***1***	2	**2**
32	14	13	16	23	11	14	15	11	13	***1***	6	**6**
33	15	15	17	23	11	15	13	11	13	***1***	5	**5**
34	14	14	18	23	11	15	13	11	13	***1***	2	**2**
35	14	14	16	22	10	14	14	11	13	***1***	6	**6**
36	14	14	18	23	11	14	14	11	13	***1***	2	**2**
37	14	14	18	23	11	16	14	12	14	***1***	2	**2**
	DYS19[1)]	**DYS389I**[2)]	**DYS389II**	**DYS390**	**DYS391**	**DYS392**	**DYS393**	**DYS385a**	**DYS385b**	**Кол-во людей**		**Число мутаций на выборку**
Мутаций на локус	**2**	**7**	**84**	**5**	**11**	**38**	**8**	**5**	**9**	***172***		**169**

Примечания:
[1)] Известен также как **DYS394.**
[2)] С учетом трехкопийного TCTG повтора [2]

Каржавин С.П.
E-mail: karzhavin@inbox.ru

«Математический анализ возможности определения исторического времени жизни основателя генеалогического древа по гаплотипам его мужских потомков»

Резюме

1. Разработана математическая модель изменения в результате мутаций количества повторов в маркерах (локусах) гаплотипа в непрерывной мужской генеалогической линии от первопредка до современного потомка. Изменение количества повторов в маркерах представлено в виде целочисленного осциллирующего блуждания с возможностью изменения на +1, -1, +2 или -2 в результате мутации. Поток мутаций на генеалогической линии от первопредка до финального потомка представлен пуассоновским потоком.

2. Проведена проверка адекватности разработанной математической модели и реальных данных (наборы гаплотипов из пяти гаплогрупп I1a, J2, R1b, R1b1, R1b1c в количестве от 1300 до 7770 гаплотипов в каждом наборе). Степень адекватности модели и реальности определялась по критерию хи-квадрат в процессе сравнения гистограмм повторов в маркерах совокупности реальных гаплотипов и гистограмм теоретических распределений. Модель показала очень высокую степень адекватности.

3. На основе разработанной математической модели был синтезирован и программно реализован алгоритм имитации совокупности гаплотипов финальных потомков, который использовался для статистических расчетов точностей оценки количества мутаций, а также для верификации алгоритмов вычисления количества мутаций, осуществившихся со времени жизни первопредка.

4. Синтезированы и программно реализованы статистические алгоритмы вычисления по совокупности гаплотипов максимально правдоподобных оценок реально осуществившихся мутаций в каждом маркере (с учетом возвратных мутаций) и реконструкции гаплотипа первопредка.

5. Создана методика определения точности оценки времени до первопредка по набору гаплотипов исследуемой гаплогруппы, а также рассчитаны соответствующие таблицы доверительных интервалов для этого времени.

Содержание

Введение

В данном исследовании рассмотрены вопросы оценки количества мутаций, которые возникают в локусах (маркерах) нерекомбинирующего сегмента Y-хромосомы.

Упрощенно каждый маркер состоит из несколько раз повторяющейся собственной специфической комбинации нуклеотидов, причем, количество повторов в каждом маркере разное и лежит в пределах от 9 до 40. В качестве примера можно привести квадруплет нуклеотидов ТАГА (тинин-аденин-гуанин-аденин), или триплет ТАТ (тинин-аденин-тинин), и т.д. Значения количества повторов в маркерах меняются в достаточно ограниченных интервалах значений. Например, маркер DYS#19 содержит от 9 до 20, маркер DYS#390 – от 19 до 29, маркер DYS#391 – от 7 до 14, маркер DYS#388 – от 9 до 19. Заметим, что это не абсолютные границы, но 99% всех значений в маркерах в них укладываются.

Если бы не было мутаций, то во всех поколениях и у всех потомков данная хромосома имела бы один и тот же состав и порядок нуклеотидов. На самом же деле в момент возникновения хромосомы потомка иногда возникает мутация, которая изменяет в отдельных локусах (маркерах) количество повторов соответствующей комбинации нуклеотидов. Измененное количество повторов в маркере передается дальше потомкам, пока очередная мутация не изменит это состояние на новое.

Чем больше прошло поколений, тем больше из-за прошедших мутаций начинают различаться по количеству повторов одни и те же маркеры у разных потомков. Очевидно, что чем больше различие в количестве повторов у потомков, тем древнее их общий прародитель. Следовательно, если удастся оценить среднее количество мутаций, прошедшее от прародителя до последнего потомка, то можно вычислить и количество поколений, отделяющее прародителя и его современных потомков. Почему именно среднее количество мутаций? Дело в том, что на каждой генеалогической линии (определенной последовательности потомков, соединяющей прародителя и конкретного последнего потомка) количество осуществившихся мутаций разное. Эта разница возникает из стохастической природы факторов, вызывающих мутации.

Следует сказать еще об одном виде редко осуществляющихся мутаций на другом участке Y-хромосомы, которые «навечно» закрепляются за всеми

последующими потомками. Именно по наличию таких специфических мутаций производится отбор потомков, у которых гарантированно один общий прародитель, у которого данная мутация и произошла. Следовательно, в дальнейших исследованиях мы всегда будем иметь дело с выборками данных, гарантированно взятых у потомков единого прародителя.

Нам не важны ни конкретные комбинации нуклеотидов в маркерах, ни физическая природа, приводящая к мутациям. Нас будет интересовать только изменение количества повторов комбинаций нуклеотидов в результате мутации. Одним словом, каждый маркер будем описывать лишь целым неотрицательным числом, соответствующим количеству повторов собственной уникальной комбинации нуклеотидов.

В настоящее время у каждого индивида в лабораториях анализируется не один маркер, а группа – от 6 до 67. Поэтому далее под индивидом будем понимать лишь упорядоченный набор чисел (по одному на каждый маркер), соответствующих количеству повторов комбинаций нуклеотидов в маркерах. Этот числовой набор далее называется гаплотипом и является математическим образом генома конкретного индивида мужского пола.

Пример совокупности 12-ти маркерных гаплотипов, соответствующих восьми мужчинам представлен ниже (сверху даны обозначения маркеров)

DYS# 393	DYS# 390	DYS# 19	DYS# 391	DYS# 385a	DYS# 385b	DYS# 426	DYS# 388	DYS# 439	DYS# 389/1	DYS# 392	DYS# 389/2
12	23	16	11	11	15	12	12	10	12	11	30
13	23	15	11	9	15	12	12	10	13	11	29
12	23	15	11	11	14	12	11	10	12	11	30
12	24	16	10	11	15	11	12	10	13	10	28
12	23	16	11	11	15	12	12	10	13	11	30
12	23	16	11	11	14	12	12	11	14	11	30
12	23	16	11	10	16	12	12	10	13	11	31
12	22	14	11	11	15	10	12	10	14	12	30

Из таблицы видно, что в 1-м маркере (1-й столбец) количество повторов в разных гаплотипах меняется в пределах некоего среднего значения 12, во 2-м маркере (2-й столбец) — в пределах 23-х, в третьем маркере — в пределах 15-ти, и так далее. В дальнейшем именно такие серии чисел по каждому маркеру (столбцы таблицы) и будут служить основой для статистического анализа.

Теперь можно изложить постановку задачи: по достаточно большой совокупности гаплотипов требуется вычислить среднее количество мутаций,

которое прошло от первопредка до настоящего времени, а также определить количество повторов в каждом маркере, которое было у первопредка, т.е., «восстановить» его гаплотип. Второстепенный, но вполне практический интерес представляет также необходимое количество гаплотипов, которое нужно использовать для получения результатов с приемлемой точностью.

Анализируемую совокупность гаплотипов представим в виде следующей числовой матрицы:

$$\begin{matrix} m_1(1) & m_1(2) & .... & m_1(M) \\ m_2(1) & m_2(2) & .... & m_2(M) \\ .... & .... & .... & .... \\ m_N(1) & m_N(2) & .... & m_N(M) \end{matrix} \qquad \text{(i.1)}$$

в которой

N – число гаплотипов в выборке;

M – число маркеров, по которым данные гаплотипы построены (в настоящее время обычные значения числа маркеров в гаплотипе – 12,25,37 и даже 67);

$m_i(j)$ – целое неотрицательное число, соответствующее количеству повторов в i -м маркере (столбце матрицы) j -го гаплотипа (строки матрицы).

Каждый столбец матрицы (i.1) представляет собой выборку числовых значений одного и того же маркера у разных индивидов (финальных потомков каждой генеалогической ветви общего генеалогического древа прародителя).

Таким образом, для анализа имеется только матрица чисел (i.1), и ничего более. Ни точное количество мутаций, ни моменты их возникновения не известны, поскольку происходили много поколений назад. Чтобы оценить среднее количество мутаций, прошедшее от первопредка до финальных потомков, необходимо создать математическое описание процесса мутаций, приводящего к возникновению числовой матрицы (i.1), а также создать вычислительный алгоритм, позволяющий на основе этой матрицы получить искомые параметры.

Изложим постулаты, положенные в основу создания математического описания процесса формирования наблюдаемой выборки гаплотипов:

1. Постулаты о гаплотипах

1.1. Под базовой единицей данных понимается гаплотип, представляющий собой строго упорядоченный набор из M целых неотрицательных чисел. Таким образом, отдельный гаплотип можно представить вектором в M - мерном пространстве, а конкретные числовые значения маркеров – это координаты вектора (гаплотипа). Размер гаплотипа, вообще говоря, может быть различным: от $M=1$ (один маркер) до 100 и более. В настоящее время приняты строго определенные наборы маркеров для построения гаплотипов, причем количество маркеров в гаплотипе может быть 6,12, 25, 37 и 67.

1.2. Считается, что исследуемый набор гаплотипов происходит от одного базового гаплотипа, который назовем гаплотипом первопредка данной гаплогруппы. Гаплотипы, которые существовали от первопредка до любого из гаплотипов в исследуемой выборке, составляют непрерывную цепочку.

1.3. Наблюдению (непосредственному измерению количества повторов в позициях-маркерах) доступны только конечные гаплотипы (наличие определенного количества одновременно и отцовских и сыновьих гаплотипов ощутимо статистических выводов не меняет). Гаплотип потомка формируется из гаплотипа его непосредственного предка (отца) простым дублированием количества повторов в маркерах (т.е., дублированием гаплотипа), но очень редко в отдельных маркерах в процессе дублирования возможно изменение в результате мутации количества повторов на единицу, как в сторону уменьшения, так и в сторону увеличения, а значительно реже допускается увеличение/уменьшение на два.

1.4. Все гаплотипы от первопредка до наблюдаемых гаплотипов можно представить графом типа дерева с одним корнем, причем, ребрами являются гаплотипы, а вершинами – моменты появления нового гаплотипа из предыдущего. Несмотря на вариации длительностей жизни индивидов, при большом количестве поколений можно считать, что каждого современного потомка (конечного гаплотипа в дереве) отделяет от первопредка примерно одно и то же количество ребер (поколений), по четыре на каждое столетие, условно принимая поколение за 25 лет.

2. Постулаты о мутациях

2.1. Будем для простоты считать, что мутация в любом маркере происходит только в момент зачатия потомка и затем в течение жизни мутаций в данном маркере уже не происходит.

2.2. Под мутацией понимается событие мгновенного изменения количества повторов в некотором маркере гаплотипа по сравнению с его предковым (отцовским) гаплотипом, причем, с вероятностью p_1 количество повторов увеличивается на единицу, с вероятностью p_2 увеличивается на два, с вероятностью q_1 – уменьшается на единицу, и, наконец, с вероятностью q_2 – уменьшается на два. Все четыре события несовместны и составляют полную группу событий.

2.3. Появление мутации в каком-либо маркере гаплотипа потомка (точнее, сына) не зависит от того, появилась ли мутация в ином маркере данного гаплотипа сына. При этом направление изменения количества повторов в сторону увеличения или уменьшения не зависит от того, в какую сторону изменялось количество повторов в предыдущей мутации, а определяется только вероятностями p_1 , p_2 , q_1 , q_2 .

2.4. Поток событий (мутаций) развивается в дискретном времени, причем, дискретом является поколение. Каждый индивидуальный гаплотип существует только одно поколение (один дискрет времени). В гаплотипе сына могут сохраниться все координаты (числовые значения маркеров) или измениться любая из них. Таким образом, можно говорить о потоке мутаций в дискретном времени в каждой из цепочек (генеалогических линий) гаплотипов от первопредка исследуемой гаплогруппы до конечного (наблюдаемого) гаплотипа.

2.5. Поток мутаций считается стационарным (т.е., его базовые характеристики не зависят от времени).

2.6. Поток мутаций считается ординарным (т.е., вероятность одновременного возникновения двух и более мутаций считается величиной второго порядка малости, т.е., поток редкий).

2.7. В потоке мутаций отсутствует последействие (т.е., вероятность возникновения мутации в каком-либо маркере не зависит от того, когда в нем происходила предыдущая мутация).

2.8. Для любой пары наугад выбранных цепочек гаплотипов дерева (генеалогических линий) всегда есть некая общая часть («нижняя», т.е., начинающаяся от первопредка и далее к потомкам), т.е., имеется общий набор мутаций. В этом смысле потоки мутаций для каждой из цепочек частично взаимозависимы. Тем не менее, данная корреляция не влияет на итоговое численное значение оценки частоты мутаций, а влияет на степень уменьшения дисперсии оценки с ростом количества измерений.

* * * * * * * *

И последнее. Поскольку процессы мутаций в маркерах независимы друг от друга, то можно перейти от рассмотрения дерева гаплотипов к рассмотрению аналогичного дерева, но только для какого-нибудь конкретного маркера, т.е., искусственно «обрезаем» гаплотипы до единичной размерности (гаплотип, таким образом, содержит только один интересующий нас маркер). В этом случае ребрами дерева будут уже не «полные» гаплотипы, т.е., наборы чисел, а только числовые значения определенного маркера (фактически, одномерные гаплотипы). Всего можно построить столько идентичных по структуре деревьев, сколько маркеров в гаплотипе. По каждому из таких M деревьев в модели численно определяется один и тот же набор базовых параметров (вероятности p_1 , p_2 , среднее количество мутаций μ , прошедших от первопредка до наблюдаемого гаплотипа и т.д.). Для простоты изложения в дальнейшем, под гаплотипам часто понимается только конкретный исследуемый маркер гаплотипа, т.е, *одномерный гаплотип*.

ГЛАВА 1. Математическая модель количества повторов в маркерах нерекомбинирующего сегмента Y-хромосомы

В данной главе рассматривается математическая модель поведения количества повторов групп нуклеотидов в отдельно взятом маркере (локусе) нерекомбинирующего сегмента Y-хромосомы.

Все математические соотношения, полученные в данной главе, полностью вытекают из постулатов, изложенных во Введении, и не требуют дополнительных условий.

1.1. Вид распределений количества повторов в маркере при наличии мутаций

Пусть начальное количество повторов в маркере равно m_0 . В результате ряда мутаций количество повторов уже будет отличаться от начального значения. Это изменение от мутации к мутации удобно записывать в виде следующего рекуррентного соотношения:

$$m_n = m_{n-1} + z_n \quad , \qquad (1.1.1)$$

где

m_{n-1} – количество повторов в маркере до момента n -й мутации;

z_n – величина, на которую изменилось количество повторов в результате данной (n -й) мутации.

Таким образом, поведение количества повторов в маркере в каком-либо генеалогическом направлении (генеалогической линии) можно представить в виде последовательности сумм

$$m_n = \sum_{k=1}^{n} z_k \quad , \quad n \geq 0 \quad , \quad z_0 = 0 \qquad (1.1.2)$$

независимых одинаково распределенных случайных величин z_k с функцией распределения $F(x)$ $(0<F(x)<1)$ и определяет случайные блуждания на прямой. Величину z_k называют шагами (скачками) блуждания, суммы m_n определяют положение блуждания после n шагов (мутаций).

Существует только три вида блужданий: осциллирующий, уходящий в положительную бесконечность, уходящий в отрицательную бесконечность. В свою очередь, среди осциллирующих случайных блужданий имеются как возвратные, так и невозвратные.

Если пренебречь вероятностями p_2 и q_2 изменения количества повторов сразу на +2 или -2 соответственно, и считать, что z_k в момент мутации принимает значение +1 с вероятностью $p_1=p$, и значение -1 с вероятностью $q_1=q=1-p$, то количество повторов в маркере становится случайным блужданием по схеме Бернулли. При $p_1=p$ бернуллиевское случайное блуждание уходит в положительную бесконечность, а при $p<q$ – в отрицательную бесконечность. При $p=q=0{,}5$ мы имеем осциллирующее и возвратное блуждание. При допущении скачков на +2 и -2 в случае равенства $p_1+p_2=q_1+q_2$ блуждание остается осциллирующим, а при дополнительном условии, что число мутаций, попадающих в каждый конечный интервал равно бесконечности с вероятностью единица (верно для пуассоновского потока), блуждание возвратное.

За начальное значение можно взять любое число. Все теоретические результаты зависят только от разности шагов (мутаций) между новым начальным значением и тем, на котором остановились. Вследствие этого любая ветвь дерева, в качестве корня которого берется какой-либо промежуточный предок, может рассматриваться как абсолютно независимое дерево (гаплогруппа).

Чтобы проиллюстрировать процесс перераспределения вероятностей состояний количества повторов с каждым шагом (мутацией), рассмотрим процесс подробно.

Пусть имеем дискретную случайную величину m , значение которой последовательно подвергается изменению (мутации), причем, с вероятностью p увеличивается на единицу, а с вероятностью $q=1-p$ уменьшается на единицу.

Пусть прошло n шагов (мутаций). В результате значение m может с определенными вероятностями находиться в состояниях от m_0-n до m_0+n (

m_0 – начальное значение), причем, с каждым очередным шагом диапазон возможных состояний расширяется, а вероятности состояний перераспределяются.

Итак, после первой мутации величина m может принять следующие значения: m_0-1 и m_0+1. Значения m_0 после мутации, естественно быть не может. Таким образом, вероятность $P(m/n=1)$ нахождения случайной дискретной величины m в состояниях m_0-1, m_0, m_0+1 соответственно будет:

$$\begin{aligned} P(m_0-1/n=1) &= q = 1\,p^0 q^1 \\ P(m_0/n=1) &= 0 \\ P(m_0+1/n=1) &= p = 1\,p^1 q^0 \end{aligned} \quad . \qquad (1.1.3)$$

Схема, поясняющая формирование вероятностей после 1-й мутации

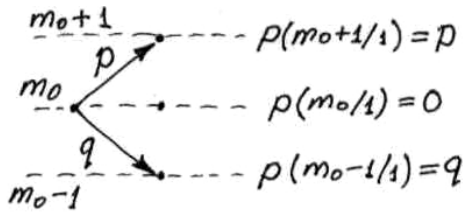


Рис.1-1. Возможные направления изменения исходного количества повторов m_0 после первой мутации (в состояние m_0+1 с вероятностью p, в состояние m_0-1 с вероятностью q)

После второй мутации $(n=2)$ величина m может уже быть в пяти состояниях m_0-2, m_0-1, m_0, m_0+1, m_0+2. Вероятности перечисленных состояний $P(m/n=2)$ будут следующие:

$$P(m_0-2/n=2) = q\cdot q = 1\,p^0q^2$$
$$P(m_0-1/n=2) = 0$$
$$P(m_0/n=2) = q\cdot p+p\cdot q = 2\,p^1q^1 \quad \text{(есть переход и из } m_0-1 \text{ , и из } m_0+1 \text{)}$$
$$P(m_0+1/n=2) = 0$$
$$P(m_0+2/n=2) = p\cdot p = 1\,p^2q^0$$

(1.1.4)

Схема, поясняющая формирование вероятностей после 2-й мутации

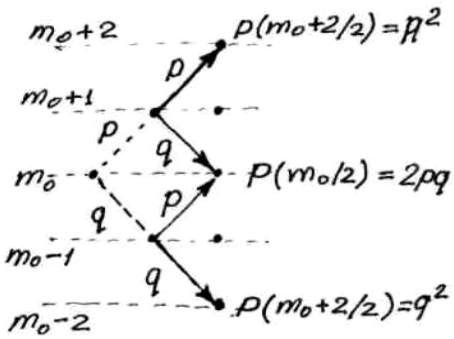


Рис.1-2. Возможные направления изменения исходного количества повторов m_0 после второй мутации (в состояние m_0+2 с вероятностью p^2, в состояние m_0-2 с вероятностью q^2, в начальное состояние m_0 с вероятностью 2pq)

После третьей мутации $(n=3)$ величина m может уже быть в семи состояниях m_0-3 , m_0-2 , m_0-1 , m_0 , m_0+1 , m_0+2 , m_0+3 . Вероятности перечисленных состояний $P(m/n=3)$ будут следующие:

$$
\begin{aligned}
P(m_0-3/n=3) &= (q\cdot q)q = 1q^3p^0 \\
P(m_0-2/n=3) &= 0 \\
P(m_0-1/n=3) &= 2(p\cdot q)q+(q\cdot q)p = 3q^2p^1 \\
P(m_0/n=3) &= 0 \\
P(m_0+1/n=3) &= (2\cdot p\cdot q)p+(p\cdot p)q = 3q^1p^2 \\
P(m_0+2/n=3) &= 0 \\
P(m_0+3/n=3) &= (p\cdot p)p = 1q^0p^3
\end{aligned}
\qquad (1.1.5)
$$

Схема, поясняющая формирование вероятностей после 3-й мутации

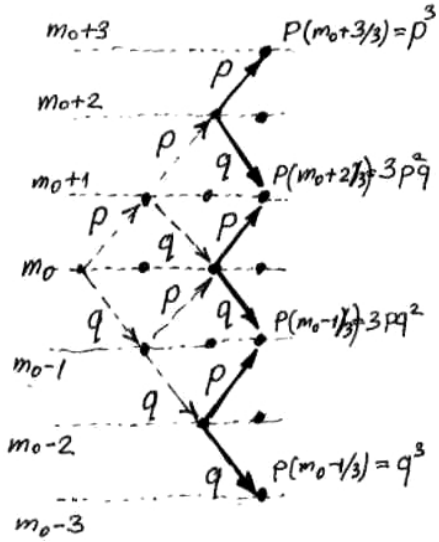


Рис.1-3. Возможные изменения количества повторов в маркере после третьей мутации с начального состояния m_0

Еще раз подробно напишем результат после четвертой мутации $(n=4)$. Величина m может уже быть в девяти состояниях m_0-4, m_0-3, m_0-2, m_0-1, m_0, m_0+1, m_0+2, m_0+3, m_0+4. Вероятности перечисленных состояний $P(m/n=4)$ будут следующие:

$$
\begin{aligned}
P(m_0-4/n=4) &= (q\cdot q\cdot q)q = q^4 = 1q^4p^0 \\
P(m_0-3/n=4) &= 0 \\
P(m_0-2/n=4) &= (q\cdot q\cdot q)p+(3\cdot q\cdot q\cdot p)q = 4q^3p^1 \\
P(m_0-1/n=4) &= 0 \\
P(m_0/n=4) &= (3\cdot q\cdot q\cdot p)p+(3\cdot p\cdot p\cdot q)q = 6q^2p^2 \\
P(m_0+1/n=4) &= 0 \\
P(m_0+2/n=4) &= (p\cdot p\cdot p)q+(3\cdot p\cdot p\cdot q)p = 4q^1p^3 \\
P(m_0+3/n=4) &= 0 \\
P(m_0+4/n=4) &= (p\cdot p\cdot p)p = p^4 = 1q^0p^4
\end{aligned}
\qquad (1.1.6)
$$

Очевидно, что коэффициентами в формулах являются биномиальные коэффициенты, а общий вид закона распределения примет следующий вид:

$$
\begin{aligned}
P(m_0-n+2i/n) &= \binom{n}{i} p^{n-i}q^i \quad , \quad (i=0,1,\ldots,n) \\
P(m_0-n+2i+1/n) &= 0 \quad , \quad (i=0,1,\ldots,n-1)
\end{aligned}
\qquad . \qquad (1.1.7)
$$

Для нас представляет интерес численные значения распределений на разных шагах (мутациях). Приведем первые шесть распределений (начальное значение $m_0=20$):

Таблица 1-1. Первые шесть распределений количества повторов в маркере после одной, двух, трех, четырех, пяти и шести мутаций

m	n=1	n=2	n=3	n=4	n=5	n=6
14	------	------	------	------	------	0.0156
15	------	------	------	------	0.0313	0.0000
16	------	------	------	0.0625	0.0000	0.0938
17	------	------	0.1250	0.0000	0.1563	0.0000
18	------	0.2500	0.0000	0.2500	0.0000	0.2344
19	0.5000	0.0000	0.3750	0.0000	0.3125	0.0000
20	0.0000	0.5000	0.0000	0.3750	0.0000	0.3125
21	0.5000	0.0000	0.3750	0.0000	0.3125	0.0000
22	------	0.2500	0.0000	0.2500	0.0000	0.2344
23	------	------	0.1250	0.0000	0.1563	0.0000
24	------	------	------	0.0625	0.0000	0.0938
25	------	------	------	------	0.0313	0.0000
26	------	------	------	------	------	0.0156

Графики распределения количества повторов (1.1.7) для разных значений параметра n (количества осуществившихся мутаций) представлены на следующих рисунках (количество повторов в маркере отложено по оси абсцисс):

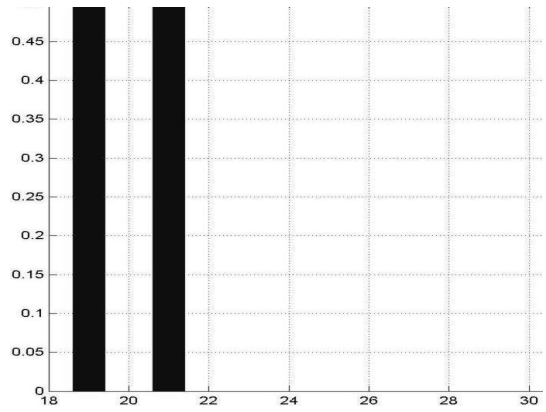


Рис.1-4. Распределение количества повторов при осуществившемся количестве мутаций $n=1$, для $p=0,5$, и при начальном количестве повторов $m_0=20$

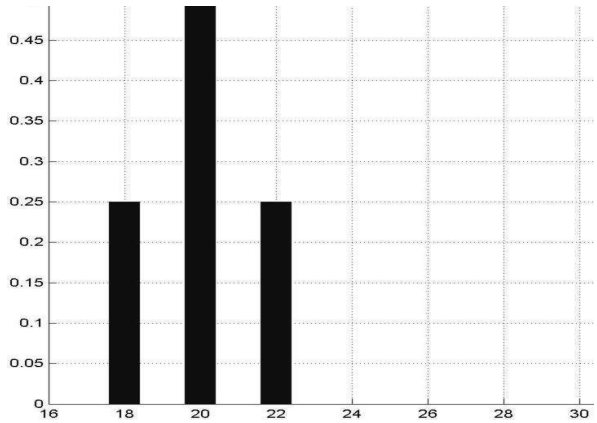


Рис.1-5. Распределение количества повторов при осуществившемся количестве мутаций $n=2$, для $p=0,5$, и при начальном количестве повторов $m_0=20$

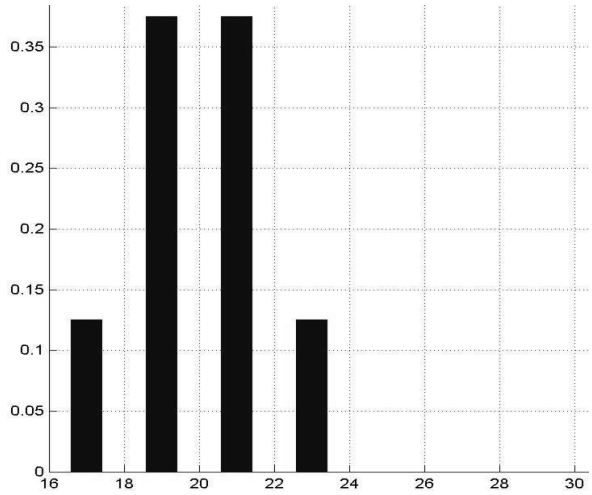


Рис.1-6. Распределение количества повторов при осуществившемся количестве мутаций $n=3$, для $p=0,5$, и при начальном количестве повторов $m_0=20$

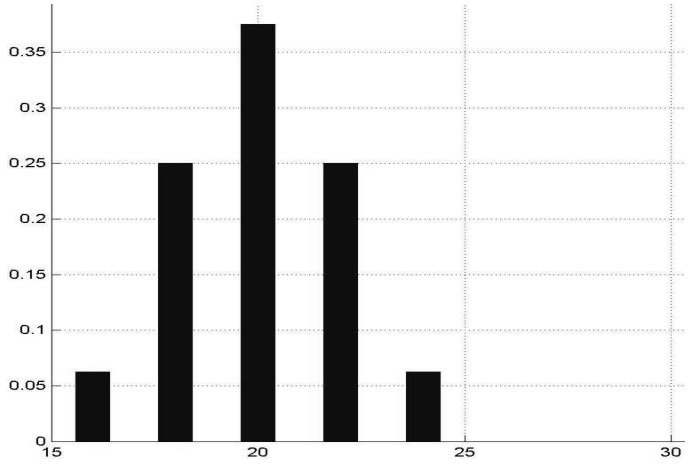


Рис.1-7. Распределение количества повторов при осуществившемся количестве мутаций $n=4$, для $p=0,5$, и при начальном количестве повторов $m_0=20$

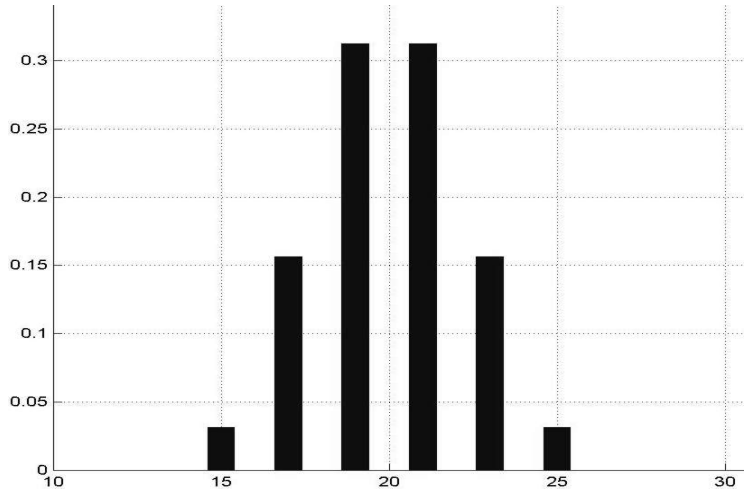


Рис.1-8. Распределение количества повторов при осуществившемся количестве мутаций $n=5$, для $p=0{,}5$, и при начальном количестве повторов $m_0=20$

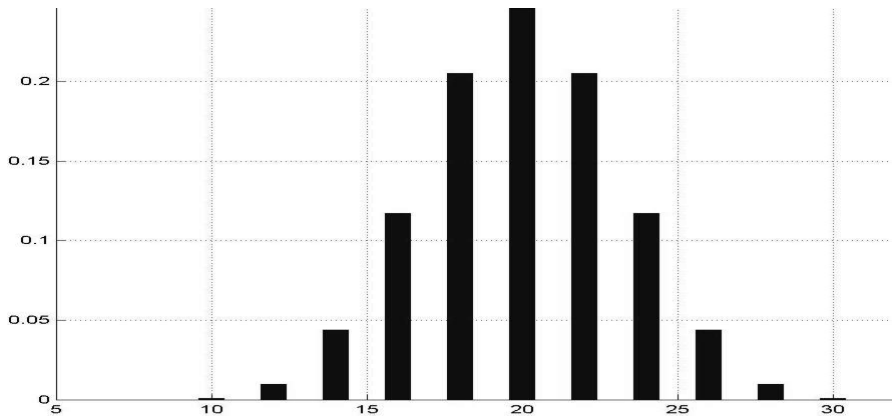


Рис.1-9. Распределение количества повторов при осуществившемся количестве мутаций $n=10$, для $p=0{,}5$, и при начальном количестве повторов $m_0=20$

Как следует из (1.1.7) и иллюстрируется таблицей 1-1, ширина распределения количества повторов в маркере линейно зависит от количества осуществившихся мутаций. Таким образом, получив экспериментальную гистограмму значений какого-либо маркера для группы «потомков» единого «предка», можно по ее ширине оценить количество осуществившихся мутаций.

На самом деле все обстоит намного сложнее. Пусть первопредок породил несколько генеалогических линий, последних потомков которых мы «одновременно» анализируем по прошествии значительного времени. В каждой из генеалогических ветвей за это время произойдет разное количество мутаций. Следовательно, конкретные значения одного и того же маркера у разных индивидов (гаплотипов) будут принадлежать распределению (1.1.7), но с разными значениями параметра n, а экспериментальная гистограмма количества повторов исследуемого маркера будет соответствовать смеси законов распределения.

Для примера рассмотрим случай, когда данные получены для двух одинаковых по количеству групп индивидов единого первопредка, причем, у одной группы от первопредка прошло 5 мутаций, а у другой - 6. Закон распределения, которому будет соответствовать гистограмма в этом случае, представлен в таблице 1-2 (предпоследняя и последняя колонка):

Таблица 1-2. Законы распределения количества повторов в маркере для смеси гаплотипов с разным количеством осуществившихся мутаций

```
*  1-й столбец:  значение маркера m (аргумент
                 закона распределения)
*  2-й столбец:  распределение вероятности P(m/n=5)
                 в случае пяти осуществившихся мутаций
*  3-й столбец:  распределение вероятности P(m/n=6)в случае
                 шести осуществившихся мутаций
*  4-й столбец:  смесь  0.25*P(m/n=5) + 0.75*P(m/n=6)
                 (в исследуемой группе четверть потомков с
                 пятью мутациями, а остальные - с шестью)
*  5-й столбец:  смесь  0.50*P(m/n=5) + 0.50*P(m/n=6)
                 (в исследуемой группе равное количество
                  потомков с пятью и шестью мутациями)

      14   ------   0.0156   0.0117   0.0078
      15   0.0313   0.0000   0.0078   0.0157
      16   0.0000   0.0938   0.0703   0.0469
      17   0.1563   0.0000   0.0391   0.0781
      18   0.0000   0.2344   0.1758   0.1172
      19   0.3125   0.0000   0.0781   0.1563
      20   0.0000   0.3125   0.2344   0.1563
      21   0.3125   0.0000   0.0781   0.1563
      22   0.0000   0.2344   0.1758   0.1172
      23   0.1563   0.0000   0.0391   0.0781
      24   0.0000   0.0938   0.0703   0.0469
      25   0.0313   0.0000   0.0078   0.0157
      26   ------   0.0156   0.0117   0.0078
```

Распределение равнопропорциональной смеси (последняя колонка) стало «гладким» и без нулевых «дырок». В предпоследней колонке представлена смесь законов распределения, но для случая, когда группа с пятью мутациями втрое меньше группы с шестью мутациями. «Гладкость» распределения пропадает, но вместо «дырок», тем не менее, имеют место ненулевые значения.

Следует также обратить внимание на тот факт, что наиболее вероятным состоянием является состояние m_0 - случай «отсутствия» мутации. Несмотря на то, что процесс с течением времени все более раскачивается относительно среднего значения, но все равно возвращается в исходное состояние. Из столбцов для $n=2,4,6$ таблицы 1-1 видно, что вероятности состояния $m_0=20$ оказываются достаточно большими, т.е., серия мутаций, приведшая к возврату состояния m_0 , окажется ненаблюдаемой.

Формула (1.1.7) справедлива и для несимметричных случаев, когда величины p и q не равны друг другу. В таблице 1-3 представлены распределения для 6 шагов при разных вероятностях p :

Таблица 1-3. Распределение вероятностей количества повторов в маркере $P(m/n=6)$ для шести осуществившихся мутаций при разных вероятностях p

m	p=0,5	p=0.25	p=0,125
14	0.0156	0.1780	0.4488
15	0.0000	0.0000	0.0000
16	0.0938	0.3560	0.3847
17	0.0000	0.0000	0.0000
18	0.2344	0.2966	0.1374
19	0.0000	0.0000	0.0000
20	0.3125	0.1318	0.0262
21	0.0000	0.0000	0.0000
22	0.2344	0.0330	0.0028
23	0.0000	0.0000	0.0000
24	0.0938	0.0044	0.0002
25	0.0000	0.0000	0.0000
26	0.0156	0.0002	0.0001
матожидание:	mean=20	mean=17	mean=15,5

В последней строке таблицы 1-3 даны значения матожиданий (mean) для всех трех рассмотренных случаев: симметричного ($p=0{,}5$) и несимметричных ($p=0,25$ и $p=0,125$).

Как видим, с перекосом процесса в сторону уменьшения вероятности «положительного» приращения (на +1) матожидание процесса довольно быстро «сползает» вниз от величины m_0 , равного двадцати. В результате исходное состояние m_0 прямым вычислением среднего по гистограмме осуществить уже не удастся. В значительной степени меняется и форма распределения.

Следовательно, при реконструкции базового гаплотипа следует особое внимание обратить на степень симметричности экспериментальных гистограмм. Отсутствие учета коэффициента асимметрии гистограммы приведет к смещению оценки базовых значений повторов в маркерах.

Примеры смесей законов распределения представлены на рис. 1-10 и 1-11.

