Exome+ Full Data Set: Data File Format

Overview

Helix provides hg38 aligned sequencing data in a variation of VCF called gVCF. (For more information on the VCF text file format, see https://samtools.github.io/hts-specs/VCFv4.2.pdf.) gVCF is very similar to VCF version 4.2 except for one key difference: it represents every base of the 24 chromosomes and the mitochondrial genome. This is accomplished through the inclusion of reference blocks.

A reference block is the stretch of DNA where a user has the same sequence as the reference genome. Normally, a VCF entry indicates that the position described is different from the reference genome in some way. By only reporting variants, typical VCFs do not differentiate between instances where the absence of a variant is due to the lack of variation from the reference or where it is due to a lack of coverage (sensitivity) at that location. Explicit reference calls and no-calls are indistinguishable. However, gVCF corrects for this loss of information by including reference block records to represent areas of reference or low quality/coverage. For more information about the gVCF format, see https://software.broadinstitute.org/gatk/documentation/article.php?id=4017.

Example of a header for the Full Data Set gVCF:

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in
the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF
block">
##INFO=<ID=END,Number=1,Type=Integer,Description="Stop position of the interval">
##contig=<ID=chr1,length=248956422,assembly=hg38>
##contig=<ID=chr2,length=242193529,assembly=hg38>
##contig=<ID=chr3,length=198295559,assembly=hg38>
##contig=<ID=chr4,length=190214555,assembly=hg38>
##contig=<ID=chr5,length=181538259,assembly=hg38>
##contig=<ID=chr6,length=170805979,assembly=hg38>
##contig=<ID=chr7,length=159345973,assembly=hg38>
##contig=<ID=chr8,length=145138636,assembly=hg38>
##contig=<ID=chr9,length=138394717,assembly=hg38>
##contig=<ID=chr10,length=133797422,assembly=hg38>
##contig=<ID=chr11,length=135086622,assembly=hg38>
##contig=<ID=chr12,length=133275309,assembly=hg38>
##contig=<ID=chr13,length=114364328,assembly=hg38>
##contig=<ID=chr14,length=107043718,assembly=hg38>
##contig=<ID=chr15,length=101991189,assembly=hg38>
##contig=<ID=chr16,length=90338345,assembly=hg38>
##contig=<ID=chr17,length=83257441,assembly=hg38>
##contig=<ID=chr18,length=80373285,assembly=hg38>
##contig=<ID=chr19,length=58617616,assembly=hg38>
##contig=<ID=chr20,length=64444167,assembly=hg38>
##contig=<ID=chr21,length=46709983,assembly=hg38>
##contig=<ID=chr22,length=50818468,assembly=hg38>
##contig=<ID=chrX,length=156040895,assembly=hg38>
##contig=<ID=chrY,length=57227415,assembly=hg38>
##contig=<ID=chrM,length=16569,assembly=hg38>
##contig=<ID=chrM_rsrs,length=16569,assembly=hg38>
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phaseset ID">
##FILTER=<ID=ANEUPLOID,Description="Called genotype does not agree with expected ploidy">
##FILTER=<ID=HETEROPLASMY,Description="Called heterozygous genotype on mitochondrial contig">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Estimated Genotype Probability">
##FILTER=<ID=IMP,Description="Set if true: IMP==1">
##FILTER=<ID=BOOSTED,Description="Set if true: BOOSTED==1">
##FILTER=<ID=LOWDP,Description="Set if GQ>20 and 10<=DP<=20">
##FILTER=<ID=LOWQ,Description="Set if GQ<=20 or DP<10">
##FILTER=<ID=NOTVALIDATED,Description="Set if variant falls outside of analytic range">
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype likelihoods">
##FORMAT=<ID=VAR_TYPE,Number=.,Type=String,Description="Variant type: SNV, INSERTION, DELETION,
SUBSTITUTION, MNV, COMPLEX">
##FORMAT=<ID=VAR_CONTEXT,Number=.,Type=String,Description="Variant genomic context: STR-
expansion, STR-contraction, STR-proximal">
##FORMAT=<ID=STR_MAX_LEN,Number=1,Type=Integer,Description="Maximum observed STR sequence
length">
##FORMAT=<ID=STR_PERIOD,Number=1,Type=Integer,Description="Repetition period for STR variants">
##FORMAT=<ID=STR_TIMES,Number=1,Type=Float,Description="Number of repetition for STR variants">
##pipeline=helix-v2.6.1
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  YOUR_UNIQUEID
```

Detailed explanation of header and fields in the gVCF:

fileformat	VCF file format. Will be `VCFv4.2`. For more details, see https://samtools.github.io/hts-specs/VCFv4.2.pdf.
contig	Chromosome name. This file includes the 22 autosomal chromosomes as well as chrX, chrY, chrM, and chrM_rsrs. All locations are aligned to the hg38 reference genome (as stated in each contig header line).
ALT = <NON_REF>	Alternate sequences are anything but the reference sequence. Since gVCF files cover reference blocks, entries exist to represent the reference (GT = `0/0`) or not-called regions (GT = `./.`). VCF format requires something in the ALT field; `<NON_REF>` was introduced to represent "everything except reference". Note: `<NON_REF>` should never be in the ALT field if GT > 0 (for example: `0/1, 1/1, 1/2,` etc.).
FILTER = PASS	Pass. A call made from observed read data that passes the Helix Exome+ minimum quality thresholds is assigned this value. This will be the most common FILTER flag that occurs for a called variant. It means that the variant is called with high confidence. Only PASS variants are clinically validated and land within the Helix analytical range. The main criteria is that GQ > 20 and DP >=20 in autosomes or DP >=10 in chrX & chrY in males.
FILTER = IMP	Imputed. The variant was statistically inferred based on the genotype of other surrounding variants. However, DP = 0 over the variant itself.
FILTER = BOOSTED	Boosted. The variant was statistically inferred based on the genotype of other surrounding variants. It also has some coverage over the variant (0 < DP < 20) supporting the variant call, although this coverage was not high enough to confidently call the variant on its own.
FILTER = LOWDP	Low depth. Set if: GQ > 20, 10 <= DP < 20. This is a variant that at least 10 but less than 20 high quality reads aligned to it, while genotype quality is greater than 20.

FILTER = LOWQ	Low quality. Set if GQ <=20 or DP < 10 and not IMP or BOOSTED. This includes two different types of variants: 1. Variants that have a genotype call (GQ > 20) but fewer than 10 reads. 2. Variants that are no-calls (GT = ./.) because of low GQ (GQ <= 20).
FILTER = ANEUPLOID	Aneuploid. Ploidy is defined as the number of sets of chromosomes. Most people have two copies of every chromosome (except X and Y in males and mitochondrial). This flag occurs when the called genotype does not agree with the expected ploidy.
FILTER = HETEROPLASMY	Heteroplasmy. Called heterozygous genotype on mitochondrial contig. At this time, heteroplasmy is not supported and so this flag occurs when the called genotype does not agree with the expected ploidy for the mitochondrial chromosome specifically.
FILTER = NOTVALIDATED	Not validated. Set if variant falls outside of analytic range due to genomic context, such as position within a short tandem repeat.
INFO = END	Reference block end. If the INFO field has END in it, then this line of the gVCF refers to a block where the POS is the start and the END is the end (inclusive). Metrics are included in the FORMAT field to facilitate information about this block (ex: MIN_DP). If the INFO field has "`.`", then the line is not referring to a block of the genome and treat it as a normal VCF line.
FORMAT = GT	Genotype called. The length of this array is equal to the number of copies of each chromosome present in the individual (i.e. ploidy). The values are: * `.` if an allele was not called (GQ <=20). * `0` if the reference allele was observed. * `1` if the first alternative allele (from ALT) was observed. * `2` if the second alternative allele was observed. If no genotype assignment could be made for the individual, the value of this field will be `./.` for autosomes and GQ will be <=20. For diploid bi-allelic variants, i.e. variants with only one alternative allele, the value of this field will be 0/0 for two copies of the reference, 0/1 for one copy of the reference and one copy of the

	alternate, or 1/1 for two copies of the alternate allele. For diploid heterozygous alternate variants, the values will be 1/2 indicating that the individual has one copy of each of the two alternative alleles. For mitochondria, the Y chromosome, and the X chromosome in men, genotype would have only one element: `.`, `0`, or `1`. If the variants are in phase with one another (see PS for more details), then the `/` will be replaced with `|`.
FORMAT = DP	Read depth. Number of reads mapped to this location.
FORMAT = AD	Allele depth(s). For the REF and each ALT, the value of this field gives the total number of reads that harbor the allele at the genomic position(s) of the variant. The first value corresponds to the REF allele, the second to the first alternative allele, the third to the second ALT allele, and so on.
FORMAT = MIN_DP	Minimum depth. The minimum number of reads within the reference block overlapping with the POS and END coordinates.
FORMAT = GQ	Genotype quality. Conditional genotype quality. Encoded as a phred quality –10*log10 p (genotype call is wrong, conditioned on the site's being variant).
FORMAT = GP	Genotype probability. The probabilities that the imputed genotype is correct. The range is 0 to 1. If there is only one ALT, there will be 3 numbers referring to genotypes: `0/0`, `0/1`, and `1/1`.
FORMAT = GL	Genotype likelihood. log10-scaled likelihoods for all possible genotypes given the set of alleles defined in the REF and ALT fields. The most likely genotype will have a genotype likelihood of 0 and each increasingly less likely call will have a decreasing negative value. In short, if REF is A and ALT is C, then the ordering of diploid genotype likelihoods corresponds to genotypes AA, AC, CC. For heterozygous alternate alleles with ALT is C,G, the ordering is: AA, AC, CC, AG, CG, GG. See Section 1.6.2 of https://samtools.github.io/hts-specs/VCFv4.2.pdf for further information.

FORMAT = PS	Phaseset ID. The identification number of the phaseset, which indicates the starting location of the phaseset (number only, does not include contig name).
FORMAT = VAR_TYPE	Variant type. Indicates the type of variant observed in the record. Can be SNV, INSERTION, DELETION, MNV, SUBSTITUTION, or COMPLEX.
FORMAT = VAR_CONTEXT	Variant context. Identifies if the variant is near a STR (short tandem repeat, 1-3 bp unit repeat length). Possible values are STR-expansion (expands the STR), STR-contraction (reduces the STR), STR-proximal (variant does not follow the pattern of the STR).
FORMAT = STR_PERIOD	STR period. Repetition period for short tandem repeat (STR) variants.
FORMAT = STR_TIMES	STR times. Number of repetition for short tandem repeat (STR) variants.
FORMAT = STR_MAX_LEN	STR maximum length. Maximum observed short tandem repeat (STR) sequence length.
pipeline	Pipeline version. Helix's bioinformatic pipeline version that was used to create this gVCF.