BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 7 2014, pages 1006–1007
doi:10.1093/bioinformatics/btt730

Genome analysis

Advance Access publication December 18, 2013

CrossMap: a versatile tool for coordinate conversion between genome assemblies

Hao Zhao[1], Zhifu Sun[2], Jing Wang[1], Haojie Huang[3], Jean-Pierre Kocher[2,*] and Liguo Wang[2,*]

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, [2]Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MN 55905, USA and [3]Department of Biochemistry and Molecular Biology, Mayo Clinic College of Medicine, Rochester, MN 55905, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: Reference genome assemblies are subject to change and refinement from time to time. Generally, researchers need to convert the results that have been analyzed according to old assemblies to newer versions, or vice versa, to facilitate meta-analysis, direct comparison, data integration and visualization. Several useful conversion tools can convert genome interval files in browser extensible data or general feature format, but none have the functionality to convert files in sequence alignment map or BigWig format. This is a significant gap in computational genomics tools, as these formats are the ones most widely used for representing high-throughput sequencing data, such as RNA-seq, chromatin immunoprecipitation sequencing, DNA-seq, etc.

Results: Here we developed CrossMap, a versatile and efficient tool for converting genome coordinates between assemblies. CrossMap supports most of the commonly used file formats, including BAM, sequence alignment map, Wiggle, BigWig, browser extensible data, general feature format, gene transfer format and variant call format.

Availability and implementation: CrossMap is written in Python and C. Source code and a comprehensive user's manual are freely available at: http://crossmap.sourceforge.net/.

Contact: Kocher.JeanPierre@mayo.edu or wang.liguo@mayo.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 28, 2013; revised on December 4, 2013; accepted on December 12, 2013

1 INTRODUCTION

The Human Genome Project brought into the world the first human reference genome that has been widely used as basis for biomedical research. This reference genome has been and is still being refined and improved. The first working draft assembly became available in 2000 with ~150 000 gaps (Lander *et al.*, 2001). Subsequently, several assemblies were released. In 2010, the human reference genome was in its 19th version (hg19 or GRCh37) with remaining 271 gaps (239 Mb), and a new human reference assembly (GRCh38) is to be released at the end of 2013.

Hundreds of thousands of studies have been performed, and sequence data have been analyzed according to different versions of the reference genome. As more accurate versions of the reference genome become available, the compatibility and therefore the ability to integrate new and old genomics results becomes a major barrier. When dealing with such heterogeneous datasets, two approaches can be considered. The first consists of re-aligning all the data to the same reference genome. However, redoing the whole analyses is time-consuming, especially for high-throughput sequencing (HTS) data. In some cases, re-aligning is impossible if the original raw data were not available. The second approach, faster and more convenient, consists of converting genome coordinates between assemblies. Tools like the University of California, Santa Cruz (UCSC) liftOver (Kuhn *et al.*, 2013), Ensembl assembly converter (Spudich *et al.*, 2010) and Galaxy (Giardine *et al.*, 2005) are able to convert plain text files containing genome intervals in browser extensible data (BED), general feature format (GFF)/browser extensible data (GTF) format. However, most HTS data are in BAM, SAM, BigWig format rather than in plain text files because binary compressed files such as BAM or BigWig facilitate data storage, fast transferring, efficient retrieving and visualization. For instance, nearly all processed data generated by ENCODE and TCGA were represented in BAM or BigWig format or both. To the best of our knowledge, no bioinformatics tool is available to convert these "big data" files.

Here we introduce CrossMap, a versatile tool to convert coordinates or annotation files between genome assemblies. It supports the most commonly used file types, including BAM, BED, BigWig, GFF, GTF, SAM, Wiggle (WIG) and VCF (variant cell format) formats. For large plain text file types, such as BED, GFF, GTF and VCF, reading from remote servers and compressed files is supported (see online manual). This versatile tool will greatly facilitate meta-analysis and integration of HTS data.

2 FEATURES AND METHODS

All CrossMap modules require a chain format file, which describes pairwise alignment between two genome assemblies. Chain files for human (*Homo sapiens*) reference genomes (hg17, hg18, hg19) and mouse (*Mus musculus*) reference genomes

*To whom correspondence should be addressed.

(9, 10 mm) are available from our project Web site. Chain files for other species are available from the UCSC genome browser.

