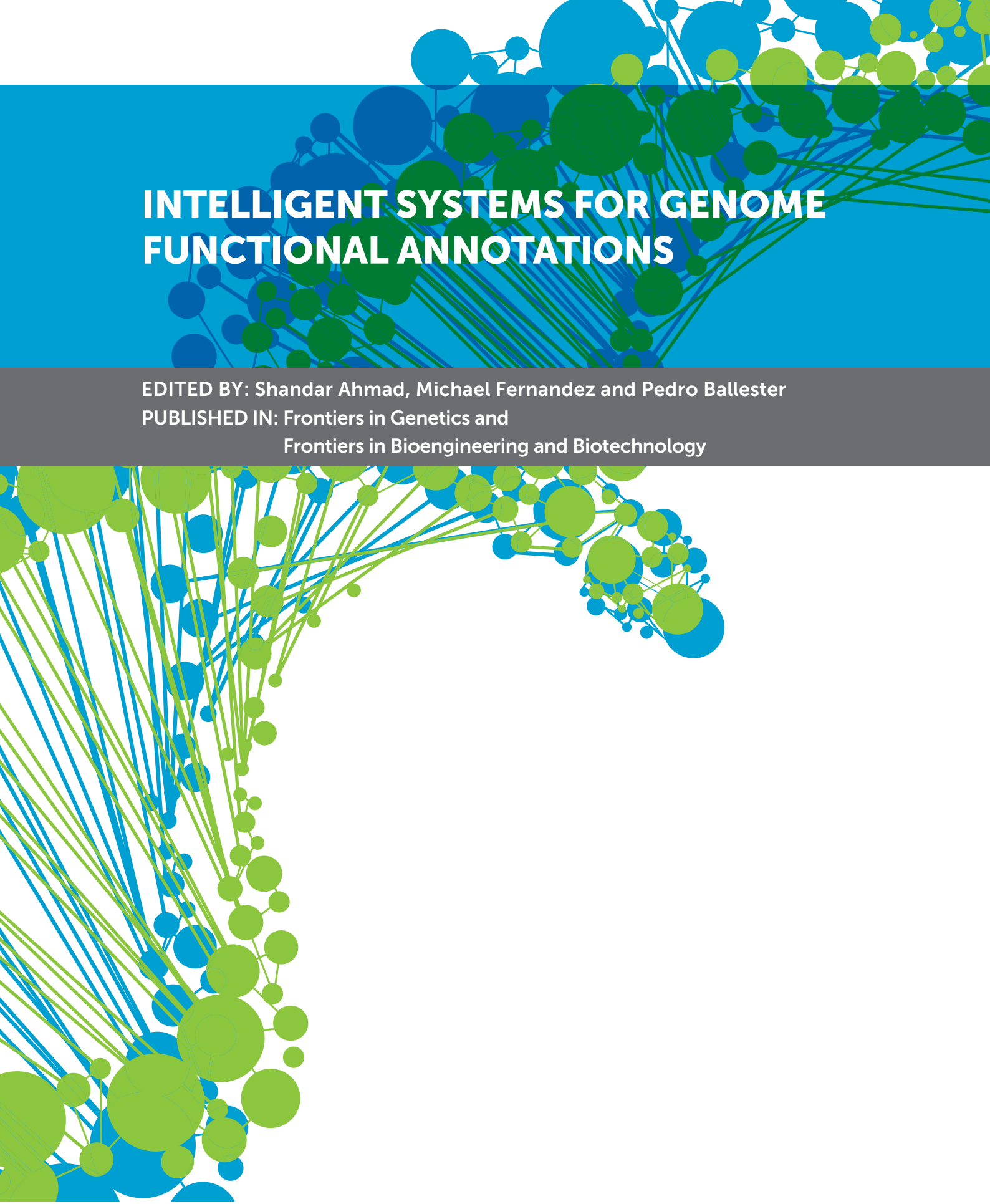

INTELLIGENT SYSTEMS FOR GENOME FUNCTIONAL ANNOTATIONS

EDITED BY: Shandar Ahmad, Michael Fernandez and Pedro Ballester

PUBLISHED IN: Frontiers in Genetics and
Frontiers in Bioengineering and Biotechnology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-090-2
DOI 10.3389/978-2-88966-090-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

INTELLIGENT SYSTEMS FOR GENOME FUNCTIONAL ANNOTATIONS

Topic Editors:
Shandar Ahmad, Jawaharlal Nehru University, India
Michael Fernandez, The Vancouver Prostate Centre, Canada
Pedro Ballester, INSERM U1068 Centre de Recherche en Cancérologie de Marseille (CRCM), France

Citation: Ahmad, S., Fernandez, M., Ballester, P., eds. (2020). Intelligent Systems for Genome Functional Annotations. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-090-2

Table of Contents

frontiers in Genetics

EDITORIAL
published: 25 August 2020
doi: 10.3389/fgene.2020.00915

Editorial: Intelligent Systems for Genome Functional Annotations

Shandar Ahmad[1]*, *Pedro J. Ballester*[2,3,4,5] *and Michael Fernandez*[6]

[1] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India, [2] Cancer Research Center of Marseille, INSERM U1068, Marseille, France, [3] Institut Paoli-Calmettes, Marseille, France, [4] Aix-Marseille Université, Marseille, France, [5] CNRS UMR7258, Marseille, France, [6] Department of Urologic Sciences, Faculty of Medicine, Vancouver Prostate Centre, University of British Columbia, Vancouver, BC, Canada

Keywords: functional annotation, protein-protein interaction (PPI), machine learning, gene annotation, intelligent system applications

Editorial on the Research Topic

Intelligent Systems for Genome Functional Annotations

Functional annotation of an entire genome is critical to the understanding of any biological process or pathway. Yet, large parts of the human genome, and much more in the non-model organisms, remain without annotations. Simple, sequence-similarity based annotations have been found to be grossly inadequate for this purpose. More complex models, often based on intelligent systems, such as Machine Learning (ML) have proved to be very helpful. Indeed, ML models have key properties that makes them particularly useful for genome annotation (Yip et al., 2013). In their basic formulation, ML techniques have found their way into the field of biological functional annotation quite early. For example, secondary structure prediction using ML was carried out as early as in mid-1980's and many other areas of biological sequence, structure and/or function prediction have seen great advances in terms of the complexity of techniques, feature engineering, and other principles of data-driven analytics.

Several computational techniques have been developed exclusively for solving functional annotation problems. However, most of the growth has been in terms of the application of emerging and established computational techniques in the biological domain. ML software has often been used as a blackbox tool, while researchers focus on the biological concept of the problem and its solution. More recently, deep learning based neural networks methods have made rapid progress and have shown particular success with problems associated with large amounts of biological data and complex system representations. Typically popular amongst them have been convolutional neural networks (CNN), multi-layer perceptrons (MLP) and long short-term memory networks (LSTM), with widely different forms and learning strategies. On the other hand the biological understanding of molecular function and organization of knowledge on this subject has also undergone rapid advances. Instead of scattered and ambiguous labeling of function, systematic annotations in terms of biological (disease, gene, and protein etc.) ontologies with hierarchical and nested labels from semantic models have made the task of annotation learning and prediction of biological function much more robust. Clearly, much has been achieved on biological and technical aspects of functional annotations, but many hurdles remain. It was under this context, this special issue was proposed to promote the reporting of various aspects of biological functional annotations where different types of intelligent systems/ML have been used to solve functional annotation problems.

In the eight research papers forming this special issue, a special aspect of functional annotation i.e., for predicting drugability was reported. There is a need to annotate gene products to indicate whether these are likely to be druggable or not. Ghadermarzi et al. argued that the majority of

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

***Correspondence:**
Shandar Ahmad
shandar@jnu.ac.in

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *17 July 2020*
Accepted: *23 July 2020*
Published: *25 August 2020*

Citation:
Ahmad S, Ballester PJ and Fernandez M (2020) Editorial: Intelligent Systems for Genome Functional Annotations. Front. Genet. 11:915. doi: 10.3389/fgene.2020.00915

the druggable human proteome is yet to be annotated and explored. To advance on that front, these authors collected the data from three types of protein targets: druggable, non-druggable, and possibly druggable. Both new and established markers for each protein were extracted from its protein sequence or names/identifiers. They discovered that the possibly druggable proteins have significantly higher abundance of alternative splicing isoforms, relatively large number of domains, higher degree of centrality in the protein-protein interaction networks, and lower numbers of conserved and surface residues, when compared with the non-druggable proteins. These markers can be helpful to find novel druggable human proteins and provide interesting insights into the cellular functions and subcellular locations of both current drug targets and potentially druggable proteins.

The genome can also be annotated with those regions whose alterations in tumor cells are found to control patient response to a drug treatment. Thus, the contribution from Bomane et al. looked at this problem from a precision oncology perspective. In particular, authors investigated the extent to which it is possible to predict breast cancer patient response to the mitotic inhibitor paclitaxel using the US National Cancer Institute's Genomic Data Commons. These datasets comprised the responses of breast cancer patients to paclitaxel along with six molecular profiles of their tumors. Ten ML algorithms were applied to each of these profiles and the resulting 60 classifiers evaluated on the held-out patients. Only three of these 60 models were at all predictive, highlighting the crucial importance of a broad search to avoid suboptimal results. DNA methylation and miRNA profiles were the most informative overall. In combination with these two profiles, the ML algorithms selecting the smallest subset of molecular features were found to generate the most predictive classifiers.

In addition to supervised ML models for function annotation, unsupervised ML methods have shown to be extremely successful to interpret experiments when labeled experiments are not available. For example, annotations from newly invented single-cell RNA sequencing (scRNA-seq) technology (Sasagawa et al., 2019) is incredibly challenging because of the lack of labels for individual cells. Purely unsupervised clustering methods, such as t-distributed stochastic neighbor embedding and uniform manifold approximation and projection, have been employed to obtain low-dimensional embedding of cell–cell relationships with the foreseeable drawback of highly dependent upon genes selected for clustering. Taguchi and Turki contribution explored tensor decomposition (TD)-based unsupervised feature extraction (FE) (Taguchi, 2019) to integrate two scRNA-seq expression profiles that measure human and mouse midbrain development. TD-based unsupervised FE showed to be a promising method to effectively integrate two scRNA-seq profiles while outperforming other popular unsupervised selection methods.

The integration of biological experiments also requires informatics tools and platforms that combine and analyse different sources of biological data (Triplet and Butler, 2014). TargetMine is an integrative data analysis platform for target prioritization and broad-based biological knowledge discovery. The recent improvement of the platform described by Chen et al. forms a contribution in that direction and highlights newly modeled biological data types and the improvement of new analytical and visual tools. Enhanced coverage of gene–gene relations, and small molecule metabolite to pathway mappings are now implemented in TargetMine together with an improved literature survey feature. The platform also incorporated *in silico* predictions of gene functional associations such as protein–protein interactions and global gene co-expression. Authors demonstrated how the newer enhancements in TargetMine provides a more expansive coverage of the biological data space and can help interpret genotype–phenotype relations.

Finding new biological targets is key for designing potential new drugs. Computational biology can help identifying targets by sorting the parasite's metabolic pathways that pins out proteins essential for its survival. Bora and Jha contributed a kinetic modeling for determining targets against *Leishmania donovani*, a deadly human pathogen responsible for causing *Visceral Leishmaniasis.* Metabolic pathway and Protein-Protein Interactions (PPI) were integrated to analyse the "purine salvage" pathway, which is mandatory for parasite survival. Available experimental data was used to develop a kinetic model of Purine salvage pathway that helped marking of crucial enzymes involved in the synthesis of the metabolites. Additionally, PPI analysis of the pathway assisted in building a static interaction network for selected proteins. Dynamic Modeling and Topological analysis of the PPI network through centrality measures were combined to detected targets. ADSL (Adenylosuccinate lyase) and IMPDH (Inosine-50-monophosphate dehydrogenase) enzymes appeared to be crucial and further modeling of three dimensional structure of ADSL enzyme aided toward the search for antiparasitic drugs for the treatment of *Visceral Leishmaniasis*.

In terms of specific issues discussed in this special issue, Wang et al. have reported the use of one of the top unsupervised machine learning technique known as overlapping cluster generator (OCG) for the functional characterization of hitherto poorly annotated P311. They propose that the proteins on the interface of OCGs represent multifunctional property and based on this PPI characterization propose that P311 may be involved in inflammatory responses, cell proliferation, and coagulation. While protein-wise functional annotation is a key biological problem of interest, more detailed characterization of proteins to gain general insights into their structure and function are critically important. In this regards, amino acid repeats in proteins play an important role in their structures and by consequence their functions. Thus, Rajathei et al. have reported an analysis of protein repeat regions and their role in structure and function of proteins. They also report that repeat regions of longer than 15 residues are present in about 67% of proteins in the Uniprot, when viewed in a non-redundant manner. Biological function annotation cannot be complete without looking at the sensitivities at the epigenetic and single nucleotide polymorphism-driven functional diversity. Database and resources form the centerstage in any such analysis. Ma et al. have presented a database FeatSNP that focuses specifically on the SNPs in epigenetic factors in human brain. In the absence of a thorough understanding of human brain and proteins

involved in performing cognitive and behavioral functions, such a database will be of immense value for people working on understanding genomic perspectives of protein function in brain and its related disorders.

Overall, the special issue covered various aspects of ML both supervised and unsupervised with classical clustering, overlapping clustering and general statistical principles, neural networks, tensor decomposition and related techniques. The biological systems investigated included generalized druggability, P331 systems, brain associated proteins and *Leishmania donovani*. Thus, special issue brings together typical systems in which ML and intelligent systems have helped gaining predictive value and biological insights into the vast area of biological function annotation. We hope the readers from computational biology and the domain specific researchers will be benefitted by reading articles included in this special issue.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by a grant from Department of Science and Technology, Government of India under the project DST/ICPS/Cluster/DataScience/2018/General/11 to SA.

REFERENCES

Sasagawa, Y., Hayashi, T., and Nikaido, I. (2019). "Strategies for converting RNA to amplifiable cDNA for single-cell RNA Sequencing Methods," in *Single Molecule and Single Cell Sequencing. Advances in Experimental Medicine and Biology*, ed Y. Sasagawa (Singapore: Springer), 1017. doi: 10.1007/978-981-13-6037-4_1

Taguchi, Y. (2019). Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data. *BMC Bioinformatics* 19:298. doi: 10.1186/s12859-018-2395-8

Triplet, T., and Butler, G. (2014). A review of genomic data warehousing systems. *Brief. Bioinform.* 15, 471–483. doi: 10.1093/bib/bbt031

Yip, K. Y., Cheng, C., and Gerstein M. (2013). Machine learning and genome annotation: a match meant to be?. *Genome Biol.* 14:205. doi: 10.1186/gb-2013-14-5-205

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 February 2019
doi: 10.3389/fgene.2019.00109

Reconstruction and Functional Annotation of P311 Protein–Protein Interaction Network Reveals Its New Functions

Song Wang[1], Xiaorong Zhang[1], Fen Hao[1], Yan Li[2], Chao Sun[3], Rixing Zhan[1], Ying Wang[1], Weifeng He[1], Haisheng Li[1,4] and Gaoxing Luo[1]**

[1] Institute of Burn Research, State Key Laboratory of Trauma, Burn and Combined Injury, Southwest Hospital, Third Military Medical University, Chongqing, China, [2] Laboratory Center of Southwest Hospital, Third Military Medical University, Chongqing, China, [3] The Sixth Resignation Cadre Sanatorium of Shandong Province Military Region, Qingdao, China, [4] The 324th Hospital of Chinese People's Liberation Army, Chongqing, China

OPEN ACCESS

Edited by:
Rosalba Giugno,
University of Verona, Italy

Reviewed by:
Hauke Busch,
Universität zu Lübeck, Germany
Lingyun Zou,
Third Military Medical University, China
Vincenzo Bonnici,
University of Verona, Italy

***Correspondence:**
Haisheng Li
lee58427@163.com
Gaoxing Luo
logxw@yahoo.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 19 September 2017
Accepted: 30 January 2019
Published: 19 February 2019

Citation:
Wang S, Zhang X, Hao F, Li Y, Sun C, Zhan R, Wang Y, He W, Li H and Luo G (2019) Reconstruction and Functional Annotation of P311 Protein–Protein Interaction Network Reveals Its New Functions. Front. Genet. 10:109. doi: 10.3389/fgene.2019.00109

P311 is a highly conserved multifunctional protein. However, it does not belong to any established family of proteins, and its biological function has not been entirely determined. This study aims to reveal the unknown molecular and cellular function of P311. OCG (Overlapping Cluster Generator) is a clustering method used to partition a protein-protein network into overlapping clusters. Multifunctional proteins are at the intersection of relevant clusters. DAVID is an analytic tool used to extract biological meaning from a large protein list. Here we presented OD2 (OCG + DAVID + 2 human PPI datasets), a novel strategy to increase the likelihood to identify biological functions most pertinent to the multifunctional proteins. The principle of OD2 is that OCG prepares the protein lists from multifunctional protein relevant overlapping clusters, for a functional enrichment analysis by DAVID, and the similar functional enrichments, which occurs simultaneously when analyzing two human PPI datasets, are supposed to be the predicted functions. By applying OD2 to two reconstructed human PPI datasets, we supposed the function of the P311 in inflammatory responses, cell proliferation and coagulation, which were confirmed by the following biological experiments. Collectively, our study preliminarily found that P311 could play a role in inflammatory responses, cell proliferation and coagulation. Further studies are required to validate and elucidate the underlying mechanism.

Keywords: P311, protein–protein interaction networks, inflammatory response, cell proliferation, coagulation

INTRODUCTION

P311, with the official gene symbol NERP (neuronal regeneration related protein), is a highly conserved 8-kDa intracellular protein. The 68-amino acid sequence of P311 contains a PEST domain (rich in Pro, Glu, Ser, and Thr) in the N-terminus (Stradiot et al., 2018). The domain is also in short-lived proteins such as transcription factors, cytokines and signal molecules, which implies that P311 might belong to one of the protein families (Sommer and Wolf, 2014; Varshavsky, 2014). However, no more evidence was found to ascribe P311 to one of the protein families. So far,

P311 does not belong to any established family of proteins; therefore, it fails to provide any clues on its function. Studler et al. (1993) first reported that P311 was highly expressed in embryonic mouse brains, and then other groups demonstrated that P311 was expressed in motoneurons (Fujitani et al., 2004), glioblastomas (McDonough et al., 2005), smooth muscle cells (Badri et al., 2013), and fibroblasts (Tan et al., 2010; Cheng et al., 2017). Furthermore, these studies showed that P311 was involved in nerve and lung regeneration (Fujitani et al., 2004; Zhao et al., 2006), glioma invasion ((Mariani et al., 2001; McDonough et al., 2005), blood pressure homeostasis (Badri et al., 2013), myofibroblast differentiation (Pan et al., 2002), amoeboid-like migration (Shi et al., 2006), behavioral responses in learning and memory (Taylor et al., 2008) and the affective, but not the sensory component of pain (Sun et al., 2008).

Our group has been focused on P311 since 2004, when we found that the expression of P311 increased dramatically in hypertrophic scars through gene expression profiling and a comparative proteomics analysis (Wu et al., 2004). We then found that P311 induced fibroblast differentiation via enhancing the TGFβ1 signaling pathway in a human hypertrophic scar (Tan et al., 2010), and it also was a new inducer of EpMyT (Epidermal stem cell transdifferentiate into myofibroblasts) in wound healing (Li et al., 2016). Furthermore, we demonstrated that P311 played a crucial role in renal fibrosis via TGFβ1/Smad signaling (Yao et al., 2015). Recently we showed that P311 accelerated skin wound re-epithelialization by promoting epidermal stem cell migration through RhoA and Rac1 activation (Yao et al., 2017) and that P311 deficiency leads to attenuated angiogenesis in cutaneous wound healing (Wang et al., 2017). These studies have aroused our tremendous and sustained interest in P311 and we therefore continue to study its biological function.

Protein-protein interaction (PPI) networks can highlight the modularity of cellular processes and allow the deciphering of protein functions at the cellular level, as proteins tend to interact with each other when they are involved in the same molecular complex, pathway, or biological process (Hartwell et al., 1999). Meanwhile, a PPI network can be represented as a simple graph in which vertices correspond to proteins and edges, to direct physical interactions, which allow the graph partition method to highlight clusters of densely connected vertices (Aittokallio and Schwikowski, 2006). The identified clusters stand for groups of proteins involved in the same molecular complex, pathway, or biological process. Further analyzing the proteins from each group can predict the function of uncharacterized proteins (Sharan et al., 2007). OCG (Overlapping Cluster Generator) is a graph partition that decomposes a protein-protein network into overlapping clusters. Multifunctional proteins are at the intersection of relevant clusters (Becker et al., 2012). DAVID consists of a comprehensive biological knowledgebase as well as analytical tools designed to systematically extract biological meaning from a large gene/protein list (Huang et al., 2008).

In this study, we presented OD2 (OCG + DAVID + 2 human PPI datasets), a promising strategy to predict the function of P311. By applying OD2 to two reconstructed human PPI datasets, we supposed the function of P311 in inflammatory responses, cell proliferation and coagulation. Finally, we conducted relevant biological experiments to confirm the functions of P311.

MATERIALS AND METHODS

Datasets

A PPI network dataset (named Dataset 1) involving 80,930 binary interactions between 10,229 proteins (**Figure 1A** and **Supplementary Material 1**) was constructed by (1) eight interactions from our previous literature (Peng et al., 2012) and (2) 80,922 binary interactions from the PPI network dataset assembled by Bossi and Lehner (2009).

Additionally, another high confidence dataset (named Dataset 2) of 110,707 binary interactions involving 9,606 proteins (**Figure 1A** and **Supplementary Material 3**) was built by fusing the eight interactions from our previous literature (Peng et al., 2012) with 110699 binary interactions, whose combine-score is greater or equal to 0.5, from the STRING database (Szklarczyk et al., 2014). In the STRING dataset, each protein-protein interaction is annotated with a combine-score. The score does not necessarily indicate the strength or specificity of the interaction, instead, it indicates confidence, i.e., how likely STRING judges an interaction to be true, given the available evidence. All scores rank from 0 to 1, with 1 being the highest possible confidence.

OCG (Overlapping Cluster Generator) Algorithm

The OCG algorithm was carried out by the software available in Becker et al. (2012) (**Supplementary Material 4**). The principle of OCG is to build a tree in which the leaves are introductory classes that are progressively and hierarchically joined (**Figure 1B**).

Initial Overlapping Class System

This program begins with building initial overlapping clusters from a simple unweighted graph $G = (V, E)$. Let $|V| = n$ and $|E| = m$. To obtain the initial overlapping class system, there are three strategies to cover the graph.

The covering class system can be:

(1) Set of edges: The Newman's algorithm (Newman, 2006) continues to search for two clusters V_i and V_j such that $\forall(x, y) \in V_i \times V_j$, $(x, y) \in E$ to maximizes the modularity Q of the system.
(2) Set of maximal cliques: When computing the maximal cliques, a class system with maximal modularity Q is formed. Any class fusion will cause Q to decrease until $Q_{\min} = \sum_{x=1-n} d_x^2$.
(3) Set of centered cliques: For each vertex $x \in G$, a greedy polynomial algorithm is used to construct a clique. As long as a clique is generated, the vertices adjacent to x are arranged in descending order of the relative degree. The resulting clique that contains x is not necessarily the maximal one, as there may be a larger one containing x.

The centered clique system has been chosen for further studies, because its graph density is better and is less time and memory consuming.

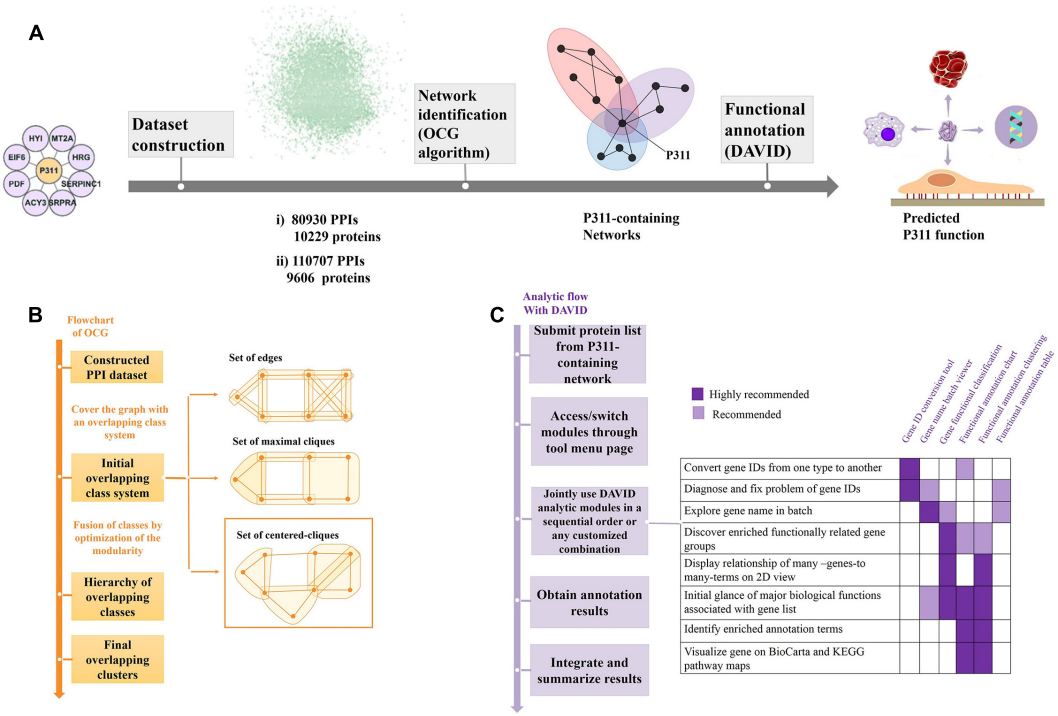

FIGURE 1 | Flow diagram of the bioinformatic analysis. **(A)** Flow diagram of OD2 approach. The initial P311-containing network identified in the Y2H screen (Peng et al., 2012) were integrated with the human PPI dataset gathered by Bossi and Lehner (2009) and dataset from the STRING database (Szklarczyk et al., 2014), separately. The resulting datasets were decomposed and identified the reconstructed P311-containing network using an OCG algorithm. The proteins in each reconstructed P311-containing network were analyzed with DAVID. Finally, predicted functions of P311 were identified. **(B)** Flowchart of the OCG (Overlapping Cluster Generator) algorithm. A graph was covered by the initial overlapping class system with centered cliques. Based on optimization of modularity of the partition, a hierarchy of nested clusters was built iteratively, leading to the final overlapping clusters (Becker et al., 2012). **(C)** Analytic flow with DAVID (Huang et al., 2008).

Hierarchy of Overlapping Class

To obtain the hierarchy overlapping class systems, the need to be fused with the clusters by optimization of the modularity Q. In each step, the connected clusters are the ones that maximize the average gain. This average is defined as the global modular gain divided by the number of newly connected vertex pairs. The chain effect can thus be avoided, which is caused by adding elements one by one and producing clusters that are not suitable for the following functional prediction.

Final Overlapping Clusters

Final overlapping clusters come out when the system of clusters reaches the maximized modularity. Alternatively, if the researcher sets either a minimum number of clusters or the maximum allowed cluster cardinality, the final overlapping cluster that maximizes the overall modularity within those constraints come out.

P311 PPI Networks Reconstruction and Analysis

As shown in **Figure 1A**, through the OCG (Overlapping Cluster Generator) algorithm, the dataset was partitioned into overlapping PPI networks (**Supplementary Material 5**). Among those, we picked out the networks containing protein P311 (NERP) as P311-containing networks.

Following the analysis flow with DAVID (**Figure 1C**), all constituents in each P311-containing network were analyzed, separately, to find the significant terms. The Q-values < 0.05 were invoked as the threshold from which to choose the significant terms (Katsogiannou et al., 2014). Q-values are hypergeometric p-values corrected for multiple testing according to the Benjamini and Hochberg procedure. Cytoscape was utilized to visualize all the networks (Smoot et al., 2010). Significant GO terms occurring in both terms from dataset 1 derived P311-containing networks and dataset 2 derived P311-containing networks, were picked out as the common terms, which were supposed to be the functions of P311.

Following the bioinformatic analysis above, we identified the common predicted function as the function which would be confirmed in the follow-up experiments.

Animals

P311 WT and P311 KO mice were kindly gifted by Professor Gregory A Taylor (Taylor et al., 2008). All mice grew up in the animal Institute of Third Military (Army) Medical University. The mice were maintained in a specific, pathogen-free environment under controlled light, temperature, and humidity.

Culture of Mouse Primary Fibroblasts

Cells were cultured as previously described (Varani et al., 2004). Briefly, after incubation in 0.25% Dispase II (04942078001. Roche) overnight at 4°C, the dermis was separated and minced into tissue fragments. Then primary fibroblasts were grown from the fragments and cultured in Dulbecco's modified Eagle's medium (DMEM) (11965118, Gibco) supplemented with 10% fetal bovine serum (FBS) (10099141, Gibco), 100 U/mL of penicillin and streptomycin (15140122, Gibco).

Proliferation Assay

Mouse primary fibroblasts (MPFs) were seeded in three replicates in 96-well plates in DMEM supplemented with 10% FBS. After 1, 2, 3, 4, 5, and 6 days, according to the manufacturer, the absorbance was measured at 450nm using an enzyme-linked immunosorbent assay reader with the addition of 10 μl/well of CCK8 reagent (CK04, Dojindo).

Full-Thickness Excisional Skin Wound Model

The model was prepared as previously described (Wang et al., 2018). Briefly, hairs on the dorsal surface of mice were shaved with an electric shaver and cleaned with 75% alcohol before surgery. A full-thickness excisions skin wound was made on the dorsal surface, with a 4 mm round skin biopsy punch, while anesthetized, with 1% pentobarbital via an intraperitoneal injection.

Superficial Second-Degree Burn Mouse Model

As previously described (Li et al., 2016), mice were anesthetized with intraperitoneal pentobarbital (35 mg/kg) and dorsal skin hairs were shaved 2 days before the burn. The scald apparatus (YSL-5Q) was then used to make the second-degree thermal burn injury with the condition (80°C, 3 s under a pressure of 500 g weight). Two wounds were produced on each mouse along the posterior median line, and the distance between the two wounds was 1.0 cm. The burn depth was confirmed by pathology.

Hematoxylin-Eosin (H&E) Staining

The mice were sacrificed on the 3rd-day post-surgery, and the wound tissues were then carefully harvested, fixed with 4% paraformaldehyde, embedded in paraffin, sliced and stained with H&E. The wound area and inflammatory cell count on the hematoxylin and eosin (H&E) stained sections were determined by the ImageJ 1.41 software provided by the National Institute of Health.

RNA Isolation and Quantitative Real-Time PCR

Total RNA was extracted from mouse skin with the RNeasy Mini Kit (QIAGEN, 74104), and cDNA was synthesized with the cDNA Synthesis Kit (Toyobo, FSK-100). Real-time PCR was performed with the SYBR Green Master Mix (Toyobo, QPK-201) on a 7500 Real-Time PCR System (Applied Biosystems Instruments). The following primers were used:

CD14, 5′-ACATCTTGAACCTCCGCAACGTGT-3′ and 5′TTGAGCGAGTGTGCTTGGGCAATA-3′; CD16, 5′-TTGCAGTGGACACGGGCCTTTATT-3′ and 5′TTGTCTTGAGGAGCCTGGTGCTTT-3′; β-actin, 5′-CGTGCGTGACATCAAAGAGAA-3′ and 5′-TGGATGCCACAGGATTCCCAT-3′; GAPDH, 5′-CGTGCCGCCTGGAGAAAC-3′ and 5′-AGTGGGAGTTGCTGTTGAAGTC-3′; P311, 5′-GAGGCTTCCTAAGGGAAGACTT-3′ and 5′-AAGTGGAGGTAAC TGATTCTTGG-3′.

Flow Cytometry

After isolating cells from the dermal sheet, the appropriate primary antibody was added for 1 h at 4°C to PerCP CY5.5-conjugated mAb specific F4/80 (Cell Signaling). After transduction with Ad-P311 or Ad-Vector for 24 h, MPFs (mouse primary fibroblasts) were scraped off the plates in PBS containing 5% BSA. Cells were then fixed in 70% ethanol in 4°C overnight, washed in PBS two additional times, and then stained for 30 min at 37°C in a 50 mg/ml propidium iodide (Sigma, United States) solution containing 200 mg/ml RNase A and 0.1% Triton-X-100. All the prepared cells were analyzed with the Attune Acoustic Focusing Cytometer (Applied Biosystems, Life Technologies, CA, United States), and the data were analyzed using the Flow Jo software (Tree Star Incorporation, United States). Experiments were replicated at least three times using the same conditions and settings.

Thromboelastography (TEG)

Thromboelastography (TEG) is a method of testing the efficiency of blood coagulation (Branco et al., 2014). Blood was gathered from the retro-orbital plexus of the P311 WT-burn, P311 KO-burn, P311 WT-sham-burn, and the P311 KO-sham-burn mice on the 7th-day post-burn, in tubes containing buffered sodium citrate. A minimum of 1 mL was achieved in each case. The previously described (Bolliger et al., 2012), the TEG analysis at 37°C using the 'Citrated Native' protocol with TEG 5000 Thromboelastography Hemostasis Analyzer System (Haemonetics Corporation, Braintree, MA, United States) was used. Descriptions of the TEG parameters are provided in **Table 1**.

TABLE 1 | Description of thromboelastography (TEG) parameters.

Parameter	Description
Reaction time (R, min)	The speed of initial clot formation
Angle (α- Angle, degrees)	A measure of clot rate, the tangent of the curve at 2 mm amplitude
Maximum amplitude (MA, mm)	A reflection of clot strength, the maximum clot strength
G-value (*G*,*k*)	A measure of clot strength, calculated overall clot strength
Per cent lysis at 30 min (LY30, %)	A measure of blood clot dissolution, the measure of per cent clot lysis 30 min after MA
Per cent lysis at 60 min (LY60, %)	A measure of blood clot dissolution, the measure of per cent clot lysis 60 min after MA

Statistical Analysis

For DAVID analysis, the entire genome-wide genes of humans were the default background, and the significance of the gene-term enrichment was analyzed by a modified Fisher's exact test (EASE score). For follow-up experiments, the data were described as mean ± SD (standard deviation) and analyzed by an unpaired, two-tailed Student's *T*-test with SPSS 18.0 software. $P < 0.05$ was considered as statistically significant.

RESULTS

Functional Enrichment Analysis of P311 PPI Networks Predicts Its New Functions

Previously, our group identified eight proteins that might interact with P311, utilizing the yeast two-hybrid (Y2H) technique. These proteins are HRG, SERPINC, MT2A, SRPR, HYI, ACY3, EIF6, and PDF (**Supplementary Material 2**) (Peng et al., 2012). The nine proteins then constructed the initial P311-containing network. Following the analytic flow of DAVID (**Figure 1C**), according to functional annotation and enrichment, the proteins, with a *q*-value more than 0.05, were enriched in the negative regulation of endopeptidase activity ($Q = 0.9404$).

As shown in **Figure 1A**, the initial P311-containing network was merged with the human PPI network dataset assembled by Bossi and Lehner (2009) and the human PPI network dataset downloaded from STRING database (Szklarczyk et al., 2014), was used separately, to build two datasets. We obtained two reconstructed human PPI datasets, one (named Dataset 1) with 80,930 binary interactions between 10,229 proteins (**Figure 1A** and **Supplementary Material 1**), another (named Dataset 2) with 110,707 binary interactions involving 9,606 proteins (**Figure 1A** and **Supplementary Material 3**).

The two large human PPI networks were partitioned by OCG using the centered clique system to initially cover the graph. The final overlapping clusters emerged when the maximal modularity was reached (**Figure 1B**). Finally, 732 overlapping clusters, four of which contained P311, were obtained from Dataset 1. The four reconstructed P311-containing networks were named M1, M2, M3, M4 (**Supplementary Materials 5, 6**). Meanwhile, we obtained four reconstructed P311-containing networks, among 1588 overlapping clusters, obtained from Dataset 2. The four reconstructed P311-containing networks were named N1, N2, N3, N4 (**Supplementary Materials 5, 7**). Overall, we obtained eight reconstructed P311-containing networks.

All constituents in each reconstructed P311-containing network were then analyzed by DAVID, separately (**Figure 1C**). In the analysis of dataset 1 derived P311-containing networks, according to functional annotation and enrichment, these functions range from biological processes already reported (**Supplementary Material 8**) to novel ones (**Supplementary Material 9**), such as the GPI anchor biosynthetic process, glucose metabolic process, peptidyl-serine phosphorylation, chemokine-mediated signaling pathway, monocyte chemotaxis, cellular response to interferon-gamma,G1/S transition of mitotic cell cycle, DNA replication, platelet activation and the positive regulation of the establishment of protein localization to the plasma membrane. The top ten functions of each reconstructed P311-containing network are shown in **Figure 2A**.

In the analysis of dataset 2 derived P311-containing networks, according to functional annotation and enrichment, the predicted functions also included the already reported ones (**Supplementary Material 8**) and the new ones (**Supplementary Material 10**). The novel functions were enriched in the carboxylic acid catabolic process, monocarboxylic acid catabolic process, rRNA catabolic process, rRNA processing, the G1/S transition of mitotic cell cycle, DNA replication initiation and so on. The top 10 functions of each reconstructed P311-containing network are shown in **Figure 2B**.

To improve the confidence of the bioinformatic analysis, we compared the predicted functions from dataset 1 and dataset 2 to find the ones occurring on both sides. As shown in **Table 2**, one identified function (positive regulation of cell migration) (McDonough et al., 2005; Yao et al., 2017) was predicted by the analysis system. Another 13 unidentified ones were reported.

Integrated with the predicted functions above, and the phenomenon we observed before during our experiments, we supposed the function of P311 in inflammatory responses, cell proliferation and coagulation as the most confident ones.

The Contribution of P311 to the Inflammatory Responses During Wound Healing

Inflammatory responses are the hallmark pathophysiological procedure for wound healing. After injuries such as trauma, burns or surgery, the prompt recruitment of inflammatory cells, such as monocytes and macrophages, which express CD14 and CD16 markers, occur in the wound site, followed by neutrophils and lymphocytes. These cells control the inflammatory response to wounds (Eming et al., 2007). Histological analysis of the wounds was carried out to check the microscopic appearances of wounds and the number of inflammatory cells that have infiltrated into the subcutaneous areas. The results showed that on the 3rd-day after the injury, the number of inflammatory cells in P311 WT was significantly higher than that in P311 KO mice (231.5 ± 63.9/field vs. 141.4 ± 39.2/field, $p < 0.05$) (**Figures 3A,B**). Further, we detected the mRNA levels of CD14 and CD16 expressed in wounds from the two groups on the 1st- and 3rd- day after the injury. We found more than a 150-fold change in CD14 expression and more than a 50-fold change of CD16 expression in wounds in P311 WT mice compared with non-wounded skin ($p < 0.05$) (**Figures 3C,D**). Additionally, changes in P311 KO mice were considerably smaller than those in P311 WT mice. On the 3rd day after the injury, the levels of CD14 and CD16 decreased compared with those on the 1st day after injury, but the expression levels in P311 WT mice were still significantly higher than that in P311 KO mice ($p < 0.05$) (**Figures 3C,D**). Detail CT values are provided in **Supplementary Material 11**. Finally, we utilized flow cytometry to detect the percentage of F4/80+ inflammatory cells in the wounds. Consistent with the findings above, we found that wounds from P311 KO mice showed a significant reduction

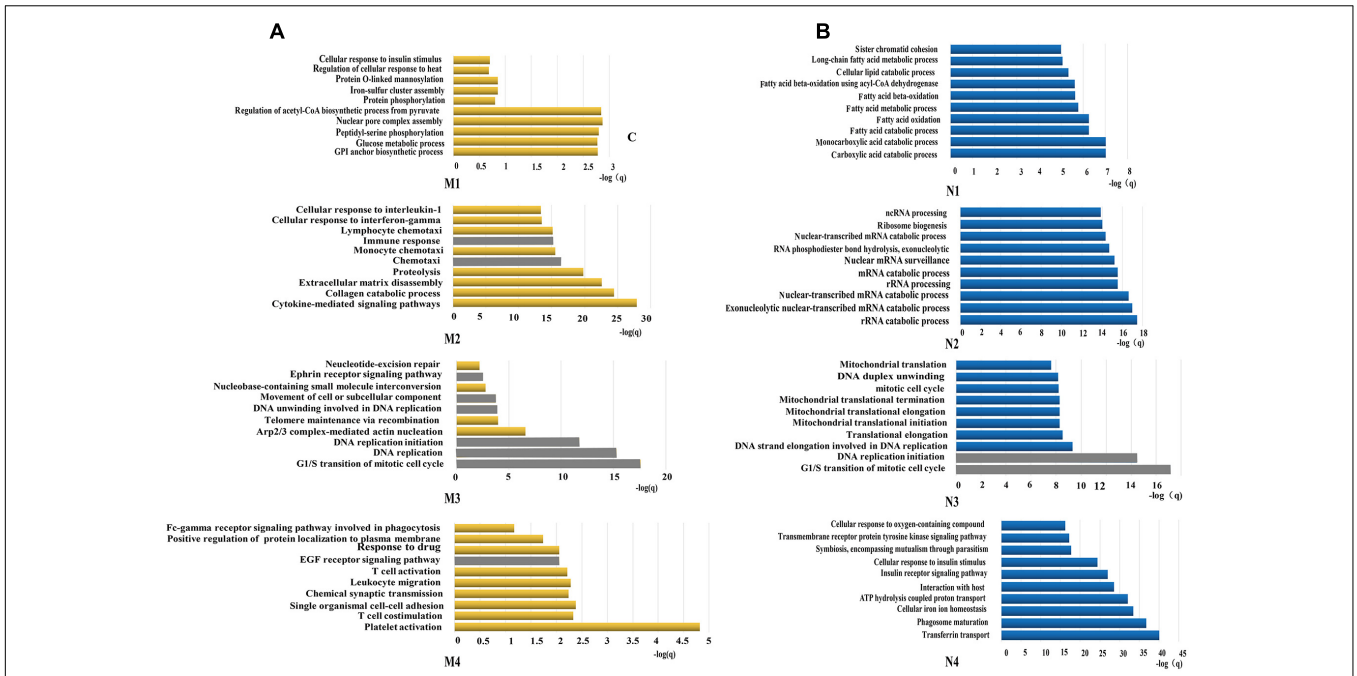

FIGURE 2 | Functions predicted by the bioinformatic analysis. **(A)** Represents the top 10 predicted functions of P311 in modules from dataset 1. **(B)** Represent the top 10 predicted functions of P311 in modules from dataset 2. Gray squares stand for the predicted common functions.

in the percentage of F4/80+ inflammatory cells in the wounds ($13.3 \pm 1.3\%$ vs. $9.0 \pm 1.4\%$, $p < 0.05$) on the 3rd day after injury (**Figures 3E,F**). All these findings confirmed that P311 influences the inflammatory responses during wound healing, by affecting the initial inflammatory cell recruitment.

TABLE 2 | Predicted function of P311.

New predicted functions of P311	Known functions of P311
• The G1/S transition of the mitotic cell cycle (M3, N3).	• Positive regulation of cell migration (M2, N2)
• Regulation of transcription involved in the G1/S transition of mitotic cell cycle 90 (M3, N3).	
• DNA replication (M3, N3).	
• DNA unwinding involved in DNA replication (M3, N3).	
• DNA replication initiation (M3, N3).	
• Movement of a cell or subcellular component (M3, N2).	
• Chemotaxis (M2, N2).	
• Immune response (M2, N2).	
• Epidermal growth factor receptor signaling pathway (M4, N2).	
• Blood coagulation (M2, N2).	
• Cellular protein metabolic process (M2, N2).	
• Innate immune response (M2, N2).	
• Ephrin receptor signaling pathway (M3, N2).	

M, N stand for the P311-containing network from dataset one and dataset two, respectively.

P311 Overexpression Promotes Proliferation of Mouse Primary Fibroblasts (MPF)

To confirm the function of P311 in proliferation, P311 was cloned into a pAdEasy vector which expresses GFP (green fluorescent protein), and then MPFs were transfected with a recombinant adenovirus vector (P311) or a negative vector (GFP). Before flow cytometry analysis and proliferation assay, the transfection efficiency was quantified and verified by observing GFP expression using a fluorescent microscope, and P311 mRNA expression levels were checked in real-time PCR. After transfection for 48 h, we observed that more than 90% of MPFs were transfected in both groups (**Figures 4A,B**). Real-time PCR result displayed that the mRNA level of P311 in the recombinant adenovirus vectors (P311) group was more than 10000-fold higher than that in the negative vector (GFP) group ($p < 0.05$), which confirmed the over-expression of P311 in the recombinant adenovirus vectors (P311) group. Detailed CT values are appended in the **Supplementary Material 11**. Further, proliferation capacity of these transfected cells was assessed using a proliferation assay. The result showed that from day 3, cells in the recombinant adenovirus vectors (P311) group had a higher proliferation capacity than cells in the negative vector (GFP) group ($p < 0.05$), and this tendency continued to the end ($p < 0.05$), which indicated that P311 over-expression significantly promoted the proliferation of MPFs (**Figure 4D**).

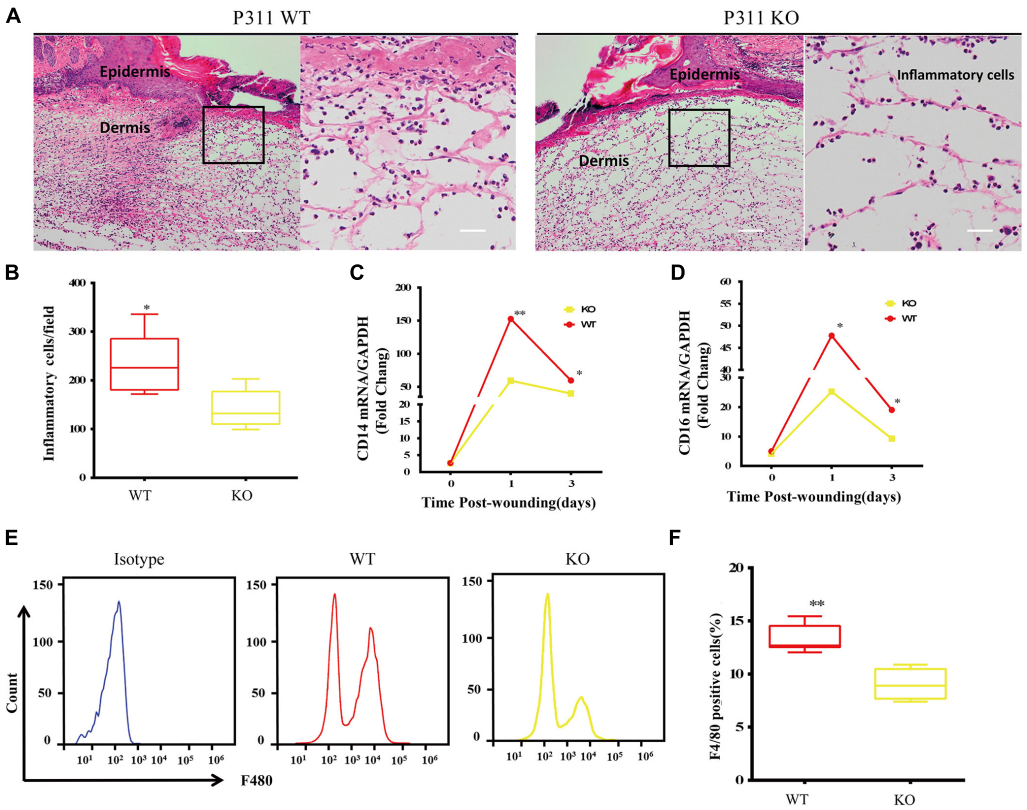

FIGURE 3 | Effect of P311 on the inflammatory response during wound healing. **(A)** Representative histological analysis of a skin wound by H&E staining on the 3rd day. The area marked by the black box in the left column is enlarged to display the mononuclear inflammatory cells in the right column. Scar bar = 200 μm, 50 μm. **(B)** Quantitation of mononuclear inflammatory cells in the granulation tissues from H&E-stained wound sections on the 3rd day (n = 3 animals per genotype). CD14 **(C)** and CD16 **(D)** mRNA expression in wounds on days 0, 1, and 3 after injury (n = 3 animals per time point and genotype). GAPDH and β-actin are the housekeeping genes used to perform the qPCR analysis. **(E)** Representative flow cytometry of F4/80. **(F)** Quantitation data of flow cytometry (n = 3 animals per genotype). Data represented mean ± SD, Student T-test, $^*P < 0.05$, $^{**}P < 0.01$, P311 KO vs. WT.

To understand how P311 promoted the proliferation of MPFs, we determined the cell cycle status of transfected cells by PI (Propidium iodide) staining. We found that the negative vector (GFP) group had a higher proportion of cells in the G0/G1 phrase (51.71 ± 4.21% vs. 41.05 ± 3.49%, $p < 0.05$), while the proportion of cells in the S phase was significantly lower in the negative vector (GFP) group than in the recombinant adenovirus vectors (P311) group (25.16 ± 4.92% vs. 36.68 ± 3.56%, $p < 0.05$) (**Figure 4F**). Moreover, there was no difference in the proportion of cells in the G2/M between the two groups ($p > 0.05$). These finding were consistent with the prediction by bioinformatics analysis. All of which indicated that P311 promoted MPFs proliferation via enhancing cells into the S phase.

The Coagulation Profile of Blood Obtained From P311 WT Versus P311 KO Mice Immediately and 7-Days Post Burned

To assess the function of P311 in coagulation, we established a superficial second degree burn mouse model, as burn injury is traditionally thought to be a common triggering cause of coagulopathy, ranging from activation of coagulation to disseminated intravascular coagulation (DIC) (Curreri et al., 1975). Thromboelastography (TEG) was then used to monitor the coagulation profile of blood after trauma, as growing evidence shows that TEG is better than the conventional laboratory tests in evaluating the coagulation profile, platelet dysfunction, and fibrinolysis after trauma (Branco et al., 2014; Pekelharing et al., 2014).

Figure 5A represented a schematic TEG tracing. Results of TEG analysis are shown in **Table 3**, while **Figure 5B** represents TEG tracings from the two groups. On the 7th-day post-injury, no difference was detected between the P311 WT-sham and the P311 KO-sham injury. The clot formation time (R 8.850 ± 1.115 vs. 8.167 ± 1.291 min, $p = 0.955$) was similar, and the clot rate (α- angle 6.075 ± 1.656° vs. 6.275 ± 1.819°, $P = 0.876$) was also almost the same. No fibrinolysis was observed for both the P311 WT or the P311 KO mice. Comparing the P311 WT-burn to the P311 KO-burn, except for the clot formation time (R 8.850 ± 1.541 vs. 7.733 ± 1.210 min, $p = 0.57$), other parameters (Angle, MA, and G) were all significantly different ($p < 0.05$), which implied that P311 might regulate the process

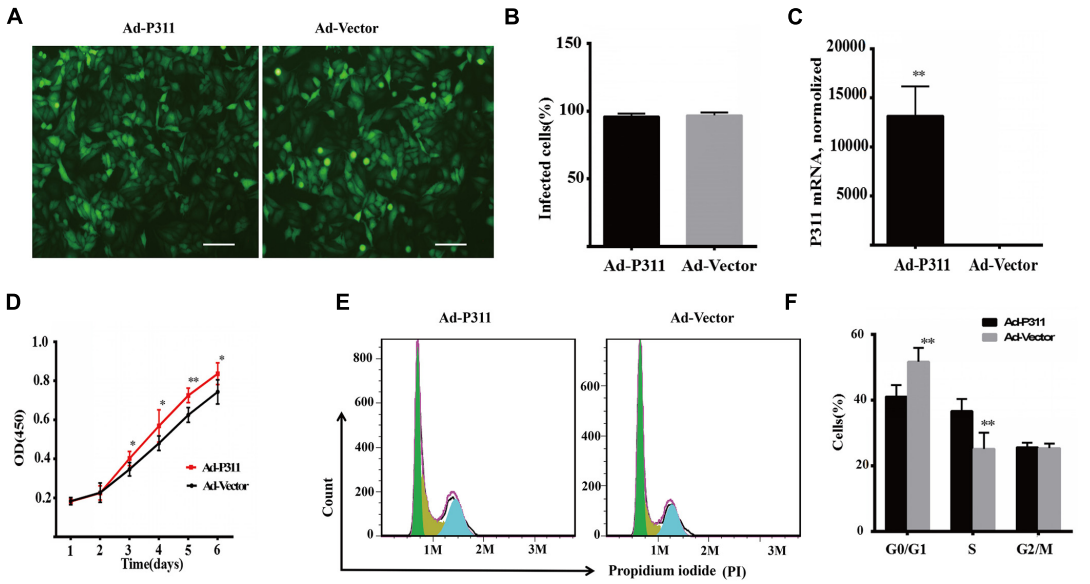

FIGURE 4 | P311 overexpression promotes proliferation of primary mouse fibroblasts (MPF). **(A)** Representative morphology of primary mouse fibroblasts (MPF) transfected with Ad-P311 or a negative vector (GFP). Transfected for 48 h and then observed under a fluorescence microscope to confirm the infection efficiency by visualizing GFP expression. Scale bar = 100 mm. **(B)** Quantitation data of GFP^+ cells fraction in P311 transfected MPFs and vector group to determine the transfection efficiency. **(C)** P311 mRNA level in P311 transfected MPFs and vector group (n = 4 per group). **(D)** Representative flow cytometry cell cycle. **(E)** Quantitation data of flow cytometry cell cycle (n = 6). **(F)** The CCK8 assay was performed to assess the effect of P311 on the proliferation (n = 6). $^*P < 0.05$, $^{**}P < 0.01$, Ad-P311 vs. Ad-Vector.

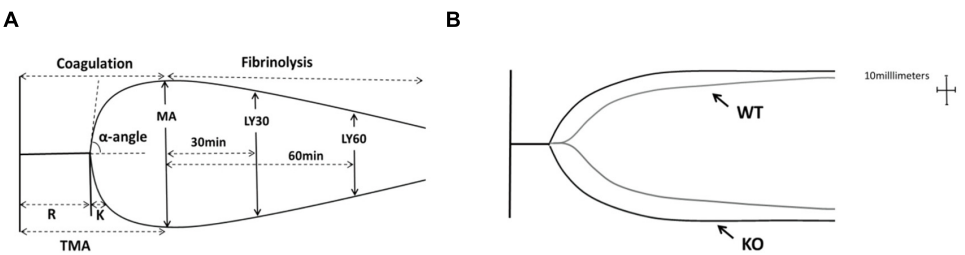

FIGURE 5 | The impact of P311 knock-out on murine blood coagulation. **(A)** A schematic TEG tracing. The parameters are described in **Table 1**. **(B)** Quantitation of the neo-epidermal length (n = 6). Representative TEG tracing from a P311 WT-burn mouse (gray line) and P311 KO-burn mouse (black line). The details of the result of TGE analysis is showed in **Table 2**.

of coagulation. Altogether, the experiment demonstrated that P311 knockout significantly impacts burn-induced coagulopathy, suggesting a potential target for therapy.

DISCUSSION

In this study, bioinformatic and experimental approaches were combined to predict and validate the function of P311. Firstly, by applying OD2 to two reconstructed human PPI datasets, we predicted the function of P311 in inflammatory responses, cell proliferation and coagulation. The principle of OD2 is that OCG prepares the protein lists from multifunctional protein relevant overlapping clusters, for functional enrichment analysis by DAVID. Similar functional enrichments, which occur simultaneously when analyzing two human PPI datasets, were chosen to be the predicted functions. Finally, we conducted relevant biological experiments to confirm these functions of P311.

Integrating Initial P311 PPI Network With the Human Interactome Strengthens the Functional Landscape of P311

As proteins tend to interact with each other when they are involved in the same molecular complex, pathway, or biological process, the understanding of protein function is intrinsically tied to the understanding of this network (Hao et al., 2016). These networks are functional units of protein–protein interaction (PPI) networks and allow function prediction when involving unidentified proteins (Brun et al., 2003; Sharan et al., 2007). DAVID consists of a comprehensive biological knowledgebase

TABLE 3 | Comparison of P311 WT and P311-KO mice after burned and sham injury.

	WT-sham (n = 6)	KO-sham (n−6)	p	WT-burned (n = 6)	KO-burned (n = 6)	p
R (min)	8.850 ± 1.115	8.167 ± 1.291	0.955	8.850 ± 1.541	7.733 ± 1.210	0.57
Angle, a (degrees)	6.075 ± 1.656	6.275 ± 1.819	0.876	9.075 ± 3.143	19.833 ± 2.079	0.002
MA (mm)	16.775 ± 3.132	17.100 ± 3.110	0.888	18.525 ± 3.648	34.200 ± 7.233	0.049
G(*k*)	377.750 ± 155.663	405.650 ± 135.393	0.796	1197.875 ± 742.662	3170.033 ± 582.418	0.005
LY30 (%)	0 ± 0	0 ± 0	NS	0 ± 0	0 ± 0	NS
LY60 (%)	0 ± 0	0 ± 0	NS	0 ± 0	0 ± 0	NS

and analytical tools designed to systematically extract biological meaning from a large gene/protein list (Huang et al., 2008). To decipher the unknown function of P311, we performed functional annotation and enrichment analysis of constituent proteins in the initial P311-containing network, using DAVID. The binary analysis result showed that P311 might have the function to regulate endopeptidase activity, which was not observed in our follow-up experiments. This implied that the strategy to analyze the PPI network, which consists of binary interactions, has its own limitations, as it did not include the extended functional information hiding in the whole human PPI network, which is modular and consists of groups of highly related proteins in the same cellular function (Hartwell et al., 1999; Brun et al., 2003). Katsogiannou et al. (2014) reported that merging the initial PPI network with human interactome can enhance the functional information. Following their strategy, we merged the initial P311-containing network identified by our group, with two existing human PPI datasets, respectively (**Figure 1**).

Regarding a PPI network as a simple graph, in which vertices correspond to proteins and edges to direct physical interactions, allows the graph partition method to identify the networks (Newman, 2006). OCG is a clustering method used to identify the networks containing proteins involved in the same molecular complex, pathway, or biological process, and multifunctional proteins were identified at the intersection of the overlapping networks. Moreover, OCG had a better trade-off between sensitivity and specificity than the CFfinder and Link communities did (Becker et al., 2012). By applying OD (OCG + DAVID) to PPI networks, we found that the result of analyzing different datasets by OD may be a little different, as the datasets were built at a different time by different researchers, which may cause them to not all have the same data in the dataset. Reversely, this may demonstrate the reliability of the result. To improve the confidence of the bioinformatic analysis, we chose the common predicted functions occurring on both sides to be the predicted function. So the bioinformatic system was developed into OD2 (OCG + DAVID + 2 human PPI datasets) with the principle that OCG prepared the protein lists from multifunctional protein relevant overlapping clusters for functional enrichment analysis by DAVID. The similar functional enrichments, which occurred simultaneously when analyzing two human PPI datasets, were chosen to be the predicted functions (**Figure 1**).

Through OD2, we predicted that P311 was involved in one well-known function-positive regulation of cell migration, which supported a reliable prediction function of the system. It has been reported that P311 accelerated cell migration mainly by enhancing the activity of GTPase. Mechanisms, involving P311 in enhancing the activity of GTPase, were cell specific. In epidermal stem cells, it enhanced the activity of Rho A and Rac1 (Yao et al., 2017). While in myofibroblasts, it enhanced the activity of Ral A (Shi et al., 2006). Other known functions of P311 like the regulation of blood pressure homeostasis (Badri et al., 2013), the regulation of development of fibrosis (Yao et al., 2015; Cheng et al., 2017), the regulation of wounds (Yao et al., 2017), angiogenesis (Wang et al., 2017) and so on are shown in **Supplementary Material 8**. P311–TGF-β axis signaling was demonstrated to be important signaling that is related to the processes above. Moreover, mechanisms involving P311 in the expression of TGF-β were also found to be cell-specific. In NIH3T3 fibroblasts, P311 downregulated the expression of TGF-β1 and TGF-β2 binds to the LAP (latency associated protein) (Paliwal et al., 2004). In vascular smooth muscle cells, P311 binds eukaryotic translation initiation factor 3 subunit b (eIF3b) to promote the translation of the transforming growth factor β1–3 (TGF-β 1–3) (Badri et al., 2013; Yue et al., 2014). In EpSCs, P311 stimulated TGFβ1 expression by promoting TGFβ1 promoter methylation and by activating the TGFβ1 5′/3′UTR (Li et al., 2016). However, eIF6, an interacting protein of P311, downregulated the expression of TGFβ1 via H2A.Z occupancy and Sp1 recruitment in fibroblasts (Yang et al., 2015). Additionally, 13 predicted functions, in which P311 had never been implicated before, provided new insight into the function of P311 (**Table 2**).

Predicted Functions Are Confirmed by Biological Experiments

With the proper biological experiments, we verified the predicted functions preliminarily. We confirmed the role of P311 in an inflammatory response, which was related to chemotaxis and immune responses (**Table 2** and **Supplementary Materials 9, 10**), by checking the recruitment of inflammatory cells in the wound healing model. Compared to the P311 WT mice, the number of inflammatory cells was much lower in P311 KO mice, while the mRNA level of CD14 and CD16 was also much lower in P311 KO mice (**Figure 3**). It has been proven that P311 could regulate wound healing by accelerating reepithelization (Yao et al., 2017) or by affecting angiogenesis (Wang et al., 2017). It therefore implies that P311 might also regulate wound healing by affecting the inflammation response during wound healing.

The role of P311 in proliferation was identified by a proliferation assay and flow cytometry analysis of the cell cycle. The results showed that P311 promoted MPFs proliferation by enhancing cells into the S phase (**Figure 4**). The expression of P311 was also found in small-cell lung carcinoma (SCLC), large-cell neuroendocrine carcinoma (LCNEC) (Jones et al., 2004) and glioblastoma (Mariani et al., 2001), whose cells had a really high proliferation capability, which indirectly implied that P311 might regulate the proliferation of cells. This all indicated that P311 might regulate the proliferation of cells. However, further studies are required to validate and elucidate the underlying mechanisms.

Moreover, we also verified a role of P311 in coagulation. As we had observed the difference in the formation of scabs between P311 WT and P311 KO mice when we made the mice wound healing model (Wang et al., 2017), the bioinformatic analysis predicted the function of coagulation. These strongly implied a function of P311 in coagulation. We utilized Thromboelastography (TEG) to monitor the coagulation profile of blood after the burn injury. The result displayed that Angle, MA and G in the P311 KO-burn group were all higher than those in the P311 WT-burn group, which preliminarily verified that P311 might regulate coagulation. Further studies are needed to validate and elucidate the underlying mechanisms, which may suggest P311 as a potential therapeutic target for coagulation related diseases.

