HTSanalyzeR: a R/Bioconductor package for integrated network analysis of high-throughput RNAi screens

Camille Terfve 1,†, John C. Rose 1,†, Xin Wang 2,3 and Florian Markowetz 2,3,*
1Cambridge Computational Biology Institute and Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
2Department of Oncology, University of Cambridge, UK
3Cancer Research UK Cambridge Research Institute, Robinson Way, Cambridge CB2 0RE, UK
Received on XXXXX; revised on XXXXX; accepted on XXXXX

ABSTRACT

Motivation: High-throughput screens (HTS) by RNAi or small molecules are among the most promising tools in functional genomics. They enable researchers to observe detailed reactions to experimental perturbations on a genome-wide scale. While there is a core set of computational approaches used in many publications to analyze these data, a specialized software combining them and making them easily accessible has so far been missing.

Results: Here we describe HTSanalyzeR, an integrated analysis pipeline for HTS data that contains over-representation analysis, gene set enrichment analysis, comparative gene set analysis and rich sub-network identification. HTSanalyzeR directly builds on commonly used pre-processing packages for HTS data and presents its statistical results as HTML pages and network plots.

Availability: Our software is written in R language and freely available via the Bioconductor project at http://www.bioconductor.org.
Contact: florian.markowetz@cancer.org.uk

1 INTRODUCTION

In recent years several technological advances have pushed gene perturbation screens to the forefront of functional genomics. Combining high-throughput screening (HTS) techniques with rich phenotypes enables researchers to observe detailed reactions to experimental perturbations on a genome-wide scale. This makes HTS one of the most promising tools in functional genomics.

Although the phenotypes in HTS data correspond to single genes, it becomes more and more important to analyze them in the context of cellular pathways and networks to understand how genes work together. Network analysis of HTS data depends on the dimensionality of the phenotypic readout (Markowetz, 2010). While specialized analysis approaches exist for high-dimensional phenotyping (e.g. Fröhlich et al., 2008), analysis approaches for low-dimensional screens have so far been spread out over diverse softwares and online tools like DAVID (Huang et al., 2009) or gene set enrichment analysis (GSEA, Subramanian et al., 2005)).

Here we provide an integrated analysis pipeline for HTS data that contains gene set and network analysis approaches commonly used in many papers (as reviewed by Markowetz, 2010). HTSanalyzeR is written in R (R Development Core Team, 2009) and freely available via the Bioconductor project (Gentleman et al., 2004). The software interfaces directly with existing HTS pre-processing packages like cellHTS2 (Boutros et al., 2006) or RNAither (Rieber et al., 2009). Additionally, our software is in the process of being fully integrated in a web-interface for the analysis of HTS data (Pelz et al., 2010) and will thus be easily accessible to non-programmers.

2 AN INTEGRATED ANALYSIS PIPELINE FOR HIGH-THROUGHPUT SCREENING DATA

HTSanalyzeR takes as input HTS data that has already undergone preprocessing and quality control (e.g. by using cellHTS2). It then functionally annotates the hits by gene set enrichment and network analysis approaches (see Figure 1 for an overview).

Gene set analysis. HTSanalyzeR implements two approaches: (i) hypergeometric tests for surprising overlap between hits and gene sets, and (ii) cut-off free gene set enrichment analysis which measures if a gene set shows a concordant trend to stronger phenotypes. HTSanalyzeR uses gene sets from MSigDB (Subramanian et al., 2005), the Gene Ontology (Ashburner et al., 2000) and KEGG (Kanehisa et al., 2006). The accompanying vignette explains how user-defined gene sets can easily be included.

Network analysis. In a complementary approach strong hits are mapped to a network and enriched subnetworks are identified. Networks can come from different sources, especially protein interaction networks are often used. In HTSanalyzeR we use networks defined in the BioGRID database (Stark et al., 2006), but other user-defined networks can easily be included in the analysis. To identify rich subnetworks, we use the BioNet package (Beisser et al., 2010), which in its heuristic version is fast and produces close-to-optimal results.

Comparing phenotypes. A goal we expect to become more and more important in the future is to compare phenotypes for the same genes in different cellular conditions. HTSanalyzeR supports comparative analyses for gene sets and networks. Differentially enriched gene sets are computed by comparing GSEA enrichment scores or alternatively by a Wilcoxon test statistic. Subnetworks

* to whom correspondence should be addressed. † These authors contributed equally.
rich for more than one phenotype can easily be found with BioNet (Beisser et al., 2010).

Parameters and report. Each of these analysis methods depends on several input parameters. While every one of them can be changed in the package, HTSanalyzeR also implements a standard analysis option using default parameters that we have found to work well in many applications. Results are presented in an HTML format similar to cellHTS2. Overrepresentation and enrichment results are presented as tables, where gene sets are linked to their descriptions at EBI and KEGG pages. GSEA results are accompanied by enrichment plots similar to the ones in (Subramanian et al., 2005).

3 AN EXAMPLE SESSION

HTSanalyzeR directly interfaces seamlessly with cellHTS2, the most widely used preprocessing package for HTS data. We exemplify its use on the example data set included in cellHTS2, a genome-wide RNAi screen for viability in Drosophila melanogaster.

First, the user needs to load the packages HTSanalyzeR, org.Dm.eg.db, GO.db, KEGG.db, and cellHTS2. The cellHTS2 vignette describes how to generate an object \( x \) containing the normalized data. This object is the input to HTSanalyzeR, as the following code chunk shows:

```r
HTSanalyzeR(
  x=xn,
  annotationColumn="GeneID",
  species ="Drosophila melanogaster",
  initialIDs="FlybaseCG",
  listOfGeneSetCollections= gsc,
  networkObject= netobj
)
```

The input gsc contains a user-defined list of gene sets that can e.g. be generated from GO or KEGG using the functions GOGeneSets or KeggGeneSets included in HTSanalyzeR. The input netobj describes the network and by default loads protein-interactions from BioGRID.

ACKNOWLEDGEMENTS

We thank Oliver Pelz and Michael Boutros for integrating HTSanalyzeR into the web-cellHTS interface.

Funding: This work was supported by Cancer Research UK. CT was funded by the Fondation Philippe Wiener - Maurice Anspach.

REFERENCES


