GEVALT Documentation

Table of Contents

Chapter 1. Using GEVALT

Loading a Dataset

Data can be loaded in three formats along with an optional marker info file (except for HapMap files which do not require an info file). Further options are presented on the load screen:

- GEVALT saves time by only computing pairwise LD statistics for markers within a certain distance of each other. The default is 500KB. Enter a value of zero to force all pairwise computations. For dense files with many markers it is recommended to lower the distance in order to reduce memory requirements and computing time.
- GEVALT excludes individuals with less than 50% complete genotypes. This threshold can be adjusted in the load dialog. Additional details about excluded individuals are available from the marker check tab.
- When loading a file dumped from the HapMap project website, it is possible to automatically display SNP and gene tracks from the HapMap above the data by checking the "Download and show HapMap info track" box. More information is available with the LD Display help. [hapmap file only]
- If you wish to perform association tests, you must inform the program now and select either family trios or case/controls. More details are available under association. [pedfile only]
- If your data is completely phased (and without missing data) you can choose not to run Gerbil by checking the "Don't run Gerbil" box. [haps file only]

GEVALT allocates 512MB of memory by default. This is usually sufficient to handle datasets with up to a few thousand markers. If you are running the program on larger datasets you may need to force more memory (presuming your computer has sufficient resources available). This can be accomplished using the following command:

```
java -jar Gevalt.jar -memory 2000
```

Where "2000" in this case specifies 2000 megabytes of memory and can be adjusted as necessary.

Saving And Loading Status

It is possible to save current phasing results and load them in another time. This allows to avoid running Gerbil on the same data file for many times. This option is located under the “file” menu. GEVALT saves the current checked markers and the phasing results of Gerbil. Operations on other tabs (Stampa, Association, etc.) are not saved.

Data Quality Checks

Marker Checks

After loading a linkage or HapMap format file, GEVALT shows some basic data quality checks for the markers. Markers are filtered out based on some default criteria which can be adjusted as necessary. Markers can be added or removed from analyses by hand via the checkboxes.

- # is the marker number.
- **Name** is the marker ID specified (only if an info file is loaded).
- **Position** is the marker position specified (only if an info file is loaded).
- **ObsHET** is the marker's observed heterozygosity.
- **PredHET** is the marker's predicted heterozygosity (i.e. 2*MAF*(1-MAF)).
- **HWpval** is the Hardy-Weinberg equilibrium p value, which is the probability that its deviation from H-W equilibrium could be explained by chance.
- **%Geno** is the percentage of non-missing genotypes for this marker.
- **FamTrio** is the number of fully genotyped family trios for this marker (0 for datasets with unrelated individuals).
- **MendErr** is the number of observed Mendelian inheritance errors (0 for datasets with unrelated individuals).
- **PhaseErr** is the number of phasing errors detected for this marker (see the Phased Genotypes tab for details regarding phasing errors).
- **MAF** is the minor allele frequency (using founders only) for this marker.
- **Rating** is checked if the marker passes all the tests and unchecked if it fails one or more tests (highlighted in red).

You can adjust the filtering thresholds and click "Rescore" to refilter the markers using the new values. Any marker which fails one of the quality tests will have the relevant field(s) highlighted in red. Markers can also be selected/unselected by hand by clicking the "Rating" check box or by using the “Select All” and “Select Range” buttons.

Running Gerbil

Notice that only the selected markers are being phased by Gerbil.
After you complete the selection of markers you should click on "Run Gerbil" to start the phasing of the genotypes.
After the initial phasing, markers can still be selected or unselected. Every time you change the selected markers new phasing is done by gerbil (the re-phasing happens only when you move to another tab and not every time you click a "Rating" check box).

Gerbil's Long Phasing Options

When trying to phase more than 300 SNPs, GEVALT breaks the chromosomes into smaller segments and Gerbil phases each segment separately. Every two consecutive segments are ligated by phasing a short “linker” segment that contains SNPs from the end of the first segment and from the beginning of the second segment. You can configure the long phasing parameters in "Long phasing options" under the Gerbil menu. There are three parameters:

- **Break file with more then <X> SNPs** - If the number of checked markers is larger than X, then the long phasing procedure will be used. The default value is 300. Notice – gerbil cannot be used normally for files longer the 300 SNPs therefore this value cannot exceed 300.

- **Break input file after <X> SNPs** - The length of the segments. High values result less segments and therefor higher accuracy but longer running time. The default value is 100. From our experiments, higher values than 100 result much longer running time and provide very little increase in accuracy, if any.

- **Maximum linker length –** GEVALT chooses linkers with minimal length that provide enough information for determining the correct ligation (i.e. at least 2 heterozygous SNPs in each individual in each side of the linker). This parameter determines the maximal length of the linker from each size (the maximal length of a linker can be twice this parameter value, which means that this value must be equal or lower than the previous parameter).

Duplicate Markers

If two markers in an input file have the same chromosomal position, GEVALT will ignore the less completely genotyped marker by default and highlight both in yellow on the check markers panel. When running in nogui mode GEVALT always ignores the less completely genotyped version of two markers with the same position. If you want to use both from the command line, you'll need to adjust one of the positions.

If two markers in an input file have the same name, GEVALT renames the second one in the file by appending ".X" to the filename, where "X" is a running integer count starting with 1. So if you have marker1, marker1 and marker2, GEVALT would adjust this to: marker1, marker1.1 and marker2. Note that if the markers with the same name have different positions, GEVALT won't deselect any of them; if they do have identical positions, it will filter all but one out as described above.

Filtered Individuals

The top of the tab contains information about individuals filtered during the loading of the file. It will show overview information about the number of singletons and trios used and the number of independent families loaded. Further information about filtered individuals can be shown by clicking the "Show Excluded Individuals" button. This will present a list of excluded individuals as well as the reason for exclusion. Details about individual filtering can be found in Haploview's FAQ.

