1000 Genomes

A Deep Catalog of Human Genetic Variation

Home About Data Analysis Participants Contact Browser Wiki

Search

Home ›

Encoding Structural Variants in VCF (Variant Call Format) version 4.0

1. Introduction

This web page describes the conventions and extensions adopted by the 1000 Genomes Project for encoding structural variations in VCF 4.0 format.

The encoding of structural variants in VCF is guided by two principles:

a) When breakpoints / alleles of structural variants are precisely known, then the format should be completely compatible with the format used for smaller indels. b) When the position, length and/or base composition of the variant is not known, we want to encode as much useful information as possible about the variant.

For precisely known variants, the REF and ALT fields should contain the full sequences for the alleles, following the usual VCF conventions. For imprecise variants, the REF field may contain a single base and the ALT fields should contain descriptive alleles (e.g. <ID>), described in more detail below. Imprecise variants should also be marked by the presence of an IMPRECISE flag in the INFO field.

In both cases, the POS field should specify the 1-based coordinate of the base before the variant or the best estimate thereof. When the position is ambiguous due to identical reference sequence, the POS coordinate is based on the leftmost possible position of the variant.

2. Example

Examples of structural variants encoded in VCF:

USER LOGIN

Username: *

Password: *

Log in

Request new password

```
##fileformat=VCFv4.0
##fileDate=20100501
##reference=1000GenomesPilot-NCBI36
##assembly=ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/sv/breakpoint_assemblies.fasta
##INFO=<ID=BKPTID,Number=-1,Type=String,Description="ID of the assembled alternate allele in the assembly file"
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=-1,Type=Integer,Description="Length of base pair identical micro-homology at event bre
##INFO=<ID=HOMSEQ,Number=-1,Type=String,Description="Sequence of base pair identical micro-homology at event br
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=MEINFO,Number=4,Type=String,Description="Mobile element info of the form NAME,START,END,POLARITY">
##INFO=<ID=SVLEN,Number=-1,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DEL:ME:ALU,Description="Deletion of ALU element">
##ALT=<ID=DEL:ME:L1,Description="Deletion of L1 element">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=DUP:TANDEM,Description="Tandem Duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=INS:ME:ALU,Description="Insertion of ALU element">
##ALT=<ID=INS:ME:L1,Description="Insertion of L1 element">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">
#CHROM  POS   ID  REF ALT   QUAL  FILTER  INFO  FORMAT  NA00001
1 2827693   . CCGTGGATGCGGGGACCCGCATCCCCTCTCCCTTCACAGCTGAGTGACCCACATCCCCTCTCCCCTCGCA  C . PASS  SVTYPE=DEL;END=
2 321682    . T <DEL>   6 PASS    IMPRECISE;SVTYPE=DEL;END=321887;SVLEN=-105;CIPOS=-56,20;CIEND=-10,62  GT:GQ 0
2 14477084  . C <DEL:ME:ALU>  12  PASS  IMPRECISE;SVTYPE=DEL;END=14477381;SVLEN=-297;MEINFO=AluYa5,5,307,+;CIPO
3 9425916   . C <INS:ME:L1> 23  PASS  IMPRECISE;SVTYPE=INS;END=9425916;SVLEN=6027;CIPOS=-16,22;MIINFO=L1HS,1,60
3 12665100  . A <DUP>   14  PASS  IMPRECISE;SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500,500;CIEND=-500,500
4 18665128  . T <DUP:TANDEM>  11  PASS  IMPRECISE;SVTYPE=DUP;END=18665204;SVLEN=76;CIPOS=-10,10;CIEND=-10,10  G
```

The example shows in order:

- A precise deletion with known breakpoint, a one base micro-homology, and a sample that is homozygous for the deletion.
- An imprecise deletion of approximately 105 bp.
- An imprecise deletion of an ALU element relative to the reference.
- An imprecise insertion of an L1 element relative to the reference.
- An imprecise duplication of approximately 21Kb. The sample genotype is copy number 3 (one extra copy of the duplicated sequence).
- An imprecise tandem duplication of 76bp. The sample genotype is copy number 5 (but the two haplotypes are not known).

3. Meta-information lines

The following meta-information lines are used with structural variations:

Breakpoint assemblies:

##assembly=*url*

The URL field specifies the location of a fasta file containing breakpoint assemblies referenced in the VCF records via the BKPTID key.

Descriptive alternate alleles for imprecise variants:

##ALT=<ID=*type*,Description=*description*>

The ID field indicates the type of structural variant, and can be a colon-separated list of types and subtypes. The first level type must be one of the following:

- DEL Deletion relative to the reference
- INS Insertion of novel sequence relative to the reference
- DUP Region of elevated copy number relative to the reference
- INV Inversion of reference sequence
- CNV Copy number variable region (may be both deletion and duplication)

The CNV category should not be used when a more specific category can be applied. Reserved subtypes include:

- DUP:TANDEM Tandem duplication
- DEL:ME Deletion of mobile element relative to the reference
- INS:ME Insertion of a mobile element relative to the reference

4. INFO keys used for structural variants

The following INFO keys are reserved for encoding structural variants. In general, when these keys are used by imprecise variants, the values should be best estimates. When a key reflects a property of a single alt allele (e.g. SVLEN), then when there are multiple alt alleles there will be multiple values for the key corresponding to each alelle (e.g. SVLEN=-100,-110 for a deletion with two distinct alt alleles).

##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">

##INFO=<ID=NOVEL,Number=0,Type=Flag,Description="Indicates a novel structural variation">

##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">

For precise variants, END is POS + length of REF allele - 1, and the for imprecise variants the corresponding best estimate.

##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">

Value should be one of DEL, INS, DUP, INV, CNV. This key can be derived from the REF/ALT fields but is useful for filtering.

##INFO=<ID=SVLEN,Number=-1,Type=Integer,Description="Difference in length between REF and ALT alleles">

One value for each ALT allele. Longer ALT alleles have positive values, shorter ALT alleles negative values.

##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">

##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">

##INFO=<ID=HOMLEN,Number=-1,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">

##INFO=<ID=HOMSEQ,Number=-1,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">

##INFO=<ID=BKPTID,Number=-1,Type=String,Description="ID of the assembled alternate allele in the assembly file">

For precise variants, the consensus sequence the alternate allele assembly is derivable from the REF and ALT fields. However, the alternate allele assembly file may contain additional information about the characteristics of the alt allele contigs.

##INFO=<ID=MEINFO,Number=4,Type=String,Description="Mobile element info of the form NAME,START,END,POLARITY">

##INFO=<ID=METRANS,Number=4,Type=String,Description="Mobile element transduction info of the form CHR,START,END,POLARITY">

##INFO=<ID=DGVID,Number=1,Type=String,Description="ID of this element in Database of Genomic Variation">

##INFO=<ID=DBVARID,Number=1,Type=String,Description="ID of this element in DBVAR">

##INFO=<ID=DBRIPID,Number=1,Type=String,Description="ID of this element in DBRIP">

5. FORMAT keys used for structural variants

##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">

##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise events">

These keys are analogous to GT/GQ and are provided for genotyping imprecise events by copy number (either because there is an unknown number of alternate alleles or because the haplotypes cannot be determined). CN specifies the integer copy number of the variant in this sample. CNQ is encoded as a phred quality -10log_10p(copy number genotype call is wrong). When possible, GT/GQ should be used instead of (or in addition to) these keys.

up

Last updated by flicek@ebi.ac.uk on Thursday 25th November, 2010 15:37.
Originally submitted by flicek@ebi.ac.uk on Wednesday 3rd November, 2010 11:20.