BMEG3102 Bioinformatics

Lecture 9. Functional Annotations



Qi Dou Email: qidou@cuhk.edu.hk Office: Room 1014, 10/F, SHB

BMEG3102 Bioinformatics The Chinese University of Hong Kong



1. Genome annotations

2. Gene ontology and biological pathways

3. Functional enrichment analysis



Part 1

Genome Annotations

Genomic elements



- Genome annotation is the process to catalog all functional elements in a genome and characterize their properties
- Types of functional genomic elements:
 - Protein coding genes
 - Transcripts, exons, introns, coding sequences (CDSs), untranslated regions (UTRs), ...
 - Non-coding RNAs

- ...

• They are also called "biotypes"

Genomic annotations

- For each element, what do we record?
 - -Its location
 - Chromosome, start, end (strand)
 - Reference-specific (e.g., hg19 for human)
 - -Its name
 - Standard IDs (linking to other databases)
 - Official name
 - Aliases
 - -Its sequence
 - **—**…
 - -Confidence of annotation

Human lists of genomic elements



- Where to find the whole list of genomic elements in human?
 - NCBI Reference Sequence (RefSeq)
 - Integrated
 - Ensembl
 - Automatic annotation
 - <u>UCSC</u>

- <u>Gencode</u>
 - HAVANA manual annotation + Ensembl automatic annotation
 - Multiple levels of confidence
 - Level 1: verified
 - Level 2: manually annotated
 - Level 3: automatically annotated





- Comparing different sets and different versions of a set:
 - –A lot more in recent versions of Gencode due to large amount of high-throughput experimental data produced



Image credit: Harrow et al., Genome Research 22(9):1760-1774, (2012)

Annotation files

- GFF format (from http://genome.ucsc.edu/FAQ/FAQformat.html): tab-delimited. Fields:
 - 1. seqname The name of the sequence. Must be a chromosome or scaffold.
 - 2. source The program that generated this feature.
 - 3. feature The name of this type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
 - 4. start The starting position of the feature in the sequence. The first base is numbered 1.
 - 5. end The ending position of the feature (inclusive).
 - 6. score A score between 0 and 1000. If the track line useScore attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). If there is no score value, enter ".".
 - 7. strand Valid entries include '+', '-', or '.' (for don't know/don't care).
 - 8. frame If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
 - 9. group All lines with the same group are linked together into a single item.

Annotation files



- GFF format (from http://genome.ucsc.edu/FAQ/FAQformat.html): tab-delimited. Fields:
 - 1. seqname The name of the sequence. Must be a chromosome or scaffold.
 - 2. source The program that generated this feature.
 - 3. feature The name of this type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
 - 4. start The starting position of the feature in the sequence. The first base is numbered 1.
 - 5. end The ending position of the feature (inclusive).
 - 6. score A score between 0 and 1000. If the track line useScore attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). If there is no score value, enter ".".
 - 7. strand Valid entries include '+', '-', or '.' (for don't know/don't care).
 - 8. frame If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
 - 9. group All lines with the same group are linked together into a single item.
- GTF format: Similar to GFF, except that the group field is replaced by a list of attributes in <name>, <value> pairs





• Gencode v12 GTF file:

chr1 ENSEMBL exon 17021 17055 . - . gene_id "ENSG00000227232.3"; transcript_id "ENST00000430492.2"; gene_type "pseudogene"; gene_status "KNOWN"; gene_name "WASH7P"; transcript_type "unprocessed_pseudogene"; transcript_status "KNOWN"; transcript_name "WASH7P-202"; level 3; havana_gene "OTTHUMG0000000958.1";

chr1 HAVANA gene 29554 31109 . + . gene_id "ENSG00000243485.1"; transcript_id "ENSG00000243485.1"; gene_type "antisense"; gene_status "NOVEL"; gene_name "MIR1302-11"; transcript_type "antisense"; transcript_status "NOVEL"; transcript_name "MIR1302-11"; level 2; tag "ncRNA_host"; havana_gene "OTTHUMG0000000959.2";

•••

chr1 HAVANA gene 34554 36081 . - . gene_id "ENSG00000237613.2"; transcript_id "ENSG00000237613.2"; gene_type "protein_coding"; gene_status "KNOWN"; gene_name "FAM138A"; transcript_type "protein_coding"; transcript_status "KNOWN"; transcript_name "FAM138A"; level 2; havana_gene "OTTHUMG0000000960.1";

chr1 HAVANA transcript 34554 36081 . - . gene_id "ENSG00000237613.2"; transcript_id "ENST00000417324.1"; gene_type "protein_coding"; gene_status "KNOWN"; gene_name "FAM138A"; transcript_type "protein_coding"; transcript_status "KNOWN"; transcript_name "FAM138A-001"; level 2; havana_gene "OTTHUMG0000000960.1"; havana_transcript "OTTHUMT00000002842.1";

chr1 HAVANA exon 35721 36081 . - . gene_id "ENSG0000237613.2"; transcript_id "ENST00000417324.1"; gene_type "protein_coding"; gene_status "KNOWN"; gene_name "FAM138A"; transcript_type "protein_coding"; transcript_status "KNOWN"; transcript_name "FAM138A-001"; level 2; havana_gene "OTTHUMG0000000960.1"; havana_transcript "OTTHUMT0000002842.1"; chr1 HAVANA CDS 35721 35736 . - 0 gene_id "ENSG0000237613.2"; transcript_id "ENST00000417324.1"; gene_type "protein_coding"; gene_status "KNOWN"; gene_name "FAM138A"; transcript_type "protein_coding"; transcript_status "KNOWN"; transcript_name "FAM138A-001"; level 2; havana_gene "OTTHUMG0000000960.1"; havana_transcript "OTTHUMT0000002842.1"; chr1 HAVANA start_codon 35734 35736 . - 0 gene_id "ENSG0000237613.2"; transcript_id "ENST00000417324.1"; gene_type "protein_coding"; gene_status "KNOWN"; gene_name "FAM138A"; transcript_type "protein_coding"; transcript_status "KNOWN"; transcript_name "FAM138A-001"; level 2; havana_gene "OTTHUMG00000000960.1"; havana_transcript "OTTHUMT0000002842.1"; chr1 HAVANA start_codon 35734 35736 . - 0 gene_id "ENSG0000237613.2"; transcript_id "ENST00000417324.1"; gene_type "protein_coding"; gene_status "KNOWN"; gene_name "FAM138A"; transcript_type "protein_coding"; transcript_id "ENST00000417324.1"; gene_type "protein_coding"; gene_status "KNOWN"; gene_name "FAM138A"; transcript_type "protein_coding"; transcript_status "KNOWN"; gene_name "FAM138A"; transcript_type "protein_coding"; transcript_status "KNOWN"; transcript_name "FAM138A-001"; level 2; havana_gene "OTTHUMG00000000960.1"; havana_transcript "OTTHUMT00000002842.1"; "OTTHUMT00000002842.1";

