Prioritization of candidate disease genes for metabolic syndrome by computational analysis of its defining phenotypes

Nicki Tiffin,1 Ikechi Okpechi,2 Carolina Perez-Iratxeta,3 Miguel A. Andrade-Navarro,3,4 and Rajkumar Ramesar1

1Division of Human Genetics, MRC Human Genetics Research Unit, Institute for Infectious Diseases and Molecular Medicine, 2Nephrology Division and Department of Medicine, Groote Schuur Hospital, University of Cape Town, Cape Town, South Africa; 3Ontario Genomics Innovation Centre, Ottawa Health Research Institute; and 4Department of Cellular and Molecular Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

Submitted 12 May 2008; accepted in final form 7 July 2008

Tiffin N, Okpechi I, Perez-Iratxeta C, Andrade-Navarro MA, Ramesar R. Prioritization of candidate disease genes for metabolic syndrome by computational analysis of its defining phenotypes. Physiol Genomics 35: 55–64, 2008. First published July 8, 2008; doi:10.1152/physiolgenomics.90247.2008.—There is a rapid increase in the world-wide burden of disease attributed to metabolic syndrome, as defined by co-occurrence of an array of phenotypes including abdominal obesity, dysglycemia, hypertriglyceridemia, low levels of high density lipoprotein cholesterol, and hypertension. Familial studies clearly indicate a genetic component to the disease and many linkage studies have identified a large number of linked loci. No disease-causing genes, however, have been conclusively identified, most likely because this is a multifactorial disease for which effects of many causative genes may be small and combined with environmental effects. To assist empirical identification of metabolic syndrome associated genes, we present here a novel computational approach to prioritize candidate genes. We have used linkage studies and the clinical and population-specific presentation of the disease to select a final candidate gene list of 19 most likely disease-causing genes. These are predominantly involved in chylomicron processing, transmembrane receptor activity, and signal transduction pathways. We propose here that information about the clinical presentation of a complex trait can be used to effectively inform computational prioritization of disease-causing genes for that trait.

Candidate genes; complex disease

METABOLIC SYNDROME IS AN INCREASINGLY common complex disease that is defined by co-occurrence of an array of metabolic phenotypes. These phenotypes have been classified in multiple ways (Ref. 2 and reviewed in Ref. 49) to provide a list of simple clinical markers that can diagnose individuals with the syndrome. The guidelines of the 2001 National Cholesterol Education Program-Adult Treatment Panel III are in common use and characterize metabolic syndrome as the co-occurrence of any three of the five following phenotypes: abdominal obesity, dysglycemia including insulin resistance/hyperinsulinemia, increased triglycerides, decreased high-density lipoprotein (HDL) cholesterol, and hypertension (2, 17, 22, 44). Currently, ~25% of North Americans are thought to be affected by metabolic syndrome (15). Obesity is the prevalent phenotype and is rising as a global epidemic (1) that is likely to result in a concurrent increase in metabolic syndrome incidence (42, 54) and related morbidity and mortality.

Familial studies clearly indicate a heritable component to metabolic syndrome (reviewed in Ref. 49); however, multiple linkage studies have generated a large number of metabolic syndrome-associated loci (Supplementary Data File S1) without subsequent identification of etiological genes. To date, existing approaches have in general failed to identify genes underlying complex diseases or traits such as metabolic syndrome, as they often present with a wide range of phenotypes and generally involve multiple etiological mechanisms and contributing genes (20). Also, the number of genes falling within the identified loci is generally far too large for further empirical analysis (47).

As a complementary approach, many computational methods have been developed to prioritize candidate genes by mining existing biological data (28). A selection of such methods that analyze gene and protein sequence data, ortholog data, expression data, Gene Ontology (GO) (23) and eVOC (29) annotation and biomedical literature has been previously used in concert in a study to select and/or prioritize most likely candidate disease genes from regions identified by linkage analyses in a case study of Type 2 diabetes and the related trait obesity (57). The rationale for using multiple computational methods is to select candidate genes that fulfill as many disease-relevant criteria as possible in a complementary approach. In prioritizing genes that have been selected by all methods we prioritize candidate genes that fulfill all criteria used to select most likely disease-causing genes across all the methods. In the current study, we build on this approach by additionally incorporating complex clinical presentation data with computational analyses of candidate genes, for the complex disease metabolic syndrome and its defining phenotypes. We include in our study the additional analysis of the individual phenotypes described in metabolic syndrome and the relative frequencies with which the different symptoms occur in several patient populations. Although data-mining of biological and disease-related data entails inherent bias to existing knowledge, this knowledge can be applied to effectively guide the prioritization of most likely disease candidates for further empirical analysis, particularly in the scenario where no other discriminating characteristics identify candidate genes. The

1 The online version of this article contains supplemental material.

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.
phenotype complexity of metabolic syndrome makes it an excellent candidate for such computational analysis to identify most likely disease gene candidates for subsequent empirical validation.

