ConceptGen: A gene set enrichment and gene set relation mapping tool

Concept Building

Gene Ontology, KEGG pathway, Biocarta Pathway, Panther Pathway, and Pfam information were downloaded from their respective sources**.** For Pfam, Uniprot IDs were mapped to Entrez Gene IDs, thus resulting in all transcripts for a gene being mapped to the same Pfam terms. Chromosomal location was determined by NCBI cytoband assignment, and gene expression signatures were defined as detailed below. Testing cytobands allows one to identify differential expression clustered in a particular location, possibly indicating copy number change in that cytoband. Testing gene expression signatures enables one to identify combinations of processes/pathways that are not will characterized by any other concept type. The additional concept types were built as detailed here.

Literature-derived concepts: Gene2MeSH (http://gene2mesh.ncibi.org) identifies gene-MeSH term pairs by testing whether the number of co-occurrences of each pair in the literature is significantly more than expected at random. MeSH terms were generated from Gene2MeSH using a p<0.001 cutoff for gene-MeSH term pairs, and were manually curated to remove uninformative terms, such as continents and experimental method protocols. Testing against MeSH allows one to generate hypotheses related to diseases, processes, other genes, confounders such as populations and experimental techniques, etc. based on knowledge from the literature that may not yet be formally described in any other concept type.

Human diseases: Online Mendelian Inheritance in Man® (OMIM®) was downloaded and manually curated to form unique concept names. Testing against OMIM allows one to identify relationships in their data with known genetic changes in disease.

Drug Targets: "DrugCard" flat files were downloaded directly from DrugBank and parsed by drug target (Wishart, et al., 2008). Testing against DrugBank allows identification of drugs and compounds that bind to a significant number of input genes.

Transcription factor targets: To build the transcription factor targets concepts, KnownGene, KnownToLocusLink, and TfbsConsSites tables were obtained from UCSC Genome browser (Mar. 2006, NCBI36). For each known gene, the Entrez Gene ID (formerly known as Locus Link ID) is assigned using the KnownToLocusLink table, and the list of transcription factors that bind to a gene promoter region (±2,000 bp of TSSs) was generated using minimal overlap.

Protein-interactions: MiMI concepts are defined by the protein-protein interactions centered at a specified gene. The protein interactions were formed by merging multiple protein-interaction data sources based on all known molecule identifiers (Tarcea, et al., 2008). Testing against MiMI protein interactions allows one to generate hypotheses about hub proteins that may be activated/de-activated in ways other than differential expression (or other than what was measured to create the user-input set).

MicroRNA targets: For miRNA concepts, the TargetScanS table containing 54,199 conserved miRNA target sites in human Refseq genes predicted by TargetScanHuman5.1 was obtained from the UCSC genome browser, as well as the target sites predicted by miRanda using the newly improvised mirSVR algorithm from microRNA.org. The later source has predicted over 16 million miRNA target sites in 34,911 distinct 3'UTR in human, and for downstream analysis, 3,155,472 non-conserved and 1,047,672 conserved miRNA sites (totalling approximately 4 million sites) with good mirSVR scores to be considered further. To provide high-quality miRNA targets, only the target sites predicted by both targetScan and miRanda algorithms within 3'UTR regions of the human genome were included in the LRpath database, covering 36,015 sites with 153 different miRNAs. The other concepts were created as described previously for ConceptGen software (Sartor et al., 2010) (Table 2).

Metabolite-centered concepts: Metabolite concepts were defined using Edinburgh Human Metabolic Network database (Ma, et al., 2007). Each metabolite concept is comprised of genes encoding metabolic enzymes that catalyze reactions involving the respective metabolite. Testing against the Metabolite concept type enables one to generate hypotheses about what compounds may be driving the changes observed in the user-input set or that may be affected by the observed changes.

Gene Expression analysis: In order to define expression-based concepts, we developed a gene expression analysis pipeline that uses a carefully chosen, statistical method for each step. The gene expression concept type is populated with human Affymetrix experiments in Gene Expression Omnibus (GEO). The analysis pipeline performs the following tasks:

1. Downloads the raw Affymetrix CEL files from GEO and the relevant experimental design information.

2. Using R, probes are mapped to probe sets based on the relevant Entrez ID centered CDF package (Dai, et al., 2005), which provides a unique probe set for each Entrez ID.

3. Data are pre-processed and normalized using RMA (Robust Multi-array Average).

4. Quality control output from AffyQCReport is manually observed, and only data passing our quality control standards proceed

5. Based on automatically extracted sample names provided by the author and other GEO dataset annotation, comparisons are set up manually through a pipeline interface.

6. Differential expression is tested using an empirical Bayes, intensity-based moderated t-test (IBMT) (Sartor, et al., 2006), which provides better estimates of variance and improved ranking of significant genes compared to a standard t-test, especially for experiments of small sample size

7. Gene sets (concepts) are defined by the top ranked genes, ranked by p-value, using the criteria fold change > 20% and p-value < 0.05, and limited to no more than 1000 genes. The limitation of 1000 genes avoids situations having several thousands of differentially expressed genes, taking only the most significantly differentially expressed, and limits the difference in power between very large and small gene sets. *Note*: We do not use the adjusted p-value for concept creation because, as opposed to identifying individual genes as significantly differentially expressed (when adjusting the p-values is necessary), identifying enriched concepts is improved by using a more relaxed significance cut-off (Sartor, et al., 2009).

References

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., Watson, S.J. and Meng, F. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data, *Nucleic Acids Res*, **33**, e175.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. and Enright, A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature, *Nucleic Acids Res*, **34**, D140-144.

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O. and Goryanin, I. (2007) The Edinburgh human metabolic network reconstruction and its functional analysis, *Mol Syst Biol*, **3**, 135.

Sartor, M.A., Leikauf, G.D. and Medvedovic, M. (2009) LRpath: A logistic regression approach for identifying enriched biological groups in gene expression data, *Bioinformatics*.

Sartor, M.A., Tomlinson, C.R., Wesselkamper, S.C., Sivaganesan, S., Leikauf, G.D. and Medvedovic, M. (2006) Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments, *BMC.Bioinformatics.*, **7**, 538.

Sreekumar, A., Poisson, L.M., Rajendiran, T.M., Khan, A.P., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R.J., Li, Y., Nyati, M.K., Ahsan, A., Kalyana-Sundaram, S., Han, B., Cao, X., Byun, J., Omenn, G.S., Ghosh, D., Pennathur, S., Alexander, D.C., Berger, A., Shuster, J.R., Wei, J.T., Varambally, S., Beecher, C. and Chinnaiyan, A.M. (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression, *Nature*, **457**, 910-914.

Tarcea, V.G., Weymouth, T., Ade, A., Bookvich, A., Gao, J., Mahavisno, V., Wright, Z., Chapman, A., Jayapandian, M., Ozgur, A., Tian, Y., Cavalcoli, J., Mirel, B., Patel, J., Radev, D., Athey, B., States, D. and Jagadish, H.V. (2008) Michigan molecular interactions r2: from interacting proteins to pathways, *Nucleic Acids Res*.

Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B. and Hassanali, M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets, *Nucleic Acids Res*, **36**, D901-906.