Gene prioritization based on biological plausibility over genome wide association studies renders new loci associated with type 2 diabetes

Silvia Sookoian, MD, PhD, Tomas Fernández Gianotti, MSc, Mariano Schuman, MSc, and Carlos Jose Pirola, PhD

Purpose: We present an approach to prioritize single nucleotide polymorphisms for further follow-up in genome-wide association studies of type 2 diabetes. **Method:** The proposed method combines both the use of open data access from two type 2 diabetes-genome-wide association studies (granted by the Diabetes Genetics Initiative and the Welcome Trust Case Control Consortium) and the comprehensive analysis of candidate regions generated by the freely accessible ENDEAVOUR software. **Results:** The algorithm prioritized all genes of the whole genome in relation to type 2 diabetes. There were six of 1096 single nucleotide polymorphisms in five genes potentially associated with type 2 diabetes: tachykinin receptor 3 (rs1384401), anaplastic lymphoma receptor tyrosine kinase (rs4319896), calcium channel, voltage-dependent, L type, alpha 1D subunit (rs12487452), *FOXO1A* (rs10507486 and rs7323267), and v-akt murine thymoma viral oncogene homolog 3 (rs897959). We estimated the fixed effect and *P* values of each single nucleotide polymorphism in the combined dataset by Mantel-Haenszel meta-analysis and we observed significant *P* values for all single nucleotide polymorphisms except for rs897959 at v-akt murine thymoma viral oncogene homolog 3. **Conclusion:** The proposed strategy may be used as an alternative tool for optimizing the information of the nearly 500,000 gene variants in which markers with modest significant *P* value for disease association are currently disregarded. Additionally, the said single nucleotide polymorphisms may be incorporated into the replication of the multistage design involved in the genome-wide association studies. ***Genet Med* 2009:11(5):338–343.**

Key Words: *genome-wide association studies, genes, type 2 diabetes, ENDEAVOUR, candidate genes*

Genome-wide association studies (GWAs) using a dense map of single nucleotide polymorphism (SNP) markers enable scientists to detect common genetic variants that influence susceptibility to complex diseases, enlightening both disease mechanisms, and the translation of this knowledge into clinical benefit for diagnosis, prognosis, and therapy.

Based on common human genetic variation information provided by the HapMap Project, the whole-genome scans using microarrays with 500,000 SNPs are making remarkable progress in the understanding of the genetic architecture of human diseases, including the constellation of complex diseases such as type 2 diabetes, dyslipidemias, central obesity, arterial hypertension, and fatty liver disease that gather in the metabolic syndrome among others.

Interestingly, one of the major outcomes of these studies is the elucidation of important aspects of disease pathogenesis through the discovery of novel genes or genomic regions previously unrelated to a disease. For instance, researchers from the Diabetes Genetics Initiative (DGI),[1] the Finland-United States Investigation of NIDDM Genetics (FUSION),[2] and the Welcome Trust Case Control Consortium (WTCCC)[3] reported the results of a consortium GWA study revealing the role of several novel loci such as *CDKN2B*, *CDKAL1* and *IGF2BP2* in the genetic risk of type 2 diabetes.

Surprisingly, data about previous published loci associated with type 2 diabetes were not sufficiently powerful to reach a significant *P* value in individual scans. For example, variants at *SLC30A8* and *PPARG* were significantly associated with type 2 diabetes only when pooling all the GWAs data, whereas in a single genome scan (DGI), no gene showed a positive signal (*P* value: 0.92 and 0.83, respectively). Thus, this may suggest that GWAs are still underpowered to find SNPs with small effect size.

In this regard, a challenging issue is the selection of "statistically significant" associations with those SNPs that show the most extreme *P* values (as small as 10^{-7})[4] followed by a robust replication that enables identification of a true positive signal. In principle, this approach strongly increases the weight of selected markers and has the virtue of avoiding the potential of false positive results. However, a worrying drawback is that if the *P* value for association for a given SNP in the initial study is not sufficiently small, the SNP will not be carried forward in the second stage of the analysis.[5] Thus, constraining the results of the variant selection to the small *P* values for association may exclude those SNPs that are biologically important for the disease and that are excluded from either the replication or the confirmation, and are also disregarded as potential predictors of a true effect.

