**Variant Classification**

**Author:** Mike Thiesen, Golden Helix, Inc.

## Overview

Sequencing pipelines are able to identify rare variants not found in catalogs such as dbSNP. As a result, variants in these datasets often lack useful gene-based information that is present in most micro-array based marker-maps. Variant classification examines the interactions between variants and gene transcripts in order to classify variants based on their potential effect on genes. This provides insight into which variants are most likely to have functional effects.

**Note**: Two different tables are used to identify codon amino acids, for all variants except for those in the mitochondrial (MT) "chromosome" the standard codon amino acid table is used. Otherwise, the MT specific table is used, see: https://www.mun.ca/biology/scarr/MGA2-03-28_mtDNA_code.jpg

## Recommended Directory Location

Save the script to the following directory:

*..**\Application Data\Golden Helix SVS\UserScripts\Spreadsheet\DNA_Seq**

**Note:** The **Application Data** folder is a hidden folder on Windows operating systems and its location varies between XP and Vista. The easiest way to locate this directory on your computer is to open SVS and go to **Tools > Open Folder > User Scripts Folder**. If saved to the proper folder, this script will be accessible from the spreadsheet's **DNA-Seq**.

## Using the Script

From a marker mapped spreadsheet go to **DNA-Seq > Variant Classification.**

### Classification

Golden Helix SVS examines the interaction of every variant with all nearby transcripts. Each interaction is classified based on the variant's position relative to the transcript. The classifications are:

| Classification | Priority | Description |
|---|---|---|
| **Coding** | 7 | Variant is in the coding exonic region of a protein coding transcript. |
| **Splicing** | 7 | Variant affects a nucleotide that is in a splicing region of a coding transcript. See *Splice Sites* for details. |
| **UTR5** | 6 | Variant is in an exon of a coding transcript but is on the 5' side of the start codon. |
| **UTR3** | 6 | Variant is in an exon of a coding transcript but is on the 3' side of the stop codon. |

| Classification | Priority | Description |
|---|---|---|
| **NCSplicing** | 5 | Variant affects a nucleotide that is in a splicing region of a non-coding transcript. See *Splice Sites* for details. |
| **NCExonic** | 4 | Variant is in an exon for a non-coding transcript. |
| **Intronic** | 3 | Variant lies within an intron. |
| **Upstream** | 2 | Variant is within 1000 bp of the transcript start site on the 5' side. This distance can be modified with the *Upstream distance* option. |
| **Downstream** | 2 | Variant is within 1000 bp of the transcript stop site on the 3' side. This distance can be modified with the *Downstream distance* option. |
| **Intergenic** | 1 | Variant does not interact with any gene transcripts. |

**Priority**

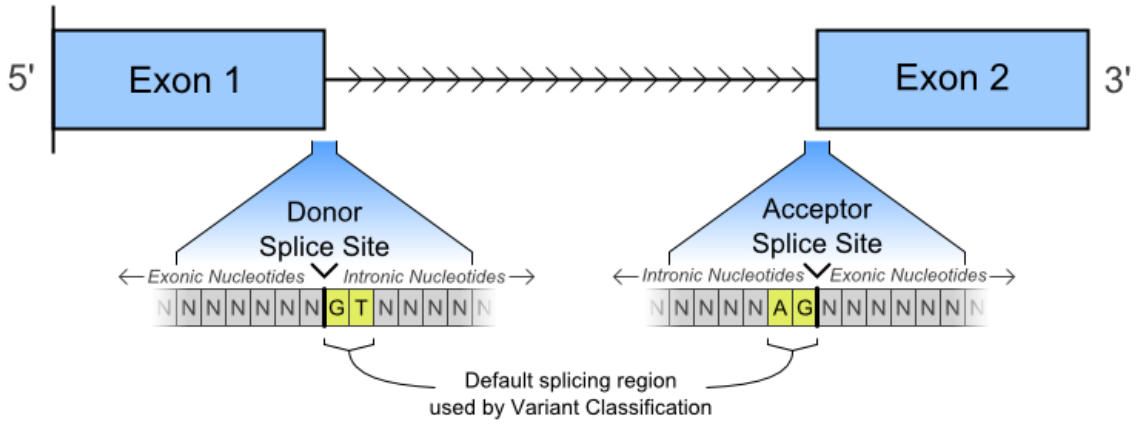
It is important to note that genes may overlap one another. Furthermore, genes often have multiple transcripts. This means that a single variant may interact with multiple transcripts and possibly multiple genes. SVS uses a priority system to decide how to classify a variant with multiple interactions. For example, if a variant lies within an intron of one transcript and within a coding region of another transcript, it will be reported as a coding variant because the "coding" classification takes a higher priority than "intronic". Reporting only high priority interactions helps to keep the report size manageable.

