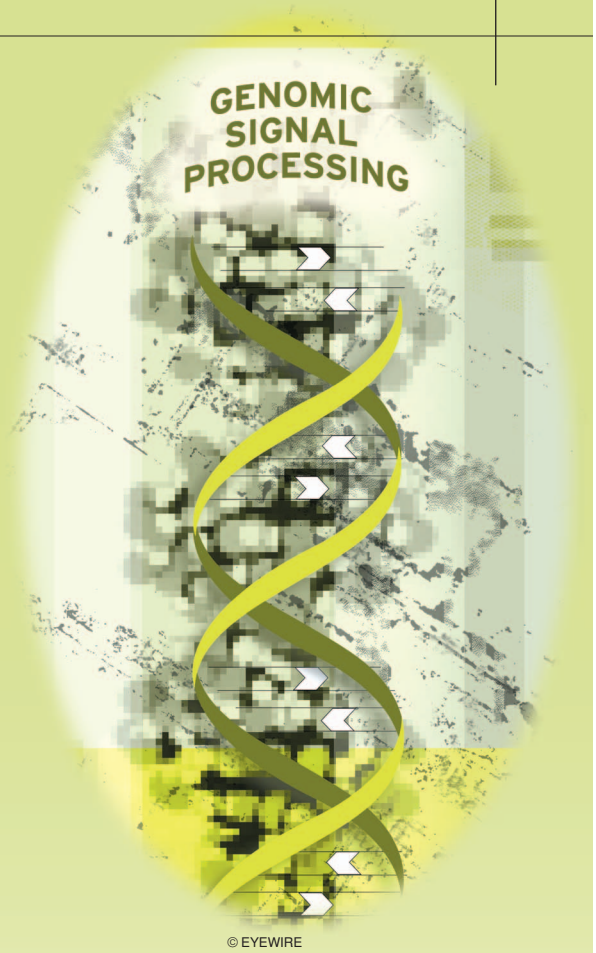
Michel Sarkis, Bernhard Goebel, Zaher Dawy,
Joachim Hagenauer, Pavol Hanus, and Jakob C. Mueller

Gene Mapping of Complex Diseases



A comparison of methods from statistics, information theory, and signal processing

The goal of gene mapping is to identify the genetic loci that are responsible for apparent phenotypes such as complex diseases. With the fast enhancement of genotyping techniques and equipment, the research focus is moving towards population-based studies. These studies usually concentrate on sequences of single nucleotide polymorphisms observed within a population. The variations of these nucleotides are believed to be responsible for the majority of genetic differences among given individuals [1]–[3].

In this work, a review and comparison of three different gene mapping methods is presented along with the assumptions and imposed constraints. The first method relies on standard statistical techniques and follows a probabilistic approach to deal with the problem. The second method is based on a well-known concept from information theory. The idea is to model the available data as random variables and measure the dependence between them using mutual information. The last method explores blind source separation techniques by finding a suitable model that involves mixing various sources so that independent component analysis can be applied.

BIOLOGICAL BACKGROUND

The size of the human genome is about 3 billion base pairs (bp). The DNA of a typical human somatic (nongerm line) cell is arranged in 23 chromosome pairs, whereas one chromosome in each pair is inherited from the maternal side and one from the paternal side. The chromosome pairs 1–22 are common to both sexes and the sex chromosome pair X-X in a female are homologous and undergo recombination. The sex chromosome pair X-Y in a male is nonhomologous and only recombines in a small region.

SINGLE NUCLEOTIDE POLYMORPHISMS

In the world's human population, about 10 million sites (that is, one per 300 bases on average) vary with an observed minor allele frequency of $\geq$1% [4]. Such varying sites in a population are referred to as single nucleotide polymorphisms (SNPs). The terms marker and locus refer to a SNP's position within the genome. An allele is in this context the term for the specific nucleotide observed in an individual on a particular chromosome for a particular SNP locus. It is very unlikely that more than one mutation would have occurred at the same locus during the short human evolution, thus SNPs are usually biallelic, which means that only two different nucleotides (alleles) can be observed at each SNP locus throughout the population [1], [2]. A typical DNA sequence fragment of one individual may look as follows:

```
paternal ...ATGTCCTGCATTGCTAGACTGGGTACT
         GAGAGTCGTGTAC...
maternal ...ATGTCCTGCTTTGCTAGACTGGGTACAGAGAGTC
         GTGTAC...
```

The two lines represent the paternal and maternal chromosome sets. The SNP marker have been highlighted. Assuming biallelic SNPs (e.g., for a SNP with observed alleles A and T), there exist two homozygous (AA, TT) and two heterozygous genotypes (AT, TA) at each marker. However, the heterozygous genotypes, where different allele has been inherited from each parental side, cannot be easily distinguished from each other by standard genotyping techniques. Instead it has to be determined using statistical means and is referred to as gametic phase estimation.

An individual's genome is its genotype, whereas its outward appearance is called phenotype. Connecting phenotype with genotype, or determining the DNA parts that cause a particular trait, is called gene mapping. Based on the assumption that the difference in the observed phenotypes is of genetic origin, it is natural for the ongoing gene mapping research to concentrate primarily on SNPs.

CONNECTING PHENOTYPE WITH GENOTYPE, OR DETERMINING THE DNA PARTS THAT CAUSE A PARTICULAR TRAIT, IS CALLED GENE MAPPING.

As the mutation rate is very low relative to the number of generations since the most recent common ancestor of any two humans, each new allele is initially associated with the other alleles that happened to be present on the particular chromosomal background on which it arose. This specific set of alleles is referred to as a haplotype. The coinheritance of SNP alleles on these haplotypes leads to associations between these alleles in the population (known as linkage disequilibrium, LD) [4]. Because the likelihood of recombination between two SNPs increases with the distance between them, on average such associations between SNPs decline with distance. These correlations make it possible to limit genotyping to only a few carefully chosen SNPs in any particular region and to identify this region even if the causal SNP happened not to be among the investigated SNPs. Due to linkage disequilibrium most of the information about genetic variation represented by the 10 million common SNPs in the population could be provided by genotyping 200,000–1 million tag SNPs across the genome.

COMPLEX DISEASES

The term complex disease (or trait) is used to describe diseases caused by multiple genetic markers, possibly interacting in a complex and unknown manner as opposed to Mendelian traits which are caused by one single major genetic marker (and are therefore also referred to as monogenic diseases). Examples of complex diseases are Parkinson's, schizophrenia, diabetes type 1, and Alzheimer's.

Trait variables are usually divided into discrete traits and quantitative traits. The latter are traits that occur in various grades usually measured on a linear scale. Examples of quantitative traits are body height and blood concentrations. It is sometimes a question of interpretation whether a particular trait is regarded as discrete or quantitative. Hair loss, for instance, may be seen as a binary or dichotomous trait (affected or nonaffected) or as a quantitative one (size of surface area affected by hair loss).

GENE-MAPPING METHODS

Gene-mapping methods are traditionally divided into two categories, pedigree-based and population-based methods [1], [2], [5].

Pedigree-based linkage studies test for cosegregation of disease and nearby marker alleles among affected and nonaffected members of one family and are often referred to as linkage studies. The main advantage is that family members share similar environments and genes; however the number of individuals recruited is usually very small. Thus, pedigree-based methods are mostly limited to coarse-map simple genetic diseases. In linkage analysis, individuals are genotyped at random markers spread across the genome. If a disease gene is close to one of the markers, then, within the pedigree, the inheritance pattern at the marker will mimic the inheritance pattern of the disease itself.

In population-based association studies, the investigated individuals are subdivided into two groups. One group, the cases, is affected by the disease under investigation, while the other, the controls, is not affected. The two groups are checked for genetic differences, which are assumed to appear close to the true causal loci of the disease. Subsequently, an SNP set from this area is genotyped and investigated further for the purpose of fine-scale mapping.

Several high-throughput techniques have been developed for genotyping SNPs in a high number of individuals. The locus specificity is obtained by site specific hybridization of oligonucleotide probes or primers. Sometimes genotyping fails at some loci for a particular individual. In such cases, the missing values can be estimated. Thanks to the possibility of genotyping a large number of individuals, the focus is shifting towards population-based association studies to fine map the complex traits. The actual costs for one SNP could become as low as one cent [5], making genotyping assays designed for testing a patients predisposition to a particular disease affordable. Such assays comprise roughly about thousand SNPs in genome regions known to be associated with the disease.

There are several aspects that need to be taken into account when doing an association study. It is essential that, except for the disease status, the groups of cases and controls should be as homogenous as possible in the sense, that the control sample should reflect the ethnic composition of the case sample. This prevents detecting spurious and masking true associations [6]. Additionally, both groups should preferably have approximately equal size and the sample size should be as large as possible.

STATISTICS-BASED METHODS

Two classes of statistical methods widely used in gene mapping are presented in this section. These techniques test the null hypothesis of no association by evaluating the sample data of a case-control study to find out which SNP or group of

SNPs is associated with a given phenotype. The first class is based on the evaluation of contingency tables while the second one is based on regression analysis between genotype and phenotype [1]–[3], [7].

CONTINGENCY-TABLES-BASED METHODS

The term contingency refers to the relationship between two variables. By setting up a contingency table, statistical analysis can be performed to determine the dependence or independence of the variables under study. In the context of gene mapping, the two random variables are the genotype represented by the SNPs and the phenotype reflected by the condition of the individuals under investigation.

GENE-MAPPING METHODS ARE TRADITIONALLY DIVIDED INTO TWO CATEGORIES, PEDIGREE-BASED AND POPULATION-BASED METHODS.

Here, the individuals' genotypes are sorted into a table, e.g., 2×3 for a case control study investigating one SNP. Dividing the SNP's realizations by the total number of individuals, an estimate for the joint distribution of phenotype and genotype is obtained. Consequently, it is possible to determine two marginal probability mass functions (mPMFs). This classification can be easily extended to a set of M SNPs. This requires building a contingency table for each SNP which results in a total of M tables. In some cases, this is not enough since this analysis only reflects single SNP association with the phenotype while a combination of SNP genotypes might be associated with the disease. Hence, in these cases it is required to build contingency tables reflecting multiple contributing loci.

After building these tables, it is necessary to test the association between the genotype and the phenotype. Therefore, the two marginal PMFs obtained from the contingency tables are compared using a test variable, e.g., the χ^2 or the log-likelihood ratio($2\hat{I}$) test variable. These deliver a number (test statistic) which for the case of independence has a χ^2 distribution with f degrees of freedom, where $f =$ (number of rows -1)*(number of columns -1) Then, a significance level is chosen, e.g., 0.05, which means 95% confidence level. As a consequence, it will be possible to reject the null hypothesis depending on the number obtained from the independence test and the defined significance level [1], [2].

This approach, however, suffers from two major limitations. The first one relates to the number of multiple contributing loci that can at most be taken into account. This will increase the number of contingency tables reflecting these multiple loci relationship with the phenotype, e.g., to investigate the all possible pairs of SNPs jointly for a sequence of 30 SNPs, a combination of two out of $30 = 435$ contingency tables have to be evaluated. The second limitation is that only discrete phenotypes can be treated. This is because the contingency tables cannot be defined for continuous phenotypes. To overcome such limitations, regression-based methods can be employed.

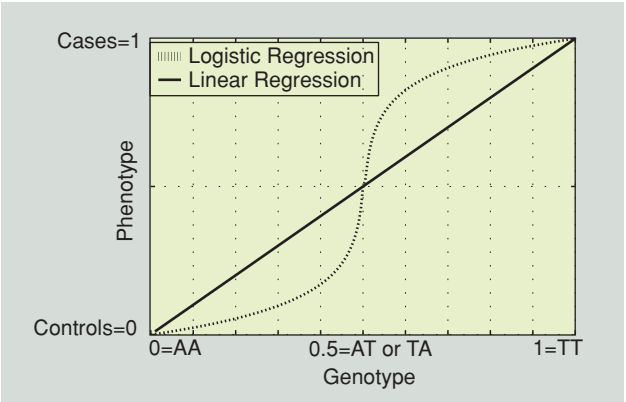
REGRESSION-BASED METHODS

Regression analysis is usually used to fit a set of measured points to an available response. This can be applied to the problem of gene mapping. The measured points are, in this case, computed from the genotype while the available response is the phenotype. The task is to find the curve that best fits the genotype to the phenotype measurements. The strength of the genotype-phenotype association relates to the measure of goodness of the fit [1], [2], [7], [8].

Consider, for example, a sample of individuals where the allele combinations (genotypes) can vary among AA, AT, TA, and TT. The genotypes AA and TT are homozygous and hence are given two different codes (0 and 1), while the genotypes AT and TA are heterozygous and are given the same code (0.5). The phenotype is assumed to be 0 for controls and 1 for cases. The linear regression between these two variables tries to find the curve that best fits the measured data, e.g., see the straight line of Figure 1.

The regression analysis works for both discrete and continuous phenotypes. Because the phenotype is dichotomous, i.e., binary, in case-control studies, logistic regression (LR) is often applied instead [9]. LR allows the prediction of a discrete outcome from a set of variables that may be continuous, discrete, or a mixture of both. It does not make any assumption about the distribution of the independent variables. In contrast to the linear regression analysis, the relation between the phenotype and the genotype in LR is not linear but logit-transformed. The LR between phenotype and genotype of the previous example is shown in the dotted curve of Figure 1. In this case, the regression curve asymptotically approaches the two realizations so that the error in the results is minimized [9].

Once the regression coefficients are obtained, they are tested for significance. This is done in a similar manner to the contingency-based methods. In other words, the regression coefficients are tested for significance using a test variable, whose distribution in the case of independence is known. Then, the SNPs with the significant coefficients are considered to be associated with the phenotype.



[FIG1] Regression analysis example.

Multiple linear regression is a systematic approach to a variable selection problem. However, it suffers from the same problem as table-based methods: too few individuals, not sufficient statistics. Hence, the SNP selection process must be simplified. One way to accomplish it is by employing the stepwise regression with the sliding window. The assumption by using the sliding window technique is that only the neighboring SNPs can be jointly causal [7]. The advantage of this technique is its stability and ability to provide the knowledge of the relationships and strengths among the variables, due to the stepwise elimination process. In addition, the stepwise regression method provides a more favorable approach than the contingency tables based methods since they usually are less complex to evaluate.

PEDIGREE-BASED LINKAGE STUDIES TEST FOR COSEGREGATION OF DISEASE AND NEARBY MARKER ALLELES AMONG AFFECTED AND NONAFFECTED MEMBERS OF ONE FAMILY AND ARE OFTEN REFERRED TO AS LINKAGE STUDIES.

INFORMATION-THEORY-BASED METHOD

Information theory is the science started by Claude E. Shannon in 1948 as the mathematical theory of communications [10]. Shannon defines the entropy of a random variable X, $H(X)$, as a measure for its randomness or uncertainty. The mutual information (MI) $I(X;Y)$ between two random variables X and Y is then defined as the reduction of uncertainty of one variable after observing the other: $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. MI turns out to be a suitable measure of dependence between two random variables: MI is symmetric and nonnegative and is equal to zero only if X and Y are statistically independent. While the concept of MI originated from problems in communications engineering, it can be also used for gene mapping of complex diseases [11]. The idea here is to regard the phenotype by a random variable P and a given SNP by a random variable S. The amount of information shared between P and S, i.e., the mutual information $I(P;S)$, yields the degree of association between the two in bits. In case-control studies, where the phenotype is binary, the maximum amount of information that can be observed is one bit. The basic idea to calculate the mutual information between the phenotype P and each SNP S separately, yielding M different MI values, where M is the number of SNP markers in the study. Note that complex diseases, however, are usually not caused by single markers. Rather, markers at multiple loci are suspected to jointly influence the disease. Thus, it makes sense to evaluate the MI between combinations of markers and the phenotype. As an example, a two-locus model of two SNPs with three possible states each will yield 3^2 possible combinations.

