Research Article
Open Access

# 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 2diabetes 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 up 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

Volume 1 Issue 4 - 2014

Ana Paula Delgado, Pamela Brandao, Ramaswamy Narayanan
Department of Biological Sciences, Charles E. Schmidt College of Science, Florida Atlantic University, USA

**Correspondence:** Ramaswamy Narayanan, Department of Biological Sciences, Charles E. Schmidt College of Science, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA, Tel +15612972247, Fax +15612973859, Email rnarayan@fau.edu

**Received:** June 19, 2014 | **Published:** July 19, 2014

**Abbreviations:** eQTL, expression quantitative trait loci; GAD, genetic association database; GtEx, genotype-tissue expression; HapMap, haploid map; HGNC, human genome nomenclature committee; HPRD, human protein reference database; HPA, human protein atlas; MOPED, model organism protein expression database; PHEGENI, phenotype-genotype integrator; ORF, open Reading Frame; RefSeq, reference sequence; SNP, single nucleotide polymorphism

## 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 2diabetes. It is estimated that 382million 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 drug gable 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 drug gable classes of proteins (enzymes, receptors, transporters) were encoded by the diabetes ORFs. Further, secreted ORF biomarkers unique to type 1 and type 2diabetes 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

i. broad phenotypes,

ii. disease classes and

iii. *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 druggableness and biomarker potential for diabetes.

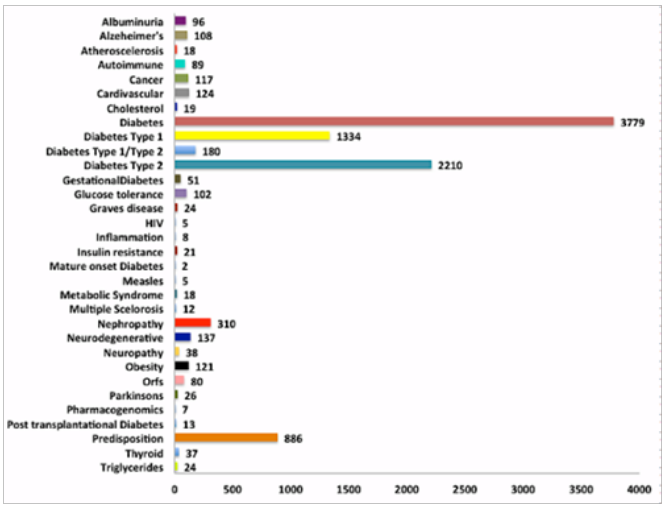


**Figure 1** Disease association profile of the diabetes genes. The Genetic Association Database (GAD) was enriched by text mining-based filtering to identify diabetes-associated polymorphisms in related disorders. The numbers indicate the polymorphisms associated with the indicated disorders.

### 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 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 druggableness 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.

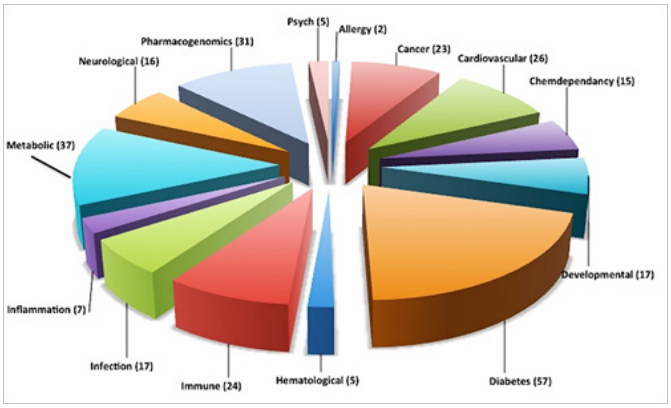


**Figure 2** Association of the diabetes ORFs with diverse diseases. The association of the diabetes-related uncharacterized Open Reading Frames (ORFs) with diverse diseases and disorders is shown. The numbers in parentheses indicate the ORFs associated with the indicated disorders.

