American Journal of Epidemiology
Vol. 176, No. 5
DOI: 10.1093/aje/kws123
Advance Access publication:
August 2, 2012

Original Contribution

Integrating Genetic Association, Genetics of Gene Expression, and Single Nucleotide Polymorphism Set Analysis to Identify Susceptibility Loci for Type 2 Diabetes Mellitus

Danielle M. Greenawalt*, Solveig K. Sieberts, Marilyn C. Cornelis, Cynthia J. Girman, Hua Zhong, Xia Yang, Justin Guinney, Lu Qi, and Frank B. Hu

* Correspondence to Dr. Danielle M. Greenawalt, Department of Genetics, Merck Research Laboratories, 33 Avenue Louis Pasteur, Boston, MA 02115 (e-mail: danielle_greenawalt@merck.com).

Initially submitted July 6, 2011; accepted for publication January 30, 2012.

Large-scale genome-wide association studies (GWAS) have identified over 40 genomic regions significantly associated with type 2 diabetes mellitus. However, GWAS results are not always straightforward to interpret, and linking these loci to meaningful disease etiology is often difficult without extensive follow-up studies. The authors expanded on previously reported type 2 diabetes mellitus GWAS from the nested case-control studies of 2 prospective US cohorts by incorporating expression single nucleotide polymorphism (SNP) information and applying SNP set enrichment analysis to identify sets of SNPs associated with genes that could provide further biologic insight to traditional genome-wide analysis. Using data collected between 1989 and 1994 in these previous studies to form a nested case-control study, the authors found that 3 of the most significantly associated SNPs to type 2 diabetes mellitus in their study are expression SNPs to the lymphocyte antigen 75 gene (*LY75*), the ubiquitin-specific peptidase 36 gene (*USP36*), and the phosphatidylinositol transfer protein, cytoplasmic 1 gene (*PITPNC1*). SNP set enrichment analysis of the GWAS results identified enrichment for expression SNPs to the macrophage-enriched module and the Gene Ontology (GO) biologic process fat cell differentiation human, which includes the transcription factor 7-like 2 gene (*TCF7L2*), as well as other type 2 diabetes mellitus-associated genes. Integrating genome-wide association, gene expression, and gene set analysis may provide valuable biologic support for potential type 2 diabetes mellitus susceptibility loci and may be useful in identifying new targets or pathways of interest for the treatment and prevention of type 2 diabetes mellitus.

expression single nucleotide polymorphism; gene set enrichment analysis; genome-wide association study; integrative genomic analysis; single nucleotide polymorphism; type 2 diabetes

Abbreviations: eSNP, expression-associated single nucleotide polymorphism; GSEA, gene set enrichment analysis; GWAS, genome-wide association study(ies); HapMap, International Haplotype Map Project; KEGG, Kyoto Encyclopedia of Genes and Genomes; SNP, single nucleotide polymorphism; SSEA, single nucleotide polymorphism set enrichment analysis.

Type 2 diabetes afflicts an estimated 270 million adults worldwide and is projected to increase to 416.5 million by 2030, an increase of 54% (1). In the United States, estimates of projected increases from 2005 to 2050 are as high as 174% in men and 220% in women (2). Genome-wide association studies (GWAS) have now identified greater than 70 single nucleotide polymorphisms (SNPs) spanning more than 40 genomic regions, each with a small contribution to an individual's susceptibility to develop type 2 diabetes mellitus (3). The most widely investigated and replicated locus maps to the region of the transcription factor 7-like 2 gene (*TCF7L2*), which has been linked to diabetes, prediabetes, insulin response, and complications such as retinopathy (4–7). However, linking the SNPs identified in GWAS to a gene of interest and then biologically validating this gene in the context of the disease is often difficult. More recently, integrative analyses have been reported that seek to leverage gene ontology/pathway information to provide context for GWAS results (8, 9) similar to gene set enrichment analysis (GSEA) for gene expression results. GSEA was first proposed to analyze data from gene expression-profiling experiments (10) and was later used for analysis of GWAS data by selecting genes based on their proximity to the GWAS SNPs (11). We have recently extended this analytical method to incorporate gene expression data with SNP association results through the use of expression-associated SNPs (eSNPs). This analysis is referred to as SNP set enrichment analysis (SSEA) (9). Combining the genetics of gene expression with GSEA allows for a direct association between a set of SNPs and a gene set of interest. This allows identification of sets of genes whose surrogate SNPs show enriched association with the endpoints of interest, even though individual SNPs may not have reached genome-wide significance.

