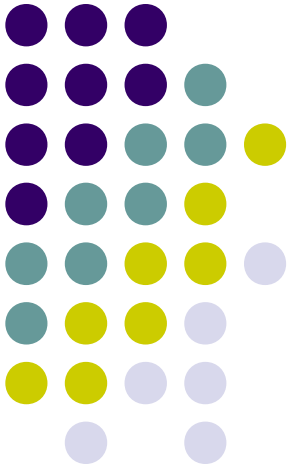
Lecture #10

Classification

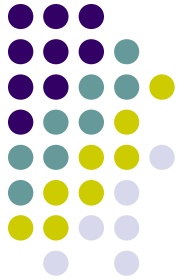
Annotation (GO/KEGG Pathways)



Annotation

- **Annotation** uses biological research databases of many types to interpret analytic results.
- It closes the gap between knowledge of sequence and knowledge of function.
- This knowledge is widely scattered and encoded in many different formats.
- Challenge is to develop tools that can be used to access these data, and integrate them.

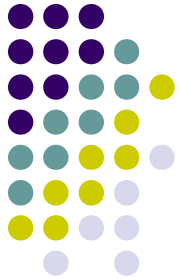
Some Uses of Annotation



- Perform dimension-reduction at an early stage
- Introduce constraints on relationships between statistical model parameters during model building
- To interpret discovered patterns at the conclusion
- …

The data evolve rapidly, and it is important to track the version number.

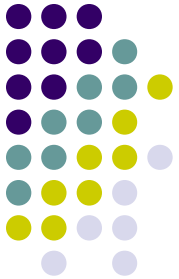
Annotation Resources



Major classes of resources:

- Genes and gene products
- Pathways and gene clusters
- Biochemical pathway elements
- Scientific literature
- Assay-oriented resources that link probe identifiers with associated sequence catalog entries

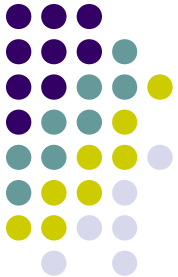
Bioconductor



Provides access to annotation through:

1. Direct real-time queries to Web services
   - May not be reproducible later
1. Curated, downloadable modules
   - Reproducible, but may have obsolete information

National Center for Biotechnology Information (NCBI)



EntrezGene: a catalog of genetic loci

http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

UniGene defines sequence clusters

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

RefSeq: a non-redundant data set of sequences, including DNA, transcripts and proteins

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

Other resources

Enzyme Commision (EC) numbers are assigned to different enzymes and linked to genes though EntrezGene.

Gene Ontology (GO) is a structured vocabulary of terms describing gene products

PubMed tool for working with published journal articles related to medicine and health

...

Pathway annotation packages

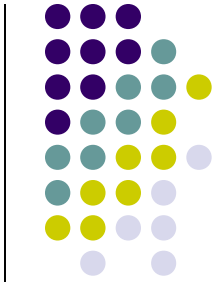
The **Kyoto Encyclopedia of Genes and Genomes (KEGG):**

can map EntezGene IDs to KEGG pathways.

The **cancer Molecular Analysis Project (cMAP):**

a project providing software and data for exploration of data relevant to cancer. Pathway data is from Biocarta and KEGG

Bioconductor

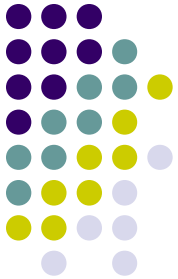


MetaData node of the Bioconductor portal lists R packages that encode annotation resources.

http://www.bioconductor.org/data/metaData.html

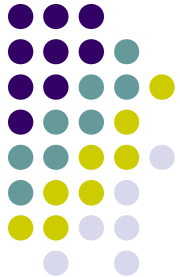
Important!! The associations and data are likely to undergo constant change and you will want to update the meta-data packages on a regular basis.

Annotating a platform: HG-U95av2



```
> biocLite('hgu95av2.db')
> library("hgu95av2.db")
Loading required package: org.Hs.eg.db
> ls("package:hgu95av2.db")
 [1] "hgu95av2"              "hgu95av2_dbconn"       "hgu95av2_dbfile"
 [4] "hgu95av2_dbInfo"       "hgu95av2_dbschema"     "hgu95av2ACCNUM"
 [7] "hgu95av2ALIAS2PROBE"   "hgu95av2CHR"           "hgu95av2CHRLENGTHS"
[10] "hgu95av2CHRLOC"        "hgu95av2CHRLOCEND"     "hgu95av2ENSEMBL"
[13] "hgu95av2ENSEMBL2PROBE" "hgu95av2ENTREZID"      "hgu95av2ENZYME"

[16] "hgu95av2ENZYME2PROBE"  "hgu95av2GENENAME"      "hgu95av2GO"
[19] "hgu95av2GO2ALLPROBES"  "hgu95av2GO2PROBE"      "hgu95av2MAP"
[22] "hgu95av2MAPCOUNTS"     "hgu95av2OMIM"          "hgu95av2ORGANISM"
[25] "hgu95av2ORGPKG"        "hgu95av2PATH"          "hgu95av2PATH2PROBE"
[28] "hgu95av2PFAM"          "hgu95av2PMID"          "hgu95av2PMID2PROBE"
[31] "hgu95av2PROSITE"       "hgu95av2REFSEQ"        "hgu95av2SYMBOL"
[34] "hgu95av2UNIGENE"       "hgu95av2UNIPROT"
```

Number of Mapped Probes



> **hgu95av2()**
Quality control information for hgu95av2:

