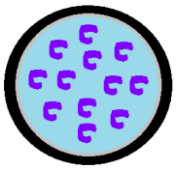
PAEA: Principle Angle Enrichment Analysis



Web Application User Manual

Written by Zichen Wang and Avi Ma'ayan

Updated on 4/21/2015

Table of Contents

Abstract

Functional analysis of genome-wide differential expression is central to biological investigations. Here we present a new multivariate approach to gene-set enrichment called Principal Angle Enrichment Analysis (PAEA). PAEA uses the geometrical concept of the principal angle to quantify gene-set enrichment. We find that PAEA outperforms a selection of commonly used gene set enrichment methods including GSEA. To benchmark PAEA with other enrichment methods we use real data. We examined the ranking of transcription factors by performing enrichment analysis on gene expression signatures from many studies that knocked-down, knocked-out or over-expressed transcription factors, and performed the enrichment analysis with a library of gene sets created from ChIP-Seq data profiling the same transcription factors. We also found that PAEA was able to rank better aging-related phenotype-terms from a collection of gene expression profiling studies where tissue from young adults was compared to tissue of elderly subjects. PAEA is implemented as a user-friendly R Shiny gene-set enrichment web application with over 70 gene set libraries available for enrichment analysis. Canned enrichment analysis for over 700 disease signatures extracted from GEO is provided with the application which is freely available at: http://amp.pharm.mssm.edu/PAEA.

Installation and Requirements

The PAEA web application requires Internet connection and a modern browser capable of supporting HTML5. Such browsers include Google Chrome 10, FireFox 3.6, Opera 11.01, Safari 5 and IE10 or higher versions. To test your browser compatibility you should be able to see the following screen view (Fig1) when pointing your browser to http://amp.pharm.mssm.edu/PAEA.

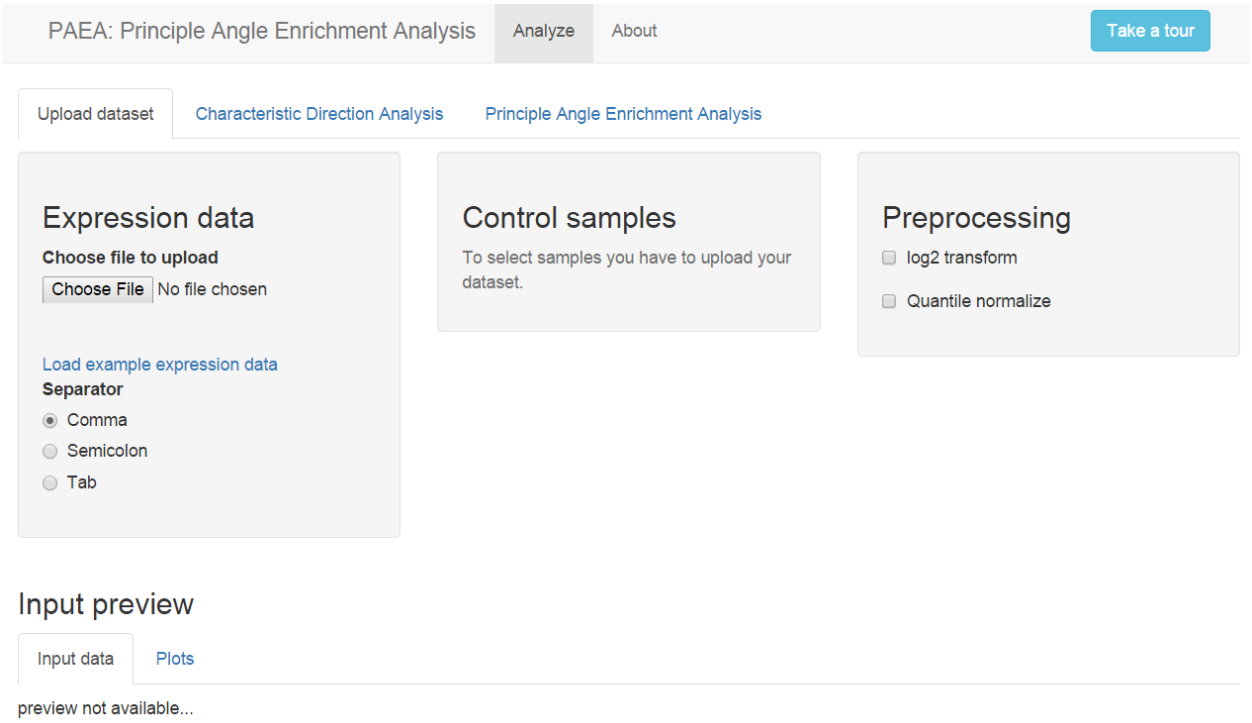


Figure 1 The PAEA web application user interface.

PAEA Use Cases

The PAEA web application can be used for analyzing custom gene/protein expression datasets as well as exploring the enriched biological terms for over 700 disease signatures extracted from the Gene Expression Omnibus (GEO).

Analyzing your own gene or protein expression data

To analyze your own gene or protein expression dataset, the expression data should be uploaded to the PAEA web application through the "Upload dataset" tab. The dataset should be stored in a plain-text file. The data in the text file should be organized as a table where entries are tab, semi-colon or comma delimited. In this data file, the first column should be gene names, and the other columns should be expression values across samples. The first row should be a header specifying the names of samples. A valid example dataset file is shown in Fig2. Although opened in Excel, it is a text file.

	A	B	C	D	E	F	G
1	IDENTIFIER	GSM526561	GSM526562	GSM526602	GSM526603	GSM526605	GSM526678
2	Pdhb	6.78172	6.67057	6.67767	6.61841	6.38248	6.71026
3	Lypla1	10.0028	10.0212	9.98037	9.95486	9.94111	10.0634
4	Tcea1	9.00629	9.04831	8.98056	9.10134	9.00104	8.83338
5	Atp6v1h	8.35819	8.35274	8.36684	8.33537	8.36566	8.35614
6	Oprk1	4.12316	3.88417	4.00814	3.81399	3.97681	3.99824
7	Rb1cc1	7.50064	7.44986	7.42371	7.32109	7.42521	7.47707
8	Fam150a	2.79769	2.89569	3.02865	3.06864	2.82011	2.98995
9	St18	4.70729	4.86906	5.02481	5.0828	4.93599	4.96627
10	Pcmtd1	8.81027	8.89916	8.8256	8.76545	8.71039	8.78622
11	Ahcy	12.9397	12.9809	12.9608	12.9495	12.8176	12.8718
12	Rrs1	6.58592	6.55599	6.51913	6.58703	6.73596	6.70705

Figure2 Snapshot of a properly formatted dataset file that can be used as input to PAEA

Once the dataset file is successfully uploaded, the uploaded data is available for preview in a searchable table as well as through various plots (Fig. 3).

With these plots users can examine the distribution of the expression values of genes within each sample. It is highly recommended to perform log2 transformation at the "Preprocessing" step if the data is not normally distributed. Control samples should be chosen using the check boxes. Unselected samples are automatically considered experimentally perturbed samples. There should be at least two samples (replicates) in each condition. This is a requirement for all statistical tests that perform differential expression analysis including the Characteristic Direction method [1].

