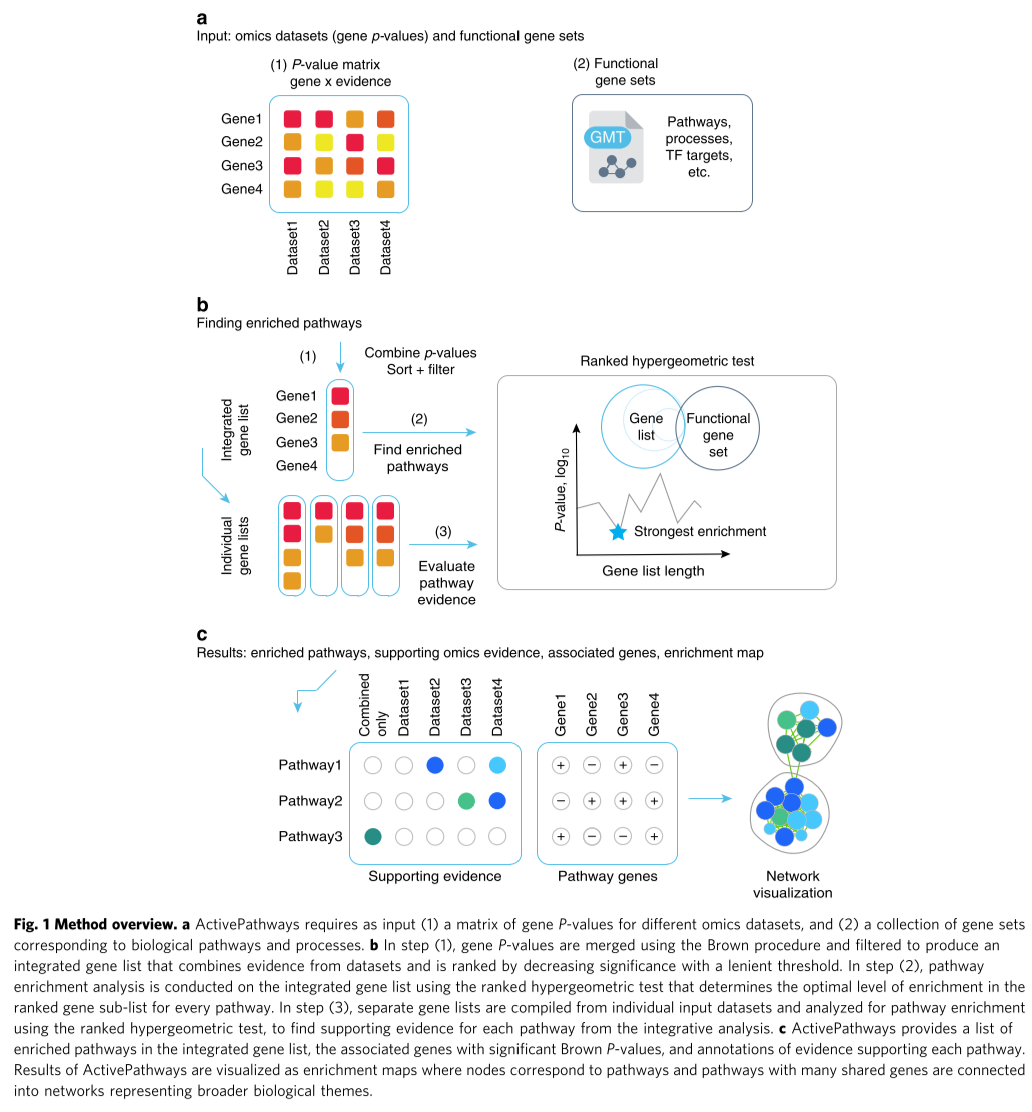
**Integrative pathway enrichment analysis of multivariate omics data**

1. Abstract:

Multi-omics datasets represent distinct aspects of the central dogma of molecular biology. Such high-dimensional molecular profiles pose challenges to data interpretation and hypothesis generation. ActivePathways is an integrative method that discovers significantly enriched pathways across multiple datasets using statistical data fusion, rationalizes contributing evidence and highlights associated genes. As part of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium, which aggregated whole genome sequencing data from 2658 cancers across 38 tumor types, we integrated genes with coding and non-coding mutations and revealed frequently mutated pathways and additional cancer genes with infrequent mutations. We also analyzed prognostic molecular pathways by integrating genomic and transcriptomic features of 1780 breast cancers and highlighted associations with immune response and anti-apoptotic signaling. Integration of ChIP-seq and RNA-seq data for master regulators of the Hippo pathway across normal human tissues identified processes of tissue regeneration and stem cell regulation. ActivePathways is a versatile method that improves systems-level understanding of cellular organization in health and disease through integration of multiple molecular datasets and pathway annotations.

