dbSNP VCF Submission Format Guidelines

Contact: snp-admin@ncbi.nlm.nih.gov

Last update: June 19, 2015

Introduction

dbSNP Submissions

dbSNP is a public database of short genetic variations. The data can be from any species, and from any part of a genome. dbSNP has been designed to include a broad collection of simple genetic variations such as single-base nucleotide substitutions, small-scale multi-base deletions or insertions, retrotransposable element insertions, and microsatellite repeats. Submissions can include genotype and allele frequency data if those data are available. dbSNP accepts submissions for all classes of simple molecular variation, including common variations as well as rare variations of germline or somatic origin that are clinically significant. Large-scale insertion/deletion, inversion and translocation data that are over 50bp long should be submitted to dbVar, the NCBI database of genomic structural variation.

The Variant Call Format (VCF)

The Variant Call Format, or VCF, was developed for the 1000 Genomes Project as a standardized format for storing large quantities of sequence variation data (SNPs, indels, larger structural variants, etc.) and any accompanying genotype data and annotation. A VCF file contains a header section and a data table section. Since the metadata lines in the header section can be altered to fit the requirements of the data to be submitted, you can use VCF to submit many different kinds of common variations (as well as their associated genotypes and annotation) that are contained within one reference sequence. VCF files are compressed (using bgzip), and easily accessed. See Danecek, et. al. for a concise overview of VCF, and the official 1000 Genomes site for a detailed description of the VCF format. Submissions to dbSNP currently use VCF format version 4.1.

NOTE: Please do not use the VCF format if you have human mutations or variations with clinical significance or phenotype. They should be submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) or contact dbSNP (snp-admin@ncbi.nlm.nih.gov) if you have any questions.

When should I use the VCF format for dbSNP Submissions?

Use dbSNP's Variant Call Format (VCF) to submit large or small numbers of short genetic variations that have asserted positions on genome or reference sequences[a]. Large scale submitters especially will find dbSNP's VCF submission format a very useful submission tool since it allows for the submission of numerous variations generated by high-throughput sequencing (HTS) projects over multiple populations, as well as a wide variety of associated data. The VCF file for dbSNP submissions, as opposed to the standard VCF format as defined by the 1000 Genomes project, includes additional fields and attributes that describe dbSNP-specific submission and variation properties, and may include tags that are different than those used in standard VCF.

[a]dbSNP prefers that all variant asserted positions submitted using the VCF format are submitted either on a sequence accession that is part of an assembly housed in the NCBI Assembly Resource or as an asserted location on an INSDC sequence housed in DDBJ, ENA, and GenBank.

Submission Overview

1. Check to see if your lab already has a handle assignment from NCBI. If it does not, request a handle using the dbSNP online handle request form.

2. Prepare your submission:
    a. The VCF file format required for dbSNP submissions is based on the 1000 Genomes Project VCF format guidelines with the addition of dbSNP specific fields. These additional fields describe dbSNP submission and variation properties.

    b. Create required metadata (meta) files for the publication, method, population, and assay information associated with the submission.

c. Create a VCF Submission file for your data. Include:

- a properly formatted dbSNP VCF file header
- a data table that contains the required INFO tag for the variants you are submitting
- optional INFO tags that will describe your data more fully.

d. We suggest you compress the VCF file using gzip (http://www.gzip.org/) and send the compressed file by email or FTP:

- File size less than 10MB
  1. Email your submission at attachments to snp-sub@ncbi.nlm.nih.gov.
- File size more than 10MB:
  1. Request a FTP account from NCBI for uploading by sending your handle confirmation information to snp-sub@ncbi.nlm.nih.gov.
  2. Upload your submission files to your assigned FTP account and notify snp-sub@ncbi.nlm.nih.gov when the upload is complete.

See the appendix of this document for an example of a VCF formatted dbSNP submission.

Required Metadata Files

In addition to VCF formatted variation files, dbSNP also requires VCF submissions to include separate Meta file(s).

- The required Meta files are: Publication, Method, Population, and Assay.

- You can submit these Meta files separately or combine them into a single text file for submission.

- Specifications for each Meta file is available in the "How to Submit" documentation for dbSNP. Links to the specific sections of the document that provide the required specifications are provided above.

