

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

Atanas Kamburov<sup>1,\*</sup>, Rachel Cavill<sup>2,3,\*</sup>, Timothy M. D. Ebbels<sup>3</sup>, Ralf Herwig<sup>1</sup> and Hector C. Keun<sup>3,\*</sup>

<sup>1</sup>Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany, <sup>2</sup>Department of Toxicogenomics, Maastricht University, Universiteitssingel 50, Maastricht, The Netherlands and <sup>3</sup>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 dysregulated on the transcriptional level, the metabolic level or both. Evidence of pathway dysregulation 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 dysregulation, 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.

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

