Introduction to SVS 8 Tutorial

Release 8.4.0

Golden Helix, Inc.

July 16, 2015
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Requirements</td>
<td>1</td>
</tr>
<tr>
<td><strong>Welcome Screen</strong></td>
<td>2</td>
</tr>
<tr>
<td>Log In and Register</td>
<td>2</td>
</tr>
<tr>
<td>Viewer Mode</td>
<td>3</td>
</tr>
<tr>
<td>Welcome Screen for Logged in and Activated Licenses</td>
<td>4</td>
</tr>
<tr>
<td><strong>Project Navigator</strong></td>
<td>6</td>
</tr>
<tr>
<td>Create New Project</td>
<td>6</td>
</tr>
<tr>
<td>3. Import Data Files</td>
<td>8</td>
</tr>
<tr>
<td>Import Text</td>
<td>8</td>
</tr>
<tr>
<td>Import Third Party</td>
<td>9</td>
</tr>
<tr>
<td>Drag-and-Drop DSF</td>
<td>9</td>
</tr>
<tr>
<td><strong>Combining Datasets</strong></td>
<td>11</td>
</tr>
<tr>
<td>Join/Merge Spreadsheet</td>
<td>11</td>
</tr>
<tr>
<td><strong>Spreadsheet Manipulation</strong></td>
<td>14</td>
</tr>
<tr>
<td>Change Single Column/Row State</td>
<td>14</td>
</tr>
<tr>
<td>Activation by Column Data</td>
<td>14</td>
</tr>
<tr>
<td>Other types of Activation</td>
<td>15</td>
</tr>
<tr>
<td><strong>Editing Spreadsheets</strong></td>
<td>17</td>
</tr>
<tr>
<td>Spreadsheet Editor</td>
<td>17</td>
</tr>
<tr>
<td>Recode Genotypes</td>
<td>17</td>
</tr>
<tr>
<td><strong>Filter by Gene List</strong></td>
<td>19</td>
</tr>
<tr>
<td><strong>Conclusion</strong></td>
<td>22</td>
</tr>
<tr>
<td>Congratulations!</td>
<td>22</td>
</tr>
</tbody>
</table>
Introduction

One of the greatest challenges of any genetic analysis project is the seemingly endless formatting, manipulation, and editing of data that takes place in order to properly perform analysis. This challenge is significantly compounded when the project involves whole genome data with millions to billions of data points. SNP & Variation Suite 8 (SVS 8) eliminates much of this hassle with streamlined data import of virtually any file format, as well as real-time spreadsheet manipulation and editing on a grand scale.

This tutorial seeks to guide you through basic importing and spreadsheet manipulation. The steps outlined below aim to make you a power user by providing example of functions that are often under-utilized but can make spreadsheet manipulation a breeze.

Requirements

To complete this tutorial you will need to download and unzip the following file, which includes four files.

Download
Introduction to SVS.zip

File included in the above ZIP file:

- **simulated_phenotype.csv** - comma separated text file
- **sample_data.xls** - Excel file containing more sample phenotype information
- **gene_list.txt** - text file with one column of gene names
- **Affy 6.0 HapMap Example Data.dsf** - Affymetrix Genome-Wide Human SNP Array 6.0 genotype example data in Golden Helix DSF Format.
Welcome Screen

Log In and Register

When you open SVS for the first time, the first thing you will notice is the Welcome Screen (Figure 1-1). If you already have a Golden Helix Account you can login by entering your email and password. If you do not already have an account you can create one by clicking on the Register tab. See Figure 1-1.

- (A) Enter Credentials: Enter in an existing Golden Helix User Account credentials (email and password) to log in to SVS. If your license allows, you may stay logged in by checking the Stay logged in option.

- (B) Register Tab: Click on this tab to register to create a Golden Helix User Account to obtain access to SVS Viewer or activate a full license key.

- (C) Support Contact Information: Contact us for technical support! Clicking on support@goldenhelix.com will prepopulate an email directly to the support team, all you need to do is fill in your question and hit send, or
give us a call at one of the phone numbers listed.

- (D) **Product Version Information**: Current installed version of SVS. If an update is available it will be indicated here.
- (E) **News Feed**: The news feed will contain the latest technical support bulletins and blog posts to stay informed about the newest features and latest events at Golden Helix and in regards to SVS.

**Viewer Mode**

Once you log in, if you haven’t already activated a license key, you will be in SVS Viewer mode. This mode allows you to open projects, view resulting spreadsheets, explore data in GenomeBrowse, and export results. You are free to share the SVS installer with your colleagues so they can examine projects in viewer mode. All that is required is registering and creating a Golden Helix User account.

If you have purchased SVS you may activate the key to unlock the full SVS program. See *Figure 1-2*.

- (F) **Viewer Label**: This label identifies SVS as being in the “Viewer” mode which allows access to projects to view, examine and export data from projects as well as download data from the data repository.
- (G) **Request Evaluation**: To request an evaluation of SVS to try the fully featured version of SVS, click on this link.
- (H) **License String**: The license string indicates what features are available in SVS. In viewer mode, this indicates that the version is unlicensed and provides a link to activate a license key. If you have a license key to activate, click on this link and enter the key in the dialog that opens. This can also be accessed through the Help menu, **Help > Activate a SVS License Key**.

---

**Figure 1-2 SVS Viewer Welcome Screen with section labels**
Welcome Screen for Logged in and Activated Licenses

After logging in and activating the Welcome Screen looks very similar to that of the Welcome Screen in Viewer mode with the exception of the “Viewer” label and the license information. More options are now available in the toolbar and new projects can be created and existing projects can be modified with a licensed version of SVS. See Figure 1-3.

Figure 1-3 SVS Welcome Screen with section labels

- **Toolbar and shortcut bar**: Displays several actions that are available to the user. Hovering over a shortcut icon displays the action the button represents.

**Resources Menu**
- Add-on scripts are written in Python and are used to implement new features within days.
- Webcasts may include recorded presentations from experts at Golden Helix or demonstrations of how some of our customers are using SVS.
- Our Blog is a platform we use to discuss current issues in genetic analysis or introduce new features.
- You can also find links to the tutorials

**Help Menu**
- Links to the latest version of the SVS Manual.
- Edit user profile information such as the email address, password.
- Activate a license key.
• **(J) Getting Started...**: Create a new project or follow one of the online tutorials. Plus a list of recently used projects.

For more information, please see SVS Manual
Project Navigator

Create New Project

First you will need to create a new project. Click on Create New Project from the Welcome Screen. Name the project Introduction to SVS and save it in a convenient location.

![Create New Project Dialog]

**Figure 2-1 Create New Project Dialog**

**Note:** Make sure and save the project to a directory that has the proper write permissions.

**Note:** You can also change the Default Genome Assembly in this step. By default, the genome assembly is set to the human GRCh_37 build (g1k). The data we will be using in this tutorial corresponds to this build so we will leave the default selection.

**Note:** For more details on creating and using Genome Assemblies in SVS please see the Creating and Using Genome Assemblies tutorial.

Your window should look similar to *Figure 2-1*. Click OK.

Now the main screen that you see is the Project Navigator. Currently, it does not contain any spreadsheets. Like the Welcome Screen, the Project Navigator has a toolbar and shortcut bar.
Node Change Log

The Node Change Log (Figure 2-2) keeps track of everything you do in the SVS project. Clicking on a spreadsheet will change the log to display information specific to that spreadsheet. The spreadsheet-specific log will list all actions performed on that spreadsheet as well.

You can view the complete Node Log by clicking on **Tools > View Project Log**. You can choose to view them chronologically or by each node (or spreadsheet).

User Notes and Node Tags

The User Notes Log (Figure 2-3) is located below the Node Change Log. This feature is meant for users to take project specific notes that they may want to refer to at a later date. The node is tagged with an N when a note is added in the user note log.

You may add colored tags to any node to facilitate navigation through the project by right clicking on the node and selecting one of the four available colors. A note tag with no color selected will appear on a white background.