If a variant has more than one interaction tied for highest priority, then those interactions will be combined. For example, if a variant is downstream from one gene and upstream of another gene, it will be classified as "downstream, upstream" since both interactions are tied for highest priority.

**Splice Sites**

RNA splicing is a complicated process and is still the topic of much research. However, it is known that certain nucleotides near splice junctions are highly conserved [Sheth2006]. Thus, we assume that changes to these nucleotides are likely to affect splicing behavior. Based on this assumption, Variant classification uses a simple method to determine if a variant affects splicing: If a variant affects nucleotides in a splicing region of at least one transcript, then it is classified as "Splicing" (or "NCSplicing" for non-coding transcripts).

By default, a splicing region is defined as the two nucleotides on the intronic side of a splice junction:




A typical “GT-AG” intron, nucleotides considered part of the splicing region are highlighted in yellow.

We chose this default because these nucleotides are almost perfectly conserved at splice sites in model organisms (see [Sheth2006]). However, other nucleotides near splice sites are also highly conserved, so we added options to expand the splicing region further into the intronic and exonic side of the splice site (see *Options*).

**Protein Coding Variants**

If a variant affects nucleotides within the coding region of a transcript, then Variant Classification will reconstruct the amino acid sequence for both the reference and alternate version of the variant and predict how the change affects the protein. This is an accurate process, but it should be noted that Variant Classification can not account for post-transcriptional mRNA editing, which alters the mRNA sequence after transcription.

Splicing variants are also included in coding variant reports since they are likely to alter the protein. However, Variant classification is not able to predict exactly how (or if) a splicing variant will affect the protein sequence.

Coding variants are further classified based on the predicted effect on the amino acid sequence. The classifications are:

| Classification | Priority | Description |
|---|---|---|
| **Splicing** | 9 | Variant affects a nucleotide that is in a splicing region of a coding transcript. See *Splice Sites* for details. |
| **Init Codon** | 8 | Variant changes the start codon |
| **Frameshift Ins** | 7 | An insertion that causes a shift in the codon reading frame. |

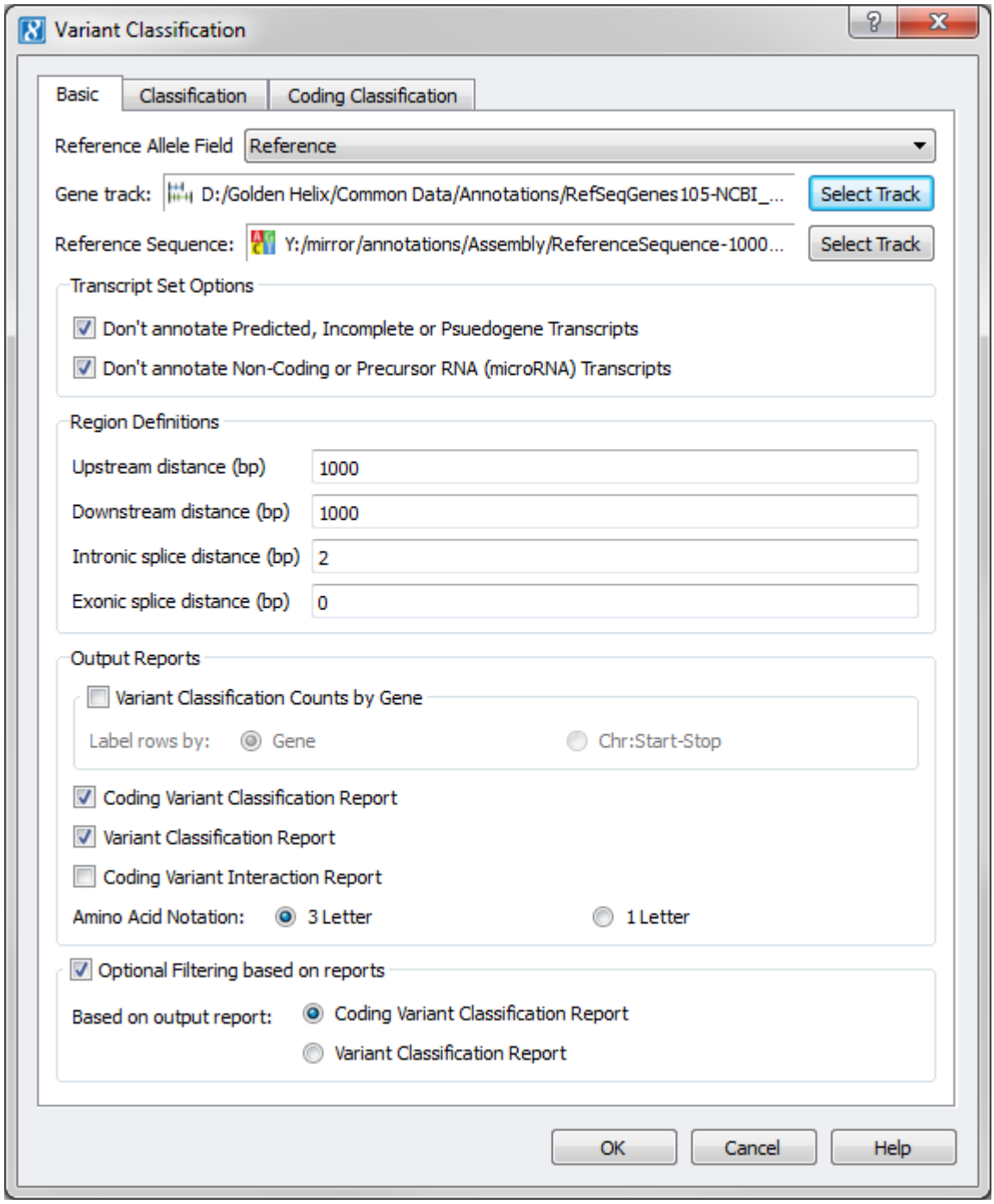
| Classification | Priority | Description |
|---|---|---|
| **Frameshift Del** | 7 | A deletion that causes a shift in the codon reading frame. |
| **Frameshift Sub** | 7 | A substitution that causes a shift in the codon reading frame. |
| **Stopgain** | 6 | Variant causes a stop codon to be created at the variant site. |
| **Stoploss** | 5 | Variant changes a stop codon to something else. |
| **Ins** | 4 | An insertion that does not cause a frameshift. |
| **Del** | 4 | A deletion that does not cause a frameshift. |
| **Sub** | 4 | A substitution that does not cause a frameshift. |
| **Nonsyn SNV** | 3 | A single nucleotide variant that changes the amino acid produced by a codon. |
| **Synonymous** | 2 | A variant affecting 1 or more nucleotides that does not change the amino acid sequence. |
| **Unknown** | 1 | A problem was found with the protein coding sequence, See *Invalid Transcripts*. |

**Input**

Variant Classification requires a column marker mapped spreadsheet with a reference allele field. The spreadsheet should contain 'AGCT' encoded genotypes.

**Options**

Options and other fields within the Variant Classification tool are described below (see *Variant Classification Window*).




**Basic Tab:**

**Reference Allele Field**

This parameter describes which marker map field contains the reference allele.

**Gene Track**

This specifies which gene definitions to use when classifying variants. For performance reasons, this must be a local annotation track. See *The Data Source Library* and

*Downloading Data* to learn more about downloading annotation tracks to a local machine.

**Reference Sequence**

This specifies the IDF track that contains the reference nucleotide sequence to use when classifying variants. For performance reasons, this must be a local annotation track. See *The Data Source Library* and *Downloading Data* to learn more about downloading annotation tracks to a local machine.

**Upstream distance**

The upstream region includes this many base pairs on the upstream side of the transcript start site. Variants that affect nucleotides in this region will be classified as Upstream.

**Downstream distance**

The downstream region includes this many base pairs on the downstream side of the transcript stop site. Variants that affect nucleotides in this region will be classified as Downstream.

**Intronic splice distance**

The splice region extends this many base pairs into the intronic side of a splice junction. Variants that affect nucleotides in this region will be classified as Splicing.

**Exonic splice distance**

The splice region extends this many base pairs into the exonic side of a splice junction. Variants that affect nucleotides in this region will be classified as Splicing.

**Output Reports**

This section allows users to choose which reports to generate.

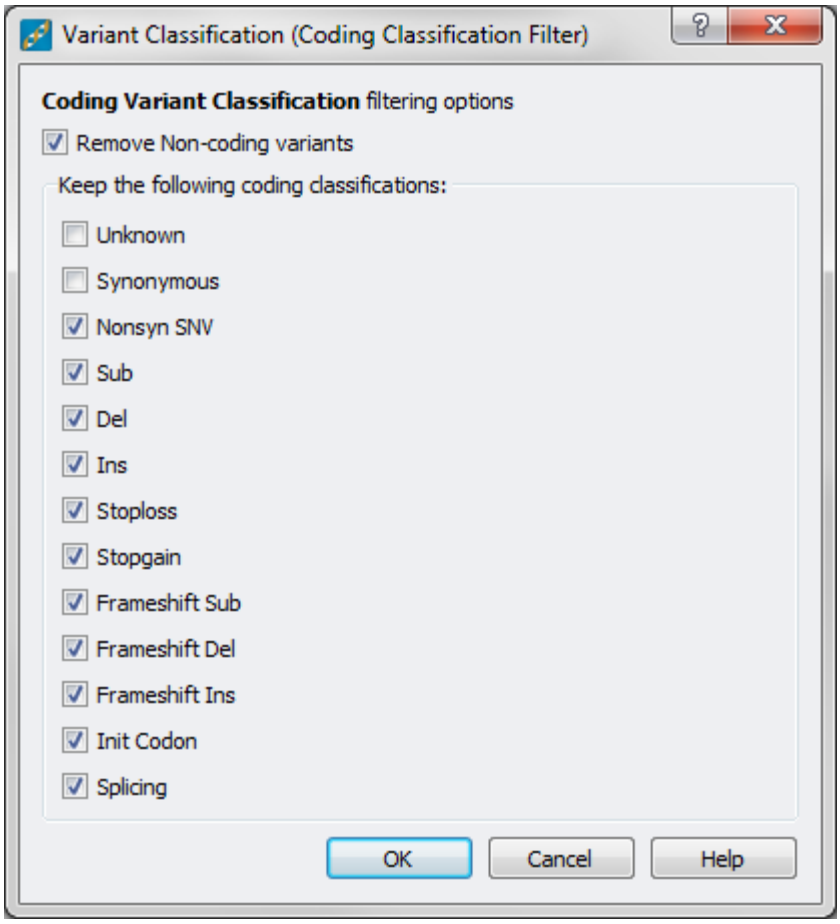
**Amino Acid Notation**

Users can choose whether to use 1 letter or 3 letter amino acid codes when producing HGVS protein notation.

**Optional Filtering based on reports**

Users can optionally select to filter the original genotype spreadsheet based on the variant classification results found on either the *Coding Variant Classification Report* or the *Variant Classification Report*.

If filtering is selected once the user clicks **OK** on the Variant Classification window a dialog will appear with the filtering options for the report selected.




**Classification Priorities Tab:**

This section allows users to customize the priorities of the various classifications. Larger numbers indicate higher priority. Unchecking a classification will cause interactions of that type to be ignored by reports. For example, one may wish to uncheck “Intergenic” upon discovering that reports are cluttered with many uninteresting intergenic variants.

**Coding Classification Priorities Tab:**

This is the same as “Classification Priorities”, except it affects the coding variant classifications.

**Output**

**Variant Classification Counts By Gene**

This report summarizes which variant classifications were found in each gene. Each row represents a gene and each column represents a variant priority level. The first three cells describe the genomic region for each gene. The rest of the cells contain the number of variants for a gene with that column’s priority. Within each gene, a variant will only be counted once. For example, if a variant has an intronic interaction with one transcript of a gene, and a coding interaction with another transcript of that same gene, it will only be counted as “coding” since coding has a higher priority.

**Variant Classification**

This report describes the classifications found for each variant. Each row describes one variant. The columns are:

**Row Labels**

These contain the column headers from the input spreadsheet.

**Classification**

The highest priority classification(s) for each variant. Multiple classifications are separated by commas, see *Priority* for details.

**Priority**

This is the highest priority level found for each variant.

**Gene(s)**

A list of genes that a variant has high priority interactions with. Genes are comma-separated and sorted alphabetically.

**Transcript(s)**

A list of transcripts that a variant has high priority interactions with. Transcripts are comma-separated and sorted alphabetically.

**HGVS Description**

The description of this variant's effect on DNA using HGVS-style notation. See *HGVS notes*.

**Coding Variant Classification**

This report provides additional details for variants in coding regions. Each row represents a variant. Variants that do not interact with coding regions are not included in this report. The first two columns are Classification and Priority. These apply to all displayed interactions.

**Row Labels**

These contain the column headers from the input spreadsheet.

**Classification**

The highest priority classification(s) for each variant. Multiple classifications are separated by commas, see *Priority* for details.

**Priority**

This is the highest coding priority level found for each variant.

The rest of the columns are arranged into groups. Group headings are followed by *n*, where *n* is the arbitrarily assigned number of the described interaction. Only high priority interactions are reported. If a variant has fewer than *n* high priority interactions, then group *n* will be filled with "?" (missing). Each group describes:

**Gene *n* :**

The gene for the *n* th interaction.

**Transcript *n* :**

The transcript for the *n* th interaction.

**Exon *n* :**

This is the exon containing the variant. Exon numbering starts at 1 and advances in the direction of gene transcription (5' to 3' on the sense strand).

**HGVS Coding *n* :**

The description of this variant's effect on DNA using HGVS-style notation. See *HGVS notes*.

**HGVS Protein *n* :**

The description of this variant's effect on protein using HGVS-style notation. See *HGVS notes*.

**Coding Variant Interaction Report**

This report describes all interactions between transcripts and coding variants, regardless of priority. This report can sometimes get very large because low priority interactions are not filtered. Each row describes one interaction, so a variant will have multiple rows if it interacts with multiple genes or transcripts. The columns are:

**Row Labels**

These contain the column headers from the input spreadsheet.

**Classification**

The classification for the current interaction.

**Gene**

The gene for the current interaction.

**Transcript**

The transcript for the current interaction.

**Exon**

This is the exon containing the variant. Exon numbering starts at 1 and advances in the direction of gene transcription (5' to 3' on the sense strand).

**HGVS Coding**

The description of this variant's effect on DNA using HGVS-style notation. See *HGVS notes*.

**HGVS Protein**

The description of this variant's effect on protein using HGVS-style notation. See *HGVS notes*.

**Invalid Variants**

This report contains a list of every variant that could not be analyzed and a message describing the problem. This report is skipped if all variants are valid.

**Invalid Transcripts**

This report contains a list of every transcript that could not be analyzed and a message describing the problem. This report is skipped if all transcripts are valid. It is fairly common for some invalid transcripts to be found in a gene annotation track. Usually this is due to one of the following:

- Errors or uncertainties that prevent the coding region from being determined.
- Errors in the reference sequence that cause unexpected frameshifts or stop codons.
- Post-transcriptional mRNA editing alters the transcript.
- Unusual start or stop codons in the coding region

**Classification Filter Applied**

If filtering is selected a new genotype spreadsheet will be created with the filter applied to the variants based on the user selected options.

**HGVS notes**

Our variant descriptions use a style similar to the HGVS but have some differences from the HGVS Nomenclature. HGVS advises that duplicating insertions be written with "dup" rather than "ins"; however, our notation always uses "ins". HGVS guidelines also state that variants should be reported at the most 3' possible position if there is any ambiguity. Our protein notation follows these guidelines. However our genomic and coding sequence notation always honors the position from the input spreadsheet's marker map (for coding variant descriptions, the marker map position is transformed into the coding sequence space).

**Tips and Tricks**

To see all of the low priority interactions that would normally be excluded, set all the priority levels to 1 under the *Classification* and *Coding Classification* tabs. This will cause all interactions to be treated as high priority (because they are all tied for highest priority). Be warned that this may produce a very large output spreadsheet. Likewise, to prevent variants from being assigned multiple classifications, assign each classification a unique priority number.