Additional information



- The information in GFF/GTF files seems to be very basic
- What additional information do we know about genes?
 - -Functional annotation: gene ontology, biological pathways, etc.
- Where to find the information?
 - -Some useful sites:
 - The three main sequence databases:
 - –<u>DDBJ</u>
 - –<u>EMBL</u>
 - –<u>GenBank</u>
 - <u>GeneCards</u>
 - UCSC Genome Browser
 - <u>UniProt</u>



Part 2

Gene Ontology and Biological Pathways



- How to systematically organize knowledge about genomic elements?
 - -Define the concepts associated with standard terms
 - E.g., A gene is a genomic region that can transcribe RNA

- How to systematically organize knowledge about genomic elements?
 - -Define the concepts associated with standard terms
 - E.g., A gene is a genomic region that can transcribe RNA
 - -Define the properties of each concept
 - E.g., A gene has an official name, and zero or more aliases

- How to systematically organize knowledge about genomic elements?
 - -Define the concepts associated with standard terms
 - E.g., A gene is a genomic region that can transcribe RNA
 - -Define the properties of each concept
 - E.g., A gene has an official name, and zero or more aliases
 - -Define the relationships between different concepts
 - E.g., A protein-coding gene is a gene the RNA of which can be translated into proteins
 - E.g., An exon is part of a gene

- How to systematically organize knowledge about genomic elements?
 - -Define the concepts associated with standard terms
 - E.g., A gene is a genomic region that can transcribe RNA
 - -Define the properties of each concept
 - E.g., A gene has an official name, and zero or more aliases
 - -Define the relationships between different concepts
 - E.g., A protein-coding gene is a gene the RNA of which can be translated into proteins
 - E.g., An exon is part of a gene
- We want to define an "ontology":
 - -"The philosophical study of the nature of being, existence, or reality as such, as well as the basic categories of being and their relations" (Wikipedia)

- An ontology for gene products produced by the Gene Ontology Consortium —Most frequently-used biological ontology
- 3 sub-ontologies:
 - -Molecular function
 - Low-level functions of a gene product
 - -Biological process
 - High-level processes that a gene product is involved
 - -Cellular component
 - Which compartment can the gene product be found

- An ontology for gene products produced by the Gene Ontology Consortium –Most frequently-used biological ontology
- 3 sub-ontologies:
 - -Molecular function
 - Low-level functions of a gene product
 - -Biological process
 - High-level processes that a gene product is involved
 - -Cellular component
 - Which compartment can the gene product be found
- 2 parts:
 - -The ontologies
 - Directed acyclic graphs (DAG) Some terms have multiple parents
 - Edges indicate three types of relationships: is-a, part-of, regulates
 - -Organism-specific instances (each gene can have 0, 1 or more annotated terms)

Tree view



🗾 💭 🔪 🖓 🖓 trup://aniigo.geneo/itology.org/cgi-on/aniigo/orowse.cg/action=pius_nouextarget=t 📺 🖄 🚺 🚺 Google	
Eile Edit View Favorites Iools Help) × 🍕
p Favorites 🎬 AmiGO: Tree Browser	e • <u>S</u> afety • T <u>o</u> ols • ᠙ • 🔀
▼ Filter tree view 🛿	
Filter by ontology Filter Gene Product Counts	Set filters
Ontology Data source Species Tree view © Full © Compa	Remove all filters
biological process ASAP Arabidopsis thaliana	
cellular component AspGD Aspergillus fumig	
all : all [561268 gene products] ⊾	Actions
GO:0008150 : biological_process [428333 gene products]	Last action: Opened GO:0005488
GO:0005575 : cellular_component [391265 gene products]	Graphical View
GO:0003674 : molecular_function [460073 gene products]	Permalink
GO:0016209 : antioxidant activity [2928 gene products]	OBO
GO:000000025 : and binding [218034 gene products]	RDF-XML
GO:00000055 : altyr binding [7 gene products]	Graphviz dot
GO:0072528 : and binding [5 gene products]	
GQ:0043176 : amine binding [1714 gene products]	
GO:0003823 : antigen binding [293 gene products]	
GO:0060090 : binding, bridging [3721 gene products]	
GO:0046904 : calcium oxalate binding [1 gene product]	
🗉 📕 GO:0030246 : carbohydrate binding [6070 gene products]	
GO:0070025 : carbon monoxide binding [22 gene products]	
🗉 🗉 GO:0031406 : carboxylic acid binding [2596 gene products]	
🗉 📕 GO:0043498 : cell surface binding [379 gene products]	
Image: GO:0003682 : chromatin binding [2170 gene products]	
GO:0048037 : cofactor binding [12507 gene products]	
GO:0051871 : dihydrofolic acid binding [2 gene products]	
GO:0035731 : dinitrosyl-iron complex binding [5 gene products]	
GO:0008144 : drug binding [1002 gene products]	
GO:0097247 : epigaliocatechin 3-gallate binding [0 gene products]	
\square \square GO:0035040 : extracellular matrix binding [240 gene products]	
\square \square G0:0042562 : hormone binding [374 gene products]	
E G0:0042302 : Normone binding [374 gene products]	
G0:0046848 : hvdroxvapatite binding [9 gene products]	
■ G0:0043167 : ion binding [72695 gene products]	
■ G0:0043515 : kinetochore binding [36 gene products]	
GO:0001530 : lipopolysaccharide binding [75 gene products]	

Directed acyclic graph view





Image source: http://elbo.gs.washington.edu/courses/GS_559_11_wi/slides/11A-GeneAnnotation.pdf

Evidence codes



- Specifying how terms are used to annotate particular instances (<u>http://www.geneontology.org/GO.evidence.shtml</u>)
 - -Very important, but usually neglected

Experimental Evidence Codes EXP: Inferred from Experiment IDA: Inferred from Direct Assay **IPI: Inferred from Physical Interaction IMP:** Inferred from Mutant Phenotype IGI: Inferred from Genetic Interaction **IEP: Inferred from Expression Pattern Computational Analysis Evidence Codes** ISS: Inferred from Sequence or Structural Similarity ISO: Inferred from Sequence Orthology **ISA: Inferred from Sequence Alignment** ISM: Inferred from Sequence Model IGC: Inferred from Genomic Context IBA: Inferred from Biological aspect of Ancestor IBD: Inferred from Biological aspect of Descendant **IKR: Inferred from Key Residues IRD:** Inferred from Rapid Divergence RCA: inferred from Reviewed Computational Analysis

Author Statement Evidence Codes TAS: Traceable Author Statement NAS: Non-traceable Author Statement Curator Statement Evidence Codes IC: Inferred by Curator ND: No biological Data available Automatically-assigned Evidence Codes IEA: Inferred from Electronic Annotation Obsolete Evidence Codes NR: Not Recorded



- Gene ontology provides a simple relationship between different objects: they are both annotated with a common term
- Pathways describe detailed mechanistic relationships between the objects
 - -E.g., a metabolic pathway records how metabolites are converted to other metabolites through the actions of enzymes
- Wikipedia: "A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell"





- KEGG: Kyoto Encyclopedia of Genes and Genomes Pathway Database
- One of the most commonly used pathway databases
- Provides non-species-specific reference pathways, as well as species-specific versions
- Different types of pathways:
 - -Metabolic pathways
 - -Genetic information processing
 - -Environmental information processing
 - -Cellular processes
 - -Organismal systems
 - -human diseases
 - -Drug development

The global map for metabolic pathways





.

Glycolysis/gluconeogenesis



• The reference pathway:



Glycolysis/gluconeogenesis



• The reference pathway:



• The human version (in green):





Part 3

Functional Enrichment Analysis

Gene set analysis



- A lot of times we obtain a certain set of genes
 - -Genes with co-expression (or simultaneous differential expression)
 - Genes of which the promoters are bound by a common transcription factor
 - -Genes that harbor some mutations in patients of a certain disease
- We want to study if these genes have any relationships

17 - Carl (1)		ALPE	ELU	CDC15	SPO	RT D C	DX			
	1			ų.			В	STU2 DES1 SEP1 SPC42 CNM67 CLM4 CDC10 CDC3 CLM3 AJPC4 CDC16 SEP11	CYTOSIGULETON DRA REPAIR CYTOSIGULETON CYTOSIGULETON CYTOSIGULETON CYTOSIGUETON CYTOSIGUETON CYTOSIGUETIN CYTOSIGUETIN CHIL CYCLE CHLL CYCLE CHLL CYCLE CHLL CYCLE	PIELD PCL BOX CONCERNT EXEMPLANE PCL BOX CONCERNT ACTU TILANDER CHARTLATION ACTU TILANDER CHARTLATION ACTU TILANDER CHARTLATION DIA COLLANDER DIA COLLANDER DIA COLLANDER ADALANDERING AL MONTANIE CONFLEX SUBJECT MARINAR-INSERVICE CONFLEX SUBJECT
	\					a constant a	c	UTPL 17901 17902 17971 18975 19124 19125 19155 191	PROTECT COMPARENT OF COMPARENT	
	1	e a di si Sena di si Nata da si		i e ini Li i j	-		Ð	P096 CAF16 XES1 SMD3 TAF40 PEF30 PEF1 F3F24 STD1 N2M1 P2F15 SLS1	THAN PROCESSING TRANSPORT WEAR SPACE WEAR SPACE TRANSPORTFOR UNER, TOTAL TRANSCRIPTION WEAR, STARLINT WEAR, STARLINT WEAR, STARLINT WEAR, STARLINT WEAR, STARLINT WEAR, STARLINT WEAR, STARLINT WEAR, STARLINT WEAR, STARLINT WEAR, STARLINT	RAME PLAN INLER HOF DUPORT AND INLER HOF DUPORT DOWN
							E	TP11 GPR1 PGE1 TDE3 TDE3 TDE3 FB02 FB02 FB02 FB02 FB02 FB02 FB02 FB02	OL VTOCHTEG GLYTOCHTEG	TICORFUNCTIONAL I CORRANA TICORFUNCTIONAL I CORRANA CARTANACATORIA CANA CARTANACATORIA CANA CARTANACATORIA CONTRACTA MARKANA MARKANA TAMANTANA TAMANTANA MARKANA MAR
					a set and a		F	28156 MSS51 MSF1 MSF1 MSF1 MSF1 MSF1 MSF1 MSF1 MSF	UNICOLOGICAL DISCONTRACTS PROFERSI PROFESSIONE PROFESSI	HIGHT.MARTEALE STORE ST
							G	2022 2024 2024 2025 2029 2029 2029 2029 2029 2029 2029	CLODOR FLEMENTION OLIDALTY PROSPERA OLIDALTY PROSPERA OLIDALTY PROSPERA ATV SUPPOSIS ATV SUPPOSI	GACTH REPLAYED FERMATE FUNCTIONED CONTACT STATE TWOCHDONG-CONTACT STATE TWOCHDONG-CONTACT STATE THAN THAT STATE THAT STATE THAT THAT THAT THAT THAT THAT THAT THAT
	1						н	HUTTA HUTTAL HTTAL HTTAL HTTAL	CERCHATEN ETHICTURE CERCHATEN ETHICTURE CERCHATEN ETHICTURE CERCHATEN ETHICTURE CERCHATEN ETHICTURE CERCHATEN ETHICTURE	REFYOR BJ REFYOR HA REFYOR KA REFYOR KD REFYOR KD REFYOR KD
	1		ana M				1 10 10 100		PIOTEIN STATUEIS	REPORTED PROTEINS / MONTANTON PACTORS*
	1			16		24	J	CDC54 HCH3 HCH2 CDC47 DBF2	DNA REPLICATION DNA REPLICATION DNA REPLICATION DNA REPLICATION CELL CYCLS	NCM ENTITATION COMPLEX NCM INITIANOR COMPLEX NCM INITIANOR COMPLEX NCM INITIANOR COMPLEX INITIANOR COMPLEX INITIANOR COMPLEX
	/	ar ar Mill	, FI				ĸ	FOR1 SOUR SIDEL SED11 QCH7 CO25 S151 CO25 S151 CO215 CO215 CO25 S154 S1555 S154 S1555 S1555 S1555 S1555 S1555 S1555 S1555 S15555 S1555 S15555 S15555 S1555 S15555 S155555555	TANDERST TATALOG UTLIANTON TATALOG UTLIANTON UTLIANTUP PROSPONDIL REFINATION REFINATION UTLIANTUP PROSPONDIL UTLIANTUP PROSPONDIL TCA CTCLS TCA CTCLS UTLIANTUP PROSPONDIL UTLIANTUP PROSPONDIL	KINCOLORILLA OFEN REFEARE PALT VELCATURA RECENTRA RECENTATIONEL A COLORIENTE RECENT PORTOCIONEL A COLORIENTE PORTOCIONEL A DEL CALLA PORTOCIONEL A DEL CALLA PORTOCIONEL A DEL CALLA REFERENCE COLORIELTO DE PORTOCIONEL RECENTATIONE COLORIELTO DE PORTOCIONEL RECENTATIONE PORTOCIONEL ANDIOL PORTOCIONE PORTOCIONEL ANDIOLO PORTOCIONEL ANDIOLO PORTOCIONEL PORTOCIONEL ANDIOLO PORTOCIONEL ANDIOLO PORTOCIONEL PORTOCIONEL ANDIOLO PORTOCIONEL ANDIOLO PORTOCIONEL ANDIOLO PORTOCIONEL PORTOCIONEL ANDIOLO PORTOCIONEL PORTOCIONEL PORTOCIONEL ANDIOLO PORTOCIONEL ANDIOLO PORTOCIONEL PORTOCIONELO PORTOCIONELE PORTOCIONEL PORTOCIO PORTOCIONEL POR

Functional enrichment



- If each gene is annotated with some standard terms (e.g., from GO), we can look for enriched terms
 - -Enrichment: statistically significant
 - -Contingency table for a term (e.g., binding):

	Genes in target set	Genes not in target set	Total
Genes annotated with a term	m ₁	m ₂	m ₁ +m ₂
Genes not annotated with a term	n ₁ -m ₁	n ₂ -m ₂	$n_1 + n_2 - m_1 - m_2$
Total	n ₁	n ₂	n ₁ +n ₂

-Null hypothesis H₀: the term and gene set are independent

$$Pr(m_1|H_0) = \frac{\binom{n_1}{m_1}\binom{n_2}{m_2}}{\binom{n_1 + n_2}{m_1 + m_2}}$$

–Compute one-sided p-value: $Pr(m \ge m_1 \mid H_0)$; Define cutoff (usually p<0.05 or p<0.01 is considered statistically significant)

Interpreting analysis results



- Generating hypotheses based on enrichment results:
 - -If many genes in the target set are annotated with a certain term, the term may be related to the phenotypic observation
 - -If many genes in the target set are annotated with a certain term, the other genes in the set may also be annotated with this term -- "guilt by association"
 - -If the genes are annotated with two terms, the terms may be related
 - -If no statistically significant terms can be found, the phenotypic observation may be non-genetic, largely affected by other (e.g., environmental) factors, or is affected by many loci

Problems



- Choosing a proper background
 - -Whole genome vs. all genes studied in the experiment
 - Suppose your experiment involves 100 genes, 20 of which (i.e., 20%) are annotated with a certain GO term, but overall only 10% of all human genes are annotated with that term. Using a different background to test your gene set could give very different results.
- Different evidence codes
 - Filtering
- Hierarchical relationships between different terms
 - Problems:
 - Redundancy
 - Different levels of detail/number of annotated genes affecting p-values
 - Different curators used terms at different levels: deeper if more is known
 - -Some proposed solutions:
 - GO slims
 - Fixed level (e.g., level 3)
 - More complex analysis involving term distances, information content, etc.
- Multiple hypothesis testing



- If a gene is annotated with a term, and another gene is annotated with a sub-term (child node in the DAG), they do not share terms but are clearly related
- Some ideas for tackling the problem based on the graph:



- –Use shortest-path in the graph as distance. Consider distance instead of common terms
 - Add edge weights based on number of instances annotated with the two terms
- -Consider the lowest common ancestor

Multiple hypothesis testing



- Suppose you have a set of genes from a certain experiment (e.g., those with a 2-fold increase of expression in cancer samples)
- You perform enrichment analysis using GO terms, KEGG biological pathways, OMIM disease annotations, etc., and find a term with p-value 0.001
 - -Recap of the meaning: Say you have n_1 genes in your set and m_1 of them are annotated with this term, if the term is randomly assigned to m_1+m_2 genes among all n_1+n_2 genes, the probability of having m_1 or more genes annotated with the term in a random set of n_1 genes is 0.001
 - —Is it statistically significant?
 - –Is it biologically significant?

Multiple hypothesis testing



- The famous "stock prediction email" metaphor:
 - —On day 1, you receive an email predicting the price of a stock will go up on day 2, and it really happens
 - -On day 2, you receive an email again predicting the price of the stock will go down on day 3, and it really happens
 - Similar things happen for 10 days. If every day the probability for the stock price to go up or down is independent of the other days and equals 0.5, then the p-value, i.e., the probability that in 10 random guesses there are 10 or more correct is 0.5¹⁰ ≈ 0.001
 - -Is it statistically significant?
 - —Is it financially significant?

Multiple hypothesis testing: The problem



- If you are the only one receiving emails, the predictions are indeed surprisingly accurate
- But if many have receive the emails with different predictions, together the results may not be that surprising
 - –If 1,024 people have received the emails, each with a different sequence of predictions, then one of them must be correct for all 10 days (p-value=1)
 - -If 1,024 people have received the emails and their contents are independently generated, one person is expected to have got 10 correct predictions
- Similarly, if you have tested your gene set with many terms, it would be not surprising if a term has p-value 0.001.

