Variant Call Format, Binary Variant Call Format and VCFtools

Anthony Marcketta and Adam Auton

Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461

One of the main uses of next-generation sequencing is to discover variation amongst large populations of related samples. The Variant Call Format (VCF) has been growing in popularity as a standardized format for storing sequence variations, including SNPs, indels and larger structural variants, together with rich annotations. The format was developed for the 1000 Genomes Project, and has also been adopted by other projects such as UK10K, dbSNP, or the NHLBI Exome Project.

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. However the size of these files is rapidly increasing as more and more individuals are being sequenced and called for genetic variation. Large scale VCF files with thousands of samples are increasingly common, can be hundreds of GBs, and can be very slow to process. There is a pressing need to keep the size of these files small, while increasing the accessibility of the data if possible. The BCF (binary variant call format) has been proposed to do exactly those things. This format will contain the same information as VCF, but represent it in binary values instead of ASCII text values which are slow to parse. These binary files will also be compressed using block compression so they can be indexed for quick access.

VCFtools is a software suite that implements various utilities for processing VCF and BCF files. The goal of the VCFtools project is to provide users with an easy to use command-line interface for investigating their genetic variation files in a way that minimizes memory usage so it can be run on an average computer. The tools provided will be used to mainly to summarize data, run calculations on data, filter through data, and convert data into other useful file formats. Many of the analyses it can perform are listed below. The VCF and BCF specifications as well as the VCFtools source code are available from http://vcftools.sourceforge.net

Example VCF file

```
##fileformat=VCFv4.1
##fileDate=20120707
##source=VCFtools
##reference=NCBI36
##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 (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FILTER=<ID=q10,Description="Quality below 10”>
##CONTIG=<ID=19,length=59128983,assembly=b37,species="Homo Sapiens">
##CONTIG=<ID=20,length=63025520,assembly=b37,species="Homo Sapiens">
##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
19       12430 rs1   ACG     A,AT  100   PASS     .        GT:DP      1/2:13      0/0:29
19       29835 rs59  C       T,CT   35   PASS     H2;AA=T  GT:GQ      0|1:100     2/2:70
20       19453 .     A       G       5   q10      .        GT:GQ      1|0:77      1/1:95
20       89283 .     T       A     100   PASS     END=300  GT:GQ:DP   1/1:12:3    0/0:20
```

Optional information and annotations about the file in general

Meta-data to explain the annotations used in body of the file

Mandatory Header for the essential information and the individual names

Entries to represent the information at each variant site for each individual

Some useful functionality found in VCFtools

Calculations:

- Calculate per site allele frequencies
- Calculate mean coverage depth of each individual
- Measure the heterozygosity of each individual
- Report a Hardy-Weinberg p-value for every site
- Determine Linkage Disequilibrium statistics for each pair of sites
- Estimate levels of nucleotide diversity (pi)
- Calculate F_{ST} between multiple populations
- Locate differences between VCF files

Filter data by:

- Position of sites
- Lists of SNP names
- Variant type (SNP or indel)
- Filter or info tags in the file
- Site quality
- Allele frequency
- P-value
- Ploidy
- Amount of missing data
- Coverage depth

Example VCFtools Usage

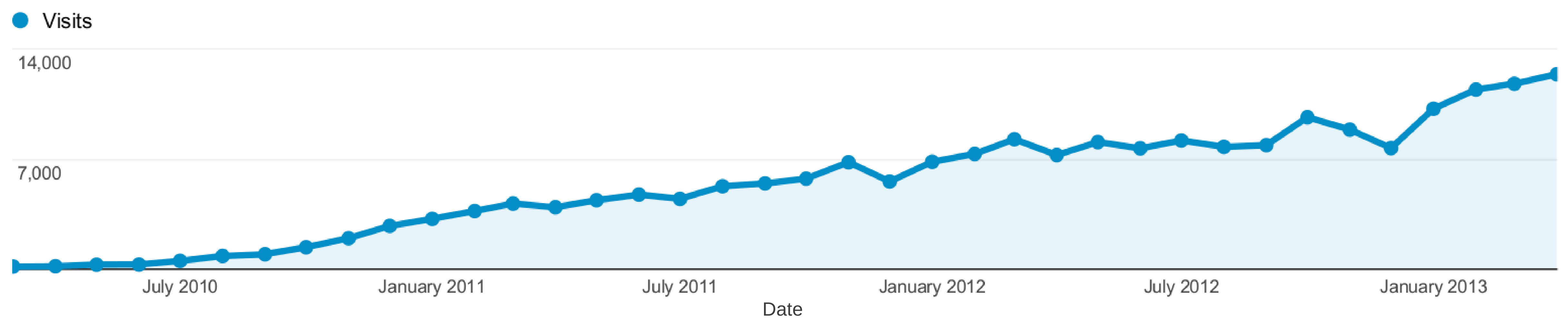
```
./vcftools  --gzvcf input_file.vcf.gz  --freq  --chr 1  --out chr1_analysis

./vcftools  --vcf input_file.vcf  --remove-indels  --recode  --out SNPs_only

./vcftools  --gzvcf input_file1.vcf.gz  --gzdiff input_file2.vcf.gz  --out in1_v_in2

./vcftools  --gzvcf input_file1.vcf.gz --remove-filtered-all --recode-to-stream  |
     gzip  >  output_PASS_only.vcf.gz
```

