11

Meta-Analysis of Genome-Wide Association Studies to Understand Disease Relatedness

Stephanie N. Lewis, Elaine O. Nsoesie, Charles Weeks, Dan Qiao and Liqing Zhang
Virginia Tech, Blacksburg, VA
USA

1. Introduction

Genome-wide association studies (GWAS) have become a popular method of surveying haplotype variations within populations. The recent explosion and success of these studies has allowed for identification of multiple gene variations and non-genetic risk factors that are often involved in pathogenesis of many diseases (Xavier&Rioux, 2008). Efforts to archive these single nucleotide polymorphisms (SNPs) and make the information publicly available have been made possible by the International Haplotype Map Project (HapMap) (The International HapMap Consortium, 2005; The International HapMap Consortium, 2007) and development of GWAS databases (Johnson&O'Donnell, 2009) such as Genomes.gov (Hindorff et al., 2009). The HapMap database of genetic variants and the ever progressing technology involved in identifying genetic disease susceptibility markers has allowed for identification of shared genetic associations that were undetectable with previous methods for identifying deleterious mutations effects for individual genes (Xavier&Rioux, 2008). We are now capable of detecting common susceptibility markers between previously unassociated diseases with the ability to assess combined association signals shared by biological pathways (Wang et al., 2011).

Research of immune-mediated disease susceptibility has benefited from the discovery of shared haplotypes. GWAS with a focus on autoimmune diseases, which included celiac disease, Crohn's disease, multiple sclerosis, rheumatoid arthritis, systemic lupus erythematosus, and type 1 diabetes (Lettre&Rioux, 2008), have shed light on shared genetic markers. Such markers can be exploited to identify biomedical traits that translate to improved diagnostic and treatment techniques (McCarthy et al., 2008). Under the common disease/common variant hypothesis (Wang et al., 2005), one would assume that shared variants result in shared disease phenotypes, and this commonality could serve as a global target for effective treatment options. It is under this assumption that many disease association studies are conducted. The Wellcome Trust Case Control Consortium (WTCCC) conducted a study in which nearly 2000 individuals were examined for coronary artery disease (CAD), hypertension, type II diabetes (T2D), rheumatoid arthritis (RA), Crohn's disease (CD), type I diabetes (T1D) and bipolar disorder (BD) susceptibility against a shared set of about 3000 controls (The Wellcome Trust Case Control Consortium, 2007). The study revealed several association loci for the seven diseases, with some of these indicating risk for more than one of the studied diseases (The Wellcome Trust Case Control Consortium, 2007).

Huang et al. used the data from the WTCCC study to see if associations could be made between the seven diseases given the loci and collections of other data regarding disease susceptibility (Huang et al., 2009). Huang et al. performed analyses at four levels (nucleotide, gene, protein, and phenotype) to determine the existence of overlap across SNPs associated with the seven diseases and constructed protein-protein interaction networks to visualize similarities between diseases (Huang et al., 2009). The group found strong associations across all four levels of analysis for the autoimmune group (CD, RA, and T1D), while no genetic associations were found at any level within the metabolic/cardiovascular group (CAD, hypertension and T2D) (Huang et al., 2009). These results reasserted some expectations derived from clinical literature in the case of the autoimmune group, and suggested inappropriate disease grouping in the case of the metabolic/cardiovascular group (Huang et al., 2009).

For this study, we proposed a large-scale disease and phenotype comparison based on the WTCCC and Huang et al. studies. To this end, we have combined data from GWAS with expression pattern data to determine if genetic and expression similarities exist between diseases. A total of 61 human diseases and phenotypes were assessed. Disease relatedness networks (DRNs) were constructed to visually assess associations on a larger scale. We also took advantage of high-throughput molecular assay technologies to incorporate mRNA expression profiles of diseases, and thus added another dimension of analysis toward assessing disease relationships. Gene expression is an indicator of cellular state, and gene expression profiles can be considered as quantitative traits that are highly heritable. The link between organismal complex traits, such as disease-related phenotype, and gene expression variation has been theoretically accepted (Goring et al., 2007; Moffatt et al., 2007; Chen et al., 2008; Emilsson et al., 2008). With the declining per-sample costs of high-throughput microarray experiments, the amount of gene expression data in international repositories has grown exponentially. The availability of these datasets for many different diseases provides an opportunity to use data-driven approaches to improve our understanding of disease relationships. Hu and Agarwal (Hu&P., 2009) determined disease-disease and disease-drug networks using large-scale gene expression data. Very recently, Suthram et al. (Suthram et al., 2010) presented a quantitative framework to compare and contrast diseases by combining both disease-related mRNA expression data and human protein interaction data. Although GWAS provide comprehensive views of disease interrelationships at the DNA level, the insights from the gene expression aspect, which reflects cellular phenotype, will further advance and strengthen the understanding of this issue. A large-scale disease comparison study such as this has the potential to uncover relationships between diseases and phenotypes that are often overlooked in single disease SNP data analysis.

