Schizophrenia as a Disorder of Molecular Pathways

Supplement 1

Bioinformatic Tools

Gene Ontology (GO) classification is perhaps the oldest and most widely used gene classification strategy to date (1). Initially three independent ontologies were developed classifying genes based on biological process, molecular function, and cellular component. Over the last decade this classification expanded and an impressive collection of tools were developed by the GO Consortium and various third parties. These tools are currently accessible through NeuroLex, the Neuroscience Lexicon (http://neurolex.org), encompassing >60,000 concepts that span gross anatomy, cells of the nervous system, subcellular structures, diseases, functions, and techniques (2). Supported by The Neuroscience Information Framework and the International Neuroinformatics Coordinating Facility, this regularly updated resource represents an essential source of information for any knowledge-based gene classification. A good, simple way to apply these classifications to transcriptome studies comes from Database for Annotation, Visualization and Integrated Discovery (DAVID) (3, 4), which particularly excels in identifying enriched biological GO terms in the dataset. Furthermore, this scientific tool is able to visualize genes on BioCarta & KEGG pathway maps (see below).

Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/) (5) is a database resource for understanding high-level functions and utilities of the biological system from the level of the cell all the way to the ecosystem. It is a digital, visual, and database representation of life processes, integrating genomic, proteomic, and chemical knowledge into reaction and relation-based networks. Today, KEGG's integrated database resource consists of the sixteen main databases that can be broadly classified into systems information, genomic information, and chemical information, with each subdivided into further subclasses (6). Initially released in 1995, the first and best known is KEGG PATHWAY, which represents a collection of manually drawn pathways that map our knowledge on the

molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases drug development. Perhaps the best feature on this database is its ability to map large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics to KEGG pathway maps for biological interpretation of higher-level systemic functions. While KEGG offers an impressive analysis toolset (7, 8) (including KEGG Mapper and KEGG Atlas), it is important to note that investigators are not restricted to the use of these tools: KEGG pathway analysis is also an integral part of the most commonly used transcriptome-proteome-metabolome analysis software packages (e.g. DAVID and GenePattern) (4, 9). Unfortunately, as of lately KEGG pathways are now not provided free of charge, and it is noteworthy that the DAVID database has not updated its KEGG pathway definitions since 2009.

BioCarta (http://www.biocarta.com/genes/index.asp) is an open source, interactive graphic model of molecular and cellular pathways. This community-fed forum constantly integrates emerging proteomic information from the scientific community, providing information for over 120,000 genes from multiple species. Many of these pathways have been integrated into commonly used gene expression analysis tools, including GenePattern (9).

Weighted Correlation Network Analysis (WGCNA) (10) is relatively new approach for analysis of large scale data-driven experiments, yet it is perhaps the most informative method for analysis of large scale transcriptome data: WGCNA does not require or rely on *a priori* knowledge. Rather, WGCNA, using eigengene network methodology, can be used for finding clusters (modules) of highly correlated genes and for relating these modules to one another and to external sample traits. Thus, one can identify transcripts that are correlated and behave similarly within and across the samples or conditions, enabling discovery of previously unknown, putatively functional relationships between transcripts. WGCNA groups genes into similarly functioning modules based on their expression, and those modules are generally tested via secondary pathway analysis tools to determine what biological processes they represent.

Bioconductor (http://www.bioconductor.org/) is a gateway for packages related to any type of bioinformatics analysis, including, but not limited to gene expression and pathway analyses. It is an open source software platform providing tools for the analysis and comprehension of high-throughput genomic data. For analysis of gene expression, it integrates most of the available analytical approaches into one comprehensive package (11, 12). Using the R statistical programming language, Bioconductor, with its several hundred software packages, can perform complex analysis of common microarray platforms, (e.g. Affymetrix, Illumina, Nimblegen, Agilent) supporting exon, copy number, SNP, methylation, and other assays. This software system also excels in pre-processing, quality assessment, differential expression, clustering and classification. In addition, various gene set enrichment analyses from multiple sources are a crucial component of the software package, allowing performing the above-mentioned GO, BioCarta, KEGG, and WGCNA analyses through a single platform.

Cytoscape (www.cytoscape.org) is an open source software platform for visualizing complex networks and data integration across different sources of information (13). It is very versatile in network integrating, inference customization, literature mining, topological clustering, functional enrichment, network comparison, and programmatic access. It excels in integration of global datasets and functional annotations, including curated human pathway datasets such as KEGG, Reactome (www.reactome.org) and WikiPathways (www.wikipathways.org).

Supplemental References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al.* (2000): Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 25:25-29.

2. Larson SD, Martone ME (2013): NeuroLex.org: an online framework for neuroscience knowledge. *Front Neuroinform*. 7:18.

3. Huang da W, Sherman BT, Lempicki RA (2009): Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 4:44-57.

4. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, *et al.* (2003): DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*. 4:P3.

5. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999): KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 27:29-34.

6. Tanabe M, Kanehisa M (2012): Using the KEGG database resource. *Curr Protoc Bioinformatics*. Chapter 1:Unit1 12.

7. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, *et al.* (2008): KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*. 36:W423-426.

8. Wrzodek C, Drager A, Zell A (2011): KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics*. 27:2314-2315.

9. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP (2006): GenePattern 2.0. *Nat Genet*. 38:500-501.

10. Langfelder P, Horvath S (2008): WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 9:559.

11. Reimers M, Carey VJ (2006): Bioconductor: an open source framework for bioinformatics and computational biology. *Methods Enzymol*. 411:119-134.

12. Okoniewski MJ, Miller CJ (2008): Comprehensive analysis of affymetrix exon arrays using BioConductor. *PLoS Comput Biol*. 4:e6.

13. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, *et al.* (2012): A travel guide to Cytoscape plugins. *Nat Methods*. 9:1069-1076.