Available in latest version from SVN repository – BCF functionality

```
./vcftools  --bcf input_file.bcf  --het  --geno 1.0  --out output_noMissing
```

VCFtools is increasing in popularity and has worldwide visibility

VCFtools Sourceforge Page Visits per Month

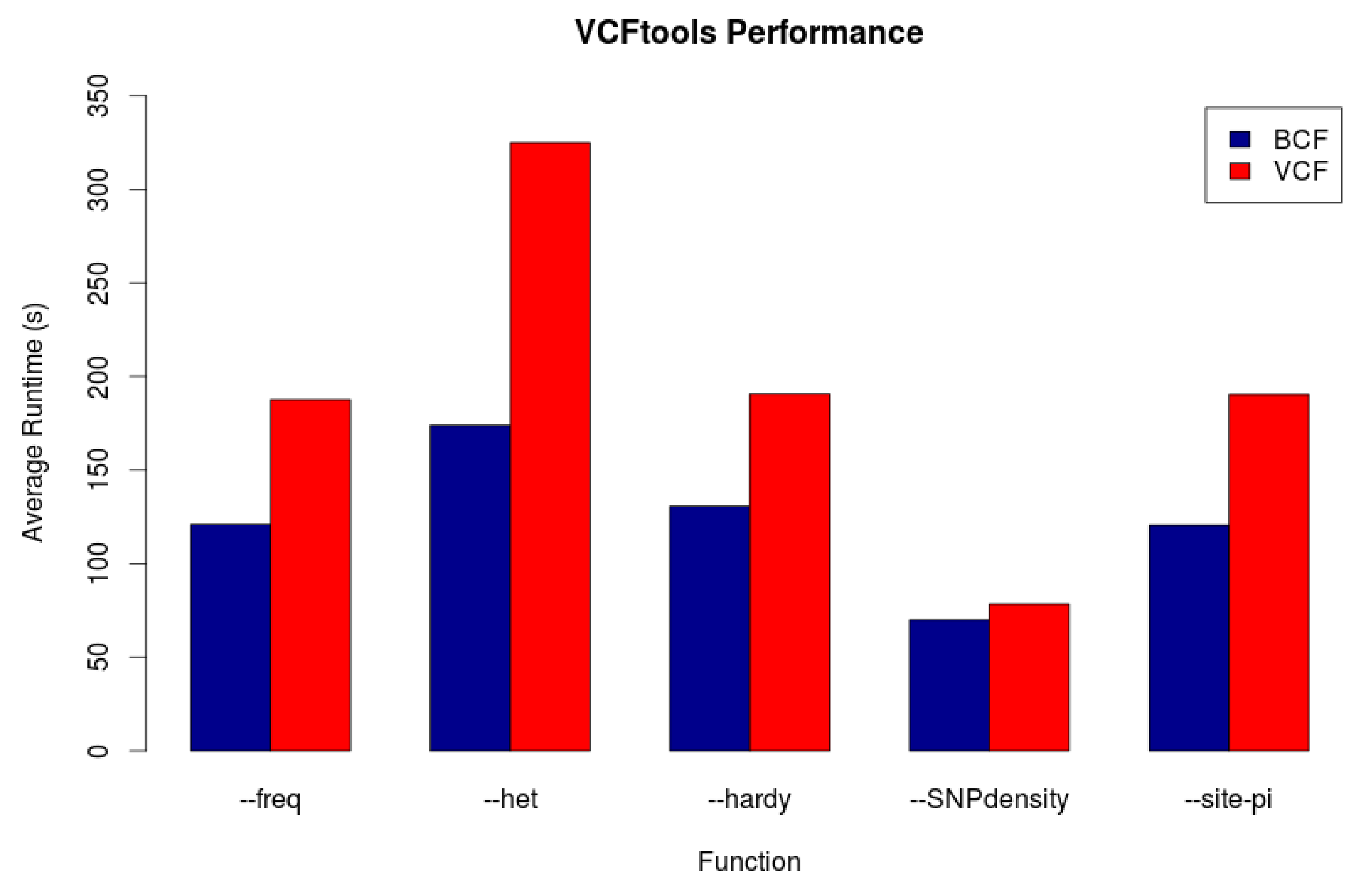


Data from Google Analytics

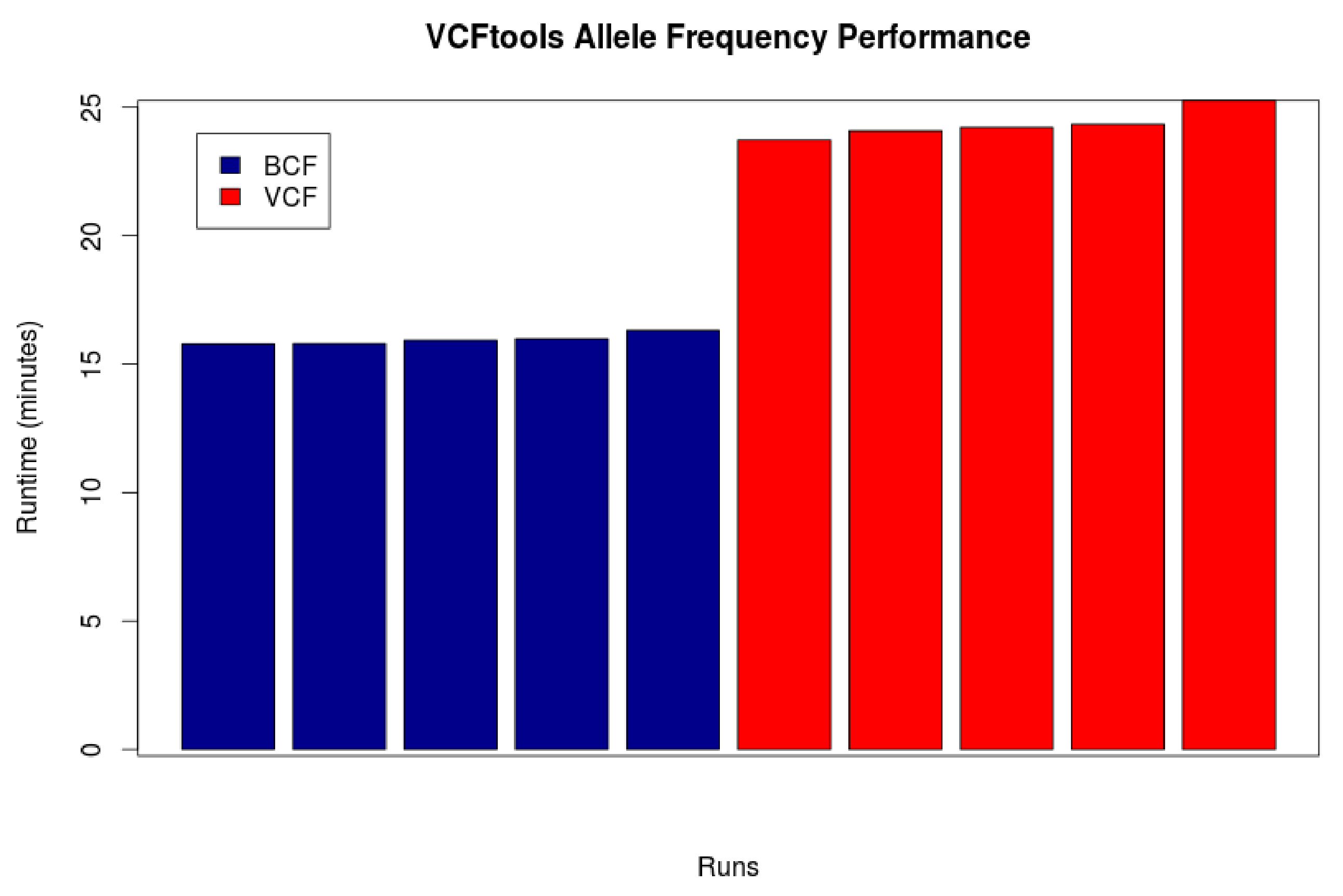
VCFtools Sourceforge Page Visits since March 1, 2010

Country	Visits
United States	87,539
United Kingdom	27,003
Germany	9,677
Canada	6,661
China	5,279
India	5,187
France	5,090
Japan	5,009
Netherlands	4,362
Spain	4,182

Better data accessibility helps increase VCFtools performance on BCF files in comparison to VCF files



Performance was measured using equivalent VCF and BCF files containing 5,000 individuals and ~150,000 sites over several runs



Performance was measured using equivalent VCF and BCF files containing 1,092 individuals and ~3,000,000 sites (all 1000 Genomes Phase 1 calls for chromosome 1)

Subversion Command

```
svn checkout svn://svn.code.sf.net/p/vcftools/code/
vcftools-code
```

Acknowledgements

Petr Danecek, Wellcome Trust Sanger Institute
Mark DePristo, Broad Institute
Heng Li, Broad Institute

Reference

Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin and 1000 Genomes Project Analysis Group. **The variant call format and VCFtools**. Bioinformatics, 2156-2156 (01 August 2011).