Каталог геномного разнообразия планеты

Надежда Маркина

Результаты финальной стадии проекта «1000 геномов»

Представлены итоги проекта «1000 геномов», который продолжался 27 лет. Секвенированы геномы и экзомы для 2504 индивидов из 26 популяций пяти регионов (полные геномы с низким покрытием, экзомы – с высоким покрытием). Описано свыше 88 млн генетических вариаций, в том числе структурные вариации, затрагивающие большие участки ДНК. Дана характеристика изученных популяций в целом, для отдельных регионов, групп популяций и популяций, в том числе и генетических вариаций, ассоциированных с заболеваниями. Создана модель реконструкции демографической истории изученных популяций и найдены новые мишени положительного естественного отбора.

В журнале Nature месяц назад были опубликованы результаты финальной стадии проекта «1000 геномов» — глобального исследования геномного разнообразия методом полногеномного секвенирования. Заявленная в названии задача – секвенировать 1000 геномов – перевыполнена, так как усилиями международного консорциума (The 1000 Genomes Project Consortium) исследованы 2504 геномов из 26 популяций пяти регионов: Африки, Европы, Южной Азии, Восточной Азии и Америки. Эту работу возглавили специалисты Европейской молекулярно-биологической лаборатории (European Molecular Biology Laboratory (EMBL) и Вашингтонского университета. Проект «1000 геномов» начался в 2008 г. результаты его промежуточных стадий публиковались постепенно, а финальные результаты представлены в двух статьях в Nature.

Проект «1000 геномов» направлен на оценку геномного разнообразия в глобальном масштабе — выявление того, как разного рода генетические вариации распределены по миру (в первом приближении). Но кроме фундаментальной научной задачи ставилась и прикладная – биомедицинская: оценить, как распределены по популяциям те генетические вариации, которые могут быть связаны с заболеваниями.

Все 2504 геномов были секвенированы с низким покрытием — в среднем 7.4x. Это число показывает, сколько раз в среднем был прочитан каждый нуклеотид и оценивает надежность прочтения генома. В настоящее время стандартом является покрытие 60x- 70x для современных геномов (для древней ДНК требования ниже из-за ее плохой сохранности). Однако экзомы (кодирующие белки части генома) в проекте «1000 геномов» были секвенированы с высоким покрытием (в среднем 65.7x).

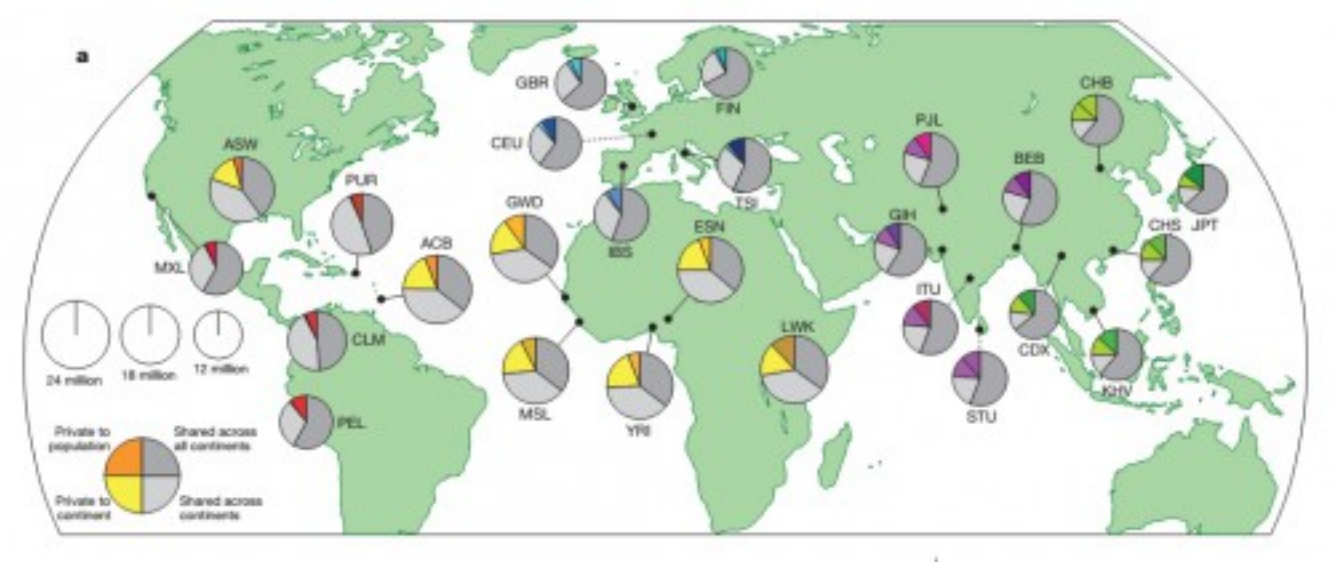
У всех индивидов (а, по возможности, и у их родственников в первом поколении) были изучены генетические вариации в масштабах всего генома. В общей сложности описано свыше 88 млн генетических вариаций: 84,7 млн точек однонуклеотидного полиморфизма (SNP) – единичных замен нуклеотидов; 3,6 млн вставок/выпадений фрагментов ДНК (инсерции/делеции, которые объединяют под общим название Indels); 60 тыс. структурных вариаций (SV) – это более крупные изменения в строении ДНК, затрагивающие большое число нуклеотидов.

О полноте полученных данных говорит такая цифра: авторы указывают, что охвачено 99% SNP и 85% Indels, встречающихся с частотой >1%, а для более редких вариаций, встречающихся с частотой не меньше 0,5%, эти цифры составляют 95% и 80% соответственно. То есть, создан первый вариант «каталога» генетического разнообразия человека. При этом найдено большое число новых генетических вариаций. Например, в изученных популяциях Южной Азии они составили 24% от всех вариаций, в Африке – 28%.

Первая статья сфокусирована на коротких изменчивых фрагментах ДНК, длина которых не больше 500 пар оснований. Изучив разнообразие таких фрагментов, авторы определили, что «типичный» геном отличается от референсного генома (избранная стандартная последовательность, используемая для сравнения) по 4,1 млн — 5,0 млн сайтов. Более 99,9% этих отличий представлены однонуклеотидным полиморфизмом (SNP) и короткими вставками-выпадениями (Indels). В качестве референсного генома исследователи использовали геном GRCh38, который был предложен Консорциумом генома человека (Genome Reference Consortium) в декабре 2013 года.

Во второй статье анализируются оставшиеся 0.01% отличий от референсного генома – но это уже крупные структурные изменения (SV), включающие до полумиллиона пар нуклеотидов. Именно в этом проекте они были изучены наиболее полно благодаря технологиям полногеномного секвенирования, что позволило выдвинуть гипотезы о их связи с развитием заболеваний. Типичный геном содержит от 2100 до 2500 структурных вариаций (SV): примерно 1000 крупных делеций, 160 вариаций числа копий, 915 Alu вставок, 128 L1 вставок, 51 SVA вставок, 4 NUMTs и около 10 инверсий (поворот фрагмента на 180 градусов). Суммарно они затрагивают около 20 млн пар оснований.

Разнообразие геномов – какое и где



На карте указаны изученные популяции. Каждая круговая диаграмма обозначает внутрипопуляционное разнообразие. Диаграмма поделена на четыре сектора: ярким цветом обозначены генетические вариации, уникальные для популяции, бледным цветом – вариации, общие для данной группы популяций, светло серым – вариации, общие для данного континента, и темно серым – вариации, общие для всех континентов. Размер кружка указывает на размер популяции (см. легенду).

