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=<name=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species=Homo sapiens,taxond
##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=.,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      NA00001        NA
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,5
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,5
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|2:21:6:23,2
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,6
20     1234567 microsat1 GTC    G,GTCTC 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

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 3 bases (TCT), and the other an insertion of one base (A). 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, often as 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):

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

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. However, if the INFO field describes a pair of numbers, then this value should be 2 and so on. If the number of possible values varies, is unknown, or is unbounded, then this value should be '.'. Possible Types are: Integer, Float, Character, String and Flag. 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 and cannot itself contain a double-quote.

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.

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.

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. All entries for a specific CHROM should form a contiguous block within the VCF file. (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. (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 InDels or larger structural variants, the reference String must include the base before the event (which must be reflected in the POS field). 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>*”). 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) 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 (especially for structural variants)
   - 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, mitochondrion, only one allele value should be given. 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. 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 (Numeric)
- 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 (Numeric)
- GQ : conditional genotype quality, encoded as a phred quality -10log_10p(genotype call is wrong, conditioned on the site's being variant) (Numeric)
- HQ : haplotype qualities, two comma separated phred qualities (Numeric)
- 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) (Numeric)
- EC : expected alternate allele count (typically used in association analyses) (Numeric)

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.

How do I represent example variation in VCF records?

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     2 .         TCGCG      TCG,TCGCGCG    .   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 the equivalence among bases isn't provided by the VCF. Technically the following is an equivalent alignment:

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.

5. 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=-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).

6. 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 (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=-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">

7. 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=-1,Type=Float,Description="Copy number genotype likelihood for imprecise events">

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.

---

up

Last updated by ebanks@broadinstitute.org on Tuesday 22nd February, 2011 15:09.
Originally submitted by flicek@ebi.ac.uk on Tuesday 15th February, 2011 20:56.