LD Display

Perusing the LD Display

- The color scheme option (Display menu) allows you to choose among several LD color schemes. The following tables provide details on the color schemes, and a key to the meaning of the currently selected scheme can be dropped down from the "Key" menu in the upper right corner of the screen.

Table 1.1. Standard Color Scheme

	D' < 1	D' = 1
LOD < 2	white	blue
LOD ≥ 2	shades of pink/red	bright red

Table 1.2. Confidence Bounds Color Scheme

Strong Evidence of LD	dark grey
Uninformative	light grey
Strong Evidence of Recombination	white

Table 1.3. r^2 Color Scheme

$r^2 = 0$	white
$0 < r^2 < 1$	shades of grey
$r^2 = 1$	black

Table 1.4. Alternate D'/LOD Color Scheme

	Low D'	High D'
Low LOD	white	shades of pink
High LOD	white	black

(r^2 and Alt D'/LOD courtesy of Will Fitzhugh)

Table 1.5. 4 Gamete Color Scheme

4 distinct 2-marker haplotypes	white
< 4 distinct 2-marker haplotypes	black

- In order to help keep the display uncluttered, D prime values of 1.0 are never shown (the box is empty). The r^2 and Alternate D'/LOD schemes do not show any values in the boxes.
- The LD values are calculated using the phased genotypes. You can calculate the values using the unphased genotypes by choosing that option in the Analysis menu.
- The zoom option (Display menu) allows you to select one of three zoom modes. The two zoomed out versions can be useful for browsing large datasets.
- Large datasets also show a "map" in the lower left corner which gives an overview of the D prime display and allows you to navigate quickly. Clicking on an area of the map will cause the main display to jump to that area. This map also shows the currently defined blocks as small black lines across the top.
- Markers with additional notes (as loaded from the info file) are highlighted (the names are green in the zoomed-in view and the lines from the SNP position to the LD chart are green in the zoomed-out view. Details can be viewed by right clicking on the marker number (as mentioned below).
- Markers which were chosen as tag-SNPs in the Stampa tab are highlighted in blue (overrides the green highlighting).
- Right clicking on the marker number (or the equivalent space in the zoomed out views) shows the marker name, minor allele frequency and any additional notes

specified in the info file. This can be especially helpful in the zoomed out views which do not display marker names.
- Right clicking on any pairwise LD comparison will show a more detailed summary of the LD between the two markers in question.

Additional Data Tracks

Analysis Track

A graph of any variable versus chromosomal location can be added above the LD plot with the "Load Analysis Track" option. Simply create a file with two columns: <position> <value> . GEVALT will plot the values in a continuous line along the top of the screen.

HapMap Gene/SNP Track

The "Download HapMap info track" option (with an internet connection) allows you to connect to the HapMap Project server and download and display a track with HapMap genotyped SNPs and gene names. If an info file is specified, the default boundaries are the positions of the first and last markers (which is only valid if the info file is in genomic coordinates). You must specify the proper chromosome in the dialog box. If you are using a file downloaded from the HapMap website the program will specify the correct default chromosome and start/end positions. This track display can be configured with the "HapMap Info Track Options" item in the "Display" menu. Available tracks include HapMap SNPs, gene names, mRNAs, contigs, gaps and GC content.

Blocks and Haplotypes

Blocks

GEVALT generates blocks whenever a file is opened, but these blocks can be edited and redefined in a number of ways. In the Analysis menu, you can clear all the blocks in order to start over, define blocks based on one of several automated methods or customize the parameters of those algorithms. Additionally, the blocks can be edited by hand.

Gerbil [DEFAULT]

While phasing the genotypes, Gerbil also generates blocks. These blocks are the default selection. When the long phasing procedure is used the blocks are generated in each segment separately. It means that blocks at the beginning and end of each segment might be inaccurate.

Confidence Intervals

This algorithm is taken from Gabriel et al, Science, 2002. 95% confidence bounds on D prime are generated and each comparison is called "strong LD", "inconclusive" or "strong recombination". A block is created if 95% of informative (i.e. non-inconclusive) comparisons are "strong LD". This method by default ignores markers with MAF < 0.05. The MAF cutoff and the confidence bound cutoffs can be edited by choosing "Customize Block Definitions" (Analysis menu). This definition allows for many overlapping blocks to be valid. The default behavior is to sort the list of all possible blocks and start with the largest and keep adding blocks as long as they don't overlap with an already declared block.

Four Gamete Rule

This is a variant on the algorithm described in Wang et al, Am. J. Hum. Genet., 2002. For each marker pair, the population frequencies of the 4 possible two-marker haplotypes are computed. If all 4 are observed with at least frequency 0.01, a recombination is deemed to have taken place. Blocks are formed by consecutive markers where only 3 gametes are observed. The 1% cutoff can be edited to make the definition more or less stringent.

Solid Spine of LD

This internally developed method searches for a "spine" of strong LD running from one marker to another along the legs of the triangle in the LD chart (this would mean that the first and last markers in a block are in strong LD with all intermediate markers but that the intermediate markers are not necessarily in LD with each other).

Markers can be removed from blocks by clicking on the marker number (along the top of the D prime graph). Blocks can be defined by hand by clicking and dragging along the marker number row. Any block which overlaps with an existing block will take precedence and delete the existing block.

Haplotypes

Display

View haplotypes for selected blocks by clicking on the "Haplotypes" tab or selecting "Haplotypes" from the Display menu. Haplotypes are estimated simply by using the phased genotypes.

The haplotype display shows each haplotype in a block with its population frequency and connections from one block to the next. In the crossing areas, a value of multiallelic D' is shown. This represents the level of recombination between the two blocks. Note that the value of multiallelic D' is computed for only the haplotypes ("alleles") currently displayed. This usually does not have a strong effect, as the rare haplotypes contribute only slightly to the overall value. Above the haplotypes are marker numbers along with a tick beneath haplotype tag SNPs (htSNPs).

