... Continued from previous page

Field	Number	Type	Description
LADR	LR	Integer	Local-allele representation of ADR
LEC	LA	Integer	Local-allele representation of EC
LGL	LG	~~Integer~~ Float	Local-allele representation of GL
LGP	LG	~~Integer~~ Float	Local-allele representation of GP
LPL	LG	Integer	Local-allele representation of PL
LPP	LG	Integer	Local-allele representation of PP
M[0-9]+[ACGTUN]	M	Float	Fraction of bases modified with the given ChEBI ID.
DPM[0-9]+[ACGTUN]	M	Integer	Total read depth for reads able to detect the base modification with the given ChEBI ID.
ADM[0-9]+[ACGTUN]	M	Integer	Read depth for reads with the base modification with the given ChEBI ID.
M5mC	M	Float	Alias for M27551C 5-Methylcytosine
DPM5mC	M	Integer	Alias for DPM27551C
ADM5mC	M	Integer	Alias for ADM27551C
M5hmC	M	Float	Alias for M76792C 5-Hydroxymethylcytosine
DPM5hmC	M	Integer	Alias for DPM76792C
ADM5hmC	M	Integer	Alias for ADM76792C
M5fC	M	Float	Alias for M76794C 5-Formylcytosine
DPM5fC	M	Integer	Alias for DPM76794C
ADM5fC	M	Integer	Alias for ADM76794C
M5caC	M	Float	Alias for M76793C 5-Carboxylcytosine
DPM5caC	M	Integer	Alias for DPM76793C
ADM5caC	M	Integer	Alias for ADM76793C
M5hmU	M	Float	Alias for M16964T 5-Hydroxymethyluracil
DPM5hmU	M	Integer	Alias for DPM16964T
ADM5hmU	M	Integer	Alias for ADM16964T
M5fU	M	Float	Alias for M80961T 5-Formyluracil
DPM5fU	M	Integer	Alias for DPM80961T
ADM5fU	M	Integer	Alias for ADM80961T
M5caU	M	Float	Alias for M17477T 5-Carboxyluracil
DPM5caU	M	Integer	Alias for DPM17477T
ADM5caU	M	Integer	Alias for ADM17477T
M6mA	M	Float	Alias for M28871A 6-Methyladenine
DPM6mA	M	Integer	Alias for DPM28871A
ADM6mA	M	Integer	Alias for ADM28871A
M8oxoG	M	Float	Alias for M44605G 8-Oxoguanine
DPM8oxoG	M	Integer	Alias for DPM44605G
ADM8oxoG	M	Integer	Alias for ADM44605G
MXaoN	M	Float	Alias for M18107N Xanthosine
DPMXaoN	M	Integer	Alias for DPM18107N
ADMXaoN	M	Integer	Alias for ADM18107N
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

Continued on next page...

Table 2: Reserved genotype keys

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 <*> allele and the FORMAT LEN field. 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).

Positions implicitly called by a preceding <*> for a sample must have *GT* set to the missing value ('.') and have no FORMAT fields other than *LAA* present. If *LAA* is present and a reference block start is being defined for a given sample, the <*> allele must be included as an *LAA* allele for that sample even though the *GT* is 0/0.

Reference blocks were originally introduced by the gVCF file format[¶]. Unfortunately, gVCF has issues scaling to many samples as the use of INFO END to encode the reference block length requires the reference block length to be the same for all samples.

To retain backwards compatibility with with gVCF, the symbolic allele <NON_REF> should be treated as an alias of <*> and a missing FORMAT LEN field should be inferred from the INFO END tag if present.

An example with both FORMAT LEN and a redundant INFO END is given below:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample
1	4370	.	G	<*>	.	.	END=4383	GT:DP:GQ:MIN_DP:PL:LEN	0/0:25:60:23:0,60,9
1	4384	.	C	<*>	.	.	END=4388	GT:DP:GQ:MIN_DP:PL:LEN	0/0:25:45:25:0,42,6
1	4389	.	T	TC ~~,<*>~~	~~213.73~~ .	.	GT:DP:GQ:PL ~~:LEN~~	0/1:23:99:51,0,36,93,92,86	
1	4390	.	C	<*>	.	.	END=4390	GT:DP:GQ:MIN_DP:PL:LEN	0/0:26:0:26:0,0,315
1	4391	.	C	<*>	.	.	END=4395	GT:DP:GQ:MIN_DP:PL:LEN	0/0:27:63:27:0,63,9
1	4396	.	G	C ~~,<*>~~	~~0~~ .	.	GT:DP:GQ:MIN_DP:PL ~~:LEN~~	0/0:24:52:0,52,95,66,95,97	
1	4397	.	T	<*>	.	.	END=4416	GT:DP:GQ:MIN_DP:PL:LEN	0/0:22:14:22:0,15,5

When base modification information is present in the FORMAT field of a reference block record, the base modification information apply to all applicable bases covered by that reference block.

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

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 VCFv4.4 Errata

- Fixed typos in gVCF example.

7.2 Changes between VCFv4.5 and VCFv4.4

- Added base modification support (FORMAT M5mC, M5hmC, M6mA, etc.).
- Reserved all FORMAT keys of the form $M[0-9]+$ as base modification fields.
- Added Number=P support for fields with cardinality matching sample ploidy/local copy number.
- Added local allele support (Number=LA, LG, LR; FORMAT LAA, LAD, LADF, LADR, LEC, LGL, LGP, LPL, LPP) to reduce the size of multi-sample VCFs and enable lossless merging.