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.

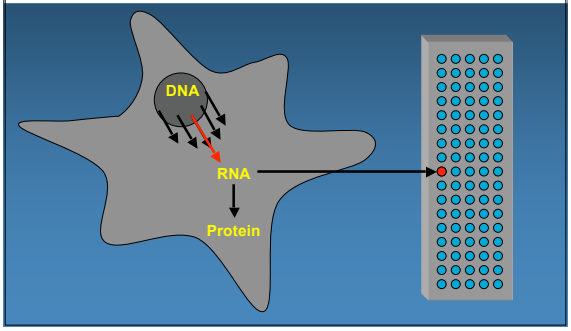
Tasks necessary for gene expression analysis:

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.

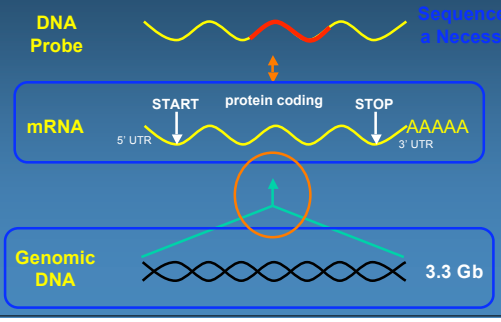
Cellular Biology, Gene Expression, and Microarray Analysis



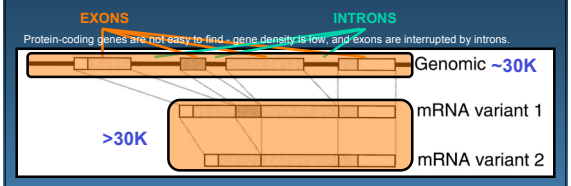
Gene: Protein coding unit of genomic DNA with an mRNA intermediate.

~30K genes

Sequence is a Necessity

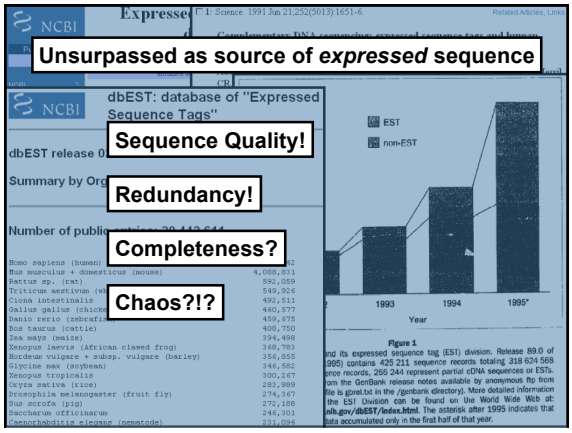


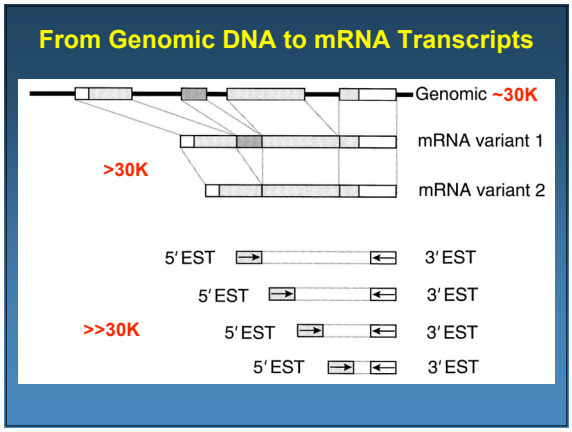
From Genomic DNA to mRNA Transcripts

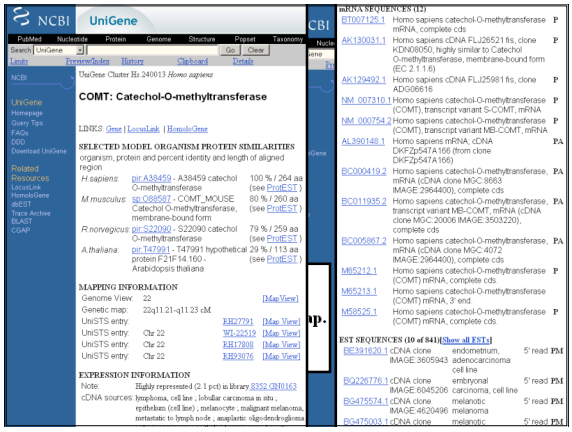


Alternative splicing
Alternative start & stop sites in same RNA molecule
RNA editing & SNPs

Transcript coverage
Homology to other transcripts
Hybridization dynamics
3' bias


Unsurpassed as source of *expressed* sequence
dbEST: database of "Expressed Sequence Tags"
Sequence Quality!
Redundancy!
Completeness?
Chaos?!?
EST
non-EST
1993
1994
1995*
Year


From Genomic DNA to mRNA Transcripts
Genomic ~30K
mRNA variant 1
mRNA variant 2
>30K
>>30K
5′ EST
3′ EST


UniGene
COMT: Catechol-O-methyltransferase
SELECTED MODEL ORGANISM PROTEIN SIMILARITIES
MAPPING INFORMATION
EXPRESSION INFORMATION
mRNA SEQUENCES (12)
EST SEQUENCES (10 of 841)

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):

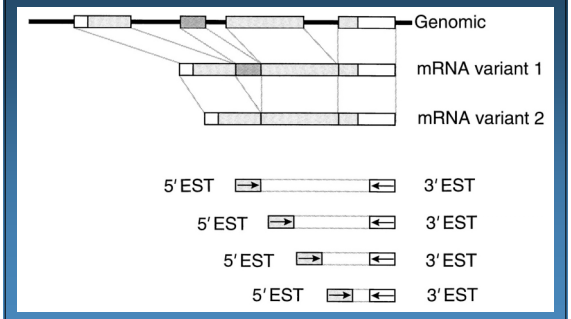
Sequence <> GenBank <> UniGene <> HomoloGene

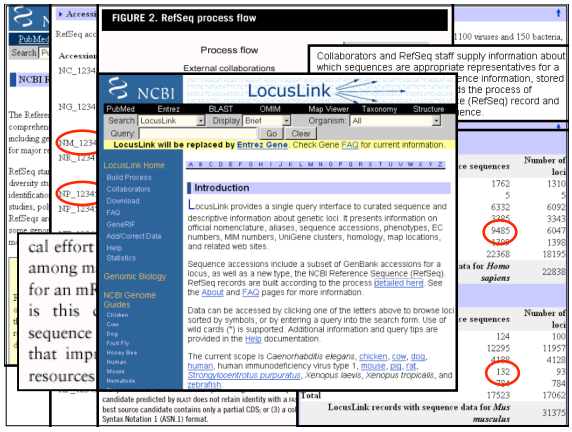
Possible mis-referencing:

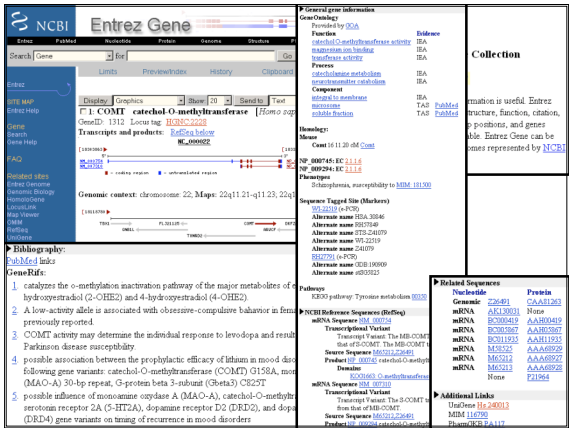
- Genomic GenBank Acc.#'s
- Referenced ID has more NT's than probe
- Old DB builds
- DB or table errors – copying and pasting 30K rows in excel …

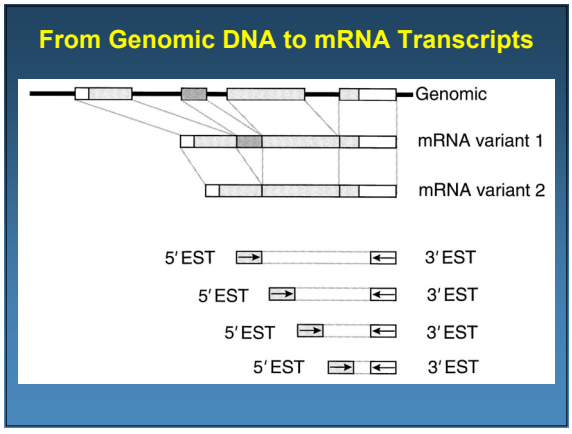
Using RefSeq's can help.

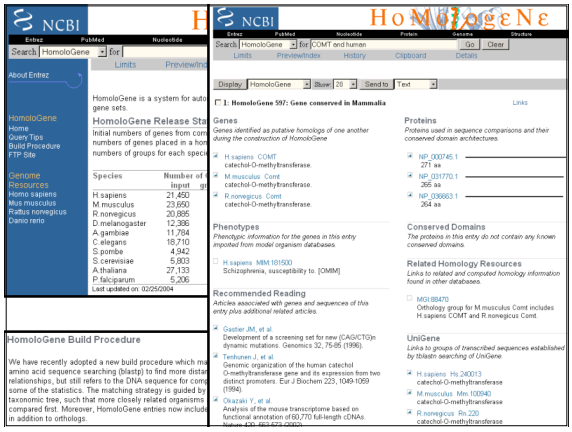
From Genomic DNA to mRNA Transcripts

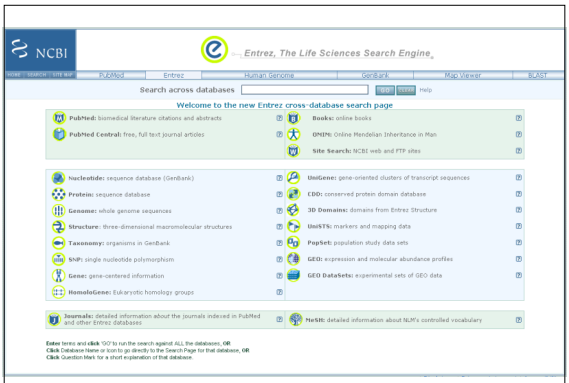





NCBI
Entrez Gene
Collection


From Genomic DNA to mRNA Transcripts
Genomic
mRNA variant 1
mRNA variant 2
5′ EST
3′ EST


