Gene Annotation and GO

BST 226
Statistical Methods for Bioinformatics
Slide Sources

• www.geneontology.org
• Jane Lomax (EBI)
• David Hill (MGI)
• Pascale Gaudet (dictyBase)
• Stacia Engel (SGD)
• Rama Balakrishnan (SGD)
The Gene Ontologies

A Common Language for Annotation of Genes from Yeast, Flies and Mice

…and Plants and Worms

…and Humans

…and anything else!
Gene Ontology Objectives

• GO represents categories used to classify specific parts of our biological knowledge:
  – Biological Process
  – Molecular Function
  – Cellular Component

• GO develops a common language applicable to any organism

• GO terms can be used to annotate gene products from any species, allowing comparison of information across species
Expansion of Sequence Info

Billions of Bases in Genbank/WGS

Date

Total Bases in Genbank (Billions)

January 29, 2014

Statistical Methods for Bioinformatics
Expansion of Sequence Info

Billions of Bases in Genbank/WGS (Log Scale)

January 29, 2014
# Entering the Genome Sequencing Era

<table>
<thead>
<tr>
<th>Eukaryotic Genome Sequences</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Yeast</strong> (<em>S. cerevisiae</em>)</td>
<td>1996</td>
</tr>
<tr>
<td><strong>Worm</strong> (<em>C. elegans</em>)</td>
<td>1998</td>
</tr>
<tr>
<td><strong>Fly</strong> (<em>D. melanogaster</em>)</td>
<td>2000</td>
</tr>
<tr>
<td><strong>Plant</strong> (<em>A. thaliana</em>)</td>
<td>2001</td>
</tr>
<tr>
<td><strong>Human</strong> (<em>H. sapiens, 1st Draft</em>)</td>
<td>2001</td>
</tr>
</tbody>
</table>
Baldauf et al. (2000)
Science 290:972
These proteins form a hexamer in the species that have been examined.
Outline of Topics

• Introduction to the Gene Ontologies (GO)

• Annotations to GO terms

• GO Tools

• Applications of GO
What is Ontology?

- Dictionary: A branch of metaphysics concerned with the nature and relations of being.
- Barry Smith: The science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality.
So what does that mean?

From a practical view, ontology is the representation of something we know about. “Ontologies" consist of a representation of things, that are detectable or directly observable, and the relationships between those things.
The ontology. Dividing human knowledge into a clean set of categories is a lot like trying to figure out where to find that suspenseful black comedy at your corner video store. Questions inevitably come up, like are Movies part of Art or Entertainment? (Yahoo! lists them under the latter.) -Wired Magazine, May 1996
The 3 Gene Ontologies

- **Molecular Function** = elemental activity/task
  - the tasks performed by individual gene products; examples are carbohydrate binding and ATPase activity

- **Biological Process** = biological goal or objective
  - broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions

- **Cellular Component** = location or complex
  - subcellular structures, locations, and macromolecular complexes; examples include nucleus, telomere, and RNA polymerase II holoenzyme
Example:
Gene Product = hammer

**Function** (what)  **Process** (why)

Drive nail (into wood)  Carpentry

Drive stake (into soil)  Gardening

Smash roach  Pest Control

Clown’s juggling object  Entertainment
Biological Examples

<table>
<thead>
<tr>
<th>Biological Process</th>
<th>Molecular Function</th>
<th>Cellular Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lactate dehydrogenase 1.1.1.27</td>
<td>Propionate</td>
<td>Lactobacillus</td>
</tr>
<tr>
<td>Sour milk</td>
<td>Lactate</td>
<td>Propionibacterium</td>
</tr>
<tr>
<td>Sauerkraut</td>
<td>2[H+]</td>
<td>Streptococcus (10 μm)</td>
</tr>
<tr>
<td>Yoghurt</td>
<td></td>
<td>DNA</td>
</tr>
<tr>
<td>DNA</td>
<td></td>
<td>Cell wall</td>
</tr>
</tbody>
</table>

A. Lactic acid and propionic acid fermentation

<table>
<thead>
<tr>
<th>Biological Process</th>
<th>Molecular Function</th>
<th>Cellular Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyruvate decarboxylase [TPP] 4.1.1.1</td>
<td>Pyruvate</td>
<td>Yeast (Saccharomyces cerevisiae)</td>
</tr>
<tr>
<td>Alcoholic fermentation</td>
<td></td>
<td>Vacuole</td>
</tr>
</tbody>
</table>

B. Alcoholic fermentation

<table>
<thead>
<tr>
<th>Biological Process</th>
<th>Molecular Function</th>
<th>Cellular Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol dehydrogenase [Zn²⁺] 1.1.1.1</td>
<td>CO₂</td>
<td>Septum</td>
</tr>
<tr>
<td></td>
<td>Ethanal (Acetaldehyde)</td>
<td>Daughter cell</td>
</tr>
<tr>
<td></td>
<td>Ethanol</td>
<td>Nucleus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ER</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cell wall</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tonoplast</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mitochondrion</td>
</tr>
</tbody>
</table>
Terms, Definitions, IDs

term: MAPKKK cascade (mating sensu Saccharomyces)

goid: GO:0007244

definition: MAPKKK cascade involved in transduction of mating pheromone signal, as described in Saccharomyces

definition_reference: PMID:9561267
Ontology

Includes:

1. A vocabulary of terms (names for concepts)
2. Definitions
3. Defined logical relationships to each other
chromosome

organelle

nucleus

[other types of chromosomes]

nuclear chromosome

[other organelles]
Ontology Structure

Ontologies can be represented as graphs, where the **nodes** are connected by **edges**

- Nodes = **terms** in the ontology
- Edges = **relationships** between the concepts
A child is a subset or instances of a parent’s elements

Parent-Child Relationships

Chromosome

Cytoplasmic chromosome
Mitochondrial chromosome
Nuclear chromosome
Plastid chromosome
Ontology Structure

• The Gene Ontology is structured as a hierarchical directed acyclic graph (DAG)

• Terms can have more than one parent and zero, one or more children

• Terms are linked by two relationships
  – is-a
  – part-of
Directed Acyclic Graph (DAG)

- chromosome
- organelle
- nucleus
- nuclear chromosome

- is-a
- part-of

[other types of chromosomes]

[other organelles]
http://www.ebi.ac.uk/ego

Gene Ontology

biological_process

physiological_process

development

response to stimulus

organ development

organogenesis

appendage morphogenesis

wound healing

tissue regeneration

fin regeneration
Evidence Codes for GO Annotations

Evidence codes

Indicate the type of evidence in the cited source* that supports the association between the gene product and the GO term

*capturing information
Types of evidence codes

• Experimental codes – EXP, IDA, IMP, IGI, IPI, IEP

• Computational codes - ISS, ISO, ISA, IGC, IBA, IBD, IKR, IRD, RCA, IEA

• Author statement - TAS, NAS

• Other codes - IC, ND
Experimental Evidence Codes

Inferred from Experiment (EXP)
Inferred from Direct Assay (IDA)
Inferred from Physical Interaction (IPI)
Inferred from Mutant Phenotype (IMP)
Inferred from Genetic Interaction (IGI)
Inferred from Expression Pattern (IEP)
Computational Evidence Codes

Inferred from Sequence or structural Similarity (ISS)
Inferred from Sequence Orthology (ISO)
Inferred from Sequence (ISA)
Inferred from Sequence Model (ISM)
Inferred from Genomic Context (IGC)
Inferred from Biological aspect of Ancestor (IBA)
Inferred from Biological aspect of Descendant (IBD)
Inferred from Key Residues (IKR)
Inferred from Rapid Divergence (IRD)
inferred from Reviewed Computational Analysis (RCA)
Author Statement Codes

Traceable Author Statement (TAS)
Non-traceable Author Statement (NAS)

Curatorial Statement Evidence Codes

Inferred by Curator (IC)
No biological Data available (ND)

Automatically Assigned Evidence Codes

Inferred from Electronic Annotation (IEA)
IDA

Inferred from Direct Assay

- direct assay for the function, process, or component indicated by the GO term
  - Enzyme assays
  - In vitro reconstitution (e.g. transcription)
  - Immunofluorescence (for cellular component)
  - Cell fractionation (for cellular component)
IMP
Inferred from Mutant Phenotype

• variations or changes such as mutations or abnormal levels of a single gene product
  • Gene/protein mutation
  • Deletion mutant
  • RNAi experiments
  • Specific protein inhibitors
  • Allelic variation
IGI

Inferred from Genetic Interaction

• Any combination of alterations in the sequence or expression of more than one gene or gene product
  • Traditional genetic screens
    - Suppressors, synthetic lethals
  • Functional complementation
  • Rescue experiments

• An entry in the ‘with’ column is recommended
IPI
Inferred from Physical Interaction

- Any physical interaction between a gene product and another molecule, ion, or complex
  - 2-hybrid interactions
  - Co-purification
  - Co-immunoprecipitation
  - Protein binding experiments

- An entry in the ‘with’ column is recommended
IEP

Inferred from Expression Pattern

• Timing or location of expression of a gene
  – Transcript levels
    • Northern, microarray, RNA-Seq

• Exercise caution when interpreting expression results
ISS

Inferred from Sequence or structural Similarity

- Sequence alignment, structure comparison, or evaluation of sequence features such as composition
  - Sequence similarity
  - Recognized domains/overall architecture of protein

- An entry in the ‘with’ column is recommended
RCA

Inferred from Reviewed Computational Analysis

• non-sequence-based computational method
  – large-scale experiments
    • genome-wide two-hybrid
    • genome-wide synthetic interactions
  – integration of large-scale datasets of several types
  – text-based computation (text mining)
IGC
Inferred from Genomic Context

• Chromosomal position

• Most often used for Bacteria - operons
  – Direct evidence for a gene being involved in a process is minimal, but for surrounding genes in the operon, the evidence is well-established
IEA

Inferred from Electronic Annotation

• depend directly on computation or automated transfer of annotations from a database
  – Hits from BLAST searches
  – InterPro2GO mappings

• No manual checking

• **Entry in ‘with’ column is allowed (ex. sequence ID)**
TAS
Traceable Author Statement

• publication used to support an annotation doesn't show the evidence
  – Review article

• *Would be better to track down cited reference and use an experimental code*
NAS

Non-traceable Author Statement

- Statements in a paper that cannot be traced to another publication
ND

No biological Data available

• Can find no information supporting an annotation to any term
• Indicate that a curator has looked for info but found nothing
  – Place holder
  – Date
IC
Inferred by Curator

- annotation is not supported by evidence, but can be reasonably inferred from other GO annotations for which evidence is available
- ex. evidence = transcription factor (function)
  - IC = nucleus (component)
Choosing the correct evidence code

Ask yourself:

What is the experiment that was done?
Using the Gene Ontology (GO) for Expression Analysis
What is the Gene Ontology?

• Set of biological phrases (terms) which are applied to genes:
  – protein kinase
  – apoptosis
  – membrane
What is the Gene Ontology?

• Genes are linked, or associated, with GO terms by trained curators at genome databases
  – known as ‘gene associations’ or GO annotations
• Some GO annotations created automatically
GO annotations

GO database

associated genes

gene -> GO term

gene -> GO term

gene -> GO term

gene -> GO term

gene -> GO term

genome and protein databases
What is the Gene Ontology?

• Allows biologists to make inferences across large numbers of genes without researching each one individually
ASPARAGINE UTILIZATION
CRYSTAL VIOLET RESISTANCE
LYSINE BIOSYNTHESIS
CELL WALL CATABOLISM
OXIDATIVE STRESS RESPONSE
GLUCOSE REPRESSION
AGING
RIBOSE METABOLISM
PROTEIN FOLDING
ANTIPROLIFERATIVE PROTEIN
RNA PROCESSING
UBIQUINONE BIOSYNTHESIS
TRANSCRIPTION
1 PROTEIN SYNTHESIS

GO structure

- GO isn’t just a flat list of biological terms
- terms are related within a hierarchy
GO structure

- all: all (168775)
  - GO:0008150: biological_process (118690)
    - GO:0009987: cellular_process (71171)
    - GO:0050875: cellular_physiological_process (85087)
    - GO:0044237: cellular_metabolism (41106)
    - GO:0006139: nucleobase, nucleoside, nucleotide and nucleic acid metabolism (16561)
    - GO:0006259: DNA_metabolism (4671)
    - GO:0006260: DNA_replication (1115)
GO structure

- This means genes can be grouped according to user-defined levels
- Allows broad overview of gene set or genome
How does GO work?

• GO is species independent
  – some terms, especially lower-level, detailed terms may be specific to a certain group
    • e.g. photosynthesis
  – But when collapsed up to the higher levels, terms are not dependent on species
How does GO work?

What information might we want to capture about a gene product?

- What does the gene product do?
- Where and when does it act?
- Why does it perform these activities?
GO structure

• GO terms divided into three parts:
  – cellular component
  – molecular function
  – biological process
Cellular Component

- where a gene product acts
Cellular Component
Cellular Component

[Diagram of a virus with labels for Capsid, Spikes, and DNA]
Cellular Component

- Enzyme complexes in the component ontology refer to places, not activities.
Molecular Function

- activities or “jobs” of a gene product

glucose-6-phosphate isomerase activity
Molecular Function

insulin binding insulin receptor activity
Molecular Function

• A gene product may have several functions; a function term refers to a single reaction or activity, not a gene product.

• Sets of functions make up a biological process.
Biological Process

a commonly recognized series of events

Preprophase

- Centriole
- Intranuclear condensation of chromosomes

Prophase

- Mitotic spindle
- Individualization of chromosomes, initiation of mitotic spindle, rupture of nuclear envelope

Metaphase

- Chromosomes arranged in equatorial plane, spindle completed, disappearance of nuclear envelope and nucleolus

Telophase

- Nuclear restitution, nuclear envelope and nucleolar formation, end of cell division

Late anaphase

- Aggregation of chromosomes at the poles, beginning of cell division, initiation of cleavage furrow

Early anaphase

- Longitudinal splitting of chromosomes and migration to poles

cell division
Biological Process

**Initiation**

- Promoter
- Factor sigma
- DNA helix
- RNA Polymerase

**Elongation**

- 3'
- RNA
- 5'
- Stop sequence

**Termination**

- (c) Chemis
- RNA
- σ
regulation of gluconeogenesis
Biological Process

limb development
Ontology Structure

• Terms are linked by two relationships
  – is-a
  – part-of
Ontology Structure

cell

membrane          chloroplast

mitochondrial        chloroplast
membrane             membrane

is-a

part-of
Ontology Structure

- Ontologies are structured as a hierarchical directed acyclic graph (DAG)
- Terms can have more than one parent and zero, one or more children
Ontology Structure

