CS-E5875 High-Throughput Bioinformatics
Gene and SNP set enrichment analysis

Harri Lähdesmäki

Department of Computer Science
Aalto University

November 21, 2023

Contents

Motivation 1

- Consider e.g. a gene expression analysis between two groups:
  - One of the most common use of gene expression studies (e.g. RNA-seq)
  - Determine which genes are differentially expressed between two classes, say healthy and diseased groups
- At the end, statistical analysis of the experimental data gives:
  - A list of differentially expressed genes between the two classes
  - This list can be empty, short (tens), long (hundreds), or very long (thousands)
- Nobody knows/remembers the function of all genes
  - E.g. human genome contains around 20,000 genes
  - Interpreting/Understanding such gene lists is challenging

$\rightarrow$ Interpret the resulting gene list(s) collectively (not gene-by-gene) with the help of computational tools

Motivation 2

- If only a few replicate measurements exist, then gene-wise differential expression tests give results that
  - Have low statistical power and, thus, possibly contain only a few genes
  - May be unreliable at the level of individual genes
- Interpreting the resulting gene set collectively can help making the correct biological conclusion
- Can switch back and forth between gene level and gene set level analysis/interpretation, depending on their purpose:
  - For choosing a drug target we need gene level information
  - For understanding e.g. global dysregulation in complex diseases, gene sets can be more helpful

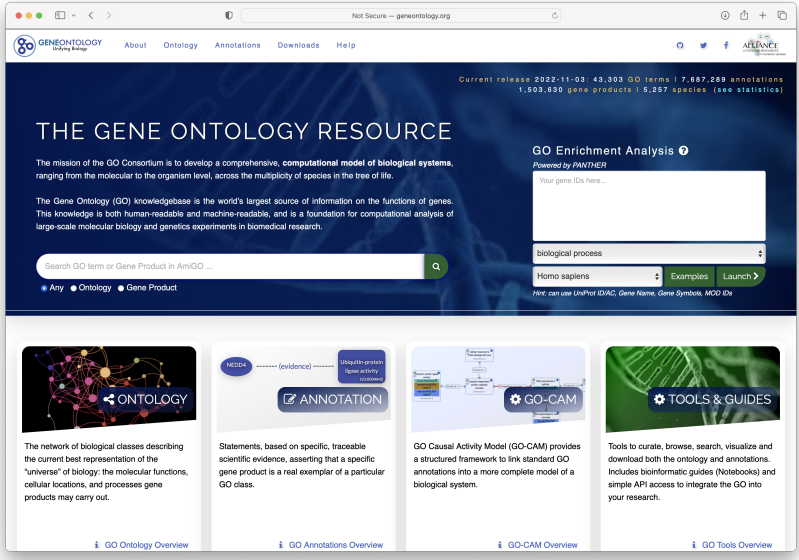
Interpreting the list of differentially expressed genes

- A typical goal: find the biological processes that are affected between the study groups, e.g., between healthy and diseased samples
- Address this question by assessing the genes collectively that are differentially expressed between the groups
- Examples of biological processes:
  - Protein translation
  - Cell death
  - Signal transduction
  - Response to stress
  - ...
- Biological processes can be described at multiple levels
  - Higher-level = more general process: multitude of genes
  - Lower-level = more detailed process: a few specific genes

Assigning genes to ontologies

- Gene Ontology (GO): The GO project is a collaborative, international effort to address the need for consistent and systematic functional annotation of gene products: `http://www.geneontology.org/`

Assigning genes to ontologies



Assigning genes to ontologies

GO offers three separate ontologies (term hierarchies):

1. Biological process: describes a biological objective to which the gene or gene product contributes
   - E.g. cell growth, cell death, signal transduction, protein translation
2. Molecular function: refers to the biochemical activity of gene products
   - E.g. enzyme, transporter, ligand
3. Cellular component: specifies in which compartment or location of a cell the active gene product can be found
   - E.g. ribosome, nuclear membrane, Golgi apparatus

Ontology structure example

- A set of terms under the biological process node pigmentation

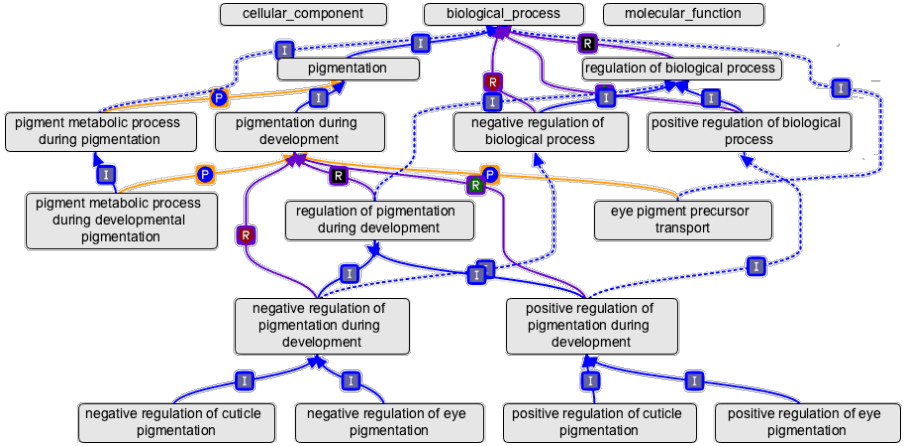


Figure from `http://www.geneontology.org/`

Constructing gene categories from GO terms

- The set of genes S associated with any particular GO term can be considered as a gene category or gene set of interest for subsequent analysis
- For example: Gene ontology term (GO:0008219) called `Cell Death` contains genes:
  - PDCD2L, BAD, DELE1, CD274, ...
  - Altogether 1103 genes for Homo sapiens

Other annotation resources

- MSigDB (Molecular signatures database)
  - Sets based on curated pathway information from 9 databases
  - Sets based on DNA motif occurrence
  - Sets based on computation analysis/predictions (expression similarity etc.)
  - Sets based on GO
  - Sets based on chromosomal location
- PANTHER database (mainly signaling pathways)
- KEGG and KEGG pathways
  - Molecular interaction and reaction networks

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways

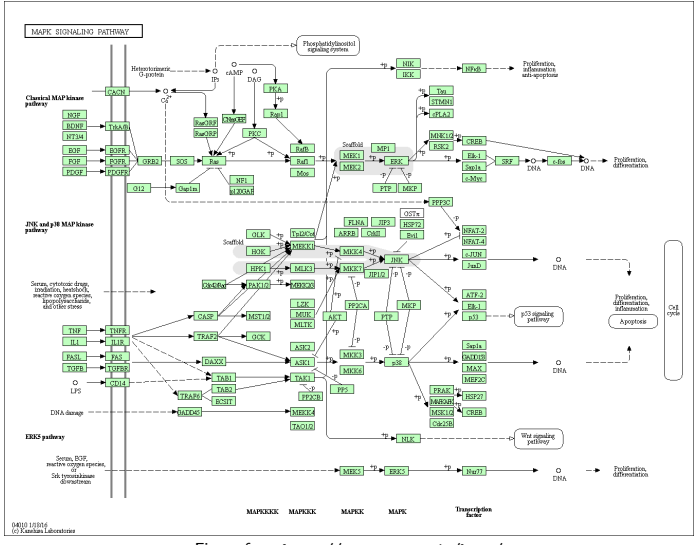


Figure from `http://www.genome.jp/kegg/`

Contents

Enrichment of a gene set

- Assume we have obtained a list of genes G_0 e.g. from statistical analysis of RNA-seq data
  - The gene list contain genes that, based on our data, are statistically significantly differentially expressed e.g. between our two study groups
- Question: is a gene ontology term overrepresented among the genes in the gene list?
  - A gene ontology term corresponds to a set of genes S
  - In other words, do the genes in the gene set S occur in the list of statistically significant genes G_0 more often than would be expected by chance
- The most common setting for enrichment analysis

Enrichment of a gene set

- Assume we want to evaluate the enrichment for a gene category (e.g. a biological process) S among differentially expressed genes G_0
  - G: all genes, $|G| = N$ in total
  - $G_0 \subseteq G$: differentially expressed genes, $|G_0| = n \leq N$ (often $n \ll N$)
  - $S \subseteq G$: a known set of genes annotated with a biological process, $|S| = m \leq N$
  - k: genes that are differentially expressed and belong to S, i.e., $|G_0 \cap S| = k$
- Null hypothesis H_0 : Assume that our differentially expressed genes are independent of the biological process
- Test statistic k: the number of genes that are in both S and G_0, i.e. overlap

