WGAViewer: A Software for Genomic Annotation of Whole Genome Association Studies

Supplementary Material

WGAViewer User's Guide

Whole Genome Association Annotation Software Package

Version 1.25

Dongliang Ge

David Goldstein

Duke Institute for Genome Sciences & Policy
Center for Population Genomics & Pharmacogenetics

Contents

0. Introduction and motivation

Dramatic advances in genotyping technologies have allowed the development of affordable products that simultaneously genotype a genome-wide set of polymorphisms that are known to represent most of the common genetic variants in specific human population groups. Because of these developments, the use of HapMap data (The International HapMap Consortium. 2005) and other genome resources has shifted from upstream SNP-selection tasks to the downstream tasks of interpreting the observed genotype-phenotype associations (Telenti and Goldstein 2006). Ideally these resources should be used not only to help distinguish real associations from false positives ones, but should also help to generate hypotheses concerning the possible biological bases of observed associations.

These expectations create an immediate need to develop approaches that facilitate interpretation of a large set of P values in the context of known genomic features and also in the context of other studies of similar phenotypes. The ultimate goal is to allow investigators to consider the full set of P values resulting from an association study rather than simply looking at the few "top" polymorphisms with the lowest P values. This software package allows the researcher to visualize and consider other supporting evidence, such as the genomic context of the SNP, linkage disequilibrium (LD) with ungenotyped SNPs and the evidence from other Whole Genome Association projects alongside the P value of association when determining the potential importance of an individual SNP. Most importantly, it would highlight possible mechanisms, for example by directly or indirectly implicating a polymorphism with an apparent link to gene expression, splicing, non-coding RNAs or other possibilities that would suggest specific functional follow-up.

1. About this software

WGAViewer is a free software tool that is designed to provide a user-friendly interface to annotate, visualize, and help interpret the full set of P values indicating evidence of association resulting from a Whole Genome Association (WGA) study. The full set of resulting P values (however calculated) will be referred to here as the WGA results. WGAViewer is a program developed in JAVA language.

The current version offers six classes of annotation:

1) Chromosome view of WGA results allowing
   (1) Region selection; zoom in/out; search for gene/SNP; etc.;
   (2) Support for multiple databases enabling cross check for evidence;
   (3) Top hits sorting with individual SNP annotation;

2) Genic annotation of WGA results with explicit reference to:
   (1) Alignment using the latest Genome build version;
   (2) Gene/transcripts structure and related information (Hubbard et al. 2007);
   (3) Linkage disequilibrium context(The International HapMap Consortium. 2005);
   (4) Evidence of selection (Voight et al. 2006);
   (5) Hyperlink for other available information for genes, transcripts, exons, and SNPs.

3) Annotation for SNPs:
   In addition to genic annotation in the surrounding region, this annotation also includes:
   (1) Indication of LD score for all genotyped and non-genotyped HapMap SNPs in specified region(The International HapMap Consortium. 2005);
   (2) Test and plot of association with specified gene expression, using the GENEVAR data (Stranger et al. 2005; Stranger et al. 2007);
   (3) Other available SNP-related information, such as ancestral allele, function (synonymous, non-synonymous, splice-site, etc) (Hubbard et al. 2007).
   These annotations could be either performed on certain selected SNPs individually, or on a set of SNPs with high ranks (top hits) automatically.

4) Gene/SNP finding:
   These functions offer a convenient way to locate and annotate candidate genes/genes of specific interest in a WGA project, and align with the physical coordinates from the latest genome build. This annotation also enables an easy search for specific SNPs and/or their LD proxies if they are not present in a WGA project. These functions make an easy and reliable comparison with existing reports.

5) Evidence from multiple genome scans (or smaller datasets):
   This software allows the user to load multiple databases simultaneously, with one of them as the "core" database to be considered. The supporting databases could include replication studies, projects with related phenotypes, other publicly available projects, even studies with different marker sets or different phenotypes. These datasets can then be listed and plotted alongside the core database as concurrent evidence.

6) Supporting/QC databases:
   Supporting information, for example, HWE P values, effect size, effect direction, QC scores, or other user-customized data, can be loaded and listed alongside the main result set to assist the interpretation of the findings.

In addition to these annotations, WGAViewer also provides a convenient platform to directly access and annotate result sets from WGA and other association studies hosted in and released from Duke Institute for Genome Sciences and Policy, through **the Mart for IGSP Data from Association Studies (MIDAS)**.

Furthermore, WGAViewer also offers several classes of **useful annotation tools without requirement of a WGA result set or any other set of P values**. These include:

(1) Test for association of SNP genotype with gene expression
(2) SNP Annotation
(3) SNP Linkage Disequilibrium testing

WGAViewer is supported by the Center for HIV-AIDS Vaccine Immunology (CHAVI, http://www.chavi.org/) and the Duke Institute for Genome Sciences & Policy (IGSP, http://www.genome.duke.edu/).

Comments, suggestions, and bug reports are welcome. For these purposes please send an email to the author: d.ge@duke.edu

Downloading WGAViewer Version 1.10b:

WGAViewer can be downloaded without charge at:

http://www.genome.duke.edu/centers/pg2/downloads/wgaviewer.php

2. Installation

(2.1) System requirements

(2.1.1) Hardware

A computer system with processor equal to or faster than 1.0GHz, and RAM equal to or greater than 1GB is required.

(2.1.2) JAVA

This software is written in JAVA language, therefore a JAVA environment is required. JAVA software can be downloaded free of charge from: http://www.java.com/en/download/index.jsp .

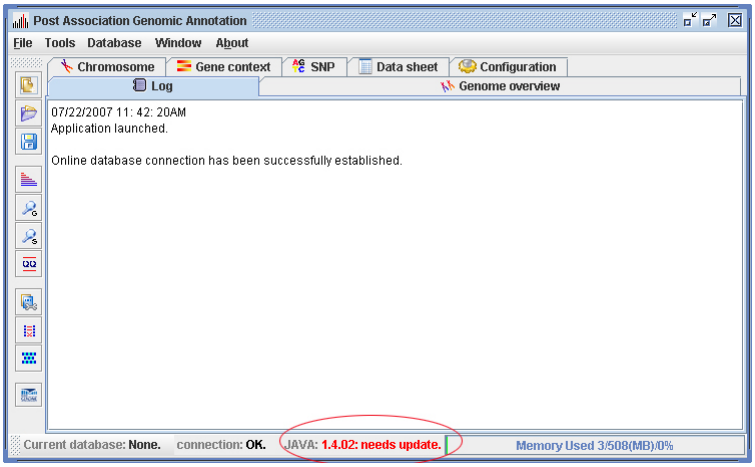


Figure 2.1.2-1. Warning message on JAVA version

An appropriate version of the JAVA software must be successfully installed before this software can be executed correctly. The version of the JAVA software should be equal to or later than "Java Runtime Environment Version 5.0 Update 7" (1.5.07). This software cannot run properly on a computer system with JAVA environment earlier than the specified version. It will show a warning if the JAVA version is out-of-date (See trouble shooting (1), figure 2.1.2-1).

Important note: check your JAVA version before you install and run this program. To do this:

For windows: click 'start' -> click 'run' -> type in 'cmd' -> hit 'ok' -> type in 'java -version' -> hit 'enter'.
For Linux/Unix: type in terminal window 'java -version'.
For Mac OS: From the desktop, go into Applications -> Utilities, and open Terminal.app. Type in'java -version'.

Box 2.1.2-1. Determine JAVA version.

This software may be used on systems that support JAVA environment, including Windows, LINUX, UNIX, and Macintosh.

(2.1.3) Database connection

Most of the annotation is based on a MySQL connection with the Ensembl (Hubbard et al. 2007) database servers , and a HTTP connection with the HapMap (The International HapMap Consortium. 2005) database servers established by the WGAViewer software on your local machine. Therefore an internet connection is necessary to perform the annotation. For many users from academic institutions, this connection can be directly through port 3306, the standard MySQL port. If this direct connection cannot be established (for example, if you work behind a proxy server) WGAViewer provides an option to establish the connection through a HTTP-MySQL bridge service hosted in Duke University Institute for Genome Sciences and Policy. This option can be accessed through menu "Options -> Online database connections", as illustrated in Figure 2.1.3-1.

WGAViewer has been setup to automatically test this connection when it is launched. If the connection cannot be established under the current internet profile (as illustrated in Figure 2.1.3-1), a warning will

be shown (See trouble shooting (2), figure 2.1.3-2). In this case you may still load your dataset, for example, the example set released with this package. But you may not perform further annotation tasks.

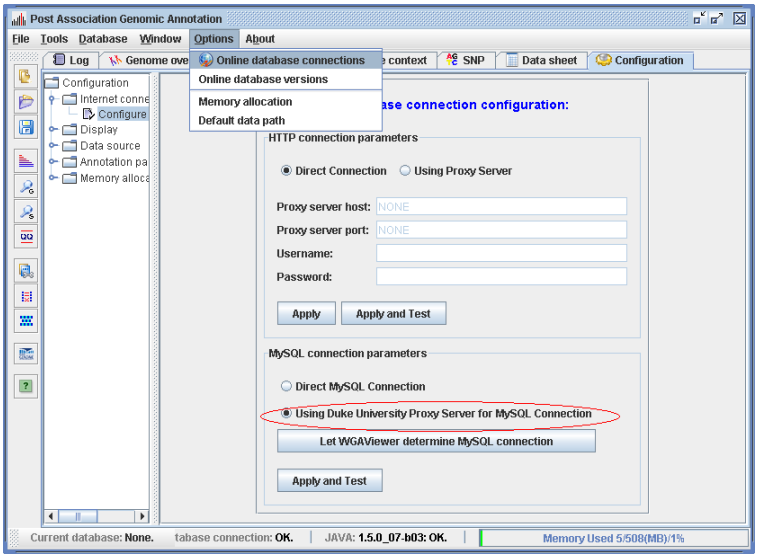


Figure 2.1.3-1. Database connection options.

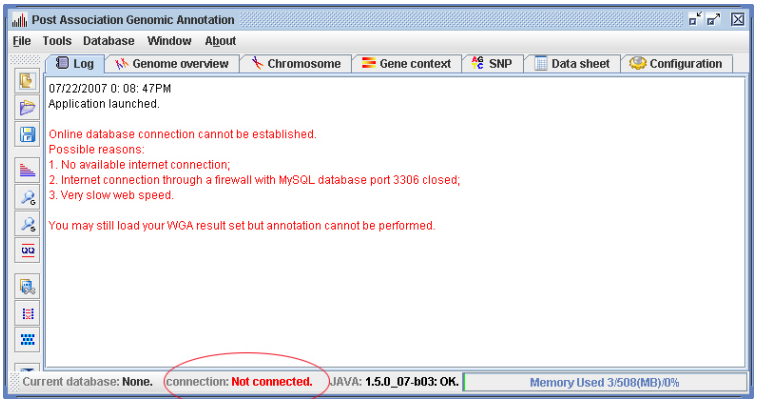


Figure 2.1.3-2. Warning message on database connection.

(2.2) Installation

The WGAViewer package comes in one ZIP file. Unzip (extract) this file to a folder (home folder). No additional installation is necessary. The ZIP file has to be unzipped to run the program. WGAViewer cannot be launched from a ZIP folder (for example, opened by Windows explorer).

The size of the full functional release is about 180MB. A "Lite" version is also offered for downloading with a reduced size of around 40MB. Without the support database for selection and GENEVAR, this "Lite" version can perform genic and LD annotation but not annotations for gene expression or selection.

Windows users should double click "WGAViewer.exe" in the home folder to launch the program.

For LINUX/MAC users, there is a "WGAViewer.sh" in the home folder. Change the permission of this file to be runnable, for example by typing: "chmod 777 WGAViewer.sh" in the terminal window. And then double click this file to launch the program.

Alternatively, double click a jar file in the home folder: WGAViewer.jar. to start the program, or type "java –jar WGAViewer.jar" in a terminal window.

(2.3) Memory allocation

The default startup will allocate up to 512MB of memory, which should be sufficient to load most genome-wide data files with up to 550K markers. A real time memory information bar monitors the memory allocation and usage (lower right corner, Figure 2.3-1). Green color indicates the memory allocation is adequate, while yellow or red indicates the program is running out of memory. In this case you may want to increase the default memory allocation by clicking on the memory monitoring bar (Figure 2.3-2). The program needs to be restarted to effect the change.

Usually the allocated memory should be at least greater than four-fold of the data file size, for example, if the data file is 30MB, the roughly estimated memory requirement should be at least 128M. This allocated number should not exceed the amount allowed by the system, that is, the "free" memory. Usually you may not allocate more than one half of your total memory. This means your computer should be equipped with at least 1GB RAM for a reasonable capacity and performance.

Figure 2.3-1. Memory monitoring bar.

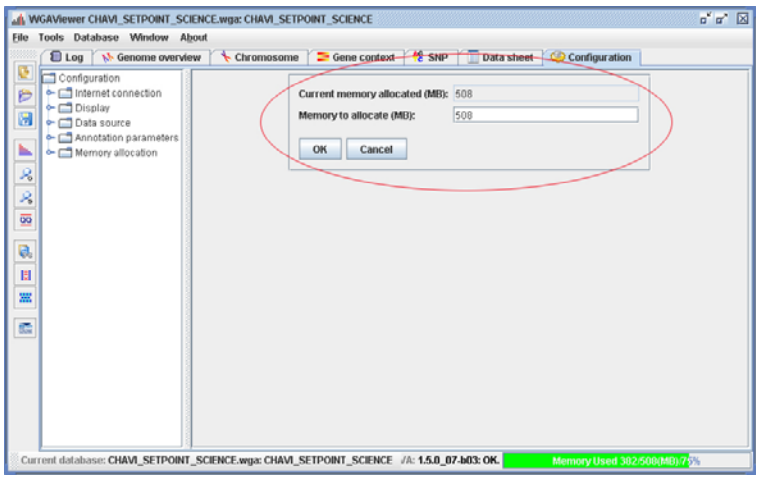


Figure 2.3-2. Change memory allocation.

(2.4) Trouble shooting
(2.4.1) The main program window does not show up, or you see a JAVA version warning (Figure 2.1.2-1).

This problem is caused by an older JAVA version. Try to uninstall your current JAVA software and download a current one. If you have manually specified your CLASSPATH environmental variable, make sure to include the current directory (".") .

Also see Installation->Software.

(2.4.2) A connection warning shows up after launch (Figure 2.1.3-2).

WGAViewer has been setup to automatically test the connection to the HapMap and Ensembl databases when it is launched. If the connection cannot be established, this warning will be shown. In this case you may still load your dataset, for example, the example set released with this package, but you may not perform further annotation tasks.

Sometimes this warning may also be shown because of a very slow web speed. Try to click on the circled warning message (Figure 2.1.3-2) to retest the connection.

If you work behind a proxy server and/or a username and password is required to access the internet, try to set the internet configuration of WGAViewer to through "Duke University Proxy Server" (Figure 2.1.3-1).

If you are sure you can access internet but this message still shows, consult with your internet administrator or write to author of this software.

If these solutions do not help, please email the author: d.ge@duke.edu

3. Using the software

(3.0) Structure of this guide; questions to be answered

We hope that this guide will not only offer practical and easy-to-follow advice as to how to use this software, but also provide a platform that we can share our own thoughts and experiences of trying to interpret the statistical results yielded from a WGA project. Therefore, instead of listing the functions of each menu item, button, and interactive graphical components one by one, we have constructed this guide as a tour through a real WGA annotation process recently completed in our own group (Fellay et al. 2007).

The following are examples of questions that came up during this study and that led to specific features now implicated in WGAViewer:

- What are the top hits and their P values?
- Are these top hits located in or near any gene?
- If they are located in a gene, what type of SNPs are they? Are they of known function?
- If they are not non-synonymous coding SNPs, nor located in a known splice site, how far are they from the closest exon?
- If they are not in a known gene, how far are they from the closest known gene?
- What exactly is the genic context for each hit? What are the surrounding genes?
- Is there any evidence for evolutionary conservation/ selection of the surrounding region?
- What are the P values of the surrounding SNPs?
- What is the LD context among these SNPs?
- How far does the LD extend for each hit? Does this LD extension cover other genes?
- Are there (perhaps ungenotyped) proxies for the associated SNP that are in a more interesting genomic context?
- Do these hits or their proxies show any association with available functional data, for example, gene expression levels?
- After all, is there a way to conveniently annotate these hits in an automatic and batch manner? Is there a way to automatic filter their proxies by their function?

- There are many candidate gene studies published on the same or related phenotypes. What are the P values for SNPs in and around these associated genes in our WGA project? Can we replicate previous findings?
- Can we replicate previous associations of particular SNPs?
- If the previously -associated SNPs have not been included in our WGA project, are there any correlated proxies or tags for these candidate SNPs? What are their P values?

- Is there evidence of population stratification effects?

- We have association data for replication cohorts. There are also cohorts with related but not identical phenotypes. Is there a way to compare them easily?
- We have genome-wide HWE test results. We have effect size, effect direction, etc. Is there a way to list them alongside our association findings?

- I don't have a WGA set. But I want to annotate a SNP in such a way too. I also want to test LD among a list of SNPs. I want to test SNP-gene expression associations. Are there any convenient bioinformatic tools that WGAViewer can offer?

We hope this structure will make this user guide more useful. In the following guide, instead of going through buttons and menus, we will mainly go through these questions using real data.

(3.1) Example datasets

To illustrate this software, we use our recently completed study on host control of HIV-1 viral load during the asymptomatic set point period (Fellay et al. 2007) as an example (CHAVI_SETPOINT_SCIENCE). All the illustrations in this documentation are based on this example.

To load this example, click on menu "File->Open an example dataset" (Figure 3.1-1). Two types of datasets are released with this package. One is the pre-annotation dataset (.wr file, text-based) and the other is the post-annotation dataset (.wga file, binary). Figure 3.1-1 shows that the annotated dataset is selected, and Figure 3.1-2 shows the window after this example dataset is loaded. Figure 3.1-2 also shows two of the genome-wide significant hits (red vertical lines) that we recently reported (Fellay et al. 2007). This real dataset is based on the Illumina HumanHap 550K SNP chip.

For demonstrating the usage of multiple databases discussed later in this user guide, we also include a simulated dataset (Illumina_HumanHap300_sim.wr) in this software package. This is a simulated set based on Illumina HumanHap 300K.

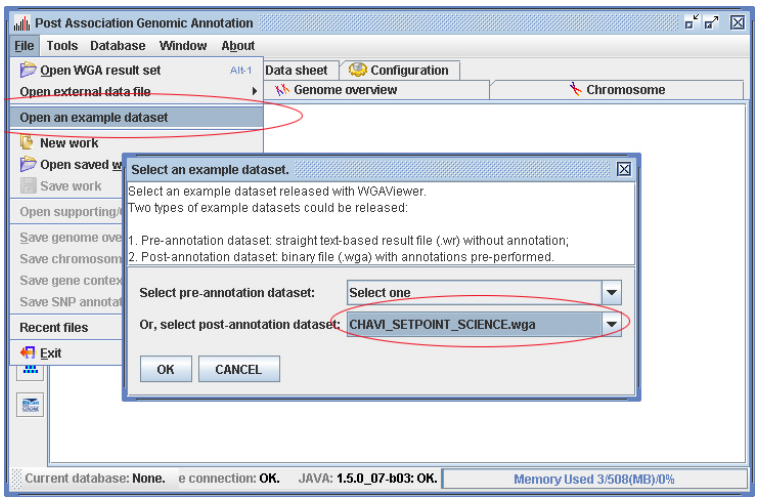


Figure 3.1-1. Load example dataset.

Figure 3.1-2. Example dataset loaded.

(3.2) Data input

WGAViewer directly loads text-based datasets (.wr file) and can also save the annotation work session to a reloadable binary project file (.wga file). WGAViewer can make .wr file from different types of text-based data sets, including output from PLINK (Purcell et al. 2007).

(3.2.1) Text –based datasets (.wr)

To make the text-based input file for WGAViewer, click on menu "File->Open external data file -> make input file for WGAViewer".

WGAViewer also directly supports the outputs from PLINK (Purcell et al. 2007). To open PLINK output, click on menu "File->Open external data file->Open PLINK output".

These processes will generate and save a ".wr" input file for WGAViewer. This file will be automatically loaded after generation. Click "File -> Open WGA result set" to reload the generated ".wr" file at a later time.

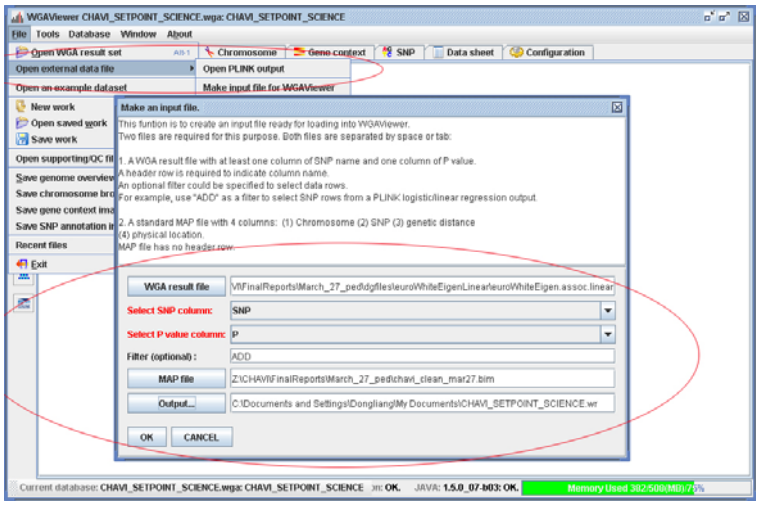


Figure 3.2.1-1. Load external data file.