NCBI
HomoloGene
HomoloGene Build Procedure



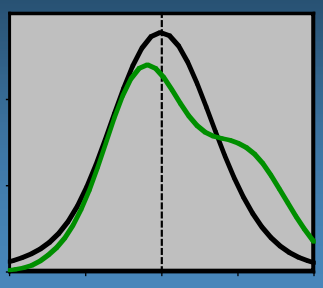
http://www.ncbi.nlm.nih.gov/Entrez/

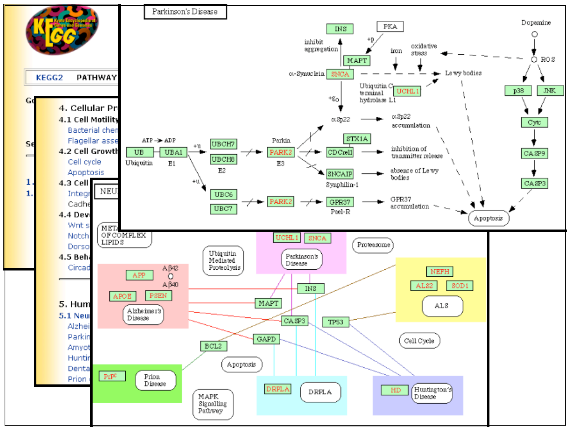
Functional Annotation of Lists of Genes

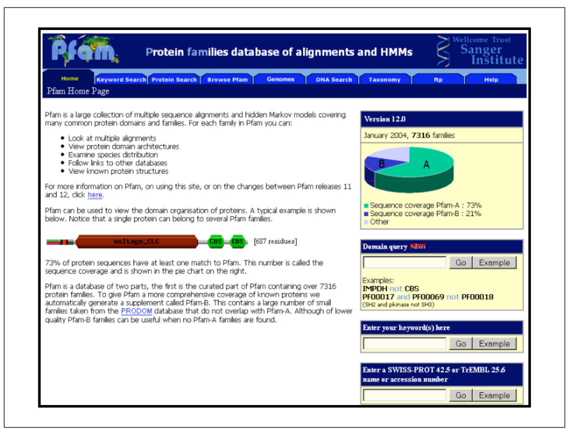
KEGG
PFAM
SWISS-PROT
GO

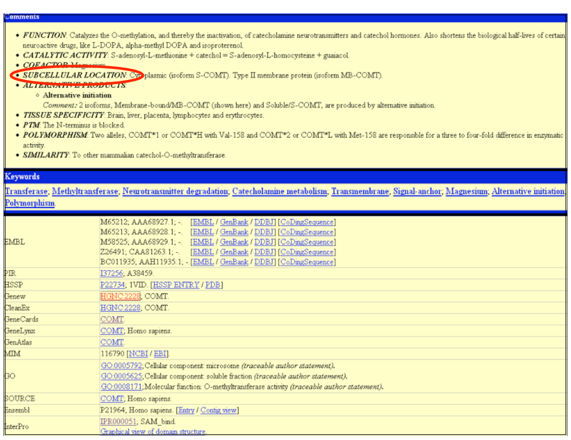
DRAGON
DAVID
BioConductor

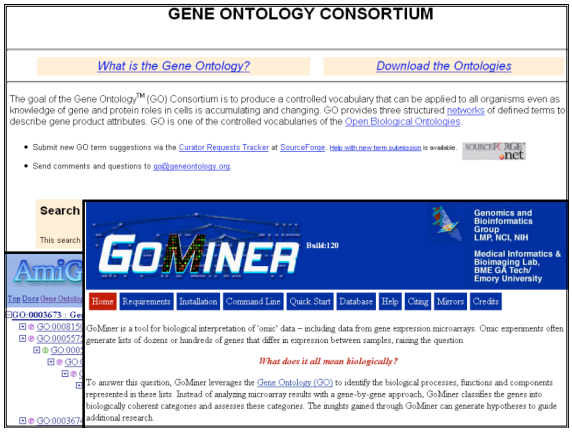
Analysis of Functional Gene Groups




KEGG2 PATHWAY
Parkinson's Disease
4. Cellular Pr
4.1 Cell Motility
4.2 Cell Growth
Cell cycle
Apoptosis
5. Hum
Dopamine
ROS
Lewy bodies
Ubiquitin
Apoptosis
Proteasome
Alzheimer's Disease
Parkinson's Disease
ALS
Cell Cycle
Prion Disease
MAPK Signalling Pathway
DRPLA
Huntington's Disease


Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute
Home
Keyword Search
Protein Search
Browse Pfam
Genomes
DNA Search
Taxonomy
ftp
Help
Pfam Home Page
Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. For each family in Pfam you can:
Look at multiple alignments
View protein domain architectures
Examine species distribution
Follow links to other databases
View known protein structures
Version 12.0
January 2004, 7316 families
Sequence coverage Pfam-A : 73%
Sequence coverage Pfam-B : 21%
Other
Domain query
Go
Example
Enter your keyword(s) here
Enter a SWISS-PROT 42.5 or TrEMBL 25.6 name or accession number


Comments
FUNCTION
CATALYTIC ACTIVITY
COFACTOR
SUBCELLULAR LOCATION
ALTERNATIVE PRODUCTS
TISSUE SPECIFICITY
PTM
POLYMORPHISM
SIMILARITY
Keywords
Transferase, Methyltransferase, Neurotransmitter degradation, Catecholamine metabolism, Transmembrane, Signal-anchor, Magnesium, Alternative initiation, Polymorphism
EMBL
PIR
HSSP
Genew
CleanEx
GeneCards
GeneLynx
GenAtlas
MIM
GO
SOURCE
Ensembl
InterPro



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 HGU95A chips	"41046_s_at"
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`.
- `DPExplorer`: 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.nih.gov/entrez/query.fcgi?tool=bioconductor&cmd=Search&db=Nucleotide&term=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.nih.gov/LocusLink/LocRpt.cgi?l=9203

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

```
locuslinkQuery("zinc finger")
```

http://www.ncbi.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")
```

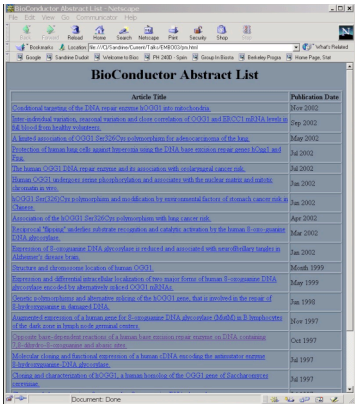
http://www.ncbi.nih.gov/entrez/query.fcgi?tool=bioconductor&cmd=Retrieve&db=PubMed&list_uids=10486218%2c9205841%2c8817323

```
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")
```

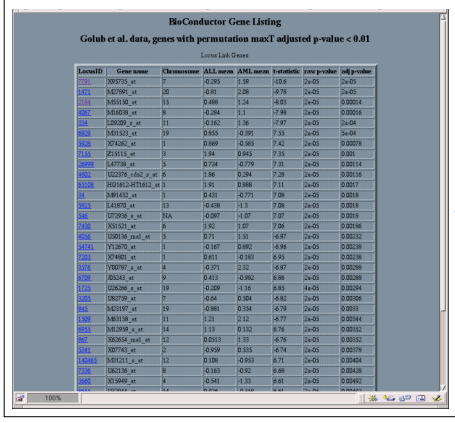


pmAbst2html function from **annotate** package

pm.html

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.**



htmlpage function from **annotate** package

genelist.html

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] "COL11A2" "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=nucleoti
de&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

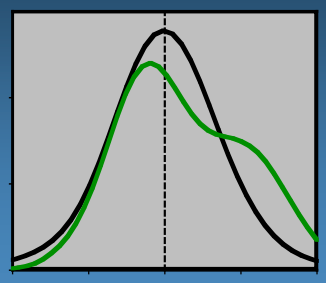
```
> aaf.handler()
 [1] "Probe"            "Symbol"            "Description"
 [4] "Function"          "Chromosome"         "Chromosome Location"
 [7] "GenBank"           "LocusLink"          "Cytoband"
[10] "UniGene"           "PubMed"            "Gene Ontology"
[13] "Pathway"

