COMBINING PHENOTYPIC AND GENOTYPIC DATA TO DISCOVER MULTIPLE DISEASE GENES

Hannu Toivonen, Saara Hyvönen, Petteri Sevon

Department of Computer Science, University of Helsinki
P.O.Box 68, FI–00014 University of Helsinki, FINLAND,
{hannu.toivonen, saara.hyvonen, petteri.sevon}@cs.helsinki.fi

ABSTRACT

Mapping genes for common, polygenic diseases is challenging due to the number of genes involved. Typical mapping methods search for one gene at a time, but their marginal effects may be too weak to be discovered in isolation. On the other hand, practically all methods for simultaneous mapping of multiple genes suffer from an exponential growth of time complexity as the number of genes grows.

We explore the combination of phenotypic and genotypic data in multiple gene mapping. Given a data set of individuals and observations of several phenotypes, as well as genetic marker data from the same individuals, our aim is to discover a set of genetic loci that together explain the observed phenotypes as well as possible.

We formulate three increasingly complex variants of the problem. We illustrate the problem by a regularized linear multiple regression model, where the most important components of a least squares solution are output as the predicted gene loci. Our initial results indicate that already this simple model can be more powerful than looking for one gene at a time. We discuss the problem variants and identify research problems for more powerful methods.

1. INTRODUCTION

The goal of gene mapping is to locate disease susceptibility genes for a given disease. The existence and importance of a genetic component in the etiology of the disease has usually already been identified, so the question now is about where the disease susceptibility gene or genes are located.

In a typical form of gene mapping, a single phenotype has been recorded for a set of individuals. This phenotype can be simply a dichotomous indicator "diseased" vs. "healthy", or it can be a quantitative measurement, such as blood pressure. Additionally, the individuals are genotyped at a number of marker loci. Given these data, the task is to find those loci that are most strongly associated with the phenotype.

A related problem is the analysis of candidate genes: a set of most promising genes identified in gene mapping studies are genotyped in a subsequent study, and the task is to identify those candidate genes that actually contribute to the phenotype.

Diseases are usually complex, i.e., affected by several genes, as well as by environmental factors. The power to detect one gene at a time may be weak, as their marginal effects can be small or even non-existent (Hoh and Ott, 2003; Marchini et al., 2005). Any straightforward generalization of the basic method to consider several loci at the same time, instead of just one, has exponential time complixity: given d, the total number of loci, and k, the number of loci to be considered in conjunction, there are $O(d^k)$ different combinations of loci to consider. Since d can be large, in the order of thousands or more, this approach does not easily generalize beyond small values of k.

The size of the other important dimension, number of phenotypes, varies from one to tens. With the advance of high throughput genotyping technology, large data sets collected for other purposes than gene mapping, e.g., for epidemiology, have become more relevant for genetical studies and gene mapping. These data sets can contain large numbers of phenotypes related to the same disease, obtained by clinical tests or questionaries, or extracted from medical records. Since different genes, all related to the same disease, are likely to have different effects on different phenotypes, this rich data could be powerful for mapping of multiple genes.

In this paper, we consider the problem of identifying multiple disease susceptibility loci (or quantitative trait loci) at once. We formulate three different variants of the problem (Section 2). We then describe a least squares solution to one of them and experimentally compare it to a single gene approach, to illustrate the nature of the problem (Section 3). In Section 4 we discuss the described problems further and identify a number of research problems. Section 5 contains our conclusions.

2. PROBLEM FORMULATION

Given both phenotypic and genotypic data for a sample of individuals, the task is to find a model that explains the observed phenotypes with (a subset of) the genotype data. As an illustration, in a simple linear model the observed phenotypes (B) are the product of genotypes (A) and (un-

known) effects of genes (X),

$$B = AX + \text{noise}. \quad (1)$$

With this formulation, the task essentially is to solve the equation for X. We will return to the subtleties in the problem definitions below.

Phenotypes are given in an $n \times m$ matrix B, where the n rows represent individuals and the m columns their phenotypes. We assume that the phenotypes are related to the same disease or family of diseases. For instance, the diagnosis of asthma is based on a number of different alternative indicators of symptoms, and all these could be used as phenotypes, as well as, e.g., any other respiratory symptoms. Genotype data is given as a 0/1 matrix A, with n rows for the individuals and d columns that correspond to genetic loci. The genotype data consists of candidate genes, genetic markers, or haplotypes. Since these all represent possible disease gene loci, the terms "gene" and "locus" are interchangable in most of our text.

An assumption is that the same disease, manifested in a number of phenotypes, is affected by a number of genes, but these k genes are only a small subset of all d given candidate loci. Each of the k true genes typically contributes to several phenotypes, and conversely any given phenotype can be affected by several genes. Each person has a mutation in zero or more genes, and his or her phenotypes are the result of these genes, plus a number of other factors, such as the environment.

In current datasets, the number of individuals, n, is typically in the order or hundreds, while the number of phenotypes, m, is in the order of tens. The number of genetic loci, d, can range from tens to thousands depending on the study. The number of genes with a real effect on any of the phenotypes is largely an open question, but it is commonly believed that many complex diseases have few important genes (less than ten) and more genes with minor effects.

We next formulate increasingly complex variants of the multiple gene mapping problem. For convenience, we describe them using notation from linear algebra; in reality, however, the phenomena cannot be assumed to behave linearly.

- **Problem 1a** For the first problem variant we assume that the genotypes in matrix A are obtained for a set of named candidate genes; each column of A corresponds exactly (up to genotyping accuracy) to the presence or absence of a specific allele in a specific gene.

  Given phenotypes B and genotypes A of named candidate genes, the task is to explain (all) the observed phenotypes in terms of a subset of k candidate loci from the whole set of d loci. Assuming that the effects of genes on the phenotypes can be described with a linear additive model, the task is to solve X in (1), where X is a $d \times m$ matrix where only k rows have non-zero elements: these rows give the contributions of the k genes on the phenotypes.

- **Problem 1b** In another variant of the first problem, the available genotype data A consists of alleles of markers, not specific genes. The closer a marker is to a disease susceptibility gene, the larger is the linkage disequilibrium between these, and consequently also their correlation. In this case some of the columns of A correlate with the presence of an actual gene.

  Formally this variant can be expressed as Problem 1a. A subtle difference is, however, that in this case the genotype data A can be considered noisy, and this can have an influence on the methods for solving X in (1). This can be viewed as a total least squares problem (Golub and Van Loan, 1996; Golub et al., 1999), where an additional goal is to also estimate the true disease susceptibility gene genotypes, $\hat{A}$.

- **Problem 2** A more challenging problem is to use only the phenotype data B to simultaneously estimate genotypes at k unspecified gene loci and the effects of the genes. This corresponds to factorizing B into A and X so that

  $$B \approx AX, \quad (2)$$

  where A is a 0/1 genotype matrix of size $n \times k$ and X a gene effect matrix of size $k \times m$. The utility of this approach is that once the genotypes A have been estimated, each gene can be mapped individually in the normal way, i.e., by finding a locus that tends to occur in those individuals identified by the corresponding column in A.

Obviously, with Problem 2 we have the least amount of information at our use, whereas with Problem 1a we are the most informed. We next address Problem 1a as an example case, and then return to discuss all the problems.

3. EXAMPLE: LINEAR MULTIPLE GENE MAPPING FOR PROBLEM 1A

We now illustrate the multiple gene mapping problem by providing a least squares solution to the problem variant 1a and by experimentally comparing its performance to the a single gene approach.

3.1. A least squares solution

Problem 1a is formulated in linear algebra as follows: given the matrices A and B with dimensions $n \times d$ and $n \times m$ respectively, find the $d \times m$ matrix X that solves the problem

$$AX = B. \quad (3)$$

When $n > d$ the system is *overdetermined* and usually has no exact solution, so instead we solve the least squares problem

$$\min_X \|AX - B\|_2. \quad (4)$$

More commonly this is expressed in the vector form: for each column b_j

$$x_{\cdot j} = \arg\min_x \|Ax - b_j\|_2. \quad (5)$$

Here we denote the column vectors of the matrix X by $x_{\cdot j}$ and the rows by $x_{i\cdot}$. For details on the solution of the least squares problem see e.g. Golub and Van Loan, 1996. For ill-conditioned problems, where the columns of A are nearly dependent, this is very sensitive to noise. The standard way to overcome this problem is to use some form of regularization. The most common form is Tikhonov regularization (Tikhonov and Arsenin, 1977; Hanke and Hansen, 1993), which in the general case takes the following form: for each column b_j,

