

Современные программные решения для преобразования геномных координат между различными версиями сборки генома человека

Бобрик П.Ю.*, Вертёлко В.Р., Гурьянова И.Е.

Республиканский научно-практический центр детской онкологии, гематологии и иммунологии, Боровляны, Республика Беларусь

* bobrykpavel2001@gmail.com

Ключевые слова: преобразование геномных координат; liftOver; GRCh37; GRCh38

Мотивация и цель: Референсная геномная сборка является необходимым ресурсом при работе с данными высокопроизводительного секвенирования. На сегодняшний день существуют две актуальные версии геномных сборок человека – GRCh37 и GRCh38. Самой свежей считается версия GRCh38, одним из основных обновлений в которой стало закрытие многочисленных пробелов, присутствующих в предыдущей сборке [1]. Несмотря на это, многие исследователи не решаются перейти на актуальную версию геномной сборки, так как подобный переход может потребовать значительных изменений в алгоритме обработки данных: установка/настройка дополнительных программных решений, скачивание/интеграция новых баз данных, а также редактирование общего пайплайна обработки. В свою очередь, использование GRCh37 накладывает ряд ограничений на возможность работы с актуальными базами данных, так как их авторы сфокусировались на свежей сборке GRCh38. Если рассмотреть задачу поиска вариантов, то, в случае наличия сырых FASTQ-файлов, можно повторно провести процедуру маппинга и вызова вариантов для актуальной GRCh38. Данный процесс может потребовать дополнительных вычислительных и временных ресурсов, что зачастую не всегда возможно. Как альтернативу данному решению можно рассмотреть наиболее эффективную в плане используемых ресурсов процедуру, а именно – конвертацию геномных координат. Учитывая все вышесказанное, а также то, что наиболее популярный инструмент для преобразования геномных координат NCBI's Remap Tool прекратил свою работу в ноябре 2023 г., была поставлена цель провести обзор существующих программных решений, имеющих схожий функционал.

Методы и алгоритмы: Были протестированы следующие биоинформатические инструменты для преобразования геномных координат: CrossMap, LiftoverVcf (Picard), transanno, Assembly Converter (Ensembl). Для оценки работоспособности программ использовались 10 VCF-файлов (≈1.5 Гб каждый), содержащих в среднем 4.5 млн замен, выявленных в ходе полногеномного секвенирования, а также файл GRCh37_to_GRCh38.chain, необходимый для выполнения конвертации. Тестирование осуществлялось в ОС Ubuntu (22.04).

Результаты: Программа CrossMap смогла обработать до 98,70 % всех замен в среднем за 2 минуты 17 секунд (±10 с) с минимальным использованием вычислительных ресурсов. Инструмент transanno показал схожие результаты: было преобразовано 98.72 % вариантов за 2 минуты 7 секунд (±6 с). Программа

LiftoverVcf (Picard) конвертировала 98.67 % замен в среднем за 1 минуту 15 секунд (± 5 с), однако потребление оперативной памяти в пиковый момент достигало 12.8 Гб. Веб-инструмент Assembly Converter (Ensembl) дает возможность запустить преобразование координат в удобном графическом интерфейсе. Стоит отметить, что максимальный размер загружаемого в Assembly Converter файла не должен превышать 50 Мб, что накладывает ограничение на работу с данными полноэкзомного и полногеномного секвенирования.

Выводы: Было рассмотрено четыре инструмента для преобразования геномных координат. CrossMap и transanno можно назвать наиболее универсальными вариантами для данной задачи, так как данные программы легко устанавливаются и не требуют больших вычислительных мощностей. Assembly Converter является хорошим вариантом для специалистов с небольшим опытом работы в командной строке. LiftoverVcf (Picard) удобен в случаях, когда нужно провести преобразование в более сжатые сроки за счет дополнительных вычислительных ресурсов.

Modern software solutions for converting genomic coordinates between different versions of human genome assemblies

Bobryk P.Y.*, Vertelko V.R., Guryanova I.E.

Belarusian Research Center for Pediatric Oncology, Hematology and Immunology, Borovlyany, Belarus

* bobrykpavel2001@gmail.com

Key words: genomic coordinates transformation; liftOver; GRCh37; GRCh38

Motivation and Aim: The reference genome assembly is a necessary resource for processing high-throughput sequencing data. Today, there are two current versions of human genome assemblies, GRCh37 and GRCh38. The newest version is GRCh38, with one of the major updates being the closure of numerous gaps present in the previous assembly [1]. Despite this, many researchers hesitate to switch to the current version of the genomic assembly, as such a switch may require significant changes in the data processing algorithm: installation/configuration of additional software solutions, downloading/integrating new databases, and editing the overall processing pipeline. In turn, using GRCh37 imposes a number of limitations on the ability to work with actual databases, as their authors focused on a fresh build GRCh38. If we consider the task of variant calling, it is possible to re-run the mapping and variant calling procedure for the up-to-date GRCh38 if raw FASTQ files are available. This process may require additional computational and time resources, which is often not always possible. As an alternative to this solution, we can consider the most resource-efficient procedure, namely genomic coordinate conversion. Considering the above, as well as the fact that NCBI's Remap Tool for genomic coordinate transformation stopped its work in November 2023, the goal was to review existing programs with similar functionality.

Methods and Algorithms: The following bioinformatic tools for genomic coordinates conversion were tested: CrossMap, LiftoverVcf (Picard), transanno, Assembly Converter (Ensembl). To evaluate the performance of the programs, we used 10 VCF files (≈ 1.5 GB each) that contained an average 4.5 million variants detected through full genome sequencing, as well as the GRCh37_to_GRCh38.chain file required to perform the conversion. Testing was performed in Ubuntu OS (22.04).

Results: The CrossMap was able to convert 98.70 % of all variants in an average of 2 minutes and 17 seconds (± 10 seconds) with minimal use of computational resources. The transanno tool showed similar results: 98.72 % of all variants were converted in 2 minutes 7 seconds (± 6 seconds). The LiftoverVcf (Picard) converted 98.67 % of the variants in 1 minute 15 seconds (± 5 seconds) on average, but the RAM usage reached 12.8 GB at the peak. The web-based tool Assembly Converter (Ensembl) gives the opportunity to run coordinate conversion in a user-friendly graphical interface. It should be noted that the maximum size of a file uploaded to Assembly Converter should not exceed 50 Mb, which imposes limitations on work with full-exome and full-genome sequencing data.

Conclusion: Four tools for genomic coordinates transformation were reviewed. CrossMap and transanno can be called the most universal choices for this task, since these programs are easy to install and do not require large computing power. Assembly Converter is a good option for specialists with little command line experience. LiftoverVcf (Picard) is convenient in cases when you need to perform conversion faster at the expense of additional computational resources.

Список литературы/References

1. Ormond C. et al. Converting single nucleotide variants between genome builds: from cautionary tale to solution. *Briefings Bioinf.* 2021;22(5):bbab069. doi 10.1093/bib/bbab069