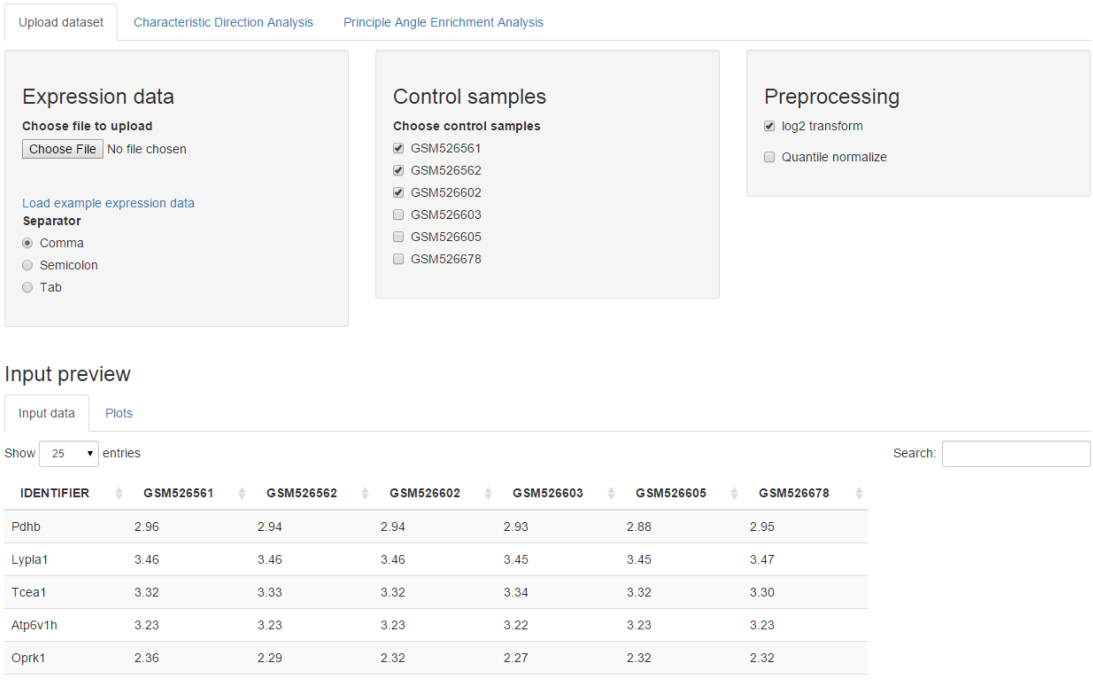


Figure 3 Preview of a table that displays the uploaded dataset

The Characteristic Direction (CHDIR) analysis can be performed in the second tab of the PAEA web-application to compute the differentially expressed genes (DEGs). Parameters for the CHDIR include: Gamma, which is the shrinkage parameter used for regularization. This should be a number between 0 and 1. Nnull is the number of random vectors used to estimate the significance observed CHDIR vector; and random seed, which is needed for the reproducibility of the results. Once the CHDIR computation is complete, the results are displayed in a bar graph showing the top DEGs (Fig. 4). A positive coefficient denotes up-regulated genes, and a negative coefficient denotes down-regulated genes. Prioritized DEGs can also be downloaded as text file for analysis with other tools.

