ADER18

LO 10 - Perform simple **functional enrichment analysis** and understand the concepts involved

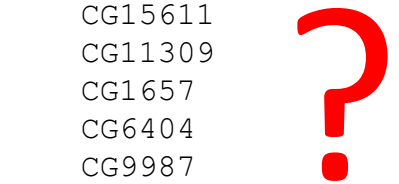
Daniel Faria

LO 10.1 - From genes to gene functions

We've got differentially expressed genes—what now?

- RNAseq experiments result in **sets of genes** of interest (that are differentially, over- or under-expressed)
- Such sets are opaque—it is hard to understand much from gene codes or even names, and even if we could, we are seldom interested in individual genes
- We usually want to understand phenomena at a more abstract level, such as the functional level, rather than the gene level

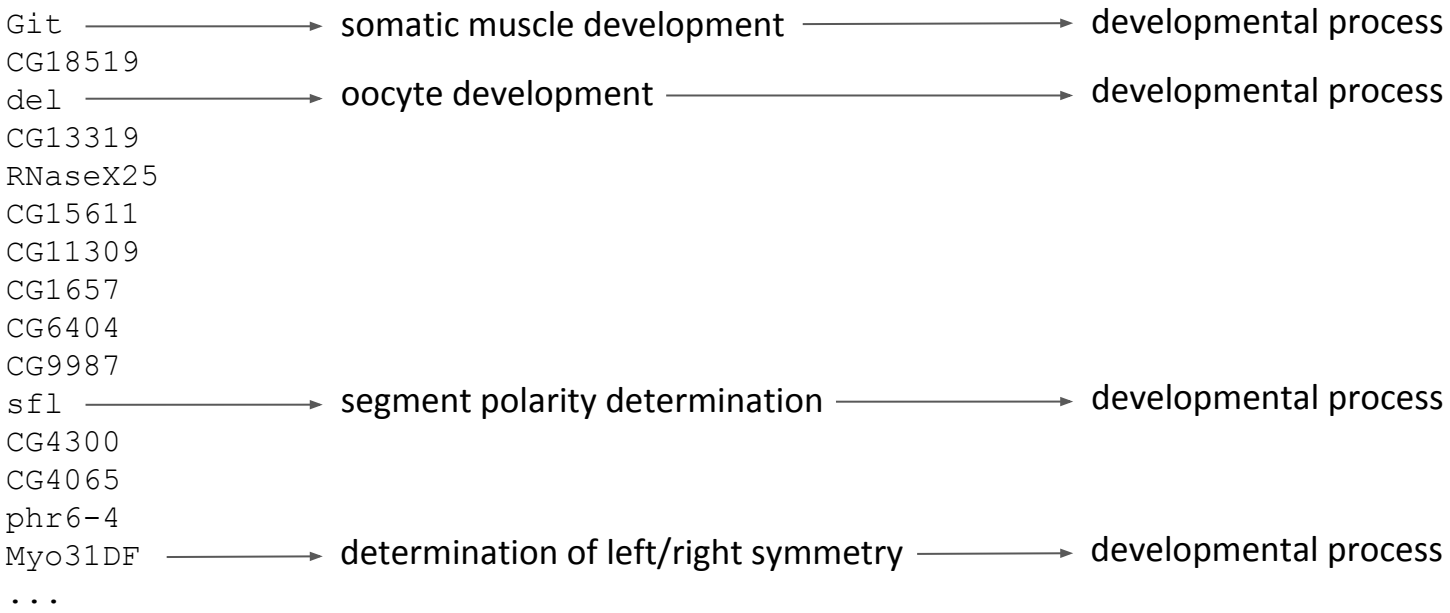
```
Git
CG18519
del
CG13319
RNaseX25
CG15611
CG11309
CG1657
CG6404
CG9987
sfl
CG4300
CG4065
phr6-4
Myo31DF
...
```



Abstracting from the gene level

- To abstract from the gene level, we need **annotations** of our genes according to a **classification schema** that covers the aspects we're interested in, which are typically **functional aspects**
- For some problems, a flat classification is sufficient (e.g., if all you care about are transcription factors)
- But usually a **hierarchical** classification schema is best to enable **integration** and pattern discovery, as fine-grained functions typically occur in only a few genes

Abstracting from the gene level



Without a hierarchical classification, this pattern would be hard to uncover!!!