Multiple hypothesis testing: The solution

- Need to adjust the way to calculate or interpret the p-value
- Many ways:
 - –Bonferroni
 - Multiply the p-value by the number of terms tested
 - If you have tested 100 terms and you want to use a cutoff of p=0.01 to define statistical significance, then a term is significantly enriched in your gene set only if it gets a p-value of 0.01/100 = 0.0001
 - –Benjamini-Hochberg
 - Less conservative
 - —…
- Beware of cherry picking: shrinking the set of tested terms a posteriori
 - -Scientifically and ethically wrong

• AmiGO

- -Default software on the Gene Ontology web site
- BiNGO (A Biological Network Gene Ontology tool)
 - -Flanders Interuniversitary Institute for Biotechnology
 - -Plugin for the Cytoscape network viewer (next lecture)
- DAVID (Database for Annotation, Visualization and Integrated Discovery)
 - -National Cancer Institute
 - -One of the most cited tool for functional enrichment analysis
- GSEA (Gene Set Enrichment Analysis)

-Broad Institute

• ... (see http://www.geneontology.org/GO.tools_by_type.term_enrichment.shtml







- Functionality:
 - -ID conversion
 - -Enrichment analysis
 - Disease
 - Functional categories
 - Gene ontology
 - Protein domains
 - Pathways
 - ...

-Correction for multiple hypothesis testing



Sample results



• Functional annotation chart:

D		ВСИТОВЛЯЕ	DAVID Bioinfo National Institute of Allergy	rmat	ics Res	ources 6.7 Diseases (NIAID)	, NIH			
Deer	Function Current Gen Current Bac 155 DAVID © Options	nal Annotation Chart ne List: demolist1 ckground: Homo sapiens IDs						<u>Help</u>	and Manu	<u>ial</u>
Reru		Create Sublist								wolcod File
Sublist	Category	Term	A DT Genes	Coun	éit≜ ph.4	DT≜%≜P_Value	<u>Fold</u>	Bonferron	Benjamin	t FDR <u>Fisher</u>
		signal pentide		50	146 3250	19113 32 3 6 55-7	Enrichmen	4 2E-4	4 2E-4	9 8E-4 3 2E-7
	SP PIR KEYWORDS	signal	RT	50	148 3250	19235 32.3 8.6E-7	2.0	2.8F-4	2.8E-4	1.2E-3 4.2E-7
	UP SEO FEATURE	disulfide bond	RT	45	146 2819	19113 29.0 1.2E-6	2.1	8.1E-4	4.0E-4	1.9E-3 5.8E-7
	SP_PIR_KEYWORDS	disulfide bond	RT	46	148 2924	19235 29.7 1.7E-6	2.0	5.4E-4	2.7E-4	2.3E-3 8.1E-7
	GOTERM_CC_FAT	extracellular region	RT	40	124 2010	12782 25.8 6.9E-6	2.1	1.5E-3	1.5E-3	8.8E-3 3.3E-6
	GOTERM_CC_FAT	extracellular region part	<u>RT</u>	24	124 960	12782 15.5 3.8E-5	2.6	8.0E-3	4.0E-3	4.8E-2 1.4E-5
	GOTERM_MF_FAT	oxygen binding	<u>RT</u>	6	117 43	12983 3.9 3.8E-5	15.5	1.4E-2	1.4E-2	5.3E-2 2.2E-6
	SP_PIR_KEYWORDS	heme	<u>RT</u>	8	148 121	19235 5.2 4.0E-5	8.6	1.3E-2	4.3E-3	5.4E-2 4.4E-6
	SP_PIR_KEYWORDS	iron	<u>RT</u>	11	148 285	19235 7.1 6.9E-5	5.0	2.2E-2	5.6E-3	9.4E-2 1.3E-5
	SP_PIR_KEYWORDS	Secreted	<u>RT</u>	29	148 1689	19235 18.7 7.2E-5	2.2	2.3E-2	4.6E-3	9.7E-2 3.1E-5
	GOTERM CC FAT	extracellular space	RT	19	124 685	12782 12.3 9.4E-5	2.9	2.0E-2	6.5E-3	1.2E-1 3.2E-5

Sample results



• Functional annotation chart:

Su	blist	<u>Category</u>	term	, ¢RT	Genes	Cour	¢ LT	<u>PH</u> =	<u>PT</u> (<u>%</u>	P-Value	Fold Enrichment	Bonferront	<u>Benjamint</u>	FDR \$	Fisher Exact
]	UP_SEQ_FEATURE	signal peptide	RT		50	146	3250	19113	32.3	6.5E-7	2.0	4.2E-4	4.2E-4	9.8E-4	3.2E-7

- RT: related terms
- Count: number of genes on your list annotated with the term
- LT: number of genes on your list
- PH: number of genes in the whole genome annotated with the term
- PT: number of genes in the whole genome
- %: Count / LT
- P-Value: uncorrected p-value based on a procedure similar to Fisher's exact test (more conservative)
- Fold Enrichment: (Count / LT) / (PH / PT)
- Bonferroni: p-value corrected for multiple hypothesis testing based on the Bonferroni procedure
- Benjamini: p-value corrected for multiple hypothesis testing based on the Benjamini-Hochberg procedure
- FDR: False discovery rate (highly related to the Benjamini-adjusted p-value)
- Fish Exact: uncorrected p-value based on Fisher's exact test

Thresholding issue



- Sometimes the way to define the gene set is quite arbitrary
 - Differential expression Should we use 1.5-fold or 2-fold?
 - TF binding How far away from a gene should we still consider a site as a promoter of a gene?
- An illustration:
 - -2-fold over-expressed:

 Over-expressed	Yes	No
"binding" Yes	1	2
No	1	6

One-sided p-value: 0.5333

-1.5-fold over-expressed:

Over-expressed "binding"	Yes	Νο
Yes	2	1
Νο	1	6

One-sided p-value: 0.1833

Gene	Expression fold change in cancer	Annotated with term "binding"
1	2.7	Yes
2	2.5	No
3	1.9	Yes
4	1.2	No
5	1.1	No
6	0.9	No
7	0.7	Yes
8	0.6	No
9	0.4	No
10	0.2	No





• GSEA tries to avoid using arbitrary thresholds to call gene sets

• Ideas:

- -For each term, find the set of genes annotated with it
- -Check the ranks of these genes based on a phenotypic measure for calling gene sets (e.g., expression fold-change). Compute a statistic (enrichment score) based on these ranks: increase score when a gene with the term is encountered, decrease score otherwise
 - If the genes annotated with the term is randomly distributed across the whole list of genes, this process is essentially a random walk. The point with maximum distance can be used for evaluating the deviation from this random case.

-Evaluate statistical significance of the score by permuting phenotypic measure values



• Example: Are genes highly expressed in cancer related to binding?

