SOFTWARE TOOL ARTICLE

GenRank: a R/Bioconductor package for prioritization of candidate genes [version 1; referees: 2 not approved]

Chakravarthi Kanduri, Irma Järvelä

Department of Medical Genetics, Haartman Institute, University of Helsinki, Helsinki, Finland

v1 **First published:** 11 Apr 2017, **6**:463 (doi: 10.12688/f1000research.11223.1)
Latest published: 11 Apr 2017, **6**:463 (doi: 10.12688/f1000research.11223.1)

Abstract

Modern high-throughput studies often yield long lists of genes, a fraction of which are of high relevance to the phenotype of interest. To prioritize the candidate genes of complex genetic traits, our R/Bioconductor package GenRank ranks genes based on convergent evidence obtained from multiple layers of independent evidence. We implemented three methods to rank genes that integrate gene-level data generated from multiple layers of evidence: (a) the convergent evidence (CE) method aggregates evidence based on a weighted vote counting method; (b) the rank product (RP) method performs a meta-analysis of microarray-based gene expression data, and (c) the traditional method combines p-values. The methods are implemented in R and are available as a package in the Bioconductor repository (http://bioconductor.org/packages/GenRank/).

This article is included in the Bioconductor gateway.

Open Peer Review

Referee Status: ✗ ✗

	Invited Referees 1	2
version 1 published 11 Apr 2017	✗ report	✗ report

1 **Joshua W. K. Ho**, The Victor Chang Cardiac Research Institute (VCCRI), Australia

2 **Emma E. Laing**, University of Surrey, UK
Carla Moller-Levet, University of Surrey, UK
Huihai Wu, University of Surrey, UK

Discuss this article

Comments (1)

Corresponding author: Chakravarthi Kanduri (chakra.kanduri@gmail.com)

Competing interests: No competing interests were disclosed.

How to cite this article: Kanduri C and Järvelä I. **GenRank: a R/Bioconductor package for prioritization of candidate genes [version 1; referees: 2 not approved]** *F1000Research* 2017, **6**:463 (doi: 10.12688/f1000research.11223.1)

Grant information: The study has been funded by the University of Helsinki (Grant number: 73603104).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 11 Apr 2017, **6**:463 (doi: 10.12688/f1000research.11223.1)

Introduction

Genetic studies employ multiple independent lines of investigation spanning pan-omics approaches to holistically understand the molecular background of complex genetic traits. This includes studying the roles of various forms of genomic variation (e.g. SNPs, InDels, and CNVs) and gene expression in multiple tissues, and the regulation of a single phenotype across single or multiple species (e.g, humans and other relevant model organisms). One of the common objectives of performing such diverse experimental assays across multiple types of cells, tissues, treatments, time-points and species is to find the causal genes underlying a specific disease or trait. Integration of data from such diverse experimental assays (hereafter referred to as evidence layers) would enable prioritization of genes that are most relevant to the phenotype. Meta-analytic approaches that integrate gene-level data from multiple evidence layers have been shown to be successful in identifying and prioritizing candidate genes for complex genetic traits (Ayalew *et al.*, 2012). However, no implementation of candidate gene prioritization methods existed in the Bioconductor project at the time this package was written, which otherwise offers a seamless framework to perform various statistical analyses in biomedical research. The majority of the existing meta-analysis related packages in Bioconductor have been exclusively developed to integrate microarray gene expression data, but do not serve the purpose of integrating gene-level data from multiple study types. Here, we implemented three methods to rank genes by integrating gene-level data generated from multiple evidence layers.

Methods

Operation

The methods are implemented in R and available as a package in the Bioconductor repository (http://bioconductor.org/packages/GenRank/). The package requires R version 3.2.3 or later versions and runs on all operating systems. Figure 1 shows an overview of the workflow of the GenRank package.

Implementation

GenRank provides three methods to prioritize gene-level data obtained through multiple independent evidence layers. It requires a tab-delimited text file with three required fields: gene symbols or IDs, type of evidence layer and a significance statistic (e.g., p-value or effect-size). The first two fields are sufficient for the convergent evidence method. Summary statistics to prioritize the genes are computed as follows.

The convergent evidence (CE) method

The convergent evidence (CE) method aggregates ranks of genes based on a weighted vote counting method. A conceptually similar gene-level integration has been successfully used to prioritize candidate genes in neuropsychiatric diseases (Ayalew *et al.*, 2012).

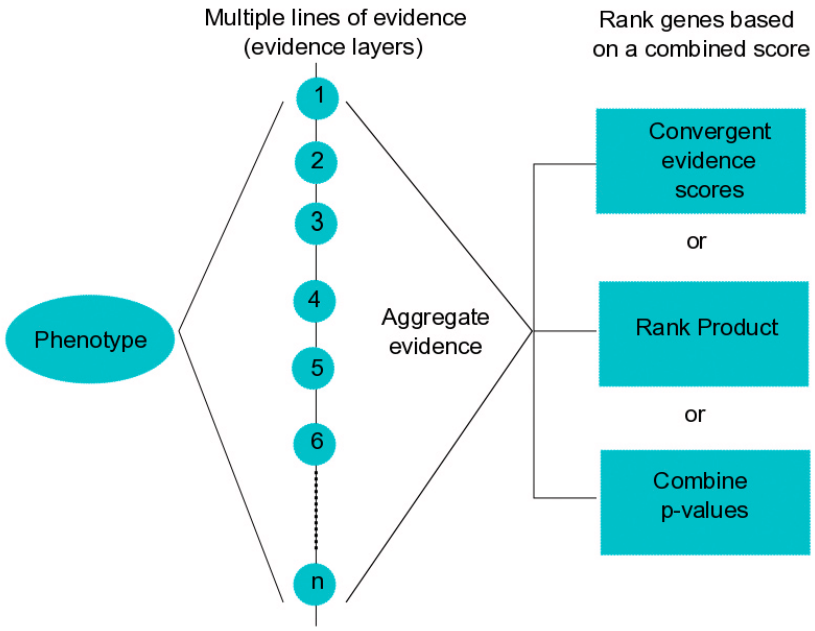

Figure 1. Overarching theme of GeneRank Bioconductor package. To obtain convergent evidence for the molecular basis of phenotypes, GenRank bioconductor package implements three methods to integrate gene-level data generated from multiple independent experiments. Examples of evidence layers are experiment assay-type (e.g., GWAS, RNAseq, ChIPseq), tissue-type (e.g., blood, liver, intestine), cell-type (e.g., neutrophils, lymphocytes), time-series (e.g., 0h, 2h, 6h), species-type (e.g., human, mouse, drosophila), treatment-type (e.g., control, dexamethasone, lipopolysachharide).

Here, to rank genes, we compute convergent evidence scores. The convergent evidence score of gene G is given by

$$CE(G) = CE(G_{L_1})/n(L_1)+....+CE(G_{L_n})/n(L_n)$$

Here $CE(G_{L_i})$ refers to the self-importance of evidence layer-i, while $n(L_i)$ refers to the number of genes within evidence layer-i. Additionally, we propose two other ways to compute convergent evidence scores. One of them is to ignore the number of genes within each layer, thus

$$CE(G) = CE(G_{L_1})+....+CE(G_{L_n})$$

In this case, the convergent evidence score would be equivalent to the primitive vote counting. Another alternative method enables the researchers to determine the importance of each layer based on their own intuition. This involves assigning custom weights to each evidence layer based on their expert knowledge in the field. For example, when a researcher knows that a specific technology could yield less reproducible findings, such evidence layer could be given relatively less weight compared to the other evidence layers. Another objective way of assigning custom weights to each evidence layer could be based on the sample sizes of each evidence layer. In this case the convergent evidence score is

$$CE(G) = CE(G_{L_1}) * w(L_1)+....+CE(G_{L_n}) * w(L_n)$$

where $w(L_i)$ refers to the custom weight assigned to evidence layer-i. Figure 2 shows an illustration of how CE scores are computed.

The rank product (RP) method

The rank product (RP) method has been used widely to perform differential expression analysis in microarray-based gene expression datasets. This biologically motivated method is simple, yet powerful and ranks genes that are consistently ranked highly in replicated experiments, based on the geometric mean (Breitling *et al.*, 2004). This method has been implemented earlier as a Bioconductor package to perform meta-analysis of gene expression experiments (Hong *et al.*, 2006). We adapted the rank product method to identify genes that are consistently highly ranked across evidence layers. The rank product is computed and compared to a permutation-based distribution of rank product values to estimate the proportion of false predictions (pfp; equivalent to FDR).

Combining p-values

Combining p-values has been one of the traditional methods of meta-analysis. To combine p-values of a gene from multiple evidence layers, the p-values should have been estimated from the same null hypothesis. Popular methods to combine p-values include Fisher's and Stouffer's methods, where the latter incorporates custom weights (e.g. sample sizes). These popular methods have already been implemented in the Bioconductor package *survcomp* (Schröder *et al.*, 2011). Here, we built a wrapper around those methods to suit the overarching theme of this package (integrating gene-level data from multiple evidence layers). Missing p-values in some evidence layers could lead to a potential bias when combining p-values. To handle this issue, our implementation returns the combined p-values of only those

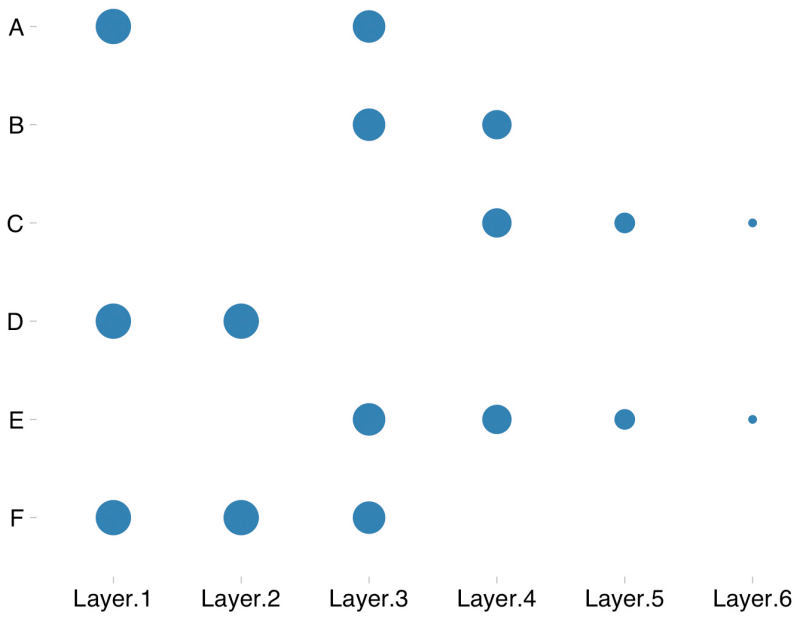

Figure 2. An example of computing convergent evidence scores. This illustration shows six evidence layers (Layer.1–Layer.6). The point indicates the detection of a gene in an evidence layer, while the size of the point indicates the importance of an evidence layer (custom weights assigned by the user). Here, genes A, B and D are detected twice each. However, based on a weighted vote counting method, gene D would get a better rank than genes A and B.

genes, for which p-values are available at least across half of the evidence layers. However, it would be an ideal scenario to have p-values available across all evidence layers.

To avoid a potential bias owing to duplicated genes, duplicated genes are counted only once (as a single vote) within each evidence layer in all the three methods implemented in this package. When retaining duplicated genes, those with significant test statistic (e.g low p-values or high effect-size) were retained.

Use cases

The use cases are explained in detail, with example data in the package vignette available at the package webpage here:
https://www.bioconductor.org/packages/devel/bioc/vignettes/GenRank/inst/doc/GenRank_Vignette.html

Oikkonen *et al.* (2016) serves as an interesting use case that used convergent evidence scores to prioritize candidate genes obtained through diverse experiment types in a complex genetic trait.

bioRxiv

An earlier version of this article can be found on bioRxiv at (http://biorxiv.org/content/early/2016/04/12/048264)

Software availability

The GenRank package is hosted on Bioconductor at:
http://bioconductor.org/packages/GenRank/.

Latest source code:
https://github.com/Bioconductor-mirror/GenRank

Archived source code as at the time of publication:
http://doi.org/10.5281/zenodo.439738. (Kanduri & Järvela, 2017)

License: Artistic-2.0 license.

Author contributions

CK and IJ conceived the study and drafted the manuscript. CK carried out the implementation.

Competing interests

No competing interests were disclosed.

Grant information

The study has been funded by the University of Helsinki (Grant number: 73603104).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Ayalew M, Le-Niculescu H, Levey DF, *et al.*: **Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction.** *Mol Psychiatry.* 2012; **17**(9): 887–905.
PubMed Abstract | Publisher Full Text | Free Full Text

Breitling R, Armengaud P, Amtmann A, *et al.*: **Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments.** *FEBS Lett.* 2004; **573**(1–3): 83–92.
PubMed Abstract | Publisher Full Text

Hong F, Breitling R, McEntee CW, *et al.*: **RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.** *Bioinformatics.* 2006; **22**(22): 2825–2827.
PubMed Abstract | Publisher Full Text

Kanduri C, Järvela I: **GenRank: Bioconductor package for candidate gene prioritization based on convergent evidence [Data set].** *Zenodo.* 2017.
Data Source

Oikkonen J, Onkamo P, Järvelä I, *et al.*: **Convergent evidence for the molecular basis of musical traits.** *Sci Rep.* 2016; **6**: 39707.
PubMed Abstract | Publisher Full Text | Free Full Text

Schröder MS, Culhane AC, Quackenbush J, *et al.*: ***survcomp*: an R/Bioconductor package for performance assessment and comparison of survival models.** *Bioinformatics.* 2011; **27**(22): 3206–8.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✗ ✗

Version 1

Referee Report 03 May 2017

doi:10.5256/f1000research.12108.r21806

✗ **Emma E. Laing** [1], **Carla Moller-Levet** [2], **Huihai Wu** [2]

[1] Department of Microbial Sciences, School of Biosciences and Medicine, Faculty of Health and Medical Sciences, University of Surrey, Guildford, UK

[2] Core Bioinformatics Facility, Faculty of Health and Medical Sciences, University of Surrey, Guildford, UK

Being able to combine evidence from multiple sources to prioritize genes associated with a particular scientific question is very desirable. *GenRank* is a Bioconductor package that aims to integrate gene-level data generated from multiple layers of evidence (e.g. multiple study types, tissues, analysis tools) to prioritize candidate genes (**it is not clear for what**). Three methods are implemented via easy-to-use functions. The Convergent Evidence (CE) method counts the number of times a gene is present in each layer of evidence. Counts can be weighted relative to the total number of genes per layer of evidence and further weighted by the type of evidence. The Rank Product (RP) method applies the rank product strategy originally developed for the analysis of microarray data to the p-values or effect sizes across layer of evidence for a set of genes. The third method is a wrapper of the **combine.test** function from the *survcomp* package, which combines p-values estimated from the same null hypothesis in different studies.

The manuscript and package, in their current form, are of limited value. Each of the functions 'wrap' existing methods and is a task easily achieved by any proficient bioinformatician i.e. a bioinformatician would simply use the existing packages. Thus, the likely users of the package are biologists with limited experience or R/programming who want simple to use tools. Whilst the package offers simple to use functions there is limited discussion on how to achieve such analysis nor the 'weight' of each parameter, how to approach such analysis, how to merge the data (so no missing data), etc. For example, no advice is offered for the parameters of 'z.transform' or 'logit' for combing p-values. Whilst this information may be available in the original *survcomp* package this defeats the idea of having an easy 'out-of-the-box' package which *GenRank* aims to be. Without such information it is not easy to see the contribution of this work to the field.

In light of the above, and the technical aspects picked up below, we believe this manuscript and package requires a substantial amount of work before it can be indexed and make a contribution to the scientific community.

Technical aspects: The tool was installed, manual and vignette read, and all examples successfully run. The tool was also tested with in-house data and there were no problems.
Technical issues are:

1. *GenRank* package indicates a dependency on R (>= 3.2.3), however, I could only install *GenRank* on R 3.3.3. It looks like *GenRank* depends on *survcomp*, which depends on *SuppDists*, which is only available for R 3.3.3.

   Details:
   In R 3.2.3, I tried to install *GenRank* as follows:

   sudo R CMD INSTALL GenRank_1.2.0.tar.gz

   I got an error indicating that I needed dependency *survcomp*. I tried to install *survcomp* but I got an error indicating that I needed several dependencies. I was able to install all dependencies successfully (except for *SuppDists*) by running:

   install.packages("package_name", repos="http://cran.cnr.berkeley.edu", dependencies=TRUE)

   For *SuppDists* I then tried

   sudo R CMD INSTALL SuppDists_1.1-9.4.tar.gz

   which produced the error

   ERROR: this R is version 3.2.3, package 'SuppDists' requires R >= 3.3.0

   In addition, the *SuppDists* package's maintenance status is orphaned, i.e. the maintainer is unresponsive (dated 2013-03-22).

2. In the abstract the RP method is described as: "the rank product (RP) method performs a meta-analysis of microarray-based gene expression data" however, in the context of the manuscript, the method is not restricted to microarray data. The reference manual has a more suitable description under the **ComputeRP** function: "the rank product (RP) method returns ranks of the genes based on rank product method".

3. The writing style of the Vignette could be improved, in particular the RP tutorial section.

4. The *PC* argument of the **ComputeCE** function can have three different values, which correspond to each of the three different ways of computing the CE score. It would very useful if the meaning of these options were included in the R help documentation.

5. The *method* argument of the **ComputeP** function can have three different values, which correspond to each of the three different ways of computing the combined p-values. It would be very useful if a short description of each method was included in the R help documentation. Please see the description provided in **combine.test** function.

6. There are a few assumptions and requisites described in the vignette that could be included in the R help documentation, for example:

- In the RP method, the gene-list or the number of genes should be the same across all evidence layers.
- In the Combining p-values method, the p-values must be one-sided.

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Referee Report 19 April 2017

doi:10.5256/f1000research.12108.r21804

Joshua W. K. Ho
The Victor Chang Cardiac Research Institute (VCCRI), Darlinghurst, NSW, Australia

In this manuscript, Kanduri and Jarvela present a bioconductor R package that facilitates integration of multiple layers of experimental data to prioritise disease- or phenotype-associated genes. This fairly simple package contains three methods: convergence evidence (CE), rank product (RE) and combination of p-values (p). These methods were intended to combine multiple sources of experimental evidence if the evidence are in the form of presence/absence of detection of genes (CE), ranking of genes (RE) and p-values of genes (p).

This paper did not present any theoretical justification or empirical evaluation of these methods, and the 'Use cases' presented in their R package's vignette is based on some very simple toy examples. There is no evidence in this paper or on the github repository that directly supports their claim that these methods can 'prioritize the candidate genes of complex genetic traits' (Abstract).

After further careful examination of their source code, I believe their methods have important flaws, and their description in the text contains errors.

The major flaw is that they fail to consider two important implicit assumptions: (1) each evidence layer is independent, and (2) the same number of genes are tested in each evidence layer. All the methods described in this manuscript are only potentially valid if these two assumptions are satisfied. Nonetheless,

considering the wide range of applications described in their Introduction, it is very easy to imagine these assumptions will be violated in practice. In fact, the failure to consider differences in the gene universe in different evidence layer (assumption 2) is a particularly problematic issue. For example, when combining data from different detection platforms (custom microarrays, targeted or non-targeted proteomic experiments, and NGS-based data), the number of genes that are probed in each experiment can vary a lot. Their CE method implicitly assumes the gene universe to be identical. Their RP method assumes that any missing genes are imputed with rank (n+1) where n is the number of detectable genes in that evidence layer (described in the online Vignette of the package). Their p method excludes genes that have too many missing entries. None of these approaches are entirely appropriate to address the issues related to the violation of these assumptions.

Both RP and the p-value combination methods were designed for other more specific purposes, and have been implemented in other bioconductor packages. They were not specifically designed for performing the type of integrative meta-analysis proposed by the authors in this manuscript.

The CE method is essentially a very simple weighted sum of presence/absence detection across multiple layers. Even if the two implicit assumptions are satisfied, I still find this CE method rather useless. There is no statistical significance associated with the CE score, and the inclusion of a 'custom weight' is rather arbitrary. In essence, the entire method can be implemented in 2-3 lines of R code. It does not seem necessary to develop a whole bioconductor package for this.

I also found a technical error in weighted CE equation on page 3. Based on their source code (https://github.com/KanduriC/GenRank/blob/master/R/compute_CE.R), the equation should have been:

CE(G) = [CE(G_L1)*w(L1)+...+CE(G_Ln)*w(Ln)]/ [w(L1)+...+w(Ln)].

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
No

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Referee Expertise: Bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Discuss this Article

Version 1

Reader Comment (*Member of the F1000 Faculty*) 17 Apr 2017

Peter Uetz, Center for the Study of Biological Complexity, Virginia Commonwealth University, USA

Not clear from title what the prioritization is for.

Competing Interests: No competing interests were disclosed.