1000 Genomes

A Deep Catalog of Human Genetic Variation

Home About Data Analysis Participants Contact Browser Wiki FTP search

Search

Home ›

VCF (Variant Call Format) version 4.1

0. Example

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome.

There is an option whether to contain genotype information on samples for each position or not.

Example:

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens
##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      NA000
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:4
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:4
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|2:2
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:5
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:3
```

This example shows in order a good simple SNP, a possible SNP that has been filtered out because its quality is below 10, a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error), a site that is called monomorphic reference (i.e. with no alternate alleles), and a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T). Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.

1. Meta-information lines

File meta-information is included after the ## string and must be key=value pairs.

A single 'fileformat' field is always required, must be the first line in the file, and details the VCF format version number. For example, for VCF version 4.1, this line should read:

##fileformat=VCFv4.1

It is strongly encouraged that information lines describing the INFO, FILTER and FORMAT entries used in the body of the VCF file be included in the meta-information section. Although they are optional, if these lines are present then they must be completely well-formed.

INFO fields should be described as follows (all keys are required):

##INFO=<ID=*ID*,Number=*number*,Type=*type*,Description="*description*">

Possible Types for INFO fields are: Integer, Float, Flag, Character, and String.

USER LOGIN

Username: *

Password: *

Log in

Request new password

The Number entry is an Integer that describes the number of values that can be included with the INFO field. For example, if the INFO field contains a single number, then this value should be 1; if the INFO field describes a pair of numbers, then this value should be 2 and so on. If the field has one value per alternate allele then this value should be 'A'; if the field has one value for each possible genotype (more relevant to the FORMAT tags) then this value should be 'G'. If the number of possible values varies, is unknown, or is unbounded, then this value should be '.'. The 'Flag' type indicates that the INFO field does not contain a Value entry, and hence the Number should be 0 in this case. The Description value must be surrounded by double-quotes. Double-quote character can be escaped with backslash (\") and backslash as \\.

FILTERs that have been applied to the data should be described as follows:

##FILTER=<ID=*ID*,Description="*description*">

Likewise, Genotype fields specified in the FORMAT field should be described as follows:

##FORMAT=<ID=*ID*,Number=*number*,Type=*type*,Description="*description*">

Possible Types for FORMAT fields are: Integer, Float, Character, and String (this field is otherwise defined precisely as the INFO field).

Symbolic alternate alleles for imprecise structural 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. ID values are case sensitive strings and may not contain whitespace or angle brackets. 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

In addition, it is highly recommended (but not required) that the header include tags describing the reference and contigs backing the data contained in the file. These tags are based on the SQ field from the SAM spec; all tags are optional (see the VCF example above).

Breakpoint assemblies for structural variations may use an external file:

##assembly=*url*

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

As with chromosomal sequences it is highly recommended (but not required) that the header include tags describing the contigs referred to in the VCF file. This furthermore allows these contigs to come from different files. The format is identical to that of a reference sequence, but with an additional URL tag to indicate where that sequence can be found. For example:.

##contig=<ID=ctg1,URL=ftp://somewhere.org/assembly.fa,...>

As explained below it is possible to define sample to genome mappings:

##SAMPLE=<ID=S_ID,Genomes=G1_ID;G2_ID; ...;GK_ID,Mixture=N1;N2; ...;NK,Description=S1;S2; ...; SK >

As well as derivation relationships between genomes using the following syntax:

##PEDIGREE=<Name_0=G0-ID,Name_1=G1-ID,...,Name_N=GN-ID>

or a link to a database:

##pedigreeDB=<url>

2. The header line syntax

The header line names the 8 fixed, mandatory columns. These columns are as follows:

1. #CHROM

2. POS
3. ID
4. REF
5. ALT
6. QUAL
7. FILTER
8. INFO

If genotype data is present in the file, these are followed by a FORMAT column header, then an arbitrary number of sample IDs. The header line is tab-delimited.

3. Data lines

Fixed fields

There are 8 fixed fields per record. All data lines are tab-delimited. In all cases, missing values are specified with a dot ("."). Fixed fields are:

1. CHROM chromosome: an identifier from the reference genome or an angle-bracketed ID String ("<ID>") pointing to a contig in the assembly file (cf. the ##assembly line in the header). All entries for a specific CHROM should form a contiguous block within the VCF file. The colon symbol (:) must be absent from all chromosome names to avoid parsing errors when dealing with breakends. (String, no white-space permitted, Required).
2. POS position: The reference position, with the 1st base having position 1. Positions are sorted numerically, in increasing order, within each reference sequence CHROM. It is permitted to have multiple records with the same POS. Telomeres are indicated by using positions 0 or N+1, where N is the length of the corresponding chromosome or contig. (Integer, Required)
3. ID semi-colon separated list of unique identifiers where available. If this is a dbSNP variant it is encouraged to use the rs number(s). No identifier should be present in more than one data record. If there is no identifier available, then the missing value should be used. (String, no white-space or semi-colons permitted)
4. REF reference base(s): Each base must be one of A,C,G,T,N (case insensitive). Multiple bases are permitted. The value in the POS field refers to the position of the first base in the String. For simple insertions and deletions in which either the REF or one of the ALT alleles would otherwise be null/empty, the REF and ALT Strings must include the base before the event (which must be reflected in the POS field), unless the event occurs at position 1 on the contig in which case it must include the base after the event; this padding base is not required (although it is permitted) for e.g. complex substitutions or other events where all alleles have at least one base represented in their Strings. If any of the ALT alleles is a symbolic allele (an angle-bracketed ID String "*<ID>*") then the padding base is required and POS denotes the coordinate of the base preceding the polymorphism. Tools processing VCF files are not required to preserve case in the allele Strings. (String, Required).
5. ALT comma separated list of alternate non-reference alleles called on at least one of the samples. Options are base Strings made up of the bases A,C,G,T,N, (case insensitive) or an angle-bracketed ID String ("*<ID>*") or a breakend replacement string as described in the section on breakends. If there are no alternative alleles, then the missing value should be used. Tools processing VCF files are not required to preserve case in the allele String, except for IDs, which are case sensitive. (String; no whitespace, commas, or angle-brackets are permitted in the ID String itself)
6. QUAL phred-scaled quality score for the assertion made in ALT. i.e. -10log_10 prob(call in ALT is wrong). If ALT is "." (no variant) then this is -10log_10 p(variant), and if ALT is not "." this is -10log_10 p(no variant). High QUAL scores indicate high confidence calls. Although traditionally people use integer phred scores, this field is permitted to be a floating point to enable higher resolution for low confidence calls if desired. If unknown, the missing value should be specified. (Numeric)
7. FILTER : PASS if this position has passed all filters, i.e. a call is made at this position. Otherwise, if the site has not passed all filters, a semicolon-separated list of codes for filters that fail. e.g. "q10;s50" might indicate that at this site the quality is below 10 and the number of samples with data is below 50% of the total number of samples. "0" is reserved and should not be used as a filter String. If filters have not been applied, then this field should be set to the missing value. (String, no white-space or semi-colons permitted)
8. INFO additional information: (String, no white-space, semi-colons, or equals-signs permitted; commas are permitted only as delimiters for lists of values) INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: <key>=<data>[,data]. Arbitrary keys are permitted, although the following sub-fields are reserved (albeit optional):
   - AA : ancestral allele
   - AC : allele count in genotypes, for each ALT allele, in the same order as listed
   - AF : allele frequency for each ALT allele in the same order as listed: use this when estimated from primary data, not called genotypes
   - AN : total number of alleles in called genotypes
