SeattleSeq annotation file column details

Columns:

1. inDBSNPOrNot: whether the variation is in the dbSNP database
2. chromosome: chromosome number
3. position: location on the chromosome, 1-based
4. referenceBase: input from the user or calculated if not provided
5. sampleGenotype: input from the user
6. sampleAlleles: same information as sampleGenotype, with the ambiguity code (see Table 1 at the end of this document resolved into alleles; if there are multiple individuals, this is a list of all alleles present
7. allelesDBSNP: list of alleles for all populations and individuals in dbSNP, derived from the HGVS notation
8. accession: NCBI transcript identifier
9. functionGVS: GVS class of variation function, using reference genome and your submitted alleles (Table 2)
10. functionDBSNP: dbSNP class of variation function
11. rsID: dbSNP identifier for the variation, 0 if not in dbSNP
12. aminoAcids: list of amino acids for the codon, starting with that of the reference base, coding SNVs only
13. proteinPosition: the position of the amino acid in the protein, beginning at the N-terminal with the first amino acid at position 1, followed by the total number of amino acids in the protein; the total includes a count for the stop codon
14. cDNAPosition: the location within the sequence of coding bases; NA if not coding; so far only SNVs
15. polyPhen: amino acid substitution impacts from PolyPhen-2. The score is a number between 0 and 1, where 1 is the most damaging.
16. GranthamScore: the Grantham score of any amino acid changes (Grantham, 1974). Ranges from 5 to 215 with the higher scores representing more radical changes in amino acid properties
17. consScoreGERP: rejected-substitution score from the program GERP, Stanford University, range of -12.3 to 6.17, with 6.17 being the most conserved (Davydov et al., 2010)
18. scoreCADD: phred-like Combined Annotation Dependent Depletion scores from Kircher et al., University of Washington, range 0 though 99) (Kircher et al, 2014)
19. chimpAllele: from UCSC alignments
20. geneList: HUGO names, any for which the transcription region overlaps the variation
21. dbSNPValidation: dbSNP validation status codes, dealing with e.g. whether the variation has been seen at least twice
22. repeatMasker: for identfitying repeats
23. tandemRepeat: for identfitying repeats
24. clinicalAssociation: links to NCBI pages and PubMed

25. distanceToSplice: how close the variation is to a donor or acceptor splice site
26. microRNAs: EMBL IDs of any microRNAs in which the variation is found
27. keggPathway: the Kyoto Encyclopedia of Genes and Genomes biological pathways for the gene
28. cpgIslands: whether in a region where CpGs are present at a high level, from the UCSC genome annotation database
29. genomesESP: the allele counts observed in the Exome Sequencing Project, optionally split by two ancestries
30. genomesExAC: the allele counts of the Exome Aggregation Consortium, optionally split by 7 populations
31. PPI: protein-protein interactions with experimental confidence scores from the STRING 9.05 database)
32. proteinSequence: NCBI protein accession ID

Table 1-Amiguity code for nucleotide bases

UPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Table 2-FunctionGVS annotations

functionGVS	**description**
missense	Leading to an amino acid change
5-prime-UTR	Affects 5'-untranslated region
intron	Affects intron
intergenic	Affects region between genes
non-coding-exon	Non-coding exon annotation involves the identification and annotation of genetic variants within exonic regions of a gene that do not contribute to the coding sequence of the protein

upstream-gene	Upstream-gene annotation involves the identification and annotation of genetic variants located in the upstream region of a gene. Outside transcribed regions, but within 5000 bp of a transcription region
synonymous	Leading to no amino acid change
synonymous near splice	Leading to no amino acid change but possibly affecting splice site
3-prime-UTR	Affects 3'-untranslated region
downstream-gene	Genetic variants located in the downstream region of a gene. Outside transcribed regions, but within 5000 bp of a transcription region
splice acceptor	Affects splicing acceptor site, i.e., any of the two bases at the 3' end of an intron (splice-acceptor)
splice donor	Affects splicing donor site, i.e., any of the two bases at the 5' end of an intron (splice-donor)
stop-gained	Leading to gain of a stop codon
stop-lost	Leading to a loss of a stop codon

References:

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010 Dec 2;6(12):e1001025. doi: 10.1371/journal.pcbi.1001025. PMID: 21152010; PMCID: PMC2996323.

Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. *Science*, *185*(4154), 862–864. http://www.jstor.org/stable/1739007

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310-5. doi: 10.1038/ng.2892. Epub 2014 Feb 2. PMID: 24487276; PMCID: PMC3992975.