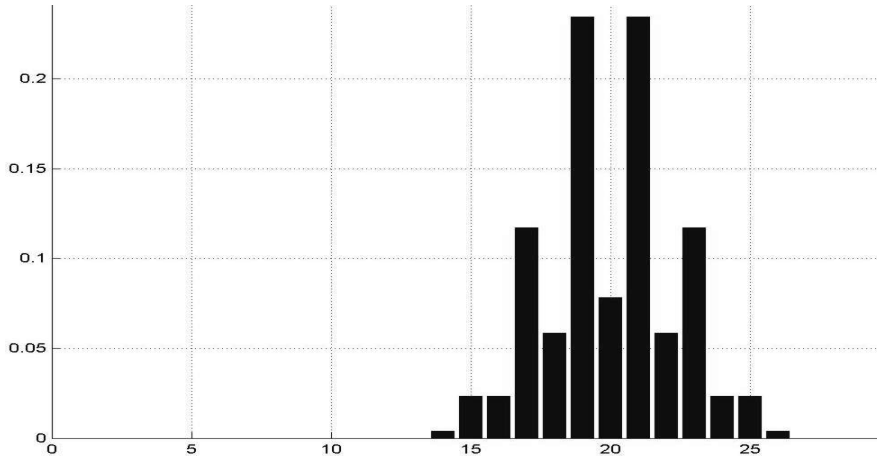


Рис.1-10. Неравновесная смесь двух законов распределения количества повторов для $n=5$ осуществившихся мутаций и $n=6$ осуществившихся мутаций (начальное количество повторов до мутаций было 20)

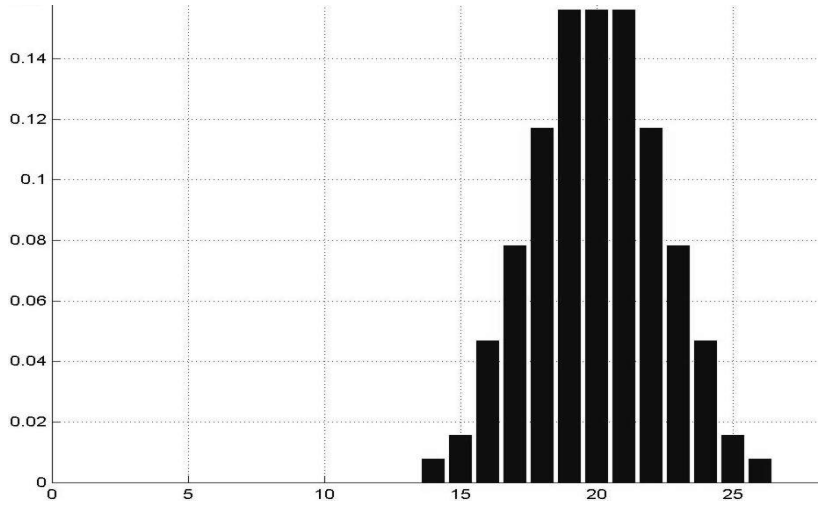


Рис.1-11. Равновесная смесь двух законов распределения количества повторов для $n=5$ осуществившихся мутаций и $n=6$ осуществившихся мутаций (начальное количество повторов до мутаций было 20)

На рис.1-12 в качестве примера представлено распределение со значительно меньшей вероятностью увеличения количества повторов в маркере $p=0,125$ по сравнению с вероятностью уменьшения количества повторов $q=0,875$ (третья колонка таблицы 1-3):

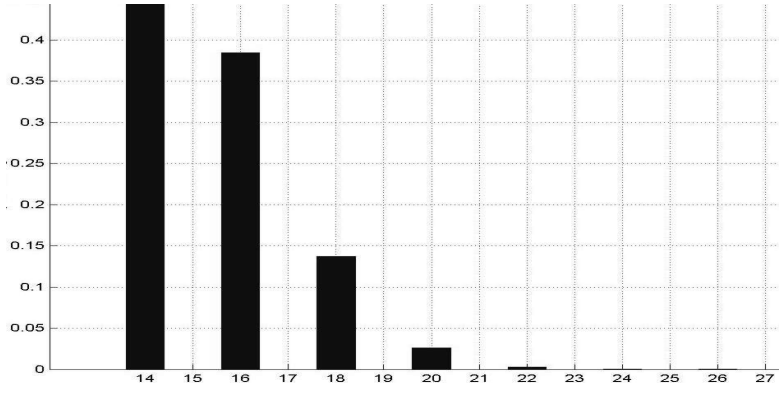


Рис.1-12. Распределение при осуществившемся количестве мутаций $n=6$, для $p=0,125$, и при начальном количестве повторов $m_0=20$

При увеличении вероятности p от 0,5 до 1 искажение формы распределения будет осуществляться аналогичным образом, только в другую сторону, чем на рис.1-12.

* * * * * * * *

В завершение данного параграфа следует подчеркнуть ряд неприятных особенностей случайного блуждания, которые фактически заставляют привлекать большие объемы данных для получения достоверных результатов.

Случайное блуждание в известном смысле начинается сначала каждый раз, когда количество повторов в маркере возвращается в начало. «Время до r -го возвращения есть, таким образом, сумма времен ожидания, которые можно интерпретировать как «результаты измерения одной и той же физической величины в одинаковых условиях». Считается, что среднее из r таких наблюдений должно сходиться к «истинному значению». Однако в данном случае сумма является величиной такого же порядка, как r^2 r^2 , и поэтому среднее увеличивается примерно пропорционально r . Более глубокий анализ показывает, что одно из r времен ожидания является величиной такого же порядка, что и вся сумма, а именно, r^2 . На практике такое явление часто приписывается «ошибке эксперимента» или отбрасывается как «постороннее». Трудно заметить то, что не ожидалось увидеть» [1, т.1,с.110].

«… в симметричном случайном блуждании математическое ожидание числа попаданий в состояния $k \geq 1$ (состояние k – это означает, что количество повторов на данном шаге равно k) до первого возвращения в нуль (т.е., стало равным исходному, как у первопредка) равно единице при всех k . Фантастичность этого результата можно лучше уяснить в терминологии игры с бросанием монеты. Наше утверждение состоит в том, что до первого возвращения на нулевой уровень (или равное исходному у первопредка) средний накопленный выигрыш … принимает когда-либо любое значение k … В симметричном случайном блуждании Бернулли (бросание монеты) каждое значение +1 или -1 достигается с вероятностью единица, однако математическое ожидание времени ожидания для каждого из этих событий бесконечно» [1, т.2, с.447].

Если рассмотреть только одну генеалогическую линию, то, как уже говорилось, оказывается, что каждое возвращение количества повторов в

маркере в исходное состояние полностью «отрезает» всю предысторию, и процесс начинается сначала (т.е., текущее значение маркера стало «первопредковым» для всей последующей части генеалогической линии). Таким образом, по одному гаплотипу доступна анализу только оставшаяся часть от последнего «возвращения» до настоящего времени. Да и в этом случае доступность относительна, поскольку локальные возвраты с максимума также «отрезают» безвозвратно часть информации.

Но в случае наблюдения массы гаплотипов (точнее, конкретного маркера у массы гаплотипов) оказывается, что у какого-то гаплотипа вообще не было возвратов, у какого-то было строго линейное нарастание (убывание) количества повторов и так далее. И чем больше гаплотипов, тем больше реализовано было различных «траекторий» блуждания количества повторов. В этом случае общая гистограмма уже несет в себе и истинное количество шагов (мутаций) от первопредка, и общее количество возвратов и т.д.

Вероятность возвращения в начало в момент n -й мутации ($n=2,4,6,...,2k$) [2]:

$$U(n)=\binom{n}{\frac{n}{2}}\cdot 2^{-n} \quad . \qquad (1.1.8)$$

Вероятность того, что первое возвращение в начало произошло на n -й мутации, дается следующим соотношением (равносильна вероятности, что до этой мутации не было ни одного возвращения):

$$F(n)=\frac{1}{n-1}U(n) \quad . \qquad (1.1.9)$$

Вероятность того, что до данной мутации было хотя бы одно возвращение:

$$G(n)=1-F(n) \quad . \qquad (1.1.10)$$

В таблице 1-4 приведены расчеты вероятностей $G(\nu)$, $F(\nu)$, $U(\nu)$.

Таблица 1-4. Распределения вероятностей процесса блуждания

n – номер шага от начала (количество осуществившихся мутаций)

$U(\nu)$ – вероятность возврата в состояние предка на n-й мутации.

$F(\nu)$ – вероятность того, что первое возвращение в начало произошло на n-й мутации

$G(\nu)$ – вероятность того, что до данной мутации было хотя бы одно возвращение

n	$U(\nu)$	$F(\nu)$	$G(\nu)$
2	0.500000	0.500000	0.500000
4	0.375000	0.125000	0.875000
6	0.312500	0.062500	0.937500
8	0.273438	0.039063	0.960938
10	0.246094	0.027344	0.972656
12	0.225586	0.020508	0.979492
14	0.209473	0.016113	0.983887
16	0.196381	0.013092	0.986908
18	0.185471	0.010910	0.989090
20	0.176197	0.009274	0.990726

Из таблицы 1-4 следует, что даже на 20-й мутации вероятность возврата в исходное состояние довольно велика, а на более ранних шагах (4-я и 6-я мутации) эта вероятность превышает 30%. Следовательно, уже на 4-й и 6-й мутациях по 30% блужданий приходят к начальному состоянию, «отрезая» для наблюдения по 4 и 6 мутаций соответственно.

1.2. Влияние параметров потока мутаций на окончательный вид закона распределения количества повторов в маркере

В реальности совокупность потомков, подвергаемая исследованию, всегда состоит из особей с различным количеством мутаций, осуществившихся за время, прошедшее от первопредка. В силу

стохастичности мутаций в исследуемой совокупности встретятся варианты с любым количеством мутаций - от 0 до сколь угодно большого значения. В этом случае закон распределения количества повторов в маркере представляет собой смесь законов распределения (1.1.7) с разным значением параметра n (количества мутаций) у каждой составляющей:

$$P(m)=C_0P(m/n=0)+C_1P(m/n=1)+...+C_jP(m/n=j)+... \qquad (1.2.1)$$

где m - аргумент (количество повторов) суммарного закона распределения.

Необходимо как-то вычислять бесконечный набор коэффициентов C_0 , C_1 , C_0 ,... Очевидно, что эти коэффициенты фактически определяют доли гаплотипов в выборке, у которых маркер не мутировал (C_0), мутировал только один раз (C_1), мутировал ровно два раза (C_2) и так далее. А от чего эти доли зависят? Очевидно, что если интенсивность мутаций мала, то коэффициент C_0 будет большим (много немутировавших маркеров), а остальные коэффициенты будут монотонно уменьшаться по величине (например, мутировавших два раза больше, чем мутировавших пять раз). Оказывается, свойства самого потока мутаций прямым образом влияют на соотношение коэффициентов , C_0 , C_1 , C_2 ,....

Непосредственно поток мутаций мы наблюдать не можем по причине исключительной редкости осуществления мутаций (единицы за тысячу лет в непрерывной линии потомков мужского рода). Измерению доступно лишь количество повторов в маркерах, которое возникло за счет мутаций. Т.е., мы анализируем лишь следы, оставленные мутациями. Следовательно, необходимо делать определенные предположения о свойствах потока мутаций во времени.

Итак, выберем некий интервал времени T . Рассмотрим, какие события могут произойти за этот интервал. Под событием будем понимать возникновение мутации, неважно в какую сторону при этом изменился маркер (увеличилось или уменьшилось количество повторов в нем).

На интервале T может не возникнуть ни одного события, может появиться ровно одно событие (неважно, в какой момент), ровно два события, ровно три события и так далее до бесконечности. Заметим, что нас не интересует взаимное положение событий на интервале T , а интересует только их количество. В рассматриваемом случае интервал времени T формируется непрерывной цепочкой мужских потомков от некоего

первопредка до финального потомка (например, нашего современника), и, таким образом, интервал T соответствует времени существования какой-либо генеалогической линии.

Очевидно, что событие полного отсутствия мутаций на интервале T, событие появления ровно одной мутации, событие появления ровно двух мутаций, ровно трех мутаций и так далее, составляют полную группу событий. Следовательно, сумма вероятностей всех этих событий равна единице.

Обозначим вероятность полного отсутствия событий на интервале времени T как $P_{П}(n=0,T)$ (здесь через $n=0$ обозначено событие, что количество мутаций n равно нулю). Очевидно, что чем больше интервал T, тем меньше эта вероятность.

Аналогично обозначим вероятности появления ровно одной мутации на интервале T как $P_{П}(n=1,T)$, ровно двух мутаций – $P_{П}(n=2,T)$ и так далее. Вероятность $P_{П}(n,T)$ является функцией двух параметров. И как уже говорилось, сумма всех вероятностей равна единице:

$$P_{П}(0,T)+P_{П}(1,T)+...+P_{П}(j,T)+...=1 \qquad (1.2.2)$$

Но как вычислить вероятности $P_{П}(0,T)$, $P_{П}(1,T)$,...? Для этого нам помогут свойства потока мутаций.

Примем следующие предположения относительно потока мутаций:

- поток мутаций редкий, т.е., вероятность одновременного осуществления мутаций конкретного маркера – величина второго порядка малости (поток ординарный);
- темп мутаций не зависит от времени (поток стационарный);
- вероятность осуществления мутации не зависит от того, когда происходила предыдущая мутация (поток без последействия).

Следовательно, поток мутаций – это пуассоновский поток. Для пуассоновского потока вероятность $P_{П}(n,T)$ того, что за интервал времени T осуществится ровно n мутаций, описывается следующим выражением:

$$P_{П}(n,T)=\frac{(\lambda\cdot T)^{n}}{n!}\mathrm{e}^{-\lambda\cdot T} \quad , \qquad (1.2.3)$$

где λ – интенсивность потока, имеющая размерность количества мутаций, произошедших за определенный интервал времени на каком либо маркере, например [*мутации/(маркер×время)*].

Заметим, что параметры λ и T входят в (1.2.3) в виде произведения $\lambda \cdot T$, причем, нам, как правило, не известны оба параметра. Вместо них введем один обобщенный параметр

$$\mu = \lambda \cdot T \quad , \qquad (1.2.4)$$

размерность которого:

[*мутации/(маркер×время)*]×[*время*] = [*мутации/маркер*]

Выражение (1.2.3), таким образом, принимает следующий вид:

$$P_{П}(n/\mu) = \frac{\mu^n}{n!} e^{-\mu} \quad , \qquad (1.2.3)$$

Обобщенный параметр μ имеет смысл среднего количества мутаций, произошедших на данном маркере на интервале времени T, прошедшем от «первопредка» до настоящего времени.

Обобщенный параметр μ является одновременно и матожиданием распределения (1.2.3) и численно равен дисперсии этого же распределения (1.2.3).

Представим себе, что на интервале времени от первопредка до конкретного индивида (гаплотипа) произошло ровно n мутаций. Какова вероятность того, что значение маркера (количество повторов) изменится из состояния m в одно из $2(m+n)+1$ новых состояний $\{(m-n),(m-n-1),...,(m),(m+1),(m+2),...,(m+n)\}$? А эти вероятности мы уже нашли в самом начале (формула 1.1.7).

Пусть теперь имеется достаточно большой набор гаплотипов потомков от одного первопредка (для простоты будем считать этот набор бесконечным). Напомним, что рассуждения касаются одного конкретного маркера в гаплотипе.

Очевидно, что будет какая-то часть гаплотипов, у которых значение маркера так и останется m, у какой-то части будет $(m-n)$, у какой-то части – $(m-n-1)$, у какой-то части – $(m+1)$, и так далее до $(m+n)$. Одним словом, всевозможные значения от $(m-n)$ до $(m+n)$.

Подчеркнем, что для конкретного числа мутаций закон распределения количества повторов на данном маркере свой и отличается от распределения для другого количества мутаций (см. формулу (1.1.7)).

Представим себе, что мы точно знаем, сколько мутаций было в каждом гаплотипе. Тогда рассортируем гаплотипы по количеству произошедших мутаций. Очевидно, что доля каждого типа гаплотипов будет соответствовать значениям вероятности пуассоновского потока $P(0,T)$ (для гаплотипов, в которых не произошло ни одной мутации), $P(1,T)$ (для совокупности гаплотипов, в которых произошла ровно одна мутация), $P(2,T)$ (для совокупности гаплотипов, в которых произошло ровно две мутации), и так далее.

Легко видеть, что закон распределения значения маркера m при одновременном наличии множества гаплотипов с разным количеством мутаций на одном и том же интервале T, представляет собой смесь законов распределения

$$P(m)=C_0P(m/n=0)+...+C_jP(m/n=j)+... = \\ = P_П(0,T)P(m/n=0)+...+P_П(j,T)P(m/n=j)+... \qquad (1.2.5)$$

где роль коэффициентов $C_0, C_1, C_2, ...$, удовлетворяющих условиям нормировки $C_0+C_1+C_2+...+C_j+...=1$ выполняют соответствующие вероятности $P_П(0,T), P_П(1,T), ..., P_П(j,T), ...$ пуассоновского распределения (1.2.3) (сумма которых, естественно, также равна единице).

1.3. Закон распределения количества повторов в маркере в «двухшаговой» модели мутации

В предыдущих параграфах под мутацией понималось событие, при котором изменения значения маркера (количества повторов) происходят на единицу вверх или вниз («одношаговая» мутация). Усложним явление мутации возможностью изменения количества повторов не только на +1 или –1, но и на +2 или –2 («двухшаговая» мутация). Таким образом, мутация влечет за собой одно из следующих четырех событий в маркере (эти события составляют полную группу событий):

- изменение маркера на +1 с вероятностью p_1;
- изменение маркера на +2 с вероятностью p_2;
- изменение маркера на –1 с вероятностью $q_1=1-p_1$;
- изменение маркера на –2 с вероятностью $q_2=1-p_2$.

Очевидно, что $p_1+q_1+p_2+q_2=1$. Из опыта известно, что величина $p_2 << p_1$. Понятно, что сумма $p_1+p_2=p$ – это вероятность изменения маркера на +1 в предыдущей «одношаговой» модели, а $q_1+q_2=q$ – аналогичная вероятность изменения маркера на –1. В новой «двухшаговой» модели вероятности p и q «расщепились» на две составляющие каждая.

На практике величину p_1 удобно задавать как долю вероятности p , которая теперь имеет смысл увеличения значения маркера на один и более шагов:

$p_1=p\cdot\eta$ здесь доля вероятности задается коэффициентом η , который лежит в пределах $0<\eta<1$.

Соответственно, $p_2=p\cdot(1-\eta)$.

Аналогично для вероятностей уменьшения значения маркера:

$q_1=q\cdot\eta$,

$q_2=q\cdot(1-\eta)$.

Нас, как и прежде, интересуют законы распределения значения маркера в случае одной мутации, двух мутаций подряд, трех мутаций, четырех, и так далее. Для «одношаговой» модели эти вероятности задавались выражением (1.1.7).

Для «двухшаговой» модели получить простое выражение, аналогичное (1.1.7), затруднительно. Тем не менее, можно разработать точную вычислительную процедуру. Итак. Пусть маркер после n -й мутации находится в состоянии m (т.е., имеет количество повторов, равное m). В результате следующей $(n+1)$ -й мутации маркер изменяет свое состояние в одно из четырех состояний:

- из состояния m в состояние $(m+2)$ с вероятностью p_2
- из состояния m в состояние $(m+1)$ с вероятностью p_1 (1.3.1)
- из состояния m в состояние $(m-1)$ с вероятностью q_1
- из состояния m в состояние $(m-2)$ с вероятностью q_2

Сразу же заметим, что, например, в состояние $(m+2)$ может с вероятностью q_2 прийти состояние $(m+4)$, с вероятностью q_1 – состояние $(m+3)$, с вероятностью p_1 – состояние $(m+1)$. Точно так же обстоит дело и с другими состояниями.

Обозначим вероятность нахождения маркера в m -м состоянии после n-й мутации как $P(m/n)$, где m – некое начальное значении маркера до 1-й мутации (т.е., после «нулевой» мутации). Тогда доли вероятности в новом распределении $P(.../n+1)$ от m -го состояния распределения $P(m/n)$, будут следующими:

$$\begin{array}{lll} \text{состояние} \quad (m+2) & : & P(m/n)\cdot p_2 \\ \text{состояние} \quad (m+1) & : & P(m/n)\cdot p_1 \\ \text{состояние} \quad (m-1) & : & P(m/n)\cdot q_1 \\ \text{состояние} \quad (m-2) & : & P(m/n)\cdot q_2 \end{array} \qquad (1.3.2)$$

Пусть начальное состояние маркера – m . Вероятностное распределение на «нулевом» шаге

$$P(m/0) = 1 \quad . \qquad (1.3.3)$$

После первой мутации:

$$\begin{array}{l} P(m+2/1)\cdot p_2 \\ P(m+1/1)\cdot p_1 \\ P(m/1) \ =0 \\ P(m-1/1)\cdot q_1 \\ P(m-2/1)\cdot q_2 \end{array} \qquad (1.3.4)$$

После второй мутации вероятностное распределение имеет следующий вид:

$$\begin{array}{l} P(m+4/2)=p_2 p_2 \\ P(m+3/2)=p_1 p_2+p_2 p_1 \\ P(m+2/2)=p_1 p_1 \\ P(m+1/2)=p_2 q_1+q_1 p_2 \\ P(m/2)=p_1 q_1+p_2 q_2+q_1 p_1+q_2 p_2 \\ P(m-1/2)=p_1 q_2+q_2 p_1 \\ P(m-2/2)=q_1 q_1 \\ P(m-3/2)=q_1 q_2+q_2 q_1 \\ P(m-4/2)=q_2 q_2 \end{array} \qquad (1.3.5)$$

Из (1.3.3), (1.3.4) и (1.3.5) следует простая рекуррентная вычислительная процедура распределений «двухшаговой» модели для любого количества мутаций n (значком := вместо знака равенства обозначена операция присвоения, которая в данном случае соответствует добавлению к уже существующему значению нового значения):

$$
\begin{aligned}
&P(m+2/n+1) := P(m+2/n+1) + P(m/n)\cdot p_2 \\
&P(m+1/n+1) := P(m+1/n+1) + P(m/n)\cdot p_1 \\
&P(m/n+1) \ := P(m/n+1) \ + 0 \quad \text{для всех} \quad m = m-n, \ldots m+n \qquad (1.3.6) \\
&P(m-1/n+1) := P(m-1/n+1) + P(m/n)\cdot q_1 \\
&P(m-2/n+1) := P(m-2/n+1) + P(m/n)\cdot q_2
\end{aligned}
$$

Начальное значение для запуска рекуррентной процедуры, очевидно, $P(m/0) := 1$. Далее все вероятностные распределения формируются последовательно с помощью (1.3.6).

Попутно заметим, что выражение (1.1.7) тоже можно вычислять с помощью рекуррентной процедуры (1.3.6), положив вероятности q_2 и p_2 равными нулю, а вероятности q_1 и p_1 – величинам p и q соответственно. То есть, процедура (1.3.6) является общей для «одношаговой» и «двухшаговой» моделей.

В таблице 1-5 представлены тройками таблицы закона распределения (1.3.6) в случае двух ($n=2$) и семи ($n=7$) осуществившихся мутаций при различном соотношении вероятностей p_1 и p_2 , задаваемых коэффициентом K (напомним, что при $K=1$ вероятности $p_1=p$ и $p_2=0$). Суммарная вероятность $p=p_1+p_2=0{,}5$ (закон распределения симметричный). За базовое значение маркера взято $m=15$ повторов.

В таблице 1-6 представлены те же самые законы распределения, но для несимметричного случая, когда $p=p_1+p_2=0{,}8$.

Таблица 1-5. Распределение вероятностей количества повторов для p=0.5

	n=2				n=7		
	K=1	K=0.95	K=0.9		K=1	K=0.95	K=0.9
5	------	------	------	5	------	------	------
6	------	------	------	6	------	0.0001	0.0003
7	------	------	------	7	------	0.0021	0.0033
8	------	------	------	8	0.0078	0.0065	0.0070
9	------	------	------	9	0.0000	0.0122	0.0186
10	------	------	------	10	0.0547	0.0415	0.0363
11	------	0.0006	0.0025	11	0.0000	0.0325	0.0489
12	------	0.0238	0.0450	12	0.1641	0.1206	0.0972
13	0.2500	0.2256	0.2025	13	0.0000	0.0528	0.0792
14	0.0000	0.0238	0.0450	14	0.2734	0.1991	0.1560
15	0.5000	0.4525	0.4100	15	0.0000	0.0609	0.0912
16	0.0000	0.0238	0.0450	16	0.2734	0.1991	0.1560
17	0.2500	0.2256	0.2025	17	0.0000	0.0528	0.0792
18	------	0.0238	0.0450	18	0.1641	0.1206	0.0972
19	------	0.0006	0.0025	19	0.0000	0.0325	0.0489
20	------	------	------	20	0.0547	0.0415	0.0363
21	------	------	------	21	0.0000	0.0122	0.0186
22	------	------	------	22	0.0078	0.0065	0.0070
23	------	------	------	23	------	0.0021	0.0033
24	------	------	------	24	------	0.0001	0.0003
25	------	------	------	25	------	------	------

Таблица 1-6. Распределение вероятностей количества повторов для p=0.8

	n=2				n=7		
	K=1	K=0.95	K=0.9		K=1	K=0.95	K=0.9
5	------	------	------	5	------	------	------
6	------	------	------	6	------	------	------
7	------	------	------	7	------	------	------
8	------	------	------	8	------	------	------
9	------	------	------	9	------	0.0001	0.0001
10	------	------	------	10	0.0004	0.0003	0.0004
11	------	0.0001	0.0004	11	0.0000	0.0008	0.0013
12	------	0.0038	0.0072	12	0.0043	0.0034	0.0032
13	0.0400	0.0361	0.0324	13	0.0000	0.0046	0.0069
14	0.0000	0.0152	0.0288	14	0.0287	0.0213	0.0177
15	0.3200	0.2896	0.2624	15	0.0000	0.0160	0.0239
16	0.0000	0.0152	0.0288	16	0.1147	0.0834	0.0651
17	0.6400	0.5776	0.5184	17	0.0000	0.0374	0.0557

Продолжение таблицы 1-6.

18	------	0.0608	0.1152	18	0.2753	0.1988	0.1518
19	------	0.0016	0.0064	19	0.0000	0.0646	0.0959
20	------	------	------	20	0.3670	0.2650	0.2024
21	------	------	------	21	0.0000	0.0815	0.1207
22	------	------	------	22	0.2097	0.1536	0.1223
23	------	------	------	23	------	0.0543	0.0804
24	------	------	------	24	------	0.0024	0.0075
25	------	------	------	25	------	------	------

В случае $p<0{,}5$ перекос законов распределения будет в обратную сторону по сравнению с представленными в таблице 1-6.

Вид распределений количества повторов в маркере (1.3.6) представлен на следующих рисунках. Слева – «одношаговая» модель ($p=0{,}5$), а справа – «двухшаговая» модель ($p_1=0{,}8\,p$ и $p_2=0{,}2\,p$ соответственно):

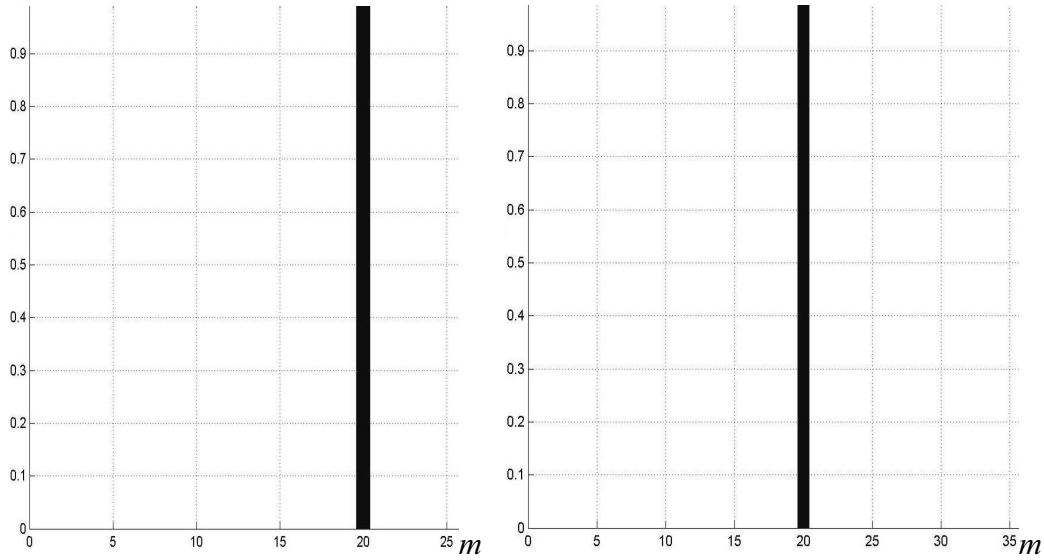


Рис.1-13. слева: $P(m/n=0)\quad m_0=20\quad p=0{,}5$
справа: $P(m/n=0)\quad m_0=20\quad p_1=0{,}4\quad p_2=0{,}1$

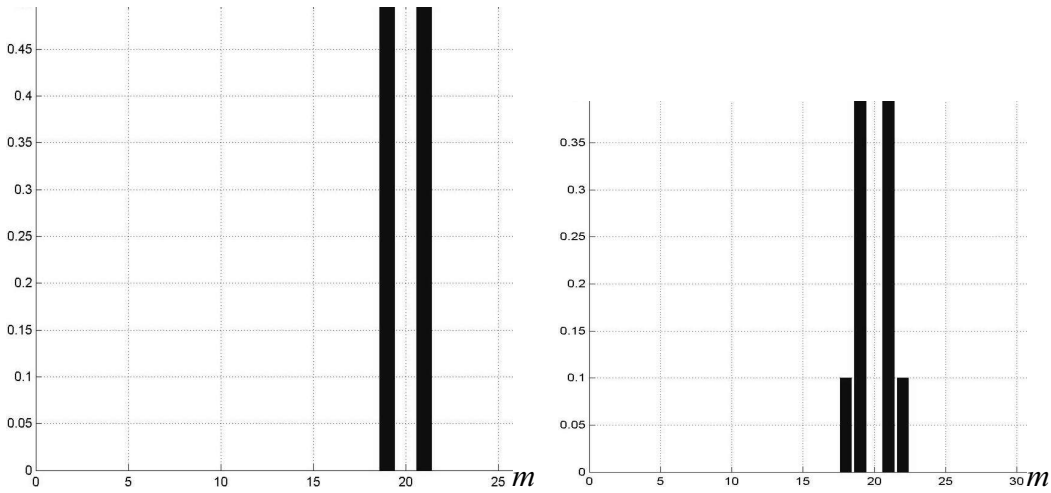


Рис.1-14. слева: $P(m/n=1)\ \ m_0=20\ \ p=0{,}5$
справа: $P(m/n=1)\ \ m_0=20\ \ p_1=0{,}4\ \ p_2=0{,}1$

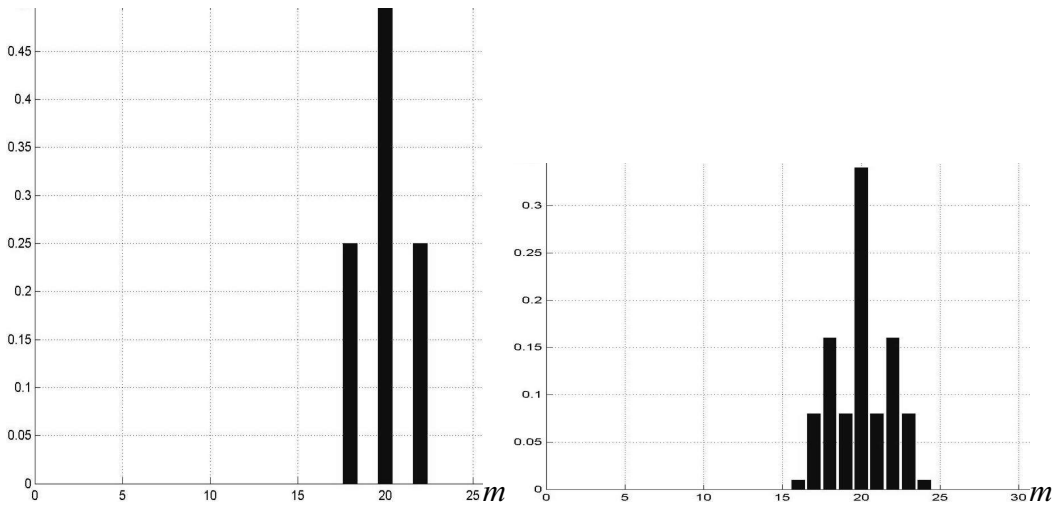


Рис.1-15. слева: $P(m/n=2)\ \ m_0=20\ \ p=0{,}5$
справа: $P(m/n=2)\ \ m_0=20\ \ p_1=0{,}4\ \ p_2=0{,}1$

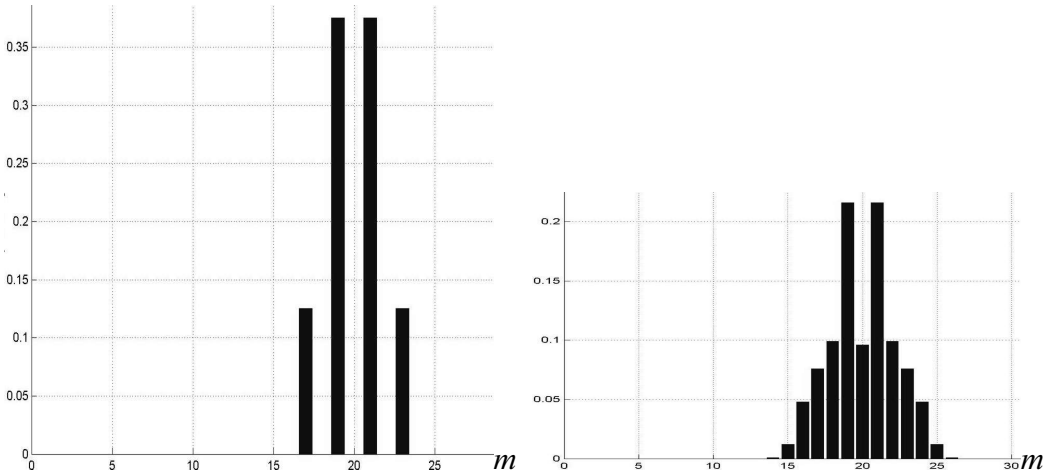


Рис.1-16. слева: $P(m/n=3)\quad m=20\quad p=0{,}5$
справа: $P(m/n=3)\quad m=20\quad p_1=0{,}4\quad p_2=0{,}1$

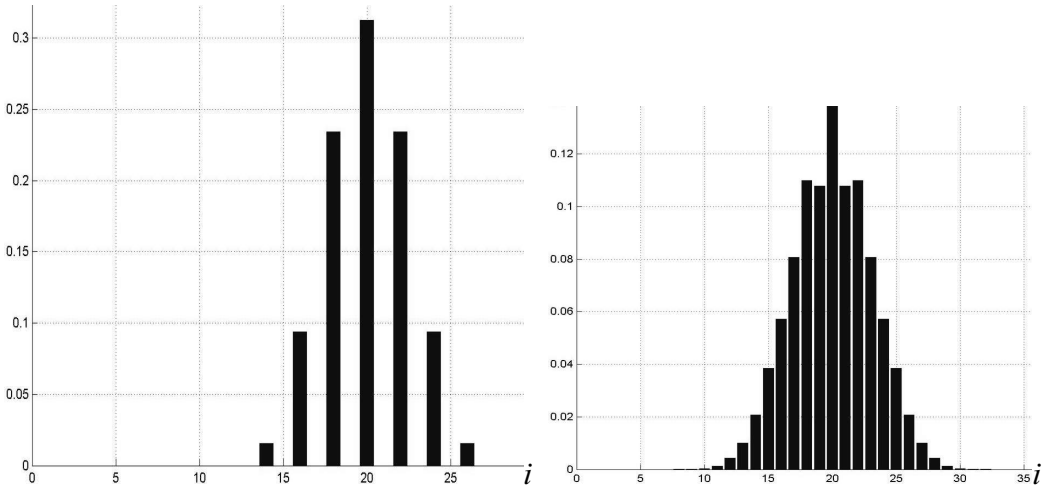


Рис.1-17. слева: $P(i/n=6)\quad m_0=20\quad p=0{,}5$
справа: $P(i/n=6)\quad m_0=20\quad p_1=0{,}4\quad p_2=0{,}1$

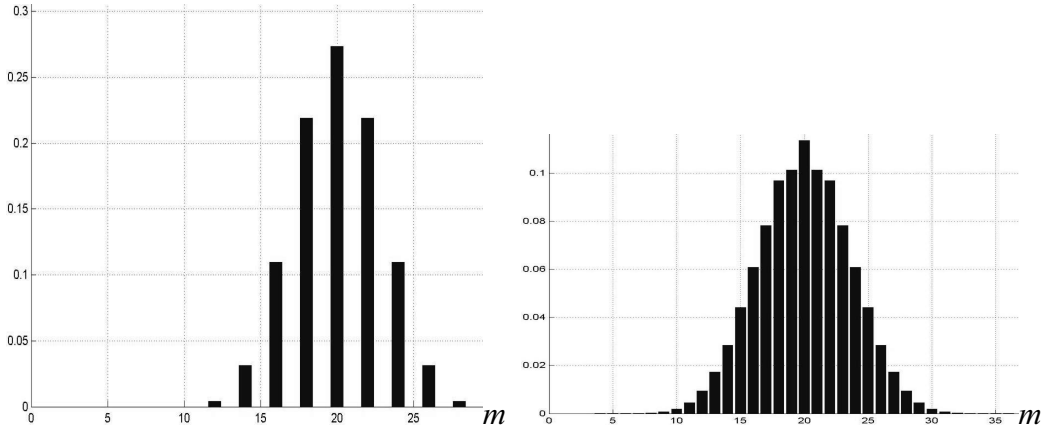


Рис.1-18. слева: $P(m/n=8)\ \ m_0=20\ \ p=0{,}5$
справа: $P(m/n=8)\ \ m_0=20\ \ p_1=0{,}4\ \ p_2=0{,}1$

В случае неравенства вероятностей p и q вид распределений количества повторов в маркере несколько усложняется. Ниже в качестве примера даны две пары графиков для $p=0{,}75\ (q=0{,}25)$ и, соответственно $p_1=0{,}6$ и $p_2=0{,}15$:

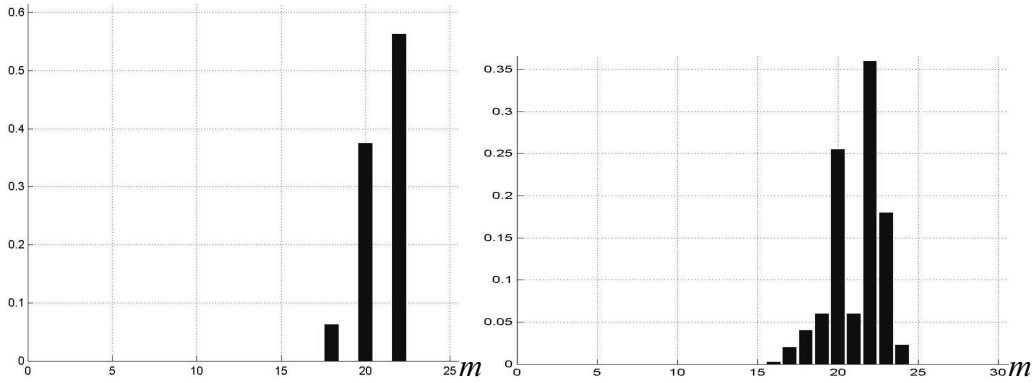


Рис.1-19. слева: $P(m/n=2)\ \ m_0=20\ \ p=0{,}75$
справа: $P(m/n=2)\ \ m_0=20\ \ p_1=0{,}6\ \ p_2=0{,}15$

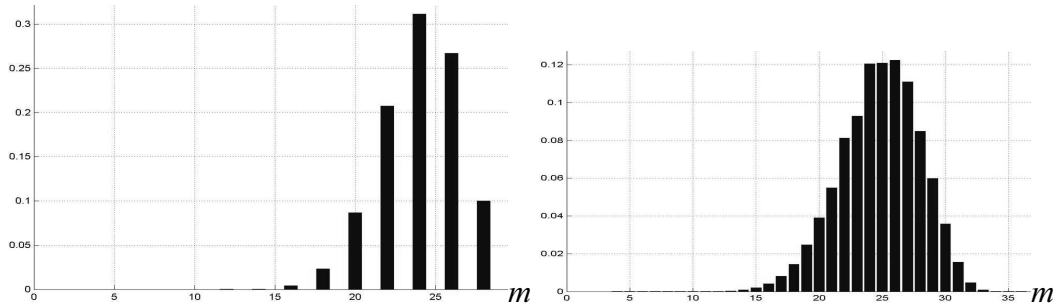


Рис.1-20. слева: $P(m/n=8)\ \ m_0=20\ \ p=0{,}75$
справа: $P(m/n=8)\ \ m_0=20\ \ p_1=0{,}6\ \ p_2=0{,}15$

1.4. Выводы

1. Разработана математическая модель изменения в результате мутаций количества повторов в маркерах гаплотипа в непрерывной мужской генеалогической линии от первопредка до современного потомка.

