Tel Aviv University
Raymond & Beverly Sackler
Faculty of Exact Sciences
School of Computer Sciences

Methods and tools in computational human genetics

Thesis submitted as part of the requirements
for the degree of Master of Science

By
Ofir Davidovich

Under the supervision of Prof. **Ron Shamir**

September, 2007

Acknowledgments

First and foremost, I would like to thank my advisor, Prof. Ron Shamir, for walking me through the world of science. By patiently guiding me and setting an example he succeeded in turning me from a student to a researcher. Throughout this thesis I tried to adopt (and hopefully succeeded) his values of scientific integrity and thoroughness.

A great thanks goes to Dr. Gad Kimmel who was practically my co-advisor. Gadi had the patience to introduce me to the world of human genetics and answer my questions even from the far away Berkeley in awkward hours.

I want to thank my parents and two sisters for giving me the best support a family can give. Especially I thank my mother, Dr. Liema Davidovich, for inspiring me to choose the academic path. I also want to thank my girlfriend, Rachel, for giving me the support I needed on rainy days, and for cleverly putting things in perspective.

I want to thank all my lab mates: Yonit Halperin, Igor Ulitsky, Daniela Raijman, Michael Gutkin, Oded Apel, Chaim Linhart, Adi Maron-Katz, Israel Steinfeld, Irit Gat-Viks, Michal Ozery-Flato, Michal Ziv-Ukelson, Rani Elkon, Seagull Shavit and Ofer Lavi. Thanks for the fruitful conversations, advices, and especially for the laughs, arguments, and the great atmosphere in the lab.

Last, but not least, I would like to thank my collaborators: Dr. Eran Halperin (International Computer Science Institute, Berkeley), Dr. Arie Levine and Dr. Esther Leshinsky-Silver (Wolfson Medical Center), Dr. Amir Karban (Rambam Medical Center), and Prof. Margret Hoehe (Max Plank Institute).

Abstract

The availability of the human genome sequence has revolutionized human genetics research. By studying the differences in the genomes between sick and healthy individuals one can associate particular differences with the disease. Such association is the first step towards understanding disease cause, diagnostics and eventually therapy. In this thesis we study several problems related to the most common type of sequence differences, single nucleotide polymorphisms (SNPs).

We first explore tag SNP selection criteria. We compare two algorithms that use different criteria for selecting tag SNPs: the r^2 criterion and the *prediction accuracy* criterion. By performing extensive simulations with real haplotypes we show that tags selected according to the prediction accuracy criterion provide higher power to association studies. We show that the magnitude of the advantage in power is dependent on the tag density. When choosing tags at high density, both methods provide comparable and very high power, but as the tag density decreases, the advantage of the prediction accuracy criterion grows larger.

We next describe a software package for genotype analysis that we developed. The software package combines state-of-the-art algorithms for genotype phasing, tag SNP selection, and association testing, with convenient visualizations. By streamlining the application of the algorithms, which were only available as batch executables previously, we make the algorithms accessible to the broad community of researchers in genetics.

Lastly, we describe three studies on Crohn's Disease performed in collaboration with gastroenterologists from Wolfson and Rambam medical centers. In these studies we applied common methods for statistical analysis and developed necessary ad-hoc adjustments, in order to identify genotype-phenotype and phenotype-phenotype associations. The studies and methods used are briefly summarized here while full information is given in articles published in the literature.

Contents

Chapter 1

Introduction and Summary

The availability of the human genome sequence has revolutionized human genetics research. By studying the differences in the genomes between sick and healthy individuals one can associate particular differences with the disease. Such association is the first step towards understanding disease cause, diagnostics and eventually therapy. In this thesis we study several problems related to the most common type of sequence differences, single nucleotide polymorphisms (SNPs).

Chapter 2 provides biological background and discusses several key challenges in human genetics. It also describes three of the most common tasks in human genetics research: genotype phasing, tag SNP selection, and association studies.

In **Chapter 3** we explore tag SNP selection criteria. We compare two algorithms that use different criteria for selecting tag SNPs: Tagger [12], a very popular and widely used program, which uses the r^2 criterion, and STAMPA [21] which uses the *prediction accuracy* criterion. We ask the following question: given the two tag SNP selection algorithms, which one would give the highest power to association studies that use tag SNPs selected by it? In order to answer the question we perform extensive simulations with real haplotypes. We choose tag SNPs with both methods, simulate case/control panels, and perform association tests to evaluate the power of each method. Several scenarios and tag densities were tested. We show that STAMPA attains significantly and consistently higher power than Tagger. We also show that the strong advantage in the power is explained primarily by the selection criterion of STAMPA, prediction accuracy. When choosing tags at high density, both methods provide comparable and very

high power, but as the tag density decreases, STAMPA's advantage grows larger. This study, which was done jointly with Dr. Gad Kimmel and Dr. Eran Halperin, was submitted for publication and is currently under review.

In **Chapter 4** we describe a software package for genotype analysis that we developed. The software package, called GEVALT, combines state-of-the-art algorithms for genotype phasing (GERBIL [27]), tag SNP selection (STAMPA [21]), and association testing (RAT [28]), with convenient visualizations. By streamlining the application of GERBIL, STAMPA and RAT together with strong visualization for assessment of the results, GEVALT makes the algorithms accessible to the broad community of researchers in genetics. Most of these results were published in *BMC Bioinformatics* [10].

Lastly, **Chapter 5** describes three studies on Crohn's Disease (CD) performed in collaboration with gastroenterologists from Wolfson and Rambam medical centers. In the first study, published in *Inflammatory Bowel Diseases* [32], we searched for genotype–phenotype and phenotype–phenotype correlations in a large pediatric cohort consisting of patients from Israel, USA, and Italy. We found a correlation between age of onset of the disease and the disease location. Interestingly, this correlation was also dependent on genotype. In the second study, published in *The American Journal of Gastroenterology* [25], we studied perianal Crohn's disease (PD) which is a frequent complication of CD. We found two factors to be associated with PD: rectal inflammation and ethnicity (Sephardic origin). All SNPs tested were not associated with PD. The third study, published in *International Journal of Colorectal Disease* [31], searched for an association between a SNP in the NFKBIA gene and CD. This SNP was previously found to be associated with CD in a German cohort. We did not find the SNP to be associated with CD in our Israeli cohort. Neither associations with other phenotypes nor interactions with other SNPs were found.

Chapter 2

Background

This chapter provides background on common terms and tasks in human genetics. The chapter is organized as follows. Section 2.1 defines biological terms that are in frequent use throughout this thesis. Section 2.2 discusses a key aspect in the organization of the genome, known as linkage disequilibrium. The last three sections describe three of the most common tasks in human genetics research: genotype phasing, tag SNP selection, and association studies.

2.1 Biological Background

DNA (Deoxyribonucleic acid) is a nucleic acid molecule that contains the genetic instructions used in the development and functioning of all known living organisms. Chemically, DNA is a long polymer of four simple units called *nucleotides* or *bases*. The four bases are adenine, cytosine, guanine, and thymine, abbreviated as A, C, G, and T, respectively. The DNA is organized in separate linear molecules called *chromosomes*. The human genome, for example, contains approximately three billion bases divided into 23 chromosomes. Most sexually reproducing organisms are *diploid*, i.e., have a duplicate set of genetic material consisting of paired chromosomes, one from each parent. Such paired chromosomes, called *homologous*, are essentially identical, having only small differences originating from the variability present in the population.

One main source of variation comes from *Single Nucleotide Polymorphisms (SNPs)* (see [61] for a review). A SNP is a single nucleotide position in the genome that differs across members of a species. The different possible bases in

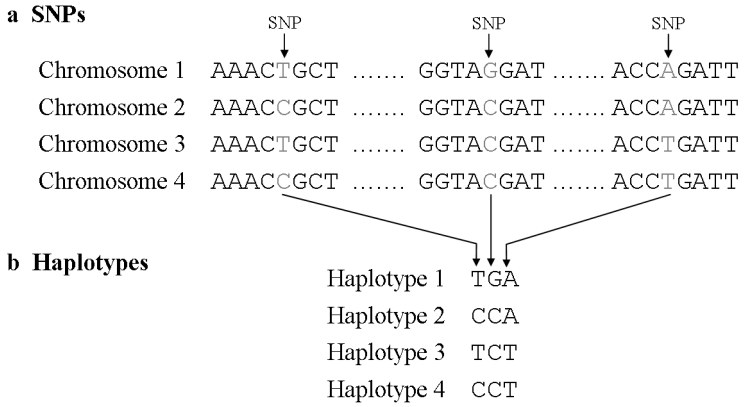


Figure 2.1: SNPs and haplotypes. (a) Four chromosomes from the population are shown. The chromosomes are identical in all positions except three positions indicated by an arrow. These positions are called *Single Nucleotide Polymorphisms (SNPs)*. Each SNP has two *alleles*. For example, the leftmost SNP has alleles T and C. (b) The four chromosomes are shown as *haplotypes*. A haplotype is the sequence of alleles in contiguous SNP positions along a chromosomal region.

a site are called *alleles* (see **Figure 2.1a**). Almost all SNPs have only two alleles (out of four possible). In the human genome, around five millions of SNPs have been detected [47, 63], out of an estimated total of ten million common SNPs.

As mentioned above, homologous chromosomes of an individual are almost identical with some small differences. These differences occur mainly at SNPs, although other types of polymorphisms occur (e.g. microsatellites [46], restriction fragment length polymorphisms [6], rearrangements [35], and copy number variations [50]). An individual is said to be *homozygous* for a SNP if the same allele is present at both homologous chromosomes, and *heterozygous* otherwise. The sequence of alleles in contiguous SNP positions along a chromosomal region is called a *haplotype* (see **Figure 2.1b**). Conflating the two haplotypes of an individual (from both homologous chromosomes) creates the individual's *genotype* (see Section 2.3). A major source for variation of haplotypes in the population comes from *recombination*. Generally, recombination is the process by which a strand of DNA is broken and then joined to the end of a different DNA molecule. In this context recombination occurs as *chromosomal crossover* between paired homologous chromosomes (see **Figure 2.2**). This process occurs naturally in meiosis, and leads to "mixing" of the two chromosomes in each parent, before

one of them is passed to the offspring.

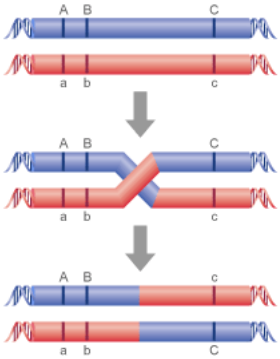


Figure 2.2: Chromosomal crossover. A short region of two homologous chromosomes is shown. This individual is heterozygous for three SNPs with alleles denoted by a,A,b,B,c,C. Its two haplotypes are: ABC and abc. After the crossover event two new haplotypes are created: ABc and abC.

2.2 Linkage Disequilibrium

Linkage Disequilibrium (LD) is a term used for the non-random association of alleles at two or more loci in the population. In this section we will focus on LD between two SNPs. Denote the two alleles of the first SNP as A and a, and the two alleles of the second SNP as B and b. If the two SNPs are independent, we will expect the frequency of the AB haplotype in the population to be approximately the product of the frequencies of alleles A and B. If this is true for all four haplotypes we say that the SNPs are in *linkage equilibrium*. Deviations from the expected frequencies of haplotypes suggest that the SNPs are in *linkage disequilibrium*. The larger the deviation from these frequencies, the larger the LD between the SNPs. A wide variety of statistics have been proposed to measure the amount of LD between a pair of SNPs [13, 4]. The most widely used are D, D' and r^2, described below:

$$D = p(AB) - p(A)p(B),$$

where $p(\cdot)$ is the frequency of the allele/haplotype. Note that $|D|$ is independent of the choice of the major and minor alleles in A and B.

$$D' = D/D_{max},$$

where

$$D_{max} = \begin{cases} \min\{p(A)p(b), p(a)p(B)\} & D \geq 0 \\ \max\{-p(A)p(B), -p(a)p(b)\} & D < 0 \end{cases},$$

and

$$r^2 = \frac{D^2}{p(A)p(a)p(B)p(b)}.$$

Both D' and r^2 take values in the range $[0, 1]$, where 0 corresponds to linkage equilibrium. $D' = 1$ can be described as "complete LD" because the correlation between the SNPs is as strong as possible ($D = D_{max}$), given the allele frequencies of the two SNPs. r^2 equals 1 only when the two SNPs are completely identical, and therefore described as "perfect LD".

Another common statistic is the *logarithm of odds (LOD) score*. This statistic compares the likelihood of the data under two hypotheses: H_0 – linkage equilibrium, and H_1 – linkage disequilibrium. The likelihood of the data is calculated assuming a multinomial distribution as follows:

$$L = \frac{(n_{AB} + n_{Ab} + n_{aB} + n_{ab})!}{n_{AB}!n_{Ab}!n_{aB}!n_{ab}!} p(AB)^{n_{AB}} p(Ab)^{n_{Ab}} p(aB)^{n_{aB}} p(ab)^{n_{ab}}$$

where n_{ij} is the observed number of occurrences of the ij haplotype. The LOD score is $log_{10}(L_1/L_0)$, where L_1 is the likelihood of the data under H_1, and L_0 is the likelihood under H_0. L_1 is calculated by plugging in the observed haplotype frequencies into the equation above. L_0 is calculated by plugging in the expected haplotype frequencies under linkage equilibrium (i.e. $p(AB) = p(A)p(B)$ etc.).

All LD statistics described above require knowing the haplotype frequencies in the population. It is straightforward to calculate these frequencies when haplotypes are given but in most cases only genotypes are known. The haplotypes of an individual that is heterozygous in both SNPs are unknown since he can have two possible pairs of haplotypes: AB and ab, or Ab and aB. So in the case where only genotypes are given the maximum-likelihood estimators of the haplotype frequencies can be estimated by a simple Expectation-Maximization (EM) procedure [15]. EM iteratively estimates the haplotype frequencies in the population and assigns the most probable haplotypes to double heterozygous individuals.

LD tends to decay with physical distance between SNPs, so high values of LD are expected mainly between SNPs located closely on the same chromosome. This can be explained by the fact that the main cause for LD between two SNPs is the rate of recombination between the two sites which increases with physical distance. Other causes for LD are random drift, non-random mating, and interactions between loci.

A major recent discovery is that the recombination rate along chromosomes is not fixed, but tends to vary widely [17, 40]. It seems that recombination is very frequent in some narrow regions called *hotspots*, with very little recombination occurring between these hotspots. The regions between hotspots are called *blocks of LD* since the SNPs inside those regions tend to be in high LD (see **Figure 2.3**).

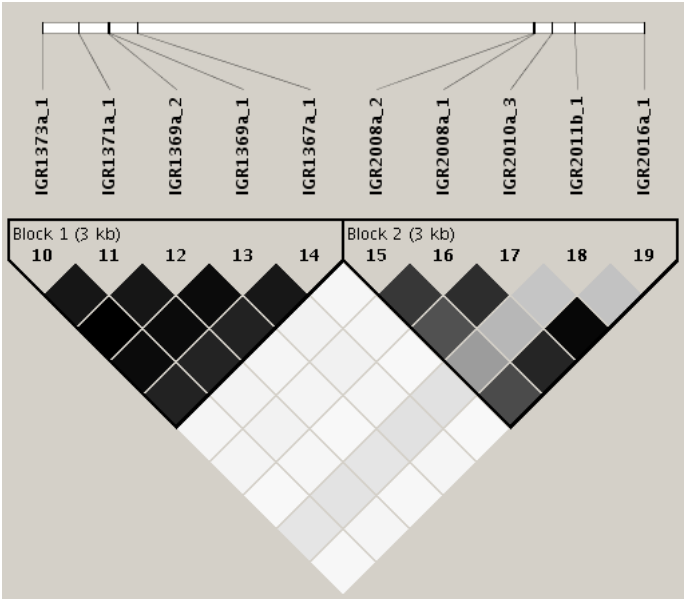


Figure 2.3: Blocks of LD. Pairwise LD measures between ten SNPs are shown. The pairwise LD is color coded from white ($r^2 = 0$) to black ($r^2 = 1$). Each square is color coded with the LD between the two SNPs on the two diagonals it belongs to. Two blocks are identified separating the first five SNPs from the last five. It can be seen that the amount of LD between SNPs from the same block is much higher than between SNPs from separate blocks. The figure was created with the GEVALT software [10].

2.3 Phasing

Most current genotyping techniques produce the genotypes, and not haplotypes, of the tested individuals. Since haplotypes are more informative than genotypes (e.g. for LD calculations as described in the previous section) it is desirable to estimate the haplotypes. The task of estimating the individuals' haplotypes from their observed genotypes is known as *phasing*. Formally, the input to the problem is the genotypes of a set of unrelated individuals. The goal is to find the most probable pair of haplotypes for each individual (see **Figure 2.4** for an example).

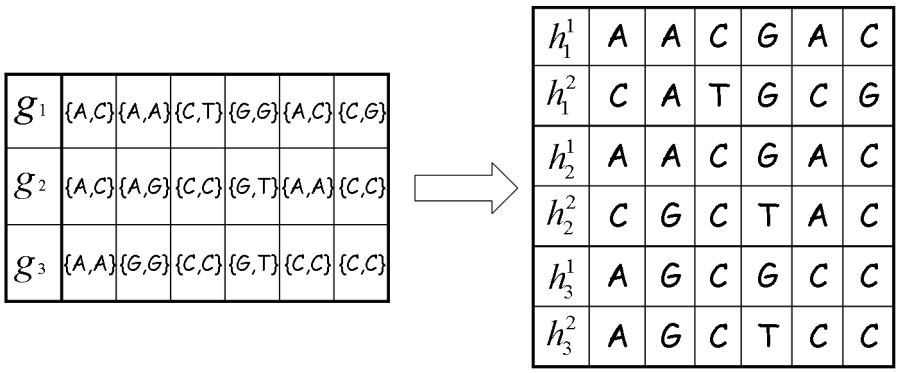


Figure 2.4: Three individuals' genotypes (g_1, g_2, g_3) comprising of 6 SNPs are shown (left) as unordered pairs of alleles. After phasing (right), every genotype, g_i, is split into two haplotypes (h_i^1, h_i^2).

A genotype with i heterozygous sites can be phased into 2^{i-1} different pairs of haplotypes. Hence, the solution space grows exponentially with i. Methods for genotype phasing try to phase all individuals simultaneously and make use of the LD structure between the SNPs in order to achieve accurate results. The first methods for phasing assumed that the input data consisted of one block with no recombination events inside it. It suggests that before applying these phasing algorithms, blocks should be identified and then each block should be phased separately. These algorithms include Clark's parsimony-based algorithm [9], Likelihood-based Expectation - Maximization (EM) algorithms [15, 34], MCMC-based methods [55, 37], and methods based on the perfect phylogeny model [14, 2].

More recent algorithms can handle recombination events. Greenspan and

Geiger's HaploBlock algorithm [20] performs phasing while taking into account the block structure. The method is based on a Bayesian network model. A very accurate and widely used software is PHASE [54]. It is based on the coalescent model [29] and performs phasing while simultaneously estimating the fine-scale recombination rate between the SNPs. Recently, a faster and more flexible model of PHASE, called fastPHASE [49] was suggested. Kimmel and Shamir [26, 27] suggested a model-based approach for simultaneous phasing and block partitioning. They implemented their method in a software package called GERBIL.

2.4 Tag SNPs

The cost of a genetic study is directly influenced by the number of SNPs genotyped. Considering the vast amount of SNPs in the human genome (estimated at ten million) it is impractical to genotype all SNPs. Also, the high correlation between nearby SNPs suggests that only a subset of all SNPs needs to be typed. An important challenge in every genetic study is to decide which SNPs to type. The main idea is to choose SNPs that best "represent" or "tag" all other SNPs. These SNPs are therefore called *tag SNPs*.

One problem that faces designers of algorithms for tag SNP selection is that the goal of the algorithm is not clearly defined. What is the meaning of a subset of SNPs that best "represent" the other SNPs? Should the emphasis be on minimizing the number of tag SNPs needed, on the quality of the "representation", or some tradeoff between the two?

The main scheme of all tag SNP selection algorithms is as follows. A training set of genotypes or haplotypes is accepted as input. This training set usually contains all known SNPs in a small set of individuals (e.g. individuals from the HapMap resource [59]). The algorithm selects tag SNPs that best "represent" all the SNPs in the training set. Eventually, only the tag SNPs are typed in a large cohort of interest (e.g. patients having some disease and matching healthy controls), and the typed SNPs are used for subsequent analysis (e.g. test of association, haplotypes inference, etc.).

Many of the tag SNP selection algorithms adopt the following method: partition the dataset into haplotype blocks and select, within each block, a subset of "haplotype tagging" SNPs sufficient to reconstruct the diversity within that block. The algorithms differ in their method for defining blocks and in the criteria

for tag selection within a block. Avi-Itzhak et al. [1] developed a simple numerical algorithm that selects the minimal subset of SNPs required to capture the diversity of a haplotype block. Patil et al. [40] defined a consecutive set of SNPs as a block if the common haplotypes (haplotypes with frequency $> \beta$) account for at least α percent of all the observed haplotypes. They select tag SNPs in a block as to minimize the number of SNPs that can distinguish at least α percent of all the observed haplotypes. A greedy algorithm is used in order to partition the entire data into blocks such that the total number of tags is minimized. Zhang et al. [69] replaced the greedy algorithm with a dynamic programming algorithm. In a follow up study [70] they combined the dynamic programming algorithm with the partition-ligation-expectation-maximization (PL-EM) algorithm for haplotype inference [45]. This allows their algorithm, HapBlock, to accept genotype data as input. There are two drawbacks to such methods. First, it is not always clear how to define blocks, and different definitions might result in different block partitioning, which consequentially affects the tag SNP selection. Second, correlation between blocks (which is known to exist) is ignored.

Many "block free" methods, which select tag SNPs without partitioning the data into blocks, have also been developed. Bafna et al. [3] described a new measure of *informativeness* of a SNP, that provides a direct measure of how a SNP, or a set of SNPs, can be used to characterize another SNP or a set of SNPs. They showed that finding a minimal set of informative SNPs is NP-hard and provided an algorithm that solves a special case of the problem where the predictive SNPs are in close proximity to their target. Carlson et al. [7] suggested a greedy algorithm based on LD between markers. Their algorithm selects tags that capture all SNPs with an r^2 above some threshold. de Bakker et al. [12] further extended the algorithm and implemented it in a program called Tagger. Tagger is described in detail in Section 2.4.1. Halperin et al. [21] proposed a novel measure for tag selection, called *prediction accuracy*, and developed a dynamic programming algorithm to select tags with maximal prediction accuracy. Their algorithm, STAMPA, is described in detail in Section 2.4.2. He and Zelikovsky [23] used multiple linear regression in order to predict the non-tag SNPs based on all tag SNPs. They suggested a greedy algorithm that selects tag SNPs with maximal prediction accuracy. Eyheramendy et al. [16] proposed a method for predicting the non-tags based on the tags which uses the Li and Stephens model for haplotype data [33]. They developed a method to select tags with

maximum prediction.

2.4.1 Tagger

Tagger, developed by de Bakker et al. [12], is based on the r^2 criterion (see section 2.2). It uses a greedy algorithm to select tags that capture all other SNPs with r^2 exceeding a prescribed threshold (actually, this is an implementation of the ldSelect [7] algorithm of Carlson et al.). In addition, to improve efficiency, it can carry out an aggressive search attempting to replace each tag with a specific multi-marker predictor (on the basis of the remaining tags). We will first describe the standard mode (called pairwise mode) and then the aggressive mode.

The input of Tagger is a set of haplotypes or genotypes as a training set, and a threshold θ. First, the pairwise r^2 between every pair of SNPs is calculated. In case the input contains genotypes, a simple two marker EM is used in order to estimate the maximum-likelihood values of the four gamete (AB, Ab, aB, ab) frequencies needed for the calculation of r^2. We say that SNP i *captures* SNP j if the r^2 between them exceeds θ. Then, a single SNP capturing a maximum number of other SNPs is identified. This SNP and its associated SNPs are grouped as a bin. Within the bin, every SNP capturing all SNPs in the bin is specified as a "potential tag SNP for the bin". Not all SNPs within the bin will be selected because pairwise association is not a transitive property: if $r^2 > \theta$ for SNP pairs A/B and B/C, $r^2 < \theta$ for SNP pair A/C is possible. Only one of the potential tag SNPs in the bin will eventually be reported as a tag SNP. It can be selected randomly or by some other properties such as coding vs. non-coding or ease of assay design. The binning process is iterated, analyzing all unbinned SNPs at each round, until all SNPs are binned. Finally, the selected tag SNPs are reported. Also, for every tag SNP, a list of SNPs that reside in the same bin are reported (these SNPs are said to be *tagged* by this tag SNP, and the tag SNP is called the *proxy* of these SNPs). Note that the SNPs in a bin need not be consecutive.

Tagger's aggressive mode is an extension of the pairwise mode. After completing the steps described above, Tagger attempts to replace each tag with a specific multimarker predictor. Specifically, for each tag Tagger searches for a specific combination of tag SNPs (i.e., a haplotype of tag SNPs) that can capture the tag and all the SNPs captured by that tag. If such a combination is found

then the tag is discarded from the list of tags, and instead, the combination of tags (termed *test*) is added to the list. See **Figure 2.5** for an example. The maximal number of tags comprising such test is user defined (tests may consist of up to six SNPs). To minimize risk of overfitting, tags in a specified test are forced to be in strong LD (defined as LOD score greater than some user defined threshold) with one another and with the discarded tag.

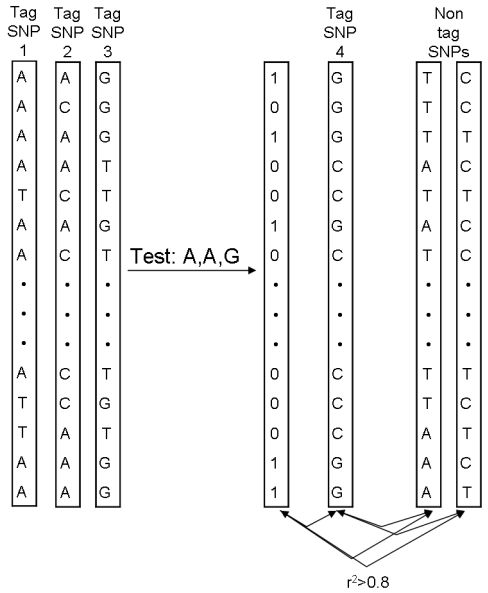


Figure 2.5: An example of the operation of Tagger's aggressive mode. Tag SNPs 1-3 create a multimarker predictor that captures tag SNP 4. This predictor (test) is the haplotype AAG. More specifically, a new vector is created by assigning "1" to the AAG haplotype of the combination of tag SNPs 1-3, and assigning "0" to all other haplotypes. This vector captures tag SNP 4 and all non-tag SNPs that were captured by it, and therefore can replace tag SNP 4.

Tagger can also select a user predefined number of tag SNPs (called the 'best N' method). In this method Tagger prioritizes the tags by the number of SNPs for which they can serve as a proxy. The first N tag SNPs are reported.

2.4.2 STAMPA

The main idea behind STAMPA [21] is to view tag SNPs as predictors of the non-tag SNPs. After typing the tag SNPs in the study cohort, the geneticist can

use the tags to predict the values of the non-tag SNPs and then use all SNPs (typed and predicted) for the analysis. Hence, STAMPA selects tags according to their *prediction accuracy*. Before describing the STAMPA algorithm we will briefly present a formal definition of the tag SNP selection problem, needed for the presentation of the algorithm.

We assume that each haplotype is represented by a binary string. Thus, a haplotype of length m is a sequence over $\{0,1\}^m$. A genotype of length m is represented by a $\{0,1,2\}$ sequence, where 0 and 2 stand for the homozygous types {0,0} and {1,1}, respectively, and 1 stands for a heterozygous type. We are given a training set of n genotypes $g_1, \ldots, g_n$ of length m each. We use $g_{i,j}$ to denote the j-th component (0,1, or 2) of the vector g_i.

Consider a genomic region that spans a set of m SNPs. The frequencies of the genotypes in that region across the entire population are determined by some unknown distribution function $\Pr(g_i \in \mathcal{G})$, where $\mathcal{G}$ is the sample space of all genotypes in the population. Let S be the set of all SNPs, and suppose a set T of t tags was selected (i.e., $|T| = t$). A prediction algorithm is a function $f : \{0,1,2\}^t \to \{0,1,2\}^m$. Informally, the prediction algorithm uses the genotype values of the tag SNPs in T in order to predict the values of the rest of the SNPs. For a given vector $w \in \{0,1,2\}^t$ of tag SNPs values, let $[f(w)](j)$ be the j-th component of that vector. Finally, let $z_T : \{0,1,2\}^m \to \{0,1,2\}^t$ be the restriction of the genotype to the tag SNPs in T.

Formally, for a given number t, our objective is to find a set T of t tag SNPs, and a prediction function f, such that the following expression is minimized.

$$\eta = \sum_{j=1}^{m} \Pr[[f(z_T(g))](j) \neq g(j)], \tag{2.1}$$

where the probability is over the sample space given by $\Pr(g \in \mathcal{G})$. In other words, for a randomly picked individual from the population, we want to minimize the expected number of prediction errors.

The prediction algorithm used in STAMPA is based on the observation that the correlation between SNPs tends to decay as the physical distance increases. This monotonicity does not always hold but is a good practical approximation of biological reality. Formally, the algorithm assumes that given the genotype values of two SNPs, the probabilities of the values at any intermediate SNPs do not change by knowing the values of additional distal ones. Thus, the prediction

ALGORITHM Predict(i, j_1, j_2, a_1, a_2)

Input: $i, j_1, j_2 \in \{1, \ldots, m\}$, and $a_1, a_2 \in \{0, 1, 2\}$.

Output: An integer $v \in \{0, 1, 2\}$ which is a predicted value of a SNP in position i, given that in position j_1 and j_2 the values are a_1 and a_2 respectively.

1. For every $(x, y, z) \in \{0, 1\}^3$ we let $C(x, y, z) = \{(j, p) \mid h^p_{jj_1} = x, h^p_{jj_2} = y, h^p_{ji} = z\}$ be the set of haplotypes having the values x, y, z in positions j_1, i and j_2 respectively.
2. Let $A(x, y) = z \in \{0, 1\}$, where $|C(x, y, z)| \geq |C(x, y, 1 - z)|$ breaking ties arbitrarily.
3. Let $c(x, y) = |C(x, y, 0)| + |C(x, y, 1)|$.
4. We compute the values of two variables x, y using the following case analysis.
   - If $a_1 < 2$ and $a_2 < 2$, then we set $x = y = A(a_1, a_2)$.
   - If $a_1 = 2$, $a_2 = 2$ and $c(0,0) \cdot c(1,1) \geq c(0,1) \cdot c(1,0)$, then $x = A(0,0)$ and $y = A(1,1)$.
   - If $a_1 = 2$, $a_2 = 2$ and $c(0,0) \cdot c(1,1) < c(0,1) \cdot c(1,0)$, then $x = A(0,1)$ and $y = A(1,0)$.
   - If $a_1 = 1$, $a_2 = 2$ ($a_2 = 1, a_1 = 2$), then we set $x = A(1,1)$ and $y = A(1,0)$ ($y = A(0,1)$).
   - If $a_1 = 0$, $a_2 = 2$ ($a_2 = 0, a_1 = 2$), then we set $x = A(0,0)$ and $y = A(0,1)$ ($y = A(1,0)$).
5. If $x \neq y$ output 2, else output x.

Figure 2.6: The procedure Predict. We implicitly assume that the training set and its phase are given. The variables x and y computed by the case analysis represent the majority votes for the two haplotypes induced by the values a_1 and a_2. Note that the output value is determined by simply counting the frequencies of different partial haplotypes in the training set that match a_1 and a_2 and taking the majority vote. Source: [21]

function predicts a SNP value using only the values of the two closest tag SNPs flanking it, while ignoring the values of all other tags. Given a set of tag SNPs $T = (s_1, \ldots, s_t)$, we use the procedure Predict given in **Figure 2.6** to predict the value of SNP i given the value of its flanking tag SNPs. We assume that we are given the training set of genotypes $g_1, \ldots, g_n$ together with their corresponding haplotypes $h^1_1, h^2_1, h^1_2, \ldots, h^2_n$, where $h^k_i = (h^k_{i1}, \ldots, h^k_{im}) \in \{0, 1\}^m$ for $k = 1, 2$ (if only genotypes are given, the haplotypes can be resolved as described in Section 2.3). Let j_1 and j_2, $j_1 < i < j_2$ be the positions of the tag SNPs closest to position i on both sides. If there is no tag SNP in position $j_2 > i$, then j_1 and j_2 are the two rightmost tag SNPs, and if there is no tag SNP in position $j_1 < i$ then j_1 and j_2 are the two leftmost tag SNPs. The procedure $Predict(i, j_1, j_2, a_1, a_2)$ uses a majority vote in order to determine which value is more likely to appear in position i given that positions j_1 and j_2 have the values $a_1 \in \{0, 1, 2\}$ and $a_2 \in \{0, 1, 2\}$, respectively.

The goal of the algorithm is to find a set of tag SNPs T of size t, such that expression 2.1 is minimized when the genotype is randomly picked from the training set. Practically, we want to minimize X_T where $X_T = |\{(i, j)|g_{i,j} \neq$

$Predict(j, j_1, j_2, g_{i,j_1}, g_{i,j_2}))\}|$. STAMPA uses dynamic programming in order to solve this problem to optimality. See [21] for details.

Since correlation between SNPs usually decays with distance between the SNPs, tag SNPs will only be useful when predicting SNPs in their close neighborhood. So STAMPA accepts as input a parameter indicating the maximum allowed distance between tags. By considering only pairs of SNPs within the specified distance a major reduction in running time is achieved.

2.5 Association Studies

The holy grail of human genetics is finding genetic variations that are associated with a certain disease. The abundance of SNPs and the ease of SNP genotyping make them the genetic markers of choice for most association studies. In these studies researchers look for certain alleles that predispose their carriers to a certain disease. High throughput genotyping techniques are rapidly progressing and as much as one million SNPs can be genotyped today with Affymetrix's Genome-Wide Human SNP Array 6.0 [62] and Illumina's human1M beadchip [64]. In parallel, the cost of per SNP genotyping is dramatically decreasing, making association studies with thousands of patients a reality. For example, the Wellcome Trust Case Control Consortium recently performed an association study with a total of 14,000 patients from seven common diseases [60].

Generally, association studies are divided into two categories: *family-based* and *population-based*. In the family-based studies, affected individuals and their parents are collected. Then, one searches for alleles that are transmitted from parents to their affected child more often than would be expected by chance. The population based studies are comprised of unrelated affected individuals (called *cases*) and unrelated healthy individuals (called *controls*). In this study design one searches for alleles whose frequency among the cases is different from their frequency among the controls. In both study designs, finding associated SNPs is not the end of the story. An associated SNP may be the direct cause for the disease, but alternatively, it may only be genetically linked to the causal SNP. Therefore, further investigation and fine-mapping of the areas around associated SNPs are usually necessary.

Many different tests have been proposed and used in association studies. The most widely used test for family-based studies is the transmission disequilibrium

test (TDT), introduced by Spielman et al. [53]. The TDT measures the over-transmission of an allele from heterozygous parents to affected offsprings. Let the two alleles of a SNP be denoted as A and a. Then the number of transmitted and non-transmitted alleles from parents can be summarized in a 2-by-2 table (Table 2.1).

	non-transmitted	allele
Transmitted allele	A	a
A	b	c
a	d	e

Table 2.1: TDT counts table

Under the null hypothesis of no association between the SNP and the disease, every heterozygote parent has an equal chance of transferring allele A or a to its child. When summing over all heterozygous parents it translates to $c = d$. This hypothesis can be tested using a binomial (asymptotically χ^2) test with one degree of freedom (df):

$$X^2 = \frac{(c-d)^2}{c+d}$$

An extended version of the TDT test [52] can also handle multi-allelic markers (e.g. microsattellites).

In population-based studies the most common test for single SNPs, suggested by Olsen et al. [39], is based on building a contingency table of genotypes vs. disease status (case/control). The paired observations, expressed in the contingency table, are tested for independence by the standard Pearson χ^2 test. There are several ways to build the contingency table on which the test is performed. The most general and conservative approach is to build a 3×2 contingency table of genotypes vs. disease status (Table 2.2). The χ^2 test on this table has two dfs and it can capture association with the disease under any kind of disease model. Merging the first two rows of the table by summing up their values creates a 2×2 table. The 1-df χ^2 test on this table corresponds to a dominant disease model (Allele A is the dominant one). A test under recessive disease model is achieved similarly by merging the last two rows. A widely used approach is building a 2×2 table that counts alleles instead of genotypes (Table 2.3). The χ^2 test here corresponds to a multiplicative penetrance model of disease, meaning that each

copy of allele A multiplies the risk by the same factor. The 2-df test has the advantage of being able to capture all kinds of disease models but with reduced power compared to the 1-df tests.

Genotype	Case	Control
AA	b	c
Aa	d	e
aa	f	g

Table 2.2: Case/control 3×2 contingency table – counting genotypes

Allele	Case	Control
A	$2b + d$	$2c + e$
a	$2f + d$	$2g + e$

Table 2.3: Case/control 2×2 contingency table – counting alleles

Another approach widely used is *logistic regression*. Logistic regression methods model the probability of disease p as a function of the genotype. The most general model codes a genotype with two variables x_1 and x_2. $x_1 = 1$ for a heterozygote and 0 otherwise. $x_2 = 1$ for a homozygote AA and 0 otherwise. The linear model is

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The maximum likelihood estimates of β_0, β_1, and β_2 can be calculated with standard generalized linear models tools. No association between the genotype and the disease translates in this model to $\beta_1 = \beta_2 = 0$. The log-likelihood ratio (LLR) test, whose statistic is $2(l_1 - l_0)$, is used to test for association. l_1 denotes the log-likelihood of the full model (i.e. , the maximum likelihood estimates of β_0, β_1, and β_2 are found and the log-likelihood of the model under these parameters is calculated) and l_0 denotes the log-likelihood of the null model (forcing $\beta_1 = \beta_2 = 0$). The test is asymptotically distributed as χ^2 with 2 dfs, and for large sample sizes it is equivalent to the Pearson χ^2 2-df test. It is easy to fit specific disease models to the logistic regression model. For example, forcing $\beta_2 = 2\beta_1$ corresponds to a multiplicative model. It is also easy to add covariates, such as age and sex, and to handle multiple SNPs, including interactions between them (all at the expense of adding variables to the model).

The block-like structure of the human genome suggests using another promising strategy – haplotype-based methods. The general scheme in these methods is to partition the data into blocks of little recombination, infer the haplotypes within each block and test these haplotypes for association with the disease. The simplest analysis involves treating a block as a k-allelic marker, where k is the number of haplotypes in the block. Then a $2 \times k$ contingency table is built and a $k-1$ df χ^2 test is performed. However, this method ignores the uncertainty of haplotype assignment for individuals. Zaykin et al. [67] suggested a method in which haplotype frequencies are estimated through the EM algorithm, and each individual in the sample is expanded into all possible haplotype configurations with corresponding probabilities, conditional on their genotype. A regression-based approach is then used to relate inferred haplotype probabilities to the case/control status. Schaid et al. [48] used generalized linear models to develop efficient score statistics for haplotype specific tests. Both methods, and others, were implemented in the software package WHAP (Haplotype-based association analysis package) [44].

One common problem with the above haplotype-based methods is what should be done about rare haplotypes. Including them in the analyses can lead to loss of power because there are too many degrees of freedom, but excluding them may result in loss of valuable information. One possible remedy is to use clustering to identify sets of haplotypes that are assumed to share recent common ancestry and therefore convey a common disease risk that is more frequent among cases. This approach (often called cladistic) was first introduced by Templeton et al. [57], who recently developed a method called Tree Scanning [58]. This method uses evolutionary trees of haplotypes to study phenotypic associations by exhaustively examining all possible biallelic partitions of the tree.

Association between SNPs and disease status is not the only connection researchers are after. Finding association between SNPs and other phenotypes, including continuous (or quantitative) traits such as blood pressure, height, and weight is also of great importance. There are several statistical tools for finding association between genomic loci and quantitative traits (such loci are termed Quantitative Trait Loci – QTLs). The most natural one is analysis of variance (ANOVA). In this analysis the mean values of the trait in each of the three genotype groups are compared. The null hypothesis of no association corresponds to equal means. An alternative to ANOVA is linear regression where a linear

relationship between mean value of the trait and genotype are assumed. In either case, tests require the trait to be approximately normally distributed for each genotype, with a common variance. The Kolmogorov-Smirnov (KS) test is another alternative. In this test the cumulative distributions of the trait among the different genotype groups are compared.

The number of SNPs tested in association studies is rapidly increasing. As mentioned above, DNA chips of one million SNPs are already available. When using such chips, the number of statistical tests performed is extremely high and spurious associations might arise if the method to declare statistical significance does not take the multiple testing problem into consideration. One typical correction is the Bonfferoni correction, where each marker's p-value is multiplied by the number of tests. However, this correction does not take into account the dependence of linked maker loci, and may lead to over-conservative conclusions. A widely used alternative procedure is permutation testing, suggested for disease association by Zhang et al. [68]. The procedure works as follows: the χ^2 score of each marker is calculated, and the maximum value over all markers is chosen as the test statistic. Then, the same statistic is calculated for many data sets with the same genotypes and randomly permuted labels of the disease status. The corrected p-value of the highest scoring marker is the fraction of times the value of the statistic in the permuted data exceeds the original value. The advantage of this test is that correlations between markers are maintained, so the corrected p-value tends to be less conservative than the Bonfferoni corrected p-value. The disadvantage is that performing the test with k permutations gives p-values not smaller than $1/k$. Therefore, it can become prohibitively slow to compute low (high significance) p-values in large studies. Kimmel and Shamir [28] developed a faster algorithm for calculating low p-values which is based on importance sampling.

Chapter 3

The impact of tag SNP selection criteria on the power of association

Prior to an association study, a large number of SNPs are typed in a small sample of individuals from the population. Based on this sample, tag SNPs are selected, and only these tag SNPs are typed in a larger set of cases and controls. Thus, the use of tag SNPs reduces the overall cost of the study. The eventual goal of the study is to find sites in the genome that are significantly associated with the disease. An important issue that has not been fully addressed so far in tag SNP selection is the power: How does the selection method affect the power of the subsequent association study?

Here we demonstrate that the prediction accuracy criterion produces better tag SNPs than the r^2 criterion, in terms of the power of association studies using the tags. We performed extensive simulations on 5,000 case-control panels which were generated from SNPs of the HapMap dataset. We tested two tag SNP selection algorithms: STAMPA – which uses the prediction accuracy, and Tagger – a state of the art algorithm that uses the r^2 criterion. Our simulations show that STAMPA attains higher power than Tagger, over a wide range of significance levels. When the tag SNPs are selected at high density, both methods show high power. As the density of selected tag SNPs decreases, STAMPA's advantage grows larger, reaching up to 18% increase in the relative power. Hence, the method for tag SNP selection has a major effect on the eventual chances of

finding the disease association.

This study, which was done jointly with Dr. Gad Kimmel and Dr. Eran Halperin, was submitted for publication and is currently under review.

3.1 Introduction

Genome-wide disease association studies are becoming a reality due to recent technological advances, and many such studies are under way. As described in Section 2.5, the goal of such studies is to identify the genetic causes of complex diseases, by finding sites in the human genome that are correlated with the disease. In these studies, differences in the frequencies of SNPs between cases and controls are evaluated statistically. These discrepancies serve as evidence for the association of particular SNPs with the disease. The *significance* of a test is the probability to declare association when the SNPs are not associated with the disease. The *power* of a test is the probability to correctly conclude that there is association when one of the SNPs is associated with the disease.

The statistical significance and power of a study are directly affected by the number of tested individuals and by the number of SNPs typed. When more SNPs are typed, more information is obtained, and the chance to find an association, if such exists, increases. On the other hand, resources are limited so only a certain number of SNPs can be typed. Therefore, in every study a given number of tag SNPs are to be selected and typed. A key problem is to find a set of tag SNPs of a given size that would have a sufficiently high power.

The tag SNP selection problem has been under study for about seven years, and is still very important for association mapping, in three scenarios. First, there are still many candidate gene studies, e.g., [18, 30, 42, 8, 56], which seek association around a focused number of genes rather than genome-wide. Such studies have two advantages: their cost is smaller than the cost of whole genome association studies, and they are less prone to power loss due to large multiple hypothesis testing. Second, in many cases, following a genome wide association study, one wants to obtain fine mapping of suspected regions or genes. Third, new tag SNPs must be selected in the course of developing a new SNP chip.

There are many methods for selecting a desirable set of tag SNPs (see Section 2.4). Recently, de Bakker et al. [12] suggested a method called *Tagger* for

tag SNP selection based on the r^2 criterion. The method is explained in detail in Section 2.4.1. Briefly, Tagger uses a greedy algorithm to select tags that capture all other SNPs with r^2 exceeding a prescribed threshold. In addition, to improve efficiency, it can carry out an aggressive search attempting to replace each tag with a specific multi-marker predictor (on the basis of the remaining tags). Tagger is implemented in the Haploview software [5] developed as part of the HapMap project, and is very popular among geneticists. The method was applied in a recent important study [11] to show that the HapMap DNA samples can be used to select tags for genome-wide association studies that would remain informative when using other samples from different populations around the world.

The power of Tagger in association tests was not fully addressed so far. Moreover, while there are theoretical reasons that justify using r^2 as a measure for selecting tag SNPs [43], Tagger's greedy heuristic for choosing the tag SNPs provides no guarantee for optimality.

Halperin et al. [21] proposed a novel measure for tag selection, called *prediction accuracy*, which directly evaluates the average SNP prediction quality. They implemented tag SNP selection and prediction algorithms that optimize this criterion, in a program called *STAMPA*, described in Section 2.4.2. Under certain biologically realistic restrictions, STAMPA was shown to find an optimal set of tag SNPs using a dynamic programming algorithm.

STAMPA was compared to two state of the art algorithms that were available at the time of its publication: ldSelect [7] (an algorithm very similar to Tagger, which also uses the r^2 criterion) and HapBlock [70]. STAMPA consistently outperformed both of these methods on several different data sets in terms of prediction accuracy. However, although intuitively the prediction accuracy is a reasonable optimization criterion, the power of an association study that uses the tags is a more meaningful optimization criterion, as it directly measures the chance of success of the study. So far no comparison of tag SNP selection algorithms in terms of power was made. This raises the following question: given several tag SNP selection algorithms, which one would give the highest power to association studies that use tag SNPs selected by it?

In this chapter we study in a systematic fashion the relation of tag SNP selection, association and power. We would like to choose a specific number of tag SNPs that represent the studied population, out of a larger set of reference

SNPs in that population. These tags are to be typed in cases and controls of new individuals collected from the population. For a fixed significance level (say, 5%), the goal is to choose the tag SNPs such that the power is maximized. Which tag SNP selection method would give higher power?

To empirically test the power of Tagger and STAMPA, we conducted extensive simulations in a similar way to de Bakker et al. [12]. In order to have a realistic scenario, we worked with real SNPs from the HapMap project [63]. We sampled 100 disjoint regions of 1,000 SNPs each from chromosome 1, and selected tag SNPs in each region. We then simulated cases and controls according to a multiplicative model, where the causal SNP is randomly chosen. We performed permutation tests to evaluate the significance of each experiment. In total, 5,000 case-control panels were generated and 155,000 association tests were performed. We tested both STAMPA and Tagger (in both its pairwise and aggressive modes). Our simulations show that STAMPA attains higher power than Tagger, across a broad range of significance levels. When Tagger picks tags that capture the entire region with $r^2 \geq 0.8$, the tag SNPs density is relatively high (one tag per 5-8 kb), resulting in a comparable high power for both algorithms. At lower densities of selected tag SNPs, STAMPA attains an advantage that grows larger as the tags density decreases. The relative power of STAMPA is up to 18% higher than Tagger's when the tag SNP density is low (one tag per 40 kb).

Our results show that the method for tag SNP selection has a major effect on the eventual chances of finding the disease association. The optimization of the prediction accuracy by STAMPA yields a clear advantage in terms of power, and particularly, it is more powerful than Tagger.

3.2 Methods

3.2.1 Datasets

We used phased data from parents of CEU (west European ancestry) trios from the HapMap resource (release 21) [63]. We first extracted the 120 haplotypes of chromosome 1. Then, SNPs with minor allele frequency $< 5\%$ were removed, and the dataset was partitioned into 100 contiguous regions of 1,000 SNPs per region (spanning the first 172Mb of chromosome 1). The average distance between adjacent SNPs was 1,364bp. For each region, the matrix of 120 haplotypes $\times$

1000 SNPs constitutes a *reference panel*.

3.2.2 Comparisons

We compared three algorithms: STAMPA (as implemented in the software package GEVALT [10], see also Chapter 4), Tagger (as implemented in Haploview) and *Rand* – an algorithm that randomly selects tag SNPs. Tagger was tested in both the standard mode and in its "aggressive mode", which uses haplotypes (we call the latter here *Tagger-A*). Several tagging scenarios were tested (see Results section). The following procedure was repeated for each tagging scenario (see **Figure 3.1** for a flow chart):

1. Repeat for each of the 100 reference panels:
   (a) Select tag SNPs by each algorithm using the reference panel as the training set.
   (b) Repeat 50 times:
      i. Select a causal SNP at random and simulate a *test panel* of genotypes for 1000 cases and 1000 controls as described below.
      ii. For each set of tag SNPs selected in step 1.a, hide the values of the non-tag SNPs in the test panel and compute the significance of association between the disease status and the values of the tag SNPs in the test panel. A detailed description of this computation appears below.
      iii. Compute the significance of association between the disease status and the values of all SNPs in the test panel (i.e. without hiding SNPs).
2. Compute the *relative power* of each algorithm as described below.

For simulating case/control panels (step 1.b.i) we used the HAPGEN software [36]. HAPGEN accepts as input a set of known haplotypes, an estimate of the fine-scale recombination rate across a region, the causal SNP, and the disease model parameters in terms of relative risk. Genotypes for cases and controls at the causal SNP are simulated under the disease model, and data at flanking SNPs are simulated using the known haplotypes. These haplotypes are recombined to

form the simulated haplotypes using a Hidden Markov Model approximation to the Li-Stephens population genetics model [33]. For the disease model we used a standard multiplicative model [70] with relative risk of 1.5 (i.e. heterozygous relative risk of 1.5 and homozygous relative risk of 2.25). As the set of known haplotypes we used the 120 HapMap haplotypes. The fine-scale recombination rates were estimated by the PHASE program (rates were downloaded from HAPGEN's website).

In step 1.b.ii, for each panel and for the tag SNPs selected by each algorithm, χ^2 allelic tests with one degree of freedom were computed and the p-values of association were computed using a permutation test (done with the RAT software [28]). For STAMPA, the set of tag SNPs was used to predict the values of the other SNPs (as described in Section 2.4.2), and the entire set of SNPs was tested for association. For Tagger-A, the set of tests was generated as described in Section 2.4.1. Since Tagger-A uses haplotypes in the case-control panel to generate the tests, we also tested a version of STAMPA that uses haplotypes instead of genotypes in the prediction, in order to have a fair comparison. For Tagger and Rand, the χ^2 tests were applied only to the tag SNPs.

After completing step 1 of the testing procedure we have a list of 5000 p-values for each algorithm (100 regions and 50 test panels per region). In step 2 we compute the relative power of each algorithm as follows. For a fixed significance level p_0, we count the number of test panels in which the algorithm had a significant ($\leq p_0$) p-value. We then divide this number by the number of panels that are detected as significant when all SNPs (including the non-tags) are available (computed in step 1.b.iii). Hence, this measure evaluates how well an algorithm performs in comparison to an algorithm that has full genotype (or haplotype) information. In order to test if the difference in relative power between two algorithms is significant we use a paired, 1-tailed T-Test between 0/1 vectors indicating the success on each panel. Significance levels p_0=0.05, 0.01, 0.005, and 0.001 were used.

3.3 Results

Several tagging scenarios were explored. The first scenario used Tagger's default parameters ($r^2 \geq 0.8$) and pairwise tagging in order to capture the entire set of SNPs. On average, 26.4% (standard deviation (SD) 6.3%) of the SNPs were

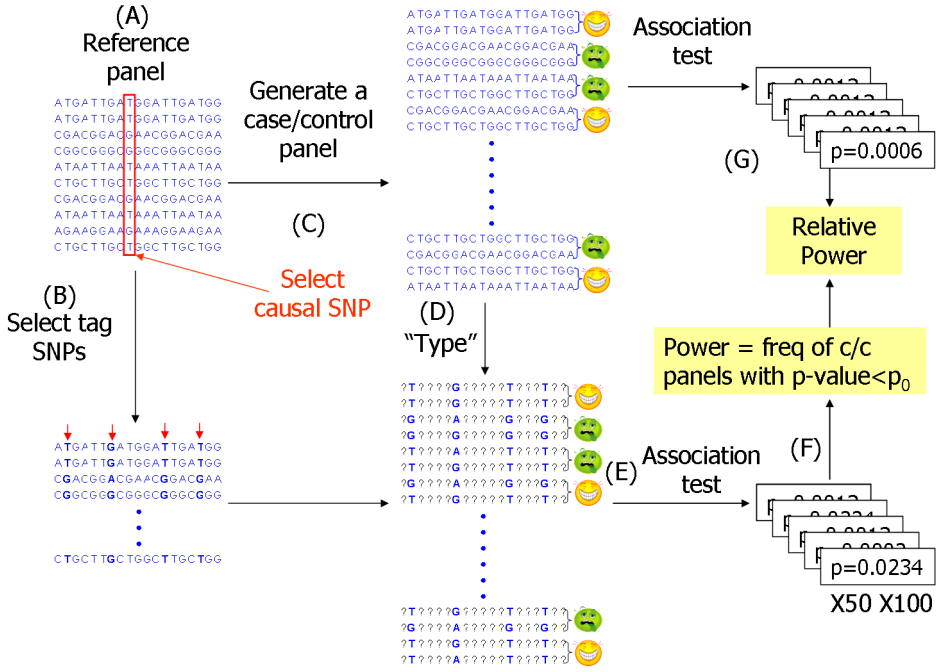


Figure 3.1: A flow chart of the testing procedure. The procedure for one test panel in one region and for one tag selection algorithm is shown. (A) We start with a reference panel containing the 120 HapMap haplotypes. (B) Tag SNPs are chosen. (C) A causal SNP is randomly selected and a panel of 1000 cases and 1000 controls is simulated. (D) The non-tag SNPs are hidden. (E) The case/control panel is tested for association. (F) Steps C-E are repeated for 50 times. The whole procedure is done in each of the 100 regions, resulting in a list of 5000 p-values from which the power of the tag SNP selection algorithm is calculated. (G) Each case/control panel is also tested for association without hiding the non-tag SNPs. This allows the calculation of the relative power of each algorithm.

selected as tags (an average of one tag SNP per 5.4kb). **Figure 3.2A** shows the power of the three algorithms. Both STAMPA and Tagger obtain a very high relative power (over 96% under all four significance levels) with non-significant differences between the methods (T-test p-values > 0.14). Rand has significantly lower power than both, but interestingly, it attains quite high relative power of over 90%.

We next applied Tagger's aggressive method (using 2 and 3-marker haplotypes, and default parameters $r^2 \geq 0.8$, LOD threshold 3.0) in order to capture the entire set of SNPs. The average fraction of SNPs selected as tags was 21.2% (SD 5.3%), i.e., on average one tag SNP per 6.7kb. **Figure 3.2B** shows the comparison between the algorithms. Here STAMPA has a significant advantage over Tagger for p-value cutoffs 0.001 and 0.005 (T-test p-values of $8.9 * 10^{-5}$ and 0.04 respectively).

The amount of tag SNPs needed by Tagger to capture all the SNPs in a region was quite high: around 25%. We next explored scenarios where the tag density is lower. We used Tagger's 'best N' method, which, for a prespecified N, seeks the N tags that capture the maximum number of SNPs with the prescribed r^2 threshold (see Section 2.4.1). We set N to be 34, 45, 68, and 136, which corresponds to an average of one tag SNP per 40, 30, 20, and 10kb, respectively. We tested both Tagger and Tagger-A (with default parameters $r^2 \geq 0.8$, LOD threshold 3.0). Results are shown in **Figure 3.3**. All algorithms show a linear decrease in power as the tag density decreases, but the rate of decrease in STAMPA's performance is much lower than Tagger's. In all tests STAMPA attains a higher power than Tagger but with varying differences. The advantage in the relative power is 11% at a tag distance of 40kb. STAMPA's advantage is smaller but statistically significant also at tag distances of 20 and 30kb under all significance levels (T-test p-values < 0.007), but not at 10kb. Interestingly, Tagger-A had essentially the same performance as Tagger. When STAMPA uses haplotypes instead of genotypes in the prediction STAMPA's advantage is larger, and the differences are statistically significant at all tag distances. The highest advantage is achieved at a tag distance of 40kb and a p-value cutoff of 0.001 where there is 18% improvement in power.

We wanted to verify that the gain in STAMPA's advantage is due to the different tags chosen and not merely because STAMPA predicts the values of the non-tag SNPs and Tagger does not. For that, we selected tags with Tagger, predicted the non-tag SNPs based on them using STAMPA's prediction method, and computed the association for each tag and for each predicted non-tag. The results for the 'best N' scenario can be seen in **Figure 3.3**. The prediction algorithm does improve Tagger's relative power (all T-test p-values < 0.05) but only by a modest factor (average increase of 1.1%), and still well below STAMPA.

Generally, the same trend was observed under all four significance levels

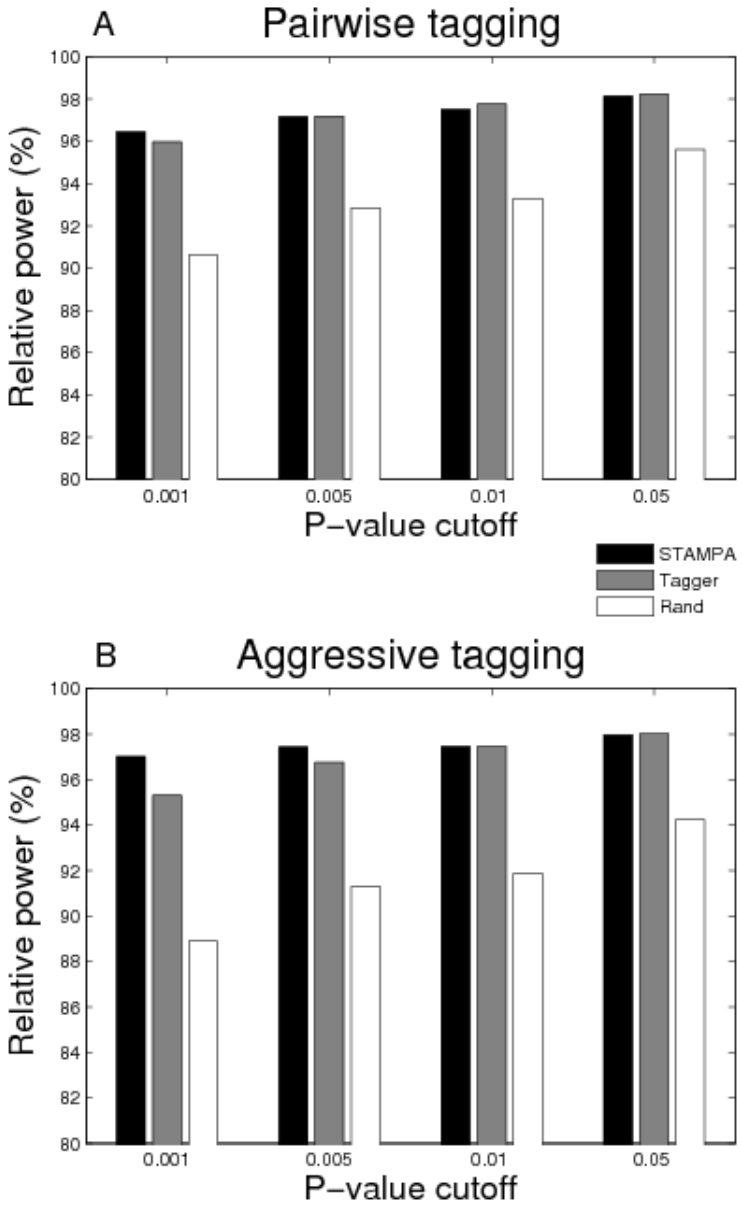


Figure 3.2: Comparison of relative power when all SNPs are captured by Tagger. Relative power is shown for STAMPA, Tagger , and Rand, calculated for four different association significance levels (p-value cutoffs). A: Comparisons to Tagger. B: comparisons to Tagger aggressive. Note that the number of SNPs selected in each case was determined by Tagger, and then used in the two other algorithms.

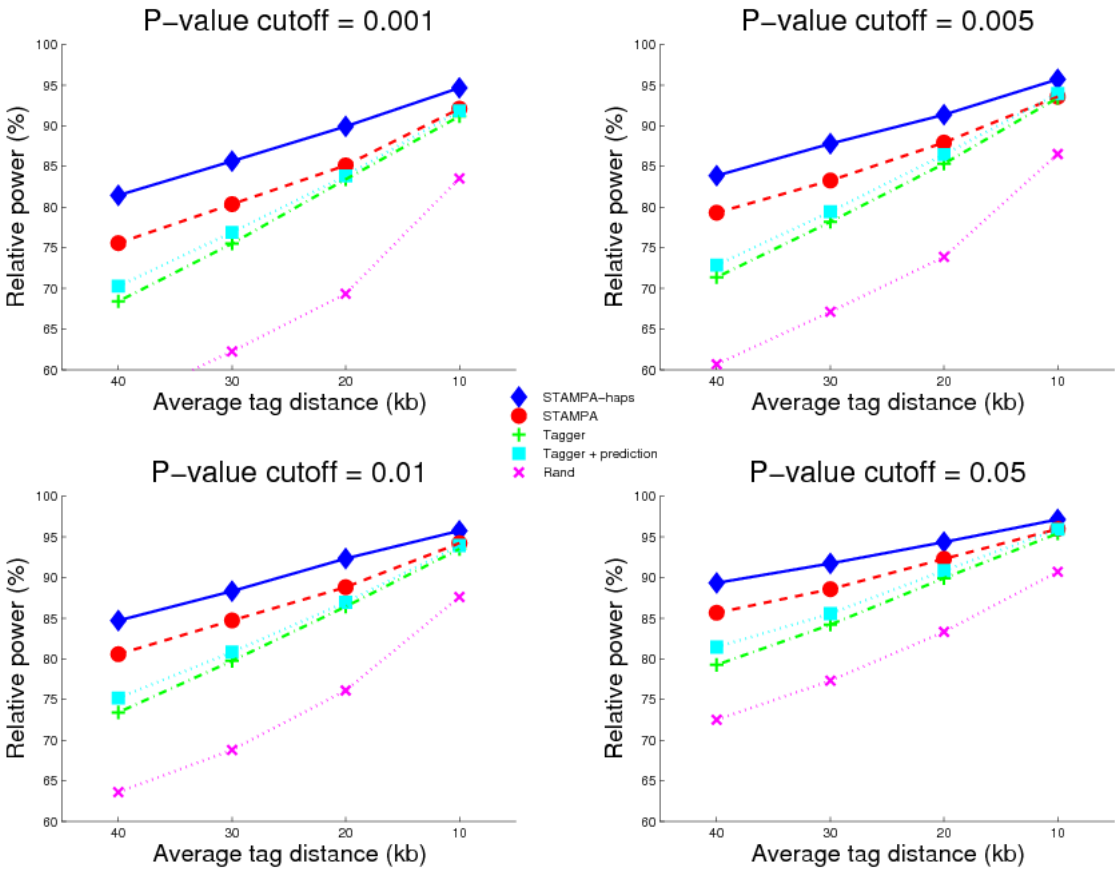


Figure 3.3: Comparison of relative power when the number of tag SNPs is limited. Relative power is shown for STAMPA, Tagger, Rand, STAMPA using haplotypes, and Tagger using prediction. The number of tag SNPs per tested region was limited to 34, 45, 68, and 136, which corresponds to an average distance between tags of 40, 30, 20, and 10kb, respectively. The power is shown in four different p-value cutoffs. Tagger-A gave virtually identical results to Tagger and is therefore omitted from the figure.

tested. However, STMAPA's advantage is slightly larger under low significance levels (0.001 and 0.005).

Since STAMPA and Tagger use different selection criteria we wanted to test how STAMPA performs in terms of the r^2 criterion. To test this, we calculated the r^2 coverage of tags selected by each method: For each threshold θ, we calculated the fraction of non-tag SNPs that are captured by some tag with $r^2 \geq \theta$. **Figures 3.4A,B** show that Tagger has a higher r^2 coverage. This is expected, since Tagger uses the r^2 criterion in the optimization process.

We also compared the ability of each method to predict the non-tags, in the following way. For each set of tags, we predicted the non-tags in the original HapMap haplotypes and calculated the r^2 between each real non-tag and its prediction. **Figures 3.4C,D** demonstrate a large advantage of STAMPA over Tagger. In fact, Tagger's prediction ability is very similar to that of random selection of tags. The large difference is also manifested by the error rates when prediction was applied on the simulated panels. The error rate is defined as the fraction of wrongly predicted alleles. At tag distances of 10, 20, 30, and 40kb, the average error rates of STAMPA were 3.9, 5.2, 6.2, and 7.1%, respectively, while Tagger's error rates were almost twofold higher: 8.9, 9.4, 10.5, and 11.5%, respectively.

3.4 Discussion

In this chapter, we compared the prediction accuracy and r^2 as criteria for tag SNP selection when the goal is to maximize the power of association studies. Our experimental results show that STAMPA, which optimizes the prediction accuracy, attains significantly and consistently higher power than Tagger, which optimizes r^2, in extensive simulations with real haplotypes. This strong advantage in the power is explained primarily by the selection criterion of STAMPA, prediction accuracy. An alternative possible explanation is that the advantage is mainly due to the fact that STAMPA guarantees optimality, while Tagger uses a local search heuristic which does not guarantee optimality. The fact that STAMPA has a higher power although its tags have a lower r^2 coverage (**Figures 3.4A,B**) rejects the latter explanation. Moreover, applying the prediction algorithm on Tagger's tags only marginally improves Tagger's power (**Figure 3.3**), which implies that the choice of tag SNP selection algorithm is important.

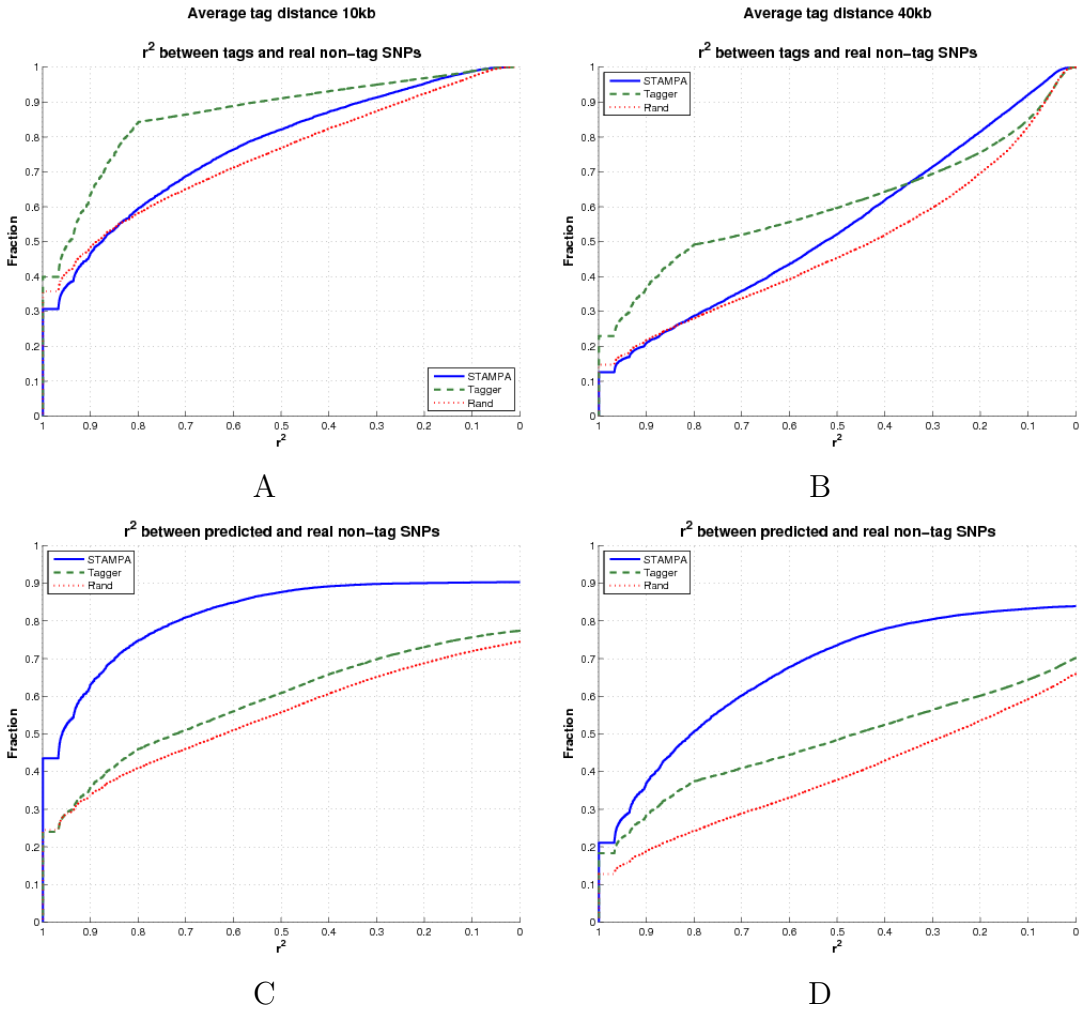


Figure 3.4: Average SNP coverage as a function of the r^2 threshold. A,B: For every value of r^2 (X-axis) the plots show the fraction of non-tag SNPs that have equal or higher r^2 with some tag SNP. C,D: For every value of r^2 (X-axis) the plots show the fraction of non-tag SNPs that have equal or higher r^2 with their predicted values as computed on the HapMap haplotypes. The curves do not reach 100% since some of the SNPs are predicted to be monoallelic, which makes the r^2 measure undefined. Results are shown for the datasets with 10kb (A,C) and 40kb (B,D) tag distance. Tagger-A gives almost identical results to Tagger (not shown).

As expected, when the density of tag SNPs is smaller, the power advantage of STAMPA grows larger. As the tag densities decrease, both methods show a linear decline in power, but STAMPA's decline is much more moderate than Tagger's. When the density of tags is high, capturing all SNPs with $r^2 \geq 0.8$, both Tagger and STAMPA obtain high power (above 96%). On the other hand, if only a small fraction of the SNPs is to be selected and typed, the algorithm is of great importance, with a prominent advantage to STAMPA. It is worthwhile mentioning that STAMPA maintains quite good power even when the selected tags are sparse. For example, using significance threshold of 0.05, STAMPA's relative power is 86% when the tag distance is 40kb.

When using haplotypes instead of genotypes in the prediction, STAMPA's advantage is larger, and the differences are statistically significant for all tag distances. It suggests that in real scenarios, if experimental information on phasing is not available (e.g. from trios), one may gain power by doing the association test on haplotypes that have been phased computationally. Testing this scenario in our simulations was not possible due to the added computational burden of phasing many thousand panels. On a limited number of examples we did perform phasing using fastPHASE [49], and measured the error rate of the phasing algorithm. Our tests showed that the switch error rate [49] is small at tag distances 10kb or less (2-4%) (data not shown). Therefore, the results of STAMPA with computationally derived haplotypes are expected to be only marginally less accurate than if we were to use the true haplotypes, and the results of STAMPA with haplotypes are close to a valid representation of this real scenario. Hence, even at high tag densities STAMPA probably has an advantage over Tagger if phasing of the tags is performed prior to predicting the non-tags.

Since the same haplotypes used to select tag SNPs were also used to create the case/control data, there is a possible risk of overfitting in our results. Due to the scarcity of large-scale published haplotypes we could not separate completely the data for the training and the test phases. As a partial remedy to this problem, we created the test panels with the HAPGEN software [36]. HAPGEN simulates case and control individuals conditional upon a set of known haplotype data using an estimate of the fine-scale recombination rate across a region. This approach uses the Li-Stephens model [33], and is preferable over a direct resampling approach [12], which produces new haplotypes that are copies of the original HapMap haplotypes. Even if there is a modest amount of overfitting

in the reported power levels, since the same simulated data sets were used by STAMPA and Tagger, the relative performance should be accurate.

One might argue that the tag SNP selection problem is currently of limited interest, since today many studies use standard high-density SNP chips (e.g. of Illumina and Affymetrix), and have no control over the tag selection; moreover, genotyping costs are rapidly decreasing, and high density tags can be achieved. Still, finding a better set of tags remains an extremely important issue for three main reasons: First, certain platforms (such as Illumina's Golden-Gate and Affymetrix's Custom SNP Kits) are designed for customized genotyping, and require ad-hoc tag selection for each new study. Many focused association studies cannot use standard arrays and have to select tags for specific genomic areas, often at modest density due to budget constraints (this large need is demonstrated, for example, by over 150 studies citing Tagger since October 2005). Second, each new generation of standard high-density chips involves the selection of a new set of tag SNPs, and therefore the future chips produced by these technologies may benefit considerably from using more powerful selection methods. While average tag densities in such chips will grow, there are always sparse genomic regions showing low LD, on which prediction-based tag selection will have an advantage.

Recently, several works [36, 66, 51, 41] used algorithms for predicting untyped SNPs based on typed tag SNPs. The idea was to use all the imputed SNPs when performing the association test. This approach was shown to have higher power than using only the tags, without the predicted SNPs. This gives additional support to using tag selection algorithms based on prediction accuracy: Since a prediction procedure is performed on the tags, and only then the association is tested, it makes sense to choose the tags that predict the rest of the SNPs with minimal error rate (recall the differences between STAMPA and Tagger in **Figures 3.4C,D**). In other words, we claim that if STAMPA would be used to select the tags in such methods instead of r^2 based selection, a higher power will be achieved.

Every association study is preceded by selection of the tag SNPs to be typed, whether explicitly by the researchers or previously, by chip producers. The success of the study critically depends on its power. Here, we demonstrated that using the prediction accuracy in the tag selection stage is preferable over the widely used r^2 measure. In our experiments we assumed that association is tested on each SNP separately, which is the common practice. However, different measures

for tag SNP selection may give higher power in other testing approaches. Even in the scenario tested here we do not claim that STAMPA gives optimal power, only that it outperforms today's common practice. Therefore, the connection between tag SNP selection and power deserves further exploration.

Chapter 4

GEVALT

Most researchers in genetics prefer using software with graphical user interface (GUI) over using batch executables. Since algorithms developed in our lab for genotypes phasing (GERBIL), and tag SNP selection (STAMPA) were only available as batch executables, we developed an integrated software tool, called GEVALT, that streamlines the application of GERBIL and STAMPA. A description of GEVALT (version 1.1) was published as an article in *BMC Bioinformatics* [10].

Section 4.1 briefly describes GEVALT 1.1 and the full published paper appears as Appendix A. Section 4.2 presents additional features that were developed after the publication of the paper and were implemented in the next version of GEVALT (version 2.0).

We would like to thank Daniela Amann and Dr. Edna Ben-Asher from the Weizmann Institute of Science for helpful discussions and comments in the first stages of the project.

4.1 Summary of GEVALT 1.1

GEVALT is an integrated software providing easy access to the GERBIL and STAMPA algorithms as well as to some other tools for genotype analysis. It is based on Haploview [5] 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.

The input of GEVALT is genotype data that can be loaded as either unphased

or phased genotypes. Genotype data dumps from the HapMap website [63] can also be loaded, as well as additional information such as disease status and marker positions. The data can consist of unrelated individuals and two-generation pedigrees. Upon loading a dataset, GEVALT first phases the genotypes (if necessary) by using GERBIL. After phasing is completed, GEVALT generates several displays and option menus, and each of these is shown on a separate tab, allowing the user to move from one to the other. The phased genotypes of each individual are displayed with different colors to indicate alleles phased by GERBIL, missing data and Mendelian errors.

The STAMPA option menu enables the user to run STAMPA, select the desired number of tag SNPs according to the expected prediction accuracy, and also calculate the prediction accuracy of a custom set of tags. The association tab, taken from Haploview, displays association scores and p-values (χ^2 test for unrelated individuals and TDT for trios) for each marker and each haplotype in a block.

GEVALT also allows performing permutation testing in order to obtain a multiple testing corrected p-value. The permutation test in GEVALT was implemented in C++ which makes it about 20 times faster than the JAVA implementation in Haploview. All other Haploview tabs are also available in GEVALT, including an LD plot showing the LD between each pair of SNPs, and a tab showing common haplotypes and their frequencies. The LD plot can be calculated using either the phased or the unphased genotypes. In addition, GEVALT calculates summary statistics for each individual, including the percentage of missing genotypes, the percentage of heterozygous markers, the percentage of minor alleles and a tally of Mendelian inheritance errors.

GEVALT was implemented in JAVA based on the open source code of Haploview version 3.2. The analysis algorithms (GERBIL, STAMPA and permutation testing) were implemented in C++. Both Linux and Windows versions of GEVALT are available and can be downloaded from http://acgt.cs.tau.ac.il/gevalt

4.2 Improvements in version 2.0

A few features were developed and added to GEVALT in version 2.0. Because GERBIL is prohibitively slow when phasing more than 300 SNPs, GEVALT 1.1

was limited to handling only 300 SNPs or less. In order to enable the phasing of more SNPs we developed a long phasing procedure, described in Section 4.2.1. Section 4.2.2 describes an improvement to STAMPA that allows user defined SNPs to be included or excluded from the solution. Finally, an option to use the RAT software for fast permutation testing [28] was added to GEVALT in addition to the standard permutation testing. All features were coded by Oded Apel.

4.2.1 Long Phasing

We implemented a long phasing procedure that breaks the data into smaller regions, uses GERBIL to phase each region separately, and ligates the resulting haplotypes. The program has two parameters: b and l, the break factor and the maximal linker size, respectively. First the data is divided into contiguous regions of b SNPs each (except perhaps the last region). Then each region is phased by GERBIL. The haplotypes of every individual are ligated by using *linkers*. A linker is a region that covers the end of one region and the beginning of the region following it. The linkers are also phased by GERBIL and the resulting haplotypes are used for the ligation of haplotypes from consecutive regions. See **Figure 4.1** for a detailed example.

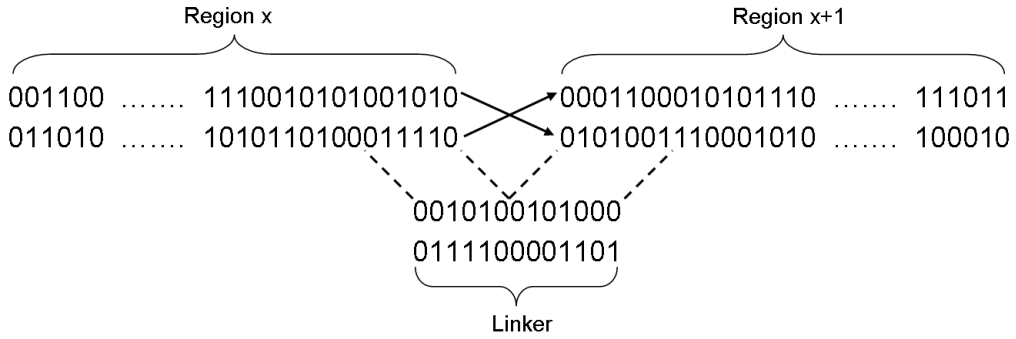


Figure 4.1: An example of the ligation process. The haplotypes of one individual in two consecutive regions are shown. The linker region is six SNPs long to its left side and seven SNPs long to its right. The black arrows indicate the selected ligation solution. This solution results in only one disagreement (last SNP of the linker), while the alternative solution produces two disagreements.

The ligation of haplotypes of an individual from consecutive regions requires

that its linker contains heterozygous sites from both sides of the linker (i.e. from both regions it ligates). Therefore, the length of each linker is set to the minimal length that contains at least two heterozygous sites on each side for every individual. The length on each side is limited to l in order to avoid very long linkers.

We tested the long phasing program on haplotypes from the HapMap resource. We used datasets consisting of 300 SNPs of 60 individuals in order to compare phasing results with and without the long phasing procedure. The long phasing program (with parameters $b = 100$ and $l = 50$) gave practically identical results to GERBIL. The running time of the long phasing program is dramatically influenced by the break factor parameter. For example, phasing 1000 SNPs of 1000 individuals with parameters $b = 100$ and $l = 50$ takes less than three hours on a stand alone PC, and more than eight hours with $b = 200$.

4.2.2 STAMPA improvements

The ease and quality of genotyping varies from SNP to SNP. Therefore, a useful utility for tag SNP selection programs is to allow the user to specify SNPs to be included or excluded from the solution. In order to allow it in STAMPA we modified its dynamic programming algorithm.

Recall from Section 2.4.2 that our objective was to minimize X_T where $X_T = |\{(i,j)|g_{i,j} \neq \text{Predict}(j, j_1, j_2, g_{i,j_1}, g_{i,j_2}))\}|$. We will first describe the original dynamic programming equations and then present the changes made. Let $X_T^{i,j} = 1$ if $g_{i,j} \neq \text{Predict}(j, j_1, j_2, g_{i,j_1}, g_{i,j_2})$ and let $X_T^{i,j} = 0$ otherwise. Then $X_T = \sum_{i,j} X_T^{i,j}$. For every pair of SNPs $m_1 < m_2$ we next define three auxiliary score functions. We assume that $m_1, m_2 \in T$ and that for each j, $m_1 < j < m_2$, $j \notin T$. Then, we define

$$\text{score}(m_1, m_2) = \sum_{i=1}^{n} \sum_{j=m_1+1}^{m_2-1} X_T^{i,j}$$

$\text{score}(m_1, m_2)$ is the total number of prediction errors in SNPs $m_1+1, \ldots, m_2-1$, given that m_1 and m_2 are tag SNPs, and that there are no tag SNPs between m_1 and m_2.

For scoring SNPs at the very end of the list we define a $\text{score}_1(m_1, m_2)$, where we assume that m_1 and m_2 are the two rightmost tag SNPs. Then, the score is

defined as

$$\text{score}_1(m_1, m_2) = \sum_{i=1}^{n} \sum_{j=m_1+1}^{m} X_T^{i,j}$$

Similarly, for $\text{score}_2(m_1, m_2)$ we assume that m_1 and m_2 are the two leftmost tag SNPs, and sum over SNPs $j = 1, \ldots, m_2 - 1$.

The function $f(m^*, l)$, which will be used in the dynamic programming, is defined for $l \geq 2$ and $1 \leq m^* \leq m$. For $l < t$, the function $f(m^*, l)$ represents the minimum number of prediction errors in SNPs $1, 2, \ldots, m^*$, given that the l-th tag SNP is in position m^*. For $l = t$, the function $f(m^*, l)$ represents the minimum number of prediction errors in all SNPs given that the last tag SNP is in position m^*. Formally, we define $f(m^*, l)$ in the following way:

- For $l = t$, $f(m^*, l) = \sum_{i=1}^{n} \sum_{j=1}^{m} X_T^{i,j}$ when the last tag SNP is in position m^*.
- For $2 \leq l < t$, $f(m^*, l) = \sum_{i=1}^{n} \sum_{j=1}^{m^*-1} X_T^{i,j}$ when the l-th tag SNP is in position m^*.

The dynamic programming recurrence formula (assuming $t > 2$) is as follows:

$$f(m^*, l) = \begin{cases} \min_{1 \leq m' < m^*} \text{score}_2(m', m^*) & l = 2 \\ \min_{l-1 \leq m' < m^*} \{f(m', l-1) + \text{score}(m', m^*)\} & 2 < l < t \\ \min_{t-1 \leq m' < m^*} \{f(m', t-1) + \text{score}_1(m', m^*)\} & l = t \end{cases}$$

Now suppose some tag SNPs are dictated by the user and must be used, others must be excluded, and the total number of tag SNPs is t. Excluding specific SNPs from the solution is straightforward by simply ignoring these positions during the recurrence. Including specific SNPs (termed *forced* SNPs) requires some changes in the dynamic programming algorithm. First, the definition of $f(m^*, l)$ slightly changes. For $l < t$, the function $f(m^*, l)$ represents the minimum number of prediction errors in SNPs $1, 2, \ldots, m^*$, given that the l-th tag SNP is in position m^* and all forced SNPs smaller than m^* were selected as tags. For $l = t$, the function $f(m^*, l)$ represents the minimum number of prediction errors in all SNPs given that the last tag SNP is in position m^* and all forced SNPs were selected as tags.

Let $s(m^*)$ be the closest forced SNP smaller than m^*. Formally, $s(m^*) = \max(i | i \text{ is a forced SNP and } i < m^*)$. If no such i exists then $s(m^*) = 1$. $s(m^*)$ can be easily calculated for every $1 \leq m^* \leq m$ as a preprocess step.

The recurrence formula now changes as follows. For $l = 2$:

$$f(m^*, l) = \begin{cases} \infty & k(m^*) > 1 \\ \text{score}_2(s(m^*), m^*) & k(m^*) = 1 \\ \min_{1 \le m' < m^*} \text{score}_2(m', m^*) & k(m^*) = 0 \end{cases}$$

where $k(m^*)$ is the number of forced SNPs smaller than m^*.

For $l > 2$ we have:

$$f(m^*, l) = \begin{cases} \min_{\delta(m^*, l) \le m' < m^*} \{f(m', l-1) + \text{score}(m', m^*)\} & 2 < l < t \\ \min_{\delta(m^*, t) \le m' < m^*} \{f(m', t-1) + \text{score}_1(m', m^*)\} & l = t \end{cases}$$

where $\delta(m^*, l) = \max(s(m^*), l - 1)$. Thus, the search for the optimal neighbor tag of m^* is limited by the closest forced SNP if such one exists.

Since the use of dynamic programming guarantees an optimal solution, forced excluded/included tags can only reduce the prediction accuracy of the selected tags. In our experience, even increasing the total number of tags by adding extra forced tags usually reduces the prediction accuracy. This is because of the nature of the prediction algorithm that predicts a SNP by only considering its two closest tags. Therefore it is recommended to include or exclude tags only when truly needed because of external considerations.

Chapter 5

Crohn's Disease

Crohn's disease (CD) is a chronic inflammatory disease of the gastrointestinal (GI) tract associated with dysregulation of the immune response. It can affect any area of the GI tract, from the mouth to the anus, but it most commonly affects the lower part of the small intestine, called the ileum, and the large intestine (colon). See **Figure 5.1**. Several theories exist about what causes CD, but none have been proven. It is thought to be caused by a combination of environmental and genetic factors. CD can occur in people of all age groups, but it is more often diagnosed in people between the ages of 20 and 30. There are differences in CD prevalence among different populations. For example, people of Jewish heritage have an increased risk of developing CD, while African Americans are at decreased risk.

The NOD2/CARD15 gene on chromosome 16 was the first susceptibility gene identified in patients with CD [24, 38, 22]. The NOD2/CARD15 protein (Nucleotide-binding Oligomerization Domain containing 2 / CAspase Recruitment Domain family, member 15) is an intracellular pattern recognition receptor, which recognizes molecules containing the specific structure called muramyl dipeptide (MDP) that is found in certain bacteria, and thus acts as part of the immune mechanism against foreign bacteria. Three well described mutations, causing loss of function, were found to be associated with the disease. Carriers of these mutations have a tendency to be affected mainly in the ileum. Currently, 13 associated loci are listed in OMIM [65].

In the next section we describe three studies performed in collaboration with gastroenterologists from Wolfson and Rambam medical centers. Some of the

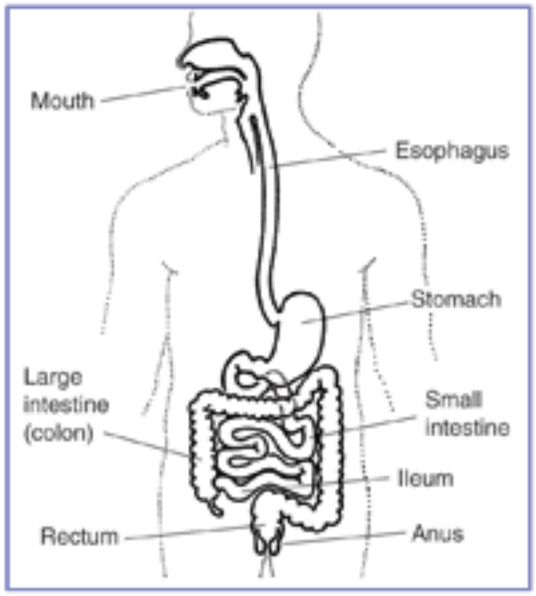


Figure 5.1: The digestive system. Crohn's disease most commonly affects the ileum and/or the colon.

common statistical methods used in these papers are described in Section 5.2.

5.1 Summary of Articles

1. **Pediatric Onset Crohn's Colitis is Characterized by Genotype Dependent Age–Related Susceptibility**
Arie Levine, Subra Kugathasan, Vito Annese, Vincent Biank, Esther Leshinsky–Silver, Ofir Davidovich ,Gad Kimmel, Ron Shamir , Palmieri Orazio, Amir Karban, Ulrich Broeckel, Salvatore Cucchiara
Published in *Inflammatory Bowel Diseases* [32].

Crohn's disease patients can be classified into four categories according to the location of the disease: ileum (denoted L1), colon (L2), both ileum and colon (L3), and upper gastro-intestinal (L4). This classification is called the Vienna classification [19].

Pediatric onset CD is associated with more colitis (L2) in comparison to adult onset CD. Differences in disease site by age may suggest a different genotype, or different host responses such as decreased ileal susceptibility or increased susceptibility of the colon. In this paper we evaluated 721 pediatric onset CD patients

from three cohorts (from USA, Israel and Italy). The patients were evaluated for interactions between age of onset, NOD2/CARD15 genotype and disease location.

NOD2/CARD15 mutations were highly associated with ileal disease location, as expected. We found no association between NOD2/CARD15 mutations and age of onset (AOO). In order to correctly test for association between disease location and AOO we had to neutralize the strong effect of NOD2/CARD15 mutations on disease location. We subdivided the cohorts into patients carrying a NOD2/CARD15 mutation and those with wild type NOD2/CARD15. In the wild type NOD2/CARD15 cohort, we found a high tendency for isolated colitis (L2) in first decade patients compared with second decade patients. In first decade patients, the rate of isolated colitis is higher as AOO decreases. Among the carriers of NOD2/CARD15 mutations, a similar trend was observed, but for colonic involvement (L2+L3) instead of isolated colitis. See **Figure 5.2**.

Together, these findings demonstrate that pediatric onset CD may be characterized by different genes that predispose to early onset and isolated colitis. Recently, following our work, a new locus was found to be associated with isolated colitis and early age of onset (Arie Levine, personal communication).

2. **Risk Factors for Perianal Crohn's Disease: The Role of Genotype, Phenotype, and Ethnicity**
Amir Karban, Maza Itay, Ofir Davidovich, Esther Leshinsky-Silver, Gad Kimmel, Herma Fidder, Ron Shamir, Matti Waterman, Rami Eliakim, and Arie Levine
Published in *The American Journal of Gastroenterology* [25].

As mentioned before, Crohn's disease can affect any part of the gastrointestinal tract from mouth to anus. Perianal Crohn's disease (PD) occurs when Crohn's disease affects the anus, and its surrounding areas, such as the rectum, vagina and skin. PD is a frequent complication of Crohn's disease. Recent studies have found genetic and clinical distinctions between cases with or without PD. These observations may suggest that perianal CD is a distinct phenotype with possible specific susceptibility genes or environmental factors. This study was undertaken to evaluate the role of genotype, clinical, and demographic characteristics with PD in an Israeli cohort.

Phenotypic data (e.g. disease location, smoking, rectal involvement) on 121 CD patients with PD and 179 patients without PD were carefully character-

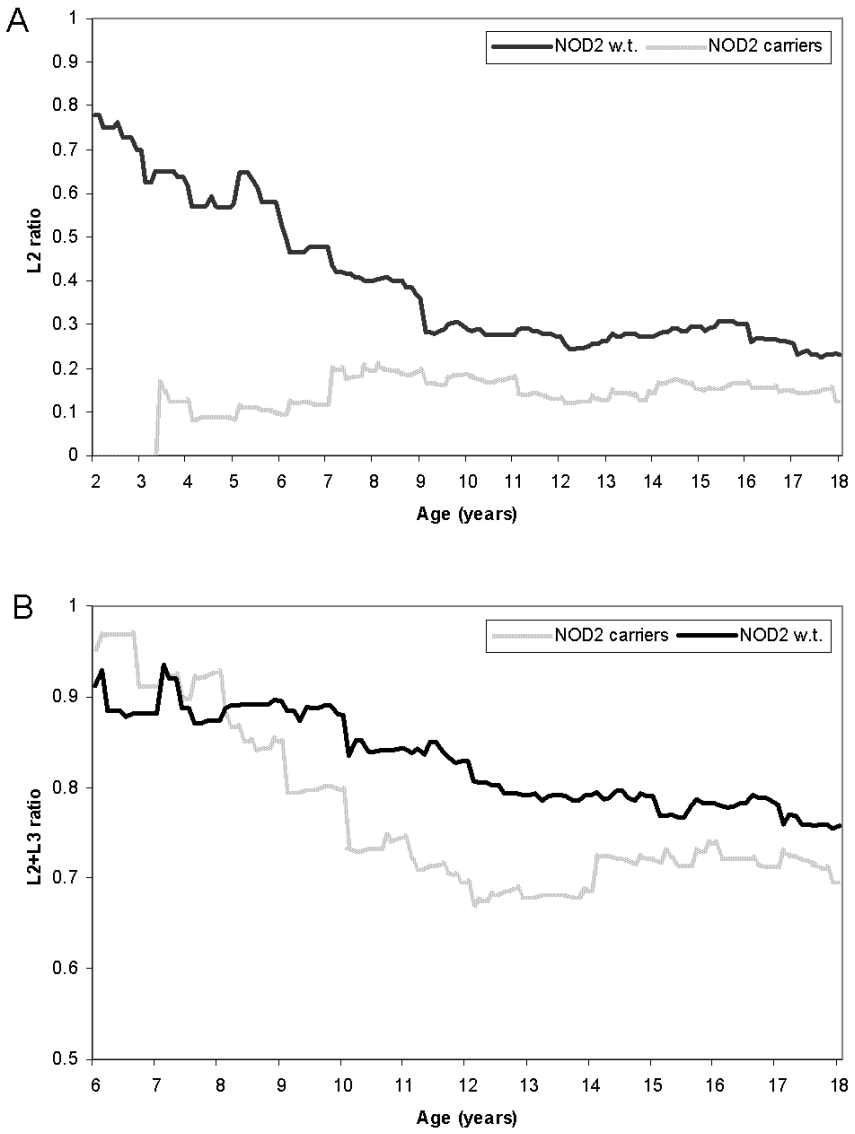


Figure 5.2: Colitis ratio and colonic involvement ratio vs. AOO in patients with and without NOD2/CARD15 mutations.
X-axis - AOO; Y-axis: ratio for patients with AOO within +/- 2 years of the X value. (A) Isolated colitis (L2) ratio. (B) Colonic involvement (L2+L3) ratio. Windows with median age below 6 were excluded since they contained less than 30 patients. Source: [32].

ized. The patients were also genotyped for disease-associated OCTN1/2 and NOD2/CARD15 variants and the TNF-α promoter polymorphisms. Analysis was performed to evaluate the differences in phenotype and genotype frequencies between the PD group and the non-PD group.

PD was found to be associated with rectal involvement and with Sephardic (non-Ashkenazi) ethnicity (the two associations were independent). **Table 5.1** presents all phenotypes tested and their association with PD. No association was found among the studied OCTN, NOD2, TNF-α variants and the risk for PD. The association of ethnicity with PD may suggest that there are as yet unknown genetic variants that are associated with PD.

	All CD (N = 300)	Perianal CD (N = 121)	Nonperianal CD (N = 179)	*P* Value
Male patients	167 (55.6%)	70 (57.8%)	97 (54.2%)	0.26
Smoking: currently	68/296 (23%)	32/119 (26.9%)	36/177 (20.3%)	0.44
Smoking: past	28/296 (9.5%)	11/119 (9.2%)	17/177 (9.6%)	0.73*
Age at diagnosis (yr)	23.48 ± 12.2	22.12 ± 9.5	24.35 ± 13.5	0.17
Family Hx (first degree)	55/297 (18.5%)	22/118 (18.6%)	33/179 (18.4%)	0.51
Duration of disease	9.2 ± 6.7	9.95 ± 6.7	8.73 ± 6.7	0.33
Urban	226/262 (86.2%)	70/84 (83.3%)	156/178 (87.6%)	0.35
Ashkenazi Jews†	144/296 (48.6%)	49/119 (41.2%)	95/177 (53.7%)	0.0457
Sephardic Jews†	98/296 (33.1%)	46/119 (38.7%)	52/177 (29.4%)	
Extent: terminal ileum	104/300 (34.7%)	37/121 (30.6%)	67/179 (37.4%)	0.17
Extent: colon only	40/300 (13.3%)	21/121 (17.4%)	19/179 (10.6%)	0.41
Extent: ileocolon	126/300 (42%)	53/121 (43.8%)	73/179 (40.8%)	0.15
Extent: upper GI	30/300 (10%)	10/121 (8.2%)	20/179 (11.2%)	0.41
Extraintestinal disease	77/300 (25.6%)	33/121 (27.3%)	44/179 (24.6%)	0.97
Rectal involvement	71/284 (25%)	37/106 (34.9%)	34/178 (19.1%)	0.01

*Patients were divided into three smoking categories: never, currently, and past.
†Fifty-four patients have mixed Ashkenazi–Sephardic ethnicity.

Table 5.1: Demographic and clinical characteristics in CD patients with or without perianal disease. Source: [25].

3. **Lack of association of the 3'-UTR polymorphism in the NFKBIA gene with Crohn's disease in an Israeli cohort**
Esther Leshinsky-Silver, Amir Karban, Sara Cohen, Marcelo Fridlander, Ofir Davidovich, Gad Kimmel, Ron Shamir and Arie Levine
Published in *International Journal of Colorectal Disease* [31].

Patients with CD have a TH1-type inflammatory response characterized by nuclear factor kappa B (NFκB) activation. Mutations in the bacterial pattern recognition receptors NOD2/CARD15 and Toll-like receptor 4 (TLR4) genes, which lead to activation of NFκB under normal circumstances, have been associated with increased susceptibility for CD. NFκB plays a critical role in the immune response and is down-regulated by NFκB inhibitor α (NFKBIA). Re-

cently, NFKBIA was found to be a susceptibility gene for German CD patients lacking NOD2/CARD15 mutations.

In this study a cohort of 231 Israeli CD patients, previously genotyped for the SNPs in the CARD15 and TLR4 susceptibility genes for CD, was analyzed for the 3'-untranslated region (UTR) SNP of the NFKBIA gene. The same SNP was genotyped in a group of 100 healthy ethnically matched controls. We searched for association between the NFKBIA SNP and CD, as well as with several phenotypes: age of onset, disease location, and disease behavior. We also evaluated the association between interactions of SNPs (NFKBIA, NOD2/CARD15, and TLR4) and the phenotypes. Finally, the same associations were checked in a subset of the cohort lacking NOD2/CARD15 mutations. We did not identify a significant difference in allele and genotype frequencies between either groups or an effect on phenotype.

5.2 General Methods

In all three studies some common methods for statistical analysis were used with some necessary ad-hoc adjustments. Below is a short summary of the methods. See the Methods sections of the papers for a more detailed explanation.

Haplotype inference of NOD2/CARD15 SNPs Since all three NOD2/CARD15 mutations are known to be deleterious we did not consider each one separately, but created a new artificial locus called "Any NOD" as follows. The NOD genotypes were phased into haplotypes using the software GERBIL [27]. Any haplotype that contained at least one NOD mutation was considered as an "Any NOD" mutation, and a haplotype without mutations was considered as an "Any NOD" wild type allele. This method was applied in papers 1 and 3.

Association analysis The association between SNPs and a phenotype was evaluated as follows. We first calculated a score for each SNP. For a discrete trait (e.g., case/control status, disease location) we used the Pearson χ^2 score as the test statistic. The contingency table for the Pearson χ^2 score was built based on allele counting (2×2 table). This statistic assumes the multiplicative model for penetrance (see Section 2.5). For a continuous trait (e.g., AOO) both ANOVA and Kolmogorov-Smirnov (KS) scores were used. We used the permutation test [69] described in Section 2.5 in order to derive a multiple testing

corrected p-value for the highest scoring SNP. The permutation test was used even when only one SNP was tested. Although it is possible to derive a p-value directly from the statistics (e.g. using the χ^2–test for the χ^2 statistic, and the F–test for the ANOVA statistic) using the permutation test has the advantage of creating a null model free from any assumptions (unlike F–test, for example, that assumes a normal distribution of the trait).

Interactions In addition to association of single SNPs, one often wishes to test interactions, i.e., pairs of SNPs that together have association with the phenotype. The permutation test can be readily generalized to handle this case. For a specific pair of SNPs, a 4×2 contingency table is built by counting haplotypes of the pair, and the Pearson χ^2 score is calculated as the test statistic. When haplotypes are unknown (SNPs are far away or on different chromosomes) then for double heterozygotes, a value of 0.5 is added in each of the four cells in the relevant column of the table. This statistic is calculated for every pair of SNPs and multiple testing corrected p-value for the highest scoring pair is calculated with the permutation test. This method was applied in paper 3.

Population correction A potential contributor to a false-positive association is confounding effect due to differences in genetic background, termed *population stratification*. Population stratification refers to differences in allele frequencies between cases and controls due to differences in ancestry between the two groups rather than association of SNPs with disease. Since each study contained two or more different population groups, all permutation tests were corrected as follows: The association score was calculated for each population separately, and the test statistic was defined to be the weighted average of these scores, where the weight of each score is the fraction of its corresponding population. In calculating the p-value, permutations were generated by randomly permuting the labels within each population independently. In all studies, Israeli patients were divided into populations according to ethnicity – Ashkenazim and Sephardim. Patients with mixed or unknown ethnicity were excluded from the cohort when applying the population correction.

Power calculations The power of an association study is defined as the probability to correctly conclude that there is association when one of the SNPs is truly associated with the disease. It is especially important to evaluate the power of a study when no associations are found. We calculated the power of the studies by simulations. The power of an association test between a single SNP

and a disease was calculated as follows. Several differences in allele frequency between cases and controls were tested (e.g. 5, 10, 15, and 20%). For each of these differences, the following test was performed: 1,000 data sets of cases and controls (with the same number of case and control individuals as in the real study) were simulated, where alleles were drawn according to the defined allele frequency. The permutation test described above was applied to each of these data sets. The power of the test was calculated as the fraction of the data sets that received a p-value lower than 0.05. Since no association between SNPs and the tested phenotype were detected in papers 2 and 3, we applied this method in order to evaluate the power of the tests. Similar tests can evaluate the power of association when interactions or haplotypes are considered.

List of abbreviations

ANOVA (ANalysis Of Variance) – a statistical procedure used to test hypotheses concerning means of several populations.
CD (Crohn's Disease) – a chronic inflammatory disease of the gastrointestinal tract associated with dysregulation of the immune response.
GERBIL (GEnotype Resolution and Block Identification using Likelihood) – a software for simultaneously phasing genotypes into haplotypes and block partitioning. [27]
GEVALT (GEnotype Visualization and ALgorithmic Tool) – an integrated software tool for genotype analysis. GEVALT combines the visual (and algorithmic) abilities of Haploview with the STAMPA, GERBIL, and RAT algorithms. [10]
HAPGEN – a program that simulates case control datasets from a given set of haplotypes. [36]
Haploview – a very popular software providing a common interface to several tasks relating to haplotype analysis. [5]
HapMap Project – a multi-country effort to identify and catalog genetic similarities and differences in human beings. Contains millions of SNPs of 270 individuals from four different origins. [59]
KS-test (Kolmogorov-Smirnov test) – used to determine whether two probability distributions differ.
LD (Linkage Disequilibrium) – the non-random association of alleles at two or more loci in the population.
PD (Perianal Crohn's Disease) – a variant of Crohn's disease affecting the anus, and its surrounding areas. PD is a frequent complication of Crohn's disease.
RAT (Rapid Association Test) – a software for testing association between SNPs and a disease. [28]
SNP (Single Nucleotide Polymorphism) – a single nucleotide position in the genome that differs across members of a species.
STAMPA (Selection of TAg SNPs to Maximize Prediction Accuracy) – a software for tag SNP selection based on the prediction accuracy criterion. [21]
Tagger – a popular software for tag SNP selection based on the r^2 criterion. Implemented in Haploview. [12]

Bibliography

[1] H. I. Avi-Itzhak, X. Su and F. M. D. L. Vega. Selection of minimum subsets of single nucleotide polymorphisms to capture haplotype block diversity. In *Proceedings of Pacific Symposium on Biocomputing (PSB 03)*. volume 8, 466–477 (2003).

[2] V. Bafna, D. Gusfield, G. Lancia and S. Yooseph. Haplotyping as perfect phylogeny: A direct approach. *J. Comput. Biol.* **10(3-4)**, 323–340 (2003).

[3] V. Bafna, B. V. Halldorsson, R. Schwartz, A. Clark and S. Istrail. Haplotypes and informative SNP selection algorithms: Don't block out information. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 03)* (The Association for Computing Machinery), 19–27 (2003).

[4] D. J. Balding, M. Bishop and C. Cannings. *Handbook of Statistical Genetics, Second Edition* (John Wiley and Sons, inc.) (2003).

[5] J. C. Barrett, B. Fry, J. Maller and M. J. Daly. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**(2), 263–265 (2005).

[6] D. Botstein, R. White, M. Skolnick and R. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**, 314–331 (1980).

[7] C. S. Carlson, M. A. Eberle, M. Rieder, Q. Yi, L. Kruglyak and D. A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium. *Am. J. Hum. Genet.* **74**(1), 106–120 (2004).

[8] K. Choudhury, A. McQuillin, V. Puri, J. Pimm, S. Datta, S. Thirumalai, R. Krasucki, J. Lawrence, N. J. Bass, D. Quested, C. Crombie, G. Fraser, N. Walker, H. Nadeem, S. Johnson, D. Curtis, D. St. Clair and H. M. D. Gurling. A Genetic Association Study of Chromosome 11q22−24 in Two Different Samples Implicates the FXYD6 Gene, Encoding Phosphohippolin, in Susceptibility to Schizophrenia. *Am J Hum Genet* **80**(4), 664–672 (2007).

[9] A. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7(2)**, 111–122 (1990).

[10] O. Davidovich, G. Kimmel and R. Shamir. Gevalt: An integrated software tool for genotype analysis. *BMC Bioinformatics* **8:36** (2007).

[11] P. I. W. de Bakker, N. P. Burtt, R. R. Graham, C. Guiducci, R. Yelensky, J. A. Drake, T. Bersaglieri, K. L. Penney, J. Butler, S. Young, R. C. Onofrio, H. N. Lyon, D. O. Stram, C. A. Haiman, M. L. Freedman, X. Zhu, R. Cooper, L. Groop, L. N. Kolonel, B. E. Henderson, M. J. Daly, J. N. Hirschhorn and D. Altshuler. Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* **38**, 1298–1303 (2006).

[12] P. I. W. de Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly and D. Altshuler. Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223 (2005).

[13] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).

[14] E. Eskin, E. Halperin and R. M. Karp. Large scale reconstruction of haplotypes from genotype data. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 03)* (The Association for Computing Machinery), 104–113 (2003).

[15] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12(5)**, 912–917 (1995).

[16] S. Eyheramendy, J. Marchini, G. McVean, S. Myers and P. Donnelly. A model-based approach to capture genetic variation for future association studies. *Genome Res.* **17**(1), 88–95 (2007).

[17] S. B. Gabriel, S. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly and D. Altshuler. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).

[18] I. Gal, G. Kimmel, R. Gershoni-Baruch, M. Z. Papa, E. Dagan, R. Shamir and E. Friedman. A specific RAD51 haplotype increases breast cancer risk in Jewish non-Ashkenazi high-risk women. *Eur. J. Cancer* **42(8)**, 1129–1134 (2006).

[19] C. Gasche, J. Scholmerich, J. Brynskov, G. D'Haens, S. B. Hanauer, E. J. Irvine, D. P. Jewell, D. Rachmilewitz, D. B. Sachar, W. J. Sandborn and L. R. Sutherland. A simple classification of crohn's disease: Report of the working party for the world congresses of gastroenterology, Vienna 1998. *Inflammatory Bowel Diseases* **6**(1), 8–15 (2000).

[20] G. Greenspan and D. Geiger. Model-based inference of haplotype block variation. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 03)* (The Association for Computing Machinery), 131–137 (2003).

[21] E. Halperin, G. Kimmel and R. Shamir. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics* **21(1)**, 195–203 (2005).

[22] J. Hampe, A. Cuthbert, P. J. P. Croucher, M. M. Mirza, S. Mascheretti, S. Fisher, H. Frenzel, K. King, A. Hasselmeyer, A. J. S. MacPherson, S. Bridger, S. van Deventer, A. Forbes, S. Nikolaus, J. E. Lennard-Jones, U. R. Foelsch, M. Krawczak, C. Lewis, S. Schreiber and C. G. Mathew. Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet* **357**, 1925–1928 (2001).

[23] J. He and A. Zelikovsky. MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression. *Bioinformatics* **22**(20), 2558–2561 (2006).

[24] J. P. Hugot, M. Chamaillard, H. Zouali, S. Lesage, J. P. Cezard, J. Belaiche, S. Almer, C. Tysk, C. A. O'Morain, M. Gassull, V. Binder, Y. Finkel,

A. Cortot, R. Modigliani, P. Laurent-Puig, C. Gower-Rousseau, J. Macry, J. F. Colombel and M. Sahbatou. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).

[25] A. Karban, M. Itay, O. Davidovich, E. Leshinsky-Silver, G. Kimmel, H. Fidder, R. Shamir, M. Waterman, R. Eliakim and A. Levine. Risk factors for perianal Crohn's disease: The role of genotype, phenotype, and ethnicity. *The American Journal of Gastroenterology* **102**(8), 1702–1708 (2007).

[26] G. Kimmel and R. Shamir. 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), 2–9 (2004).

[27] G. Kimmel and R. Shamir. GERBIL: Genotype resolution and block identification using likelihood. *P. Natl. Acad. Sci. USA* **102**, 158–162 (2005).

[28] G. Kimmel and R. Shamir. A fast method for computing high significance disease association in large population-based studies. *Am. J. Hum. Genet.* **79**, 481–492 (2006).

[29] J. F. C. Kingman. The coalescent. *Stochastic Processes and Their Applications* **13**, 235–248 (1982).

[30] M. Koren, G. Kimmel, E. Ben-Asher, I. Gal, M. Z. Papa, J. S. Beckmann, D. Lancet, R. Shamir and E. Friedman. ATM haplotypes and breast cancer risk in Jewish high risk women. *Br. J. Cancer* **94**, 1537–1543 (2006).

[31] E. Leshinsky-Silver, A. Karban, S. Cohen, M. Fridlander, O. Davidovich, G. Kimmel, R. Shamir and A. Levine. Lack of association of the 3'-UTR polymorphism in the NFKBIA gene with Crohn's disease in an Israeli cohort. *International Journal of Colorectal Disease* doi:10.1007/s00384-007-0287-x (2007).

[32] A. Levine, S. Kugathasan, V. Annese, V. Biank, E. Leshinsky-Silver, O. Davidovich, G. Kimmel, R. Shamir, P. Orazio, A. Karban, U. Broeckel and S. Cucchiara. Pediatric onset Crohn's colitis is characterized by genotype dependent age–related susceptibility. *Inflammatory Bowel Diseases* doi: 10.1002/ibd.20244 (2007).

[33] N. Li and M. Stephens. Modelling linkage disequilibrium and identifying recombinations hotspots using SNP data. *Genetics* **165**, 2213–2233 (2003).

[34] J. Long, R. C. Williams and M. Urbanek. An EM algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56(3)**, 799–810 (1995).

[35] J. R. Lupski. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics* **14(10)**, 417–422 (1998).

[36] J. Marchini, B. Howie, S. Myers, G. McVean and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906–913 (2007).

[37] T. Niu, Z. S. Qin, X. Xu and J. S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **70(1)**, 157–169 (2002).

[38] Y. Ogura, D. K. Bonen, N. Inohara, D. L. Nicolae, F. F. Chen, R. Ramos, H. Britton, T. Moran, R. Karaliuskas, R. H. Duerr, J. P. Achkar, S. R. Brant, T. M. Bayless, B. S. Kirschner, S. B. Hanauer, G. Nunez and J. H. Cho. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).

[39] J. M. Olson and E. M. Wijsman. Design and sample size considerations in the detections of linkage disequilibrium with a disease locus. *Am. J. Hum. Genet.* **55**, 574–580 (1994).

[40] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor and D. R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).

[41] I. Pe'er, P. de Bakker, J. Maller, R. Yelensky, D. Altshuler and M. Daly. Evaluating and improving power in whole genome association studies using fixed marker sets. *Nat Genet* **38(6)**, 663–667 (2006).

[42] P. D. P. Pharoah, J. Tyrer, A. M. Dunning, D. F. Easton and B. B. A. J. Ponder. Association between common variation in 120 candidate genes and breast cancer risk. *PLoS Genetics* **3(3)** (2007).

[43] J. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).

[44] S. Purcell, M. J. Daly and P. C. Sham. WHAP: haplotype-based association analysis. *Bioinformatics* **23**(2), 255–256 (2007).

[45] Z. S. Qin, T. Niu and J. S. Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet* **71**, 1242–1247 (2002).

[46] D. C. Queller, J. E. Strassmann and C. R. Hughes. Microsatellites and kinship. *Trends in Ecology and Evolution* **8**, 285–288 (1993).

[47] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **291**, 1298–2302 (2001).

[48] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobson and G. A. Poland. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70**, 425–434 (2002).

[49] P. Scheet and M. Stephens. A fast and flexible statistical model for large−scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**(4), 629–644 (2006).

[50] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg and M. Wigler. Large-Scale Copy Number Polymorphism in the Human Genome. *Science* **305**(5683), 525–528 (2004).

[51] B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* (2007).

[52] R. S. Spielman and W. J. Ewens. The TDT and other family based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* **59**, 983–989 (1996).

[53] R. S. Spielman, R. E. McGinnis and W. J. Ewens. Transmission test for linkage disequilibrium: the Insulin gene region and Insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).

[54] M. Stephens and P. Donnelly. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* **73(6)**, 1162–1169 (2003).

[55] M. Stephens, N. J. Smith and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68(4)**, 978–989 (2001).

[56] M. W. Sun, J. Y. Lee, P. I. de Bakker, N. P. Burtt, P. Almgren, L. Rastam, T. Tuomi, D. Gaudet, M. J. Daly, J. N. Hirschhorn, D. Altshuler, L. Groop and J. C. Florez. Haplotype Structures and Large-Scale Association Testing of the 5' AMP-Activated Protein Kinase Genes PRKAA2, PRKAB1, and PRKAB2 With Type 2 Diabetes. *Diabetes* **55**(3), 849–855 (2006).

[57] A. R. Templeton, E. Boerwinkle and C. F. Sing. A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in drosophila. *Genetics* **117**, 343–351 (1987).

[58] A. R. Templeton, T. Maxwell, D. Posada, J. H. Stengard, E. Boerwinkle and C. F. Sing. Tree Scanning: A Method for Using Haplotype Trees in Phenotype/Genotype Association Studies. *Genetics* **169**(1), 441–453 (2005).

[59] The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).

[60] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

[61] A. Vignal, D. Milan, M. SanCristobal and A. Eggen. A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* **34**, 275–305 (2002).

[62] http://www.affymetrix.com/products/arrays/specific/genome_wide_snp6/genome_wide_snp_6.affx.

[63] http://www.hapmap.org.

[64] http://www.illumina.com/pages.ilmn?ID=209.

[65] Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.

[66] N. Zaitlen, H. M. Kang, E. Eskin and E. Halperin. Leveraging the HapMap Correlation Structure in Association Studies. *Am J Hum Genet* **80**(4), 683–691 (2007).

[67] D. V. Zaykin, P. H. Westfall, S. S. Young, M. A. Karnoub, M. J. Wagner and M. G. Ehm. Testing association of statistically inferred haplotypes with discrete and continues traits in samples of unrelated individuals. *Hum. Hered.* **53**, 79–91 (2002).

[68] K. Zhang, P. Calabrese, M. Nordborg and F. Sun. Haplotype block structure and its applications to association studies power and study designs. *Am. J. Hum. Genet.* **71**, 1386–1394 (2002).

[69] K. Zhang, M. Deng, T. Chen, M. Waterman and F. Sun. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. USA* **99(11)**, 7335–9 (2002).

[70] K. Zhang, Z. Qin, J. Liu, T. Chen, M. S. Waterman and F. Sun. Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res.* **14(5)**, 908–916 (2004).

Appendix A

GEVALT's paper

BMC Bioinformatics

BioMed Central

Software

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.

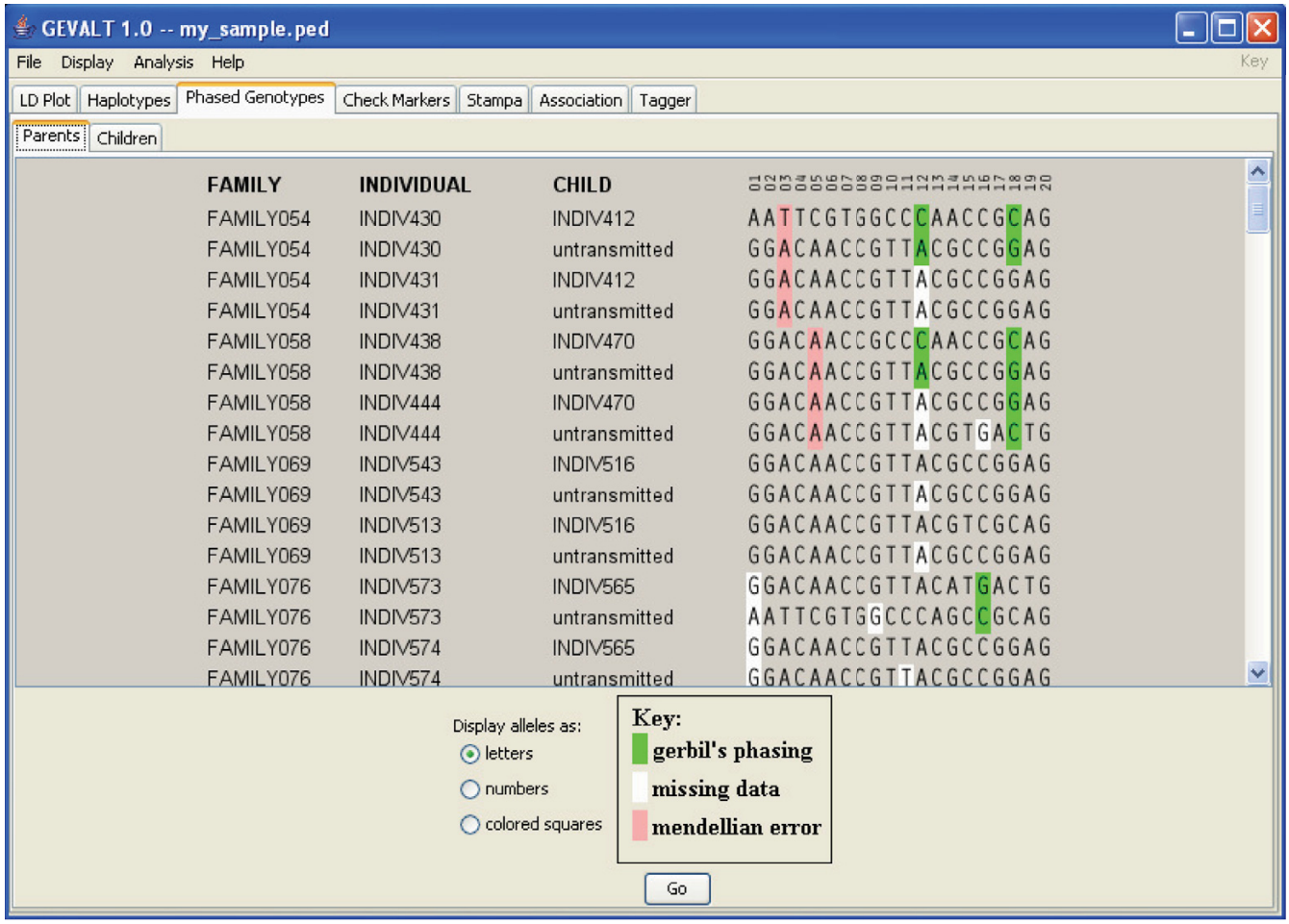


Figure 1
The Phased Genotypes display. The phased genotypes display is divided into two tabs – parents and children. In this image the parents tab is displayed. For each parent, its two chromosomes are displayed, with an indication which of them was transmitted to the child that was included in the trio. 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

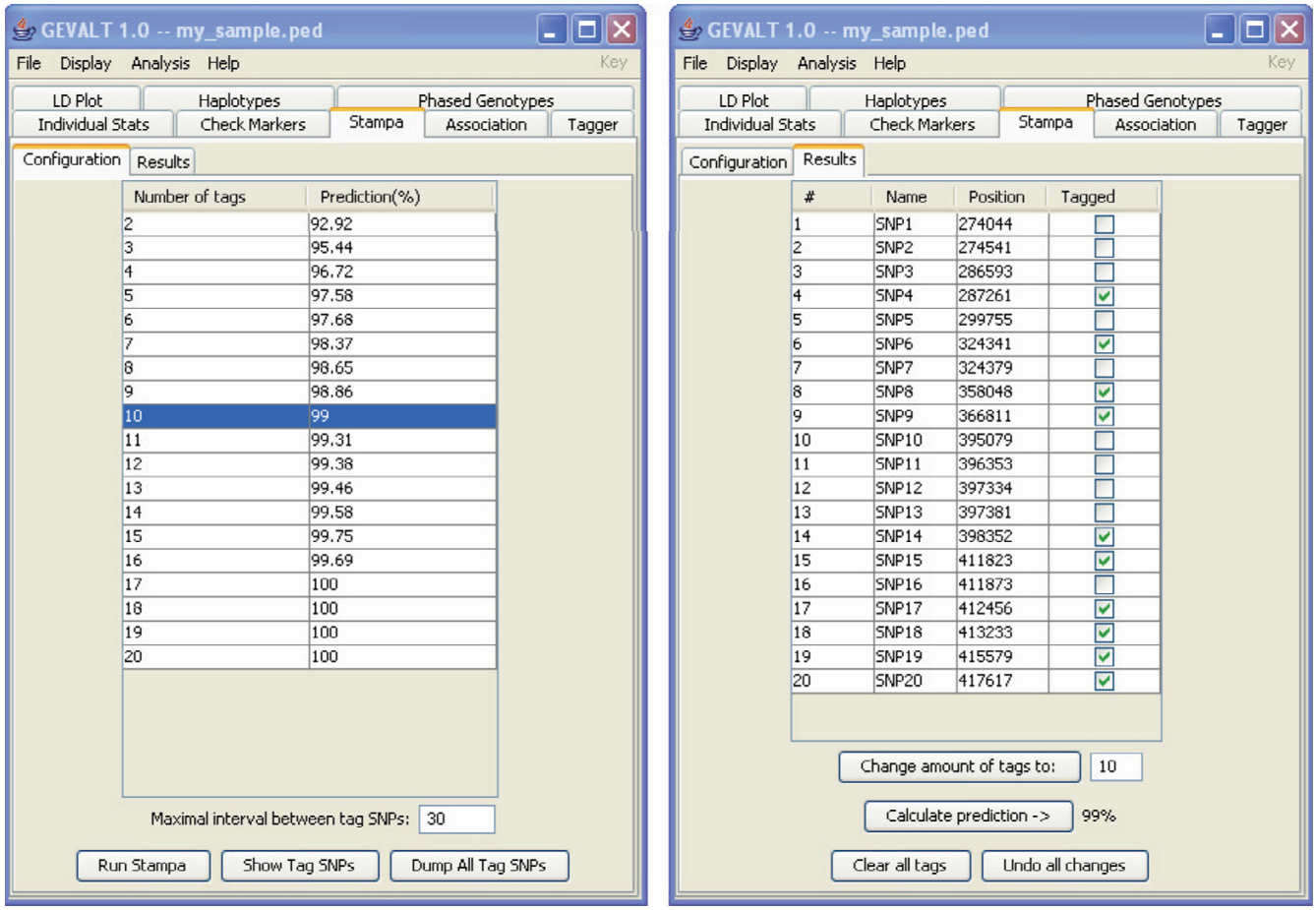
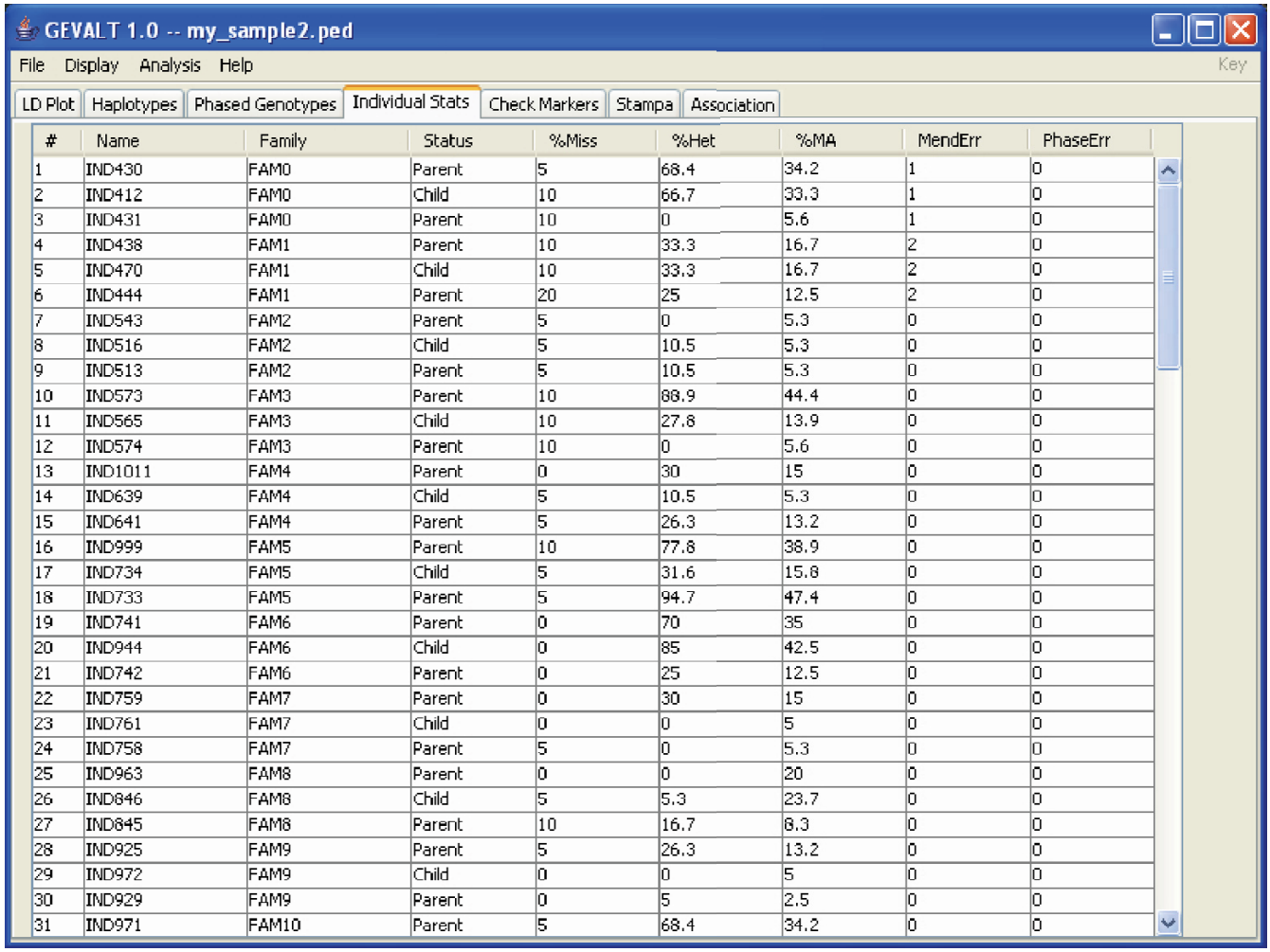


Figure 2
The Stampa display. Left – The Stampa configuration menu. After running Stampa, a table is displayed, showing for every number of tag SNPs, the prediction accuracy of the best set of tag SNPs of that size. The user can choose the number of tags to display according to the required prediction 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.

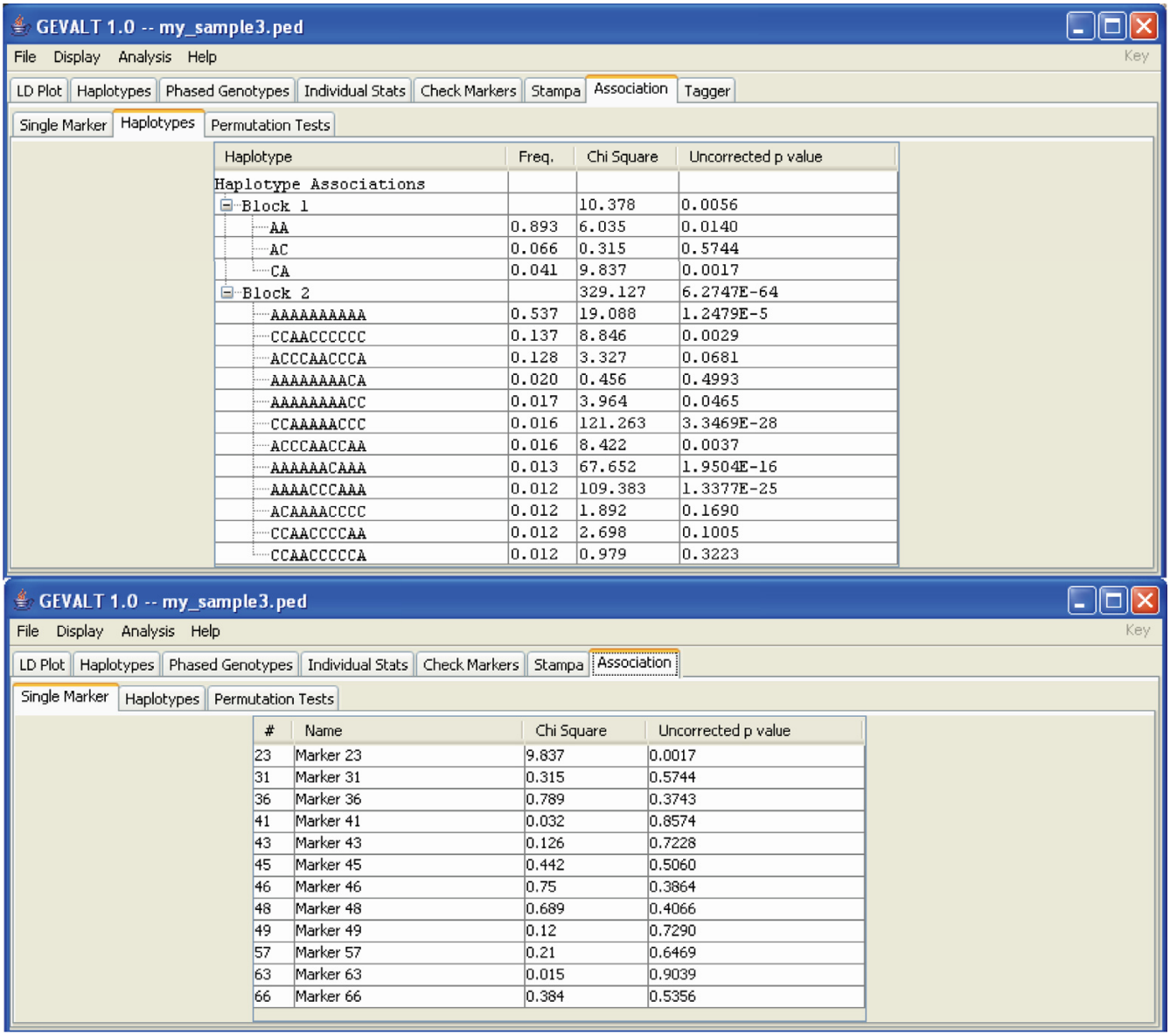


Figure 4

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.

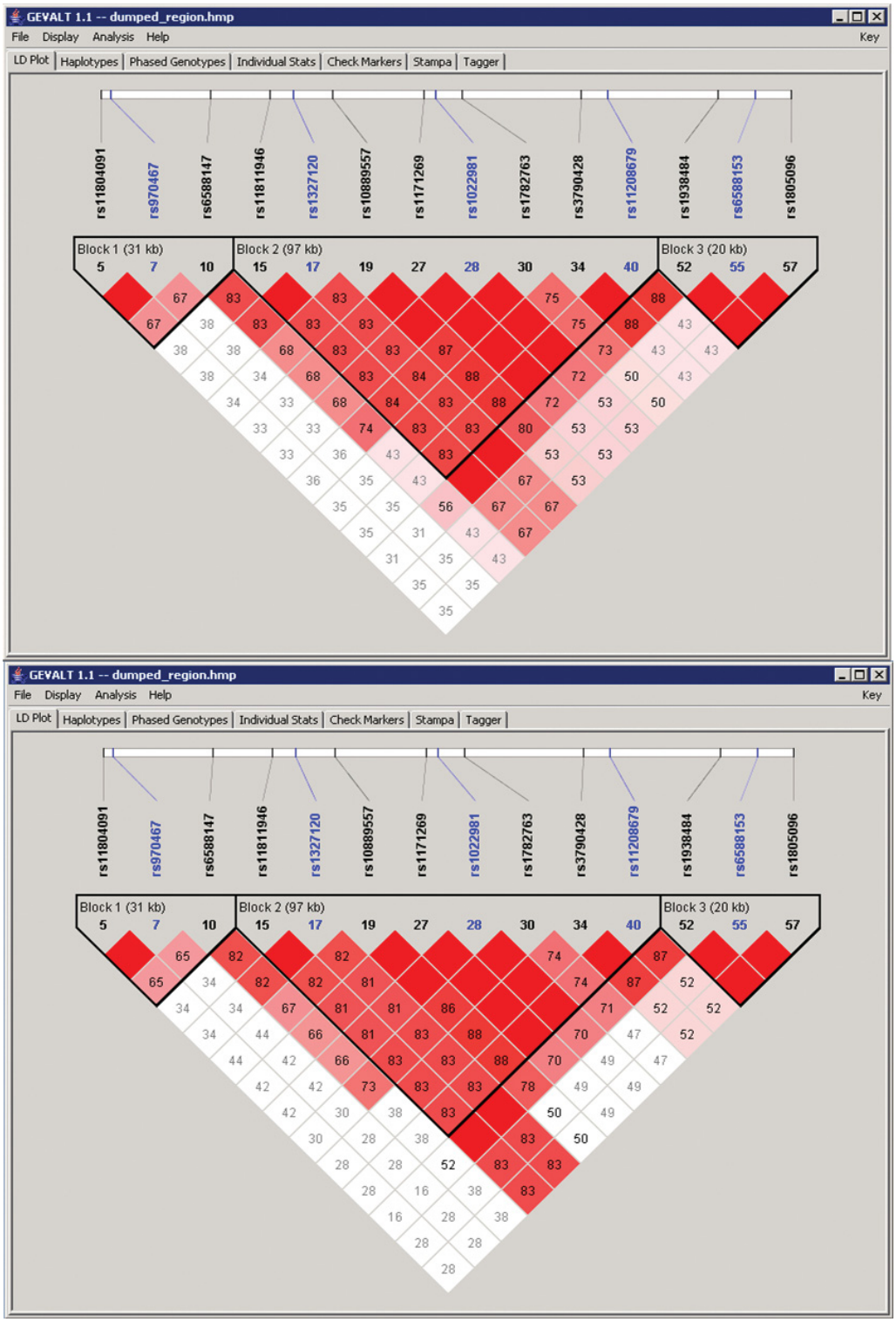


Figure 5

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.