Данные по 26 популяциям, из которых были получены образцы, представлены на сайте проекта http://www.1000genomes.org/about . Вот список популяций и их условные обозначения на рисунках:

CDX – китайцы Сишуанбаньна-Дайский автономный округ

CHB – китайцы , Пекин

CHS – южные китайцы

JPT – японцы, Токио

KHV – вьетнамцы

BEB – бенгальцы, Бангладеш

GIH – гуджаратцы (индоарийский народ), Хьюстон

ITU – индийцы телугу, Великобритания

PJL – панджаби, Пакистан

STU – ланкийские тамилы, Шри-Ланка

ASW – афроамериканцы, юго-запад США

ACB – вест-индское негритянское население, о-в Барбадос

ESN – ишаны, Нигерия

GWD – население Гамбии

LWK – лухья, Кения

MSL – менде, Сьерра-Леоне

YRI – йоруба, Нигерия

GBR – британцы, Англия и Шотландия

FIN – финны, Финляндия

IBS – иберийцы, Испания

TSI – тосканцы, Италия

CEU – американцы северо- и западноевропейского происхождения, штат Юта США

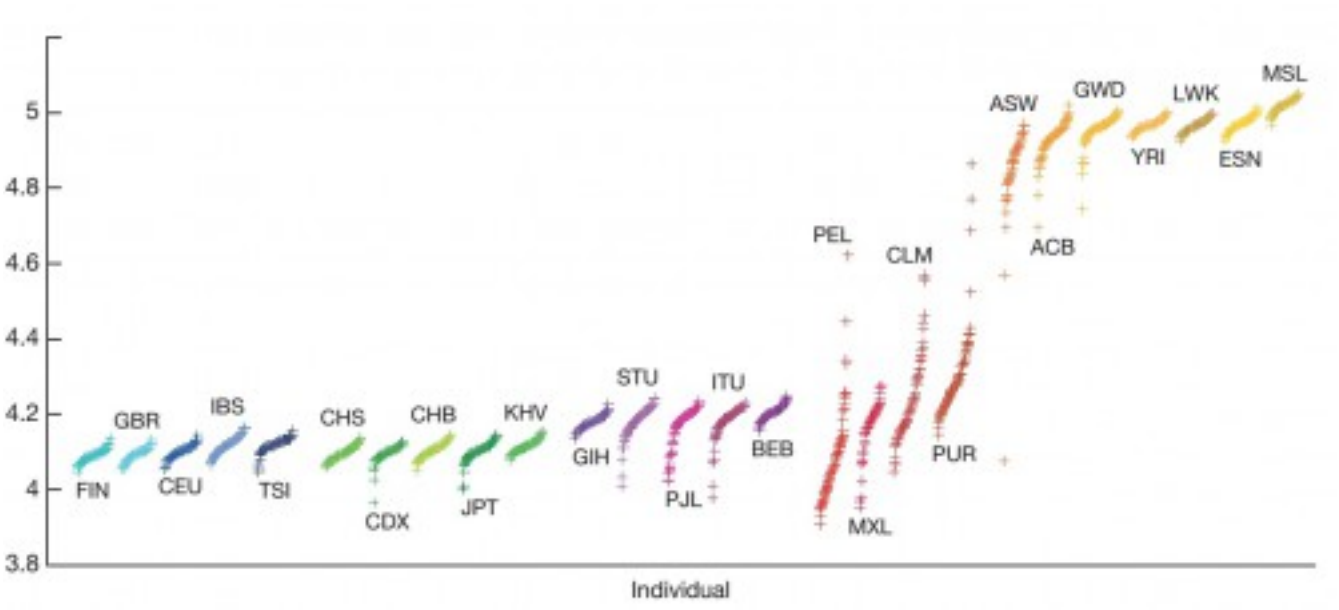
CLM – колумбийцы, Меделлин, Колумбия

MXL – американцы мексиканского происхождения, Лос-Анджелес, Калифорния, США

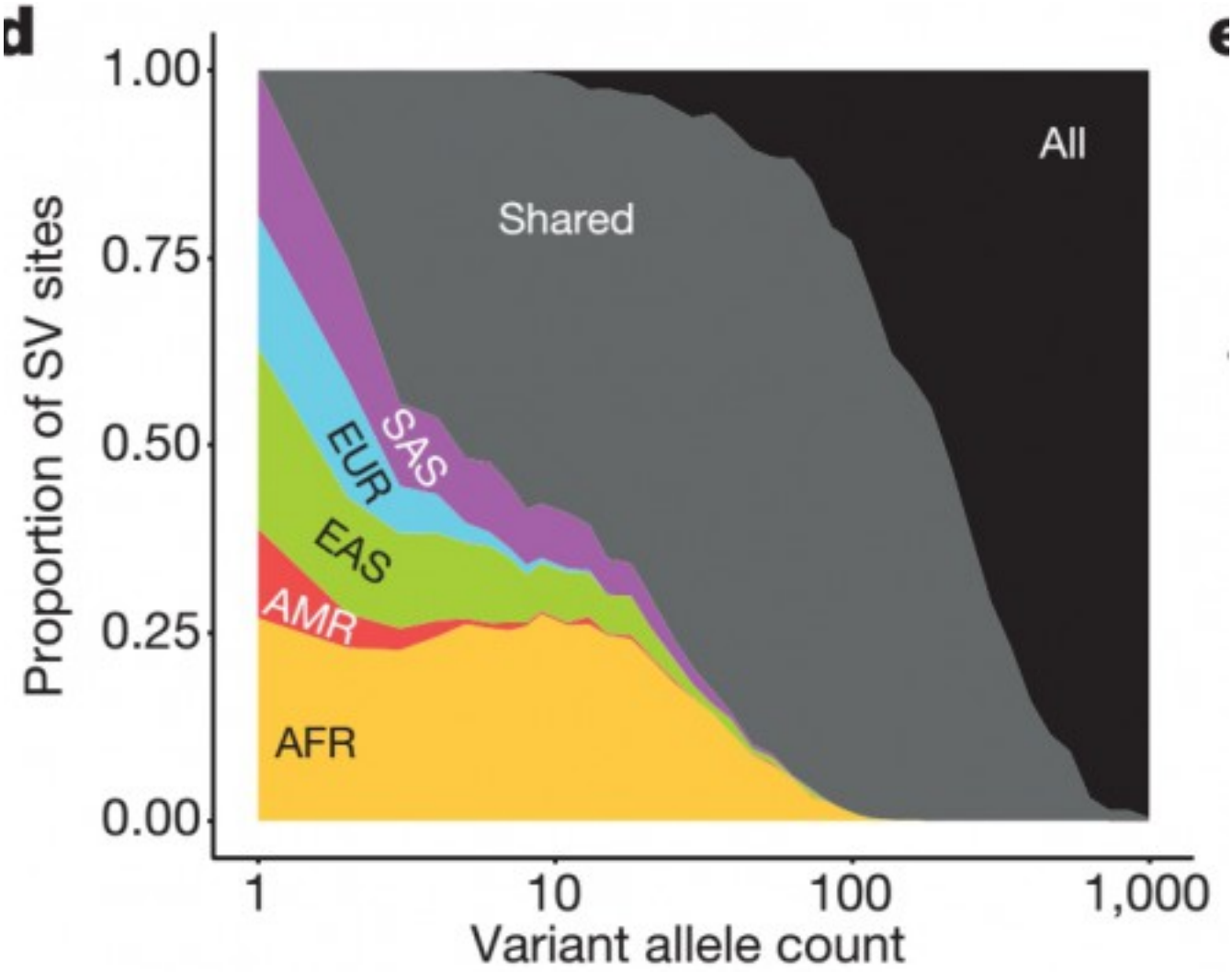
PEL – перуанцы, Лима, Перу

PUR – пуэрториканцы, Пуэрто-Рико

Общее число сайтов, отличных от референсного генома, неодинаково у разных популяций. Больше всего их обнаружено в геномах из африканских популяций. Это предсказуемо, исходя из того, что все современное человечество вышло из Африки – наибольшее разнообразие обычно наблюдается в месте происхождения. Более того, в других популяциях число сайтов, отличных от стандартного генома, по грубым оценкам, пропорционально вкладу африканских предков в геном.



Количество вариабельных сайтов (по оси Y) на один геном в разных популяциях. Регионы обозначены разным цветом: голубой-синий – Европа, зеленый – Восточная Азия, фиолетовый-розовый – Южная Азия, красный – Америка, желтый-оранжевый – Африка. Обозначения популяций см. выше.

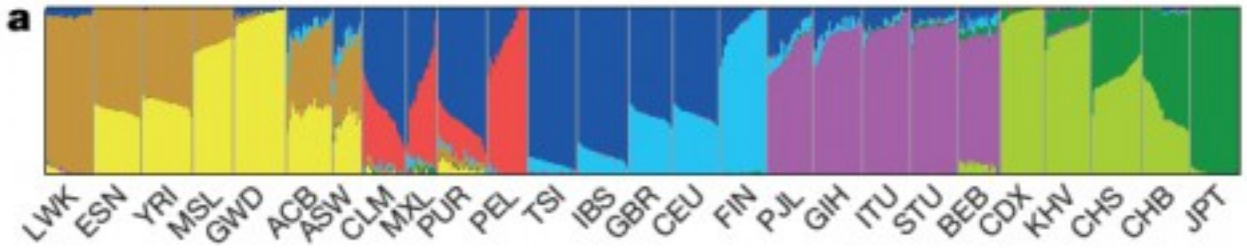


На следующем рисунке показано соотношение числа генетических вариаций (аллелей) и доля в них структурных вариаций (CV). Как видно, если по общему числу вариаций лидирует Африка, то относительная доля структурных вариаций (CV) больше в Европе и Южной Азии.

В итоговой базе генетических вариаций большая часть относится к редким — около 64 млн аутосомных вариаций имеют частоту <0,5%, около 12 млн – частоту между 0,5% и 5% и около 8 млн – частоту >5%. Поиск редких вариаций, подчеркивают авторы, примерно вдвое более эффективен при использовании глубокого секвенирования.

Общие и локальные вариации

Большая часть генетических вариаций из глобальной базы встречается практически во всех изученных популяциях. Но некоторые вариации приурочены к определенному континенту, или группе популяций или специфичны для конкретной популяции. В пределах регионов обнаружились градиенты частот вариаций: в Африке и Восточной Азии они идут по оси восток-запад, а в Европе, Африке и Америке – по оси север-юг. Редкие генетические вариации обычно ограничены родственными популяциями.

