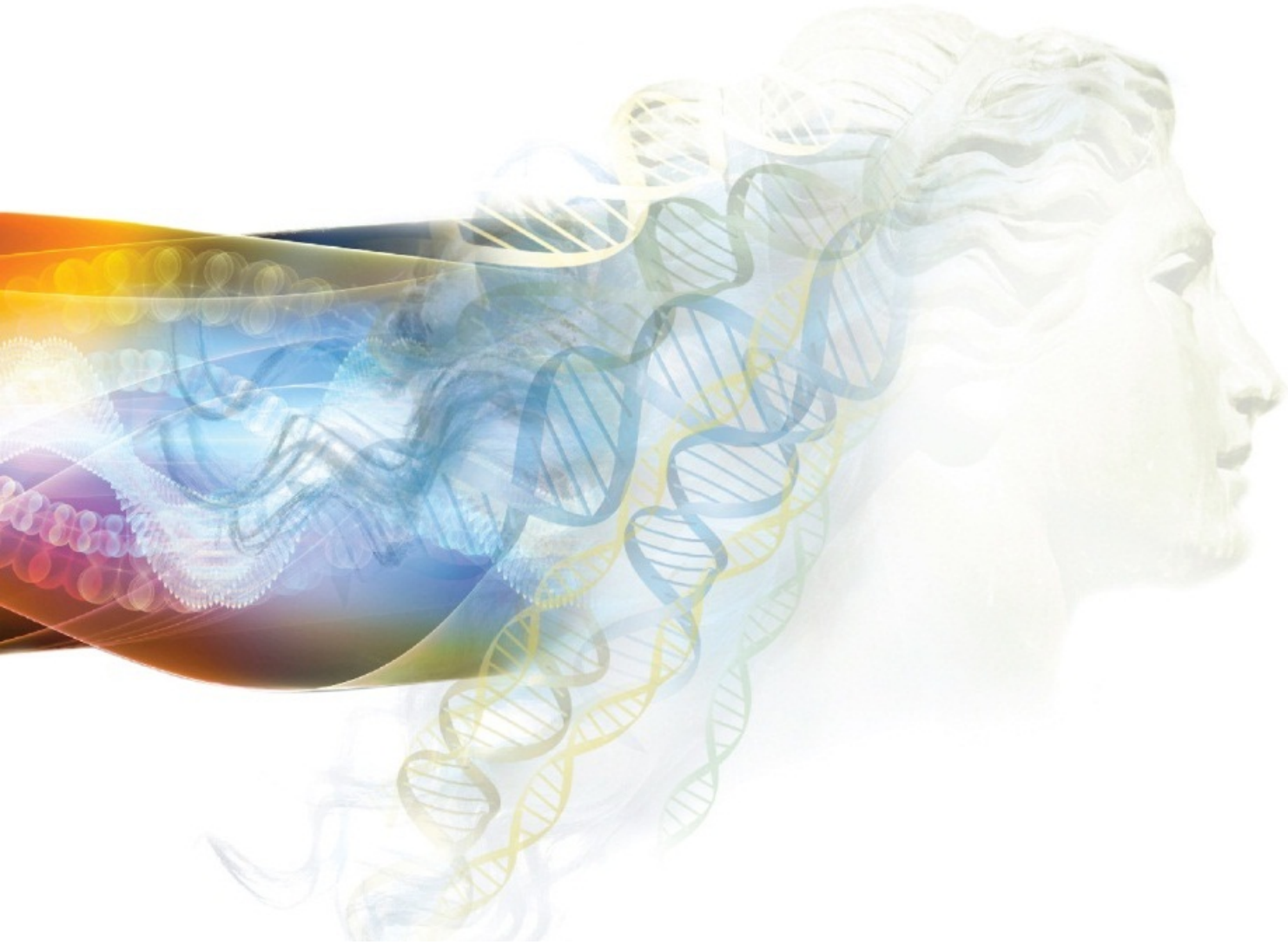


Persephone

GENOME VISUALIZATION AND EXPLORATION

Capabilities

ceres

Persephone
GENOME VISUALIZATION AND EXPLORATION

Table of Contents

Persephone™
Genome Visualization and Exploration Software

Developer

Ceres, Inc. is an agricultural biotechnology company that has spent over a decade developing, collecting, collating, and curating a wide range of genomics data from both internal and public datasets. Our vision of easy accessibility, long-term consistency and integration of experimental data has been realized in the Persephone software, we have developed and deployed for both in-house and external use.

Persephone™ is a next-generation genomics visualization platform.
Just as Google Maps offers easy access to large amounts of geographic information, Persephone provides this same type of functionality for genomic information. Persephone makes visualizing and exploring through diverse genomics data types easy and enjoyable with *gaming-type performance*. In addition, the platform architecture has been optimized for high-volume data visualization (e.g., you can typically visualize 30 million SNPs in under 30 seconds).

Persephone enables users and institutions to manage ever-increasing genomic datasets. The application can integrate new public and private data onto already established genomic information. This ability to stack multiple datasets on top of each other provides consistent, long-term accessibility and transparency for experiments and data.

Key Features and Benefits

Speed enables new insights

- Persephone's data visualization engine is optimized for displaying high volumes of information and enabling **real-time investigation** of large data sets.
- Instant access to data promotes a **deeper exploration** of the data and enables researchers to unearth valuable insights that would otherwise remain hidden due to the size and complexities of the data.
- **Fast and intuitive navigation** can be performed across, as well as between, multiple layers of information, such as
    - whole genomes,
    - their relationship to other genomes,
    - individual genes,
    - their activity (expression profile), and
    - their mutations.

Integration of multiple data sources

- Persephone enables transformative research opportunities by providing **integrated dynamic views** of information from multiple public and proprietary data sources and types.
- Data architecture is optimized for data integration and rapid visualization.
- Persephone manages **diverse data types.** You can utilize established analysis pipelines with minor data loading modification. Underlying primary data is consistent among technologies (e.g., nucleotides, SNP, coverage).
- **Concurrent visualization** of secondary and tertiary data (e.g., marker-trait associations, expression, protein translations) provides a deeper understanding of the effects caused by primary data (sample) differences.

Context-aware data

- Data is organized by type rather than by file to ensure **scalability and long-term data consistency**.
- Organization by data type allows Persephone to provide **context-aware functionality** (containing information about what is being visualized) vs. rendered representations (only drawing instructions).
- Context-aware data enables rendering of the same data in alternative ways depending on the data context the user chooses.

Extensibility

- Data in Persephone can contain links to external resources, such as web, alternate databases and files.
- External applications can link into specific objects in Persephone via URLs.
- Persephone's database can be accessed and updated directly by external applications.
- **Database engine** provides powerful facilities for performance optimization, access control and replication across data centers.

Performance Statistics

Below are examples from normal use of the software.
These are not performance limits.

- Application has been optimized to remain under 1 GB of RAM in most circumstances (e.g., visualizing 500 genotypes).
- Search through 100,000+ genotypes in real-time (e.g., by name or trait).
- Display 100,000+ elements on screen simultaneously (e.g., genes and markers).
- Load and visualize 30 million SNPs in 30 seconds.
- Search through 100,000+ maps (physical, genetic, scaffold) in the database in real-time.
- Zoom in and out from 300+ million to a single nucleotide in real-time.

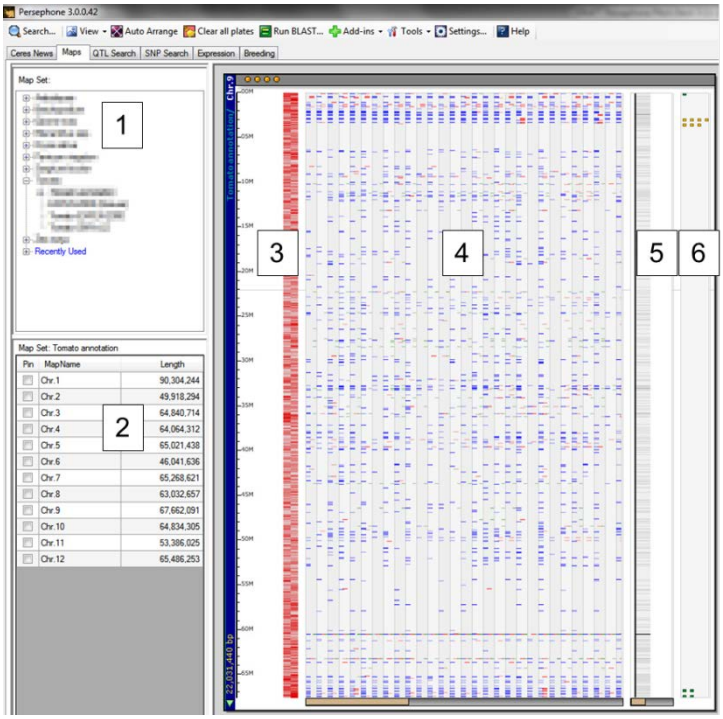
Sample Screenshots and Uses

Stacking Information

The layout of the software is separated by stages and each stage can display a variety of different data types.

Any number of species can be easily navigated through and each species can hold any type of map set, including genetic maps, consensus maps, physical maps, and others (*see 1*).

Once a map set is selected for a given species, all chromosomes, linkage groups, or scaffolds are displayed for the user to select *(see 2)*.

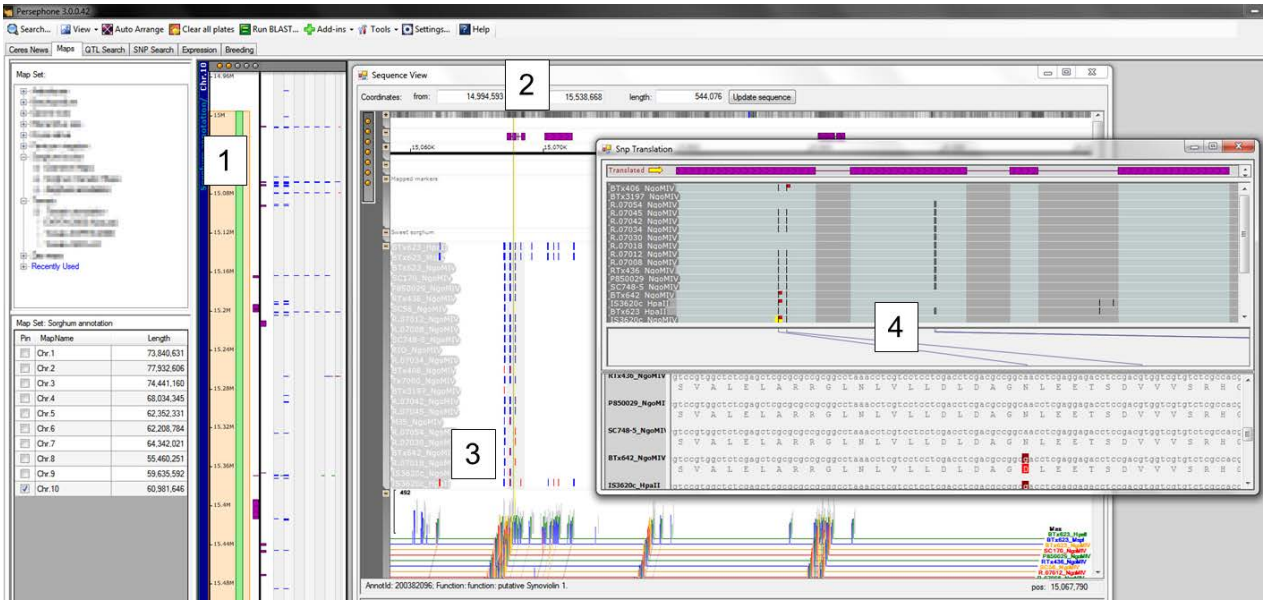


In this example, a user has selected tomato chromosome 9 (entire chromosome), and all annotated genes/features (red lines) are now visible *(see 3)* Further zooming to the nucleotide level is possible.

The user then selects a set of genotypes that have previously run through a Genotype-by-Sequencing experiment that identified thousands of segregating SNP markers on each genotype. The user sees these SNP markers mapped to the selected reference chromosome and color coded based on user-selected context and allele (*see 4*).
The user also decides to add public and internal markers, so the new alleles can be compared with what is already available (*see 5).*

And finally, the user wants to see all regions/genes that have been associated (correlated) with any traits so to better refine and target a region to investigate further (*see 6).*

Narrowing Down Targets

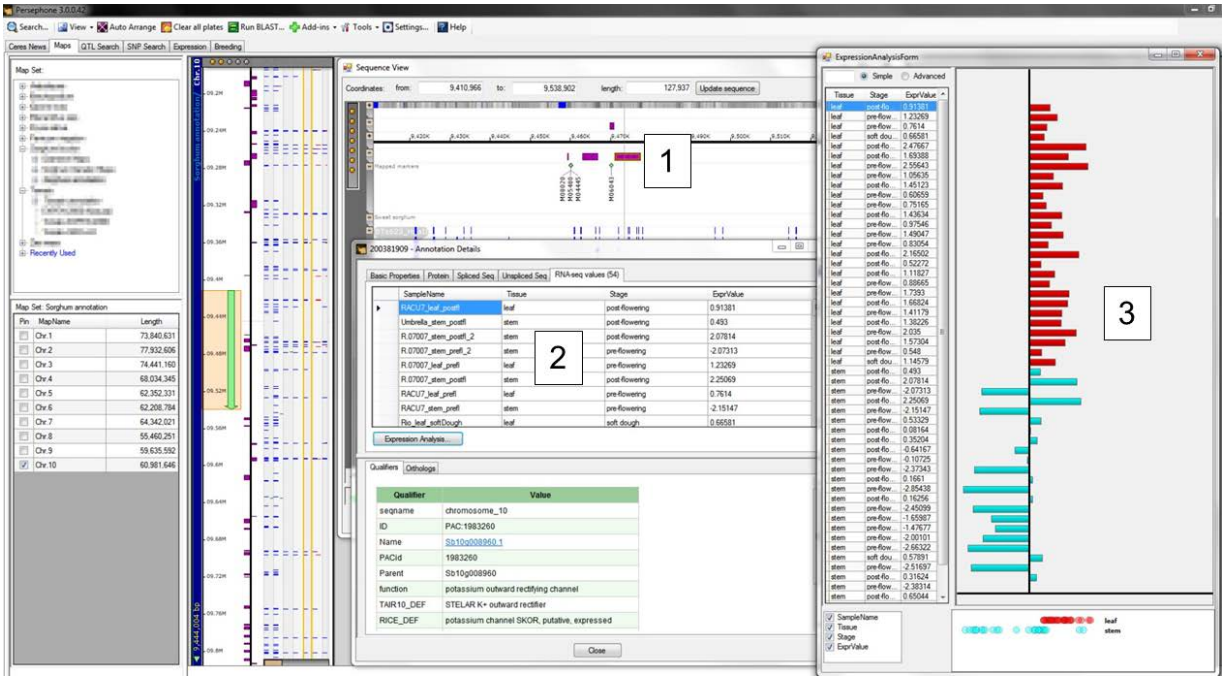


In this example, a user has selected a targeted region for further investigation (green bar above) (*see 1*) and now wishes to visualize this region in a more granular way horizontally (*see 2*).

To do this, the user selects this region, and the horizontal view appears (*see 3*) and allows the user to zoom down to the nucleotide level. The settings and data-types selected in the vertical view have all been maintained, including the selected genotypes, alleles, and now, coverage data.

The user can now further re-configure the displayed datasets. In this view (*see 4*), it is easy to explore genes, find mutations (SNP/indels), and perform real-time gene translations to identify alleles and genotypes in the selected region which contain mutations that affects the translated amino acid sequence. The user can easily design new SNP assays with a click of the right-mouse button.

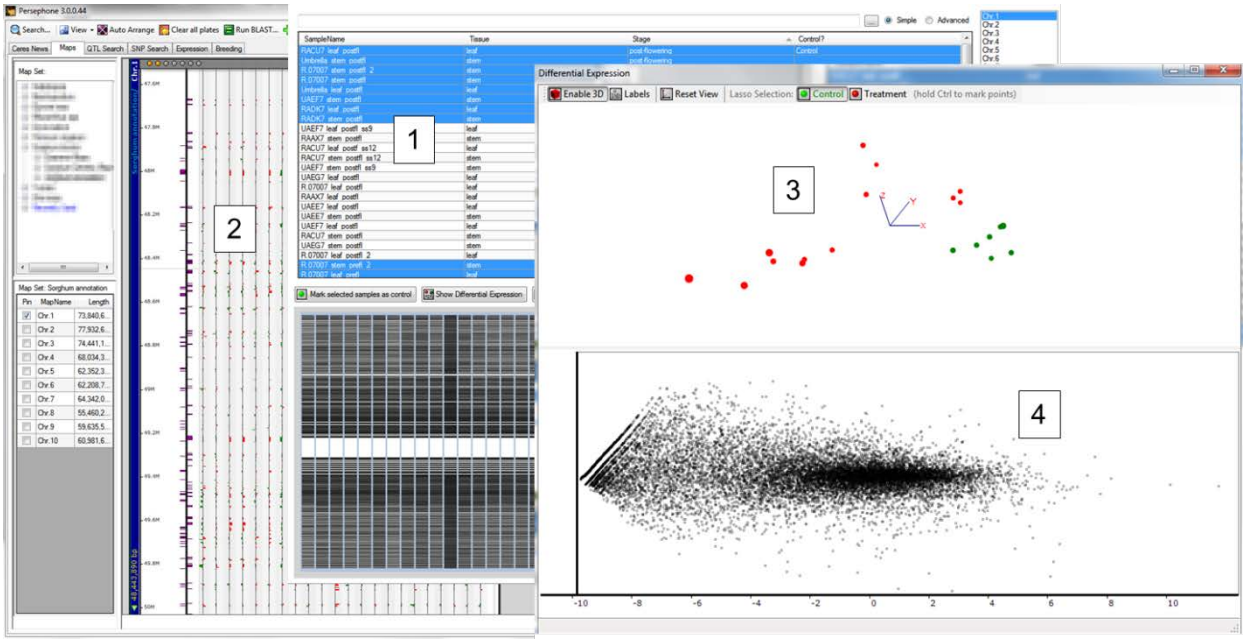
Investigating Genes



In this example, a user has identified that markers are present within this gene and many of the selected genotypes are segregating for these markers (*see 1*).

Here, the user has used the “gene information card (*see 2*),” which provides additional information, including any available ortholog or homolog data which may have other research information. These information cards are fully customizable to manage any type of data associated with the given element, including links that can be added for navigating to external or alternate internal sites. The user also has selected the differential expression data available for this gene (tab on the gene card) which the user can filter based on any number of variables within the experiments (*see 3*). The visualization of the data makes it clear that the selected gene is differentially expressed between different tissue types (red vs. aqua).

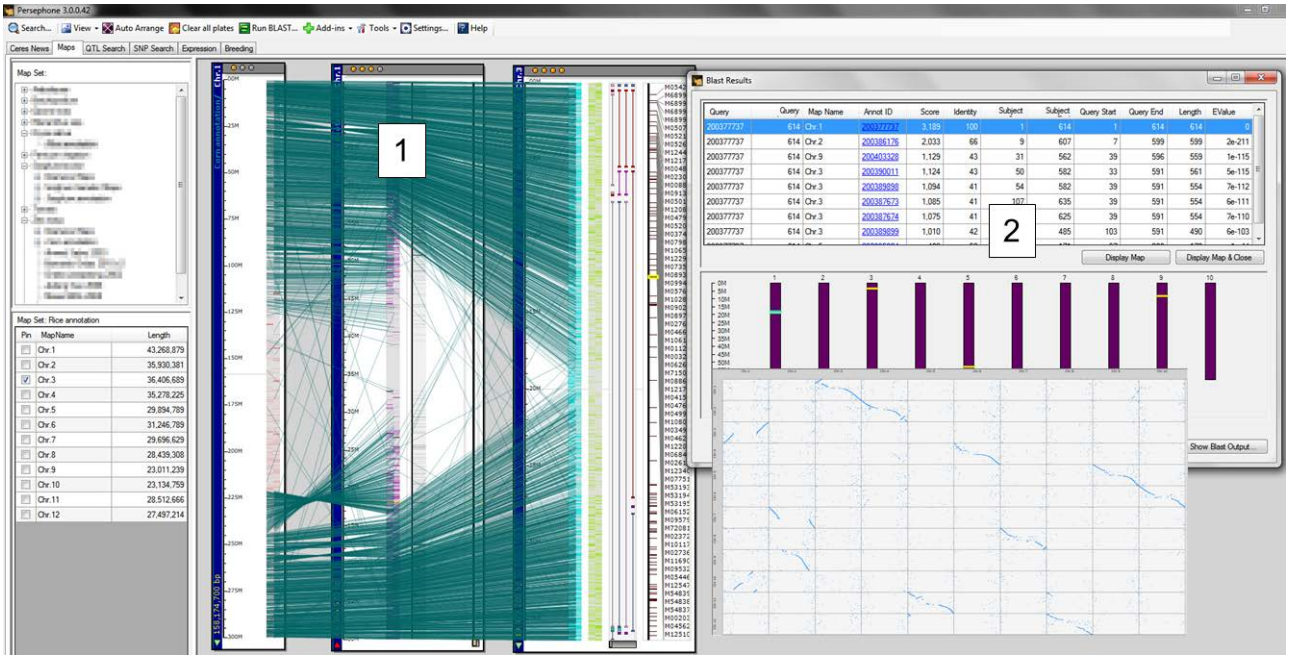
Experiments Directing Exploration



In this case, a user has decided to start a research project by searching through RNA-seq (expression) experiments. The first step is to filter samples from a set of available expression experiments (*see 1*). The expression of the selected samples is compared on ~30,000 genes and this result is visualized on the chromosome and gene level immediately (*see 2*) or alternatively a 3D MDS (multi-dimensional scaling) plot is created allowing rotation and zooming (*see 3*). The user can now easily differentiate groups of samples or experimental variables that have similar or different genome wide expression profiles.

At this point, the user can select different control and treatment sample sets and re-run the real-time analysis as many times as needed. Once a final selection has been made, a dot graph shows the genes that are more highly differentially expressed between treatment and control (*see 4*). The user can then circle selected dots on the graph, which provides identification of the gene, chromosome position and function. The user can then review the gene information and quickly navigates from this information to visualizing the gene on the chromosome vertical view previously illustrated. From here, the user can add additional data, including gene information, genotype information, allelic differences, and many other data types.

Comparative Genomics



In this example, a user has some research results which they would like to further validate or compare by using available reference model organisms.

Here, the user has selected corn chromosome 1 as the main target and found syntenic relationships with chromosome 1 in sorghum and chromosome 3 in rice (*see 1*). The display shows the syntenic regions and genes across these species. Rice has a large set of QTLs/trait associations that the user can now utilize to better understand their target in corn. In addition, the user can easily identify many small and large inversions which may provide greater insight. The user can also utilize the real-time BLAST functionality to identify genes or sequences that are similar across organisms with a graphical display to provide visual reference for the BLAST results (*see 2*).

FAQS

How do I get more information about Persephone?

- Please navigate to www.Persephone.net where you can find general information about the software. For more detailed information, please use the Contact Us form or email address found on www.Persephone.net.

Can I try Persephone?

- A 30 Day Free Trial is available (with no credit card needed). The data is a variety of publically available data that is hosted through AWS (Amazon web services). Please navigate to www.Persephone.net and select "Get a 30-day FREE TRIAL".

What support is available with Persephone?

- Persephone is a fully supported commercial software package. Support is provided for customers of the software. A dedicated email address is available (PersephoneHelp@ceres.net) and a Persephone team member will get back to you with an answer.

Can I import data into Persephone?

- For software administrators, minor modifications to your procedures may be necessary to load data into the Persephone database. Currently several data loaders are available and can be modified. This is available on the client/server software licensing.
- For all, our current functionality allows imports of generic excel sheets containing feature information and chromosome coordinates. This excel file can be drag-n-dropped onto the Persephone screen and the excel data will be viewed in conjunction (stacked) with all available data in the system.
- VCF files are also handled by the software. Example of VCF and Excel files can be found @ www.persephone.net

How can I export data from Persephone?

- Persephone has several ways data can be exported from cut-n-paste to a true export into excel/text. All major features can be exported with ease from whole chromosomes, to selected regions, to individual genes/features, and down to single nucleotides.