## 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 proteomics 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 data 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:

i. a SNP marker

ii. The gene expression levels, as measured by a probe or sequence information, and

iii. 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 PheGenI 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 type1 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 eosinophils (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 (C6orf25 and C6orf27) 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, C6orf217, C14orf70). 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-Biedel 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 C15orf41|Uncharacterized protein harbors pathogenic mutations for Congenital Dyserythropoietic Anemia, Type 1b.[67,68]

### Novel druggable targets for type 2 diabetes and biomarkers for type I 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) (Table 3). These included druggable targets such as enzymes (C7orf10, C10orf2), receptor/cell adhesion molecules (C10orf112), transporters (C1orf87, C16orf70), secreted immunoglobulins (C1orf204, C6orf25) and other secreted proteins (C4orf52, C6orf27, C6orf57 and C14orf64). These novel ORF proteins present a valuable opportunity to open new avenues for diabetes drug discovery and diagnostic marker development.

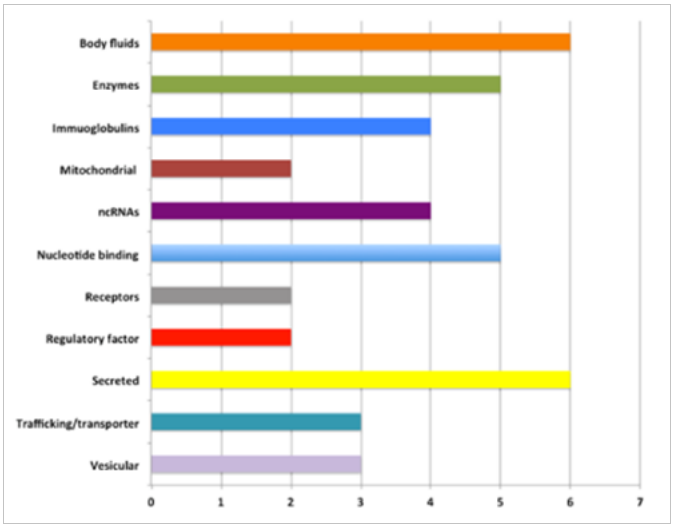


**Figure 3** Novel druggable targets and biomarkers for diabetes. A summary of the gene ontology and protein motif- and domain-based prediction of the protein classes for the diabetes ORFs is shown.

## Discussion

We have used GAD to stratify diabetes-associated genes and genetic polymorphisms. Diverse diabetes-associated complications and disorders (albuminuria, 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 NAA25) 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

**Citation:** Delgado AP, Brandao P, Narayanan R. Diabetes associated genes from the dark matter of the human proteome. *MOJ Proteomics Bioinform.* 2014;1(4):86–92. DOI: 10.15406/mojpb.2014.01.00020

cancer, head and neck cancers and bipolar disorder (Supplemental 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:

i. receptor C5orf23|Atrial natriuretic peptide clearance receptor,[73]

ii. enzymes (C7orf10|baiF CoA-transferase family protein and C12orf30|N(alpha)-acetyltransferase 25,[58] NatB auxiliary subunit) and

iii. 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 druggableness 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 2diabetes.

## 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.

## Acknowledgements

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 Next Bio 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.

## Conflict of interest

Author declares that there is no conflict of interest.

## References

1. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
2. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science*. 2001;291(5507):1304–1351.
3. Wheeler DA, Wang L. From human genome to cancer genome: the first decade. *Genome Res*. 2013;23(7):1054–1062.
4. Narayanan R. Bioinformatics approaches to cancer gene discovery. *Methods Mol Biol*. 2007;360:13–31.
5. Pertea M, Salzberg SL. Between a chicken and a grape: estimating the number of human genes. *Genome Biol*. 2010;11(5):206.
6. Blaxter M. Genetics: revealing the dark matter of the genome. *Science*. 2010;330(6012):1758–1759.
7. Martin L, Chang HY. Uncovering the role of genomic "dark matter" in human disease. *J Clin Invest*. 2012;122(5):1589–1595.
8. Brylinski M. Exploring the "dark matter" of a mammalian proteome by protein structure and function modeling. *Proteome Sci*. 2013;11(1):47.
9. Delgado AP, Brandao P, Hamid S, et al. Mining the dark matter of the cancer proteome for novel biomarkers. *Current Cancer Therapy Reviews*. 2013;9(4):265–277.
10. Delgado AP, Hamid S, Brandao P, et al. A novel transmembrane glycoprotein cancer biomarker present in the X chromosome. *Cancer Genomics Proteomics*. 2014;11(2):81–92.
11. Delgado AP, Hamid S, Brandao P, et al. Open reading frames associated with cancer in the dark matter of the Human genome. *Cancer Genomics Proteomics*. 2014;11(2):81–92.
12. Danaei G, Finucane MM, Lu Y, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980:systematic analysis of health examination surveys and epidemiological studies with 370 country–years and 2.7million participants. *Lancet*. 2011;378(9785):31–40.

