Genome Variation Format

Table of Comments

Summary

Version 1.05
19 January 2011

The Genome Variation Format (GVF) is a file format for describing sequence variants at nucleotide resolution relative to a reference genome. The GVF format was published in *Reese et al., Genome Biol., 2010;11(8):R88* A standard variation file format for human genome sequences. We would like to acknowledge the contributing groups for their support.

Karen Eilbeck	Biomedical Informatics University of Utah
Paul Flicek	ENSEMBL EBI
Gabor Marth	Boston University 1000 Genomes Project
Martin Reese	Omicia Inc.
Lincoln Stein	Ontario Cancer Institute

News

March 2011 SOBA workshop presented at GMOD Spring Training.

Feb 11th 2011 The CVS repository on sourceforge is back online after the hacking incident.

January 2011 SO presented at Synthetic Biology Workshop.

December 2010 10Gen Data Set v1.04 available.

December 2010 GVF Specification v1.04 available.

November 2010 Release 2.4.4 available.

September 2010 A standard variation file format for human genome sequences in Genome Biology.

August 2010 'Toward a Richer Representation of Sequence Variation in the Sequence Ontology' with Mike Bada accepted by AIMM Workshop.

June 2010 Release 2.4.3 available.

May 2010 SOBA paper available from Nucleic Acids Research.

April 2010 Release 2.4.2 available.

March 2010 New SO paper out: 'Evolution of the Sequence Ontology terms and relationships' in the Journal of Biomedical Informatics. A pre-print of this paper is available here.

February 2010: Sequence Ontology Bioinformatics Analysis (SOBA) is available.

Mark Yandell	Department of Human Genetics University of Utah

GVF is a type of GFF3 file with additional pragmas and attributes specified. The GVF format has the same nine-column tab delimited format as GFF3 and all of the requirements and restrictions specified for GFF3 apply to the GVF specification as well. In addition GVF adds additional constraints to some of these columns as described below.

See the GFF3 Specification for more details about GFF3.

Quick Example of GVF Content

A few lines of single nucleotide variants (SNV) are shown below as an example of a very simple GVF file. Scroll right to see the complete lines or view a full screen text version.

```
##gvf-version 1.05
##feature-ontology http://sourceforge.net/projects/song/files/Sequence%20Ontology/
##genome-build NCBI B36.3
##sequence-region chr16 1 88827254

chr16   samtools        SNV     49291141        49291141        .       +       .
chr16   samtools        SNV     49291360        49291360        .       +       .
chr16   samtools        SNV     49302125        49302125        .       +       .
chr16   samtools        SNV     49302365        49302365        .       +       .
chr16   samtools        SNV     49302700        49302700        .       +       .
chr16   samtools        SNV     49303084        49303084        .       +       .
chr16   samtools        SNV     49303156        49303156        .       +       .
chr16   samtools        SNV     49303427        49303427        .       +       .
chr16   samtools        SNV     49303596        49303596        .       +       .
```

Likewise, a few lines of a more complete GVF file are shown below with the functional consequence of each variant annotated relative to their effect on coding for mRNA/protein annotations (text version):

```
##gvf-version 1.05
##feature-ontology http://sourceforge.net/projects/song/files/Sequence%20Ontology/
##genome-build NCBI B36.3
##sequence-region chr16 1 88827254

chr16   samtools        SNV     49291141        49291141        .       +       .
chr16   samtools        SNV     49302125        49302125        .       +       .
chr16   samtools        SNV     49302365        49302365        .       +       .
chr16   samtools        SNV     49302700        49302700        .       +       .
chr16   samtools        SNV     49303084        49303084        .       +       .
chr16   samtools        SNV     49303156        49303156        .       +       .
chr16   samtools        SNV     49303427        49303427        .       +       .
chr16   samtools        SNV     49303596        49303596        .       +       .
```

GVF Specification Archive

Previous versions of the GVF specification are archived. Please see the GVF Specification Archive to find these older versions of the specification.

GVF Wiki Pages

This page provides the official description of the GVF format with some simple examples. However, several GVF wiki pages have been set up to describe the GVF format in a more tutorial manner, to provide additional examples and to document implementations of GVF files by the GVF community.

Data Sets

10Gen Data set

The SNVs from ten of the first sequenced individual human genomes have been converted to GVF format and are made available as the 10Gen Data set.

Genome annotations in GFF3 can be uploaded and analysed.

2009 December Release 2.4.1 available.

2009 October Release 2.4 available.

2009 July SO presented at the International Conference of Biomedical Ontology.

2009 June Chris Conley - Undergrad from BYU joins SO for the summer

2009 May Graduate student, John Naylor joins SO.

2009 February 22: A new SO related paper Quantitative Measures for the Management and Comparison of Annotated Genomes K Eilbeck *et. al.*

2009 January 4-6: SO represented at the RNA Ontology consortium meeting in Cambridge UK.

2008 December 1: A paper from the BioSapiens project describing protein features in SO is published. The Protein Feature Ontology: A tool for the unification of protein feature annotations. GA Reeves *et. al.*

Data from ENSEMBL

```
ENSEMBL is providing GVF files for their variant data sets at:
ftp://ftp.ensembl.org/pub/current/variation/
```

Data from NCBI

```
The dbVar database at NCBI is providing GVF files for their structural
variant data at: (Links are temporarily unavailable)
```

Column Descriptions

```
Sequence alterations are described in a GVF file with 9 tab delimited
columns.  The columns of a GVF file are inherited from GFF3 with
additional constraints placed on some columns.  A brief descriptions
of each column follows:

Column 1: "seqid" The ID of the landmark used to establish the
          coordinate system for the current feature. IDs may contain
          any characters, but must escape any characters not in the
          set [a-zA-Z0-9.:^*$@!+_?-|]. In particular, IDs may not
          contain unescaped white space and must not begin with an
          unescaped ">".

Column 2: "source" The source is a free text qualifier intended to
          describe the algorithm or operating procedure that generated
          this feature. Typically this is the name of a piece of
          software, such as "MAQ" or a database name, such as "dbSNP."
          The source may also be used to describe the individual
          carrying the variant such as "NA_18507" if multiple
          individuals are combined in a single GVF file.

Column 3: "type" The type of the feature.  This is constrained to be
          either: (a) the SO term sequence_alteration (SO:0001059),
          (b) a child term of sequence_alteration, (c) the SO term gap
          (SO:0000730), or (d) the SO accession number for any of the
          previous terms.  The gap feature, while not a
          sequence_alteration, provides a way to distinguish regions
          where variant information is unknown (low coverage no-call
          regions) from regions that have no variation.

Columns 4 & 5: "start" and "end" The start and end of the feature, in
          1-based integer coordinates, relative to the landmark given
          in column 1. Start is always less than or equal to end. For
          features that cross the origin of a circular feature
          (e.g. most bacterial genomes, plasmids, and some viral
          genomes), the requirement for start to be less than or equal
          to end is satisfied by making end = the position of the end
          + the length of the landmark feature. For zero-length
          features, such as insertion sites, start equals end and the
          implied site is to the three-prime of the indicated base in
          the direction of the landmark.

Column 6: "score" The score of the feature, an integer or floating point
          number. The semantics of the score are not defined, however
          it is strongly recommended that a Phred scaled quality
          score be used whenever possible.

Column 7: "strand" The strand of the feature. + for positive strand
          (relative to the landmark), - for minus strand, and . for
          features that are not stranded. In addition, ? can be used
          for features whose strandedness is relevant, but unknown.

Column 8: The phase column is not used in GVF, but is maintained with
          the placeholder '.' for compatibility with GFF3 and tools
          that conform to the GFF3 specification.
```

GVF Column 9 Attributes

Attribute Summary

```
Column 9 in GFF3 allows use of tag-value pairs to define attributes of
the feature.  All of the GFF3 attribute tag-value pairs are supported
in GVF as well.  Note especially that the GFF3 attribute tags ID,
Alias Dbxref are useful in GVF.  GVF specifies the additional tag
definitions given below.  As specified by GFF3, all attribute tags
beginning with upper-case letters are reserved for future use by the
```

GFF3 and GVF specifications, however applications are free to use attributes beginning with lower-case letters to provide custom data.

Community Attributes Collection

The Sequence Ontology provides a collection of descriptions of attributes in use by the GVF community. These descriptions are of organization specific attributes (those attributes beginning with a lower case letter that are specific to that organizations data sets) or they provide clarification about how the organizations data sets fit within the existing attribute structure. Note that existing attributes should not be redefined here. This information is provided as a wiki. Please refer to GVF Community Supported Attributes for more information. If you would like to maintain a description of attributes in use by your organization please contact the Sequence Ontology group to set up an account on the wiki.

Attribute Tag Definitions

ID

While the GFF3 specification considers the ID tag to be optional, GVF requires it. As in GFF3 this ID must be unique within the file and is not required to have meaning outside of the file. If no ID is available, it can default to seqid:source:type:start adding an incremented value at the end to provide uniqueness if necessary. For example chr1:Soap:SNP:12345 or chr1:Soap:SNP:12345:001.

Alias

A secondary name for the feature. It is suggested that this attribute be used if needed to describe the variant in other nomenclature systems such as the HGVS nomenclature.

Dbxref

A database cross reference. This can be useful to associate this variant with the same variant previously described in variant databases such as dbSNP or OMIM. See the Database Cross References section below

Variant_seq

All sequences found in this individual (or group of individuals) at a variant location are given with the Variant_seq tag. These sequences are often referred to as alleles, although allele has a more general meaning. Note that the reference sequence should be given here as well when appropriate.

```
Variant_seq=A;
Variant_seq=A,T;
Variant_seq=A,C,G,T;
```

If the sequence is longer than 50 nt the sequence may be (but is not required to be) abbreviated as '~'. In this case the location of the variant, but not it's actual sequence will be known. In the case where the variant represents a deletion of sequence relative to the reference, the variant_seq is given as '-'. This tag is optional, but should always be given where applicable and should contain a value for each unique sequence called for this location including the sequence that is identical to the reference sequence if applicable.

Reference_seq

The sequence from the reference sequence corresponding to the start and end coordinates of this feature. In the case where the variant represents an insertion relative to the reference, the Reference_seq is given as '-'. This tag is optional as software should be able to access this data from the FASTA sequence data associated with the GVF file, however providing this tag allows some analyses to be done on the GVF file without accessing the sequence data.

Variant_reads

The number of reads supporting each variant at this location given in the form reads. Read counts for variants longer than one

nucleotide should be average read counts for each position over the length of the variant. The order of the reads must be in the same order that the corresponding sequences in the Variant_seq tag. A '.' character may be used as a placeholder to signify missing data. This tag is optional, but strongly encouraged.

Total_reads

The total number of reads covering a variant. Total read counts for variants longer than one nucleotide should be average read counts for each position over the length of the variant. The sum of reads given in Variant_reads is not required to be the same as Total_reads. This tag is optional, but strongly encouraged.

Genotype

The genotype of this locus where genotype is heterozygous, homozygous or hemizygous. This tag is optional, but strongly encouraged.

Variant_freq

A real number describing the frequency of the variant in a population. The details of the source of the frequency should be described in an attribute-method pragma as discussed above. The order of the values given must be in the same order that the corresponding sequences occur in the Variant_seq tag. A '.' character may be used as a placeholder to signify missing data. This tag is optional.

Variant_effect

The effect of a variant on a sequence features that overlaps it is given in the form:

sequence_variant index feature_type feature ID,feature ID

The value of the Variant_effect tag has four white-space delimited fields:

The first field is a term that describes the effect of the variant on a sequence feature and must be the SO term sequence_variant or one of its children.

The second field is an index value (0-based) that identifies which Variant_seq is being described. For example, if this Variant_effect tag is describing the effect of the first sequence in the Variant_seq tag then the index value would be 0, for a Variant_effect tag describing the second sequence in the Variant_seq tag the index value would be 1 and so on (see the examples below). The index is necessary because multiple Variant_effect tags may refer to the same Variant_seq.

The third field is a term describing the sequence feature that is being affected. This term must be the SO term sequence_feature or one of its children.

The fourth field is a single (or a comma separated list of) feature ID(s). Ideally these feature IDs will correspond to ID attributes in another GFF3 file that describes the sequence features annotated for the genome.

White space is not allowed within any of the four fields. If IDs have white space they must be escaped as hex escape %20. Both of the SO terms can be given either as a valid SO term name or the SO ID. Each Variant_effect tag can only describe one effect for one variant sequence. If a single Variant_seq has different effects on different sequence features (for example it causes a non_synonymous_codon change in one mRNA, but a synonymous_codon change in another mRNA) then multiple Variant_effect tags are required. This tag is optional.

Variant_copy_number

For regions on the variant genome that exist in multiple copies, this tag represents the copy number of the region as an integer value in the form Variant_copy_number=integer. This tag is optional.

Reference_copy_number

```
    For regions on the reference genome that exist in multiple copies,
    this tag represents the copy number of the region as an integer in
    the form Reference_copy_number=integer.  This tag is optional.

Start_range End_range

    The Start_range and End_range attributes describe ambiguity in the
    start and end coordinates given in column 4 and 5.  The
    description below is given for Start_range, but the same rules
    apply for End_range.  The value is required to be two integers (or
    a '.'  for unknown values) separated by a comma.  The first value
    defines a range of ambiguity less-than the value given in column 4
    and must be less-than or equal-to that coordinate (or '.')  The
    second value defines a range of ambiguity greater-than the
    coordinate specified in column 4 and must be greater-than or
    equal-to the value of that coordinate (or '.').  If either value
    is equal-to the coordinate in column 4 then there is no ambiguity
    in that direction.  The values given for these attributes are
    always relative to the landmark feature just as they are for the
    coordinates given in columns 4 and 5.

Phased

    The Phased tag can be used to define a set of variants whose
    genotypes are phased.  A phased genotype indicates that for
    heterozygous loci it is known which chromosome each variant
    sequence belongs to.  That is, if one variant has a
    Variant_seq=A,C attribute and another variant on the same landmark
    feature (chromosome) has a Variant_seq=T,G attribute then the A
    and T occur together on one copy of the chromosome and the C and G
    occur together on the other copy of that chromosome.  If an entire
    GVF file (or a defined subset of it - for example all SNVs) is
    phased, then use the '##phased-genotypes' pragma described above.
    However, if only regions of the genome are phased then the Phased
    attribute allows you to specify those features and to describe
    which features are phased together if mutliple phased regions
    exist.  The Phased attribute only has meaning for loci that are
    heterozygous, but may be given for any variant.  The value of the
    Phased attribute can be any alphanumeric value.  This value will
    serve as an ID for features that are phased together and is not
    required to have meaning outside the file or in other contexts
    within the file.  If multiple regions of the chromosome are phased
    then separate values should be used to group each set of phased
    features.  To avoid amibuity the same value should not be used for
    more than one phased region even if they lie on different landmark
    features (chromosomes).
```

GVF Feature Examples

```
    A heterozygous SNV with sequences G and C having 17 and 16 reads
    respectively.  This SNV falls within the CO'S of the RefSeq mRNAs
    NM_012345 and NM_543210 and the variant sequence creates an allele
    of this gene with a non-conservative substitution.  The second
    variant sequence 'C' is the same as the reference.

chr1    SOAP    SNV     15883   15883   36.5    +       .       ID=chr1:SOAP:SNV:1

    A homozygous deletion in the individual genome relative to the
    reference genome.  The region deleted is longer than 50
    nucleotides and thus the GVF simply has a'~' as the
    Reference_sequence value.  There are a total of 27 reads on
    average spanning this region, of which 26 on average supported the
    deletion.

chr1    Celera  nucleotide_deletion     8834426 8834497 .       +       .       ID

    A Copy number variant created by expansion of a segmental
    duplication that was already present in the reference genome.

chr10 PennCNV   copy_number_variation   3922747 3923761 .       +       .       ID

    A Copy number variant from NCBI dbVar for which there is some
    ambiguity in the start and end coordinates that is described with
    the Start_range and End_range tags.

NC_000010.9 dbVar copy_number_variation 51580055 51580298 . . . ID=nssv8537;Name=n
```

```
    Another copy number variant as above.  In this example the outer
    boundaries of the ranges are given with nucleotide resolution, but
    the inner boundaries are unknown.

NC_000024.7 dbVar copy_number_variation 9188753 9995409 . . . ID=nssv27813;Name=ns

    Another copy number variant as above.  In this example the inner
    boundaries of the ranges are given with nucleotide resolution, but
    the outer boundaries are unknown.

NC_000001.9 dbVar copy_number_variation 213446726 220244033 . . . ID=nsv491682;Nat

    A complex example where a t(8;21)(q22;q22.3) translocation, an
    inversion and an SNV overlap.  These are actually three separate
    variants each mapping to a common region of the reference genome
    and thus are described as three separated records in the GVF file
    - each relative to it's location on the reference genome.

chr21   BreakDancer     translocation   41400000        46944323        .       +
chr21   DGV     inversion       42061144        42083169        .       +       .
chr21   samtools        SNV     42071394        42071394        .       +       .

    The two SNVs below are phased.  In this case the individual is a
    compound heterozygote with the Variant_seq G on one copy of the
    chromosome (first value given in the Variant_seq attribute) and
    the other Variant_seq A on the other copy of the chromosome
    (second value given in the Variant_seq attribute).

chr1    GATK    SNV     15883   15883   .       +       .       ID=SNV001;Referenc
chr1    GATK    SNV     15765   15765   .       +       .       ID=SNV002;Referenc
```

Pragmas

```
GFF3 allows for pragmas that define file-wide directives to processing
software.  Pragmas should be given at the top of a GVF file before any
feature lines are given.  All pragmas from GFF3 are included in GVF.
Note especially the following GFF3 pragmas which are required for GVF:

##gvf-version
##sequence-region
##feature-ontology
##genome-build

The following describes a set of tag-value pairs for use with GVF
specific pragmas.
```

Pragma Tag-Value Pairs

```
The pragmas described by GVF may refer to the entire file, or they may
limit their scope by use of tag-value pairs for any combination of the
following tags: Seqid, Source, Type.  For example if a pragma only
applies to SNVs, insertions and deletions that were called by
Gigabayes on chromosome 13 then the following:

Seqid=chr13;Source=Gigabayes;Type=SNV,insertion,deletion;

would indicate the scope for the given pragma.  Omitting any
scope-limiting tag implies that the scope of the pragma includes all
values for that tag.

The Dbxref tag within a GVF pragma takes values of the form "DBTAG:ID"
as described below in the Database Cross References section. The
Dbxref tag provides a reference for the information given by the
pragma whether that be the location of sequence files or a link to a
paper describing a method.

The Comment tag is allowed for all pragmas described here to provide a
more human readable description.

Tags beginning with uppercase letters are reserved for future use
```

within the GVF specifications. Applications are free to
provide additional tags beginning with lower case letters.

Pragma Descriptions

##gvf-version

    The version of the GVF specification that this file conforms to.

```
##gvf-version 1.05
```

##file-version

    The file-version pragma allows specification of the version of a
    file. The tag is provided for the case when an individual's
    variants are described in GVF and then at a later date changes to
    the data or the software require an update to the file. An
    increment of the file-version could signify such a change. The
    format and interpretation of the version is left to the
    application, although a scheme of an incrementing decimal number
    where the fractional part of the number represents minor changes
    to formatting and documentation and changes to the integral part
    of the number represent actual changes to the variant data.

    This is version 1.0 of data represented in this file.

```
##file-version 1.0
```

##file-date

    The file-date pragma is included as a method to describe the date
    when the file was created. The ISO 8601 standard for dates in the
    form YYYY-MM-DD is required for the value.

    The data in this file was created/updated on Feb. 8th 2010.

```
##file-date Date=2010-02-08;Comment=Variants converted to GVF on Feb. 8, 2010;
```

##individual-id

    This pragma provides details about the individual whose variants
    are described in the file. Tags: Dbxref, Gender, Population,
    Comment. Note that the Gender tag is a convenience for the human
    reader as this can be stated explicitly in the phenotype-description
    pragma described below. Tags: Seqid, Source, Type, Dbxref,
    Comment, Display_name.

    A male of Yoruba origin whose ID in the Coriell database is
    NA18507

```
##individual-id Dbxref=Coriell:NA18507;Gender=male;Population=Yoruba;Comment=Yorub
```

##score-method

    This pragma provides details about the algorithms or methodologies
    used to generate the score of a feature. A typical use would be to
    provide a Dbxref link to a journal article or website describing
    the software or algorithm used to calculate the score. Tags:
    Seqid, Source, Type, Dbxref, Comment.

```
##score-method Comment=Scores are Phred scaled probabilities of an incorrect varia
```

##source-method

    This pragma provides details about the algorithms or methodologies
    used to generate data for a given source in the file. This is
    used for example to document how a particular type of variant was
    called. A typical use would be to provide a Dbxref link to a
    journal article describing software used for calling the variant
    data with the given source tag. Tags: Seqid, Source, Type,
    Dbxref, Comment.

Features on chromosome 1 whose source is MAQ and type is SNV had
SNVs called by MAQ which is referenced with PubMed ID 18714091.

```
##source-method Seqid=chr1;Source=MAQ;Type=SNV;Dbxref=PMID:18714091;Comment=MAQ SN
```

Features on all chromosomes with source SOAP and type SNV were
called by SOAPsnp which is referenced with PubMed ID 18227114.

```
##source-method Source=SOAP;Type=SNV;Dbxref=PMID:18227114,PMID:18987735;Comment=Sh
```

##attribute-method

This pragma provides details about algorithms or methodologies for
a given attribute tag in the file. This is used to document how a
particular type of attribute value (i.e. Genotype,
Variant_effect) was calculated. Tags: Seqid, Source, Type,
Attribute, Dbxref, Comment.

The value for the Genotype attributes for all features with a
source SOLiD and a type SNV were passed through into the GVF file
'as-is' from the original data file. A Dbxref could have been
included as well to reference the paper of the original study.

```
##attribute-method Source=SOLiD;Type=SNV;Attribute=Genotype;Comment=Genotype is re
```

##technology-platform

This pragma provides details about the technologies
(i.e. sequencing or micro array) used to generate the primary
data. Tags: Seqid, Source, Type, Dbxref, Comment

For sequencing technologies:
Read_length, Read_type, Read_pair_span, Platform_class,
Platform_name, Average_coverage.
For micro array technologies:
???

All features with source SOAP and type SNV were called from data
that was generated by an Illumina GA short read sequencer. There
was both a single read (fragment) library and two paired-end
libraries. The read lengths for both libraries is 35 base pairs,
and the paired-end libraries had an average length of 135 and 440
From beginning of the first read to the end of the second. The
average depth of coverage for sequencing from all libraries
combined was 36x haploid coverage against the reference genome.

```
##technology-platform Source=SOAP;Type=SNV;Dbxref=URI:http://www.illumina.com;Plat
```

All features with source AFFY_SNP_6 and type SNV are called from
data that was generated by SNP Array genotyping with an Affymetrix
Human SNP Array 6.0.

```
##technology-platform Seqid=chr1;Source=AFFY_SNP_6;Type=SNV;Dbxref=URI:http://www.
```

##data-source

This pragma provides details about the source data for the
variants contained in this file. This could be links to the
actual sequence reads in a trace archive, or links to a variant
file in another format that have been converted to GVF. Tags:
Seqid, Source, Type, Dbxref, Data_type, Comment.

The data for all features with source MAQ and type SNV are based
on DNA sequence data from NCBI Short Read Archive ID SRA008175.

```
##data-source Source=MAQ;Type=SNV;Dbxref=SRA:SRA008175;Data_type=DNA sequence;Comm
```

The data for all features with source Crossbow and type SNV were
converted to GVF format from variant information that can be
downloaded from the given URI.

```
##data-source Source=MAQ;Type=SNV;Dbxref=URI:ftp://ftp.kobic.kr/pub/KOBIC-KoreanGe
```

```
##phenotype-description

    A description of the phenotype of the individual. This pragma can
    contain either ontology constrained terms, or a free text
    description of the individual's phenotype or both.  In the first
    case the constraining ontology is given with the Ontology tag which
    is a URI pointing to an ontology file.  The IDs or terms given
    with the Term tag are then required to be in the given ontology.
    A free text description of the phenotype may be given in the
    Comment tag to supplement or replace the ontology description.
    Tags: Ontology, Term, Comment.

    A text description of a 50 year old individual with AML.
##phenotype-description Comment=Individual presented at 50 years with AML;


    An ontology description of an individual with AML.
##phenotype-description Term=acute myloid leukemia;Ontology=http://www.human-phenc


    A HPO term used to describe an individual with AML.
##phenotype-description Term=HPO:0004808;Ontology=http://www.human-phenotype-ontol


    A mature sterile fly

##phenotype-description Term=sterile,mature;Ontology=http://obo.cvs.sourceforge.ne


##phased-genotypes

    This pragma indicates that the genotypes in the file are phased.
    Tag-value pairs can be used to limit the scope of the pragma.  For
    example, you may want the pragma to define only SNV features or
    only features with a given source value as phased.  Phased
    genotypes indicate that for heterozygous loci it is known which
    chromosome each variant sequence belongs to.  That is, if one
    variant has a Variant_seq=A,C attribute and another variant on the
    same landmark feature (chromosome) has a Variant_seq=T,G attribute
    then the A and T occur together on one copy of the chromosome and
    the C and G occur together on the other copy of that chromosome.
    This information can come from pedigree data, population data or
    complete or partial chromosome assembly.  When the
    '##phased-genotypes' pragma is given for a file, the sequences
    given in the Variant_seq attributes for all features (except as
    limited by the tag-value pairs described above) are required to be
    ordered.  That is, the first sequence given is on the same copy of
    the chromosome as the first sequence given in all other
    Variant_seq tags covered and the second sequence given is always
    on the other copy of the chromosome.  Note that you can use the
    Phased attribute (see below) to indicate that individual features
    are phased.  The phased attribute would be used when only regions
    of the genome are phased.

    All heterozygous SNV features are phased in this file.  That is,
    if one SNV has Variant_seq=A,C and another SNV has Variant_seq=T,G
    then the A and T are on the same chromosome and the C and G are on
    the other.

##phased-genotypes Type=SNV;
```

Comments

```
Other less structured details can be included in the form of
comments. And since they are for human reading only they can scroll
over multiple lines with a comment identifier '#' at the beginning of
each line.

# laboratory-description
# This is individual X035 that was originally mislabeled as X053.  As of 03/08/10
```

Database Cross References

```
For all tags in GVF that have a Dbxref tag, the value is in the form
database:ID.  For example an SNV might have a cross reference to
dbSNP as 'Dbxref=dbSNP:rs113993958,OMIM:602421.0004'.  The
format of each type of ID varies from database to database. An
authoritative list of databases, their DBTAGs, and the URI
transformation rules that can be used to fetch the objects given their
IDs can be found at this location:

ftp://ftp.geneontology.org/pub/go/doc/GO.xrf_abbs.

Further details can be found here:

ftp://ftp.geneontology.org/pub/go/doc/GO.xrf_abbs_spec

In addition, database:ID pairs can point to a stable URN, URL or URI
with Dbxref=URL:http//www.example.com for example.
```

Acknowledgments

We would like to thank the NHGRI for funding this work (R44HG2991, R44HG3667).

```
We would like to
```

Change Log

```
1.05 Wed Jan 19 16:26:14 MST 2011
    -Modified the description and examples for the Start_range and
     End_range tag slightly as per discussions with NCBI, Ensembl and
     UCSC.
    -Modified the wording under Community Attributes Collection
     slightly.
    -Several typo fixes that were caught by Bob Kuhn and John Lopez

1.04 Tue Dec 14 14:51:41 MST 2010
    -Added Display_id tag to the Individual-id pragma.

1.03 Wed Nov 24 10:45:18 MST 2010
    -Specified constraint on column 3 to be a sequence_alteration, one
     of it's children or gap.
    -Added ##phased-genotype pragma and example.
    -Added Start_range and End_range attributes and examples.
    -Cleaned up the 'Quick Example of GVF Content' section to make it
     simpler.
    -Replaced Nomenclature attributes with Alias attributes in the
     'Quick Example of GVF Content' section.
    -Added the ##score-method pragma as a method to describe the score
     calculated in column 6.
    -Added links to the GVF wiki pages.

1.02 Sat Jul 24 08:17:39 MDT 2010
    -Updated format of Variant_effect tag and updated the examples to
     comply.
    -Minor updates to the homozygous deletion example.
    -Added link to the 10Gen Data set.

1.01 Fri Mar 19 16:56:20 MDT 2010
     - Added documentation to, and fixed inconsistencies with pragma
       examples.
     - Added hemizygous as a valid Genotype tag value.
     - Removed Intersected_feature tag.
     - Modified Variant_effect tag to incorporate the function of the
       Intersected_feature tag.
     - Clarified the definition of the Variant_copy_number and
       Reference_copy_number tags.
     - Modified the GVF examples to reflect the changes in
       Variant_effect and Intersected_feature tags.
```