However, there were still several predicted functions, which were confirmed to not be associated with P311, through biological experiments. This indicated again that the bioinformatic analysis provides clues but does not validate it as truth. The strategy reported by our group can provide us with specific clues for follow-up biological experiments. Our study preliminarily found that P311 could be involved in inflammatory response, cell proliferation and coagulation. Further studies are required to validate and elucidate the underlying mechanism.

ETHICS STATEMENT

All protocols involving animals were reviewed and approved by the Southwestern Hospital Institutional Review Board.

AUTHOR CONTRIBUTIONS

HL and GL made substantial contributions to the conception and design of the work. SW and XZ performed the majority of the experiments. RZ, YW, CS, and FH contributed to the collection, analysis and interpretation of the data for this study. SW wrote the first draft of the manuscript. YL, WH, and GL revised and edited the manuscript critically with important intellectual contributions. GL was responsible for obtaining funds. All authors read and approved the manuscript.

FUNDING

This work was supported by grants from China's NSFC grants program (81630055).

ACKNOWLEDGMENTS

We wish to thank Prof. Gregory A. Taylor from the Duke University Medical Center for kindly providing the P311 KO mice.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00109/full#supplementary-material

REFERENCES

Aittokallio, T., and Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.* 7, 243–255. doi: 10.1093/bib/bbl022

Badri, K. R., Yue, M., Carretero, O. A., Aramgam, S. L., Cao, J., Sharkady, S., et al. (2013). Blood pressure homeostasis is maintained by a P311-TGF-beta axis. *J. Clin. Invest.* 123, 4502–4512. doi: 10.1172/jci69884

Becker, E., Robisson, B., Chapple, C. E., Guénoche, A., and Brun, C. (2012). Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics* 28, 84–90. doi: 10.1093/bioinformatics/btr621

Bolliger, D., Seeberger, M. D., and Tanaka, K. A. (2012). Principles and practice of thromboelastography in clinical coagulation management and transfusion practice. *Trans. Med. Rev.* 26, 1–13. doi: 10.1016/j.tmrv.2011.07.005

Bossi, A., and Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* 5:260. doi: 10.1038/msb.2009.17

Branco, B. C., Inaba, K., Ives, C., Okoye, O., Shulman, I., David, J.-S., et al. (2014). Thromboelastogram evaluation of the impact of hypercoagulability in trauma patients. *Shock* 41, 200–207. doi: 10.1097/SHK.0000000000000109

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guénoche, A., and Jacq, B. (2003). Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol.* 5:R6. doi: 10.1186/gb-2003-5-1-r6

Cheng, T., Yue, M., Aslam, M. N., Wang, X., Shekhawat, G., Varani, J., et al. (2017). Neuronal protein 3.1 deficiency leads to reduced cutaneous scar collagen deposition and tensile strength due to impaired transforming growth factor-beta1 to -beta3 translation. *Am. J. Pathol.* 187, 292–303. doi: 10.1016/j.ajpath.2016.10.004

Curreri, P. W., Wilterdink, M. E., and Baxter, C. R. (1975). Coagulation dynamics following thermal injury: effect of heparin and protamine sulfate. *Ann. Surg.* 181:161.

Eming, S. A., Krieg, T., and Davidson, J. M. (2007). Inflammation in wound repair: molecular and cellular mechanisms. *J. Invest. Dermatol.* 127, 514–525. doi: 10.1038/sj.jid.5700701

Fujitani, M., Yamagishi, S., Che, Y. H., Hata, K., Kubo, T., Ino, H., et al. (2004). P311 accelerates nerve regeneration of the axotomized facial nerve. *J. Neurochem.* 91, 737–744. doi: 10.1111/j.1471-4159.2004.02738.x

Hao, T., Peng, W., Wang, Q., Wang, B., and Sun, J. (2016). Reconstruction and application of protein–protein interaction network. *Int. J. Mol. Sci.* 17:E907. doi: 10.3390/ijms17060907

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402(6761 Suppl.), C47–C52. doi: 10.1038/35011540

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Jones, M. H., Virtanen, C., Honjoh, D., Miyoshi, T., Satoh, Y., Okumura, S., et al. (2004). Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *Lancet* 363, 775–781. doi: 10.1016/S0140-6736(04)15693-6

Katsogiannou, M., Andrieu, C., Baylot, V., Baudot, A., Dusetti, N. J., Gayet, O., et al. (2014). The functional landscape of Hsp27 reveals new cellular processes such as DNA repair and alternative splicing and proposes novel anticancer targets. *Mol. Cell. Proteom.* 13, 3585–3601. doi: 10.1074/mcp.M114.041228

Li, H., Yao, Z., He, W., Gao, H., Bai, Y., Yang, S., et al. (2016). P311 induces the transdifferentiation of epidermal stem cells to myofibroblast-like cells by stimulating transforming growth factor beta1 expression. *Stem Cell Res. Ther.* 7:175. doi: 10.1186/s13287-016-0421-1

Mariani, L., McDonough, W. S., Hoelzinger, D. B., Beaudry, C., Kaczmarek, E., Coons, S. W., et al. (2001). Identification and validation of P311 as a glioblastoma invasion gene using laser capture microdissection. *Cancer Res.* 61, 4190–4196.

McDonough, W. S., Tran, N. L., and Berens, M. E. (2005). Regulation of glioma cell migration by serine-phosphorylated P311. *Neoplasia* 7, 862–872. doi: 10.1593/neo.05190

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Paliwal, S., Shi, J., Dhru, U., Zhou, Y., and Schuger, L. (2004). P311 binds to the latency associated protein and downregulates the expression of TGF-beta1 and TGF-beta2. *Biochem. Biophys. Res. Commun.* 315, 1104–1109. doi: 10.1016/j.bbrc.2004.01.171

Pan, D., Zhe, X., Jakkaraju, S., Taylor, G. A., and Schuger, L. (2002). P311 induces a TGF-beta1-independent, nonfibrogenic myofibroblast phenotype. *J. Clin. Invest.* 110, 1349–1358. doi: 10.1172/jci15614

Pekelharing, J., Furck, A., Banya, W., Macrae, D., and Davidson, S. (2014). Comparison between thromboelastography and conventional coagulation tests after cardiopulmonary bypass surgery in the paediatric intensive care unit. *Int. J. Lab. Hematol.* 36, 465–471. doi: 10.1111/ijlh.12171

Peng, X., Yuan, S., Tan, J., Ma, B., Bian, X., Xu, C., et al. (2012). Identification of ITGB4BP as a new interaction protein of P311. *Life Sci.* 90, 585–590. doi: 10.1016/j.lfs.2012.02.008

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol. Syst. Biol.* 3:88. doi: 10.1038/msb4100129

Shi, J., Badri, K. R., Choudhury, R., and Schuger, L. (2006). P311-induced myofibroblasts exhibit ameboid-like migration through RalA activation. *Exp. Cell Res.* 312, 3432–3442. doi: 10.1016/j.yexcr.2006.07.016

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2010). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675

Sommer, T., and Wolf, D. H. (2014). The ubiquitin-proteasome-system. *Biochim. Biophys. Acta* 1843:1. doi: 10.1016/j.bbamcr.2013.09.009

Stradiot, L., Mannaerts, I., and van Grunsven, L. A. (2018). P311, friend, or foe of tissue fibrosis? *Front. Pharmacol.* 9:1151. doi: 10.3389/fphar.2018.01151

Studler, J. M., Glowinski, J., and Lévi-Strauss, M. (1993). An abundant mRNA of the embryonic brain persists at a high level in cerebellum, hippocampus and olfactory bulb during adulthood. *Eur. J. Neurosci.* 5, 614–623. doi: 10.1111/j.1460-9568.1993.tb00527.x

Sun, Y. G., Gao, Y. J., Zhao, Z. Q., Huang, B., Yin, J., Taylor, G. A., et al. (2008). Involvement of P311 in the affective, but not in the sensory component of pain. *Mol. Pain* 4:23. doi: 10.1186/1744-8069-4-23

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Tan, J., Peng, X., Luo, G., Ma, B., Cao, C., He, W., et al. (2010). Investigating the role of P311 in the hypertrophic scar. *PLoS One* 5:e9995. doi: 10.1371/journal.pone.0009995

Taylor, G. A., Rodriguiz, R. M., Greene, R. I., Daniell, X., Henry, S. C., Crooks, K. R., et al. (2008). Behavioral characterization of P311 knockout mice. *Genes Brain Behav.* 7, 786–795. doi: 10.1111/j.1601-183X.2008.00420.x

Varani, J., Schuger, L., Dame, M. K., Leonard, C., Fligiel, S. E., Kang, S., et al. (2004). Reduced fibroblast interaction with intact collagen as a mechanism for depressed collagen synthesis in photodamaged skin. *J. Invest. Dermatol.* 122, 1471–1479. doi: 10.1111/j.0022-202X.2004.22614.x

Varshavsky, A. (2014). Discovery of the biology of the ubiquitin system. *JAMA* 311, 1969–1970. doi: 10.1001/jama.2014.5549

Wang, S., Yan, C., Zhang, X., Shi, D., Chi, L., Luo, G., et al. (2018). Antimicrobial peptide modification enhances the gene delivery and bactericidal efficiency of gold nanoparticles for accelerating diabetic wound healing. *Biomater. Sci.* 6, 2757–2772. doi: 10.1039/c8bm00807h

Wang, S., Zhang, X., Qian, W., Zhou, D., Yu, X., Zhan, R., et al. (2017). P311 deficiency leads to attenuated angiogenesis in cutaneous wound healing. *Front. Physiol.* 8:1004. doi: 10.3389/fphys.2017.01004

Wu, J., Ma, B., Yi, S., Wang, Z., He, W., Luo, G., et al. (2004). Gene expression of early hypertrophic scar tissue screened by means of cDNA microarrays. *J. Trauma* 57, 1276–1286. doi: 10.1097/01.TA.0000108997.49513.DC

Yang, S. S., Tan, J. L., Liu, D. S., Loreni, F., Peng, X., Yang, Q. Q., et al. (2015). Eukaryotic initiation factor 6 modulates myofibroblast differentiation at transforming growth factor-beta1 transcription level via H2A.Z occupancy and Sp1 recruitment. *J. Cell Sci.* 128, 3977–3989. doi: 10.1242/jcs.174870

Yao, Z., Li, H., He, W., Yang, S., Zhang, X., Zhan, R., et al. (2017). P311 accelerates skin wound reepithelialization by promoting epidermal stem cell migration through RhoA and Rac1 activation. *Stem Cells Dev.* 26, 451–460. doi: 10.1089/scd.2016.0249

Yao, Z., Yang, S., He, W., Li, L., Xu, R., Zhang, X., et al. (2015). P311 promotes renal fibrosis via TGFbeta1/Smad signaling. *Sci. Rep.* 5:17032. doi: 10.1038/srep17032

Yue, M. M., Lv, K., Meredith, S. C., Martindale, J. L., Gorospe, M., and Schuger, L. (2014). Novel RNA-binding protein P311 binds eukaryotic translation initiation factor 3 subunit b (eIF3b) to promote translation of transforming growth factor beta1-3 (TGF-beta1-3). *J. Biol. Chem.* 289, 33971–33983. doi: 10.1074/jbc.M114.609495

Zhao, L., Leung, J. K., Yamamoto, H., Goswami, S., Kheradmand, F., and Vu, T. H. (2006). Identification of P311 as a potential gene regulating alveolar generation. *Am. J. Respir. Cell Mol. Biol.* 35, 48–54. doi: 10.1165/rcmb.2005-0475OC

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer LZ declared a shared affiliation, with no collaboration, with several of the authors, SW, XZ, FH, YL, RZ, YW, WH, HL, and GL, to the handling Editor at the time of review.

frontiers in Genetics

CORRECTION
published: 06 November 2019
doi: 10.3389/fgene.2019.00818

Corrigendum: Reconstruction and Functional Annotation of P311 Protein–Protein Interaction Network Reveals Its New Functions

Song Wang[1], Xiaorong Zhang[1], Fen Hao[1], Yan Li[2], Chao Sun[3], Rixing Zhan[1], Ying Wang[1], Weifeng He[1], Haisheng Li[1,4] and Gaoxing Luo[1]**

[1] Institute of Burn Research, State Key Laboratory of Trauma, Burn and Combined Injury, Southwest Hospital, Third Military Medical University, Chongqing, China, [2] Laboratory Center of Southwest Hospital, Third Military Medical University, Chongqing, China, [3] The Sixth Resignation Cadre Sanatorium of Shandong Province Military Region, Qingdao, China, [4] The 324th Hospital of Chinese People's Liberation Army, Chongqing, China

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

***Correspondence:**
Haisheng Li
lee58427@163.com
Gaoxing Luo
logxw@yahoo.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 31 July 2019
Accepted: 07 August 2019
Published: 06 November 2019

Citation:
Wang S, Zhang X, Hao F, Li Y, Sun C, Zhan R, Wang Y, He W, Li H and Luo G (2019) Corrigendum: Reconstruction and Functional Annotation of P311 Protein–Protein Interaction Network Reveals Its New Functions. Front. Genet. 10:818. doi: 10.3389/fgene.2019.00818

Keywords: P311, protein–protein interaction networks, inflammatory response, cell proliferation, coagulation

A Corrigendum on:

Reconstruction and Functional Annotation of P311 Protein–Protein Interaction Network Reveals Its New Functions

By Wang S, Zhang X, Hao F, Li Y, Sun C, Zhan R, Wang Y, He W, Li H and Luo G (2019) Front. Genet. 10:109. doi: 10.3389/fgene.2019.00109

There is an error in the ***Funding*** statement. The correct number for "China's NSFC grants program" is "81630055." A correction has therefore been made to the ***Funding*** statement and should read:

"This work was supported by grants from China's NSFC grants program (81630055)."

Additionally, in the original article, the reference for "Becker et al., 2011" was incorrectly written as "Becker, E., Robisson, B., Chapple, C. E., Guénoche, A., and Brun, C. (2011). Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics 28, 84–90. doi: 10.1093/bioinformatics/btr621". It should be "Becker, E., Robisson, B., Chapple, C. E., Guénoche, A., and Brun, C. (2012). Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics 28, 84–90. doi: 10.1093/bioinformatics/btr621".

Lastly, the citation of Becker's work was missing in the **Supplementary Material 4**. Corrections have been made by placing a copy of the code from Becker et al., Boinformatics, 2012 Jan 1;28(1):84–90, in the supplementary material along with an edited README file. Thus, corrections have also been made to the *Materials* and *Methods*, subsection *OCG (Overlapping Cluster Generator) Algorithm*, by removing the last sentence, which was redundant and to the first paragraph:

"The OCG algorithm was carried out by the software available in (Becker et al., 2012) (**Supplementary Material 4**)."

The authors apologize for these errors and state that they do not change the scientific conclusions of the article in any way. The original article has been updated.

ACKNOWLEDGMENTS

We wish to thank Prof. Gregory A. Taylor from the Duke University Medical Center for kindly providing the P311 KOmice.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00818/full#supplementary-material

frontiers in Genetics

METHODS
published: 02 April 2019
doi: 10.3389/fgene.2019.00262

FeatSNP: An Interactive Database for Brain-Specific Epigenetic Annotation of Human SNPs

Chun-yu Ma[1], Pamela Madden[2], Paul Gontarz[1], Ting Wang[3] and Bo Zhang[1]**

[1] Center of Regenerative Medicine, Department of Developmental Biology, Washington University School of Medicine, St. Louis, MO, United States, [2] Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, United States, [3] Department of Genetics, The Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine, St. Louis, MO, United States

OPEN ACCESS

Edited by:
Shandar Ahmad,
Jawaharlal Nehru University, India

Reviewed by:
Wei-Hua Chen,
Huazhong University of Science and Technology, China
Sandeep Kumar Dhanda,
La Jolla Institute for Immunology (LJI), United States

***Correspondence:**
Ting Wang
twang@wustl.edu
Bo Zhang
bzhang29@wustl.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 19 November 2018
Accepted: 08 March 2019
Published: 02 April 2019

Citation:
Ma C-y, Madden P, Gontarz P, Wang T and Zhang B (2019) FeatSNP: An Interactive Database for Brain-Specific Epigenetic Annotation of Human SNPs. Front. Genet. 10:262. doi: 10.3389/fgene.2019.00262

FeatSNP is an online tool and a curated database for exploring 81 million common SNPs' potential functional impact on the human brain. FeatSNP uses the brain transcriptomes of the human population to improve functional annotation of human SNPs by integrating transcription factor binding prediction, public eQTL information, and brain specific epigenetic landscape, as well as information of Topologically Associating Domains (TADs). FeatSNP supports both single and batched SNP searching, and its interactive user interface enables users to explore the functional annotations and generate publication-quality visualization results. FeatSNP is freely available on the internet at FeatSNP.org with all major web browsers supported.

Keywords: SNP, database, epigenetics, brain, transcription factor

INTRODUCTION

Genome-wide association studies (GWAS) and expression quantitative trait loci (eQTL) analyses have identified thousands of genetic variants that are associated with a wide range of human phenotypes, shedding lights on the understanding of the genetic effect to human diseases. However, a key challenge for scientists in the human genetics community is to understand the molecular mechanism connecting significant genetic variant and specific phenotype. More than 90% of SNPs associated with human phenotypes are located in non-protein-coding regions, and cannot be explained by alteration of amino acid sequence of proteins (Welter et al., 2014). Recently, mounting evidence suggests that disease-associated non-coding SNPs are highly enriched in tissue-specific regulatory elements including enhancers, which can be detected and defined by specific chromatin modifications (Carey et al., 2015;

Zhou et al., 2015; Agrawal et al., 2018). Moreover, some non-coding SNPs are found to be located within transcription factor (TF) binding motifs, which affect the TF binding affinity and result in allele switching and/or allele-specific regulation of target genes (Andersson et al., 2014; Roadmap Epigenomics et al., 2015; Nelson et al., 2016). These evidences underscore the potential causal role of non-coding genetic variants in affecting human diseases and phenotypes through regulation of gene expression (Claussnitzer et al., 2015).

Here we introduce FeatSNP, an online tool and database which provides an interactive user interface (UI) for inquiring brain-specific functional and epigenetic annotation of human SNPs. Unlike traditional SNP functional annotation databases, such as RegulomeDB (Boyle et al., 2012) and HaploReg (Ward and Kellis, 2012), FeatSNP focuses on the collection and curation of brain-specific functional genomics data, including epigenomes, transcriptomes, and eQTL data, to better annotate the regulatory potential of single SNP. Specifically, FeatSNP supplies a series of new features to facilitate research understanding the functional annotation of SNP on human brain (**Supplementary Table S1**). FeatSNP uses human brain transcriptomes to improve and refine the prediction of allele-specific TF binding motifs. The expression correlation between SNP-associated gene and predicted SNP-associated TFs was used to determine the best allele-associated TF candidate. The interactive UI allows the users easily to browse functional annotation and generate analysis results and high quality figures.

METHODS

FeatSNP consists of a front end UI implemented with HTML/PHP/JavaScript, and a backend NoSQL database implemented with MongoDB (v3.2.7) as shown in **Figure 1**. The current SNP dataset contains 81,144,876 bi-allelic SNPs from dbSNP (V144), with SNP accession number as unique identifier in the database. Human dbSNP build 144 was downloaded from ftp.ncbi.nih.gov/snp, which includes 84,435,229 SNPs records, 1,591,294 insertions records, 2,595,517 deletions records, 33,234 indel records, and 110 Multiple Nucleotide Polymorphisms (MNPs) records. After filtering redundant records, 81,144,876 of 84,435,229 biallelic SNPs were used to generate functional annotations and were curated by the FeatSNP database. The genome coordinates (hg19) of 81,144,876 SNPs were used to associate the SNPs with their nearest genes based on 56,642 records of GENOCDE gene annotation Release 19 (GRCh37.p13).

To predict impact of allele-specific TF binding affinity by SNPs, the Position Weight Matrix (PWM) of 519 vertebrate TFs were collected from JASPAR (Core Vertebrate 2016) (Mathelier et al., 2016). After evaluating the motif weight PWM of 519 TFs at base-pair resolution (**Supplementary Figure S2**), the reference and alternate alleles for every SNP with flanking 10 bp of genomic sequences both upstream and downstream were obtained from the UCSC Genome Browser. FIMO (Grant et al., 2011) was used to scan the 21 bp sequence to identify binding motifs matching any of the 519 TF PWMs, and calculate the TFBS motif scores. Only instances where a motif in the sequence (i) passed the threshold of $P < 1e\text{-}2$ in either the reference or the alternate allele, and (ii) contained the SNP location and (iii) the difference of motif scores between the reference and the alternate allele was greater than 2, were recorded in the database.

1,259 transcriptome datasets of 13 brain tissues generated by the GTEx consortium (Gibson, 2015) were used to calculate the Pearson correlation between each SNP associated gene and predicted binding TFs. The lowly expressed gene and TFs (expression of all samples in one tissue less than 0.2RPKM) were removed. The correlation and gene expression in 13 brain tissues were visualized by using JavaScript package Highcharts (v5.0.2).

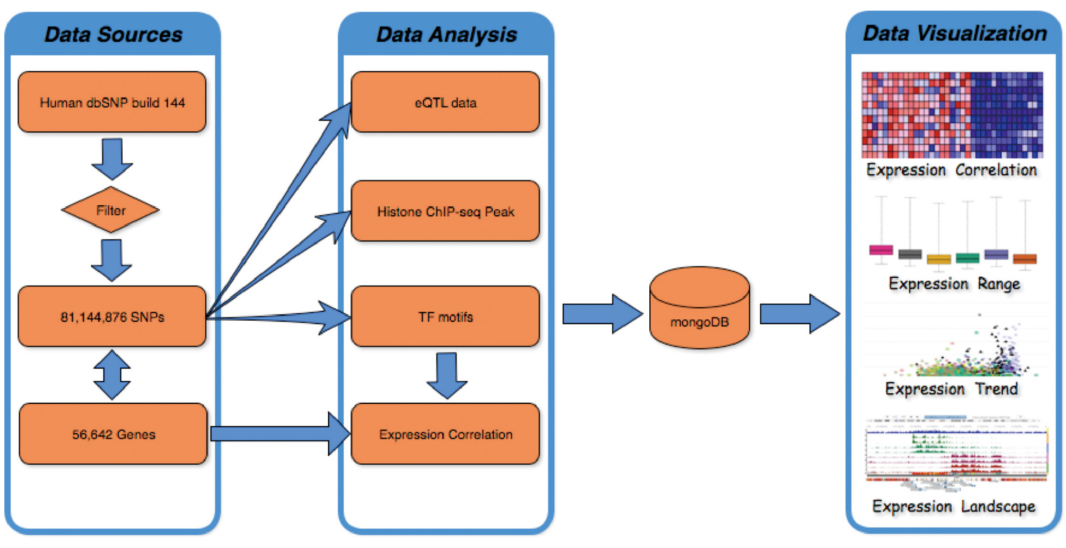

FIGURE 1 | Analysis flowchart and infrastructure of FeatSNP.

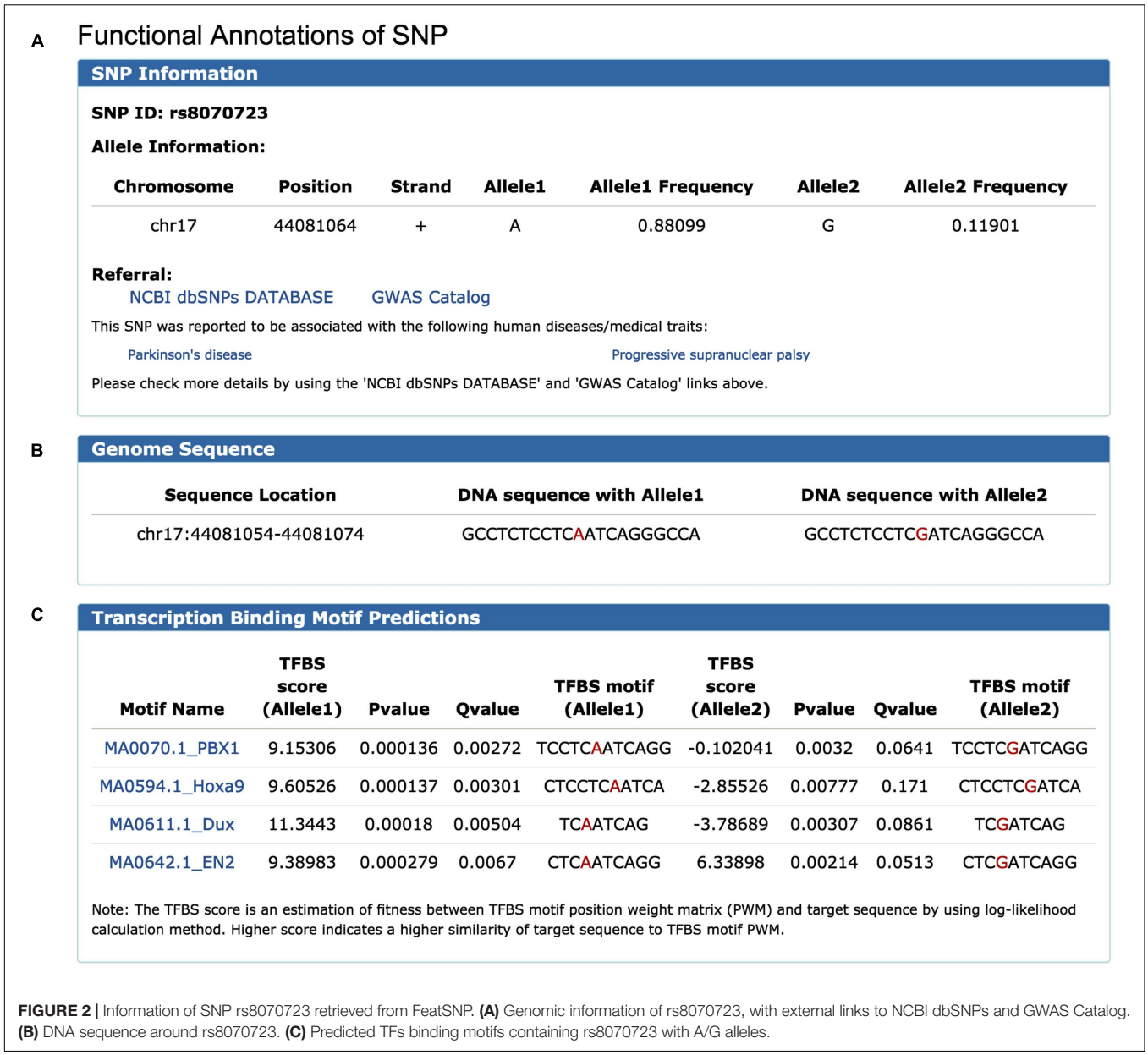

FIGURE 2 | Information of SNP rs8070723 retrieved from FeatSNP. **(A)** Genomic information of rs8070723, with external links to NCBI dbSNPs and GWAS Catalog. **(B)** DNA sequence around rs8070723. **(C)** Predicted TFs binding motifs containing rs8070723 with A/G alleles.

eQTL data of 10 brain tissues generated by GTEx consortium were negative-log10 transformed and further visualized by using Highcharts (v5.0.2).

Histone modification ChIP-seq data of 10 brain tissues were downloaded from NIH Roadmap Epigenomics data portal. Bedtools was used to identify SNPs residing in peaks of 7 histone modification marks (H3K4me3, H3K36me3, H3K27me3, H3K4me1, H3K27ac, H3K9me3, and H3K9ac) that were identified by macs2 (Zhang et al., 2008) with default parameters. To enhance the user experience, the WashU epigenome browser (Zhou et al., 2015) was embedded in the UI to display epigenetic landscape in a 200 bp region surrounding each SNP. The browser also displays DNA methylation data (Whole Genome Bisulfite Sequencing) of 4 neuronal progenitor and brain tissues generated by Roadmap Epigenomics Project, enhanced epilogos visualization[1] of all 127 epigenomes, and topologically associating domains (TAD) data of GM12878, IMR90, and Hap1 cell lines (Rao et al., 2014; Sanborn et al., 2015). eQTL data of 10 brain tissues generated by GTEx consortium were also visualized on the embedded WashU epigenome browser.

The association records of SNP and human disease/traits (V1.0.2) were downloaded from GWAS Catalog. 33,894 associations with *p*-value smaller then 5E-8 were kept and classified based on 1,374 human disease/traits categories. The functional annotations of these 33,894 SNPs were

[1] epilogos.altiusinstitute.org

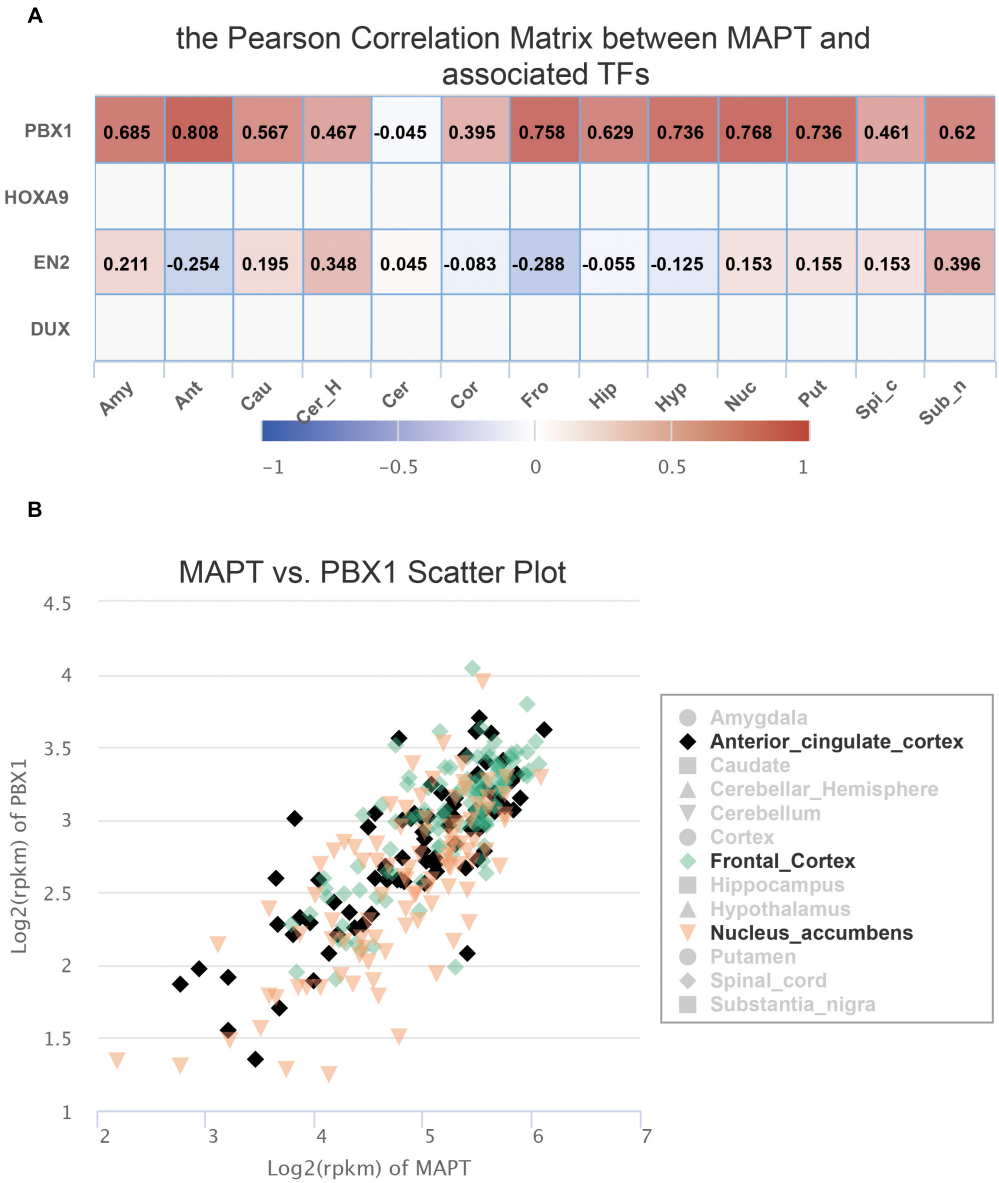

FIGURE 3 | Expression information of SNP rs8070723 tagged gene retrieved from FeatSNP. **(A)**. Pearson Correlation Matrix between rs8070723 tagged MAPT and potential bound TFs. **(B)**. Scatter plot of gene expression of MAPT and PBX1 in anterior cingulate cortex, frontal cortex, and nucleus accumbens.

reported on FeatSNP.org/html_file/disease_classification.html (**Supplementary Figure S3**).

RESULTS

To illustrate the use of FeatSNP, we performed the analysis using rs8070723 as an example. rs8070723 is an intronic A/G SNP (major allele A frequency 0.881, minor allele G frequency 0.119) in *MAPT*, the gene that encodes the microtubule-associated protein tau, and is associated with Progressive Supranuclear Palsy (Hoglinger et al., 2011) and with Parkinson's Disease (UK Parkinson's Disease Consortium et al., 2011). To better understand the regulatory potential of this human disease-associated SNP, we inquired the epigenetic annotation of rs8070723 in FeatSNP through Single SNP ID Searching function on SNP Query Page (**Supplementary Figure S4**). The database first reported the basic information of SNP rs8070723, including genomic location, allelic frequency, surrounding DNA sequence, and associated gene (**Figures 2A–C**). Users can further access the genetic information and associated human disease or traits of inquired SNPs on dbSNP and GWAS Catalog through external links.

FeatSNP found four potential TF binding motifs harboring rs8070723 with A allele, including *PBX1*, *Hoxa9*, *Dux*, and *EN2*. All four TF binding motifs had high TFBS scores in A allele,

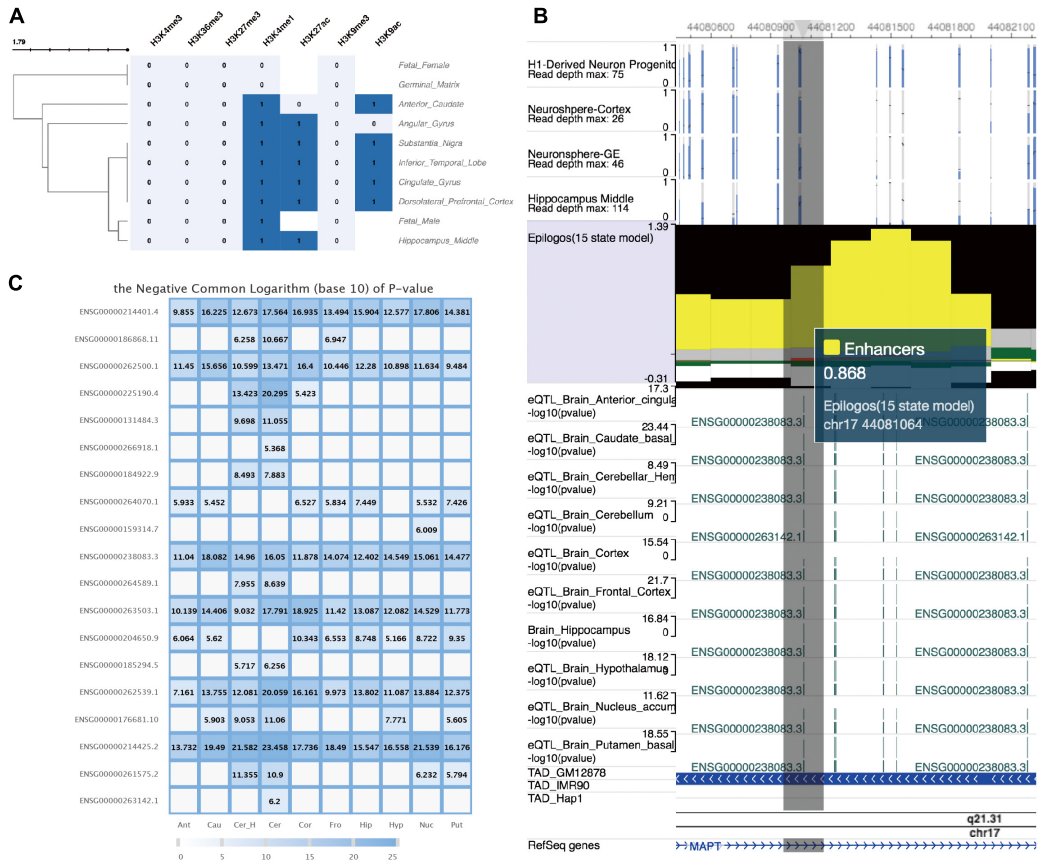

FIGURE 4 | Epigenetic annotation and eQTL information of SNP rs8070723 tagged genomic loci retrieved from FeatSNP. **(A)** Clustering visualization of epigenetic annotation to the genomic loci tagged by SNP rs8070723. **(B)** WashU EpiGenome Brower view of genomic loci tagged by rs8070723. Top: DNA methylation level of CpG sites in four neuronal cells. Middle track: Epilogos visualization of chromHMM predicted chromatin status. Followed eQTL tracks: log10 transformed *p*-value of eQTL in 10 brain regions. TAD track: Topological Associated Domain tracks of GM12878, IMR90, and Hap1. Bottom: RefSeq gene annotation track. **(C)** Complete eQTL information of SNP rs8070723 in 10 brain regions.

and the TFBS motifs were destroyed with G allele with low TFBS scores (**Figure 2C**). *PBX1* encodes a nuclear protein that belongs to the *PBX* homeobox family of transcriptional factors, and studies suggested *PBX1* regulates the patterning of the cerebral cortex (Golonzhka et al., 2015) and its transcriptional network controls dopaminergic neuron development in Parkinson's disease (Villaescusa et al., 2016). *EN2* encodes homeodomain-containing proteins and has been implicated in the control of pattern formation during development of the central nervous system (Genestine et al., 2015). *Hoxa9* is an important homeobox transcription factor and plays important roles in myeloid leukemogenesis (Siriboonpiputtana et al., 2017). Dux-family transcription factors were recently identified to regulate zygotic genome activation in placental mammals (De Iaco et al., 2017). Thus, *PBX1* and *EN2* could be the potential master TFs affected by the SNP rs8070723.

Since FeatSNP curated 1,259 transcriptome data of 13 brain tissues generated by the GTEx consortium (Gibson, 2015), we were able to further check the expression level of *PBX1* and *EN2* in multiple brain regions in FeatSNP database. *EN2* was only expressed in the cerebellum of the brain (**Supplementary Figure S1A**) and did not correlate with expression level of *MAPT* (**Figure 3A**). We found that PBX1 highly expressed in different brain regions (**Supplementary Figure S1B**), and we also found the expression of *MAPT* had strong and specific correlation with *PBX1* in multiple brain regions (**Figure 3A**), especially in anterior cingulate cortex (r = 0.808), nucleus accumbens (r = 0.768), and frontal cortex (r = 0.768) (**Figure 3B**), which were considered as major affected regions of Progressive Supranuclear Palsy (Salmon et al., 1997).

We further explored the epigenetic annotation of the genomic regions tagged by rs8070723 in 10 brain regions by using epigenome data generated from Roadmap Consortium, which were also curated in FeatSNP database. We found the regions tagged by SNP rs8070723 enriched for strong active histone modification signals including H3K4me1, H3K9ac, and H3K27ac in 8 brain tissues (**Figure 4A**). Such active histone modifications were generally associated with active enhancer and promoter functions. Chromatin epigenetic status prediction based on chromHMM (Ernst and Kellis, 2012) suggested that the regions tagged by SNP rs8070723 could be considered

as strong enhancers (**Figure 4B**). Finally, we explored the eQTL data in 13 brain tissues, and found rs8070723 was associated with several genes' expression, including *MAPT* (**Figures 4B,C**). *MAPT* gene mutations have been associated with several neurodegenerative disorders such as Alzheimer's disease and Parkinson's disease. Our result suggests that rs8070723 G allele might influence *MAPT* expression level by reducing the binding affinity of upstream regulatory protein *PBX1*, therefore providing a mechanistic association with neurodegenerative diseases including Progressive Supranuclear Palsy and Parkinson's Disease.

CONCLUSION

In summary, FeatSNP is an interactive database providing brain-specific functional genomics resources to investigate the regulatory potential of human SNPs. This database provides a multitude types of functional annotations, including TF binding motif prediction, epigenetic landscape, expression correlation and eQTL information. We anticipate that this database will facilitate scientists to investigate the functional impact of their candidate genetic variants in a more streamlined, rapid, and efficient fashion.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: http://www.roadmapepigenomics.org/.

AUTHOR CONTRIBUTIONS

C-yM and BZ performed the data analysis, C-yM and PG developed the database and website. PM, TW, and BZ designed and supervised the study.

FUNDING

This work was supported by National Institutes of Health grant DA027995, HG007175, HG007354, and Goldman Sachs Philanthropy Fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00262/full#supplementary-material

REFERENCES

Agrawal, A., Chou, Y. L., Carey, C. E., Baranger, D. A. A., Zhang, B., Sherva, R., et al. (2018). Nabis dependence. *Mol. Psychiatry* 23, 1293–1302. doi: 10.1111/den.13366

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461. doi: 10.1038/nature12787

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1797. doi: 10.1101/gr.137323.112

Carey, C. E., Agrawal, A., Zhang, B., Conley, E. D., Degenhardt, L., Heath, A. C., et al. (2015). Monoacylglycerol lipase (MGLL) polymorphism rs604300 interacts with childhood adversity to predict cannabis dependence symptoms and amygdala habituation: evidence from an endocannabinoid system-level analysis. *J. Abnorm. Psychol.* 124, 860–877. doi: 10.1037/abn0000079

Claussnitzer, M., Dankel, S. N., Kim, K. H., Quon, G., Meuleman, W., Haugen, C., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* 373, 895–907. doi: 10.1056/NEJMoa1502214

De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., and Trono, D. (2017). DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* 49, 941–945. doi: 10.1038/ng.3858

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216. doi: 10.1038/nmeth.1906

Genestine, M., Lin, L., Durens, M., Yan, Y., Jiang, Y., Prem, S., et al. (2015). Engrailed-2 (En2) deletion produces multiple neurodevelopmental defects in monoamine systems, forebrain structures and neurogenesis and behavior. *Hum. Mol. Genet.* 24, 5805–5827. doi: 10.1093/hmg/ddv301

Gibson, G. (2015). Human genetics. GTEx detects genetic effects. *Science* 348, 640–641. doi: 10.1126/science.aab3002

Golonzhka, O., Nord, A., Tang, P. L. F., Lindtner, S., Ypsilanti, A. R., Ferretti, E., et al. (2015). Pbx regulates patterning of the cerebral cortex in progenitors and postmitotic neurons. *Neuron* 88, 1192–1207. doi: 10.1016/j.neuron.2015.10.045

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi: 10.1093/bioinformatics/btr064

Hoglinger, G. U., Melhem, N. M., Dickson, D. W., Sleiman, P. M., Wang, L. S., Klei, L., et al. (2011). Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* 43, 699–705. doi: 10.1038/ng.859

Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C. Y., Denay, G., Lee, J., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 44, D110–D115. doi: 10.1093/nar/gkv1176

Nelson, E. C., Agrawal, A., Heath, A. C., Bogdan, R., Sherva, R., Zhang, B., et al. (2016). Evidence of CNIH3 involvement in opioid dependence. *Mol. Psychiatry* 21, 608–614. doi: 10.1038/mp.2015.102

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., and Robinson, J. T. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680. doi: 10.1016/j.cell.2014.11.021

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Salmon, E., Van der Linden, M. V., and Franck, G. (1997). Anterior cingulate and motor network metabolic impairment in progressive supranuclear palsy. *Neuroimage* 5, 173–178. doi: 10.1006/nimg.1997.0262

Sanborn, A. L., Rao, S. S., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. I., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6456–E6465. doi: 10.1073/pnas.1518552112

Siriboonpiputtana, T., Zeisig, B. B., Zarowiecki, M., Fung, T. K., Mallardo, M., Tsai, C. T., et al. (2017). Transcriptional memory of cells of origin overrides beta-catenin requirement of MLL cancer stem cells. *EMBO J.* 36, 3139–3155. doi: 10.15252/embj.201797994

UK Parkinson's Disease Consortium, Wellcome Trust Case Control Consortium, Spencer, C. C., Plagnol, V., Strange, A., Gardner, M., et al. (2011). Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Hum. Mol. Genet.* 20, 345–353. doi: 10.1093/hmg/ddq469

Villaescusa, J. C., Li, B., Toledo, E. M., Rivetti di Val Cervo, P., Yang, S., Stott, S. R., et al. (2016). A PBX1 transcriptional network controls dopaminergic neuron development and is impaired in Parkinson's disease. *EMBO J.* 35, 1963–1978. doi: 10.15252/embj.201593725

Ward, L. D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934. doi: 10.1093/nar/gkr917

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137. doi: 10.1186/gb-2008-9-9-r137

Zhou, X., Li, D., Zhang, B., Lowdon, R. F., Rockweiler, N. B., Sears, R. L., et al. (2015). Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nat. Biotechnol.* 33, 345–346. doi: 10.1038/nbt.3158

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 September 2019
doi: 10.3389/fgene.2019.00864

Tensor Decomposition-Based Unsupervised Feature Extraction Applied to Single-Cell Gene Expression Analysis

Y-h. Taguchi[1]* *and Turki Turki*[2]

[1] Department of Physics, Chuo University, Tokyo, Japan, [2] Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia

OPEN ACCESS

Edited by:
Michael Fernandez, The Vancouver Prostate Centre, Canada

Reviewed by:
Yunierkis Perez, University of the Americas, Ecuador
Jose Ignacio Abreu Salas, Catholic University of the Most Holy Conception, Chile

***Correspondence:**
Y-h. Taguchi
tag@granular.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *26 June 2019*
Accepted: *19 August 2019*
Published: *19 September 2019*

Citation:
Taguchi Y-h and Turki T (2019) Tensor Decomposition-Based Unsupervised Feature Extraction Applied to Single-Cell Gene Expression Analysis. Front. Genet. 10:864. doi: 10.3389/fgene.2019.00864

Although single-cell RNA sequencing (scRNA-seq) technology is newly invented and a promising one, but because of lack of enough information that labels individual cells, it is hard to interpret the obtained gene expression of each cell. Because of insufficient information available, unsupervised clustering, for example, *t*-distributed stochastic neighbor embedding and uniform manifold approximation and projection, is usually employed to obtain low-dimensional embedding that can help to understand cell–cell relationship. One possible drawback of this strategy is that the outcome is highly dependent upon genes selected for the usage of clustering. In order to fulfill this requirement, there are many methods that performed unsupervised gene selection. In this study, a tensor decomposition (TD)-based unsupervised feature extraction (FE) was applied to the integration of two scRNA-seq expression profiles that measure human and mouse midbrain development. TD-based unsupervised FE could select not only coincident genes between human and mouse but also biologically reliable genes. Coincidence between two species as well as biological reliability of selected genes is increased compared with that using principal component analysis (PCA)-based FE applied to the same data set in the previous study. Since PCA-based unsupervised FE outperformed the other three popular unsupervised gene selection methods, highly variable genes, bimodal genes, and dpFeature, TD-based unsupervised FE can do so as well. In addition to this, 10 transcription factors (TFs) that might regulate selected genes and might contribute to midbrain development were identified. These 10 TFs, BHLHE40, EGR1, GABPA, IRF3, PPARG, REST, RFX5, STAT3, TCF7L2, and ZBTB33, were previously reported to be related to brain functions and diseases. TD-based unsupervised FE is a promising method to integrate two scRNA-seq profiles effectively.

Keywords: tensor decomposition, enrichment analysis, single-cell RNA-sequencing, midbrain development, inter-species analysis

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) (Sasagawa et al., 2019) is a newly invented technology that enables us to measure the amount of RNA in a single-cell basis. In spite of its promising potential, it is not easy to interpret the measurements. The primary reason of this difficulty is the lack of sufficient information that characterizes individual cells. In contrast to the huge number of cells measured, which is often as many as several thousands, the number of labeling is limited, for example, measurement of conditions as well as the amount of expression of key genes measured by fluorescence-activated cell sorting, whose number is typically as little as tens. This prevents us from selecting genes that characterize the individual cell properties.

In order to deal with samples without suitable numbers of labeling, unsupervised method is frequently used, since it does not make use of labeling information directly. *K*-means clustering and hierarchical clustering are popular methodologies that are often applied to gene expression analysis. The popular clustering methods specifically applied to scRNA-seq are *t*-distributed stochastic neighbor embedding (tSNE) (van der Maaten and Hinton, 2008) and uniform manifold approximation and projection (UMAP) (McInnes et al., 2018), which are known to be useful to get low-dimensional embedding of a set of cells. In spite of that, the obtained clusters are highly dependent upon genes used for clustering. Thus, the next issue is, without labeling (i.e., pre-knowledge), to select genes that might be biologically meaningful.

The various unsupervised gene selection methods applicable to scRNA-seq were invented, for example, highly variable genes, bimodal genes, dpFeature, and principal component analysis (PCA)-based unsupervised feature extraction (FE) (Murakami et al., 2012; Taguchi and Okamoto, 2012; Taguchi and Murakami, 2013; Ishida et al., 2014; Kinoshita et al., 2014; Murakami et al., 2014; Taguchi, 2014; Taguchi and Murakami, 2014; Umeyama et al., 2014; Murakami et al., 2015; Taguchi, 2015; Taguchi et al., 2015a; Taguchi et al., 2015b; Taguchi et al., 2015c; Taguchi et al., 2016; Taguchi, 2016a; Taguchi, 2016b; Taguchi, 2016c; Taguchi, 2016d; Taguchi and Wang, 2017; Taguchi et al., 2017; Taguchi, 2017d; Taguchi and Wang, 2018a; Taguchi, 2018a; Taguchi and Wang, 2018b). Chen et al. (2018) recently compared genes selected by these methods and concluded that the genes selected are very diverse and have their own (unique) biological features. In this sense, it is required to invent more advanced unsupervised gene selection methods that can select more biologically relevant genes.

In this paper, we propose the application of tensor decomposition (TD)-based unsupervised FE (Taguchi, 2017a; Taguchi, 2017b; Taguchi, 2017c; Taguchi, 2017e; Taguchi, 2017f; Taguchi and Ng, 2018; Taguchi, 2018b; Taguchi, 2018c; Taguchi, 2019a). It is an advanced method of PCA-based unsupervised FE for scRNA-seq analysis. For more details about PCA-based unsupervised FE and TD-based unsupervised FE, see the recently published book (Taguchi, 2019b). Especially focusing on the integration of two scRNA-seq profiles, the advantages of TD-based unsupervised FE when compared with PCA-based unsupervised FE are as follows: The former can integrate more than two gene expressions prior to the analysis, while the latter can only integrate the results obtained by applying the method to individual data sets.

In the following, based on the previous study (Taguchi, 2018a) where PCA-based unsupervised FE was employed, we try to integrate human and mouse midbrain development gene expression profiles to obtain key genes that contribute to this process, by applying TD-based unsupervised FE. It turned out that TD-based unsupervised FE can identify biologically more relevant and more common genes between human and mouse than can PCA-based unsupervised FE that outperformed other compared methods.

METHODS AND MATERIALS

scRNA-seq Data

Midbrain Development of Humans and Mice

The first scRNA-seq data used in this study were downloaded from Gene Expression Omnibus (GEO) under the GEO ID GSE76381; the files named "GSE76381_EmbryoMoleculeCounts.cef.txt.gz" (for human) and "SE76381_MouseEmbryoMoleculeCounts.cef.txt.gz" (for mouse) were downloaded. These two gene expression profiles were generated from scRNA-seq data set: One represents human embryo ventral midbrain cells between 6 and 11 weeks of gestation (287 cells for 6 weeks, 131 cells for 7 weeks, 331 cells for 8 weeks, 322 cells for 9 weeks, 509 cells for 10 weeks, and 397 cells for 11 weeks, for a total of 1,977 cells). Another is a set of mouse ventral midbrain cells at six developmental stages between E11.5 and E18.5 (349 cells for E11.5, 350 cells for E12.5, 345 cells for E13.5, 308 cells for E14.5, 356 cells for E15.5, 142 cells for E18.5, and 57 cells for unknown, for a total of 1,907 cells).

Mouse Hypothalamus With and Without Acute Formalin Stress

The second scRNA-seq data used in this study were downloaded from GEO under GEOID GSE74672; the file named "GSE74672_expressed_mols_with_classes.xlsx.gz" was downloaded. It is generated from scRNA-seq data set that measures mouse hypothalamus with and without acute formalin stress. Various meta-data, which are included in the first 11 rows of the data set, are available. The meta-data available include sex, age, cell types [astrocytes, endothelial, ependymal, microglia, neurons, oligos, and vascular smooth muscle (VSM)], control vs stressed samples, and so on.

TD-Based Unsupervised FE

Midbrain Development of Humans and Mice

TD-based unsupervised FE is a recently proposed method successfully applied to various biological problems. TD-based unsupervised FE can be used for integration of multiple measurements applied to the common set of genes. Suppose $x_{ij} \in \mathbb{R}^{N\times M}$ and $x_{ik} \in \mathbb{R}^{N\times K}$ are the ith expression of the jth and kth cells under the two distinct conditions (in the present study, they are human and mouse), respectively. Then the three-mode tensor, $x_{ijk} \in \mathbb{R}^{N\times M\times K}$, where N (= 13,889) is total number of common genes between human and mouse, which share gene symbols,

M (= 1,977) is the number of human cells, and K (= 1,907) is total number of mouse cells, is defined as

$$x_{ijk} = x_{ij} \cdot x_{ik}. \tag{1}$$

It is Case II Type I tensor (Taguchi, 2017e). Since it is too large to be decomposed, it is further transformed into Type II tensor, as follows:

$$x_{jk} = \sum_{i=1}^{N} x_{ijk}, \tag{2}$$

where $x_{jk} \in \mathbb{R}^{M \times K}$ is now not a tensor but a matrix. In this case, TD is equivalent to singular value decomposition (SVD). After applying SVD to x_{jk}, we get SVD,

$$x_{jk} = \sum_{\ell=1}^{\min(M,K)} \lambda_\ell u_{\ell j} v_{\ell k}, \tag{3}$$

where $u_{\ell j} \in \mathbb{R}^{M \times M}$ and $v_{\ell k} \in \mathbb{R}^{K \times K}$ are singular value vectors attributed to cells of human scRNA-seq and those of mouse scRNA-seq, respectively. Here, Case II means that tensor is generated such that two matrices share the genes, while Type II means that summation is taken over as in Eq. (2). On the other hand, the tensor before taking summation as in Eq. (1) is Type I.

Singular value vectors attributed to genes of human and mouse scRNA-seq, $u_{\ell i} \in \mathbb{R}^{N \times M}$ and $v_{\ell i} \in \mathbb{R}^{N \times K}$ are defined as respectively.

$$u_{\ell i} = \sum_{j=1}^{M} u_{\ell j} x_{ij}, \tag{4}$$

$$v_{\ell i} = \sum_{k=1}^{K} v_{\ell k} x_{ik}, \tag{5}$$

In order to find genes associated with biological functions, we need to select $u_{\ell j}$ and $v_{\ell k}$ which are coincident with biological meaning. In this study, we employ time points of measurements as biological meanings. In other words, we seek for genes associated with time development. Since we would like to find any kind of time dependence, we simply deal with time points as un-ordered labeling. Thus, we apply categorical regression

$$u_{\ell j} = a_\ell + \sum_{t=1}^{T} a_{\ell t} \delta_{jt}, \tag{6}$$

(T = 6; t = 1 to T, which correspond to 6, 7, 8, 9, 10, and 11 weeks; see *Methods and Materials*) or

$$v_{\ell k} = b_\ell + \sum_{t=1}^{T} b_{\ell t} \delta_{kt}, \tag{7}$$

(T = 7; t = 1 to T, which correspond to E11.5, E12.5, E13.5, E14.5, F15.5, E18.5, and unknown; see *Methods and Materials*), where $\delta_{jt}\left(\delta_{kt}\right) = 1$ when the jth (kth) cell is taken from the tth time point otherwise $\delta_{jt}\left(\delta_{kt}\right) = 0$. a_ℓ, $a_{\ell t}$, b_ℓ and $b_{\ell t}$ are the regression coefficients.

P-values are attributed to ℓth singular value vectors using the above categorical regression [`lm` function in R (R Core Team, 2018) is used to compute P-values]. P-values attributed to singular value vectors are corrected by Benjamini-Hochberg (BH) criterion (Benjamini and Hochberg, 1995). Singular value vectors associated with corrected P-values of less than 0.01 are selected for the download analysis. Hereafter, the set of selected singular value vectors of human and mouse is denoted as $\Omega_\ell^{\text{human}}$ and $\Omega_\ell^{\text{mouse}}$, respectively.

P-values are attributed to genes with assuming χ^2 distribution for the gene singular value vectors, $u_{\ell i}$ and $v_{\ell i}$, corresponding to the cell singular value vectors selected by categorical regression as

$$P_i^{\text{human}} = P_{\chi^2}\left[> \sum_{\ell \in \Omega_\ell^{\text{human}}} \left(\frac{u_{\ell i} - \left\langle u_{\ell i} \right\rangle}{\sigma_\ell^{\text{human}}} \right)^2 \right] \tag{8}$$

for human genes and

$$P_i^{\text{mouse}} = P_{\chi^2}\left[> \sum_{\ell \in \Omega_\ell^{\text{mouse}}} \left(\frac{v_{\ell i} - \left\langle v_{\ell i} \right\rangle}{\sigma_\ell^{\text{mouse}}} \right)^2 \right] \tag{9}$$

for mouse genes, respectively. Here,

$$\left\langle u_{\ell i} \right\rangle = \frac{1}{N} \sum_{i=1}^{N} u_{\ell i} \tag{10}$$

and

$$\left\langle v_{\ell i} \right\rangle = \frac{1}{N} \sum_{i=1}^{N} v_{\ell i}. \tag{11}$$

$\sigma_\ell^{\text{human}}$ and $\sigma_\ell^{\text{mouse}}$ are the standard deviations of ℓth gene singular value vectors for human and mouse, respectively, $\Omega_\ell^{\text{human}}$ and $\Omega_\ell^{\text{mouse}}$ are sets of ℓs, selected by categorical regression for human [Eq. (6)] and mouse [Eq.(7)], respectively. $P_{\chi^2}[> x]$ is the cumulative probability of χ^2 distribution when the argument takes values larger than x. P_i^{human} and P_i^{mouse} are corrected by BH criterion, and genes associated with corrected P-values of less than 0.01 are selected.

Mouse Hypothalamus With and Without Acute Formalin Stress

The application of TD-based unsupervised FE to mouse hypothalamus is quite similar to that of mouse and human midbrain. There are also two matrices, $x_{ij} \in \mathbb{R}^{N \times M}$ and $x_{ik} \in \mathbb{R}^{N \times K}$ which correspond to the ith expression of the jth and kth

cells under the two distinct conditions (in the present case, they are without and with acute formalin stress, respectively); N=24,341,M=1,785 and K=1,096. Case II Type II tensor, x_{jk}, was also generated using Eqs. (1) and (2), and SVD was applied to x_{jk} as Eq. (3). Then singular value vectors attributed to genes of samples without and with acute formalin stress, $u_{\ell i}$ and $v_{\ell i}$, were computed by Eqs. (4) and (5). We also applied categorical regressions to $u_{\ell i}$. and $v_{\ell i}$, although categories considered here are not time points but cell types. Then categorical regressions applied to $u_{\ell i}$ and $v_{\ell i}$ in mouse hypothalamus without and with acute formalin stress are

$$u_{\ell j} = a_\ell + \sum_{s=1}^{7} a_{\ell s}\delta_{js}, \tag{12}$$

$$v_{\ell k} = b_\ell + \sum_{s=1}^{7} b_{\ell s}\delta_{ks}, \tag{13}$$

where s stands for one of seven cell types mentioned in Methods and Materials and $\delta_{js}\left(\delta_{ks}\right)=1$ when the jth (kth) cell is taken from the sth cell types otherwise $\delta_{js}\left(\delta_{ks}\right)=0$. **Table 1** lists the number of cells in these categories. The remaining procedures to select genes associated with identified cell type dependency are exactly the same as those in midbrain case.

Enrichment Analyses

Various enrichment analysis methods are performed with separate uploading selected human and mouse gene symbols, or genes selected commonly between samples without and samples with acute formalin stress, to Enrichr (Kuleshov et al., 2016).

RESULTS

Midbrain Development of Humans and Mice

As a result, following the procedure described in the *Methods and Materials*, we identified 55 and 44 singular value vectors attributed to cells, $u_{\ell j}$s and $v_{\ell k}$s for human and mouse, respectively. One possible validation of selected $u_{\ell j}$s and $v_{\ell k}$s is coincidence. Although cells measured are not related between human and mouse at all, if SVD works well, corresponding singular value vectors (i.e., $u_{\ell j}$ and $v_{\ell k}$ sharing the same ℓs) attributed to cells should share something biological, for example, time dependence. This suggests that it is more likely that corresponding singular value vectors attributed to cells, $u_{\ell j}$ and $v_{\ell k}$, are simultaneously associated with significant P-values computed by categorical regression. As expected, they are highly significantly correlated. **Table 2** shows confusion matrix of the coincidence of selected singular value vectors between human and mouse. For human cells, only the top 1,907 singular value vectors among all 1,977 singular value vectors are considered, since the total number of singular value vectors attributed to mouse cells is 1,907.

Figure 1 shows the coincidence of selected singular value vectors between human and mouse. Singular value vectors with smaller ℓs, that is, with more contributions, are more likely selected and coincident between human and mouse. This can be the side evidence that guarantees that TD-based unsupervised FE successfully integrated human and mouse scRNA-seq data.

Next, we selected genes with following the procedures described in *Methods and Materials*. (The list of genes is available as **Supplementary Data Sheet 1** and **2**). The first validation of selected genes is the coincidence between human and mouse. In Taguchi's previous study (Taguchi, 2018a), more number of common genes were selected by PCA-based unsupervised FE than other methods compared, that is, highly variables genes, bimodal genes, and dpFeature. **Table 3** shows the confusion matrix that describes the coincidence of selected genes between human and mouse. Odds ratio is as large as 133, and P-value is 0 (i.e., less than numerical accuracy), which is significantly better than coincidence of selected genes between human and mouse (53 common genes between 116 genes selected for human and 118 genes selected mouse), previously achieved by PCA-based unsupervised FE (Taguchi, 2018a), which outperformed other methods, that is, highly variable genes, bimodal genes, and dpFeature.

On the other hand, most of the genes selected by PCA-based unsupervised FE in the previous study (Taguchi, 2018a) are included in the genes selected by TD-based unsupervised FE in the present study. One hundred two genes are selected by TD-based unsupervised FE among 116 human genes selected by PCA-based unsupervised FE in the previous study (Taguchi, 2018a),

TABLE 1 | The number of cells that belong to either without or with acute formalin stress or cell types.

Cell types	Without	With
	Acute stress	
Astrocytes	135	132
Endothelial	169	71
Ependymal	211	145
Microglia	34	14
Neurons	628	270
Oligos	570	431
VSM	38	33

VSM, vascular smooth muscle.

TABLE 2 | Confusion matrix of coincidence between selected 55 singular value vectors selected among all 1,977 singular value vectors, $u_{\ell j}$, attributed to human cells and 44 singular value vectors selected among all 1907 singular value vectors, $v_{\ell o}$, attributed to mouse cells.

		Human	
		Not selected	Selected
Mouse	Not selected	1,833	12
	Selected	23	32

Selected: corrected P-values, computed with regression analysis [Eqs. (6) and (7)], are less than 0.01. Not selected: otherwise. Odds ratio is as many as 227, and P-values computed by Fisher's exact test are 1.44×10^{-44}.

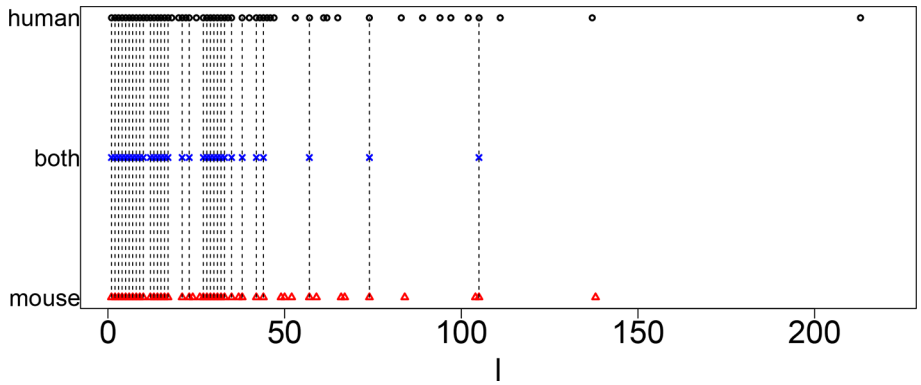

FIGURE 1 | Coincidence between singular value vectors shown in **Table 2**. Horizontal axis: singular value vector numbering ℓ. Black open circles, ℓ selected for human; blue crosses, ℓ selected for both human and mouse; red open triangles, ℓ selected for mouse. Vertical black broken lines connect ℓ selected for both human and mouse.

TABLE 3 | Confusion matrix of coincidence between selected 456 genes for human and selected 505 genes for mouse among all 13,384 common genes.

		Human	
		Not selected	**Selected**
Mouse	Not selected	13,233	151
	Selected	200	305

Selected: corrected P-values, computed with χ^2 distribution [Eqs. (8) and (9)], are less than 0.01. Not selected: otherwise. Odds ratio is as many as 133, and P-values computed by Fisher's exact test are 0 (i.e., less than numerical accuracy).

while 91 genes are selected by TD-based unsupervised FE among 118 mouse genes by PCA-based unsupervised FE. Thus, TD-based unsupervised FE is quite consistent with PCA-based unsupervised FE.

Biological significance tested by enrichment analysis is further enhanced (Full list of enrichment analysis is available as **Supplementary Tables 1** and **2**). Most remarkable advance achieved by TD-based unsupervised FE is "Allen Brain Atlas," to which only downregulated genes were enriched in the previous study (Taguchi, 2018a). As can be seen in **Table 4**, now much enrichment is associated with upregulated genes. In addition to this, most of the five top-ranked terms are related to paraventricular nucleus, which is adjusted to midbrain. This suggests that TD-based unsupervised FE successfully identified genes related to midbrain.

In addition to this, "Jensen TISSUES" (**Table 5**) for Embryonic_brain is highly enhanced [i.e., more significant (smaller), with P-values $\sim 10^{-100}$ which were as large as 10^{-10} to 10^{-20} in the previous study (Taguchi, 2018a)]. On the other hand, "ARCHS4 tissues" also strongly supports the biological reliability of selected genes (**Table 6**). The term "MIDBRAIN" is enriched highly, and it is top ranked for both human and mouse.

There is some brain-related enrichment found in other categories, although it is not strong enough compared with that of the top three. Brain-related terms in "GTEx Tissue Sample Gene Expression Profiles up" (**Table 7**) are also enhanced for mouse brain (top three terms are brain), although no brain terms are enriched within five top-ranked terms for human (this discrepancy cannot be understood at the moment). On the contrary, brain-related terms in "MGI Mammalian Phenotype

TABLE 4 | Five top-ranked terms from "Allen Brain Atlas up" by Enrichr for selected 456 human genes and 505 mouse genes.

Human			
Term	**Overlap**	***P*-value**	**Adjusted *P* -value**
Paraventricular hypothalamic nucleus, magnocellular division, medial magnocellular part	31/301	2.68×10^{-12}	2.91×10^{-9}
Paraventricular hypothalamic nucleus, magnocellular division	31/301	2.68×10^{-12}	2.91×10^{-9}
Paraventricular hypothalamic nucleus, magnocellular division, posterior magnocellular part	28/301	3.39×10^{-10}	1.47×10^{-7}
Paraventricular hypothalamic nucleus	29/301	7.02×10^{-11}	5.08×10^{-8}
Paraventricular nucleus, dorsal part	27/301	1.57×10^{-9}	4.88×10^{-7}
Mouse			
Paraventricular hypothalamic nucleus, magnocellular division, medial magnocellular part	31/301	4.03×10^{-11}	2.19×10^{-8}
Paraventricular hypothalamic nucleus, magnocellular division	31/301	4.03×10^{-11}	2.19×10^{-8}
Paraventricular hypothalamic nucleus, magnocellular division, posterior magnocellular part	31/301	4.03×10^{-11}	2.19×10^{-8}
Lower dorsal lateral hypothalamic area	29/301	8.40×10^{-10}	3.65×10^{-7}
Paraventricular hypothalamic nucleus, magnocellular division, posterior magnocellular part, lateral zone	31/301	4.03×10^{-11}	2.19×10^{-8}

TABLE 5 | Enrichment of embryonic brain by "JENSEN TISSUES" in Enrichr.

Term	Overlap	*P* -value	Adjusted *P* -value
Human			
Embryonic_brain	330/4936	3.36×10^{-104}	4.30×10^{-102}
Mouse			
Embryonic_brain	366/4936	3.59×10^{-115}	4.59×10^{-113}

TABLE 6 | Enrichment of embryonic brain by "ARCHS4 Tissues" in Enrichr.

Term	Overlap	*P* -value	Adjusted *P* -value
Human			
MIDBRAIN	248/2316	1.02×10^{-129}	1.11×10^{-127}
Mouse			
MIDBRAIN	248/2316	1.44×10^{-99}	1.56×10^{-97}

2017" (**Table 8**) are enhanced for human brain (fourth and fifth ranks), although no brain terms are enriched within the five top-ranked terms for mouse (this discrepancy also cannot be understood at the moment). The above observations suggest that TD-based unsupervised FE could identify genes related to mouse and human embryonic midbrain.

We also uploaded selected 456 human genes and 505 mouse genes to STRING server (Szklarczyk et al., 2014), which evaluates protein–protein interaction (PPI) enrichment. Among 456 human genes, 7,488 PPI are reported, while the expected number of PPI is as small as 3,524 (*P*-value is less than 1×10^{-6}). Among 505 mouse genes, 6,788 PPI are reported, while the expected number of PPI is as small as 3,290 (*P*-value is again less than 1×10^{-6}). Thus, TD-based unsupervised FE can successfully identify significantly interacting protein-coding genes.

Finally, we checked if transcription factors (TFs) that target selected genes are common between human and mouse (**Table 9**). These TFs are associated with adjusted *P*-values of less than 0.01 in "ENCODE and ChEA Consensus TFs from ChIP-X" of Enrichr. They are highly overlapped between human and mouse

TABLE 7 | Five top-ranked terms from "GTEx Tissue Sample Gene Expression Profiles up" by Enrichr for selected 456 human genes and 505 mouse genes. Brain-related terms are asterisked.

Human			
Term	**Overlap**	***P* -value**	**Adjusted *P* -value**
GTEX-QCQG-1426-SM-48U22_ovary_female_50-59_years	105/1165	3.56×10^{-35}	1.04×10^{-31}
GTEX-RWS6-1026-SM-47JXD_ovary_female_60-69_years	116/1574	7.96×10^{-31}	7.74×10^{-28}
GTEX-TMMY-1726-SM-4DXTD_ovary_female_40-49_years	117/1582	2.97×10^{-31}	4.33×10^{-28}
GTEX-RU72-0008-SM-46MV8_skin_female_50-59_years	94/1103	1.99×10^{-31}	1.45×10^{-26}
GTEX-R55E-0008-SM-48FCG_skin_male_20-29_years	111/1599	3.67×10^{-27}	1.78×10^{-24}
Mouse			
*GX-WVLH-0011-R4A-SM-3MJFS_brain_male_50-59_years	139/1957	1.93×10^{-30}	5.63×10^{-27}
*GX-X261-0011-R8A-SM-4E3I5_brain_male_50-59_years	135/1878	5.24×10^{-30}	7.65×10^{-27}
*GX-T5JC-0011-R4A-SM-32PLT_brain_male_20-29_years	129/1948	3.51×10^{-25}	3.42×10^{-22}
GTEX-R55E-0008-SM-48FCG_skin_male_20-29_years	109/1599	4.93×10^{-22}	2.40×10^{-19}
GTEX-TMMY-1726-SM-4DXTD_ovary_female_40-49_years	107/1582	2.37×10^{-21}	7.69×10^{-19}

TABLE 8 | Five top-ranked terms from "MGI Mammalian Phenotype 2017" by Enrichr for selected 456 human genes and 505 mouse genes. Brain-related terms are asterisked.

Human			
Term	**Overlap**	***P* -value**	**Adjusted *P* -value**
MP:0002169_no_abnormal_phenotype_detected	82/1674	2.52×10^{-11}	5.53×10^{-8}
MP:0001262_decreased_body_weight	63/1189	3.40×10^{-10}	3.72×10^{-7}
MP:0001265_decreased_body_size	46/774	3.20×10^{-9}	$2.33 \times s10^{-6}$
*M0009937_abnormal_neuron_differentiation	15/106	1.81×10^{-8}	9.90×10^{-6}
*M0000788_abnormal_cerebral_cortex_morphology	17/145	3.64×10^{-8}	1.60×10^{-5}
Mouse			
MP:0002169_no_abnormal_phenotype_detected	89/1674	$1.36 \times s10^{-11}$	3.09×10^{-8}
MP:0011091_prenatal_lethality,_complete_penetrance	27/272	1.68×10^{-9}	1.91×10^{-6}
MP:0001262_decreased_body_weight	65/1189	3.93×10^{-9}	2.97×10^{-6}
MP:0011100_preweaning_lethality,_complete_penetrance	42/674	8.55×10^{-8}	3.88×10^{-5}
MP:0001265_decreased_body_size	46/774	8.22×10^{-8}	3.88×10^{-5}

TABLE 9 | TFs enriched in "ENCODE and ChEA Consensus TFs from ChIP-X" by Enrichr for human and mouse. Bold TFs are common.

Human	BCL3, **BHLHE40**, **EGR1**, **GABPA**, **IRF3**, **PPARG**, **REST**, **RFX5**, SP1, SP2, SRF, **STAT3**, **TCF7L2**, TRIM28, TRIM28, **ZBTB33**
Mouse	**BHLHE40**, CTCF, E2F4, E2F6, **EGR1**, ESR1, ETS1, FLI1, **GABPA**, **IRF3**, NFIC, NRF1, **PPARG**, RCOR1, **REST**, **RFX5**, SPI1, **STAT3**, **TCF7L2**, USF1, USF2, YY1, **ZBTB33**, ZNF384

TFs, transcription factors.

(there are 10 common TFs between 16 TFs found in human and 24 TFs found in mouse). Although selected TFs are very distinct from those in the previous study (Taguchi, 2018a), they are highly interrelated with each other (see below). These TFs are uploaded to the regnetworkweb server (Liu et al., 2015), and TF networks shown in **Figure 2** are identified. Clearly, even partially, these TFs interact highly with each other.

We also checked if the 10 commonly selected TFs (in bold in **Table 9**) are related to brains. Lack of BHLHE40 was found to result in brain malfunction (Hamilton et al., 2018). The function of EGR1 was found in embryonic rat brain (Wells et al., 2011). GABPA is essential for human cognition (Reiff et al., 2014). IRF3 is related to brain disease (Schultz et al., 2019). PPAR, which PPARG belongs to, is believed to be the therapeutic target of neurodegenerative diseases (Warden et al., 2016). REST is a master regulator of neurogenesis (Mozzi et al., 2017). RFX5 is known to be expressive in fetal brain (Sugiaman-Trapman et al., 2018). STAT3 promotes brain metastasis (Priego et al., 2018). TCF7L2 regulates brain gene expression (Shao et al., 2013). ZBTB33 affects the mouse behavior through regulating brain gene expression (Kulikov et al., 2016). Thus, all 10 commonly selected TFs are related to brains.

Mouse Hypothalamus With and Without Acute Formalin Stress

Although the effectiveness of the proposed strategy toward scRNA-seq is obvious in the results shown in the previous subsection, one might wonder if it is accidental. In order to dispel such doubts, we apply TD-based unsupervised FE to yet another scRNA-seq data set: mouse hypothalamus with and without acute formalin stress. Contrary to the data set analyzed in the previous subsection where very distant two data sets were analyzed, the data sets analyzed here are very close to each other. Both data sets are taken from the same tissue of mouse, hypothalamus. The only difference is if they are stressed by formalin dope or not. The motivation why we here specifically apply TD-based unsupervised FE to two close data sets is as follows: When two data sets are too close, it might be difficult to identify which genes are commonly altered by additional condition, in this case, the dependence upon cell types, because all genes might behave equally between the two. Thus, it is not a bad idea to check if TD-based unsupervised FE can work well when not only very distant data sets are analyzed but also very close data sets are analyzed.

With following the procedure described in the Materials and Methods, we identified 30 and 24 singular value vectors attributed to cells, $u_{\ell j}$s and $v_{\ell k}$s, without and with acute formalin stress, respectively. We again applied Fisher's exact test (**Table 10**). Although odds ratio is 10 times larger than that in **Table 2**, P -value is even smaller than that in **Table 2**; this suggests that TD-based unsupervised FE could

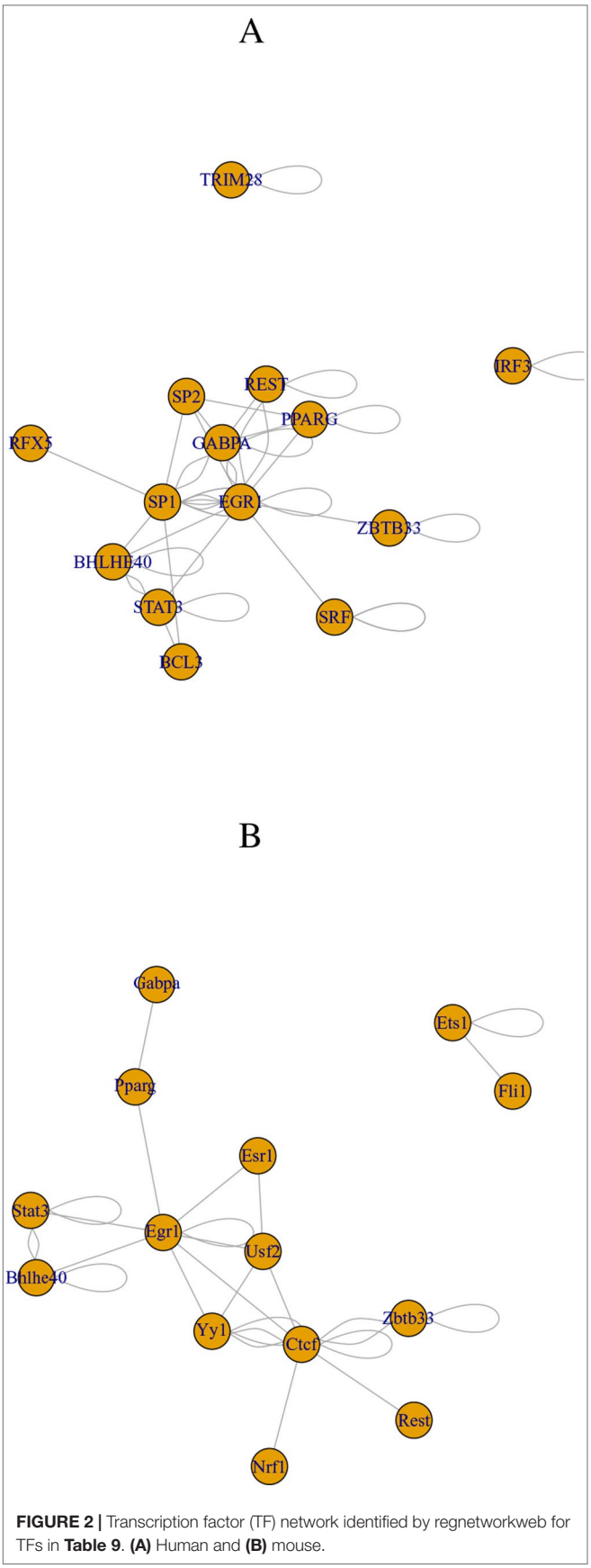

FIGURE 2 | Transcription factor (TF) network identified by regnetworkweb for TFs in **Table 9**. **(A)** Human and **(B)** mouse.

identify not all of genes but only limited genes as being common between two experimental conditions: without and with stress.

Figure 3 shows the coincidence of selected singular value vectors between samples without and with stress. Singular value vectors with smaller ℓs, that is, with more contributions, are more likely selected and coincident between samples without and with stress. This can be the side evidence that guarantees that TD-based unsupervised FE successfully integrated scRNA-seq data taken from samples without and with stress while avoiding to regard that all are coincident between two samples.

Next, we selected genes with following the procedures described in the *Methods and Materials*. The first validation of selected genes is the coincidence between human and mouse. **Table 11** shows the confusion matrix that describes the coincidence of selected genes between samples without and with stress. Odds ratio is as large as 270, and *P*-value is 0 (i.e., less than numerical accuracy). Thus, as expected, TD-based unsupervised FE could not identify all genes but only a limited number of genes associated with cell-type dependence.

Finally, we tried to evaluate if genes selected are tissue type specific, that is, hypothalamus. We have uploaded 3,324 commonly selected genes to Enrichr. "GTEx Tissue Sample Gene Expression Profiles up" suggest that all five top-ranked terms are brain with high significance (**Table 12**, adjusted *P*-values are less than 1×10^{-130}). This suggests that TD-based unsupervised FE successfully identified limited number of genes related to brains even using closely related samples. In order to be more specific, we checked "Allen Brain Atlas up" in Enrichr. Then we found that all five top-ranked terms are hypothalamic (**Table 13**). It is interesting that TD-based unsupervised FE could successfully identify hypothalamus-specific genes by only using scRNA-seq retrieved from hypothalamus. It is usually required to use data taken from other tissues in order to identify tissue-specific genes because we need to compare targeted tissues and not targeted tissues in order to identify genes expressed specifically in target tissues. The successful identification of genes specific to something without using the comparison with other samples was also observed previously during an attempt to identify tumor-specific genes by TD-based unsupervised FE (Taguchi, 2017c). In this sense, TD-based unsupervised FE methods are effective not only when genes common between two distinct conditions are sought but also when genes common between two closely related conditions are sought. Thus, it is unlikely that the success of a TD-based unsupervised method applied to scRNA-seq is accidental.

TABLE 10 | Confusion matrix of coincidence between selected 30 singular value vectors selected among all 1,096 singular value vectors, $u_{\ell j}$,attributed to samples without stress and 24 singular value vectors selected among all 1,096 singular value vectors, $v_{\ell k}$ attributed to samples with stress.

		Not selected	Selected
Without stress	Not selected	1,065	1
	Selected	7	23

For samples without stress, only the top 1,096 singular value vectors among all 1,785 singular value vectors are considered, since total number of singular value vectors attributed to samples without stress is 1,096. Selected: corrected P -values, computed with regression analysis (Eqs. (12) and (13)), are less than 0.01. Not selected: otherwise. Odds ratio is as many as 2,483, and P-values computed by Fisher's exact test are 1.92×10^{-40}.

TABLE 11 | Confusion matrix of coincidence between selected 4,150 genes for samples without stress and selected 3,621 genes for samples with stress among all 24,341 genes.

		With stress	
		Not selected	Selected
Without stress	Not selected	19,894	297
	Selected	826	3,324

Selected: corrected P -values, computed with χ^2s atribution that corresponds to Eqs. (8) and (9) in human and mouse midbrain study, are less than 0.01. Not selected: otherwise. Odds ratio is as many as 270, and P-values computed by Fisher's exact test are 0 (i.e., less than numerical accuracy).

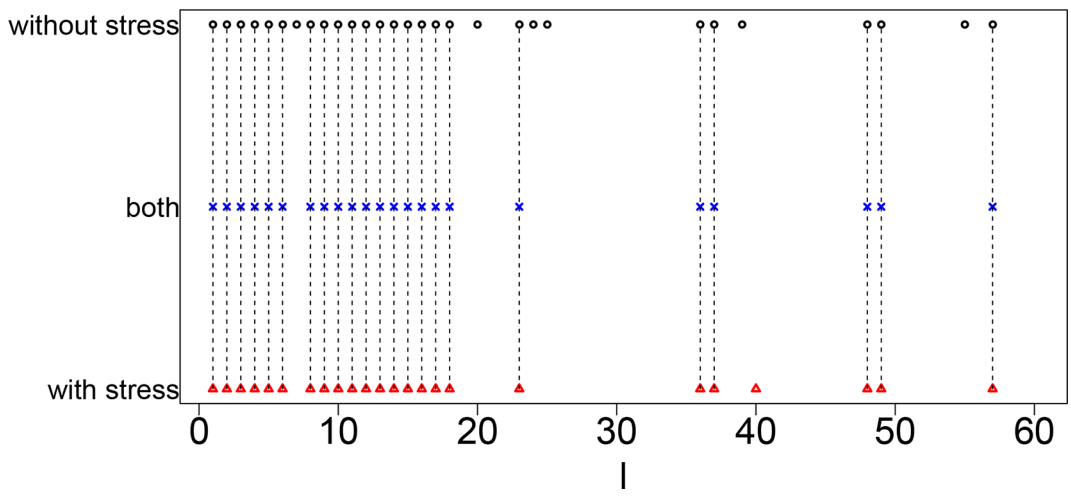

FIGURE 3 | Coincidence between singular value vectors shown in **Table 10**. Horizontal axis: singular value vector numbering ℓ. Black open circles, ℓs selected for samples without stress; blue crosses, ℓs selected for both samples without and with stress; red open triangles, ℓs selected for samples with stress. Vertical black broken lines connect ℓs selected for both samples without and with stress.

TABLE 12 | Five top-ranked terms from "GTEx Tissue Sample Gene Expression Profiles up" by Enrichr for 3,324 genes selected commonly between samples without and with stress.

Term	Overlap	*P* -value	Adjusted *P* -value
GTEX-WWYW-0011-R10A-SM-3NB35_brain_female_50-59_years	1006/2885	2.7880×10^{-151}	8.135×10^{-148}
GTEX-T6MN-0011-R1A-SM-32QOY_brain_male_50-59_years	859/2317	2.9865×10^{-144}	4.3575×10^{-141}
GTEX-QVUS-0011-R3A-SM-3GAFD_brain_female_60-69_years	963/2759	6.8195×10^{-144}	6.6325×10^{-141}
GTEX-T2IS-0011-R3A-SM-32QPB_brain_female_20-29_years	967/2792	5.5265×10^{-142}	4.0315×10^{-139}
GTEX-WZTO-0011-R3B-SM-3NMC6_brain_male_40-49_years	991/2972	2.6805×10^{-133}	1.5645×10^{-130}

TABLE 13 | Five top-ranked terms from "Allen Brain Atlas up" by Enrichr for 3,324 genes selected commonly between samples without and with stress.

Term	Overlap	*P* -value	Adjusted *P* -value
Paraventricular hypothalamic nucleus	120/301	3.38×10^{-22}	7.41×10^{-19}
Paraventricular hypothalamic nucleus, parvicellular division	119/301	1.15×10^{-21}	1.27×10^{-18}
Paraventricular hypothalamic nucleus, parvicellular division, medial parvicellular part, dorsal zone	117/301	1.29×10^{-20}	9.42×10^{-18}
Paraventricular nucleus, cap part	116/301	4.22×10^{-20}	2.31×10^{-17}
Paraventricular hypothalamic nucleus, magnocellular division	115/301	1.36×10^{-19}	5.96×10^{-17}

DISCUSSIONS AND FUTURE WORK

In this study, we applied TD-based unsupervised FE to the integration of scRNA-seq data sets taken from two species: human and mouse. In the sense of identification of biologically more relevant set of genes, TD-based unsupervised FE can outperform PCA-based unsupervised FE that previously (Taguchi, 2018a) could outperform three more popular methods: highly variable genes, bimodal genes, and dpFeature. Thus, it is expected that TD-based unsupervised FE can do so, too.

For the purpose of integration of two scRNA-seq data sets, TD-based unsupervised FE has many advantages than the other four methods, that is, PCA-based unsupervised FE, highly variable genes, bimodal genes, and dpFeature. At first, TD-based unsupervised FE can integrate two scRNA-seq data sets, not after but before the selection of genes. This enabled us to identify more coincident gene sets between two scRNA-seq in this study of human and mouse. As a result, we were able to identify more coincident results between human and mouse.

The criteria of gene selection are quite robust; they should be dependent upon time points when they are measured. We did not have to specify how they are actually correlated with time. It is another advantage of TD-based unsupervised FE.

By applying enrichment analysis to the genes selected, we found many valuable insights about the biological process. As a result, we identified 10 key TFs that might regulate embryonic midbrain developments. All of the 10 selected TFs turned out to be related to brains.

TD-based unsupervised FE turned out to be quite effective to integrate two scRNA-seq data sets. This method should be applied to various scRNA-seq data sets considering broader scope of investigations.

In future work, we plan to (1) utilize the proposed TD-based unsupervised FE under the transfer learning setting; (2) extend the proposed approach to handle the data integration from multiple related tasks; and (3) investigate the performance of the proposed approach when coupled with machine and deep learning algorithms.

DATA AVAILABILITY

The data sets analyzed for this study can be found in the GEO.

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76381.

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74672

AUTHOR CONTRIBUTIONS

Y-HT planned the research, performed analyses, and wrote a paper. TT discussed the results and wrote a paper.

FUNDING

This study was supported by KAKENHI (17K00417 and 19H05270) and Okawa Foundation (grant number 17-10). This project was also funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. KEP-8-611-38. The authors, therefore, acknowledge with thanks DSR for technical and financial support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00864/full#supplementary-material

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Chen, B., Herring, C. A., and Lau, K. S. (2018). pyNVR: investigating factors affecting feature selection from scRNA-seq data for lineage reconstruction. *Bioinformatics* 35, 2335–2337. doi: 10.1093/bioinformatics/bty950

Hamilton, K. A., Wang, Y., Raefsky, S. M., Berkowitz, S., Spangler, R., Suire, C. N., et al. (2018). Mice lacking the transcriptional regulator Bhlhe40 have enhanced neuronal excitability and impaired synaptic plasticity in the hippocampus. *PLoS One* 13, 1–22. doi: 10.1371/journal.pone.0196223

Ishida, S., Umeyama, H., Iwadate, M., and Taguchi, Y.-h. (2014). Bioinformatic screening of autoimmune disease genes and protein structure prediction with FAMS for drug discovery. *Protein Pept. Lett.* 21, 828–839. doi: 10.2174/0929866513209990052

Kinoshita, R., Iwadate, M., Umeyama, H., and Taguchi, Y.-h. (2014). Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as candidate drug targets. *BMC Syst. Biol.* 8 Suppl 1, S4. doi: 10.1186/1752-0509-8-S1-S4

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Kulikov, A. V., Korostina, V. S., Kulikova, E. A., Fursenko, D. V., Akulov, A. E., Moshkin, M. P., et al. (2016). Knockout zbtb33 gene results in an increased locomotion, exploration and pre-pulse inhibition in mice. *Behav. Brain Res. Ser. Test Content1* 297, 76–83. doi: 10.1016/j.bbr.2015.10.003

Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* bav095. doi: 10.1093/database/bav095

McInnes, L., Healy, J., and Melville, J. (2018). Umap: uniform manifold approximation and projection for dimension reduction. *ArXiv* 1802, 03426. doi: 10.21105/joss.00861

Mozzi, A., Guerini, F. R., Forni, D., Costa, A. S., Nemni, R., Baglio, F., et al. (2017). REST, a master regulator of neurogenesis, evolved under strong positive selection in humans and in non human primates. *Scie. Rep.* 7, 9530. doi: 10.1038/s41598-017-10245-w

Murakami, Y., Toyoda, H., Tanahashi, T., Tanaka, J., Kumada, T., Yoshioka, Y., et al. (2012). Comprehensive miRNA expression analysis in peripheral blood can diagnose liver disease. *PLoS One* 7, e48366. doi: 10.1371/journal.pone.0048366

Murakami, Y., Tanahashi, T., Okada, R., Toyoda, H., Kumada, T., Enomoto, M., et al. (2014). Comparison of hepatocellular carcinoma miRNA expression profiling as evaluated by next generation sequencing and microarray. *PLoS One* 9, e106314. doi: 10.1371/journal.pone.0106314

Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., et al. (2015). Comprehensive analysis of transcriptome and metabolome analysis in intrahepatic cholangiocarcinoma and hepatocellular carcinoma. *Sci. Rep.* 5, 16294. doi: 10.1038/srep16294

Priego, N., Zhu, L., Monteiro, C., Mulders, M., Wasilewski, D., Bindeman, W., et al. (2018). STAT3 labels a subpopulation of reactive astrocytes required for brain metastasis. *Nat. Med.* 24, 1024–1035. doi: 10.1038/s41591-018-0044-4

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria R: R Foundation for Statistical Computing.

Reiff, R. E., Ali, B. R., Baron, B., Yu, T. W., Ben-Salem, S., Coulter, M. E., et al. (2014). METTL23, a transcriptional partner of GABPA, is essential for human cognition. *Hum. Mol. Genet.* 23, 3456–3466. doi: 10.1093/hmg/ddu054

Sasagawa, Y., Hayashi, T., and Nikaido, I. (2019). *Strategies for converting RNA to amplifiable cDNA for single-cell RNA sequencing methods*. Singapore: Springer Singapore, 1–17. doi: 10.1007/978-981-13-6037-4

Schultz, K. L. W., Troisi, E. M., Baxter, V. K., Glowinski, R., and Griffin, D. E. (2019). Interferon regulatory factors 3 and 7 have distinct roles in the pathogenesis of alphavirus encephalomyelitis. *J. Gen. Virol.* 100, 46–62. doi: 10.1099/jgv.0.001174

Shao, W., Wang, D., Chiang, Y.-T., Ip, W., Zhu, L., Xu, F., et al. (2013). The Wnt signaling pathway effector TCF7L2 controls gut and brain proglucagon gene expression and glucose homeostasis. *Diabetes* 62, 789–800. doi: 10.2337/db12-0365

Sugiaman-Trapman, D., Vitezic, M., Jouhilahti, E.-M., Mathelier, A., Lauter, G., Misra, S., et al. (2018). Characterization of the human RFX transcription factor family by regulatory and target gene analysis. *BMC Genom.* 19, 181. doi: 10.1186/s12864-018-4564-6

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Taguchi, Y.-h. (2014). "Integrative analysis of gene expression and promoter methylation during reprogramming of a non-small-cell lung cancer cell line using principal component analysis-based unsupervised feature extraction," in *Intelligent Computing in Bioinformatics*. Eds. D.-S. Huang, K. Han, and M. Gromiha (Heidelberg: Springer International Publishing). vol. 8590 of *LNCS*. 445–455 ICIC2014. doi: 10.1007/978-3-319-09330-7_52

Taguchi, Y.-h. (2015). Identification of aberrant gene expression associated with aberrant promoter methylation in primordial germ cells between E13 and E16 rat F3 generation vinclozolin lineage. *BMC Bioinform.* 16 (Suppl 18), S16. doi: 10.1186/1471-2105-16-S18-S16

Taguchi, Y.-h. (2016a). Identification of more feasible microRNA–mRNA interactions within multiple cancers using principal component analysis based unsupervised feature extraction. *Int. J. Mol. Sci.* 17, E696. doi: 10.3390/ijms17050696

Taguchi, Y. H. (2016b). "MicroRNA–mRNA interaction identification in Wilms tumor using principal component analysis based unsupervised feature extraction," in *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, 71–78. doi: 10.1109/BIBE.2016.14

Taguchi, Y.-h. (2016c). Principal component analysis based unsupervised feature extraction applied to budding yeast temporally periodic gene expression. *BioData Min.* 9, 22. pmid27366210. doi: 10.1186/s13040-016-0101-9

Taguchi, Y. H. (2016d). Principal component analysis based unsupervised feature extraction applied to publicly available gene expression profiles provides new insights into the mechanisms of action of histone deacetylase inhibitors. *Neuroepigenetics* 8, 1–18. doi: 10.1016/j.nepig.2016.10.001

Taguchi, Y.-H. (2017a). "Identification of candidate drugs for heart failure using tensor decomposition-based unsupervised feature extraction applied to integrated analysis of gene expression between heart failure and DrugMatrix datasets," in *Intelligent Computing Theories and Application* (Springer International Publishing), 517–528. Taguchi. doi: 10.1007/978-3-319-63312-1_45

Taguchi, Y.-H. (2017b). Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets. *Sci. Rep.* 7, 13733. doi: 10.1038/s41598-017-13003-0

Taguchi, Y.-H. (2017c). "One-class differential expression analysis using tensor decomposition-based unsupervised feature extraction applied to integrated analysis of multiple omics data from 26 lung adenocarcinoma cell lines," in *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, 131–138. doi: 10.1109/BIBE.2017.00-66

Taguchi, Y.-h. (2017d). Principal components analysis based unsupervised feature extraction applied to gene expression analysis of blood from dengue haemorrhagic fever patients. *Sci. Rep.* 7, 44016. doi: 10.1038/srep44016

Taguchi, Y.-H. (2017e). Tensor decomposition-based unsupervised feature extraction applied to matrix products for multi-view data processing. *PLoS One* 12, e0183933. doi: 10.1371/journal.pone.0183933

Taguchi, Y.-H. (2017f). Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases. *BMC Med. Genom.* 10, 67. InCob2017. doi: 10.1186/s12920-017-0302-1

Taguchi, Y.-h. (2018a). "Principal component analysis-based unsupervised feature extraction applied to single-cell gene expression analysis," in *Intelligent Computing Theories and Application*. Eds. D.-S. Huang, K.-H. Jo, and X.-L. Zhang (Cham: Springer International Publishing), 816–826. doi: 10.1007/978-3-319-95933-7_90

Taguchi, Y.-H. (2018b). Tensor decomposition-based unsupervised feature extraction can identify the universal nature of sequence-nonspecific off-target regulation of mRNA mediated by microRNA transfection. *Cells* 7, 54. doi: 10.3390/cells7060054

Taguchi, Y.-H. (2018c). Tensor decomposition/principal component analysis based unsupervised feature extraction applied to brain gene expression and methylation profiles of social insects with multiple castes. *BMC Bioinform.* 19, 99. APBC2018. doi: 10.1186/s12859-018-2068-7

Taguchi, Y.-h. (2019a). Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data. *BMC Bioinform.* 19, 388. doi: 10.1186/s12859-018-2395-8

Taguchi, Y.-h. (2019b). *Unsupervised Feature Extraction Applied to Bioinformatics*. Switzerland: Springer International.

Taguchi, Y.-h., and Murakami, Y. (2013). Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers. *PLoS One* 8, e66714. doi: 10.1371/journal.pone.0066714

Taguchi, Y.-h., and Murakami, Y. (2014). Universal disease biomarker: can a fixed set of blood microRNAs diagnose multiple diseases? *BMC Res. Notes* 7, 581. doi: 10.1186/1756-0500-7-581

Taguchi, Y. H., and Ng, K.-L. (2018). "Tensor decomposition-based unsupervised feature extraction for integrated analysis of TCGA data on micrRNA expression and promoter methylation of genes in ovarian cancer," in *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, 195–200. doi: 10.1109/BIBE.2018.00045

Taguchi, Y.-h., and Okamoto, A. (2012). "Principal component analysis for bacterial proteomic analysis," in *Pattern Recognition in Bioinformatics*. Eds. T. Shibuya, H. Kashima, J. Sese, and S. Ahmad (Heidelberg: Springer International Publishing). vol. 7632 of *LNCS*. 141–152 prib2012. doi: 10.1007/978-3-642-34123-6_13

Taguchi, Y. H., and Wang, H. (2017). Genetic association between amyotrophic lateral sclerosis and cancer. *Genes* 8, 243. doi: 10.3390/genes8100243

Taguchi, Y.-h., and Wang, H. (2018a). Exploring microRNA biomarker for amyotrophic lateral sclerosis. *Int. J. Mol. Sci.* 19. doi: 10.3390/ijms19051318

Taguchi, Y.-h., and Wang, H. (2018b). Exploring microRNA biomarkers for Parkinson disease from mRNA expression profiles. *Cells* 7, 245. doi: 10.3390/cells7120245

Taguchi, Y.-h., Iwadate, M., and Umeyama, H. (2015a). "Heuristic principal component analysis-based unsupervised feature extraction and its application to gene expression analysis of amyotrophic lateral sclerosis data sets," in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on.* 1–10. doi: 10.1109/CIBCB.2015.7300274

Taguchi, Y. H., Iwadate, M., and Umeyama, H. (2015b). Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for posttraumatic stress disorder-mediated heart disease. *BMC Bioinform.* 16, 139. doi: 10.1186/s12859-015-0574-4

Taguchi, Y.-h., Iwadate, M., Umeyama, H., Murakami, Y., and Okamoto, A., (2015c). "Heuristic principal component analysis-based unsupervised feature extraction and its application to bioinformatics," in *Big Data Analytics in Bioinformatics and Healthcare*. Eds. B. Wang, R. Li, and W. Perrizo, 138–162. IGI Global, Pennsylvania. doi: 10.4018/978-1-4666-6611-5.ch007

Taguchi, Y.-h., Iwadate, M., and Umeyama, H. (2016). SFRP1 is a possible candidate for epigenetic therapy in non-small cell lung cancer. *BMC Med. Genom.* 9, 28. doi: 10.1186/s12920-016-0196-3

Taguchi, Y. H., Iwadate, M., Umeyama, H., and Murakami, Y. (2017). "Principal component analysis based unsupervised feature extraction applied to bioinformatics analysis," in *Computational Methods with Applications in Bioinformatics Analysis*, vol. 8 . Eds. J. J. P. Tsai and K.-L. Ng (Singapore: World Scientific), 153–182. doi: 10.1142/9789813207981_0008

Umeyama, H., Iwadate, M., and Taguchi, Y.-h. (2014). TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer. *BMC Genom.* 15 Suppl 9, S2. doi: 10.1186/1471-2164-15-S9-S2

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Warden, A., Truitt, J., Merriman, M., Ponomareva, O., Jameson, K., Ferguson, L. B., et al. (2016). Localization of PPAR isotypes in the adult mouse and human brain. *Scie. Rep.* 6, 27618. doi: 10.1038/srep27618

Wells, T., Rough, K., and Carter, D. (2011). Transcription mapping of embryonic rat brain reveals EGR-1 induction in SOX2+ neural progenitor cells. *Front. Mol. Neurosci.* 4, 6. doi: 10.3389/fnmol.2011.00006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 08 October 2019
doi: 10.3389/fbioe.2019.00250

Identification and Analysis of Long Repeats of Proteins at the Domain Level

*David Mary Rajathei, Subbiah Parthasarathy and Samuel Selvaraj**

Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli, India

Amino acid repeats play an important role in the structure and function of proteins. Analysis of long repeats in protein sequences enables one to understand their abundance, structure and function in the protein universe. In the present study, amino acid repeats of length >50 (long repeats) were identified in a non-redundant set of UniProt sequences using the RADAR program. The underlying structures and functions of these long repeats were carried out using the Gene3D for structural domains, Pfam for functional domains and enzyme and non-enzyme functional classification for catalytic and binding of the proteins. From a structural perspective, these long repeats seem to predominantly occur in certain architectures such as sandwich, bundle, barrel, and roll and within these architectures abundant in the superfolds. The lengths of the repeats within each fold are not uniform exhibiting different structures for different functions. We also observed that long repeats are in the domain regions of the family and are involved in the function of the proteins. After grouping based on enzyme and non-enzyme classes, we observed the abundant occurrence of long repeats in specific catalytic and binding of the proteins. In this study, we have analyzed the occurrence of long repeats in the protein sequence universe apart from well-characterized short tandem repeats in sequences and their structures and functions of the proteins at the domain level. The present study suggests that long repeats may play an important role in the structure and function of domains of the proteins.

Keywords: long repeats, protein, domain, protein family, enzyme and non-enzyme classes, structural fold

OPEN ACCESS

Edited by:
Shandar Ahmad,
Jawaharlal Nehru University, India

Reviewed by:
Michael Gromiha,
Indian Institute of Technology Madras, India
Vladimir N. Uversky,
University of South Florida, United States
Tadashi Satoh,
Graduate School of Pharmaceutical Sciences, Nagoya City University, Japan
Bratati Kahali,
Indian Institute of Science, India

****Correspondence:***
Samuel Selvaraj
selvarajsamuel@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *18 April 2019*
Accepted: *16 September 2019*
Published: *08 October 2019*

Citation:
Rajathei DM, Parthasarathy S and Selvaraj S (2019) Identification and Analysis of Long Repeats of Proteins at the Domain Level. Front. Bioeng. Biotechnol. 7:250. doi: 10.3389/fbioe.2019.00250

INTRODUCTION

Amino acid repeats are ubiquitous in protein sequences that often correspond to structural and functional units of proteins. The length of these repeats varies considerably from shorter units of homo repeats of single amino acid (Jorda and Kajava, 2010), oligopeptide repeats of 2–20 residues (Fraser and MacRae, 1973) and solenoid repeats of 20–40 residues to larger repetitions of length >50 called domain repeats (Andrade et al., 2001). These repeats occur as a single pair or as multiple copies in a tandem/non-tandem manner that are useful for structural packing or for one or more interactions with ligand (Katti et al., 2000; Luo and Nijveen, 2014). It has been observed that many proteins of length >500 contain internal repeats, suggesting the importance of repeats in producing larger proteins (Marcotte et al., 1998). However, these repeats possess weak identities due to extensive divergence, but retain similar folds and functions of the proteins (Holm and Sander, 1993). It has also been found out that long stretches of perfect repetitions are infrequent in protein

sequences even though they are folded into recurrent structural motifs (Turjanski et al., 2016). Many methods and algorithms, such as Fourier transformation, short string extension, sequence-sequence alignment, and sequence profiles comparison have been introduced for the identification of such diverged sequence repeats with insertion and deletion without prior knowledge. Web based servers such as the Internal Repeat Finder, RADAR, REPRO, TRUST, XSTREAM, HHRepID, T-REKS, and PTRStalker (Pellegrini et al., 1999, 2012; George and Heringa, 2000; Heger and Holm, 2000; Szklarczyk and Heringa, 2004; Newman and Cooper, 2007; Biegert and Söding, 2008; Jorda and Kajava, 2009) have been developed by implementing the above techniques to detect amino acid repeats in proteins.

Earlier, proteins containing homo repeats (Jorda and Kajava, 2010), fibrous repeats (Fraser and MacRae, 1973) and different well-characterized repeats types, namely tetratricopeptide, leucine-rich, ankyrin and armadillo/heat etc. (Fraser and MacRae, 1973; Yoder et al., 1993; Groves and Barford, 1999; Kobe and Kajava, 2001), possessing different structures and functions have been analyzed (Andrade et al., 2001). Further, short units of repeats in tandem that form repeats in the structural folds of solenoids (α, β, α/β), β-trefoil (Murzin et al., 1992; Ponting and Russell, 2000), β-prisms (Chothia and Murzin, 1993; Bourne et al., 1999), and β-propellers (Bork and Doolittle, 1994; Neer et al., 1994) have been reviewed (Kajava, 2012). Recently, a detailed analysis and classification of β-hairpin repeat structures has been carried out (Roche et al., 2017). Also, it has been pointed out that short tandem repeats accumulate in the intrinsically disordered regions (IDR) (van der Lee et al., 2014) and play an important role in protein interactions and stability (Tompa, 2012; Habchi et al., 2014).

Analysis of larger proteins has demonstrated that significant portions of proteins are composed of domains. They are the conserved parts of proteins which can fold and function independently. The folded domains can either serve as modules for building up large assemblies or provide specific catalytic enzyme functions or bindings of the proteins. It has been found that repeats of a length >50 residues often correspond to conserved regions that are present in proteins as single or multiple copies for the function of the proteins (Hemalatha et al., 2007). Our analysis of sequence repeats of the proteins with known 3D structures in the PDB (Berman et al., 2014) has shown that they retain similar folds in spite of divergences, in order to conserve the structure and function of the proteins and, repeats that are in the single/two domains from the same family contain conserved motifs for the function of the proteins (Mary Rajathei and Selvaraj, 2013). Further, the conservation of inter-residues interactions in domain repeats have been analyzed in terms of long-range contact, surrounding hydrophobicity and pair-wise interaction energy (Mary et al., 2015). A database IR-PDB for repeats in the sequence of the proteins in the PDB has been developed for the analysis of impact of repeats in proteins (Selvaraj and Rajathei, 2017).

The widely used sequence database UniProtKB (UniProt Consortium T, 2017) contains more than 500,000 sequences that are annotated with well-characterized repeats of tetratricopeptide, leucine-rich repeats, ankyrin, and armadillo/heat etc. However, there has been no survey of repeats of length>50 in the UniProt sequences, which may provide insights into their role in the structure, function and evolution of the proteins. In the present study, we have analyzed the occurrence of long repeats and their underlying structures and functions in a non-redundant set of UniProt sequences. Since repeats of size exceeding 50 residues are large enough to fold independently into stable domains (Kajava, 2012), we used Gene3D for structural domains, Pfam for functional domains and enzyme and non-enzyme functions for specific catalytic and binding for their structure and function of long repeats proteins. It was found that long repeats occur in about 23% of the considered proteins. Analysis of the structure of long repeats reveals that these repeats are predominantly observed in the structural folds of sandwich, bundle, barrel and roll. We observed that repeats in the domains for the function of the proteins. Further, we observed that long repeats tend to occur both in enzyme and non-enzyme functions of proteins. While long repeats are found in all the major enzyme classes, these are more abundant among both ligases and isomerases. Among the non-enzyme proteins, such as DNA binding, metal binding, calcium binding, and Nucleotide binding (NP), these repeats are observed more in Nucleotide binding and DNA binding proteins. The present analysis shows that the occurrence of long repeats and their structures and functions of the proteins at the domain level.

MATERIALS AND METHODS

Data Collection

A collection of 555,100 proteins along with their assigned UniProt ID, amino acid sequence, protein name, protein family, enzyme function, and non-enzyme functions such as DNA binding, calcium binding, metal binding, and NP binding, as well as other annotation of the sequences from the databases of Pfam, Gene3D, PDB, and DisProt, was downloaded from UniProtKB/Swiss-Prot (UniProt Consortium T, 2017) and stored in a file. The Pfam is a database of protein domain families that assigns the domains, as well as their functional regions (Finn et al., 2014). Gene3D (Lewis et al., 2018), is a database that assigns the structure of the protein according to CATH hierarchy of class, architecture and fold in numerical values (Dawson et al., 2017). At the class level (C), the numerical value 1 is for all alpha class, 2 for all beta and 3 for a mixture of alpha and beta. Likewise, the numerical values are assigned for Architecture level (A) based on secondary structure arrangement in 3-D space and for Topology/Fold level (T) based on the connection of secondary structural elements. The PDB ID's of the 3D structure known proteins were obtained from the PDB database (http://www.rcsb.org/pdb/home/home.do). The intrinsic disordered regions of the proteins that were extracted from the literature are available in the DisProt database (Piovesan et al., 2017). A non-redundant representative set of 126,945 sequences that share <50% sequence identity was obtained by clustering the 555,100 sequences using the web server CD-HIT (Fu et al., 2012). The overall work-flow is summarized as a flowchart (**Figure 1**).

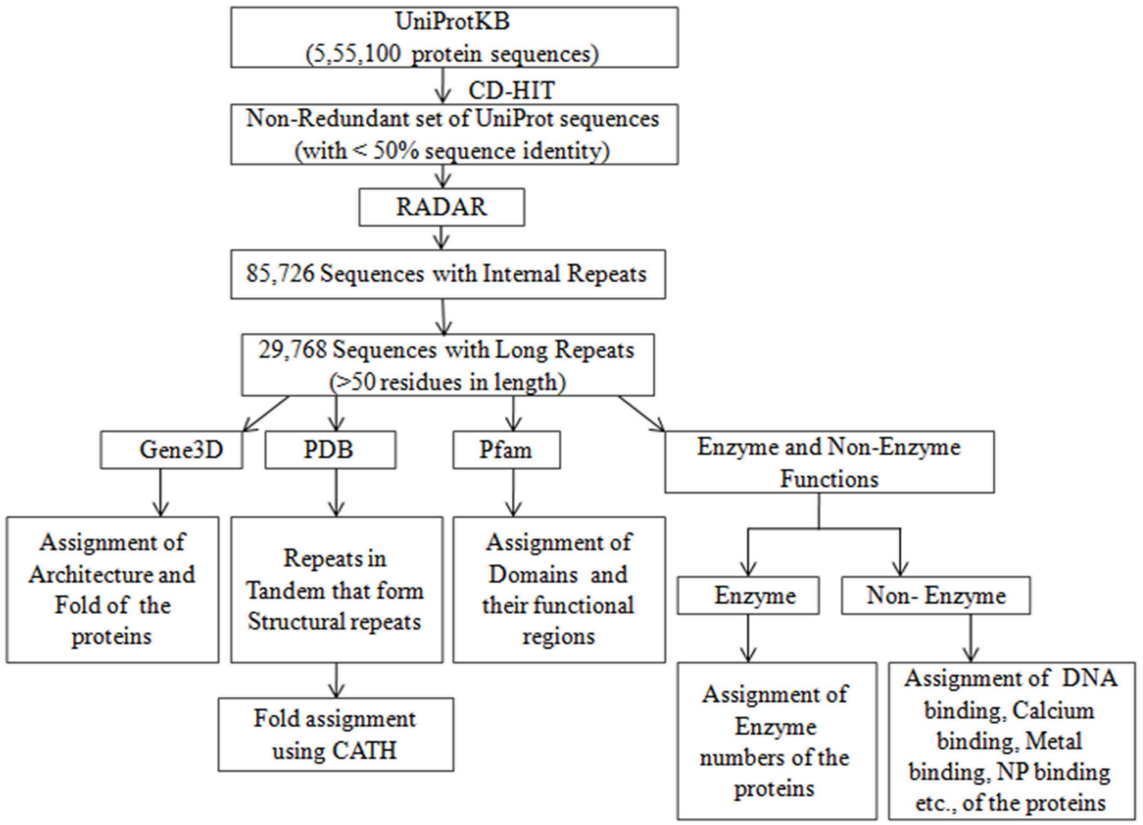

FIGURE 1 | Flow diagram of identification and analysis of Long repeats from non-redundant set of UniProt sequences.

Finding Sequence Repeats of the Proteins Using RADAR

The presence of internal repeats in each protein sequences was identified using the repeat detection program RADAR (Heger and Holm, 2000), which was downloaded from the URL (https://sourceforge.net/projects/repeatradar). The RADAR program is efficient for *ab initio* detection of repeats of length >15 in a single sequence by aligning the sequence against itself, as well as by generating the sequence profile using multiple sequence alignment. RADAR evaluates the statistical significance of the observed repeats by measuring a Z-score for each repeat unit (McLachlan, 1983; Heringa and Argos, 1993). The *Z*-score of a repeat unit is the number of standard deviations of the repeat unit score above the mean. The score of each unit is determined from a profile derived from the multiple alignment of repeat unit without considering end-gaps. Repeats with *Z*-scores threshold of > 6 are reported by the RADAR program. An in-house Perl program that incorporated the RADAR executable was written to detect internal repeats of all sequences in the dataset in a single run. Proteins containing repeats of length >50 were considered for further analysis.

Finding the Structure of Long Repeats Proteins

The UniProt ID's of proteins having long repeats were extracted and their Gene3D structural domain-based assignments of the proteins were extracted using a Perl program. Then, the name of class, architecture and fold of the protein was found out by using CATH search and grouped according to their name for the further analysis of architecture and fold of the protein with repeats.

Finding the Functional Domains of Long Repeats Proteins

The UniProt ID's of long repeat proteins were extracted and their assigned Pfam domains of the sequences were identified. The domain regions and their functional residues information of the proteins were found out using Pfam database search (Finn et al., 2014), and repeats in the domain regions were identified by manual search. The level of similarity of the repeats within a protein and within a protein family was found out in terms of % sequence identity through using the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970) implemented in the ggsearch36 program of the FASTA-36.3.5b package (Henikoff and Henikoff, 1992). Needleman–Wunsch alignment scores were calculated using the BLOSUM50 scoring matrix (Pearson, 2000) with a penalty of −12 for gap opening and −2 for gap extension. Further, the repeats in domains of the proteins were also analyzed for their functional involvement at the structure of the proteins using the server PDBsum by giving PDB ID as input (Laskowski et al., 2018).

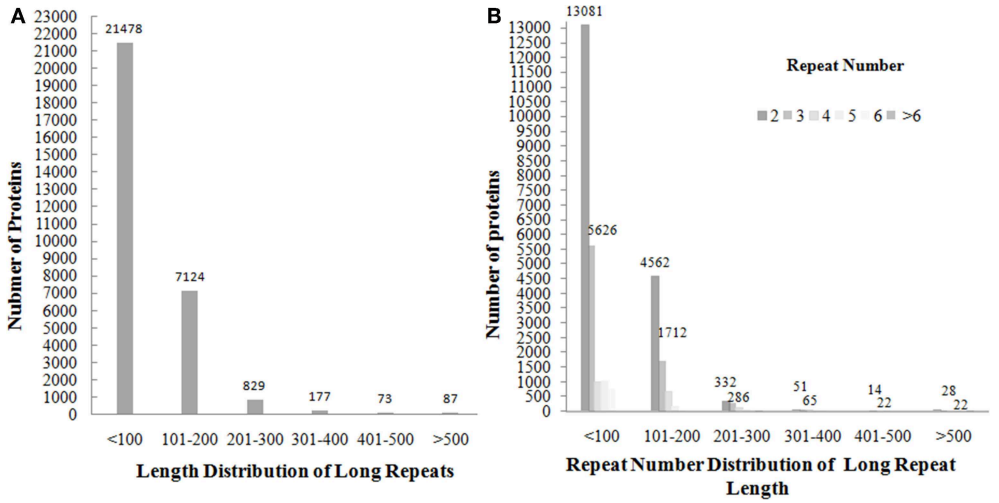

FIGURE 2 | The plotting of number of proteins against the distribution of long repeats of length >50 in the range of <100, 101–200, 201–300, 301–400, 401–500, and>500 shows that most of the longrepeat lengths fall in the range of <200 **(A)** and repeat number distribution of long repeats shows that repeat numbers of 2 and 3 in most of long repeat proteins **(B)**.

Finding the Enzyme and Non-enzyme Functions of Long Repeats Proteins

The assigned enzyme numbers (EC) of long repeats proteins were extracted. The EC number of the protein at the first level corresponds to seven enzyme classes of Oxidoreductases (EC 1), Transferases (EC 2), Hydrolases (EC 3), Lyases (EC 4), Isomerases (EC 5), Ligases (EC 6), and finally, Translocases (EC 7). The enzyme numbers were extracted and grouped according their numbers for further analysis. The non-enzyme proteins that are assigned with DNA binding, calcium binding, metal binding and NP binding were also extracted and grouped according to their name.

RESULTS

Abundance of Proteins Having Long Repeats

The presence of amino acid repeats of length >15 was found out in 85,726 (67%) out of non-redundant set of 126,945 UniProt protein sequences (**Supplementary Data File 1**). The long repeats were found out in 29,768 (35%) proteins. These repeats are present as a single pair or multiple copies of repeats in tandem/non-tandem manner. For example, N-acetylmuramoyl-L-alanine amidase Rv3717 protein (UniProt ID: I6Y4D2) of length 241 contains a single pair of tandem repeats of length 96 in the continuous region of 12–116/118–226. Complement control protein C3 (P68639) of length 263 has three copies of repeats of length 55 in tandem (81–142/143–200/201–254), whereas Transcriptional regulatory protein TyrR (UniProt ID: P44694) of length 318 contains a single pair of non-tandem repeats of length 76 in the discontinuous region of (19–99/200–279). The length of the repeats varies in the range of 51–1759 and lengths of >1,000 are mostly found out in enzyme proteins. For example, the non-ribosomal peptide synthetase 1 (Q4WT66) of sequence length 6,269 contains repeats of length 1,546 (277–905/906–2,334/2,335–3,465/3,466–4,590/4,593–5,634). Through the analysis of length distribution of long repeats, as well as their repeat number distribution of long repeats against the number of proteins (**Figure 2**), we observed that the lengths <200 are observed in more than 90% of the proteins with an average of 100 residues, and repeated in 2–5 number of times with repeat numbers of 2 (61%) and 3 (26%) in most of the long repeat proteins. The *Z*-score values of the repeats were extracted and found that 74,089 out of 74,154 repeat units have *Z*-scores >6. Among these, 66,400 repeat units have *Z*-scores of >20. This suggests that most of the observed repeats are statistically significant.

Analysis the Structure of Long Repeats Proteins

The structural class, architecture and fold of the 14,176 proteins (48%) have been found out using structural domain based Gene3D assignments. Among these, some proteins are having two or more Gene3D assignments. In this study, 10,504 proteins that contained a single Gene3D assignment were considered for further analysis (**Supplementary Data File 2**). For example, Annexin A1 (P04083) protein contains a single Gene3D assignment of 1.10.220, which means that this protein belongs to class alpha (1) of orthogonal bundle architecture (10) with Annexin V domain fold (220).

Analysis of Long Repeats at the Architecture Level

According to CATH domain-based hierarchy (http://www.cathdb.info/browse/tree), the presence of long repeats in different architectures of alpha (α), beta (β), and alpha/beta (α/β) class proteins was observed (**Figure 3**). Out of five architectures of α class, these were observed in the four architectures, namely orthogonal bundle, up-down bundle, α horseshoe and

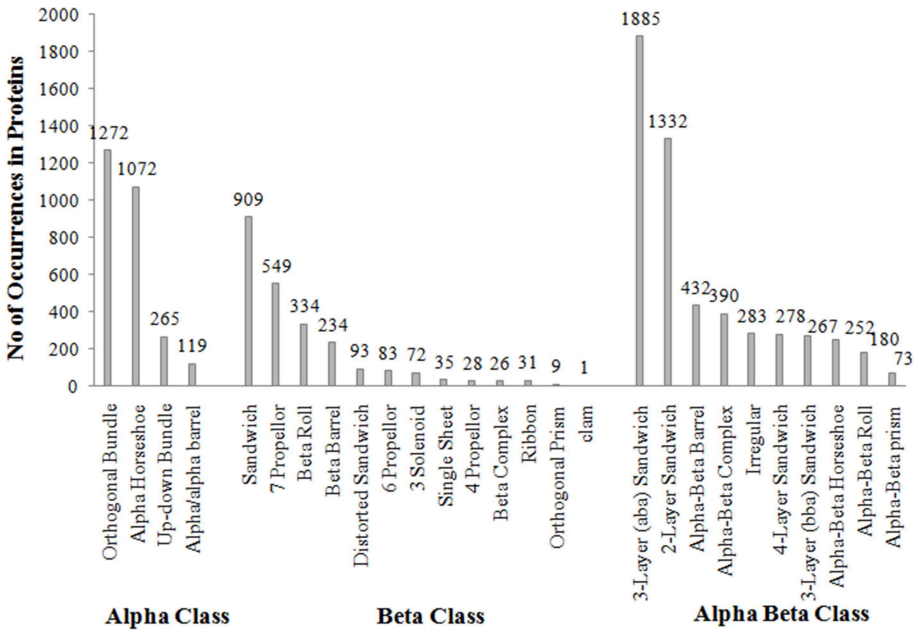

FIGURE 3 | Number of Long repeats containing proteins assigned with different architectures of α, β, and α/β class using CATH.

α/α barrel. Among these, substantial numbers were present in the architectures of bundle and horseshoe. Under β class, the repeats were present in 13 out of 20 architectures and the sandwich, propeller, roll and barrel were observed most. Likewise, repeats were found in 10 out of 14 architectures of α/β class and the architectures of 3-layer (αβα) sandwich, 2-layer sandwich, α/β barrel and αβ-complex were observed most. By combining the architectures from different classes of proteins, repeats in specific architectures of sandwich, bundle, barrel and roll compared to other architectures were found out.

Analysis of Long Repeats at the Fold Level

The existence of repeats in different folds of sandwich, bundle, barrel and roll architectures was found out. Repeats were observed in 84 out of 287 folds in orthogonal bundle and 32 out of 101 folds in up-down bundle of α class. At the β class, 12 out of 43 folds in β sandwich, 18 out of 48 folds in β barrel, and 13 out of 40 folds in β roll architecture of the proteins were having repeats. Under α/β class, repeats in 47 out of 126 folds under 3-layer (αβα) sandwich, 57 out of 224 under 2-layer sandwich, 6 out of the 18 folds under α/β barrel and 16 out of 58 folds under α/β roll were observed. Among that, some folds were observed in a greater number of proteins compared to other folds (**Figure 4**). In α class, the Arc Repressor Mutant Subunit A fold and four Helix Bundle fold of bundle architectures were observed in most of the proteins compared to other folds (**Table 1**). Under β class, the Immunoglobulin-like fold and Jelly Roll fold of β-sandwich, PH-domain fold of β-roll, and OB fold of β-barrel were observed most. The Rossmann fold in 3-layer (αβα) sandwich, TIM Barrel in αβ barrel and Herpes Virus-1 followed by Alpha-Beta plaits fold of 2-layer Sandwich, and Ubiquitin-like (UB roll) of αβ roll were observed most. The results reveal the predominant occurrence of long repeats in the diverse structure exhibiting folds of the proteins.

Analysis of Long Repeats for Structural Repeats

The long repeats in proteins with known 3-D structure (as available from UniProt annotation) were analyzed for structural repeats. The proteins with tandem repeats were found out and analyzed at the structural level. We observed long tandem repeats form structural repeats in the folds of up-down and orthogonal bundle of α-class, Immunoglobulin, Jelly Roll and OB fold of β-class, Rossmann fold, TIM barrel, α/β plait, and UB roll of α/β class. **Figure 5** shows the structural repeats of the proteins in the folds of up-down and orthogonal bundle of α-class, Immunoglobulin, Jelly Roll and OB fold of β-class, Rossmann fold, TIM barrel, α/β plait, and UB roll of α/β class. Further, we found out that the lengths of the repeats are not uniform and vary considerably within each fold. **Figure 6** shows the considerable variation in lengths, as well as in the secondary structures of different proteins possessing the Rossmann fold that usually contains βαβαβ secondary structure arrangements. The Desulfovibrio vulgaris CbiK(P) Cobaltochelatase (PDB ID: 2XVY) contains two repeats of βαβαβαβ secondary structure arrangement of length 103 (**Figure 6A**) (Malay et al., 2009), whereas, another protein Thermoplasma volcanium Phosphoribosyl pyrophosphate synthetase (PDBID: 3MBI) contains two repeats

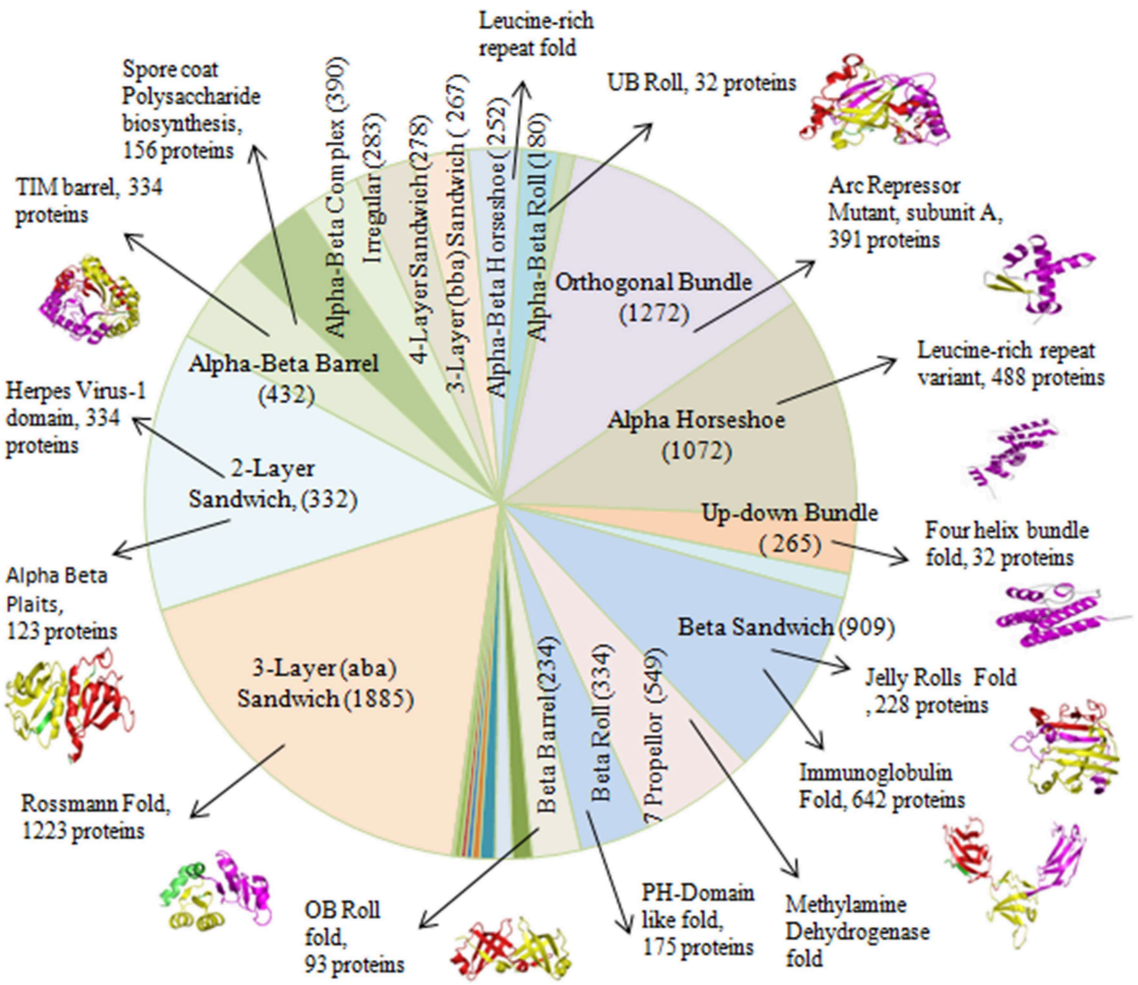

FIGURE 4 | The occurrences of certain folds of Arc Repressor Mutant subunit A of Orthogonal bundle architecture, Four helix bundle of Up-down bundle, Jelly Rolls and Immunoglobulin of Beta sandwich, PH-Domain like fold of Beta Roll, OB Roll of Beta Barrel, Rossmann fold of 3-layer sandwich, Alpha Beta plaits, and Herpes Virus-1 domain of 2-layer sandwich, TIM barrel of Alpha-Beta Barrel and UB Roll of Alpha-Beta Roll in a substantial numbers of Long repeats proteins.

TABLE 1 | Number of proteins containing long repeats in the architectures and folds of the proteins.

Class	Architecture	Number of proteins	Fold	Number of proteins
Alpha (α) class	Orthogonal bundle	1,272	Arc repressor mutant, subunit A	391
	Alpha horseshoe	1,072	Leucine-rich repeats variant	488
	Up-down bundle	265	Four helix bundle	32
Beta (β) class	Beta sandwich	909	i) Jelly Rolls ii) Immunoglobulin	228 642
	7 Propeller	549	Methylamine dehydrogenase	549
	Beta roll	334	PH-domain like	175
	Beta barrel	234	OB Roll	93
Alpha Beta (αβ) class	3-Layer (aba) Sandwich	1,885	3-layer(αβα) sandwich	1,223
	2-Layer sandwich	1,332	Alpha beta plaits	123
	Alpha-beta barrel	432	TIM barrel	334
	Alpha-beta complex	390	Spore coat polysaccharide biosynthesis protein SpsA	156
	Alpha-beta roll	180	UB Roll	32

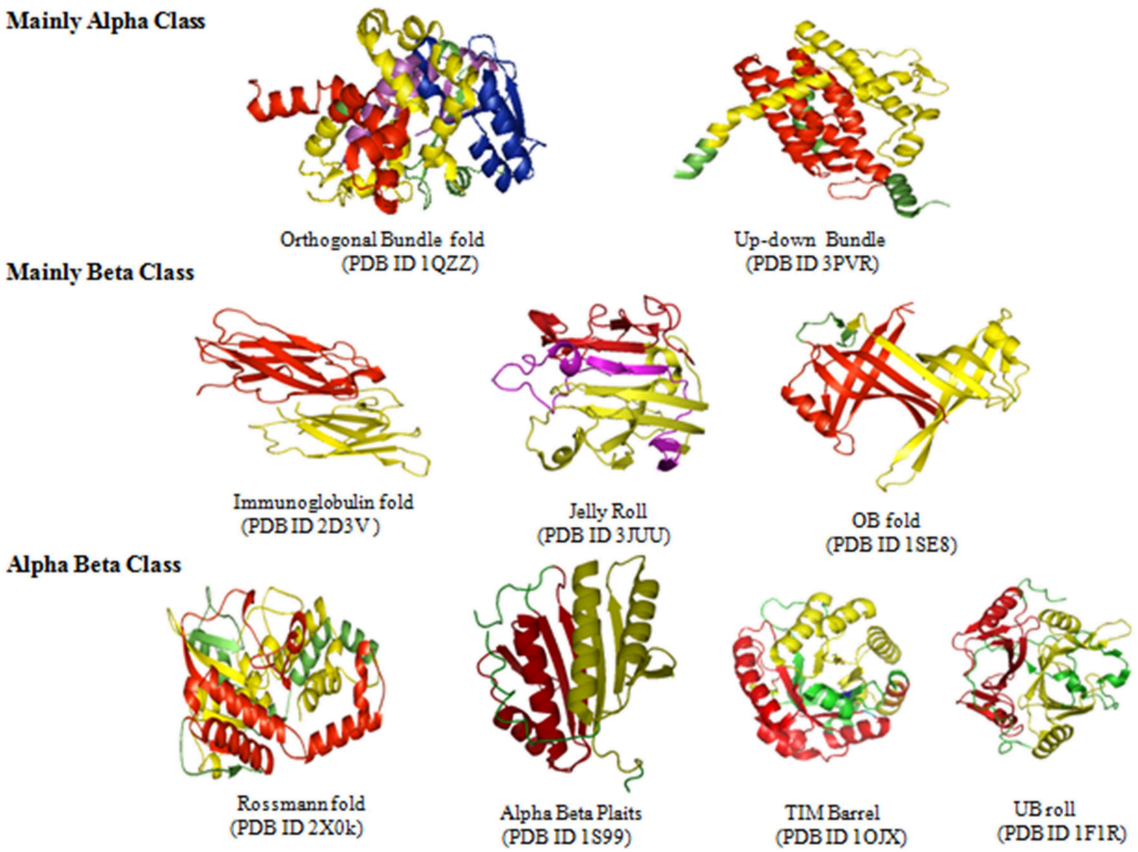

FIGURE 5 | Long repeats that form structural repeats in the folds of Orthogonal bundle, Up-down bundle of alpha class, Immunoglobulin fold, Jelly Roll, OB fold of beta class, Rossmann fold, Alpha Beta Plait, TIM barrel, and UB roll of alpha-beta class.

of βαβαβαβα of length 121 in **Figure 6B** (Cherney et al., 2011). The analysis results suggest that the length variations of repeats within the Rossmann fold lead to the presence of additional α-helices, β-strands, and coil regions. Thus, longer repeats of different lengths provide the structural differences within a fold of the proteins.

Analysis of Long Repeats for Intrinsic Disordered Region

The intrinsically disordered regions (IDR) for 51 (<1%) of long repeats proteins were found out using DisProt database. While analyzing the predisposition of long repeats for IDR, most of the repeats were identified in the structured regions. However, we also identified long repeats in an IDR. For example, Nucleoporin NUP1 (P20676) protein of length 1,076 contains tandem repeats of length 62 in the region of (352–399/403–462/522–564/666–728/731–778/779–840/849–906/907–972/978–1,031), which has been identified as an IDR (300–1,078). This analysis suggests that long repeats are generally structured in most of the proteins while few of them may have IDRs.

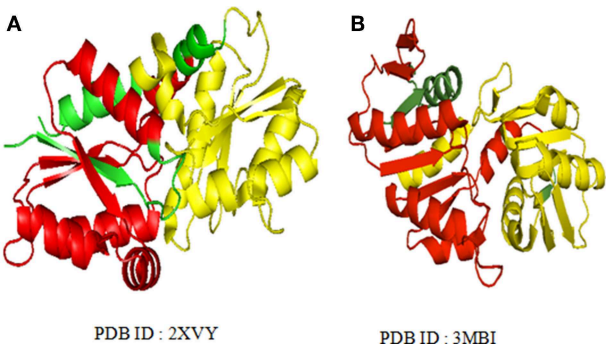

FIGURE 6 | Varying larger repeat lengths are observed in the Rossmann fold of the **(A)** Cobalt chelatase CbiK (2xvyA) with repeat length 103 and **(B)** Phosphoribosyl pyrophosphate synthetase (3 mbiA) with repeat length 121. The repeats regions are highlighted with different colors.

Analysis of Functions of Long Repeats at the Domain Level

The Pfam domain assignments in 26,750 (90%) of proteins were found and suggested the occurrence of repeats in the functional domain families containing proteins. While grouping

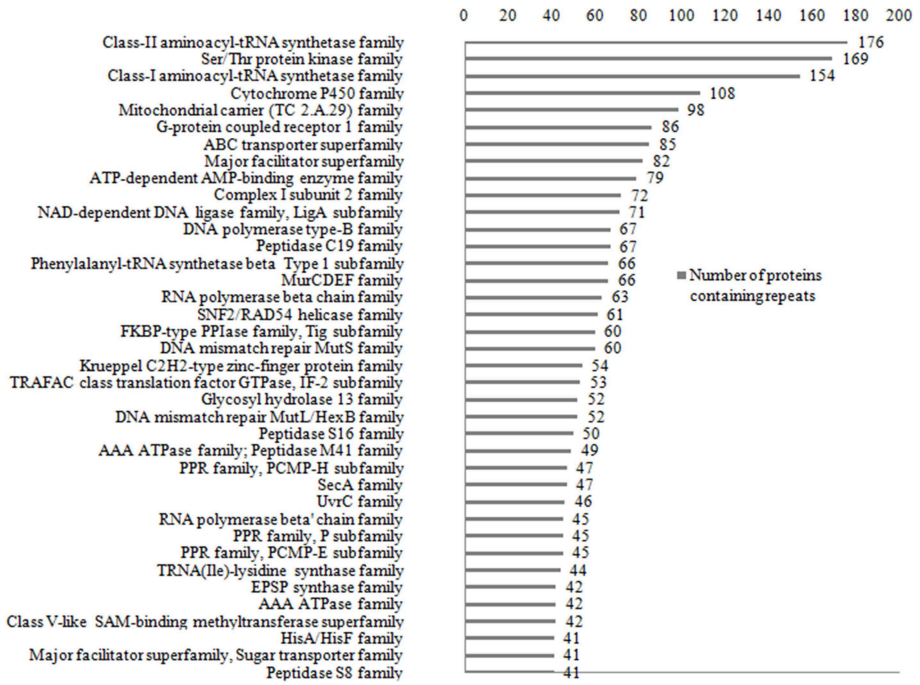

FIGURE 7 | List of the 36 protein families that are having long repeats in more than 40 member proteins.

by protein family, the existence of repeats in 5,258 distinct protein families was found out. Some of the protein families are having long repeats in a greater number of their member proteins (**Supplementary Data File 3**). **Figure 7** shows the list of 36 protein families such as Class II aminoacyl-tRNA synthetase, Ser/Thr Protein kinase, Class I aminoacyl-tRNA synthetase, Cytochrome P450, Mitochondrial carrier (TC 2.A.29), G-protein coupled receptor 1, and ABC transporter that are having repeats in more than 40 member proteins of the family. We observed long repeats in the domains of the family with varying lengths. For example, the Peptidase S8 family proteins contained long repeats in 41 member proteins of the family (**Table 2**). Among these, 38 protein repeats were in the Peptidase S8 domains with varying repeat lengths. **Figure 8** shows some of the proteins' repeat regions as well as their alignment that covers the Peptidase S8 domain regions. The level of similarity between the repeats in the Peptidase S8 domain within a protein and within the member proteins of the Peptidase S8 domain family was computed in terms of % sequence identity. For example, the sequence identity of 29% was observed for the repeats (157–215/228–313), within the Peptidase domain (157–401) of the Aqualysin-1 protein (P08594) (**Table 2**). Further, the sequence similarities of repeat unit (157–215) of this protein, with the repeat units in the Peptidase S8 domain of the 37 member proteins, were also computed. We observed that 65 % of repeats were in the range of 20–40% sequence identity and the remaining protein repeats were in the range of 10–20% identity. This observation suggests that the repeats within a protein, as well as within a protein family, are considerably diverged.

Further, repeats in the domains are involved in the function through functional residues (highlighted in red color). For example, the regions (162–173) and (197–207) of repeats (157–215/228–313) of Aqualysin-1 (UniProt ID P08594) have contained functional residues VYVIDTGIRTTH and HGTHVAGTIGG for Serine proteases (**Figure 8**). The functional involvement of the repeats was also found out in the structure of the proteins using PDBsum. For example, the functionally involved residues (highlighted red in color) of repeats (157–215/228–313) in the structure of Aqualysin-1 (PDB ID 4DZT) were found out using PDBsum search (**Figure 9**). This suggests that these repeats occur in the domains of the family for the function of the proteins.

Analysis of Enzyme and Non-enzyme Functions of Long Repeats

Further, the enzyme functions in 13,333 proteins and non-enzyme functions in 2,437 proteins, of a total of 15,770 (53%) of long repeats proteins, were also found out. Of a total of 13,333 enzymes having long repeats, Ligases (35.91%) have the maximum number of repeats followed by Isomerases (28.98%), Translocases (11.81%), Transferases (11.12%), Lyases (5.12%), Hydrolases (4.74%), and Oxidoreductases (2.32%). Among the non-enzymes in 2,437 proteins, NP binding proteins (48.09%) have the maximum number of repeats followed by DNA binding (30.44%), metal binding (16.94%), and calcium binding (4.51%). These observations suggest the importance of long repeats in both the catalytic and binding function of proteins apart from serving as modules of large assemblies.

TABLE 2 | List of 41 member proteins of the Peptidase S8 family long repeat region's and their function domain regions assigned using Pfam.

S. NO.	Protein name (UniProt ID and length)	Long repeats regions and their length	Peptidase S8 domain regions and their length	Other domains and their regions
1	Aqualysin-1 (P08594 514)	157–215/228–313 (57)	157–401 (244)	Inhibitor_I9 (54–125)
2	Bacillopeptidase F (P16397 1434)	i) 198–279/280–352/355–436/437–529 (85) ii) 568–609/615–701/1,044–1,167 (79)	218–504 (286)	Peptidase_M6 (667–801) Inhibitor_I9 (68–178)
3	Calcium-dependent protease (Q59149 663)	219–313/315–412 (93)	228–530 (302)	P_proprotein (547–662)
4	Cell wall-associated protease (P54423 895)	737–803/818–885 (67)	458–729 (271)	
5	Cuticle-degrading protease (P29138 389)	72–171/172–270/277–355 (84)	139–383 (244)	Inhibitor_I9 (41–107)
6	Extracellular serine protease (P29805 1046)	i) 158–240/241–381/385–491 (137) ii) 509–585/586–685/687–754/771–831 (83)	71–397 (326)	Autotransporter (769–1,045)
7	Microbial serine proteinase (P31339 622)	167–237/389–458 (67)	89–411 (322)	P_proprotein (491–572)
8	Minor extracellular protease vpr (P29141 807)	18–148/149–208/340–473/475–532/658–711 (183)	184–594 (410)	Inhibitor_I9 (57–143); PA superfamily (355–497); FlgD_ig superfamily (712–792)
9	Minor extracellular protease Epr (P16396 646)	39–112/169–240/249–327 (74)	137–380 (243)	
10	MycP4 protease (I6YC58 456)	25–208/222–407 (159)	86–389 (303)	
11	MycP1 protease (A0QNL1 450)	91–172/173–327/332–428 (127)	83–381 (298)	
12	Nisin leader peptide-processing serine protease (Q07596 683)	228–281/379–418/504–557 (52)	255–546 (291)	
13	PIII-type proteinase (P15292 1963)	156–206/208–209/295–380 (82)	212–698 (486)	
14	Proprotein convertase subtilisin/kexin type 9 (Q80W65 695)	467–537/540–611/616–682 (140)	185–423 (238)	Inhibitor_I9 (80–152)
15	Pyrolysin (P72186 1399)	i) 225–274/276–339/341–400 (62) ii) 959–1,006/1,011–1,158/1,169–1,296 (126)	i) 174–380 (206) ii) 408–654 (246)	
16	Putative subtilisin-like proteinase 1 (Q8SQJ3 466)	23–92/94–160/165–195 (67)	144–422 (278)	Inhibitor_I (919–90)
17	Putative subtilisin-like proteinase 2(Q8SS86 536)	106–165/278–336/362–390 (60)	272–452 (180)	
18	Probable subtilase-type serine protease DR_A0283 (Q9RYM8 729)	84–126/131–209/232–310/320–378 (78)	183–470 (287)	Peptidase_M14NE-CP-C_like(486–558); PPC (624–693)
19	Subtilase-type proteinase psp3 (Q9UTS0 452)	217–283/349–407 (56)	202–429 (227)	Inhibitor_I9 (80–162)
20	Subtilase-type proteinase RRT12 (P25381 492)	53–107/269–320 (52)	156–389 (233)	
21	Subtilisin-like protease SBT3.13 (Q8GUK4 767)	320–431/608–719 (107)	153–588 (453)	Inhibitor_I9 (41–119); PA_Superfamily (384–485)
22	Subtilisin-like protease SBT4.4 (Q9FGU3 742)	65–226/416–581 (150)	137–581 (444)	Inhibitor_I9(34–112); PA(338–458)
23	Subtilisin-like protease SBT4.10 (Q9FIM8 694)	138–284/387–534 (139)	138–526 (388)	Inhibitor_I9 (35–113);PA (332–371)
24	Subtilisin-like protease SBT4.14 (Q9LLL8 750)	202–333/336–464/467–596 (129)	141–594 (453)	Inhibitor_I9 (38–115); PA (346–467)
25	Subtilisin-like protease SBT2.4 (F4HYR6 833)	245–361/362–547/548–736 (178)	169–691 (522)	Inhibitor_I9 (70–138);PA (389–533)
26	Subtilisin-like protease SBT4.15 (Q9LZS6 767)	284–379/450–550 (92)	137–590 (453)	Inhibitor_I9 (35–113); PA (342–474)
27	Subtilisin-like protease SBT3.18 (Q9STQ2 780)	i) 179–224/495–575/707–756 (76) ii) 318–388/405–476 (65)	137–613 (476)	Inhibitor_I9 (30–109); PA_Superfamily (361–482)
28	Subtilisin-like protease SBT6.1 (Q0WUG6 1039)	556–644/812–901 (86)	208–486 (278)	
29	Subtilisin-like protease SBT2.2 (Q9SUN6 857)	163–222/226–283 (54)	184–674 (490)	Inhibitor_I9 (98–159); PA superfamily (406–548)

(Continued)

TABLE 2 | Continued

S. NO.	Protein name (UniProt ID and length)	Long repeats regions and their length	Peptidase S8 domain regions and their length	Other domains and their regions
30	Subtilisin-like protease SBT2.6 (Q9SZV5 817)	155–186/195–224/315–398 (59)	151–635 (484)	Inhibitor_I9 (61–124); Pasuperfamily (374–511); fn3_5 superfamily (698–810)
31	Subtilisin-like protease SBT3.6 (Q8L7I2 779)	216–306/307–392 (71)	138–593 (455)	Inhibitor_I9 (34–113); PA superfamily (365–493)
32	Subtilisin-like protease SBT1.2 (O64495 776)	150–259/527–633 (101)	127–587 (460)	Inhibitor_I9 (27–112); PA superfamily (353–481)
33	Subtilisin-like protease SBT1.4 (Q9LVJ1 778)	169–268/435–532 (86)	133–589 (456)	Inhibitor_I9 (32–110); PA superfamily (355–474)
34	Serotype-specific antigen 1 (P31631 933)	378–433/466–588/594–709 (115)	54–408 (354)	Autotransporter superfamily (673–916)
35	Subtilisin-like protease 12 (D4AQA9 417)	253–300/305–372 (64)	145–399 (254)	Inhibitor_I9 (35–116)
36	Subtilisin-like protease CPC735_047380 (C5PFR5 401)	84–138/140–196 (53)	143–363 (220)	Inhibitor_I9 (35–114)
37	Tripeptidyl-peptidase 2 (Q09541 1375)	370–469/688–782 (87)	89–559 (470)	TPPII (832–1,017)
38	Tripeptidyl-peptidase 2 homolog (Q9UT05 1275)	i) 265–379/413–522 (96) ii) 637–807/820–958 (121)	90–545 (450)	TPPII (837–1,008)
39	Thermophilic serine proteinase (Q45670 402)	121–202/203–282/283–358 (79)	151–392 (241)	
40	Tripeptidyl-peptidase 2 (F4JVN6 1381)	i) 143–207/343–403/717–760 (62) ii) 1,043–1,188/1,236–1,380 (135)	140–620 (480)	TPPII (897–1,078) SMC_N (1,140–1,355)
41	Subtilisin-like protease (Q00139 371)	20–72/80–135 (52)	83-255 (172)	P–proprotein (240–370)

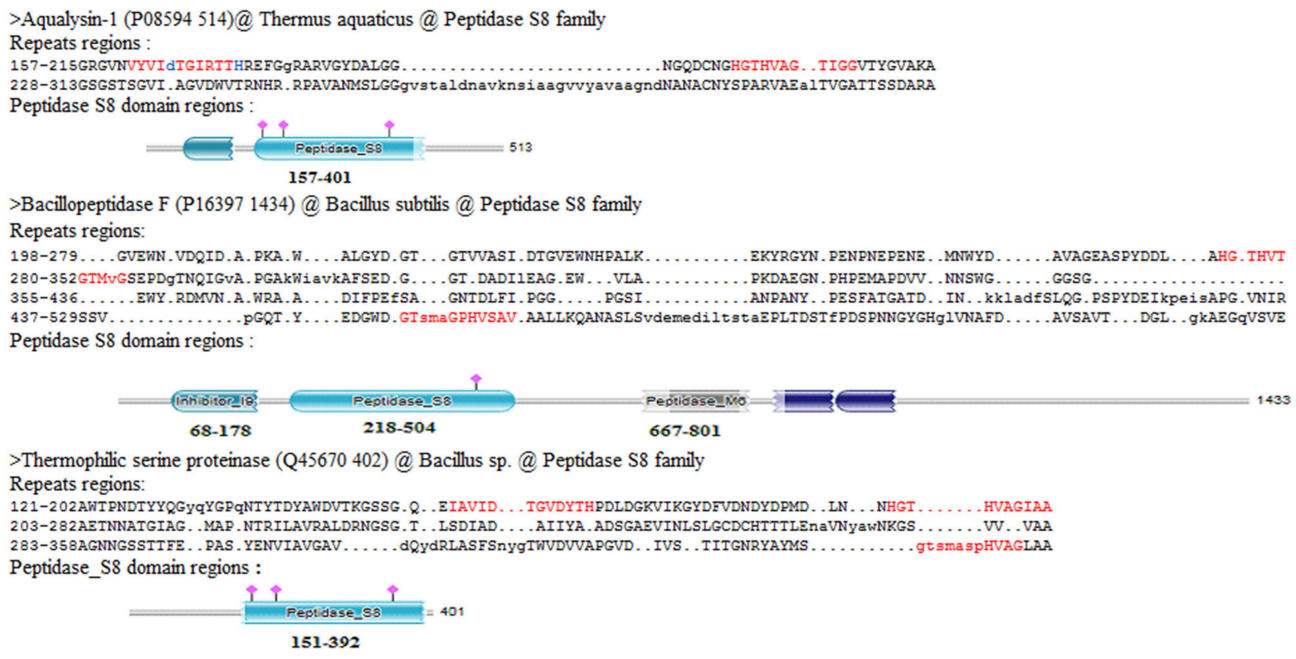

FIGURE 8 | The Aqualysin-1(P085594), Bacillopeptidase F(P16397), and Thermophilic serine proteinase (Q45670) protein's repeats regions and their alignments that are in the Peptidase S8 domain region assigned by Pfam.

DISCUSSION

Our survey of long repeats in a non-redundant set of UniProt sequences has highlighted the occurrence of these repeats that play an important role in the structure and function of domains of the proteins. Previous studies have focused on structural and functional implications of proteins with homo repeats (Uthayakumar et al., 2012), fibrous repeats (Parry, 2005) and different well-characterized repeats of length 5–50 (Andrade et al., 2001). Therefore, an in-depth study of long repeats in

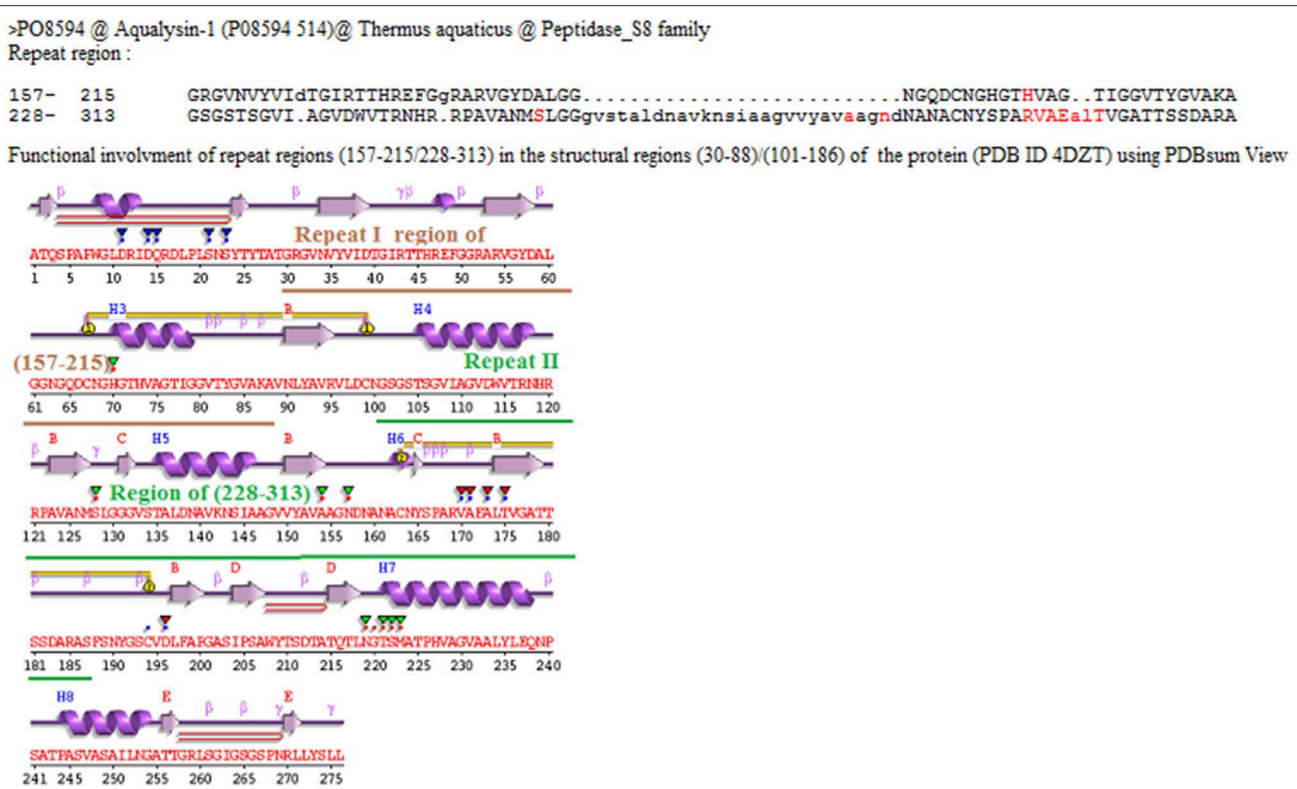

FIGURE 9 | The functional residues of repeats (157–215/228–313) in the structural regions (30–88)/(101–186) (highlighted by red and green inverted triangles with red color dots) of Aqualysin-1 proteins are found out using PDBsum search.

UniProt sequences was carried out for a better understanding of the correspondence of repeat sequences with their structures and functions. In this study, we used the RADAR program for internal repeat detection, since it often detects both tandem and interspersed repeats in larger size. Our earlier studies for repeats analysis (Mary Rajathei and Selvaraj, 2013; Mary et al., 2015) have shown the ability of RADAR to detect repeats of length > 50 that are structurally similar and conserved in a 3D structure environment. Further, the sensitivity and accuracy of RADAR repeats, by comparison with Pfam, indicate good coverage, accurate alignments, and reasonable repeat borders (Heger and Holm, 2000). The identified repeats vary in the range of 50–1,759 of lengths and diverged with more insertions and deletions, but the calculated z-scores by RADAR have shown their statistical significance.

From a structural perspective, long repeats tend to occur abundantly in certain architectures of sandwich, barrel, bundle, and roll. Within these architectures, they are predominately observed in the super folds of up-down and orthogonal bundle of α-class, Immunoglobulin, Jelly Roll and OB fold of β-class, Rossmann fold, TIM barrel, α/β plait, and UB roll of α/β class of the proteins. The adoption of classic super secondary elements (αα, βαβ, ββ) and incorporation of repetitive duplication of a small stable unit may be the possible reasons for abundance of larger duplication in these folds (Thornton et al., 1999). For example, the evolution of the $(\beta\alpha)_8$ repeat in the TIM barrel is through repetitive duplication of a small stable unit (βα) (Lang et al., 2000). It has been observed that repeats in the folds may fulfill the physical demand (stable and fast folding conformation) of the protein chain during the process of evolution, in order to meet the cellular function (Lupas et al., 2001). Further, it has been shown that the existence of structural symmetries in the super-folds (6 out of 10) may also require larger duplication during evolution of the proteins (Brych et al., 2003). Kim et al. (2010), through their SymD (detecting symmetry in protein structures) method, have identified 33 folds that contain 10 or more symmetric domains. There is considerable overlap between the symmetry in the folds they identified and those observed in the present work (**Figure 5**). We observed that long repeats of different lengths within a fold provide the structural differences of the proteins for different functions. Further, the analysis of predisposition of long repeats for disordered regions has shown that long repeat proteins are mostly structured to form stable folds. However, it has been observed that short tandem repeats are highly disordered, which do not adopt a single defined configuration for specific function (Tompa, 2012; Habchi et al., 2014; van der Lee et al., 2014).

Further, repeats have been analyzed for a specific domain of the family, in which protein function could be found out through the domain (Rentzsch and Orengo, 2013). We found that repeats in the domain regions of the family are involved in the function through functional residues. Earlier, we analyzed

the repeats in the individual proteins of PDB and found that the existence of repeats in single/two domains from the same family, for the function of the proteins and that are not in the domains, are also involved in the function of the proteins (Mary Rajathei and Selvaraj, 2013). We observed that the lengths of repeats in the domains of the family are not uniform. Further, the computation of sequence identity of the repeats within a protein and within a family of Peptidase S8 domain shows lower similarity, which may be the consequence of their divergences over a period. Earlier, it was observed that repeat proteins are indeed repetitive in their families, exhibiting abundant stretches of short perfect repetitions (Turjanski et al., 2016). The repeats of varying lengths in the structures of the fold, as well as in the functional domains of the family, have suggested that long repeats are considerably diverged and may not be overlapped. However, further studies would be needed to understand the conservation of long repeats of the proteins in the structure and function of the proteins.

Further, we observed the existence of long repeats in all seven enzyme classes of the proteins and are especially more abundant in ligases and isomerases. Among the non-enzyme proteins, long repeats are observed in DNA binding, calcium binding, metal binding and NP binding proteins with NP binding and DNA binding in a greater number of proteins. However, further studies are needed to understand why certain enzyme classes and non-enzyme classes are having long repeats in more numbers. This shows that the occurrence of long repeats, not only serves as modules of large assemblies, but also in the catalytic function or binding of the proteins.

While commenting on the evolution of the well-characterized short tandem repeats in many evolutionary lineages, it has been postulated that repeat-containing proteins are cheap to evolve, rather than the *de nova* sequence evolution, as the repeat units are thermodynamically stable (Andrade et al., 2001; Andersson et al., 2015). Through our analysis, we observed the occurrence of long repeats in the stable folds for different functions of the proteins and suggested that long repeats may play a role in the evolution of proteins with stable folds and novel functions.

CONCLUSIONS

The present large scale study has focused on the presence of long repeats in a non-redundant set of the entire annotated UniProtKB/Swiss-Prot database and reveals that long repeats are found in 23% of the proteins. Regarding their three-dimensional structures, they are found in certain structural folds that are incorporated with repetitive duplication of small stable folds. Further, the long repeats of different lengths within each fold are observed in different structures of the proteins. From a functional perspective, these repeats are found in both enzyme and non-enzyme functions containing proteins. Hence, long repeats may have a role in the evolution of proteins with stable folds and novel functions.

DATA AVAILABILITY STATEMENT

The UniProt annotated sequence files and the RADAR output files were analyzed for this study. Major results are available as **Supplementary Material**.

AUTHOR CONTRIBUTIONS

DR developed the computer programs in perl platform for this study and drafted the manuscript. SP supported to analyze and computation the data. SS conceived the idea and helped in the preparation of the manuscript.

ACKNOWLEDGMENTS

DR acknowledges University Grants Commission, India for a research fellowship through UGC Post-doctoral Fellow for Women Research Grant No. F.15-1/2017-18/PDFWM-2017-18-TAM-44286/(SA-II).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00250/full#supplementary-material

REFERENCES

Andersson, D. I., Jerlström-Hultqvist, J., and Näsvall, J. (2015). Evolution of new functions *de novo* and from preexisting genes. *Cold Spring Harb. Perspect. Biol.* 7:a017996. doi: 10.1101/cshperspect.a017996

Andrade, M. A., Perez-Iratxeta, C., and Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* 134, 117–131. doi: 10.1006/jsbi.2001.4392

Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2014). The Protein Data Bank archive as an open data resource. *J. Comput. Aided Mol. Des.* 28, 1009–1014. doi: 10.1007/s10822-014-9770-y

Biegert, A., and Söding, J. (2008). *De novo* identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics* 24, 807–814. doi: 10.1093/bioinformatics/btn039

Bork, P., and Doolittle, R. F. (1994). Drosophila kelch motif is derived from a common enzyme fold. *J. Mol. Biol.* 236, 1277–1282. doi: 10.1016/0022-2836(94)90056-6

Bourne, Y., Zamboni, V., Barre, A., Peumans, W. J., Van Damme, E. J., and Rougé, P. (1999). Helianthus tuberosus lectin reveals a widespread scaffold for mannose-binding lectins. *Structure* 7, 1473–1482. doi: 10.1016/S0969-2126(00)88338-0

Brych, S. R., Kim, J., Logan, T. M., and Blaber, M. (2003). Accommodation of a highly symmetric core within a symmetric protein superfold. *Protein Sci.* 12, 2704–2718. doi: 10.1110/ps.03374903

Cherney, M. M., Cherney, L. T., Garen, C. R., and James, M. N. (2011). The structures of Thermoplasmavolcanium phosphoribosyl pyrophosphate synthetase bound to ribose-5-phosphate and ATP analogs. *J. Mol. Biol.* 413, 844–856. doi: 10.1016/j.jmb.2011.09.007

Chothia, C., and Murzin, A. G. (1993). New folds for all-beta proteins. *Structure* 1, 217–222. doi: 10.1016/0969-2126(93)90010-E

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., et al. (2017). CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* 45, D289–D295. doi: 10.1093/nar/gkw1098

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Fraser, R. D. B., and MacRae, T. P. (1973). *Conformation in Fibrous Proteins and Related Synthetic Polypeptides*. New York, NY: Academic Press. doi: 10.1016/B978-0-12-266850-0.50022-6

Fu, L., Niu, B., Zhu., Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

George, R. A., and Heringa, J. (2000). The REPRO server: finding protein internal sequence repeats through the Web. *Trends Biochem. Sci.* 25, 515–517. doi: 10.1016/S0968-0004(00)01643-1

Groves, M. R., and Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* 9, 383–389. doi: 10.1016/S0959-440X(99)80052-9

Habchi, J., Tompa, P., Longhi, S., and Uversky, V. N. (2014). Introducing protein intrinsic disorder. *Chem. Rev.* 114, 6561–6588. doi: 10.1021/cr400514h

Heger, A., and Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* 41, 224–237. doi: 10.1002/1097-0134(20001101)41:2<224::AID-PROT70>3.0.CO;2-Z

Hemalatha, G. R., Rao, D. S., and Guruprasad, L. (2007). Identification and analysis of novel amino-acid sequence repeats in Bacillus anthracis str. ames proteome using computational tool. *Comp. Funct. Genomics* 2007:47161. doi: 10.1155/2007/47161

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci U.S.A.* 89, 10915–10919. doi: 10.1073/pnas.89.22.10915

Heringa, J., and Argos, P. (1993). A method to recognize distant repeats in protein sequences. *Proteins* 17, 391–411. doi: 10.1002/prot.340170407

Holm, L., and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233, 123–138. doi: 10.1006/jmbi.1993.1489

Jorda, J., and Kajava, A. V. (2009). T-REKS: identification of Tandem REpeats in sequences with a K-means based algorithm. *Bioinformatics* 25, 2632–2638. doi: 10.1093/bioinformatics/btp482

Jorda, J., and Kajava, A. V. (2010). Protein homorepeats: sequences, structures, evolution and functions. *Adv. Protein Chem. Struct. Biol.* 79, 59–88. doi: 10.1016/S1876-1623(10)79002-7

Kajava, A. V. (2012). Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.* 179, 279–288. doi: 10.1016/j.jsb.2011.08.009

Katti, M. V., Sami-Subbu, R., Ranjekar, P. K., and Gupta, V. S. (2000). Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci.* 9, 1203–1209. doi: 10.1110/ps.9.6.1203

Kim, C., Basner, J., and Lee, B. (2010). Detecting internally symmetric proteins structures. *BMC Bioinformatics* 11:303. doi: 10.1186/1471-2105-11-303

Kobe, B., and Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* 11, 725–732. doi: 10.1016/S0959-440X(01)00266-4

Lang, D., Thoma, R., Henn-Sax, M., Sterner, R., and Wilmanns, M. (2000). Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science* 289, 1546–1550. doi: 10.1126/science.289.5484.1546

Laskowski, R. A., Jabłonska, J., Pravda, L., Vareková, R. S., and Thornton, J. M. (2018). PDBsum: structural summaries of PDB entries. *Protein Sci.* 27, 129–134. doi: 10.1002/pro.3289

Lewis, T. E., Sillitoe, I., Dawson, N., Lam, S. D., Clarke, T., Lee, D., et al. (2018). Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res.* 4, D435–D439. doi: 10.1093/nar/gkx1187

Luo, H., and Nijveen, H. (2014). Understanding and identifying amino acid repeats. *Brief Bioinformatics* 15, 582–591. doi: 10.1093/bib/bbt003

Lupas, A. N., Ponting, C. P., and Russell, R. B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relices of an ancient peptide world? *J. Struct. Biol.* 134, 191–203. doi: 10.1006/jsbi.2001.4393

Malay, A. D., Bessho, Y., Ellis, M. J., Antonyuk, S. V., Strange, R. W., Hasnain, S. S., et al. (2009). Structure of glyceraldehyde-3-phosphate dehydrogenase from the archaeal hyperthermophile *Methanocaldococcus jannaschii*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* 65, 1227–1233. doi: 10.1107/S1744309109047046

Marcotte, E. M., Pellegrini, M., Yeates, T. O., and Eisenberg, D. (1998). A Census of protein repeats. *J. Mol. Biol.* 293, 151–160. doi: 10.1006/jmbi.1999.3136

Mary Rajathei, D., and Selvaraj, S. (2013). Analysis of sequence repeats of proteins in the PDB. *Comput. Biol. Chem.* 47, 156–166. doi: 10.1016/j.compbiolchem.2013.09.001

Mary, R. D., Saravanan, M. K., and Selvaraj, S. (2015). Conservation of inter-residue interactions and prediction of folding rates of domain repeats. *J. Biomol. Struct. Dyn.* 33, 534–551. doi: 10.1080/07391102.2014.894944

McLachlan, A. D. (1983). Analysis in gene duplication repeats in the myosin rod. *J. Mol. Biol.* 169, 15–30. doi: 10.1016/S0022-2836(83)80173-9

Murzin, A. G., Lesk, A. M., and Chothia, C. (1992). Beta-Trefoil fold. Patterns of structure and sequence in the Kunitz inhibitors interleukins-1 beta and 1 alpha and fibroblast growth factors. *J. Mol. Biol.* 223, 531–543. doi: 10.1016/0022-2836(92)90668-A

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453. doi: 10.1016/0022-2836(70)90057-4

Neer, E. J., Schmidt, C. J., Nambudripad, R., and Smith, T. F. (1994). The ancient regulatory-protein family of WD-repeat proteins. *Nature* 371, 297–300. doi: 10.1038/371297a0

Newman, A. M., and Cooper, J. B. (2007). XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* 8:382. doi: 10.1186/1471-2105-8-382

Parry, D. A. (2005). Structural and functional implications of sequence repeats in fibrous proteins. *Adv. Protein Chem.* 70, 11–35. doi: 10.1016/S0065-3233(05)70002-4

Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132, 185–219. doi: 10.1385/1-59259-192-2:185

Pellegrini, M., Marcotte, E. M., and Yeates, T. O. (1999). A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins* 35, 440–446.

Pellegrini, M., Renda, M. E., and Vecchio, A. (2012). Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. *BMC Bioinformatics* 13:S8. doi: 10.1186/1471-2105-13-S3-S8

Piovesan, D., Tabora, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., et al. (2017). DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* 45, D219–D227. doi: 10.1093/nar/gkw1056

Ponting, C. P., and Russell, R. B. (2000). Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *J. Mol. Biol.* 302, 1041–1047. doi: 10.1006/jmbi.2000.4087

Rentzsch, R., and Orengo, C. A. (2013). Protein function prediction using domain families. *BMC Bioinformatics* 14:S5. doi: 10.1186/1471-2105-14-S3-S5

Roche, D. B., Viet, P. D., Bakulina, A., Hirsh, L., Tosatto, S. C. E., and Kajava, A. V. (2017). Classification of β-hairpin repeat proteins. *J. Struct. Biol.* 201, 130–138. doi: 10.1016/j.jsb.2017.10.001

Selvaraj, S., and Rajathei, M. (2017). A web database IR_PDB for sequence repeats of proteins in the Protein Data Bank. *Int. J. Knowl. Discov. Bioinformatics* 7, 1–10. doi: 10.4018/IJKDB.2017070101

Szklarczyk, R., and Heringa, J. (2004). Tracking repeats using significance and transitivity. *Bioinformatics* 20, i311–i317. doi: 10.1093/bioinformatics/bth911

Thornton, J. M., Orengo, C. A., Todd, A. E., and Pearl, F. M. (1999). Protein folds, functions and evolution. *J. Mol. Biol.* 293, 333–342. doi: 10.1006/jmbi.1999.3054

Tompa, P. (2012). Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.* 37, 509–516. doi: 10.1016/j.tibs.2012.08.004

Turjanski, P., Parra, R. G., Espada, R., Becher, V., and Ferreiro, D. U. (2016). Protein repeats from first principles. *Sci. Rep.* 6:23959. doi: 10.1038/srep23959

UniProt Consortium T(2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 46, D158–D169. doi: 10.1093/nar/gkw1099

Uthayakumar, M., Benazir, B., Patra, S., Vaishnavi, M. K., Gurusaran, M., Sureka, K., et al. (2012). Homepeptide repeats: implications for protein structure, function and evolution. *Genomics Proteomics Bioinformatics* 10, 217–225. doi: 10.1016/j.gpb.2012.04.001

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., et al. (2014). Classification of intrinsically disordered regions and proteins. *Chem. Rev*. 114, 6589–6631. doi: 10.1021/cr400525m

Yoder, M. D., Lietzke, S. E., and Jurnak, F. (1993). Unusual structural features in the parallel beta-helix in pectate lyases. *Structure* 1, 241–251. doi: 10.1016/0969-2126(93)90013-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 October 2019
doi: 10.3389/fgene.2019.00934

The TargetMine Data Warehouse: Enhancement and Updates

Yi-An Chen†, Lokesh P. Tripathi*†, Takeshi Fujiwara, Tatsuya Kameyama, Mari N. Itoh and Kenji Mizuguchi**

Laboratory of Bioinformatics, National Institutes of Biomedical Innovation, Health and Nutrition, Osaka, Japan

Biological data analysis is the key to new discoveries in disease biology and drug discovery. The rapid proliferation of high-throughput 'omics' data has necessitated a need for tools and platforms that allow the researchers to combine and analyse different types of biological data and obtain biologically relevant knowledge. We had previously developed TargetMine, an integrative data analysis platform for target prioritisation and broad-based biological knowledge discovery. Here, we describe the newly modelled biological data types and the enhanced visual and analytical features of TargetMine. These enhancements have included: an enhanced coverage of gene–gene relations, small molecule metabolite to pathway mappings, an improved literature survey feature, and *in silico* prediction of gene functional associations such as protein–protein interactions and global gene co-expression. We have also described two usage examples on trans-omics data analysis and extraction of gene-disease associations using MeSH term descriptors. These examples have demonstrated how the newer enhancements in TargetMine have contributed to a more expansive coverage of the biological data space and can help interpret genotype–phenotype relations. TargetMine with its auxiliary toolkit is available at https://targetmine.mizuguchilab.org. The TargetMine source code is available at https://github.com/chenyian-nibio/targetmine-gradle.

Keywords: data warehouse, integrative data analysis, multi-omics data analysis, gene prioritisation, drug discovery, data mining, knowledge discovery

OPEN ACCESS

Edited by:
Shandar Ahmad, Jawaharlal Nehru University, India

Reviewed by:
Hamed Bostan, North Carolina State University, United States
Marco Brandizi, Rothamsted Research (BBSRC), United Kingdom

****Correspondence:***
Yi-An Chen
chenyian@nibiohn.go.jp
Lokesh P. Tripathi
lokesh@nibiohn.go.jp
Kenji Mizuguchi
kenji@nibiohn.go.jp

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *18 June 2019*
Accepted: *05 September 2019*
Published: *09 October 2019*

Citation:
Chen Y-A, Tripathi LP, Fujiwara T, Kameyama T, Itoh MN and Mizuguchi K (2019) The TargetMine Data Warehouse: Enhancement and Updates. Front. Genet. 10:934. doi: 10.3389/fgene.2019.00934

INTRODUCTION

The rapid proliferation of high-throughput omics technologies has revolutionised biological research by significantly adding new omics data. However, as the experimental datasets increase in size and complexity, extraction of meaningful biological knowledge becomes qualitatively more difficult, expensive and labourious. Therefore, there is an ever widening gulf between data generation and the rate at which it can be properly analysed (Greene et al., 2014). Proper mining and curation of large biological datasets are necessary to develop an improved understanding of living systems and of disease pathogenesis.

An integrative multi-omics approach combines different types of biological data into a single analytical framework to understand the relationships between different cellular components (Zhu et al., 2012; Yan et al., 2018). Such analyses are useful to develop analytical models that can interpret genotype–phenotype relationships, garner the knowledge of pathways involved in cellular events and diseases, help pinpoint targets (such as gene and proteins) of biological and therapeutic interest and potentially develop intervention methods than can counteract undesirable phenotypic progression (i.e. diseases) (Sun and Hu, 2016; Hasin et al., 2017).

A major challenge in multi-omics data analysis is the availability of clean and usable biological data. We have previously developed TargetMine, an integrated data warehouse based on the object-oriented InterMine data warehouse framework (Smith et al., 2012; Kalderimis et al., 2014; Triplet and Butler, 2014), which models biological entities (such as genes and proteins) as 'objects' that are described by a set of attributes and their relationships with other objects are modelled as 'references'. The InterMine system allows for integration of different types of biological databases, and it comes pre-equipped with data integration features that are able to directly parse the data from commonly used data formats and sources (such as UniProt, OBO, FASTA and BioPAX). InterMine also allows the users to design their own data parsers (Smith et al., 2012; Lyne et al., 2007; Kalderimis et al., 2014). The TargetMine data model was developed by extending a customised version of the core InterMine data model. When integrating similar types of data from heterogeneous data sources, we first identified common attributes (gene identifiers for instance) which are then used to merge the overlapping datasets into a suitable data model. The data sources are prioritised based on their reliability, and the stored identifiers are constantly revised to update or discard outdated identifiers with every database update (Lyne et al., 2007; Smith et al., 2012).

TargetMine was initially developed and optimised for target discovery and prioritisation of candidate genes, especially in early stage drug discovery (Chen et al., 2011). We have continued to make significant additions and refinements to the TargetMine system to transform TargetMine into an integrative data analysis platform that can more effectively interpret information-rich omics datasets for biological knowledge discovery (Chen et al., 2019). Aside from periodically updating the existing datasets, these new developments have involved assimilation of newer biological data types and a new auxiliary toolkit to assist with data analysis and visualisation that addresses the limitations in the core InterMine framework (Chen et al., 2016).

Here, we describe our progressive efforts to enhance TargetMine as a data analysis platform that can better assist multi-omics data analysis and biological knowledge discovery especially in disease biology. These efforts broadly fall into three categories: (1) upgrading the existing data types with the up-to-date information available from the source repositories; (2) assimilating new data types, especially those data types that help to examine different types of gene-gene relations; and (3) augmenting the auxiliary toolkit to better analyse and visualise biological data.

We will now describe our efforts below individually.

ADDITIONAL DATA SOURCES AND DATA MODELS IN TARGETMINE PROVIDE A DEEPER COVERAGE OF THE BIOLOGICAL DATA SPACE

A comprehensive coverage of the biological data space is necessary for drug discovery and related research. To achieve this, we have continuously expanded the repertoire of data types in TargetMine. Since the last major release, we have included within TargetMine new biological data associated with three major areas—drug-target interactions, gene-disease associations and biological mechanisms. The inclusions of these data types have offered deeper insights into gene-gene relations and have also enabled the users to perform more probing biological queries with TargetMine (**Table 1**). To enable the user to quickly and easily perform complex queries, TargetMine contains a library of 'templates' that consist of predefined queries with a simple form and description and are categorised by data types (Chen et al., 2011; Chen et al., 2016; Chen et al., 2019).

KEGG Relations

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a collective repository of genes, genomes, pathways, diseases and chemical compounds that provides a comprehensive

TABLE 1 | Key enhancements and updates in TargetMine since the last published iteration (2016).

Data types, data models and features	New and/or enhanced data types and features	Existing data types and features
Protein–protein interactions	KEGG relations; Post-translation modifications (phosphorylation); PSOPIA-likelihood scores for all PPIs; Gene co-expression scores for HCDPs (GCE-HCDP)	Combined PPI repository from iRefindex and BioGRID, literature; Classification of PPIs as HCs and HCDPs
Metabolomics	KEGG COMPOUND – pathway mapping; KEGG reactions	KEGG COMPOUND
Gene-disease relations	ClinVar variations; dbSNP publications; DisGeNET associations	GWAS data from NHGRI
Literature mining	MeSH term descriptors; Publication abstracts	NCBI PubMed links
TF-target interactions	~400,000 human and mouse TF-target annotation from ENCODE	Amadeus; ORegAnno; HTRIdb
TargetMine auxiliary toolkit		
Composite interaction network	Filter PPIs by HCDPs; Filter interaction types by expressed-tissues and GEL; Add directed PPIs from KEGG; Restrict interaction types to within the user-supplied gene list	Include multiple interaction types
Association heatmap	Dendrogram of hierarchically assembled associations with distances; Expressed-tissue feature	Two-colour grid of squares

mapping of the biological systems (Kanehisa et al., 2016). The relation element in KEGG typically specifies relationship between two entities (proteins and compounds) in KEGG pathways. KEGG relations largely correspond to signalling pathway maps and encode regulatory information such as 'A activates B', 'A inhibits B' and 'A phosphorylates B'. The inclusion of KEGG relations is useful, since they often provide an additional context to interactions between gene products that are not always evident from standard PPI analysis. This integration has enabled the users to reconstruct probable signal transduction paths by performing queries such as 'Given a pair of genes A and B, find an intermediate gene and relations from gene A to gene B'.[1]

Post-Translational Modifications

Post-translational modifications (PTMs) are events that involve covalent addition of functional groups to proteins or their proteolytic processing during and after their biosynthesis. PTMs amplify the functional diversity of proteins and expand their influence over various cellular processes. Therefore, identifying and understanding PTMs help in a deeper understanding of cellular functions and in disease biology (Mnatsakanyan et al., 2018; Thygesen et al., 2018). We retrieved PTM associations from PhosphoSitePlus (Hornbeck et al., 2015), a knowledgebase of mammalian PTMs, and we carefully parsed the UniProt sequence annotation (features) that describe regions or sites of interest in proteins, to create an integrated repository of PTMs in TargetMine. The inclusion of PTMs in TargetMine enables the users to perform complex queries such as 'Given a list of proteins, identify upstream kinases that may phosphorylate them'[2] or 'Given a protein and a specific residue position, identify any PTMs mapped to that residue'.[3]

KEGG Reactions Compound-Pathway Mappings

Metabolites are the low molecular weight compounds such as amino acids, sugars and lipids, which are typically substrates and by-products of biological processes and enzymatic reactions; they are widely involved in feedback regulatory processes in the cell and, being the downstream products, often directly influence the phenotype. Thus, the metabolome is often regarded as the link between genotype and phenotype (Krumsiek et al., 2016). To facilitate a more effective metabolomics analysis with TargetMine, we first extended the existing compound class to create a new KEGG COMPOUND class. Subsequently, we referenced the KEGG COMPOUND class both with the existing Enzyme and Pathway classes using the relationships extracted from the KEGG COMPOUND database. We also defined a new Reaction class to describe the biochemical reactions in the KEGG reaction database, and this class was referenced with all of the KEGG COMPOUND, pathway and enzyme classes. Given a list of compounds, the users can now retrieve the corresponding enzymes involved in their metabolism, the enzymatic reactions involving these metabolites and, even map them to the corresponding pathways and diseases associated with the pathways. Users can also perform enrichment analyses to prioritise the enzymes/genes and pathways specifically associated with their metabolites of interest (see example below).

Disease-Gene Mappings

A deeper understanding of disease pathogenesis requires a mapping of links between genes, pathways and specific diseases, but they are difficult to obtain in general. Recently, we have enhanced the integration of genetic linkages to diseases by improving the existing GWAS data model and adding the variation annotations from ClinVar (Landrum et al., 2015), and the disease associations that were extracted from the associated publications in dbSNP (Chen et al., 2019). We have also included gene-disease associations compiled in DisGeNET, an integrative platform of curated gene-disease associations (Pinero et al., 2017). This integration has enabled the users to perform queries such as 'Given a gene, find the related SNPs and any diseases associated with these SNPs'.[4]

Scientific Literature Survey

Literature survey is indispensable to annotating gene information, interpreting gene sets and facilitating further research. However, scientific literature is increasing exponentially, making it difficult for the researchers to find, study and understand new publications of interest. To facilitate an easier sharing of scientific knowledge, we have incorporated document representations such as MeSH (Medical Subject Headings) (Rogers, 1963) descriptors (such as general article, review, clinical study, case report, etc.) and abstracts into TargetMine. This implementation allows the users to quickly screen for scientific texts (based on their MeSH descriptors) associated with their gene(s) of interest. For example, users may restrict their query to retrieving only those publications classified as 'case report' by constraining the 'Mesh Terms'-> 'Identifier' attribute for the MeSH Terms identifier 'D002363' (case reports). Researchers typically rely on abstracts to assess an article for further reading and often; abstracts are the only source of information that are freely available (Germini et al., 2017). To allow the users to easily access and scan article abstracts of interest, we leveraged the attribute 'Abstract Text' within the 'Publication' class. This implementation allows the users to retrieve publications associated with their gene(s) of interest along with their corresponding abstracts; this implementation also allows the users to quickly and easily scan multiple abstracts in a single webpage, instead of visiting the individual 'Publication report' pages and clicking on the available PMID links to access the corresponding abstracts on NCBI PubMed (as was the case previously).

INCLUSION OF COMPUTATIONALLY PREDICTED ASSOCIATIONS AND SCORES IN TARGETMINE

TF-Target Associations

Transcription factor (TF)–target gene interactions determine gene expression patterns, and therefore, regulate cellular functions. Previously, we had included expert-curated experimentally validated human TF-target gene interaction data from Amadeus

(Linhart et al., 2008), ORegAnno (Griffith et al., 2008) and HTRIdb (Bovolenta et al., 2012) to create a combined repository in TargetMine (Chen et al., 2016). To expand the coverage of TF-target gene interactions, we examined and processed the vast amounts of TF-binding site data compiled by the Encyclopedia of DNA Elements (ENCODE) consortium (see *Methods*). Over 200,000 new TF-target gene interactions corresponding to 23 human TFs were incorporated into TargetMine in this manner. We also incorporated nearly 200,000 TF-target gene interactions corresponding to 39 TFs in the mouse genome, thereby providing a detailed coverage of gene-regulatory associations in mouse that were not available in the previous iterations of TargetMine.

PPI Confidence Scores Using Predicted Likelihood of PPIs

PPIs are vital to virtually every cellular process, and their dysregulation typically leads to cellular dysfunction including diseases. However, it is necessary to assess the PPI data properly to ensure the robustness of PPIN-based analyses in investigating disease-causing biological pathways and to discover druggable target proteins. We had previously performed a confidence assessment of our combined PPI repository and defined a reliable high-quality subset termed 'high-confidence direct physical PPIs' (HCDPs) (Chen et al., 2016). HCDPs have been helpful in analysing network topological properties and identifying key components of the presently characterised interactome maps, namely, network 'hubs' and 'bottlenecks' (Tripathi et al., 2013; Chen et al., 2016). However, HCDPs constitute only a small proportion of all available PPIs, and using HCDPs alone for PPI-based network analysis may often exclude potentially useful PPIs. This is especially true for the mouse and rat interactomes in TargetMine, where HCDPs are rather sparse. Therefore, we have included an additional measure for the assessment of PPI reliability. In our group, we had previously developed prediction server of protein–protein interactions (PSOPIA), an integrative averaged one-dependence estimators (AODE)–based method to predict the likelihood of interaction between a pair of proteins based on experimentally characterised homologous PPIs (Murakami and Mizuguchi, 2014). We employed PSOPIA to evaluate all PPIs within TargetMine, and the output PSOPIA scores were tagged to the individual PPIs as a new attribute PsopiaScore. This implementation has enabled the users to query the interacting partners of a gene/protein or a list of genes/proteins of interest and to infer overall PPI networks involving these genes/proteins consisting of all interactions judged to be of sufficiently high quality by the user, either based on their HCDP status and/or their PSOPIA scores.

Gene Co-Expression Analysis for Prediction of Novel DNA-Binding Proteins and for Improved PPI-Based Network Analysis

Gene expression refers to the process where the genetic information encoded in a gene is transcribed into a functional gene product—RNA or the eventual protein. Gene expression analysis involves mapping and analysing collective gene expression patterns that dictate cellular function under different environments. The spread of technologies that can map global gene expression profiles has led to an abundance of genome-wide gene expression data (transcriptome) for many cell and tissue types and physiological conditions (Barrett et al., 2012; Kolesnikov et al., 2015). Global gene expression profiles have been variously analysed to search for genes that are differentially expressed in different cellular and physiological conditions (such as development, infection and diseases) and to predict functions for genes of unknown function. A guiding principle of function prediction using gene expression is *guilt-by-association*, which assumes that genes with related functions are more likely to have correlated properties such as interactions and expression patterns (Singer et al., 2004).

In this study, we have sought to leverage global gene co-expression (GCE) patterns to minimise biological noise and to further refine and improve PPI-based network analysis (**Figure 1**). To assess the effectiveness of this approach, we performed GSFE analysis on a multiple gene sets gathered from literature and then repeated the tests on modified gene sets that included both co-expressed genes and randomly selected unrelated genes (biological noise) (see methods). Among the 298 curated gene sets that were tested, 81% (240) were associated with overall higher F1 scores when HCDPs or globally co-expressed HCDPs (GCE-HCDP) where added to the initial gene list, compared with initial gene sets (**Figure 2**; **Table 2**, **S1**). Furthermore, of the 240 gene sets where GCE showed an improved prioritisation performance, in 170 of them (~70%), inclusion of GCE-HCDP contributed to a better performance in terms of F1 scores as compared to inclusion of HCDPs alone (**Figure 2**; **Table 2**).

Our observations suggested that the inclusion of HCDPs and GCE-HCDPs contributed to improved target prioritisation and gene set analysis with TargetMine.

ENHANCED DATA ANALYSIS AND VISUALISATION WORKFLOW WITH TARGETMINE

We had previously developed an auxiliary toolkit to assist with data analysis and visualisation in TargetMine without any scripting and/or programming efforts on the part of the user (Chen et al., 2016) (https://targetmine.mizuguchilab.org/tutorials/auxiliary-toolkit/). We have subsequently added new analytical and visualisation features to further enhance the ability of TargetMine as a data analysis platform. For instance, we have now added a dendrogram to the association heatmap function, which permits users to quickly and more easily identify clusters of genes that share significant functional attributes. We have also introduced the 'Expressed Tissue' feature that allows the users to hierarchically assemble a heatmap of user-supplied genes and the cell/tissues where they are highly expressed and thereby obtain a contextual view of their expression patterns. We have also enhanced the efficacy of the network visualisation function. In the present form, the function would permit the users to supply a list of genes and construct and visualise a composite interaction network that

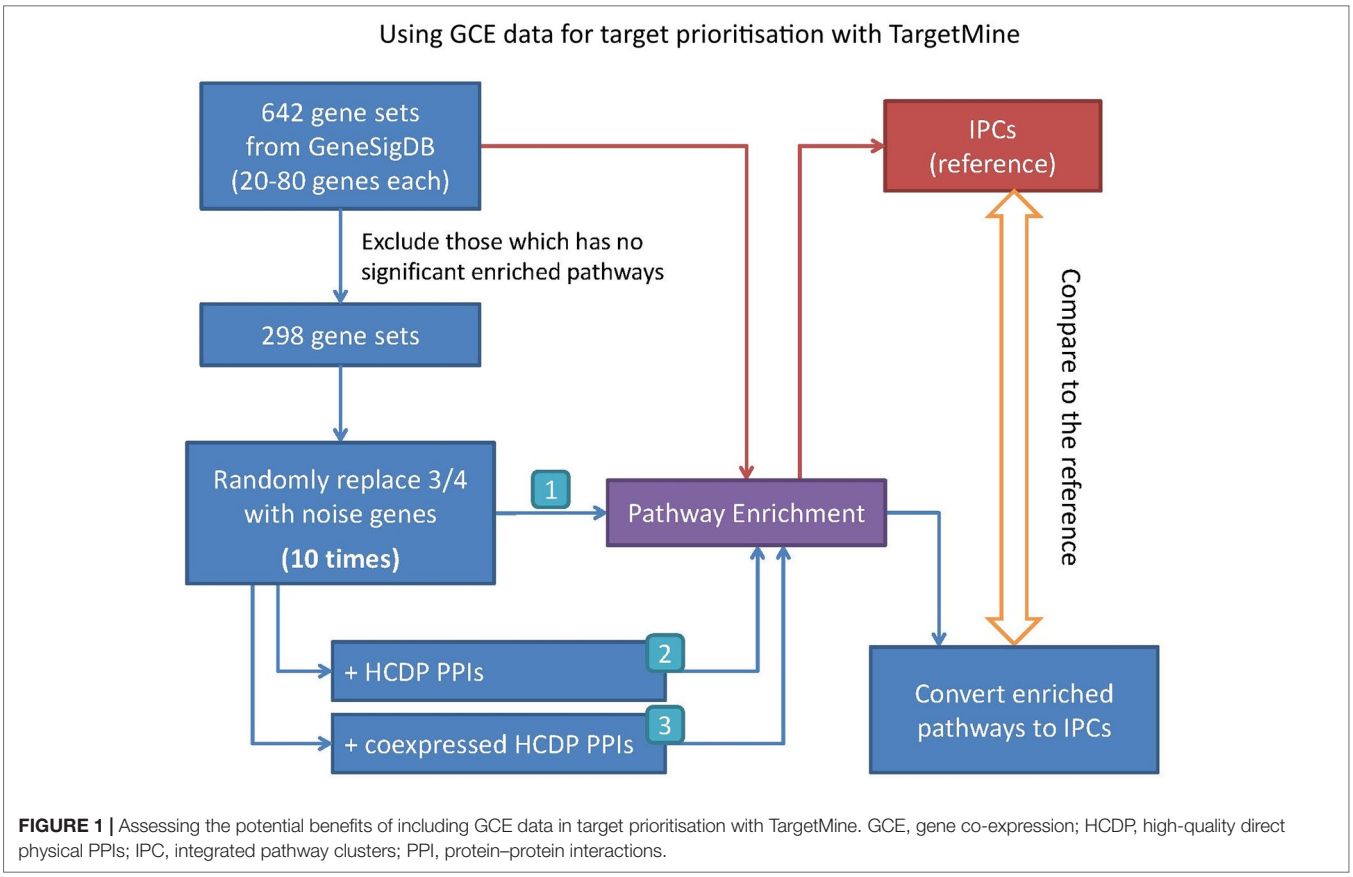

FIGURE 1 | Assessing the potential benefits of including GCE data in target prioritisation with TargetMine. GCE, gene co-expression; HCDP, high-quality direct physical PPIs; IPC, integrated pathway clusters; PPI, protein–protein interactions.

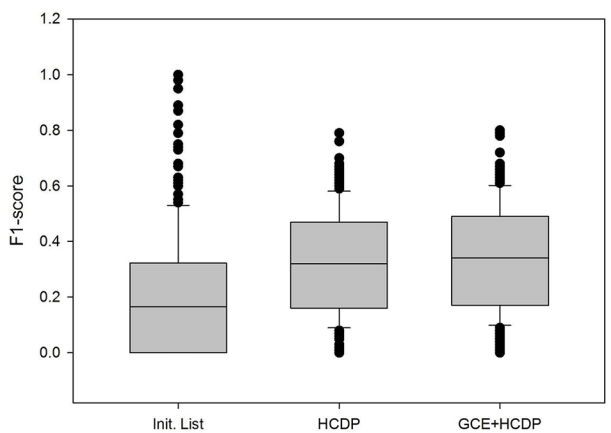

FIGURE 2 | The inclusion of HCDPs filtered by Gene co-expression (GCE-HCDP) to generate extended gene sets led to an overall improved target prioritisation performance when compared with the inclusion of unfiltered HCDPs and un-extended gene sets.

TABLE 2 | The inclusion of HCDPs filtered by gene co-expression (GCE-HCDP) to generate extended gene sets led to an overall improved target prioritisation performance when compared with the inclusion of unfiltered HCDPs and un-extended gene sets.

a.	**Average F1-score**	
Original test	0.211	
+HCDP	0.327	
+Co-exp-HCDP	0.341	
b.	**T-test**	
	Original test	**+HCDP**
+HCDP	5.77×10^{-18}	
+Co-exp-HCDP	5.93×10^{-22}	6.14×10^{-20}

includes all the biomolecular interactions within TargetMine, i.e. PPIs, MTIs, PCIs/drug-target interactions and TF-target gene interactions that are associated with the query genes. However, adding too many interactions can also render the network very dense and complex, therefore becoming difficult to load and visualise properly in the browser. To address these concerns, we have added a series of features to select and filter biomolecular interactions by qualitative assessment and/or contextual information. For instance, the 'Interaction Network' feature allows the user to restrict the PPI selection to 'HCDPs' or expand them to include 'All' PPI types by selecting the corresponding circles. We have also introduced features that permit the users to filter the interaction types by expressed tissues and GELs. We have also introduced a feature to allow the users to specifically include and visualise directed gene–gene relations parsed from KEGG. Moreover, we have improved the network feature to restrict the MTIs,

PCIs/drug-target interactions and TF-target gene interactions to the user supplied gene list.

APPLICATIONS WITH USE CASES

Trans-Omics Data Analysis

To demonstrate the effectiveness of TargetMine in assisting multi-omics data analysis, we re-examined a previously published multi-omics data on mitochondrial links to liver metabolism, and the effects of a high-fat diet on it (Williams et al., 2016). We first retrieved the biomolecules (110 transcripts, 27 proteins and 25 metabolites) that were differentially expressed in high-fat diet (HFD) fed mice relative to control (see methods). Next, the differentially expressed transcript, protein and metabolite sets were transformed into the corresponding transcriptome, proteome and metabolome differentially expressed gene (DEG) sets, respectively (see **Supplementary File S1** for the detailed methodology with the help of an example). The inferred DEG sets (containing 84, 29 and 62 genes, respectively; **Supplementary Table S1**, **Supplementary Figure S1**) were first compared (using the list operations in TargetMine) to identify overlapping genes. Three DEGs (Ces2a, Cyp3a11 and Csad) were shared across transcriptome and proteome DEG sets, and a solitary gene, cysteine sulfinic acid decarboxylase (Csad), was downregulated across all the three DEG sets (**Supplementary File S1B**). Csad is an enzyme that plays a key role in generating taurine from cysteinesulfinate in liver, and its hepatic expression and abundance are typically downregulated by bile acids responsible for modulating lipid metabolism (Kerr et al., 2014). Our observations, therefore, clearly suggested that the fluctuations in Csad levels were likely to be modulated by dietary fat *via* bile acids. Additionally, we were also able to prioritise signatures that were consistent with an HFD model such as cytochrome p450 subunit Cyp3a11 that functions in retinoic acid metabolism.

Next, the individual DEG sets were then subjected to pathway enrichment analysis (see methods). Five enriched pathways were associated with the transcriptome DEG set, 11 enriched pathways with the proteome DEG set, and 17 enriched pathways were associated with the metabolome DEG set, respectively. KEGG pathway sub-categories 'Lipid metabolism', 'Global and overview maps', 'Cancers: Overview' and 'Metabolism of cofactors and vitamins' were commonly represented in the enriched pathways across all the three DEG sets (**Supplementary File S1B**). Specifically, KEGG pathway 'Metabolic pathways' was commonly enriched in all the three DEG sets; 'Steroid hormone biosynthesis', 'Linoleic acid metabolism', 'Retinol metabolism', 'Arachidonic acid metabolism' and 'Chemical carcinogenesis' were commonly enriched in Transcriptome and Proteome DEG sets (associated with Cyp3a11), and 'Drug metabolism—other enzymes' was commonly enriched in Proteome and Metabolome DEG sets, respectively (although no gene overlap was observed). Taken together, our observations have suggested that higher levels of dietary fat are responsible for dysregulation of cellular factors and pathways associated with lipid metabolism and as such have provided promising candidates for further research.

Extracting Gene-Disease Associations From Literature Using Mesh Descriptors

A vast amount of untapped associations between genes and diseases are scattered across biomedical literature. A quick and efficient mining of such information can help interpret genotype-phenotype relationships and also speed up database curation. The inclusion of MeSH descriptors now allows the TargetMine users to easily survey for annotated gene-disease associations for their gene(s) of interest. As a case study, we extracted literature-embedded gene-disease associations for PPARγ (peroxisome proliferator–activated receptor gamma), a nuclear receptor that is implicated in the pathology of numerous diseases including obesity, diabetes and cancer. Next, we sought to retrieve all the publications that were indicative of the involvement of PPARγ in disease pathogenesis by constraining the query for MeSH term attribute 'Diseases'. We retrieved 397 unique disease associations for PPARG in this manner (**Supplementary File S2**).

CONCLUSIONS

TargetMine is a versatile data analysis platform that provides a unified, homogenous representation of diverse types of omics and other biological data; it allows the users to query and navigate across the stored data types and analyse them in a singular interface. In this study, we have described the augmentation of the TargetMine by progressively improving and expanding the coverage of data types and by adding new and improved analytical features. We have also demonstrated how the extension of TargetMine system has significantly boosted its capabilities to survey the biological target space, to assist multi-omics data analysis, to interpret novel genotype-phenotype relationships and to facilitate biologically relevant knowledge discovery, especially in disease biology.

For the future developments, we will continue to accommodate new and emerging data types of interest and expand the analytical features. We also aim to introduce a workflow function for multi-omics data analysis that will allow the users to more easily and effectively analyse and interpret their omics datasets and advance their research.

METHODS

Tf-Target Gene Associations From Encode

The binding events (peaks) for human and mouse TFs with binding profiles in different cell types were downloaded from the ENCODE resource (Davis et al., 2017). To accommodate the additional TF-target information, we redefined the erstwhile protein-DNA interaction class into a new transcriptional regulation class. The promoter region was defined as 10,000-bp upstream of the transcriptional start site (TSS). We extracted binding site positions within this hypothesised promoter region, and we identified the corresponding genes

by mapping the genomic coordinates downstream of the TSS to the genomic coordinates stored within TargetMine. Next, we mapped the TFs whose binding sites were identified in this manner with the downstream genes to generate new TF-target gene associations.

Gene Co-Expression Analysis

Data Sets

The gene sets were retrieved from GeneSigDB, a database of curated gene signatures (Culhane et al., 2012). To obtain a reasonable size of candidates, we selected only the human, mouse and rat gene sets that consisted of 30~80 genes; 642 gene sets were selected in this manner. The genes in the so-called 'standardised' gene list in GeneSigDB were represented by Ensembl identifier and symbol; for further analysis, we mapped them to Entrez Gene identifier (Gene ID) using TargetMine (build 20160629). The Ensembl identifiers that were not mapped to a corresponding Gene ID were excluded from the list, thereby marginally reducing the sizes of the gene sets. Next, we performed pathway enrichment analysis on each of these gene sets, and only those gene sets (298 out of 642) that were associated with at least one enriched pathway (pathways were judged to be significant if the adjusted *p*-value was 0.05 or less) were taken up for subsequent analyses (**Figure 1**).

Global Gene Co-Expression Analysis

Global gene expression profiles for human genes were retrieved from gene expression omnibus (GEO) (Barrett et al., 2012; Kolesnikov et al., 2015), and gene co-expression levels were computed as described earlier (Ahmad et al., 2018).

Gene Prioritisation With PPI and GCE

The aim of the gene prioritisation is to identify a relatively important subset of genes form a list of candidates for further analysis. The first step of target prioritisation within TargetMine involves uploading a list of initial candidate genes or proteins (e.g. a set of differentially expressed genes or a set of proteins that interact with a given protein) to create a TargetMine gene list. Next, the enrichment of specific biological themes such as KEGG/Reactome pathways, integrated pathway cluster (IPC) (Chen et al., 2014), gene ontology terms etc. is estimated by performing Fisher's exact test here followed by multiple testing correction to control the false discovery rate. The genes associated with the most significantly enriched biological associations (that satisfied, in this instance, a condition of $p \leq 0.05$ after a multiple test correction with the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995; Benjamini et al., 2001)) are judged to be highly important to the biological phenomenon under study and therefore selected for further analyses.

For each of 298 gene sets, we replaced at random 75% genes with an equal number of unrelated randomly selected genes from the corresponding genome to generate test gene sets to incorporate biological noise. To avoid any bias incurred due to the selection of random genes, the process was repeated 10 times to infer 10 test gene sets for each curated gene list. Next, the HCDPs for the genes within the test gene sets were retrieved from TargetMine and were appended to the initial test gene sets to create extended test gene sets. Independently, co-expressed HCDPs for the test gene sets were retrieved from TargetMine and were appended to the initial test gene sets to create extended test gene sets. Only the interacting partners that had a GCE value greater than 0.03 or less than −0.03 with the genes in the test gene sets were considered. Finally, the prioritisation tests (**Figure 1**) were then performed for each test gene set.

Evaluating the Performance of GCE-Filtered HCDPs in Target Prioritisation

To evaluate the protocols, we compared the enriched pathways among the reference gene sets, the 'noisy' gene sets and the extended noisy gene sets that were independently generated with the inclusion of HCDPs and GCE-HCDPs.

The enriched pathways (p-value < 0.05) were then mapped to their corresponding IPC (Chen et al., 2014). These enriched IPCs from the reference gene set were defined as the true positives (TPs) in this instance, and the rest were defined as false positives (FPs). TPs which were not found in the test results were defined as false negatives (FNs). For each gene set, the F1-score was estimated as follows:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

The test was performed 10 times for each gene set (from the step that we randomly generated the 'noisy' gene set). A student t-test was also performed to compare the significance of the differences between the two approaches. If the new prioritisation protocol achieved an overall higher F1-score than standard enrichment analysis, it was assumed to have provided an improved prioritisation performance, even the difference was trivial.

Gene Set Inference and Pathway Enrichment for Multi-Omics Data Analysis

The biomolecules (transcripts/genes, proteins and metabolites) were judged to be differentially expressed if they were statistically significantly ($p \leq 0.05$; t-test) increased or decreased more than 1.5-fold i.e. if the fold change (FC) ≥1.5 (upregulated) or FC ≤ 0.667 (downregulated) in mice fed with high-lipid diet relative to the control. The biological pathway data from KEGG were used to assign functional annotations to the DEGs, using TargetMine. Statistical significance of the pathway enrichment was determined by Fisher's exact test, and the p-values were corrected for multiple testing using the Benjamini–Hochberg procedure. The enriched pathways were considered statistically significant if the adjusted $p \leq 0.05$.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the manuscript/**Supplementary Files**.

AUTHOR CONTRIBUTIONS

Y-AC and LT contributed equally to this work. Y-AC, LT, TF, TK, MI, and KM were responsible for data gathering and validation. Y-AC, LT, TF and TK performed all the data analysis. Y-AC, LT, and KM were responsible for overall data analysis and interpretation and wrote the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This work was in part supported by Grants-in-Aid for Scientific Research from the Japan Agency for Medical Research and Development (Grant Number 19ak0101068h0003; "The adjuvant database project" Grant Number 16ak0101010h0005) and from the Japan Society for the Promotion of Science (Grant Number 17K07268) to KM.

ACKNOWLEDGMENTS

The authors thank the members of the Mizuguchi laboratory for their critical review of the study and the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00934/full#supplementary-material

REFERENCES

Ahmad, S., Prathipati, P., Tripathi, L. P., Chen, Y. A., Arya, A., Murakami, Y., et al. (2018). Integrating sequence and gene expression information predicts genome-wide DNA-binding proteins and suggests a cooperative mechanism. *Nucleic Acids Res.* 46, 54-70. doi: 10.1093/nar/gkx1166

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate- a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.* 57, 289-300.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* 125, 279-84. doi: 10.1016/s0166-4328(01)00297-2

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* 13, 405. doi: 10.1186/1471-2164-13-405

Chen, Y. A., Tripathi, L. P., and Mizuguchi, K. (2011). TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One* 6, e17844. doi: 10.1371/journal.pone.0017844

Chen, Y. A., Tripathi, L. P., and Mizuguchi, K. (2016). An integrative data analysis platform for gene set analysis and knowledge discovery in a data warehouse framework. *Database (Oxford)* 2016, baw009. doi: 10.1093/database/baw009

Chen, Y. A., Yogo, E., Kurihara, N., Ohno, T., Higuchi, C., Rokushima, M., et al. (2019). Assessing drug target suitability using TargetMine. *F1000Research* 8, 233. doi: 10.12688/f1000research.18214.2

Chen, Y.-A., Tripathi, L. P., and Mizuguchi, K., (2019). "Data warehousing with TargetMine for omics data analysis," in *Microarray Bioinformatics*. Eds. V. Bolón-Canedo and A. Alonso-Betanzos (New York, NY: Springer New York), 35–64. doi: 10.1007/978-1-4939-9442-7_3

Chen, Y. A., Tripathi, L. P., Dessailly, B.H., Nystrom-Persson, J., Ahmad, S., and Mizuguchi, K. (2014). Integrated pathway clusters with coherent biological themes for target prioritisation. *PLoS ONE* 9, e99030. doi: 10.1371/journal.pone.0099030

Culhane, A. C., Schroder, M. S., Sultana, R., Picard, S. C., Martinelli, E. N., Kelly, C., et al. (2012). GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res.* 40, D1060–D1066. doi: 10.1093/nar/gkr901

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2017). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801. doi: 10.1093/nar/gkx1081

Germini, F., Marcucci, M., Fedele, M., Galli, M. G., Mbuagbaw, L., Salvatori, V., et al. (2017). Quality of reporting in abstracts of RCTs published in emergency medicine journals: a protocol for a systematic survey of the literature. *BMJ Open* 7, e014981. doi: 10.1136/bmjopen-2016-014981

Greene, C. S., Tan, J., Ung, M., Moore, J. H., and Cheng, C. (2014). Big data bioinformatics. *J. Cell. Physiol.* 229, 1896–1900. doi: 10.1002/jcp.24662

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., et al. (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.* 36, D107–D113. doi: 10.1093/nar/gkm967

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18, 83. doi: 10.1186/s13059-017-1215-1

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* 43, D512–D520. doi: 10.1093/nar/gku1267

Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., et al. (2014). InterMine: extensive web services for modern biology. *Nucleic Acids Res.* 42, W468–W472. doi: 10.1093/nar/gku301

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kerr, T. A., Matsumoto, Y., Matsumoto, H., Xie, Y., Hirschberger, L. L., Stipanuk, M. H., et al. (2014). Cysteine sulfinic acid decarboxylase regulation: a role for farnesoid X receptor and small heterodimer partner in murine hepatic taurine metabolism. *Hepatol Res.* 44, E218–E228. doi: 10.1111/hepr.12230

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., et al. (2015). ArrayExpress update—simplifying data submissions. *Nucleic Acids Res.* 43, D1113–D1116. doi: 10.1093/nar/gku1057

Krumsiek, J., Bartel, J., and Theis, F. J. (2016). Computational approaches for systems metabolomics. *Curr. Opin. Biotechnol.* 39, 198–206. doi: 10.1016/j.copbio.2016.04.009

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al. (2015). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44, D862–D868. doi: 10.1093/nar/gkv1222

Linhart, C., Halperin, Y., and Shamir, R. (2008). Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.* 18, 1180–1189. doi: 10.1101/gr.076117.108

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., et al. (2007). FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.* 8, R129. doi: 10.1186/gb-2007-8-7-r129

Mnatsakanyan, R., Shema, G., Basik, M., Batist, G., Borchers, C. H., Sickmann, A., et al. (2018). Detecting post-translational modification signatures as potential biomarkers in clinical mass spectrometry. *Exp. Rev. Proteomics* 15, 515–535. doi: 10.1080/14789450.2018.1483340

Murakami, Y., and Mizuguchi, K. (2014). Homology-based prediction of interactions between proteins using averaged one-dependence estimators. *BMC Bioinform.* 15, 213. doi: 10.1186/1471-2105-15-213

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Rogers, F. B. (1963) Medical subject headings. *Bull. Med. Libr. Assoc.* 51, 114–116.

Singer, G. A. C., Lloyd, A. T., Huminiecki, L. B., and Wolfe, K. H. (2004). Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol. Biol. Evol.* 22, 767–775. doi: 10.1093/molbev/msi062

Smith, R. N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F., et al. (2012). InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics* 28, 3163-5. doi: 10.1093/bioinformatics/bts577

Sun, Y. V., and Hu, Y. J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.* 93, 147–190. doi: 10.1016/bs.adgen.2015.11.004

Thygesen, C., Boll, I., Finsen, B., Modzel, M., and Larsen, M. R. (2018). Characterizing disease-associated changes in post-translational modifications by mass spectrometry. *Exp. Rev. Proteomics* 15, 245–258. doi: 10.1080/14789450.2018.1433036

Tripathi, L. P., Kambara, H., Chen, Y. A., Nishimura, Y., Moriishi, K., Okamoto, T., et al. (2013). Understanding the biological context of NS5A-host interactions in HCV infection: a network-based approach. *J. Proteome Res.* 12, 2537–2551. doi: 10.1021/pr3011217

Triplet, T., and Butler, G. (2014). A review of genomic data warehousing systems. *Brief Bioinform* 15, 471–483. doi: 10.1093/bib/bbt031

Williams, E. G., Wu, Y., Jha, P., Dubuis, S., Blattmann, P., Argmann, C. A., et al. (2016). Systems proteomics of liver mitochondria function. *Science* 352, aad0189. doi: 10.1126/science.aad0189

Yan, J., Risacher, S. L., Shen, L., and Saykin, A. J. (2018). Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform*. 19, 1370–1381. doi: 10.1093/bib/bbx066

Zhu, J., Sova, P., Xu, Q., Dombek, K. M., Xu, E. Y., Vu, H., et al. (2012). Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS Biol.* 10, e1001301. doi: 10.1371/journal.pbio.1001301

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with the authors.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 October 2019
doi: 10.3389/fgene.2019.01041

Paclitaxel Response Can Be Predicted With Interpretable Multi-Variate Classifiers Exploiting DNA-Methylation and miRNA Data

*Alexandra Bomane, Anthony Gonçalves and Pedro J. Ballester**

Cancer Research Center of Marseille, CRCM, INSERM, Institut Paoli-Calmettes, Aix-Marseille Univ, CNRS, Paris, France

To address the problem of resistance to paclitaxel treatment, we have investigated to which extent is possible to predict Breast Cancer (BC) patient response to this drug. We carried out a large-scale tumor-based prediction analysis using data from the US National Cancer Institute's Genomic Data Commons. These data sets comprise the responses of BC patients to paclitaxel along with six molecular profiles of their tumors. We assessed 10 Machine Learning (ML) algorithms on each of these profiles and evaluated the resulting 60 classifiers on the same BC patients. DNA methylation and miRNA profiles were the most informative overall. In combination with these two profiles, ML algorithms selecting the smallest subset of molecular features generated the most predictive classifiers: a complexity-optimized XGBoost classifier based on CpG island methylation extracted a subset of molecular factors relevant to predict paclitaxel response (AUC = 0.74). A CpG site methylation-based Decision Tree (DT) combining only 2 of the 22,941 considered CpG sites (AUC = 0.89) and a miRNA expression-based DT employing just 4 of the 337 analyzed mature miRNAs (AUC = 0.72) reveal the molecular types associated to paclitaxel-sensitive and resistant BC tumors. A literature review shows that features selected by these three classifiers have been individually linked to the cytotoxic-drug sensitivities and prognosis of BC patients. Our work leads to several molecular signatures, unearthed from methylome and miRNome, able to anticipate to some extent which BC tumors respond or not to paclitaxel. These results may provide insights to optimize paclitaxel-therapies in clinical practice.

Keywords: biomarker discovery, machine learning, artificial intelligence, precision oncology, tumor profiling

OPEN ACCESS

Edited by:
Rakesh Kaundal, Utah State University, United States

Reviewed by:
Deepak Singla, Punjab Agricultural University, India
Haiquan Li, University of Arizona, United States

***Correspondence:**
Pedro J. Ballester
pedro.ballester@inserm.fr

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *18 June 2019*
Accepted: *30 September 2019*
Published: *25 October 2019*

Citation:
Bomane A, Gonçalves A and Ballester PJ (2019) Paclitaxel Response Can Be Predicted With Interpretable Multi-Variate Classifiers Exploiting DNA-Methylation and miRNA Data. Front. Genet. 10:1041. doi: 10.3389/fgene.2019.01041

INTRODUCTION

Breast cancer (BC) is the most common type of cancer in women worldwide resulting in half a million deaths annually (Golubnitschaja et al., 2016). BC is a disease presenting substantial inter-tumor heterogeneity (Russnes et al., 2011). Cytotoxic drugs are used to eradicate tumor cells, to complement surgery or radiotherapy as well as to alleviate cancer symptoms. Paclitaxel is a BC-approved cytotoxic drug from the taxane family, which acts by interfering with the normal function of microtubules during cell division (Perez, 1998). As with other cancer drugs (Brown and Böger-Brown, 1999; Cardoso et al., 2002; Ribeiro et al., 2012; Housman et al., 2014), resistance to paclitaxel have been regularly observed in BC patients (Flint et al., 2009; Ajabnoor et al., 2012).

Precision oncology requires predictors to guide the optimization of drug therapies for patients (Peck, 2016; Schwartzberg et al., 2017). Indeed, it is now well-established that gene polymorphisms and other genomic alterations play important roles in the observed heterogeneous response to drugs (Wang et al., 2011; Harper and Topol, 2012; Kadra et al., 2012). This has led to the identification of clinical biomarkers of drug response from molecular profiles of the patients' tumors (Huang et al., 2014). These predictive biomarkers now guide patient-specific treatment selection during clinical trials and are also used in clinical practice (Mandrekar and Sargent, 2009; Biankin et al., 2015). Most commonly, single-gene markers are used to discriminate between therapy responders and non-responders (Prahallad et al., 2012; Rodríguez-Antona and Taron, 2015), typically consisting of an actionable mutation (e.g. single-nucleotide variant) of a specific gene in the tumor sample.

Single-gene markers that are able to predict the efficacy of cytotoxic drugs are rare (Felip and Martinez, 2012), especially for taxanes (Murray et al., 2012; Bartlett et al., 2015; Norimura et al., 2018). For instance, Marsh et al. (2007) have proposed that the point mutation *CYP1B1*3* could be an important factor that helps to differentiate between sensitive and resistant BC patients to paclitaxel. However, Gehrmann et al. (2008) have raised doubts about the association between this alteration and paclitaxel-treated patient prognosis, concluding that CYP1B1 alone is not sufficient to predict tumor response to paclitaxel, and that it could interact with still unknown factors involved in paclitaxel sensitivity. This is an example of inter-patient variability in drug response not being fully captured by the mutational status of single gene, as it has also been seen *in vitro* in a range of drugs (Naulaerts et al., 2017).

Machine Learning (ML) can be used to build *in silico* models able to predict tumor response to a given drug by combining multiple tumor features in an optimal manner (Libbrecht and Noble, 2015; Ali and Aittokallio, 2018). The scarcity of suitable clinical data to build such predictors has been a major roadblock, which has made predictors based on cancer cell line data thrive (Costello et al., 2014; Ding et al., 2016). Fortunately, response data from paclitaxel-treated BC patients along with comprehensive molecular profiles of their tumors are increasingly available. Such datasets represent an opportunity to improve our ability to anticipate which BC patients will respond to paclitaxel. We obtained them from the recently created Genomic Data Commons (GDC) of the US National Cancer Institute (NCI) (Jensen et al., 2017). The GDC provides a unified data repository enabling data sharing across cancer genomic studies in support of precision medicine. The GDC feeds from several cancer genome programs at the NCI Center for Cancer Genomics, notably The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013), and offers a range of information-rich genomic, transcriptomic and epigenomic profiles, as well as clinical drug response data.

These datasets, however, pose the challenge of being high-dimensional. Each profile typically contains between hundreds and many thousands of features, but only tens of profiled tumors of the same cancer type and treated with the same drug. For example, a community challenge intended to predict drug response employed 53 BC cell lines (Costello et al., 2014), while thousands of features from DNA copy-number variation, transcript expression, mutations, DNA methylation, and protein abundance profiles were considered. In another study (Tripathi et al., 2016), predictive models of response to cytotoxic drugs were achieved using 60 pancancer cell lines and gene variants as features. A further example of predictive drug-sensitivity models is a study employing 60 diverse cell lines and protein abundances as features (Ma et al., 2006). Small sample sizes are not only typical of preclinical studies, but also of clinical studies addressing the same problem. For instance, gene expression signatures were identified and evaluated using 81 melanoma patients to predict their response to PD-1 checkpoint inhibitors (Ayers et al., 2017).

In this study, we will investigate whether it is possible to anticipate the response of BC patients to paclitaxel using GDC data. We also aim at discovering the molecular factors that, collectively, best discriminate between paclitaxel-resistant and paclitaxel-sensitive BC patients. High-dimensional data promotes model overfitting, which in turn results in poorer predictions. As predictive performance differences between ML algorithms are strongly problem-dependent (Tan and Gilbert, 2003; Fernández-Delgado et al., 2014), considering a range of algorithms to identify those that are most suitable for paclitaxel-treated BC patients is appealing. To this end, we apply 10 ML methods to build predictive models in combination with each available molecular profile. Some of the resulting multi-variate predictors are highly interpretable in that they can answer questions such as why this particular patient is non-responsive. This information should permit formulating hypothesis about the molecular mechanisms of BC patient resistance to paclitaxel.

MATERIAL AND METHODS

GDC Data

GDC molecular profiling and clinical data from the TCGA Breast Invasive Carcinoma or BRCA (https://portal.gdc.cancer.gov/projects/TCGA-BRCA) provide the basis for this study. Molecular profiles and clinical data come from release version 4.0, except for miRNA and miRNA isoform (isomiR) expressions coming from release version 8.0 (Release Notes - GDC Docs).

TCGA-BRCA project gathers data from 1,098 patients, resulting in almost 13,000 files (around 130 GB). These datasets were retrieved and downloaded using the GDC Application Programming Interface (API). **Table S1** reports information about files collected from the GDC that have been used to generate datasets.

Processing Clinical Data for Modelling

Patient population included primary or secondary advanced breast cancer receiving single-agent paclitaxel. For some patients it was observed that different drugs have the same or

very close treatment start and end time. These entries may form part of a drug combination. However, available drug response annotations do not allow to check this information. Therefore, possible effects due to drug combinations are ignored in this study when identifying paclitaxel-treated patients. Patients with missing paclitaxel response were not retained. To only consider baseline tumors' molecular profiles, patient records were only retained if no treatment was administered before resection and the time of sample procurement is indicated. We assumed that a baseline molecular profile can explain drug responses observed in a given patient even if paclitaxel was administered at any time after sample resection (Geeleher et al., 2014). After these curation steps, 61 paclitaxel-treated BC patients with valid records remained (**Table S2** reports information about treatments and biospecimens). Annotated patient responses are divided into four categories based on the RECIST standard (Therasse et al., 2000): "Complete Response" (CR), "Partial Response" (PR), "Stable Disease" (SD), and "Clinical Progressive Disease" (CPD). We further classified clinical responses into two categories, namely "responder" (CR or PR) and "non-responder" (SD or CPD).

Processing Molecular Profiles for Modelling

The GDC works on harmonization of raw genomic data developing specific workflows to provide consistent and up-to-date molecular profiles (GDC Reference Files | NCI Genomic Data Commons, Genomic Data Harmonization | NCI Genomic Data Commons). Available profiles comprise copy-number variation (CNV), DNA methylation, mRNA, miRNA and isomiR (Ameres and Zamore, 2013) expressions. In order to produce suitable inputs for ML algorithms and/or extract some specific information, we processed some of them. More details are available in the homonym section of Supplementary Methods. All the datasets produced from these molecular profiles are made of real-valued features.

Predicting Drug Response Using ML Algorithms With Embedded Feature Selection

Most classifiers were built with ML algorithms capable of embedded feature selection to mitigate the impact of high-dimensional data on their generalization on unseen data. Implementations of Classification And Regression Tree (CART) (Breiman et al., 1984) and Random Forest (RF) (Breiman, 2001) were taken from the python library *Scikit-learn* version 0.19.1, while the ones of XGBoost (XGB) (Chen and Guestrin, 2016) version 0.6 and LightGBM (LGBM) (Ke et al., 2017) version 2.0.10 were respectively downloaded from https://github.com/dmlc/xgboost and https://github.com/Microsoft/LightGBM. We also applied a Deep Neural Network (DNN) algorithm (Bengio, 2009) implemented with the python library *Keras* version 2.2.4 using the TensorFlow backend. In addition to these nonlinear models, linear models were generated with Logistic Regression (LR) (Ranstam et al., 2016), which is also implemented in *Scikit-learn*. The visualization of Decision Trees (DTs) was done with the python package *dtreeviz* version 0.2.2. The homonym section in Supplementary Methods provides more details.

Predicting Drug Response Using the Optimal Model Complexity (OMC)

OMC is a strategy to build ML models employing only the most relevant features (Nguyen et al., 2018). More details are available in the homonym section of **Supplementary Methods**.

Measuring the Predictive Performance of a Classifier

This is a binary classification problem, as each patient belongs to one of two classes, responder and non-responder, with the responder considered as the positive class. As it is customary with problems with a small number of data instances (**Table S3**), we are using LOO (Leave-One-Out) CV (Cross-Validation) to evaluate the developed classifiers. Several types of LOOCV will be used: standard for "all-features models", nested for "OMC models", and permutated for "permutation models". As with any other CV (Kohavi, 1995), each data instance (patient here) is exactly once in the test set. Thus, CV performance of a model is exclusively based on the prediction of test instances that were not used in any way to train or select the model (any feature selection is hence carried out on training folds only). Employing nested CV on algorithms requiring model selection (those employing OMC) provides an unbiased estimate of model performance, as it has been shown elsewhere (Cawley and Talbot, 2010; Varma and Simon, 2006).

Once known and predicted classes are compared for all held-out samples, true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are counted among BC patients. These counts are used for calculating classification metrics that summarize the predictive performance reached by a classifier: precision, recall, F1-scores (Van Rijsbergen, 1979), and Matthews Correlation Coefficient (MCC) (Matthews, 1975; Boughorbel et al., 2017). More details about these metrics can be found in the homonym section of Supplementary Methods. Classification scores and contingency matrices obtained from all produced classifiers are stored in **Tables S4** and **S5**, respectively.

RESULTS

Benchmarking of All-Features Models (RF, XGB, LGBM, DNN, LR) Reveals Some Informative Molecular Profiles, But the Resulting Classifiers Perform Marginally Better Than Random

Figure 1 shows that most of the all-features ML classifiers perform worse than random classification reaching slightly negative median MCCs (from -0.19 to -0.05). Poor performance was also obtained when using linear models: LR models perform randomly at best (median MCCs range

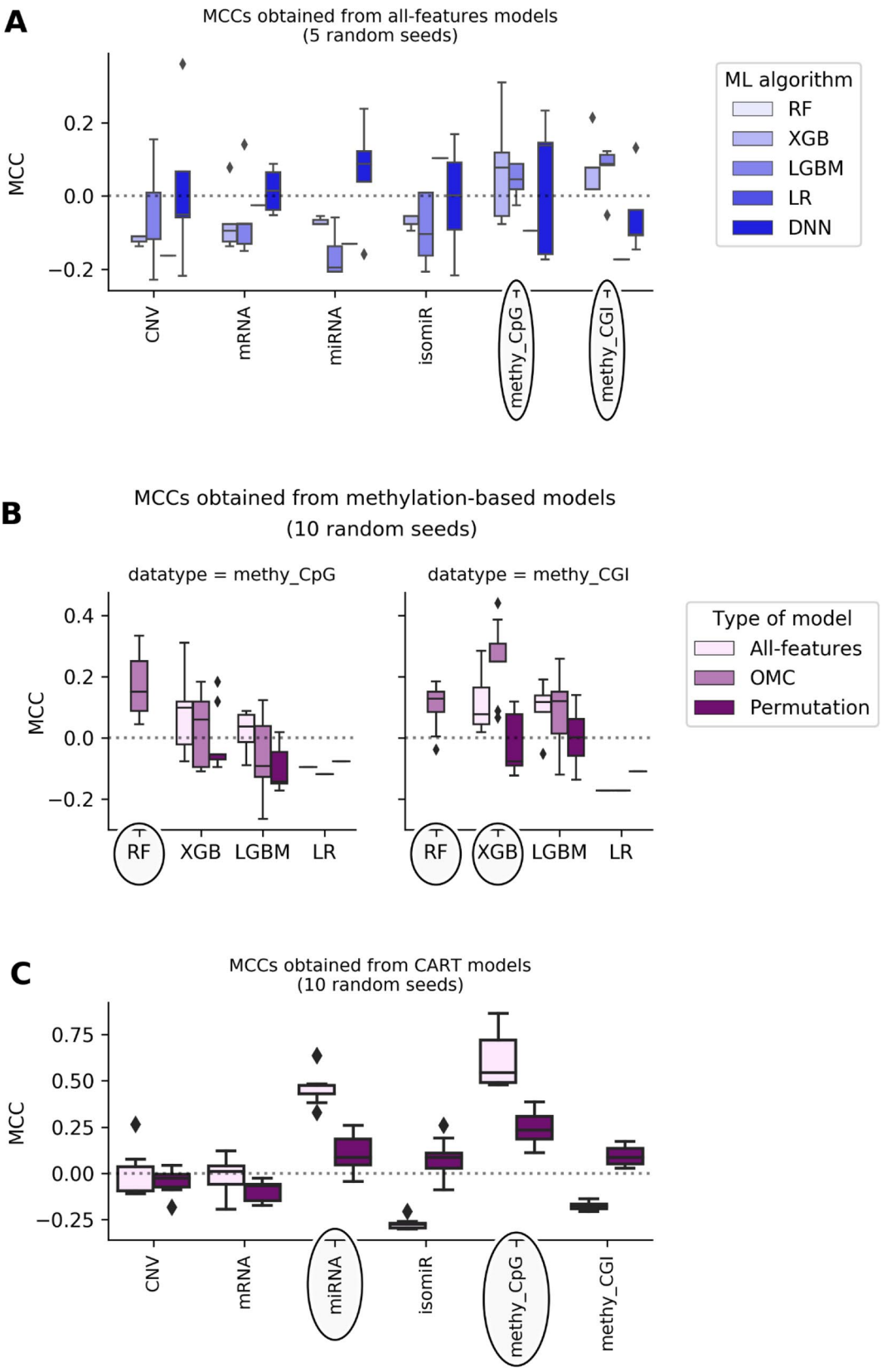

FIGURE 1 | Continued

FIGURE 1 | DNA methylation and miRNA expression lead to the most predictive ML models. Each MCC of a given model is calculated by LOOCV. The experiment is repeated several times, each time with a different random seed, giving rise to a boxplot of MCCs for each case. Permutation models were generated after shuffling class labels on the considered training set. As RF models give undefined MCCs, blanks are found in bins where boxplots are expected. Substantial variability is observed, showing that this problem is both profile- and classification method-dependent. The dashed line shows the expected MCC from random classification. **(A)** Predictive performance of all-features models. All-features models are those in principle employing all the features in the profile to generate the prediction. Models are built with ML algorithms using the default operating threshold (0.5) to calculate the predicted class label from the predicted class probability. Five random seeds were set for each ML algorithm; thus, MCC values come from five runs of standard LOOCV. *x*-axis shows the employed molecular profile, while *y*-axis displays the MCCs obtained by classifiers. From the lightest to the darkest blue, boxplots summarize the distributions of MCCs obtained by XGB,LGBM, LR, and DNN models, respectively. Ellipses indicate which profiles employed by models obtain better-than-random predictive performance: DNA methylation profiles are the most predictive. This also suggests that the other profiles are less informative for the prediction of BC tumor response to paclitaxel. **(B)** Predictive performance applying OMC to methylation-based models. 10 random seeds were used to investigate further the most predictive profiles. OMC models had their hyperparameters complexity and operating threshold tuned and thus required nested LOOCV. Horizontal axes show the employed ML algorithms to process CpG site (left) and CGI (right) methylation datasets, while vertical axes display MCCs achieved by classifiers. Light-pink, medium-pink, and indigo boxplots summarize the distributions of MCCs obtained by all-features, OMC and permutations models, respectively. Circles indicate ML algorithms releasing models with predictive performance improved using OMC. This shows that predictive accuracy depends on both the molecular profile and the ML algorithm. Here, the best models found are CpG site methylation-based RF-OMC, CGI methylation-based RF-OMC, and CGI methylation-based XGB-OMC. **(C)** Predictive performance of CART models. These models (light-pink boxes) were built considering all features in the profile with no hyperparameter tuning. Permutation models (indigo boxes) were trained after that shuffling class labels in the training set. Each MCC is calculated by standard LOOCV, a process repeated with 10 different random seeds. *x*-axis shows the molecular profiles ('CNV' is short for copy-number variation, 'methy_CpG' for CpG site methylation, and 'methy_CGI' for CGI methylation), while *y*-axis displays the LOOCV MCCs achieved by each classifier. The dashed line shows the expected MCC from random classification. Ellipses indicate molecular profiles processed by CART models obtaining the highest predictive performance. These results reveal that CpG sites methylation-based and miRNA expression-based CART models are the most predictive. Predictive accuracy is substantially higher than that provided by all-features or OMC models (in **A** and **B**), which demonstrates that the CART learning algorithm is more suitable for these problem instances.

from -0.17 to 0.1 depending on the profile). Poor predictive performance is primarily caused by FPs (i.e. misclassification of non-responders). This problem was particularly noticeable in RF models, which misclassified every non-responsive patient regardless of the employed profile, leading to undefined MCCs. The latter shows that all-features RF handles class imbalance poorly on these particular problem instances.

DNA methylation-based XGB, LGBM, and DNN models achieve median MCCs slightly higher than 0.0, and they perform hardly better than permutation models. On the one hand, CpG site methylation-based XGB, LGBM, and DNN models obtain a median MCC of 0.08 (p-value from two-sided paired Student's t-test obtained by class-permutation test = $1.05{\cdot}10^{-1}$), 0.04 (p-value = $1.41{\cdot}10^{-1}$), and 0.14 (p-value = $4.08{\cdot}10^{-1}$), respectively. On the other hand, CpG island (CGI) methylation-based XGB and LGBM models achieve a median MCC of 0.08 (p-value = $2.33{\cdot}10^{-1}$) and 0.09 (p-value = $8.04{\cdot}10^{-2}$), respectively. Moreover, miRNA and mRNA expression-based DNN models had a median MCC of 0.088 (p-value = $4.84{\cdot}10^{-2}$) and 0.015 (p-value = $4.76{\cdot}10^{-1}$), respectively.

Complexity-Optimized ML Models (RF-OMC, XGB-OMC, and LGBM-OMC) Provide Better Prediction and Extract Relevant Factors for Paclitaxel Response From CGI Methylation Data

Using OMC allows both to reduce considerably the number of features considered during model training and to adjust the operating threshold for assigning class labels to data instances. This leads to some OMC models that perform better than those considering all features from dataset (**Table S12** in **Supplementary Results**). This is especially the case for some methylation-based models that have been improved using OMC (**Figure 1**), unlike for models based on other profiles (**Figure S5**). The improvement of OMC over the all-features approach is ML algorithm-dependent.

CGI methylation-based OMC models have obtained improved predictive performance, using either RF or XGB. For instance, XGB-OMC models obtain a median MCC of 0.25, which is significantly better than both permutation and all-features models (p-values equal $9.30 \times{\cdot}10^{-4}$ and $2.16 \times{\cdot}10^{-2}$, respectively). In order to extract a robust subset of molecular factors potentially involved in paclitaxel response, the most informative features selected by these models were investigated (**Table S13** and **S14**). It results in 7 out of the 11,644 CGI coordinates encoded as CGI_ID.24217, CGI_ID.15915, CGI_ID.6919, CGI_ID.5276, CGI_ID.5459, CGI_ID.16043, and CGI_ID.11903. Moreover, we notice that 5 of them are common to the features used by the RF-OMC models. Consulting indices provided in **Tables S7** and **S8** (more details in **Supplementary Methods**), we found that these coordinates are related to the following 16 genes: CYP2D6, NDUFA6-AS1, RP4-669P10.19 (or C6orf108 pseudogene), MBTPS2, YY2, C2orf40 (or ECRG4), UXS1, IKZF1, APOBEC4, RGL1, ARPC5, NCF2, SMG7, C1orf177 (or LEXM), RP11-631M21.6 (or FAM166A pseudogene 7), and TUBB8 (**Table S14**).

Transparent ML Models (CART) Capture CpG Methylation Sites and Mature MIRNAS Relevant for the Sensitivity to Paclitaxel and Show How They Are Combined to Explain Drug Response

Most of the available profiles led to poor classification of test set patients when modelled with CART (**Figure 1**). By contrast, CART classifiers based on miRNA expression and CpG site methylation data provided high to very high predictive performance in the context of this problem (in 10 LOOCV runs, median MCCs of 0.43 and 0.54 were obtained, respectively) and performed

significantly better than random models (p-values from two-sided paired Student's t-test obtained by class-permutation test equal $4.57 \cdot \times 10^{-6}$ and $2.86 \cdot \times 10^{-4}$, respectively; see **Figure 1** and **Table 1**). For each case, the best model is defined as that obtaining the highest MCC in 10 standard LOOCVs of the full dataset (i.e. all data instances and all available features). **Figure S6** shows that the performance of these models is robust to different sizes of both training set and test set.

As observed in **Figure 2** CART models strongly reduce the number of features involved in the predictions. The miRNA expression-based model found that 4 out of 337 mature miRNAs were the most informative features (MIMAT0004985 or miR-942-5p, MIMAT0000084 or miR-27a-3p, MIMAT0000274 or miR-217, and MIMAT0004657 or miR-200c-5p), while the CpG-site methylation model identified 2 out of 22,941 CpG sites as the most informative features (cg09691574, which is related to the protein coding genes MRGPRX4 and SAA2-SAA4, and to the lincRNA RP11-113D6.6 also called antisense to MRGPRX4; and cg12542281, which is related to the protein coding gene N4BP2L2). The DTs represented in **Figure 2** show directly the interactions between independent features leading to the predictions. They also reveal the molecular types associated to paclitaxel-sensitive and paclitaxel-resistant BC tumors (the CpG site index is provided in **Table S7**).

Lastly, integrating different molecular profiles has sometimes been found to provide small predictive accuracy gains, e.g. see **Figure 4** in this study (Costello et al., 2014). Thus, since both miRNA and methy_CpG profiles led to the most predictive models, it makes sense to integrate these data sets and train CART models on the features of the resulting hybrid profile. Using the same 10 random seeds as the methy_CpG-based CART models (median LOOCV MCC of 0.54), the hybrid CART models obtained slightly worse accuracy (median LOOCV MCC of 0.52). The resulting CART tree is identical to that in **Figure 2**, suggesting that miRNA features were overshadowed by methy_CpG features during CART induction.

DISCUSSION

Owing to the wealth of curated data offered by the GDC, we could evaluate six profiles. The exhaustive evaluation of the 60 predictive models obtained, employing 10 ML algorithms with each profile, reveal strong variability in predictive performance (**Figure 3**). These results show the importance of considering multiple profiles and ML algorithms, the latter being always possible. For example, we could have carried out this study using the standard all-features versions of tree-ensemble, LR and DNN algorithms. However, this would have only resulted in models with near-random predictability despite using six profiles and thus, we could have concluded that precision oncology is not possible for paclitaxel-treated BC patients. Instead, we also tested algorithms generating models requiring only a handful of features (OMC-based and CART), which in addition, provided the best performance on these problem instances. Note that the most predictive of these models achieved an over 10,000-fold reduction in the number of features (**Table 1**).

Identifying a concise list of predictive molecular features is indeed beneficial for interpretability. The CGI methylation-based XGB-OMC model employs a dramatically reduced number of features (11 of the considered 11,644). The increased predictive performance comparing to all-features model (**Figure 1**) shows that the selected subset of features contains the information relevant for predictions (**Figure S2**). Therefore, applying OMC not only offers better predictivity, but also better interpretability of response to paclitaxel, as it revealed a molecular signature able to discriminate sensitive and resistant BC tumors from high-dimensional data. The best CART models reached the highest predictive performance among the generated predictors (**Figure 1**). Moreover, these models allow going further in the interpretation of response to paclitaxel (**Figure 2**). For example, the CpG-methylation DT unveils two rules employing only two features to predict which are the paclitaxel-sensitive BC tumors (**Figure 2**). The other example is the miRNA DT, which carries out these predictions using four induced rules based on only four features (**Figure 2**). Thus, the application of these rules to forthcoming tumors should improve paclitaxel treatment for BC patients. To facilitate such application, we are providing two python scripts in the supplementary materials, each implementing the rules for one of these predictive profiles.

Our best classifier obtained a median MCC of 0.54 in 10 LOOCV runs (an average MCC of 0.62, with MCC ranging from

TABLE 1 | Best CART models.

Tumor profiling data	Number of considered features	Number of selected features	Median MCC(CART trained on original data)	Median MCC(CART trained on class-permutated data)	p-value(original vs permutated)
miRNA	337	4	0.43	0.09	$4.57 \cdot 10^{-6}$
methy_CpG	22,941	2	0.54	0.23	$2.86 \cdot 10^{-4}$

*The predictive performance of CART models was presented in **Figure 1**. Here we summarize the characteristics of the two best models (i.e. those exploiting miRNA expression and CpG methylation profiles). A median MCC was calculated with the 10 MCCs coming from LOOCV experiments (each with a different random seed). This five additional LOOCV runs with respect to those presented in **Figure 3** were carried out to better characterize the performance of the best models found in our study. The small difference found in median MCC (0.52 in **Figure 3** versus 0.54 here) suggests that this performance metric is quite robust to the number of LOOCV runs for CART. The training sets were also class-permutated during cross-validation as explained in the Methods section and CART trained on the resulting data to provide a second set of 10 MCCs per profile. The p-value (two-sided paired Student's t-test) of this class-permutated test shows how likely are the MCCs of the CART models to arise by chance. The first model was trained on miRNA expression: 4 out of 337 mature miRNAs were retained to build this model reaching a median MCC of 0.43 and performing significantly better than models based on permutated data (p-value = $4.57 \cdot 10^{-6}$). The second model is obtained processing CpG site methylation (shorten as 'methy_CpG'): 2 out of 22,941 CpG sites were retained to build this model achieving a median MCC of 0.54 and performing significantly better than permutation models (p-value = $2.86 \cdot 10^{-4}$).*

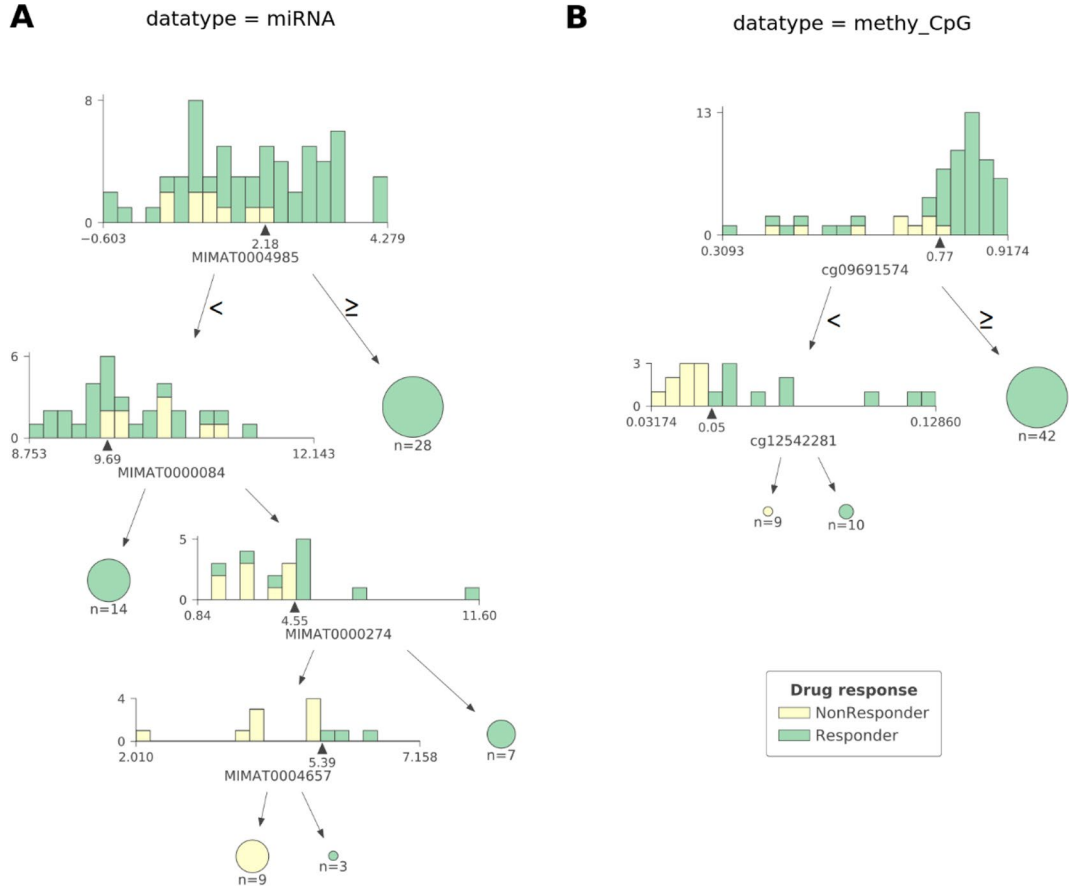

FIGURE 2 | The most predictive CART models offer high interpretability of BC tumors response to paclitaxel. Visualization of the most predictive CART models seen in **Figure 1** exploiting **(A)** miRNA expression and **(B)** CpG site methylation data. Each DT node has a histogram showing the distribution of patients at that node against the selected feature (the proportion of responders vs non-responders in each feature bin is also shown). The triangle under the histogram marks the value of the best split for the selected feature, whose name can be found under the histogram as well. Each node has two leaves: to the left (patients with a feature value lower than that of the best split) and to the right (the rest of the patients). Terminal nodes (or leaves) are displayed as pie charts. The proportions of non-responders and responders are respectively colored yellow and green. The log2-transformed miRNA expression-based DT shown in **(A)** reveals four different molecular types of sensitive BC tumors and one molecular type associated to resistant BC tumors involving four mature miRNAs: MIMAT0004985, MIMAT0000084, MIMAT0000274, and MIMAT0004657 (also known as miR-942-5p, miR-27a-3p, miR-217, and miR-200c-5p, respectively). Thus, a tumor is classified as responsive to paclitaxel if: 1) the expression value of MIMAT0004985 is higher than 2.18, or 2) the expression value of MIMAT0004985 is lower than 2.18 and that of MIMAT0000084 is lower than 9.69, or 3) the expression value of MIMAT0004985 is lower than 2.18, and that of MIMAT0000084 is higher than 9.69 and that of MIMAT0000274 is higher than 4.55, or 4) the expression value of MIMAT0004985 is lower than 2.18, and that of MIMAT0000084 is higher than 9.69, and that of MIMAT0000274 is lower than 4.55, and that of MIMAT0004657 is higher than 5.39. Otherwise, the tumor is classified as non-responsive. The DT based on CpG site methylation (shortened as 'methy_CpG') shown in **(B)** unveils two different molecular types of sensitive BC tumors and one type of resistant BC tumors involving two CpG sites represented by probes cg09691574, which is related to the protein coding genes, MRGPRX4 and SAA2-SAA4; and to the lincRNA RP11-113D6.6, also called antisense, to MRGPRX4; and to cg12542281, which is related to the protein coding gene N4BP2L2. Thus, a tumor is predicted to be sensitive to paclitaxel if: 1) the beta value associated to the methylation of cg09691574 is higher than 0.77, or 2) the beta value associated to the methylation of cg09691574 is lower than 0.77, and that of cg12542281 is higher than 0.05. Otherwise, the tumor is predicted to be resistant. Both DTs were found to give pure leaves (i.e. all data instances that are in terminal nodes belong to the same class).

0.48 to 0.87 in these runs as it can be seen in **Figure 1**). To put these predictive accuracies in the context of what is typically achieved when predicting tumor response to a drug from omics profiles, we have looked at other test set performances reported at the literature for this problem. One study (Kim et al., 2016) applied a range of ML algorithms to predict pancancer cell line response from transcriptomic profiles and obtained MCCs below 0.6 in all cases (see **Figure 1** in that paper). Maximum MCCs slightly above 0.5 and 0.3 were also obtained using RF with transcriptomic profiles (Nguyen et al., 2017) and genomic profiles (Naulaerts et al., 2017), respectively. Another study (Xu et al., 2019) also predicted drug response using many hundreds of pancancer cell lines using several ML algorithms from various omics profiles (gene expression, copy-number alterations, single-nucleotide mutations). Depending on the considered data resource, average MCCs across drug and profiles range from 0.15 to 0.31 or from 0.22 to 0.45 (see Tables 2 and 3 in that paper). Yet another example is by (Tripathi et al., 2016) using gene variants as features, where MCCs range across

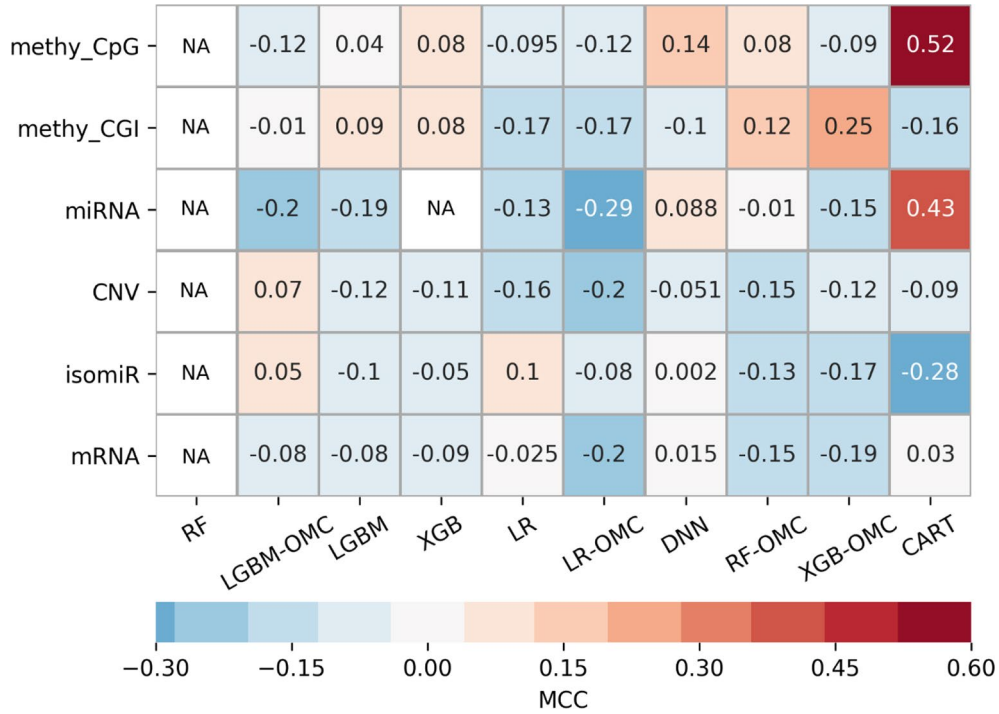

FIGURE 3 | Employing multiple ML algorithms and tumor profiles increase the likelihood of discovering models able to predict BC patient response to paclitaxel. ML algorithms include the unaltered version of tree-ensemble and linear algorithms using all available features (RF, XGB, LGBM, and LR) and their OMC versions (RF-OMC, XGB-OMC, LGBM-OMC, and LR-OMC). The 9th algorithm was CART, employed to generate simpler and more interpretable classification models. The 10th algorithm was DNN, employed to generate more sophisticated but less interpretable models. Each of these algorithms was evaluated on each of the six molecular profiles, which resulted in 60 classifiers on the same BC patients. LOOCV evaluation was performed 5 times setting a different random seed for the employed ML algorithm, leading to 5 MCC determinations quantifying predictive performance. The heatmap shows the median MCC per classifier. Rows show the processed molecular profiles ('CNV' is short for copy-number variation, 'methy_CpG' for CpG site methylation, and 'methy_CGI' for CGI methylation), while columns display the employed ML algorithms. Thus, each cell corresponds to the median MCC of a given predictive model. Cells are colored in light-blue and dark-blue when this model reaches a negative or very negative median MCC (i.e. classification worse than random); in grey when it reaches a median MCC very close to 0.0 (i.e. random classification); in light-brown and dark-brown when it reaches a positive or very positive median MCC (i.e. classification better than random or close to perfect); in white and labelled NA (i.e. not available) when it reaches an undefined median MCC (i.e. misclassification of non-responders within several or all iterations). These results show that DNA methylation is the most informative profile (it leads to 2 of the 3 classifiers with a median MCC of a least 0.25). The choice of ML algorithm also affects the predictive performance. For example, none of the RF or LGBM classifiers obtain an MCC of at least 0.10. Thus, predictive performance depends strongly on of algorithm- profile combination: only one XGB-OMC models is predictive (that based on CGI methylation) and it is among the best predictors (median MCC of 0.25). Two other examples are the CART classifiers based on CpG methylation and miRNA expression, with median MCC of 0.52 and 0.43, respectively. **Figures S4** and **S5** further characterizes the performance of the best classifiers.

drugs from 0.32 to 0.56 or from 0.30 to 0.44 depending on data resource (see **Tables 1** and 2 in that paper). Lastly, single-gene drug response markers identified by MANOVA and Chi-Square tests on pancancer cell lines obtained maximum MCCs of 0.30 and 0.31, respectively (Dang et al., 2018).

The alteration of gene expression due to epigenetic modifications triggers the development of cancers, including BC. DNA methylation changes, occurring both within and around CGIs, can impact transcriptional activity of genes or transcription factors involved in malignant phenotypes (Esteller, 2002; Irizarry et al., 2009; Levenson, 2010; Deaton and Bird, 2011; Manjegowda et al., 2017; Stirzaker et al., 2017). It has been shown that biomarkers for prognosis and treatment can be unearthed from DNA methylation profiles (Xiang et al., 2013; Mikeska and Craig, 2014; Stirzaker et al., 2014; Li et al., 2015; Pouliot et al., 2015). Furthermore, it has been found that DNA methylation can interfere in chemo-resistance to paclitaxel (Wang et al., 2012; Ignatov et al., 2014; Yun et al., 2015; He et al., 2016; Zhang et al., 2018). Our DNA-methylation-based predictors selected CpG sites and CGIs related to genes previously found individually involved in cancer development and with transcriptional activity regulated by methylation (e.g. MBTPS2, YY2, ECRG4, IKZF1). Selected features by these models are also related to genes associated to response to cytotoxic drugs such as N4BP2L2 (paclitaxel), CYP2D6 (tamoxifen), APOBEC4 (tamoxifen, doxorubicin, and etoposide), and TUBB8 (paclitaxel) (**Table S15**).

miRNAs also play a key role in cancer development by acting as tumor suppressors or oncogenes. These molecules can be used as biomarkers, and modulation of their specific activities provides insight for therapeutic investigations (Hayes et al., 2014; Peng and Croce, 2016). Furthermore, the expression of some miRNAs has been associated to the sensitivity to paclitaxel (Zhou

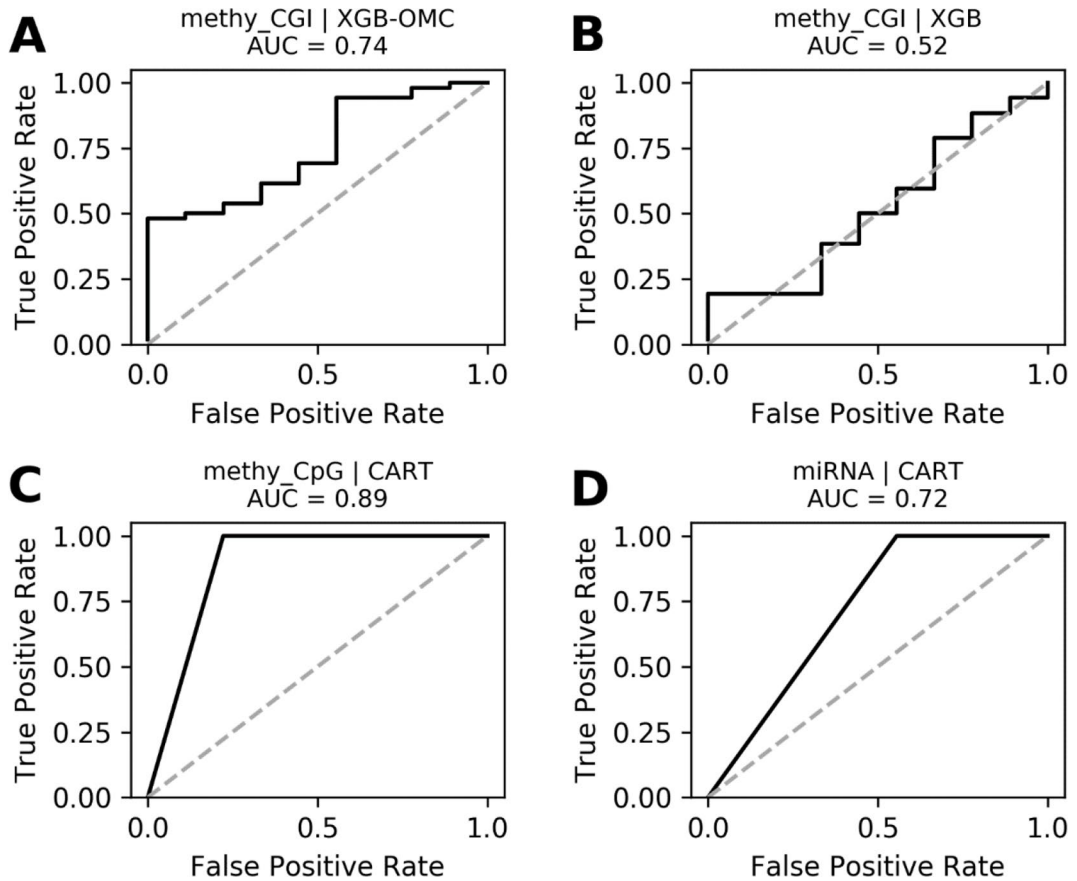

FIGURE 4 | ROC curves of the most predictive case of best models. ROC curves obtained plotting the true positive rates against the false positive rates calculated from the models presented in **Figures S4**. The AUCs were calculated from the predictions that came out from the nested and standard LOOCV runs and were respectively carried out for OMC and CART models. We notice that AUCs follow the same trend as MCCs and that models shown in **(A)**, **(C)**, and **(D)** are very robust. The dashed line delimitates the expected AUC from random classification.

et al., 2010; Chen et al., 2014; He et al., 2016; Lu et al., 2017). The miRNA expression-based CART model combines miR-27a-3p, miR-217, miR-200c-5p, and miR-942-5p to predict which BC tumors are paclitaxel-responsive with high accuracy (**Figures 1** and **2A**). Individually, each of these miRNAs have been linked to paclitaxel response and BC prognosis: the first three are related to paclitaxel resistance, whereas the last one is associated to shorter survival of BC patients (**Table S15**).

Our study has some limitations to be pointed out. First, for a given patient, molecular profiles were obtained from the primary tumor, while clinical response was registered later following tumor evolution. Both tumors may present some differences at the molecular level, due to temporal or spatial tumor heterogeneity, as well the possible impact of the treatment administered after tumor resection. Second, while we reported predictive accuracy on BC tumors not used in any way to build or select the model, an additional independent evaluation on a second cohort would shed further light into how general these models are. The latter is currently not possible due to the scarcity of paclitaxel-treated BC patients with DNA methylation or miRNA profiles of their tumors.

Yet, our work provides very predictive (in the context of the considered problem), robust (**Figure S4** and **Figure 4**), and even interpretable models to identify primary BC tumors sensitive to paclitaxel. These results also suggest that tumor methylomes and miRNomes can be a source of multi-variate models to predict prognosis and treatment response. Indeed, our predictive models reveal molecular features that can collectively anticipate which BC tumors are sensitive or resistant to paclitaxel. Previous studies have experimentally validated the involvement in BC development, and even in the resistance to paclitaxel, of these molecular factors individually, which further supports the applicability of these classifiers. Furthermore, our results also suggest novel predictive factors such as the antisense to MRGPRX4; the pseudogenes (Poliseno et al., 2015; Xiao-Jie et al., 2015) C6orf108 and FAM166A; and the coding genes NDUFA6-AS1, UXS1, RGL1, and LEXM.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the https://portal.gdc.cancer.gov/.

AUTHOR CONTRIBUTIONS

PB conceived the study and designed the experiments. AB and PB wrote the manuscript with the assistance of AG. AB carried out the numerical experiments. All authors analyzed the results and contributed to their discussion.

FUNDING

This work was supported by the Institut Paoli-Calmettes (grant number 305/2016 to PB).

ACKNOWLEDGMENTS

Computing resources for this study were partly provided by the computing facilities DISC (Datacenter IT and Scientific Computing) of the Centre de Recherche en Cancérologie de Marseille.

SUPPLEMENTARY MATERIALS

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01041/full#supplementary-material

REFERENCES

Ajabnoor, G., Crook, T., and Coley, H. (2012). Paclitaxel resistance is associated with switch from apoptotic to autophagic cell death in MCF-7 breast cancer cells. *Cell Death Dis.* 3, e260. doi: 10.1038/cddis.2011.139

Ali, M., and Aittokallio, T. (2018). Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys. Rev.* 11, 1–9. doi: 10.1007/s12551-018-0446-z

Ameres, S. L., and Zamore, P. D. (2013). Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.* 14, 475–488. doi: 10.1038/nrm3611

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D. R., et al. (2017). IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* 127, 2930–2940. doi: 10.1172/JCI91190

Bartlett, J. M. S., Nielsen, T. O., Gao, D., Gelmon, K. A., Quintayo, M. A., Starczynski, J., et al.. (2015). TLE3 is not a predictive biomarker for taxane sensitivity in the NCIC CTG MA.21 clinical trial. *Br. J. Cancer*. 113, 722–728. doi: 10.1038/bjc.2015.271

Bengio, Y. (2009). "Learning Deep Architectures for AI," in *Found. Trends® Mach. Learn*. Hanover, MA, USA. doi: 10.1561/2200000006

Biankin, A. V., Piantadosi, S., and Hollingsworth, S. J. (2015). Patient-centric trials for therapeutic development in precision oncology. *Nature*. 526, 361–370. doi: 10.1038/nature15819

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One*. 12, e0177678. doi: 10.1371/journal.pone.0177678

Breiman, L. (2001). Random Forests. *Mach. Learn.* Boca Raton, FL, USA. 45, 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Brown, R., and Böger-Brown, U., (1999). *Cytotoxic Drug Resistance Mechanisms*. New Jersey: Humana Press. doi: 10.1385/1592596878

Cardoso, F., Di Leo, A., Lohrisch, C., Bernard, C., Ferreira, F., and Piccart, M. J. (2002). Second and subsequent lines of chemotherapy for metastatic breast cancer: what did we learn in the last two decades? *Ann. Oncol.* 13, 197–207. doi: 10.1093/annonc/mdf101

Cawley, G. C., and Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* 11, 2079–2107.

Chen, N., Chon, H. S., Xiong, Y., Marchion, D. C., Judson, P. L., Hakam, A., et al. (2014). Human cancer cell line microRNAs associated with in vitro sensitivity to paclitaxel. *Oncol. Rep.* 31, 376–383. doi: 10.3892/or.2013.2847

Chen, T., and Guestrin, C. (2016). "XGBoost," in *Reliable Large-scale Tree Boosting System*. ACM New York, NY, USA. doi: 10.1145/2939672.2939785

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* 32, 1202–1212. doi: 10.1038/nbt.2877

Dang, C. C., Peón, A., and Ballester, P. J. (2018). Unearthing new genomic markers of drug response by improved measurement of discriminative power. *BMC Med. Genomics* 11, 10. doi: 10.1186/s12920-018-0336-z

Deaton, A. M., and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev.* 25, 1010–1022. doi: 10.1101/gad.2037511

Ding, Z., Zu, S., and Gu, J. (2016). Evaluating the molecule-based prediction of clinical drug responses in cancer. *Bioinformatics* 32, 2891–2895. doi: 10.1093/bioinformatics/btw344

Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*. 21, 5427–5440. doi: 10.1038/sj.onc.1205600

Felip, E., and Martinez, P. (2012). Can sensitivity to cytotoxic chemotherapy be predicted by biomarkers? *Ann. Oncol.* 23 Suppl 1, x189–x192. doi: 10.1093/annonc/mds309

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 15, 3133–3181.

Flint, M. S., Kim, G., Hood, B. L., Bateman, N. W., Stewart, N. A., and Conrads, T. P. (2009). Stress hormones mediate drug resistance to paclitaxel in human breast cancer cells through a CDK-1-dependent pathway. *Psychoneuroendocrinology*. 34, 1533–1541. doi: 10.1016/j.psyneuen.2009.05.008

GDC Reference Files | NCI Genomic Data Commons.

Geeleher, P., Cox, N. J., and Huang, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines. *Genome Biol.* 15, R47. doi: 10.1186/gb-2014-15-3-r47

Gehrmann, M., Schmidt, M., Brase, J. C., Roos, P., and Hengstler, J. G. (2008). Prediction of paclitaxel resistance in breast cancer: is CYP1B1*3 a new factor of influence? *Pharmacogenomics*. 9, 969–974. doi: 10.2217/14622416.9.7.969

Genomic Data Harmonization | NCI Genomic Data Commons.

Golubnitschaja, O., Debald, M., Yeghiazaryan, K., Kuhn, W., Pešta, M., Costigliola, V., et al. (2016). Breast cancer epidemic in the early twenty-first century: evaluation of risk factors, cumulative questionnaires and recommendations for preventive measures. *Tumor Biol.* 37, 12941–12957. doi: 10.1007/s13277-016-5168-x

Harper, A. R., and Topol, E. J. (2012). Pharmacogenomics in clinical practice and drug development. *Nat. Biotechnol.* 30, 1117–1124. doi: 10.1038/nbt.2424

Hayes, J., Peruzzi, P. P., and Lawler, S. (2014). MicroRNAs in cancer: biomarkers, functions and therapy. *Trends Mol. Med.* 20, 460–469 doi: 10.1016/j.molmed.2014.06.005

He, D. X., Gu, F., Gao, F., Hao, J. J., Gong, D., Gu, X. T., et al. (2016). Genome-wide profiles of methylation, microRNAs, and gene expression in chemoresistant breast cancer. *Sci. Rep.* 6, 24706. doi: 10.1038/srep24706

Housman, G., Byler, S., Heerboth, S., Lapinska, K., Longacre, M., Snyder, N., et al. (2014). Drug resistance in cancer: an overview. *Cancers (Basel)*. 6, 1769–1792. doi: 10.3390/cancers6031769

Huang, M., Shen, A., Ding, J., and Geng, M. (2014). Molecularly targeted cancer therapy: some lessons from the past decade. *Trends Pharmacol. Sci.* 35, 41–50. doi: 10.1016/j.tips.2013.11.004

Ignatov, T., Eggemann, H., Costa, S. D., Roessner, A., Kalinski, T., and Ignatov, A. (2014). BRCA1 promoter methylation is a marker of better response to platinum-taxane-based therapy in sporadic epithelial ovarian cancer. *J. Cancer Res. Clin. Oncol.* 140, 1457–1463. doi: 10.1007/s00432-014-1704-5

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., et al.. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* 41, 178–186. doi: 10.1038/ng.298

Jensen, M. A., Ferretti, V., Grossman, R. L., and Staudt, L. M. (2017). The NCI genomic data commons as an engine for precision medicine. *Blood*. 130, 453–459. doi: 10.1182/blood-2017-03-735654

Kadra, G., Finetti, P., Toiron, Y., Viens, P., Birnbaum, D., Borg, J.-P., et al. (2012). Gene expression profiling of breast tumor cell lines to predict for therapeutic response to microtubule-stabilizing agents. *Breast Cancer Res. Treat.* 132, 1035–1047. doi: 10.1007/s10549-011-1687-8

Ke, G., Meng, Q., Wang, T., Chen, W., Ma, W., Liu, T.-Y., et al. (2017). LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* 30, 3147–3155.

Kim, S., Sundaresan, V., Zhou, L., and Kahveci, T. (2016). Integrating domain specific knowledge and network analysis to predict drug sensitivity of cancer cell lines. *PLoS One* 11, e0162173. doi: 10.1371/journal.pone.0162173

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc. *14th Int. Jt. Conf. Artif. Intell.* 2, 1137–1143. doi: 10.1067/mod.2000.109031

Levenson, V. V. (2010). DNA methylation as a universal biomarker. *Expert Rev. Mol. Diagn.* 10, 481–488. doi: 10.1586/erm.10.17

Li, Y., Melnikov, A. A., Levenson, V., Guerra, E., Simeone, P., Alberti, S., et al.. (2015). A seven-gene CpG-island methylation panel predicts breast cancer progression. *BMC Cancer*. 15, 417. doi: 10.1186/s12885-015-1412-9

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Lu, C., Xie, Z., and Peng, Q. (2017). MiRNA-107 enhances chemosensitivity to paclitaxel by targeting antiapoptotic factor Bcl-w in non small cell lung cancer. *Am. J. Cancer Res.* 7, 1863–1873.

Ma, Y., Ding, Z., Qian, Y., Shi, X., Castranova, V., Harner, E. J., et al.. (2006). Predicting cancer drug response by proteomic profiling. *Clin. Cancer Res.* 12, 4583–4589. doi: 10.1158/1078-0432.CCR-06-0290

Mandrekar, S. J., and Sargent, D. J. (2009). Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J. Clin. Oncol.* 27, 4027–4034. doi: 10.1200/JCO.2009.22.3701

Manjegowda, M. C., Gupta, P. S., and Limaye, A. M. (2017). Hyper-methylation of the upstream CpG island shore is a likely mechanism of GPER1 silencing in breast cancer cells. *Gene*. 614, 65–73. doi: 10.1016/j.gene.2017.03.006

Marsh, S., Somlo, G., Li, X., Frankel, P., King, C. R., Shannon, W. D., et al.. (2007). Pharmacogenetic analysis of paclitaxel transport and metabolism genes in breast cancer. *Pharmacogenomics J.* 7, 362–365. doi: 10.1038/sj.tpj.6500434

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta. Protein Struct.* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Mikeska, T., and Craig, J. M. (2014). DNA methylation biomarkers: cancer and beyond. *Genes (Basel)*. 5, 821–864. doi: 10.3390/genes5030821

Murray, S., Briasoulis, E., Linardou, H., Bafaloukos, D., and Papadimitriou, C. (2012). Taxane resistance in breast cancer: mechanisms, predictive biomarkers and circumvention strategies. *Cancer Treat. Rev.* 38, 890–903. doi: 10.1016/j.ctrv.2012.02.011

Naulaerts, S., Dang, C., and Ballester, P. J. (2017). Precision and recall oncology: combining multiple gene mutations for improved identification of drug-sensitive tumours. *Oncotarget* 8, 97025–97040. doi: 10.18632/oncotarget.20923

Nguyen, L., Dang, C. C., and Ballester, P. J. (2017). Systematic assessment of multi-gene predictors of pan-cancer cell line sensitivity to drugs exploiting gene expression data. *Research* 5, 2927. doi: 10.12688/f1000research.10529.2

Nguyen, L., Naulaerts, S., Bomane, A., Bruna, A., Ghislat, G., and Ballester, P. (2018). Machine learning models to predict *in vivo* drug response *via* optimal dimensionality reduction of tumour molecular profiles. *bioRxiv* 277772, 1–34. doi: 10.1101/277772

Norimura, S., Kontani, K., Kubo, T., Hashimoto, S.-I., Murazawa, C., Kenzaki, K., et al. (2018). Candidate biomarkers predictive of anthracycline and taxane efficacy against breast cancer. *J. Cancer Res. Ther.* 14, 409–415. doi: 10.4103/jcrt.JCRT_1053_16

Peck, R. W. (2016). The right dose for every patient: a key step for precision medicine. *Nat. Rev. Drug Discovery*. 15, 145–146. doi: 10.1038/nrd.2015.22

Peng, Y., and Croce, C. M. (2016). The role of microRNAs in human cancer. *Signal Transduct. Target. Ther.* 1, 15004. doi: 10.1038/sigtrans.2015.4

Perez, E. A. (1998). Paclitaxel in Breast Cancer. *Oncologist* 3, 373–389.

Poliseno, L., Marranci, A., and Pandolfi, P. P. (2015). Pseudogenes in human cancer. *Front. Med.* 2, 68. doi: 10.3389/fmed.2015.00068

Pouliot, M. C., Labrie, Y., Diorio, C., and Durocher, F. (2015). The role of methylation in breast cancer susceptibility and treatment. *Anticancer Res.* 35, 4569–4574. doi: 10.1007/s13566-015-0216-5

Prahallad, A., Sun, C., Huang, S., Di Nicolantonio, F., Salazar, R., Zecchin, D., et al. (2012). Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature* 483, 100–103. doi: 10.1038/nature10868

Ranstam, J., Cook, J. A., and Collins, G. S. (2016). Clinical prediction models. *Br. J. Surg.* 103, 1886. doi: 10.1002/bjs.10242

Release Notes – GDC Docs Available at: https://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/ [Accessed March 11, 2019].

Ribeiro, J. T., Macedo, L. T., Curigliano, G., Fumagalli, L., Locatelli, M., Dalton, M., et al. (2012). Cytotoxic drugs for patients with breast cancer in the era of targeted treatment: back to the future? *Ann. Oncol.* 23, 547–555. doi: 10.1093/annonc/mdr382

Van Rijsbergen, C. J. (1979) *Information Retrieval*. Butterworths, London, UK. doi: 10.1016/j.pestbp.2006.07.008

Rodríguez-Antona, C., and Taron, M. (2015). Pharmacogenomic biomarkers for personalized cancer treatment. *J. Int. Med.* 277, 201–217. doi: 10.1111/joim.12321

Russnes, H. G., Navin, N., Hicks, J., and Borresen-Dale, A. L. (2011). Insight into the heterogeneity of breast cancer through next-generation sequencing. *J. Clin. Invest.* 121, 3810–3818. doi: 10.1172/JCI57088

Schwartzberg, L., Kim, E. S., Liu, D., and Schrag, D. (2017). Precision oncology: who, how, what, when, and when not? *Am. Soc. Clin. Oncol. Educ. B.* 37, 160–169. doi: 10.14694/EDBK_174176

Stirzaker, C., Song, J. Z., Ng, W., Du, Q., Armstrong, N. J., Locke, W. J., et al. (2017). Methyl-CpG-binding protein MBD2 plays a key role in maintenance and spread of DNA methylation at CpG islands and shores in cancer. *Oncogene*. 36, 1328–1338. doi: 10.1038/onc.2016.297

Stirzaker, C., Taberlay, P. C., Statham, A. L., and Clark, S. J. (2014). Mining cancer methylomes: prospects and challenges. *Trends Genet.* 30, 75–84. doi: 10.1016/j.tig.2013.11.004

Tan, A. C., and Gilbert, D. (2003). An empirical comparison of supervised machine learning techniques in bioinformatics. *Proc. First Asia-Pacific Bioinforma. Conf. Bioinforma.* 19, 219–222.

Therasse, P., Arbuck, S. G., Eisenhauer, E. A., Wanders, J., Kaplan, R. S., Rubinstein, L., et al. (2000). New guidelines to evaluate the response to treatment in solid tumors. *J. Natl. Cancer Inst.* 92, 205–216. doi: 10.1093/jnci/92.3.205

Tripathi, S., Belkacemi, L., Cheung, M. S., and Bose, R. N. (2016). Correlation between gene variants, signaling pathways, and efficacy of chemotherapy drugs against colon cancers. *Cancer Inform.* 15, 1–13. doi: 10.4137/CIN.S34506

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7, 91. doi: 10.1186/1471-2105-7-91

Wang, L., McLeod, H. L., and Weinshilboum, R. M. (2011). Genomics and drug response. *N. Engl. J. Med.* 364, 1144–1153. doi: 10.1056/NEJMra1010600

Wang, N., Zhang, H., Yao, Q., Wang, Y., Dai, S., and Yang, X. (2012). TGFBI promoter hypermethylation correlating with paclitaxel chemoresistance in ovarian cancer. *J. Exp. Clin. Cancer Res.* 31, 6. doi: 10.1186/1756-9966-31-6

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113. doi: 10.1038/ng.2764

Xiang, T. X., Yuan, Y., Li, L. L., Wang, Z. H., Dan, L. Y., Chen, Y., et al. (2013). Aberrant promoter CpG methylation and its translational applications in breast cancer. *Chin. J. Cancer*. 32, 12–20. doi: 10.5732/cjc.011.10344

Xiao-Jie, L., Ai-Mei, G., Li-Juan, J., and Jiang, X. (2015). Pseudogene in cancer: Real functions and promising signature. *J. Med. Genet.* 52, 17–24. doi: 10.1136/jmedgenet-2014-102785

Xu, X., Gu, H., Wang, Y., Wang, J., and Qin, P. (2019). Autoencoder based feature selection method for classification of anticancer drug response. *Front. Genet.* 10, 233. doi: 10.3389/fgene.2019.00233

Yun, T., Liu, Y., Gao, D., Linghu, E., Brock, M. V., Yin, D., et al. (2015). Methylation of CHFR sensitizes esophageal squamous cell cancer to docetaxel and paclitaxel. *Genes Cancer*. 6, 38–48. doi: 10.18632/genesandcancer.46

Zhang, J., Zhang, J., Xu, S., Zhang, X., Wang, P., Wu, H., et al. (2018). Hypoxia-Induced TPM2 methylation is associated with chemoresistance and poor prognosis in breast cancer. *Cell. Physiol. Biochem*. 45, 692–705. doi: 10.1159/000487162

Zhou, M., Liu, Z., Zhao, Y., Ding, Y., Liu, H., Xi, Y., et al. (2010). MicroRNA-125b confers the resistance of breast cancer cells to paclitaxel through suppression of pro-apoptotic Bcl-2 antagonist killer 1 (Bak1) expression. *J. Biol. Chem*. 285, 21496–21507. doi: 10.1074/jbc.M109.083337

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 November 2019
doi: 10.3389/fgene.2019.01075

Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins

Sina Ghadermarzi[1], Xingyi Li[2], Min Li[2] and Lukasz Kurgan[1]**

[1] Department of Computer Science, Virginia Commonwealth University, Richmond, VA, United States, [2] School of Computer Science and Engineering, Central South University, Changsha, China

Recent research shows that majority of the druggable human proteome is yet to be annotated and explored. Accurate identification of these unexplored druggable proteins would facilitate development, screening, repurposing, and repositioning of drugs, as well as prediction of new drug–protein interactions. We contrast the current drug targets against the datasets of non-druggable and possibly druggable proteins to formulate markers that could be used to identify druggable proteins. We focus on the markers that can be extracted from protein sequences or names/identifiers to ensure that they can be applied across the entire human proteome. These markers quantify key features covered in the past works (topological features of PPIs, cellular functions, and subcellular locations) and several novel factors (intrinsic disorder, residue-level conservation, alternative splicing isoforms, domains, and sequence-derived solvent accessibility). We find that the possibly druggable proteins have significantly higher abundance of alternative splicing isoforms, relatively large number of domains, higher degree of centrality in the protein-protein interaction networks, and lower numbers of conserved and surface residues, when compared with the non-druggable proteins. We show that the current drug targets and possibly druggable proteins share involvement in the catalytic and signaling functions. However, unlike the drug targets, the possibly druggable proteins participate in the metabolic and biosynthesis processes, are enriched in the intrinsic disorder, interact with proteins and nucleic acids, and are localized across the cell. To sum up, we formulate several markers that can help with finding novel druggable human proteins and provide interesting insights into the cellular functions and subcellular locations of the current drug targets and potentially druggable proteins.

Keywords: drug targets, druggability, druggable human proteome, drug-protein interactions, protein-protein interactions, intrinsic disorder

OPEN ACCESS

Edited by:
Shandar Ahmad, Jawaharlal Nehru University, India

Reviewed by:
Marcelo Adrian Marti, University of Buenos Aires, Argentina
Yuanyuan Wang, St. Jude Children's Research Hospital, United States

****Correspondence:***
Min Li
limin@mail.csu.edu.cn
Lukasz Kurgan
lkurgan@vcu.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *30 May 2019*
Accepted: *09 October 2019*
Published: *15 November 2019*

Citation:
Ghadermarzi S, Li X, Li M and Kurgan L (2019) Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins. Front. Genet. 10:1075. doi: 10.3389/fgene.2019.01075

INTRODUCTION

Knowledge of the drug-target interactions is essential for numerous applications including screening of drug candidates (Schneider, 2010; Núñez et al., 2012; Dalkas et al., 2013; Tseng and Tuszynski, 2015), drug repositioning and repurposing (Chong and Sullivan, 2007; Haupt and Schroeder, 2011; Oprea and Mestres, 2012; Hu and Bajorath, 2013; Li et al., 2016), characterization and mitigation of side-effects of drugs (Lounkine et al., 2012; Wang et al., 2012b; Kuhn et al., 2013; Tarcsay and Keserű, 2013; Hu et al., 2014), and prediction of novel protein-drug interactions (Wang et al., 2016a; Lotfi

Shahreza et al., 2017; Ezzat et al., 2018; Hao et al., 2019; Wang and Kurgan, 2019; Wang and Kurgan, 2018; Wang et al., 2019). Recent analysis reveals that over 95% of the currently known drug targets are proteins and that these proteins facilitate about 93% of known drug-target interactions (Santos et al., 2017). Thus, we focus on the drug-protein interactions and we use the term "drug target" as a synonym for the protein drug target. While earlier works report about 400 drug targets (Hopkins and Groom, 2002; Russ and Lampel, 2005), subsequent studies annotate as many as over 600 drug targets in human (Santos et al., 2017). Furthermore, the druggable human proteome, defined as the full complement of the human drug targets (Hopkins and Groom, 2002; Russ and Lampel, 2005; Rask-Andersen et al., 2014; Cimermancic et al., 2016; Hu et al., 2016), is expected to be much larger. Early estimates place the number of human drug targets at around 3,000 (Hopkins and Groom, 2002; Russ and Lampel, 2005). A more recent analysis approximates this number at 4.5 thousand (Finan et al., 2017), which corresponds to about 22% of the human genome. While the historically typical drug targets include G-protein coupled receptors, nuclear receptors, ion channels, and some of the enzymes (Overington et al., 2006; Imming et al., 2007), recent works suggest that many of the non-enzymes (e.g., scaffolding, regulatory, and structural proteins) and proteins involved in specific protein-protein interactions (PPIs) should be targeted by drugs (Makley and Gestwicki, 2013; Ozdemir et al., 2019), effectively expanding the list of potential drug targets. These observations point to the fact that many of the drug targets remain to be discovered and characterized. The search for these proteins relies on the concept of druggability, which was originally defined based on the presence of structure that favors interactions with drug-like compounds where the corresponding interactions provide desired therapeutic effects (Hopkins and Groom, 2002; Russ and Lampel, 2005; Keller et al., 2006). In a purely structural context, druggability is related to binding of a compound to a given protein target with high affinity (< 1 μM) (Sheridan et al., 2010; Radusky et al., 2014). We focus on the former definition where both the interactions and the therapeutic effects are considered.

One of the key elements in the quest to find druggable proteins is to identify functional and structural characteristics that differentiate drug targets from the non-drug targets (Zheng et al., 2006; Lauss et al., 2007; Bakheet and Doig, 2009; Zhu et al., 2009b; Zhu et al., 2009c; Bull and Doig, Mitsopoulos et al., 2015; 2015; Feng et al., 2017; Kim et al., 2017). In one of the earliest works, Chen *et al.* concentrated on the analysis of structural fold types, target family representation and similarity, pathway associations, tissue distribution, and chromosome location for the drug targets (Zheng et al., 2006). A similar analysis that considered cellular functions, pathway associations, tissue distribution, and subcellular and chromosome location of the drug targets was published soon after by Lauss and colleagues (Lauss et al., 2007). More recent studies have shifted the focus towards characteristic features of the target protein sequence and structure. Bakheet and Doig used a relatively small set of 148 targets to analyze several sequence properties (chain length, hydrophobicity, charge, and isoelectric point), putative secondary structure and transmembrane regions, inclusion of signal peptides, selected set of post-translational modifications (PTMs), as well as the previously studied subcellular location and functions (Bakheet and Doig, 2009). Subsequently, Bull and Doig investigated a similar set of characteristics using a much larger set of 1324 drug targets (Bull and Doig, 2015). They considered a similar set of sequence properties, native secondary structure and signal peptides, selected PTMs, and a few new properties: the number of germline variants, expression levels, and the number of PPIs (Bull and Doig, 2015). The most recent study by Park, Lee, and colleagues expanded the above list of characteristics by inclusion of gene essentiality and tissue specificity (Kim et al., 2017). Moreover, several articles narrowly focused on characteristics that quantify topological features of the underlying PPI networks (Zhu et al., 2009b; Zhu et al., 2009c; Mitsopoulos et al., 2015; Feng et al., 2017). While these studies have considered a broad range of functional and structural features of drug targets, they identified the drug target-specific characteristics by comparing the drug targets against the other human proteins (non-drug targets). However, many of these non-drug targets could be in fact druggable, i.e., as many as 22% according to (Finan et al., 2017). Using the non-drug targets to represent the non-druggable proteins in order to define characteristic features of the druggable targets ultimately creates a bias toward describing the currently known drug targets. Consequently, this reduces our ability to use these characteristics to identify a complete set of druggable proteins.

We address the abovementioned shortcoming of the prior works by comparing sequence-derived characteristics of the drug targets, possibly druggable proteins, and non-druggable proteins using a large and well-curated dataset of human proteins. Our study is novel in four ways. First, we contrast the drug targets (D dataset) not only against all non-drug targets (N dataset), which was also done in prior studies, but also against non-druggable non-drug targets (Nn dataset; the non-drug targets that exclude disease associated proteins) and against possibly druggable non-drug targets (Nd dataset; the non-drug targets that are associated with multiple diseases). The association of the non-drug targets with diseases is necessary for the druggable proteins to exert therapeutic effects. Second, we further compare the D, N, Nd, and Nd proteins against highly promiscuous drug targets that interact with many drugs (Dh dataset) and drug targets that interact with low number of drugs (Dl dataset). This full-spectrum analysis allows us to pinpoint characteristics that differentiate between drug targets, possibly druggable proteins and non-druggable proteins, as well as features that are specific to promiscuous *vs.* non-promiscuous drug targets. Third, we focus on the characteristics that can be quantified directly from the protein sequence or protein name/identifier. This facilitates their use as potential markers for druggability across the entire human proteome. This is in contrast to several related studies that are limited to a relatively small subset of human proteins with solved structures (Hambly et al., 2006; Bull and Doig, 2015; Hu et al., 2016; Wang et al., 2016a; Wang et al., 2019). Fourth, we include several important sequence/protein-derived characteristic that were missed in the past studies including putative intrinsic disorder, residue-level conservation, presence and number of alternative splicing isoforms, inclusion of domains, and solvent

accessibility (surface area). Moreover, we cover some of the key characteristics from the prior works, such as the topological features of PPIs, cellular functions, and subcellular locations.

MATERIALS AND METHODS

Datasets

Datasets of Drug Targets (D Dataset), Highly Promiscuous Drug Targets (Dh Dataset), and Low-Interaction Drug Targets (Dl Dataset)

We collect a comprehensive set of drug targets by combining interaction information extracted from several large bioactive compounds-protein interaction databases. We filter these bioactive compounds to include only approved and experimental drugs. Furthermore, we focus on human proteins by excluding protein fragments and proteins from other organisms. We maximize the coverage by first collecting an inclusive set of interactions (including all bioactive compounds and protein chains) and then applying the two filters to obtain a high quality and large set of drugs and proteins.

The data collection protocol follows the work in (Wang and Kurgan, 2019; Wang and Kurgan, 2018). We extract the source data from three large repositories: Drug2gene (Roider et al., 2014), TTD (Zhu et al., 2009a), and GtP (Harding et al., 2017). Drug2gene is one of the most inclusive repositories that aggregates 19 source databases including TTD and GtP and several other major databases like ChEMBL (Gaulton et al., 2016) and DrugBank (Wishart et al., 2017). However, Drug2gene includes older and substantially smaller version of the TTD and GtP resources. Therefore, we integrated the latest versions of these two databases into our dataset. These databases provide a list of drug-protein pairs that use different identifiers and which include other information that could be useful to identify these molecules (like drug structure). The arguably most popular way to identify drugs and proteins are the PubChem CIDs and UniProt accession numbers, respectively. We use these identifiers to map data between the resources. We also merged the drugs with different PubChem CID but identical *simplified molecular-input line-entry system* (SMILES) structures. First, we remove the data collected from TTD and GTP that lacks PubChem CID or UniProt identifiers. Next, we map the proteins in Drug2gene that are represented by Entrez Gene ID into the corresponding UniProt accession numbers. After mapping and combining these datasets and removing duplicates, we obtain 2,490,057 interactions for 591,684 bioactive compounds and 4,128 proteins. Next, we filter this list of compounds using the list of drugs obtained from the DrugBank and ChEMBL. We remove the compounds that do not have the same CID or SMILES structure when compared to the list of DrugBank and ChEMBL drugs. Finally, we remove non-human proteins using a reference human proteome from UniProt. At the end, the set of drug targets (D dataset) includes 33,104 interactions between 4,405 drugs (PubChem CID) and 1,638 protein (UniProt identifiers). We provide the complete D dataset in the **Supplementary Material**. Moreover, we generate an expanded set of human and human-like drug targets that includes proteins in the D dataset plus proteins from other organisms that share high sequence similarity to the human proteins (D+ dataset). More specifically, following recent works (Hu et al., 2014; Wang et al., 2016a; Wang et al., 2019), human proteins that share at least 90% sequence identity quantified using BLAST with default parameters (Altschul et al., 1997) to any of the drug targets were added into the D+ dataset. Consequently, the D+ dataset has 1,762 proteins including 124 proteins that were included based on the high similarity; we list these proteins in the **Supplementary Material**. The number of drug targets in our dataset is slightly higher than the sizes of the datasets used in related studies (in the inverse chronological order): 1604 in (Feng et al., 2017), 1578 in (Kim et al., 2017), 1324 in (Bull and Doig, 2015), and 1,030 in (Rask-Andersen et al., 2014). Compared to popular databases, such as KEGG DRUG and DrugBank, our dataset features a more complete set of interactions (33,104 *vs.* 14,222 and 23,380, respectively (Wang and Kurgan, 2019) while focusing on a smaller and relevant set drugs that specifically target human proteins [4,405 *vs.* 5,045 and 10,562, respectively (Wang and Kurgan, 2019).

Drug targets in our dataset interact with as few as 1 drug and as many as 443 drugs. We investigate whether sequence-derived and functional characteristics of highly promiscuous drug targets are different from the drug targets that interact with a few proteins. To do that we extracted two subsets of the drug targets, the highly promiscuous targets (Dh dataset) that correspond to the top quartile of the targets with the highest interaction counts, and the low-interaction drug targets (Dl dataset) that include the bottom quartile of the drug targets with the lowest numbers of interactions.

Dataset of Non-Drug Targets (N Dataset)

We contrast the sequence-derived and functional characteristics of the proteins in the D, D+, Dh, and Dl datasets against the proteins that are not current drug targets. We collect these non-drug targets (N dataset) by selecting proteins from the UniProt's human proteome that are not in the D dataset. The selection process follows two rules. First, we match the size of the N dataset to the size of the D dataset to ensure robust statistical comparisons between different datasets. Second, when down-sampling the human proteins we ensure that the selected proteins have similar size as the proteins in the D dataset. More specifically, for each protein in the D dataset we pick a human non-drug target at random (without replacement) that has a matching sequence length (with 10% tolerance). We introduce the latter rule since the amount of intrinsic disorder in proteins is dependent on proteins length (HOWELL et al., 2012). The same selection process was used in several related studies (Meng et al., 2015b; Na et al., 2016; Meng et al., 2018) to eliminate protein size bias when studying intrinsic disorder. We provide the list of the 1,638 size-matched proteins that constitute the N dataset in the **Supplementary Material**. Moreover, Section "Non-druggable and possibly druggable proteins" describes how the N dataset is used to derive the dataset of Non-druggable non-drug targets (Nn dataset; the non-drug targets that exclude disease associated

proteins) and the dataset of possibly druggable non-drug targets (Nd dataset; the non-drug targets that are associated with multiple diseases).

Characterization of Protein Properties

We characterize a broad collection of characteristics of human proteins that include their disease associations, structural properties derived from the sequence (putative intrinsic disorder and surface), sequence properties (domain annotations, alternative splicing, and residue-level conservation), topological properties of the corresponding PPI network (centrality measures and hubs), and functional properties (GO annotations and predicted protein-binding regions). We extract these characteristics directly from the protein sequence or protein names/identifiers. This means that they could be used as potential markers for druggability that cover the entire human proteome.

Disease Associations

The protein-disease association data were collected from DisGeNET (Gutiérrez-Sacristán et al., 2016). DisGeNET integrates several curated databases and offers arguably one of the most complete levels of coverage for human diseases. This database provides association between disease MeSH IDs and Entrez Gene IDs and also provides a mapping between Entrez Gene IDs and UniProt identifiers. We mapped these annotations to our dataset using the UniProt identifiers.

Sequence-Derived Structural Properties

We annotate two relevant structural properties that we can accurately derive from the protein sequences: intrinsic disorder and solvent accessibility. We are unable to directly collect structural data since significant majority of the proteins in the D, D+, and N datasets do not have solved structures.

Intrinsically disordered proteins and protein regions lack a stable tertiary structure in isolation (Dunker et al., 2013; Habchi et al., 2014; Uversky, 2014a). Proteins with disordered regions are crucial for many key cellular functions including molecular recognition and assembly, cell cycle and cell death regulation, signal transduction, transcription, translation, and viral cycle (Dyson and Wright, 2005; Uversky et al., 2005; Liu et al., 2006; Xie et al., 2007; Peng et al., 2012; Xue et al., 2012; Peng et al., 2013; Uversky et al., 2013; Fan et al., 2014; Fuxreiter et al., 2014; Peng et al., 2014b; Xue and Uversky, 2014; Dolan et al., 2015; Meng et al., 2015a; Meng et al., 2015b; Varadi et al., 2015; Babu, 2016; Na et al., 2016; Yan et al., 2016; Wang et al., 2016b; Kjaergaard and Kragelund, 2017). They are also the main contributors to the dark proteome (Hu et al., 2018; Kulkarni and Uversky, 2018). Intrinsic disorder is abundant in the human proteins. Computational studies estimate that about 19% amino acids in eukaryotic proteins are intrinsically disordered (Peng et al., 2015) and over 40% human proteins have at least one long disordered region with 30 or more consecutive residues (Oates et al., 2013). These proteins are particularly relevant to this study since they are associated with several human diseases (Uversky et al., 2008; Babu, 2016; Uversky et al., 2014; Uversky, 2014b) and since they attract recent interest as potent drug targets (Cheng et al., 2006; Uversky, 2012; Dunker and Uversky, 2010; Ambadipudi and Zweckstetter, 2016; Tantos et al., 2015). Intrinsic disorder can be predicted accurately from protein sequence using computational methods (Peng and Kurgan, 2012; Walsh et al., 2015; Lieutaud et al., 2016; Meng et al., 2017a; Meng et al., 2017b). We use one of the leading disorder predictors, IUPred (Dosztányi et al., 2005; Dosztanyi, 2018). This selection is motivated by the fact that IUPred is computationally efficient (i.e., it can be used to process large datasets of proteins, such as the D and N datasets) and since it provides accurate predictions (Peng and Kurgan, 2012; Walsh et al., 2015). We use the IUPred's results to compute the disorder content (fraction of disordered residues in a given protein) and the length of the putative disordered regions.

Solvent accessibility provides a crucial context for the analysis of the residue-level conservation since it allows us to separate conserved residues that are localized on the surface (which include residues that are instrumental for the drug-protein interaction) from those located in the protein core (which are likely responsible for structural stability of the protein). We predict the relative accessible surface area using the ASAquick method (Faraggi et al., 2014). This method predicts relative solvent accessibility from a single sequence (without alignment), and thus it much faster than the other predictors that require calculation of multiple sequence alignment. It also provides accurate prediction, which is why it was recently used in related studies (Zhang et al., 2017; Amirkhani et al., 2018; Meng and Kurgan, 2018). We convert the numeric relative solvent accessibility of residues into a binary annotation (solvent exposed *vs*. buried) using a threshold of 0.15. This value adequately splits the bimodal distribution of solvent accessibility values for the residues in the combined D and N datasets (**Figure S2** in the **Supplementary Material**). We use these results to quantify the fraction of the putative surface residues in a given protein.

We assess quality of these predictions by comparing values of the fraction of the native surface residues that are computed using a limited set of proteins that have structures against the fraction of the predicted surface residues for the same set of proteins. We utilize mapping generated with the SIFTS resource (Velankar et al., 2013) that is available in UniProt to identify structures of the human proteins from the D and N datasets in the PDB database (Berman et al., 2000). We consider structures that cover at least 90% of the corresponding full protein sequences collected from UniProt to ensure that they correspond to a similar set of residues that are covered by the predictions which rely on the full protein chains. We compute the native solvent accessibility from these structures in three steps. First, we remove other molecules (including other protein chains) from the PDB structures. Second, we use DSSP (Kabsch and Sander, 1983; Joosten et al., 2010) to compute solvent accessibility values. Third, we convert the solvent accessibility into the relative solvent accessibility values using the normalization procedure that is described in the ASAquick article (Faraggi et al., 2014). We were able to collect the native solvent accessibility values for 373 drug targets (including 343 proteins from the D dataset, 55 from the Dh dataset, and 103 from the Dl dataset) and 73 proteins non-drug targets (including 39 from the Nd dataset and 12 from the Nn dataset). This corresponds to (373 + 73)/(1762 + 1,638) = 13% structural

coverage of the human proteins in our datasets. **Figure S3** compares the distributions of the fractions of the surface residues computed from the protein structures against the fractions that are based on the predicted solvent accessibility for the seven considered datasets. The distributions that rely on the native *vs*. putative solvent accessibility for each of the seven dataset are very similar. The differences are not statistically significant (p-values range between 0.17 for the N dataset and 0.88 for the Nd dataset). This results suggests that the solvent accessibility predicted with ASAquick provides an accurate approximation of the native fraction of the surface residues.

Protein Sequence Properties

We use the proteins sequences to annotate the domains, alternative splicing isoforms, and sequence conservation. We collect the domain annotations from Pfam (Calderone et al., 2013) using UniProt identifiers, and we use these annotations to compute the domain boundaries (fraction of the domain-assigned residues) and the number of domains per protein. We obtain the number of alternative splicing isoforms from the UniProt database (UniProt: the universal protein knowledgebase, 2016). We calculate residue-level conservation scores using the relative entropy measure (Wang and Samudrala, 2006) from the PSSMs generated with PSI-BLAST (Altschul et al., 1997). We use a threshold to convert the numeric conservation scores to binary, i.e., a given residue is either conserved (if its conservation score > threshold) or non-conserved (otherwise). We selected the threshold that corresponds to the 80th percentile of the distribution of the conservation scores for the residues in the combined D and N datasets (**Figure S1** in the **Supplementary Material**). The corresponding threshold value of 0.63 corresponds to an inflection point in the distribution tail where the conserved residues should be located. Using these annotations, we quantify the rate of the conserved residues in the protein sequence and among the residues located on the putative protein surface, given that this is where the drug-protein interaction occurs.

Topological Properties of the Protein-Protein Interaction Network

Motivated by work in (Zhu et al., 2009b; Zhu et al., 2009c; Mitsopoulos et al., 2015; Feng et al., 2017), we quantify the topological characteristics of drug targets and non-drug targets in the human PPI network. We collected the interaction network from the MENTHA resource (Calderone et al., 2013) and directly mapped it to our datasets using UniProt identifiers. MENTHA integrates data coming from several popular databases of PPIs, such as IntAct (Orchard et al., 2014), MINT (Licata et al., 2012), DIP (Salwinski et al., 2004), BioGRID (Oughtred et al., 2019), and MatrixDB (Launay et al., 2015), providing arguably one of the most comprehensive coverage levels. Several different centrality measures can be used to define topological characteristics of proteins in PPI networks (Wang et al., 2013a). We considered a comprehensive set of measures including betweenness centrality (Freeman, 1977), eigenvector centrality (Bonacich, 1987), closeness centrality (Bavelas, 1950), information centrality (Stephenson and Zelen, 1989), degree centrality (Jeong et al., 2001), subgraph centrality (Estrada and Rodriguez-Velazquez, 2005), network centrality (Wang et al., 2012a), and local average connectivity (Li et al., 2011). We reduced this set by removing measures that are redundant (highly correlated). The corresponding subset of four measures (eigenvector, closeness, betweenness and information centrality) has relatively low mutual correlations (< 0.6) while being highly correlated (> 0.8) with at least one of the removed measures. We give the corresponding correlations between these measures on our datasets in **Table S1** in the **Supplementary Material**. The eigenvector centrality is an extension of the node degree in which connections to more important nodes have more impact on the score. The nodes that are connected to many highly connected nodes end up having higher score than nodes which are connected to the same number of less-connected nodes (Bonacich, 1987). The closeness centrality measures the average length of the shortest path from the node to other nodes. The nodes with higher closeness centrality on average have smaller distance to the other nodes (Bavelas, 1950). The betweenness centrality quantifies the frequency with which a given node appears in the shortest paths between nodes in the network. Thus, removal of nodes with high betweenness centrality has big impact on the shortest paths between nodes (Freeman, 1977). Finally, information centrality is based on information along the paths from a given node to the other nodes (Stephenson and Zelen, 1989).

Besides quantifying several different topological features, we also annotate hub proteins, defined as proteins that interact with many proteins (Jeong et al., 2001). While early works on hub proteins defined them using a fixed minimal number of (Jeong et al., 2001), more recent studies use a floating threshold defined as a certain percentage of the most connected nodes in a given interactome (Han et al., 2004; Batada et al., 2006; Dosztányi et al., 2006). This results in different cut-offs that define hubs for different interactomes (different organisms) and emphasizes the fact that hubs are a property of the whole interactome system rather than a property of individual proteins. We used the latter definition using the cut-off that corresponds to the 90th percentile of the interaction counts in the complete human PPI network, which is consistent with several recent studies (Han et al., 2004; Batada et al., 2006; Dosztányi et al., 2006). Therefore, we annotate hub proteins as those that have the number of PPIs in the complete interactome collected from MENTHA that is higher than this threshold (i.e., ≥ 77 interactions).

Hub proteins have increased levels of intrinsic disorder (Meng et al., 2015b; Patil et al., 2010) and the disordered regions are often employed to carry out PPIs (Mohan et al., 2006; Vacic et al., 2007; Yan et al., 2016). The disordered protein-binding regions are also linked to certain human diseases (Uversky, 2018). Thus, we also annotate putative disordered protein binding regions. We use ANCHOR (Dosztányi et al., 2009) to predict the disordered protein-binding residues and we aggregate this information to compute the content of disordered protein binding residues for the proteins in our datasets. The selection of this method is motivated by the fact that it is accurate and popular, and provides fast predictions (i.e., is capable of processing our large datasets) (Meng et al., 2017; Katuwawala et al., 2019).

Functional Properties

We annotate cellular functions and subcellular locations of the drug targets and the non-drug targets using the Gene Ontology (GO) terms (Consortium, 2004), which we collect using the PANTHER system (Muruganujan et al., 2018). We annotate and separately analyze the molecular functions, biological processes, and cellular components, where the latter define the subcellular locations.

Statistical and Similarity Analyses

We compare the sequence-derived and functional characteristics between the drug targets, non-drug targets, and possibly druggable proteins using statistical tests of significance of differences. We quantify the significance of the differences using the *t*-test if the underlying measure of the sequence-derived/functional property has normal distribution, and Wilcoxon rank-sum test otherwise. We used the Anderson-Darling test with the *p*-value cutoff of 0.05 to test normality. We use the Fisher's exact test when comparing binary characteristics, including disease associations and presence of hubs.

We annotate the cellular functions and subcellular locations associated with a particular set of proteins using enrichment analysis offered by the PANTHER system (Muruganujan et al., 2018). This system generates a list of annotations that are statistically over-represented when compared with the annotations present in the whole human proteome. PANTHER quantifies the ratios of enrichment and the corresponding *p*-values for each GO term when compared with the reference human proteome. We focus on the GO terms that occur at least 10 times in our datasets (to ensure robustness of statistical analysis), and we annotate a given term as associated with a particular set of proteins if its ratio > 2 (at least two fold increase) and the associated *p*-value (quantified using the False Discovery Rate correction) is < 0.05.

We measure similarity between two sets of proteins by comparing the cellular function and subcellular location GO terms associated with these two protein sets. We calculate this similarity using the GOSemSim package (Li et al., 2010) with default parameters [Wang et al. measure (Wang et al., 2007)] and the reference set to human.

RESULTS AND DISCUSSION

Non-Druggable and Possibly Druggable Proteins

The set of the non-drug targets likely includes a relatively large number of druggable proteins. The ability to characterize properties that differentiate the drug targets and druggable proteins from the non-drug targets hinges on the annotation of the non-druggable and possibly druggable proteins in the set of these non-drug targets. Druggability of proteins requires that they interact with a drug-like compound and that this interaction provides a desired therapeutic effects (Hopkins and Groom, 2002; Russ and Lampel, 2005; Keller et al., 2006). Thus, one way to annotate possibly druggable and non-druggable proteins is to analyze protein-disease associations. **Figure 1** shows the fractions of the proteins associated with different classes of diseases among the drug targets and the non-drug targets. As expected, the

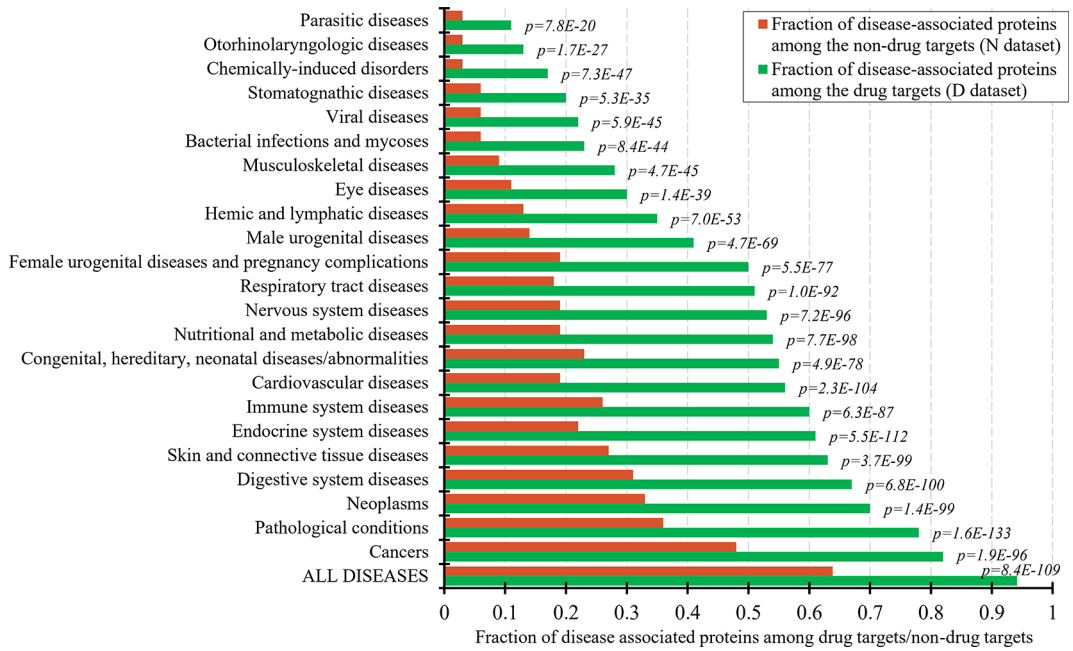

FIGURE 1 | Fraction of drug targets and non-drug targets associated with different classes of diseases. The green and red bars show the fraction of disease associated proteins among the drug targets and non-drug targets for each disease class. The *p*-values quantify the significance of the differences between the two fractions using the Fisher's exact test. The disease classes are sorted by the value of the fraction of the drug targets.

number of the disease associated proteins is significantly higher among the drug targets compared to the non-drug targets. This difference is statistically significant for each of the 23 diseases classes (p-values < 0.0001). About 94% of the drug targets are associated with at least one disease, attesting to the relatively high coverage of these annotations and supporting the fact that the drug targets exert therapeutic effects. The largest fraction of the drug targets (82%) is associated with cancers. To compare, only about 64% of the non-drug targets are disease-associated. The latter suggests that the non-drug targets include both non-druggable proteins (those that lack association with any of the diseases) and possibly druggable proteins (those that are associated with diseases). We note that the use of the diseases associations provides a partial support for their druggability since it does not address the ability of the possibly druggable proteins to interact with drug-like molecules.

Figure 2 analyzes relation between the drug targets, non-drug targets, and disease associations. **Figure 2A** reveals that the disease-associated proteins are likely to be drug targets. About 60% of proteins that are associated with at least one disease are drug targets. The fraction of drug targets increases for the proteins that are associated with more disease. This increase is sharper for a lower number of diseases and plateaus for proteins with about 10 or more disease associations. Therefore, we hypothesize that the non-drug targets with a relatively large number of disease associations can be used as a proxy for possibly druggable proteins. We use the inflection point in **Figure 2A**, which corresponds to proteins with ≥13 disease associations among which 75% are drug targets, to define the set of possibly druggable proteins. **Figure 2B** is a Venn diagram that visualizes overlap between the disease associated proteins (black borders), the drug targets (dataset D; green border), and the non-drug targets (dataset N; red border). We define the set of the non-drug targets that are associated with 13 or more diseases as possibly druggable proteins (Nd dataset; orange area in **Figure 2B**). **Figure 2B** also shows that virtually all drug targets are associated with at least one disease (black border with number of diseases K ≥ 1), while a large portion of the non-drug targets lacks any disease associations (brown area in **Figure 2B**). The latter set of proteins constitutes the set of the non-druggable proteins (Nn dataset).

We test reliability of annotations of the possibly druggable and non-druggable proteins using the 124 human-like drug targets from the D+ dataset that were annotated based on their high sequence similarity to drug targets in other organisms. We found only 4% (5 of the 124) of the human-like drug targets among the 4,869 non-drug targets that are not associated with diseases compared to 67% (83 human-like drug targets) that are among the 4,287 non-drug targets that are associated with 13 or more diseases. The high degree of the latter overlap suggests that the Nd dataset should include a substantial number of druggable proteins. We note that the 4% overlap with the non-drug targets

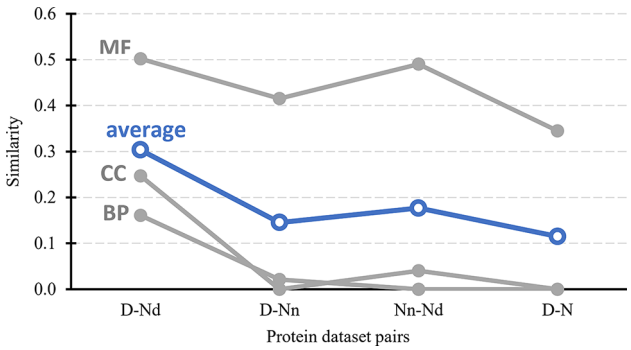

FIGURE 3 | Similarity in cellular processes and subcellular locations between the drug targets (D dataset), possibly druggable proteins (Nd dataset), non-druggable proteins (Nn dataset), and non-drug targets (N dataset). We measure similarity for four pairs of these datasets (D *vs*. Nd, D *vs*. Nn, D *vs*. N, and Nn *vs*. Nd) based on the comparison of the corresponding sets of GO terms associated with these datasets, i.e., GO terms over-represented in a given dataset when compared to the entire human proteome. The GO terms are divided into three categories: MF (molecular functions), BP (biological processes), and CC (cellular components). Similarity was measured with the GOSemSim package (Li et al., 2010). We describe details of these calculations in section "Statistical and similarity analyses". The gray markers show the similarity for each GO-term category while the blue markers are the average across the three categories.

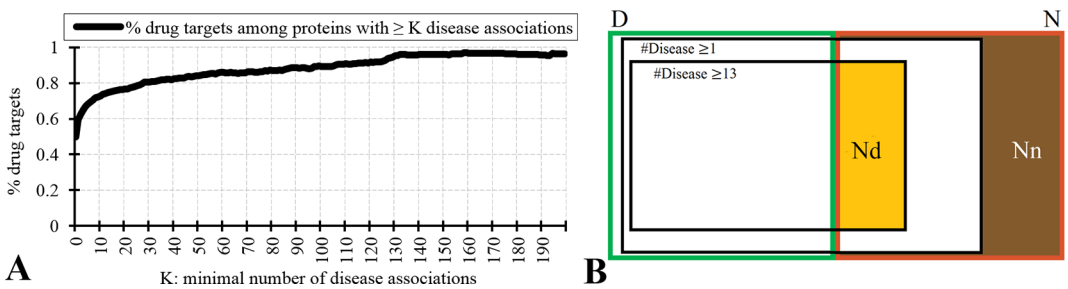

FIGURE 2 | Relation between drug targets, non-drug targets and diseases associations. Panel **A** shows the fraction of the drug targets among proteins associated with a given minimal number of diseases K. Panel **B** is a Venn diagram that visualizes overlap between the disease associated proteins (with K = 1 and K = 13), the drug targets (dataset D; green border), and the non-drug targets (dataset N; red border). Among the non-drug targets we define the Nn dataset of non-druggable proteins (brown area), i.e., the non-drug targets that are not associated with any disease, and the Nd dataset of possibly druggable proteins (orange area), i.e., the non-drug targets that are associated with 13 or more diseases.

that lack diseases associations likely stems from incompleteness of the diseases association data.

Figure 3 further tests the validity of the hypothesis that the Nd and Nn datasets include the possibly druggable and the non-druggable proteins, respectively. It quantifies similarity in the context of cellular functions and subcellular location between the drug targets, possibly druggable proteins, non-druggable proteins, and the non-drug targets. First, we generate a set of GO terms that are associated with each of these datasets, i.e., GO terms over-represented in a given dataset when compared to the human proteome. We perform this analysis separately for each of the three GO terms categories: molecular functions, biological processes, and cellular components; the latter is a proxy for the subcellular location. Next, we calculate similarity between the corresponding sets of dataset-specific GO terms; we describe the details in section "Statistical and similarity analyses". The gray lines in **Figure 3** shows the similarity values for each GO term category while the blue lines show the average across the three categories. The left-most set of results reveals that the cellular functions and subcellular location of the drug targets (D dataset) are similar to the possibly druggable proteins (Nd dataset), which aligns with our hypothesis that the Nd dataset in fact includes druggable proteins. The second set of results, which compares the drug targets against the non-druggable proteins (Nn dataset), shows lack of similarity in the biological processes and subcellular locations and modestly reduced levels of similarity in the molecular functions. The corresponding average similarity = 0.145 is lower by a factor of two when compared with the similarity = 0.303 between the drug targets and possibly druggable proteins. The other two sets of results, which compare the possibly druggable against the non-druggable proteins and the drug targets against the non-drug targets, similarly reveal the lack of similarity in the biological processes and subcellular locations, while showing similarity in the molecular functions. The average similarities for these two dataset pairs are low and equal 0.177 and 0.115, respectively, suggesting that the corresponding two pairs of datasets include proteins involved in distinct cellular processes and subcellular locations. To sum up, the above analysis demonstrates that drug targets and the possibly druggable proteins share much higher levels of functional and subcellular location similarity compared to the similarity between possibly druggable proteins, non-druggable proteins, and non-drug targets. This finding, which uses an independent source of information compared to the approach we used to annotate the possibly druggable proteins, supports validity of our annotations of the possibly druggable and the non-druggable proteins.