Figure 3.2.1-1 shows an example of the configuration of this process. In this example we directly load a WGA result set generated by linear regression analysis from PLINK. The external data set to be loaded should be separated by either space or tab key. It should have one row of header line indicating each column name, and at least one column for SNP and one column for P value (as shown from Figure 3.2.1-1). Every data row should have the same number of columns with the header row, otherwise the program will show an error message (Figure 3.2.1-2). A standard MAP file (or PLINK BIM file) is also required. A MAP file has no header line, and has four columns: chromosome, SNP, genetic distance (could be 0 and will not be used), and physical chromosomal location (could be from older genome build and will be annotated later). For an example see Box 3.2.1-1.

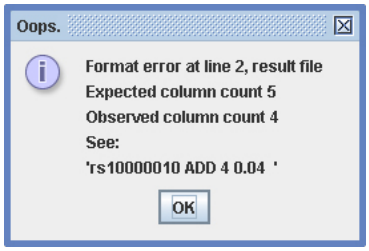


Figure 3.2.1-2. Error message for column inconsistency.

2 rs10495761 0 28147606
2 rs10495766 0 28783336
2 rs10495767 0 29198843

Box 3.2.1-1. An example of a MAP file.

Mitochondrial SNPs should be marked as "M" in the "Chromosome" column; SNPs on chromosome X or Y must be marked as either "X" or "Y", respectively. The acceptable "Chromosome" values are either integers 1-22, or string "X", "Y"/"XY", or "M". The case does not matter. It is necessary that each SNP has a map location as a positive integer. Any SNP that has a "Map" value other than a positive integer, for example, "NA" or "-9", will be skipped.

In some types of WGA output, for example, logistic or linear regression outputs from PLINK (as shown from Figure 3.2.1-1), one SNP could have more than one row of results (Box 3.2.1-2). That is, each covariate will have one row of result marked with SNP name too.

CHR	SNP	TEST	NMISS	BETA	STAT	P
0	MitoC182T	ADD	851	-0.18	-0.8233	0.4106
0	MitoC182T	COV1	851	-0.3226	-4.058	5.403e-05
0	MitoC182T	COV2	851	0.00942	3.168	0.001588

Box 3.2.1-2. An example of data file with multiple rows for each marker.

In the example shown in Box 3.2.1-2 a filter of "ADD" needs to be specified to only include selected genetic effects. WGAViewer will also try to roughly check the duplication of marker in the dataset. If a SNP apparently appears twice while no filter is specified, the program will show an error message (Figure 3.2.1-3).

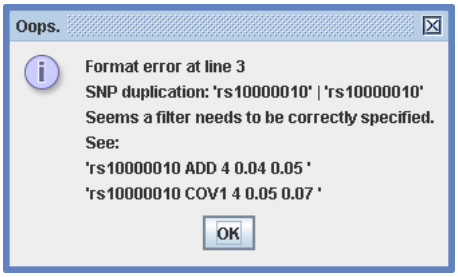


Figure 3.2.1-3. Error message for filter misspecification.

(3.2.2) Binary datasets or saved work session (.wga)

WGAViewer can save the annotation work session into binary file, namely, the .wga file.

To load a binary dataset or saved work session (.wga file), click "File -> Open saved work".
To save a work session any time during the annotation, click "File -> Save work".

(3.3) Batch annotation for top hits

After the statistical analyses of a WGA project, usually the first important question that most researchers want to answer is: what are the top hits and their P values? After this they may be even more eager to know, are these top hits located in or close to any gene? If they are located in some gene, then what type of SNPs are they? Are they functionally relevant? --- if they are not non-synonymous coding, not splice changing, then how far are they from the closest exon? If unfortunately they are not in a known gene, how far are they from the closest known gene?

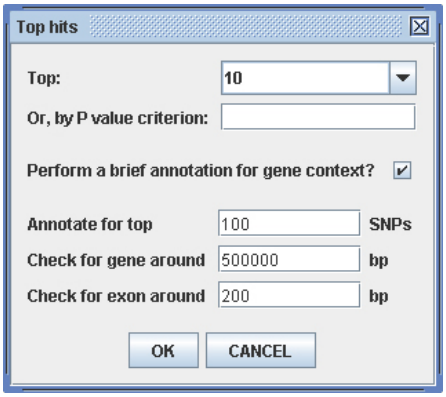


Figure 3.3.1-1. Parameters of first-round annotation for top hits.

(3.3.1) Fast first-round annotation

To address these immediate questions, WGAViewer offers a fast (around 0.5-2.0 second per SNP, depending on web and CPU speed) first-round annotation procedure.

As an illustration for this function we load the pre-annotation example dataset by clicking on menu "File->Open an example dataset" and select "Pre-annotation dataset" "CHAVI_SETPOINT_SCIENCE.wr". Click on menu "Tools->Top hits", an annotation dialog will present itself (Figure 3.3.1-1). In this fast annotation process, using the latest genome build and coordinates available at the time of the annotation, WGAViewer will search for any known transcripts and exons located within an adjustable upstream and downstream region. WGAViewer will also annotate each SNP for their consequence type from the following 14 categories:

ESSENTIAL_SPLICE_SITE, STOP_GAINED, STOP_LOST, FRAMESHIFT_CODING, NON_SYNONYMOUS_CODING, SPLICE_SITE, SYNONYMOUS_CODING, REGULATION_REGION, 5PRIME_UTR, 3PRIME_UTR, INTRONIC, UPSTREAM, DOWNSTREAM, INTERGENIC.

This process is based on an annotated and updated physical location so the accuracy of the annotation can be ensured even if the MAP file, which has been used to generate the input .wr file for WGAViewer, is from an out-of-date genome build. This policy also applies for all the following annotation procedures.

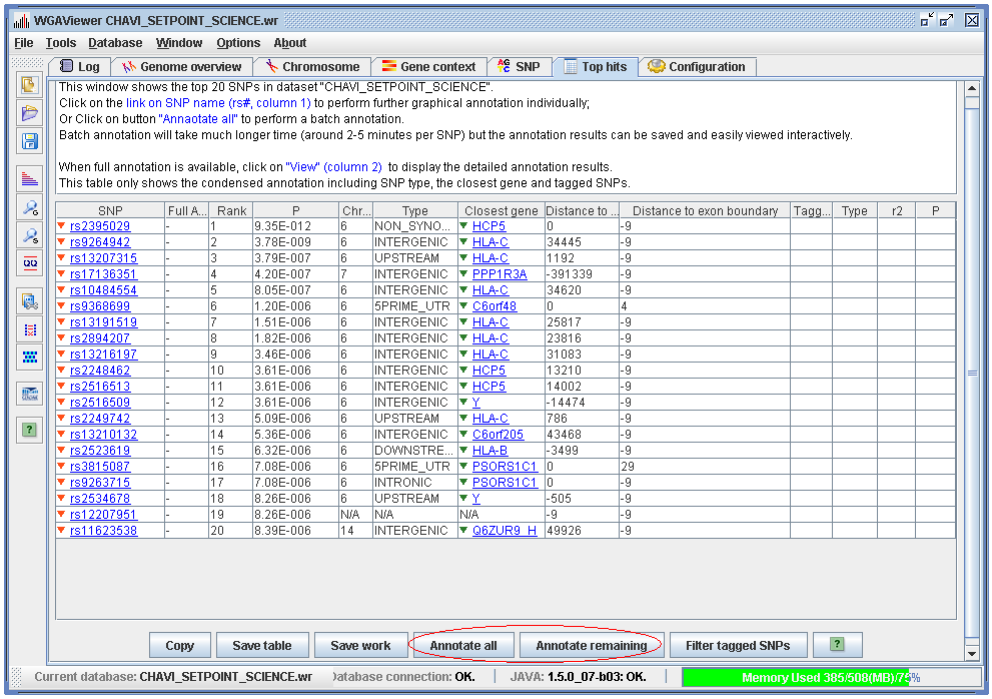


Figure 3.3.1-2. Results of first-round annotation for top hits.

Figure 3.3.1-2 shows the results of this fast first-round annotation using data on host control of HIV-1 viral load. As shown from this figure, most of the hits among the top 20 are on chromosome 6 and clustered in MHC region. The first hit, rs2395029, is non-synonymous coding in gene HCP5. The

second hit, rs9264942, is classified as intergenic but only 35Kbp upstream to HLA-C gene. This first-round annotation provides a very brief but explicit insight into the top hits. But one also has the option to skip this step and directly enter into the next stage by un-checking the "Perform a brief annotation for gene context" box in Figure 3.3.1-1.

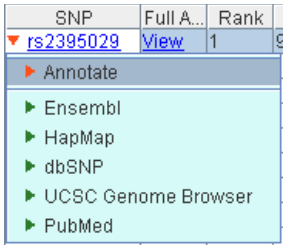


Figure 3.3.1-3. Resource menu for top hits

Click on the blue hyperlink of each hit will bring up a resource menu for top hits (Figure 3.3.1-3). Click on the resource entries (for example, dbSNP) to visit the webpage for the hit in the relevant public databases, or click on "Annotate" to perform individual comprehensive annotation, as we will discuss in the following section.

Likewise, click on the blue hyperlink of each gene will bring up a resource menu for the closest genes (Figure 3.3.1-4).

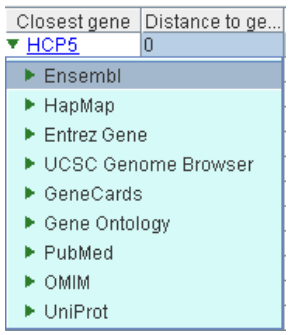


Figure 3.3.1-4. Resource menu for the closest gene

(3.3.2) Comprehensive annotation

In addition to these fast annotations, there is however one class of annotation that goes beyond convenient summaries of work that could be carried out manually and individually. The manual annotations using disparate databases are not feasible for any but the several top discoveries in any study. What happens with a polymorphism that only achieves a P value of 10^{-5}, but which is itself in strong LD with an ungenotyped SNP that is annotated as definitely functional? This SNP would warrant special consideration, but would be missed in most manual settings. We have therefore established a slow annotation (typically 2-5 minutes per SNP) meant to pluck out such suggested associations and this is one of the few features of WGAViewer that can be viewed as moving toward real automated consideration of the full sets of results, as opposed to only expediting and summarizing analyses that would in any event have been completed. In addition to the gene context, LD extension, and expression annotation routines for each SNP, this process will also automatically check and filter the functions for both the original genotyped SNPs and their LD proxies. Once the annotation is done, the interactive filtering and other annotation features can be saved for convenient later use.

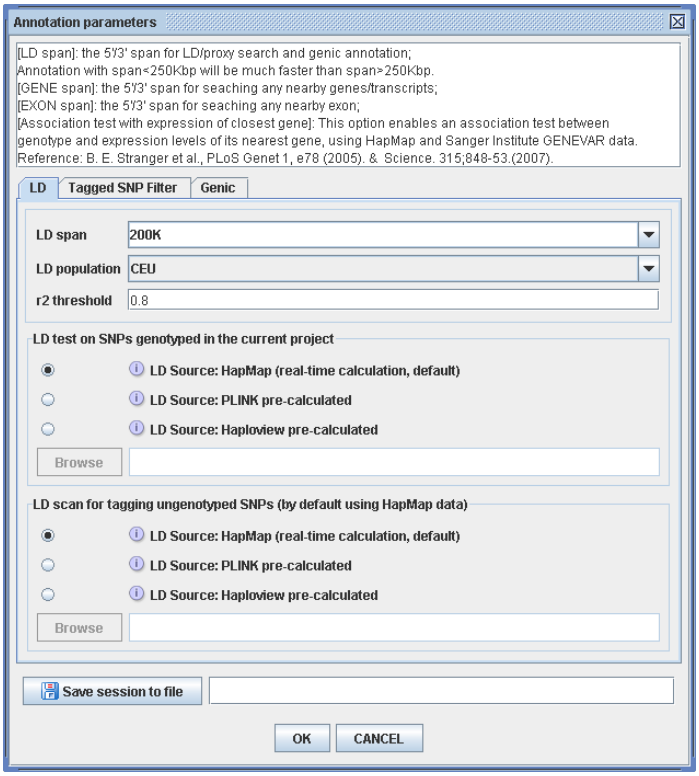


Figure 3.3.2-1. Parameters of comprehensive annotation for top hits.

The questions that may be answered by this process include: what is the genomic context for each hit? What are those genes surrounding these hits? What are the SNPs and their P values around these hits? What is the LD context among these SNPs? How far does the LD extend for each hit? Do these hits represent proxies for functionally more relevant but ungenotyped variations? Do these hits or their proxies show any association with available functional data, for example, gene expression level? Etc.

To perform this annotation, click on the blue hyperlink for each SNP rs# in Figure 3.3.1-2 individually, or click button "**Annotate all**" or "**Annotate remaining**" to launch a batch job. Either way will bring up a dialog for annotation parameters (Figure 3.3.2-1). The user then needs to specify the annotation parameters or leave the default values. "**LD span**" determines the main window size for the following

annotation. **"r2 threshold"** will be used as the criterion to determine tagged SNPs for each hit. By default the LD test will be performed using HapMap data, but one has the options to use pre-calculated LD dataset as **LD source**, using either PLINK or Haploview. **Gene span and exon span** have the same definition as shown in figure 3.3.1-1, and will affect only those that have not been briefly annotated. The LD context is based on HapMap data. The specification of population and the cutoff for tagging can be setup in the "configuration" panel and will be discussed later in this documentation. If a batch annotation job is selected, an **output file** (.wga work session file) is mandatory. This is to avoid loss due to accidental interruption of the annotation procedure, for example, power surge, internet connection failure, etc. During this process WGAViewer will save the work session from time to time, so if the annotation is interrupted the saved work session can still be loaded and the annotation process can be restored by clicking on button "Annotate remaining" (Figure 3.3.1-2). Using the default parameters shown in Figure 3.3.2-1, around 2 minutes are required to annotate each SNP.

After the annotation is successfully performed (either individually or in a batch), a clickable "**View**" hyperlink will appear next to the column of rs# for each SNP with annotation results (Figure 3.3.2-2). Clicking on this "View" hyperlink will immediately bring up the saved annotation results with no further wait.

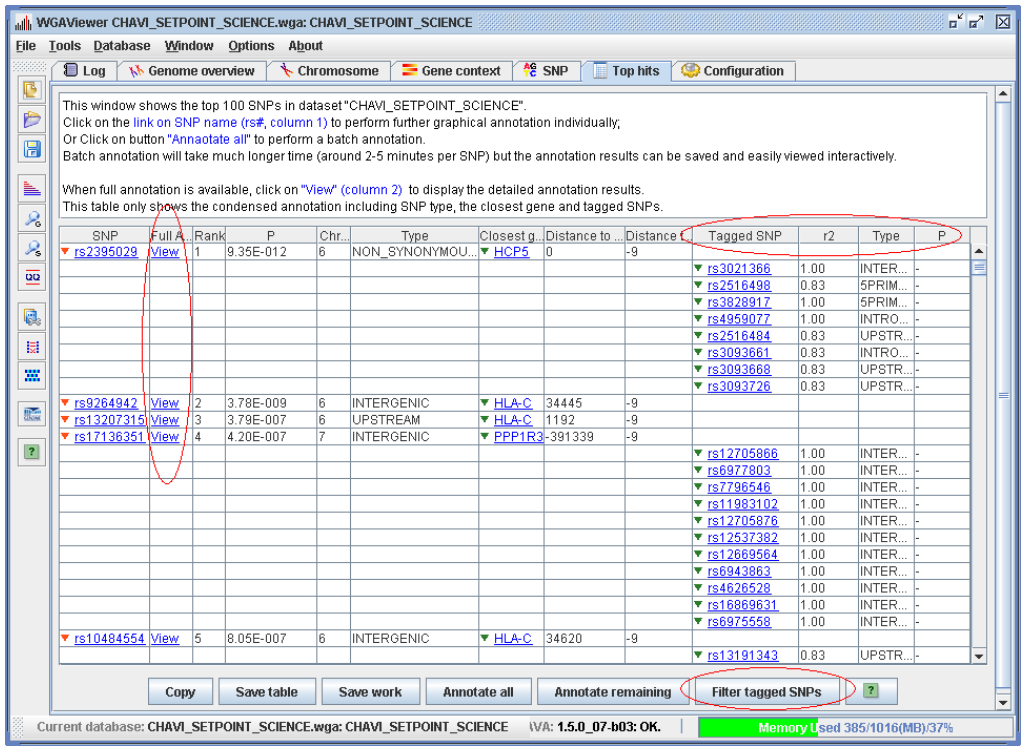


Figure 3.3.2-2. Results of comprehensive annotation for top hits.

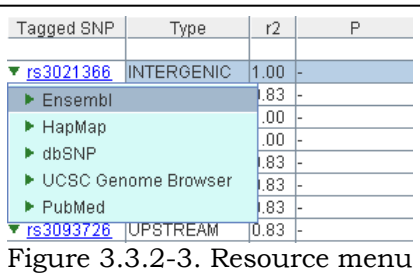


Figure 3.3.2-3. Resource menu for tagged SNPs

Like the resource menu for top hit and genes (Figure 3.3.1-3, 3.3.1-4), click on the tagged SNPs will bring up a resource menu which can direct to public databases (Figure 3.3.2-3).

The user also has the option to save the contents of this annotation summary table to a comma-separated text file (.CSV) by clicking on the "**Save table**" button, or to copy tab-separated data into the system clip board by clicking on the "**Copy**" button for pasting into any text editor or Microsoft EXCEL. To save the work session, click on button "**Save work**". All the interactive features, annotation results can be reloaded from a saved work session (.wga file).

WGAViewer also offers an easy way to filter the SNPs tagged by the top hits. By clicking on button "**Filter tagged SNPs**" (Figure 3.3.2-2) the user can bring up a dialog for the filtering parameters (Figure

3.3.2-4). This dialog lists the 14 categories that are discussed in 3.3.1. This function provides a convenient data-mining method to screen the functionally relevant findings, for example, by excluding all the INTERGENIC HapMap SNPs that have been tagged by the user's WGA project.

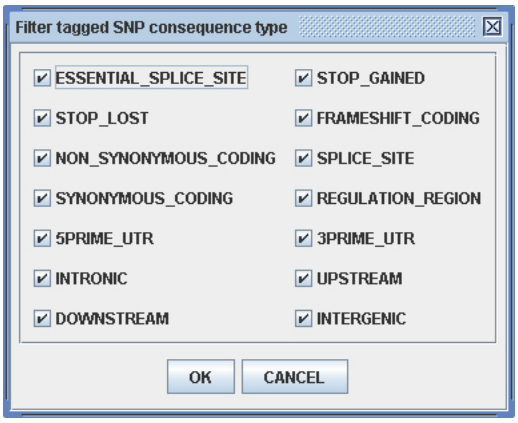


Figure 3.3.2-4. Filtering tagged SNPs.

As an illustration for this function we load the post-annotation example dataset by clicking on menu "File->Open an example dataset" and select "Post-annotation dataset" "CHAVI_SETPOINT_SCIENCE.wga". Click on menu "Windows->Top hits viewer" to display the comprehensively annotated top 100 hits. Figure 3.3.2-2 shows the results. Different from the brief annotation results (Figure 3.3.1-2), the "Full annotation" column now has clickable hyperlinks "View". And for some SNPs, a list of tagged SNPs and the type of the tagged SNPs are also shown in the table.

If one clicks on the hyperlink "View" (Figure 3.3.2-2), for example, on SNP rs9264942, the results of the comprehensive graphical annotation can then be shown sequentially. This involves the following three graphical panels.

(3.3.2.a) Chromosome view (Figure 3.3.2.a-1)

Click on the "Chromosome" tab to display the chromosome view panel (Figure 3.3.2.a-1). This panel shows the SNP and P values spanning the annotation window around the annotated top hit, that is, as shown from Figure 3.3.2.a-1, 200Kbp upstream and downstream to rs9264942. Each P value line responds to mouse movement and a small stick with SNP rs number and P value will appear over the targeted SNP. The chromosome ideogram also shows the approximate position of the annotated region with a transparent red rectangle. All the P value lines plotted in this panel are spaced evenly. If the whole panel cannot accommodate all the P value lines, the program will automatically determine a resolution and displays only the lowest P value for the adjacent n SNPs. The resolution will be shown on the right upper corner: "Current resolution".

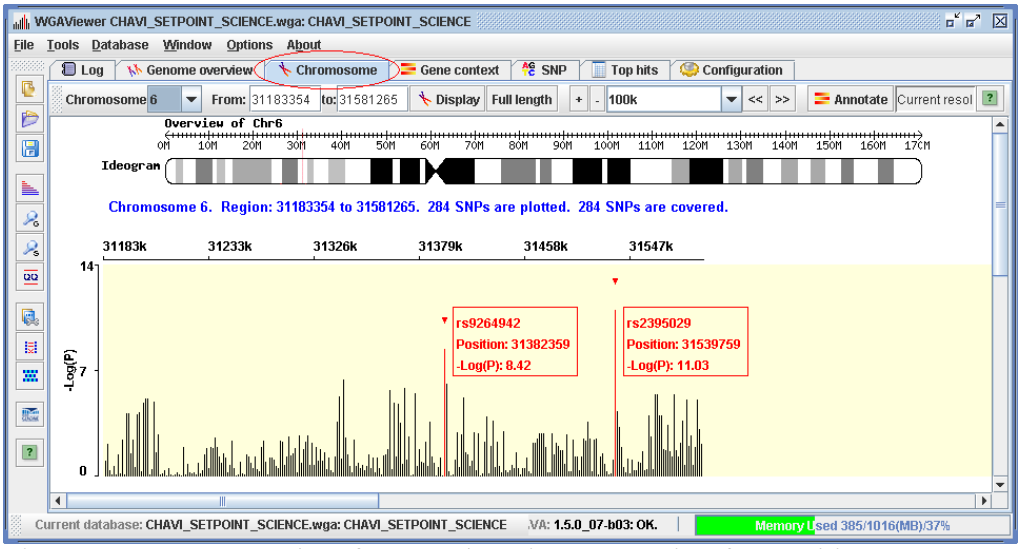


Figure 3.3.2.a-1. Results of comprehensive annotation for top hits: rs9264942, chromosome view.

(3.3.2.b) Gene view (Figure 3.3.2.b-1)

Click on the "Genic context" tab to display the gene view panel (Figure 3.3.2.b-1). This panel shows the annotation of the selected region of chromosome view with transcripts and LD structure. It consists of 8 parts:

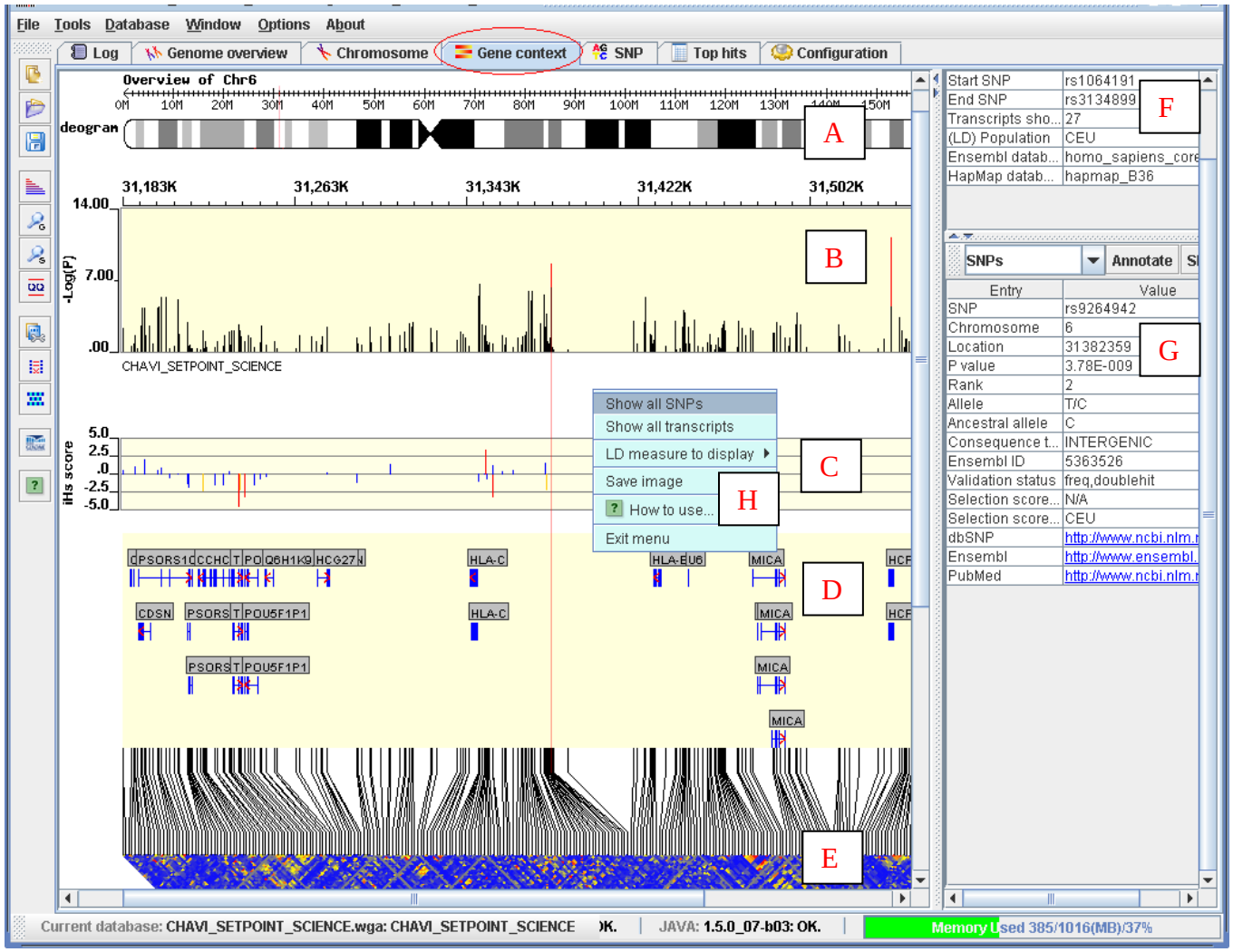


Figure 3.3.2.b-1. Results of comprehensive annotation for top hits: rs9264942, genic view.
A: Chromosome ideogram; B: SNP P value lines; C: Recent selection score; D: transcripts; E: LD matrix; F: Description for annotated region; G: Dynamic data sheet; H: Popup menu.

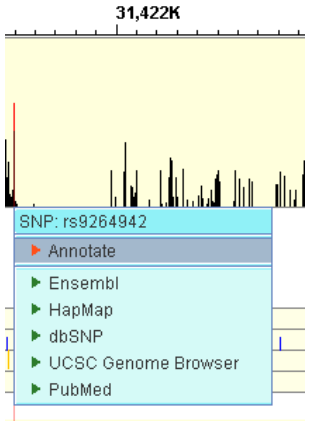


Figure 3.3.2.b-2.
Resource menu for SNPs

A: Chromosome ideogram: shows the annotated region on a chromosome with a transparent red rectangle;

B: Association results: show the SNPs with successful association P values. Different from chromosome view (3.3.2.a), these lines are spaced according to their actual physical location based on the latest genome build (Hubbard et al. 2007). This panel will always plot every SNP P value line, no matter how many SNPs to be plotted, therefore the P value lines could be overlapped but the highest –LogP (lowest P value) can always be seen and be highlighted. Each SNP P value line will respond to mouse movement and will plot a red highlight line towards part D to show the detailed information for each SNP, together with the hyperlink to external databases, in an dynamic data sheet (part G); **Click** on each SNP line to bring up a resource menu (Figure 3.3.2.b-2). For lines too

dense to easily pick up by mouse movement, **press key "<"/","** to move the highlighted lines backward (left, towards smaller chromosome coordinates), or **key ">"/"."** to move the highlighted lines forward (right, towards larger chromosome coordinates), and then **press key "enter"** to bring up this menu.

C: Recent selection score (Voight et al. 2006) where available for SNPs shown in part B;

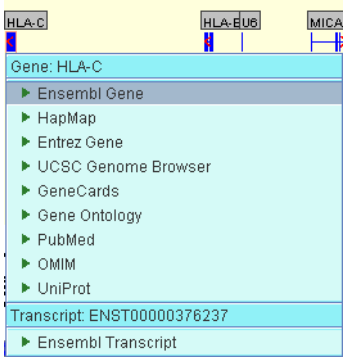


Figure 3.3.2.b-3. Resource menu for transcripts

D: Transcripts: shows transcripts located in the annotated region. Exons are plotted as blue rectangles with a red arrow representing the strand. Each transcript will respond to mouse movement and will show the detailed information for each SNP, together with the hyperlink to external databases, in an dynamic data sheet (part G). Alternative transcripts are plotted with detailed exon information shown in the dynamic data sheet too (part G); **Click** on each transcript to bring up a resource menu (Figure 3.3.2.b-3). For transcripts too dense to easily pick up by mouse movement, **press key "<"/","** to move backward (left, towards smaller chromosome coordinates), or **key ">"/"."** to move forward (right, towards larger chromosome coordinates), and then **press key "enter"** to bring up this menu.

E: LD matrix: shows the LD data (The International HapMap Consortium. 2005) among the SNPs plotted in part B. based on selected HapMap population. The r2 for the color scheme is: blue 0-0.2; yellow 0.2-0.6; red 0.6-1.0. Missing values are coded as -9 and plotted in gray. Each LD cell responds to mouse movement and will show detailed LD information, including the names of the two SNPs, r2 and D', together with HapMap population in the dynamic data sheet (part G);

F: Description for annotated region: shows the physical location and landmark for the start and end of the annotated region. The version of the Ensembl and HapMap databases are also shown in this data sheet. Differing from the dynamic data sheet (part G), the contents of this data sheet are fixed, unless other annotation results are selected. The coordinate discrepancies between the latest Ensembl and HapMap genome builds are adjusted automatically and the coordinates from the latest Ensembl databases are shown in all annotation reports. This enables an accurate alignment between Ensembl variation/transcript data and HapMap LD data.

G: Dynamic data sheet: shows the detailed information for the highlighted item in part B-E. Therefore the contents of this data sheet will change according to which type of item is highlighted in the main graphical region. This data sheet has also a fixed tool bar including a drop-down menu for all the SNPs shown in part B, sorted by the rs#. The user can select any SNP and show the detailed information in the dynamic data sheet. A red highlight line will also then be plotted on part B to D to show which SNP has been clicked. Sometimes this is more convenient than directly pointing the mouse to a specific SNP in part B, because when the SNP density is higher it is difficult to conveniently highlight a SNP among the overlapped lines. For SNP and transcript, this data sheet always offers hyperlinks to external databases, including Ensembl and NCBI, for a convenient reference for data not shown.

H: Popup menu: Clicking on blank region (other than hotspots, for example SNPs or transcripts) will activate this popup menu. It can then be dismissed by click on menu item "Exit". This popup menu offers four functions:

H.1 Show all SNPs: click on this menu item will activate a popup window and show the detailed information for all the SNPs (Figure 3.3.2.b-4) plotted in part B, instead of one by one in dynamic data sheet (part G). This data sheet window offers four methods to sort the SNP collection: by location, by P value, by rs #, or by type. Select the sorting method and then click on button "Sort by".

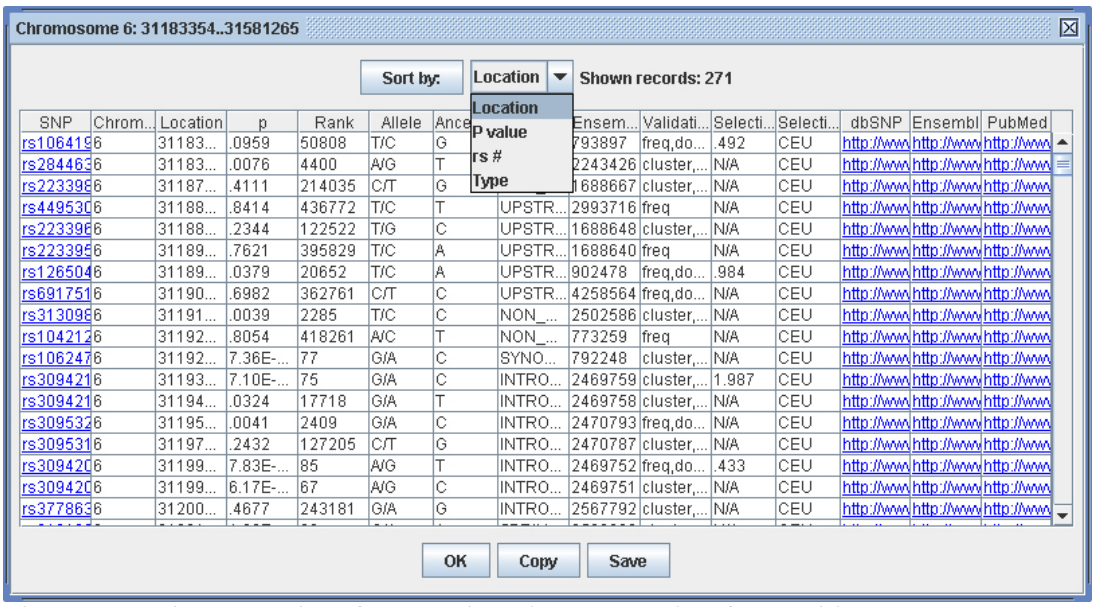


Figure 3.3.2.b-4. Results of comprehensive annotation for top hits: rs9264942, genic view, data sheet for showing all SNPs.

Like the dynamic data sheet, this data sheet window also offers the clickable hyperlink navigating to dbSNP, Ensembl, and PubMed. In addition to this, if one clicks on rs# for each SNP, a red highlight line will be drawn on part B to D to indicate the location of the selected SNP. To dismiss this window, click on the "OK" button. The user also has the option to save the contents of this table to a comma separated text file (.csv) by clicking on the "Save" button, or to copy the tab-separated contents to system clip board by clicking on the "Copy" button and pasting the data into external software, for example, any text editor or Microsoft EXCEL.

H.2 Show all transcripts: clicking on this menu item will activate a popup window and show the detailed information for all the transcripts (Figure 3.3.2.b-5) plotted in part D, instead of one by one in dynamic data sheet (part G). This data sheet window offers up to six methods to sort the transcript collection: by location, by gene symbol (if available), by biotype (if available), by gene ontology (GO) term (if available), by Ensembl gene ID, or by Ensembl transcript ID. Select the sorting method and then click on button "Sort by".

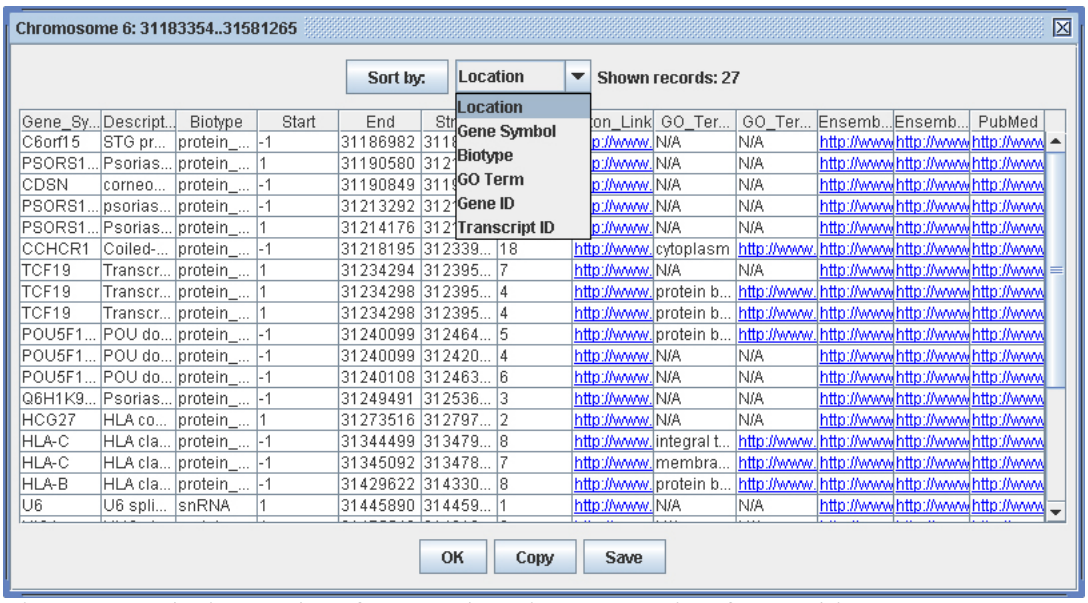


Figure 3.3.2.b-5. Results of comprehensive annotation for top hits: rs9264942, genic view, data sheet for showing all transcripts.

Like the dynamic data sheet, this data sheet window also offers the clickable hyperlink navigating to the Ensembl transcript (and then to the Entrez Gene if needed), the Ensembl exon, and PubMed. To dismiss this window, click on the "OK" button. One also has the option to save the contents of this table to a comma separated text file (.csv) by clicking on the "Save" button, or to copy the tab-separated contents to the system clip board by clicking on the "Copy" button and pasting the data into external software, for example, any text editor or Microsoft EXCEL.

H.3 LD measures to display: click on this menu item to choose whether to display D' or r^2 in LD matrix.

H.4 Save image: click on this menu item to save the entire chromosome view panel as an image file (Figure 3.3.2.b-6). Four image formats are supported: bmp, jpg, png, or eps (for publishing).

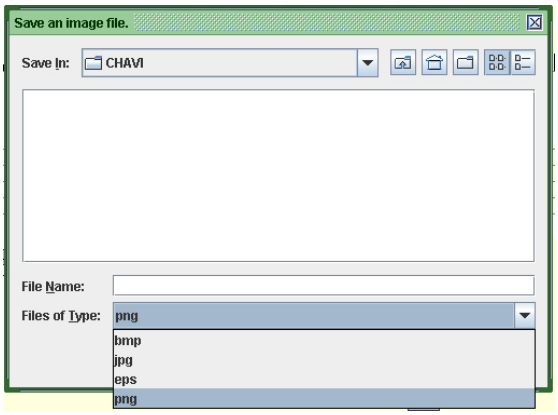


Figure 3.3.2.b-6. Results of comprehensive annotation for top hits: rs9264942, genic view, save image.

(3.3.2.c) SNP view (Figure 3.3.2.c-1)

Click on tab "SNP" to display the SNP view panel (Figure 3.3.2.c-1). This panel shows the annotation of the selected SNP (in this case the selected top hit) with LD extension, recent selection score, gene context, and association with gene expression levels. It consists of 8 parts:

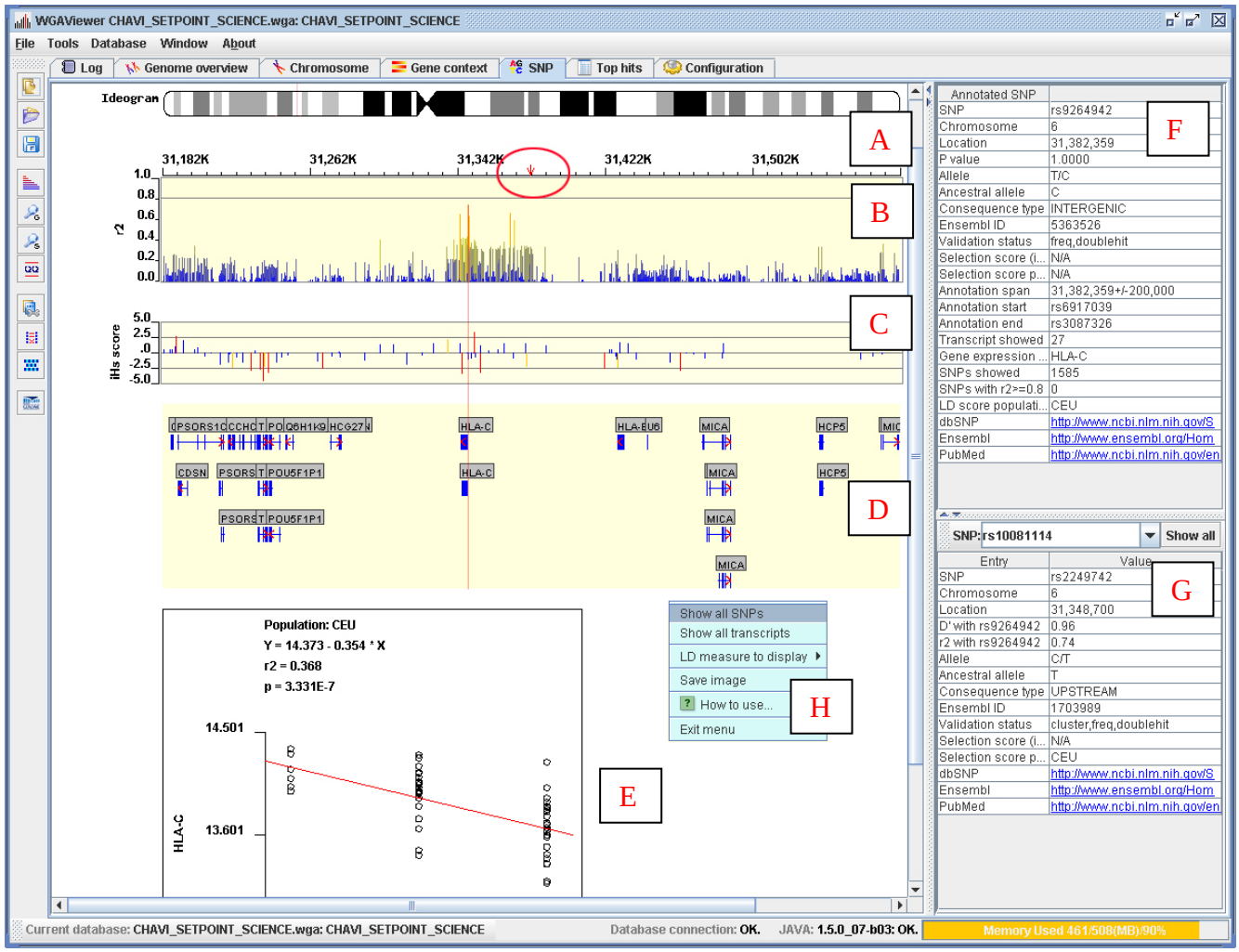


Figure 3.3.2.c-1. Results of comprehensive annotation for top hits: rs9264942, SNP view.
A: Chromosome ideogram; B: LD extension; C: Recent selection score; D: transcripts; E: association with (the closest) gene expression levels; F: Description for annotated SNP; G: Dynamic data sheet; H: Popup menu.
Part E shows a strong correlation between the annotated top hit, rs9264942, and HLA-C expression levels in immortalized B-lymphocytes. Circled red arrow denotes the physical location of the annotated SNP (rs9264942). The red highlight line through part B to D represents rs2246742, which is in LD with the annotated SNP rs9264942 (r2=0.74) and is also located in promoter region of HLA-C.

A: Chromosome ideogram: shows the annotated region on a chromosome with a transparent red rectangle;

B: LD extension: shows the pair-wise LD extension between the annotated top hit and all the available HapMap SNPs in the annotated region defined by Figure 3.3.2-1 (200Kbp up- and down-stream in this example). The color scheme for r2 is the same with gene view, LD matrix (3.3.2.a, Figure 3.3.2.b-1, part E): blue 0-0.2; yellow 0.2-0.6; red 0.6-1.0. Missing values are coded as -9 and plotted in gray. Each line will respond to mouse movement and show the detailed information, including the pair-wise LD

scores, for each HapMap SNPs in the dynamic data sheet (part G). The LD cutoff is 0.8 by default, but can be set before the annotation (see: section 3.3.2, Figure 3.3.2-1).

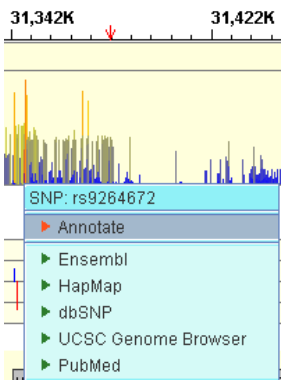


Figure 3.3.2.c-2. Resource menu for SNPs

Click on each SNP LD score line to bring up a resource menu (Figure 3.3.2.c-2). For lines too dense to easily pick up by mouse movement, **press key "<"/","** to move the highlighted lines backward (left, towards smaller chromosome coordinates), or **key ">"/"."** to move the highlighted lines forward (right, towards larger chromosome coordinates), and then **press key "enter"** to bring up this menu.

C: Available recent selection score (Voight et al. 2006) for HapMap SNPs shown in part B;

D: Transcripts: show transcripts located in the annotated region. Exons are plotted as blue rectangles with a red arrow representing the strand. Each transcript will respond to mouse movement and will show the detailed information for each SNP, together with the hyperlink to external databases, in an dynamic data sheet (part G). Alternative transcripts are plotted with detailed exon information shown in the dynamic data sheet too (part G).

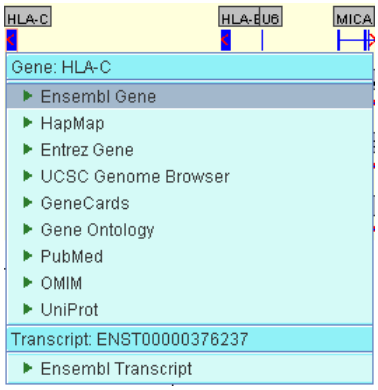


Figure 3.3.2.c-3. Resource menu for transcripts

Click on each transcript to bring up a resource menu (Figure 3.3.2.c-3). For transcripts too dense to easily pick up by mouse movement, **press key "<"/","** to move backward (left, towards smaller chromosome coordinates), or **key ">"/"."** to move forward (right, towards larger chromosome coordinates), and then **press key "enter"** to bring up this menu.

E: Association test between genotype and gene expression: This test is based on the genome-wide gene expression database from Sanger Institute, GENEVAR project (Stranger et al. 2005; Stranger et al. 2007) and genotype data from HapMap database (The International HapMap Consortium. 2005). The gene expression has been quantified in immortalized B-lymphocytes. As shown from the example dataset in Figure 3.3.2.c-1, there is a strong correlation between one of the top hits, rs9264942, and HLA-C expression levels (Fellay et al. 2007). This creates an immediate working hypothesis that this genetic variation might function through controlling the expression levels of the HLA-C. This immediately leads to the functional follow up and the hypothesis has been verified by data from other independent cohorts.

F: Description of annotated SNP: shows the detailed information for the annotated SNP. If the annotated SNP (top hit) tags any other HapMap SNPs in the specified HapMap population with the specified pair-wise r2 cutoff, the tagged SNPs and the r2 values will be listed in this data sheet too. This data sheet also shows a summary of the spanning region for annotation, including annotation span, start and end landmark, the number of SNPs and transcripts that are showed, etc. If the SNP annotation is based on the exact same region of gene view, annotation span will be shown as "0" in this data sheet. Different from the dynamic data sheet (part G), the contents of this data sheet are fixed.

G: Dynamic data sheet: Similar to the dynamic data sheet in gene view panel (Figure 3.3.2.a-1). This sheet shows the detailed information for the highlighted item in parts B-D. Therefore the contents of this data sheet will change according to which type of item is highlighted in the main graphical region. Like the dynamic data sheet in gene view, this data sheet has also a fixed tool bar including a drop-down menu for all the HapMap SNPs shown in part B, sorted by the rs#. One has the option to select a SNP and show the detailed information in the dynamic data sheet. A red highlight line will also then be plotted on part B to D to show which SNP has been selected. Sometimes this is more convenient than directly pointing the mouse to a specific SNP in part B, because when the SNP density is higher it is difficult to conveniently highlight a SNP among the overlapped lines. For SNP and transcript, this data sheet always offers hyperlinks to external databases, including Ensembl and NCBI, for a convenient reference for data not shown.

H: Popup menu: Similar to the popup menu in gene view panel (Figure 3.3.2.b-1). Clicking on blank region (other than hotspots, for example SNPs or transcripts) will activate this popup menu. It can then be dismissed by click on menu item "Exit". Similarly, this popup menu offers four functions:

H.1 Show all SNPs: click on this menu item will activate a popup window and show the detailed information for all the HapMap SNPs (Figure 3.3.2.c-1) plotted in part B, instead of one by one in the dynamic data sheet (part G). This data sheet window offers five methods to sort the SNP collection: by location, by r2, by D', by rs #, or by type. Select the sorting method and then click on button "Sort by". Sorting by r2 is probably the most useful one to quickly find the tagged SNPs in this region by the annotated top hit.

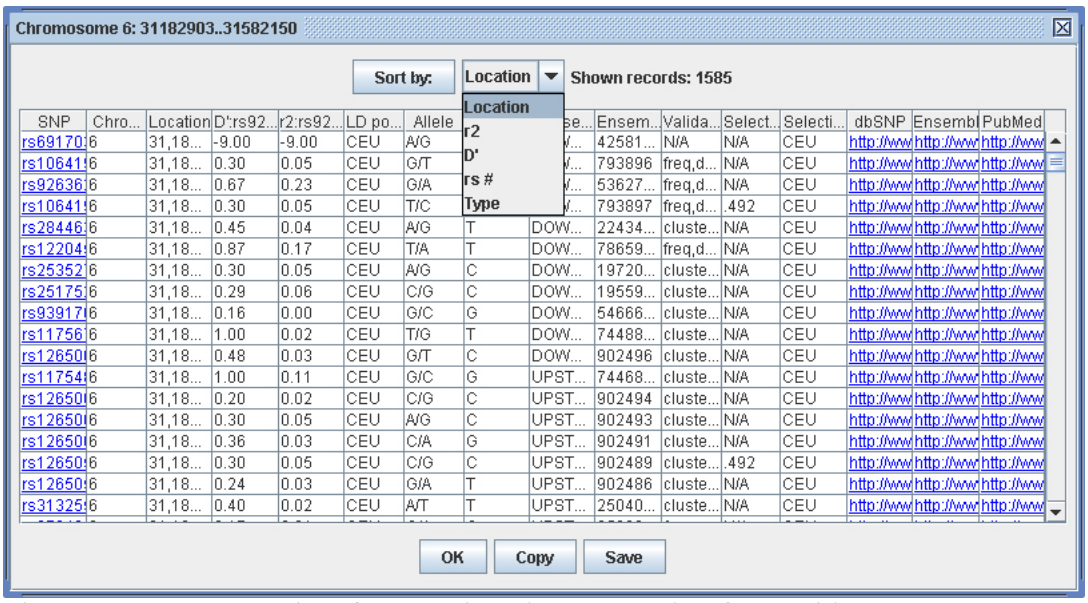


Figure 3.3.2.c-4. Results of comprehensive annotation for top hits: rs9264942, SNP view, data sheet for showing all HapMap SNPs.

Very similar to the data sheet window in gene view (Figure 3.3.2.b-4), this data sheet window also offers the clickable hyperlink navigating to dbSNP, Ensembl, and PubMed. If one clicks on rs# for each SNP, a red highlight line will be drawn on part B to D to indicate the location of the selected HapMap SNP. To dismiss this window, click on the "OK" button. One also has the option to save the contents of this table to a comma separated text file (.csv) by clicking on the "Save" button, or to copy the tab-separated contents to system clip board by clicking on the "Copy" button and pasting the data into external software, for example, any text editor or Microsoft EXCEL.

H.2 Show all transcripts: clicking on on this menu item will activate a popup window that shows the detailed information for all the transcripts (Figure 3.3.2.c-5) plotted in part D, instead of one by one in the dynamic data sheet (part G). This function is equivalent to the one in gene view (see 3.3.2.b, H2, Figure 3.3.2.b-5.).

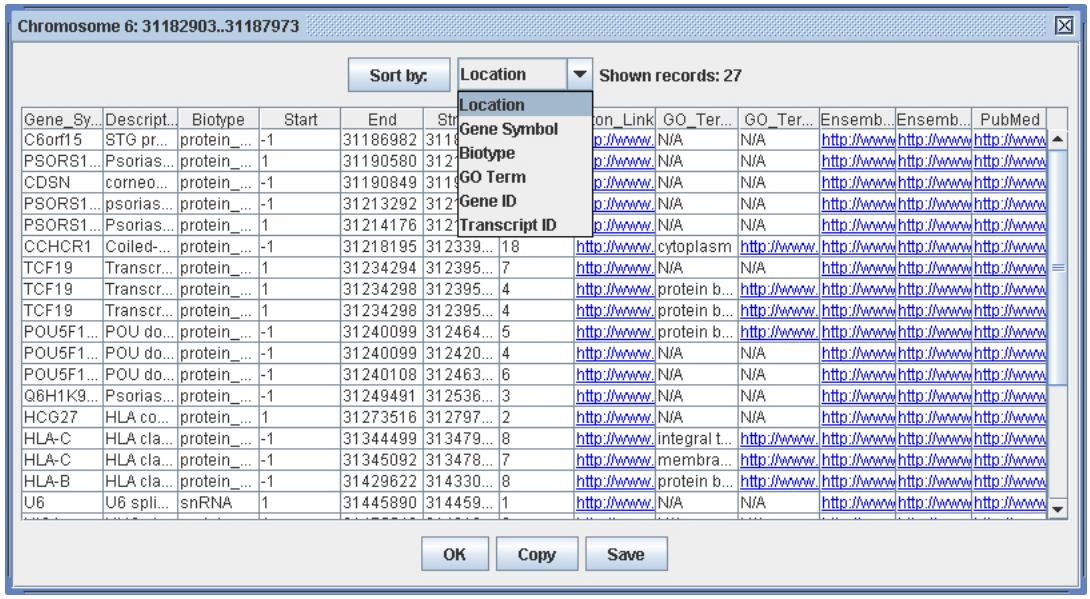


Figure 3.3.2.c-5. Results of comprehensive annotation for top hits: rs9264942, SNP view, data sheet for showing all transcripts.

H.3 LD measures to display: click on this menu item to choose whether to display D' or r^2 in LD extension.

H.4 Save image: This function is similar to the function described in gene view (see: section 3.3.2.b, H4, Figure 3.3.2.b-6). The difference is this function will save the SNP view as an image file instead.

(3.4) Finding a gene

After the check for the top hits, another natural question is to assess the evidence of association in particular genes which may have strong priori probability of influencing a phenotype (for example because they have been associated to the phenotype in other studies,, or because the gene is a target of a drug in a pharmacogenetic study, and so on).

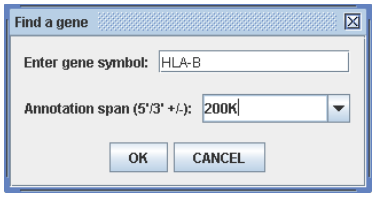


Figure 3.4-1. Parameters for finding a gene.

To manually answer this question is time-consuming and error-prone. The first problem is, are the physical coordinates listed in the MAP file used for the WGA project up to date? What version of genome build are they based on? Can they be directly used to align with the gene coordinates in the currently available public databases? And, further, what version of genome build are the public databases based on? Another issue is how to consider all of the alternative transcripts.

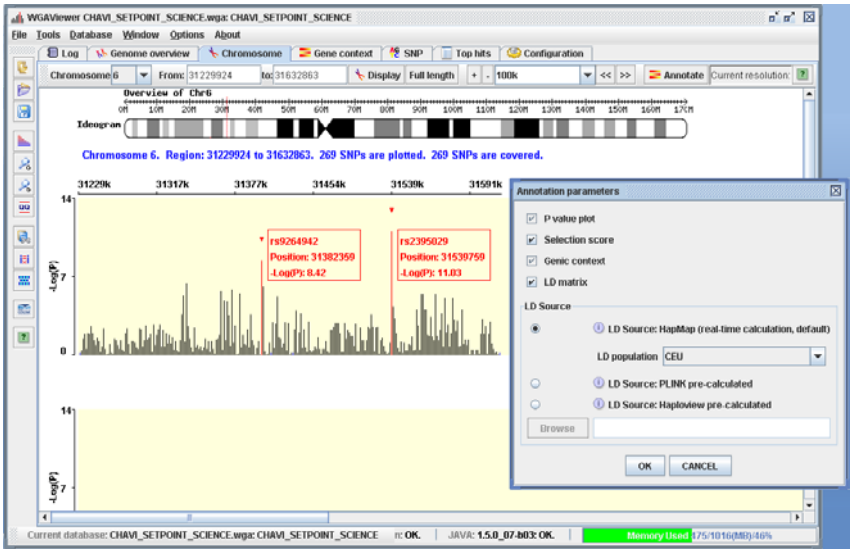


Figure 3.4-2. Parameters for annotating chromosomal region.

WGAViewer offers an accurate and efficient way to solve these problems by always applying the latest available genome build coordinates to all data that will be used to align with each other. More specifically, all the annotated coordinates listed as annotation results by this software are always based on the latest Ensembl Core, Variation, and GO databases. Every coordinate from other sources, for example, the coordinates input from the MAP file, or coordinates from HapMap database (based on NCBI b36, dbSNP b136 as in Aug, 2007) will always be compared with core databases and then aligned. This enables an accurate annotation and solves the problem of discrepancies of genome builds from different sources, especially when trying to align the WGA result set with gene context. Furthermore, this process takes advantage of the efficient monthly updating system of Ensembl.

For alternative transcripts, WGAViewer reports the longest form, alongside all the available short forms based on the Ensembl core database, plus the up- and down-stream span specified by the user when trying to locate a gene and annotate the related genomic region.

For example, HLA-B variants have been related to HIV-related phenotypes in a number of previous publications. To find HLA-B in the WGA dataset, click on menu "**Tools->Find a gene**". This will bring up dialogs for parameters (Figure 3.4-1, 3.4-2). One has the options to exclude the annotation for selection score and LD matrix, and the options for the sources of LD calculation.

Figure 3.4-3 shows the annotation results for finding HLA-B in the example dataset. As shown, there seems no particularly interesting finding located within HLA-B. However, not far from HLA-B, there are two genome-wide significant hits, one is close to HLA-C (rs9264942), the other is located within HCP5 (rs2395029).

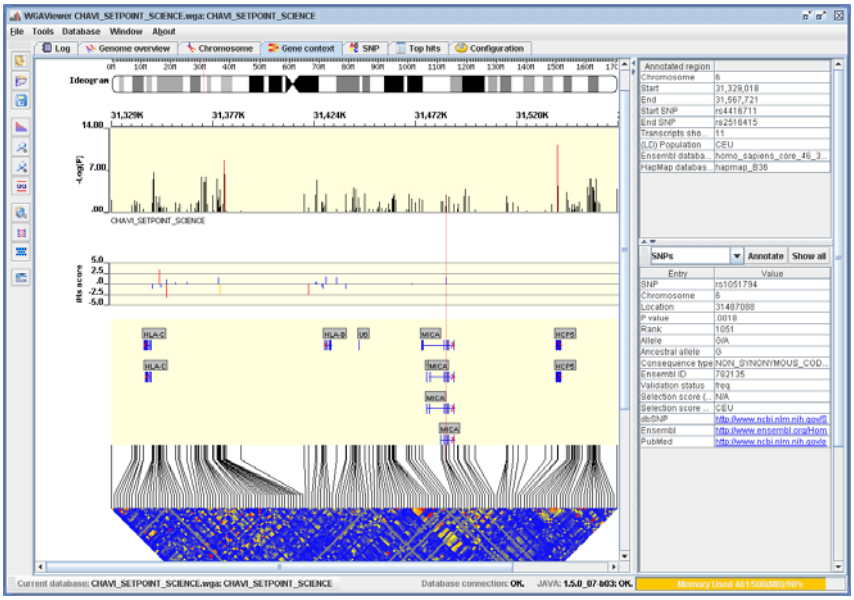


Figure 3.4-3. Finding a gene: annotation results.

(3.5) Finding a SNP

Very similar to the questions we asked when finding a gene, it is often important to assess association for individuals SNPs. Furthermore, if such SNPs are not present in the WGA dataset, it would then be important to know whether there are LD proxies for these SNPs, and what are their P values.

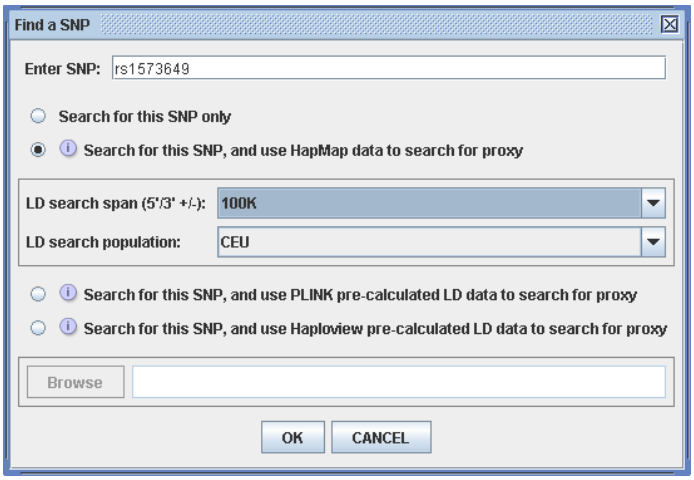


Figure 3.5-1. Parameters for finding a SNP.

We take the pre-annotation dataset "CHAVI_SETPOINT_SCIENCE. wr" as an example. Suppose we would like to know about the results for an interesting SNP rs1573649, which is a nonsynonymous coding SNP located on exon 1 of HLA-DQB2 gene. Clicking on menu "**Tools->Find a SNP**" will activate a dialog for annotation parameters (Figure 3.5-1). In this example we use the HapMap database to search for the LD proxies. The LD search span is the region around the target SNP that will be checked for proxies. This process will then lead to an annotation result as shown in Figure 3.5-2, where it is explicitly shown that not only the target SNP itself (rs1573649), but also its proxies (for example, rs2301271, r2=0.817) were genotyped in the WGA project, and none are significantly associated with the set point phenotype.

One has the option to search for the LD proxies using pre-calculated LD dataset by Haploview or PLINK, instead of using the HapMap data. There is also an option to search for specific SNPs only, not including their LD proxies (Figure 3.5-1). Both of these two options do not require an internet access.

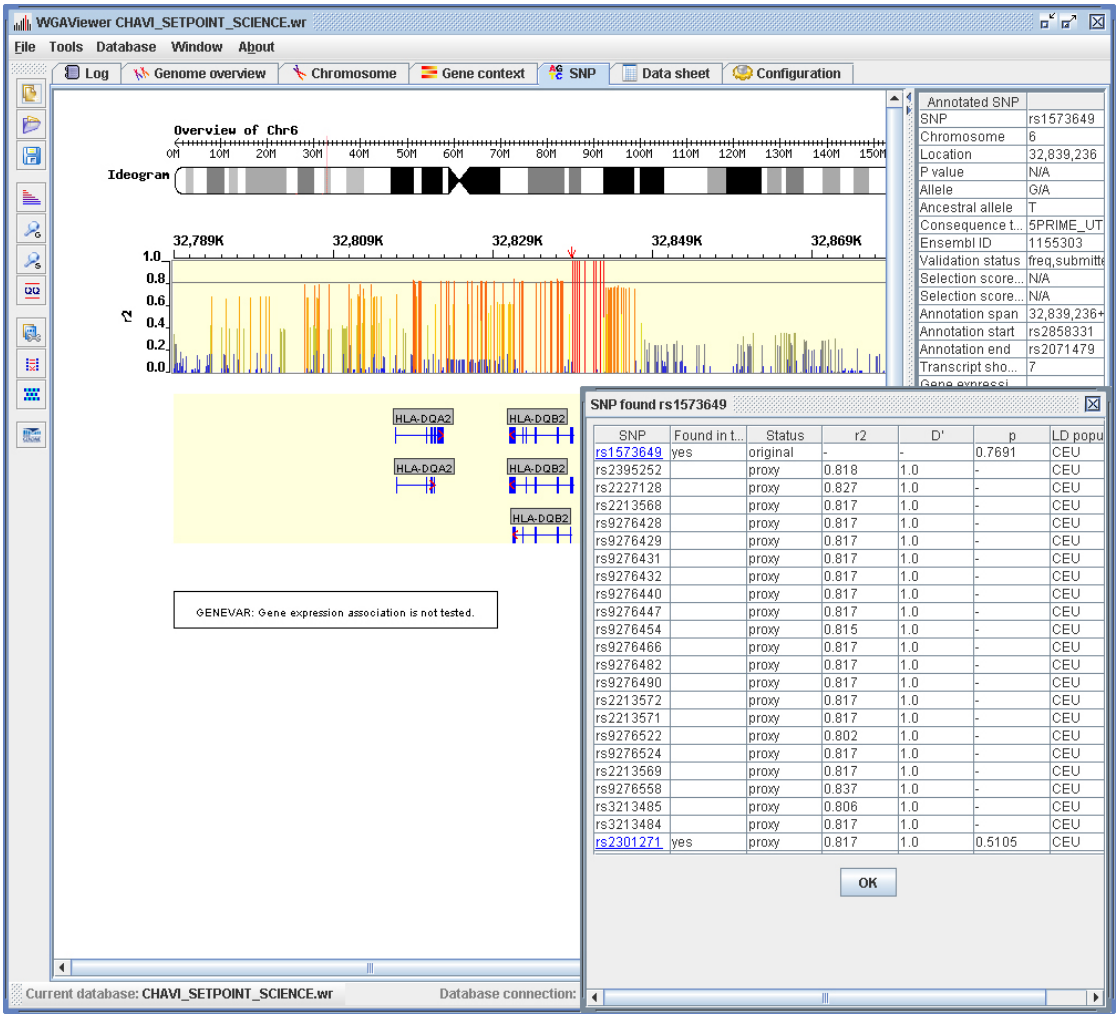


Figure 3.5-2. Finding a SNP: annotation results.

(3.6) From chromosome to SNP: another way to look at the data.

In the above sections from 3.3 to 3.5 we discussed the most interesting annotation pathways to most WGA analyzers, i.e., the annotation for top hits, as well as the searching for specific genes and SNPs. Alternative to this, WGAViewer offers a more general and intuitive look at the WGA data. This pathway, as illustrated in Figure 3.6-1, starts from genome overview, goes through chromosome view and genic view, and finally annotates selected SNPs.

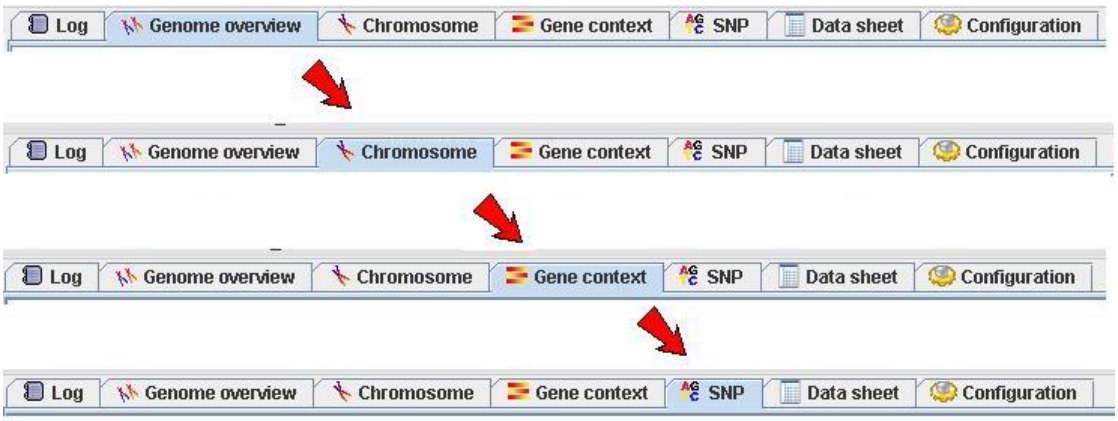


Figure 3.6-1. From chromosome to SNP.

Take our pre-annotation set "CHAVI_SETPOINT_SCIENCE.wr" as an example. Figure 3.6-2 shows a **genome overview** after this dataset is successfully loaded.

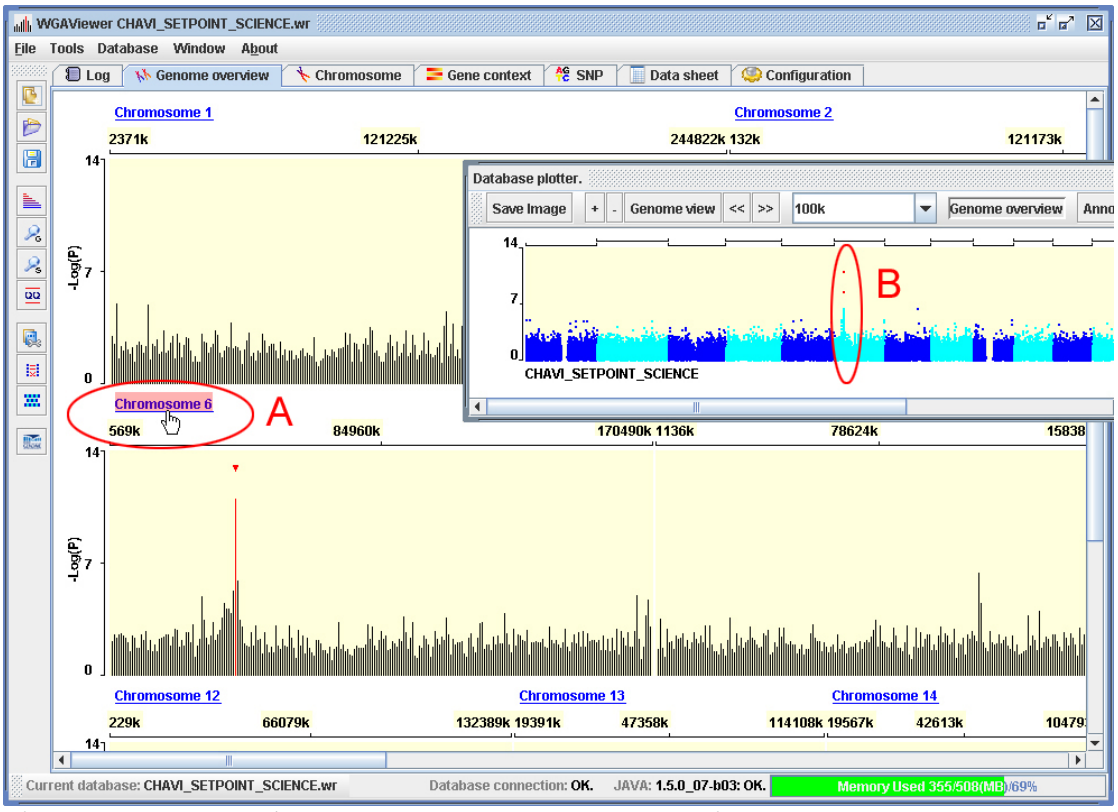


Figure 3.6-2. From chromosome to SNP: genome overview.

Tab "Genome overview" (panel with "A" marked) plots P values as evenly-spaced lines. WGAViewer will automatically determine a resolution and only plots the lowest P values in 'n' adjacent data points. In this tab panel, every data line will respond to mouse movement and will show a small stick presenting the SNP position and P value. Every chromosome label (circled and marked as "A") will respond to

mouse click and will bring the user interface into chromosome view (Figure 3.6-3). Alternative to this panel, the database plotter plots all data points (marked as "B"). Each portion of the plotting region can also respond to a mouse click, which will then result in an improved resolution in the database plotter window. Clicking on the "annotate" button in this database plotter window with a high enough resolution can also switch the user interface into chromosome view (Figure 3.6-3).

Either way, as a consequence, the **chromosome view** is then brought forward (Figure 3.6-3). The –logP lines in this panel are also evenly-spaced and can respond to mouse movement and click. One has the option to either drag the mouse and select a region, or type in the chromosome region into the text field in the tool bar and click on button "Display", to get an improved resolution. With a reasonable resolution, for example, including around 200 SNPs as shown from Figure 3.6-3 (too many SNPs will take a long time to annotate, especially for LD calculation), one then may click the "Annotate" button to annotate this region.

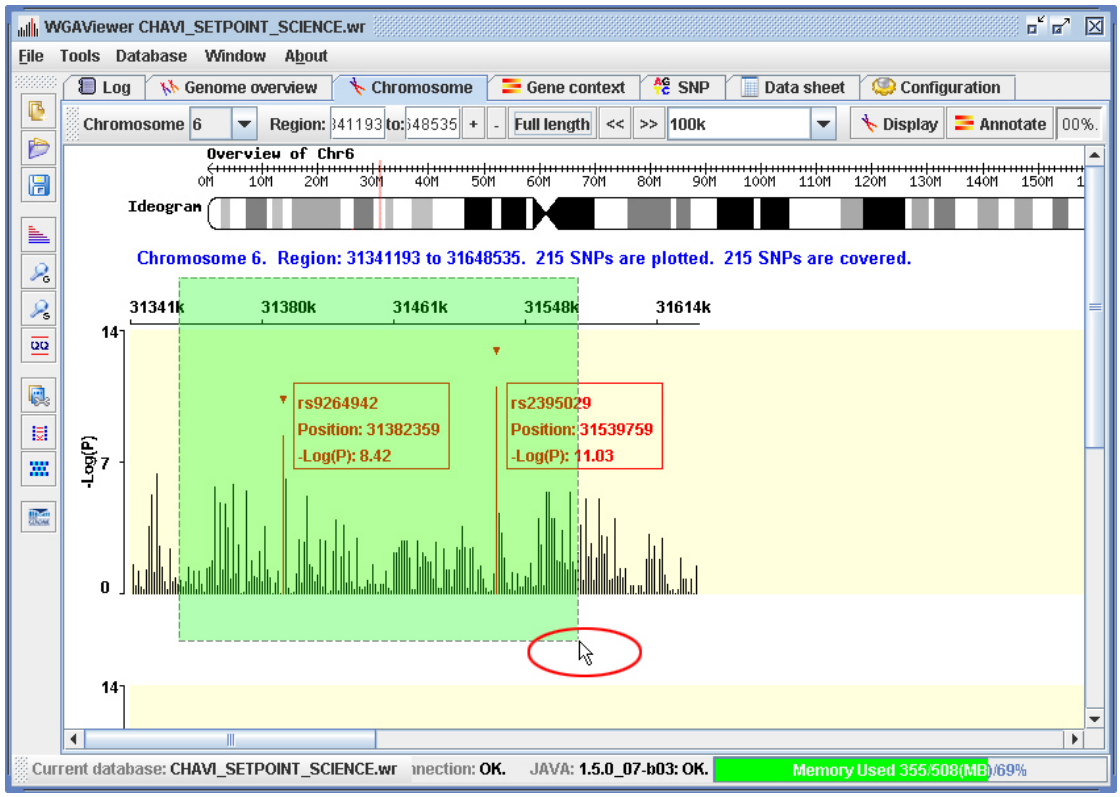


Figure 3.6-3. From chromosome to SNP: chromosome view.

Figure 3.6-4 shows the annotation results in **genic view** panel. The details of this panel have already been discussed (Figure 3.3.2.a-1). One has the options to either click on SNP P value lines and then click on item "Annotate SNP" in the pop-up menu, or select a SNP from a drop-down menu and click on "Annotate" button to annotate a SNP (circled in Figure 3.6-4). And as a consequence, **SNP view** panel (Figure 3.6-5) becomes the focus showing the annotation results, which then enables the further annotation for each HapMap SNPs, or one may go back to chromosome view or genic view to annotate other WGA SNPs. The details of SNP view have also been discussed (Figure 3.3.2.c-1).

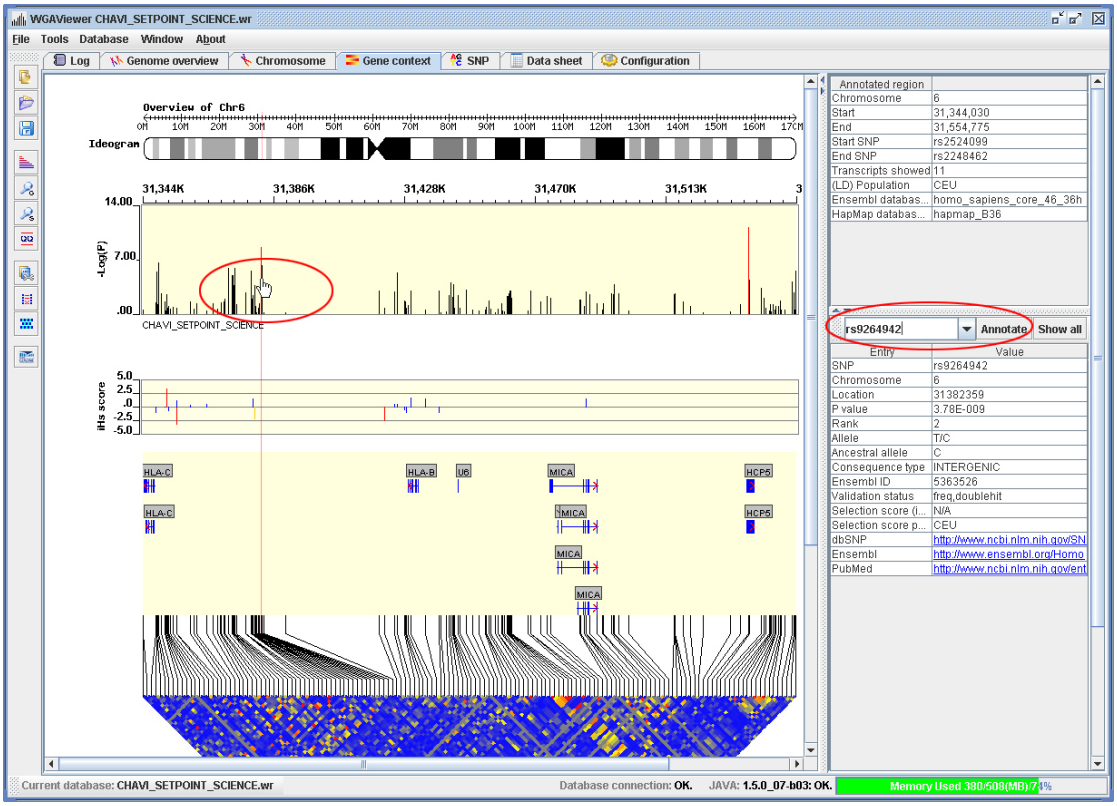


Figure 3.6-4. From chromosome to SNP: genic view.

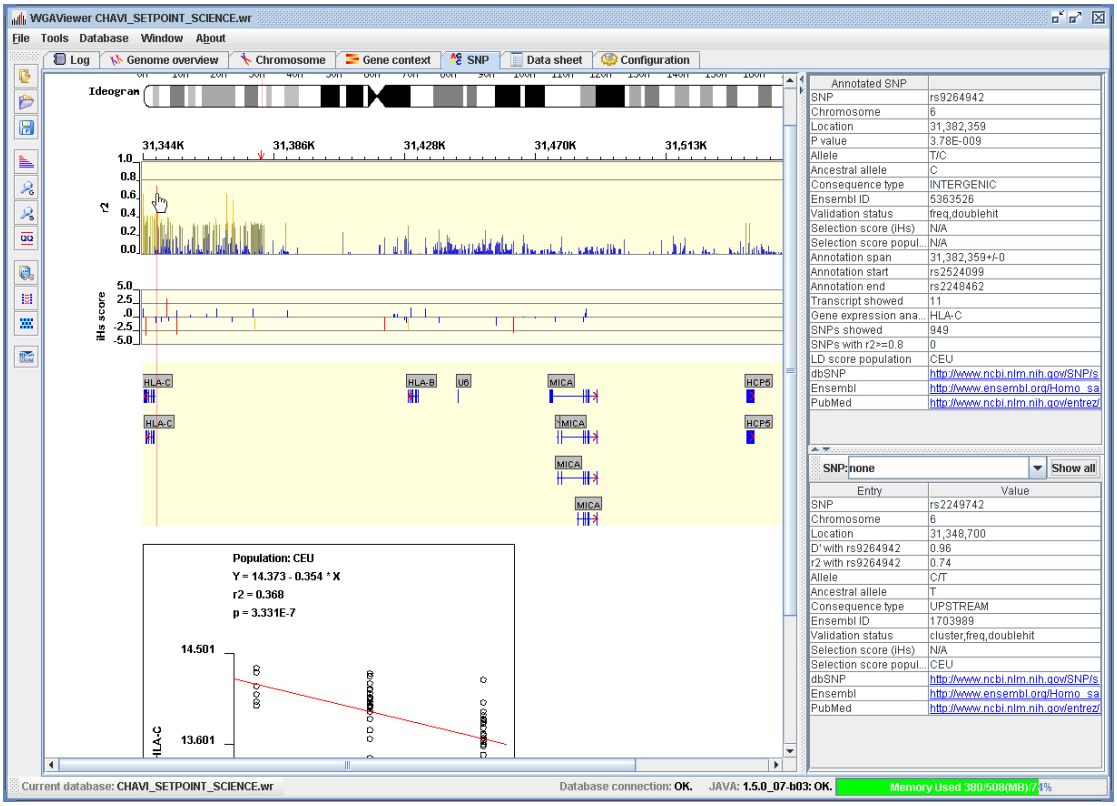


Figure 3.6-5. From chromosome to SNP: SNP view.

(3.7) Effect of population stratification

WGAViewer offers an easy graphical way to inspect the possible population stratification effects by comparing the distribution of observed P values with expected distribution through a Q-Q plot. WGAViewer also calculates a lambda value for quantifying the population stratification effects assuming a 2 df chi distribution of the -2log(e)P measures (Clayton et al. 2005) (Weale 2007). This feature can be especially useful for confirming that methods for controlling stratification have been properly implemented.

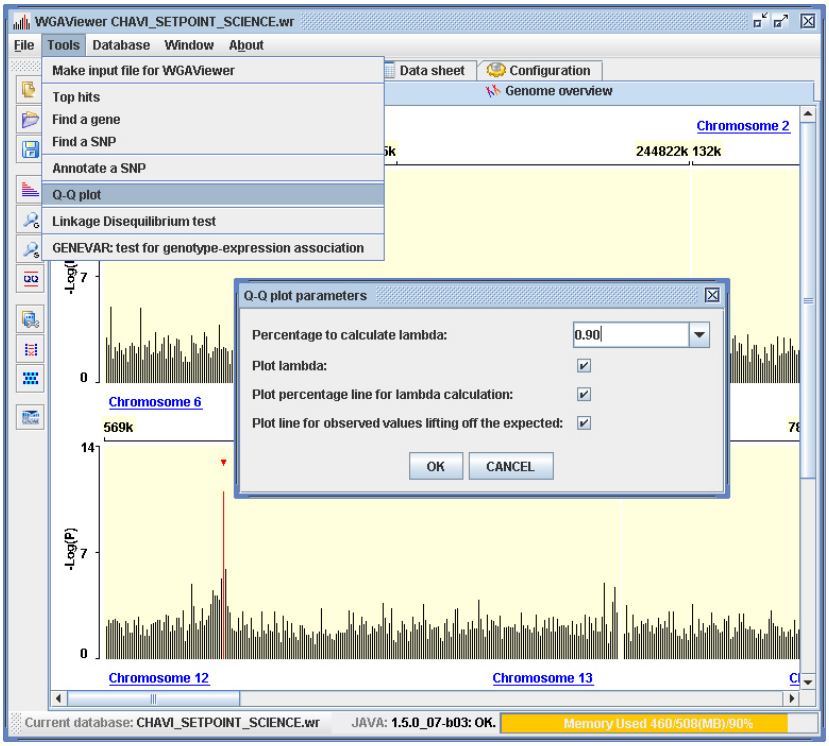


Figure 3.7-1. Parameters for Q-Q plot.

To perform this function, click on menu "**Tools->Q-Q plot**". This will activate a dialog for plotting parameters (Figure 3.7-1). One has the option to plot percentage line and lifting line to help inspect the P value distribution. Figure 3.7-2 shows the plot result. The lower red line denotes the 90th percentile, while the upper one indicates the point where the P values lift -up line from the expected line. As discussed in (Fellay et al. 2007), we applied a principal component method (Price et al. 2006) to control for population stratification effects. The P value distribution here therefore takes account of the correction. As from Figure 3.7-2, it is clearly shown that over 90% of the P values distribute in accordance with random expectation, only 323 P value data points lift from the expectation distribution. The lambda value of 1.0053 also indicates the residual population stratification effect (after correction) is minimal.

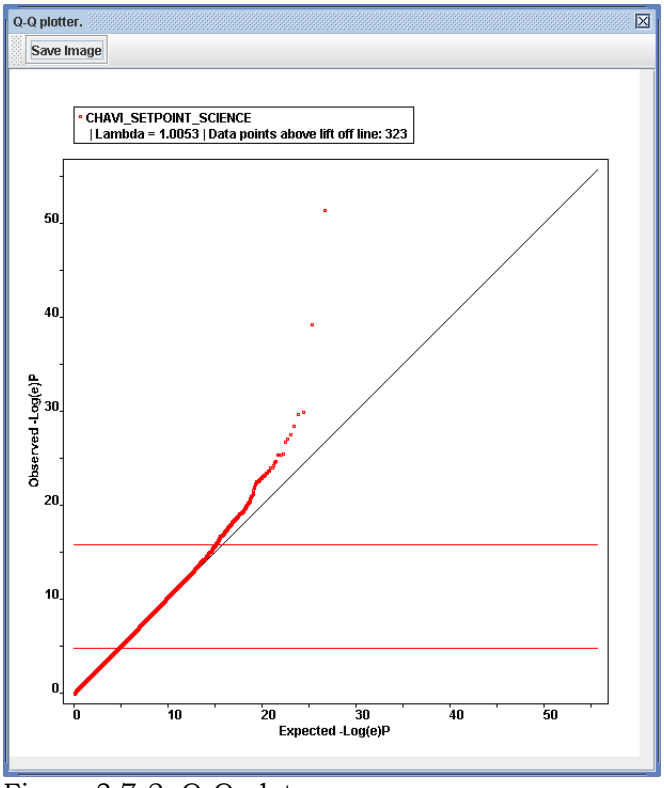


Figure 3.7-2. Q-Q plot.

Another usage of this plot is to easily inspect how the top hit P values depart from the random distribution, in addition to any statistical method used for correcting for multiple testing. As from Figure 3.7-2, the top two data points (rs2395029, HLA-B/HCP5 and rs9264942, HLA-C) can be clearly distinguished from any other points plotted. These two are the two genome-wide significant hits that we reported for setpoint phenotype.

(3.8) Multiple databases

This function suite has been designed to easily compare between multiple databases. These databases can be replication cohorts, cohorts with the same phenotype but different SNP sets, cohorts with related phenotypes, or the same cohort with different phenotypes. The aim of this function is to easily obtain concurrent or supporting evidence.

To illustrate this function, we add a simulated dataset (Illumina_HumanHap300_sim.wr) into the loaded CHAVI_SETPOINT_SCIENCE.wga set. Each database to be loaded has to be first transformed into WGAViewer .wr file as discussed in section "data input".
Click on menu "**Database -> Database manager**". And then load the simulated dataset from folder "./examples" by clicking "**Add a reference dataset**" button in the interface of database manager (circled in Figure 3.8-1).

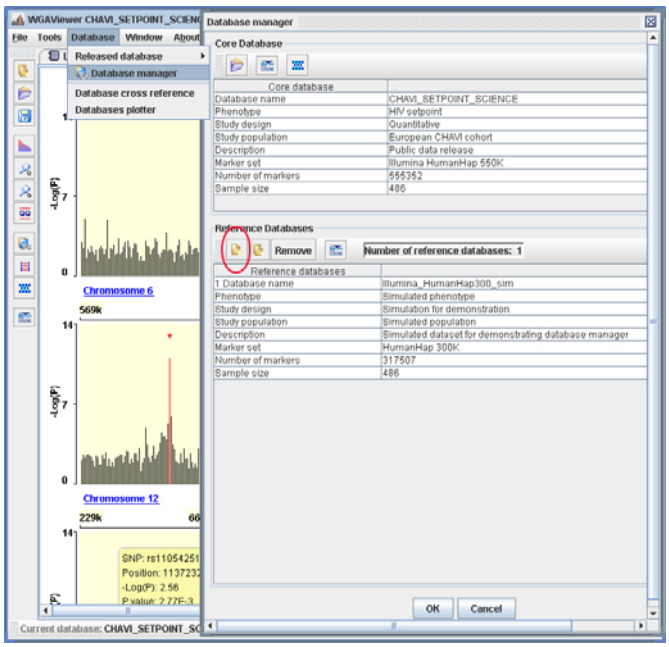


Figure 3.8-1. Loading a supporting database.

One has the option to add or remove specific reference databases to list alongside with the core database. One also has the option to edit the description for each database as shown from Figure 3.8-1.

(3.8.1) Multiple databases plotting

Once reference database(s) is successfully loaded, click on menu "**Database->Database plotter**" to plot the multiple databases, as shown from Figure 3.8.1-1. Zoom in or zoom out by moving and clicking the mouse, or through a navigation bar located on the top of the window. With appropriate resolution (for example, when number of SNPs in this region is less than 500 to ensure reasonable speed) click on the button "**Annotate**" to annotate this region.

Figure 3.8.1-2 shows the annotation results. Different from the gene view panel for a single database, this view plots the P values (part B) for the reference database(s) too. In addition to this, for each SNP that is highlighted in part A, the P values in the core database and in the reference databases (if available) will both be listed in the dynamic data sheet (part C).

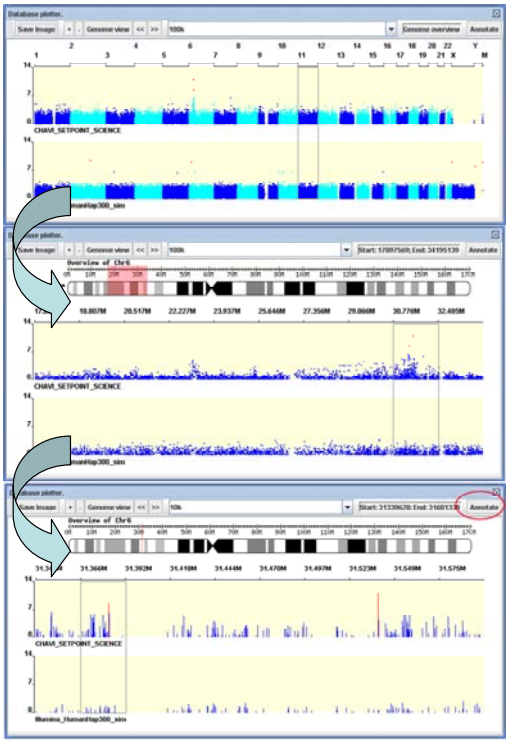


Figure 3.8.1-1. Database plotter.

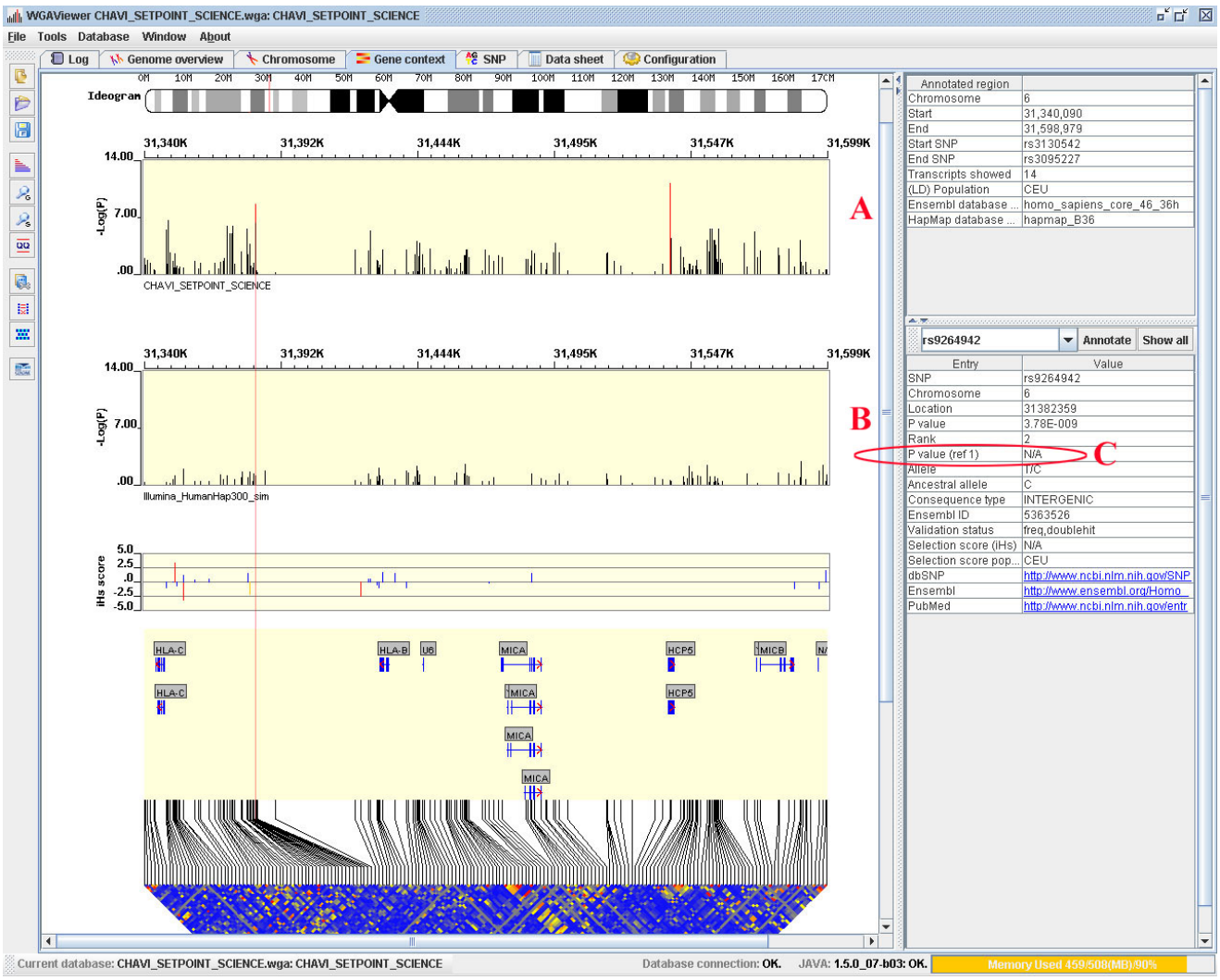


Figure 3.8.1-2. Multiple databases: annotation results on the level of gene view.
A: Association results for core database; B: Association results for reference database(s); C: P value(s) in reference databases (if available) for highlighted SNP in part A (shown in this example is rs9264942).

(3.8.2) Searching for association across multiple databases

It is reasonable to assume that some polymorphisms will influence multiple related phenotypes, for example a single polymorphism might influence multiple autoimmune disease, resistance to different infectious agents, or different neuropsychiatric conditions. For this reason it is convenient to be able to identify polymorphisms that show evidence of association, at a user defined threshold, in two or more datasets.

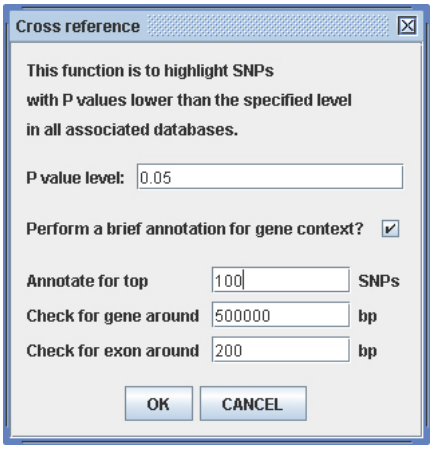
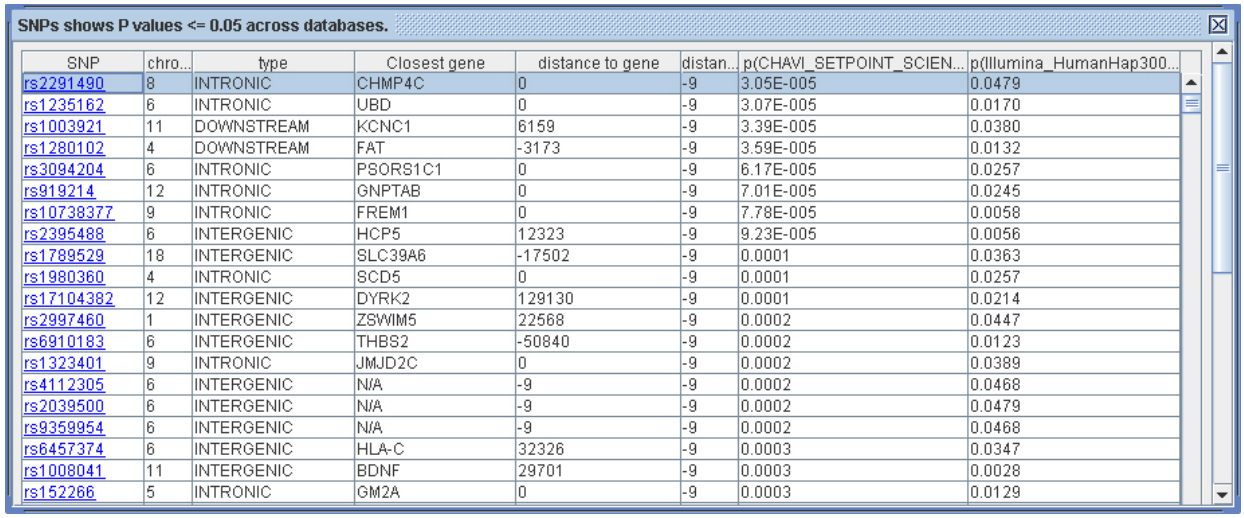


Figure 3.8.2-1. Parameters for cross reference searching.

To do this, click on menu “Database -> Database cross reference” and a dialog for searching parameters will present itself (Figure 3.8.2-1). The most important parameter here that needs to be specified should be the P value level that will be used as a cutoff for including into the report. In the example shown (Figure 3.8.2-1), only SNPs with P values equal to or lower than 0.05 in **all** core and reference databases will be included in the consequent report. The report will be sorted by the P

values in the core dataset. One can also specify a brief annotation on a subset (or all) of the cross-referenced SNPs, for example, annotating the top 100 (sorted by core dataset P values) as shown from Figure 3.8.2-1.

This will then give a report as illustrated in Figure 3.8.2-2. The columns from left to right are: SNP name, chromosome, SNP type, the closest gene, the distance to the closest gene, the distance to the closest exon, P values in the core dataset, and P values in the reference dataset. The SNP name is clickable and will launch an annotation for the surrounding chromosome region.



SNPs shows P values <= 0.05 across databases.

SNP	chro...	type	Closest gene	distance to gene	distan...	p(CHAVI_SETPOINT_SCIEN...	p(Illumina_HumanHap300...
rs2291490	8	INTRONIC	CHMP4C	0	-9	3.05E-005	0.0479
rs1235162	6	INTRONIC	UBD	0	-9	3.07E-005	0.0170
rs1003921	11	DOWNSTREAM	KCNC1	6159	-9	3.39E-005	0.0380
rs1280102	4	DOWNSTREAM	FAT	-3173	-9	3.59E-005	0.0132
rs3094204	6	INTRONIC	PSORS1C1	0	-9	6.17E-005	0.0257
rs919214	12	INTRONIC	GNPTAB	0	-9	7.01E-005	0.0245
rs10738377	9	INTRONIC	FREM1	0	-9	7.78E-005	0.0058
rs2395488	6	INTERGENIC	HCP5	12323	-9	9.23E-005	0.0056
rs1789529	18	INTERGENIC	SLC39A6	-17502	-9	0.0001	0.0363
rs1980360	4	INTRONIC	SCD5	0	-9	0.0001	0.0257
rs17104382	12	INTERGENIC	DYRK2	129130	-9	0.0001	0.0214
rs2997460	1	INTERGENIC	ZSWIM5	22568	-9	0.0002	0.0447
rs6910183	6	INTERGENIC	THBS2	-50840	-9	0.0002	0.0123
rs1323401	9	INTRONIC	JMJD2C	0	-9	0.0002	0.0389
rs4112305	6	INTERGENIC	N/A	-9	-9	0.0002	0.0468
rs2039500	6	INTERGENIC	N/A	-9	-9	0.0002	0.0479
rs9359954	6	INTERGENIC	N/A	-9	-9	0.0002	0.0468
rs6457374	6	INTERGENIC	HLA-C	32326	-9	0.0003	0.0347
rs1008041	11	INTERGENIC	BDNF	29701	-9	0.0003	0.0028
rs152266	5	INTRONIC	GM2A	0	-9	0.0003	0.0129

Figure 3.8.2-2. Reports for cross reference.

(3.9) Supporting/QC information

In addition to the annotation based on available data from public databases, many other interesting and important questions could be raised. For example, what are the Hardy-Weinberg Equilibrium (HWE) results for the top hits? What are the effect sizes, effect directions, etc, for these SNPs of most interest?

WGAViewer offers a convenient interface to allow the user to load many kinds of customized supporting/QC data to list alongside the SNP lists of interest, in addition to the annotation from public databases. We illustrate this function using our pre-annotation example set "CHAVI_SETPOINT_SCIENCE.wr" as the core dataset. Click on menu "File -> Open supporting/QC file" and a dialog will present itself (Figure 3.9-1). Select a database that the supporting file is associated with, which is CHAVI_SETPOINT_SCIENCE in this case. Click on button "Support/QC file" to locate the text file containing the supporting/QC dataset (a file containing HWE results for cases in Figure 3.9-1). And then select the column for the SNP, and the columns to be included as supporting/QC information (P values for HWE as in Figure 3.9-1). Multiple-column selection is allowed. Click on the "OK" button and if the data is successfully loaded, a message box as in Figure 3.9-2 will present itself. If not, there may be unsupported data lines in file, for example, inconsistent column counts.

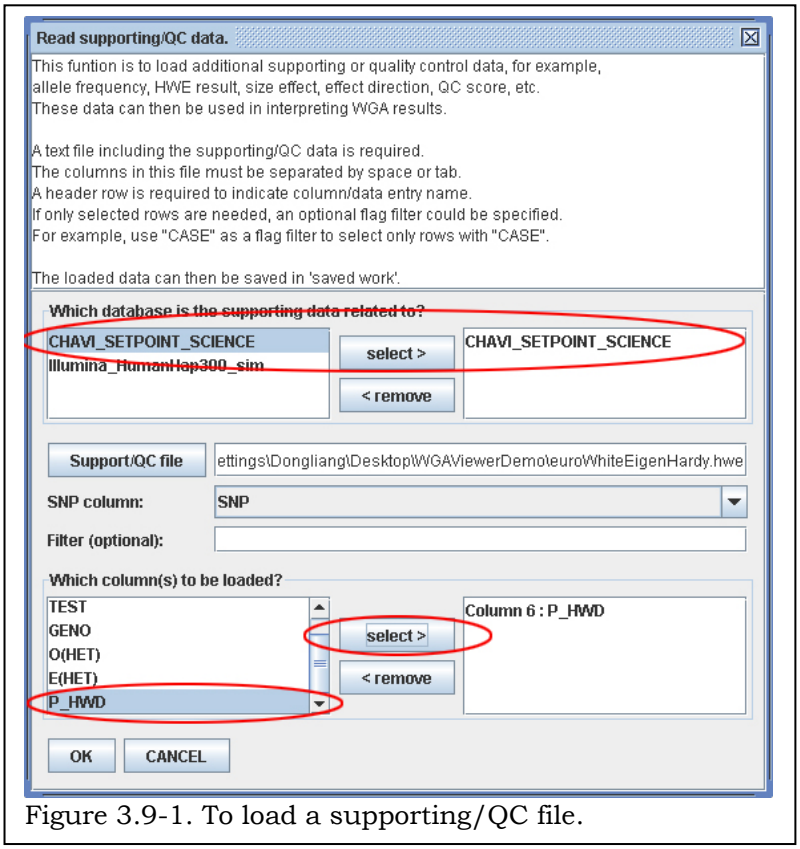


Figure 3.9-1. To load a supporting/QC file.

After the inclusion of supporting/QC files, the annotation for top hits (chapter 3.3) will automatically include columns for such information. For example, click on menu "Tools->Top hits" and annotate the top hits as discussed in chapter 3.3, the annotation results will be like Figure 3.9-3, with a column for HWE P value alongside each SNP.

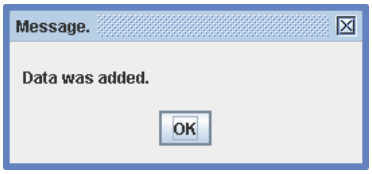


Figure 3.9-2. A supporting/ QC file is successfully

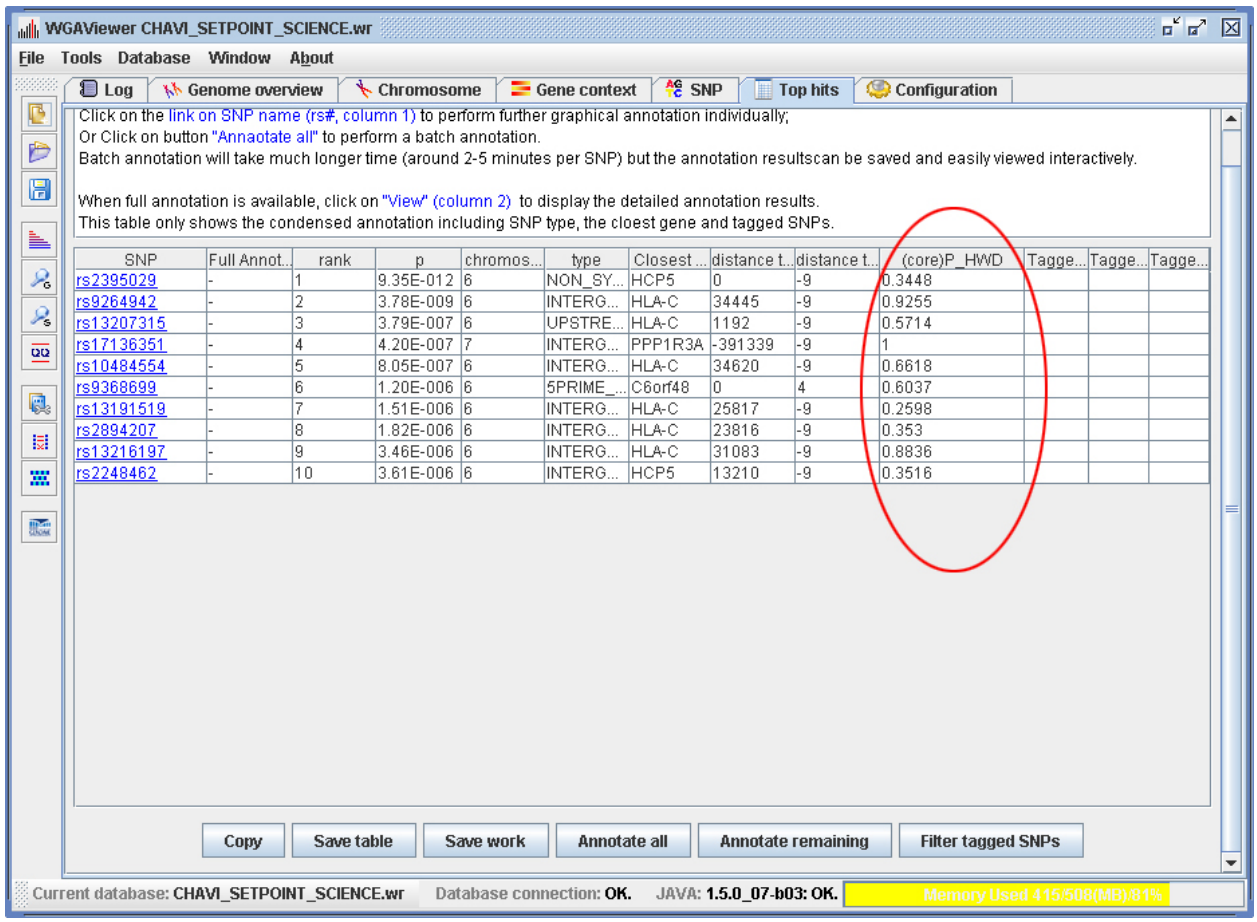


Figure 3.9-3. Briefly annotated top hits with supporting information. Circled are Hardy-Weinberg Equilibrium P values.

(3.10) Mart for IGSP Data from Association Studies (MIDAS)

WGAViewer offers an easy way to access the data for WGA or other association studies hosted by and released from Duke Institute for Genome Sciences & Policy. To access the Mart for IGSP Data from Association Studies (MIDAS), click on menu "Database->Released database->MIDAS online". Figure 3.10-1 shows the datasets that are released from Duke Institute for Genome Sciences & Policy. Simply click on button "Download" to download these datasets. The dataset are fully compatible with WGAViewer and will be automatically loaded after the download. Annotations discussed in this user's guide can then be performed.

WGAViewer and MIDAS use the query engine provided by the Biomart Data Integration System (http://www.biomart.org/). All the downloads are free of charge.

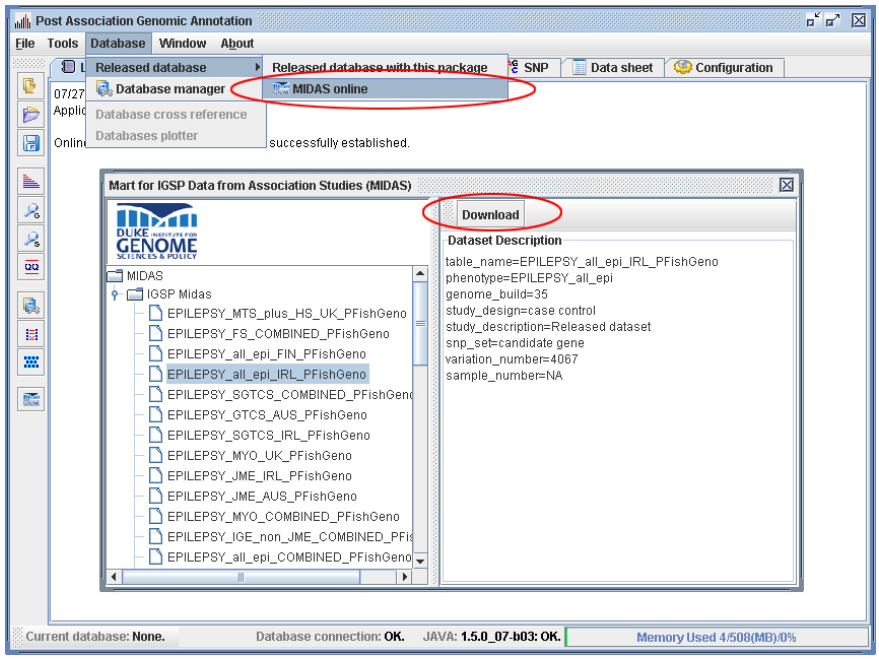


Figure 3.10-1. Mart for IGSP Data from Association Studies (MIDAS).

(3.11) Useful tools that do not require an association result set

In addition to the support and annotation for WGA result sets, WGAViewer also offers several classes of annotation tools that do not require such data sets. These include:

(3.11.1) Tests for association of SNP with gene expression

WGAViewer offers a convenient tool to easily test association between SNP genotypes and gene expression data, quantified in immortalized B-lymphocytes, using the databases from the Sanger Institute GENEVAR project (Stranger et al. 2005; Stranger et al. 2007) and the HapMap project (The International HapMap Consortium. 2005).

To do this, click on menu "Tools-**>GENEVAR: test for genotype-expression association**" and this will activate a dialog for inputs (Figure 3.8.1-2). Two types of inputs are supported: (1) directly typing in SNP rs# and gene symbol, as shown from Figure 3.11.1-1; or (2) selecting a text file containing the SNP-gene pairs between which association will be tested, for example, as shown in Box 3.11.1-1.

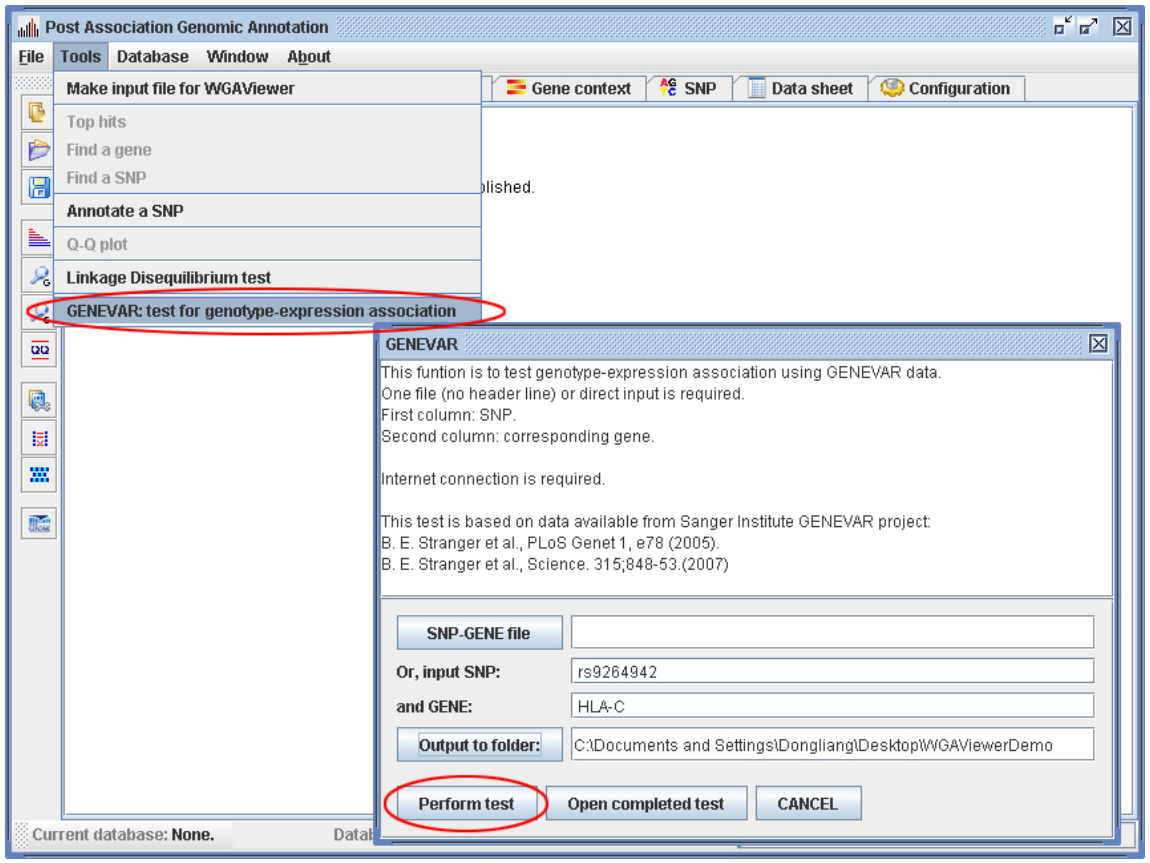
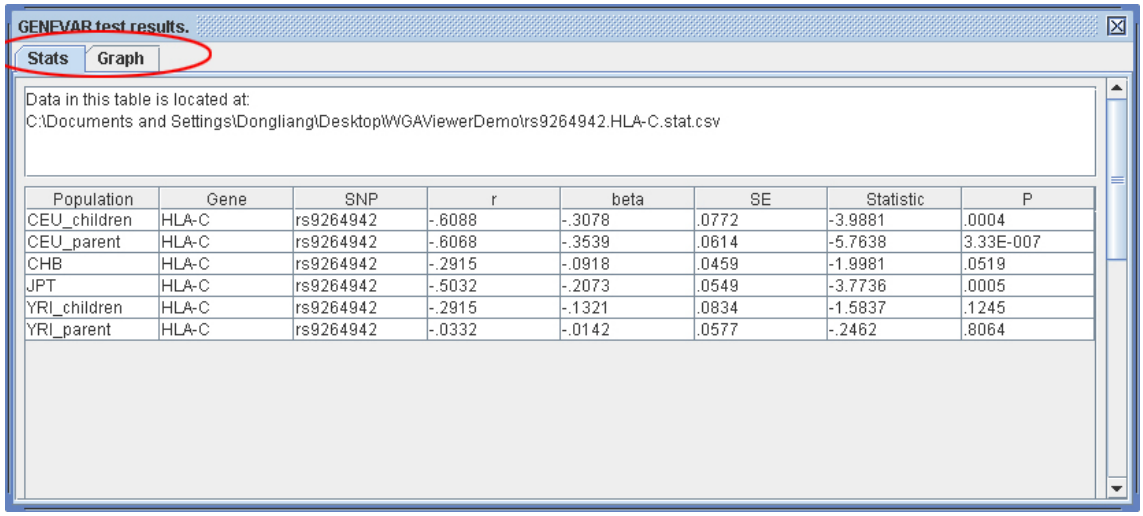


Figure 3.11.1-1. Test for genotype-gene expression association

```
rs2395029    HCP5
rs9264942    HLA-C
```

Box 3.11.1-1. An example of input file for genotype-gene expression association tests

Click on button "Perform test". Figure 3.11.1-2 shows the results based on inputs from Figure 3.11.1-1. Click on the "Stats" tab for a sheet of statistics, or click on the "Graph" tab for graphs (Figure 3.11.1-3). If multiple SNP-gene pairs were entered (from an input file), one has the option to select to particular genes and SNPs from two drop-down menus (circled in Figure 3.11.1-3).



Population	Gene	SNP	r	beta	SE	Statistic	P
CEU_children	HLA-C	rs9264942	-.6088	-.3078	.0772	-3.9881	.0004
CEU_parent	HLA-C	rs9264942	-.6068	-.3539	.0614	-5.7638	3.33E-007
CHB	HLA-C	rs9264942	-.2915	-.0918	.0459	-1.9981	.0519
JPT	HLA-C	rs9264942	-.5032	-.2073	.0549	-3.7736	.0005
YRI_children	HLA-C	rs9264942	-.2915	-.1321	.0834	-1.5837	.1245
YRI_parent	HLA-C	rs9264942	-.0332	-.0142	.0577	-.2462	.8064

Figure 3.11.1-2. Text-based results for test for genotype-gene expression association

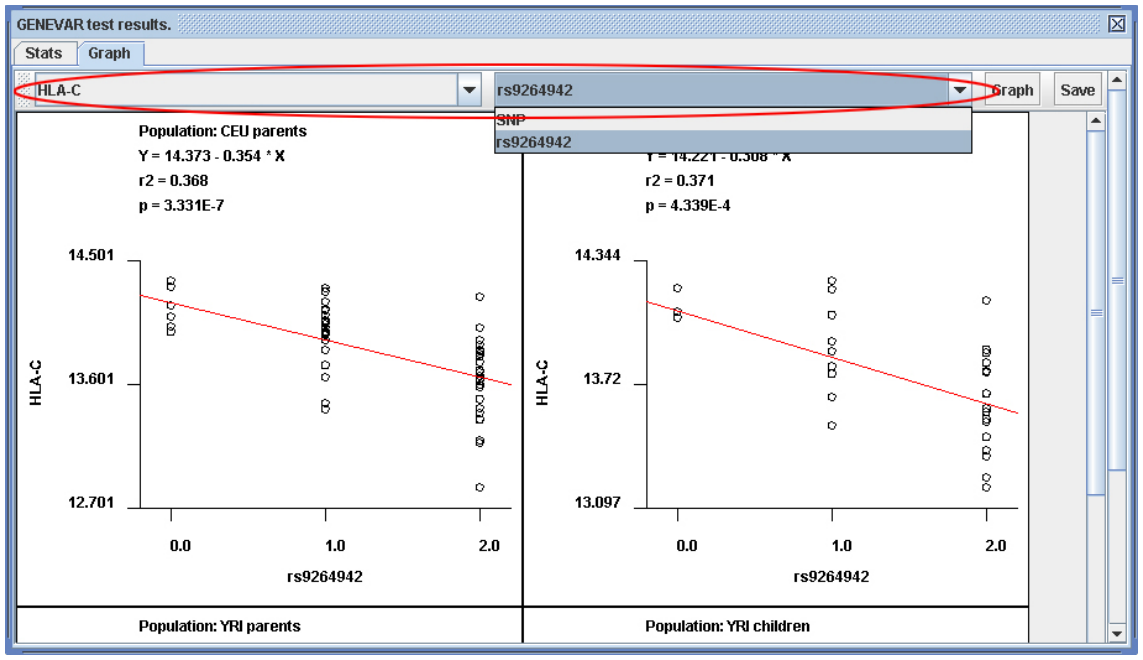


Figure 3.11.1-3. Graphical results for test for genotype-gene expression association

(3.11.2) Annotation for a SNP

The function of this tool is very similar to the SNP annotation procedures in WGAViewer, except that no WGA results are involved.

Click on menu "Tools -> Annotate a SNP". Setup the annotation parameters, for example, as in Figure 3.11.2-1, to annotate SNP rs9264942 with an up- and down-stream span of 100kbp, and a genotype-expression association test with gene HLA-C. One has the option to use HapMap data for LD calculation, or use pre-calculated LD dataset by PLINK or Haploview. This will result in a typical SNP annotation as discussed in this user guide (see: section 3.3.2.c, Figure 3.3.2.c-1), however with no WGA results involved (Figure 3.11.2-2).

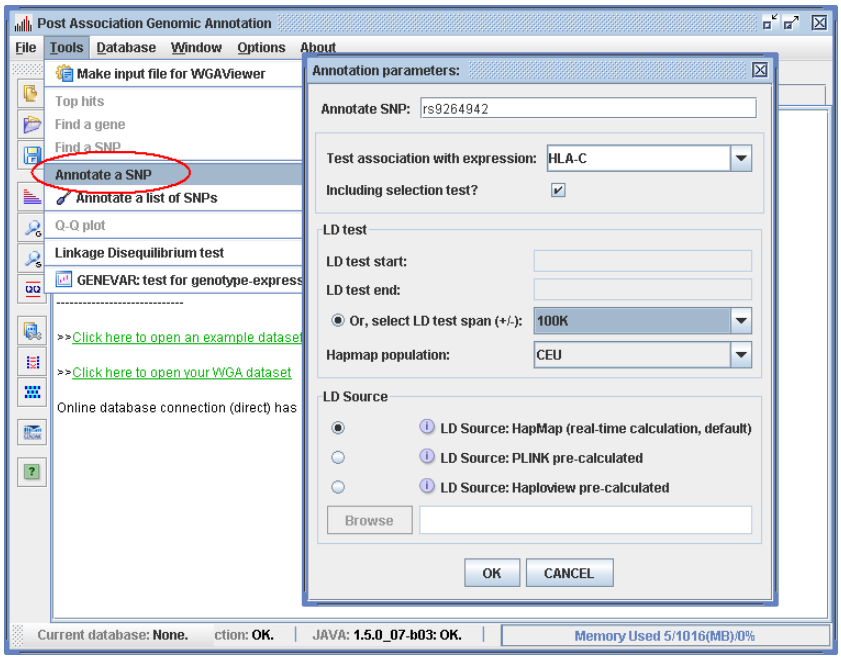


Figure 3.11.2-1. Tools: Annotation for a SNP.

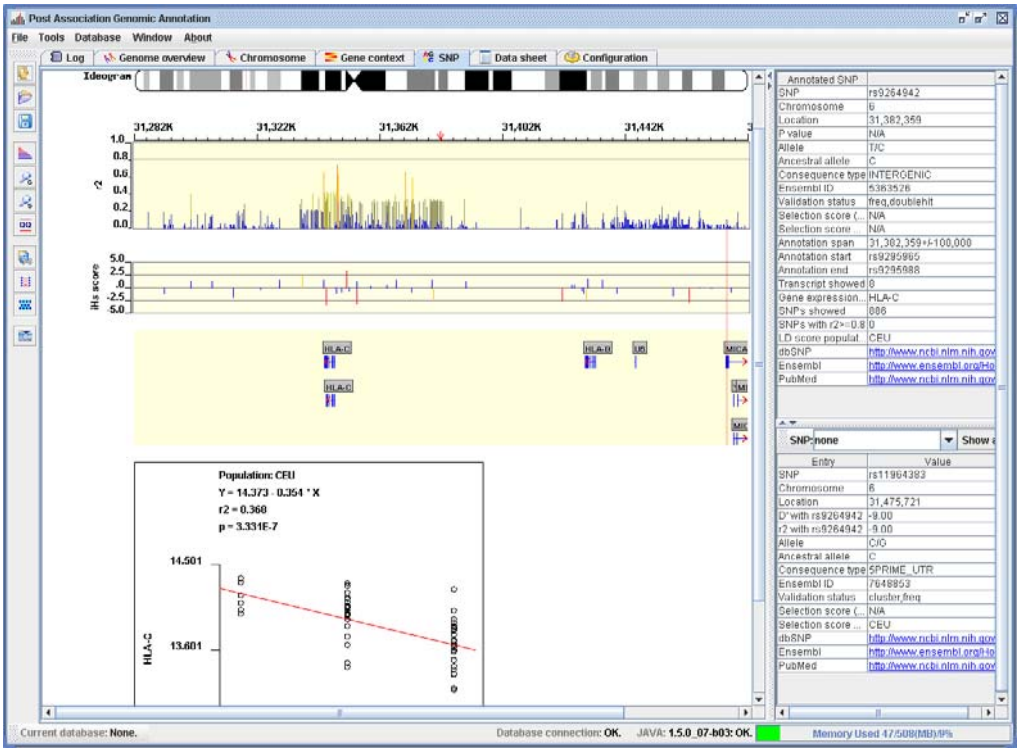


Figure 3.11.2-2. Tools: Annotation for a SNP: results.

(3.11.3) Linkage Disequilibrium test for a list of SNPs

This function is a very convenient tool for performing linkage disequilibrium tests for a list of SNPs, using HapMap data (The International HapMap Consortium. 2005).

The main utility of this tool is that the user does not need to know the detailed information about the SNPs for LD tests. For example, it is not necessary to have information on which chromosomes, in what HapMap populations (all HapMap populations will be tested), etc. All the information required is a simple list of SNPs, as shown in Box 3.11.3-1. Sometimes this is very useful and convenient, especially when one has a long list of SNPs of interest (for example, top hits from a WGA project) and it would be tedious to group these SNPs into different chromosomes, to download the related HapMap genotype data, and to calculate LD one by one.

rs2395029
rs9264942
rs13207315
rs17136351
rs10484554
rs11623538
rs4131373
rs481830
rs12082157
rs3131003
rs1891060
rs7539708
rs12889327

Box 3.11.3-1. Contents of the example input file: LDTest_SNP_List.txt

To use this tool, click on menu "**Tools->Linkage Disequilibrium test**". Two types of inputs are supported: (1) A text file simply listing the SNPs (Box 3.11.3-1); or (2) Direct typing in two SNPs to check. An example is given in Figure 3.11.3-1. In this case, a text file (Box 3.11.3-1) was used as an input. Select an output file and click on the "Perform test" button. Figure 3.11.3-2 shows the results, in which WGAViewer automatically groups the list of SNPs into different chromosomes, and performs LD test within each chromosome. One has also the option to filter the results based on chromosome, HapMap population, or certain SNPs. Click on the "Show" button to view the filtered results.

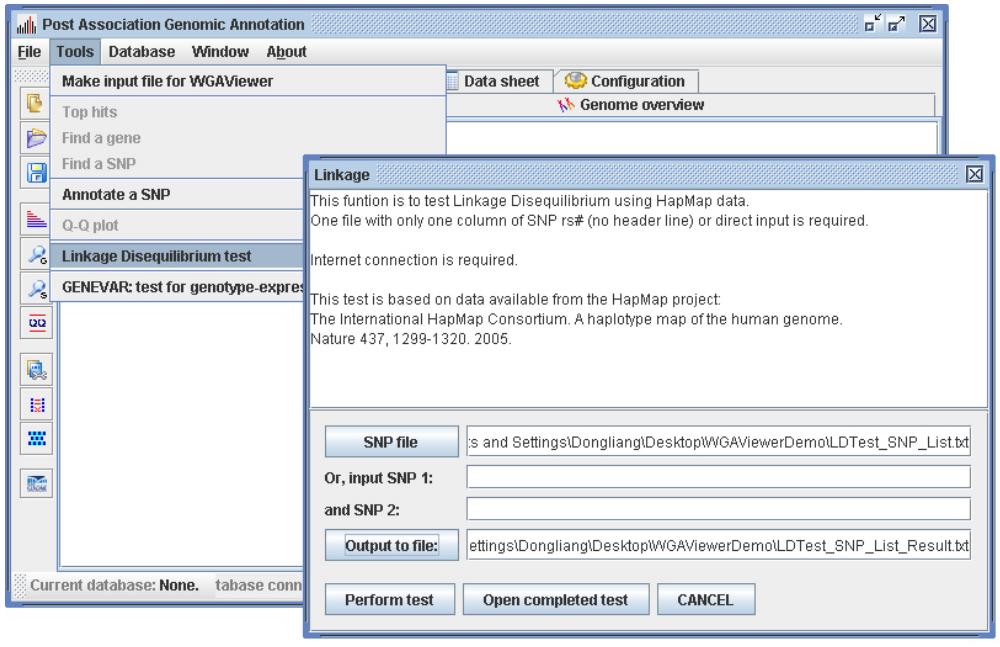
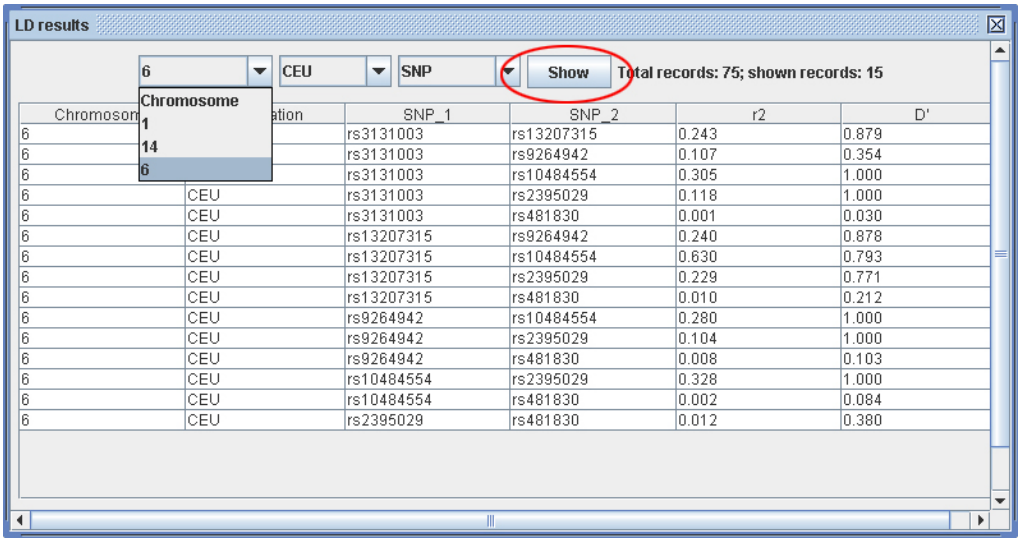


Figure 3.11.3-1. Tools: LD test for a list of SNPs



Chromosom	ation	SNP_1	SNP_2	r2	D'
6		rs3131003	rs13207315	0.243	0.879
6		rs3131003	rs9264942	0.107	0.354
6		rs3131003	rs10484554	0.305	1.000
6	CEU	rs3131003	rs2395029	0.118	1.000
6	CEU	rs3131003	rs481830	0.001	0.030
6	CEU	rs13207315	rs9264942	0.240	0.878
6	CEU	rs13207315	rs10484554	0.630	0.793
6	CEU	rs13207315	rs2395029	0.229	0.771
6	CEU	rs13207315	rs481830	0.010	0.212
6	CEU	rs9264942	rs10484554	0.280	1.000
6	CEU	rs9264942	rs2395029	0.104	1.000
6	CEU	rs9264942	rs481830	0.008	0.103
6	CEU	rs10484554	rs2395029	0.328	1.000
6	CEU	rs10484554	rs481830	0.002	0.084
6	CEU	rs2395029	rs481830	0.012	0.380

Figure 3.11.3-2. Tools: LD test for a list of SNPs: results

(3.12) Saving the work session.

All the data sets, figures, and interactive features during a work session can be saved to a binary ".wga" file. To do this, click "File -> Save work". To open the saved work session file, click "File->Open saved work". If you annotate further after loading a saved work session, you need to save the further work to another work session file and load them separately.

If you load your work session file, you don't need to load the text data set file (.wr) separately.

If you only load your work session file to view and perform no further annotation, the program does not need the internet connection to do so.

(3.13) Graphical user interface components

(3.13.1) Main menu

The main menu consists of four menu groups: File, Tools, Database, Window, and About. The menu items will be activated (colored in black instead of gray) when the conditions for performing the corresponding functions are satisfied.

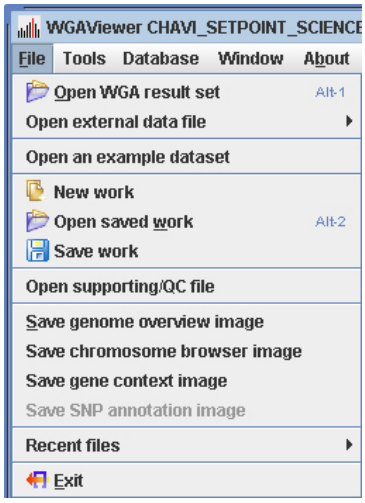


Figure 3.13.1-1. File menu

File menu:

1. Open WGA result set: will load a ".wr" (pre-annotation) or ".wga" (post-annotation) file;
2. Open external data file: will load a text-based WGA result set, including outputs from PLINK (Purcell et al. 2007). See chapter 3.2;
3. Open an example dataset: will load an example dataset released with WGAViewer. In this current version we use our recently completed study on host control of HIV-1 viral load during the asymptomatic set point period (Fellay et al. 2007) as example datasets. See chapter 3.1;
4. New work: will start a new annotation project;
5. Open saved work: will load the post-annotation file (.wga). For example see chapter 3.1;
6. Open supporting/QC file: will load supporting/QC files to help interpret WGA results. See chapter 3.9;
7. Save genome overview image: will save "Genome overview" tabbed panel as an image file. See chapter 3.6;
8. Save chromosome browser image: will save "Chromosome" tabbed panel as an image file. See chapter 3.6;
9. Save gene context image: will save "Gene context" tabbed panel as an image file. See chapter 3.6;
10. Save SNP annotation image: will save "SNP" tabbed panel as an image file. See chapter 3.6;
11. Recent files: includes a list of recently opened files. Click to open them again;
12. Exit: will exit WGAViewer.

Tools menu:

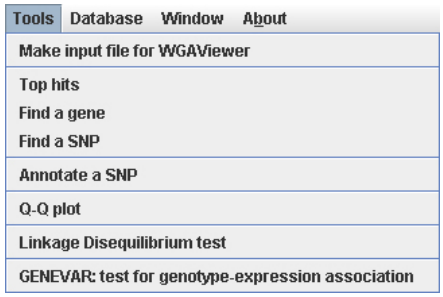


Figure 3.13.1-2. Tools menu

1. Make input file for WGAViewer: will make text-based .wr file for WGAViewer from text-based WGA result set. See chapter 3.2;
2. Top hits: will sort the dataset, find and annotate the top hits with specified number. See chapter 3.3;
3. Find a gene: will locate and annotate a gene region with specified span in the loaded WGA result set. See chapter 3.4;
4. Find a SNP: will check the coverage of specified SNP and its LD proxies in the loaded WGA result set. See chapter 3.5;
5. Annotate a SNP: will annotate a SNP not necessarily based a WGA result set. See chapter 3.11.2;
6. Q-Q plot: will perform a Q-Q plot and calculate a lambda value to inspect the effects of population stratification, and to visually inspect how the top hit P values depart from the random distribution. See chapter 3.7;
7. Linkage disequilibrium test: will perform LD test for a list of SNPs using HapMap (The International HapMap Consortium. 2005) data, without detailed information necessary for inputs. See chapter 3.11.3;
8. GENEVAR: test for genotype-expression association: will perform a genotype-gene expression association test based on specified SNPs and genes, using Sanger Institute GENEVAR database (Stranger et al. 2005)and HapMap database (The International HapMap Consortium. 2005).

Database menu:

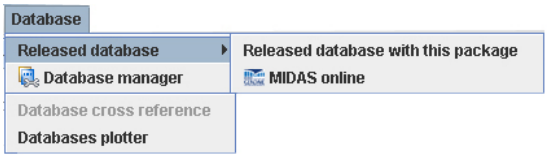


Figure 3.13.1-3. Database menu

1. Released database with this package: will load the released .wga database with this current WGAViewer package;
2. MIDAS online: will access the Mart for IGSP Data from Association Studies database. See chapter 3.10.
3. Database manager: will plot all data points from all core and reference databases in a separate window, and allows an interactive zoom in/out as well as annotation. See chapter 3.8.1;
4. Database cross reference: will search for all SNPs that show concurrent association signals at a specified level in core and all reference databases. This function requires at least one reference database to be loaded. See chapter 3.8.2;
5. Database plotter: will plot all data points from all core and reference databases in a separate window, and allows an interactive zoom in/out as well as annotation. See also chapter 3.8.1;

Window menu:

This menu allows a revisit to the saved annotation results after the related data windows are closed. These menu items will be activated if such annotation results are already available.

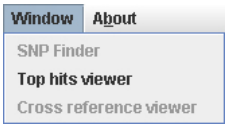


Figure 3.13.1-4. Window menu

1. SNP finder: will show the results from "Find a SNP". See menu "Tools-> Find a SNP" and chapter 3.5;
2. Top hits viewer: will show the annotation results for top hits. See menu "Tools->Top hits" and chapter 3.3;
3. Cross reference viewer: will show the results for database cross reference. See menu "Database-> Database cross reference" and chapter 3.8.2;

About menu:

1. Help: will display help window;
2. Check for Updates: will check and download the latest online WGAViewer version. The WGAViewer software needs a restart to take effect the updates;
3. About this software: will display software information (Figure 3.13.1-6).

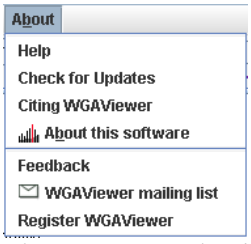


Figure 3.13.1-5. About menu

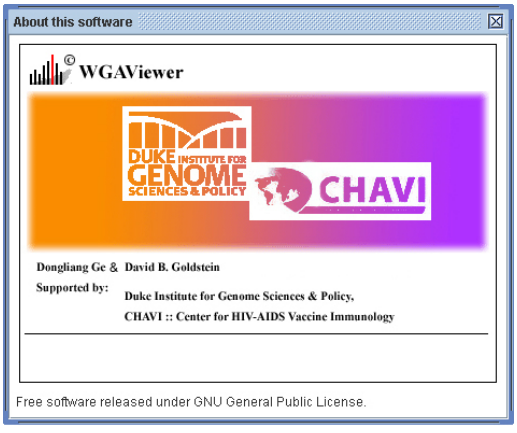


Figure 3.13.1-6. About this software.

(3.13.2) Tool bar

A tool bar is also available for convenient access for the most important functions offered by WGAViewer (Figure 3.13.2-1). Every tool bar button has an equivalent menu item. From up to down, these tool buttons are:

Figure 3.13.2-1. Tool bar

1. Create new work: equivalent to menu item "File->New work";
2. Open project file: will load a ".wr" (pre-annotation) or ".wga" (post-annotation) file. Equivalent to menu item "File->Open WGA result set";
3. Save work session: will save the work session to a binary ".wga" file with all interactive features included. Equivalent to menu item "File->Save work";
4. Top hits: will sort the dataset, find and annotate the top hits with specified number. Equivalent to menu item "Tools -> Top hits". See also chapter 3.3;
5. Find a gene: will locate and annotate a gene region, with specified span. Equivalent to menu item "Tools->Find a gene". See also chapter 3.4;
6. Find a SNP: will check the coverage of specified SNP and its LD proxies in the loaded WGA data set. Equivalent to menu item "Tools->Find a SNP". See also chapter 3.5;
7. Q-Q plot: will perform a Q-Q plot and calculate a lambda value to inspect the effects of population stratification and to visually inspect how the top hit P values depart from the random distribution. Equivalent to menu item "Tools->Q-Q plot". See also chapter 3.7;
8. Database manager: will allow a user to load, remove, and add description to core and reference databases. One also may use this interface to plot database and access MIDAS database. Equivalent to menu item "Database -> Database manager". See also chapters 3.8 and 3.10;
9. Cross reference: will search for all SNPs that show concurrent association signals at a specified level in core and all reference databases. Equivalent to menu item "Database -> Database cross reference". See also chapter 3.8.2;
10. Database plotter: will plot all data points from all core and reference databases in a separate window, and allows an interactive zoom in/out as well as annotation. Equivalent to menu "Tools -> Databases plotter". See also chapter 3.8.1;
11. Mart for IGSP Data from Association Studies (MIDAS): will access the MIDAS online database. See also chapter 3.10.
12. Help: will bring up a help window for each panel that currently has the focus.

(3.13.3) Navigation bar

Figure 3.13.3-1. Navigation bar

A zooming bar is available to select chromosomal region in the "Chromosomal" panel (Figure 3.13.3-1). From left to right, these tool buttons are:

1. "Chromosome" box: selects chromosome;
2. "From" and "to" text field; selects chromosomal region;
3. "Display" button: press to display the relative chromosomal region, based on the genomic coordinates **as inputted by the user** (and will be annotated in the "Gene context" panel);
4. "Full length" button: press to display the full length of the current selected chromosome;
5. "+" button: zooms in, by the step specified by the "step" box;
6. "-" button: zooms out, by the step specified by the "step" box;
7. "Step" box: specifies the step for zooming in/out, or moving backward/forward;
8. "<<" button: moves backward (to smaller chromosomal coordinates), by the step specified by the "step" box;
9. ">>" button: moves forward (to larger chromosomal coordinates), by the step specified by the "step" box;
10. "Annotate' button: annotates the currently showed chromosomal region;

11. Information field: shows the current resolution;
12. “How to use” button: shows help information on the “Chromosome” panel.

(3.13.4) Status bar

Figure 3.13.4-1. Status bar

There is a status bar located at the bottom of the main window showing the current status of the program (Figure 3.13.4-1). From left to right, these status are:

1. Current database: shows the currently loaded database;
2. Software version: shows a red star (as in Figure 3.13.4-1) when WGAViewer detects an update. Click it to download the update. A blue star indicates that the version you are using is current.
3. Database connection: shows the status of online database connection. Most of the annotation cannot be performed if this does not state “OK”. Click on this status label to retest the database connection;
4. Memory monitor: monitors the memory allocation and usage. The color will turn yellow/red when the memory is not adequate. Click to reallocate memory to be used by WGAViewer.

(3.13.5) Tabbed panels

Figure 3.13.5-1. Tabbed panels

The tabbed panels as shown from Figure 3.13.5-1 are the main working interface of WGAViewer. The titles of these tabbed panels are largely self-explanatory. Click on titles to switch the tabs. See also chapter 3.6.

(3.14) Configurations

Several display/usage parameters can be set in a "Configuration" tabbed panel. These parameters include:

Internet: proxy name and port;
Display: threshold for highlighting a SNP; Y axis (-LogP) scale (Figure 3.14-1);
Data source: default data folder. This configuration will influence all "Open file" and "Save file" dialogs in WGAViewer software. The default is the example folder "./examples". For example see Figure 3.14-2.
Annotation: LD criterion for considering as high association; HapMap population used to calculate LD;
Memory allocation: to allocate computer memory to WGAViewer software. Restart WGAViewer software to take effect. See chapter 2.3.

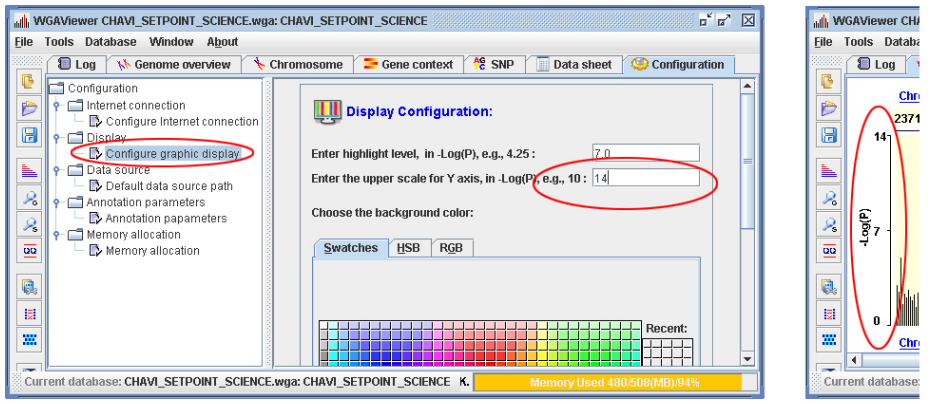


Figure 3.14-1. Y axis scale.

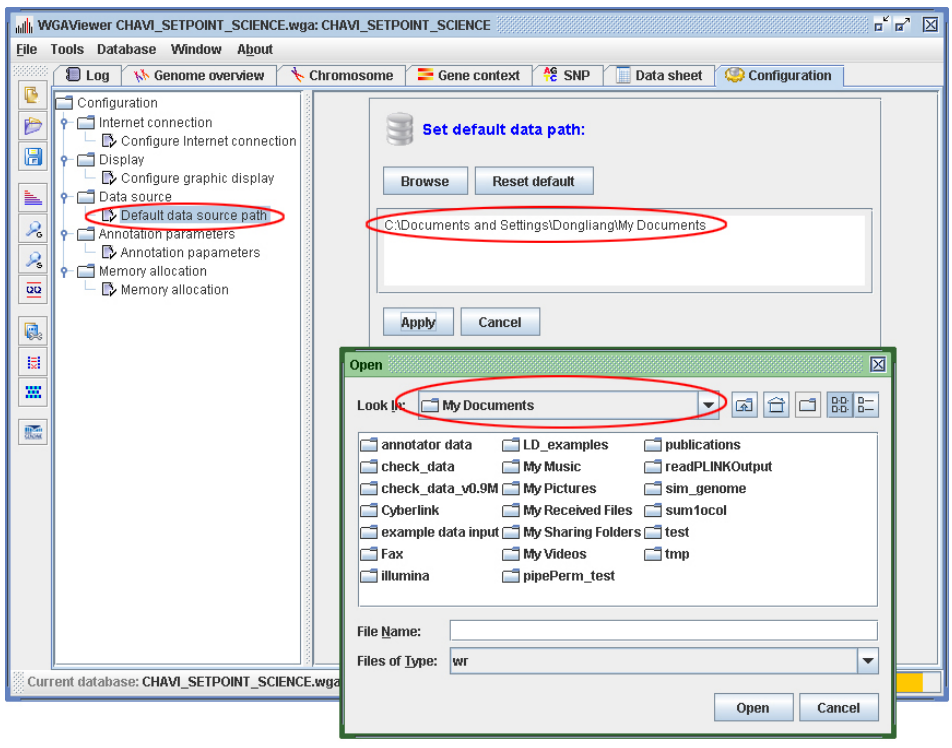


Figure 3.14-2. Default data path.

(3.15) Checking WGAViewer updates

Click on menu “About->Check for Updates” to check and download the latest online WGAViewer version. The WGAViewer software needs a restart to take effect the updates;

(3.16) Online databases version control

For each session of annotation, WGAViewer will automatically check the latest genome build, HapMap release, etc and align the database loaded with the latest genome build. The version of the databases used will be displayed in the upper right table in Genic Annotation (Gene context panel). No manual setup is necessary.

WGAViewer uses a “Dynamic Genome Build Anchoring” algorithm to control for the genome coordinate discrepancies during the version updates. Instead of directly using genome coordinates from diverse sources, we have treated the genetic markers as anchors and always used a fast hashtable-matching to map these anchors into the uniform Ensembl build version. This uniform build version, together with the version from other sources, is stored in the annotated project file and can be referred to in the later interpretation. This dynamic procedure helps to interpret the WGA results using the most updated transcripts and SNP coordinates, and avoids discrepancy between different major builds used in different sources (for example, build 36 and build 35), or even between different subversions (for example, build 36: Ensembl version 46.36h and 43.36e). This procedure is implemented in all the annotation routines, most importantly, in the gene and SNP searching functions. We noticed that sometimes these discrepancies could be up to ~200kb, which could potentially lead to very different gene/SNP contexts and hence result in different interpretations and hypotheses.

4. Future plans

WGAViewer is an ongoing project under continuous development. Our future development plans currently include:

1) transcription factor bindings sites
2) miRNA sites
3) comparative genetics
4) splicing regulators; regulatory sequences

We also would welcome any suggestions and comments.

5. Credits and sources of supports

WGAViewer has mainly been developed by Drs. Dongliang Ge and David Goldstein in Duke IGSP. Some colleagues have also participated in the development, including:

Drs. Kunlin Zhang & Amalio Telenti [a]; Olivier Martin [b]
comparative genetics module, quality control module.
[a] Institute of Microbiology at the University of Lausanne, Switzerland.
[b] Swiss Institute of Bioinformatics, Lausanne, Switzerland

The authors would thank Dr. Mike E. Weale from University College London for help in the population stratification module.

The authors also thank Darin London, Robert J. Wagner, and Mark R DeLong from Duke Institute for Genome Sciences & Policy for their work on setting up the Mart for IGSP Data from Association Studies (MIDAS). WGAViewer uses the query engine provided by the Biomart Data Integration System (http://www.biomart.org/). The MIDAS data are also accessible from the MIDAS BioMart instance at http://midas.genome.duke.edu/biomart/martview.

This project is supported by the Center for HIV-AIDS Vaccine Immunology (CHAVI, http://www.chavi.org/) and the Duke Institute for Genome Sciences & Policy (IGSP, http://www.genome.duke.edu/).

6. Citing this software

Ge, D. & Goldstein, D.B. WGAViewer: Package of Whole Genome Association Annotation. 1.10 edn (Durham, NC, 2007).

7. Projects using WGAViewer

Fellay, J., K.V. Shianna, D. Ge, S. Colombo, B. Ledergerber, M. Weale, K. Zhang, C. Gumbs, A. Castagna, A. Cossarizza, A. Cozzi-Lepri, A. De Luca, P. Easterbrook, P. Francioli, S. Mallal, J. Martinez-Picado, J.M. Miro, N. Obel, J.P. Smith, J. Wyniger, P. Descombes, S.E. Antonarakis, N.L. Letvin, A.J. McMichael, B.F. Haynes, A. Telenti, and D.B. Goldstein. 2007. A whole-genome association study of major determinants for host control of HIV-1. Science 317: 944-947.

Cavalleri, G.L., M.E. Weale, K.V. Shianna, R. Singh, J.M. Lynch, B. Grinton, C. Szoeke, K. Murphy, P. Kinirons, D. O'Rourke, D. Ge, C. Depondt, K.G. Claeys, M. Pandolfo, C. Gumbs, N. Walley, J. McNamara, J.C. Mulley, K.N. Linney, L.J. Sheffield, R.A. Radtke, S.K. Tate, S.L. Chissoe, R.A. Gibson, D. Hosford, A. Stanton, T.D. Graves, M.G. Hanna, K. Eriksson, A.M. Kantanen, R. Kalviainen, T.J. O'Brien, J.W. Sander, J.S. Duncan, I.E. Scheffer, S.F. Berkovic, N.W. Wood, C.P. Doherty, N. Delanty, S.M. Sisodiya, and D.B. Goldstein. 2007. Multicentre search for genetic susceptibility loci in sporadic epilepsy syndrome and seizure types: a case-control study. Lancet Neurol 6: 970-980.

8. References

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37:1243-1246

Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, Zhang K, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317:944-947

Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, et al. (2007) Ensembl 2007. Nucleic Acids Res 35:D610-617

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81

Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavare S, Deloukas P, Dermitzakis ET (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1:e78

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848-853

Telenti A, Goldstein DB (2006) Genomics meets HIV-1. Nat Rev Microbiol 4:865-873

The International HapMap Consortium. (2005) A haplotype map of the human genome. Nature 437:1299-1320

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72

Weale M (2007) Personal communications.

Online databases and resources used in this program:

Ensembl : http://www.ensembl.org
HapMap: http://www.hapmap.org
Haploview: http://www.broad.mit.edu/mpg/haploview/
GENEVAR (GENe Expression VARiation): http://www.sanger.ac.uk/humgen/genevar/
Selection score: http://hg-wen.uchicago.edu/selection/haplotter.htm
dbSNP : http://www.ncbi.nlm.nih.gov/SNP/
UCSC Genome Bioinformatics Site: http://genome.ucsc.edu/
Clustal W software (1.83) multiple sequence alignments: http://www.ebi.ac.uk/clustalw/
BioMart: http://www.biomart.org/

9. License and Copyright

WGAViewer may be used freely for research or non-commercial purposes. Source code is available at: https://sourceforge.net/projects/wgaviewer/. If you wish to incorporate parts of this program into other software whose distribution conditions are different, please write to the authors for permission. For use in a commercial environment, please communicate with Dongliang Ge and David Goldstein. The authors reserve the copyright of this software.

WGAViewer is free software released under GNU General Public License, v2 (see COPYING.txt).

November, 2007