BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 3 2015, pages 413–415
doi:10.1093/bioinformatics/btu636

Genome analysis

Advance Access publication October 1, 2014

WhopGenome: high-speed access to whole-genome variation and sequence data in R

Ulrich Wittelsbürger[1], Bastian Pfeifer[1] and Martin J. Lercher[1,2,*]

[1]Institute for Computer Science, Heinrich Heine University, D-40255 Düsseldorf, Germany and [2]Cluster of Excellence on Plant Sciences CEPLAS, D-40255 Düsseldorf, Germany

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The statistical programming language R has become a *de facto* standard for the analysis of many types of biological data, and is well suited for the rapid development of new algorithms. However, variant call data from population-scale resequencing projects are typically too large to be read and processed efficiently with R's built-in I/O capabilities. WhopGenome can efficiently read whole-genome variation data stored in the widely used variant call format (VCF) file format into several R data types. VCF files can be accessed either on local hard drives or on remote servers. WhopGenome can associate variants with annotations such as those available from the UCSC genome browser, and can accelerate the reading process by filtering loci according to user-defined criteria. WhopGenome can also read other Tabix-indexed files and create indices to allow fast selective access to FASTA-formatted sequence files.

Availability and implementation: The WhopGenome R package is available on CRAN at http://cran.r-project.org/web/packages/WhopGenome/. A Bioconductor package has been submitted.

Contact: lercher@cs.uni-duesseldorf.de

Received on May 13, 2014; revised on August 29, 2014; accepted on September 18, 2014

1 INTRODUCTION

Population-scale whole-genome sequencing projects produce information on single-nucleotide polymorphisms (SNPs), InDels and structural variations across thousands of individuals. These projects commonly use the variant call format (VCF) (1000 Genomes Project Analysis Group, 2011) text files for data storage. The resulting files often contain millions of variant sites and may fill tens of gigabytes. The environment for statistical computing R (R Core Team, 2013) has established itself as a *de facto* standard for general statistics and for the analysis of different types of sequencing data, and has efficient functions to process large vectorized data. However, routinely reading gigabyte-sized VCF files into R is not realistic with R's built-in I/O capabilities.

VCF files are typically compressed and then indexed with Tabix (Li, 2011). Tabix produces an index file for appropriately formatted data files; the index can be used to quickly locate, decompress and extract selected portions of the data. Although several R packages are capable of reading VCF files [VariantAnnotation (Obenchain *et al.*, 2014), seqminer (http://cran.r-project.org/web/packages/seqminer/), Rplinkseq (https://atgu.mgh.harvard.edu/plinkseq/)], these lack the desirable speed, ease of use or completeness of support for Tabix files. Further, these implementations post-process the text returned by Tabix in R, which incurs a sizeable overhead especially for repeated and large-scale processing.

Here, we present WhopGenome, an R package for fast, straightforward and flexible processing of genomic variation data in VCF format. WhopGenome is also capable of compressing files with BGZF and indexing any suitably formatted file with Tabix, allowing efficient selective access. Indexing is possible on any data organized and sorted into entries uniquely identifiable by index pairs: a group name (e.g. a chromosome) and a number (e.g. a chromosomal position). With WhopGenome's generic Tabix interface, users can process, for example, GFF or BED files to access them efficiently from within R.

The same selective access functionality exists also for FASTA files through WhopGenome's interface to FaIdx (the indexing solution included in samtools) (1000 Genome Project Data Processing Subgroup, 2009). Using this interface to preprocess FASTA files facilitates quick selective retrieval of DNA or amino acid sequence regions. WhopGenome can thus efficiently integrate information from associated sequence, genome annotation and population-scale variation files for a given chromosomal region for joint processing.

2 FEATURES AND IMPLEMENTATION

All indexed data files can reside either on local hard disks or on remote HTTP or FTP servers. Thus, WhopGenome can, for example, selectively read data from the 1000 Genomes Project directly from the NCBI servers (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release).

WhopGenome provides functionality to relate genomic loci to annotation data. A wide range of different annotation data types is accessible through the UCSC Genome Browser (Kent *et al.*, 2002), the AmiGO Gene Ontology database (Carbon *et al.*, 2009) and through BioConductor's (Gentleman *et al.*, 2004) org.XX.eg.db annotation packages. WhopGenome includes user-friendly interfaces to the UCSC Genome Browser and the AmiGO servers, vastly simplifying the construction of the necessary SQL queries and the communication with the remote servers. WhopGenome also provides a comfortable way to select and download the BioConductor annotation packages.

*To whom correspondence should be addressed.

Table 1. Comparison overview

	WhopGenome	SeqMiner	VariantAnnotation
Average time[a]			
100 000 single SNPs	0 min 32 s	15 min 34 s	1 h 56 min 12 s
Matrix of 1000	0 min 6 s	0 min 29 s	1 min 27 s
Average lines/second[a]			
100 000 single SNPs	3126.66	107.02	14.34
Matrix of 1000	16 677.23	3460.28	1144.00
Estimated time for entire file[b]			
Reading single SNPs	15 min 47 s	7 h 48 min 18 s	57 h 21 min 53 s
Matrix of 1000	2 min 58 s	14 min 16 s	43 min 9 s
Pre-filtering	Y	N	N
Result formats[c] R,V,M,L	Y,Y,Y,N	Y,N,N,Y	N,N,Y,Y
Read via HTTP/FTP	Y	N	Y
Create indices	Y	Y[d]	N[e]

Timings ran on an Intel Core-i7 3.5 GHz, 16 GB RAM, Linux kernel 3.11 using R 3.0.1.
A minimal stand-alone C++ program takes 54.9 s to read the entire reference file, indicating an average rate of 53 952 lines / s as the performance baseline.
All timings rounded to nearest full second.
[a]Average over five runs.
[b]Estimate using factor of 29.62 182 = 2 962 182 total lines in file / 100 000 in measurement.
[c]Supported result formats: Raw string, Vector, SNP Matrix, List.
[d]SeqMiner can create indices only on compressed files; WhopGenome can compress, too.
[e]Not specifically for VCF files.

To link genomic variation to pedigree data, WhopGenome includes support for .PED files (a simple text-based table format used, e.g. by PLINK). Users can load these data into a matrix, modify it, save it back and locate individuals with certain family relationships. This is mainly useful for selecting samples or correlating them with phenotypes, populations or other information.

When reading from VCF files with WhopGenome, a typical workflow would be as follows. The function VCF_open() creates a handle to the VCF file. This handle is required to select samples (individuals), genomic regions and filtering steps, as well as for reading data.

Users can choose to get their results in a variety of R data types. Besides reading each data field independently, it is also possible to read only information on single-nucleotide polymorphisms (SNPs) and store the data fields for each SNP in a vector. Especially useful for sliding window analyses are the matrix variants, which can return SNP genotypes in four different representations, either numeric or textual. To maximize speed gains, we wrote a dedicated read function for each result format. If specific areas of research would benefit from additional data representations in R, we will implement these in future versions of WhopGenome.

After setting a region by specifying a chromosome (or contig) and start and end positions, the next read call will return data for the first variant within that region. Active prefilters exclude lines depending on a list of user-defined rules. Rules are specified with simple function calls in R, but are run in compiled C++.

Each rule tests a specific property, either the value of a column (e.g. QUAL column's phred score) or the value of a key-value pair in a column (e.g. INFO column's allele frequency key). The tests can compare numerical values to reference values or against ranges, or test for the presence of a key (e.g. H2 in INFO indicates that a variation is found in HapMap2).

The .PED file support for pedigrees, Gene Ontology queries, UCSC Genome Browser database queries and Bioconductor genome annotation is implemented in R. All time-critical code is written in C/C++. To avoid losing time by allocating memory, many read functions expect an R variable as a parameter in which to store the results. This improves speed dramatically especially if the data are read into matrices.

3 EVALUATION

We compared WhopGenome in terms of speed, features and ease of use with two other R packages that make use of Tabix: SeqMiner and VariantAnnotation (Table 1). We did not make a comparison with Rplinkseq, as Rplinkseq could not be compiled without manual code changes and because using Rplinkseq requires extensive manual interaction with the external PLINK software (Purcell *et al.*, 2007).

As reference file, we chose the 1000 Genomes Project's chromosome 1 consensus VCF file, describing >2.9 million variants in 1094 individuals, stored in 49 GB of text, compressed down to 1.4 GB. SeqMiner requires additional annotation in the VCF files, which is not present in the 1000 Genomes Project files. We thus ran all benchmarks on the same, preprocessed file for better comparability (to preprocess the input file for SeqMiner, we needed to uncompress the original file and install additional software).

Although all three packages rely on the Tabix library, their usage and feature sets differ substantially. For both benchmark types, VariantAnnotation was slowest, whereas WhopGenome was the fastest. All programs provide matrix representations of

genotypes, but only WhopGenome offers four alternative genotype codings in matrix form. Also, its prefiltering capabilities have no direct equivalent in the other packages. With regards to the learning curve, we consider our solution to be easier to understand than VariantAnnotation, while SeqMiner's limited feature set makes its usage somewhat simpler, but also much less powerful.

The VCF functionality of WhopGenome has been successfully used by the population genomics software PopGenome (Pfeifer *et al.*, 2014), which implements a broad range of population genetics analyses for individual loci, sliding windows and genomic feature sets such as exons.

Besides its ability to efficiently read VCF and other Tabix-indexed files, WhopGenome can also index and access FASTA-formatted sequence files efficiently. With its auxiliary feature set covering pedigree, genome annotation and fast prefiltering, we expect WhopGenome to substantially accelerate the development and application of genomic and population genomic analyses in R.

ACKNOWLEDGEMENTS

WhopGenome makes use of Tabix and FaIdx by Heng Li, of BGZF written by Bob Handsaker and modified by Heng Li and of zlib by Jean-loup Gailly and Mark Adler.

Funding: This work was supported by the German Research Foundation [DFG grants EXC 1028 and CRC 680 to M.J.L.].

Conflict of interest: none declared.

REFERENCES

1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

1000 Genomes Project Analysis Group. (2011) The variant call format and vcftools. *Bioinformatics*, **27**, 2156–2158.

Carbon,S. *et al.* (2009) Amigo: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Li,H. (2011) Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics*, **27**, 718–719.

Obenchain,V. *et al.* (2014) Variantannotation: a bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, **30**, 2076–2078.

Pfeifer,B. *et al.* (2014) Popgenome: an efficient swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.*, **31**, 1929–1936.

Purcell,S. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

R Core Team. (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.