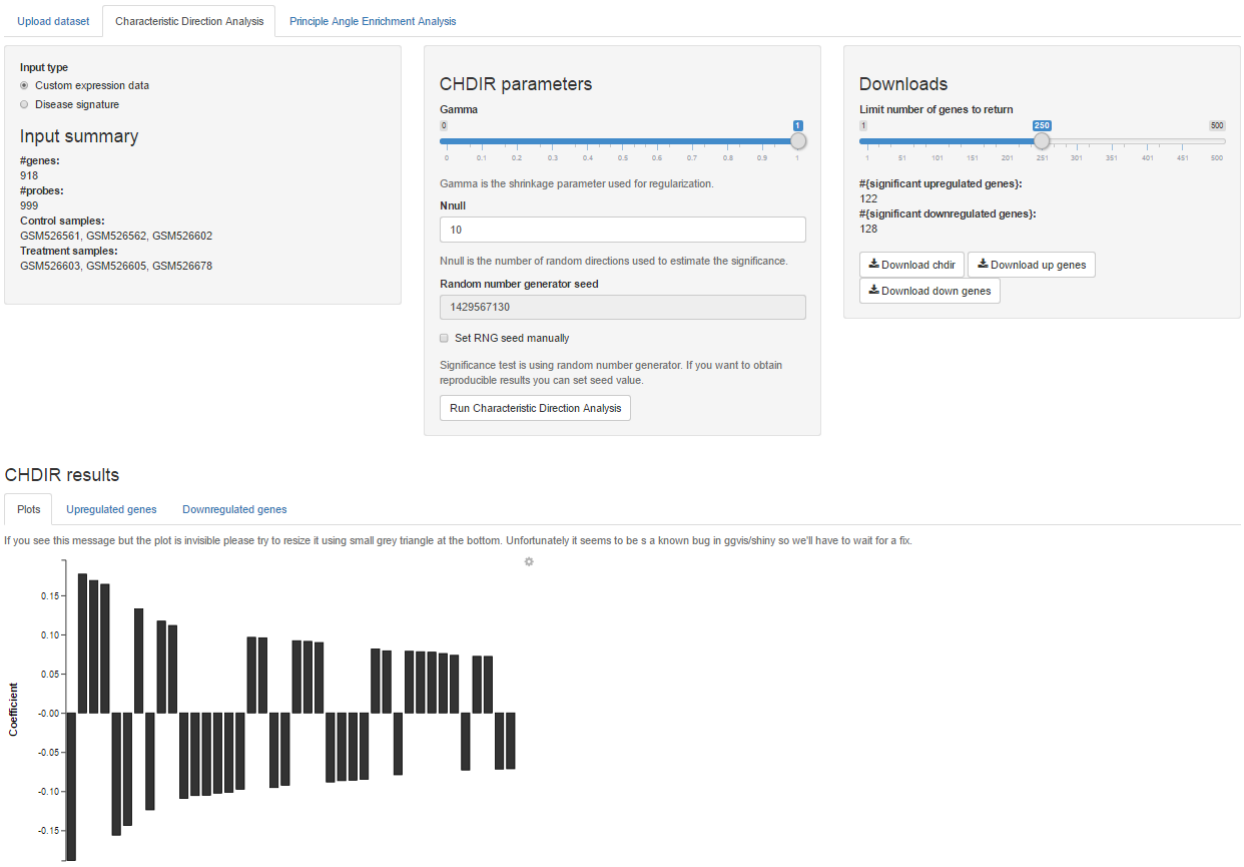
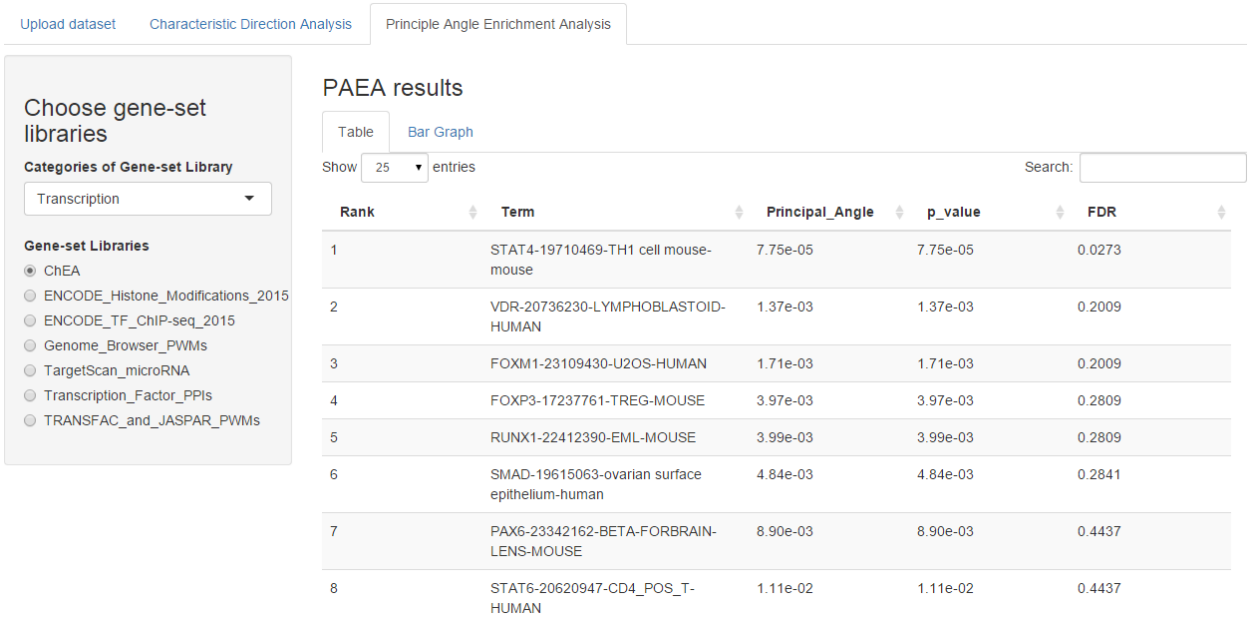


Figure 4 Visualization of the CHDIR results

Once the CHDIR analysis is completed, users can switch to the Principle Angle Enrichment Analysis tab to perform enrichment analysis applied to the computed expression signature (Fig. 5). There are over 70 gene-set libraries supported by PAEA covering different the following categories: Transcription, Pathways, Ontologies, Drugs/Diseases, Cell Type and Miscellaneous. These libraries are borrowed from the tool Enrichr [2] and use the same underlying database. PAEA analysis will be automatically performed once a gene-set library is selected. Enrichment analysis for some gene-set libraries may take over >10 seconds to run, and thus a progress indicator is provided. Once the PAEA analysis is complete,

enriched biological terms are sorted by their significance and displayed in a searchable table. A bar graph is provided as an alternative visualization of the results, showing the top enriched terms.



