SNPrints: Defining SNP signatures for prediction of onset in complex diseases

Linda Liu, Biomedical Informatics, Stanford University
Daniel Newburger, Biomedical Informatics, Stanford University
Grace Tang, Bioengineering, Stanford University
Emidio Capriotti, Bioengineering, Stanford University (not in CS229)

1. Introduction

Complex genetic diseases are a major cause of human morbidity and mortality, and their prevalence and severity place a tremendous burden on patients and medical facilities [1,2]. Preventive care studies have shown that identification of patients at risk for disease and prediction of patient age at disease onset improve patient outcome and reduce health care costs [2]. However, the development of accurate predictive methods remains in preliminary stages.

Recent research suggests that methods for analyzing combinatorial interactions of single nucleotide polymorphisms (SNPs) can lead to effective predictors for disease [3]. SNPs, which are single allele mutations in the genomic sequence of an organism, are responsible for about 90% of all human DNA variation and play an important role in human evolution, drug sensitivity, and disease susceptibility [4]. Due to advances in high-throughput experimental techniques for SNP identification and the resulting data explosion, several machine learning methods have been applied to study the relationship between SNPs and disease [3,5]. Algorithms such as MegaSNPHunter achieve good performance by avoiding the computationally intractable combinatorial search space, but they are limited by the inability to use a large number of SNPs in disparate genomic locations [3]. Other machine learning approaches have been successfully applied to disease risk prediction using SNP data [5], but these methods have not yet been applied to onset prediction.

Therefore, we will leverage a multiple-SNP approach to create a novel predictive model of both disease risk and age of disease onset. Our project aims to improve performance of disease risk and age of onset assessment, and to bring us closer to personalized preventative treatment for complex diseases.

2. Methods

2.1 Dataset

We have obtained SNP data from genome-wide association studies (GWAS) performed by the Wellcome Trust Case Control Consortium (WTCCC). This dataset is comprised of the SNP genotypes for 3,000 healthy controls and 14,000 diseased patients, all genotyped at 500,568 genomic locations [1]. The patient populations are equally sized for seven complex genetic diseases (Table 1). Age of disease onset is available for three of the diseases and is binned by decade. There is sufficient spread in age of onset to enable subpopulation studies.

Table 1. WTCCC Study participant characteristics**.**

# Patients	Cohort	Age of Onset								
		0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	Unknown
1,998	Bipolar Disorder (BD)	-	-	-	-	-	-	-	-	-
1,991	Coronary Artery Disease (CAD)	0%	0%	1%	12%	39%	40%	8%	0%	0%
2,001	Hypertension (HT)	-	-	-	-	-	-	-	-	-
2,009	Crohn's Disease (CD)	2%	21%	36%	16%	9%	6%	3%	1%	7%
1,999	Rheumatoid Arthritis (RA)	-	-	-	-	-	-	-	-	-
2,000	Type 1 Diabetes (T1D)	-	-	-	-	-	-	-	-	-
1,999	Type 2 Diabetes (T2D)	0%	0%	2%	15%	32%	38%	13%	0%	0%
3,004	Controls	-	-	-	-	-	-	-	-	-

2.2 Data filtering

We have removed patient data and SNP data from our set for the following reasons: 1) patients missing more than 3% of SNP data, 2) genotype calls that disagree between the two calling algorithms used by the WTCCC, and 3) satisfying other exclusion criteria specified by the WTCCC (poor data quality, incorrect genotyping, etc).

2.3 Classification

We used the LIBSVM software package [6] to build support vector machine classifiers. All binary classifiers and multi-class classifiers were built using C-support vector classification, which solves the primal problem (1) with C set to 2. We used a radial basis kernel function (2) with gamma set to 2^{-15}. Parameter values were optimized for a single binary classifier using a grid search over a range of values. Due to the computational complexity of this optimization problem, these parameters were not re-optimized for each classifier.

$$(1)\quad \min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

$$\text{subject to}\quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, i = 1, \ldots, m$$

$$\xi_i \geq 0, i = 1, \ldots, m$$

$$(2)\quad K(x, z) = \exp(-\gamma \|x - z\|^2)$$

2.3.1 Disease risk:

For each disease, we trained a binary classifier to distinguish between disease and control. To balance training set sizes for disease samples and control samples, we selected a random sample of patients from the larger group. In each case, the total training set size averaged around 3600 individuals (half disease, half control).

