dbVar VCF Submission Format Guidelines

Contact: dbvar@ncbi.nlm.nih.gov

Last updated: September 23, 2014

Introduction

dbVar Submissions

dbVar is a public database of large structural variation. The data in dbVar can be from any species, and from any part of a genome. dbVar has been designed to include a broad collection of structural variations such as copy number variants (CNVs), insertions, duplications, inversions, translocations, and complex variants. Submissions can include genotype and allele frequency data if those data are available. dbVar accepts submissions for all classes of structural variation, including common variations as well as rare variations of germline or somatic origin that are clinically significant. In most cases, variants shorter than 50bp should be submitted to dbSNP, the NCBI database of single nucleotide variation.

The Variant Call Format (VCF)

The Variant Call Format, or VCF, was developed by 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 dbVar are currently based on VCF format version 4.1, with some additional changes and requirements as detailed below.

NOTE: **If you have human mutations or variations with clinical significance or phenotype, submit your data directly to ClinVar** (http://www.ncbi.nlm.nih.gov/clinvar/). Once your variants have been curated by the ClinVar staff, they will be forwarded to dbVar where they will be assigned dbVar IDs and displayed alongside non-clinical variants. If you have questions regarding your submission, contact dbVar at dbvar@ncbi.nlm.nih.gov

When should I use the VCF Submission Format?

Use dbVar's Variant Call Format (VCF) to submit large or small numbers of structural variations that have asserted positions on genomes or on reference sequences[a]. Large scale submitters especially will find dbVar'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 dbVar submissions, as opposed to the standard VCF format as defined by the 1000 Genomes project, includes additional fields and attributes that describe dbVar-specific submission and variation properties, and may include tags that are different than those used in standard VCF.

[a]dbVar 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, or GenBank.

VCF Submission Overview

A complete dbVar VCF submission consists of two files: a VCF-formatted file describing your structural variation data, and the standard dbVar Submission Template with the following sections completed: Contacts, Study, Samplesets, Samples, and Experiments (the Variant Calls and Variant Regions sections can be omitted). Instructions and requirements for completing the standard submission template are available in our online documentation and in the Excel version of the submission template; there is no pre-formatted template for the VCF file. It is important to note that some data in the VCF file and the submission template will cross-reference each other (e.g., sample_ids, sampleset_ids, experiment_ids, validation results, etc.), so be sure to double-check that all cross-referenced data are in agreement before submitting your files to dbVar.

If dbVar finds any conflict between cross-referenced material in your submission, your submission files will be returned to you for correction and resubmission.

VCF Submission Steps

1. Download a copy of the dbVar Submission Template – in Excel, tab-delimited, or XML format – and complete the following sections: **Contacts**, **Study**, **Samplesets**, **Samples**, and **Experiments**. Some of the information you record here will be needed for the VCF file. Do **NOT** complete the **Variant Calls** and **Variant Regions** sections – your variants will be recorded in VCF using the instructions that follow.

   NOTE: NCBI Variation Resources will soon be changing over to a Common Submission Portal for data submissions. Users will be able to log in using their MyNCBI ID and upload their submissions rather than emailing them. Watch the following URL for progress: https://submit.ncbi.nlm.nih.gov/subs/variation/**]**

2. Create a plain-text VCF file consisting of a header section and a data table section. Requirements and guidelines are provided below; be sure to include all required metadata and tags in your VCF file.

3. Enter your data in the data table section, one variant per row. In addition to the standard required fields, include any optional INFO and FORMAT tags needed to describe the variants in detail. Remember to copy complete tag definitions for any tags you use, from Appendix A: Library of dbVar VCF Tag Definitions into the header section of your VCF file.

4. Double-check that any cross-referenced data between your files is in agreement. Any conflicting data will result in your files being returned to you.

5. Email both your VCF file and submission template to dbvar@ncbi.nlm.nih.gov. If the files are too large to email send us a link to the files stored on your local FTP or cloud storage service instead, or contact us at dbvar@ncbi.nlm.nih.gov to arrange for FTP upload of your files.

dbVar VCF Submission Format

The VCF portion of your submission should contain a header section and a data table section.

dbVar's VCF submission specification is very similar to the 1000 Genomes VCF specification v.4.1, and adheres to it wherever possible. Occasionally, dbVar has had to modify these requirements to obtain additional information from our submitters.