**Note:** Your screen should look similar to Figure 2-3 after importing data files in the next step of this tutorial.
3. Import Data Files

If you take a peak at the Import menu, you'll see several different functions for importing different types of data. Some of the importers are specific to certain platforms, for example the Affymetrix and Illumina submenus. This tutorial demonstrates basic text and third party import. More details about each specific importer can be found in the Import Section of the SVS Manual.

Import Text

- Click Import > Text to open the text import dialog. This tutorial will provide the appropriate options for the example data. For more specific information about these options, see Text File Import.
- Click Browse and navigate to the previously downloaded files. Choose the simulated_phenotype.csv file.
- Leave the rest of the default values on the Input File tab as they are.
- Click on the Advanced Options tab and note the Missing Encoding Options section. In this file, the missing value indicator is ‘?’ which SVS will recognize automatically. None of these options need to be changed for this file.
- Click OK.

The resulting spreadsheet contains one binary column along with the row labels. The column type is specified by a blue B (for binary) to the left of the column number.
Import Third Party

- Click **Import > Third Party** to open the third party dialog. See Third Party Import for more information about the specific options.

- Click **Browse** and make sure **Excel (*.xls)** is selected after **File name**:. Choose **sample_data.xls** and click **Open**.

- Keep the default values for all other options on this page and click **Next**.

This importer will scan the file and provide a list of column headers found in the file. You can then choose which column is appropriate to use as the row labels. In this case the **Sample** column is selected by default.

- Click **Finish**

The resulting spreadsheet contains two columns in addition to the row labels. The first column **Population** is categorical (C) and the second **Age** is integer-valued (I).

### Drag-and-Drop DSF

DSF is the standard SVS file format that Golden Helix uses to store and share datasets. These files can be imported via the import menu (**Import > Golden Helix DSF**) or simply by dragging dropping the file into an open project.

- Navigate to directory where the previously downloaded files were saved. Click on **Affy 6.0 HapMap Example Data.dsf**, drag and drop it into the open project.

The resulting spreadsheet will have a marker map attached to it (note the green bar and the green spreadsheet node icon). Open the spreadsheet and click on the green Map button in the top left corner. The marker map information should now be visible (*Figure 3-1*).

---

*Figure 3-1 Affy 6.0 HapMap Example Data with Marker Map*
Note: Marker Maps can also be applied manually through File > Apply Genetic Marker Map. Marker Maps at minimum must contain chromosome and position information for each uniquely named marker. Depending on the source of your genotype data there can be several array specific fields that may not be available from other sources. For example, in the above screenshot the Associated Gene field is provided by Affymetrix in the annotation documentation for each specific array, and so will not be available for Illumina provided data.
Combining Datasets

Join/Merge Spreadsheet

Having all of the phenotype and genotype data in the same spreadsheet is often useful for analysis and visualization. Combining spreadsheets in SVS is straightforward, as long as the labels match in each spreadsheet.

- Open simulated_phenotype Dataset - Sheet 1 and choose File > Join or Merge Spreadsheets.
- Choose sample_data Dataset - Sheet 1 from the list and click OK.

The next dialog contains several options for merging datasets. In this example, row labels are appropriate matching criteria, there are no unmatched rows and no columns with the same name. In this merge, the resulting spreadsheet will be a child of the current spreadsheet.

- Leave the default options as they are and click OK.

**Note:** Use a custom sort order will use row number to match. The user can specify which labels to use in the merged spreadsheet.

The resulting spreadsheet has all of the phenotype and sample data in one spreadsheet.

**Note:** In a mapped spreadsheet the phenotype data and any unmapped genotype data will appear first, in the order in which it was added, followed by the mapped genotype data. Due to this ordering, when joining phenotype and genotype information you should always start with the phenotype spreadsheet or this information can appear in the middle of the spreadsheet.

A second merge will combine all datasets.

- From simulated_phenotype Dataset + sample_data Dataset - Sheet 1 choose File > Join or Merge Spreadsheets.
- Select Affy 6.0 HapMap Example Data - Sheet 1 from the list and click OK.
- After New dataset name: type Pheno + Geno. Choose Project root after Spreadsheet as Child of. The window should look like Figure 4-1. Click OK.

The resulting spreadsheet will have three phenotype columns followed by 56,222 mapped genotypic columns. Now that all of the data is combined into one spreadsheet all of the open windows can be closed and the other three dataset nodes can be collapsed to reduce clutter.

- From the Project Navigator Window go to Window > Close All
- Right-click on the simulated_phenotype Dataset root node and select Collapse All Children Nodes.
- Repeat for sample_data Dataset and Affy 6.0 HapMap Example Data.

**Note:** You can also click on the arrow to the left of the root node. This shortcut will hide all child spreadsheets.
The project navigator should now look like Figure 4-2.

**Note:** Just as Join/Merge combines dataset using row matching, append combines datasets using column header matching. This function is available under **File > Append Spreadsheets**. In general if you have more columns (phenotype or genotype) to add, use **Join/Merge** but if you have more rows (samples) to add, use **Append**.
Figure 4-2 Project Navigator with Nodes Hidden
Spreadsheet Manipulation

Change Single Column/Row State

The default state for every column and row in a spreadsheet is Active. You can change the state of a column or row by clicking on the column header or row label.

- From Pheno + Geno - Sheet 1 left-click once on the first column name header Case/Control.
- Left-click twice on the second column name header Population.

The Case/Control column is now colored magenta, signifying its dependent status. Since there is only one dependent column, the column header is also listed as dependent in the top-right corner of the spreadsheet. The Population column is inactive. This means that it will be ignored by all actions in SVS, including analysis, filtering and plotting.

Note: Many analysis functions require a binary dependent column or a quantitative (integer or real) dependent column.

The top-right corner of the spreadsheet always lists the total number of rows and columns (All: 100 x 56,225) and the number of active/dependent rows and columns Case/Control (Case/Control), 100 x 56,224. In this example there is one dependent and 56,224 active columns. See Figure 5-1.

Note: Rows can also be made inactive by left-clicking once on the row label. Rows cannot be made dependent.

Activation by Column Data

In order to investigate the data in a spreadsheet, a user may want to only activate rows that correspond to certain values in a data column. The activation mechanism available depends on the column type. Real and integer valued columns have Activate by Threshold available, while binary, categorical and genotypic columns allow you to Activate by Category.

- First activate all of the columns. Choose Select > Column > Activate All Columns.
- Right-click on Age and choose Activate by Threshold.
- Change the direction to >= and enter 55.
- Click OK.

45 samples are either 55 or older (as specified in the top right corner).

- Active all of the data once more, Select > Activate All.
- Right-click on Population and choose Activate by Category.
- Select CEU from the list and click OK.
9 samples are reported to be Caucasian based on the reported CEU value listed in the population column. You can also view all of the column values and their counts very easily.

- Choose Select > Activate All.
- Right-click on Population and choose Value Counts. See Figure 5-2.

Note: The value counts are also stored in the Node Change Log.

**Other types of Activation**

Another useful function for inactivating based on column data is Inactivate Missings, which is available in any binary, integer or real-valued column’s menu.

The Select menu features a few other useful activation functions as well. Select > Activate by Chromosomes will be available if the spreadsheet contains mapped data (does not have to be genotypic). Select > Column > Activate by Column Type is available in all spreadsheets.

Note: Make sure all data is active before proceeding. <CTRL> + A is the keyboard shortcut to do this.
Figure 5-2 Value Counts of Population
Editing Spreadsheets

Spreadsheet Editor

The Spreadsheet Editor in SVS allows you to edit single values, labels/headers, convert column types, and move or add columns. Only active data will be available and saved in the spreadsheet editor.

- From Pheno + Geno - Sheet 1 choose Edit > Edit This Spreadsheet.

Similar to Spreadsheet view, the spreadsheet editor functions that work on a whole-spreadsheet level are available in the spreadsheet editor menus. Functions that work on a column-data level are available in the column menu (accessed by right-clicking on a column header). For example, there is a Find and Replace function in the Edit menu (Edit > Replace...) that will search for values across the entire spreadsheet. Right-clicking on a column and choosing Find and Replace... will only consider values in that column.