2.3.2 Early onset risk:

For the three diseases with onset data (CAD, CD, and T2D), we trained a binary classifier to identify patients at risk for early onset. We partitioned the dataset into early onset and late onset groups to train the classifier. To balance training set sizes for early and late onset groups, we used multiple iterations of bootstrapping and ensured performance was not based on the selected individuals.

2.4 Feature Representation

For each SNP, major and minor alleles were defined based on allele frequencies in the control population (where the major allele is the more frequently observed allele). These annotations provided a consistent allele nomenclature for representing all patient SNP vectors. Representation of each SNP required two values, the first of which indicated the presence/absence of genotype information and the second of which encoded the genotype. Presence and absence corresponded to values of 100 and 0 respectively. These values provided a means of accounting for missing data without disrupting our genotype representations. Genotype information (minor/minor, minor/major, major/major) corresponded to values of 100, 50, and 0 respectively. This labeling scheme makes the assumption that the phenotypic effect of a SNP is linearly dependent upon the major (or minor) allele.

2.5 Feature Selection

Our feature space of 500,568 SNPs far exceeds the number of individuals available for training our machine learning algorithms. We therefore reduced our feature dimensionality by filtering SNPs based on strength of disease association and on chromosomal proximity, which prevents overrepresentation of genomic loci. We measured disease association for each SNP by calculating a chi-square p-value for the difference between the SNP genotype distributions of diseased patients versus that of control individuals. These measures of significance allowed us to rank SNPs for feature selection for both disease risk prediction and early onset risk prediction for individual diseases. To filter by chromosomal proximity, we first clustered our set of top ranked SNPs by single linkage clustering using HapMap linkage disequilibrium r^2 values as our pair-wise distances [7]. We then selected the SNPs with the most significant p-value within a given cluster and filtered out all other SNPs.

2.5.1 Disease risk using binary classifiers

As described above, SNPs were ranked by chi-square p-value for each disease, where SNPs with the lowest p-value received the highest ranking. The number of top ranked SNPs selected from each disease was optimized by empirical testing using multiple iterations of SVM training. The top 30 SNPs from T1D and the top 75 SNPs from all other diseases were selected as our final feature vectors by the disease association ranking step. 75 SNPs gave the best performance for all diseases except T1D, in which case more than 30 SNPs did not improve the performance.

2.5.2 Early onset risk using binary classifiers

SNPs were ranked by chi-square p-value, where p-values were calculated based on the SNP genotype distributions for 'early-onset' versus 'late-onset' groups. The top 30 SNPs were selected as our preliminary feature vector. The number 30 was chosen to prevent overfitting, as the smallest training set had around 300 patients. The linkage disequilibrium filtering step was then applied to the preliminary vectors to obtain final feature vectors.

2.6 Validation

We performed 20-fold cross validation on the disease risk and age of onset classifiers discussed above. We calculated performance metrics including prediction accuracy, false positive rate, ROC curves, and AUC (area under ROC curve) to assess classifier performance. We also performed classification with 20 random sets of 24 SNPs (permutation testing) and compared the performance of our feature sets with the random sets. This process allowed us to determine the baseline performance for our learning method and whether our selected features outperformed this baseline significantly.

To evaluate the biological significance of our SNP profiles for disease risk and age of onset prediction, we built a pipeline to identify genes, pathways, and other biological features associated with our SNP feature vectors (Figure 1). We used Ensembl Biomart [12] to generate the list of Ensemble Gene IDs associated with a given SNP vector and then used the Clone/Gene ID Converter [13] to determine the Kegg pathways in which these genes are involved [14]. Finally, we manually examined the retrieved Kegg pathways to look for biological relevance with respect to the original classification problem.

