Methods for Analysis and Visualization of SNP Genotype Data for Complex Diseases

A. Tsalenko, A. Ben-Dor, N. Cox, Z. Yakhini

Pacific Symposium on Biocomputing 8:548-561(2003)

METHODS FOR ANALYSIS AND VISUALIZATION OF SNP GENOTYPE DATA FOR COMPLEX DISEASES

ANYA TSALENKO[1], AMIR BEN-DOR[1], NANCY COX[2] AND ZOHAR YAKHINI[1]

[1] *Agilent Laboratories, 3500 Deer Creek Road,Palo Alto CA, 94304.*
E-mail: anya_tsalenko@agilent.com
amir_ben-dor@agilent.com
zohar_yakhini@agilent.com

[2] *Department of Human Genetics, University of Chicago, 920 East 58th Str., Chicago, IL 60637.*
E-mail: ncox@genetics.bsd.uchicago.edu

SNP markers are becoming central for studying genetic determinants of complex diseases. Large SNP data collected in such studies call for the development of specialized analysis tools. We present methods for selecting sets of SNPs that can be associated to sample properties in case/control studies. We also describe how scoring and selection can be statistically tested. This is done at the single locus as well as at the set level.

1 Introduction

Much of human DNA sequence variation is due to *single nucleotide polymorphisms* (SNPs), which are single base pair positions in genomic DNA at which different sequence alternatives (*alleles*) exist in normal individuals in some population(s). They are distinguished from rare sequence variations by a requirement for the least abundant allele to have a frequency above 1% in the population. The density of SNPs differ between different genomic regions and different populations, but the overall frequency, for the global human population, is estimated to be around 1 in every 200-600 base pairs. The importance of SNPs in various areas of clinical medicine is gaining increasing attention [14]. Association studies, using polymorphic markers (such as SNPs), in genome-wide scans have been advocated as the most efficient way of identifying genetic regions or genes implicated in common complex diseases and traits [15]. The collection of SNP variants/allelles that an individual possesses in a number of key genes (termed his/her *genotype* over this set of loci) are assumed to play an important role in conferring drug response variability. Therefore, *pharmacogenomics* (and other) association studies are expected to reveal sets of SNPs that separate phenotypically distinct classes of samples according to their genotype signatures.

The sequencing of most of the human genome, the improved understanding of this sequence, mostly of its coding part, and the fast development of parallel measurement platforms such as microarrays are three important recent advances in molecular biology. The combination of these advances drive an increasing interest and activity in measuring gene expression profiles of different cell types and disease stages or types as well as in understanding the role of human sequence variation in influencing disease and treatment susceptibility. Gene expression profiling data are accumulating at a fast rate and the association between profile properties and clinical attributes is being explored [13,5,10,17]. Such studies reveal sets of genes that separate phenotypically distinct classes of samples according to their expression signatures. The study of naturally occurring DNA sequence variations and the relationship between genetic variants and clinically meaningful phenotypes precedes the interest in expression profiling by many years. The recent developments mentioned above, however, bring them together and allow for the exploitation of common characteristics and for the study of joint properties.

In this paper we describe statistical methods, visualization tools and algorithmic approaches to questions that arise in pursuing correlations between SNPs as well as *sets* of SNPs and sample properties. Some of the methods draw on the common characteristics of expression data [2,4,3] and genotyping data in case/control studies.

A pioneering effort to positionally clone a gene that affects susceptibility to type 2 diabetes in Mexican Americans is reported by Horikawa *et al* [12]. The authors show that certain combinations of polymorphisms in the gene encoding calpain-10 are associated with the risk of type 2 diabetes. By indicating a role for a calpain protease, these findings propose a fundamentally new hypothesis for diabetes research.