One attractive methodology to circumvent the puzzle of choosing either a hypothesis-driven or an exploratory research may be the strategy of gene prioritization offered by the new bioinformatics tools based on the biological plausibility of a gene-disease association and on knowledge of the protein function.[6]

We propose an approach for expanding the selection of genes or loci of interest and prioritizing associations over GWAs related with genetic susceptibility to type 2 diabetes. The proposal profits from the recent initiatives of data sharing of the genome scan results that make the information publicly available as soon as they are generated and checked for quality. Both

From the Department of Molecular Genetics and Biology of Complex Diseases, Institute of Medical Research. A. Lanari, University of Buenos Aires-CONICET, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina.

Dr. Carlos J. Pirola, Instituto de Investigaciones Médicas A. Lanari., Combatiente de Malvinas 3150, Buenos Aires (1427), Argentina. E-mail: carlospirola@ciudad.com.ar or pirola.carlos@lanari.fmed.uba.ar.

Disclosure: The authors declare no conflict of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.geneticsinmedicine.org).

Submitted for publication October 15, 2008.

Accepted for publication December 19, 2008.

DOI: 10.1097/GIM.0b013e31819995ca

the DGI and the WTCCC are committed to embracing these principles as they made available all the phenotype-genotype data for type 2 diabetes.

MATERIALS AND METHODS

The proposed method combines both the use of the aforementioned open data access of the GWAs on the web sources and the comprehensive analysis of candidate regions generated by the freely accessible ENDEAVOUR software available at http://homes.esat.kuleuven.be/∼bioiuser/endeavor/endeavor.php.

ENDEAVOUR is a software application for the computational prioritization of candidate genes underlying biological processes or diseases, based on their similarity to known genes involved in a disease as previously described.[7] The hypothesis of prioritization by ENDEAVOUR is that candidate test genes are ranked based on their similarity with a set of known training genes. The rationale of this approach and the underlying principle for choosing different sets of 'training genes' for further gene prioritization are explained in detail in the **Appendix, Supplemental Digital Content 1**, http://links.lww.com/A1049. In addition, the complete list of the training genes, including both the Gene HGNC symbol, and gene name are shown in the **Appendix, Supplemental Digital Content 1**, http://links.lww.com/A1049. Moreover, from the freely available site http://www.broad.mit.edu/diabetes/, we downloaded the results of the GWA study in 3000 Scandinavian individuals about the genetic variants that influence the risk of type 2 diabetes (1464 patients with type 2 diabetes and 1467 matched controls). We also included in the analysis the results of the GWA study of type 2 diabetes performed by the WTCCC (2000 patients and 3000 control samples), which were downloaded from the freely available site http://www.wtccc.org.uk/info/summary_stats.shtml. Both consortiums used the GeneChip 500K Mapping Array Set (Affymetrix Human chip) for sample genotyping.

Type 2 diabetes cases in the DGI study were selected according to American Diabetes Association definitions of type 2 diabetes: fasting plasma glucose >7.0 mM or 2 hour postload glucose during an oral glucose tolerance test (OGTT >11.1 mM). To avoid confounding with type 1 diabetes, glutamate decarboxylase antibody Ab positive patients were excluded. Maturity-onset diabetes of the young (MODY) subjects from families with mutations in known MODY diabetes genes and diabetic individuals with onset age <35 years were excluded.