Two nested case-control studies were used to investigate the genetics of type 2 diabetes mellitus. These samples were identified from Nurses' Health Study and Health Professionals Follow-up Study prospective cohorts (12). We extended the analysis performed by Qi et al. (12) by imputing approximately 2,500,000 genotypes in the 5,773 individuals and performed single SNP association analysis. We then used eSNP information from various relevant tissues, generated from independent cohorts, to provide biologic context in terms of gene associations to our single SNP association results. Single SNP results are then linked to eSNPs of genes in gene sets to perform SSEA. SSEA allows us to identify gene sets or biologic pathways enriched with GWAS association signals to provide further insight to type 2 diabetes mellitus etiology. SSEA identified gene sets for which none of the constituent eSNPs showed genome-wide significance associations, as one might expect when each member of the gene set or pathway contributes small effects but whose combined effect is significant, allowing us exploratory mining of single SNP association results that otherwise may not have been included in deeper mining efforts.

MATERIALS AND METHODS

The Nurses' Health Study and Health Professionals Follow-up Study cohorts have been described previously (12–14). Briefly, these are prospective studies for which biennial follow-up through self-administered questionnaires is used to update information about health and disease, as well as dietary intake and lifestyle. Blood collection was performed on 32,826 women (Nurses' Health Study) and 18,159 men (Health Professionals Follow-up Study) between 1989 and 1994. It was from these subjects that the nested case-control study was formed. Diabetes incidence was identified by self-report on biennial follow-up questionnaires and confirmed by a medical record-validated supplementary questionnaire regarding diabetes symptoms, diagnostic tests, and treatments. Validation through medical records demonstrated high reliability (>98%) of the self-report of diabetes in the supplementary questionnaires (15). Diabetes-free controls were selected to be matched on gender, year of birth, month of blood collection, and fasting status, although the matching was later broken because not all subjects gave informed consent for the genome-wide association study.

Genotype data

Genotyping was done on the Affymetrix 6.0 chip (Affymetrix, Santa Clara, California) and is available in the database of genotypes and phenotypes known as "dbGAP." Quality control was performed as described previously (12). The current analyses expanded the set of SNPs by imputing approximately 2.5 million International Haplotype Map Project (HapMap) SNPs in the 3,286 and 2,487 (genetically inferred) samples of European ancestry from the Nurses' Health Study and the Health Professionals Follow-up Study, respectively. Imputation was performed by using MACH HapMap CEU reference II r22 b36 (16).

Single SNP analyses

Genome-wide association analyses were first conducted separately for the Nurses' Health Study and Health Professionals Follow-up Study type 2 diabetes mellitus case-control studies, followed by a meta-analysis of the 2 studies. Association analyses of the directly assayed SNPs were performed in PLINK version 1.07 (17) and the imputed doses in ProbABEL (18). Meta-analyses of the 2 case-control studies were performed by using a fixed-effect model in PLINK. In order to assess and control for population structure, principal component analysis was done prior to association, separately in the 2 case-control studies, by using EIGENSTRAT (19). Results of the principal component analysis were used as covariates in the association analysis along with age and body mass index. Only the first eigenvector from the principal component analysis was used in this model. However, 2 additional models were run as sensitivity analyses and did not change the results of the analysis materially. The first sensitivity model included no covariates, and the second included age, body mass index, and the first 3 (Nurses' Health Study) or 4 (Health Professionals Follow-up Study) principal components. In all 3 analyses, an additive genetic model was assumed.

Expression-associated SNPs

eSNPs have been identified in 3 independent cohorts by performing whole-genome genotyping by SNP array and RNA profiling in multiple tissues from the sample individual. In these studies, an individual's genotype at each locus is correlated to the expression of each transcript profiled, and significant associations were identified at a false discovery rate of 10% (20–22). Our analysis focused on eSNPs that were within 500 kilobases of the gene of interest or within 1 megabase of the transcript of interest. Results of our type 2 diabetes mellitus single SNP analysis were queried against each of the eSNP data sets representing metabolic related tissues including liver, subcutaneous adipose, and omental adipose. eSNPs were selected from metabolic tissues rather than other eSNP cohorts because they have been shown to be more highly associated with obesity and type 2 diabetes mellitus traits (21, 23).

