Supporting Information

Identifying high-priority proteins across the human diseasome using semantic similarity

Edward Lau[1], Vidya Venkatraman[2], Cody T Thomas[3], Joseph C Wu[1], Jennifer E Van Eyk[2,*], Maggie PY Lam[3,*]

1 Stanford Cardiovascular Institute, Stanford University, Palo Alto, CA.

2 Advanced Clinical Biosystems Research Institute, Department of Medicine and The Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA.

3 Department of Medicine, Division of Cardiology, Consortium for Fibrosis Research and Translation, Anschutz Medical Campus, University of Colorado Denver, CO.

*** Correspondence**

Jennifer E Van Eyk, PhD

Cedars-Sinai Medical Center

Advanced Health Sciences Pavilion, 9th Floor

127 S. San Vicente Boulevard

Los Angeles, CA 90048, USA

Email: Jennifer.VanEyk@cshs.org

Maggie Pui Yu Lam, PhD

University of Colorado Denver - Anschutz Medical Campus

Mail Stop B139, Research Complex 2

12700 E. 19th Avenue

Aurora, CO 80045, USA

Email: maggie.lam@ucdenver.edu

Contents

1 Supplementary Figure S1 **S-3**

2 Supplementary Figure S2 **S-4**

3 Supplementary Figure S3 **S-5**

4 Supplementary Figure S4 **S-6**

5 Supplementary Data 1 **S-7**

1 Supplementary Figure S1

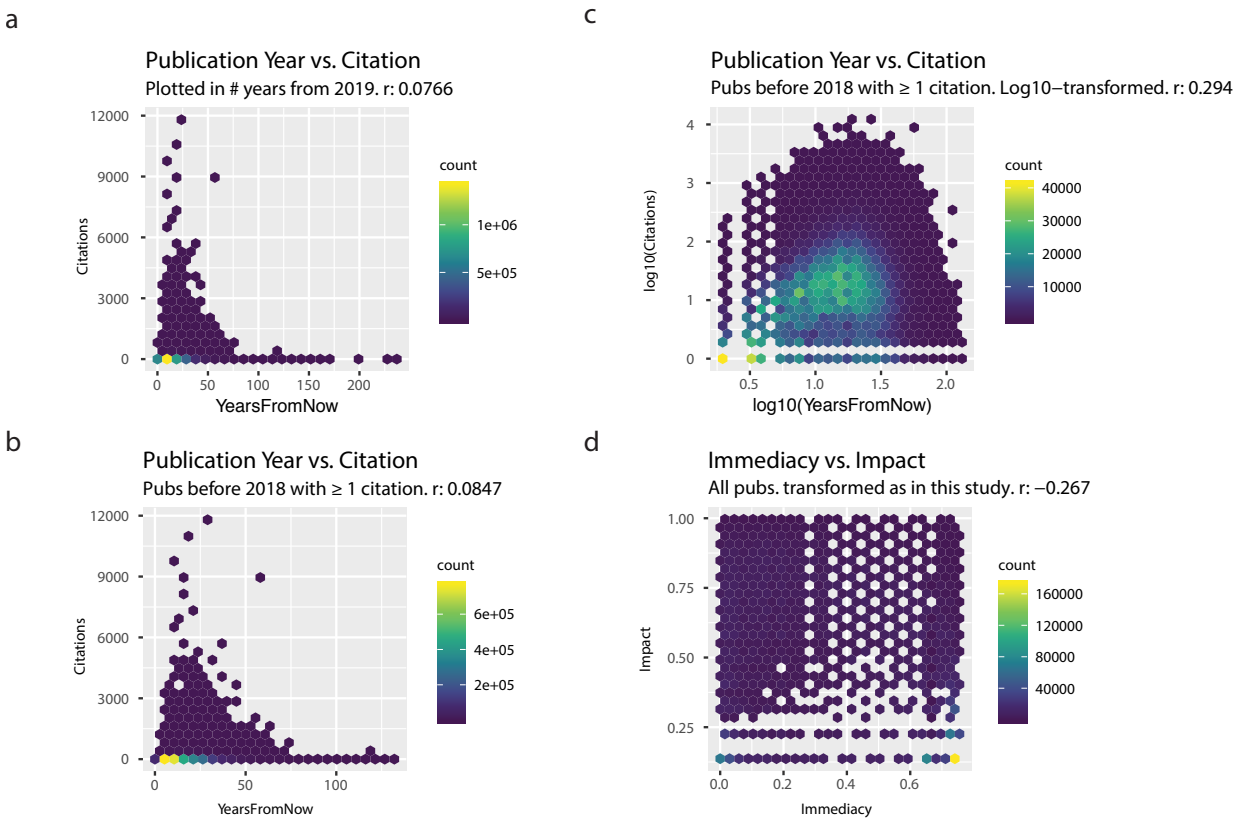

Figure S1. Correlation between immediacy and impact values. **a** Scatterplot showing distribution of publication year (x-axis, plotted as number of years counted backward from 2019) vs. number of citations from 3,565,789 annotated publications. **b** As in panel a, but including only 2,976,187 annotated publications published up to 2017 and cited at least once. **c** As in panel b, but with log10-transformed values. **d** Scatterplot showing distribution of immediacy and impact values of 3,565.789 annotated publications as calculated in the Methods section.

2 Supplementary Figure S2

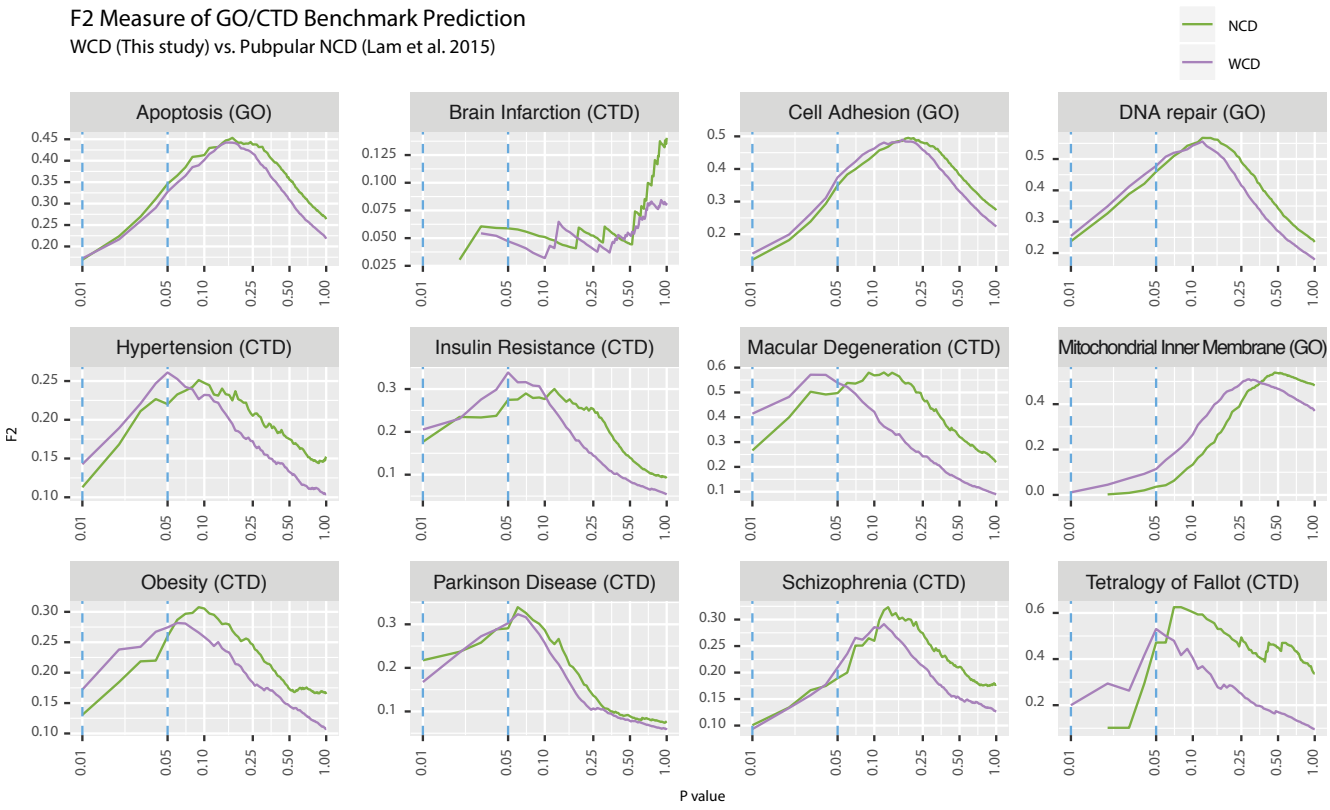

Figure S2. Comparison of WCD and NCD against benchmark gene/protein lists. The F_2 measure is selected as a metric to compare the recall and precision between weighted co-publication distance (WCD) (This study) vs. unadjusted normalized co-publication distance (NCD) methods[2] against curated benchmark protein lists retrieved from Gene Ontology (GO) and the Comparative Toxicogenomics Database (CTD). Calculations were carried out using identical PubMed query results and annotation data. WCD (purple lines) yielded greater F_2 values than unweighted NCD (green lines) for 10 of 12 terms at $P \leq 0.05$ and for 9 of 12 terms at $P \leq 0.01$.

3 Supplementary Figure S3

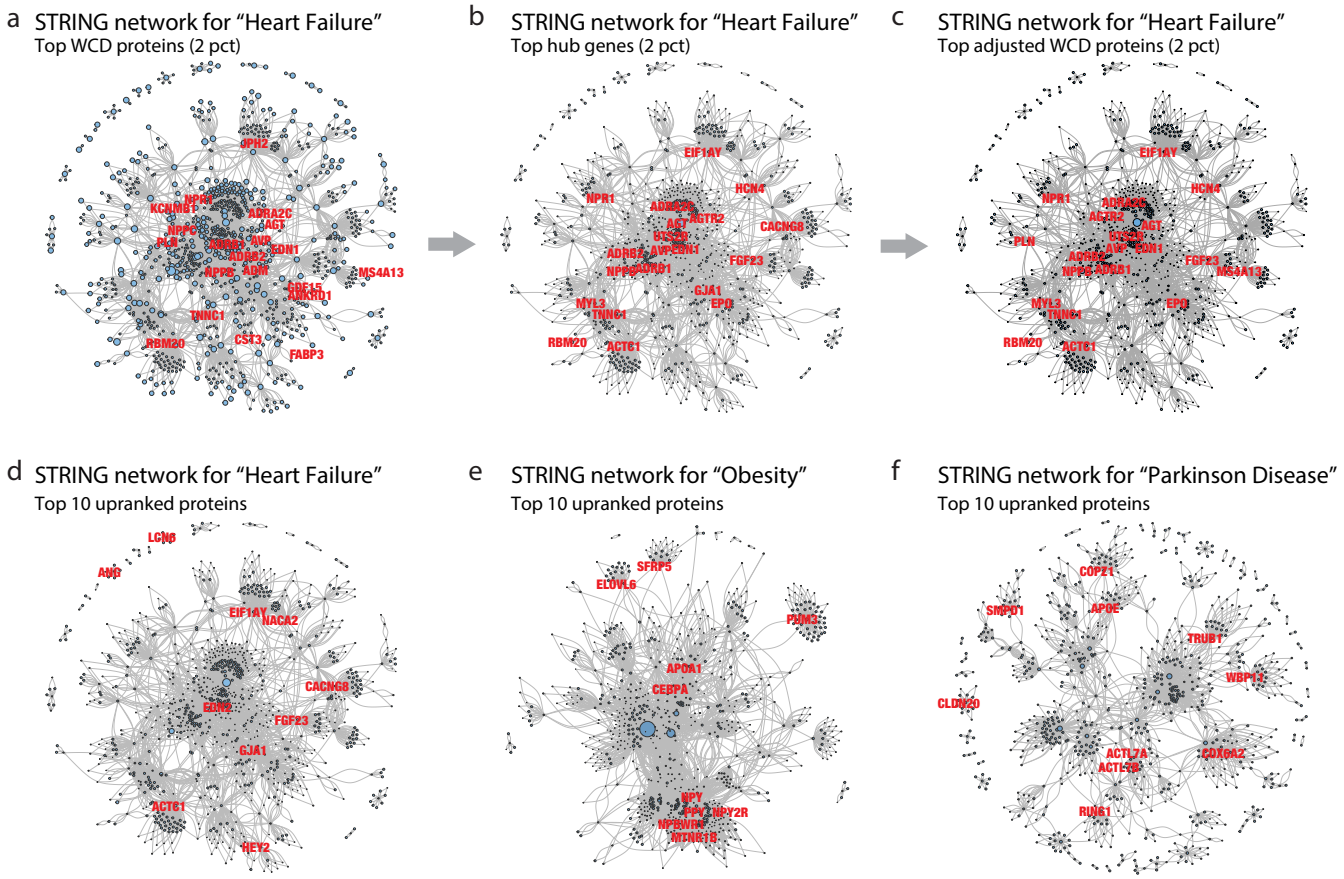

Figure S3. Identifying under-studied proteins by popularity overlaid on protein association graphs. **a** Protein-protein association networks from STRING db (STRING score ≥ 500) involving proteins with at least one publication in "Heart Failure" query. Vertex sizes scale with WCD; top 2 percentile of proteins with highest popularity are labeled. **b** Hub genes/proteins in the network are labeled. **c** Proteins are re-scored using the PageRank algorithm. Top 2 percentile of popular proteins following reranking are labeled. **d** After reranking the popular protein lists, the top 10 proteins that gained the most in ranking are labeled. **e** As in panel d, but for analysis on the query term "Obesity". **f** As in panel e, but for analysis on the query term "Parkinson Disease".

4 Supplementary Figure S4

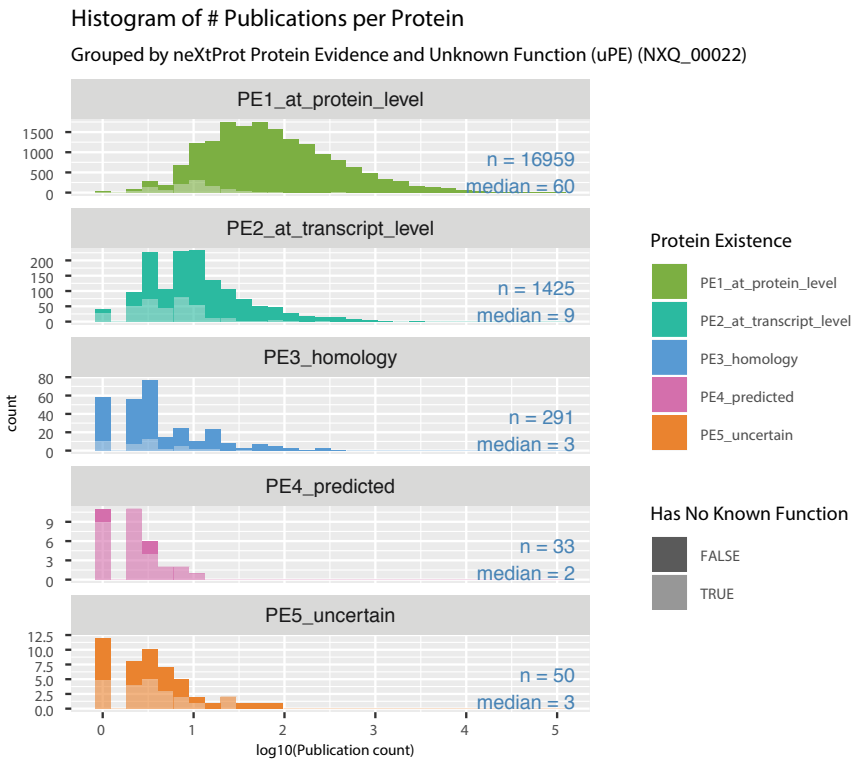

Figure S4. Number of associated publications per protein across protein evidence and functional categories. Proteins are grouped by neXtProt[31] Protein Evidence (PE) levels (PE1 to PE5 fill color) as well as whether the protein has no known function (transparency). Protein function was queried via SPARQL NXQ_00022. PE1 proteins (known to be expressed at protein levels) are associated with more publications (median 60 publications per protein) than PE2-5 proteins (median publications 2-9 per protein). PE1 proteins with unknown function (uPE1) have similar publication distribution as PE2-5 proteins.

5 Supplementary Data 1

Popular Proteins in the Human Diseasome. Collection of popular proteins across 10,129 human diseases as defined by the Disease Ontology, 10,642 disease phenotypes defined by Human Phenotype Ontology, and 2,370 cellular pathways defined by Pathway Ontology. Accessible on figshare at https://doi.org/10.6084/m9.figshare.6378485.v2.