Display Controls

The display can be edited using the controls at the bottom of the screen to display only more common haplotypes or to adjust the connecting lines. By default, alleles are displayed using A,C,G,T. The display can also be changed to show the alleles numerically from 1-4, or as blue and red boxes, with blue being the major allele and red the minor.

Tag SNPs

The htSNPs are selected on a block-by-block basis. This means that the end set is not necessarily the most parsimonious one for the entire dataset, but it provides for a much more thorough testing set if you plan to move from an initial sample used to pick htSNPs to a much larger sample, that is, it is much more likely to catch variation in the new, large dataset that was not observed in the initial dataset. Preference for this method or another depends on experimental design.

Specifically, the SNPs in each block are ranked in order of completeness of genotyping and then a set is selected which defines all haplotypes above a certain (adjustable) frequency threshold. The method ensures the most efficient possible set for strict 4 gamete blocks, but occasionally features redundancies of larger and more loosely defined blocks (where some recombination has occurred).

For a parsimonious set for the entire dataset use the Stampa algorithm in the Stampa tab.

Phased Genotypes

Display

View the phased genotypes by clicking on the "Phased Genotypes" tab or selecting "Phased Genotypes" from the Display menu.

The phased genotypes display shows two chromosomes for every individual. For family trios, only the two parents are shown. The "child" field states for each chromosome whether it was transmitted (the name of the child to whom it was transmitted appears) or untransmitted.
Alleles whose phasing was uncertain and were phased by gerbil have a green or white background. white for alleles that were originally missing and green for an heterogeneous locus.

If the input file was in linkage format, then the display is divided into two tabs: parents and children. The parents tab shows the parents in every family (including singletons) as described above. The children tab shows only the children in every family. For every family, the first child shown is the child that was chosen for the phasing of his parents. then the rest of the children in the family are shown (these children were excluded and marked as "Not a member of maximum unrelated subset.").

Alleles coloring

The alleles have different background colors, according to the way they were phased:

- gray - phasing is certain (a homozigous locus or phasing was deduced from a trio).
- green - heterogeneous locus that can't be deduced from a trio. phasing is done by gerbil.
- white - data is missing and can't be deduced from a trio. the data is completed by gerbil.
- pink - a Mendelian error was found in this locus. data is considered missing and is completed by gerbil.
- cyan (children tab only) - phasing of the children that were marked as "Not a

member of maximum unrelated subset." is done by finding a chromosome from their parents that best fits their genotype. Still, that chromosome can differ from the child's chromosome in several markers. These markers have a cyan background.

Output files

The "Dump phased genotypes" button exports a file with the phased haplotypes in the format accepted as input by GEVALT.

The "Export Tab to Text" option in the File menu will export a summary file showing the phased genotypes as they appear on screen.

Display Controls

By default, alleles are displayed using A,C,G,T. The display can also be changed to show the alleles numerically from 1-4, or as blue and red boxes, with blue being the major allele and red the minor.

Individual Statistics

View summary statistics for every individual by clicking on the "Individual stats" tab or selecting "Individual stats" from the Display menu.

All individuals are shown, except individuals that were excluded because of low percentage of complete genotypes.
The calculated Statistics include:

- **%Miss** - percentage of missing genotypes.
- **%Het** - percentage of heterogeneous markers (out of the non-missing markers).
- **%MA** - percentage of minor alleles (out of the non-missing alleles).
- **MendErr** - number of observed Mendelian inheritance errors.
- **PhaseErr** - number of phasing errors (see the Phased Genotypes tab for details regarding phasing errors).

N.B. Mendelian inheritance errors are also counted as missing genotypes. This is because gerbil considers these genotypes as missing and completes them regardless of the original genotypes.

Stampa

Introduction

The Stampa algorithm finds a set of tag SNPs for the entire data set. Details on the algorithm can be found in Stampa's paper (see references).

Stampa Configuration Panel

This panel shows all SNPs available for tag selection. SNPs which are deselected in the Check Markers tab will not be in this list. Each SNP can be either “force included” or “force excluded”, meaning that the SNP will be chosen as a tag SNP or not chosen, respectively, in all of Stampa's solutions. The forced tags configuration can also be loaded from a file.

Below the marker list are several additional tagging options. You can set the minimum and maximum number of tags. Stampa finds solutions for every number of tags between these two parameters. The default is from 2 to the number of markers. When the number of markers is high, limiting the solution to a reasonable number of tags will speed up the running time of Stampa. You can also set the maximal allowed distance between tags. Longer distance can produce better results on the expense of a longer running time. When no info file is provided, the distance is measured in number of SNPs.

Clicking "Run Stampa" will run Stampa. When finished it will switch from the Configuration to the Results Panel. When Stampa is running it is possible to stop the run by pushing the “Stop Stampa” button (the “run Stampa” button changes into “Stop Stampa”).

Stampa Results Panel

This panel is composed of two tables: prediction table and tags table.

The prediction table contains Stampa's prediction accuracy for every number of tags. If there is no possible solution for a certain number of tags then “---” will be displayed. Select the required number of tags in this table and press the “Show tags” button in order to see the selected tags in the tags table.

In the tags table you can also pick or unpick tags manually. Pressing the “Calculate prediction” button will calculate the prediction accuracy for the current set of tags. Notice that in most cases adding tags manually will reduce the prediction accuracy because the prediction of each non-tag SNP is made by using its two flanking tags. Therefor, it is recommended to use Stampa's solutions instead of adding or removing tags manually.

The selected tags for every number of tags and their prediction accuracy can be saved to a file by pressing the “Dump tag SNPs to file” button.