2. Methods

2.1 SNP-based genetic analysis

Five populations were considered for this expansion study: Han Chinese (CHB), Japanese (JPT), a combined CHB and JPT population (CHB+JPT), Yoruba (YRI), and U.S. residents with northern and western European ancestry (CEU). SNP dataset 2009-02_rel24 (The International HapMap Consortium, 2005; The International HapMap Consortium, 2007) was downloaded from the HapMap site and the SNP set was expanded by means of linkage disequilibrium (LD). SNPs with an r^2 greater than or equal to 0.5 were included. SNPs were divided by associated disease or phenotype (listed in Table 1) and the divisions were

maintained for each succeeding level of analysis. SNPs were divided into blocks based on an r^2 greater than or equal to 0.1. Gene names from Ensembl (Birney et al., 2004) were assigned to blocks if the genetic location was within 2 kilobases up- or downstream of the gene of interest or within the start and end bases for the gene. Gene data were cross-referenced against pathway-specific gene lists generated from the KEGG database (Kanehisa&Goto, 2000; Kanehisa et al., 2006; Kanehisa et al., 2010) in order to assign genes to identified pathways. Pairwise comparisons for each level were conducted to see if diseases and phenotpyes shared SNPs, blocks, genes, or pathway designations. Jaccard index values were calculated for each comparison at each level to assess similarity. Using the Jaccard indexes, DRNs were constructed to visualize the strength of relatedness between diseases. DRNs were visually inspected to identify the strongest relationships. Suggested associations were verified by principal components analysis (PCA) and minor data mining for clinical relevance. Complete details of these methods were previously described by Lewis et al (Lewis et al., 2011).

2.2 Gene expression dataset

The gene expression data used in this analysis was obtained from the NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2009). Not all of the 61 diseases were represented by expression data on the GEO site. Data for a subset of diseases was found by scanning the experimental context of a collection of GEO data (or GEO Series, GSE) for microarrays that were assigned to human disease conditions. Only those microarrays that were curated and reported in the GEO Datasets (or GDS) were used in our analysis. The data set was also restricted to those GSEs in which both the disease and the corresponding control condition (from healthy tissue samples) were measured in the same tissue. For consistency, we further restricted the GSEs to only those datasets which used Affymetrix Gene Chip Human Genome U133 Array Set HG-U133A (GPL96), HG-U133B (GPL97) and HG-U133plus2 (GPL570), which are among the most commonly used platforms. Probes for these platforms were mapped to the current gene identifiers (Chen et al., 2007). This process yielded nineteen diseases for the final GEO analysis.

2.3 Expression measurement

To quantitatively compare expression data, we first normalized the data in each microarray sample using the Z-score transformation to make the expression values across various microarray samples and diseases comparable. Next, we performed an unpaired two-sample Student t-test to compute the t-test statistic and *p*-value of each gene between the disease and control groups. We only used the most appropriate Affymetrix probe set in which a single probe was representative of each gene. The most appropriate Affymetrix probe set was adopted from the work of Hu et al. (Hu&Agarwal, 2009) as many genes were represented by multiple probe sets in Affymetrix U133 microarray chips. This modification avoided correlation and scoring biases brought on by over-representation of those genes. 18,600 most appropriate probes/genes for each of nineteen diseases were identified. The genes were grouped with statistically significant high t-test statistics ($p<0.05$) as "up-regulated genes" and statistically significant low t-test statistics ($p<0.05$) as "down-regulated genes". Instead of using a *p*-value threshold as a cutoff to identify significantly changed genes, the 200 and 1000 most changed genes were designated as the disease-associated significantly changed genes for each disease state. The lowest p-values in each category

Abbre-viation	Disease/ Phenotype	Abbre-viation	Disease/ Phenotype	Abbre-viation	Disease/ Phenotype	Abbre-viation	Disease/ Phenotype
AD	Alzheimer's disease	EO	Early onset extreme obesity	LM	Lipid measurements	QT	Cardiac repolarization (QT interval)
AF	Atrial Fibrillation/ At-rial Flutter	GCA	General cognitive ability	LOAD	Late-onset Alzheimer's disease	RA	Rheumatoid Arthritis
ALS	Amyotrophic Lateral Sclerosis	GD	Gallstone disease	LONG	Longevity and age-related phenotypes	RLS	Restless Leg Syndrome
BA	Brain aging	GLA	Glaucoma	MHA	Minor histocompat-ibility antigenicity	SA	Subclinical atherosclerosis
BC	Breast cancer	HAE	Hepatic adverse events with thrombin inhibitor ximelagatran	MI	Myocardial infarction	SALS	Sporadic Amyotrophic lateral Sclerosis
BD	Bipolar disorder	HBF	Adult fetal hemoglobin levels (HbF) by F cell levels	MS	Multiple sclerosis	SCP	Sleep and circadian phenotypes
BL	Blood lipids	HEI	Height	ND	Nicotine dependence	SLCL	Serum LDL cholesterol levels
BMG	Bone mass and geometry	HEM	Human episodic memory	NEU	Neuroticism	SLE	Systemic Lupus Erythematosus
BPAS	Blood pressure and arterial stiffness	HIV1	HIV-1 disease progression	OBE	Obesity-related traits	SP	Schizophrenia
CA	Childhood asthma	HT	Haematological (blood) traits	PA	Polysubstance addiction	SPBC	Sporadic post-menopausal breast cancer
CAD	Coronary Artery Disease	HYP	Hypertension	PC	Prostate cancer	SPM	Skin pigmentation
CC	Colorectal cancer	IC	Iris color	PD	Parkinson's disease	STR	Stroke
CD	Crohn's disease	IMAN	Immunoglo-bulin A nephropathy	PF	Pulmonary function phenotypes	T1D	Type I Diabetes
CDI	Celiac disease	IS	Ischemic stroke	PR	Psoriasis	T2D	Type II Diabetes
CS	Coronary spasm	KFET	Kidney function and endocrine traits	PSP	Progressive Supranuclear Palsy	TG	Triglycerides
CVD	Cardiovascular Disease outcomes						

Table 1. List of diseases and phenotypes considered for this study and the previous study (Lewis et al., 2011) with corresponding abbreviations.

