

Gene Annotation

Contributions from Carlo Colantuoni and
Robert Gentleman

What We Are Going To Cover

Cells, Genes, Transcripts → Genomics Experiments

Sequence Knowledge Behind Genomics Experiments

Annotation of Genes in Genomics Experiments

Biological Setup

Every cell in the human body contains the entire human genome: 3.3 Gb or ~30K genes.

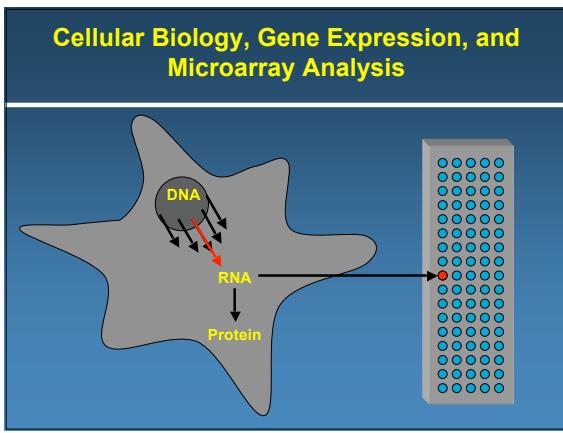
The investigation of gene expression is meaningful because different cells, in different environments, doing different jobs express different genes.

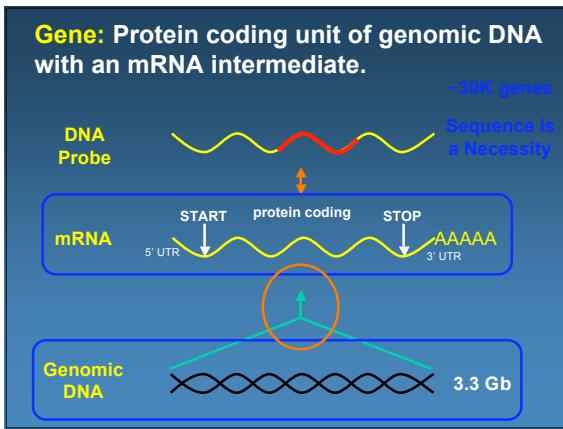
Tasks necessary for gene expression analysis:

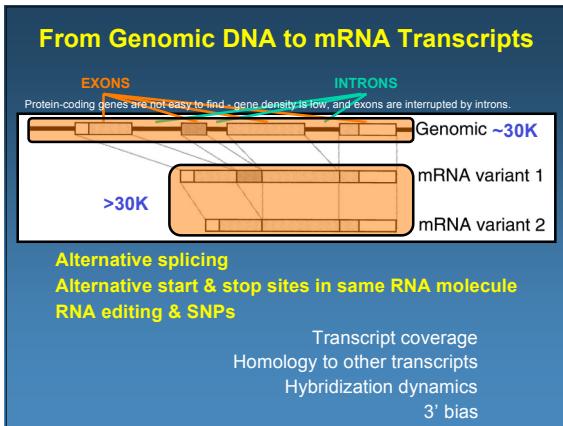
Define what a gene is.

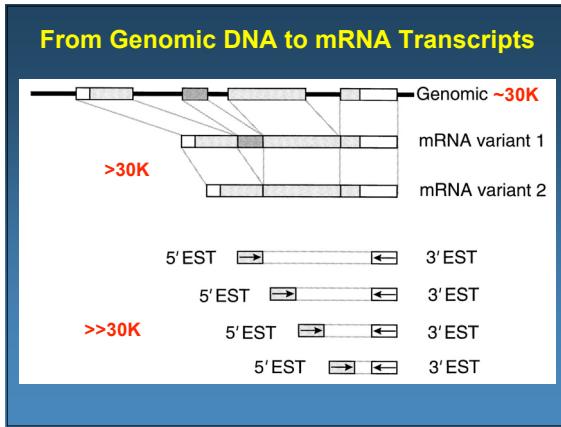
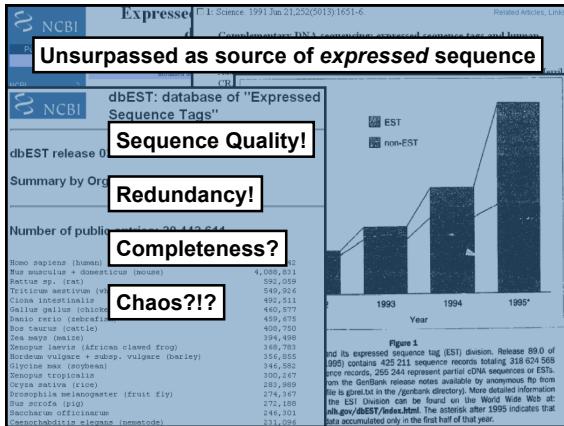
Identify genes in a sea of genomic DNA where <3% of DNA is contained in genes.

Design and implement probes that will effectively assay expression of ALL (most? many?) genes simultaneously. Cross-reference these probes.









Design of Gene Expression Probes

Content: UniGene, Incyte, Celera *Expressed vs. Genomic*

Source: cDNA libraries, clone collections, oligos

Cross-referencing of array probes (across platforms):

Sequence <> GenBank <> UniGene <> HomoloGene

Possible mis-referencing:

- Genomic GenBank Acc.#'s
- Referenced ID has more NT's than probe
- Old DB builds
- DB or table errors – copying and pasting 30K rows in excel ...

Using RefSeq's can help.

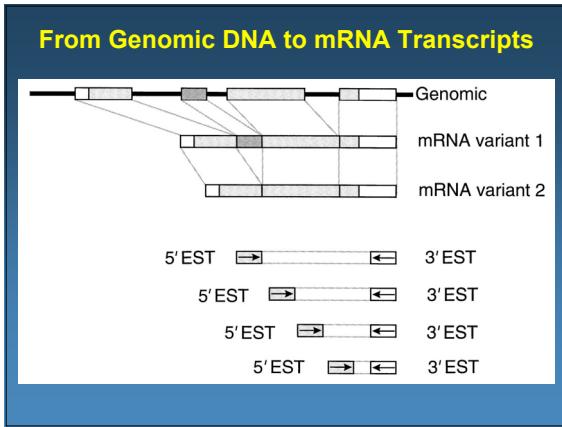


FIGURE 2. RefSeq process flow

The RefSeq database contains many entries, including genes for major model organisms and many others. RefSeq staff work on many different types of data, including RefSeq records for many species, RefSeq annotations for many other species, and RefSeq records for many other types of data.

LocusLink

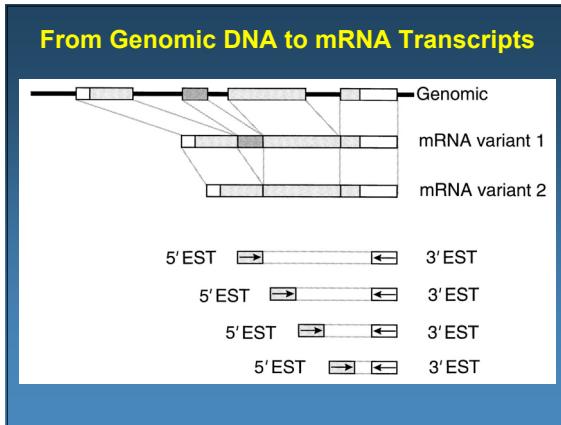
LocusLink provides a single query interface to curated sequence and descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, cross-references, UniGene clusters, homology, map locations, and related web sites.

Sequence accessions include a subset of GenBank accessions for a locus, as well as a new type, the NCBI Reference Sequence (RefSeq). RefSeq records are built according to the process [detailed here](#). See also the [RefSeq FAQ](#).

Data can be accessed by clicking one of the letters above to browse loci sorted by symbol, or by entering a query into the search form. Use of wild cards (*) is supported. Additional information and query tips are provided in the [Help](#) documentation.

The current scope is *Caenorhabditis elegans*, *chicken*, *cow*, *dog*, *human*, *honey bee*, *mouse*, *pig*, *rat*, *zebrafish*, *Xenopus laevis*, *Xenopus tropicalis*, and *petunia*.

Species	Number of RefSeqs
<i>Caenorhabditis elegans</i>	22838
<i>chicken</i>	1310
<i>cow</i>	6092
<i>dog</i>	3343
<i>human</i>	6047
<i>honey bee</i>	1398
<i>mouse</i>	18195
<i>pig</i>	22368
<i>rat</i>	106
<i>zebrafish</i>	11957
<i>Xenopus laevis</i>	4128
<i>Xenopus tropicalis</i>	93
<i>petunia</i>	784
Total	17062
LocusLink records with sequence data for <i>Alosa</i> musculus	31375



Homologene is a system for automatic gene set analysis.

Homologene Release Statistics

Total initial numbers of genes from complete sets of homologous genes, and the numbers of groups for each species.

Species	Number of genes
H. sapiens	31,450
M. musculus	23,650
R. norvegicus	11,764
D. melanogaster	12,336
A. gambiae	10,710
C. elegans	9,492
S. cerevisiae	5,803
A. thaliana	27,133
P. patens	5,209

Last updated: 02/25/2004

Homologene Build Procedure

We have recently adopted a new build procedure which matches amino acid sequence searching (blast) to find more distant homologs. This has led to a significant increase in the number of the statistics. The matching strategy is guided by taxonomic trees, so that more closely related organisms are matched first. Homologene entries now include additional to orthologs.

Entrez **PubMed** **Nucleotide** **Search** **Prokaryotes** **Viruses** **Human/Chimp** **Bird/Comt and human** **Primates** **Custom** **Stardust**

Display **Homologene** **Search** **Send to** **Text**

1. Homologene 50%: Gene conserved in Mammalia

Links

Proteins

Proteins used in sequence comparisons and their conservation architectures:

- H sapiens COMT
- cation-coupled-methyltransferase
- M musculus COMT
- cation-coupled-methyltransferase
- R norvegicus COMT
- D melanogaster COMT
- A gambiae COMT
- C elegans COMT
- S cerevisiae COMT
- A thaliana COMT
- P patens COMT

Phylogeny

Phylogenetic information for the genes in this entry imputed from model organism databases.

- H sapiens MM100100 Schizosaccharomyces, susceptibility to [MM100]

Conserved Domains

The proteins in this entry do not contain any known conserved domains.

Related Homology Resources

Links to related and computed homology information from other databases.

Entrez **PubMed** **OmeGene** for protein M. musculus COMT includes H. sapiens COMT and R. norvegicus COMT.

UniGene

Links to groups of transcribed sequences established by library cloning of UniGene.

- H. sapiens Hs_240013
- M. musculus Mc_10594
- cation-coupled-methyltransferase
- R. norvegicus Rn_10595
- cation-coupled-methyltransferase

NCBI Entrez, The Life Sciences Search Engine

PubMed Entrez Human Genome GenBank Map Viewer BLAST

Search across databases Help

Welcome to the new Entrez cross-database page

PubMed: biomedical literature citations and abstracts
PubMed Central: free, full text journal articles
Books online books
OMIM: Online Mendelian Inheritance in Man
Site Search: NCBI web and FTP sites

Nucleotide sequence database (GenBank)
Protein sequence database
Genomes: whole genome sequences
Structure: three-dimensional macromolecular structures
Taxonomy: organisms in GenBank
SNPs: single nucleotide polymorphism
Gene: gene-centered information
HomoloGene: Eukaryotic homology groups

Journals: detailed information about the journals indexed in PubMed and other Entrez databases
MeSH: detailed information about NLM's controlled vocabulary

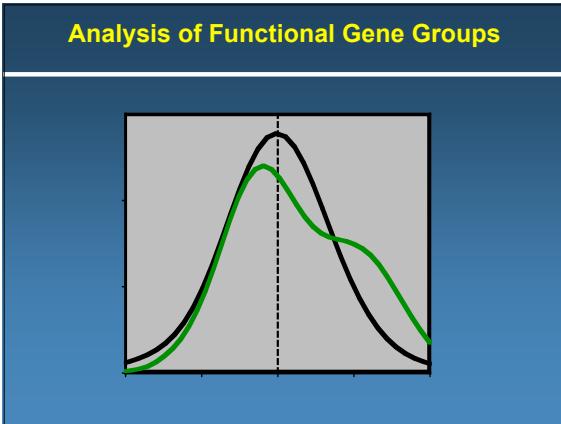
Enter terms and click 'GO' to run the search against ALL the databases, OR
Click 'Search' to run the search against the specific database, OR
Click 'Question Mark' for a direct search of the database

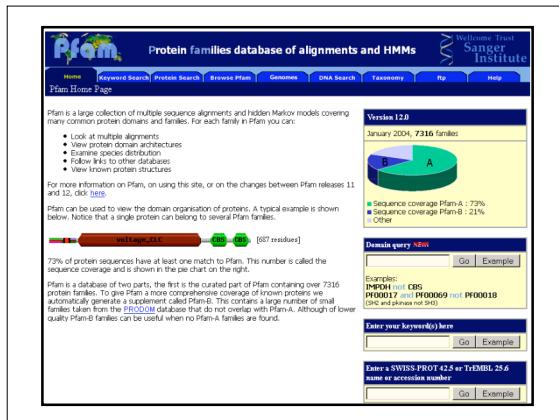
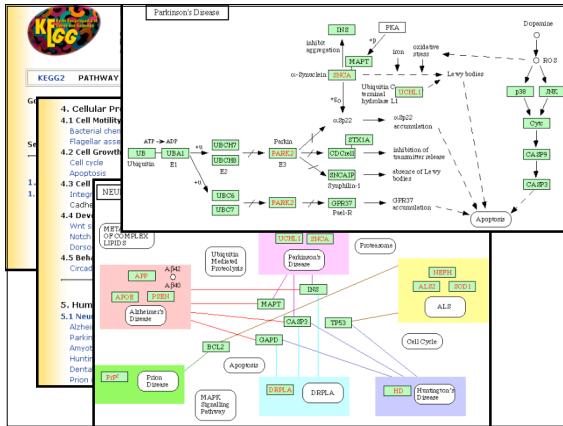
<http://www.ncbi.nlm.nih.gov/Entrez/>

Functional Annotation of Lists of Genes

KEGG
PFAM
SWISS-PROT
GO

DRAGON
DAVID
BioConductor





FUNCTION	Catalyzes the O-methylation, and thereby the inactivation, of catecholamine neurotransmitters and catechol hormones. Also shortens the biological half-lives of certain psychotropic drugs. Like L-DOPA, alpha-methyl DOPA and reserpine.
CATALYTIC ACTIVITY	catechol-O-methyltransferase + catechol + S-adenosyl-L-homocysteine → guanidino
SUPERCELLULAR LOCATION	cytosolic; membrane; plasma (intron S-COMT); Type II membrane protein (intron MB-COMT)
ALTERNATIVE NAMES	O-alternative nomenclature
EC NUMBER	EC 2.1.1.66
TISSUE SPECIFICITY	Membrane-bound COMT (dienol form) and Soluble/S-COMT, are produced by alternative splicing.
PTM	The N-terminal is blocked.
POLYMORPHISM	Two alleles, COMT ¹ or COMT ^{1H} with Val-158 and COMT ² or COMT ^{1L} with Ile-158 are responsible for a three to four-fold difference in enzymatic activity.
SIMILARITY	To other mammalian catechol-O-methyltransferase
Keywords	
Enzyme	Transferring Methyl groups; Neurotransmitter degeneration; Catecholamine metabolism; Transmembrane; Signal anchor; Magnesium; Alternative isoform; Polycomb
Protein	M0311C; AAA99927_1; [EMBL/GenBank/DBJ] [CDGeneSequence]; MG3523; AAA69328_1; [EMBL/GenBank/DBJ] [CDGeneSequence]; MG3525; AAA69329_1; [EMBL/GenBank/DBJ] [CDGeneSequence]; 2S9P; CADB1261_1; [EMBL/GenBank/DBJ] [CDGeneSequence]; BC011928; AAH1935_1; [EMBL/GenBank/DBJ] [CDGeneSequence]
PR	13756; A3459
HSSP	P22724; IVID; [HSPEI ENTRY / PUB]
Review	HEC-2223; COMT
Res	HEC-2226; COMT
Card	
Def	COMT; Homo sapiens
Domains	COMT; Homo sapiens
DM	10779; [NCBI / BAA]
DO	Q0005653; Cellular component; membrane (traceable author statement); Q0005653; Cellular component; soluble fraction (traceable author statement); Q0008117; Molecular function; O-methyltransferase activity (traceable author statement).
GO	COMT; Homo sapiens
GOA	P2194; Homo sapiens; [Entry / Catalog view]
GOB	HEC-2226; SAM; COMT
Pro	Q96KJ2; [Protein function catalog]

GENE ONTOLOGY CONSORTIUM

[What Is the Gene Ontology?](#) [Download the Ontologies](#)

The goal of the Gene Ontology™ (GO) Consortium is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. GO provides three structured *networks* of defined terms to describe gene product attributes. GO is one of the controlled vocabularies of the [Open Biological Ontologies](#).

- Submit new GO term suggestions via the [Curator Requests Tracker](#) at [SourceForge](#). Help with new item submission is available at [SourceForge.net](#)
- Send comments and questions to go@geneontology.org

Search: This search is powered by AmiGO

GO MINER Build 120

Genomics and Bioinformatics Group
NCI, NIH
Medical Informatics & Biologimaging Lab,
BME GA Tech
Georgia Institute of Technology

Home Requirements Installation Command Line Quick Start Database Help Citing Mirrors Credits

Top Doge Gene Catalog

GO 0003673: Gr...

Bi @ GO:0008110
Bi @ GO:0005525
Bi @ GO:0003673
Bi @ GO:0003673
Bi @ GO:0003673
Bi @ GO:0003673
Bi @ GO:0003673

What does it all mean biologically?

GoMiner is a tool for biological interpretation of 'omic' data – including data from gene expression microarrays. Omic experiments often generate lists of dozens or hundreds of genes that differ in expression between samples, raising the question: "What does it all mean biologically?" To answer the question, GoMiner leverages the [Gene Ontology \(GO\)](#) to identify the biological processes, functions and components associated with these lists. Instead of analyzing microarray result with a gene-by-gene approach, GoMiner classifies the genes into biologically coherent categories and assesses these categories. The insights gained through GoMiner can generate hypotheses to guide additional research.

Annotation with Bioconductor

Annotation

- One of the largest challenges in analyzing genomic data is associating the experimental data with the available **biological metadata**, e.g., sequence, gene annotation, chromosomal maps, literature.
- And making that data available for computation
- Bioconductor provides three main packages for this purpose:
 - **annotate** (**end-user**);
 - **AnnBuilder** (**developer**)
 - **annaffy** (**end-user** – will see a name change)

WWW resources

- Nucleotide databases: e.g. GenBank.
- Gene databases: e.g. Entrez Gene, UniGene.
- Protein sequence and structure databases: e.g. SwissProt, Protein DataBank (PDB).
- Literature databases: e.g. PubMed, OMIM.
- Chromosome maps: e.g. NCBI Map Viewer.
- Pathways: e.g. KEGG.
- [Entrez](#) is a search and retrieval system that integrates information from databases at NCBI (National Center for Biotechnology Information).
- if you know of some we should be using – please let us know

annotate: matching IDs

Important tasks

- Associate manufacturers or in-house probe identifiers to other available identifiers.
E.g.
 - Affymetrix IDs → Entrez Gene IDs
 - Affymetrix IDs → GenBank accession number.
- Associate probes with biological data such as chromosomal position, pathways.
- Associate probes with published literature data via PubMed (need PMID).

annotate: matching IDs

Affymetrix identifier	"41046_s_at"
HGU95A chips	
Entrez Gene ID	"9203"
GenBank accession #	"X95808"
Gene symbol	"ZNF261"
PubMed, PMID	"10486218" "9205841" "8817323"
Chromosomal location	"X", "Xq13.1"

Annotation data packages

- The Bioconductor project provides annotation data packages, that contain many different mappings to interesting data
 - Mappings between Affy IDs and other probe IDs: hgu95av2 for HGU95AV2 GeneChip series, also, hgu133a, hu6800, mgu74a, rgu34a, YG.
 - Affy CDF data packages.
 - Probe sequence data packages.
- These packages are updated and expanded regularly as new data become available.
- They can be downloaded from the Bioconductor website and also using `installDataPackage`.
- DPEexplorer: a widget for interacting with data packages.
- AnnBuilder: tools for building annotation data packages.

annotate: matching IDs

- Much of what `annotate` does relies on matching symbols.
- This is basically the role of a hash table in most programming languages.
- In R, we rely on environments.
- The annotation data packages provide R environment objects containing key and value pairs for the mappings between two sets of probe identifiers.
- Keys can be accessed using the R `ls` function.
- Matching values in different environments can be accessed using the `get` or `multiget` functions.

annotate: matching IDs

```
> library(hgu95av2)
> get("41046_s_at", env = hgu95av2ACCNUM)
[1] "X95808"
> get("41046_s_at", env = hgu95av2LOCUSID)
[1] "9203"
> get("41046_s_at", env = hgu95av2SYMBOL)
[1] "ZNF261"
> get("41046_s_at", env = hgu95av2GENENAME)
[1] "zinc finger protein 261"
> get("41046_s_at", env = hgu95av2SUMFUNC)
[1] "Contains a putative zinc-binding
motif (MYM)|Proteome"
> get("41046_s_at", env = hgu95av2UNIGENE)
[1] "Hs.9568"
```

annotate: matching IDs

```
> get("41046_s_at", env = hgu95av2CHR)
[1] "X"
> get("41046_s_at", env = hgu95av2CHRLOC)
  X
-68692698
> get("41046_s_at", env = hgu95av2MAP)
[1] "Xq13.1"
> get("41046_s_at", env = hgu95av2PMID)
[1] "10486218" "9205841" "8817323"
> get("41046_s_at", env = hgu95av2GO)
  TAS      TAS      IEA
"GO:0003677" "GO:0007275" "GO:0016021"
```

annotate: matching IDs

- Instead of relying on the general R functions for environments, new user-friendly functions have been written for accessing and working with specific identifiers.
- E.g. `getGO`, `getGODesc`, `getLL`, `getPMID`, `getSYMBOL`.

annotate: matching IDs

```
> getSYMBOL("41046_s_at", data="hgu95av2")
  41046_s_at
  "ZNF261"
> gg<- getGO("41046_s_at", data="hgu95av2")
> getGODesc(gg[[1]], "MF")
$"GO:0003677"
"DNA binding activity"
> getLL("41046_s_at", data="hgu95av2")
  41046_s_at
  9203
> getPMID("41046_s_at", data="hgu95av2")
  $"41046_s_at"
[1] 10486218 9205841 8817323
```

annotate: querying databases

- The `annotate` package provides tools for
- Searching and processing information from various WWW biological databases
 - GenBank,
 - LocusLink,
 - PubMed.
 - Regular expression searching of PubMed abstracts.
 - Generating nice HTML reports of analyses, with links to biological databases.

annotate: WWW queries

- Functions for querying WWW databases from R rely on the `browseURL` function
`browseURL("www.r-project.org")`
- Other tools: `HTMLPage`, `class`, `getTDRows`,
`getQueryLink`, `getQuery4UG`, `getQuery4LL`,
`makeAnchor`.
- The `XML` package is used to parse query results.

annotate: querying GenBank

www.ncbi.nlm.nih.gov/Genbank/index.html

- Given a vector of GenBank accession numbers or NCBI UIDs, the `genbank` function
 - opens a browser at the URLs for the corresponding GenBank queries;
 - returns an `XMLdoc` object with the same data.

```
genbank("X95808", disp="browser")
http://www.ncbi.nlm.nih.gov/Genbank/query.fcgi?db=nonredundant&cmd=Search&d=NCBInet&t=X95808
genbank(1430782, disp="data",
        type="uid")
```

annotate: querying LocusLink

www.ncbi.nlm.nih.gov/LocusLink/

- **locuslinkByID:** given one or more LocusIDs, the browser is opened at the URL corresponding to the first gene.

locuslinkByID("9203")
<http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?I=9203>

- **locuslinkQuery:** given a search string, the results of the LocusLink query are displayed in the browser.

locuslinkQuery("zinc finger")
<http://www.ncbi.nlm.nih.gov/LocusLink/list.cgi?Q=zinc.finger&ORG=Hs&V=0>

- **getQuery4LL.**

annotate: querying PubMed

www.ncbi.nlm.nih.gov

- For any gene there is often a large amount of data available from PubMed.
- The **annotate** package provides the following tools for interacting with PubMed
 - **pubMedAbst:** a class structure for PubMed abstracts in R.
 - **pubmed:** the basic engine for talking to PubMed (`pmidQuery`).

annotate: pubMedAbst class

Class structure for storing and processing PubMed abstracts in R

- pmid
- authors
- abstText
- articleTitle
- journal
- pubDate
- abstUrl

annotate: high-level tools for querying PubMed

- `pm.getabst`: download the specified PubMed abstracts (stored in XML) and create a list of `pubMedAbst` objects.
- `pm.titles`: extract the titles from a list of PubMed abstracts.
- `pm.abstGrep`: regular expression matching on the abstracts.

annotate: PubMed example

```
pmid <- get("41046_s_at", env=hgu95aPMID)
pubmed(pmid, disp="browser")

http://www.ncbi.nih.gov/entrez/query.fcgi?tool=biocore&cmd=Retrieve&db=PubMed&list\_uids=10486218%2c9205841%2c8817323

absts <- pm.getabst("41046_s_at", base="hgu95a")
pm.titles(absts)
pm.abstGrep("retardation", absts[[1]])
```

annotate: PubMed HTML report

- The new function `pmAbst2HTML` takes a list of `pubMedAbst` objects and generates an HTML report with the titles of the abstracts and links to their full page on PubMed.

```
pmAbst2HTML(absts[[1]],
            filename="pm.html")
```

The screenshot shows a Microsoft Internet Explorer window with the following details:

- Title Bar:** BioConductor Abstract List - NetPage
- Address Bar:** http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&term=bioconductor abstract
- Search Bar:** Search
- Navigation Buttons:** Back, Forward, Stop, Home, Refresh
- Toolbar:** Favorites, File, Edit, View, Insert, Search, Metrics, Part, Security, Stop
- Status Bar:** "What's This?"

The main content area displays a list of abstracts from the BioConductor Abstract List:

Article Title	Publication Date
Conditional targeting of the TNSA region upstream of OSMRbeta.	Nov 2002
Induced pluripotent stem cells: reprogramming and clonal expansion of CD34+ and CD38+ hematopoietic progenitors.	Sep 2002
A human cDNA clone of OSMR γ with polycomb recruitment for silencing of the gene.	May 2002
Inhibition of long-term self-renewal by overexpression of the DNA base excision repair gene XPCR1.	Jul 2002
The human OSMR mRNA is nuclear and its association with nuclear matrix is gene specific.	Jul 2002
Osmr $\alpha\beta\gamma$ transcripts encode phosphorylated and associates with the nuclear matrix and interacts with the nuclear envelope.	Aug 2002
Osmr $\alpha\beta\gamma$ mRNA polymorphisms and methylation as environmental factors of stomach cancer risk in Chinese.	Jan 2002
Identification of OSMR $\alpha\beta\gamma$ gene promoter polymorphisms with high cancer risk.	Apr 2002
Phosphatidylserine and sphingomyelin content increase the human OSM $\alpha\beta\gamma$ mRNA expression.	Mar 2002
Expression of human OSM receptor is reduced and associated with non-Hodgkin lymphoma.	Jan 2002
Prognostic and therapeutic location of human OSMR α .	Month 1999
Expression and differential spatial distribution of two major forms of human OSM receptor, OSMR α and OSMR β .	May 1999
Genomic organization and alternative splicing of the OSMR α gene. It is involved in the rescue of T-cell responses in cultured L3T4 cells.	Jan 1998
Structural organization of a human gene for OSM receptor α subunit (OSMR α) in 3'-flanking region of the gene.	Mar 1997
Human OSM receptor α gene: a human Y染色体 recessive marker on XIST2, containing a functional OSMR α gene.	Oct 1997
Molecular cloning and functional analysis of a human OSM receptor α gene.	Jul 1997
Cloning and characterization of OSMR α , a human homolog of the OSMR α gene of <i>Zacharowski et al.</i>	Jul 1997

annotate: analysis reports

- A simple interface, [htmlpage](#), can be used to generate an HTML report of analysis results.
 - The page consists of a table with one row per gene, with links to Entrez Gene, Affymetrix, SwissProt, UniGene, or OMIM.
 - Entries can include various gene identifiers and statistics.

pmAbst2html
function from
annotate package

pm.html

10

annaffy

- Provides simplified mappings between Affymetrix IDs and annotation data
- Relies on chip-level annotation packages created by AnnBuilder
- Supplies functions to produce mappings for almost all environments in a given annotation package

annaffy:Interactive

```
> symbol <- aafSymbol(probids, "hgu95av2")
> getText(symbol)
[1] "COL1A2" "FLT3"  "BDNF"  "CD19"  "GSTT2" "FGFR2" "IL18"
[8] "IFNB1" "RAB5B" "TAF11"

> gos <- aafGO(probids, "hgu95av2")
> gos[[3]]
An object of class "aafGO"
[[1]]
An object of class "aafGOItem"
@id  "GO:0007399"
@name "neurogenesis"
@type "Biological Process"
@evid "TAS"
```

annaffy:Interactive

```
> gbs <- aafGenBank(probids, "hgu95av2")
> getURL(gbs[[3]])
[1]
"http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=search&db=nucleotide&term=M61176%5BACCN%5D&doptcmdl=GenBank"
> browseURL(getURL(gbs[[3]]))

This will open a browser pointing to this particular GenBank ID
```

annaffy:Non-interactive

- Primary function of annaffy is to produce very nice HTML or text tables
 - These tables can contain:
 - Links to databases
 - Statistics
 - Expression measures
 - Color-coded to intensity for easy viewing

annaffy:HTML Table

```

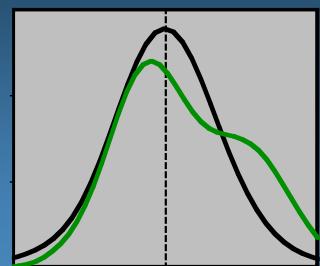
> aaf.handler()
[1] "Probe"
[4] "Function"
[7] "GenBank"
[10] "UniGene"
[13] "Pathway"

> annatable <- aafTableAnn(probids[1:10], "hgu95av2", aaf.handler())[c(1:3,
10)]
> stattable <- aafTable("t-stat" = rnorm(10), "p-value" = runif(10))
> exportable <- aafTableIn(afExpr, probeids = probids[1:10])
> table <- merge(annatable, stattable)
> table <- merge(table, exportable)
> saveHTML(table, "faketable.HTML", title = "Some Fake Results")

```

Supplemental Slides

Analysis of Functional Gene Groups



Functional Gene/Protein Networks

DIP

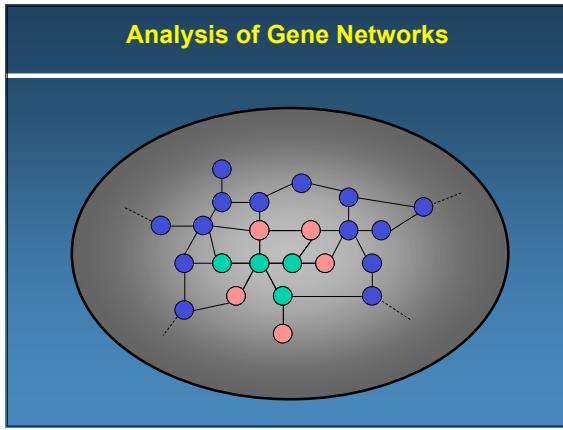
BIND

MINT

HPRD

PubGene

Predicted Protein Interactions



Blueprint

B MOUNT SINAI BIOCLOUD

bioinformatics	Contact Us	Help	Search	Mount Sinai
Home	Sequenator	Protein/Folding	Downloads	Exhibitions/News
About	Products	Services	Technical Support	Publications/People
Bind				Jobs

About BIND

The Biocolecular Interaction Network Database (BIND) is a collection of interactions between biological molecules. It contains over 100,000 high-quality, manually curated, and peer-reviewed data submissions and hand-curated collections gathered from the scientific literature.

BIND is an interaction database with three classifications of interactions that associate with each other to form **Interactions**, **Molecular Complexes**, or **Pathways**. An interaction is defined as a specific sequence of two or more interactions that occur sequentially.

A molecular complex is a collection of one or more molecules that interact with each other to form a stable, functional object that is believed to occur in a living organism. A biological object can be a protein, RNA, DNA, ligand, molecular complex, gene, photon, or an unclassified entity. Molecular complexes are often found in the literature and have been shown experimentally and published in at least one peer-reviewed journal. A molecular complex can be associated with experimental evidence that supports the associated interaction.

Interactions are the basic units of BIND and can be linked together to form molecular complexes or pathways.

A molecular complex is a collection of two or more molecules that associate with each other to form a stable, functional object. In BIND, these are represented as a molecular complex object, formed by linking two or more interaction objects. Molecular complexes can be annotated with additional information such as complex topology and pathway (represented with additional objects) involved in their formation.

A pathway is a collection of one or more interactions that occur sequentially within a living organism. In BIND, pathway records that are formed by linking two or more interaction objects. Pathways are often associated with the progression of some stage of the cell cycle; the pathway exists independently of the cell cycle.

The object-oriented design of BIND has allowed diversity of interactions in a format that is efficient and used by software engineers. The BIND community of curators and users has been growing exponentially since it inception in 1997. The BIND database is now the largest public resource for biological interactions used by researchers worldwide to validate a hypothesis or predict a new finding.

Current BIND Database Statistics

Record Type	Count
Interactions (All)	77573
Interactions (Spoke Model)	7654
Molecular Complexes	1522
Pathways	10
Organisms Represented	863
Sequences (Gifs)	32385
Publications	9283

Accession Query

Start:	NCBI Taxonomy ID
End:	Submit

Results for NCBI Taxonomy ID(s): 9606

TaxonID 9606 is found in:
 [1] 9606 BING Interaction Links | [4] BING Pathway Links | [116 BING Molecular Complex Links]

9606 is the Taxonomy ID for Homo Sapiens

MINT		STATISTICS	
		Last update March 31, 2004, 5:00 am	
		Number of interactions	
		Number of interactions	
		Number of interactions on two yeast partner	
		12594	
		Number of interactions with at least one mammalian partner	
		3013	
		Total interactions	
		4096	
		Number of proteins	
		4662	
		Number of species	
		Number of species	
		Number of interactions by organism	
		interactions with at least one <i>Arabidopsis thaliana</i> (Mouse-ear cress) partner	
		26	
		interactions with at least one <i>Bacteriophage T7</i> partner	
		27	
		interactions with at least one <i>Bos taurus</i> (Bovine) partner	
		105	
		interactions with at least one <i>Caenorhabditis elegans</i> partner	
		4500	
		interactions with at least one <i>Cani familiaris</i> (Dog) partner	
		20	
		interactions with at least one <i>Drosophila melanogaster</i> (Fly) partner	
		2046	
		interactions with at least one <i>Escherichia coli</i> partner	
		50	
		interactions with at least one <i>Escherichia coli</i> O157:H7 partner	
		29	
		interactions with at least one <i>Gallus gallus</i> (Chicken) partner	
		38	
		interactions with at least one <i>Homo sapiens</i> (Human) partner	
		2054	
		interactions with at least one <i>Mus musculus</i> (Mouse) partner	
		1002	
		interactions with at least one <i>Oryctolagus cuniculus</i> (Rabbit) partner	
		20	
		interactions with at least one <i>Ovis aries</i> (Sheep) partner	
		22	
		interactions with at least one <i>Rattus norvegicus</i> (Rat) partner	
		301	
		interactions with at least one <i>Saccharomyces cerevisiae</i> (Baker's yeast) partner	
		12611	
		interactions with at least one <i>Sus scrofa</i> (Pig) partner	
		20	
		interactions with at least one <i>Vaccinia virus</i> (strain WR) partner	
		30	
		interactions with at least one <i>Xenopus laevis</i> (African clawed frog) partner	
		28	

Human Protein Reference Database

You are at Home

Highlights

OMIM entries
List of 154 human genes implicated in human diseases listed in OMIM. The first interface in Man (OMIM) database have been annotated in HPRD.

Statistics

Proteins: 8782
Protein-Protein: 13000

Molecular Function: Enzyme: Acyltransferase 60
Biological Process: Energy and metabolism 60

Catechol-O-methyltransferase

Showing Interaction

ALTERNATE NAMES DISEASES PMID & SUBSTRATES EXTERNAL LINKS

SEQUENCE INTERACTIONS

INTERACTIONS

INTERACTING PROTEIN

Name of Interactor	Domain in Interacting Protein	Region	Domain	Region	Experiment Type	Type
Arginine-converting enzyme					In vivo	Direct

Predicted Human Protein Interactions

Research
A first-draft human protein-interaction map
Ben Lehner and Andrew G Fraser

Address: The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.
Correspondence: Andrew G Fraser. E-mail: agf@sanger.ac.uk

Published: 13 August 2004
Genome Biology 2004, 5:R63
The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2004/5/9/R63>

Received: 7 May 2004
Revised: 23 June 2004
Accepted: 20 July 2004

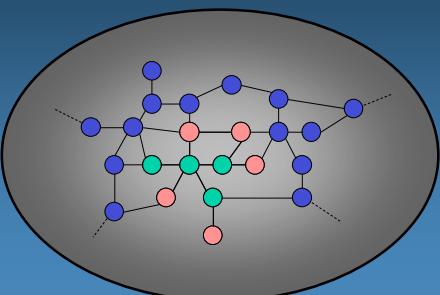
Predicted Human Protein Interactions

Used high-throughput protein interaction experiments from fly, worm, and yeast to predict human protein interactions.

Human protein interaction is predicted if both proteins in an interaction pair from other organism have high sequence homology to human proteins.

>70K Hs interactions predicted
>6K Hs genes

Analysis of Gene Networks



NCBI Web Links

<http://www.ncbi.nlm.nih.gov>
<http://www.ncbi.nlm.nih.gov/Entrez/>
<http://www.ncbi.nlm.nih.gov/Genbank/>
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide>
<http://www.ncbi.nlm.nih.gov/dbEST/>
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein>
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome>
<http://www.ncbi.nlm.nih.gov/LocusLink/>
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene>
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>
<http://www.ncbi.nlm.nih.gov/PubMed/>
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=cdd>
<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp>
<http://www.ncbi.nlm.nih.gov/SNP/>
<http://utils.ncbi.nlm.nih.gov/entrez/query/static/advancedentrez.html>
<http://www.ncbi.nlm.nih.gov/geo/>
<http://www.ncbi.nlm.nih.gov/RefSeq/>

```

graph TD
    Nucleotide["NUCLEOTIDE:  
http://genome.ucsc.edu/  
http://www.embl-heidelberg.de/  
http://www.ensembl.org/  
http://www.ebi.ac.uk/  
http://www.gdb.org/  
http://blipinfo.weizmann.ac.il/cards/index.html  
http://www.genome.ucsc.edu/cgi-bin/men德尔法律/geneSpI"/>
    Pathways["PATHWAYS and NETWORKS:  
http://www.genome.ad.jp/kegg/  
http://ftp.ncbi.nlm.nih.gov/KEGG/KEGG14/KEGG14.tar.gz  
http://ftp.ebi.ac.uk/pub/KEGG/KEGG14/KEGG14.tar.gz  
http://www.genome.ad.jp/norftp/  
http://dip.doe-mbl.edu/  
http://dip.doe-mbl.edu/dip/Download.cgi  
http://www.blueprint.org/bind/  
http://www.blueprint.org/bind/bind_downloads.html  
http://160.80.34.4/minitindex.php  
http://160.80.34.4/minit/release/main.php  
http://www.hprd.org/  
http://www.hprd.org/FAQ?selectedtab=DOWNLOAD+REQUESTS  
http://www.pubgene.org/ [also .com]"]
    Protein["PROTEIN:  
http://us.expasy.org/  
http://www.expasy.org/  
http://www.sanger.ac.uk/Software/Pfam/  
http://www.sanger.ac.uk/Software/Pfam/ftp.shtml  
http://smart.embl-heidelberg.de/  
http://www.ebi.ac.uk/interpro/  
http://us.expasy.org/prosite/  
http://us.expasy.org/databases/prosite/"]
    Bioconductor["Bioconductor  
http://www.bioconductor.org/  
http://rapsid.missouri.gov/david/  
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi  
http://idecovar.net.nih.gov/idecovar/index.jsp  
http://pubmatrx.pro.missouri.gov/  
http://pivesherlab.kennedykrieger.org/dragon.htm"]
    MoreLinks["More Web Links  
http://www.ncbi.nlm.nih.gov/david/  
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi  
http://idecovar.net.nih.gov/idecovar/index.jsp  
http://pubmatrx.pro.missouri.gov/  
http://pivesherlab.kennedykrieger.org/dragon.htm"]

```