tmod: an R package for general and multivariate enrichment analysis

January Weiner 3rd[1] **and Teresa Domaszewska**[1]

[1]**Max Planck Institute for Infection Biology, Chariteplatz 1, 10117 Berlin, january@mpiib-berlin.mpg.de**

ABSTRACT

"Omics" studies generate long lists of genes, proteins, metabolites or other features which can be difficult to decipher. Feature set enrichment analysis utilizing annotated groups/classes of features (such as pathways, gene ontology terms or gene/metabolic modules) can provide a powerful gateway to associate data to phenotypes such as disease process or treatment progression. At the same time, the increasing use of technologies to generate multidimensional omics data sets based on specific cell types or responses to stimuli increases the number and breadth of annotated feature sets available for enrichment analysis, facilitating the ability to draw biologically relevant conclusions. However, existing tools and applications for enrichment analysis are adapted specifically to gene set enrichment and lack functionalities to analyze rapidly growing amounts of metabolomics and other data. Moreover, such tools often provide only a limited range of statistical methods, rely on permutation tests, lack suitable visualization tools to facilitate result interpretation in complex experimental setups, and lack standalone versions usable in semi-automatized workflows. Here, we present tmod, an R package which implements powerful statistical methods for enrichment analysis. Tmod includes definitions of widely used feature sets for transcriptomic and metabolomic profiling and also allows use of custom user-provided feature sets. Moreover, it provides novel and intuitive visualization methods which facilitate interpretation of complex data sets. The implemented statistical tests allow the significance of enrichment within sorted feature lists to be calculated without randomization tests and thus are suitable for combining functional analysis with multivariate techniques.

Keywords: Gene set enrichment analysis,transcription,metabolomics,metabolic profiling

INTRODUCTION

To begin to decipher the complex spatial, temporal and environmentally induced regulation and interaction of genes, proteins, metabolites within a cell or organism, biologists continue to annotate groups of molecules which are co-regulated, interact or have similar functions.
For example, there have been efforts to characterize functional associations between genes or metabolites on the basis of common patterns of their regulation in response to environmental challenges (Berry et al., 2010; Pascual, Chaussabel & Banchereau, 2010). These annotated "feature sets" equip biologists with a powerful method for predicting programs activated by an organism in response to a given stimulus, whereby novel data sets can be elucidated by detecting enrichments of previously described feature sets. In the context of infectious diseases and immune responses, identifying transcriptional signatures or metabolic profiles that allow discrimination between infected and healthy individuals helps to monitor disease progression and response to treatment and can reveal defense mechanisms activated by the organism.

In order to generate feature sets based on gene co-expression, several studies have attempted to determine meaningful relationships between genes based on their regulation; by measuring changes in gene expression under various environmental conditions in distinct organisms they have provided models for the discovery of transcriptional modules (Bar-Joseph et al., 2003; Liu et al., 2007), resulting in a broad range of gene classification collections. Chaussabel et al. (Chaussabel et al., 2008) developed a strategy for blood microarray data analysis which identified genes co-expressed across multiple disease datasets and classified them into 28 blood transcriptional modules. The genes belonging to such functionally related clusters can be used to generate a disease-specific transcription signature which may serve for diagnostics or treatment prognosis (Berry et al., 2010). In another approach, 334 blood transcriptional modules (BTMs) were annotated according to biological function or tissue-specific expression by Li et al. (Li et al., 2014). Transcriptional modules were defined as groups of highly connected genes belonging to context-specific sub networks. BTMs have proved successful in immunological applications, e.g. for autoimmune diseases (Pascual, Chaussabel &

Banchereau, 2010) or predicting response to pyogenic bacteria in patients carrying mutations in pathways responsible for pathogen sensing (Alsina et al., 2014). Other widely used gene collections useful for interpretation of transcriptome data sets include annotated gene sets used with GSEA (Gene Set Enrichment Analysis) software available in the MSigDB database (Subramanian et al., 2005), gene ontology (GO) annotations, and pathway annotations such as KEGG pathways.