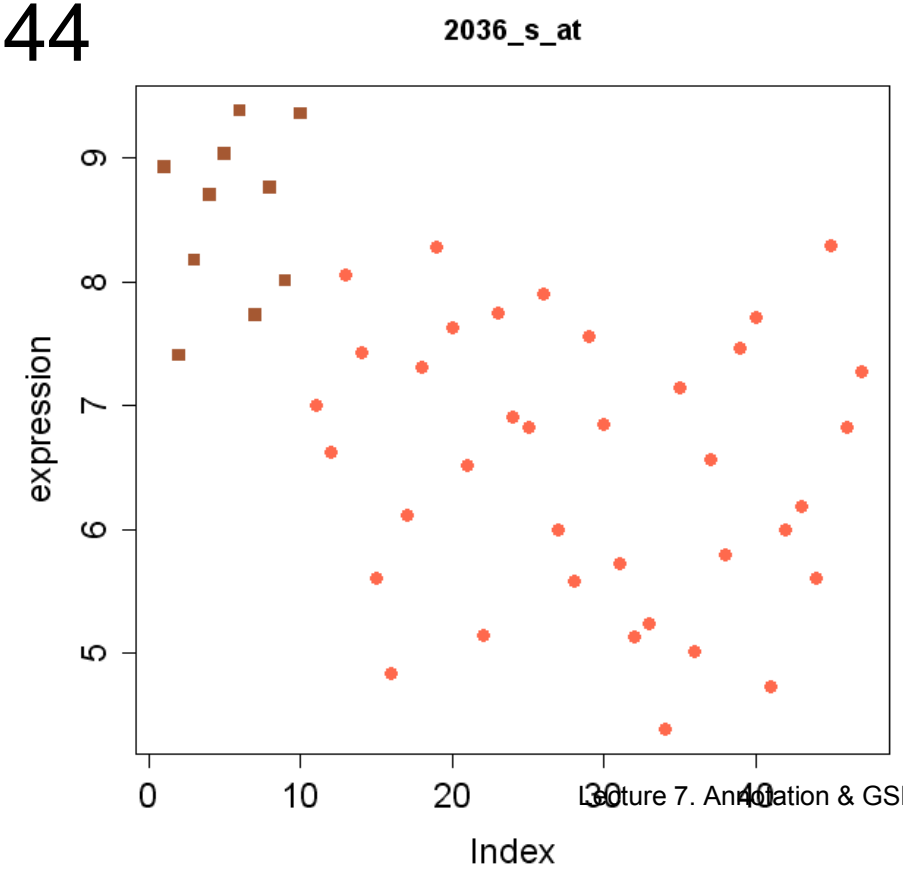
This package has the following mappings:
hgu95av2ACCNUM has 12625 mapped keys (of 12625 keys)
hgu95av2ALIAS2PROBE has 37969 mapped keys (of 110391 keys)
hgu95av2CHR has 11721 mapped keys (of 12625 keys)
hgu95av2CHRLENGTHS has 93 mapped keys (of 93 keys)
hgu95av2CHRLOC has 11609 mapped keys (of 12625 keys)
hgu95av2CHRLOCEND has 11609 mapped keys (of 12625 keys)
hgu95av2ENSEMBL has 11468 mapped keys (of 12625 keys)
hgu95av2ENSEMBL2PROBE has 9072 mapped keys (of 19971 keys)
hgu95av2ENTREZID has 11724 mapped keys (of 12625 keys)
hgu95av2ENZYME has 2046 mapped keys (of 12625 keys)
hgu95av2ENZYME2PROBE has 747 mapped keys (of 912 keys)
hgu95av2GENENAME has 11724 mapped keys (of 12625 keys)
…

Additional Information about this package:

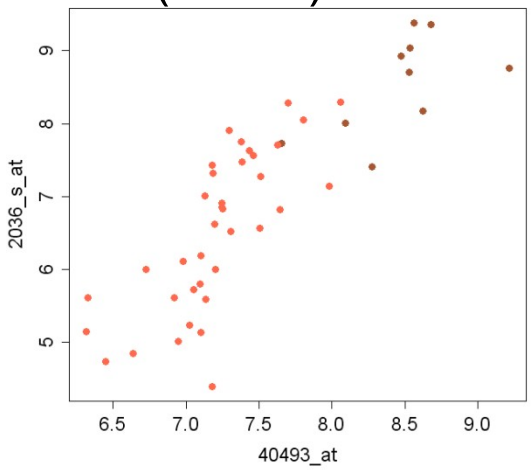
DB schema: HUMANCHIP_DB
DB schema version: 2.1
Organism: Homo sapiens
Date for NCBI data: 2010-Mar1
Date for GO data: 20100320
Date for KEGG data: 2010-Feb28
Date for Golden Path data: 2009-Jul5
Date for IPI data: 2010-Feb10
Date for Ensembl data: 2010-Mar3