This study exemplifies the long process and the various stages involved in finding genetic determinants of any clinically meaningful condition. Significant evidence for linkage of type 2 diabetes to the distal long arm of human chromosome 2 was reported in 1996 by Hanis *et al* [9] and the locus was designated *NIDDM1.* The implicated region was large, with the 1-lod support interval, which is expected to contain the responsible gene in 80 to 90% of the cases, extending over 12 cM. In later studies this region was narrowed down to 7 cM, corresponding, in this case, to a relatively short span of 1.7 million base pairs, rather than the expected 7 million [a]. Finding the causative gene(s) in 1.7 Mb of sequence is a difficult task: this particular interval contains at least

[a] the average ratio of physical distance to genetic distance across the human genome is approximately 1 million base pairs per cM

7 known genes and 15 ESTs; none of these are obvious candidates. Horikawa *et al.* [12] chose to screen polymorphisms in the region for association with diabetes, relying on linkage disequilibrium (LD). Further investigation of the results of the said screening led to genotyping 63 SNPs in a larger set of about 100 diabetic cases and controls. The authors applied simulation based statistical tests to the results and implicated *CAPN10* as associated with increased risk of diabetes. We demonstrate our methods on data that further extends this study, finding interactions between genes in different chromosomes and identifying a specific set of SNPs and a specific genotype profile for these that helps explain evidence for linkage in the *CAPN10* region.

The main contribution of the current work is in providing methods for selecting and for statistically benchmarking sets of SNPs (as opposed to single SNPs) that jointly associate with a property of interest. An approach to identifying sets of SNPs, associated with disease, is described by Ott *et al* in [11]. In this pioneering work SNPs were ordered by a score that combined allele association, Hardy-Weinberg equilibrium and evidence for genotyping errors. Contributions from the highest scoring SNPs were combined to form a single genome-wide test. The statistical significance of the scores is assessed by simulations. Our approach differs in the way we score individual SNPs and sets of SNPs and in the way we model and compute the related statistics. We also apply less greedy selection methods. These typically perform better when the features are highly dependent.

The current paper and [11] share an emphasis on rigorous and critical statistical assessment of conclusions based on the data. As genotyping and association studies explore new frontiers it is important that methods remain grounded in sound statistics.

The paper is organized as follows. We start by describing the running example data, in Section 2. In Section 3 we describe an information theory driven method for scoring SNPs for their relevance to a partition of the set of samples. Assigning statistical meaning to this score is also discussed and an application introduced in Section 4. Data visualization is demonstrated in Section 5. Finding SNP *s*et association, including a statistical test, is the topic of Section 6. We conclude with methods applicable for quantitative traits, in Section 7, and a discussion in Section 8.

2 Data

We analyzed 216 SNPs typed in Mexican-Americans (from Starr County, Texas) with type 2 diabetes [12]. Of these, 88 SNPs were on chromosome 2 in the *NIDDM1* region, 63 SNPs were on chromosome 15 in the CYP19 region and 65 SNPs were on chromosome 7. In previous studies these regions had shown at least nominally significant evidence for linkage [9], moreover it was shown that there is statistical interaction between genes on chromosomes 2 and 15 [7]. The study consisted of 170 families, 330 possible affected sib-pairs. One patient from each affected sib-pair was selected into a set of representatives for the analysis, but not all representatives from all families were typed for all SNPs in the study. 108 families were typed for SNPs on chromosomes 2 and 15, one member from each of 96 families was typed on chromosome 7. The overlap between families typed on chromosomes 2, 15 and 7 is 57 families. A random sample of 112 individuals from Starr County, Texas not diagnosed with diabetes at the time of the study was also typed for most of the SNPs.

