MIXTURE MODEL FOR SUB-PHENOTYPING IN GWAS

David Warde-Farley[3], Michael Brudno[2], Quaid Morris[2,4], Anna Goldenberg[1*]

[1] *Genetics and Genome Biology, SickKids Research Institute, 101 College Street, Toronto, ON M5G 1L7*
[2] *Department of Computer Science, University of Toronto, 6 King's College Rd., Toronto ON M5S 3G4*
[3] *Départment d'Informatique et de Recherche Operationelle, Université de Montréal, CP 6128 succ. Centre-ville, Montréal QC H3C 3J7*
[4] *Donnelly Centre, University of Toronto, 160 College Street, Toronto ON M5S 3E1, Canada*

Genome Wide Association (GWA) studies resulted in discovery of genetic variants underlying several complex diseases including Chron's disease and age-related macular degeneration (AMD). Still geneticists find that in majority of studies the size of the effect even if it is significant tends to be very small. There are several factors contributing to this problem such as rare variants, complex relationships among SNPs (*epistatic effect*), and heterogeneity of the phenotype. In this work we focus on addressing phenotypic heterogeneity. We introduce the problem of identifying, from GWAS data, separate genotypic markers from overlapping mixtures of clinically indistinguishable phenotypes. We propose a generative model for this scenario and derive an expectation-maximization (EM) procedure to fit the model to data, as well as a novel screening procedure designed to identify skew specific to certain phenotypic regimes. We present results on several simulated datasets as well as preliminary findings in applying the model to type 2 diabetes dataset.

Keywords: GWAS; mixture modeling; sub-phenotyping

1. Introduction

With the current revolution in genome sequencing technologies many of the challenges in acquisition of genomic data and discovery of genomic variants have been overcome, supplanted by the challenge of understanding and assigning functions to the discovered variants. Genome Wide Association Studies (GWAS) have shown promise for discovery of disease-related variants, where usually large cohorts are used to find correlations between genomic variants (typically Single Nucleotide Polymorphisms, or SNPs) and disease risk or prognosis. Despite many successes, in majority of studies the size of the effect tends to be very small. Obtaining larger cohorts, which is rather costly, does not necessarily yield much better results. There are several factors contributing to this problem such as rare variants, complex relationships among SNPs (epistasis), and heterogeneity of the phenotype. While current efforts are targeted at collecting rare variant data and there is rich literature in modeling epistatic effect,[1,2] heterogeneity of the phenotype is often overlooked. Misclassification of disease phenotypes, especially when multiple distinct phenotypes are classified as a single disease, where each phenotype could potentially be modulated by different genetic factors, is prominent in cancers and spectrum disorders, such as Autism Spectrum Disorder (ASD).[3] In the presence of such misclassification, strong genotypic effect in a small homogeneous sub-population will appear as very small or even negligible in the whole population. Analyzing richer, continuous phenotypes such as age of onset, as opposed to a binary variable (disease, no disease) may help with discovery of those more homogeneous phenotypic subgroups.

[*] to whom correspondence should be addressed

A similar situation exists in diabetes: while Types 1 and 2 typically affect children and adults respectively, Maturity-Onset Diabetes of the Young (MODY) is a separate genetic disease that affects patients between 30 and 50 years old. Young MODY patients are often misdiagnosed as having Type 1 diabetes, if a patient is older and obese – as Type 2, and gestational if the patient is pregnant.[4] If we view the age of onset of various types of diabetes together, we observe a mixture, forming distinct but overlapping groups, caused by different genetic (and environmental) conditions. In this paper we develop methodology for simultaneous sub-phenotyping and phenotype-genotype association of such mixtures.

Simultaneous study of quantitative phenotypes and their underlying genotypes also form the basis for identification of expression quantitative trait loci (eQTLs[5,6]). The recent paper of Kim and Xing[7] models the correlation among multiple genes in a Quantitative Trait Network, and identifies SNPs that affect the expression of multiple genes that are tightly connected within the network. Other research[8] stratifies the patient population into groups based on ethnic heritage with assumption that identical phenotypes may be caused by different SNPs in different populations. Our work addresses the case where the phenotype is composed of a set of clinically indistinguishable, but distinct continuous traits, as in spectrum disorders. Both works, ours and,[8] draw additional power by subdividing the population, however our model takes into account information contained in a continuous phenotypic variable and determines genetic variants simultaneously with data partitioning. There have been a few major works that deal with data partitioning according to a phenotypic variable.[9–11] These works perform data stratification (akin a screening step proposed in our paper), mostly deal with discrete traits and do not propose a generative model of the data.