Obviously, the number of possible combinations grows exponentially with the number of markers combined, rendering this approach impractical in many cases. One idea to overcome this limitation is to use a sliding window approach. This approach, however, introduces a restriction as the assumption that jointly causal markers are also genetic neighbors may not be always justified for complex diseases.

Another approach is to use the concept of relevance chains [11]. Here, all single markers are checked for significant MI shared with the phenotype. In a second round, the conditional MI between each marker and the phenotype given the marker found S_j in the first step $I(S_i; P|S_j)$ is calculated. This procedure is repeated until no markers bearing significant additional mutual information are found. Using this method, patterns of jointly causal markers can be revealed, while the number of calculations does not grow exponentially with the number of markers involved.

Remarkably, MI is closely related to two other statistical measures of association, the χ^2 test and the log-likelihood ratio or $2\hat{I}$ test. The $2\hat{I}$ test variable differs from the mutual information only by a constant factor (of twice the sample size). The χ^2 test, in turn, is a second-order Taylor series approximation to the logarithmic $2\hat{I}$ test, being χ^2 distributed under the null hypothesis of no association [12], [13].

SIGNAL-PROCESSING-BASED METHOD

Consider a scenario where several speakers are talking together at the same time in a room, and several microphones are placed at different positions recording their activities. The aim of blind-source separation is to extract the signal of each speaker from the recorded mixtures. This situation is usually referred to as the cocktail party problem. One way to extract the signals is by employing independent component analysis (ICA) that tries to find a representation of the mixture of signals where the original sources are as independent as possible. The only assumptions to be made are that the original signals are independent, which is the case in almost every application, in addition to being non-Gaussian distributed [14], [15].

GENE-MAPPING ALGORITHM

Given a study comprising a total of N individuals (samples) divided among cases and controls where each individual's genotype is represented by a SNP genotype set of size M. As seen in the previous sections, our aim is to locate the SNPs associated with a given complex disease. In general, ICA tries to extract the useful information from mixtures of data where the prior knowledge about these mixtures is not available and have to be estimated [16]. In a similar manner, the genetic information of an organism regulates several unknown synergistical intracellular processes that result in a certain phenotype. Furthermore, it is assumed that the SNPs interact in an unknown environment to form new entities called SNP expressions. For simplicity, an example is illustrated in Figure 2 where SNPs 1 and 4 are supposed to transform to SNP

expression 1 after getting multiplied by the weighting factors a_1 and a_4. Similarly, SNP 7 transforms to expression 2 and SNPs 2, 6, and 10 transforms to SNP expression 3. In addition, assuming that the SNP expressions are the independent sources and the transformation process is the mixing environment, the problem changes to an ICA-based problem where a set of SNP expressions that are almost independent of each other have to be estimated along with some mixing environment.

The ICA-based gene mapping problem reflected in Figure 2 can be generalized mathematically as

$$\mathbf{E} = \mathbf{SA} \Rightarrow \mathbf{S} = \mathbf{EA}^{+}, \tag{1}$$

where $\mathbf{S} \in \mathbb{R}^{N \times M}$ contains the given SNP sequences of all individuals to be investigated, $\mathbf{E} \in \mathbb{R}^{N \times P}$ is the matrix of independent SNP expressions to be determined, $\mathbf{A} \in \mathbb{R}^{M \times P}$ is the SNP coefficient matrix to be determined, N is the number of individuals, M is the number of SNPs, P is the number of estimated SNP expressions, and $+$ is the pseudoinverse sign. As a direct property of ICA, the determined SNP expressions in $\mathbf{E}$ are almost independent from each other [14], [16]. Consequently, the SNPs that affect the same expression will be taken to be dependent, e.g., SNPs 1 and 4 in Figure 2, while the ones that affect different expressions are (almost) independent, e.g., SNPs 1, 7, and 10 in Figure 2. Furthermore, the entries of $\mathbf{A}$ define the magnitude of the contribution of each SNP to the corresponding SNP expression in $\mathbf{E}$.

The main task of ICA is to estimate the SNP expressions matrix $\mathbf{E}$ along with the mixing matrix $\mathbf{A}^{+}$ from the SNPs matrix $\mathbf{S}$. In other words, ICA will explore the provided genotype data in $\mathbf{S}$ to classify the SNPs into independent SNP expressions in $\mathbf{E}$ and obtain the magnitude of contribution of each SNP to the SNP expressions in $\mathbf{A}$. To relate the SNP contribution to the phenotype and, hence determine the associated SNPs, a regression analysis is performed between the SNP expressions and the phenotype.