			Actua	al data	Rando	om set 1	Rando	om set 2	Rando	om set 3	Rando	om set 4
		Expression	Annotated		Annotated		Annotated		Annotated		Annotated	
		fold change	with term	Enrichment	with term	Enrichment	with term	Enrichment	with term	Enrichment	with term	Enrichment
en	ne	in cancer	"binding"	score	"binding"	score	"binding"	score	"binding"	score	"binding"	score
	1	2.7	Yes	1	No	-1	No	-1	No	-1	Yes	1
	2	2.5	No	0	No	-2	Yes	0	No	-2	Yes	2
	3	1.9	Yes	1	Yes	-1	No	-1	Yes	-1	No	1
	4	1.2	No	0	No	-2	No	-2	Yes	0	No	0
	5	1.1	No	-1	Yes	-1	Yes	-1	No	-1	Yes	1
	6	0.9	No	-2	No	-2	No	-2	No	-2	No	0
	7	0.7	Yes	-1	No	-3	No	-3	No	-3	No	-1
	8	0.6	No	-2	No	-4	No	-4	No	-4	No	-2
	9	0.4	No	-3	Yes	-3	Yes	-3	Yes	-3	No	-3
	10	0.2	No	-4	No	-4	No	-4	No	-4	No	-4
	Rande	om set 5	Rando	om set 6	Rando	m set 7	Rando	m set 8	Rando	m set 9	Randor	mset 10
	Annotated		Annotated		Annotated		Annotated		Annotated		Annotated	
	with term	E na mi a la mara mat										
		Enrichment	with term	Enrichment	with term	Enrichment	with term	Enrichment	with term	Enrichment	with term	Enrichment
	"binding"	score	with term "binding"	Enrichment score	with term "binding"	Enrichment score	with term "binding"	Enrichment score	with term "binding"	Enrichment score	with term "binding"	Enrichment score
	"binding" No	score -1	with term "binding" Yes	Enrichment score	with term "binding" No	Enrichment score -1	with term "binding" Yes	Enrichment score	with term "binding" No	Enrichment score -1	with term "binding" No	Enrichment score -1
	"binding" No No	score -1	with term "binding" Yes No	Enrichment score 1 0	with term "binding" No Yes	Enrichment score -1 0	with term "binding" Yes No	Enrichment score 1 0	with term "binding" No No	Enrichment score -1 -2	with term "binding" No Yes	Enrichment score -1 0
	"binding" No No Yes	score -1 -2 -1	with term "binding" Yes No No	Enrichment score 1 0 -1	with term "binding" No Yes No	Enrichment score -1 0 -1	with term "binding" Yes No No	Enrichment score 1 0 -1	with term "binding" No No Yes	Enrichment score -1 -2 -1	with term "binding" No Yes No	Enrichment score -1 0 -1
	"binding" No No Yes No	score -1 -2 -1 -2 -1	with term "binding" Yes No No Yes	Enrichment score 1 0 -1	with term "binding" No No No	Enrichment score -1 0 -1 -2	with term "binding" Yes No No No	Enrichment score 1 0 -1 -2	with term "binding" No No Yes Yes	Enrichment score -1 -2 -1 0	with term "binding" No Yes No No	Enrichment score -1 0 -1 -2
	"binding" No No Yes No Yes	Enrichment score -1 -2 -1 -2 -1 -2 -1	with term "binding" Yes No Yes No	Enrichment score 1 0 -1 0 -1	with term "binding" No Yes No No No	Enrichment score -1 0 -1 -2 -3	with term "binding" Yes No No No No	Enrichment score 1 0 -1 -2 -3	with term "binding" No No Yes Yes No	Enrichment score -1 -2 -1 0 -1	with term "binding" No Yes No No Yes	Enrichment score -1 0 -1 -2 -2 -1
	"binding" No No Yes No Yes No	-1 score -1 -2 -1 -2 -1 -2 -1 -2 -2	with term "binding" Yes No Yes No No	Enrichment score 1 0 -1 0 -1 -1 -2	with term "binding" No Yes No No Yes	Enrichment score -1 0 -1 -2 -3 -2	with term "binding" Yes No No No Yes	Enrichment score 1 0 -1 -2 -3 -2	with term "binding" No No Yes No Yes	Enrichment score -1 -2 -1 0 -1 0	with term "binding" No Yes No Yes No	Enrichment score -1 0 -1 -2 -1 -2 -2
	"binding" No Yes No Yes No No	Enficiment score -1 -2 -1 -2 -1 -2 -1 -2 -1 -2 -3	with term "binding" Yes No Yes No No Yes	Enrichment score 1 0 -1 0 -1 -1 -2 -1	with term "binding" No Yes No No Yes No	Enrichment score -1 0 -1 -2 -3 -2 -3 -2 -3	with term "binding" Yes No No No Yes No	Enrichment score 1 0 -1 -2 -3 -2 -3	with term "binding" No Yes Yes No Yes No	Enrichment score -1 -2 -1 0 -1 0 -1	with term "binding" No Yes No Yes No Yes	Enrichment score -1 0 -1 -2 -1 -2 -1 -2 -1
	"binding" No Yes No Yes No No No	Enficiment score -1 -2 -1 -2 -1 -2 -1 -2 -3 -3 -4	with term "binding" Yes No Yes No Yes No Yes No	Enrichment score 1 0 -1 0 -1 -1 -2 -1 -2	with term "binding" No Yes No No Yes No No	Enrichment score -1 0 -1 -2 -3 -2 -3 -2 -3 -2 -3 -4	with term "binding" Yes No No Yes No Yes	Enrichment score 1 0 -1 -2 -3 -3 -2 -3 -2 -2	with term "binding" No Yes Yes No Yes No No	Enrichment score -1 -2 -1 0 -1 0 -1 -1 -1 -2	with term "binding" No Yes No Yes No Yes No	Enrichment score -1 0 -1 -2 -1 -2 -1 -2 -1 -2 -1
	"binding" No Yes No Yes No No No No	Enficiment score -1 -2 -1 -2 -1 -2 -1 -2 -3 -3 -4 -5	with term "binding" Yes No Yes No Yes No No No No No	Enrichment score 1 0 -1 0 -1 -1 -2 -1 -2 -3	with term "binding" No Yes No No Yes No No Yes	Enrichment score -1 0 -1 -2 -3 -2 -3 -2 -3 -2 -3 -4 -4	with term "binding" Yes No No No Yes No Yes No	Enrichment score 1 0 -1 -2 -3 -2 -3 -3 -2 -3 -2 -3	with term "binding" No Yes Yes No Yes No No No	Enrichment score -1 -2 -1 0 -1 0 -1 -1 -2 -2 -3	with term "binding" No Yes No Yes No Yes No No No	Enrichment score -1 0 -1 -2 -1 -2 -1 -2 -1 -2 -3
	"binding" No No Yes No No No No Yes	Enficiment score -1 -2 -1 -2 -1 -1 -2 -1 -2 -3 -4 -4 -5 -4	with term "binding" Yes No Yes No Yes No No No No No No	Enrichment score 1 0 -1 0 -1 -1 -2 -1 -2 -1 -2 -3 -3 -4	with term "binding" No Yes No No Yes No No Yes No	Enrichment score -1 0 -1 -2 -3 -2 -3 -3 -4 -4 -3 -4	with term "binding" Yes No No Yes No Yes No No No No No	Enrichment score 1 0 -1 -2 -3 -3 -2 -3 -2 -3 -2 -3 -2 -3 -2 -3	with term "binding" No Yes Yes No Yes No No No No No	Enrichment score -1 -2 -1 0 -1 0 -1 -1 -2 -3 -3	with term "binding" No Yes No Yes No Yes No No No	Enrichment score -1 0 -1 -2 -1 -2 -1 -2 -1 -2 -3 -3 -4