$$x_{\cdot j} = \arg\min_x(\|Ax - b_j\|_2^2 + \lambda_j\|x\|_2^2). \quad (6)$$

Here the regularization parameter λ_j controls the size of the solution. The solution $x_{\cdot j}^{\lambda}$ to (6) solves the problem

$$(A^TA + \lambda_j I)x = A^Tb_j. \quad (7)$$

Various methods for choosing the optimal regularization parameter exist. One way is to use generalized cross-validation (GCV), see e.g., Hansen, 1994, and references therein. If there is no reason to expect the optimal regularization parameter to differ from column to column, a reasonable approach is to use the mean of the λ_j obtained by GCV for different columns as the regularization parameter of the whole problem.

Ideally our solution X is a $d \times n$ matrix, the elements of which give the contribution of each gene to each phenotype. Thus, only the k rows corresponding to the true genes have elements that significantly differ from zero. We can identify the important genes by looking at the norms of each row $x_{i\cdot}$ of X.

A baseline against which to compare the above method is provided by the obvious approach of simply considering one gene at a time and testing how well it explains the phenotype data with no interaction with other genes. This corresponds to replacing the matrix A in (4) by each column a_j of A one at a time, and solving for $x_{i\cdot}$, where each $1 \times m$ vector $x_{i\cdot}$ corresponds to one gene. This corresponds to simply taking the average phenotype vector of those individuals carrying gene i. Unlike in the case described above, the vectors $x_{i\cdot}$ do not directly tell us how the genes contribute to the phenotype. Still, a comparison of vectors $x_{i\cdot}$ indicates relative strengths of genes on phenotypes. Again we can order the $x_{i\cdot}$ according to their norms, and select as significant the genes with the largest norms.

3.2. Experiments

We next describe experimental results with the above described least squares method. We use synthetic phenotype data, generated with a linear model involving a number of disease susceptibility genes. Given the occurrence vectors of these real genes plus a large number of irrelevant candidate genes, we measure the power of the method to discover the correct genes among all genes.

Parameter	Values
k, number of true genes	1 – 30
d, total number or loci	50 – 2000
m, number of phenotypes	1 – 20
n, number of individuals	50 – 1000
fraction of unobserved genes	0 – 90%
noise/signal ratio	0 – 300%

Table 1. Parameter value ranges in simulations

Data simulation The synthetic genotype and phenotype data are generated under the following assumptions: (1) All genotypes are mutually independent and they have frequency 0.3, i.e., the genotype matrix A is a random 0/1 matrix of size $n \times d$ with 30% of ones. (2) The effects of true disease susceptibility genes on the phenotypes are non-negative and mutually independent, i.e., the gene effect matrix X is a non-negative random matrix of size $d \times m$. In our experiments, the elements of the k non-zero rows of X are initially drawn uniformly from $[0, 1]$. To simulate genes of different importances while avoiding unnecessary randomness in the experiments, the effects of true genes (rows of X) were then weighted by values $1/k, ..., k/k = 1$. (3) Each observed phenotype vector is the sum of the effects of the genes that the person has plus Gaussian noise, i.e., $B = AX +$ noise.

Probably none of the above assumptions is fully realistic. A particularly strong assumption is the use of a simple additive model of gene effects: in reality, there can be complex interactions between genes, and between genes and environmental factors. For Problem 1b another severe assumption would be that the genotype matrix has no noise.

We generate data using wide ranges of parameters (Table 1). In each setting, we simulate 100 independent data sets and report the average results. Despite our strong assumptions we believe this data and the range of experiments is indicative of the nature of the problem.

Evaluation methodology We evaluate the success of the method by measuring how well the k true disease susceptibility loci rank when there are a number of irrelevant loci included. This corresponds to the situation where there is a large number of candidate loci, and the task is to identify a small subset that is likely to include the true ones.

Using the least squares solution, we rank the loci i (row $x_{i\cdot}$) by $\|x_{i\cdot}\|_2$, the Euclidean norms of their effects on the phenotypes. In the optimal case, the true loci are the k best ranking ones. As a measure of the quality of the result, we use the fraction of true k loci among the best ranking k loci. We will later briefly consider methods for choosing a good value for k. Additional information about the quality of the result could be obtained by comparing the estimated gene effects (rows $x_{i\cdot}$) to the real effects used when simulating the data. This is, however, of secondary interest: the primary goal is to identify a small set

Parameter	Setting A	Setting B
k, number of true genes	5	10
d, total number or loci	500	1000
m, number of phenotypes	10	5
n, number of individuals	500	300
fraction of unobs. genes	0%	20%
noise/signal ratio	5%	50%

Table 2. Fixed parameter settings (Figure 1)

Parameter	Setting C
k, number of true genes	10
d, total number or loci	100
m, number of phenotypes	10
n, number of individuals	120
fraction of unobs. genes	0%
noise/signal ratio	10%, 100%

Table 3. Parameter settings for norm plots (Figure 2)

of loci for further investigation.

We empirically evaluate the effect of one parameter of Table 1 at a time, while keeping all other parameters constant. For the constant parameters we fix two sets of values (Table 2). In Setting A, the task is to find $k = 5$ true disease loci from $d = 500$ loci in total. There are $m = 10$ phenotypes and the number of individuals is $n = 500$. There is modest noise; the variance of phenotype noise is 5% of the variance of the noiseless ideal phenotypes.

Setting B is designed to be very challenging, and it is more difficult in terms of all our parameters. In particular, in Setting B, 20% of the true genes are unobserved, i.e., they are not included in the genotype matrix A. This should make the task of multiple gene mapping more difficult—but also more realistic, since in real applications there are likely to be numerous factors that are not included in the data nor the model.

Results An overview of the fraction of highly ranked true genes is given in Figure 1. Each of the panels shows the power of the method as a function of one of our parameters. Results for Setting A are drawn with solid lines, results for Setting B with dashed lines; thick (blue) lines are for the multiple gene approach (regularized least squares method) and thin (red) lines for the baseline of finding a single gene at a time. For a reference of how the fixed values in Settings A and B relate to the range of values tested in each subfigure, results in Settings A and B are denoted by an asterisk and a plus, respectively.

We make two general observations. First, in most of the tests the multiple gene approach achieves very good results in Setting A, ranking all or almost all of the k true genes among the best k genes. Setting B is much more challenging, and powers are generally between 0.5–0.8. Second, in the two different settings and over the relatively wide ranges of parameter values, the multiple gene approach outperforms the single gene approach, except for high fractions of unobserved true genes.

One of the most striking results is the effect of the number k of true genes on the power to detect those genes (Figure 1.A). Since the magnitudes of gene effects are (deterministically) uniformly distributed in $[1/k, 1]$, there will be more very minor genes when their number is larger, and the task of finding all of them becomes more difficult. This seems true especially if genes are searched for individually: the single gene approach experiences a clear drop also for the easier setting A as the number of genes increases.

The total number of genes (Figure 1.B) has a relatively slight adverse effect on the power, as does the fraction of unobserved true genes (Figure 1.C). Unobserved genes theoretically have no effect on the single gene approach, and deviations from a horizontal line are due to random effects.

Noise obviously has an adverse effect (Figure 1.D), but not a very strong one. The variance of noise has to be larger than the variance of signal to have a clear effect. The multiple gene approach seems to suffer slightly more from noise, so that with signal-to-noise ratio 1/3 ("300%" in the figure) the two approaches perform equally well.

The two parameters, n and m, that govern the amount of data (dimensions of matrix B) give interesting results. First, in our settings, increasing the number m of phenotypes beyond 5 does not affect the results much (Figure 1.E). On the other hand, increasing the number n of individuals steadily improves the results throughout the whole range (Figure 1.F). Note that here we have occasionally violated the assumption $n \geq d$, but for our problem this does not seem to affect the performance of the regularized least squares method.

As an example of how the multiple gene and single gene approaches rank loci, Figure 2 shows the Euclidian norms $\|x_{i\cdot}\|_2$ of the estimated gene effects on the phenotypes, sorted in descending order. For illustration, the data comes from Setting C with a smaller number genes ($k = 10, d = 100$; Table 3). The left column of the figure shows results with the baseline of finding genes individually, the right column with the multiple gene approach (regularized least squares method).