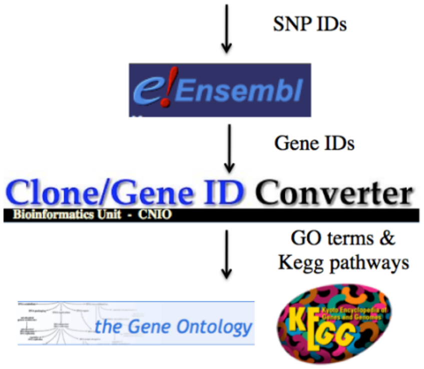


Figure 1. Pipeline for biological validation of SNP subsets

3. Results and discussion

3.1 Disease risk

The ROC curves for the seven binary disease predictors are shown in Figure 2. The classifier for Type 1 diabetes has the best performance, while the other classifiers have only moderate performance. These results are likely due to the fact that several genomic regions contribute strongly towards the T1D phenotype. This conjecture is supported by the fact that T1D had a few SNPs with very significant p-values (on the order of 1E-200) while the other diseases had less significant SNP p-values. The quality of the T1D result, which matches or exceeds the predictive accuracy achieved by prior methods [8], justifies our approach of ranking SNPs by p-value in order to capture the most discriminating features. The performance of the other disease classifiers suggests that the 500K genotyped SNPs does not include those that co-segregate with highly influential genetic loci.

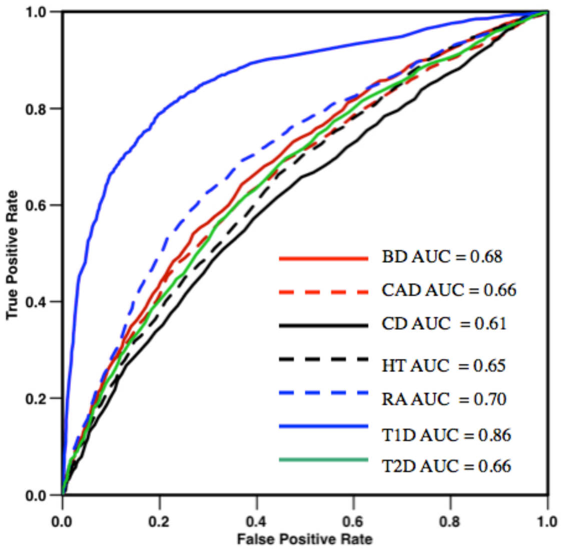


Figure 2. ROC curves for binary SVM classifiers for disease risk

3.2 Early onset risk

Using values from recent literature, which linked SNPs and clinical findings to disease onset groups [9-11], we derived medically relevant cutoff ages to partition the dataset into early onset and late onset groups (Figure 3). We tested variations in our training data where we shifted the cutoff by one decade and where we removed training data for patients within one decade of the cutoff. The reasoning behind this second method is that onset-differentiating SNPs may present a stronger signal between the extremes of the onset populations. Furthermore, because literature definitions for early onset versus late onset were imprecise, removing the patients in the age categories adjacent to our cutoffs produced a training set with higher quality class labels. For all partitioned datasets, we performed permutation testing and found the random sets of SNPs to achieve a mean AUC no greater than 0.49 with standard deviation 0.05. For each disease, we then selected the best performing onset classifier from the above variations (highlighted in Figure 4). The large AUC values for these classifiers indicate that the selected SNP subsets have strong predictive power for early versus late onset.

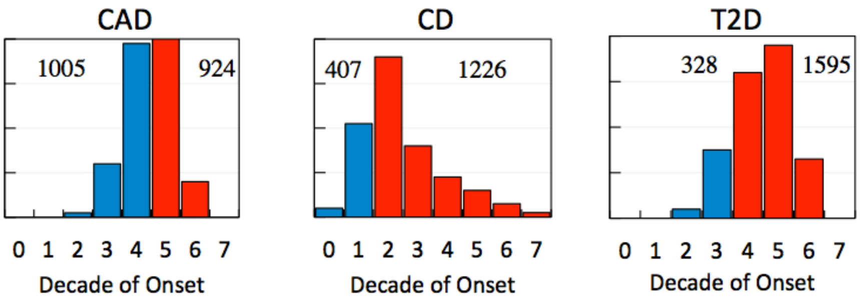


Figure 3. Early and late onset categorization from literature

(A)

Disease			**CAD**		
Partition	Cutoff 4	Cutoff 5	Exclude 4	Exclude 5	Exclude 4,5
$\mathbf{n}_{\textbf{early onset}}$	239	1005	239	1005	239
$\mathbf{n}_{\textbf{late onset}}$	1690	924	924	161	161
$\boldsymbol{\mu}_{\mathbf{AUC}} \pm \boldsymbol{\sigma}_{\mathbf{AUC}}$	0.69 ± 0.02	0.69 ± 0.00	0.73 ± 0.01	0.74 ± 0.03	0.82 ± 0.01
Disease			**CD**		
Partition	Cutoff 2	Cutoff 3	Exclude 2	Exclude 3	Exclude 2,3
$\mathbf{n}_{\textbf{early onset}}$	407	1038	407	1038	407
$\mathbf{n}_{\textbf{late onset}}$	1226	595	595	329	329
$\boldsymbol{\mu}_{\mathbf{AUC}} \pm \boldsymbol{\sigma}_{\mathbf{AUC}}$	0.69 ± 0.02	0.66 ± 0.01	0.73 ± 0.01	0.73 ± 0.02	0.77 ± 0.01
Disease			**T2D**		
Partition	Cutoff 4	Cutoff 5	Exclude 4	Exclude 5	Exclude 4,5
$\mathbf{n}_{\textbf{early onset}}$	328	944	328	944	328
$\mathbf{n}_{\textbf{late onset}}$	1595	979	979	253	253
$\boldsymbol{\mu}_{\mathbf{AUC}} \pm \boldsymbol{\sigma}_{\mathbf{AUC}}$	0.70 ± 0.02	0.66 ± 0.00	0.72 ± 0.02	0.75 ± 0.01	0.77 ± 0.02

(B)

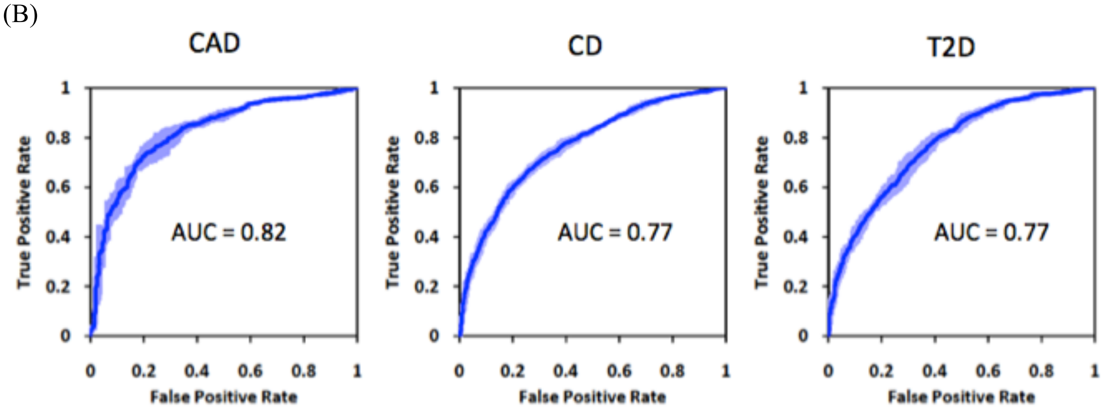
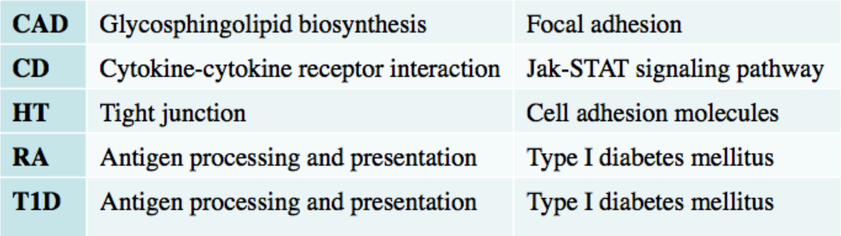


