
# GenomeBrowse Manual

**Release 2.0.5**

**Golden Helix, Inc.**

**September 09, 2014**

# CONTENTS

The GenomeBrowse visualization tool delivers stunning visualizations of your genomic data that give you the power to see what is occurring at each base pair in your samples. A high performance backend is paired with an intuitive user interface to make sure that your discovery process is fluid and streamlined.

**Acknowledgments**

**Golden Helix GenomeBrowse®** visualization tool would not exist without the generous contributions of many minds and hearts. We would particularly like to thank all of the GenomeBrowse beta testers and everyone who has given GenomeBrowse a try and provided feedback.

# CHAPTER ONE

# INSTALLING AND INITIALIZING

## 1.1 Installation Under Windows

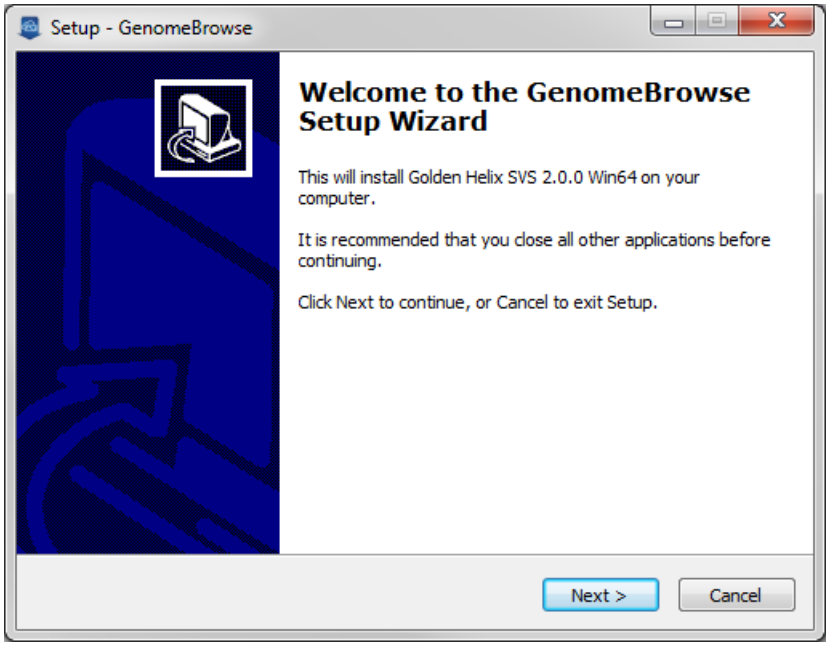



Figure 1.1: GenomeBrowse Setup Program

Double-click on the self-extracting executable (GenomeBrowse-Win32-x.x.x.exe or GenomeBrowse-Win64-x.x.x.exe). The x.x.x will be a version number, e.g. GenomeBrowse-Win32-2.0.0.exe. It will open up an install dialog screen.

The default is to install the software in the users application data folder which is approximately `C:\Users\<User Name>\AppData\Local\Golden Helix\GenomeBrowse\Application\`.

Under Windows, you also have the option of adding a GenomeBrowse program icon to your desktop, a Quick Launch icon, and having GenomeBrowse available from the Start Menu.

Once GenomeBrowse is installed you can run the program either by using the shortcuts or by running the executable. This will launch the login and registration dialog. If you already have a GenomeBrowse account enter in your credentials to launch the application. Otherwise register an email address and to create an account to launch the application.

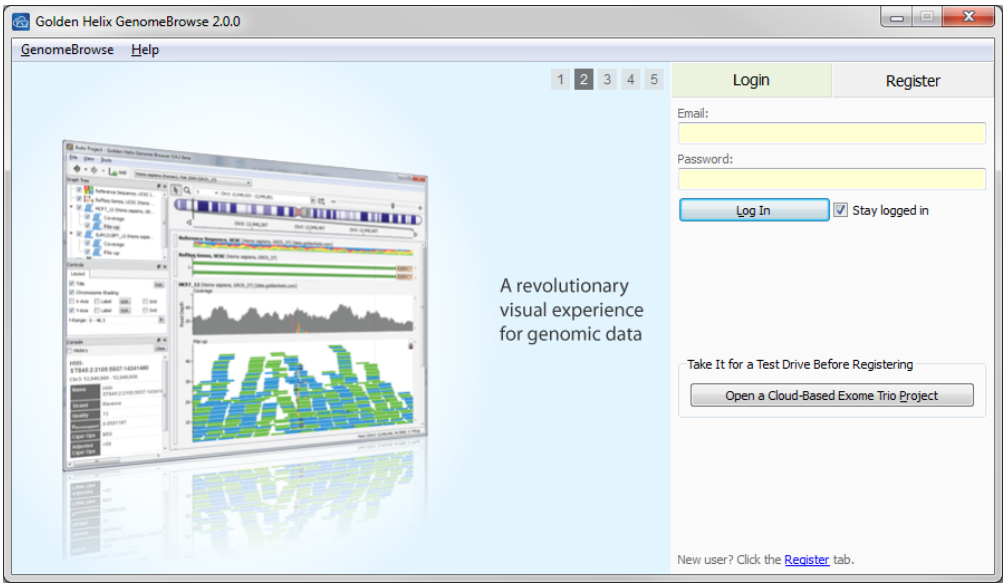



Figure 1.2: Login or register for a GenomeBrowse account

## 1.2 Installation Under Linux Distributions other than RHEL5 and CentOS5

1. Download the GenomeBrowse archive for Lin32 or Lin64 to a convenient user directory.
2. The bundle is a ".tar.gz" archive file that will extract a folder called "GenomeBrowse". You can extract the archive using Linux GUI or command line tools. On the command line perform the following command:

   ```
   ~/Programs/> tar -xzvf GenomeBrowse-Lin64-2.0.0.tar.gz
   ```

3. Once extracted, you can move the resulting GenomeBrowse folder to another location of your choice at any time.
4. Go into the GenomeBrowse folder and run the GenomeBrowse App. The program will launch and present the login and registration dialog. If you already have a GenomeBrowse account enter in your credentials to launch the application. Otherwise register an email address and to create an account to launch the application. You can do this with the following commands:

   ```
   ~/Programs/> cd "GenomeBrowse"
   ~/Programs/GenomeBrowse> ./GenomeBrowse
   ```

**Note:** You may create a symbolic link to the GenomeBrowse program and even put it on your path so that it can be launched from any directory. This also allows you to run scripts from various directories.

Our Linux binaries are compiled on Ubuntu Hardy 8.04 for maximum compatibility with most Linux distributions. Known incompatibilities exist with Red Hat Enterprise Linux (RHEL) and CentOS version 4 and earlier. RHEL versions 7.x and newer are supported as well as Ubuntu version 7.04 and newer. If you have RHEL/CentOS 5.x or 6.x see the instructions below *Installation Under RHEL5 and CentOS5 Linux Distributions*.

Please contact support if your experience issues running GenomeBrowse on your Linux machine.

## 1.3 Installation Under RHEL5 and CentOS5 Linux Distributions

1. Download the GenomeBrowse archive for RHEL5 to a convenient user directory.

2. The bundle is a ".tar.gz" archive file that will extract a folder called "GenomeBrowse". You can extract the archive using Linux GUI or command line tools. On the command line perform the following command:

```
~/Programs/> tar -xzvf GenomeBrowse-RHEL5-7.6.7.tar.gz
```

3. Once extracted, you can move the resulting GenomeBrowse folder to another location of your choice at any time.
4. Go into the GenomeBrowse folder and run the GenomeBrowse App. The program will launch and present the login and registration dialog. If you already have a GenomeBrowse account enter in your credentials to launch the application. Otherwise register an email address and to create an account to launch the application. You can do this with the following commands:

```
~/Programs/> cd "GenomeBrowse"
~/Programs/GenomeBrowse> ./GenomeBrowse
```

Our RHEL5 binary is compiled on CentOS5 for compatibility with RHEL/CentOS versions 5.x and 6.x.

Please contact support if your experience issues running GenomeBrowse on your Linux machine.

## 1.4 Installation Under Mac OS X

1. Download the GenomeBrowse app bundle for Mac to a convenient temporary directory.
2. Run the installer by double clicking on the package and click and drag the GenomeBrowse application into the Applications folder.
3. Go into the GenomeBrowse folder and run the GenomeBrowse App. The program will launch the login and registration dialog. If you already have a GenomeBrowse account enter in your credentials to launch the application. Otherwise register an email address and to create an account to launch the application.

# CHAPTER TWO

# RELEASE NOTES

## 2.1 2.0.5

- Added a "slice" function to the expression editor to strip out characters from data. This would allow values such as "[A/G]" to be treated like A/G for styling purposes.
- Feature List now remembers the previous maximum number of features displayed for the same source. Now add additional features in chunks of up to 10,000 features.
- Data Source Library export a VCF file in the VCF format now correctly copying format field information. Some numeric array fields were being classified as string arrays.
- Fixed font in data console for source information.
- Fixed Amino Acid rendering in MT Chromosome. Now, the amino acids in MT use the MT table instead of the standard amino acid table. See: https://www.mun.ca/biology/scarr/MGA2-03-28_mtDNA_code.jpg
- **Convert Source Wizard** bugs fixed:
  - Allow segments/scaffolds/chromosomes to be skipped in the segment list.
  - Fixed conversion for GTF files with no exon features to an annotation source this should enable the Convert Source Wizard to work for GenCode GTF files.
  - Fixed detection of list/array fields if the first occurrence of a list is not in the first 1000 lines, also the converter now properly handles a field that contains a mixture of no values, a single value and lists of values.
- Suppress incorrect "GRCh_38" assembly from the assembly list if present.
- Prevent crash when sorting by a categorical array field in the Feature List.
- Fixed **Export to VCF** from an annotation source, prevented a crash when the doc string is missing.
- Fixed downloading annotation sources from the Data Source Library on Mac OS X.
- The shipped **RefSeq Genes 105, NCBI** source was indexed to allow searching on gene and transcript fields.
- Prevented searching non-indexed annotation sources to avoid cluttering the Search and Location Bar.
- Fixed crash when visualizing a region that included a coverage transform at the exact boundary of a chunk of data.
- Updated the default recent genome assembly list to include the three most recent human genome assemblies, and the most recent mouse and rat assemblies.

## 2.2 2.0.4

- New default RefSeq Gene Annotation Source based on GRCh37_g1k build curated directly from NCBI.
- Changed system default genome assembly to GRCh_g1k to take advantage of the best mitochondrial reference sequence and the new RefSeq gene annotation source with updated mitochondrial gene annotations.
- GenomeBrowse Source Convert Wizard
  - can now handle enumerated or string lists for features.
  - can compute extents for very small float64 values (i.e. p-values) on Linux/ Mac OSX.
- GenomeBrowse bugs fixed:
  - Save as Image will now respect the current zoom if the region was obtained by entering the location in the genomic location bar without also scrolling to zoom.
  - Deleting one or more plots by selecting them in the plot view and pressing the delete key no longer crashes the program.

## 2.3 2.0.3

- Fixed GTF source conversion to adjust CDS start/stop based on frame and allow for CDS to jump introns.
- Fixed crash when reordering plots on MacOSX by clicking and dragging one or more plots in the plot view and clicking on the GenomeBrowse window before the plots finished drawing.
- Added option for variant sources to left-align indels using a reference sequence source.

## 2.4 2.0.2

- Added the ability to save annotation sources as VCF, XLS, FASTA, and WIG files.
- Updated and expanded the functionality of the Expression Editor for filtering and added it to the Source Convert Wizard. Now fields can be computed based on existing fields for either filtering or adding to annotation sources.
- Fixed the following GTF source conversion problems:
  - Exon starts and stops were listed in strand order instead of genomic order. This only resulted in issues when performing Variant Classification using these files.
  - Accounted for out of frame transcripts to result in fewer invalid transcripts.
  - Added source field GTF file source info.
- Fixed **Convert GFF Files to Annotation Track** to work on files that do not have mrna features.
- Replaced tabix with htslib to handle symbol collision. This should fix all crashes that were occurring when trying to compute coverage and index on BAM files and indexing and compressing VCF files.
- Fixed help links for data source library, source convert wizard and on plot help links.
- Fixed click issues for Marker Blocks on LD plots. It should now be easier to select the correct marker and marker block.
- GAserver has been updated to handle alias chromosome naming for remote sources.
- Added the "*.fna" extension to FASTA file options for the source convert wizard.

## 2.5 2.0.1

- Added Active/Passive FTP option for loading BAM files from URL.
- Delayed launching Evernote Note until after an authenticated connection is made. This fixes a problem when the note for a project failes to load on startup for some users.
- Fixed loss of dock size information on minimize/restore of GenomeBrowse main window.
- Fixed computation of BAM indexes and coverage files by upgrading to the latest htslib library.
- Fix regression where only one computation was run on a source such that a source requiring an index and coverage computation would not kick off the coverage after the index finished.
- Pileups now update after changing options without needing to force an update.
- Made it possible for interval tracks with zero width features to label all non-zero width features.
- Have assembly aliases used consistently in readers to re-map segment names.
- Made sure that each GenomeBrowse project remembers whether or not the feature list was shown or hidden.
- Fixed bug that prevented the tabix TBI file from being generated for a >2GB bgzip compressed VCF file on windows.

## 2.6 2.0.0

In version 2 we have rebuilt most of our backend and almost all of our frontend of GenomeBrowse with too many changes to count. Here are the large features you will notice that are new and changed.

- The "Add" sources dialog has been redesigned to easily allow managing multiple local and remote sources in one view. It features hierarchical repositories, a information pane for selected sources and the ability to plot other fields in a source other than the default (for example, plotting the allele frequencies from 1000 Genomes track instead of the variants).
- From the Add dialog, there is a new Convert Wizard to convert any tabular file as well as all our supported file formats to the new compressed and index TSF format. All public annotations are now also in TSF, which can be 1/10 the size of their previous versions. This includes the ability to import custom genomes from FASTA files and define a new genome assembly.
- The public data repository now is organized with folders to make it easier to find a specific type of data. Also, by default only the latest version of an annotation source is shown.
- Added ability to save plots as images. You can right click on individual plots or save all the plots in a project.
- Added documentation in the form of a GenomeBrowse manual as well as on plot help controls to jump to specific sections of the manual referring to the plot type.
- Added Evernote integration. You can create new notes or open existing ones from your notebook once you connect your Evernote account. The note editor is attached to the project and can be used to add bookmarks and insert images from your current context.
- Added IGV Session file support. All sources in the IGV session file that are supported by GenomeBrowse are added to the current project.
- Added support for streaming BAM from HTTP and FTP. Just click the URL button in the Add dialog and enter the URL of either a single source, or the index file or directory that contains BAM files. It's required that each BAM has a corresponding ".bai" file.
- Support for loading sources and changing the zoom of GenomeBrowse from other programs has been added in a number of forms. The installer now registers the genomebrowse: protocol and URLs can control a new

or existing GenomeBrowse process. Also, under the Program tab of the Options dialog, you can enable a *Link Server* that by default listens on the same localhost port that IGV runs on, allowing programs that control IGV through http://localhost:60151/ urls to control GenomeBrowse.

- Numerous other optimizations, polishes and bug fixes.

## 2.7 1.1.2

- Added support for reading sorted BED files directly! The "track" line is respected to support display name, coloring and the bedDetail format variant.
- A new table view of a source's features is available on the top-left of plots when you hover over them. You can view the features based on the current zoom or from the start of the source. Selecting features in this view jumps the zoom to that position.
- Added more details to the data console output for BAM coverage plots. Percentage of total reads is now an additional column for each nucleotide and insertion detected at the current position.
- Added an additional option to choose whether to automatically display of Coverage and/or Pile-up plots when adding new BAM sources. If either coverage or the pile-up plot is not displayed automatically, it can be shown by checking the appropriate box in the plot tree view.
- The "garead" and "gautil" command line utilities are bundled with GenomeBrowse. Running `gautil precompute <source.bam>` will produce index and coverage files. This may be useful for pipeline environments to prepare files for GenomeBrowse visualization.
- Some glitches with parsing the input to the location bar have been resolved. It has also been improved to be able to take whitespace delimited region descriptions like "chr7 123,456 444,555" as well as many others. As you type, the normalized range that was detected is shown in the results drop down so you know exactly what range will be used when you hit Enter.

## 2.8 1.1.1

- Overhauled the touch pad input handling on Mac. Zooming is significantly more responsive. Two finger scroll up and down zooms in and out while side to side pans the view.
- Before registering, you can now open a cloud-based demo project to give GenomeBrowse a test drive.
- Show transcript names as a hover-label when looking at whole genes.
- Fixed some crashes and unexpected behavior with the new downloader engine.
- Fixed issue where you could not log in if the "Stay Logged In" checkbox was not checked.
- Fix display of data for VCF coverage at chromosome level scale for sparsely populated sources.
- Reverse-strand hard-clipped reads were not being aligned correctly and thus looked like they had a lot of mismatches.
- Updated the links for gene names in the data console.

## 2.9 1.1.0

- Completely new downloader engine that supports download acceleration, pause/resume, auto-resume and can be minimized to the system tray for long downloads.
- Files are now downloaded with their precomputed coverage or indexes so they are immediately displayable.

- From pipeline.goldenhelix.com, ghdownload will now open the new GenomeBrowse downloader.
- Support for VCF Files! New rendering mode for multi-sample "Variant Maps", as well as auto-compression/indexing of VCF to Bgzip/Tabix format required to read files directly for rendering.
- Single sample VCFs, multi-sample VCFs and even "site"-based VCFs with no sample data are now supported.
- SNV and InDels supported. We will do our best to draw Structural Variants, but their representation in VCF format is less standardized.
- GenomeBrowse should now start faster once you have downloaded the reference sequence track for your existing project.
- We have experimental support for "Value" tracks. More support coming in the near future.
- Many polishes and bug-fixes, including improved controls for changing labels on features.
- Gene tracks have been much enhanced. Exon numbers and transcript names are labeled at appropriate zoom levels as well as more details are provided while hovering over codons.
- The reference sequence track gives you forward, reverse and all theoretical amino acid encodings as you expand its height to give it more space to draw.
- Improved Y-axis labels, which is especially important when looking at VCF tracks with sample names as Y-axis labels.

## 2.10 1.0.7

- Added the "g1k" Human Reference build, which is GRCh37 with the rCRS MT as specified by the 1000 genomes project.
- Updated the genome assembly file format to handle more meta-data. Also improved when warnings are displayed for a source not matching the selected build.
- Added ability to change the field used to label features for Interval and Variant plots.
- Polished the label drawing system for all plot types.

## 2.11 1.0.6

- Added ability to filter low quality alignments from pileups and coverage plots.
- Polished rendering of BAM plots and mouse-hover capabilities.
- Performance improvements when rendering views with tens of thousands of reads.
- Significantly improved "gear" menu and control panel capabilities of plots.
- Fixed some crashes in pre-compute reported by users.
- Added an Update Available notification.

## 2.12 1.0.5

- Add ability to have project documentation.
- Have "demo" mode for putting on thumb drives.

## 2.13 1.0.4

- Added project management! Create new projects, save and open recent projects.
- Multi-window support: can open multiple projects in multiple windows.
- Added Illumina BaseSpace integration in the 'Add' dialog.
- Amino-acid information now drawn on genes. Also labeling of gene parts enhanced.
- Fixed crashes associated with attempting to index or run precompute on an unsorted BAM file.
- Fixed crashes when precomputing on certain BAM files.
- Prevent cache system by being overloaded by 30K+ read-depth regions. This should help with out-of-memory errors when viewing areas ultra-high-read-depth.

## 2.14 1.0.3

- Fixed issue where the directory to store settings and the default project was not created. You should now not need to log in every time you start GenomeBrowse and your project should save on close.
- The message boxes and progress dialogs have been updated to have a more consistent style.

## 2.15 1.0.2

- Made the Windows installer able to be run by a non-Administrator user.
- On Mac, use the native directory chooser so you can pick mounted volumes.
- Prompt the user to download a reference sequence track if there is not one locally when switching genomes or on first use.

## 2.16 1.0.1

- Fixed some crashes on Mac OS X when switching to and from GenomeBrowse.
- Fixed plots that got stuck in "Initializing..." mode.
- A few fixes of handling views of edge cases.

## 2.17 1.0.0 - Yea!

- Complete overhaul of the zoom system. Scroll wheel zooming on pile-up beautifully scales the content to an ideal aspect ratio.
- Login and registration window in place.
- Improved drag-n-drop of plots by their handles.
- Per-base quality information is used to shade mismatches in both pile-ups and coverage views.
- Many tweaks, polish and optimizations.

## 2.18 0.9.9

- Improved interactive mouse support for insertions. Labels displayed on hover for pileup and coverage plots.
- Polish and performance improvements to BAM rendering plots.
- While Y-axis zoom lock is enabled, smoothly adjust the y-limits as you pan into areas with high and low coverage.
- Significant performance improvement for Mac renderings, and improved handling of touchpad inputs like pinch-zoom and panning.

## 2.19 0.9.8

- Added interactive mouse support for the labeling system, allow you to hover over genes or variants and see their label even if it was hidden by default.
- Fixed crash when viewing certain regions of the 1000 Genomes track from our annotation server.
- Certain types of BAM files were creating invalid BAI files, which is now corrected.
- Fixed UI freeze that occurred when you tried to delete a plot that had a computation in the queue to be run.

## 2.20 0.9.7

- Many stylistic improvements and user interface polish and tweaks.
- Detect sample.bai as valid BAI file instead of just sample.bam.bai
- Larger zoom limits in pile-up mode before switching to drawing coverage

## 2.21 0.9.6

- Updated logo and name

## 2.22 0.9.5

- New <a href="http://pipeline.goldenhelix.com/">pipeline.goldenhelix.com</a> account integration. Re-designed the "Add" dialog to have tabs for Annotations, EA Pipeline Account and Example Samples.
- The Landmark view displays the Cytobands track by default, but can be set manually to show other plots.
- Totally new labeling system for genes and variant tracks. You'll notice transcripts are grouped by their gene and have labels on the exons and utrs. Genes also collapse transcripts when vertical space is tight.
- The BAM pile-up plot now has a "gear" icon in the top left that allows you to switch coloring modes to emphasize mismatches versus strands. Also "Split by Strand" stacks forward and reverse strands above and below the X-axis.
- Tools->Proxy Settings allows you to auto-detect or manually enter proxy settings.
- Many stylistic improvements and user interface polish and tweaks.

# CHAPTER THREE

# LAUNCHING GENOMEBROWSE

When you launch GenomeBrowse you may encounter a few dialogs before a project opens up or is created. These dialogs are outlined below.

## 3.1 Login and Register Dialog

If you are opening GenomeBrowse for the first time on a machine (or for the first time period) the first dialog you will see on launching GenomeBrowse is the dialog to login or register for a GenomeBrowse account. You may also see this dialog if you choose not to have your credentials saved or have previously logged out of GenomeBrowse. See *Login or register for a GenomeBrowse account*.

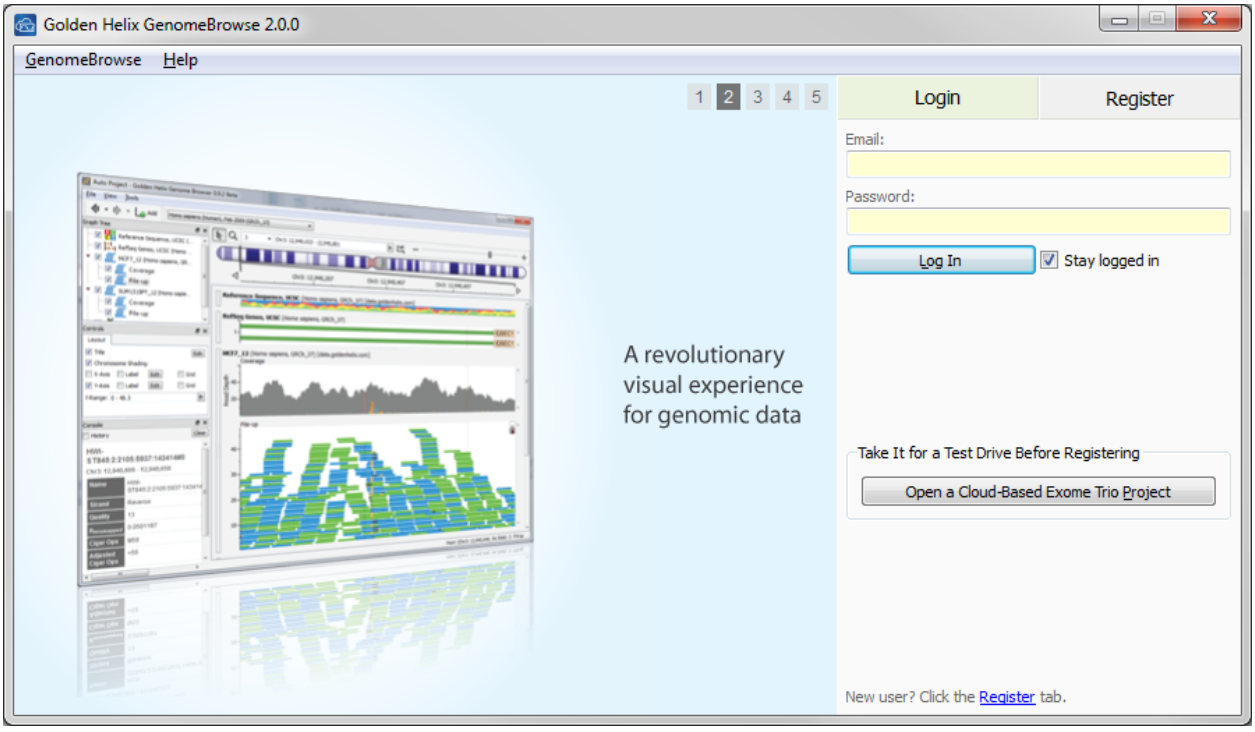



Figure 3.1: Login or register for a GenomeBrowse account

### Login

If you have an existing GenomeBrowse account, enter in your email address and password. You can optionally uncheck the "Stay logged in" option to have GenomeBrowse not store credentials locally and require logging in again after relaunching the software. After the account information is filled in, click **Log In**.

To recover a lost password go to http://answers.goldenhelix.com/account/signin/?next=/ and click on the link to "send your email a password reset link".

### Register

If you do not have an existing GenomeBrowse account, click on the **Register** tab and fill out the registration form. Once the required fields have been filled in and the license agreement has been accepted the **Register & Log In** button will become active. You can optionally uncheck the "Stay logged in" option to have GenomeBrowse not store credentials locally and require logging in again after relaunching the software.

The GenomeBrowse account will also be the credentials used to log into the answers.goldenhelix.com community site.

### Example Project

To try GenomeBrowse without registering first, click on the **Open a Cloud-Based Exome Trio Project** button. The project is fully usable but changes cannot be saved.

## 3.2 Adjusting Proxy Settings

To allow GenomeBrowse to communicate with our data servers, if your institution has a proxy server, the proxy settings need to be set appropriately. Please request assistance from your IT department if needed for the proxy settings.

To change the proxy settings, from the **Login and Register** dialog, go to **GenomeBrowse > Proxy Settings**. Adjust the proxy settings as required and click **OK**.

If you have already logged in and are in a project, you can still adjust the proxy settings from **Tools > Proxy Settings**. You may need to restart the program for all uses of the network to properly pick up the new settings.

## 3.3 Create a New Project on Launch

If an existing project is not detected, then on launch the **New Project** dialog will be presented. This asks for the project name as well as the genome build to use in the new project.

Set the project name and select the species and build to use from the genome assembly drop-down menu.

If the species or build is not included in the list, you can create a new genome assembly using the **Convert Source Wizard** as long as there is a reference sequence available for that particular species and build. See either *Convert a 2Bit File* or *Converting a FASTA File* for more information. The **Convert Source Wizard** can only be launched from the Data Source Library or from the GenomeBrowse File menu in an existing project, so specify a temporary genome for a temporary project. Once the new genome assembly is created a new project can be created. See *Open a New Project from an Existing Project*.

### Download Reference Sequence Question

If a reference sequence for the genome selected exists on the public data annotations repository, and the reference sequence is not detected in the User Annotations folders, a question dialog will appear asking whether or not the reference sequence should be downloaded. See *Download reference sequence question*.

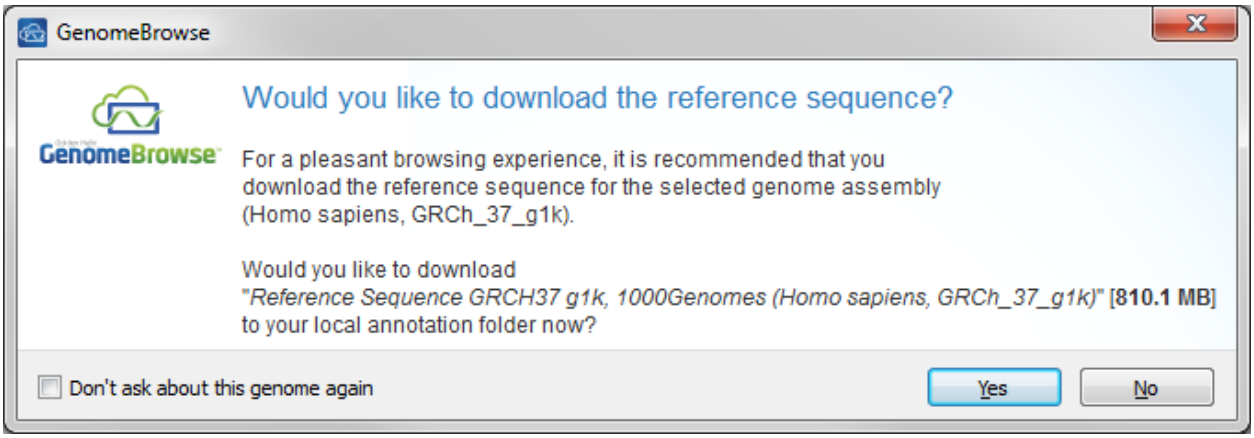



Figure 3.2: Download reference sequence question

To prevent this dialog from reappearing for the specific genome, check the "Don't ask about this genome again" box at the bottom of the dialog.

Otherwise, to download the reference sequence in a background process, click **Yes**. To not download the reference sequence, click **No**.

For more information on downloading annotation sources see *Download Window*.

# CHAPTER FOUR

# NAVIGATING THE GENOMEBROWSE WINDOW

On the GenomeBrowse Window there are several regions to note and become familiar with:

- The Project/Window Title Bar
- Menu Bar
- Toolbars:
  - Location
  - Mode
  - General
  - Genome
  - Zoom
  - Project
- Domain View
- Plot View
- Plot Tree
- Control Panel
- Console
- Feature List
- Evernote Text Editor

## 4.1 Project Title Bar

The title bar reflects the GenomeBrowse project name and version.

## 4.2 Menu Bar

The menu bar consists of the following menus: File, View, Tools and Help.

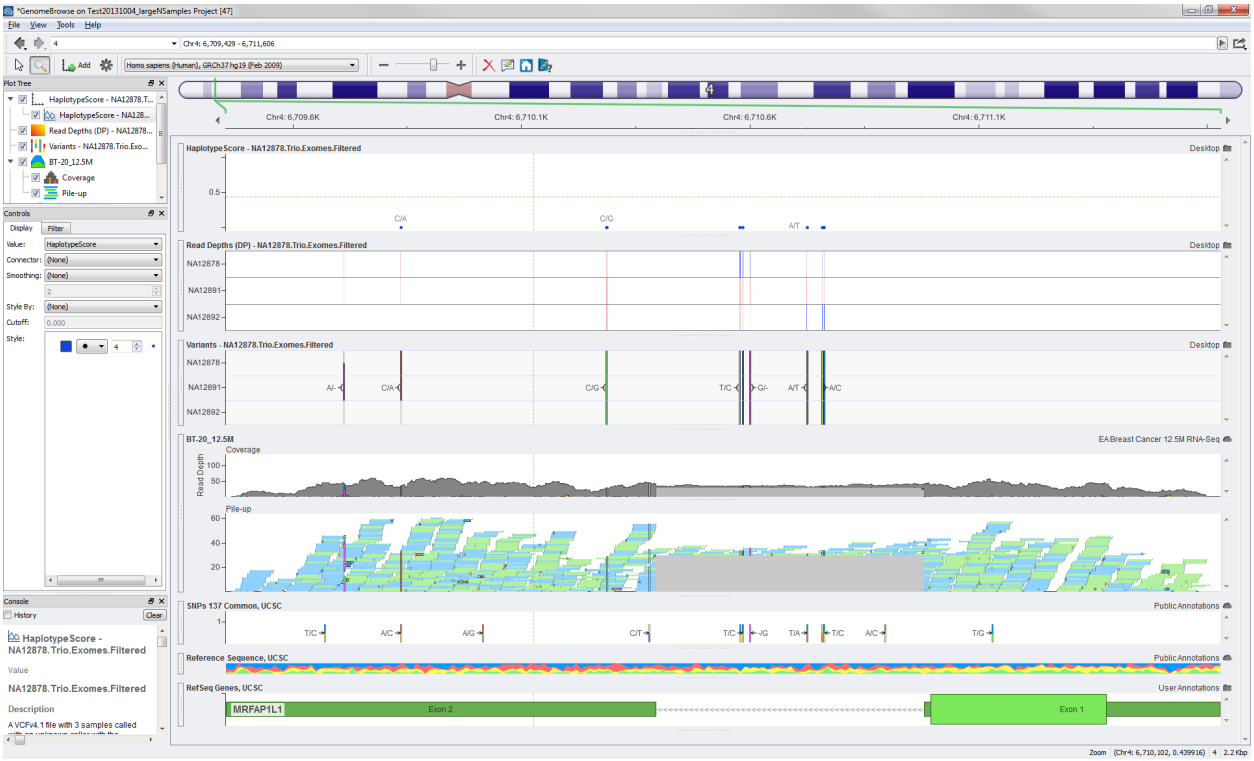


Figure 4.1: Using GenomeBrowse to visualize data from a VCF plot and a BAM file

## File Menu

The file menu contains the following items:

- **New Project...**: Close the existing project and create a new project. See *Open a New Project from an Existing Project*.
- **Open Project...**: Close the current project and open an existing project.
- **Open Recent Project**: A list of recent projects. Selecting a project from this list will close the current project and open the recent project.
- **Save Project**: Saves the current project.
- **Save Project As...**: Saves the current project to a new file.
- **Add...**: Opens the Data Source Library to add plots from data sources to the current project. See *The Data Source Library* for more information.
- **Convert..**: Opens the Convert Source Wizard to convert data sources to TSF format. See *Annotation Convert Source Wizard* for more information.
- **Import IGV Session...**: Imports an IGV session file and creates a GenomeBrowse project from the session file. See *Import IGV Session* for more information.
- **Evernote**: This submenu contains options for Evernote. See *Evernote: Cloud-based Project Documentation* for more information. The menu items will change depending on whether GenomeBrowse has been connected to an Evernote account.

- **Save As Image...**: Save one or more plots as images. See *Saving Plots from a GenomeBrowse Window* for more information.
- **Log out <user email>**: Log out the current GenomeBrowse account and return to the log in and register window.
- **Exit**: Exit GenomeBrowse.

### View Menu

The **View** menu contains controls to hide or show the various dock windows and toolbars.

- **Dock Window**: Windows that can be moved, docked and hidden in GenomeBrowse. Dock windows include:
  - **Plot Tree**: See *Plot Tree*
  - **Controls**: See *Control Panel*
  - **Console**: See *Console*
  - **Feature List**: See *Feature List*
- **Toolbar**: Toolbars that can be shown or hidden in GenomeBrowse. See *Toolbars* for more information.

### Tools Menu

The **Tools** menu contains various GenomeBrowse tools and utilities including:

- **Downloads**: Launches the download manager. See *Download Window* for more information.
- **Open Folder**:
  - **User Data Folder**: Launches (or puts into focus, if one already exists) a file browser window from your operating system which is open to the User Data folder. GenomeBrowse projects are saved in the User Data folder by default.
  - **Program Folder**: Launches (or puts into focus, if one already exists) a file browser window from your operating system which is open to the program installation folder.
  - **Annotations Folder**: Launches (or puts into focus, if one already exists) a file browser window from your operating system which is open to the local annotations folder.
- **Proxy Settings**: Dialog to adjust the proxy settings to allow GenomeBrowse connect with the Golden Helix data servers. See *Adjusting Proxy Settings* for more information.
- **Options...**: General GenomeBrowse options. See *General Options for GenomeBrowse* for more information.

### Window Menu

The **Window** menu contains a list of all open project windows as well as the **New Window** menu item. Selecting **New Window** launches a new GenomeBrowse instance open by default to the current project. Multiple projects can be open at the same time by selecting a window and switching the open project.

### Help Menu

The **Help** menu contains various menu items to provide assistance with GenomeBrowse including:

- **GenomeBrowse Manual**: Launches the GenomeBrowse manual.

- **GenomeBrowse PDF Manual**: Opens the GenomeBrowse PDF manual.
- **Ask Questions**: Launches the answers.goldenhelix.com community support site, a place where you can ask or answer questions.
- **Online Help**: Launches the GenomeBrowse online resources page.
- **Release Notes**: Opens the release notes.
- **About...**: Provides information regarding the version of the software you have installed and important license information.

## 4.3 Toolbars

### Location Toolbar

The location and navigation bar features back/forward buttons to retrieve previous zoom locations, a chromosome selector, a genomic location bar, and an external browser link list. Each feature is described below.

- **Back/Forward** buttons: These buttons act similar to Internet browsers, they enable you to jump back to the previous zoom or go forward to the next zoom. A slow click or right-click will bring up a zoom history to jump back or forward to a specific zoom from some point in the immediate history.
- **Chromosome Selector**: This dialog contains a drop down list of chromosome names. The whole genome view can be obtained by selecting *All* from the list. Also, a chromosome name(s) can be typed in the box to jump to it rather than selecting it from the list. Examples of input include *1*, *1-3*, etc.
- **Genomic Location Bar**: A chromosome and position string can be entered into this dialog to jump to the region. Cytobands and gene names can also be entered into this dialog. Once the region has been entered, either click on the "Go" or "Play" button, press "Enter", or select a region from a list of potential zooms when applicable to jump to the desired genomic coordinates.
- **External Browser Links**: Multiple external browsers can be accessed for the zoom region specified in the Genomic Zoom Region box. Regions spanning multiple chromosomes are not supported by the external browsers and will result in a warning when attempting to link out.

### Mode Toolbar

There are two zoom modes, navigation pointer mode and zoom mode.

- Navigation pointer mode allows for click-and-drag operations in plots or scales to pan the view. Right-click-and-drag operations cause the view to be scaled around the initial click point. The scroll wheel zooms the view in and out. This mode can be accessed using the hot-keys **a** or **p**.
- Zoom mode allows for click-and-drag selection of a region in plots or scales. When the drag operation is concluded, the selected region will become the new view. Right-click-and-drag selection will zoom the view in such a way that the previous view fits inside the selected region in the new view. In other words, selecting a small region with the right mouse button causes the view to zoom out a lot. Selecting a large region with the right mouse button causes the view to zoom out a little bit. The scroll wheel zooms the view in and out in this mode as well. This mode can be accessed using the hot-key **z**.

Any drag operation can be canceled by pressing the opposite mouse button during the drag. Cancel a left-drag by pressing the right mouse button. Cancel a right-drag by pressing the left mouse button.

### General Toolbar

The general toolbar contains controls for adding plots, saving all plots in the view as an image, and setting default options for a GenomeBrowse window.

- **Add**: See *The Data Source Library as an Add Dialog* for more information on the Add dialog.
- **Save As Image**: See *Saving Plots from a GenomeBrowse Window* for more information on the Save As Image dialog.
- **Options**: See *General Options for GenomeBrowse* for more information on the Options dialog.

### Genome Toolbar

The genome assembly (species/build) can be changed through the genome toolbar. Recently used or common genomes are at the top of the list, otherwise all genome assemblies available are listed alphabetically by the scientific name.

### Zoom Toolbar

The zoom tool bar offers a zoom slider and buttons for zooming the view in and out only on the x-axis. Slide the bar to the left (toward the minus **-** button) to zoom out, slide the bar to the right (toward the plus **+** button) to zoom in. The buttons zoom view in (**+**) or out (**-**) on the x-axis in steps.