(up-regulated, down-regulated, and combined) for the top 200 or 1000 genes were pooled for each disease. All of the genes with significant expression changes were grouped together and Jaccard index values were calculated. Gene lists for each disease were compared pair-wise for each of the three expression categories. Here, a high Jaccard index implied a high degree of commonality between diseases/phenotypes. The Jaccard indexes were normalized to produce Z-scores, which were then used as a measure of disease relatedness. The significantly changed genes shared by two diseases were also subjected to Gene Ontology (GO) term enrichment analysis using the web-based Gene Ontology enrichment analysis and visualization (GOrilla) tool (Eden et al., 2007; Eden et al., 2009).

2.4 Medical subject headings (MeSH) term mapping

MeSH is the National Library of Medicine's controlled vocabulary thesaurus (Bodenreider et al., 1998). It consists of sets of terms associated with descriptors in a hierarchical structure. For the nineteen GEO validation diseases (Table 2), the MeSH trees were downloaded and the first level of each tree was used as the disease category. The category that could best indicate the cause of the disease was taken as the disease category.

3. Results

3.1 Summary of significant disease associations for screening of 61 diseases and phenotypes

Jaccard index values were used to assess similarity between diseases and phenotypes within each level of analysis. Correlation between the levels was also assessed using the Spearman correlation method. High correlation was seen between the SNP and block data sets, while low correlation was seen between the pathway data and the other three levels of analysis. The progression from SNP to block, block to gene, and gene to pathway levels resulted in a grouping of susceptibility markers. Visualization of the associations by means of DRNs suggested the grouping translated to an increase in the strength of associations between diseases. This was also reflected in the distribution of Jaccard indexes for each level. Figure 1 shows a slight distribution shift to the right from SNP level to pathway level.

The DRNs suggested consistent association between several diseases for the SNP, block, and gene levels. The strongest associations seen for all populations were observed between (multiple sclerosis [MS], T1D, and RA), with noticeable association between (haematological traits [HT] and adult fetal hemoglobin levels [HBF]) and (serum low-density lipopolysaccharide cholesterol levels [SLCL] and lipid measurements [LM]). Several other less significant associations were suggested by the DRNs as well, but these associations were not consistent in significance for all populations. The qualitative assessments made by examining the DRNs were verified using PCA, which allowed for quantitative isolation of the strongest relationships. The PCA results matched the visual assessment for all levels, and suggested additional strong associations unique to specific populations were present. For example, an association between (LM and triglyceride levels [TG]) that was unique to the JPT population was suggested that was not outwardly apparent by visual inspection of the DRNs. This association was found in the CHB+JPT populations, but not the CHB population. JPT was also missing the (HBF and HT) association that was observed in the other populations. Further details regarding the results of this portion of the study were previously submitted for publication (Lewis et al., 2011).

Disease	platform	GEO record	Sample Size		MeSH category
			Disease	Control	
AD	GPL96	GSE1297	22	9	Nervous System Diseases [C10] Mental Disorders [F03]
ALS	GPL96 and 97	GSE3307	9	16	Nervous System Diseases [C10] Nutritional and Metabolic Diseases [C18]
BD	GPL96	GSE5388	30	31	Mental Disorders [F03]
BC	GPL96 and 97	GSE6883	6	3	Neoplasms [C04] Skin and Connective Tissue Diseases [C17]
CD	GPL96	GSE3365	59	42	Digestive System Diseases [C06]
IS	GPL96	GSE1869	6	10	Cardiovascular Diseases [C14]
OBE	GPL96	GSE474	16	8	Nutritional and Metabolic Diseases [C18]
PD	GPL96	GSE6613	50	22	Nervous System Diseases [C10]
PR	GPL96	GSE6710	13	13	Skin and Connective Tissue Diseases [C17]
SLE	GPL96 and 97	GSE11909	103	12	Skin and Connective Tissue Diseases [C17] Immune System Diseases [C20]
CAD	GPL96	GSE12288	110	120	Cardiovascular Diseases [C14]
T1D	GPL570	GSE10586	12	15	Nutritional and Metabolic Diseases [C18] Endocrine System Diseases [C19] Immune System Diseases [C20]
T2D	GPL96 and 97	GSE9006	12	24	Nutritional and Metabolic Diseases [C18] Endocrine System Diseases [C19]
CA	GPL570	GSE8052	268	136	Respiratory Tract Diseases [C08] Immune System Diseases [C20]
CC	GPL570	GSE9348	70	12	Neoplasms [C04] Digestive System Diseases [C06]
ND	GPL570	GSE11208	6	5	Disorders of Environmental Origin [C21] Mental Disorders [F03]
SP	GPL570	GSE4036	14	14	Mental Disorders [F03]
AF	GPL96 and 97	GSE2240	10	5	Cardiovascular Diseases [C14]
PSP	GPL96	GSE6613	6	22	Nervous System Diseases [C10] Eye Diseases [C11]

Table 2. List of nineteen diseases in gene expression analysis and their MeSH classification.

