BMC Bioinformatics

BioMed Central

Software

Open Access

GEVALT: An integrated software tool for genotype analysis

Ofir Davidovich*, Gad Kimmel and Ron Shamir

Address: School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel

Email: Ofir Davidovich* - offirdav@post.tau.ac.il; Gad Kimmel - kimmel@cs.berkeley.edu; Ron Shamir - rshamir@post.tau.ac.il
* Corresponding author

Published: 1 February 2007

BMC Bioinformatics 2007, **8**:36 doi:10.1186/1471-2105-8-36

Received: 3 November 2006
Accepted: 1 February 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/36

Abstract

Background: Genotype information generated by individual and international efforts carries the promise of revolutionizing disease studies and the association of phenotypes with alleles and haplotypes. Given the enormous amounts of public genotype data, tools for analyzing, interpreting and visualizing these data sets are of critical importance to researchers. In past works we have developed algorithms for genotypes phasing and tag SNP selection, which were shown to be quick and accurate. Both algorithms were available until now only as batch executables.

Results: Here we present GEVALT (GEnotype Visualization and ALgorithmic Tool), a software package designed to simplify and expedite the process of genotype analysis, by providing a common interface to several tasks relating to such analysis. GEVALT combines the strong visual abilities of Haploview with our quick and powerful algorithms for genotypes phasing (GERBIL), tag SNP selection (STAMPA) and permutation testing for evaluating significance of association. All of the above are provided in a visually appealing and interactive interface.

Conclusion: GEVALT is an integrated viewer that uses state of the art phasing and tag SNP selection algorithms. By streamlining the application of GERBIL and STAMPA together with strong visualization for assessment of the results, GEVALT makes the algorithms accessible to the broad community of researchers in genetics.

Background

Genotype information generated by individual and international efforts carries the promise of revolutionizing disease studies and the association of phenotypes with alleles and haplotypes. Given the enormous amounts of public genotype data, tools for analyzing, interpreting and visualizing these data sets are of critical importance to researchers.

In past works we have developed the following analysis algorithms:

1. **GERBIL** [1,2] – an algorithm for simultaneously phasing genotypes into haplotypes and block partitioning. The algorithm is based on a stochastic model for recombination-poor regions ("blocks"), in which haplotypes are generated from a small number of core haplotypes, allowing for mutations, rare recombinations and errors. The genotype phasing and block partitioning is solved by an expectation-maximization algorithm. Gerbil accepts genotype data as input and outputs the phased genotypes for each individual, the block structure of the entire population and the common haplotypes in each block. As part of the algorithm, Gerbil also accurately completes missing data according to the common haplotypes found. Gerbil

was shown to be quick and accurate even for many hundreds of individuals [1].

2. STAMPA [3] – an algorithm for tag SNP selection. The algorithm finds a set of tag SNPs with maximal prediction accuracy. The prediction accuracy of a set of tag SNPs is the expected accuracy of predicting untyped SNPs, given the tag SNPs. Dynamic programming is used in order to efficiently find the set of tag SNPs. Halperin et. al tested Stampa on many different genotype datasets from different sources, and showed that it finds tag SNPs with considerably better prediction ability than two other state-of-the-art tag SNP selection algorithms [3].

Both GERBIL and STAMPA were available until now only as batch executables. In this work we introduce GEVALT (GEnotype Visualization and ALgorithmic Tool). GEVALT (Version 1.1) is an integrated software providing easy access to the GERBIL and STAMPA algorithms as well as to some other tools for genotype analysis. GEVALT is based on Haploview version 3.2 [4] and it maintains the user-friendly interface and strong visualization capabilities of Haploview, as well as its other functionalities, including computation of marker quality statistics and LD information.

Implementation

GEVALT is implemented in JAVA based on the open source code of Haploview version 3.2. The analysis algorithms (GERBIL, STAMPA and permutation testing) are implemented in C++. Both Linux and Windows versions of GEVALT are available for download, as well as the JAVA source code.

Results and Discussion

GEVALT accepts input in a variety of formats. Genotype data can be loaded as unphased genotypes in the standard linkage format, or as either partially or fully phased chromosomes. Genotype data dumps from the HapMap website [5] can also be loaded. When using the standard linkage format, the user can specify family structure as well as disease status. The user can also specify marker information, including name and location. Upon loading a dataset, GEVALT first phases the genotypes in the following manner: For data consisting of unrelated individuals, GEVALT uses Gerbil to phase the genotypes. For data consisting of two-generation pedigrees, GEVALT first creates a set of trios, one per family, where each trio contains the child with least missing data. In each trio phasing is done, if possible, according to Mendelian rules. Then only the children's genotypes are phased using Gerbil and each parent's haplotypes are deduced from its child's haplotypes. Gerbil is then run again on the set of the parents' genotypes only to complete the missing data. The haplotypes of the children that were not included in the trios are deduced from their parents' haplotypes.

After phasing is completed, GEVALT generates several displays and option menus including the following:

• **Phased genotypes**: The phased genotypes of each individual are displayed. Different colors are used to indicate alleles phased by GERBIL, missing data and Mendelian errors (Figure 1). For data consisting of pedigrees, the phased genotypes are divided into two groups, parents and children, and each group is displayed separately.

• **Stampa**: Select tag SNPs using the Stampa algorithm (Figure 2). The user can specify the desired number of tag SNPs, and the algorithm finds an optimal set of tag SNPs and reports its prediction accuracy. The user can add or remove tag SNPs from the set manually and GEVALT recalculates the prediction accuracy of the new set.

• **Individual Stats**: Summary statistics for each individual are displayed (Figure 3). These include the percentage of missing genotypes, the percentage of heterozygous markers, the percentage of minor alleles and a tally of Mendelian inheritance errors.

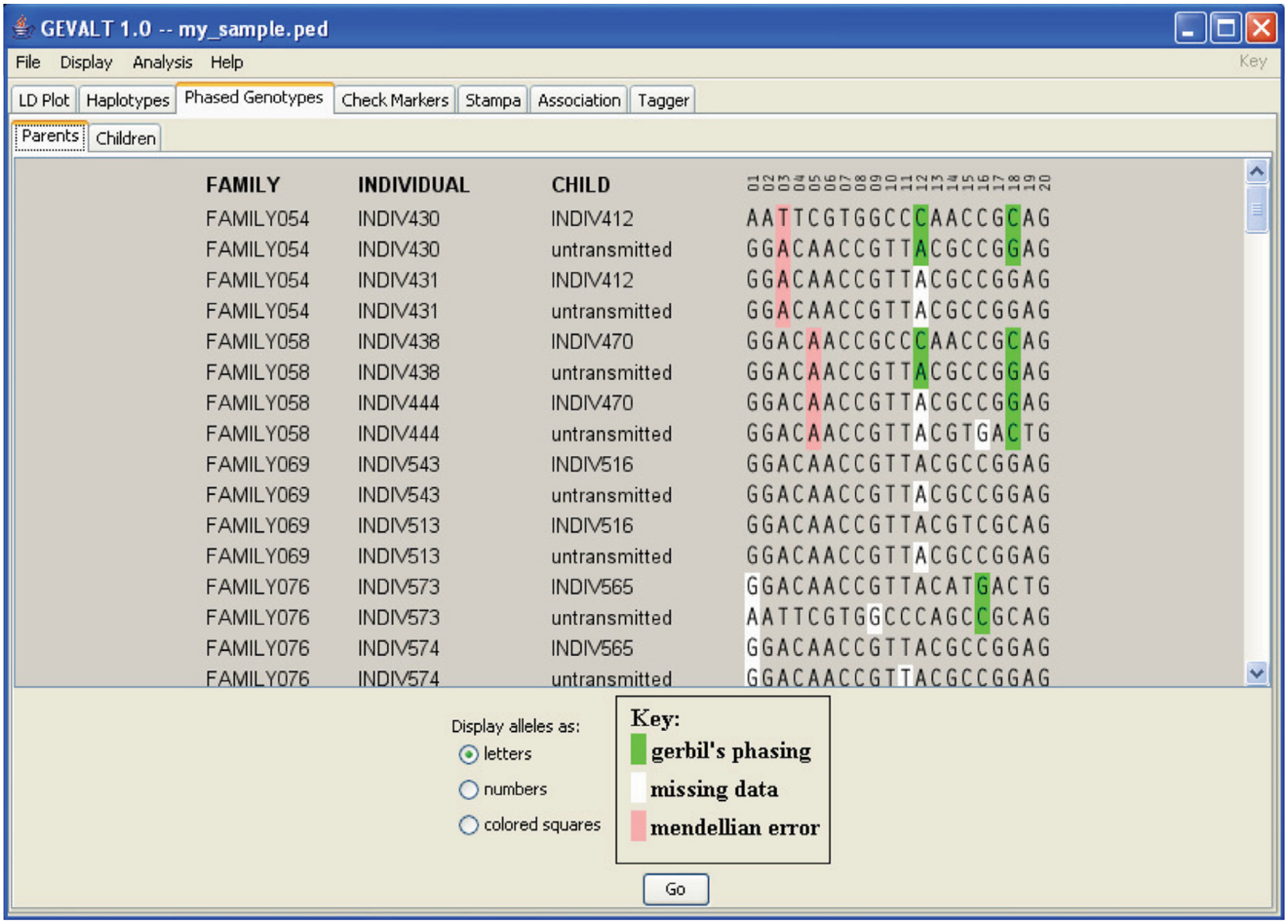
All of Haploview's displays and option menus are available. See [4] for a full description of these features. In addition, the following changes and extensions are introduced:

• **Association**: GEVALT contains a faster implementation of the permutation test in C++ instead of JAVA. The new implementation runs about 20 times faster than the JAVA implementation in Haploview. In the haplotype associations tab, a p-value is calculated for each block and for each haplotype (Figure 4).

• **LD Plot**: The LD between each pair of SNPs can be calculated using either the phased or the unphased genotypes (Figure 5). The block structure displayed by default is that found by GERBIL. The user can still employ any of Haploview's block identification methods or select blocks manually. Markers that were chosen as tag SNPs in the Stampa tab are highlighted in this display.

• **Haplotypes**: The common haplotypes in each block are computed based on the Gerbil solution.

• **Check markers**: Phasing by Gerbil is done only for the set of picked markers. Whenever the set of picked markers changes, GEVALT recalculates the phasing, and all displays are updated accordingly.



FAMILY	INDIVIDUAL	CHILD	01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20
FAMILY054	INDIV430	INDIV412	AATTCGTGGCCCAACCGCAG
FAMILY054	INDIV430	untransmitted	GGACAACCGTTACGCCGGAG
FAMILY054	INDIV431	INDIV412	GGACAACCGTTACGCCGGAG
FAMILY054	INDIV431	untransmitted	GGACAACCGTTACGCCGGAG
FAMILY058	INDIV438	INDIV470	GGACAACCGCCCAACCGCAG
FAMILY058	INDIV438	untransmitted	GGACAACCGTTACGCCGGAG
FAMILY058	INDIV444	INDIV470	GGACAACCGTTACGCCGGAG
FAMILY058	INDIV444	untransmitted	GGACAACCGTTACGTGACTG
FAMILY069	INDIV543	INDIV516	GGACAACCGTTACGCCGGAG
FAMILY069	INDIV543	untransmitted	GGACAACCGTTACGCCGGAG
FAMILY069	INDIV513	INDIV516	GGACAACCGTTACGTCGCAG
FAMILY069	INDIV513	untransmitted	GGACAACCGTTACGCCGGAG
FAMILY076	INDIV573	INDIV565	GGACAACCGTTACATGACTG
FAMILY076	INDIV573	untransmitted	AATTCGTGGCCCAGCCGCAG
FAMILY076	INDIV574	INDIV565	GGACAACCGTTACGCCGGAG
FAMILY076	INDIV574	untransmitted	GGACAACCGTTACGCCGGAG

Display alleles as:
● letters
○ numbers
○ colored squares

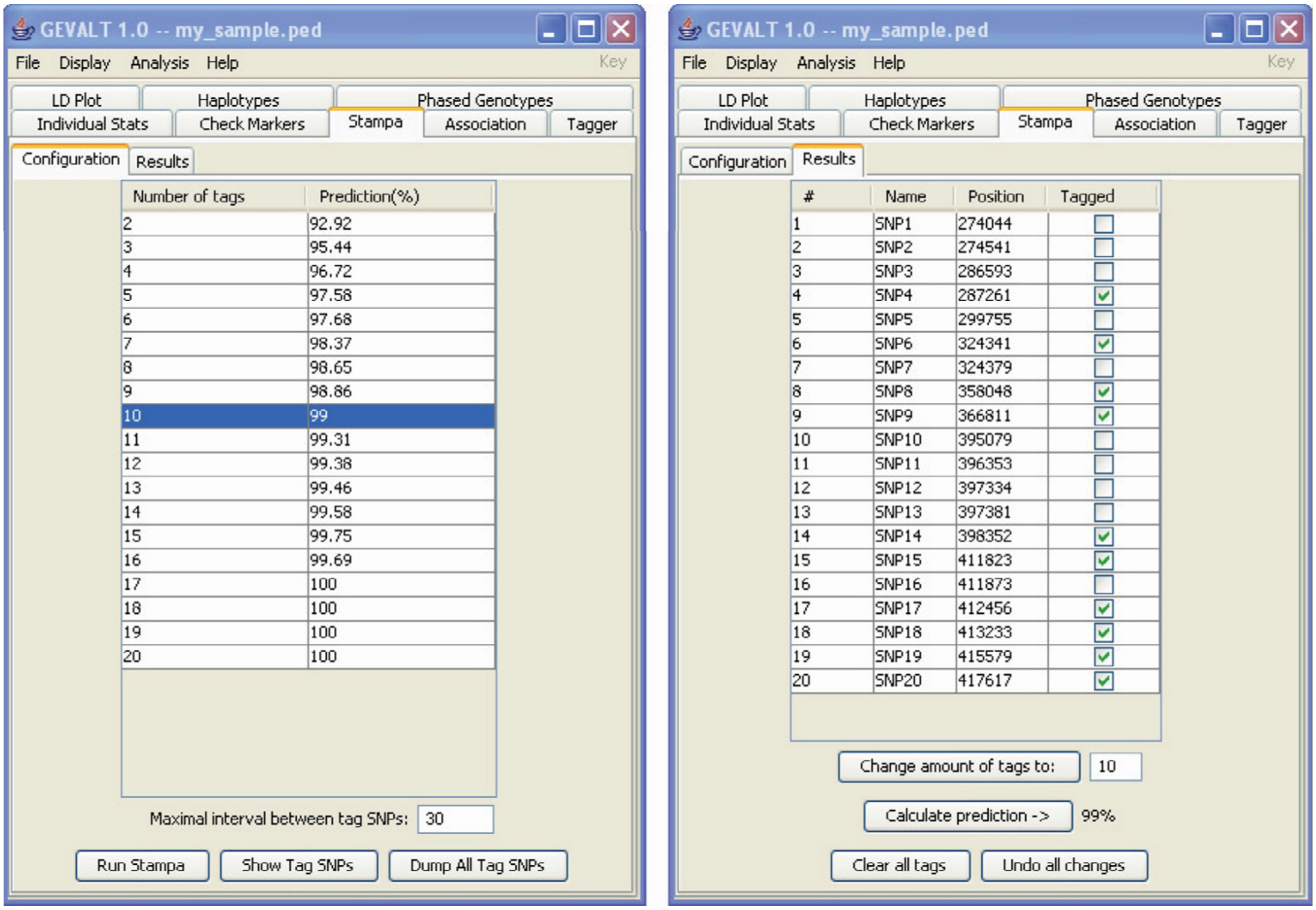
Key:
🟩 gerbil's phasing
□ missing data
🟥 mendellian error