Functional classification schemes

- There are several suitable functional classification schemas in use for genes, e.g.:
  - Enzyme Commission (EC) classification for enzymatic functions
  - KEGG for metabolic pipelines
  - **Gene Ontology** (GO)—the broadest and richest option, and thus the most widely used
- Most genetic databases include annotations to these classification schemas

The Gene Ontology

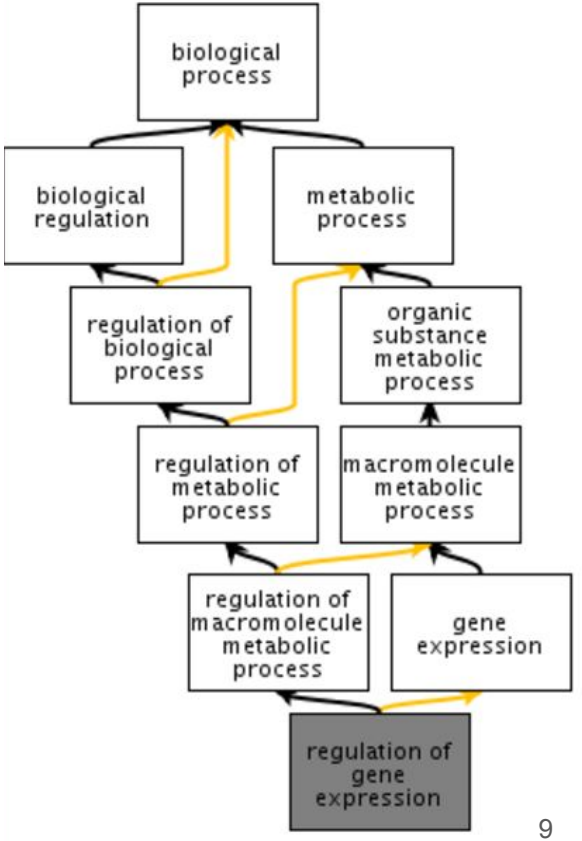
- GO is an **ontology** that covers three levels of gene function, called GO types or aspects:
  - ***Molecular function***: the individual gene functional level (e.g., GTPase)
  - ***Biological process***: the cellular and/or organismal functional level (e.g., signalling, muscle development)
  - ***Cellular component***: the locational level (e.g., nucleus)

Aside: what is an ontology?

- An ontology is a specification of a conceptualization
- In plainer English: it is a formal, structured representation of a domain of knowledge, including the description of the concepts therein and the relations between them
- In even plainer English: it is a classification scheme for any given subject

The Gene Ontology

- Each GO type is structured as a **directed acyclic graph** (a hierarchy with multi-parenting)
- In addition to subclass (‘is a’) relations, there are ‘part of’, ‘regulates’, and ‘occurs in’ relations
- GO types are ‘is a’ orthogonal, but *molecular functions* can be ‘part of’ *biological processes*, and both can ‘occur in’ *cellular components*



GO annotations

- Like for other classification schemes, genes are associated with GO terms via annotations
- A gene may have multiple annotations, even of the same GO type
- According to the **true path rule**, a gene annotated to a term is implicitly annotated to each ancestor of that term
- Annotations have evidence codes that encode the type of evidence supporting them

Getting GO annotations

- You can get individual GO annotations in most genetic databases (GeneBank, UniProt, specific organism genome dataases)
- You can download GO annotations in bulk for a given organism from the Gene Ontology download page or from BioMart
- GO and its annotations are updated monthly; it is important to use up-to-date versions but above all, to use a version of the annotations that matches the version of the ontology you're using

Tasks

- Task 1 – Go to the Gene Ontology download page:
  - Download the GO in OBO format (right-click save)
  - Download the GO annotations for *Drosophila melanogaster*
- Task 2 – Go to BioMart:
  - Download the GO annotations for *Mus musculus* (select Gene Stable ID plus GO term accession; save output in TSV)

LO 10.2 - Understand the concept of functional enrichment analysis, and the statistics involved

We’ve got annotated genes—what now?