3.2 Clustering of genetic associations

Based on the observations made using the DRNs, agglomerative hierarchical clustering was used to find groups of diseases. At each level, the 61 diseases/phenotypes were clustered into ten groups. The number of clusters was set to ten based on visual inspection of the

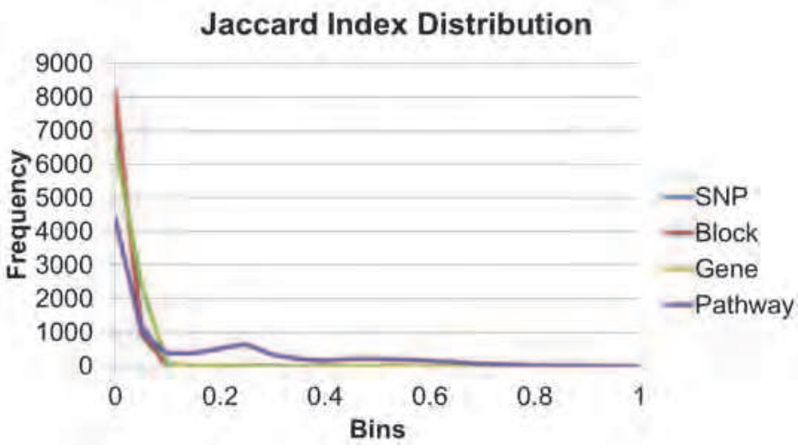


Fig. 1. Graphical representation of histogram data showing distribution of all Jaccard indexes for all populations at each level of analysis. Index values were grouped and then divided into twenty bins across the range zero to one. (N = 9150 for each analysis level)

hierarchical branching of the trees. Representative clustering results are shown for the CHB+JPT population in Figure 2. The CHB+JPT population showed a high correlation to most populations at all levels of analysis based on the Rand Index for similarity. The Rand Index for similarity was used to compare the clustering across populations at each level. The diseases within each cluster were least similar at the SNP level for all populations and most similar at the gene level across most of the populations. At the SNP level groupings, associations between (MS, RA, and T1D), (HBF and HT), and (breast cancer [BC] and sporadic post-menopausal breast cancer [SPBC]) were found for all populations (Figure 2A). The grouping of (RA and T1D), (BC and SPBC), (HBF and HT), (Amyotrophic Lateral Sclerosis [ALS] and Parkinson's disease [PD]) and (colorectal cancer [CC] and prostate cancer [PC]) were consistent at the block level for all populations (Figure 2B). At the gene level, the number of diseases/phenotypes included in each cluster increased with consistent groups again observed for all populations. These groups included (MS, RA, and T1D), (ALS, PD, CAD, Alzheimer's disease [AD] and T2D), and (neuroticism [NEU], brain aging [BA], and sleep and circadian phenotypes [SCP]) (Figure 2C). Clusters at the pathway level were also much larger than at the other levels. No consistent relationships were seen for the clusters containing a larger number of diseases, but the smaller groupings consistently showed relationships between (longevity and age-related phenotypes [LONG] and early onset extreme obesity [EO]), (cardiovascular disease outcomes [CVD], CD, and NEU) and (blood lipids [BL], LM, and Restless Leg Syndrome [RLS]) (Figure 2D). Four populations suggested clustering of (LONG, EO, and T1D), while one, YRI, showed a relationship between (LONG, EO, and SLCL).

3.3 Gene expression analysis

The gene expression profiles showed some patterns for the three expression categories (up-regulated, down-regulated, and combined), with the number of strong associations increasing with cutoff type (top 200 most changed genes, top 1000 most changed genes, and changes with a *p*-value less than 0.05). Jaccard indexes for each disease/phenotype pair were calculated and used to construct DRNs, which are shown in Figure 3. Strong associations between (PD, Progressive Supranuclear Palsy [PSP], and nicotine dependence

[ND]), (ischemic stroke [IS], CC, and CD), and (CAD and childhood asthma [CA]) were observed under all three cutoff scenarios for all three expression categories of analysis. Of these, the (CAD and CA) pair showed the most variation in association strength for all the variables considered.

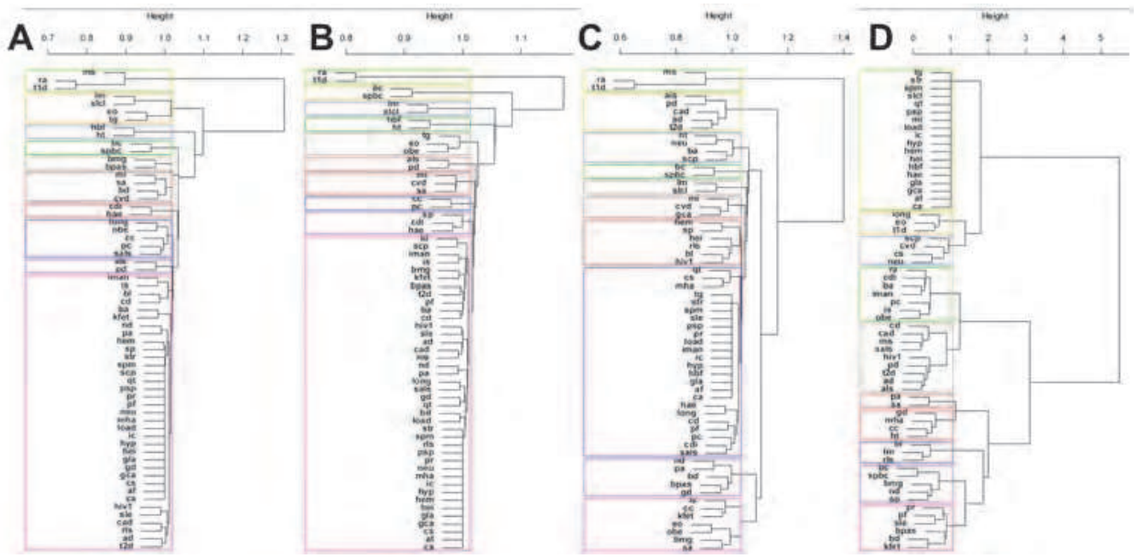


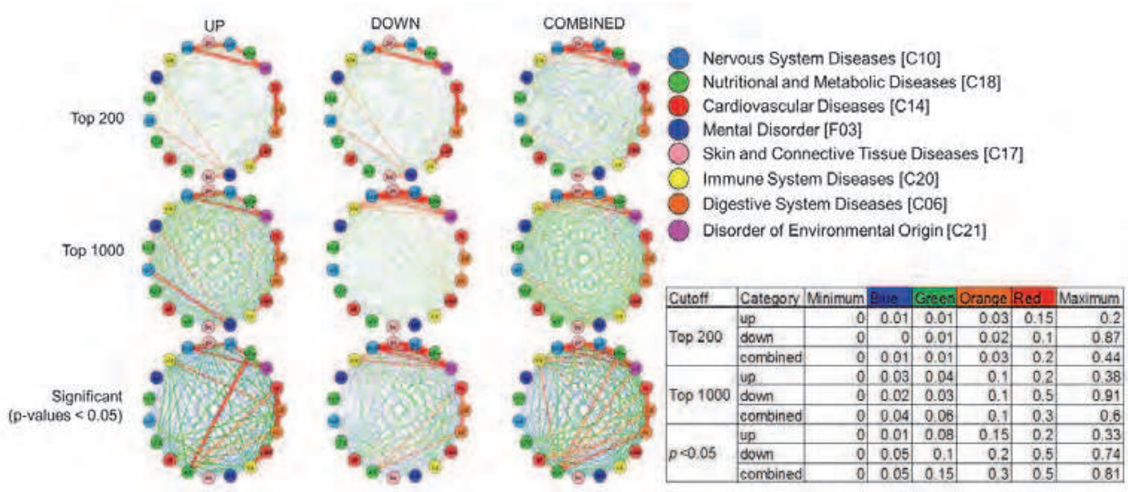
Fig. 2. Clustering dendrogram for 61 disease/phenotype comparisons at the (A) SNP, (B) block, (C) gene, and (D) pathway levels. Colored boxes indicate the clusters derived from Rand Index analysis. Results for the CHB+JPT population are shown as a representative data set for all populations.

Links between disease classifications were also seen. Connections between nervous system diseases and disorders of environmental origin (i.e., (PSP and ND) and (PD and ND)) were seen in all three expression categories and cutoff types. Associations between nervous system and mental disorders (e.g., AD and BD) were seen for the top 200 and top 1000 groups, but this association was masked in the *p*-value-derived group. For the p-value group, predominate associations between metabolic, cardiovascular, digestive, and immune system diseases were found. One unexpected classification association was the nervous system-metabolic disease link exemplified by (PSP and OBE) and (PD and OBE) for the down-regulation and subsequently combined expression groups with the top 1000 and p-value cutoffs.

As expected, the number of significant associations increased as the threshold criteria increased given that the quantity of data available for comparison was greater. Seemingly strong associations observed at the top 200 cutoff, such as the (AD and BD) and (BD and SP) associations were masked in the *p*-value cutoff data as other stronger associations were present. The increase in maximum Jaccard index for the combined expression data set from 0.44 to 0.81 agreed with this observation. Though we saw an increase in relationship strength with less stringent cutoff thresholds, the additional comparison data resulted in reduction in significant associations. Therefore, the expression categories for the *p*-value cutoff group were used to compare with the SNP-based data in order to avoid assigning an arbitrary cutoff for the expression data and to ensure enough data was available for the nineteen-disease comparison.

3.4 Comparison of the SNP and expression data for nineteen diseases

Correlation between data sets may have been influenced by the data sources. Both the SNP and block levels encompassed data from the HapMap site. The gene level data was obtained by cross referencing the HapMap data against the Ensembl database of gene names. The pathway data was obtained by cross referencing the Ensembl-derived data against the KEGG database. Given that the amount of data available through each of these sources is not consistent, there was loss of data in the transition from blocks to genes and genes to pathways. Of the reduced set of nineteen diseases and phenotypes compared, only atrial fibrillation/atrial flutter (AF) did not contain gene data for the SNP-based comparisons. The number of missing diseases/phenotypes increased to four at the pathway level (i.e., AF, CA, psoriasis [PR], and PSP). Despite the missing disease associations for AF, the gene level of analysis was used for comparison to the expression data. The range of Z-scores for this dataset was closest to the range seen for the expression data, and intuitively, the gene data should show some correlation to gene expression.



Cutoff	Category	Minimum	Blue	Green	Orange	Red	Maximum
Top 200	up	0	0.01	0.01	0.03	0.15	0.2
	down	0	0	0.01	0.02	0.1	0.87
	combined	0	0.01	0.01	0.03	0.2	0.44
Top 1000	up	0	0.03	0.04	0.1	0.2	0.38
	down	0	0.02	0.03	0.1	0.5	0.91
	combined	0	0.04	0.06	0.1	0.3	0.6
p <0.05	up	0	0.01	0.08	0.15	0.2	0.33
	down	0	0.05	0.1	0.2	0.5	0.74
	combined	0	0.05	0.15	0.3	0.5	0.81

Fig. 3. DRNs for expression data for the three cutoff levels (top 200, top 1000, and significance with p-value < 0.05) and three expression categories (up-regulated, down-regulated, and combined). Disease nodes are color coded to show grouping of diseases based on MeSH classification. Edges are color coded according to increasing strength of disease association. Values for the color scale are listed in the inserted table.

