DRAFT SPEC SUBJECT TO CHANGE
The Variant Call Format Specification

VCFv4.5 and BCFv2.2

20 Apr 2024

The master version of this document can be found at `https://github.com/samtools/hts-specs`. This printing is version 214572b from that repository, last modified on the date shown above.

... Continued from previous page

Key	Number	Type	Description
VALIDATED	0	Flag	Validated by follow-up experiment
1000G	0	Flag	1000 Genomes membership

Table 1: Reserved INFO keys

- END: End reference position (1-based), indicating the variant spans positions POS–END on reference/contig CHROM. Normally this is the position of the last base in the REF allele, so it can be derived from POS and the length of REF, and no END INFO field is needed. However when symbolic alleles are used, e.g. in gVCF or structural variants, an explicit END INFO field provides variant span information that is otherwise unknown. If a record containing a symbolic structural variant allele does not have an END field, it must be computed from the SVLEN field as per Section 3.

  This field is used to compute BCF's `rlen` field (see 6.3.1) and is important when indexing VCF/BCF files to enable random access and querying by position.

1.6.2 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 FORMAT keys matching the regular expression `^[A-Za-z_][0-9A-Za-z_.]*$`, duplicate keys are not allowed). This is followed by one data block per sample, with the colon-separated data corresponding to the types specified in the format. The first key must always be the genotype (GT) if it is present. If LGT key is present, it must be after GT (if also present) and before all others. There are no required keys. Additional Genotype keys can be defined in the meta-information, however, software support for them is not guaranteed.

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. If a field contains a list of missing values, it can be represented either as a single MISSING value ('.') or as a list of missing values (e.g. '.,.,.' if the field was Number=3). Trailing fields can be dropped, with the exception of the GT field, which should always be present if specified in the FORMAT field.

As with the INFO field, there are several common, reserved keywords that are standards across the community. See their detailed definitions below, as well as Table 2 for their reference Number, Type and Description. See also Section 4 for a list of genotype keys reserved for structural variants.

Field	Number	Type	Description
AD	R	Integer	Read depth for each allele
ADF	R	Integer	Read depth for each allele on the forward strand
ADR	R	Integer	Read depth for each allele on the reverse strand
DP	1	Integer	Read depth
EC	A	Integer	Expected alternate allele counts
END	1	Integer	End position on CHROM (used with multi-sample <*> alleles)
FT	1	String	Filter indicating if this genotype was "called"
GL	G	Float	Genotype likelihoods
GP	G	Float	Genotype posterior probabilities
GQ	1	Integer	Conditional genotype quality
GT	1	String	Genotype
HQ	2	Integer	Haplotype quality
LAA	.	Integer	Strictly increasing indices into REF and ALT, indicating which alternate alleles are relevant (local) for the current sample
LAD	.	Integer	Read depth for each of the local alternate alleles listed in LAA

Continued on next page...

Table 2: Reserved genotype keys

... Continued from previous page

Field	Number	Type	Description
LGT	.	String	Genotype against the local alleles
LPL	.	Integer	Phred-scaled genotype likelihoods rounded to the closest integer for genotypes that involve the local alternative alleles listed in LAA
MQ	1	Integer	RMS mapping quality
PL	G	Integer	Phred-scaled genotype likelihoods rounded to the closest integer
PP	G	Integer	Phred-scaled genotype posterior probabilities rounded to the closest integer
PQ	1	Integer	Phasing quality
PS	1	Integer	Phase set
PSL	P	String	Phase set list
PSO	P	Integer	Phase set list ordinal
PSQ	P	Integer	Phase set list quality