3 Scoring SNP for relevance

Consider a partition of the samples into disjoint classes $C = \{C_1, ..., C_n\}$ (for example affected/unaffected individuals). Denote the number of samples in each class by $d_1, d_2, ...d_n$ respectively, and the total number of samples by D. We want to score each SNP locus l according to its relevance to the partition C. One way of mathematically defining locus relevance (with respect to C), is by the *mutual information score.* Let G denote the partition of the tissues induced by the genotypes at l. The *mutual information* of the partitions G and C is defined as the difference between the measure-theoretic entropy of the partition C and the conditional entropy of C conditioned on G:

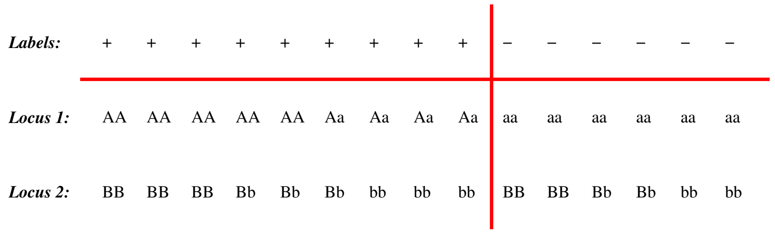
$$s = H(C) - H(C|G), \tag{1}$$

where H is the entropy [6], i.e. $H(C) = -\Sigma_{i=1}^{n} d_i/D \cdot log(d_i/D)$. This score measures the amount of information the genotype at the locus under consideration gives about membership in each one of the sets $C_1, C_2, ...C_n$. Figure 1 shows an example of loci with high and low mutual information scores.

Together with the mutual information score we compute the corresponding significance level (p-value) in the following way. Consider a random assignment of the samples to n groups of the appropriate sizes $d_1, d_2, ..., d_n$. We call such assignments *admissible.* For a locus l, let S be a random variable obtained by computing (1) for a partition uniformly drawn over the set of all admissible assignments and fixed partition G defined by the genotypes at l. Then given a score s we define:

$$p\text{-value}(l, s) = Prob(S \geq s). \tag{2}$$

Note that this way of scoring SNPs, and of computing their p-values, allows to compare relevance of SNPs on the same scale, since it takes into account possible missing genotype data. In the example of Figure 1, the p-value for locus 1 is 0.0002, the p-value for locus 2 is 1. Whenever no confusion arises we will denote p-value(l, s) by p-value(s).



Labels:	+	+	+	+	+	+	+	+	+	–	–	–	–	–	–
Locus 1:	AA	AA	AA	AA	AA	Aa	Aa	Aa	Aa	aa	aa	aa	aa	aa	aa
Locus 2:	BB	BB	BB	Bb	Bb	Bb	bb	bb	bb	BB	BB	Bb	Bb	bb	bb

Figure 1. Illustration of the mutual information score for two classes: one with 9 individuals labels by '+', the other one with 6 individuals labelled by '-'. At locus 1, all people in the first class have genotypes AA or Aa, and all people in the second class have genotypes aa. This locus is informative and has the score of 0.97 (p-value=0.0002). At locus 2, there is no difference between genotype frequencies in different classes, the mutual information score for this locus is 0 (p-value=1).

For modest size data sets like the one we are considering, these p-values can be computed exactly, exhaustively counting over all possible admissible label assignments. For data sets with many samples, for micro-satellite loci or for haplotype data these p-values can be estimated by simulations.

In the diabetes data, we considered several ways to partition the samples. We compared all patients with all controls, and we compared subgroups of patients. One way of partitioning the patients was generated using linkage signal. Patients were partitioned into 3 groups: patients from families that showed evidence for linkage (NPL score $>$ 0.6) in the *NIDDM1* region, patients from families with evidence against linkage (NPL score $<$ -0.6) and patients from families with no linkage signal (-0.6 $\leq$ NPL score $\leq$ 0.6). The first group has 38 patients, the second group has 16 patients and the third group has 54 patients. Analogous partitions were considered using NPL scores on chromosomes 15 and on chromosome 7. We will use the 'linkage' partition L generated by NPL scores on chromosome 2 in the examples below. SNPs with high scores for L may help explain evidence for linkage in this region and point to genes related to diabetes susceptibility.