While transcriptomic studies focus on mRNA levels being produced in a tissue, metabolic profiling focuses on metabolites that are the product of cellular processes providing an alternative snapshot of the tissue's state and condition. Whereas mRNAs cannot move outwith the cells in which they are expressed, metabolites can be produced in one cell and subsequently be detected in other tissues of the organism; for example metabolites produced at the site of infection (e.g. in the lung) can be detected in blood. During infectious disease, the metabolic profile can change directly with pathogen entrance as multiple host metabolic pathways are being altered. This, as well as the pathogen's own specific metabolites can already be detected before a pathology develops which carries a promise for early detection of complex and slow progressing diseases, like tuberculosis (Weiner 3rd et al., 2012). Metabolic profiling in the context of diagnostics and disease or treatment outcome prediction is currently being broadly implemented thanks to advanced and sensitive methods for detecting and classifying metabolites. Given a wide scope of meaningful biological annotation and manual curation, custom-created gene or metabolite sets can serve to i) functionally annotate observed changes in gene or metabolite regulation, ii) pre-select transcripts or metabolites playing significant roles in a disease, and as iii) a diagnostic tool for detection of disease (Berry et al., 2010), its progression, possible treatment outcomes or best vaccination type.

For example, in patients suffering from systemic lupus erythematosus, follow-up transcriptional studies based on blood transcriptional modules (BTMs) predicted the severity of disease more accurately than the currently used SLEDAIC score ((Pascual, Chaussabel & Banchereau, 2010)), and in tuberculosis patients anti-inflammatory, immunosuppressive and stress responses have been revealed by changes in amino acid, lipid and nucleotide metabolism (Weiner 3rd et al., 2012).

The output from initial analysis of omics data is generally a long list of features (e.g. genes or metabolites) with their associated fold-changes and p-values. Annotated feature classifications can be applied to these data sets using feature set enrichment analysis (FSEA). Two main approaches for FSEA exist. Firstly, a commonly used method divides genes or metabolites (later referred to as 'features') into two sets: "foreground" with differentially regulated features and "background" with all others, and than applies a hypergeometric test to test for enrichment. This approach has been implemented in numerous packages for enrichment analysis (for example in the GOstats R package (Falcon & Gentleman, 2007)). However, it requires setting of an arbitrary threshold based on p-value and fold change. Changing these arbitrary thresholds can result in different enrichments. Moreover, selecting the cutoffs is indeed very arbitrary since p-values depend on the sample size and fold changes can be dependent on the platform used. This can have dramatic effects on the enrichment statistics depending on the sample size used. A similar problem occurs if the enrichment statistics is calculated based on statistics from differential gene analysis, for example by combining the p-values obtained for genes using Fisher's method or other related approaches (for example, as implemented in the Piano package, (Vremo, Nielsen & Nookaew, 2013)).

Alternatively, initial list of features can be ordered and enrichment occuring towards the top of the list can be detected using statistical methods. Statistics may be derived by ranking features by their changes between experimental conditions or by the correlation with experimental groups. This approach has been implemented in the widely-used GSEA analysis of MSigDB collections using randomization tests to obtain p-values (Subramanian et al., 2005). Randomization tests are commonly used as they can be applied to any statistics but they require sufficient sample size to work effectively. GSEA uses sample-wise randomization tests for large samples, replaced by gene-wise randomizations for small samples. This results in poor performance for small sample sizes, and cannot be applied in combination with multivariate processing. For integration in a high-throughput setting, GSEA requires high memory and CPU requirements due to the high load required by randomizations. Moreover, GSEA does not allow an easy integration with differential expression analysis, for example from the R package Limma (Ritchie et al., 2015), but in itself does not have the full flexibility of Limma's linear models.

In order to overcome some of the pitfalls of GSEA and other approaches, we set out to design a FSEA solution with the following features:

- statistical test for enrichment in ordered lists of features (thus requiring no arbitrary cutoffs),
- based on an analytical solution rather than permutation or randomization tests, with acceptable sensitivity and specificity, suitable for integration with multivariate approaches;

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2420v1 | CC BY 4.0 Open Access | rec: 4 Sep 2016, publ: 4 Sep 2016

- straight-forward integration of user-defined sets in the analysis, especially of BTM's and metabolic profiling feature sets;
- testing independent of the method used for differential analysis (in order to combine with methods such as limma or edgeR);
- implementation which allows semi-automatized processing (preferably as in the R language);
- integration with multivariate methods (e.g. using FSEA to interpret principal components);
- visualization strategy giving an overview over several enrichment analyses (e.g. for tracking changes of enrichment in a time series analysis).

We have integrated the above concepts in a novel package, tmod, an open, standalone framework implemented in R, containing set definitions for transcriptional modules from both Li et al. and Chaussabel et al. (Chaussabel et al., 2008; Li et al., 2014) and metabolite sets based on metabolic profiles published by Weiner et al. (Weiner 3rd et al., 2012).

Importantly, tmod provides a choice of statistical methods to assign the significance of enrichment analysis. Aside from an implementation of hypergeometric tests, tmod includes Mann-Whitney non-parametric U test and the highly sensitive CERNO test (Yamaguchi et al., 2008), which is a new application of Fisher's method for ranked list in FSEA. Unlike the Fisher's method implementation in packages such as Piano, here feature ranks, scaled by the number of features, are treated as probabilities and combined with Fisher's method to calculate the statistics.

Both U and CERNO tests can be performed on ranked lists of features, eliminating the need to apply arbitrary log-fold change or p-value cutoffs. The included feature set definitions allow the use of any other test for the analysis. Moreover, tmod can be used to combine multivariate analyses of the datasets with functional enrichment analysis. Calculation results are returned in a form of clear, interpretable visualizations, which are calibrated to illustrate multiple-group comparisons, changes in regulation over time and interactions. Finally, tmod allows the user to directly analyse and visualize results returned by functions from the Limma package (Ritchie et al., 2015). Our package has already proven to be a valuable tool in the analysis of heterologous and multivariate data sets (Esterhuyse et al., 2015).

RESULTS

Tmod functionality

The tmod package includes three basic functions to test the enrichment of metabolite or gene sets. The tmodHGtest function performs a hypergeometric test on two groups of features, foreground and background, defined by the user on the basis of differential regulation analysis. tmodUtest operates on a ranked set of features and tests the significance of the area under the curve (AUC) statistics by performing a nonparametric Mann-Whitney test on groups of features belonging, as well as not belonging to a set. This approach allows analysis of enrichment independent of arbitrary choice of threshold p- or logarithm of fold change (logFC) value for defining the set of regulated features. Furthermore, we have implemented the CERNO test based on Fisher's method (Fisher's combined probability test) in the tmodCERNOtest function, which is the first implementation of this powerful statistical method (Yamaguchi et al., 2008). Both U and CERNO are statistics with known distributions and therefore do not require a randomization approach to calculate the p-values. However, randomization tests can also be conveniently used with tmod if required, as described in the package vignette. The functions return data frame objects of variables enriched in subsequent modules together with their p-values, as well as test-specific statistics (the number of features in the foreground, background and the total number of features from the analyzed dataset belonging to listed modules for tmodHGtest, calculated U statistics for tmodUtest and statistic for tmodCERNOtest).

CERNO statistic: a powerful approach to FSEA

In the tmodCERNOtest, we have implemented and tested a modification of Fisher's method of combining probabilities to FSEA, described by (Yamaguchi et al., 2008). The method uses scaled ranks of features and combines them in a statistic which is directly used to compute the p-values. Compared to a U-test, this method weights low-ranking features more than intermediate, resulting in higher p-values for feature sets which have a low effect size, but a large number of features. Naturally, it does not require a randomization approach. Using a randomization approach, we have tested the specificity and sensitivity of this method compared to the GSEA and found that its performance is similar, but faster by several orders of magnitude in terms of computation time, for large sample sizes; for small sample sizes we see a marked improvement in performance over GSEA. Importantly, the results from tmod/CERNO were more robust and less dependent on the sample size (Figure 1).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2420v1 | CC BY 4.0 Open Access | rec: 4 Sep 2016, publ: 4 Sep 2016

The tmod/CERNO results for small samples derived from a large pool of samples were simlar to those derived directly from a large sample (Spearman's > 0.9 for sample size of 5), whereas the results for GSEA for small sample sizes were variable (Spearman's rho < 0.1 for sample sizes of five or less).

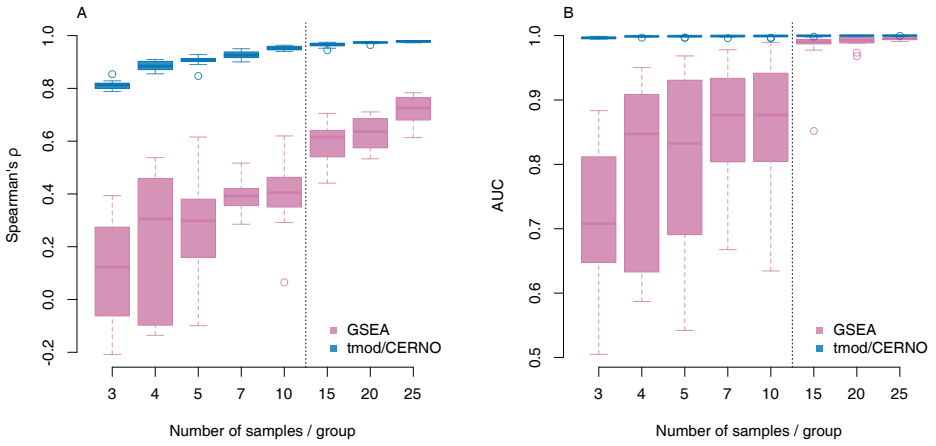

Figure 1. Performance of tmod/CERNO compared to GSEA. Boxplots summarize the results of 10 random replicates. Dotted vertical line denote the difference in the permutation method for GSEA for small sample sizes. **A**, Spearman's coefficient between the results derived for the small subsample compared with results generated for the full sample set; **B** AUC for a given sample calculation relative to a set of "true" positives and negatives defined by conservative threshold and detected by both methods on the full sample set.

For reporting the effect size in enrichment, we have decided to use area under curve (AUC), as it has a straightforward interpretation and visualization, and is similar to the U-statistic.

Visualization of results

The visualization of results includes the receiver-operator characteristic (ROC) curve allowing detailed investigation of enrichment of specific sets. The ROC curve is a graphical method which assesses significance of enrichment visualizing it on a simple and straightforward plot (Figure 2).

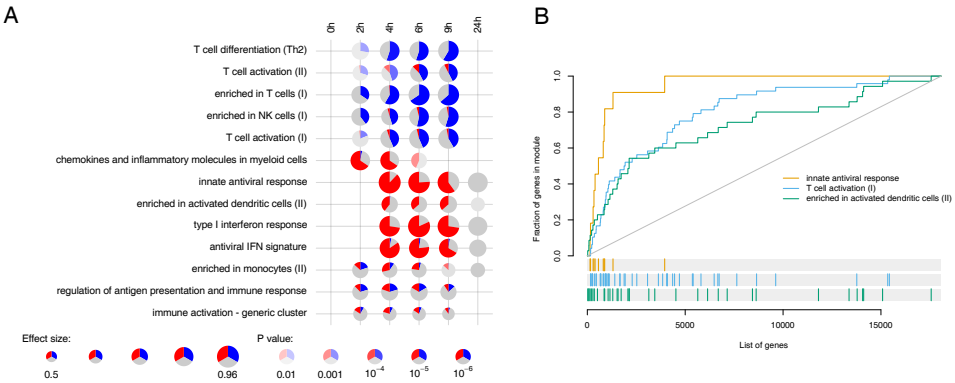

Figure 2. Serial FSEA using limma and tmod. Whole blood samples from patients after endotoxin injection have been analysed at 5 different time points. **A**, enriched gene modules at every time point. Red represents up-regulated and blue down-regulated genes. **B**, ROC curves for three selected modules significantly enriched at time point 4h.

The data objects included in the tmod package can also be used to analyze enrichment with one of the many other methods published, including geneSetTest and CAMERA from the limma package

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2420v1 | CC BY 4.0 Open Access | rec: 4 Sep 2016, publ: 4 Sep 2016

or maxmean method from the GSA package (Smyth, 2005; Efron & Tibshirani, 2007; Wu & Smyth, 2012).

Single genes or metabolites enriched in a given module can be visualized using the evidencePlot function. The output is a ROC curve, where the area under the curve corresponds to the Mann-Whitney test statistics reported by tmodUtest (Figure 2, B). Another visualization method is the panel plot, which allows the depiction of module enrichment on the span of measured time points and conditions (Figure 2, A). For each module or variable set, it shows its enrichment in each of the performed analyses in terms of effect size and p-value, simultaniously showing numbers of up- or down-regulated features.

Modules and Gene sets

The tmod package provides gene module definitions based on HGNC (HUGO Gene Nomenclature Committee) and metabolite set definitions based on HMDB (Wishart et al., 2012) identifiers. The classification of 260 BTMs created by Chaussabel et al. (Chaussabel et al., 2008) was downloaded from the online resource and is based on studies of 9 datasets from 410 patients suffering from 9 different diseases. The classification of 334 BTMs created by Li et al. was downloaded from the supplementary information to the original publication (Li et al., 2014) and is based on 30000 transcriptome profiles of human blood samples from over 500 publicly available studies. The metabolome modules were defined by Weiner et al. (Weiner 3rd et al., 2012). Objects used for data storage in the package include two data frames (MODULES and GENES) containing original module and general gene information, and two lists (MODULES2GENES and GENES2MODULES) containing module mapping to genes and gene mapping to modules, respectively.

Functionality of tmod includes integration of other feature sets for the enrichment analyses, notably the MSigDB collections. Once an MSigDB collection file is downloaded, tmod integrates it into a compatible R object with tmodImportMSigDB function. The object is then ready for enrichment testing with tmod.

Manual definition of feature sets

Although multiple collections of annotated modules are available online, broad scope of biological contexts in which enrichment can be studied may favor another approach than the annotated sets. For this purpose tmod provides tools to manually create desired module sets. Accordingly, any signaling pathway, interaction, disease association or other collection from external or user-supported source can be tested with tmod. The exact guide for module creation is described in detail in the vignette for the package and contains an example of implementing the WikiPathways (Kelder et al., 2012) pathway.

Functional multivariate analyses

Multivariate transformations are broadly implemented in analyses of multidimensional data. Principal component analysis (PCA) or Independent Component Analysis (ICA) highlight main factors influencing variability in gene and metabolite regulation. Feature set annotation can help to understand the biological meaning of calculated multivariate transformations. Using tmod, components can be tested for how well they correspond to specific feature sets taking advantage of the fact that variables which influence the position of a sample along a given component have a larger absolute weight for that component. One of the applications of tmod that proves useful for PCA interpretation is to sort variables by their weight in a given component, and use the tmodUtest or tmodCERNOtest function to test for the enrichment of modules. However, tmod offers also a robust tmodPCA function which automatically analyzes PCA results and visualizes the result on an annotated plot (Figure 3). For example, using data from (Weiner 3rd et al., 2012) we applied tmodPCA, and, using this streamlined approach, confirmed the findings originally obtained using manual evaluation of hundreds of statistical tests.

Serial FSEA with limma

To facilitate analysis of complex experimental set ups and of results of multivariate analyses such as PCA, tmod includes several functions for serial analysis and visualization of set enrichments. For transcriptional data, the functions tmodLimmaTest and tmodLimmaDecideTests allows the user to quickly analyze the enrichment in all coefficients included in a limma differential analysis (Smyth, 2005). The functions tmodSummary and tmodPanelPlot allow the creation of, respectively, a tabular or a visual summary of the results of enrichment analysis respectively (Figures 2 and 3). The output of tmodPanelPlot function is tailored to visualize changes in transcriptomic/metabolomic data over time and in complex comparisons and allows applying custom p-value or effect size thresholds.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2420v1 | CC BY 4.0 Open Access | rec: 4 Sep 2016, publ: 4 Sep 2016

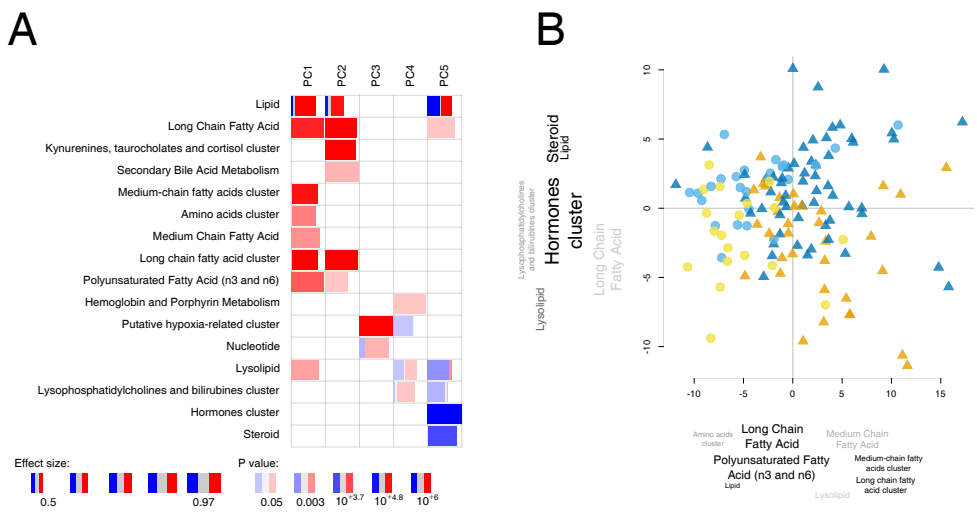

Figure 3. Multivariate functional analysis of serum metabolic profiles of TB patients and healthy controls (data from (Weiner 3rd et al., 2012)). **A**, tmod panel plot showing FSEA for the first 5 components in a PCA analysis of the data set. Width of the squares corresponds to the effect size (AUC), color corresponds to sign of the PCA score and shading reflects the p-value as shown in the legend. Components 1 and 2 correspond to the differences between TB patients and healthy individuals, showing changes in lipid metabolism, amino acids and a cluster including cortisol and kynurenine. **B**, PCA plot augmented by FSEA enrichment tag clouds. Components 1 (corresponding to differences between TB patients and healthy controls) and 5 (corresponding to the differences between males and females) were selected for visualization. Blue, females; yellow, males; circles, healthy; triangles, TB.

Usage example

As an illustration of the package's featured attributes, we present an example of functional interpretation of PCA using tmodCERNOtest together with resulting plots generated with the tagcloud package [18] (Figure 3). The data has been published by Weiner et al. (Weiner 3rd et al., 2012) and was generated from samples taken from tuberculosis (TB) patients and healthy controls. Here, for each of the first five components, metabolites were ordered by their absolute loadings and tested for enrichment using tmod/CERNO (Figure 3, A). While the first three components correspond to the previously described differences between TB patients and healthy controls (notably including a cluster of amino acids in component 1 and kynurenine/cortisol cluster in component 3), component 5 corresponds to gender differences. This has been visualized on a plot of components 1 and 5 (Figure 3, B), where the plot axes have been annotated by tag clouds corresponding to the results of enrichment tests. In tmod, this plot is generated with the command tmodPCA.

MATERIALS AND METHODS

We have developed and tested tmod using R version 3.2, as well as the development versions of R. The tmod package is available from CRAN (http://cran.r-project.org/web/packages/tmod/) and from the projects web site http://bioinfo.mpiib-berlin.mpg.de/tmod/.

For comparison with GSEA, we have used a large data set described by Kaforou et al. (2013) derived from tuberculosis patients and healthy controls. We tested the performance of both methods for small sample sizes (3-25 per group) compared to the full data set (over 50 samples / group), using the full MSigDB (v. 5.0). First, we tested the Spearman correlation between the p-values obtained for a given method for each of the small sample sizes compared with the full sample. Next, for the full sample set, we have defined a list of bona fide true positives and negatives by choosing gene sets which either had a q-value < 0.01 in both methods ("positive") or > 0.2 in both methods ("negative"). We then calculated the area under the receiver-operator curve (AUC) for each sample size, as compared to the true positives / negatives in the full sample set.

Differential expression of transcriptome data (Figure 2) was calculated using limma for gene expression changes between controls and individuals at time points 2, 4, 6, 9 and 24 hours after endotoxin injection for study published by Calvano et al. (2005). Gene set enrichment analysis was

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2420v1 | CC BY 4.0 Open Access | rec: 4 Sep 2016, publ: 4 Sep 2016

performed using tmod and the module set annotated by Li et al. and visualized with tmodPanelPlot. ROC curves for three of the enriched modules in time point 4h have been created using evidencePlot function.

To demonstrate multivariate functional analyses, we have used the data set published by Weiner et al. (2012), also provided with the tmod R package. To generate the plots, tmod functions tmodPanelPlot and tmodPCA were used.

DISCUSSION

We have created the R package tmod for enrichment analysis integrating transcriptomic and metabolomic data and implemented a simple web interface which can be accessed at http://bioinfo.mpiib-berlin.mpg.de/tmod/. The interface allows choice of set definitions (Chaussabel et al., 2008; Li et al., 2014) or gene collections from MSigDB and different types of statistical tests provided by tmod package, which enables quick comparison of those approaches, which are not available in any previously published tools.

The R package for tmod is available online on CRAN and equipped with a vignette presenting illustrative practical applications, including step by step instructions for enrichment analysis in an built-in transcriptional dataset (Maertzdorf et al., 2011), use of other sets of modules, performing functional multivariate analysis, visualization of results and application and creation of custom sets of modules. As an illustration of the package's featured attributes, we presented an example of principal component analysis result interpretation with use of tmodUtest function.

Built-in BTM and metabolome sets are integrated as tmod objects. In contrary to previously available standalone tools for overrepresentation analyses (many of which are not directly compatible for use with R) (Subramanian et al., 2005), tmod does not demand manual upload of set annotation. At the same time, application of functions allowing easy and straightforward integration of other gene or metabolite sets broadens the scope of biological analyses supported by this single package, which otherwise would demand use of GSEA (Subramanian et al., 2005) software or manual transformations of each set derived from different source. In comparison to other tools, tmod first allows manual creation of modules, which can be of extreme importance for scientists studying transcriptome or metabolome regulation in very specific and not broadly explored biological contexts. Implementation of three statistical tests called by single functions makes statistical testing easily accessible for biologists, and different test statistics shown imply easier interpretability of calculated gene enrichment. In contrast to the statistic used in the GSEA approach (Subramanian et al., 2005), the U and CERNO test are statistics with known distributions and therefore do not require a randomization approach to calculate the p-values. This is advantageous in case of a low number of samples and in other settings such as functional interpretation of principal components. We show that the CERNO function within tmod outperforms GSEA in terms of speed for data sets with large sample sizes, and in overall performance for datasets with less than 15 samples.

Finally, an otherwise rather complex PCA functional interpretation is contained in a single function, tmodPCA. No other online or standalone tools to our knowledge possess such a broad functionality, which makes tmod especially useful in application for biologists or immunologists investigating complex events such as disease diagnosis, progression or treatment, which require simple tools giving robust and biologically significant results. The package contains efficient and intuitive data visualization tools.

CONCLUSION

We have introduced tmod, a flexible R package for standalone analysis of enrichment analysis of transcriptomic and metabolomics data. Considering widespread use of R by scientists, the package provides a useful tool that otherwise has not yet been accessible for R users, containing useful application of three statistical tests assessing significance of the enrichment, PCA analysis and data visualization tools.

List of abbreviations

AUC – area under curve; BTM – blood transcriptional module; FSEA – Feature Set Enrichment Analysis; GSEA – Gene Set Enrichment Analysis; HGNC - HUGO Gene Nomenclature Committee; ICA – independent component analysis; logFC - logarithm of fold change; PCA – principal component analysis; ROC – receiver-operator characteristic.

Competing interests

Authors declare no conflicts of interests.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2420v1 | CC BY 4.0 Open Access | rec: 4 Sep 2016, publ: 4 Sep 2016

Acknowledgements
We would like to thank Gayle McEwen for invaluable help with the manuscript and many important insights, and Jeroen Maertzdorf and Stefan H.E. Kaufmann for helpful discussions.

REFERENCES

Alsina L., Israelsson E., Altman MC., Dang KK., Ghandil P., Israel L., Bernuth H von., Baldwin N., Qin H., Jin Z., others. 2014. A narrow repertoire of transcriptional modules responsive to pyogenic bacteria is impaired in patients carrying loss-of-function mutations in myd88 or irak4. *Nature immunology* 15:1134–1142.

Bar-Joseph Z., Gerber GK., Lee TI., Rinaldi NJ., Yoo JY., Robert F., Gordon DB., Fraenkel E., Jaakkola TS., Young RA., others. 2003. Computational discovery of gene modules and regulatory networks. *Nature biotechnology* 21:1337–1342.

Berry MP., Graham CM., McNab FW., Xu Z., Bloch SA., Oni T., Wilkinson KA., Banchereau R., Skinner J., Wilkinson RJ., others. 2010. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 466:973–977.

Calvano SE., Xiao W., Richards DR., Felciano RM., Baker HV., Cho RJ., Chen RO., Brownstein BH., Cobb JP., Tschoeke SK., others. 2005. A network-based analysis of systemic inflammation in humans. *Nature* 437:1032–1037.

Chaussabel D., Quinn C., Shen J., Patel P., Glaser C., Baldwin N., Stichweh D., Blankenship D., Li L., Munagala I., others. 2008. A modular analysis framework for blood genomics studies: Application to systemic lupus erythematosus. *Immunity* 29:150–164.

Efron B., Tibshirani R. 2007. On testing the significance of sets of genes. *The annals of applied statistics*:107–129.

Esterhuyse MM., Weiner J., Caron E., Loxton AG., Iannaccone M., Wagman C., Saikali P., Stanley K., Wolski WE., Mollenkopf H-J., others. 2015. Epigenetics and proteomics join transcriptomics in the quest for tuberculosis biomarkers. *mBio* 6:e01187–15.

Falcon S., Gentleman R. 2007. Using gostats to test gene lists for go term association. *Bioinformatics* 23:257–258.

Kaforou M., Wright VJ., Oni T., French N., Anderson ST., Bangani N., Banwell CM., Brent AJ., Crampin AC., Dockrell HM., others. 2013. Detection of tuberculosis in hiv-infected and-uninfected african adults using whole blood rna expression signatures: A case-control study. *PLoS Med* 10:e1001538.

Kelder T., Iersel MP van., Hanspers K., Kutmon M., Conklin BR., Evelo CT., Pico AR. 2012. WikiPathways: Building research communities on biological pathways. *Nucleic acids research* 40:D1301–D1307.

Li S., Rouphael N., Duraisingham S., Romero-Steiner S., Presnell S., Davis C., Schmidt DS., Johnson SE., Milton A., Rajam G., others. 2014. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature immunology* 15:195–204.

Liu X., Jessen WJ., Sivaganesan S., Aronow BJ., Medvedovic M. 2007. Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and chip-chip data. *BMC bioinformatics* 8:283.

Maertzdorf J., Ota M., Repsilber D., Mollenkopf HJ., Weiner J., Hill PC., Kaufmann SH. 2011. Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. *PloS one* 6:e26938.

Pascual V., Chaussabel D., Banchereau J. 2010. A genomic approach to human autoimmune diseases. *Annual review of immunology* 28:535.

Ritchie ME., Phipson B., Wu D., Hu Y., Law CW., Shi W., Smyth GK. 2015. Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*:gkv007.

Smyth GK. 2005. Limma: Linear models for microarray data. In: *Bioinformatics and computational biology solutions using r and bioconductor*. Springer, 397–420.

Subramanian A., Tamayo P., Mootha VK., Mukherjee S., Ebert BL., Gillette MA., Paulovich A., Pomeroy SL., Golub TR., Lander ES., others. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102:15545–15550.

Vremo L., Nielsen J., Nookaew I. 2013. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic acids research*:gkt111.

Weiner 3rd J., Parida SK., Maertzdorf J., Black GF., Repsilber D., Telaar A., Mohney RP., Arndt-

Sullivan C., Ganoza CA., Fa KC., others. 2012. Biomarkers of inflammation, immunosuppression and stress with active disease are revealed by metabolomic profiling of tuberculosis patients. *PloS one* 7:e40221.

Wishart DS., Jewison T., Guo AC., Wilson M., Knox C., Liu Y., Djoumbou Y., Mandal R., Aziat F., Dong E., others. 2012. HMDB 3.0—the human metabolome database in 2013. *Nucleic acids research*:gks1065.

Wu D., Smyth GK. 2012. Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic acids research* 40:e133–e133.

Yamaguchi KD., Ruderman DL., Croze E., Wagner TC., Velichko S., Reder AT., Salamon H. 2008. IFN-β-regulated genes show abnormal expression in therapy-nave relapsing–remitting ms mononuclear cells: Gene expression analysis employing all reported protein–protein interactions. *Journal of neuroimmunology* 195:116–120.