Directed Acyclic Graph (DAG) - multiple parentage allowed
Anatomy of a GO term

<table>
<thead>
<tr>
<th>id</th>
<th>unique GO ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>name: gluconeogenesis</td>
<td>term name</td>
</tr>
<tr>
<td>namespace: process</td>
<td>ontology</td>
</tr>
<tr>
<td>def: The formation of glucose from noncarbohydrate precursors, such as pyruvate, amino acids and glycerol.</td>
<td>definition</td>
</tr>
<tr>
<td>[<a href="http://cancerweb.ncl.ac.uk/omd/index.html">http://cancerweb.ncl.ac.uk/omd/index.html</a>]</td>
<td></td>
</tr>
<tr>
<td>exact_synonym: glucose biosynthesis</td>
<td>synonym</td>
</tr>
<tr>
<td>xref_analog: MetaCyc:GLUCONEO-PWY</td>
<td>database ref</td>
</tr>
<tr>
<td>is_a: GO:0006006</td>
<td>parentage</td>
</tr>
<tr>
<td>is_a: GO:0006092</td>
<td></td>
</tr>
</tbody>
</table>

BST 226 Statistical Methods for Bioinformatics
GO tools

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

  http://www.geneontology.org/GO.tools
GO tools

• Affymetrix also provide a Gene Ontology Mining Tool as part of their NetAffx™ Analysis Center which returns GO terms for probe sets
GO tools

• Many tools exist that use GO to find common biological functions from a list of genes:

GO tools

• Most of these tools work in a similar way:
  – input a gene list and a subset of ‘interesting’ genes
  – tool shows which GO categories have most interesting genes associated with them i.e. which categories are ‘enriched’ for interesting genes
  – tool provides a statistical measure to determine whether enrichment is significant
Microarray process

- Treat samples
- Collect mRNA
- Label
- Hybridize
- Scan
- Normalize
- Select differentially expressed genes
- Understand the biological phenomena involved
Traditional analysis

Gene 1
Apoptosis
Cell-cell signaling
Protein phosphorylation
Mitosis
...

Gene 2
Growth control
Mitosis
Oncogenesis
Protein phosphorylation
...

Gene 3
Growth control
Mitosis
Oncogenesis
Protein phosphorylation
...

Gene 4
Nervous system
Pregnancy
Oncogenesis
Mitosis
...

Gene 100
Positive ctrl. of cell prolif
Mitosis
Oncogenesis
Glucose transport
...
Traditional analysis

• gene by gene basis

• requires literature searching

• time-consuming
Using GO annotations

• But by using GO annotations, this work has already been done for you!

Gene Symbol
- A1AA_HUMAN
  ATGCC / GOst
- A4_HUMAN
  ATGCC / GOst
- AA2A_HUMAN
  ATGCC / GOst
- ABS_HUMAN
  ATGCC / GOst
- AG22_HUMAN
  ATGCC / GOst
- AHR_HUMAN
  ATGCC / GOst
- APAF_HUMAN
  ATGCC / GOst
- APGB_HUMAN
  ATGCC / GOst
- ARH6_HUMAN
  ATGCC / GOst
- BCLB_HUMAN
  ATGCC / GOst
- BCLX_HUMAN
  ATGCC / GOst
- BRC1_HUMAN
  ATGCC / GOst
- CAR4_HUMAN
  ATGCC / GOst
- CD14_HUMAN
  ATGCC / GOst

GO:0006915 : apoptosis
Grouping by process

- **Apoptosis**
  - Gene 1
  - Gene 53

- **Mitosis**
  - Gene 2
  - Gene 5
  - Gene 45
  - Gene 7
  - Gene 35
  - ...

- **Glucose transport**
  - Gene 7
  - Gene 3
  - Gene 6
  - ...

- **Positive ctrl. of cell prolif.**
  - Gene 7
  - Gene 3
  - Gene 12
  - ...

- **Growth**
  - Gene 5
  - Gene 2
  - Gene 6
  - ...

---

January 29, 2014

BST 226 Statistical Methods for Bioinformatics
GO for microarray analysis

• Annotations give ‘function’ label to genes

• Ask meaningful questions of microarray data e.g.
  – genes involved in the same process, same/different expression patterns?
Using GO in practice

- statistical measure
  - how likely your differentially regulated genes fall into that category by chance

microarray
1000 genes

experiment

100 genes differentially expressed

mitosis – 80/100
apoptosis – 40/100
p. ctrl. cell prol. – 30/100
glucose transp. – 20/100

mitosis apoptosis positive control of cell proliferation glucose transport

0 10 20 30 40 50 60 70 80

January 29, 2014

BST 226 Statistical Methods for Bioinformatics

86
Using GO in practice

• However, when you look at the distribution of all genes on the microarray:

<table>
<thead>
<tr>
<th>Process</th>
<th>Genes on array</th>
<th># genes expected in 100 random genes</th>
<th>occurred</th>
</tr>
</thead>
<tbody>
<tr>
<td>mitosis</td>
<td>800/1000</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>apoptosis</td>
<td>400/1000</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>p. ctrl. cell prol.</td>
<td>100/1000</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>glucose transp.</td>
<td>50/1000</td>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>
AmiGO

- Web application that reads from the GO Database (mySQL)
- [http://amigo.geneontology.org/cgi-bin/amigo/go.cgi](http://amigo.geneontology.org/cgi-bin/amigo/go.cgi)
- Allows us to
  - browse the ontologies
  - view annotations from various species
  - compare sequences (GOst)
- Ontologies are loaded into the database from the gene_ontology.obo file
- Annotations are loaded from the gene_association files submitted by the various annotating groups
  - Only ‘Non-IEA’ annotations are loaded
AmiGO
http://www.godatabase.org

Node has children, can be clicked to view children

January 29, 2014
BST 226 Statistical Methods for Bioinformatics
Some basics

- Leaf node or no children

- Node has children, can be clicked to view children
- Node has been opened, can be clicked to close

Is_a relationship
Part_of relationship

pie chart summary of the numbers of gene products associated to any immediate descendants of this term in the tree.
## Searching the Ontologies

<table>
<thead>
<tr>
<th>GO Term</th>
<th>GO ID</th>
<th>Match Synonym</th>
<th>Ontology</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA ligation during DNA-dependent DNA replication</td>
<td>GO:0051104</td>
<td>DNA replication accessory factor and 1 more</td>
<td>P</td>
<td>The re-formation of a broken phosphodiester bond in the DNA backbone, carried out by DNA ligase, as occurs during DNA-dependent DNA replication.</td>
</tr>
<tr>
<td>DNA replication</td>
<td>GO:0006260</td>
<td>DNA replication accessory factor and 1 more</td>
<td>P</td>
<td>The process whereby new strands of DNA are synthesized. The template for replication can either be DNA or RNA. Definition: The re-formation of a broken phosphodiester bond in the DNA backbone, carried out by DNA ligase, as occurs during DNA-dependent DNA replication. Comments: See also the biological process terms 'DNA-dependent DNA replication' (GO:0006261) and 'RNA-dependent DNA replication' (GO:0006278).</td>
</tr>
<tr>
<td>DNA replication checkpoint</td>
<td>GO:0000076</td>
<td></td>
<td>P</td>
<td>A signal transduction based surveillance mechanism that prevents the initiation of mitosis until DNA replication is complete, thereby ensuring that progeny inherit a full complement of the genome.</td>
</tr>
<tr>
<td>DNA replication factor A complex</td>
<td>GO:0002662</td>
<td></td>
<td>C</td>
<td>A conserved heterotrimeric complex that binds nonspecifically to single-stranded DNA and is required for multiple processes in eukaryotic DNA metabolism, including DNA replication, DNA repair, and...</td>
</tr>
</tbody>
</table>
Term Tree View

DNA replication

- GO:0006260: DNA replication (1108)
  - GO:0006277: DNA amplification (29)
  - GO:0007311: DNA synthesis during DNA repair (19)
  - GO:0006261: DNA-dependent DNA replication (507)
  - GO:0006279: premeiotic DNA synthesis (8)
  - GO:0001326: replication of extrachromosomal circular DNA (0)
  - GO:0006278: RNA-dependent DNA replication (169)
- GO:0007592: physiological process (72074)
  - GO:00050875: cellular physiological process (84265)
  - GO:0004437: cellular metabolism (40939)
    - GO:0006139: nucleobase, nucleoside, nucleotide and nucleic acid metabolism (16364)
    - GO:0006259: DNA metabolism (4868)
- GO:0007582: biological process (1168032)
  - GO:00090987: cellular process (70985)
    - GO:00050875: cellular physiological process (84265)
  - GO:0004437: cellular metabolism (40939)
    - GO:0006139: nucleobase, nucleoside, nucleotide and nucleic acid metabolism (16364)
    - GO:0006259: DNA metabolism (4868)

- GO:0008152: metabolism (44331)
  - GO:0004437: cellular metabolism (40939)
    - GO:0006139: nucleobase, nucleoside, nucleotide and nucleic acid metabolism (16364)
    - GO:0006259: DNA metabolism (4868)

- GO:00043170: macromolecule metabolism (23502)
  - GO:0043283: biopolymer metabolism (13452)
  - GO:0006259: DNA metabolism (4868)

Comment:
The process whereby new strands of DNA are synthesized. The template for replication can either be DNA or RNA.

See also the biological process terms 'DNA-dependent DNA replication' (GO:0006261) and 'RNA-dependent DNA replication' (GO:0006278).
Click on the term name to view term details and annotations.
Annotations associated with a term

Annotation data are from the gene_associations file submitted by the annotating groups

### All Gene Product Associations

(1281 results)

**Filter Associations**

<table>
<thead>
<tr>
<th>Datasource</th>
<th>Evidence Code</th>
<th>Species</th>
<th>Data View</th>
</tr>
</thead>
<tbody>
<tr>
<td>FlyBase</td>
<td>All Curator Approved</td>
<td>All</td>
<td></td>
</tr>
<tr>
<td>SGD</td>
<td>All</td>
<td>All, A. japonica, A. niger</td>
<td></td>
</tr>
</tbody>
</table>

**Get ALL associations here:**

- All Associations

---

### All Gene Product Associations Table

<table>
<thead>
<tr>
<th>Qualifier</th>
<th>Symbol</th>
<th>Information</th>
<th>Source</th>
<th>Evidence</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27H14.40</td>
<td>DNA polymerase delta catalytic subunit, putative, protein from Trypanosoma brucei</td>
<td>TIGR_Tba1</td>
<td>ISS</td>
<td>TIGR_Tba1:annotation</td>
</tr>
<tr>
<td></td>
<td>ABF1</td>
<td>ARS1 binding protein, transcriptional activator, gene from Saccharomyces cerevisiae</td>
<td>SGD</td>
<td>TAS, IDA</td>
<td>PMID:11756546, PMID:11756546</td>
</tr>
<tr>
<td></td>
<td>ACES_HUMAN</td>
<td>ACH protein, protein from Homo sapiens</td>
<td>UniProt</td>
<td>TAS</td>
<td>PMID:11283752</td>
</tr>
<tr>
<td></td>
<td>AT1G07270.1</td>
<td>gene from Arabidopsis thaliana</td>
<td>TAIR</td>
<td>ISS</td>
<td>TAIR:Communication:501714663</td>
</tr>
<tr>
<td></td>
<td>AT1G09815.1</td>
<td>gene from Arabidopsis thaliana</td>
<td>TAIR</td>
<td>ISS</td>
<td>TAIR:Communication:501714663</td>
</tr>
<tr>
<td></td>
<td>AT1G68730.1</td>
<td>gene from Arabidopsis thaliana</td>
<td>TAIR</td>
<td>ISS</td>
<td>TAIR:Communication:501714663</td>
</tr>
</tbody>
</table>
Searching by gene product name

<table>
<thead>
<tr>
<th>Rcl1, RNA terminal phosphate cyclase-like 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Term</strong></td>
</tr>
<tr>
<td>35S primary transcript processing</td>
</tr>
<tr>
<td>nucleolus</td>
</tr>
<tr>
<td>not RNA-3'-phosphate cyclase activity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RCL1, gene from Saccharomyces cerevisiae, data from SGD (S00005370)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Term</strong></td>
</tr>
<tr>
<td>35S primary transcript processing</td>
</tr>
<tr>
<td>nucleolus</td>
</tr>
<tr>
<td>molecular function unknown</td>
</tr>
<tr>
<td>not RNA-3'-phosphate cyclase activity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RCL1_HUMAN, RNA 3'-terminal phosphate cyclase-like protein</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Term</strong></td>
</tr>
<tr>
<td>biological process unknown</td>
</tr>
<tr>
<td>nucleolus</td>
</tr>
<tr>
<td>RNA-3'-phosphate cyclase activity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sperdl1, SPARC-like 1 (mast9, hevin)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Term</strong></td>
</tr>
<tr>
<td>extracellular space</td>
</tr>
</tbody>
</table>
Advanced search

AmiGO

Search GO
- DNA replication
  - Exact Match
  - Terms
  - Gene Symbol/Name
- Submit Query

Gene Product Filters
- Species
  - All
  - A. japonica
  - A. niger
- Datasource
  - All
  - FlyBase
  - SGD
- Evidence Code
  - All Curator Approved
  - IC
  - IMP

Submit Query

Basic Query

Advanced Query
Query By Sequence

Fields
- Name and Synonym
- Name and Symbol

Gene Products

Species
- All
  - A. japonica
  - A. niger

Datasource
- All
  - FlyBase
  - SGD

Evidence Code
- All Curator Approved
  - IC
  - IMP

Help
GOfst
The Gene Ontology
GO Request
AmiGO Request
GOST-Gene Ontology blaST

- Blast a protein sequence against all gene products that have a GO annotation
- Can be accessed from the AmiGO entry page (front page)
GOst can also be accessed from the annotations section.

<table>
<thead>
<tr>
<th>Qualifier</th>
<th>Symbol</th>
<th>Information</th>
<th>Source</th>
<th>Evidence</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATL5</td>
<td></td>
<td>A. THALIANA RIBOSOMAL PROTEIN LS, gene from Arabidopsis thaliana</td>
<td>TAIR</td>
<td>IDA</td>
<td>PMID:12711688</td>
</tr>
<tr>
<td>BRX1</td>
<td></td>
<td>gene from Saccharomyces cerevisiae</td>
<td>SGD</td>
<td>IPI</td>
<td>PMID:11864606</td>
</tr>
<tr>
<td>CBU1840</td>
<td></td>
<td>ribosomal 5S rRNA E-loop binding protein CioL25TL5, protein from Coxieilla burnetii RSA 493</td>
<td>TIGR_CM</td>
<td>ISS</td>
<td>PMID:12704232</td>
</tr>
</tbody>
</table>

**Your GOst Query**

Threshold: 0.001

**Query Sequence**

```
>SGD|S000005437 symbol:BRX1  Ncbi:NP_014565
MSSYVHAKS5KDKNSEEKQCNVKQFMQKNRQTL1ISSGYNVRHRLIQ
LKGSLGPHRKXPEXLDTHKMQQNLNTALELYNCDNWEFARQBDLYL11
LSKPPNPGPKYQINLHMDENPTCNCLGSRVLSDPQTFPSBHYQ
LIKELYVINFCVPPARKXPFIDEVMSF1YVDRKVRVTY1813S7TKNL
BEYEDGIDISLGVEIGPRFMVTVIL1818GPGP1REVKQYVSPNVRA
QIKQAAEAKRSAEAAVERXIKRRNVLADFLSNDALFK
```

**High Scoring Gene Products**

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Score(P)</th>
<th>Datasource</th>
<th>Associated Terms</th>
<th>Aspect</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRX1</td>
<td>2.0e-143</td>
<td>SGD</td>
<td>5S rRNA binding nucleolus ribosomal large subunit assembly and maintenance rRNA primary transcript binding</td>
<td>molecular_function</td>
<td>IPI</td>
</tr>
<tr>
<td>brx1</td>
<td>4.0e-74</td>
<td>GeneDB_Spombe</td>
<td>nucleolus ribosome biogenesis and assembly RNA binding rRNA processing</td>
<td>cellular_component</td>
<td>IPI</td>
</tr>
<tr>
<td>BXDC2_HUMAN</td>
<td>2.9e-64</td>
<td>UniProt</td>
<td>molecular_function unknown</td>
<td>molecular_function</td>
<td>ND</td>
</tr>
<tr>
<td>AT2G15460.1</td>
<td>1.9e-58</td>
<td>TAIR</td>
<td>biological process unknown cellular component unknown molecular function unknown</td>
<td>biological_process</td>
<td>ND</td>
</tr>
<tr>
<td>AT1G52930.1</td>
<td>5.1e-58</td>
<td>TAIR</td>
<td>cellular component unknown</td>
<td>cellular_component</td>
<td>ND</td>
</tr>
<tr>
<td>CG11583</td>
<td>2.3e-55</td>
<td>FlyBase</td>
<td>molecular function unknown nucleolus ribosomal large subunit biogenesis</td>
<td>molecular_function</td>
<td>ND</td>
</tr>
<tr>
<td>zgc:77150</td>
<td>5.9e-07</td>
<td>ZFIN</td>
<td>biological process unknown cellular component unknown molecular function unknown</td>
<td>biological_process</td>
<td>ND</td>
</tr>
<tr>
<td>AT4G01560.1</td>
<td>9.0e-05</td>
<td>TAIR</td>
<td>biological process unknown cellular component unknown molecular function unknown</td>
<td>biological_process</td>
<td>ND</td>
</tr>
</tbody>
</table>
Analysis of Gene Expression Data

- The usual sequence of events is to conduct an experiment in which biological samples under different conditions are analyzed for gene expression.
- Then the data are analyzed to determine differentially expressed genes.
- Then the results can be analyzed for biological relevance.
The Missing Link

- Biological Knowledge
- Expression Experiment
- Statistical Analysis
- Biological Interpretation
Gene Set Enrichment Analysis (GSEA)

• Given a set of genes (e.g., zinc finger proteins), this defines a set of probes on the array.
• Order the probes by smallest to largest change (we use p-value, not fold change).
• Define a cutoff for “significance” (e.g., FDR p-value < .10).
• Are there more of the probes in the group than expected?
<table>
<thead>
<tr>
<th></th>
<th>Not in gene set</th>
<th>In gene Set</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not significant</td>
<td>30</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>91%/75%</td>
<td>9%/38%</td>
<td></td>
</tr>
<tr>
<td>Significant</td>
<td>10</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>67%/25%</td>
<td>33%/62%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>8</td>
<td>48</td>
</tr>
</tbody>
</table>

P-value 0.0947

BST 226 Statistical Methods for Bioinformatics
GSEA for all cutoffs

• If one does GSEA for all possible cutoffs, and then takes the best result, this is equivalent to an easily performed statistical test called the Kolmogorov-Smirnov test for the genes in the set vs. the genes not in the set.

• Programs on www.broad.mit.edu/gsea/

• However this requires a single summary number for each gene, such as a p-value.
An Example Study

• This study examined the effects of relatively low-dose radiation exposure in-vivo in humans with precisely calibrated dose.

• Low LET ionizing radiation is a model of cellular toxicity in which the insult can be given at a single time point with no residual external toxic content as there would be for metals and many long-lived organics.
The study design

• Men were treated for prostate cancer with daily fractions of 2 Gy for a total dose to the prostate of 74 Gy.
• Parts of the abdomen outside the field were exposed to lower doses.
• These could be precisely quantitated by computer simulation and direct measurements by MOSFETs.
• A 3mm biopsy was taken of abdominal skin before the first exposure, then three more were taken three hours after the first exposure at sites with doses of 1, 10, and 100 cGy.

• RNA was extracted and hybridized on Affymetrix HG U133 Plus 2.0 whole genome arrays.

• The question asked was whether a particular gene had a linear dose response, or a response that was linear in (modified) log dose (0, 1, 10, 100 -> -1, 0, 1, 2).
Why is this difficult?

• For a single patient, there are only 4 data points, so the statistical test is not very powerful.

• With 54,675 probe sets, very apparently significant results can happen by chance, so the barrier for true significance is very high.

• This happens in any small sized array study.
• There are reasons to believe that there may be inter-individual variability in response to radiation.
• This means that we may not be able to look for results that are highly consistent across individuals.
• One aspect is the timing of transcriptional cascades.
• Another is polymorphisms that lead to similar probes being differentially expressed, but not the same ones.
Gene 1 → Gene 2 → Gene 3

Gene 1 → Gene 2 → Gene 3

3 Hours
The ToTS Method

• For a gene group like zinc finger proteins, identify the probe sets that relate to that gene group.
• ToTS = Test of Test Statistics
• For each probe set, conduct a statistical test to try to show a linear dose response.
• This yields a t-statistic, which may be positive or negative.
• Conduct a statistical test on the group of t-statistics, testing the hypothesis that the average is zero, vs. leaning to up-regulation or leaning to down-regulation
• This could be a t-test, but we used in this case the Wilcoxon test.
• This can be done one patient at a time, but we can also accommodate inter-individual variability in a study with more than one individual by testing for an overall trend across individuals.

• This is not possible using GSEA, so the ToTS method is more broadly applicable.

• This was published in October, 2005 in *Bioinformatics*. 
Integrity and Consistency

• For zinc finger proteins, there are 799 probe sets and 8 patients for a total of 6,392 different dose-response t-tests
• The Wilcoxon test that the median of these is zero is rejected with a calculated p-value of 0.00008.
• We randomly sampled 2000 sets of probe sets of size 799, and in no case got a more significant result. We call this an empirical p-value (0.000 in this case).
• This is needed because the 6,392 tests are all from 32 arrays
<table>
<thead>
<tr>
<th>Patient</th>
<th>Direction</th>
<th>EPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Up</td>
<td>0.125</td>
</tr>
<tr>
<td>2</td>
<td>Down</td>
<td>0.044</td>
</tr>
<tr>
<td>3</td>
<td>Down</td>
<td>0.001</td>
</tr>
<tr>
<td>4</td>
<td>Up</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>Up</td>
<td>0.003</td>
</tr>
<tr>
<td>6</td>
<td>Up</td>
<td>0.000</td>
</tr>
<tr>
<td>7</td>
<td>Up</td>
<td>0.000</td>
</tr>
<tr>
<td>8</td>
<td>Up</td>
<td>0.039</td>
</tr>
<tr>
<td>All</td>
<td>Up</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Major Advantages

• More sensitive to weak or diffuse signals
• Able to cope with inter-individual variability in response
• Conclusions are solidly based statistically
• Can use a variety of types of biological knowledge
Assessing Significance

• For each gene set, hypergeometric = Fisher’s exact test.
• Not robust to correlations.
• Simple to implement
• Requires specific cutoff
• GSEA KS test is a generalization if used with the standard KS significance points
• Must be adjusted (say, by FDR) if many gene sets are used.
Assessing Significance

• Array permutation, compare significance of set to significance of same set under permutations.

• If there are 12 control and 12 treatment arrays, then there are 2,704,156 ways to choose 12 arrays from the 24 without regard to treatment assignment. P-values can be down to $4 \times 10^{-7}$.

• Can only test the complete null if there is more than one factor.
Assessing Significance

• Gene permutation can test any hypothesis.
• Compare given gene set to random gene sets from the same set of arrays.
• This tests if the given gene set is extreme from a random gene set.
• Array permutation tests if a given gene set is surprising regardless of other gene sets.
• These are different hypotheses, but both may be useful.
Exercise

• Take the top 10 genes from the keratinocyte gene expression study and map their go annotations using AMIGO.
• Are there any obvious common factors?
• Do you think this would work better if you looked at all the significant genes and all the GO annotations, or would this be too difficult?