Type 2 diabetes cases in the WTCCC study were defined as follows: in each case, the diagnosis of diabetes was based on current prescribed treatment with sulfonylureas, biguanides, other oral agents and insulin or in the case of individuals treated with diet alone, historical or contemporary laboratory evidence of hyperglycaemia (as defined by the World Health Organization). Other forms of diabetes (for example, MODY, mitochondrial diabetes, and type 1 diabetes) were excluded by standard clinical criteria based on personal and family history. Criteria for excluding autoimmune diabetes included absence of first-degree relatives with T1D, an interval of ≥1 year between diagnosis and institution of regular insulin therapy and negative testing for antiglutamate decarboxylase.[8]

The public dataset of both GWAs shows the following information regarding the gene variants: SNP annotation dbSNP ID (rs) and physical mapping location, allele frequencies in affected and unaffected individuals, test of Hardy-Weinberg equilibrium, minor allele frequency, genotype counts in cases and controls, and association analysis results (*P* values). In addition, the DGI discloses the nearest gene name (HGNC Symbol) as provided by the genotyping platform (the DGI dataset has a column regarded as "GENE LIST" that belongs to the annotation of genes within 30 kb of the SNP). To select the SNPs in the prioritized genes, we used the information available on the DGI dataset to couple the same list to the SNPs of WTCCC dataset. This process guarantees that the SNPs selection in potential type 2 diabetes-candidate genes is made by the same strategy as the nearest gene information is given as we said before by the genotyping platform. We also double checked the SNPs information looking at the physical mapping location in both datasets, information also available on the NCBI SNP Reference Assembly.

A total of 386,731 markers were analyzed from the DGI database in both cases and controls for type 2 diabetes. Nominal *P* values for each subset of data were converted to *Z* scores based on the magnitude of significance and the direction of effect (based on the odds ratio estimated for each subset of data) as described by the authors.

In the WTCCC database, we incorporated for the analysis, a total number of 459,653 SNPs (including those SNPs that passed their quality control filters as did a study MAF >1%). The authors of the WTCCC reported the *P* values for both the additive and the general genetic model. The analyzed DNA samples by the WTCCC were restricted to the white 97% of the cohort. The samples tested are therefore estimated to be 99.8% white.

The different number of SNPs that pass the quality criteria of the genotyping assay in each study may explain differences on the number of markers between the two GWAs.

After applying the ENDEAVOUR algorithm for gene prioritization, we performed a search in the GWAs open data for the prioritized genes that showed, in both GWAs, a sign with a *P* value for the test of association smaller than 0.05 in one database and at least <0.08 in the other as a primary screening strategy (we name this *P* value as screening *P* value). This is the step where we joined both datasets and the screening *P* value cutoff was a condition that was required to be simultaneously present in both datasets to continue with the analysis.

To gain a better estimation of the effect, results from the different populations were combined by Mantel-Haenszel meta-analysis. Heterogeneity was evaluated with *Q* statistic and the I^2 statistic, a transformation of *Q* that estimates the percentage of the variation in effect sizes that is due to heterogeneity. *P* values were obtained by comparing the statistic with a χ^2 distribution with $k - 1$ degree of freedom (where *k* is the number of studies).[9] An I^2 value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity. For the combined analysis, control for multiple testing was done when applicable by Bonferroni correction, to obtain an empirical *P* value.

RESULTS

The tool ENDEAVOUR makes use of statistics to compute a ranking of test genes according to their similarity to the training genes. In a subsequent step, these rankings are integrated into a single ranking by making use of order statistics. In the **Appendix, Supplemental Digital Content 2**, http://links.lww.com/A1050, we show the list of the first 20 prioritized genes of 241 from the whole human genome (23.712 genes) with a significant association with the training set. The whole list is presented in the **Appendix, Supplemental Digital Content 3**, http://links.lww.com/A1051.

The ranking of the test genes is built by ENDEAVOUR after integrating the data into a mathematical model based on its similarity with the training genes. Vector-based data are scored by the Pearson correlation between a test profile and the training average, whereas attribute-based data are scored by Fisher's

Table 1 Results of SNPs with *P* values less than 0.08 for association with type 2 diabetes either in the DGI or in the WTCCC GWA study database for genes prioritized by the ENDEAVOUR software