In this data set, three of the true genes rank badly in the single gene approach, largely regardless of the amount of noise. In the multiple gene approach, adding noise seems to affect the weakest genes more gradually.

The norms of gene effects could potentially be used in some cases to estimate the number k of true disease susceptibility genes. Based on Figure 2, in the case of the multiple gene approach it is fairly easy to get a reasonable estimate for k by looking at where the norms level off. The situation is less clear for the single gene approach, where the distribution of norms is much smoother.

The unregularized least squares solution has a poor performance, unless there are enough individuals and no or very little noise (results not shown).

Conclusions from results The results indicate that already the simple linear multiple regression model as a multiple gene approach can be more powerful than a similar single gene approach. However, since the data was

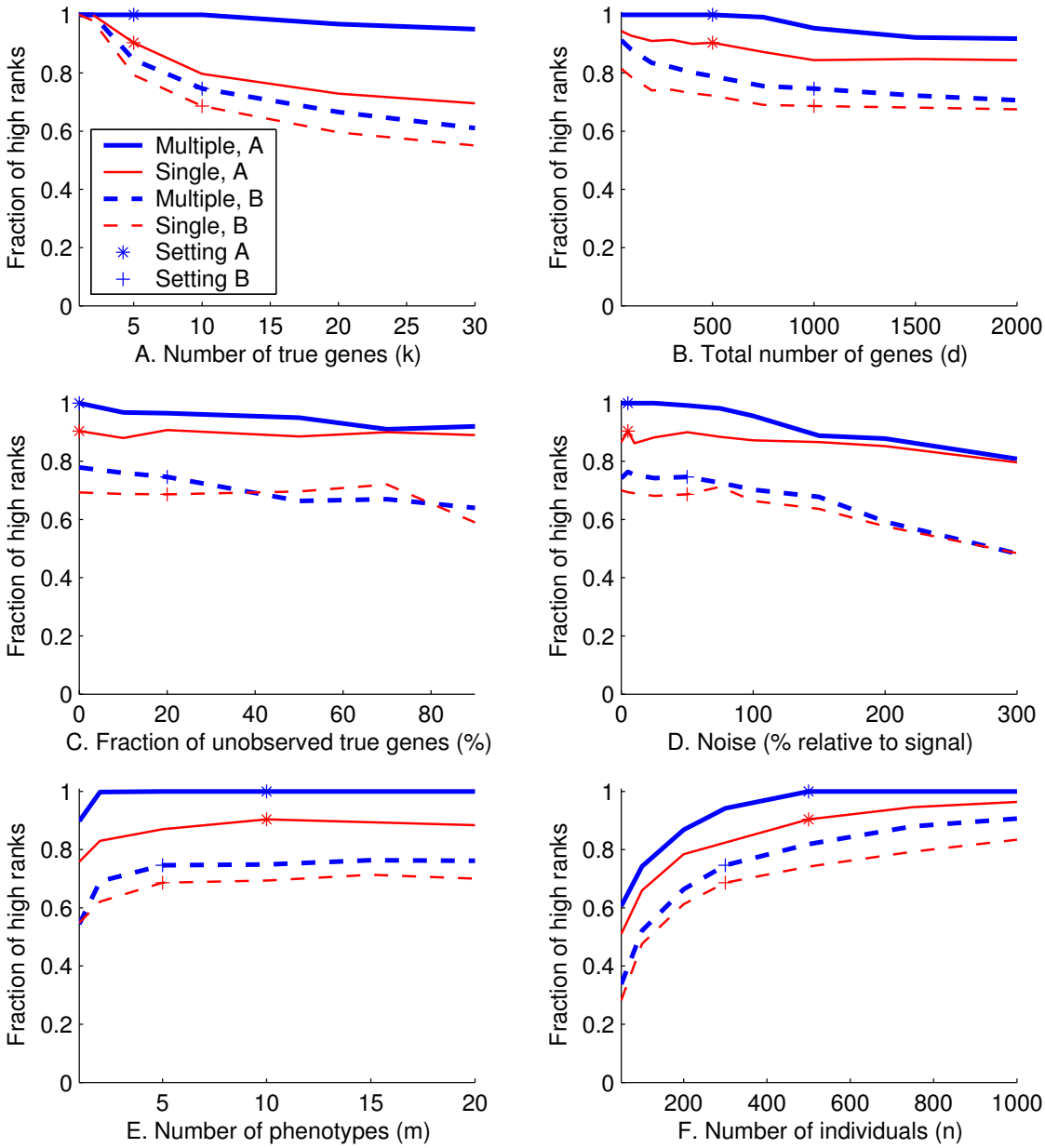

Figure 1. Fraction of highly ranked real genes as a function of various parameters.

generated with the same model that was used to analyze it, these conclusions are at most indicative. On the other hand, no gene-gene interactions were simulated, making the task easier also for the single gene approach.

Our goal here has been to illustrate the problem by applying a relatively simple multiple gene mapping method to it, and comparing it to the single gene approach. Under our assumptions—which we know are not fully realistic—it seems that when possible, increasing the number of individuals is more useful than increasing the number of phenotypes, and that increasing the total number of loci only has a modest adverse effect. The multiple gene approach seems more powerful, especially for large numbers of true genes in otherwise easier settings. As expected, and illustrated by the norms in Figure 2, the single gene approach has more problems identifying genes with small effects.

4. RESEARCH TOPICS

We now move on to discuss research topics related to solving the three problem variants we introduced. Many methods for multivariate quantitative trait loci (QTL) analysis (essentially our Problems 1a and b) based on linear models have already been proposed (Jiang and Zeng, 1995; Henshall and Goddard, 1999; Caliński et al., 2000; Knott and Haley, 2000; Bjørnstad et al., 2004). To the best of our knowledge, Problem 2 has not been addressed in the literature.

Gene selection For Problems 1a and b, a central task is the identification of the set of true disease susceptibility genes underlying the observed phenotypes. The simplest possibility in our example least squares model is to take the k best ranking loci, where k is a parameter given