Results for the highest scoring chromosome 2 SNPs and the partition L are shown in Table 1. We identified 10 informative SNPs on chromosome 2 with p-values less than 0.05. Interestingly, 7 of these are tightly linked

Table 1. Top scoring SNPs from chromosome 2. The nucleotide indicates the location of the SNP relative to the A of the ATG of the initiator Met of the *CAPN10*.

Chromosome number	SNP number	Nucleotide (bp)	Score	p-value
2	30	11098	0.15	0.0058
2	48	15111	0.16	0.0062
2	19	7917	0.14	0.0082
2	59	6018	0.17	0.0099
2	56	5415	0.13	0.0124
2	28	23527	0.12	0.0154
2	65	23527	0.11	0.0269
2	33	(65kb)	0.11	0.0403
2	31	(60kb)	0.10	0.0414
2	43	4852	0.08	0.0451
2	27	(100kb)	0.06	0.0619
2	18	(330kb)	0.08	0.0657
2	36	41509	0.08	0.0807
2	49	41943	0.08	0.0806
2	51	41959	0.07	0.1038

polymorphisms in the calpain-10 gene; others are linked SNPs from the region between *GPR35* and *ATSV* on chromosome 2.

4 Overabundance analysis

To estimate the overall significance of a set of SNPs with respect to a given partition we compare the observed number of SNPs with score $\geq s$ for each score level s with the expected number of SNPs with scores $\geq s$ for random admissible labellings of the samples. The higher the gap is between the observed number of significant SNPs (high mutual information score) and the expected number of significant SNPs, the more significant the sample partition is.

At a given score level s, let $p = p$-value(s). Suppose that in the data we observe $N(s)$ SNPs with score $\geq s$. In a data set with N SNPs we expect to see $E(s) = pN$ SNPs with a score better than s. Moreover, the number of SNPs with score $\geq s$ we observe for uniformly and independently drawn labellings is a random variable $n(s)$, with $n(s) \sim Binom(N, p)$. The surprise rate at s is defined as

$$\sigma(s) = Prob(n(s) \geq N(s)) = \sum_{k=N(s)}^{N} \binom{N}{k} p^k (1-p)^k .$$

Finally, the maximum surprise score for the partition is

$$\Omega = \max_s(-\log(\sigma(s))).$$

Figure 2 shows results of overabundance analysis for chromosome 2 SNPs and the partition into diabetes patients and random sample.

Note that the exact numerical value of the partition score Ω should not be taken literally, as it relies on strong statistical assumption (namely, that the different loci are not linked and thus can be treated as independent). Nevertheless, this score is useful when we want to compare two partitions of the samples (to select which is more supported by the SNP data). Another application for the partition score is class discovery [3].

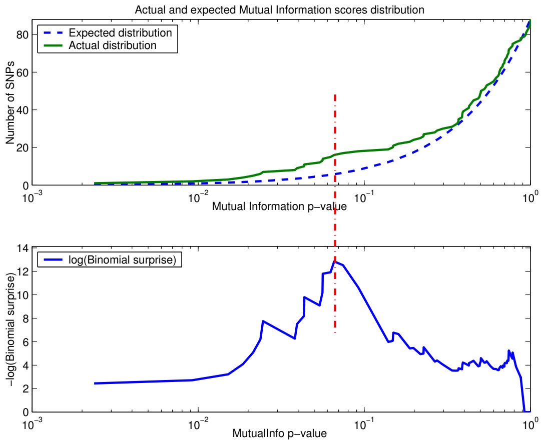


Figure 2. Overabundance analysis of SNPs of chromosome 2 with respect to the affected/random sample partition. In the upper part of the figure we plot the distribution of SNPs p-values with respect to this partition (green curve) vs. the expected distribution of p-values for random admissible partition (blue curve). The gap between the two curves show that there is an overabundance of significant SNPs. The red line (at p-value =0.0659) correspond to the max-surprise score 12.84 (depicted at the lower half of the figure).

5 Visualization of the data

In this Section we present a few figures that allow us to visualize the SNP data with respect to different sample partitions, and different SNP ordering methods [b]. As a result we can highlight different aspects of the data.

[b] These figures were generated by `SNPTool`, part of `BioTools`, a software package developed at Agilent Labs, for internal research and scientific collaborations.

Figure 3 shows data for SNPs from chromosome 2 with highest mutual information scores. Note that this Figure shows that the top SNPs 2_59, 2_48, 2_30, 2_19, 2_19, 2_65 are very similar. Indeed these SNPs are polymorphisms (see Table 1) from the *CAPN10* gene in strong linkage disequilibrium.

Figure 4 shows the same data, but now each row is sorted by genotype with each class. This way of plotting illustrates well the mutual information score. The first 5 SNPs got high scores because individuals from 'not-linked' group do not have homozygous genotypes for the rare allele at these SNPs, plotted in yellow.

SNPs could also be sorted by their chromosomal location, which together with their scores may provide additional insight to the interesting genomic regions that may be related to disease susceptibility.

6 Selection of SNP subsets for classification

Assume we are given a sample partition C (e.g., diabetics/random samples or 'linked'/'not-linked' families). For complex, multi-genic diseases like diabetes, it is not necessarily the case that a single SNP would suffice to explain the genetic origin of the disease. In this Section we describe an approach to select a set of SNPs that *together* contains a strong *genotype signature* of the sample class. To rigorously define the SNP-subset search problem, we need to choose an appropriate optimization criteria, and search for a set of SNPs that maximize it. In this paper we describe a simple approach that is based on the *classification accuracy* of the SNP subset. Intuitively, a good set of SNPs, A, is such that knowing the genotypes over A for an unknown sample will allow us to make a good guess about the class membership of this sample. We first show how to construct a classifier (Naive Bayesian Classifier) using a SNP set A. We shortly describe how the classification accuracy of such a classifier can be estimated (using LOOCV approach). We then describe simple search heuristics to select the set of SNPs, and evaluate the success rate of these methods with respect to the 'linked'/'not-linked' partition. Finally, we conclude the section with a comparison of the best classification accuracy achieved, and a classification accuracy for a random admissible assignment of classes.

6.1 The naive Bayesian classifier

Consider a set of SNPs A. One simple way to construct a classifier is to consider the naive Bayesian classifier based on the probabilistic approach to this problem [8]. For each SNP, we compute the probability of a given label, given genotypes of the training set of samples. Then these one-SNP classifiers are combined together to predict the labels of test samples.

For a sample x with unknown label, in the case of two classes (e.g. 'linked'/'not linked') labelled by '+' and '-', applying Bayes rule to the set of SNPs:

$$\begin{aligned}\log\frac{P(+|x)}{P(-|x)} &= \log\frac{P(+)}{P(-)} + \log\frac{P(x|+)}{P(x|-)} = \log\frac{P(+)}{P(-)} + \sum_{a\in A}\log\frac{P(x_a|+)}{P(x_a|-)} \\ &= \log\frac{P(+)}{P(-)} + \sum_{a\in A}\left(\log\frac{P(x_a|+)}{P(x_a|-)} - \log\frac{P(+)}{P(-)}\right), \qquad (3)\end{aligned}$$

where x_a is the genotype of sample x at locus a. In the above formula we assumed independence of SNP loci in the set A given the partition of the samples. For positive $\log\frac{P(+|x)}{P(-|x)}$, we predict that the label of x should be '+', otherwise the label is '-'.

The accuracy of the classifier can be measured by the number c of correct predictions it makes for the test samples, and we can use it to find the 'best' subset of SNPs. Training and test sets of samples can be defined using leave one out cross-validation technique (LOOCV). On each step of the LOOCV algorithm we 'hide' one sample and construct a classifier using the remaining samples. Then this classifier is used to predict the label of the 'hidden' sample, and the procedure is repeated for every samples in the data. LOOCV method can be modified to hide more than one sample at a time and can be applied to more than 2 classes.

6.2 SNP subset selection

As it is not computationally feasible to exhaustively try all possible SNP subsets we describe here a few simple efficient methods to select SNP subsets.

One approach to find the SNP subset is to order SNPs by their scores, e.g. by mutual information score, and consider classifiers using k top scoring SNPs A_k for each $k = 1, 2, ..., N$. Then, we choose the subset A_{k_0} for which the classifier with k_0 SNPs makes the biggest number of correct predictions $c_{k_0} = max_{1\leq k\leq N} c_k$. In the Mexican-American diabetes data set this approach did not work well, because genotypes at many SNPs are not independent since many SNPs in this data set are very close to each other and are in strong linkage disequilibrium. The number of correct predictions for 'linked'/'not linked' classes was close to the prior probability of making a correct prediction 0.7 (Figure 5). This approach is very computationally efficient, since the calculation time is linear with the number of SNPs, and will probably work best in the data sets with independent SNPs.

Another approach is to select the subset using forward (backward) sequential search [16,1]. On the first step of the forward sequential search, we select a SNP a_1 out of the whole set of SNPs A such that the corresponding classifier makes the biggest number of correct predictions. We set $A_1 = \{a_1\}$. On each step $k = 2, ..., N$, we find a SNP a_k such that the classifier corresponding to the set $A_k = A_{k-1} \cup \{a_k\}$ makes the biggest number of correct predictions among the classifiers with subsets $A_{k-1} \cup \{a\}, a \in A \backslash A_{k-1}$. The 'best' subset is defined by k_0 for which the classifier with the set A_{k_0} makes the biggest number of correct predictions. Backward sequential search works in the reverse direction, i.e. we start with the set of all SNPs, and on each step of the algorithm remove a SNP such that the classifier build using remaining subset of SNPs makes the biggest number of correct predictions.

Using backward sequential search we identified a subset of 11 SNPs from chromosome 2 with combined genotypes that predict the 'linkage status' with 87% accuracy (Figure 5). Similarly, a set of 11 SNPs from chromosomes 2 and 2 SNPs from chromosome 15 was found that predicts the 'linkage status' with 90% accuracy. Another interesting result was found for chromosome 7 and 15 SNPs. Out of 57 families typed for SNPs on both chromosomes, 22 showed evidence for linkage on chromosome 7 (NPL > 0.6), 17 showed evidence against linkage (NPL < -0.6). A set of 7 SNPs from chromosomes 7 and 15 was found that predicts 'linkage status' on chromosome 7 correctly in 38 out of 39 samples (Figure 6).

Note that the number of steps in forward/backward sequential searches may become quite big for data sets with many SNPs. To save on the computational time, we can limit the searches to top scoring SNPs only.

6.3 Statistical significance

We estimated the significance of LOOCV results by simulations, i.e. we simulated random admissible labels of the samples and ran LOOCV with the corresponding set selection method for each labelling. Then we compared the observed probability of the maximal number of correct predictions for random labels with the number of correct predictions for the original labels.

In the current example, 100 random admissible labellings of the samples were simulated and we ran LOOCV with backward sequential search algorithm for these labels. The probability of finding a subset of SNPs on chromosome 2 that better predicts set membership than the original subset is 0.04. Also observe that prediction quality of 'linkage' status is consistently better than the average prediction quality for random labellings (Figure 5).

7 Quantitative traits

In many studies, clinical quantitative information is available for the samples. This information may help define more homogeneous subgroups of patients for which SNP association with disease susceptibility will be easier to detect. Similar to (1), we can score each SNP l, and each quantitative measurement q, in the following way. We find a threshold t such that the mutual information score of l with respect to the partition of samples into samples with $q \geq t$ and samples with $q < t$ is maximal. Analogous to (2) we can compute the p-value for this score, counting over all possible assignments of samples to groups. This computation can be effectively and efficiently carried out using dynamic programming methods, as described in the supplementary information (http://dogbert.cs.technion.ac.il).

8 Discussion

We presented methods for analyzing and visualizing SNP case/control data. In particular we described processes for selecting subsets of loci that jointly correlate with sample properties. The selection process an be compared to a null-model by means of simulations. In future work we will further address this crucial statistical testing. We also briefly discussed similar methods that apply to quantitative traits. We are planning to test these and other means of scoring quantitative trait predictors on biological data. More scientific activity in this space will drive the emergence of appropriate methodology.

Acknowledgments

Supported by DK-55889, DK-47486, DK-20595 and Agilent Laboratories. Authors would like to thank Professor Graeme Bell for useful discussions and SNP genotype data on chromosomes 2 and 15 and Professor Pat Concannon for SNP genotype data on chromosome 7.

References

1. Aha D and Bankert R. A Comparative Evaluation of Sequential Feature Selection Algorithms. *Artificial Intelligence and Statistics, D. Fisher and J. H. Lenz, New York* (1996).

2. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M and Yakhini Z. Tissue classification with gene expression profiles. Journal of Computational Biology **7**, 3-4:559-83 (2000).
3. Ben-Dor A, Friedman N and Yakhini Z. Class Discovery in Gene Expression Data. *Recomb* (2002).
4. Ben-Dor A, Friedman N and Yakhini Z. Scoring genes for relevence. *Agilent technical report*, http://www.labs.agilent.com/resources/techreports.html (2001).
5. Bittner M *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature **406**, 6795:536-40 (2000).
6. Cover T and Thomas J. Elements of Information Theory. *Jon Wiley Sons Inc* (1993).
7. Cox N, Frigge M *et al.* Loci on chromosomes 2 (*NIDDM1*) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nature Gen **21**, 213-215 (1999).
8. Duda R and Hart P. Pattern Classification and Scene Analysis. *New York, Jorn Wiley and Sons*(1973).
9. Hanis C *et al.* A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveales a major susceptibility locus on chromosome 2. Nature Gen **13**, 161-166 (1996).
10. Hedenfalkl I *et al.* Gene expression profiles in hereditary breast cancer. N Engl J Med **344**, 8:539-48 (2001).
11. Hoh J, Wille A and Ott J. Trimming, Weighting, and Grouping SNPs in Human Case-Control Association Studies. Genome Research **11**, 2115-2119 (2001).
12. Horikawa Y *et al.* Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nature Gen **26**, 2:163-75 (2000).
13. Golub T, Slonim D *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**, 5439:531-7 (1999).
14. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science **273**, 1516-1517 (1996).
15. Risch N. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. American Journal of Human Genetics **46**, 242-253 (1990).
16. Pudil P, Novovicova J and J. Kittler. Floating search methods in feature selection.Pattern Recognigion Letters **15**, 1119-1125 (1994).
17. Shipp M *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Med **8**, 1:68-74 (2002).

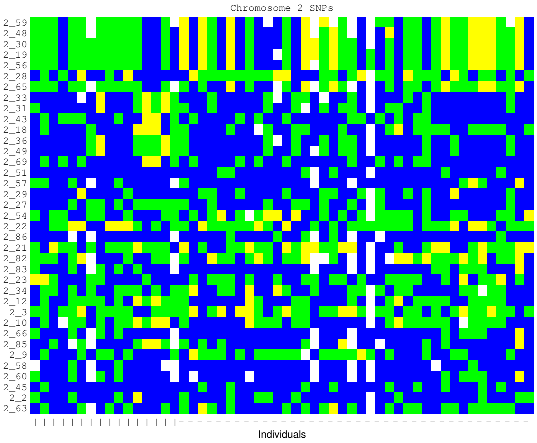


Figure 3. Graphical representation of the highest scoring SNPs from chromosome 2 and 'linked'/'not-linked' partition. Each column represents all genotypes for a given person; each row represents all genotypes for a given SNP. Blue corresponds to homozygous genotype for common allele, yellow corresponds to homozygous genotype for rare allele and green corresponds to heterozygous genotype. White corresponds to missing data. Loci are ordered with respect to mutual information score. Columns marked by '|' on the x-axis correspond to patients from 'not linked' group, columns marked by '-' on the x-axis correspond to patients from 'linked' group.

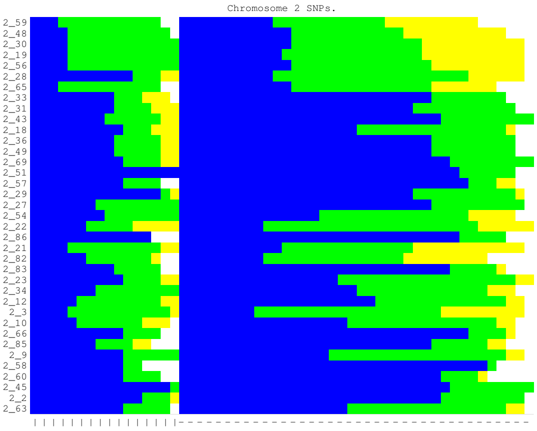


Figure 4. Same data as in Figure 3, now data in each row is sorted by genotypes within each group. This plot helps to visually assess mutual information score, since it is clear that the top 5 SNPs got high scores because individuals in 'not-linked' group do not have homozygous genotypes for the rare alleles at these SNPs. Note that columns no longer correspond to a particular individual.

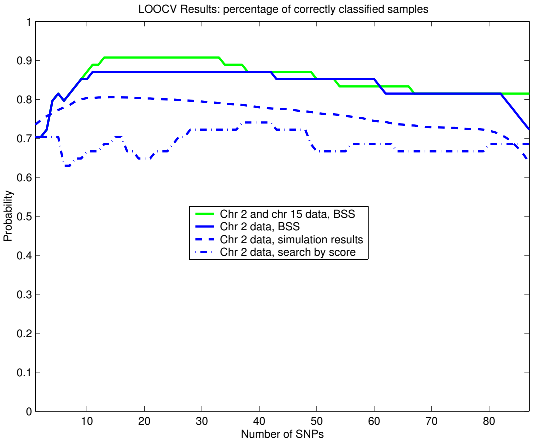


Figure 5. LOOCV results for SNPs on chromosomes 2 and 15. ’Linkage’ status was correctly predicted for 91% of patients, using subset of SNPs identified by backward sequential selection method. Using SNPs from chromosome 2 only ’linkage’ status was correctly predicted for 87% of patients.

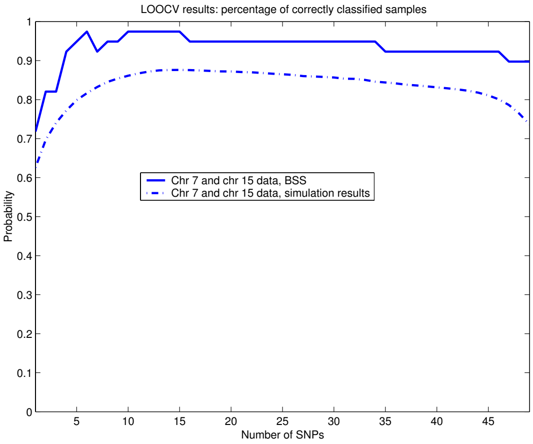


Figure 6. LOOCV results for SNPs on chromosomes 7 and 15. ’Linkage’ status on chromosome 7 was correctly predicted for 98% of patients using subset of SNPs identified by backward sequential selection method.