Table 2: Reserved genotype keys

- AD, ADF, ADR (Integer): Per-sample read depths for each allele; total (AD), on the forward (ADF) and the reverse (ADR) strand.
- DP (Integer): Read depth at this position for this sample.
- EC (Integer): 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.
- END (Integer): end position of the <*> reference block for this sample.
- FT (String): 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 semicolon-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. No whitespace or semicolons permitted.
- GQ (Integer): Conditional genotype quality, encoded as a phred quality $-10log_{10}$ p(genotype call is wrong, conditioned on the site's being variant).
- GP (Float): Genotype posterior probabilities in the range 0 to 1 using the same ordering as the GL field; one use can be to store imputed genotype probabilities.
- GT (String): Genotype, encoded as allele value preceded by either of / or | depending on whether that allele is considered phased. The first phasing indicator may be omitted and is implicitly defined as / if any phasing indicators are / and | otherwise. 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, /0/1, or 1/2, etc. Haploid calls, e.g. on Y, male non-pseudoautosomal X, or mitochondria, should be indicated by having only one allele value. A triploid call might look like 0/0/1, and a partially phased triploid call could be |0/1/2 to indicate that the first allele is phased with another variant in the VCF. If a call cannot be made for a sample at a given locus, '.' must be specified for each missing allele in the `GT` field (for example './.' for a diploid genotype and '.' for haploid genotype). The meanings of the phasing indicators are as follows (see the `PS` and `PSL` fields below for more details on incorporating phasing information into the genotypes):
  - ○ / : allele is unphased
  - ○ | : allele is phased (according to the phase-set indicated in `PS` or `PSL`)

  For symbolic structural variant alleles, GT=0 indicates the absence of any of the ALT symbolic structural variants defined in the record. Implementer should note that merging a VCF record containing only symbolic structural variant ALT alleles with a record containing other alleles will result a change of the meaning of the GT=0 haplotypes from the record containing only symbolic SVs.

- GL (Float): Genotype likelihoods comprised of comma separated floating point log_{10}-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; without GT field, diploidy is assumed.

  Genotype Ordering. In general case of ploidy P and N alternate alleles (0 is the REF and $1 \ldots N$ the alternate alleles), the ordering of genotypes for the likelihoods can be expressed by the following pseudocode with as many nested loops as ploidy: [‡]

```
for $a_P = 0 \ldots N$
  for $a_{P-1} = 0 \ldots a_P$
      ...
      for $a_1 = 0 \ldots a_2$
          println $a_1 a_2 \ldots a_P$
```

  Alternatively, the same can be achieved recursively with the following pseudocode:

```
Ordering($P$, $N$, suffix=""):
    for $a$ in $0 \ldots N$
        if ($P == 1$) println str($a$) + suffix
        if ($P > 1$) Ordering($P$-1, $a$, str($a$) + suffix)
```

  Conversely, the index of the value corresponding to the genotype $k_1 \leq k_2 \leq \ldots \leq k_P$ is

  $$\texttt{Index}(k_1/k_2/\ldots/k_P) = \sum_{m=1}^{P} \binom{k_m+m-1}{m}$$

  Examples:

  - for P=2 and N=1, the ordering is 00,01,11
  - for P=2 and N=2, the ordering is 00,01,11,02,12,22
  - for P=3 and N=2, the ordering is 000, 001, 011, 111, 002, 012, 112, 022, 122, 222
  - for P=1, the index of the genotype a is a
  - for P=2, the index of the genotype "a/b", where $a \leq b$, is $b(b+1)/2 + a$
  - for P=2 and arbitrary N, the ordering can be easily derived from a triangular matrix

$b \setminus a$	0	1	2	3
0	0			
1	1	2		
2	3	4	5	
3	6	7	8	9

- HQ (Integer): Haplotype qualities, two comma separated phred qualities.

- LAA is a sorted list of n distinct integers, where $0 \leq n \leq |\text{ALT}|$, giving the indices of the alleles that are observed in the sample. In callsets with many samples, sites may grow to include numerous alternate alleles at the same POS. Usually, few of these alleles are actually observed in any one sample, but each genotype must supply fields like PL and AD for all of the alleles—a very inefficient representation as PL's size is quadratic in the allele count. Similarly, in rare sites, which can be the bulk of the sites, the vast majority of the samples are reference. To prevent this growth in VCF size, one can choose to specify the genotype, allele depth and the genotype likelihood against a subset of "Local Alleles". LAA is the strictly increasing index into REF and ALT, pointing out the alleles that are actually in-play for that sample. 0 indicates the REF allele and should always be included with the subsequent values being 1-based indexes into ALT. LAD is the depth of the local alleles, LPL is subset of the PL array that pertains to the alleles that are referred to by LAA, LGT is the genotype but referencing the local alleles rather than the global ones. For example, if REF is G, ALT is A,C,T,`<*>` and a genotype only has information about G, C, and `<*>`, one can have LAA=[0,2,4] and thus LPL will be interpreted as pertaining to the alleles [G, C, `<*>`] and not contain likelihood values for genotypes that involve A or T. In this case LGT=0/1 means that the sample is G/C. GQ is still the genotype quality, even when the genotype is given against the local alleles. Note that reordering might be required and care need to be taken to reorder LAD and LPL appropriately. LAA is required in order to interpret LAD, LPL, and LGT. In the following