Below is an example of a Meta file that combines all four Meta file types into a single file:

```
TYPE:   CONT
HANDLE: MYSEQ_SNP
NAME:   Jim Johnson
FAX:    111 111 1111
TEL:    222 222 2222
EMAIL:  jj@nih.gov
LAB:    NCBI
```

```
INST:   NCBI, NIH
ADDR:   9600 Rockville Pike, Bethesda, MD 20892
||
TYPE: PUB
HANDLE: MYSEQ_SNP
PMID: 123456
TITLE: Variation discovery in European and African Populations
AUTHORS: Jim Johnson
YEAR: 2014
STATUS: 1
||
TYPE:    METHOD
HANDLE:    MYSEQ_SNP
ID:    AgilentWholeExome
METHOD_CLASS:    Sequence
TEMPLATE_TYPE:    DIPLOID
METHOD:
Solution hybridization exome capture was carried out using the Human All Exon
System. The captured regions totaled approximately 38 or 50 Mb depending on
the kit used. Flow cell preparation and paired end read sequencing were
carried out on GAIIx and HiSeq2000 sequencers (Illumina Inc, San Diego CA).
Sequence reads were aligned with the diagCM aligner and genotypes were called
with bam2mpg (Teer et al, Systematic comparison of three genomic enrichment
methods for massively parallel DNA sequencing, Genome Res. 2010
Oct;20(10):1420-31).
||
TYPE:    POPULATION
HANDLE:    MYSEQ_SNP
ID:    EUROPEAN
POPULATION:    This population includes 712 participants of European descent.
||
TYPE:    POPULATION
HANDLE:    MYSEQ_SNP
ID:    AFRICAN
POPULATION:    This population includes 600 participants of African descent.
||
TYPE:    SNPASSAY
HANDLE:    MYSEQ_SNP
BATCH:    Exome_SNP_Discovery
MOLTYPE:    Genomic
METHOD:    AgilentWholeExome
ORGANISM:    Homo sapiens
||
TYPE:    SNPPOPUSE
HANDLE:    MYSEQ_SNP
BATCH:    Exome_SNP_MAF
METHOD:    AgilentWholeExome
||
```

dbSNP VCF Submission Format

VCF Submission File Header

Required VCF Header Metadata

The VCF file header for a dbSNP submission should start with the following metadata:

```
##fileformat=    {The current VCF version ID: i.e. VCF v4.1}
##fileDate=      {The date that the file was generated or the date when the
file was
                  updated. Use YYYYMMDD
                  format:i.e.20120201}
##handle=        {Your registered dbSNP submission handle}
##batch=         {A unique local batch ID. Use the same value placed in
                  The BATCH field of the Meta file SNPASSAY section; dbSNP
                  uses the local batch ID to associate the VCF submission
                  with the ASSAY, PUBLICATION, and METHOD meta data}
##bioproject_id= {A registered BioProject ID if available}
##biosample_id=  {A comma separated list of registered BioSample IDs
                  (https://www.ncbi.nlm.nih.gov/biosample/). We encourage
                  submitter to register their samples with BioSample and
                  provide detail descriptions such as traits and phenotype.
                  In this example of two Biosample records, the ID numbers
                  are 423 and 1595}
##reference=     {The RefSeq Assembly accession.version on which the
                  variation position is based: i.e. GCF_000001405.12. You
                  can find this ID by accessing NCBI's Genome Assembly
                  Resource (https://www.ncbi.nlm.nih.gov/assembly/) and search for the
                  record of the specific
                  assembly. You can use the organism or assembly name(e.g.
            GRCh37) as your search term: the assembly record
                  for GRCh37 shows the RefSeq ID is GCF_000001405.12.  Only
                  the accession.version for a fully assembled genome can
                  be reported here. For unassembled and unplaced contigs,
                  leave this tag blank and use the reporting method for
                  INSDC sequence coordinates as shown in the example below
                  (VCF Data Table Examples B) for the CHROM column.}
```

Example of dbSNP Metadata in a VCF formatted file:

```
##fileformat=VCFv4.1
##fileDate=20120215
##handle=MYSEQ_SNP
##batch=Exome_SNP_Discovery
##bioproject_id=60153
##biosample_id=423, 1595
##reference=GCF_000001405.12
```

INFO Tag Descriptions

The VCF header continues with tag/value descriptions for required and optional dbSNP INFO tags. These descriptions should be placed in the header following the required metadata.

The INFO tag/value descriptions you provide in the VCF header will serve to define the data you place in the INFO column of the data table. These descriptions are an important part of the VCF header as they will allow users viewing your data in VCF format to identify a tag you placed in the INFO column and see definitions for values of that tag. The data you present in the INFO column of the data table will be meaningless to some users without the inclusion of the tag/value descriptions in the VCF header for those data.

Descriptions for Required INFO Tag

Currently, the only required INFO tag for a dbSNP submission is the Variation Type (VRT) tag. Place the VRT tag description in the VCF file header after the required metadata. The VRT tag is required for each variant submitted in VCF format. **Failure to include this required INFO tag will result in the delay of your submission**.

See the dbSNP INFO Tag Descriptions and Examples section of this document for example tag descriptions you can cut and paste into the VCF file header for both the required INFO tags and the optional INFO tags.

Descriptions for Optional INFO Tags

Place descriptions for the optional INFO tags in the VCF file header after the required metadata. These descriptions identify and define the optional INFO tags you have elected to use in the data table portion of the file.

See the dbSNP INFO Tag Descriptions and Examples section of this document for example tag descriptions you can cut and paste into the VCF file header for both the required INFO tags and the optional INFO tags.

Submission Data Table

Create a tab-delimited table to house your variations and variation data for your submission. The table header should include these six fixed, mandatory columns (in order):

```
#CHROM    POS    ID    REF    ALT    INFO
```

The above columns represent six fixed fields that must be filled out for each submitted variant. If you do not have data for a particular field, use a dot (".") to represent the missing value.

VCF Data Table Examples

A) Reporting positions using chromosome coordinates (please provide the 'reference' tag in the header if the assembly and version is known).

#CHROM	*POS*	*ID*	*REF*	*ALT*	*QUAL*	*FILTER*	*INFO*
23	**135498962**	**NG_021219.1:g.120841A>G**	**A**	**G**	**29**	**PASS**	**VRT=1**
23	**135499109**	**NG_021219.1:g.120988G>A**	**G**	**A**	**40**	**PASS**	**VRT=1**
23	**135499270**	**NG_021219.1:g.121149C>T**	**C**	**T**	**51**	**PASS**	**VRT=1**
23	**135499419**	**NG_021219.1:g.121298G>C**	**G**	**C**	**68**	**PASS**	**VRT=1**

B) Reporting positions using INSDC sequence (i.e. GenBank sequence with accession) coordinates if assembly is not known

#CHROM	*POS*	*ID*	*REF*	*ALT*	*QUAL*	*FILTER*	*INFO*
NG_021219.1	***140860***	***SNV1***	***T***	***C***	***29***	***PASS***	***VRT=1***
NG_021219.1	***140879***	***SNV2***	***A***	***G***	***40***	***PASS***	***VRT=1***
NG_021219.1	***140921***	***SNV3***	***T***	***C***	***51***	***PASS***	***VRT=1***
NG_021219.1	***140939***	***SNV4***	***C***	***T***	***68***	***PASS***	***VRT=1***

C) Reporting positions using non-INSDC sequence coordinates or sequence yet to be submitted to GenBank

Users that submit variations with an asserted location on a sequence that is being submitted simultaneously to GenBank or has a pending GenBank submission accession assignment can temporarily report the variant's asserted location based on the local (user-defined) sequence ID with the following additional requirements:

- Provide the 5' and 3' flanking sequence surrounding the variation. A minimum of 25bp is required for each 5' and 3' flanking sequence provided in the INFO tag FLANK-5 and FLANK-3, respectively (see example below)

- Upon receiving GenBank accession numbers, the submitter can: 1) update the VCF and replace the local sequence ID with the corresponding