• Maximum values highlighted in blue:

		Actua	al data	Rando	om set 1	Rando	om set 2	Rando	m set 3	Rando	m set 4
	Expression	Annotated		Annotated		Annotated		Annotated		Annotated	
	fold change	with term	Enrichment	with term	Enrichment	with term	Enrichment	with term	Enrichment	with term	Enrichment
ene	in cancer	"binding"	score	"binding"	score	"binding"	score	"binding"	score	"binding"	score
	1 2.7	Yes	1	No	-1	No	-1	No	-1	Yes	1
	2 2.5	No	0	No	-2	Yes	0	No	-2	Yes	2
	3 1.9	Yes	1	Yes	-1	No	-1	Yes	-1	No	1
	4 1.2	No	0	No	-2	No	-2	Yes	0	No	0
	5 1.1	No	-1	Yes	-1	Yes	-1	No	-1	Yes	1
	6 0.9	No	-2	No	-2	No	-2	No	-2	No	0
	7 0.7	Yes	-1	No	-3	No	-3	No	-3	No	-1
	8 0.6	No	-2	No	-4	No	-4	No	-4	No	-2
1	9 0.4	No	-3	Yes	-3	Yes	-3	Yes	-3	No	-3
			4	No	4	No	1	No	_1	No	_1
1	0 0.2	No	-4	INO	-4	NU	-4	NO		NO	-4
1 Ran	0 0.2 ndom set 5	No Rando	-4 om set 6	Rando	-4 m set 7	Rando	-4 m set 8	Rando	m set 9	Randor	n set 10
1 Ran Annotated	0 0.2 ndom set 5 d	No Rando Annotated	-4 om set 6	Rando Annotated	-4 m set 7	Rando Annotated	m set 8	Rando Annotated	m set 9	Randor Annotated	n set 10
1 Ran Annotated with term	0 0.2 ndom set 5 d Enrichment	No Rando Annotated with term	-4 om set 6 Enrichment	Rando Annotated with term	m set 7 Enrichment	Rando Annotated with term	m set 8 Enrichment	Rando Annotated with term	m set 9 Enrichment	Randor Annotated with term	n set 10 Enrichment
1 Ran Annotated with term "binding"	0 0.2 Idom set 5 d Enrichment score	No Rando Annotated with term "binding"	-4 om set 6 Enrichment score	Rando Annotated with term "binding"	m set 7 Enrichment score	Rando Annotated with term "binding"	m set 8 Enrichment score	Rando Annotated with term "binding"	m set 9 Enrichment score	Randor Annotated with term "binding"	n set 10 Enrichment score
1 Ran Annotated with term "binding" No	0 0.2 ndom set 5 d Enrichment score -1	No Rando Annotated with term "binding" Yes	-4 om set 6 Enrichment score 1	Rando Annotated with term "binding" No	m set 7 Enrichment score -1	Rando Annotated with term "binding" Yes	m set 8 Enrichment score	Rando Annotated with term "binding" No	m set 9 Enrichment score -1	Randor Annotated with term "binding" No	n set 10 Enrichment score -1
1 Ran Annotated with term "binding" No No	0 0.2 ndom set 5 d Enrichment score -1 -2	No Rando Annotated with term "binding" Yes No	Enrichment score	Rando Annotated with term "binding" No Yes	m set 7 Enrichment score -1 0	Rando Annotated with term "binding" Yes No	m set 8 Enrichment score 1	Rando Annotated with term "binding" No No	m set 9 Enrichment score -1 -2	Randor Annotated with term "binding" No Yes	n set 10 Enrichment score -1
1 Ran Annotated with term "binding" No No Yes	0 0.2 ndom set 5 d Enrichment score -1 -2 -1	No Rando Annotated with term "binding" Yes No No	-4 om set 6 Enrichment score 1 0 -1	Rando Annotated with term "binding" No Yes No	m set 7 Enrichment score -1 0 -1	Rando Annotated with term "binding" Yes No No	m set 8 Enrichment score 1 0 -1	Rando Annotated with term "binding" No No Yes	m set 9 Enrichment score -1 -2 -1	Randor Annotated with term "binding" No Yes No	n set 10 Enrichment score -1 0 -1
1 Ran Annotated with term "binding" No No Yes No	0 0.2 ndom set 5 d Enrichment score -1 -2 -1 -2	No Randc Annotated with term "binding" Yes No No Yes	-4 om set 6 Enrichment score 1 0 -1	Rando Annotated with term "binding" No Yes No No	m set 7 Enrichment score -1 0 -1 -2	Rando Annotated with term "binding" Yes No No No	m set 8 Enrichment score 1 0 -1 -2	Rando Annotated with term "binding" No No Yes Yes	m set 9 Enrichment score -1 -2 -1 0	Randor Annotated with term "binding" No Yes No No	n set 10 Enrichment score -1 0 -1 -1 -2
1 Ran Annotated with term "binding" No No Yes No Yes	0 0.2 ndom set 5 d Enrichment score -1 -2 -1 -2 -1 -2 -1	No Randc Annotated with term "binding" Yes No Yes No	-4 om set 6 Enrichment score 1 0 -1	Rando Annotated with term "binding" No Yes No No No	m set 7 Enrichment score -1 0 -1 -2 -3	Rando Annotated with term "binding" Yes No No No No	m set 8 Enrichment score 1 0 -1 -2 -3	Rando Annotated with term "binding" No No Yes Yes No	m set 9 Enrichment score -1 -2 -1 0 -1	Randor Annotated with term "binding" No Yes No No Yes	n set 10 Enrichment score -1 0 -1 -2 -1