In this work we assume that there are multiple genetic diseases, each caused by a single variant, where each disease corresponds to a Gaussian-distributed quantitative phenotype (e.g., age of onset of the disease), with an unknown mean and variance. The diseases cannot be sub-divided, as they are clinically indistinguishable, hence the phenotype of all patients can be modeled as a mixture of Gaussians. We jointly model the cohesiveness of the sub-phenotypes (likelihood of the Gaussians explaining the phenotypes) and associations (SNPs explaining the sub-phenotypes). We optimize our model using expectation-maximization, and present the results of applying our model to simulated data and to the analysis of Gene-Environment Association Studies (GENEVA) genome-wide association scans for type 2 diabetes.[12]

2. Methods

To address the problem of identifying causal variants from phenotype mixtures we developed an algorithm SNPMix. The algorithm consists of three sub-parts: (1) The screening procedure, where we identify a small set of SNPs that are distributed abnormally in a sub-range of the continuous phenotype; (2) an Expectation Maximization procedure that identifies the sub-phenotypic groups, and the SNPs responsible for these, simultaneously; (3) model selection, that identifies a parsimonious subset of the screened SNPs and corresponding sub-phenotypes.

2.1. *SNP Screening*

Genome Wide Association studies typically genotype millions of SNPs. Only a handful of these are assumed to be associated with the disease of interest. Most approaches that look at multiple SNPs simultaneously *screen* the SNPs in order to identify the subset that is most likely to be associated with the disease for further in-depth analysis. Normally, a statistical test such as χ^2 or Fisher's Exact test is used to find SNPs that have significant deviations in observed allele frequencies between the patient and control populations. In our setting our aim is to identify SNPs that have unlikely frequencies in sub-populations with respect to the control. We thus modify the usual screening procedure: we stratify the population in the case cohort according to their phenotypic value in increasing order and test subsets (halfs, thirds, quarters, etc., halting at eighths) of the stratified population against the control. This way we are able to coarsely capture the modes of our hypothesized Gaussian distributions in one or several of those subsets. We retain the lowest of the p-values from tests made across all subsets ($1 + 2 + 3 + 4 + \ldots + 8 = 36$ tests) for each SNP. We assume that we have N patients and S SNPs. The pseudocode for this algorithm is presented in Table 1.

Table 1: SNP screening procedure

```
INPUT:    list of SNPs to be screened in $N \times S$ matrix:
                      $N$ – patients, $S$ – snps
          test level threshold $\alpha$
          z - phenotype variable
OUTPUT:   $M$ SNPs – a small subset of the original SNPs

(1) Sort people according to z in ascending order
(2) For each SNP
    (a) calculate the frequencies of 0's, 1's, 2's in the control sample
    (b) for each contiguous (as sorted by phenotype) half, third, fourth, fifth, sixth, seventh
        and eighth of the patients in the case sample
        i. calculate the frequencies in this portion of the case data
        ii. perform a $\chi^2$ test against the control
    (c) record the minimum $p$-value over all splits
(3) Rank SNPs according to their $p$-values
```

We perform a rough correction based on the Benjamini-Hochberg FDR[13] for the number of SNPs being tested. It is an approximation, since we do not correct for the 36 dependent tests per SNP. The goal of the screening is merely to obtain a superset of the responsible SNPs, and in practice, models fit to real data will include the top K most significant p-values rather than thresholding at a particular significance level.

2.2. *SNPMix Model*

Given the candidate set of screened SNPs, the core of the SNPMix algorithm simultaneously identifies the sub-phenotypic groups and the SNPs responsible for each group by modelling the phenotype and SNPs jointly. We assume that we have M SNPs remaining after the screening procedure, and a continuous phenotype z that follows a mixture-of-Gaussians distribution where each component represents a sub-phenotypic group. For each individual, a given SNP is encoded as a number $j \in \{0, 1, 2\}$, where j is the count of minor alleles: 0 indicates that the given genomic locus is homozygous major, 1 is heterozygous, and 2 is homozygous minor. Thus SNP information is represented as an $N \times M$ table, each entry being 0,1 or 2. Each SNP is modeled as a multinomial distribution. Since we do not know which sub-phenotypic component each patient belongs to, we introduce a latent variable c responsible for the sub-phenotypic class assignment which has as many states as there are components (K). The graphical representation of the model in Figure 1 captures our notion of conditional independence: information about the class de-couples the two, making phenotype and SNP frequencies dependent only on the class and independent of each other. The SNPs that are not corresponding to any sub-phenotypic groups are independent of the phenotype. To summarize, we have N patients, M SNPs, J SNP values ($J = \{j : j \in \{0, 1, 2\}\}$). The phenotype variable z is represented by sets of mean and standard deviation $\{(\mu_k, \sigma_k)\}$ parameters, where $k \in [1 \ldots K]$. We can then write our two conditional distributions as follows:

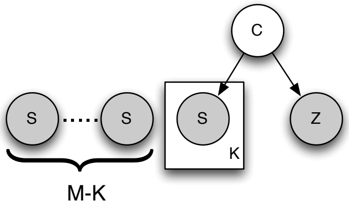


Fig. 1: Graphical model representation. z – phenotype, s – SNP, c – sub-phenotypic class assignment. Our generative model is different from the natural direction of causality, where SNPs cause disease ($s \rightarrow z$, in our notation). What we capture instead is that heterogeneity in the data (sub-phenotypic indicator c) confounds the notion of the dependence of phenotype on genotype. Given this information, we can disambiguate the dependence.

$$\begin{aligned} p(s_i^n|c = k) &= \prod_{j=1}^{J} {p_{ijk}}^{\delta_{j,s_i^n}}, \quad i \in [1, S] \\ p(z_n|c = k) &= \mathcal{N}(\mu_k, \sigma_k^2) \end{aligned}$$

s_i^n indicates the value of SNP $i \in [1 \ldots M]$ for person $n \in [1 \ldots N]$, δ_{j,s_i^n} is an indicator function which takes the value of 1, if $s_i^n = j$ and is 0 otherwise and $\{(\mu_k, \sigma_k)\}$ are parameters for each of the K mixture components.

Then the joint probability can be written as

$$p(\vec{s}, \vec{z}, c = k^n) = \prod_{k=1}^{K} \left(\alpha_k \mathcal{N}(\mu_k, \sigma_k^2) \prod_{i=1}^{S} \prod_{j=1}^{J} {p_{ijk}}^{\delta_{j,s_i^n}} \right)^{\delta_{k,k^n}} \qquad (1)$$

where α_k is a mixture proportion of the k^{th} component, k^n is the class of the person n and δ_{k,k^n} is an indicator function which is 1 if $k = k^n$, $S \in [K \ldots M]$ is the current number of SNPs in the model, generalizing to the case where each group can have more than one associated SNP. Let $\Theta = \{\{p_{ijk}\}, \vec{\alpha}, \vec{\mu}, \vec{\sigma}\}$. All of the Θ parameters are estimated as indicated below.

2.3. *Optimization and Model Selection*

We optimize the parameters of the latent variable model developed in the previous section using an expectation-maximization (EM) algorithm by fixing a posterior $Q = p(c|\vec{s}, z; \Theta^{old})$ in the E step and optimizing the *expected complete log likelihood* under Q:

$$\mathbb{E}_Q[\ell(\Theta)] = -\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log\sigma_k + \frac{(z_n-\mu_k)^2}{2\sigma_k^2}\right) + \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log\alpha_k + \sum_{s=1}^{S}\sum_{j=1}^{J}\delta_{j,s_i^n}\log p_{ijk}\right) \quad (2)$$

where r_{nk} is the responsibility (i.e. $p(c = k|\vec{s}_n, z_n; \Theta^{old})$) assigned to person n by component k, i.e. the model's estimate of the probability that patient n is from the component k. Note that the expectation now looks very similar to the standard mixture of Gaussians model,[14] where the means and the variances of the Gaussian components depend on the multinomial distributions of the SNPs through the responsibilities.

To obtain the M-step update equations, we differentiate Equation 2 and solve for the parameters which jointly maximize the likelihood given the current estimate of the responsibilities. The updates for μ_k, σ_k and α_k are the standard mixture of Gaussians updates, with p_{ijk} being updated as follows:

$$p^{t+1}_{ijk} = \frac{\sum_{n:s_i^n=j} r^t_{nk}}{\sum_m r^t_{mk}} \qquad\qquad p^{t+1}_{ij,k\neq i} = \frac{\sum_{k\neq i}\sum_{n:s_i^{(m)}=j} r^t_{nk}}{\sum_{k\neq i}\sum_m r^t_{mk}}$$

The second equation corresponds to the probability of each SNP's out-of-component distribution, which is the same for all of the components that the SNP is not associated with.

Because the screening procedure results in a selection of a large set of SNPs a subset of which is likely to be responsible for the sub-phenotypic components, we need to perform model selection to eliminate the false positives. In this work, we select $K \leq M$ SNPs that represent sub-phenotypic groups, one per group. We use the greedy model selection procedure described below to select the SNPs, though extensions where many SNPs represent a single component and other model selection techniques, e.g. BIC, are possible. The SNPs that have been selected but are not modeled as associated with the phenotype (indicated by a set of independent SNPs $M' = M - K$ on Figure 1) result in an additive component to our log likelihood:

$$\sum_{n=1}^{N}\sum_{i=K+1}^{M'} \log\begin{pmatrix}2\\ s_i^n\end{pmatrix} + s_i^n\log\gamma_i + (2 - s_i^n)\log(1-\gamma_i) \quad (3)$$

where γ_i and 1 - γ_i here correspond to the usual Hardy-Weinberg p and q values[15] and can be estimated once, in closed form, prior to running EM.

