Journal of proteome research

Cite This: J. Proteome Res. XXXX, XXX, XXX−XXX

Article

pubs.acs.org/jpr

Systematic Protein Prioritization for Targeted Proteomics Studies through Literature Mining

Kun-Hsing Yu,[†,‡] Tsung-Lu Michael Lee,[§] Chi-Shiang Wang,[∥] Yu-Ju Chen,[⊥] Christopher Ré,[#] Samuel C. Kou,[‡] Jung-Hsien Chiang,[*,∥] Isaac S. Kohane,[*,†] and Michael Snyder[*,▽]

[†]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, United States

[‡]Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, United States

[§]Department of Information Engineering, Kun Shan University, Tainan City 710-03, Taiwan

[∥]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan City 701-01, Taiwan

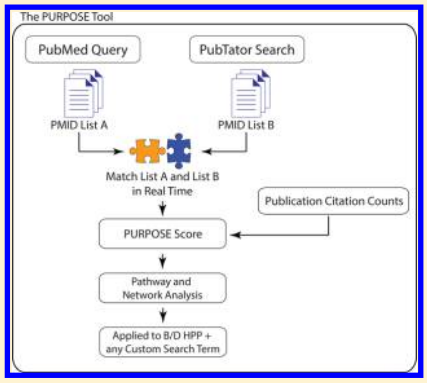
[⊥]Institute of Chemistry, Academia Sinica, Taipei 115-29, Taiwan

[#]Department of Computer Science, Stanford University, Stanford, California 94305, United States

[▽]Department of Genetics, Stanford University, Stanford, California 94305, United States

Supporting Information

ABSTRACT: There are more than 3.7 million published articles on the biological functions or disease implications of proteins, constituting an important resource of proteomics knowledge. However, it is difficult to summarize the millions of proteomics findings in the literature manually and quantify their relevance to the biology and diseases of interest. We developed a fully automated bioinformatics framework to identify and prioritize proteins associated with any biological entity. We used the 22 targeted areas of the Biology/Disease-driven (B/D)-Human Proteome Project (HPP) as examples, prioritized the relevant proteins through their Protein Universal Reference Publication-Originated Search Engine (PURPOSE) scores, validated the relevance of the score by comparing the protein prioritization results with a curated database, computed the scores of proteins across the topics of B/D-HPP, and characterized the top proteins in the common model organisms. We further extended the bioinformatics workflow to identify the relevant proteins in all organ systems and human diseases and deployed a cloud-based tool to prioritize proteins related to any custom search terms in real time. Our tool can facilitate the prioritization of proteins for any organ system or disease of interest and can contribute to the development of targeted proteomic studies for precision medicine.



KEYWORDS: *proteomics, literature mining, bioinformatics, Human Proteome Project, information retrieval*

INTRODUCTION

Proteomics studies provide significant insights into the dynamics of biological processes related to normal physiology and disease pathology.[1] There are more than 3.7 million published papers describing proteins implicated in various human health and disease status.[2] These studies elucidate the biological pathways related to diseases,[3−7] describe potential biomarkers,[8] and reveal potential drug targets.[9] Recently, the Precision Medicine Initiative launched population-scale studies of the human genome and proteome, which promoted and inspired international collaborations to generate high-throughput biomedical data, to compile health information on millions of participants, and to publish biomedical insights gleaned from these studies.[10−12] As an illustration, the well-established Human Proteome Organization (HUPO) Biology/Disease-driven (B/D)-Human Proteome Project (HPP) is a coordinated international effort to systematically study the human proteome under different conditions. The goal of the HPP is to make proteomics a standard companion to other high-throughput "omics" studies, including genomics, transcriptomics, epigenomics, and metabolomics, in the integrative investigations of all diseases. It is expected to further expand our knowledge and scientific literature on the proteomic landscapes of the 22 targeted fields, including cancer, diabetes, infectious diseases, the cardiovascular system, the liver, mitochondria, and plasma.[13−15]

The accumulation of biomedical publications in the PubMed database presents an opportunity to identify the important proteins associated with known human biology and diseases. However, the distribution of this collective knowledge in millions of PubMed articles makes it very difficult for investigators to track and summarize the literature in real time; currently, there are 27 million publications in total in the PubMed database, with approximately 800 000 added every year.[16,17] A significant proportion of recent publications

Received: October 28, 2017
Published: March 5, 2018

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

describes the functions of proteins, genes, and their relations to phenotypes or diseases.[2] As is standard practice in many fields, many researchers have periodically summarized the most significant findings in their areas in the form of review articles. However, as the number of proteomic and other publications increases rapidly, it is becoming ever more difficult to keep up with the current literature through conventional approaches. An automated method for prioritizing proteins of interest for research in normal biology and diseases is needed.

Previous studies have shown that text mining algorithms can help prioritize proteins of interest and facilitate biomedical investigations.[18] Several groups of researchers employed text-mining strategies to extract protein−protein interactions,[19] to identify the relations between human diseases, genes, drugs, and metabolites,[20] and to discover novel drug-gene interactions and gene−gene interactions.[21] A recent study designed copublication metrics to prioritize proteins in six organ systems.[22] These studies demonstrated the feasibility of literature mining in retrieving useful and nontrivial information from published work.

However, very few studies have systematically investigated and summarized the proteins relevant to organ systems or diseases or compared the importance of a protein in different organ systems, which hinders the progress of protein array design and targeted proteomics studies.[18] In the recent decades, various proteomics platforms were developed for the characterization of thousands of proteins.[23] However, there are more than tens of thousands of proteins presented in organisms, and targeted proteomics assays cannot cover all of the protein candidates. A comprehensive literature mining tool is needed to facilitate the development of targeted proteomics assay by prioritizing the protein candidates for any given biological system or disease of interest. The development of an effective literature-based protein prioritization platform will greatly facilitate biomarker discovery or drug target investigations.

We designed an automated literature mining platform that systematically retrieves biomedical literature, analyzes the relations between proteins and any organ system, phenotypes, or diseases in various species, and generates a ranked list for any topic of interest. To demonstrate the utility of our platform, we presented our protein prioritization results of the 22 B/D-HPP targeted fields, which encompassed 22 prevalent human diseases and important organ-systems with ongoing systematic proteomics studies on each of these areas.[15] In addition to publication frequency, our approach quantified the specificity of a protein to the topic of interest and accounted for the number of citations of each article. We generated a ranked list of proteins related to each of the B/D-HPP targeted areas and compared the protein prioritization results with a curated database to validate the relevance of the identified proteins to the B/D of interest. The platform is freely available as a cloud service, allowing researchers of all fields to query any proteins, diseases, phenotypes, or other biological entities in human and the common model organisms of interest. This platform will facilitate the comprehensive investigations on the key areas of the B/D-HPP as well as expedite targeted proteomics studies on human and other organisms, which will promote the progression from proteomics-based discovery to translational research for precision medicine. Our approach can also be generalized beyond proteomics to analyze other areas of biomedical investigations, such as metabolomics, protein modifications, and alternative splicing isoforms.

MATERIALS AND METHODS

Identification of the B/D-HPP Targeted Fields, Human Organ Systems, and Human Diseases

B/D-HPP targeted fields, human organ systems, and human diseases were identified as the target areas for protein prioritization. The 22 targeted fields of the B/D-HPP were obtained from the HPP Web site, which include brain, cancer, cardiovascular, diabetes, extreme conditions, EyeOME, food and nutrition, glycoproteomics, immune-peptidome, infectious diseases, kidney and urine, liver, mitochondria, model organisms, musculoskeletal, pathology, PediOme (the human pediatric proteome), plasma, protein aggregation, rheumatic disorders, stem cells, and toxicoproteomics.[24] Eleven systems of the human anatomy, including cardiovascular/circulatory, digestive/excretory, endocrine, integumentary, lymphatic/immune, muscular, nervous, renal/urinary, reproductive, respiratory, and skeletal systems were identified.[25] Phenome-wide association scan (PheWAS) codes were used to identify the phenotypical groups of human diseases.[26] All 1866 PheWAS codes, which summarized 15 558 International Classification of Diseases version 9 (ICD-9) codes, were included in our analysis.[26]

Retrieval of Publications Associated with Proteins and Topical Queries

PubTator files were obtained from the National Center for Biotechnology Information (NCBI) PubTator portal, which contains annotations of genes, species, and diseases.[2] Tagged genes, diseases, organ systems, and species were extracted from the PubTator database. Because the PubTator files update every few weeks, an automated file downloader was implemented to obtain and process the updated files periodically. The name and symbol of each protein were retrieved through the NCBI gene information tool.[27] For each protein, the PubMed identifiers (PMIDs) of the papers tagged with the protein were identified. The title, authors, journal, publication years, and citation numbers of the publications were retrieved by the NLM Entrez Programming Utilities (E-utilities).[28] Because older publications may accumulate more citations than recently published papers with equal importance, the numbers of citations were normalized against the number of years the papers had been published. With this method, the recently discovered protein−biology associations would not be heavily penalized when compared with well-established associations. In addition, the classic protein−biology associations will still receive substantial weight given their consistent trend of publication and citation.

For each topical query (T), the NLM E-utilities were used to retrieve the PMIDs associated with T, and the list of PMIDs retrieved from T was intersected with the PMIDs associated with each protein tagged by PubTator. A summary of our method is shown in Figure 1A.

