Diabetes Associated Genes from the Dark Matter of the Human Proteome

Abstract
The human genome offers an attractive starting point for diabetes biomarker discovery. We have undertaken a survey of the Genetic Association Database (GAD) to develop a comprehensive genetic profiling of the type 1 and type 2 diabetes phenotypes. Using text mining, the GAD was explored for diabetes-associated genetic polymorphisms and a working database for type 1 and type 2 diabetes was established. In addition to well-characterized genes, 57 novel, uncharacterized Open Reading Frames (ORFs) encompassed in the dark matter of the human proteome were identified. Diverse bioinformatics and proteomics tools were used to characterize these ORFs for gene expression, protein motifs and domain information. Distinct protein classes including secreted products, enzymes, transporters, and receptors were encoded by these ORFs. Using expression Quantitative Traits Loci, Clinical Variations and the Genome-Phenome Integrator tools, 50 novel ORFs associated with phenotypes for both type 1 and type 2 diabetes were identified. These results open new avenues for better understanding type 1 and type 2 diabetes and may provide novel therapy targets for type 2 diabetes and associated disorders.

Keywords
Autoimmune disease; Bioinformatics; Diabetes; Proteomics; Genetic association; Phenotype; Druggable genes; Biomarkers; Dark matter proteome; Open reading frames

Introduction
The human genome project is an attractive starting point for novel gene discovery for diverse diseases [1-4]. In the past, gene discovery approaches focused on one gene at a time, which was time consuming and inefficient. The ready availability of numerous meta-analysis bioinformatics tools has greatly enhanced our ability to mine the genome globally to identify genes involved in multiple diseases. A significant number of the human proteins in the genome, however, remain uncharacterized [5]. These uncharacterized proteins together with the noncoding RNAs (ncRNAs) have been termed the Dark Matter of the human genome [6-8].

Development of new molecular entities for therapy and diagnosis for various diseases requires novel targets. Reasoning that such novel targets may emerge from characterizing the dark matter proteome, we have embarked on a systematic dissection of the uncharacterized proteins, the Open Reading Frames (ORFs) in the genome [9-11]. Our recent development of a cancer-associated fingerprint from the dark matter proteome, the OncoORFs [11], provided a framework for expanding our approaches to other diseases.

We have next undertaken mining the human genome with a view towards novel biomarker discovery for type 1 and type 2 diabetes. It is estimated that 382 million people suffer from diabetes, for a global prevalence of 8.3% [12]. Diabetes affects a large number of people in the world and is a major healthcare challenge [13,14]. The complications associated with diabetes involve numerous other disorders such as cardiovascular, developmental, immune, metabolic, neurodegenerative, obesity, renal and vision [15-19]. Both of the two major forms of diabetes, type 1, an autoimmune disease [20,21] and type 2, a metabolic disorder [22-25], require novel approaches to early diagnosis and therapy [26-29]. Notwithstanding the availability of several classes of anti-diabetic drugs, it is often difficult to maintain long-term glycemic control and many current agents have treatment-limiting side effects. Discovery of targets affecting multiple pathways in the diabetes-associated disorders would greatly facilitate the development of novel therapeutics for type 2 diabetes [26,30].

The Genetic Association Database (GAD) provides an efficient way to mine the human genome for disease association studies [31]. Association data regarding both the known and the uncharacterized proteins are available in the GAD, which can be readily mined to establish genetic polymorphism-associated disease phenotypes.

Reasoning that the GAD would enable us to discover type 1 and type 2 diabetes-associated novel biomarkers and druggable...
targets [32,33], which might also be relevant to diabetes-related disorders, the GAD was searched for diabetes-related entries by text mining. The genetic polymorphisms associated with both type 1 and type 2 diabetes and related disorders were classified. In addition to known genes, numerous ORFs were found to be associated with both type 1 and type 2 diabetes. These diabetes-associated ORFs (referred to as diabetes ORFs) were also found to be genetically associated to numerous other diseases. Using diverse bioinformatics and proteomics tools we demonstrate that novel druggable classes of proteins (enzymes, receptors, transporters) were encoded by the diabetes ORFs. Further, secreted ORF biomarkers unique to type 1 and type 2 diabetes were detected in the body fluids including blood, urine and pancreatic juice. These results provide a framework for discovery of novel biomarkers for diabetes type 1 and type 2 and to further develop a better understanding of the diabetes-associated disorders.

Materials and Methods

The bioinformatics and proteomics tools used in the study have been described elsewhere [9-11]. In addition, the following genome-wide association tools were used: the Genetic Association Database, GAD [31]; the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 from the NCBI [34]; GeneALaCart (LifeMap discovery) from the Genecards [35]; the Phenotype-GenoType Integrator [36]; the Database of Genomic Variants, DGV [37]; Clinical Variations, ClinVar [38]; the International HapMap project, the type 1 diabetes database and the type 2 genetic association database, T2D-db [39].

All of the bioinformatics mining was verified by two independent experiments. Big data was downloaded two independent times and the output verified for consistency. Big data verification was performed by two independent investigators. Only statistically significant results per each tool’s requirement are reported. Prior to using a bioinformatics tool, a series of control query sequences was tested to evaluate the predicted outcome of the results.

Results

Disease association profile of the diabetes genes

We have undertaken a comprehensive profiling of the diabetes-associated genetic polymorphisms of the human genome. The GAD is a comprehensive archive of human genetic association studies of complex diseases and disorders [31,40]. The association data for both the known and uncharacterized proteins are present in this database. The identification of clinically relevant polymorphisms from the large volume of polymorphism and mutational data is possible with the GAD. The entire GAD database as of 2014-03-08 was downloaded to provide a basis for mining the diabetes genome.

From the 65,536 entries in the complete GAD, 4,665 diabetes-related entries were found. These entries were filtered using the advanced filter option of Excel and gene entries related to diabetes associated diseases and disorders were enriched (Figure 1). Genetic polymorphisms related to cardiovascular, metabolic, immune, infection, cancer, neurodegenerative disorders etc. associated with diabetes were found. The entire data is shown in Supplemental Table 1.

Multiple entries for each gene were found which represented different polymorphic, Single Nucleotide Polymorphism (SNP) rs numbers. The GAD entries were enriched for diabetes using three filters 1) broad phenotypes, 2) disease classes and 3) Medical Subject Headings (MeSH) terms, yielding 57 diabetes-associated ORFs. In view of the strong association of these ORFs with diabetes-related disorders, these ORFs are termed the “Diabetes ORFs”. These novel ORFs provided the framework for detailed characterization studies to establish druggability and biomarker potential for diabetes.

Uncharacterized proteome of the diabetes genome

Recently we demonstrated the usefulness of a streamlined approach to mining the GAD database to identify a cancer-related fingerprint, the OncoORFs [11]. The availability of multiple batch analysis tools such as the GeneALaCart from the Genecards [35], DAVID [34], the canSAR Integrated Drug Discovery platform [41], and numerous protein expression analysis tools such as the Model Organism Protein Expression Database (MOPED) [42], the Human Protein Reference Database (HPRD) [43], the Human Protein Atlas (HPA) [44] and the recently described Clinical Proteomic Tumor Analysis Consortium (CPTAC) database greatly facilitated big data handling approaches.

In order to establish an initial framework for characterization studies, the 57 diabetes ORFs were batch analyzed using the GeneAlaCart, DAVID and canSAR integrated bioinformatics tools. Information related to gene descriptions, IDs (mRNA and proteins), chromosomal map positions, putative function and gene ontology were obtained. These analyses enabled an initial protein class prediction for some of these diabetes ORFs (Table 1 included as supplementary). Noncoding RNAs, putative

Figure 1: Disease association profile of the diabetes genes. The Genetic Association Database (GAD) was enriched by text mining-based filtering to identify diabetes-associated polymorphisms in related disorders. The numbers indicate the polymorphisms associated with the indicated disorders.
enzymes, secreted proteins, cell cycle and trafficking proteins were inferred from the gene descriptions, the gene ontology (GO) and the UniProt summary. Thirty of the diabetes ORFs were uncharacterized proteins. Encouraged by the possible druggability and biomarker potential of these uncharacterized diabetes ORFs, a comprehensive analysis and characterization was undertaken.

Association of the diabetes ORFs with diverse diseases

Diabetes type 1 and type 2 represents a complex set of associated diseases and disorders [45-47]. Hence, it was of interest to investigate the relationship of the diabetes ORFs with other diseases. The diabetes ORFs from the GAD were analyzed using the MeSH and the broad terms filters. To augment the disease data output from the GAD, a disease-oriented database, the Malacards [48] and the NextBio Meta analysis tool were used to establish a comprehensive disease profiling of these ORFs. Data was also generated from the NextBio for most correlated characteristics (tissues, drug interactions and genes perturbed) of the diabetes ORFs (Supplemental Table 2). As shown in Figure 2, the 57 diabetes ORFs were associated with various disorders and diseases, which often accompany both type 1 and type 2 diabetes. Many overlapping diseases were seen for these ORFs, implying a complex landscape of involvement.

Gene expression profile of the diabetes ORFs

The mRNA and protein expression data provide an important clue to the specificity of the ORFs. Hence, the diabetes ORFs’ expression in human normal and tumor tissues was investigated using the MOPED, HPA and HPRD and the National Cancer Institute (NCI) CPTAC protein expression tools. The diabetes ORF data was enriched from the complete HPRD and HPA downloaded databases; the MOPED and the NCI clinical databases were batch analyzed using the diabetes ORFs. The tissue-restricted mRNA expression was inferred from UniGene and HPA tools (Table 2 included as supplementary). Distinct expression profiles for numerous diabetes ORFs were detected in diverse tissues and body fluids: blood (C1orf167, C1orf204, C6orf25, C12orf63, C14orf64, C15orf62, C20orf27), liver secretion (C1orf167, C11orf9), pancreatic juice (C20orf27, C11orf9), serum (C4orf41), sperm (C6orf10) and urine (C6orf1). Tissue-restricted expression was seen for brain (C4orf50), lung (C1orf87, C9orf171), small intestine (C10orf112, C17orf78), testis (C1orf87, C6orf10, C8orf85, C9orf171, C1orf167, C17orf50) and fetus (C18orf56). Pancreatic expression was seen for C3orf65, C4orf32, C4orf52, C6orf1, C6orf10, C6orf47, C6orf173, C10orf2 and C16orf70. The C20orf27 protein expression was seen in both blood platelets and pancreatic juice. The C11orf9 protein was detected in both the pancreatic juice and in liver secretion, while expression of the C6orf10 protein was seen in the sperm and pancreatic tissues. The detection of several of the diabetes ORFs in diverse body fluids highlights the biomarker potential for these ORFs to enhance the pipeline of diagnostic markers for diabetes type 1 and type 2.

Motif and domain analysis of the diabetes ORFs

To develop further insight into the nature of the diabetes-related proteins, the ORFs were analyzed for protein motifs and domains. The GeneAlacart and DAVID tools were used to batch analyze the diabetes ORFs for the InterPro/UniProt Domains and Families, Panther and Procyte protein motifs. In addition, the NCBI Conserved Domain Database, CDD [49], the InterProscan [50], the Protein Family, PFAM [51,52] and SignalP [53] bioinformatics tools were used to analyze the diabetes ORFs (Table 3 included as supplementary). The post-translational modification sites, binary interactions and protein architecture and complexes were obtained from the HPRD database batch analysis. From these analyses, the diabetes ORFs were grouped into classes of proteins. Protein families including immunoglobulins, secreted products, antigens, cell cycle proteins, enzymes, nucleotide/metal binding, receptors, transporter/sorting proteins, vesicular proteins and noncoding RNA (ncRNAs) were identified among the diabetes ORF encoded proteins. The binary interaction data, post-translational modification as well as the protein architecture from the HPRD provides additional information regarding the nature of the diabetes ORF proteins.

From these results, five ORF proteins were predicted as secreted proteins based on the presence of signal peptide sequence at the N-Terminus using the p signal tool (C1orf204, C6orf25, C6orf27, C6orf57 and C14orf64). Two of these proteins, C6orf25 and C6orf27, were specific to type 1 diabetes. On the other hand, the C6orf57 protein was specific to type 2 diabetes. The expression of C6orf25 and C1orf204 proteins was also detected in the blood (Table 2 included as supplementary). These five ORFs provide a rationale for development of novel diagnostic markers.

Diabetes-associated traits of the diabetes ORFs by eQTL analysis

The Phenotype-GenoType Integrator (PheGenI) merges National Human Genome Research Institute (NHGRI) Genome-Wide Association Studies (GWAS) data with several databases including Gene, database of Genotypes and Phenotypes (dbGaP), Online Mendelian Inheritance in Man (OMIM), the Genotype-
Tissue Expression project (GTEx) and the Single Nucleotide Polymorphism database (dbSNP) [36].

An expression Quantitative Trait Locus (eQTL) represents a marker (locus) in the genome in which variation between individuals is associated with a quantitative gene expression trait, measured as mRNA abundance. Three parameters are used to verify eQTL results: 1) a SNP marker, 2) the gene expression levels, as measured by the probe or sequence information, and 3) a measure of the statistical association between the two in a study population, such as the P-value. The eQTL browser provides an approach to query the eQTL database [54].

The eQTLs can be cis, where the genotyped marker is near the expressed gene, or trans, in which the genotyped marker is distant from the expressed gene either in the same or on another chromosome. Currently only the cis-eQTLs [55] are available. In order to establish diabetes-associated eQTL results for the diabetes ORFs, the PheGene tool was used to batch analyze the ORFs. Genotypes for the diabetes ORFs were selected for exons, introns, near gene and Untranslated Region (UTR). From the output of results, diabetes traits were enriched. The eQTL data for the diabetes ORFs are shown in Table 4 (included as supplementary).

Four diabetes ORFs showed strong eQTL association evidence with diabetes with significant P-values. The ORF C6orf10 was associated with type 1 diabetes, systemic lupus [56] and multiple sclerosis [57]. Other diseases showing association with C6orf10 included rheumatoid arthritis, drug-induced liver injury, Graves disease, asthma, psoriasis, glomerulonephritis, IGA, systemic scleroderma, bone density, diabetic nephropathy, heart rate, vitiligo and eosinophilis (Supplemental Table 2). The ORFs C6orf27 and C6orf47 were associated exclusively with type 1 diabetes [58-60], whereas ORF C6orf57 was associated with type 2 diabetes and CD40 ligand [61,62].

**Diabetes ORFs and subtypes of diabetes**

A summary of key findings related to the two major subtypes of diabetes is shown in Table 5. Thirty of the 57 ORFs remain uncharacterized proteins. Both type 1 and type 2 diabetes were found to be associated with distinct as well as common diabetes ORFs. Fifteen ORFs were associated with type 1 diabetes and 35 ORFs were linked to type 2 diabetes. Three eQTL ORFs associated with type 1 (C6orf10, C6orf25 and C6orf47) were identified. A single ORF, C6orf57, was found to be associated only with type 2 diabetes.

The ncRNA class of the diabetes ORFs was associated with both type 1(C6orf208) and type 2 diabetes (C6orf217, C14orf70). Diabetes type-specific secreted ORF proteins were also identified for type 1 (C6orf225 and C6orf227) and for type 2 (C6orf57). Strong genetic association for phenotypes was seen for both type 1 (C6orf27, C6orf47, C6orf173, C6orf208) and type 2 (C1orf204, C6orf47, C6orf57, C6orf17, C1orf70). Clinical variations were identified using the ClinVar tool. The C2orf86| WD repeat containing planar cell polarity effector is a risk factor for Meckel Syndrome Type 6 and Bardet-Biedl Syndrome 12 and is pathogenic for Bardet-Biedl Syndrome 15 [63]. The C4orf32| CTBP1 antisense RNA 2 (head to head) is pathogenic for developmental delay [64]. The C10orf55|Uncharacterized protein is a risk factor for susceptibility to late-onset Alzheimer’s Disease [65,66], whereas the C1orf41|Uncharacterized protein harbors pathogenic mutations for Congenital Dyserythropoietic Anemia, Type 1b [67,68].

**Novel druggable targets for type 2 diabetes and biomarkers for type 1 Diabetes**

Using diverse bioinformatics and proteomics tools (gene ontology, motif and domain analysis, protein expression data in normal tissues and body fluids), putative protein classes were assigned for 42/57 diabetes ORFs (Figure 3 and Table 3). These included druggable targets such as enzymes (C7orf10, C10orf2), receptor/cell adhesion molecules (C10orf12), transporters (C1orf87, C16orf70), secreted immunoglobulins (C1orf204, C6orf25) and other secreted proteins (C4orf52, C6orf27, C6orf57 and C1orf64). These novel ORF proteins present a valuable opportunity to open new avenues for diabetes drug discovery and diagnostic marker development.

**Discussion**

We have used GAD to stratify diabetes-associated genes and genetic polymorphisms. Diverse diabetes-associated complications and disorders (albinomia, alzheimer’s disease, autoimmune, cardiovascular, glucose intolerance, infection, inflammation, insulin resistance, metabolic syndrome, neurodegenerative, neoplasm, obesity, pharmacogenomics, predisposition, subtypes of diabetes) were segregated into a distinct set of gene-associated polymorphisms. In addition to known genes, over 50 uncharacterized ORF proteins were associated with type 1 and type 2 diabetes. These ORFs also showed association with diverse diseases and complications that often accompany both type 1 and type 2 diabetes, suggesting a complex landscape of disease involvement for these proteins. Currently, it is not possible to separate the associated disorders specific to type 1 versus type 2 diabetes. However, the type-specific ORFs predicted in this study should provide a starting point for such an analysis in the future.

Identification of five new and uncharacterized ORF proteins with signal peptides and their detection in body fluids adds to the pipeline of potential biomarkers for both type 1 and type 2 diabetes. Further, 13 new druggable genes encompassing receptors, transporters and enzymes motifs were identified. It is likely that some of these novel ORFs may provide a basis for drug discovery efforts for diabetes type 2 and associated diseases.

Novel links to type 1 and type 2 diabetes with cancer is predicted from this study. The association results for C12orf30 (N-terminal acetyltransferase B complex subunit NA2A25) indicate that individuals with increased susceptibility to type 1 or 2 diabetes have a decreased risk of developing prostate cancer [69]. A long intergenic non-protein coding RNA (C6orf208) showed a strong association with renal cell carcinoma and type 1 diabetes [70]. Although the precise relationship between...
diabetes and prostate and renal cancer is unclear; these results underscore the importance of linking unrelated human diseases. Additional studies on the C6orf208 ncRNA may provide valuable clues for understanding the link between both types of diabetes and cancer.

The C2orf65 (meiosis 1-arresting protein, M1AP), C6orf10 (testis-specific basic protein, TSBP) and C8orf85 (alanine-and arginine-rich domain-containing protein, AARD) showed association with rheumatoid arthritis, coronary artery disease, Crohn’s disease, diabetes mellitus type 2, type 1 insulin-dependent diabetes mellitus, and hypertension [71]. These three ORFs may provide further insight into the association of type 1 and type 2 diabetes with cardiovascular and inflammatory diseases.

Interestingly, numerous type 2 diabetes ORFs were found to be associated with a pharmacogenomics potential in thiazolidinedione-induced edema [72]. It is tempting to speculate that these ORFs may form a core pharmacogenomic signature for the treatment of type 2 diabetes. Additional experiments are needed to verify these findings.

The ncRNAs are increasingly becoming an important component of the dark matter of the human genome [6,7]. Our study demonstrates that distinct ncRNAs were associated with type 1 diabetes (C6orf208) versus type 2 (C6orf217, C14orf70). In addition to its association with type 1 diabetes, C6orf208 was also linked with allergic disorders, disorders of the lung and viral infections. On the other hand, C14orf70 was associated with viral infections, liver transplant disorder and neuroblastoma. C6orf217 was associated with virus infections, breast cancer, head and neck cancers and bipolar disorder (Supplementary Table 2). A common association with viral infections was seen in these three distinct ncRNAs. It is possible they are involved in a common pathway. Further studies on these three diabetes type-specific ncRNAs are warranted.

Fifteen of the diabetes ORFs were uniquely associated with type 1 diabetes and 35 were uniquely associated with type 2 diabetes (Table 5 Included as supplementary). These unique ORFs included secreted factors for Type 1 (C6orf25|Secreted immunoglobulin, and C6orf27| von Willebrand factor A domain-containing 7 protein) and for type 2 (C6orf57|Protein of unknown function). The type 1-specific secreted ORFs, if verified as an early stage marker, offer early intervention potential.

In addition, the diabetes ORFs encompassed a class of druggable proteins: 1) receptor C5orf23|Atrial natriuretic peptide clearance receptor [73], 2) enzymes (C7orf10|baIF CoA-transferase family protein and C12orf30|N(alpha)-acetyltransferase 25 [58], NatB auxiliary subunit) and 3) transporter (C16orf70|lin-10 homolog|post-Golgi vesicle-mediated transport). These three targets may provide novel opportunities for drug discovery approaches for type 2 diabetes.

The C5orf23 is associated with hypertension, obesity, asthma, thyroiditis, and lung and neuroendocrine tumors [74-76]. The C12orf30 is associated with type 1 diabetes, hypothyroidism, arthritis, systemic lupus erythematosus, other autoimmune disorders and prostate cancer [77-79]. The druggability of the C12orf30 (enzyme) offers an attractive target for type 1 diabetes-associated disorders and complications. The C16orf70 is associated with type 2 diabetes and chronic lymphocytic leukemia and may offer a response therapy target for the treatment of edema among individuals who receive rosiglitazone [72].

The eQTL evidence identifies the C6orf10|Testis-specific basic protein|TSBP as a key gene involved in vitiligo, type 1 diabetes, multiple sclerosis, systemic lupus erythematosus, rheumatoid arthritis and various other immune disorders [80-83]. Generalized vitiligo is an autoimmune disease characterized by patchy depigmentation of skin, hair, and mucous membranes resulting from loss of melanocytes from involved areas [84]. Strong association evidence was seen for C6orf10 in vitiligo, and this gene may provide a biomarker potential.

These results support our starting premise that mining the uncharacterized diabetes proteome using bioinformatics and proteomics approaches can identify novel molecular targets for better understanding the etiology. It is reasonable to predict that from the druggable class of the ORFs identified in this study, new drug targets may emerge for the treatment of type 2 diabetes and related diseases. The discovery of novel diabetes type 1 and type 2 specific ORFs, expression validation and protein motif characterization should facilitate functional studies for these genes in the future. Further studies on these diabetes ORFs with functional genomics should shed light on their relevance to the complex etiology of both type 1 and type 2 diabetes.

**Conclusion**

In summary, these results demonstrate the usefulness of mining the human genome for novel biomarker discovery for both type 1 and type 2 diabetes. Identification of novel secreted proteins in the body fluids and druggable genes encompassing...
enzymes, receptors and transporters provides a rationale for new biomarkers and therapeutic targets discovery for type 2 diabetes and related disorders. Understanding the gene network and pathway interactions with other genes with these novel diabetes-associated ORFs is likely to provide new knowledge about function of these ORFs. The diabetes type-specific ORFs discovered in this study should provide a basis for follow-up studies toward a better understanding of the complex etiology of both type 1 and type 2 diabetes and related diseases that often accompany diabetes.

Acknowledgement

RN was responsible for the overall execution of the project and data generation. APD was responsible for the data mining, visualization and verification. PB performed the data mining of the disease-oriented databases. This work was supported in part by the Genomics of Cancer Fund, Florida Atlantic University Foundation. We thank Dr. Stein of the GeneCards team for generous permission to use the powerful GeneALaCart tool; Dr. Montague, Kolker Laboratory of the MOPED Team for batch analysis of the ORFs; the NextBio Meta analysis for generous permission to use the tool, the CanSar, the Human Protein Atlas, Human Protein Reference Database and DAVID functional annotations tools for various datasets. We thank Jeanine Narayanan for editorial assistance.

References

Diabetes Associated Genes from the Dark Matter of the Human Proteome


Citation: Delgado AP, Brando R, Narayanan R (2014) Diabetes Associated Genes from the Dark Matter of the Human Proteome. MOJ Proteomics Bioinform 1(4): 00020. DOI: 10.15406/mojpb.2014.01.00020