ruby-htslib and HTS.cr

kojix2 Naohisa Goto

24 May 2022

Summary

We present ruby-htslib and HTS.cr.

Ruby-htslib is the Ruby bindings to HTSlib, a C library for processing high throughput sequencing (HTS) data. It will provide APIs to read and write file formats such as SAM/BAM and VCF/BCF.

In recent years, next-generation sequencing (NGS) technologies for reading DNA and RNA sequences have become popular in the life science field. We will provide a way to manipulate the HTS file formats from Ruby.

- Code of ruby-htslib : https://github.com/kojix2/ruby-htslib
- Code of HTS.cr : https://github.com/bio-cr/hts.cr

Statement of need

The Ruby language is an object-oriented programming language. It is a general-purpose programming language used primarily in the field of web application development. Ruby has also been used in the bioinformatics field, and the BioRuby project (Goto et al. 2010) provides access to many file formats.

In recent years, the volume of biological data generated by sequencing technologies has increased. file formats such as SAM, BAM, CRAM, VCF, and BCF have become widely used with the spread of next-generation sequencers. SAM, BAM, and CRAM are file formats for alignments, while VCF and BCF are file formats for variants. The specifications for these file formats are defined in hts-spec. Samtools and Bcftools were created to manipulate HTS files. And the core part of samtools became a library called HTSlib. We can use HTSLib (Bonfield et al. 2021) to read, write and query HTS files.

However, the ways to manipulate HTS files from the Ruby language have been limited. BioRuby does not include a module to work with HTS files. Bio-Samtools (Etherington, Ramirez-Gonzalez, and MacLean 2015) was originally developed as a samtools binding. However, the binding stopped working while samtools was ported to htslib and now calls samtools directly from standard streams using open3.

Ruby-htslib is a binding for htslib. It provides access to comprehensive HTS files from the Ruby language. This allows the Ruby language to analyze genomes and create applications.

Benchmark

(Pedersen and Quinlan 2018)

Examples

Reference

Bonfield, James K., John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M. Davies. 2021. "HTSlib: C Library for Reading/Writing High-Throughput Sequencing Data." *GigaScience* 10 (2): giab007. https://doi.org/10.1093/gigascience/giab007.

Etherington, Graham J., Ricardo H. Ramirez-Gonzalez, and Dan MacLean. 2015. "Bio-Samtools 2: A Package for Analysis and Visualization of Sequence and Alignment Data with SAMtools in Ruby: Fig. 1." *Bioinformatics* 31 (15): 2565–67. https://doi.org/10.1093/bioinformatics/btv178.

Goto, Naohisa, Pjotr Prins, Mitsuteru Nakao, Raoul Bonnal, Jan Aerts, and Toshiaki Katayama. 2010. "BioRuby: Bioinformatics Software for the Ruby Programming Language." *Bioinformatics* 26 (20): 2617–19. https://doi.org/10.1093/bioinformatics/btq475.

Pedersen, Brent S, and Aaron R Quinlan. 2018. "Hts-Nim: Scripting High-Performance Genomic Analyses." Edited by Inanc Birol. *Bioinformatics* 34 (19): 3387–89. https://doi.org/10.1093/bioinformatics/bty358.