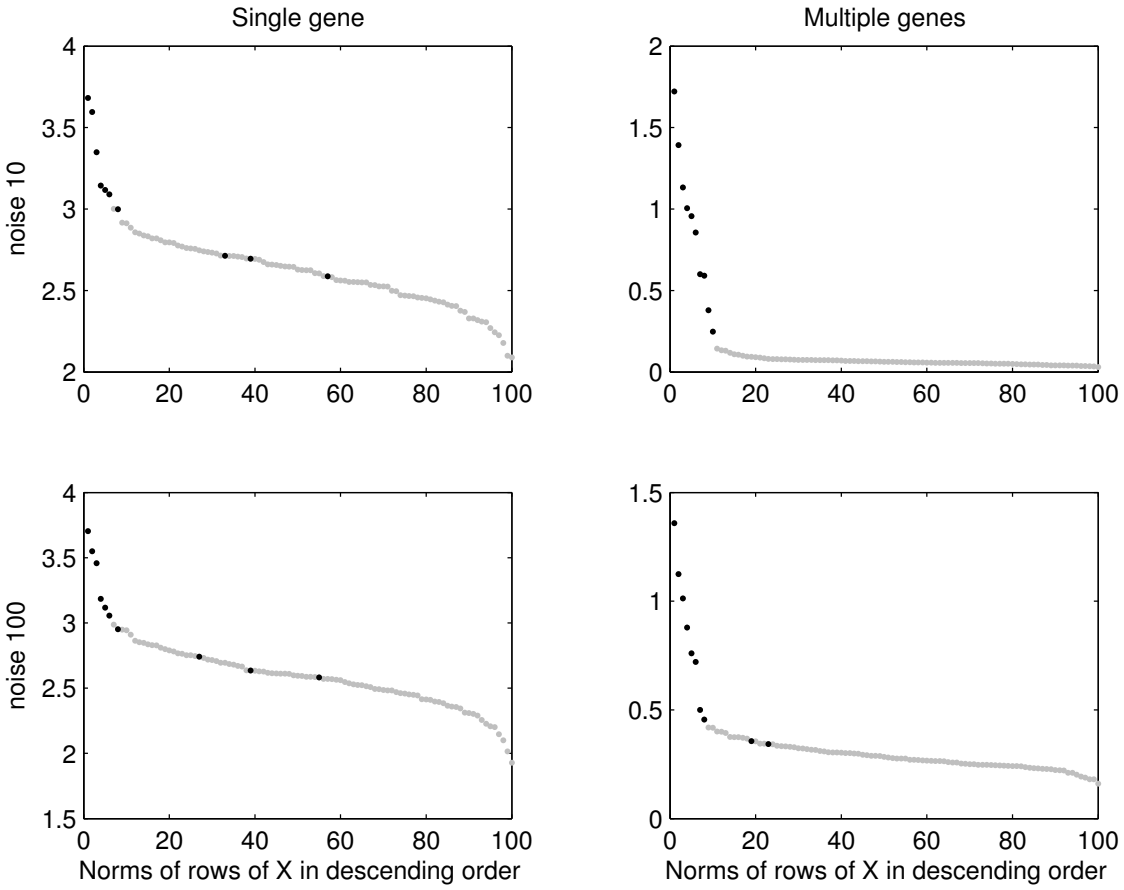

Figure 2. Norms of the gene effects (rows of X) sorted in ascending order. The points corresponding to true genes are in black, genes with no effect in gray.

by the user. Automatic selection of a suitable k could be based on an analysis of the norms of the effect of genes, as mentioned in the previous section.

It is probably more powerful to choose k as part of the model fitting process. Stepwise methods for fitting a multiple linear model produce a nested set of models; in each step either the variable (gene) that best explains the residual is added to the set of independent variables (forward selection), or the least significant independent variable is removed from the set (backward elimination).

The least squares method without regularization is equivalent to maximizing likelihood over X in a model with Gaussian noise in the observed phenotypes. Regularization adds a prior Gaussian distribution for X, and the task is then equivalent to finding the *Maximum a Posteriori* solution for X, where the regularization parameter λ determines the ratio between the variance of noise and the variance of the prior distribution. With the probabilistic interpretation, statistical criteria (e.g., AIC (Akaike, 1973) or BIC (Schwarz, 1978)) for model selection can be applied.

Improvements to the least squares model It is relatively easy to allow for a mixture of quantitative and dichotomous phenotypes in the least squares model by using logistic regression for the dichotomous variables. However, when dichotomous variables are included, the absolute values of the variances of the observed phenotypes and the prior of X become fixed.