- Abstracting from the gene level via functional annotations may enable us to find patterns in our gene set
- But we need to assess how significant the patterns we’re observing are in order to substantiate any inference of meaning
- That is precisely the purpose of **enrichment analysis**

Enrichment analysis

- Enrichment analysis is the application of **statistical tests** to ascertain whether a **sample** set of entities is **enriched** in relation to the overall **population** w.r.t. particular features

- By enriched, we mean that the **sample frequency** of the feature is greater than would be expected by chance given the **population frequency**

- The appropriate statistical test is the **one-tailed** variant of **Fisher's exact test**, a.k.a. hypergeometric test for over-representation

Fisher’s exact test

- Fisher’s exact test is a **statistical test** that applies to **sampling events**, and calculates the probability that the feature(s) of the sample are the product of chance alone, given their frequency in the population (null hypothesis)

- In the **one-tailed** version, which measures enrichment, we compute the probability of observing at least the sample frequency, given the population frequency

- The test relies on the **hypergeometric distribution**

The hypergeometric distribution

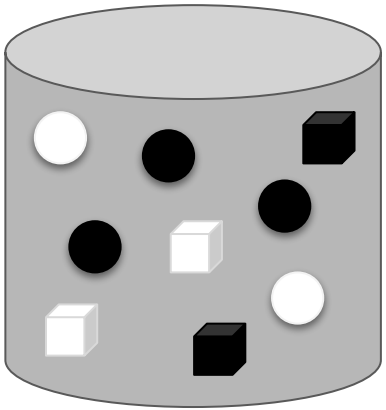
- The hypergeometric distribution describes the probability of *k* successes in *n* random draws, without replacement, from a finite population of size *N* that contains exactly *K* “successful” objects:

$$\mathrm{P}(\mathrm{X} = k) = \frac{\binom{K}{k}\binom{N-n}{K-k}}{\binom{N}{n}}$$

- The probability of getting at least k successes corresponds to the one-tailed **Fisher test p-value**

The hypergeometric distribution

- Example:
  - If I draw 3 items from my pool, what is the probability of drawing:
    - All spheres: $P(X=3|3,5,9) = 12\%$
    - All white: $P(X=3|3,4,9) = 5\%$
    - At least 2 black items: $P(X\geq 2|3,5,9) = 60\%$
    - At least 2 black spheres: $P(X\geq 2|3,3,9) = 23\%$



Applying Fisher's test to RNA-seq gene sets

Sample:

- The set of differentially or over- or under-expressed genes, depending on the biological question being addressed

Population:

- The transcriptome (i.e., all genes present in the RNA-seq experiment with meaningful counts)
- **We should only place in the population genes for which we could determine status w.r.t. inclusion in the sample**

Applying Fisher's test to RNA-seq gene sets

Frequencies (*k*, *K*):

- Count of genes in the sample/population that have the feature we're testing; if hierarchical, count also genes have subclasses of that feature (true path rule)

Sizes (*n*, *N*):

- Total count of genes in the sample/population **that have any known feature** under our classification schema—**we cannot count genes whose status w.r.t. the feature of interest is undetermined**

Applying Fisher's test to RNA-seq gene sets

- You can only use Fisher's test to compare a sample with the population whence it was extracted
- **You cannot use it to compare two samples directly**
- You can compare the enrichment analysis results of two or more samples to gauge the differences between them
- If you need to statistically assess whether two samples are functionally different, you'd need a different test (e.g., Wilcoxon) but that is beyond the scope of this course

Performing functional enrichment analysis

- Given an **RNA-seq experiment**, a **functional classification schema** and corresponding **functional annotations**:
  - Determine what should be the sample and population sets of genes
  - Compute all inferred annotations (if the schema is hierarchical)
  - Compute *n* and *N* (genes in the sample/population that have any annotation)
  - For each functional annotation (of interest) that occurs in the study set
    - Get the counts *k* and *K*
    - Compute the one-tailed Fisher p-value

