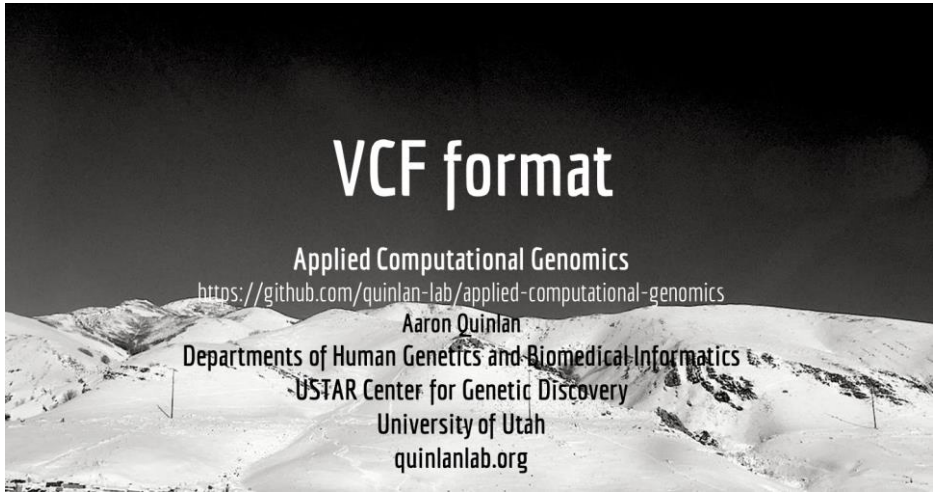
Cold Spring Harbor Laboratory

VCF Format

Arpad Danos, Felicia Gomez, Obi Griffith, Malachi Griffith,
My Hoang, Mariam Khanfar, Chris Miller, Kartik Singhal



Some slides are adapted from Dr. Aaron Quinlan

Variant Calling Overview

→ FASTQ → Align (BWA) → BAM → Detect SNP/INDELs (GATK or FreeBayes) → VCF

VCF format

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.

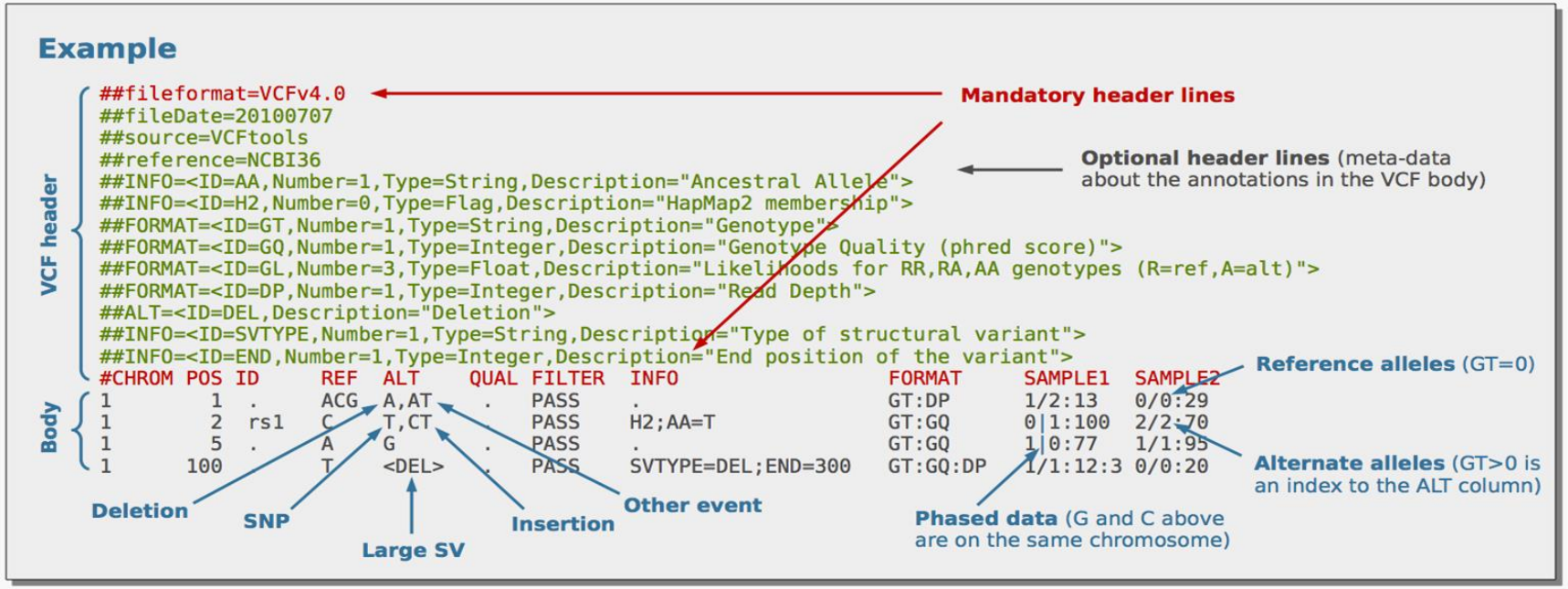
Availability: http://vcftools.sourceforge.net

Contact: rd@sanger.ac.uk

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.