NCBI SNP reference[a]	Type	*Z P* value by GC DGI	*P* value for additive genetic model WTCCC	Chr.	Gene (HGNC symbol)	Physical position (NCBI build 35)	Minor allele	Major allele	MAF	Validation status[b]	Gene function[c]
rs1384401	Intronic	0.0107	0.0147	4	*TACR3*	104922682	A	G	0.3395	Proven by cluster, frequency, submitter, double hit	Unknown
rs4319896	Intronic	0.01487	0.0394	2	*ALK*	30051500	A	C	0.4459	Proven by cluster, frequency, double hit	Orphan receptor with a tyrosine-protein kinase activity
rs12487452	Intronic	0.01826	0.0241	3	*CACNA1D*	53767451	G	C	0.1939	Proven by frequency, double hit	Calcium voltage-gated channel
rs10507486	Intronic	0.0676	0.00149	13	*FOXO1A*	40084501	A	G	0.2155	Proven by cluster, frequency	Negative regulator of insulin sensitivity
rs7323267	Intronic	0.07172	0.0013907	13	*FOXO1A*	40102015	C	T	0.2058	Proven by cluster, frequency, double hit	
rs897959	Intronic	0.022	0.080	1	*AKT3*	240223815	C	T	0.3552	Proven by cluster, frequency	Key regulator for cell growth, cell survival and metabolic insulin action

[a] NCBI SNP reference.

[b] SNP tested and validated by a noncomputational method. Validation status according to Ensemble.

[c] Gene function according to Gene Atlas and OMIM.

GC, *P* value of Test statistic after correction by Genomic Control; Gene, annotation of genes within 30kb of the SNP; MAF, minor allele frequency. Gene (HGNC symbol).

omnibus analysis on statistically over represented training attributes.[7] Therefore, we analyzed the data regarding 1096 SNPs in the 241 prioritized genes.

Six of 1096 SNPs located in five prioritized genes showed the screening *P* value cutoff ($P < 0.08$) for association with type 2 diabetes in the DGI and WTCCC datasets (Table 1 lists the SNPs information).

The tachykinin receptor 3 of unknown function (*TACR3*) showed a potential association with type 2 diabetes represented by 1 SNP in both databases (rs1384401). In addition, *FOXO1A*, a negative regulator of insulin sensitivity in liver, adipocytes, and pancreatic beta cells that acts downstream of the insulin signaling pathway, showed two highly correlated SNPs potentially associated with type 2 diabetes in the DGI and in the WTCCC databases (rs10507486 and rs7323267). Calcium channel, voltage-dependent, L type, alpha 1D subunit (*CACNA1D*), anaplastic lymphoma receptor tyrosine kinase (*ALK*), and v-akt murine thymoma viral oncogene homolog 3 (*AKT3*, protein kinase B, gamma) showed 1 SNP, which screening *P* values are also shown in Table 1. Because genotype frequencies were available in both the WTCCC and the DGI datasets, we further investigated whether the risk allele (and then the effect direction) for each SNP potentially associated with type 2 diabetes was the same in both datasets. To strengthen the results, we estimated the fixed effect and *P* value of each SNP in the combined dataset by Mantel-Haenszel meta-analysis. We observed significant *P* values for all SNPs except for rs897959 of *AKT3* (Table 2), without evidence of heterogeneity in over 3300 cases and 4300 controls. It is worth noting that there was heterogeneity between both studies for this marker but several SNPs of the *AKT3* gene were associated with HOMA index in the DGI dataset (data not shown).

Finally, after testing different training set of genes as explained in Methods section, we observed that four of five genes initially prioritized by the ENDEAVOUR software not only remained in the same ranking position after applying a diverse training set but also some of them (*FOXO1A* and *AKT3*) improved the ranking position when using the new version of the software that make use of a training set only for type 2 diabetes (data not shown). We tested by both Kruskal-Wallis Test and Mann-Whitney test whether the difference in the ranking position was statistically significant and both tests showed that they were not statistically different ($P = 0.9$ and $P < 0.06$, respectively).

DISCUSSION

Compared with traditional ways of identifying disease-associated genes, GWAs generate an amount of data of four or five orders of greater magnitude because they assess roughly 500,000 SNPs in a single sample. Although GWAS are not intended to be hypothesis oriented, the data generated by them can be used to test candidate-gene associations. However, at that point, a comprehensive empirical analysis of candidate gene sets is impractical and requires statistical approaches for the analysis and interpretation of the data.

