BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 20 2011, pages 2917–2918
doi:10.1093/bioinformatics/btr499

Systems biology Advance Access publication September 4, 2011

Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA

Atanas Kamburov[1,*,†], Rachel Cavill[2,3,*,†], Timothy M. D. Ebbels[3], Ralf Herwig[1] and Hector C. Keun[3,*]

[1]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany, [2]Department of Toxicogenomics, Maastricht University, Universiteitssingel 50, Maastricht, The Netherlands and [3]Biomolecular Medicine, Department of Surgery and Cancer, Imperial College London, London SW7 2AZ, UK

Associate Editor: John Quackenbush

ABSTRACT

Summary: Pathway-level analysis is a powerful approach enabling interpretation of post-genomic data at a higher level than that of individual biomolecules. Yet, it is currently hard to integrate more than one type of omics data in such an approach. Here, we present a web tool 'IMPaLA' for the joint pathway analysis of transcriptomics or proteomics and metabolomics data. It performs over-representation or enrichment analysis with user-specified lists of metabolites and genes using over 3000 pre-annotated pathways from 11 databases. As a result, pathways can be identified that may be disregulated on the transcriptional level, the metabolic level or both. Evidence of pathway disregulation is combined, allowing for the identification of additional pathways with changed activity that would not be highlighted when analysis is applied to any of the functional levels alone. The tool has been implemented both as an interactive website and as a web service to allow a programming interface.

Availability: The web interface of IMPaLA is available at http://impala.molgen.mpg.de. A web services programming interface is provided at http://impala.molgen.mpg.de/wsdoc.

Contact: kamburov@molgen.mpg.de; r.cavill@imperial.ac.uk; h.keun@imperial.ac.uk

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 9, 2011; revised on July 26, 2011; accepted on August 11, 2011

1 INTRODUCTION

Systems biology aims at the concerted analysis of biological systems at different levels, for example the combination of transcriptomics, proteomics and metabolomics. Biochemical pathways are the primary focus of systems biology. Pathways are extensively used to interpret omics data, for example to gain mechanistic insight into gene disregulation, which is causative or indicative of complex diseases. In particular, pathway over-representation (ORA) and enrichment analyses have become important tools for the interpretation of data from transcriptomics (Riedel *et al.*, 2008) and metabolomics (Sabatine *et al.*, 2005) experiments.

Several web-based tools exist for such pathway analyses on transcriptomic or metabolomic data separately (Chagoyen and Pazos, 2011; Huang *et al.*, 2008; Kamburov *et al.*, 2011; Xia and Wishart, 2010), but to our knowledge no tools exist yet for integrated pathway analysis with both types of data simultaneously. In a recent study (Cavill *et al.*, 2011), a method for the integrated analysis of transcriptomic and metabolomic data was proposed that exploits the fact that genes and metabolites are linked through biochemical reactions and thus are contained in many pathways. Here, we report the implementation of this method, providing a web server called Integrated Molecular Pathway-Level Analysis (IMPaLA) for the combined analysis of gene/protein and metabolite datasets using a comprehensive basis of biochemical pathways currently taken from 11 publicly available resources. We illustrate the use of the website and exemplify the benefit of combined analysis of transcriptomics and metabolomics data. IMPaLA is accessible through an interactive website or through web services.

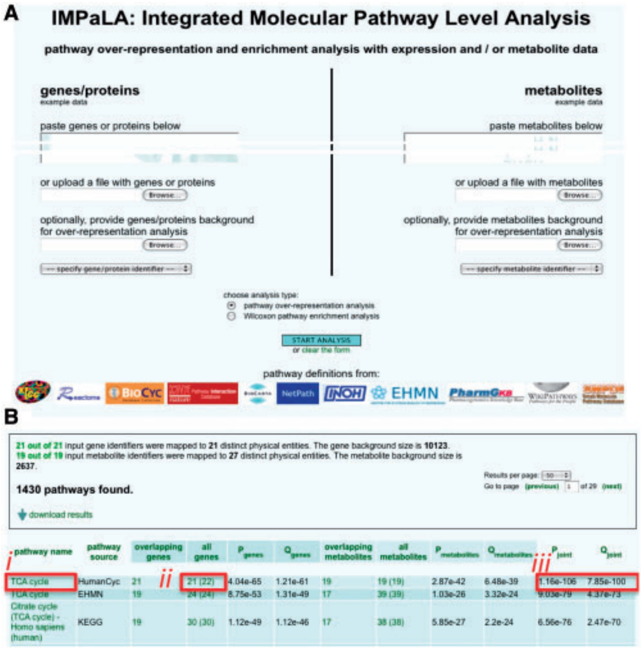
2 DESCRIPTION

Web interface: the web interface at http://impala.molgen.mpg.de gives the user the possibility to upload genes and/or metabolites in several identifier namespaces and to perform over-representation or Wilcoxon enrichment analysis (WEA) (Adjaye *et al.*, 2005) with the available pathways (Fig. 1A). These pathways (currently 3073) originate from 11 public databases such as Reactome (http://www.reactome.org), KEGG (http://www.genome.jp/kegg/) or Wikipathways (http://www.wikipathways.org). For ORA, the user uploads lists of identifiers that typically represent genes/proteins/metabolites significantly associated with the effect of interest. In addition, background lists of identifiers representing all measured genes/proteins and/or metabolites can be uploaded. This is especially useful when the list of measured entities is small compared with the number of genes/proteins or metabolites in the organism, to avoid potential bias. If no background lists are specified by the user, all entities present in pathways and annotated in the user-specified identifier namespaces are used as default background lists. Based on the uploaded lists, the hypergeometric distribution is used to assess the significance of each pathway in terms of its overlap with those lists.

For WEA, the user uploads lists of all measured genes/proteins and/or metabolites (rather than just the significant ones as in ORA) with either one or a pair of numerical values per entity. These values

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.


A
IMPaLA: Integrated Molecular Pathway Level Analysis
pathway over-representation and enrichment analysis with expression and / or metabolite data
genes/proteins
example data
paste genes or proteins below
or upload a file with genes or proteins
Browse...
optionally, provide genes/proteins background
for over-representation analysis
Browse...
--- specify gene/protein identifier ---
metabolites
example data
paste metabolites below
or upload a file with metabolites
Browse...
optionally, provide metabolites background
for over-representation analysis
Browse...
--- specify metabolite identifier ---
choose analysis type:
● pathway over-representation analysis
○ Wilcoxon pathway enrichment analysis
START ANALYSIS
or clear the form
pathway definitions from:
KEGG
Reactome
BioCyc
Pathway Interaction Database
BioCarta
NetPath
INOH
EHMN
PharmGKB
WikiPathways
Small Molecule Pathway Database
B
Results per page: 50
Go to page (previous) 1 of 29 (next)
download results
i
ii
iii

21 out of 21 input gene identifiers were mapped to 21 distinct physical entities. The gene background size is 10123.
19 out of 19 input metabolite identifiers were mapped to 27 distinct physical entities. The metabolite background size is 2637.

1430 pathways found.

pathway name	pathway source	overlapping genes	all genes	P_{genes}	Q_{genes}	overlapping metabolites	all metabolites	$P_{metabolites}$	$Q_{metabolites}$	P_{joint}	Q_{joint}
TCA cycle	HumanCyc	21	21 (22)	4.04e-65	1.21e-61	19	19 (19)	2.87e-42	6.48e-39	1.16e-106	7.85e-100
TCA cycle	EHMN	19	24 (24)	8.75e-53	1.31e-49	17	39 (39)	1.03e-26	3.32e-24	9.03e-79	4.37e-73
Citrate cycle (TCA cycle) - Homo sapiens (human)	KEGG	19	30 (30)	1.12e-49	1.12e-46	17	38 (38)	5.85e-27	2.2e-24	6.56e-76	2.47e-70