Aside: correcting for multiple testing

- Whenever we're testing for multiple hypotheses simultaneously, we generally should correct for multiple testing:
  - Statistical testing is based to the probability of erroneous rejection of the null hypothesis being low
  - With multiple related tests, the probability of at least one of them being a false positive increases
  - E.g.: if you flip 10 coins, the odds of getting 10 heads are only 0.1%; but if you repeat this 1000 times, you expect to observe 10 heads once

Aside: correcting for multiple testing

- Even though:
  - The only stochastic event—sampling of genes—typically has already been the subject of statistical testing and multiple test correction
  - The transformation from genes to functions is deterministic
- We can only consider a functional aspect statistically significant if it occurs more often than would be expected by chance, which includes the consideration that we are performing multiple tests

Aside: correcting for multiple testing

- **Family-wise error rate (FWER)**: control the probability of making at least one false discovery—more conservative but safer
  - Bonferroni correction: multiply the p-values by the number of tests to obtain corrected p-values
- **False discovery rate (FDR)**: control the ratio of false discoveries—more powerful
  - Benjamini-Hochberg correction: step-wise correction; produces q-values, which indicate the ratio of false discoveries

Particulars of GO enrichment analysis

- GO is actually three independent classification schemas, so we should carry out enrichment analysis independently for each (or just for the one we are interested in)
- This affects the sizes (*n* and *N*) as genes may have annotations in one GO type and not another
- It also affects multiple test corrections—only tests of the same GO type should be considered related for this purpose

GO enrichment analysis tools

- There are many tools available:
  - **Webtools**: GOrilla, GO
  - **Stand-alone & Galaxy tools**: **GOEnrichment**, Ontologizer
  - **R tools**: gsea, GOstats, topGO
- Choose tools that enable you to define the version of GO and the annotation set used!

Alternatives to Fisher's test

- GOrilla and a few other tools offer the option of "enrichment analysis" of a single ranked list of genes, using a *minimum hypergeometric* score (or variant thereof) to compare top genes in the list with the rest of the list
- Rank typically lacks biological meaning—the p-values of the differential expression test only provide validation, and the log fold-changes in expression are too imprecise to meaningfully rank our genes
- So we're better off sticking with Fisher's test

Task 1

- Perform functional enrichment analysis using the GOEnrichment tool on Galaxy, on the differential expression results from Trapnell et al (300 random differentially expressed *D. melanogaster* genes). Use the sample file and Drosophila annotation file provided in the functional_enrichment folder, as well as the GO OBO file you downloaded earlier. Set “summarize output” to off and otherwise use default options. For now, don’t use a population file.
    - Do you see enriched functions? Should you?

Task 2

- Repeat the previous analysis, but this time use the population file provided in the functional_enrichment folder (the set of genes that have a numeric adjusted FDR, not 'NA', or non-zero base expression)
    - Are there still functionally enriched genes?

This demonstrates that selecting an adequate background population set is critical to obtaining accurate and statistically correct results

LO 10.3 - Interpreting the results of functional enrichment analysis

What can we get out of enrichment analysis results?

- Interpretation hinges heavily on the biological context of the study and on the motivation to do the analysis
- Enrichment analysis can be used for:
  - **Validation** (e.g., of a protocol for extracting membrane proteins)
  - **Characterization** (e.g., of the effects of a stress in an organism)
  - **Elucidation** (e.g., of the functions impacted by the knock-out of a transcription factor)

Caveats

- Keep in mind that **statistically significant ≠ biologically meaningful**!
- But statistically enriched terms often provide some biological or technical insight about the underlying experiment, even if it isn't readily apparent (e.g., "binding" being enriched in the nasal epithelium)
- Terms that are very generic are difficult to interpret, whereas those that are very specific are usually not integrative

Caveats

- **Outliers may occur!** We're making a statistical test (of enrichment) on top of another (of differential expression) which in turn is preceded by a statistical normalization. Errors propagates across steps, and even fine differences in each step can affect the final results.