Rank	Term	Principal_Angle	p_value	FDR
1	STAT4-19710469-TH1 cell mouse-mouse	7.75e-05	7.75e-05	0.0273
2	VDR-20736230-LYMPHOBLASTOID-HUMAN	1.37e-03	1.37e-03	0.2009
3	FOXM1-23109430-U2OS-HUMAN	1.71e-03	1.71e-03	0.2009
4	FOXP3-17237761-TREG-MOUSE	3.97e-03	3.97e-03	0.2809
5	RUNX1-22412390-EML-MOUSE	3.99e-03	3.99e-03	0.2809
6	SMAD-19615063-ovarian surface epithelium-human	4.84e-03	4.84e-03	0.2841
7	PAX6-23342162-BETA-FORBRAIN-LENS-MOUSE	8.90e-03	8.90e-03	0.4437
8	STAT6-20620947-CD4_POS_T-HUMAN	1.11e-02	1.11e-02	0.4437

Figure 5 Table displaying PAEA results using the ChEA gene set library [3]

Exploring the mechanisms of disease with the canned disease expression signatures

The PAEA web application comes with canned analysis for over 700 disease gene expression signatures extracted from GEO. These signatures were extracted manually by identifying genome-wide gene expression studies where normal healthy tissue was compared with disease tissue samples. To explore mechanisms of disease, and identify potential drugs that can reverse expression in disease, users can directly switch to the CHDIR tab (Fig. 6) and check the "Disease signature" option. Once this option is checked, users can search for diseases by name. Metadata about each signature includes the tissue/cell-line and the GEO accession numbers of the study that performed the expression experiments. Once the disease signature is loaded, users can switch to the PAEA tab to perform enrichment analyses.

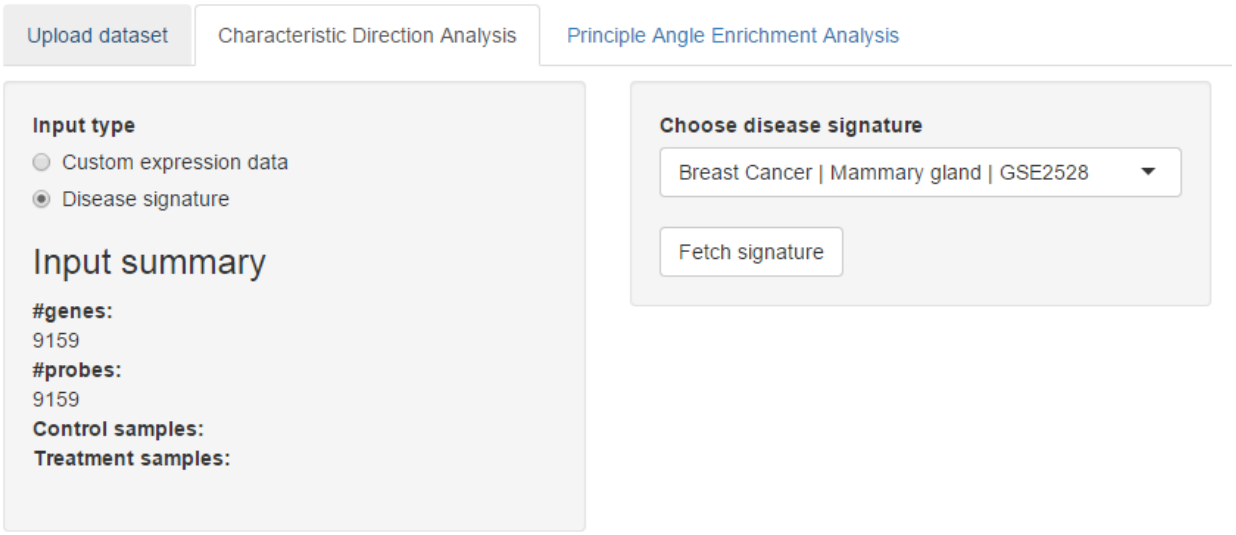


Figure 5 Loading of a disease signatures

References

1. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, Ma'ayan A: **The characteristic direction: a geometrical approach to identify differentially expressed genes**. *BMC Bioinformatics* 2014, **15**:79.
2. Chen E, Tan C, Kou Y, Duan Q, Wang Z, Meirelles G, Clark N, Ma'ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool**. *BMC bioinformatics* 2013, **14**(1):128.
3. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A: **ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments**. *Bioinformatics* 2010, **26**(19):2438-2444.