Enrichment of a gene set

- Under the null, the probability of having overlap of exactly k genes, by chance, can be computed from the hypergeometric distribution

$$P(\text{overlap} = k) = \frac{\binom{N-m}{n-k}\binom{m}{k}}{\binom{N}{n}}$$

Enrichment of a gene set

- Under the null, the probability of having overlap of exactly k genes, by chance, can be computed from the hypergeometric distribution

$$P(\text{overlap} = k) = \frac{\binom{N-m}{n-k}\binom{m}{k}}{\binom{N}{n}}$$

- Alternative hypothesis H_1 : differentially expressed genes are not independent of the biological process
- The probability of an overlap of at least k genes is

$$p\text{-value} = P(\text{overlap} \geq k) = \sum_{l=k}^{\min\{n,m\}} \frac{\binom{N-m}{n-l}\binom{m}{l}}{\binom{N}{n}}$$

Enrichment of a gene set: illustration

- An example
  - 100 genes in total, $N = 100$
  - 20 are differentially expressed, $n = 20$
  - S contains 10 genes, $m = 10$
  - 5 differentially expressed genes are in S, $k = 5$
  - $P(\text{overlap} = 5) = 0.0215$
  - $P(\text{overlap} \geq 5) = \sum_{i=5}^{10} P(\text{overlap} = i) = 0.0255$

Enrichment of a gene set: illustration 2

- Another example
  - 20000 genes in total, $N = 20000$
  - 500 are differentially expressed, $n = 500$
  - S contains 100 genes, $m = 100$
  - 10 differentially expressed genes are in S, $k = 10$
  - $P(\text{overlap} = 10) = 0.0001611$
  - $P(\text{overlap} \geq 10) = \sum_{i=10}^{100} P(\text{overlap} = i) = 0.00020185$

Enrichment of a gene set

- The above hypothesis testing corresponds to the Fisher's exact test of association
- It is simple, accurate and can be applied in various contexts
- On the other hand, it requires setting a p-value threshold or FDR threshold for differential expression, and assumes that observations for each gene are independent
- Several different computational methods have been proposed for enrichment analysis

Contents

Gene set enrichment analysis (GSEA)

- Aim of GSEA: determine whether the members of S are randomly distributed throughout a ranked list L or primarily found at the top or bottom of the list
- No-cutoff strategy: find enriched annotations (gene categories) without having to specify a threshold for differentially expressed genes
  - Uses the whole information obtained from gene expression experiments
- Basic idea in gene set enrichment tests:
  - Start from ranked list of all genes (from up-regulated to down-regulated) and compute enrichment score for each gene set
  - Estimate statistical significance (p-value) of an enrichment score by permuting phenotype labels (e.g. randomly shuffle the case-control label of a subject) and recomputing differentially expressed genes as well as the enrichment score

Gene set enrichment analysis (GSEA)

1. Rank genes according to differential expression, set a running-sum statistic to 0
2. Compute Enrichment Score (ES):
   - Go down the gene list and
     - Increment a running-sum statistic if the gene belongs to set S
     - Decrease the running-sum statistic a gene if not in S
   - ES is the maximum deviation from 0 (a type of a Kolmogorov-Smirnov statistic)
3. Calculate empirical null distribution for ES:
   - Permute phenotype labels R times
   - Re-compute ES for each permutation: $\mathrm{ES}^{(1)}, \ldots, \mathrm{ES}^{(R)}$
4. Compute empirical p-value from empirical null distribution by counting the number of times the ES score is as large or even larger than for the observed data

$$p - \text{value} = \frac{1}{R} \sum_{i=1}^{R} I(\mathrm{ES}^{(i)} \geq \mathrm{ES})$$

5. Repeat the analysis for all sets S, adjust for multiple hypothesis testing

Gene set enrichment analysis (GSEA)