[‡] Note that we use inclusive `for` loop boundaries.

example, the records with the same POS encode the same information (some columns removed for clarity):

POS	REF	ALT	FORMAT	sample
1	G	A,C,T,<*>	LAA:LGT:LAD:LPL	0,2,4:1/1:20,30,10:90,80,0,100,110,120
1	G	A,C,T,<*>	GT:AD:PL	2/2:20,.,30,.,10:90,.,.,80,.,0,.,.,.,.,100,.,110,.,120
2	A	C,G,T,<*>	LAA:LGT:LAD:LPL	0,3:0/1:15,25:40,0,80
2	A	C,G,T,<*>	GT:AD:PL	0/3:15,.,.,25,.:40,.,.,.,.,.,0,.,.,80,.,.,.,.,.
3	C	G,T,<*>	LAA:LGT:LAD:LPL	0,3:0/0:30,1:0,30,80
3	C	G,T,<*>	GT:AD:PL	0/0:30,.,.,1:0,.,.,.,.,.,30,.,.,80
4	G	A,T,<*>	LAA:LGT:LAD:LPL	0:0/0:30:0
4	G	A,T,<*>	GT:AD:PL	0/0:30,.,.,.:0,.,.,.,.,.,.,.,.,.

- LAD: is a list of n integers giving read depths (as per AD) for each of the local alleles as listed in LAA.

- LGT: is the genotype, encoded as allele indexes separated by either of / or |, as with GT, however, the indexes are into the alleles referenced by LAA. So that in the case that LAA is 0,2,3, LGT=0/2 is equivalent to GT=0/3 and LGT=1/2 is equivalent to GT=2/3 (see example above).

- LPL: is a list of $\binom{n}{\text{Ploidy}}$ integers giving phred-scaled genotype likelihoods (rounded to the closest integer; as per PL) for all possible genotypes given the set of alleles defined in the LAA local alleles. The precise ordering is defined in the GL paragraph.

- MQ (Integer): RMS mapping quality, similar to the version in the INFO field.

- PL (Integer): The phred-scaled genotype likelihoods rounded to the closest integer, and otherwise defined in the same way as the GL field.

- PP (Integer): The phred-scaled genotype posterior probabilities rounded to the closest integer, and otherwise defined in the same way as the GP field.

- PQ (Integer): 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.

- PS (non-negative 32-bit Integer): Phase set, 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).

- PSL (List of Strings): The list of phase sets, one for each allele specified in the `GT` or `LGT`. Unphased alleles (without a | separator before them) must have the value '.' in their corresponding position in the list. Unlike `PS` (which is defined per CHROM), records with different CHROM but the same phase-set name are considered part of the same phase set. If an implementation cannot guarantee uniqueness of phase-set names across the VCF (for example, phasing a streaming VCF or each CHROM is processed independently in parallel), new phase-set names should be of the format CHROM*POS*ALLELE-NUMBER of the "first" allele which is included in this set, with ALLELE-NUMBER being the index of the allele in the `GT` field, since multiple distinct phase-sets could start at the same position. [§] A given sample-genotype must not have values for both PS and PSL. In addition, PS and PSL are not interoperable, in that a PS mentioned in one variant cannot be referenced in a PSL in another, since when used in PS it isn't connected to any specific haplotype (i.e. first or second), but PSL is.

  Example:

<table>
<tr><th>#CHROM</th><th>POS</th><th>ID</th><th>REF</th><th>ALT</th><th>QUAL</th><th>FILTER</th><th>INFO</th><th>FORMAT</th><th>SAMPLE1</th></tr>
<tr><td>chr19</td><td>5</td><td>.</td><td>T</td><td>G</td><td>.</td><td>PASS</td><td>DP=100</td><td>GT:PSL</td><td><code>|0/1:chr9*5*1,.</code></td></tr>
<tr><td>chr20</td><td>10</td><td>.</td><td>A</td><td>T,G</td><td>.</td><td>PASS</td><td>DP=100</td><td>GT:PSL</td><td><code>|1/2|3:chr20*10*1,.,chr9*5*1</code></td></tr>
<tr><td>chr20</td><td>15</td><td>.</td><td>G</td><td>C</td><td>.</td><td>PASS</td><td>DP=100</td><td>GT:PSL</td><td><code>1|2:.,chr20*10*1</code></td></tr>
</table>

[§] The '*' character is used as a separator since ':' is not reserved in the CHROM column.

5.5 Representing unspecified alleles and REF-only blocks (gVCF)

In order to report sequencing data evidence for both variant and non-variant positions in the genome, the VCF specification allows to represent blocks of reference-only calls in a single record using the END INFO tag, an idea originally introduced by the gVCF file format[¶].

The convention adopted here is to represent reference evidence as likelihoods against an unknown alternate allele represented as <*>. Think of this as the likelihood for reference as compared to any other possible alternate allele (both SNP, indel, or otherwise). The <*> representation is preferred over the symbolic allele <NON_REF>.

Example records are given below:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample
1	4370	.	G	<*>	.	.	END=4383	GT:DP:GQ:MIN_DP:PL	0/0:25:60:23:0,60,900
1	4384	.	C	<*>	.	.	END=4388	GT:DP:GQ:MIN_DP:PL	0/0:25:45:25:0,42,630
1	4389	.	T	TC,<*>	213.73	.	.	GT:DP:GQ:PL	0/1:23:99:51,0,36,93,92,86
1	4390	.	C	<*>	.	.	END=4390	GT:DP:GQ:MIN_DP:PL	0/0:26:0:26:0,0,315
1	4391	.	C	<*>	.	.	END=4395	GT:DP:GQ:MIN_DP:PL	0/0:27:63:27:0,63,945
1	4396	.	G	C,<*>	0	.	.	GT:DP:GQ:P	0/0:24:52:0,52,95,66,95,97
1	4397	.	T	<*>	.	.	END=4416	GT:DP:GQ:MIN_DP:PL	0/0:22:14:22:0,15,593

5.5.1 Multi-sample REF-only blocks

When handling VCFs with multiple samples, the length of the <*> reference blocks can differ. To account for this, a sample-specific END can be specified via the FORMAT END field. If any FORMAT END value exists, the INFO END must be present and equal the largest FORMAT END value. Positions implicitly called by a preceding <*> for a sample must have *GT*/*LGT* set to the missing value ('.') and have no other FORMAT fields present. If *LAA* is present and a reference block is defined for a given sample, the <*> allele must be included as an *LAA* allele for that sample even though the *LGT* is 0/0.

For example, the genotype-only version of the above example with a second sample with no variants:

POS	REF	ALT	INFO	FORMAT	SampleA	SampleB
4370	G	<*>	END=4416	LGT:LAA:END	0/0:0,1:4388	0/0:0,1:4416
4389	T	TC	.	LGT:LAA:END	0/1:0,1:.	.
4390	C	<*>	END=4416	LGT:LAA:END	0/0:0,1:4416	.

[¶] https://help.basespace.illumina.com/articles/descriptive/gvcf-files/

5.6 Representing copy number variation

To encode copy number variation, VCF uses <CNV>, <DEL> and <DUP> symbolic structural variant alleles, CN INFO and FORMAT fields.