2. Показано, что изменение количества повторов в маркерах может быть представлено в виде целочисленного осциллирующего блуждания с возможностью изменения на +1, -1, +2 или -2 в результате мутации. Для блуждания с изменением на +1 или -1 распределение количества повторов в маркерах в результате мутации получено в явном виде (1.1.7), а для более сложного случая изменения на +1, -1, +2 или -2 в результате мутации получено удобное рекуррентное выражение (1.3.6), позволяющее точно вычислять требуемые вероятности для любых значений параметров распределения.

3. Поток мутаций на генеалогической линии от первопредка до финального потомка аппроксимирован пуассоновским потоком. Таким образом, вся математическая модель опирается на соотношения (1.1.7), (1.3.6), (1.2.3) и (1.2.5), которые были использованы и в алгоритмах вычисления максимально правдоподобных оценок параметров модели по совокупностям реальных гаплотипов (см. Приложение 2).

4. На основе разработанной математической модели был синтезирован и программно реализован алгоритм имитации совокупности гаплотипов финальных потомков, который использовался для статистических расчетов точности оценки количества мутаций, а также для верификации алгоритмов расчетов количества мутаций (см. главу 2).

ГЛАВА 2. Оценка количества мутаций, прошедших от первопредка до настоящего времени

В данной главе рассматриваются вопросы синтеза вычислительной процедуры, позволяющей на основе совокупности гаплотипов, представленных в виде числовой матрицы (см. Введение), получить по каждому маркеру среднее количество мутаций, произошедшее со времени от первопредка и реставрировать количество повторов в маркере, которое должно было быть у первопредка. В основу метода положена статистическая процедура поиска максимально правдоподобных оценок параметров (МП-оценки).

Также рассматриваются вопросы точности получаемых оценок. Исследование точности осуществлялось путем проведения статистических испытаний, для чего был разработан программный имитатор наборов гаплотипов, которые использовались в качестве тестовых данных. Для исследования метода оценки параметров предкового гаплотипа по большой совокупности гаплотипов потомков была создана специальная управляющая программа, обеспечивающая связку имитатора гаплотипов и рабочей программы исследуемого метода, а также позволяющая вычислить статистические характеристики (матожидание, дисперсию, доверительные интервалы и пр.) получаемых оценок параметров.

Одновременно с оценкой точности исследуемых методов на тех же самых тестовых выборках гаплотипов проводилась работа по верификации программного обеспечения.

2.1. Максимально правдоподобные оценки параметров распределения количества повторов в маркере

Для нас особый интерес представляют два параметра модели – среднее количество мутаций в маркере μ, прошедшее от первопредка исследуемой выборки гаплогрупп до финальных потомков, а также количество повторов в маркере m_0, которое должно было быть у первопредка (реставрированный гаплотип первопредка). Параметры модели p_1 и p_2 в задаче исторической

датировки времени жизни первопредка не используются, хотя точное их значение необходимо для определения первых двух параметров.

Таким образом, следует осуществить подбор совокупности четырех параметров $\{ p_1 , p_2 , m_0 , \mu \}$. Будем искать максимально правдоподобные оценки $\{ \hat{p}_1 , \hat{p}_2 \hat{m}_0 , \hat{\mu} \}$ параметров $\{ p_1 , p_2 , m_0 \mu \}$, которые вычисляются непосредственно из совокупности измерений $m_1, m_2, \ldots, m_N$ количества повторов m исследуемого маркера в каждом из N гаплотипов, составляющих исследуемую выборку.

Всю используемую в вычислениях выборку из N штук гаплотипов запишем в матрицу (см. Введение)

$$\begin{matrix} m_1(1) & m_1(2) & .... & m_1(M) \\ m_2(1) & m_2(2) & .... & m_2(M) \\ .... & .... & .... & .... \\ m_N(1) & m_N(2) & .... & m_N(M) \end{matrix} ,$$

из которой будем поочередно выбирать столбцы, соответствующие отдельному маркеру. Например, для некоего j -го маркера совокупность измерений $m_1, m_2, \ldots, m_N$ формируется следующим образом:

$$\begin{matrix} m_1 = m_1(j) \\ m_2 = m_2(j) \\ ..... \\ m_N = m_N(j) \end{matrix} .$$

Итак, пусть $P_j(m_i)$ - вероятность того, что j -м маркере отдельного i -го гаплотипа используемой выборки имеет место m_i повторов. Эта вероятность, как было найдено раньше, последовательно вычисляется по формулам (1.3.6) и (1.2.2). Вероятность $P_j(m_1, ..., m_N / p_1, p_2, m_0, \mu)$ того, что наблюденные повторы в исследуемом j -м маркере во всей выборке из N измерений (гаплотипов) составят конкретную совокупность значений $\{m_1, m_2, \ldots, m_N\}$ может быть представлена в виде следующего произведения (в силу независимости мутаций у отдельных гаплотипов):

$$P_j(m_1,\ldots,m_N/p_1,p_2,m_0,\mu)=\prod_{i=1}^{N}P_j(m_i/p_1,p_2,m_0,\mu) \qquad (2.1.1)$$

Совокупность оцененных значений параметров { $\hat{p}_1$, $\hat{p}_2$ $\hat{m}_0$, $\hat{\mu}$ }, при которых достигается максимум вероятности (2.1.1), получается совместным решением соответствующих уравнений правдоподобия, которые можно получить подстановкой в (2.1.1) выражений (1.3.6), (1.2.2) и (1.2.3):

$$P_j(m_1,\ldots,m_N/p_1,p_2,m_0,\mu)=\prod_{i=1}^{N}\sum_{n=0}^{\infty}P_j(m_i/n;p_1,p_2,m_0)\cdot P_{П}(n/\mu) \qquad (2.1.2)$$

где

$P_{П}(n/\mu)=\frac{\mu^n}{n!}e^{-\mu}$ - распределение количества мутаций n при заданном значении среднего количества мутаций μ в маркере и на интервале времени T (см. 1.2.3)

$P_j(m_i/n;p_1,p_2,m_0)$ - набор вероятностных распределений маркера для различных значений количества мутаций n (см. рекуррентное соотношение 1.3.6).

К сожалению, получается довольно громоздкая конструкция, из которой аналитически получить оценочные значения параметров { $\hat{p}_1$, $\hat{p}_2$ $\hat{m}_0$, $\hat{\mu}$ } затруднительно. Решено было использовать численный метод нахождения максимума вероятности (2.1.1) путем полного перебора по всем параметрам в пределах их возможных диапазонов. В этом случае точность вычисляемых таким образом оценок { $\hat{p}_1$, $\hat{p}_2$ $\hat{m}_0$, $\hat{\mu}$ } зависит лишь от производительности используемых вычислительных средств, к тому же гарантированно будет найден глобальный максимум.

Следует отметить некоторые особенности вычислительной процедуры. Умножение нескольких тысяч значений, каждое из которых меньше нуля, причем, некоторые из них вообще от нуля практически не отличаются, приводит к значительным погрешностям. Поэтому вместо поиска максимума выражения (2.1.2) велся поиск максимума гладкой функции от (2.1.2), в качестве которой был взят натуральный логарифм, при этом произведение (2.1.2) превращается в сумму:

$$\max_{p_1,\,p_2,\,m_0,\,\mu}\{\ln P_j(m_1,\ldots,m_N/\,p_1,\,p_2,\,m_0,\,\mu)\}=$$
$$=\max_{p_1,\,p_2,\,m_0,\,\mu}\{\sum_{i=1}^{N}\sum_{n=0}^{\infty}\ln\left(P_j(m_i/n;\,p_1,\,p_2,\,m_0)\cdot\frac{\mu^n}{n!}\mathrm{e}^{-\mu}\right)\} \qquad (2.1.3)$$

Рассмотрим теперь правила выбора диапазонов изменения параметров { p_1 , p_2 , m_0 μ ,} для вычислительной процедуры. Сначала для каждого исследуемого маркера по всем N измерениям строится гистограмма количества повторов $H(k,1{:}50)$. Здесь k - номер маркера $(k=1,\ldots,12)$, а (1:50) - диапазон возможных значений количества повторов (диапазон можно взять с большим запасом). Затем по гистограмме каждого маркера вычисляются среднее количество повторов m' и грубая оценка количества мутаций, произошедших от первопредка до настоящего времени, и приходящихся на данный маркер и на один гаплотип, по следующей формуле:

$$L(k)=\sum_r\left|H(k\,,m'+r)\right| \quad , \qquad (2.1.4)$$

где $r=-R,\ldots,0,\ldots,+R$, а величина R определяется из границ гистограммы (т.е., суммирование ведется по всем ненулевым значениям гистограммы).

Таким образом, диапазон изменения m_0 следующий: от $(m'-2)$ до $(m'+2)$, не более (итого - 5 значений), а в подавляющем большинстве случаев можно ограничиться только тремя значениями $(m'-1,m',m'+1)$.

Диапазон изменения параметра μ можно задавать от $L(k)/5$ до $5\mathrm{L}(k)$. Дискретности в 500 или 1000 значений в указанном диапазоне будет более чем достаточно.

Вероятность p_1 , очевидно, лежит в пределах от 0 до 1 (реально, от 0,1 до 0,95). Указанный диапазон можно разбить на 50-100 значений, что даст вполне приемлемую точность.

Параметр p_2 в большинстве случаев вообще можно исключить (сделать равным нулю), ограничившись только вероятностью изменения маркера на +1 или -1 в результате мутации. Тем не менее, можно и оставить параметр p_2 , ограничившись 3-5 значениями, близкими к нулю.

Итак, учитывая независимость изменения всех четырех параметров, общее количество переборов не превышает величины $5\times1000\times100\times5=2{,}5$

миллиона. Число переборов получилось небольшим, учитывая несложность вычислений для каждого варианта.

И, наконец, правило вычисления вероятностей $P_j(m_i/p_1, p_2, m_0, \mu)$ следующее. Для каждого набора параметров $\{p_1, p_2, m_0\}$ вычисляется спектр вероятностей (1.3.6) для разных возможных значений количества повторов m. Затем, для каждого слагаемого (2.1.4) из этого спектра выбирается то значение $P_j(m_i/...)$, которое соответствует величине m_i. Учитывая, что величины m_i дискретные и лежат обычно в довольно узком диапазоне относительно m_0, не превышающем десяти «вверх» или «вниз», формирование такого ограниченного спектра вероятностей также проблемы не представляет.

Варьирование параметра μ влияет на значения коэффициентов в формуле (1.2.2) в соответствии с (1.2.3) и (1.2.4), которые используются при вычислении вероятности $P_j(m_i/p_1, p_2, m_0, \mu)$.

Описание алгоритма вычисления МП-оценок параметров $\{\hat{p}_1, \hat{p}_2 \hat{m}_0, \hat{\mu}\}$ методом перебора приведено в Приложении 2.1.

2.2. Определение точности МП-оценки среднего количества мутаций, прошедших от первопредка до настоящего времени

Исследованию подвергалась точность метода максимально правдоподобной оценки (МП-оценки) среднего количества мутаций μ в отдельном маркере, прошедших от первопредка до настоящего времени (теоретическое описание метода дано в п.2.1). В качестве исходной статистики использовалась большая совокупность гаплотипов, заведомо принадлежащих одной гаплогруппе.

В первую очередь интерес представляют как наличие смещения этой оценки, так и границы доверительного интервала. Поскольку аналитически оценить данные параметры пока не представилось возможным, то эту задачу будем решать с помощью статистического моделирования значительного количества выборок одномаркерных гаплотипов, которые упаковываются в виде числовой матрицы, структура которой описана во Введении.

2.2.1. Описание алгоритма

Для получения среднего значения, дисперсии и доверительных интервалов МП-оценок μ и p_1 необходимо выполнить следующие шаги:

ШАГ 1. Задаем исходные значения параметров модели одномаркерных гаплотипов:

1.1. μ_0 ,

1.2. p_1 .

1.3. $p_2=0$,

1.4. Конкретное значение параметра m_0 значения не имеет, но для определенности зададим его равным 20.

1.5. Размер N моделируемой совокупности одномаркерных гаплотипов.

1.6. Размер M (количество маркеров) совокупности оценок параметра μ .

ШАГ 2. Используя рекуррентные соотношения (1.3.6) формируем двумерную таблицу вероятностей (предельное значение n берем с запасом, например, 20)

$$\begin{aligned}
&P(m+2/n+1):=P(m+2/n+1)+P(m/n)\cdot p_2\\
&P(m+1/n+1):=P(m+1/n+1)+P(m/n)\cdot p_1\\
&P(m/n+1) \ :=P(m/n+1) \ +0 \quad \text{для всех} \quad m=(m-n),\ldots,(m+n)\\
&P(m-1/n+1):=P(m-1/n+1)+P(m/n)\cdot q_1\\
&P(m-2/n+1):=P(m-2/n+1)+P(m/n)\cdot q_2
\end{aligned} \qquad (2.2.1)$$

Из таблицы (2.1.1) формируем новую таблицу, содержащую интегральные законы распределения для всех n от 1 до 20:

$$F(m/n)=\sum_{r=1}^{m} P(r/n) \quad , \quad (n=0,1,\ldots) \qquad . \qquad (2.2.2)$$

ШАГ 3. Организуем цикл по моделируемым одномаркерным гаплотипам (по k от 1 до N) для получения массива значений оценок $\{\hat{\mu}_1,\ldots,\hat{\mu_N}\}$ параметра μ .

3.1. С помощью датчика случайных чисел rand(), равномерно распределенных на интервале (0,1), получаем случайное число n, распределенное по Пуассону с параметром μ. Процедура основана на том факте, что интервалы между пуассоновскими событиями распределены экспоненциально. Суммируется столько интервалов, сколько нужно, чтобы выйти за границы общего интервала длиной μ (в данном случае суммирование экспоненциально распределенных чисел заменено произведением равномерно распределенных чисел, но общий интервал при этом вычисляется как $\exp(-\mu)$).

```
b := exp(-Mu0);
r := 0;
s := rand();
до тех пор, пока s>b делать:
      s=s*rand();
      r=r+1;
Иначе
      n = r;
Конец процедуры получения случайного пуассоновского числа
```

3.2. С помощью датчика случайных чисел, равномерно распределенных на интервале (0,1), формируем случайное число x.

3.3. Проверяем, в какой из интервалов попало x : $\left[0, F_n(1)\right]$, $\left[F_n(1), F_n(2)\right]$, ... , $\left[F_n(n),1\right]$.
Если $\left[F_n(i-1)<x\le F_n(i)\right]$, то в качестве количества повторов в k-м гаплотипе принимаем дискретную величину i ($m(k)=i$).

3.4. Конец цикла по моделируемым гаплотипам.

3.5. Методом подбора по смоделированной на шаге 3.1 совокупности N одномаркерных гаплотипов $\{m_1,\dots,m_N\}$ вычисляем очередные максимально правдоподобные оценки { $\rho_1(k)=\hat{p}_1$, $\rho_2(k)=\hat{p}_2$ $\hat{m}_0(k)$, $\hat{\mu}_k$, используя алгоритм, описанный в параграфе 2.1. Здесь каждое значение z_k есть количество повторов в единственном маркере моделируемого гаплотипа.

3.6. Конец цикла по массиву значений оценок k.

ШАГ 4. По полученному массиву МП-оценок среднего количества мутаций $\{\hat{\mu}_1, ..., \hat{\mu}_N\}$:

4.1. вычисляем среднее значение МП-оценки среднего количества мутаций $\hat{\mu}_0(N)$ в отдельном маркере (здесь и далее (N) означает, что величина $\hat{\mu}_0$ является функцией длины выборки N):

$$\hat{\mu}_0(N) = \frac{1}{N}\sum_{k=1}^{N}\hat{\mu}_k \qquad (2.2.3)$$

4.2. вычисляем дисперсию и среднеквадратическое отклонение МП-оценки среднего количества мутаций в маркере

$$D_\mu(N) = \frac{1}{N-1}\sum_{k=1}^{N}(\hat{\mu}_k - \hat{\mu}_0(N))^2 \qquad (2.2.4a)$$

$$\tilde{\sigma}_\mu(N) = \frac{\sqrt{D_\mu(N)}}{\mu_0} \qquad (2.2.4b)$$

4.3. вычисляем относительное смещение МП-оценки относительно «истинного» значения μ_0 :

$$\tilde{S}_\mu(N) = \frac{\hat{\mu}_0(N) - \mu_0}{\mu_0} \quad . \qquad (2.2.5)$$

Здесь следует сделать одно замечание. Число n , распределенное по Пуассону, которое мы моделируем, имеет чуть заниженные (на 1-2%) значения среднего и дисперсии за счет того, что последний интервал, выходящий за пределы общего интервала, завершающий процедуру вычисления n , на самом деле не закончился последним событием. Чтобы учесть этот факт, в (2.2.5) вместо значения $\hat{\mu}_0$ подставляется вычисленное по смоделированным пуассоновским числам среднее значение.

4.4. Границы доверительных интервалов вычисляем упрощенно по уровню $\pm\sqrt{D_\mu(N)}$ и по уровню $\pm 2\sqrt{D_\mu(N)}$ относительно среднего значения $\hat{\mu}_0(N)$, предполагая «нормальность» распределения оценки величины μ .

При этом уровни значимости α считаются равными 15,85% и 2,25% соответственно.

ШАГ 5. По полученному массиву МП-оценок $\{\rho_1,\dots,\rho_N\}$ вероятности p_1 :

5.1. вычисляем среднее значение МП-оценки

$$\hat{p}_1(N)=\frac{1}{N}\sum_{k=1}^{N}\rho_k \qquad (2.2.6)$$

5.2. вычисляем дисперсию и среднеквадратическое отклонение МП-оценки

$$D_p(N)=\frac{1}{N-1}\sum_{k=1}^{N}(\rho_k-\hat{p}_1(N))^2 \qquad (2.2.7a)$$

$$\tilde{\sigma}_p(N)=\frac{\sqrt{D_p(N)}}{p_1} \qquad (2.2.7b)$$

5.3. вычисляем относительное смещение МП-оценки относительно «истинного» значения p_1 :

$$\tilde{S}_p(N)=\frac{\hat{p}_1(N)-p_1}{p_1} \quad . \qquad (2.2.8)$$

ШАГ 6. Конец алгоритма.

2.2.2. Верификация алгоритма

В алгоритме используется стандартная процедура rand() формирования случайных некоррелированных чисел, равномерно распределенных на интервале [0,1]. Данную процедуру верификации не подвергаем.

На шаге 3.1.1 формируются целые неотрицательные случайные числа, распределенные по Пуассоновскому закону со средним значением, соответствующим заданному значению μ . Чтобы проверить правильность работы данной процедуры, вычисляется среднее значение и дисперсия

последовательности из N таких случайных чисел, которые формируются в одном цикле вычисления МП-оценок.

Для N =3000 были получены следующие результаты:

- истинное значение параметра μ = 0.5000
- матожидание N моделируемых «пуассоновских» чисел 0.4993
- дисперсия N моделируемых «пуассоновских» чисел 0.4992

Очевидно, что имеет место практически точное совпадение матожидания и дисперсии моделируемой последовательности «пуассоновских» чисел, что и должно было быть. Чуть меньшее значение среднего значения моделируемой последовательности по сравнению с заданным значением 0.5 объясняется тем фактом, что моделируются не сами пуассоновские числа (количества пуассоновских событий), а экспоненциально распределенные интервалы между событиями. Последний интервал незавершенный, поскольку процедура накопления на шаге 3.1.1 прерывается по условию `s>b`.

Процесс моделирования количества повторов в маркере по заданным рекуррентным соотношениям (1.3.6) проверяется фактически по результатам совпадения (малого различия) вычисленных МП-оценок параметров μ, p_1, m_0 и заданных значений этих же параметров в качестве исходных, поскольку одни и те же процедуры (1.3.6) и (2.2.5) используются как при моделировании на шаге 3.1.1, так и при вычислении МП-оценок на шаге 3.2 алгоритма.

2.3. Анализ свойств максимально правдоподобных оценок параметров модели потока мутаций

В данном параграфе приведены результаты исследований свойств МП-оценок параметров μ, m_0 и суммарной вероятности $p = p_1 + p_2$, получаемых с помощью алгоритма обработки большой совокупности гаплотипов (см. п.2.1).

Одной из проблем является определение минимального объема выборки гаплотипов, по которой можно вычислить искомые параметры с приемлемой точностью. К сожалению, не всегда для исследования можно

получить достаточное количество гаплотипов, принадлежащих заведомо интересующей нас популяции. Поэтому в качестве базового параметра (аргумента) возьмем количество гаплотипов N в исследуемой выборке.

Особое внимание уделим таким свойствам максимально правдоподобных оценок (МП-оценок), как *смещенность* и *состоятельность* (в смысле поведения дисперсии оценки с ростом количества гаплотипов N в исследуемой выборке).

Чтобы выявить основные особенности поведения МП-оценок, использовался достаточно большое количество значений исходных параметров.

2.3.1. Максимально правдоподобная оценка среднего количества мутаций

Значение обобщенного параметра μ (среднего количества мутаций в исследуемом маркере) задавалось в широких пределах от 0,1 до 2,0. Вероятность p_1 задавалась тремя значениями: 0.25, 0.5 (симметричное блуждание) и 0.75. Поскольку в большинстве случаев вероятность p_2 была или слишком мала, или вообще равнялась нулю, то значения вероятностей p_1 и p в большинстве случаев – одно и то же.

На рис. 2-1а представлены зависимости относительного смещения $\tilde{S}_\mu(N)$ (2.2.5) максимально правдоподобной оценки (МП-оценки) среднего количества мутаций в маркере μ как функции количества гаплотипов N в выборке при дополнительном параметре (вероятности) p_1 .

Для удобства анализа относительные смещения $\tilde{S}_\mu(N)$ представлены в процентах. В качестве примеров приведены две группы графиков для $\mu=0{,}1$ и $\mu=1{,}5$:

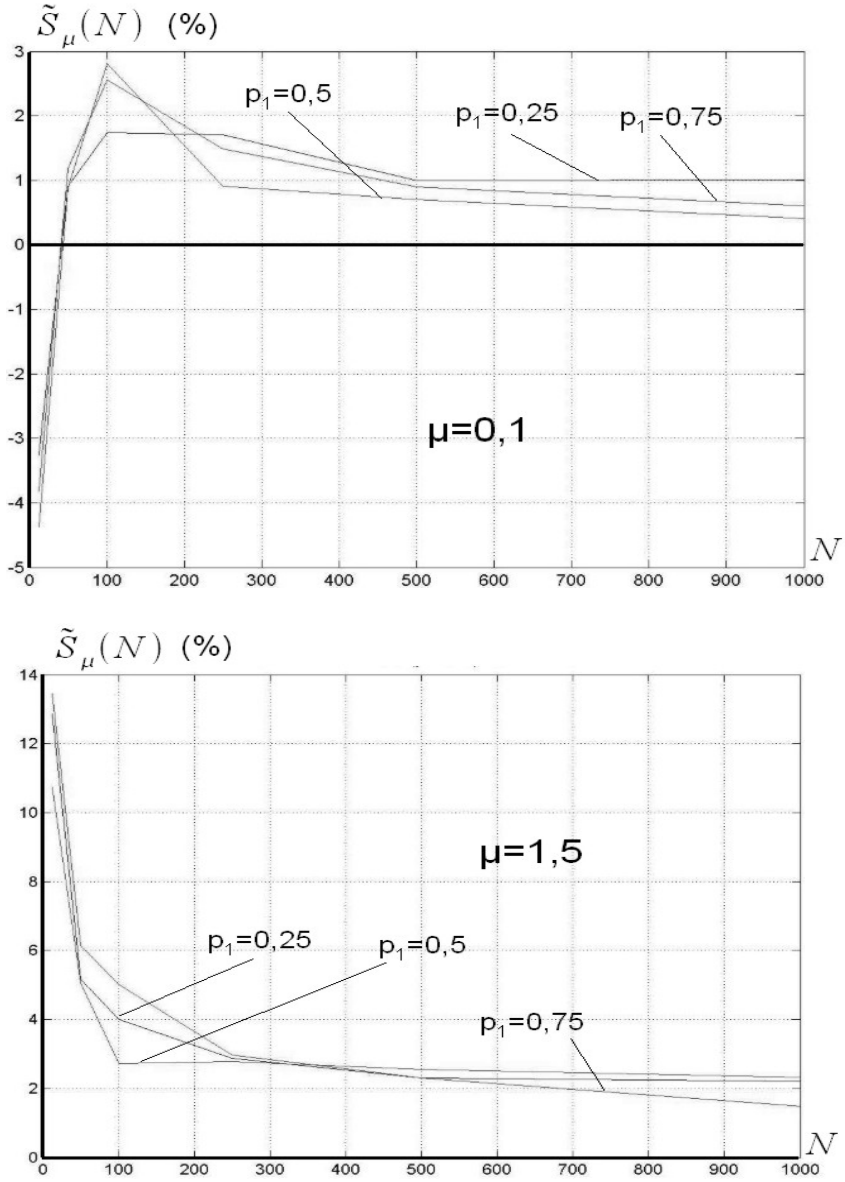


Рис. 2-1а. Зависимости относительного смещения $\tilde{S}_\mu(N)$ МП-оценки количества мутаций μ для различных «перекосов» гистограмм (вероятность p_1) с ростом количества гаплотипов N

Очевидно, что какой-то определенной зависимости величины смещения оценки среднего количества мутаций μ от значения вероятности p_1 не наблюдается. Поэтому имеет смысл усреднить все три графика (для

$p_1 = 0{,}25\ \ 0{,}5\ \ 0{,}75$) и представить в виде одной зависимости для каждого конкретного значения среднего количества мутаций $\mu = 0{,}1$ и $\mu = 1{,}5$:

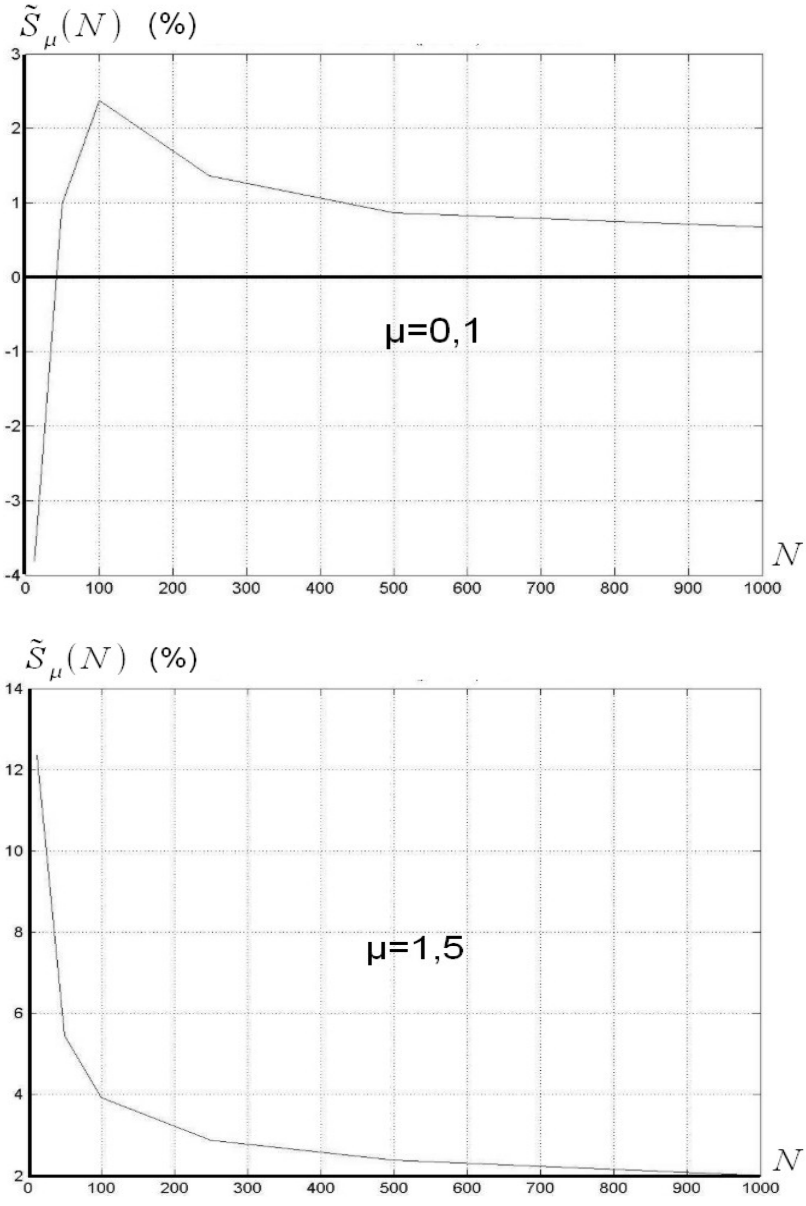


Рис. 2-1b. Усредненная зависимость относительного смещения $\tilde{S}_\mu(N)$ МП-оценки количества мутаций μ с ростом количества гаплотипов N

На рис. 2-2 представлены зависимости относительного среднеквадратического отклонения $\tilde{\sigma}_{\mu}(N)$ МП-оценки параметра μ (2.2.4а), причем, также в виде функций от количества N гаплотипов в выборке. В результате моделирования выяснилось, что данная характеристика практически не зависит от вероятности p_1. Поэтому имеет смысл усреднить все три графика (для $p_1 = 0{,}25\ \ 0{,}5\ \ 0{,}75$) и представить в виде одной зависимости для каждого конкретного значения среднего количества мутаций $\mu = 0{,}1$ и $\mu = 1{,}5$. Для удобства анализа усредненное таким образом среднеквадратическое отклонение дополнительно переведено в проценты:

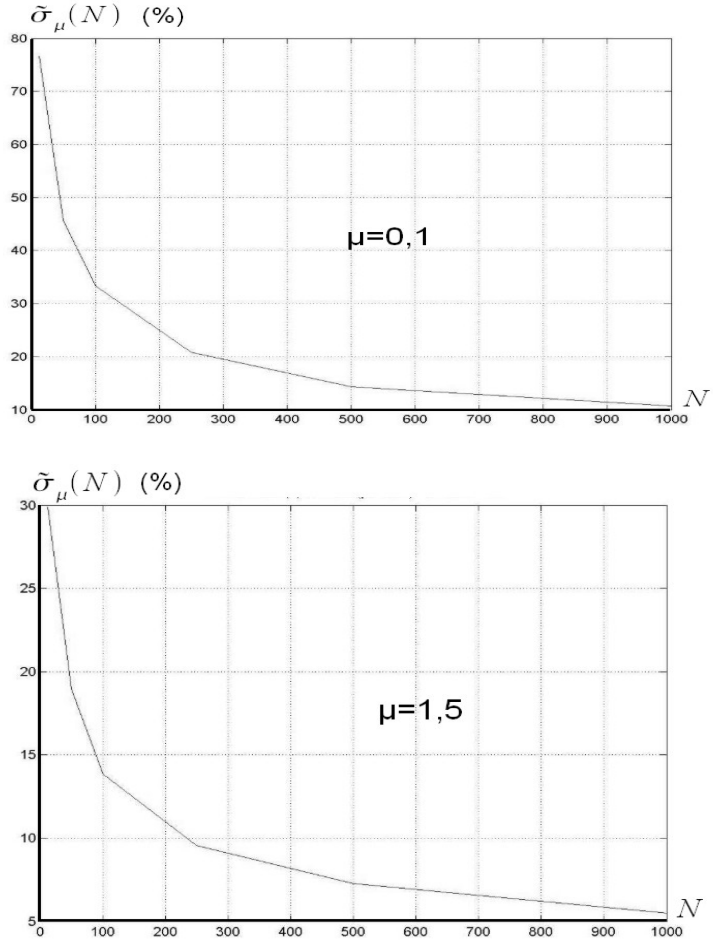


Рис.2-2. Зависимость относительного среднеквадратического отклонения $\tilde{\sigma}_{\mu}(N)$ МП-оценки количества мутаций μ с ростом количества гаплотипов N

Зависимость, представленная на рис.2-2, фактически дает полуширину доверительного интервала, вычисленную по так называемому уровню «одной сигмы», т.е., по уровню $\pm\sqrt{D_\mu(N)}$, что для нормального распределения соответствует уровню значимости $\alpha = 15{,}85$ %. В случае использования уровня «двух сигм» ($\pm 2\sqrt{D_\mu(N)}$), это будет соответствовать уровню значимости $\alpha = 2{,}25$ %.

Как следует из рис.2-2, с ростом объема выборки среднеквадратическое отклонение монотонно уменьшается, следовательно, МП-оценка среднего количества мутаций μ является *состоятельной*.

На рис.2-3 представлены сразу несколько зависимостей относительного смещения $\tilde{S}_\mu(N)$ МП-оценок среднего количества мутаций μ для различных его истинных (т.е., моделируемых) значений (в процентах):

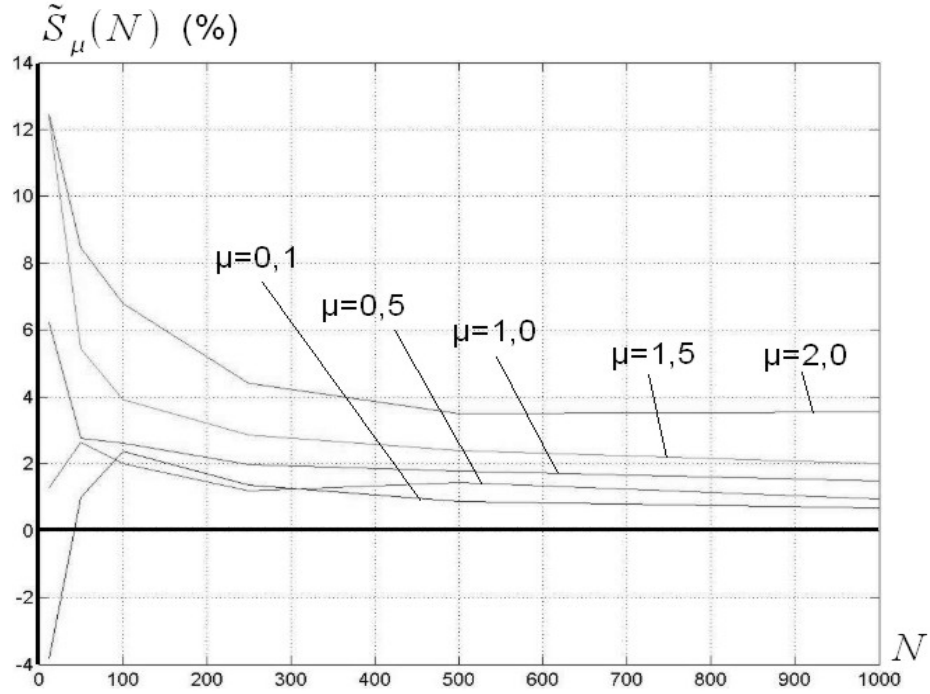


Рис.2-3. Изменение относительного смещения $\tilde{S}_\mu(N)$ МП-оценок среднего количества мутаций μ с ростом количества гаплотипов N и при различных истинных значениях μ

Как следует из рис.2-3, МП-оценка среднего количества мутаций μ является смещенной, причем, смещение $\tilde{S}_\mu(N)$ растет с ростом истинного

значения параметра μ . Вместе с тем, смещение при больших значениях среднего количества мутаций μ не превысит 5% от истинного значения.

Из рис.2-3 также следует, что для объемов выборки гаплотипов $N = 0 \div 200$ качество МП-оценки среднего количества мутаций μ довольно низкое (смещение резко возрастает со значений в несколько процентов при $N = 200$ и до десяти и более при $N < 20$).

На рис. 2-4 представлены сразу несколько зависимостей относительного среднеквадратического отклонения МП-оценки среднего количества мутаций μ :

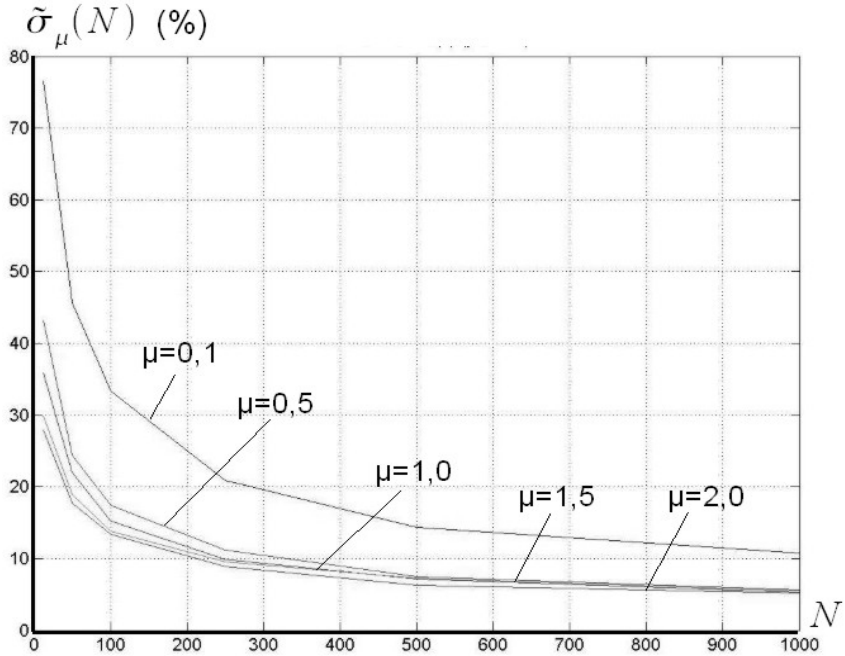


Рис. 2-4. Относительное среднеквадратическое отклонение $\tilde{\sigma}_\mu(N)$ МП-оценки среднего количества мутаций μ с ростом количества гаплотипов N в выборке и при различных его истинных значениях μ

Как следует из рис.2-4, зависимость относительного среднеквадратического отклонения $\tilde{\sigma}_\mu(N)$ с ростом среднего количества мутаций μ от величины 0.5 и выше практически уже не изменяется. Только для малых величин μ имеет место увеличение дисперсии оценки среднего количества мутаций.

