БИОФИЗИКА, 2023, том 68, № 2, с. 263–267

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 519.876.5

ИНСТРУМЕНТЫ ДЛЯ ВИЗУАЛИЗАЦИИ СИГНАЛА ПОКРЫТИЯ СЕКВЕНИРОВАНИЯ

© 2023 г. И.В. Бездворных*, Н.А. Черкасов*, А.А. Канапин*, А.А. Самсонова*, #

**Санкт-Петербургскии государственный университет, Университетская набережная, 7–9, Санкт-Петербург, 199034, Россия*

#E-mail: a.samsonova@spbu.ru

Поступила в редакцию 22.11.2022 г.
После доработки 22.11.2022 г.
Принята к публикации 07.12.2022 г.

Данные полногеномного секвенирования позволяют не только получать информацию о генетических вариантах, но также оценивать общую стабильность генома. Сигнал покрытия секвенирования, понимаемый как количество выравненных фрагментов в данной точке генома может быть использован в качестве ценного источника данных как о местонахождении структурных перестроек, об общем состоянии генома и о точности предсказания структурного варианта вычислительным алгоритмом. Последнее особенно важно, т.к. методы поиска перестроек в геноме часто дают очень противоречивую информацию о их нахождении. Однако до недавнего времени валидация предсказанных вариантов была затруднена, во многом из-за отсутствия информационных ресурсов, позволяющих напрямую работать с сигналами покрытия и визуализировать их с высокой степенью детализации. В работе представлен SCOPE (Sequence COverage ProfilEs) – прототип ресурса такого рода, включающий в себя базу данных, веб-интерфейс и набор программ для обработки данных секвенирования, извлечения и хранения профилей сигнала покрытия. Вычислительная платформа и интерфейс реализованы в программном коде открытого доступа и могут быть развернуты на локальном узле, что дает возможность пользователям проводить обработку и анализ собственных данных.

Ключевые слова: полногеномное секвенирование, база данных, визуализация, сигнал покрытия.

DOI: 10.31857/S0006302923020072, **EDN:** CAPGJL

Данные полногеномного секвенирования – незаменимый источник информации не только о генетических вариациях, но и о состоянии генома в целом. Сигнал покрытия секвенирования, представляет собой последовательность величин, соответствующих числу фрагментов секвенирования, выравненных на данную позицию референсного генома. Анализ сигнала покрытия широко используется для обнаружения структурных вариантов, особенно вариантов копийности, поскольку данная величина не зависит от протоколов и способов секвенирования, а также от длины фрагментов секвенирования и их типа (см., например, работу [1]). Однако до сих пор не существует надежного решения задачи поиска структурных вариантов в данных секвенирования [2–4]. Геномные координаты структурных вариантов, полученные с помощью различных алгоритмов, плохо согласуются друг с другом, поэтому дополнительно приходится применять так называемые meta-callers т.е. программы позволяющие получить консенсусное предсказание на основе результатов работы нескольких программ [5]. Все это существенно затрудняет разработку диагностических тестов на основе анализа структурных вариантов.

Одним из факторов, затрудняющих создание новых методов поиска структурных вариантов, является отсутствие открытых ресурсов, содержащих в достаточном объеме информацию о форме и характере сигнала покрытия, соответствующего тем или иным структурным вариациям. В данной работе мы представляем SCOPE (Sequence COverage ProfilEs) – базу данных и программный интерфейс к ней, позволяющие хранить, анализировать и визуализировать сигнал покрытия, полученный из данных полногеномного секвенирования. База данных реализована в формате SQLite и может быть развернута на пользовательском компьютере вместе с интерфейсом. Также мы разработали набор программных средств для обработки данных секвенирования и подготовки их к загрузке в базу данных.

ИСТОЧНИКИ ДАННЫХ

В качестве демонстрационных данных для базы были использованы два набора данных секвенирования:

1. Один из эталонных образцов консорциума GIAB (genome-in-a-bottle) [6, 7], а именно HG002 [8] и соответствующие ему структурные варианты.

2. Три генома детей из семей YRI, CHS и PUR, опубликованных в работе [9]

Выбор данных образцов обусловлен следующими факторами. В течение последнего десятилетия консорциумом GIAB создается и поддерживается эталонная коллекция геномов человека. Образцы ДНК секвенируются с использованием различных протоколов, после чего в данных секвенирования осуществляется поиск генетических вариаций: однонуклеотидных замен и структурных вариантов. Образцы консорциума представляют собой *de facto* стандарт данных секвенирования и используются для сравнения производительности методов поиска вариантов, а также для оценки их точности. Образец HG002 примечателен тем, что для него впервые была проведена масштабная проверка точности предсказания структурных вариантов с помощью кураторов [10]. Анализ трех геномов, опубликованных в работе [9] и составляющих вторую часть данных, проводился с помощью протоколов секвенирования длинными и короткими фрагментами, что позволяет обеспечить приемлемую точность предсказания структурных вариантов, за счет сравнения и интеграции результатов, полученных при помощи разных протоколов секвенирования. Более того, так как в работе был проведен анализ геномов семейных трио, в отличие от большинства геномных исследований, генотип детей не был предсказан при помощи вычислительных процедур, а определен точно на основании данных родителей.

Нами были использованы файлы выравнивания в формате BAM, и информация о структурных вариантах в формате VCF, опубликованные авторами данных исследований. Данные находятся в открытом доступе и могут свободно использоваться и распространяться.

ИЗВЛЕЧЕНИЕ СИГНАЛА ПОКРЫТИЯ И КООРДИНАТ ТОЧЕК РАЗРЫВА

Для хранения в базе данных и последующей визуализации были использованы профили сигнала покрытия в окрестности ±256 нуклеотидов от точки разрыва, соответствующей тому или иному структурному варианту. Сигнал покрытия хранится в специальном двоичном формате BCOV [11]. Программный код предусматривают адаптацию ресурса под задачи пользователя за счет манипуляции параметрами генерации файлов базы данных и построения графического интерфейса.

Более того, путем подбора и установки соответствующих параметров можно создать базу данных сигналов покрытия для произвольных типов профилей сигнала (т.е., например. не только для структурных вариаций, но и для локусов однонуклеотидных замен) и в интервалах произвольной длины. Информация о геномных координатах точки разрыва извлекается из VCF-файлов, тип структурного варианта и его длина определяются по флагам SVTYPE и SVLEN соответственно при помощи программы import_vcf.py (см. репозиторий проекта на Github: https://github.com/comp-bio/scope).

При анализе профиля сигнала покрытия с целью обнаружения структурного варианта желательно его «обособить» для упрощения задачи поиска точки. Иначе говоря, необходимо избегать ситуаций, когда в окно поиска попадает несколько кандидатов или так называемых «сложных» событий, когда они наслаиваются друг на друга в результате процесса хромотрипсиса [12, 13]. Как следствие, для визуализации мы отбираем в базу данных только те точки разрыва, для которых в указанную окрестность (256 п.о.) не попадают никакие другие варианты. Если размер структурной вариации слишком мал (т.е. меньше, чем 30 нуклеотидов), то она по умолчанию будет пропущена. Последнее решение мотивируется тем, что программы для поиска коротких структурных вариантов основываются на алгоритмах нахождения однонуклеотидных замен и основаны на других принципах нежели методы поиска длинных событий. Однако вполне возможно, что сигналы, возникающие в результате коротких перестроек, могут представлять научный интерес, поэтому импортировать сигналы коротких структурных вариаций можно, указав в параметре программы «special» их минимальный размер. В результате указанные геномные перестройки будут добавлены в базу данных с маркером «spSV» (специального типа).Предусмотрено четыре типа маркеров, связанных с типом точки разрыва: «L» и «R» – соответственно для левой и правой границ варианта, «BP» – присваивается для так называемых перестроек «нулевой длины». Перестройки «нулевой длины» – это сбалансированые перестройки, такие как, например, инсерции, инверсии и транслокации, где технически обе точки разрыва попадают в одну координату в референсном геноме. Тип «SpSV» соответствует вариантам, длина которых меньше, чем ширина региона, используемого для извлечения сигнала покрытия (512 п.о. по умолчанию). Разметка хромосом доступна для обеих версий человеческого генома (GRCh38 и GRCh37).

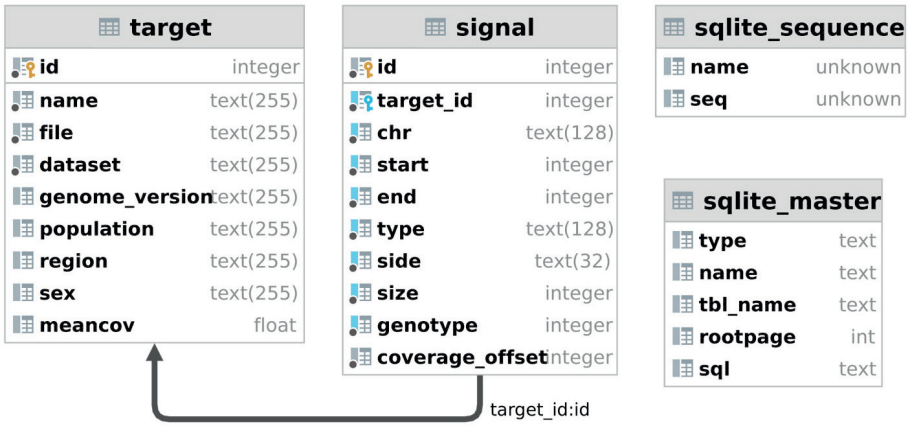

Рис. 1. Схема базы данных профилей сигнала покрытия SCOPE.

Сигнал покрытия извлекается из файлов выравнивания в формате BAM или CRAM с помощью программы mosdepth [14], определяющей численные значения покрытия для каждого основания в геноме. В базе данных профили сигнала секвенирования сохраняются в формате BCOV, специально разработанном для быстрого доступа к значениям покрытия в любой точке. Идея формата максимально проста: для каждой позиции в хромосоме отводится ровно два байта. Это свойство позволяет получить доступ к любой части генома, используя побайтовый отступ в файле покрытия, равный удвоенной координате искомого покрытия. Такое кодирование не позволяет сохранять глубину прочтения, превышающую 65536 прочтений на одну пару оснований. Таким образом, в случае обнаружения покрытия со значениями, превышающими это число, значения будут принудительно ограничены сверху. Обработка данных покрытия проводится программой import_coverage.py (см. репозиторий проекта на Github).

Для функционирования базы данных также необходимы метафайлы, содержащие следующую информацию: гистограммы покрытия для различных групп образцов (например, популяций), статистика количества точек разрыва различных типов и наконец хромосомная карта плотности точек разрыва в геноме. Генерация метафайлов осуществляется программой overview.py (см. репозиторий проекта на Github).

СТРУКТУРА И ИНТЕРФЕЙС БАЗЫ ДАННЫХ

SCOPE состоит из трех ключевых частей – инструментов, базы данных и интерфейса, каждая из которых максимально изолирована и независима от прочих. Это фактически означает, что у исследователей есть возможность использовать один интерфейс и подключать его к нескольким базам данных, расположенным локально или на удаленных серверах и смонтированных через инструменты удаленного доступа (например, SSH). Базы данных и интерфейс поддерживают объединение нескольких наборов данных в одном месте. Схема базы данных приведена на рис. 1.

Готовая к визуализации база данных вместе с файлом сигналов покрытия и метафайлом может распространяться как отдельный набор данных, например:

EXAMPLE/index.db — база данных SQLite;

EXAMPLE/storage.bcov — все сигналы со значениями покрытий в BCOV-формате;

EXAMPLE/overview.json — мета-информация о проекте.

Подключение интерфейса к указанной базе производится через параметр «db» (db:EXAMPLE).

Пользовательская часть интерфейса реализована на JavaScript (React, SASS, Webpack, D3.js), серверная – на Python3 (Flask, gevent, tslearn). Полностью скомпилированная версия находится прямо в репозитории (раздел build), для ее развертывания на локальном узле требуется установка следующих модулей языка програмирования Python, версия3:

pip3 install flask tslearn gevent.

Запуск сервера осуществляется командой:

python3 server.py db:EXAMPLE_DB.

Интерфейс позволяет визуализировать извлеченные профили, осуществлять поиск профилей по их геномным координатам, типу структурных вариантов и точек разрыва, а также по генотипу. Пример графического представления данных показан на рис. 2.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

SCOPE – комплекс программного обеспечения для визуализации сигнала покрытия секве-

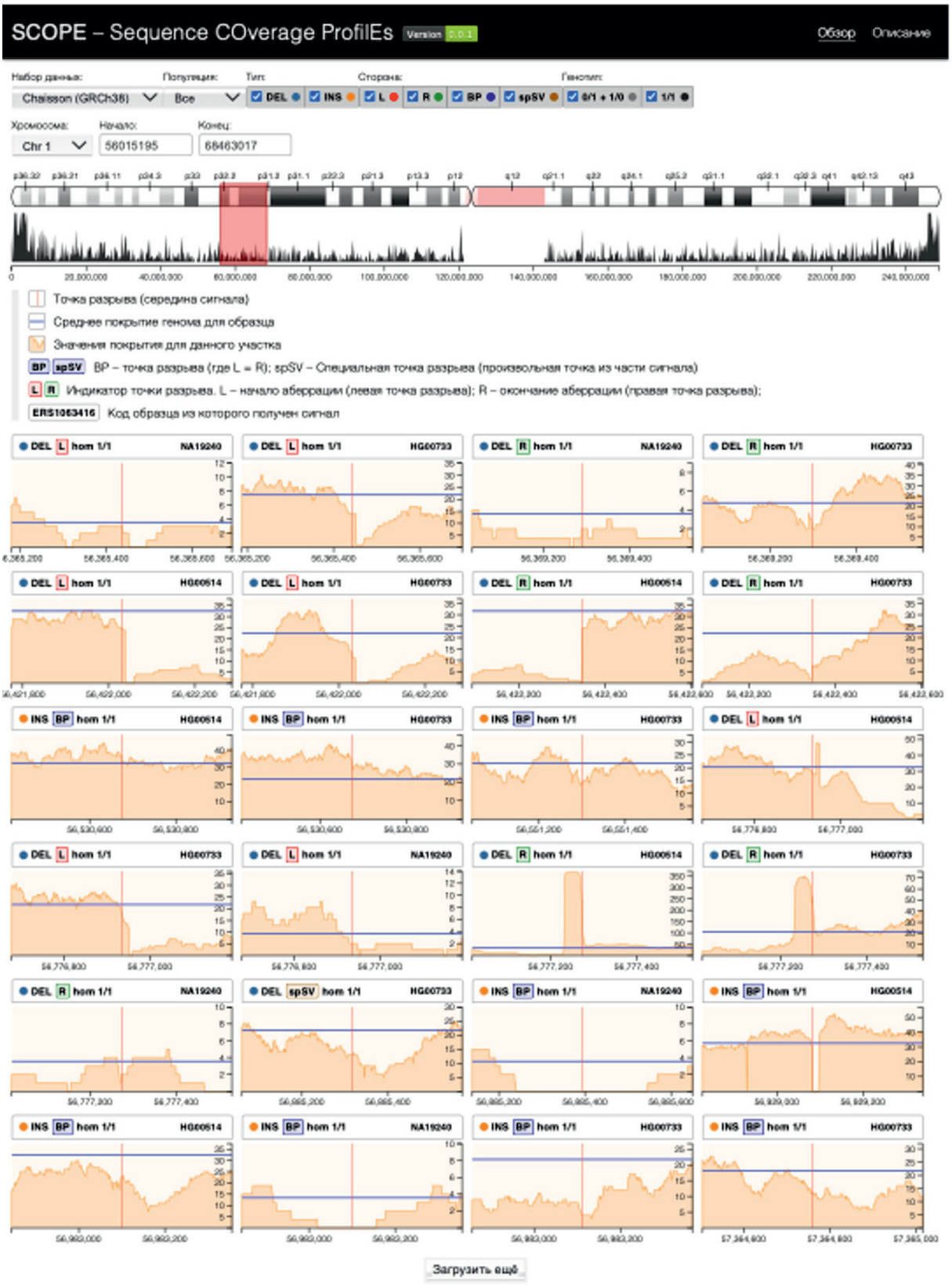

Рис. 2. Графический интерфейс ресурса SCOPE (scope.compbio.ru).

нирования, включающий в себя базу данных, веб-интерфейс и набор программ для обработки данных. Репозиторий открытого кода находится по адресу https://github.com/comp-bio/scope.

Доступ к ресурсу с полнофункциональной версией интерфейса SCOPE, включающей визуализацию профилей сигнала, поиск по геномным координатам и фильтрацией по типам точек разрыва осуществляется по адресу: https://scope.compbio.ru/.

Созданный ресурс открывает новые возможности для исследования нового типа данных, получаемых при полногеномном секвенировании. Данные такого типа могут быть использованы как для разработки новых методов поиска структурных вариантов, так и для анализа механизмов поддержания стабильности генома.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при финансовой поддержке Российского научного фонда (грант № 20-14-00072).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. A. Abyzov, et al., Genome Res., **21** (6), 974 (2011).
2. S. Kosugi, et al., Genome Biol., **20** (1), 117 (2019).
3. Z. Liu, et al., Genome Biol., **23** (1), 68 (2022).
4. M. Mahmoud, et al., Genome Biol., **20** (1), 1 (2019).
5. A. Kuzniar, J. Maassen, S. Verhoeven, et al., PeerJ, **18**, e8214 (2020). DOI: 10.7717/peerj.821
6. J. M. Zook, et al., Sci. Data, **3**, 160025 (2016).
7. J. M. Zook, et al., Nat. Biotechnol., **32** (3), 246 (2014).
8. A. Shumate, et al., Genome Biol., 1 (2020).
9. M. J. P. Chaisson, et al., Nat. Commun., 10 (1), 1 (2019).
10. L. M. Chapman, et al., PLoS Comput. Biol. **16** (6), e1007933-20 (2020).
11. I. Bezdvornykh, A. Kanapin, and A. Samsonova, In *Abst. Book of the Thirteenth Int. Multiconf. on Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022)* (2022), p. 762.
12. J. O. Korbel and P. J. Campbell, Cell, **152** (6), 1226 (2013).
13. A. Aguilera and T. García-Muse, Annu. Rev. Genetics, **47** (1), 1 (2013).
14. B. S. Pedersen and A. R. Quinlan, Bioinformatics, **34** (5), 867 (2018).

A Toolbox for Visualization of Sequencing Coverage Signal

I.V. Bezdvornykh*, N.A. Cherkasov*, A.A. Kanapin*, and A.A. Samsonova*

**St. Petersburg State University, Universitetskaya nab. 7–9, St. Petersburg, 199034 Russia*

Whole genome sequencing data allow access not only to information about genetic variation, but also provide an opportunity to evaluate the overall genome stability. Sequencing coverage signal considered as the number of fragments alligned to a given region within the genome can be used as a trustworthy source of data both on discovery of genomic rearrangements and the current state of whole genome sequencing as well as on precision of structural variant predictions by computational algorithms. The latter is of utmost importance as conflicting data on gene rearrangement events obtained by tools for finding gene rearrangements often appear. However, until recently, validation of predicted variants may present a significant challenge mainly due to the lack of information sources that may assist researchers with direct work with coverage signals and signal visualization with high precision. The present study proposes Sequence COverage ProfilEs (SCOPE), a prototype toolset that includes databases, web-interface and a series of programs for the processing of sequencing data, visualizing and storing of signal coverage profiles. The computer platform and interface is equipped with open-source software, supports local host deployment and allows users to process and analyze their own sequencing data.

Keywords: whole genome sequencing, database, visualization, coverage signal