Figure 4. (A) AUC table for all onset cutoff/leave-out variations (B) ROC curves for the best-performing binary SVM classifiers for early onset risk (shaded regions indicate one standard deviation from the mean.

3.3 Biological validation

For each disease classifier, our biological feature pipeline described in 2.6 yielded biological pathways involved in that disease's mechanism, and for each onset classifier, the pipeline yielded pathways implicated in aging and diet. These pathway associations present strong evidence that our SNP vectors have biological relevance and are not artifacts of the learning process (Table 2).

Table 2. Pathways associated with SNP features for disease classification



CAD	Glycosphingolipid biosynthesis	Focal adhesion
CD	Cytokine-cytokine receptor interaction	Jak-STAT signaling pathway
HT	Tight junction	Cell adhesion molecules
RA	Antigen processing and presentation	Type I diabetes mellitus
T1D	Antigen processing and presentation	Type I diabetes mellitus

Conclusion

We have developed a diagnostic tool to predict both disease risk and risk of early disease onset given an individual's genetic information. Our results indicate a low dimensional patient SNP profile can be used for effective risk assessment for type 1 diabetes, and that the WTCCC patient data set contains sufficient information for the construction of disease onset classifiers. Further work on onset classification promises to yield effective early onset prediction and preventative methods both for clinical use and for the rapidly expanding field of personalized medicine.

References

1. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 (7145), 661-678 (2007).
2. Ambady, R., *et al*. Early diagnosis and prevention of diabetes in developing countries. *Rev Endocr Metab Disord*. 3, 193-201 (2008).
3. Wan, X. *et al.*, MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study. *BMC Bioinformatics* 10, 13 (2009).
4. Calabrese, R. *et al*. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat.* 30 (8), 1237-44 (2009).
5. Schaub, M.A. *et al.*, A Classifier-based approach to identify genetic similarities between diseases. *Bioinformatics* 25 (12), i21-29 (2009).
6. Chih-Chung Chang and Chih-Jen Lin. LIBSVM : a library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
7. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851-861 (2007).
8. Wei, Z. *et al.,* From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. *Plos Genetics* 5 (10), (2009).
9. Prudente, S. The TRIB3 Q84R polymorphism and risk of early-onset type 2 diabetes. *J Clin Endocrinol Metab*. 94 (1),190-6 (2008).
10. Brant, S.R. Linkage heterogeneity for the IBD1 locus in Crohn's disease pedigrees by disease onset and severity. *Gastroenterology*. 119 (6), 1483-90 (2000).
11. A genomewide scan for early-onset coronary artery disease in 438 families: the GENECARD Study. *Am J Hum Genet*. 2004 75(3), 436-47 (2004).
12. Flicek P. *et al*. Ensembl's 10th year. *Nucleic Acids Res.* 2009 Nov 11 [Epub ahead of print]
13. Alibés, A. et al. IDconverter and IDClight: Conversion and annotation of gene and protein IDs. *BMC Bioinformatics* 8, 9 (2007).
14. Kanehisa, M. *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354-357 (2006).

Acknowledgments

We would like to acknowledge CS Foo, Marc Schaub, Marina Sirota, Joel Dudley, Lucile Packard Children's Hospital, and SimTK for assistance and support.