The disease-associated SNPs are often measured assuming the stringent criteria of choosing a cutoff *P* value for association as small as 10^{-7} by applying Bonferroni's criteria.[5,10]

Table 2 Estimation of the effect of each SNP associated with type 2 diabetes in the combined dataset (DGI and WTCCC GWA study)

NCBI SNP reference	Gene (HGNC symbol)	Genotype	Risk allele[a]	No. cases/ controls	OR (95% CI)	Cumulative OR (95% CI)	Nominal *P* value	Empirical *P* value	Test for heterogeneity[b]
rs1384401	*TACR3*	AA*	G	3374/4394			3.9×10^{-4}	0.0023	0.57
		AG			1.24 (1.07–1.46)				
		GG			1.34 (1.15–1.58)	1.17 (1.07–1.27)			
rs4319896	*ALK*	CC*	A	3379/4387			3.1×10^{-3}	0.018	0.64
		AC			1.15 (1.02–1.28)				
		AA			1.23 (1.08–1.39)	1.15 (1.05–1.25)			
rs12487452	*CACNA1D*	GG*	C	3339/4368			1.2×10^{-3}	0.007	0.76
		GC			1.16 (0.87–1.56)				
		CC			1.34 (1.01–1.78)	1.18 (1.06–1.30)			
rs10507486	*FOXO1A*	AA*	G	3382/4399			5.1×10^{-4}	0.003	0.45
		AG			1.09 (0.87–1.37)				
		GG			1.27 (1.02–1.59)	1.18 (1.07–1.29)			
rs7323267	*FOXO1A*	CC*	T	3382/4397			2.8×10^{-4}	0.001	0.56
		CT			1.07 (0.85–1.36)				
		TT			1.27 (1.01–1.60)	1.19 (1.08–1.31)			
rs897959	*AKT3*	CC*	T/C	3382/4396			0.91	5.46	8.4×10^{-3}
		CT			1.03 (0.89–1.20)				
		TT			1.02 (0.88–1.19)	0.99 (0.91–1.08)			

[a] Risk allele in both datasets.

[b] *P* value for heterogeneity. Odds ratios (OR) and 95% confidence intervals (95% CI) toward the first genotype (the reference genotype as indicated by an asterisk) for the other two genotypes is indicated. *P* value stands for one-sided alternative (cases > controls) significance from the extended Mantel-Haenszel (MH) test for trend for the combined studies. MH cumulative OR (95% CI) for the combined studies is also shown. Cumulative OR using proportional odds model [22] stands for the cumulative effect of the two genotypes (heterozygous and homozygous for the risk allele (and then risk allele copy) in comparison with homozygous for the nonrisk allele. Control for multiple testing was done by Bonferroni correction to obtain an empirical *P* value. Risk allele for rs897959 is not the same in both datasets (T is the risk allele in the DGI dataset and C is the risk allele in the WTCCC dataset).

An interesting strategy that may help to distinguish a chromosomal region in which a disease causal gene is expected to lie is the use of computational instruments that can score, based on likelihood, candidate genes involved in a disease or biological process by combining an optional number of heterogeneous sources of information.[7]

To take advantage of the open availability of GWAs data we investigated, whether the computational disease gene prioritization method proposed by the bioinformatic tool ENDEAVOUR could help differentiate, in the bulk of GWA, genes that are more or less likely to influence genetic susceptibility to type 2 diabetes in white subjects.

Thus, this study yielded an additional list of candidate genes potentially associated with type 2 diabetes that were overlooked in the original GWAs as they did not show a sufficiently small *P* value for association. It is important to mention that we are showing just five additional SNPs worthy of follow-up and replication in other studies; thus, we are not claiming significant association with the disease.

