EVA only accepts VCF files that conform to VCF version 4.1 and above.

VCF Specification

All VCF files must be valid, i.e. they must pass validation using a VCF validator, such as vcftools. In addition, all VCF files must include a reference tag (##reference). Any tag that contains white space, (e.g. spaces), must be enclosed in double quotes, e.g.

##EVA_PipelineDescription="This is a testing piepline!" OK
##EVA_PipelineDescription=This is a testing piepline! NOT OK

EVA Specific headers

An EVA submission automatically creates the required objects in ENA. A VCF file belongs to an Analysis, which in turn belongs to a Project. Currently each VCF file can only belong to one Analysis, and each Analysis can only contain one VCF file. The Analysis allows the VCF file to be linked to the Samples or SampleGroup (a group of samples).

Samples

If a Sample or SampleGroup already exists in ENA, then the VCF file can reference this object. This can be done in 3 ways:
1) Using the ENA Sample accession in the header of the VCF file.
2) Use the ##EVA_Sample information line to to link the sample ID in the VCF file to the Sample accession using the following specification:
##EVA_Sample=<ID=sample_id,ACCESSION=ena_accession>
3) Supply a sample mapping file(s) along with the VCF files, which links the VCF sample IDs to the Sample accession. The filename for the sample mapping file is given using the ##EVA_SampleMappingFile information line in the VCF file. Multiple files can be used, but each sample ID from the VCF must only occur once in all of the specified sample mapping files. The format to specify sample mapping files is:
##EVA_SampleMappingFile=<NAME=filename,MD5=md5sum>

Samples can have the following attributes:

Gender	REQUIRED
Phenotype	REQUIRED
Taxonomy ID	REQUIRED
Title	REQUIRED
Description	REQUIRED
Sample Type	OPTIONAL
Subject	OPTIONAL
Disease Site	OPTIONAL
Strain	OPTIONAL

The format for specifying the attributes is:
##EVA_Sample=<ID=sample_id,TAXID=taxonomy_id,TITLE="title",GENDER=gender,PHENOTYPE=phenotype,DESCRIPTION="description",SAMPLETYPE=sampletype,SUBJECT=subject,STRAIN=strain,SITE=diseasesite>

Samples can be linked to external resources, e.g. CORIELL. This is done using the ##EVA_SampleLink tag. e.g.
##EVA_SampleLink=<ID=HG00097,DB=CORIELL,DB_ID=NA18501,LABEL="Link to Coriell">
ID and DB are required (if the VCF ID is the same as the external resource ID). If the external resource ID is different to the VCF ID, this can be specified using the DB_ID sub-tag. LABEL is optional.
The ##EVA_Scope tag defines the study type. This tag is optional, and if used must correspond to the number of samples listed in the VCF file. If there is 1 sample, the ##EVA_Scope tag is 'single-isolate', otherwise it is 'multi-isolate'.

Analysis

Each VCF file must belong to an Analysis object which describes the analysis used to produce the VCF file. Currently and Analysis object can only contain 1 VCF file.
The following tags are used to describe the analysis:

##reference	REQUIRED
##EVA_AnalysisTitle	REQUIRED
##EVA_PipelineDescription	REQUIRED
##EVA_AnalysisCenter	OPTIONAL
##EVA_AnalysisDate	OPTIONAL
##EVA_CallingAlgorithm	OPTIONAL
##EVA_Platform	OPTIONAL
##EVA_ExperimentType	REQUIRED
##EVA_RunAccession	OPTIONAL
##EVA_AnalysisAccession	OPTIONAL

##reference
This is a 'highly recommended' tag according to VCFv4.1 specifications, and EVA require it for all VCF files. It can refer to either:
1) An INSDC assembly name or accession,e.g.
##reference=GRCh37
or
##reference=GCA_000001405.1
2) or a location of the reference file, e.g.
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
or
##reference=http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit
EVA also requires INSDC accessions or names for each of the chromosomes (or sequences) referenced in column 1 of the VCF. This should be of the form:
##contig=<ID=1,accession=CM000663.1,length=249,250,621,species="Homo sapiens",taxonomy=9606>
or
##contig=<ID=1,URL=http://www.ncbi.nlm.nih.gov/nuccore/224384768?report=fasta>
##EVA_AnalysisTitle
This is the title displayed for the accession Analysis object, e.g.
##EVA_AnalysisTitle="1000 Genomes Phase 1 Variation Calls Analysis"
##EVA_PipelineDescription

Description of analysis pipeline. Free text field, e.g.
##EVA_PipelineDescription="Final release set from derived from data from 7 different centers. See PUBMED:23128226"
##EVA_AnalysisCenter
The center that performed the analysis
##EVA_AnalysisCenter="European Bioinformatics Institute"
##EVA_AnalysisDate
The date the analysis was performed (in DD/MM/YYYY-hh:min:secs-offest, where offset and hh:min:secs are optional, hr:min:secs defaults to 12:00:00, and offset indicates the hour offset from UTC defaulting to 0), e.g. for an analysis performed at 9.30 am, 6 hours before UTC on 30th May 2011:
##EVA_AnalysisDate="30/05/2011-09:30-06"
##EVA_CallingAlgorithm
Details of the calling algorithm, version, and parameters, e.g.
##EVA_CallingAlgorithm="GATKv2.3, java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R reference/human_g1k_v37.fasta -I HG00096.bam -I HG00097.bam --dbsnp dbSNP.vcf -stand_call_conf -stand_emit_conf 10.0 -o output.raw.snps.indels.vcf"
##EVA_Platform
The platform used, e.g.
##EVA_Platform="Illumina Genome Analyser IIe"
##EVA_ExperimentType
Tag to indicate the type of experiment, choose from:
Whole genome, Exome, Array
e.g.
##EVA_ExperimentType="Whole Genome"
##EVA_RunAccession
Allows the VCF and associated Analysis object to be linked to an existing run accession, e.g.
##EVA_RunAccession=DRR000003