Finding candidate loci In a genomewide analysis without any prior candidate genes (Problem 1b), the genotypes are typically obtained for a large number of marker loci. Because of linkage disequilibrium, markers can be used as surrogates for nearby genes. Haplotypes of several adjacent markers are more informative—and potentially better surrogates—than alleles at a single marker. It is possible to enumerate all haplotypes with some reasonable constraint on their length occurring in the genotype data with at least at a given threshold frequency (Toivonen et al., 2000).

This set of haplotypes may be very large, and it may be necessary to prune it further. One can reject haplotypes with weak individual phenotypic association prior to fitting the linear model, with the risk of losing low marginal effect loci. Another option is to prune similar haplotypes, and so find a smaller set of haplotypes that represents well the whole (analyzed area of the) genome as well as different distributions.

Matrix factorization The multiple gene mapping problem is considerably more challenging if also the genotype matrix A is unknown (Problem 2). Matrix fac-

torization methods such as *Nonnegative Matrix factorization* (NMF) (Lee and Seung, 2001) or *Independent Component Analysis* (ICA) (Hyvärinen et al., 2001) may be applied. NMF assumes non-negativity of the matrices, which may be a useful constraint. Without loss of generality, we can assume that phenotypes are non-negative; a usual assumption—although not strictly necessary—is that that mutations in genes (1's in matrix A) predispose to the disease and higher values of phenotypes.

Neither NMF nor ICA as such account for the binary nature of the genotype matrix A, and modifications for this task would be needed. As we expect the matrices A and B to be fairly sparse, that is to have a low proportion of non-zero elements, a sparse NMF approach (Hoyer, 2004) may prove an attractive alternative.

