v 0.9, preprint of Jan 14, 2016 (Systems Genetics: *Methods in Molecular Biology*, Schughart K, Williams RW eds; Humana Press, in press)

GeneNetwork – A Toolbox for Systems Genetics

Megan K. Mulligan, Khyobeni Mozhui, Pjotr Prins, Robert W. Williams

Summary

The goal of systems genetics is to understand the impact of genetic variation across all levels of biological organization, from mRNAs, proteins, and metabolites, to higher order physiological and behavioral traits. This approach requires the accumulation and integration of many types of data, and also requires the use of many types of statistical tools to extract relevant patterns of covariation and causal relations as a function of genetics, environment, stage, and treatment. In this protocol we explain how to use the GeneNetwork web service, a powerful and free online resource for systems genetics. We provide workflows and methods to navigate massive multiscalar data sets and we explain how to use an extensive systems genetics toolkit for analysis and synthesis. Finally, we provide two detailed case studies that take advantage of human and mouse cohorts to evaluate linkage between gene variants, addiction, and aging.

1. Introduction

GeneNetwork (www.genenetwork.org, GN) is a web service for systems genetics. It started in 2001 as *WebQTL*—an online version of Ken Manly's *Map Manager QT* program [1] combined with data sets in the *Portable Dictionary of the Mouse Genome* [2]. GN is a data repository and analytic platform for systems genetics that integrates large and diverse molecular and phenotype data sets. Just over 1000 papers listed in Google Scholar have used GN in many different ways.

GN was initially used as a traditional forward genetics tool to map quantitative trait loci (QTLs) and expression QTLs (eQTLs) in sets of recombinant inbred (RI) strains and standard genetic test crosses, including F2 intercrosses and backcrosses [3]. As the number and variety of data types grew it became practical to implement multivariate type analysis in GN—namely, the genetic covariation among large numbers of phenotypes [4-6]. This kind of assembly, analysis, and integration of sets of phenotypes and even entire phenomes is a hallmark of systems genetics and is the forerunner and experimental companion of personalized health genomics and precision medicine. Thanks to recent breakthroughs in sequencing technology, GN can now also be used for novel reverse genetics approaches such as phenome-wide association studies (PheWAS). In a typical reverse genetics approach, gene function is determined through manipulation, either by gene deletion (knockout), addition of altered sequence (knock-in), silencing (RNA interference or RNAi), or gene editing (e.g. clustered regularly-interspaced short palindromic repeats or CRISPRs). Similar to these more traditional approaches, a PheWAS begins with known genes and sequence variants and then tracks down sets of linked biomarkers and phenotypic consequences [7-9].

At its most basic level, GN is a tool for studying covariation and causal connections among traits and DNA variants. This sounds simple enough, but it can be challenging to know how to get started and how to navigate and use the many program modules and options. Here we provide detailed instructions for using GN along with "worked" examples and some test questions (and answers) that should ease entry into this resource. All examples and figures were taken from the production version 1 of GN (late 2015). While the interface may change in the next few years (GN version 2, GN2), all of the logic, data types, and procedures described here will still be applicable.

The potential scope of GN analysis tools is broad—well organized collections of genetic, genomic, and trait data from different species can be integrated easily—either

as private or open data. At this point GN includes curated data sets for a variety of model organisms and plant species, including humans, monkeys, rodents, *Drosophila*, and *Arabidopsis*, soy, and barley. Data are usually open and exportable, and data typically include information for hundreds to thousands of individuals with matched genotypes for thousands to millions of markers (usually SNPs), array or RNA-sequencing (RNA-seq) data for tens of thousands of transcripts, and in a growing number of cases, proteomic, metabolomic, metagenomic, behavioral, and morphological data.

Massive omics data sets are unwieldy to access, normalize, and analyze. Even those skilled in bioinformatics spend more than half of their time simply wrangling, reformatting, and error checking data sets to match the requirements of different workflows. GN spares the user most of these problem. Data are formatted and normalized, and usually come with good metadata (often in the form of links to more information). This greatly simplifies QTL and eQTL analysis, candidate gene discovery, coexpression analysis, and hypothesis testing [3,10]. The GN toolkit includes many search functions, tools to study correlation and partial correlation, multiple QTL mapping methods (including R/qtl, PLINK, and GEMMA, and FaST-LMM in GN2), and powerful dimension-reduction techniques (principle component analysis and weighted gene coexpression analysis), network construction, enrichment analysis, variant analysis, and links to key informatics resources such as NCBI (www.ncbi.nlm.nih.gov), the UCSC Genome Browser (genome.ucsc.edu), BioGPS (biogps.org), the GWAS Catalog (www.ebi.ac.uk), Gemma (www.chibi.ubc.ca), the Allen Brain Atlas (www.brain-map.org), and GeneWeaver (GeneWeaver.org).

In this chapter we introduce the basic architecture of GN (section **2**) and work through two detailed cases studies (sections **4.1** and **4.2**) that analyze both mouse and human data sets. We also explain how GN links to other web sites that provide complementary resources and analysis tools (section **3**). Throughout the chapter we provide a series of questions that can be used to test your proficiency. Answers are provided at the end of the protocol in the **Notes** section. Both **Case Study 4.1** and **4.2** provide detailed protocols needed to exploit GN data resources and to test specific hypotheses. Work through both of these examples and use the notes to gain an excellent understanding of the range of applications and types of questions that can be addressed and often answered using a systems genetics approach.

2. Organization

The first challenge in using GN is to locate cohorts (groups of subjects or samples) and associated data sets. The hierarchical organization of GN's main **Select and Search** menu is simple and makes it relatively easy to find relevant data sets (**Fig. 1**). To get data, after opening the browser, select the most appropriate **Species** from the dropdown menu. For an open-ended search of phenotypes you can also select **All Species** at the bottom of the menu. The next steps are to select the **Group**, **Type**, and **Data Set** from the drop-down menus. For many groups, a combination of phenotypes, genotypes, and molecular data are available. This makes it possible to perform QTL mapping and the analysis of trait and gene covariation. **Table 1** provides a sample of human and rodent data sets that are amenable to these types of analyses.

As a navigation aid in this protocol, all active links in GN (buttons and linked text) and all data that you type into search fields such as **Get Any** are displayed using ***bold italic*** font. In contrast, page names, titles, column headers, and static menu items are displayed using **bold** font.

2.1. Types of Data

Almost all human data sets in GN include gene expression measurements (**Table 1**). In addition, several data sets also include genotypes and can therefore also be used for eQTL analyses. Examples include all of the human GTEx data sets and several human brain and liver expression data sets: *Brain, Aging: AD, Normal Gene Expression with Genotypes (Myers)* and *Liver: Normal Gene Expression with Genotypes (Merck*) (**Table 1**). Concerns about subject confidentiality sometimes limit the amount of data available for human cohorts. Non-human cohorts, such as rodent populations, do not suffer from these restrictions and often contain more levels of data (**Table 1**). The rodent cohort with the most extensive data collection is currently the BXD family of strains derived from a cross between C57BL/6J (B) and DBA/2J (D) [11]. Inbred panels and RI strains represent stable populations that allow for deep resampling of individual genotypes and the accumulation of many different levels of data over time, and across laboratories and research communities enabling replication of research and the study of the pleiotropic actions of variants. The BXD set, for example, includes a wide variety of trait measurements collected over the last four decades [12]. Other populations commonly used for quantitative genetics and systems genetics include F2 intercrosses and outbred populations such as heterogeneous stock (HS) mice. For most F2, outbred, and human populations, each individual is truly unique and collecting multiple levels of data and studying gene-by-environmental (GXE) interactions and lab-to-lab replication is usually not practical.

2.2. Starting an Analysis

The main GN search page and an overview of a typical workflow are shown in **Figure 2**. Data sets are selected based on **Species**, **Group**, and **Type** (**Fig. 1**). Detailed information and metadata can often be reached by clicking the ***Info*** buttons to the left (**Fig. 2**). **Data Sets** are queried using either **Get Any** or the **Combined** options. Searching with **Get Any** performs matches to entered text using the logical OR operator. For example, if the term "*alcohol ethanol*" is entered into **Species = Mouse**, **Group = BXD**, **Type = Phenotypes**, and **Data Set = BXD Published Phenotypes**, then the search will return all matches for "*alcohol*" or for "*ethanol*" (>500 results). In contrast a search for "*alcohol consumption*" in the **Combined** search option (**Fig. 2B**) uses the logical AND operation and generates far fewer results. Very long lists of gene symbols or probe set IDs—a thousand or more—will fit into these search boxes.

To get started, experiment with the **Quick HELP Examples** located just below the **Combined** search option (**Fig. 2**, center). Test whether you can find all genes on human chromosome (Chr) 21 that have high expression (>4.0 $\log_2$ RPKM) in the frontal cortex (**see Note 1**). Search queries are dependent on the **Data Set** type. For example, genotype data sets can be searched by marker name or marker position; phenotype data sets can be searched by phenotype description or authors' names; gene and protein expression data sets can be searched based on expression level, gene location, gene symbol, a Gene Ontology category (GO), or even by NCBI *Gene Reference into Function* (GeneRIF) text string.

To compare and improve compatibility across data sets, most array data have been $\log_2$ transformed and rescaled to an average of 8 and a standard deviation of ± 2 units. This is true of Affymetrix and Illumina array data. However, Agilent data report gene expression as the $\log_{10}$ of the ratio between a specific tissue compared against a reference pool of multiple tissues (mlratio). RNA-seq data is usually normalized to $\log_2(\text{RPKM} + 1)$.

Many of GN data sets can be searched for traits or transcripts based on QTL position and significance levels (LRS or LOD score). Transcripts or proteins that

are controlled by variants in or near their parent gene produce so-called cis-acting expression QTLs (cis eQTLs) whereas those that are controlled by a more distant locus, usually on a different chromosome, produce trans-acting QTLs (trans eQTLs) (**Fig. 3**). Test whether you can find a set of proteins in the mouse liver that are strongly controlled by trans eQTLs (**see Note 2**).

2.3. Create a Trait Collection

Once you have selected a **Data Set** and submitted a search, results will appear in a **Search Results** page (**Fig. 2C** and **Fig. 4**). From this page, select the individual traits, transcripts, or gene markers for additional analysis by adding them to a **Trait Collection** (**Fig. 2D**). Do this from the **Search Results** page either by selecting all rows with the *Select* icon (**Fig. 4A**) or by selecting a subset of rows with the **Add** icon (**Fig. 4B**). Trait collections are usually restricted to a single species and group. Comparisons across groups and species are possible, but in most cases this involves assembling several **Trait Collections**—one for each group.

From either **Search Results** or from a **Trait Collection** (**Fig. 5**) you can inspect traits in greater detail by clicking on their *Record ID* or *Trait ID*. This will direct you to the **Trait Data and Analysis** page (**Fig. 6**), which contains links to other web resources and GN tools.

The GN banner *Search* pull-down lists additional options, each of which is reviewed briefly below (**Fig. 2E**). *Search Databases* and *Trait Collections* are simply navigations aids to quickly get back to these two pages.

Tissue Correlation computes correlations of gene expression level across sets of 26 different tissues or 32 different brain regions from inbred (isogenic) strains of mice. Variation in expression is purely due to differences among cell and organ systems rather than being due to genetic or environmental factors. The output tables and graphs are particularly useful when studying genes with minimal annotation or when testing the hypothesis that expression of two or more genes are jointly regulated across tissues.

SNP Browser, *Interval Analyst, QTLminer* all provide three different ways to screen for genes and gene variants within defined genomic regions—but currently only for the mouse genome. *QTLminer* is the most comprehensive of the three tools, and takes advantage of the many levels of data available in GN. This tool can provide output tables that includes many types of QTL information, data on gene expression, and genetic variation across multiple mouse groups [13].

GeneWiki allows anyone to add notes on genes, transcripts, or proteins to GN. It is essentially an open public notepad with good search functions. *GeneWiki* incorporates current NCBI GeneRIF annotations.

GenomeGraph provides a way to review global genetic modulation for many gene expression data sets. This tool plots the physical position of each gene against the position of the highest linkage score for the corresponding transcript or probe (this function is not yet available for human data sets). *GenomeGraph* provides two complementary overviews (see the tabs) of the distribution of cis- and trans-eQTLs. One of these is suitable for figures, while the other is interactive and enables zooming and clicking on individual transcript/marker coordinates. The *GenomeGraph* is used to detect both the cis-acting eQTLs and prominent trans eQTL bands—loci that modulate the expression of large numbers of transcripts or proteins [14]. Can you use this tool to check for trans eQTL bands in mouse liver (**see Note 3**)?

Scriptable Interface is a more complex option that enables direct queries of GN databases using a set of keywords and commands—an application programming interface (API) that can be used to link one web resource with another. It is possible to access or download data and tools using R, Python or other code and scripts. The

API consists of a query that returns results in a JSON format that is easily loaded locally. The R/qtl package, for example, can read GN REST API data by default. Examples of such functionality are:

1. Fetch all genotype data belonging to a cross or sample.
2. Fetch all phenotype data belonging to an experiment or population.
3. Get the genome scan results for a particular phenotype.
4. Get a list of phenotype correlates and their correlations.
5. Get a list of phenotypes with a QTL in a given interval.
6. Get a list of genes matching a QTL in a given interval.

The final three pull-down items—***Database Information***, ***Data Sharing***, and ***Annotations***—provide documentation and download tools.

In addition there are several useful resources available under the ***Help*** tab in the banner menu (**Fig. 2E**). Useful guides and tutorials outlining how to use the GN web resource can be accessed under the ***Movies***, ***Tutorials***, and ***HTML Tour*** options. Extremely useful explanations to frequently asked questions and for terms and tools used in GN can be found in the ***FAQ*** and ***Glossary of Terms***. The glossary has been hand curated since the inception of GN and is a great companion guide for all new users.

3. The GeneNetwork Toolbox

Now that you are familiar with the organization of data and typical search workflows, we can introduce resources available for trait analysis in the extensive GN toolbox. We will explore these tools first at the level of a single trait, and then at the level of multiple traits.

3.1. Tools for single trait analysis

The **Trait Data and Analysis** page is key to using GN and includes many useful tools for studying single traits (**Fig. 6**). Options differ by data type and species. A trait such as body weight has very different **Resource Links** than mRNA, protein, metabolite, and genotype data. Most data sets that include transcript or protein assay measurements include links to resources that provide information about function, homology, expression across tissues, and genomic location. These include ***Gene*** pages at NCBI, ***OMIM***, ***HomoloGene***, ***UCSC*** Genome Browser, and ***BioGPS***. Other links are focused on protein structure and function, including ***STRING***, ***PANTHER***, and ***Wiki-PI***. ***Gemma*** and ***ABA*** provide access and analysis of thousands of transcriptome and *in situ* expression data sets, respectively. ***EBI GWAS*** searches human genome-wide association studies for matches to selected transcripts or proteins.

The row of icons labeled ***Add***, ***Find***, ***Verify***, ***GeneWiki*** etc. link to large GN database resources. The ***Add*** icon is used to build up collections of traits for network analysis in a **Trait Collection**. ***Find*** locates similar expression traits in other data sets and other species. ***GeneWiki*** provides a summary of gene and protein function based on notes made by GN users and published data. It is simple to add your own notes to GN by selecting ***GeneWiki*** and then ***New GeneWiki Entry***. ***SNPs*** links to a **Variant Browser** that is identical to the ***SNP Browser*** accessed from the GN banner under the ***Search*** tab. ***Verify***, ***RNA-seq***, and ***Probes*** provide quality control

information about transcripts and peptides. Both *Verify* and *RNA-seq* link to GN mirrors of the Genome Browser.

The *Verify* and *RNA-seq* tools uses the transcript, peptide or probe sequence to align against the reference genome. The BLAT reanalysis results and annotations at the top of the *Trait Data and Analysis* page should match, but mismatches are frequent and arise from poor annotation, poor sequence selection, or ambiguous alignment. The *RNA-seq* tool performs the same type of BLAT alignment but includes tracks with data on all genomic variants segregating between the parents of the BXD mouse cohort [15], and expression profiles from whole brain [7] and striatum [16] generated by RNA-seq. Sequence variants are displayed in the **DBA/2J Sequence and Structural Variation** track and RNA-seq data from brain (B, D, and BXD strains) and striatum (B and D strains) are displayed in the **RNA-seq: Brain (BR) ABI, N tags/nt, adjusted** track and the **RNA-seq: Striatum (STR) ILM, N tags/nt, adjusted** track, respectively. These data are useful for visualizing variants within genes that may affect expression, and can also be used to determine whether variants overlap probe sequences. Array platforms have all been designed based on the genome of a single reference genome (C57BL/6J in the case of mice, Brown-Norway in the case of rats). The use of a single genome for design purposes can result in biased hybridization in array studies and biased alignment in RNA-seq studies [17]. The RNA-seq data is also useful for validating expression differences detected using array platforms. The related *Probes* tool is useful only for Affymetrix data sets and is used to evaluate the performance of individual array probes.

3.2. Analysis and mapping methods for single trait analysis

The lower set of four panels (**Fig. 6C**) on the **Trait Data and Analysis** page include the core computational functions of GN—**Basic Statistics**, **Calculate Correlations**, **Mapping Tools**, and **Review and Edit Data.**

Basic Statistics is used to summarize statistical properties of single (univariate) traits. Open this section (click on the bar) and select the **Basic Table** tab or **Probability Plot** or **Bar Graph** tabs. These options are reviewed below in detail in **Case Studies 4.1** and **4.2**.

Calculate Correlations is used to compute the bivariate correlations between the reference trait and any other set of traits that has been measured in the same **Group**. Open this section and select a target **Database**, the number of correlations to **Return** (default is **top 500**, but the range is between **100** and **20,000**), and the method of correlation—**Pearson** or **Spearman Rank**. Note the tabs: GN can compute three types of correlation—**Sample r**, **Literature r**, and **Tissue r**. **Sample r** does what you expect. It computes correlations using values listed at the bottom of the page. **Literature r** computes correlations between genes based on their shared vocabularies in PubMed. The same method is applied when using the *GCAT* tool (http://binf1.memphis.edu/gcat/help.html, [18]). Finally, **Tissue r** computes correlations based on variation in expression of genes across about 30 tissues and organs in mouse (identical to the *Tissue Correlation* tool). All correlation output results are displayed in a **Correlation Table**. Any of the rows in these tables can be evaluated in their own **Trait Data and Analysis** page by selecting the *Record ID*, or large sets of rows and covariates can be analyzed as a group using tools at the top of **Correlation Table** page. Use either the *Index* check boxes or the *Select*, *Deselect*, *Invert*, and *Add* icons to move traits into a collection.

Mapping Tools includes a number of on-line "live" QTL mapping methods. The association function in PLINK is currently the default for human GWAS. **Interval** mapping is the default for almost all plant and non-human cohorts. Interval mapping exploits Haley-Knott regression equations to evaluate the linkage across all autosomes and chromosome (Chr) X. Linkage is displayed either as a

likelihood ratio statistic (LRS) or the log of the odds ratio (LOD). Both scores provide an estimate of the statistical strength of linkage and the LRS is derived from the LOD score by multiplying by 4.61. A linkage probability of 0.001 is roughly equivalent to a LOD of 3 and an LRS of 13.8. Genome-wide association studies (GWAS) in humans often use a $-\log_{10}(P)$ value where P is the probability of linkage between differences in genotype and differences in trait or disease severity.

Mapping Tools also include **Marker Regression**, a very simple method that computes statistics only for individual marker genotypes. Composite interval mapping (*Composite*) is a variant of simple interval mapping that enables control for one or more other markers. It is equivalent to mapping the results of a partial correlation. *Pair-Scan* is an experimental mapping option implemented for larger RI sets (samples of 50 or more strains) that searches for epistatic interactions among loci.

Review and Edit Data contains a working copy of the trait values for each case. Outliers, if any, are highlighted in yellow. Users can manually change trait values, select subsets of individuals for further analysis, exclude outlier values, export values for analysis offline, or reset to the original values.

3.3. Tools for multiple trait analysis

A key feature of GN is access to several different levels of data that all originate from well defined groups of subjects or cases. The levels can range from genotypes to behavior, but can also include different treatments, developmental stages or laboratory settings. Users can assemble computationally coherent collections of traits to explore joint gene control, gene-by-treatment, gene-by-lab, and gene-by-environmental interactions. Users may want to examine expression for a single gene, gene families, or members of a biological pathway across multiple tissues. To accomplish these tasks it is necessary to find the data types and then assemble them into a single collection. This is done using the **Search Results** page, the **Trait Data and Analysis** page, and several other tables generated by tools in GN, particularly **Correlation Tables**. Once these multiscalar data sets have been assembled, a number of new tools are available for joint analysis from the **Trait Collection** (**Fig. 4**). Basic actions are similar to those found in the **Search Results** page, including *Select*, *Deselect*, and *Invert*. Other actions include *Remove* and *Export*.

Analysis tools that are optimized for large collections of genes and proteins include *Gene Weaver*, *GCAT*, *Gene Set* analysis (WebGestalt), and *BNW* (Bayesian Network Webserver). *GCAT* uses text mining to determine if a list is functionally coherent and related based on the literature [18]. *Gene Set* searches for significant enrichment based on GO categories (functional annotations describing gene function or location) and *Graph*, *Matrix*, *Partial*, and *Compare* are tools that leverage correlations to identify patterns and relations among traits. The *Graph* tool is used to construct and visualize correlation networks from selected traits. The lines or edges connecting trait nodes can be filtered and exported to the open source Cytoscape software platform or graph images can be reconfigured and saved as a PDF. *Matrix* generates correlation matrices from any number of traits using both Pearson and Spearman coefficients. Scatter plots can be generated for each pairwise comparison. Principal component analysis (PCA), a data reduction and pattern detection technique, is also performed and eigenvectors are generated for the principal components that capture the majority of the variation in expression of selected traits. Eigenvector values can be added to the **Trait Collection** and are handled by GN in the same way as other traits. The pattern of expression captured across cases by each eigenvector trait can be used for mapping, to find additional correlates, or to check for technical artifacts.

The *Partial* correlation tool computes correlation between traits after controlling for other traits, markers, or cofactors such as age or sex. Partial correlations can be calculated for a subset of traits in a **Trait Collection** or against an entire data set. Select at least one **Primary** trait (X), one or more **Target** traits (Y), and a set of **Control** traits (Z). Again you have the option of computing either Pearson's r or Spearman's rho partial correlations.

The final correlation tool is ***Compare***. This tool is used to identify intersecting sets of traits across data sets from the same **Group** that are correlated with selected traits in the **Trait Collection** based on a user defined threshold. It will essentially compute the intersecting values of a Venn diagram using 2 to 20 or more variables in the collection.

Tools for exploring the genetic control and mapping of multiple traits from the same collection include ***QTL Map*** and ***Heat Map***. The ***QTL Map*** tool allows users to compare QTLs for up to ten traits globally or by single chromosome. This tool is useful to visually explore traits that may be modulated by the same chromosomal position. The ***Heat Map*** tool is used to compare global patterns of genetic modulation for up to 500 traits at a time. Individual traits are represented by columns with genomic position shown by row. Significant QTLs are indicated for each trait as intense blue or red bands depending on whether expression is increased by the maternal or paternal allele (blue and red respectively for the BXD RI set).

The tools available for individual or multiple trait analysis in GN are designed for users to explore data sets and detect relations among traits that are driven by genetic and non-genetic factors. The underlying genetic variants responsible for some of these associations and their potential impact on higher-order phenotypic variation can then be evaluated. We provide two case studies below that put these tools and data sets into context, and that illustrate how they can be used in a systems genetics approach.

4. Case Studies and Workflows

In this section we have provided case studies for both mouse and human data sets that illustrate the utility of GN. Other case and use studies can be found in this book and other publications [19].

4.1. Mouse Case Study

The BXD family of strains and their parents—C57BL/6J (B) and DBA/2J (D)—differ greatly in their preference and sensitivity to alcohol and many other drugs. As a result, the BXDs have been used as a genetic model system to map loci and define gene variants that may be involved in addiction. Using data and tools in GN we can ask whether there are any gene variants associated with addiction and whether gene expression varies as a function of strain and genotype. We can also test the possible causes and consequences of variation in gene sequence and gene expression. This case study takes you through the main steps in this process.

1. Navigate to the **Select and Search** page at www.genenetwork.org.
2. Choose an expression database by picking the following options. **Species** = Mouse, **Group** = BXD, **Type** = Hippocampus mRNA, **Data Set** = Hippocampus Consortium M430v2 (Jun06) RMA (the third data set in this menu). For this example we will use an Affymetrix hippocampus expression data set that uses the RMA normalization method. This is the most commonly used normalization method for Affymetrix arrays and is therefore the best choice for comparing across tissue and even species data sets. The hippocampus is one of many brain regions important for episodic memory formation and

spatial navigation. It is also particularly sensitive to many types of environmental and pharmacological perturbations. For more information (metadata) about how this and other data sets were generated, click the ***Info*** button to the right of the data set name.

3. Search for genes. Enter the following search string in the **Combined** option: "*Mean=(8 16) cisLRS=(10 99 10) RIF=addiction*" (remove the double quotes). This search will return all transcripts (in this case also called probe sets) that have a mean $\log_2$ expression between 8 and 16 units and whose expression is modulated by a cis-acting eQTL with an LRS between 10 and 99 that have also been linked to addiction. By using the **Combined** search field, all three components of the query have been combined automatically using a Boolean AND operator. The first component—*Mean=(8 16)*—limits the search to transcripts that have moderate to very high expression level. Eight is the average $\log_2$ expression level for most array expression data sets in GN while 16 is very high. Typically, a trait with an average $\log_2$ expression value less than 6 is not considered expressed.

   The second component of the query—*cisLRS=(10 999 10)*—limits the search to those transcripts associated with a cis eQTL LRS value between 10 and 99. An LRS score of 10 corresponds to a LOD of 2.2 and is roughly associated with a nominal (point-wise) *p* value of 0.01. Similarly, an LRS of 99 is equivalent to a LOD of 21.5. The third parameter (also 10) included in the query limits how far the eQTL location can be from the corresponding gene associated with the mRNA. In this case we set a 10 Mb exclusion limit. Finally, the third query term—*RIF=addiction*—limits the search to genes that have been annotated with the term "*addiction*" in NCBI GeneRif collection.
4. Click on the ***Search*** button to explore the results of this query. The search returns 31 records (November 2015). The **Symbol** and **Description** columns provide the gene symbol and full name. The **Record ID** column gives the probe, exon, or transcript ID that has been used to measure expression. The particular part of the mRNA that is the target of the assay is often listed in the **Description** column after the gene name (e.g., "distal 3' UTR)". Gene location is given in the **Location Chr and Mb** column, whereas the location of highest LRS associated with the trait is given in the **Max LRS Location Chr and M**b column. The last **Add** column lists the additive effect of alleles at the **Max LRS Location**. In this case, the positive and negative values of **Add** indicate that expression is increased by the paternal (*D*) or maternal (*B*) allele, respectively. All of these **Search Result** columns can be sorted. Initially the list is sorted alphabetically by **Symbol** but can also be sorted by probe set genomic location (**Location Chr and Mb**) or by eQTL strength (**Max LRS**). The top 10 unique genes sorted by **Max LRS** include *Rb1*, *Csnk1e*, *Cntnap2*, *Cdkn1b*, *Mpdz*, *Gria1*, *Comt*, *Gabra2*, *Kcnj3*, and *Slc1a2*. ***Select*** all and then click ***Add*** to move all of the search results into a **BXD Trait Collection** for further analysis.
5. To study the expression of the *Rb1* transcript in greater detail, select its **Record ID** or **Trait ID** (***1417850_at***) to navigate to the **Trait Data and Analysis** page (**Fig. 7**). Each trait can be examined in more detail in this manner, whether it is a transcript, peptide, metabolite, genotype, or behavioral trait. There are a number of tools for single trait analysis on the **Trait Data and Analysis** page. We now will take you through many of these in the next few steps.
6. Examine the expression of *Rb1* across all of the BXD family members included in the data set using the **Basic Statistics** track. Expand the track by clicking the "**+**" symbol or in the gray bar. Under the **Include** drop-down menu select "BXD Only". The **Basic Table** provides simple univariate statistics such as **N of**

Samples, **Mean**, and **Range**. This particular data set includes 71 samples with a **Range (fold)** of 2.34 fold on this $\log_2$ scale.

The **Probability Plot** tab is a critical tool for detecting outliers and for reviewing the distribution of trait values. If the distribution is close to normal then the observed **Trait values** on the Y-axis will line up well with the **Expected Z score**s on the X-axis. Deviations from the expected straight line of normality—an S-shape, a set of abrupt breaks (as here), or a set of ripples—indicate that one or more large effects may be influencing the distribution. A strong QTL or a sex difference can produce such effects. For an example of a sex effect (and potential confounder), review the expression of the Xist gene (probe set 1436936_s_at).

Another means to visualize data distributions are with **Bar Graph (by rank)** and **Bar Graph (by name)**. By selecting **Bar Graph (by rank)** you can see that expression of *Rb1* is reasonably close to expectation (a normal distribution), although there are two or three small breaks. This could indicate the presence of one or more loci that have a modest impact on expression and that are segregating among the BXD family members. In this case there are no outliers.

Had outliers been detected it would have been necessary to handle them in the **Review and Edit Data** section toward the bottom of the page. This part of the **Trait Data** page contains a working copy of the data values. Values can be deleted or blocked with an *X*. Data can be modified, winsorized, or truncated to make them less extreme. Even a single outlier can have a very adverse impact on genetic mapping—often increasing the risk of false-positive QTLs and producing Pearson correlations that are inflated. The original values can be ***Reset*** or downloaded using the ***Export*** function.

7. Perform QTL mapping using the **Mapping Tools** track, below the basic statistics and calculate correlations tracks. Very fast interval mapping is a powerful feature of GN that makes it possible to carry out complex trait analysis of most cohorts in real time. Click on the ***Compute*** button under the **Interval** tab using the default options. We already noted that the distribution of *Rb1* expression had some breaks. We can now explore possible causes of these disruptions to the expected normal distribution by mapping trait variance.

   The results of whole genome interval mapping are displayed as a graphical map with chromosome number and megabase position displayed at the top and bottom of the map, respectively. You can change to a genetic map measured in centimorgans (cM), but this is rarely useful when a physical map is available. The ***LRS*** linkage score is displayed on the left Y-axis. **Blue**, **red** and **green** lines plot the **LRS**, the additive coefficient for the *B* allele (inherited by roughly half of the strains from C57BL/6J) and *D* allele across the genome, respectively. The horizontal red and grey lines show the threshold for significant and suggestive linkage scores based on mapping 5000 permutations (see the **Histogram of Permutation Test**). A permutation is simply the random rearrangement of elements in an ordered list (in this case a list of genotypes and associated trait values). A permutation test is a method for evaluating statistical significance by randomly reshuffling and recomputing scores for list elements. To achieve a significance of $p = 0.05$, the original association score between genotype and trait expression must be greater than at least 95% of all permuted associations. All of these calculations, including the default 5001 genome scans, and the display, usually take less than a minute to generate.

   The visual display of the graph can be altered by changing the attributes in the box above the graph. Note the purple arrowhead at the bottom of the X-axis that indicates the position of the cognate gene. Here we see strong and highly significant linkage between expression of *Rb1* and a locus on Chr 14 that

overlaps the physical location of the *Rb1* gene, a cis eQTL. Change the units to **LOD** in the attribute box above the map and click on the Chr 14 icon to zoom in and replot the map using a LOD score scale.

To look at the relationship between gene expression, genotype, and the segregation pattern of parental alleles in greater detail, check the **Haplotype Analyst** box and change the **View** to 70 to 80 Mb in the attributes box and then select ***Remap***. This will zoom in and show the pattern of inheritance for each BXD strain with the location of gene models shown at the top of the plot followed by a map of the chromosome for each strain (strain name to the right) and the corresponding trait value sorted from highest to lowest (value to the right of the strain name). The vertical black lines represent the location of genotyped markers that reveal whether that position in the genome was inherited from the maternal or paternal strain (the corresponding marker names are shown at the bottom of the chromosome map). Similar genotypes across a set of adjacent markers define a haplotype and are represented here as large blocks of green (inherited from the paternal strain) and red (inherited from the maternal strain) with intervening undefined grey regions. Somewhere within the grey interval a recombination event occurred and more markers will be needed to resolve the haplotype blocks more completely. Blue areas are or were heterozygous when the strains were genotyped last. You may have already noticed the striking segregation of green haplotype blocks at the top and red haplotype blocks to the bottom of the chromosome map. Parental alleles at this locus are strongly associated with expression variation and this can be seen here as BXD strains that have inherited the paternal *D allele* (in green) have high expression of *Rb1* and those strains that have inherited the maternal *B allele* (in red) have lower expression (expression values shown for each strain at the far right).

It is often useful to define a confidence interval in which the candidate variant or gene driving trait variation is likely to be located based on the mapping results. One rough estimate of the confidence interval is the 1.5 LOD drop-off which is defined as the interval bordered to the left and right of the peak QTL in which the LOD score (represented by the blue line) drops by 1.5 LOD units. In this example, that would be the point on the blue line to the left and right of the peak that represents a value of 15.5 LOD. This can be roughly approximated visually from the graph such that the 1.5 LOD confidence interval defining the cis eQTL is roughly between 73 and 75 Mb on Chr 14.

To view the precise association score for any single marker and the corresponding chromosomal position, click the '***Download result in a tab-delimited text format***' link toward the top left side of the **Map Viewer** page. Note that the peak marker is rs3701623 located on Chr 14 at 73.597 Mb. To estimate the amount of trait variance that is genetic and captured by this single QTL, navigate back to the main GN **Select and Search** page (use the ***Search Databases*** option under the **Search** dropdown in the banner or click on ***GeneNetwork*** in the top left corner of the browser window). Enter the marker 'rs3701623' using the **Get Any** query under **Group** = BXD, **Type** = Genotypes, **Data Set** = BXD Genotypes and select ***Search***. This query will return information about genotypes at this marker. Select the marker and ***Add*** it to the **Trait Collection**. The collection should now contain all 31 genes from the previous search results and the marker rs3701623. Select the marker and the *Rb1* probe set, and then choose the ***Matrix*** tool. We will learn more about the matrix tool later, but for now we have just generated the ***Pearson*** (left value) and ***Spearman Rank*** (right value) correlation coefficient for our expression trait and marker. The Pearson r is 0.83 and the corresponding r^2 is ~0.7. In other words,

about 70% of the variation in hippocampal *Rb1* expression among BXD strains is explained by a cis eQTL.

8. Verify that *Rb1* is linked in to addiction or substance abuse in ***GeneWiki***. *Rb1* is a tumor suppressor with high expression in hippocampus. But is there a link to addiction of the type we expect? From the ***Wiki*** pages perform a search for the work "addiction". This will highlight entry 276. However, *Rb1* is linked to addiction in a different context: the acute need of cells for *Myc* expression to survive. Try this using another gene from the original list—*Cdkn1b* (**see Note 4**).
9. As shown above, quality control is critical. Both the ***Verify*** and the ***RNA-seq*** tools on the **Trait Data and Analysis** page are used to confirm the correct identity of probe sequences and detect possible problems associated with local sequence variants. Probe set 1450486_a_at (*Oprl1*) is a good example of how sequence variants can interfere with expression measurements. Select *Oprl1* probe set 1450486_a_at from the **Trait Collection** and link to the corresponding **Trait Data and Analysis** page.

   Confirm involvement of this gene in addiction by clicking the ***GeneWiki*** link and performing the same analysis as in Step 8. Note that the term "addiction" appears in three separate ***GeneRIF*** entries. From the **Trait Data and Analysis** page perform quality control by selecting the ***RNA-seq*** tool. This tool is similar to ***Verify*** in that it uses UCSC BLAT to align the probe set to the reference genome. The **BLAT Search Results** page (**Fig. 8**) summarizes alignment scores. Click on the far left ***browser*** link of the top row.

   The RNA-seq browser page displays many tracks (**Fig. 8** bottom). These include the alignment of the 11 probes (black rectangles), the region of the gene targeted by the probes (the 3' UTR, exons, or in rare cases, the introns), DBA/2J sequence variants, and RNA-seq expression measurements. Confirm that the probes target the right gene (*Oprl1*) and determine if any variants overlap probes and might interfere with expression measurements (**Fig. 8**).

   Note that the probe set targets *Oprl1* correctly. However, several probes overlap SNPs (probes 299709 452573; **Fig. 8**). These SNPs could impact measurements of expression in strains that inherit the *D* allele. To check whether or not expression differs between probes that overlap SNPs, use the ***Probes*** tool in the **Trait Data and Analysis** page for *Oprl1* (probe set 1450486_a_at). Affymetrix microarrays feature multiple probes whose expression is then summarized to get a measure of cognate gene expression. The ***Probes*** tool allows you to explore individual probe expression, genetic mapping, and covariation. In the case of the M430 array used here, expression is based on hybridization of 11 perfect match (PM) and 11 mismatch (MM) probes (**Fig. 9**). Use the ***Select PM*** button to select the perfect match probes and then select the ***Heat Map*** icon to look at the eQTL profile for all 11 probes (**Fig. 9**). The heat map shows the location and strength of eQTLs for each probe. A strong cis eQTL indicating higher expression in BXD strains that have inherited the *B* allele of *Oprl1* (blue, **Fig. 9**) is only associated with probes overlapping SNPs (299709 and 452573). The strong cis eQTL detected for *Oprl1* is actually a technical artifact caused by sequence variants that disrupt the hybridization of probes to their target RNA sequence in strains other than those with the reference *B* haplotype. When exploring eQTLs it is good practice to determine: (1) That the assay targets the right genes, and (2) Whether or not measurements might be impacted by sequence variants. Try this analysis on *Kcnj3*, probe set 1455374_at (**see Note 5**).

   Thus far we have searched and returned a list of genes whose expression is likely modulated by local sequence variants segregating in the BXD cohort that

may play a role in addiction. We identified two genes (*Rb1* and *Oprl1*) whose presence on the list is due to different types of technical errors. What about the remaining genes? Are these genes connected in any other way?

10. Select the top nine genes from our **Search Results** page (1417176_at, 1434045_at, 1422798_at, 1418664_at, 1448972_at, 1449183_at, 1421738_at, 1439940_at, 1437920_at, 1421202_at) and ***Add*** them to the **Trait Collection** (**Fig. 10**).

    We can now explore whether these traits are connected at the level of genetic regulation or gene expression. Select all traits and then select the ***Matrix*** tool. The output is a correlation matrix comprised of pair-wise correlations for each selected probe set (**Fig. 11**) and the results of a PCA that will be described below (**Fig. 12**). From the correlation matrix at the top of the page, we can explore whether the expression of these traits are correlated in the hippocampus of 71 BXD strains. With this number of individuals, a correlation of ~ |0.3| will be significant at a p-value less than 0.01, however, only correlation coefficients greater than |0.5| are highlighted in the matrix. For each pair-wise correlation, it is possible to generate a scatterplot that also displays the associated p-value by clicking on each correlation (**Fig. 11**). Note that nine pairwise correlations are significant ($p < 0.01$) within this gene set.

    Embedded in the ***Matrix*** tool is a module to compute principal components (PCs) and eigenvector scores. PCA is used to extract shared patterns of variation from larger numbers of traits that covary for different reasons. For example, the first PC could represent a technical error or batch effect, a second PC could correspond to sex differences, and a third PC could correspond to variation produced by a gene variant. In many cases, PCs will not correspond to any obvious single source of variance. Scores can be assigned to each subject in the analysis for each of the PCs. These PC scores (also known as eigenvector scores or even "eigengene" score in transcriptome studies) are similar to residuals and have a mean of 0. The **Scree Plot** describes the fraction of variance that is explainable by each of the PCs in descending order. For a set of randomly selected transcripts as much as 25% of the variance may be described by the first PC—often an indicator of an uncorrected batch effect. The **Factor Loadings Plot** describes how each trait loads onto, or is correlated with the first and second PCs (**Fig. 12**). In this example the first factor, or PC1, explains ~28% of the variance in expression of the nine top transcripts from our search. The PC scores can be used as composite traits and entered into GN collections and workflows just like any other trait. To perform mapping and analysis of the PC scores, select the PCA Traits link under **PCA Traits** (e,g., **PC01**) then review the scores in the corresponding **Trait Data and Analysis** page (**Fig. 12**). In this example two PCs capture most of the variation in expression. Use the **Interval** tab in the **Mapping Traits** track to perform standard QTL interval mapping. This common source of variation is not derived from a single genetic locus as there are no strong QTLs modulating either PC.

11. Construct a network graph from the **Trait Collection** using the ***Graph*** tool.
    Additional tools are available in the **Trait Collection** to analyze relations among the top genes (probe sets) in our list. Select all nine traits and the ***Graph*** tool. This tool constructs a network graph that shows all possible correlations among selected traits at a given threshold (**Fig. 13A**). Users can control the way the graph is displayed using the options provided. The type of network can be changed using the **Select Graph Method** dropdown menu. In addition, line color and style, correlation type and threshold, and node label, font, and shape are all customizable. High quality PDF or GIF files can also be generated. In

our example, *Mpdz* is the highest connected gene in the network and has four connections at a correlation of |0.3| or better (**Fig. 13**A), in contrast, *Comt* is not connected at all. Highly connected genes, sometimes called network hubs or hub genes, are thought to have important biological roles, although this is a topic of much debate in systems biology. In less complex systems (flies, worms, and yeast), such hub genes are often essential genes required for survival. However, in higher organisms the role of such hub genes is less clear. Note, that our network of nine genes (or nodes) is much too small to make grand biological conclusions, but is sufficient for an exploratory analysis and tutorial.

12. Test whether a subset of selected expression traits is enriched for biological function using the ***Gene Set*** tool. Variation or covariation, such as that observed using the ***Matrix*** (pair-wise correlations) and ***PCA*** (data reduction and pattern analysis) or the ***Graph*** tool (covariation) can indicate underlying genetic control or shared biological function. The ***Gene Set*** tool in the **Trait Collection** page can be used to investigate whether selected sets of genes share common biological functions. Select *Mpdz* and its correlates (*Chrna4*, *Gria1*, *Csnk1e*, and *Cntnap2*) and the ***Gene Set*** tool (**Fig. 13B**). This tool uses WebGestalt to compare functional GO anotations within the selected genes compared to a background gene list that includes all of the genes (probe sets) included on the M430 microarray used to generate this data set. Select ***View results*** to display a directed acyclic graph of significantly enriched functional categories (**Fig. 13C**). Even though the gene list submitted is quite small (only five genes), several categories are enriched at an adjusted *p*-value less than 0.05. These categories include signaling (*Chrna4*, *Cntnap2*, *Mpdz*, and *Csnk1e*), part of neuron projection (*Chrna4*, *Cntnap2*, and *Mpdz*), and regulation of action potential (*Chrna4* and *Mpdz*). Click on the ***Trait ID*** of each gene in the **Trait Collection** and use the ***GeneWiki*** tool to explore their function in more detail. These genes function in overlapping biological pathways, play a critical role in synaptic and intracellular signaling, and have been linked to addiction. In addition, expression of all genes is correlated and the expression of each is variable in BXD hippocampus—likely due to the presence of local sequence variants that modulate expression.
13. Perform a reverse systems genetics analysis to dissect the consequences of genomic variation on higher order traits by selecting the link for Trait ID 1449183_at (*Comt*) to navigate to the **Trait Data and Analysis** page.

Now that we have initiated a functional search and explored variation and covariation among sets of genes, let us use the vast data resources available in GN to perform a reverse systems genetics analysis to dissect the consequences of genomic variation on higher order traits. From the **Trait Data and Analysis** page for *Comt*, navigate to the ***GeneWiki*** entry. This gene has been extensively studied in human populations and in the BXD cohort. A common polymorphism in humans results in the substitution of the amino acid valine (*Val*) to methionine (*Met*), and a decrease in activity. *COMT* is involved in the degradation of catecholamines, including the neurotransmitters adrenaline, noradrenaline and dopamine. *COMT* alleles have been associated with subtle differences in risk of psychiatric disease and difference in cognition and attention. A *Comt* polymorphism also segregates among the BXD population such that the maternal strain and those BXD progeny that have inherited the *B* allele have a ~200 bp insertion (a type of mutational event in which additional DNA is added to the genomic sequence) in the 3' UTR that leads to truncation when compared to the paternal haplotype (*D* allele) [7]. Interestingly, for some *Comt* probe sets (1449183_at) this mutation leads to higher expression in those strains that have inherited the *B* allele, unless the probe sets target the most

distal part of the 3' UTR (1418701_at) that is not expressed in those cases. In the latter case, higher expression is observed in those strains that have inherited the *D* allele. To look at this interesting discordance between probe sets, use the ***Find*** tool to identify probe sets targeting *Comt* in multiple expression data sets from BXD. Using the tools introduced to you earlier in this case study, compare where each *Comt* probe set (1418701_at and 1449183_at) aligns to the reference genome, the strain distribution of expression for each probe set, and the difference in cis eQTL mapping (**see Note 6**). Note the different ***Record IDs*** for *Comt* that correspond to different probes or probe sets across different microarray platforms. Different regions of the *Comt* gene are being targeted by each probe or probe set, and this is generally true for most genes and microarray platforms. The ***Find*** tool can also be used to find corresponding probe sets for the same gene in human and rat data sets.

We know that the expression of *Comt* varies across the BXD set and we now know from ***GeneWiki*** that the causal mutation underlying this variation is an insertion. We can use GN data sets to determine the functional consequences of this variation. In other words, we can ask what phenotypes are controlled by the genetic variation at the *Comt* locus. To do this we can navigate back to the **Select and Search** page and identify phenotypes from the BXD Phenotypes BXD Published Phenotypes data set that map back to the *Comt* locus. In the **Combined** search option enter "*LRS=(9 99 chr16 16 22)"* to identify all phenotypes that have a peak QTL located within 2 Mb of the *Comt* locus on Chr 16 at 18.4 Mb. This should return at least 12 traits that we can add to our collection. Do the traits returned make sense given the role of *Comt* in the regulation of catecholamine (epinephrine, norepinephrine, and dopamine) levels? The expression of these phenotypes is controlled by a QTL that precisely overlaps the location of *Comt*. To compare the overlap in QTL mapping among these phenotypes and with the *Comt* probe set, select all phenotype traits and the expression trait in the **Trait Collection** and select the ***Heat Map*** tool. For finer mapping resolution up to 10 traits can be mapped together using the ***QTL Map*** tool.

In many cases this type of a reverse genetic analysis is complicated by the linkage disequilibrium inherent in the BXD population, which has an average haplotype block of about 50 Mb and an eQTL mapping resolution of around 1 Mb. This often results in the presence of several genes and variants within a QTL confidence interval that could control trait expression. In our case, *Comt* is the only gene within a 4 Mb interval that contains a variant. Thus, traits that map back to this locus are controlled by the variation in *Comt*. You can also use this same search query in different BXD expression data sets to find downstream expression traits (probe sets that map back to the *Comt* locus or are controlled by a trans eQTL that originates from the *Comt* locus) or to find phenotypes or expression traits that correlate with *Comt* expression.

In the preceding series of examples we have illustrated how to query the GN database and use some of the many tools available to perform systems level analyses, including genetic mapping, exploring patterns of covariation and performing a reverse genetics systems analysis to uncover the functional impact of sequence variation. All examples rely on a large and well characterized genetic reference population, the BXD cohort. In the next example we will explore some of the ways to search human data sets available in GN.

4.2. Human Case Study

In this example we will make use of a publicly available multi-level data set collected from a human cohort. As in the mouse case study, navigate to the **Select and Search** page and this time select **Species** = Human, and **Group** = Liver: Normal Gene Expression with Genotype (Merck). Clicking on the ***Info*** button will show that this

data set was originally published in 2008 [20] and then in 2010 [21] and was specifically used to examine gene expression and cytochrome P450 activity in human liver. Click on the **Type** dropdown menu to see the types of data that are available for this group. You will see that there are two data types available for this group. The Phenotypes data set (named as HLC Published Phenotypes) consists of phenotypes collected from this population that can be used for genetic mapping. Additionally, for some of the human cohorts including this particular group, the Phenotypes category can also include some individual level demographic data such as age, race, socio-economic status, etc. The other data type for this group is microarray gene expression data for the liver (Liver mRNA). Additionally, there is genotype data available for this cohort and users can perform basic genetic association analysis within GN using PLINK.

Using a simple workflow, we will demonstrate how functions in GN enable secondary analysis of published human data. We start out with basic demographic data—the age of subjects—and examine what we can learn about age-related gene expression changes in the liver.

1. Select **Type** = Phenotype and enter the wildcard symbols * or *?* in the ***Get Any*** search box. These wildcards will retrieve all records available for this cohort in the database. As of November 2015, there are 17 records in the Phenotype category for this group and include three demographic variables, twelve metabolic and physiologic traits, and two morphometric traits. Can you now use the ***Matrix*** and ***Graph*** tools that were described in the above mouse case study to inspect the correlation structure among these demographic variables and the different phenotypes (**see Note 7**)?
2. Click on the ***Record ID 10001*** (Demographics, age: Age [year]) to open the Trait **Data and Analysis** page for the age data. Notice that the layout of the page is similar to that of the expression traits described in the mouse case study, but without the **Resource Links** and probe tools that are relevant to gene expression traits. Examine the descriptive statistics and distribution profiles for this data using the **Basic Statistics** track. You will see that the mean age is about 50 years (±17 SD) and ranges from 1 to 94 years.
3. Given this wide range in sample age, we can now query if age is associated with differences in gene expression in the liver. Open the **Calculate Correlations track and** Select **Database** = GSE9588 Human Liver Normal (Mar11) Both Sexes. It is also possible to stratify the analysis by sex by choosing either the male or female expression data. For this example, we will retrieve the top 500 transcripts that have the highest correlation with age in both sexes. Select **Pearson** and click ***Compute***. The result of this analysis will be displayed in the **Correlation Table** page. The top of this page will display actions and tools as in the **Trait Collection** page (**Fig. 5**). The main correlation results are in the **Sample r** and **Sample p(r)** columns (Pearson correlations and p-values, respectively) (**Fig. 14A**). To access individual correlation plots, click on an *r* value and this will display a **Sample Correlation Scatterplot** with the trait on the X-axis (in this case, age) and the mRNA expression on the Y-axis (**Fig. 14B**). For this example, click on the correlation (*r* value) for the 12th transcript in the list (mitochondrial ribosomal protein L9, *MRPL9*) and we see that the expression of this mitochondrial ribosomal protein (MRP) gene is negatively correlated with age. You can customize the scatterplot by selecting ***Show Options*** in the **Sample Correlation Scatterplot** and setting your own preferences. For instance, in this example (**Fig. 14B**), the axes have been renamed from the default and the sample ID tag hidden.
4. The entire correlation results table can be exported by clicking on the ***Download Table*** button. Additionally, you can also select a set of records based on

correlation values using AND/OR operators by clicking the ***More Options*** button and setting the selection criteria. In the example in **Figure 14**, all transcripts that are negatively correlated with age are selected by setting the Pearson correlations to range between $r > -1.0$ AND $r < 0$ (**Fig. 14C**).

5. As described above, the ***GeneWeaver*** (http://ontologicaldiscovery.org), ***GCAT***, and ***Gene Set*** buttons at the top of the page allows users to seamlessly connect with other external bioinformatics tools for additional analyses. After selecting by correlation range, click the ***Gene Set*** tool to import your gene list from the GN correlation table directly to WebGestalt for GO enrichment analysis. This will reveal if the transcripts that are negatively correlated with age are enriched for any biologically relevant functions. Select ***View results*** and carefully examine the graph of enriched functional categories. The most enriched GO categories in this list of transcripts that are negatively correlated with age include mRNA metabolic process and ribonucleoprotein complex components (**Fig. 15A**). Now go back to the **Correlation Table** page that has the negatively correlated transcripts selected. From here, clicking the ***GCAT*** icon exports your selections as a gene list for a network analysis that examines imputed functional relatedness based on published abstracts and text mining (**Fig. 15B**). This quick analysis indicates that ribosomal genes are down-regulated in expression during aging. The negative correlation between *MRP* genes and age is striking, and members of this family of genes modulate aging and lifespan in mice and *C. elegans* [22].

