BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 14 2014, pages 2076–2078
doi:10.1093/bioinformatics/btu168

Sequence analysis

Advance Access publication March 28, 2014

VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants

Valerie Obenchain[1,*], Michael Lawrence[2], Vincent Carey[3], Stephanie Gogarten[4], Paul Shannon[1] and Martin Morgan[1]

[1]Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, [2]Bioinformatics and Computational Biology, Genentech, South San Francisco, CA 94080, [3]Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115 and [4]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: VariantAnnotation is an R / Bioconductor package for the exploration and annotation of genetic variants. Capabilities exist for reading, writing and filtering variant call format (VCF) files. VariantAnnotation allows ready access to additional R / Bioconductor facilities for advanced statistical analysis, data transformation, visualization and integration with diverse genomic resources.

Availability and implementation: This package is implemented in R and available for download at the Bioconductor Web site (http://bioconductor.org/packages/2.13/bioc/html/VariantAnnotation.html). The package contains extensive help pages for individual functions and a 'vignette' outlining typical work flows; it is made available under the open source 'Artistic-2.0' license. Version 1.9.38 was used in this article.

Contact: vobencha@fhcrc.org

Received on May 2, 2013; revised on February 20, 2014; accepted on March 25, 2014

Major products of DNASeq and other high-throughput experiments are catalogs of called variants [e.g. single-nucleotide polymorphisms (SNPs), indels] saved in variant call format (VCF) (The 1000 Genomes Project Consortium, 2012) files. VCF files contain data lines with position and genotype information on samples. VariantAnnotation enables users to explore these data in R.

1 AVAILABLE FUNCTIONALITY

Important operations available with the VariantAnnotation package are summarized in Table 1; we illustrate these operations using a subset of chr7 breast cancer variants for a tumor/normal pair (Drmanac and Sparks, 2010).

1.1 Reading, writing and filtering

readVcf reads data from a VCF file into a VCF R object. Genomic locations are stored as a GRanges object, with REF, ALT, FILTER, QUALITY and INFO fields as metadata columns. The GRanges object is a convenient format for manipulating range data and is compatible with extensive and well-developed Bioconductor (Gentleman *et al.*, 2004) tools for discovering overlaps and matching between ranges (Lawrence *et al.*, 2013). Genotype data are parsed into arrays and stored in reference classes to avoid multiple data copies. A VCF object can be written out as a tabix-indexed (Heng, 2010) VCF file with writeVcf.

One strategy for processing large tabix-indexed files is to use scanVcfHeader to identify INFO or FORMAT fields of interest, formulate range-based queries and load the data with readVcf. Memory use can be tuned by setting a yieldSize and iterating over the data in chunks.

```
> library(VariantAnnotation)
> fl <- system.file(''extdata", ''chr7-sub.vcf.gz",
+ package="VariantAnnotation'')
> hdr <- info(scanVcfHeader(fl)) ##'info' fields
```

Table 1. Example functions available in VariantAnnotation

Function	Description
Reading, writing and filtering	
scanVcfHeader	Retrieve information about file content
ScanVcfParam	Select fields to input
readVcf	Read a VCF file into an R object
readGeno, readInfo, readGT	Read a single field into an R object
writeVcf	Write an R object to a VCF file
filterVcf	Filter one VCF file to another
Annotation	
locateVariants	Identify variants overlapping ranges
predictCoding	Predict amino acid consequences
summarizeVariants	By range and sample
SNPs	
genotypeToSnpMatrix	Genotypes as SnpMatrix objects
snpSummary	Counts and distribution statistics
Manipulation	
expand	Convert R VCF representations
cbind, rbind	Combine variants or samples

*To whom correspondence should be addressed.

```
> param <- ScanVcfParam(info="CGA_BF", geno="AD'')
> tabix <- TabixFile(fl, yieldSize=100000)
> vcf <- readVcf(tabix, ''hg19", param) ## chunk 1
```

readInfo, readGeno and readGT retrieve individual fields as standard R objects. filterVcf identifies records satisfying predefined and *ad hoc* criteria, creating a new VCF file.

1.2 Annotating and transforming variants

locateVariants associates variants with coding, intron, splice site, promoter, UTR or intergenic regions.

```
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)
> txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
> vcf <- renameSeqlevels(vcf, c('7'='chr7'))
> loc <- locateVariants(vcf, txdb, IntronVariants())
```

The gene, transcript and coding region identifiers provided in the output can be used with other Bioconductor resources to map to additional identifiers such as protein families database (PFAM) or gene ontology project (GO).

```
> library(org.Hs.eg.db)
> select(org.Hs.eg.db, loc$GENEID, c(''PFAM", ''GO''))
```

predictCoding computes amino acid coding changes for non-synonymous variants that overlap coding regions. Reference sequences are retrieved from a BSgenome package or FASTA file. Variant sequences are constructed by substituting or inserting variant alleles into the reference sequence. Custom genomes can be imported as a Transcriptdb object with one of the makeTranscriptDb functions available in the GenomicFeatures package.

```
> library(BSgenome.Hsapiens.UCSC.hg19)
> predictCoding(vcf, txdb, Hsapiens)
```

genotypeToSnpMatrix performs probability-based encoding of the genotype calls in a VCF object to create an SnpMatrix object for use in downstream packages. snpSummary provides counts and distribution statistics.

1.3 Integration and comparison with other resources

VariantAnnotation offers highly flexible tools to interrogate and transform VCF files into R objects for exploration and analysis. In contrast to programs such as VCFtools (Danecek *et al*., 2011) or PLINK/SEQ, VariantAnnotation provides an interactive environment for integrated portable analysis and methods development. The ensemblVEP package is an interface to the VEP (McLaren *et al*., 2010) tool, while functions in VariantAnnotation allow close integration with SNP analysis routines in packages such as snpStats. I/O capabilities are compatible with upstream alignment and variant calling R packages such as gmapR and VariantTools, as well as VCFs produced by VarScan (Koboldt *et al*., 2012), GATK (McKenna *et al*., 2010), etc. The ability to transform and output VCF subsets enables creation of files for use in tools such as ANNOVAR (Wang *et al*., 2010). R VCF objects can be visualized with packages such as ggbio (Yin *et al*., 2012).

Table 2. VariantAnnotation and Rplinkseq runtimes (min)

Function	Range (All fields)	Range (Select fields)	Iterate
Rplinkseq			
load.vcf	359.8	NA	NA
var.fetch	291.8	NA	NA
meta.fetch	NA	120.9	NA
var.iterate	NA	NA	1583.1
VariantAnnotation			
scanVcf	359.1	35.5	50.3

VariantAnnotation has good performance relative to other R tools operating on VCF files, e.g. Rplinkseq (http://atgu.mgh.harvard.edu/plinkseq/r-intro.shtml), as illustrated using a compressed indexed VCF (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/ALL.chr22.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz) (494 328 records, 1092 samples and 22 INFO and 3 genotype fields). Testing was done on a 64-bit 387 Gb 2.90 GHz Linux server; test script is available in inst/scripts/ of the built tarball or scripts/of the installed package. Runtimes for four Rplinkseq functions and scanVcf from VariantAnnotation are summarized in Table 2. NA values indicate the function could not perform the abstraction.

A range of 63 088 records and two INFO and two genotype fields were arbitrarily chosen for testing. VariantAnnotation outperformed load.vcf when reading the range with all fields and meta.fetch when reading in specific INFO and genotype fields. VariantAnnotation was ~30× faster than Rplinkseq when iterating over all records in the file. Input times for scanVcf scale linearly with the number of variants or samples.

2 CONCLUSIONS

This Note introduces the VariantAnnotation package to flexibly interrogate, annotate and transform VCF files. The package integrates with Bioconductor packages for advanced SNP and variant analysis, gene and genome annotation and rich tools for range-based queries. VariantAnnotation is performant compared with other R solutions and scales to handle large files with reasonable memory requirements. Read/write capabilities allow ready integration with third party software.

Funding: National Human Genome Research Institute of the National Institutes of Health (U41HG004059 to M.M.).

Conflict of Interest: none declared.

REFERENCES

Danecek,P. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Drmanac,R. and Sparks,A.B. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.

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

Heng,L. (2010) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.

Koboldt,D. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Lawrence,M. *et al.* (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.

McKenna,A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

McLaren,W. *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *BMC Bioinformatics*, **26**, 2069–2070.

The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Yin,T. *et al.* (2012) ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.*, **13**, R77.