Variant Calling

Arrhythmia Panel Analysis

Anish S. Shah, MD, MS

2023-03-01

Overview

We are using a specific patient example named `UIC0003` in this case.

This is an example of what is contained within an VCF file. Primarily will review the header information here.

```
class: CollapsedVCF dim: 165521 1 rowRanges(vcf): GRanges with 5
metadata columns: paramRangeID, REF, ALT, QUAL, FILTER info(vcf):
DataFrame with 44 columns: AC, AF, AN, AS_BaseQRankSum, AS_FS,
AS_FilterSt… info(header(vcf)): Number Type Description
AC A Integer Allele count in genotypes, for each … AF A Float Allele Fre-
quency, for each ALT allel… AN 1 Integer Total number of alleles in called ge…
AS_BaseQRankSum A Float allele specific Z-score from Wilcoxo… AS_FS
A Float allele specific phred-scaled p-value… AS_FilterStatus A String Filter
status for each allele, as as… AS_InbreedingCoeff A Float Allele-specific
inbreeding coefficie… AS_MQ A Float Allele-specific RMS Mapping Quality
AS_MQRankSum A Float Allele-specific Mapping Quality Rank… AS_QD
A Float Allele-specific Variant Confidence/Q… AS_QUALapprox 1 String
Allele-specific QUAL approximations
AS_RAW_BaseQRankSum 1 String raw data for allele specific rank
su… AS_RAW_MQ 1 String Allele-specfic raw data for RMS Mapp…
AS_RAW_MQRankSum 1 String Allele-specfic raw data for Mapping …
AS_RAW_ReadPosRankSum 1 String allele specific raw data for rank
su… AS_ReadPosRankSum A Float allele specific Z-score from Wilcoxo…
AS_SB_TABLE 1 String Allele-specific forward/reverse read… AS_SOR A
Float Allele specific strand Odds Ratio of… AS_VQSLOD A String For each alt
allele, the log odds of… AS_VarDP 1 String Allele-specific (informative) depth
… AS_culprit A String For each alt allele, the annotation … BaseQRankSum 1
Float Z-score from Wilcoxon rank sum test … DB 0 Flag dbSNP Membership
DP 1 Integer Approximate read depth; some reads m… END 1 Integer Stop
position of the interval
ExcessHet 1 Float Phred-scaled p-value for exact test … FS 1 Float Phred-
```

scaled p-value using Fisher’s … InbreedingCoeff 1 Float Inbreeding coefficient as estimated … MLEAC A Integer Maximum likelihood expectation (MLE)… MLEAF A Float Maximum likelihood expectation (MLE)… MQ 1 Float RMS Mapping Quality

MQRankSum 1 Float Z-score From Wilcoxon rank sum test … MQ_DP 1 Integer Depth over variant samples for bette… NEGATIVE_TRAIN_SITE 0 Flag This variant was used to build the n… POSITIVE_TRAIN_SITE 0 Flag This variant was used to build the p… QD 1 Float Variant Confidence/Quality by Depth

QUALapprox 1 Integer Sum of PL[0] values; used to approxi… RAW_GT_COUNT 3 Integer Counts of genotypes w.r.t. the refer… RAW_MQandDP 2 Integer Raw data (sum of squared MQ and tota… ReadPosRankSum 1 Float Z-score from Wilcoxon rank sum test … SOR 1 Float Symmetric Odds Ratio of 2x2 continge… VQSLOD 1 Float Log odds of being a true variant ver… VarDP 1 Integer (informative) depth over variant gen… culprit 1 String The annotation which was the worst p… geno(vcf): List of length 11: GT, AD, DP, GQ, MIN_DP, PGT, PID, PL, PS, RGQ, SB geno(header(vcf)): Number Type Description

GT 1 String Genotype

AD R Integer Allelic depths for the ref and alt alleles in the o… DP 1 Integer Approximate read depth (reads with MQ=255 or with b… GQ 1 Integer Genotype Quality

MIN_DP 1 Integer Minimum DP observed within the GVCF block

PGT 1 String Physical phasing haplotype information, describing … PID 1 String Physical phasing ID information, where each unique … PL G Integer Normalized, Phred-scaled likelihoods for genotypes … PS 1 Integer Phasing set (typically the position of the first va… RGQ 1 Integer Unconditional reference genotype confidence, encode… SB 4 Integer Per-sample component statistics which comprise the …

Header information

class: VCFHeader samples(1): UIC0003 meta(9): fileformat source … GATKCommandLine contig fixed(2): FILTER ALT info(44): AC AF … VarDP culprit geno(11): GT AD … RGQ SB

- Name/number of sample(s) = UIC0003

There is a **meta** region as well.

DataFrameList of length 9 names(9): fileformat source source.1 … source.5 GATKCommandLine contig DataFrame with 1 row and 1 column Value fileformat VCFv4.2 DataFrame with 1 row and 1 column Value source ApplyVQSR DataFrame with 1 row and 1 column Value source.1 GenomicsDBImport DataFrame with 1 row and 1 column Value source.2 GenotypeGVCFs DataFrame with 1 row and 1 column Value source.3 HaplotypeCaller DataFrame with 1 row and 1 column Value source.4 ReblockGVCF DataFrame with 1 row and 1 column Value source.5 VariantFiltration

There appear to be multiple sources from how variants were called.

DataFrame with 6 rows and 3 columns CommandLine Version Date ApplyVQSR “ApplyVQSR –recal-f..”4.1.8.0” “February 22, 2022 3.. GenomicsDBImport”GenomicsDBImport –.. “4.1.8.0” “February 22, 2022 2.. GenotypeGVCFs”GenotypeGVCFs –out.. “4.2.3.0” “February 22, 2022 2.. HaplotypeCaller”HaplotypeCaller –c.. “4.1.8.0” “May 4, 2021 1:08:40.. ReblockGVCF”ReblockGVCF –outpu.. “4.2.2.0” “September 20, 2021 .. VariantFiltration”VariantFiltration -.. “4.1.8.0” “February 22, 2022 2.. DataFrame with 3366 rows and 2 columns length assembly chr1 248956422 38 chr2 242193529 38 chr3 198295559 38 chr4 190214555 38 chr5 181538259 38 … … … HLA-DRB1*15:01:01:04 11056 38 HLA-DRB1*15:02:01 10313 38 HLA-DRB1*15:03:01:01 11567 38 HLA-DRB1*15:03:01:02 11569 38 HLA-DRB1*16:02:01 11005 38