Bioinformatics, 32(12), 2016, 1891–1894
doi: 10.1093/bioinformatics/btw030
Advance Access Publication Date: 1 February 2016
Applications Note

Gene expression

GiANT: gene set uncertainty in enrichment analysis

Florian Schmid[1], Matthias Schmid[2], Christoph Müssel[1], J. Eric Sträng[1], Christian Buske[3], Lars Bullinger[4], Johann M. Kraus[1,†] and Hans A. Kestler[1,5,*,†]

[1]Institute of Medical Systems Biology, Ulm University, Ulm 89069, Germany, [2]Institut für Medizinische Biometrie, Informatik und Epidemiologie, Universität Bonn, Bonn 53127, Germany, [3]Institute of Experimental Cancer Research, [4]Department of Internal Medicine III, Ulm University, Ulm 89069, Germany and [5]Leibniz Institute on Ageing – Fritz Lipmann Institute and FSU Jena, Jena 07745, Germany

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Jonathan Wren

Received on 14 August 2015; revised on 28 December 2015; accepted on 12 January 2016

Abstract

Summary: Over the past years growing knowledge about biological processes and pathways revealed complex interaction networks involving many genes. In order to understand these networks, analysis of differential expression has continuously moved from single genes towards the study of gene sets. Various approaches for the assessment of gene sets have been developed in the context of gene set analysis (GSA). These approaches are bridging the gap between raw measurements and semantically meaningful terms.

We present a novel approach for assessing uncertainty in the definition of gene sets. This is an essential step when new gene sets are constructed from domain knowledge or given gene sets are suspected to be affected by uncertainty. Quantification of uncertainty is implemented in the R-package GiANT. We also included widely used GSA methods, embedded in a generic framework that can readily be extended by custom methods. The package provides an easy to use front end and allows for fast parallelization.

Availability and implementation: The package GiANT is available on CRAN.

Contacts: hans.kestler@leibniz-fli.de or hans.kestler@uni-ulm.de

1 Introduction

Differential expression analysis investigates the association of measurements of single genes to a predefined phenotype (e.g. tumour versus inflammation). Frequently yielding thousands of differentially expressed genes, such analyses are often hard to interpret in a biological context (Zeeberg *et al.*, 2003). Gene set analyses aim at assigning a meaning to differentially expressed genes by comparing them to sets of genes whose relevance in processes or pathways is known. Such gene sets can be derived, e.g. from Gene Ontology (Gene Ontology Consortium, 2000), KEGG (Kanehisa and Goto, 2000), AgeFactDb (Hühne *et al.*, 2014), Reactome (Joshi-Tope *et al.*, 2005), WikiPathways (Pico *et al.*, 2008) or other collections of gene sets (Glez-Peña *et al.*, 2009; Huang *et al.*, 2007; Subramanian *et al.*, 2005). Albeit frequently used, gene set analysis suffers from pitfalls concerning the significance assessment (Goeman and Bühlmann, 2007; Maciejewski, 2013). Also the large variety of methods makes the choice of which analysis to use often difficult. Beside that, issues arise from the crafting of the gene sets. For example, rapidly changing knowledge may not only affect the members of sets, but also the mapping of probe sets to gene identifiers (Bleazard *et al.*, 2015; Retraction for Dixson *et al.*, 2014; Sedeño-Cortés and Pavlidis, 2014) and the validity of included genes under a certain condition

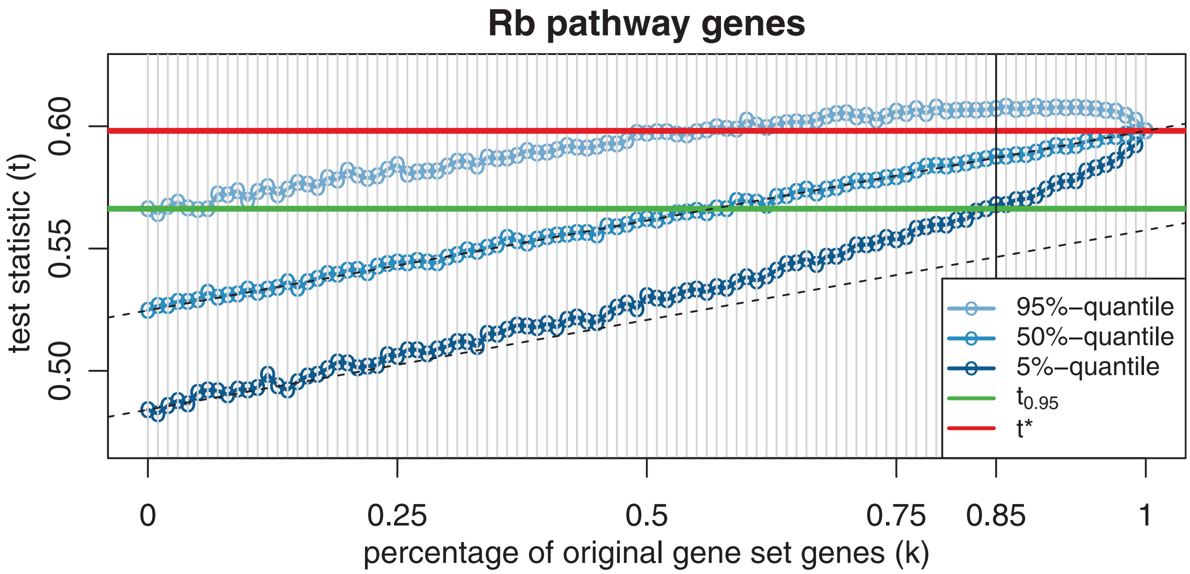

Fig. 1. Quantification of the uncertainty in the Rb pathway. The lines with circles show the tested degrees of uncertainty (*k*) for the Rb pathway. The dots in each column give the quantiles ({0.05, 0.5, 0.95}) of the test statistic values obtained by resampling a percentage of genes from the Rb pathway (*k*) and the remaining genes ($1-k$) from the set of all genes in the dataset. As a test statistic the mean absolute correlation (Spearman) of the gene set genes to the class-label has been used. The values for $k=0$ give the quantiles of the null distribution, with the lower horizontal line corresponding to the 95% quantile. The upper horizontal line shows the value of the test-statistic for the original Rb pathway. The estimated uncertainty of the gene set is the minimum value of *k* that has a non overlapping credibility interval with the credibility interval of the null distribution. For the Rb pathway this value is 85%. An estimate of the upper bound of certainty can be given by calculating the slope of the dotted line by using the median of null distribution and t^*. When shifted to the 95%-quantile the point of intersection with $t_{0.95}$ is the upper bound (Color version of this figure is available at *Bioinformatics* online.)

(e.g. if a gene set is only enriched in the data because of genes that have been wrongly assigned to the gene set).

To address these issues of generating gene sets, we present a novel method to quantify the uncertainty in gene set analyses. This method assesses the impact of changes in the gene set definition on the result of a gene set analysis. We apply a bootstrap-type resampling strategy in which parts of the original gene set are replaced by randomly choosen genes. By analyzing the derived credibility intervals, an estimate of the certainty in the definition of the gene set can be derived. Such robustness assessments are essential for the validation of custom hand-crafted gene sets and are also of interest for sets extracted from the knowledge bases mentioned above.

2 Method and application

2.1 Method

In a robust gene set, slight changes in the definition of the set should not have strong effects on its statistical significance. We therefore implemented a robustness evaluation for gene set analysis that rates the certainty in the definition of a hand-crafted gene set. This evaluation is based on repeated gene set analyses with slightly modified versions of the set in order to measure how strongly uncertainty in the gene set affects statistical significance.

To assess robustness, a large number of perturbed gene sets $\mathrm{GS}_{\mathrm{pert}}$ are generated from the original gene set *GS*, keeping a percentage *k* of the genes $x \in GS$ and replacing the remaining percentage $(1-k)$ by randomly chosen genes $x' \in DS \backslash GS$, where *DS* denotes the whole set of available genes:

$$GS_{\mathrm{pert}}(k) = \cup(\{x_i | x_i \in GS, i \in \{1, \ldots, k \cdot |GS|\}\},$$

$$\{x'_i | x_i' \in DS \backslash GS, i \in \{1, \ldots, (1-k) \cdot |GS|\}\}).$$

To evaluate the fuzziness of *GS*, we use the following three-step approach: In the first step, the test statistic of interest (denoted by *t*) is computed for each of the samples $GS_{\mathrm{pert}}(k)$, resulting in estimates of the distribution of *t* at various values of *k*. Setting $k=0$, i.e. including none of the original gene set genes, results in the 'null distribution' of *t*. In the second step, 90% credibility intervals for *t* (denoted by $\mathrm{CI}_{0.9}(k)$) are constructed for each *k* by evaluating the 5 and 95% quantiles of the estimated distributions. In the final step, the minimum value of *k* for which the credibility interval $\mathrm{CI}_{0.9}(k)$ does not overlap with the null interval $\mathrm{CI}_{0.9}(0)$ is calculated. We use this value, which we denote by k^*, as an estimate of the certainty in the definition of GS, as large values of k^* indicate a high sensitivity of *t* with regard to the random replacement of genes in GS and conversely low values of k^* indicate a high robustness (see Fig. 1).

For the definition of the gene set uncertainty proposed above we can formally derive a lower bound. Assume the null distribution with mean μ and standard deviation σ and a second distribution with a fixed percentage of gene set genes in each sample, mean μ' and standard deviation σ'. Then $\kappa(\sigma') \leq \kappa(\sigma)$ with κ giving the value of the test statistic corresponding to a quantile of the underlying distribution. From that it can be seen that $\kappa(\sigma)$ gives a lower bound for the uncertainty of a gene set:

$$\mu + \kappa(\sigma) = \mu' + \kappa(\sigma') \Rightarrow \mu - \mu' \leq 2\kappa(\sigma). \tag{1}$$

We implemented the R-package GiANT for analyzing the uncertainty in the definition of gene sets. Following Ackermann and Strimmer (2009), the package also includes a toolbox for gene set analysis that is modularized into four steps. The first step is the analysis of differential expression. The resulting gene-level statistic may be transformed for subsequent steps and is then summarized for a specified gene set, i.e. a gene set statistic is calculated. The significance of this statistic with respect to certain null hypotheses is finally assessed in a resampling-based testing procedure. Importantly, the `GiANT` package is based on a highly generic framework that allows users to create custom analyses by replacing some or all steps by their own implementations. It also supports parallelization of

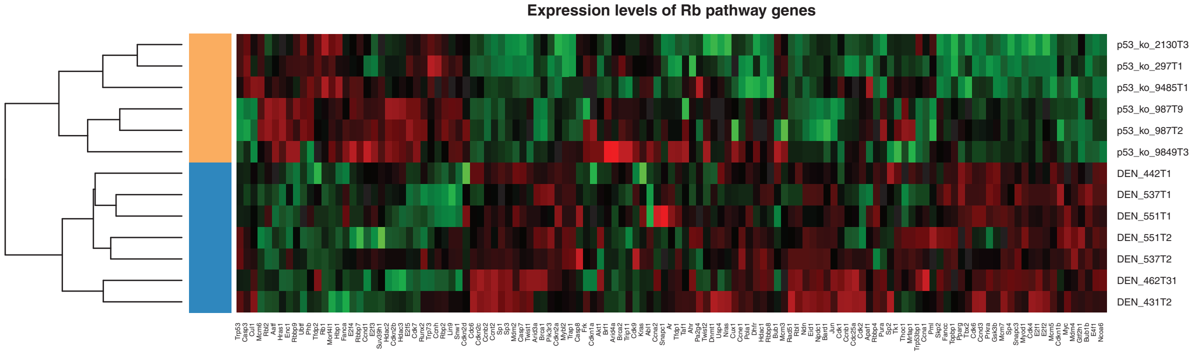

Fig. 2. Heatmap of the 123 genes associated to the Rb pathway. Hierarchical clustering of samples (complete linkage) in the feature subspace of the Rb pathway coincides with the known classes DEN and P53 (lower and upper labeling). Expression values have been mean centred and scaled within each gene

calculations, which allows for the analysis of gene set collections, benchmark studies or approaches that combine the results of several gene set analyses (Väremo *et al.*, 2013). Furthermore, the package also ensures that parallelized random number generation has no effect on outcome (L'Ecuyer *et al.*, 2002).

2.2 Application

We demonstrate the handling of the `GiANT` package using a dataset of six p53-deficient liver tumor samples and seven samples of DEN-induced liver tumors in mice. The dataset includes 30 278 genes and no missing values. Following Katz *et al.* (2012) we extracted the gene expression values and performed a preprocessing of the dataset using the standard workflow. For normalization, baseline and percentile (75%) shifts were performed after $\log_2$ transformation. Katz *et al.* (2012) showed a significant enrichment of a hand-crafted Rb pathway gene set in this dataset. The set consists of 123 up-stream genes, interaction partners and downstream targets related to the Rb pathway. All genes were collected from literature using the PubMed database. Expression levels of the Rb pathway genes in the data are illustrated in Figure 2. In the following we evaluate the uncertainty in the definition of this hand-crafted Rb pathway gene set. As a test statistic the average absolute (Spearman) correlation to the class label of all genes in the set is calculated. Statistical significance is of the gene set is based on a computer intensive resampling test:

```
evaluateGeneSetUncertainty(
  dat = dataset$data,
  labs = dataset$labs,
  geneSet = rbPathway,
  analysis = gsaTools.averageCorrelation(),
  method = "spearman",
  numSamplesUncertainty = 1000,
  numSamples = 1000,
  k = seq(0.01,0.99, by = 0.01))
```

A detailed step by step example is give in the vignette of the `GiANT` package. Figure 1 visualizes the distributions with different degrees of uncertainty: For each value of k the quantiles of the resulting distribution are given. $k = 0$ gives the null distribution with the lower horizontal (green) line showing the corresponding 95%-quantile. The upper horizontal (red) line gives the value of the test statistic for the Rb pathway t^*. The black vertical line indicates the degree of uncertainty where the two distributions (null distribution and distribution with a fixed degree of uncertainty) do not overlap. As we can see, even for 15% of random genes, the perturbed gene sets achieve higher scores than the 95% quantile of the null distribution and are thus statistically significant with respect to the null distribution. This indicates that the Rb pathway gene set is robustly enriched ($k = 0.85$) in the dataset.

3 Conclusion

Sets are the core of gene set enrichment analyses. Every analysis is based on their correct definition. Motivated by critical articles like Sedeño-Cortés and Pavlidis (2014), Retraction for Dixson *et al.* (2014) and Bleazard *et al.* (2015), we developed a new robustness analysis that assesses and quantifies the performed enrichment analyses using a partial resampling approach. These anylses are especially important for newly found pathways, but also well established and curated databases suffer from errors and uncertainty in their gene set definitions. Our method allows now a quantification of the gene set uncertainty and therefore an assessment of the validity of the enrichment analyses. The approach is implemented in the R package `GiANT`, which also provides a comprehensive toolkit for generic gene set analysis. Apart from standard methods like GSEA, user-defined workflows can be constructed readily within the flexible pipeline mechanism. This allows the user to build new high-level analyses, adapted to the specific context of use.

Funding

The research leading to these results received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 602783, the DFG (SFB 1074 project Z1), and the German Federal Ministry of Education and Research (BMBF, Gerontosys II, Forschungskern SyStaR, project ID 0315894A) all to H.A.K.

Conflict of Interest: none declared.

References

Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinf.*, **10**, 47.

Bleazard,T. *et al.* (2015) Bias in microRNA functional enrichment analysis. *Bioinformatics*, **31**, 1592–1598.

Glez-Peña,D. *et al.* (2009) WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis. *Nucleic Acids Res.*, **37**, W329–W334.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Huang,D.W. *et al.* (2007) DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.*, **35**, W169–W175.

Hühne,R. *et al.* (2014) AgeFactDB – the JenAge Ageing Factor Database – towards data integration in ageing research. *Nucleic Acids Res.*, **42**, D892–D896.

Joshi-Tope,G. *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*., **33**, D428–D432.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **28**, 27–30.

Katz,S. *et al*. (2012) Disruption of trp53 in livers of mice induces formation of carcinomas with bilineal differentiation. *Gastroenterology*, **142**, 1229–1239.e3.

L'Ecuyer,P. *et al*. (2002) An object-oriented random-number package with many long streams and substreams. *Oper. Res*., **50**, 1073–1075.

Maciejewski,H. (2013) Gene set analysis methods: statistical models and methodological differences. *Briefings Bioinf*., **15**, 504–518.

Pico,A.R. *et al*. (2008) WikiPathways: pathway editing for the people. *PLoS Biol*., **6**, e184.

Retraction for Dixson *et al*. (2014) Identification of gene ontologies linked to prefrontalhippocampal functional coupling in the human brain. *Proc. Natl. Acad. Sci*., **111**, 13582.

Sedeño-Cortés,A.E. and Pavlidis,P. (2014) Pitfalls in the application of geneset analysis to genetics studies. *Trends Genet*., **30**, 513–514.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: A knowledgebased approach for interpreting genome-wide expression profiles. *PNAS*, **102**, 15545–15550.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Väremo,L. *et al*. (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res*., **41**, 4378–4391.

Zeeberg,B. *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*., **4**, R28.