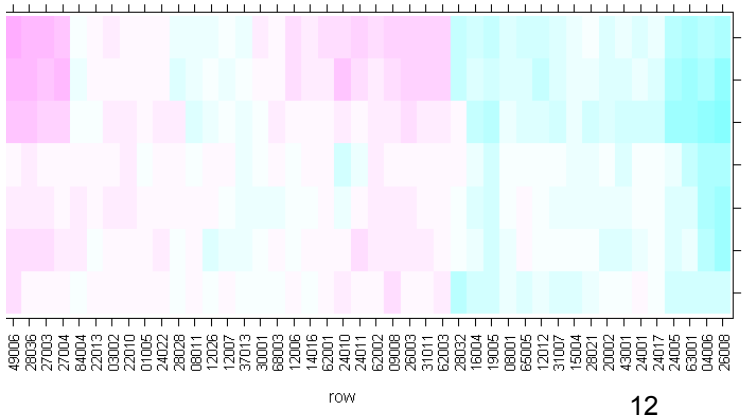
Identify Probes by EntrezGene ID

- Expression profile for 1 probe from CD44

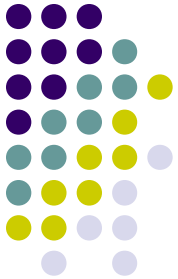


- Expression profile for multiple probe sets (CD44)





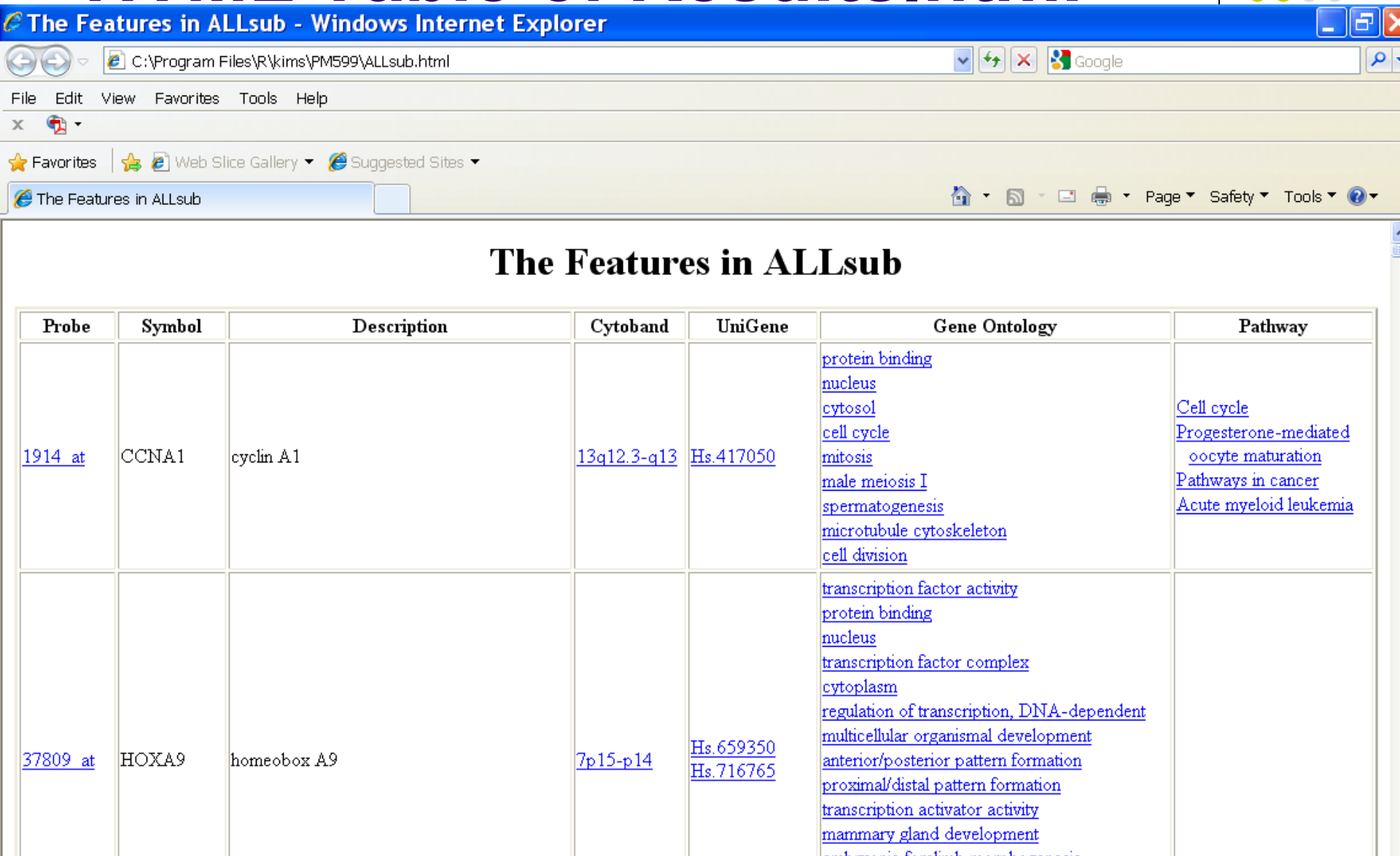
Characterize Gene Lists



1. Identify subset of interesting genes.

   e.g. use differential expression analysis
   - Perform non-specific filtering
   - Univariate T-tests
   - Pick subset of top ranking *genes*
2. Characterize subset
   - Produce linked HTML table of results
   - Distribution across chromosomes
   - Distribution across predefined gene sets