SNP set enrichment analysis

Fifty-three gene sets were selected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (24), Gene Ontology, or relevant literature, that were linked to type 2 diabetes mellitus or other metabolic traits of interest (Web Table 1, the first of 3 Web tables and 2 Web figures posted on the *Journal*'s Web site (http://aje.oupjournals.org/)). The gene sets also included gene coexpression modules, which in previous studies have been found to be correlated to insulin, glucose, homeostatic model assessment of insulin resistance, body mass index, or leptin levels. The modules are named by a reference color, the tissue, and cohort from which they were derived (21). Similar to methods described previously by Zhong et al. (9), for each gene set, a set of SNPs that had previously been associated with the expression of each gene in the gene set (eSNPs) was selected. If more than one SNP was found to be associated with a gene in a gene set, the eSNP with the most significant association to the gene was chosen for inclusion in the analysis. In the case where no eSNP was identified for a gene in a gene set, no SNP was selected, and the gene was not represented in the test. With the SNPs representing each gene set, we then used the type 2 diabetes mellitus meta-analysis *P* values to perform a one-sample, one-sided Kolmogorov-Smirnov test for departure from uniform, in which a significant result indicates that the distribution of *P* values for the gene set is stochastically less than that expected under a uniform (0, 1) distribution. In order to assess the empirical null distribution of these Kolmogorov-Smirnov test *P* values, which may differ from the theoretical, asymptotic distribution due to departures from the implicit assumptions of the test, we performed 10,000 permutations of the genotypes relative to the phenotypes, and the distributions of the permuted SNP set test statistic (i.e., the Kolmogorov-Smirnov test *P* values) were compared with those for the observed data in order to obtain SNP set permutation *P* values. This procedure is done on a gene set-by-gene set basis. The analysis and permutations were repeated independently with 141 KEGG pathways, 140 of which had 5 or more genes represented by eSNPs similar to the analysis performed by Zhong et al. (9). In order to address the multiple testing due to the number of gene sets examined, we computed a false discovery rate by using the Benjamini-Hochberg method (25).

RESULTS

Single SNP analysis

After genotype quality control and population inference, 3,214 (1,464 cases, 1,750 controls) individuals in the Nurses' Health Study and 2,307 (1,063 cases, 1,244 controls) individuals in the Health Professionals Follow-up Study were used in the analysis. The most significant associations with type 2 diabetes mellitus identified from the meta-analysis were in the region of *TCF7L2* ($P = 3.26 \times 10^{-13}$) (Web Figure 1; Table 1). The second most significant region identified in our analysis was in the ADAM metallopeptidase with thrombospondin type 1 motif, 9 gene (*ADAMTS9*) (chromosome 3, $P = 1.28 \times 10^{-07}$). Both of these regions have been significantly associated with type 2 diabetes mellitus susceptibility previously (26–28). Another region identified from the meta-analysis was on chromosome 2 in the RNA-binding motif, single-stranded interacting protein 1

Table 1. Type 2 Diabetes Mellitus Genome-wide Association Study Results[a,b]

Chromosome	Position	Most Significant SNP	Meta *P* Value	Odds Ratio	In Gene	Closest Gene	eSNP Gene
10	114748339	rs7903146	3.26×10^{-13}	0.72	*TCF7L2*		
2	160922421	rs6718526	7.65×10^{-7}	1.30	*RBMS1*		*LY75*
3	64711617	rs2371765	8.90×10^{-7}	1.22		*ADAMTS9*	
17	74333763	rs1531797	1.22×10^{-6}	0.82	*USP36*		*USP36*
1	202547459	rs16853272	2.47×10^{-6}	1.90	*PLEKHA6*		
15	240948199	rs12148430	3.90×10^{-6}	0.73		*GABRB3*	
14	71978830	rs2283381	5.86×10^{-6}	1.24	*RGS6*		
20	60009590	rs17750066	5.98×10^{-6}	0.64	*TAF4*		
21	37158147	rs7282868	6.63×10^{-6}	1.27	*HLCS*		
17	62804441	rs8866	7.17×10^{-6}	1.24	*PITPNC1*		*PITPNC1*
1	183539106	rs1208517	8.73×10^{-6}	1.29	*IVNS1ABP*		

Abbreviations: eSNP, expression-associated single nucleotide polymorphism; rs, reference single nucleotide polymorphism (identification number); SNP, single nucleotide polymorphism.

[a] If multiple SNPs in linkage disequilibrium in the region were identified, the most significant SNP is noted, and the eSNP gene to any SNP in the region is noted. The *P* values reported are based on meta-analysis results from the Nurses' Health Study and the Health Professionals Follow-Up Study.

[b] *ADAMTS9*, ADAM metallopeptidase with thrombospondin type 1 motif, 9 gene; *GABRB3*, gamma-aminobutyric acid A receptor, beta 3 gene; *HLCS,* holocarboxylase synthetase gene; *IVNS1ABP,* influenza virus NS1A binding protein gene; *LY75*, lymphocyte antigen 75 gene; *PITPNC1,* phosphatidylinositol transfer protein, cytoplasmic 1 gene; *PLEKHA6,* pleckstrin homology domain containing, family A member 6 gene; *RBMS1,* RNA-binding motif, single-stranded interacting protein 1 gene; *RGS6,* regulator of G-protein signaling 6 gene; *TAF4,* TAF4 RNA polymerase II gene; *TCF7L2,* transcription factor 7-like 2 gene; *USP36,* ubiquitin-specific peptidase 36 gene.

gene (*RBMS1*). In silico replication (i.e., replication sought in available GWAS data from collaborators) of the SNP associations identified in this region was reported previously (12). The top associated SNPs in this region are not eSNPs but in tight linkage disequilibrium ($r^2 = 0.86$) to an eSNP (rs10929981) to the lymphocyte antigen 75 gene (*LY75*) in blood and adipose tissue (20). A region on chromosome 17 identified in our meta-analysis ($P = 1.22 \times 10^{-6}$) has been found previously as suggestive in the Wellcome Trust Case-Control Consortium type 2 diabetes mellitus case-control study ($P = 1 \times 10^{-4}$) (4). The SNPs identified in our study are eSNPs with strong associations with the ubiquitin-specific peptidase 36 gene (*USP36*) in multiple tissues (rs2279308; eSNP $P = 1.35 \times 10^{-99}$ (omental adipose), $P = 9.51 \times 10^{-79}$ (subcutaneous adipose), $P = 1.01 \times 10^{-69}$ (liver)). Another region of interest from our analysis on chromosome 17 (rs8866; $P = 7.17 \times 10^{-6}$) is an eSNP to the phosphatidylinositol transfer protein, cytoplasmic 1 gene (*PITPNC1*). Differences in results from those reported by Qi et al. (12) likely stem from the covariates used in this analysis, specifically age and body mass index, and the fact that these results are from the meta-analysis of imputed SNPs from the 2 studies. eSNP data for GWAS results with $P < 0.001$ are reported in Web Table 2 for reference.

SNP set enrichment analysis

Three gene sets were identified as significant at a false discovery rate threshold of 0.05. These gene sets are as follows:

- Human homolog of the Macrophage-enriched Metabolic Network (also known as the "MEM Network"; uncorrected $P = 0.0005$, false discovery rate = 0.027) (29);
- Gene Ontology "fat cell differentiation human" (uncorrected $P = 0.002$, false discovery rate = 0.035); and
- "Oxidative stress" (uncorrected $P = 0.001$, false discovery rate = 0.034) (30).

An additional, notable gene set is suggestive at a false discovery rate = 0.083: the black module derived from the Roux en Y Gastric Bypass (RNGB) cohort (21) (Table 2). The black module is a gene coexpression module, which was found to be correlated with body mass index and leptin levels from a cohort of extremely obese individuals.

In addition to the 53 gene sets tested, we also performed SSEA with 141 KEGG pathways, for comparison with recently reported results for the Wellcome Trust Case-Control Consortium type 2 diabetes mellitus cohort (9). Some of these pathways were included in our original 53 gene sets, such as the peroxisome proliferator-activated receptor gene (*PPAR*) signaling because of their known associations with type 2 diabetes mellitus. Eight pathways had an uncorrected significance level of $P < 0.05$ (Web Table 3). However, none of the KEGG pathways tested was found to be significantly enriched in analysis of our cohort with a false discovery rate of less than 0.2.

Table 2. Single Nucleotide Polymorphism Set Enrichment Analysis Results[a]

Gene Set Name	Source, Year (Reference No.)	Permutation *P* Value	FDR
MEM Network human homolog	Chen et al., 2008 (29)	0.0005	0.027
Oxidative stress	Furukawa et al., 2005 (30)	0.0013	0.034
Fat cell differentiation human	GO biologic process	0.0020	0.035
RYGB black module	Greenawalt et al., 2011 (21)	0.0063	0.083

Abbreviations: FDR, false discovery rate; GO, Gene Ontology; MEM, macrophage-enriched module; RYGB, Roux-en-Y gastric bypass.

[a] The results reported are from the meta-analysis of the primary model. Additional description of the gene sets can be found in Web Table 1 posted on the *Journal*'s Web site (http://aje.oupjournals.org/).

DISCUSSION

We have built on previously reported GWAS for type 2 diabetes mellitus through imputation, genetics of gene expression, and SSEA in an attempt to provide context for SNPs shown by previous studies to be associated with genes and pathways that are related to the development of type 2 diabetes mellitus that could be targeted for intervention. Imputation of HapMap SNPs to include >2,500,000 genotypes and use of additional covariates, such as body mass index and age, in our analysis strengthened the results in a number of regions of the genome including around *TCF7L2* and *ADAMTS9*, which have been implicated in type 2 diabetes mellitus GWAS previously. Between our analysis and that of a previously reported GWAS on this data set, associations in the region of *TCF7L2* were strengthened from 2.13×10^{-9} to 3.26×10^{-13}, by including imputed SNPs and body mass index as a covariate in the analysis. It has recently been reported that there is a significant interaction between the SNPs around *TCF7L2* and body mass index in this cohort when gene/environment interactions are explored (31). The region of *ADAMTS9* has been linked to type 2 diabetes mellitus in large GWAS previously (4). None of the significant SNPs identified within the region of *TCF7L2* or *ADAMTS9* to our knowledge has been associated with human gene expression in any genetics study of gene expression to date.