1. Introduction

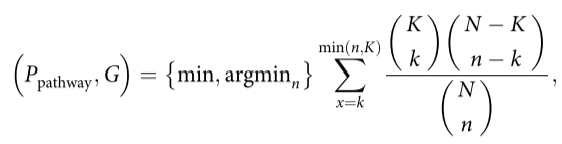
* GSEA detects up- and downregulated pathways in gene expression data sets
* ActivePathways method uses data fusion techniques to address challenge of integrative pathway analysis of multi-omics data
* It detects significantly enriched pathways across multiple datasets, including those pathways that are not apparent in any individual dataset
* Three step method
* Requires two input datasets: one table of P-values with genes as rows. Secondly, a collection of gene sets representing collective knowledge of gene function and interactions (pathways)
* First step: derive an integrated gene list that for each input gene aggregates significance from multiple omics datasets. Compiled by fusion of gene significance from different omics datasets using Brown’s extension. Ranking by decreasing significance and filtered
* Second step: Pathway enrichment analysis using a ranked hypergeometric test and collection of pathways
* Third step: similar analysis on the gene lists of individual omics datasets separately to determine the omics evidence supporting the integrative pathway analysis results determined in step 2
* Visualisation using enrichment map



1. Methods

I**ntegrated and evidence-based gene lists.** The main input of ActivePathways is a matrix of P-values where rows include all genes of a genome and columns correspond to evidence from omics datasets. To interpret multiple omics datasets, a combined P-value is computed for each gene using a data fusion approach, resulting in an integrated gene list. The integrated gene list is computed by merging all P-values of a given gene into one combined P-value using the Brown’s extension18 of the Fisher’s combined probability test that accounts for overall covariation of P-values from different sources of evidence. The integrated gene list of Brown P-values is then ranked in order of decreasing significance and filtered using a lenient threshold (unadjusted P < 0.1 by default). Evidence-based gene lists representing different omics datasets are based on ranked P-values from individual columns of the input matrix and filtered using the same significance threshold.

**Statistical enrichment of pathways.** Statistical enrichment of pathways in ranked lists of candidate genes is carried out with the ranked hypergeometric test. The test considers one pathway gene set at a time and analyses increasing subsets of input genes from the top of the ranked gene list. The same procedure is used for integrated and evidence-based gene lists. At each iteration, the test computes the hypergeometric enrichment statistic and P-value for the set of genes shared by the pathway and top sub-list of the input gene list. For optimal processing speed, only gene lists ending with a pathway-related gene are considered. The ranked hypergeometric statistic selected the input gene sub-list that achieves the strongest enrichment and the smallest P-value as the final result for the given pathway, as:



where Ppathway stands for the hypergeometric P-value of the pathway enrichment at the optimal sub-list of the significance-ranked candidate genes, G represents the length of the optimal sub-list, i.e., the number of top genes from the input gene list, N is the number of protein-coding genes with annotations in the pathway database, i.e., in Gene Ontology and Reactome, K is the total number of genes in a given pathway, n is the number of genes in a given gene sub-list considered, and k is the number of pathway genes in the considered sub-list. For a conservative estimate of pathway enrichment, we consider as background N the universe of genes contained in input gene sets (terms from pathway databases and ontologies) rather than the complete repertoire of protein-coding genes. To obtain candidate genes involved in the pathway of interest, we intersect pathway genes with the optimal sub-list of candidate genes. The ranked hypergeometric P-value is computed for all pathways and resulting P-values are corrected for multiple testing using the HolmBonferroni method of family-wise error rate (FWER)19. Significant pathways are reported by default (Q < 0.05).