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

Some Fake Results

Probe	Symbol	Description	UniGene	t-stat	p-value	Ctrl 1	Ctrl 2	Ctrl 3	Ctrl 4	Exp 1	Exp 2	Exp 3	Exp 4
1027_at	COL11A2	collagen, type XI, alpha 2	Hs.390171	-1.45594	0.670306	6.78988	6.95535	7.01725	7.03271	6.87766	6.91575	7.08804	6.97782
1065_at	FLT3	fms-related tyrosine kinase 3	Hs.507590	-0.220808	0.631552	4.75681	4.90082	4.58507	4.435	4.66974	4.82038	4.96138	4.74451
1088_at	BDNF	brain-derived neurotrophic factor	Hs.502182	-0.51622	0.865794	2.94753	3.12375	3.14666	3.40514	3.18515	3.22673	3.09794	3.73931
1096_g_at	CD19	CD19 antigen	Hs.96023	-0.674515	0.409788	7.80312	6.4948	7.71022	8.07147	7.69353	6.82171	7.22894	7.35078
1099_s_at	GSTT2	glutathione *S*-transferase theta 2	Hs.1581	0.741629	0.830797	3.75509	3.82736	3.65213	3.63096	3.63285	3.61646	3.66221	3.80546
1144_at	FGFR2	fibroblast growth factor receptor 2 (bacteria-expressed kinase, keratinocyte growth factor receptor, craniofacial dysostosis 1, Crouzon syndrome, Pfeiffer syndrome, Jackson-Weiss syndrome)	Hs.533683	0.560647	0.0985497	5.05875	5.24738	4.94823	5.2071	5.11945	5.44314	4.80758	4.77517
1165_at	IL18	interleukin 18 (interferon-gamma-inducing factor)	Hs.83077	1.31626	0.460907	5.07169	5.09287	4.41281	4.79389	4.76078	5.07755	4.79351	4.72389
1214_s_at	IFNB1	interferon, beta 1, fibroblast	Hs.93177	1.20016	0.829474	4.81134	4.91741	4.78125	4.55761	4.77543	4.64006	4.8158	4.84777
1221_at	RAB5B	RAB5B, member RAS oncogene family	Hs.77690	1.33816	0.115838	5.66729	5.87053	5.65699	5.4345	5.63707	5.85552	5.52444	5.65217
131_at	TAF11	TAF11 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 28kDa	Hs.112444	-0.71489	0.16326	5.8401	6.12606	6.3367	5.80433	5.89106	6.00132	6.35426	5.71814

Supplemental Slides

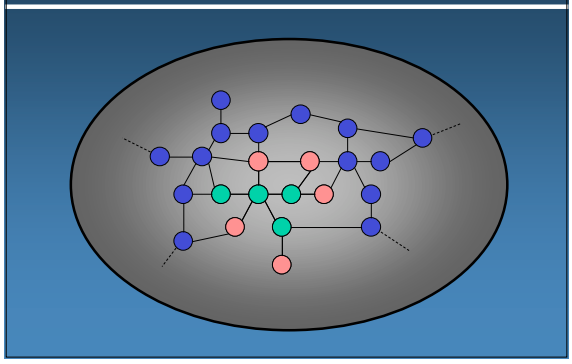
Analysis of Functional Gene Groups

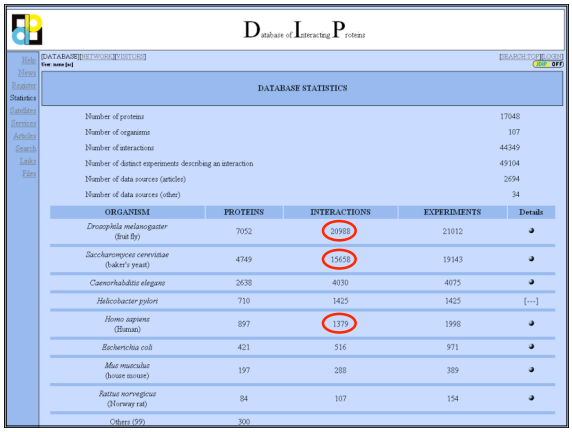


Functional Gene/Protein Networks

DIP
BIND
MINT
HPRD
PubGene
Predicted Protein Interactions

Analysis of Gene Networks

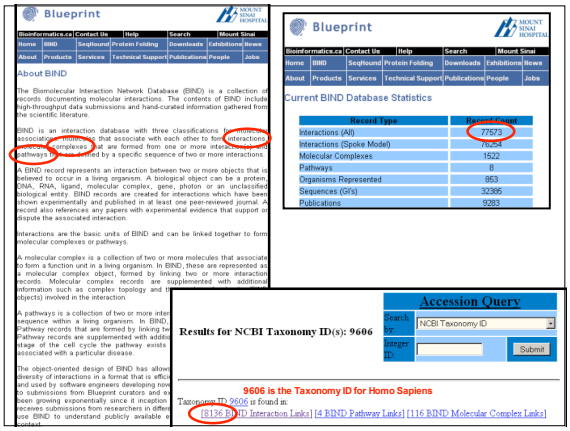




DATABASE STATISTICS

Number of proteins	17048
Number of organisms	107
Number of interactions	44349
Number of distinct experiments describing an interaction	49104
Number of data sources (articles)	2694
Number of data sources (other)	34

