Для чего служит анализ ADMIXTURE и как он работает

Елена Лукьянова, биоинформатик

ADMIXTURE (буквально: примесь) – это компьютерная программа (анализ), позволяющая выявлять смешанность состава некоего набора индивидов на основе данных о генотипах и тем самым строить предположения о происхождении популяции.

Принцип работы ADMIXTURE.

Рассмотрим принцип работы ADMIXTURE на примере образцов и популяций из проекта HapMap.

Всего у нас N = 324 образца/индивида, каждый из которых относится к одной из четырех нижеперечисленных популяций:

1. АФРИКА (ASW) – Африканские предки из Юго-Западной части США
2. ЮТА (CEU) – жители штата Юта США с корнями из Северной и Западной Европы
3. МЕКСИКА (MEX) – Мексиканцы, Лонг-Айленд США
4. ЙОРУБА (URI) – Йоруба, Нигерия

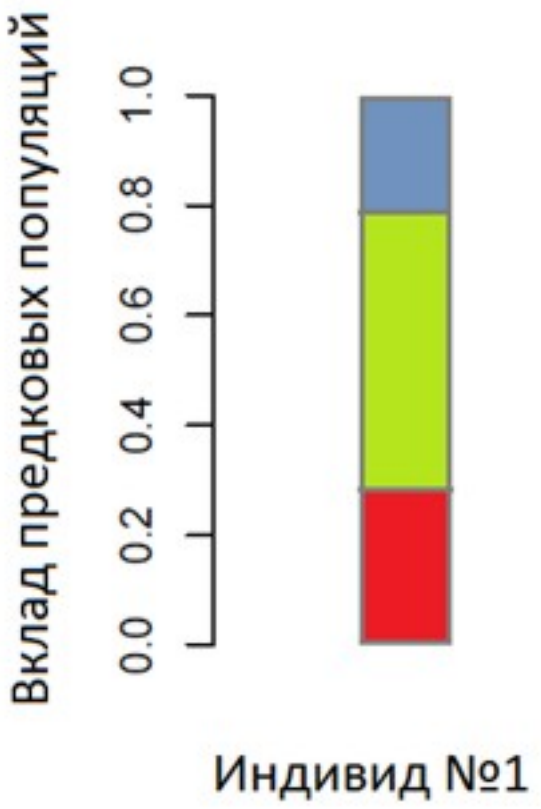
Для удобства дальнейшего изложения будем называть эти популяции «известными».

Также мы предполагаем, что они произошли от К разных предковых популяций (мы не знаем от каких именно). В дальнейшем будем называть эти предковые популяции «предполагаемыми предковыми». Этих «предполагаемых предковых» популяций на самом деле не существует, у них нет общепризнанных названий и характеристик. И на этом этапе мы даже не знаем какие образцы к какой из этих К популяций могут быть отнесены. Теоретически возможно, что образцы из одной и той же «известной» популяции могут принадлежать к двум разным «предполагаемым предковым» популяциям.

Пример 1.

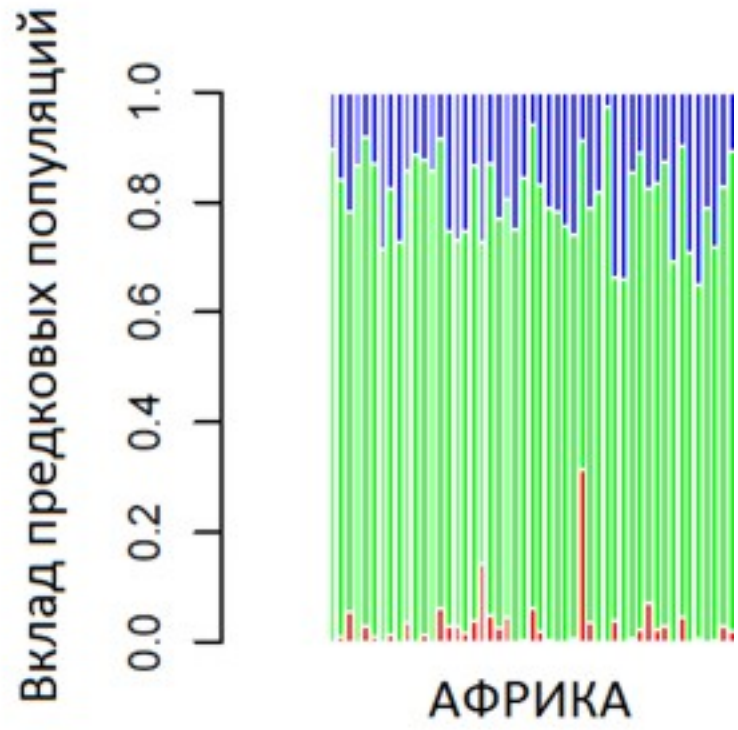
Предположим, что К = 3.