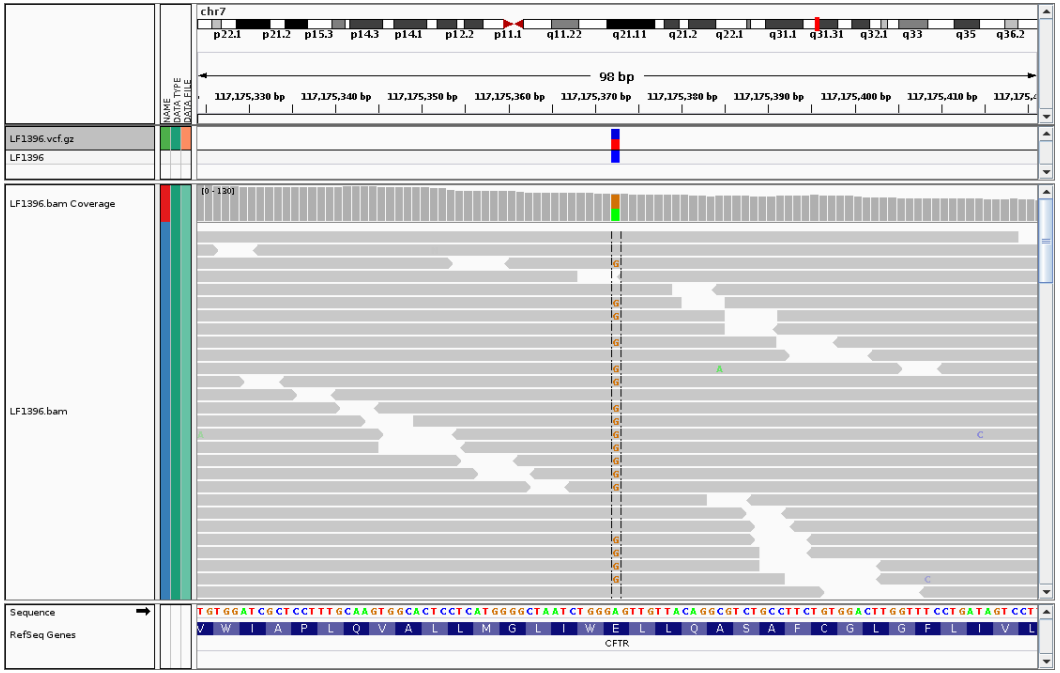
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/pdf/btr330.pdf

VCF format



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/pdf/btr330.pdf

VCF format. A basic example



Heterozygous A/G. The REF allele is allele "0", ALT is allele "1"

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr7	117175373	.	A	G	90	PASS	AF=0.5

FORMAT	LF1396
GT	0/1

Genotypes

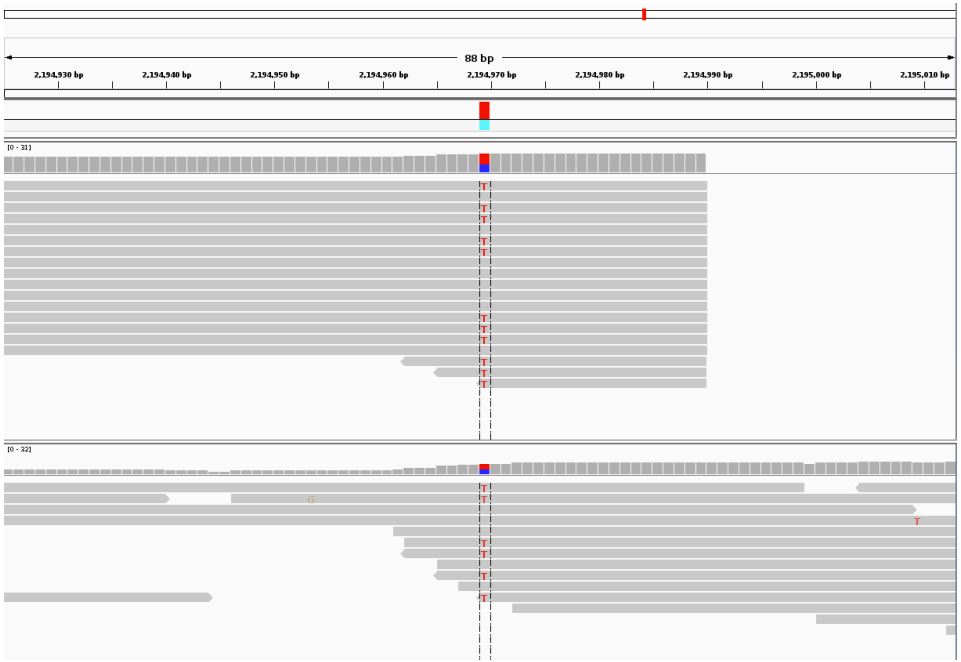
#CHROM	POS		ID	REF	ALT	QUAL		
	FILTER	INFO	FORMAT	LF1396				
chr7		117175373	.	A	G	90	PASS	Hom. Ref.
		AF=0.0 GT		0/0				
chr7		117175373	.	A	G	90	PASS	Het.
		AF=0.5 GT		0/1				
chr7		117175373	.	A	G	90	PASS	Hom. Alt.
		AF=1.0 GT		1/1				
chr7		117175373	.	A	G	0	PASS	Unknown
		AF=0.0 GT		./.				

Why would a genotype be unknown?

Multi-sample VCF

Mom

Kid



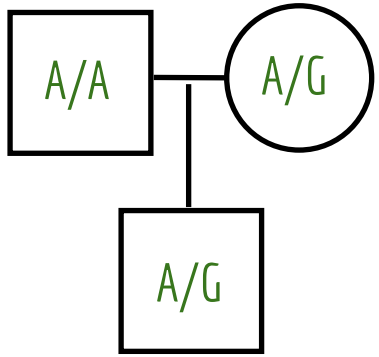
Heterozygous C/T.

#CHROM	POS		ID	REF	ALT	QUAL	FILTER	INFO
		FORMAT	MOM		KID			
chr7		2194169	.	C	T	210	PASS	
	AF=0.5	GT		0/1		0/1		

VCF format example

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=variantcallerXYZ
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOM	DAD	KID
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

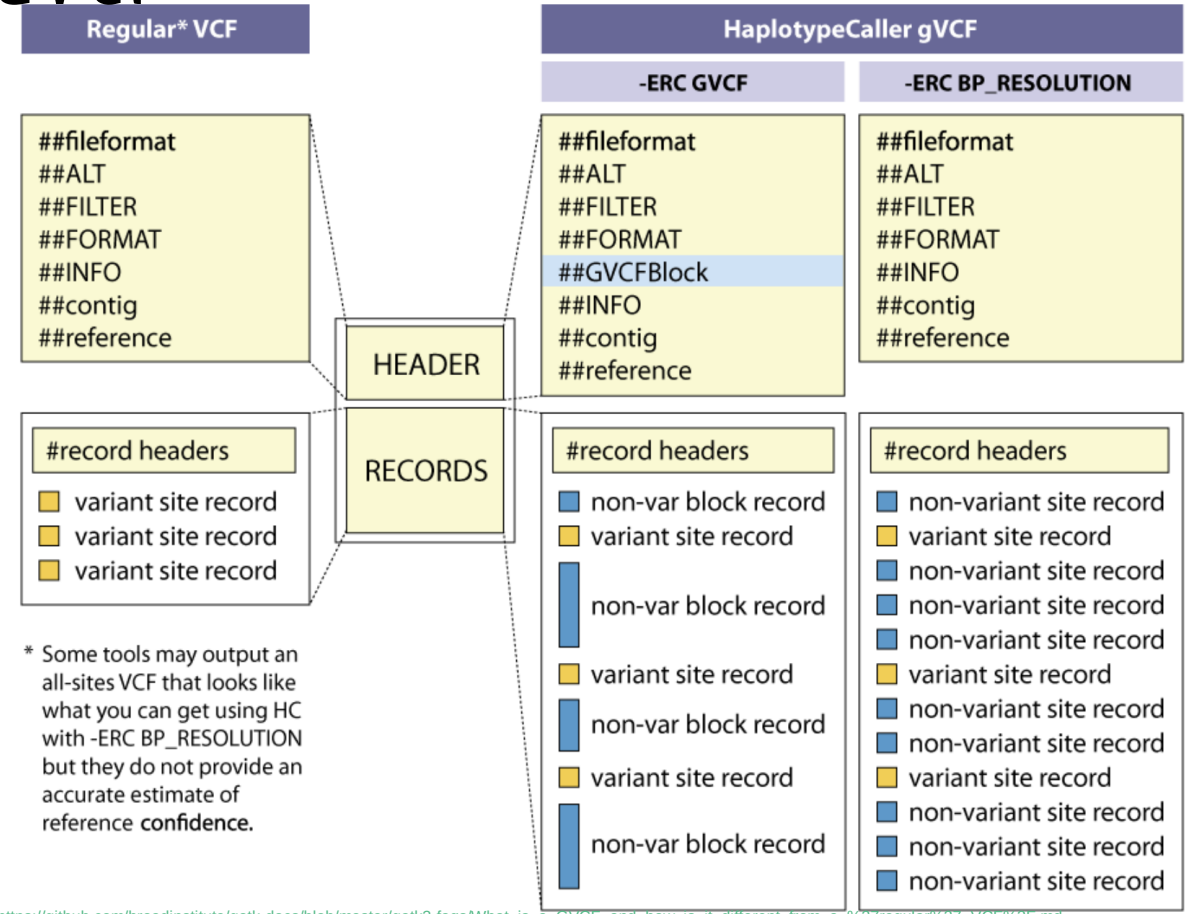


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/pdf/btr330.pdf

GVCF

- Genomic VCF but contains extra information
- The key difference between a regular VCF and a gVCF is that the gVCF has records for all sites, whether there is a variant call there or not.
- The goal is to have every site represented in the file in order to do joint analysis of a cohort in subsequent steps.
- The records in a gVCF include an accurate estimation of how confident we are in the determination that the sites are homozygous-reference or not.
- Two types; ERC: GVCF and BP_RESOLUTION

VCF vs GVCF



https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-faqs/What_is_a_GVCF_and_how_is_it_different_from_a_%27regular%27_VCF%3F.md

GVCF

```
##GVCFBlock=minGQ=0(inclusive),maxGQ=5(exclusive)
##GVCFBlock=minGQ=20(inclusive),maxGQ=60(exclusive)
##GVCFBlock=minGQ=5(inclusive),maxGQ=20(exclusive)
```

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  NA12878
20  10000000    .   T   <NON_REF>   .   .   END=10000116    GT:DP:GQ:MIN_DP:PL  0/0:44:99:38:0,89,1385
20  10000117    .   C   T,<NON_REF> 612.77  .   BaseQRankSum=0.000;ClippingRankSum=-0.411;DP=38;MLEAC=1,0;MLEAF=0.500,0.00;MQ=221.39;
20  10000118    .   T   <NON_REF>   .   .   END=10000210    GT:DP:GQ:MIN_DP:PL  0/0:42:99:38:0,80,1314
20  10000211    .   C   T,<NON_REF> 638.77  .   BaseQRankSum=0.894;ClippingRankSum=-1.927;DP=42;MLEAC=1,0;MLEAF=0.500,0.00;MQ=221.89;
20  10000212    .   A   <NON_REF>   .   .   END=10000438    GT:DP:GQ:MIN_DP:PL  0/0:52:99:42:0,99,1403
20  10000439    .   T   G,<NON_REF> 1737.77 .   DP=57;MLEAC=2,0;MLEAF=1.00,0.00;MQ=221.41;MQ0=0 GT:AD:DP:GQ:PL:SB   1/1:0,56,0:56:99:
20  10000440    .   T   <NON_REF>   .   .   END=10000597    GT:DP:GQ:MIN_DP:PL  0/0:56:99:49:0,120,1800
20  10000598    .   T   A,<NON_REF> 1754.77 .   DP=54;MLEAC=2,0;MLEAF=1.00,0.00;MQ=185.55;MQ0=0 GT:AD:DP:GQ:PL:SB   1/1:0,53,0:53:99:
20  10000599    .   T   <NON_REF>   .   .   END=10000693    GT:DP:GQ:MIN_DP:PL  0/0:51:99:47:0,120,1800
20  10000694    .   G   A,<NON_REF> 961.77  .   BaseQRankSum=0.736;ClippingRankSum=-0.009;DP=54;MLEAC=1,0;MLEAF=0.500,0.00;MQ=106.92;
20  10000695    .   G   <NON_REF>   .   .   END=10000757    GT:DP:GQ:MIN_DP:PL  0/0:48:99:45:0,120,1800
20  10000758    .   T   A,<NON_REF> 1663.77 .   DP=51;MLEAC=2,0;MLEAF=1.00,0.00;MQ=59.32;MQ0=0  GT:AD:DP:GQ:PL:SB   1/1:0,50,0:50:99:
20  10000759    .   A   <NON_REF>   .   .   END=10001018    GT:DP:GQ:MIN_DP:PL  0/0:40:99:28:0,65,1080
20  10001019    .   T   G,<NON_REF> 93.77   .   BaseQRankSum=0.058;ClippingRankSum=-0.347;DP=26;MLEAC=1,0;MLEAF=0.500,0.00;MQ=29.65;M
20  10001020    .   C   <NON_REF>   .   .   END=10001020    GT:DP:GQ:MIN_DP:PL  0/0:26:72:26:0,72,1080
20  10001021    .   T   <NON_REF>   .   .   END=10001021    GT:DP:GQ:MIN_DP:PL  0/0:25:37:25:0,37,909
20  10001022    .   C   <NON_REF>   .   .   END=10001297    GT:DP:GQ:MIN_DP:PL  0/0:30:87:25:0,72,831
```