The relationships between candidate genes in the final selected list and their involvement in common pathways and protein interactions are studied with a variety of online tools: the genes selected are mostly involved in lipid transport and processing, the chylomicron, and signal transduction at the plasma membrane. Such functions are appealing targets for the search for genetic causes of metabolic syndrome as each signal transduction pathway has multiple potential activating events and a variety of downstream targets and effects and could therefore be responsible for the diverse array of phenotypes seen in metabolic syndrome. Also, these functions are fundamental to maintaining an appropriate balance between the intracellular and extracellular environments, and disruption could lead to the homeostatic imbalances seen in metabolic syndrome.

In summary, we combine multiple computational approaches for candidate disease gene selection and give candidate genes a clinically relevant score that is based on the defining phenotypes of the disease and their frequency of occurrence in several populations. We use this score to select a short list of most likely candidate genes for metabolic syndrome and discuss their etiological relevance.

**METHODS**

**Overview.** An overview of the analysis is shown in Fig. 1. First, we use multiple computational methods (described below) to analyze the starting set of candidate genes. The methods access an array of biological databases and analyze the candidates according to different criteria. Candidate genes are scored and ranked according to the number of methods that select them (as described in Collation of gene lists and construction of a scoring matrix, below). We have selected the top-scoring genes to generate lists of the most likely candidates for metabolic syndrome and, in parallel, each of the phenotypes defining this syndrome (abdominal obesity, dysglycemia, hypertension, hypertriglyceridemia, low HDL cholesterol - see Refs. 2, 49 for full definitions of phenotypes). Next, we assigned a score to each gene in each phenotype candidate list, using a weighting system that depends on the frequency with which each phenotype appears in patients diagnosed with metabolic syndrome. Finally, every gene is given a total score depending on the score it received in each of the phenotype candidate lists. This total score is used to rank the genes. As the phenotype frequencies vary between populations we have repeated this analysis using weightings appropriate to the white population and to the black population as reported by Kraja et al. 2005 (32). Additionally, phenotype frequencies vary between males and females, and we have repeated the analysis using the corresponding frequencies reported by McCarthy et al. 2003 (39). We repeated the analysis for black/white/male/female population groups using the data from女孩 and Frank et al. (2002) (17) and generated a slightly more stringent list containing matching candidates. We have used the more inclusive analysis in this study. Thus, the final candidate gene list consists of genes that are selected in the search for candidate genes for metabolic syndrome and are also independently selected in the parallel selection of candidate genes for the five phenotypes, depending on the frequencies with which the phenotypes occur in four population groups.

**Generating the starting set of candidate genes.** Linkage studies for metabolic syndrome were used as a primary filter to select the starting set of candidate genes (see Supplementary Data File S1 for references and LOD scores). Cytogenetic loci were used to define linked regions from which candidates were selected, to ensure that potential candidate genes were not excluded due to changes or inconsistencies in marker annotation and identification. Genes were compiled from linked cytogenetic loci using the Ensembl database Ensembl_mart_41 into a local database for further analysis. Database versions differ due to ongoing database updates and differences occurring between gene sets in different databases. This is accommodated in part in this study by the scoring system that discriminates between genes that are rejected as poor candidates and those that are not analyzed at all by a method, avoiding biased negative scoring against those genes not measured by all methods because of database differences (see Collation of gene lists and construction of a scoring matrix).

**Compilation of search terms and training sets of disease genes.** Many of the methods require various user-defined inputs, and these are shown in table 1. These include disease search terms for querying abstracts in PubMed and anatomy terms used for GeneSeeker queries. Also shown are the top ranking eVOC terms selected by text-mining of PubMed abstracts containing the disease name. Disease name synonyms were used as described in the Online Mendelian Inheritance in Man (OMIM) database (40). The MeSH terms used for each disease by O2D are as follows. For metabolic syndrome, we used all MeSH terms annotated in the literature references from OMIM entry 605552. For its phenotypes, we used “obesity,” “hypertriglyceridemia,” “hypertension,” “diabetes mellitus,” and “cholesterol, HDL.” The SUSPECTS software (4) automatically retrieves genes implicated in a disorder from OMIM, the Human Gene Mutation Database (55), and Genetic Association Database (8). These training sets were used for each of the phenotypes, using the search terms outlined in Table 1, column 6. For “dysglycemia,” no associated genes were found and a training set of “diabetes”-associated genes was used, as this is a very closely related condition. For “low HDL cholesterol” no associated genes were found, and “HDL cholesterol”-associated genes were used.
Table 1. Summary of search terms and training sets for metabolic syndrome and associated phenotypes

<table>
<thead>
<tr>
<th>Disease/Phenotype</th>
<th>Search Terms for PubMed</th>
<th>Abstracts, n</th>
<th>Terms for GeneSeeker</th>
<th>Top eVOC Terms</th>
<th>SUSPECTS and Endeavour Training Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominal obesity</td>
<td>“abdominal obesity,” “increased waist circumference”</td>
<td>1,098</td>
<td>adipose, kidney, heart, arteries, intestine</td>
<td>“cardiovascular vascular,” “blood peripheral blood,” “heart,” “oral cavity”</td>
<td>“abdominal obesity” NR3C1, LPL, ADRA2A</td>
</tr>
<tr>
<td>Hypertriglyceridemia</td>
<td>“hypertriglyceridemia,” “elevated triglycerides”</td>
<td>5,543</td>
<td>adipose, liver, intestine, arteries</td>
<td>“cardiovascular vascular,” “blood peripheral blood,” “heart,” “hepatocellular liver”</td>
<td>“hypertriglyceridemia” APOA1, APOC3</td>
</tr>
<tr>
<td>Low HDL cholesterol</td>
<td>“low HDL cholesterol”</td>
<td>755</td>
<td>adipose, liver, intestine, arteries</td>
<td>“cardiovascular vascular,” “blood peripheral blood,” “heart,” “blood peripheral blood,” “artery”</td>
<td>“HDL cholesterol” PPARG, GHR, LIPG, CETP, SCARB1, ABCA1</td>
</tr>
<tr>
<td>Hypertension</td>
<td>“hypertension,” “high blood pressure”</td>
<td>most recent 5,000</td>
<td>kidney, arteries, heart</td>
<td>“cardiovascular vascular,” “blood peripheral blood,” “heart,” “artery”</td>
<td>“Hypertension” AGT, PNMT, AGTR1, GNB3, HSD11B2, NPR3, NOS3, SAH, CYP1B2, GNAS, TNF, LEPROT, ADRA2B, NOS2A, RETN, TRHR, BDKRB2, ADIPOQ, PPARγ, KCNJ11, IL6, IFNG, CYP1A1, CTLA4, B2M, ADRA2B, VDR, RETN, PAX4, CRP, ICAM1, PTPN22, IL18, TAP, NOS2A, SUMO4, GH1, TCF2, IPF1, ADIPQ, KCNJ9, IRS2, HK2, ISL1, FABP3, GCK, SORBS1, CAPN10, NOS3, FOXA2, MAPK14, IP3R1, GIPR, FOXP3, PTPN22, IL18, FABP3, CD38, AGTR1, STX1A, ABCC8, IAPP, TCF7, STX1A, LPL, ADRA2A</td>
</tr>
<tr>
<td>Dysglycemia</td>
<td>“impaired fasting glucose,” “dysglycemia,” “diabetes”</td>
<td>Most recent 5,000</td>
<td>pancreas, liver, kidney, muscle, adipose, intestine</td>
<td>“cardiovascular vascular,” “blood peripheral blood,” “heart,” “oral cavity”</td>
<td>“diabetes” IRS1, TCF1, INSR, INS, AQP2, AVPR2, SLCO1A1, AVP, HNF4A, SLCO2A1, GCR, GPD2, GYS1, HLA-DQB3, UCP3, PPARG, KCNJ11, IL6, IFNG, CYP1A1, CTLA4, B2M, ADRA2B, VDR, RETN, PAX4, CRP, ICAM1, PTPN22, IL18, TAP, NOS2A, SUMO4, GH1, TCF2, IPF1, ADIPQ, KCNJ9, IRS2, HK2, ISL1, FABP3, GCK, SORBS1, CAPN10, NOS3, FOXA2, MAPK14, IP3R1, GIPR, FOXP3, PTPN22, IL18, FABP3, CD38, AGTR1, STX1A, ABCC8, IAPP, TCF7, STX1A, LPL, ADRA2A</td>
</tr>
<tr>
<td>Metabolic syndrome</td>
<td>“metabolic syndrome”</td>
<td>5,132</td>
<td>adipose, kidney, pancreas</td>
<td>“cardiovascular vascular,” “blood peripheral blood,” “heart,” “hepatocellular liver”</td>
<td>“metabolic syndrome” ACE, LMNA, RXRG, PTGS2, LIPE, PPARA, UCP1, WFS1, ATP2A3, RXRG, PTGS2, LIPE, genes identified by Matsunaga and Muramatsu (38)</td>
</tr>
</tbody>
</table>

Query terms used to search biomedical literature for disease-relevant abstracts and terms identified by text mining as most commonly associated with disease/phenotype (eVOC method) are also shown.

for the training set. There are no known disease genes for “metabolic syndrome” and very few genes that are currently considered strong candidate genes, so the training set was compiled from a set of genes that have been previously linked with “metabolic syndrome” in the manner described by Matsunaga and Muramatsu (38) and derived by a sophisticated knowledge-based analysis involving text-mining of PubMed abstracts.

Computational methods for disease gene prioritization. Seven existing candidate disease gene prioritization methods were used to analyze the starting set of candidate genes: Disease Gene Prediction (DGP) (http://cgg.ebi.ac.uk/services/dgp/) (34) assigns probabilities to all genes (from Ensembl version 15.33.1), indicating their likelihood to mutate based on specific sequence property patterns; PROSPECTR (http://www.genetics.med.ed.ac.uk/prospectr/) (3) differentiates between genes likely to be involved in disease and those that are not, according to sequence-based features; the eVOC method (58) selects candidate genes by determining disease relevance of their expression profile; GeneSeeker (http://www.cmbi.ru.nl/geneseeker/) (60, 61) filters positional candidate disease genes based on expression and phenotypic data from human and mouse databases; SUSPECTS (http://www.genetics.med.ed.ac.uk/suspects/) (4) scores genes according to annotation data from GO, Intpro, and expression libraries, and sequence analysis by the method PROSPECTR; G2D (http://www.ogic.ca/projects/g2d_2/) (46–48) scores all terms in GO according to annotation data from GO, Interpro, and expression library sequences; and finally Endeavour (http://homes.esat.kuleuven.be/~bioisert/endavour/endavour.php) (5), which ranks candidate genes according to Entrez Gene data, GO data, InterPro and BIND protein-protein interaction data, KEGG pathway data, microarray and expressed sequence tag (EST)-based expression data, TFBS cis-regulatory modules, and sequence similarity by BLAST. More detailed descriptions of all methods, including validation and threshold values, are given in the Supplementary Data File S3.

Collation of gene lists and construction of a scoring matrix. Gene lists were analyzed using Microsoft Excel. Visual Basic software was written to compile the data, assign scores to the candidates, generate final scores for candidates, and graphically present and rank them (as illustrated in Supplementary Data File S4). The computational methods used access a variety of data sources, and consequently not all candidate genes may be analyzed by all methods. This can create a bias in the scoring system, for example an excellent candidate may be selected by six of the seven methods and not analyzed by the seventh.

Physiol Genomics • VOL 35 • www.physiolgenomics.org

Downloaded from http://physiolgenomics.physiology.org/ by 10.220.33.3 on April 10, 2017
method, although the seventh method would have similarly selected it if it had been present in the database(s) analyzed by this method. This candidate should score more highly than a candidate that was analyzed by all seven methods and found to be a good candidate by six methods but found to be an unlikely candidate by the seventh method.

To reflect this, our scoring system distinguishes between candidates that were not analyzed by a method and candidates that were analyzed by a method and found to be poor candidates: a likely candidate ("positive"), is assigned a score of 1, an unlikely candidate ("negative") is assigned a score of 0, and if the method does not analyze the gene ("null"), the score assigned is 0.5. Thus a gene that is not considered by a method has a marginally better ranking than a gene that is actively found to be "unlikely" by that method. For this analysis, top-scoring genes selected by all seven methods were selected for each of the phenotypes and for metabolic syndrome. Phenotype frequency of occurrence is used for each population group (black, white, male, and female), and the following scoring system is used: genes selected for each phenotype are given a value of the frequency of occurrence of that phenotype within the population. The frequency values are shown in Table 2. For each gene, a final score is calculated as the sum of all scores assigned to it for that population.

**Automated analysis of final gene list.** Enrichment for specific regulatory pathways and functional annotation was investigated with the Database for Annotation, Visualization and Integrated Discovery (DAVID) [http://david.abcc.ncifcrf.gov/ (13)] and the reaction pathway database REACTOME [www.reactome.org (63)]. Additionally, GOstat was used [http://gostat.wehi.edu.au/ (9)] to measure overrepresentation of GO annotations [http://www.geneontology.org (23)] in the final candidate gene list. To identify known relationships between the candidates described in the biomedical literature, the CHILIBOT data-mining software was used [www.chilibot.net (12)], and the results obtained using CHILIBOT were manually curated for accuracy. Additionally, the genes were investigated for known relationships to metabolic syndrome with CHILIBOT. Known or inferred protein-protein interactions between the candidate genes were investigated with STRING [Search Tool for Retrieval of Interacting Proteins; http://string.embl.de (64)]. The interactions identified from STRING were manually curated for sufficiency of evidence.

**RESULTS**

The numbers of genes selected by each method for metabolic syndrome and its five defining phenotypes are shown in Table 3 (see METHODS). Also shown are the numbers of genes selected by all methods for each phenotype and those in common to both metabolic syndrome and all phenotypes. Appropriately weighted according to frequency of occurrence of each phenotype (see METHODS for details). In brief, for "metabolic syndrome," 54 genes were selected by all methods used. For each of the phenotypes, the following numbers of genes were selected in common to all methods: abdominal obesity - 48; dysglycemia - 58; hypertension - 69; hypertriglyceridemia - 80; low HDL cholesterol - 14. Genes were then scored according to their selection for the individual phenotypes in a weighted system whereby genes for more common phenotypes received higher scores. Using the phenotype frequency weightings, we selected the following numbers of top-scoring genes for each population: white population - 34 (score > 113), black population - 41 (score > 103), males - 36 (score > 128), females - 34 (score > 144). There were 34 genes common to all these analyses, and of these 19 were also common to the set selected for metabolic syndrome. These 19 were investigated further as a final candidate gene list. They are briefly described in Supplementary Data File S5, and the loci in which they fall are detailed in Supplementary Data File S8. The genes selected using population-based weighting were compared with a selection made without weighting, i.e., all genes that appeared in at least three of the phenotype-specific sets were selected regardless of phenotype frequency. There were 34 genes that appeared in three or more phenotype sets, and these were the same 34 as for the white and female population-weighted data. However, six of these genes did not appear in the list selected using black population data. Seven additional unique genes appeared in the black population data set, and two additional unique genes were selected using the male population data (Supplementary Data File S4). This illustrates the extent to which population-specific frequency data affects the final set of selected candidate genes. Reanalysis using population-specific frequency data generated by Ford et al. (2002) (17) generated a list containing the same genes except for the exclusion of TIE1, AXL, and UTRN. We have analyzed the more inclusive list.

Several existing utilities, described below, were used to analyze the selected set of candidate genes, to give an overview of the types of genes selected according to their functional annotation (DAVID and GOstat), pathways represented (REACTOME SkyPainter), existing relationships previously reported between the candidate genes (CHILIBOT), and protein-protein interactions and networks (STRING).

**Clustering of genes by functional annotation by DAVID.** The DAVID software analyzes similarity of functional annotation of candidate genes, indicating the gene functions most commonly represented in the final candidate group of 19 genes (13). When using high stringency, we found the most enriched functional cluster, with a score of 3.9, contained annotation terms relating to lipid metabolism and transportation. The terms, shown with number of genes and P values in parentheses, were "lipid transporter activity" (6; 1.3 × 10^-8), "lipid metabolism" (5; 6.0 × 10^-7), "transporter activity" (6; 2.6 × 10^-7). The second cluster, with a score of 2.91, contained the terms, with number of genes and P values shown in parentheses, "signal" (10; 5.9 × 10^-5), “signal peptide” (10; 3.8 × 10^-5).
Table 4. Gene Ontology terms most overrepresented in the final candidate gene list, as determined using GOstat

<table>
<thead>
<tr>
<th>GO ID</th>
<th>GO Term</th>
<th>Genes</th>
<th>Count</th>
<th>Total</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005319</td>
<td>lipid transporter activity</td>
<td>APOE, LDLR, APOB, HSD17B4, LPL, LRPI</td>
<td>6</td>
<td>87</td>
<td>3.23 × 10⁻⁵</td>
</tr>
<tr>
<td>GO:0032501</td>
<td>multicellular organismal process</td>
<td>APOE, INSR, LDLR, ITGB1, APOB, GIPC1, IFG1, LPL,</td>
<td>14</td>
<td>3,870</td>
<td>1.09 × 10⁻⁷</td>
</tr>
<tr>
<td>GO:0042627</td>
<td>chylomicron</td>
<td>APOE, APOB, LPL</td>
<td>3</td>
<td>8</td>
<td>1.40 × 10⁻⁶</td>
</tr>
<tr>
<td>GO:0007275</td>
<td>multicellular organismal development</td>
<td>APOE, INSR, ITGB1, IFG1, PTPN11, TIE1, UTRN, FYN, FOXO1A, LRPI</td>
<td>10</td>
<td>2,289</td>
<td>1.31 × 10⁻⁵</td>
</tr>
<tr>
<td>GO:0007154</td>
<td>cell communication</td>
<td>APOE, INSR, AXL, ITGB1, APOB, GIPC1, IFG1, PTPN11, TIE1, UTRN, FYN, FOXO1A, LRPI</td>
<td>13</td>
<td>5,554</td>
<td>6.38 × 10⁻³</td>
</tr>
<tr>
<td>GO:0048731</td>
<td>system development</td>
<td>APOE, INSR, TIE1, UTRN, FYN, FOXO1A, IFG1, PTPN11, TIE1, UTRN, FYN, FOXO1A, LRPI</td>
<td>8</td>
<td>1,623</td>
<td>1.09 × 10⁻⁴</td>
</tr>
<tr>
<td>GO:0007165</td>
<td>signal transduction</td>
<td>APOE, INSR, AXL, ITGB1, APOB, GIPC1, IFG1, PTPN11, TIE1, UTRN, FYN, FOXO1A, LRPI</td>
<td>12</td>
<td>5,125</td>
<td>1.85 × 10⁻⁴</td>
</tr>
<tr>
<td>GO:0005887</td>
<td>integral to plasma membrane</td>
<td>INSR, TIE1, AXL, ADCY7, LDLR, ITGB1, LRPI</td>
<td>7</td>
<td>1,358</td>
<td>3.00 × 10⁻⁴</td>
</tr>
<tr>
<td>GO:0031226</td>
<td>intrinsic to plasma membrane</td>
<td>INSR, TIE1, AXL, ADCY7, LDLR, ITGB1, LRPI</td>
<td>7</td>
<td>1,370</td>
<td>0.0003</td>
</tr>
<tr>
<td>GO:0048856</td>
<td>anatomical structure development</td>
<td>APOE, INSR, TIE1, UTRN, FYN, FOXO1A, IFG1, PTPN11, TIE1, UTRN, FYN, FOXO1A, LRPI</td>
<td>3</td>
<td>78</td>
<td>0.000465</td>
</tr>
<tr>
<td>GO:0044459</td>
<td>plasma membrane part</td>
<td>INSR, TIE1, AXL, UTRN, ADCY7, LDLR, ITGB1, LRPI</td>
<td>10</td>
<td>3,887</td>
<td>0.000532</td>
</tr>
<tr>
<td>GO:0008201</td>
<td>heparin binding</td>
<td>APOE, APOB, LPL</td>
<td>8</td>
<td>2,386</td>
<td>0.000812</td>
</tr>
<tr>
<td>GO:0002502</td>
<td>developmental process</td>
<td>APOE, INSR, ITGB1, IFG1, PTPN11, TIE1, UTRN, FYN, FOXO1A, LRPI</td>
<td>4</td>
<td>323</td>
<td>0.000812</td>
</tr>
</tbody>
</table>

Total number of candidate genes = 19, total number of annotated human genes used for statistical analysis = 33,781. Gene Ontology (GO) ID and corresponding term are shown. "Count" = number of candidate genes with that annotation, "Total" = total number of all genes with that annotation (GOstat, http://gostat.wehi.edu.au/ Ref. 9).
genes, thus giving an indication of the degree of relevance and/or novelty of the selected candidates. The disease name “metabolic syndrome” was associated with 9/19 candidate gene names, and on inspection 8/19 were found to be real associations (APOB, APOE, FOXO1A, IGF1, INSR, LDLR, LPL, and PTPN11; the relationship indicated between “metabolic syndrome” and CAT was spurious). The nature of these associations is described in Table 5. Potential indirect relationships with metabolic syndrome through one or more of the other short-listed candidates were indicated for all genes except ADCY7. Analysis of literature associating the candidate genes with each of the phenotypes indicates that only one of the candidates has been previously associated with all five phenotypes (APOB), five of the candidates have been associated with all phenotypes except dysglycemia (IGF1, LDLR, APOE, INSR, LPL), and three of the candidates have not been independently associated with any of the phenotypes previously (ADCY7, GIPC1, PFKM).

Protein binding and protein network analysis by STRING. STRING analysis (64) indicates two main groups of interacting proteins within the final candidate list (Fig. 2A). The first centers around PTPN11 (also called SHP2), and interactions between PTPN11 and FYN, TIE1, AXL, CAT, and INSR are indicated for all genes except dysglycemia (IGF1, LDLR, APOE, INSR, LPL). Two other short-listed candidates were indicated for all genes except dysglycemia (FOXO1A, IGF1, INSR, LDLR, APOE, and LPL), and these interactions are also described in the biomedical literature and pathway databases (see below). FOXO1A (also called FOXO1 and FKHR) lies between these two clusters, and we present below a possible connection between the two clusters facilitated by this transcription factor. Proteins PFKM and ADCY7 are not included in the interacting groups.

The potential interactions between the candidates’ protein products are represented in Fig. 2B, illustrating the predominance of genes in the final candidate set that are involved in chylomicron processing, transmembrane receptor activity and signal transduction. These are described below.

Hepatic chylomicron processing is mediated by binding of an extracellular complex containing APOB, APOE, and LPL to the transmembrane receptors LDLR and LRP1. The APOE and APOB components interact with LDLR, and the APOE and LPL components bind to LRP1 to facilitate chylomicron remnant uptake (11, 24, 35, 39).

Three distinct groups of transmembrane proteins can be identified in the candidate set. LRP1 forms a complex with GIPC1 (21). UTRN/ITGB1, and FYN form a second (50, 67). Finally, INSR binds extracellular IGF1 as well as intracellular PTPN11 (also commonly called SHP2) (7, 30) to form the third complex through their interactions with PTPN11. Additionally, the ability of FYN to interact with ITGB1/UTRN as well as INSR/PTPN11 suggests it may link these two complexes.

Three distinct signaling pathways are likely to be activated by genes in the final candidate set: The LRP1 receptor can activate G protein signaling through its interaction with GIPC1 (21); UTRN/ITGB1 in complex with FYN activates the MAP/ERK pathway (50), and IGF1 binding to INSR activates the PI3K pathway (26). LPL has also been implicated in INSR signaling, although the mechanism of this interaction is unclear (31, 53).

FOXO1A, also commonly referred to as FKHR or FOXO1, is the only nuclear transcription factor in the final candidate list and is regulated via PTPN11 and IGF1/INSR binding and activation of the PI3K pathway (51, 68, 69). FOXO1A can in turn upregulate expression of LPL (27), an integral protein for chylomicron processing (24) and a proposed regulator of insulin signaling (31, 53), thus providing a possible direct regulatory connection between transmembrane receptor activation/signal transduction and the processes of chylomicron processing and insulin receptor signaling.

**Table 5. Genes previously associated with metabolic syndrome in the biomedical literature, as detected by Chilibot**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Summary of Association With Metabolic Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>APOB</td>
<td>The serum level of ApoB showed a strong correlation with metabolic syndrome in an Asian population (43). In metabolic syndrome, increases in fatty acid delivery to the liver result in increased secretion of apoB-containing lipoproteins (14).</td>
</tr>
<tr>
<td>APOE</td>
<td>Genetically obese Ay ApoE(−/−) knockout mouse model fails to develop obesity and metabolic syndrome features (19).</td>
</tr>
<tr>
<td>FOXO1A (FOXO1, FKHR)</td>
<td>FoxO1 promotes insulin sensitivity and lipid synthesis in addition to glucose production. This dual role could explain the admixture of insulin resistance and sensitivity that is commonly observed in the metabolic syndrome (37).</td>
</tr>
<tr>
<td>IGF1</td>
<td>Individuals with metabolic syndrome had lower IGF-1 levels than subjects without metabolic syndrome with an almost linear decline of IGF-1 levels as the number of fulfilled criteria of the metabolic syndrome (none to 5) increased (16).</td>
</tr>
<tr>
<td>INSR</td>
<td>A muscle-specific insulin receptor knockout exhibits features of the metabolic syndrome of NIDDM (10, 25).</td>
</tr>
<tr>
<td>LDLR</td>
<td>Polymorphisms in LDLR were associated with metabolic syndrome in coronary heart disease patients (LDLR_1 P = 0.008; LDLR_2 P = 0.018) (39). An LDLR knockout mouse is used as a model for metabolic syndrome (52).</td>
</tr>
<tr>
<td>LPL</td>
<td>Decreased expression of LPL possibly causes insulin resistance, in addition to hypertriglyceridemia, in metabolic syndrome (53).</td>
</tr>
<tr>
<td>PTPN11 (SHP2)</td>
<td>Uptregulation of cytosolic PTPN11 expression may contribute to the hemodynamic action as well as metabolic action in hypertensive metabolic syndrome (66).</td>
</tr>
</tbody>
</table>

Common gene synonyms are shown in parentheses (Chilibot, www.chilibot.net, Ref. 12).

**DISCUSSION**

In summary, we have used computational methods to select most likely candidate genes for metabolic syndrome, starting with a set of genes selected from all loci previously associated with the disease. We have used the five phenotypes that define...
metabolic syndrome and used phenotype and population-specific filters to refine the starting set of genes. We selected the top-scoring 19 genes from the final candidate list as most likely disease gene candidates. These candidates are found to most commonly represent pathways involving metabolism of lipids and lipoproteins, as well as transmembrane signaling and subsequent activation of signal transduction pathways.

Given the diversity of the five phenotypes associated with metabolic syndrome and the variety of the phenotype combinations that present clinically, it would seem unlikely that genes belonging to specific and restricted pathways are causative of this syndrome. Instead, disease candidates are more likely to belong to generic pathways with multiple downstream effects and thus be capable of generating a diverse array of phenotypic effects if disrupted. The predominance of receptor and signaling molecules in the candidate set fits this model: signal transduction pathways can have a large variety of initiating events and multiple downstream targets, regulated through a common signal transduction pathway. Additionally, because of the cascade effect in signal transduction events, even small changes in top-level signaling events can have large and diverse downstream effects, such as the variety of phenotypes seen in metabolic syndrome. Of interest, the only transcription factor candidate, FOXO1A, appears to connect chylomicron remnant uptake with insulin signaling and possibly functions in a feedback loop: FOXO1A is regulated by insulin signaling and upregulates LPL expression; and LPL in turn activates chylomicron processing and is involved in regulation of insulin signaling. Additional trans-regulatory and secondary downstream effects of this transcription factor may also lead to the diverse phenotypes seen in metabolic syndrome. Many of the genes selected are also annotated with terms relating to development (as seen in results from GOstat analysis, Table 4).

This most likely reflects a role for lipids and insulin in regulating genes involved in development (for example, neural development as described in Ref. 21) rather than a role for developmental genes in metabolic syndrome. Additionally, however, adipose tissue is increasingly understood as an active endocrine and paracrine organ that secretes molecules involved in the inflammatory process, coagulation, insulin resistance, diabetes, and atherosclerosis (33, 41, 65). Differentiation of preadipocytes to mature fat cells is accompanied by release of many adipokines that appear to modulate both the insulin resistance of metabolic syndrome and endothelial dysfunction present in obesity, and many of these adipokines affect a variety of cellular processes that can lead to the array of abnormalities that characterize metabolic syndrome (6).

Individually, eight of the candidates have been previously implicated in metabolic syndrome, and notably, polymorphisms in LDLR have been associated with metabolic syndrome previously (39). Only one of the candidates (APOB) has been previously linked to all five phenotypes. Three of the candidates show no prior connection with these phenotypes, and one of these also has no direct or indirect association with metabolic syndrome (ADCY7) and is therefore a totally novel candidate. This selection of the top-ranking 19 most likely candidates therefore presents an array of genes including a gene containing known disease-associated single nucleotide polymorphisms (SNPs), some genes previously associated with metabolic syndrome and/or its phenotypes, and some entirely new candidates for further investigation. This demonstrates that novel predictions can be made from existing data using computational approaches and that this approach is truly synthetic and not simply sophisticated data-mining of existing knowledge.
Certain issues intrinsic to computational analysis do arise. These include differences between gene sets in different databases or versions of the same database, as described in METHODS. This is accommodated in part in this study by the scoring system that avoids biased negative scoring against genes not measured by all methods. Also to be considered, genes with better annotation are generally more likely to be selected, which results in a bias toward selection of well-investigated genes (including disease-related genes) rather than novel predicted or uncharacterized genes. Several of the selected genes have previously been well characterized as disease-causing genes, including LDLR (OMIM #606945), ApoE (OMIM #107741), ApoB (OMIM #107730), and LPL (OMIM #609708) for a variety of phenotypes. Although selection of these genes may be biased due to their extensive annotation, it is equally possible that these genes are predisposed to mutation and/or are sufficiently crucial metabolic genes that their disruption is highly likely to result in a clinical phenotype. The two latter possibilities support the selection of these genes as good candidates, and, within the model of a multigenic complex disease, novel combinations of mutations in these genes could cause the metabolic syndrome phenotypes. To further investigate novel or poorly characterized gene candidates, however, it is possible to look exclusively at scoring by DGP (34) and PROSPECTR (3), analysis by G2D (46–48), which employs sequence similarity to analyze uncharacterized genes, and expression annotation by eVOC, which is based on mapping of EST sequences from cDNA libraries regardless of characterization and annotation of the gene (29, 58).

Another issue that arises is the use for some methods of suboptimal training sets of “known” or disease-associated genes. For example, for metabolic syndrome no known disease genes have been identified, and the training set of genes has been identified through a previous text-mining study as having been considered in the context of the disease (38) rather than being known disease-causing genes. Also, the use of training sets in selection limits the search to fit existing paradigms regarding the type of gene causing the disease and limits the possibility of finding novel disease mechanisms. It would, however, be possible to investigate candidates selected without the use of training sets by excluding the methods that require a training set of known disease genes. Finally, thresholds and cut-offs for selection of most likely candidates can be arbitrary or an “informed guess” for methods that rank genes (in comparison to straightforward inclusion/exclusion). Where possible, thresholds in this study have been based on results from test data sets, as described in the seminal reference for each method (see Supplementary Data File S3).

Metabolic syndrome, in conjunction with obesity, is becoming a serious health issue globally (1, 42, 54). Linkage studies have failed to date to identify causative genes underlying the syndrome, despite clear evidence for heritability (49), and this is most likely due to complex disease genetics whereby many etiological genes of low penetrance, rather than a single causative gene, result in the syndrome. With multiple contributing phenotypes as well as many permutations of those phenotypes, the syndrome lends itself to being deconstructed for systematic computational analysis. Here, we have addressed the identification of most likely disease genes for metabolic syndrome using computational rather than empirical methods. We have used the results of published linkage studies as a primary filter to select our starting set of candidate disease genes and then used multiple computational methods to select most likely candidate genes in a combinatorial approach informed by the clinical presentation of the syndrome. With the increasing use of genome-wide association studies to identify loci associated with specific phenotypes (45), however, our approach may be additionally adapted to pinpoint most likely candidate genes associated with assayed tagging SNPs in these studies. Computational selection of appropriate candidate genes associated with the SNPs assayed is likely to be highly successful because of the much smaller starting sets of candidate genes under analysis.

Whether the selected genes are truly associated with metabolic syndrome can only be shown empirically [as, for example, disease-associated polymorphisms have been previously identified in the selected candidate gene LDLR (39)], and it is not possible to validate our approach to selection of metabolic syndrome candidate disease genes without further empirical research. However, our study is instead a novel combination of independently validated methods combined with disease phenotype data, and we propose a novel set of candidate causative genes for metabolic syndrome as a starting point for further empirical research. All genes selected by any one of the methods already form a subset of likely candidate disease genes, as determined by these independently verified methods. We then take this subset and filter it further by selecting genes that match the most of the criteria tested across all methods and apply additional filters based on the array of phenotypes that define metabolic syndrome and the frequencies of those phenotypes in several patient populations.

Given the increasing urgency of understanding and managing metabolic syndrome effectively, we show here that computational analysis can offer new insights into genetic determinants and regulatory pathways that may be involved and presents a selection of genes that includes both novel and identified candidates for further empirical analysis. Exploring new avenues can also lead to a better understanding of environmental and genetic interactions underlying the syndrome.

We believe that accurately understanding disease phenotype and clinical presentation can lead to success in identifying genetic determinants for complex disease. A complete understanding of the complex clinical presentation of a disease can be used to ask intelligent questions of the existing genetic data and can subsequently be employed to make sense of the pathways and processes represented by the candidates selected. Because no disease-causing genes have been identified for metabolic syndrome to date, it is timely to utilize the phenotypic data in novel and informative ways to determine the genetics of the disease.

ACKNOWLEDGMENTS

M. A. Andrade-Navarro is a Canada Research Chair in Bioinformatics. We thank A. J. Swift for Visual Basic programming.

GRANTS

I. Okpechi is funded by Wyeth South Africa (pty) Limited, funds are administered by the University of Cape Town. M. A. Andrade-Navarro and C. Perez-Iratxeta acknowledge funding from the Canadian Institutes of Health Research, Canadian Foundation for Innovation, and Ontario Innovation Trust. R. Ramesar is funded by the Medical Research Council of South Africa.
REFERENCES