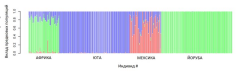
ADMIXTURE далее работает с образцами (их генотипами) и заданным нами числом К = 3. Имея сведения о генотипах и предположение о количестве «предполагаемых предковых» популяций (К) ADMIXTURE строит свою модель (предположение) того, каков вклад каждой из «предполагаемых предковых» популяций в каждый индивид. В результате мы имеем для каждого индивида 3 цифры: количественный вклад каждой из трех популяций (или образно говоря, на сколько процентов данный индивид состоит из первой «предполагаемой предковой» популяции, на сколько – из второй и на сколько – из третьей). При этом может быть и такая ситуация, что у конкретного индивида в составе отсутствует какая-то из «предполагаемых предковых» популяций, даже возможно, что он принадлежит только к одной из «предполагаемых предковых» поуляций. Предположим, для индивида №1 эти цифры такие: 0.3, 0.5 и 0.2. Что эти цифры означают? Означают они доли каждой из «предполагаемых предковых» популяций (ППП) в индивиде №1, т.е. индивид состоит на 30% из первой ППП, на 50% — из второй и 20% — из третьей. Чем больше вклад каждой ППП в индивида, тем больше индивид является «носителем» данной популяции и ее представителем. Эти цифры можно визуализировать (bar plot), например, так:



При этом мы сами можем задать любой цвет для каждой из трех «предполагаемых предковых» популяций. На рисунке выше мы выбрали: красный, зеленый и синий.

Далее мы группируем этих индивидов по «известным» популяциям и получаем генетические профили каждой из популяций. Например, для Африканской популяции:



А после этого мы можем сравнивать известные популяции по этим генетическим профилям между собой: 

Важен как количественный состав, так и качественный, т.е. какие «предполагаемые предковые» популяции (из трех – желтой, зеленой и красной) присутствуют в индивидах и в каком соотношении. Обычно сравнение происходит «на глаз», чисто визуально. Одним исследователям может показаться, что профили похожи, другим может показаться, что профили совершенно различны.

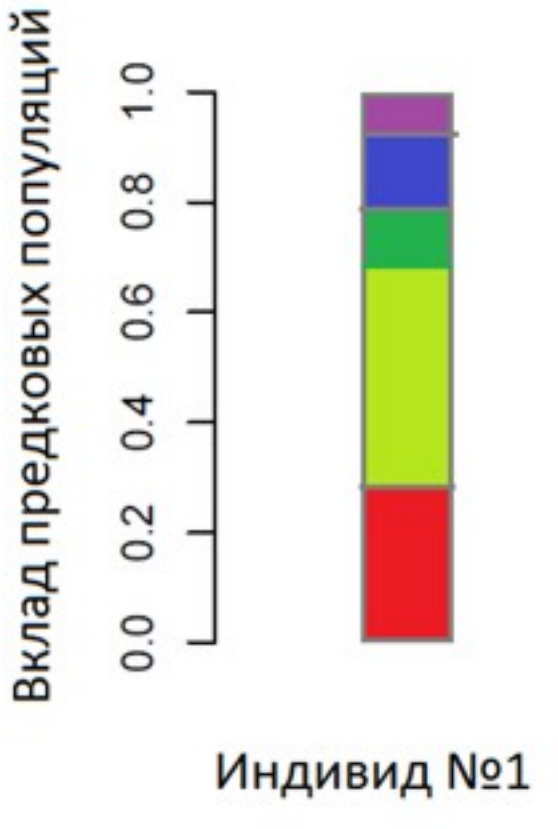
Пример 2.

Теперь рассмотрим ситуацию, когда число «предполагаемых предковых» популяций K = 5. Мы можем варьировать это значение сколько угодно.

