GrASP

A Graphical Assessment of Sliding P-values

USER'S GUIDE
(v.0.6 beta)

(June, 2006)

Rasika A Mathias
Inherited Disease Research Branch,
NHGRI/NIH,
333 Cassell Drive, Suite 1200,
Baltimore, MD 21224
Tel: 410-550-7120.
(rmathias1@mail.nih.gov)

Janet L Goldstein
Center for Inherited Disease Research,
333 Cassell Drive, Suite 2000,
Baltimore, MD 21224
Tel: 410-550-2819
(jgold@cidr.jhmi.edu)

Table of Contents

1. Overview
   - 1.1. Introduction
   - 1.2. Features of GrASP
2. Installing GrASP
   - 2.1. Installing Excel Macro of GrASP
   - 2.2. Installing Perl for the NCBI query
3. Executing GrASP
   - 3.1. The GUI
   - 3.2. Specifying the P-value sliding window tracks
   - 3.3. Specifying the physical track
   - 3.4. Specifying the color scheme
4. Data Files
   - 4.1. Input P-value data file
   - 4.2. Input Gene location file
   - 4.3. GrASP output file
5. Example
   - 5.1. Example Data Files
   - 5.2. Example Output
6. Error Messages
7. Copyright, Disclaimer and References
8. Coming soon to GrASP

1. Overview

1.1. Introduction

In statistical testing for association between regions with multiple markers in linkage disequilibrium (LD) and traits of interest, the use of haplotypes may be far more efficient than the statistical analysis of individual SNPs. In the current age of high-throughput genotyping and highly dense SNP data, the sliding window approach to analysis of haplotypes has gained much importance both on the methodological and application fronts. In this approach, windows of varying size are examined, generally beginning with the first SNP and sliding the windows down the map one SNP at a time. This has its main advantage in being a simple and efficient way to comprehensively screen a dense region of genotyping (typically SNPs) for association with the trait of interest. This advantage**,** however, is not without its issues: exceedingly high numbers of tests, the daunting task of assimilating thousands of test results, attempting to create haplotypes across SNPs too far from each other, and prioritizing regions based on the association signals from these analyses. Graphical Assessment of Sliding P-values, or GrASP, is an attempt to provide one modest but potentially efficient solution to some of these issues.

GrASP is a graphical tool to display and assess p-values from sliding window tests. It can present thousands of p-values from the sliding window tests as a simple graphic that uses varying levels of user-specified color to indicate the width of the sliding windows and the varying levels of significance. It therefore allows the user to identify regions/blocks of interest from these sliding windows, based jointly on the absolute p-value of the tests from these windows and the building of haplotypes of significance in the region. GrASP is executed as an Excel macro and is written in Excel's built-in version of Visual Basic for Applications. It is freely available at: http://research.nhgri.nih.gov/GrASP/ .

.

1.2. Features of GrASP

GrASP has two main features:

(A) It extracts the information on the sliding window tests from user-provided input and creates a graphical overview to illustrate the size of each window and the p-value of the test for each window. This is referred to as the **P-value Track**. The P-value Track is always drawn by default.

(B) It illustrates the physical location of the SNPs and gene locations alongside the graphical overview of the sliding windows using either user provided input or by querying public databases (specifically Entrez). This is referred to as the **Physical Track.** This track is optional and is not drawn by default.

2. Installing GrASP

GrASP has two components: an Excel macro written in Excel's built-in version of Visual Basic for Applications and a Perl script that allows the user to query SNP positions and known genes from the National Center for Biotechnology Information (NCBI) public databases (currently available only in the Windows version). It is freely available for download at http://research.nhgri.nih.gov/GrASP/ . If you would like to be informed about new releases and updates to GrASP, please e-mail the authors at rmathias1@mail.nih.gov to be added to the GrASP e-mail list.

2.1. Installing the GrASP Excel Macro

To set up GrASP

1. Place the `GrASP.xla` file in your Excel add-ins folder.

   On a PC, this will be

   `C:\Documents and Settings\`*`username`*`\Application Data\Microsoft\AddIns`

   where "username" is your user name that you use to log on.

   On a Macintosh, this will be

   `Applications : Microsoft Office X : Office : Add-Ins`

2. Open Excel.

3. You must enable macros in order to use GrASP. To enable macros, select **Macro** from the **Tools** menu, then **Security** from the pop-up menu. Set the security level to medium or low. If you are using a PC with an automatic virus scanner installed, the Low setting is recommended. Choosing the Medium security level will cause Excel to ask if you wish to enable macros at every execution as long as you have GrASP installed.

4. Select **Add-Ins** from the **Tools** menu.

5. If GrASP does not appear in the list of add-ins, add it using the Browse button.

6. Check the box next to GrASP, then close the add-ins list.

7. Now a **GrASP** item appears permanently at the bottom of your **Tools** menu within Excel.

2.2 Installing the Perl components for the NCBI query

To use the Physical Track feature in GrASP that queries map positions for SNPs and known genes from NCBI databases, the user needs to have Perl installed, because this feature of GrASP is implemented as a Perl script launched by the GrASP Excel macro. (At present, this functionality is available only in the Windows version of GrASP. The authors invite advice from OS X gurus as to how to run an external program from Excel X.) The most recent Perl distribution is available from ActiveState at ftp://ftp.activestate.com/ActivePerl/Windows/. As of this writing, the latest version is 5.8.7; the Perl component of GrASP was developed with version 5.8.4, which can be obtained from ftp://ftp.activestate.com/ActivePerl/Windows/5.8/ActivePerl-5.8.4.810-MSWin32-x86.msi. Download and execute the Windows installer (MSI) package and follow the onscreen instructions. (Windows 2000 Professional users: this requires administrator privileges on your local PC.)

Place the `get_SNP_gene_posns.pl` file in your Excel add-ins folder (see above for location of this folder).

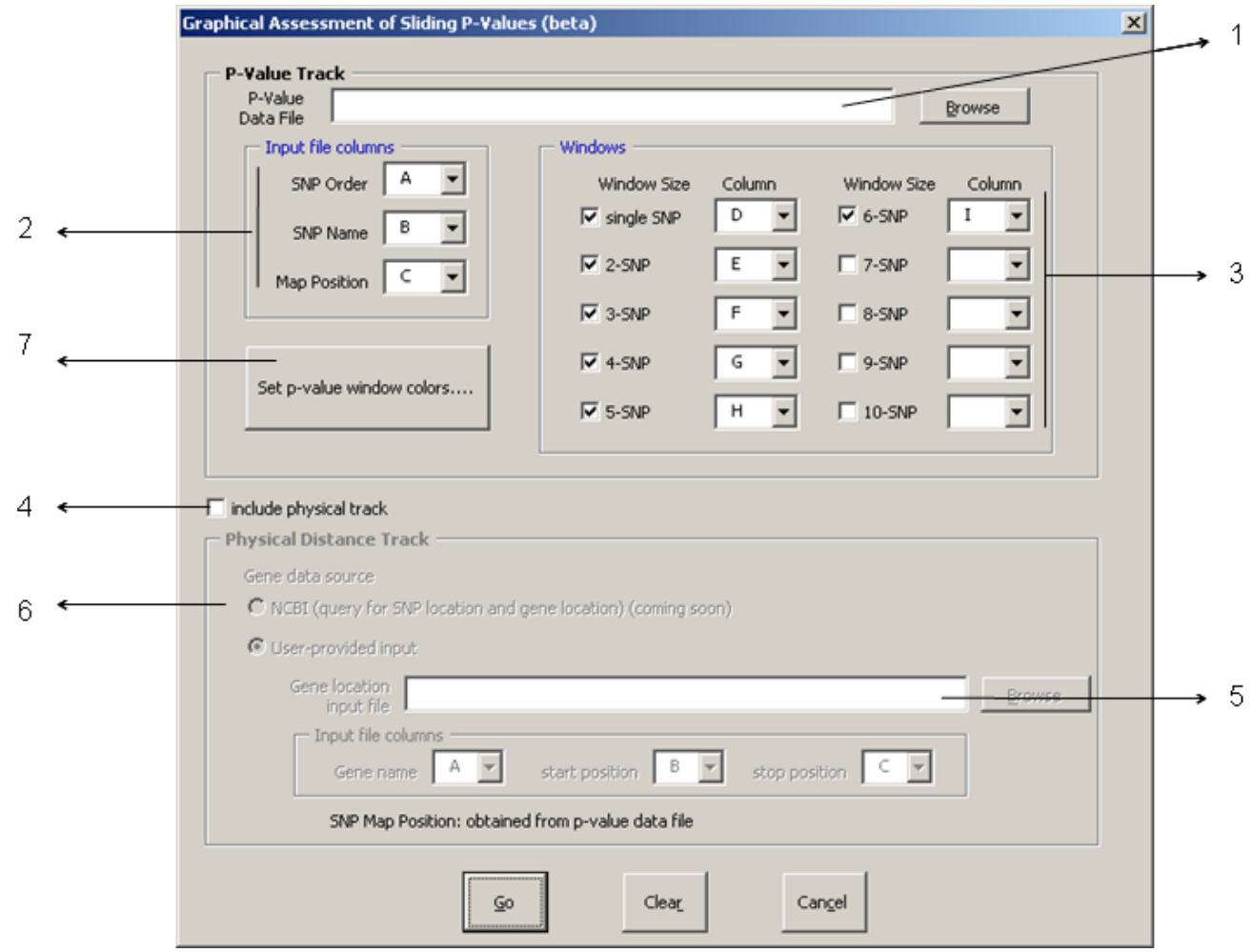
3. Executing GrASP

Once installed as an add-in, GrASP is available to the user under the Tools menu within Excel. To run GrASP simply select it from the Tools menu.

3.1. The GUI

The main Graphical User Interface (GUI) of GrASP as illustrated in Figure 1 allows the user to specify input file locations and names, column specification for data, vary the color scheme, and elect to draw the physical track containing SNP locations and gene locations and names. Specific details on the various features of the main GUI are presented below.

Figure 1: The main graphical user interface of GrASP.
(Flagged locations are referred to in sections 3.2 to 3.4 below)



3.2. Specifying the P-value Track (refer to Figure 1, location 1, 2, and 3)

The graphical overview of the sliding window test results is performed by default. Window sizes may range from 2 SNPs to 10 SNPs (the larger windows allow the user to use GrASP even in situations of dense SNPs in candidate genes, where one may analyze windows of larger sizes). The user must have data on at least one window for GrASP. The user may also have the test results for the single-SNP tests if they want the graphical overview to include these test results which may be useful.

GrASP requires as input one Excel file containing SNP order, name, location, and p-values from tests performed. Details on the structure of this file are specified in section 4.1. The location and name of the file are specified in the P-Value Data File textbox (**location 1** in Figure 1); the user may navigate to the file's location using the Browse button. Details on the columns containing the SNP information (i.e. SNP order, SNP name and SNP position in base pairs) can be specified in the Input File Columns section of the GUI (**location 2)**. The column locations of the single-SNP and sliding window p-values are specified in the Window Sizes section of the GUI (**location 3)**. Default column values are pre-set for the SNP information and the column locations for several of the p-value columns, however these may changed by the user.

3.3. Specifying the Physical Track (refer to Figure 1, location 4, 5, and 6)
The Physical Track is not drawn by default, and the user needs to check the "Include physical track" checkbox on the GUI (location 4 in Figure 1) to enable this feature. There are two alternate sources to the Physical Track data:

3.3.1. User-provided input: For this option, the user needs to specify a list of genes with their corresponding start and stop positions in base pairs that are to be drawn in this track. Details on the structure of this file are specified in section 4.2. The name and location of this user-provided file are to be specified in the "Gene location input file" text box of the GUI (location 5). In this case, the map positions of the SNPs for this track are obtained from those specified in the P-value data file (see section 4.1).

Note: It is possible for the user to draw only the physical locations of the SNPs and not include any genes in the physical track. This is done by checking the "Include physical track" checkbox (location 4) on the GUI, and not specifying any gene location filename in the text box.

3.3.2. Public Database input: For this option (location 6), GrASP will query NCBI's public database Entrez for current locations for all SNPs in the p-value file and all known genes between the first and last SNP. In this case, the locations of the SNPs for this track will be obtained from Entrez and the locations specified in the p-value data file will be disregarded.

This option requires an active Internet connection. Once this option is selected, and the user selects "Go", a window will pop up that requests the chromosome that the SNPs are on for verification of SNP location.

Note: It is not uncommon that SNP positions and location may change between the time of analysis and time of summary of results (i.e. between two successive genome builds). To accommodate this:

- The order of the SNPs will be retained in the graphical P-value Track. This is important because the SNPs in any window of a particular size are determined by the order in which they were originally analyzed.
- The Physical Track will illustrate the current location of SNP with appropriately drawn anchoring lines. Hence, anchoring lines may cross over when SNPs are no longer in the same order as which they were originally run, or anchoring lines will be absent when a SNP no longer is exists in the databases on the specified chromosome.

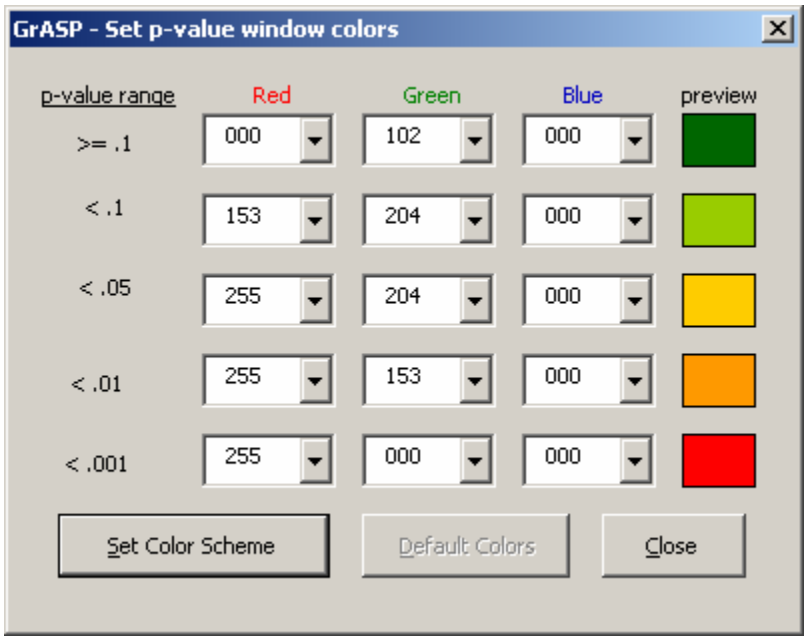
3.4. Specifying the color scheme (refer to Figure 1, location 7)