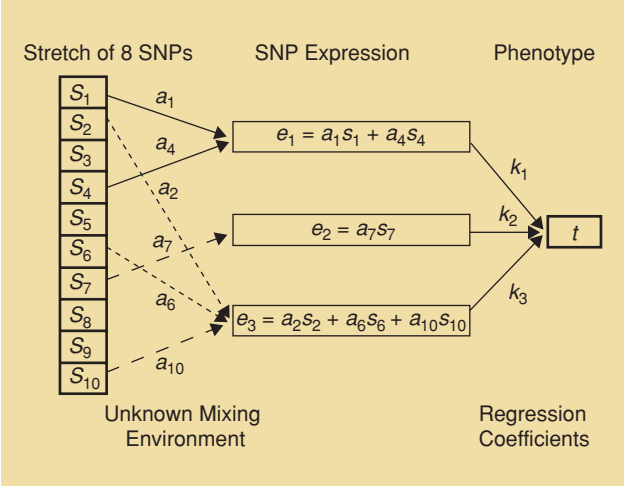
What is left to be done is to choose the regression coefficients and hence SNPs that are relevant to the phenotype. This is performed by choosing the SNP contributions where the regression coefficients have p-values less than a threshold, e.g., 0.05 for 95% confidence level, and disregarding all the others. (p-value is the lowest level of significance at which the test statistic is significant. It is a measure of probability that a difference between groups during an experiment happened by chance. A p-value of 0.05 means that the experiment has less than a 5% chance it has occurred by accident or 95% confidence level [1].) Following this methodology, it is possible to locate the associated SNPs [1], [2], [14].

IN POPULATION-BASED ASSOCIATION STUDIES, THE INVESTIGATED INDIVIDUALS ARE SUBDIVIDED INTO TWO GROUPS; ONE GROUP IS AFFECTED BY THE DISEASE UNDER INVESTIGATION, WHILE THE OTHER IS NOT AFFECTED.

To the contrary of the methods stated earlier, the analysis between the phenotype and the genotype is done indirectly, i.e., after transforming the SNPs into SNP expressions. The advantage of such transformation is that it relates clusters of SNPs represented by the expressions to the phenotype instead of the SNPs themselves. This will help in determining the dependence/independence of the SNPs causing the complex disease as will be seen later in the results without the restriction that only neighboring SNPs can cause a disease as in sliding window technique [7]. However, it will be interesting if some nonlinear techniques can be used to map the SNPs to the expressions prior to ICA as done in [16] in ordinary signal processing. Such a study might better model the interactions among the SNPs.

PREPROCESSING THE DATA

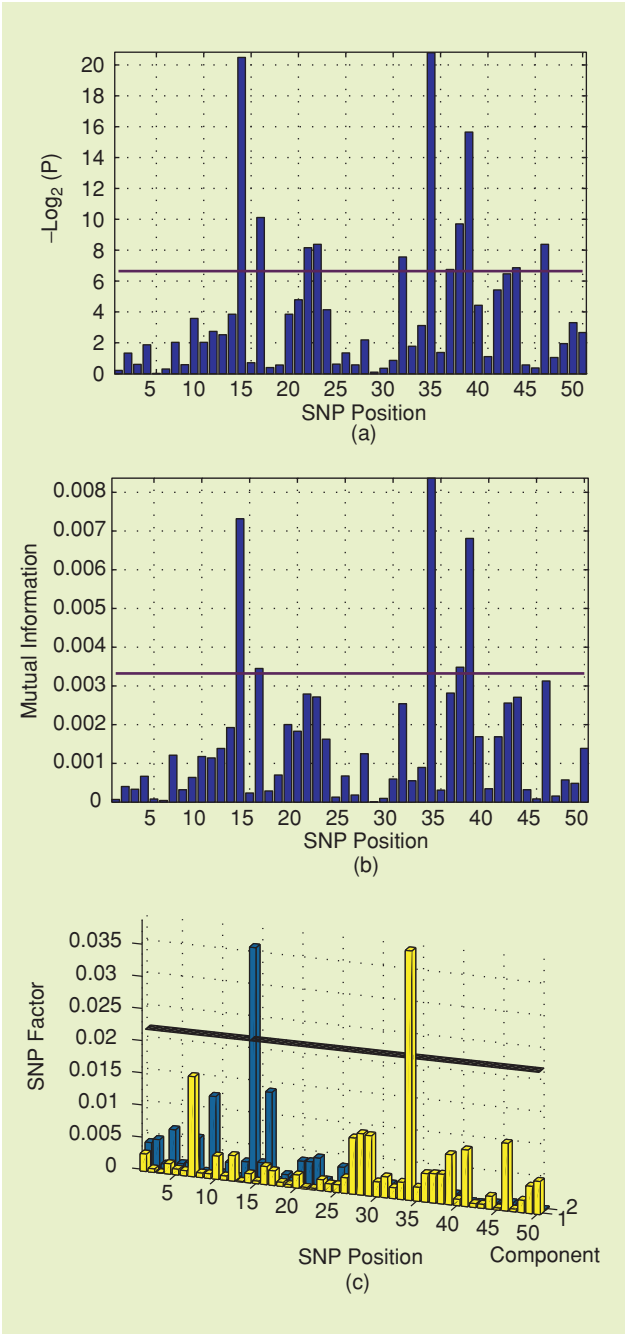
The SNP genotypes are usually acquired by standard genotyping. In addition, it is known that any data acquisition process suffers almost always from missing values due to many reasons such as errors of the instruments or low DNA quality. Such inconsistencies must be dealt with before processing the data. The methods described here generally solve this problem by omitting the missing SNP values in a given sample when building the data statistics. This solution is not favorable in linear algebra, since it requires not only omitting the missing SNP values in the sample of the individual, but the whole data set of the individual for the computations to be performed. Another way to overcome this problem is to estimate the missing values. Several estimation algorithms have been proposed in the literature for the task of DNA microarrays (continuous



[FIG2] Example of an individual's SNPs transforming to SNP expressions which are then fitted to the phenotype. In this example, three expressions, one individual, and ten SNPs are given. The numbers are only illustrative.

values) [17], [18]; however, none of the methods is specifically dedicated to missing genotype value estimation (discrete values). Theoretically, any estimation technique should work if the final value is rounded to the nearest value used to code the SNPs, e.g., 0, 0.5, or 1. Nevertheless, the estimation error will differ depending on the used method [14].

Another issue that should be dealt with is the number of SNP expressions that has to be estimated; i.e., the method that determines that only three SNP expressions should be estimated by ICA in the example illustrated in Figure 2. It is known in blind source separation that choosing all the possible components, the number of SNPs in our case, will make the solution deviate from the optimal one since the eigenvalues of the expressions with low magnitude will contain noise rather than information [16]. One might choose all the expressions that have eigenvalues larger than one [19]; nevertheless, this solution is also not very practical since it does not consider the structure of the data and reduces in many cases too much the data's dimension. The best methodology that should be followed is to employ a model-order selection technique that takes into account the structure of the data along with the number of samples available. Such techniques include Akaike information criterion (AIC), Bayesian information criterion (BIC), minimum description length (MDL), Laplace principle component analysis, and residual correlation technique (RCT) [20]–[23].



[FIG3] Association results of the simulated data set obtained by (a) LR, (b) MI, and (c) ICA.

RESULTS AND DISCUSSION

This section presents the results obtained using the gene mapping algorithms described earlier when applied to two different data sets. The first one is a simulated data set where the locations of the causal loci are known. A haplotype population was generated using a neutral coalescent model [24]. These haplotypes were then used to separate cases and controls according to selected causal loci using the software package SNAP [25]. The second set contains (unpublished) clinical data of individuals with the schizophrenia disease and of age, sex, and geographically matched control subjects. Logistic regression test was used in the comparisons with the mutual information and ICA based methods.

TWO LOCI SIMULATED MULTIPLICATIVE RISK DATA SET

The simulated data set was obtained using a double-loci full multiplicative risk model where the two loci affect the phenotype without interaction. The set contains 2,000 samples equally divided between cases and controls. The causal polymorphisms are located between positions 11–12 and 37–38, respectively, and were removed from the sequence. This is usually done to test the robustness of the algorithms since, in real life, the causal SNPs are not known when genotyping; however, the causal SNPs can still be determined by way of correlation between SNPs (linkage disequilibrium). The results of the algorithms is depicted in Figure 3. In graph (a), the logarithm of the p-values obtained by LR are plotted along with the significance level which was obtained in a similar way to the contingency-based methods which were previously described. In graph (b), the mutual information between the SNPs and the phenotype are shown along with the 99% significance level which was determined analytically [11]. In graph (c), only the two components that correspond to p-values less than 0.01 were plotted and the SNP factors (contributions) correspond to the entries of the SNP coefficient matrix A found by ICA multiplied by the regression coefficients. The 99% significance level was obtained in this

case by the permutation test [14]. The significance levels were plotted in each case to determine the decision border upon which a SNP is considered to be associated or not. As a first impression, it can be noticed that the three algorithms have determined successfully the locations of the causal regions. The highest peaks are at the locations 14 and 34, respectively. One might think that the algorithms are not accurate since the real locations are defined between 11–12 and 37–38. However, the causal SNPs were removed from the simulated data and the peaks found correspond to the SNP locations that are in high correlation with the skipped ones and thus justifying this small deviation. Note that the ICA based method was able to additionally determine the independence of the two causal SNPs since each one belongs to a different component (SNP expression).