Вместе с тем, следует отметить факт очень слабого убывания с ростом объема выборки гаплотипов, начиная с величин $N=500$ и более, причем, уровень в 5% практически предельно достижимый. Таким образом, для уровня «двух сигм» размер доверительного интервала будет составлять ± 10 % при больших объемах выборки гаплотипов, а при объемах в $N=300$ будет значительно большим - ± 40 % (!).

2.3.2. Максимально правдоподобная оценка вероятности мутации количества повторов «вверх»

Несмотря на то, что в нашей задаче параметр p_1 является вспомогательным, тем не менее, от правильности его вычисления зависит и правильность получаемой оценки главного параметра - среднего количества мутаций μ .

На рис.2-5 представлены две группы зависимостей отношения среднего значения МП-оценки вероятности $\hat{p}_1(N)$ к ее истинному значению p_1 (для трех значений $p_1=0.25,\ 0.5,\ 0.75$).

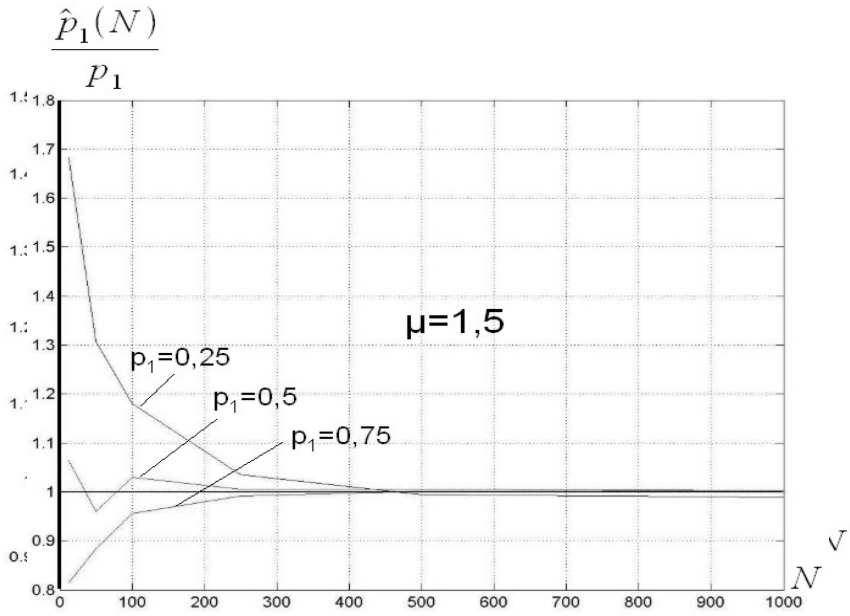


Рис. 2-5. Поведение отношения среднего значения МП-оценок вероятности p_1 к ее истинному значению с ростом объема выборки гаплотипов N

На рис. 2-6 представлена зависимость $\tilde{\sigma}_p(N)$ относительного среднеквадратического отклонения МП-оценки вероятности p_1 (2.2.7а), усредненная по трем исходным значениям параметра $p_1 = 0.25,\ 0.5,\ 0.75$ (как показало статистическое моделирование, все три графика среднеквадратического отклонения практически идентичны). Также для удобства зависимость дана в процентах от истинного (моделируемого) значения вероятности p_1 .

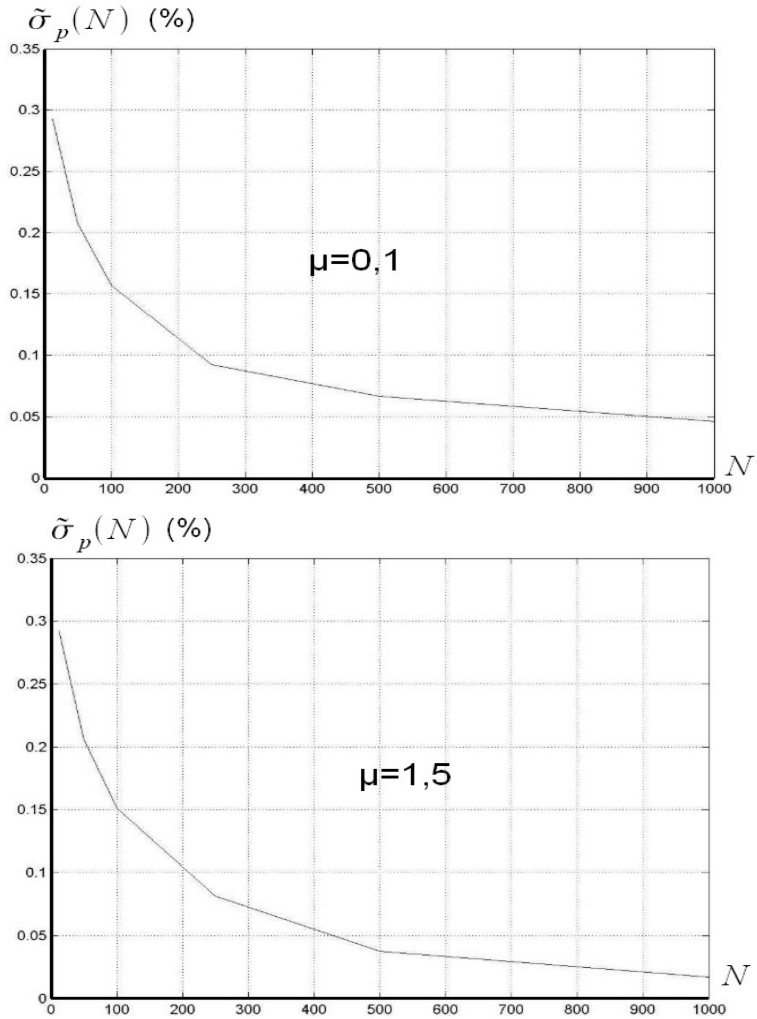


Рис. 2-6. Поведение относительного среднеквадратического отклонения $\tilde{\sigma}_p(N)$ МП-оценки параметра p_1 с ростом объема выборки гаплотипов N

Как следует из рис.2-5 и 2-6, только при объемах выборки гаплотипов не менее $150 \div 200$ штук оценка вероятности p_1 становится близкой к ее истинному (моделируемому) значению.

На рис. 2-7 представлены сразу несколько зависимостей относительного среднеквадратического отклонения $\tilde{\sigma}_p(N)$ МП-оценки вероятнсти p_1 :

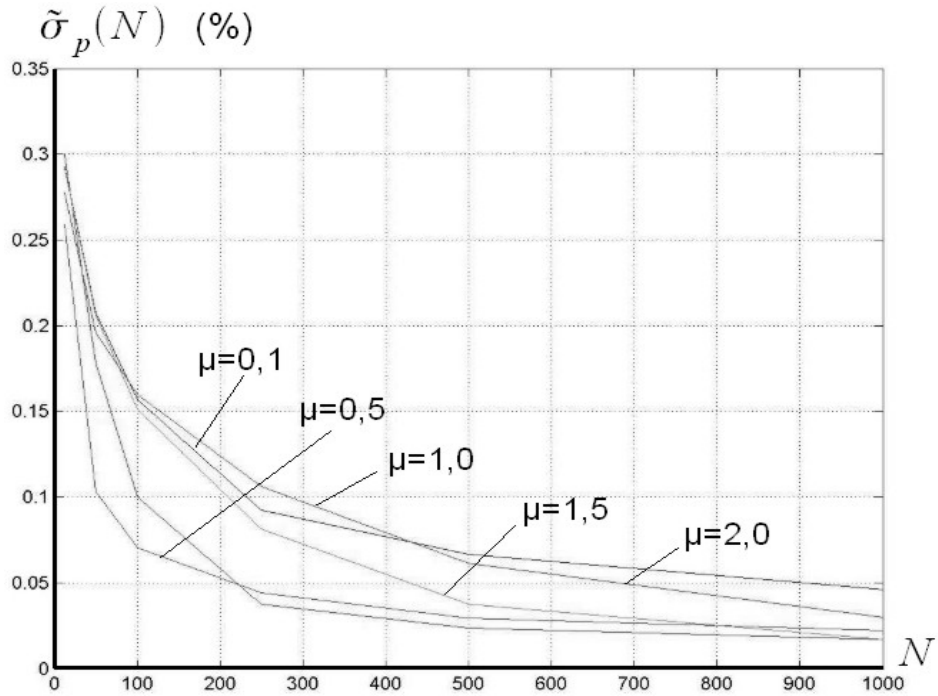


Рис.2-7. Изменение относительного среднеквадратического отклонения $\tilde{\sigma}_p(N)$ МП-оценки вероятности p_1 с ростом объема выборки гаплотипов N

Как следует из рис.2-7, какой-либо монотонности поведения относительного среднеквадратического отклонения $\tilde{\sigma}_p(N)$ оценки вероятности p_1 для одного и того же значения N , но с ростом среднего количества мутаций μ , не наблюдается. Вместе с тем, сами значения $\tilde{\sigma}_p(N)$ довольно малы и разброс лежит в пределах статистических погрешностей эксперимента.

2.3.3. Максимально правдоподобная оценка количества повторов в маркере у первопредка

И, наконец, рассмотрим качество определения количества повторов m_0 в маркере, которое должно было быть у первопредка. Если бы не было перекосов формы гистограмм количество повторов маркеров, то эта задача решалась бы весьма просто, а именно, нахождением матожидания, значение которого являлось бы одновременно и модой этого распределения. На самом деле всегда присутствуют перекосы, причем встречаются довольно значительные. В этом случае количество повторов в маркере у первопредка, вычисленное как матожидание, может отличаться от истинного.

Для определенности при моделировании истинное значение параметра m_0 у первопредка всегда задавалось равным 20.

При малых значениях параметра μ (меньше 0.5) для любых объемов выборки от $N=12$ до $N=1000$ всегда точно определялось значение $m_0=20$, а дисперсия оценки всегда равна нулю.

Для значений параметра μ от 0.5 и до 2.0 монотонно нарастают ошибки в определении параметра m_0, причем, для симметричной гистограммы количества повторов (симметричность имеет место, как мы помним, при $p_1=0{,}5$) результаты оценки наилучшие, а для $p_1=0{,}25$ и $p_1=0{,}75$ результаты хуже, причем, для $p_1=0{,}25$ получается заниженная оценка параметра m_0, а для $p_1=0{,}75$ – ровно настолько же завышенная.

Как и следовало ожидать, с ростом объема выборки N для одного и того же параметра μ качество оценки параметра m_0 растет, и дисперсия оценки соответственно уменьшается.

Для иллюстрации приведем результаты для больших значений параметра $\mu=1.0,\ 1.5\ 2.0$ [мутаций/маркер]. Левый предел в парах чисел – для количества гаплотипов в выборке $N=12$, а правый предел – для $N=1000$. Верхняя пара – пределы матожидания МП-оценки параметра m_0, а нижняя пара – относительное среднеквадратическое отклонение оценки параметра m_0 (в процентах):

Таблица 2-1. Качество оценки количества повторов в гаплотипе первопредка

Среднее количество мутаций в маркере, «отделяющее» гаплотип первопредка от настоящего времени	$p_1 = 0{,}25$ («горб» гистограммы смещен влево)	$p_1 = 0{,}5$ (симметричная гистограмма)	$p_1 = 0{,}75$ («горб» гистограммы смещен вправо)
$\mu = 1{,}0$	19.65 – 20 1.77% - 0%	19.988 – 20 0.06% - 0%	20.26 – 20 -1.31% - 0%
$\mu = 1{,}5$	19.46 – 19.998 2.7% - 0.04%	19.94 – 19.999 0.32% - 0.03%	20.5 – 20.01 -2.58% - -0.05%
$\mu = 2{,}0$	19.3 – 19.996 3.54% - 0.02%	20.01 – 19.994 -0.06% - 0.03%	20.8 – 20.004 -3.83% - -0.02%

В качестве пояснения рассмотрим нижнюю строку таблицы 2-1 (для $\mu = 2{,}0$) Как видим, при $p_1 = 0{,}25$ и при $N = 12$ матожидание количества повторов m_0 равно величине 19.3, т.е., в 70% случаев алгоритм выдавал значение $m_0 = 19$, и лишь в 30% случаев — истинное (моделируемое) значение $m_0 = 20$.

Аналогичная картина имеет место и для $p_1 = 0{,}75$, только в другую сторону: в 80% случаев алгоритм выдавал значение $m_0 = 21$, и только в 20% случаев — истинное значение $m_0 = 20$. При $p_1 = 0{,}5$ даже малое количество гаплотипов в выборке ($N = 12$) приводит практически всегда к правильной оценке количества повторов $m_0 = 20$ (матожидание равно 20.01, а, значит, только в 1% случаев имеет место принятие решения о количестве повторов, равном 21 вместо 20).

Таким образом, наличие сильного перекоса формы в гистограмме количества повторов приводит к значительному отличию количества повторов в гаплотипе первопредка (в один и даже два повтора) вычисленному как матожидание гистограммы, и с помощью алгоритма построения максимально правдоподобных оценок.

* * * * * * * * * *

В заключение следует отметить очень важное обстоятельство, а именно, максимально правдоподобные оценки (МП-оценки) параметров μ , p_1 и m_0 во всех моделируемых случаях достигались одновременно, что указывает на определенную монотонность многомерной функции правдоподобия $P_j(m_1, ..., m_N / p_1, m_0, \mu)$, а также отсутствие седловин, гребней и прочих сложностей рельефа, которые могли бы приводить к неправильным результатам по какому-нибудь одному из трех параметров при правильных значениях двух остальных.

2.4. Выводы

1. Разработан и программно реализован статистический алгоритм вычисления по совокупности гаплотипов следующих максимально правдоподобных оценок:

- количество повторов m_0 в исследуемом маркере гаплотипа первопредка;
- количество μ реально осуществившихся мутаций в каждом маркере (с учетом возвратных мутаций) со времени, прошедшем от первопредка.

2. С помощью статистического моделирования на имитаторе гаплотипов были изучены свойства вычисляемых максимально правдоподобных оценок параметров μ и m_0 .

3. Хорошее качество оценок получается для совокупности гаплотипов не менее $200 \div 300$ в выборке. Это, по-видимому, связано с неприятной особенностью процесса изменения количества повторов в маркере, заключающейся в периодическом возвращении процесса в исходное состояние и «отрезании» таким образом предыстории процесса.

ГЛАВА 3. Оценка среднего количества мутаций до первопредка по единственному гаплотипу

Иногда возникает задача исторической датировки останков индивида, генетический анализ которых точно выявил гаплотип и гаплогруппу. Как и для совокупности гаплотипов, принадлежащих одной и той же гаплогруппе, анализ времени жизни будет проводиться по количеству осуществившихся мутаций, прошедших от первопредка до времени жизни исследуемого индивида.

Само по себе количество повторов в маркерах единичного гаплотипа не несет никакой информации о прошедших мутациях. Иное дело, если мы имеем хотя бы два гаплотипа, заведомо принадлежащих одной и той же гаплогруппе, причем, находящихся на единой линии родства (их соединяет последовательность ребер генеалогического древа гаплогруппы). В этом случае можно уже говорить о количестве мутаций, произошедших за время между ними. Естественно в качестве опорного принять гаплотип первопредка.

Для поиска среднего количества мутаций, отделяющих время жизни первопредка и исследуемого индивида используем численный поиск максимально правдоподобной оценки (МП-оценки).

3.1. Максимально правдоподобная оценка среднего количества мутаций по единственному гаплотипу

Итак, нам понадобится следующая исходная информация:

- тестируемый гаплотип потомка $\{m^{(1)}(1), m^{(1)}(2), \ldots, m^{(1)}(M)\}$;
- гаплотип первопредка $\{m^{(0)}(1), m^{(0)}(2), \ldots, m^{(0)}(M)\}$ (M – количество маркеров);
- среднее количество мутаций в маркерах, прошедшее со времени первопредка до нашего времени, для гаплогруппы в целом $\{\mu_1, \mu_2, \ldots, \mu_M\}$;

- набор вероятностей p_1 по маркерам для гаплогруппы в целом $\{p_1(1), p_1(2), \ldots, p_1(M)\}$;
- набор вероятностей p_2 по маркерам для гаплогруппы в целом $\{p_2(1), p_2(2), \ldots, p_2(M)\}$.

Вся вышеуказанная информация по гаплогруппе и предковому гаплотипу, кроме тестируемого гаплотипа, вычисляется по методике, изложенной ранее в параграфе 2.1.

В силу характера блужданий, которому подчиняются изменения количества повторов в маркерах, можно количества повторов в маркерах гаплотипа первопредка обнулить, а в качестве повторов тестируемого гаплотипа сделать следующие разности (возникающие таким образом отрицательные количества повторов ничуть не хуже положительных значений):

$$\begin{aligned} m_1 &= m^{(1)}(1) - m^{(0)}(1) \\ m_2 &= m^{(1)}(2) - m^{(0)}(2) \\ &\ldots\ldots\ldots\ldots \\ m_M &= m^{(1)}(M) - m^{(0)}(M) \end{aligned} \quad . \qquad (3.1.1)$$

Пусть имеется набор $\{m_1, m_2, \ldots, m_M\}$. Вероятность возникновения данного набора в одном гаплотипе

$$P\left(m_1, m_2, \ldots, m_M / p_1(1), \ldots, p_1(M); p_2(1), \ldots, p_2(M); \lambda_1, \ldots, \lambda_M; T\right) \quad . \qquad (3.1.2)$$

В (3.1.2) вместо обобщенных параметров $\{\mu_1, \mu_2, \ldots, \mu_M\}$ использованы интенсивности мутаций $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ [мутации/(маркер×время)] и как отдельный параметр – время T , прошедшее от первопредка до тестируемого гаплотипа (это время общее для всех маркеров и именно его мы ищем).

Очевидно, что среди всех параметров в (3.1.2) только время T является варьируемым (остальные параметры считаются вычисленными на этапе реконструкции гаплотипа первопредка, и представляют собой константы) и именно его требуется найти, то максимально правдоподобная оценка времени T находится максимизацией выражения (3.1.2). Поскольку процессы мутаций в маркерах не зависят друг от друга, то максимум функции правдоподобия (3.1.2) можно представить в следующем виде:

$$\max_{T}\{\prod_{j=1}^{M} P(m_j / p_1(j), p_2(j), \lambda_j; T)\} \quad . \qquad (3.1.3)$$

Из (1.3.6) следует, что вероятность возникновения количества повторов m в маркере в результате мутации напрямую никак не связана с параметрами λ_j и T, а зависит только от наборов вероятностей p_1 и p_2, а также от реально осуществившегося количества мутаций n. Но именно величина n, в свою очередь, связана с интенсивностью мутаций λ, и со временем T, которое прошло от первопредка до исследуемого гаплотипа.

Варьируя величиной T в процессе поиска максимума в (3.1.3), мы одновременно изменяем интервал времени накопления мутаций в каждом маркере, а, следовательно, синхронно изменяем среднее количество мутаций в маркерах, которые могут осуществиться за этот интервал времени. Но при одном и том же среднем количестве мутаций в маркере $\mu = \lambda \cdot T$ реальное количество мутаций n, прошедшее за один и тот же интервал времени, может быть разным и подчиняется распределению Пуассона (1.2.3). И, с другой стороны, как мы уже знаем, одно и то же количество повторов m в маркере может быть при разном количестве мутаций n (естественно, при ограничении, что количество повторов m не может быть меньше количества мутаций n). Следовательно, необходимо усреднение по n для одного и того же значения m. Тогда выражение (3.1.3) можно представить в следующем виде:

$$\max_{T}\{\prod_{j=1}^{M} \sum_{n=0}^{\infty} P(m_j / n; p_1(j), p_2(j)) \cdot P_{П}(n / \lambda_j; T)\} \quad , \qquad (3.1.4)$$

где, исходя из факта пуассоновости потока мутаций,

$$P_{П}(n / \lambda_j; T) = \frac{(\lambda_j \cdot T)^n}{n!} e^{-\lambda_j \cdot T} \quad . \qquad (3.1.5)$$

а вероятности $P(m_j / n; p_1(j), p_2(j)) \quad (j = 1, \ldots, M)$ вычисляются в соответствии с рекуррентными соотношениями (1.3.6).

Заметим, что в (3.1.4) можно не пользоваться отдельно вычисленными константами λ_j и варьируемым параметром T, а использовать единый варьируемый параметр μ_j (среднее количество мутаций до первопредка в

j -м маркере, т.е., с размерностью [*мутации/маркер*]), значения которого для всех маркеров гаплотипа первопредка мы уже определили (как видим, параметры λ_j и T входят в (3.1.5) только в виде произведения). Тем самым, можно не учитывать результаты оценки интенсивности мутаций λ, которые могут меняться в процессе уточнения статистических данных.

Для удобства вычислений введем набор коэффициентов $\{\zeta_1, \zeta_2, \dots, \zeta_M\}$, которые задают относительные величины параметров μ. Например, для наибольшего значения μ_{max} коэффициент ζ примем равным единице, а для остальных маркеров – в соответствующее количество раз меньше:

$$\zeta_j = \frac{\mu_{max}}{\mu_j} \quad , \tag{3.1.6}$$

а затем варьировать будем только обобщенный параметр μ_{max} (соответствующий максимально «быстрому» маркеру с наибольшим значением μ), а для остальных маркеров текущее в процессе подбора значение обобщенного параметра будем вычислять по формуле

$$\mu_j = \frac{\mu_{max}}{\zeta_j} \quad . \tag{3.1.7}$$

Таким образом, выражение (3.1.4) приводится к следующему виду:

$$\max_{\mu_{max}} \{ \prod_{j=1}^{M} \sum_{n=0}^{\infty} P(m_j / n; p_1(j), p_2(j)) \frac{(\mu_j)^n}{n!} \mathrm{e}^{-\mu_j} \} \quad , \tag{3.1.8}$$

Удобнее искать максимум натурального логарифма, и в этом случае произведение заменим суммой (так же, как и в параграфе 2.1):

$$\max_{\mu_{max}} \{ \sum_{j=1}^{M} \sum_{n=0}^{\infty} \ln(P(m_j / n; p_1(j), p_2(j)) \frac{(\mu_j)^n}{n!} \mathrm{e}^{-\mu_j}) \} \quad , \tag{3.1.9}$$

Вычисляя (3.1.9) для всех возможных значений μ_{max}, определим то значение параметра $\hat{\mu_{max}}$, при котором максимум (3.1.9) достигается. Далее,

используя (3.1.6), с помощью вычисленных коэффициентов $\{\zeta_1, \zeta_2, ..., \zeta_M\}$, найдем искомые оценки количества мутаций $\hat{\mu}_j$ для каждого маркера

$$\hat{\mu}_j = \frac{\hat{\mu}_{max}}{\zeta_j} \quad , \tag{3.1.10}$$

* * * * * * * * *

В заключение параграфа следует подчеркнуть пару важных обстоятельств.

1. Конкретные наборы вероятностей $\{p_1(1), p_1(2), ..., p_1(M)\}$ и $\{p_2(1), p_2(2), ..., p_2(M)\}$, приводящие к «перекосам» гистограмм количества повторов в маркерах, отражают особенности процентного содержания потомков от предков, гаплотип каждого из которых имел свои особенности. Если нет уверенности, что исследуемый гаплотип относится к какой-нибудь генеалогической линии исследованной подгруппы из «большой» гаплогруппы (как правило, именно такая ситуация имеет место), то следует все вероятности $\{p_1(1), p_1(2), ..., p_1(M)\}$ сделать равными 0.5, т.е., принять гипотезу о высокой степени симметричности процесса мутаций в гаплотипах. Таким образом, точностные характеристики реально будут несколько хуже заявленных далее в параграфе 3.3.

2. Как следует из всего вышеизложенного, в качестве характеристики используемых маркеров достаточно лишь набора коэффициентов $\{\zeta_1, \zeta_2, ..., \zeta_M\}$, определяющих соотношения интенсивностей мутаций в маркерах, а не абсолютные значения интенсивностей. Таким образом, в качестве исходных данных остаются лишь:

- тестируемый гаплотип $\{m^{(1)}(1), m^{(1)}(2), ..., m^{(1)}(M)\}$;
- гаплотип первопредка $\{m^{(0)}(1), m^{(0)}(2), ..., m^{(0)}(M)\}$;
- коэффициенты, определяющие соотношения интенсивностей мутаций в маркерах (относительные интенсивности) $\{\zeta_1, \zeta_2, ..., \zeta_M\}$.

3.2. Определение точности МП-оценки среднего количества мутаций, прошедших от первопредка до исследуемого одиночного гаплотипа

Методика определения точности максимально правдоподобной оценки (МП-оценки) среднего количества мутаций, прошедших от первопредка до исследуемого одиночного гаплотипа, по своей структуре аналогична методике для оценки качества «группового» алгоритма, которая была изложена в параграфе 2.2.

Описание алгоритма

ШАГ 1. Задаем исходные значения параметров модели M -маркерных гаплотипов:

1.1. Среднее количество мутаций μ для каждого маркера, полученное для первопредка $\{\mu_0(1), \mu_0(2), \ldots, \mu_0(M)\}$, где M - количество маркеров.
1.2. Вероятности p_1 для каждого маркера $\{p_1(1), p_1(2), \ldots, p_1(M)\}$.
1.3. Вероятности p_2 для каждого маркера равны нулю.
1.4. Конкретное значение параметра m_0 значения не имеет, но для определенности зададим его равным 20 для всех маркеров.
1.5. Размер N моделируемой совокупности M -маркерных гаплотипов.

ШАГ 2. Используя рекуррентные соотношения (1.3.6), формируем трехмерную таблицу вероятностей (предельное значение n берем с запасом, например, 20) для всех маркеров (и для всех m от $(m-n)$ до $(m+n)$:

$$\begin{aligned}
P_j(m+2/n+1) &:= P_j(m+2/n+1) + P_j(m/n)\cdot p_2 \\
P_j(m+1/n+1) &:= P_j(m+1/n+1) + P_j(m/n)\cdot p_1 \\
P_j(m/n+1) &:= P_j(m/n+1) + 0 \\
P_j(m-1/n+1) &:= P_j(m-1/n+1) + P_j(m/n)\cdot q_1 \\
P_j(m-2/n+1) &:= P_j(m-2/n+1) + P_j(m/n)\cdot q_2
\end{aligned} \qquad (3.2.1)$$

Здесь j - текущий номер маркера $(j=1,\dots,M)$.

Из таблицы (3.2.1) формируем новую трехмерную таблицу, содержащую интегральные законы распределения для всех n от 1 до 20 и для всех маркеров:

$$F_j(m/n)=\sum_{r=1}^{m} P_j(r/n) \quad , \quad (n=0,1,\dots; \quad j=1,\dots,M; \quad m=0,1..) \qquad (3.2.2)$$

ШАГ 3. Организуем цикл по моделируемым гаплотипам (по i от 1 до N) для получения массива значений оценок параметра μ :

$$\begin{array}{c} \hat{\mu}_1(1)\ \ \hat{\mu}_2(1)\ \ \dots\ \ \hat{\mu}_M(1) \\ \hat{\mu}_1(2)\ \ \hat{\mu}_2(2)\ \ \dots\ \ \hat{\mu}_M(2) \\ \dots\dots\dots\dots \\ \hat{\mu}_1(N)\ \ \hat{\mu}_2(N)\ \ \dots\ \ \hat{\mu}_M(N) \end{array}$$

3.1. Организуем цикл по маркерам по j от 1 до M

3.1.1. С помощью датчика случайных чисел rand(), равномерно распределенных на интервале (0,1), получаем случайное число n , распределенное по Пуассону с параметром μ (аналогично шагу 3.1.1 в п.2.2).

3.1.2. С помощью датчика случайных чисел, равномерно распределенных на интервале (0,1), формируем случайное число x .

3.1.3. Проверяем, в какой из интервалов попало x : $\left[0, F_n(1)\right]$, $\left[F_n(1), F_n(2)\right]$, ... , $\left[F_n(n),1\right]$ (здесь и далее в п.3 индекс «j» для простоты опускаем).

Если $\left[F_n(i-1)<x\le F_n(i)\right]$, то в качестве количества повторов в j -м маркере принимаем дискретную величину i $(m(j)=i)$.

3.1.4. Конец цикла по маркерам. Очередной i -й гаплотип $\{m_1(i), m_2(i), \dots, m_M(i)\}$ сформирован.

3.2. Методом подбора вычисляем по формуле (3.1.9) для текущего i -го гаплотипа $\{m_1(i), m_2(i), \dots, m_M(i)\}$ свои максимально правдоподобные оценки $\{\hat{\mu}_1(i), \hat{\mu}_2(i), \dots, \hat{\mu}_M(i)\}$ (среднее количество мутаций до первопредка по каждому маркеру), используя алгоритм, описанный в параграфе 3.1.

3.3. Конец цикла по i (по массиву моделируемых гаплотипов).

ШАГ 4. По полученному массиву МП-оценок среднего количества мутаций

$$\begin{array}{c} \hat{\mu}_1(1)\ \ \hat{\mu}_2(1)\ \ .\ .\ .\ \ \hat{\mu}_M(1) \\ \hat{\mu}_1(2)\ \ \hat{\mu}_2(2)\ \ .\ .\ .\ \ \hat{\mu}_M(2) \\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ . \\ \hat{\mu}_1(N)\ \ \hat{\mu}_2(N)\ \ .\ .\ .\ \ \hat{\mu}_M(N) \end{array} \qquad (3.2.3)$$

и исходному вектору моделируемых значений $\{\mu_0(1), \mu_0(2), \ldots, \mu_0(M)\}$ первопредка вычисляем:

4.1. среднее значение МП-оценки среднего количества мутаций по каждому маркеру (здесь и далее помним, что величины $\hat{\mu}_j$, D_μ , $\tilde{\sigma}_\mu$ являются функциями длины выборки N)

$$\hat{\mu}_j = \frac{1}{N}\sum_{k=1}^{N} \hat{\mu}_j(k) \qquad (j = 1, \ldots, M) \qquad [\textit{мутации/маркер}] \qquad (3.2.4)$$

4.2. суммарное значение по всем маркерам МП-оценок среднего количества мутаций

$$\hat{\mu}_\Sigma = \sum_{j=1}^{M} \hat{\mu}_j \qquad [\textit{мутации/гаплотип}] \qquad (3.2.5)$$

4.3. дисперсию и среднеквадратическое МП-оценки среднего количества мутаций для каждого маркера

$$D_\mu(j) = \frac{1}{N-1}\sum_{k=1}^{N} (\hat{\mu}_j(k) - \mu_0(j))^2 \qquad (j = 1, \ldots, M) \qquad (3.2.6a)$$

$$\tilde{\sigma}_\mu(j) = \frac{\sqrt{D_\mu(j)}}{\mu_0} \qquad (j = 1, \ldots, M) \qquad (3.2.6b)$$

4.4. суммарное значение по всем маркерам дисперсии и среднеквадратического отклонения МП-оценок среднего количества мутаций

$$D_\mu = \frac{1}{M}\sum_{j=1}^{M} D_\mu(j) \qquad (3.2.7a)$$

$$\tilde{\sigma}_\mu = \sqrt{D_\mu} \qquad (3.2.7b)$$

4.5. абсолютное и относительное смещения МП-оценки относительно «истинного» значения μ_0 для каждого маркера:

$$S_\mu(j)=\hat{\mu}_j-\mu_0(j) \quad . \qquad (j=1,\dots,M) \tag{3.2.8a}$$

$$\tilde{S}_\mu(j)=\frac{\hat{\mu}_j-\mu_0(j)}{\mu_0(j)}=\frac{S_\mu(j)}{\mu_0(j)} \qquad (j=1,\dots,M) \tag{3.2.8b}$$

4.6. границы доверительных интервалов вычисляем упрощенно по уровню $\pm\sqrt{D_\mu}$ и по уровню $\pm 2\sqrt{D_\mu}$ относительно среднего значения $\hat{\mu_\Sigma}$, предполагая «нормальность» распределения оценки величины μ. При этом уровни значимости α считаются равными 15,85% и 2,25% соответственно.

4.7. суммарное среднее количество мутаций для первопредка (сумма исходных значений, использованных для моделирования):

$$\mu_0=\sum_{j=1}^{M}\mu_0(j) \tag{3.2.9}$$

Данная величина μ_0 необходима для сравнения суммарного моделируемого количества мутаций с величиной $\hat{\mu_\Sigma}$, определенной методом максимального правдоподобия.

ШАГ 5. Конец алгоритма.

3.3. Результаты тестирования на модельных данных

Интерес представляет качество получаемой оценки времени, прошедшего от первопредка до индивида, которому принадлежал исследуемый гаплотип. Под качеством, как и ранее будем понимать набор статистических параметров оценки, которые получим посредством статистического моделирования.

В соответствии с алгоритмом моделирования, описанным в в параграфе 3.2 (шаги 1,2, и 3.1 алгоритма) формируем несколько тестовых наборов 12-ти маркерных гаплотипов по 5000 штук каждый. Каждый набор представляется в

соответствии в виде матрицы, описанной во Введении данной работы (выражение i.1):

$$\begin{matrix} m_1(1) & m_1(2) & \ldots & m_1(M) \\ m_2(1) & m_2(2) & \ldots & m_2(M) \\ \ldots & \ldots & \ldots & \ldots \\ m_N(1) & m_N(2) & \ldots & m_N(M) \end{matrix} \qquad (3.3.1)$$

в которой

N – число гаплотипов в выборке;

M – число маркеров, по которым данные гаплотипы построены (в нашем случае – 12);

$m_i(j)$ – целое неотрицательное число, соответствующее количеству повторов в i -м маркере (столбце матрицы) j -го гаплотипа (строки матрицы).

Для простоты при моделировании начальное значение количества повторов m_0 у всех маркеров задавалось равным 20 (конкретное значение не принципиально, допускаем даже ситуацию, когда в результате мутаций количество повторов в маркере станет отрицательным). Как уже ранее было показано для случайного процесса типа блуждания важна только разность между начальным и промежуточным значениями.

На первом этапе параметры исследуемой гаплогруппы задавались в виде тестовых значений, удобных для анализа качества, а именно:

- среднее количество мутаций μ в маркере, прошедшее от первопредка (фактически определяющее «скорость» изменения маркера), для всех маркеров задавалось различным – от 0,05 до 10,0;
- вероятности p_1 во всех маркерах задавались одинаковыми (вероятности p_2 для простоты обнулялись) и равными 0.25 или 0.5, или 0.75.

На втором этапе в качестве исходных параметров были взяты параметры гаплогруппы R1b (а именно, вероятности p_1 и p_2 для всех маркеров, а также количество мутаций μ для всех маркеров), которые были вычислены как максимально правдоподобные оценки при обработке 7780 реальных 12-ти маркерных гаплотипов класса R1b (см. главу 5).

Итак, в качестве сравниваемой пары гаплотипов берутся «предковый» гаплотип, состоящий из 12-ти целых чисел, каждое из которых равно 20 (как мы знаем, начальное значение количества повторов в маркерах роли не играет), а в качестве изучаемого гаплотипа – полученный моделированием (по алгоритму параграфа 2.2), и у которого значения повторов в маркере, естественно, уже «разбросаны» относительно среднего значения 20 в соответствии с параметрами p_1 , p_2 , μ ? вычисленными по реальной гаплогруппе R1b. И таких анализируемых пар – 5000 для каждого варианта исследований.

После проведения 5000 «анализов» вычисляются статистические характеристики ошибок определения среднего количества мутаций (по каждому маркеру), прошедшего от исследуемого (смоделированного) гаплотипа до гаплотипа первопредка (состоит из 12-ти чисел значением 20). В качестве оценочных параметров рассматриваются матожидание вычисленного среднего количества мутаций по каждому маркеру и суммарное по всем маркерам., Интерес представляет и дисперсия разброса всех 5000 оценок относительно среднего значения. Кроме этого вычислены доверительные интервалы для четырех значений доверительной вероятности 0.01, 0.05, 0.1 и 0.2.

3.3.1. Вероятность p_1 =0.5

3.3.1.1. Среднее количество мутаций

Среднее количество мутаций до первопредка в маркерах $\{\mu_0(1), \mu_0(2), \ldots, \mu_0(M)\}$, которые задавались в качестве исходных (верхняя строка чисел), и среднее $\hat{\mu}_j$ (для каждого из 12-ти маркеров) наиболее правдоподобного количества мутаций от первопредка до проверяемого гаплотипа (формула 3.2.4) (нижняя строка чисел):

№:	1	2	3	4	5	6	7	8	9	10	11	12
	0.0500	0.1000	0.2500	0.5000	0.7500	1.0000	1.2500	1.5000	1.7500	2.0000	2.5000	5.0000
	0.0524	0.1048	0.2620	0.5240	0.7860	1.0480	1.3100	1.5720	1.8340	2.0960	2.6200	5.2400

3.3.1.2. Смещение оценки среднего количества мутаций

Смещение оценки среднего количества мутаций $S_\mu(j)$ (формула 3.2.8a) для каждого из 12тти маркеров:

№:	1	2	3	4	5	6	7	8	9	10	11	12
	0.0024	0.0048	0.0120	0.0240	0.0360	0.0480	0.0600	0.0720	0.0840	0.0960	0.1200	0.2400

Относительное смещение оценки среднего количества мутаций $\tilde{S}_\mu(j)$ (формула 3.2.8b) для всех маркеров примерно одинаковое и равно величине 0.048 (4,8% от истинного).