The set of informative SNPs in the model will contain those most relevant to the phenotype, however much of it can be explained by genetic factors. The false positive SNPs found through screening should not explain more than a small proportion of the data. We thus propose a greedy procedure which iteratively removes the SNP that explains the least amount of data. The iterative procedure starts by fitting a component to all of M screened SNPs, and runs the EM algorithm once to obtain the joint model corresponding to the screening result. We remove the SNP where the corresponding component explains the lowest fraction of the data, and iterate with the smaller number of SNPs. The procedure stops when all of the components explain more than a preset threshold, depending on the expected number of components.

3. Results

We will first illustrate our approach on a simulation study and then report our findings on the type 2 diabetes dataset. In all of the experiments below we set $\alpha = 0.05$ and the minimum explained threshold to $\max(1/(K' + 2), .1)$, where K' is the expected number of components (e.g., if the number of components is expected to be 3, the minimum explained threshold is .2, i.e. each component has to explain at least 20% of the data to be accepted).

3.1. *Simulation*

To simulate our data, given $N = 1000$ individuals, $M = 100,000$ SNPs and K Gaussian subphenotypes we

(1) Sample q - the frequency of the minor allele in the population following the empirical distribution observed by the HapMap population as discussed in[16]

(2) Generate two $N \times M$ SNP tables for cases (patients) and controls (healthy population) according to Hardy-Weinberg (HW) principle[15]

(3) Generate mixture proportions uniformly U(a,b) on the interval [.1, .8] (we will use U(.1,.8))

(4) Sample individuals from the mixture to belong to K Gaussians

(5) Sample K SNPs from M to be responsible for each of the K components

(6) For each of the K components

    (a) Generate phenotype z from k^{th} Gaussian with random mean sampled from U(0,K+1) and variance from U(1,2)

    (b) Sample k^{th} SNP values from the empirical multinomial of the control population: with probability .9 inside the Gaussian and with probability .1 – outside.

We refer in the following discussion to "true positives" as those simulated SNPs that are indeed mixture-modulated in the simulated ground truth and are discovered as such by the iterative model selection procedure, and "false positives" as those SNPs assigned a mixture component despite not being generated as such in the simulation.

3.1.1. *Robustness of the model screening procedure*

We first tested the ability of the screening procedure to capture simulated SNPs by generating data from our model for $K = 2$. One would expect to capture subphenotype-modulating SNPs in groups that are not too small and do not have very high overlap. Indeed, the screening procedure contains a sorting step where high overlap between components would imply dilution of the signal for each of the corresponding SNPs. The results are shown in Figure 2. The screening process is able to perfectly recover mixture-dependent simulated SNPs for components with up to half the size of the case cohort with up to 30% overlap. Furthermore, with a slight loss in accuracy we can recover SNPs with up to 70% overlap.

We have then tested the effect of the potentially inaccurate screening procedure on the results of the SNPMix by considering simulated data with 3 to 7 components ($N = 1000, M =$

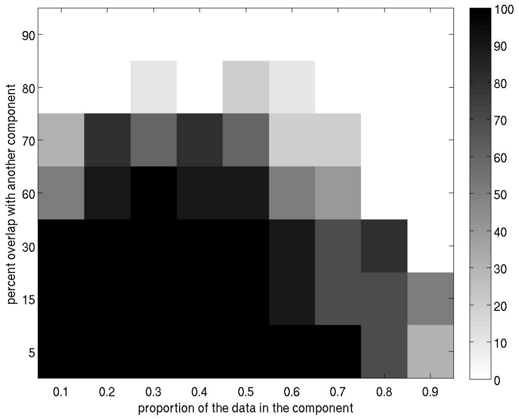
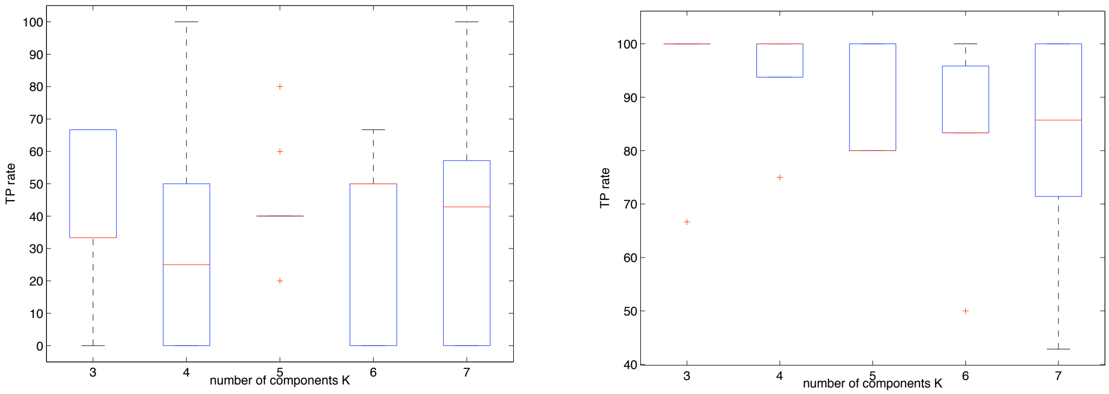


