# A NOVEL GENE MAPPING ALGORITHM BASED ON INDEPENDENT COMPONENT ANALYSIS

*Zaher Dawy*[1,3]*, Michel Sarkis*[1,4]*, Joachim Hagenauer*[1]*, and Jakob C. Mueller*[2]

[1]Munich University of Technology, Institute for Communications Engineering (LNT), Arcisstr. 21, 80290 Munich, Germany
[2]National Research Center for Environment and Health, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
[3]American University of Beirut, Department of Electrical and Computer Engineering, Beirut, Lebanon
[4]Munich University of Technology, Institute for Data Processing (LDV), Arcisstr. 21, 80290 Munich, Germany

## ABSTRACT

Identifying the causal genetic markers responsible for complex diseases is a main aim in human genetics. In the context of complex diseases, which are believed to have multiple causal loci of largely unknown effect and position, there is a need to develop advanced methods for gene mapping. In this work, we propose a novel algorithm based on independent component analysis for gene mapping. To apply the algorithm, we model the intra-cellular interactions as a mixing process of multiple sources. Results prove the superiority of the proposed algorithm over conventional statistical based methods, and demonstrate yet another successful application of a well known signal processing technique to an important problem in the field of human genetics.

## 1. INTRODUCTION

Gene mapping tries to identify the causal genetic markers that are responsible for apparent phenotypes such as complex diseases [1, 2]. With the development of rapid and cost-effective genotyping methods, the focus of research is shifting towards population-based case-control studies. These studies usually investigate sequences of Single Nucleotide Polymorphisms (SNPs) which are the predominant form of polymorphisms in the human genome. A number of standard statistical methods, e.g. Chi-squared or Haplotype Trend Regression (HTR) tests, are normally employed for gene mapping purposes [3, 4]. These methods do not follow an analytical approach to model the problem and are mainly based on a single SNP analysis which does not reveal possible interactions among the SNPs.

Recently, there has been an elevated interest in the interdisciplinary field of genomic signal processing which aims at applying signal processing techniques to problems in the field of human genetics, e.g. see [5, 6, 7]. In this direction of research, we propose a novel algorithm based on Independent Component Analysis (ICA) to locate SNPs that are causal loci for complex diseases. ICA is a well known signal processing technique, e.g. see [8, 9]. The essence of the algorithm is based on finding a suitable model that involves mixing of various sources so that ICA can be applied. The proposed algorithm is shown to perform better than conventionally used techniques and to be robust against missing values that occur due to genotyping failures.

Section 2 presents how the data was modeled to use ICA. Section 3 shows the different steps required to perform the proposed algorithm. Section 4 illustrates the results obtained by testing the algorithm on simulated and clinical data sets. Finally, Section 5 draws some conclusions.

## 2. PROBLEM MODEL

Given a study comprising a total of $N$ individuals (samples) divided among cases (sick) and controls (healthy) where for each individual a SNP sequence of length $M$ is provided. We assume that the SNPs have been transformed by unknown means to form some SNP expressions which will later affect a given phenotype (or disease), see Fig. 1. These expressions may be independent gene expressions or proteins in any living organism. Consequently, assuming that the expressions are independent sources and the transformation process is a 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.

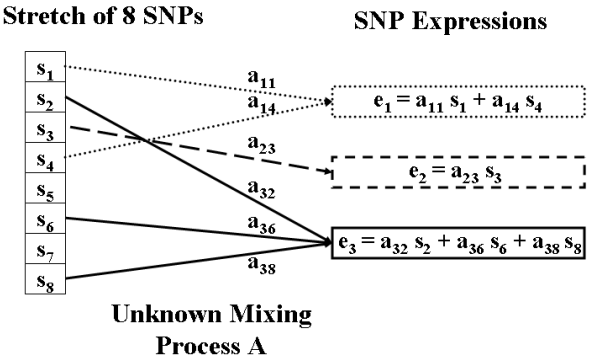


**Fig. 1**. SNPs transforming to SNP expressions.

Let $\mathbf{S} \in \mathbb{R}^{N \times M}$ contain the SNP sequences of all individuals. The general form of the problem is expressed as:

$$\mathbf{E} = \mathbf{SA}, \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{N \times P}$ is the matrix of independent SNP expressions, $\mathbf{A} \in \mathbb{R}^{M \times P}$ is the mixing matrix, and $P$ is the num-

ber of SNP expressions. Each column of **E** corresponds to the independent expression to be determined while each row presents the group of expressions that a person has. Similarly, matrix **A** or the *SNP coefficient matrix* determines the magnitude of contribution of each SNP to the corresponding expression. Each column of **A** contains the dependent SNPs' contribution to one of the expressions; therefore, the SNPs that contribute to one single expression are dependent while the ones that affect different expressions are almost independent. This verity will help determining if the SNPs are dependent or not for a multiple loci disease.

However, we are given the genotype data and not the SNP expressions. Hence, (1) should be rewritten as:

$$\mathbf{S} = \mathbf{E}\mathbf{A}^{+} = \mathbf{E}\mathbf{D} \tag{2}$$

where **D**, the demixing matrix, is the pseudo-inverse of **A**. The reason the problem was not derived directly as in (2) is the biological process itself. Usually, the SNPs determine the expressions and not vice versa. Consequently, ICA has to estimate **D** and **E** from the SNP matrix **S**. This model will use the provided genotype data to classify the SNPs into independent SNP expressions and then map the results to the phenotype to locate the causal SNPs. The specific steps of the proposed algorithm are detailed in Section 3.

Note that the only assumption made is on the distribution of the SNP expressions. The independent components to be estimated by ICA must be non-Gaussian distributed for the algorithm to work [8]. Moreover, the model is assumed to be linear. This is motivated by recent results obtained for microarray data analysis where the authors validate the accuracy of a linear model for ICA [10, 11].

## 3. THE GENE MAPPING ALGORITHM

The proposed gene mapping algorithm is composed of several steps, in addition to ICA, as shown in Fig. 2.

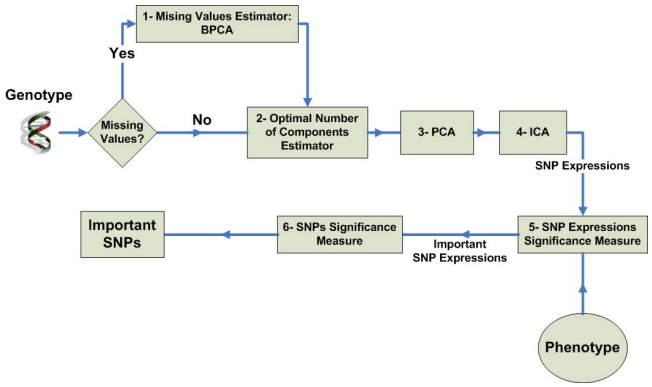


**Fig. 2**. Flow chart of the proposed gene mapping algorithm.

### 3.1. Missing Values Estimator

During genotyping, the nucleotides for some SNP locations are normally missing due to genotyping failures. To overcome this restriction, these values have either to be estimated or the samples that contain missing elements must be skipped. The second solution is not favorable since the amount of available data in clinical data sets is scarce and does not accommodate the elimination of extra samples.

Several solutions have been proposed in the literature to estimate the missing values [12]. The method that we adopt is Bayesian PCA (BPCA) because it gives lower estimation error-rates when compared to other methods, Moreover, it does not require any assumption of an underlying model and converges almost always to one solution [13].

BPCA starts by performing PCA while filling the missing values by the SNP-wise row average. Then, the distribution that best fits the missing values is determined using the variational Bayes algorithm. Consequently, these values are filled according to the found distribution. BPCA was originally designed for microarray data so its results have continuous values. Hence, we quantize its outcome as SNPs belong to a finite set of values, e.g. ternary or quaternary.