Summarizing Protein Publication and Citation with the Protein Universal Reference Publication-Originated Search Engine (PURPOSE) Score

To quantify the strength of the associations between proteins and topics, a Protein Universal Reference Publication-Originated Search Engine (PURPOSE) score was designed. The PURPOSE score is defined as follows

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

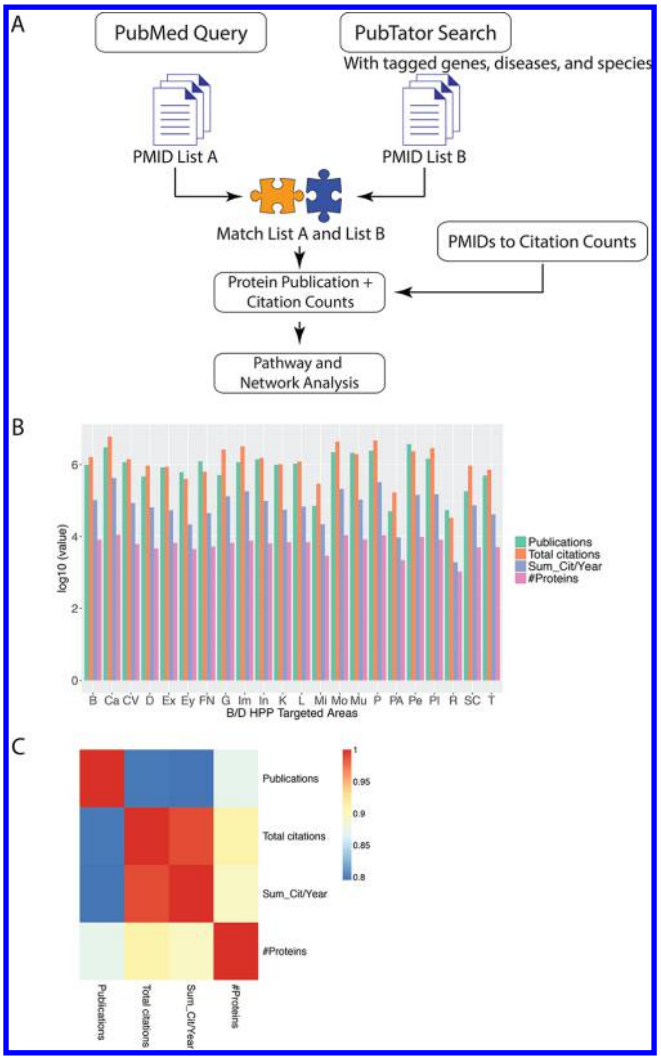


Figure 1. Protein prioritization for the 22 targeted areas of the Biology/Disease-driven Human Proteome Project (B/D-HPP). (A) Developed bioinformatics workflow for real-time literature mining. (B) Summary statistics of the literature pertaining to the B/D-HPP targeted fields. The number of publications, the total number of citations, the number of citations per year (Sum_Cit/Year), and the number of unique protein counts (#Proteins) are shown. B: brain, Ca: cancer, CV: cardiovascular, D: diabetes, Ex: extreme conditions, Ey: EyeOME, FN: food and nutrition, G: glycoproteins, Im: immunopeptidome, In: infectious diseases, K: kidney and urine, L: liver, Mi: mitochondria, Mo: model organisms, Mu: musculoskeletal, P: pathology, Pe: PediOme, Pl: plasma, PA: protein aggregation, R: rheumatic disorders, SC: stem cells, T: toxicoproteomics. (C) Correlation matrix of the number of publications, the number of citations, the number of citations per published years, and the number of proteins copublished with each of the B/D-HPP fields.

$$\left(1 + \log_{10} nTP + \log_{10} \frac{Sum\left(\frac{Cit}{Yr}\right) + 1}{10}\right) \times \left(1 + \log_{10} \frac{nU}{nT} + \log_{10} \frac{nU}{nP}\right)$$

where nTP is the number of papers related to both the protein and the topic (TP), $Sum(Cit/Yr)$ is the sum of the annualized citation number of TP, nU is the number of publications in the PubMed database, nT is the number of publications regarding the topic of interest, and nP is the number of publications regarding the protein of interest. This score is analogous to the term frequency-inverse document frequency (tf-idf) measurement commonly used in text mining:[29] The first part of the protein prioritization score is proportional to the strength of TP copublications weighted by the annualized citation, and the second part penalizes well-published topics and proteins without significant proportions of copublications, thereby prioritizing topic-specific proteins with substantial copublication strengths. The number of citations per year was included to account for the visibility and importance of each publication. To avoid potential bias, the number of citations per year was calculated at the paper level and was not aggregated to the journal level. The PURPOSE score served as a foundation for measuring the importance of different proteins associated with each topic of interest. For any given protein, the score was also used to compare the relevance of different organ systems or HPP areas.

Prioritizing Proteins Associated with the 22 B/D-HPP Targeted Areas and Genetic Variations with Clinical Significance

To identify the proteins associated with the B/D-HPP targeted areas, the protein PURPOSE score was calculated for each human protein in each of the 22 B/D-HPP topics. The query term associated with each B/D-HPP area is shown in Table S-1. For each targeted area, the proteins with the highest scores were identified, and their numbers of publications and citations were compared.

To characterize the copublication patterns of proteins and genetic variations of clinical significance in cancers, the most studied mutated genes in cancers were identified by using "mutation cancer" as a PubMed search term and ranking the associated proteins by their PURPOSE scores. For specific genetic mutations, the associated publications were retrieved by querying the mutation as a keyword. As an illustration, "EGFR mutation" was used as the PubMed search term to retrieve publications associated with EGFR mutation. Thus the search results would encompass all mutational variants of the queried gene. For cancer-type-specific analysis, the ten most prevalent cancers worldwide were identified from the global cancer statistics,[30] and the mutated genes frequently copublished with each type of cancer were determined. The prevalence rates of the frequently published genetic mutations in cancer were obtained from the pan-cancer study in the Memorial Sloan Kettering Cancer Center.[31] The identified mutated genes were used as a search term, and the PURPOSE scores of the copublished proteins were calculated and ranked.

Gene Ontology (GO)[32] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[33] pathway enrichment analyses were performed on the retrieved protein list of each topic of interest to identify the significantly associated biological entities and pathways. Benjamini–Hochberg procedure was used to correct for multiple tests in the enrichment analyses. Protein–protein interactions were visualized using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)[34] database tool. Significantly higher numbers of protein–protein interactions indicated that the identified proteins likely participated in related biological pathways. All analysis results were retrieved on October 20, 2017.

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

Protein-based Search across the 22 B/D-HPP Targeted Areas

To better understand the publication trend of human proteins, the proteins with the highest number of PubMed publications in human (*Homo sapiens*) were identified, and their PURPOSE scores across the 22 B/D-HPP targeted areas were compared. The numbers of publications and citations for each B/D-HPP area were calculated and visualized. The known functions of these top proteins were retrieved from neXtProt[35] and UniProt[36] and compared with the known biological mechanisms associated with the B/D-HPP topics.

Comparison of the Top-Ranked Proteins in Popular Model Organisms

Many model organisms, including mouse (*Mus musculus*), rat (*Rattus norvegicus*), common fruit fly (*Drosophila melanogaster*), roundworm (*Caenorhabditis elegans*), and yeast (*Saccharomyces cerevisiae*), are routinely used in proteomic studies.[37−39] To characterize the publication patterns of proteins in these model organisms, the proteins with the largest numbers of publication in each of these organisms were identified. The ortholog relations among proteins in different species were identified using the Protein ANalysis THrough Evolutionary Relationships (PANTHER) database.[40,41] The associations between the top proteins in human and those in the model organisms were quantified using the Spearman correlation coefficient.

Evaluation of the Prioritized Lists

Biologist-curated lists of known protein−topic relations from the Comparative Toxicogenomics Database (CTD)[42] were used to evaluate the prioritized lists of proteins retrieved from the PURPOSE system. "Cardiovascular," "kidney," "liver," and "lung" were selected as targets for evaluation because they are among the B/D-HPP targeted areas and are terms of general interest. The CTD gene−disease association table was used, and diseases associated with the targeted terms were retrieved by a MeSH tree search. A detailed list of search parameters and results is included in Data S-1. Precision, recall, and the F1 score were employed to quantify the performance of the PURPOSE algorithm. The performance of existing tools with similar functionalities, such as GLAD4U[43] and PubPular,[22] was also evaluated and compared with that of the PURPOSE algorithm. In addition, a variant of PURPOSE algorithm that did not incorporate citation counts but was otherwise identical was included in the comparison. To mimic users' search behavior, the terms "cardiovascular", "kidney", "liver", and "lung" were inputted into the systems under evaluation on the same day (February 2, 2018). The evaluation is restricted to human proteins due to the fact that the CTD curations are mostly about human. Because the PubPular tool only shows the top 500 proteins associated with the search query, the comparison among PURPOSE, GLAD4U, and PubPular was limited to the top 500 proteins.

Interactive User Interface

A cloud-based user interface that enables real-time protein queries was built using the R Shiny application. The R packages "shinySky", shiny bootstrap ("shinyBS"), "shinyjs", and an interface to Google's open-source JavaScript engine ("V8") were employed to establish an interactive web interface. The interactive plots were generated by R packages "ggplot2" and "plotly." All statistical analysis was performed and visualized with R version 3.3.3. The established system runs the user-defined queries in real time, performs on-demand PubMed searches on the most updated database using a java code interfaced with PubMed E-utilities, and summarizes and visualizes the protein prioritization results for each query in 1−30 s. For each protein associated with the query, the number of publications, number of citations per year, and the PURPOSE score are calculated and reported. The distribution of the scores of the proteins is visualized with scatterplots. Pathway analyses using the Database for Annotation, Visualization and Integrated Discovery (DAVID)[44,45] and protein−protein interaction analysis with the STRING tool[34] are conducted on demand through their application programming interfaces (APIs). A ranked list of relevant proteins, the distribution of PURPOSE scores, numbers of publications and citations, and the most-cited publications associated with the top proteins are shown in the interactive user interface. The source codes are available at http://rebrand.ly/proteinpurposesourcecode.