Through eSNP and network information, we were able to inform a number of single SNP associations with gene information that have only previously been identified as suggestive or associated with the nearest gene to the GWAS SNP in previous studies. Previous analysis of the Nurses' Health Study/Health Professionals Follow-up Study implicated *RBMS1* and the integrin beta 6 gene (*ITGB6*) as the genes of interest on chromosome 2q24; our most significant SNP, rs6718526, is in *RBMS1*. However, eSNP analysis identified an association between this SNP and *LY75*. *LY75* is a transmembrane receptor that plays a role in antigen processing and presentation, cellular defense response, and receptor-mediated endocytosis. The gene is overexpressed in the thyroid, blood, gut, pancreas, and subcutaneous adipose. We reviewed connections to *LY75* in mouse

Bayesian networks, which have previously been constructed from mouse cross quantitative trait loci studies for metabolic disorders (32). *LY75* was found to be directly connected to the expression of the pyruvate dehydrogenase kinase, isozyme 1, gene (*PDK1*), which is known to have a role in insulin resistance (Web Figure 2). Network analysis of the other implicated gene in the region of rs6718526, *RBMS1*, did not identify such associations.

The 2 suggestive associations identified on chromosome 17, rs1531797 and rs8866, were eSNPs to *USP36* and *PITPNC1*, respectively. *USP36* is a deubiquitinating enzyme that acts in regulation of nucleolar structure and function. *USP36* expression has also been shown to be controlled by the master regulators hepatocyte nuclear factor transcription factor, in the pancreas, which has been shown to contribute to type 2 diabetes mellitus, and the hepatocyte nuclear factor 4, alpha gene (*HNF4A*) (33). *PITPNC1* is a cytoplasmic phosphatidylinositol transfer protein, involved in lipid transport. To our knowledge, *PITPNC1* has not been associated with type 2 diabetes mellitus in the literature previously. While eSNP and network information does not provide a causal link between the SNP and the gene, it provides further clarity when exploring the biologic relevance of an SNP identified in a GWAS to a disease. This method is also useful for prioritizing SNPs/genes for replication or further laboratory-based follow-up.

Because of the current design of association studies, GWAS analysis will only identify relatively common variants with moderately large effects. For complex diseases in which a number of genes have small contributions to disease risk, the signal for most of these genes will not be detected above the false positives, given the multiple testing in a genome-wide analysis. For this reason, GSEA and SSEA, which test for enrichment of low *P* values in prehypothesized sets of genes, can be powerful when all or some of the genes in a pathway contribute small risk effects. Given the *P* values that must be achieved to reach genome-wide significance in a GWAS, SSEA allows interrogating suggestive SNPs in a biologically meaningful manner. SSEA allowed us to extend single SNP association results to gene sets of interest, through the use of SNPs that had previously been associated with the expression of genes of interest.

The MEM Network of genes was identified by Chen et al. in 2008 (34) as a group of genes that were causal for obesity in a mouse intercross between the apolipoprotein E null C57BL/6J mouse strain and the apolipoprotein E null C3H/HeJ mouse strain raised on a high-fat diet in order to induce metabolic and vascular disorder phenotypes. A similar gene set was later found to be associated with obesity in 2 human data sets (20, 21). The module contains 580 genes, 341 of which are represented by eSNPs. The gene set is enriched for genes involved in immune response (Fisher's exact test, corrected *P* value $(E) = 8.48 \times 10^{-14}$), defense response ($E = 3.29 \times 10^{-12}$), and response to stress ($E = 8.52 \times 10^{-14}$). It has long been known that type 2 diabetes mellitus is a proinflammatory disease, which involves activation of the innate immune system (for review, refer to Reference 35).

The Gene Ontology biologic process, fat cell differentiation human, includes 39 genes, 23 of which are represented by eSNPs (Table 3). This gene set includes a number of well-known type 2 diabetes mellitus-associated genes including the peroxisome proliferator-activated receptor gamma gene (*PPARG*), *TCF7L2*, and the insulin receptor gene (*INSR*). In our analysis, *TCF7L2* and *INSR* were not represented by eSNPs, but this gene set was still found to be significant, supporting the power and utility of this method. The oxidative stress gene set was selected because of its relation with insulin secretion and glucose transport (30). Interestingly, 2 of the gene sets identified as significant through SSEA are associated with immune response. However, gene sets that one might expect to find significant in a type 2 diabetes mellitus cohort analysis including the pathways type 2 diabetes mellitus, insulin signaling, and *PPAR* signaling did not achieve statistical significance. This could be due to a bias of the genes included in the analysis due to representation by eSNPs. However, 2 previous GSEA and SSEA studies on independent type 2 diabetes mellitus cohorts did not find these genes sets to be significant after multiple testing corrections either. The PPAR signaling pathway was previously found to have an enrichment $P = 0.029$ (false discovery rate = 0.2) by SSEA, and the type 2 diabetes mellitus pathway had a $P = 0.04$ (false discovery rate = 0.78) by GSEA (9, 36). Our analysis replicated the results found for the gene sets, "antigen processing" and "presentation and ether lipid metabolism" identified by Zhong et al. (9). Pathway analysis of the Wellcome Trust Case-Control Consortium on type 2 diabetes mellitus results through GSEA identified the Wingless-type mouse mammary tumor virus integration site family (WNT) signaling pathway as the top associated pathway; however, no other KEGG pathways were found to be significantly enriched (36). In our current study, the WNT signaling gene set achieved an uncorrected $P = 0.056$. Our study was conducted in a cohort that was predominately of European ancestry and, given its size, had relatively low power to identify significant, but weaker associations. Increased power in the GWAS would likely strengthen our SSEA results. Therefore, it is important to replicate these results in an independent larger cohort.

Our analysis included 53 gene sets, including KEGG pathways, gene ontologies, and gene sets of interest, linked to type 2 diabetes mellitus or associated traits. This is a supervised approach to gene selection that strengthened associations previously reported with SSEA and GSEA, which looked at KEGG pathways or gene ontology gene sets without filtering for relevance to type 2 diabetes mellitus. Although agnostic exploration of data can be useful, in this case directed exploration of known or suspected type 2 diabetes mellitus gene sets reduced the multiple testing burden and allowed us to identify 4 gene sets that were highly interesting. This supervised gene set selection method has not been tested on an independent type 2 diabetes mellitus cohort, and the results should be replicated to confirm their validity.

GWAS identify a small number of common variants that associate with susceptibility for disease. However, it is often difficult to link identified SNPs with biology relevant to the disease of interest. Through the use of eSNPs, we were able to link SNPs with suggestive significance to type

Table 3. Fat Cell Differentiation Human Gene Set and Accompanying eSNP Information[a]

Gene Symbol[b]	eSNP Tissue	eSNP rs id	GWAS *P* Value
CTSS	Omental adipose	rs10305724	0.05
ADAM12	Subcutaneous adipose	rs4962528	0.05
PPARD	Subcutaneous adipose	rs6906237	0.06
TAF8	Omental adipose	rs9381135	0.06
ADRB1	Liver	rs7915120	0.07
ERAP1	Omental adipose	rs13160562	0.07
HOXC4	Subcutaneous adipose	rs11614913	0.11
CCND1	Subcutaneous adipose	rs12808959	0.13
ADRB2	Omental adipose	rs7729953	0.24
BMPR1B	Omental adipose	rs1434546	0.26
GPD1	Liver	rs7532	0.29
KLF7	Liver	rs6732724	0.29
CBY1	Liver	rs4820345	0.34
ENPP1	Liver	rs9493120	0.34
SMAD3	Omental adipose	rs7166081	0.35
WWTR1	Omental adipose	rs10513355	0.38
PPARG	Liver	rs1797912	0.39
STEAP4	Liver	rs3745178	0.43
SOCS1	Omental adipose	rs7197119	0.44
UCP1	Omental adipose	rs17005845	0.65
MAP3K5	Omental adipose	rs9373173	0.69
LRRC8C	Omental adipose	rs2224652	0.74
SFRP1	Liver	rs6998193	0.79
ADIG			
BMP7			
CEBPA			
CEBPB			
HMGA2			
IL11			
INSR			
JARID1A			
KLF6			
MED1			
NFATC4			
NR5A2			
PPARGC1A			
RUNX1T1			
SAFB			
TCF7L2			

Table continues

Table 3. Continued

Abbreviations: dbSNP, National Institutes of Health single nucleotide polymorphism database; eSNP, expression-associated single nucleotide polymorphism; GWAS, genome-wide association studies; rs id, reference single nucleotide polymorphism identifier; SSEA, single nucleotide polymorphism set enrichment analysis.

[a] Genes represented by an eSNP in the SSEA are noted by tissue and dbSNP rs id information.

[b] *ADAM12*, ADAM metallopeptidase domain 12 gene; *ADIG*, adipogenin gene; *ADRB1*, adrenergic, beta-1-receptor gene; *ADRB2,* adrenergic, beta-2-receptor; *BMP7*, bone morphogenetic protein 7 gene; *BMPR1B*, bone morphogenetic protein receptor, type IB gene; *CBY1*, chibby homolog 1 gene; *CCND1*, cyclin D1 gene; *CEBPA*, CCAAT/enhancer binding protein, alpha gene; *CEBPB*, CCAAT/enhancer binding protein, beta gene; *CTSS*, cathepsin S gene; *ENPP1*, ectonucleotide pyrophosphatase/phosphodiesterase 1 gene; *ERAP1*, endoplasmic reticulum aminopeptidase 1 gene; *GPD1*, glycerol-3-phosphate dehydrogenase 1 gene; *HMGA2,* high mobility group AT-hook 2 gene; *HOXC4*, homeobox C4 gene; *IL11*, interleukin 11 gene; *INSR*, insulin receptor gene; *JARID1A,* lysine-specific demethylase 5A gene; *KLF6,* Kruppel-like factor 6 gene; *KLF7*, Kruppel-like factor 7 gene; *LRRC8C*, leucine-rich repeat containing 8 family, member C gene; *MAP3K5*, mitogen-activated protein kinase kinase kinase 5 gene; *MED1*, mediator complex subunit 1 gene; *NFATC4*, nuclear factor of activated T-cells calcineurin-dependent 4 gene; *NR5A2*, nuclear receptor subfamily 5, group A, member 2 gene; *PPARD*, peroxisome proliferator-activated receptor delta gene; *PPARG*, peroxisome proliferator-activated receptor gamma gene; *PPARGC1A*, peroxisome proliferator-activated receptor gamma, coactivator 1 alpha gene; *RUNX1T1*, runt-related transcription factor 1 gene; translocated to 1, gene; *SAFB*, scaffold attachment factor B gene; *SFRP1*, secreted frizzled-related protein 1 gene; *SMAD3*, SMAD family member 3 gene; *SOCS1*, suppressor of cytokine signaling 1 gene; *STEAP4*, STEAP family member 4 gene; *TAF8,* TAF8 RNA polymerase II gene; *TCF7L2*, transcription factor 7-like 2 gene; *UCP1*, uncoupling protein 1 gene; *WWTR1*, WW domain containing transcription regulator 1 gene.

2 diabetes mellitus to a number of genes of interest in insulin signaling and lipid metabolism. Although *TCF7L2* was not represented by an eSNP, in our analysis the gene set "fat cell differentiation human" was identified as significant, which contains *TCF7L2* as well as a number of other genes that have been linked to type 2 diabetes mellitus previously, further supporting the relevance of our analysis to type 2 diabetes mellitus. Further analysis of the other genes in this gene set may identify interesting genes or pathways to target for treatment of type 2 diabetes mellitus. As more combination therapies are developed, taking a pathway approach to drug development will likely increase success in identifying targets for treatment of complex disorders such as type 2 diabetes mellitus.

GWAS can be used to identify disease-associated variants that are relatively common and with relatively large effects. The rarer the variant or the smaller the effect, the larger the sample size needed, and sample sizes will always be limited by cost and feasibility. In some cases, adequate sample size can only be achieved through consortia and meta-analysis of numerous individual studies. On the other hand, in vivo work, model organisms, and other, more traditional, methods can create hypotheses about drivers of disease, which are difficult to validate in humans. We hope that our exploratory methods integrating GWAS data for a disease of interest, along with data on the genetics of gene expression for relevant tissues and hypotheses about gene sets of interest based on cell or animal model research, will allow further insight into the mechanism of disease in humans, beyond the information available from the top genome-wide significant SNPs.

ACKNOWLEDGMENTS

Author affiliations: Department of Genetics, Merck Research Laboratories, Boston, Massachusetts (Danielle M. Greenawalt); Sage Bionetworks, Seattle, Washington (Solveig K. Sieberts, Hua Zhong, Xia Yang, Justin Guinney); Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts (Marilyn C. Cornelis, Lu Qi, Frank B. Hu); Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts (Marilyn C. Cornelis, Lu Qi, Frank B. Hu); and Department of Epidemiology, Merck Research Laboratories, North Wales, Pennsylvania (Cynthia J. Girman).

Danielle M. Greenawalt and Solveig K. Sieberts contributed equally.

The Nurses' Health Study/Health Professionals Follow-up Study type 2 diabetes GWAS (U01HG004399) is a component of a collaborative project that includes 13 other GWAS funded as part of the Gene Environment-Association Studies (GENEVA) under the National Institutes of Health (NIH) Genes, Environment, and Health Initiative (GEI) (U01HG004738, U01HG004422, U01HG004402, U01H G004729, U01HG004726, U01HG004735, U01HG004415, U01HG004436, U01HG004423, U01HG004728, RFAHG 006033), with additional support from individual NIH Institutes (National Institute of Dental and Craniofacial Research: U01DE018993, U01DE018903; National Institute on Alcohol Abuse and Alcoholism: U10AA008401; National Institute on Drug Abuse: P01CA089392, R01DA013423; National Cancer Institute: CA63464, CA54281, CA136792, Z01CP010200). Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Genotyping was performed at the Broad Institute of Massachusetts Institute of Technology and Harvard, with funding support from the NIH GEI (U01HG04424), and the Johns Hopkins University Center for Inherited Disease Research, with support from the NIH GEI (U01HG004438) and the NIH contract, "High throughput genotyping for studying the genetic contributions to human disease" (HHSN 268200782096C).

The authors would like to thank Dr. Christine Suver, Dr. Pek Yee Lum, Dr. Cliona Molony, Dr. Ke Hao, and Dr. Eric Schadt for their support in planning and execution of this project.

Conflict of interest: none declared.

REFERENCES

1. *The Diabetes Atlas*. 4th ed. Brussels, Belgium: International Diabetes Federation; 2009.
2. Narayan KM, Boyle JP, Geiss LS, et al. Impact of recent increase in incidence on future diabetes burden: U.S., 2005–2050. *Diabetes Care*. 2006;29(9):2114–2116.
3. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106(23):9362–9367.
4. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*. 2008;40(5):638–645.
5. Tong Y, Lin Y, Zhang Y, et al. Association between *TCF7L2* gene polymorphisms and susceptibility to type 2 diabetes mellitus: a large Human Genome Epidemiology (HuGE) review and meta-analysis. *BMC Med Genet*. 2009;10(1):15. (doi:10.1186/1471-2350-10-15).
6. Cauchi S, El Achhab Y, Choquet H, et al. *TCF7L2* is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis. *J Mol Med (Berl)*. 2007;85(7): 777–782.
7. Yan Y, North KE, Heiss G, et al. Transcription factor 7-like 2 (*TCF7L2*) polymorphism and context-specific risk of impaired fasting glucose in African American and Caucasian adults: the Atherosclerosis Risk in Communities (ARIC) Study. *Diabetes Metab Res Rev*. 2010;26(5):371–377.
8. Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet*. 2010;86(6):929–942.
9. Zhong H, Yang X, Kaplan LM, et al. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet*. 2010;86(4):581–591.
10. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.
11. Raychaudhuri S, Plenge RM, Rossin EJ, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. International Schizophrenia Consortium. *PLoS Genet*. 2009; 5(6):e1000534. (doi:10.1371/journal.pgen.1000534).
12. Qi L, Cornelis MC, Kraft P, et al. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Meta-Analysis of Glucose and Insulin-related traits Consortium (MAGIC); Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium. *Hum Mol Genet*. 2010;19(13): 2706–2715.
13. Rimm EB, Giovannucci EL, Willett WC, et al. Prospective study of alcohol consumption and risk of coronary disease in men. *Lancet*. 1991;338(8765):464–468.
14. Colditz GA, Hankinson SE. The Nurses' Health Study: lifestyle and health among women. *Nat Rev Cancer*. 2005; 5(5):388–396.
15. Hughes TR, Mao M, Jones AR, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol*. 2001;19(4):342–347.
16. Li Y, Willer C, Sanna S, et al. Genotype imputation. *Annu Rev Genomics Hum Genet*. 2009;10(1):387–406.
17. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–575.
18. Aulchenko YS, Struchalin MV, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed

data. *BMC Bioinformatics*. 2010;11(1):134. (doi:10.1186/1471-2105-11-134).

19. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904–909.
20. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. *Nature*. 2008;452(7186):423–428.
21. Greenawalt DM, Dobrin R, Chudin E, et al. A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res*. 2011;21(7):1008–1016.
22. Schadt EE, Molony C, Chudin E, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*. 2008;6(5):e107. (doi:10.1371/journal.pbio.0060107).
23. Zhong H, Beaulaurier J, Lum PY, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet*. 2010;6(5):e1000932. (doi:10.1371/journal.pgen.1000932).
24. Ogata H, Goto S, Sato K, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 1999;27(1):29–34.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B*. 1995;57(1):289–300.
26. Dina C, Meyre D, Gallina S, et al. Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet*. 2007;39(6):724–726.
27. Salonen JT, Uimari P, Aalto JM, et al. Type 2 diabetes whole-genome association study in four populations: the DiaGen Consortium. *Am J Hum Genet*. 2007;81(2):338–345.
28. Grant SF, Thorleifsson G, Reynisdottir I, et al. Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat Genet*. 2006;38(3):320–323.
29. Chen Y, Zhu J, Lum PY, et al. Variations in DNA elucidate molecular networks that cause disease. *Nature*. 2008;452(7186):429–435.
30. Furukawa S, Fujita T, Shimabukuro M, et al. Increased oxidative stress in obesity and its impact on metabolic syndrome. *J Clin Invest*. 2004;114(12):1752–1761.
31. Cornelis MC, Tchetgen Tchetgen EJ, Liang L, et al. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am J Epidemiol*. 2012;175(3):191–202.
32. Bhattacharya S, Dey D, Roy SS. Molecular mechanism of insulin resistance. *J Biosci*. 2007;32(2):405–413.
33. Odom DT, Zizlsperger N, Gordon DB, et al. Control of pancreas and liver gene expression by HNF transcription factors. *Science*. 2004;303(5662):1378–1381.
34. Wang S, Yehya N, Schadt EE, et al. Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet*. 2006;2(2):e15. (doi:10.1371/journal.pgen.0020015).
35. Guest CB, Park MJ, Johnson DR, et al. The implication of proinflammatory cytokines in type 2 diabetes. *Front Biosci*. 2008;13(1):5187–5194.
36. Perry JR, McCarthy MI, Hattersley AT, et al. Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Wellcome Trust Case Control Consortium. *Diabetes*. 2009;58(6):1463–1467.