Sequence analysis

The variant call format and VCFtools

Petr Danecek1,†, Adam Auton2,†, Goncalo Abecasis3, Cornelis A. Albers4, Eric Banks4, Mark A. DePristo5, Robert E. Handsaker4, Gerton Lunter2, Gabor T. Marth2, Stephen T. Sherry6, Gilean McVean2,7, Richard Durbin1,∗ and 1000 Genomes Project Analysis Group‡

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

Associate Editor: John Quackenbush

© The Author(s) 2011. Published by Oxford University Press.

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 (CHR), 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
Variant call format

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

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.
The full format specification is available at the VCFtools web site.

### 2.1.3 VCF tools 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


### 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 tabsix, a generic indexer for TAB-delimited files. Both programs, bgzip and tabsix, are part of the samtools software package and can be downloaded from the SAMtools web site (http://samtools.sourceforge.net).

The full specification requires that the data lines appear in their chromosomal order. Missing values are represented with a dot. For practical reasons, the VCF specification requires that the DNA sequences in the INFO column are aligned to a specific reference genome. Where this is ambiguous, a dot is used. 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 tabsix, a generic indexer for TAB-delimited files. Both programs, bgzip and tabsix, are part of the samtools software package and can be downloaded from the SAMtools web site (http://samtools.sourceforge.net).

### 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 tabsix, a generic indexer for TAB-delimited files. Both programs, bgzip and tabsix, are part of the samtools software package and can be downloaded from the SAMtools web site (http://samtools.sourceforge.net).