Bioinformatics Advance Access published May 14, 2014

RNASeqExpressionBrowser - A web interface to browse and visualize high-throughput expression data

Thomas Nussbaumer [1], Karl G. Kugler [1], Kai C. Bader [1], Sapna Sharma [1], Michael Seidel [1], and Klaus F.X. Mayer [1,*]

[1]Institute for Bioinformatics and Systems Biology (IBIS), Helmholtz Center Munich, National Research Center for Environment and Health, Ingolstaedter Landstrasse 1, 85764 Neuherberg

Associate Editor: Prof. Ivo Hofacker

ABSTRACT

Motivation: RNA-seq techniques generate massive amounts of expression data. Several pipelines (e.g. Tophat and Cufflinks) are broadly applied to analyse these data sets. However, accessing and handling the analytical output remains challenging for non-experts.

Results: We present the RNASeqExpressionBrowser, an open-source web interface that can be used to access the output from RNA-seq expression analysis packages in different ways as it allows browsing for genes by identifiers, annotations or sequence similarity. Gene expression information can be loaded as long as it is represented in a matrix like format. Additionally, data can be made available by setting up the tool on a public server. For demonstration purposes, we have set up a version providing expression information from the barley genome.

Availability: The source code and a show case are accessible at: `http://mips.helmholtz-muenchen.de/plant/RNASeqExpressionBrowser/`.

Contact: k.mayer@helmholtz-muenchen.de

1 INTRODUCTION

Current high-throughput technologies, such as RNA-seq, allow researchers to generate massive amounts of gene expression data in short time. However, data analysis and intuitive visualization often remain a bottleneck for making efficient use of generated data. For biologists with limited programming experience, accessing data output from RNA-seq pipelines in a meaningful manner poses a challenge: Manual search for genes of interest is time-intensive if feasible at all, and generating non-automated visualizations for a list of genes of interest is a painstaking task. This task has become even more challenging for highly repetitive genomes like barley (The International Barley Sequencing Consortium, 2012) or wheat (Brenchley *et al.*, 2012) where *de novo* transcriptome assemblies result in hundred thousands of transcripts. The Tophat and Cufflinks (Trapnell *et al.*, 2009) workflow presents one of the standard workflows for analyzing RNA-seq data and is broadly applied in life sciences' research. Packages to analyze data from these workflows that help to reduce the efforts for accessing the output data are available, e.g. CummeRbund (Goff *et al.*, 2012), RobiNA (Lohse *et al.*, 2012), or STAR (Wang *et al.*, 2013). However, accessing genes by sequence-information or annotations still requires some programming skills. Additionally, most tools work only locally, which makes sharing of the results difficult in cooperative research projects. In this paper we present the RNASeqExpressionBrowser, a web-based tool that can be used to easily access the results of expression analysis.

*to whom correspondence should be addressed

2 METHODS

A crucial requirement for analyzing expression data is to establish efficient methods for accessing genes of interest. RNASeqExpressionBrowser tackles this issue by enabling several search methods: It allows searching genes by gene annotation (e.g. Gene Ontology (Gene Ontology Consortium, 2004), Interpro), keyword search or via sequence similarity. In addition the user can provide a list of genes as an input in order to inspect and download underlying expression values.

2.1 Data schema

High-throughput data from gene expression assays can be loaded, as long as it can be represented in a matrix-like format, with rows containing feature information (genes or isoforms) and columns harboring sample information. We tested on output generated by Cuffdiff to calculate the expression of transcripts serving as input for the RNASeqExpressionBrowser, but in principle other formats that represent gene expression can be integrated. Following Cufflinks' notation, an experiment comprises a set of conditions (e.g. tissues, time-points or treatments). One condition contains the summarized expression of one or more replicates. In Cufflinks, for each condition all the transcripts are represented by a fragments per kilobase of transcript per Million mapped reads (FPKM) value. In the RNASeqExpressionBrowser several types of annotations can be provided for querying the genes (e.g. GO terms, domains, links to other databases).

2.2 Implementation and installation

The open-source software is intended to run on a Linux operating systems and was implemented in the programming language Python. A MySQL database serves as a back-end, with Google Visualization API and Javascript as main technologies for data representation. For each experiment a separate database, containing expression information and annotation is created. For installing the RNASeqExpressionBrowser a Python-based installation script is provided.

2.3 Finding genes of interest

Within a particular experiment, genes can be searched by using several options: Users can search by gene identifiers, keywords or by annotation

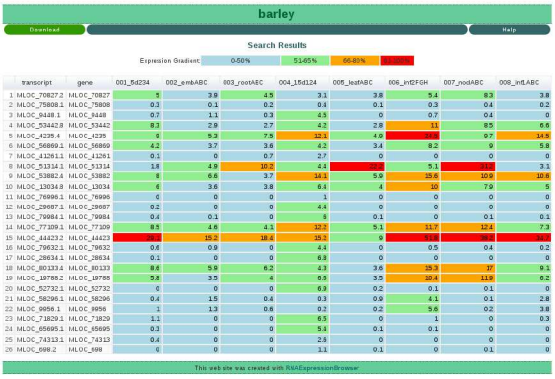

Figure 1. The query results page displays the color coded expression values for the retrieved genes.

information (e.g. GO, Interpro). Since gene homology is often used for finding genes of interest, RNASeqExpressionBrowser also provides a sequence similarity search using BLAST (Altschul *et al.*, 1997). Genes which have been retrieved by the search are listed on a search result page together with their color-coded expression values (Fig. 1).

2.4 Single gene reports

After selecting a gene from the search results page, a detailed gene report is shown, which contains a graphical and tabular summary of the expression data. Furthermore, annotation information and links to external databases are automatically included if provided during the installation.

2.5 Show case data

As a show case, the IBSC Barley expression data (The International Barley Sequencing Consortium, 2012) was used. This data comprises expression from eight tissues and annotations based on GO and Interpro. To demonstrate the linkage to external databases, we include links to PlantsDB (Nussbaumer *et al.*, 2013). This show case is also provided in the installation package.

3 DISCUSSION AND CONCLUSION

Detecting candidate genes that are differentially expressed in a particular condition is an important and frequently used route to approach genes involved in processes of interest. Good expression data is one aspect on the route to approach genes of interest, but of equal importance are robust methods for detecting them (e.g. Trapnell *et al.* (2013); Robinson *et al.* (2010); Anders and Huber (2010)). RNASeqExpressionBrowser provides a bridge between pre-processing and the final steps of the analysis. While many researchers report their findings in a spreadsheet format, an interactive application like RNASeqExpressionBrowser enables to access existing information about predicted genes and their functions very efficiently. In addition, it allows linking the data to existing databases. We are confident that for the broader community this tool will be beneficial for their work with transcriptomics data.

4 AVAILABILITY

On the project website (`http://mips.helmholtz-muenchen.de/plant/RNASeqExpressionBrowser/`), we provide the source code as well as the show case data and additional information regarding the installation. The source code is freely accessible as open source under the LGPL.

ACKNOWLEDGEMENT

We want to thank Wolfgang Schweiger (University of Natural Resources and Life Sciences, Tulln) for valuable feedback during implementing the tool.
Funding: We acknowledge financial support of the work by the FWF (SFB Fusarium) and the Deutsche Forschungsgemeinschaft (DFG)(SFB 924).

REFERENCES

Altschul, S. F. *et al.* (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, **17**, 3389–3402.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*, **11**(10), R106.

Brenchley, R. *et al.* (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**(7426), 705–710.

Gene Ontology Consortium (2004). The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, **32**(suppl 1), D258–D261.

Goff, L., Trapnell, C., and Kelley, D. (2012). *cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data.* R package version 2.0.0.

Lohse, M. *et al.* (2012). Robina: A user-friendly, integrated software solution for rna-seq-based transcriptomics. *Nucleic Acids Research*.

Nussbaumer, T. *et al.* (2013). Mips plantsdb: a database framework for comparative plant genome research. *Nucleic Acids Research*, **41**, 1144–51.

Robinson, M. D. *et al.* (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

The International Barley Sequencing Consortium (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**(7426), 711–6.

Trapnell, C. *et al.* (2009). Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C. *et al.* (2013). Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature Biotechnology*, **28**, 46–53.

Wang, T. *et al.* (2013). STAR: an integrated solution to management and visualization of sequencing data. *Bioinformatics*, **29**(24), 3204–3210.