**Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap**

1. DEFINITIONS

**Pathway.** Genes that work together to carry out a biological process.

**Gene set.** A set of related genes. A ‘pathway gene set’ includes all genes in a pathway. Gene sets can be based on various relationships between genes, such as cellular localization (e.g., nuclear genes) or enzymatic function (e.g., protein kinases). Details such as protein interactions are not included.

**Gene list of interest.** The list of genes derived from an omics experiment that is input to pathway enrichment analysis. Ranked gene list. In many omics data (e.g., that from RNA-seq for gene expression), genes can be ranked according to some score (e.g., level of differential expression) to provide more information for pathway enrichment analysis. Pathways enriched in genes clustered at the top of a ranked list would score higher than if the pathway genes are randomly scattered across the ranked list.

**Pathway enrichment analysis.** A statistical technique to identify pathways that are significantly represented in a gene list or ranked gene list of interest.

**Multiple testing correction.** Thousands of pathways may be individually tested for enrichment, and this could lead to significant enrichment P values appearing by chance alone. Multiple testing correction is a statistical technique to correct the P values from individual enrichment tests to address this problem and reduce the chance of false-positive enrichment.

**Leading-edge gene.**

A subset of genes found in the ranking at or just before the maximal ES in a GSEA analysis. This subset of genes often accounts for a pathway being defined as enriched.

**Molecular Signatures Database (MSigDB).** MSigDB is a database of gene sets based on GO, pathways, curation, individual omics studies, sequence motifs, chromosomal position, oncogenic and immunological expression signatures, and various computational analyses maintained by the GSEA team (http://www.msigdb.org). A relatively non-redundant collection of ‘hallmark’ gene sets is available. The data can be used with many pathway enrichment methods.

**KEGG.** The KEGG database is most useful for its intuitive pathway diagrams. It contains multiple types of pathways, some of which are not normal pathways but are rather disease-associated gene sets, such as ‘pathways in cancer’ (http://www.genome.jp/kegg/). Up-to-date GMT files for KEGG pathways are currently not freely available because of data licensing restrictions.

**Pathway Commons.** This database collects detailed pathway descriptions from multiple originating pathway databases. It collects information from other pathway databases and provides it in a standardized format (http://www.pathwaycommons.org).

**Statistical tests in pathway enrichment analysis.** Common: Fisher’s exact test based on hypergeometric distribution. Determines whether fraction of genes of interest in pathway is higher compared to fraction of genes outside pathway. Many improved test build on Fisher’s exact test, e.g.: Ranked versus non-ranked tests, Exact versus permutation-based tests, competitive versus self-contained tests. Generally: If genes in data can be ranked -> ranked test. Fisher’s exact test chosen for non-ranked gene lists.

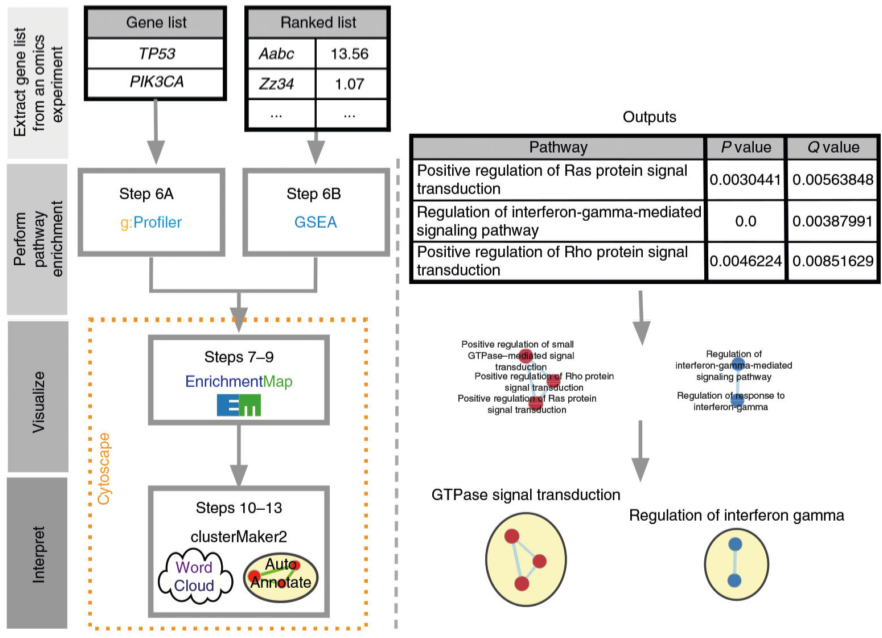
1. ABSTRACT

Pathway enrichment analysis helps researchers gain mechanistic insight into gene lists generated from genome-scale (omics) experiments. This method identifies biological pathways that are enriched in a gene list more than would be expected by chance. We explain the procedures of pathway enrichment analysis and present a practical step-by-step guide to help interpret gene lists resulting from RNA-seq and genome-sequencing experiments. The protocol comprises three major

steps: definition of a gene list from omics data, determination of statistically enriched pathways, and visualization and interpretation of the results. We describe how to use this protocol with published examples of differentially expressed genes and mutated cancer genes; however, the principles can be applied to diverse types of omics data. The protocol describes innovative visualization techniques, provides comprehensive background and troubleshooting guidelines, and uses freely available and frequently updated software, including g:Profiler, Gene Set Enrichment Analysis (GSEA), Cytoscape and EnrichmentMap. The complete protocol can be performed in ~4.5 h and is designed for use by biologists with no prior bioinformatics training.

1. INTRODUCTION

* Standard approach: pathway enrichment analysis
  + Summarizes a large gene list as smaller list of easily interpretable pathways
* Pathways are statistically tested for over-representation in the experimental gene list relative to what is expected by chance by several common statistical tests
  + Number of genes in experience
  + Relative ranking
  + Number of genes annotated to a pathway of interest



Overview:

1. definition of a gene list of interest using omics data
   * data processing
   * 2 ways to define gene list: list or ranked list
   * Ranked for example by differential gene expression score
   * Differential gene expression analysis results include P value of significance of differential expression, adjusted p value (Benjamini-Hochberg), effect size and direction of expression change (log-transformed fold-change)
   * Gene list is then ranked by one or more of these values (e.g., —log10 P value multiplied by the sign of log-transformed fold-change) and studied using pathway enrichment analysis
2. Pathway enrichment analysis
   * The paper uses two different software at this point, I read the GSEA part
   * Threshold free method, analyses all genes on basis of their differential expression rank or score without prior filtering
   * Recommended when ranks are available
   * GSEA searches for pathway whose genes are enriched at top or bottom of ranked gene list
   * E.g.: topmost differentially expressed genes involved in cell cycle -> cell cycle pathway upregulated. If genes are randomly distributed in ranked list -> pathway not upregulated
   * Calculation of enrichment score (ES): GSEA examines genes from top to bottom of ranked list, ES increases if gene is part of pathway / decreases when not part of pathway
   * Running sum values are weighted -> enrichment in very top/bottom is amplified <-> enrichment in genes with moderate rank not amplified
   * ES score = maximum value of running sum (normalized to pathway size -> NormalizedES) 🡪 Reflects enrichment in list
   * Positive/negative NES -> enrichment at top/bottom list
   * Permutation based p value is computed and corrected for multiple testing 🡪 permutation-based FDR value: Q [0 (highly significant); 1 (not significant)]
   * Same analysis performed from bottom to top
   * Resulting pathways are selected using FDR Q value threshold (e.g., < 0.05) and ranked using NES
3. Visualisation and interpretation of pathway enrichment analysis results
   * Many pathways share genes -> pathway enrichment analysis often highlights multiple versions of same pathway 🡪 Visualization methods: EnrichmentMap
   * Enrichment map: network visualization that represents overlaps among enriched pathways
   * Helps to identify interesting pathways or themes (also positive control)
   * Pathways and themes with the strongest ESs should be studied first, followed by progressively weaker signals
   * Additional pathway diagram (Pathway commons, KEGG)

Advantages and limitations:

* Interesting for colloquium? (p. 8)