Figure 1
The Phased Genotypes display. The phased genotypes display is divided into two tabs – parents and children. In this image the parents tab is displayed. For each parent, its two chromosomes are displayed, with an indication which of them was transmitted to the child that was included in the trio. Different colors are used to indicate alleles phased by GERBIL, missing data that were completed by GERBIL, and Mendelian errors.

Each of these displays and option menus is shown on a separate tab, allowing the user to move from one to the next. Interactive modifications made by the user in any panel are reflected in all the others. The information on each panel can also be exported to a PNG image file or to a text file. Additionally, the program has a command-line mode, which allows the user to run all the analyses without opening the GUI on one or more files at once.

The running time of GEVALT is dominated primarily by that of Gerbil and Stampa (see the references for detailed reports on the running times of these programs). All other operations, such as parameter adjustments and display changes, are done with no noticeable delay even for data sets with hundreds of markers and hundreds of individuals. Gerbil can currently handle up to 300 markers, while Stampa can handle thousands of markers. Both algorithms can handle thousands of individuals.

To the best of our knowledge, only two extant programs offer both algorithmic and visualization tools for genotype analysis: Haploview [4] and HapScope [6]. As described above, GEVALT maintains the popular, user-friendly interface of Haploview, but replaces its standard EM algorithm for phasing with the Gerbil algorithm. This allows a more accurate estimation of the phased haplotypes and a visualization of each individual's inferred haplotypes (and not just common haplotypes as in Haploview). Besides the Tagger algorithm for tag SNPs selection implemented in Haploview (and also in



Number of tags	Prediction(%)
2	92.92
3	95.44
4	96.72
5	97.58
6	97.68
7	98.37
8	98.65
9	98.86
10	99
11	99.31
12	99.38
13	99.46
14	99.58
15	99.75
16	99.69
17	100
18	100
19	100
20	100

Maximal interval between tag SNPs: 30

#	Name	Position	Tagged
1	SNP1	274044	□
2	SNP2	274541	□
3	SNP3	286593	□
4	SNP4	287261	☑
5	SNP5	299755	□
6	SNP6	324341	☑
7	SNP7	324379	□
8	SNP8	358048	☑
9	SNP9	366811	☑
10	SNP10	395079	□
11	SNP11	396353	□
12	SNP12	397334	□
13	SNP13	397381	□
14	SNP14	398352	☑
15	SNP15	411823	☑
16	SNP16	411873	□
17	SNP17	412456	☑
18	SNP18	413233	☑
19	SNP19	415579	☑
20	SNP20	417617	☑