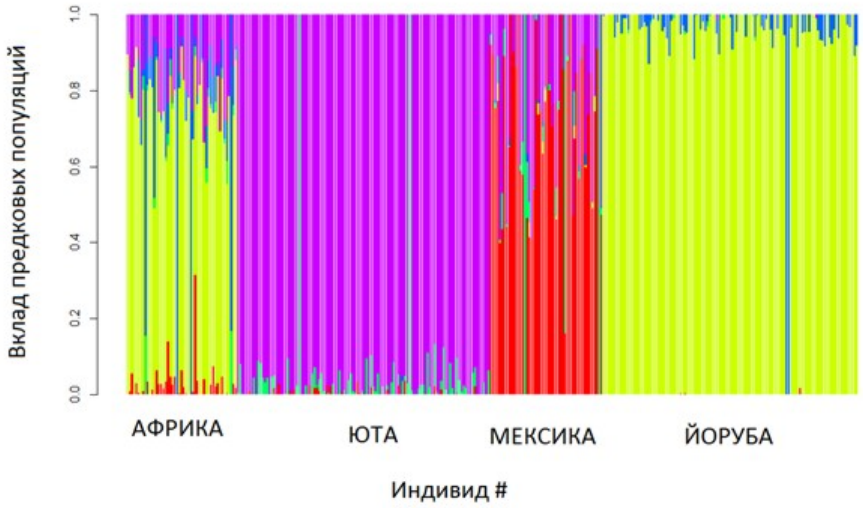
В этом случае у нас вместо 3 уже 5 разных «предполагаемых предковых» популяций, для каждой из которых мы задали свои 5 разных цветов – красный, салатовый, зеленый, синий и фиолетовый. Цвета выбираются произвольно и «предполагаемая предковая» популяция из первого примера, которую мы окрасили в красный цвет, не имеет ничего общего с нынешней «предполагаемой предковой» популяцией того же цвета. Данный цвет актуален только для конкретно этой модели.

Но, если рассуждать в рамках одной модели, то если какой-то индивид не показывает смешанности и весь принадлежит одному цвету, в принципе можно считать этот цвет цветом его этноса. Похожие заключения можно сделать и для всей популяции, если все индивиды данной популяции принадлежат одному цвету. Хотя эти все эти определения имеют смысл только в том случае, если мы действительно построили хорошую модель, подобрали оптимальное значение K (об этом см. подробное описание ниже в разделе «Подбор оптимального значения K»).

Теперь мы получим для каждого индивида 5 разных цифр – вклад каждой из 5 «предполагаемых предковых» популяций в данный образец. И генетический профиль для одного индивида будет выглядеть, например, таким образом:



Далее индивиды, как и в предыдущем примере, группируются по «известным» популяциям, после чего сравниваются профили «известных» популяций между собой:



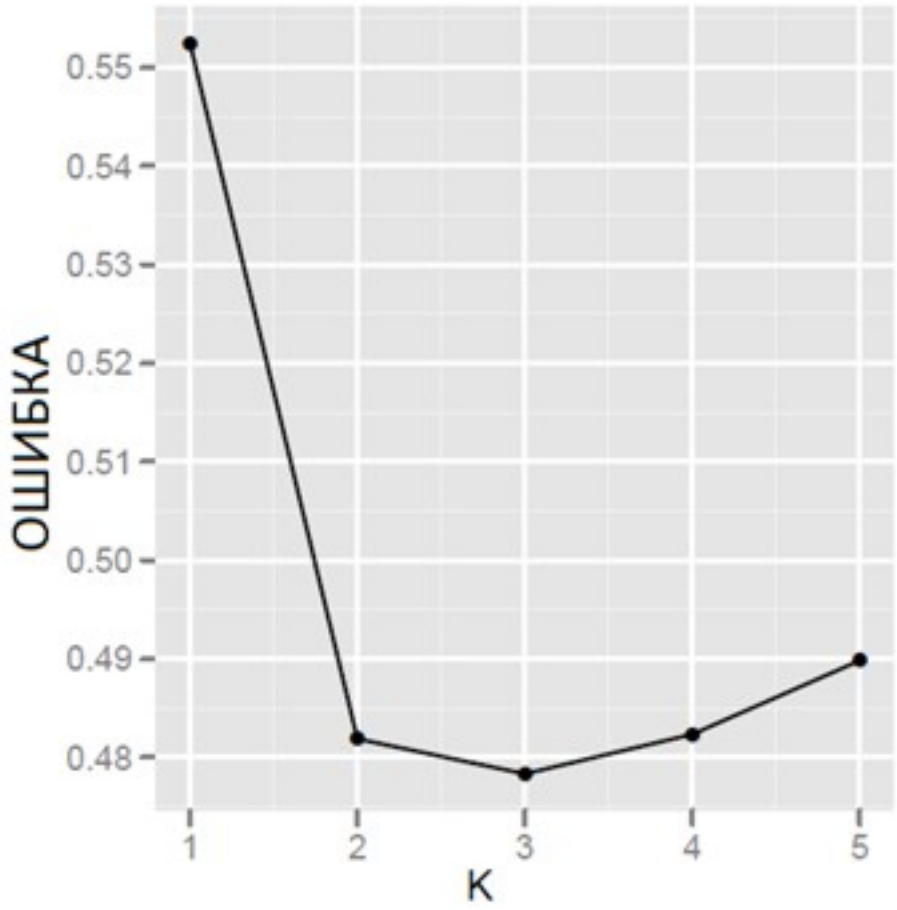
<u>Подбор оптимального значения K.</u>

Следующий важный вопрос, на который необходимо ответить: «Какое из наших двух предположений о количестве «предполагаемых предковых» популяций наиболее соответствует действительности? Что же выбрать в качестве итогового результата для статьи? K = 3 или K= 5? А может, K = 10? Как узнать какое K для этих данных было бы оптимальным?».

Оптимальное число K «предполагаемых предковых» популяций может быть подобрано эмпирическим путем. Каждый раз, когда ADMIXTURE строит свою модель (некое «предположение») о структуре «предполагаемых предковых» популяциц на основании данных о генотипах индивидов (т.е. о том, каков вклад каждой из «предполагаемых предковых» популяций в каждый индивид) для заданного нами числа K, она в конце проводит сравнение своей модели с реальными данными, а именно, проверяет, насколько хорошо входные данные описываются построенной моделью. Предположим, что мы подали на вход наши данные из проекта HapMap и задали число K = 1. Конечно, ADMIXTURE построит свою модель исходя из того предположения, что все индивиды произошли из одной предковой популяции, но будет ли это соответствовать столь сильно различающимся между собой по факту генотипам наших индивидов? И насколько лучше наши данные описываются моделью,

построенной на предположении о трех предковых популяциях? А пяти? Поэтому в дополнение к результатам о вкладе каждой из «предполагаемых предковых» популяций в каждый индивид мы получаем еще одну величину – ошибку – т.е. некую характеристику несоответствия построенной модели и реальных данных. Чем больше значение этой величины, тем хуже наше предположение о количестве предковых популяций. Стоит помнить о том, что идеала мы не добьемся и нулевой ошибки, соответственно, тоже.

Но как же выбрать тогда оптимальное значение K? А вот как: построим для каждого K из интервала от 1 до 5 свою модель (запустим программу ADMIXTURE 5 раз для одних и тех же индивидов но с пятью разными K) и получим для каждого K свое значение ошибки. Отобразим эти значения на графике:



Мы видим, что минимальная ошибка при K = 3, а дальше, при K = 4 и K = 5, ошибка начинает расти, значит оптимальное значение K – 3.

Если на нашем графике мы не видим минимума (например, с каждым следующим значением K ошибка уменьшается), то нужно строить модель и выбирать новое K до тех пор, пока ошибка не начнет увеличиваться.

Корректность выдаваемых результатов при оптимально подобранном числе K.

Стоит отметить, что результаты данного анализа очень сильно ограничены свойствами входного набора индивидов:

1. Образцы должны быть между собой неродственны.
2. Участки однонуклеотидного полиморфизма (SNP), по которым производится генотипирование образцов, должны быть равномерно распределены по геному с достаточно высокой плотностью
3. Аллели SNP находятся в равновесном сцеплении, это значит, что появление данного аллеля у конкретного индивида зависит только от его частоты в популяции, но не зависит от других аллелей.

Возможна ситуация, когда аллели SNP не находятся в равновесном сцепленни на самом деле, но только «выглядят таким образом». В этом случае результаты анализа будут некорректны.

Таким образом, от одного набора данных до другого набора данных результаты теста могут очень сильно варьировать. Аккуратность и корректность данного анализа в настоящий момент еще не достигла уровня, на котором можно использовать его как «доказательство» своей гипотезы. Но использовать вкупе с остальными анализами для получения более подробной информации о наборе входных данных (индивидов) – можно.

Степень близости между реальностью и построенной моделью может быть такой же, как и степень близости между «сферическим конем в вакууме» и реальной лошадью, а может быть и такой же, как и между двумя реальными лошадьми. Скажем так, результаты ADMIXTURE стоит принимать во внимание, но на них нельзя рассчитывать.