   - BQ : RMS base quality at this position
   - CIGAR : cigar string describing how to align an alternate allele to the reference allele
   - DB : dbSNP membership
   - DP : combined depth across samples, e.g. DP=154
   - END : end position of the variant described in this record (for use with symbolic alleles)
   - H2 : membership in hapmap2
   - H3 : membership in hapmap3
   - MQ : RMS mapping quality, e.g. MQ=52
   - MQ0 : Number of MAPQ == 0 reads covering this record
   - NS : Number of samples with data
   - SB : strand bias at this position
   - SOMATIC : indicates that the record is a somatic mutation, for cancer genomics
   - VALIDATED : validated by follow-up experiment

- 1000G : membership in 1000 Genomes

The exact format of each INFO sub-field should be specified in the meta-information (as described above).

Example for an INFO field: DP=154;MQ=52;H2. Keys without corresponding values are allowed in order to indicate group membership (e.g. H2 indicates the SNP is found in HapMap 2). It is not necessary to list all the properties that a site does NOT have, by e.g. H2=0.

See below for additional reserved INFO sub-fields used to encode structural variants.

Genotype fields

If genotype information is present, then the same types of data must be present for all samples. First a FORMAT field is given specifying the data types and order (colon-separated alphanumeric String). This is followed by one field per sample, with the colon-separated data in this field corresponding to the types specified in the format. The first sub-field must always be the genotype (GT) if it is present. There are no required sub-fields.

As with the INFO field, there are several common, reserved keywords that are standards across the community:

- GT : genotype, encoded as allele values separated by either of ”/” or “|”. The allele values are 0 for the reference allele (what is in the REF field), 1 for the first allele listed in ALT, 2 for the second allele list in ALT and so on. For diploid calls examples could be 0/1, 1|0, or 1/2, etc. For haploid calls, e.g. on Y, male non-pseudoautosomal X, or mitochondrion, only one allele value should be given; a triploid call might look like 0/0/1. If a call cannot be made for a sample at a given locus, ”.”should be specified for each missing allele in the GT field (for example "./." for a diploid genotype and "." for haploid genotype). The meanings of the separators are as follows (see the PS field below for more details on incorporating phasing information into the genotypes):
  - / : genotype unphased
  - | : genotype phased
- DP : read depth at this position for this sample (Integer)
- FT : sample genotype filter indicating if this genotype was “called” (similar in concept to the FILTER field). Again, use PASS to indicate that all filters have been passed, a semi-colon separated list of codes for filters that fail, or ”.” to indicate that filters have not been applied. These values should be described in the meta-information in the same way as FILTERs (String, no white-space or semi-colons permitted)
- GL : genotype likelihoods comprised of comma separated floating point log10-scaled likelihoods for all possible genotypes given the set of alleles defined in the REF and ALT fields. In presence of the GT field the same ploidy is expected and the canonical order is used; without GT field, diploidy is assumed. If A is the allele in REF and B,C,... are the alleles as ordered in ALT, the ordering of genotypes for the likelihoods is given by: F(j/k) = (k*(k+1)/2)+j. In other words, for biallelic sites the ordering is: AA,AB,BB; for triallelic sites the ordering is: AA,AB,BB,AC,BC,CC, etc. For example: GT:GL 0/1:-323.03,-99.29,-802.53 (Floats)
- GLE : genotype likelihoods of heterogeneous ploidy, used in presence of uncertain copy number. For example: GLE=0:-75.22,1:-223.42,0/0:-323.03,1/0:-99.29,1/1:-802.53 (String)
- PL : the phred-scaled genotype likelihoods rounded to the closest integer (and otherwise defined precisely as the GL field) (Integers)
- GP : the phred-scaled genotype posterior probabilities (and otherwise defined precisely as the GL field); intended to store imputed genotype probabilities (Floats)
- GQ : conditional genotype quality, encoded as a phred quality -10log_10p(genotype call is wrong, conditioned on the site's being variant) (Integer)
- HQ : haplotype qualities, two comma separated phred qualities (Integers)
- PS : phase set. A phase set is defined as a set of phased genotypes to which this genotype belongs. Phased genotypes for an individual that are on the same chromosome and have the same PS value are in the same phased set. A phase set specifies multi-marker haplotypes for the phased genotypes in the set. All phased genotypes that do not contain a PS subfield are assumed to belong to the same phased set. If the genotype in the GT field is unphased, the corresponding PS field is ignored. The recommended convention is to use the position of the first variant in the set as the PS identifier (although this is not required). (Non-negative 32-bit Integer)
- PQ : phasing quality, the phred-scaled probability that alleles are ordered incorrectly in a heterozygote (against all other members in the phase set). We note that we have not yet included the specific measure for precisely defining "phasing quality"; our intention for now is simply to reserve the PQ tag for future use as a measure of phasing quality. (Integer)
- EC : comma separated list of expected alternate allele counts for each alternate allele in the same order as listed in the ALT field (typically used in association analyses) (Integers)
- MQ : RMS mapping quality, similar to the version in the INFO field. (Integer)

If any of the fields is missing, it is replaced with the missing value. For example if the FORMAT is GT:GQ:DP:HQ then 0|0:.:23:23,34 indicates that GQ is missing. Trailing fields can be dropped (with the exception of the GT field, which should always be present if specified in the FORMAT field).

See below for additional genotype fields used to encode structural variants.

Additional Genotype fields can be defined in the meta-information. However, software support for such fields is not guaranteed.

4. Understanding the VCF format and the haplotype representation

VCF records use a single general system for representing genetic variation data composed of:

- Allele: representing single genetic haplotypes (A, T, ATC).
- Genotype: an assignment of alleles for each chromosome of a single named sample at a particular locus.
- VCF record: a record holding all segregating alleles at a locus (as well as genotypes, if appropriate, for multiple individuals containing alleles at that locus).

VCF records use a simple haplotype representation for REF and ALT alleles to describe variant haplotypes at a locus. ALT haplotypes are constructed from the REF haplotype by taking the REF allele bases at the POS in the reference genotype and replacing them with the ALT bases. In essence, the VCF record specifies a-REF-t and the alternative haplotypes are a-ALT-t for each alternative allele.

5. INFO keys used for structural variants

When the INFO keys reserved for encoding structural variants are used for 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).

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, BND. This key can be derived from the REF/ALT fields but is useful for filtering.

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

One value for each ALT allele. Longer ALT alleles (e.g. insertions) have positive values, shorter ALT alleles (e.g. deletions) have 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=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">

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

##INFO=<ID=BKPTID,Number=.,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">

##INFO=<ID=MATEID,Number=.,Type=String,Description="ID of mate breakends">

##INFO=<ID=PARID,Number=1,Type=String,Description="ID of partner breakend">

##INFO=<ID=EVENT,Number=1,Type=String,Description="ID of event associated to breakend">

##INFO=<ID=CILEN,Number=2,Type=Integer,Description="Confidence interval around the length of the inserted

material between breakends">

##INFO=<ID=DP,Number=1,Type=Integer,Description="Read Depth of segment containing breakend">

##INFO=<ID=DPADJ,Number=.,Type=Integer,Description="Read Depth of adjacency">

##INFO=<ID=CN,Number=1,Type=Integer,Description="Copy number of segment containing breakend">

##INFO=<ID=CNADJ,Number=.,Type=Integer,Description="Copy number of adjacency">

##INFO=<ID=CICN,Number=2,Type=Integer,Description="Confidence interval around copy number for the segment">

##INFO=<ID=CICNADJ,Number=.,Type=Integer,Description="Confidence interval around copy number for the adjacency">

6. 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">

##FORMAT=<ID=CNL,Number=.,Type=Float,Description="Copy number genotype likelihood for imprecise events">

##FORMAT=<ID=NQ,Number=1,Type=Integer,Description="Phred style probability score that the variant is novel with respect to the genome's ancestor">

##FORMAT=<ID=HAP,Number=1,Type=Integer,Description="Unique haplotype identifier">

##FORMAT=<ID=AHAP,Number=1,Type=Integer,Description="Unique identifier of ancestral haplotype">

These keys are analogous to GT/GQ/GL 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). CNL specifies a list of log10 likelihoods for each potential copy number, starting from zero. When possible, GT/GQ/GL should be used instead of (or in addition to) these keys.

7. How do I represent example variation in VCF records?

SNPs and Small Indels

For example, suppose we are looking at a locus in the genome:

```
Ref: a t C g a // C is the reference base
   : a t G g a // C base is a G in some individuals
   : a t - g a // C base is deleted w.r.t. the reference
   : a t CAg a // A base is inserted w.r.t. the reference sequence
```

In the above cases, what are the alleles and how would they be represented as a VCF record?

* First is a SNP polymorphism of C/G → { C , G } → C is the reference allele

```
20     3 .          C       G        .    PASS   DP=100
```

* Second, 1 base deletion of C → { tC , t } → tC is the reference allele

```
20     2 .          TC       T       .    PASS   DP=100
```

* Third, 1 base insertion of A → { tC ; tCA } → tC is the reference allele

```
20     2 .          TC       TCA     .    PASS   DP=100
```

Suppose I see a the following in a population of individuals and want to represent these three segregating alleles:

```
Ref: a t C g a // C is the reference base
   : a t G g a // C base is a G in some individuals
   : a t - g a // C base is deleted w.r.t. the
```

How do I represent this? There are three segregating alleles: { tC , tG , t } with a corresponding VCF record:

```
20     2 .         TC      TG,T    .   PASS  DP=100
```

Now suppose I have this more complex example:

```
Ref: a t C g a // C is the reference base
   : a t - g a
   : a t - - a
   : a t CAg a
```

There are actually four segregating alleles: { tCg , tg, t, and tCAg } over bases 2-4. This complex set of allele is represented in VCF as:

```
20     2 .         TCG      TG,T,TCAG    .   PASS  DP=100
```

Note that in VCF records, the molecular equivalence explicitly listed above in the per-base alignment is discarded, so the actual placement of equivalent g isn't retained.

For completeness, VCF records are dynamically typed, so whether a VCF record is a SNP, Indel, Mixed, or Reference site depends on the properties of the alleles in the record.

What do example VCF records indicate as variation from the reference?

SNP VCF record

Suppose I receive the following VCF record:

```
20     3 .         C      T    .   PASS  DP=100
```

This is a SNP since its only single base substitution and there are only two alleles so I have the two following segregating haplotypes:

```
Ref: a t C g a // C is the reference base
   : a t T g a // C base is a T in some individuals
```

Insertion VCF record

Suppose I receive the following VCF record:

```
20     3 .         C      CTAG    .   PASS  DP=100
```

This is a insertion since the reference base C is being replaced by C [the reference base] plus three insertion bases TAG. Again there are only two alleles so I have the two following segregating haplotypes:

```
Ref: a t C - - - g a // C is the reference base
   : a t C T A G g a // following the C base is an insertion of 3 bases
```

Deletion VCF record

Suppose I receive the following VCF record:

```
20     2 .         TCG      T    .   PASS  DP=100
```

This is a deletion of two reference bases since the reference allele TCG is being replaced by just the T [the reference base]. Again there are only two alleles so I have the two following segregating haplotypes:

```
Ref: a t C g a // C is the reference base
   : a t - - a // following the C base is a deletion of 2 bases
```

Mixed VCF record for a microsatellite

Suppose I receive the following VCF record:

```
20     4 .         GCG      G,GCGCG    .   PASS  DP=100
```

This is a mixed type record containing a 2 base insertion and a 2 base deletion. There are are three segregating alleles so I have the three following haplotypes:

```
Ref: a t c g c g - - a  // C is the reference base
   : a t c g - - - - a  // following the C base is a deletion of 2 bases
   : a t c g c g c g a  // following the C base is a insertion of 2 bases


Note that in all of these examples dashes have been added to make the haplotypes clearer but of course

Ref: a t c g - - c g a  // C is the reference base
   : a t c g - - - - a  // following the C base is a deletion of 2 bases
   : a t c g c g c g a  // following the C base is a insertion of 2 bases
```

Encoding Structural Variants in VCF

This section describes additional rules for encoding structural variation in VCF 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 symbolic 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.

Structural Variant Example

Examples of structural variants encoded in VCF:

```
##fileformat=VCFv4.1
##fileDate=20100501
##reference=1000GenomesPilot-NCBI36
##assembly=ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/sv/breakpoint_assemblies.fasta
##INFO=<ID=BKPTID,Number=.,Type=String,Description="ID of the assembled alternate allele in the assembl
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise varian
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise varian
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record"
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at ev
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at e
##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,POLA
##INFO=<ID=SVLEN,Number=.,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=String,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
1 2827694   rs2376870  CGTGGATGCGGGGACCCGCATCCCCTCTCCCTTCACAGCTGAGTGACCCACATCCCCTCTCCCCTCGCA  C . PASS
2 321682    .  T   <DEL>         6     PASS    IMPRECISE;SVTYPE=DEL;END=321887;SVLEN=-205;CIPOS=-56,20;
2 14477084  .  C   <DEL:ME:ALU>  12    PASS    IMPRECISE;SVTYPE=DEL;END=14477381;SVLEN=-297;MEINFO=AluY
3 9425916   .  C   <INS:ME:L1>   23    PASS    IMPRECISE;SVTYPE=INS;END=9425916;SVLEN=6027;CIPOS=-16,22
3 12665100  .  A   <DUP>         14    PASS    IMPRECISE;SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500
4 18665128  .  T   <DUP:TANDEM>  11    PASS    IMPRECISE;SVTYPE=DUP;END=18665204;SVLEN=76;CIPOS=-10,10;
```

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).

Specifying Complex Rearrangements with Breakends

An arbitrary rearrangement event can be summarized as a set of novel **adjacencies**.

Each adjacency ties together 2 **breakends**. The two breakends at either end of a novel adjacency are called **mates**.

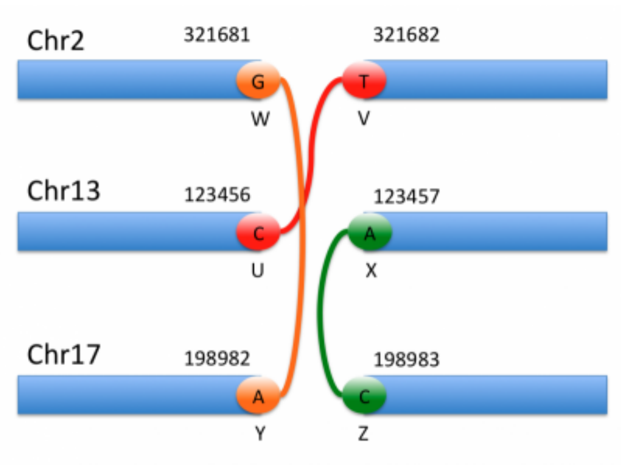
There is one line of VCF (i.e. one record) for each of the two breakends in a novel adjacency. A breakend record is identified with the tag SYTYPE=BND" in the INFO field. The REF field of a breakend record indicates a base or sequence s of bases beginning at position POS, as in all VCF records. The ALT field of a breakend record indicates a replacement for s. This "breakend replacement" has three parts:

1. the string t that replaces places s. The string t may be an extended version of s if some novel bases are inserted during the formation of the novel adjacency.
2. The position p of the mate breakend, indicated by a string of the form "chr:pos". This is the location of the first mapped base in the piece being joined at this novel adjacency.
3. The direction that the joined sequence continues in, starting from p. This is indicated by the orientation of square brackets surrounding p.

These 3 elements are combined in 4 possible ways to create the ALT. In each of the 4 cases, the assertion is that s is replaced with t, and then some piece starting at position p is joined to t. The cases are:

```
REF   ALT    Meaning
s     t[p[   piece extending to the right of p is joined after t
s     t]p]   reverse comp piece extending left of p is joined after t
s     ]p]t   piece extending to the left of p is joined before t
s     [p[t   reverse comp piece extending right of p is joined before t
```

The following example shows a 3-break operation involving 6 breakends. It exemplifies all possible orientations of breakends in adjacencies. Notice how the ALT field expresses the orientation of the breakends.



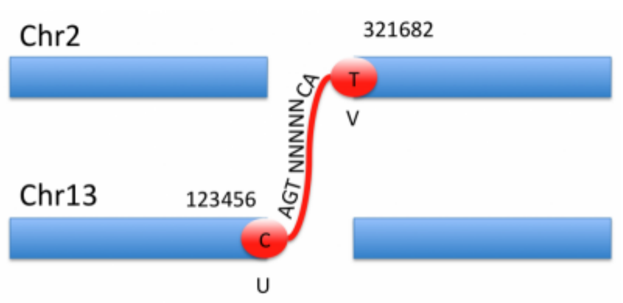
```
#CHROM POS    ID    REF ALT           QUAL FILT INFO
2      321681 bnd_W  G   G]17:198982]  6    PASS SVTYPE=BND
2      321682 bnd_V  T   ]13:123456]T  6    PASS SVTYPE=BND
13     123456 bnd_U  C   C[2:321682[   6    PASS SVTYPE=BND
13     123457 bnd_X  A   [17:198983[A  6    PASS SVTYPE=BND
17     198982 bnd_Y  A   A]2:321681]   6    PASS SVTYPE=BND
17     198983 bnd_Z  C   [13:123457[C  6    PASS SVTYPE=BND
```

On the same data, it is possible to explicitly state the pairing between mate breakends:

```
#CHROM POS    ID    REF ALT           QUAL FILT INFO
2      321681 bnd_W  G   G]17:198982]  6    PASS SVTYPE=BND;MATEID=bnd_Y
2      321682 bnd_V  T   ]13:123456]T  6    PASS SVTYPE=BND;MATEID=bnd_U
13     123456 bnd_U  C   C[2:321682[   6    PASS SVTYPE=BND;MATEID=bnd_V
13     123457 bnd_X  A   [17:198983[A  6    PASS SVTYPE=BND;MATEID=bnd_Z
17     198982 bnd_Y  A   A]2:321681]   6    PASS SVTYPE=BND;MATEID=bnd_W
17     198983 bnd_Z  C   [13:123457[C  6    PASS SVTYPE=BND;MATEID=bnd_X
```

Inserted Sequence

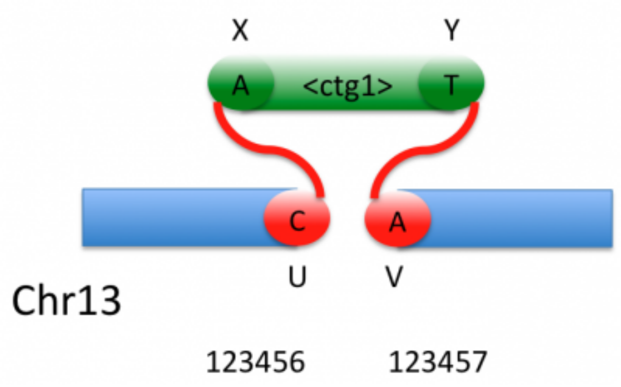
Sometimes, as shown below, some bases are inserted between the two breakends, this information is also carried in the ALT column:



```
#CHROM POS    ID     REF ALT                         QUAL FILT INFO
2      321682 bnd_V  T   ]13:123456]AGTNNNNNCAT       6    PASS SVTYPE=BND;MATEID=bnd_U
13     123456 bnd_U  C   CAGTNNNNNCA[2:321682[        6    PASS SVTYPE=BND;MATEID=bnd_V
```

Large Insertions

If the insertion is too long to be conveniently stored in the ALT column, as in the 329 base insertion below, it can be represented by a contig from the assembly file:



```
#CHROM POS    ID     REF ALT              QUAL FILT INFO
13     123456 bnd_U  C   C[<ctg1>:1[      6    PASS SVTYPE=BND
13     123457 bnd_V  A   ]<ctg1>:329]A    6    PASS SVTYPE=BND
```

Note: In the special case of the complete insertion of a sequence between two base pairs, it is recommended to use the shorthand notation described above:

```
#CHROM POS    ID     REF ALT              QUAL FILT INFO
13     123456 INS0   C   C<ctg1>          6    PASS SVTYPE=INS
```

If only a portion of <ctg1>, say from position 7 to position 214, is inserted, the VCF would be:

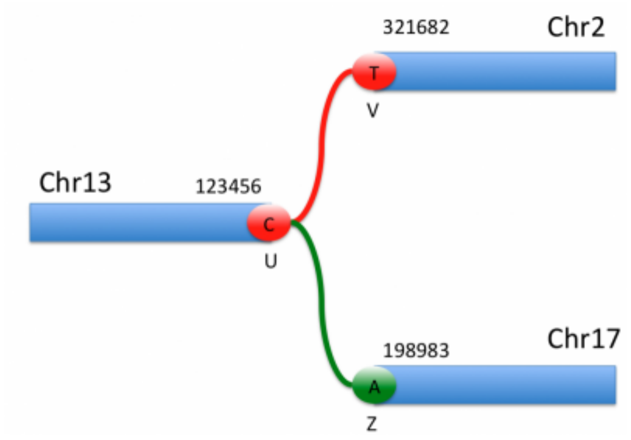
```
#CHROM POS    ID     REF ALT              QUAL FILT INFO
13     123456 bnd_U  C   C[<ctg1>:7[      6    PASS SVTYPE=BND
13     123457 bnd_V  A   ]<ctg1>:214]A    6    PASS SVTYPE=BND
```

If <ctg1> is circular and a segment from position 229 to position 45 is inserted, i.e. continuing from position 329 on to position 1, this is represented by adding a circular adjacency:

```
#CHROM POS    ID     REF ALT              QUAL FILT INFO
13     123456 bnd_U  C   C[<ctg1>:229[    6    PASS SVTYPE=BND
13     123457 bnd_V  A   ]<ctg1>:45]A     6    PASS SVTYPE=BND
<ctg1> 1      bnd_X  A   ]<ctg1>:329]A    6    PASS SVTYPE=BND
<ctg1> 329    bnd_Y  T   T[<ctg1>:1[      6    PASS SVTYPE=BND
```

Multiple mates

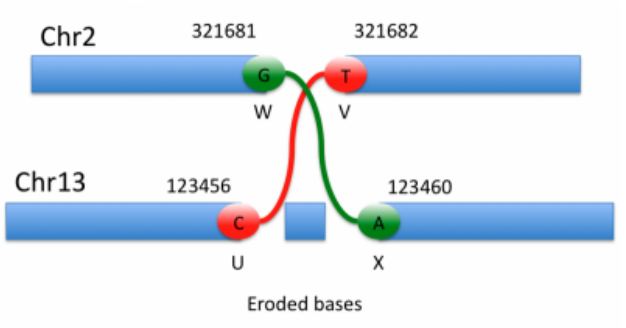
If a breakend has multiple mates (either because of breakend reuse or of uncertainty in the measurement), these alternate adjacencies are treated as alternate alleles:



```
#CHROM POS     ID      REF ALT                        QUAL FILT INFO
2      321682 bnd_V   T   ]13:123456]T               6    PASS SVTYPE=BND;MATEID=bnd_U
13     123456 bnd_U   C   C[2:321682[,C[17:198983[   6    PASS SVTYPE=BND;MATEID=bnd_V,bnd_Z
17     198983 bnd_Z   A   ]13:123456]A               6    PASS SVTYPE=BND;MATEID=bnd_U
```

Explicit partners

Two breakends which are connected in the reference genome but disconnected in the variants are called **partners**. Each breakend only has one partner, typically one basepair left or right. However, it is not uncommon to observe loss of a few basepairs during the rearrangement. It is then possible to explicitly name a breakend's partner. E.g.:

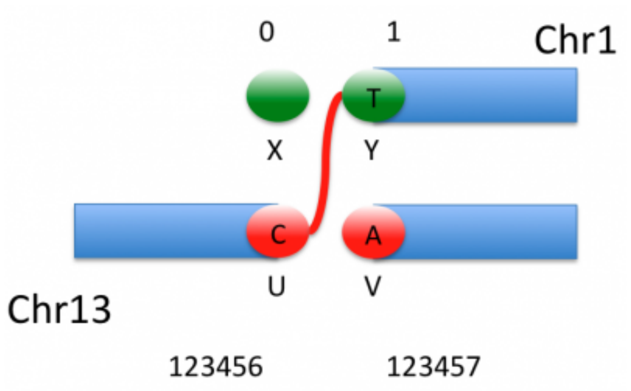


```
#CHROM  POS       ID      REF  ALT             INFO
2       321681    bnd_W   G    G[13:123460[    PARID=bnd_V;MATEID=bnd_X
2       321682    bnd_V   T    ]13:123456]T    PARID=bnd_W;MATEID=bnd_U
13      123456    bnd_U   C    C[2:321682[     PARID=bnd_X;MATEID=bnd_V
13      123460    bnd_X   A    ]2:321681]A     PARID=bnd_U;MATEID=bnd_W
```

Telomeres

For a rearrangement involving the telomere end of a reference chromosome, we define a virtual telomeric breakend that serves as a breakend partner for the breakend at the telomere. That way every breakend has a partner. If the chromosome extends from position 1 to N, then the virtual telomeric breakends are at positions 0 and N+1.

For example, to describe the reciprocal translocation of the entire chromosome 1 into chromosome 13, as illustrated here:

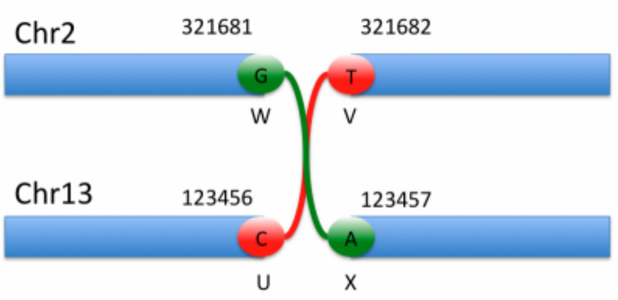


the records would look like:

```
#CHROM POS     ID      REF ALT              QUAL FILT INFO
1      0       bnd_X   N    .[13:123457[     6    PASS SVTYPE=BND;MATEID=bnd_V
1      1       bnd_Y   T    ]13:123456]T     6    PASS SVTYPE=BND;MATEID=bnd_U
13     123456  bnd_U   C    C[1:1[           6    PASS SVTYPE=BND;MATEID=bnd_Y
13     123457  bnd_V   A    ]1:0]A           6    PASS SVTYPE=BND;MATEID=bnd_X
```

Event Identifiers

As mentionned previously, a single rearrangement event can be described as a set of novel adjacencies. For example, a reciprocal rearrangement such as:

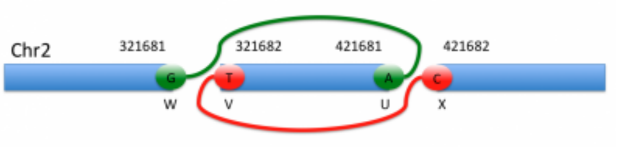


would be described as:

```
#CHROM POS     ID      REF ALT            QUAL FILT INFO
2      321681  bnd_W   G   G[13:123457[   6    PASS SVTYPE=BND;MATEID=bnd_X;EVENT=RR0
2      321682  bnd_V   T   ]13:123456]T   6    PASS SVTYPE=BND;MATEID=bnd_U;EVENT=RR0
13     123456  bnd_U   C   C[2:321682[    6    PASS SVTYPE=BND;MATEID=bnd_V;EVENT=RR0
13     123457  bnd_X   A   ]2:321681]A    6    PASS SVTYPE=BND;MATEID=bnd_W;EVENT=RR0
```

Inversion

Similarly an inversion such as:



can be described equivalently in two ways. Either one uses the short hand notation described previously (recommended for simple cases):

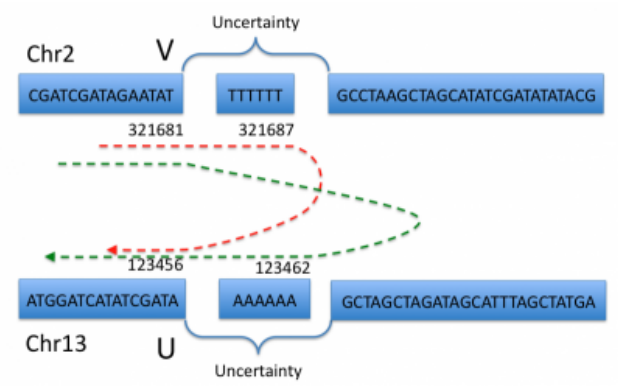
```
#CHROM POS     ID      REF ALT            QUAL FILT INFO
2      321682  INV0    T   <INV>          6    PASS SVTYPE=INV;END=421681
```

or one describes the breakends:

```
#CHROM POS     ID      REF ALT            QUAL FILT INFO
2      321681  bnd_W   G   G]2:421681]    6    PASS SVTYPE=BND;MATEID=bnd_U;EVENT=INV0
2      321682  bnd_V   T   [2:421682[T    6    PASS SVTYPE=BND;MATEID=bnd_X;EVENT=INV0
2      421681  bnd_U   A   A]2:321681]    6    PASS SVTYPE=BND;MATEID=bnd_W;EVENT=INV0
2      421682  bnd_X   C   [2:321682[C    6    PASS SVTYPE=BND;MATEID=bnd_V;EVENT=INV0
```

Uncertainty around breakend location

It sometimes is difficult to determine the exact position of a break, generally because of homologies between the sequences being modified. The breakend is then placed arbitrarily at the left most position, and the uncertainty is represented with the CIPOS tag. The ALT string is then constructed assuming this arbitrary breakend choice.



The figure above represents a nonreciprocal translocation with microhomology. Even if we know that breakend U is rearranged with breakend V, actually placing these breaks can be extremely difficult. The red and green dashed lines represent the most extreme possible recombination events which are allowed by the sequence evidence available. We therefore place both U and V arbitrarily within the interval of possibility:

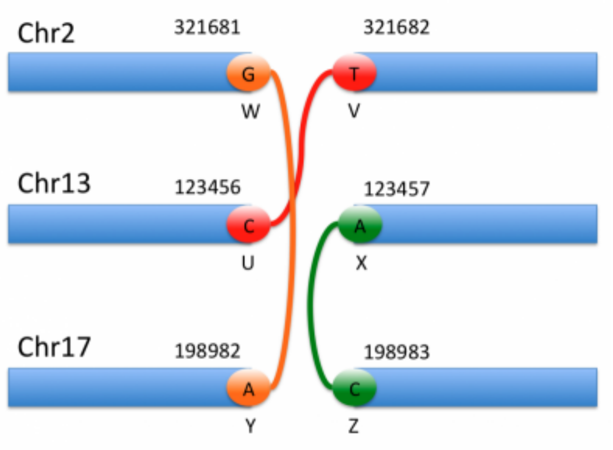
```
CHROM POS     ID      REF ALT            QUAL FILT INFO
2      321681 bnd_V  T    T]13:123462]   6    PASS SVTYPE=BND;MATEID=bnd_U;CIPOS=0,6
13     123456 bnd_U  A    A]2:321687]    6    PASS SVTYPE=BND;MATEID=bnd_V;CIPOS=0,6
```

Note that the coordinate in breakend U's ALT string does not correspond to the designated position of breakend V, but to the position that V would take if U's position were fixed (and vice-versa). The CIPOS tags describe the uncertainty around the positions of U and V.

The fact that breakends U and V are mates is preserved thanks to the MATEID tags. If this were a reciprocal translocation, then there would be additional breakends X and Y, say with X the partner of V on Chr 2 and Y the partner of U on Chr 13, and there would be two more lines of VCF for the XY novel adjacency. Depending on which positions are chosen for the breakends X and Y, it might not be obvious that X is the partner of V and Y is the partner of U from their locations alone. This partner relation ship can be specified explicitly with the tag PARID=bnd_X in the VCF line for breakend V and PARID=bnd_Y in the VCF line for breakend U, and vice versa.

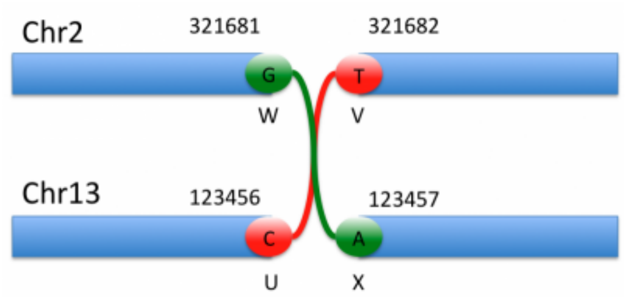
Single breakends

We allow for the definition of a breakend that is not part of a novel adjacency, also identified by the tag SVTYPE=BND. We call these single breakends, because they lack a mate. Breakends that are unobserved partners of breakends in observed novel adjacencies are one kind of single breakend. For example, if the true situation is known to be either as depicted in the figure below, and we only observe the adjacency (U,V), and no adjacencies for W, X, Y, or Z, then we cannot be sure whether we have a simple reciprocal translocation or a more complex 3-break operation. Yet we know the partner X of U and the partner W of V exist and are breakends. In this case we can specify these as single breakends, with unknown mates. The 4 lines of VCF representing this situation would be:



```
#CHROM POS     ID      REF ALT            QUAL FILT INFO
2      321681 bnd_W  G    G.             6    PASS SVTYPE=BND
2      321682 bnd_V  T    ]13:123456]T   6    PASS SVTYPE=BND;MATEID=bnd_U
13     123456 bnd_U  C    C[2:321682[    6    PASS SVTYPE=BND;MATEID=bnd_V
13     123457 bnd_X  A    .A             6    PASS SVTYPE=BND
```

On the other hand, if we know a simple reciprocal translocation has occurred as in below, then even if we have no evidence for the (W,X) adjacency, for accounting purposes an adjacency between W and X may also be recorded in the VCF file. These two breakends W and X can still be crossed-referenced as mates. The 4 VCF records describing this situation would look exactly as below, but perhaps with a special quality or filter value for the breakends W and X.



Another possible reason for calling single breakends is an observed but unexplained change in copy number along a chromosome.

```
#CHROM POS   ID    REF ALT   QUAL FILT INFO
3      12665 bnd_X A   .A    6    PASS SVTYPE=BND;CIPOS=-50,50
3      12665 .     A   <DUP> 14   PASS IMPRECISE;SVTYPE=DUP;END=13686;CIPOS=-50,50;CIEND=-50,50
3      13686 bnd_Y T   T.    6    PASS SVTYPE=BND;CIPOS=-50,50
```

Finally, if an insertion is detected but only the first few base-pairs provided by overhanging reads could be assembled, then this inserted sequence can be provided on that line, in analogy to paired breakends:

```
#CHROM POS   ID    REF ALT   QUAL FILT INFO
3      12665 bnd_X A   .TGCA 6    PASS SVTYPE=BND;CIPOS=-50,50
3      12665 .     A   <DUP> 14   PASS IMPRECISE;SVTYPE=DUP;END=13686;CIPOS=-50,50;CIEND=-50,50
3      13686 bnd_Y T   TCC.  6    PASS SVTYPE=BND;CIPOS=-50,50
```

8. Sample Mixtures

It may be extremely difficult to obtain clinically perfect samples, with only one type of cell. Let's imagine that two samples are taken from a cancer patient: healthy blood, and some tumor tissue with an estimated 30% stromal contamination. This would then be expressed in the header as:

```
##SAMPLE=<ID=Blood,Genomes=Germline,Mixture=1.,Description="Patient germline genome">
##SAMPLE=<ID=TissueSample,Genomes=Germline;Tumor,Mixture=.3,.7,Description="Patient germline genome;Pat
```

Because of this distinction between sample and genome, it is possible to express the data along both distinctions. For example, in a first pass, a structural variant caller would simply report counts per sample. Using the example of the inversion just above, the VCF code could become:

```
#CHROM POS    ID     REF ALT            QUAL FILT INFO                                FORMAT     Blood
2      321681 bnd_W  G   G]2:421681]    6    PASS SVTYPE=BND;MATEID=bnd_U;EVENT=INV0  GT:DPADJ   0:32
2      321682 bnd_V  T   [2:421682[T    6    PASS SVTYPE=BND;MATEID=bnd_X;EVENT=INV0  GT:DPADJ   0:29
13     421681 bnd_U  A   A]2:321681]    6    PASS SVTYPE=BND;MATEID=bnd_W;EVENT=INV0  GT:DPADJ   0:34
13     421682 bnd_X  C   [2:321682[C    6    PASS SVTYPE=BND;MATEID=bnd_V;EVENT=INV0  GT:DPADJ   0:31
```

However, a more evolved algorithm could attempt actually deconvolving the two genomes and generating copy number estimates based on the raw data:

```
#CHROM POS    ID     REF ALT            QUAL FILT INFO                                FORMAT     Blood
2      321681 bnd_W  G   G]2:421681]    6    PASS SVTYPE=BND;MATEID=bnd_U;EVENT=INV0  GT:CNADJ   0:1
2      321682 bnd_V  T   [2:421682[T    6    PASS SVTYPE=BND;MATEID=bnd_X;EVENT=INV0  GT:CNADJ   0:1
13     421681 bnd_U  A   A]2:321681]    6    PASS SVTYPE=BND;MATEID=bnd_W;EVENT=INV0  GT:CNADJ   0:1
13     421682 bnd_X  C   [2:321682[C    6    PASS SVTYPE=BND;MATEID=bnd_V;EVENT=INV0  GT:CNADJ   0:1
```

9. Clonal derivation relationships

In cancer, each VCF file represents several genomes from a patient, but one genome is special in that it represents the germline genome of the patient. This genome is contrasted to a second genome, the cancer tumor genome. In the simplest case the VCF file for a single patient contains only these two genomes. This is assumed in most of the discussion of the sections below.

In general there may be several tumor genomes from the same patient in the VCF file. Some of these may be secondary tumors derived from an original primary tumor. We suggest the derivation relationships between genomes in a cancer VCF file be represented in the header with PEDIGREE tags.

Analogously, there might also be several normal genomes from the same patient in the VCF (typically double normal studies with blood and solid tissue samples). These normal genomes are then considered to be derived from the original germline genome, which has to be inferred by parsimony.

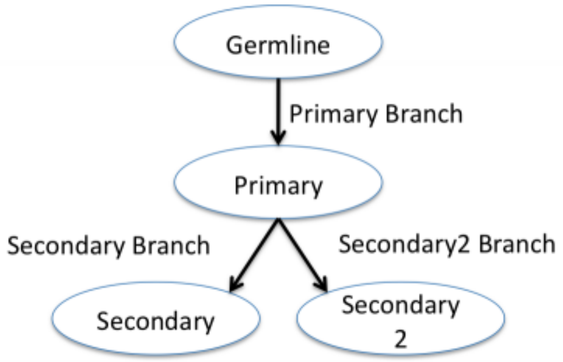
The general format of a PEDIGREE line describing asexual, clonal derivation is:

```
PEDIGREE=<Derived=ID2,Original=ID1>
```

This line asserts that the DNA in genome is asexually or clonally derived with mutations from the DNA in genome . This is the asexual analog of the VCF format that has been proposed for family relationships between genomes, i.e. there is one entry per of the form:

```
PEDIGREE=<Child=CHILD-GENOME-ID,Mother=MOTHER-GENOME-ID,Father=FATHER-GENOME-ID>
```

Let's consider a cancer patient VCF file with 4 genomes: germline, primary_tumor, secondary_tumor1, and secondary_tumor2. The primary_tumor is derived from the germline and the secondary tumors are each derived independently from the primary tumor, in all cases by clonal derivation with mutations. The PEDIGREE lines would look like:



```
##PEDIGREE=<Derived=PRIMARY-TUMOR-GENOME-ID,Original=GERMLINE-GENOME-ID>
##PEDIGREE=<Derived=SECONDARY1-TUMOR-GENOME-ID,Original=PRIMARY-TUMOR-GENOME-ID>
##PEDIGREE=<Derived=SECONDARY2-TUMOR-GENOME-ID,Original=PRIMARY-TUMOR-GENOME-ID>
```

Alternately, if data on the genomes is compiled in a database, a simple pointer can be provided:

```
##pedigreeDB=<url>
```

The most general form of a pedigree line is:

```
##PEDIGREE=<Name_0=G0-ID,Name_1=G1-ID,...,Name_N=GN-ID>
```

which means that the genome Name_0 is derived from the N >= 1 genomes Name_1, ..., Name_N. Based on these derivation relationships two new pieces of information can be specified.

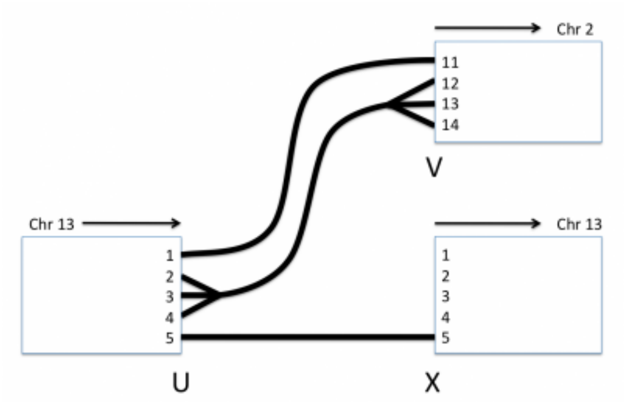
Firstly, we wish to express the knowledge that a variant is novel to a genome, with respect to its parent genome. Ideally, this could be derived by simply comparing the features on either genomes. However, insufficient data or sample mixtures might prevent us from clearly determining at which stage a given variant appeared. This would be represented by a mutation quality score.

Secondly, we define a **haplotype** as a set of variants which are known to be on the same chromosome in the germline genome. Haplotype identifiers must be unique across the germline genome, and are conserved along clonal lineages, regardless of mutations, rearrangements, or recombination. In the case of the duplication of a region within a haplotype, one copy retains the original haplotype identifier, and the others are considered to be novel haplotypes with their own unique identifiers. All these novel haplotypes have in common their **haplotype ancestor** in the parent genome.

10. Phasing adjacencies in an aneuploid context

In a cancer genome, due to duplication followed by mutation, there can in principle exist any number of haplotypes in the sampled genome for a given location in the reference genome. We assume each haplotype that the user chooses to name is named with a numerical haplotype identifier. Although it is difficult with current technologies to associate haplotypes with novel adjacencies, it might be partially possible to deconvolve these connections in the near future. We therefore propose the following notation to allow haplotype-ambiguous as well as haplotype-unambiguous connections to be described. The general term for these haplotype-specific adjacencies is **bundles**.

The following diagram will be used to support examples below:



In this example, we know that in the sampled genome

1. a reference bundle connects breakend U, haplotype 5 on chr13 to its partner, breakend X, haplotype 5 on chr13,
2. a novel bundle connects breakend U, haplotype 1 on chr13 to its mate breakend V, haplotype 11 on chr2, and finally,
3. a novel bundle connects breakend U, haplotypes 2, 3 and 4 on chr13 to breakend V, haplotypes 12, 13 or 14 on chr2 without any explicit pairing.

These three are the bundles for breakend U. Each such bundle is referred to as a haplotype of the breakend U. Each allele of a breakend corresponds to one or more haplotypes. In the above case there are two alleles: the 0 allele, corresponding to the adjacency to the partner X, which has haplotype (1), and the 1 allele, corresponding to the two haplotypes (2) and (3) with adjacency to the mate V.

For each haplotype of a breakend, say the haplotype (2) of breakend U above, connecting the end of haplotype 1 on a segment of Chr 13 to a mate on Chr 2 with haplotype 11, in addition to the list of haplotype-specific adjacencies that define it, we can also specify in VCF several other quantities. These include:

1. the depth of reads on the segment where the breakend occurs that support the haplotype, e.g. the depth of reads supporting haplotype 1 in the segment containing breakend U,
2. the estimated copy number of the haplotype on the segment where the breakend occurs,
3. the depth of paired-end or split reads that support the haplotype-specific adjacencies, e.g. that support the adjacency between haplotype 1 on Chr 13 to haplotype 11 on Chr 2
4. the estimated copy number of the haplotype-specific adjacencies, and
5. an overall quality score indicating how confident we are in this asserted haplotype.

These are specified using the using the DP, CN, BDP, BCN, and HQ subfields, respectively. The total information available about the three haplotypes of breakend U in the figure above may be visualized in a table as follows.

Allele	1	1	0
Haplotype	1>11	2,3,4>12,13,14	5>5
Segment Depth	5	17	4
Segment Copy Number	1	3	1
Bundle Depth	4	0	3
Bundle Copy Number	1	3	1
Haplotype qual.	30	40	40

up

Last updated by bob@charlemont.com on Monday 18th February, 2013 18:00.
Originally submitted by flicek@ebi.ac.uk on Tuesday 15th February, 2011 20:56.