Comparative Analysis of the Sequence-Derived Structural and Functional Characteristics of the Drug Targets, Possibly Druggable, and Non-Druggable Proteins

Our ability to identify novel druggable proteins relies on the understanding of functional and sequence-derived characteristics that differentiate drug targets from the non-drug targets. We focus specifically on the characteristics that can be quantified from the protein sequence and/or identifier, which allows for a proteome-wide deployment. We compare a broad range of these characteristics between the drug targets, non-drug targets, possibly druggable proteins, and non-druggable proteins. We also investigate differences between the above protein sets and the expanded set of drug targets that includes human and human-like targets (D+ dataset), highly promiscuous drug targets that interact with many drugs (Dh datasets), and drug targets that interact with a low number of drugs (Dl dataset).

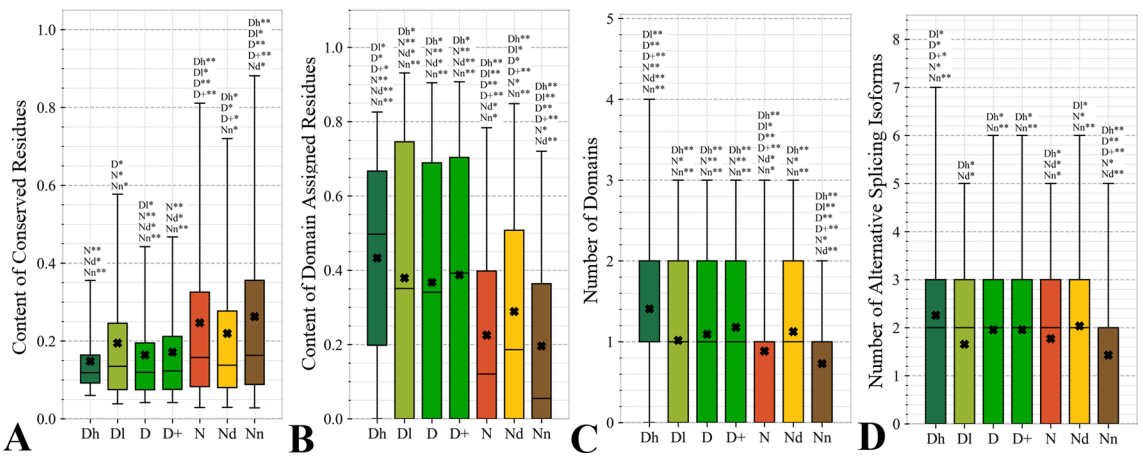

FIGURE 4 | Distributions of the values of the sequence-derived characteristics for the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). (Panels **A**) shows the amount of conserved residues. Panels **B** and **C** focus on the protein domains while Panel **D** quantifies the number of splicing isoforms. The whiskers show the 5 and 95 percentiles, the top and bottom of the box correspond to the first and third quartiles, the middle bar is the median, and the cross marker is the average. The annotation above the whiskers show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section "Statistical and similarity analyses".

Characteristics Derived From the Protein Sequence

Figure 4 focuses on the characteristics derived directly from the protein sequence, including the residue-level conservation (content of conserved residues in protein chains), number of domains and the content of domain-annotated residues, and the number of the alternative splicing isoforms. **Figure 4A** shows that the drug targets (both D and D+ datasets) have significantly fewer conserved residues than the non-drug targets, possibly druggable proteins and the non-druggable proteins (p-value < 0.05). The possibly druggable proteins (orange bars) have significantly lower numbers of conserved residues compared to the non-druggable proteins (brown bars) (p-value < 0.05). Moreover, the highly promiscuous drug targets have significantly lower numbers of the conserved amino acids than the non-drug targets and the non-druggable proteins (p-value < 0.05), while maintaining similar levels compared to the possibly druggable proteins. Altogether, relatively low numbers of the conserved residues are characteristics for the drug targets and these numbers are also relatively low among the possibly druggable proteins. Interestingly, the residue-level conservation of the residues on the protein surface, where the protein-drug interaction occurs, follows the same pattern (**Figure 5E**). This finding complements prior results that show that drug targets have lower evolutionary rates and higher similarity to orthologous genes (Lv et al., 2016).

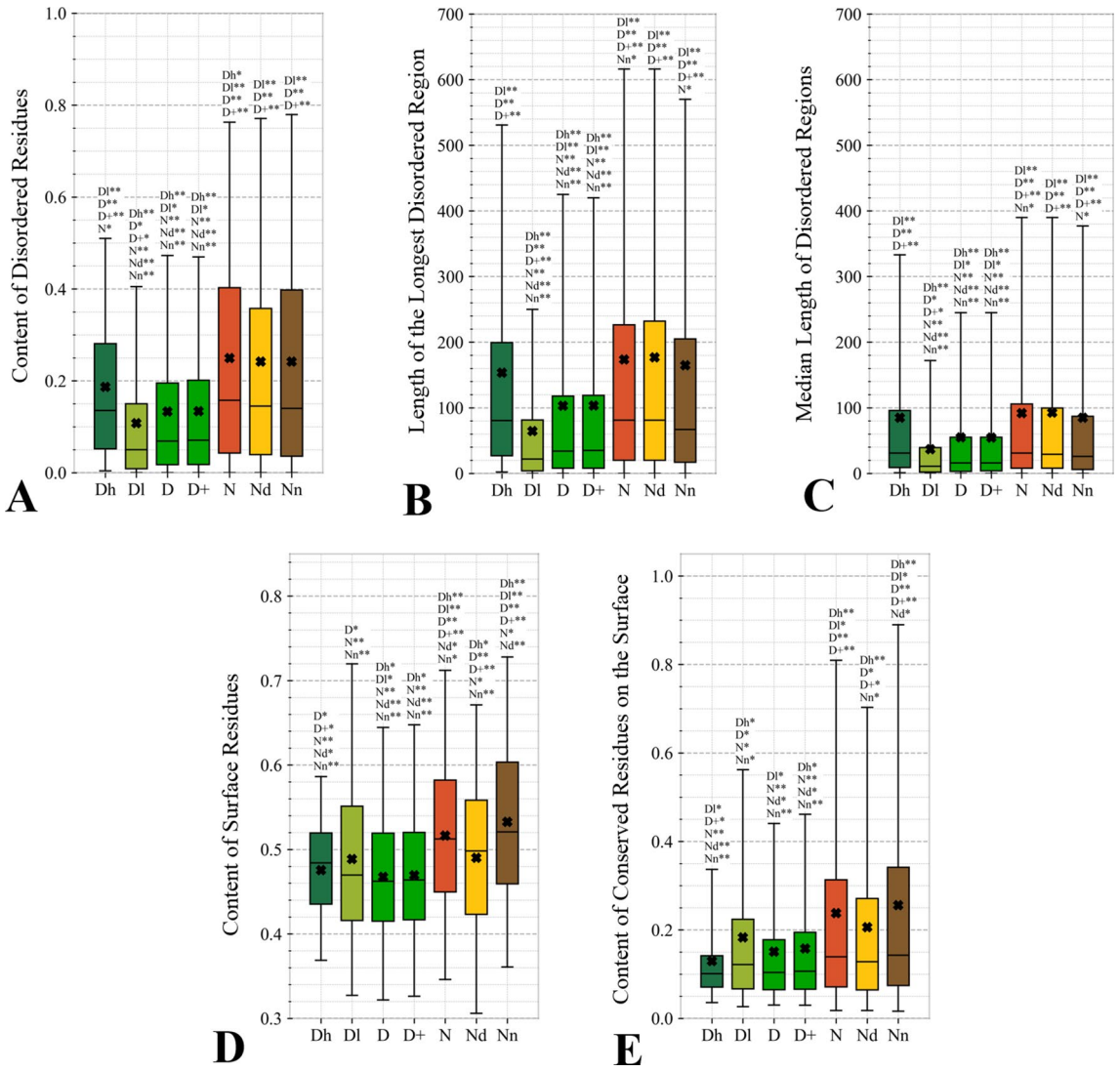

FIGURE 5 | Distributions of the values of the sequence-derived structural characteristics predicted from the protein sequence for the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). Panels **A**, **B**, and **C** quantify the abundance of intrinsic disorder while Panels **D** and **E** quantify the amount of surface and the amount of conserved residues on the surface, respectively. The whiskers show the 5 and 95 percentiles, the top and bottom of the box correspond to the first and third quartiles, the middle bar is the median, and the cross marker is the average. The annotation above the whiskers show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section "Statistical and similarity analyses".

Figures 4B, C reveal that the drug targets (both D and D+ datasets) have substantially more domains and have larger amounts of domain-annotated residues when compared to the non-druggable proteins (p-value < 0.0001). At the same time, they a similar number domains when contrasted with the possibly druggable proteins. Furthermore, the possibly druggable proteins have significantly higher levels of domain annotations when contrasted against the non-druggable proteins (p-value < 0.0001). The underlying reasons for this enrichment could be two-fold. First, there could be proportionally more multi-domain proteins among the drug targets and the possibly druggable proteins. Consequently, inclusion of a larger number of domains could increase the likelihood that these proteins host at least one druggable domain. However, our result could also mean that these proteins are more studied and understood, and thus their domain annotations are more complete. Moreover, the fact that at least close to half of proteins in all considered datasets have domain annotations, which suggests that they are functionally annotated, suggests that our functional similarity analysis in **Figure 3** should be robust.

The drug targets (both D and D+ datasets) and the possibly druggable proteins have significantly more splicing isoforms compared to the non-druggable proteins (p-value < 0.05) and this increase is even higher for the promiscuous drug targets (p-value < 0.001). This suggests that enrichment in the number of alternative splicing variants could serve as a marker for druggability. The alternative splicing was found to contribute to drug resistance (Siegfried and Karni, 2018; Zhao, 2019), which supports veracity of our result. Interestingly, recent studies suggest that targeting alternative splicing events could lead to therapeutic opportunities (Le et al., 2015; Siegfried and Karni, 2018). Our analysis also reveals that majority of the drug targets and the possibly druggable proteins have multiple isoforms. Thus, gene level analysis of drug targets may not be adequate, considering that these genes would encode multiple proteins.

Overall, we identified three potential sequence-derived markers of druggability. The drug targets and possibly druggable proteins share lower numbers of conserved residues and are more likely to have multiple domains and isoforms when compared to the non-druggable proteins. We also note that the results for the original set of human drug targets (D dataset) are consistent with the results for the expanded set of drug targets (D+ dataset).

Sequence-Derived Structural Properties

This study is the first to analyze two relevant sequence-derived structural characteristics that can be accurately predicted from the protein sequence: intrinsic disorder and solvent accessibility. Proteins with disordered regions are associated with a wide range of human diseases (Uversky et al., 2008; Uversky et al., 2014; Uversky, 2014b; Babu, 2016) while solvent accessibility determines protein surface where the drug-protein interaction happens. We note that while authors in (Kim et al., 2017) computed putative solvent accessibility, they only used it to analyze results concerning enrichment in the PTMs.

Figures 5A–C quantify two key aspects of the disorder: the overall content of disordered residues and the length of disordered regions. Proteins with higher disorder content are functionally distinct from structured proteins while long disordered regions are thought to correspond to disordered protein domains (Tompa et al., 2009; Pentony and Jones, 2010; Peng et al., 2014a). We observe that drug targets (both D and D+ datasets) are significantly less disordered (by a factor of two) and include much shorter disordered regions when compared with the non-drug targets, including both possibly druggable and non-druggable proteins (p-value < 0.001). This is in agreement with a recent study that demonstrates that the current drug targets are biased to exclude disordered proteins (Hu et al., 2016). There are several reasons for this bias. The protein structures are used during the rational drug design process (Gane and Dean, 2000; Lundstrom, 2006; Mavromoustakos et al., 2011; Lounnas et al., 2013) and to gain mechanistic insights into the protein-drug interactions (Pielak et al., 2009; Tan et al., 2013; Christopoulos, 2014) (Altschul et al., 1997; Wang and Samudrala, 2006; Calderone et al., 2013; Orchard et al., 2014; UniProt: the universal protein knowledgebase, 2016). The structures are also indispensable for modeling associated with drug repurposing and repositioning (Moriaud et al., 2011; Ma et al., 2013). This is while proteins with disordered regions are much less likely to have structures (Hu et al., 2018), partly because since they are explicitly avoided in the structural genomics pipeline (Linding et al., 2003; Oldfield et al., 2005; Mizianty et al., 2014). Interestingly, the highly promiscuous drug targets are enriched in disorder when contrasted with the overall set of drug targets and the low promiscuity drug targets (p-value < 0.0001), while their disorder levels are comparable to the possibly druggable proteins. This coincides with the observation that disordered regions are capable of interactions with multiple partners (Oldfield et al., 2008; Hu et al., 2017). Our results suggests that although low disorder amounts are a strong marker for the current drug targets, the set of possibly druggable proteins includes large amounts of disorder. In fact, the disordered proteins may become the key to unlocking a substantial portion of yet to be discovered druggable targets (Uversky, 2012; Hu et al., 2016), especially given their association with numerous human diseases (Uversky et al., 2008; Uversky et al., 2014; Uversky, 2014b; Babu, 2016).

The amount of the putative surface residues for the drug targets (both D and D+ datasets) is significantly smaller that for the non-drug targets, including the possibly druggable and non-druggable proteins (p-value < 0.0001), see **Figure 5D**. This could be driven by the fact that drug targets are often membrane proteins (Yildirim et al., 2007; Rajendran et al., 2010), which means that they have relatively low surface area compared to other proteins. They are also mostly structured proteins (Hu et al., 2016) that are more likely to have globular shape with more buried residues compared to more irregularly shaped/elongated disordered proteins (Peng et al., 2014b; Uversky, 2017). Moreover, presence of disordered regions on the protein surface also leads to an increase of the surface area compared to structured conformations (Wu et al., 2015). Interestingly, the possibly druggable proteins have comparable content of the putative surface residues with the low promiscuity drug targets, which is also significantly smaller when contrasted with the non-druggable proteins (p-value < 0.0001). This again, like in the case of the results in **Figure 4**, shows that the possibly druggable proteins are more similar to drug targets than to the non-druggable proteins. Finally, we observe that the

number of conserved residues on the putative surface (**Figure 5E**) maintains the same relation between the different protein sets as the overall number of conserved residues shown in **Figure 4A**, i.e., significantly lower for drug targets (both D and D+ datasets), and lower for the possibly druggable proteins compared to the non-druggable proteins (p-value < 0.05).

Topological Features of the Protein-Protein Interaction Networks

Topological features of the PPI networks are among the most studied characteristics of the drug targets (Zhu et al., 2009b; Zhu et al., 2009c; Bull and Doig, 2015; Mitsopoulos et al., 2015; Feng et al., 2017; Kim et al., 2017). A unique aspect of our analysis is that we focus on a set of orthogonal measures, i.e., measures that have low mutual correlations. This offers a more focused and balanced analysis given the high degree of similarity between many of these measures. **Figure 6** reveals that the entire set of four measures of centrality has significantly higher values for the drug targets (both D and D+ datasets) compared to the non-druggable proteins (p-value < 0.0001). Our results are in line with several prior studies that correspondingly show that drug targets have more connected and denser local network neighborhoods (Zhu et al., 2009b; Zhu et al., 2009c; Mitsopoulos et al., 2015; Lv et al., 2016). This finding suggests that drug targets are possibly more relevant biologically or are at a higher point of control and thus can better modify physiology, making them better therapeutic targets. The novel element in our study is that we find that all considered network centrality measures for the possibly druggable are even higher than for the drug targets (orange *vs.* green bars in **Figure 6**; p-value < 0.05). Consequently, they are also significantly higher than for the non-druggable proteins (orange *vs.* brown bars in **Figure 6**; p-value < 0.0001). Thus, our study suggests that these measures can be used as markers of druggability.

Figure 7 analyzes the abundance of the PPI network hubs among the drug targets, possibly druggable and non-druggable proteins. Approximately 17% of the drug targets (for both D and D+ datasets) are hubs and this rate is significantly higher than the 12% rate for the non-drug targets (green *vs.* red bars; p-value < 0.0001). Similarly large difference was observed in (Mitsopoulos et al., 2015). Our study reveals additional important details. We observe

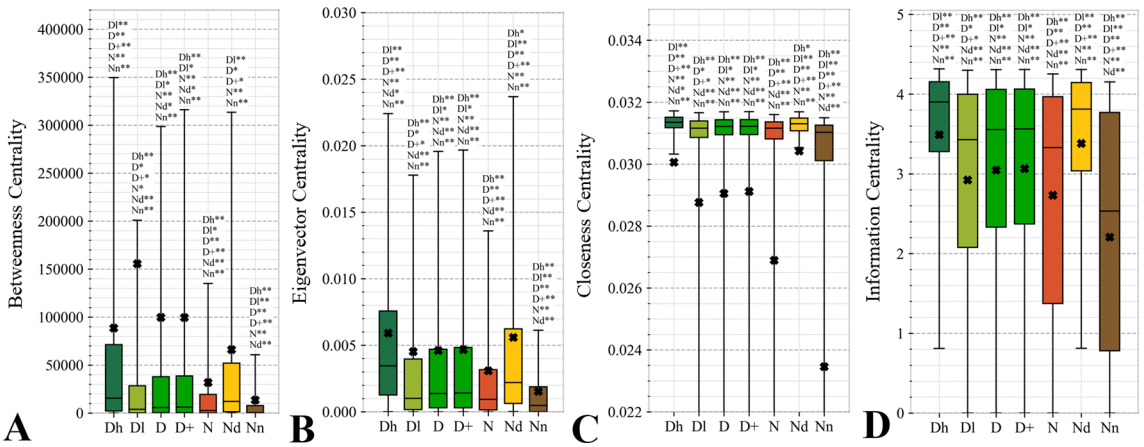

FIGURE 6 | Distributions of the values of the selected orthogonal PPI network properties for the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). Panels **A**, **B**, **C**, and **D** concern the betweenness centrality, eigenvector centrality, closeness centrality, and information centrality measures, respectively. The whiskers show the 5 and 95 percentiles, the top and bottom of the box correspond to the first and third quartiles, the middle bar is the median, and the cross marker is the average. The annotation above the whiskers show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section "Statistical and similarity analyses".

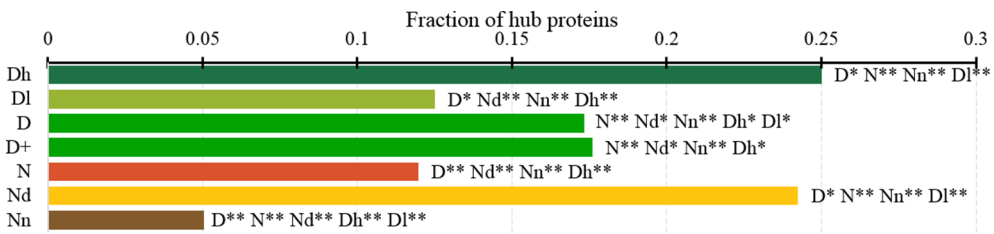

FIGURE 7 | Fraction of hub proteins among the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). The annotation next to the bars show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section "Statistical and similarity analyses".

that the rate of hubs is very high among the highly promiscuous drug targets (25%) and the possibly druggable proteins (24%), and these rates are significantly higher than the 12% rate for the non-drug targets (p-value < 0.0001) and the 5% rate for the non-druggable proteins (p-value < 0.0001). This suggests that high connectivity in the PPI network is a strong marker for druggability.

Functions and Subcellular Locations of Drug Targets and Possibly Druggable Proteins

Several studies analyzed cellular functions and subcellular locations of the drug targets (Lauss et al., 2007; Bakheet and Doig, 2009; Wang et al., 2013b). The green bars in **Figure 8** provide a list of significantly enriched functions and locations for our set of drug targets. Our results indicate that most of the drug targets are enzymes, including kinases and oxidoreductases, followed by substatial numbers of channels, and in particular ion channels. They are often involved in binding, signalling, regulation, and transport. These finding are in close agreement with the results in (Bakheet and Doig, 2009). **Figure 8** also shows that drug targets are primarily found in membranes, with a large numbers also found in the cytoplasm and the intracellular space. Consistent results are found in (Bakheet and Doig, 2009; Wang et al., 2013b), and these subcellular locations also agree with the observation that membrane proteins are the prime targets for the development of therapeutics (Yildirim et al., 2007; Rajendran et al., 2010).

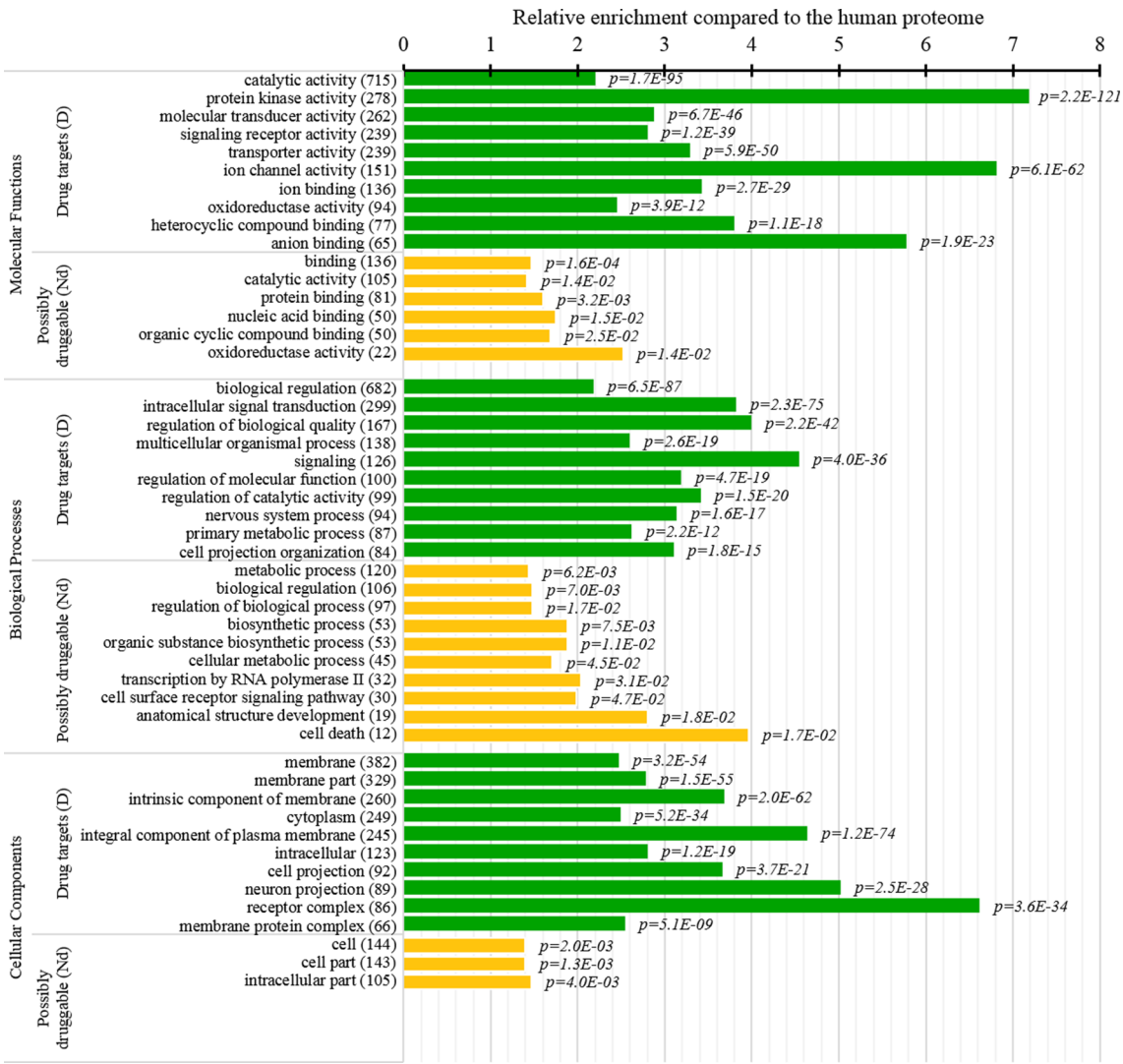

FIGURE 8 | Molecular functions, processes, and subcellular locations that are enriched among the drug targets (D dataset) and the possibly druggable proteins (Nd dataset). We show the top 10 (with the highest counts) over-represented/significantly enriched GO terms for the drug targets (green bars) and the possibly druggable proteins (orange bars). The bars quantify the ratios of enrichment relative to the human proteome and the corresponding p-values are shown on the right. GO terms are identified on the left, including their names and the number of the correspnding proteins in the given dataset. We explain calculation of statistical tests in section "Statistical and similarity analyses".

This study is the first to perform this type of analysis for the possibly druggable proteins (orange bars in **Figure 8**). Our analysis suggests that the possibly druggable proteins share functional similarities with the drug targets. They are similarly involved in the catalysis, signaling, and binding. However, the possibly druggable proteins tend to bind proteins and nucleic acids, instead of anions and ions which are the main partners for the drug targets. Moreover, the possibly druggable proteins are often involved in the metabolic and biosynthesis processes, and in the cell death cycle. The preference for the protein-protein and protein-nucleic acids binding and the cell death cycle involvement are supported by their significant enrichment in the intrinsic disorder (compared to the drug targets, see **Figures 5A, B**), and the fact that disordered regions are known to facilitate these types of functions (Vuzman and Levy, 2012; Uversky et al., 2013; Fuxreiter et al., 2014; Peng et al., 2015; Basu and Bahadur, 2016; Wang et al., 2016b; Hu et al., 2017; Srivastava et al., 2018). We further investigate this in **Figure 9** that analyzes the differences in the content of the putative disordered protein-protein binding regions. These results confirm the enrichment in the corresponding functional annotations for the possibly druggable proteins. The possibly druggable proteins include a substantial amount of the disordered protein-binding regions, on average about 14% of residues. Moreover, the drug targets (both D and D+ datasets) are significantly depleted in these protein-binding regions (on average only 7% of residues) when compared with the possibly druggable proteins (p-value < 0.0001). Interestingly, **Figure 8** also reveals that the possibly druggable proteins are localized across the cell and they do not have a specifically associated subcellular location, unlike the drug targets that are found mostly in the membranes and cytoplasm. Overall, our empirical analysis provides new insights into the cellular functions and subcellular locations of the druggable proteins.

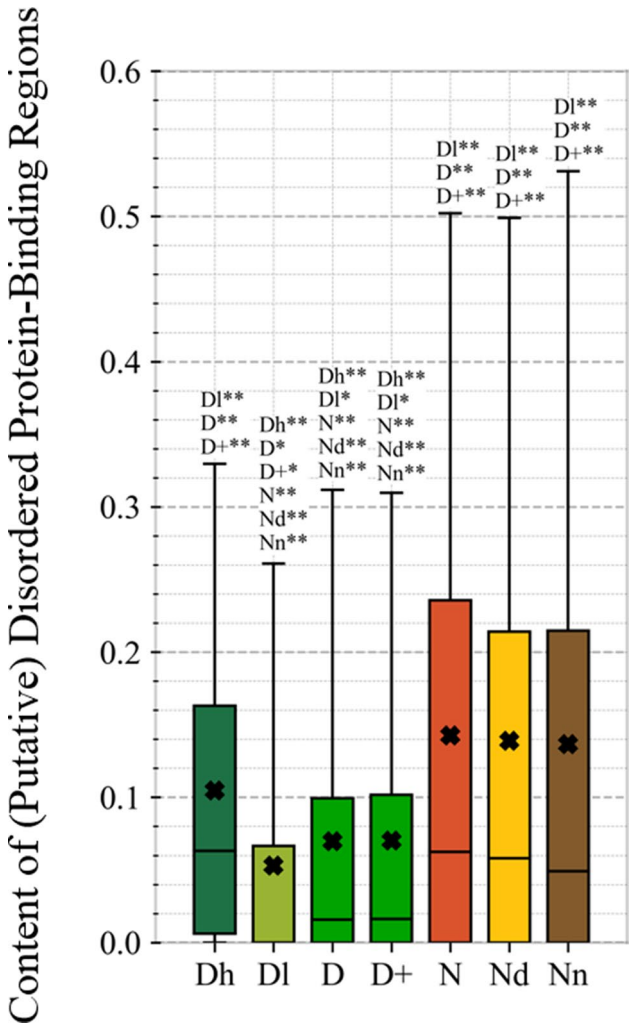

FIGURE 9 | Content of putative protein binding regions in the highly promiscuous drug targets (Dh), drug targets that interact with a low number of drugs (Dl), all drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). The annotation next to the bars show the significance of differences with the other protein sets; only significant differences are listed where N* means p-value 0.05 and N** means p-value 0.0001 when compared with the N dataset. We explain calculation of statistical tests in section "Statistical and similarity analyses".

SUMMARY AND CONCLUSIONS

Recent research approximates that the druggable human proteome has about 4,500 proteins (Finan et al., 2017), while there are about 1,600 current drug targets (1,750 drug targets if we include proteins that share high sequence similarity to drug targets that were annotated in other organisms). Annotation of the remaining druggable human proteins would facilitate development and screening of drugs, drug repurposing and repositioning, understanding and mitigation of drug side-effects, and prediction of drug–protein interactions. We contrast the drug targets against the possibly druggable and non-druggable proteins to identify markers that could be used to identify novel druggable proteins. This is in contrast to the prior studies that compare drug targets against non-drug targets (Zheng et al., 2006; Lauss et al., 2007; Bakheet and Doig, 2009; Zhu et al., 2009b; Zhu et al., 2009c; Bull and Doig, 2015; Mitsopoulos et al., 2015; Feng et al., 2017; Kim et al., 2017), thus producing markers that describe current drug target and which implicitly exclude the druggable proteins that are included in the non-drug target set. We annotate the possibly druggable and non-druggable proteins based on the presence and promiscuity of disease associations, and we validate these annotations *via* functional similarity analysis.

We cover a wide range of sequence-derived characteristics to define these markers. These characteristics can be computed across the entire human proteome, allowing for a complete sweep of all potential candidate proteins. We investigate several important characteristic that were missed in the past studies including putative intrinsic disorder, residue-level conservation, presence and number of alternative splicing isoforms, inclusion of domains, and putative solvent accessibility (surface area), as well as the key features from the prior works, such as the topological features of PPIs, cellular functions and subcellular locations. **Figure 10** summarizes the results. It shows the difference in the values of the key markers when comparing the possibly druggable proteins (in orange), the non-druggable proteins (in brown), all non-drug targets (in red), and the expanded set of human and human-like drug targets (in light

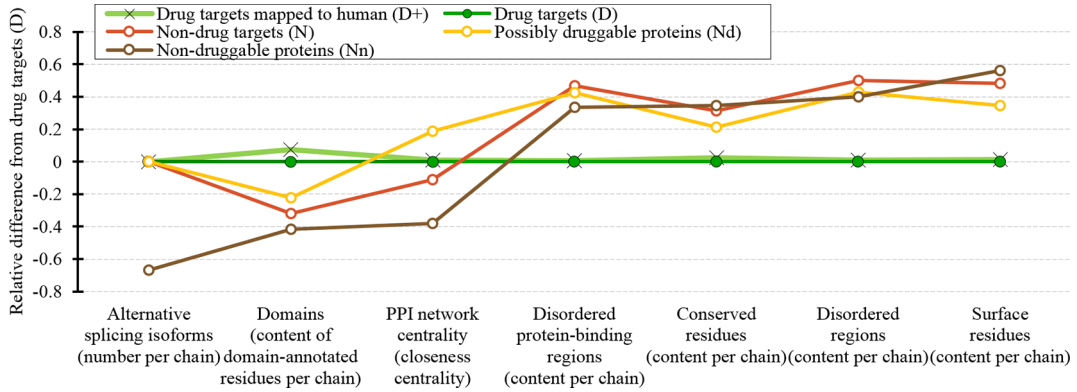

FIGURE 10 | Overview of the sequence-derived markers for the drug targets (D), all human and human-like targets (D+), non-drug targets (N), possibly druggable proteins (Nd), and non-druggable proteins (Nn). The y-axis quantifies the relative difference of the values of a given protein set X compared to the values of the drug targets (D) set defined as: [median(X)-median(D)]/IQR(D), where IQR means the interquartile range. The markers are sorted in the ascending order by the difference for the non-druggable proteins (in brown).

green) against the human drug targets (in dark green). We observe that the possibly druggable proteins are significantly more similar to the drug targets than the non-druggable proteins for majority of the markers. These markers include high abundance of alternative splicing isoforms, relatively large number of domains, higher degree of centrality in the corresponding PPI network (and correspondingly much higher rate of hubs), lower number of conserved residues, and lower number of residues on the putative (sequence-derived) surface. Thus, these factors could serve as high-quality markers for druggability. "Results and discussion" section discusses these findings in the context of the current literature. Moreover, **Figure 10** shows that drug targets (both D and D+ datasets) have significantly depleted levels of intrinsic disorder and intrinsically disordered protein-binding regions when compared with the much higher and comparable levels among the possibly druggable and non-druggable proteins. This suggests that the high levels of disorder combined with the presence of the abovementioned markers should be used together to effectively enlarge the current collection of drug targets. This is in accord with several recent studies that postulate inclusion of the disorder-enriched proteins into the set of druggable proteins (Cuchillo and Michel, 2012; Uversky, 2012; Chen and Tou, 2013; Joshi and Vendruscolo, 2015; Ambadipudi and Zweckstetter, 2016; Hu et al., 2016; Yu et al., 2016).

Our analysis also shows that the possibly druggable proteins are functionally similar to the drug targets, being involved in the catalysis, signaling, and binding. The main difference is that the possibly druggable proteins target interactions with proteins and nucleic acids, unlike the current drug targets that favor interactions with anions and ions. **Figure 10** points to the high amount of the disordered protein-binding regions for the possibly druggable proteins compared to the drug targets, which is in concert with the disordered nature of the druggable proteins. This is in agreement with the literature that shows that disordered regions often facilitate PPIs (Mohan et al., 2006; Vacic et al., 2007; Fuxreiter et al., 2014; Yan et al., 2016; Hu et al., 2017). Finally, we show that the possibly druggable proteins are involved in the metabolic and biosynthesis processes and that they are localized across the cell, without a preference for specific subcellular locations. This is unlike the current drug targets that are located primarily in the membranes.

To sum up, our empirical analysis has led us to formulate several markers that may help with identifying novel druggable human proteins and has produced interesting insights into the cellular functions and subcellular locations of potentially druggable proteins.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

LK conceptualized the study. LK and ML designed the study. SG organized the source databases. SG and XL performed acquisition of data. SG and LK organized and performed statistical analysis. All authors organized, analyzed and interpreted the results. LK and SG wrote the first draft of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version, and provided approval for publication of the content.

FUNDING

This research was supported in part by the Robert J. Mattauch Endowment funds to LK, the National Natural Science Foundation of China (No. 61832019), and the 111 Project (No. B18059).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01075/full#supplementary-material

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi: 10.1093/nar/25.17.3389

Ambadipudi, S., and Zweckstetter, M. (2016). Targeting intrinsically disordered proteins in rational drug discovery. *Expert Opin. Drug Discov.* 11 (1), 65–77. doi: 10.1517/17460441.2016.1107041

Amirkhani, A., Kolahdoozi, M., Wang, C., and Kurgan, L. (2018). Prediction of DNA-binding residues in local segments of protein sequences with Fuzzy Cognitive Maps. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2018.2890261

Babu, M. M. (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc Trans.* 44 (5), 1185–1200. doi: 10.1042/BST20160172

Bakheet, T. M., and Doig, A. J. (2009). Properties and identification of human protein drug targets. *Bioinformatics* 25 (4), 451–457. doi: 10.1093/bioinformatics/btp002

Basu, S., and Bahadur, R. P. (2016). A structural perspective of RNA recognition by intrinsically disordered proteins. *Cell Mol. Life Sci.* 73 (21), 4075–4084. doi: 10.1007/s00018-016-2283-1

Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hurst, L. D., et al. (2006). Stratus not altocumulus: a new view of the yeast protein interaction network. *PloS Biol.* 4, 1720–1731. doi: 10.1371/journal.pbio.0040317

Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *J. Acoust. Soc Am.* 22, 725–730. doi: 10.1121/1.1906679

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1), 235–242. doi: 10.1093/nar/28.1.235

Bonacich, P. (1987). Power and centrality: A family of measures. *Am. J. Sociol.* 92 (5), 1170–1182. doi: 10.1086/228631

Bull, S. C., and Doig, A. J. (2015). Properties of protein drug target classes. *PloS One* 10 (3), e0117955. doi: 10.1371/journal.pone.0117955

Calderone, A., Castagnoli, L., and Cesareni, G. (2013). Mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods* 10 (8), 690. doi: 10.1038/nmeth.2561

Chen, C. Y., and Tou, W. I. (2013). How to design a drug for the disordered proteins? *Drug Discov. Today* 18 (19-20), 910–915. doi: 10.1016/j.drudis.2013.04.008

Cheng, Y., Legall, T., Oldfield, C. J., Mueller, J. P., Van, Y. Y., Romero, P., et al. (2006). Rational drug design *via* intrinsically disordered protein. *Trends Biotechnol.* 24 (10), 435–442. doi: 10.1016/j.tibtech.2006.07.005

Chong, C. R., and Sullivan, D. J. (2007). New uses for old drugs. *Nature* 448 (7154), 645–646. doi: 10.1038/448645a

Christopoulos, A. (2014). Advances in G protein-coupled receptor allostery: from function to structure. *Mol. Pharmacol.* 86 (5), 463–478. doi: 10.1124/mol.114.094342

Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., et al. (2016). CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *J. Mol. Biol.* 428 (4), 709–719. doi: 10.1016/j.jmb.2016.01.029

Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32 (suppl_1), D258–D261. doi: 10.1093/nar/gkh036

Cuchillo, R., and Michel, J. (2012). Mechanisms of small-molecule binding to intrinsically disordered proteins. *Biochem. Soc. Trans.* 40 (5), 1004–1008. doi: 10.1042/BST20120086

Dalkas, G. A., Vlachakis, D., Tsagkrasoulis, D., Kastania, A., and Kossida, S. (2013). State-of-the-art technology in modern computer-aided drug design. *Briefings Bioinform.* 14 (6), 745–752. doi: 10.1093/bib/bbs063

Dolan, P. T., Roth, A. P., Xue, B., Sun, R., Dunker, A. K., Uversky, V. N., et al. (2015). Intrinsic disorder mediates hepatitis C virus core-host cell protein interactions. *Protein Sci.* 24 (2), 221–235. doi: 10.1002/pro.2608

Dosztanyi, Z. (2018). Prediction of protein disorder based on IUPred. *Protein Sci.* 27 (1), 331–340. doi: 10.1002/pro.3334

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21 (16), 3433–3434. doi: 10.1093/bioinformatics/bti541

Dosztányi, Z., Chen, J., Dunker, A. K., Simon, I., and Tompa, P. (2006). Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* 5, 2985–2995. doi: 10.1021/pr060171o

Dosztányi, Z., Mészáros, B., and Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* 25 (20), 2745–2746. doi: 10.1093/bioinformatics/btp518

Dunker, A. K., and Uversky, V. N. (2010). Drugs for 'protein clouds': targeting intrinsically disordered transcription factors. *Curr. Opin. Pharmacol.* 10 (6), 782–788. doi: 10.1016/j.coph.2010.09.005

Dunker, A. K., Babu, M. M., Barbar, E., Blackledge, M., Bondos, S. E., Dosztányi, Z., et al. (2013). What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disord. Proteins* 1 (1), e24157. doi: 10.4161/idp.24157

Dyson, H. J., and Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6 (3), 197–208. doi: 10.1038/nrm1589

Estrada, E., and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E.* 71 (5), 056103. doi: 10.1103/PhysRevE.71.056103

Ezzat, A., Wu, M., Li, X. -L., and Kwoh, C. -K. (2018). Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings Bioinf.* 20 (4), 1337–1357. doi: 10.1093/bib/bby002

Fan, X., Xue, B., Dolan, P. T., Lacount, D. J., Kurgan, L., and Uversky, V. N. (2014). The intrinsic disorder status of the human hepatitis C virus proteome. *Mol. Biosyst.* 10 (6), 1345–1363. doi: 10.1039/C4MB00027G

Faraggi, E., Zhou, Y., and Kloczkowski, A. (2014). Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins: Struct. Funct. Bioinform.* 82 (11), 3170–3176. doi: 10.1002/prot.24682

Feng, Y., Wang, Q., and Wang, T. (2017). Drug Target Protein-Protein Interaction Networks: A Systematic Perspective. *Biomed. Res. Int.* 2017, 1289259. doi: 10.1155/2017/1289259

Finan, C., Gaulton, A., Kruger, F. A., Lumbers, R. T., Shah, T., Engmann, J., et al. (2017). The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* 9 (383), eaag1166. doi: 10.1126/scitranslmed.aag1166

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry* 40, 35. doi: 10.2307/3033543

Fuxreiter, M., Toth-Petroczy, A., Kraut, D. A., Matouschek, A., Lim, R. Y., Xue, B., et al. (2014). Disordered proteinaceous machines. *Chem. Rev.* 114 (13), 6806–6843. doi: 10.1021/cr4007329

Gane, P. J., and Dean, P. M. (2000). Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* 10 (4), 401–404. doi: 10.1016/S0959-440X(00)00105-6

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2016). The ChEMBL database in 2017. *Nucleic Acids Res.* 45 (D1), D945–D954. doi: 10.1093/nar/gkw1074

Gutiérrez-Sacristán, A., Bravo, À., Centeno, E., Sanz, F., Piñero, J., García-García, J., et al. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45 (D1), D833–D839. doi: 10.1093/nar/gkw943

Habchi, J., Tompa, P., Longhi, S., and Uversky, V.N. (2014). Introducing Protein Intrinsic Disorder. *Chem. Rev.* 114 (13), 6561–6588. doi: 10.1021/cr400514h

Hambly, K., Danzer, J., Muskal, S., and Debe, D. A. (2006). Interrogating the druggable genome with structural informatics. *Mol. Divers.* 10 (3), 273–281. doi: 10.1007/s11030-006-9035-3

Han, J. D. J., Berlin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88–93. doi: 10.1038/nature02555

Hao, M., Bryant, S. H., and Wang, Y. (2019). Open-source chemogenomic data-driven algorithms for predicting drug–target interactions. *Brief. Bioinform.* 20 (4), 1465–1474. doi: 10.1093/bib/bby010

Harding, S. D., Sharman, J. L., Faccenda, E., Southan, C., Pawson, A. J., Ireland, S., et al. (2017). The iuphar/BPS Guide to pharmacology in 2018: updates and expansion to encompass the new guide to immunopharmacology. *Nucleic Acids Res.* 46 (D1), D1091–D1106. doi: 10.1093/nar/gkx1121

Haupt, V. J., and Schroeder, M. (2011). Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Briefings Bioinform.* 12 (4), 312–326. doi: 10.1093/bib/bbr011

Hopkins, A. L., and Groom, C. R. (2002). The druggable genome. *Nat. Rev. Drug Discovery* 1 (9), 727–730. doi: 10.1038/nrd892

Howell, M., Green, R., Killeen, A., Wedderburn, L., Picascio, V., Rabionet, A., et al. (2012). Not that rigid midgets and not so flexible giants: on the abundance and roles of intrinsic disorder in short and long proteins. *J. Biol. Syst.* 20 (04), 471–511. doi: 10.1142/S0218339012400086

Hu, Y., and Bajorath, J. (2013). Compound promiscuity: what can we learn from current data? *Drug Discovery Today* 18 (13–14), 644–650. doi: 10.1016/j.drudis.2013.03.002

Hu, G., Wang, K., Groenendyk, J., Barakat, K., Mizianty, M. J., Ruan, J., et al. (2014). Human structural proteome-wide characterization of Cyclosporine A targets. *Bioinformatics* 30 (24), 3561–3566. doi: 10.1093/bioinformatics/btu581

Hu, G., Wu, Z., Wang, K., Uversky, V. N., and Kurgan, L. (2016). Untapped potential of disordered proteins in current druggable human proteome. *Curr. Drug Targets* 17 (10), 1198–1205. doi: 10.2174/1389450116666150722141119

Hu, G., Wu, Z., Uversky, V., and Kurgan, L. (2017). Functional analysis of human hub proteins and their interactors involved in the intrinsic disorder-enriched interactions. *Int. J. Mol. Sci.* 18 (12), 2761. doi: 10.3390/ijms18122761

Hu, G., Wang, K., Song, J., Uversky, V. N., and Kurgan, L. (2018). Taxonomic landscape of the dark proteomes: whole-proteome scale interplay between structural darkness, intrinsic disorder, and crystallization propensity. *Proteomics* 18 (21–22), 1800243. doi: 10.1002/pmic.201800243

Imming, P., Sinning, C., and Meyer, A. (2007). Drugs, their targets and the nature and number of drug targets (vol 5, 2006). *Nat. Rev. Drug Discovery* 6 (2), 126–126, pg 821. doi: 10.1038/nrd2132

Jeong, H., Mason, S. P., Barabási, A. -L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411 (6833), 41–42. doi: 10.1038/35075138

Joosten, R. P., Te Beek, T. A., Krieger, E., Hekkelman, M. L., Hooft, R. W., Schneider, R., et al. (2010). A series of PDB related databases for everyday needs. *Nucleic Acids Res.* 39 (suppl_1), D411–D419. doi: 10.1093/nar/gkq1105

Joshi, P., and Vendruscolo, M. (2015). Druggability of intrinsically disordered proteins. *Adv. Exp. Med. Biol.* 870, 383–400. doi: 10.1007/978-3-319-20164-1_13

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Orig. Res. Biomol.* 22 (12), 2577–2637. doi: 10.1002/bip.360221211

Katuwawala, A., Peng, Z., Yang, J., and Kurgan, L. (2019). Computational Prediction of MoRFs, short disorder-to-order transitioning protein binding regions. *Comput. Struct. Biotechnol. J.* 17, 454–462. doi: 10.1016/j.csbj.2019.03.013

Keller, T. H., Pichota, A., and Yin, Z. (2006). A practical view of 'druggability'. *Curr. Opin. Chem. Biol.* 10 (4), 357–361. doi: 10.1016/j.cbpa.2006.06.014

Kim, B., Jo, J., Han, J., Park, C., and Lee, H. (2017). In silico re-identification of properties of drug target proteins. *BMC Bioinform.* 18 (Suppl 7), 248. doi: 10.1186/s12859-017-1639-3

Kjaergaard, M., and Kragelund, B. B. (2017). Functions of intrinsic disorder in transmembrane proteins. *Cell. Mol. Life Sci.* 74 (17), 3205–3224. doi: 10.1007/s00018-017-2562-5

Kuhn, M., Al Banchaabouchi, M., Campillos, M., Jensen, L. J., Gross, C., Gavin, A. C., et al. (2013). Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.* 9, 663. doi: 10.1038/msb.2013.10

Kulkarni, P., and Uversky, V. N. (2018). Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome. *Proteomics* 18, 21–22. doi: 10.1002/pmic.201800061

Launay, G., Salza, R., Multedo, D., Thierry-Mieg, N., and Ricard-Blum, S. (2015). MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. *Nucleic Acids Res.* 43 (Database issue), D321–D327. doi: 10.1093/nar/gku1091

Launay, G., Salza, R., Multedo, D., Thierry-Mieg, N., and Ricard-Blum, S. (2007). Characterization of the drugged human genome. *Pharmacogenomics* 8 (8), 1063–1073. doi: 10.2217/14622416.8.8.1063

Lauss, M., Kriegner, A., Vierlinger, K., and Noehammer, C. (2007). Characterization of the drugged human genome. *Pharmacogenomics* 8, 1063–1073.

Le, K. Q., Prabhakar, B. S., Hong, W. J., and Li, L. C. (2015). Alternative splicing as a biomarker and potential target for drug discovery. *Acta Pharmacol. Sin.* 36 (10), 1212–1218. doi: 10.1038/aps.2015.43

Li, F., Yu, G., Wang, S., Bo, X., Wu, Y., and Qin, Y. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26 (7), 976–978. doi: 10.1093/bioinformatics/btq064

Li, M., Wang, J., Chen, X., Wang, H., and Pan, Y. (2011). A local average connectivity-based method for identifying essential proteins from the network level. *Comput. Biol. Chem.* 35 (3), 143–150. doi: 10.1016/j.compbiolchem.2011.04.002

Li, J., Zheng, S., Chen, B., Butte, A. J., Swamidass, S. J., and Lu, Z. (2016). A survey of current trends in computational drug repositioning. *Brief. Bioinform.* 17 (1), 2–12. doi: 10.1093/bib/bbv020

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., et al. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 40 (Database issue), D857–D861. doi: 10.1093/nar/gkr930

Lieutaud, P., Ferron, F., Uversky, A. V., Kurgan, L., Uversky, V. N., and Longhi, S. (2016). How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. *Intrinsically Disord. Proteins* 4 (1), e1259708. doi: 10.1080/21690707.2016.1259708

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure* 11 (11), 1453–1459. doi: 10.1016/j.str.2003.10.002

Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006). Intrinsic disorder in transcription factors. *Biochemistry* 45 (22), 6873–6888. doi: 10.1021/bi0602718

Lotfi Shahreza, M., Ghadiri, N., Mousavi, S. R., Varshosaz, J., and Green, J.R. (2017). A review of network-based approaches to drug repositioning. *Briefings Bioinf.* 19 (5), 878–892. doi: 10.1093/bib/bbx017

Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486 (7403), 361–367. doi: 10.1038/nature11159

Lounnas, V., Ritschel, T., Kelder, J., Mcguire, R., Bywater, R. P., and Foloppe, N. (2013). Current progress in structure-based rational drug design marks a new mindset in drug discovery. *Comput. Struct. Biotechnol. J.* 5, e201302011. doi: 10.5936/csbj.201302011

Lundstrom, K. (2006). Structural genomics: the ultimate approach for rational drug design. *Mol. Biotechnol.* 34 (2), 205–212. doi: 10.1385/MB:34:2:205

Lv, W., Xu, Y., Guo, Y., Yu, Z., Feng, G., Liu, P., Luan, M., et al. (2016). The drug target genes show higher evolutionary conservation than non-target genes. *Oncotarget* 7 (4), 4961–4971. doi: 10.18632/oncotarget.6755

Ma, D. L., Chan, D. S., and Leung, C. H. (2013). Drug repositioning by structure-based virtual screening. *Chem. Soc. Rev.* 42 (5), 2130–2141. doi: 10.1039/c2cs35357a

Makley, L. N., and Gestwicki, J. E. (2013). Expanding the number of 'druggable' targets: non-enzymes and protein-protein interactions. *Chem. Biol. Drug Des.* 81 (1), 22–32. doi: 10.1111/cbdd.12066

Mavromoustakos, T., Durdagi, S., Koukoulitsa, C., Simcic, M., Papadopoulos, M. G., Hodoscek, M., et al. (2011). Strategies in the rational drug design. *Curr. Med. Chem.* 18 (17), 2517–2530. doi: 10.2174/092986711795933731

Meng, F., Murray, G. F., Kurgan, L., and Donahue, H. J. (2018). High-throughput prediction of disordered moonlighting regions in protein sequences. *Proteins* 86 (10), 1097–1110. doi: 10.1002/prot.25590

Meng, F., Na, I., Kurgan, L., and Uversky, V. (2015b). Compartmentalization and functionality of nuclear disorder: intrinsic disorder and protein-protein interactions in intra-nuclear compartments. *Int. J. Mol. Sci.* 17 (1), 24. doi: 10.3390/ijms17010024

Meng, F., Badierah, R.A., Almehdar, H.A., Redwan, E.M., Kurgan, L., and Uversky, V.N. (2015a). Unstructural biology of the Dengue virus proteins. *FEBS J.* 282 (17), 3368–3394. doi: 10.1111/febs.13349

Meng, F., Uversky, V., and Kurgan, L. (2017a). Computational prediction of intrinsic disorder in proteins. *Curr. Protoc. Protein Sci.* 88, 2 16 1–2 16 14. doi: 10.1002/cpps.28

Meng, F., Uversky, V. N., and Kurgan, L. (2017b). Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell Mol. Life Sci.* 74 (17), 3069–3090. doi: 10.1007/s00018-017-2555-4

Meng, F., Badierah, R. A., Almehdar, H. A., Redwan, E. M., Kurgan, L., and Uversky, V. N. (2018). Functional and structural characterization of osteocytic MLO-Y4 cell proteins encoded by genes differentially expressed in response to mechanical signals in vitro. *Sci. Rep.* 8 (1), 6716. doi: 10.1038/s41598-018-25113-4

Mitsopoulos, C., Schierz, A. C., Workman, P., and Al-Lazikani, B. (2015). Distinctive behaviors of druggable proteins in cellular networks. *PloS Comput. Biol.* 11 (12), e1004597. doi: 10.1371/journal.pcbi.1004597

Mizianty, M. J., Fan, X., Yan, J., Chalmers, E., Woloschuk, C., Joachimiak, A., et al. (2014). Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallogr. D. Biol. Crystallogr.* 70 (Pt 11), 2781–2793. doi: 10.1107/S1399004714019427

Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., et al. (2006). Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* 362 (5), 1043–1059. doi: 10.1016/j.jmb.2006.07.087

Moriaud, F., Richard, S. B., Adcock, S. A., Chanas-Martin, L., Surgand, J. S., Ben Jelloul, M., et al. (2011). Identify drug repurposing candidates by mining the protein data bank. *Brief Bioinform.* 12 (4), 336–340. doi: 10.1093/bib/bbr017

Muruganujan, A., Ebert, D., Mi, H., Thomas, P. D., and Huang, X. (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* 47 (D1), D419–D426. doi: 10.1093/nar/gky1038

Na, I., Meng, F., Kurgan, L., and Uversky, V. N. (2016). Autophagy-related intrinsically disordered proteins in intra-nuclear compartments. *Mol. Biosyst.* 12 (9), 2798–2817. doi: 10.1039/C6MB00069J

Núñez, S., Venhorst, J., and Kruse, C. G. (2012). Target–drug interactions: first principles and their application to drug discovery. *Drug Discovery Today* 17 (1), 10–22. doi: 10.1016/j.drudis.2011.06.013

Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., et al. (2013). D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res.* 41 (Database issue), D508–D516. doi: 10.1093/nar/gks1226

Oldfield, C. J., Ulrich, E.L., Cheng, Y., Dunker, A. K., and Markley, J. L. (2005). Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins* 59 (3), 444–453. doi: 10.1002/prot.20446

Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N., and Dunker, A. K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* 9 Suppl 1, S1. doi: 10.1186/1471-2164-9-S1-S1

Oprea, T. I., and Mestres, J. (2012). Drug repurposing: far beyond new targets for old drugs. *AAPS J.* 14 (4), 759–763. doi: 10.1208/s12248-012-9390-1

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42 (Database issue), D358–D363. doi: 10.1093/nar/gkt1115

Oughtred, R., Stark, C., Breitkreutz, B. J., Rust, J., Boucher, L., Chang, C., et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47 (D1), D529–D541. doi: 10.1093/nar/gky1079

Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). Opinion - How many drug targets are there? *Nat. Rev. Drug Discovery* 5 (12), 993–996. doi: 10.1038/nrd2199

Ozdemir, E. S., Halakou, F., Nussinov, R., Gursoy, A., and Keskin, O. (2019). Methods for Discovering and Targeting Druggable Protein-Protein Interfaces and Their Application to Repurposing. *Methods Mol. Biol.* 1903, 1–21. doi: 10.1007/978-1-4939-8955-3_1

Patil, A., Kinoshita, K., and Nakamura, H. (2010). Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. *Protein Sci.* 19 (8), 1461–1468. doi: 10.1002/pro.425

Peng, Z. L., and Kurgan, L. (2012). Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci.* 13 (1), 6–18. doi: 10.2174/138920312799277938

Peng, Z., Mizianty, M. J., Xue, B., Kurgan, L., and Uversky, V. N. (2012). More than just tails: intrinsic disorder in histone proteins. *Mol. Biosyst.* 8 (7), 1886–1901. doi: 10.1039/c2mb25102g

Peng, Z., Xue, B., Kurgan, L., and Uversky, V. N. (2013). Resilience of death: intrinsic disorder in proteins involved in the programmed cell death. *Cell Death Differ.* 20 (9), 1257–1267. doi: 10.1038/cdd.2013.65

Peng, Z., Oldfield, C. J., Xue, B., Mizianty, M. J., Dunker, A. K., Kurgan, L., et al. (2014b). A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. *Cell Mol. Life Sci.* 71 (8), 1477–1504. doi: 10.1007/s00018-013-1446-6

Peng, Z., Mizianty, M. J., and Kurgan, L. (2014a). Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins* 82 (1), 145–158. doi: 10.1002/prot.24348

Peng, Z., et al. (2015). Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol. Life Sci.* 72 (1), 137–151. doi: 10.1007/s00018-014-1661-9

Pentony, M. M., and Jones, D. T. (2010). Modularity of intrinsic disorder in the human proteome. *Proteins* 78 (1), 212–221. doi: 10.1002/prot.22504

Pielak, R. M., Schnell, J. R., and Chou, J. J. (2009). Mechanism of drug inhibition and drug resistance of influenza A M2 channel. *Proc. Natl. Acad. Sci. U.S.A.* 106 (18), 7379–7384. doi: 10.1073/pnas.0902548106

Radusky, L., Defelipe, L.A., Lanzarotti, E., Luque, J., Barril, X., Marti, et al. (2014). TuberQ: a Mycobacterium tuberculosis protein druggability database. *Database-the Journal of Biological Databases and Curation*. doi: 10.1093/database/bau035

Rajendran, L., Knolker, H. J., and Simons, K. (2010). Subcellular targeting strategies for drug design and delivery. *Nat. Rev. Drug Discov.* 9 (1), 29–42. doi: 10.1038/nrd2897

Rask-Andersen, M., Masuram, S., and Schioth, H. B. (2014). The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annu. Rev. Pharmacol. Toxicol.* 54, 9–26. doi: 10.1146/annurev-pharmtox-011613-135943

Roider, H. G., Pavlova, N., Kirov, I., Slavov, S., Slavov, T., Uzunov, Z., et al. (2014). Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinform.* 15 (1), 68. doi: 10.1186/1471-2105-15-68

Russ, A. P., and Lampel, S. (2005). The druggable genome: an update. *Drug Discovery Today* 10 (23–24), 1607–1610. doi: 10.1016/S1359-6446(05)03666-4

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 32 (Database issue), D449–D451. doi: 10.1093/nar/gkh086

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., et al. (2017). A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discovery* 16 (1), 19–34. doi: 10.1038/nrd.2016.230

Schneider, G. (2010). Virtual screening: an endless staircase? *Nat. Rev. Drug Discovery* 9 (4), 273–276. doi: 10.1038/nrd3139

Sheridan, R. P., Maiorov, V. N., Holloway, M. K., Cornell, W. D., and Gao, Y. D. (2010). Drug-like density: a method of quantifying the "Bindability" of a protein target based on a very large set of pockets and drug-like ligands from the protein data bank. *J. Chem. Inf. Model.* 50 (11), 2029–2040. doi: 10.1021/ci100312t

Siegfried, Z., and Karni, R. (2018). The role of alternative splicing in cancer drug resistance. *Curr. Opin. Genet. Dev.* 48, 16–21. doi: 10.1016/j.gde.2017.10.001

Srivastava, A., Ahmad, S., and Gromiha, M. M. (2018). Deciphering RNA-recognition patterns of intrinsically disordered proteins. *Int. J. Mol. Sci.* 19 (6), 1595. doi: 10.3390/ijms19061595

Stephenson, K., and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Soc Networks* 11 (1), 1–37. doi: 10.1016/0378-8733(89)90016-6

Tan, Q., Zhu, Y., Li, J., Chen, Z., Han, G. W., Kufareva, I., et al. (2013). Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* 341 (6152), 1387–1390. doi: 10.1126/science.1241475

Tantos, A., Kalmar, L., and Tompa, P. (2015). The role of structural disorder in cell cycle regulation, related clinical proteomics, disease development and drug targeting. *Expert Rev. Proteomics* 12 (3), 221–233. doi: 10.1586/14789450.2015.1042866

Tarcsay, Á., and Keserű, G. M. (2013). Contributions of Molecular Properties to Drug Promiscuity. *J. Med. Chem.* 56 (5), 1789–1795. doi: 10.1021/jm301514n

Tompa, P., Fuxreiter, M., Oldfield, C. J., Simon, I., Dunker, A. K., and Uversky, V. N. (2009). Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays* 31 (3), 328–335. doi: 10.1002/bies.200800151

Tseng, C. Y., and Tuszynski, J. (2015). A unified approach to computational drug discovery. *Drug Discovery Today* 20 (11), 1328–1336. doi: 10.1016/j.drudis.2015.07.004

UniProt: the universal protein knowledgebase (2016). *Nucleic Acids Res.* 45 (D1), D158–D169. doi: 10.1093/nar/gkw1099

Uversky, V. N. (2012). Intrinsically disordered proteins and novel strategies for drug discovery. *Expert Opin. Drug Discovery* 7 (6), 475–488. doi: 10.1517/17460441.2012.686489

Uversky, V. N. (2014a). Introduction to intrinsically disordered proteins (IDPs). *Chem. Rev.* 114 (13), 6557–6560. doi: 10.1021/cr500288y

Uversky, V. N. (2014b). The triple power of D(3): protein intrinsic disorder in degenerative diseases. *Front. Biosci. (Landmark Ed.)* 19, 181–258. doi: 10.2741/4204

Uversky, V. N. (2017). Intrinsically disordered proteins in overcrowded milieu: membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.* 44, 18–30. doi: 10.1016/j.sbi.2016.10.015

Uversky, V. N. (2018). Intrinsic disorder, protein-protein interactions, and disease. *Adv. Protein Chem. Struct. Biol.* 110, 85–121. doi: 10.1016/bs.apcsb.2017.06.005

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* 18 (5), 343–384. doi: 10.1002/jmr.747

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* 37, 215–246. doi: 10.1146/annurev.biophys.37.032807.125924

Uversky, A. V., Xue, B., Peng, Z., Kurgan, L., and Uversky, V. N. (2013). On the intrinsic disorder status of the major players in programmed cell death pathways. *F1000Res* 2, 190. doi: 10.12688/f1000research.2-190.v1

Uversky, V. N., Dave, V., Iakoucheva, L. M., Malaney, P., Metallo, S. J., Pathak, R. R., et al. (2014). Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. *Chem. Rev.* 114 (13), 6844–6879. doi: 10.1021/cr400713r

Vacic, V., Oldfield, C. J., Mohan, A., Radivojac, P., Cortese, M. S., Uversky, V. N., et al. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.* 6 (6), 2351–2366. doi: 10.1021/pr0701411

Varadi, M., Zsolyomi, F., Guharoy, M., and Tompa, P. (2015). Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PloS One* 10 (10), e0139731. doi: 10.1371/journal.pone.0139731

Velankar, S., Dana, J. M., Jacobsen, J., Van Ginkel, G., Gane, P. J., Luo, J., et al. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* 41 (Database issue), D483–D489. doi: 10.1093/nar/gks1258

Vuzman, D., and Levy, Y. (2012). Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Mol. Biosyst.* 8 (1), 47–57. doi: 10.1039/C1MB05273J

Walsh, I., Giollo, M., Di Domenico, T., Ferrari, C., Zimmermann, O., and Tosatto, S. C. (2015). Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* 31 (2), 201–208. doi: 10.1093/bioinformatics/btu625

Wang, C., and Kurgan, L. (2018). Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome. *Brief Bioinform*. doi: 10.1093/bib/bby069

Wang, C., and Kurgan, L. (2019). Survey of similarity-based prediction of drug-protein interactions. *Curr. Med. Chem.* 25, 1–1. doi: 10.2174/0929867325666181101115314

Wang, K., and Samudrala, R. (2006). Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinform.* 7, 385. doi: 10.1186/1471-2105-7-385

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C. -F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23 (10), 1274–1281. doi: 10.1093/bioinformatics/btm087

Wang, J., Li, M., Wang, H., and Pan, Y. (2012a). Identification of essential proteins based on edge clustering coefficient. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 9 (4), 1070–1080. doi: 10.1109/TCBB.2011.147

Wang, J., Li, Z. -X., Qiu, C.-X., Wang, D., and Cui, Q. -H. (2012b). The relationship between rational drug design and drug side effects. *Brief. Bioinform.* 13 (3), 377–382. doi: 10.1093/bib/bbr061

Wang, J., Peng, W., and Wu, F. X. (2013a). Computational approaches to predicting essential proteins: a survey. *PROTEOMICS–Clin. Appl.* 7 (1-2), 181–192. doi: 10.1002/prca.201200068

Wang, X., Wang, R., Zhang, Y., and Zhang, H. (2013b). Evolutionary survey of druggable protein targets with respect to their subcellular localizations. *Genome Biol. Evol.* 5 (7), 1291–1297. doi: 10.1093/gbe/evt092

Wang, C., Hu, G., Wang, K., Brylinski, M., Xie, L., and Kurgan, L. (2016a). PDID: database of molecular-level putative protein-drug interactions in the structural human proteome. *Bioinformatics* 32, 579–586. doi: 10.1093/bioinformatics/btv597

Wang, C., Uversky, V. N., and Kurgan, L. (2016b). Disordered nucleiome: abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics* 16 (10), 1486–1498. doi: 10.1002/pmic.201500177

Wang, C., Brylinski, M., and Kurgan, L. (2019). "PDID: Database of Experimental and Putative Drug Targets in Human Proteome," in In Silico Drug Design, ed. K. (London, United Kingdom: Academic Press), 827–847. doi: 10.1016/B978-0-12-816125-8.00028-6

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2017). DrugBank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.* 46 (D1), D1074–D1082. doi: 10.1093/nar/gkx1037

Wojcik, S., Birol, M., Rhoades, E., Miranker, A. D., and Levine, Z. A. (2018). Targeting the intrinsically disordered proteome using small-molecule ligands. *Methods Enzymol.* 611, 703–734. doi: 10.1016/bs.mie.2018.09.036

Wu, Z., Hu, G., Yang, J., Peng, Z., Uversky, V. N., and Kurgan, L. (2015). In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces. *FEBS Lett.* 589 (19 Pt A), 2561–2569. doi: 10.1016/j.febslet.2015.08.014

Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Uversky, V. N., et al. (2007). Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* 6 (5), 1882–1898. doi: 10.1021/pr060392u

Xue, B., and Uversky, V. N. (2014). Intrinsic disorder in proteins involved in the innate antiviral immunity: another flexible side of a molecular arms race. *J. Mol. Biol.* 426 (6), 1322–1350. doi: 10.1016/j.jmb.2013.10.030

Xue, B., Mizianty, M. J., Kurgan, L., and Uversky, V. N. (2012). Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cell Mol. Life Sci.* 69 (8), 1211–1259. doi: 10.1007/s00018-011-0859-3

Yan, J., Dunker, A. K., Uversky, V. N., and Kurgan, L. (2016). Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.* 12 (3), 697–710. doi: 10.1039/C5MB00640F

Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabasi, A. L., and Vidal, M. (2007). Drug-target network. *Nat. Biotechnol.* 25 (10), 1119–1126. doi: 10.1038/nbt1338

Yu, C., Niu, X., Jin, F., Liu, Z., Jin, C., and Lai, L. (2016). Structure-based inhibitor design for the intrinsically disordered protein c-Myc. *Sci. Rep.* 6, 22298. doi: 10.1038/srep22298

Zhang, J., Ma, Z., and Kurgan, L. (2017). Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief. Bioinform.* p, 1–19. doi: 10.1093/bib/bbx168

Zhao, S. (2019). Alternative splicing, RNA-seq and drug discovery. *Drug Discov. Today* 24, 1258–1267. doi: 10.1016/j.drudis.2019.03.030

Zheng, C. J., Han, L. Y., Yap, C. W., Ji, Z. L., Cao, Z. W., and Chen, Y. Z. (2006). Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol. Rev.* 58 (2), 259–279. doi: 10.1124/pr.58.2.4

Zhu, F., Han, B., Kumar, P., Liu, X., Ma, X., Wei, X., et al. (2009a). Update of TTD: therapeutic target database. *Nucleic Acids Res.* 38 (suppl_1), D787–D791. doi: 10.1093/nar/gkp1014

Zhu, M., Gao, L., Li, X., and Liu, Z. (2009b). Identifying drug-target proteins based on network features. *Sci. China C. Life Sci.* 52 (4), 398–404. doi: 10.1007/s11427-009-0055-y

Zhu, M., Gao, L., Li, X., Liu, Z. C., Xu, C., Yan, Y. Q., et al. (2009c). The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network. *J. Drug Targeting* 17 (7), 524–532. doi: 10.1080/10611860903046610

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 March 2020
doi: 10.3389/fgene.2020.00179

In silico Metabolic Pathway Analysis Identifying Target Against Leishmaniasis – A Kinetic Modeling Approach

*Nikita Bora and Anupam Nath Jha**

Computational Biophysics Laboratory, Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, India

The protozoan *Leishmania donovani*, from trypanosomatids family is a deadly human pathogen responsible for causing Visceral Leishmaniasis. Unavailability of proper treatment in the developing countries has served as a major threat to the people. The absence of vaccines has made treatment possibilities to rely solely over chemotherapy. Also, reduced drug efficacy due to emerging resistant strains magnifies the threat. Despite years of formulations for an effective drug therapy, complexity of the disease is also unfortunately increasing. Absence of potential drug targets has worsened the scenario. Therefore exploring new therapeutic approach is a priority for the scientific community to combat the disease. One of the most reliable ways to alter the adversities of the infection is finding new biological targets for designing potential drugs. An era of computational biology allows identifying targets, assisting experimental studies. It includes sorting the parasite's metabolic pathways that pins out proteins essential for its survival. We have directed our study towards a computational methodology for determining targets against *L. donovani* from the "purine salvage" pathway. This is a mainstay pathway towards the maintenance of purine amounts in the parasitic pool of nutrients proving to be mandatory for its survival. This study represents an integration of metabolic pathway and Protein-Protein Interactions analysis. It consists of incorporating the available experimental data to the theoretical methods with a prospective to develop a kinetic model of Purine salvage pathway. Simulation data revealed the time course mechanism of the enzymes involved in the synthesis of the metabolites. Modeling of the metabolic pathway helped in marking of crucial enzymes. Additionally, the PPI analysis of the pathway assisted in building a static interaction network for the proteins. Topological analysis of the PPI network through centrality measures (MCC and Closeness) detected targets found common with Dynamic Modeling. Therefore our analysis reveals the enzymes ADSL (Adenylosuccinate lyase) and IMPDH (Inosine-5′-monophosphate dehydrogenase) to be important having a central role in the modeled network based on PPI and kinetic modeling techniques. Further the available three

OPEN ACCESS

Edited by:
Shandar Ahmad,
Jawaharlal Nehru University, India

Reviewed by:
Junjie Yue,
Biotechnology Research Institute (CAAS), China
Vahab Ali,
Rajendra Memorial Research Institute of Medical Sciences, India

***Correspondence:**
Anupam Nath Jha
anjha@tezu.ernet.in

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *27 June 2019*
Accepted: *14 February 2020*
Published: *06 March 2020*

Citation:
Bora N and Jha AN (2020) In silico Metabolic Pathway Analysis Identifying Target Against Leishmaniasis – A Kinetic Modeling Approach. Front. Genet. 11:179. doi: 10.3389/fgene.2020.00179

dimensional structure of the enzyme "ADSL" aided towards the search for potential inhibitors against the protein. Hence, the study presented the significance of integrating methods to identify key proteins which might be putative targets against the treatment of Visceral Leishmaniasis and their potential inhibitors.

