The Amino Acid Variant Format (AAVF) Format Version 1.0 Specification

September 5, 2018

Contents

1. The AAVF specification

AAVF is a text file format, inspired by the Variant Call Format (VCF) format. It contains meta-information lines, a header line, and then data lines each containing information about a position in a gene within a genome.

1.1 An example

```
##fileformat=AAVFv1.0
##fileDate=20180501
##source=myProgramV1.0
##reference=hxb2.fas
##INFO=<ID=RC,Number=1,Type=String,Description="Reference Codon">
##INFO=<ID=AC,Number=.,Type=String,Description="Alternate Codon">
##INFO=<ID=ACF,Number=.,Type=Float,Description="Alternate Codon Frequency, for each Alternate Codon, in
    the same order as listed.">
##FILTER=<ID=af0.01,Description="Set if True; alt_freq<0.01">
#CHROM  GENE    POS     REF    ALT     FILTER  ALT_FREQ  COVERAGE INFO
hxb2    RT      48      S      *       af0.01  0.0031    324      RC=tca;AC=tAa;ACF=0.0031
hxb2    RT      103     K      N       PASS    0.0779    154      RC=aaa;AC=aaC;ACF=0.0779
hxb2    RT      117     S      Q       af0.01  0.0033    299      RC=tca;AC=CAa;ACF=0.0033
hxb2    RT      118     V      F       af0.01  0.0065    306      RC=gtt;AC=Ttt;ACF=0.0065
hxb2    RT      174     Q      K       af0.01  0.0091    659      RC=caa;AC=Aaa;ACF=0.0091
hxb2    RT      212     W      G       af0.01  0.0044    1133     RC=tgg;AC=Ggg;ACF=0.0044
hxb2    RT      248     E      K       af0.01  0.0022    1394     RC=gaa;AC=Aaa;ACF=0.0022
```

1.2 Meta-information lines

File meta-information is included after the ## string and must be key=value pairs. It is strongly encouraged that information lines describing the INFO and FILTER entries used in the body of the AAVF file be included in the meta-information section. Although they are optional, if these lines are present then they must be completely well-formed.

1.2.1 File format

A single 'fileformat' field is always required, must be the first line in the file, and details the AAVF format version number. For example, for AAVF version 1.0, this line should read:

```
##fileformat=AAVFv1.0
```

1.2.2 Information field format

INFO fields should be described as follows (first four keys are required, source and version are recommended):

```
##INFO=<ID=ID,Number=number,Type=type,Description="description",Source="source",Version="version">
```

Possible Types for INFO fields are: Integer, Float, Flag, Character, and String. The Number entry is an integer that describes the number of values that can be included with the INFO field. For example, if the INFO field contains a single number, then this value should be 1; if the INFO field describes a pair of numbers, then this value should be 2 and so on. If the number of possible values varies, is unknown, or is unbounded, then this value should be '.'.

The 'Flag' type indicates that the INFO field does not contain a Value entry, and hence the Number should be 0 in this case. The Description value must be surrounded by double-quotes. The double-quote character can be escaped with '"' and the backslash character with '\'. Source and Version values likewise should be surrounded by double-quotes and specify the annotation source (case-insenstive, e.g. "sdrm") and exact version (e.g. "2009"), respectively for computational use.

1.2.3 Filter field format

FILTERs that have been applied to the data should be described as follows:

```
##FILTER=<ID=ID,Description="description">
```

1.3 Header line syntax

The header line names the 9 fixed, mandatory columns. These columns are as follows:

1. #CHROM
2. GENE
3. POS
4. REF
5. ALT
6. FILTER
7. ALT_FREQ
8. COVERAGE
9. INFO

1.4 Data lines

1.4.1 Fixed lines

There are 9 fixed fields per record. All data lines are tab-delimited. In all cases, missing values are specified with a dot (‘.’). Fixed fields are:

1. CHROM - chromosome: An identifier from the reference genome. All entries for a specific CHROM should form a contiguous block within the AAVF file. The colon symbol (:) must be absent from all chromosome names to avoid parsing errors when dealing with breakends. (String, no white-space permitted, Required)
2. GENE - gene: An identifier for a coding sequence within the CHROM. All entries for a specific GENE should form a contiguous block within the AAVF file. The colon symbol (:) must be absent from all chromosome names to avoid parsing errors when dealing with breakends. (String, no white-space permitted, Required)
3. POS - position: The reference position within the gene specified, with 1st amino acid in the gene having position 1. Positions are sorted numerically, in increasing order, within each GENE sequence. It is permitted to have multiple records with the same POS. (Integer, Required)
4. REF - reference amino acid(s): Each amino acid must be one of A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y,X,* (case insensitive). Multiple amino acids are permitted. The value in the POS fields refers to the position of the first amino acid in the String. For simple insertions and deletions in which either the REF or one of the ALT alleles would otherwise be null/empty, the REF and ALT Strings must include the amino acid before the event (which must be reflected in the POS field), unless the event occurs at position 1 on the contig in which case it must include the amino after the event. (String, Required)
5. ALT - alternate amino acid(s): Each amino acid must be one of A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y,X,* (case insensitive), where ‘X’ represents an ambiguous amino acid and ’*’ represents a stop amino acid. (String, Required)
6. FILTER - filter status: PASS if this position has passed all filters, i.e. a call is made at this position. Otherwise if the site has not passed all filters, a semicolon-separated list of codes for filters that fail. e.g. “af0.01” indicates that at this site the ALT_FREQ is below 0.01. ‘0’ is reserved and should not be used as filter String. If filters have been applied, then this field should be set to the missing value. (String, no white-space or semi-colons permiited)
7. ALT_FREQ - alternate amino acid frequency: Frequency of the alternate allele. (Float, Required)
8. COVERAGE - coverage at that position: Number of reads that cover the POS. (Integer, Required)
9. INFO - additional information: (String, no white-space, semi-colons, or equals-signs permitted; commas are permitted only as delimiters for list of values) INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: `<key>=<data>[,data]`. Arbitrary keys are permitted, although the following sub-fields are reserved (albeit optional):

- RC: reference codon, the codon that makes up the REF amino acid(s).
- AC: alternate codon, the codon that makes up the ALT amino acid(s).
- ACC: alternate codon count (number of reads containing that codon) for each alternate codon, in the same order as listed
- ACF: alternate codon frequency, for each alternate codon, in the same order as listed

2 Understanding the AAVF format

AAVF records use a single general system for representing genetic variation data composed of:

- Allele: representing single genetic haplotypes
- AAVF record: a record holding all the segregating alleles at a locus

AAVF records use a simple haplotype representation for REF and ALT alleles to describe variant haplotypes at a locus. ALT haplotypes are constructed from the REF haplotype by taking the REF allele amino acids at the POS in the gene within the reference genotype and replacing them with the ALT amino acids. In essence, the AAVF record specifies a-REF-t and the alternative haplotypes are a-ALT-t for each alternative allele.

3 Representing variation in AAVF records

3.1 Creating AAVF entries for Synonymous and Non-synonymous mutations

3.1.1 Example 1

For example, suppose we are looking at a locus within the **a** gene in the **my_chrom** genome:

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l K k s	gga ctc AAA aaa tcc	K is the reference amino acid
1	g l K k s	gga ctc AAG aaa tcc	K has a silent mutation w.r.t. to the reference sequence
2	g l N k s	gga ctc AAT aaa tcc	K amino acid is a N, in a portion of the viri

Representing these as AAVF records would be done as follows:

```
#CHROM    GENE    POS     REF    ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       3       K      K       PASS    0.95      1000     RC=aaa;AC=aaa,aaG;ACF=0.75,0.20
my_chrom  a       3       K      N       PASS    0.05      1000     RC=aaa;AC=aaT;ACF=0.05
```

3.2 Decoding AAVF entries for Synonymous and Non-synonymous mutations

3.2.1 Synonymous mutation AAVF record

Suppose I received the following AAVF record:

```
#CHROM    GENE    POS     REF    ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       2       L      L       PASS    1.0       1000     RC=ctc;AC=ctc,ctT;ACF=0.75,0.25
```

This is a synonymous mutation since the alt amino acid is the same as the reference amino acid, and the 'AC' INFO field contains a codon which is difference from the reference codon, so I have the two following haplotypes:

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g L k k s	gga CTC aaa aaa tcc	L is the reference amino acid
1	g L K k s	gga CTT aaa aaa tcc	L has a silent mutation w.r.t. to the reference sequence

3.2.3 Non-synonymous mutation AAVF record

Suppose I received the following AAVF record:

```
#CHROM    GENE    POS     REF     ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       4       K       I       PASS    0.75      1000     RC=aaa;AC=aTa;ACF=0.75
```

This is a non-synonymous mutation since the alt amino acid differs from the reference amino acid, so I have the two following haplotypes:

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l k K s	gga ctc aaa AAA tcc	K is the reference amino acid
1	g l k I s	gga ctc aaa ATA tcc	K amino acid is a I, in a portion of the virus population

3.3 Creating AAVF entries for Insertions and Deletions

3.3.1 Example 1

For example, suppose we are looking at a locus with the **a** gene in the **my_chrom** genome:

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l K k s	gga ctc AAA aaa tcc	K is the reference amino acid
1	g l - k s	gga ctc — aaa tcc	K amino acid is deleted w.r.t. to the reference sequence
2	g l KKk s	gga ctc AAAAAA aaa tcc	K amino acid is inserted w.r.t. to the reference sequence

Representing these as AAVF records would be done as follows:

1. A single amino acid deletion of K at position 3 becomes REF=LK, ALT=L
2. A single amino acid insertion of K after position 3 becomes REF=K, ALT=KK

Note: that the positions must be sorted in increasing order:

```
#CHROM    GENE    POS     REF     ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       2       LK      L       PASS    0.5       1000     RC=ctcaaa;AC=ctc
my_chrom  a       3       K       KK      PASS    0.5       1000     RC=aaa;AC=aaaaaa;ACF=0.5
```

3.4 Decoding AAVF entries for Insertions and Deletions

3.4.1 Insertion AAVF record

Supposed I receive the following AAVF record:

```
#CHROM    GENE    POS     REF    ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       3       K      KK      PASS    0.5       1000     RC=aaa;AC=aaaaaa;ACF=0.5
```

This is an insertion since the reference amino acid K is being replaced by K [the reference amino acid] plus one insertion amino acid K in such a way that a gap is opened in the reference. Again there are only two alleles so I have the two following segrating haplotypes.

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l K k s	gga ctc AAA — aaa tcc	K is the reference amino acid
1	g l KKk s	gga ctc AAA AAA aaa tcc	K amino acid is inserted w.r.t. to the reference sequence

3.4.2 Deletion AAVF record

Supposed I receive the following AAVF record:

```
#CHROM    GENE    POS     REF    ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       2       LK     L       PASS    0.5       1000     RC=ctcaaa;AC=ctc
```

This is a deletion of one reference amino acid since the reference allele LK is being replaced by just the L [the reference amino acid]. Again there are only two alleles so I have the two following seggregating haplotypes.

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l K k s	gga ctc AAA aaa tcc	K is the reference amino acid
1	g l - k s	gga ctc — aaa tcc	K amino acid is deleted w.r.t. to the reference sequence