HTML Table of Results.html

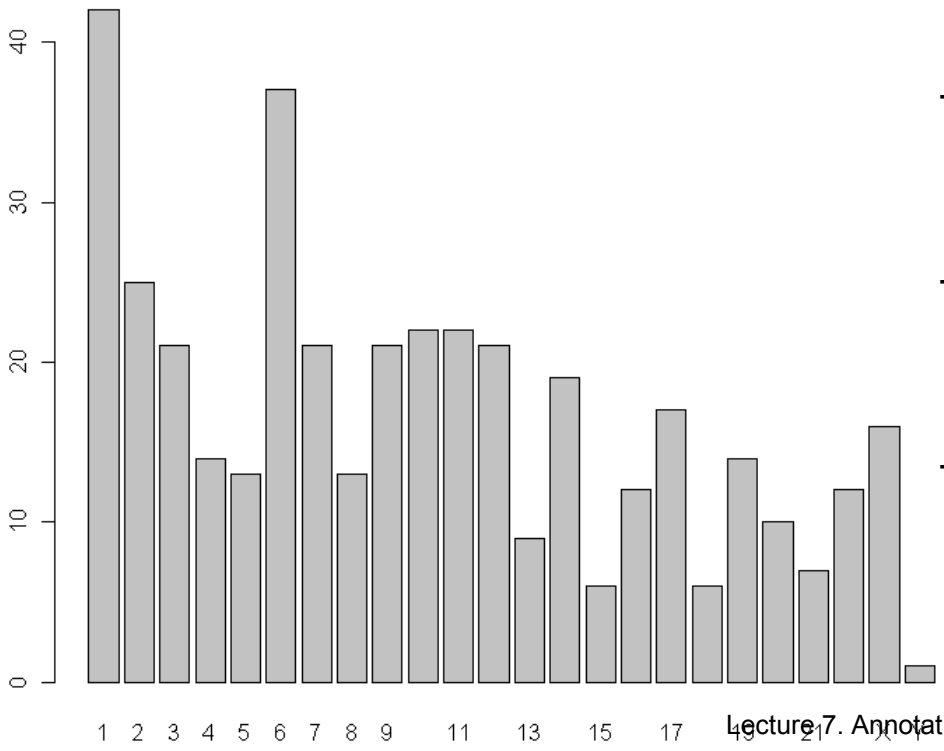


The Features in ALLsub

Probe	Symbol	Description	Cytoband	UniGene	Gene Ontology	Pathway
1914_at	CCNA1	cyclin A1	13q12.3-q13	Hs.417050	protein binding nucleus cytosol cell cycle mitosis male meiosis I spermatogenesis microtubule cytoskeleton cell division	Cell cycle Progesterone-mediated oocyte maturation Pathways in cancer Acute myeloid leukemia
37809_at	HOXA9	homeobox A9	7p15-p14	Hs.659350 Hs.716765	transcription factor activity protein binding nucleus transcription factor complex cytoplasm regulation of transcription, DNA-dependent multicellular organismal development anterior/posterior pattern formation proximal/distal pattern formation transcription activator activity mammary gland development	

Gene list vs Chromosome

- Number of genes by Chromosome (n=400 genes)

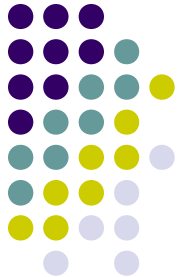


- Test for association with chromosome (overrepresentation)

```
Chromosomes: 1-22,X,Y
Top400     1   2   3   4   5   6   7
  TRUE    42  25  21  14  13  37  21
  FALSE  881 538 455 318 378 482 387

Top400     8   9  10  11  12  13  14
  TRUE    13  21  22  22  21   9  19
  FALSE  294 304 299 497 467 150 267

Top400    15  16  17  18  19  20 21
  TRUE     6  12  17   6  14  10  7
  FALSE  244 351 501 123 530 220 92

Top400    22   X  Y
  TRUE    12  16  0
  FALSE  235 365 16
```

Fisher's exact Test p=0.09

Gene Set Analysis



Goal: Use predefined sets of genes in order to better interpret results from an experiment.

Predefined sets can be based on functional annotation, or prior experiments:

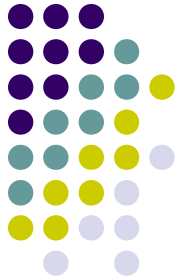
- Gene Ontology Annotation project (GOA)
- KEGG
- Chromosome bands
- Protein domains,…

Approach: Test for over- or under-representation of gene sets in your gene subset.

Warning! Multiple testing is a concern.

Gene Ontology (GO) Consortium

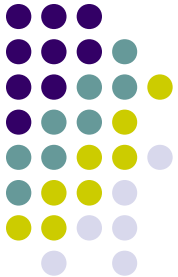
http://www.geneontology.org/GO.teaching.resources.shtml#post (Clark)



Objectives: To build **controlled vocabularies** that allow researchers to describe gene products in a consistent way.

1) To support the **annotation** of genomes, genes and gene products to these ontologies;
2) To provide **open and public access** to the ontologies;
3) To extend the **community** of people using GO.

Three Ontologies in GO

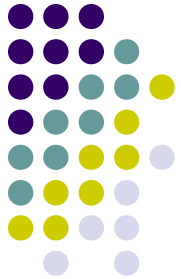


The GO Consortium produces three ontologies covering the concepts that could be described as:

- ***Molecular Function***: elemental activity or task: DNA binding
- ***Biological Process***: broad objective or goal: mitosis, signal transduction.
- ***Cellular Component***: location or complex: nucleus, ribosome

http://www.geneontology.org/GO.teaching.resources.shtml#post

Gene Ontology (GO)



GO ontologies are structured as directed acyclic graphs (DAGs), that represent a network in which each term may be the ***child*** of more than 1 ***parent***

Child terms are more specific than parents.

Child-parent relationship can be either

- *is a* relation

  e.g. “nuclear chromosome” is a child of “chromosome”

- *is a (part of)* relation

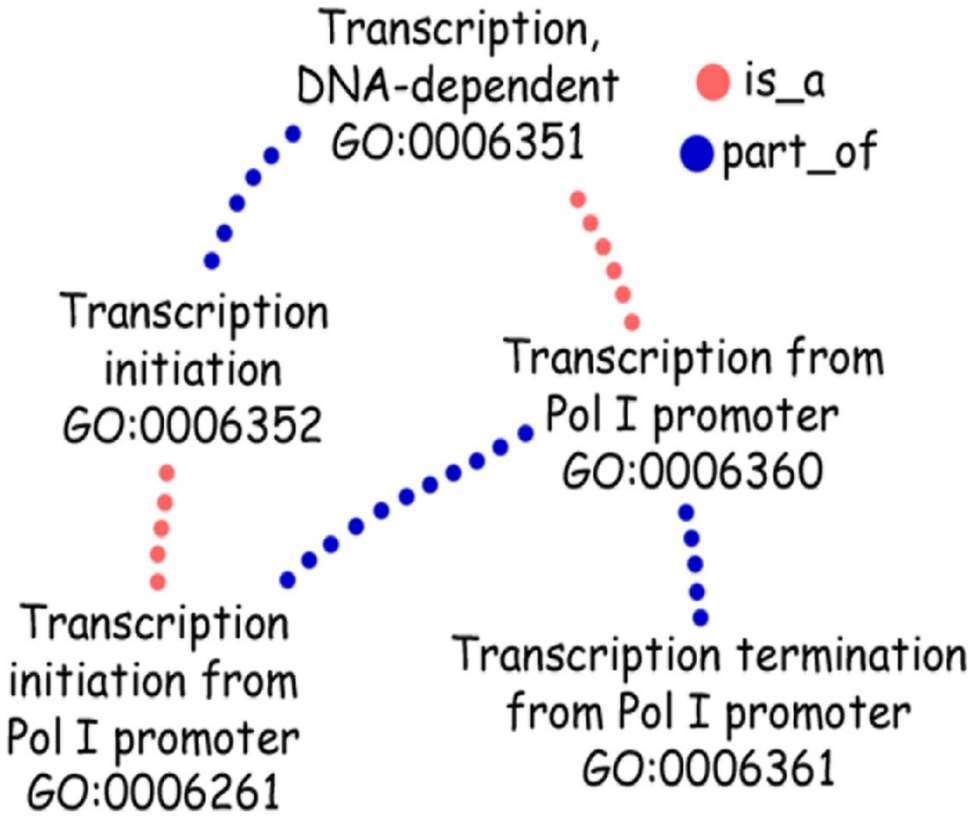
  e.g. “nucleus” is a child of “cell”

What is in a GO term?

term: transcription initiation

id: GO:0006352
(GO:7 digit code)

definition: Processes involved in starting transcription, the synthesis of RNA by RNA polymerases using a DNA template.



http://www.geneontology.org/GO.teaching.resources.shtml#post

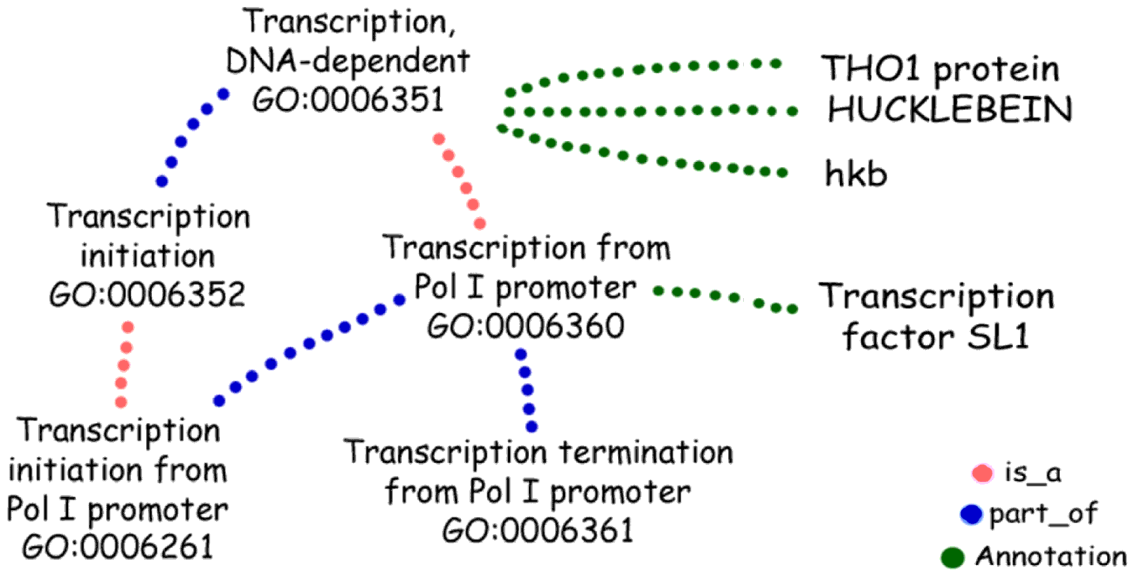
Annotation of Gene Products

Mapping gene products to terms is not part of GO.

The gene products are **electronically and then manually annotated** to appropriate gene products.

See Gene Ontolo

This annotation shows that HUCKLEBEIN protein is involved in the process of DNA-dependent transcription. Term characteristics are inherited, so Transcription factor SL1 is understood also to be involved in DNA-dependent transcription and its parents.



http://www.geneontology.org/GO.teaching.resources.shtml#post

Gene Set Analysis using GO

Goal: Determine if the frequency of genes annotated at a GO term are overrepresented in your subset of "interesting" genes.

