# **PW-pipeline**

PathWay pipeline using GWAS summary statistics, named analogously after FM-pipepline I have implemented.

## **INTRODUCTION**

Pathway analysis becomes an important element in GWAS. Broadly, it involves SNP annotation, such as Variant Effect Predictor (VEP), gene analysis such as VEGAS2, and gene set analysis. Visualisation of a particular region has been facilitated with LocusZoom, while network(s) from pathway analysis via gephi or Cytoscape, which uses genes and a collection of edges, directed or undirected, to build a network. Aspects to consider include part or all databases, individual level genotype data vs GWAS summary statistics, computing speed, with and without tissue enrichment.

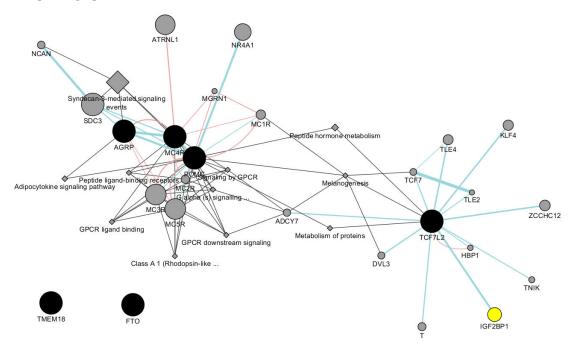


diagram from CytoScape/GeneMANIA

## **INSTALLATION**

This pipeline involves several software for pathway analysis using GWAS summary statistics, as shown below,

Full name	Abbreviation	Reference
Meta-Analysis Gene-set Enrichment of variaNT Associations	MAGENTA	Segre, et al. (2010)
Multi-marker Analysis of GenoMic Annotation	MAGMA	de Leeuw, et al. (2015)

PAthway SCoring ALgorithm PASCAL Lamparter, et al.

(2016)

Data-Driven Expression Prioritized Integration for DEPICT Pers, et al.(2015)

**Complex Traits** 

The full functionality of the pipeline requires availability of individual software for pathway analysis, which should fulfil their requirements, e.g., Matlab for MAGENTA, PLINK. It is useful to install XpdfReader or ImageMagick to produce Excel workbook. By default Sun grid engine is used but this can be any other mechanism such as GNU parallel [note with its --env to pass environment variables]. As usual, R is required.

The pipeline itself can be installed from GitHub in the usual way.

git clone https://github.com/jinghuazhao/PW-pipeline

## **USAGE**

The pipeline requires users to specify both software and database to be used. It is possible that a given database can be used for several software when appropriate.

The syntax is

bash pwp.sh <input file>

## Input

The input will be GWAS summary statistics described at SUMSTATS in that order without the header,

## Output

The output will be available from individual directories named after the software you choose, and optionally in case all software are used the output can also be an Excel workbook containing combined results.

For DEPICT databases, it is possible to call <a href="network-plot.py">network diagram</a> and perform cluster analysis.

#### **EXAMPLES**

The bmi.txt and ST4 from SUMSTATS can be called as follows,

pwp.sh bmi.txt

and

pwp.sh ST4 &

## **ADDITIONAL TOPICS**

The wiki page contains the following information,

- Databases
- Features
- Tissue and network plots
- Result collection

#### **RELATED LINKS**

- BioGRID: an interaction repository with data compiled through comprehensive curation efforts.
- Osprey: Network Visualization System.
- GeneMANIA: Imports interaction networks from public databases from a list of genes with their annotations and putative functions.
- rGREAT: Client for GREAT Analysis
- VisANT: Visual analyses of metabolic networks in cells and ecosystems.

#### **ACKNOWLEDGEMENTS**

The work drives from comparison of software performances using our own GWAS data. The practicality of a common DEPICT database to all software here was due to PASCAL developer(s). At the end of our implementation it came to our attention that similar effort has been made, e.g., DEPICT-pipeline and other adaptations.

## **SOFTWARE AND REFERENCES**

## **DEPICT** (GitHub)

Pers TH et al.(2015) Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun. 6:5890. doi: 10.1038/ncomms6890.

#### **MAGENTA**

Segre AV, et al (2010). Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. PLoS Genet. 12;6(8). pii: e1001058. doi: 10.1371/journal.pgen.1001058

#### MAGMA

de Leeuw C, et al. (2015) MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput Biol. 11(4): e1004219. doi: 10.1371/journal.pcbi.1004219

## PASCAL (GitHub)

Lamparter D, et al. (2016) Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLoS Comput Biol. 2016 Jan 25;12(1):e1004714. doi: 10.1371/journal.pcbi.1004714