### Project Toolbar

The project tool bar contains the help icon. Clicking on this icon will open the GenomeBrowse Manual.

## 4.4 Domain View

The domain view provides genome wide context. If a cytoband source for the current genome assembly is available in any "bookmarked" location (See *The Location Panel* for more information) it will be set as the domain view plot by default. Any plot can be set as the domain view plot by right clicking on the source and selecting **Set as Domain View Plot**. If a cytoband source is available but not currently set as the domain view plot, the option to restore it will be provided in the right-click menu for the domain view plot.

The domain view also includes chromosome labels, the domain view funnel which illustrates the mapping of the current view from the x-axis scale onto the domain view plot, and the current x-axis scale. Each of these components can be hidden or shown by right clicking on the domain view and checking the appropriate entry in the menu.

## 4.5 Plot View

The main portion of the GenomeBrowse window is the plot view. This is the region of the window where the sources are drawn in plots. Each plot in the plot view has several special features.

- **Rearrange Plots**: Plots in the plot view can be rearranged by clicking, dragging, and dropping the grab bar on the left of the plot.
- **Controls**: Some common plot controls are accessible for many plots by hovering over the upper left corner of the plot and clicking on the gear icon. The complete set of controls is available in the control panel (docked

in the middle of the left side of the GenomeBrowse window by default). The control panel acts on the current selection and may be composed of multiple tabs depending on what is selected.

- **Feature List**: The feature list can be shown or updated to include information for the first data source in a plot by hovering over the upper left corner of the plot and clicking on the table icon. The feature list can also be shown or updated to the first data source in the plot by right clicking on the plot and selecting **Feature List**. Once the feature list is shown, a different data source can be selected by name from the drop-down box in its upper left corner.
- **Title Editing**: The title can be edited in place by double clicking on the plot title. It can also be edited by right clicking on the title and selecting **Edit Title...**, or through the control panel.
- **Data Source Location**: The location of the data source for the plot is displayed above the upper right corner of plot.
- **Scale Label Editing**: The labels for both the x-axis and y-axis can be modified. When visible, they can be edited in place by double clicking on them. They can also be edited by right clicking and selecting **Edit [X/Y]-Axis Label...**, or through the control panel.
- **Y-axis Zoom Mode**: The current y-axis zoom mode can be identified by hovering over a plot. The second button from the top in the upper right corner will show a letter designating the current mode. The current y-axis zoom mode can be changed by clicking on that button and choosing one of the available modes from the menu. There are four available y-axis zoom modes:
  - **Manual - The y-axis zoom is controlled manually and all zoom controls are** enabled. This mode can be accessed using the hot-keys **r** or **m**.
  - **Hold - The y-axis zoom is controlled manually but vertical panning on the** plot canvas is disabled, protecting against accidental changes to the y-axis zoom. All zoom controls are enabled. This mode can be accessed using the hot-keys **e** or **h**.
  - **Fit Data - The y-axis zoom is changed dynamically as the x-axis zoom changes** to show all the data on the vertical axis. All vertical zoom controls are disabled. This mode can be accessed using the hot-keys **w** or **f**.
  - **Auto - The y-axis zoom is changed dynamically as the x-axis zoom changes.** When zooming in close on the x-axis the y-axis will be zoomed in as well to automatically improve the detail of the vertical axis in proportion to the horizontal axis. This mode is only available on Heat Map, Alignment Pile-up, Value, and Variant Map plots. It can be accessed using the hot-key **q**.

  By default the y-axis zoom mode is either Fit Data or Auto depending on whether the plot type supports Auto zoom. If the zoom mode is changed to either Manual or Hold zoom, adjustments to the y-axis will become possible. In particular this will make click-and-drag operations on the y-axis scale function. It will also cause the y-axis zoom slider to appear along the right edge of the plot (assuming the plot is tall enough). The y-axis zoom slider is analogous to the x-axis zoom slider provided by the Zoom tool bar. Drag the slider toward the minus (**-**) button to zoom out, and toward the plus (**+**) button to zoom in. The buttons zoom in (**+**) or out (**-**) in steps.
- **Mouse Anchor**: A vertical line that is drawn over all plots in the plot view can be set by right clicking on a plot and selecting **Place Mouse Anchor** or by pressing **Ctrl** - **‘**. This can be used as a visual reference, a bookmark or a ruler. To jump to the current mouse anchor press **‘** (also the tilde key). To clear the current mouse anchor right click and select **Clear Mouse Anchor** or press **Alt** - **‘**. When a mouse anchor is set, the distance between it and the current mouse pointer location will be displayed in the status bar. Note that the precision of the measurement is dependent on the current zoom.
- **Searchable**: The data sources for a plot can be set to be searchable by right clicking on the plot and selecting **Searchable**. This will add the plot's data sources to the list that gets searched when text is typed into the location bar.

- **Feature Labels**: Feature labels on some plots can be hidden or shown by right clicking on the plot and selecting the **Feature Labels** in the **Show** menu.
- **Reload**: A plot can be reloaded by right clicking and selecting **Reload**.
- **Hide**: A plot can be hidden by hovering over the upper right corner of a plot and clicking on the minimize icon, or by right clicking on a plot and selecting **Hide**. To restore a hidden plot re-check it in the plot tree or right-click on empty space in the plot view and re-check its entry in the **Show Plot/Group** menu.
- **Delete**: A plot can be deleted by selecting it in the plot view and pressing the delete button on a keyboard or right clicking on the plot and selecting **Delete** from the menu.
- **Add**: Plots can be added by right clicking on empty space in the plot view (anywhere that is not a plot) and selecting **Add**. Additional items can be added to value plots by right clicking on the plot and selecting **Add Item(s)**.
- **Resize Plots**: A plot can be resized by clicking on the grab bar underneath it and dragging up or down to the desired size. Plots have a minimum height but there is practically no maximum.

## 4.6 Plot Tree

The plot tree contains all of the data sources that have been added to the GenomeBrowse window. This view provides access to the settings of each plot, plot group, or plot item, as well as allowing easy access to rearranging, deleting, hiding, showing or adding new plots and items.

Nodes in the plot tree all have basic features in the right click menu. These features include editing the title, and hiding or deleting the node.

Plot nodes also offer the reload action, and some types of plot nodes can be set as searchable, which causes all the data sources in the plot to be searched when text is typed into the location bar.

Value plots may contain multiple items. Right clicking on a value plot provides options to add additional data sources or lines to the plot.

BAM or Read Alignment Sources are represented by two plots in a plot group. Plots cannot be added to read alignment plot groups.

Clicking on a node in the plot tree also updates the control panel (if visible) to show controls for that node.

## 4.7 Control Panel

Contains the controls for the selected plot, plot group, or plot item. See the specific plot type for information on the available controls. (*Options for Specific Data Source Types*)

## 4.8 Console

Displays information about plots and their data. Clicking on a plot, a node in the plot tree, or a feature within a plot will update the information in the console. Please see the specific plot type for descriptions of the information that can be displayed in the console.

Checking the **History** check box will cause information for multiple recent clicks to be accumulated in the console by inserting the most recent information at the top. Clicking the **Clear** button will empty all information from the console.

## 4.9 Feature List

The feature list is a tabular view of all data fields available for each feature in a data source. Up to 1000 features can be read at a time, and can be read from either the start of the genome assembly, or the current zoom.

A different data source available from the current GenomeBrowse window can be selected from the drop-down list in the upper left corner of the feature list. This will cause the table to be re-populated with features from the selected source.

Selecting the **All** radio button will update the feature list with features from the start of the genome assembly. Selecting the **Zoom** radio button will update the feature list with features from the current x-axis zoom. Selecting one or more features in the list will zoom the view to fit the selected feature(s).

The list of features can be sorted by clicking on the field names (column headers). Once for ascending order, twice for descending order.

Selected row(s) can be copied to the clipboard. Click on a row or use Ctrl+left-click or Ctrl-A to select multiple or all rows. Then right-click and select **Copy Selected Row(s)** to Clipboard or **Copy Selected Row(s) with Headers**. This table can then be pasted into a spreadsheet program or text file.

The mouse anchor can be set to the center of any feature in the table by right clicking on its row and selecting **Place Mouse-Anchor At ...**

If there are more than 1000 features available, the **read more** link will be shown in the lower left corner. Up to 1000 more features will be added to the table each time it is clicked.

## 4.10 Open a New Project from an Existing Project

To open a new project from an existing project go to **File > New Project...**. Enter in the new project name and select the desired genome. Click **OK**.

You may be asked to save the current project and/or to download the reference sequence before the new project is created.

## 4.11 Download Window

The download window lists all recent and active downloads. Downloads will continue whether or not this window is open. If the window is closed downloads will continue in the background. If GenomeBrowse is closed the downloads will resume when GenomeBrowse is reopened.

## 4.12 General Options for GenomeBrowse

The options dialog contains various controls for setting global GenomeBrowse options. These options will apply to all GenomeBrowse windows in all projects. Available controls include a choice for where new plots are added to the plot view, which axes position tracking is enabled on, download save target, and color customization.

## 4.13 Link Server

Under the Program tab, there is an option to run a link server. This will be mostly compatible with the IGV link server (and on the same port by default). See the section titled Link to Load Data under IGV's documentation.

GenomeBrowse also has a registered protocol on Windows and Mac for launching and adding sources. On Linux, a Link Server is the most reliable way to remotely control a running instance.

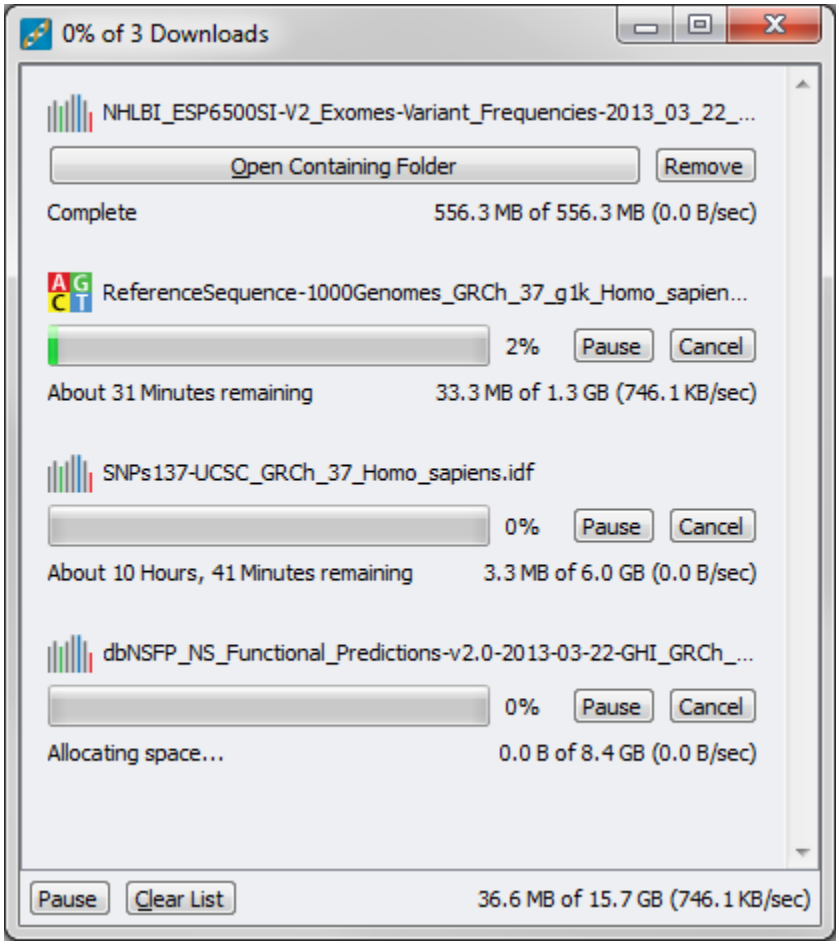



Figure 4.2: Download Window showing 1 finished download and 3 active downloads

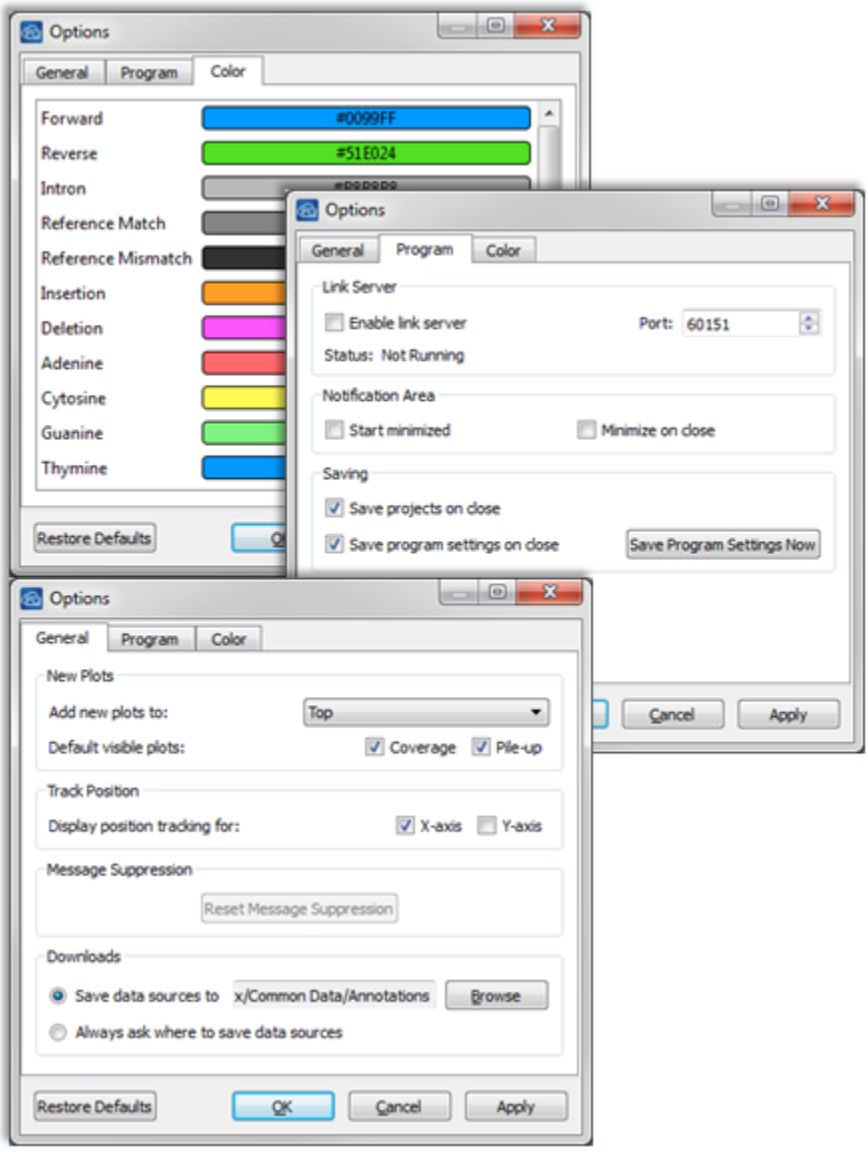



Figure 4.3: General options window

# CHAPTER FIVE

# OPTIONS FOR SPECIFIC DATA SOURCE TYPES

The options dialog contains various controls for setting global GenomeBrowse options. Available controls include a choice for where new plots are added to the plot view, which axes position tracking is enabled on, download save target, and color customization.

The various types of data sources that can be visualized in GenomeBrowse are listed below.

## 5.1 Value Plot

The Value Plot is a plot of the genomic coordinates on the X-axis and the value from the field plotted on the Y-axis.This plot type is used to look for trends associated with genomic position or the values for one variable.

### Plot Description

Value plots can be drawn from any field consisting of numeric data from an annotation source or file that can be visualized in GenomeBrowse.

### Controls

#### Display Tab for Plot Container

On the **Display** tab, the controls include:

- Labels
- Value
- Connector
  - Type
  - Size
- Smoothing
  - Type
  - Window radius
- Y-Range

The **Labels** control provides the ability to change the data field that provides the feature labels that are drawn on the plot. More labels will appear the closer the plot is zoomed in.

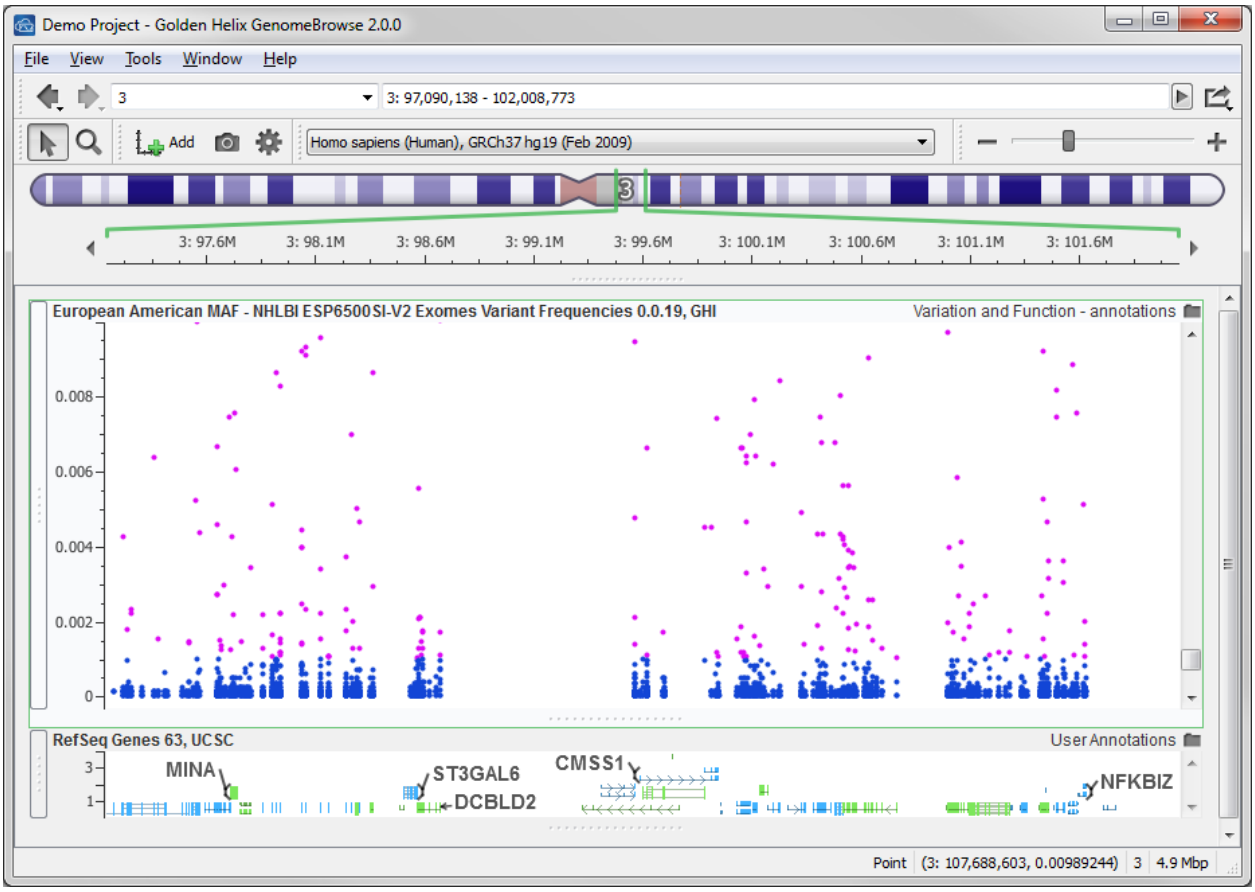



Figure 5.1: Value Plot

The **Value** control allows the data field that provides the data points for the plot to be changed. By default this value is the field selected when the plot was created. If there are multiple items in the plot, changing the value control at the top level will set the value for each plot item to the same value.

The **Connector** control allows for connecting the data points drawn in the plot. Options include:

- **None**: (Default) Data points are not connected
- **Drop Line**: Connect all data points to the x-axis with a vertical line.
- **Line**: Connect all data points with a line
- **Left Step**: Connect all data points by first stepping vertically and then horizontally to the next data point.
- **Mid Step**: Connect all data points by placing the vertical step at the midpoint of the horizontal step. I.e. step half-way horizontally, then the full vertical step, then the remaining half-step horizontally.

The connector width is specified in the integer selector box beside the Connector control option. This value indicates the thickness of the connector lines.

The **Smoothing** control allows for smoothing data based on a specified range of values. The smoothing options include:

- **None**
- **Mean Symmetric**
- **Median symmetric**
- **Mean Asymmetric**
- **Median Asymmetric**

The window radius value is specified in the integer selector box beside the Smoothing control option. This value indicates the number of points to use for smoothing on either side of the point being smoothed. For example, a window value of 2 replaces each point with a 5 point median or mean value.

The difference between Symmetric and Asymmetric smoothing is how the boundary cases are handled.

For the other controls that are common with most plot types, see *Display Controls*.

### Style Tab

On the **Style** tab, the controls include:

- Style By
  - Field
  - Save
- Style
  - Color
  - Shape
  - Size
- Restyle
  - Method
  - Various Styling Options

The **Style By** control enables the user to select a single dimension in which colors can be used to discriminate between complementary data categories. A dimension can be selected by clicking the "Style By" button. Fields available from the source that can be used for styling are available in the list. Selecting a numeric field will enable the **Cutoff** control to specify a threshold value to use for splitting the style of the data. To save the style, click the **Save** button.

The **Style** list allows for the specification of the style of the data drawn in the plot. There are controls for changing, the color, shape and size of the data points. If a field is specified to **Style By** then there will be controls for each group as determined by the field and threshold selected.

To change the shape and size of all categories, first change the shape and size with *Style By = None* then set the *Style By* to the desired field such as *Chromosome*.

The **Restyle** control allows for all styles in all selected plot items to be recolored or reshaped incrementally. When a single plot is selected, it has the same effect as selecting all of its items. The available methods include:

- **From Current**: Uses the first style as the starting point and increments the colors and shape by the specified amount for each remaining style. An increment of 0 sets all of the colors and/or shapes to the starting values.
- **Color Gradient**: Set the starting color and then specify the Hue, Saturation and Value increments.
- **Color From**: Set the starting color then specify the color increment.
- **Shape From**: Set the starting shape then specify the shape increment.

### Filter Tab

A **Filter** can be used to control which features are drawn in the plot.

To add a filter either click on **Insert** or right-click anywhere in the Filter list box and select **Insert**.

Please see *Filter Controls* for more information.

### Layout Tab

On the **Layout** tab general plot controls can be changed. See *Layout Controls* for more information.

### Add Tab

On the **Add** tab there are buttons for adding additional items or line items to the plot. Clicking on the **Add Item(s)** button opens up the data source library add data sources dialog. Clicking on the **Add Line Item(s)** button opens up the *Line Parameters* dialog to add a horizontal, vertical or line with a slope and intercept. See *Line Items* for more information.

### Display Tab for Plot Items

On the **Display** tab, the controls include:

- Labels
- Value
- Connector
- Smoothing

The **Labels** control provides the ability to change the data field that provides the feature labels that are drawn on the plot. More labels will appear the closer the plot is zoomed in.

The **Value** control allows the data field that provides the data points for the item to be changed. By default this value is the field selected when the item was created.

The **Connector** control allows for connecting the data points drawn in the plot. Options include:

- **None**: (Default) Data points are not connected
- **Drop Line**: Connect all data points to the x-axis with a vertical line.
- **Line**: Connect all data points with a line
- **Left Step**: Connect all data points by first stepping vertically and then horizontally to the next data point.
- **Mid Step**: Connect all data points by placing the vertical step at the midpoint of the horizontal step. I.e. step half-way horizontally, then the full vertical step, then the remaining half-step horizontally.

The connector width is specified in the integer selector box beside the Connector control option. This value indicates the thickness of the connector lines.

The **Smoothing** control allows for smoothing data based on a specified range of values. The smoothing options include:

- **None**
- **Mean Symmetric**
- **Median symmetric**
- **Mean Asymmetric**
- **Median Asymmetric**

The window radius value is specified in the integer selector box beside the Smoothing control option. This value indicates the number of points to use for smoothing on either side of the point being smoothed. For example, a window value of 2 replaces each point with a 5 point median or mean value.

The difference between Symmetric and Asymmetric smoothing is how the boundary cases are handled.

For the other controls that are common with most plot types, see *Display Controls*.

### Style Tab for Plot Items

- Style By
  - Field
  - Save
- Style
  - Color
  - Size
  - Shape
- Restyle
  - Method
  - Various Styling Options

The **Style By** control enables the user to select a single dimension in which colors can be used to discriminate between complementary data categories. A dimension can be selected by clicking the "Style By" button. Fields available from the source that can be used for styling are available in the list. Selecting a numeric field will enable the **Cutoff** control to specify a threshold value to use for splitting the style of the data. To save the style, click the **Save** button.

The **Style** list allows for the specification of the style of the data drawn in the plot. There are controls for changing, the color, shape and size of the data points. If a field is specified to **Style By** then there will be controls for each group as determined by the field and threshold selected.

To change the shape and size of all categories, first change the shape and size with *Style By = None* then set the *Style By* to the desired field such as *Chromosome*.

The **Restyle** control allows for styles in all selected plot items to be recolored or reshaped incrementally. The available methods include:

- **From Current**: Uses the first style as the starting point and increments the colors and shape by the specified amount for each remaining style. An increment of 0 sets all of the colors and/or shapes to the starting values.
- **Color Gradient**: Set the starting color and then specify the Hue, Saturation and Value increments.
- **Color From**: Set the starting color then specify the color increment.
- **Shape From**: Set the starting shape then specify the shape increment.

### Filter Tab for Plot Items

A **Filter** can be used to control which features are drawn in the plot.

To add a filter either click on **Insert** or right-click anywhere in the Filter list box and select **Insert**.

See *Filter Controls* for more information.

## Data Console

Clicking on a plot container for value plots will print out information about the data source including all of the fields in the source.

Clicking on a data point in the plot will result in the value, the label for the feature as well as any applied styling.

## Line Items

Line items can be added to value plots by either right clicking on the plot and selecting **Add Line Item(s)** or by clicking on the **Add Tab** for the plot container and clicking on the line item(s) button.

The **Line Parameters** dialog includes the following controls:

- Line Type Selection
- Slope/Intercept
- Color
- Width

The **Line Type Selection** control allows of of the three available line types to be selected. There are three different types of lines that can be added: Horizontal, Vertical and Slope/Intercept.

The **Slope/Intercept** controls will change depending on the selected line type.

- A Horizontal line is specified by a numeric y-intercept value.

- A Vertical line is specified by genomic coordinates (Chr#:Position) or by an x-intercept value.
- A Slope/Intercept line is specified by a numeric slope and numeric y-intercept value.

The **Color** control sets the color or the line.

The **Width** control sets the thickness of the line. This thickness is absolute, so it will be the same size on screen regardless of the current zoom.

### Special Features

If a numeric feature has both a chromosome start and stop position then the value will be drawn as an interval. In this case it is recommended that the shape drawn for the values be changed to one of the shapes that stretches better such as a rectangle or plus sign.

## 5.2 Variant Maps

A Variant Map provides a visual interpretation of genotypic data for one or more samples. Variant maps can be created from variant call format (VCF) files or annotation sources with sample level variant calls.

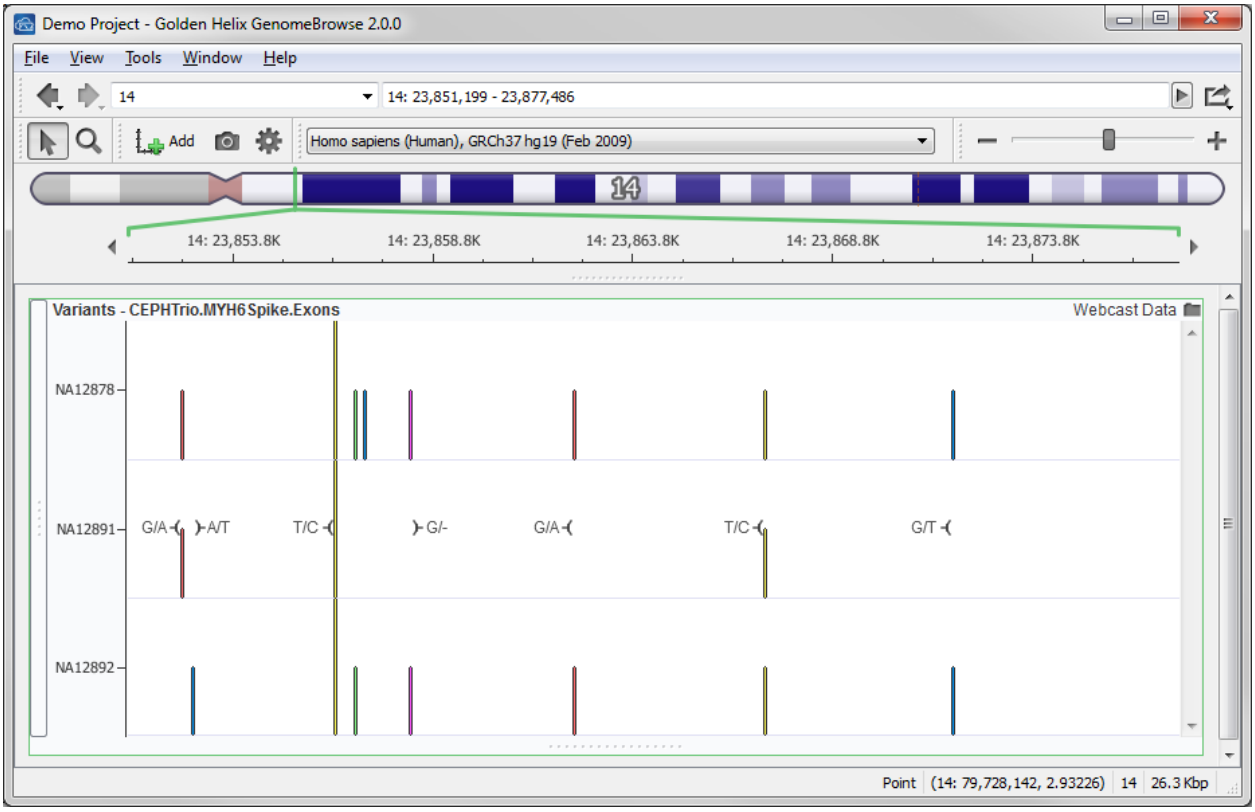



Figure 5.2: Variant Map Plot

## Plot Description

The variants for each sample in a variant map are displayed in rows along the genomic x-axis. For the cases of deletions, and substitutions, the variant may be drawn to cover multiple bases.

Variant data that matches the reference (non-variant) is downplayed in visual significance at close zooms. This makes the actual variants more obvious. Optionally, reference allele matches can be hidden at all zoom levels.

For wide-zoom views, the variants are displayed as a gray scale density plot, indicating the locations of variant data.

When zoomed in close enough, the variants are colored according to the GenomeBrowse global color options. Each sample's row is split vertically to allow for indication of zygosity. A variant of a single color therefore indicates a homozygous alternate variant call, whereas a two color variant indicates a heterozygous variant call. Missing calls are displayed as question marks (*?*) in light gray. In this way missing variants and half-called variants can be indicated as well. By default variants are labeled with the variant call or "genotype".

- Single Nucleotide Variations (SNVs) variants are colored as a single base.
- Insertions are drawn with a zero-width I bar at the location of the inserted base(s).
- Deletions are represented as a magenta solid block representing the missing base(s).
- Substitution variants are drawn with each base indicating the variant alleles at that position.

The y-axis corresponds to the **Sample Label**.

## Controls

### Display Tab

On the **Display** tab, the controls include:

- Reference Alleles
- Labels

The **Reference Alleles** control will draw a line for samples that have the reference allele(s) at close zooms if checked. Otherwise, reference allele matches are not drawn.

The **Labels** control allows for selection of the data field that provides labels for the marker or column of variants when zoomed in close enough.

For the other controls that are common with most plot types, see *Display Controls*.

### Filter Tab

On the **Filter** tab, there are two filter boxes. The top filter box can be used to filter variants based on field data, such as a variant level quality score or another *INFO* field in a VCF file. The bottom filter box can be used to filter samples based on sample names.

To add a filter either click on **Insert** or right-click anywhere in the appropriate Filter list box and select **Insert**.

Please see *Filter Controls* for more information.

### Group By Tab

If there is sample field meta data available from the plot source that can be used for grouping, on the **Group By** tab, these fields can be selected. If a numeric field is chosen a cutoff or group split-point can be specified as well. The resulting groups will be displayed as rows in the box below. Each group can be hidden or shown using the check box, and its color can be changed by clicking on the color button and choosing a new color.

#### Layout Tab

On the **Layout** tab general plot controls can be changed. See *Layout Controls* for more information.

### Data Console

Clicking a marker or column of variants in a variant map displays information about the marker from the data source in the data console. The information will include, if available, the marker label, a list of alleles found at the marker, all associated data fields from the data source, and a list of each sample's genotype at the marker.

When zoomed in close enough, a particular sample's variant can also be clicked. The information displayed is the same as for the marker but will also include the clicked genotype near the top of the report.

## 5.3 Linkage Disequilibrium

The LD plot is a triangular heat map of the LD statistics ($D'$, $R^2$) between pairs of variants across the genome in a variant annotation source.

**Note:** Linkage Disequilibrium (LD) is the non-random association of alleles in a population. LD is useful during analysis to identify linkage relationships of interesting variants. LD can be calculated from most sources with more than one sample. In particular, multiple sample VCF files.

### Plot Description

The variants are the individual pentagons along the spine of the LD graph. As LD is a pairwise computation, LD is not available at large zooms. At large zooms the plot will show the density of the variants in a "rug plot". To see LD values zoom into regions of interest, the maximum zoom is $2 * 10^6 - 1$.

### Controls

#### Display Tab

On the Display tab, the controls specific to LD are:

- Statistic
- Labels

The **Statistic** control gives the option to either display the LD $R^2$ or $D'$ statistic in the plot.

The **Labels** control allows for selection of the data field that provides labels for the markers.

See *Display Controls* for more information on controls available for all plot types.

#### Filter Tab

A **Filter** can be used to control which features are drawn in the plot.

To add a filter either click on **Insert** or right-click anywhere in the Filter list box and select **Insert**.

Please see *Filter Controls* for more information.

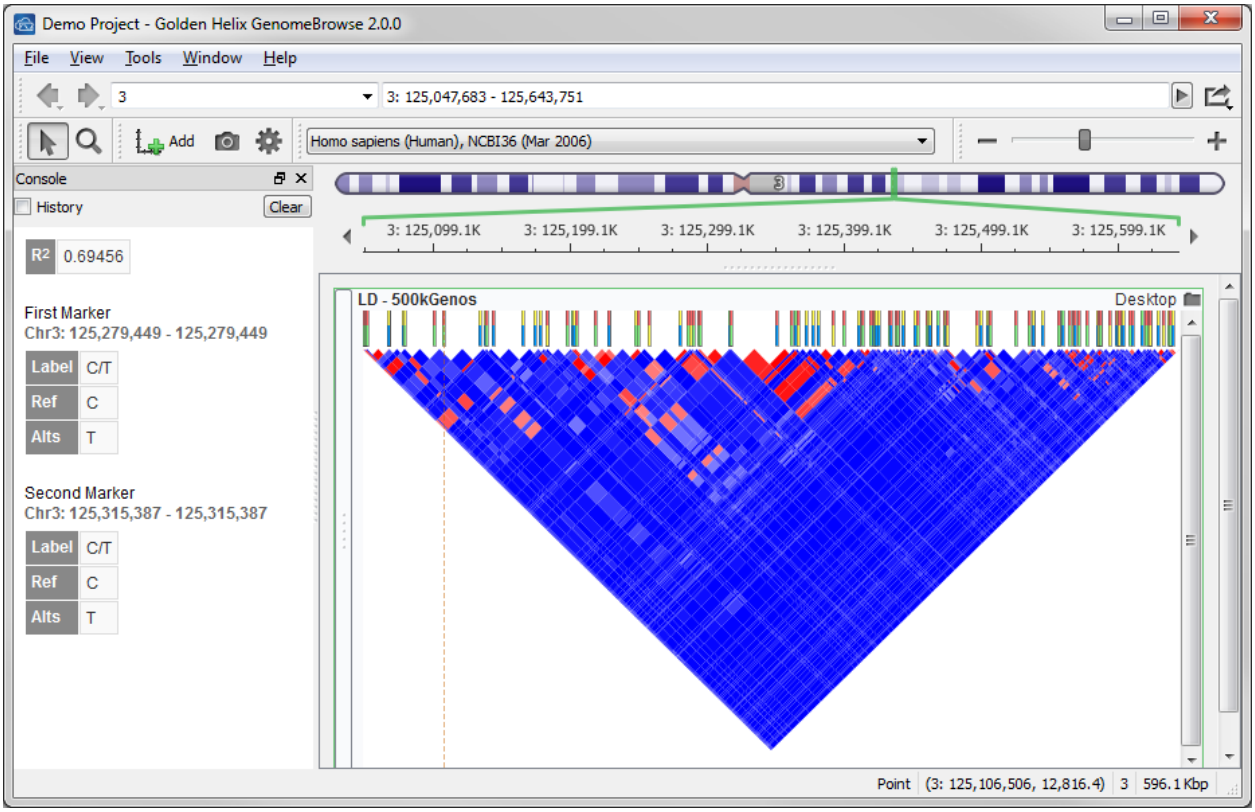



Figure 5.3: Linkage Disequilibrium for region in Chromosome 3

### Layout Tab

On the **Layout** tab general plot controls can be changed. The control specific to LD is:

- Invert

The **Invert** control allows for the specification of the location of the spine of the visualization to be either along the top or the bottom of the plot. This control is checked by default, corresponding to the spine at the top of the plot.

See *Layout Controls* for more information on controls available for all plot types.

## Data Console

The Data Console provides a detailed html formatted text output. There are 2 elements comprising an LD graph:

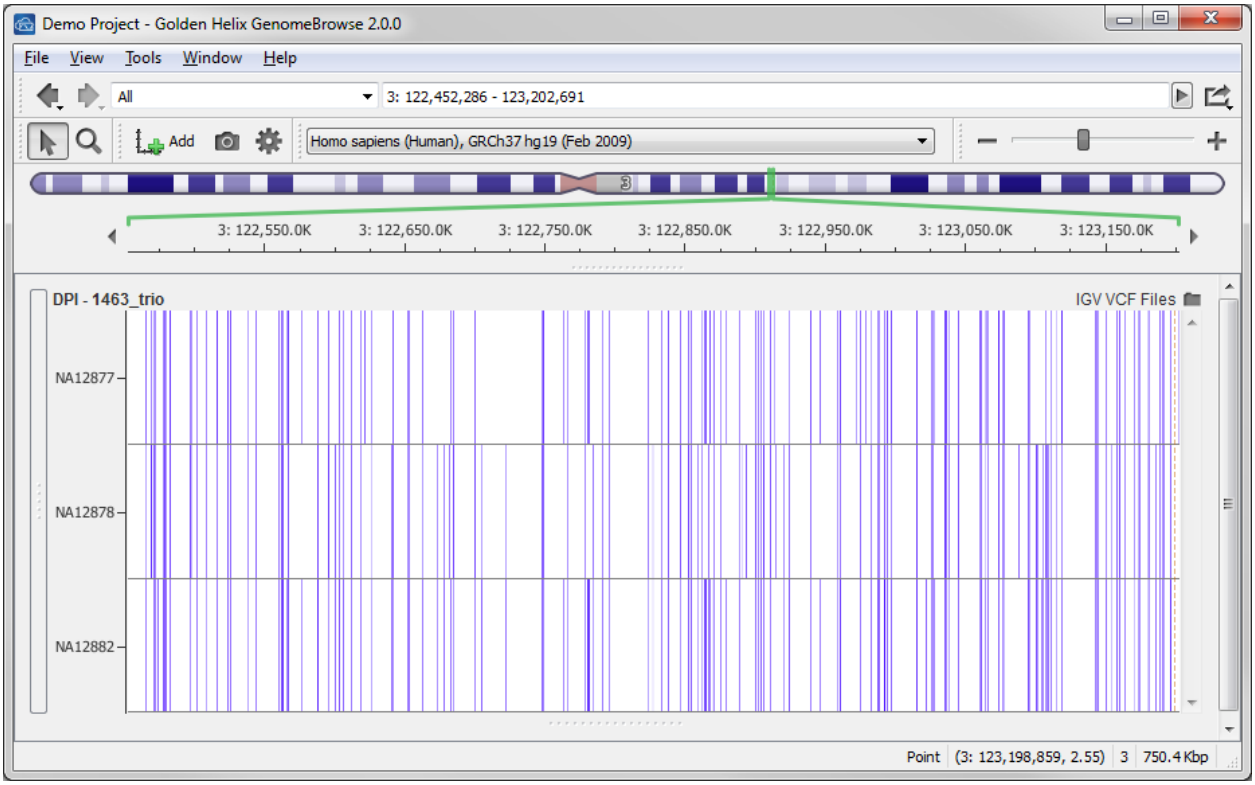
- LD Value
- Summary information for each variant or marker

The summary information includes the name of the variant or maker, and the allele information for each contributor to the LD calculation.