The default color scheme allows for five levels of significance to be displayed: p >= 0.1, p < 0.1, p < 0.05, p < 0.01, and p < 0.001. The less significant tests display green; the more significant tests display yellow, orange and red. By clicking the "Set P-Value Window Colors" button, the user can specify the Red (R) Green (G) and Blue (B) values (ranging from 0 to 255 for each) for each of these 5 categories. This allows the user to

- select any color of choice
- vary the granularity of the significance display. For example, setting p >= 0.1, p < 0.1, and p < 0.05 to the same RGB color value makes the p < 0.01 and p < 0.001 windows stand out.

Clicking the "Set P-value Window Colors" button (location 7 in Figure 1) opens a second window (Figure 2) in which the colors of the five significance levels can be altered. As the RGB values are varied, the user will be able to preview the specified color in this window. Once the desired color scheme is obtained, the user must set the color scheme using the "set color scheme" button. This color scheme will be retained for the current session of GrASP and will revert back to the default scheme when GrASP is closed.

Figure 2: GrASP color specification screen.



4. Input Data Files

There are two input data files: a required file providing the p-values from the sliding window tests, and an optional file providing the gene names and locations.

4.1 P-value data file

If one has the following three pieces of information, one simple file will store all test results from sliding window-based tests:

(1) The first SNP of each window
(2) The size of that window (number of SNPs in the window)
(3) The SNP order that was used in the sliding windows

To illustrate: Assume you have 10 SNPs analyzed in map order 1 though 10, and you are doing 3- and 4-SNP slides. You know that a 3-SNP test where the first SNP was #4~~,~~ went from SNP 4 to SNP 6. Similarly, you also know that a 4-SNP test in which the first SNP was #5 would have gone from SNP 5 to SNP 8, etc. This is the principle around which the user-provided p-value data file is designed, and it requires the user to have only as many rows as there are SNPs with as many columns as there are windows, as described below. (See example data file in section 5.1).

The p-value data file is required, and contains the following columns:

- **SNP order:** Required; contains the order in which the SNPs were used in the analysis. This column cannot have empty cells and must be numeric. The input file must be sorted on SNP order. If the SNPs are out of order in the input file, GrASP will terminate.
- **SNP name**: Required; contains the name of the first SNP of the window. This column cannot have empty cells and is alphanumeric. Since GrASP expects a header row, generally the first SNP will be on row 2, the second SNP on row 3, … the nth SNP on the n^{th+1} row.
- **Map position**: Required if physical track is to be drawn with user-provided input; optional if data will be queried from NCBI; contains the map position in base pairs for each SNP in the SNP name column. This column can have empty cells (if map position is not known) and is numeric. If SNPs are not in ascending map position or if map position is missing, GrASP will warn the user but will not terminate. This is to accommodate changes in map positions between builds.
- **Single SNP:** Optional; contains the p-value of the single SNP test (note: this is the equivalent of a window of size 1).
- **2-SNP:** Optional; contains the p-value of the 2-SNP window that began with the corresponding SNP in the SNP name field. For example: the 2-SNP p-value in the row corresponding to the first SNP is the p-value from the window that included the first and second SNPs. Again, the 2-SNP p-value in the row corresponding to

the fifth SNP is the p-value of the window that included the fifth and sixth SNPs. (See example section 7.1).

- **3-SNP:** Optional; contains the p-value of the 3-SNP window that began with the corresponding SNP in the SNP name field. For example: the 3-SNP p-value in the row corresponding to the first SNP is that 3-SNP window that included the first, second and third SNPs. Again, the 3-SNP p-value in the row corresponding to the fifth SNP is the p-value from the window that included the fifth, sixth and seventh SNPs. (See example section 7.1).
- **4-SNP to 10-SNP:** These columns are defined exactly as above (see 2- and 3-SNP) and correspond to 4-, 5-, 6-, 7-, 8-, 9- and 10-SNP sliding windows.

Requirements of the p-value data file:

- There must be a header row with column names. While GrASP expects the header row, it does not actually use the column names in the graphical output, so the user may name columns as desired.
- GrASP assumes that the data is in the first sheet of the Excel workbook.
- While not all columns for the p-values are required, the user must have at least one of the sliding windows tests from the 2- to 10-SNP tests included, or GrASP will terminate.
- The p-value columns need to be numeric where tests were performed and have blank fields (truly blank cells not filled with spaces) where a test was not performed. For example, the cells corresponding to the last 5 rows in the 6-SNP sliding window will be blank since there can be no 6-SNP window test conducted that had as its first SNP anything after the n-6th SNP, where n is the total number of SNPs in the dataset. Similarly, if the user performed sliding windows only for selected segments of contiguous SNPs, the corresponding cells can be left blank.

4.2 Gene location file

This file is optional and needs to be specified only if the gene positions are to be drawn for the physical track user user-provided input. Note: This file is not required when drawing the physical track using the NCBI query option.

- **Gene name:** Optional; contains the name of the gene. It can be blank if gene name is unknown, and it is alphanumeric.
- **Start position**: Required; contains the start of the gene in base-pairs. It cannot be blank and it must be numeric.
- **Stop position**: Required; contains the end of the gene in base-pairs. It cannot be blank and it must be numeric.

Requirements of the gene location file:

- There must be a header row with column names. While GrASP expects the header row, it does not actually use the column names in the graphical output, so the user

may name columns as desired. GrASP assumes that the data is in the first sheet of the Excel workbook.

- This file needs to have at least one gene included. **Note:** To have only the SNP locations drawn on the physical track and not genes, check the "Include physical track" checkbox and leave the gene location filename textbox blank.

4.3. GrASP output files

4.3.1. Main GrASP output

The output from GrASP is in the form of an active Excel workbook with the output from GrASP on the first worksheet. The user can save the workbook as an Excel file or copy the image (or selected parts of the image) and paste it into any program that supports graphics. The best option is to paste the image as a bitmap.

The P-value Track: The p-value track has the output from the windows arranged in descending order left to right (i.e. larger windows on left and smaller windows on right) annotated in the header row. For each window size there are as many columns as there are SNPs in a window to allow the tiles (each tile representing one sliding window test) to stagger. For example, there are 2 columns represented the 2-SNP window slides. The first column has windows beginning at even numbered SNPs (i.e., 2, 4, 6, etc.) and the second column has windows beginning at odd numbered SNPs (i.e., 1, 3, 5, etc.). Similarly, there are three columns representing the 3-SNP window: the first column having windows beginning at SNP 3, 6, 9, etc., the second column having windows beginning at SNP 2, 5, 8, etc., and the third column having windows beginning at SNP 1, 4, 7, etc.. There are blank columns between the graphical overview for each window size.

Each test/p-value is represented by a single tile: a box with a black outline containing as many cells vertically as there are SNPs in that window. The vertical length of the tile represents the size and range of the rows corresponds to the SNPs in that window. The p-value range is represented by the user-specified color. Each cell contains the numeric value of the p-value it represents; select the cell of interest and the p-value will appear in the formula bar.

The Physical Track: If the physical track is drawn, each SNP in the p-value track is anchored to a gray line based on its physical distance. The start and stop of the chromosomal location represented by this gray line is annotated in base pairs. The light pink boxes represent the genes contained in the gene location file (or all known genes in the region as obtained from NCBI public databases) drawn to scale. The boxes are anchored to their respective gene names.

4.3.2. Additional GrASP output

When the NCBI query is performed, GrASP will also produce the following comma-delimited files in the directory in which the original user-provided P-value input file is located:

P-value input filename.csv: This is a comma delimited file containing the original data from the P-value input file.

P-value input filename.csv_ncbi_out.csv This is a comma delimited file containing the SNP locations and the known genes obtained from NCBI.

P-value input filename.log: This is the log file from the Perl script. It contains useful information on errors and warnings from the Perl script (see Error Messages for further details).

5. Example

5.1. Example Data Files

Two input data files are provided:

(1) **pvalue_sample.xls** is a sample input file with 25 SNPs. The file includes the SNP order, SNP name, map position, the single SNP p-value and sliding window p-values from 2-SNP to 6-SNP windows. The blank fields in the p-value columns are where no tests were performed.

(2) **genes_sample.xls** is a sample input file with all known genes in the region corresponding to that of the 25 SNPs in the pvalue_sample.xls file, along with their start and stop positions.

5.2. Example Output

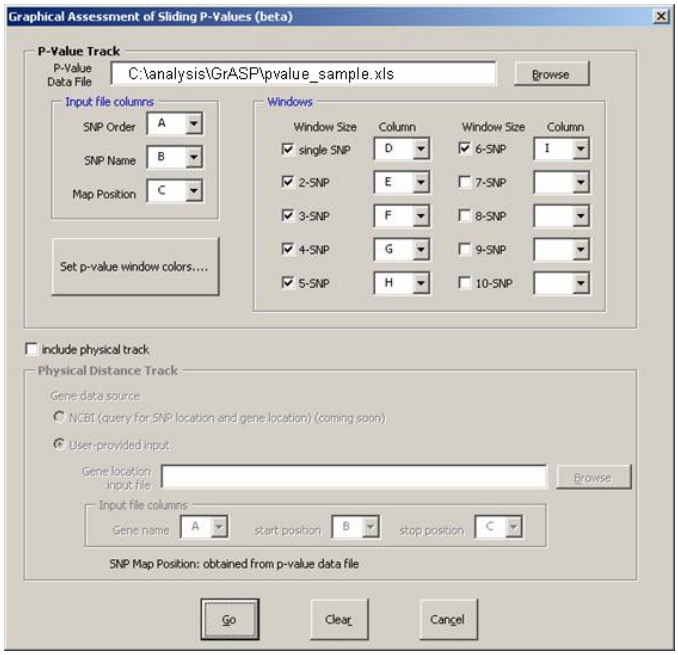
Three sample outputs along with the GUI options used to create these outputs are illustrated below:

(1) output_sample1 is the GUI and output obtained when only the graphical overview of sliding windows p-values is drawn, and not the physical track, from the user-provided input pvalue_sample.xls (Figure 3.1).

(2) output_sample 2 is the GUI and output obtained when the graphical overview of sliding windows p-values *and* the physical track are drawn for only the SNP locations from the user-provided input pvalue_sample.xls (Figure 3.2).

(3) output_sample 3 is the GUI and output obtained when the graphical overview of sliding windows p-values *and* the complete physical track are drawn from the user-provided input pvalue_sample.xls and gene_sample.xls (Figure 3.3).

Figure 3.1: GUI and output for drawing only the p-value track.



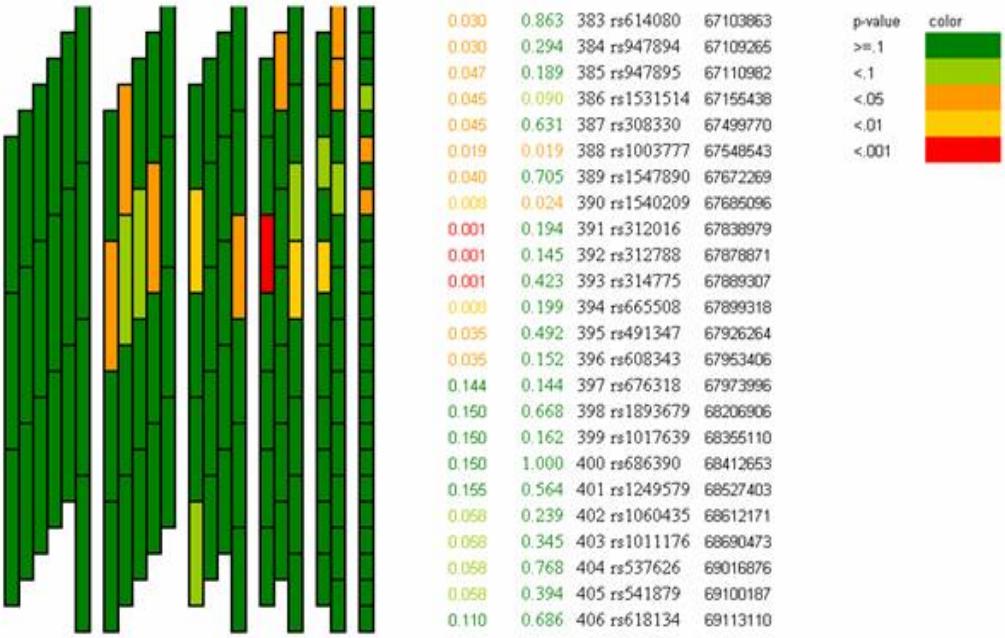
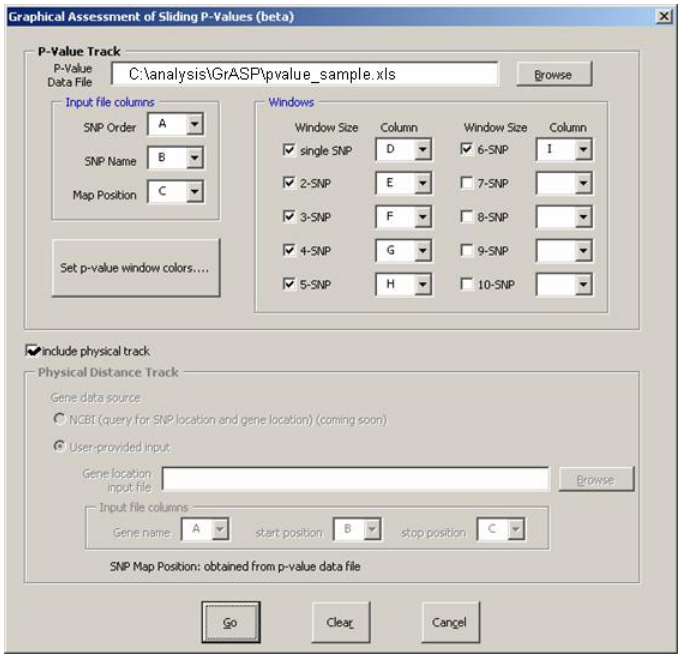


Figure 3.2: GUI and output for drawing the p-value track and only SNPs in the physical track.



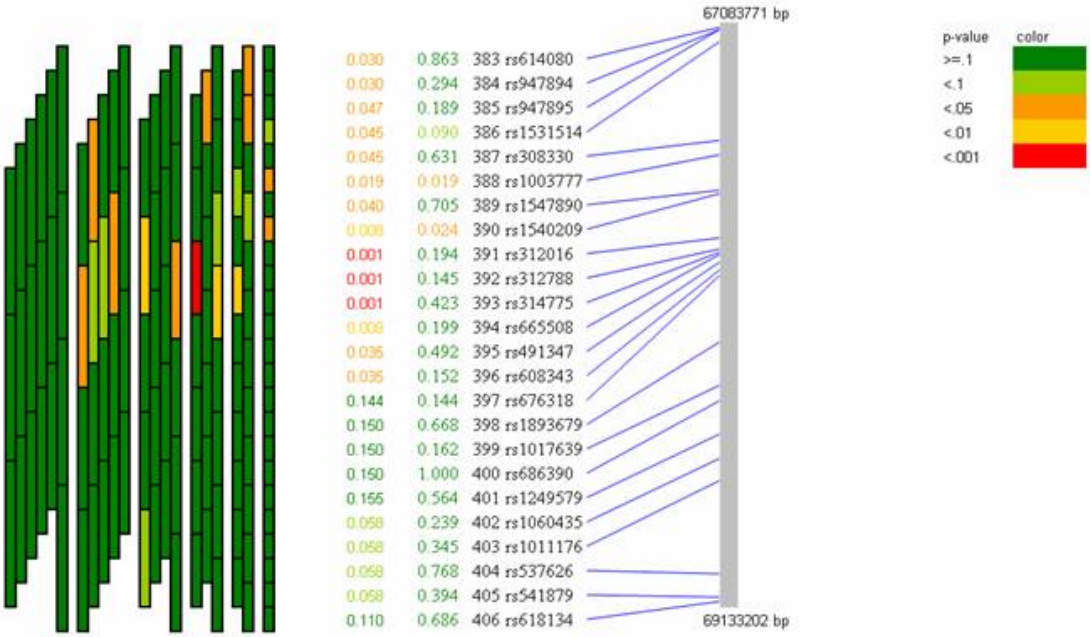
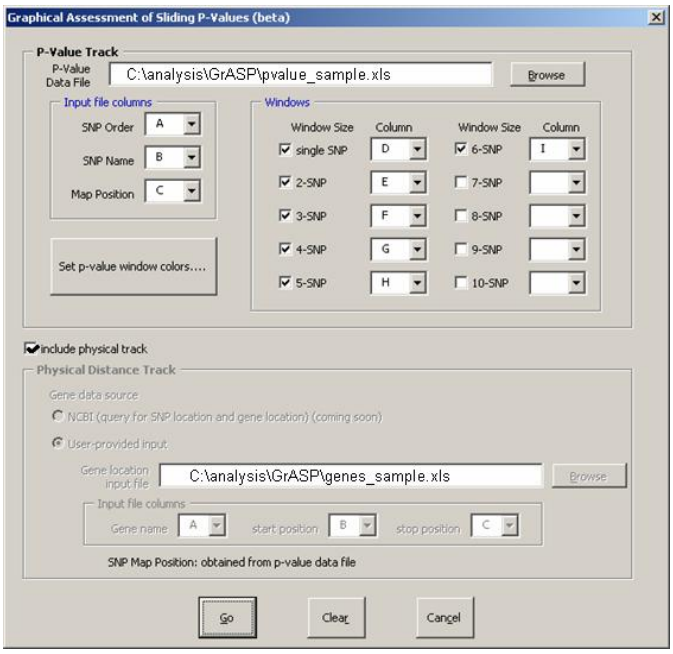
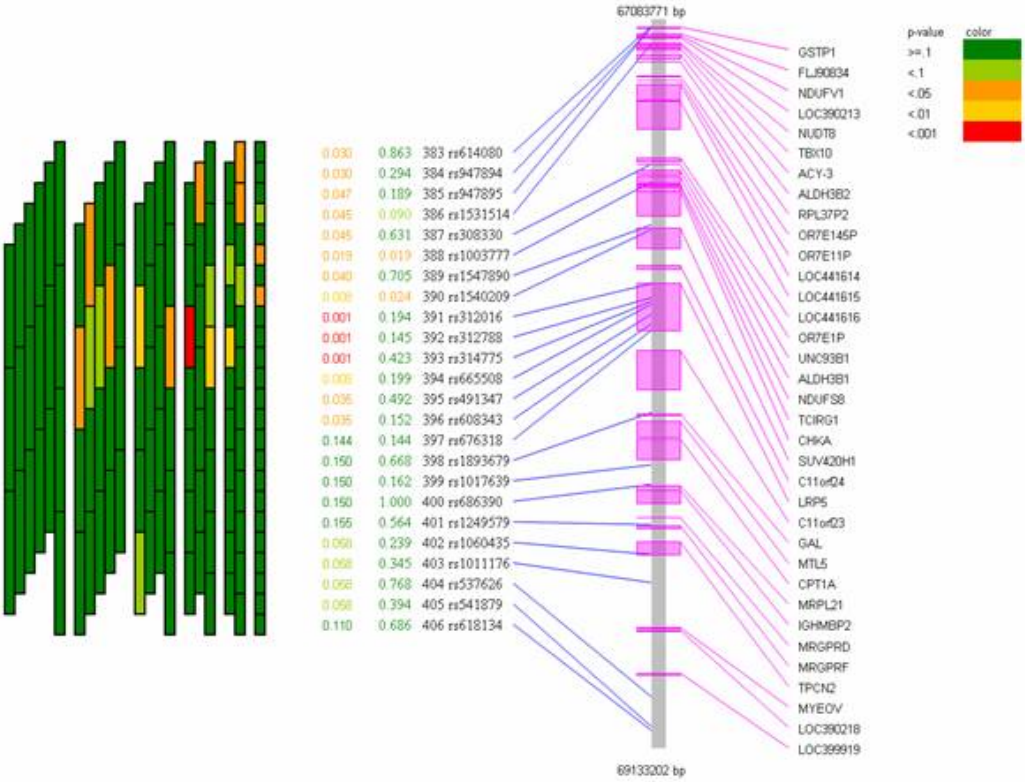


Figure 3.3: GUI and output for drawing the p-value track and the complete physical track (including both SNP locations and genes).





6. Error Messages

Below is a list of error messages from GrASP along with their descriptions:

Error	Description
Missing SNP order on line *n* of p-values input file	Cannot have a missing cell in the SNP order column. GrASP will terminate.
Non-numeric SNP order on line *n* of p-values input file	Cannot have a non-numeric cell in the SNP order column. GrASP will terminate.
SNP order is mixed up	P-value data file must be sorter on SNP order column. GrASP will terminate.
Missing SNP on line *n* of p-value data file	Should not have missing SNP names. GrASP will warn user and proceed.
The following SNPs are out of map order:	Map positions are not in ascending order or map position is missing in p-value data file. GrASP will warn user and continue.
Non numeric p-values at row *n* of column *m*	All cell values in any of the p-value columns in p-value data file must be numeric. GrASP will not execute completely (physical track will be incomplete)
Start position missing for gene *xxxx*	A missing cell value in the start position column of gene location file. GrASP will not execute completely (physical track will be incomplete)
Stop position missing for gene *xxxx*	A missing cell value in the stop position column of gene location file. GrASP will not execute completely (physical track will be incomplete)
Problem with gene *xxxx*: start position > stop position	The start position is higher than the stop position in the gene location file. GrASP will not execute completely (physical track will be incomplete)

Other Issues

- If a tile is seen with only color and no box around it, it indicates that a p-value is present in the p-value data file for a window size that is not allowed. This is typically seen at the terminal SNPs. An example is a 5-SNP p-value present on the row corresponding to the last SNP.

Perl Log File Errors

- `Getting position of rsXXXX....not found on chromosome X`

This error indicates that the referenced SNP is no longer located on the specified chromosome.

- `Getting position of rsXXXX....map data not found in SNP query response`

This error indicates that the referenced SNP is not found in NCBI's public dbSNP database.

- `Can't connect to www.ncbi.nlm.nih.gov:80 (Bad hostname 'www.ncbi.nlm.nih.gov')`

This error indicates that you have a problem with your internet connection. Check the connection and then continue.

7. Copyright, Disclaimer and References

Copyright: Programs and documentation of GrASP are within the public domain. GrASP may be freely distributed and copied. However, it is requested that in any subsequent use of this work, appropriate acknowledgment be given as below.

Disclaimer: No warranty, either expressed or implied, is made with respect to the functioning and accuracy of this program. No responsibility is assumed by the authors. Please report any problems or bugs to the authors.

Acknowledgment: Please acknowledge use of GrASP with the following citation in any publication that uses the software.

Mathias RA, Gao P, Goldstein JL, Wilson AF, Pugh EW, Furbert-Harris P, Dunston G, Malveaux F, Togias A, Barnes KC, Beaty TH, Huang SK. A graphical assessment of p-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. BMC Genet. 2006 Jun 14;7(1):38.

8. Coming Soon to GrASP

Currently, three additional components to GrASP are under development and will be added in the near future:

1) Perl scripts to assist sliding-window based analysis: to include (a) scripts to create input files for sliding-windows based analyses for several of the most commonly used analysis programs; and (b) scripts to summarize the corresponding output and format the results in the form of the input file for GrASP.

2) Sub-selection option: this will allow the user to draw just a restricted portion of the original input files (e.g. draw just SNPs 200-250 in a 600-SNP file) rather than re-create the input file each time (e.g. re-create the input file to now only contain data on SNPs 200-250).

3) The ability to input haplotype block information that will be drawn on the Physical Track. Note: Because the nature of the graphical component to GrASP and Haploview are similar (i.e. the P-value track and the LD graphics are drawn with equal

distance between the SNPs) it is relatively easy for the user in the interim to merge figures from Haploview and GrASP and therefore incorporate block information.