Differences between dbVar VCF and Standard VCF

dbVar VCF differs from standard VCF in two ways:

1. dbVar VCF introduces several new INFO tags (some required) and one FORMAT tag. These are explained below.

2. dbVar uses a slightly different system for reporting coordinates of imprecise variants. Details provided below.

1. New dbVar INFO Tags

The following INFO tags are not in the standard (1000 Genomes) VCF specification, but are often needed in dbVar submissions. For a complete list of all standard and dbVar-specific tags and definitions please consult Appendix A: Library of dbVar VCF Tag Definitions.

EXPERIMENT	***(Required)*** ID of the experiment that identified the variant*
SAMPLE	***(Required)*** ID of the sample in which the variant was observed*
SAMPLESET	***(Required)*** ID of the sampleset in which the variant was observed*
POSrange	A pair of coordinates which defines the **POS** breakpoint region
ENDrange	A pair of coordinates which defines the **END** breakpoint region
valEXPERIMENT	ID of the validation experiment(s), and the variant's Pass/Fail status*
DESC	Description; any additional information you wish to include
ORIGIN	Genetic origin of the observed allele (see Appendix A for allowed terms)
PHENO	Phenotypic information associated with the variant
LINKS	Links to external databases — e.g., GEO:GPL4040. See dbVar submission template (LINKS section) for more examples.
refCN	(**FORMAT** tag) Reference copy number when calling CNV variants

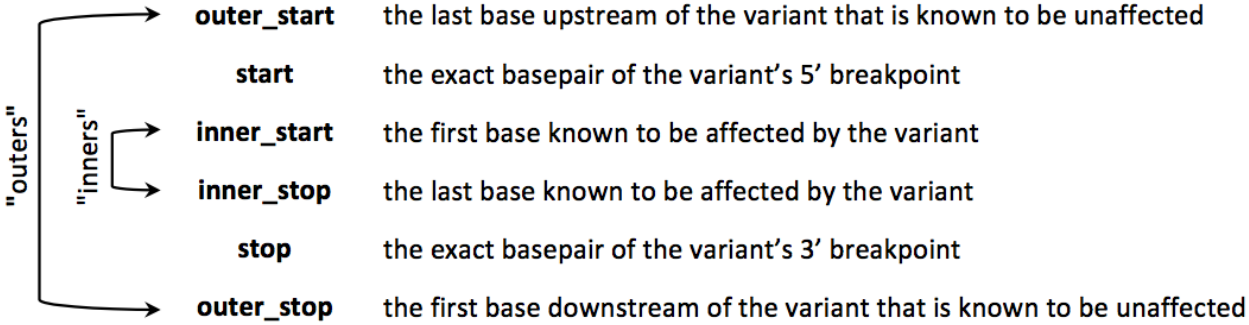
***NOTE**: The EXPERIMENT, SAMPLE and SAMPLESET values provided in the INFO tags must match those listed in the dbVar submission Excel file that accompanies your VCF submission.

2. Reporting Coordinates for Imprecise Variants

Structural variation data includes a variety of variant types of differing sizes and properties, and frequently it is not possible to determine the precise locations of breakpoints. Describing structural variation data, therefore, requires that the specification capture details of the variant's properties, including its location and the extent of ambiguity in its breakpoint boundaries.

When describing imprecise variants whose breakpoints are not known to basepair resolution, standard VCF instructs submitters to use "best estimates" for **`POS`** and **`END`** values, and to provide confidence intervals (**`CIPOS`** and **`CIEND`**) to give an approximation of the region where the actual breakpoints are likely to fall. This approach is appropriate when the underlying data are statistical in nature. It is often the case, however, that the data may provide precise coordinates clearly defining the boundaries of the region within which the breakpoints must fall. This can happen, for example, in array-based or paired-end mapping experiments. To capture these precise data, the dbVar data model uses three pairs of coordinates to define the boundaries limiting breakpoint uncertainty:

dbVar Data Model



NOTES:

1. These terms are used in the dbVar data model - ***they are not VCF tags***.
2. **start** and **stop** coordinates are "absolutes," and should only be used when exact breakpoints are known (or in standard VCF fashion, with **`CIPOS`** and **`CIEND`**).
3. **inners** and **outers** can be used in isolation or together when describing a variant.

4. 5’ and 3’ breakpoints may have differing degrees of ambiguity, as determined by the data.

Representing Inners and Outers in VCF

dbVar introduces two new INFO tags — **`POSrange`** and **`ENDrange`** — to represent inner and outer coordinates in VCF.

`POSrange`: This INFO tag includes a pair of comma separated values that define the left and right boundaries of the **`POS`** breakpoint region. These two values are equivalent to dbVar's **outer_start** and **inner_start**, respectively.

`ENDrange`: This INFO tag includes a pair of comma separated values that define the left and right boundaries of the **`END`** breakpoint region. These two values are equivalent to dbVar's **inner_stop** and **outer_stop**, respectively.

NOTES:

1. The value to the left of the comma should always be less than the value to the right
2. One **`POSrange`** value must be the same as the **`POS`** value.
3. One **`ENDrange`** value must be the same as the **`END`** value
4. Use a dot (".") placeholder to represent unknown values

Example:

Suppose an oligo array CGH experiment detects a series of contiguous probes showing a 50% increase in signal, indicative of a duplication event. The 5' breakpoint of the duplication has to fall somewhere between the 3’ end of the last probe demonstrating normal intensity and the 5’ end of the first probe showing increased intensity. Similarly, the 3' breakpoint of the duplication must fall between the 3’ end of the last probe showing increased intensity and the 5’ end of the first probe downstream that shows normal (baseline) intensity.

outer_start = 2500000
inner_start = 2501000
inner_stop = 3499000
outer_stop = 3500000

Depending on which values are available in the raw data (only outers, only inners, or both), VCF fields would be populated as follows:

	`POS`	**`POSrange`**	**`ENDrange`**	**`END`**

outers only	2500000	2500000,.	.,3500000	3500000
inners only	2501000	.,2501000	3499000,.	3499000
outers & inners	2500000	2500000, 2501000	3499000, 3500000	3500000

The VCF entry for the last row above might look something like this:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	2500000	your_id	T	<DUP>	.	PASS	SVTYPE=DUP;END=3500000;POSrange=2500000,2501000;ENDrange=3499000,3500000

Note that in all cases, **POS** = the minimum provided coordinate and **END** = the maximum provided coordinate.

IMPORTANT: The standard VCF tags **CIPOS** and **CIEND** are permitted in dbVar VCF and should be used when the underlying data is statistical in nature. Use whichever pair of terms — **POSrange** and **ENDrange** or **CIPOS** and **CIEND** — is best supported by your underlying data.

dbVar VCF Submission File

Required VCF Header Metadata

The VCF file header for a dbVar submission must begin with the following three metadata lines:

```
##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}
##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 (http://www.ncbi.nlm.nih.gov/assembly/) and
                 search for the record of the specific assembly. 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 may 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.}
```

NOTE: If you do not place the required metadata in the VCF file header, your submission files will be returned to you for correction and resubmission.

ALT Tags

Placement of ALT Tag/Value Descriptions

The VCF header continues with tag/value descriptions for the VCF ALT tags. Place the ALT tag/value descriptions in the header following the required metadata; they will serve to define the data you place in the ALT column of the submission data table.

The ALT tags used in dbVar VCF are the same as those used in standard (1000 Genomes) VCF: see VCF v.4.1 for detailed information regarding the use of ALT tags or see Appendix A: Library of dbVar VCF Tag Definitions in this document for example ALT tag descriptions you can cut and paste into the VCF file header.

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

NOTE: Although use of ALT tags is not required, it is strongly recommended when it is not possible to list the full sequence of the submitted structural variation.

INFO Tags

Following the ALT tags, the dbVar VCF header continues with tag/value descriptions for both required and optional dbVar INFO tags.

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

See Appendix A: Library of VCF Tags and Descriptions for a list of dbVar VCF required and optional INFO tags as well as example tag descriptions you can cut and paste into the VCF file header for both the

required INFO tags and the optional INFO tags. See also VCF v.4.1 for detailed information regarding the use of the INFO tags used in standard VCF that are also used in dbVar VCF

Submission Data Table

The submission data table is a tab-delimited table that houses the variations and variation data that you will be submitting. The table header should include the eight fixed, mandatory columns (in order) shown below:

#CHROM POS ID REF ALT QUAL FILTER INFO

dbVar requires submitters to be specific about the nature and extent of breakpoint ambiguity present in their data when reporting imprecise structural variants. While the 1000 Genomes specification asks submitters to populate POS and END with their "best estimate" supported by confidence intervals (using CIPOS and CIEND), dbVar's specification as described below requests submitters to actually specify inner and outer start-stop coordinates for imprecise structural variants.

See the Examples of Structural Variation in dbVar VCF Format section of this document for case studies showing how to use dbVar's VCF specifications to provide greater break point specificity for imprecise variants.

For definitions and basic information regarding inner and outer start-stop coordinates, see the Capturing Variant Information section of dbVar's Structural Variation Overview.

Data Table Field Values

CHROM

This field contains the chromosome identifier from the reference genome where the variant is located (cf. the ##reference line in the header). Chromosomes are sorted numerically in increasing order (1 – 23, X, Y). Alternatively, the sequence accession and version can be used for this field if the variation position is based on a non-chromosomal sequence.

POS

This field contains the reference position of the variant, which is the 1st base of the variation event. All coordinates should be 1-based. Positions are sorted numerically within each reference sequence chromosome (CHROM) so that entries for a specific CHROM form a contiguous block within the VCF file. You are permitted to have multiple records of different structural variation types (SVTYPE) at the same POS – list shortest variants first. Telomeres are indicated by using positions 1 (p-arm) or chromosome length (q-arm).

NOTE: Single nucleotide variants and small (< 50 bp) insertions and deletions must be submitted to dbSNP.

ID

This field contains a unique identifier (ID) for the variant and is required.

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, or N (case insensitive).

Although standard VCF specification requires that literal sequence representing the REF allele should be provided, it can be difficult to represent the REF allele as literal sequence since structural variants can be quite large and there is often ambiguity in the locations of their breakpoints. It is therefore acceptable to list only the first base of the REF allele.

ALT

Although standard VCF specification requires that literal sequence representing the ALT allele should be provided, it can be difficult to represent the ALT allele as literal sequence since structural variants can be quite large and there is often ambiguity in the locations of their breakpoints. It is therefore preferable to provide one of several ALT tags, surrounded by angle brackets, to indicate the nature of variation at the ALT allele.

See Appendix A: Library of VCF Tag Descriptions for a detailed description of available ALT tags, instructions on how to use ALT tags to identify structural variants, and example ALT tag descriptions you can cut and paste into the VCF file header.

NOTE: Although use of ALT tags is not required, it is strongly recommended when it is not possible to list the full sequence of the submitted structural variation.

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>

See Appendix A: Library of VCF Tag Descriptions for a detailed description of required and optional INFO tags, instructions on how to use INFO tags to identify variant coordinates, and example tag descriptions you can cut and paste into the VCF file header for both the required INFO tags and the optional INFO tags. A list of new, dbVar-specific INFO tags is also included in the section above titled, "New dbVar INFO Tags."

Examples of Structural Variation in dbVar VCF Format

Insertion (precise)

An insertion of 981 base pairs immediately to the right of coordinate 14588694.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
3	14588694	1	C	<INS>	.	PASS	SVTYPE=INS;END=14588694;SVLEN=981;SAMPLE=NIN044;EXPERIMENT=2

Insertion (imprecise)

An insertion of approximately 1500 base pairs, as determined by a paired-end mapping (PEM) experiment. All we know from the data is that the insertion is roughly 1500 bases and took place somewhere between the coordinates corresponding to the mapped sequence reads. Neither the sequence content nor the precise size of the insertion are known.

Even though CHROM:POS (chr5:2588032) corresponds to a real base in the reference, it would be meaningless to assign this base to the "REF" field because it is not where the insertion took place. So a dot placeholder (".") is used for "REF" instead.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
5	2588032	2	G	<INS>	.	PASS	SVTYPE=INS;END=2589946;POSrange=2588032,.;ENDrange=.,2589946;SVLEN=1500;IMPRECISE;SAMPLE=Walt07;EXPERIMENT=3

Deletion (precise)

A deletion of 387 base pairs (the deleted bases are 2599384 thru 2599770).

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
6	2599384	3	A	<DEL>	.	PASS	SVTYPE=DEL;END=2599770;SVLEN=387;SAMPLE=NIN044;EXPERIMENT=2

Deletion (imprecise)

A deletion of approximately 8000 base pairs, as determined by a paired-end mapping (PEM) experiment. All we know from the data is that the deletion is roughly 8000 bases and took place somewhere

between the coordinates corresponding to the mapped sequence reads. Neither the sequence content nor the precise size of the deletion are known.
Even though CHROM:POS (chr14:14885937) corresponds to a real base in the reference, it would be meaningless to assign this base to the "REF" field because it is not where the deletion took place. So a dot placeholder (".") is used for "REF" instead.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
14	14885937	4	A	<DEL>	.	PASS	SVTYPE=DEL;END=14899924;POSrange=14885937,.;ENDrange=.,14899924;SVLEN=8000;IMPRECISE;SAMPLESET=1;EXPERIMENT=1

Inversion (precise)

An inversion of 863 base pairs.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
17	2553052	5	G	<INV>	.	PASS	SVTYPE=INV;END=2553914;SVLEN=863;SAMPLE=Jesse08;EXPERIMENT=2

Duplication (Imprecise)

A duplication determined by oligo array CGH. The nature of arrays is such that breakpoints cannot be determined to base pair resolution, only to a range defined by probes on the array. dbVar encourages the submission of both inner and outer coordinates if known, but will also accept inners or outers separately.

Inners only:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	2501000	6	C	<DUP>	.	PASS	SVTYPE=DUP;END=3499000;IMPRECISE;POSrange=.,2501000;ENDrange=3499000,.;SAMPLESET=3;EXPERIMENT=3

Outers only:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	2500000	6	A	<DUP>	.	PASS	SVTYPE=DUP;END=3500000;POSrange=2500000,.;ENDrange=.,3500000;SAMPLESET=3;EXPERIMENT=3

Inners and Outers:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	2500000	6	A	<DUP>	.	PASS	SVTYPE=DUP;END=3500000;POSrange=2500000,2501000;ENDrange=3499000,3500000;SAMPLESET=3;EXPERIMENT=3

Deletion (or "copy number loss") (imprecise)

Same to Duplication (imprecise) above, except ALT and SVTYPE both equal <DEL> instead of <DUP>.

Appendix A: Library of VCF Tag Descriptions

ALT Tags

```
##ALT=<ID=DEL,Description="Deletion relative to the reference">
##ALT=<ID=DUP,Description="Region of elevated copy number relative to the reference">
##ALT=<ID=INS,Description="Insertion of sequence relative to the reference">
##ALT=<ID=INV,Description="Inversion of reference sequence">
##ALT=<ID=CNV,Description="Copy number polymorphic region">
##ALT=<ID=DUP:TANDEM,Description="Tandem duplication">
##ALT=<ID=INS:NOVEL,Description="Insertion of sequence that does not map to the
    reference">
##ALT=<ID=INS:ME,Description="Insertion of a mobile element relative to the
    reference">
##ALT=<ID=INS:ME:ALU,Description="Insertion of an Alu mobile element relative to the
    reference">
##ALT=<ID=DEL:ME,Description="Deletion of a mobile element relative to the reference">
##ALT=<ID=DEL:ME:ALU,Description="Deletion of an Alu mobile element relative to the
    reference">
```

etc., substituting other mobile element families for **ALU**

FORMAT Tags

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise
    variants">
##FORMAT=<ID=refCN,Number=1,Type=Integer,Description="Reference copy number used as a
    baseline when calling DELs, DUPs, and CNVs; if not specified, value is assumed to
    be '2' (i.e. a unique diploid locus)">
```

INFO Tags

```
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for
    imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for
    imprecise variants">
##INFO=<ID=DESC,Number=1,Type=String,Description="Any additional information about
    this call that is not covered elsewhere. Free text enclosed in double quotes.">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant
    described in this record">
##INFO=<ID=ENDrange,Number=2,Type=Integer,Description="For imprecise variants, if END
    represents an inner_stop or outer_stop coordinate, use a comma-delimited list to
    indicate this; e.g., if END is an inner_stop and its value is 108442336, you would
    enter '108442336,.">
##INFO=<ID=EXPERIMENT,Number=1,Type=Integer,Description="experiment_id from dbVar
    submission of the experiment that generated this call">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=LINKS,Number=.,Type=String,Description="Link(s) to external database(s),
    e.g., 'GEO:GPL4010'. See LINKS tab of dbVar submission template for more
    examples">
```