Change amount of tags to: 10

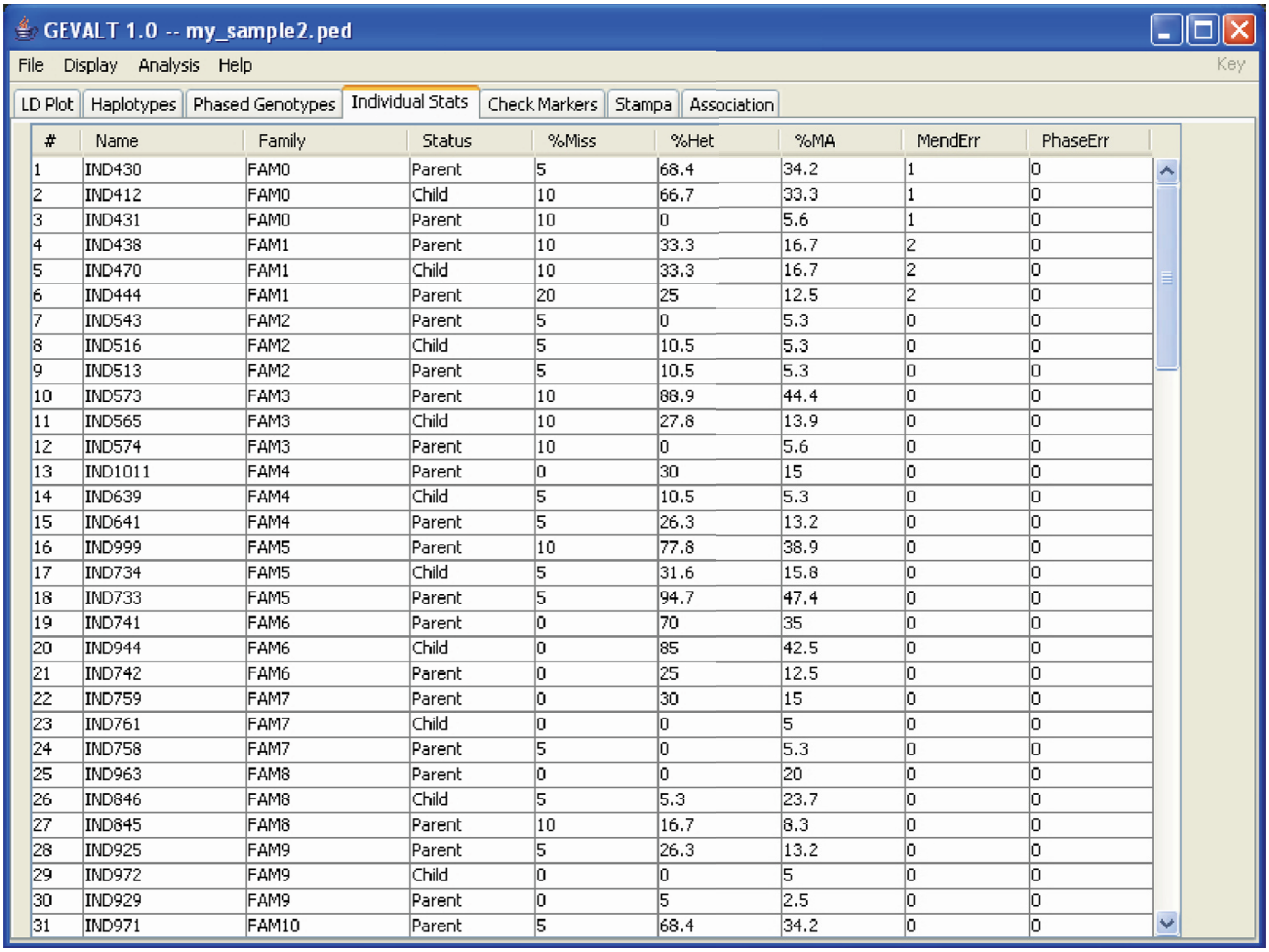
Calculate prediction -> 99%

Figure 2

The Stampa display. Left – The Stampa configuration menu. After running Stampa, a table is displayed, showing for every number of tag SNPs, the prediction accuracy of the best set of tag SNPs of that size. The user can choose the number of tags to display according to the required prediction accuracy. Right – The Stampa results menu. The selected tag SNPs are marked. The user can add or remove tag SNPs manually and recalculate the prediction accuracy of the new set.

GEVALT), GEVALT also offers STAMPA, which is not only very efficient, but also allows the user to choose the amount of tag SNPs. Other advantages and improvements over Haploview are listed above. A current limitation of Gerbil is allowing at most 300 markers. We intend to remove this limitation in the future (see below). The HapScope software includes analysis programs and a visualization tool. Most of the analyses are done separately using the command line and the results are then loaded into the visualization tool. In contrast, in GEVALT all the analyses are done within the graphical user interface, which makes it more user friendly and easy to use. HapScope uses PHASE [7] or SNPHAP [8] as its phasing algorithm. PHASE was shown to be slightly more accurate than Gerbil but much slower [1]. In contrast to HapScope, GEVALT facilitates association tests and can handle family structures. On the other hand, only HapScope includes modules for reference sequence annotation, SNP mapping and SNP classification.

We intend to continue the development of GEVALT. In particular, we shall extend Gerbil to handle more SNPs, and improve Stampa so that it incorporates into its solution predefined tag SNPs. We also intend to incorporate a new algorithm for evaluating the significance of disease association, which is dramatically faster than the standard permutation test [9].



#	Name	Family	Status	%Miss	%Het	%MA	MendErr	PhaseErr
1	IND430	FAM0	Parent	5	68.4	34.2	1	0
2	IND412	FAM0	Child	10	66.7	33.3	1	0
3	IND431	FAM0	Parent	10	0	5.6	1	0
4	IND438	FAM1	Parent	10	33.3	16.7	2	0
5	IND470	FAM1	Child	10	33.3	16.7	2	0
6	IND444	FAM1	Parent	20	25	12.5	2	0
7	IND543	FAM2	Parent	5	0	5.3	0	0
8	IND516	FAM2	Child	5	10.5	5.3	0	0
9	IND513	FAM2	Parent	5	10.5	5.3	0	0
10	IND573	FAM3	Parent	10	88.9	44.4	0	0
11	IND565	FAM3	Child	10	27.8	13.9	0	0
12	IND574	FAM3	Parent	10	0	5.6	0	0
13	IND1011	FAM4	Parent	0	30	15	0	0
14	IND639	FAM4	Child	5	10.5	5.3	0	0
15	IND641	FAM4	Parent	5	26.3	13.2	0	0
16	IND999	FAM5	Parent	10	77.8	38.9	0	0
17	IND734	FAM5	Child	5	31.6	15.8	0	0
18	IND733	FAM5	Parent	5	94.7	47.4	0	0
19	IND741	FAM6	Parent	0	70	35	0	0
20	IND944	FAM6	Child	0	85	42.5	0	0
21	IND742	FAM6	Parent	0	25	12.5	0	0
22	IND759	FAM7	Parent	0	30	15	0	0
23	IND761	FAM7	Child	0	0	5	0	0
24	IND758	FAM7	Parent	5	0	5.3	0	0
25	IND963	FAM8	Parent	0	0	20	0	0
26	IND846	FAM8	Child	5	5.3	23.7	0	0
27	IND845	FAM8	Parent	10	16.7	8.3	0	0
28	IND925	FAM9	Parent	5	26.3	13.2	0	0
29	IND972	FAM9	Child	0	0	5	0	0
30	IND929	FAM9	Parent	0	5	2.5	0	0
31	IND971	FAM10	Parent	5	68.4	34.2	0	0

