Государственное образовательное учреждение высшего образования

Санкт-Петербургский институт генетики

Статистическая популяционная геномика

Материалы девятой всероссийской научно-практической конференции
с международным участием

г. Санкт-Петербург, 25 и 26 ноября 2021 года

Электронное текстовое издание

Под редакцией д.б.н. О.М. Макарова

Санкт-Петербург
Наукоемкие технологии
2021

ISBN 978-5-907610-00-0

УДК 300.42
ББК 40.06
С23

Рецензенты:
П. Н. Фоков – доктор биологических наук, профессор, институт Московского университета;
В. Г. Ресовский – доктор биологических наук, профессор, Российский университет

Статистическая популяционная геномика [Электронный ресурс]: Материалы девятой всероссийской научно-практической конференции с международным участием; Санкт-Петербург, 25, 26 ноября 2021 г. / под ред. д.б.н. О.М. Макарова. – СПб: Наукоемкие технологии, 2021. – 121 с. – URL: http://publishing.intelgr.com/archive/Statistical-Population-Genomics.pdf.

ISBN 978-5-907610-00-0

В настоящее издание вошли материалы девятой всероссийской научно-практической конференции с международным участием, прошедшей на базе Санкт-Петербургского института генетики (Санкт-Петербург, 25, 26 ноября 2021 г.).

Издание предназначено для специалистов, интересующихся темой конференции, а также студентов и аспирантов профильных вузов.

Исследование выполнено при финансовой поддержке РФФИ
в рамках научного проекта № 20-09-40004.

УДК 300.42
ББК 40.06

ISBN 978-5-907610-00-0

Девятая всероссийская научно-практическая конференция
с международным участием

Статистическая популяционная геномика

Научное издание

Статистическая популяционная геномика

Материалы девятой всероссийской научно-практической конференции
с международным участием
г. Санкт-Петербург, 25 и 26 ноября 2021 года

Электронное текстовое издание

Компьютерная верстка – Е. С. Сидоренко

Подписано к использованию: 05.12.2021 г.
Объём издания – 11,6 Мб.

Издательство «Наукоемкие технологии»
ООО «Корпорация «Интел Групп»
http://publishing.intelgr.com
E-mail: publishing@intelgr.com
Тел.: +7 (812) 945-50-63

СОДЕРЖАНИЕ

Раздел I

Обработка и анализ выравниваний нескольких геномов

Глава 1

Обработка и анализ выравниваний нескольких геномов с помощью фильтра Maf

Хасаев Виктор Евгеньевич

Институт инновационных биотехнологий, г. Москва

Аннотация

По мере увеличения числа доступных последовательностей генома как близкородственных видов, так и особей внутри вида появились теоретические и методологические сходства между областями филогеномики и популяционной геномики. Популяционная геномика обычно фокусируется на анализе вариантов, в то время как филогеномика в значительной степени опирается на выравнивание генома. Однако они играют все более важную роль в исследованиях на популяционном уровне. Множественные выравнивания генома индивидуумов используются, когда структурные вариация представляет первостепенный интерес и тогда, когда архитектура генома позволяет собирать последовательности генома de novo. Здесь я описываю MafFilter, управляемую командной строкой программу, позволяющую обрабатывать выравнивания генома в формате множественного выравнивания (MAF). Используя конкретные примеры, основанные на общедоступных наборах данных, я демонстрирую, как MafFilter можно использовать для разработки эффективных и воспроизводимых конвейеров с гарантией качества для последующего анализа. Далее я покажу, как MafFilter можно использовать для выполнения обоих основных и расширенный популяционный геномный анализ с целью определения закономерностей нуклеотидного разнообразия в геномах.

Ключевые слова множественное выравнивание генома, синтез, постобработка выравнивания, качественная фильтрация, формат множественного выравнивания

Abstract

As the number of available genome sequences from both closely related species and individuals within species increased, theoretical and methodological convergences between the fields of phylogenomics and population genomics emerged. Population genomics typically focuses on the analysis of variants, while phylogenomics heavily relies on genome alignments. However, these are playing an increasingly important role in studies at the population level. Multiple genome alignments of individuals are used when structural variation is of primary interest and when genome architecture permits to assemble de novo genome sequences. Here I describe MafFilter, a command-line-driven program allowing to process genome alignments in the Multiple Alignment Format (MAF). Using concrete examples based on publicly available datasets, I demonstrate how MafFilter can be used to develop efficient and reproducible pipelines with quality assurance for

downstream analyses. I further show how MafFilter can be used to perform both basic and advanced population genomic analyses in order to infer the patterns of nucleotide diversity along genomes.

