Open Reading Frames Associated with Cancer in the Dark Matter of the Human Genome

ANA PAULA DELGADO, PAMELA BRANDAO, MARIA JULIA CHAPADO, SHEILIN HAMID and RAMASWAMY NARAYAN

Department of Biological Sciences, Charles E. Schmidt College of Science, Florida Atlantic University, Boca Raton, FL, U.S.A.

Abstract. Background: The uncharacterized proteins (open reading frames, ORFs) in the human genome offer an opportunity to discover novel targets for cancer. A systematic analysis of the dark matter of the human proteome for druggability and biomarker discovery is crucial to mining the genome. Numerous data mining tools are available to mine these ORFs to develop a comprehensive knowledge base for future target discovery and validation. Materials and Methods: Using the Genetic Association Database, the ORFs of the human dark matter proteome were screened for evidence of association with neoplasms. The Phenome-Genome Integrator tool was used to establish phenotypic association with disease traits including cancer. Batch analysis of the tools for protein expression analysis, gene ontology and motifs and domains was used to characterize the ORFs. Results: Sixty-two ORFs were identified for neoplasm association. The expression Quantitative Trait Loci (eQTL) analysis identified thirteen ORFs related to cancer traits. Protein expression, motifs and domain analysis and genome-wide association studies verified the relevance of these OncoreORFs in diverse tumors. The OncoreORFs are also associated with a wide variety of human diseases and disorders. Conclusions: Our results link the OncoreORFs to diverse diseases and disorders. This suggests a complex landscape of the uncharacterized proteome in human diseases. These results open the dark matter of the proteome to novel cancer target research.

Uncharacterized proteins in the human genome offer a vast untapped potential (1-6). Whereas known genes are studied extensively, the novel unknown proteins, due to the challenge inherent in their paucity of information, have been scarcely studied (7-9). We have recently embarked on elucidating the nature of the uncharacterized proteins, the dark matter of the human proteome. Two open reading frames (ORFs) were characterized using diverse bioinformatics and proteomics approaches. An ORF termed C1orf87 was characterized as an EF-Hand containing, calcium-binding protein with putative transporter function (10). This ORF was termed as the carcinoma related EF-hand protein (CREF) and has been implicated as a putative biomarker for liver, breast and lung carcinomas. Furthermore, another ORF, CXorf66, was validated as a secreted glycoprotein linked to chromosome X (SGPX) in brain, lung, liver and prostate carcinomas and leukemia (11).

These findings prompted us to datamine the Genetic Association Database, GAD (12), for a cancer-associated fingerprint in the dark matter proteome. The studies revealed strong correlative evidence for cancer association among the uncharacterized ORFs. The cancer-related ORFs were also implicated in diverse diseases and disorders. A fingerprint of 62 cancer ORFs emerged from these investigations. These ORFs were termed the OncoreORFs due to their cancer association.

Using the genome-phenome analysis tool, the Phenome-Genome Integrator (PheGenI) (13), we demonstrate disease-associated traits for the OncoreORFs. By means of the Integrated Cancer Research and Drug Discovery Platform CanSar (14), Model Organism Protein Expression Database (MOPED) (15), the Human Protein Reference Database (HPRD) (16), GeneALaCart, GeneCards (17), the DAVID functional annotation tool (18) and the Human Protein Atlas (HPA) (19), putative classes for the ORF proteins were predicted for the OncoreORFs. These results allowed us to develop a rationale for druggability and biomarker potential for the OncoreORFs. A landscape of human diseases in the context of the OncoreORFs

Correspondence to: Dr. Ramaswamy Narayan, Department of Biological Sciences, Charles E. Schmidt College of Science, Florida Atlantic University, Boca Raton, FL, U.S.A. Tel: +561 2972247, Fax: 561-297-3859, e-mail: nnarayan@fau.edu

Key Words: Phenome-genome, expression quantitative trait loci, landscape of diseases, genome-wide association, batch analysis, druggable proteome, biomarkers, open reading frames, uncharacterized proteins, cancer association.
emerged from these studies implicating the neoplasm genes in other diseases. Our results help demystify the dark matter of the human proteome and provide new insights into novel target discovery for cancer and other diseases.

**Materials and Methods**

The bioinformatics and proteome tools used in the study have been described elsewhere (10, 11). In addition, the following genome-wide association tools were used: the Genetic Association Database, GAD (12), the DAVID functional annotation tool (18), GeneALaCart from the GeneCards (17), the Phenotype-Genotype Integrator (PheGeni, http://www.ncbi.nlm.nih.gov/gap/phegeni), the Database of Genomic Variants (DGV) (20), Clinical Variations (ClinVar) (21) and the International HapMap project (http://hapmap.ncbi.nlm.nih.gov/).

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

**Results**

**Landscape of cancer ORFs: Identification of the cancer fingerprint.** We have undertaken a comprehensive profiling of the cancer-associated fingerprint of the uncharacterized proteome. The GAD (http://geneticassociationdb.nih.gov/) archives human genetic association studies of complex diseases and disorders (12, 22). Association studies data from both the known and uncharacterized proteins are available in this database. The GAD enables the identification of clinically relevant polymorphism from the large volume of polymorphism and mutational data. The GAD encompasses a comprehensive summary of data extracted from peer-reviewed publications on candidate gene and Genome-Wide Association Studies (GWAS) (23).

The entire GAD database as of 2014-03-08 was downloaded and the uncharacterized open reading frames (ORFs) were manually curated. From the 65,536 entries in the complete GAD, 1,859 ORF-related entries were found. Multiple entries for each ORF represented different polymorphic, Single Nucleotide Polymorphism (SNP) rs numbers. These ORF entries were enriched for cancer fingerprint using three filters i) broad phenotypes, ii) disease classes and iii) Medical Subject Headings (MESH) terms, and 235 cancer-associated ORFs were identified (shown bolded in Table I, http://www.science.fau.edu/biology/faculty/table_1.pdf). Use of one of the three filters for term enrichment was ineffective as many ORFs associated with cancers were missed; hence it was necessary to use all three filters simultaneously. In view of the strong association for these ORFs with diverse cancer types, these ORFs are termed the Oncology Open reading frames “ OncoORFs”. The tumor type distribution of the OncoORFs is shown in Figure 1. The OncoORFs’ association was seen with multiple solid tumors and hematological malignancies. Some of the OncoORFs were associated with multiple tumor types. These OncoORFs provided the framework for detailed characterization studies to establish druggability and biomarker potential for cancer.

**OncoRF association with other diseases and disorders.** Mining the GAD also allowed us to infer the association of the cancer ORFs with other diseases and disorders including cardiovascular, chem dependency, developmental, hematological, immune, metabolic, neurological, pharmacogenomic and psych-related phenotypes (Table I) (Figure 2). Two of the 62 OncoORFs (C11orf53, uncharacterized protein, and C11orf93, cancer susceptibility candidate 3, colorectal cancer associated-2) were specific to colorectal cancer. The remaining OncoORFs’ association with other diseases and disorders ranged from simple to complex. For example, cancer and developmental disorders showed a simple association. On the other hand, cardiovascular, hematological or neurological ORFs showed a highly complex overlap of association of other diseases with neoplasms. These results raised the possibility that mining the uncharacterized cancer proteome (OncoORFs) might provide a framework to study other diseases and disorders as well.

**Characterization of OncoORFs.** From these studies sixty-two OncoORFs emerged which showed cancer/neoplasm association. An initial characterization was undertaken to develop a preliminary hint for the putative classes of the OncoORF proteins. The OncoORFs were batch-analyzed using the GeneALaCart tool (http://www.genecards.org/) from GeneCards. This batch analysis tool allowed us to query a large number of genes for multiple parameters according to the GeneCards data entries. The results were further verified using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 from the NCBI, Human Genome Nomenclature Committee (HGNC) and National Center for Biotechnology Information (NCBI) Gene. From the aliases and description, putative classes/nature of the ORFs were inferred. From these analyses, distinct protein classes including antigen, binding proteins, carrier, enzymes, pseudogenes, secreted, sorting, transporter, ncRNAs, vacuolar and Zinc finger containing putative nature of the coding sequences were identified for 50/62 OncoORFs (data not shown). Thirteen OncoORFs (C11orf53, C13orf18, C15orf59, C1orf94, C1orf95, C2orf43, C2orf47, C2orf80, C3orf14, C5orf36, C6orf15, C6orf99 and Cxorf67) are currently uncharacterized. The enzyme and transporter functions offer druggable targets and the secreted ORFs provide a rationale for diagnostic biomarkers. These results suggested that it is possible to develop a knowledgebase for the uncharacterized ORFs of the dark matter cancer proteome. Encouraged by these findings, we have undertaken a detailed characterization of the OncoORFs.
for motifs and domains, protein expression and mutational analyses in cancers.

**Elucidation of target potential for the OncoORFs.** Gene Ontology (GO) analysis provides an opportunity to predict function and putative class of the OncoORFs. The integrative cancer-focused knowledgebase, canSAR (http://cansar.icr.ac.uk) brings together data across diverse areas (biology, chemistry, pharmacology, structural biology, cellular networks and clinical annotations) to provide insight into analysis of proteins at different levels. The canSAR Meta analysis tool was used to perform a batch analysis of the OncoORFs to establish GO information: i) cell location, ii) function and iii) process. The results from these analyses were verified using the GeneAlaCart tool, the DAVID functional annotation tool, the MOPED and HPRD databases, the UCSC genome browser and diverse protein motif and domain analysis tools. The merged database was manually curated to extract the GO features. Figure 3 shows the GO feature distribution of the OncoORFs. The OncoORFs were classified into i) location (nuclear, integral membrane, cytoplasmic, mitochondrial, extracellular/secreted, etc.), panel A; ii) process (apoptosis, development, differentiation, cell cycle, inflammation, immune, metabolic, etc.), panel B; and iii) function (enzyme, protein binding, nucleotide/metal binding, transport, etc.), panel C. These results provide a starting point to elucidate the drug target and biomarker potential of the OncoORFs.

**OncoORF motif and domain characterization.** To develop further insight into the nature of the OncoORFs, the ORFs were analyzed for protein motifs and domains. The GeneAlaCart tool was used to batch analyze the OncoORFs for the InterPro/UniProt Domains and Families (24). In addition, the NCBI Conserved Domain Database, CDD (25), the Protein Family, PFAM (26), (27), the Biosequence analysis using profile hidden Markov models, HMMER (28), the Protein Domain Analysis, ProDom (29), UCSC Genome Browser (30) and SignalP (31) bioinformatics tools were used to analyze the OncoORFs (Table II, http://www.science.fau.edu/biology/faculty/table_II.pdf). 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 OncoORFs were grouped into classes of proteins. Protein families including antigens, carrier proteins, enzymes, nucleotide/metal binding, receptors, mitochondrial chaperones, phosphoproteins, secreted glycoproteins, selenoproteins, transporter/sorting proteins, vacuolar proteins and Zinc finger proteins were identified among the OncoORF proteins. The binary interaction data, post-translational modification as well as the protein architecture from the HPRD and the UCSC 3D model analysis (Table II), provided additional clues to the nature of the OncoORFs.

**OncoORF protein expression.** The mRNA and protein expression data provide a valuable clue to the specificity of the ORFs. Hence, the OncoORFs’ expression in normal and tumor tissues was investigated using the MOPED, HPA and HPRD and the National Cancer Institute (NCI) Clinical Proteomic Tumor Analysis Consortium (CPTAC, http://www.humanproteomemap.org) protein expression tools. The OncoORF data was enriched from the complete HPRD and HPA downloaded databases; the MOPED and the NCI clinical proteomics databases were batch analyzed using the OncoORFs. The tissue-restricted mRNA expression was inferred from UniGene and HPA tools (Table III, http://www.science.fau.edu/biology/faculty/table_3.pdf). Distinct expression profiles for numerous OncoORFs were detected in diverse tissues and body fluids. A tissue-enriched mRNA expression profile was seen for adrenal gland (C7orf16); brain (C2orf80, C2orf85, C7orf16); connective tissues (C20orf61); liver (C1orf38); placenta (CXorf67); skin (C6orf15); and testis (C1orf94, C20orf61, C20orf79, C7orf16, CXorf61, CXorf66, CXorf67). Protein expression in body fluids was noted for C1orf94 (blood plasma) and C3orf10 (semen) suggesting a possible secreted nature. The NCI clinical proteomics tool provided additional evidence for the expression of 24/62 OncoORFs in normal fetal and adult tissues. The OncoORF mRNA expression was also seen in the National Cancer Institute 60 (NCI60) cancer cell lines (data not shown).

**Cancer-associated traits of the OncoORFs.** The Phenotype-Genotype Integrator (PheGenI, http://www.ncbi.nlm.nih.gov/gap/phegeni) merges NHGRI 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). This phenotype-oriented resource facilitates follow-up studies from GWAS and allows prioritization of variants.

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, often measured as mRNA abundance. The three components critical to validating eQTL results include 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 (32).

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** (33) are available. In order to establish cancer associated eQTL results for the OncoORFs, the PheGeni tool was used to batch analyze

---

*Delgado et al: OncoORFs of the Dark Matter Proteome*
the ORFs. Genotypes for the OncoORFs were selected for exons, introns, near gene and Untranslated Region (UTR). From the output of results, neoplasm traits were enriched. The eQTL data for the OncoORFs are shown in Table IV, http://www.science.fau.edu/biology/faculty/table_4.pdf.

Thirteen OncoORFs showed strong eQTL association evidence with neoplasms with significant p-values (breast, colorectal, esophageal, gall bladder, head and neck, liver, ovary, prostate, renal and non-small-cell lung cancers, neuroblastoma and myeloid leukemia).

A common SNP (rs# 8170, cds, syn) in the OncoORF C19orf62 | BABAM1, a new member of the BRCA1 complex, was associated with breast and ovarian cancers. On the other hand, distinct SNPs in the OncoORF, C6orf97, CCDC170, a coiled-coil domain-containing protein was associated with myeloid leukemia (rs# 4869742, intron) and breast cancers (rs# 3757318, intron). Interestingly, an intron SNP of the OncoORF C6orf36 | CEP85L, a serologically-defined breast cancer antigen, NY-BR-15, was associated with renal cell carcinoma (SNP, rs# 25422). The OncoORF C10orf11, an autosomal recessive oculocutaneous albinism 7 gene (OCA7), was associated with tamoxifen resistance in breast cancer patients (SNP rs# 83938, intron).

Association was also found for druggable enzymes (probable E3 ubiquitin-protein ligase C12orf51, HECTD4, C3orf21, xyloside xylosyltransferase 1 and C12orf30, N(alpha)-acetyltransferase 25, NatB auxiliary subunit) with esophageal, non-small-cell lung and head and neck neoplasms. The eQTL browser allows the identification of statistically significant SNPs across diverse populations from the population genomics (HapMap) project. The Genotype-Tissue Expression (GTEx) tool from the eQTL browser currently includes tissue expression data related to lymphoblastoid, liver and brain tissues (34). This tool enables the linking of phenotype with population-related polymorphism across diverse populations. One of the eQTL hits, C11orf93 | colorectal cancer-associated 2 | COLCA2, showed strong correlation across multiple studies: YRI-Yoruba in Ibadan, Nigeria and CEU-CEPH Utah residents with ancestry from Northern and Western Europe (35).

OncoORF mutational analysis. The Catalogue of somatic mutations in cancer database (COSMIC, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) has comprehensive mutation data for both the known and uncharacterized proteins (36). The entire COSMIC database was downloaded and enriched for the OncoORFs. Fifty-six OncoORFs harbored diverse mutations (nonsense, missense, deletions, insertions, frameshifts, in frames, homozygous and heterozygous) as shown in Figure 4. Substitution missense mutations were the largest class of mutations found for the OncoORFs. In addition, homozygous mutations (n=105) were present for sixteen of the OncoORFs (C1orf127, C1orf94, C2orf43, C3orf58, C6orf15, C6orf99, C7orf10, C7orf16, C12orf51, C17orf57, C18orf8, C19orf62, C19orf63, C20orf108, CXorf66, CXorf67). The homozygous mutational types included substitution missense, deletions and insertions and were present in malignant melanomas, endometrial, cervical, breast, renal and stomach carcinomas (37). Heterozygous mutations (N=1,080) were also present in the OncoORFs (37-41).
Three of the homozygous mutant OncoORFs included the eQTL neoplasm association: i) C19orf62, new component of the BRCA1 A complex, breast cancer; ii) C2orf43, UPF0554 protein, prostate cancer; and iii) C12orf51, HECT domain-containing E3 ubiquitin protein ligase, esophageal neoplasm.


**OncoORF alterations across multiple tumor types.** The cBioPortal Meta analysis tool (http://www.cbioportal.org) allows the cross-cancer analysis of multiple query genes for amplifications, deletions and mutations (42). The OncoORFs were batch analyzed using multiple tumor datasets (Figure 5). In 69 studies, 60 of the 62 OncoORFs were simultaneously perturbed at the levels of amplifications, deletions and somatic mutations. A large number of tumor types in the database harbored amplifications (15-78%) for multiple OncoORFs (shown in red). In addition, somatic mutations and deletions were also present for the OncoORFs across solid tumors and hematological malignancies.

**Genome-wide association of the OncoORFs.** The results of the association studies (GWAS) on the OncoORFs using the NCBI tools (PheGenI and eQTL analysis) predicted the OncoORFs’ association with susceptibility loci to diverse tumors (Tables I and IV). Structural variants for the eQTL neoplasm hits were analyzed using the Database of Genomic Variants (DGV). Complex arrays of variants for seven of the thirteen eQTL hits were identified (Table V, http://www.science.fau.edu/biology/faculty/table_5.pdf). Clinical variants were identified using the NCBI ClinVar Database for two of the OncoORFs (C1orf94, uncharacterized protein and C10orf11, oculocutaneous albinism 7, OCA7, autosomal recessive). The C1orf94 variant (p.Gly580Glu) was associated with malignant melanoma and the C10orf11 variants (p.Ala23Argfs*39 and p.Arg194*) with germline. These two mutations are shown to be pathogenic for autosomal recessive albinism (43).

The spectrum of simple somatic mutations of the eQTL neoplasm hits was next investigated using the International Cancer Genome Consortium (ICGC) database (Table VI). Mutations were identified for most of the eQTL OncoORFs (14/16) across multiple samples from around the world. Mutational subtypes included missense, nonsense, insertions and deletions (data not shown). The OncoORF mutations...
were seen in diverse tumor types including the tumor types associated with the eQTL traits (indicated by * in Table VI, http://www.science.fau.edu/biology/faculty/table_6.pdf). Neoplasm-associated evidence for the OncoORFs is summarized below.

Breast cancer. In breast cancer, four OncoORFs showed evidence of a strong association. The C10orf11 (oculocutaneous albinism 7, autosomal recessive, OCA7) was associated with a successful clinical outcome of adjuvant therapy with tamoxifen in Japanese patients (44). These studies showed that the SNP rs10509373 in C10orf11 gene on 10q22 was significantly associated with recurrence-free survival in the replication study [log-rank \( p=0.0002 \)], and a combined analysis indicated a strong association of this SNP with recurrence-free survival in breast cancer patients treated with tamoxifen [log-rank \( p=1.26 \times 10^{-10} \)].

The C17orf37 (migration and invasion enhancer 1, HBV XAg-transactivated protein 4) located near the ERBB2 gene locus in chromosome 17q21 was frequently associated with heightened breast cancer risk in British women (45, 46) and the OncoORF CXorf61 (cancer/testis antigen 83, Kita-kyushu lung cancer antigen 1) was associated with susceptibility loci to the breast and prostate cancers (47).

Figure 3. OncoORF gene ontology. The OncoORFs were analyzed for GO using the canSAAR, GeneAlacart (GeneCards), the Model Organism Protein Expression Database (MOPED) and the UCSC genome browser by batch analysis. The number of OncoORFs with extracted GO features from these analyses are shown. Panel A, cellular compartment; panel B, function and panel C, process.

Prostate cancer. The OncoORFs C9orf3 (aminopeptidase OIAOPEP), C2orf43 (UPF0554 protein), CXorf61 (cancer/testis antigen 83 I Kita-kyushu lung cancer antigen1) and CXorf67 are associated with prostate cancer susceptibility loci (47, 49-53). The C9orf3 (aminopeptidase O, AOPEP) is associated with the development of erectile dysfunction in African-American males following radiation therapy for prostate cancer (50). The C2orf43 (UPF0554 protein) was found to be one of five new susceptibility loci for prostate cancer in the Japanese population (51). The
C2orf80, on the other hand, is linked to aggressive (but not indolent) prostate cancer risk (54). The CXorf61 (cancer/testis antigen 83 | Kita-kyushu lung cancer antigen 1) was also associated with breast cancer, suggesting a possible link to hormonal cancers (47).


Von Hippel-Lindau (VHL) syndrome is a dominantly inherited familial cancer syndrome caused by mutations in the VHL gene. Large VHL gene deletions associated with a contiguous loss of C3orf10 (haematopoietic stem/progenitor cell protein 30) were associated with a significantly lower lifetime risk of renal cell carcinoma (RCC) than deletions that did not involve C3orf10. These results strongly suggest that the C3orf10 may be predictor of risk for RCC (61).

Leukemia. In childhood acute lymphoblastic leukemia (ALL), association was seen with the C18orf2 (chromatin-modifying protein 1B | vacuolar protein-sorting 46-2) (62); C2orf47 (mitochondrial) with treatment response (63); C14orf118 (G patch domain-containing protein 2-like) (64), C18orf2, charged multivesicular body protein 1b, CHMP1B (62) and C6orf97, coiled-coil domain-containing protein 17 with chronic myeloid leukemia (65).

Landscape of OncorFs in other diseases and disorders. Polymorphic traits in OncorFs were also found to be associated with other diseases and disorders including cardiac, hematological, immune, metabolic, neurological,
psychiatric, developmental and chemedependency (see Figure II). Two OncoORFs (C11orf53, uncharacterized protein and C11orf93, cancer susceptibility candidate 13 | colorectal cancer-associated 2) were uniquely associated with colorectal cancers (35, 66). The vast majority of the OncoORFs, however, were associated with multiple diseases and disorders.

**Pharmacogenomics.** A pharmacogenomics (response to therapy) potential of the OncoORFs can be inferred for C10orf11 (oculocutaneous albinism 7, OCA7) in breast cancer response to tamoxifen (44); C2orf47 (UPF0554 protein) in acute lymphoblastic leukemia (67); C7orf10 (caib/baf1 CoA-transferase family protein) in AIDS disease progression (59); and C1orf97 (long intergenic non-protein coding RNA 467) in hematological diseases (68).

**Diabetes.** The association results for C12orf30 (N-terminal acetyltransferase B complex subunit NAA25) indicate that individuals with increased susceptibility to type I or II diabetes have a decreased risk for prostate cancer development (69). The OncoORF C16orf15, protein STG, taste bud-specific protein is associated with systemic lupus, Behcet syndrome, leprosy and follicular lymphoma (70-73).

A long intergenic non-protein coding RNA (C6orf208) showed a strong association with RCC and type I 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. The C18orf8 (colon cancer-associated protein Mic1) is associated with inflammatory processes and prostate cancer susceptibility (74). This provides strong evidence of a link between inflammation and cancer (75-76). Perturbation of inflammatory cytokines and cell adhesion molecules is a common event in numerous tumor types (77-79).

**Other diseases and disorders.** The C20orf54 (solute carrier family 52, riboflavin transporter, member 3) associated esophageal neoplasms and squamous cell carcinoma with multiple sclerosis (23). The C1orf127, on the other hand, associated Ewing’s sarcoma with multiple sclerosis (80). The OncoORF C7orf10 (caib/baf1 CoA-transferase family protein) showed association of pancreatic cancer with type II diabetes and AIDS (81-82). The OncoORF C6orf97 | coiled-coil domain-containing protein 170 links breast cancer and leukemia with stroke and osteoporosis (65, 83).

The OncoORF C8orf42, testis development-related protein, TDRP2 is associated with neuroblastoma, prostate carcinoma, stroke, hematological and cardiac diseases (84-85).

The OncoORF C7orf10, dermal papilla-derived protein 1 links pancreatic cancer with aging, type II diabetes and cardiac diseases (82). The line RNAs C6orf115 and C20orf61 associate pancreatic cancer, squamous cell carcinoma and esophageal carcinoma with Alzheimer’s disease (86, 87).

Two OncoORFs (C1orf94, uncharacterized protein and C14orf118, G patch domain-containing protein 2-like) link breast cancer with attention deficit disorder and alcoholism (64, 88).

**Discussion**

**Cancer fingerprint of the OncoORFs.** Uncharacterized proteins in the human genome offer a biomarker and drug target potential (4, 7, 89). In addition, new functional knowledge may emerge from studies involving the dark matter of the human genome (1, 10, 11). Our recent results with two of the uncharacterized ORFs as cancer biomarkers (10, 11) encouraged us to explore the rest of the uncharacterized ORF proteins in the human genome for cancer association.

The GAD database offered us an opportunity to develop a landscape of uncharacterized ORFs in diverse diseases including neoplasms. The genome-phenome association studies provided an attractive starting point for further mining the uncharacterized ORFs. Use of batch analysis tools, such as the GeneALaCart from the GeneCards, the DAVID functional annotation tool from NCBI (18) the canSA R integrated database (14), and the protein analysis tools MOPED (15), HPRD (16) and HPA (19) greatly aided our efforts to demystify the dark matter of the human proteome.

Our results indicate the presence of a core group of cancer-related ORFs, the OncoORFs, among the uncharacterized proteins. Sixty-two uncharacterized cancer-related ORFs (OncoORFs) have been characterized in this study. We demonstrate strong association of these OncoORFs with diverse solid tumors and hematological neoplasms. It is tempting to speculate that the 62 newly characterized OncoORFs may provide a novel multi-tumor fingerprint and have the potential to facilitate rational drug discovery and diagnosis. Additional experiments are needed to verify these findings.

The OncoORF fingerprint broadly encompassed enzymes, membrane receptors, transporters, DNA/nucleotide/metal binding proteins and secreted proteins. Valuable clues to the nature of the OncoORFs were obtained from multiple proteomic tools. The availability of 3-D template models from the UCSC genome browser allows for molecular modeling of the OncoORFs for further analysis. Reduction of the uncharacterized protein ORFs into putative classes with hints of protein motifs and domains will facilitate the future prioritization of these ORFs for follow-up studies. Efforts are underway to develop the OncoORFs as a novel cancer signature.

The GWAS results demonstrate strong correlative evidence of cancer association for the OncoORFs within the human dark matter proteome. Furthermore, the results from the COSMIC, ICGC, DGV and cBioPortal mutational analyses
Cross-cancer alteration summary for ADAM20P1, BABAM1, BRK1, C100RF11, C110RF53, C120RF57, C150RF59, C180RF8, C190RF12, C10RF127, C10RF94, C10RF95, C20RF43, C20RF47, C20RF80, C30RF14, C30RF58, C50RF38, C60RF15, C60RF99, C90RF5, CCDC170, CCDC178, CEP128, CEP85L, COLCA2, CT83, CXorf66, CXorf67, EFCA811, EFCA813, EMC10, EMC7, FAM210B, FAM2128, HECTD4, ISM1, KIAA0226L, KIAA0825, LINC00469, LINC00470, LINC00472, LINC00520, LINC00523, LINC00574, LINC00687, LRRCC37A5P, MIE1, NAA25, P5P1R17, RTP5, SCP2D1, SLC52A3, SMIM19, SUGCT, TCAIM, TDFR, TEFM, THEMIS2, XXYL1 (69 studies / 60 genes)

Figure 5. Cross-cancer alterations of the OncoORFs. The eBioPortal tool was used to batch analyze the OncoORFs for gene amplifications (red), deletions (blue) and mutations (green). The alteration frequency of the OncoORFs in indicated study is shown. Copy number variations (CNV) and mutational data if available are indicated (+/−). Tumor types are color coded.
suggest cancer relevance across a diverse population. Additional detailed association studies are warranted to develop the OncoORFs as a novel multi-tumor fingerprint.

The eQTL analysis provides valuable clues to link particular OncoORFs with cancer and other diseases. The thirteen OncoORFs (C12orf51, C19orf62, C3orf21, C12orf30, C10orf11, C2orf43, C6orf97, C11orf93, C1orf94, C14orf143, C18orf34, C6orf204, C5orf36) that emerged from the eQTL evidence (Table IV) require further extensive studies to establish evidence of a strong neoplasm association.

Complex landscape of OncoORFs in other diseases and disorders. Considerable evidence indicates that genes can be associated with more than one disease or disorder (90-94). Sixty of the 62 OncoORFs showed polymorphic association with phenotypes for diverse disorders and diseases such as addiction, AIDS, aging, Alzheimer's disease, alcoholism, ADD, asthma, cancer, cardiac diseases, diabetes types I and II, immune disorders, osteoporosis, multiple sclerosis, neurodegenerative diseases, rheumatoid arthritis, systemic lupus erythematosus and stroke (Table 1) (Figure 2). Some of these associations were simple – for example, cancer with neurological, immune, or developmental disorders. In contrast, the association of the OncoORFs with cardiovascular, hematological and metabolic conditions was complex and involved more than one disease or disorder. Thus, the same OncoORF is likely to be involved in multiple pathways affecting diverse cell and tissue types.

Uncharacterized ORFs. Some OncoORFs (12/62) were uncharacterized. The use of multiple bioinformatics tools allowed a detailed characterization of nine of the uncharacterized ORFs. Functional classes were identified for the novel OncoORFs, including enzymes (C11orf53, C1orf95, C2orf43, C6orf99, C15orf59); RNA/nucleotide binding (C13orf18, C2orf47); secreted proteins (C15orf24, C19orf63, C6orf15); transporter (C5orf36) and Ras/actin binding (C2orf43, C3orf14). These OncoORFs provide a valuable opportunity to develop novel druggable targets and diagnostic biomarkers.

These results offer new insight into the role of the OncoORFs across a complex spectrum of diseases. Future understanding of the OncoORFs' functions and the pathways involved can help develop a comprehensive framework to link different diseases through common genes. These findings open a novel avenue for target discovery in the complex landscape of human diseases. This can make the drug discovery process more effective to target multiple therapeutics.

Summary

The results presented herein should provide strong impetus to further investigate the uncharacterized proteins of the human genome. The association of the uncharacterized ORFs with diverse diseases that we have demonstrated in the present study provides an attractive starting point for understanding the role of the dark matter proteome in the human disease landscape. In view of the US National Cancer Institute's recent focus on the druggability of the dark matter proteome, our results provide an early example of systematically demystifying the dark matter by means of diverse bioinformatics and proteomics approaches. We speculate that science is on the threshold of a new era of disease target discovery within the dark matter proteome.

Conflicts of Interest

None.

Acknowledgements

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 CanSar, cBioPortal, the TCGA and COSMIC for the mutations dataset, the Human Protein Atlas, Human Protein Reference Database and DAVID functional annotations tools for various datasets. This work was supported in part by the Genomics of Cancer Fund, Florida Atlantic University Foundation. We thank Jeanine Narayanan for editorial assistance.

References