RESULTS

Summary of Publication and Citation Distribution in the PubMed Database

We designed a fully automated tool to retrieve publications associated with any topic in the PubMed Database, linked them with mentions of specific genes and proteins in the titles and abstracts of the papers, summarized the number of publications and citations of each topic-protein pair, and calculated the PURPOSE scores to prioritize the proteins associated with each topic (Figure 1A).

We first summarized the publication and citation distribution in the PubMed database. For each of the 22 B/D-HPP topics, we calculated the number of the associated PubMed publications mentioning specific proteins and genes (Figure 1B). The 22 topics include brain, cancer, cardiovascular, diabetes, extreme conditions, EyeOME, food and nutrition, glycoproteomics, immune-peptidome, infectious diseases, kidney and urine, liver, mitochondria, model organisms, musculoskeletal, pathology, PediOme (the human pediatric proteome), plasma, protein aggregation, rheumatic disorders, stem cells, and toxicoproteomics. Results showed that the most published topics among the 22 B/D-HPP targeted areas are PediOme (3.74 million publications), cancer (3.03 million), pathology (2.43 million), model organisms (2.22 million), and musculoskeletal system (2.13 million), demonstrating the scale of publications in these broad and popular research topics. For the least studied topics among the 22 B/D-HPP topics, such as protein aggregation, there are still more than 50 000 associated publications discussing specific proteins or genes. For all of the 22 B/D-HPP topics, there are more than 1000 human proteins with at least one publication mentioning both the field and the protein, indicating the richness of the current PubMed database. The B/D-HPP topics with the largest number of copublished proteins are cancer (11 213 proteins), model organisms (10 936 proteins), and pathology (10 876 proteins).

We observed that topics with more publications generally had a higher total number of citations (Spearman's correlation coefficient = 0.85) and a higher number of citations per year (Spearman's correlation coefficient = 0.83). Similarly, the topics with more publications were associated with more copublished proteins (Spearman's correlation coefficient = 0.87). The number of total citations was strongly correlated with the number of citations per year (Spearman's correlation coefficient = 0.99; Figure 1C). These results indicated that the heavily

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

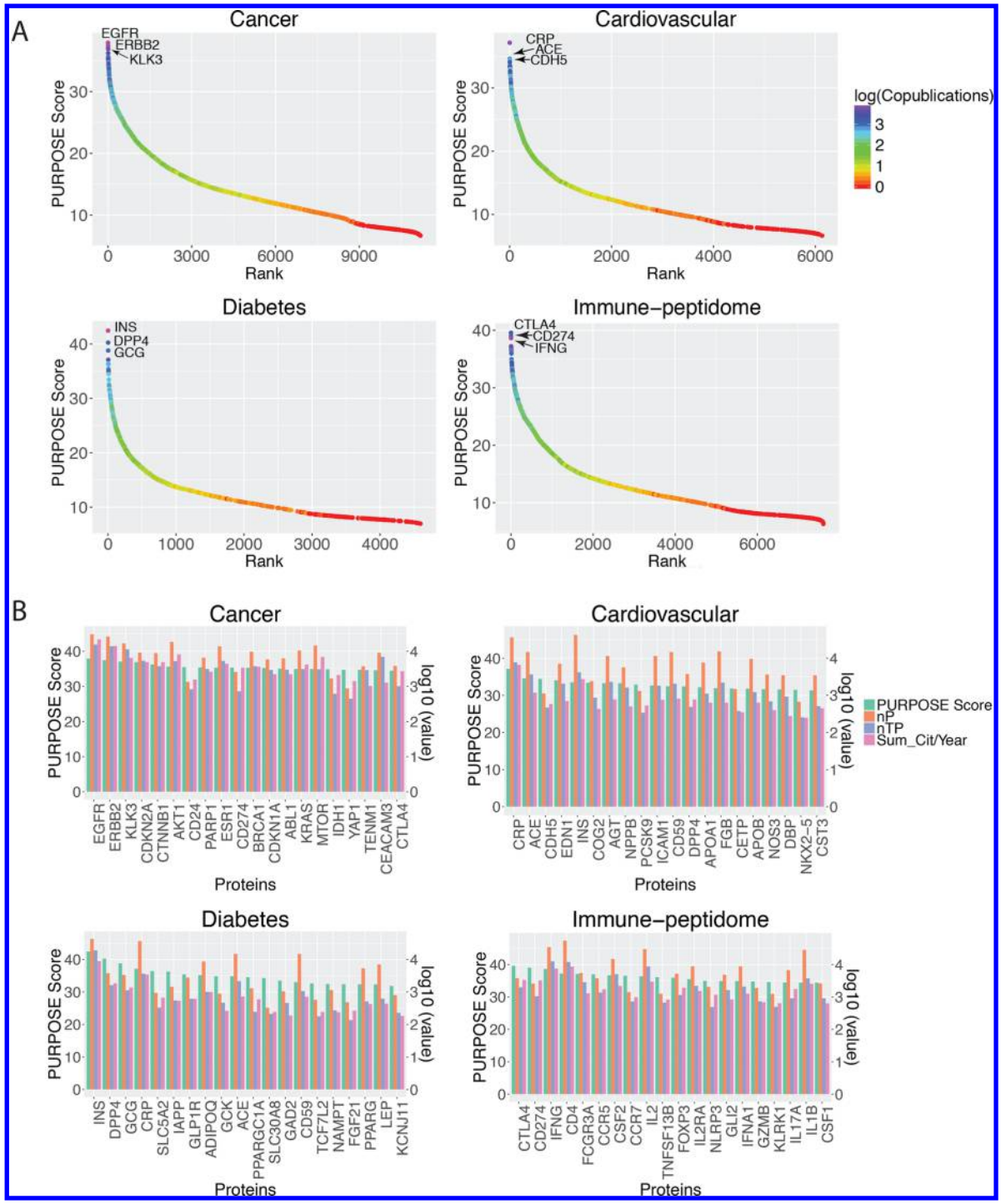


Figure 2. continued

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

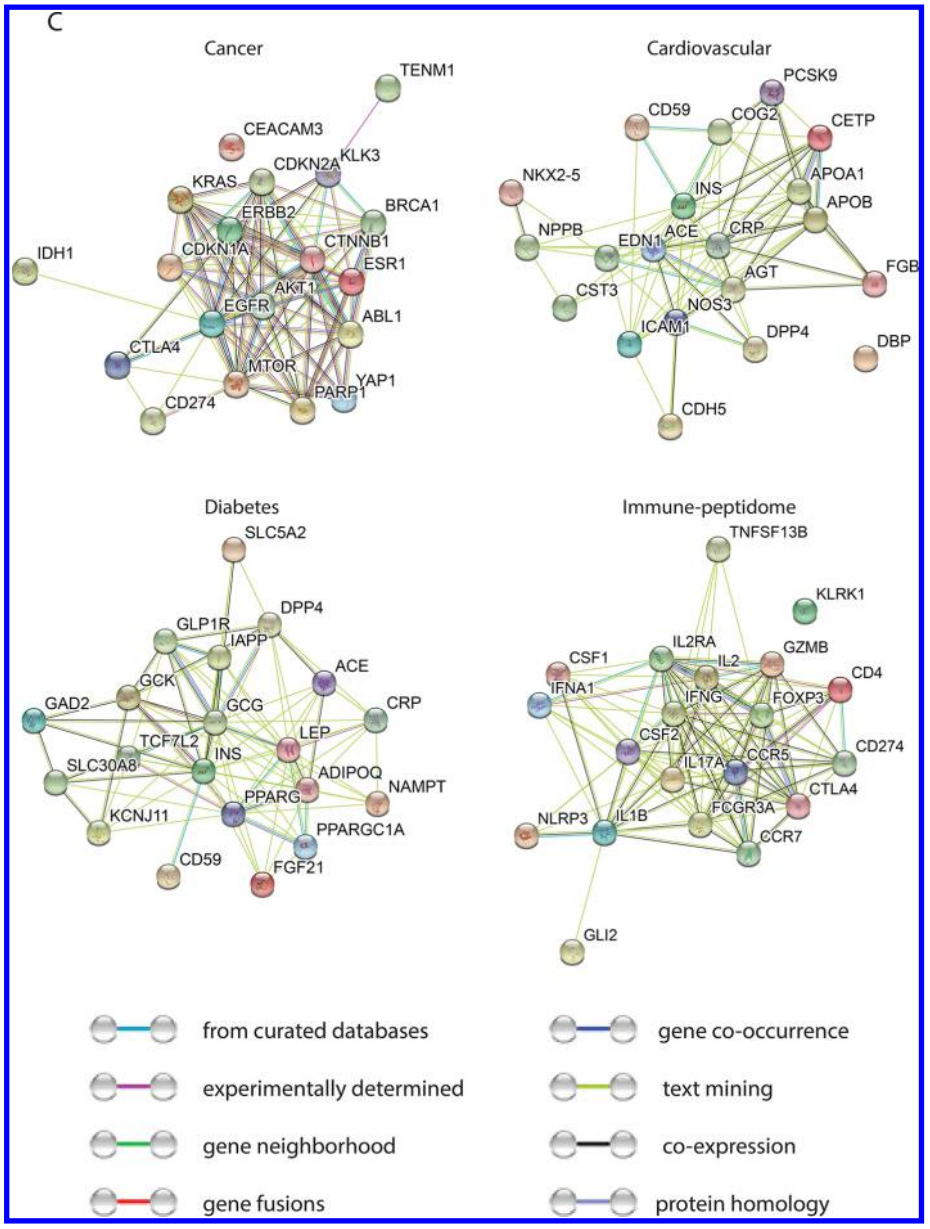


Figure 2. Distribution of the protein PURPOSE scores in the selected B/D-HPP targeted areas. (A) Protein publication scores of cancer, cardiovascular, diabetes, and immune-peptidome are shown, and the top three proteins with the highest scores are labeled. (B) Distribution of the protein PURPOSE score and its components for cancer, cardiovascular, diabetes, and immune-peptidome. nTP: the number of papers related to both the protein and the topic (TP); nP: the number of publications regarding the protein of interest; Sum_Cit/Year: the sum of the annualized citation number of TP. *Y* axis on the left: PURPOSE score; *Y* axis on the right: log10(value) of all other variables. For each topic, only the top 20 proteins with the highest PURPOSE score are shown. (C) Network analyses revealed significant protein−protein interactions among the identified proteins, and enrichment analyses confirmed their relevance to the B/D-HPP fields. Selected analyses on cancer, cardiovascular, diabetes, and immune-peptidome are shown. Please note that the CD24 protein identified in panel B in cancer was not present in the STRING database.

published topics may represent the popular fields of study and thus attracted more citations and had more copublished proteins in the PubMed literature database.

Prioritizing Proteins in the 22 Targeted Fields of the Biology/Disease-Driven Human Proteome Project

To identify the proteins associated with the 22 targeted fields of the B/D-HPP, we computed the PURPOSE score for each protein copublished with each of these fields. The design of the PURPOSE score follows the weighting principle used by tf-idf, which balances the strength of copublication and the specificity of the retrieved associations. (Please see the Materials and Methods section.) The PURPOSE score accounts for the specificity of the associations between the protein under investigation and the topic of interest while integrating the number of copublications and citations. Our algorithm calculates the PURPOSE score for thousands of proteins within seconds once the required bibliographical information is retrieved through our bioinformatics workflow.

The PURPOSE score effectively identified the proteins relevant to each topic in the B/D-HPP. Well-established protein−topic associations were successfully retrieved (Figure 2A and Data S-2). For instance, three well-known oncoproteins, EGFR (epidermal growth factor receptor; score = 37.93), ERBB2 (erb-b2 receptor tyrosine kinase 2; score = 37.45), and KLK3 (kallikrein related peptidase 3; score = 37.05) were

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

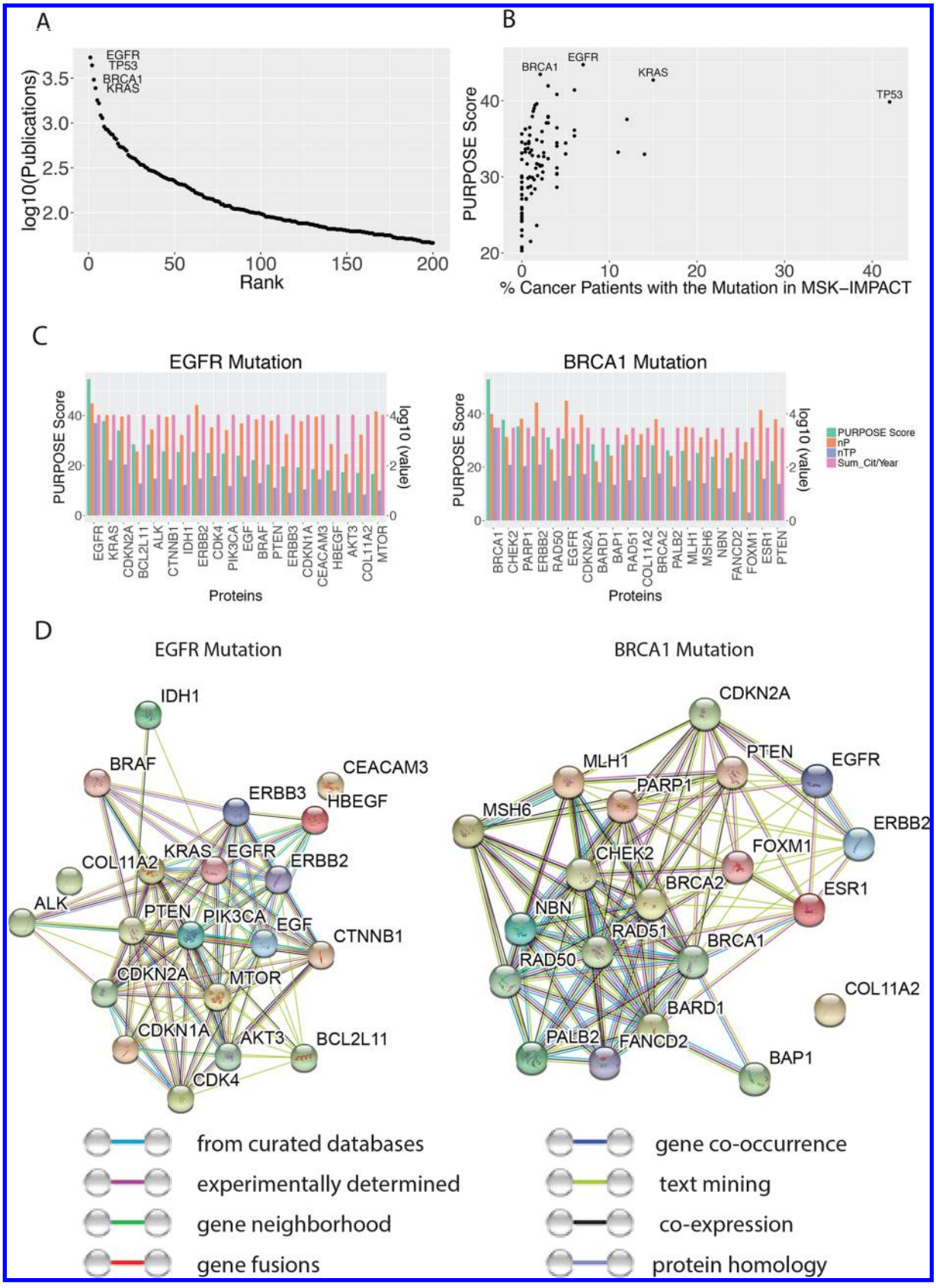


Figure 3. Distribution of the protein PURPOSE scores in selected genetic mutations with known clinical significance. (A) Distribution of the numbers of publications for the most frequently published genetic mutations in cancer. EGFR, TP53, BRCA1, and KRAS mutations are the four most frequently published genetic mutations in cancer. (B) PURPOSE scores of the 100 most published genetic mutations in cancer are moderately associated with their prevalence rates in the recent pan-cancer study (MSK-IMPACT) (Spearman's correlation coefficient = 0.64). (C) Distribution of the protein PURPOSE score and its components for EGFR and BRCA1 mutations. nTP: the number of papers related to both the protein and the topic (TP); nP: the number of publications regarding the protein of interest; Sum_Cit/Year: the sum of the annualized citation number of TP. *Y* axis

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

Figure 3. continued

on the left: PURPOSE score; *Y* axis on the right: log10(value) of all other variables. For each genetic mutation, only the top 20 proteins with the highest PURPOSE score associated with it are shown. (D) Network analyses identified significant protein−protein interactions among the identified proteins, and enrichment analysis confirmed their relevance to the genetic mutation.

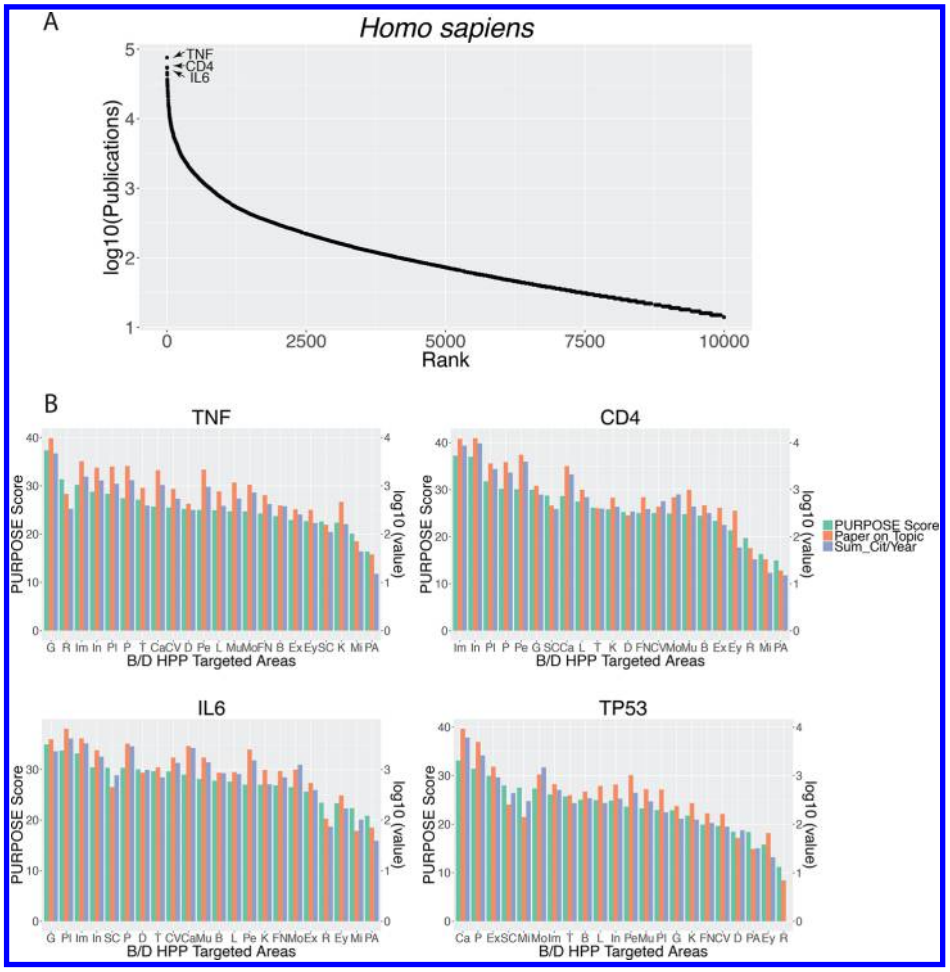


Figure 4. Well-studied proteins in human and their PURPOSE scores in the B/D-HPP targeted areas. (A) Distribution of the number of publications associated with each human protein. The top three proteins with the largest numbers of publications are annotated. (B) PURPOSE scores of the well-studied proteins across the 22 B/D-HPP targeted areas. Sum_Cit/Year: the sum of the annualized citation number of papers related to both the protein and the topic. *Y* axis on the left: PURPOSE score; *Y* axis on the right: log10(value) of all other variables. B: brain, Ca: cancer, CV: cardiovascular, D: diabetes, Ex: extreme conditions, Ey: EyeOME, FN: food and nutrition, G: glycoproteins, Im: immune-peptidome, In: infectious diseases, K: kidney and urine, L: liver, Mi: mitochondria, Mo: model organisms, Mu: musculoskeletal, P: pathology, Pe: PediOme, Pl: plasma, PA: protein aggregation, R: rheumatic disorders, SC: stem cells, T: toxicoproteomics.

found to have the highest copublication scores in the cancer category; CRP (C-reactive protein; score = 37.16), ACE (angiotensin I converting enzyme; score = 34.60), and CDH5 (cadherin 5; score = 34.44) occupy the top of the ranked list for cardiovascular; INS (insulin; score = 42.46), DPP4 (dipeptidyl peptidase 4; score = 40.26), and GCG (glucagon; score = 38.81) achieve the highest scores for diabetes; and CTLA4 (cytotoxic T-lymphocyte associated protein 4; score = 39.57), CD274 (score = 39.02), and IFNG (interferon gamma; score = 38.63) are strongly associated with the immune system (Figure 2B).

Enrichment analysis revealed that the proteins with high PURPOSE scores are highly correlated with the known biological pathways of the organ system. As an illustration, the top 20 proteins associated with cancer (Figure 2B) are significantly enriched in the positive regulation of cell−cell adhesion, regulation of signal transduction, and regulation of cell proliferation (corrected *P* value <0.0001). KEGG pathways analysis demonstrated that the same set of proteins is associated with pathways in cancer, microRNAs in cancer, and proteoglycans in cancer, glioma, and prostate cancer (corrected *P* value <0.0001), and these proteins have significant protein−protein interactions as compared with a random set of proteins ($P < 0.001$; Figure 2C). As another example, the 20 proteins with the highest PURPOSE scores for diabetes are enriched in the regulation of gluconeogenesis, regulation of cellular ketone metabolic process, glucose homeostasis, regulation of protein secretion, and regulation of hormone secretion (corrected *P* value <0.0001). These proteins participate in many KEGG pathways, including insulin secretion, type II diabetes mellitus, the PPAR signaling pathway, maturity onset diabetes of the young, and the AMPK signaling pathway (corrected *P* value

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

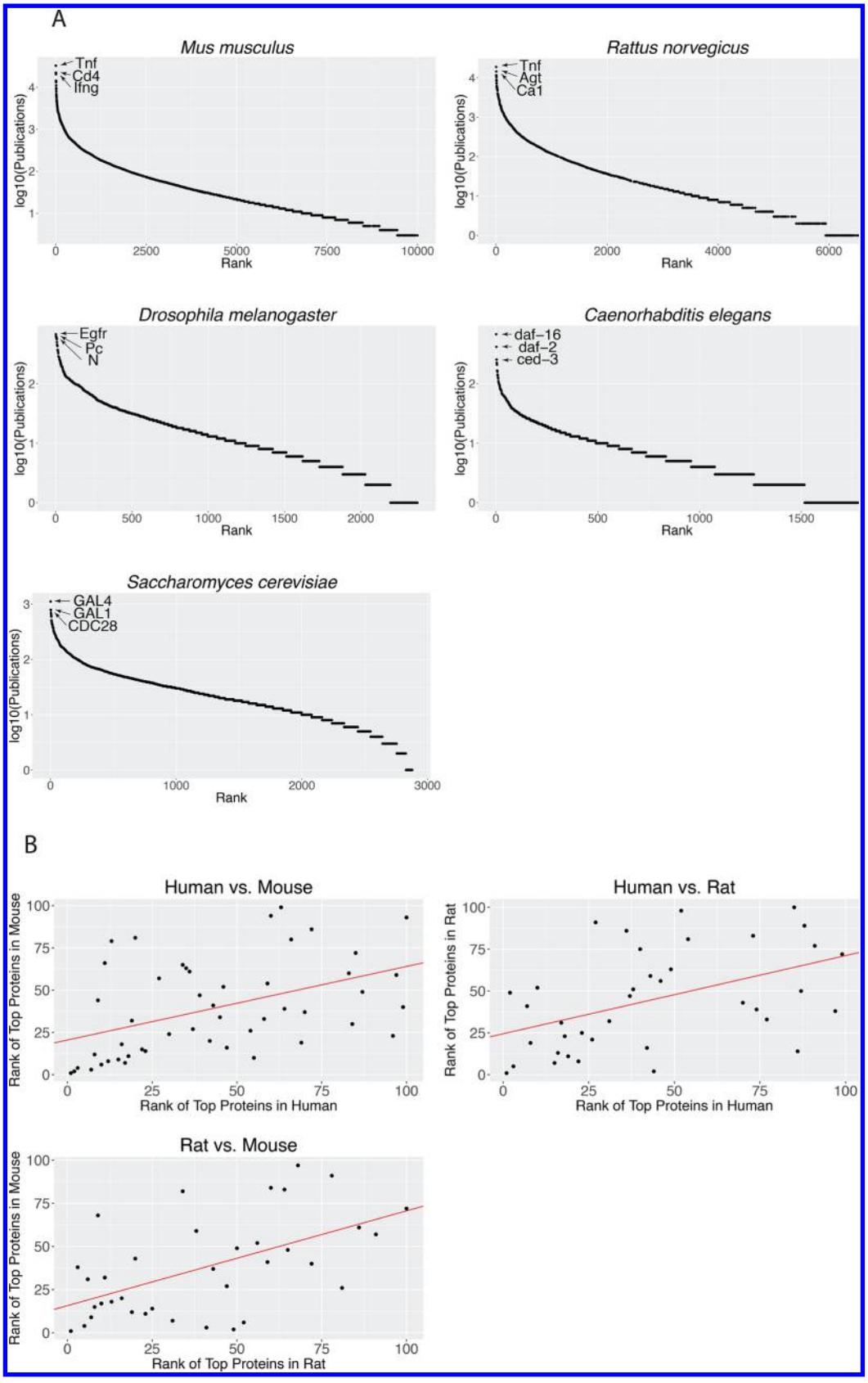


Figure 5. Comparisons among the most published proteins in common model organisms, including mouse (*Mus musculus*), rat (*Rattus norvegicus*), common fruit fly (*Drosophila melanogaster*), roundworm (*Caenorhabditis elegans*), and yeast (*Saccharomyces cerevisiae*). (A) Distribution of the number of publications associated with each protein in each organism is shown. (B) Comparisons of common proteins in human, mouse, and rat by their rank of publication numbers in each species. The most-published proteins among the three species have moderate correlations (Spearman correlation coefficients: human versus mouse = 0.53; human versus rat = 0.51; mouse versus rat = 0.55).

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

<0.0005) and have significant protein−protein interactions (P < 0.001; Figure 2C). The top-ranking proteins in other B/D-HPP topics, including cardiovascular and immune-peptidome, were also related to the known pathways associated with the topic. These results validated the associations between the prioritized proteins and the queried B/D-HPP targeted areas.

Prioritizing Proteins Associated with Genetic Mutations in Cancers

Many somatic genetic mutations in cancers are associated with patients' prognosis and treatment response. To identify the proteins commonly copublished with the clinically relevant mutations, we computed the PURPOSE score for each protein associated with the gene mutations frequently copublished with cancer. The phrase "mutation cancer" was used as the search term to retrieve the commonly mutated genes tagged by PubTator. The proteins copublished with each of the identified genetic mutations were retrieved, respectively.

Results showed that EGFR (5378 publications), TP53 (4392 publications), BRCA1 (3039 publications), and KRAS (2445 publications) are the four genes whose mutations had the highest number of cancer-related publications in the PubMed database (Figure 3A). Among the 100 most published genetic mutations in cancer, their PURPOSE scores were moderately associated with their prevalence rates in a recent pan-cancer study cohort (MSK-IMPACT) in the Memorial Sloan Kettering Cancer Center[31] (Spearman's correlation coefficient = 0.64; Figure 3B). A list of frequently copublished mutations with specific tumor types is shown in Table S-2.

We demonstrated that the proteins copublished with the known genetic variations participated in related biological pathways. As an illustration, KRAS, CDKN2A, BCL2L11, ALK, and CTNNB1 are frequently copublished with EGFR mutation (Figure 3C). These proteins are highly enriched in protein kinase activity and fibroblast growth factor receptor signaling pathway (adjusted hypergeometric test P value <0.01) and form a tight protein−protein interaction network (P < 0.0001; Figure 3D). Known protein−protein interactions, such as the biological pathway association between EGFR and KRAS[46,47] as well as the coexpression between EGFR and CDKN2A,[48] were successfully identified in the network. As another example, CHEK2, PARP1, ERBB2, and RAD50 are commonly copublished with BRCA1 mutation (Figure 3C). These proteins are significantly enriched in many molecular functions including DNA repair and the regulation of the DNA metabolic process as well as in the KEGG mismatch repair and cancer pathways (adjusted hypergeometric test P value <0.0001). These proteins also form a significant protein−protein interaction network (P < 0.0001; Figure 3D), which involves known interactions such as the coexpression between BRCA1 and CHEK2[34] and the regulation of BRCA1 by PARP1.[49]

Pan-B/D-HPP Analysis on the Well-Studied Proteins

To identify the most well-studied proteins and their relations with B/D-HPP topics, we first characterized the distribution of publication frequency for each human protein and determined the ones with the most publications in PubMed. The results showed that the number of publications generally follows the Zipf's law,[50] where the publication frequency of a keyword is inversely proportional to its rank in the publication frequency table (Figure 4A).

The 10 most published human proteins were TNF (75 731 publications), CD4 (54 739), IL6 (53 611), TP53 (46 130), INS (42 836), CRP (36 788), IFNG (34 431), VEGFA (32 733), EGFR (30 695), and IL2 (30 132). These proteins had high PURPOSE scores (more than 35) for the known B/D-HPP areas associated with them and possessed lower PURPOSE scores (approximately 10 to 15) in the less relevant fields (Figure 4B). For instance, among all 22 B/D-HPP topics, CD4 had the highest PURPOSE scores in immune-peptidome (score = 37.23), infectious diseases (score = 37.00), plasma (score = 31.77), pathology (score = 30.20), and PediOME (score = 30.09); similarly, TNF had the highest scores in glycoproteomics (score = 37.36), rheumatic disorders (score = 31.37), immune-peptidome (score = 30.18), infectious diseases (score = 28.77), and plasma (score = 28.34); and TP53 had the highest scores in cancer (score = 33.11), pathology (score = 31.43), extreme conditions (score = 29.99), stem cells (score = 27.99), and mitochondria (score = 27.53). These results showed that the PURPOSE score is highest in the areas relevant to the proteins. Areas with relatively fewer publications, such as glycoproteomics, still have high PURPOSE scores in the proteins significantly associated with them.

Identification of Proteins in Model Organisms

We extended our automated analytic platform to include common model organisms, such as mouse (*Mus musculus*), rat (*Rattus norvegicus*), common fruit fly (*Drosophila melanogaster*), roundworm (*Caenorhabditis elegans*), and yeast (*Saccharomyces cerevisiae*).[37−39] We compared the results of protein retrieval in the model organisms with those in human. Figure 5A summarized the distribution of the number of publications of each protein in these model organisms. Results showed that different species had different numbers of proteins presented in the literature and tagged by PubTator (human: 14 643; rat: 6519; mouse: 11 255, fly: 2371, roundworm: 1778; yeast: 2878), which is consistent with the fact that different species possess different numbers of genes and proteins.

The most published proteins in mouse are Tnf (32 739 publications), Cd4 (22 385), and Ifng (21 013); the most popular ones for rat are Tnf (18 872), Agt (14 480), and Ca1 (11 709); in common fruit fly, they are Egfr (679), Pc (625), and N (Notch) (619); in roundworms, they are daf-16 (677), daf-2 (416), and ced-3 (254); in yeasts, they are GAL4 (1122), GAL1 (797), and CDC28 (775).

To investigate the associations among the protein publication patterns for human, mouse, and rat, we performed a correlation analysis of the 100 most-published proteins of these species (Figure 5B). Results showed that the ranks of the most-published proteins in human, mouse, and rat are moderately correlated (Spearman correlation coefficients: human versus mouse = 0.53; human versus rat = 0.51; mouse versus rat = 0.55). Table S-3 shows a comparison of the top 20 proteins associated with cancer, the cardiovascular system, diabetes, and the liver in human, rat, and mouse. Correlations between the most-published proteins in human and common fruit fly and between human and roundworm were weak. (Spearman correlation coefficients were 0.19 and −0.02, respectively.) The negative correlation observed between human and yeast (Spearman correlation coefficients = −0.39) might be explained by the fact that many of the well-studied proteins in yeast, such as GAL1 and GAL4, are involved in the galactose metabolism pathways of biochemical interest, whereas the most widely studied proteins in human are disease-oriented, such as TNF, CD4, IL6, and TP53.

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

Evaluation of the Prioritization Performance

To quantify the performance of our PURPOSE algorithm, precision, recall, and the F1 score were calculated for four topics with curated protein−topic relations in CTD,[42] including cancer, cardiovascular, liver, and kidney. The performance of PURPOSE, a variant of PURPOSE without integrating citation count, and two other methods, GLAD4U[43] and PubPular,[22] was compared. Results showed that compared with GLAD4U and PubPular, PURPOSE algorithm had a 2.8−25.6% improvement in precision and recall and a 3.8−21.5% improvement in the F1 score (Table S-4). In addition, the variant of PURPOSE without incorporating citation counts performed slightly worse (1−3.0% decrease in precision, recall, and the F1 score) than the PURPOSE algorithm in most topics (Table S-4). These results indicated the improved performance of our methods and the value of integrating citation counts in the prioritization algorithm.

Deployment of the Automated Protein Prioritization System

To accommodate custom searches for any topic and protein of interest in real time, we deployed an automated protein prioritization system using the PURPOSE score. The system effectively summarizes the copublication patterns between proteins and any topic of interest (such as organ systems, diseases, and genetic variations), visualizes the distributions of both the PURPOSE score and the number of publications and citations, conducts enrichment and protein−protein interaction analysis on the selected proteins, and shows a summary of the highly cited publications associated with the top proteins/genes and the topic of interest. These modules direct interested users to the original publications and database entries related to their queries. The cloud-based interactive user interface is freely available for academic and nonprofit use at http://rebrand.ly/proteinpurpose.

DISCUSSION

To our knowledge, this is the first study to systematically prioritize proteins in all targeted fields of B/D-HPP. In addition, we implemented a fully automated and accessible bioinformatics tool to identify the copublication patterns of proteins and topics. We further extended the pipeline to include all human diseases, organ systems, as well as popular model organisms and deployed the system to the cloud. The cloud-based user interface enables any custom search in real time. This bioinformatics workflow can facilitate targeted proteomics studies in human and in other species.

We demonstrated that there is a great deal of existing literature associated with the 22 B/D-HPP targeted areas, which justified the use of literature mining in prioritizing proteins of interest. In addition, results revealed that the protein PURPOSE score effectively summarized the relevance of the protein to the topic of interest by incorporating the specificity of the association, the strength of copublication, and the number of citations of those publications. Our PURPOSE score extends the established works on tf-idf and quantified the strength of association between two terms, whereas the convention design of tf-idf only reflects the importance of a term to a document.[29,51,52] Among the top-ranked proteins in the B/D-HPP targeted areas, many of them were highly specific to a certain topic but did not necessarily have the most publications. For instance, the numbers of publications for IDH1 on cancer (612 publications) or DPP4 on diabetes (1,613 publications) were less than 1/10 those of many frequently published proteins on the same topic. However, our PURPOSE score was able to prioritize them due to their specificity to the topic. Pathway and network analysis results validated the pertinence of the identified protein lists of the B/D-HPP topics. Similar approaches can be applied to identify the post-translational modifications and alternative splicing isoforms[53,54] related to the topics. Protein-based searches revealed that the distribution of PURPOSE scores of a protein in various topics correlated with the importance of the protein in each topic. This strategy effectively summarized the relevance of the selected proteins to each B/D topic in the current literature.

Compared with previous literature mining tools,[22,43,55] our approach leveraged the automated and systematic PubTator tagging of each PubMed article, which helps us recover many papers not marked with the MeSH terms of proteins.[2] In addition, our methods account for the number of citations, which quantified the visibility and popularity of each research article systematically.[56] Although citation frequency could be biased, it provides a metric for attention to a particular article. The PURPOSE score is an extension of the tf-idf measurement that accounts for the popularity and specificity of the proteins simultaneously and does not require extensive renormalization after the score is computed. The cloud-based user interface we built automatically retrieves millions of PubMed publications on demand and summarizes the importance of thousands of proteins related to any disease or organ system in seconds, which informs biomedical researchers and medical practitioners of the ranked relevance of proteins to their topics of interest. The enrichment analysis and protein−protein interaction visualization modules of our cloud-based platform allow researchers to confirm the relevance of the retrieved lists, and the "View Publications" module points interested readers to the data sources of the retrieved protein−topic relations. This platform will facilitate the development of B/D-HPP and can be applied to other targeted proteomics studies.

One limitation of our approach is the potential publication bias in the current literature.[57] Our algorithm ranks proteins by their relevance found in the literature, which would prioritize the well-published and well-cited proteins (such as AFP and CYP3A4 in the liver) over the less-discussed ones (such as HSD11B1 and CYP4F2 in the liver). In extreme cases, previously undescribed protein−topic associations may not get a high prioritization score, regardless of their true biological roles. Similarly, the undescribed protein−topic associations would not be annotated in CTD or any other curated databases, which could bias the ground truth we used in evaluating PURPOSE and other tools. With more data-driven high-throughput studies being published, the publication bias could be reduced.[58] In addition, CTD only curated the associations between proteins and a number of diseases. A comprehensive data set on protein−organ and protein−disease relations is needed to assess the performance of the PURPOSE system exhaustively.

Another limitation is that misnomers of a protein or gene will strongly bias the results. For example, Brain Type Natriuretic Peptide (NPPB) is one of the top proteins in a coassociation search with brain, although it plays a more significant role in many other organ systems, including cardiovascular and immune systems, than in the brain. As such, we implemented a manual filter that excludes those instances in which the coassociations appear in the protein or gene name. Our platform also allows users to up-vote or down-vote any

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

retrieved protein−topic associations. With more voting results collected from our users, we could further incorporate the number of upvotes and downvotes into the protein prioritization score.

Overall, this study demonstrated the utility of prioritizing proteins using objective bibliographical measurements, and the results were consistent with the established knowledge on the proteins in the targeted fields of B/D-HPP. In addition, we showed the extensibility of our methods in investigating the key proteins in other organ systems, disease entities, and organisms. With the exponential growth of biomedical literature, this method will effectively summarize the current knowledge about proteins and suggest future research directions. We also note that our approach can be readily modified to identify proteins that are expressed but not readily studied in a particular target area, thus enabling researchers to devote attention to under-studied proteins. In this way, both popular and under-studied proteins can be garnered using this general approach.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00772.

Table S-1. Query terms for the 22 B/D-HPP targeted areas. Table S-2. Most frequent genetic mutations copublished with the common cancer types. Table S-3. Comparison of proteins associated with cancer, the cardiovascular system, diabetes, and the liver in human, rat, and mouse. Table S-4. Comparison of the precision, recall, and F1 score of PURPOSE, a variant of PURPOSE without integrating citation counts, GLAD4U, and PubPular. (PDF)

Data S-1. (XLSX)

Data S-2. (XLS)

■ AUTHOR INFORMATION

Corresponding Authors

*M.S.: Tel: (650) 736-8099. E-mail: mpsnyder@stanford.edu.
*I.S.K.: Tel: (617) 432-2144. E-mail: Isaac_Kohane@hms.harvard.edu.
*J.-H.C.: Tel: +886 6-2757575, ext. 62534. E-mail: jchiang@mail.ncku.edu.tw.

ORCID

Kun-Hsing Yu: 0000-0001-9892-8218
Yu-Ju Chen: 0000-0002-3178-6697

Author Contributions

K.-H.Y. designed and conducted the analysis, interpreted the results, implemented the cloud-based query system, and drafted the manuscript. T.-L.M.L. implemented the backend literature mining framework, interpreted the results, and revised the manuscript. C.-S.W. implemented a generalizable literature mining workflow. Y.-J.C., C.R., S.C.K., J.-H.C., I.S.K., and M.S. interpreted the results and revised the manuscript. M.S., I.S.K., and J.-H.C. supervised the work. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

Our cloud-based real-time search tool is freely available for academic and nonprofit use at http://rebrand.ly/proteinpurpose.

■ ACKNOWLEDGMENTS

We express appreciation to Professor Jochen Schwenk for his feedback on protein prioritization, Professor Griffin Weber for his insight into citation counts, Mr. Alex Ratner, Dr. Jared Dunnmon, Ms. Paroma Varma, Mr. Chen-Rui Liu, and Dr. Stephen Bach for their valuable advice on literature mining and suggestions on the manuscript, and Dr. Mu-Hung Tsai for pointing out the literature mining resources. We thank the anonymous reviewers for their insightful feedback. We thank the AWS Cloud Credits for Research, Microsoft Azure Research Award, and the NVIDIA GPU Grant Program for their support on the computational infrastructure. K.-H.Y. is a Harvard Data Science Fellow. This work was supported in part by grants from National Human Genome Research Institute, National Institutes of Health, grant number 5P50HG007735, National Cancer Institute, National Institutes of Health, grant number 5U24CA160036, the Defense Advanced Research Projects Agency (DARPA) Simplifying Complexity in Scientific Discovery (SIMPLEX), grant number N66001-15-C-4043, the Data-Driven Discovery of Models, contract number FA8750-17-2-0095, and the Ministry of Science and Technology Research Grant, Taiwan, grant number MOST 103-2221-E-006-254-MY2.

■ ABBREVIATIONS

B/D-HPP: biology/disease-driven human proteome project; GO: Gene ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; PURPOSE: Protein Universal Reference Publication-Originated Search Engine; tf-idf: term frequency-inverse document frequency

■ REFERENCES

(1) Jungblut, P. R.; Zimny-Arndt, U.; Zeindl-Eberhart, E.; Stulik, J.; Koupilova, K.; Pleissner, K. P.; Otto, A.; Muller, E. C.; Sokolowska-Kohler, W.; Grabher, G.; Stoffler, G. Proteomics in human disease: cancer, heart and infectious diseases. *Electrophoresis* **1999**, *20* (10), 2100−10.

(2) Wei, C. H.; Kao, H. Y.; Lu, Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* **2013**, *41* (W1), W518−W522.

(3) Mertins, P.; Mani, D. R.; Ruggles, K. V.; Gillette, M. A.; Clauser, K. R.; Wang, P.; Wang, X.; Qiao, J. W.; Cao, S.; Petralia, F.; Kawaler, E.; Mundt, F.; Krug, K.; Tu, Z.; Lei, J. T.; Gatza, M. L.; Wilkerson, M.; Perou, C. M.; Yellapantula, V.; Huang, K. L.; Lin, C.; McLellan, M. D.; Yan, P.; Davies, S. R.; Townsend, R. R.; Skates, S. J.; Wang, J.; Zhang, B.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Ding, L.; Paulovich, A. G.; Fenyo, D.; Ellis, M. J.; Carr, S. A. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **2016**, *534* (7605), 55−62.

(4) Zhang, B.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M. C.; Zimmerman, L. J.; Shaddox, K. F.; Kim, S.; et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **2014**, *513* (7518), 382−7.

(5) Zhang, H.; Liu, T.; Zhang, Z.; Payne, S. H.; Zhang, B.; McDermott, J. E.; Zhou, J. Y.; Petyuk, V. A.; Chen, L.; Ray, D.; et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **2016**, *166* (3), 755−765.

(6) Yu, K. H.; Berry, G. J.; Rubin, D. L.; Re, C.; Altman, R. B.; Snyder, M. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma. *Cell Syst* **2017**, *5* (6), 620−627.e3.

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

(7) Chen, R.; Mias, G. I.; Li-Pook-Than, J.; Jiang, L.; Lam, H. Y.; Chen, R.; Miriami, E.; Karczewski, K. J.; Hariharan, M.; Dewey, F. E.; Cheng, Y.; Clark, M. J.; Im, H.; Habegger, L.; Balasubramanian, S.; O'Huallachain, M.; Dudley, J. T.; Hillenmeyer, S.; Haraksingh, R.; Sharon, D.; Euskirchen, G.; Lacroute, P.; Bettinger, K.; Boyle, A. P.; Kasowski, M.; Grubert, F.; Seki, S.; Garcia, M.; Whirl-Carrillo, M.; Gallardo, M.; Blasco, M. A.; Greenberg, P. L.; Snyder, P.; Klein, T. E.; Altman, R. B.; Butte, A. J.; Ashley, E. A.; Gerstein, M.; Nadeau, K. C.; Tang, H.; Snyder, M. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **2012**, *148* (6), 1293−307.

(8) Yu, K. H.; Levine, D. A.; Zhang, H.; Chan, D. W.; Zhang, Z.; Snyder, M. Predicting Ovarian Cancer Patients' Clinical Response to Platinum-Based Chemotherapy by Their Tumor Proteomic Signatures. *J. Proteome Res.* **2016**, *15* (8), 2455−65.

(9) Oda, Y.; Owa, T.; Sato, T.; Boucher, B.; Daniels, S.; Yamanaka, H.; Shinohara, Y.; Yokoi, A.; Kuromitsu, J.; Nagasu, T. Quantitative chemical proteomics for identifying candidate drug targets. *Anal. Chem.* **2003**, *75* (9), 2159−65.

(10) Collins, F. S.; Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **2015**, *372* (9), 793−5.

(11) Ashley, E. A. The precision medicine initiative: a new national effort. *JAMA* **2015**, *313* (21), 2119−20.

(12) Yu, K. H.; Fitzpatrick, M. R.; Pappas, L.; Chan, W.; Kung, J.; Snyder, M. Omics AnalySIs System for PRecision Oncology (OASISPRO): A Web-based Omics Analysis Tool for Clinical Phenotype Prediction. *Bioinformatics* **2018**, *34* (2), 319−320.

(13) Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J.; Kussman, M.; Qin, J.; Omenn, G. S. Highlights of B/D-HPP and HPP Resource Pillar Workshops at 12th Annual HUPO World Congress of Proteomics: September 14−18, 2013, Yokohama, Japan. *Proteomics* **2014**, *14* (9), 975−88.

(14) Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J. Proteome Res.* **2013**, *12* (1), 23−7.

(15) Van Eyk, J. E.; Corrales, F. J.; Aebersold, R.; Cerciello, F.; Deutsch, E. W.; Roncada, P.; Sanchez, J. C.; Yamamoto, T.; Yang, P.; Zhang, H.; Omenn, G. S. Highlights of the Biology and Disease-driven Human Proteome Project, 2015−2016. *J. Proteome Res.* **2016**, *15* (11), 3979−3987.

(16) U.S. National Library of Medicine. *Key MEDLINE Indicators*. https://www.nlm.nih.gov/bsd/bsd_key.html (accessed September 24, 2018).

(17) Corlan, A. D. *Medline Trend: Automated Yearly Statistics of PubMed. Results for Any Query*, 2004. http://dan.corlan.net/medline-trend.html.

(18) Lam, M. P.; Venkatraman, V.; Cao, Q.; Wang, D.; Dincer, T. U.; Lau, E.; Su, A. I.; Xing, Y.; Ge, J.; Ping, P.; Van Eyk, J. E. Prioritizing Proteomics Assay Development for Clinical Translation. *J. Am. Coll. Cardiol.* **2015**, *66* (2), 202−4.

(19) Huang, M.; Zhu, X.; Hao, Y.; Payan, D. G.; Qu, K.; Li, M. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics* **2004**, *20* (18), 3604−12.

(20) Cheng, D.; Knox, C.; Young, N.; Stothard, P.; Damaraju, S.; Wishart, D. S. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.* **2008**, *36* (Web Server issue), W399−W405.

(21) Percha, B.; Garten, Y.; Altman, R. B. Discovery and explanation of drug-drug interactions via text mining. *Pac Symp. Biocomput* **2012**, 410−21.

(22) Lam, M. P.; Venkatraman, V.; Xing, Y.; Lau, E.; Cao, Q.; Ng, D. C.; Su, A. I.; Ge, J.; Van Eyk, J. E.; Ping, P. Data-Driven Approach To Determine Popular Proteins for Targeted Proteomics Translation of Six Organ Systems. *J. Proteome Res.* **2016**, *15* (11), 4126−4134.

(23) Kusebauch, U.; Campbell, D. S.; Deutsch, E. W.; Chu, C. S.; Spicer, D. A.; Brusniak, M. Y.; Slagel, J.; Sun, Z.; Stevens, J.; Grimes, B.; Shteynberg, D.; Hoopmann, M. R.; Blattmann, P.; Ratushny, A. V.; Rinner, O.; Picotti, P.; Carapito, C.; Huang, C. Y.; Kapousouz, M.; Lam, H.; Tran, T.; Demir, E.; Aitchison, J. D.; Sander, C.; Hood, L.; Aebersold, R.; Moritz, R. L. Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell* **2016**, *166* (3), 766−778.

(24) The Human Proteome Organization. *Biology/Disease-driven HPP*. https://www.hupo.org/B/D-HPP (accessed September 24, 2017).

(25) Widmaier, E. P.; Raff, H.; Strang, K. T. *Vander's Human Physiology*; McGraw-Hill: New York, 2006; Vol. 5.

(26) Denny, J. C.; Ritchie, M. D.; Basford, M. A.; Pulley, J. M.; Bastarache, L.; Brown-Gentry, K.; Wang, D.; Masys, D. R.; Roden, D. M.; Crawford, D. C. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **2010**, *26* (9), 1205−10.

(27) Maglott, D.; Ostell, J.; Pruitt, K. D.; Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **2011**, *39* (Database), D52−D57.

(28) Sayers, E. The E-utilities In-Depth: Parameters, Syntax and More. In *Entrez Programming Utilities Help*; National Center for Biotechnology Information (US): Bethesda, MD, 2009. http://www.ncbi.nlm.nih.gov/books/NBK25499.

(29) Leskovec, J.; Rajaraman, A.; Ullman, J. D. *Mining of Massive Datasets*; Cambridge University Press: 2014.

(30) Torre, L. A.; Bray, F.; Siegel, R. L.; Ferlay, J.; Lortet-Tieulent, J.; Jemal, A. Global cancer statistics, 2012. *Ca-Cancer J. Clin.* **2015**, *65* (2), 87−108.

(31) Zehir, A.; Benayed, R.; Shah, R. H.; Syed, A.; Middha, S.; Kim, H. R.; Srinivasan, P.; Gao, J.; Chakravarty, D.; Devlin, S. M.; Hellmann, M. D.; Barron, D. A.; Schram, A. M.; Hameed, M.; Dogan, S.; Ross, D. S.; Hechtman, J. F.; DeLair, D. F.; Yao, J.; Mandelker, D. L.; Cheng, D. T.; Chandramohan, R.; Mohanty, A. S.; Ptashkin, R. N.; Jayakumaran, G.; Prasad, M.; Syed, M. H.; Rema, A. B.; Liu, Z. Y.; Nafa, K.; Borsu, L.; Sadowska, J.; Casanova, J.; Bacares, R.; Kiecka, I. J.; Razumova, A.; Son, J. B.; Stewart, L.; Baldi, T.; Mullaney, K. A.; Al-Ahmadie, H.; Vakiani, E.; Abeshouse, A. A.; Penson, A. V.; Jonsson, P.; Camacho, N.; Chang, M. T.; Won, H. H.; Gross, B. E.; Kundra, R.; Heins, Z. J.; Chen, H. W.; Phillips, S.; Zhang, H.; Wang, J.; Ochoa, A.; Wills, J.; Eubank, M.; Thomas, S. B.; Gardos, S. M.; Reales, D. N.; Galle, J.; Durany, R.; Cambria, R.; Abida, W.; Cercek, A.; Feldman, D. R.; Gounder, M. M.; Hakimi, A. A.; Harding, J. J.; Iyer, G.; Janjigian, Y. Y.; Jordan, E. J.; Kelly, C. M.; Lowery, M. A.; Morris, L. G. T.; Omuro, A. M.; Raj, N.; Razavi, P.; Shoushtari, A. N.; Shukla, N.; Soumerai, T. E.; Varghese, A. M.; Yaeger, R.; Coleman, J.; Bochner, B.; Riely, G. J.; Saltz, L. B.; Scher, H. I.; Sabbatini, P. J.; Robson, M. E.; Klimstra, D. S.; Taylor, B. S.; Baselga, J.; Schultz, N.; Hyman, D. M.; Arcila, M. E.; Solit, D. B.; Ladanyi, M.; Berger, M. F. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **2017**, *23* (6), 703−713.

(32) Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32* (Database), D258−D261.

(33) Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28* (1), 27−30.

(34) Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; Kuhn, M.; Bork, P.; Jensen, L. J.; von Mering, C. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43* (D1), D447−D452.

(35) Gaudet, P.; Michel, P. A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; Rech de Laval, V.; Lin, J.; Nikitin, F.; Schaeffer, M.; Teixeira, D.; Lane, L.; Bairoch, A. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **2017**, *45* (D1), D177−D182.

(36) The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45* (D1), D158−D169.

(37) Muller, B.; Grossniklaus, U. Model organisms−A historical perspective. *J. Proteomics* **2010**, *73* (11), 2054−63.

(38) Krijgsveld, J.; Ketting, R. F.; Mahmoudi, T.; Johansen, J.; Artal-Sanz, M.; Verrijzer, C. P.; Plasterk, R. H.; Heck, A. J. Metabolic

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX

labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nat. Biotechnol.* **2003**, *21* (8), 927−31.
(39) Gouw, J. W.; Krijgsveld, J.; Heck, A. J. Quantitative proteomics by metabolic labeling of model organisms. *Mol. Cell. Proteomics* **2010**, *9* (1), 11−24.
(40) Mi, H.; Muruganujan, A.; Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **2012**, *41* (D1), D377−D386.
(41) Thomas, P. D.; Campbell, M. J.; Kejariwal, A.; Mi, H.; Karlak, B.; Daverman, R.; Diemer, K.; Muruganujan, A.; Narechania, A. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **2003**, *13* (9), 2129−2141.
(42) Davis, A. P.; Grondin, C. J.; Johnson, R. J.; Sciaky, D.; King, B. L.; McMorran, R.; Wiegers, J.; Wiegers, T. C.; Mattingly, C. J. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D972−D978.
(43) Jourquin, J.; Duncan, D.; Shi, Z.; Zhang, B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics* **2012**, *13* (Suppl8), S20.
(44) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4* (1), 44−57.
(45) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **2009**, *37* (1), 1−13.
(46) Croft, D.; O'Kelly, G.; Wu, G.; Haw, R.; Gillespie, M.; Matthews, L.; Caudy, M.; Garapati, P.; Gopinath, G.; Jassal, B.; Jupe, S.; Kalatskaya, I.; Mahajan, S.; May, B.; Ndegwa, N.; Schmidt, E.; Shamovsky, V.; Yung, C.; Birney, E.; Hermjakob, H.; D'Eustachio, P.; Stein, L. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **2011**, *39* (Database), D691−D697.
(47) Hart, T.; Chandrashekhar, M.; Aregger, M.; Steinhart, Z.; Brown, K. R.; MacLeod, G.; Mis, M.; Zimmermann, M.; Fradet-Turcotte, A.; Sun, S.; Mero, P.; Dirks, P.; Sidhu, S.; Roth, F. P.; Rissland, O. S.; Durocher, D.; Angers, S.; Moffat, J. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **2015**, *163* (6), 1515−26.
(48) Holland, E. C.; Hively, W. P.; DePinho, R. A.; Varmus, H. E. A constitutively active epidermal growth factor receptor cooperates with disruption of G1 cell-cycle arrest pathways to induce glioma-like lesions in mice. *Genes Dev.* **1998**, *12* (23), 3675−85.
(49) Hu, Y.; Petit, S. A.; Ficarro, S. B.; Toomire, K. J.; Xie, A.; Lim, E.; Cao, S. A.; Park, E.; Eck, M. J.; Scully, R.; Brown, M.; Marto, J. A.; Livingston, D. M. PARP1-driven poly-ADP-ribosylation regulates BRCA1 function in homologous recombination-mediated DNA repair. *Cancer Discovery* **2014**, *4* (12), 1430−47.
(50) Li, W. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Trans. Inf. Theory* **1992**, *38* (6), 1842−1845.
(51) Aizawa, A. An information-theoretic perspective of tf−idf measures. *Inf. Process. Manage.* **2003**, *39* (1), 45−65.
(52) Ramos, J. Using TF-IDF to determine word relevance in document queries. *Proceedings of the First instructional Conference on Machine Learning* **2003**, *2003*, 133−142.
(53) Li, H. D.; Omenn, G. S.; Guan, Y. A proteogenomic approach to understand splice isoform functions through sequence and expression-based computational modeling. *Briefings Bioinf.* **2016**, *17* (6), bbv109.
(54) Li, H. D.; Menon, R.; Omenn, G. S.; Guan, Y. The emerging era of genomic data integration for analyzing splice isoform function. *Trends Genet.* **2014**, *30* (8), 340−7.
(55) Tsuruoka, Y.; Tsujii, J.; Ananiadou, S. FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics* **2008**, *24* (21), 2559−60.
(56) De Bellis, N. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*; Scarecrow Press: 2009.
(57) Dickersin, K. The existence of publication bias and risk factors for its occurrence. *JAMA* **1990**, *263* (10), 1385−9.
(58) Becker, K. G.; Barnes, K. C.; Bright, T. J.; Wang, S. A. The genetic association database. *Nat. Genet.* **2004**, *36* (5), 431−2.

DOI: 10.1021/acs.jproteome.7b00772
J. Proteome Res. XXXX, XXX, XXX−XXX