13. Abegunde DO, Mathers CD, Adam T, et al. The burden and costs of chronic diseases in low–income and middle–income countries. *Lancet*. 2007;370(9603):1929–1938.

14. Gillum LA, Gouveia C, Dorsey ER, et al. NIH disease funding levels and burden of disease. *PLoS One*. 2011;6(2):e16837.

15. Rasool M, Malik A, Qazi AM, et al. Current view from Alzheimer disease to type 2 diabetes mellitus. *CNS Neurol Disord Drug Targets*. 2014;13(3):533–542.

16. Nikfarjam M, Low N, Weinberg L, et al. Total pancreatectomy for the treatment of pancreatic neoplasms. *ANZ J Surg*. 2014;84(11):823–826.

17. Guerrero–Berroa E, Ravona–Springer R, Heymann A, et al. Decreased motor function is associated with poorer cognitive function in elderly with type 2 diabetes. *Dement Geriatr Cogn Dis Extra*. 2014;4(1):103–112.

18. Aslam A, Singh J, Rajbhandari S. Pathogenesis of painful diabetic neuropathy. *Pain Res Treat*. 2014;2014:412041.

19. Lin YC, Thuy TD, Wang SY, et al. Type 1 diabetes, cardiovascular complications and sesame (Zhi Ma). *J Tradit Complement Med*. 2014;4(1):36–41.

20. Burn P. Type 1 diabetes. *Nat Rev Drug Discov*. 2010;9(3):187–188.

21. van Belle TL, Coppieters KT, von Herrath MG. Type 1 diabetes: etiology, immunology, and therapeutic strategies. *Physiol Rev*. 2011;91(1):79–118.

22. Maruthur NM, Gribble MO, Bennett WL, et al. The pharmacogenetics of type 2 diabetes: a systematic review. *Diabetes Care*. 2014;37(3):876–886.

23. Lin Y, Sun Z. Current views on type 2 diabetes. *J Endocrinol*. 2010;204(1):1–11.

24. Brunetti A, Chiefari E, Foti D. Recent advances in the molecular genetics of type 2 diabetes mellitus. *World J Diabetes*. 2014;5(2):128–140.

25. Kato N. Insights into the genetic basis of type 2 diabetes. *J Diabetes Investig*. 2013;4(3):233–244.

26. Verspohl EJ. Novel pharmacological approaches to the treatment of type 2 diabetes. *Pharmacol Rev*. 2012;64(2):188–237.

27. Aicher TD, Boyd SA, McVean M, et al. Novel therapeutics and targets for the treatment of diabetes. *Expert Rev Clin Pharmacol*. 2010;3(2):209–229.

28. Mullard A. Drug makers and NIH team up to find and validate targets. *Nat Rev Drug Discov*. 2014;13(4):241–243.

29. Flannick J, Thorleifsson G, Beer NL, et al. Loss–of–function mutations in SLC30A8 protect against type 2 diabetes. *Nat Gen*. 2014;46(4):357–363.

30. Moller DE. New drug targets for type 2 diabetes and the metabolic syndrome. *Nature*. 2001;414(6865):821–827.

31. Zhang Y, De S, Garner JR, et al. Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. *BMC Med Genomics*. 2010;3:1.

32. Sioud M, Leirdal M. Druggable signaling proteins. *Methods Mol Biol*. 2007;361:1–24.

33. Russ AP, Lampel S. The druggable genome: an update. *Drug Discov Today*. 2005;10(23–24):1607–1610.

34. Huang da DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2008;4(1):44–57.

35. Safran M, Dalah I, Alexander J, et al. GeneCards Version 3:the human gene integrator. *Database (Oxford)*. 2010;2010:baq020.

36. Ramos EM, Hoffman D, Junkins HA, et al. Phenotype–GenoType 1ntegrator (PheGenI): synthesizing genome–wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet*. 2014;22(1):144–147.

37. MacDonald JR, Ziman R, Yuen RK, et al. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(Database issue):D986–D992.

38. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(Database issue):D980–D985.

39. Lim JE, Hong KW, Jin HS, et al. Type 2 diabetes genetic association database manually curated for the study design and odds ratio. *BMC Med Inform Decis Mak*. 2010;10:76.

40. Becker KG, Barnes KC, Bright TJ, et al. The genetic association database. *Nat Genet*. 2004;36(5):431–432.

41. Halling–Brown MD, Bulusu KC, Patel M, et al. canSAR: an integrated cancer public translational research and drug discovery resource. *Nucleic Acids Res*. 2012;40(Database issue):D947–D956.

42. Kolker E, Higdon R, Haynes W, et al. MOPED:Model Organism Protein Expression Database. *Nucleic Acids Res*. 2012;40(Database issue):D1093–D1039.

43. Keshava Prasad TS, Goel R, Kandasamy K, et al. Human Protein Reference Database—2009 update. *Nucleic Acids Res*. 2009;37(Database issue):D767–D772.

44. Uhlen M, Oksvold P, Fagerberg L, et al. Towards a knowledge–based Human Protein Atlas. *Nat Biotechnol*. 2010;28(12):1248–1250.

45. Koivula RW, Tornberg AB, Franks PW. Exercise and diabetes–related cardiovascular disease: systematic review of published evidence from observational studies and clinical trials. *Curr Diab Rep*. 2013;13(3):372–380.

46. Cereda E, Barichella M, Pedrolli C, et al. Diabetes and risk of Parkinson's disease: a systematic review and meta–analysis. *Diabetes care*. 2011;34(12):2614–2623.

47. Colosia AD, Palencia R, Khan S. Prevalence of hypertension and obesity in patients with type 2 diabetes mellitus in observational studies: a systematic literature review. *Diabetes Metab Syndr Obes*. 2013;6:327–338.

48. Rappaport N, Nativ N, Stelzer G, et al. MalaCards: an integrated compendium for diseases and their annotation. *Database (Oxford)*. 2013;2013:bat018.

49. Marchler–Bauer A, Zheng C, Chitsaz F, et al. CDD: conserved domains and protein three–dimensional structure. *Nucleic Acids Res*. 2013;41(Database issue):D348–D352.

50. Mulder NJ, Apweiler R, Attwood TK, et al. InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform*. 2002;3(3):225–235.

51. Punta M, Coggill PC, Eberhardt RY, et al. The Pfam protein families database. *Nucleic Acids Res*. 2012;40(Database issue):D290–D301.

52. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42(Database issue):D222–D230.

53. Petersen TN, Brunak S, von Heijne G, et al. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*. 2011;8(10):785–786.

54. Liang L, Morar N, Dixon AL, et al. A cross–platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res*. 2013;23(4):716–726.

55. Stranger BE, Montgomery SB, Dimas AS, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet. 2012;8(4):e1002639.

56. Hom G, Graham RR, Modrek B, et al. Association of systemic lupus erythematosus with C8orf13–BLK and ITGAM–ITGAX. *N Engl J Med*. 2008;358(9):900–909.

57. International Multiple Sclerosis Genetics Consortium, Hafler DA, Compston A, et al. Risk alleles for multiple sclerosis identified by a genome wide study. *The New England journal of medicine*. 2007;357(9):851–862.

58. Cooper JD, Smyth DJ, Smiles AM, et al. Meta–analysis of genome–wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet*. 2008;40(12):1399–1401.

59. Todd JA, Walker NM, Cooper JD, et al. Robust associations of four new chromosome regions from genome–wide analyses of type 1 diabetes. *Nat Genet*. 2007;39(7):857–864.

60. 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(7145):661–678.

61. Sim X, Ong RT, Suo C, et al. Transferability of type 2 diabetes implicated loci in multi–ethnic cohorts from Southeast Asia. *PLoS Genet*. 2011;7(4):e1001363.

62. Benjamin EJ, Dupuis J, Larson MG, et al. Genome–wide association with select biomarker traits in the Framingham Heart Study. *BMC Med Genet*. 2007;8(Suppl 1):S11.

63. Kim SK, Shindo A, Park TJ, et al. Planar cell polarity acts through septins to control collective cell movement and ciliogenesis. *Science*. 2010;329(5997):1337–1340.

64. Imamura M, Maeda S, Yamauchi T, et al. A single–nucleotide polymorphism in ANK1 is associated with susceptibility to type 2 diabetes in Japanese populations. *Hum Mol Genet*. 2012;21(13):3042–3049.

65. Finckh U, van Hadeln K, Muller–Thomsen T, et al. Association of late–onset Alzheimer disease with a genotype of PLAU, the gene encoding urokinase–type plasminogen activator on chromosome 10q22.2. *Neurogenetics*. 2003;4(4):213–217.

66. Ertekin–Taner N, Ronald J, Feuk L, et al. Elevated amyloid beta protein (Abeta42) and late onset Alzheimer's disease are associated with single nucleotide polymorphisms in the urokinase–type plasminogen activator gene. *Hum Mol Genet*. 2005;14(3):447–460.

67. Babbs C, Roberts NA, Sanchez–Pulido L, et al. Homozygous mutations in a predicted endonuclease are a novel cause of congenital dyserythropoietic anemia Type 1. *Haematologica*. 2013;98(9):1383–1387.

68. Ahmed MR, Zaki M, Sabry MA, et al. Evidence of genetic heterogeneity in congenital dyserythropoietic anaemia Type 1. *Br J Haematol*. 2006;133(4):444–445.

69. Pierce BL, Ahsan H. Genetic susceptibility to type 2 diabetes is associated with reduced prostate cancer risk. *Hum Hered*. 2010;69(3):193–201.

70. Bradfield JP, Qu HQ, Wang K, et al. A genome–wide meta–analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet*. 2011;7(9):e1002293.

71. Feng T, Zhu X. Genome–wide searching of rare genetic variants in WTCCC data. *Hum Genet*. 2010;128(3):269–280.

72. Bailey SD, Xie C, Do R, et al. Variation at the NFATC2 locus increases the risk of thiazolidinedione–induced edema in the Diabetes REduction Assessment with ramipril and rosiglitazone Medication (DREAM) study. *Diabetes Care*. 2010;33(10):2250–2253.

73. Aoi N, Soma M, Nakayama T, et al. Variable number of tandem repeat of the 5'–flanking region of type–C human natriuretic peptide receptor gene influences blood pressure levels in obesity–associated hypertension. *Hypertens Res*. 2004;27(10):711–716.

74. Vassalle C, Andreassi MG, Prontera C, et al. Influence of ScaI and natriuretic peptide (NP) clearance receptor polymorphisms of the NP system on NP concentration in chronic heart failure. *Clin Chem*. 2007;53(11):1886–1890.

75. Fox AA, Collard CD, Shernan SK, et al. Natriuretic peptide system gene variants are associated with ventricular dysfunction after coronary artery bypass grafting. *Anesthesiology*. 2009;110(4):738–747.

76. Pitzalis MV, Sarzani R, Dessi–Fulgheri P, et al. Allelic variants of natriuretic peptide receptor genes are associated with family history of hypertension and cardiovascular phenotype. *J Hypertens*. 2003;21(8):1491–1496.

77. Eriksson N, Tung JY, Kiefer AK, et al. Novel associations for hypothyroidism include known autoimmune risk loci. *PLoS One*. 2012;7(4):e34442.

78. Prahalad S, Hansen S, Whiting A, et al. Variants in TNFAIP3, STAT4, and C12orf30 loci associated with multiple autoimmune diseases are also associated with juvenile idiopathic arthritis. *Arthritis Rheum*. 2009;60(7):2124–2130.

79. McKay JD, Truong T, Gaborieau V, et al. A genome–wide association study of upper aerodigestive tract cancers conducted within the INHANCE consortium. *PLoS Genet*. 2011;7(3):e1001333.

80. Stammers M, Rowen L, Rhodes D, et al. BTL–II: a polymorphic locus with homology to the butyrophilin gene family, located at the border of the major histocompatibility complex class II and class III regions in human and mouse. *Immunogenetics*. 2000;51(4–5):373–382.

81. Hirota T, Takahashi A, Kubo M, et al. Genome–wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. *Nat Genet*. 2011;43(9):893–896.

82. Stahl EA, Raychaudhuri S, Remmers EF, et al. Genome–wide association study meta–analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet*. 2010;42(6):508–514.

83. McKinnon E, Morahan G, Nolan D, Jet al. Association of MHC SNP genotype with susceptibility to type 1 diabetes: a modified survival approach. *Diabetes Obes Metab*. 2009;11(Suppl 1):92–100.

84. Jin Y, Birlea SA, Fain PR, et al. Genome–wide analysis identifies a quantitative trait locus in the MHC class II region associated with generalized vitiligo age of onset. *J Invest Dermatol*. 2011;131(6):1308–1312.