GenBank accession or 2) provide a tab-delimited file containing the GenBank accession for each the local sequence ID per row.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
My_Seq_ID_1	140860	SNV1	T	C	29	PASS	VRT=1;FLANK-5=TGCAACAATCTGGGCTATGAGATCA;FLANK-3=TAAAGTCAGAGCCAAAAGAAGCAGC
My_Seq_ID_1	140979	SNV2	A	G	40	PASS	VRT=1;FLANK-5=TTAACTAGCTTGGTTGCTGAACACC;FLANK-3=GGTTAGGCTCTCAAATTACCCTCTG
My_Seq_ID_1	141921	SNV3	T	C	51	PASS	VRT=1;FLANK-5=TGCAACAATCTGGGCTATGAGATCA;FLANK-3=AGGCTGGTGAGCATTCTGGGCTAAA
My_Seq_ID_1	149939	SNV4	C	T	68	PASS	VRT=1;FLANK-5=GACACCATGGTGCATCTGACTCCTG;FLANK-3=GGAGAAGTCTGCCGTTACTGCCCTG

VCF Data Table Field Values

#CHROM

This field contains the chromosome identifier from the reference genome where the variant is located or an angle-bracketed ID String ("<ID>") pointing to a contig in the assembly file. (cf. the ##assembly line in the header). Entries for a specific CHROM should form a contiguous block within the VCF file. Alternatively, the sequence accession and version can be used for this field if the variation position is based on a non-chromosomal sequence (see example B above). Do not use the colon symbol (:) in a chromosome name.

#POS

This field contains the reference position of the variant, which is the 1st base of the variation event. Positions are sorted numerically within each reference sequence chromosome (CHROM) in increasing order. You are permitted to have multiple records of different variation type (VRT) at the same POS. Telomeres are indicated by using positions 0 or N+1, where N is the length of the corresponding chromosome or contig.

Note: For short, simple insertions and deletions in which the REF or one of the ALT alleles would otherwise be null/empty, the POS field must contain the coordinates of the base preceding the indel event. See the Submission Data Table Special Case Examples section of this document for instruction on reporting insertion/deletion POS values.

Large indels and structural variants must be submitted to dbVAR

ID

This field contains the unique local ID (LID) of the variant, and is a required value (cannot be NULL).The LID provided here combined with the handle must be unique for a particular submitter. You can use an HGVS expression (http://www.hgvs.org/mutnomen/recs.html) for the variant ID if you do not have a unique identifier of your own.

REF

This field contains the reference allele of the variant. The bases representing the reference allele can be any of the following: A, C, G, T (case insensitive).
Note: In order for the variant to be included in dbSNP, the maximum length for the REF allele is 51bp.

Note: For short, simple insertions and deletions in which the REF or one of the ALT alleles would otherwise be null/empty, the REF and ALT Strings must include the base preceding the indel event. See the Submission Data Table Special Case Examples section of this document for instruction on reporting indel reference (REF) alleles.

ALT

This field contains a comma separated list of alternate, non-reference alleles that you have called in at least one sample. You can use A, C, G, or T (case insensitive) or you can use an angle-bracketed ID String (”<ID>”).. **Note: In order for the variant to be included in dbSNP, the maximum length of each ALT allele is 51bp.**

Note: For short, simple insertions and deletions in which the REF or one of the ALT alleles would otherwise be null/empty, the REF and ALT Strings must include the base preceding the indel event. See the Submission Data Table Special Case Examples section of this document for instruction on reporting indel alternate (ALT) alleles.

QUAL

This field contains the quality score for the assertion if available.

FILTER

This field contains the filter status if available.

INFO

This field contains additional information for the reported variation. INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: <key>=<data>[,data] See the INFO Tag Descriptions and Examples section of this document for examples of the required and optional INFO Tags that dbSNP supports.

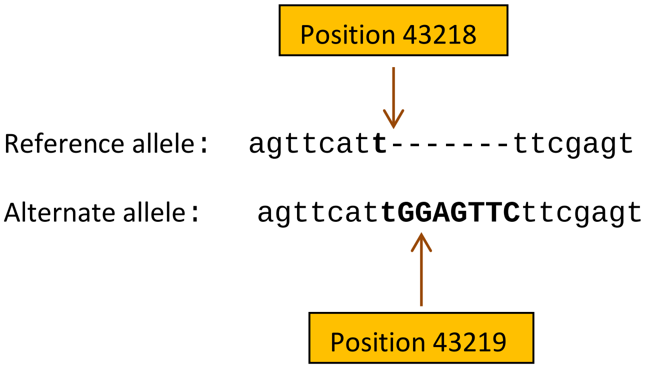
Submission Data Table Special Case Examples: Reporting POS, REF and ALT for insertion/deletion variants

For simple insertions and deletions where either the REF or one of the ALT alleles would otherwise be null/empty, include the base preceding the variation event (a "padding base") in the REF and ALT allele Strings, and report the coordinates of this "padding base" in POS.

The "padding base" is not required for complex substitutions or other events where all alleles have at least one base represented in their Strings.

Insertion Example

Sequence: TCAGTCTCACCATGAAAGTTCATT[-/GGAGTTC]TTCGAGTAAATGGTTCCCAGCGGG



Position 43218

Reference allele: agttcat**t**-------ttcgagt

Alternate allele: agttcat**tGGAGTTC**ttcgagt

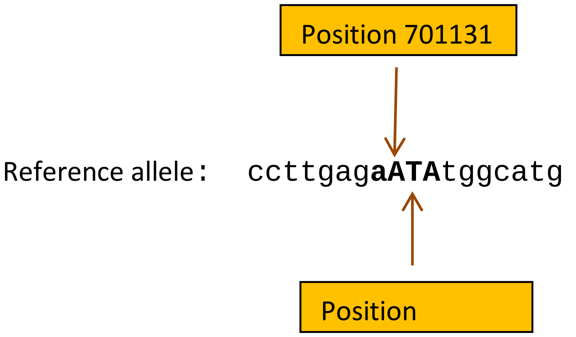
Position 43219

If the coordinates of first base of the insertion event ("G" at position 43219) in the above sequence were used as the reference position (POS) of this event, the REF field would have no value since the inserted bases are only present in the ALT allele. In such a case, report the coordinates of the base that precedes the insertion event— the "t" at position 43218 — for POS and include this "padding base" in the REF and ALT Strings:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
10	43218	NC_0000010.10:g.43218_43219insGGAGTTC	T	TGGAGTTC	.	.	VRT=2;ANC=T;NIO=12;SSR=0;SAO=0;SCS=0

Deletion Example

Sequence: AGAGATTCACAGCCTCACCTTGAGA[ATA/-]TGGCATGGAGAATATTTTGGATAAT



Reference allele: ccttgag**aATA**tggcatg

Alternate allele: ccttgag**a**---tggcatg

If the coordinates of first base of this deletion event (“A” at position 701132) in the above sequence were used as the reference position (POS) of this variant, the ALT field would have no value since the deleted bases are only present in the reference (REF) allele. In such a case, report the coordinates of the base that precedes the deletion event— the “a” at position 701131 — for POS and include this “padding base” in the REF and ALT Strings:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
15	701131	NC_000015.9:g.701132_701134delATA	AATA	A	.	.	VRT=2;ANC=A;NIO=5;SSR=0;SAO=0;SCS=0

INFO Tag Descriptions and Examples

Required dbSNP VCF INFO Tag

Place the required tag in the INFO column of the data table and place the corresponding tag description in the file header.

Variation Type (VRT) INFO Tag

The required "VRT" INFO tag allows you to define the kind of variation you are submitting to dbSNP. We use this information to verify position and that the reported alleles are consistent with reported variation type.

Note: Only one variation type (VRT) can be reported per row. For instance, if you have a deletion variation and a SNV at the same location, they should be reported in two separate rows with the corresponding VRT value.

Failure to include this required INFO tag will result in the delay of your submission.

VRT Tag/Value Description

```
##INFO=<ID=VRT,Number=1,Type=Integer,Description="Variation type,1 - SNV:
single nucleotide variation,2 - DIV: deletion/insertion variation,3 -
HETEROZYGOUS: variable, but undefined at nucleotide level,4 - STR: short
tandem repeat (microsatellite) variation, 5 - NAMED: insertion/deletion
variation of named repetitive element,6 - NO VARIATON: sequence scanned for
variation, but none observed,7 - MIXED: cluster contains submissions from 2
or more allelic classes (not used),8 - MNV: multiple nucleotide variation
with alleles of common length greater than 1,9 - Exception">
```

VRT Data Format Example

#CHROM	*POS*	*ID*	*REF*	*ALT*	*QUAL*	*FILTER*	*INFO*
1	140860	NC_000001.10:g140860T>C	T	C	.	.	***VRT=1***

Optional dbSNP VCF INFO Tags

The following INFO tags are optional and need only be used if they describe available data. If you want to include any of the following INFO tags with your submitted data, place the tag in the INFO column of the data table and place the corresponding tag description in the file header. Optional VCF INFO tags for dbSNP submissions include:

Alternate Designations

Ancestral Allele

Free Text for Comment

LinkOut

Mutation

Number of Independent Observations

OMIM/OMIA Record

Population Frequency

Population ID

PubMed ID

Variant Allele Origin

Variant Suspect Reason

Alternate Designations (AD) or Names

The optional “AD” INFO tag allows you to provide dbSNP with a (comma separated) set of alternative names or common names used to describe the same submitted variant.

AD Tag/Value Description

```
##INFO=<ID=AD,Number=1,Type=String,Description=”Alternate designations; a set
of (comma separated)alternative names used to describe the same variant”>
```

AD Tag/Value Example

#CHROM	*POS*	*ID*	*REF*	*ALT*	*QUAL*	*FILTER*	*INFO*
10	134017295	NC_000010.10:g.134017295A>G	A	G	.	.	VRT=1;ANC=T;NIO=12;AD=SNP-12313,chr10:134017295A>G;

Ancestral Allele (ANC)

The optional “ANC” INFO tag allows you to provide dbSNP with the ancestral allele (if you know it) for a variant.

ANC Tag/Value Description

```
##INFO=<ID=ANC,Number=1,Type=String,Description=”Provide Ancestral Allele if
known”>
```

ANC Tag/Value Example

#CHROM	***POS***	***ID***	***REF***	***ALT***	***QUAL***	***FILTER***	***INFO***
8	19863	NC_000008.10:g.19863G>C	G	C	.	.	VRT=1;ANC=T;

Free Text for Comment (CMT)

The optional “CMT” INFO tag allows you to provide dbSNP with text about any additional important information that cannot be described (e.g. phenotypic information) using the other available INFO tags

CMT Tag/Value Description

```
##INFO=<ID=CMT=1,Type=String,Description=”Comment”>
```

CMT Data Format Example

#CHROM	***POS***	***ID***	***REF***	***ALT***	***QUAL***	***FILTER***	***INFO***
8	19863	NC_000008.10:g.19863G>C	G	C	.	.	VRT=1;CMT=”A variant ident SLC10A1 gene with possible correlation to disease susceptibilities(PMID: 124

LinkOut (LKO)

The optional “LKO” INFO tag allows you to point to this variant on your organization’s web site or to other relevant online information about your submission.

LKO Tag/Value Description

```
##INFO=<ID=LKO,Number=1,Type=String,Description=”A link out URL for this
variant on the submitting organization's website”>
```

LKO Data Format Example:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO

8	19863	NC_000008.10:g.19863G>C	G	C	.	.	VRT=1;ANC=T;
							LKO=http://variantgps.nci
							08;

Number of Independent Observations (NIO)

The optional “NIO” INFO tag allows you to provide dbSNP with the number of times you observed this variant occur independently in your experimental analysis.

NIO Tag/Value Description

```
##INFO=<ID=NIO,Number=1,Type=Integer,Description=”Number of Independent
Observations;the number of times the submitter observed this variant
occurring independently”>
```

NIO Tag/Value Example

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
8	19863	NC_000008.10:g.19863G>C	G	C	.	.	VRT=1;ANC=T;NIO=12;

OMIM and OMIA (OMIM/OMIA) Records

The optional “OMIM” and “OMIA” INFO tags allow you to provide dbSNP with any available OMIM or OMIA record and variant ID (if available) associated with a variant.

OMIM and OMIA Tag/Value Descriptions

OMIM:

```
##INFO=<ID=OMIM,Number=1,Type=String,Description=”Provide OMIM
(http://omim.org)record and variant ID if available i.e. 300746.0001”>
```

OMIA:

```
##INFO=<ID=OMIA,Number=1,Type=String,Description=”Provide OMIA
(https://www.ncbi.nlm.nih.gov/omia)record and variant ID if available i.e.
000011-9615”>
```

OMIM and OMIA Data Format Example

#CHROM	*POS*	*ID*	*REF*	*ALT*	*QUAL*	*FILTER*	*INFO*
16	919982	NC_000016.9:g.919982G>C	G	C	.	.	VRT=1;ANC=T;OMIM=300746.0001;

Population IDs (for Allele Frequency, Genotype Frequency, or Observed Heterozygosity data submissions)

If you intend to report allele frequency, genotype frequency, or observed heterozygosity in your VCF formatted dbSNP submission, place the population ID for each assayed population in the VCF header after the INFO Tag/Value descriptions, and before your data table. The POP IDs you will provide in the VCF header are the same ones you placed in the ID field of the Meta File.

Population_ID Tag/Value Description

```
##population_id=<A unique local population ID e.g. “HapMap”, “Case”,
“Control”, “Healthy Blood Donors”>  Use the same value placed in the ID field
of the Meta file POPULATION section where the population details are
described.
```

Population_ID Example

```
##INFO=<ID=GEN_FRQ,Number=1,Type=string,Description=”Report population,
sample size (number of distinct chromosomes assayed), and frequency for each
genotype
##population_id=EUROPEAN
##population_id=AFRICAN
```

Genotype Format Example

The format for reporting genotypes is found in the genotype submission example provided by 1000 Genome Project in their description of VCF version 4.1.

Allele Frequency Format Examples

The format for reporting allele frequency follows the convention for reporting for genotype.

- Add a reporting FORMAT column to specify the data type and order. Suggested data types are listed below.

```
##FORMAT=<ID=NA,Number=1,Type=Integer,Description="Number of alleles for
the population."
##FORMAT=<ID=NS,Number=1,Type=Integer,Description="Number of samples for
the population."
##FORMAT=<ID=FRQ,Number=.,Type=Float,Description="Frequency of each
alternate allele."
##FORMAT=<ID=AC,Number=.,Type=Integer,Description="Allele count for each
alternate allele."
```

- Add additional column for each population
- Report under the population column the total allele count (NA) or population samples (NS) follow by the allele frequency (FRQ) or allele count (AC) separated by a colon ':'
- Below are examples for reporting allele frequencies for a novel variant (row 1) and for a known variant with a dbSNP RS number (row 2) reported in the ID column.

#CHROM	*POS*	*ID*	*REF*	*ALT*	*QUAL*	*FILTER*	*INFO*	*FORMAT*	**EUROPEAN**	**AFRICAN**
X	140879	.	A	G	.	.	VRT=1; LID=SNV:chrX:140879	NA:FRQ	1424:0.001	1200:0.05
8	19962213	rs328	C	G	.	.	.	NA:FRQ	178:0.101	224:0.045

PubMed ID (PMID) INFO Tag

The optional "PMID" INFO tag allows you to provide dbSNP with the PubMed ID (if available) for an original publication associated with a variant. If multiple PubMed IDs (PMID) are available for a single variant, report them using a comma separated list (see example below). Report PMIDs for multiple variants as a batch in the ASSAY and PUBLICATION meta files.

PMID Tag/Value Description

```
##INFO=<ID=PMID,Number=.,Type= Integer,Description=”PubMed ID linked to
variation if available”>
```

PMID Data Format Example

#CHROM	*POS*	*ID*	*REF*	*ALT*	*QUAL*	*FILTER*	*INFO*
16	919982	NC_000016.9:g.919982G>C	G	C	.	.	VRT=1;ANC=T;PMID=21840003;

Variant Allele Origin (SAO) INFO Tag

The optional “SAO” or “Variant Allele Origin” INFO tag allows you to provide dbSNP with the source of the sample from which the variant was derived.

Note: Although the name we use to refer to Allele Origin has changed from “SNP Allele Origin” (SAO) to “Variant Allele Origin” to emphasize that the dbSNP database contains both rare and polymorphic variants, the database itself still uses the acronym “SAO”.

SAO Tag/Value Description

```
##INFO=<ID=SAO,Number=.,Type=Integer,Description=”Variant Allele Origin: 0 -
unspecified, 1 - Germline, 2 - Somatic, 3 – Both”>
```

SAO Data Format Example

#CHROM	*POS*	*ID*	*REF*	*ALT*	*QUAL*	*FILTER*	*INFO*
16	919982	NC_000016.9:g.919982G>C	G	C	.	.	VRT=1;ANC=T;SAO=1;

Note: If you are providing more than one allele origin value, place the allele origin values in a comma separated list in the order that they appear in the submission. List the value for the for the reference allele first, followed by the allele origin value for the 1st alternate allele, 2nd alternate allele, etc.:

Variant Suspect Reason (SSR) INFO Tag

The optional “SSR” or “SNP Suspect Reason” INFO tag allows you to provide dbSNP with the reason you suspect that a variant is a false positive. Evidence for false positives can include information indicating the presence of a paralogous sequence in the genome (Musumeci et al. 2010) (Sudmant et al. 2010), or evidence of sequencing error or computation artifacts.

Note: Although the name we use to refer to the Suspect Reason code has changed from “SNP Suspect Reason” (SSR) to “Variant Suspect Reason” to emphasize that the dbSNP database contains both rare and polymorphic variants, the database itself still uses the acronym “SSR”.

SSR Tag Description

```
##INFO=<ID=SSR,Number=.,Type=Integer,Description=”Variant Suspect Reason
Code, 0 - unspecified, 1 - Paralog, 2 - byEST, 3 - Para_EST, 4 - oldAlign, 5
– other”>
```

SSR Data Format Example

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
16	919982	NC_000016.9:g.919982G>C	G	C	.	.	VRT=1;ANC=T;SSR=1;

Appendix: Example of a VCF Formatted dbSNP Submission

```
##fileformat=VCFv4.1
##fileDate=20120215
##handle=MYSEQ_SNP
##batch=Exome_SNP_Discovery
##reference=GCF_000001405.12
##INFO=<ID=VRT,Number=1,Type=Integer,Description=”Variation type, 1 - SNV:
single nucleotide variation, 2 - DIV: deletion/insertion variation, 3 -
HETEROZYGOUS: variable, but undefined at nucleotide level, 4 - STR: short
tandem repeat (microsatellite) variation, 5 - NAMED: insertion/deletion
variation of named repetitive element, 6 - NO VARIATON: sequence scanned for
variation, but none observed, 7 - MIXED: cluster contains submissions from 2
or more allelic classes, 8 - MNV: multiple nucleotide variation with alleles
of common length greater than 1, 9 – Exception”>
##INFO=<ID=LID, Number=1,Type=string, Description=”Unique local variation ID
or name for display.  The LID provided here combined with the handle must be
unique for a particular submitter.  An HGVS expression
(http://www.hgvs.org/mutnomen/recs.html) can be used here”>
##FORMAT=<ID=NA,Number=1,Type=Integer,Description="Number of alleles for the
population."
##FORMAT=<ID=NS,Number=1,Type=Integer,Description="Number of samples for the
population."
##FORMAT=<ID=FRQ,Number=.,Type=Float,Description="Frequency of each alternate
allele."
```

```
##FORMAT=<ID=AC,Number=.,Type=Integer,Description="Allele count for each alternate allele."
##population_id=EUROPEAN
##population_id=AFRICAN
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	EUROPEAN	AFRICAN
X	140860	.	T	C	.	.	VRT=1;LID=SNV:chrX:140860;	NA:FRQ	1424:0.056	.
X	140879	.	A	G	.	.	VRT=1;LID=SNV:chrX:140879;	NA:FRQ	1424:0.001	1200:0.05
X	140921	.	T	C	.	.	VRT=1;LID=SNV:chrX:140921;	NA:FRQ	1424:0.003	1200:0.002
X	140939	.	C	T	.	.	VRT=1;LID=SNV:chrX:140939;	NA:FRQ	1424:0.01	.