Stability and uniqueness of the results For any method to solve the problems, it would be useful to be able to assess how stable or unique the solution is. If equally good results can be obtained with different sets of genes or different phenotype effects, is it because there are no clear gene effects, or because they can be explained in different ways?

For instance, collinearity is a potential problem. If there are strong correlations between columns of A, the corresponding rows of X are unstable. The issue is more serious in the case where matrix A consists of a large number of haplotype patterns. Bjørnstad et al., 2004, addressed the problem using partial least squares regression, in which the matrix is regressed onto a smaller set of uncorrelated columns. Another option is using a stepwise method with a constraint limiting the correlation between the independent variables. The combinatorial nature of the problem calls for use of a less greedy algorithm, e.g., beam search.

Non-linear methods Our discussion so far has been limited to the case where each gene has an independent linear contribution to the phenotypes. This is not a realistic assumption for the gene effects—even if such models may sometimes be sufficient for finding genes.

Pairwise interactions could be easily incorporated, but with a considerable computational cost because the number of possible interactions is quadratic in the number of loci. Considering only interactions within the already selected genes is more affordable, but some gene effects might be only observable through the interactions and such genes would be missed. An intermediate alternative is to first find all loci with individual phenotypic effect meeting some low threshold, and then considering all pairwise interactions within this reduced set of loci (Marchini et al., 2005; strategy III).

A very liberal model of arbitrary gene interactions would allow the effects of all gene combinations to be mutually independent. Problem 2 would then reduce to clustering. Namely, given k there are 2^k different gene combinations (of which some maybe do not occur in the data). Assuming individuals with a given combination are relatively homogeneous, the task now is to find (at most) 2^k clusters from the phenotype data B. This connection to clustering is interesting but not very useful as such: there is no direct way to assign any particular subsets of the k genes to the clusters. Suitable constraints on the clusters and the related genes, such as requiring that adding a gene never decreases the expected phenotype values, might help finding a clustering that is more likely to correspond to carriers of different sets of genes.

5. CONCLUSION

Simultanous mapping of multiple genes, using a combination of phenotypic and genotypic data, can be necessary for the localization of genes with small or non-existing marginal effects. In this paper we formulated three variants of the multiple gene mapping problem. We illustrated the problem by developing a regularized linear least squares solution to one variant, and by experimentally comparing it to the single gene approach. Our simulations were simple, but they demonstrate how a multiple gene approach can be more powerful in detecting genes with small effects.

We discussed several research issues related to the multiple gene mapping problem and identified a number of research opportunities potentially leading to more powerful mapping methods. We believe that significant advances can be made in this problem.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In *Proceedings of the second international symposium on information theory*, pages 267–281.

Bjørnstad, Å., Westad, F., and Martens, H. (2004). Analysis of genetic marker-phenotype relationships by jack-knifed partial least squares regression (PLSR). *Hereditas*, 141:149–165.

Caliński, Z., Kaczmarek, P., Frova, C., and Sari-Gorla, M. (2000). A multivariate approach to the problem of QTL localization. *Heredity*, 84:303–310.

Golub, G. H., Hansen, P. C., and O'Leary, D. P. (1999). Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.*, 21(1):185–194.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press, 3^{rd} edition.

Hanke, M. and Hansen, P. C. (1993). Regularization methods for large-scale problems. *Surv. Math. Ind.*, 3:253–315.

Hansen, P. C. (1994). Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems. *Numerical Algorithms*, 6:1–35.

Henshall, J. M. and Goddard, M. E. (1999). Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics*, 151:885–894.

Hoh, J. and Ott, J. (2003). Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Review Genetics*, 4:701–709.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. John Wiley & Sons.

Jiang, C. and Zeng, Z.-B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140:1111–1127.

Knott, S. A. and Haley, C. S. (2000). Multitrait least squares for quantitative trait loci detection. *Genetics*, 156:899–911.

Lee, D. D. and Seung, H. S. (2001). Algorithm for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, pages 556–562.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4):413–417.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solution of Ill-Posed Problems*. Wiley, New York.

Toivonen, H. T. T., Onkamo, P., Vasko, K., Ollikainen, V., Sevon, P., Mannila, H., Herr, M., and Kere, J. (2000). Data mining applied to linkage disequilibrium mapping. *The American Journal of Human Genetics*, 67:133–145.