SCHIZOPHRENIA DATA SET

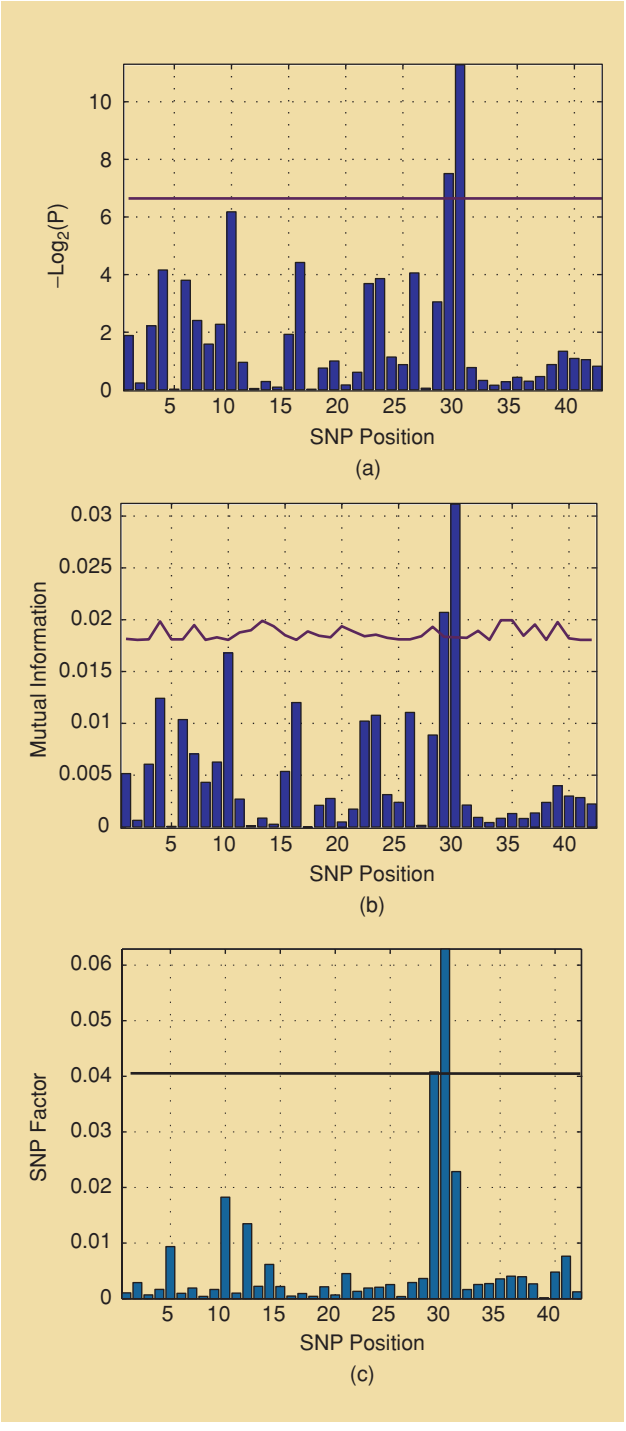
In this set, there are 42 candidate SNPs collected from different coding regions on different chromosomes. The set contains 368 samples equally divided between cases and controls. The causal SNPs are not known yet and need to be determined. Of all the available genotype values, 3% are missing due to genotyping failure and are estimated for the ICA-based gene mapping method according to [18]. The outcome of the different algorithms is depicted in Figure 4, where the values plotted are similar to the ones described in the previous data set. The significance level in the case of MI is not a straight line due to the presence of missing values in the data set [11]. As can be noticed, the highest peaks correspond in the three results and do not contradict each other (SNPs 29 and 30), emphasizing again the equivalence of these methods. It can be seen that the ICA-based method contains less side peaks than the two other techniques. This is due to the fact that ICA seeks for clusters of SNPs, and then the regression analysis eliminates the ones that are insignificant. In other words, the other peaks were eliminated since they belong to insignificant clusters. Note that the results of the different gene mapping methods are quite similar due to the relationship between MI and statistics [26] and between ICA and MI [16].

CONCLUSION

This work explains different methods to gene mapping. The main issue in these methods is to explore the variation among the genotype and relate it to the phenotype under study. The first method is based on statistical analysis which tries to relate the genotype and the phenotype using a probabilistic approach. The second method uses MI to measure the distance between two random variables and, thus, determine the association. The third method is based on signal processing theory. It redefines the process of gene mapping to make use of ICA to find a hidden entity called SNP expression. This entity is then related to the phenotype using regression analysis to find the association. Results obtained show that the three methods are able to give similar results when tested on simulated and clinical data sets.

AUTHORS

Michel Sarkis (michel@tum.de) received a bachelor's degree in electrical engineering at the American University of Beirut (AUB), Lebanon, in 2002 and a M.Sc. in electrical engineering at Munich University of Technology (TUM), Germany in 2004. Currently, he is pursuing his Ph.D. TUM. His research interests include image processing, time-varying systems, computer



[FIG4] Association results of the clinical data set Schizophrenia obtained by: (a) LR, (b) MI, and (c) ICA.

vision, signal processing, and statistical genetics. He is a graduate student member at the IEEE.

Bernhard Goebel (bernhard.goebel@tum.de) received the Diplom-Ingenieur degree in electrical engineering from Munich University of Technology (TUM) in 2004 . In 2002–2003, he spent graduate semesters at the University of Southampton, UK, and Siemens Corporate Research, Princeton, New Jersey, doing a research internship in the field of medical imaging. Currently, he is a research assistant and a Ph.D. student in optical communications at TUM working on information-theoretic optimization of fiber-optic communication systems. He is a graduate student member of the IEEE.

Zaher Dawy (zaher.dawy@aub.edu.lb) received a B.E. degree in computer and communications engineering from the American University of Beirut z in 1998. He received his M.Sc. and Dr.-Ing. degrees in Electrical Engineering from Munich University of Technology in 2000 and 2004, respectively. Since September 2004, he has been an assistant professor in the Electrical and Computer Engineering Department at AUB. His research interests are in the areas of wireless communications, ad hoc networks, multiuser information theory, and computational biology.