DRNs comparing the gene level of analysis for the CEU, CHB+JPT, and YRI populations to the expression data are shown in Figure 4. The JPT and CHB populations are not shown since the CHB+JPT population is highly representative of the individual populations. A Spearman correlation was calculated between each population for the SNP-based data set and the expression data (Table 3). A weak negative correlation was observed between the genetic and expression data, suggesting no significant relationships were shared between the two data sets. A qualitative analysis of the networks and clustering from the SNP-based data analysis suggested a high degree of similarity between the predicted associations for all population. However, the strong associations observed in the genetic analysis were not seen in the expression data. Rather, a seemingly reciprocal relationship appeared between the

genetic and expression DRNs. The strongest expression-based association was between ALS and obesity-related traits (OBE), which was in the weakest associations group for the SNP-based associations. An examination of the genetic DRNs suggested the strongest associations between (ALS and PD), (AD and T2D), and (T1D and SLE). These associations were weak for the expression data. Some associations near the middle of the Z-score range appeared more common between the data sets, such as the (IS and CC), (AD and BD), and (OBE and CC) pairs.

61 diseases	CEU	CHB	JPT	CHB+JPT	YRI
CEU	1	0.9599	0.9595	0.9574	0.9447
CHB		1	0.9779	0.9925	0.9726
JPT			1	0.9858	0.9556
CHB+JPT				1	0.9686
YRI					1
19 diseases					
GEO	-0.1367	-0.1228	-0.1278	-0.1254	-0.1176

Table 3. Spearman correlation coefficients between populations and between each population and the GEO data. The Spearman correlation is a comparison of the ranked Z-scores for each data set.

Despite the overall lack of correlation between the genetic and expression analyses, several unexpected links between neurological and cardiovascular/metabolic diseases were observed in both data sets (i.e., (AD and T2D) and (PD and OBE)). These potentially novel disease relationships may primarily rely on genetic similarity or genomic expression similarity instead of phenotypic classification, but this idea would need to be further explored.

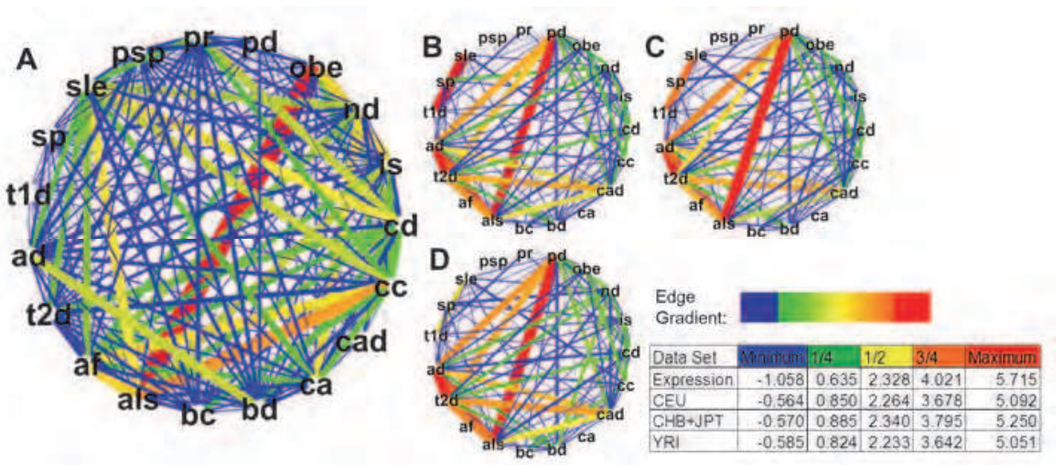


Fig. 4. DRNs based on Z-scores for three populations and expression data. DRNs for (A) combined expression data for significantly changed genes (p<0.05), (B) CEU gene level, (C) CHB+JPT gene level, and (D) YRI gene level are shown. Edge live color and width correspond to strength of association between disease pairs. The gradient and corresponding values are listed in the inserted table.

4. Discussion

The results from this study suggested it is possible to elucidate genetic similarities that can be overlooked during single disease GWAS. Several expected associations supported by literature were found (e.g. association between (SLE and RA) and (EO and SLCL)) while some unexpected associations were also observed. The unexpected neurological-cardiovascular/metabolic disease associations were observed for both the genetic analysis and the expression profile analysis. Though the origin and symptoms associated with diseases in each category may be different, the results suggest genetic similarities. Possible explanations for these associations cannot be elucidated solely from this study given the broad nature of the comparison. A detailed SNP-by-SNP and gene-by-gene examination may indicate the reason behind the neurological-cardiovascular/metabolic relatedness. Those relationships are particularly interesting and may indicate some common underlying molecular mechanism among these disease groups that has not yet been widely studied.

Clinical evidence supports the strongest relationships identified from the expression data. PSP and PD share some common symptoms such as stiffness, and movement difficulties which could explain the common expression pattern indicating some degree of relatedness between the two. On the other hand, explaining the relationship between PSP and ND is more difficult. Several studies have shown that smokers have a lower risk of developing Parkinson's disease (Soto-Otero et al., 1998; Hernan et al., 2001; Quik, 2004). One recently published paper showed that smoking for a greater number of years may reduce the risk of the disease (Chen et al., 2010). An earlier study suggested that younger patients with CD might be under an increased risk of IS (Andersohn et al., 2010). Extensive studies have demonstrated a strong association between CD and CC (Gillen et al., 1994). The relationship between (IS and CC) and (CAD and CA) is also unclear, but shared immune-dependent responses may be the common link.

