Take-home
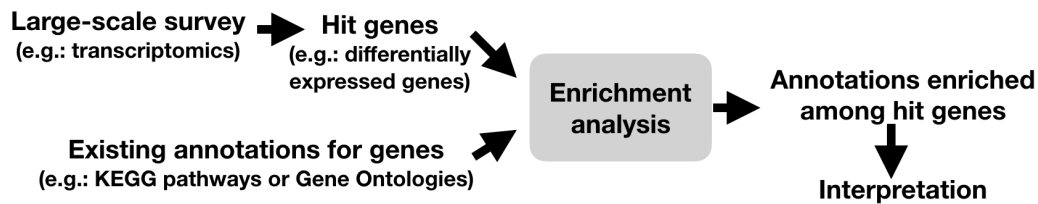
1. Become comfortable developing own tools.
2. Appreciate that tools include non-obvious assumptions.

What is the role of enrichment analysis?

Large-scale survey
(e.g.: transcriptomics) → **Hit genes**
(e.g.: differentially expressed genes)

Enrichment analysis → **Annotations enriched among hit genes**

Existing annotations for genes
(e.g.: KEGG pathways or Gene Ontologies)

Interpretation

Why not use an existing tool?

- **Do use existing tools, if it fits purpose (e.g.: excellent https://david.ncifcrf.gov)!**
- Gain control over assumptions.
- Have fun building it.

Our construction map: Bold indicates functions

1) complete_path
access files

2) read_go_and_turn_to_annotation
All pairs of Gene Ontology (GO) annotations and genes

Select for human (NCBI taxon: 9606)

Build filter for different categories within GO

Only keep gene annotations without negating evidence

3) functional_enrichment
Uses hits (e.g.: genes identified in study), background (e.g.: genes that could have been measured in experiment) and pairs of annotations and genes (e.g.: GO)

lead up to Fisher's exact test with
[hit_genes_in_annotation, not_hit_genes_in_annotation],
[hit_genes_not_in_annotation, not_hit_genes_not_in_annotation]

compute enrichment
perform testing for multiple hypothesis
(e.g.: Bonferroni, Benjamini-Hochberg)

4) Example gene list: from Crow et al. 2019:
Predictability of human gene expression
consider the top 300 genes as hits

5) Bonus: read_gwas_and_turn_to_annotation
use EBI-GWAS collection and NCBI gene info

Material

- Code: https://github.com/tstoeger/enrichment_tool
- Gene Ontology annotations: https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz
- Gene list: https://github.com/maggiecrow/DEprior/blob/master/DE_Prior.txt
- EBI GWAS: https://www.ebi.ac.uk/gwas/api/search/downloads/full
- Human Gene Info:
  https://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz