Canadian Bioinformatics Workshops

www.bioinformatics.ca
Module 1
Introduction to Gene Lists

Gary Bader
Pathway and Network Analysis of -omics Data
June 4-6, 2014

Interpreting Gene Lists

- My cool new screen worked and produced 1000 hits! ...Now what?
- Genome-Scale Analysis (Omens)
  - Genomics, Proteomics
- Tell me what’s interesting about these genes

Ranking or clustering

GenMAPP.org

Module 1: Introduction to Gene Lists
bioinformatics.ca
Interpreting Gene Lists

• My cool new screen worked and produced 1000 hits! ...Now what?
• Genome-Scale Analysis (Omics)
  – Genomics, Proteomics
• Tell me what’s interesting about these genes
  – Are they enriched in known pathways, complexes, functions

From genotype to phenotype via pathways

Whole genome

Predict effect

Molecular, physiological phenotype

Experiments, Predictions (PPIs)

Databases

Literature

Experts
Pathway and Network Analysis

- Any type of analysis that involves pathway or network information
- Most commonly applied to help interpret lists of genes
- Most popular type is pathway enrichment analysis, but many others are useful
- Helps gain mechanistic insight into ‘omics data

Benefits of Pathway Data vs. transcripts, proteins, SNPs...

- Improves statistical power
  - Fewer tests
- More reproducible
  - E.g. gene expression signatures
- Easier to interpret
  - Familiar concepts e.g. cell cycle
- Identifies mechanism
  - Can explain cause
Before Analysis

- Normalization
- Background adjustment
- Quality control (garbage in, garbage out)

- Use statistics that will increase signal and reduce noise specifically for your experiment
- Gene list size
- Make sure your gene IDs are compatible with software

Pathway analysis workflow

1. Collect genomics data (e.g. mRNA expression)
2. Normalize and score (e.g. compute differential expression)
3. Generate gene list
4. Learn about underlying cellular mechanism using pathway and network analysis
5. Visualize and identify interesting pathways and networks
6. Drill down to understand molecular mechanism
7. Publish model explaining data
Autism Spectrum Disorder (ASD)

• Genetics
  – highly heritable
    • monozygotic twin concordance 60-90%
    • dizygotic twin concordance 0-10%
      (depending on the stringency of diagnosis)
  – known genetics:
    • 5-15% rare single-gene disorders and chromosomal re-arrangements
    • de-novo CNV previously reported in 5-10% of ASD cases
    • GWA (Genome-wide Association Studies) have been able to explain only a small amount of heritability


Rare copy number variants in ASD

• Rare Copy Number Variation screening (Del, Dup)
  – 889 Case and 1146 Ctrl (European Ancestry)
  – Illumina Infinium 1M-single SNP
  – high quality rare CNV (90% PCR validation)
    • identification by three algorithms required for detection
      – QuantiSNP, iPattern, PennCNV
    • frequency < 1%, length > 30 kb

• Results
  – average CNV size: 182.7 kb, median CNVs per individual: 2
  – > 5.7% ASD individuals carry at least one de-novo CNV
  – Top ~10 genes in CNVs associated to ASD
Pathways Enriched in Autism Spectrum

Zoom of CNS-Development

Module 1: Introduction to Gene Lists
Where Do Gene Lists Come From?

- Molecular profiling e.g. mRNA, protein
  - Identification → Gene list
  - Quantification → Gene list + values
  - Ranking, Clustering (biostatistics)
- Interactions: Protein interactions, microRNA targets, transcription factor binding sites (ChIP)
- Genetic screen e.g. of knock out library
- Association studies (Genome-wide)
  - Single nucleotide polymorphisms (SNPs)
  - Copy number variants (CNVs)

What Do Gene Lists Mean?

- Biological system: complex, pathway, physical interactors
- Similar gene function e.g. protein kinase
- Similar cell or tissue location
- Chromosomal location (linkage, CNVs)
Biological Questions

• Step 1: What do you want to accomplish with your list (hopefully part of experiment design! 😊)
  – Summarize biological processes or other aspects of gene function
  – Perform differential analysis – what pathways are different between samples?
  – Find a controller for a process (TF, miRNA)
  – Find new pathways or new pathway members
  – Discover new gene function
  – Correlate with a disease or phenotype (candidate gene prioritization)

Biological Answers

• Computational analysis methods we will cover
  – Regulatory network analysis: find controllers
  – Pathway enrichment analysis: summarize and compare
  – Network analysis: predict gene function, find new pathway members, identify functional modules (new pathways)
Pathway Enrichment Analysis

- Gene identifiers
- Pathways and other gene annotation
  - Gene Ontology
    - Ontology Structure
    - Annotation
  - BioMart + other sources

Gene List Enriched Pathways

Also:
Function Phenotypes

GSEA, g:Profiler

Gene and Protein Identifiers

- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
  - E.g. Social Insurance Number, Entrez Gene ID 41232
- Gene and protein information stored in many databases
  - Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
  - Important to recognize the correct record type
  - E.g. Entrez Gene records don’t store sequence. They link to DNA regions, RNA transcripts and proteins e.g. in RefSeq, which stores sequence.
Module 1: Introduction to Gene Lists

NCBI Database Links

NCBI: U.S. National Center for Biotechnology Information
Part of National Library of Medicine (NLM)


For your information

Common Identifiers

Gene
- Ensembl ENSG00000139618
- Entrez Gene 675
- Unigene Hs.34012

RNA transcript
- GenBank BC026160.1
- RefSeq NM_000059
- Ensembl ENST00000380152

Protein
- Ensembl ENSP00000369497
- RefSeq NP_000050.2
- UniProt BRCA2_HUMAN or A1YBP1_HUMAN
- IPI IPI00412408.1
- EMBL AF309413
- PDB 1MIU

Species-specific
- HUGO HGNC BRCA2
- MGI MGI:109337
- RGD 2219
- ZFIN ZDB-GENE-060510-3
- FlyBase CG9097
- WormBase WBGene00002299 or ZK1067.1
- SGD S000002187 or YDL029W

Annotations
- InterPro IPR015252
- OMIM 600185
- Pfam PF09104
- Gene Ontology GO:0000724
- SNPs rs28897757

Experimental Platform
- Affymetrix 208368_3p_s_at
- Agilent A_23_P99452
- CodeLink GE60169
- Illumina GI_4502450-S

Red = Recommended
Identifier Mapping

- So many IDs!
  - Software tools recognize only a handful
  - May need to map from your gene list IDs to standard IDs
- Four main uses
  - Searching for a favorite gene name
  - Link to related resources
  - Identifier translation
    - E.g. Proteins to genes, Affy ID to Entrez Gene
  - Merging data from different sources
    - Find equivalent records

Module 1: Introduction to Gene Lists

ID Challenges

- Avoid errors: map IDs correctly
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol: TP53
- Excel error-introduction
  - OCT4 is changed to October-4 (paste as text)
- Problems reaching 100% coverage
  - E.g. due to version issues
  - Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80
Retraction: Hes1 is a target of microRNA-23 during retinoic-acid-induced neuronal differentiation of NT2 cells

Hiroaki Kawasaki & Kazunari Taira


In this Article, the messenger RNA that is identified to be a target of microRNA-23 (miR-23) is from the gene termed human 'homolog of ES1' (HES1), accession number Y07572, and not from the gene encoding the transcriptional repressor 'Hairy enhancer of split' HES1 (accession number NM_00524) as stated in our paper. We incorrectly identified the gene because of the confusing nomenclature. The function of HES1 Y07572 is unknown but the encoded protein shares homology with a protein involved in isoprenoid biosynthesis. Our experiments in NT2 cells had revealed that the protein levels of the repressor Hes1 were diminished by miR-23. Although we have unpublished data that suggest the possibility that miR-23 might also interact with Hes1 repressor mRNA, the explanation for the finding that the level of repressor Hes1 protein decreases in response to miR-23 remains undefined with respect to mechanism and specificity. Given the interpretational difficulties resulting from our error, we respectfully retract the present paper. Further studies aimed at clarifying the physiological role of miR-23 will be submitted to a peer-reviewed journal subject to the outcome of our ongoing research.

ID Mapping Services

- g:Convert
  - http://biit.cs.ut.ee/gprofiler/gconvert.cgi
- Ensembl BioMart
  - http://www.ensembl.org
- PICR (proteins only)
  - http://www.ebi.ac.uk/Tools/picr/
Recommendations

• For proteins and genes
  – (doesn’t consider splice forms)
• Map everything to Entrez Gene IDs or Official Gene Symbols using a spreadsheet
• If 100% coverage desired, manually curate missing mappings
• Be careful of Excel auto conversions – especially when pasting large gene lists!
  – Remember to format cells as ‘text’ before pasting

What Have We Learned?

• Genes and their products and attributes have many identifiers (IDs)
• Genomics often requires conversion of IDs from one type to another
• ID mapping services are available
• Use standard, commonly used IDs to reduce ID mapping challenges
Module 1: Introduction to Gene Lists

Pathway Enrichment Analysis

- Gene identifiers
- Pathways and other gene annotation
  - Gene Ontology
    - Ontology Structure
    - Annotation
  - BioMart + other sources

Also:
- Function Phenotypes
- GSEA, g:Profiler

Pathways and other gene function attributes

- Available in databases
- Pathways
  - Gene Ontology biological process, pathway databases e.g. Reactome
- Other annotations
  - Gene Ontology molecular function, cell location
  - Chromosome position
  - Disease association
  - DNA properties
    - TF binding sites, gene structure (intron/exon), SNPs
  - Transcript properties
    - Splicing, 3’ UTR, microRNA binding sites
  - Protein properties
    - Domains, secondary and tertiary structure, PTM sites
  - Interactions with other genes
Pathways and other gene function attributes

- Available in databases
- Pathways
  - Gene Ontology biological process, pathway databases e.g. Reactome
- Other annotations
  - Gene Ontology molecular function, cell location
  - Chromosome position
  - Disease association
  - DNA properties
    - TF binding sites, gene structure (intron/exon), SNPs
  - Transcript properties
    - Splicing, 3’ UTR, microRNA binding sites
  - Protein properties
    - Domains, secondary and tertiary structure, PTM sites
  - Interactions with other genes

What is the Gene Ontology (GO)?

- Set of biological phrases (terms) which are applied to genes:
  - protein kinase
  - apoptosis
  - membrane
- Dictionary: term definitions
- Ontology: A formal system for describing knowledge
- www.geneontology.org
GO Structure

- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child

What GO Covers?

- GO terms divided into three aspects:
  - cellular component
  - molecular function
  - biological process

Cell division
Part 1/2: Terms

- Where do GO terms come from?
  - GO terms are added by editors at EBI and gene annotation database groups
  - Terms added by request
  - Experts help with major development
  - 37104 terms, with definitions
    - 23074 biological_process
    - 2994 cellular_component
    - 9392 molecular_function
    - As of June 2012

Part 2/2: Annotations

- Genes are linked, or associated, with GO terms by trained curators at genome databases
  - Known as ‘gene associations’ or GO annotations
  - Multiple annotations per gene
- Some GO annotations created automatically (without human review)
Annotion Sources

• Manual annotation
  – Curated by scientists
    • High quality
    • Small number (time-consuming to create)
  – Reviewed computational analysis

• Electronic annotation
  – Annotation derived without human validation
    • Computational predictions (accuracy varies)
    • Lower ‘quality’ than manual codes

• Key point: be aware of annotation origin

Module 1: Introduction to Gene Lists  bioinformatics.ca

For your information

Evidence Types

• Experimental Evidence Codes
  • EXP: Inferred from Experiment
  • IDA: Inferred from Direct Assay
  • IPI: Inferred from Physical Interaction
  • IMP: Inferred from Mutant Phenotype
  • IGI: Inferred from Genetic Interaction
  • IEP: Inferred from Expression Pattern

• Author Statement Evidence Codes
  • TAS: Traceable Author Statement
  • NAS: Non-traceable Author Statement

• Curator Statement Evidence Codes
  • IC: Inferred by Curator
  • ND: No biological Data available

• Computational Analysis Evidence Codes
  • ISS: Inferred from Sequence or Structural Similarity
  • ISO: Inferred from Sequence Orthology
  • ISA: Inferred from Sequence Alignment
  • ISM: Inferred from Sequence Model
  • IGC: Inferred from Genomic Context
  • RCA: Inferred from Reviewed Computational Analysis

• IEA: Inferred from electronic annotation

Species Coverage

- All major eukaryotic model organism species and human
- Several bacterial and parasite species through TIGR and GeneDB at Sanger
- New species annotations in development
- Current list:
  - http://www.geneontology.org/
  - GO.downloads.annotations.shtml

Variable Coverage

For your information

**Contributing Databases**

- Berkeley Drosophila Genome Project (BDGP)
- *dictyBase* (*Dictyostelium discoideum*)
- *FlyBase* (*Drosophila melanogaster*)
- *GeneDB* (*Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Leishmania major* and *Trypanosoma brucei*)
- UniProt Knowledgebase (Swiss-Prot/TrEMBL/PIR-PSD) and *InterPro* databases
- *Gramene* (grains, including rice, *Oryza*)
- Mouse Genome Database (MGD) and Gene Expression Database (GXD) (*Mus musculus*)
- Rat Genome Database (RGD) (*Rattus norvegicus*)
- Reactome
- *Saccharomyces Genome Database* (SGD) (*Saccharomyces cerevisiae*)
- The Arabidopsis Information Resource (TAIR) (*Arabidopsis thaliana*)
- The Institute for Genomic Research (TIGR): databases on several bacterial species
- *WormBase* (*Caenorhabditis elegans*)
- Zebrafish Information Network (ZFIN): (*Danio rerio*)

**GO Slim Sets**

- GO has too many terms for some uses
  - Summaries (e.g. Pie charts)
- GO Slim is an official reduced set of GO terms
  - Generic, plant, yeast

GO Software Tools

- GO resources are freely available to anyone without restriction
  - Includes the ontologies, gene associations and tools developed by GO
- Other groups have used GO to create tools for many purposes

Accessing GO: QuickGO

http://www.ebi.ac.uk/QuickGO/
Other Ontologies

- cell
  - cell in vivo
  - cell by organism
  - eukaryotic cell
    - Mycetozoa cell
    - fungal cell
      - hyphal cell
      - vegetative cell (sensu Fungi)
    - spore
  - heterokaryon
  - dikaryon
  - animal cell
  - plant cell
  - prokaryotic cell
  - spore
  - cell by class
  - stem cell

http://www.ebi.ac.uk/ontology-lookup

Module 1: Introduction to Gene Lists  bioinformatics.ca

Pathway Databases

- http://www.pathguide.org/lists ~550 pathway related databases
- http://www.pathwaycommons.org/ collects major ones

Module 1: Introduction to Gene Lists  bioinformatics.ca
Pathways and other gene function attributes

- Available in databases
- Pathways
  - Gene Ontology biological process, pathway databases e.g. Reactome
- Other annotations
  - Gene Ontology molecular function, cell location
  - Chromosome position
  - Disease association
  - DNA properties
    - TF binding sites, gene structure (intron/exon), SNPs
  - Transcript properties
    - Splicing, 3' UTR, microRNA binding sites
  - Protein properties
    - Domains, secondary and tertiary structure, PTM sites
  - Interactions with other genes

Sources of Gene Attributes

- Ensembl BioMart (general)
  - http://www.ensembl.org
- Entrez Gene (general)
- Model organism databases
  - E.g. SGD: http://www.yeastgenome.org/
- Many others: discuss during lab
### Ensembl BioMart

- Convenient access to gene list annotation

![Ensembl BioMart Interface](image)

- Select genome
- Select filters
- Select attributes to download

![Gene List Interface](image)

www.ensembl.org

### What Have We Learned?

- Pathways and other gene attributes in databases
  - Pathways from Gene Ontology (GO) and pathway databases
  - Gene Ontology (GO)
    - GO is a classification system and dictionary for biological concepts
    - Annotations are contributed by many groups
    - More than one annotation term allowed per gene
    - Some genomes are annotated more than others
    - Annotation comes from manual and electronic sources
    - GO can be simplified for certain uses (GO Slim)

- Many gene attributes available from genome databases such as Ensembl
Module 1: Introduction to Gene Lists

Pathway analysis workflow

1. Collect genomics data (e.g. mRNA expression)
2. Normalize and score (e.g. compute differential expression)
3. Generate gene list
4. Learn about underlying cellular mechanism using pathway and network analysis
5. Visualize and identify interesting pathways and networks
6. Drill down to understand molecular mechanism
7. Publish model explaining data

Measures:
- mRNA expression (e.g. RNA-Seq)
- Protein expression
- DNA methylation
- DNA methylation at gene promoters
- RNA and DNA
- miRNA
- Protein binding to DNA or RNA (e.g. ChIP, CLIP)
- SNPs or CNVs
- Sequence cancer and/or normal genomes

Analysis:
- Pathway enrichment analysis for rare events
- Rank-based pathway enrichment analysis
- List-based pathway enrichment analysis
- Map protein interaction network
- Measure SNP or CNV in cases and controls
- Identify and filter variants
- Sequence cancer and/or normal genomes

Tools:
- GeneMANIA
- GSEA
- Enrichment Map
- Cytoscape
- Pathvisio
- Reactome FI
- jActiveModules
- ClusterMaker
- GeneMANIA Network Assessor

Mechanistic drill down:
- Drill down to mechanism (e.g. overlay genomics data on pathway or network diagram)
- Develop mechanistic model explaining data
- Integrate additional information e.g. disease signatures, mRNA/TX targets, drug targets
- Measure mRNA expression (e.g. RNA-Seq)
- Measure DNA methylation
- Measure differential expression, methylation (e.g. exp vs. control, time series)
- Measure target genes
- Identify significant or recurrent gene-associated variants in cases
- Measure SNPs or CNVs in cases and controls
- Compare to known networks
- Identify interesting pathways
- Identify interesting networks
- Visualize and identify network modules and their functions
Lab: Gene IDs and Attributes

- Objectives
  - Learn about gene identifiers, Synergizer and BioMart
- Use yeast demo gene list (module1YeastGenes.txt)
- Convert Gene IDs to Entrez Gene: Use g:Profiler
- Get GO annotation + evidence codes
  - Use Ensembl BioMart
  - Summarize terms & evidence codes in a table
- Do it again with your own gene list
  - If compatible with covered tools, run the analysis. If not, instructors will recommend tools for you.

We are on a Coffee Break & Networking Session