Key words Multiple genome alignment, Synteny, Alignment post-processing, Quality filtering, Multiple alignment format

1 Введение: Множественное выравнивание генома

Множественные выравнивания генома (MGAS) регистрируют гомологические отношения между родственными последовательностями генома. В то время как обычные выравнивания последовательностей содержат информацию о нуклеотидных заменах, вставках и делециях, MGAS кодируют эволюционные события, происходящие в большем масштабе. Такие события включают хромосому слияние, расщепление и перегруппировки, которые нарушают коллинеарность между последовательностями (разрыв акасинтены). Кроме того, последовательности генома, в отличие от последовательностей генов, обычно сегментированы. Основная причина этой сегментации может быть биологической (наличие множества хромосом) или технической (последовательность генома может быть собрана только на уровне контига или каркаса).

2 Общие принципы использования маффильтра

Поскольку файлы MAF изначально использовались для выравнивания по нескольким видам, каждый входной геном упоминается как вид. В дальнейшем вид может, однако, также обозначать определенный штамм или особь в популяции. Аналогичным образом, термин "хромосома" будет использоваться в широком смысле, охватывающем каркасы и контиги, в случае несопоставленных геномных сборок (см. рис. 1).

2.1 Последовательная обработка блоков выравнивания: Фильтры

Поскольку файлы MAF организованы в серию синтетических блоков, Maf- Filter последовательно обрабатывает входные файлы по одному блоку за раз с помощью применение фильтров. Фильтр принимает блок выравнивания MAF в качестве входных данных, проводит один или несколько анализов и возвращает блок MAF. В зависимости от типа выполняемого анализа выходной блок может быть идентичен входному или иметь модифицированную версию. В некоторых случаях фильтр может вычислять дополнительную информацию, которая может быть записана в выходной файл или сохранена в виде метаданных (примеры см. в таблице 1). Фильтры комбинируются последовательно, выходные данные одного фильтра служат входными данными для следующего, что позволяет разрабатывать расширенные рабочие процессы анализа.

Формат множественного выравнивания (MAF, не путать с форматом аннотаций мутаций) описывает отношения гомологии между несколькими

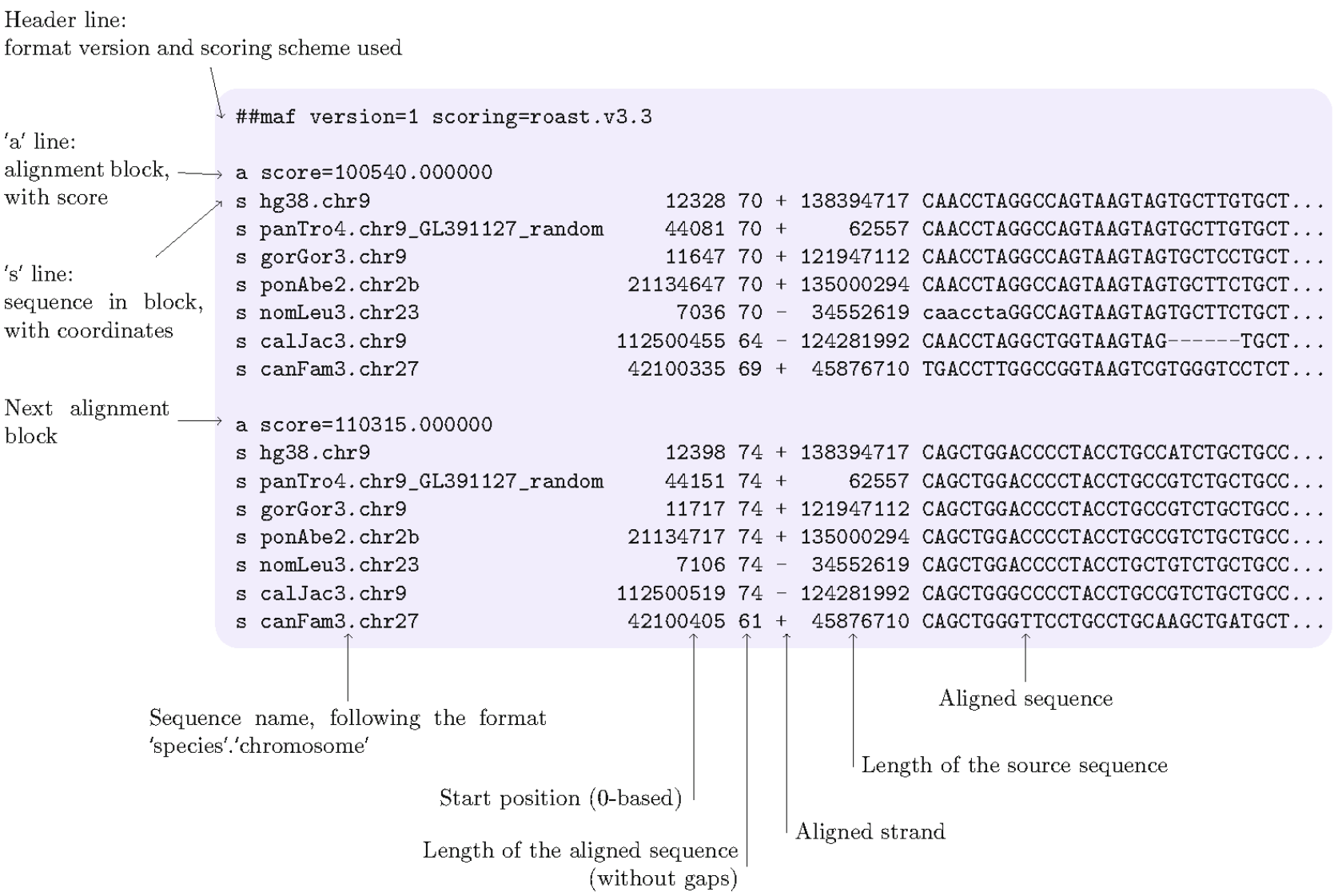

Рис. 1 Структура файла MAF. Источник данных: UCSC выравнивание хромосомы 9 человека вместе с 19 млекопитающими, среди которых 16 приматов

геномами в виде плоских текстовых файлов (см. https://genome.ucsc.edu/FAQ/FAQformat.html, последнее обращение 29/08/18). Файл MAF представляет собой список из нескольких блоков выравнивания, где составляющие последовательности находятся в синтении (см. рис. 1). В то время как структура каждого блока идентична традиционным выравниваниям последовательностей (как в форматах Clustal или Phylip), где гомологичные позиции в каждой последовательности находятся друг над другом и образуют выравнивание столбец, имена последовательностей следуют выделенному синтаксису для записи координат генома. Кроме того, может быть включено несколько строк аннотации, включая, например, оценки качества последовательности. Программы выравнивания генома, создающие файлы MAF в качестве выходных данных, включают TBA [4], Mugs [5], ROAST http://www.bx.psu.edu/~cathy/toast-roast.tmp/README.toast-roast.html (последний доступ 29/08/18), Последний [6] и лиловый [7].

2.2 Файлы опций и командная строка Аргументы

Программой MafFilter можно управлять с помощью аргументов, которые передаются из командной строки или, что более удобно, в виде файла сценария. Аргументы принимают форму операторов ‘parameter’ и 'value', которые потенциально могут быть вложенными. Аргументы также могут быть вызваны внутри скрипта, что позволяет определять глобальные переменные. Ниже приведен минималистский пример, демонстрирующий синтаксис:

Таблица 1

Примеры типов фильтров, поддерживаемых MafFilter

Имя фильтра	Функции фильтра	Вывод
MafStatistic	Вычисляет статистику по блоку	Неизмененный входной блок
Минимальная длина	Фильтр блокирует заданную длину выравнивания	Неизмененный входной блок
Подмножество	Сохраняет только подмножество видов	Блок с последовательностями из указанного набора видов
WindowSplit	Разбивает блок на более мелкие блоки заданного размера	Несколько блоков меньшего размера
DistanceEstimation	Вычисляет матрицу эволюционных расстояний из всех последовательностей в блоке	Неизмененный входной блок с матрицей расстояний, прикрепленной в качестве метаданных

```
1 # maffilter param=MinimalistExample.bpp DATA=chr9
2 input.file=../Primates/$(DATA).maf.gz
3 input.file.compression=gzip
4 input.format=Maf
5 output.log=$(DATA).maffilter.log
6 maf.filter=\
7 MinBlockLength(min_length=1000),\
8 Output(file=$(DATA).min1kb.maf.gz, compression=gzip)
```

Таблица также содержит координаты блока в соответствии с одним эталонным видом. Следующий конвейер является модификацией того, который представлен в Подзаголовок 3.1. После каждого шага добавляется фильтр статистики последовательностей, который сообщает длину (количество столбцов выравнивания) и размер (количество последовательностей) каждого блока. При этом создаются четыре файла, обобщенные на рис. 2.

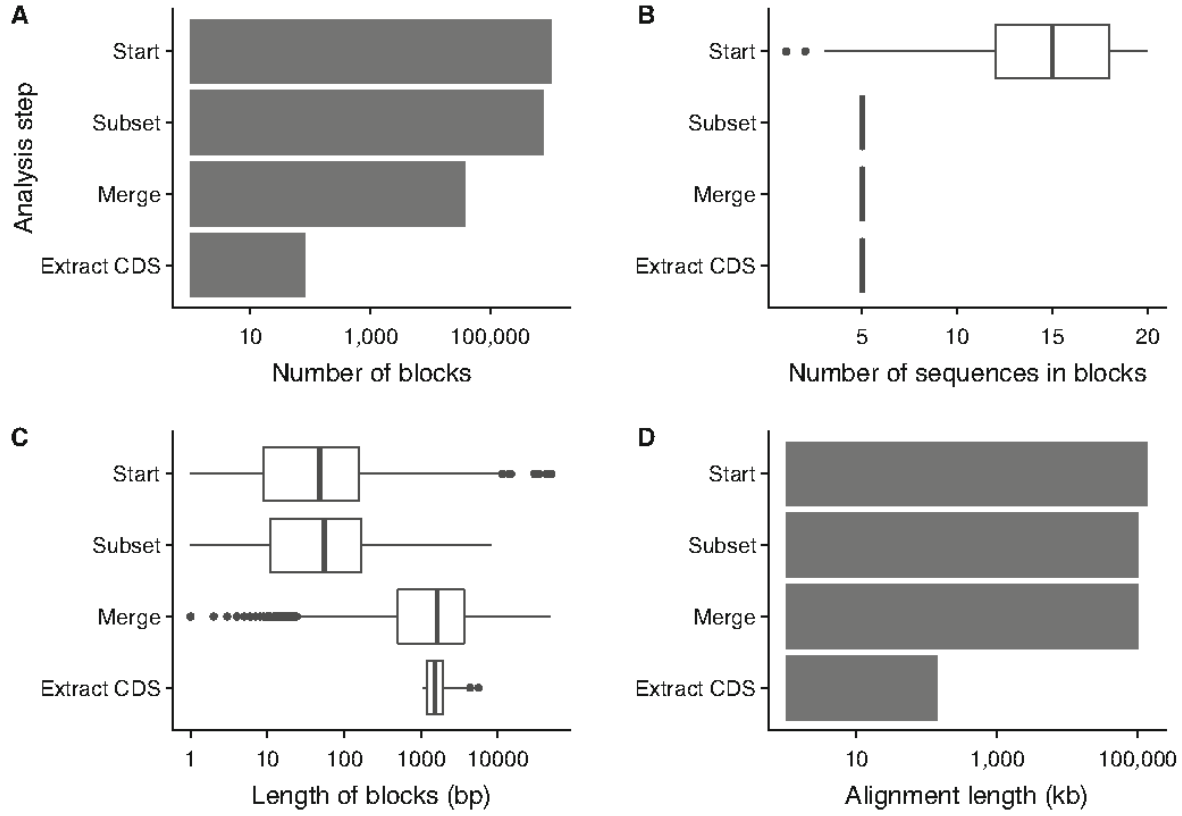

Рис. 2 Эффект фильтров извлечения данных, измеренный с помощью фильтров статистики. Показаны четыре этапа: перед фильтрацией (“Старт”), после подмножества до пяти видов приматов (“Подмножество”), после объединения блоков синтеза (“Слияние”) и после извлечения областей CDS (“Извлечение CDS”). (A) Количество блоков после каждого шага. (B) распределение размеров блоков, то есть количество видов, представленных в каждом блоке. (C) Распределение длин блоков, то есть количество столбцов выравнивания в каждом блоке. (D) Общая длина выравнивания, то есть сумма длин всех блоков

4.2 Пример Анализ 2: Вывод Филогенетический Отношения

В этом примере мы выводим филогенетические отношения пяти человекообразных обезьян. Мы используем 20-стороннее выравнивание генома UCSC, содержащее 16 Геномов приматов. Ради эффективности вычислений мы ограничиваем анализ только хромосомой 9. Мы реализуем следующий конвейер:

1) извлеките выравнивание генома человека, шимпанзе, бонобо, гориллы и орангутанга,

2) отфильтруйте выравнивание, чтобы удалить неоднозначно выровненные области,

3) установите модель эволюции последовательности для человека, шимпанзе и гориллы в группе, используя максимальное правдоподобие и выходные параметры в файл.

5 Другие полезные инструменты

MafFilter предоставляет инструменты для последовательного анализа файла MAF. Эти инструменты в первую очередь ориентированы на обработку данных для статистического анализа. Он имеет ограниченные возможности форматирования, в частности, когда задействованы операции с большим диапазоном, такие как изменение порядка блоков выравнивания. Пакеты TBA [4] и Last [6] содержат несколько полезных инструментов для этой цели, которые можно использовать в сочетании с MafFilter.

Из пакета TBA:

- программа maf_order позволяет выбирать и упорядочивать последовательности в соответствии с названиями их видов;
- программа maf_project упорядочивает блоки выравнивания в соответствии с эталонным геномом. Блоки, в которых эталонный геном находится на отрицательной цепи, будут перевернуты. Все блоки, которые не содержат ссылочных видов, будут отброшены.

Из последней упаковки:

- программа maf-join позволяет объединять несколько (отсортированных) множественные выравнивания;
- программа maf-sort позволяет сортировать выравнивания в соответствии с названиями последовательностей.

6 Заключение

Программа MafFilter позволяет эффективно обрабатывать несколько файлов выравнивания генома путем последовательного анализа блоков синтеза. Он имеет гибкий и расширяемый синтаксис, позволяющий создавать воспроизводимые конвейеры для последующей обработки данных генома. Помимо фильтрации и оценки качества, MafFilter можно использовать для анализа закономерностей разнообразия геномов, внутри видов и между ними.

Литература

1. Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, Hinrichs AS, Haeussler M, Guruvadoo L, Navarro Gonzalez J, Gibson D, Fiddes IT, Eisenhart C, Diekhans M, Clawson H, Barber GP, Armstrong J, Haussler D, Kuhn RM, Kent WJ (2018) The UCSC Genome Browser database: 2018 update. Nucleic Acids Res 46(D1): D762–D769. https://doi.org/10.1093/nar/gkx1020

2. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330

3. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7. https://doi.org/10.1186/s13742-015-0047-8

4. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, MillerW(2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14(4):708–715. https://doi.org/10.1101/gr.1933104

5. Angiuoli SV, Salzberg SL (2011) Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27(3):334–342. https://doi.org/10.1093/bioinformatics/btq665

6. Kiełbasa SM,WanR, SatoK,Horton P, FrithMC (2011) Adaptive seeds tame genomic sequence comparison. Genome Res 21(3):487–493. https://doi.org/10.1101/gr.113985.110

7. Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5(6):e11147. https://doi.org/10.1371/journal.pone.0011147

8. Stukenbrock EH, Christiansen FB, Hansen TT, Dutheil JY, Schierup MH (2012) Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. Proc Natl Acad Sci USA 109 (27):10954–10959. https://doi.org/10.1073/pnas.1201403109

9. Stukenbrock EH, Dutheil JY (2018) Fine-scale recombination maps of fungal plant pathogens reveal dynamic recombination landscapes and intragenic hotspots. Genetics 208 (3):1209–1229. https://doi.org/10.1534/genetics.117.300502

10. Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE (2015) The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. Genetics 199(4):1229–1241. https://doi.org/10.1534/genetics.115. 174664

11. Higgins DG, Sharp PM (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73 (1):237–244

12. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nat Genet 46 (8):919–925. https://doi.org/10.1038/ng.3015

Глава 2

Управление данными и сводная статистика с PLINK

Перьев Валерий Викторович

Институт проблем биохакинга, г. Омск

Аннотация

PLINK – это универсальная программа, которая поддерживает управление данными, контроль качества и общие статистические вычисления на матрицах вызовов геномных вариантов эффективным с вычислительной точки зрения способом. В популяционной геномике он часто используется для того, чтобы позаботиться об “основах”, поэтому их не нужно повторно применять, когда необходимо выполнить новый тип анализа на такой матрице. Я описываю некоторые из этих основных операций и обсуждаю способы их использования и подводные камни.

Ключевые слова частота аллелей, равновесие Харди–Вайнберга, неравновесие сцепления, главный компонент анализ, вывод о взаимоотношениях, вывод о поле, вариант формата вызова

Abstract

PLINK is a versatile program which supports data management, quality control, and common statistical computations on matrices of genomic variant calls, in a computationally efficient manner. In population genomics, it is frequently used to take care of the “basics,” so they do not need to be reimplemented when a new type of analysis needs to be performed on such a matrix. I describe several of these basic operations, and discuss uses and pitfalls.

Key words Allele frequency, Hardy–Weinberg equilibrium, Linkage disequilibrium, Principal component analysis, Relationship inference, Sex inference, Variant call format

1 Введение

Чипы для генотипирования и машины для секвенирования выдают данные в самых разнообразных форматах. Однако все они пытаются измерить одно и то же: каковы последовательности генома этих организмов? Конечно, эти последовательности будут иметь тенденцию…