2.1 Processing BED files

A BED file is the standard file format used by UCSC to describe genome regions or annotations. It consists of one line per feature, with each line containing 3–12 columns. CrossMap converts BED files with <12 columns to a different assembly by updating the chromosome and genome coordinates only; all other columns remain unchanged. Regions from old assembly mapping to multiple locations to the new assembly will be split. For 12-column BED files, all columns will be updated accordingly except the fourth column (name of BED line), fifth column (score value) and ninth column (RGB value describing the display color). Twelve-column BED files usually define multiple blocks (e.g. exons); if any of the exons fails to map to a new assembly, the whole BED line is skipped and marked as 'Failed'.

The input BED file can be plain text file, compressed file with extension of .gz, .Z, .z, .bz, .bz2 and .bzip2 or even a URL pointing to a remote file (http://, https://and ftp://). The output is in BED format with exactly the same number of columns as the original one.

To evaluate the accuracy of CrossMap, we randomly generated 10 000 genome intervals from human reference genome hg19 and converted them into hg18 using both CrossMap and liftOver. For genome intervals that were successfully converted to hg18, the genome coordinates were exactly the same between liftOver and CrossMap conversion (Supplementary Fig. S1).

2.2 Processing GTF/GFF files

GFF is another plain text file used to describe gene structure. GTF is a refined version of GFF with the first eight fields same as GFF. Plain text, compressed plain text and URLs pointing to remote files are all supported. Only chromosome and genome coordinates are updated. Output file is either GFF or GTF, depending on the input file format.

2.3 Processing BAM/SAM files

Sequence Alignment Map (SAM) format is a generic format for storing sequencing alignments, and BAM is binary and compressed version of SAM (Li *et al*., 2009). Most HTS alignments are stored in SAM/BAM format and many HTS analysis tools work with this format. CrossMap updates chromosomes, genome coordinates and header sections to the target assembly. In addition, all SAM flags indicating mapping status (mapped or not), strand and mapping quality are updated accordingly. For pair-end sequencing, insert size is also recalculated. As records, CrossMap program's name and version, names of the original BAM and the chain file were inserted into the header section. The input BAM file should be sorted and indexed properly using SAMTools (Li *et al*., 2009). BAM output will be sorted and indexed automatically.

The time CrossMap took increased linearly with the size of BAM files (Supplementary Fig. S2) and kept the memory usage constant (~32 Mb) regardless of the file size (Supplementary Table S1).

2.4 Processing BigWig/WIG files

WIG format is useful for displaying continuous data such as GC content and reads intensity of HTS data. BigWig is a self-indexed binary-format WIG file, and has the advantage of supporting random access so that only regions that need to be displayed are retrieved by genome browser, and it dramatically reduces the time needed for data transferring (Kent *et al*., 2010). BigWig format also facilitates large-scale data mining in genomics and epigenomics (e.g. genomic signal aggregation analysis at base pair resolution using a large number of datasets over a large number of genome regions).

WIG data can be in variableStep or fixedStep. Regardless of the input, the output will be always in bedGraph format. bedGraph format is similar to WIG format and can be converted into BigWig format using UCSC wigToBigWig tool. We export files in bedGraph because it is usually much smaller than the file in WIG format, and more importantly, CrossMap internally transforms WIG into bedGraph to increase the running speed.

The output file will be in BigWig format if UCSC's 'wigToBigWig' executable can be found; otherwise, the output file will be in bedGraph format.

2.5 Convert VCF format files

VCF is a flexible and extendable line-oriented text format developed by the 1000 Genome Project (http://www.1000genomes.org/). It is useful for representing single nucleotide variants, indels, copy number variants and structural variants. Chromosomes, coordinates and reference alleles are updated to a new assembly, and all the other fields are not changed.

Funding: This work was funded by the Bioinformatics Program, Center for Individualized Medicine of Mayo Clinic [C4331304 to K.JP., W.L.]; NIH Cancer Center Support Grant (P30 CA016672), Bioinformatics Shared Resources [PP-SR22 to W.J., Z.H.].

Conflict of Interest: none declared.

REFERENCES

Giardine,B. *et al*. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*., **15**, 1451–1455.

Kent,W.J. *et al*. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.

Kuhn,R.M. *et al*. (2013) The UCSC genome browser and associated tools. *Brief. Bioinform*., **14**, 144–161.

Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Li,H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Spudich,G.M. and Fernández-Suárez,X.M. (2010) Touring Ensembl: a practical guide to genome browsing. *BMC Genomics*, **11**, 295.