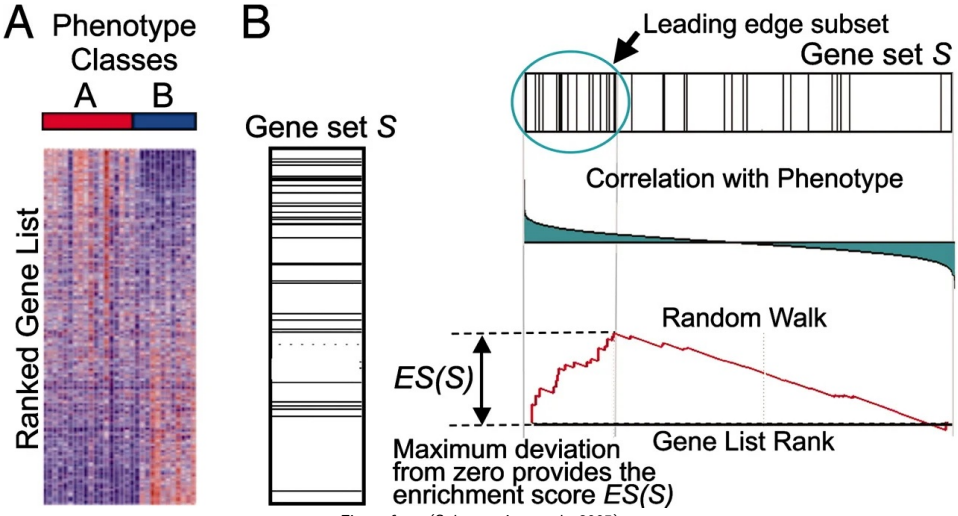


Figure from (Subramanian et al., 2005)

Example: GSEA in lung cancer studies

- Example: Two independent studies on lung cancer. Identify genes that are differentially expressed between group A and group B
  - Group A: good clinical outcome
  - Group B: poor clinical outcome
- Looking at individual genes, the two studies have little in common (12 genes among top 100 genes)
- However, there is large overlap between significantly enriched gene sets

Example: GSEA in lung cancer studies

Data set: Lung cancer outcome, Boston study	
Enriched in poor outcome	
Hypoxia and p53 in the cardiovascular system	0.050
Aminoacyl tRNA biosynthesis	0.144
Insulin upregulated genes	0.118
tRNA synthetases	0.157
Leucine deprivation down-regulated genes	0.144
Telomerase up-regulated genes	0.128
Glutamine deprivation down-regulated genes	0.146
Cell cycle checkpoint	0.216
Data set: Lung cancer outcome, Michigan study	
Enriched in poor outcome	
Glycolysis gluconeogenesis	0.006
vegf pathway	0.028
Insulin up-regulated genes	0.147
Insulin signalling	0.170
Telomerase up-regulated genes	0.188
Glutamate metabolism	0.200
Ceramide pathway	0.204
p53 signalling	0.179
tRNA synthetases	0.225
Breast cancer estrogen signalling	0.250
Aminoacyl tRNA biosynthesis	0.229

Figure from (Subramanian et al., 2005)

Contents

Understanding disease associated genetics variants

- Lecture #3 described methods to detecting SNPs using high-throughput sequencing technology
- Once genotyping has been done for a large cohort of individuals, statistical genetics methods are used to identify SNPs that are associated with the condition
  - These are generally called as genome-wide association studies (GWAS), and will be covered in other courses

Understanding disease associated genetics variants

- An illustration of GWAS studies

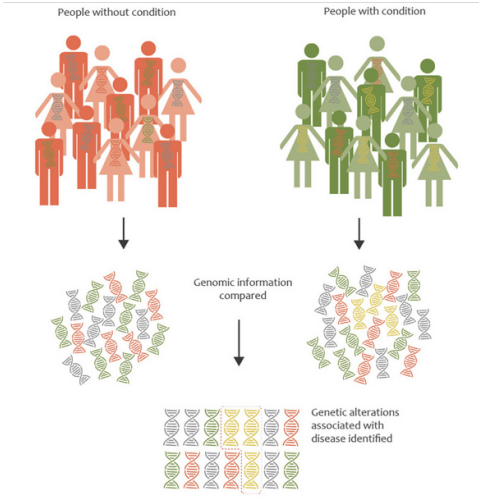


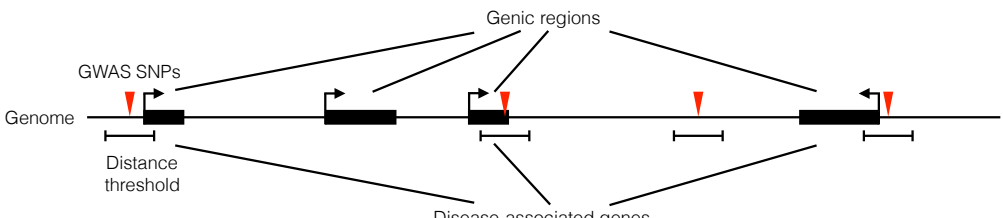
Figure from http://genetics.thetech.org/ask-a-geneticist/how-gwas-works