1. Define the "gene univer
2. Classify gene by GO ter
   & subset membersh
3. Test independence using
   'Hypergeometric testir
   e.g. use Fisher's exact test
4. Repeat 2 & 3 for each GO category **(multiple**

	in Gene Subset	Not in Gene Subset
in GO term	N_{11}	N_{12}
Not in GO term	N_{21}	N_{22}

The Importance of the Gene Universe

Universe = 1,000 genes

	in Gene Subset	Not in Gene Subset
in GO	10	30
Not in GO	390	570
% in GO	2.5%	5%

Fisher's p = 0.05

Universe = 5,000 genes

	in Gene Subset	Not in Gene Subset
in GO	10	30
Not in GO	390	4570
% in GO	2.5%	0.65%

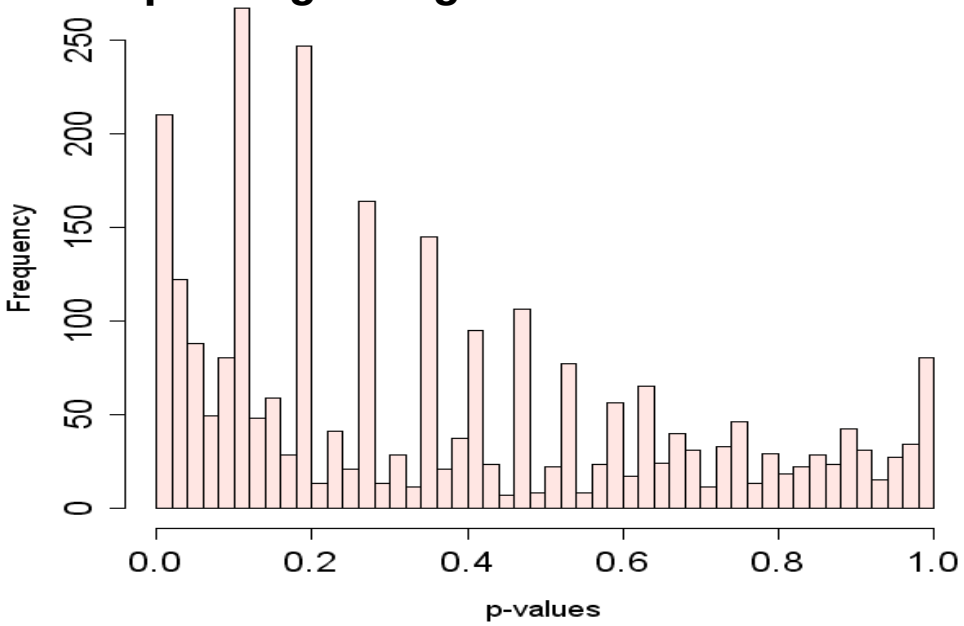
Comments:

- The gene universe should be defined by the genes that are represented on your microarray.
- It is also sensible for gene expression studies to limit the universe to genes that are expressed. (e.g. apply a non-specific filter)

See "Use and misuse of the gene ontology annotations" (Rhee et al. Nat Rev Genet, 2008)

Gene Set Analysis using GO

```
Gene to GO BP  test for over-representation
2746 GO BP ids tested (42 have p < 0.001)
Selected gene set size: 400
    Gene universe size: 3878
    Annotation package: hgu95av2
```



Gene Set Analysis using GO

Top hits for p<0.001

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0023052	6.403056e-08	1.799922	117.68953	165	1141	signaling
2	GO:0007166	2.678169e-07	2.122942	42.90872	75	416	cell surface receptor linked signaling pathway
3	GO:0006955	4.570910e-07	2.325098	29.80918	57	289	immune response
4	GO:0023033	2.816269e-06	1.949264	47.03455	77	456	signaling pathway
5	GO:0002376	4.837927e-06	1.945569	44.45591	73	431	immune system process
6	GO:0023060	4.949857e-05	1.577090	100.67045	134	976	signal transmission

Gene Set Enrichment Analysis (GSEA)

Use: To find biological themes in gene sets

Approach: Use a continuous-valued score, e.g. t-statistic, and see whether its values are associated with the gene sets of interest.

For the 2-sample scenario:

Let K denote the gene set

|K| the number of genes in the gene set

t_k the t-statistic for differential expression for gene $k \in K$

$$z_k = \frac{1}{\sqrt{|K|}} \sum_{k \in K} t_k$$

Test statistic:

If genes are independent, under H_0, $z_k \sim N(0,1)$.

Gene Set Enrichment Analysis

Comment: Despite making the same strong assumption we've made before about the genes behaving independently, the statistic seems to lead to reasonable results.

The difference from Hypergeometric testing is that we don't need to classify genes as differentially expressed or not.

Strength: increase statistical power of analyses by aggregating the signal across groups of related genes.

Pathways as Gene Sets

KEGG (Kyoto Encyclopedia of Genes and Genomes) provides mappings from genes to pathways.

In Bioconductor, we can map probes from microarrays to KEGG pathways and ask if we our differential expression analysis

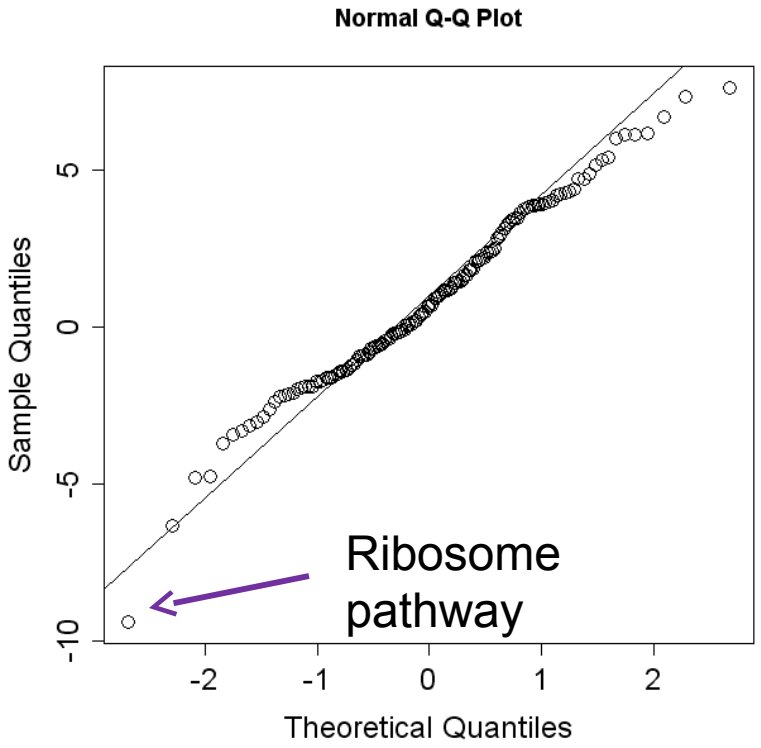
Pathway	Gene 1	Gene 2	Gene 3	…	Gene n
A	1	1	0	0	1
B	1	0	1		0
…					
ZZ	0	0	1		1

Example: ALL (Ch13)

1. Preprocess ALL data
   - filter on IQR
   - require gene in EntrezGene
   - restrict to 1 probe/gene

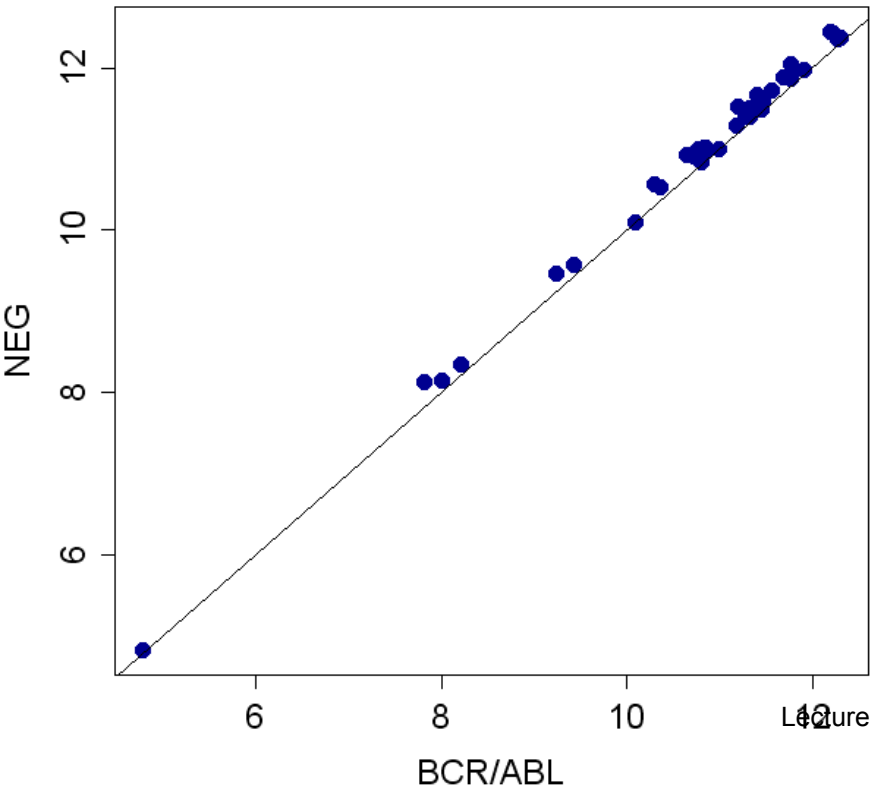
     using max IQR
2. Get all pathways that include probes found in 1.
3. Reduce data set to probes in pathways identified in 2.
4. Require pathways to

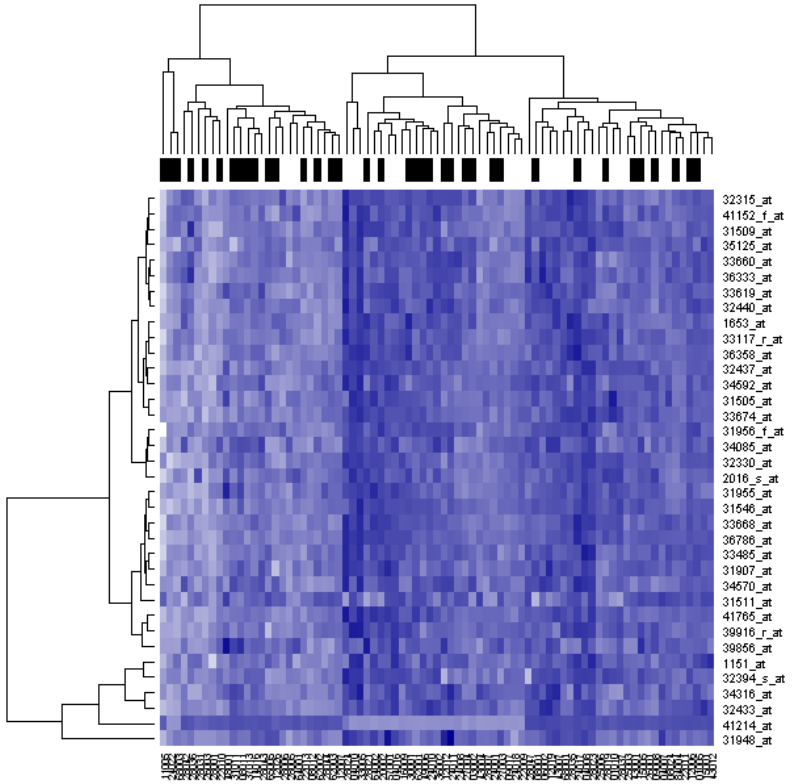
5. Compute z_k for each pathway and plot Q-Q plot.



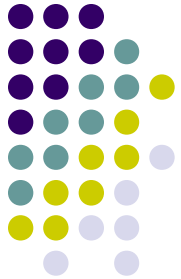
Summary Plots of Genes in the Ribosome Pathway

- Average expression level in NEG vs BCR/ABL samples
- Heatmap for genes in Ribosome Pathway





DNA-Methylation vs. Gene Expression



Some facts:

- DNA methylation is a mechanism for silencing genes
- Lots of aberrant DNA methylation is present in cancer

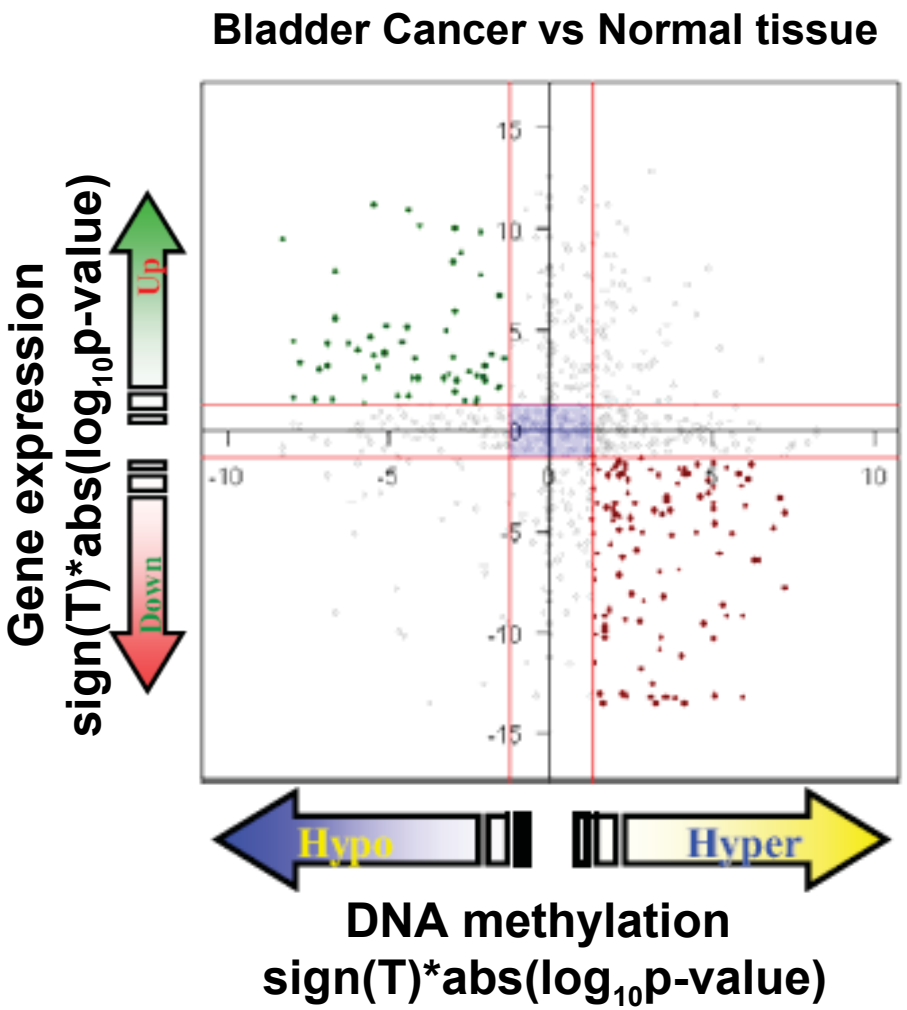
Question: In cancer, is DNA methylation associated with gene silencing?

Approach:

1. Use DNA methylation microarrays in samples of cancer and normal tissue and identify probes that are differentially methylated
2. Use gene expression microarrays in samples of cancer and normal tissue and identify probes that are differentially expressed

Compare results

DNA-Methylation vs. Gene Expression



Observed Number of Genes

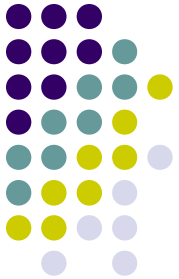
GEX	Hypo	NS	Hyper	%Total
↑	**56**	106	63	39%
NS	71	109	116	41%
↓	25	92	**117**	20%
% Total	31%	39%	30%	

Fisher's exact p=0.0001

Observed/Expected Frequencies

GEX	Hypo	NS	Hyper
↑	**1.24**	1.16	0.71
NS	1.19	0.91	1.00

Summary



Major classes of resources:

- EntrezGene ID
- Chromosome information
- GO
- KEGG
- Scientific literature
- …