## 5.4 Heat Maps

A Heat Map is an intensity plot of numeric values from a data source containing multiple samples such as a multi-sample VCF file. The X-axis consists of the variants from the source. The Y-axis consists of sample labels.

Heat maps are useful to find non-random patterns in the data, particularly for read depth or quality scores.




Heat map of DPI (Basecall depth associated with indel) from a VCF file

## Plot Description

The Y-axis corresponds to the **Sample Label**.

## Controls

### Display Tab

On the **Display** tab, the controls include:

- Aggregation [Method]
- Auto-Compute Color Values
- Color Values
- Presets

The **Aggregation** control allows for the specification of how pixels are aggregated if there are not enough pixels available to draw all of the data selected. A decision needs to be made as to what value to show for each pixel. The options available are Mean, Minimum, Maximum and Extreme.

The **Auto-Compute Color Values** check box indicates whether the colors should be assigned based on automatically computed values. The method for auto-computing the values for the colors is as follows:

$$
\begin{aligned}
n &= \text{number of color values} \\
k &= \frac{6}{n-1} \\
\bar{x} &= \text{mean of all numeric data in the source} \\
\sigma^2 &= \text{variance of the numeric data} \\
\sigma &= \sqrt{\sigma^2} = \text{standard deviation} \\
i &= \text{index of color value; } i \in [0, n) \\
a_i &= i - \frac{n-1}{2} \\
z_i &= a_i * k \\
SP_i &= \bar{x} + z_i * \sigma
\end{aligned}
$$

The **Color Values** box lists all of the colors and the values associated with the colors (whether manually specified or auto-computed. To edit the colors, double-click on the color box for the color value. To edit the value, uncheck **Auto-Compute Color Values** and right-click on a value. New points can be added or values can be deleted from the right-click menu as well.

The **Presets** buttons are different preset color combinations to use for coloring the heat map. The default color combination is *Gain/Loss*.

For the other controls that are common with most plot types, see *Display Controls*.

### Filter Tab

On the **Filter** tab, there are two filter boxes. The top filter box is to filter features based on a feature field in the source of the data. The bottom filter box is to filter features based on a sample field in the source of the data.

To add a filter either click on **Insert** or right-click anywhere in the appropriate Filter list box and select **Insert**.

Please see *Filter Controls* for more information.

#### Group By Tab

If there is sample field meta data available from the plot source that can be used for grouping, on the **Group By** tab, these fields can be selected. If a numeric field is chosen a cutoff or group split-point can be specified as well. The resulting groups will be displayed as rows in the box below. Each group can be hidden or shown using the check box, and its color can be changed by clicking on the color button and choosing a new color.

#### Layout Tab

On the **Layout** tab general plot controls can be changed. See *Layout Controls* for more information.

### Data Console

Clicking in a Heat Map value produces the number of features clicked on as well as the mean, minimum and maximum values in the heat map bin. When there is only one feature clicked on, the value for all three statistics will be the value of the feature. The sample used for computing the three summary statistics is also listed in the data console.

## 5.5 BAM File Type

Please see *Read Alignment Sources* for information on visualization of data from this file type.

### File Information

BAM files can be added into a GenomeBrowse window by either selecting it from the Add Data Sources dialog (Add dialog) or by dragging the file into an open GenomeBrowse window.

Before visualization of the data GenomeBrowse must first compute index (BAI) and coverage (COVIDF) files for the BAM. Once the index file has been computed then visualization of the data will be available only for zoomed in regions of the plot until the coverage file is done being computed. If a BAI file was provide GenomeBrowse will not generate another index but will instead use the existing file if it is saved in the same directory as the BAM file.

BAM files need to be sorted to be loaded into GenomeBrowse. Additionally to be able to compute index and coverage files GenomeBrowse needs to be able to identify the reference sequence that corresponds to the genome build associated with the data in the BAM.

GenomeBrowse will use the BAM header information to identify the correct reference sequence by matching the chromosome names and lengths from the header exactly to that information in an available genome assembly file, see *Genome Assemblies* for more information on assembly files. Once the correct reference sequence is identified for the BAM you must download a local copy of the reference sequence for GenomeBrowse to use in computing the index and coverage files. Please see *downloadingData* for information on downloading the correct reference sequence.

If the BAM file is unsorted or if the header is not formatted correctly, it is recommended that a third party tool such as SAMtools, http://samtools.sourceforge.net, be used to edit the file into the correct format.

## 5.6 BED File Type

BED files can either contain interval data or gene information. Please see *Interval Sources* for information on interval BED files and see *Gene Sources* for information on gene sources.

### File Information

BED files need to be sorted to be loaded into GenomeBrowse, if a file is not sorted, it is recommended that a text editor or spreadsheet editor such as MS Excel be used to get the data in the correct order.

Once the file has been sorted, it can be added into a GenomeBrowse window by either selecting it from the Add Data Sources dialog (Add dialog) or by dragging the file into an open GenomeBrowse window. Adding a BED file will also compress and index the file. Once the file has been compressed and indexed then only the compressed and indexed files are needed for visualization in GenomeBrowse.

## 5.7 Cytoband Sources

Sources contain cytoband information including Giemsa stain results.

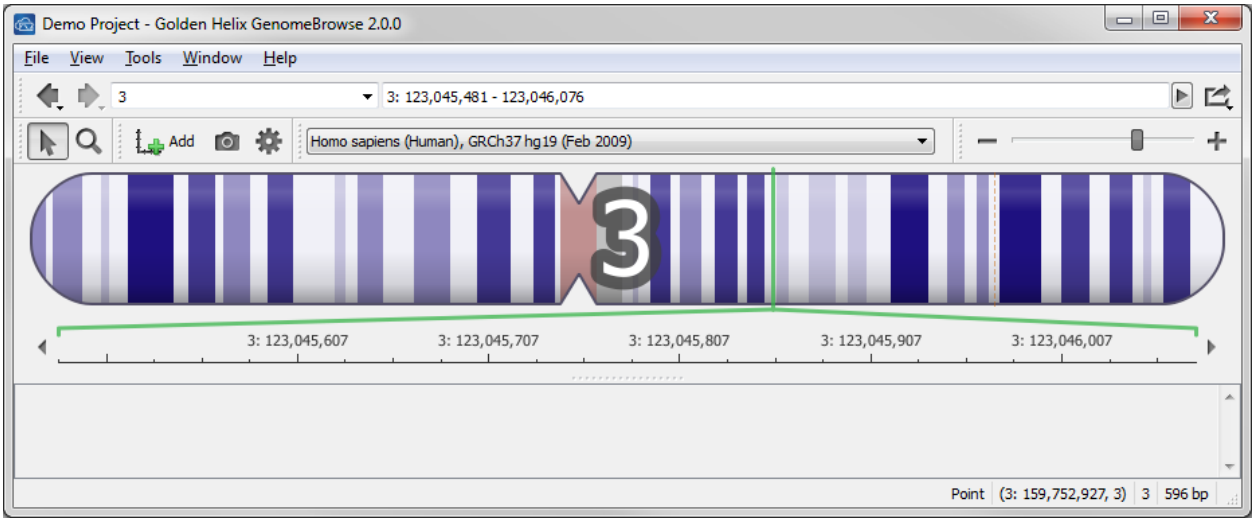



Figure 5.4: Cytoband Plot

### Plot Description

The plot displays a karyotype view of the cytobands (cytogenetic bands) within each chromosome.

### Controls

#### Display Tab

On the **Display** tab general plot controls can be changed. See *Display Controls* for more information.

#### Layout Tab

On the **Layout** tab general plot controls can be changed. See *Layout Controls* for more information.

### Special Features

If one of the human assemblies is chosen for the genome build then the appropriate cytoband source shipped with the software will be set as the domain view plot by default.

If a cytoband source is available for your species/build it is recommended that it be set as the domain view plot.

To set a plot as the domain view plot, right-click on the plot and select **Set as Domain View Plot**.

## 5.8 Interval Sources

Sources contains features of variable width and the may have multiple data fields associated with each feature interval.

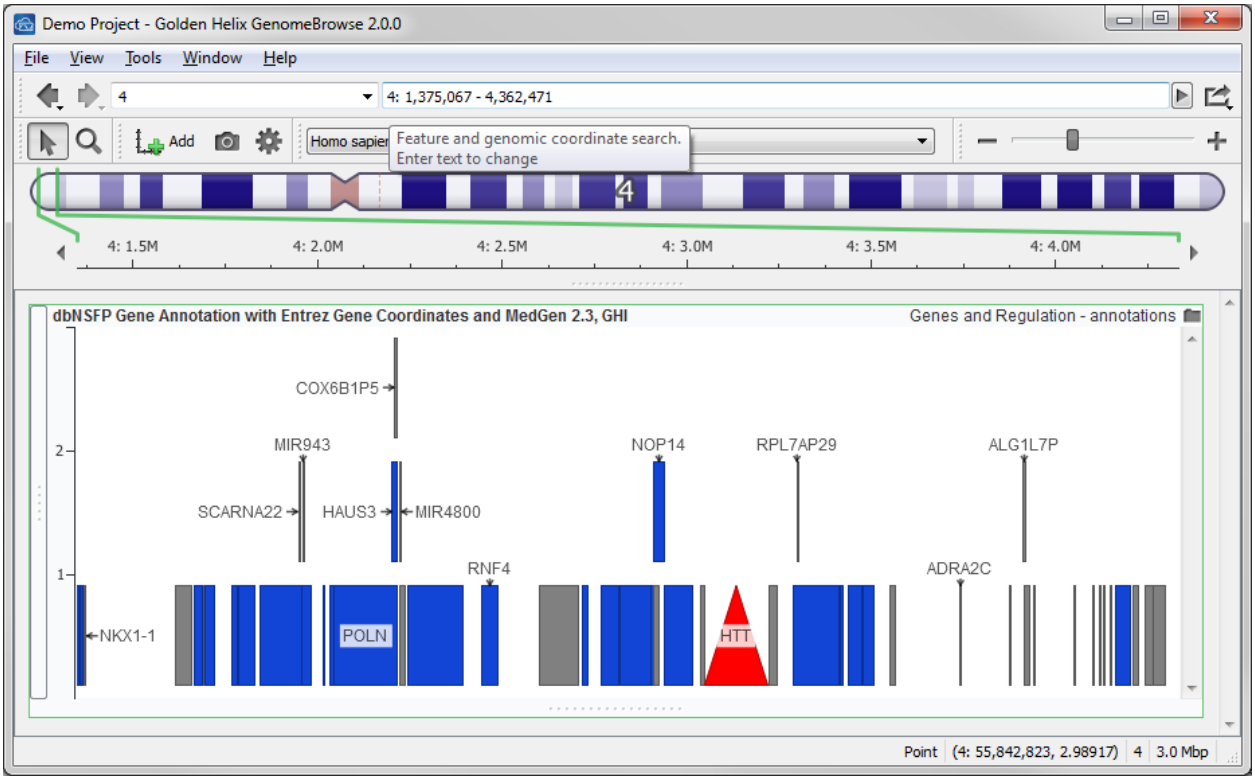



Figure 5.5: Interval Source Plot

### Plot Description

Displays information for genomic intervals. There can be multiple overlapping intervals, so they will be stacked on the vertical axis to avoid visual overlap. The style of each interval can be specified to convey meaning. For instance, the *dbNSFP Gene Annotation* source contains information about genes with nonsynonymous variants. The color of each interval is based on the *Inheritance Type* information in the source.

## Controls

### Display Tab

On the **Display** tab, the controls include:

- Labels

The **Labels** control provides the ability to change the data field that provides labels for the features. Labels will be displayed when zoomed in close enough.

For the other controls that are common with most plot types, see *Display Controls*.

### Style Tab

On the **Style** tab, the controls include:

- Style By
  - Field
  - Save
- Style
  - Color
  - Shape
- Restyle
  - Method
  - Various Styling Options

The **Style By** control enables the user to select a single dimension in which colors can be used to discriminate between complementary data categories. A dimension can be selected by clicking the "Style By" button. Fields available from the source that can be used for styling are available in the list. Selecting a numeric field will enable the **Cutoff** control to specify a threshold value to use for splitting the style of the data. To save the style, click the **Save** button.

The **Style** list allows for the specification of the style of the data drawn in the plot. There are controls for changing, the color and shape of the data points. If a field is specified to **Style By** then there will be controls for each group as determined by the field and threshold selected.

The **Restyle** control allows for styles in all selected plot items to be recolored or reshaped incrementally. The available methods include:

- **From Current**: Uses the first style as the starting point and increments the colors and shape by the specified amount for each remaining style. An increment of 0 sets all of the colors and/or shapes to the starting values.
- **Color Gradient**: Set the starting color and then specify the Hue, Saturation and Value increments.
- **Color From**: Set the starting color then specify the color increment.
- **Shape From**: Set the starting shape then specify the shape increment.

### Filter Tab

A **Filter** can be used to control which features are drawn in the plot.

To add a filter either click on **Insert** or right-click anywhere in the Filter list box and select **Insert**.

Please see *Filter Controls* for more information.

### Layout Tab

On the **Layout** tab general plot controls can be changed. See *Layout Controls* for more information.

## Data Console

The data console contains information on the feature from the data source as well as the genomic position of the interval.

## 5.9 Gene Sources

Sources contain information on genes. Most gene sources also include coding region information as well.

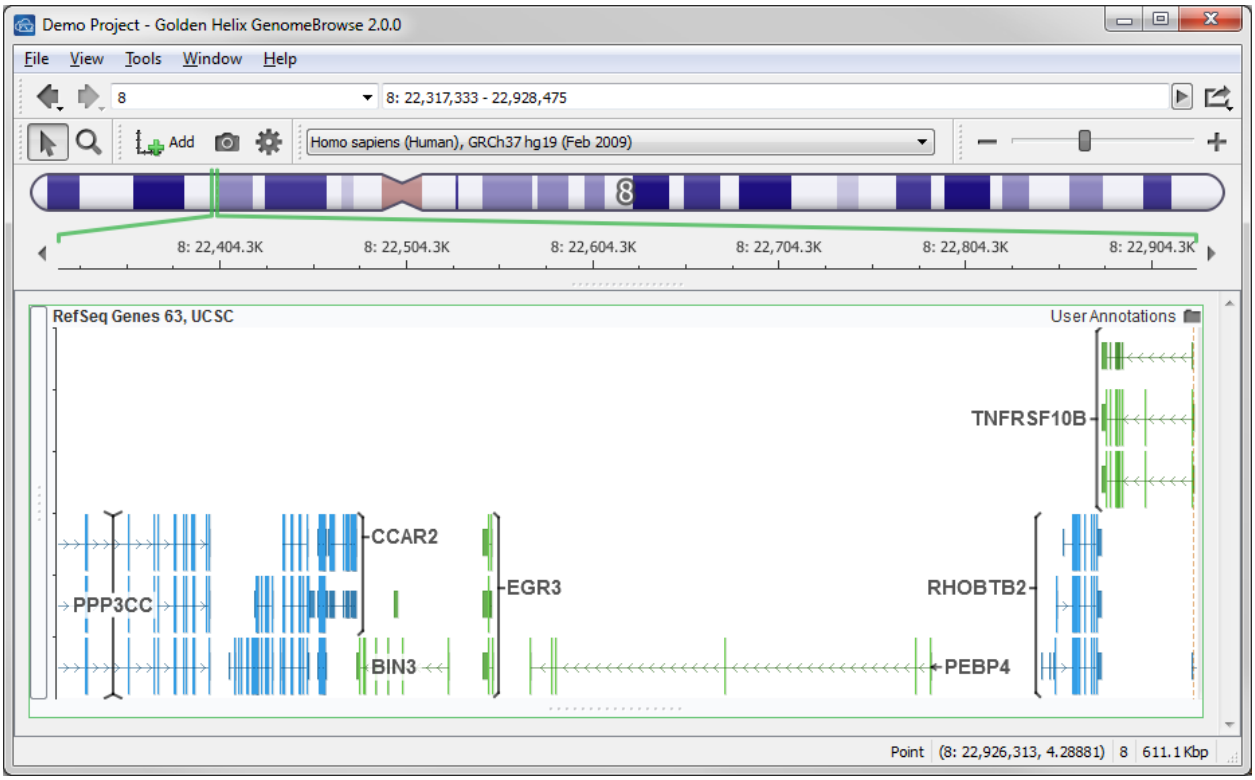



Figure 5.6: Gene Source Plot

## Plot Description

Gene sources draw genes as intervals using rectangles to indicate exons and a directional line to indicate introns. Genes on the forward strand are colored blue while genes on the reverse strand are colored green by default. UTR regions are a slightly darker shade of blue or green depending on the strand orientation of the gene.

Feature labels are drawn dynamically depending on the zoom range. Whenever possible gene names are drawn at larger zoom ranges. As the zoom range becomes smaller exon labels become visible and hover labels indicating the codon amino acids are available by mousing over the alternating light/dark codon segments of exon regions.

For mitochondrial codon amino acids, the alternate MT codon table is used to label gene features, see: https://www.mun.ca/biology/scarr/MGA2-03-28_mtDNA_code.jpg

## Controls

### Display Tab

On the **Display** tab, the controls include:

- Genes [Draw Mode]

The **Genes** control allows the user to specify the draw mode for genes. Options include:

- *Auto*: Depending on the zoom range and the vertical height of the gene plot, either show all genes and transcripts or collapse all transcripts into one gene interval drawing.
- *Compact*: Always try to collapse all transcripts into one gene interval drawing. It will not always be possible to collapse all transcripts, but the preference will be to collapse regardless of the zoom range and plot height.
- *Expanded*: Always draw all transcripts as separate gene intervals regardless of the zoom region/plot height.

The other two controls are available for most plots, see *Display Controls* for more information.

### Filter Tab

A **Filter** can be used to control which features are drawn in the plot.

To add a filter either click on **Insert** or right-click anywhere in the Filter list box and select **Insert**.

A basic filter can be added based on the fields of the source. A custom filter can also be specified using the muParserX syntax.

### Layout Tab

On the **Layout** tab general plot controls can be changed. See *Layout Controls* for more information.

## Data Console

Clicking on a gene feature will display information about the gene from the gene source in the data console. The information will include, if available, the gene and transcript names, the strand, if it is coding or not, the pre-spliced size, the post-spliced size and all of the exon region information in a tabular form.

The Exon Table includes the exon number, size, start relative to the start of the chromosome, the start relative to spliced RNA, and start relative to coding RNA.

Also included are hyperlinks to gene databases if they were included in the gene source. The hyperlinks either search the databases by gene name or transcript name. A web browser and active internet connection is required for this feature.

# 5.10 Read Alignment Sources

Sources contain reads, generally short nucleotide sequences, typically aligned to a reference genome. Currently only binary sequence alignment/map (BAM) files, generally from a secondary analysis pipeline, are supported as read alignment data sources.

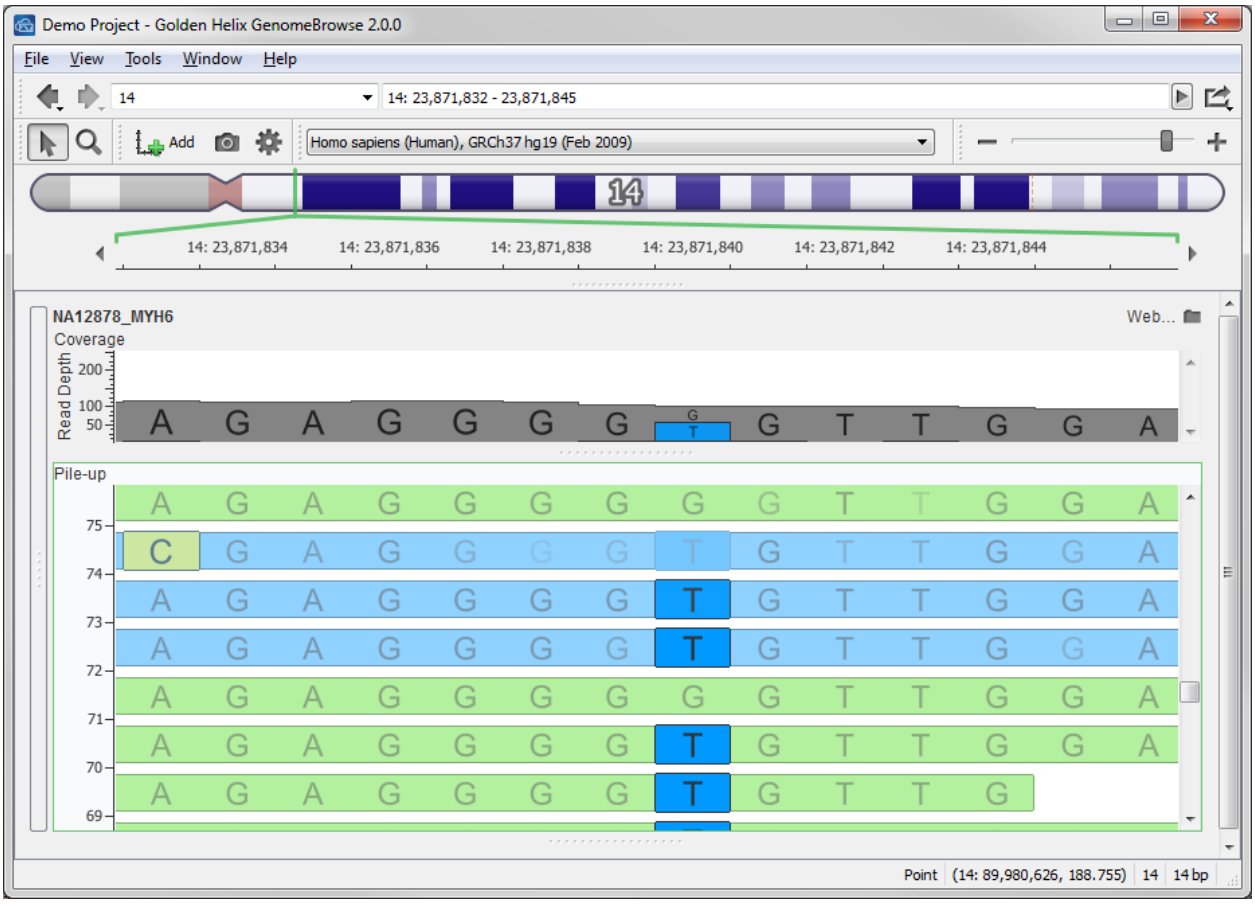



Figure 5.7: BAM File Plot

## Plot Description

Read alignment sources are visualized in two different ways coverage and pile-up plots. The coverage is a measure of read depth or read count across the genome. The pile-up plot shows individual reads from the read alignment source stacked up on the vertical axis to avoid visual overlap.

### Coverage Plot

From a whole genome or large zoom region the coverage plot shows stacked histograms of the reads (Read depth) the histograms are split into two groups to emphasize the strand of the read, either the forward or reverse strand.

When zoomed in close enough the coverage plot switches to a detail view showing stacked histograms of the nucleotide counts for the bins. Each bin is just one base-pair wide and the histograms count all of the nucleotides from the reads spanning that base-pair.

### Pile-up Plot

The pile-up plot displays all the reads from the read alignment source. Many of the reads may overlap, so the reads are stacked or piled up along the y-axis. The depth of the stack is often similar to the read depth, but there often empty spaces in the stack causing it to be taller than the actual number of reads spanning a given position. If the reads provided by the read alignment source are RNA-Seq reads aligned to a DNA reference sequence, they may be aligned across introns (DNA intervals that are not represented in RNA). Such read alignments are displayed with thin (gray by default) bars between the two ends of the read.

For a whole genome or wide zoom view, the pile-up plot displays a shape that approximates the way the stacks of reads will look as the view is zoomed in. Intron spanning reads are piled at the bottom of the stacks so that the collection of reads that span an intron region will appear with a large (gray by default) rectangle between them. The wide zoom views are primarily useful for navigating to regions which contain data and maintaining a spatial reference while adjusting the view.

The way read alignments are displayed in the pile-up plot can be changed to better suit various browsing needs. None of these changes affect wide zoom views. The coloring can be changed to emphasize either mismatches from the reference sequence, or the read strand (forward or reverse). The stacking can be entirely above the axis or split so that forward reads are stacked above the axis and reverse reads are stacked below. Read pairings can also be indicated (by thin light gray lines between reads), but for paired alignments to be displayed all reads must be stacked above the axis.

## Controls

### Display Tab

On the **Display** tab, the controls include:

- Emphasize:
  - Mismatches
  - Strand
- Stack:
  - Above Axis
  - Split By Strand
  - Paired Ends

- Per-base Quality Shading

The **Emphasize** control provides the ability to change the features emphasized. If **Mismatches** is selected, then plot coloring will be adjusted so that alleles that do not match the reference allele will be more visible. If **Strand** is emphasized, plot coloring will be adjusted so that the read's strand color will be bolder.

The **Stack** control provides the ability to change how the reads are stacked. If **Above Axis** is selected all reads are stacked above the X-axis. If **Split By Strand** is selected reads on the forward strand are placed above the X-axis whereas reads on the reverse strand are placed below the Y-axis. If **Paired Ends** is selected, reads are stacked to connect paired end reads.

The **Per-base Quality Shading** control enables or disables quality proportional blending of each base's color. High quality bases will be rendered more vibrantly and low quality bases will be blended into their background read color. This results in higher quality mismatches being visually more obvious than low quality ones.

#### Filter Tab

On the **Filter** tab, the controls include:

- Flag Zero Quality Alignments
- Filter Multi-Mapped Alignments
  - Mapping Quality Threshold
- Filter Duplicate Alignments
- Filter Duplicate Alignments
- Filter Vendor Failed Alignments

These controls allow for different coloring or removal of certain reads that are marked with commonly used flags in the BAM file format. The flags provided are typically used by alignment programs to indicate a potential problem or lack of certainty in a read's alignment, therefore it may be useful to highlight them or remove them from the visualization.

The **Flag Zero Quality Alignments** control enables or disables special coloring for alignments marked with zero mapping quality. Such alignments will be colored light gray when the option is enabled.

The **Filter Multi-Mapped Alignments** control disables or enables display of read alignments marked as "secondary alignment" or having zero mapping quality. Such alignments are hidden when the control is checked. The **Mapping Quality Threshold** will also be enabled when the control is checked. Setting the mapping quality threshold to a value greater than one will not only filter out zero mapping quality alignments, but also those with mapping quality less than the specified value.

The **Filter Duplicate Alignments** control disables or enables display of reads marked as "PCR or optical duplicate". Such reads will be hidden when the control is checked.

The **Filter Vendor Failed Alignments** control disables or enables display of reads marked with "not passing quality controls". Such reads will be hidden when the control is checked.

#### Layout Tab

On the **Layout** tab general plot controls can be changed. See *Layout Controls* for more information.

### Data Console

Clicking on a read in the pile-up plot will display information about the read from the read alignment source in the data console. The information will include, if available, the read name, mapping quality, mate chromosome, mate position,

template length, mismap probability, strand, several flag values, and the cigar operation string in a tabular form. The entire sequence of the read will also be included along with the corresponding list of base quality scores.

For wide-zoom views, clicking on a pile-up plot will display the maximum read depth for the aggregated region under the click position.

Clicking on a stacked histogram in the coverage plot will display information about the nucleotide counts at the associated genomic position. The information will include a table of matches, mismatches and deletions by type and nucleotide along with their counts, percentages, and mean qualities. It will also include a table of insertions and non-insertions existing at the nearest base-pair boundary along with their counts, percentages, and mean qualities.

For wide-zoom views, clicking on a coverage plot will display a table of mean read depth by strand (forward or reverse) for the aggregated region under the click position. The aggregation bin size will be shown above the table.

## 5.11 Allele Sequence Sources

This source contains allele sequence information. It is also known as a reference sequence.


Figure 5.8: Allele Sequence Source

### Plot Description

The data displayed from this source depends on the zoom range and the height given to the plot. At zoom ranges greater than 500 base-pairs the proportion of AGCT alleles (nucleotides) in the sequence are displayed as stacked

histograms. At zoom ranges less than or equal to 500 base-pairs the alleles in the sequence in the forward strand are displayed by themselves.

If the height of the allele sequence plot is increased to allow more drawing room, both the forward and reverse sequences are drawn with the forward strand on the top and the reverse strand on the bottom.

As the height of the plot is further increased hypothetical codon triplets are drawn based on both the forward and reverse strand sequences.

### Controls

#### Display Tab

On the **Display** tab general plot controls can be changed. See *Display Controls* for more information.

#### Layout Tab

On the **Layout** tab general plot controls can be changed. See *Layout Controls* for more information.

### Data Console

If the zoom range is greater than 500 base-pairs the nucleotide count and percentages for the particular stacked histogram clicked on will be displayed as well as the size of the window used to compute the stacked histogram.

If the zoom range is less than or equal to 500 base-pairs, and a single nucleotide is clicked on then the nucleotide and position information will be displayed in the console.

If hypothetical codons are visible and clicked on then information about the hypothetical codon will be displayed in the console.

### Special Features

For gaps in a sequence a nucleotide of **N** will be displayed in the track. This indicates that there is no known nucleotide at that particular location.

## 5.12 Variant Sites

Sources contain categorical/string values for positions included in the track. Here the focus is on the location of the variant and the information contained for each variant as a whole.

### Plot Description

Variant sites mark the location of all variants in the source. If alleles are detected in the source, the alleles are colored based on the nucleotide bases of the variant or as an insertion "I" bar or deletion rectangle. If allele information is not detected in the source variants are drawn as gray flags marking the location.

If there are numeric fields in a variant source, these fields can be plotted as a numeric value plot. See: *Value Plot*.

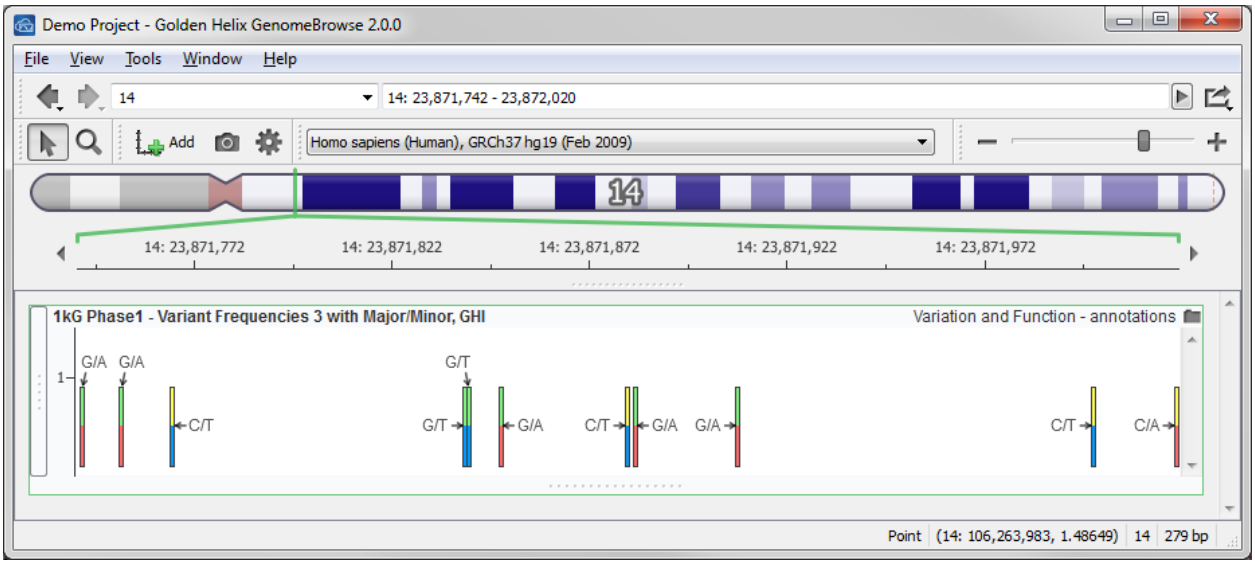



Figure 5.9: Variant Site Plot

## Controls

### Display Tab

On the **Display** tab, the controls include:

- Labels

The **Labels** control provides the ability to change the data field that provides the feature labels that are drawn on the plot at close enough zooms.

For the other controls that are common with most plot types, see *Display Controls*.

### Filter Tab

A **Filter** can be used to control which features are drawn in the plot.

To add a filter either click on **Insert** or right-click anywhere in the Filter list box and select **Insert**.

Please see *Filter Controls* for more information.

### Layout Tab

On the **Layout** tab general plot controls can be changed. See *Layout Controls* for more information.

## Data Console

Clicking on a variant plot will print out information about the data source including all of the fields in the source. Clicking on a data point in the plot will result in the value, the label for the feature as well as any applied styling.

## 5.13 General Control Panels

### Display Controls

The display controls for a plot may contain the following options:

- **Chromosome Shading**: This control enables or disables alternating background shading for adjacent chromosomes. The shading will only be visible when zoomed out far enough, regardless of this controls setting.
- **Background**: This control displays the current background color for a selected plot and allows a new background color to be set.
- **Feature Labels**: This control enables or disables labeling of features within selected plot(s) or item(s).
- **Y-Range**: Enter the numeric y-axis range to change the y-axis extents of the plot. If the current zoom mode is Fit Data or Auto, the zoom mode will be automatically changed to Hold when a new y-axis range is entered.

  Beneath the y-range control is a set of y-axis zoom mode selection buttons. Available zoom modes are:

  - **Manual - The y-axis zoom is controlled manually and all zoom controls are** enabled. This mode can be accessed using the hot-keys **r** or **m**.
  - **Hold - The y-axis zoom is controlled manually but vertical panning on the** plot canvas is disabled, protecting against accidental changes to the y-axis zoom. All zoom controls are enabled. This mode can be accessed using the hot-keys **e** or **h**.
  - **Fit Data - The y-axis zoom is changed dynamically as the x-axis zoom changes** to show all the data on the vertical axis. All vertical zoom controls are disabled. This mode can be accessed using the hot-keys **w** or **f**.
  - **Auto - The y-axis zoom is changed dynamically as the x-axis zoom changes.** When zooming in close on the x-axis the y-axis will be zoomed in as well to automatically improve the detail of the vertical axis in proportion to the horizontal axis. This mode is only available on Heat Map, Alignment Pile-up, Value, and Variant Map plots. It can be accessed using the hot-key **q**.

### Filter Controls

There are two types of filter controls available. All plot types have the ability to filter data using feature filters. Variant Maps and Heat Maps also have the ability to filter data using sample filters.

**Inputs**:

- **Field** - The list of feature fields which are available for the current source.
- **Function** - The list of functions which can be preformed on field values.
- **Constant** - The list of constants which are available for the construction of filters.

**Operations**:

- **Comparison** - Common logic operators used to compare fields and their values
- **Arithmetic** - Common arithmetic operators which can be applied to numeric fields
- **Logic** - Logic operators which can be used to combine simple expression into compound filters
- **parenthesize** - Adds parenthesis around the current selection

A simple filter takes the form:

*Chr == "1"*

Where *Chr* represents the name of one of the fields found in the source, and *"1"* is the value of the field which must be placed in double quotes for type string fields; they are separated by the == comparison operator.

Simple expressions can be combined by placing them in parenthesis and using the logic operators to create complex filters.

### General Filter

The general filter takes a feature level field such as Chromosome or other genome-wide data field available in the source and filters the data drawn based on the specified criteria.

### Sample Filter

The sample filter takes a sample-wise field such as the sample names and removes samples from the entire plot if they do not meet the specified criteria.

## Layout Controls

The layout controls for a plot may contain the following options:

- **Title**: The check box indicates whether or not to display the title. The Edit button allows title of the plot to be changed. Full font, color and styling controls are available. Quick edit of a title is available by double-clicking on the title in the plot view. Right clicking on the title in the plot view or the plot tree also provides the title editing option.
- **Location**: The check box indicates whether or not the location of the plot's data sources should be displayed at the top right corner of the plot. The location can be a useful reference point at a glance and may help to differentiate between sources with the same or similar names.
- **X-Axis**: Specifies whether or not to display an x-axis scale next to the plot.
  - *Label*: Specifies whether or not to show the x-axis label. An edit button is also provided to edit the X-axis label text and style.
- **Y-Axis**: Specifies whether or not to show the y-axis scale next to the plot.
  - *Label*: Specifies whether or not to show the y-axis label. An edit button is also provided to edit the y-axis label text and style.
- **Height**: Specify the height of the plot canvas in pixels. This control can be used to ensure that multiple plots are exactly the same height.

# CHAPTER SIX

# THE DATA SOURCE LIBRARY

Golden Helix's Data Source Library provides complete access to local, public, network, and project data sources to add to a GenomeBrowse viewer. The public network or cloud-based account sources can be downloaded through this interface as well.

The options available with the Data Source Library depend on the context in which it is accessed or used. All available features of the Data Source Library will be discussed and the specific options for each context will be presented.

## 6.1 Navigating the Data Source Library

The Data Source Library is structured like a File Browser or Windows Explorer dialog. The primary view is a list of valid data sources found in the currently selected location(s). There is a location navigation pane on the left side, showing "bookmarked" locations. The information pane at the bottom displays information about the last selected item(s). There may also be a plot data pane on the right side, which can be shown to select particular subsets or uses of the currently selected data source(s).

All of these features will be discussed in detail below.

### The Location Panel

The location panel holds buttons to add data sources by browsing or from a project (when applicable). It also displays a hierarchy of "bookmarked" locations. The **Browse** button can be used to add data from any locally accessible data source. For commonly accessed locations it may be useful to create a bookmark so that they can be viewed with a single click. Locations can also be organized into containers. The containers can help organize locations. In addition, selecting a container will display the combined listing of data sources from all of the locations it contains. There are several locations available by default. Locations can be added or removed with the **+** and **-** buttons, or by right-clicking in the location hierarchy.

#### Browse

Allows access to data sources on the computer or local network without adding a location "bookmark".

Click on the **Browse** button to reveal the path and **Browse...** buttons at the top of the Data Source Library window. A path to a data source location may be entered in directly into the path text box or a location may be navigated to by clicking on the **Browse...** button. Any valid data sources found at the specified path will be shown (if they match the current filters).

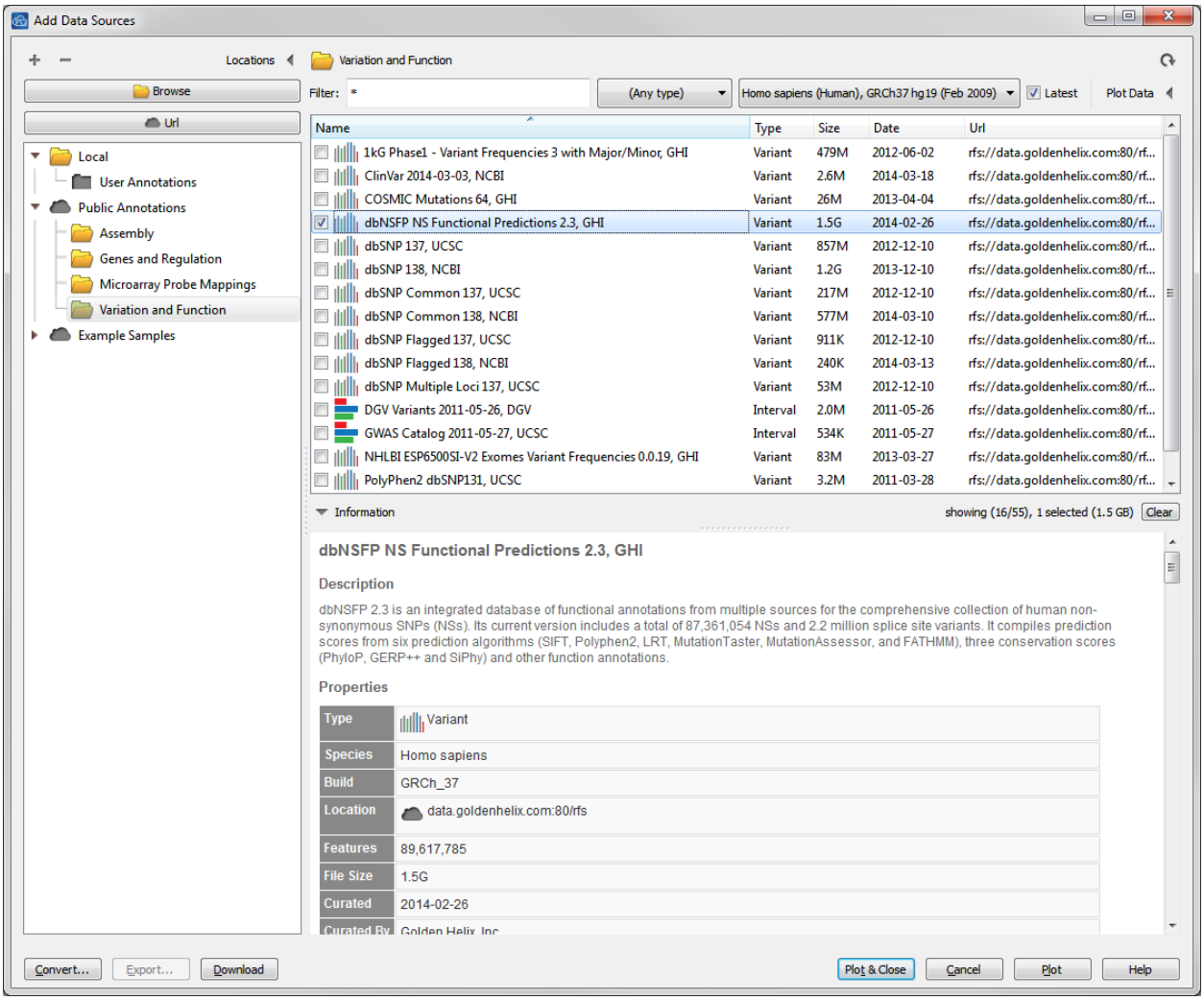



Figure 6.1: Data Source Library

### Project

Allows access to data from nodes in the current project. Select an enabled node to retrieve a list of its plottable data. Then check one or more items in the list to plot.

### Local

A location container for various local sources such as User Annotations, GenomeBrowse Annotations, and System Annotations. Additional local locations for sources can be added as well. These folders must be refreshed to see any changes that are made elsewhere. Refresh a location by selecting it and then clicking the refresh button in the upper right corner of the Data Source Library, see *Refresh Sources*.

### Public Annotations

Golden Helix has provided a large set of public annotation tracks for download or streaming. These data sources can be visualized in GenomeBrowse and utilized in various DNA-Seq workflows.

To either stream or download data a network connection is required. If the software cannot connect to the public annotations network server please check your proxy settings and adjust if necessary. See *Adjusting Proxy Settings* for more information.

### Example Samples

Golden Helix has provided several public example files for streaming and downloading directly to GenomeBrowse.

To either stream or download data a network connection is required. If the software cannot connect to the public example network server please check your proxy settings and adjust if necessary. See *Adjusting Proxy Settings* for more information.

## The Source List

The source list in the main panel lists all of the valid data sources for the selected location. To be a valid data source, a file or spreadsheet (from a project if applicable) must have one or more plottable data feature(s). For instance, a VCF file can contain numerous plottable data features such as variants (as genotypes), read depth, allelic depth, quality scores, and so on. Other data sources will only have one plottable data feature (i.e. gene data sources can only plot gene tracks).

If a source has multiple plottable features, the *Plot Data* panel will contain a list of all the plottable features. See *The Plot Data Panel* for more information.

Multiple sources can be selected simultaneously from multiple locations. The text on the right side, below the source list panel indicates how many sources are currently selected. The entire selection can be cleared by clicking on the accompanying **Clear** button.

If a specific source type is not selected, a column containing the type of each source is included in the source list. This column as well as any of the other columns in the source list can be used for sorting the list. All of the columns can be reordered as well. By default sources are sorted on their name in ascending order. To sort on a column, click on the column header. Click once for ascending order, twice for descending order.

### Filtering Tools

The filtering tools can be used to limit the data sources shown in the source list. The tools available include:

- A text box to filter down to sources with a specified string in their name,
- A type selection drop down menu to filter sources to a certain type,
- An assembly selection drop down menu to filter sources down to a specific species or build.
- A "Latest" check box to filter sources down to only the most recent version.

#### Refresh Sources

To refresh the source list, click on the refresh icon located in the upper-right corner of the Data Source Library.

### The Plot Data Panel

When applicable the plot data panel is available and allows the selection of multiple plottable features based on the data type(s) in the source.

### The Information Panel

Information available relating to the last selected data source(s) or location(s) is displayed in the information panel. In the case of a data source, this information may include the represented data's origins in the form of a link or summary, field types and descriptions, and any citation information required as part of the data's redistribution.

Local sources can have their documentation and some field information edited unless the source is locked for editing. If a source can be edited there will be an **Edit** hyperlink next to the source name in the **Information** panel. See *editAnnotationDocumentation* for more information.

## 6.2 Data Source Types Available through the Data Source Library

- **BAM:** A binary sequence alignment file from a secondary analysis DNA-Seq pipeline.
- **BED:** A tab-delimited text file containing annotation track information. See BED file format for more information.
- **IDF:** A file format specifically designed to work with Golden Helix software programs. Possible types of IDF files include:
  - *Cytoband*: Contains cytoband intervals including Giemsa stain results.
  - *Interval*: Contains general purpose intervals. Each track can define the meaning of its intervals independently and multiple data fields of any supported types may be associated with each interval.
  - *Gene*: Contains information on gene transcripts.
  - *Heat Map*: Contains numeric values on intervals for one or more samples.
  - *Allele Sequence*: Contains allele sequences. Typically represents a particular DNA reference sequence as a list of its single letter nucleotide abbreviations.
  - *Value*: Contains one or more numeric values for each interval or position of interest.
  - *Variant*: Contains genomic variant data as intervals. The represented variants may include single nucleotide variants (SNVs), insertions, and deletions.
  - *Variant Map*: Track contains genotypes (variants) for one or more samples. Genotypes are drawn to emphasize deviation from the reference allele sequence.

- **VCF:** Information from a VCF file typically contains genotypes for one or more samples which can be visualized a Variant Map, or genomic variants which can be visualized as a variant plot. Other types of data stored in VCF can also be visualized. VCF files will be indexed and compressed before drawing.

## 6.3 Downloading Data

Network data sources can be downloaded from most instances of the Data Source Library. When one or more network data sources are selected, the **Download** button will be enabled. Clicking the **Download** button will start downloading all selected network data sources in the background. The Download Manager will be activated to display the download progress. See *Download Window* for more information.

The default target download location is "Golden Helix/Common Data/Annotations". This can be changed in GenomeBrowse options. See *General Options for GenomeBrowse* for more information. Once the downloading is complete, the target location can be refreshed and the new local copy of the data source can be used just like any other local data source.

## 6.4 Exporting Data

Local data sources can be exported to delimited text, variant call 4.1 (VCF), Microsoft Excel Xls, FASTA, and wiggle track (WIG) format, depending on the source file type. Only one source can be exported at a time.

To export a source, check its selection box and click the **Export** button. A dialog will appear with the compatible export file types. Upon selection of a export file type click the **Export** button and the appropriate subdialog will appear.

### Delimited Text

Currently, only non full-coverage sources (i.e. cytoband, interval, gene, value, and variant) can be exported to delimited text.

The settings which may be changed are:

- **Header**: By default, this setting is selected and will cause the source's field names to be displayed on the first line of the file.
- **Prefix**: A string can be entered to prefix the header line of the file.
- **Delimiter**: By default, a "tab" character will delimit the columns in the file. A comma can be selected from the drop down list or a custom string can be entered.
- **Sub Delimiter**: By default, a comma will delimit a list of values in column. A "tab" character can be selected or a custom string can be entered.
- **Coordinates**: The genomic coordinates of the exported intervals can be represented as 0-based, 1-based, or a position.
  - **0-Based Interval**: The difference between the stop and the start positions defines the width of the interval. For example, an interval covering the first three positions of a chromosome in 0-based coordinates would be specified as [0, 3]. (Also known as 'half-open coordinates'.)
  - **1-Based Interval**: The difference between the stop and the start positions plus one defines the width of the interval. For example, an interval covering the first three positions of a chromosome in 1-based coordinates would be specified as [1, 3]. (Also known as 'indexed coordinates'.)
  - **Position (1bp width)**: This option outputs a single coordinate. Thus, it is only useful if all features have a single base pair width. The position is 1-based so the smallest position in a chromosome would be 1.

- **Exported Fields**: The desired fields may be selected using this option. The order in which the fields appear in the exported file may be changed by reordering the list by dragging fields up or down.
- **Output File**: Clicking the **Browse** button will bring up a dialog to select the name and location for the exported file.

### Variant Call (VCF) 4.1

Variant Call Files can be created for variant and variant map files.

The settings which may be changed are:

- **Exported Fields**: The desired fields may be selected using this option. The order of the fields in the file will follow the ordering found in the genome browse table.
- **Exported Flags**: The flags for the data features may be selected. These flags correspond to the Flags field in the genome browse table.
- **Output File**: Clicking the **Browse** button will bring up a dialog to select the name and location for the exported file.

### Microsoft Excel Xls

Currently, only non full-coverage sources (i.e. cytoband, interval, gene, value, and variant) can be exported to xls.

The settings which may be changed are:

- **Exported Fields**: The fields desired in the output xls file may be selected. The fields will appear in the output file in the same order which they appeared in the GenomeBrowse table view.
- **Output File**: Clicking the **Browse** button will bring up a dialog to select the name and location for the exported file.

**Note**: Output to the Microsoft .xls format is limited to 32767 rows and 256 columns. After 32767 rows have been written copying will stop and the rest of the input file will be truncated. Fields may be unselected to limit the number of columns in the file.

### FASTA Format

FASTA files may be written for sequence sources with valid assemblies.

The settings which may be changed are:

- **Separate Chromosome Files**: Selecting this check box will create a new file for each Chromosome in the source file.
- **Output File**: Clicking the **Browse** button will bring up a dialog to select the name and location for the exported file.

**Note**: When creating separate files for each of the chromosomes in the source file, the text "%chr%" must be included in the destination file. When the individual files are written the "%chr%" will be replaced with the chromosome corresponding to that file.

### Wiggle Track Format

Variable Step Wiggle track files can be created files with numeric fields

The settings which may be changed are:

- **Value**: the data value associated with the each chromosome position can be selected from the numeric fields in the file.
- **Base Span**: If single base span is selected a data point will be created for every base covered in the file, otherwise the span from the first feature will be used (in this case the span must be consistent across all of the features).
- **Output File**: Clicking the **Browse** button will bring up a dialog to select the name and location for the exported file.

## 6.5 Source Information Editor

To edit the documentation, field descriptions or categorical values for a source, click on the local source in the Data Source Library and then click on the **Edit** hyperlink in the **Information** panel. Alternatively, right-click on the source and select **Edit Source Info**. This will open up the **Source Information Editor**.

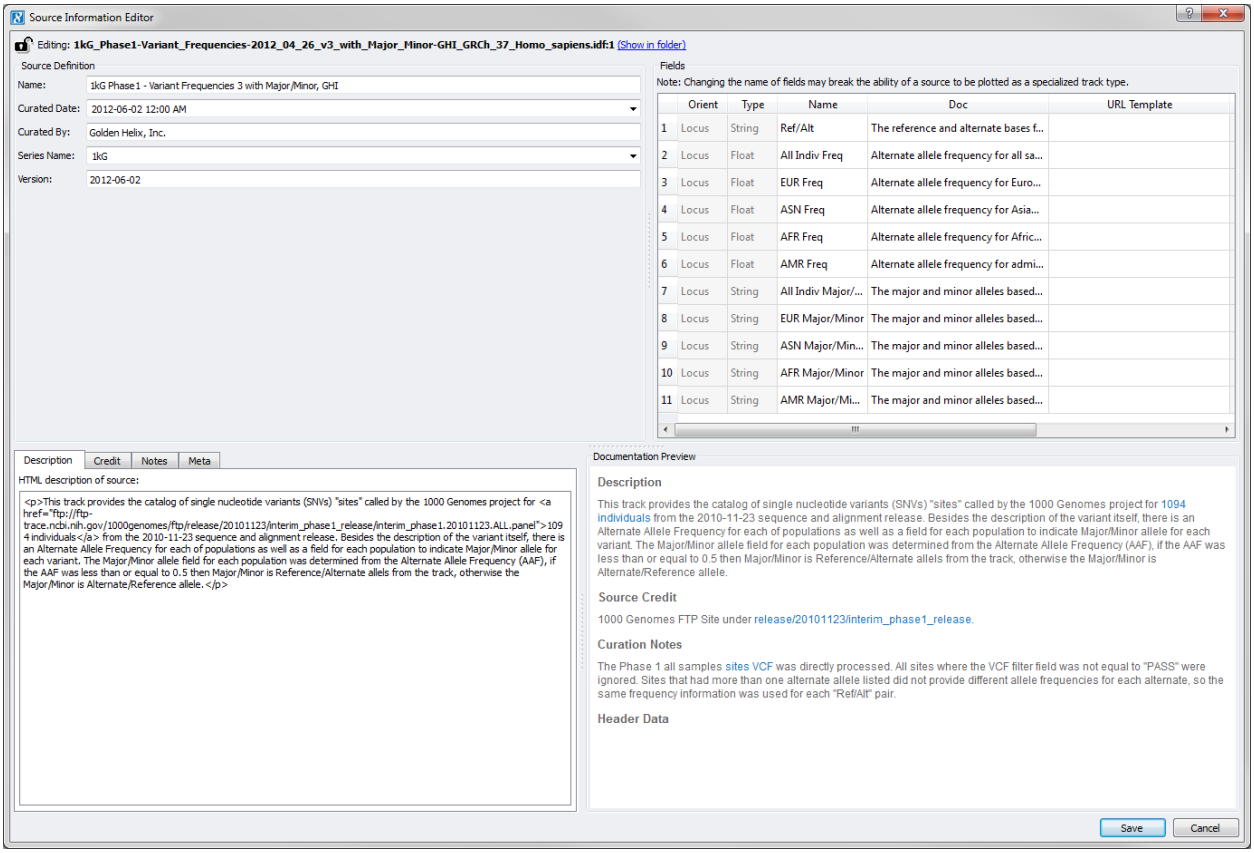



Figure 6.2: Source Information Editor for a variant source

The lock icon at the top of the window may be locked for sources provided by Golden Helix, in which case the source can not be edited. Any custom created source should not be locked and the information in the source can be edited.

Annotation sources must have a name specified. In addition to the source name, documentation on how the data was converted as well as the date and documentation for each field can be specified or edited. All documentation will be embedded into the TSF file to make sharing files and documentation easy.

There are three sections in the **Source Information Editor**:

- **Source Definition**: This information is used to identify the annotation source, and also indicate the date it was converted, who converted it and any version information.
  - **Name**: [*Required*] The name of the annotation source.
  - **Curated Date**: [*Required*] By default this was a date associated with the files converted into the annotation source. It can be modified, but a date is required.
  - **Curated By**: Name or organization of who is curating the data.
  - **Series Name**: Name of a particular group of data. This field can be used to differentiate between newer versions of the same type of data. For example, RefSeqGenes-UCSC or dbSNP.
  - **Version**: A version number or date. It is recommended if there is a particular version name or identifier that this is included in the **Name** field and that this field be used for a date associated with the particular version.
- **Fields**: The individual field descriptions can be specified in this table.
  - **Orient**: The orientation of the data (locus or sample) cannot be modified.
  - **Type**: The type of the data (cannot be modified). If the type needs to be modified, the TSF file can be converted into another TSF file using the **Convert Source Wizard**. See *Converting an IDF or TSF File* for more information.
  - **Name**: The name of fields can be modified. However, if a field name is modified for a required field with an explicit name the source may not be able to be plotted as a specialized track type.
  - **Doc**: The documentation string for the specific field.
  - **URL Template**: For fields with information that can be queried in an external site, specify the URL and two dollar signs ($$) to indicate where the text should be replaced. For example, for an "Identifier" field that contains RS ID's, the URL Template could be http://www.ncbi.nlm.nih.gov/snp/?term=$$.
  - **Categories**: Click on the **Edit** button to edit the category names and/or documentation. The documentation can be edited to rename and/or document categories for categorical field types. For instance, change the category "D" to "Damaging" for a more informative name without having to modify the source file(s).
- **HTML Documentation**: The four tabs at the bottom of the dialog are for writing HTML documentation of the source. The tabs are to guide the writing of the documentation and to provide nice headers for each section. HTML tags can be used for formatting.
  - **Description**: Description of the source and where it was obtained from.
  - **Credit**: Any required citations or credits for the source should go in this section.
  - **Notes**: Any relevant notes on pre-processing that had to be performed on the data or settings used to convert the source.

    **Note:** This section will contain statistics about the fields and data in the source.

  - **Meta**: Any meta information for the data source(s).

    **Note:** If the source was created from converted VCF files the header information from the first VCF file will be placed in this section.

# CHAPTER SEVEN

# GENOME ASSEMBLIES

A genome assembly defines the chromosomes for a particular species and build. This definition includes the chromosome names and lengths, as well as the order in which they are arranged when displayed in a genome browser. GenomeBrowse use the current genome assembly so that plotted features will be arranged according to its chromosome definition. The species and build specified by the selected genome assembly is used to manage annotation and data sources within GenomeBrowse. This functionality is intended to help prevent accidental alignment of annotation data to the wrong genome assembly.

A genome assembly can be selected from the list of bundled genome assemblies, downloaded from the genome assemblies data repository from Golden Helix, created from a marker mapped spreadsheet, or created from a marker map. This enables the user to create a genome specific to the data at hand or for a species not in the list of bundled species.

The current genome assembly can be selected through the tool bar controls of GenomeBrowse. The genomes available can be viewed through the project navigator menu item, **Tools > Manage Genome Assemblies**.

## 7.1 Bundled Genome Assemblies

Basic genome assemblies for several species are made available with GenomeBrowse. The initial list of chromosome definitions to provide was based on those available within the Integrated Genome Browser (IGB) http://www.bioviz.org, [Nicol2009]. Additional species and builds have been added to the bundled genome assemblies since then.

If genome information for a particular species and build is available and is not yet included in the bundled builds, you can convert that information into a genome assembly at the same time as converting a Reference Sequence to a TSF annotation source. See *Convert a 2Bit File* or *Converting a FASTA File*.

## 7.2 Switching Genome Assemblies

The genome assembly can be changed in one of two ways. These are through GenomeBrowse tool bar controls or through the current project options dialog. (**Tools > Current Project's Options**).

### Using Tool Bar Controls to Switch Genomes

GenomeBrowse has genome assembly information located on the tool bar. The control menu contains a list of all system and user genome assemblies. Recently used genome assemblies are listed at the top, all genomes are listed in alphabetical order under the black bar.

To change the genome to a different species or build, select the desired genome assembly from the list.

After changing the current genome assembly, the user will be asked if the reference sequence track should be downloaded if it is not found in the local annotation folder. Selecting **Yes** will start the download of the annotation track. Analysis and visualization of the data can continue while the file is being downloaded except for BAM files. Selecting **No** will not download the reference file so that coverage for BAM files will not be able to be computed.

**Note:**

1. Switching the genome assembly will cause the data to be re-plotted with the new mapping and the zoom to be reset showing all of the data.
2. If annotation tracks that do not correspond to the current species and build features are plotted the data may not be correctly aligned to the genome.

# CHAPTER EIGHT

# ANNOTATION CONVERT SOURCE WIZARD

Annotation sources exist in numerous file formats. Some are text based and easily manipulated, others are in binary formats that can only be read by specialized readers. If the format is a genome wide variant text file the size of the files can quickly become unmanageable. In addition separate documentation files need to be kept to contain the source and data curation information. Finally, if the data is not formatted subject to very specific requirements it is not possible for the files to be read and handled with ease and speed. To meet the ever increasing need for visualization of numerous file formats Golden Helix has created a very powerful convert source wizard for both Golden Helix SVS and GenomeBrowse. This converter will take nearly every data file type and convert it into a 'TSF' file. This file format contains not only the data but also the coverage information, data extents, documentation, index and manages to package it up in a compressed data format to keep the file sizes as small and as compact as possible.

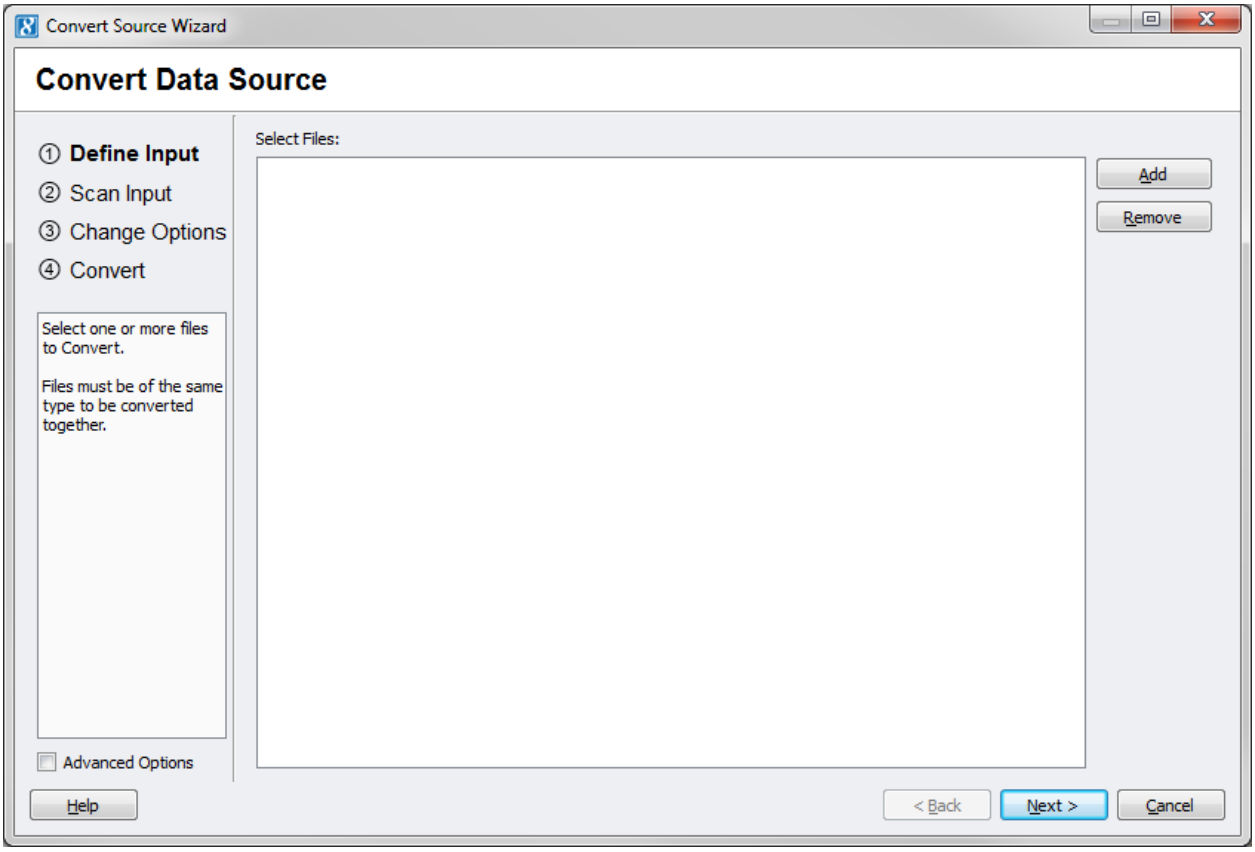



Figure 8.1: File Add Page of Convert Source Wizard

## 8.1 Opening the Convert Source Wizard

From a Data Source Library click on the **Convert...** button in the lower left hand corner of the dialog. This brings up the **Define Input** dialog which is the first step of the conversion wizard. In the left hand pane the source conversion steps are listed. Below the itemized list of steps is an information pane that will contain important instructions or tips for each step in the wizard.

The source conversion process can either use default settings or allow for more control over the process. To have more options or control over how the source is converted, check the **Advanced Options** box below the information pane.

On the right hand side is the file selection (Add) dialog. Multiple files can be selected at the same time to be converted into one source. However, all files will be added to the same TSF file and are required to be of the same type. For instance, if there is one FASTA file per chromosome, select all of the FA files, one for each chromosome. Similarly, if the source file type is VCF and there is one VCF file for each chromosome, select all of the per chromosome VCF files.

To select files, click on the **Add** button. Files can also be added to the wizard by dragging and dropping into the selected files dialog.

To remove files from the list, select the file by clicking on it and then click on the **Remove** button.

Once the desired files have been added and icon will appear in front of the file names. If the file extension is of a specific type an icon representing the expected source type will be displayed (Allele Sequence, Interval, Gene, Variant, etc.). If the file is a text file that can be converted into numerous source types a question-mark icon will be displayed in front of the file(s). If an exclamation point in a red circle is displayed, the file is either an invalid source or multiple sources of different types are selected in the same list of files. To get more information on the error, click on the **Next >** button.

If all sources are of the same type and are valid, clicking on the **Next >** button will lead to the next step in the wizard. At any point after the first page you can click **< Back** to return to previous steps. Any information changed or added will be preserved as long as the list of file sources is not modified. Clicking **Cancel** will exit the source conversion wizard.

If a source is already indexed it will not need to be scanned to determine the Genomic Coordinates and/or data types. If the file is not indexed a scanning pass will be initiated. This scan may be skipped if the genomic coordinates and/or data types are known. In general the scan will be the step immediately following the file selection, however, for text files (TXT, TSV, CSV, etc.) the delimited file characteristics will need to be specified first.

Each of the data formats are covered below in greater detail.

## 8.2 Convert a 2Bit File

A 2Bit file is a packed binary format described at http://genome.ucsc.edu/FAQ/FAQformat.html#format7 and is a very efficient method of packing ACGT sequence data. The 2Bit import requires exactly one input file in 2Bit format. The only available output type is 'Allele Sequence'. When creating an allele sequence source a new genome assembly can be created at the same time if one does not already exist for the species and build for the 2Bit file.

After selecting a 2Bit source on the file selection page of the convert wizard, click **Next >**. This brings the wizard to the **New Genome Assembly Build** page. A 2Bit source can define a genome assembly if one has not already been defined for the species and build.

**Note:** Throughout the Convert Source Wizard there are certain options considered "Advanced Options" that do not need to be selected for in most cases. To show "Advanced Options", check the box on the lower left of the dialog. Any option that is advanced will be labeled as such in the documentation.

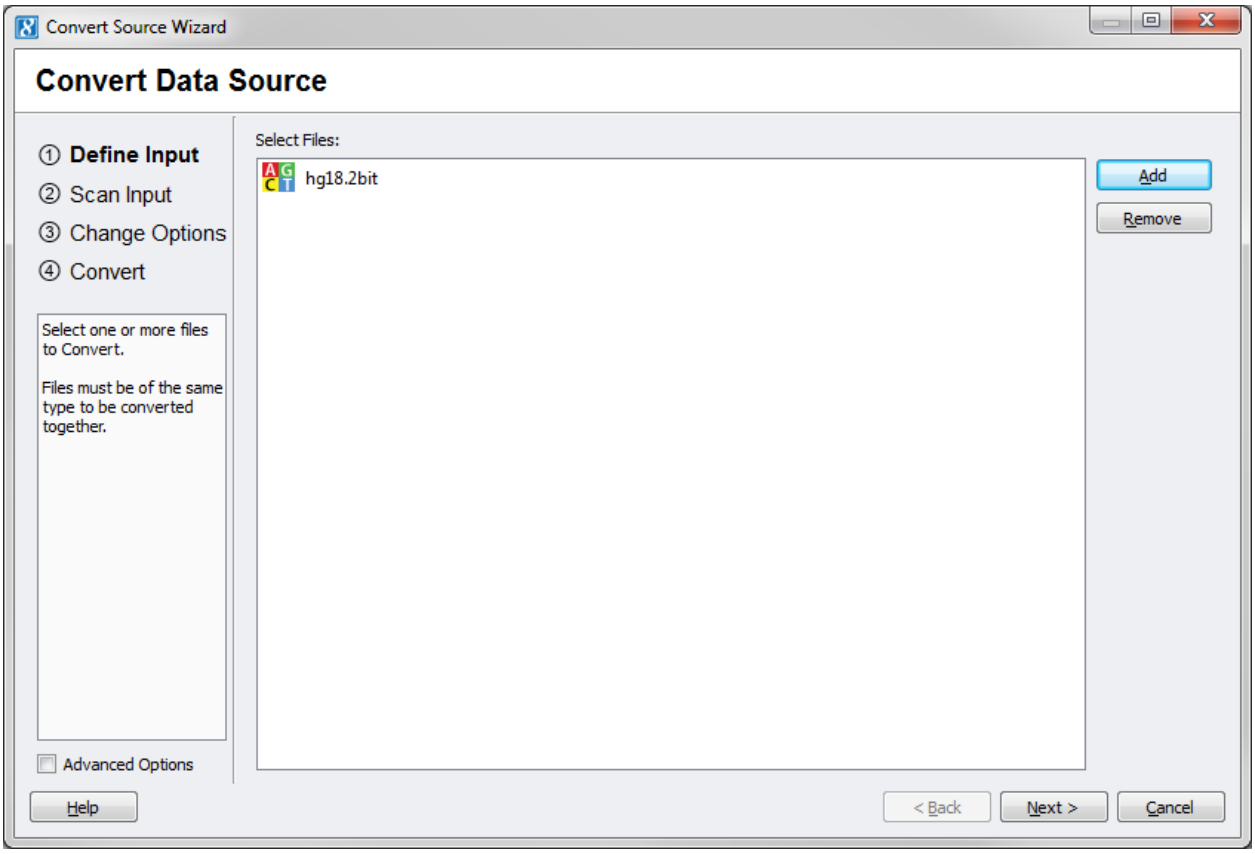



Figure 8.2: Added 2Bit human reference sequence file to convert wizard

## Select an Existing Genome Assembly from 2Bit

**ADVANCED OPTION**

To select an existing assembly for the allele sequence, make sure the **Advanced Options** option box is check and select an assembly from the drop down list. Selecting an existing assembly will inactivate all of the other fields on the Genome Assembly Build page. However, the **Source to Segment Mapping** table will remain active.

The **Source to Segment Mapping** table can be used to exclude or include segments in a new/updated genome assembly file. Segments previously included in the selected genome assembly will have a green background. Segments not included in the selected genome assembly will have a white background. Segment names that exist in the selected assembly but have different lengths between the source and selected assembly file will have a warning icon in front of the segment name. See *Define the Source to Segment Mapping* below for more information.

## Define a new Genome Assembly from 2Bit

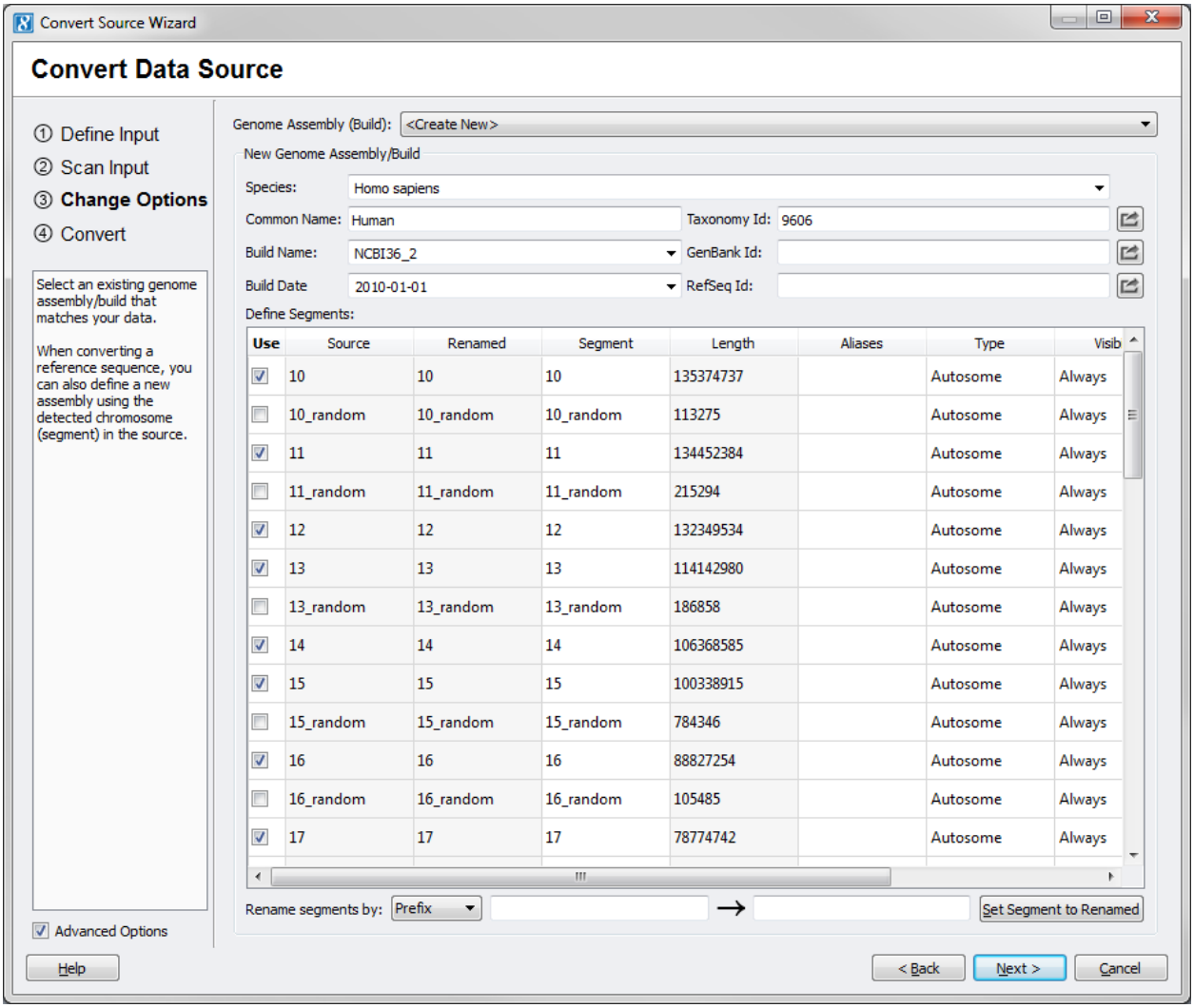



Figure 8.3: Create new assembly in convert wizard

To define a new genome assembly/build, fill in as many of the available fields as possible:

- **Species: Either select the species from the list or enter in a new one.** The scientific name is preferred. Examples include: 'Homo sapiens', 'Canis familiaris', etc.
- **Common Name: Enter in a common name for the species, such as 'Human' or** 'Dog'.
- **Build Name: The NCBI assembly name is preferred, but the assembly synonym** or UCSC assembly name can be used instead.
- **Build Date**: Submission or published date of the assembly.
- **Taxonomy Id: Taxonomy ID for the species. If the species is in the NCBI** database clicking on the link out button will open a web page on NCBI to help identify the taxonomy ID.
- **GenBank Id: GenBank Assembly ID. If the species is in the NCBI** database clicking on the link out button will open a web page on NCBI to help identify the GenBank ID. This field can be left empty if the species does not have a GenBank ID.
- **RefSeq ID: RefSeq Assembly ID. If the species is in the NCBI** database clicking on the link out button will open a web page on NCBI to help identify the RefSeq ID. This field can be left empty if the species does not have a RefSeq ID.

The **Source to Segment Mapping** table can be used to exclude or include segments in a new/updated genome assembly file. See *Define the Source to Segment Mapping* below for more information.

## Define the Source to Segment Mapping

Segments can either be excluded or renamed from the assembly using the fields in this table. The type of the segment can be set as well as the visibility of the segment.

- **Use: To include a segment in the assembly leave the 'Use' box** checked. To "Check All" or "Uncheck All" click on the "Use" column header. "Uncheck All Unmapped" will not work when creating a new assembly as there are no mapped or unmapped segments.

  **Note:** If there are more than 5000 segments only the 5000 longest segments will be included in the assembly file. If there are more than 500 segments they will be arranged in the assembly file in descending order by length.

- **Source: The name of the segment from the allele sequence file. To** rename a segment, either double click on the name in the **Segment** column, or if the segment names share the same pattern to be removed or for renaming, below the segment definition table are controls for renaming segments programmatically. The options include:

  - *RegEx*: Use regular expressions to rename the "Source".
  - *Substring*: Remove a substring from all segment names to generate the segment name.
  - *Prefix*: Remove a common prefix from all segment names.
  - *Suffix*: Remove a common suffix from all segment names.

  Enter in either the RegEx expression or the string to remove in the first text box. A preview of the renamed segment name will appear in the second text box. To apply the rename to all segments click on the **Set Segment to Renamed** button.

- **Length**: The length of each segment is displayed in this column.
- **Aliases: If a segment has an alias, it can be specified in this column.** For example mitochondrial chromosomes might be named "M" or "MT", the alternate name can be listed in the alias column.

- **Type: Set the type of segment. By default 'Autosomes' are always visible** and the rest are visible only if there is data. Options include:
  - Autosomes
  - Allosome
  - Mitochondrial
  - Fragment
  - Scaffold
  - Contig
  - Unknown
- **Visibility: The visibility of the data can be set manually. If the** segment should only be shown in GenomeBrowse if there is data the visibility can be set to "With Data". Options include:
  - Always
  - Never
  - With Data

Once the assembly has been defined click **Next >**. When the allele sequence is converted to TSF format an assembly file will be created and placed in the User Assembly folder and will be available for use in Golden Helix SVS and GenomeBrowse.

After clicking **Next >** the wizard will display the documentation page. See *Documentation Step* for more information.

## 8.3 Converting a FASTA File

A FASTA file is a text file where each character of data designates the value of a sequence base at each offset in a segment designated by its simple header. After conversion an 'Allele Sequence' source is created. When creating an allele sequence source a new genome assembly can be created at the same time if one does not already exist for the species and build for the FASTA file(s).

After selecting one or more FASTA sources (FA, FASTA, FA.GZ, FASTA.GZ, etc.) on the file selection page of the convert wizard, click **Next >**. If the data has not previously been indexed the file(s) will be scanned to obtain the genomic coordinates. This scan may not be skipped since a new genome assembly may be generated based on the genomic coordinates in the file(s). Next, the wizard will display the **New Genome Assembly Build** page. FASTA sources can define a genome assembly if one has not already been defined for the species and build.

**Note:** Throughout the Convert Source Wizard there are certain options considered "Advanced" options that do not need to be selected for in most cases. To show "Advanced" options, check the box on the lower left of the dialog. Any option that is advanced will be labeled as such in the documentation.

### Select an Existing Genome Assembly from FASTA

**ADVANCED OPTION**

To select an existing assembly for the allele sequence, make sure the **Advanced Options** option box is check and select an assembly from the drop down list. Selecting an existing assembly will inactivate all of the species, build and identifier fields on the Genome Assembly Build page. However, the **Source to Segment Mapping** table will remain active.

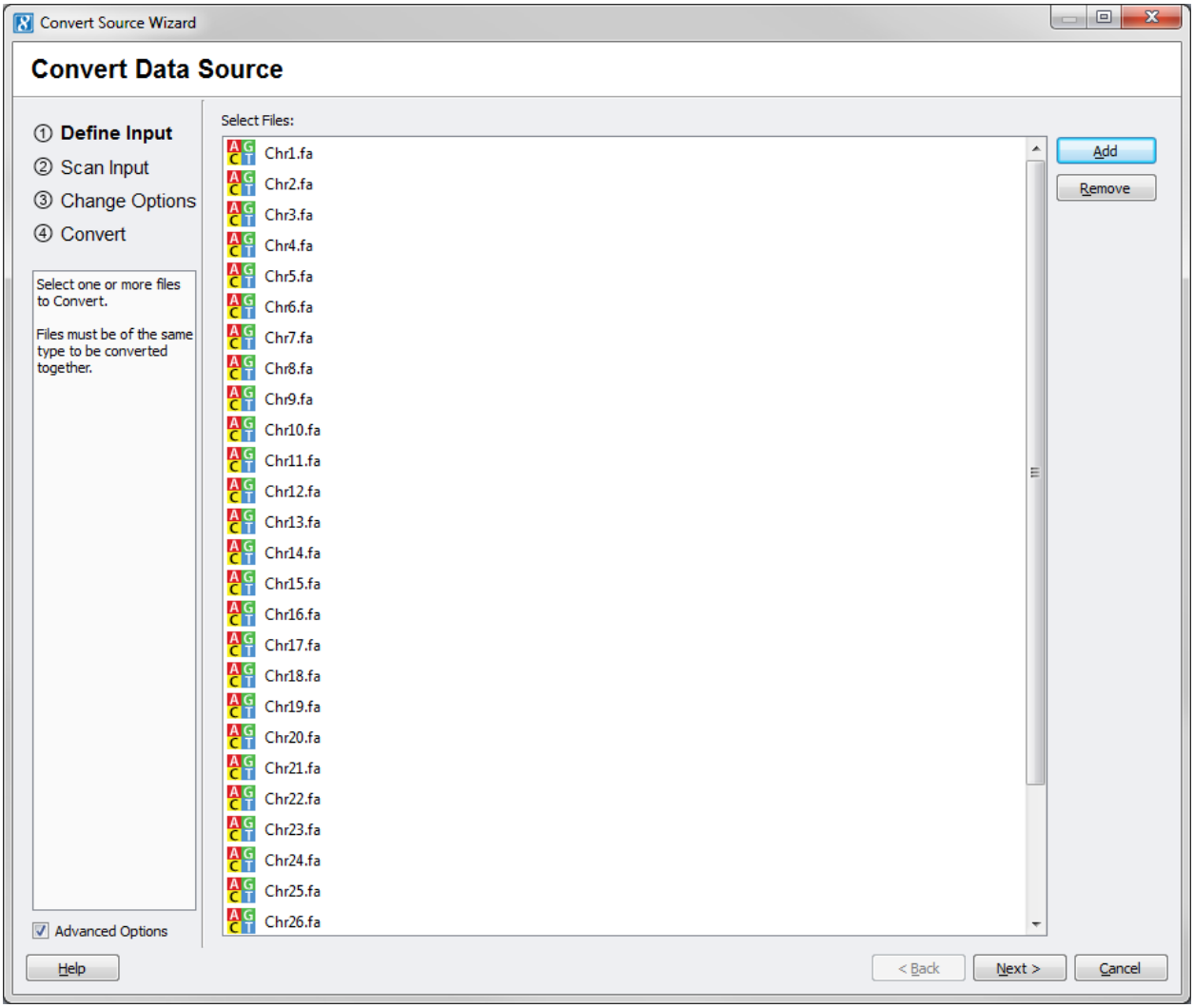



Figure 8.4: File add page with FASTA files in convert wizard

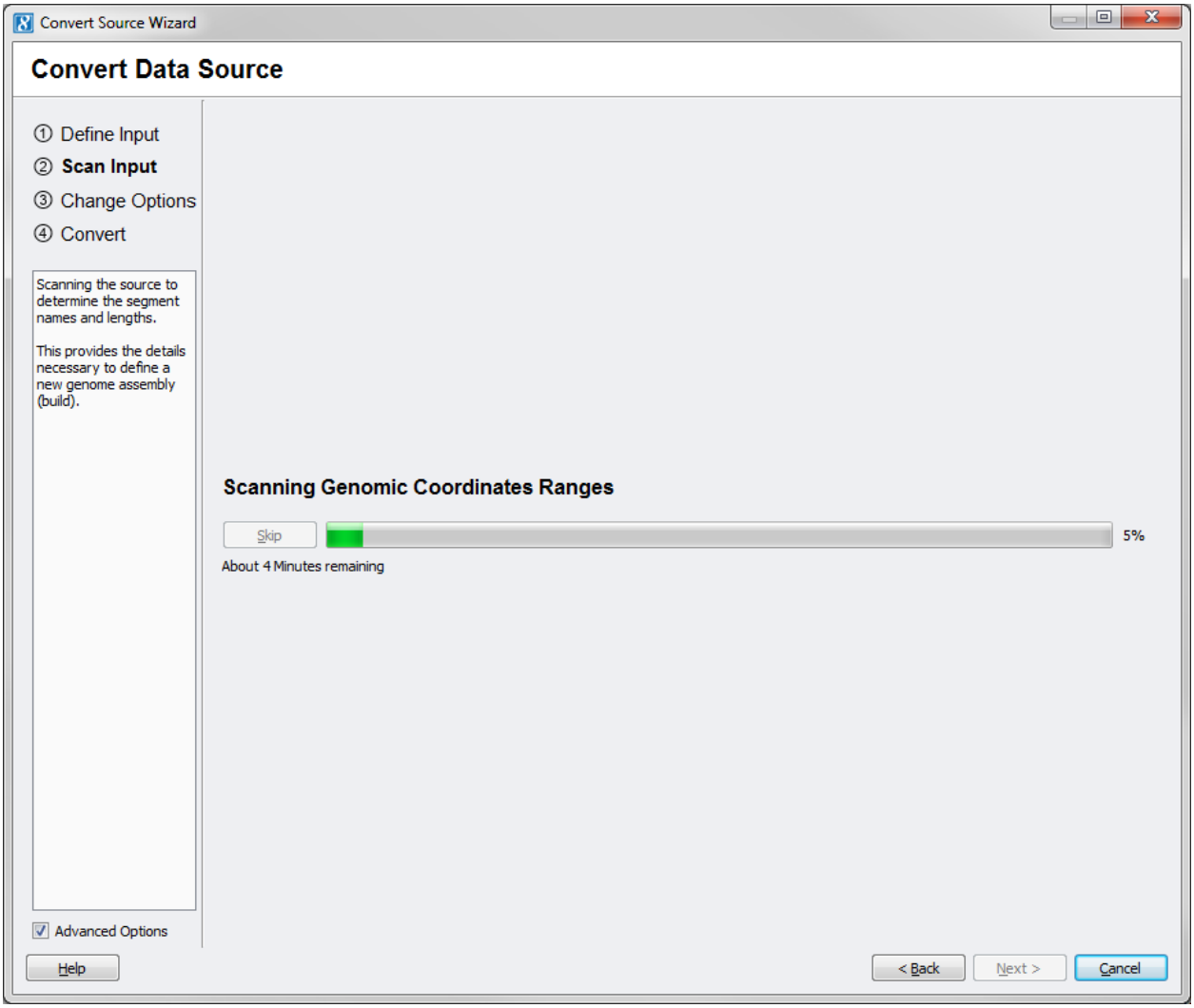



Figure 8.5: Scan FASTA files to get genomic coordinate ranges in convert wizard

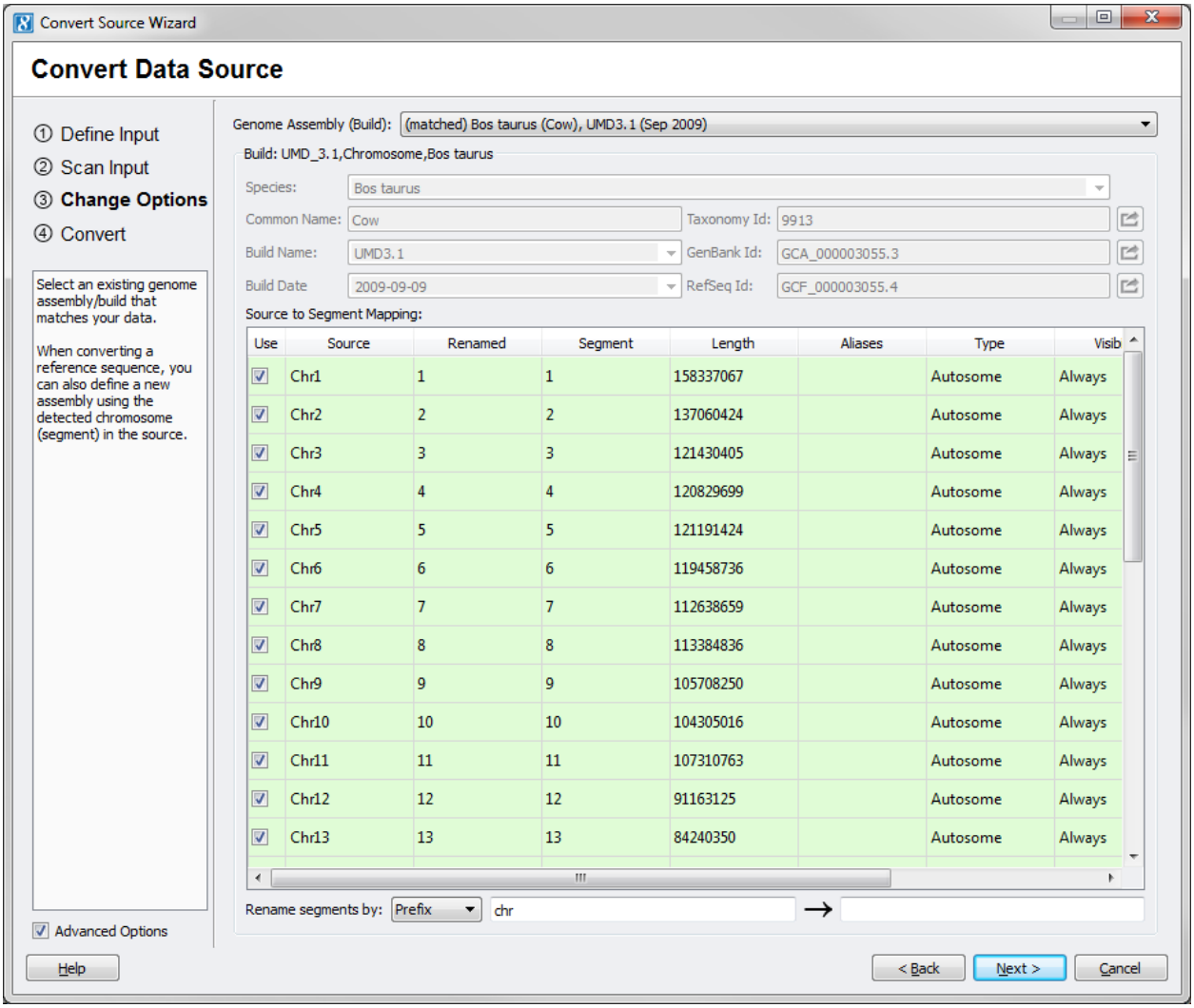



Figure 8.6: Select existing assembly for new allele sequence in convert wizard

The **Source to Segment Mapping** table can be used to exclude or include segments in a new/updated genome assembly file. Segments previously included in the selected genome assembly will have a green background. Segments not included in the selected genome assembly will have a white background. Segment names that exist in the selected assembly but have different lengths between the source and selected assembly file will have a warning icon in front of the segment name. See *Define the Source to Segment Mapping* below for more information.

## Define a new Genome Assembly from FASTA

To define a new genome assembly/build, fill in as many of the available fields as possible:

- **Species: Either select the species from the list or enter in a new one.** The scientific name is preferred. Examples include: 'Homo sapiens', 'Canis familiaris', etc.
- **Common Name: Enter in a common name for the species, such as 'Human' or** 'Dog'.
- **Build Name: The NCBI assembly name is preferred, but the assembly synonym** or UCSC assembly name can be used instead.
- **Build Date**: Submission or published date of the assembly.
- **Taxonomy Id: Taxonomy ID for the species. If the species is in the NCBI** database clicking on the link out button will open a web page on NCBI to help identify the taxonomy ID.
- **GenBank Id: GenBank Assembly ID. If the species is in the NCBI** database clicking on the link out button will open a web page on NCBI to help identify the GenBank ID. This field can be left empty if the species does not have a GenBank ID.
- **RefSeq ID: RefSeq Assembly ID. If the species is in the NCBI** database clicking on the link out button will open a web page on NCBI to help identify the RefSeq ID. This field can be left empty if the species does not have a RefSeq ID.

The **Source to Segment Mapping** table can be used to exclude or include segments in a new/updated genome assembly file. See *Define the Source to Segment Mapping* below for more information.

## Define the Source to Segment Mapping

Segments can either be excluded or renamed from the assembly using the fields in this table. The type of the segment can be set as well as the visibility of the segment.

- **Use: To include a segment in the assembly leave the 'Use' box** checked. To "Check All" or "Uncheck All" click on the "Use" column header. "Uncheck All Unmapped" will not work when creating a new assembly as there are no mapped or unmapped segments.

  **Note:** If there are more than 5000 segments only the 5000 longest segments will be included in the assembly file. If there are more than 500 segments they will be arranged in the assembly file in descending order by length.

- **Source: The name of the segment from the allele sequence file. To** rename a segment, either double click on the name in the **Segment** column, or if the segment names share the same pattern to be removed or for renaming, below the segment definition table are controls for renaming segments programmatically. The options include:
  - *RegEx*: Use regular expressions to rename the "Source".
  - *Substring*: Remove a substring from all segment names to generate the segment name.
  - *Prefix*: Remove a common prefix from all segment names.

  - *Suffix*: Remove a common suffix from all segment names.

  Enter in either the RegEx expression or the string to remove in the first text box. A preview of the renamed segment name will appear in the second text box. To apply the rename to all segments click on the **Set Segment to Renamed** button.

- **Length**: The length of each segment is displayed in this column.
- **Aliases: If a segment has an alias, it can be specified in this column.** For example mitochondrial chromosomes might be named "M" or "MT", the alternate name can be listed in the alias column.
- **Type: Set the type of segment. By default 'Autosomes' are always visible** and the rest are visible only if there is data. Options include:
  - Autosomes
  - Allosome
  - Mitochondrial
  - Fragment
  - Scaffold
  - Contig
  - Unknown
- **Visibility: The visibility of the data can be set manually. If the** segment should only be shown in GenomeBrowse if there is data the visibility can be set to "With Data". Options include:
  - Always
  - Never
  - With Data

Once the assembly has been defined click **Next >**. When the allele sequence is converted to TSF format an assembly file will be created and placed in the User Assembly folder and will be available for use in Golden Helix SVS and GenomeBrowse.

After clicking **Next >** the wizard will display the documentation page. See *Documentation Step* for more information.

## 8.4 Converting a VCF File

A VCF (Variant Call Format) file is a text file in the format specified by 1000 Genomes (http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41, https://github.com/samtools/hts-specs). VCF files will be converted into a variant source with one feature for each line in the VCF file. One field in the new annotation source will be created for each **INFO** field unless the type of the field is **FLAG**. All flags will be combined into one **FLAG** field. If the VCF file contains sample information, the information for each **FORMAT** field will be combined into a delimited list of the appropriate type (string, integer, float, etc).

After selecting one or more VCF sources (VCF, VCF.GZ) on the file selection page of the convert wizard, click **Next >**. If the data has been previously indexed or if **CONTIG** information exists in the header, then the file will not be scanned for genomic coordinates. Otherwise, the file(s) will be scanned to determine the genomic coordinates and most likely genome assembly. The scan may be skipped if the genome assembly is known and the segment naming convention is known.

**Note:** Throughout the Convert Source Wizard there are certain options considered "Advanced" options that do not need to be selected for in most cases. To show "Advanced" options, check the box on the lower left of the dialog. Any option that is advanced will be labeled as such in the documentation.

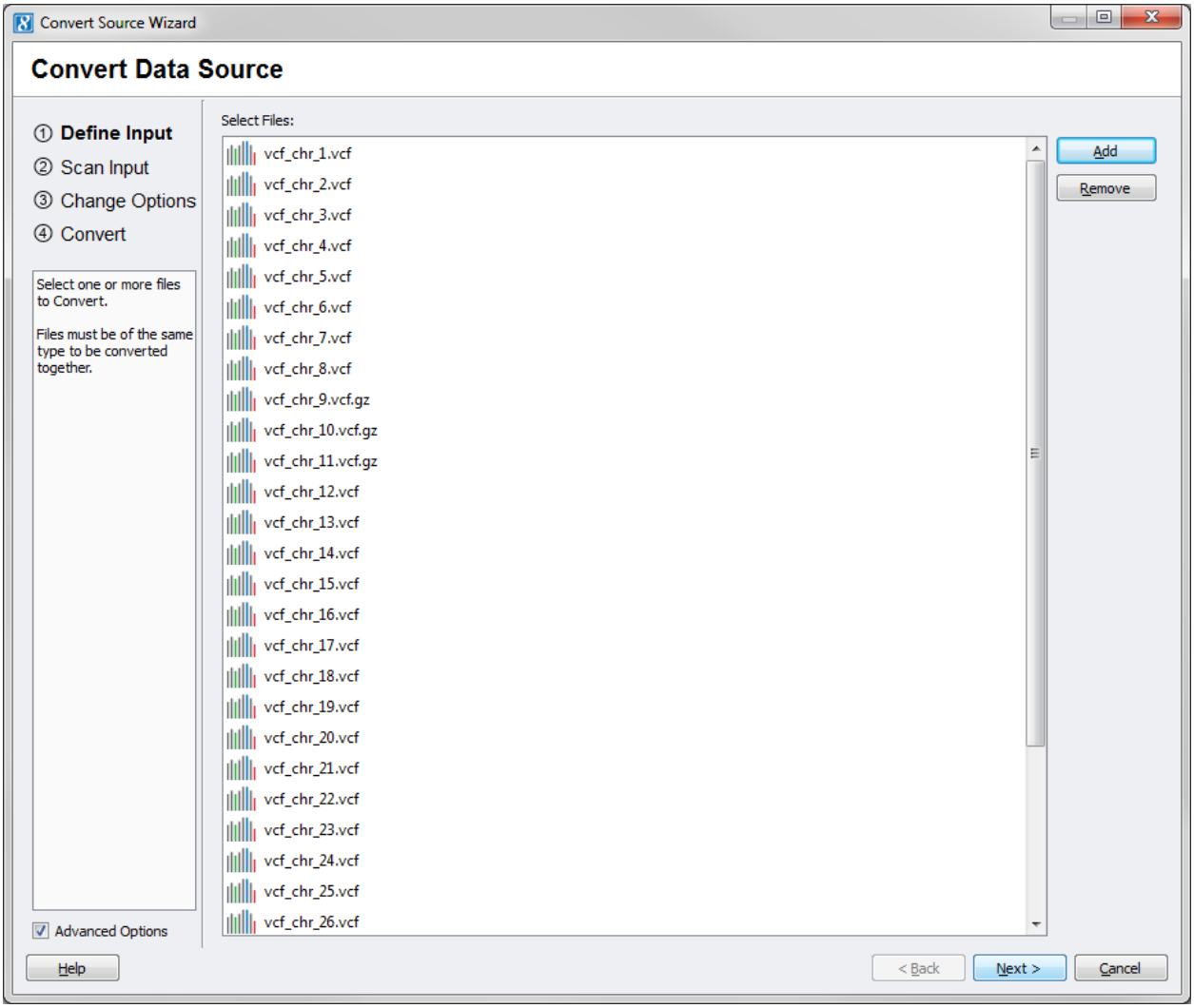



Figure 8.7: Add VCF files in convert wizard

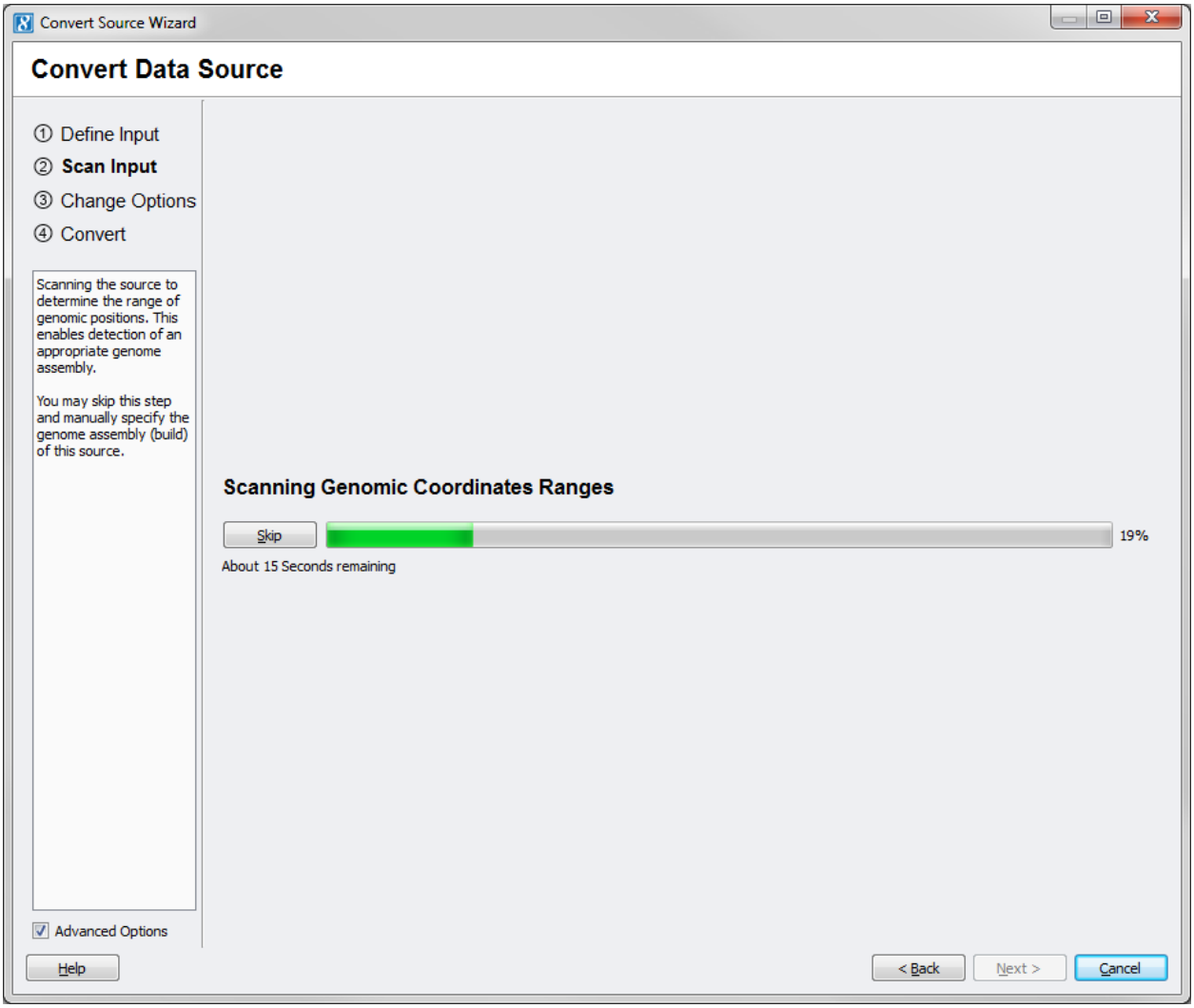



Figure 8.8: Scanning VCF files for genomic coordinate ranges

## Select the Desired Plot Type for VCF

The desired plot type will be automatically detected by default. Unless a VCF file violates the file format specifications, this will always be a Variant plot type. To change the desired plot type, change the selection in the drop down box. If the selected current fields in the file(s) do not meet the specifications for the selected plot type a warning icon will appear on the upper right. To check the required fields for the selected plot type, or to read the warning message(s), hover over the [i]nformation or [!] warning icons. The tool tips will contain the information or warning messages.

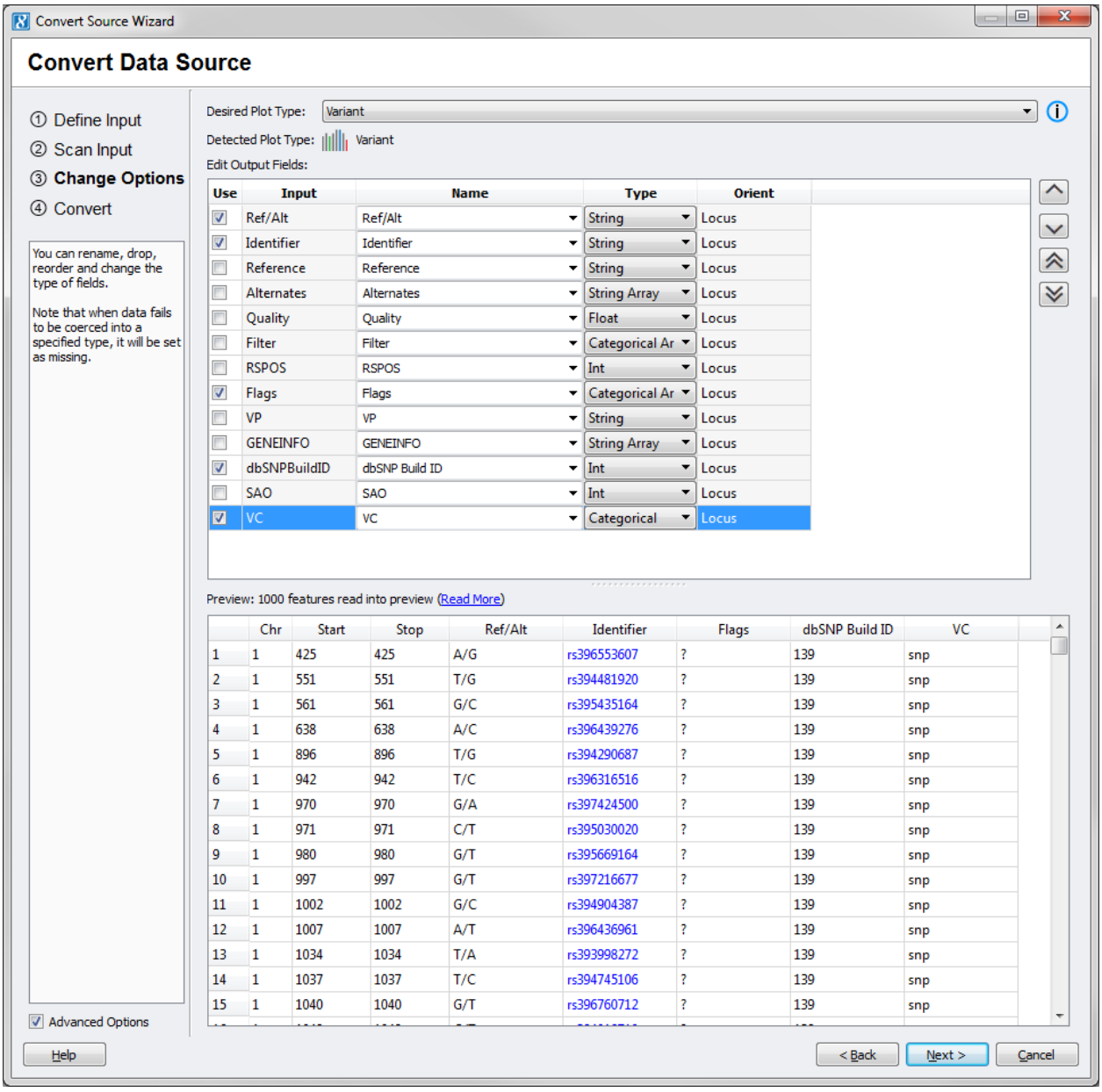



Figure 8.9: Set plot type and field names and types for VCF files in convert wizard

The output fields can be edited on this page as well.

- **Use**: To remove a field from the output uncheck the box in front of the field. For a VCF file, the "Reference" and "Alternate" fields contain redundant information and can be excluded from the source, if desired. The Chr, Start and Stop fields cannot be modified and are not included in the list of editable fields.
- **Rename**: To rename the field, either type in the **Name** box, or select a name of a field required for the plot type in the drop down box. If a required field is renamed a warning will appear for the selected plot type. For a variant source a "Ref/Alt" or "Observed" field is required. From a VCF file the "Ref/Alt" field is created by concatenating the "Reference" and "Alternate" fields.
- **Type**: The default types of the fields are specified based on the type information in the VCF headers. To change the type, select the appropriate type in the drop down **Type** box. For instance, Float64 can be changed to Float, etc. If there are only a few strings used for a string field, "Category" is a better type to use, this will enable filtering on that field. If "Category" is selected but there are too many different categories for a particular field an error will be generated on convert.
- **Reorder Fields**: To reorder fields, click on a field to select (highlight) the row and use the directional arrows on the right of the dialog to move the field either up or down in the list. The double up arrow moves the field to the top of the list. The double down arrow moves the field to the bottom of the list.

The preview pane contains the fields selected for import and will update based on changes made in the **Edit Output Fields** table. By default only the first 1000 features will be read into the preview. To read more features click on **Read More**.

Once the plot type and fields have been edited as desired, click **Next >** to move to the next page of the wizard. See *Select a Genome Assembly* for the next page.

## 8.5 Converting a BED File

A BED file is a text file in the format specified by UCSC (http://genome.ucsc.edu/FAQ/FAQformat.html#format1). It has three required fields (Chromosome, Start and Stop also called chrom, chromStart and chromEnd) and can have up to 9 optional fields. The expected order of the fields is as follows:

1. Chr (chrom)
2. Start (chromStart)
3. Stop (chromEnd)
4. Name
5. Score
6. Strand
7. thickStart
8. thickEnd
9. itemRgb
10. blockCount
11. blockSizes
12. blockStarts

If additional fields are included in the BED file, the field names can be specified on the desired plot type and field specification page.

After selecting one or more BED files (BED, BED.GZ) on the file selection page of the convert wizard, click **Next >**. If the data has not been previously indexed then the file(s) will be scanned to determine the genomic coordinates and

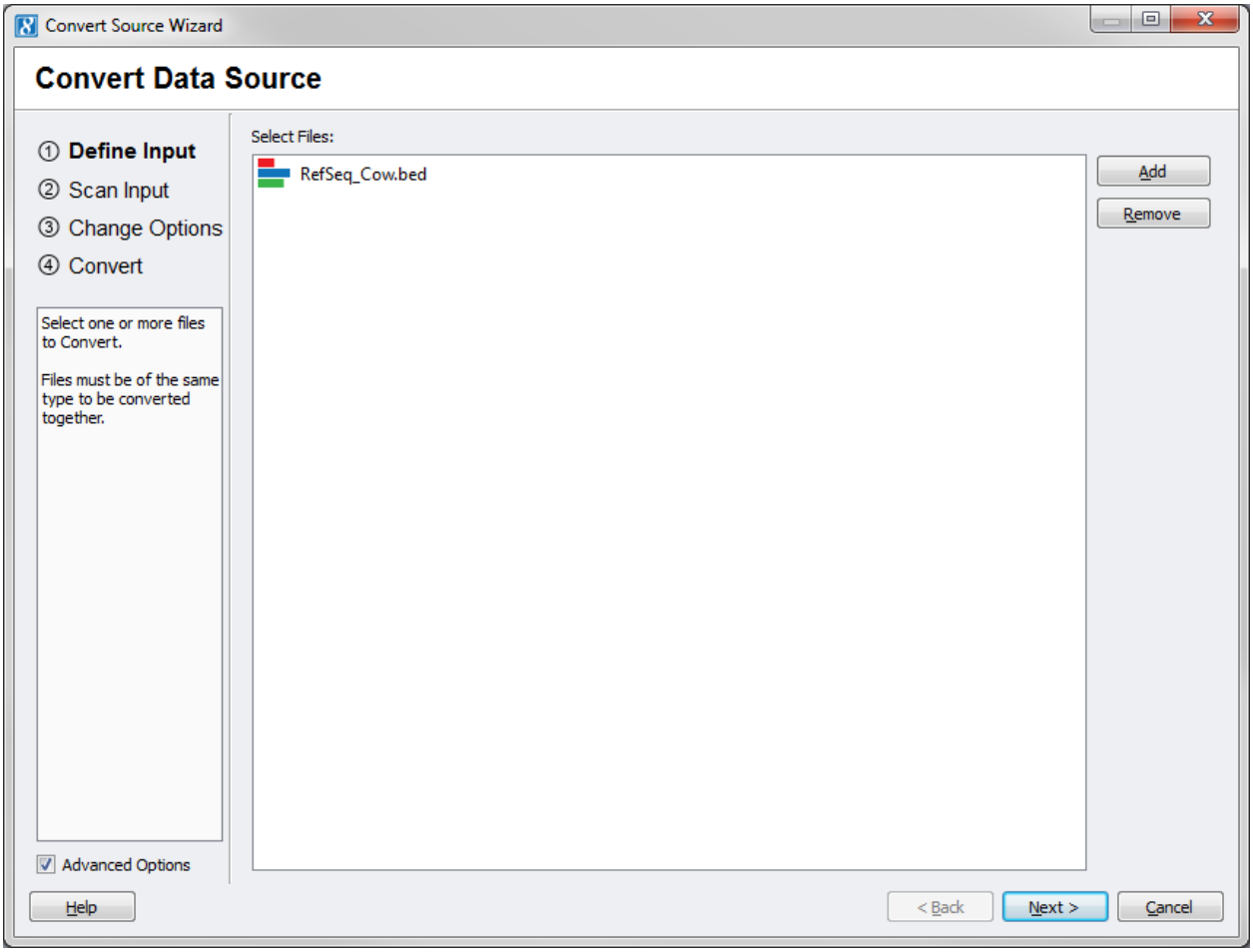



Figure 8.10: Add BED file to convert wizard

most likely genome assembly. The scan may be skipped if the genome assembly is known and the segment naming convention is known.

**Note:** Throughout the Convert Source Wizard there are certain options considered "Advanced" options that do not need to be selected for in most cases. To show "Advanced" options, check the box on the lower left of the dialog. Any option that is advanced will be labeled as such in the documentation.

## Select the Desired Plot Type for BED

The desired plot type will be automatically detected by default. Unless a BED file violates the file format specifications, this will always be a Generic Interval plot type. To change the desired plot type, change the selection in the drop down box. If the selected current fields in the file(s) do not meet the specifications for the selected plot type a warning icon will appear on the upper right. To check the required fields for the selected plot type, or to read the warning message(s), hover over the [i]nformation or [!] warning icons. The tool tips will contain the information or warning messages.

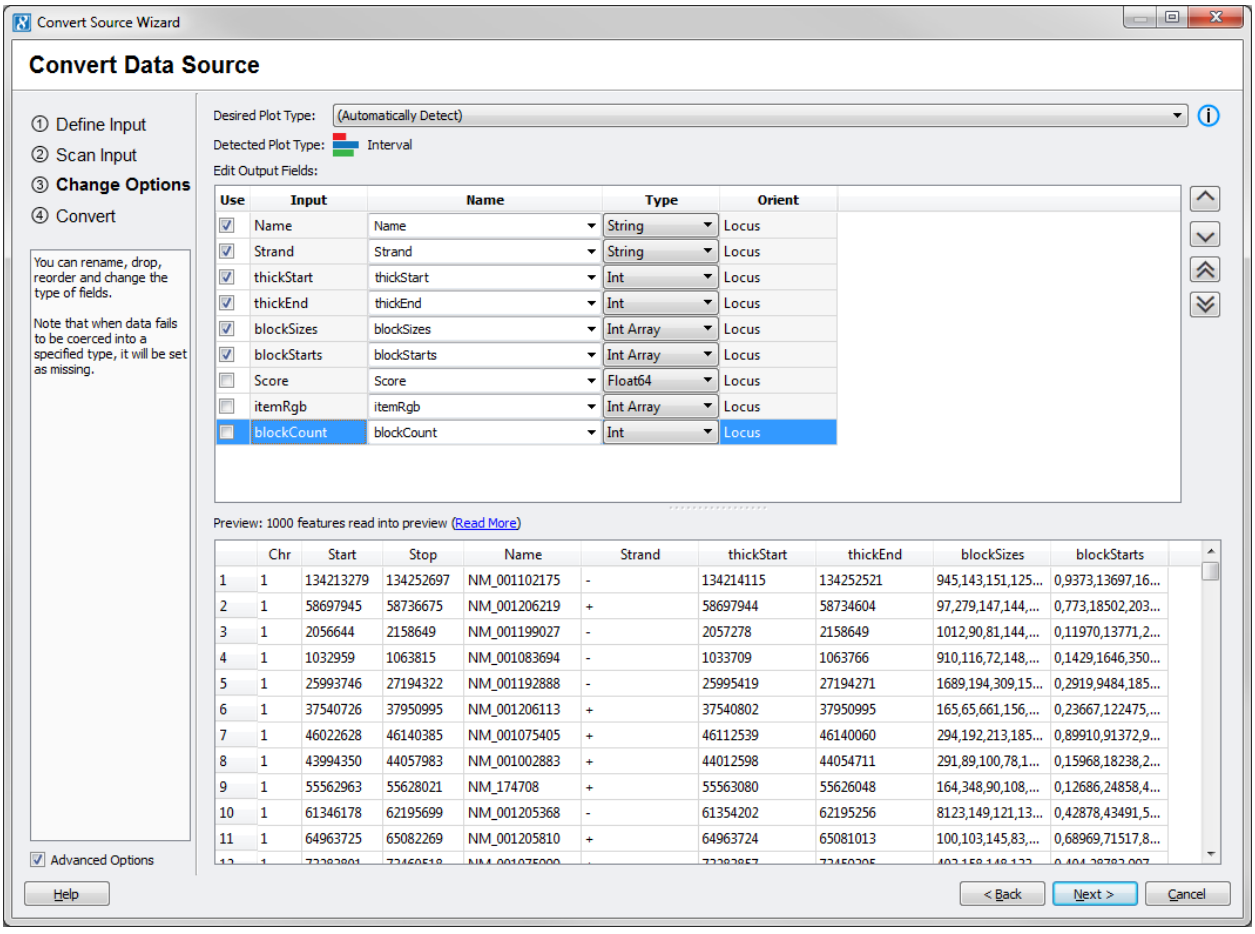



Figure 8.11: Select plot type and field specifications for BED file in convert wizard

The output fields can be edited on this page as well.

- **Use**: To remove a field from the output uncheck the box in front of the field. The Chr, Start and Stop fields cannot be modified and are not included in the list of editable fields.

- **Rename**: To rename the field, either type in the **Name** box, or select a name of a field required for the plot type in the drop down box. If a required field is renamed a warning will appear for the selected plot type. For a generic interval source a "Name" or "Identifier" field is required.
- **Type**: The default types of the fields are specified based on the types defined in the BED specifications. To change the type, select the appropriate type in the drop down box. For instance, Float64 can be changed to Float, etc. If there are only a few strings used for a string field, "Category" is a better type to use, this will enable filtering on that field. If "Category" is selected but there are too many different categories for a particular field an error will be generated on convert.
- **Reorder Fields**: To reorder fields, click on a field to select (highlight) the row and use the directional arrows on the right of the dialog to move the field either up or down in the list. The double up arrow moves the field to the top of the list. The double down arrow moves the field to the bottom of the list.

The preview pane contains the fields selected for import and will update based on changes made in the **Edit Output Fields** table. By default only the first 1000 features will be read into the preview. To read more features click on **Read More**.

Once the plot type and fields have been edited as desired, click **Next >** to move to the next page of the wizard. See *Select a Genome Assembly* for the next page.

## 8.6 Converting a GTF File

A GTF (Gene Transfer Format) file is a text file in the format specified by Washington University in St. Louis (http://mblab.wustl.edu/GTF22.html and http://genome.ucsc.edu/FAQ/FAQformat.html#format4). Two requirements are that the GTF file contain a gene_id and transcript_id for each feature. If a gene_name field is also available the gene_name will be preferred over the gene_id for labeling the features in the annotation source. If the GTF file does not follow the rigid specifications try using the *Convert GFF Files to Annotation Track* tool.

After selecting one or more GTF files (GTF, GTF.GZ) on the file selection page of the convert wizard, click **Next >**. If the data has not been previously indexed then the file(s) will be scanned to determine the genomic coordinates and most likely genome assembly. The scan may be skipped if the genome assembly is known and the segment naming convention is known.

**Note:** Throughout the Convert Source Wizard there are certain options considered "Advanced" options that do not need to be selected for in most cases. To show "Advanced" options, check the box on the lower left of the dialog. Any option that is advanced will be labeled as such in the documentation.

### Select the Desired Plot Type for GTF

The desired plot type will be automatically detected by default. Unless a GTF file violates the file format specifications, this will always be a Gene plot type. To change the desired plot type, change the selection in the drop down box. If the selected current fields in the file(s) do not meet the specifications for the selected plot type a warning icon will appear on the upper right. To check the required fields for the selected plot type, or to read the warning message(s), hover over the [i]nformation or [!] warning icons. The tool tips will contain the information or warning messages.

The output fields can be edited on this page as well.

- **Use**: To remove a field from the output uncheck the box in front of the field. The Chr, Start and Stop fields cannot be modified and are not included in the list of editable fields.
- **Rename**: To rename the field, either type in the **Name** box, or select a name of a field required for the plot type in the drop down box. If a required field is renamed a warning will appear for the selected plot type. For a gene source the following are required "Gene Name", "Transcript Name", "Strand", "CDS Start", "CDS Stop", "Exon Starts" and "Exon Stops" are required.

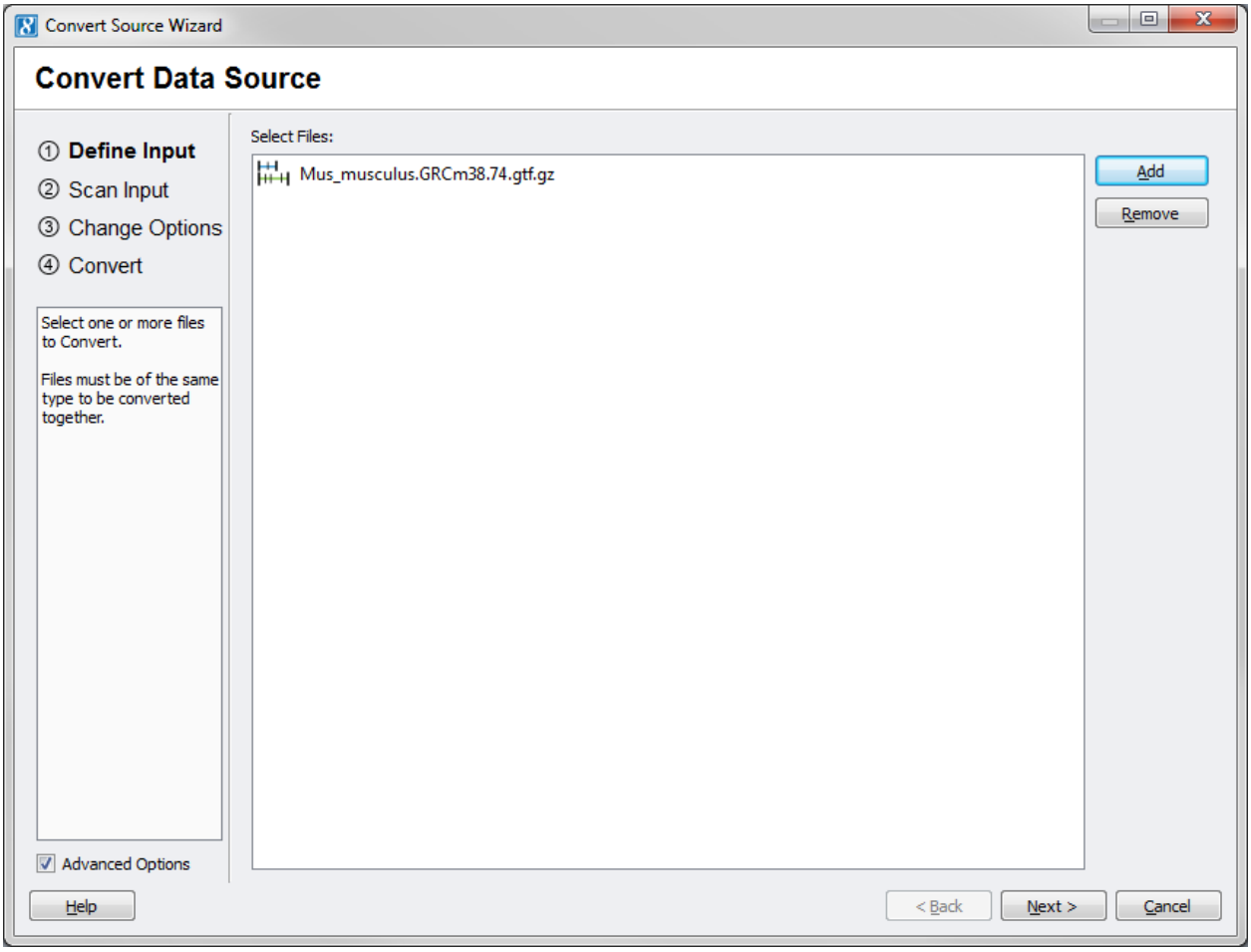



Figure 8.12: Add GTF file to convert wizard

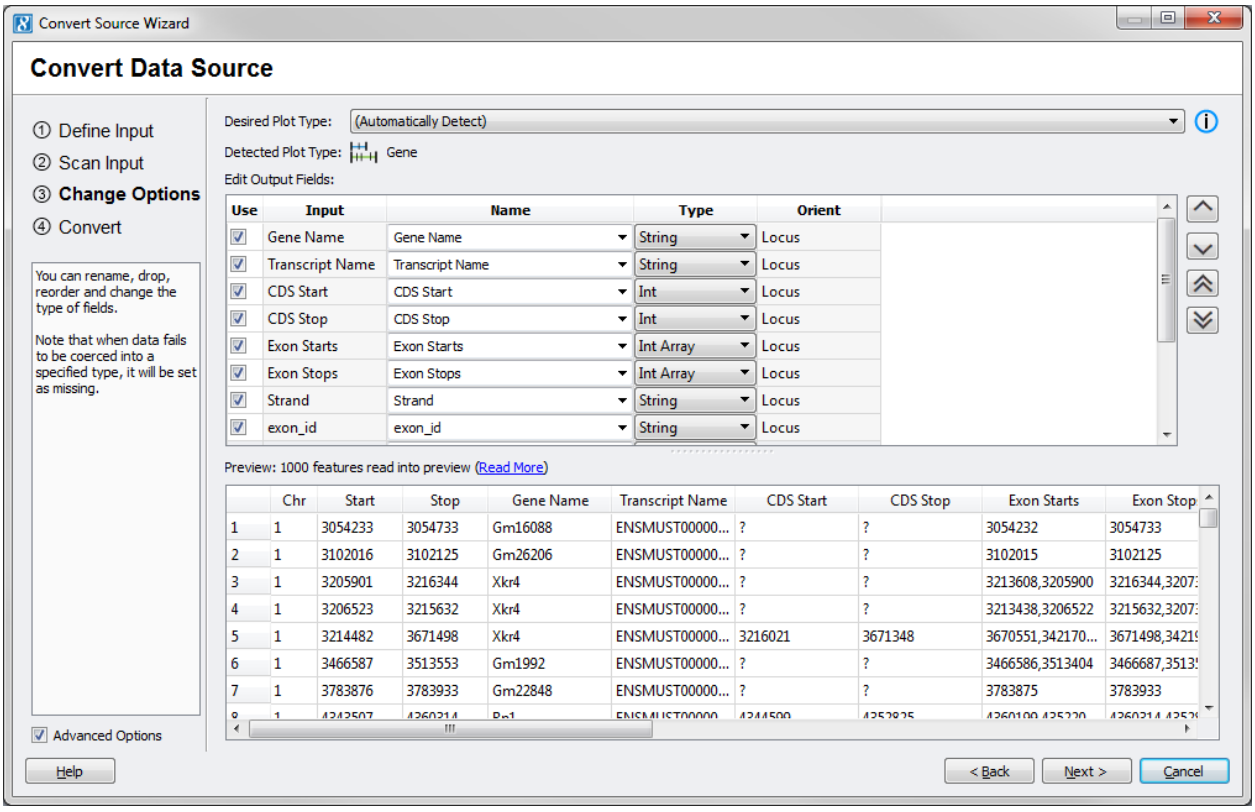



Figure 8.13: Specify plot type and field specifications for GTF in convert wizard

- **Type**: The default types of the fields are specified based on the types detected in the scan pass. To change the type, select the appropriate type in the drop down box. For instance, Float64 can be changed to Float, etc. If there are only a few strings used for a string field, "Category" is a better type to use, this will enable filtering on that field. If "Category" is selected but there are too many different categories for a particular field an error will be generated on convert.
- **Reorder Fields**: To reorder fields, click on a field to select (highlight) the row and use the directional arrows on the right of the dialog to move the field either up or down in the list. The double up arrow moves the field to the top of the list. The double down arrow moves the field to the bottom of the list.

The preview pane contains the fields selected for import and will update based on changes made in the **Edit Output Fields** table. By default only the first 1000 features will be read into the preview. To read more features click on **Read More**.

Once the plot type and fields have been edited as desired, click **Next >** to move to the next page of the wizard. See *Select a Genome Assembly* for the next page.

## 8.7 Converting a WIG (Fixed or Variable Step) File

A WIG file is a text file which assigns a floating point value to each position or interval of interest in genomic space. Wiggle files can either be formatted as variable step or fixed step (variableStep and fixedStep respectively). See http://genome.ucsc.edu/goldenPath/help/wiggle.html for more information on the format.

After selecting one or more WIG files (WIG, WIG.GZ, WIGFIX, WIGFIX.GZ) on the file selection page of the convert wizard, click **Next >**. If the data has not been previously indexed then the file(s) will be scanned to determine the genomic coordinates and most likely genome assembly. The scan may be skipped if the genome assembly is known and the segment naming convention is known. For WIG files, the only available output is a data sequence source so the next page of the convert wizard is the genome assembly specification page. See *Select a Genome Assembly* for information on the next page.

**Note:** Throughout the Convert Source Wizard there are certain options considered "Advanced" options that do not need to be selected for in most cases. To show "Advanced" options, check the box on the lower left of the dialog. Any option that is advanced will be labeled as such in the documentation.

## 8.8 Converting a Delimited Text File

A delimited text file is a text file consisting of rows and columns of data delimited by special characters or strings which are not allowed in the data values themselves. The delimited text import can take multiple files as long as they are formatted in the same manner. Because the input format is customizable, a delimited file characteristics page is included in the convert source wizard to allow the user to indicate how the file(s) should be parsed.

After selecting one or more text files (TXT, TXT.GZ, TSV, TSV.GZ, CSV, CSV.GZ, etc.) on the file selection page, the icon next to the files will be a question mark. This is because text files can contain data for numerous plot types and that cannot be determined by the file extension. If the files are a format that cannot be read or processed by the user or are of inconsistent types a warning will appear or prevent progressing to the next page of the convert wizard. To move to the next step of the conversion process, click **Next >**.

**Note:** Throughout the Convert Source Wizard there are certain options considered "Advanced" options that do not need to be selected for in most cases. To show "Advanced" options, check the box on the lower left of the dialog. Any option that is advanced will be labeled as such in the documentation.

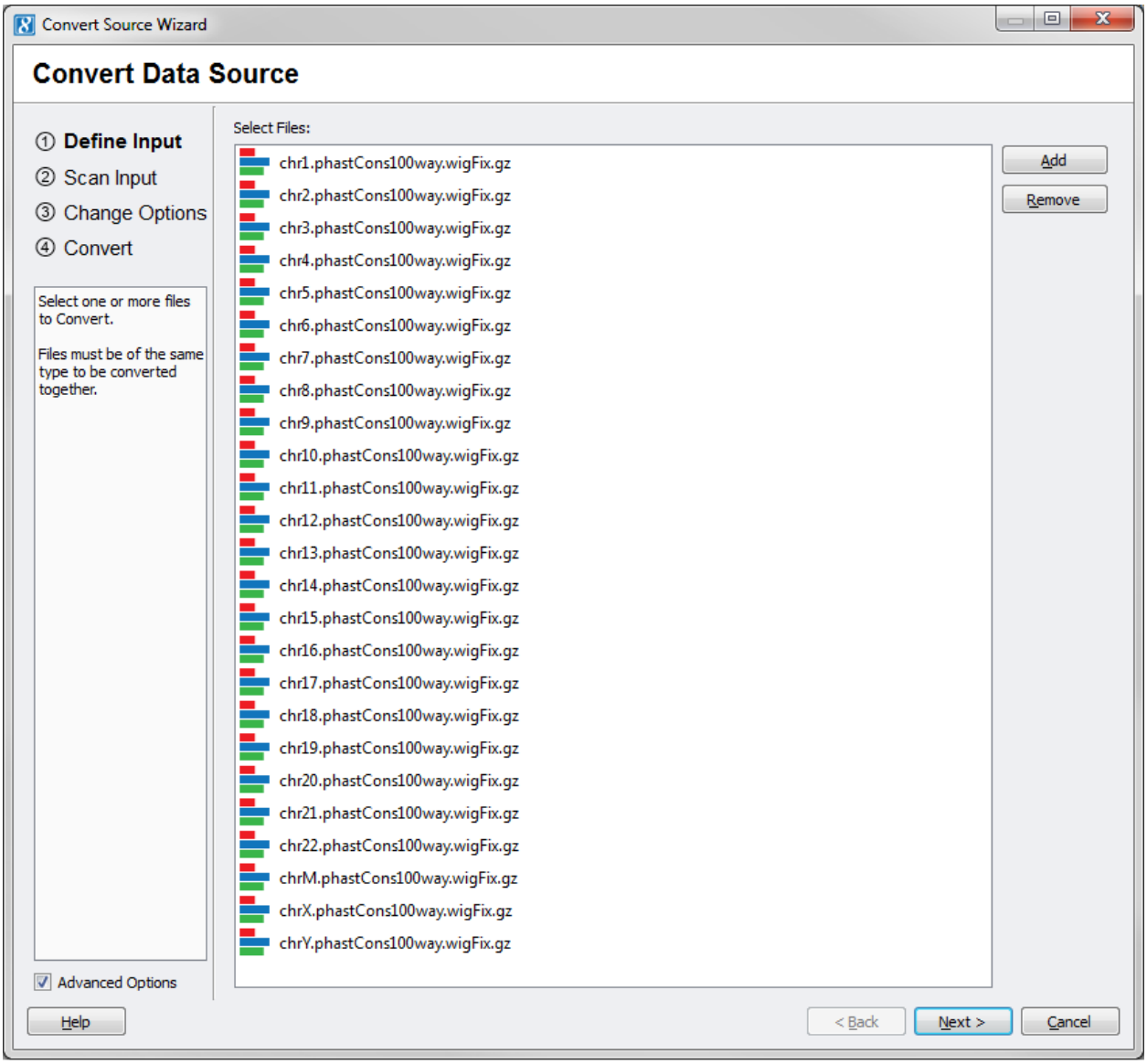



Figure 8.14: Add WIG files to convert wizard

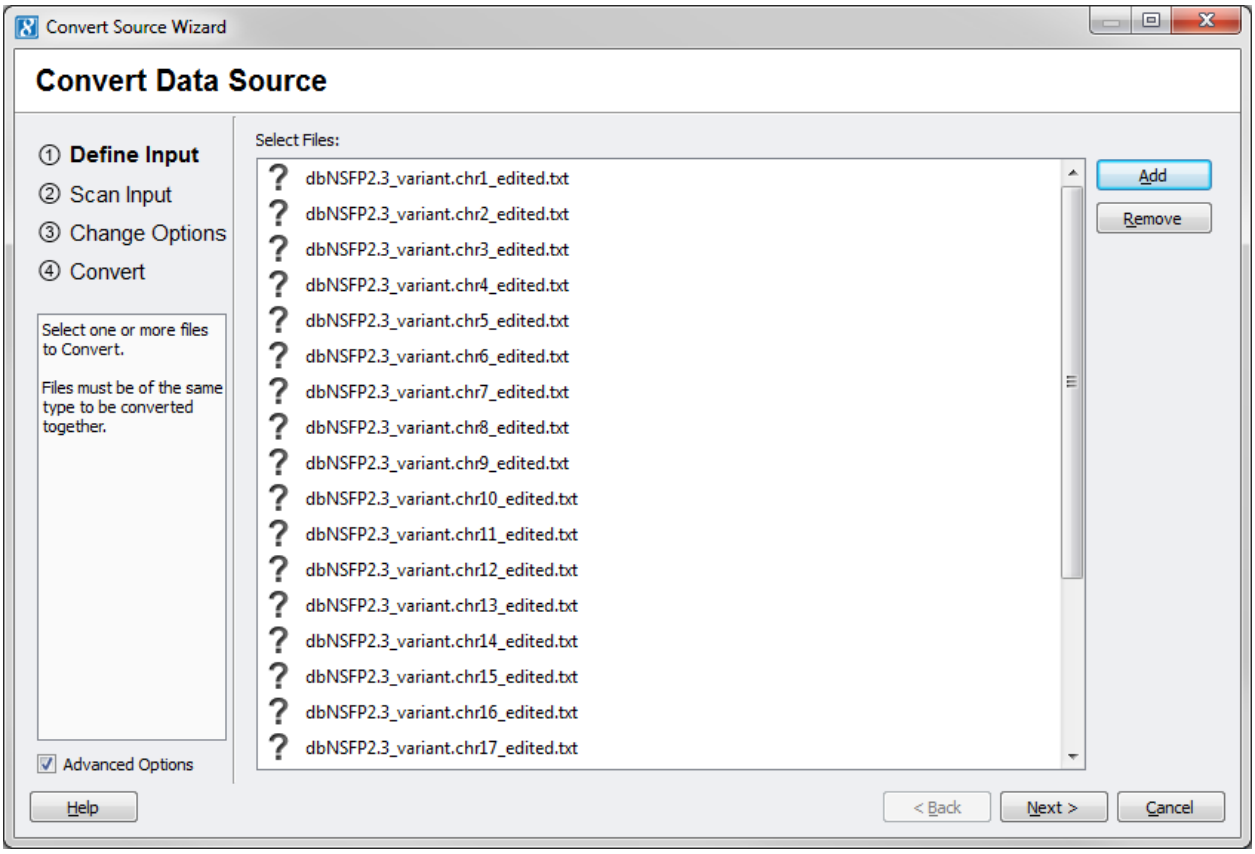



Figure 8.15: Add TXT files to convert wizard

## Specify the Delimited File Characteristics

As delimited text files can come in many different formats, additional information is required to parse the file(s). This information is specified on the **Delimited Text File Characteristics** page of the convert wizard. The options that can be specified are described below.

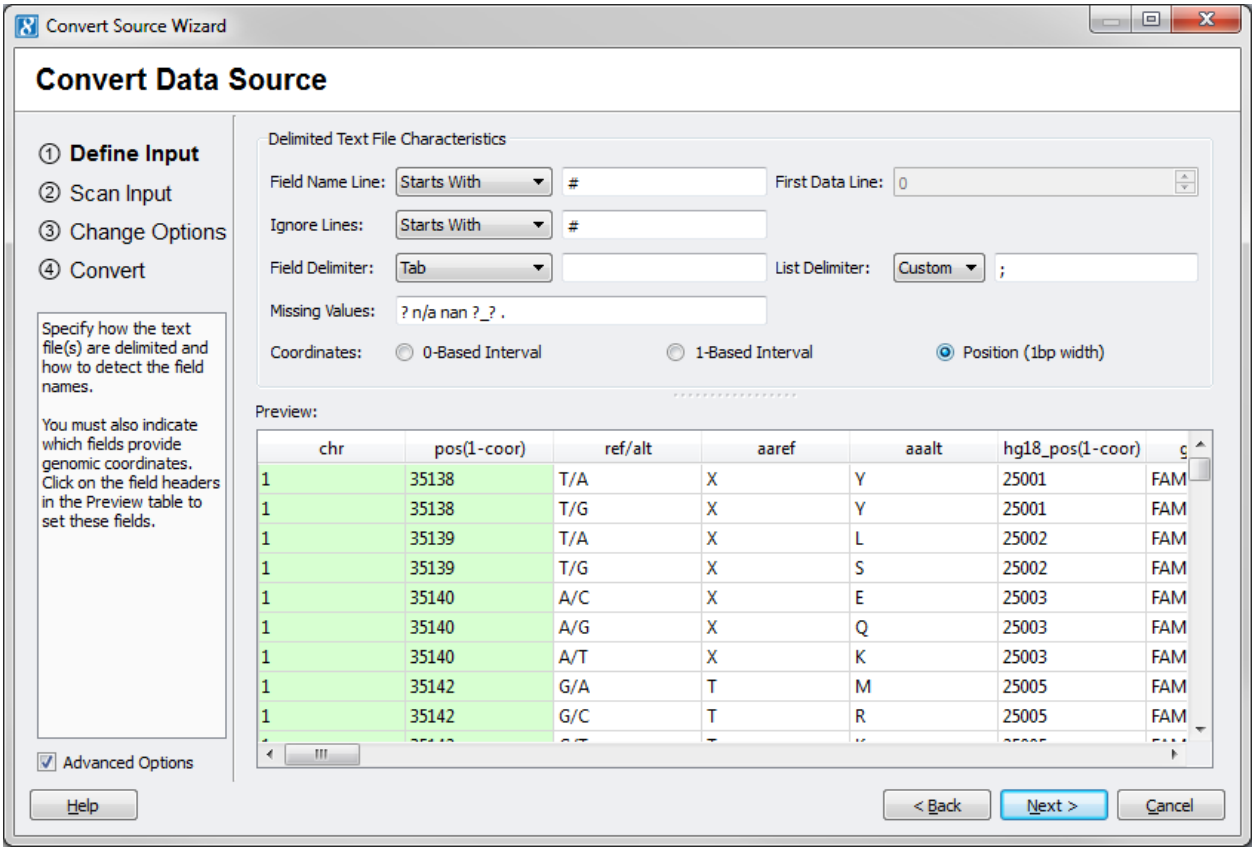



Figure 8.16: Specify the options for the format of the delimited text file(s)

- **Field Name Line**: The files may contain a line which defines the names of the data fields in the file(s). The header line must be near the top of the file(s) and any text above the header line will be ignored during import. The header line may be detected by:
  - **Starts With**: A single or multiple character string
  - **Line Before Data**: The line immediately proceeding the first data line. Selecting this option activates the **First Data Line** option. If not in *Advanced* mode, it is assumed that the first data line is the second line in the file. [Line 1]
  - **Manual Names**: If no header line exists, select this option to auto generate field names that can be edited on the plot type/output field name specification page. Selecting this option activates the **First Data Line** option. If not in *Advanced* mode, it is assumed that the first data line is the first line in the file. [Line 0]
- **First Data Line**: [*Advanced Option*] Available when **Line Before Data** or **Manual Names** are selected for **Field Name Line**. Data lines are in 0-based indexes, i.e. the first line of the file is line number 0, the second line is line number 1, etc.
- **Ignore Lines**: [*Advanced Option*] The input file(s) may also contain any number of lines which should be

ignored during conversion. Such lines are often referred to as comments because they are intended to include additional notes but do not include data. The controls used to indicate which lines should be treated as comments are similar to the header line controls. Comments may be detected by:

  - **Starts With**: A single or multiple character string at the beginning of the line.
  - **Ends With**: A single or multiple character string at the end of the line.
  - **Contains**: Any line that contains the specified string.
  - **Equals**: Any line that is exactly equal to the specified string.
  - **Wildcard**: Any line that contains the wildcard character.
  - **RegEx**: A regular expression can be used to indicate lines to ignore using a more complicated pattern.
  - **Don't Ignore**: Don't ignore lines that contain the specified string.

- **Field Delimiter**: The string that separates data values on each line of the input file(s). For correct alignment of the converted data, the delimiter should not occur within any of the data values.
- **List Delimiter**: [*Advanced Option*] The list delimiter specifies the string that separates data values items within a single field. By default this delimiter is a comma if the file is tab delimited or a custom delimiter and a semi-colon if the file is comma delimited.
- **Missing Values**: [*Advanced Option*] A list of common missing value indicators is included in the conversion wizard by default. To view this list or to add or remove a missing value indicator, select **Advanced Options** and edit the space delimited list of common missing value strings.
- **Coordinates**: The coordinates option is used to specify whether the intervals defined in the input data are 0-based intervals, 1-based intervals or positions.
  - **0-Based Interval**: The difference between the stop and the start positions defines the width of the interval. For example, an interval covering the first three positions of a chromosome in 0-based coordinates would be specified as [0, 3]. (Also known as 'half-open coordinates'.)
  - **1-Based Interval**: The difference between the stop and the start positions plus one defines the width of the interval. For example, an interval covering the first three positions of a chromosome in 1-based coordinates would be specified as [1, 3]. (Also known as 'indexed coordinates'.)
  - **Position (1bp width)**: For files with only a single position/coordinate select this option for the coordinates. This option assumes all features have a single base pair width. The position is 1-based so the smallest position in a chromosome would be 1.
- **Preview**: The chromosome, start, stop or position columns are selected by default when ever possible. However, it is likely that the wrong fields are selected for the chromosome, start, stop or position. To change the fields used for the genomic coordinates, click on the field name and specify whether the field is the *Chromosome (Segment) Field*, *Start Field* (for interval coordinates only), *Stop Field* (for interval coordinates only), or *Position Field* (for position coordinates only).

Once the delimited text file characteristics and fields to use to define genomic coordinates are specified, click **Next >**. If the data has not been previously indexed then the file(s) will be scanned to determine the data types and genomic coordinates and most likely genome assembly. The scan may be skipped if the data types, genome assembly, and the segment naming convention are known.

### Select the Desired Plot Type for Delimited Text

The desired plot type will be automatically detected by default. To change the desired plot type, change the selection in the drop down box. If the selected current fields in the file(s) do not meet the specifications for the selected plot type a warning icon will appear on the upper right. To check the required fields for the selected plot type, or to read

the warning message(s), hover over the [i]nformation or [!] warning icons. The tool tips will contain the information or warning messages.

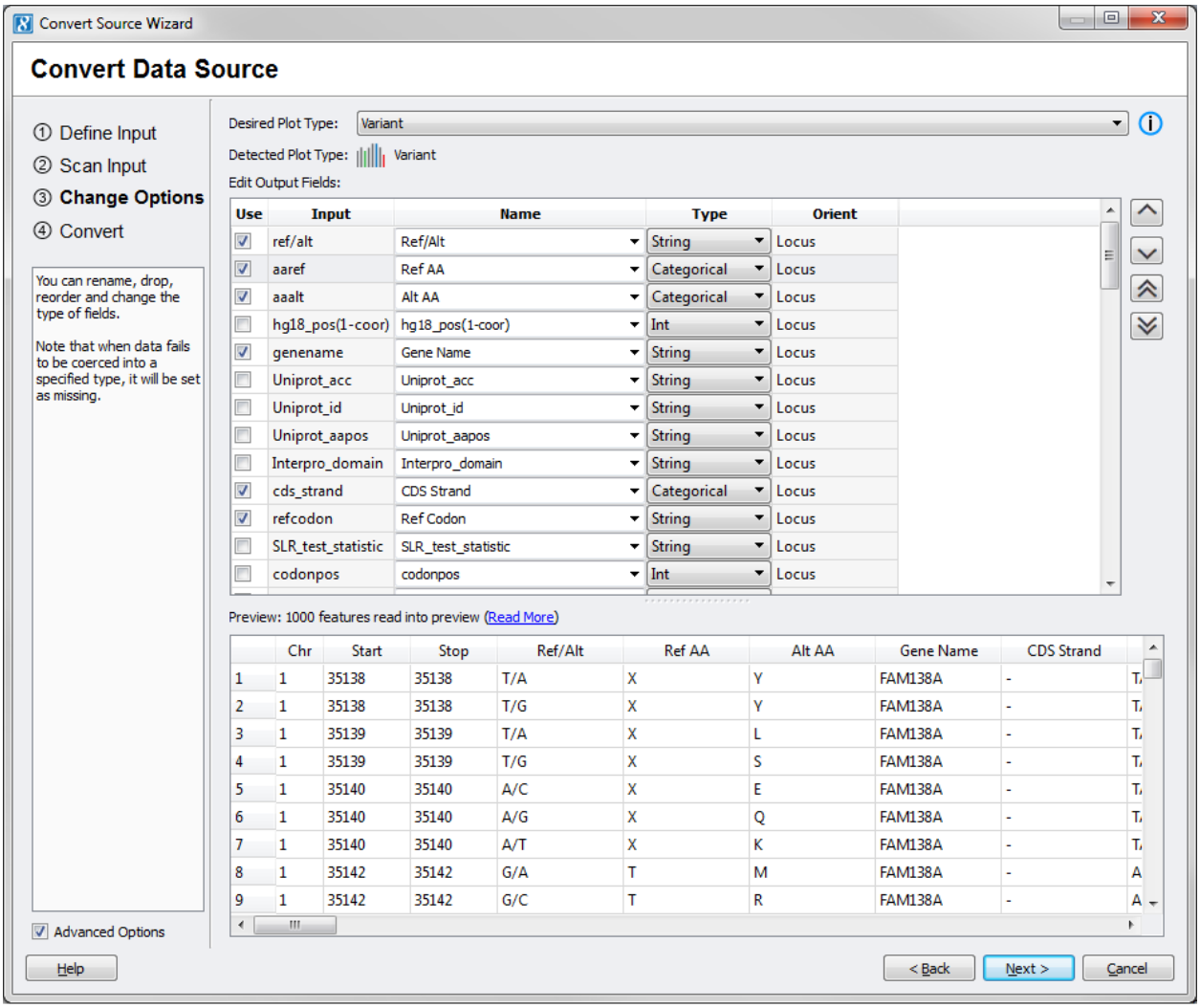



Figure 8.17: Specify the file type and output fields for delimited text file conversion

The output fields can be edited on this page as well.

- **Use**: To remove a field from the output uncheck the box in front of the field. The Chr, Start and Stop fields cannot be modified and are not included in the list of editable fields.
- **Rename**: To rename the field, either type in the **Name** box, or select a name of a field required for the plot type in the drop down box. If a required field is renamed a warning will appear for the selected plot type.
- **Type**: The default types of the fields are specified based on the types detected in the scan pass. To change the type, select the appropriate type in the drop down box. For instance, Float64 can be changed to Float, etc. If there are only a few strings used for a string field, "Category" is a better type to use, this will enable filtering on

that field. If "Category" is selected but there are too many different categories for a particular field an error will be generated on convert.

- **Reorder Fields**: To reorder fields, click on a field to select (highlight) the row and use the directional arrows on the right of the dialog to move the field either up or down in the list. The double up arrow moves the field to the top of the list. The double down arrow moves the field to the bottom of the list.

The preview pane contains the fields selected for import and will update based on changes made in the **Edit Output Fields** table. By default only the first 1000 features will be read into the preview. To read more features click on **Read More**.

Once the plot type and fields have been edited as desired, click **Next >** to move to the next page of the wizard. See *Select a Genome Assembly* for the next page.

## 8.9 Converting an IDF or TSF File

An IDF file is an annotation track source from SVS 7. It can be converted to the new TSF format to optimize data storage as well as take advantage of the embedded genome assembly, coverage, and documentation.

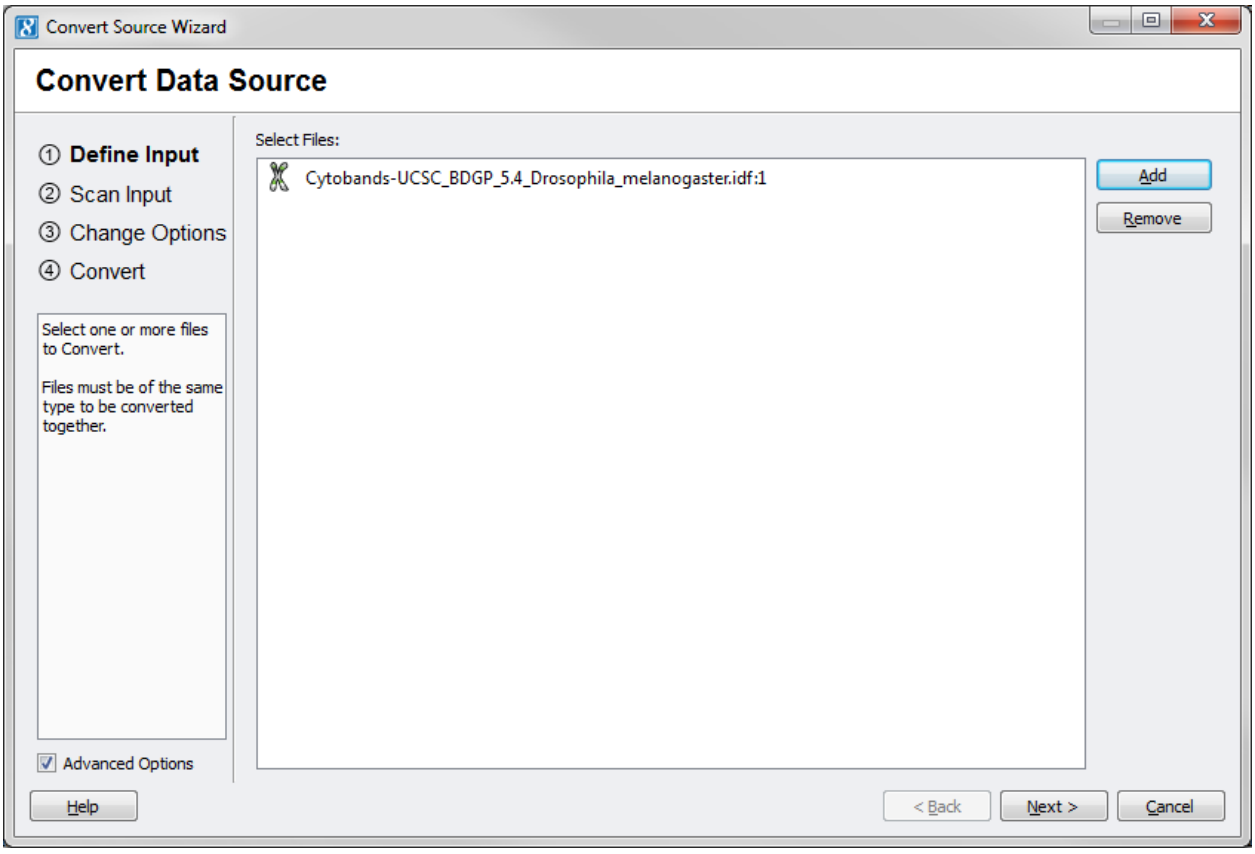



Figure 8.18: Add an IDF file to convert wizard

If the IDF file is an allele sequence source, the file will be treated like a 2Bit or FASTA file in the convert wizard. See *Convert a 2Bit File* or *Converting a FASTA File* for more information.

If the IDF file is any other source type, the file will be treated like an indexed file and the next page in the convert wizard will be specifying the plot type and output fields.

**Note:** Throughout the Convert Source Wizard there are certain options considered "Advanced" options that do not need to be selected for in most cases. To show "Advanced" options, check the box on the lower left of the dialog. Any option that is advanced will be labeled as such in the documentation.

## Select the Desired Plot Type for IDF or TSF

The desired plot type will be automatically set to the type of the IDF or TSF source being converted. To change the desired plot type, change the selection in the drop down box. If the selected current fields in the file does not meet the specifications for the selected plot type, a warning icon will appear on the upper right. To check the required fields for the selected plot type, or to read the warning message(s), hover over the [i]nformation or [!] warning icons. The tool tips will contain the information or warning messages.

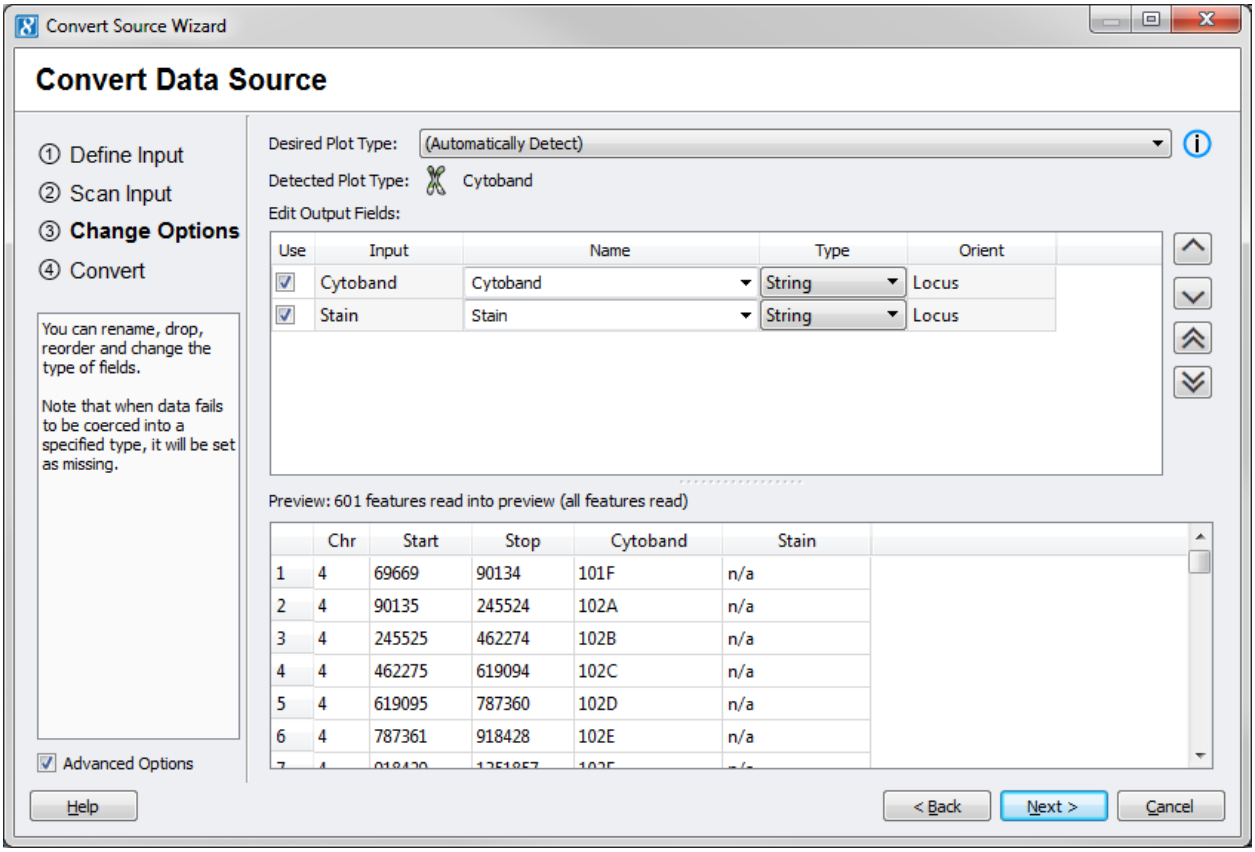



Figure 8.19: Specify type and output field options for IDF cytoband conversion

The output fields can be edited on this page as well.

- **Use**: To remove a field from the output uncheck the box in front of the field. The Chr, Start and Stop fields cannot be modified and are not included in the list of editable fields.
- **Rename**: To rename the field, either type in the **Name** box, or select a name of a field required for the plot type in the drop down box. If a required field is renamed a warning will appear for the selected plot type.
- **Type**: The default types of the fields are specified based on the types detected in the scan pass. To change the type, select the appropriate type in the drop down box. For instance, Float64 can be changed to Float, etc. If there are only a few strings used for a string field, "Category" is a better type to use, this will enable filtering on

that field. If "Category" is selected but there are too many different categories for a particular field an error will be generated on convert.

- **Reorder Fields**: To reorder fields, click on a field to select (highlight) the row and use the directional arrows on the right of the dialog to move the field either up or down in the list. The double up arrow moves the field to the top of the list. The double down arrow moves the field to the bottom of the list.

The preview pane contains the fields selected for import and will update based on changes made in the **Edit Output Fields** table. By default only the first 1000 features will be read into the preview. To read more features click on **Read More**.

Once the plot type and fields have been edited as desired, click **Next >** to move to the next page of the wizard. See *Select a Genome Assembly* for the next page.

## 8.10 Select a Genome Assembly

Depending on how much of the file(s) are scanned the convert source wizard will pick the most likely genome assembly based on the genomic coordinates (chromosomes and largest positions) as the default genome assembly. If this default is not the correct species or build, select the correct assembly from the drop down list. Note, all species that matched the features read will be at the top of the list. Scrolling down past the list of matches will present the available assemblies in alphabetical order based on the scientific name.

Once the assembly has been selected, the source to segment mapping can be modified as required.

### Modify the Source to Segment Mapping

Segments can either be excluded or renamed from the annotation source using the fields in this table.

- **Use: To include a segment in the assembly leave the 'Use' box** checked. To "Check All" or "Uncheck All" click on the "Use" column header. "Uncheck All Unmapped" can be used to remove contigs or segments not found in the genome assembly.

  **Note:** If there was more than 5000 segments in the allele sequence only the 5000 longest segments would have been included in the assembly file.

- **Source: The name of the segment from the allele sequence file. To** rename a segment, either double click on the name in the **Segment** column, or if the segment names share the same pattern to be removed or for renaming, below the segment definition table are controls for renaming segments programmatically. The options include:

  - *RegEx*: Use regular expressions to rename the "Source".
  - *Substring*: Remove a substring from all segment names to generate the segment name.
  - *Prefix*: Remove a common prefix from all segment names.
  - *Suffix*: Remove a common suffix from all segment names.
  - *Manual*: Rename segments manually, one at a time. If manual mode is selected you can rename directly in the **Renamed** cell in the **Source to Segment Mapping** table.

  Enter in either the RegEx expression or the string to remove in the first text box. A preview of the renamed segment name will appear in the second text box. To apply the rename to all segments click on the **Set Segment to Renamed** button.

- **Length**: The length of each segment is displayed in this column.
- **Aliases**: If a segment has an alias listed in the genome assembly it will be listed in this field.

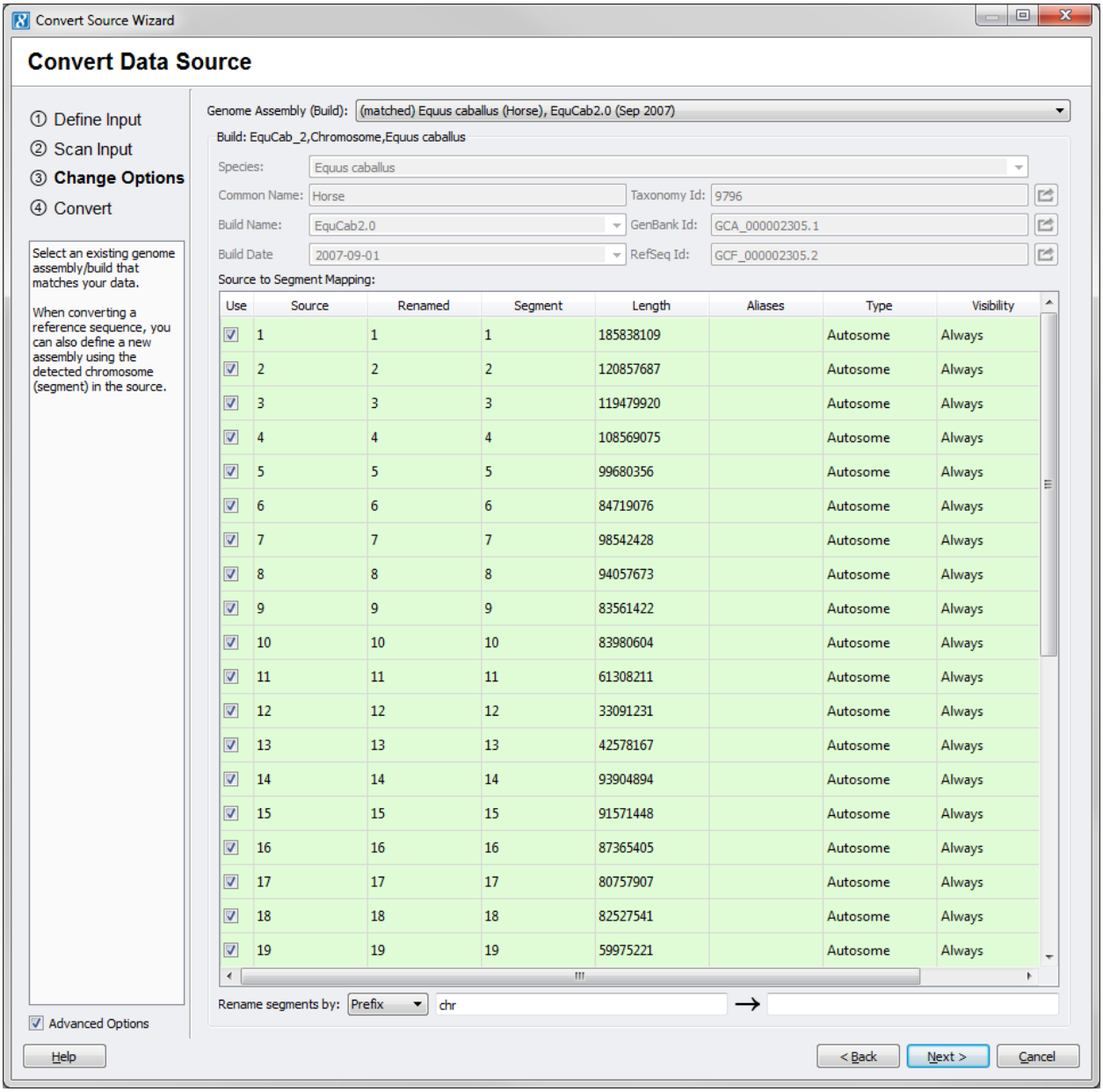



Figure 8.20: Select genome assembly for VCF files in convert wizard

- **Type: The type of segment. By default 'Autosomes' are always visible** and the rest are visible only if there is data. Options include:
  - Autosomes
  - Allosome
  - Mitochondrial
  - Fragment
  - Scaffold
  - Contig
  - Unknown
- **Visibility**: The visibility of the data. If the segment will only be shown in GenomeBrowse if there is data the visibility will be set to "With Data". Options include:
  - Always
  - Never
  - With Data

Once the assembly has been selected and the segments mapped to the assembly, click **Next >**.

After clicking **Next >** the wizard will display the documentation page. See *Documentation Step* for more information.

## 8.11 Documentation Step

Annotation sources must have a name specified. In addition to the source name, documentation on how the data was converted as well as the date and documentation for each field can be specified. All documentation will be embedded into the TSF file to make sharing files and documentation easy.

There are three sections in the documentation specification page of the convert source wizard:

- **Source Definition**: This information is used to identify the annotation source, and also indicate the date it was converted, who converted it and any version information.
  - **Name**: [*Required*] The name of the annotation source.
  - **Curated Date**: [*Required*] By default this is a date associated with the files being converted. It can be modified, but a date is required.
  - **Curated By**: Name or organization of who is curating the data.
  - **Series Name**: Name of a particular group of data. This field can be used to differentiate between newer versions of the same type of data. For example, RefSeqGenes-UCSC or dbSNP.
  - **Version**: A version number or date. It is recommended if there is a particular version name or identifier that this is included in the **Name** field and that this field be used for a date associated with the particular version.
- **Fields**: The individual field descriptions can be specified in this table.
  - **Orient**: The orientation of the data (locus or sample) cannot be modified.
  - **Type**: The type of the data (cannot be modified). If the type needs to be modified, click **< Back** to go back to the desired plot type and field specification page. Other changes will not be lost.
  - **Name**: The name of fields can be modified. However, if a field name is modified for a required field with an explicit name the source may not be able to be plotted as a specialized track type.

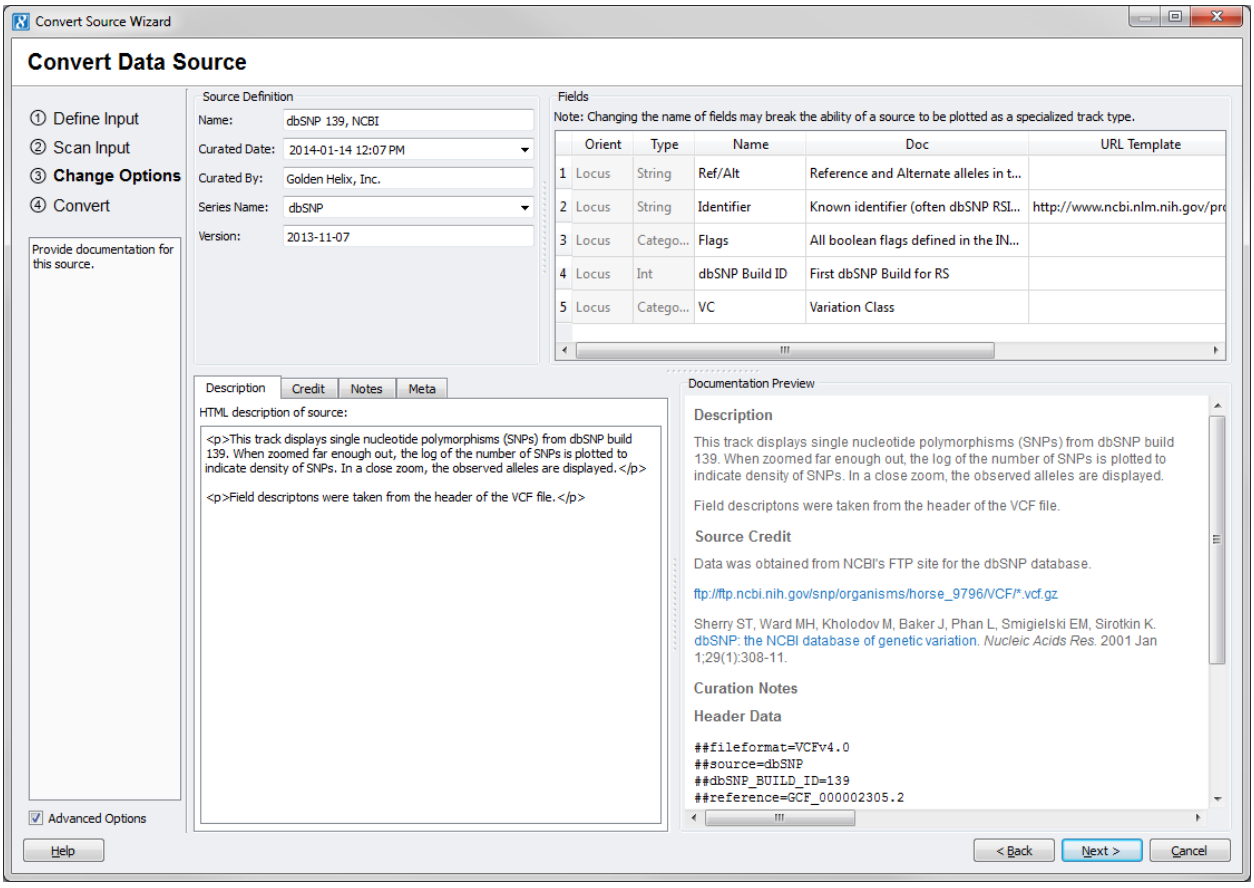



Figure 8.21: Document source for VCF files in convert wizard

  - **Doc**: The documentation string for the specific field.
  - **URL Template**: For fields with information that can be queried in an external site, specify the URL and two dollar signs ($$) to indicate where the text should be replaced. For example, for an "Identifier" field that contains RS ID's, the URL Template could be http://www.ncbi.nlm.nih.gov/snp/?term=$$.
  - **Categories**: Click on the **Edit** button to edit the category names and/or documentation. For some data sources including VCF files this option is immediately available as the data types and categories are specified in the header/meta information. For other data sources, after the source is converted, the documentation can be edited to rename and/or document categories for categorical field types. See *editAnnotationDocumentation* for more information.

- **HTML Documentation**: The four tabs at the bottom of the dialog are for writing HTML documentation of the source. The tabs are to guide the writing of the documentation and to provide nice headers for each section. HTML tags can be used for formatting.
  - **Description**: Description of the source and where it was obtained from.
  - **Credit**: Any required citations or credits for the source should go in this section.
  - **Notes**: Any relevant notes on pre-processing that had to be performed on the data or settings used to convert the source.

    **Note:** After the file has been converted statistics about the fields and data in the source will be placed in this section.

  - **Meta**: Any meta information for the data source(s).

    **Note:** If VCF files are converted the header information from the first VCF file will be placed in this section.

Once the documentation has been filled in as desired, click **Next >** to go to the confirmation and conversion step.

## 8.12 Confirmation of the Specified Parameters

- **Ready to convert message**: Contains information about the number of input files, the total size of the data to be converted (not the final size of the new file), the number of fields to be included in the converted source, the track type, the assembly and type of coverage to be computed. If any of the information in this message is not correct, back up through the wizard to adjust the parameters.
- **Field Indexing**: [*Advanced Option*] String fields can be indexed to enable searching or looking up information from the source in the location bar in GenomeBrowse. Fields should only be indexed if there is a reasonable expectation that the names in the field apply to just a few features. Such as RS IDs, gene names or transcript names. By default Identifier, Gene Name, and Transcript Name fields will be indexed. These fields can unchecked if it is not desired to have them indexed.
- **Left Align**: [*Advanced Option*] Using the reference and alternate fields, insertions and deletions are shifted to their left most possible representation. The genome assembly selected for the input files will be used to find the sequence surrounding each variant, so that it may be realigned. This option is only available for variant and variant map source types.
- **File Name**: [*Advanced Option*] The file name will be auto-generated based on the source name, version, species and build. If a different file name is desired, that can be changed in this advanced option.
- **Path**: [*Advanced Option*] By default the converted source will be saved in the user's default annotation folder. If a different location is desired, this can be specified by entering in a path or clicking on **Browse...**.

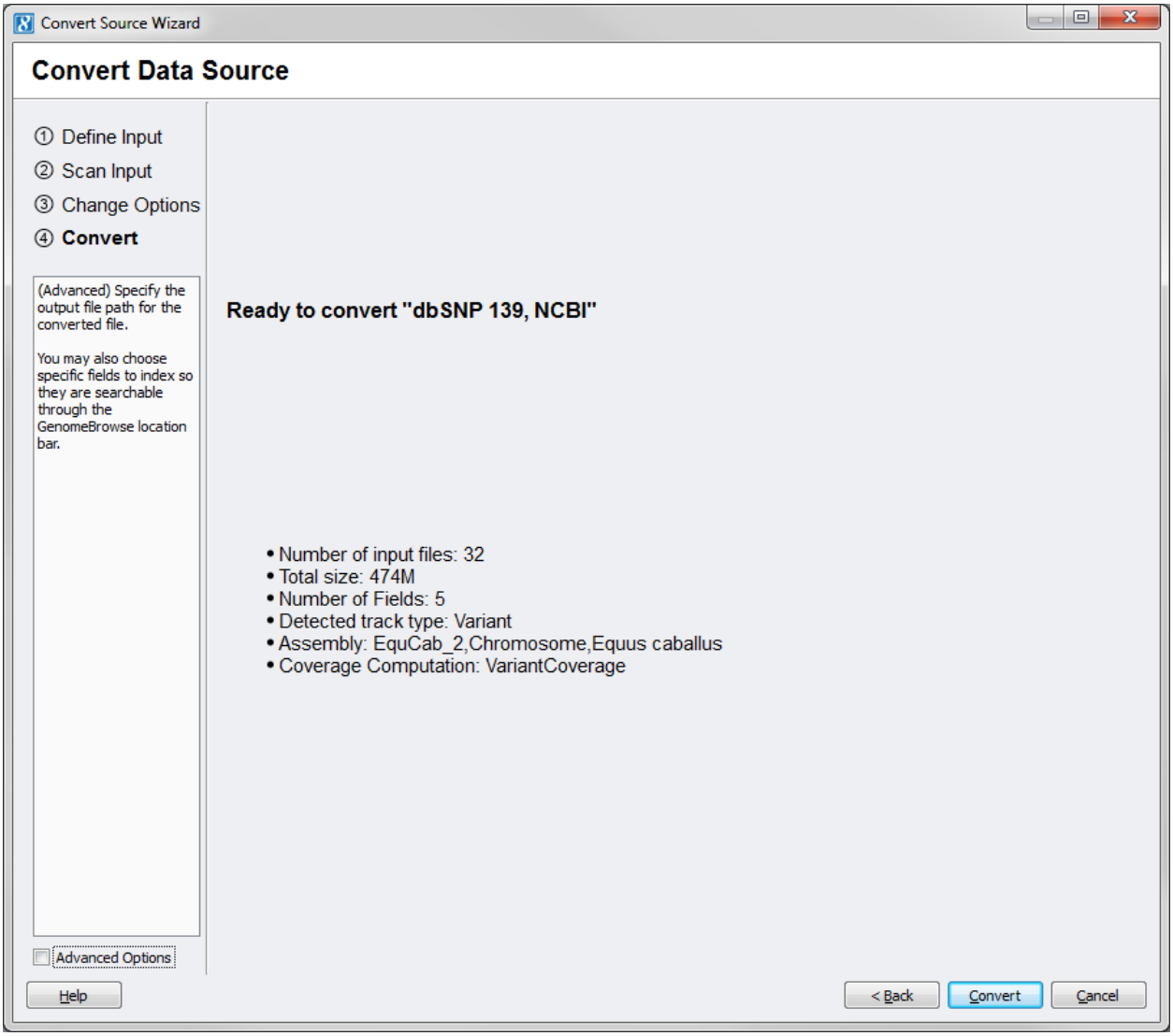



Figure 8.22: Confirm VCF sources in convert wizard

- **Add Path to Library**: [*Advanced Option*] If the path is changed, the path can be saved to the library so the source can be easily accessed in the Data Source Library.

If an option does not appear correct, click **< Back** to go back to fix the option. Any information specified will not be lost when backing up through the wizard unless changing an option changes how the data is read.

Once all of the options look correct click **Convert** to convert the data source(s) into the Annotation Source TSF format.

## 8.13 Converting Data Sources to Annotation Source

The convert source wizard will become a progress dialog during the last step, the converting step. An approximate time remaining will be displayed as well. Clicking **< Back** will stop the conversion but let it be restarted without specifying all of the options again. Clicking **Cancel** will exit out of the convert source wizard completely.

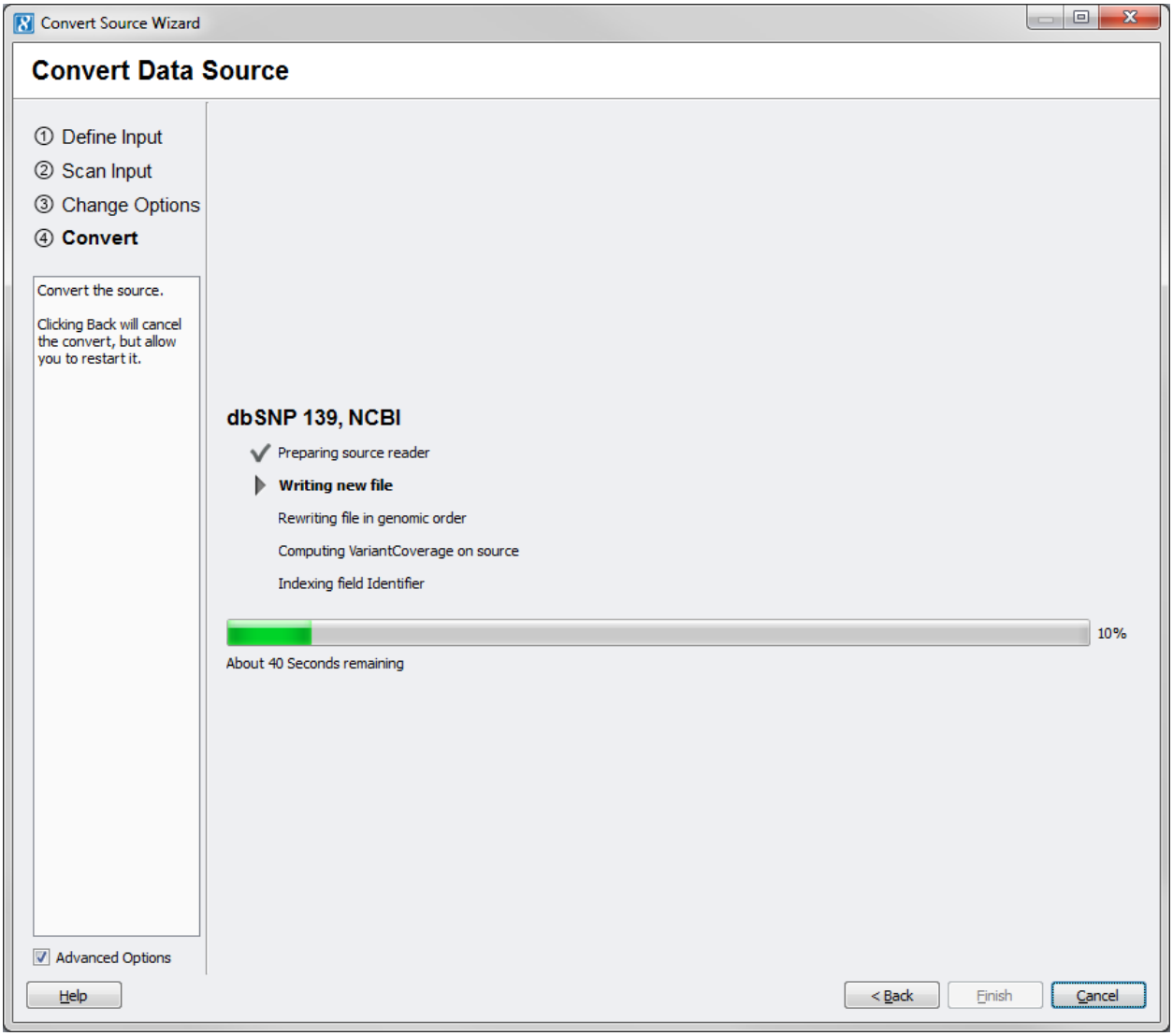



Figure 8.23: Converting progress for VCF files to dbSNP source in convert wizard

# CHAPTER NINE

# SAVING PLOTS FROM A GENOMEBROWSE WINDOW

Plots can be exported to several image formats.

When saving a plot, only non-hidden graphs in the GenomeBrowse window will appear in the output along with the optional currently visible Domain View plot.

## 9.1 Saving GenomeBrowse Plots to Image Formats

To save all visible plots to an image file, select **File > Save As Image**. This opens the **Save As Image** dialog which includes a preview of the image that will be saved as well as various options applicable to saving an image.

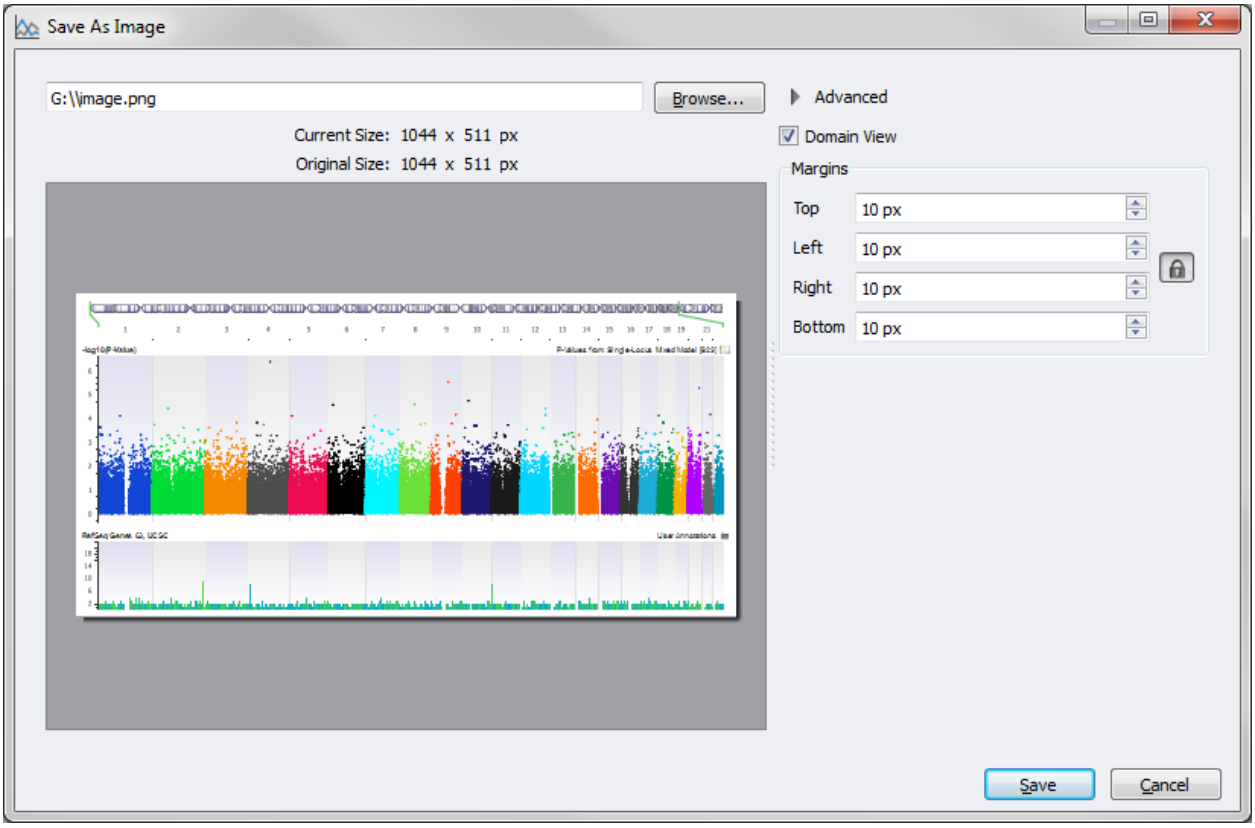



Figure 9.1: GenomeBrowse Save As Image dialog

**Note:** You can save a single plot or a set of selected plots to an image file by right-clicking on a plot and

selecting **Save As Image**.

- **Select Output File**: Select the output file by clicking **Browse...**. Available image types may include:
  - *PNG*: Portable Network Graphic
  - *BMP*: Windows Bitmap
  - *JPG*: Joint Photographic Experts Group
  - *PPM*: Portable Pixmap
  - *SVG*: Scalable Vector Graphic
  - *TIF*: Tagged Image File Format
  - *XBM*: X11 Bitmap
  - *XPM*: X11 Pixmap

- **Image Size**: The current size of the plot window in pixels is displayed. The size is dependent on the original size of the GenomeBrowse window. To change the image size adjust the original GenomeBrowse window before launching the *Save As Image* dialog.
- **Preview of Image**: A scaled view of the image to be saved is displayed in the dialog. Larger plots may take longer to draw, if the image has not finished drawing before *Save* is pressed an additional progress dialog will appear. Once the rendering of the image has finished the image will be saved.
- **Advanced Options**
  - *Domain View*: The domain view plot is shown at the top of the image. This view can be shown by checking this box.
  - *Margins*: The top, bottom, left and right margins (in pixels) can be altered independently. Click the *Unlock* button to the right of the current selections and adjust the appropriate control box.

To save the image click **Save**. To cancel and return to the GenomeBrowse window, click **Cancel**. If an output file has not been selected, a reminder to choose an output file before proceeding will appear.

# CHAPTER TEN

# IMPORT IGV SESSION

This feature allows data sources from Integrated Genomics Viewer (IGV) session files to be imported. These files are stored as XML (*.xml*), as outlined in the Broad Institute Specifications (http://www.broadinstitute.org/software/igv/Sessions.)

## 10.1 Importing an IGV Session File

- Follow: **File >> Import IGV Session**
- Clicking **Import IGV Session** will launch a file explorer dialogue. Use the file explorer to navigate to, and open, the IGV session file.
- After clicking open, a dialogue box will open providing a summary of information concerning any files that were not able to be imported.

# CHAPTER ELEVEN

# EVERNOTE: CLOUD-BASED PROJECT DOCUMENTATION

Evernote is an online note taking platform. Notes can be written in rich text and contain links, images, and other types of attachments. Your notes are available through any of Evernote's interfaces including the web.

In the context of GenomeBrowse, you can create a link to a genomic region, save an image of that region and note why that region was of interest. Later, you can share this note with others, who can explore the region on their own and add their own findings to the note.

## 11.1 Linking to an Evernote Account

Before notes can be created and saved, you must link GenomeBrowse with Evernote. If you do not have an Evernote account, visit https://evernote.com/sign-up/ and register for a free account.

To connect your account, select **File > Evernote > Connect to Evernote**. Enter the email address and password you used when you registered with Evernote. Next, confirm that you want to allow GenomeBrowse to have access to your notebooks.

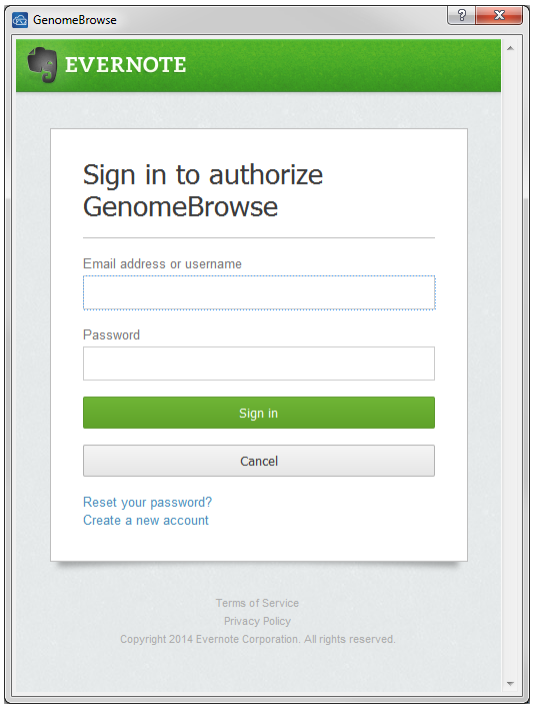



Figure 11.1: GenomeBrowse Connect to Evernote Dialog

## 11.2 Creating a New Note

To create a new note select **File > Evernote > New Note**. A dialog will appear which will allow you to select the notebook in which the note will be stored, enter the note's title, and add any tags that you would like to apply to the note. Multiple tags can be entered in the **Tags:** field using a comma to separate each tag.

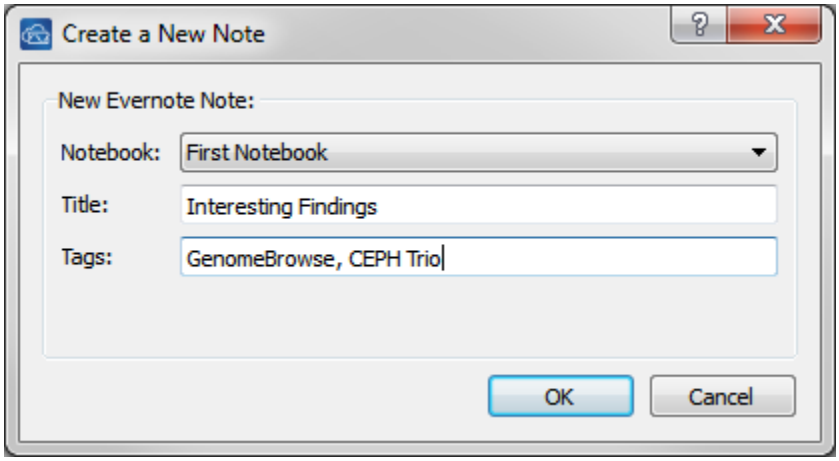



Figure 11.2: GenomeBrowse New Note Dialog

## 11.3 Opening an Existing Note

An existing note may be opened by selecting **File > Evernote > Open Note**. In the resulting dialog, select the desired notebook from the dropdown menu. A list of the twenty most recently edited notes in the selected notebook will appear in the list. Select the desired note and click **Ok** to open the note.

If the note you'd like to open does not appear in the initial note listing, you may search for it. By default, whole words will be matched against the title, contents or tags of notes in the selected notebook. A wild card can be used at the end of a search term to find all items starting with the term (e.g. gen* will match gene, genome, genetics, etc).

## 11.4 Editing a Note

After opening an existing note or creating a new note, a note editor will be shown. If an existing note is being edited, the editor will be pre-populated with that note's content.

Text can be entered and formated using the dialog's controls, just as you would when using a word processor.

The Evernote dock window contains the following items (from left to right):

- **Status message**: The note's status. Notes are automatically saved to your Evernote notebook whenever you stop typing.
- **Insert Bookmark**: Inserts a genomic coordinate bookmark. These bookmarks are links within GenomeBrowse that will jump to the region of the genome that was in view when the bookmark was set.

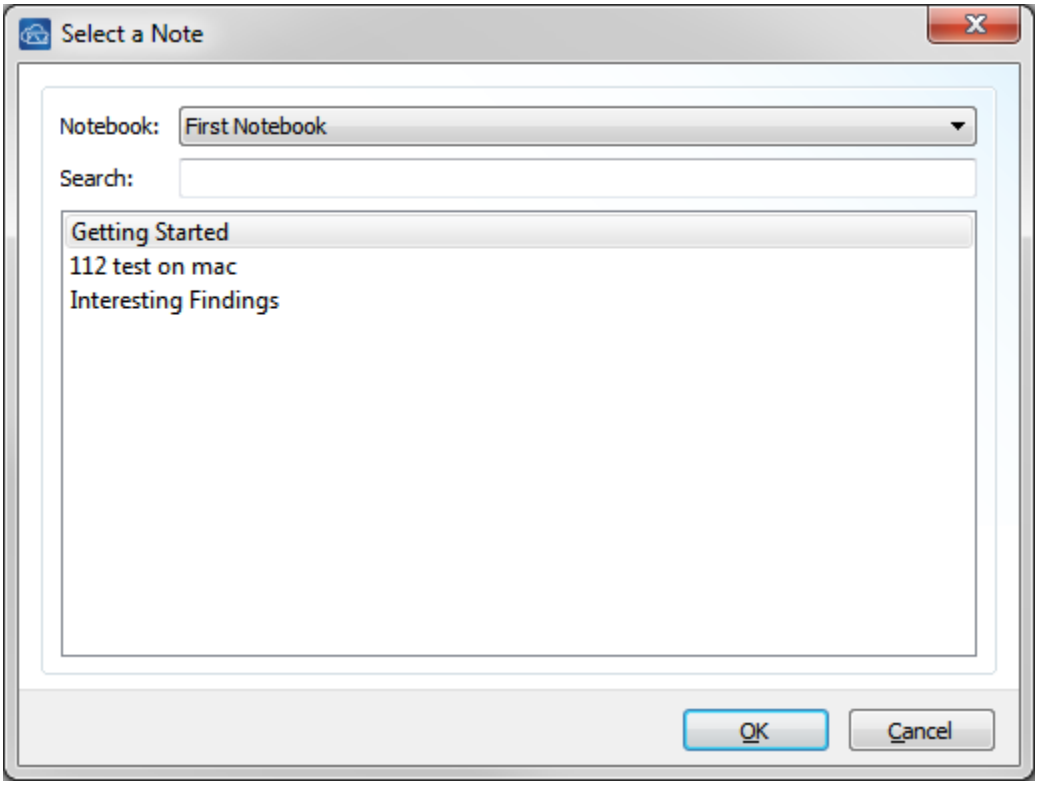



Figure 11.3: GenomeBrowse Open Note Dialog

- **Insert Screenshot**: Inserts a cleaned up screenshot of all active plots at the current zoom into the note. This "screenshot" displays the plots via the **Save as Image** controls. See *Saving Plots from a GenomeBrowse Window* for more information. The image is inserted at the cursor's current position or at the end of the note is no cursor is present.
- **Insert Screenshot of Selected**: Inserts a cleaned up screenshot of the selected plot(s) at the current zoom into the note. This "screenshot" displays the plots via the **Save as Image** controls. The image is inserted at the cursor's current position or at the end of the note is no cursor is present.
- **Share Note with Public URL**: Opens a dialog with a generated URL that can be copied and pasted to share a notebook as an HTML page. This facilitates sharing a note or results with someone who does not have Evernote access.

  **Note:** As the note is edited or updated the content in the shared HTML page will be updated as well.

The rest of the toolbar contains standard text editing controls.


Figure 11.4: GenomeBrowse Edit Note Dialog

# CHAPTER TWELVE

# GAUTIL DOCUMENTATION

This command line utility provides a collection of file processing tools. These tools allow for tasks such as converting between different file types, and the preprocessing of files prior to analysis.

## 12.1 gautil help

Provides a top level overview of the commands available and their specific usage.

## 12.2 gautil coverage

Create the .vcf.gz.covtsf file, which is used to draw zoomed out coverage

```
gautil coverage  <source>


Arguments:

   source    The source file name


Flags:

  --refFolder=<refrence_file> the folder with a reference sequence matching

                              the source
```

## 12.3 gautil fieldindex

Index the input file based on a selected field

```
gautil fieldindex <source>


Arguments:

   source    The source file name
```

```
Flags:

  --index=<index>  integer value of field to be indexed

  --path=<path>   path of the index file to be created
```

## 12.4 gautil index

Create the .vcf.gz and .vcf.gz.tbi version of the input file that is genomically indexed

```
gautil index <source>


Arguments:

   source    The source file name


Flags:

  --refFolder=<refrence_file> the folder with a reference sequence matching
                              the source
```

## 12.5 gautil precompute

Build an index/coverage for a BAM/IDF file

```
gautil precompute <source>


Arguments:

   source    The source file name
```

## 12.6 gautil s3url

Generate an Amazon s3url based on input parameters

```
gautil s3url <S3Bucket> <S3Key> <S3AccessKey> <S3SecretKey> <ExpiresInNSeconds>


  Arguments:

  S3Bucket           Container of addressed objects

  S3Key              Object identifier in the bucket

  S3AccessKey        Amazon account identification
```

```
  S3SecretKey        Amazon account password

  ExpiresInNSeconds  The number of seconds the link will be live
```

## 12.7 gautil schema

Ouput the schema in json for the source (if it's a directory, it will output a list of schema's)

```
gautil schema <source>
```

```
 Arguments:

   source    The source file name
```

## 12.8 gautil writefasta

Output the file into the .fasta format

```
gautil writefasta <source> <destination>
```

```
 Arguments:

   source         The source file

   destination    The destination file
```

## 12.9 gautil writewig

Output the file into .wig format

```
gautil writewig <source> <destination>
```

```
Arguments:

   source         The source file name

   destination    The destination file name
```

## 12.10 gautil writetsf

Output the input file to the .tsf file format

```
gautil writetsf <source> <destination> --query=<subset_query>
```

```
Arguments:

   source         The source file name

   destination    The destination file name


Optional Argument:

   query          Optional query such as chr1:851549-891052 that subsets

                  source to the query string. Requires that destination

                  is set. This is useful in creating subsets of files for

                  testing mostly.
```

## 12.11 gautil writevcf

Output the file into .vcf format

```
gautil writevcf <source> <destination>


Arguments:

   source         The source file name

   destination    The destination file name
```

## 12.12 gautil lock

Lock and unlock a file, multiple files may be locked using an asterisk '*'. If a flag is not included in the command it will default to the lock command.

```
gautil lock --flag <source>


Arguments:

  source    The source file name, the file name may be replaced

            with '*' to operate on all of the files in the directory

            with the given file extension.


Flags:

  --set     Lock the files for editing
```

```
  --unset   Unlock the files for editing
```

## 12.13 gautil leftalign

Left align the features based on the included reference sequence.

```
gautil leftalign <source> <destination>


Arguments:

   source         The source file name

   destination    The destination file name


Flags:

  --refFolder=<refrence_file> the folder with a reference sequence matching
                              the source
```

# CHAPTER THIRTEEN

# METHODS

## 13.1 Haplotype Frequency Estimation Methods

The alleles of multiple markers transmitted from one parent are called a haplotype. Haplotype analysis of safety and efficacy data can incorporate the information from multiple markers from the same gene or genes, which are physically close on a specific chromosome. Genotypic data from unrelated individuals do not contain information on which alleles were transmitted from each parent, but haplotype frequencies can be estimated using several existing *in silico* methodologies such as the Expectation Maximization (EM) algorithm and the Composite Haplotype Method (CHM).

**Note:** GenomeBrowse only computes LD using the CHM method.

### About Haplotype Inference

A common type of genetic data is genetic information independently scored at several markers along a chromosome. Each subject has two copies of the same chromosome, one chromosome from the mother, the other one from the father. It is not clear which alleles at different markers reside on the maternal, and which are on the paternal copy of the chromosome. Only individuals that are heterozygous at most at a single marker can be resolved into a pair of haplotypes unambiguously. This problem, called "genetic phase uncertainty" is an example of a more general problem of statistical inference in the presence of missing data.

The expectation-maximization (EM) algorithm, formalized by [Dempster1977], is a popular iterative technique for obtaining maximum likelihood estimates of sample haplotype frequencies (see [Excoffier1995] for details of obtaining haplotype frequencies by EM).

Determining the probability of each haplotype for each sample from the overall haplotype probabilities is based on the computation method being used, and does not relate to any estimate of LD between any pair of markers involved.

CHM diplotype probabilities are based on the average of the probabilities of the two haplotypes, while EM diplotype probabilities are based on the product of the probabilities of the two haplotypes. The process of finding the EM diplotype probabilities is equivalent to the "expectation" step of each EM iteration.

### Composite Haplotype Method (CHM)

The CHM is based on the idea of the genotypic LD coefficient, $\Delta_{AB}$, [Weir1996]. Estimation of $\Delta_{AB}$ involves calculation of di-genic frequencies. In the two-locus bi-allelic case, they are estimated as

$$\frac{1}{n}\eta_{AB} = \frac{2n_{AABB} + n_{AABb} + n_{AaBB} + n_{AaBb}/2}{n},$$

where $n_{AaBb}$, for example, is the number of individuals with genotype Aa/Bb, and n is the sample size. The composite disequilibrium is defined as a sum of inter- and intra-gametic components,

$$\Delta_{AB} = D_{AB} + D_{A/B} = P_{AB} + P_{A/B} - 2p_A p_B.$$

Under random mating, $P_{A/B} = p_A p_B$, and so assuming random mating, $\Delta_{AB}$ is an unbiased estimate of the LD parameter, $D_{AB}$. Also, if we do not wish to separate the inter- and intra-gametic components, we may define

$$\rho_{AB} = \frac{P_{AB} + P_{A/B}}{2},$$

which is an observable quantity.

[Zaykin2001] extended the definition of di-genic frequencies to multiple loci and alleles. For the i-th individual multilocus genotype $g_i$, let $H(g_i)$ be the number of single-locus heterozygotes in $g_i$. Define weights as

$$w(g_i) = \frac{1}{2^{H(g_i)-1}} = 2^{1-H(g_i)}.$$

Sample composite haplotype counts are calculated from summing over individual contributions,

$$\eta_{abc,\ldots} = \sum_{i=1}^{n} w(g_i) I(a, b, c, \ldots \subset g_i),$$

where n is the sample size, and $I(\cdot)$ is the indicator function, defined as

$$I(a, b, c, \ldots \subset g_i) = \begin{cases} 1 & \text{if i-th individual genotype } g_i \text{ has alleles } a, b, c, \ldots \\ 0 & \text{otherwise} \end{cases}$$

Thus, if the i-th individual has at least one copy of all required alleles, it is counted with weight $w(g_i)$. The composite haplotype frequencies are given by

$$\rho_{abc\ldots} = \frac{1}{2n} \eta_{abc\ldots}.$$

Note that $\rho_{abc\ldots}$ includes both inter- and intra-gametic component frequencies.

In a two-locus, two-allele case, composite haplotype counts simplify to Weir's definition, $\eta_{AB} = 2n_{AABB} + n_{AABb} + n_{AaBB} + n_{AaBb}/2$. In a single-locus case, they are the usual definition of the allele count:

$$n_i = 2n_{ii} + \sum_{i \neq j} n_{ij}.$$

## 13.2 Formulas for Computing Linkage Disequilibrium (LD)

The approach used for computing linkage disequilibrium (LD) in GenomeBrowse is the composite haplotype method (CHM).

## Computing LD using the Composite Haplotype Method (CHM)

### Multi-Allelic

If there are $k$ alleles in the first marker and $m$ alleles in the second, where either $k > 2$ or $m > 2$ or both, and using the same notation for $p_i$ and $q_j$ as above, a chi-squared distribution with $(k-1)(m-1)$ degrees of freedom may be written as

$$\chi^2 = n \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\Delta_{ij}}{2p_i q_j},$$

where

$$\Delta_{ij} = \frac{\eta_{ij}}{n} - 2p_i q_j,$$

and $\eta_{ij}$ is defined as in *ftCHM*. Here, we are effectively using

$$\rho_{ij} = \frac{1}{2n} \eta_{ij},$$

which includes both inter- and intra-gametic component frequencies, as our haplotype frequencies.

$R^2$ may then be computed by taking the p-value as

$$p = \text{chisqr}(\chi^2, (k-1)(m-1))$$

and obtaining $R^2$ from the inverse distribution for one degree of freedom as

$$R^2 = \frac{F^{-1}(p)}{n}.$$

### Bi-Allelic

For the two-locus two-allele case, and using the notation of *ftCHM*, we compute $R^2$ using the following direct formula

$$R^2 = \frac{\Delta_{AB}^2}{(p_A(1-p_A) + D_{AA})(q_B(1-q_B) + D_{BB})},$$

where $D_{AA}$ and $D_{BB}$ are the Hardy-Weinberg coefficients for allele $A$ of the first marker and allele $B$ of the second marker, respectively. This formula may be thought of as putting a "Hardy-Weinberg correction" onto the formula

$$R^2 = \frac{\Delta_{AB}^2}{p_A(1-p_A)q_B(1-q_B)},$$

which is only completely accurate under the special circumstance of random mating (perfect Hardy-Weinberg equilibrium over the two-marker haplotypes), for which $p_{A/B}$ approximates

$$p_A p_B$$

and $\Delta_{AB}$ is an unbiased estimate of $D_{AB}$.

It may be shown that for the circumstance of perfect linkage disequilibrium, the result of using the "Hardy-Weinberg correction" formula is equivalent to

$$R^2 = \frac{D_{AB}^2}{p_A(1-p_A)q_B(1-q_B)}.$$

## The D-Prime Statistic

If the minor allele frequencies of the respective markers are small, the magnitude of the $D_{ij}$ statistic cannot get very large, even if the marker is in almost complete linkage disequilibrium, compared to the magnitude it could have had if the allele frequencies of the markers were almost equal.

The D-prime statistic was designed to compensate for this. $D'_{ij}$ is defined as $D_{ij}$ normalized by the maximum possible value that $D_{ij}$ could possibly have given the allele frequencies in each of the markers.

Specifically,

$$D'_{ij} = \frac{D_{ij}}{\min(p_i q_j, (1-p_i)(1-q_j))}$$

if $D_{ij} < 0$, and

$$D'_{ij} = \frac{D_{ij}}{\min((1-p_i)q_j, p_i(1-q_j))}$$

otherwise.

The overall D-prime statistic is defined as

$$D' = \sum_{i=1}^{k}\sum_{j=1}^{m} p_i q_j |D'_{ij}|.$$

### Computing D-Prime

For multi-allelic CHM, we use

$$D'_{ij} = \frac{\Delta_{ij}}{\min(p_i q_j, (1-p_i)(1-q_j))}$$

if $\Delta_{ij} < 0$, and

$$D'_{ij} = \frac{\Delta_{ij}}{\min((1-p_i)q_j, p_i(1-q_j))}$$

otherwise, with the overall D-prime statistic being defined as

$$D' = \sum_{i=1}^{k} \sum_{j=1}^{m} p_i q_j |D'_{ij}|.$$

For bi-allelic CHM, we use the same formulas as for multi-allelic CHM, except that for the final $D'$, we take the original overall $D'$ obtained as above and use a Hardy-Weinberg correction on it:

$$D' = D'_{uncorrected} \sqrt{\frac{p_A(1-p_A)p_B(1-p_B)}{(p_A(1-p_A)+D_{AA})(p_B(1-p_B)+D_{BB})}},$$

where $A$, $B$, $D_{AA}$ and $D_{BB}$ are defined as in *Bi-Allelic*.

# CHAPTER FOURTEEN

# APPENDIX

## 14.1 Getting Started Guide

Golden Helix GenomeBrowse®

# Getting Started Guide

## How do I add a plot?

- You can drag and drop any BAM, BED, IDF, TSF or VCF file directly into the Plot View. Additionally, if a BED or VCF file is bgzipped you can drag and drop the BED.GZ or VCF.GZ file into the Plot View as well. GenomeBrowse also supports many other file formats.
- You can also stream BAM and TSF files by clicking on the Add button (  ) in the top-left corner of GenomeBrowse. From here you can stream TSF annotation tracks from the cloud, add BAM files from your EA Pipeline account, or add example data. Data from the cloud can also be downloaded to your local computer.
- You can also do the same as above by right-clicking in empty space in either the Plot Tree or the Plot View and selecting the Add button.
- You can change the default visible plots for Read Alignment (BAM) files by opening the **Options...** dialog in the **Tools** menu or by clicking on the Gear icon in the tool bar (  ) and checking or unchecking the **Coverage** or **Pile-up** options.
- You can change where new plots are added in the **Options** dialog. By default new plots are added to the **Top** of the plot view, this can be changed to have plots added to the **Bottom** of the view.

## How do I navigate?

- There are two navigation modes in GenomeBrowse:
  - In Navigation Pointer Mode (  ), you can pan any plot or axis by holding down the left mouse button. Dragging with the right mouse button scales the plot.
  - Zoom Mode (  ) will give you access to the rubber band zoom feature.
- You can use the scroll wheel to zoom in and out in any mode.
- By default, all plots have y-axis fit-data or auto zoom enabled. To change the y-axis zoom mode, hover over a plot and click the zoom mode button on the right edge (labeled with a letter).
- You can jump to a specific region or gene by typing it into the Zoom Toolbar. This includes typing in a gene or transcript name. If there are multiple options, the first 10 results from each searchable track will be displayed in a drop down list.
- There is a zoom slider for the x-axis in the toolbar between the – and + buttons. These controls scale the x-axis on all plots.
- There are left and right arrows framing the Domain Scale. These controls pan the x-axis on all plots.
- Keyboard shortcuts:
  - Arrow keys allow you to pan on the current plot.
  - + and - zoom in and out.
  - Shift + scroll wheel lets you zoom on just the y-axis.
  - Ctrl + Shift + scroll wheel lets you zoom on both axes.
  - Double-click on the y-axis to zoom to y-extents.
  - Double-click on the x-axis (or on a chromosome in the Domain View) to zoom to x-extents or the current chromosome.
- Right-click before you complete a zoom selection to cancel.

## How do I see my previous view?

- Clicking the Back and Forward arrows (  ) allows you to traverse zoom history.
- Right-clicking or holding the mouse button down on one of the buttons for a few seconds gives you a list of recent zoom history states, so that you can jump back by multiple steps.
- Keyboard shortcuts: Alt+left and Alt+right.

## How do I change the species and/or genome build?

- You can choose a new build from the drop down menu in the Genome Toolbar.

## How do I move and resize plots?

- You can drag plots around in the Plot View using their handles (at the left of a plot) or drag-and-drop nodes in the Plot Tree.
- To resize the plot, you can mouse over the handle (at the bottom of a plot) and drag up or down.

## How do I change how my data is displayed in a plot?

- In some plots, a gear icon will appear in the top-left corner when view options are available. Click on this gear icon to view plot-type specific options.
- If the control panel is already open, you just need to select a plot to have the control options update with those relevant to that plot type.
- To open the control panel, right-click on a plot and choose **Controls...**
- In pile-up plots, you can choose to **Emphasize Strand** or **Emphasize Mismatches** and **Stack Above Axis**, **Stack Split by Strand** or **Stack Paired-ends**. Multi-mapped alignments can be filtered by clicking on **Filter Multi-Mapped Alignments**. More controls are available by clicking on the **More Controls...** button. This brings up more display and filtering options.
- In a gene track, you can view regions in **Compact** or **Expanded** mode. **Auto** blends the two modes to adjust based on plot height.
- In variant maps, variant, interval and value plots, you can hide and show the **Feature labels** and change the label field.
- Value plots have the additional feature of being able to change the **Value** displayed based on the values available from the source.

### How do I edit plot labels?

- Double-click on any plot title or axis label to edit the text.
- Right-click on any plot title or axis label and select **Edit...**
- Right-click on a Plot Tree node and choose **Edit Title**.
- Double-click any Plot Tree node's text.

### How do I change the colors?

- Choose **Options...** in the **Tools** menu and click on the **Color** tab.
- **Reference Mismatch** defines the color of variant plots when zoomed out.

### How do I delete a plot?

- Select the plot and press the delete key.
- Right-click on the plot in either the Plot Tree or the Plot View and choose **Delete**.

### How do I change the Domain View (default is a cytoband)?

- Right-click on any plot in the Plot Tree or Plot View and choose **Set As Domain View Plot.**
- To change back to a cytoband, right-click on the Domain View and choose **Restore Cytoband Plot**.

### What can I do with the mouse anchor?

- Right-click anywhere in a plot and choose **Place Mouse Anchor** to mark a location.
- Keyboard shortcuts: Ctrl+` to set, Alt+` to remove, and ` to center the view on it.

### How do I save my project?

- By default, the project is saved on close. To disable this feature, open **Options...** dialog in the **Tools** menu and uncheck **Save projects on close** on the **General** tab.
- Similarly, the project settings are saved on close. To disable this feature, uncheck **Save program settings on close** in the **Options** dialog.
- Click **Save Program Settings Now** to save GenomeBrowse in its current state.

### How do I manage downloads from Public Data Sources?

- Once files are selected for download the **Download** window will open indicating progress of downloads. If GenomeBrowse is closed while files are being downloaded the Download manager will be minimized to the tray and downloads will continue.
- The **Download** manager can be accessed from the **Tools** menu.

### How do I view a list of features or find features in a sparse annotation track?

- If you hover over the upper-left corner of a plot view, a table icon will appear (  ). If you click on this icon the **Feature List** will open. By default up to 1000 features from the current zoom will be displayed.
- You can jump to the start of the genome by clicking on the **Start** button, and then back to the current zoom by clicking on the **Zoom** button.
- Features can be sorted by clicking on the field names (column headers). Once for ascending order, twice for descending order.
- Click on a feature to jump (zoom) to that feature in the plot view.
- Selected row(s) can be copied to the clipboard. Click on a row or use Ctrl+click or Ctrl+A to select multiple or all rows. Then right-click and select **Copy Selected Row(s) to Clipboard** or **Copy Selected Row(s) with Headers**. This table can then be pasted into a spreadsheet program or text file.
- A mouse anchor can be set at any of the features in the list by right clicking on a row and selecting **Set Mouse-Anchor At...**
- If there are more than 1000 features you can click on the **Read More...** link in the lower left corner to add more features to the list.

### How do I create a new project, save or open a recent project?

- In the **File** menu there are options to create a new project, open a recent project, open a project by browsing the file system, save or save as a new project. When you create a new project you will be prompted to select the project genome assembly and to download the reference sequence if you do not already have it locally.

### How do I know when there is an update available?

- There will be an **Update Available** message and a download icon (  ) in the lower-right corner of the GenomeBrowse window indicating that there is a newer version available. Click on the message to open the download page in a web browser to obtain the latest version. NOTE: This feature is only availalbe in version 1.06 and later.

**For more help, visit:**

goldenhelix.com/GenomeBrowse/online_help

Golden Helix GenomeBrowse®

# Reference Sheet

## Mouse

Double-click on a title, scale, or node label to edit it in-place

Use the scroll wheel to zoom in or out
- On canvas—affects x-axis only unless auto-zoom is on or a modifier key is used
- On scale—affects the scale axis only unless auto-zoom is on or a modifier key is used

Double-click on the domain view or an x-axis scale:
- Zoom to target chromosome
- If already zoomed to a chromosome, zooms out to full data extents

Double-click on a y-axis scale to zoom to full data extents

Double-click on a plot feature to zoom to the feature extents

**Pointer Navigation Mode**
Left-drag to pan canvas or scale
Right-drag to stretch canvas or scale

Click the other mouse button to cancel a drag-to-zoom operation in progress

**Zoom Box Mode**
Left-drag to zoom into selected area

Right-drag to zoom out, fitting the previous view in the selected area

Click the other mouse button to cancel a drag-to-zoom operation in progress

*Zoom modifiers remain restricted while y-axis auto-zoom is enabled. When both axes are enabled, dragging in a straight line affects only the axis parallel to the drag line.*

## Keys

| | |
|---|---|
| Alt+[Key] | Menu access |
| Ctrl+J | Show download window |
| Alt+Left | Previous view (back) |
| Alt+Right | Next view (forward) |
| F1 | GenomeBrowse Manual |
| F2 | Edit title of current node (Plot Tree) |
| Del | Delete selected nodes (Plot Tree or Plot View) |
| Ctrl+A | Select all (Plot Tree or Plot View) |
| Ctrl+C | Copy selected text (Console or Plot Tree or Editor) |
| Ctrl+Z | Undo (in any text box) |
| Ctrl+N | Create new project |
| Ctrl+O | Open project |
| Ctrl+S | Save project |
| Ctrl+Shift+S | Save project as... |
| Ctrl+Shift+T | Show/Hide Plot Tree |
| Ctrl+Shift+C | Show/Hide Controls |
| Ctrl+Shift+O | Show/Hide Console |
| Ctrl+Shift+F | Show/Hide Feature List |
| Ctrl+Shift+N | Show/Hide Note Editor |
| A or P | Switch to pointer navigation mode |
| Z | Switch to zoom box mode |
| Q | Set y-axis auto zoom mode on selected plots |
| F or W | Set y-axis fit-data zoom mode on selected plots |
| H or E | Set y-axis hold zoom mode on selected plots |
| M or R | Set y-axis manual zoom mode on selected plots |
| F5 | Reload selected plots |
| Arrow keys | Pan current plot |
| +/- | Zoom current plot |
| Back tick (`) | Jump to mouse anchor |
| Ctrl+' | Place/move mouse anchor |
| Alt+' | Clear mouse anchor |

## Modifiers

Click to select plot or Plot Tree node

| | |
|---|---|
| Ctrl+ | Multiple selection |
| Shift+ | Select range between first selected and target |
| Ctrl+Mouse Wheel | Scroll the plot view instead of zooming |

*Zoom modifiers remain restricted in some y-axis zoom modes*

Zoom with mouse wheel or keyboard (+/-)

| | |
|---|---|
| Shift+ | Apply to opposite axis (e.g. y instead of x over canvas) |
| Ctrl+Shift | Apply to both axes (over canvas only) |
| Right-hold | Apply to both axes (scroll wheel only over canvas) |

Double-click domain view or x-axis scale

| | |
|---|---|
| Shift+ | Zoom to full data extents regardless of current zoom |

## General Use

Left-drag to move a plot group by dragging its move handle (left side)

Left-drag to resize a plot by dragging its size handle (bottom edge)

Right-click on a specific target (such as the title) to get a list of relevant options

Left-drag on the funnel in the Domain Plot to pan the selection in the Plot View

Left-drag on the edge of the funnel to resize it

Left-drag outside of the current funnel to jump to new location

Drag the zoom slider or use adjacent +/- buttons to zoom in/out on the x-axis

To change the y-axis zoom mode, hover over a plot and click the zoom mode button on the right edge (labeled with a letter).

**Search Box/Location Bar**
Click once to select all text in the search box

Double-click to select a portion of the search box value

With a cursor in the search box, triple-click to select all text

Type to replace selection with new chromosomes, coordinates, or feature name
- Zoom to a specific range by entering two end points
- Jump to a small surrounding area by entering one position
- Jump to a gene by typing in the gene name

While typing, the search box will search all tracks that are set as "searchable"

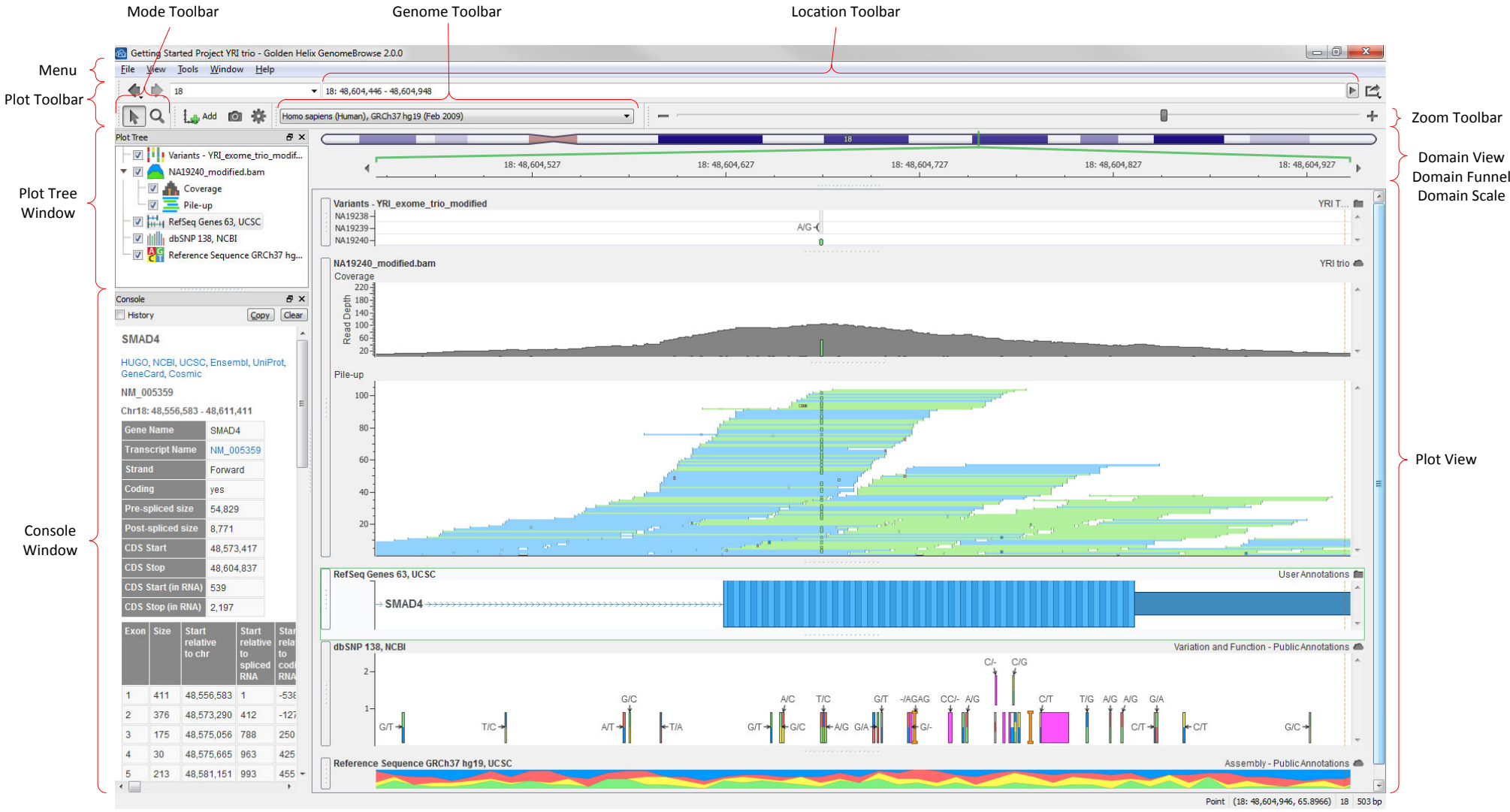
To set a track as searchable, right-click on it (Plot Tree or Plot VIew). Searchable data will show a "searchable" option. If checked, it is already searchable. If not, check it.

The link button to the right of the search box includes quick links to the current zoom for a few online genome browsers

The link button also includes a copy function for the current zoom in a format compatible with most genome browsers

*Note: Use the Cmd key on Macs instead of Ctrl*

Golden Helix GenomeBrowse™



## 14.2 Platform Notes

This section will contain notable information regarding the use of GenomeBrowse on specific platforms.

Under some platforms the behavior of GenomeBrowse could vary slightly in specific situations. Also, on some platforms certain system settings can be used to improve the performance of program.

### Microsoft Windows

#### Memory Usage

##### Allow up to 3GB of memory usage under 32~bit Windows

By default, 32~bit Windows versions allow applications to run in a 2GB memory space. Applications that attempt to use more than this 2GB limit will crash. When working with very large datasets, it may no longer be possible to fit the required data into the default memory space supplied by Windows. Using the /3GB switch in boot.ini allows certain applications to access up to 3GB of virtual address space leaving 1GB for the windows kernel.

To allow GenomeBrowse to use more than 2GB of memory, you can edit your system's boot.ini file.

To open the boot.ini file:

1. Click **Start**, click **Run**
2. Enter sysdm.cpl, and click **OK**

On the resulting window, select the **Advanced** tab, and click **Settings** under **Startup and Recovery**. Next, click the **Edit** button in the **System startup** group.

Now the boot.ini file should be open in a Notepad editor. Before you edit the file, it is recommended that you create a backup of the original. To do this: select **File > Save As...**, and choose a location to create the backup file. Close the Notepad editor, and click the **Edit** button again. Now you should be ready to edit the file.

To allow GenomeBrowse to use up to 3GB of memory add the entry **/3GB** to the end of the line under the [operating systems] section corresponding to your current configuration. Save the file to keep this option. A reboot is required before the change will take effect.

##### Memory Availability Under 64~bit Windows

Windows XP Professional x64 Edition (for PCs with x86~64 processors) is capable of running GenomeBrowse in its 32~bit emulation mode efficiently and allows the application to address up to 4 GB of virtual memory if available. Windows XP 64~bit Edition (built for Intel's IA~64 Itanium processors) is discontinued as an operating system as of 2005, and is not supported by Golden Helix, Inc.

# CHAPTER FIFTEEN

# EULA

GOLDEN HELIX SOFTWARE END USER LICENSE AGREEMENT

This End User License Agreement (the "Agreement") is a legal document. Please read it carefully before you install and use the Software.

To complete your downloading and or installation of the Software, you must accept the terms and conditions of this Agreement by electronically checking the box labeled "I Accept the License Agreement" that is displayed below and then clicking the "Next" button; but before doing so, we urge you to read this Agreement.

If for any reason you check the "I Accept the License Agreement" box and click the "Next" button without having read and understood this Agreement (which is a course of conduct which we do not recommend you follow), you should revisit and read this Agreement as soon as possible, and in any case you should do so before you commence using the Software.

Use of the Software is undertaken on the basis of these terms and conditions. If you choose not to read these terms and conditions, but use the Software, then you are deemed to have accepted these terms and conditions as though you had read them, understood them, and explicitly agreed to them.

This Agreement governs your use of the Software as defined hereunder and all related components thereof, and is a legally binding agreement which creates important legal obligations.

If you do not accept this Agreement, click the button labeled "Cancel" and do not complete the installation process. The entire software package should then be deleted if received electronically or, if delivered physically, returned to Golden Helix, Inc. at 203 Enterprise Blvd., Suite One, Bozeman, Montana, 59718, United States of America.

1. Definitions.

(a) "Software" means HelixTree®, SNP and Variation Suite ™, Golden Helix GenomeBrowse®, Optimus RP ™, ChemTree®, and/or other software that you are licensing from Golden Helix, Inc., and includes any additional packages, modules, upgrades, modified versions, updates, additions, and copies of such software, and any third party software Golden Helix, Inc. is licensed to include in the Software.

(b) "You" means the licensee. If the licensee is a company, then "you" includes those employees of the company who will be using or evaluating the Software.

3. "We," "us," and "our" means Golden Helix, Inc.

(d) "Documentation" means all of the explanatory written materials that accompany the Software.

2. Terms of License.

We hereby grant you a non-exclusive, non-transferable license to install and use the Software as described below in the section pertinent to the type of license granted to you, as specified on the applicable invoice. The appropriate number of machine-specific license keys will be provided to you, based on the license type or as stated on your invoice. These keys will expire upon termination of your right to use the Software. Unless you have purchased the Universal Server License Option, as described in 2(e), below, this is not a license to store or use the Software on a network server

computer. This Software is licensed as a single product and its component parts may not be separated for use on more than one computer.

(a) Limited Time Evaluation License. If this license is for evaluation purposes, the license will expire at the end of the evaluation period. The length of the evaluation period will be dependent upon your particular arrangement with us and will be confirmed in a separate correspondence between you and us. If other people within your company would like to evaluate the Software, they may request separate license keys for their machines at no additional charge. The scope of the evaluation license is limited to evaluation purposes internal to your company.

(b) Single Named User License. If you purchased a Single Named User License, then you may install the Software on one computer to be used exclusively by the individual specified on the applicable invoice. The Software cannot be shared with or used by any other individual.

(c) Site License. If this license is for a machine covered by a separate Site License Agreement with us, you may install additional copies of the Software up to the number of machines specified in that Site License Agreement. The Site License is for the term of one (1) year from the applicable invoice date unless stated otherwise in the Site License Agreement or on the applicable invoice. A machine-specific key is required for each computer that is added to the Site License.

(d) Lab License. If you purchased a Lab License, then you may install the Software on one computer for use by any individual in your organization. The software is not to be used by people outside of your organization. The Lab License does not allow you to put the Software on a server, nor any other device that enables remote access.

(e) Universal Server License. If you have purchased the Universal Server license, then you may install the Software on one server that allows any individual in your organization to access the Software from within your organization. The Software must be installed in such a way that it is inaccessible by anyone outside of your organization.

(f) Concurrent User License. If you have purchased one or more Concurrent User Licenses, the software can be installed on as many computers in your organization as desired, for use by any individual in that organization. However, use of the license is strictly limited to individuals employed by your organization. Access to the software will be limited to the number of concurrent users for which licenses have been purchased. For example, if you purchase one Concurrent License, access to the software will be granted to only one person at a time. If you buy two Concurrent User Licenses, two individuals will be able to access the software at the same time, and so on. The user's computer must be connected to the internet and have access to our website for the software to work. A machine-specific key is required for each computer that is to have access to the Concurrent User License(s).

(g) Annual License. If the license is not for a limited time evaluation or a monthly subscription license, then the license is for a term of one (1) year from the applicable invoice date, unless stated otherwise on your invoice, and you may use the Software for any internal purpose. This license does not include the right to market, sell, distribute, or sublicense the Software to any third parties.

(h) Renewal. Unless special terms have been detailed on your invoice, pricing for renewing your license will be based on the renewal prices and policies in effect as of the date you actually renew. The start date of your renewal license will be the day following the expiration of your prior license, and not the date you actually renew. If you do not renew your license within 30 days of its expiration, you will no longer be eligible for renewal pricing and will have to buy a license based on the then-current pricing policies for new licenses.

(i) Monthly Subscription License. If you have purchased a monthly subscription license, then the license will be active for a term of one (1) calendar month, unless stated otherwise on your invoice, and the license will remain active for additional one (1) month increments under the conditions outlined in the SNP & Variation Suite Software Monthly Subscription Agreement signed by you. You may use the Software for any internal purpose. This license does not include the right to market, sell, distribute, or sublicense the Software to any third parties.

(j) GenomeBrowse Standalone License. If you have installed and logged into the GenomeBrowse standalone product, your license is available for authorized use under this agreement while using the current version of the software. Updates to the Software may include amendments to this Agreement which you will be required to agree to prior to continued use of the Software. In order to download and install GenomeBrowse, you will be required to create an account that will be used to login to GenomeBrowse. Should you wish to install GenomeBrowse on additional computers, you may do so. Should other individuals wish to use GenomeBrowse, they may register for an account to

login and/or download the software. This license does not include the right to market, sell, distribute or sublicense the Software to any third parties, or to repackage the Software for any purpose.

If you are an individual from a commercial company, you are free to download, install, and use GenomeBrowse inside your organization without any cost or restrictions except the general use restrictions outlined in this Agreement. If you are interested in adding GenomeBrowse as a "value add-on" to your product, including but not limited to links within your product to GenomeBrowse and/or full instances of the software, and/or if you or your organization intend to use GenomeBrowse for any monetary or other business benefit, you must contact Golden Helix at info@goldenhelix.com or 406-585-8137 for a separate licensing agreement.

3. Use of Software.

You may use this Software for any internal purpose permissible by law. You may publish, reproduce, and distribute Software screen displays, or any derivative thereof, in any media.

4. Payment.

If the license granted pursuant to the preceding section is for evaluation purposes only, then it shall be free. If the license granted pursuant to the preceding paragraph is not for evaluation purposes, then you shall pay us or, if you are in a market served by a Golden Helix distributor, you shall pay our distributor, in accordance with the quotation or invoice previously provided to you and incorporated herein by reference.

5. Title.

This Agreement shall not constitute a sale of the Software or any copy thereof, nor of the magnetic or other physical media upon which the Software and Documentation are recorded or fixed. We will remain at all times the owner of the Software and Documentation on the original media and subsequent copies thereof regardless of the form in which or medium upon which such subsequent copies may exist. Any product(s) or chemical compound(s) developed through the use of this Software, except those that infringe on the copyrights and patent rights of Golden Helix, Inc. or third party licensors to Golden Helix, Inc., remain your product(s).

6. Things You May Not Do.

By accepting this Agreement, you agree not to, nor to allow anyone else to, do any of the following:

1. Distribute the Documentation outside your company.

2. Copy the Software, except for one copy for back-up purposes.

(c) Modify or adapt the Software or merge it into another program without our written permission.

(d) Reverse engineer, disassemble, decompile or make any attempt to discover the source code of the Software.

(e) Place the Software onto a server so that it is accessible via a public network, except as provided for in Paragraph 2(e) herein.

(f) Sublicense, rent, lease or lend any portion of the Software or Documentation.

7. Modify the Documentation without our written permission.

(h) Circumvent the license manager to use the software outside of the time period of your license code.

(i) Export or re-export the Software in violation of any export provisions of the United States or any other applicable jurisdiction.

(j) Use a screen display of the Software, or any derivative thereof, to register or claim any copyright or trademark rights.

7. Disclaimer of Warranties; Limitation of Remedies.

The software and documentation are being provided to you "as is," and you agree to assume the entire risk as to the quality and performance of the licensed software. We hereby disclaim any and all warranties, whether expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. In no event shall we be liable for any damages whatsoever (including, but not limited to, damages for loss of use of the

software, loss of profits, loss of savings, business interruption, or loss of data or other business information, or other incidental or consequential damages) arising out of the use of or inability to use the software, even if we have been advised of the possibility of such damages.

8. Term and Termination of Agreement.

This Agreement takes effect upon the applicable invoice date and remains effective as long as your license to use the Software has not expired. If your license is for evaluation purposes, as referenced in section 2(a) herein, then this Agreement takes effect on the date on which you install the Software and remains effective as long as your license to use the Software has not expired. Unless your license is for evaluation purposes, you may renew your license for successive terms after expiration of the initial license period under the terms described in Paragraph 2(f). If you elect not to renew your license, all physical copies of the licensed Software and the Documentation must be either destroyed or returned to us within thirty (30) days after the expiration of this Agreement, and all Software copies in Customer's computer(s) must be erased.

WE RESERVE THE RIGHT TO TERMINATE THIS AGREEMENT, WITHOUT REFUND, IN THE EVENT THAT YOU MATERIALLY BREACH THIS AGREEMENT AND HAVE NOT CURED SUCH BREACH WITHIN THIRTY (30) DAYS AFTER RECEIVING WRITTEN NOTICE OF SUCH BREACH FROM US.

NOTWITHSTANDING THE ABOVE, IF YOUR LICENSE HAS BEEN USED IN A WAY THAT VIOLATES THE USAGE TERMS SPECIFIC TO YOUR LICENSE TYPE AS DESCRIBED IN SECTION 2, WE WILL, AT OUR SOLE DISCRETION, EITHER A) IMMEDIATELY INVOICE YOU FOR THE FEES REQUIRED TO UPGRADE YOUR LICENSE TO THE TYPE APPROPRIATE BASED ON YOUR USAGE, OR B) IMMEDIATELY TERMINATE YOUR LICENSE WITH NO REFUND GIVEN FOR ANY TIME LEFT ON YOUR LICENSE.

9. Confidentiality.

The Software is being made available to you in strict confidence. You agree to maintain the confidentiality of the Software and Documentation and any and all trade secrets or other proprietary or confidential information contained in the Software and Documentation (collectively, the "Confidential Information") to the degree exercised by you with respect to your own proprietary and confidential materials or to a reasonable degree, whichever is greater. You further agree not to disclose any Confidential Information to any third parties without our written consent and to inform any of your agents and employees who will be using or evaluating the Software of their obligations to maintain the confidentiality of the Confidential Information. Notwithstanding the foregoing, your obligation of confidentiality hereunder shall not apply to any information:

(a) which, at the time of disclosure, is publicly available or in the public knowledge;

(b) which, after disclosure, lawfully becomes part of the public knowledge through publication or otherwise, but through no fault of yours;

(c) which you possess at the time of the disclosure of such information by us to you and which was not acquired, directly or indirectly, from us; or

(d) acquired by you from a third party who has a right to disclose such information.

The obligations set forth in this section shall survive the termination of this Agreement.

10. Indemnification.

You hereby agree to indemnify, defend and hold us, our officers, directors, employees and agents harmless from and against any and all claims, actions, causes of action, demands or expenses (including, but not limited to, attorneys' fees) relating to or arising from your use of the Software; provided, however, that this indemnification provision shall not apply to claims that the Software or Documentation, as provided by us to you, violates the intellectual property rights of a third party. The obligations set forth in this section shall survive the termination of this Agreement.

11. Notice of Patent and Copyright.

This Software and Documentation are protected by United States patent and copyright laws and international treaties. Copies are to be made only in accordance with Section 6(b) hereof.

12. Trademarks and Proprietary Names.

"HelixTree," "SNP & Variation Suite," "SVS," "Golden Helix Genome Browse," "ChemTree," "Optimus RP," "Accelerating the Quest for Significance," "CNAM," "Copy Number Analysis Module," and "The power of personalized medicine" are trademarks of Golden Helix, Inc. Any other product names mentioned in Documentation may be trademarks or proprietary names of other corporations and are used in Documentation for identification purposes only.

13. Miscellaneous Provisions.

(a) This Agreement supersedes any and all prior negotiations, understandings, proposals or verbal agreements and any other communications between us relating to the subject matter of this Agreement.

(b) This license agreement may be modified only by a written agreement signed by you and us.

(c) This Agreement is governed by the laws of the state of Montana, United States of America.

(d) You agree to submit to the personal jurisdiction of a state or federal court in Montana in the event that you breach this Agreement.

(e) This Agreement may not be assigned without our prior written consent, which shall not be unreasonably withheld.

(f) In the event that either party materially breaches this Agreement, the non-breaching party shall be entitled to recover from the breaching party its reasonable attorneys' fees incurred in pursuing a claim or claims against the breaching party, regardless of whether a lawsuit is actually filed.

BY CLICKING THE "I ACCEPT THE LICENSE AGREEMENT" BOX BELOW AND THEN CLICKING THE "NEXT" BUTTON, YOU ACKNOWLEDGE THAT YOU HAVE READ AND UNDERSTAND THIS AGREEMENT, AND YOU AGREE TO BE BOUND BY ITS TERMS.

# CHAPTER SIXTEEN

# REFERENCES

# BIBLIOGRAPHY

[Dempster1977] Dempster, A. P., Laird, N. M., Rubin D., (1977), 'Maximum likelihood from incomplete data via the EM algorithm.' J of the Royal Stat Soc B 39: 1-38.

[Excoffier1995] Excoffier L, Slatkin M (1995) 'Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.' Molecular Biology and Evolution 12: 921–927.

[Nicol2009] Nicol JW, Helt GA, Blanchard SG Jr, Raja A, Loraine AE (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics. 25(20):2730-1. PubMed PMID: 19654113; PubMed Central PMCID: PMC2759552

[Nielsen1998] Nielsen D, Ehm M, Weir BS (1998) 'Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus.' Am J Hum Genet 63: 1531–1540.

[Weir1996] Weir BS (1996) 'Genetic Data Analysis II.' Sinauer Associates.

[Zaykin2000] Zaykin DV, Nielsen DM (2000) 'Hardy-Weinberg disequilibrium (HWD) fine mapping for case-control samples.' Am J Hum Genet 67: 1238(S).

[Zaykin2001] Zaykin DV, Ehm, MG, Weir BS (2001) 'Evaluating new haplotyping methods for predicting clinical response using dense maps of single nucleotide polymorphisms (SNPs).' Work in progress. Presented at Bioinformatics Seminar Series, Research Triangle Institute, NC.