



MAF Data File Format

vcf2maf.pl version 6

August 2012

CGA Tools, cPAL, and DNB are trademarks of Complete Genomics, Inc. in the US and certain other countries. All other trademarks are the property of their respective owners.

Disclaimer of Warranties. COMPLETE GENOMICS, INC. PROVIDES THESE DATA IN GOOD FAITH TO THE RECIPIENT "AS IS." COMPLETE GENOMICS, INC. MAKES NO REPRESENTATION OR WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTY OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE OR USE, OR ANY OTHER STATUTORY WARRANTY. COMPLETE GENOMICS, INC. ASSUMES NO LEGAL LIABILITY OR RESPONSIBILITY FOR ANY PURPOSE FOR WHICH THE DATA ARE USED.

Any permitted redistribution of the data should carry the Disclaimer of Warranties provided above.

Data file formats are expected to evolve over time. Backward compatibility of any new file format is not guaranteed.

Complete Genomics data is for Research Use Only and not for use in the treatment or diagnosis of any human subject. Information, descriptions and specifications in this publication are subject to change without notice.

Copyright © 2011-2012 Complete Genomics Incorporated. All rights reserved.

RM_MAF_2.2-01

MAF File Format

The Complete Genomics analysis pipeline for TARGET (version 2.1) natively produces variant calls and somatic status in a VCF (Variant Call Format) file. At the NCI's request, Complete Genomics has provided a script (`vcf2maf.pl`) to convert these VCF files to Mutation Annotation Format (MAF) files. This document describes the data format of the MAF files as generated by version 6 of this conversion script.

Content Description

Column Name	Description
1 TARGET_CASE_ID	This column contains the TARGET disease code and 6-character patient identifier, e.g., "10-PAEAKL".
2 Trio	This column is unpopulated by the Complete Genomics conversion script. It is intended to be used by the NCI upon merging primary tumor and relapse tumor MAF files.
3 Hugo_Symbol	If the variant overlaps the footprint of a gene (including 7.5 kpb upstream), the HUGO identifier of the gene is provided. If the variant overlaps the footprint of more than one gene, each gene is reported, separated by the pipe () symbol. The ordering of genes is maintained for the <i>Variant_Classification</i> column.
4 Variant_Classification	<p>The location or predicted impact of the variant on the gene is contained in this column. If the variant overlaps the footprint of multiple genes, the pipe symbol () is used as a delimiter, preserving the order stated in the <i>Hugo_symbol</i> column. If the variant has different impacts on the isoforms of a single gene, the list of possible impacts is presented as a comma separated list.</p> <p>The allowed values are:</p> <ul style="list-style-type: none"> ▪ INTRON: Region of nucleotides within a gene that is removed before translation of mRNA. ▪ DONOR or ACCEPTOR: Indicates that the variation falls inside the 6 bases of the splice donor site or the 15 bases of the splice acceptor site, but a clear case of splice disruption or repair is not observed (see DISRUPT and NO-CHANGE). ▪ TSS-UPSTREAM: Indicates that the variation falls within the 7.5 kb region upstream of 5' transcription start site of a gene. ▪ SPAN5, SPAN3, or SPAN: SPAN5 and SPAN3 indicate that the variation overlaps an exon and another component, such as, ACCEPTOR and CDS, or TSS-UPSTREAM and UTR5. SPAN5 indicates that the 5' end of the exon is one of the components. SPAN3 indicates that the 3' end of the exon is one of the components. SPAN indicates that the variation overlaps an entire exon. ▪ UTR5, UTR3, or UTR: Indicates that the variation falls inside the 5' untranslated region (UTR5) or 3' untranslated region (UTR3) of protein coding genes, or genes with no known coding region (UTR). ▪ NO-CHANGE: The sequence of this allele is identical to the canonical transcript sequence (which may or may not be identical to the reference sequence used in the assembly). Also, non-GT/AG conserved splice site sequences or AT/AC rare splice site sequences become canonical sequences. ▪ SYNONYMOUS: The DNA sequence for this transcript has changed, but there is no change in the protein sequence: the altered codon codes for the same amino acid. ▪ MISSENSE: The DNA sequence for this transcript has changed and there is a change in the protein sequence as well, since the codon codes for a different amino acid. There is no change in size of the protein. ▪ NONSENSE: The DNA sequence for this transcript has changed and has resulted in a STOP codon (TGA, TAG, or TAA), resulting in an early termination of the protein translation. ▪ NONSTOP: The DNA sequence for this transcript has changed and has