### 3.2. Optimal Number of Components Estimator

The second critical issue that should be dealt with is the number of SNP expressions (components) that has to be estimated by ICA. Choosing all the components of the data will increase the noise since the eigenvalues with low power comprise noise more than information [8]. Another proposal would be to choose all components that have eigenvalues larger than one [14]. This solution is also not very practical since it does not consider the structure of the data and reduces most of the time too much the data's dimension.

The methodology followed in this work is based on statistical fit. The idea is similar to the one introduced by [15] in factor analysis. This method is adapted to fit in the context of PCA. What has to be done here is to compute the covariance matrix $\mathbf{C}_S$ of the SNP matrix **S**, perform Singular Value Decomposition (SVD) on it, set the smallest eigenvalue to zero, compute the approximate covariance matrix $\widetilde{\mathbf{C}}_S$, and finally calculate the error difference between the two matrices. This process is repeated until a dimension $m$ is found where the standard deviation of the error becomes greater than the standard deviation of a distribution with zero correlation, i.e. $\sigma_{r=0} = N^{-1/2}$. Consequently, the number of SNP expressions $P$ that has to be estimated by ICA is nothing but $(m-1)$. This dimension should be given to PCA (step 3) for sphering and dimension reduction of the data. Then, the new transformed data should be handed to ICA (step 4) to determine the SNP expressions.

### 3.3. SNP Expressions Significance Measure

The obtained SNP expressions do not relate to the complex disease under study unless their relative importance to the phenotype is measured. Linear least squares regression is used in this work to measure the distance of each SNP expression to the phenotype. The larger the regression coefficients, the more important are the SNP expressions to the phenotype. Nevertheless, the aim of the algorithm is to find

out the contribution of each SNP; thus, each term of the regression coefficients must be multiplied by the corresponding column of the SNP coefficient matrix $\mathbf{A}$ as follows:

$$\mathbf{W} = \left( \mathbf{a}_1 \cdot k_1 \ \ldots \ \mathbf{a}_P \cdot k_P \right), \tag{3}$$

where matrix $\mathbf{W} \in \mathbb{R}^{P \times M}$ is the weighted SNP coefficient matrix, $\{\mathbf{a}_1, \ldots, \mathbf{a}_P\}$ are the columns of matrix $\mathbf{A}$, and $\{k_1, \ldots, k_P\}$ are the regression coefficients.

What is left to be done is to choose the columns of matrix $\mathbf{W}$ that are most relevant to the phenotype. This can be simply done by choosing the columns in which the regression coefficients have p-values less than $0.01$ for $99\%$ confidence level and disregarding all the others.

### 3.4. SNPs Significance Measure

In the final step of this algorithm, the significant SNPs that are most probably causing the disease have to be located from the relevant columns of $\mathbf{W}$. A common solution in genetics is the permutation test that was proposed by [16]. This test requires at each step the permuting of either the phenotype or the genotype data and then applying the gene mapping algorithm. This is done to destroy any relationship between the genotype and the phenotype of the marker loci in the results to obtain a distribution that should be similar to the one when there is no link between the genotype and the phenotype. The procedure is repeated tens of thousands of times. Then, the results are mapped into non-overlapping intervals allowing to build the required statistics at each SNP location. Hence, the border line that determines the significance of each SNP is nothing but the probability of $(1 - \alpha)$, where $\alpha$ is the significance level to be chosen. A typical value of $\alpha$ is $0.01$ for $99\%$ confidence level.

## 4. RESULTS AND ANALYSIS

This section presents the results of the proposed algorithm applied to three data sets. The first two are simulated data sets where the locations of the causal loci are known. They were generated using the SNAP software [17]. The third set contains (unpublished) clinical data of individuals with the Schizophrenia disease. In the displayed results, the ICA computations are performed using the FastICA algorithm [9].

### 4.1. Single Locus Multiplicative Data Set

The set *sim1locus* simulates a single locus multiplicative disease where the causal SNP is located between locations 23-24. Among the 62 candidate SNPs, the causal locus was removed from the simulation to test if the algorithm can detect the dependency of this SNP on the neighboring ones. Fig. 3 presents the results of our algorithm where the optimal number of components is found to be 17 and the remaining variance is 73% of the total. Every SNP location is plotted versus its contribution, SNP factor, for each of the components. By considering the p-values of the regression coefficients, only component number 12 has a p-value less than 0.01 (right plot). For 99% confidence level, i.e. the border line in the right plot, only SNP position 23 is selected as significant which reflects the correctness of the algorithm.

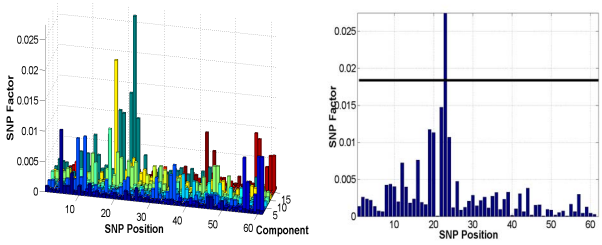


**Fig. 3**. Outcome of the proposed algorithm for the simulated data set *sim1locus*. Left: Plot of all the components (17 components, 73% variance). Right: Plot of the most relevant component.

### 4.2. Two Loci Multiplicative Data Set

The simulated set *sim2loci* imitates a double-loci multiplicative disease where the loci affect the phenotype independently. The causal polymorphisms are located between positions 11-12 and 37-38, respectively. The final outcome of the blind algorithm is illustrated in Fig. 4. As can be seen, two independent components were detected to be significant where each one has a different color. For 99% confidence level, the significant SNPs are 14 and 34, respectively. One might think that ICA has not performed correctly in this set since the estimated loci are not the exact locations; nonetheless, the causal polymorphisms have been removed. Thus, ICA has determined the SNPs that are in correlation with the removed ones and, as a consequence, the causal loci will be determined by dependency. This verity can also be shown in the output of the Haplotype Trend Regression (HTR) test in the right plot of Fig. 4 because these locations are also the most significant ones there. However, HTR was unsuccessful in capturing the *independence* of the two loci. This demonstrates the superiority of the proposed algorithm. Note that in HTR each SNP is plotted vs. the logarithm of the p-value found in least squares regression.

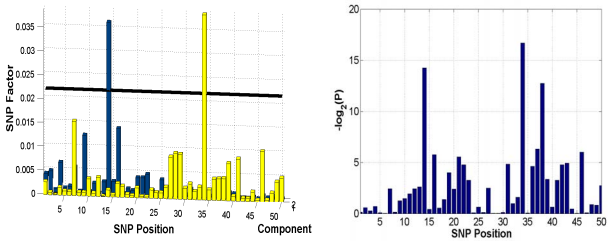


**Fig. 4**. Results for the simulated data *sim2loci*. Left: Outcome of the proposed algorithm (most relevant two components). Right: Outcome of HTR.

### 4.3. Schizophrenia Data Set

In this set, there are 42 candidate SNPs where the causal SNPs are not known yet and need to be determined. From all the genotype values, 3% are missing due to genotyping

failure and were estimated by BPCA. The final outcome of the blind algorithm is depicted in Fig. 5. For 99% confidence level the most suspicious SNPs have the locations 29 and 30. These two peaks do not contradict with the ones in the HTR result. However, the latter seems to have more significant peaks in its outcome than that of the proposed algorithm. Nevertheless, these peaks are removed from the final result with ICA because they belong to different components which have p-values more than 0.01. This once more demonstrates the superiority of the proposed algorithm.

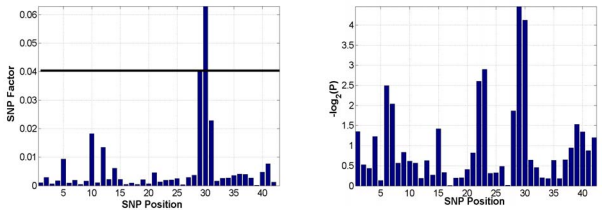


**Fig. 5**. Results for the *Schizophrenia* disease. Left: Outcome of the proposed algorithm. Right: Outcome of HTR.

## 5. CONCLUSION

This paper exploited the use of BSS techniques in gene mapping of complex diseases. The main advantage of this model is that it mimics to some extent the biological process. The SNPs get mixed in an unknown environment to produce a signal, the SNP expression, which will later cause the phenotype. Another advantage, is the revealment of the dependent and independent polymorphisms. This is due to the famous property of ICA that tries to find components that are as much possible as independent from each other. In addition, the algorithm is able to estimate the missing values eliminating the need to neglect some samples and it determines the number of components automatically taking into account the structure of the data. The only assumption that has to be made in this model is the distribution of the SNP expressions. They are supposed to be non-Gaussian distributed for the ICA algorithm to work.

The blind algorithm proved to be more accurate than the HTR since the latter cannot determine if the SNPs are dependent or not and consequently the contribution of the SNP clusters, or SNPs belonging to one component, to the phenotype. This property along with the fact that ICA is an unsupervised technique will help in eliminating SNPs that are considered to be significant by statistical methods through considering the p-values of the regression coefficients.

Other ICA algorithms can be used to include the prior knowledge (phenotype) in order to constraint the ICA solution. This might eliminate the need for the regression analysis and enhance the performance of the blind algorithm.

## 6. REFERENCES


[1] J. Hoh, A. Wille, and J. Ott, "Trimming, weighting, and grouping SNPs in human case-control association studies," *Genome Research*, vol. 11, no. 12, pp. 2115–2119, December 2001.

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

[3] D. Zaykin, P. Westfall, S. Young, M. Karnoub, M. Wagner, and M. Ehm, "Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals," *Human Heredity*, vol. 53, no. 2, pp. 79–91, May 2002.

[4] J. Percus, *Mathematics of genome analysis*, Cambridge University Press, Cambridge, England, 2002.

[5] J. Chen, H. Li, K. Sun, and B. Kim, "How will bioinformatics impact signal processing research?," *IEEE Signal Processing Magazine*, November 2003.

[6] P. P. Vaidyanathan and B. Yoon, "The role of signal processing concepts in genomics and proteomics," *Journal of the Franklin Institute, special issue in Genomics*, 2004.

[7] E. R. Dougherty, I. Shmulevich, and M. L. Bittner, "Genomic signal processing: The salient issues," *Eurasip Journal on Applied Signal Processing*, vol. 2004, no. 1, January 2004.

[8] A. Hyvaerinen, J. Karhunen, and E. Oja, *Independent component analysis*, Wiley, New York, USA, 2001.

[9] A. Hyvaerinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, May 1999.

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

[11] S. Lee and S. Batzoglou, "Application of independent component analysis to microarrays," *Genome Biology*, vol. 4, no. 76, October 2003.

[12] 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.

[13] 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, November 2003.

[14] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, vol. 23, pp. 187–200, 1958.

[15] A. Machado, J. Gee, and M. Campos, "Visual data mining for modeling prior distributions in morphometry," *IEEE Signal Processing Magazine*, vol. 21, no. 3, pp. 20–27, May 2004.

[16] G. A. Churchill and R. W. Doerge, "Empirical threshold values for quantitative trait mapping," *The Genetics Society of America*, vol. 138, no. 3, pp. 963–971, November 1994.

[17] M. Nothnagel, "Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-cotrol data by supervised learning methods," *American Journal of Human Genetics*, vol. 71 (Suppl.), no. A2363, October 2002.