The rationale of the proposed approach may be extended to populations with different genetic backgrounds. However, even though many SNPs associated with type 2 diabetes in the three major type 2 diabetes GWAs (DGI, FUSION, and WTCCC) were replicated in non-white population,[11] the results we are showing regarding additional type 2 diabetes association signals should be particularly confined to white subjects. Differences in genetic background, linkage disequilibrium structure, and environmental exposures may differ across populations and may thus explain that a true susceptibility gene for type 2 diabetes in one population might not be readily replicated in other population.[12]

Although the identification and characterization of variants associated with type 2 diabetes was thoroughly evaluated in the previously mentioned GWAs, a potentially associated sign for two intronic SNPs in the transcription factor *FOXO1A* suggests that *FOXO1A* may be considered a putative candidate gene. For instance, by combining the genotype information of both databases by Mantel-Haenszel meta-analysis, we observed an effect (cumulative odds ratio using proportional odds model 1.182, 95% confidence interval: 1.07–1.29, $P = 0.001$, without heterogeneity) similar to the one observed for those genes that were reported as significantly associated with type 2 diabetes in both studies.[1,8] Supporting this observation, it was reported in animal models that the haploinsufficiency of the *FOXO1A* restores insulin sensitivity and rescues the diabetic phenotype in insulin-resistant mice and that, conversely, a gain-of-function *FOXO1A* mutation results in diabetes.[13] This may be the reason that explains *FOXO1A* is regarded as a potential therapeutic target for improving insulin resistance[14] and also may be a contributor to type 2 diabetes.

Similarly, other genes were prioritized: *ALK*, *CACNA1D,* and *TACR3*. *ALK* was originally identified as a member of the insulin receptor subfamily of receptor tyrosine kinases, an orphan receptor in vertebrates, and seems to be particularly expressed in hypothalamic and sympathetic chain neurons and the ganglion cells of the gut indicating its possible role in regulating energy balance.[15]

The *CACNA1D* gene encoding calcium channel, voltage-dependent, L type, alpha 1D subunit may participate in the regulation of insulin secretion.[16]

The *TACR3* gene encodes the tachykinin NK (3) receptor of still unknown function. But the presence of tachykinin3 and its receptor (TACR3) in a wide variety of peripheral tissues argue for a still unexplored role of this system in mediating visceral effects of tachykinins.[17]

Although the *AKT3* variant was not significantly associated with type 2 diabetes in the pooled dataset, more studies may be necessary to determine its role in the disease, because the protein encoded by the *AKT3* gene is a member of the AKT, also called PKB, serine/threonine protein kinase family. AKT kinases are known to be regulators of cell signaling in response to insulin and growth factors. They are involved in a wide variety of biological processes, including cell proliferation, differentiation, apoptosis, tumorigenesis, and glycogen synthesis and glucose uptake.[18]

A criticism of our approach might be that we searched by "gene name" for the prioritized genes in the databases (DGI) to perform thereafter a joined analysis in both GWAs coupling the same list of SNPs in the prioritized genes. This list is based on genes nearby the SNPs based on certain criteria. However, this is not always correct. Often it is not clear to what gene a SNP belongs, and it is not necessarily the closest one. Because of linkage desequilibrium (LD) structures in the genome, SNPs are sometimes linked to multiple genes, and it is not clear which gene is functionally causal. Nevertheless, as the raw datasets are not available to the public, we could not calculate haploblocks and map SNPs back to genes, based on LD. Nevertheless, an important point is that in the Mapping 500K Array Set, SNP annotation includes dbSNP ID, nearest gene, physical map location, cytoband, and allele frequencies in multiple populations, a reasonable strategy to offer accuracy in the SNP localization.

The advantage of our approach is that the predisposing SNP selection does not only rely on the most extreme P values but also reinforces the plausible biological relevance of the association. In searching for 240 genes (roughly 1000 SNPs), some associations with $P < 0.05$ may emerge by chance, because 1% of the SNPs meet this condition in both databases. However, we believe that the concurrent biological plausibility is extremely improbable. Furthermore, the correlation between P values of both databases is extremely low (for instance, for type 2 diabetes it is NS, Spearman R: -0.0028, $P = 0.2496$). Hence, we propose to give special treatment to the analysis of the markers at the prioritized genes across the genome to assist in further selection of SNPs that will be considered for replication.