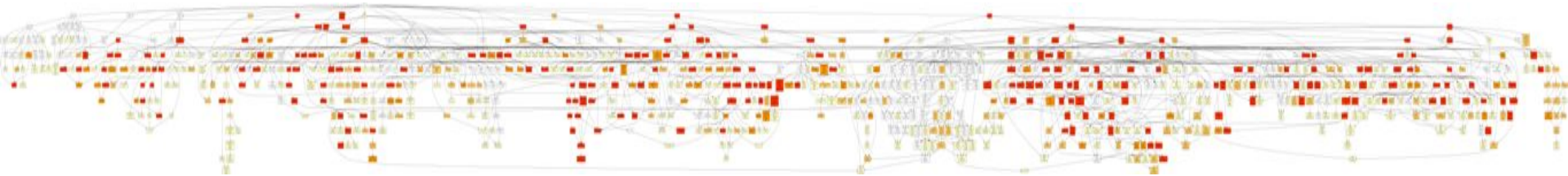
- **Enrichment analysis is qualitative**, rather than quantitative: we're treating genes as either "on" or "off" and consequently only assessing which functional aspects are statistically affected, rather than by how much they are affected.

Interpreting GO enrichment analysis results

- The p-value provides validation, but the sample frequency and the semantics of the GO term (definition + structure) are the keys for interpretation
- We can get the frequency from results tables, but the semantics requires graph views of the results. These enable us to view enrichment results in context, and highlight enriched ontology branches
- Evidently, interpretation varies with GO type

Too much information!

- The size and complexity of GO often lead to huge sets of enriched terms with different levels of specificity, so it helps to group related enriched terms into clusters when analysing the results
- Graph views are also essential for this, but sometimes even the graph view can become overwhelmed by the size of the results...



Reducing the complexity of GO enrichment results

- We can reduce the number of tests performed to avoid getting overwhelmed:
  - **Ignore singletons**: functions that occur in a single sample gene may be enriched (e.g., if they occur in no other genes in the population) but aren't integrative
  - **Skip dependent tests**: testing a superclass when its sample frequency is the same as one of its subclasses is redundant (we gain neither specificity nor integration)

Reducing the complexity of GO enrichment results

- A more extreme reduction can be achieved by using **GO slims** (“trimmed” versions of the ontology) instead of the full GO:
    - They will lead to much simpler results, but also to a substantial loss in specificity which may be unsatisfactory
    - They require that the GO annotations be converted from full GO to GO slim

Reducing the complexity of GO enrichment results

- Alternatively, we may simplify/summarize the results *a posteriori*, using:
  - The family-based clustering algorithm integrated into GOEnrichment which reduces complexity while keeping branch information, but loses some specificity
  - The semantic similarity-based REVIGO tool, which not only loses specificity but may merge branches
  - An *ad hoc* filter

Applying an *ad hoc* filter

- We can consider that our initial enrichment analysis was exploratory, and focus only on the parts of GO we are interested
- As long as our criteria for selecting those parts are independent of the p-value (e.g., we can make vertical or horizontal cuts of GO)
- If that is the case, we can even recompute the multiple test corrections according to the resulting number of selected tests

Task 1

- Pick up the differential expression results from mouse brain vs. heart
- Generate a population file and two sample files, one with overexpressed genes and the other with underexpressed genes
- Run GOEnrichment as previously, for both the over- and underexpressed sample files (use the mouse GO annotation file you downloaded earlier), then analyze the biological process results tables and graph files
    - Can you guess the order of the tissues in the original differential expression test?

Task 2

- Repeat the GOEnrichment run for the underexpressed genes with the set “summarize output” to on
- Download the generic GO Slim; use the GOSlimmer tool in Galaxy to convert your annotation file from GO to GO Slim, then repeat the GOEnrichment run using the GO Slim and slim annotations (set “summarize output” to off)
    - Analyze the BP graphs. Can you still tell that this sample is brain tissue?
    - How do the results compare w.r.t. simplicity and specificity?

Task 3

- Analyze the GO enrichment results of a few of the single cell clusters you identified yesterday that are included in the functional_enrichment/single_cell folder
    - Can you guess to which cell cluster each GO enrichment graph corresponds?