Similarities and differences were observed between the three categories (up-regulated, down-regulated and combined) of gene expression analysis (see Figure 3). The different association patterns may be due to the use of a single rule to identify disease associated genes for all kinds of diseases, which over simplifies the problem. Theoretically, variance of gene expression can be considered as a quantitative trait inherited from genetic variation. It is possible that a combined DNA variant and expression phenotype can better explain genetic architecture with reduced environmental and biological noise (Dermitzakis, 2008). However, the precise and reliable estimation of molecular link between functional genomic effects and complex organism phenotypes depends on a large number of pooled variant and gene expression data from corresponding tissues or cell types, since tissue-specific differences can be found widely (Dermitzakis, 2008). A combined genetic and gene expression profile study, as presented here, can shed light on disease relatedness from different perspectives. Parikh et al. performed a more direct comparison of GWAS and expression data in an effort to prioritize T2D susceptibility genes (Parikh et al., 2009). The group isolated SNPs from GWAS, searched for associated genes, and then found corresponding tissue-specific expression profiles for a subset of all the SNP-associated genes (Parikh et al., 2009). Parikh et al. were able to identify five genes common to individuals with T2D and twelve genes with differentiating expression patterns in individuals with versus without the disease (Parikh et al., 2009). Rather than focusing on a single disease to identify targets, we strove for a more global comparison of genetic and expression data.

Even though discrepancies between our data sets were observed, it is possible that the reduction in data between the gene and pathway level could have excluded some genes common to multiple diseases. With the increased density of GWAS and gene expression studies, the discrepancies and anomalies observed in this study might be better understood. We set out to support the idea that diseases potentially share phenotype similarity as a result of genetic factors, pathway associations, expression regulation, or some combination of these three ideas. Within the autoimmune disease group, we observed diseases that possessed some genetic similarity. We saw expected strong associations between T1D, MS, and RA, as well as less expected associations between AD and T2D. It would appear that systemic inflammation responses may be the key to shared susceptibility among many of the diseases and phenotypes for which we observed relatedness. Clinical studies suggested individuals with one immune-mediated disease, such as T1D, may be more susceptible to pathogenesis of another (Dorman et al., 2003; Nielson et al., 2006; Toussirot et al., 2006; Doran, 2007). It has also been clinically suggested that inflammation plays a role in neurological diseases like AD (Akiyama et al., 2000; Perry, 2004) and PD (Perry, 2004). We also know that cardiovascular and metabolic diseases, such as atherosclerosis, T2D, and OBE have links to chronic inflammatory responses (Stienstra et al., 2006; Tontonoz&Spiegelman, 2008). In all of these cases, our results suggest the clinical manifestations may have genetic relevance and the unexpected cardiovascular/neurological links may be important. Given the broad scope of this study, the conclusions made here are suggestions for where genetic commonality could be found without specific identification of the related targets. A more detailed disease-by-disease analysis similar to the study conducted by Parikh et al. (Parikh et al., 2009) would need to be conducted to identify specific genes of interest shared by diseases. The methods used in the Parikh et al. study can be specifically applied to the study of T1D by performing a detailed step-by-step comparison between this disease and other possibly related diseases in order to elucidate genetic commonalities to T1D. The results from our study and from one tailored specifically for T1D could influence current treatment options and suggest new approaches for managing and treating the disease. We feel our study is a strong example of how GWAS and expression data can be used conjunctively to predict significant disease associations relevant to improving and unifying diagnoses and treatment options for multiple immune-mediated diseases.

5. Acknowledgements

The authors would like to acknowledge the faculty and students of the Spring 2010 Genetics, Bioinformatics, and Computational Biology Problem Solving course for feedback regarding the progress of this study.

6. References

Akiyama, H., S. Barger, et al. (2000). "Inflammation and Alzheimer's." Neurobiology of Aging 21(3): 383-421.

Andersohn, F., M. Waring, et al. (2010). "Risk of ischemic stroke in patients with Crohn's disease: A population-based nested case-control study." Inflammatory Bowel Disease 16(8): 1387-1392.

Barrett, T., D. Troup, et al. (2009). "NCBI GEO: archive for high-throughput functional genomic data." Nucleic Acids Research 37: D885-D890.

Birney, E., T. D. Andrews, et al. (2004). "An overview of Ensembl." Genome Research 14(5): 925-928.

Bodenreider, O., S. Nelson, et al. (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Annual Symposium of the American Medical Informatics Association, Orlando, FL, Hanley & Belfus, Inc.

Chen, H., X. Huang, et al. (2010). "Smoking, duration, intensity, and risk of Parkinson disease." Neurology 74(11): 878-884.

Chen, R., L. Li, et al. (2007). "AILUN: reannotating gene expression data automatically." Nature Methods 4(11): 879.

Chen, Y., J. Zhu, et al. (2008). "Variations in DNA elucidate molecular networks that cause disease." Nature 452(7186): 429-435.

Dermitzakis, E. T. (2008). "From gene expression to disease risk." Nature genetics 40(5): 492-493.

Doran, M. (2007). "Rheumatoid arthritis and diabetes mellitus: evidence for an association?" The Journal of Rhematology 34(3): 460-462.

Dorman, J. S., A. R. Steenkiste, et al. (2003). "Type 1 Diabetes and Multiple Sclerosis." Diabetes Care 26(11): 3192-3193.

Eden, E., D. Lipson, et al. (2007). "Discovering Motifs in Ranked Lists of DNA Sequences." PLoS Computational Biology 3(3): e39.

Eden, E., R. Navon, et al. (2009). "GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists." BMC Bioinformatics 10: 48.

Emilsson, V., G. Thorleifsson, et al. (2008). "Genetics of gene expression and its effect on disease." Nature 452(7186): 423-428.

Gillen, C. D., H. A. Andrews, et al. (1994). "Crohn's disease and colorectal cancer." Gut 35(5): 651-655.

Goring, H. H. H., J. E. Curran, et al. (2007). "Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes." Nature Genetics 39: 1208-1216.

Hernan, M. A., S. M. Zhang, et al. (2001). "Cigarette Smoking and the Incidence of Parkinson's Disease in Two Prospective Studies." Annals of Neurology 50(6): 780-786.

Hindorff, L. A., P. Sethupathy, et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." PNAS 106(23): 9362-9367.

Hu, G. and P. Agarwal. (2009). "Human disease-drug network based on genomic expression profiles." PLoS One 4(8): e6536.

Huang, W., P. Wang, et al. (2009). "Indentifying disease associations via genome-wide association studies." BMC Bioinformatics 10: 1-11.

Johnson, A. D. and C. J. O'Donnell (2009). "An Open Access Database of Genome-wide Association Results." BMC Medical Genetics 10: 1-6.

Kanehisa, M. and S. Goto (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes." Nucleic Acids Research 28(27-30).

Kanehisa, M., S. Goto, et al. (2010). "KEGG for representation and analysis of molecular networks involving diseases and drugs." Nucleic Acids Research 38: D355-D360.

Kanehisa, M., S. Goto, et al. (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic Acids Research 34: D354-D357.

Lettre, G. and J. D. Rioux (2008). "Autoimmune diseases: insights from genome-wide assocation studies." Human Molecular Genetics 17(2): R116-R121.

Lewis, S. N., E. Nsoesie, et al. (2011). "Prediction of Disease and Phenotype Associations from Genome-Wide Association Studies." PLoS One Submitted for Review.

McCarthy, M. I., G. R. Abecasis, et al. (2008). "Genome-wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges." Nature 9: 356-369.

Moffatt, M. F., M. Kabesch, et al. (2007). "Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma." Nature 448(7152): 470-473.

Nielson, N. M., T. Westergaard, et al. (2006). "Type 1 Diabetes and Multiple Sclerosis: A Danish Population-Based Cohort Study." Archives of Neurology 63(7): 1001-1004.

Parikh, H., V. Lyssenko, et al. (2009). "Prioritizing genes for follow-up from genome wise assocation studies using information on gene expression in tissues relevant for type 2 diabetes mellitus." BMC Medical Genetics 2(72).

Perry, V. H. (2004). "The influence of systemic inflammation on inflammation in the brain: implications for chronic neurodegenerative disease." Brain, Behavior, and Immunity 18(5): 407-413.

Quik, M. (2004). "Smoking, nicotine and Parkinson's disease." Trends in Neurosciences 27(9): 561-568.

Soto-Otero, R., E. Mendez-Alvarez, et al. (1998). "Studies on the interaction between 1,2,3,4-tetrahydro-B-carboline and cigarette smoke: a potential mechanism of neuroprotection for Parkinson's disease." Brain Research 802(1-2): 155-162.

Stienstra, R., C. Duval, et al. (2006). "PPARs, Obesity, and Inflammation." PPAR Research 2007: 1-10.

Suthram, S., J. T. Dudley, et al. (2010). "Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets." PLoS Computational Biology 6(2): 1-10.

The International HapMap Consortium (2005). "A Haplotype Map of the Human Genome." Nature 437(7063): 1299-1320.

The International HapMap Consortium (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature 449: 851-861.

The Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447: 661-678.

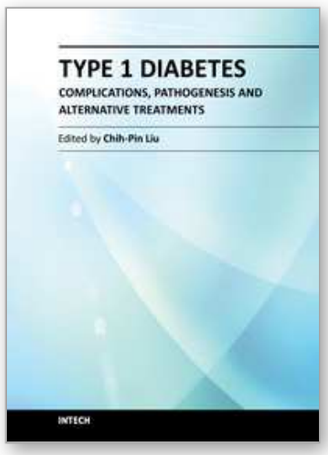
Tontonoz, P. and B. M. Spiegelman (2008). "Fat and Beyond: The Diverse Biology of PPARγ." Annual Review of Biochemistry 77: 289-312.

Toussirot, E., E. Pertuiset, et al. (2006). "Association of Rheumatoid Arthritis with Multiple Sclerosis: Report of 14 Cases and Discussion of Its Significance." The Journal of Rhematology: Correspondence 33(5): 1027-1028.

Wang, L., P. Jia, et al. (2011). "An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide assocation studies." Bioinformatics 27(5): 686-692.

Wang, W. Y. S., B. J. Barratt, et al. (2005). "Genome-Wide Association Studies: Theoretical and Practical Concerns." Nature Reviews Genetics 6: 109-118.

Xavier, R. J. and J. D. Rioux (2008). "Genome-wide association studies: a new window into immune-mediated diseases." Nature Reviews Immunology 8: 631-643.



Type 1 Diabetes - Complications, Pathogenesis, and Alternative Treatments

Edited by Prof. Chih-Pin Liu

ISBN 978-953-307-756-7
Hard cover, 470 pages
Publisher InTech
Published online 21, November, 2011
Published in print edition November, 2011

This book is intended as an overview of recent progress in type 1 diabetes research worldwide, with a focus on different research areas relevant to this disease. These include: diabetes mellitus and complications, psychological aspects of diabetes, perspectives of diabetes pathogenesis, identification and monitoring of diabetes mellitus, and alternative treatments for diabetes. In preparing this book, leading investigators from several countries in these five different categories were invited to contribute a chapter to this book. We have striven for a coherent presentation of concepts based on experiments and observation from the authors own research and from existing published reports. Therefore, the materials presented in this book are expected to be up to date in each research area. While there is no doubt that this book may have omitted some important findings in diabetes field, we hope the information included in this book will be useful for both basic science and clinical investigators. We also hope that diabetes patients and their family will benefit from reading the chapters in this book.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen