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

Atanas Kamburov1,*,†, Rachel Cavill2,3,*,†, Timothy M. D. Ebbels3, Ralf Herwig1 and Hector C. Keun5,*

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

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.

*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.
WEA analyses are presented by IMPaLA as tables listing pathways in Cavill et al. number of all pathway entities as in the source database; (iii) the 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-fluorouracil (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 S5 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 Pjoint = 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. (PI037712 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