11 Ran Annotated with term "binding" No No Yes No Yes No	0 0.2 ndom set 5 d Enrichment score -1 -2 -2 -1 -2 -2 -1 -2 -2 -1 -2 -2 -1 -2 -2 -1 -2 -2 -1 -2 -2 -2 -1 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2	No Rando Annotated with term "binding" Yes No No Yes No No No	-4 om set 6 Enrichment score 1 0 -1 0 -1 -1 -2	Rando Annotated with term "binding" No Yes No No No Yes	-4 m set 7 Enrichment score -1 0 -1 -2 -3 -2	Rando Annotated with term "binding" Yes No No No No Yes	m set 8 Enrichment score 1 0 -1 -2 -3 -2	Rando Annotated with term "binding" No Yes Yes No Yes	m set 9 Enrichment score -1 -2 -1 0 -1 0	Randor Annotated with term "binding" No Yes No Yes No	n set 10 Enrichment score -1 0 -1 -2 -1 -2 -1 -2
Annotated with term "binding" No Yes No Yes No No No	0 0.2 ndom set 5 d Enrichment score -1 -2 -3 -3 -2 -1 -2 -3 -3 -2 -3 -3 -3 -2 -3 -3 -3 -3 -3 -2 -3 -3 -3 -3 -2 -3 -3 -3 -3 -3 -3 -3 -3 -3 -3	No Rando Annotated with term "binding" Yes No Yes No No Yes No Yes	-4 om set 6 Enrichment score 1 0 -1 0 -1 -2 -1	Rando Annotated with term "binding" No Yes No No Yes No No	m set 7 Enrichment score -1 0 -1 -2 -3 -2 -3	Rando Annotated with term "binding" Yes No No No Yes No	m set 8 Enrichment score 1 0 -1 -2 -3 -2 -3 -2 -3	Rando Annotated with term "binding" No Yes Yes No Yes No	m set 9 Enrichment score -1 -2 -1 0 -1 0 -1	Randor Annotated with term "binding" No Yes No Yes No Yes No Yes	n set 10 Enrichment score -1 0 -1 -2 -1 -2 -1 -2 -1
Annotated with term "binding" No No Yes No Yes No No No No No	0 0.2 ndom set 5 d Enrichment score -1 -2 -3 -3 -4 -4 -2 -3 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4	No Rando Annotated with term "binding" Yes No Yes No Yes No Yes No	-4 om set 6 Enrichment score 1 0 -1 0 -1 -2 -1 -2	Rando Annotated with term "binding" No Yes No No Yes No No No No	-4 m set 7 Enrichment score -1 0 -1 -2 -3 -2 -3 -2 -3 -2 -3 -2	Rando Annotated with term "binding" Yes No No No Yes No Yes	m set 8 Enrichment score 1 0 -1 -2 -3 -2 -3 -2 -3 -2 -3 -2	Rando Annotated with term "binding" No Yes Yes No Yes No Yes No No No	m set 9 Enrichment score -1 -2 -1 0 -1 0 -1 0 -1 -1 -2	Randor Annotated with term "binding" No Yes No Yes No Yes No Yes No	n set 10 Enrichment score -1 0 -1 -2 -1 -2 -1 -2 -1 -2 -1 -2
11 Ran Annotated with term "binding" No No Yes No Yes No No No No No	0 0.2 ndom set 5 d Enrichment score -1 -2 -3 -3 -4 -4 -4 -5 -4 -4 -5 -2 -3 -4 -4 -5 -4 -4 -5 -4 -5 -5 -4 -5 -5 -5 -5 -5 -5 -5 -5 -5 -5	No Rando Annotated with term "binding" Yes No Yes No Yes No Yes No No No No	-4 om set 6 Enrichment score 1 0 -1 0 -1 -1 -2 -1 -2 -1 -2 -3	Rando Annotated with term "binding" No Yes No No No Yes No No Yes	-4 m set 7 Enrichment score -1 0 -1 -1 -2 -3 -2 -3 -2 -3 -2 -3 -2 -3 -2 -3 -2 -3 -3 -2 -3 -3 -2 -3 -3 -3 -3 -3 -3 -3 -3 -3 -3 -3 -3 -3	Rando Annotated with term "binding" Yes No No No Yes No Yes No	m set 8 Enrichment score 1 0 -1 -2 -3 -2 -3 -2 -3 -2 -3 -2 -3	Rando Annotated with term "binding" No No Yes No Yes No No No No No	m set 9 Enrichment score -1 -2 -1 0 -1 0 -1 0 -1 -1 -2 -2 -3	Randor Annotated with term "binding" No Yes No Yes No Yes No No No No	n set 10 Enrichment score -1 0 -1 -2 -1 -2 -1 -2 -1 -2 -1 -2 -3



• Example:

- –Since in 3 of the 10 random sets, the maximum score \geq 1, the observed maximum score for the real data is not significant (p=0.3)
- It is possible to compute the p-value without performing the simulation



GSEA (cont'd)



• Illustration and example:



Image credit: Subramanian et al., PNAS 102(43):15545-15550, (2005)



Epilogue

Summary and Further Readings

Summary



• Gene annotations

- -Location (Chromosome, start, end, strand)
- -Biotype
- -Attributes
- Gene ontology
 - -Molecular function
 - -Biological process
 - -Cellular component
- Biological pathways
- Functional enrichment analysis for gene sets

Further readings



- Very good review papers on the topics covered in this lecture:
 - -Rhee et al., Use and Misuse of the Gene Ontology Annotations. *Nature Reviews Genetics* 9(7):509-515, (2008)
 - –Noble, How Does Multiple Hypothesis Testing Correction Work? *Nature Biotechnology* 27(12):1135-1137, (2009)