Tagger

Introduction

Haploview have developed a tagging strategy that combines the simplicity of pairwise methods with the potential efficiency of multimarker approaches. It avoids overfitting and unbounded haplotype tests in the association phase by (*a*) using only those multiallelic combinations in which the alleles are themselves in strong LD, and (*b*) explicitly recording the allelic hypotheses that are to be tested in the subsequent association analysis. Attractive practical features include the ability to force in or exclude sets of tags.

It is based on Paul de Bakker's *Tagger*. It and more information are available at the

Tagger website. There are a number of differences between the implementations, although they are constructed around the same concept. Tagger currently searches a much broader space of available multi-marker tests (up to 6-mers) whereas this version allows only 2- or 3-marker tests in the interest of computational efficiency.

Features

Tagger operates in either pairwise or aggressive mode. In either case it begins by selecting a minimal set of markers such that all alleles to be captured are correlated at an r^2 greater than a user-editable threshold with a marker in that set. Certain markers can be forced into the tag list or explicity prohibited from being chosen as tags. You can also specify which markers in the dataset you want to be captured.

Aggressive tagging introduces two additional steps. The first is to try to capture SNPs which could not be captured in the pairwise step (N.B. these must have been "excluded" since otherwise they would simply be chosen to capture themselves) using multi-marker tests constructed from the set of markers chosen as pairwise tags. After this, it tries to "peel back" the tag list by replacing certain tags with multi-marker tests. Tagger avoids overfitting by only constructing multi-marker tests from SNPs which are in strong LD with each other, as measured by a pairwise LOD score. This LOD cutoff can be adjusted to loosen or tighten this requirement; in general, the default cutoff of 3.0 is appropriate for selecting tags from a HapMap-sized reference panel of 120 chromosomes.

Much more information about the development of this algorithm is available at the Tagger website.

Tagger Configuration Panel

N.B. Tagger requires either an info file or a hapmap style input file, because it references the marker names specified in those files. If you load a pedigree or phased haplotypes input file without an info file, the Tagger panels will not be available.

This panel shows all SNPs available for tag selection. SNPs which are deselected in the Check Markers tab will not be in this list. There are three checkboxes for each SNP:

Force Include

> Checking this box will force this SNP to be chosen as a tag SNP.

Force Exclude

> Checking this box will prohibit this SNP from being chosen as a tag SNP.

Capture this Allele?

> If this box is checked, GEVALT will include this SNP in the list of alleles to be captured by the chosen tag set.

N.B. The include and exclude checkboxes are mutually exclusive, and "Capture this Allele" must be checked in order to either include or exclude a marker.

Below the marker list are several additional tagging options. You can choose from among pairwise and two aggressive tagging strategies discussed above. You can also set the r^2 and LOD thresholds as previously mentioned. Clicking "Run Tagger" will run the tagging algorithm. When finished it will switch from the Configuration to the Results Panel.

Tagger Results Panel

This panel is split into a "Tests" section on the left and a marker-by-marker report on the right. The marker report lists all SNPs, the test which best captures them, and their r^2 with that test. SNPs which were unchecked from the "Capture this allele?" list on the Configuration panel are greyed out. SNPs which could not be successfully tagged are shown in red.

The first list in the "Tests" section shows all the tests (both single marker and multi-marker alleles) chosen by GEVALT. Selecting tests in this list will show which alleles are captured by those tests in the second list in the panel. Beneath these lists is a summary of the tagging results.

Captured N alleles with mean r^2 of X.

> This shows how many of the SNPs in the dataset have been successfully tagged by the set of chosen tests. The mean r^2 represents the mean for only those SNPs successfully captured.

Captured N percent of alleles with r^2 >0.8

> This shows what fraction of the alleles captured by the tests have an r^2 >= 0.8. Of course, if your tagging r^2 threshold is >= 0.8 this value will always be 100%.

Using N SNPs in M tests.

> This shows that N unique SNPs have been chosen to create M tests, which can either be one of the set of N SNPs or some combination of those SNPs.

The "Dump Tests File" button exports a file with the list of tests in the format used by Tagger's export.

The "Export Tab to Text" option in the File menu will export a summary file showing the best tag for each marker and the list of tests along with the alleles tagged by each test.

Association Tests

If selected when loading the data, GEVALT computes single locus and multi-marker haplotype association tests. For case/control data, the chi square and p-value (uncorrected) for the allele frequencies in cases vs. control are shown. For family trios, all probands (affected individual with genotyped parents) are used to compute TDT values (including individuals that were exluded for the reason of "Not a member of maximum unrelated subset").

The haplotype association test is performed on the set of blocks selected on the LD and haplotype tabs. Results are calculated only for those haplotypes above the display threshold on the haplotype tab. Haplotypes below that threshold are counted as their most similar unfiltered haplotype.

Note that in the case/control test missing data is being completed by gerbil.

GEVALT is not intended to be the only way of testing association results, but to provide a straightforward way to do simple association tests. It's always a good idea to try out multiple approaches to analyzing your data.

Permutation Testing

GEVALT provides a framework for permuting your association results in order to obtain a measure of significance corrected for multiple testing bias. You can choose to permute one of two test sets:

Single Markers Only

> Permute just association tests to the individual SNPs in your dataset.

Single Markers and Blocks

> Permute the individual SNPs as above, along with all the blocks shown in the Haplotypes tab.

Specify how many permutations to do and press the "Do Permutations" button to start the permutations.

Once the permutations are complete, GEVALT displays:

- A table listing all tests (single SNP and blocks) along with their association chi squares.
- The permutation p-value.

In a case\control test with a Single Markers Only test set you can choose the “Rapid Test” option which runs the RAT software instead of the normal permutation algorithm. Only a few permutations (around 100) are enough to achieve a very accurate p-value regardless of its value. You can choose to sample columns in order to further reduce the running time. See RAT's paper for details about the algorithm.

You can stop the permutations at any time with the "Stop" button, but then no results will be displayed.
You can save the permutation summary by using the "Export Tab to Text" option in the File menu.

Chapter 2. Files

Input File Formats

GEVALT currently accepts input data in three formats, standard linkage format, completely or partially phased haplotypes and HapMap Project data dumps. It also takes in a separate file with marker position information, as well as several auxiliary input files, described below. The four formats are explained in depth below.

Linkage Format

Linkage data should be in the Linkage Pedigree (pre MAKEPED) format, with columns of family, individual, father, mother, gender, affected status and genotypes. The file should not have a header line (i.e. the first line should be for the first individual, not the names of the columns). Please note that GEVALT can only interpret biallelic markers — markers with greater than two alleles (e.g. microsatellites) will not work correctly. A sample line from such a file might look something like:

```
3    12    8    9    1    2    1 2    3 3    0 0    4 2
a     b    c    d    e    f    -----------g------------
```

(a) pedigree name

A unique alphanumeric identifier for this individual's family. Unrelated individuals should not share a pedigree name.

(b) individual ID

An alphanumeric identifier for this individual. Should be unique within his family (see above).

(c) father's ID

Identifier corresponding to father's individual ID or "0" if unknown father. Note that if a father ID is specified, the father must also appear in the file.

(d) mother's ID

Identifier corresponding to mother's individual ID or "0" if unknown mother Note that if a mother ID is specified, the mother must also appear in the file.

(e) sex

Individual's gender (1=MALE, 2=FEMALE).

(f) affectation status

Affectation status to be used for association tests (0=UNKNOWN, 1=UNAFFECTED, 2=AFFECTED).

(g) marker genotypes

Each marker is represented by two columns (one for each allele, separated by a space) and coded 1-4 where: 1=A, 2=C, 3=G, T=4. A 0 in any of the marker genotype position (as in the the genotypes for the third marker above) indicates missing data.

It is also worth noting that this format can be used with non-family based data. Simply use a dummy value for the pedigree name (1, 2, 3...) and fill in zeroes for father and mother ID. It is important that the "dummy" value for the ped name be unique for each individual. Affectation status can be used to designate cases vs. controls (2 and 1, respectively).

Files should also follow the following guidelines:

- Families should be listed consecutively within the file (i.e. all the lines with the same pedigree ID should be adjacent)
- If an individual has a nonzero parent, the parent should be included in the file on his own line.

Phased Haplotypes

Haplotype data for GEVALT's input must be formatted in columns of Family, Individual and Genotypes. There should be two lines (chromosomes) for each individual. This is the standard format of Genehunter's TDT output. See the sample below:

```
FAM1    FAM1M01    0    4    2    2
FAM1    FAM1M01    0    4    2    2
FAM1    FAM1F02    3    h    1    2
FAM1    FAM1F02    3    h    1    2
```

The data format uses the numerals 1-4 to represent genotypes, the number zero to represent missing data, and the letter "h" to represent a heterozygous allele. That is, if an individual is heterozygous at a locus, both alleles should be "h" if the phasing (which allele falls on which chromosome) is uncertain.

HapMap Project Data Dumps

Data from the HapMap Project can be dumped by region using the GBrowse interface. Downloading data requires user registration and agreement to the terms of use. The saved data file is in a marker-per-line format which can be loaded in GEVALT.

GBrowse dumps only one file, which has one marker per line and which includes familial relationships among the HapMap samples as well as marker position information. The file format has several header lines (beginning with "#") which GEVALT parses. Open the file by selecting "Browse HapMap Data" option and selecting the downloaded file.

Marker Information File

The marker info file is two columns, marker name and position. The positions can be either absolute chromosomal coordinates or relative positions. It might look something like this:

```
marker01 190299
marker02 190950
marker03 191287
```

An optional third column can be included in the info file to make additional notes for specific SNPs. SNPs with additional information are highlighted in green on the LD display. For instance, you could make note that the first SNP is a coding variant as follows:

```
marker01 190299 CODING_SNP
marker02 190950
marker03 191287
```

Batch Load File

The "-batch" flag on the command line allows you to run GEVALT automatically (in nogui mode) on several files. Batch input files should have one genotype file per line, along with an info file (if desired) separated by a space. Filenames must conform to the following rules:

- Pedfile names must end in ".ped"
- Phased haplotype file names must end in ".haps"
- HapMap file names must end in ".hmp"
- Info file names must end in ".info"

The following example shows 2 pedfiles (with info files) and a hapmap file:

```
sample1.ped   sample1.info
sample2.ped   sample2.info
sample3.hmp
```

Output Files

For any given tab the information in the display can be saved. For the data check, stampa, tagger and association test tabs, a simple tab-delimited text file is generated from the tables. For the LD, Phased Genotypes and Haplotype tabs, data can either be dumped to text files or the image can be saved to a PNG.

LD Text Output File

LD text output is a tab delimited set of columns containing the various measures of LD used by the program. Details for each column are shown below:

- `L1` and `L2` are the two loci in question, referenced by their number or name (if marker info file is provided)
- `D'` is the value of D prime between the two loci.

- `LOD` is the log of the likelihood odds ratio, a measure of confidence in the value of `D'`
- r^2 is the correlation coefficient between the two loci
- `CIlow` is 95% confidence lower bound on `D'`
- `CIhi` is the 95% confidence upper bound on `D'`
- `Dist` is the distance (in bases) between the loci, and is only displayed if a marker info file has been loaded
- `T-int` is a statistic used by the HapMap Project to measure the completeness of information represented by a set of markers in a region

Details about additional options for this output type can be found below in the Export Options section.

LD PNG Output

When saving the LD table to a PNG, GEVALT saves an image using the current display settings. This includes color scheme, zoom and proportional spacing. Thus, in order to save a less detailed image to a PNG, first zoom out, then export the tab. Note that GEVALT cannot save large datasets at the higher zoom levels. For more information see the Export Options section below.

Haplotype Text Output File

Haplotype output shows a block, its markers, the haplotypes and their population frequencies, the crossover percentages to the next block and the multiallelic D prime. Tag SNPs are denoted with a "!". Crossover percentages are shown as a matrix with this block's haplotypes as the rows and the next block's haplotypes as the columns. An example might look like:

```
BLOCK 1.  MARKERS: 1 2 3! 4!
3312 (0.825)    |0.800  0.025   0.000|
1144 (0.163)    |0.031  0.125   0.007|
3342 (0.013)    |0.006  0.000   0.006|
Multiallelic Dprime: 0.802
BLOCK 2.  MARKERS: 10! 11! 12
441 (0.837)
222 (0.150)
242 (0.013)
```

In this example, the first block has 4 markers with 3 haplotypes displayed and the second block has 3 markers and 3 haplotypes. The tag SNPs for each block are (3,4) and (10,11) respectively. The crossover percentage matrix can be read as follows: 80% of all samples have the pattern 3312-441, 3.1% have the pattern 1144-441 and so forth.

Haplotype PNG Output

Saving the haplotype tab to a PNG produces an image using the current display settings (such as haplotype frequency cutoff).

Phased Genotypes Text Output File

The phased genotypes are saved in a text file as they appear on screen. If the alleles are displayed as colored squares then they are saved as numbers.

Phased Genotypes Dump

This file is the same format as the Phased Haplotypes input file accepted by GEVALT.

Phased Genotypes PNG Output

Saving the Phased Genotypes tab to a PNG simply produces an image similar to the image displayed on screen.

Individual Statistics Text Output File

The individual statistics table is saved in a text file as it appears on screen.

Stampa Results Text Output File

The stampa results output first shows the amount of tag SNPs selected and the prediction percentage. then the SNPs table from the results panel is saved as a tab-delimited table.

Stampa Results Dump

The selected tags for every number of tags and their prediction accuracy are saved to a file by pressing the "Dump tag SNPs to file" button.

Single Marker Association Text Output File

Single marker association results are saved in a tab-delimited text file with the following columns:

- `#` is the marker number.
- `Name` is the marker ID specified if an info file is loaded.
- `Chi Square` is the chi square value for the marker.
- `uncorrected p value` is the significance level for the above chi square.

Haplotype Association Text Output

Haplotype association text output is a tab-delimited file, broken into sections by block. The columns are:

- `Haplotype` is the sequence of alleles for this haplotype in this block.
- `Frequency` is the population frequency for this haplotype.
- `Chi Square` is the chi square value for the haplotype.
- `uncorrected p value` is the significance level for the above chi square.

Permutation Text Output File

The output from the permutations tab shows the number of permutations performed, the best observed chi-square, the permutation's p-value and then a tab-delimited table with one row per permuted test and the following columns:

- Name is the test name, which is either a marker name or a comma separated list of marker names then a tab then a comma separated set of alleles for those markers.
- Chi Square is the observed association chi square for that test.

Tagger Text Output File

The Tagger text output begins with several pieces of summary information. More details on this can be found in the Tagger section. The rest of the output is divided into two sections. The first lists each marker, with the following rows:

- Marker is the marker name.
- Best Test is the test with the highest r^2 to this marker.
- r^2 w/test is the r^2 between this marker and its test.

The second part consists of a list of the tests and the alleles they capture best.

Tagger Tests Dump

This file is the same format used by Haploview for custom association tests and exported by Tagger. It is discussed below in the auxiliary files section.

Marker Check Text Output File

The marker check data is a tab-delimited file with the following columns:

- # is the marker number.
- Name is the marker ID specified (only if an info file is loaded).
- Position is the marker position specified (only if an info file is loaded).
- ObsHET is the marker's observed heterozygosity.
- PredHET is the marker's predicted heterozygosity (i.e. 2*MAF*(1-MAF)).
- HWpval is the Hardy-Weinberg equilibrium p value, which is the probability that its deviation from H-W equilibrium could be explained by chance.
- %Geno is the percentage of non-missing genotypes for this marker.
- FamTrio is the number of fully genotyped family trios for this marker (0 for datasets with unrelated individuals).
- MendErr is the number of observed Mendelian inheritance errors (0 for datasets with unrelated individuals).
- PhaseErr is the number of phasing errors detected for this marker (see the Phased Genotypes tab for details regarding phasing errors).
- MAF is the minor allele frequency (using founders only) for this marker.
- Rating is "BAD" if the marker failed any of the above tests and blank otherwise.

Export Options

The "Export Options" item in the File Menu allows adjustment of several parameters and allows the user to save any tab without having to switch to it. Specifically, the LD tab allow the markers to be filtered to output only some of the markers:

All

> The default setting (and only one available for most tabs) is to use all the markers.

Marker Range

> Generates the LD text or PNG file for only a specific range of markers.

Adjacent Markers

> Generates the LD text file for only adjacent markers. This can be useful to view the T-int stat, which measures LD information content in the gaps between markers.

There is also an option to generate a "compressed" LD PNG, which is useful for very large datasets. The image is shrunk to an arbitrary zoom level which allows GEVALT to save the PNG with minimal memory usage.

Auxiliary Input Files

Blocks File

> You can specify a set of blocks by loading a blocks file. Each line is a space separated list of markers with one block per line. For example:

```
1 2 3 4
9 10 11 12 13 14 15
```

> Would create one block from markers 1-4 and another from 9-15. The first marker in the file is number 1 (not 0).

Analysis Track File

> You can add an analysis track along the top of the LD display by loading a file with two columns, <position> <value>. GEVALT will plot the values continuously with respect to the positions of the markers, so the positions should use the same coordinates as the marker info file. For example:

```
1000 0.3
2000 1.7
3000 11.0
4000 2.3
5000 4.6
```

Would plot a line from position 1000 to 5000. The values can be of any units or magnitude, as the GEVALT scales the analysis track to the bounds of the values.

Stampa Marker force Include/Exclude File

You can specify a list of tag SNPs numbers for Stampa to force include/exclude. This file can be created by using the "save configuration" button under the Stampa tab and can be loaded using the "load configuration" button. The following file can could be used to force include tags 5 and 6 and to force exclude tags 7 and 8.

```
forceInclude: 5 6

forceExclude: 7 8
```

Notice that both lines must appear in this specific order. In case you want only to exclude tags, your file should be look generally like this:

```
forceInclude:

forceExclude: 7 8
```

The same thing applies for the second line.

Notice that each tag number is separated from the next with a tab character.

Tagger Marker Include/Exclude File

You can specifiy a list of markers for Tagger to include or exclude from those markers available for selection as tag SNPs. In either case the format is the same: one marker name per line. The following file could be used to either include or exclude markers 1,7 and 9:

```
marker1
marker7
marker9
```

N.B. Using a Tagger Include/Exclude File requires a marker info file, since it reads the marker names as specified in the info file.

Custom Association Tests File

Custom association tests are not currently supported by GEVALT. However, this format is exported by GEVALT using the "Dump Tests" button in the Tagger Results panel and by Paul deBakker's Tagger webpage. The format is one test per line with each line containing one of the following: a single marker name, several marker names separated by commas or several comma separated names, then a tab, then comma separated alleles for each marker.

For instance, the following example would create 5 tests: markers 1, 2 and 3 individually, all the alleles (haplotypes) of the block 4,5,6 and the CAA haplotype of the block 12,13,14:

```
marker1
marker2
marker3
marker4,marker5,marker6
marker12,marker13,marker14     2,1,1
```

N.B. Using a Custom Association Tests File requires a marker info file, since the tests file reads the marker names as specified in the info file.

Chapter 3. Command Line Options

GEVALT can be run from the command line without the display in order to do processing of multiple datasets or quick computation on very large datasets. In order to run GEVALT without the display, add the "-nogui" flag. The "-help" flag shows a condensed explanation of all the command line options explained below.

General Options

-h, -help

Print help information.

-n, -nogui

Command line mode—does not launch display.

-q, -quiet

Quiet mode—minimizes output to command line.

-memory <memsize>

Allocate <memsize> megabytes of memory to the GEVLAT process (default is 512MB).

Input Options

-pedfile <filename>

Specify a genotype input file in pedigree format. This option works in GUI mode.

-hapmap <filename>

Specify a HapMap format input file. This option works in GUI mode.

-haps <filename>

Specify a phased input file. This option works in GUI mode.

-info <filename>

Specify a marker information file. This option works in GUI mode.

-batch <filename>

Specify a batch load file.

-blocks <filename>

Specify a block definition file. This will automatically use this block definition for haplotype output.

-track <filename>

Specify an analysis track file

Data Check Options

-skipcheck

Skip all the genotype data quality checks and uses all markers for all analyses.

-minMAF <threshold>

Exclude all markers with minor allele frequency below <threshold>, which must be between 0 and 0.5. Default of 0. This option works in GUI mode.

-maxMendel <integer>

Exclude markers with greater than <integer> Mendelian inheritance errors. Default of 1. This option works in GUI mode.

-minGeno <threshold>

Exclude markers with less than <threshold> fraction of nonzero genotypes. <threshold> must be between 0 and 1 with a default of 0.5. This option works in GUI mode.

-hwcutoff <threshold>

Exclude markers with a Hardy Weinberg p-value less than <threshold>, which ranges from 0 to 1 with a default of 0.001 This option works in GUI mode.

-maxDistance <distance>

Maximum intermarker distance for LD comparisons (in kilobases). Default is 500. This option works in GUI mode.

-missingCutoff <threshold> c

Exclude individuals with more than <threshold> fraction missing data, where <threshold> is a value between 0 and 1 with a default of 0.5. This option works in GUI mode.

Gerbil Options

-skipGerbil

Don't run gerbil. Only for completely phased data in haps format.

-gerbilSNPsThreshold <number>

Use the long phasing procedure if the number of markers is larger than this value (default 300)

-gerbilBreakFactor <number>

Size of the segments in the long phasing procedure (default 100)

-gerbilMaxLinker <number>

Maximal size of the linker from both sides (default 50)

Output Options

-blockoutput <type>

Generate haplotypes and population frequencies for blocks of <type>, which can be GER(Gerbil blocks), GAB (Gabriel et al), GAM (4 gamete blocks), SPI (solid spine blocks) or ALL (each of the previous 4). The default block type is Gerbil. More information can be found with the blocks help.

-blockCutHighCI <thresh>

Gabriel 'Strong LD' high confidence interval D' cutoff.

-blockCutLowCI <thresh>

Gabriel 'Strong LD' low confidence interval D' cutoff.

-blockMAFThresh <thresh>

Gabriel MAF threshold. Markers below this allele frequency will be skipped in building Gabriel blocks.

-blockRecHighCI <thresh>

Gabriel recombination high confidence interval D' cutoff.

-blockInformFrac <thresh>

Gabriel fraction of informative markers required to be in strong LD.

-block4GamCut <thresh>

4 Gamete block cutoff for frequency of 4th pairwise haplotype.

-blockSpineDP <thresh>

Solid Spine blocks D' cutoff for 'Strong LD'.

-check

Output marker quality checks to <inputfile>.CHECK

-indStats

Outputs individual statistics to <inputfile>.IND_STATS

-unphased

Calculates LD using unphased genotypes.

-dprime <BOTH>

Output pairwise LD text table to <inputfile>.LD. If BOTH is added then LD is calculated twice: using unphased genotypes and phased genotypes. the output files are <inputfile>.LD_PHASED and <inputfile>.LD_UNPHASED.
Note that -dprime and -check default to no haplotype output unless the -blockoutput flag is also specified.

-png <BOTH>

Output PNG image file of LD display to <inputfile>.LD.PNG. If BOTH is added then two files are created: <inputfile>.LD_PHASED.PNG and <inputfile>.LD_UNPHASED.PNG

-compressedpng <BOTH>

Output low-resolution (smaller file) PNG image of LD display to <inputfile>.LD.PNG. If BOTH is added then two files are created: <inputfile>.LD_PHASED.PNG and <inputfile>.LD_UNPHASED.PNG

-spacing <threshold>

Use proportional spacing for dumped LD pngs. <threshold> ranges from 0 (no spacing) to 1 (max spacing) with a default of 0.

-ldcolorscheme <type>

Use a particular color scheme for dumped LD pngs. <type> can be DEFAULT, RSQ, DPALT, GAB or GAM. More information can be found with the LD display help

-phasedGen <LET,NUM>

Output phased genotypes text to <inputfile>.PHASED_GENOTYPES. If a ped file was loaded, outputs two files: PHASED_GENOTYPES_PARENTS and PHASED_GENOTYPES_CHILDREN.

Alleles are displayed as letters(LET) or numbers(NUM). The default is letters.

-phasedGenPNG <LET,NUM,COL>

Output phased genotypes text to <inputfile>.PHASED_GENOTYPES.PNG. If a ped file was loaded, outputs two files: PHASED_GENOTYPES_PARENTS and PHASED_GENOTYPES_CHILDREN.

Alleles are displayed as letters(LET), numbers(NUM) or colored squares(COL). The default is letters.

-stampa <min number of tags> <max number of tags>

Runs stampa to find the specified range of number of tag SNPs and outputs results in <inputfile>.STAMPA.

-stampaIncludeTags <tag1, tag2, tag3...>

Comma seperated list of force included tags

-stampaIncludeTagsFile <file>

A file containing a list of force inlcluded tag SNPs (one in each line)

-stampaExcludeTags <tag1, tag2, tag3...>

Comma seperated list of force excluded tags

-stampaExcludeTagsFile <file>

A file containing a list of force excluded tag SNPs (one in each line)

-maxDistBetTags <number>

Maximal distance in kb between Stampa's tags. If an info file is not available the distance is measured in number of SNPs.

-assocCC

Output case/control association results. Saves single marker results to <inputfile>.ASSOC and haplotype results to <inputfile>.HAPASSOC. Haplotype association results are not generated if block type is set to ALL.

-rapidAssocTest

For case/control tests only, runs a rapid test using RAT program.

-rapidAssocTestSampleCols <num>

For case/control rapid test only, defines how many columns will rat sample.

-assocTDT

Output TDT association results. Saves single marker results to <inputfile>.ASSOC and haplotype results to <inputfile>.HAPASSOC. Haplotype association results are not generated if block type is set to ALL.

-permtests <numtests>

Performs <numtests> permutations on default association tests and writes to <inputfile>.PERMUT

-pairwiseTagging

Generates pairwise tagging information in <inputfile>.TAGS and .TESTS

-aggressiveTagging

As above but also allows 2- and 3-marker haplotype tags.

-includeTags <markers>

Forces in a comma separated list of marker names as tags.

-includeTagsFile <file>

Forces in a file of one marker name per line as tags.

-excludeTags <markers>

Excludes a comma separated list of marker names from being used as tags.

-excludeTagsFile <file>

Excludes a file of one marker name per line from being used as tags.

-taglodcutoff <thresh>

Tagger LOD cutoff for creating multimarker tag haplotypes.

-tagrsqcutoff <thresh>

Tagger r^2 cutoff.

-hapthresh <threshold>

Only output haplotypes with frequency ≥ <threshold>. Note that multiallelic D' and htSNPs are computed using only displayed haplotypes.

-excludeMarkers <markers>

Exclude markers in a comma separated list with ranges specified as start..end. So, to exclude markers 3, 5 and 10 through 15 you'd use "-excludeMarkers 3,5,10..15"

Chapter 4. About GEVALT

GEVALT was developed in and is maintained by Ron Shamir's lab at the Tel-Aviv University by Ofir Davidovich, Gad Kimmel, Eran Halperin, and Oded apel. Questions and comments should be addressed to: gevalt@cs.tau.ac.il

GEVALT's JAVA source code is based on Haploview's source code.

References

Gad Kimmel and Ron Shamir.
Maximum Likelihood Resolution of Multi-block Genotypes.
In Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 04), pages 2-9.
The Association for Computing Machinery, 2004.

Gad Kimmel and Ron Shamir.
GERBIL: Genotype Resolution and Block Identification Using Likelihood.
Proceedings of the National Academy of Sciences of the United States of Ameirca (PNAS) 102: 158-162, 2005.

Eran Halperin, Gad Kimmel and Ron Shamir.
Tag SNP Selection in Genotype Data for Maximizing SNP PredictioN Accuracy.
Bioinformatics 21(Suppl 1): i195-i203, 2005.

Gad Kimmel and Ron Shamir.
A Fast Method for Computing High Significance Disease Association in Large Population-Based Studies.
American Journal of Human Genetics 79: 481, 2006.

Barrett JC, Fry B, Maller J, Daly MJ.
Haploview: analysis and visualization of LD and haplotype maps.
Bioinformatics. 2005 Jan 15 [PubMed ID: 15297300]

System Requirements

It is recommended that GEVALT be run on a machine with at least 128M of memory.

GEVALT requires Java JRE 1.5 or later. It is worthwhile in any case to download the most recent Java release.

Updates

If you have an internet connection, GEVALT will automatically check for an update upon startup. If a new version is available, it will show a message in the lower right corner of the screen for a few seconds. Details can be found by using the "Check for Updates" button in the File menu.

In order to reach the downloads page without going over the registration process again, click on the "Open downloads page" button in the File menu or open the "updates.html" file in the Gevalt directory.