Project

A Project object is required to contain the Sample and Analysis objects that pertain to the submitted VCF files. The tags for Project are:
##EVA_ProjectName--REQUIRED
##EVA_ProjectDescription--REQUIRED
##EVA_Material--REQUIRED
##EVA_Selection--REQUIRED
##EVA_Publication--OPTIONAL
##EVA_RelatedProject--OPTIONAL
##EVA_ProjectLink--OPTIONAL
##EVA_ProjectName
This is a unique name for the Project,e.g.
##EVA_ProjectName="1000 Genomes Phase 1 Variation Calls"
##EVA_ProjectDescription
This is a short description describing the project,e.g.
##EVA_ProjectDescription="Final release set of the Phase 1 1000 genomes variation data."
##EVA_Material

The material used, one of:
DNA, genomic RNA, transcribed RNA, unknown, other. e.g.
##EVA_Material=DNA
##EVA_Selection
One of the following:
genome, partial genome, exome, epigenomics, PCR, clone ends, random, CDNA, tag, unknown, other. e.g.
##EVA_Selection=genome
##EVA_Publication
Any associated publications with the Project, e.g.
##EVA_Publication=<DB=PUBMED,ID=23128226,LABEL="Main 1000 Genomes Publication">
DB and ID are required, LABEL is optional.
##EVA_RelatedProject
Projects can have parent, peer, or child projects. This tag allows one to specify associated projects, e.g.
##EVA_RelatedProject=<RELATIVE=PARENT,ID=28889>
##EVA_RelatedProject=<RELATIVE=PEER,ID=28890>
##EVA_RelatedProject=<RELATIVE=CHILD,ID=28891>
##EVA_RelatedProject=<RELATIVE=CHILD,ID=28892>
The RELATIVE and ID sub-tags are both required.
##EVA_ProjectLink
This tag allows the project to be linked to an external resource, e.g.
##EVA_ProjectLink=<URL=Error! Hyperlink reference not valid."1000 Genomes website">
##EVA_ProjectLink=<DB=DGVa,ID=estd199>
DB and ID can be used for INSDC databases.
If DB is present, ID is required. If DB is absent, URL is required and ID must not be included. LABEL is OPTIONAL.

Below is a valid set of tags for a fictional submission:
##fileformat=VCFv4.1
##EVA_ProjectName="1000 Genomes Phase 1 Variation Calls"
##EVA_ProjectTitle="Test 1"
##EVA_ProjectDescription="Test 2"
##EVA_Publication=<DB=PUBMED,ID=23128226,LABEL="Main 1000 Genomes Publication">
##EVA_Publication=<DB=ePUB,ID=23128226,LABEL="Main 1000 Genomes Publication">
##EVA_Scope=multi-isolate
##EVA_Material=DNA
##EVA_Selection=other
##EVA_TaxID=9606
##EVA_RelatedProject=<RELATIVE=PARENT,ID=28889>
##EVA_RelatedProject=<RELATIVE=PEER,ID=28890>
##EVA_RelatedProject=<RELATIVE=CHILD,ID=28891>
##EVA_RelatedProject=<RELATIVE=CHILD,ID=28892>
##EVA_ProjectLink=<URL=http://www.1000genomes.org/phase1-analysis-results-directory,LABEL="FTP directory of submitted VCF files">

##EVA_ProjectLink=<URL=Error! Hyperlink reference not valid."1000 Genomes website">
##EVA_AnalysisTitle="1000 Genomes Phase 1 Variation Calls VCF Parse TEST"
##EVA_FileDescription="Final release set from derived from data from 7 different centers. See PUBMED:23128226"
##EVA_SampleMappingFile=<Name=test_file1.txt,MD5=12345678912345678912345678900> ##EVA_ChecksumFile=ch1_test_md5.txt
##EVA_PipelineDescription="This is a testing piepline!"
##EVA_CallingAlgorithm="GATKv2.3, java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R reference/human_g1k_v37.fasta -I HG00096.bam -I HG00097.bam --dbsnp dbSNP.vcf -stand_call_conf [50.0] -stand_emit_conf 10.0 -o output.raw.snps.indels.vcf" ##EVA_Platform="Illumina GA11"
##EVA_ExperimentType="Is this a CV"
##EVA_RunAccession=<ID=SRA00001,NAME=ARunName>
##EVA_AnalysisAccession=<ID=ENAA00001,NAME=AAnalysisName>
##EVA_Sample=<ID=HG00096,ACCESSION=ERSAM1>
##EVA_Sample=<ID=HG00097,GENDER=Male,PHENOTYPE=Ill,SAMPLETYPE=DNA,SUBJECT=Someone,DISEASE_SITE=Pancreas,TAXID=9606,TITLE="1000 Genomes Sample Number 97",DESCRIPTION="This is a description of HG00097">
##EVA_SampleLink=<ID=HG00097,DB=CORIELL,DB_ID=HG00097,LABEL="This is a label">