Keywords: *Leishmania donovani*, Visceral Leishmaniasis, kinetic modeling, purine salvage, protein protein interaction

INTRODUCTION

One of the primary problems arising worldwide is the increasing risk of parasitic disease mostly infecting the people and animals in the underdeveloped countries (El Kouni, 2003). It includes infections from a wide source of microorganisms like fungi, bacteria, and protozoans etc. (Wang, 1984; Kokina et al., 2019). One such parasitic disease, Visceral leishmaniasis (VL or Kala Azar), caused by the protozoan species *Leishmania donovani* has served as a major threat to these countries, increasing the rate of fatality (Desjeux, 2004; Alvar et al., 2012). VL serves to be one of the most severe forms of leishmaniasis (Sundar, 2001) with the highest death rate (Cavalli and Bolognesi, 2009). This species can infect the internal organs threatening the human health (Sharma et al., 2017). Mostly affected are the poor people from the East African and the Indian subcontinent hence leading to a higher demand for the identification, treatment and control of the infection in the low and middle income countries (Chappuis et al., 2007).

Treatment of *Leishmania* infection relies on chemotherapy (Sundar and Chatterjee, 2006), however, failure towards the available chemotherapeutic agents and treatments still prevails. The first drugs for the treatment were made available around five decades ago. However, the formulation of a single drug is not sufficient to combat the species due to the differences in drug sensitivity among the *Leishmania* sp. (Croft and Coombs, 2003). The substantial side effects (Vijayakumar and Das, 2018) and difficulty in administration has also led to the evolution of drug resistant parasitic strains contributing to the increased rate of mortality (Sundar, 2001). Further, expensive treatment strategies acts as a hurdle towards an effective drug development (Bora, 1999; Croft et al., 2005). In a conclusive manner, a major challenge still exists in identifying effective cure and treatment for the parasite disease (Freitas-Junior et al., 2012) which requires an exploitation of current technologies for identifying novel chemotherapeutics (Davis et al., 2004). The mandate is to find novel drug-targets from the parasite's proteome (Guerin et al., 2002). The identification of such targets from a pathogen's biological pathway is reported to be an important feature in the drug discovery process (Chawla and Madhubala, 2010). This has helped in exploring ways for studying the protozoan's metabolic pathways to sort out the ones unique to them (Martin et al., 2016). The information embedded in the microbe's life cycle may pave the way for understanding the pathogenesis (Smith and Romesberg, 2007). It proves to be essential in controlling the microbial infections that are becoming resistant to the drugs available for their treatment causing fatal conditions.

One of the key factors in understanding the pathogen's biological pathway requires knowledge of the underlying kinetics governing the enzymes and molecules involved in the pathway. This complex biological system can be represented into a network of interconnecting links signifying the reactions involved in the pathway (Meshram et al., 2019). *In silico* analysis of metabolic pathways through systems biology approach has been on the forefront for providing the means to understand the whole network through the availability of the experimental data. Hence, availability of experimental details paves a way for describing the pathway mathematically (Van Riel, 2006). System level analysis has been used as a tool for identifying targets against different species of *Leishmania* (Chavali et al., 2008; Mandlik et al., 2012; Sharma et al., 2017). However, providing a detailed mathematical model is still a challenge in systems biology (Steuer et al., 2006). Biological databases (Stein, 2003), provide a means to attain the knowledge of biological reactions involved in the pathways. Also with the growth of high throughput technologies (Baker and Brass, 1998), the size of these databases are increasing making the interpretation of data a major challenge in the scientific field (Guimera and Amaral, 2005).

Life cycle of *L. donovani* exists as flagellated extracellular promastigotes in the phlebotomine sandfly vector and as immotile intracellular amastigotes within cells of the infected mammalian host. The amastigote has been reported to be the cause behind the pathogen infections including Visceral Leishmaniasis (McConville et al., 2007). *L. donovani* lacks the machinery for the synthesis of purines competing with the hosts for salvaging the required purines (Boitz et al., 2012). Incapability of these organisms to synthesize the purines *de novo* has led to their dependence on the salvaging of the purines from their hosts rendering the "purine salvage" pathway essential for its survival (Boitz and Ullman, 2013; Ansari et al., 2016). Hence, the pathway has served has a backbone towards the maintenance of purine amounts in the parasitic pool of nutrients (Marr et al., 1978; Looker et al., 1983; De Koning et al., 2005). A lack of effective cure for the disease has made its importance in the scientific community opening up ways for delineation of this pathway (Boitz et al., 2012). The enzymes involved in this survival pathway opens up the exploration space for newer drug targets (Carter et al., 2008) for chemotherapeutic agents against parasitic diseases (Berg et al., 2010; Doleželová et al., 2018).

In the current work, we have carried out a twin approach – both dynamical and static, for identifying potential drug targets against the pathogen *L. donovani*. Unlike the traditional ways of identifying a target, we have carried out the *in silico* simulation of the "purine salvage" pathway for the protozoan which represents the dynamic method. The selected pathway has been represented

in the form of a mathematical model defining the reactions catalyzed by the enzymes and the pathway components. Kinetic modeling of the pathway displayed the importance of the proteins Adenylosuccinate lyase (ADSL) and Inosine-5′-monophosphate dehydrogenase (IMPDH). The static method comprises of a Protein Protein Interaction (PPI) network for the proteins involved in the purine salvage pathway. PPI network analysis has been used as a tool for studying the proteomes of *Leishmania* species like *L. braziliensis*, *L. major* and *L. infantum* (Flórez et al., 2010; Rezende et al., 2012; Chávez-Fumagalli et al., 2018; Dos Santos Vasconcelos et al., 2018). The PPI networks are types of biological network representing a set of proteins and their interactions. It was carried out to analyze the importance of the sorted proteins from the dynamical model. A topological analysis of the PPI network showed that these proteins were also found to be ranked in the top central nodes (proteins) with higher connectivity. These nodes (hub proteins) have a higher number of connections that designates its physical associations with their other proteins responsible for performing the specialized functions. Hence a destruction of these nodes will lead to a loss of the connectivity and function which might be crucial for the parasite's survival.

MATERIALS AND METHODS

A dual approach has been applied in our current work where a dynamic modeling and a static interaction network was generated. The dynamic modeling approach includes the selection of a pathway essential for the parasite survival, retrieving the enzymatic reactions and construction of a kinetic model reflecting the importance of certain enzymes which might be important targets in controlling *Leishmaniasis.* The static interaction network includes construction of a PPI and computing the hub nodes.

Selection of Metabolic Pathway

A series of reactions are indulged within the parasite which effect the host (Lambris et al., 2008). These reactions are regulated by enzymes for the generation of the desired products. Some metabolic pathways of the pathogen are proven to be essential for the survival of the organism inside the host. Analysis of those reactions is crucial in listing out the enzymes which could be potential drug targets. One such pathway considered to be mandatory for the survival of *L. donovani* is the "purine salvage" pathway. The enzymes and reactions involved in the pathway are compiled with the help of the available resources.

Biochemical Network Construction

Reactions of the Biological Pathway

The enzymes and the reactions involved in the salvaging of purines in *L. donovani* are obtained from the biological pathway resource "KEGG" (Kanehisa and Goto, 2000). The conversions of the metabolites were further cross checked with the current literature. For building of our model of interest, we have considered the reactions only occurring at the amastigote stage, i.e., the infection causing stage of the parasite (McConville et al., 2007). The metabolic pathway is then graphically represented through the pathway editor "Cell Designer" (Funahashi et al., 2003). It represents every reaction along with the substrates and the metabolites formed with a simpler view. The species transformation is represented through reactions (whether reversible or irreversible).

Kinetic Model

A kinetic rate law defines every reaction included in the model. Enzymatic mechanisms were studied to specify the rate equations for the reactions governed by the enzyme. For enzymes whose detailed mechanisms were not known, the rate equations were constrained to the basic Michaelis-Menten equation, with the basic knowledge that all enzymatic reactions follow the Michaelis-Menten equation. The kinetic model of our pathway is set up using the pathway simulator "COPASI" (Hoops et al., 2006). It constitutes the species defined in their biochemical terms. All the rate equations for the enzymes were generated through COPASI. The mathematical model of the metabolic pathway is defined in the form of Ordinary Differential equations (ODE). Defining a kinetic model implies retrieving available kinetic parameters through different sources. These parameters for the enzymes in our model is collected from curated enzyme database "Brenda" (Schomburg et al., 2004). Parameters not available in the database are collected through literature. Parameters unavailable for a few enzymes for *L. donovani* are taken from other reference organisms (Franco and Canela, 1984) assuming that the enzyme mechanisms remains conserved across species. Further integrating a system requires setting up the initial conditions. Considering this, we have set the initial concentrations for the different species in our model.

Model Simulation and Analysis

The kinetic behavior of the model has been assessed through evaluating the species with respect to time.

Steady State Analysis

Initially the steady state analysis was carried out for our model using COPASI. It enables to analyze the state at which the concentrations of the species do not vary with respect to time. The behavior of the system at that state tends to continue to be present in the future. Methods that COPASI uses for the steady state calculations are the damped Newton method and the integration method both forward and backward.

Time Course Analysis

Followed by the steady state calculation we conducted the time course analysis using the "time course utility" of COPASI. The deterministic method (LSODA) is used for simulating the dynamic behavior and carried out for a period of 500 sec.

Sensitivity Analysis

Assessing the effects of the parameters over the system variables is another crucial point in kinetic modeling. This was carried out through a sensitivity analysis of the model using the sensitivity utility of COPASI.

Protein-Protein Interaction Network

Enzymes sorted through our dynamic analysis were further checked for their importance in the pathway through a topological analysis of the Protein-Protein Interaction (PPI) Network. Compared to the kinetic models, the PPI represents the static interactions of a protein with their co-partners. These networks provide as an effective means towards understanding functional interactions, identification of important modules and prioritization of targets. Protein-Protein Interaction network are mathematically created networks where every protein is represented as a node and the interaction between two proteins as an edge. The functional protein interactions for *L. donovani* were retrieved from the available source of interaction database "STRING version 10.5" (Szklarczyk et al., 2014). The corresponding interactions of proteins involved in the purine salvage pathway were retrieved. Using this molecular interaction information, a Protein-Protein interaction (PPI) network has been constructed through the open source platform "Cytoscape" (Shannon et al., 2003) which represents the overall interaction of the selected proteins in the *L. donovani* sp. MCODE (Chin et al., 2014) clustering algorithm was applied to identify the clusters consisting of densely interconnected nodes. Proteins in a cluster tend to functionally link to each other. Topological analysis for the constructed PPI network was carried out to deduce the properties of the network. Central nodes in the network were analyzed through the Cytohubba (Chin et al., 2014) package of "Cytoscape." These nodes show a higher number of connections with the other proteins in the network and hence maintain their associated functions. Deletion of the proteins behaving as the central nodes will lead to a loss of connectivity in the network. Hence, nodes having a higher tendency to alter the topology of the network are identified. They might prove to be important targets for inhibiting the growth and survivability of the pathogen.

Inhibitor Search for the Selected Enzyme

The presence of the three dimensional structure of a protein assists towards the identification of possible inhibitors against the selected enzyme (Choudhary et al., 2019). Inhibitor search for our selected protein was carried out through pharmacophore mapping of the available ligands against the protein. In our study we highlighted two enzymes having a potential of being targets against *L. donovani*. However, the crystal structure of only the ADSL enzyme was available and hence we selected this enzyme for further study. A literature search was conducted to list out the possible inhibitors for the ADSL enzyme. These inhibitors were subjected to a pharmacophore generation through Pharmagist. The pharmacophore was then mapped against the ZINC database through the Zinc Pharmer to select out possible inhibitors for ADSL enzyme. A set of molecules were then selected from the database of molecules which were further considered for analysis. Molecular docking approach was applied to the selected molecules to observe the binding behavior of the ligands within the receptor molecule. Autodock version 4.2.6 was applied to carry out the molecular docking.

RESULTS AND DISCUSSION

Selection of Metabolic Pathway

Purine nucleotides are considered important for maintaining the vital processes of the cell. However, the *Leishmania* parasite being unable to synthesize the purine rings has developed a pathway for scavenging the required purines (Boitz et al., 2012). Literature sources showed various reactions in the incorporation of purines from the host metabolism. Summarizing all the possible transformations of the metabolites, a generalized form of the pathway is constructed including all the reaction products and the enzymes involved in salvaging of the nutrients. Also, the complexity of simulating kinetic models assists the fact that single pathways are more feasible to study. The main strategy while building our model was to reduce the system while still maintaining the overall characteristics of the biological system. Therefore in our current study we have considered a total of 15 reactions with 14 enzymes found to be involved in these reactions generating the desired metabolites. These enzymes are tabulated in **Table 1** along with their reactions retrieved from BRENDA and literature. Several classes of enzymes like lyases, kinases, transferases etc. are found to be involved in the scavenging process which is shown by their EC numbers. The schematic representation of the pathway is created through Cell Designer. Both reversible and irreversible reactions were included in our model. The reactions and conversions playing an important role in the regulation of the metabolites are schematically arranged. These enzymes (**Table 1**) along with the reactions they catalyze are shown in **Figure 1**.

Model Building

The kinetic model of the pathway was generated through the pathway simulation software "COPASI." Pathway species (enzymes, metabolites) were incorporated to generate the biochemical model followed by fixing the defined rate laws against each enzyme. The rate of the enzymes was described by the Michaelis-Menten equations which were able to capture the necessary details of the reactions and hence displaying the dynamic behavior. These rate equations depend on the concentrations of the metabolites and parameters like binding constants, maximum velocity etc. The model was subsequently fed with the collected kinetic parameters setting up their initial concentrations. This yielded a set of ODE's representing the enzyme kinetics of the system. Kinetic modeling solves these set of ODE's for their dynamic behavior. Once the model has been assembled, it is simulated for observing the variations in the system. The publicly available biochemical network simulator "COPASI" was used to simulate the constructed model and also to perform the analysis.

Model Simulation and Analysis

Steady State Analysis

A characteristic feature of metabolic network is to evolve into a steady state frequently. Computational calculation of steady state requires solving the model through solving the system of ODE's numerically. Steady states of a system represent a

TABLE 1 | Enzymes involved in the purine salvage pathway.

Sl. No	Enzymes	EC No.	Reactions
1	Adenosine kinase	2.7.1.20	Adenosine + ATP ->adenosine 5′ monophosphate + ADP
2	Adenine phosphoribosyltransferase	2.4.2.7	AMP + diphosphate = adenine + 5-phospho-alpha-D-ribose 1-diphosphate
3	Adenine aminohydrolase	3.5.4.2	Adenine + $H_2O \rightarrow$ hypoxanthine + NH_3
4	Purine nucleoside phosphorylase	3.2.2.1	Purine nucleoside + phosphate $\rightleftharpoons$ purine + alpha-D-ribose 1-phosphate
5	Adenosine deaminase	3.5.4.4	Adenosine + $H_2O \rightarrow$ inosine + NH_3
6	AMP deaminase	3.5.4.6	AMP $\rightarrow$ IMP
7	Adenylosuccinate Synthase/Succino-AMP synthetase	6.3.4.4	GTP + IMP + L-aspartate = GDP + phosphate + adenylosuccinate
8	Adenylosuccinate lyase/Succino-AMP lyase ADSL	4.3.2.2	Succino-AMP = AMP + fumarate
9	Hypoxanthine-Guanine phosphoribosyltransferase	2.4.2.8	(1) IMP + Diphosphate = Hypoxanthine + 5-phospho-alpha-D-ribose 1-diphosphate (2) GMP + Diphosphate = Guanine + 5-phospho-alpha-D-ribose 1-diphosphate
10	Xanthine phosphoribosyltransferase	2.4.2.22	XMP + Diphosphate = 5-phospho-alpha-D-ribose 1-diphosphate + xanthine
11	GMP reductase	1.7.1.7	GMP + NADPH + (H+) $\rightarrow$ IMP + (NADP+) + NH_3
12	IMPDH	1.1.1.205	IMP + (NAD+) + $H_2O \rightarrow$ XMP + NADH + (H+)
13	GMP synthase	6.3.5.2	ATP + XMP + $NH_3 \rightarrow$ AMP + diphosphate + GMP
14	Guanine deaminase	3.5.4.3	Guanine + $H_2O \rightarrow$ xanthine + NH_3

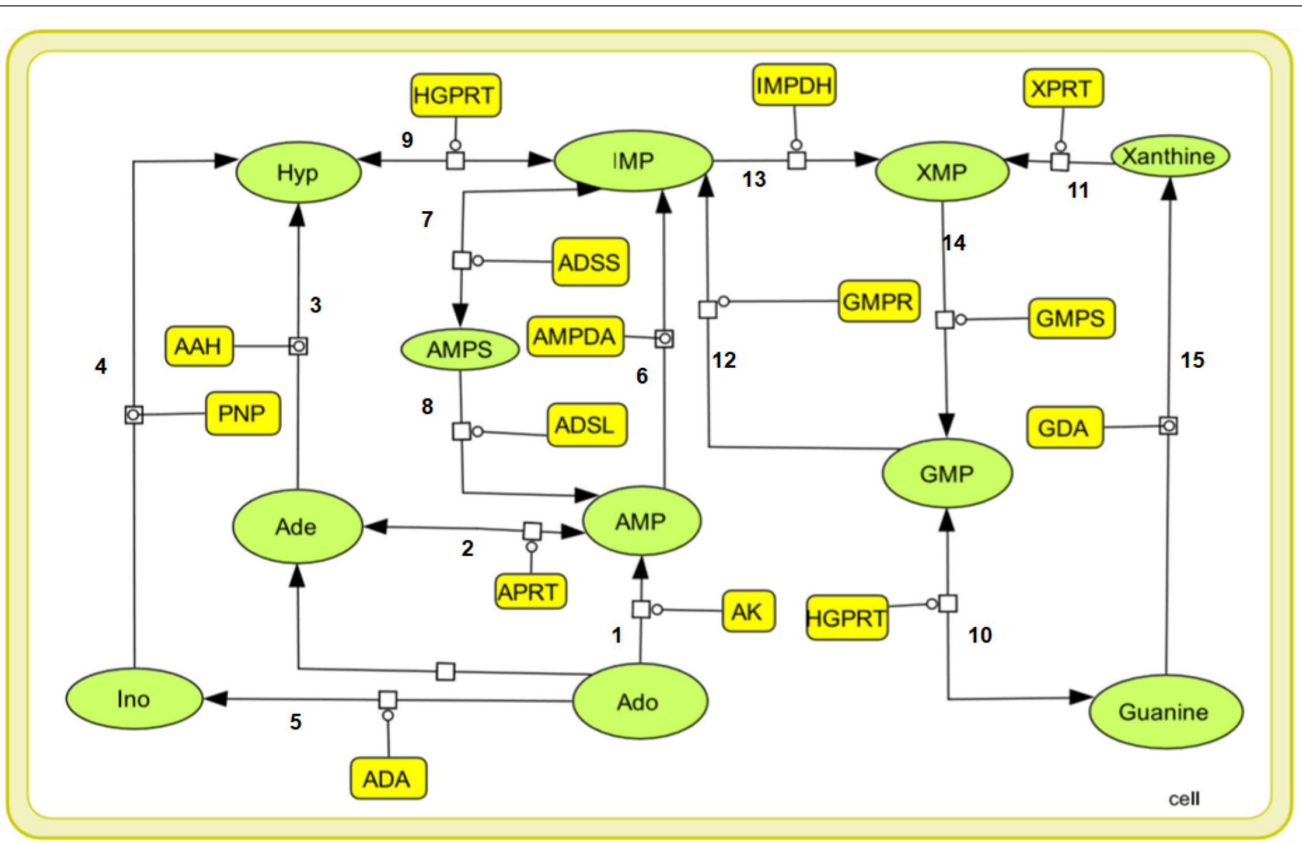

FIGURE 1 | A graphical representation of the purine salvage pathway. The enzymes are shown in yellow rectangular boxes and the oval shapes represents the metabolites: Ade, adenine, Hyp, hypoxanthine, Ino, inosine, Ado, adenosine, AMP, IMP, XMP, GMP, Xanthine, Guanine, and AMPS.

certain physiological condition of the network. Changes in the physiological conditions results in transient states as the system transfers itself to a new steady state. The steady state of our kinetic model was calculated and it showed that the model acquired a steady state with our initial conditions and set of parameters. The steady state fluxes have been shown in **Table 2**.

TABLE 2 | Steady state fluxes of Purine salvage model.

Reactions	Enzymes	Flux (mol/s)
1	Adenosine kinase	−1.00449e-15
2	Adenine phosphoribosyltransferase	4.77811e-15
3	Adenine aminohydrolase	3.22322e-15
4	Purine nucleoside phosphorylase	−4.20221e-14
5	Adenosine deaminase	1.78269e-15
6	AMP deaminase	3.90643e-18
7	Adenylosuccinate Synthase/Succino-AMP synthetase	6.24077e-15
8	Adenylosuccinate lyase/Succino-AMP lyase ADSL	6.0136e-15
9	Hypoxanthine-Guanine	−6.8775e-15
10	phosphoribosyltransferase	−1.68639e-15
11	Xanthine phosphoribosyltransferase	3.02017e-12
12	GMP reductase	−1.80154e-15
13	IMPDH	4.72791e-16
14	GMP synthase	7.62529e-20
15	Guanine deaminase	−2.32796e-15

Time Course Analysis

Time evolution of the model was carried out through the "time course analysis" utility of COPASI. The main objective of performing this analysis is to observe the dynamics of the model. A graph shows the behavior of the model with respect to time (**Figure 2**). The concentration vs. time plot is generally used to infer the attainment of a steady state. A plateau has been observed in the trajectories of the concentrations against time plot where no further changes in the variables are observed. This suggests that the model has reached the steady state characterized by the constant species concentrations. The production and the consumption of the metabolites occur at the same rate.

Sensitivity Analysis

Sensitivity analysis is essentially performed to observe the consequences of the parameters over the model variables like concentration of the species. Two plots were generated showing the effects of the concentration of the species over the variables and the parameters. **Figure 3** displays that the reactions catalyzed by the enzymes IMPDH, ADSL, ADSS, XPRT, GDA, and GMPS were sensitive to the initial concentrations of the marked variables. Whereas **Figure 4** shows that the marked parameters involved in the reactions 1, 2, 4, 8, 9, and 12 were mostly sensitive to the model. These reactions were catalyzed by the enzymes AK, APRT, PNP, ADSL, HGPRT, and the IMPDH, respectively. It is observed from both the cases that the enzymes IMPDH and ADSL are the most sensitive to the effect of concentration. Therefore, the study indicates that these specific reactions probable to be sensitive in the metabolic chain could be further analyzed through future experiments to give valuable insights into the dynamics of the system. Moreover the experimental studies could be directed towards the therapeutic importance of these reactions. Thus, the behavior observed here provides generous amount of dynamics of the "Purine salvage" pathway reflecting the importance of these reactions.

Protein-Protein Interaction Network

To further analyze the importance of the two proteins, IMPDH and ADSL, we carried out a static analysis along with the dynamics study. The available protein-protein interactions for *L. donovani* from STRING consists a total of 3707 proteins and a total of 437053 interactions. The constructed PPI network of the purine salvage pathway showed the participation of 1581 proteins (nodes) and 5735 unique interactions (list of proteins attached at the end of **Supplementary Material**). The interactions showed proteins to be involved in other functions other than the purine salvage. Clustering of the network through MCODE resulted into two clusters as tabulated in **Table 3**. Comparison of both the clusters reveals that lesser number of proteins were involved with a higher number of interactions resulting into a dense network. Proteins in a dense network often forms functional modules that contributes to cellular processes (Rasti and Vogiatzis, 2019).

Therefore knocking out of these proteins should result into a much lesser number of interactions which will lead to an overall loss of a number of functions of the proteins. These interactions may also prove to be fatal. On the other hand, in the second cluster higher number of proteins were involved having lesser number of interactions compared to cluster 1. The proteins which were found to be sensitive in the kinetic modeling approach were analyzed for their presence in the cluster. The enzymes ADSL and IMPDH were found to be having interactions with other proteins in the first cluster. Along with these two proteins, two other proteins were also found to be present in the interaction pattern. These proteins along with their STRING ids have been tabulated in **Supplementary Table S1**. Cluster 1 of the MCODE clustering method has been shown in **Supplementary Figure S1** with the tabulated enzymes being highlighted in the network. The other protein APRT (Adenine phosphoribosyltransferase) with a STRING id of XP_003861601.1 was found to be present in cluster 2 (**Supplementary Figure S2**). Rest of the proteins were not involved in formation of any interactions in these two clusters.

Network analysis enhances our understanding of the interactions of a node with other nodes in the network. As the biological network is heterogeneous, different topological parameters were used to identify the essential proteins. The MCC (Maximum Clique centrality) method of Cytohubba is reported to be a better method in identifying central nodes (Chin et al., 2014). It shows that the proteins ADSL and IMPDH were captured to be the central nodes. Another method to rank the proteins applied was global centrality based method "closeness." It showed that the two proteins IMPDH and ADSL were ranked to be essential.

Literature studies support the fact the enzyme IMPDH is a promising target for drug discovery in antibacterial, anticancer and antiviral treatments (Shu and Nair, 2008) and organism like *Pneumocystis carinii* (O'Gara et al., 1997). Further importance of this enzyme has also been demonstrated in the organism *Leishmania amazonensis* (Pitaluga et al., 2015). ADSL is known

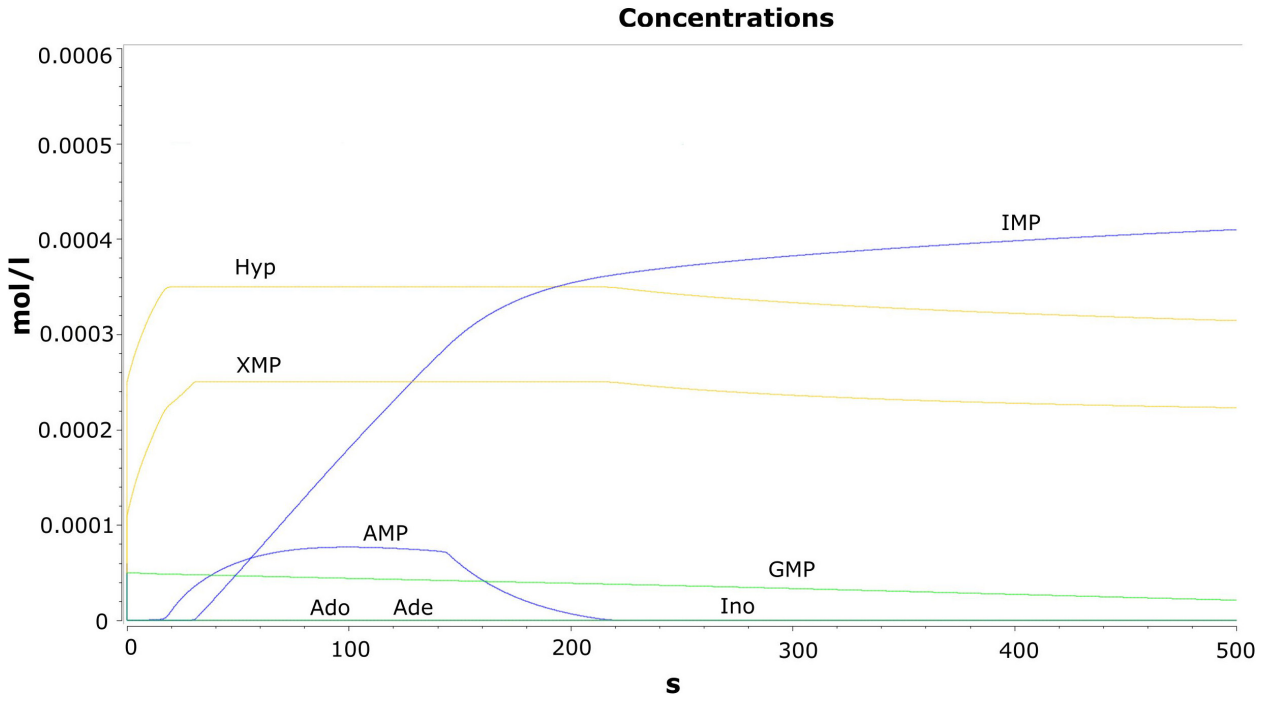

FIGURE 2 | Results of the time course analysis for the metabolites.

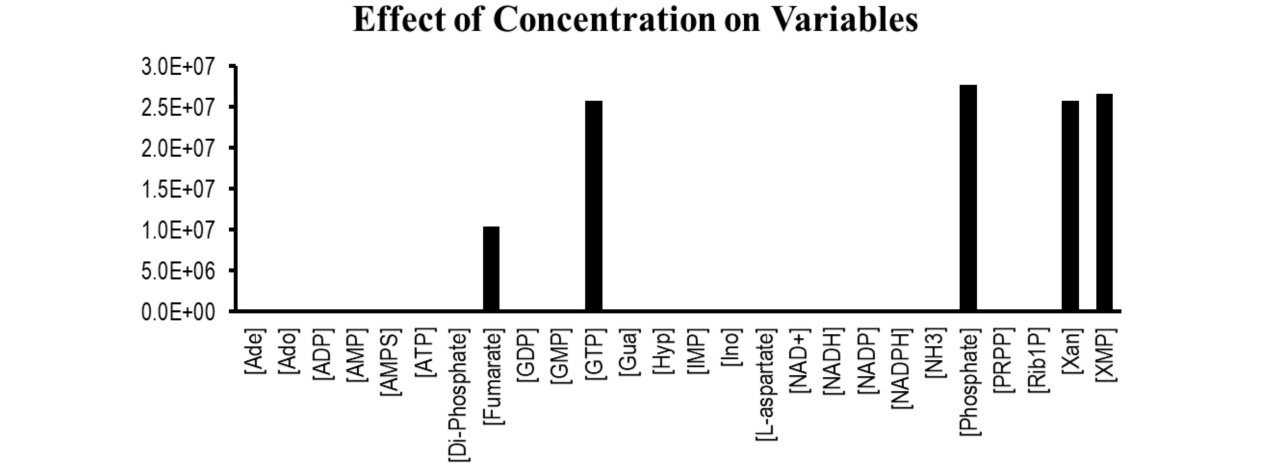

FIGURE 3 | Effect of the species concentration over all the variables in the model.

as an important target in organisms like *Plasmodium falciparum* (malaria) (Bulusu et al., 2009), *Cryptococcus neoformans* (Chitty et al., 2017), *Staphylococcus aureus* (Fyfe et al., 2010) and *Schistosoma mansoni* (schistosomiasis) (Romanello et al., 2017). Experimental groups have also reported the fact that a decrease growth rate occurs for the phenotypes when the ADSL gene is knocked out (Boitz et al., 2013). Therefore, studies could be further conducted upon targeting these proteins of interest which might lead to a loss of *Leishmania* infections (Galina et al., 2017). *In silico* mutational analysis of the ADSL protein suggested a few mutations that bought conformational changes to the catalytic site of the protein (Bora and Nath Jha, 2019).

Inhibitor Search for the Selected Enzyme

Availability of three dimensional structures for the proteins gives an insight into the molecular information of the proteins (Das et al., 2018). Exploring the protein structural repository "PDB", it was found that the crystal structure for only ADSL (PDB id: 4MX2) was available. This led to the exploration of the protein through other computational techniques highlighting it as a probable target. We carried out an analysis to

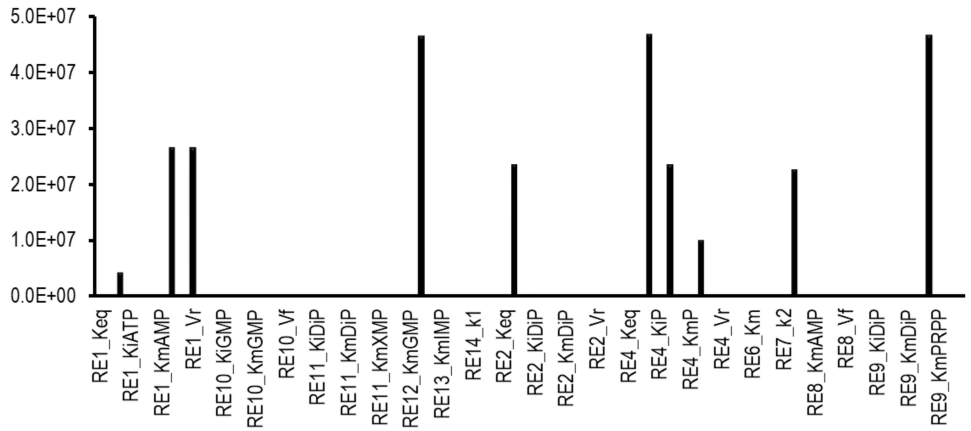

FIGURE 4 | Sensitivity of the parameters involved in the model.

identify potent inhibitors against this target molecule having an available 3D structure.

A total of 14 molecules reported to be possible inhibitors of Adenylosuccinate Lyase has been listed out (shown in **Supplementary Table S2**). Pharmacophore generation of the 14 molecules was carried out to find out the potential regions of the ligands contributing towards the catalytic activity. It was found out that 11 molecules were best aligned in generating the desired pharmacophore. Three chemical features were found out to be involved in maintaining the activity of the ligand molecules which includes two hydrogen bond acceptor regions and one aromatic region. The generated pharmacophore has been shown in **Figure 5**. It was then used as a query to search potential ligands. The reason behind the use of searching the database through a pharmacophore is to identify hits having similar chemical features to that of the query. We mapped the pharmacophore against the Zinc database. Twenty inhibitors with a root mean square deviation (rmsd) of zero were selected out as our dataset for further analysis. The list of molecules is tabulated in **Supplementary Table S3**. The molecules were found to be satisfying Lipinski's rule which designates the druglike property of the molecules. To find out the optimal binding pose of these ligands to the enzyme, the ligand molecules were docked to the active site of the ADSL enzyme. The results of the Docking have been shown in **Supplementary Table S4**.

Molecular docking analysis revealed the binding patterns of the ligand molecules within the ADSL protein. Docking results showed that the binding energy of all the molecules were quite stable but out of 20 molecules, 18 molecules were found to be forming hydrogen bonds with the receptor. The strength of binding usually depends of the interaction between the ligand atoms and the receptor atoms. Therefore the presence

TABLE 3 | MCODE clustering result.

Clusters	No. of nodes	No. of nodes	Score
Cluster 1	27	195	15
Cluster 2	67	143	4.333

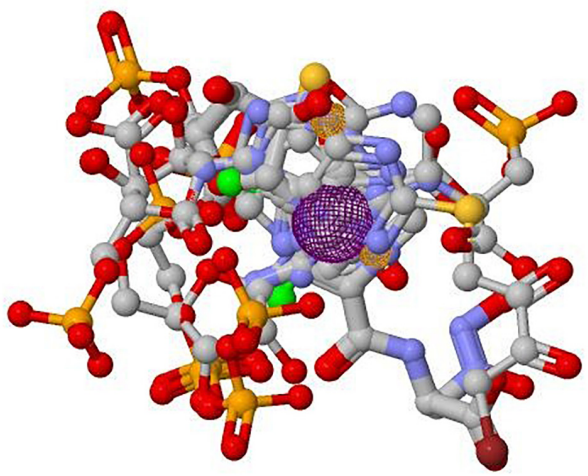

FIGURE 5 | Pharmacophore of the known inhibitors against ADSL.

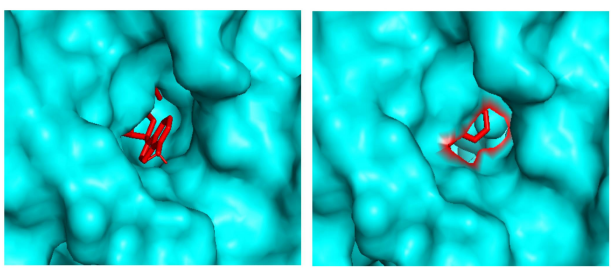

FIGURE 6 | Docking pictures of Molecule 7 and Molecule 13 within the active site of ADSL.

of hydrogen bonds aids toward the stability of the ligand-receptor complexes. Also out of the 18 molecules, 12 molecules were observed to be forming hydrogen bonds with the active site residues. These residues are reported to be involved in the catalysis process and hence the binding of these ligands to the residues showcases their importance in inhibiting the enzyme. Also molecule 7 and molecule 13 has been observed to be interacting with the maximum number of active residues. The orientation of the molecules inside the cavity of the protein has been shown in **Figure 6**. Results showed that the molecules were able to bind deep within the activity that assists them in their catalytic properties. Therefore the effective binding of the molecules to the protein display their therapeutic importance as drug molecules.

CONCLUSION

The *in silico* study reflected that a systems level analysis for a protozoan parasite "*L. donovani*" is possible resulting into the identification of drug targets. In our work, a biochemical network of a metabolic pathway "Purine salvage" for the organism *L. donovani* has been constructed. The dynamic behavior of the model is analyzed through a mathematical representation of the reactions showcasing the biological events. The dynamic simulation allowed us to define the biological pathway of interest with respect to time. Further the model stands as a benchmark for incorporation of complex enzymatic mechanisms through the availability of experimental data. Reactions showing higher sensitivity to the model have been sorted out with the listing of the enzymes regulating those reactions. These enzymes and their reactions could be further experimentally tested out for their importance in therapeutics. Also to analyze the essentiality of these proteins, a static interaction network (the PPI network) for the proteins involved in purine salvage has been constructed. Clustering analysis resulted into a much higher dense network consisting of the proteins ADSL and IMPDH. It was observed through topological analysis that the mentioned proteins were ranked among the top 30% of the hub proteins. These proteins might serve as targets to further explore the pathway mechanism shedding light on the control of infections caused by the pathogens. Availability of the three dimensional structures for ADSL marks its possibility for other computational analysis like Virtual screening of inhibitors. Search for potential inhibitors for the ADSL protein was carried out through the identification of specific chemical features i.e., the pharmacophore. Mapping of the pharmacophore to available ligands lead to the selection of molecules having similar features and within minimum variation from the query molecules. To observe the binding mode of the ligand molecules to that of the receptor, molecular docking analysis was applied. Results of the docking analysis showed that two molecules were able to effectively bind to the active site residues of the protein. Hence it displayed their therapeutic importance as inhibitors having features similar to the known inhibitors against ADSL.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

NB performed the *in silico* studies, analyzed the data, and wrote the manuscript. AJ supervised this study.

ACKNOWLEDGMENTS

We acknowledge Department of Biotechnology (BT/PR15847/NER/95/21/2015), Government of India for providing computational facilities and fellowship to NB.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00179/full#supplementary-material

REFERENCES

Alvar, J., Velez, I. D., Bern, C., Herrero, M., Desjeux, P., Cano, J., et al. (2012). Leishmaniasis worldwide and global estimates of its incidence. *PLoS One* 7:e35671. doi: 10.1371/journal.pone.0035671

Ansari, M. Y., Equbal, A., Dikhit, M. R., Mansuri, R., Rana, S., Ali, V., et al. (2016). Establishment of correlation between in-silico and in-vitro test analysis against *Leishmania* HGPRT to inhibitors. *Int. J. Biol. Macromol.* 83, 78–96. doi: 10.1016/j.ijbiomac.2015.11.051

Baker, P. G., and Brass, A. (1998). Recent developments in biological sequence databases. *Curr. Opin. Biotechnol.* 9, 54–58. doi: 10.1016/s0958-1669(98)80084-0

Berg, M., Van Der Veken, P., Goeminne, A., Haemers, A., and Augustyns, K. (2010). Inhibitors of the purine salvage pathway: a valuable approach for antiprotozoal chemotherapy? *Curr. Med. Chem.* 17, 2456–2481. doi: 10.2174/092986710791556023

Boitz, J. M., Strasser, R., Yates, P. A., Jardim, A., and Ullman, B. (2013). Adenylosuccinate synthetase and adenylosuccinate lyase deficiencies trigger growth and infectivity deficits in *Leishmania donovani*. *J. Biol. Chem.* 288, 8977–8990. doi: 10.1074/jbc.M112.431486

Boitz, J. M., and Ullman, B. (2013). Adenine and adenosine salvage in *Leishmania donovani*. *Mol. Biochem. Parasitol.* 190, 51–55. doi: 10.1016/j.molbiopara.2013.06.005

Boitz, J. M., Ullman, B., Jardim, A., and Carter, N. S. (2012). Purine salvage in *Leishmania*: complex or simple by design? *Trends Parasitol.* 28, 345–352. doi: 10.1016/j.pt.2012.05.005

Bora, D. (1999). Epidemiology of visceral leishmaniasis in India. *Natl. Med. J. Ind.* 12, 62–68.

Bora, N., and Nath Jha, A. (2019). An integrative approach using systems biology, mutational analysis with molecular dynamics simulation to challenge the functionality of a target protein. *Chem. Biol. Drug Design* 93, 1050–1060. doi: 10.1111/cbdd.13502

Bulusu, V., Srinivasan, B., Bopanna, M. P., and Balaram, H. (2009). Elucidation of the substrate specificity, kinetic and catalytic mechanism of adenylosuccinate lyase from *Plasmodium falciparum*. *Biochim. Biophys. Prot. Proteom.* 1794, 642–654. doi: 10.1016/j.bbapap.2008.11.021

Carter, N. S., Yates, P., Arendt, C. S., Boitz, J. M., and Ullman, B. (2008). Purine and pyrimidine metabolism in *Leishmania*. *Adv. Exp. Med. Biol.* 625, 141–154.

Cavalli, A., and Bolognesi, M. L. (2009). Neglected tropical diseases: multi-target-directed ligands in the search for novel lead candidates against *Trypanosoma* and *Leishmania*. *J. Med. Chem.* 52, 7339–7359. doi: 10.1021/jm9004835

Chappuis, F., Sundar, S., Hailu, A., Ghalib, H., Rijal, S., Peeling, R. W., et al. (2007). Visceral leishmaniasis: what are the needs for diagnosis, treatment and control? *Nat. Rev. Microbiol.* 5:S7.

Chavali, A. K., Whittemore, J. D., Eddy, J. A., Williams, K. T., and Papin, J. A. (2008). Systems analysis of metabolism in the pathogenic trypanosomatid leishmania major. *Mol. Syst. Biol.* 4:177. doi: 10.1038/msb.2008.15

Chávez-Fumagalli, M. A., Schneider, M. S., Lage, D. P., Tavares, G. D. S. V., Mendonça, D. V. C., Santos, T. T. D. O., et al. (2018). A computational approach using bioinformatics to screening drug targets for *Leishmania infantum* species. *Evid. Based Complement. Altern. Med.* 2018:1. doi: 10.1155/2018/6813467

Chawla, B., and Madhubala, R. (2010). Drug targets in leishmania. *J. Parasit. Dis.* 34, 1–13. doi: 10.1007/s12639-010-0006-3

Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., and Lin, C.-Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8:S11. doi: 10.1186/1752-0509-8-S4-S11

Chitty, J. L., Blake, K. L., Blundell, R. D., Koh, Y. A. E., Thompson, M., Robertson, A. A., et al. (2017). Cryptococcus neoformans ADS lyase is an enzyme essential for virulence whose crystal structure reveals features exploitable in antifungal drug design. *J. Biol. Chem.* 292, 11829–11839. doi: 10.1074/jbc.M117.787994

Choudhary, S. A., Bora, N., Banerjee, D., Arora, L., Das, A. S., Yadav, R., et al. (2019). A novel small molecule A_{2A} adenosine receptor agonist, indirubin-3′-monoxime, alleviates lipid-induced inflammation and insulin resistance in 3T3-L1 adipocytes. *Biochem. J.* 476, 2371–2391. doi: 10.1042/BCJ20190251 doi: 10.1042/bcj20190251

Croft, S. L., Barrett, M. P., and Urbina, J. A. (2005). Chemotherapy of trypanosomiases and leishmaniasis. *Trends Parasitol.* 21, 508–512. doi: 10.1016/j.pt.2005.08.026

Croft, S. L., and Coombs, G. H. (2003). Leishmaniasis–current chemotherapy and recent advances in the search for novel drugs. *Trends Parasitol.* 19, 502–508. doi: 10.1016/j.pt.2003.09.008

Das, S., Bora, N., Rohman, M. A., Sharma, R., Jha, A. N., and Roy, A. S. (2018). Molecular recognition of bio-active flavonoids quercetin and rutin by bovine hemoglobin: an overview of the binding mechanism, thermodynamics and structural aspects through multi-spectroscopic and molecular dynamics simulation studies. *Phys. Chem. Chem. Phys.* 20, 21668–21684. doi: 10.1039/c8cp02760a

Davis, A. J., Murray, H. W., and Handman, E. (2004). Drugs against leishmaniasis: a synergy of technology and partnerships. *Trends Parasitol.* 20, 73–76. doi: 10.1016/j.pt.2003.11.006

De Koning, H. P., Bridges, D. J., and Burchmore, R. J. (2005). Purine and pyrimidine transport in pathogenic protozoa: from biology to therapy. *FEMS Microbiol. Rev.* 29, 987–1020. doi: 10.1016/j.femsre.2005.03.004

Desjeux, P. (2004). Leishmaniasis: current situation and new perspectives. *Comp. Immunol. Microbiol. Infect. Dis.* 27, 305–318. doi: 10.1016/j.cimid.2004.03.004

Doleželová, E., Terán, D., Gahura, O., Kotrbová, Z., Procházková, M., Keough, D., et al. (2018). Evaluation of the *Trypanosoma brucei* 6-oxopurine salvage pathway as a potential target for drug discovery. *PLoS Negl. Trop. Dis.* 12:e0006301. doi: 10.1371/journal.pntd.0006301

Dos Santos Vasconcelos, C. R., De Lima Campos, T., and Rezende, A. M. (2018). Building protein-protein interaction networks for *Leishmania* species through protein structural information. *BMC Bioinform.* 19:85. doi: 10.1186/s12859-018-2105-6

El Kouni, M. H. (2003). Potential chemotherapeutic targets in the purine metabolism of parasites. *Pharmacol. Ther.* 99, 283–309. doi: 10.1016/s0163-7258(03)00071-8

Flórez, A. F., Park, D., Bhak, J., Kim, B.-C., Kuchinsky, A., Morris, J. H., et al. (2010). Protein network prediction and topological analysis in *Leishmania* major as a tool for drug target selection. *BMC Bioinform.* 11:484. doi: 10.1186/1471-2105-11-484

Franco, R., and Canela, E. I. (1984). Computer simulation of purine metabolism. *Eur. J. Biochem.* 144, 305–315. doi: 10.1111/j.1432-1033.1984.tb08465.x

Freitas-Junior, L. H., Chatelain, E., Kim, H. A., and Siqueira-Neto, J. L. (2012). Visceral leishmaniasis treatment: what do we have, what do we need and how to deliver it? *Inter. J. Parasitol.* 2, 11–19. doi: 10.1016/j.ijpddr.2012.01.003

Funahashi, A., Morohashi, M., Kitano, H., and Tanimura, N. (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico* 1, 159–162. doi: 10.1016/S1478-5382(03)02370-9

Fyfe, P. K., Dawson, A., Hutchison, M.-T., Cameron, S., and Hunter, W. N. (2010). Structure of *Staphylococcus aureus* adenylosuccinate lyase (PurB) and assessment of its potential as a target for structure-based inhibitor discovery. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 66, 881–888. doi: 10.1107/S0907444910020081

Galina, L., Dalberto, P. F., Martinelli, L. K. B., Roth, C. D., Pinto, A. F. M., Villela, A. D., et al. (2017). Biochemical, thermodynamic and structural studies of recombinant homotetrameric adenylosuccinate lyase from *Leishmania braziliensis*. *RSC Adv.* 7, 54347–54360. doi: 10.1039/c7ra10526f

Guerin, P. J., Olliaro, P., Sundar, S., Boelaert, M., Croft, S. L., Desjeux, P., et al. (2002). Visceral leishmaniasis: current status of control, diagnosis, and treatment, and a proposed research and development agenda. *Lancet Infect. Dis.* 2, 494–501. doi: 10.1016/s1473-3099(02)00347-x

Guimera, R., and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature* 433, 895–900. doi: 10.1038/nature03288

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., et al. (2006). COPASI—a complex pathway simulator. *Bioinformatics* 22, 3067–3074. doi: 10.1093/bioinformatics/btl485

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Kokina, A., Ozolina, Z., and Liepins, J. (2019). Purine auxotrophy: possible applications beyond genetic marker. *Yeast* 36, 649–656. doi: 10.1002/yea.3434

Lambris, J. D., Daniel, R., and Brian, V. G. (2008). Complement evasion by human pathogens. *Nat. Rev. Microbiol.* 6, 132–142. doi: 10.1038/nrmicro1824

Looker, D. L., Berens, R. L., and Marr, J. J. (1983). Purine metabolism in *Leishmania donovani* amastigotes and promastigotes. *Mol. Biochem. Parasitol.* 9, 15–28. doi: 10.1016/0166-6851(83)90053-1

Mandlik, V., Shinde, S., Chaudhary, A., and Singh, S. (2012). Biological network modeling identifies IPCS in Leishmania as a therapeutic target. *Integr. Biol.* 4, 1130–1142. doi: 10.1039/c2ib20037f

Marr, J. J., Berens, R. L., and Nelson, D. J. (1978). Purine metabolism in *Leishmania donovani* and *Leishmania braziliensis*. *Biochim. Biophys. Acta Gen. Sub.* 544, 360–371. doi: 10.1016/0304-4165(78)90104-6

Martin, J. L., Yates, P. A., Boitz, J. M., Koop, D. R., Fulwiler, A. L., Cassera, M. B., et al. (2016). A role for adenine nucleotides in the sensing mechanism to purine starvation in *Leishmania donovani*. *Mol. Microbiol.* 101, 299–313. doi: 10.1111/mmi.13390

McConville, M. J., De Souza, D., Saunders, E., Likic, V. A., and Naderer, T. (2007). Living in a phagolysosome; metabolism of *Leishmania amastigotes*. *Trends Parasitol.* 23, 368–375. doi: 10.1016/j.pt.2007.06.009

Meshram, R. J., Goundge, M. B., Kolte, B. S., and Gacche, R. N. (2019). An in silico approach in identification of drug targets in leishmania: a subtractive genomic and metabolic simulation analysis. *Parasitol. Int.* 69, 59–70. doi: 10.1016/j.parint.2018.11.006

O'Gara, M. J., Lee, C.-H., Weinberg, G. A., Nott, J. M., and Queener, S. F. (1997). IMP dehydrogenase from Pneumocystis carinii as a potential drug target. *Antimicrob. Agents Chemother.* 41, 40–48. doi: 10.1128/aac.41.1.40

Pitaluga, A., Moreira, M., and Traub-Csekö, Y. (2015). A putative role for inosine 5′ monophosphate dehydrogenase (IMPDH) in *Leishmania amazonensis* programmed cell death. *Exp. Parasitol.* 149, 32–38. doi: 10.1016/j.exppara.2014.12.006

Rasti, S., and Vogiatzis, C. (2019). A survey of computational methods in protein–protein interaction networks. *Ann. Operat. Res.* 276, 35–87.

Rezende, A. M., Folador, E. L., Resende, D. D. M., and Ruiz, J. C. (2012). Computational prediction of protein-protein interactions in Leishmania predicted proteomes. *PLoS One* 7:e51304. doi: 10.1371/journal.pone.0051304

Romanello, L., Serrão, V. H. B., Torini, J. R., Bird, L. E., Nettleship, J. E., Rada, H., et al. (2017). Structural and kinetic analysis of *Schistosoma mansoni Adenylosuccinate Lyase* (SmADSL). *Mol. Biochem. Parasitol.* 214, 27–35. doi: 10.1016/j.molbiopara.2017.03.006

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., et al. (2004). BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* 32, D431–D433. doi: 10.1093/nar/gkh081

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharma, M., Shaikh, N., Yadav, S., Singh, S., and Garg, P. (2017). A systematic reconstruction and constraint-based analysis of *Leishmania donovani* metabolic network: identification of potential antileishmanial drug targets. *Mol. Biosyst.* 13, 955–969. doi: 10.1039/c6mb00823b

Shu, Q., and Nair, V. (2008). Inosine monophosphate dehydrogenase (IMPDH) as a target in drug discovery. *Med. Res. Rev.* 28, 219–232. doi: 10.1002/med.20104

Smith, P. A., and Romesberg, F. E. (2007). Combating bacteria and drug resistance by inhibiting mechanisms of persistence and adaptation. *Nat. Chem. Biol.* 3, 549–556. doi: 10.1038/nchembio.2007.27

Stein, L. D. (2003). Integrating biological databases. *Nat. Rev. Genet.* 4, 337–345. doi: 10.1038/nrg1065

Steuer, R., Gross, T., Selbig, J., and Blasius, B. (2006). Structural kinetic modeling of metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11868–11873. doi: 10.1073/pnas.0600013103

Sundar, S. (2001). Drug resistance in Indian visceral leishmaniasis. *Trop. Med. Int. Health* 6, 849–854. doi: 10.1046/j.1365-3156.2001.00778.x

Sundar, S., and Chatterjee, M. (2006). Visceral leishmaniasis-current therapeutic modalities. *Ind. J. Med. Res.* 123:345.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452.

Van Riel, N. A. (2006). Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief. Bioinform.* 7, 364–374. doi: 10.1093/bib/bbl040

Vijayakumar, S., and Das, P. (2018). Recent progress in drug targets and inhibitors towards combating leishmaniasis. *Acta Trop.* 181, 95–104. doi: 10.1016/j.actatropica.2018.02.010

Wang, C. C. (1984). Parasite enzymes as potential targets for antiparasitic chemotherapy. *J. Med. Chem.* 27, 1–9. doi: 10.1021/jm00367a001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership