SOFTWARE TOOL ARTICLE

regionReport: Interactive reports for region-based analyses [version 1; referees: 1 approved, 1 approved with reservations, 1 not approved]

Leonardo Collado-Torres[1-3], Andrew E. Jaffe[1-4], Jeffrey T. Leek[1,3]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205, USA
[2]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, Maryland, 21205, USA
[3]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, 21205, USA
[4]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205, USA

v1 **First published:** 01 May 2015, **4**:105 (doi: 10.12688/f1000research.6379.1)
Latest published: 29 Jun 2016, **4**:105 (doi: 10.12688/f1000research.6379.2)

Abstract

regionReport is an R package for generating detailed interactive reports from regions of the genome. The report includes quality-control checks, an overview of the results, an interactive table of the genomic regions and reproducibility information. regionReport can easily be expanded with report templates for other specialized analyses. In particular, regionReport has an extensive report template for exploring derfinder results from annotation-agnostic RNA-seq differential expression analyses.

This article is included in the RPackage gateway.

This article is included in the Bioconductor gateway.

Open Peer Review

Referee Status: ✔ ? ✔

	Invited Referees 1	2	3
REVISED version 2 published 29 Jun 2016	✔ report		
version 1 published 01 May 2015	✘ report	? report	✔ report

1 **Timothy J. Triche Jr** , Keck School of Medicine of the University of Southern California, USA

2 **Karthik Ram** , University of California Berkeley, USA

3 **David Robinson** , Princeton University, USA

Discuss this article

Comments (0)

Corresponding author: Jeffrey T. Leek (jtleek@gmail.com)

Competing interests: No competing interests were disclosed.

How to cite this article: Collado-Torres L, Jaffe AE and Leek JT. **regionReport: Interactive reports for region-based analyses [version 1; referees: 1 approved, 1 approved with reservations, 1 not approved]** *F1000Research* 2015, **4**:105 (doi: 10.12688/f1000research.6379.1)

Grant information: J.T.L. was partially supported by NIH Grant 1R01GM105705, L.C-T. was supported by Consejo Nacional de Ciencia y Tecnologia Mexico 351535.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 01 May 2015, **4**:105 (doi: 10.12688/f1000research.6379.1)

Introduction

Many analyses of genomic data result in regions along the genome that associate with a covariate of interest. These genomic regions can result from identifying differentially bound peaks from ChIP-seq data[1], identifying differentially methylated regions (DMRs) from DNA methylation data[2], or performing base-resolution differential expression analyses using RNA sequencing data[3,4]. The genomic regions themselves are commonly stored in a `GRanges` object from GenomicRanges[5] when working with R or the BED file format on the UCSC Genome Browser[6], but other information on these regions, for example summary statistics on the magnitude of effects and statistical significance, also provide useful information. The usage of R in genomics is increasingly common due to the usefulness and popularity of the Bioconductor project[7], and in the latest version (3.0), 181 unique packages use `GenomicRanges` for many workflows, demonstrating the widespread utility of identifying and summarizing characteristics of genomic regions.

Here we introduce `regionReport` which allows users to explore genomic regions of interest through interactive stand-alone HTML reports that can be shared with collaborators. These reports are flexible enough to display plots and quality control checks within a given experiment, but can easily be expanded to include custom visualizations and conclusions. The resulting HTML report emphasizes reproducibility of analyses[8] by including all the `R` code without obstructing the resulting plots and tables. We envision `regionReport` will provide a useful tool for exploring and sharing genomic region-based results from high throughput genomics experiments.

Methods

Implementation

The package includes a `R` Markdown template which is processed using knitr[9] and rmarkdown[10], then styled using knitrBootstrap[11]. This package generates a HTML report that includes a series of plots for checking the quality of the results and browsing the table of regions. Each element of the report has a brief explanation, although actual interpretation of the results is dataset- and workflow-dependent. To facilitate navigation a menu is always included, which is useful for users interested in a particular section of the report. Figure 1, panel a shows the menu of the general report for a set of regions with associated p-values. The code for each plot or table is hidden by default and can be shown by clicking on the appropriate toggle as shown in Figure 2.

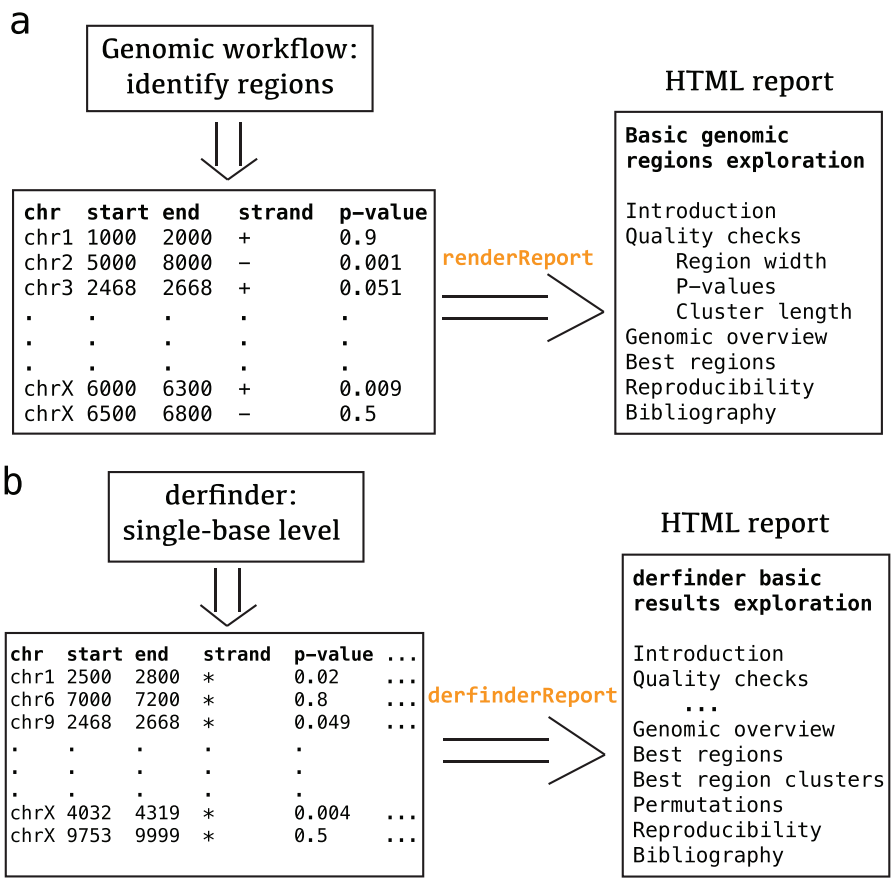

Figure 1. `regionReport` workflow. Example region input, the appropriate `regionReport` function to use, and menu of the resulting report for the general use case (panel **a**) and `derfinder` results (panel **b**).

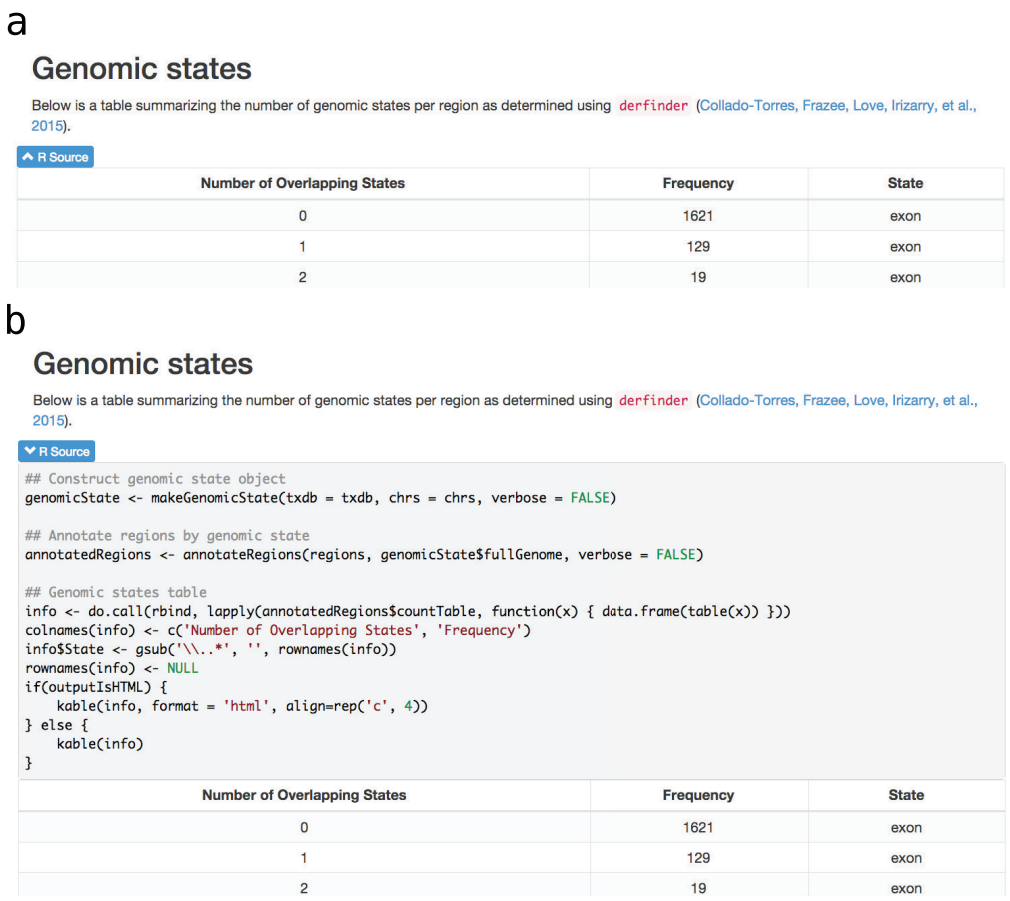

Figure 2. Interactively display the code for each table/figure in the report. View by default (panel **a**) and after clicking on the "R source" toggle (panel **b**) for a section of the general report. The full report is available at the supplementary website and includes a toggle to hide/show all the R code.

Quality checks

This section of the report includes a variety of quality control steps which help the user determine whether the results are sensible. The quality control steps explore:

- P-values, Q-values, and FWER adjusted p-values
- Region width
- Region area: sum of single-base level statistics (if available)
- Mean coverage or other score variables (if available)

A combination of density plots and numerical summaries are used in these quality checks. If there are statistically significant regions, the distributions are compared between all regions and the significant ones. For example, the distribution region widths might have a high density of small values for the global results, but shifted towards higher values for the subset of significant regions.

Genomic overview

The report includes plots to visualize the location of all the regions as well as the significant ones. Differences between them can reveal location biases. The nearest known annotation feature for each region is summarized and visually inspected in the report.

Best regions

An interactive table with the top 500 (default) regions is included in this section. This allows the user to sort the region information according to their preferred ranking option. For example, lowest p-value, longest width, chromosome, nearest annotation feature, etc. The table also allows the user to search and subset it interactively. A common use case is when the user wants to check if any of the regions are near a known gene of their interest.

Reproducibility

At the end of the report, detailed information is provided on how the analysis was performed. This includes the actual function call to generate the report, the path where the report was generated, time spent, and the detailed `R` session information including package versions of all the dependencies.

The `R` code for generating the plots and tables in the report is included in the report itself, thus allowing users to manually reproduce any section of the report, customize them, or simply change the graphical parameters to their liking.

`derfinder` report

When exploring `derfinder` results, for each of the best 100 (default) DERs a plot showing the coverage per sample is included

in the report. These plots allow the user to visualize the differences identified by `derfinder` along known exons, introns and isoforms. The plots are created using `derfinderPlot`, also available via Bioconductor.

Due to the intrinsic variability in RNA-seq coverage data or mapping artifacts, in situations where there are two candidate DERs that are relatively close there might be reasons to consider them a single candidate DER and its important to visualize them. This tailored report groups candidate DERs into clusters based on a distance cutoff. After ranking them by their area, for the top 20 (default) clusters it plots tracks with the coverage by sample, the mean coverage by group, the identified candidate DERs colored by whether they are statistically significant, and known alternative transcripts. Figure 1, panel b shows the main categories of the report generated from a richer region data set than in the general case.

Operation

Installation. `regionReport` and required dependencies can be easily installed from Bioconductor with the following commands:

```
source("http://bioconductor.org/biocLite.R")

biocLite("regionReport")
```

Input. To generate the report, the user first has to identify the regions of interest according to their analysis workflow. For example, by performing bumphunting to identify DMRs with bumphunter. The report is then created using `renderReport()` which is the main function in this package as shown in Figure 1, panel a. The argument `customCode` can be used to customize the report if necessary.

For the `derfinder` use case, the `derfinderReport()` function creates the recommended report that includes visualizations of the coverage information for the best regions and clusters of regions.

Output. A small example can be generated using:

```
example("renderReport", "regionReport", ask=FALSE)
```

The resulting HTML file will open in the users default browser. Note that alternative output formats such as PDF files can also be generated, although they are not as dynamic and interactive as the HTML format.

Use cases

The supplementary website contains reports using DiffBind, `bumphunter` and `derfinder` results. The `derfinder` use case is illustrated with data sets previously described in 3 which span simulation results, a moderately sized data set (25 samples), and a large data set with 487 samples; thus covering a wide range of scenarios.

Summary

`regionReport` creates interactive reports from a set of regions and can be used in a wide range of genomic analyses. Reports generated with `regionReport` can easily be extended to include further quality checks and interpretation of the results specific to the data set under study. These shareable documents are very powerful when exploring different parameter values of an analysis workflow or applying the same method to a wide variety of data sets. The reports allow users to visually check the quality of the results, explore the properties of the genomic regions under study, and inspect the best regions and interactively explore them.

Furthermore, `regionReport` promotes reproducibility of data exploration and analysis. Each report provides R code that can be used as the starting point for other analyses within a dataset. `regionReport` provides a flexible output for exploring and sharing results from high throughput genomics experiments.

Software availability

Software access

`regionReport` is freely available via Bioconductor at bioconductor.org.

http://leekgroup.github.io/regionReportSupp/ hosts the code for generating three types of reports as well as the resulting HTML reports generated by `regionReport`. Versions of all software used are included in the reports.

Latest source code

The latest source code is available at Bioconductor and github.com/leekgroup/regionReport via the git-svn bridge although we recommend users to install `regionReport` directly from Bioconductor.

Archived source code as at the time of publication

Archived source code available at http://dx.doi.org/10.5281/zenodo.17083

License

Artistic-2.0.

Author contributions

L.C-T. conceived and developed the `regionReport` package, supervised by A.E.J. and J.T.L. All authors wrote and approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

J.T.L. was partially supported by NIH Grant 1R01GM105705, L.C-T. was supported by Consejo Nacional de Ciencia y Tecnolog'ia México 351535.

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

References

1. Stark R, Brown G: ***DiffBind*: differential binding analysis of ChIP-Seq peak data**. 2011.
Reference Source

2. Jaffe AE, Murakami P, Lee H, *et al.*: **Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.** *Int J Epidemiol.* 2012; **41**(1): 200–209.
Publisher Full Text

3. Torres LC, Frazee AC, Love MI, *et al.*: **derfinder: Software for annotation-agnostic rna-seq differential expression analysis.** *bioRxiv.* 2015; 015370.
Publisher Full Text

4. Frazee AC, Sabunciyan S, Hansen KD, *et al.*: **Differential expression analysis of RNA-seq data at single-base resolution.** *Biostatistics.* 2014; **15**(3): 413–26.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Lawrence M, Huber W, Pages H, *et al.*: **Software for computing and annotating genomic ranges.** *PLoS Comput Biol.* 2013; **9**(8): e1003118.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Rosenbloom KR, Armstrong J, Barber GP, *et al.*: **The UCSC Genome Browser database: 2015 update.** *Nucleic Acids Res.* 2015; **43**(Database issue): D670–D681.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Huber W, Carey VJ, Gentleman R, *et al.*: **Orchestrating high-throughput genomic analysis with bioconductor.** *Nat Methods.* 2015; **12**(2): 115–121.
PubMed Abstract | Publisher Full Text

8. Sandve GK, Nekrutenko A, Taylor J, *et al.*: **Ten simple rules for reproducible computational research.** *PLoS Comput Biol.* 2013; **9**(10): e1003285.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Xie Y: **Dynamic Documents with R and knitr**. CRC Press. 2013; 216.
Reference Source

10. Rstudio. **RMarkdown: Dynamic Documents for R**. 2014.
Reference Source

11. Hester J: **Knitr Bootstrap framework, R package**. 2015.
Reference Source

Open Peer Review

Current Referee Status: ✗ ? ✓

Version 1

Referee Report 22 June 2015

doi:10.5256/f1000research.6840.r8559

✓ **David Robinson**
Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA

The authors present regionReport, an R package to produce interactive HTML reports from a genomic-region based analysis, such as those produced by derfinder, bumphunter or DiffBind. The report shows quality control summaries, interactive tables of the most significant regions, and information on how regions lie in exons, introns, and intergenic regions. Both novice and expert genomicists are sure to gain much from this, and the paper clearly and concisely explains the software and its use, while providing useful examples and instructions.

The idea of producing common reports for a Bioconductor object is ingenious, and will hopefully inspire packages for other types of biological data. One of the great strengths of the package is the reproducibility practices it follows. For example, the section at the end of the produced report that shows reproducibility information, such as the original command, the session info, and the amount of time the report took to generate, is a great idea. (Indeed, the option to add sessionInfo() and timers could probably be baked into rmarkdown, or a thin wrapper thereof). Another strength is the use of modern knitr templates, such as expandable tables. Scientists who want to develop automated reports should use this package as a guide.

Overall my concerns are minor, and mostly concern the package rather than the paper, some of which I attempt to address in a GitHub pull request.

In pull request

- If the renderReport function leaves early (for example, if it is interrupted by the user hitting Stop) it strands the user's R session in a working directory. Using the on.exit function, as described here, lets R return to the original directory instead.

- The options for customization of the report are limited, by the customCode argument, to chunks between the main text and the reproducibility section. Genomicists may wish to take advantage of these reports while customizing some of their outputs. (For example, the authors of region-finding packages may wish to wrap renderReport with a customized template for their own objects). I've added a template argument in my pull request, and go over another suggestion below.

Not in pull request

- The `template` argument is a start towards greater customization, but a further improvement would be to allow the user to provide a list of customized internal chunks (for example, density-pvalue). As it is now, these are constructed in the renderReport function and cannot be altered without rewriting the entire function. This suggests finding a way to abstract them, such as bringing them in

from a separate file, would be useful.

- As one example of an important customization I'd make: the reports show density plots of p-values and q-values, but in my experience genomicists are more accustomed to histograms (especially since bumps in density plots may be misleading, while histograms can get a better sense of which bumps are meaningful). I understand if the authors wish to keep it as a density plot, but if so I would appreciate a way to change it for my own use.

Minor issues

- The use of "smart quotes" in code within the PDF, such as source (“http://bioconductor.org/biocLite.R”), make it inconvenient to copy and paste them into an R terminal. If there's any way this could be remedied by the author or editors, it should.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 11 June 2015

doi:10.5256/f1000research.6840.r8558

Karthik Ram
Berkeley Institute for Data Science, University of California Berkeley, Berkeley, USA

Review of regionReport: Interactive reports for region based analysis.

This short software tool article describes a new R package, `regionReport`, available from Bioconductor that generates HTML reports which allows users to explore genomic regions and quickly scan quality control information. The reports also provide provenance of code used in the analysis, including detailed session information to facilitate as much reproducibility as possible.

The paper/report clearly describes functionality of the tool, potential use cases, and details on installation and operation.

Suggestions for improvement

1. Given that you are describing an HTML application, it would be really helpful to include more screenshots/figures rather than just the one, and the workflow diagram. Most readers are unlikely to install the package immediately (see further comments on 3) and so it would help to make the value proposition clear. I would also suggest annotating these figures highlighting the key parts. Happy to approve the narrative itself after this revision.

2. Given that the package primarily generates html based reports, it would be of great value to have these files in the `gh-pages` branch of a GitHub repo, such that reports could be automatically made available under `https://USERNAME.github.io/repo/file.html` much like the page that describes the supplementary material. One way to easily enable this would be to use the functionality in the `git2r` package (disclosure: I am a coauthor on the package) to programmatically create a new branch (if it doesn't already exist), generate the report, then add those files and push to GitHub (assuming the same folder in under git revision control). Obviously I am not expecting the authors to add this suggestion to the current version of the package, but as

something to consider for future versions.

3. Reduce the number of dependencies. `locfdr` is no longer on CRAN. On a slightly slower than normal connection (currently on travel) it took a fairly long time to track down and install all the dependencies. I'd recommend moving non-essential dependencies to suggests and using something like this to selectively install packages as needed using `requireNamespace(pkg, quietly = TRUE)`. It was disappointing to go down a rabbit hole of dependencies and still not be able to install and run examples. However, I found the report examples posted online (here: http://leekgroup.github.io/regionReportSupp/bumphunter-example/index.html and here: http://leekgroup.github.io/regionReportSupp/DiffBind-example/index.html) extremely useful. The use of Twitter bootstrap also adds a layer of a familiarity that I found extremely useful.

4. It would be nice to have the package generate a direct link to the bib file under the bibliography.

Competing Interests: In my suggestions to improve the software, I've recommended adding one dependency tool on which I am a coauthor. The software itself is free and I don't benefit from any citations (there is no paper associated with that software). In this case it would really help make it easier for researchers to publish their reports generated by the software described in this paper. It did not affect my review and I have offered that addition only as a suggestion (not a requirement to acceptance).

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 18 May 2015

doi:10.5256/f1000research.6840.r8554

Timothy J. Triche Jr
Jane Anne Nohl Division of Hematology, USC/Norris Comprehensive Cancer Center, Keck School of Medicine of the University of Southern California, Los Angeles, CA, USA

Needs more figures to demonstrate why a user would choose this tool. For example http://leekgroup.github.io/regionReportSupp/bumphunter-example/index.html (but even better would be to show an example, e.g. ITGB2 exon inclusion/exclusion or multiscale DMRs, where in our hands at least, nothing else short of IGV really does the job, and IGV doesn't do it that well.) The software is a firm foundation but the writeup needs work if it is to be compelling and thus influence readers to try out an unfamiliar tools.

My apologies for being harsh, but without figures, an applied paper simply will not be read. I would be less harsh if the underlying work were not compelling enough to command broader interest. A poor writeup will doom the work to obscurity.

Competing Interests: No competing interests were disclosed.

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.

Author Response 18 May 2015

Jeffrey Leek, Johns Hopkins Bloomberg School of Public Health, USA

Thanks, we will update with more figures and expand the description. This is meant to be a short description of the software but we certainly appreciate the feedback on how to increase users. We will update the draft and respond shortly.

Competing Interests: No competing interests were disclosed.