   Now that we have performed a GO analysis of the transcripts that are negatively correlated with age, repeat the analysis above with transcripts that are positively correlated and demonstrate increased expression with age (**see Note 8**).

   While mapping functions in GN are better optimized for model organisms and standard test crosses, GN also provides an interface to PLINK for performing simple GWAS in humans. Below we conclude this case study with a demonstration of this mapping tool.

6. So far, we have used a wildcard search key to retrieve all the trait data available for the Merck liver cohort and examined gene expression changes associated with age. Now to perform a genetic association analysis using the phenotype data, open the **Trait Data and Analysis** page for ***Record ID 10015***. This is CYP2C8 enzymatic activity measured in 362 cases.
7. Using the **Basic Statistics** track, note that unlike the age data, which had a normal distribution, this phenotype has a highly skewed distribution. This phenotype provides an example in which the choice between Pearson and Spearman Rank in the **Calculate Correlation** section has a significant impact on the resulting list of correlated genes. First, perform a Pearson correlation and retrieve just the top 100 correlates from the ***GSE9588 Human Liver Normal (Mar11) Both Sexes*** data. Perform the same analysis but this time select the ***Spearman Rank*** option. Compare the two correlation tables. Note that while the top gene for the Pearson correlation is *TOMM40L* (*ID 10023831160*), the top transcript computed using Spearman rho is *CYP2C8* itself (*10033668843*). The scatter plots for the Pearson *r* and Spearman rho reveals why the Spearman rank correlation is better suited for this CYP2C8 enzymatic activity data and, from the Spearman correlation table, we find that CYP2C8 enzymatic activity is correlated with the expression of a number of other cytochrome P450 genes.
8. Now we test whether variation in CYP2C8 enzyme activity and *CYP2C8* expression share common genetic causes. From the **Trait Data and Analysis** page for record *ID 10015*, navigate to the **Mapping Tools** section. This tool provides a quick but basic interface to PLINK [23]. Note that you can set the

thresholds for the minor allele frequency and as well as the *p* value. The current version of this function in GN allows only the basic genetic association tests and users cannot set the threshold for Hardy-Weinberg equilibrium or include other covariates for population structure or demographic covariates in the association model. So use this tool with these caveats in mind (and compare with GN2 which does include some of these important functions). To initiate the genetic association test, click the ***Compute Using PLINK*** button and **Keep** outliers for this preliminary test. Perform the same analysis for *CYP2C8* expression (10033668843) to identify eQTLs.

9. The mapping result will be displayed as a Manhattan plot with chromosomal location on the X-axis and the $-\log_{10}(p)$ values on the Y-axis (**Fig. 16**). For the enzyme activity phenotype, the top significant association ($p < .0000001$) is with SNP rs6508937 chromosome 19. Clicking on the **SNP Name** (**rsID**) will take you to NCBI's ***dbSNP*** page for that particular SNP which will contain additional information on the type of variation, the ancestral allele, minor allele frequency, etc. For the expression trait, the most significant association is with SNP rs10964657 on Chr 9 ($p < .0001$). Surprisingly, in this case, the comparison of the two Manhattan plots does not flag any common SNPs and therefore does not provide support for the hypothesis that covariation in expression of transcripts and enzymes are due to shared genetic causes.

5. Future Directions and Conclusions

One of the main values of GN is its vast resource of data that enables both exploratory data-mining as well as specific hypothesis testing and cross-correlations between phenotypes at many scales. At the end of 2015, GN contained 578 systems genetics data sets for eight species and well over 70 different cell, tissue, and organ types making it a 160 GB database of genotypes and well-structured phenotypes. The amount of data in GN is growing rapidly: 255 datasets have been added in the past two years, compared to ~100 in the preceding decade. With this volume of data, search is a key feature for analysis and exploration. GN allows searching through genomic, genetic and phenotype data contained in the database. Users can then select multiple datasets and perform analysis on selected genes, traits and collections. The web-browser interface allows for interactive exploration of GN resources and the use of built-in analysis tools. This allows biomedical researchers to explore the data without training in more advanced bioinformatics programming languages, such as R and Python.

GN started out as a simple database and web site that was used primarily for analysis of mouse, rat, and human genes, chromosomes, and linked phenotypes. GN has now transformed into a service for on-line QTL mapping, eQTL analysis, and systems genetics. GN allows researchers to upload and store their own research data, run analyses—including QTL mapping, GWAS, and network analysis, generate publishable figures, compare results with those of other datasets, and explore relations between QTLs, genes, and phenotypes.

In this chapter we have highlighted the potential of GN by discussing built-in functionality and providing a few use cases. GN is an evolving service. The goals and challenges are to integrate new and sophisticated mapping and analysis features while maintaining an easy user interface in a structured environment and providing a powerful REST programming interface for power users. The new version of GN (GN2) will provide greater flexibility and additional features such as the use of generalized linear mixed models (LMMs), QTL mapping with covariates, and Weighted Gene Coexpression Analysis (WGCNA) [24]. These tools are already

available in the beta-release of the next generation of GN2 (Fig. **2**). There are many packages and web services available that can do individual components of a quantitative genetics or systems genetics analysis well, such as QTL mapping or data reduction and organization. However, there are no other resources that provide both a data repository and an integrated set of tools and services for systems genetics. Because GN and its environment consist of free and open source software, the whole system is easily installed and deployed locally allowing for coexistence of both a public data resource (the heart of GN) and local (private) data. It is even possible to rebrand the webserver and make it outward facing for new projects or institute.

6. Notes

1. **Select Species = Human, Group = All Tissues..., Type = Frontal Cortex mRNA**. Click on the ***Default*** button to lock-in these settings. Now review the **Quick HELP Examples and User's Guide**. The final query string should be entered into the **Combined** search. It should look like this: ***POSITION=(chr21 0 1000) MEAN=(4 1000)*** and should generate 28 hits. To focus on genes involved in Down syndrome, also known as Trisomy 21, add ***RIF=trisomy***. This will trim the set down to four hits.

2. **Select Species = Mouse, Group = BXD, Type = Liver Proteome, Data Set = EPFL/ETHZ BXD Liver, Chow Diet...** Click on the ***Default*** button to lock-in these settings. Review the **Quick HELP Examples and User's Guide**. The query string should be entered into the **Get Any** search. It should look like this: ***transLRS=(20 999 10)*** and should generate ~136 hits. This search will return all trans QTLs with an LRS between 20 and 999 using a 10 Mb window. Sort the results by the **Max LRS Location** column and look for patterns in the types of proteins that map to the same eQTL location; e.g. Chr 5 at about 127–128 Mb and Chr 10 at 107 Mb. These are potential trans regulatory regions.

3. Yet another way to visualize whole data sets and search for regulatory regions would be to select ***GenomeGraph*** from the ***Search*** tab in the banner menu. Select the "**EPFL…**" data set described above in **Note 2** and choose the ***Mapping*** option. This should generate a graph that shows genome location on the X-axis (each block is a chromosome) and position of the gene on the Y-axis. Each red cross represents a significant association at a false discovery rate (FDR) less than 0.2 (default is set to 0.2 or a FDR of 20%). Note the vertical bands (or trans bands) that indicate a number of significant associations on several chromosomes, including Chr 5. In contrast to trans eQTLs, cis eQTLs are indicated as a red cross on the diagonal (a significant association that corresponds to the location of the gene).

4. Check the function of *Cdkn1b* by selecting ***Gene Wiki*** from the dropdown menu under the ***Search*** tab in the banner menu and entering the gene name in the box and selecting ***submit***. Inspect the entries and then perform a search for the term "*addiction*". Again, the term addiction (entry 799) is used in an interesting way, "*Data indicate that the addiction of MYCC-amplified ovarian cancer cells to MYCC differs…*".

5. The probe set for *Kcnj3* appears to align far beyond (distal to) the known limits of the gene. To verify this, perform the RNA-seq BLAT alignment,

click on the ***browser*** link (far left), and then click on the ***zoom out 10x*** button twice. Note that the RNA-seq tracks (blue and red) show intense expression in the region well beyond the standard model 3' UTR. This is not unusual; 3' UTRs are often not well annotated. The probe set actually does target the gene, and does so at the distal part of the 3' UTR. Two of the probes overlap SNPs (736871 and 725381), and both are associated with strong cis eQTL artifacts (see ***Heat Map***).

6. To get to the ***Find*** tool you must navigate to the **Trait Data and Analysis** page for the gene (or probe set) of interest. Use this page or an existing **BXD Trait Collection** if active from the mouse case study. Alternatively, start over from the main search page by searching for *Comt* in most open BXD or even human lymphoblastoid and some aging brain expression data sets (**Groups** from Meyers and Liang). For *Comt*, the ***Find*** tool will return a number of results from four human data sets, four rat data sets and over 20 mouse data sets. For many of these data sets the expression of *Comt* is measured from multiple probes. For mouse and human data sets, the expression of each probe set appears to vary despite targeting the same gene (see **Mean Expr** or Mean Expression column). Note the large number of probes for data sets annotated with the term exon; exon-level microarrays have probes designed to target each feature of a transcript (UTRs, introns and exons). Explore each probe set for *Comt* by clicking on the Record ID for ***1418701_at*** and ***1449183_at*** using **Tissue = Hippocampus** and **Dataset= Hippocampus Consortium M430 (Jun06) RMA**. You will be redirected to the **Trait Data and Analysis** page for each record where you can compare the *Comt* transcript feature targeted by each probe set using the ***Verify*** or ***RNA-seq*** tool, explore the distribution of expression across BXD strains using the **Basic Statistics** track and **Probability Plot** and **Bar Graph (by rank)** options, and compare allelic effects and cis eQTL mapping using the **Mapping Tools** track and the **Interval** mapping option. You should see that probe sets targeting the distal end of the *Comt* transcript (the distal 3' UTR, probe set 1418701_at) have a very different pattern of expression across inbred strains of mice and the BXD panel when compared to probe sets that target coding exons or more proximal regions of the 3' UTR (probe set 1449183_at).

7 Start by selecting all 17 records from the **Search Results** page and ***Add*** to **Trait Collection**. From the **Trait Collection** page, select all 17 records and then click the ***Matrix*** tool. For a graphical visualization of the correlation among the different variables select the ***Graph*** tool. Try any of the network methods from the ***Select Graph Method*** and set the correlation to |0.25| with ***Pearson*** as **Correlation Type**. By examining the correlation matrix and the network graphs, you will learn that the various enzyme activity traits form a correlated network. While ethnicity and sex show no correlation with any of the traits, age is positively correlated with weight, which, as might be expected, has a strong positive correlation with BMI.

8 Transcripts from the **Correlation Table** that are positively correlated with age can be selected by setting the ***More Options*** track to $r > 0$ AND $r < 1.0$. Alternatively, a quicker way is to simply click the ***Invert Select*** option. Send this list of genes to WebGestalt by clicking the ***Gene Set*** tool. Note that the enrichment p values for this set of genes positively correlated with age are not as significant as for those that are negatively correlated with age.

Acknowledgment

We thank Lei Yan, Arthur Centeno, and Zachary Sloan, for their many contributions to building and maintaining GN over the past decade. GN code has benefited greatly from contributions by Jintao Wang, Sam Ockman, Xiaodong Zhou, Ning Liu, and Alex G. Williams, and Drs. Rudi Alberts, Arends, Elissa J. Chesler, Kenneth Manly, Danny and Evan G. Williams. Support for GeneNetwork has been provided by NIH grants U01AA013499, U01AA16662, U01AA014425, P20DA21131, U01CA105417, and U24 RR021760. GN is also generously supported by the UT Center for Integrative and Translational Genomics, and funds from the UT-ORNL Governor's Chair.

References

1. Manly KF, Olson JM (1999) Overview of QTL mapping software and introduction to map manager QT. Mamm Genome 10: 327-334.
2. Williams RW (1994) The Portable Dictionary of the Mouse Genome: a personal database for gene mapping and molecular biology. Mamm Genome 5: 372-375.
3. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, et al. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet 37: 233-242.
4. Andreux PA, Williams EG, Koutnikova H, Houtkooper RH, Champy MF, et al. (2012) Systems genetics of metabolism: the use of the BXD murine reference panel for multiscalar integration of traits. Cell 150: 1287-1299.
5. Chesler EJ, Wang J, Lu L, Qu Y, Manly KF, et al. (2003) Genetic correlates of gene expression in recombinant inbred strains: a relational model system to explore neurobehavioral phenotypes. Neuroinformatics 1: 343-357.
6. Wang J, Williams RW, Manly KF (2003) WebQTL: web-based complex trait analysis. Neuroinformatics 1: 299-308.
7. Li Z, Mulligan MK, Wang X, Miles MF, Lu L, et al. (2010) A transposon in Comt generates mRNA variants and causes widespread expression and behavioral differences among mice. PLoS One 5: e12181.
8. Williams EG, Mouchiroud L, Frochaux M, Pandey A, Andreux PA, et al. (2014) An evolutionarily conserved role for the aryl hydrocarbon receptor in the regulation of movement. PLoS Genet 10: e1004673.
9. Wang X PA, Mulligan MK, Williams EG, Mozhui K, et al. (2016) Joint mouse-human phenome-wide association to test gene function and disease risk. Nature Communications in press.
10. Chesler EJ, Lu L, Wang J, Williams RW, Manly KF (2004) WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. Nat Neurosci 7: 485-486.
11. Peirce JL, Lu L, Gu J, Silver LM, Williams RW (2004) A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC Genet 5: 7.
12. Taylor BA, Heiniger HJ, Meier H (1973) Genetic analysis of resistance to cadmium-induced testicular damage in mice. Proc Soc Exp Biol Med 143: 629-633.
13. Alberts R, Schughart K (2010) QTLminer: identifying genes regulating quantitative traits. BMC Bioinformatics 11: 516.
14. Overall RW, Kempermann G, Peirce J, Lu L, Goldowitz D, et al. (2009) Genetics of the hippocampal transcriptome in mouse: a systematic survey and online neurogenomics resource. Front Neurosci 3: 55.
15. Wang XS, Agarwala R, Capra JA, Chen ZG, Church DM, et al. (2010) High-throughput sequencing of the DBA/2J mouse genome. Bmc Bioinformatics 11.
16. Bottomly D, Walter NA, Hunter JE, Darakjian P, Kawane S, et al. (2011) Evaluating

gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PLoS One 6: e17820.
17. Ciobanu DC, Lu L, Mozhui K, Wang X, Jagalur M, et al. (2010) Detection, validation, and downstream analysis of allelic variation in gene expression. Genetics 184: 119-128.
18. Homayouni R, Heinrich K, Wei L, Berry MW (2005) Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinformatics 21: 104-115.
19. Williams RW, Mulligan MK (2012) Genetic and molecular network analysis of behavior. Int Rev Neurobiol 104: 135-157.
20. Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6: e107.
21. Yang X, Zhang B, Molony C, Chudin E, Hao K, et al. (2010) Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. Genome Res 20: 1020-1036.
22. Houtkooper RH, Mouchiroud L, Ryu D, Moullan N, Katsyuba E, et al. (2013) Mitonuclear protein imbalance as a conserved longevity mechanism. Nature 497: 451-457.
23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
24. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.

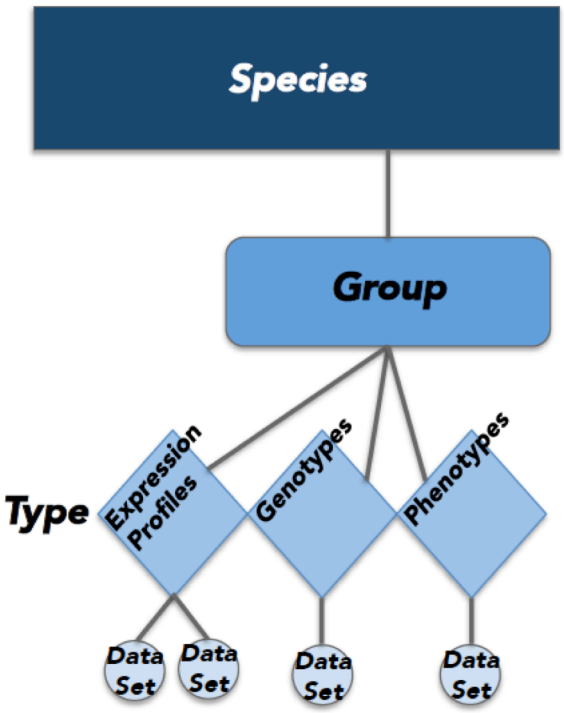


Figure 1. Organization of data sets in GeneNetwork

	Group	Genotypes	Molecular Traits	Higher Order Phenoype Traits	Description and Usage	Reference
Human	All Tissue, RNA-Seq GTEx v5	Yes	Expression profiles (RNA-seq) from 30 peripheral tissues and 11 brain subregions	No	Massive collaborative effort to explore associations between genotype and gene expression across tissues collected from up to 1,000 individuals. Suitable for quantitative genetics (QTL mapping) and systems genetics, including correlation and network analysis to compare associations between tissues and between other human or rodent data sets.	The GTEx Consortium, Science (2015)
	Brain, Aging: AD, Normal Gene Expression with genotypes (Myers)	Yes	Expression profiles (Agilent microarray) from cerebellum, prefrontal cortex, and primary visual cortex	No	A study of cortical gene expression for normal aged and Alzheimer's disease cases with ~176 cases and 187 controls per tissue. Suitable for quantitative genetics (QTL mapping) and systems genetics, including correlational and network analysis to compare associations between tissues and disease state or between other human or rodent data sets.	Webster J.A. et al., Am J of Human Gen. (2009)
	Liver: Normal Gene Expression with Genotypes (Merck)	Yes	Expression profiles (Rosetta/Merck Human 44K 1.1 microarray) from liver	Metabolic traits	Gene expression profiles from 427 human liver samples that includes measurements of activity for nine enzymes. Suitable for quantitative genetics (QTL mapping) and systems genetics, including correlational and network analysis to compare associations between tissues and disease state or between other human or rodent data sets.	Schadt E.E. et al., Plos Biology (2008)
	Brain: Normal Gene Expression (NIH/Gibbs)	No	Expression profiles (Illumina humanRef-8 v2.0 expression beadchip microarray) from cerebellum, caudal pons, frontal cortex, and temporal cortex	No	Study that includes brain expression profiles from 147 individuals. Suitable for correlational and network analysis to compare associations between tissues and between other human or rodent data sets.	Gibbs J.R. et al., Plos Genetics (2010)
	Brain, Aging: AD, HD, Normal Gene Expression (Harvard/Merk)	No	Expression profiles (Agilent microarray) from cerebellum, prefrontal cortex, and primary visual cortex	No	A study that includes brain expression profiles from 307 Alzheimers disease cases, 152 Huntington's Disease cases and 132 controls. Suitable for correlational and network analysis to compare associations between tissues and disease state and between other human or rodent data sets. Tissues provided by the Harvard Brain Tissue Resource Center (www.brainbank.mclean.org).	Zhang B. et al., Cell (2013)
	Brain, Aging: AD, Normal Gene Expression (Liang)	No	Expression profiles (Affymetrix Human genome U133 Plus 2.0 microarray) from entorhinal cortex, hippocampus, medial temporal gyrus, posterior cingulate cortex, superior frontal gyrus, primary visual cortex	No	A survey of gene expression across six brain regions for normal aged and Alzheimer's disease cases with ~ 14 biological replicates per tissue and condition. Suitable for correlational and network analysis to compare associations between tissues and disease state or between other human or rodent data sets.	Liang W.S. et al., PNAS (2008)
Mouse	BXD	Yes	Expression profiles for peripheral tissue (adipose, adrenal gland, bone, cartilage, eye and retina, gastrointestinal tract, kidney, liver, lung, muscle, and spleen), brain tissue (whole brain, amygdala, cerebellum, hippocampus, hypothalamus, midbrain, neocortex, nucleus accumbens, pituitary, prefrontal cortex, and striatum), and cell type (hematopoietic cells, hepatocytes, hippocampal precursor cells, and T-cells) measured on multiple microarray platforms and using RNA-sequencing. Some proteome data from liver is also available.	Behavioral, Metabolic, Morphological, Pharmacological, Toxicology	Recombinant inbred genetic reference population (GRP) derived by crossing a C57BL/6J (B) female with a DBA/2J (D) male. The BXD set was derived from three seperate crosses of B and D parental strains in early 1970's, late 1990's, and early 2000's. Data collection is part of a massive collaborative effort from multiple investigators. Suitable for quantitative genetics (QTL mapping) and systems genetics, including correlation and network analysis to compare associations between tissues and between other rodent or human data sets.	Peirce J.L. et al., Genetics (2004)
	Mouse Diversity Panel	Yes	Expression profiles for bone, dorsal root ganglion, hippocampus, and liver measured using microarray platforms.	Behavioral, Metabolic, Morphological, Pharmacological, and Toxicology	The Mouse Diversity Panel (MDP) is represented by multiple and genetically divergent inbred strains. This panel has a higher recombination rate, level of genetic variation, and phenotypic diversity than crosses derived from two parental inbred strains but demonstrates significant population structure. Suitable for quantitative genetics (QTL mapping) and systems genetics, including correlation and network analysis to compare associations between tissues and between other rodent or human data sets.	McClurg P. et al., Genetics (2007)
	BHF2 (Apoe Null) UCLA	Yes	Expression profiles for adipose, brain, liver and muscle measured using the agilent microarray platform.	Metabolic	This data set features a large F2 cross derived from C57BL/6J and C3H/HeJ (BHF2) of 334 individuals. Both inbred progenitors were null for ApoE resulting in a population of genetically diverse F2 individuals that lack ApoE. Loss of this gene recapitulates some of the phenotypes associated with metabolic syndrome. The F2 population was fed a high fat diet from 8 to 24 weeks of age. Suitable for quantitative genetics (QTL mapping) and systems genetics of metabolism.	Wang S. et al., Plos Genetics (2006)
	Heterogeneous Stock	Yes	Expression profiles for hippocampus, liver, and lung using the Illumina Mouse WG-6 v1, v1.1 microarray platform.	Morphological	Heterogeneous Stock (HS) mice are derived from eight different inbred strains (A/J, AKR/J, BALBc/J, CBA/J, C3H/HeJ, C57BL/6J, DBA/2J, and LP/J). This panel has a higher recombination rate, level of genetic variation, and phenotypic diversity than crosses derived from two parental inbred strains but can demonstrate significant population structure. Suitable for high resolution quantitative genetics (QTL mapping) and systems genetics, including correlation and network analysis to compare associations between tissues and between other rodent or human data sets.	Huang G.J. et al., Genome Res (2009)

Table 1. A sample of well characterized human and mouse data sets. Many of the **Data Sets** are amenable to systems genetics mapping and other methods and are accessible at GeneNetwork. The **Description and Usage** column provides details about the data set and potential usage. Note that only the first three human data sets have both genotype and gene expression data and only the third data set features genotypes, gene expression, and higher order trait data in the form of metabolic phenotypes.

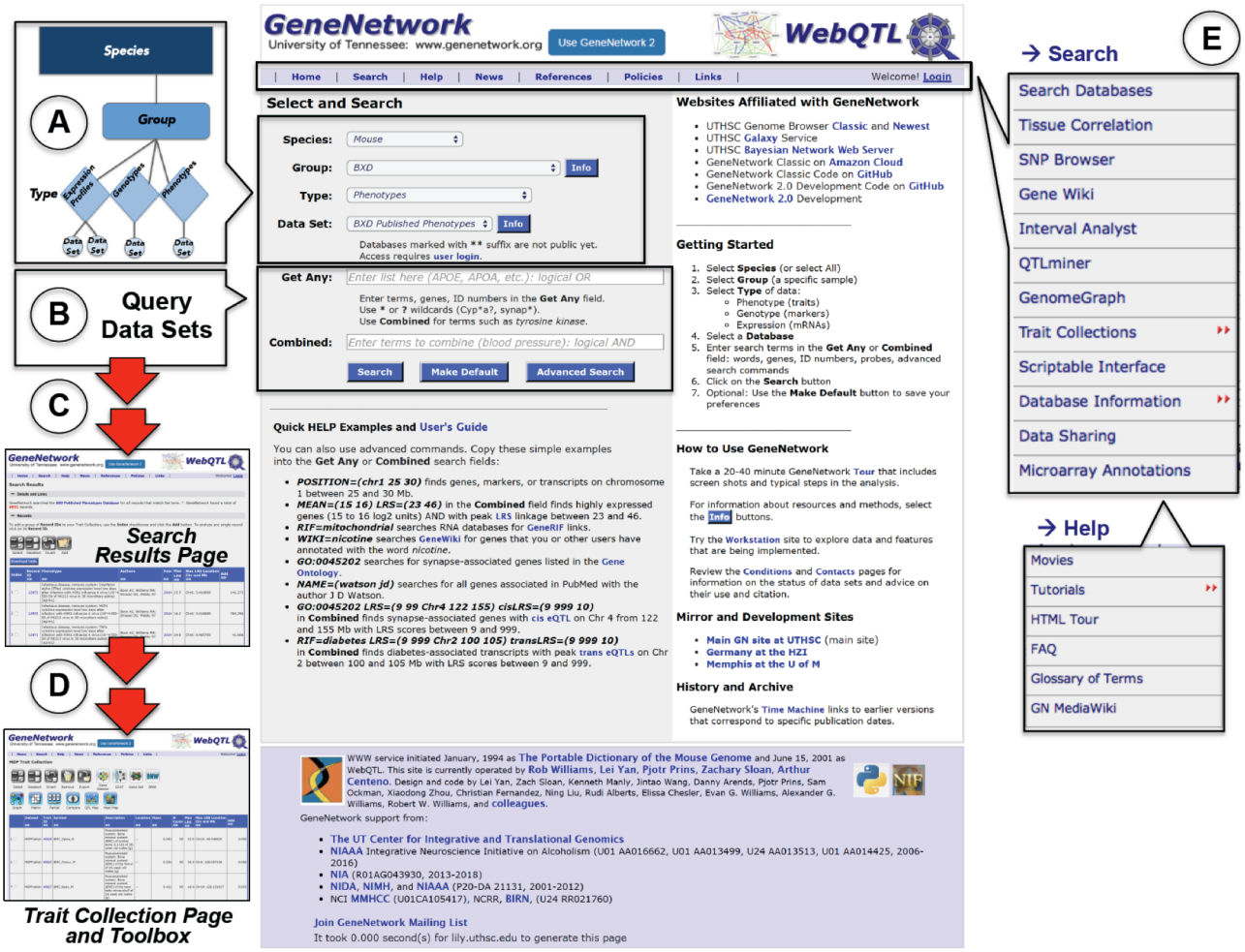


Figure 2. GeneNetwork main search page and organization. Most analyses in GeneNetwork will follow the steps shown in panels A through D. In this workflow, a data set is selected (A) and mined for traits of interest based on user search queries (B). Traits are then selected from the search (C) and placed in a collection for further inspection and quantitative analysis (D). The banner menu contains additional search options and helpful resources under the ***Search*** and ***Help*** tab, respectively (E).

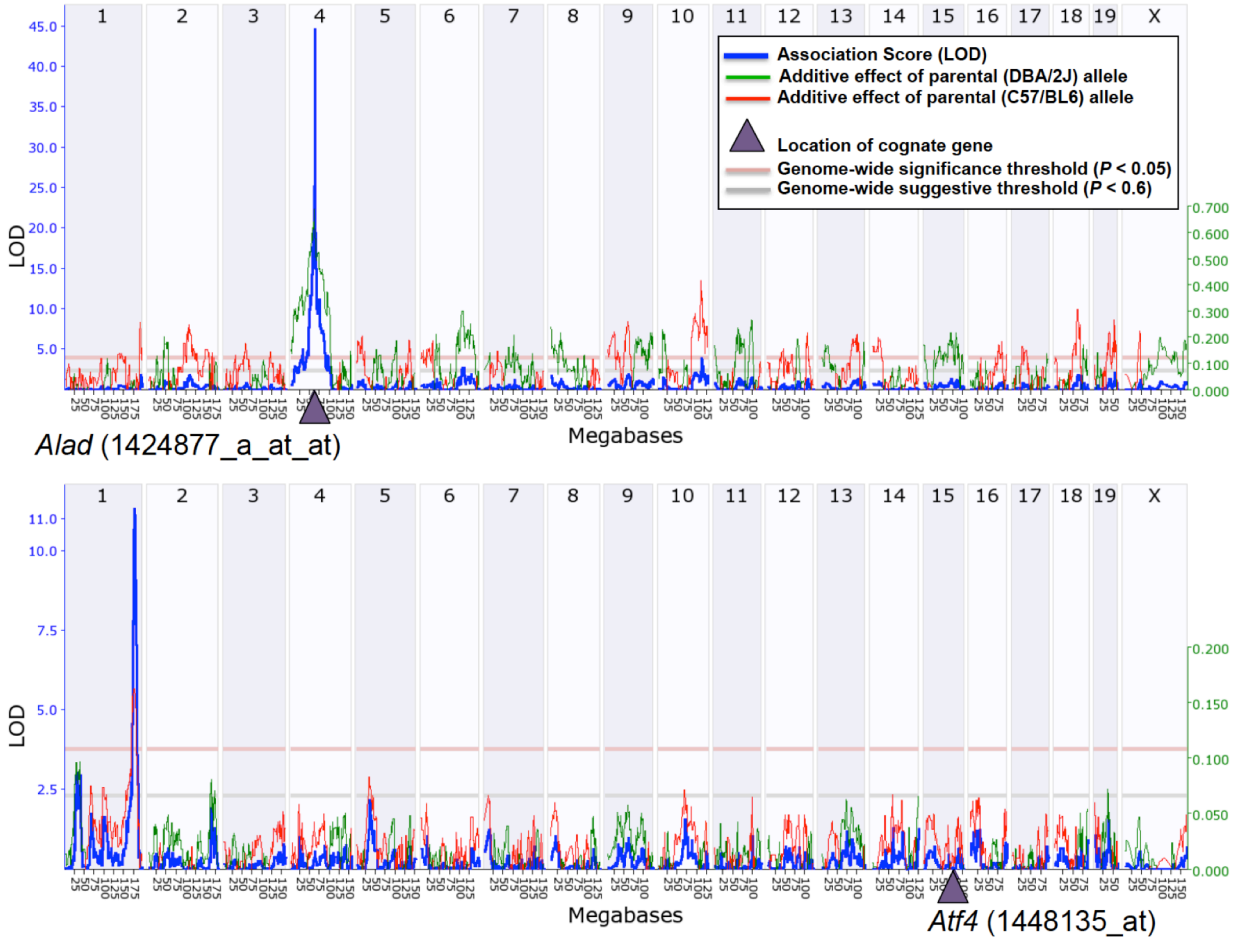


Figure 3. Local or distant modulation of gene expression in the hippocampus of BXD strains. QTL maps are shown for *Alad* and *Atf4* in the top and bottom panels with the association score (LOD) plotted on the Y axis across the genome (X-axis). Chromosomes and megabase position are shown at the top and bottom of the graph, respectively. Expression of *Alad* is modulated by a local cis eQTL whereas expression of *Atf4* is modulated by a distant trans eQTL. The sequence variant underlying expression of *Alad* is actually a copy number variant such that the parental DBA/2J strain and BXD strains that have inherited the *D allele* at this locus have additional copies of the gene and higher expression (indicated by the green line associated with the QTL peak in blue). The expression of Atf4 is modulated from a distal region on Chr 1. BXD strains that have inherited the B allele from the C57BL/6J parent at the Chr 1 locus have higher expression of Atf4. This distal region on Chr 1 (often referred to as QTL rich region 1 or *QRR1*) is a major regulatory locus of many expression and behavioral traits. The additive effect is shown in green to the right. The expression data can be accessed using **Mouse** *Species*: **Mouse,** *Group:* **BXD Phenotypes,** *Type*: **BXD** Data Set**: Hippocampus Consortium M430v2 (Jun06) RMA** and entering the probe set IDs in the **Get Any** search option.

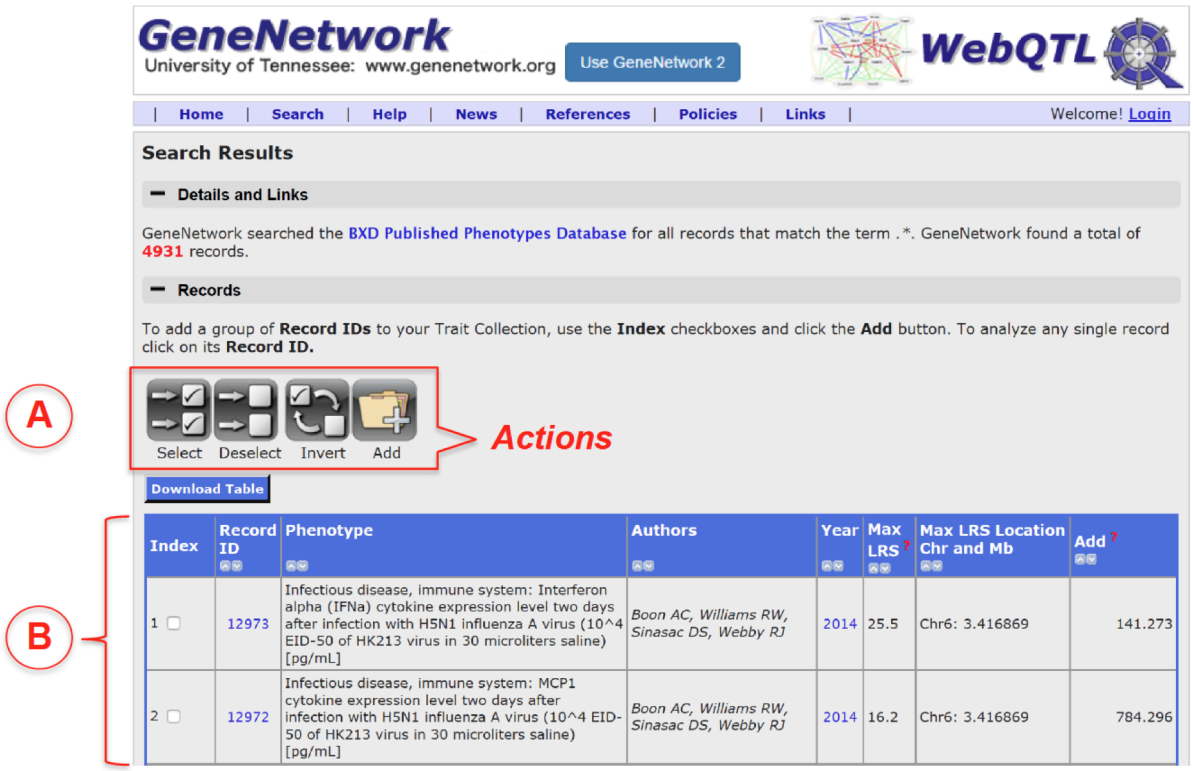


Figure 4. Overview of Search Results page. Panel A indicates actions and panel B shows indexed search results. Number of records that match search term are shown in the **Details and Links** section at the top of the page. Note that this page was generated using the **Mouse (*Species*), BXD (*Group*) Phenotypes (*Type*) BXD Published Phenotypes** Data **Set** and entering the wild card character (*) using the **Get Any** option. Summarized information for each trait varies based on data set type but, in general, **Record ID** gives a unique identifier for each data set, (e.g. a number for phenotype data sets and a probe set identifier for expression data sets), **Max LRS** and **MAX LRS Location Chr and MB** give the maximum association score for each trait, and associated peak chromosome and megabase position, respectively. **Add** gives the additive allele effect, which is the estimated effect on trait expression associated with inheritance of the maternal or paternal allele. Positive or negative values indicate higher or lower expression associated with inheritance of the paternal or maternal allele, respectively. From the **Search Results** page additional information about individual traits can be accessed by clicking the **Record ID.** Multiple traits can be selected (or deselected) using the actions options ***Select***, ***Deselect***, and ***Invert***. Selected traits can be added to a **Trait Collection** for further analysis using the ***Add*** option. The red question marks are links to additional information about column headings.

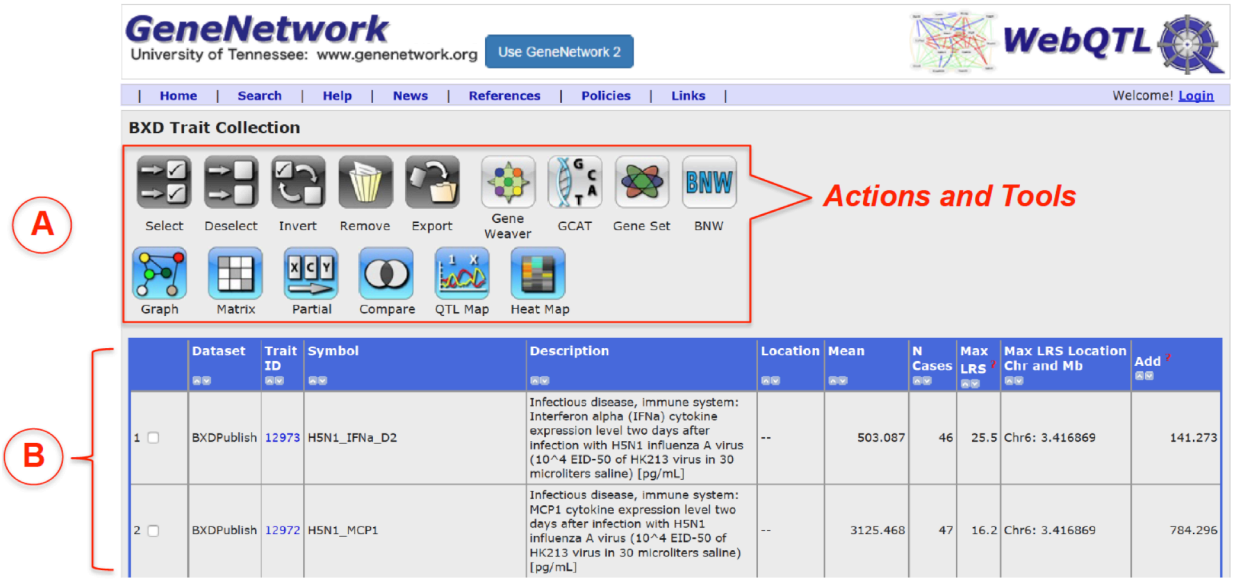


Figure 5. Overview of the Trait Collection page. Panel A shows the actions tools menu with each action or tool represented by a clickable icon. Panel B shows the indexed search results. Note that additional columns of data are shown for traits in a collection compared to traits in the **Search Results** page, including **Dataset**, **Symbol**, **Description**, **Location**, **Mean**, and **N Cases**. The **Dataset** and **Description** column provide information about which data set the trait originated from and details about the trait itself. As multiple different types of data can be added to the same **Group** collection it is useful to keep track of which data set the trait originated from, especially if exploring the expression oft he same gene across tissue types. For phenotype data sets, detailed descriptions are provided about trait measurement and for gene expression data sets, the full gene name is given along with information about the probe set used to measure the expression of that gene. The **Symbol** column gives the gene symbol for expression data sets and an abbreviated name for phenotypes. **Location** and **Mean** give the location of the gene for expression data sets and average trait expression, respectively. **N Cases** shows the number of individuals that were included in the trait measurement. The red question marks are links to additional information about column headings.

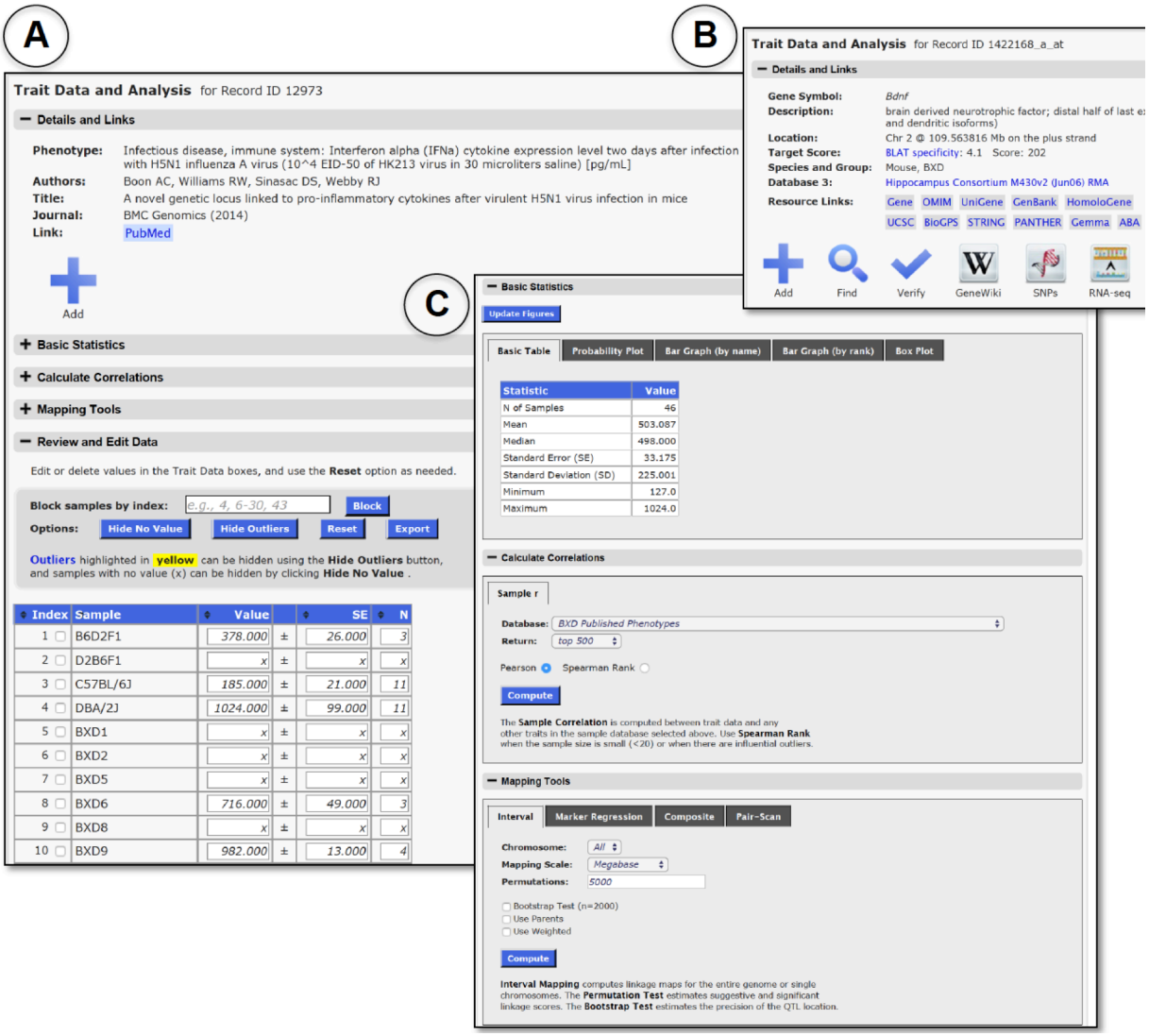


Figure 6. Layout of Trait Data and Analysis page. Users can explore individual traits in detail in the **Trait Data and Analysis** page. In the **Details and Links** track, a full description of the trait and associated actions and tools are shown. Actions and tools vary slightly depending on whether the trait is from a phenotype (A) or gene expression (B) ***Data Set***. The results in B can be generated by selecting **Mouse (***Species***), BXD (***Group***)**, **Hippocampus mRNA (***Type***), Hippocampus Consortium M430v2 (Jun06) RMA (**Data Set) and entering the gene symbol "*Bdnf*" using the **Get Any** option. Multiple links to outside resources (shown as **Resource Links**) are provided for gene expression data in addition to the GeneNetwork actions and tools ***Add***, ***Find***, ***Verify***, ***GeneWiki***, ***SNPs***, ***RNA-seq***, and ***Probes***. Both traits have a common set of tools shown in Panel C as the **Basic Statistics**, **Calculate**, **Correlations**, and **Mapping Tools** tracks. Each track gives the user options to graph the trait distribution, correlate expression of the trait with all other traits in a ***Data Set*** from the same ***Group***, or perform QTL mapping for the trait, respectively. Actual trait values are shown in the **Review and Edit Data** track.

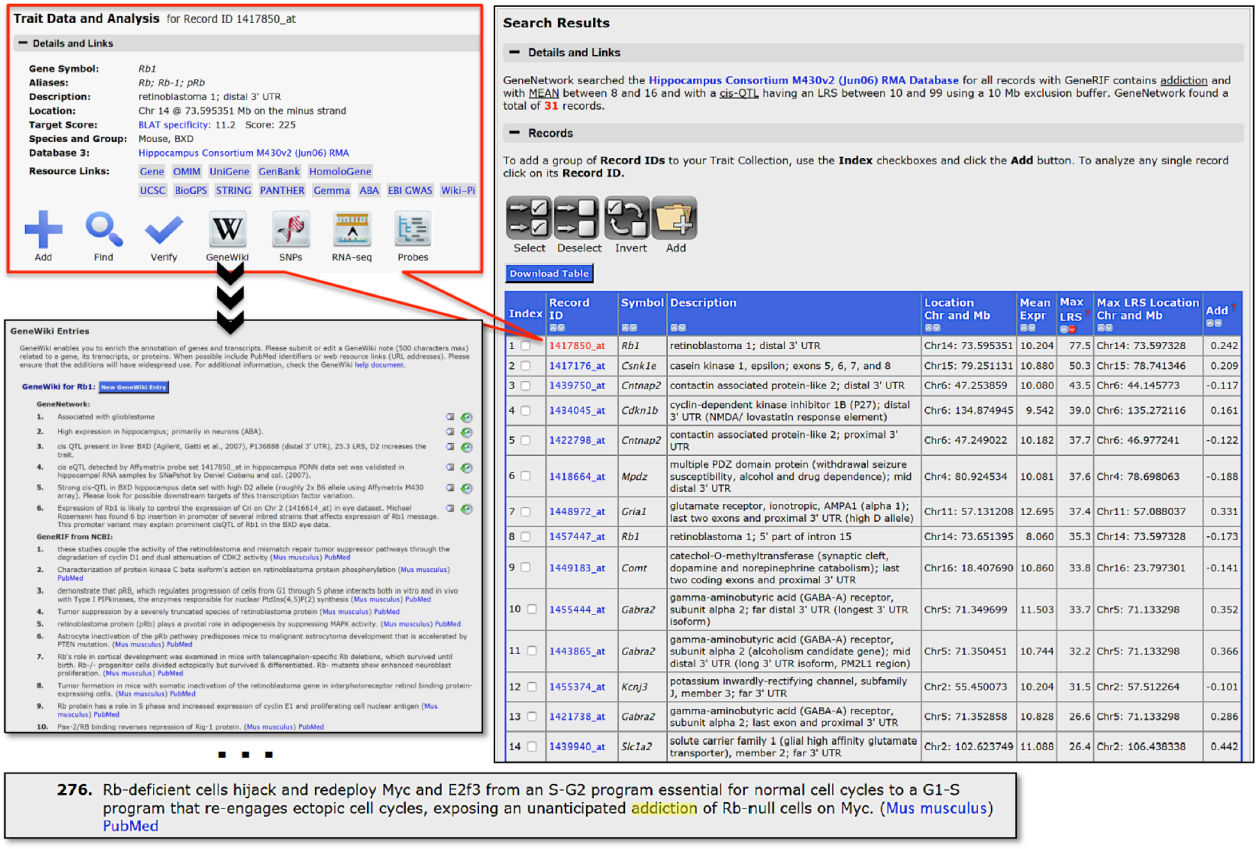


Figure 7. Exploring the function of *Rb1*. An unusual use of the term addiction in NCBI GeneRIF lead to the inclusion of *Rb1* in our search for addiction related genes whose expression is modulated by a strong cis eQTL.

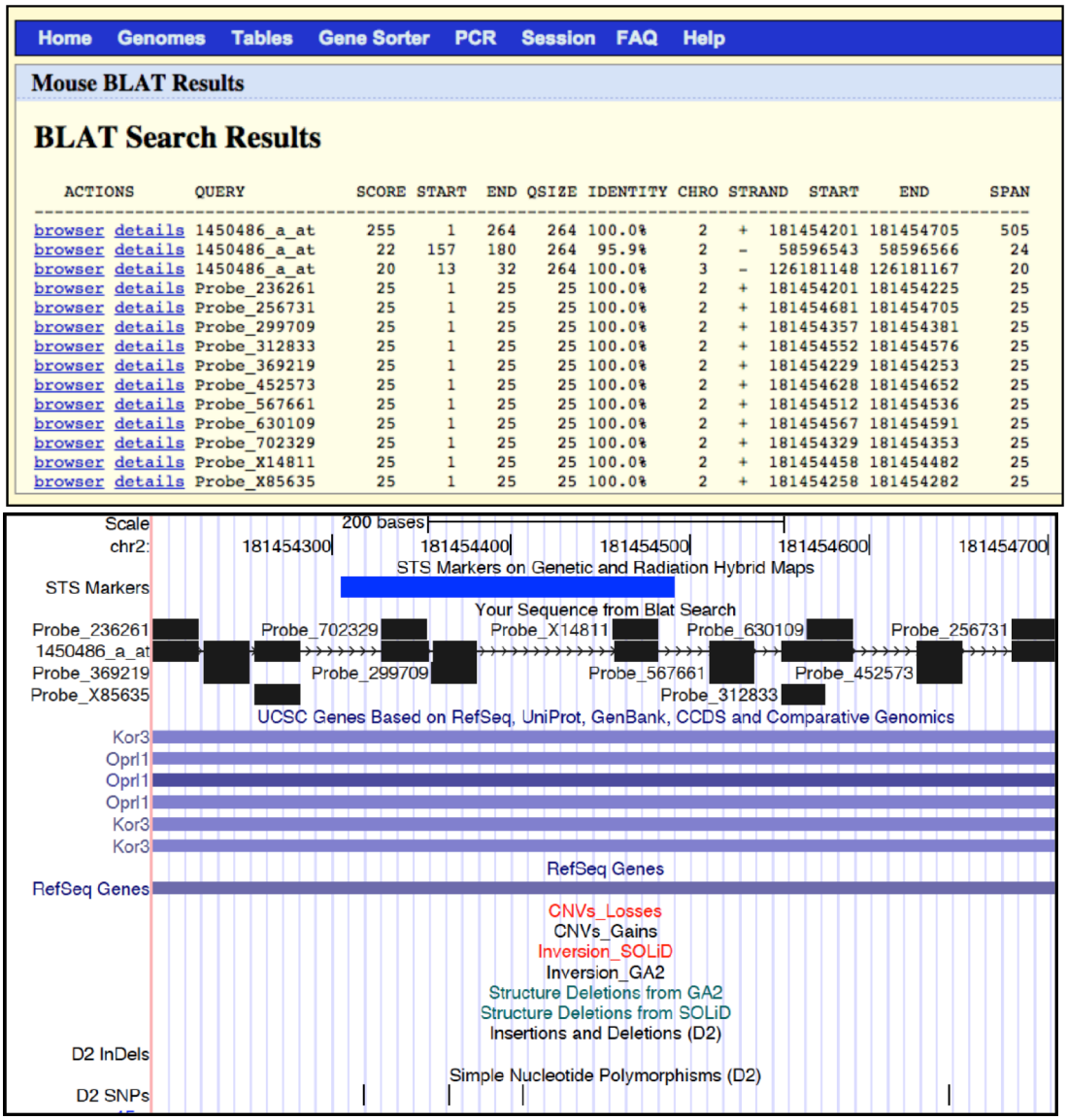
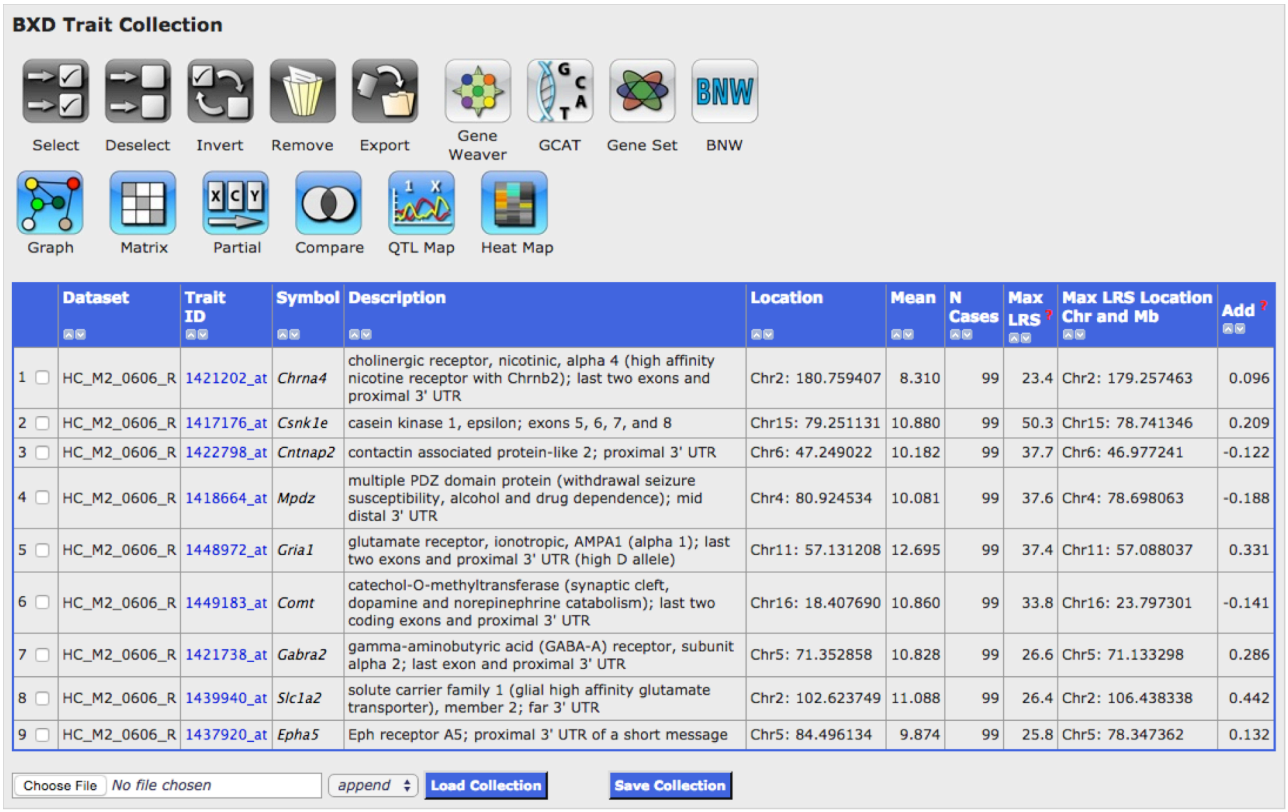


Figure 8. Probe set quality control. The RNA-seq button performs alignment of a probe set sequence against the appropriate reference genome using UCSC Genome Browser's BLAST-like alignment tool (BLAT). The results are shown for probe set 1450486_a_at in the top panel. The SCORE is a function of the size and match. For large sequences a perfect score is 255. START, END and QSIZE provide information about the size in base pairs of the query sequence. IDENTITY provides information about the match with 100% indicating a perfect match to the reference C57BL/6J genome. The location and span of the match are given by CHRO (chromosome) STRAND, START, END, and SPAN. Note that both the probe set and the 11 perfect match probes that comprise the probe set are shown and that the best match for the individual probes and entire probe set is on the positive strand on Chr 2 around 181.45 Mb. Clicking the browser link for the best match directs to a graphical display of the probe set alignment, shown in the bottom panel. The genome browser display can be cluttered fort he uninitiated. The basic layout is a display of several different *Tracks* of information. These tracks can be modified by scrolling down to the track tables at the bottom of the page. The display in the above panel was generated by selecting the hide option for all tracks EXCEPT the **Mapping and Sequencing**, **Genes and Gene Prediction**, and the **DBA/2J Sequence and Structural Variation** tracks. The position of all 11 probes and the composite probe set are shown in the bottom panel in black with the corresponding IDs shown to the left. The arrowheads designate the alignment oft he probe set on the positive (or sense) strand. The targeted gene (*Oprl1*) is shown below and indicates that the probe set is designed to target the 3' UTR according to the UCSC gene model. The location of sequence variants in the DBA/2J strain relative tot he C57BL/6J reference genome are shown in the last two tracks (D2 Indels and D2 SNPs). Note probes 299709 and 452573 overlap a DBA/2J SNP.

Figure 9. Impact of variants overlapping probe sets in microarray data sets. SNPs overlapping *Oprl1* probeset 1450486_a_at (perfect match or PM probes 299709 and 452573) lead to expression measurements that are higher in BXD strains that have inherited the *B allele* and lower in strains that have inherited the *D allele*. The **QTL Heatmap** reveals a strong eQTL with higher expression associated with inheritance of the B allele at the *Oprl1* locus (blue) only for the probes that overlap SNPs. The arrowhead indicates the genomic position of the probes. No other probes demonstrate a strong association between inheritance of alleles at this locus and gene expression. This analysis reveals that the strong cis eQTL detected for *Oprl1* is actually the result of a technical artifact resulting from sequence variants that disrupt the hybridization of probes to their target RNA sequence in strains other than the reference B6 strain (in this case the D2 strain).



	Dataset	Trait ID	Symbol	Description	Location	Mean	N Cases	Max LRS	Max LRS Location Chr and Mb	Add
1	HC_M2_0606_R	1421202_at	*Chrna4*	cholinergic receptor, nicotinic, alpha 4 (high affinity nicotine receptor with Chrnb2); last two exons and proximal 3' UTR	Chr2: 180.759407	8.310	99	23.4	Chr2: 179.257463	0.096
2	HC_M2_0606_R	1417176_at	*Csnk1e*	casein kinase 1, epsilon; exons 5, 6, 7, and 8	Chr15: 79.251131	10.880	99	50.3	Chr15: 78.741346	0.209
3	HC_M2_0606_R	1422798_at	*Cntnap2*	contactin associated protein-like 2; proximal 3' UTR	Chr6: 47.249022	10.182	99	37.7	Chr6: 46.977241	-0.122
4	HC_M2_0606_R	1418664_at	*Mpdz*	multiple PDZ domain protein (withdrawal seizure susceptibility, alcohol and drug dependence); mid distal 3' UTR	Chr4: 80.924534	10.081	99	37.6	Chr4: 78.698063	-0.188
5	HC_M2_0606_R	1448972_at	*Gria1*	glutamate receptor, ionotropic, AMPA1 (alpha 1); last two exons and proximal 3' UTR (high D allele)	Chr11: 57.131208	12.695	99	37.4	Chr11: 57.088037	0.331
6	HC_M2_0606_R	1449183_at	*Comt*	catechol-O-methyltransferase (synaptic cleft, dopamine and norepinephrine catabolism); last two coding exons and proximal 3' UTR	Chr16: 18.407690	10.860	99	33.8	Chr16: 23.797301	-0.141
7	HC_M2_0606_R	1421738_at	*Gabra2*	gamma-aminobutyric acid (GABA-A) receptor, subunit alpha 2; last exon and proximal 3' UTR	Chr5: 71.352858	10.828	99	26.6	Chr5: 71.133298	0.286
8	HC_M2_0606_R	1439940_at	*Slc1a2*	solute carrier family 1 (glial high affinity glutamate transporter), member 2; far 3' UTR	Chr2: 102.623749	11.088	99	26.4	Chr2: 106.438338	0.442
9	HC_M2_0606_R	1437920_at	*Epha5*	Eph receptor A5; proximal 3' UTR of a short message	Chr5: 84.496134	9.874	99	25.8	Chr5: 78.347362	0.132

Figure 10. Top cis modulated genes associated with addiction

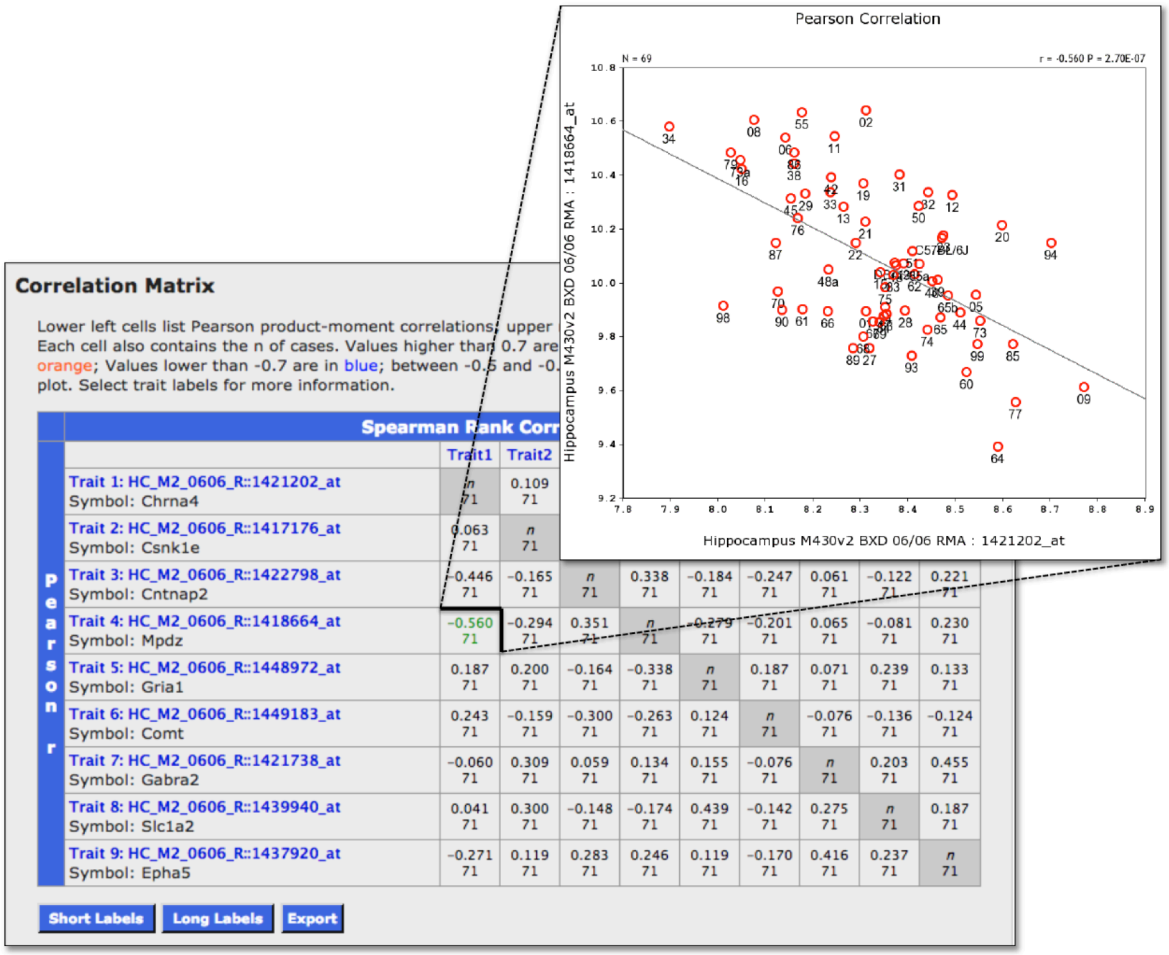


Figure 11. Exploring covariation. The matrix function allows users to investigate covariation between genes (or probe sets) in the **Trait Collection**. To display the gene symbols along with the probe set IDs, use the ***Short Labels*** button to redraw the correlation matrix. The matrix displays the correlation for each pair of genes (or probe sets) with the spearman correlation coefficient shown to the right of the diagonal and the Pearson Correlation Coefficient shown to the left (the diagonal is indicated by grey shading and would normally be represented as a 1, or the correlation of each probe set with itself). Scatterplots can be generated by clicking the correlation in the matrix. The scatterplot can be customized by selecting the ***Show Options*** icon, adjusting the settings, and replotting.

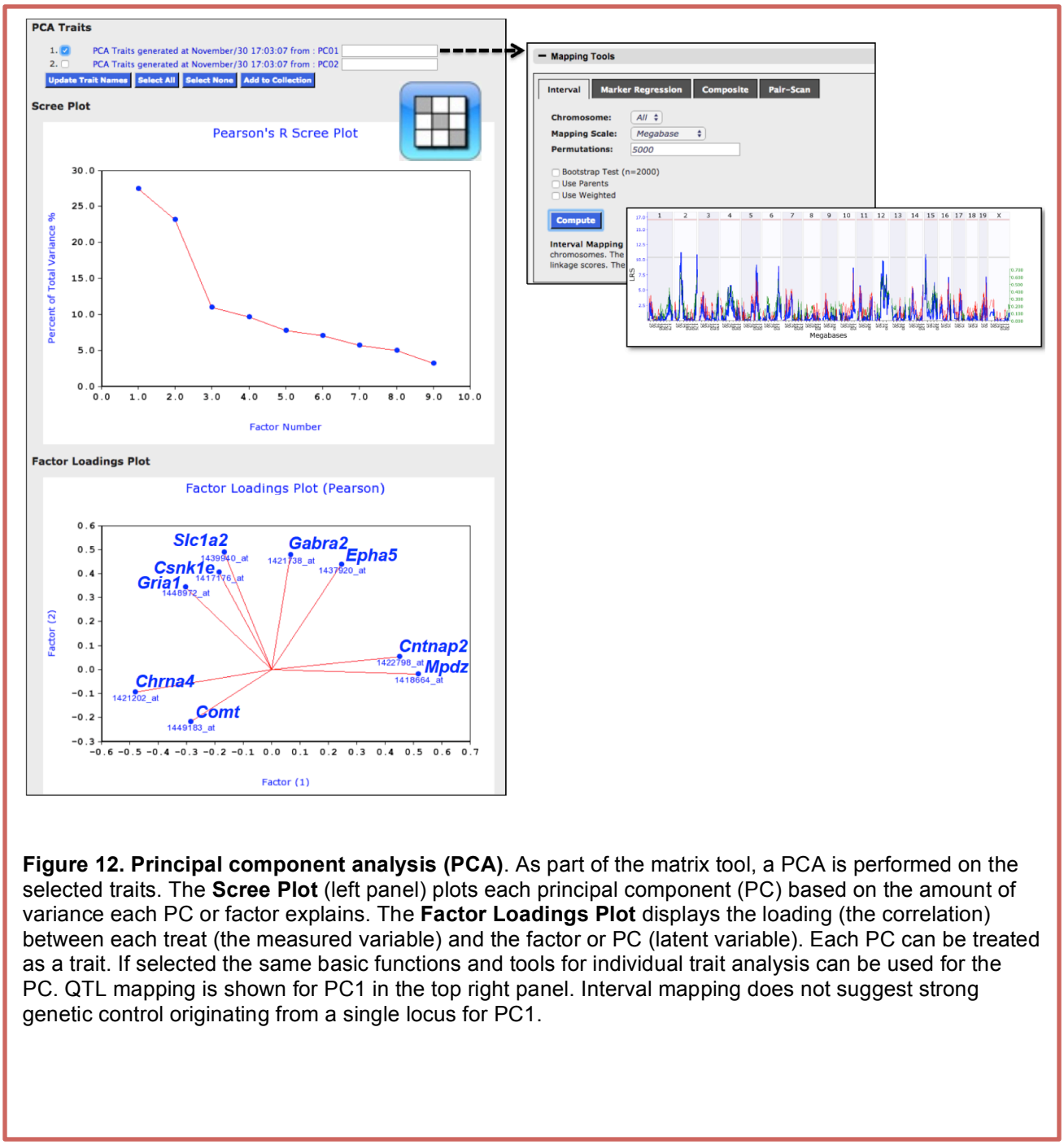


Figure 12. Principal component analysis (PCA). As part of the matrix tool, a PCA is performed on the selected traits. The **Scree Plot** (left panel) plots each principal component (PC) based on the amount of variance each PC or factor explains. The **Factor Loadings Plot** displays the loading (the correlation) between each treat (the measured variable) and the factor or PC (latent variable). Each PC can be treated as a trait. If selected the same basic functions and tools for individual trait analysis can be used for the PC. QTL mapping is shown for PC1 in the top right panel. Interval mapping does not suggest strong genetic control originating from a single locus for PC1.

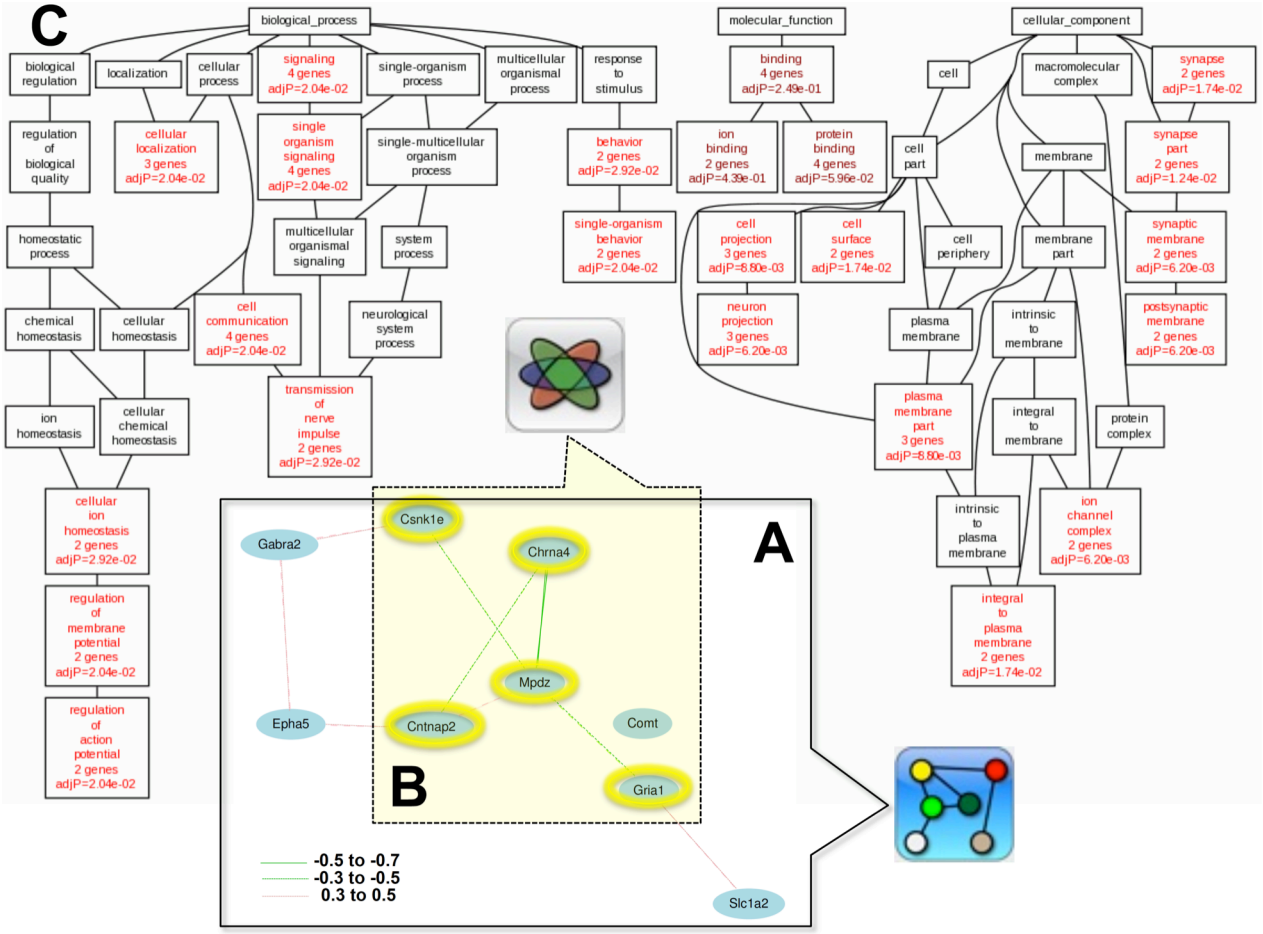


Figure 13. Creating networks and analysis of biological enrichment. From the **Trait Collection** a network graph depicting relations between gene set members can be constructed using the ***Graph*** tool. Display and correlation threshold can be adjusted using the Network Graph interface. Each node represents a gene (probe set) and the edge indicates the correlation (green for negative correlations and red for positive correlations). In this case the network shown in A was given a threshold of $r = |0.3|$ as this represents a significant correlation ($p < -0.01$) in this data set. Based on the network, a subset of genes (shown in the yellow panel in B) can be selected for enrichment analysis. Select the subset in the **Trait Collection** and select the ***Gene Set*** tool. Enrichment analysis is shown in the background (C), with significant (adjusted p-value or AdjP < 0.05) enrichment of biological function (based on GO annotations) shown in red.

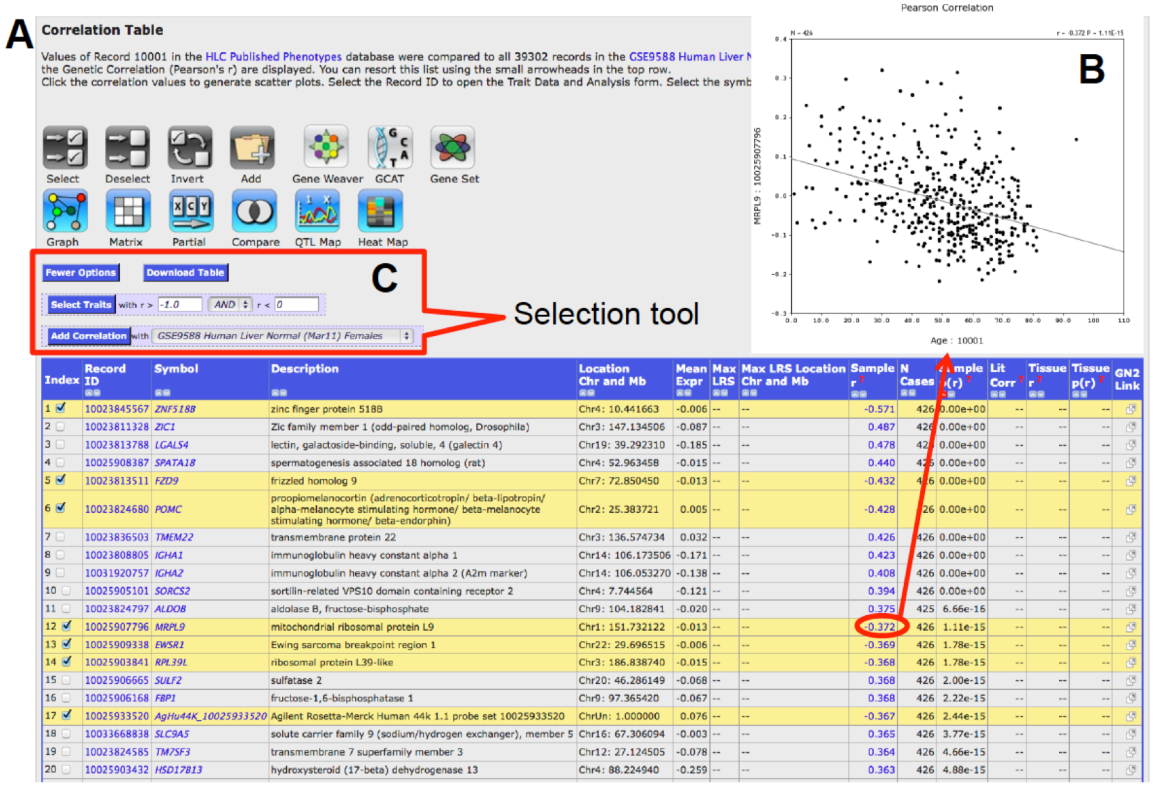
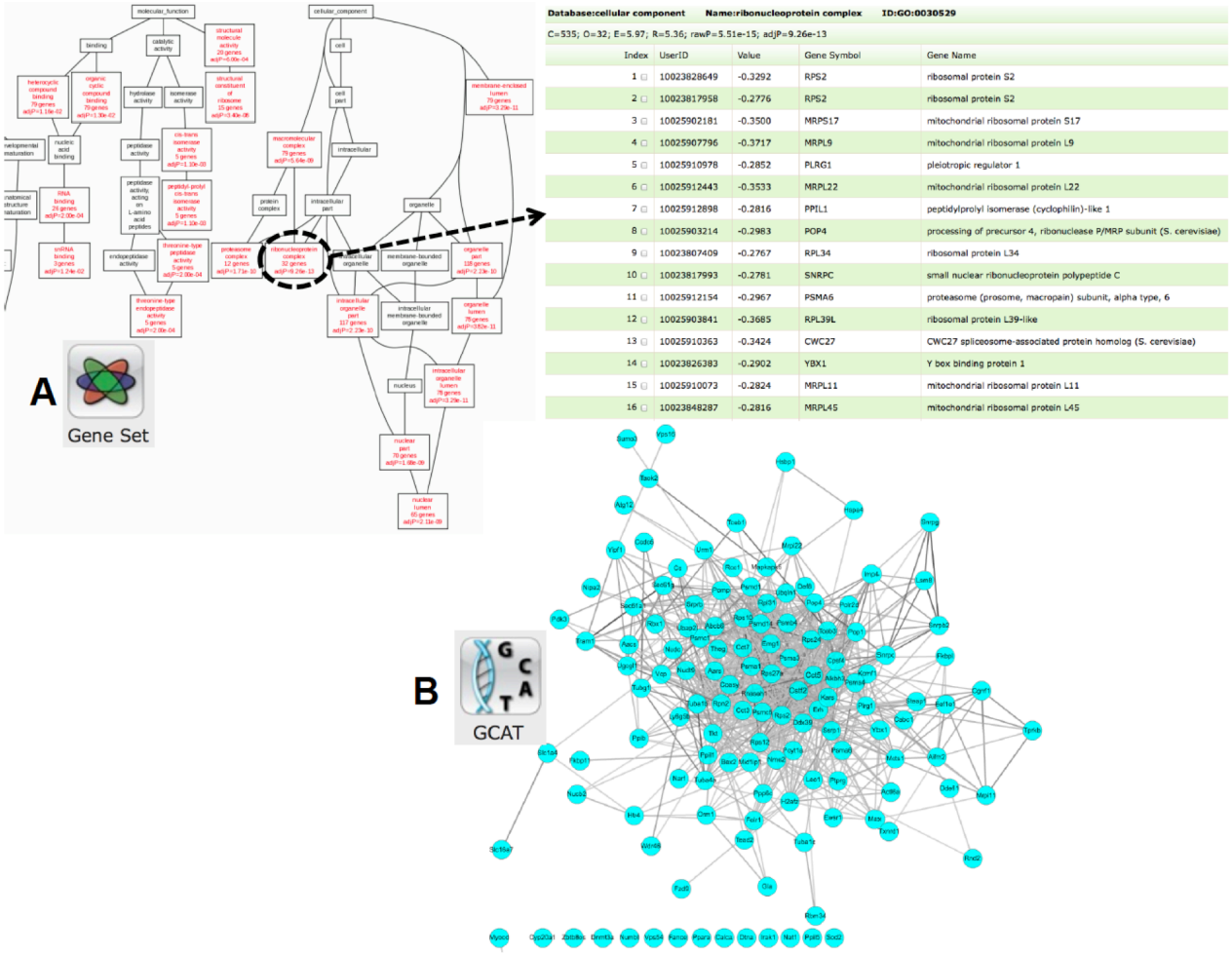


Figure 14. Correlation table and correlation scatter plot. (A) The Correlation Table displays the results of a correlation analysis between a trait or data of interest and other traits collected from the same cohort. In this case, the correlation analysis is between the demographic age data and gene expression in the liver. **(B)** Individual scatter plots can be displayed by clicking on correlation values found in the **Sample r** column. This example shows a significant negative correlation between the expression of a mitochondrial ribosomal protein gene, *MRPL9*, and age. **(C)** Users can select transcripts in the table by setting the correlation criteria using AND/OR operators.



Database:cellular component Name:ribonucleoprotein complex ID:GO:0030529

C=535; O=32; E=5.97; R=5.36; rawP=5.51e-15; adjP=9.26e-13

Index	UserID	Value	Gene Symbol	Gene Name
1	10023828649	-0.3292	RPS2	ribosomal protein S2
2	10023817958	-0.2776	RPS2	ribosomal protein S2
3	10025902181	-0.3500	MRPS17	mitochondrial ribosomal protein S17
4	10025907796	-0.3717	MRPL9	mitochondrial ribosomal protein L9
5	10025910978	-0.2852	PLRG1	pleiotropic regulator 1
6	10025912443	-0.3533	MRPL22	mitochondrial ribosomal protein L22
7	10025912898	-0.2816	PPIL1	peptidylprolyl isomerase (cyclophilin)-like 1
8	10025903214	-0.2983	POP4	processing of precursor 4, ribonuclease P/MRP subunit (S. cerevisiae)
9	10023807409	-0.2767	RPL34	ribosomal protein L34
10	10023817993	-0.2781	SNRPC	small nuclear ribonucleoprotein polypeptide C
11	10025912154	-0.2967	PSMA6	proteasome (prosome, macropain) subunit, alpha type, 6
12	10025903841	-0.3685	RPL39L	ribosomal protein L39-like
13	10025910363	-0.3424	CWC27	CWC27 spliceosome-associated protein homolog (S. cerevisiae)
14	10023826383	-0.2902	YBX1	Y box binding protein 1
15	10025910073	-0.2824	MRPL11	mitochondrial ribosomal protein L11
16	10023848287	-0.2816	MRPL45	mitochondrial ribosomal protein L45

Figure 15. Biological enrichment and network analysis. Gene lists can be sent directly from gene network to other external websites for **(A)** Gene Ontology, and **(B)** functional network analysis.

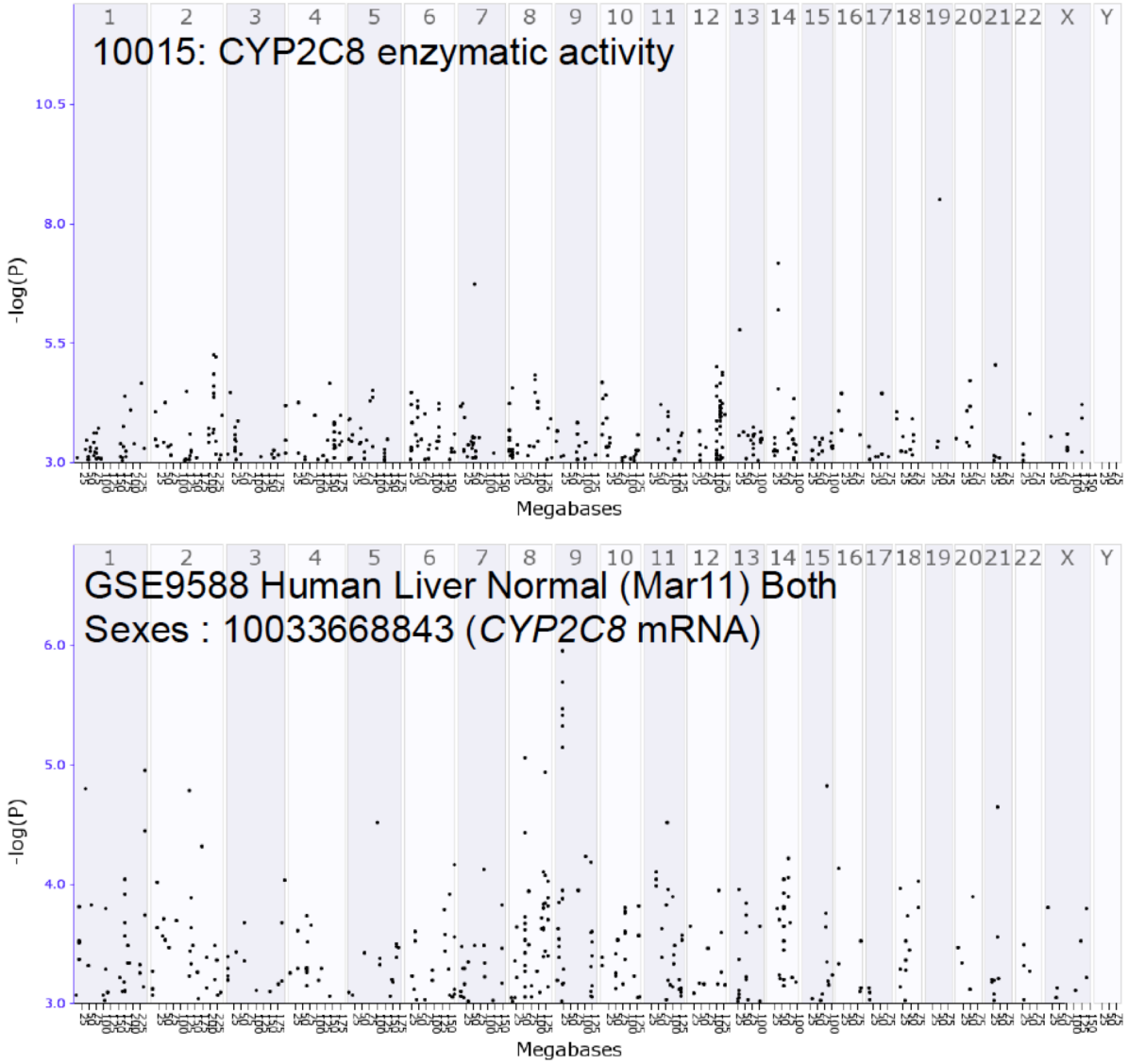


Figure 16. Manhattan Plots. Basic genetic association test is performed within GeneNetwork using PLINK and result is displayed as a standard Manhattan plot. Comparing between the GWAS results for the (top) CYP2C8 enzyme activity (Record ID *10015*), and (bottom) expression of *CYP2C8* gene in liver (*GSE9588 Human Liver Normal (Mar11) Both Sexes*: *10033668843*), we find no common genetic modulator of the two related traits.