Supplementary Information

Integrative interactome analysis for mechanistic discovery and drug repurposing in rare diseases: GPX4-related rare skeletal dysplasia

Kalyani B. Karunakaran[1], N. Balakrishnan[2] and Madhavi K. Ganapathiraju[3]*

[1] Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore, 560012; kalyanithepebble@gmail.com
[2] Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore, 560012; balki@iisc.ac.in
[3] Department of Biomedical Informatics, School of Medicine, and Intelligent Systems Program, School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA; madhavi@pitt.edu
* Correspondence: madhavi@pitt.edu

Detailed Methods

Compilation of gene lists and prediction of novel interactions

SMDS-associated genes, genes associated with X-linked spondyloepimetaphyseal dysplasia, acrocapitofemoral dysplasia, Dyggve-Melchior-Clausen disease, autosomal recessive Megarbane type SMD and genes associated with Joubert syndrome with Jeune asphyxiating thoracic dystrophy (JATD) were compiled from the DisGeNET database (Piñero et al., 2016). Genes associated with Kozlowski type SMD, spondyloenchondrodysplasia, odontochondrodysplasia, Sutcliffe/corner fractures type SMD, SMD with severe genu valgam/Schmidt type SMD, SMD with cone-rod dystrophy, SMD with retinal degeneration/axial SMD, dysspondyloenchondromatosis, achondrogenesis type 1A, schneckenbecken dysplasia and opsismodysplasia were curated from Warman et al. (Warman et al., 2011). Genes whose perturbation (gene knockout/knockdown/mutation/overexpression) affects GPX4 were collected from Knockdown Atlas (BaseSpace Correlation Engine software suite (Kupershmidt et al., 2010a)).

Protein-protein interactions (PPIs) involving these genes that have been reported in the literature ('known PPIs') were obtained from HPRD (Human Protein Reference Database (Keshava Prasad et al., 2008)) and BioGRID (Biological General Repository for Interaction Datasets (Stark et al., 2006)). To systematically expand the molecular context of GPX4 and SMDS-associated genes, computationally predicted PPIs were incorporated using the HiPPIP model, which we previously developed and validated for high precision through computational evaluations and experimental validations (Zhu et al., 2014; Ganapathiraju et al., 2016c; Karunakaran et al., 2021). Each protein (denoted N_1) was paired with each of the other human proteins say, (M_1, M_2,…M_n), and each pair (N_i, M_j) was evaluated using the HiPPIP model (Ganapathiraju et al., 2016a). Protein pairs not previously reported to interact but predicted as interacting by HiPPIP were treated as candidate novel PPIs. All PPI networks were visualized using Cytoscape (Shannon et al., 2003).

Network analysis using LENS

Lens for Enrichment and Network Studies of human proteins (LENS) was used to extract shortest paths in the human interactome connecting the gene sets compiled in this study to GPX4. LENS is a web-based tool designed to identify pathways, diseases, drugs and genome-wide association study annotations significantly enriched among user-supplied gene sets (Handen and Ganapathiraju, 2015). LENS identifies nearest neighbors of each gene in the interactome and the intermediate interactions that connect them via shortest paths. It then computes the statistical significance of the overlap of genes in the network and genes with annotations pertaining to pathways, diseases, drugs and genome-wide association studies, and reports a p-value computed using Fisher's exact test.

Identification of functional modules

Functional gene modules were identified using the HumanBase toolkit (Krishnan et al., 2016) (https://hb.flatironinstitute.org/). Under the "Analyses" interface, the "Functional Module Detection" option was selected, and the relevant gene lists were submitted. Network context was specified as either "global" or "fetus" to detect tissue-naïve and fetus-specific functional modules. The minimum module size was set to the default of eight nodes, and the FDR-corrected p-values (q-values) for enriched Gene Ontology (GO) biological process terms were recorded. HumanBase applies shared k-nearest-neighbors and the Louvain community-finding algorithm to cluster genes with similar network neighborhoods and GO annotations, as described previously (Krishnan et al., 2016). Enrichment p-values were calculated using Fisher's exact test and corrected for multiple hypotheses using the Benjamini–Hochberg procedure.

Pathophenotype enrichment using MONARCH

Genes associated with twelve major SMDS-associated pathophenotypes were compiled from the MONARCH database (Cacheiro et al., 2019). Enrichment of these phenotypes within the interactomes was evaluated against a background of 15,451 genes with phenotype annotations. The analyzed phenotypes included atrial septal defect (290 genes), atrioventricular block (54 genes), cardiorespiratory arrest (10 genes), cerebellar hypoplasia (219 genes), myocarditis (7 genes), agenesis of corpus callosum (240 genes), arrhythmia (355 genes), pachygyria (133 genes), metaphyseal cupping (18 genes), platyspondyly (109 genes), abnormal scapula morphology (4 genes) and abnormality of the ribs (85 genes).

Gene expression analysis

Gene expression profiles across 53 postnatal human tissues were extracted from GTEx (Consortium, 2015). Principal component analysis (PCA) and hierarchical clustering were performed to characterize expression-based relationships among genes within the constructed networks. Log-transformed transcripts per million (TPM) values were assembled into a matrix with genes as rows and tissues as columns. PCA was performed using ClustVis (Metsalu and Vilo, 2015). The matrix was preprocessed to allow up to 70% missing values across rows and columns. Log(TPM) values were standardized across genes to zero mean and unit variance. Singular value decomposition (SVD) with imputation was applied to estimate missing values iteratively until convergence (Metsalu and Vilo, 2015). Hierarchical clustering was conducted using Heatmapper (Babicki et al., 2016). Pairwise distances were computed using Pearson correlation, and clusters were formed using the average linkage method. Dendrograms were generated by merging tissues with minimal pairwise distances.

Identification of repurposable drugs

Drug repurposing candidates were identified using two complementary computational strategies. First, chemical compounds whose gene expression profiles were negatively correlated with four dysplasia expression datasets were compiled using the BaseSpace correlation software (https://www.nextbio.com) following a previously established protocol (Chattopadhyay and Ganapathiraju, 2017) (list 1). The analyzed datasets included tibial growth plate hypertrophic zone from Cog mice (chondroplasia) versus wildtype littermates (GSE30628 (Cameron et al., 2011)), tibial growth plate hypertrophic zone profiles from Schmid mice (chondroplasia) versus wildtype littermates (GSE30628 (Cameron et al., 2011)), skin fibroblast profiles from Schimke immuno-osseous dysplasia cell line SD60 versus healthy control (GSE35551 (Baradaran-Heravi et al., 2012)), and skin fibroblasts from Schimke immuno-osseous dysplasia cell line SD8 versus healthy control (GSE35551 (Baradaran-Heravi et al., 2012)). Second, drugs targeting at least one gene in the interactome of SMDS-associated genes (GPX4, AGRP, ARNTL, ARTN, LOH19CR1, PSD4 and RPS19) were identified using Drug Bank (list 2) (Wishart et al., 2006). Drugs appearing in both lists, i.e., those targeting proteins in the SMDS interactome and exhibiting inverse transcriptional signatures relative to dysplasia-associated expression profiles, were prioritized as candidate repurposable therapies.