BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 15 2011, pages 2156–2158
doi:10.1093/bioinformatics/btr330

Sequence analysis

Advance Access publication June 7, 2011

The variant call format and VCFtools

Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, [5]Department of Biology, Boston College, MA 02467, [6]National Institutes of Health National Center for Biotechnology Information, MD 20894, USA and [7]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Associate Editor: John Quackenbush

ABSTRACT

Summary: The variant call format (VCF) is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, together with rich annotations. VCF is usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome. The format was developed for the 1000 Genomes Project, and has also been adopted by other projects such as UK10K, dbSNP and the NHLBI Exome Project. VCFtools is a software suite that implements various utilities for processing VCF files, including validation, merging, comparing and also provides a general Perl API.

Availability: http://vcftools.sourceforge.net

Contact: rd@sanger.ac.uk

Received on October 28, 2010; revised on May 4, 2011; accepted on May 28, 2011

1 INTRODUCTION

One of the main uses of next-generation sequencing is to discover variation among large populations of related samples. Recently, a format for storing next-generation read alignments has been standardized by the SAM/BAM file format specification (Li *et al*., 2009). This has significantly improved the interoperability of next-generation tools for alignment, visualization and variant calling. We propose the variant call format (VCF) as a standardized format for storing the most prevalent types of sequence variation, including SNPs, indels and larger structural variants, together with rich annotations. The format was developed with the primary intention to represent human genetic variation, but its use is not restricted to diploid genomes and can be used in different contexts as well. Its flexibility and user extensibility allows representation of a wide variety of genomic variation with respect to a single reference sequence.

Although generic feature format (GFF) has recently been extended to standardize storage of variant information in genome variant format (GVF) (Reese *et al*., 2010), this is not tailored for storing information across many samples. We have designed the VCF format to be scalable so as to encompass millions of sites with genotype data and annotations from thousands of samples. We have adopted a textual encoding, with complementary indexing, to allow easy generation of the files while maintaining fast data access. In this article, we present an overview of the VCF and briefly introduce the companion VCFtools software package. A detailed format specification and the complete documentation of VCFtools are available at the VCFtools web site.

2 METHODS

2.1 The VCF

2.1.1 Overview of the VCF A VCF file (Fig. 1a) consists of a header section and a data section. The header contains an arbitrary number of meta-information lines, each starting with characters '##', and a TAB delimited field definition line, starting with a single '#' character. The meta-information header lines provide a standardized description of tags and annotations used in the data section. The use of meta-information allows the information stored within a VCF file to be tailored to the dataset in question. It can be also used to provide information about the means of file creation, date of creation, version of the reference sequence, software used and any other information relevant to the history of the file. The field definition line names eight mandatory columns, corresponding to data columns representing the chromosome (CHROM), a 1-based position of the start of the variant (POS), unique identifiers of the variant (ID), the reference allele (REF), a comma separated list of alternate non-reference alleles (ALT), a phred-scaled quality score (QUAL), site filtering information (FILTER) and a semicolon separated list of additional, user extensible annotation (INFO). In addition, if samples are present in the file, the mandatory header columns are followed by a FORMAT column and an arbitrary number of sample IDs that define the samples included in the VCF file. The FORMAT column is used to define the information contained within each subsequent genotype column, which consists of a colon separated list of fields. For example, the FORMAT field GT:GQ:DP in the fourth data entry of Figure 1a indicates that the subsequent entries contain information regarding the genotype, genotype quality and read depth for each sample. All data lines are TAB delimited and the number of fields in each data line must match the number of fields in the header line. It is strongly recommended that all annotation tags used are declared in the VCF header section.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

‡http://www.1000genomes.org

(a) VCF example

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                FORMAT     SAMPLE1  SAMPLE2
1       1   .      ACG  A,AT    40   PASS    .                   GT:DP      1/1:13   2/2:29
1       2   .      C    T,CT    .    PASS    H2;AA=T             GT         0|1      2/2
1       5   rs12   A    G       67   PASS    .                   GT:DP      1|0:16   2/2:20
X     100   .      T    <DEL>   .    PASS    SVTYPE=DEL;END=299  GT:GQ:DP   1:12:.   0/0:20:36
```

(Header: ##fileformat=VCFv4.1 … #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2; Body: 1 1 . ACG A,AT 40 PASS . GT:DP 1/1:13 2/2:29; 1 2 . C T,CT . PASS H2;AA=T GT 0|1 2/2; 1 5 rs12 A G 67 PASS . GT:DP 1|0:16 2/2:20; X 100 . T <DEL> . PASS SVTYPE=DEL;END=299 GT:GQ:DP 1:12:. 0/0:20:36)

(b) SNP

Alignment
```
1234
ACGT
ATGT
 ^
```

VCF representation

POS	REF	ALT
2	C	T

(c) Insertion

```
12345
AC-GT
ACTGT
  ^
```

POS	REF	ALT
2	C	CT

(d) Deletion

```
1234
ACGT
A--T
 ^^
```

POS	REF	ALT
1	ACG	A

(e) Replacement

```
1234
ACGT
A-TT
 ^^
```

POS	REF	ALT
1	ACG	AT

(f) Large structural variant

Alignment
```
   100       110       120       290       300
    .         .         .         .         .
ACGTACGTACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC
ACGT------------------------[...]----------GTAC
```

VCF representation

POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=299

(g) Resolving ambiguity

Alignment	*Possible representation*			*Possible representation*			*Recommended VCF representation*		
1234567890	POS	REF	ALT	POS	REF	ALT	POS	REF	ALT
TTTCCCTCTA	1	TTTCCCTCT	CTTACCTA	1	T	C	1	T	C
CTTACCT--A				4	C	A	4	C	A
^ ^ ^^				7	TCT	T	5	CCT	C

Fig. 1. (**a**) Example of valid VCF. The header lines ##fileformat and #CHROM are mandatory, the rest is optional but strongly recommended. Each line of the body describes variants present in the sampled population at one genomic position or region. All alternate alleles are listed in the ALT column and referenced from the genotype fields as 1-based indexes to this list; the reference haplotype is designated as 0. For multiploid data, the separator indicates whether the data are phased (|) or unphased (/). Thus, the two alleles C and G at the positions 2 and 5 in this figure occur on the same chromosome in SAMPLE1. The first data line shows an example of a deletion (present in SAMPLE1) and a replacement of two bases by another base (SAMPLE2); the second line shows a SNP and an insertion; the third a SNP; the fourth a large structural variant described by the annotation in the INFO column, the coordinate is that of the base before the variant. (**b–f**) Alignments and VCF representations of different sequence variants: SNP, insertion, deletion, replacement, and a large deletion. The REF columns shows the reference bases replaced by the haplotype in the ALT column. The coordinate refers to the first reference base. (**g**) Users are advised to use simplest representation possible and lowest coordinate in cases where the position is ambiguous.

2.1.2 Conventions and reserved keywords The VCF specification includes several common keywords with standardized meaning. The following list gives some examples of the reserved tags.

Genotype columns:

- GT, genotype, encodes alleles as numbers: 0 for the reference allele, 1 for the first allele listed in ALT column, 2 for the second allele listed in ALT and so on. The number of alleles suggests ploidy of the sample and the separator indicates whether the alleles are phased ('|') or unphased ('/') with respect to other data lines (Fig. 1).
- PS, phase set, indicates that the alleles of genotypes with the same PS value are listed in the same order.
- DP, read depth at this position.
- GL, genotype likelihoods for all possible genotypes given the set of alleles defined in the REF and ALT fields.
- GQ, genotype quality, probability that the genotype call is wrong under the condition that the site is being variant. Note that the QUAL column

gives an overall quality score for the assertion made in ALT that the site is variant or no variant.

INFO column:

- DB, dbSNP membership;
- H3, membership in HapMap3;
- VALIDATED, validated by follow-up experiment;
- AN, total number of alleles in called genotypes;
- AC, allele count in genotypes, for each ALT allele, in the same order as listed;
- SVTYPE, type of structural variant (DEL for deletion, DUP for duplication, INV for inversion, etc. as described in the specification);
- END, end position of the variant;
- IMPRECISE, indicates that the position of the variant is not known accurately; and
- CIPOS/CIEND, confidence interval around POS and END positions for imprecise variants.

Missing values are represented with a dot. For practical reasons, the VCF specification requires that the data lines appear in their chromosomal order. The full format specification is available at the VCFtools web site.

2.1.3 Variation types VCF is flexible and allows to express virtually any type of variation by listing both the reference haplotype (the REF column) and the alternate haplotypes (the ALT column). This permits redundancy such that the same event can be expressed in multiple ways by including different numbers of reference bases or by combining two adjacent SNPs into one haplotype (Fig. 1g). Users are advised to follow recommended practice whenever possible: one reference base for SNPs and insertions, and one alternate base for deletions. The lowest possible coordinate should be used in cases where the position is ambiguous. When comparing or merging indel variants, the variant haplotypes should be reconstructed and reconciled, such as in the Figure 1g example, although the exact nature of the reconciliation can be arbitrary. For larger, more complex, variants, quoting large sequences becomes impractical, and in these cases the annotations in the INFO column can be used to describe the variant (Fig. 1f). The full VCF specification also includes a set of recommended practices for describing complex variants.

2.1.4 Compression and indexing Given the large number of variant sites in the human genome and the number of individuals the 1000 Genomes Project aims to sequence (Durbin *et al*., 2010), VCF files are usually stored in a compact binary form, compressed by bgzip, a program which utilizes the zlib-compatible BGZF library (Li *et al*., 2009). Files compressed by bgzip can be decompressed by the standard gunzip and zcat utilities. Fast random access can be achieved by indexing genomic position using tabix, a generic indexer for TAB-delimited files. Both programs, bgzip and tabix, are part of the samtools software package and can be downloaded from the SAMtools web site (http://samtools.sourceforge.net).

2.2 VCFtools software package

VCFtools is an open-source software package for parsing, analyzing and manipulating VCF files. The software suite is broadly split into two modules. The first module provides a general Perl API, and allows various operations to be performed on VCF files, including format validation, merging, comparing, intersecting, making complements and basic overall statistics. The second module consists of C++ executable primarily used to analyze SNP data in VCF format, allowing the user to estimate allele frequencies, levels of linkage disequilibrium and various Quality Control metrics. Further details of VCFtools can be found on the web site (http://vcftools.sourceforge.net/), where the reader can also find links to alternative tools for VCF generation and manipulation, such as the GATK toolkit (McKenna *et al*., 2010).

3 CONCLUSIONS

We describe a generic format for storing the most prevalent types of sequence variation. The format is highly flexible, and can be adapted to store a wide variety of information. It has already been adopted by a number of large-scale projects, and is supported by an increasing number of software tools.

Funding: Medical Research Council, UK; British Heart Foundation (grant RG/09/012/28096); Wellcome Trust (grants 090532/Z/09/Z and 075491/Z/04); National Human Genome Research Institute (grants 54 HG003067, R01 HG004719 and U01 HG005208); Intramural Research Program of the National Institutes of Health, the National Library of Medicine.

Conflict of Interest: none declared.

REFERENCES

Durbin,R.M. *et al*. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

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

Reese,M.G. *et al*. (2010) A standard variation file format for human genome sequences. *Genome Biol.*, **11**, 20796305.