Allele specific copy number is specified through a <CNV> ALT allele for each distinct allelic copy number. INFO CN defines the allele specific copy number with FORMAT CN defining the overall copy number for that sample. POS and INFO SVLEN specify the genomic interval over which the copy number is defined. <DEL> and <DUP> copy number (SVCLAIM=D) alleles should be treated as <CNV> alleles that implicitly define INFO CN=0 and ~~INFO CICN~~CN=2, ~~.~~respectively. As with all symbolic structural variants, the starting position of the interval is the base immediately after POS. For example, a region on chr1 from position 101 to 130 (both inclusive) with allele-specific copy numbers of 1 and 2 can be represented as follows:

```
chr1 100 . T <CNV>,<CNV> . . END=130;SVLEN=30,30;CN=1,2 GT:CN 1/2:3
```

All <CNV> alleles in the same VCF record should have the same SVLEN. To eliminate genotype ambiguity, copy number ALT alleles should not be mixed with other ALT alleles. When only copy number ALT alleles are present in a VCF record, GT=0 is equivalent to a <CNV> ALT allele with INFO CN of 1 and should be treated identically.

If only total copy number is known, the copy number of the segment should be defined with a single <CNV> ALT allele with a missing INFO CN field. In the above example this corresponds to the following:

```
chr1 100 . T <CNV> . . END=130;SVLEN=30 GT:CN .:3
```

The granularity of copy number representation is explicitly not defined in these specifications. Copy number segmentation can be base-pair accurate with even 1bp changes deletions resulting in new copy number segments, be at a highly granular megabase level of resolution, or anywhere in between. When the bounds of a copy number segment is not known precisely, this should be encoded in the CIPOS and CILEN INFO fields.

0x33000000	l_shared as 32-bit little endian hex
0x2A000000	l_indiv as 32-bit little endian hex
0x01000000	CHROM offset is at 1 in 32 bit little endian
0x64000000	POS in 0-based 32-bit little endian
0x01000000	rlen = 1 (it's just a SNP)
0x41 0xF0 0xCC 0xCD	QUAL = 30.1 as 32-bit float
0x0400	n_info as 16-bit little-endian
0x0200	n_allele as 16-bit little-endian
0x030000	n_sample as 24-bit little-endian
0x05	n_fmt
0x57 0x72 0x73 0x31 0x32 0x33	ID = rs123
0x17 0x41	REF A
0x17 0x43	ALT C
0x11 0x00	FILTER field PASS
0x11 0x50 0x00	HM3 flag is present
0x11 0x51	AC key
0x11 0x03	with value of 3
0x11 0x52	AN key
0x11 0x06	with value of 6
0x11 0x53	AA key
0x17 0x43	with value of C
0x1101 0x21 0x020202040404	GT
0x1102 0x11 0x0A0A0A	GQ
0x1103 0x11 0x203040	DP
0x1104 0x21 0x300030200040	AD
0x1105 0x31 0x000A640A0064640A00	PL

That's quite a lot of information encoded in only 96 bytes!

6.5 BCF2 block gzip and indexing

These raw binary records may be subsequently encoded into BGZF blocks following the BGZF compression format, section 3 of the SAM format specification. BCF2 records can be raw, though, in cases where the decoding/encoding costs of bgzipping the data make it reasonable to process the data uncompressed, such as streaming BCF2s through pipes with samtools and bcftools. Here the files should be still compressed with BGZF but with compression 0. Implementations should perform BGZF encoding and must support the reading of both raw and BGZF encoded BCF2 files.

BCF2 files are expected to be indexed through the same index scheme, section 4 as BAM files and other block-compressed files with BGZF.

7 List of changes

7.1 Changes between VCFv4.5 and VCFv4.4

- Added local allele support (FORMAT LAA, LGT, LAD, LPL) to reduce the size of multi-sample VCFs and enable lossless merging.
- Added FORMAT END to support sample-specific <*> alleles.

7.2 Changes between VCFv4.4 and VCFv4.3

- Added tandem repeat support (<CNV:TR>, RN, RUS, RUL, RB, CIRB, RUC, CIRUC, RUB)
- Redefined INFO CN as allele-specific copy number and FORMAT CN as total copy number.
- Redefined INFO and FORMAT CN to support non-integer copy numbers.
- Added support for phasing and derivative chromosome reconstruction in the presence of SVs (PSL, PSO, PSQ)