Next, there will be a few examples of ways to edit data in the spreadsheet editor. These are for demonstration purposes only and the edited data will not be saved. The edited values will turn red in the spreadsheet editor.

- To edit a single value, double click on the spreadsheet cell. The cell will become a text edit box. 

Note: Changing a value to \( ? \) specifies a missing value.

- Choose Scripts > Create Column from Row Labels. The column created will always be categorical.

- Right-click on Age and choose Convert to > Binary by Threshold. Enter 55 and choose Overwrite column. Click OK.

You can also revert any changes you have made in the spreadsheet editor.

- Right-click on Age and choose Revert Column.

- Now choose File > Discard Changes and Close. Click Yes.

There are several functions available for editing columns in the spreadsheet editor and many will not be discussed in this tutorial. See Spreadsheet Editor for more information.

Recode Genotypes

SVS has developed several solutions for manipulating genotypic data in SVS. Some functions, such as Numeric Regression, require genotypes to be encoded numerically and Genotype Association Tests cannot handle multi-allelic columns and require a bi-allelic transformation.

- First encode the genotypes numerically, from Pheno + Geno - Sheet 1 choose Edit > Recode > Recode Genotypes.
Numeric encoding can be based on Major/Minor Allele specification (i.e. based on the allele frequency found in each column) or based on Reference/Alternate Allele specification. The latter requires a ‘Reference’ field in the marker map that contains the reference allele for each mapped genotypic column.

- Choose **Encode genotypes numerically based on genetic model:** and leave **Additive Model** selected below.
  The window should look like **Figure 6-1**. Click **OK**.

![Figure 6-1 Recode Genotypes](image)

The numerically encoded genotypes are created in the new spreadsheet **Pheno + Geno Numeric Genotypes Additive (Major/Minor used) - Sheet 1**.
Filter by Gene List

SVS allows for filtering a mapped genotype spreadsheet by a list of gene names. The first step is to import the gene list into a spreadsheet column in SVS.

- From the Project Navigator go to Import > Text, browse to select the gene_list.txt file. Leave the default Comma Delimited file format but select Generate generic row labels (1,2,3...) under Row Labels. The dialog should look like the figure 7-1 then click OK.

Figure 7-1 Text Import of Gene List
You will get a message saying *File does not appear to be comma-delimited, or the file has only one column. Continue Anyway?*, Click **Yes** to continue the import.

The resulting spreadsheet has one column with eight gene names listed.

![Gene List Spreadsheet](image)

**Figure 7-2 Gene List Spreadsheet**

Now to filter your genotype data down to only those that fall within the eight genes on our list.

- Open **Pheno + Geno - Sheet 1** and then go to **Select > Activate by Gene List**
- Select the **gene_list Dataset - Sheet 1** spreadsheet and click **OK**

![Select Spreadsheet with Gene List](image)

**Figure 7-3 Select Spreadsheet with Gene List**

- Select the **Gene Name** column.
- Select a gene annotation track that contains your gene names as it will be used to define start and stop positions used for activating your markers. By default the **RefSeqGenes105v2-NCBI** track for the default genome assembly list for the project will be selected.
• Make sure the dialog looks like Figure 7-4 and click **OK**

![Figure 7-4 Activate by Gene List Options](image)

For this dataset 71 markers were found within the specified genes, all those that were outside the defined regions would have been inactivated.

• Reactivate your phenotype columns by left-clicking once on each column header and then take a subset spreadsheet by going to **Select > Column > Column Subset Spreadsheet**

• Rename the subset **SNPs within Candidate Genes**

• Close all open spreadsheets by going to **Window > Close All**

• Go to **File > Quit** to close SVS and save the project.
Conclusion

Congratulations!

You have completed the Introduction to SVS tutorial! The tools and functions that you learned in this tutorial will be useful regardless of what SVS package you have. You can find more tutorials specific to your SVS package on the main tutorial page.

We welcome your feedback, comments or questions about this or any other tutorial. Email the Golden Helix support team at support@goldenhelix.com.