```
##INFO=<ID=ORIGIN,Number=1,Type=String,Description="Origin of allele if known; must be
    one of: de novo, Maternal, Paternal, Present in both parents, Not tested, Tested -
    inconclusive; must be enclosed in double quotes. See also SOMATIC">
##INFO=<ID=PHENO,Number=.,Type=String,Description="Phenotype(s) associated with this
    call; NOT for clinical assertions (which should be submitted to ClinVar). Free
    text enclosed in double quotes.">
##INFO=<ID=POSrange,Number=2,Type=Integer,Description="For imprecise variants, if POS
    represents an inner_start or outer_start coordinate, use a comma-delimited list to
    indicate this; e.g., if POS is an inner_start and its value is 2865734, you would
    enter '.,2865734">
##INFO=<ID=SAMPLE,Number=1,Type=String,Description="sample_id from dbVar submission.
    Each call must have only one of either SAMPLE or SAMPLESET">
##INFO=<ID=SAMPLESET,Number=1,Type=Integer,Description="sampleset_id from dbVar
    submission. Each call must have only one of either SAMPLE or SAMPLESET">
##INFO=<ID=SOMATIC,Number=0,Type=Flag,Description="Somatic mutation. NOT for clinical
    assertions, i.e. cancer. See also ORIGIN">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF
    and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant; must
    be one of: DEL, INS, DUP, INV, CNV">
##INFO=<ID=valEXPERIMENT,Number=.,Type=String,Description="experiment_id(s), from
    dbVar submission, of the experiment(s) used to validate this call — followed by a
    colon and 'Pass' or 'Fail'. E.g., '6:Pass,8:Fail'">
##INFO=<ID=VALIDATED,Number=0,Type=Flag,Description="Validated by follow-up
    experiment">
```

Appendix B: Example of a dbVar VCF Submission File

```
##fileformat=VCFv4.1_dbVar
##fileDate=20140803
##reference=GCF_000001405.13
##ALT=<ID=DEL,Description="Deletion relative to the reference">
##ALT=<ID=DUP,Description="Region of elevated copy number relative to the reference">
##ALT=<ID=INS,Description="Insertion of sequence relative to the reference">
##ALT=<ID=INV,Description="Inversion of reference sequence">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise
    variants">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant; must
    be one of: DEL, INS, DUP, INV, CNV">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant
    described in this record">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF
    and ALT alleles">
##INFO=<ID=SAMPLE,Number=,Type=,Description="sample_id from dbVar submission; every
    call must have SAMPLE or SAMPLESET, but NOT BOTH">
##INFO=<ID=SAMPLESET,Number=,Type=,Description="sampleset_id from dbVar submission;
    every call must have SAMPLESET or SAMPLE but NOT BOTH">
##INFO=<ID=EXPERIMENT,Number=1,Type=Integer,Description="experiment_id (from
    EXPERIMENTS tab) of the experiment that generated this call">
##INFO=<ID=PHENO,Number=.,Type=String,Description="Phenotype(s) thought to associated
    with this call; however all clinical assertions should be submitted to ClinVar.
    Use free text enclosed in double quotes.">
##INFO=<ID=LINKS,Number=.,Type=String,Description="Link(s) to external database(s) -
    see LINKS tab of dbVar submission template for examples">
#CHROM POS    ID     REF    ALT    QUAL   FILTER INFO   FORMAT
1   25883940 1      C      <DEL>  .      .      SVTYPE=DEL;END=25895965;SVLEN=12025;
    SAMPLESET=1;VALIDATED;valEXPERIMENT=5      GT:CN  0/1:1
3   25883940 2      C      <DUP>  .      .
      SVTYPE=DUP;END=25895965;SVLEN=12025;EXPERIMENT=3; SAMPLESET=1;VALIDATED;
    valEXPERIMENT =5       GT:CN  ./.:4
4   1685628  3      A      <INS>  .      .
      SVTYPE=INS;END=1685628;SVLEN=1688;EXPERIMENT=1;SAMPLESET=1;PHENO="Rapid growth,
    Heat sensitive";LINKS=GEO:GPL4010
5   8277466  4      G      <INS>  .      .
      SVTYPE=INS;END=8277466;SVLEN=2510;SAMPLESET=1;EXPERIMENT=2
13  2675081  5      A      <INV>  .      .
      SVTYPE=INV;END=2675906;SVLEN=825;SAMPLESET=1;EXPERIMENT=4
```