Fig. 1. (**A**) IMPaLA (version 1) input screen including the logos of the 11 source databases. (**B**) Output screen with a ranked list of pathways showing; (i) a link to each pathway in the source database; (ii) the size of each pathway in terms of entities also present in the background list, followed by the number of all pathway entities as in the source database; (iii) the *P*- and *Q*-values from the joint analysis with genes and metabolites, calculated as in Cavill *et al.* (2011).

may represent fold changes or average expression/concentration values for two different experimental conditions. The values are used to assess the joint expression/concentration difference of all entities contained in each pathway through Wilcoxon's signed-rank test. Even if a pathway contains no individual entities with significant differential expression or concentration, the joint expression/concentration of the group of pathway members may be significantly changed, indicating potential pathway disregulation on a low but nonetheless consistent level. Results from both ORA and WEA analyses are presented by IMPaLA as tables listing pathways that contain at least one gene and/or metabolite from the input lists (Fig. 1B). Information about pathway name, source, size and overlap with the input entities is provided along with *P*-values calculated with appropriate statistical test for each pathway. Notably, if both metabolites and genes/proteins are uploaded by the user, a joint *P*-value is given, calculated as per Cavill *et al.* (2011). To control for multiple testing, *Q*-values are calculated with the false discovery rate method (Benjamini and Hochberg, 1995). Results can be sorted on any column by clicking on the appropriate column header, and can be downloaded as a tab-delimited file. By clicking on a pathway name, the user is guided to a summary web page at the original source database, which in most cases also shows a detailed pathway diagram.

Example: using publicly available data from the NCI60 (Scherf *et al.*, 2000), we selected the genes and metabolites that were significantly correlated with the GI50 values for the common cancer therapeutic 5-flurouracil (5-FU) across the 58 cell lines as in previous work (Cavill *et al.*, 2011) (see Files S1–S4 in Supplementary Material that includes background). Full results of the IMPaLA ORA analysis on this data can be found in File 5 in Supplementary Material. The top seven pathways are based entirely on the over-representation of genes, and mainly relate to the ribosome or to eukaryotic translation. For the ABC Transporter pathway, neither genes nor metabolites alone gave $P < 0.05$, yet $P_{\text{joint}} = 0.049$. This example demonstrates that metabolic information gives added value to the genome-wide analysis enhanced the pathway recovery. For further examples and help, please see the tutorial document provided in File S7 in Supplementary Material.

Web service: in addition to the standard interactive web interface, the functionality provided can also be accessed through a SOAP web service. Here, functions are available that carry out ORA or WEA with lists of genes, of metabolites or both. The web service definition (WSDL) file and the appropriate documentation are available at http://impala.molgen.mpg.de/wsdoc.

Funding: International Max Planck Research School for Computational Biology and Scientific Computing and EU APO-SYS project (HEALTH-F4-2007-200767 to A.K.); EU CarcinoGENOMICS consortium, contract No. (PL037712 to R.C., T.M.D.E. and H.C.K.); BMBF NGFN-transfer project (01GR0809 to R.H.).

Conflict of Interest: none declared.

REFERENCES

Adjaye,J. *et al.* (2005) Primary differentiation in the human blastocyst: Comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells*, **23**, 1514–1525.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Cavill,R. *et al.* (2011) Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Comput. Biol.*, **7**, e1001113.

Chagoyen,M. and Pazos,F. (2011) MBRole: enrichment analysis of metabolomic data. *Bioinformatics*, **27**, 730–731.

Huang,D.W. *et al.* (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*, **4**, 44–57.

Kamburov,A. *et al.* (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, **39**, D712–D717.

Riedel,R.F. *et al.* (2008) A genomic approach to identify molecular pathways associated with chemotherapy resistance. *Mol. Cancer Therap.*, **7**, 3141–3149.

Sabatine,M.S. *et al.* (2005) Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation*, **112**, 3868–3875.

Scherf,U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.

Xia,J. and Wishart,D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.*, **38**, W71–W77.