Fig. 2: Heatmap indicating whether we are able to recover true (simulated) SNPs given the size and the overlap of the components

100, 000 as before), running 10 trials for each scenario. Figure 3 shows True Positive (TP) rates for SNPMix, on the default set of screened SNPs (those that pass the Benjamini-Hochberg threshold at the time of screening) shown on the left, and a set that includes all of the true positive SNPs in addition to the SNPs found through Screening step whether or not the true SNPs were found by screening procedure (on the right), i.e. if a true SNP was not present in the set of SNPs after screening it was imputed before SNPMix was run. As demonstrated in Figure 3a SNPMix+Screening achieves on average a 40-50% TP rate on the SNPs originally obtained through screening procedure, which compares well with $K/100,000 < 10^{-5}$ baseline prediction accuracy. It appears that the reduced TP rate is due to the inaccuracy of the screening procedure: when all of the true positive SNPs are made available to the model, the TP rate increases to 80-100%, as is illustrated in Figure 3b.

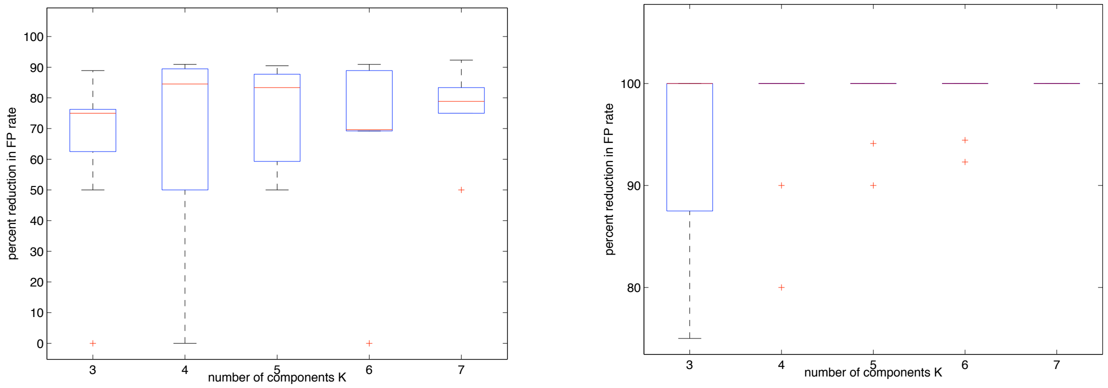


(a) using original results of the Screening step (some true positives might be missing)

(b) using SNP set after Screening augmented to insure the presence of all true positives

Fig. 3: True Positive rate for K=3..7 for the Screen+SNPMix procedure

Figure 4 shows the reduction of false positive rate after running SNPMix, on the same set of simulation data as described above. Comparing Figures 4a and 4b we see that in cases where not all true positives are available, the SNPMix performs slightly worse in reducing FP rate, trying to account for variation in the phenotype with some of the better fitting false positives. To the contrary, in the case where all true positives are available after the Screening step, we observe a 100% reduction in FP rate in most cases.



(a) using original results of the Screening step (some true positives might be missing)

(b) using SNP set after Screening augmented to insure the presence of all true positives

Fig. 4: Reduction in false positive rate for K=3..7

3.1.2. *Robustness of the model selection procedure*

Finally, we have attempted to quantify the effectiveness of our model selection procedure. Figure 5 demonstrates how the number of true (TP) and false (FP) positives (represented by red and blue lines on the graph) vary on the simulated data with K=3 SNPs (assuming all true positives are available) as a function of the size of the component (the x-axis on Figure 5 shows the proportion of the minimum number of the patients that have to be in the component for the component to be valid to the total number of patients). If any of the three, in this case, component sizes fall below the allowed threshold of the number of patients per component (as depicted on the x-axis), the component and the corresponding SNP are no longer considered as valid and are added to the pool of SNPs that are not associated with the phenotype.

Figure 5 represents the tradeoff between the number of false and true positives. When the threshold is very stringent, e.g. we require each component in the model to explain at least 33% of the data (for $K = 3$ this means that it is not very likely that all three components will pass the threshold), there are no false positives, but we are not capturing all true positives either. Whereas if we set the threshold too low, the number of false positives might not justify the certainty in retaining the true positives. Fortunately, there is a range (between .15 and .2 for $K = 3$) where the number of false positives is relatively low while all true positives are captured. Such analysis can be performed for varying K to estimate a proper threshold for our model selection procedure prior to running the SNPMix on the real world data.

So far we have not dealt with the case where there is no genotypic variant in our data that could explain the phenotype. This case is illustrated on the Type 2 diabetes dataset.

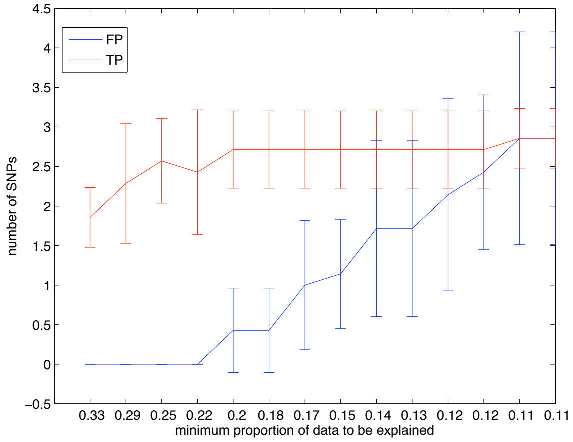


Fig. 5: Average number of SNPs in the model after running a SNPMix procedure for varying minimum-explained thresholds ($K = 3$). Error bars represent 1 std deviation.

3.2. *GENEVA Type 2 diabetes data*

We applied our methods to the Gene Environment Association Studies Initiative (GENEVA)[b] Diabetes.[12] The study contains extensive phenotype and genotype data for 6,000 patient and control individuals. Both phenotype and genotype data was obtained from the dbGAP database.[17] Here we study the $2,502$ male subjects (members of the Health Professionals Follow-up Study, HPFS), genotyped by the Affymetrix 6.0 array, including $1,161$ diabetes patients and $1,341$ controls. By observing the distributions of the various quantitative phenotypes in the study, we chose to model the patient age (at the time of the study) as our quantitative phenotype, as it appeared to be multimodal (Figure 6 hints at the presence of at least two mixture components, supported by fitting a mixture of two Gaussians). For younger patients the age of participation in the study is indicative of the age at diagnosis. Unfortunately, the information on the age at the onset of the disease was not available in this data.

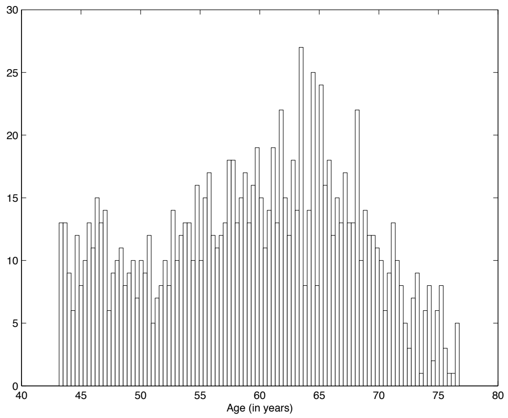


Fig. 6: Age distribution of patients in the GENEVA HPFS type 2 diabetes study

[b] http://www.genevastudy.org/

We first applied the screening procedure and selected the top 25 SNPs. In these 25 SNPs we found 2 regions of linkage disequilibrium (high correlation between SNPs in close proximity on the genome). We chose a single representative SNP for each region, resulting in 20 SNPs total. Figure 7(a) shows the responsibilities after applying the SNPMix procedure, setting the minimum explained threshold at 10% (the approximate size of the younger sub-group). The analysis revealed that the smaller subgroup of younger patients is explained by a single SNP, while a significantly larger number of SNPs are used to explain the larger component. This may be due to a complex relationship among SNPs jointly explaining the larger component (epistasis), not modeled by our method, or to a lack of genetic basis for the larger component. It is also possible that the larger component consists of a very large number of distinct Gaussian distributions. To eliminate the latter possibility, we ran 100 randomized trials of our algorithm, and found that for most individuals in the middle age group the maximum over all trials of the maximum responsibility learned by any of the components is low. At the same time, the model has much higher certainty about the SNPs corresponding to the younger (and older) cohorts, as is illustrated on Figure 7(c). To model a large sub-component with no significant genetic basis we added a "free" Gaussian, that is not explained by any genotype, with mixture proportion based on the fraction of patients for whom the average over 100 randomized trials of the maximum responsibility was below 50% (0.76). We then ran our SNPMix procedure allowing for components that explain at least 10% of the phenotype. The iterative SNPMix procedure converged on 3 Gaussians as can be seen in Figure 7(b), with 2 Gaussians corresponding to SNPs (dbSNP ids rs2002520 and rs4465650), as well as the free Gaussian.

Analyzing the SNPs we found that the minor allele of SNP rs2002520 is rare: in the 934 HapMap individuals genotyped at this position in HapMap, the frequency of the G allele was 0.05, and not a single homozygous minor individual was observed. The allele frequencies, however, were significantly different in the younger sub-group (130 patients), with an allele frequency of 0.09, and 4 individuals carrying the homozygous minor allele. This SNP is located in a region that is annotated as an enhancer by ENCODE based on CHiP-seq of Histone mark. Proximity to a DNase hypersensitivity site indicates a possible regulatory role of the region. The locus is annotated as a human body-mass index QTL by the Rat Genome Database, linked to obesity. Obesity is a common risk factor for type 2 diabetes, so our analysis possibly uncovered a SNP that affects an early onset of diabetes due to this additional risk. None of the genes in immediate proximity of the SNP are linked to diabetes, thus additional analysis is necessary to confirm the role of this variant in the disease.

rs4465650 associated with the older group is harder to justify. It could be a protective SNP that delays the onset of type 2 diabetes in people who have a certain variant of the SNP even if they are predisposed to the disease. However, since we do not have the information on the onset of the disease, we cannot know that these people have not had undetected diabetes for a while. To test whether this SNP is simply related to longevity, we would need to perform and hypothesis test on the controls with matching age distribution, which was unfortunately not available at the time the experiments were performed.

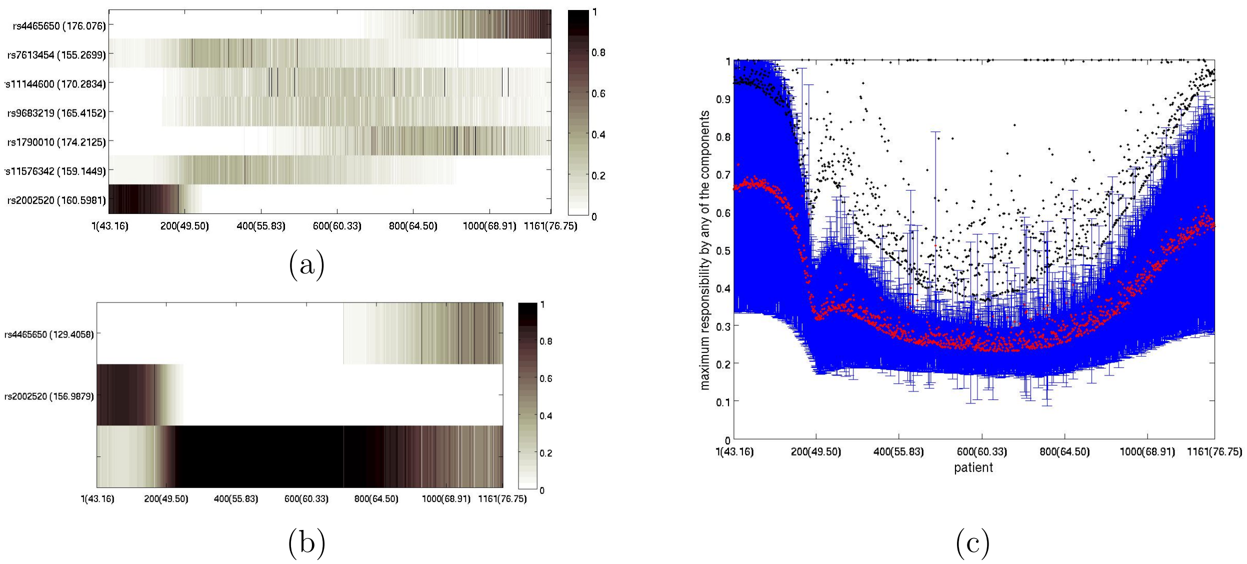


Fig. 7: SNPMix result of modeling age phenotypic variable in the GENEVA type 2 diabetes study starting from 20 independent pre-screened SNPs. (a) Y-axis indicates SNP name corresponding to each component with total responsibility in parenthesis, X-axis indicates patients sorted by age; (b) SNPMix including a phenotype-only (free) component; (c) maximum (in black), mean (in red) and standard deviation (in blue) responsibility for each patient over a 100 runs of SNPMix.

4. Discussion and Conclusion

In this study we consider a novel problem setup, where several clinically indistinguishable phenotypes are assumed to have different underlying associated genotypic variants. This problem is especially inherent in what is known as spectrum disorders, where the case cohort is comprised of several different sub-diseases, that are not easy to identify clinically. The signal for each of the sub-diseases is too weak to be revealed in the full cohort of patients. We believe that it is precisely this reason that led to very weak signal in many known GWA studies. Frustrated with the current state of standard GWAS performance, clinicians are becoming increasingly interested in more complex machine learning approaches, where fewer assumptions are made. Our probabilistic model and the associated expectation-maximization algorithm can recover dependencies on synthetic data where the causal relationships are known. We also present preliminary results on applying the method to a case control study of type 2 diabetes patients. Though the public dataset we used was somewhat simplistic (one could almost read a plausible separation of the Gaussian modes from the graph), it is the simultaneous consideration of phenotype and genetic variants that help us to identify the more accurate borders of the sub-groupings. In the field where there are almost no golden standards of sub-phenotyping, these kind of experiments serve as an indication that the method is doing something reasonable with the data. Further work is needed both to extend our model and to improve model-selection and screening procedures.

Our assumption of a single SNP being associated with each mixture component, while simplifying fitting and model selection, may poorly reflect the reality of complex diseases. The model can be altered to allow a single SNP to flexibly account for perturbations in more than one mixture component, as well as for the assignment of multiple SNPs to a single component, but such a many-to-many relationship greatly expands the space of possible models, and overfitting becomes an increasingly prominent worry. Future work should explore strategies for making principled choices about which dependences to retain, with care taken to appropriately regularize the parameters of richer, more flexible models.

Having narrowed our search space through pre-screening, the discrete choice of which SNPs to model as associated spans the space of all possible subsets of pre-screened SNPs, making exhaustive search prohibitive for even a moderate number of SNPs. Of several techniques we explored, including greedy and beam-search variants of model-pruning using standard model comparison criteria (e.g. the Bayes information criterion), none were entirely satisfactory. The pruning strategy we adopted was among the simplest, yet produced results that offered a parsimonious explanation of the data. Nevertheless, more work is needed on determining the most appropriate model selection criteria and optimization procedures to take full advantage of the class of complex genotype-mixed phenotype models studied in this work.

Acknowledgement

This study was funded by NSERC Discovery Grants to MB and QM.

References

1. H. J. Cordell, *Nat Rev Genet* **10**, 392 (Jun 2009).
2. S. K. Musani, D. Shriner, N. Liu, R. Feng, C. S. Coffey, N. Yi, H. K. Tiwari and D. B. Allison, *Hum Hered* **63**, 67 (2007).
3. P. Lichtenstein, E. Carlström, M. Råstam, C. Gillberg and H. Anckarsäter, *Am J Psychiatry* (Aug 2010).
4. A. Stride and A. T. Hattersley, *Ann Med* **34**, 207 (2002).
5. W. Cookson, L. Liang, G. Abecasis, M. Moffatt and M. Lathrop, *Nat Rev Genet* **10**, 184 (Mar 2009).
6. J. J. Michaelson, S. Loguercio and A. Beyer, *Methods* **48**, 265 (Jul 2009).
7. S. Kim and E. P. Xing, *PLoS Genet* **5** (Aug 2009).
8. K. Puniyani, S. Kim and E. P. Xing, *Bioinformatics* **26**, 208 (Jun 2010).
9. X. Qin, E. R. Hauser and S. Schmidt, *Genet Epidemiol* **34**, 407 (Jul 2010).
10. S. Macgregor, N. Craddock and P. A. Holmans, *Eur J Hum Genet* **14**, 529 (May 2006).
11. A. P. Morris, C. M. Lindgren, E. Zeggini, N. J. Timpson, T. M. Frayling, A. T. Hattersley and M. I. McCarthy, *Genet Epidemiol* **34**, 335 (May 2010).
12. C. Zhang, L. Qi, D. J. Hunter, J. B. Meigs, J. E. Manson, R. M. van Dam and F. B. Hu, *Diabetes* **55**, 2645 (Sep 2006).
13. Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society, Series B* **57**, 289 (1995).
14. A. Dempster, N. Laird and D. Rubin, *Journal of the Royal Statistical Society, Series B* **39**, 1 (1977).
15. A. Edwards, *Foundations of Mathematical Genetics* (Cambridge University Press, 1977).
16. International HapMap Consortium, *Nature* **437**, 1299 (Oct 2005).
17. M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z. Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell and S. T. Sherry, *Nat Genet* **39**, 1181 (Oct 2007).