Column Name	Description
	<p>resulted in the change of a STOP codon (TGA, TAG, or TAA) into a codon that codes for an amino acid, likely resulting in the continuation of the translation for this protein.</p> <ul style="list-style-type: none"> ▪ DELETE: The DNA sequence for this transcript has changed and the length of the deletion is a multiple of 3, resulting in deletion of amino acids in the sequence in-frame, with no neighboring amino acids modified ▪ INSERT: The DNA sequence for this transcript has changed and the length of the insertion is a multiple of 3, resulting in the insertion of amino acids in the sequence in-frame, with no neighboring amino acids modified. ▪ DELETE+: The DNA sequence for this transcript has changed and the length of the deletion is a multiple of 3, occurs out of frame, and results in the deletion of amino acid(s) with possible modification of one or both of the neighboring codons. ▪ INSERT+: The DNA sequence for this transcript has changed and the length of the insertion is a multiple of 3, occurs out of frame, and results in the insertion of amino acid(s) with possible modification of one or both of the neighboring codons. ▪ FRAMESHIFT: The DNA sequence for this transcript has changed and has resulted in a frameshift for this protein. ▪ MISSTART: The DNA sequence for this transcript has changed and resulted in the change of a START codon into a codon that codes for something other than a start codon, likely resulting in a non-functional gene. ▪ DISRUPT: GT or AG conserved donor and acceptor splice site sequence has changed to something that is incompatible. Also used if rare AT/AC sequence has changed to something that is incompatible.
5 VariantType	One of SNP, Ins(ertion), Del(etion), or Sub(stitution, or multiple nucleotide polymorphism.)
6 dbSNP_RS	Identifier for this variant in NCBI's dbSNP, if the variant is present in that database. The format for this column is dbSNP: dbsnp.<build>:<rsID> with multiple entries separated by commas. <build> indicates in which build of dbSNP this entry first appeared. For example, "dbsnp.129:rs12345".
7 Mutation_Status	One of Somatic, Germline, LOH, or Unknown.
8 PFAM_DOMAIN	Pfam identifier and domain name of the locus in which this variation falls. Format: PFAM:<identifier>:<domain name> For example, "PFAM:PF00069:Pkinase".
9 Somatic_Score	An integer that provides a total order on quality for all somatic mutations. It is equal to $-10 \cdot \log_{10}(P(\text{false})/P(\text{true}))$ under the assumption that this genome has a rate of somatic mutation equal to 1/Mb for SNVs, 1/10Mb for insertions, 1/10Mb for deletions, and 1/20Mb for substitutions. This field is empty for mutations that are not somatic. The calculation of this score is based on CGA™ Tools calldiff somatic score, using default parameters and not using the <code>-diploid</code> option. As such, it is based on a calibration of <i>varScoreVAF</i> and a mixture model where we assume half of variants are present with 20% allele fraction and half have 50% allele fraction.
10 Somatic_Rank	The estimated rank of this somatic mutation, amongst all true somatic mutations within a given <i>somaticCategory</i> . Value is a number between 0 and 1; a value of 0.012 means, for example, that 1.2% of the true somatic mutations in this <i>somaticCategory</i> have a <i>somaticScore</i> less than the <i>somaticScore</i> for this mutation. This field is empty for mutations that are not somatic.
11 Somatic_quality	Equal to SQHIGH for somatic variants where <i>somaticScore</i> ≥ -10 . Otherwise, this field is empty.
12 Tumor_ReadCount_Alt	The number of reads supporting the alternate allele in the tumor sample.
13 Tumor_ReadCount_Ref	The number of reads supporting the reference allele in the tumor sample.

Column Name	Description
14 Tumor_ReadCount_Total	Total number of reads that overlap the variant interval in the tumor sample. Note that this count also includes reads that do not strongly support one allele over the other and consequently are not accounted for in <i>Tumor_ReadCount_Alt</i> or <i>Tumor_ReadCount_Ref</i> . For loci where one of the alleles contains a no-call, this column also includes the reads that support that no-called allele. This column does not include reads that do not overlap the locus.
15 Normal_ReadCount_Alt	The number of reads supporting the alternate allele in the normal sample.
16 Normal_ReadCount_Ref	The number of reads supporting the reference allele in the normal sample.
17 Normal_ReadCount_Total	Total number of reads that overlap the variant interval in the normal sample. Note that this count also includes reads that do not strongly support one allele over the other and consequently are not accounted for in <i>Normal_ReadCount_Alt</i> or <i>Normal_ReadCount_Ref</i> . For loci where one of the alleles contains a no-call, this column also includes the reads that support that no-called allele. This column does not include reads that do not overlap the locus.
18 Cosmic	Identifier for this variant in COSMIC, if the variant is present in that database. The format for this column is <code>COSMIC.<type>:identifier</code> with multiple entries separated by the semicolon (;). <code><type></code> indicates COSMIC classification of somatic variants. For example for a non-coding variant, xRef would contain "COSMIC:ncv_id:139111".
19 Cosmic_Gene	Populated with "Cosmic_Gene" when overlapping a gene with mutations in COSMIC, whether or not this specific variant is contained in COSMIC.
20 Reference_Allele	The bases of the reference at the positions described by <i>Chromosome</i> , <i>Start_position</i> , and <i>End_position</i> , per the MAF specification.
21 TumorSeq_Allele1	Sequence of the first allele in the tumor. May contain the "?" character indicating a no-call. The field is empty when the called variant is a deletion of all bases in the locus.
22 TumorSeq_Allele2	Sequence of the second allele in the tumor. May contain the "?" character indicating a no-call. The field is empty when the called variant is a deletion of all bases in the locus.
23 Match_Norm_Seq_Allele1	Sequence of the first allele in the matched normal. May contain the "?" character indicating a no-call. The field is empty when the called variant is a deletion of all bases in the locus.
24 Match_Norm_Seq_Allele2	Sequence of the second allele in the matched normal. May contain the "?" character indicating a no-call. The field is empty when the called variant is a deletion of all bases in the locus.
25 Tumor_Sample_Barcode	Full TARGET identifier of the tumor sample.
26 Match_Normal_Sample_Barcode	Full TARGET identifier of the matched normal sample.
27 Entrez_Gene_Id	Numeric accession of this gene in the Entrez Gene database.
28 Chromosome	This is the chromosome, e.g., 1, 2, 3... 22, X, Y, M, as described in the MAF documentation.
29 Start_position	The <i>Start_position</i> is 1-based, and the <i>End_position</i> is 1-based inclusive, as per the MAF specification. For insertions we have followed the GFF v3 spec from http://www.sequenceontology.org/gff3.shtml . Thus, for insertions, both <i>Start_position</i> and <i>End_position</i> are the 1-based coordinate of the base preceding the insert. An important consequence of the conventions for <i>Start_position</i> and <i>End_position</i> is that the <i>Chromosome</i> , <i>Start_position</i> , and <i>End_position</i> do not fully determine the set of reference bases modified in a given variant. For example, a SNP at (1-based) position 100 and an insert following position 100 are both described with the same <i>Start_position</i> and <i>End_position</i> . The <i>Variation_Type</i> column distinguishes between the two.
30 End_position	

Column Name	Description
31 miRNA	miRBase identifier for any miRNAs the variant locus overlaps.
32 Verification_Status	The result of the verification experiment. One of: Somatic, BadAssay, TumorFN (only for LOH targets), TumorFP, NormalFN, NormalFP (only for LOH targets), LOH, OtherVar (variant other than the one called by WGS seen), or . (Unknown)
33 Verification_Method	Orthogonal technology used for verification. To date, either Illumina or Sanger. This field is only populated for loci selected for verification.
34 FET_Score	A statistical test—Fisher’s Exact Test (FET)—is used to determine the likelihood that the variant is somatic given the observed read counts in the tumor and normal samples. The test’s <i>p</i> -value is converted to a Phred-scaled score using the following formula: $-10 * \log_{10}(p\text{-value not somatic})$
35 TumorRefCount_VS	The number of sequencing reads observed supporting the reference allele in the tumor sample in the verification experiment. This field is only populated for loci selected for verification.
36 TumorVarCount_VS	The number of sequencing reads observed supporting the alternate allele in the tumor sample in the verification experiment. This field is only populated for loci selected for verification.
37 TumorTotalCount_VS	The total number of sequencing reads confidently mapped to the variation locus in the tumor sample in the verification experiment. This may include reads that did not support reference or the WGS called alternate allele. This field is only populated for loci selected for verification.
38 NormalRefCount_VS	The number of sequencing reads observed supporting the reference allele in the normal sample in the verification experiment. This field is only populated for loci selected for verification.
39 NormalVarCount_VS	The number of sequencing reads observed supporting the alternate allele in the normal sample in the verification experiment. This field is only populated for loci selected for verification.
40 NormalTotalCount_VS	The total number of sequencing reads confidently mapped to the variation locus in the normal sample in the verification experiment. This may include reads that did not support reference or the WGS called alternate allele. This field is only populated for loci selected for verification.