A note of caution should be added. In the analysis of the GWAs datasets, we only included the prioritized gene list built by ENDEAVOUR, and the list of genes included in the training set was not initially evaluated. However, we performed an additional evaluation of the association analyses based on the reported P values in the GWAs on the training gene list. Remarkably, the *TCF7L2* gene (represented by five SNPs significantly associated with type 2 diabetes in DGI, WTCCC, and also the FUSION study)[2] was part of the training set; however, it is most likely because *TCF7L2* was originally identified as a gene with strong T2D association.

We emphasize that our proposal complements the current statistical methods used to test true associations. Markers yielding modest P values may be considered alongside the other markers that show P values in the order of 10^{-7}. Of course, our procedure does not contemplate the unexpected markers with unknown functions, new mechanistic connections, or unsuspected contributions to the disease, as the nature of the aforementioned procedure is just oriented to those genes of some biological importance as candidate genes.

It is worth mentioning that other reports shared the concern about deciding the SNPs in GWAs that merit follow-up and further replication analysis. Chen et al.[19] recently proposed an approach for selecting SNPs based on a hierarchical model. This approach, which is not strictly based on biological plausibility of candidate's genes, allows the users to incorporate existing information about the SNPs into the analysis. For instance, the algorithm ranks P values assuming a weighting function that incorporates prior information about linkage or association evidence.

Almost simultaneously, Lewinger et al.[20] proposed a similar approach based on hierarchical regression modeling to select a subset of markers from the first stage of a GWAs. The model, rather than selecting the most significant marker-disease associations at some cutoff, is based on a prior model for the true noncentrality parameters of these associations composed of a large mass at zero and a continuous distribution of nonzero values. Resembling the previously mentioned study, the proposal of Lewinger et al. also allows the consideration of various covariates that characterize each marker, such as their location relative to genes or evolutionary conserved regions, or prior linkage or association data. But, none of these studies include in the analysis existing data from GWAs.

Finally, a recent study identified additional susceptibility loci for type 2 diabetes by performing a meta-analysis of three published GWAs.[21] As acknowledged by the authors, GWAs are limited by the modest effect sizes of individual common variants and the need for stringent statistical thresholds. Thus, by combining data involving 10,128 samples, the authors found in the initial stages of the analysis highly associated variants (they followed only 69 signals out of over 2 million meta-analyzed SNPs) with P values $<10^{-4}$ in unknown loci, and 11 of these type 2 diabetes' associated SNPs were taken forward to further stages of analysis. Large stage replication testing allowed the detection of at least six previously unknown loci with robust evidence for association with type 2 diabetes.

As a final approach to support the potential usefulness of our proposal, we performed an additional test. We chose the first 20 genes prioritized by ENDEAVOUR, and related to these genes there were 1729 SNPs. We simultaneously chose 20 genes in position 20,000 (the whole genome captured by ENDEAVOUR has 23,712 genes). In relation to these genes, there were 534 SNPs. We further explored in the Diabetes Genetics Replication And Meta-analysis Consortium data available at http://www.well.ox.ac.uk/DIAGRAM/ how many SNPs were found among the above mentioned 1729 SNPs and how many SNPs were found among the 534 SNPs. Interestingly, among the first 20 prioritized genes there were 88 SNPs with a P value <0.05 in the Diabetes Genetics Replication And Meta-analysis Consortium dataset. However, among the last prioritized genes, only 10 SNPs showed a P value <0.05. This difference was statistically significant ($P < 0.003$, χ^2). Hence, we can assume that our proposal has a potential strength to prioritize SNPs for further follow-up in GWAs.

In conclusion, the proposed strategy may be used as an alternative tool for optimizing the information of the nearly 500,000 gene variants in those markers with modest significant P values for disease association (close to 10^{-3}–10^{-4}), which are ignored as they are not satisfactorily small and are dumped into a sea of false positives results.[10] Additionally, the said SNPs may be incorporated in the replication of the multistage design involved in the GWA studies.

We wish to point out that data sharing of GWA studies offers an unprecedented opportunity to generate new hypotheses and explore the association between specific genes and disease. The strategy devised in this report is exemplified for type 2 diabetes but would be applicable to the metabolic syndrome and related traits as well.

We propose that a further gain can be achieved by data sharing with a consensus format among databases (for example by providing raw allele or genotype frequencies with clear indication of strand location as in the WTCCC dataset).

ACKNOWLEDGMENTS

This study was partially supported by Grants B119 (Universidad de Buenos Aires), PICT 05-25920 (Agencia Nacional de Promoción Científica y Tecnológica), PICT 2006-124, and PIP 5195 (Consejo Nacional de Investigaciones Científicas y Técnicas). S.S., M.S., and C.J.P. belong to Consejo Nacional de Investigaciones Científicas.

REFERENCES

1. Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;316:1331–1336.
2. Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;316:1341–1345.
3. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007;316:1336–1341.
4. Williams SM, Canter JA, Crawford DC, Moore JH, Ritchie MD, Haines JL. Problems with genome-wide association studies. *Science* 2007;316:1840–1842.
5. Hunter DJ, Kraft P. Drinking from the fire hose–statistical issues in genomewide association studies. *N Engl J Med* 2007;357:436–439.
6. Elbers CC, Onland-Moret NC, Franke L, Niehoff AG, van der Schouw YT, Wijmenga C. A strategy to search for common obesity and type 2 diabetes genes. *Trends Endocrinol Metab* 2007;18:19–26.
7. Aerts S, Lambrechts D, Maity S, et al. Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006;24:537–544.
8. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–678.
9. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–560.
10. Clark AG, Li J Conjuring SN. Ps to detect associations. *Nat Genet* 2007;39:815–816.
11. Hayes MG, Pluzhnikov A, Miyake K, et al. Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. *Diabetes* 2007;56:3033–3044.
12. Rampersaud E, Damcott CM, Fu M, et al. Identification of novel candidate genes for type 2 diabetes from a genome-wide association scan in the Old Order Amish: evidence for replication from diabetes-related quantitative traits and from independent populations. *Diabetes* 2007;56:3053–3062.
13. Nakae J, Biggs WH III, Kitamura T, et al. Regulation of insulin action and pancreatic beta-cell function by mutated alleles of the gene encoding forkhead transcription factor Foxo1. *Nat Genet* 2002;32:245–253.
14. Samuel VT, Choi CS, Phillips TG, et al. Targeting foxo1 in mice using antisense oligonucleotide improves hepatic and peripheral insulin action. *Diabetes* 2006;55:2042–2050.
15. Morris SW, Naeve C, Mathew P, et al. ALK, the chromosome 2 gene locus altered by the t(2;5) in non-Hodgkin's lymphoma, encodes a novel neural receptor tyrosine kinase that is highly related to leukocyte tyrosine kinase (LTK). *Oncogene* 1997;14:2175–2188.
16. Liu G, Jacobo SM, Hilliard N, Hockerman GH. Differential modulation of Cav1.2 and Cav1.3-mediated glucose-stimulated insulin secretion by cAMP in INS-1 cells: distinct roles for exchange protein directly activated by cAMP 2 (Epac2) and protein kinase A. *J Pharmacol Exp Ther* 2006;318:152–160.
17. Pinto FM, Almeida TA, Hernandez M, Devillier P, Advenier C, Candenas ML. mRNA expression of tachykinins and tachykinin receptors in different human tissues. *Eur J Pharmacol* 2004;494:233–239.
18. Dummler B, Hemmings BA. Physiological roles of PKB/Akt isoforms in development and disease. *Biochem Soc Trans* 2007;35:231–235.
19. Chen GK, Witte JS. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet* 2007;81:397–404.
20. Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol* 2007;31:871–882.
21. 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;40:638–645.
22. Liu IM, Agresti A. Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics* 1996;52:1223–1234.