Figure 3

The Individual Stats display. This table summarizes statistics for each individual. These include the percentage of missing genotypes, the percentage of heterozygous markers, the percentage of minor alleles, and a tally of Mendelian inheritance errors and of phasing errors.

Conclusion

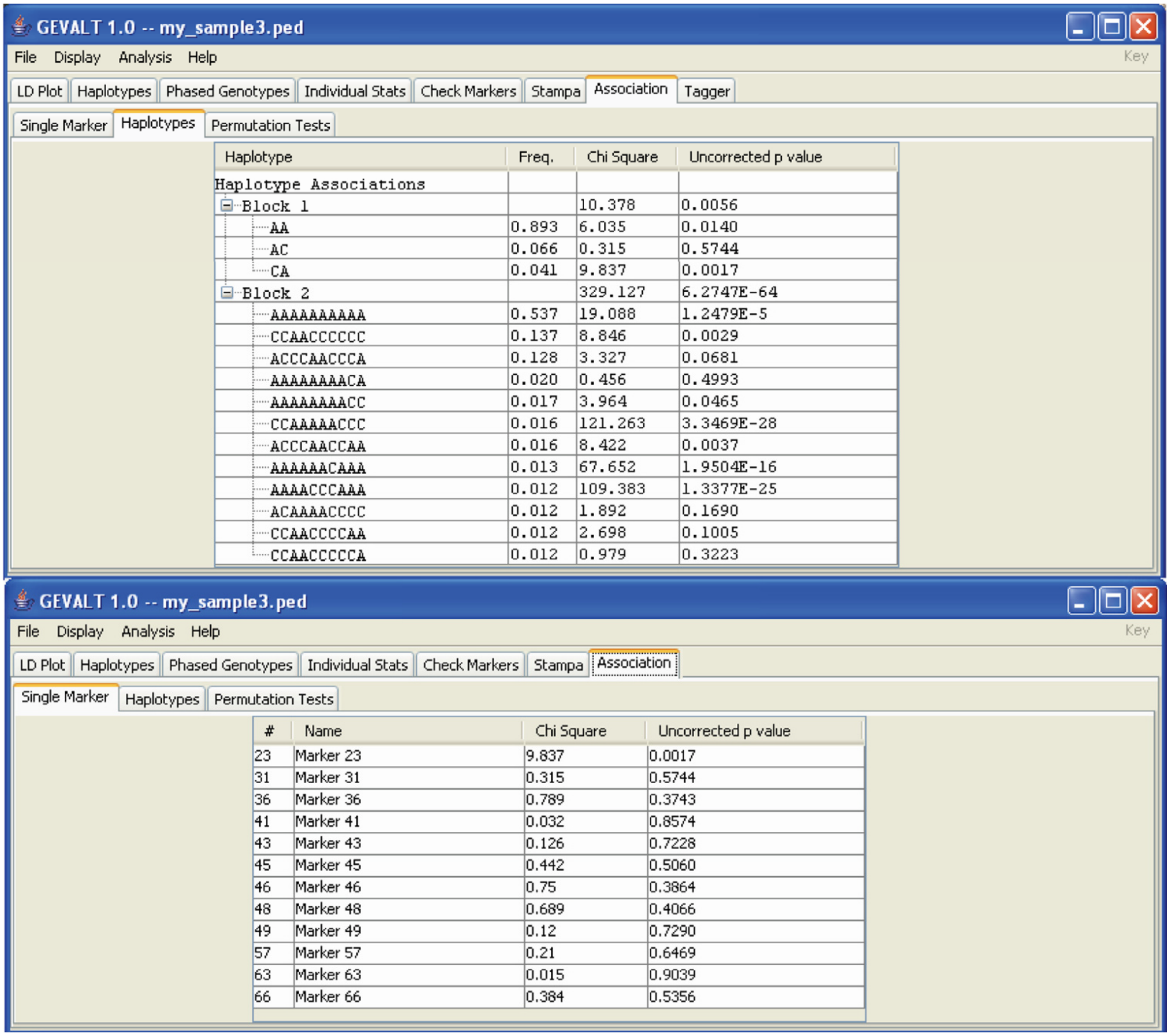
GEVALT is an integrated viewer that uses state of the art phasing and tag SNP selection algorithms. It streamlines the application of GERBIL and STAMPA, which were available until now only as batch executables, and allows using them together with the strong visualizations of Haploview for assessment of the results. Both running the algorithms and visualizing the results are done within the graphical user interface, unlike, e.g., the HapScope software [6], which only enables the latter. This makes the algorithms accessible to the broad community of researchers in genetics.

Availability and requirements

• **Project name**: GEVALT

• **Project home page**: http://www.cs.tau.ac.il/~rshamir/gevalt

• **Operating systems**: Windows and Linux.



Haplotype	Freq.	Chi Square	Uncorrected p value
Haplotype Associations			
Block 1		10.378	0.0056
AA	0.893	6.035	0.0140
AC	0.066	0.315	0.5744
CA	0.041	9.837	0.0017
Block 2		329.127	6.2747E-64
AAAAAAAAAA	0.537	19.088	1.2479E-5
CCAACCCCCC	0.137	8.846	0.0029
ACCCAACCCA	0.128	3.327	0.0681
AAAAAAAACA	0.020	0.456	0.4993
AAAAAAAACC	0.017	3.964	0.0465
CCAAAAACCC	0.016	121.263	3.3469E-28
ACCCAACCAA	0.016	8.422	0.0037
AAAAAACAAA	0.013	67.652	1.9504E-16
AAAACCCAAA	0.012	109.383	1.3377E-25
ACAAAACCCC	0.012	1.892	0.1690
CCAACCCCAA	0.012	2.698	0.1005
CCAACCCCCA	0.012	0.979	0.3223

#	Name	Chi Square	Uncorrected p value
23	Marker 23	9.837	0.0017
31	Marker 31	0.315	0.5744
36	Marker 36	0.789	0.3743
41	Marker 41	0.032	0.8574
43	Marker 43	0.126	0.7228
45	Marker 45	0.442	0.5060
46	Marker 46	0.75	0.3864
48	Marker 48	0.689	0.4066
49	Marker 49	0.12	0.7290
57	Marker 57	0.21	0.6469
63	Marker 63	0.015	0.9039
66	Marker 66	0.384	0.5356

Figure 4

The Haplotypes and Single Markers Associations displays. The Association tab contains three displays: Single Marker, Haplotypes, Permutation Tests. Top – The Haplotypes Association tab. For each block and each haplotype in a block a chi-square score is calculated (TDT test for pedigrees, case/control test for unrelated individuals) and a p-value is derived. Only haplotypes above a certain frequency threshold are considered and displayed (the threshold is set by the user in the Haplotypes tab). Bottom – The Single Marker tab for the same set of markers. In this example more significant associations are detected when testing for haplotype associations than when testing individual SNPs.

- **Programming language**: Java and C++

- **Other requirements**: Java 1.3 or higher.

- **License**: free non-commercial research use license.

- **Any restrictions to use by non-academics**: license needed for commercial use.

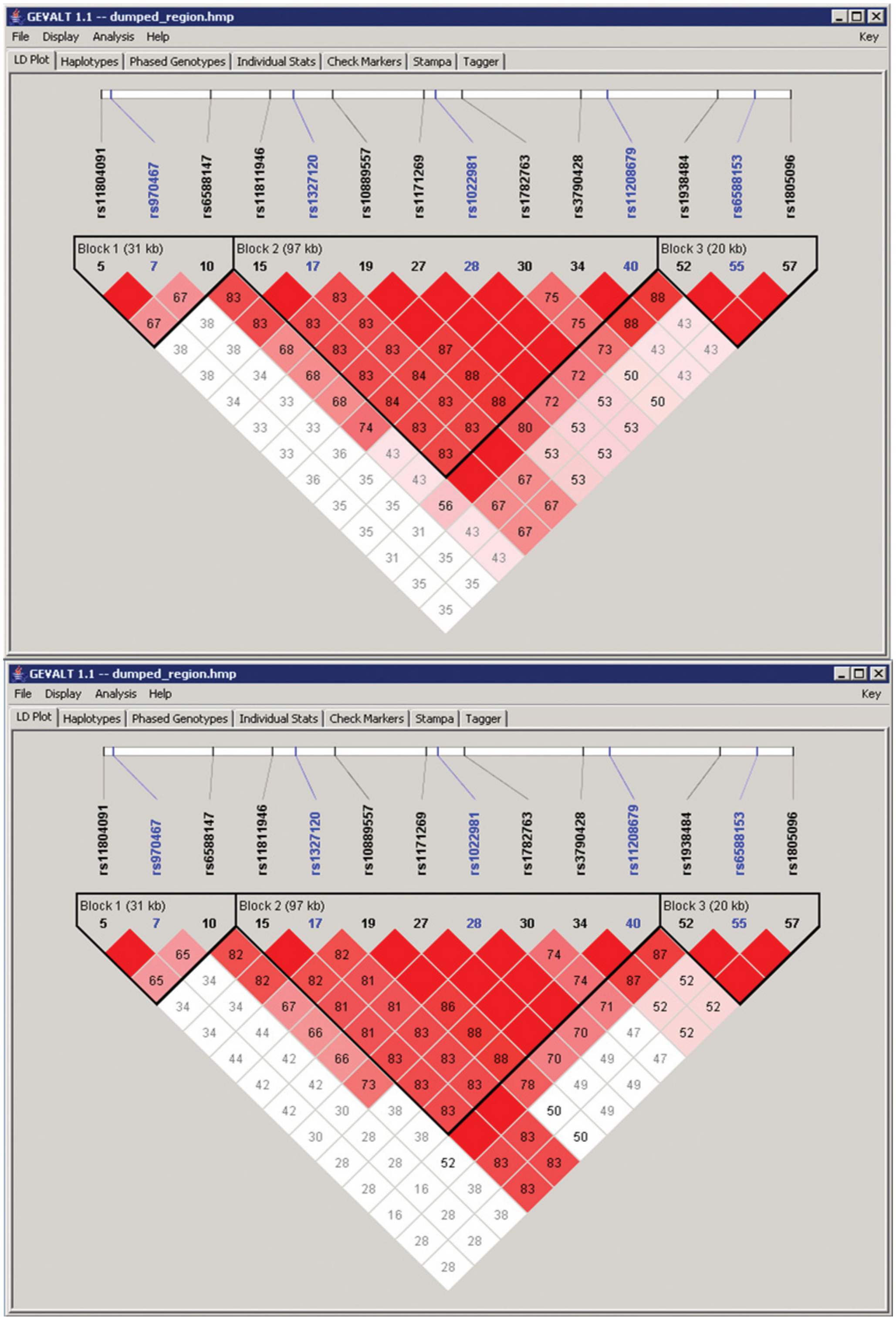

GEVALT 1.1 -- dumped_region.hmp
File Display Analysis Help
Key
LD Plot | Haplotypes | Phased Genotypes | Individual Stats | Check Markers | Stampa | Tagger
rs11804091 rs970467 rs6588147 rs11811946 rs1327120 rs10889557 rs1171269 rs1022981 rs1782763 rs3790428 rs11208679 rs1938484 rs6588153 rs1805096
Block 1 (31 kb) 5 7 10
Block 2 (97 kb) 15 17 19 27 28 30 34 40
Block 3 (20 kb) 52 55 57
67 83 83 75 88
67 38 83 83 83 75 88 43
38 38 68 83 83 87 73 43 43
38 34 68 83 84 88 72 50 43
34 33 68 84 83 88 72 53 50
33 33 74 83 83 80 53 53
33 36 43 83 53 53
36 35 43 67 53
35 35 56 67 67
35 31 43 67
31 35 43
35 35
35
GEVALT 1.1 -- dumped_region.hmp
File Display Analysis Help
Key
LD Plot | Haplotypes | Phased Genotypes | Individual Stats | Check Markers | Stampa | Tagger
rs11804091 rs970467 rs6588147 rs11811946 rs1327120 rs10889557 rs1171269 rs1022981 rs1782763 rs3790428 rs11208679 rs1938484 rs6588153 rs1805096
Block 1 (31 kb) 5 7 10
Block 2 (97 kb) 15 17 19 27 28 30 34 40
Block 3 (20 kb) 52 55 57
65 82 82 74 87
65 34 82 82 81 74 87 52
34 34 67 81 81 86 71 52 52
34 44 66 81 83 88 70 47 52
44 42 66 83 83 88 70 49 47
42 42 73 83 83 78 49 49
42 30 38 83 50 49
30 28 38 83 50
28 28 52 83 83
28 16 38 83
16 28 38
28 28
28

Figure 5

LD plots comparison. Top – An LD plot calculated using the phased genotypes (default). Bottom – An LD plot calculated using the unphased genotypes. Different LD scores are observed in many SNP pairs. These differences result from unambiguous phasing of double heterozygotes and from completed missing data. Markers that were chosen as tag SNPs in the Stampa tab are highlighted in blue.

Authors' contributions

OD and GK contributed to the design of GEVALT. OD implemented GEVALT and GK implemented the analysis algorithms. RS supervised the project. All authors participated in the drafting and revising of the manuscript, and read and approved the final manuscript.

Acknowledgements

This work is supported in part by a grant from the German-Israel Foundation (grant 237/2005). We are grateful to the Haploview Team for making their software and source code available to us, and to Eran Halperin who codeveloped Stampa.

References

1. Kimmel G, Shamir R: **GERBIL: Genotype Resolution and Block Identification Using Likelihood.** *Proc Natl Acad Sci U S A* 2005, **102:**158-162.
2. Kimmel G, Shamir R: **Maximum likelihood resolution of multi-block genotypes.** In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 04)* The Association for Computing Machinery; 2004:2-9.
3. Halperin E, Kimmel G, Shamir R: **Tag SNP Selection in Genotype Data for Maximizing SNP Prediction Accuracy.** *Bioinformatics* 2005, **21(Suppl 1):**i195-i203.
4. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21(2):**263-265.
5. **International HapMap Project** [http://www.hapmap.org]
6. Zhang J, Rowe WL, Struewing JP, Buetow KH: **HapScope: a software system for automated and visual analysis of functionally annotated haplotypes.** *Nucl Acids Res* 2002, **30(23):**5213-5221.
7. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68(4):**978-989.
8. **SNPHAP software** [http://www-gene.cimr.cam.ac.uk/clayton/software/]
9. Kimmel G, Shamir R: **A Fast Method for Computing High Significance Disease Association in Large Population-Based Studies.** *Am J Hum Genet* 2006, **79:**481-492.