Understanding disease associated genetics variants

- Lets assume that we have successfully identified SNPs that are associated with a condition/disease
- Individual disease-associated SNPs that overlap protein-coding genes:
  - Can be studied further by analyzing individual proteins, experimentally or computationally, to understand how the non-synonymous mutations (missense, nonsense) affect the protein function
- Alternatively, computational methods can be used to assess whether disease-associated loci as a group (i.e., all detected SNPs) are enriched in
  - Biological pathways
  - Genomic annotations in non-coding genome

Understanding disease associated genetics variants

- A computational strategy proceeds as follows:
  - Choose a distance threshold (e.g. 100kb)
  - Associate each disease-associated SNP to those genes that are within the distance threshold from the SNP (along the linear sequence)
  - This will give a set of disease-associated genes S



An illustration of quantifying enrichment of GWAS SNPs at genic regions.

Understanding disease associated genetics variants

- This gene set S can be interpreted as any gene ontology category and its enrichment among differentially expressed genes can be analyzed using the same methods that we just studied
- Alternatively, the set of disease-associated genes can be interpreted as a set of differentially expressed genes G and its enrichment among gene ontologies can be assessed using the Fisher's exact test of association

Understanding disease associated genetics variants

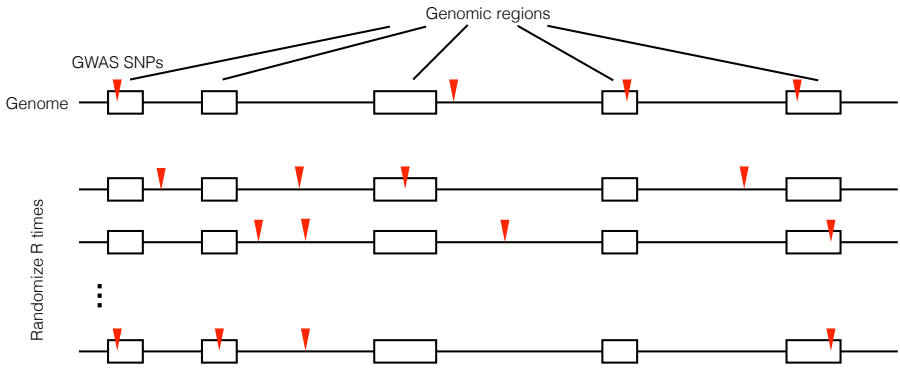
- Another computational strategy proceeds by randomizing SNPs
- Challenges in a straightforward (=uniform) randomization:
  - SNPs have a greater likelihood to overlap long genes and regions of strong linkage disequilibrium (LD)
- These biases can lead to false positive findings
  - For instance, brain pathways typically containing large genes and thus they likely appear to be overrepresented in GWAS loci

Understanding disease associated genetics variants

- The SNPsnap tool samples randomly SNPs with similar genetic properties as a set of query SNPs (i.e., the disease-associated SNPs)
- Random SNPs are matched based on
  - Minor allele frequency
  - Distance to nearest gene
  - Number of nearby genes (gene density), and
  - Number of SNPs in LD ("LD buddies")

Understanding disease associated genetics variants

- An illustration of SNPsnap tool



An illustration of quantifying enrichment of GWAS SNPs at genic regions: SNPsnap randomization.

Understanding disease associated genetics variants

- Empirical enrichment analysis for GWAS SNPs among genomic regions (=a gene ontology set)
  1. Count the overlap C of the original GWAS SNPs with the genomic regions
  2. Construct an empirical null distribution:
     - Randomize the GWAS SNPs R times using SNPsnap
     - For each randomized SNP set, count the overlap with the genomic regions, $C^{(i)}, i \in \{1, \ldots, R\}$
  3. Compute empirical p-value from the empirical null distribution by counting the number of times randomized SNP set has equal or larger overlap than the observed overlap

$$p - \text{value} = \frac{1}{R} \sum_{i=1}^{R} I(C^{(i)} \geq C)$$

  4. Repeat the analysis for all genomic regions, adjust for multiple hypothesis testing

References

- Nousiainen, Kari, et al. "snpEnrichR: analyzing co-localization of SNPs and their proxies in genomic regions." Bioinformatics (2018).
- Pers TH, Timshel P, Hirschhorn JN, SNPsnap: a Web-based tool for identification and annotation of matched SNPs, Bioinformatics, 31(3):418–420, 2015
- Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." PNAS 102.43 (2005)
- Tsagaratou, Ageliki, et al. TET proteins regulate the lineage specification and TCR-mediated expansion of iNKT cells. Nature Immunology 18.1 (2017)