Joachim Hagenauer (hagenauer@tum.de) received his degrees from the Technical University of Darmstadt, Germany. He held a postdoctoral fellowship position at the IBM T.J. Watson Research Center, Yorktown Heights, NY, a one year visiting position year at AT&T Bell Laboratories, Crawford Hill, and a research position as the director of the DLR Institute for Communications Technology. Since April 1993 he has been a full professor for Telecommunications at the Munich University of Technology (TUM). He is a Fellow of the IEEE, a past president of the IEEE Information Theory Society, and the recipient of the IEEE Alexander Graham Bell Medal.

Pavol Hanus (pavol.hanus@tum.de) received his M.Sc. in electrical engineering from the Munich University of Technology (TUM) in 2004. Currently he is a research assistant and Ph.D. student at the Institute for Communications Engineering at the TUM. His research focuses on the possibilities of applying methods and concepts from communication and information theory to the problems in molecular genetics. He also investigates parallels between the handling of genetic information on the molecular level and information processing in communication systems. He is a graduate student member of the IEEE.

Jakob C. Mueller (mueller@orn.mpg.de) received his Ph.D. degree at the Department of Population Biology of the Johannes Gutenberg-University in Mainz, Germany, where he served as an assistant professor. Since 2002 he has held a position as a senior scientist at the Institute of Human Genetics, at the Institute for Medical Statistics and Epidemiology and at the Institute for Psychiatry and Psychotherapy at the Technical University in Munich. He is also affiliated with the National Centre for Genetic Epidemiological Methods. His research interests include population and evolutionary genetics in animals and humans as well as gene mapping for complex traits.

REFERENCES

[1] *Handbook of Statistical Genetics*, D.J. Balding, M. Bishop, and C. Cannings, Eds., 1st ed. New York: Wiley, Mar. 2001.

[2] J. Percus, *Mathematics of Genome Analysis*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, Jan. 2002.

[3] H. Cordell, B. Barratt, and D. Clayton, "Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects," *Genetic Epidemiology*, vol. 26, no. 3, pp. 167–185, 2004.

[4] R.A. Gibbs, J.W. Belmont, P. Hardenbol, et al., "The international HapMap project," *Nature*, vol. 426, pp. 789–796, Dec. 2003.

[5] B. Sobrino, M. Brion, and A. Carracedo, "SNPs in forensic genetics: A review on SNP typing methodologies," *Forensic Sci. Int.*, vol. 154, pp. 181–194, Jan. 2005.

[6] L. Cardon and J. Bell, "Association study designs for complex diseases," *Nature Rev. Genetics*, vol. 2, no. 2, pp. 91–99, Feb. 2001.

[7] H. Cordell and D. Clayton, "A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in type 1 diabetes," *Amer. J. Hum. Genetics*, vol. 70, pp. 124–141, Jan. 2002.

[8] D. 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 continuous traits in samples of unrelated individuals," *Hum. Heredity*, vol. 53, no. 2, pp. 79–91, 2002.

[9] A. Afifi, V.A. Clark, and S. May, *Computer-Aided Multivariate Analysis*, 4th ed. London, U.K.: Chapman & Hall, Jan. 2004.

[10] C.E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July, Oct. 1948.

[11] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. Mueller, "Gene mapping and marker clustering using Shannon's mutual information," *IEEE/ACM Trans. Comp. Biol. Bioinformatics*, vol. 3, no. 1, pp. 47-56, Jan. 2006.

[12] K. Pearson, *On the Theory of Contingency and Its Relation to Association and Normal Correlation*, (Biometric Series, no. 1). Cambridge, U.K.: Cambridge Univ. Press, 1904.

[13] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[14] Z. Dawy, M. Sarkis, J. Hagenauer, and J.C. Mueller, "A novel gene mapping algorithm based on independent component analysis," in *Proc. IEEE Int. Conf. Acoustic Speech Signal. Processing (ICASSP)*, Philadelphia, PA, vol. 5, pp. 381-384, Mar. 2005.

[15] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, no. 1, pp. 51–60, 2002.

[16] A. Hyvaerinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, May 2001.

[17] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, June 2001.

[18] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.

[19] H.F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, vol. 23, no. 3, pp. 187–200, Sept. 1958.

[20] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 1, pp. 461–464, 1978.

[21] T.P. Minka, "Automatic choice of dimensionality for PCA," *Neural Inform. Proc. Syst.*, vol. 13, no. 1, pp. 598–604, 2000.

[22] P. Stoica and Y. Selen, "A review on information criterion rules," *IEEE Signal Processing Mag.*, vol. 21, no. 4, pp. 36–47, July 2004.

[23] M. Sarkis, Z. Dawy, F. Obermeier, and K. Diepold, "Automatic model-order selection for PCA," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, vol. 6, pp. 933-936, 2006.

[24] R. Hudson, "Generating samples under a Wright-Fisher neutral model of genetic variation," *Bioinformatics*, vol. 18, pp. 337–338, Feb. 2002.

[25] M. Nothnagel, "Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-cotrol data by supervised learning methods," *Amer. J. Hum. Genetics*, vol. 71, no. A2363, 2002.

[26] B. Goebel, Z. Dawy, J. Hagenauer, and J. Mueller, "An approximation to the distribution of finite sample size mutual information estimates," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Seoul, Korea, vol. 2, pp. 1102-1106, 2000.

SP