ORGANISM	PROTEINS	INTERACTIONS	EXPERIMENTS	Details
Drosophila melanogaster (fruit fly)	7052	20988	21012	
Saccharomyces cerevisiae (baker's yeast)	4749	15658	19143	
Caenorhabditis elegans	2638	4030	4075	
Helicobacter pylori	710	1425	1425	[---]
Homo sapiens (Human)	897	1379	1998	
Escherichia coli	421	516	971	
Mus musculus (house mouse)	197	288	389	
Rattus norvegicus (Norway rat)	84	107	154	
Others (99)	300			



Current BIND Database Statistics

Record Type	Record Count
Interactions (All)	77573
Interactions (Spoke Model)	76254
Molecular Complexes	1522
Pathways	8
Organisms Represented	853
Sequences (GIs)	32385
Publications	9283

Results for NCBI Taxonomy ID(s): 9606

Accession Query

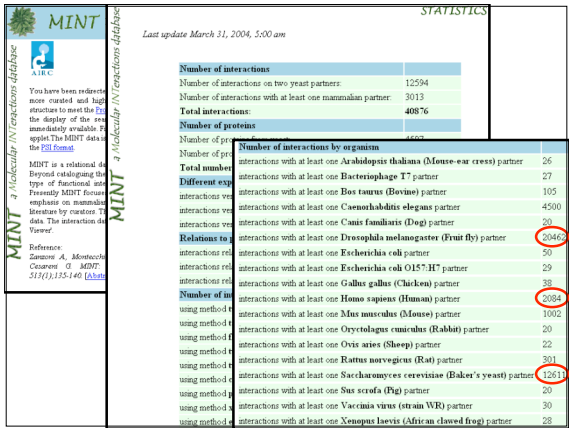
Search by: NCBI Taxonomy ID

Integer ID: Submit

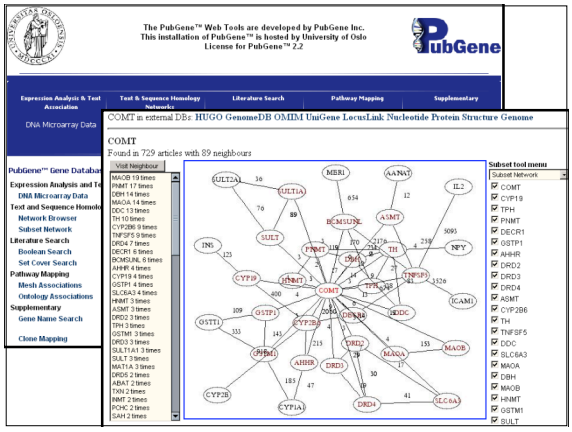
9606 is the Taxonomy ID for Homo Sapiens

Taxonomy ID 9606 is found in:

[8136 BIND Interaction Links] [4 BIND Pathway Links] [116 BIND Molecular Complex Links]


MINT
a Molecular INteractions database
STATISTICS
Last update March 31, 2004, 5:00 am
Number of interactions
Number of interactions on two yeast partners: 12594
Number of interactions with at least one mammalian partner: 3013
Total interactions: 40876
Number of proteins
Number of interactions by organism
interactions with at least one Arabidopsis thaliana (Mouse-ear cress) partner 26
interactions with at least one Bacteriophage T7 partner 27
interactions with at least one Bos taurus (Bovine) partner 105
interactions with at least one Caenorhabditis elegans partner 4500
interactions with at least one Canis familiaris (Dog) partner 20
interactions with at least one Drosophila melanogaster (Fruit fly) partner 20462
interactions with at least one Escherichia coli partner 50
interactions with at least one Escherichia coli O157:H7 partner 29
interactions with at least one Gallus gallus (Chicken) partner 38
interactions with at least one Homo sapiens (Human) partner 2084
interactions with at least one Mus musculus (Mouse) partner 1002
interactions with at least one Oryctolagus cuniculus (Rabbit) partner 20
interactions with at least one Ovis aries (Sheep) partner 22
interactions with at least one Rattus norvegicus (Rat) partner 301
interactions with at least one Saccharomyces cerevisiae (Baker's yeast) partner 12611
interactions with at least one Sus scrofa (Pig) partner 20
interactions with at least one Vaccinia virus (strain WR) partner 30
interactions with at least one Xenopus laevis (African clawed frog) partner 28