3.3.1.3. Суммарное среднее количество мутаций

Суммарное среднее количество мутаций для первопредка (формула (3.2.9))
μ_0 =16.65 [*мутации/гаплотип*]

Суммарное среднее количество мутаций по проверенным гаплотипам (формула (3.2.5))
$\hat{\mu_\Sigma}$ =17.45 [*мутации/гаплотип*]

3.3.1.4. Среднеквадратическое отклонение количества мутаций

Относительное среднеквадратическое отклонение $\tilde{\sigma}_\mu(j)$ наиболее правдоподобного количества мутаций от первопредка до проверяемого гаплотипа (формула (3.2.6b)) для каждого из 12-ти маркеров:

№:	1	2	3	4	5	6	7	8	9	10	11	12
	0.0259	0.0519	0.1297	0.2593	0.3890	0.5187	0.6483	0.7780	0.9077	1.0373	1.2967	2.5934

Суммарное среднеквадратическое отклонение количества мутаций по проверенным гаплотипам (формула (3.2.7b)) $\tilde{\sigma}_\mu$ =3.44

3.3.1.5. Доверительные интервалы

Ниже приведены размеры доверительных интервалов искомого параметра μ (вычислялись по гистограмме) в процентах от среднего

значения оценки «вправо» и «влево» (с минусом, поскольку «левая» граница по величине меньше среднего значения $\hat{\mu}_{\Sigma}$ формула (3.2.5)):

Таблица 3-1. Доверительные интервалы для оценки количества мутаций, отделяющих исследуемый гаплотип от гаплотипа первопредка

Доверительная вероятность	α =0.01	δ_{min} = -85%	δ_{max} = +152%
Доверительная вероятность	α =0.05	δ_{min} = -69%	δ_{max} = +99%
Доверительная вероятность	α =0.10	δ_{min} = -54%	δ_{max} = +68%
Доверительная вероятность	α =0.20	δ_{min} = -46%	δ_{max} = +38%

Для случаев наличия «перекоса» гистограмм из-за неравенства вероятности p_1 величине 0.5, смещение оценки пропадает, причем и размеры доверительных интервалов (и дисперсия) уменьшаются примерно на треть, т.е., качество оценки становится значительно лучше. Например:

3.3.2. Для вероятности p_1 =0.25:

3.3.2.1. Среднее количество мутаций

Среднее количество мутаций до первопредка в маркерах $\{\mu_0(1), \mu_0(2), \ldots, \mu_0(M)\}$, которые задавались в качестве исходных (верхняя строка чисел), и среднее $\hat{\mu}_j$ наиболее правдоподобного количества мутаций от первопредка до проверяемого гаплотипа в маркерах (формула 3.2.4) (нижняя строка чисел) для каждого из 12-ти маркеров:

№:	1	2	3	4	5	6	7	8	9	10	11	12
	0.0500	0.1000	0.2500	0.5000	0.7500	1.0000	1.2500	1.5000	1.7500	2.0000	2.5000	5.0000
	0.0501	0.1002	0.2506	0.5012	0.7517	1.0023	1.2529	1.5035	1.7540	2.0046	2.5057	5.0115

3.3.2.2. Суммарное среднее количество мутаций

Суммарное среднее количество мутаций для первопредка (формула (3.2.9))
μ_0 =16.65 [*мутации/гаплотип*]
Суммарное среднее количество мутаций по проверенным гаплотипам (формула (3.2.5)) $\hat{\mu}_{\Sigma}$ =16.69 [*мутации/гаплотип*]

3.3.3. Для вероятности p_1 =0.75:

3.3.3.1. Среднее количество мутаций

Среднее количество мутаций до первопредка $\{\mu_0(1), \mu_0(2), \ldots, \mu_0(M)\}$, которые задавались в качестве исходных (верхняя строка чисел), и среднее $\hat{\mu}_j$ наиболее правдоподобного количества мутаций от первопредка до проверяемого гаплотипа (формула 3.2.4) (нижняя строка чисел) для каждого из 12-ти маркеров:

№:	1	2	3	4	5	6	7	8	9	10	11	12
	0.0500	0.1000	0.2500	0.5000	0.7500	1.0000	1.2500	1.5000	1.7500	2.0000	2.5000	5.0000
	0.0504	0.1008	0.2519	0.5038	0.7557	1.0076	1.2595	1.5114	1.7633	2.0152	2.5190	5.0379

3.3.3.2. Суммарное среднее количество мутаций

Суммарное среднее количество мутаций для первопредка (формула (3.2.9))
μ_0 =16.65 [*мутации/гаплотип*]
Суммарное среднее количество мутаций по проверенным гаплотипам (формула (3.2.5))
$\hat{\mu}_\Sigma$ =16.78 [*мутации/гаплотип*]

3.4. Сравнение качества «группового» и «одиночного» алгоритмов

Интерес представляет сравнение оценки качества определения параметров гаплогруппы, полученных с помощью «группового» и «одиночного» алгоритмов. Для этого была с помощью имитатора гаплотипов смоделирована совокупность 1000 двенадцатимаркерных гаплотипов с параметрами, соответствующими гаплогруппе R1b.

3.4.1. Количество мутаций, прошедших от первопредка до проверяемого гаплотипа (отдельно по маркерам)

- Первая строка – исходные значения μ_0 12-ти маркеров для моделирования 1000 гаплотипов
- Вторая строка – результаты оценки параметров μ_0 для 12-ти маркеров «групповым» алгоритмом
- Третья строка – результаты оценки параметров μ_0 для 12-ти маркеров «одиночным» алгоритмом.

№:	1	2	3	4	5	6	7	8	9	10	11	12
	0.1100	0.5350	0.1225	0.3775	0.1700	0.5125	0.0275	0.0300	0.5125	0.2500	0.1400	0.5475
	0.0950	0.5725	0.1250	0.3925	0.1825	0.5575	0.0300	0.0300	0.4725	0.2500	0.1225	0.5500
	0.1199	0.5830	0.1335	0.4114	0.1853	0.5585	0.0300	0.0327	0.5585	0.2724	0.1526	0.5967

3.4.2. Суммарное среднее количество мутаций для первопредка как сумма всех маркеров

- рассчитано по исходным параметрам для моделирования: $\mu_0 = 3{,}335$
- рассчитано по результатам работы «группового» алгоритма: $\hat{\mu_\Sigma} = 3{,}38$
- рассчитано по результатам работы «одиночного» алгоритма: $\hat{\mu_\Sigma} = 3{,}63$

Несколько худшие значения, полученные «одиночным» алгоритмом, объясняются, во-первых, тем фактом, что в процессе подбора мы вынуждены были синхронно менять параметр μ в каждом из маркеров в соответствии со значениями нормировочных коэффициентов $\{\zeta_1, \zeta_2, ..., \zeta_M\}$ (см. формулу (3.1.6)), а в групповом алгоритме по каждому маркеру осуществляется свой независимый подбор. То есть, у нас пропала одна «степень свободы» в подборе параметров.

Во-вторых, размер статистики, используемой «одиночным» гаплотипом — количество чисел, соответствующее количеству маркеров в гаплотипе, тогда как для «группового» алгоритма может использоваться несколько тысяч гаплотипов.

3.5. Выводы

1. Разработан и программно реализован статистический алгоритм вычисления максимально правдоподобной оценки реально осуществившихся мутаций (с учетом возвратных мутаций) в маркерах единственного потомка за время от первопредка.

2. С помощью статистического моделирования были изучены свойства вычисляемых максимально правдоподобных оценок количества мутаций численно сравнены «одиночноый» и «групповой» алгоритмы.

ГЛАВА 4. Некоторые вопросы точности оценок интенсивности потока мутаций

Как следует из описания основных свойств мутаций и гаплотипов, обсужденных во Введении, поток мутаций соответствует пуассоновскому потоку случайных событий. Это обстоятельство неизбежно приводит к тому, что интервалы времени между моментами любых двух последовательных мутаций имеет случайную величину. А интервал времени, образованный несколькими подряд мутациями, представляет собой уже сумму интервалов случайной длины и, очевидно, также имеет случайную длину. Заметим, что и количество мутаций тоже случайное, поскольку вычисляется статистическими методами (см. главы 2,3). Последующий пересчет количества мутаций в физическое время осуществляется с помощью калибровочных значений скоростей (интенсивностей) мутаций, которые, как мы увидим далее, также оцениваются методами статистики и имеют определенный разброс.

Учитывая вышеизложенное, вопрос о точности окончательно вычисляемых оценок времени до первопредка имеет первостепенное значение.

4.1. Точность оценки реальной интенсивности мутаций

Поскольку получить истинные времена, когда происходили мутации, невозможно, то, воспользовавшись предположением об эргодичности процесса возникновения мутаций, можно заменить усреднение по времени усреднением по реализациям. Также будем считать, что темп мутаций не зависит от гаплогруппы, условий жизни, состояния здоровья и пр, а определяется только самой структурой нуклеотидных цепочек в маркерах Y-хромосомы.

В качестве «реализаций» были использованы пары гаплотипов отца и сына (обычно используются в задаче установления отцовства). Оказалось, что из $N=3026$ пар «отец-сын» мутации встретились у $k=54$ сыновей. Следовательно, максимально правдоподобная оценка вероятности одной мутации на гаплотип равна отношению $\hat{\theta}=54/3026=0{,}01784$. Если за базовый интервал времени взять длительность одного поколения, и учитывая,

что мутации случаются с дискретностью во времени, также равной одному поколению, то полученная оценка вероятности $\hat{\theta}$ численно равна интенсивности потока мутаций, размерность которой, соответственно, [*кол.мутаций/(гаплотип×поколение)*]. Для гаплотипов с разным количеством маркеров (от 6-ти маркерных до 67-маркерных) эти интенсивности (вернее, точечные оценки этих интенсивностей) вычислены А.А.Клёсовым и в дальнейших расчетах принимаются в качестве калибровочных параметров.

Поскольку $\hat{\theta}$, полученная как отношение величины k к N, является точечной оценкой интенсивности мутаций, то встает вопрос о ее точности.

Очевидно, вероятность получения k мутироваших гаплотипов из N наблюденных гаплотипов имеет биномиальное распределение

$$P_{bin}=C(k,N)\theta^k(1-\theta)^{N-k} \quad , \qquad (4.1.1)$$

где $C(k,N)=\frac{N!}{k!(N-k)!}$ - число сочетаний по k из N.

Нас интересует доверительный интервал, в котором лежит истинное значение параметра θ (вероятность «положительного» исхода, т.е., появления мутации в проверяемой паре «отец-сын»). Зададим вероятность α_θ «выпадения» оценочного значения за доверительный интервал «вверх» или «вниз». В этом случае вероятность нахождения оценки в доверительном интервале $P=1-2\alpha_\theta$.

Известно [2,3], что верхняя и нижняя границы доверительного интервала для параметра θ распределения (4.1.1) вычисляются как корни квадратного уравнения следующим образом:

$$\theta_{min}=\frac{k+0{,}5\,g^2+g\sqrt{k\frac{(N-k)}{N}+0{,}25\,g^2}}{N+g^2}$$
$$\theta_{max}=\frac{k+0{,}5\,g^2-g\sqrt{k\frac{(N-k)}{N}+0{,}25\,g^2}}{N+g^2} \qquad (4.1.2)$$

где g - квантиль нормального распределения с нулевым матожиданием и единичной дисперсией, соответствующий уровню значимости α_θ :

$$g = \Phi^{-1}(\alpha_\theta) \quad , \qquad (4.1.3)$$

где, в свою очередь, $\Phi^{-1}(\alpha_\theta)$ обозначает обратную функцию к интегралу функции ошибки (нормального распределения).

В таблице 4-1 представлены доверительные интервалы для количества мутаций $k = 54$, полученного при обследовании $N = 3026$ гаплотипов. Максимальные и минимальные значения θ_{min} и θ_{max} как раз и задают допустимые границы оценки параметра θ в зависимости от требуемой точности, определяемой вероятностью α_θ . Границы интервалов дополнительно даны в процентах от истинного значения (per_{min} , per_{max} - ширина доверительного интервала вверх и вниз от максимально правдоподобного значения $\hat{\theta} = k/N$, выраженная в процентах от $\hat{\theta}$).

Таблица 4-1

Колич. мутаций $k = 54$
Общее колич. пар гаплотипов $N = 3026$
Частота мутаций $\hat{\theta} = k/N = 0{,}0178$

α_θ	θ_{min}	θ_{max}	per_{min}	per_{max}	$P = 1 - 2\alpha_\theta$
0.01	0.0130	0.0244	-26.89%	36.54%	0.98
0.05	0.0143	0.0223	-19.89%	24.72%	0.90
0.10	0.0150	0.0212	-15.88%	18.81%	0.80
0.15	0.0155	0.0205	-13.05%	14.97%	0.70
0.20	0.0159	0.0200	-10.73%	12.00%	0.60
0.25	0.0163	0.0195	-8.70%	9.51%	0.50
0.30	0.0166	0.0192	-6.83%	7.32%	0.40

Теперь представим себе, что проведено обследование вдвое большего количества гаплотипов и при этом обнаружено также ровно вдвое больше мутаций. Результаты вычисления границ доверительного интервала представлены в таблице 4-2:

Таблица 4-2

Колич. мутаций $k = 108$
Общее колич. пар гаплотипов $N = 6052$
Оценка частоты мутаций $\hat{\theta} = k/N = 0{,}0178$

α_θ	θ_{min}	θ_{max}	per_{min}	per_{max}	$P = 1 - 2\alpha_\theta$
0.01	0.0143	0.0223	-19.89%	24.72%	0.98
0.05	0.0153	0.0209	-14.52%	16.94%	0.90
0.10	0.0158	0.0202	-11.51%	12.97%	0.80
0.15	0.0162	0.0197	-9.41%	10.37%	0.70
0.20	0.0165	0.0193	-7.72%	8.35%	0.60
0.25	0.0167	0.0190	-6.23%	6.64%	0.50
0.30	0.0170	0.0188	-4.88%	5.12%	0.40

Очевидно, что оценка частоты мутаций $\hat{\theta}$ осталась той же самой, что и для таблицы 4-2, а доверительный интервал для θ стал примерно на треть уже, чем в таблице 4-1, т.е., точность оценки повысилась на треть.

Как следует из таблиц 4-1 и 4-2, ослабление требований к точности (увеличение допустимой вероятности ошибки α_θ, т.е., «выпадения» оценки за интервал) уменьшает доверительный интервал.

Калибровочные значения оценки $\hat{\theta}$, которые используются в данной работе в расчетах точности, получены А.А Клёсовым для гаплотипов различного порядка M и представлены в следующей таблице:

Таблица 4-3. Калибровочные значения интенсивности мутаций $\hat{\theta}$ для гаплотипов с разным количеством маркеров M (по А.А.Клёсову)

Количество маркеров в гаплотипе	Интенсивность мутаций $\hat{\theta}$
6	0.0096
12	0.0240
25	0.0460
37	0.0900
67	0.1450

4.2. Повышение точности при сложении данных по отдельным маркерам

Интенсивность потока мутаций $\hat{\theta}=0{,}01784$, рассмотренная в предыдущем параграфе, получена как суммарная по всем маркерам гаплотипов, привлеченных для этого расчета. Но в каждом из M маркеров имеет место собственный поток мутаций со своей интенсивностью λ [*мутации/(маркер×поколение)*]. Поскольку возникновение мутации на каком-либо маркере не зависит от того, когда происходила мутация на любом другом маркере, то в этом случае на гаплотипе имеется M независимых друг от друга пуассоновских потоков мутаций с интенсивностями $\lambda_1, \ldots, \lambda_M$ соответственно. Следовательно, можно говорить о едином (суммарном) пуассоновском потоке мутаций в гаплотипе, интенсивность которого, очевидно,

$$\theta=\sum_{j=1}^{M}\lambda_j \quad [\textit{мутации/(гаплотип×поколение)}] \qquad (4.2.1)$$

В этом случае распределение вероятностей появления n мутаций у суммарного потока на интервале T принимает следующий вид:

$$P_{П}(n/\theta;T)=\frac{(\theta\cdot T)^n}{n!}\mathrm{e}^{-\theta\cdot T} \qquad (4.2.2)$$

Принципиальное различие интенсивностей θ и λ в том, что θ имеет размерность [*мутации/(гаплотип×поколение)*], а λ имеет размерность [*мутации/(маркер×поколение)*].

Оценку величины θ можно получать как по гаплотипу в целом (см. параграф 4.1), так и для каждого маркера отдельно аналогичным способом (это будут $\lambda_1, \ldots, \lambda_M$), а затем, суммируя в соответствии с (4.2.1), получить величину θ.

Пусть теперь в процессе статистической обработки получены средние количества мутаций $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M$ по всем M маркерам (напомним, что у величин $\hat{\mu}_j$ размерность — [*мутации/маркер*]). Тогда, исходя из параметров суммарного пуассоновского потока мутаций на гаплотипе,

точечная оценка интервала времени (количества прошедших поколений) до первопредка может быть получена следующим образом:

$$\hat{T} = \frac{\sum_{j=1}^{M} \hat{\mu}_j}{\sum_{j=1}^{M} \lambda_j} = \frac{\hat{\mu}_\Sigma}{\theta} \qquad [\textit{поколение}]. \tag{4.2.3}$$

где

$\hat{\mu}_\Sigma = \sum_{j=1}^{M} \hat{\mu}_j$ - суммарное количество мутаций, вычисленное по всем маркерам гаплотипа [*мутации/гаплотип*].

Определим точность оценки искомого интервала времени T. Поскольку конкретные значения $\hat{\mu}_\Sigma$ и θ в (4.2.3), получаемые в результате расчетов, случайные и имеют определенный разброс, то и результирующая оценка времени $\hat{T}$, как следует из (4.2.3), также величина случайная.

Чтобы построить доверительные интервалы для оценки $\hat{T}$, используем факт, что время между мутациями распределено экспоненциально (поток мутаций пуассоновский). Таким образом, суммарное время, за которое произошли несколько мутаций, очевидно, имеет распределение Эрланга с параметрами θ и $\hat{\mu}_\Sigma$ (здесь это должно быть целое число мутаций):

$$f(T) = \frac{\theta^{\hat{\mu}_\Sigma} \cdot T^{\hat{\mu}_\Sigma - 1} \cdot e^{-T \cdot \theta}}{(\hat{\mu}_\Sigma - 1)!} \quad . \tag{4.2.4}$$

Обычно оценка количества мутаций $\hat{\mu}_\Sigma$ — нецелое число, то вместо (4.2.4) можно воспользоваться гамма-распределением

$$f(T) = \frac{\theta^{\hat{\mu}_\Sigma} \cdot T^{\hat{\mu}_\Sigma - 1} \cdot e^{-T \cdot \theta}}{\Gamma(\hat{\mu}_\Sigma)} \quad , \tag{4.2.5}$$

частным случаем которого и является распределение Эрланга (4.2.4) при целых значениях параметра $\hat{\mu}_\Sigma$.

Матожидание $M_T = \hat{T}$ (среднее значение интервала T) распределений (4.2.4) и (4.2.5) вычисляется в соответствии с (4.2.3), а дисперсия — следующим образом:

$$D_T = \frac{\hat{\mu}_\Sigma}{\theta^2} \quad . \qquad (4.2.6)$$

В таблицах Приложения 1 представлены результирующие доверительные интервалы для значения различных уровней α_T для оценки количества поколений T, прошедших от первопредка. Поскольку в таблицах приведены расчеты только для целых значений количества мутаций $\hat{\mu}_\Sigma$, то в случае нецелого значения (что в реальности всегда и получается), можно упрощенно применять, например, линейную интерполяцию (см. примеры в Приложении 1) или использовать таблицы гамма-распределения.

4.3. Оценка интенсивности мутаций в маркерах по совокупности малых гаплогрупп

Метод максимально правдоподобного оценивания по реальным измерениям количества повторов в маркере может быть использован и для оценки интенсивности мутаций λ (размерность – [*мутации/ (маркер×поколение)*]).

Пусть мы применяем метод максимального правдоподобия для нахождения оценки $\hat{\mu}$ значения обобщенного параметра μ для какого-либо маркера (размерность – [*мутации/маркер*]), который затем используется для оценки времени t, прошедшем от первопредка до нашего времени:

$$t = \frac{\hat{\mu}}{\lambda} \qquad [\textit{поколения}], \qquad (4.3.1)$$

где λ - интенсивность мутаций на каком-либо маркере [*мутации/маркер×поколение*].

Ну а как быть, если величина λ (аналогично, суммарная интенсивность θ) пока неизвестна, или требуется ее уточнение? Нужно найти несколько первопредков (чем больше, тем лучше), чье время жизни известно достоверно. Затем для каждого первопредка выявить всех ныне живущих потомков по мужской линии.

Таким образом, получим некие аналоги гаплогрупп (далее «гаплогруппы»). Допустим, что всего нашли K таких «гаплогрупп», объемом в n_k гаплотипов каждая. И пусть времена первопредков этих «гаплогрупп» $t_1, t_2, ..., t_K$ известны. Предположим также, что интенсивность мутаций по исследуемому маркеру у всех «гаплогрупп» одинакова.

В результате примененного к каждой из «гаплогрупп» метода максимального правдоподобия (см. главу 2) получим набор значений $\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_K$ обобщенного параметра μ. Очевидно, что оценку интенсивности мутаций можно найти из любого из соотношений:

$$\begin{aligned} \lambda_1 &= \hat{\mu}_1 / t_1 \\ \lambda_2 &= \hat{\mu}_2 / t_2 \\ &\cdots\cdots\cdots \\ \lambda_K &= \hat{\mu}_K / t_K \end{aligned} \qquad (4.3.2)$$

В силу ограниченности статистики из каждого соотношения (4.3.2) мы получим разные оценки интенсивности мутаций λ, разброс значений которых может быть достаточно велик. Чтобы повысить точность, найдем величину λ как средневзвешенное арифметическое среднее:

$$\lambda = \frac{1}{N_\Sigma} \sum_{k=1}^{K} n_k \lambda_k \quad , \text{где} \quad N_\Sigma = \sum_{k=1}^{K} n_k \qquad (4.3.3)$$

Таким образом, в задаче исторической датировки мы по найденному обобщенному параметру μ ищем неизвестное время T, используя найденное заранее значение λ. А во втором случае, напротив, по найденному μ ищем неизвестное значение λ, имея точное знание о времени T.

4.4. Выводы

1. Представлена методика оценки точности расчета интенсивности мутаций, осуществленного по совокупности 3026 пар гаплотипов «отец-сын». Методика основана на общеизвестном и хорошо описанном классическом представлении схемы появления пар с мутацией из общего количества пар гаплотипов «отец-сын» как биномиальный выбор.

2. Разработана методика оценки времени до первопредка, учитывающая пуассоновские свойства потоков мутаций в маркерах и стохастический характер временных интервалов между мутациями, следующими друг за другом. Созданы таблицы, по которым можно определять на практике доверительные интервалы для времени первопредка, зная количество прошедших мутаций μ и интенсивность потока мутаций θ. Таблицы приведены в Приложении 1.

3. Рассмотрена возможность оценки интенсивности мутаций не по парам гаплотипов «отец-сын», а по нескольким группам гаплотипов потомков с заведомо известным первопредком в каждой и со строго зафиксированным временем жизни первопредков.

ГЛАВА 5. Результаты обработки реальных данных

Чтобы проверить разработанную методику, было обработано пять гаплогрупп 12-ти маркерных гаплотипов. Для проведения исследований были отобраны только те гаплогруппы, в которых количество гаплотипов не менее 1000. Достаточный объем статистики позволил выяснить все особенности отличия результатов моделирования от статистических характеристик экспериментальных данных (наборов гаплотипов). В итоге исследованию подверглись вполне представительные выборки гаплотипов из гаплогрупп I1a, J2, R1b, R1b1, R1b1c.

5.1. Анализ качества оценки среднего количества мутаций, прошедших от первопредка до настоящего времени

Параметр μ, имеющий смысл среднего количества мутаций, прошедших от первопредка до настоящего времени в отдельно взятом маркере одного гаплотипа, имеет ключевое значение в задаче исторических датировок. Поэтому особый интерес имеет изменение его значения у разных маркеров и в различных гаплогруппах.

В таблице 5-1 приведены значения МП-оценок $\hat{\mu}$ параметра μ [*среднее количество мутаций на гаплотип в отдельном маркере*].

Таблица 5-1. Значения оценок $\hat{\mu}$ обобщенного параметра μ (среднего количества мутаций в маркерах) для 12-ти маркерных гаплотипов (каждый столбец — конкретный маркер)

Гаплогруппа I1a - 1970 гаплотипов
Гаплогруппа J2 - 1120 гаплотипов
Гаплогруппа R1b - 7770 гаплотипов
Гаплогруппа R1b1 - 4625 гаплотипов
Гаплогруппа R1b1c - 4050 гаплотипов

маркер	1	2	3	4	5	6	7	8	9	10	11	12
(I1a) :	0.1600	0.3875	0.2225	0.1125	0.5025	0.4275	0.0100	0.1325	0.3075	0.1125	0.0350	0.3700
(J2) :	0.1025	0.6300	0.6650	0.2775	0.7650	1.3400	0.0100	0.5375	0.7675	0.6150	0.0650	1.4800
(R1b) :	0.1100	0.5350	0.1225	0.3775	0.1700	0.5125	0.0275	0.0300	0.5125	0.2500	0.1400	0.5475
(R1b1) :	0.1300	0.5775	0.1650	0.4050	0.2075	0.5350	0.0450	0.0350	0.5600	0.2450	0.1775	0.5550
(R1b1c):	0.1175	0.4450	0.1000	0.3850	0.1375	0.3925	0.0275	0.0300	0.5225	0.2150	0.1025	0.5250

Как следует из первых двух строк таблицы 5-1, первопредок гаплогруппы J2 старше I1a, поскольку во всех маркерах кроме первого μ у гаплогруппы J2 больше. Инверсия значения на 1-м маркере, возможно, произошла из-за узости гистограмм и, как следствие, малой точности вычисления μ .

По старшинству первопредков, иными словами, по времени возникновения, исследуемые нами ветви гаплогруппы R образуют следующую цепочку: R1b→R1b1→R1b1c . Следовательно, в любом из 12-ти маркеров должно наблюдаться уменьшение значения μ именно в указанной последовательности. Между подгруппами R1b и R1b1 такая зависимость действительно имеет место, причем, для всех маркеров. А вот значения μ у подгруппы R1b1c не подчиняются этой зависимости. Возможно, особенность молодой подгруппы R1b1c вытекает из слабой разветвленности генеалогического древа по сравнению со старой гаплогруппой J2.

5.2. Анализ поведения вероятности изменения количества повторов в маркерах возникающего в процессе мутаций

В таблице 5-2 приведены сводные результаты по параметрам вероятностного механизма мутаций на маркерах. В таблице использованы следующие обозначения:

p_1 - вероятность того, что в результате мутации количество повторов в маркере увеличится на единицу;

p_2 - вероятность того, что в результате мутации количество повторов в маркере увеличится сразу на два;

m - количество повторов в маркере, которое должно было быть у первопредка данной гаплогруппы.

Таблица 5-2. Вероятности изменения количества повторов в маркерах в результате мутаций и реставрированные значения в маркерах у первопредков групп

Гаплогруппа I1a

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
p1 :	0.7098	0.9600	0.9212	0.9100	0.8000	0.6930	0.4200	0.7553	0.8217	0.8736	0.7735	0.7954
p2 :	0.0702	0.0000	0.0188	0.0000	0.0000	0.0070	0.0000	0.0747	0.0083	0.0864	0.0765	0.0246
m :	13	22	14	10	13	14	11	14	11	12	11	28

Гаплогруппа J2

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
p1 :	0.9500	0.7300	0.9500	0.3900	0.6188	0.2548	0.6006	0.6696	0.8700	0.4700	0.8463	0.9405
p2 :	0.0000	0.0000	0.0000	0.0000	0.0612	0.0252	0.0594	0.0504	0.0000	0.0000	0.0837	0.0095
m :	12	23	14	10	13	17	11	15	11	13	11	28

Гаплогруппа R1b

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
p1 :	0.5488	0.4000	0.8064	0.1600	0.7488	0.5184	0.5568	0.7680	0.4600	0.6596	0.8064	0.6528
p2 :	0.0112	0.0000	0.0336	0.0000	0.0312	0.0216	0.0232	0.0320	0.0000	0.0204	0.0336	0.0272
m :	13	24	14	11	11	14	12	12	12	13	13	29

Гаплогруппа R1b1

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
p1 :	0.5445	0.4500	0.8330	0.1500	0.7098	0.4641	0.5700	0.7098	0.4700	0.7029	0.7695	0.7128
p2 :	0.0055	0.0000	0.0170	0.0000	0.0702	0.0459	0.0300	0.0702	0.0000	0.0071	0.0405	0.0072
m :	13	24	14	11	11	14	12	12	12	13	13	29

Гаплогруппа R1b1c

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
p1 :	0.5049	0.3400	0.7990	0.1400	0.6552	0.5460	0.5529	0.6734	0.4500	0.7000	0.6734	0.6800
p2 :	0.0051	0.0000	0.0510	0.0000	0.0648	0.0540	0.0171	0.0666	0.0000	0.0000	0.0666	0.0000
m :	13	24	14	11	11	14	12	12	12	13	13	29

Вероятности p_1 в подгруппах гаплогруппы R изменяются синхронно от маркера к маркеру, что говорит о приемлемой точности аппроксимации.

Сравнение результатов по подгруппам R1b, R1b1, R1b1c показывает хорошее совпадение вероятностей в одном и том же маркере. Некоторое их различие может быть вызвано, в частности, и некачественной статистикой (наличие ошибок при заполнении баз данных).

Сравнение вероятностей по «далеким» друг от друга гаплогруппам I1a, J2 и R показывает отсутствие корреляции между значениями вероятности p_1 в любом из выбранных маркеров, что, по-видимому, отражает индивидуальные особенности первопредков данных подгрупп. Этот вывод, в частности, подтверждается и тем фактом, что для подгрупп R1b, R1b1, R1b1c на маркере №9 вероятность p_1 лежит в пределах 0,40-0,42, а для подгрупп I1a и J2 эта вероятность на этом же маркере p_1 равна величинам 0,83 и 0,92 соответственно, причем, гистограммы для маркера №9 довольно широкие, и их форма выявляется достаточно подробно.

Отметим также, что только в половине маркеров модель требует наличия вероятности p_2 скачка на 2 в результате мутации. Вместе с тем, значение p_2 может достигать 10% от вероятности p_1 .

Пора прояснить физический смысл сильно отличающегося от 0.5 значения вероятности p_1 . Представим себе, что расплодившиеся потомки какой-нибудь ветви единого генеалогического древа, восходящего к единому корню могли (да и должны) зафиксировать в себе особенности мутаций своего прародителя. Конечно, при большой интенсивности потока мутаций (или при огромном количестве поколений) все эти «перекосы» периодически выравниваются, но появляются новые, которые в начале в силу малочисленности «погоды не делают». Как мы все постоянно видим, именно реальные гистограммы количества повторов имеют такие специфические перекосы формы, которые и выводят на вышеупомянутые соображения.

Действительно, представим себе, что у первопредка имеются два потомка, причем у одного из них произошла мутация на каком-либо маркере. Далее, «немутировавший» поток дал генеалогическое древо втрое меньшее по количеству потомков, чем мутировавший.

Пусть для простоты в обеих группах потомков вообще не было мутаций в выбранном маркере. Может такое быть? Может. Итак, построив гистограмму количества повторов в данном маркере мы увидим следующее: гистограмма будет состоять всего из двух смежных столбиков, величина одного - 0.25 (количество повторов m соответствует первопредку), а величина второго, естественно, втрое больше, т.е., равна 0.75, и с координатой $(m+1)$. Представим себе, что мы не знаем, какой именно потомок относится к той или иной из двух подгрупп. Анализируя общую гистограмму, естественно приходим к выводу, что перекос гистограммы мог возникнуть и из-за неравновесной вероятности p_1 изменения количества повторов «вверх» или «вниз».

Следовательно, вероятность p_1 может быть описана примерно так: это вероятность изменения количества повторов в маркере в результате мутации, средневзвешенная по наблюдаемой выборке из гаплогруппы. Данная вероятность учитывает совокупно как влияние ошибки считывания количества повторов в процессе формирования ДНК потомка, так и неравновесность количества потомков от равноправных (равноудаленных от корня) ветвей генеалогического древа, восходящего к первопредку исследуемой гаплогруппы на изменение количества повторов в маркере.

В результате обработки статистики 60-ти маркеров (по 12 в пяти гаплогруппах) выявилась отчетливая тенденция - вероятность p_1 , как правило, больше 0,5, что означает наличие преобладания в популяциях потомков каких-то «промежуточных» предков, с увеличенным количеством повторов в результате мутации.

5.3. Анализ качества МП-оценок по критерию хи-квадрат

Очевидно, что модель, опирающаяся на ограниченное количество параметров и базовых предположений, всегда дает результат, пусть немного, но отличающийся от реальности.

В нашем случае поток мутаций может быть чуть-чуть непуассоновским, чуть-чуть нестационарным, на вероятности мутаций могут немного влиять индивидуальные особенности людей, чьи гаплотипы были использованы в статистике, и т.п.

Конечно, если требуется создать программную модель датчика случайных чисел с заданным законом распределения, то жестокие требования критерия хи-квадрат вполне обоснованы. Также и в задаче, связанной с обнаружением сигналов на фоне шумов, где часто бывают высокие требования к форме «хвостов» распределений, влияющих на расчетные значения вероятности ложной тревоги.

Но если требуется, например, такая модель, которая бы позволяла вполне адекватно отражать некие интегральные характеристики реального процесса, то похожести формы реальной гистограммы и теоретической аппроксимации может быть вполне достаточно. У нас именно такой случай, поэтому будем в основном обращать внимание на совпадение формы гистограммы количества повторов в каком-либо маркере и гистограммой

распределения, построенного в соответствии с (1.3.6) и (1.2.5) с использованием найденных МП-оценок параметров $\{ p_1 , p_2 , m_0 , \mu \}$.

Критерий хи-квадрат вычисляется следующим образом:

$$z=\sum_{i=1}^{\nu}\frac{(N_i-N\cdot P_i)^2}{N\cdot P_i}=N\cdot\sum_{i=1}^{\nu}\frac{(G_i-P_i)^2}{P_i} \quad , \qquad (5.3.1)$$

где

z - случайная величина, имеющая распределение, асимптотически стремящееся к распределению хи-квадрат с ν степенями свободы;

G_i - экспериментальная гистограмма, которая подвергается проверке $(i=1,...,\nu)$;

P_i - гистограмма теоретического аппроксимирующего распределения $(i=1,...,\nu)$, чьи параметры были найдены в процессе численного подбора МП-оценок параметров $\{ p_1 , p_2 , m_0 , \mu \}$;

N_i - частота событий, вычисляемая как произведение $N_i=N\cdot G_i$ значений экспериментальной гистограммы G_i на количество измерений гаплогруппы (гаплотипов) N , из которых и построена экспериментальная гистограмма G_i .

Как следует из (5.3.1), значение критерия хи-квадрат (величина z) представляет собой произведение количества гаплотипов N в проверяемой выборке на сумму квадратов разности нормированных теоретической и экспериментальной гистограмм. Второй множитель (сумма) представляет собой некий критерий различия формы гистограмм, не зависящий от количества экспериментов (гаплотипов).

Очевидно, что при одинаковом интегральном различии формы проверяемой и экспериментальной гистограмм значение критерия хи-квадрат будет больше для того варианта, в котором использовалось больше исходных данных. Фактически, критерий хи-квадрат требует все более точного совпадения форм гистограмм при неограниченном увеличении количества гаплотипов N .

Введем дополнительный критерий качества, который выполнял бы роль индикатора отличия форм гистограмм. В качестве такого критерия примем сумму модулей разности гистограмм:

$$F=\sum_{i=1}^{\nu}\left|G_i-P_i\right| \quad . \qquad (5.3.2)$$

Затем величину F переведем в проценты и разделим на 2:

$$Form=F\cdot 100/2 \quad . \qquad (5.3.3)$$

Деление на 2 понадобилось для того, чтобы при абсолютном несовпадении формы двух гистограмм (теоретической и экспериментальной) в силу их нормированности значение критерия получится равным 200%, что не очень удобно при анализе, поскольку теперь критерий F меняется в пределах от 0% до 100%.

Допустим, в процессе расчетов мы при некотором значении степеней свободы ν получили значение критерия хи-квадрат, равное z . Чтобы оценить качество аппроксимации, используем довольно удобный эмпирический критерий Романовского, который вычисляется следующим образом:

$$R=\frac{z-\nu}{\sqrt{2\nu}} \quad . \qquad (5.3.4)$$

Если значение критерия $R<3$, то считается, что получено отличное согласие результатов эксперимента с теоретической аппроксимацией.

Очевидно что при $\nu>z$ (очень хорошее совпадение) величина R становится меньше нуля. В этом случае для удобства анализа при вычислениях найденное значение критерия R просто ограничивается нулем.

Результаты оценки качества аппроксимации по всем исследованным гаплогруппам приведены в таблице 5-3. в этой таблице использованы следующие условные обозначения:

xi - значение критерия хи-квадрат (формула (5.3.1));

nu - количество степеней свободы;

Rom - значение критерия Романовского (формула (5.3.4));

Forma - значение критерия отличия формы в процентах (формулы (5.3.2), (5.3.3)).

Таблица 5-3. Результаты оценки качества аппроксимации

Гаплогруппа I1a (N=1970 гаплотипов)

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
xi :	0.2	60.1	2.2	3.8	120.1	0.9	0.0	211.5	0.4	0.1	3.0	2.1
nu :	5	4	4	3	5	5	2	5	5	3	3	5
Rom:	0.00	19.82	0.00	0.32	36.41	0.00	0.00	65.32	0.00	0.00	0.00	0.00
Forma :	0.31	6.11	0.63	0.51	8.96	0.78	0.03	5.15	0.25	0.39	0.31	0.56

Гаплогруппа J2 (N=1120 гаплотипов)

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
xi :	0.3	11.3	32.4	1.7	31.1	11.0	0.3	42.3	42.2	19.1	0.8	33.4
nu :	3	5	4	4	7	7	2	6	5	4	3	6
Rom:	0.00	2.00	10.03	0.00	6.45	1.07	0.00	10.49	11.75	5.34	0.00	7.91
Forma :	0.21	4.66	7.49	2.54	6.46	3.46	0.25	5.64	8.46	9.16	0.71	7.47

Гаплогруппа R1b (N=7770 гаплотипов)

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
xi :	0.2	171.9	22.2	91.7	653.8	901.2	1.0	258.9	176.9	248.9	77.2	136.3
nu :	4	6	5	4	7	7	4	5	6	5	6	8
Rom:	0.00	47.90	5.43	31.00	172.88	238.99	0.00	80.29	49.33	77.12	20.57	32.07
Forma :	0.21	5.20	0.37	7.38	2.01	4.48	0.21	0.63	4.16	1.50	0.76	2.15

Гаплогруппа R1b1 (N=4625 гаплотипов)

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
xi :	0.4	106.0	0.6	53.4	197.6	369.4	1.1	34.9	120.0	41.6	18.5	19.8
nu :	4	6	5	4	6	7	4	4	6	4	5	6
Rom:	0.00	28.88	0.00	17.47	55.30	96.85	0.00	10.91	32.90	13.30	4.27	3.98
Forma :	0.22	5.44	0.07	8.74	1.75	4.78	0.25	1.11	4.59	2.15	1.17	2.21

Гаплогруппа R1b1c (N=4050 гаплотипов)

маркер:	1	2	3	4	5	6	7	8	9	10	11	12
xi :	0.2	74.0	2.4	37.1	255.3	1490.8	0.8	0.1	13.7	22.8	2.0	4.8
nu :	4	6	4	4	6	7	4	3	5	4	5	6
Rom:	0.00	19.62	0.00	11.71	71.96	396.56	0.00	0.00	2.76	6.66	0.00	0.00
Forma :	0.22	3.82	0.31	8.04	1.72	6.24	0.09	0.19	4.51	1.32	0.43	1.24

Таблица 5-4. Распределение критерия отличия формы по интервалам

Интервалы (%):	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	9-10
Доля (%):	55%	8.3%	10%	10%	5%	3.3%	8.4%	1.7%	0.6%

Как следует из таблицы 5-4, более чем у 50% исследованных маркеров экспериментальные гистограммы по своей форме отличаются от теоретической аппроксимации менее, чем на 1%, и лишь у 1.7% маркеров различие формы гистограмм достигает 8%.

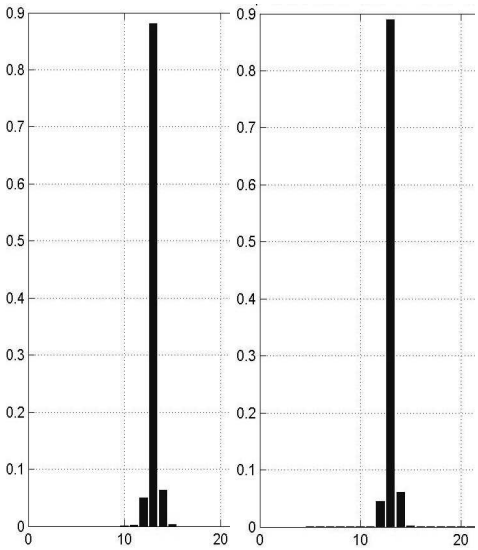
Таблица 5-5. Распределение критерия Романовского по интервалам

Интервал :	0-3	3-10	10-50	50 и более
Доля (%) :	71.7%	16.7%	10%	1.6%

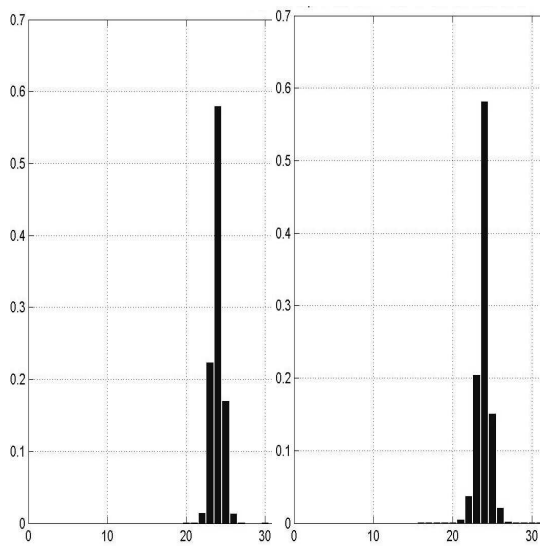
Как следует из таблицы 5-4, более чем у 70% исследованных маркеров имеет место очень хорошее совпадение теоретической аппроксимации и гистограммы экспериментов по критерию хи-квадрат, 16% - совпадение не очень хорошее, 11% - плохое (гаплогруппы R1b, R1b1, R1b1c) в маркерах 5 и 6 (см. табл. 5-3). Тем не менее, даже на этих маркерах совпадение формы удовлетворительное.

В качестве примеров качества аппроксимации реальных гистограмм модельными (с параметрами являющимися МП-оценками) далее представлена гаплогруппа R1b1. Все гистограммы даны для каждого маркера парами: слева – экспериментальная гистограмма, а справа – теоретическая аппроксимация.

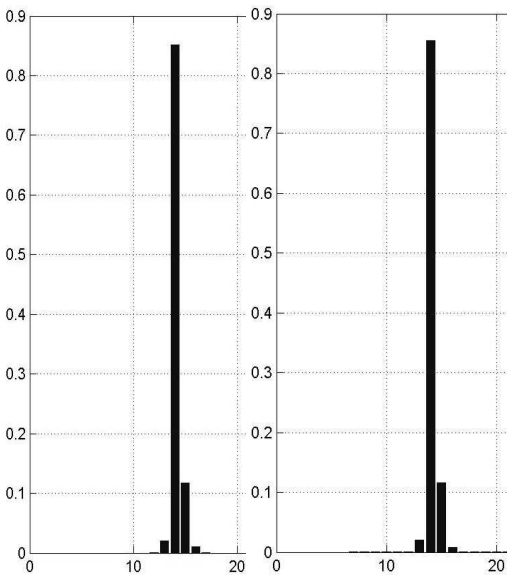
Гаплогруппа R1b1 (4620 измерений) слева – реальные гистограммы, справа – теоретическая аппроксимация



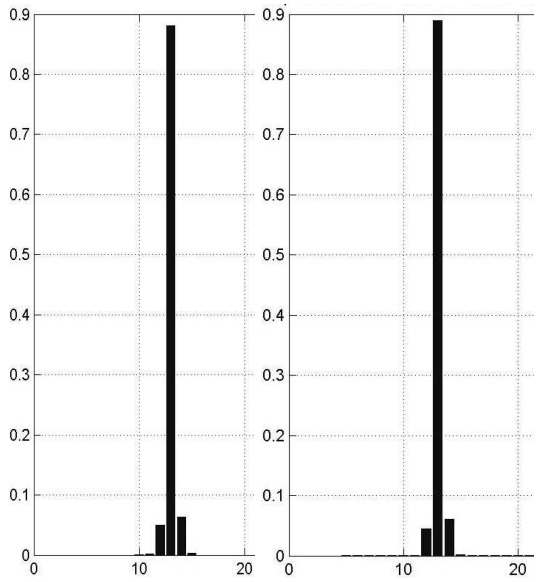
маркер №1



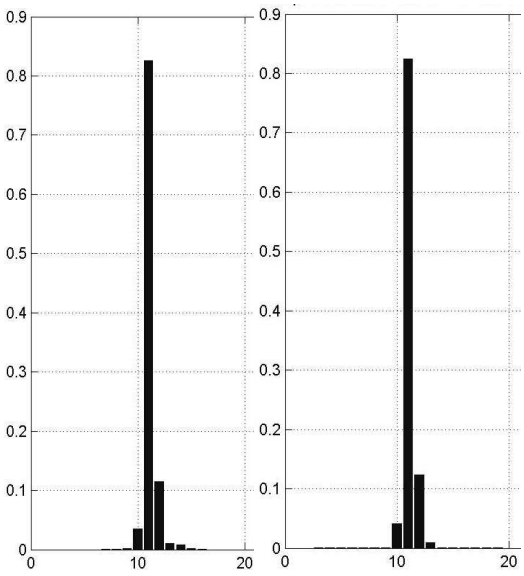
маркер №2



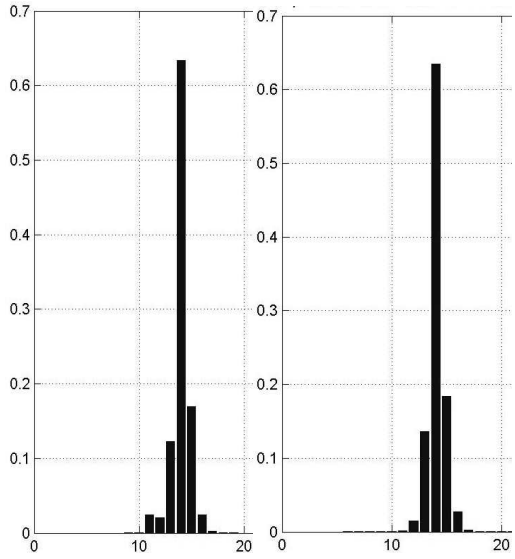
маркер №3



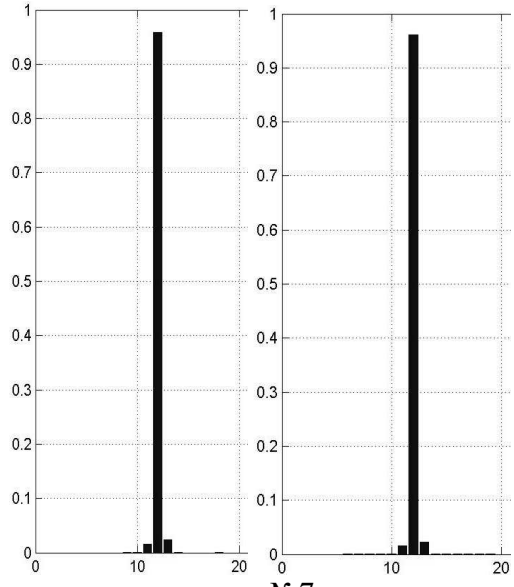
маркер №4



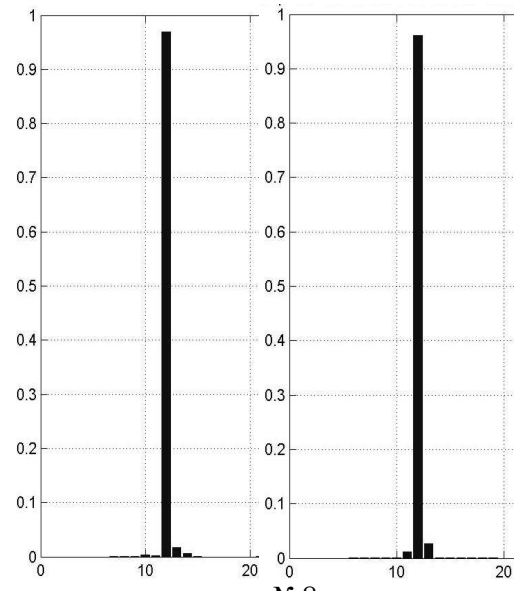
маркер №5



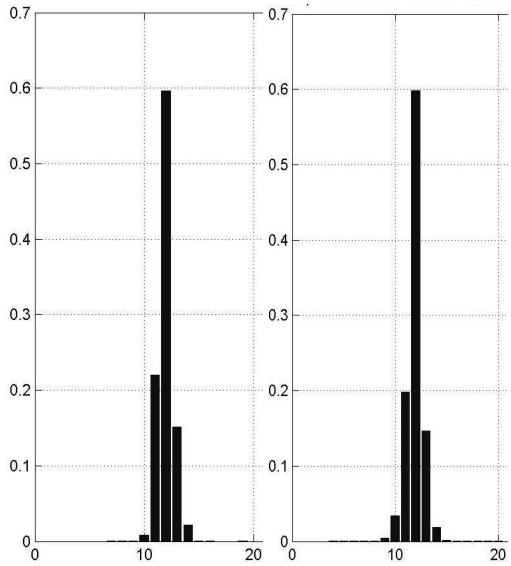
маркер №6



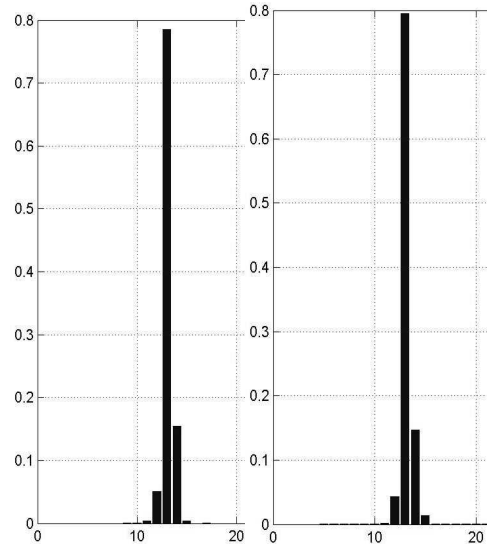
маркер №7



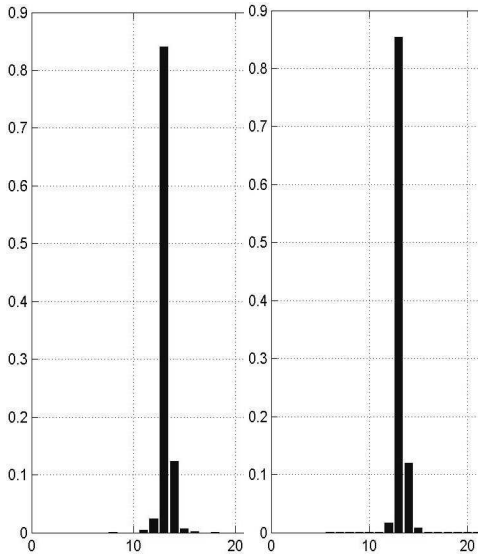
маркер №8



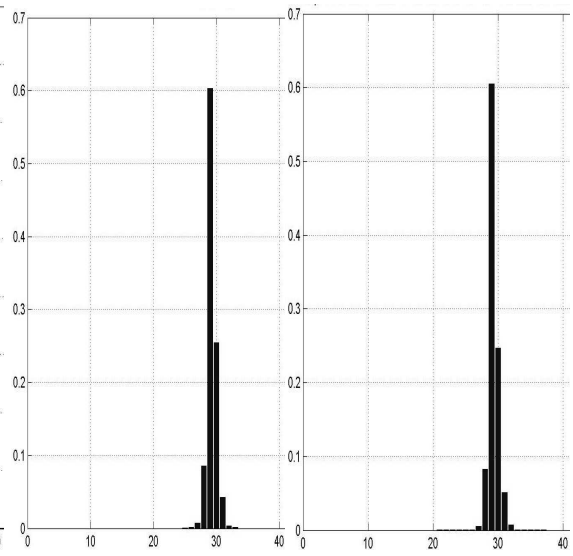
маркер №9



маркер №10



маркер №11



маркер №12

5.4. Анализ влияния ограниченности используемых выборок по сравнению с полными ансамблями данных

В данном случае под полной выборкой понимаются все существующие в настоящий момент члены определенной гаплогруппы. Очевидно, что наличие данных по всем членам гаплогрупп – дело далекого будущего, то возникает вопрос, а как влияет ограниченность выборки в предлагаемом методе на «выявление» гаплотипа истинного первопредка и на количество мутаций, отделяющего его от финальных (современных) потомков. Не получаем ли мы результаты для некоторых «промежуточных» предков (очевидно, что с ростом объема выборки мы должны все более приближаться к истинному первопредку исследуемой гаплогруппы)?

Конечно, можно использовать статистическое моделирование, но значительно больший интерес представляет обработка реальных данных. Чтобы выяснить это, данные по гаплогруппе R1b были произвольно разделены на 8 подгрупп примерно по 1000 штук в каждой, а затем была осуществлена процедура поиска МП-оценок параметров для каждой подгруппы. Результаты по МП-оценкам среднего количества мутаций μ приведены в таблице 5-6. В первой строке таблицы даны результаты, полученные сразу по всем 7780 гаплотипам, а в последней строке относительное среднеквадратическое отклонение (в процентах) от величины, полученной по 7780 гаплотипам (значения в первой строке таблицы):

Таблица 5-6.

маркер: 1	2	3	4	5	6	7	8	9	10	11	12
0.1100	0.5350	0.1225	0.3775	0.1700	0.5125	0.0275	0.0300	0.5125	0.2500	0.1400	0.5475
0.1225	0.5275	0.1300	0.4025	0.1775	0.4475	0.0275	0.0225	0.4625	0.2400	0.1325	0.5400
0.1175	0.5200	0.1200	0.3925	0.1525	0.4675	0.0325	0.0250	0.5525	0.2650	0.1425	0.5900
0.1050	0.4900	0.1200	0.4100	0.1500	0.5450	0.0200	0.0350	0.4900	0.2525	0.1575	0.5250
0.0800	0.5425	0.1200	0.3750	0.1650	0.5450	0.0300	0.0250	0.5200	0.3000	0.1350	0.5825
0.1025	0.5350	0.1275	0.3725	0.1800	0.4950	0.0250	0.0325	0.5125	0.2375	0.1375	0.5150
0.1050	0.5675	0.1050	0.3725	0.1900	0.5475	0.0250	0.0225	0.4925	0.2275	0.1300	0.5300
0.1175	0.5575	0.1300	0.3325	0.1825	0.4800	0.0325	0.0400	0.5600	0.2325	0.1325	0.5250
0.1425	0.5250	0.1250	0.3700	0.1650	0.5850	0.0200	0.0350	0.5075	0.2525	0.1600	0.5325
16.6%	4.5%	6.7%	6.4%	8.5%	9.4%	18.5%	22.5%	6.3%	9.3%	8.4%	5.2%

Как следует из последней строки таблицы 5-6, наибольшие ошибки в определении среднего количества мутаций, произошедших от первопредка, получились для «медленных» маркеров (малая величина параметра μ). Это, по-видимому, является следствием узости исходных гистограмм и соответственно плохой точности определения параметров по ним. Напротив, для более быстрых маркеров мы получили довольно маленькие ошибки. Если учесть, что в суммарную по всем маркерам величину параметра μ именно «быстрые» маркеры вносят наибольший вклад, то и суммарная ошибка по каждому из восьми подгрупп гаплотипов будет приводить к незначительному отличию параметра μ от более точного его значения, полученного по 7780 гаплотипам.

Таким образом, применение метода к выборкам более 1000 гаплотипов обеспечивает довольно устойчивые результаты и, по-видимому, близкие к истинному значению среднего количества мутаций, отделяющего первопредка от его финальных потомков.

5.5. Выводы

Проведена проверка на адекватность разработанной математической модели поведения повторов в маркерах в результате мутаций и реальных данных (больших наборов гаплотипов из пяти гаплогрупп I1a, J2, R1b, R1b1, R1b1c). Степень адекватности модели и реальности определялась в основном по величине критерия хи-квадрат, которая получалась в процессе сравнения гистограмм повторов в маркерах совокупности реальных гаплотипов и гистограмм теоретических распределений, параметры которых были получены как максимально правдоподобные оценки по тем же реальным гаплотипам. Модель показала очень высокую степень адекватности (см. таблицы 5-3, 5-4, 5-5).

ГЛАВА 6. Методики применения «группового» и «одиночного» алгоритмов

В данной главе представлены две методики вычисления времени, прошедшего от первопредка.

Первая методика реализует алгоритм поиска параметров гаплогруппы по большой совокупности гаплотипов современных индивидов, заведомо принадлежащих одной и той же гаплогруппе. В частности, вычисляется и время, прошедшее от времени жизни первопредка до настоящего времени.

Вторая методика реализует алгоритм поиска времени, отделяющего первопредка данной гаплогруппы от времени жизни исследуемого индивида. Данная методика может применяться только после использования первой методики, поскольку требуется знание параметров гаплогруппы, в которую входит исследуемый гаплотип.

Предварительно отметим, что прямое определение времени, когда жил индивид, чей гаплотип мы исследуем, невозможно. Это время вычисляется опосредованно на основании накопленных мутаций.

Весь цикл вычислений можно разбить на три этапа:

Этап 1. Определение количества мутаций, осуществившихся за время, прошедшее от гаплотипа первопредка до исследуемого гаплотипа. Если же решается задача времени жизни самого первопредка, то определяется количество мутаций за время, прошедшее от перпопредка до настоящего момента.

Этап 2. Определение частоты (интенсивности) мутаций в отдельных маркерах (локусах) Y-хромосомы.

Этап 3. Вычисление времени, разделяющего гаплотип первопредка и исследуемый гаплотип. Или же выясняется время жизни первопредка.

Если задачи 1-го и 2-го этапов решаются независимо друг от друга на основе анализа разных наборов гаплотипов, то 3-й этап осуществим только после завершения работ по 1-му и 2-му этапам.

6.1. Методика оценки времени, прошедшего от первопредка до современных потомков

Рассмотрим последовательность действий, которые необходимо осуществить для оценки времени, прошедшего от первопредка до современных потомков, чьи гаплотипы мы исследуем. Рабочий алгоритм, используемый для этой процедуры, описан в п.2.2.

Последовательность действий

ШАГ 1. Сформировать текстовый файл с числовой матрицей размера $(M \times N)$ (выражение (i.1) во Введении), соответствующей исследуемой совокупности гаплотипов (M – количество маркеров, N – количество гаплотипов).

ШАГ 2. Для всех M штук маркеров осуществить процедуру МП-оценивания параметров $\{p_1, p_2, m_0 \quad \mu\}$с помощью программно реализованного алгоритма поиска МП-оценок, изложенного в п.2.1. В результате получим набор параметров исследуемой гаплогруппы:

- Среднее количество мутаций для каждого маркера $\{\hat{\mu}_1, ..., \hat{\mu_M}\}$;
- Вероятность увеличения маркера в случае одношаговых мутаций $\{p_1(1), ..., p_1(M)\}$;
- Вероятность увеличения маркера в случае двухшаговых мутаций $\{p_2(1), ..., p_2(M)\}$;
- Гаплотип первопредка $\{m_0(1), ..., m_0(M)\}$.

ШАГ 3. Учесть поправки на смещение найденных оценок $\{\hat{\mu}_1, ..., \hat{\mu_M}\}$ параметра μ для всех маркеров в соответствии с размером выборки N , и используя зависимости на рис.2-3 (в параграфе 2.3).

Несмещенные значения вычисляются по всем маркерам в соответствии с (2.2.5), только «в обратном порядке»:

$$\mu_0(j) = \frac{\hat{\mu}_j}{\tilde{S}_\mu(N) + 1} \quad , \qquad (6.1.1)$$

где

$\hat{\mu}_j$ - МП-оценка (смещенная) параметра μ для j -го маркера;

$\tilde{S}_\mu(N)$ - поправка на относительное смещение (переведенная обратно из процентов в относительное значение путем деления на 100), взятая из рис.2-3 в соответствии со значениями N и $\hat{\mu}_j$.

ШАГ 4. Осуществить расчет точности (границы доверительного интервала) полученной оценки параметра μ в соответствии с размером выборки N , используя зависимости на рис.2-4 (в параграфе 2.3).

4.1. Вычислить оценку суммарного среднего количества мутаций

$$\hat{\mu}_\Sigma = \sum_{j=1}^{M} \hat{\mu}_0(j) \quad [\textit{мутации/гаплотип}] \quad . \qquad (6.1.2)$$

4.2. Вычислить дисперсию D_μ оценки суммарного среднего количества мутаций $\hat{\mu}_\Sigma$ с помощью графика (рис.2-4) для заданного количества N (для простоты для всех маркеров величину $\tilde{\sigma}_\mu(j)$ можно брать одной и той же):

$$D_\mu = \sum_{j=1}^{M} (\tilde{\sigma}_\mu(j) \cdot \hat{\mu}_0(j))^2 \quad . \qquad (6.1.3)$$

4.3. Вычислить границы оценки среднего количества мутаций по всему гаплотипу (например, для уровня «двух сигма»):

$$\mu_{min} = \hat{\mu}_\Sigma - 2\sqrt{D_\mu} \qquad (6.1.4)$$

$$\mu_{max} = \hat{\mu}_\Sigma + 2\sqrt{D_\mu}$$

ШАГ 5. Вычислить искомое значение времени (в количестве поколений), отделяющего первопредка от современных потомков.

Вычисляем точечную оценку времени до первопредка следующим образом:

$$\hat{T} = \frac{\hat{\mu}_\Sigma}{\hat{\theta}} \quad , \quad [\textit{поколения}]). \qquad (6.1.5)$$

где

$\hat{\mu}_{\Sigma}$ - вычисленная на шаге 3 оценка суммарного по всему гаплотипу среднего количества мутаций (размерность – [*мутации/гаплотип*]);

$\hat{\theta}$ - интенсивность мутаций [*мутации/(гаплотип×поколение)*], вычисленная по парам «отец-сын» (по методике, представленной в параграфе 4.1), причем, для гаплотипа порядка M .

Калибровочные значения оценки $\hat{\theta}$, полученные А.А Клёсовым для гаплотипов различного порядка M представлены в таблице 4-3 (см. параграф 4.1).

ШАГ 6. Осуществить расчет точности (границы доверительного интервала $T_{min} \div T_{max}$) полученной оценки времени T .

Поскольку точное распределение временного интервала T получить не удалось, то приближенно границы доверительного интервала можно вычислить следующим образом:

$$T_{min} = \frac{\mu_{min}}{\theta_{max}} \qquad T_{max} = \frac{\mu_{max}}{\theta_{min}} \tag{6.1.6}$$

где

μ_{min} и μ_{max} - границы доверительного интервала вычисленного количества мутаций, например, по уровню «двух сигма», полученные на Шаге 4;

θ_{min} и θ_{max} - границы доверительного интервала интенсивности мутаций θ для заданного уровня значимости α_{θ} , полученные в соответствии с (4.1.2) в параграфе 4.1, и представленные в таблице 4-1.

Несмотря на то, что интервалы времени между мутациями случайные, все гаплотипы в используемой статистике отстоят от первопредка *на один и тот же интервал* времени. Следовательно, при количестве гаплотипов в выборке, большем единицы, происходит усреднение, и тем сильнее, чем

больше использовано гаплотипов. При объеме выборки $N \geq 100$ гаплотипов можно считать длину интервала времени T неслучайной.

В случае повышения точности вычисления оценки параметра θ (может быть обеспечено все возрастающим объемом обрабатываемой статистики пар гаплотипов «отец-сын») доверительный интервал $\theta_{min} \div \theta_{max}$ сужается, и в пределе оба выражения в (6.1.7) принимают следующий вид:

$$T_{min} = \frac{\mu_{min}}{\hat{\theta}} \qquad T_{max} = \frac{\mu_{max}}{\hat{\theta}} \qquad (6.1.7)$$

При неограниченном возрастании количества гаплотипов в выборке обе границы μ_{min} и μ_{max} будут стремиться к величине $\hat{\mu_{\Sigma}}$. В этом случае выражения (6.1.7) превратятся в (6.1.5).

Но, как следует из графика 2-4 (параграф 2.3), уменьшение дисперсии оценки количества мутаций $\hat{\mu_{\Sigma}}$ с ростом объема выборки гаплотипов N , начиная с величины $N > 500$ сильно замедляется, поэтому пренебрежение разбросом оценки количества мутаций $\hat{\mu_{\Sigma}}$ возможно, по-видимому, при объемах выборки гаплотипов $N > 10000$.

ШАГ 7. Конец

6.2. Методика оценки времени, прошедшего от первопредка до отдельного потомка

Рассмотрим последовательность действий, которые необходимо осуществить для оценки времени, прошедшего от первопредка до отдельного потомка (гаплотипа). Рабочий алгоритм, используемый для этой процедуры, описан в п.3.2.

Последовательность действий

ШАГ 1. Загрузка исходных данных

1.1. Загрузить исследуемый гаплотип в виде набора целых неотрицательных чисел $\{m^{(1)}(1), m^{(1)}(2), \ldots, m^{(1)}(M)\}$ (M – количество маркеров).

1.2. Загрузить параметры гаплогруппы, которой соответствует исследуемый гаплотип:

- Относительные интенсивности мутаций $\{\zeta_1, \zeta_2, \ldots, \zeta_M\}$;
- Вероятность увеличения маркера в случае одношаговых мутаций $\{p_1(1), p_1(2), \ldots, p_1(M)\}$;
- Вероятность увеличения маркера в случае двухшаговых мутаций $\{p_2(1), p_2(2), \ldots, p_2(M)\}$;
- Гаплотип первопредка $\{m^{(0)}(1), m^{(0)}(2), \ldots, m^{(0)}(M)\}$.

ШАГ 2. Осуществить процедуру МП-оценивания параметров $\{ p_1 , p_2 , m_0 \mu \}$ (п. 3.2 в параграфе 3.2), в частности, получить оценку $\{\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M\}$ по всем маркерам.

ШАГ 3. Учесть поправку (абсолютную) в $S_\mu(j) = 4{,}8\%$ на смещение найденных оценок $\{\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M\}$ параметра μ в соответствии с п.1.2 параграфа 1.3. Несмещенное значение вычисляется в соответствии с (3.2.8а), только «в обратном порядке»:

$$\mu_0(j) = \hat{\mu}_j - S_\mu(j) \quad . \qquad (j = 1, \ldots, M) \qquad (6.2.1)$$

где

$\hat{\mu}_j$ - МП-оценка (смещенная) параметра μ в j -м маркере;

$S_\mu(j)$ - поправка на смещение в j -м маркере.

ШАГ 4. Вычислить суммарное количество мутаций в гаплотипе:

$$\hat{\mu}_\Sigma = \sum_{j=1}^{M} \hat{\mu}_0(j) \quad [\textit{мутации/гаплотип}] \quad . \qquad (6.2.2)$$

ШАГ 4. Осуществить расчет точности (границы доверительного интервала) полученной оценки параметра $\hat{\mu}_\Sigma$ в соответствии с таблицей 3-1 параграфа 3.3.

4.1. Для заданного уровня значимости α из таблицы 3-1 выбираем величины δ_{min} и δ_{max} .

4.2. Вычисляем границы доверительного интервала, который с вероятностью $1-2\alpha$ накрывает истинное значение количества мутаций μ , отделяющее гаплотип первопредка от исследуемого гаплотипа:

$$\mu_{min} = \hat{\mu_{\Sigma}} \cdot \left[1 - \frac{-\delta_{min}}{100} \right] \qquad (6.2.3a)$$

$$\mu_{max} = \hat{\mu_{\Sigma}} \cdot \left[1 + \frac{\delta_{max}}{100} \right] \qquad (6.2.3b)$$

ШАГ 5. Вычислить искомое значение времени (точечную оценку), отделяющего первопредка от исследуемого гаплотипа:

$$\hat{T} = \frac{\hat{\mu_{\Sigma}}}{\hat{\theta}} \qquad [\textit{поколения}], \qquad (6.2.4)$$

где

$\hat{\mu_{\Sigma}}$ - вычисленная на шаге 3 оценка суммарного среднего количества мутаций по всему гаплотипу (размерность [*мутации/гаплотип*]);

$\hat{\theta}$ - оценка частоты мутаций, вычисленная по парам «отец-сын» (по методике, изложенной в параграфе 4.1), причем, для гаплотипа порядка M .
Размерность - [*мутации/(гаплотип×поколение)*]

ШАГ 6. Осуществить расчет точности (границы доверительного интервала $T_{min} \div T_{max}$) полученной оценки времени T в соответствии с алгоритмом, изложенным в Приложении 1.

ШАГ 7. Конец

6.3. Выводы

1. Таким образом, окончательным результатом исторической датировки первопредка (первая методика) или одиночного потомка (вторая методика) является интервал времени $T_{min} \div T_{max}$ (интервальная оценка). Кроме этого в методиках вычисляются и точечные оценки времени $\hat{T}$ (см. формулы (6.1.5) и (6.2.4)).

2. При использовании методики оценки времени жизни первопредка по большой группе гаплотипов его современных потомков точность оценки времени T растет с ростом количества гаплотипов потомков (при вычислении оценки количества мутаций μ).

3. При использовании методики оценки времени T , прошедшего от первопредка до его одиночного потомка (необязательно нашего современника) точность этой оценки растет с ростом размерности гаплотипа M (количество маркеров) предка и проверяемого гаплотипа потомка.

4. Точность определения времени до одиночного первопредка значительно хуже точности определения времени жизни первопредка по большому объему гаплотипов предков.

5. В обеих рассмотренных методиках точность оценки времени T также растет с увеличением количества пар гаплотипов «отец-сын», по которым вычислялась оценка интенсивности мутаций θ .

Литература

1. Феллер В. *Введение в теорию вероятностей и ее приложения.* М., «Мир», 1984, т.1,2

2. Ван дер Варден Б.Л. *Математическая статистика.* М., Иностр. Лит-ра, 1960, гл.2. Разд.7

3. *Математическая статистика. Биномиальное распределение.* http://algolist.manual.ru /maths/ matstat/binomial/index.php

ПРИЛОЖЕНИЕ 1. Доверительные интервалы для времени, прошедшего от первопредка

В каждом параграфе содержатся таблицы границ доверительных интервалов оценки времени T, прошедшего от первопредка. Эти таблицы получены для вполне определенной интенсивности мутаций θ, которые присущи гаплотипам из конкретного набора маркеров (далее для простоты они называются 6-ти маркерные, 12-ти маркерные, и т.д.).

Интенсивности θ для различных по количеству маркеров гаплотипов были получены А.А Клёсовым на основе анализа и статистической обработки результатов ДНК-тестов на парах «отец-сын» (см. параграф 4.1 и таблицу 6-1 а параграфе 6.1).

В таблицах, представленных ниже, использованы следующие условные обозначения:

- θ_{min} - нижняя граница оценки интенсивности мутаций, вычисленной по парам “отец-сын” (см. параграф 4.1);
- θ_{max} - верхняя граница оценки интенсивности мутаций, вычисленной по парам “отец-сын” (см. Параграф 4.1);
- α_{θ} - вероятность выхода оценки интенсивности мутаций θ за доверительный интервал как «вниз», так и «вверх»;
- $M_T = \hat{T}$ - среднее количество поколений (матожидание распределения Эрланга), которое соответствует количеству мутаций $\hat{\mu_{\Sigma}}$ и для конкретного для значения интенсивности мутаций θ, вычисленному по парам «отец-сын» и для гаплотипов разной размерности (см. формулу 4.2.3);
- σ_T - стандартное отклонение (квадратный. корень из дисперсии D_T распределения Эрланга (4.2.4)) количества поколений (см. формулу 4.2.6);
- $\hat{\mu_{\Sigma}}$ - суммарное количество мутаций, которое накопилось в гаплотипе (вычислена как сумма по всем маркерам максимально правдоподобных оценок мутаций в соответствии с алгоритмами, приведенными в главах 2 и 3);
- α_T - вероятность выхода оценки искомого времени T до первопредка (измеряется в поколениях) за доверительный интервал как «вниз», так и «вверх»;

- T_{min} - нижняя граница доверительного интервала для оценки времени до первопредка T, измеренная в поколениях;
- T_{max} - верхняя граница доверительного интервала для оценки времени до первопредка T, измеренная в поколениях;

В заголовке каждой таблицы даны значения базовых параметров $\hat{\mu_\Sigma}$, M_T и σ_T.

Величины θ и M_T в нашем случае связаны соотношением (4.2.3), поэтому, например, для $\theta = 0{,}0096$ [*мутаций/гаплотип×поколение*] и $\hat{\mu_\Sigma} = 1$ [*мутаций/гаплотип*] получаем $M_T = 104{,}17$ [*поколений*] (см. 1-ю таблицу Приложения 1-3).

При формировании таблиц было решено ограничиться только двумя значениями доверительной вероятности для параметра θ, а именно: $\alpha_\theta = 0{,}1$ и $\alpha_\theta = 0{,}2$. Количество прошедших мутаций $\hat{\mu_\Sigma}$ считалось вычисленным с абсолютной точностью (т.е., не имело разброса), что близко случаю оценки времени до первопредка по большим совокупностям гаплотипов (более 1000).

Таким образом, каждая таблица получена из распределения Эрланга (4.2.4) с определенным значением параметра $\hat{\mu_\Sigma}$ (от 1 до 25) и для значения θ, соответствующего определенному гаплотипу (6-ти маркерному, 12-ти маркерному и т.п.). Строки таблиц, в свою очередь, формировались для различных доверительных вероятностей α_T искомого времени T (измеряемого в *поколениях*).

Каждая строка содержит три пары границ доверительных интервалов.
Правила формирования первой пары T_{min}, T_{max}:
1. по заданным значениям пары параметров θ и $\hat{\mu_\Sigma}$ формировалась таблица значений из распределения Эрланга.
2. Затем, для определенной величины α_T из полученного табличного представления распределения Эрланга численно определялись границы T_{min} и T_{max}.
Правила формирования второй и третьей пар T_{min}, T_{max} одинаковы:
1. по заданным значениям пары параметров θ_{min} и $\hat{\mu_\Sigma}$ формировалась первая таблица значений из распределения Эрланга.

2. Затем, для заданной величины α_T из полученного табличного представления распределения Эрланга численно определялись границы $T_{min}(1)$ и $T_{max}(1)$.

3. по заданным значениям пары параметров θ_{max} и $\hat{\mu_\Sigma}$ формировалась вторая таблица значений из распределения Эрланга.

4. Затем, для заданной величины α_T из полученного табличного представления распределения Эрланга численно определялись границы $T_{min}(2)$ и $T_{max}(2)$.

5. Из пары $T_{min}(1)$ и $T_{min}(2)$ выбиралось наименьшее значение и оно принималось за T_{min} .

6. Из пары $T_{max}(1)$ и $T_{max}(2)$ выбиралось наименьшее значение и оно принималось за T_{max} .

Доверительный интервал $T_{min} \div T_{max}$ наиболее узкий в случае пренебрежения существующим разбросом параметра θ (первая пара чисел в строке). Фактически – это предельный случай при неограниченном повышении точности оценки параметра θ . Вторая пара чисел – это границы доверительного интервала для довольно слабых требований ($\alpha_\theta = 0{,}2$) по вероятности выскакивания истинного значения интенсивности мутаций θ за границы собственного доверительного интервала $\theta_{min} \div \theta_{max}$. И, наконец, третья пара чисел получена для более жестких условий $\alpha_\theta = 0{,}1$, приводящих к наиболее широкому доверительному интервалу $T_{min} \div T_{max}$ времени до первопредка.

Следует заметить, что приведенные а таблицах расчеты границ доверительных интервалов приблизительные, поскольку распределение величины T с учетом случайности параметров θ и $\hat{\mu_\Sigma}$ несколько отличается от распределения Эрланга. Тем не менее, для практических нужд этим обстоятельством можно пренебречь.

Если необходимо учесть разброс вычисленного количества мутаций, что необходимо делать в задаче определения времени, прошедшем от первопрека до “одиночного” гаплотипа (см. главу 3), то можно поступить следующим образом. Поскольку величина $\hat{\mu_\Sigma}$ имеет разброс (см. параграф 3.2)

$$\mu_{min} = \hat{\mu_\Sigma} - 2\sqrt{D_\mu}$$

$$\mu_{max} = \hat{\mu_\Sigma} + 2\sqrt{D_\mu} \quad ,$$

или более точно, например, для гаплогруппы R1b при заданном уровне значимости α (см. результаты статистического моделирования в Таблице 3-1 параграфа 3.2):

$$\mu_{min} = \hat{\mu}_{\Sigma}\left[1 - \frac{-\delta_{min}}{100}\right]$$

$$\mu_{max} = \hat{\mu}_{\Sigma}\left[1 + \frac{\delta_{max}}{100}\right] \quad ,$$

то, необходимо искать две пары окончательных доверительных интервалов (для μ_{max} и μ_{min} отдельно), а результирующий интервал времени (в поколениях) формировать как объединение двух исходных интервалов. Поскольку таблицы в Приложении 1 даны только для целых значений параметра $\hat{\mu}_{\Sigma}$, то можно применить, например, линейную интерполяцию.

Примеры использования таблиц

Пример 1

Пусть исследуемый гаплотип имеет 12 маркеров. Вычисленное общее количество мутаций, отделяющих его от первопредка $\hat{\mu}_{\Sigma} = 4{,}3$. Требуемые доверительные вероятности $\alpha_{\theta} = 0{,}1$ и $\alpha_{T} = 0{,}15$ (вероятность выхода истинного значения времени за границы интервала равна $0{,}15 \times 2 = 0{,}3$, а вероятность того, что истинное значение накрывается интервалом, равна 0,7, то есть, 70%).

Из таблицы для $\hat{\mu}_{\Sigma} = 4$ Приложения 1-2 в строке для $\alpha_{T} = 0{,}15$ из третьей (левой) пары получаем: $T_{min} = 71{,}55$ и $T_{max} = 297{,}27$
Аналогично, из таблицы для $\hat{\mu}_{\Sigma} = 5$ Приложения 1-2 в строке для $\alpha_{T} = 0{,}15$ из третьей (левой) пары получаем: $T_{min} = 98{,}2$ и $T_{max} = 359{,}2$

Затем, линейной интерполяцией окончательно вычисляем искомые границы доверительного интервала:

$$T_{min} = 98{,}2 - (98{,}2 - 71{,}55) \times (5 - 4.3) = 79{,}55$$

$$T_{max} = 359{,}2 - (359{,}2 - 297{,}27) \times (5 - 4.3) = 315{,}85$$

Следовательно, первопредка от исследуемого индивида с вероятностью 70% отделяет временной интервал в $79.\div 315.85$ поколений.

В следующем примере 2 рассмотрен более усложненный вариант примера 2, учитывающий возможный разброс количества мутаций. В этом случае, как уже говорилось, приходится рассматривать два граничных значения величины $\hat{\mu_\Sigma}$, а именно, μ_{max} и μ_{min} .

Пример 2

Пусть исследуемый гаплотип имеет 12 маркеров. Вычисленное общее количество мутаций, отделяющих его от первопредка (точечная оценка), как и ранее, $\hat{\mu_\Sigma}=4{,}3$, а границы доверительного интервала пусть равны $\mu_{min}=3{,}62$ и $\mu_{max}=5{,}12$.

Из таблицы для $\hat{\mu_\Sigma}=3$ Приложения 1-2 в строке для $\alpha_T=0{,}15$ из третьей (левой) пары получаем: $T_{min}=47{,}35$ и $T_{max}=233{,}35$

Аналогично, из таблицы для $\hat{\mu_\Sigma}=4$ Приложения 1-2 в строке для $\alpha_T=0{,}15$ из третьей (левой) пары получаем: $T_{min}=71{,}55$ и $T_{max}=297{,}27$

Затем, линейной интерполяцией вычисляем первую пару искомых границ доверительного интервала:

$$T_{min}(1)= 71{,}55-(71{,}55-47{,}35)\times(4-3.62) = 62{,}35$$

$$T_{max}(1)= 297{,}27-(297{,}27-233{,}35)\times(4-3.62) = 272{,}98$$

Из таблицы для $\hat{\mu_\Sigma}=5$ Приложения 1-2 в строке для $\alpha_T=0{,}15$ из третьей (левой) пары получаем: $T_{min}=98{,}2$ и $T_{max}=359{,}2$.

Аналогично, из таблицы для $\hat{\mu_\Sigma}=6$ Приложения 1-2 в строке для $\alpha_T=0{,}15$ из третьей (левой) пары получаем: $T_{min}=126{,}26$ и $T_{max}=419{,}14$

Затем, линейной интерполяцией вычисляем вторую пару искомых границ доверительного интервала:

$$T_{min}(2)= 126{,}26-(126{,}26-98{,}2)\times(6-5.12) = 101{,}57$$

$$T_{max}(2)= 419{,}14-(419{,}14-359{,}2)\times(6-5.12) = 366{,}39$$

Окончательно получаем

$$T_{min} = min\{T_{min}(1)\ ;\ T_{min}(2)\} = min\{71{,}55\ ;\ 101{,}57\} = 71{,}55$$

$$T_{max} = max\{T_{max}(1)\ ;\ T_{max}(2)\} = max\{272{,}98\ ;\ 366{,}39\} = 366{,}39$$

Следовательно, первопредка от исследуемого индивида с вероятностью 80% отделяет временной интервал в $71{,}55 \div 366{,}39$ поколений.

ПРИЛОЖЕНИЕ 1-1. Доверительные интервалы для 6-ти маркерного гаплотипа

Интенсивность мутаций $\theta = 0{,}0096$ [*мутаций/(гаплотип×поколение)*]

При уровне значимости $\alpha_\theta = 0{,}1$ для данного θ :

$\theta_{min} = 0{,}0081$ $\theta_{max} = 0{,}0114$

При уровне значимости $\alpha_\theta = 0{,}2$ для данного θ :

$\theta_{min} = 0{,}0089$ $\theta_{max} = 0{,}0108$

==========

$\hat{\mu_\Sigma} = 1$ $M_T = 104.17$ $\sigma_T = 104.17$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1 - 2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	2.08	479.17	1.85	536.58	1.75	569.76
0.05	0.90	6.25	311.46	5.56	348.78	5.26	370.34
0.10	0.80	11.46	239.58	10.20	268.29	9.65	284.88
0.15	0.70	17.71	196.88	15.76	220.46	14.91	234.10
0.20	0.60	23.96	166.67	21.32	186.64	20.17	198.18

==========

$\hat{\mu_\Sigma}$ =2 M_T =208.33 σ_T =147.31

		α_θ=0,2		α_θ=0,1	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	16.67	689.58	14.83	772.21	14.03	819.96
0.05	0.90	37.50	493.75	33.37	552.91	31.57	587.10
0.10	0.80	56.25	404.17	50.06	452.59	47.35	480.58
0.15	0.70	72.92	350.00	64.90	391.94	61.38	416.17
0.20	0.60	87.50	310.42	77.87	347.61	73.65	369.10

==========

$\hat{\mu_\Sigma}$ =3 M_T =312.5 σ_T =180.42

		α_θ=0,2		α_θ=0,1	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	46.88	875.00	41.72	979.84	39.46	1040.43
0.05	0.90	87.50	653.13	77.87	731.38	73.65	776.61
0.10	0.80	115.63	553.13	102.91	619.40	97.33	657.70
0.15	0.70	140.63	490.63	125.16	549.41	118.37	583.38
0.20	0.60	162.50	443.75	144.62	496.92	136.78	527.65

==========

$\hat{\mu_\Sigma}$ =4 M_T =416.67 σ_T =208.33

		α_θ=0,2		α_θ=0,1	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	87.50	1045.83	77.87	1171.15	73.65	1243.56
0.05	0.90	145.83	804.17	129.79	900.52	122.76	956.20
0.10	0.80	183.33	695.83	163.17	779.21	154.32	827.39
0.15	0.70	212.50	625.00	189.12	699.89	178.87	743.16
0.20	0.60	241.67	570.83	215.08	639.23	203.42	678.76

==========

$\hat{\mu_\Sigma}$ =5 M_T =520.83 σ_T =232.92

				$\alpha_\theta=0,2$		$\alpha_\theta=0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	135.42	1208.33	120.52	1353.12	113.99	1436.78
0.05	0.90	208.33	953.13	185.42	1067.33	175.36	1133.32
0.10	0.80	255.21	828.13	227.13	927.35	214.82	984.69
0.15	0.70	291.67	755.21	259.58	845.70	245.51	897.99
0.20	0.60	322.92	697.92	287.39	781.54	271.82	829.87

ПРИЛОЖЕНИЕ 1-2. Доверительные интервалы для 12-ти маркерного гаплотипа

Интенсивность мутаций $\theta=0,024$ [*мутаций/(гаплотип×поколение)*]

При уровне значимости $\alpha_\theta=0,1$ для данного θ :

$\theta_{min}=0,0202$ $\theta_{max}=0,0285$

При уровне значимости $\alpha_\theta=0,2$ для данного θ :

$\theta_{min}=0,0214$ $\theta_{max}=0,0269$

==========

$\hat{\mu_\Sigma}$ = 1 M_T = 41.67 σ_T = 41.67

				$\alpha_\theta=0,2$		$\alpha_\theta=0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	0.83	191.67	0.74	214.63	0.70	227.90
0.05	0.90	2.50	124.58	2.22	139.51	2.10	148.14
0.10	0.80	4.58	95.83	4.08	107.32	3.86	113.95
0.15	0.70	7.08	78.75	6.30	88.19	5.96	93.64
0.20	0.60	9.58	66.67	8.53	74.65	8.07	79.27

==========

$\hat{\mu_\Sigma} = 2$ $M_T = 83.33$ $\sigma_T = 58.93$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	6.67	275.83	5.93	308.88	5.61	327.98
0.05	0.90	15.00	197.50	13.35	221.16	12.63	234.84
0.10	0.80	22.50	161.67	20.02	181.04	18.94	192.23
0.15	0.70	29.17	140.00	25.96	156.77	24.55	166.47
0.20	0.60	35.00	124.17	31.15	139.04	29.46	147.64

==========

$\hat{\mu_\Sigma} = 3$ $M_T = 125.00$ $\sigma_T = 72.17$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	18.75	350.00	16.69	391.94	15.78	416.17
0.05	0.90	35.00	261.25	31.15	292.55	29.46	310.64
0.10	0.80	46.25	221.25	41.16	247.76	38.93	263.08
0.15	0.70	56.25	196.25	50.06	219.76	47.35	233.35
0.20	0.60	65.00	177.50	57.85	198.77	54.71	211.06

==========

$\hat{\mu_\Sigma} = 4$ $M_T = 166.67$ $\sigma_T = 83.33$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	35.00	418.33	31.15	468.46	29.46	497.42
0.05	0.90	58.33	321.67	51.92	360.21	49.10	382.48
0.10	0.80	73.33	278.33	65.27	311.68	61.73	330.96
0.15	0.70	85.00	250.00	75.65	279.96	71.55	297.27
0.20	0.60	96.67	228.33	86.03	255.69	81.37	271.50

$\hat{\mu_\Sigma} = 5$ $M_T = 208.33$ $\sigma_T = 93.17$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	54.17	483.33	48.21	541.25	45.59	574.71
0.05	0.90	83.33	381.25	74.17	426.93	70.15	453.33
0.10	0.80	102.08	331.25	90.85	370.94	85.93	393.88
0.15	0.70	116.67	302.08	103.83	338.28	98.20	359.20
0.20	0.60	129.17	279.17	114.96	312.62	108.73	331.95

$\hat{\mu_\Sigma} = 6$ $M_T = 250.00$ $\sigma_T = 102.06$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	75.00	545.00	66.75	610.30	63.13	648.04
0.05	0.90	110.00	437.50	97.90	489.92	92.59	520.21
0.10	0.80	132.50	385.00	117.92	431.13	111.53	457.79
0.15	0.70	150.00	352.50	133.50	394.74	126.26	419.14
0.20	0.60	165.00	327.50	146.85	366.74	138.89	389.42

$\hat{\mu_\Sigma} = 7$ $M_T = 291.67$ $\sigma_T = 110.24$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	99.17	606.67	88.26	679.36	83.47	721.36
0.05	0.90	137.08	492.92	122.00	551.98	115.39	586.11
0.10	0.80	163.33	437.50	145.37	489.92	137.49	520.21
0.15	0.70	183.75	402.50	163.54	450.73	154.67	478.60
0.20	0.60	198.33	376.25	176.52	421.33	166.95	447.38

$\hat{\mu_\Sigma} = 8$ $M_T = 333.33$ $\sigma_T = 117.85$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	123.33	663.33	109.77	742.81	103.82	788.74
0.05	0.90	166.67	546.67	148.33	612.17	140.29	650.02
0.10	0.80	196.67	490.00	175.03	548.71	165.54	582.64
0.15	0.70	216.67	453.33	192.83	507.65	182.38	539.04
0.20	0.60	233.33	423.33	207.67	474.06	196.41	503.37

$\hat{\mu_\Sigma} = 9$ $M_T = 375.00$ $\sigma_T = 125.00$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	146.25	723.75	130.16	810.47	123.11	860.58
0.05	0.90	198.75	600.00	176.89	671.89	167.30	713.44
0.10	0.80	228.75	540.00	203.59	604.70	192.55	642.09
0.15	0.70	251.25	502.50	223.61	562.71	211.49	597.50
0.20	0.60	270.00	472.50	240.30	529.12	227.27	561.83

$\hat{\mu_\Sigma} = 10$ $M_T = 416.67$ $\sigma_T = 131.76$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	175.00	779.17	155.75	872.53	147.31	926.48
0.05	0.90	229.17	654.17	203.96	732.55	192.90	777.84
0.10	0.80	262.50	591.67	233.62	662.56	220.96	703.53
0.15	0.70	287.50	550.00	255.87	615.90	242.00	653.98
0.20	0.60	304.17	520.83	270.71	583.24	256.03	619.30

ПРИЛОЖЕНИЕ 1-3. Доверительные интервалы для 25-ти маркерного гаплотипа

Интенсивность мутаций $\theta = 0{,}046$ [*мутаций/(гаплотип×поколение)*]

При уровне значимости $\alpha_\theta = 0{,}1$ для данного θ :

$\theta_{min} = 0{,}0387$ $\theta_{max} = 0{,}0546$

При уровне значимости $\alpha_\theta = 0{,}1$ для данного θ :

$\theta_{min} = 0{,}0411$ $\theta_{max} = 0{,}0515$

==========

$\hat{\mu_\Sigma} = 1$ $M_T = 21.74$ $\sigma_T = 21.74$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	0.43	100.00	0.39	111.98	0.37	118.91
0.05	0.90	1.30	65.00	1.16	72.79	1.10	77.29
0.10	0.80	2.39	50.00	2.13	55.99	2.01	59.45
0.15	0.70	3.70	41.09	3.29	46.01	3.11	48.85
0.20	0.60	5.00	34.78	4.45	38.95	4.21	41.36

==========

$\hat{\mu_\Sigma} = 2$ $M_T = 43.48$ $\sigma_T = 30.74$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	3.48	143.91	3.10	161.16	2.93	171.12
0.05	0.90	7.83	103.04	6.97	115.39	6.59	122.52
0.10	0.80	11.74	84.35	10.45	94.45	9.88	100.29
0.15	0.70	15.22	73.04	13.54	81.80	12.81	86.85
0.20	0.60	18.26	64.78	16.25	72.54	15.37	77.03

$\hat{\mu_\Sigma} = 3$ $M_T = 65.22$ $\sigma_T = 37.65$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	9.78	182.61	8.71	204.49	8.23	217.13
0.05	0.90	18.26	136.30	16.25	152.64	15.37	162.07
0.10	0.80	24.13	115.43	21.48	129.27	20.31	137.26
0.15	0.70	29.35	102.39	26.12	114.66	24.70	121.75
0.20	0.60	33.91	92.61	30.18	103.71	28.55	110.12

$\hat{\mu_\Sigma} = 4$ $M_T = 86.96$ $\sigma_T = 43.48$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	18.26	218.26	16.25	244.41	15.37	259.53
0.05	0.90	30.43	167.83	27.09	187.94	25.62	199.56
0.10	0.80	38.26	145.22	34.05	162.62	32.21	172.67
0.15	0.70	44.35	130.43	39.47	146.06	37.33	155.09
0.20	0.60	50.43	119.13	44.89	133.40	42.45	141.65

$\hat{\mu_\Sigma} = 5$ $M_T = 108.70$ $\sigma_T = 48.61$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	28.26	252.17	25.15	282.39	23.79	299.85
0.05	0.90	43.48	198.91	38.70	222.75	36.60	236.52
0.10	0.80	53.26	172.83	47.40	193.53	44.83	205.50
0.15	0.70	60.87	157.61	54.17	176.49	51.24	187.41
0.20	0.60	67.39	145.65	59.98	163.10	56.73	173.19

==========

$\hat{\mu_\Sigma} = 6$ $M_T = 130.43$ $\sigma_T = 53.25$

		$\alpha_\theta = 0{,}2$				$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	39.13	284.35	34.83	318.42	32.94	338.11
0.05	0.90	57.39	228.26	51.08	255.61	48.31	271.42
0.10	0.80	69.13	200.87	61.53	224.94	58.19	238.85
0.15	0.70	78.26	183.91	69.65	205.95	65.88	218.68
0.20	0.60	86.09	170.87	76.62	191.34	72.46	203.17

==========

$\hat{\mu_\Sigma} = 7$ $M_T = 152.17$ $\sigma_T = 57.52$

		$\alpha_\theta = 0{,}2$				$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	51.74	316.52	46.05	354.45	43.55	376.36
0.05	0.90	71.52	257.17	63.65	287.99	60.20	305.80
0.10	0.80	85.22	228.26	75.84	255.61	71.73	271.42
0.15	0.70	95.87	210.00	85.32	235.16	80.70	249.70
0.20	0.60	103.48	196.30	92.10	219.83	87.10	233.42

==========

$\hat{\mu_\Sigma} = 8$ $M_T = 173.91$ $\sigma_T = 61.49$

		$\alpha_\theta = 0{,}2$				$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	64.35	346.09	57.27	387.56	54.16	411.52
0.05	0.90	86.96	285.22	77.39	319.39	73.20	339.14
0.10	0.80	102.61	255.65	91.32	286.28	86.37	303.99
0.15	0.70	113.04	236.52	100.61	264.86	95.15	281.24
0.20	0.60	121.74	220.87	108.35	247.33	102.47	262.63

$\hat{\mu_\Sigma} = 9$ $M_T = 195.65$ $\sigma_T = 65.22$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	76.30	377.61	67.91	422.85	64.23	449.00
0.05	0.90	103.70	313.04	92.29	350.55	87.29	372.23
0.10	0.80	119.35	281.74	106.22	315.50	100.46	335.00
0.15	0.70	131.09	262.17	116.67	293.59	110.34	311.74
0.20	0.60	140.87	246.52	125.37	276.06	118.58	293.13

$\hat{\mu_\Sigma} = 10$ $M_T = 217.39$ $\sigma_T = 68.75$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	91.30	406.52	81.26	455.23	76.86	483.38
0.05	0.90	119.57	341.30	106.41	382.20	100.64	405.83
0.10	0.80	136.96	308.70	121.89	345.68	115.28	367.06
0.15	0.70	150.00	286.96	133.50	321.34	126.26	341.21
0.20	0.60	158.70	271.74	141.24	304.30	133.58	323.11

$\hat{\mu_\Sigma} = 12$ $M_T = 260.87$ $\sigma_T = 75.31$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	120.00	466.96	106.80	522.91	101.01	555.24
0.05	0.90	151.30	393.91	134.66	441.11	127.36	468.39
0.10	0.80	172.17	360.00	153.23	403.14	144.93	428.06
0.15	0.70	185.22	336.52	164.84	376.84	155.91	400.14
0.20	0.60	198.26	320.87	176.45	359.32	166.89	381.53

==========

$\hat{\mu}_{\Sigma} = 14$ $M_T = 304.35$ $\sigma_T = 81.34$

		$\alpha_\theta = 0{,}2$				$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	149.13	523.48	132.73	586.20	125.53	622.45
0.05	0.90	185.65	447.39	165.23	501.00	156.27	531.98
0.10	0.80	206.96	410.87	184.19	460.10	174.21	488.55
0.15	0.70	222.17	386.52	197.73	432.84	187.02	459.60
0.20	0.60	237.39	368.26	211.28	412.39	199.82	437.88

==========

$\hat{\mu}_{\Sigma} = 16$ $M_T = 347.83$ $\sigma_T = 86.96$

		$\alpha_\theta = 0{,}2$				$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	180.87	580.87	160.97	650.47	152.25	690.69
0.05	0.90	219.13	500.87	195.03	560.88	184.45	595.56
0.10	0.80	243.48	462.61	216.69	518.04	204.95	550.07
0.15	0.70	260.87	434.78	232.17	486.88	219.59	516.98
0.20	0.60	274.78	417.39	244.56	467.40	231.30	496.30

ПРИЛОЖЕНИЕ 1-4. Доверительные интервалы для 37-ти маркерного гаплотипа

Интенсивность мутаций $\theta = 0{,}09$ [*мутаций/(гаплотип×поколение)*]

При уровне значимости $\alpha_\theta = 0{,}1$ для данного θ :

$\theta_{min} = 0{,}0757$ $\theta_{max} = 0{,}1069$

При уровне значимости $\alpha_\theta = 0{,}1$ для данного θ :

$\theta_{min} = 0{,}0803$ $\theta_{max} = 0{,}1008$

$\hat{\mu}_{\Sigma} = 1$ $M_T = 11.11$ $\sigma_T = 11.11$

α_T	$1-2\alpha_T$	T_{min}	T_{max}	$\alpha_\theta=0,2$ T_{min}	$\alpha_\theta=0,2$ T_{max}	$\alpha_\theta=0,1$ T_{min}	$\alpha_\theta=0,1$ T_{max}
0.01	0.98	0.22	51.11	0.20	57.24	0.19	60.77
0.05	0.90	0.67	33.22	0.59	37.20	0.56	39.50
0.10	0.80	1.22	25.56	1.09	28.62	1.03	30.39
0.15	0.70	1.89	21.00	1.68	23.52	1.59	24.97
0.20	0.60	2.56	17.78	2.27	19.91	2.15	21.14

$\hat{\mu}_{\Sigma} = 2$ $M_T = 22.22$ $\sigma_T = 15.71$

α_T	$1-2\alpha_T$	T_{min}	T_{max}	$\alpha_\theta=0,2$ T_{min}	$\alpha_\theta=0,2$ T_{max}	$\alpha_\theta=0,1$ T_{min}	$\alpha_\theta=0,1$ T_{max}
0.01	0.98	1.78	73.56	1.58	82.37	1.50	87.46
0.05	0.90	4.00	52.67	3.56	58.98	3.37	62.62
0.10	0.80	6.00	43.11	5.34	48.28	5.05	51.26
0.15	0.70	7.78	37.33	6.92	41.81	6.55	44.39
0.20	0.60	9.33	33.11	8.31	37.08	7.86	39.37

$\hat{\mu}_{\Sigma} = 3$ $M_T = 33.33$ $\sigma_T = 19.25$

α_T	$1-2\alpha_T$	T_{min}	T_{max}	$\alpha_\theta=0,2$ T_{min}	$\alpha_\theta=0,2$ T_{max}	$\alpha_\theta=0,1$ T_{min}	$\alpha_\theta=0,1$ T_{max}
0.01	0.98	5.00	93.33	4.45	104.52	4.21	110.98
0.05	0.90	9.33	69.67	8.31	78.01	7.86	82.84
0.10	0.80	12.33	59.00	10.98	66.07	10.38	70.15
0.15	0.70	15.00	52.33	13.35	58.60	12.63	62.23
0.20	0.60	17.33	47.33	15.43	53.00	14.59	56.28

$\hat{\mu_\Sigma} = 4$ $M_T = 44.44$ $\sigma_T = 22.22$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	9.33	111.56	8.31	124.92	7.86	132.65
0.05	0.90	15.56	85.78	13.84	96.06	13.09	101.99
0.10	0.80	19.56	74.22	17.40	83.12	16.46	88.25
0.15	0.70	22.67	66.67	20.17	74.65	19.08	79.27
0.20	0.60	25.78	60.89	22.94	68.18	21.70	72.40

$\hat{\mu_\Sigma} = 5$ $M_T = 55.56$ $\sigma_T = 24.85$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	14.44	128.89	12.86	144.33	12.16	153.26
0.05	0.90	22.22	101.67	19.78	113.85	18.71	120.89
0.10	0.80	27.22	88.33	24.23	98.92	22.91	105.03
0.15	0.70	31.11	80.56	27.69	90.21	26.19	95.79
0.20	0.60	34.44	74.44	30.66	83.36	28.99	88.52

$\hat{\mu_\Sigma} = 6$ $M_T = 66.67$ $\sigma_T = 27.22$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	20.00	145.33	17.80	162.75	16.84	172.81
0.05	0.90	29.33	116.67	26.11	130.65	24.69	138.72
0.10	0.80	35.33	102.67	31.45	114.97	29.74	122.08
0.15	0.70	40.00	94.00	35.60	105.26	33.67	111.77
0.20	0.60	44.00	87.33	39.16	97.80	37.04	103.84

$\hat{\mu_\Sigma} = 7$ $M_T = 77.78$ $\sigma_T = 29.40$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	26.44	161.78	23.54	181.16	22.26	192.36
0.05	0.90	36.56	131.44	32.53	147.19	30.77	156.30
0.10	0.80	43.56	116.67	38.76	130.65	36.66	138.72
0.15	0.70	49.00	107.33	43.61	120.19	41.25	127.63
0.20	0.60	52.89	100.33	47.07	112.36	44.52	119.30

$\hat{\mu_\Sigma} = 8$ $M_T = 88.89$ $\sigma_T = 31.43$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	32.89	176.89	29.27	198.08	27.68	210.33
0.05	0.90	44.44	145.78	39.56	163.24	37.41	173.34
0.10	0.80	52.44	130.67	46.68	146.32	44.15	155.37
0.15	0.70	57.78	120.89	51.42	135.37	48.63	143.74
0.20	0.60	62.22	112.89	55.38	126.42	52.38	134.23

$\hat{\mu_\Sigma} = 9$ $M_T = 100.00$ $\sigma_T = 33.33$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	39.00	193.00	34.71	216.13	32.83	229.49
0.05	0.90	53.00	160.00	47.17	179.17	44.61	190.25
0.10	0.80	61.00	144.00	54.29	161.25	51.35	171.22
0.15	0.70	67.00	134.00	59.63	150.06	56.40	159.33
0.20	0.60	72.00	126.00	64.08	141.10	60.61	149.82

$\hat{\mu_\Sigma} = 10$ $M_T = 111.11$ $\sigma_T = 35.14$

			$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	46.67	207.78	41.53	232.67	39.28	247.06
0.05	0.90	61.11	174.44	54.39	195.35	51.44	207.43
0.10	0.80	70.00	157.78	62.30	176.68	58.92	187.61
0.15	0.70	76.67	146.67	68.23	164.24	64.53	174.40
0.20	0.60	81.11	138.89	72.19	155.53	68.28	165.15

$\hat{\mu_\Sigma} = 12$ $M_T = 133.33$ $\sigma_T = 38.49$

			$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	61.33	238.67	54.59	267.26	51.63	283.79
0.05	0.90	77.33	201.33	68.83	225.46	65.10	239.40
0.10	0.80	88.00	184.00	78.32	206.05	74.07	218.79
0.15	0.70	94.67	172.00	84.25	192.61	79.69	204.52
0.20	0.60	101.33	164.00	90.19	183.65	85.30	195.01

$\hat{\mu_\Sigma} = 14$ $M_T = 155.56$ $\sigma_T = 41.57$

			$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	76.22	267.56	67.84	299.61	64.16	318.14
0.05	0.90	94.89	228.67	84.45	256.07	79.87	271.90
0.10	0.80	105.78	210.00	94.14	235.16	89.04	249.70
0.15	0.70	113.56	197.56	101.06	221.23	95.59	234.91
0.20	0.60	121.33	188.22	107.99	210.78	102.13	223.81

$\hat{\mu_\Sigma} = 16$ $M_T = 177.78$ $\sigma_T = 44.44$

		$\alpha_\theta = 0,2$		$\alpha_\theta = 0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	82.28	332.46	77.82	353.02
0.05	0.90	99.68	286.67	94.28	304.40
0.10	0.80	110.76	264.78	104.75	281.15
0.15	0.70	118.67	248.85	112.23	264.24
0.20	0.60	125.00	238.90	118.22	253.67

$\hat{\mu_\Sigma} = 18$ $M_T = 200.00$ $\sigma_T = 47.14$

		$\alpha_\theta = 0,2$		$\alpha_\theta = 0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	96.12	362.82	90.91	385.26
0.05	0.90	115.70	315.79	109.43	335.32
0.10	0.80	128.16	293.39	121.21	311.53
0.15	0.70	135.28	277.72	127.95	294.89
0.20	0.60	142.40	266.52	134.68	283.00

$\hat{\mu_\Sigma} = 20$ $M_T = 222.22$ $\sigma_T = 49.69$

		$\alpha_\theta = 0,2$		$\alpha_\theta = 0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	110.76	395.67	104.75	420.13
0.05	0.90	132.51	345.90	125.33	367.29
0.10	0.80	144.38	321.02	136.55	340.86
0.15	0.70	154.27	306.08	145.90	325.01
0.20	0.60	160.20	293.64	151.52	311.80

ПРИЛОЖЕНИЕ 1-5. Доверительные интервалы для 67-ти маркерного гаплотипа

Интенсивность мутаций $\theta = 0{,}145$ [*мутаций/(гаплотип×поколение)*]

При уровне значимости $\alpha_\theta = 0{,}1$ для данного θ :

$\theta_{min} = 0{,}1220$ $\theta_{max} = 0{,}1723$

При уровне значимости $\alpha_\theta = 0{,}2$ для данного θ :

$\theta_{min} = 0{,}1293$ $\theta_{max} = 0{,}1624$

==========

$\hat{\mu_\Sigma} = 1$ $M_T = 6.90$ $\sigma_T = 6.90$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	0.14	31.72	0.12	35.53	0.12	37.72
0.05	0.90	0.41	20.62	0.37	23.09	0.35	24.52
0.10	0.80	0.76	15.86	0.68	17.76	0.64	18.86
0.15	0.70	1.17	13.03	1.04	14.60	0.99	15.50
0.20	0.60	1.59	11.03	1.41	12.36	1.34	13.12

==========

$\hat{\mu_\Sigma} = 2$ $M_T = 13.79$ $\sigma_T = 9.75$

				$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	1.10	45.66	0.98	51.13	0.93	54.29
0.05	0.90	2.48	32.69	2.21	36.61	2.09	38.87
0.10	0.80	3.72	26.76	3.31	29.96	3.13	31.82
0.15	0.70	4.83	23.17	4.30	25.95	4.06	27.55
0.20	0.60	5.79	20.55	5.16	23.01	4.88	24.44

$\hat{\mu}_{\Sigma} = 3$ $M_T = 20.69$ $\sigma_T = 11.95$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	2.76	64.87	2.61	68.88
0.05	0.90	5.16	48.42	4.88	51.42
0.10	0.80	6.81	41.01	6.44	43.54
0.15	0.70	8.29	36.37	7.84	38.62
0.20	0.60	9.58	32.90	9.06	34.93

$\hat{\mu}_{\Sigma} = 4$ $M_T = 27.59$ $\sigma_T = 13.79$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	5.16	77.54	4.88	82.33
0.05	0.90	8.59	59.62	8.13	63.31
0.10	0.80	10.80	51.59	10.22	54.78
0.15	0.70	12.52	46.34	11.84	49.20
0.20	0.60	14.24	42.32	13.47	44.94

$\hat{\mu}_{\Sigma} = 5$ $M_T = 34.48$ $\sigma_T = 15.42$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	7.98	89.59	7.55	95.12
0.05	0.90	12.28	70.66	11.61	75.03
0.10	0.80	15.04	61.40	14.22	65.19
0.15	0.70	17.19	55.99	16.25	59.45
0.20	0.60	19.03	51.74	18.00	54.94

==========

$\hat{\mu_{\Sigma}} = 6$ $M_T = 41.38$ $\sigma_T = 16.89$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	11.05	101.02	10.45	107.26
0.05	0.90	16.20	81.09	15.33	86.10
0.10	0.80	19.52	71.36	18.46	75.77
0.15	0.70	22.10	65.34	20.90	69.38
0.20	0.60	24.31	60.70	22.99	64.46

==========

$\hat{\mu_{\Sigma}} = 7$ $M_T = 48.28$ $\sigma_T = 18.25$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	14.61	112.45	13.82	119.40
0.05	0.90	20.19	91.36	19.10	97.01
0.10	0.80	24.06	81.09	22.76	86.10
0.15	0.70	27.07	74.60	25.60	79.22
0.20	0.60	29.22	69.74	27.63	74.05

==========

$\hat{\mu_{\Sigma}} = 8$ $M_T = 55.17$ $\sigma_T = 19.51$

		$\alpha_\theta = 0{,}2$		$\alpha_\theta = 0{,}1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	18.17	122.95	17.18	130.55
0.05	0.90	24.55	101.32	23.22	107.59
0.10	0.80	28.97	90.82	27.40	96.44
0.15	0.70	31.92	84.03	30.19	89.22
0.20	0.60	34.37	78.46	32.51	83.32

$\hat{\mu_\Sigma} = 9$ $M_T = 62.07$ $\sigma_T = 20.69$

		$\alpha_\theta = 0,2$		$\alpha_\theta = 0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	24.21	119.79	21.54	134.15	20.38	142.44
0.05	0.90	32.90	99.31	29.28	111.21	27.69	118.09
0.10	0.80	37.86	89.38	33.70	100.09	31.87	106.28
0.15	0.70	41.59	83.17	37.01	93.14	35.01	98.90
0.20	0.60	44.69	78.21	39.77	87.58	37.62	92.99

$\hat{\mu_\Sigma} = 10$ $M_T = 68.97$ $\sigma_T = 21.81$

		$\alpha_\theta = 0,2$		$\alpha_\theta = 0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	28.97	128.97	25.78	144.42	24.38	153.35
0.05	0.90	37.93	108.28	33.76	121.25	31.93	128.75
0.10	0.80	43.45	97.93	38.67	109.67	36.57	116.45
0.15	0.70	47.59	91.03	42.35	101.94	40.06	108.25
0.20	0.60	50.34	86.21	44.81	96.54	42.38	102.51

$\hat{\mu_\Sigma} = 12$ $M_T = 82.76$ $\sigma_T = 23.89$

		$\alpha_\theta = 0,2$		$\alpha_\theta = 0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	38.07	148.14	33.88	165.89	32.04	176.14
0.05	0.90	48.00	124.97	42.72	139.94	40.40	148.59
0.10	0.80	54.62	114.21	48.61	127.89	45.98	135.80
0.15	0.70	58.76	106.76	52.29	119.55	49.46	126.94
0.20	0.60	62.90	101.79	55.98	113.99	52.94	121.04

$\hat{\mu_\Sigma} = 14$ $M_T = 96.55$ $\sigma_T = 25.80$

α_T	$1-2\alpha_T$	$\alpha_\theta=0,2$ T_{min}	$\alpha_\theta=0,2$ T_{max}	$\alpha_\theta=0,1$ T_{min}	$\alpha_\theta=0,1$ T_{max}
0.01	0.98	47.31	166.07	42.11	185.97	39.82	197.47
0.05	0.90	58.90	141.93	52.42	158.94	49.58	168.76
0.10	0.80	65.66	130.34	58.43	145.96	55.27	154.99
0.15	0.70	70.48	122.62	62.73	137.31	59.33	145.80
0.20	0.60	75.31	116.83	67.03	130.83	63.39	138.92

$\hat{\mu_\Sigma} = 16$ $M_T = 110.34$ $\sigma_T = 27.59$

α_T	$1-2\alpha_T$	$\alpha_\theta=0,2$ T_{min}	$\alpha_\theta=0,2$ T_{max}	$\alpha_\theta=0,1$ T_{min}	$\alpha_\theta=0,1$ T_{max}
0.01	0.98	57.38	184.28	51.07	206.36	48.30	219.12
0.05	0.90	69.52	158.90	61.87	177.94	58.52	188.94
0.10	0.80	77.24	146.76	68.74	164.34	65.02	174.50
0.15	0.70	82.76	137.93	73.65	154.46	69.66	164.01
0.20	0.60	87.17	132.41	77.58	148.28	73.38	157.45

$\hat{\mu_\Sigma} = 18$ $M_T = 124.14$ $\sigma_T = 29.26$

α_T	$1-2\alpha_T$	$\alpha_\theta=0,2$ T_{min}	$\alpha_\theta=0,2$ T_{max}	$\alpha_\theta=0,1$ T_{min}	$\alpha_\theta=0,1$ T_{max}
0.01	0.98	67.03	201.10	59.66	225.20	56.43	239.12
0.05	0.90	80.69	175.03	71.81	196.01	67.92	208.13
0.10	0.80	89.38	162.62	79.55	182.11	75.24	193.37
0.15	0.70	94.34	153.93	83.97	172.38	79.41	183.03
0.20	0.60	99.31	147.72	88.39	165.42	83.59	175.65

$\hat{\mu_\Sigma} = 20$ $M_T = 137.93$ $\sigma_T = 30.84$

		$\alpha_\theta = 0,2$		$\alpha_\theta = 0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	77.24	219.31	68.74	245.59	65.02	260.77
0.05	0.90	92.41	191.72	82.25	214.70	77.79	227.97
0.10	0.80	100.69	177.93	89.61	199.25	84.76	211.57
0.15	0.70	107.59	169.66	95.75	189.98	90.56	201.73
0.20	0.60	111.72	162.76	99.43	182.26	94.04	193.53

$\hat{\mu_\Sigma} = 25$ $M_T = 172.41$ $\sigma_T = 34.48$

		$\alpha_\theta = 0,2$		$\alpha_\theta = 0,1$	
α_T	$1-2\alpha_T$	T_{min}	T_{max}	T_{min}	T_{max}
0.01	0.98	103.45	262.07	92.07	293.47	87.08	311.62
0.05	0.90	120.69	232.76	107.41	260.65	101.59	276.76
0.10	0.80	131.03	217.24	116.62	243.27	110.30	258.31
0.15	0.70	137.93	206.90	122.76	231.69	116.10	246.01
0.20	0.60	143.10	200.00	127.36	223.96	120.46	237.81

ПРИЛОЖЕНИЕ 2. Алгоритмы вычисления среднего количества мутаций

ПРИЛОЖЕНИЕ 2-1. Структура вычислительной процедуры МП-оценок параметров гаплотипа первопредка и среднего количества мутаций

В данном параграфе приведено описание алгоритма рабочей вычислительной программы (обоснование данной процедуры было представлено в п.1.4), позволяющей на значительном наборе гаплотипов осуществить численный подбор максимально правдоподобных оценок обобщенного параметра μ, вероятностей p_1 и p_2, и начального количества повторов в маркере m_0, которое должно было быть у гаплотипа первопредка.

Поиск максимально правдоподобных оценок осуществляется путем полного перебора всех возможных значений параметров модели p_1, p_2, m_0, μ с определенным шагом по каждому из параметров.

В силу независимости мутаций в каждом маркере гаплотипа, для каждого маркера все необходимые вычисления осуществляются отдельно и независимо от результатов, полученных по иным маркерам.

2.1.1. Исходные данные для расчетов

В качестве исходных данных используются:
- Файл с параметрами управления процессом подбора МП-оценок;
- Файл с гаплотипами типа числовой матрицы, как приведенной в качестве примера во введении.

2.1.2. Перечень параметров управления

Mu (μ) - среднее количество мутаций, приходящееся на один гаплотип (в исследуемом маркере).
m0 (m_0) - количество повторов в исследуемом маркере, которое было у первопредка.

p1 (p_1) - вероятность изменения количества повторов на +1 в результате мутации (данный параметр получается как сопутствующий в процессе подбора аппроксимации);
p2 (p_2) - вероятность изменения количества повторов на +2 в результате мутации (или коэффициент K , учитывающий долю вероятности p_2 в общей вероятности увеличения количества повторов);

2.1.3. Результаты вычислений, которые используются в исторических исследованиях

Mu (μ) - среднее количество мутаций, приходящееся на один гаплотип (в исследуемом маркере).
m0 (m_0) - количество повторов в исследуемом маркере, которое было у первопредка.

2.1.4. Описание алгоритма вычислений

ШАГ 1. *Считываем из файла массив гаплотипов в виде матрицы целых чисел DATA[1:12,1:N], в которой строки – гаплотипы индивидов, а столбцы – маркеры.*
Для определенности гаплотипы 12-ти маркерные (M=12), и общее количество гаплотипов N.

ШАГ 2. В цикле *по маркерам* mrk от 1 до M делать:

ШАГ 2.1. *Вычисляем координату «центрального» столбика гистограммы данного маркера MARKER[mrk].*

ШАГ 2.2. *Задаем начальное значение максимума функции правдоподобия:*
```
max=-1000000000.0;
```

ШАГ 2.3. *Цикл по маркерам вблизи максимума гистограммы (фактически исследуем максимальный "столбик" гистограммы и соседа справа и слева), поскольку значение маркера первопредка могло сдвинуться на 1 по сравнению с современным" матожиданием:*
В цикле по m1 от -1 до +1 делать:

ШАГ 2.3.1. *Вычисляем текущий номер значения маркера, претендующего на значение у "первопредка":*
```
m := MARKER[mrk] + m1;
```

ШАГ 2.3.2. Вложенные друг в друга циклы:

1. *по процентному соотношению K вероятностей p1 и p2:*
2. *по возможным значениям обобщенного параметра Mu:*
3. *по возможным значениям вероятности p1:*

```
    Вычисляем текущие значения вероятностей
    p1,q1,p2,q2:
    p := pbeg + r*DELp;
    q := 1-p;
    p1 := p*K;
    p2 := p*(1-K);
    q1 := q*K;
    q2 := q*(1-K);

    Вычисляем вероятности из распределения Пуассона
    (коэффициенты смеси распределений) в
    соответствии с (1.2.3):
    C[j] := ((Mu)**j)/j!*exp(-Mu)

    Формируем распределения маркера для различных
    количеств мутаций:
    Задаем вероятность того, что на нулевом шаге
    (пока нет мутаций) количество повторов равно
    величине m:
    Pmrk[0,m]:=1.0;

    Рекуррентно строим спектр вероятностей для
    разного количества мутаций (j — количество
    мутаций, i - количество повторов в маркере) в
    соответствии с формулой (1.3.6):
    Pmrk[j,i+2] := Pmrk[j,i+2] + Pmrk[j-1,i]*p2;
    Pmrk[j,i+1] := Pmrk[j,i+1] + Pmrk[j-1,i]*p1;
    Pmrk[j,i]   := Pmrk[j,i]   + 0.0;
    Pmrk[j,i-1] := Pmrk[j,i-1] + Pmrk[j-1,i]*q1;
    Pmrk[j,i-2] := Pmrk[j,i-2] + Pmrk[j-1,i]*q2;

    Формируем смесь распределений маркеров в
    соответствии с формулой (1.2.2):
    P[i] := P[i] + C[j]*Pmrk[j,i]; (для всех j)

    Вычисление функции правдоподобия:
    sum=0.;
    В цикле по k от 1 до N делать:

        Формируем сумму логарифмов вероятностей в
        соответствии
        с формулой (1.4.2):
        sum := sum + log(P[DATA[mrk,k]]);
```

```
                        Проверяем условие, что новое значение
                        больше предыдущего (полученного при иных
                        соотношениях параметров):
                        Если sum>max, то
                              max = sum;
                              Запоминаем текущие параметры, при
                              которых получен новый максимум
                              логарифма функции правдоподобия:
                              Mu0[mrk] := Mu;
                              m0[mrk]  := m;
                              p0[mrk]  := p;
                              K0[mrk]  := K;
                        Конец условия sum>max
                  Конец цикла по гаплотипам.

            Конец цикла по "p1".
            Конец цикла по "Mu".
            Конец цикла по "K".
      Конец цикла по "m".

Конец цикла по маркерам "mrk".

ШАГ 3. Конец
```

ПРИЛОЖЕНИЕ 2-2. Структура вычислительной процедуры МП-оценки среднего количества мутаций для одиночного гаплотипа

В данном параграфе приведено описание алгоритма рабочей вычислительной программы (обоснование данной процедуры было представлено в п.4.1), позволяющего для отдельно взятого гаплотипа осуществить оценку среднего количества мутаций μ, осуществившихся за время, прошедшее между первопредком гаплогруппы и исследуемым гаплотипом.

2.2.1. Исходные данные для расчетов

В качестве исходных данных используются:
- Файл с параметрами управления процессом подбора МП-оценок;

- Файл с гаплотипами типа числовой матрицы, как приведенной в качестве примера во введении.
  - тестируемый гаплотип $\{m^{(1)}(1), m^{(1)}(2), \ldots, m^{(1)}(M)\}$;
  - гаплотип первопредка $\{m^{(0)}(1), m^{(0)}(2), \ldots, m^{(0)}(M)\}$ (M – количество маркеров);
  - среднее количество мутаций в маркерах, прошедшее со времени первопредка, для гаплогруппы в целом $\{\mu_1, \mu_2, \ldots, \mu_M\}$ (как уже отмечалось, вместо данного набора можно использовать набор $\{\zeta_1, \zeta_2, \ldots, \zeta_M\}$ с соответствующей корректировкой пункта 4.2 данного алгоритма);
  - набор вероятностей p_1 по маркерам для гаплогруппы в целом ;
  - $\{p_1(1), p_1(2), \ldots, p_1(M)\}$ набор вероятностей p_2 по маркерам для гаплогруппы в целом $\{p_2(1), p_2(2), \ldots, p_2(M)\}$.

2.2.2. Результаты вычислений, которые используются в исторических исследованиях

`MuRez[1:M]` (μ) – среднее количество мутаций по каждому маркеру, отделяющее исследуемый гаплотип от гаплотипа первопредка.

2.2.3. Описание алгоритма вычислений

ШАГ 1. *Считываем следующие исходные данные из файла:*

- `гаплотип первопредка  m0[1:M].`
- `среднее количество мутаций по маркерам у первопредка  Mu[1:M]`
- `вероятности p1 по маркерам  P1[1:M];`
- `вероятности p2 по маркерам  P2[1:M];`
- `Тестируемый гаплотип  m1[1:M].`

ШАГ 2. *Формируем «относительный» гаплотип как помаркерную разность повторов у гаплотипа первопредка и исследуемого гаплотипа. Чтобы не было отрицательных адресаций в массивах, добавляем константу «начальное количество повторов в маркерах» mbeg=20, одинаковую во всех маркерах.*

В цикле *по маркерам* от `mrk=1` до `M` делать:

```
      m[mrk] := m1[mrk] - m0[mrk] + mbeg;
Конец цикла
```

ШАГ 3. *Формируем таблицу вероятностей Pmrk, из которой в процессе вычисления значений функции максимального правдоподобия будут выбираться нужные значения вероятностей.*

В цикле *по маркерам* от mrk=1 до M делать:

3.1. *Формируем параметры вероятностного распределения количества повторов в текущем маркере:*

```
p := P1[mrk] + P2[mrk];
K := P1[mrk]/p;
q := 1-p;
p1 := p*K;
p2 := p*(1-K);
q1 := q*K;
q2 := q*(1-K);
```

3.2. *Проводим первичное обнуление массива, где рекуррентно будут "накапливаться" вероятности (j - количество мутаций, i - количество повторов в маркере):*

```
В цикле по мутациям от j=1 до Nc+1 делать
      В цикле по повторам от i=1 до Ni делать:
            Pmrk[mrk,j,i]=0.0;
Конец циклов по j и по i.
```

3.3. *Задаем вероятность того, что на нулевом шаге (пока нет мутаций) количество повторов строго равно величине mbeg, то есть, вероятность этого события равна единице:*

```
Pmrk[mrk,0,mbeg]=1.0;
```

3.4. *Рекуррентно строим спектр вероятностей для разного количества мутаций (j - количество мутаций, i - количество повторов в маркере):*

```
В цикле по мутациям от j=1 до Nc+1 делать
      В цикле от i=2 до Ni-2 делать:
            Pmrk[mrk,j,i+2] :=
                := Pmrk[mrk,j,i+2]+Pmrk[mrk,j-1,i]*p2;
            Pmrk[mrk,j,i+1] :=
                := Pmrk[mrk,j,i+1]+Pmrk[mrk,j-1,i]*p1;
            Pmrk[mrk,j,i]   := Pmrk[mrk,j,i]   + 0.0;
            Pmrk[mrk,j,i-1] :=
                := Pmrk[mrk,j,i-1]+Pmrk[mrk,j-1,i]*q1;
            Pmrk[mrk,j,i-2] :=
```

```
                        := Pmrk[mrk,j,i-2]+Pmrk[mrk,j-1,i]*q2;
        Конец цикла по i.
    Конец цикла по j.

    3.5. Перенормируем вероятности так, чтобы по
         «горизонтали» суммы были равными единице:
    В цикле от i=1 до Ni делать:
        sum=0.;
        В цикле по мутациям от j=1 до Nc+1 делать
              sum := sum + Pmrk[mrk,j,i];
        В цикле по мутациям от j=1 до Nc+1 делать
              Pmrk[mrk,j,i] := Pmrk[mrk,j,i]/sum;
    Конец циклов по j и по i.

Конец цикла по маркерам для формирования таблицы
            вероятностей.
```

ШАГ 4. *Поиск минимального и максимального значения Mu и вычисление количества шагов по вычислению МП-оценки:*

```
    4.1. Ищем максимальное значение Mu[mrk]:
maxmu := Mu[0];
w:=0;
В цикле по маркерам mrk от 1 до M делать:
      Если maxMu<Mu[mrk], то
            maxMu:=Mu[mrk];
            w:=mrk;
      Конец условия
Конец цикла по маркерам

    4.2. Вычисляем нормировочные коэффициенты, обеспечивающие
         вычисление количества мутаций n в каждом маркере
         пропорционально величинам Mu:
В цикле по маркерам mrk от 1 до M делать:
      DZETA[mrk] := Mu[mrk]/Mu[w];
Конец цикла по маркерам

    4.3. Задаем параметры цикла поиска МП-оценки:
Nmu   := 200;
Mubeg :=  0.1;
MAXmu := 10.0;

    4.4. Вычисляем шаг, с которым меняется Mu при поиске
         МП-оценки:
DELmu := MAXmu/Nnu;
```

ШАГ 5. *Вычисляем вероятности из распределения Пуассона (коэффициенты смеси распределений) Ppoiss(j,l) = ((Mu)**j)/j!*exp(-Mu).*

В цикле по значениям параметра Mu от l=1 до Nmu делать:

5.1. *Вычисляем текущее значение обобщенного параметра:*

```
Mu := Mubeg + l*DELmu;
```

5.2. *Вычисляем рекуррентно:*

```
Ppoiss[0,l]=1.0;
В цикле по мутациям от j=1 до Nc+1 делать
      Ppoiss[j,l] := (Ppoiss[j-1,l]*Mu)/j;
Конец цикла по мутациям
```

5.3. *Домножаем все составляющие на exp(-Mu):*

```
В цикле по мутациям от j=1 до Nc+1 делать
      Ppoiss[j,l] := Ppoiss[j,l]*exp(-Mu);
Конец цикла по мутациям
```

Конец цикла *по параметру Mu*

ШАГ 6. *Делаем начальные установки для поиска максимума функции правдоподобия.*

```
max := -100000000.0;
Mu0 := 0.0;
```

ШАГ 7. *Главная часть алгоритма поиска среднего количества мутаций, при котором достигается максимум функции правдоподобия.*

В цикле *по значениям параметра Mu* от l=1 до Nmu делать:

7.1. *Вычисляем текущее значение обобщенного параметра:*

```
Mu := Mubeg + l*DELmu;
```

7.2. *Вычисляем для каждого маркера вероятность P(m/n):*

```
В цикле по маркерам mrk от 1 до M делать:
      Если m[mrk]>Nc, то
            m[mrk]:=Nc;
      Конец условия
Конец цикла по маркерам
```

7.3. *Ищем суммарную вероятность, сложенную из условных вероятностей Pj(m/p1,n)*P(n/Lamj,T):*

```
P[mrk]:=0.;
В цикле по мутациям от n=|m[mrk]| до Nc делать:
      P[mrk] :=  P[mrk] +
            + Pmrk[mrk,n,m[mrk]+mbeg]*Ppoiss[n,l];
Конец цикла по мутациям
```

7.4. *Вычисление текущего значения функции правдоподобия:*

```
sum:=0;
В цикле по маркерам mrk от 1 до M делать:
      sum := sum + log(P[mrk]);
Конец цикла по маркерам
```

7.5. *Запоминаем промежуточное максимальное значение:*

```
Если sum > max , то
      max := sum;
      Запоминаем текущие параметры,
      при которых получен максимум:
      Mu0 := Mu;
      l0=l;
Конец условия
```

```
Конец цикла по параметру Mu
```

ШАГ 8. *Масштабируем все средние количества мутаций для проверяемого гаплотипа, поскольку поиск велся по «максимальному» количеству мутаций Mu0, соответствующему только одному из маркеров.*

```
В цикле по маркерам mrk от 1 до M делать:
      MuRez[mrk] := Mu0*DZETA[mrk];
Конец цикла по маркерам
```

ШАГ 9. Конец

Обращения читателей и персональные случаи ДНК-генеалогии

Часть 3
(начало – в Вестнике, том 1, № 2)

Анатолий Клёсов

Newton, Massachusetts 02459, U.S.A.
http://aklyosov.home.comcast.net

ВВЕДЕНИЕ

Ниже приведена серия ответов на письма людей, интересующихся ДНК-генеалогией, и являющихся носителями гаплотипов гаплогруппы R1a1. Эти письма поступили в ответ на веерную рассылку им призыва перейти на более высокой уровень тестирования гаплотипов – а именно, на 67-маркерные гаплотипы.

Этот призыв был разослан 16-17 августа 2008 года администратором Форума Rodstvo.ru тем, кто уже знает свои R1a1 гаплотипы (как правило, 12-, 25- или 37-маркерные), и, боле того, разместили их в базе данных YSearch . Цель рассылки была войти в контакт с этими людьми, вовлечь их в общение и обсуждение гаплотипов и гаплогруппы R1a1, расширить наши представления об истории гаплогруппы, и при получении дополнительных 67-маркерных гаплотипов расширить соответствующее дерево гаплотипов группы R1a1, детализировать его ветви и, возможно, выявить новые.

Цель этого проекта была направлена на лучшее понимание истории и современного состояния древнего рода R1a1 и ознакомление с выводами наших новых респондентов. А поскольку не все они определенно являются экспертами в ДНК-генеалогии вообще и в своей гаплогруппе (и в гаплотипах) в частности, то проект преследовал и образовательные цели. Типа – носители гаплогруппы R1a1, объединяйтесь на новом уровне знаний.

Пришло много ответов – частью недоуменные, частью раздраженные, частью даже несколько испуганные, типа откуда в далекой России знают мои гаплотипы, и что за этим кроется? Но таких было мало. Подавляющее

большинство ответов были благодарными, многие с явным энтузиазмом, некоторые прислали целые новеллы о своих предках и их возможных миграциях, передвижениях, судьбах.

Некоторые ответы показали, что их авторы имеют довольно смутное понятие об истории гаплогруппы R1a1, арийской гаплогруппы, и порой оперируют совершенными фантазиями, видимо, почерпнутыми из широких источников информации. Это еще раз показывает, что идея проекта была правильной.

Перед тем, как перейти к ответам на письма и вопросы читателей, кратко изложу современные представления об истории гаплогруппы R1a1. Сразу оговорюсь, что это не есть уже устоявшееся, отшлифованное знание. Такого нет ни у кого в мире. Это скорее рабочая гипотеза, но каждое положение ее подтверждено данными ДНК-генеалогии, и она в целом не противоречит известным археологическим и лингвистическим данным, скорее, их расширяет. Сама гипотеза, или скорее концепция, с соответствующими иллюстрациями на конкретных гаплотипах и их выборках подробно описана в недавней статье (Клёсов, «Откуда появились славяне и «индоевропейцы» и где их прародина? Ответ дает ДНК-генеалогия», данный Вестник, № 3, 2008).

Суть концепции следующая.

ДНК наших современников показывают, что самые древние корни ариев, рода R1a1, давностью 12 тысяч лет, находятся на Балканах – в Сербии, Косово, Боснии, Македонии. Через 6 тысяч лет этот род расширится на северо-восток, на Северные Карпаты, образовав праславянскую, трипольскую культуру и положив начало великому переселению народов в четвертом-третьем тысячелетии до нашей эры. В те же времена род R1a1 продвинулся и по южной дуге, и 5300 лет назад – по записям в наших ДНК – появился в Ливане. Прямые потомки тех первых переселенцев живут в Ливане и в наши дни. В Малой Азии, в Анатолии, предок современных носителей R1a1 жил 4500 лет назад, как и в Армении.

Как часть этого переселения, вызванного, видимо, развитием сельского хозяйства и переходом к его экстенсивным формам, этот же род R1a1 продвинулся 5200-4500 лет назад на запад, до Атлантики и Британских островов, и 4300 лет назад на север, в Скандинавию. Этот же род пришел на ближний север и восток – на земли современных Польши, Чехии, Словакии, Украины, Литвы, Белоруссии, России, с общим праславянским предком, жившим 4500 лет назад. Этот же предок дал выжившее потомство, живущее в настоящее время по всей Европе, от Исландии до Греции и Кипра, и распространившееся до юга Аравийского полуострова и Оманского залива.

Потомки того же предка, с тем же гаплотипом в ДНК, прошли до южного Урала, построили там городища 4000-3800 лет назад, одно из них (открытое в конце 1980-х годов) получило известность как Аркаим, и под именем ариев ушли в Индию, принеся туда 3500 лет назад свои праславянские гаплотипы. В том же 2-м тысячелетии до нашей эры довольно многочисленная группа рода R1a1, тоже называемая себя ариями, перешла из Средней Азии в Иран. Это – единственная, но значимая связка, позволяющая назвать весь род R1a1 родом ариев. Эта же связка приводит к тождеству «индоевропейцев», ариев, и рода R1a1 в рамках ДНК-генеалогии. Она же, эта связка, помещает прародину «индоевропейцев», ариев, праславян на Балканы. Эта же связка приводит в соответствие место балканской прародины, поток миграции ариев-праславян, динамическую цепь археологических культур и соответствующий поток индоевропейских языков, и показывает место и время появления там частицы «индо». Наконец, она же, эта связка, убедительно показывает, что не праславяне говорили на «индоиранских» языках, а наоборот, потомки праславян принесли свои арийские языки в Индию и Иран, причем времена появления этих языков в Индии и Иране, установленные лингвистами, полностью согласуются со временем прихода туда потомков праславян, временем, записанным в виде мутаций в ДНК наших современников рода R1a1.

Гаплогруппа R1b (в Европе – кельтская, баскская) образовалась 16 тысяч лет назад в Азии, продвинулась на запад, 12-14 тысяч лет назад они были на Кавказе, 6-8 тысяч лет назад на среднерусской равнине, и 4-5 тысяч лет назад стали заселять Европу, определенно в коллизиях с членами рода R1a1. Известная курганная культура, оказавшаяся на пути R1b на запад, и на пути R1a1 на восток, пока остается вне отнесений по гаплогруппам. Происхождение ее в отношении ДНК-генеалогии пока неясно. Поэтому всякие фантазии, что курганная культура – это R1a1, остаются фантазиями, как и «украинское убежище R1a1 15 тысяч лет назад», или «украинская гаплогруппа R1a1».

Переходим к письмам. Мы будем их группировать и приводить наиболее характерные вопросы и комментарии, и наши ответы.

ПИСЬМА, ВОПРОСЫ, КОММЕНТАРИИ

1. Если я сделаю себе тест на 67 гаплотипов – что с того буду иметь я, и что - ваша организация?

Ответ: Вы – узнаете, к какой ветви рода вы принадлежите, и когда, и возможно – откуда, с каких территорий ваш предок прибыл, и куда. Ветви часто соответствуют «субкладам», определенным подгруппам гаплотипа. Это делает более компактным круг ваших генеалогических родственников. Это вас не может не интересовать, иначе вы бы не делали ДНК-тест, результаты которого имеете сейчас. В конечном итоге, вы узнаете о роли ваших предков в истории человечества. Мы же получим более развитое дерево гаплотипов с лучшим, более детальным разделением по ветвям. В основании каждой ветви – один человек, предок ветви, для которого мы рассчитываем, в какие времена он жил. Это и дает нам базу для исторических исследований, миграций родов.

2. Вызван ли ваш запрос тем, что мои предки прибыли из России?

Ответ: Нет. Нас интересует весь род R1a1. По современному состоянию уровня знаний, вид гаплотипа группы R1a1 не говорит о том, откуда прибыл предок. Предковые гаплотипы R1a1 по территориям одинаковы в России, в Германии, Англии, Ирландии, Индии, Исландии, Швеции, Норвегии, Бельгии, Греции, на Крите и так далее. Это означает, что наши знания пока недостаточны, чтобы разделить гаплотипы группы R1a1по временам и территориям. Знания надо совершенствовать. Для этого, отчасти, и проект.

3. (Многие предоставили информацию, практически не имеющие отношения к истории гаплогруппы R1a1. Приводим некоторые сведения просто для демонстрации географии).

-- Бонжур, я из Франции
-- Я из штата Миннесота, США
-- Мои предки группы R1a1 прибыли в Норвегию, оттуда в Англию, и в первой половине 1600-х годов – в Новую Англию. Я живу в штате Мейн.
-- Привет из Калифорнии. Мой прадедушка приехал в Нью-Йорк из Богемии, в 1860-1890-х годах.
-- Я живу в Канаде. Мой предок родился в Германии в 1663-м году, и прибыл в Ю. Африку в 1687-м. Я знаю еще четверых его потомков, у всех нас одна и та же фамилия, трое знают свои гаплотипы, все трое имеют гаплогруппу R1a1. Было бы очень интересно увидеть наше положение на дереве.
-- Мой дядя живет в Мексике, он имеет гаплогруппу R1a1.
-- У моего брата – гаплогруппа R1a1, и мы происходим от предка, который был в составе войск, вторгнувшихся в Ирландию в 1169-м году.
-- Я живу в Норвегии. Наиболее удаленный во времени прямой предок родился в 1604 году, в Норвегии. Он был фермером. Я – член Норвежской группы, которая обменивается информацией о нордической R1a. Некоторые могут заинтересоваться участием в вашем проекте.
-- Живу в Калифорнии, предки из Украины, но больше о них я ничего не знаю.

4. Привет из штата Оклахома, США. Гаплогруппа R1a1 была принесена на Британские острова датскими и норвежскими викингами, а также германскими англами, фризами, ютами и саксонцами. Не исключено, что аланами, готами и сарматами, которые служили в римских войсках в Британии. Я лично (генеалогически) связан с династией Рюриковичей, и ДНК-тесты показали, что легендарный Рюрик имел финское происхождение.

Комментарий: Это – типичный взгляд на Западе на истоки гаплогруппы R1a1 в Британии. Но дело в том, что викинги, в том числе датские и норвежские, отмечены в истории между концом 8-го века НАШЕЙ эры и примерно серединой 11-го века, естественно, тоже нашей эры. Носители R1a1 появились на британских островах 4800-5200 лет назад. То есть о викингах в контексте появления гаплогруппы R1a1 и распространения ее по Европе 4-5 тысяч лет назад речь идти просто не может. По поводу римских войск – та же картина, это тысячелетия после заселения гаплогруппы R1a1 на территории, на которых сейчас расположены Англия, Ирландия, Шотландия.

Как отмечалось выше, истоки гаплогруппы R1a1 в Европе находятся на Балканах – в Боснии, Сербии, Косово, Македонии. Норвегия была заселена группой R1a1 относительно поздно, почти позже всех в Европе, наряду с Швецией 4200-4300 лет назад, на 600-900 лет позже, чем Британия. Викинги – это вообще другая эпоха, тысячелетиями позже. Но миф про викингов как «родоначальников» гаплогруппы R1a1 на атлантическом побережье упорно держится на Западе. При этом откуда R1a1 оказалась у викингов – обычно и не упоминается.

5. Мой дед родился в Англии. Профессор Оппенгеймер определил меня в гаплогруппу R1a1-2b, которая, как он полагает, появилась в Северной Норвегии 5000 лет назад. Мне очень интересно распределение гаплогруппы R1a1 в России, и хотелось бы выяснить, насколько близки гаплотипы R1a1 из России моему гаплотипу. Мой гаплотип показывает сходство с гаплотипами из Шотландии, Исландии, Норвегии и Швеции.

Комментарий: Полагаю, что оценка проф. Оппенгеймера неверна. Самый ранний предок гаплогруппы R1a1 в Норвегии датируется 4300 лет назад. Предковый гаплотип – следующий:

13 25 16 11 11 14 12 12 10 13 11 30 15 9 10 11 11 24 14 20 32 12 15 15 16

Здесь имеется только одно отклонение, на четвертом маркере, по сравнению с предковым российским гаплотипом R1a1, то есть предковым гаплотипом восточных славян (выделено жирным шрифтом):

13 25 16 **10** 11 14 12 12 10 13 11 30 15 9 10 11 11 24 14 20 32 12 15 15 16

Да и это отклонение только частичное, только на треть мутации. У славян там на самом деле величина 10.46, у норвежцев 10.76. А у индусов, например, 10.53. Так что гаплотип фактически тот же самый.

Предковый гаплотип англичан R1a1 – тоже очень близок:

13 25 **15 10** 11 14 12 12 10 13 11 30 15 9 10 11 11 24 14 20 32 12 15 15 16

Он даже более близок к славянам, чем к норвежцам – с первыми одна мутация, со вторыми – две мутации разницы. Но и эти мутации только частичные – у славян в третьем маркере слева 15.77, у англичан 15.42. Так что опять только треть единицы из 25 маркеров.

Иначе говоря, желание коллеги «Мне ... хотелось бы выяснить, насколько близки гаплотипы R1a1 из России моему гаплотипу. Мой гаплотип показывает сходство с гаплотипами из Шотландии, Исландии, Норвегии и Швеции» в настоящее время имеет простой ответ: все предковые гаплотипы R1a1 в этих странах, включая Англию, практически одинаковые, а гаплотипы современников в тех же странах различаются на нормальные статистические мутации, и перекрывают друг друга по всему диапазону. Так что и здесь проф. Оппенгеймер, судя по всему, ошибся.

6. Я в некотором замешательстве, так как у моего отца гаплогруппа R1a, а не R1a1.

Комментарий: Это одно и то же. Гаплогруппу R1a тестируют на снип SRY 10831.2, гаплогруппу R1a1 – на снипы M17 или M198. Они все сопровождают друг друга, если определяются.

7. Я знаю свой гаплотип, но не знаю, к какой ветви принадлежу. Славянской или арийской? В базе данных есть два похожих на мой гаплотипа – один из России, один из Польши.

Комментарий: Славянский и арийский - это одно и то же, с точки зрения рода. Оба – род R1a1. Просто названия относятся к разным эпохам, и обычно употребляются в разных контекстах.

8. Как вы пришли к выводам о предке гаплогруппы R1a1?

Комментарий: Собрали все доступные выборки гаплотипов R1a1 по всем странам, от Исландии до Греции и от Атлантики до Индии, а также Анатолии, Крита, Сардинии, Сицилии, Ливана, Аравийского полуострова, и всех остальных стран и территорий. Все это есть либо в базе данных YSearch, либо в научных публикациях. Таким образом, «картировали» весь мир по гаплогруппе R1a1, вычислили для каждой территории предковые 25-маркерные (где было возможно) гаплотипы, определили число мутаций от предкового гаплотипа для каждой территории, и рассчитали возраст общего предка с введением поправки на возвратные мутации. Самый старый предок оказался на Балканах (Сербия, Косово, Босния, Македония), 12200 лет назад. Относительно старые предки – 6000-5000 лет назад – в Буковине и Венгрии. Атлантика – 5200-4800 лет назад. Западная и Центральная Европа – 4400-4800 лет назад. Скандинавия – 4200-4300 лет назад. Россия-Украина – 4500 лет назад. Индия – 3800 лет назад.

9. Наконец-то я имею контакт с русским, который имеет такую же курганскую хромосому, как и я. В отличие от большинства курганцев, которые отправились на восток, мой предок направился в Скандинавию и присоединился к культуре викингов. Потом они отправились в Нормандию и вторглись в Англию под началом Вильяма-Завоевателя. Мой самый отдаленный предок, которого я выявил, родился в 1690-х годах. Приятно встретить российского родственника.

Комментарий: приведенные представления о курганской R1a1 пока практически ни на чем не основаны, как и то, что якобы курганские R1a1 из далеких степей Прикаспия отправились в Скандинавию. Гаплотипы курганной культуры пока ни в одном случае не идентифицированы. Они могут вполне оказаться и гаплогруппой R1b, кельтской или баскской. Гораздо проще и вероятнее, что R1a1 в Скандинавии оказались тем же путем, что и в Германии, Англии, Польше, Чехии, и вообще по всей Европе – а именно, продвижением с Балкан 5 тысяч лет назад.

Если гаплогруппа представителей курганной культуры окажется R1a1, то это скорее всего промежуточный пункт на миграционном пути ариев с запада на восток, предвестье андроновской культуры северного Казахстана, южного Урала, арийских культур Средней Азии и затем Ирана.

10. Мои предки в Англии живут примерно с 1200-х годов, и мои гаплотипы сходны с гаплотипами из Литвы и северной Германии. Это дает хорошую связь с курганной культурой.

К**омментарий**: как видно, фантазии про курганную культуру как источник R1a1 глубоко укоренились в Западной Европе. И что характерно – без каких-либо доказательств, свидетельств, гаплотипов или гаплогрупп. Просто так. Не исключено, что это окажется так, но не в меньшей мере не исключено, что так не окажется. Гаплогруппа R1b на своем пути из Азии по южным степям на Запад – совсем не худший кандидат на курганную культуру.

Наконец, сходство английских гаплотипов с литовскими и немецкими – совершенно естественно без всякой курганной культуры. Никакой «связи с курганной культурой» это не дает. У них один источник – балканские гаплотипы R1a1. У курганной культуры, повторяю, вообще нет связи с ДНК-генеалогией. Пока, во всяком случае.

11. Я живу в Сан Франциско, мне 82 года. Мои предки, по-видимому, из Шотландии. Я знаю, что R1a1 редки в Великобритании, и более часты в Скандинавии, Польше, Германии. Буду признателен за любую информацию в этом отношении.

Комментарий: Действительно, доля носителей гаплогруппы R1a1 на Британских островах составляет всего от 2% до 9%. Там полностью доминируют R1b (71%) и I1 (16%). В Ирландии представителей R1a1 не более 2-4%, в Шотландии от 27% (на севере, на Шетландских островах) до 2-5% на юге страны. В Норвегии R1a1 примерно 18-25% населения, в Швеции примерно столько же – 17%. В Германии столько же, в среднем 18%, но в некоторых районах достигает трети. В Чехословакии R1a1 в среднем треть, по районам от 29% до 41%. В Польше – в среднем 57%, в некоторых районах до 64%. В России в среднем 48%, в некоторых районах до 75%. Но предковые гаплотипы, как уже отмечалось, везде практически одинаковы, и ведут к общему для всех этих стран предку примерно 4500 лет назад.

12. Я живу в Южной Африке, в Пеории, и являюсь потомком голландцев, немцев и французов. Пытаюсь найти корни своей семьи, и написал об этом несколько статей в журналах. Хотел бы написать и книгу, так что информация, полученная от вас, была бы очень кстати.

Нас – пять родственников, все выходцы из Южной Африки, все гаплогруппы R1a1, и гаплотипы всех пятерых совпадают в 12-, 25 и 31-маркерном формате. Мы уже установили, что все мы – потомки одного и того же человека, Хендрика Вентера, который родился в 1663 году в Германии или во Франции, и прибыл в Южную Африку в 1687 году. В базе данных есть 135 человек, с которыми совпадают в некоторой степени наши гаплотипы, и большинство их – из вашей части мира, или оттуда прибыли их предки. Мой (ныне покойный) отец тоже говорил мне, что наши предки прибыли из Польши и России. Я не

знаю, откуда эти сведения. Поэтому ваше письмо – приятный сюрприз. Будем на связи.

Комментарий: Если общий предок жил 13 поколений назад, то не удивительно, что гаплотипы у пятерых потомков совпадают. Теоретически возможна одна мутация на пятерых 12-маркерных гаплотипах, две - на 25 маркерных гаплотипах, и три – на пятерых 37-маркерных гаплотипах – среди 185 аллелей. Но мутаций может и не быть, чисто статистически. То, что гаплотипы в какой-то степени совпадают с сотнями других – тоже неудивительно, так как предковые гаплотипы гаплогруппы R1a1 практически одинаковы по всей Европе, и гаплотипы потомков образуют одно и тоже облако гаплотипов и их мутаций. Все они перекрываются.

Более того, поскольку Польша и Россия имеют большинство представителей гаплогруппы R1a1 в Европе (примерно 12 млн в Польше, 11 млн на Украине и 35 млн в России; для сравнения – примерно 100 млн в Индии), то вероятность для предка R1a1 иметь происхождение в этих странах довольно высока.

13. У меня гаплогруппа R1a1, но маркер DYS426 имеет редкое значение аллели = 13.

Комментарий: Действительно, такое значение аллели редкое. До последнего времени это значение встречалось лишь у двоих в Германии (из 67 тестированных на R1a1), у двоих в Индии (из 84 человек), у одного на о. Крит, и у двоих по всей остальной Европе – это если считать тех, которые были тестированы на 25-маркерные гаплотипы. У остальных этот маркер дает значение аллели = 12.

14. Я из юго-западной части Индии, штат Керала. Тест ДНК я сделал, чтобы узнать, откуда прибыли мои предки. Мои родители, и предки много поколений жили в штате Керала, будучи христианами святого Фомы. Моя гаплогруппа R1a. Я полагаю, что мои предки, евреи-ашкеназы, прибыли с Ближнего Востока, и не были обращены в иудаизм из местных индусов. Поэтому мне уже не нужно ничего исследовать.

Комментарий: Это очень интересное письмо. Действительно, сирийские Малабар Насрани, исповедующие христианство святого Фомы – это отдельная этно-религиозная группа в штате Керала на юго-западе Индии, побережье Малабар. По историческим сведениям, евреи прибыли туда в 6-м веке до н.э., после разгрома Иудеи. Христианство пришло туда с прибытием святого Фомы («Фомы неверующего»), одного из двенадцати апостолов Христа, в 52-м году. Он и приехал в Кералу главным образом обращать в христианство евреев, прибывших туда ранее. Еще одна волна евреев прибыла в Кералу в 4-м веке

нашей эры. Есть версия, что само имя Насрани – это вариант слова Назарет, или назаретяне, и тоже было принесено евреями. Сама Керала упоминается в Махабхарате и как племя, и как королевство.

В итоге в Керале образовался уникальный сплав «иудейско-сирийского христианства», индуистских обрядов и правил, южно-индийской культуры и малазийского языка. Себя они обыкновенно называют «сирийскими христианами», как производное от «сирийского» (арамейского) богослужения, введенного в первом столетии нашей эры. В христианство в те далекие времена были обращены как евреи, так и местные индийцы.

Именно в связи с этим и интересно письмо, цитируемое выше. Автору, конечно, виднее, если он считает себя евреем-ашкенази, хотя это сочетание для Индии весьма своеобразно. Ашкенази – это в буквальном переводе «немецкие евреи», и сложились они в Европе в средние века, то есть во времена намного более поздние, чем начало нашей эры, тем более в Индии. Далее, евреи, прибывшие с Ближнего Востока, а точнее, из Иудеи, вряд ли имели гаплогруппу R1a1. Это были скорее гаплогруппы J1, J2, возможно, E. R1a1 – это с куда большей вероятностью была гаплогруппа, занесенная В Индию ариями с севера за тысячу-полторы лет до прибытия туда евреев. Например, южно-индийское племя Ченчу имеет в части своей носителей гаплогруппы R1a1 с общим предком в племени 2900 лет назад.

Так что наш коллега из Кералы, скорее всего, неправ, когда пишет, что «...полагаю, что мои предки, евреи-ашкеназы, прибыли с Ближнего Востока, и не были обращены в иудаизм из местных индусов. Поэтому мне уже не нужно ничего исследовать». Думаю, исследования для него только начинаются. Рискну предположить, что корни его приобретены от ариев, потомков праславян, пришедших с южного Урала. Эта гаплогруппа продолжила путь из Кашмира на юг Индии, в Кералу, и через полторы тысячи лет после прибытия ариев в Индию далекие индийские предки нашего коллеги были обращены в христианство, возможно, самим апостолом Фомой. Или после него, уже обращенными в христианство прихожанами. А элементы еврейской религии – это уже приобретенное. Предков с Ближнего Востока по мужской линии у нашего коллеги, видимо, не было.

15. Является ли ваш проект частной инициативой, или он связан с правительственными или религиозными организациями, университетами, и прочими организациями?

Ответ: это исключительно частный проект. Он направлен на расширение наших познаний в ДНК-генеалогии вообще, и истории гаплогруппы R1a1 в

частности. Проект никем и никакой организацией не финансируется, и является совершенно добровольным и некоммерческим.

Всего наилучшего.

Anatole Klyosov
Newton/Boston