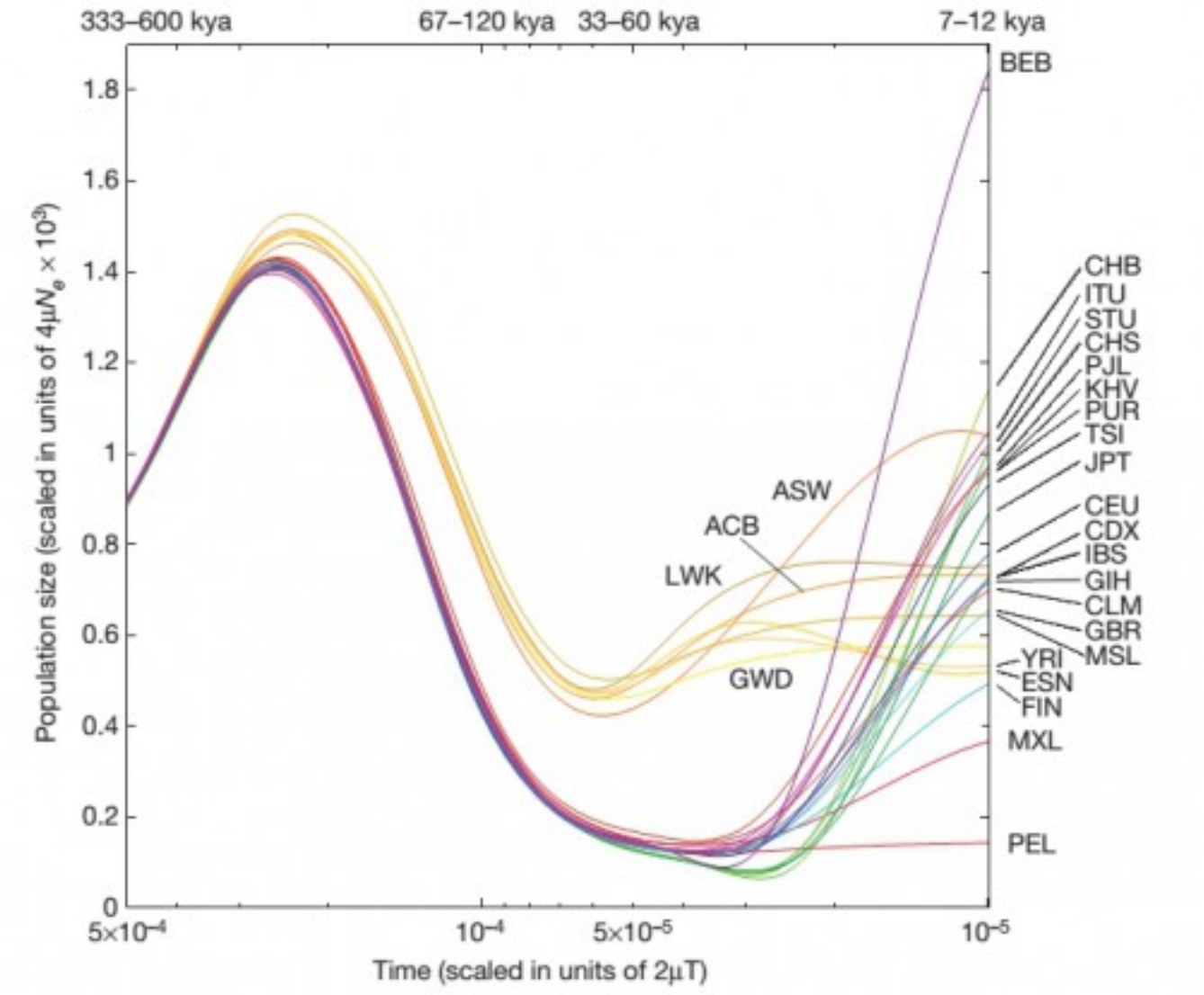


Анализ ADMIXTURE, показывающий вклад предковых популяций (при заданном числе предковых популяций k=8). Обозначения популяций см. выше.

На рисунке, показывающем результаты анализа ADMIXTURE (вклад предковых популяций в изучаемых популяциях) видно, что именно эти, указанные выше, популяции (в которых с высокой частотой встречаются вариации, редкие по глобальной базе) выделяются по составу предковых компонентов из своей региональной группы: лухья (LWK) в Африке, перуанцы (PEL) в Америке, японцы (JPT) в Восточной Азии, финны (FIN) в Европе**.** Впрочем, по мнению специалистов, в отношении финнов можно высказать предположение, что «редкость в глобальной базе» их вариантов связана с неполнотой самой базы — в ней отсутствует население Северной Евразии. Остальные редкости также могут фиксировать крупные пробелы в глобальной базе данных.

Модель реконструкции демографических событий

Авторы использовали метод Pairwise sequential Markovian coalescent (PSMC) для попытки реконструкции эффективного размера (Ne) предковых популяций. Модель показывает, что в интервале 20-15 тыс. лет назад население Европы, Азии и Америки прошло через резкое сокращение эффективного размера популяции (Ne), которое снизилось до величины менее 1500 индивидов. Иными словами, все неафриканские популяции прошли через узкое «бутылочное горлышко», в результате чего их генетическое разнообразие радикально сократилось. В этот же период африканские популяции тоже проходили через «бутылочное горлышко», но гораздо менее узкое (Ne оставалось больше 4000). Что же касается неафриканских популяций, то за сокращением последовал стремительный рост – эффективный размер популяций быстро и значительно увеличился. Правда, уточняют авторы, были и исключения: сигналов быстрого роста Ne не отмечено у перуанцев (PEL), мексиканцев (MXL) и финнов (FIN).



Динамика эффективного размера популяций (Ne) со времени выхода современного человека из Африки до неолита. График получен методом Pairwise sequential Markovian coalescent (PSMC). Обозначения популяций см. выше.

Вариации и функции

Авторы попытались оценить фукциональное значение геномных вариаций. Для этого они ограничили анализ теми, которые изменяют работу гена. И выяснили, что типичный геном содержит примерно 150-180 сайтов с вариациями, обрезающими белки, 10-12 тысяч сайтов с вариациями, меняющими аминокислотное строение белков, и 460-560 тысяч сайтов с вариациями в регуляторных областях (промоторы, инсулаторы, энхансеры и сайты связывания факторов транскрипции). По числу этих функциональных вариаций африканские геномы находятся на верхней границе «вилки». Большие структурные вариации, как правило, оказывают больший функциональный эффект, чем однонуклеотидный полиморфизм (SNP). В рамках проекта исследователи впервые составили каталог этих структурных вариаций.

Но при этом, когда авторы рассмотрели вариации, связанные с заболеваниями, выяснилось, что они подчиняются другой закономерности. Методом широкогеномного поиска ассоциаций (GWAS) они обнаружили 2000 аллельных вариаций на геном, ассоциированных с обычными заболеваниями, и 24-30 вариаций на геном, ассоциированных с редкими заболеваниями. И по этим показателям верхнюю границу «вилки» занимали геномы индивидов европейского происхождения.

Оказалось, что 35% вариаций из каталога GWAS не являются общими для континентальных групп. Этот результат имеет важное практическое значение, так как показывает важность изучения генофондов отдельных популяций для прогноза заболеваемости и планирования лечения.

С другой стороны, авторы обнаружили и совершенно неожиданную вещь. «Мы были поражены тем, что нашли более 200 генов, которые у некоторых людей просто отсутствуют», — сказал Жан Корбель (Jan Korbel), сотрудник EMBL в Гейдельберге, Германия, его слова приведены в пресс-релизе EMBL. При этом, по мнению авторов, такое отсутствие генов,

которое, казалось бы, должно быть фатальным, не оказало влияние на здоровье.

Поиск отбора

Если близкородственные популяции существенно различаются по частоте каких-то генетических вариаций, это может указать на то, что какие-то аллели дают преимущество в выживании индивидов и служат мишенями для отбора. Авторы использовали метод статистики, основанный на генетических расстояниях (Fst) для поиска генов, значительно отличающихся по частоте в парах популяций одной континентальной группы.

Такой подход выявил возможные сигналы отбора в уже известных локусах: это, например, ген SLC24A5 ассоциированный с пигментацией кожи, HERC2 – с цветом глаз, LCT – с лактозной недостаточностью и кластер FADS – с липидным обменом. Но найдено и несколько новых локусов, для которых тоже можно предположить действие отбора: TRBV9, особенно различающийся по частоте в Южной Азии, PRICKLE4, различающийся в популяциях Азии и Африки и гены кластера иммуноглобулинов, различающиеся в популяциях Восточной Азии.

Но – подчеркивают авторы – таких генов, частота которых значительно различается у родственных популяций, оказалось мало. И это говорит о том, что мощный отбор в популяциях – довольно редкое явление.

Источники:

A global reference for human genetic variation

The 1000 Genomes Project Consortium

68 | NATURE | VOL 526 | 1 OCTOBER 2015 doi:10.1038/nature15393

http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html

An integrated map of structural variation in 2,504 human genomes

авторы и аффилиация по ссылке http://www.nature.com/nature/journal/v526/n7571/full/nature15394.html

1 OCTOBER 2015 | VOL 526 | NATURE | 75 doi:10.1038/nature15394