Human Protein Reference Database
Query
Browse
Blast
Pathways
FAQs
Highlights
Statistics
Proteins 6762
Protein-Protein 13000
Catechol-O-methyltransferase
Molecular Function Enzyme: Acyltransferase
Biological Process Energy and metabolism
INTERACTIONS
Angiotensin I converting enzyme
In Vivo
Direct


The PubGene™ Web Tools are developed by PubGene Inc.
This installation of PubGene™ is hosted by University of Oslo
License for PubGene™ 2.2
PubGene
DNA Microarray Data
COMT in external DBs: HUGO GenomeDB OMIM UniGene LocusLink Nucleotide Protein Structure Genome
COMT
Found in 729 articles with 89 neighbours
Visit Neighbour
Subset tool menu
Subset Network
COMT
CYP19
TPH
PNMT
DECR1
GSTP1
AHHR
DRD2
DRD3
DRD4
ASMT
CYP2B6
TH
TNFSF5
DDC
SLC6A3
MAOA
DBH
MAOB
HNMT
GSTM1
SULT

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

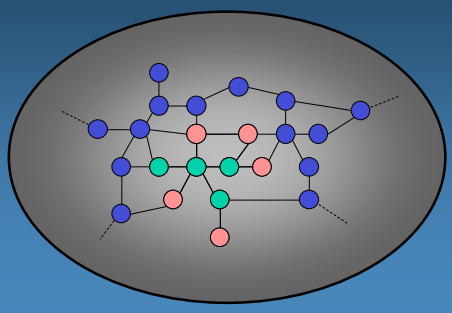
Predicted Human Protein Interactions

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.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=gene
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://eutils.ncbi.nlm.nih.gov/entrez/query/static/advancedentrez.html
http://www.ncbi.nlm.nih.gov/geo/
http://www.ncbi.nlm.nih.gov/RefSeq/

FTP:
ftp://ftp.ncbi.nlm.nih.gov/
ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene
ftp://ftp.ncbi.nih.gov/pub/HomoloGene/

NUCLEOTIDE:

http://genome.ucsc.edu/
http://www.embl-heidelberg.de/
http://www.ensembl.org/
http://www.ebi.ac.uk/
http://www.gdb.org/
http://bioinfo.weizmann.ac.il/cards/index.html
http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl

PROTEIN:

http://us.expasy.org/
ftp://us.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/
ftp://us.expasy.org/databases/prosite/

PATHWAYS and NETWORKS:

http://www.genome.ad.jp/kegg/
ftp://ftp.genome.ad.jp/pub/kegg/ (http://www.genome.ad.jp/anonftp/)
http://dip.doe-mbi.ucla.edu
http://dip.doe-mbi.ucla.edu/dip/Download.cgi
http://www.blueprint.org/bind/
http://www.blueprint.org/bind/bind_downloads.html
http://160.80.34.4/mint/index.php
http://160.80.34.4/mint/release/main.php
http://www.hprd.org/
http://www.hprd.org/FAQ?selectedtab=DOWNLOAD+REQUESTS
http://www.pubgene.org/ (also .com)

More Web Links

http://www.bioconductor.org/
http://apps1.niaid.nih.gov/david/
http://www.geneontology.org/
http://discover.nci.nih.gov/gominer/index.jsp
http://pubmatrix.grc.nia.nih.gov/
http://pevsnerlab.kennedykrieger.org/dragon.htm