Check for updates

SOFTWARE TOOL ARTICLE

restfulSE: A semantically rich interface for cloud-scale genomics with Bioconductor [version 1; referees: 2 approved]

Shweta Gopaulakrishnan[1], Samuela Pollack[1,2], BJ Stubbs[1], Hervé Pagès[3], John Readey[4], Sean Davis[5], Levi Waldron[6], Martin Morgan[7], Vincent Carey [1]

[1]Channing Division of Network Medicine, Harvard Medical School, Boston, Massachusetts, 02115, USA
[2]Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, 02115, USA
[3]Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109, USA
[4]Tools and Cloud Technology, HDF Group, Seattle, WA, 98109, USA
[5]Center for Cancer Research, National Cancer Institute, USA, Bethesda, Maryland, 20892, USA
[6]Epidemiology and Biostatistics, CUNY School of Public Health, New York, New York, 10027, USA
[7]Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, New York, 14203, USA

v1 **First published:** 07 Jan 2019, **8**:21 (https://doi.org/10.12688/f1000research.17518.1)
Latest published: 07 Jan 2019, **8**:21 (https://doi.org/10.12688/f1000research.17518.1)

Abstract
Bioconductor's SummarizedExperiment class unites numerical assay quantifications with sample- and experiment-level metadata. SummarizedExperiment is the standard Bioconductor class for assays that produce matrix-like data, used by over 200 packages. We describe the restfulSE package, a deployment of this data model that supports remote storage. We illustrate use of SummarizedExperiment with remote HDF5 and Google BigQuery back ends, with two applications in cancer genomics. Our intent is to allow the use of familiar and semantically meaningful programmatic idioms to query genomic data, while abstracting the remote interface from end users and developers.

Keywords
Bioinformatics, REST APIs, HDF5, BigQuery, Bioconductor

This article is included in the Bioconductor gateway.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 07 Jan 2019	✓ report	✓ report

1 **Dennis J. Hazelett**, Cedars-Sinai Medical Center, USA

2 **Sheila Reynolds**, Institute for Systems Biology, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Vincent Carey (stvjc@channing.harvard.edu)

Author roles: Gopaulakrishnan S: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Writing – Review & Editing; **Pollack S**: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Writing – Review & Editing; **Stubbs B**: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Writing – Review & Editing; **Pagès H**: Conceptualization, Data Curation, Formal Analysis, Methodology, Software; **Readey J**: Conceptualization, Data Curation, Methodology, Software; **Davis S**: Conceptualization, Methodology, Software, Writing – Review & Editing; **Waldron L**: Conceptualization, Data Curation, Methodology, Resources, Software, Writing – Review & Editing; **Morgan M**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Resources, Software, Writing – Review & Editing; **Carey V**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Support for the development of this software was provided by NIH grants NCI U01 CA214846 (Carey, PI), NCI U24 CA180996 (Morgan, PI), and NHGRI 1U24HG010263-01 (J Taylor, PI), and Chan Zuckerberg Initiative DAF 2018-183436 (Carey, PI).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Gopaulakrishnan S *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Gopaulakrishnan S, Pollack S, Stubbs B *et al.* **restfulSE: A semantically rich interface for cloud-scale genomics with Bioconductor [version 1; referees: 2 approved]** *F1000Research* 2019, **8**:21 (https://doi.org/10.12688/f1000research.17518.1)

First published: 07 Jan 2019, **8**:21 (https://doi.org/10.12688/f1000research.17518.1)

Introduction

Analyses of multiomic archives like The Cancer Genome Atlas (TCGA) and single-cell transcriptomic experiments such as the 10x 1.3 million mouse neuron dataset typically begin with downloads of large files and conversion of file contents into formats based on local preferences. In this paper we consider how targeted queries of large remote genomic data resources can be conducted using methods available for Bioconductor's *SummarizedExperiment* class. For large data archives that have been centralized in cloud storage, use of this approach can help diminish effort required to manage local storage, and can facilitate interactive analysis of data subsets in familiar programming idioms, without downloading entire datasets. Clients for HDF5 or Google BigQuery are available in numerous languages; our Bioconductor interface permits access to remote archives of genomic data with familiar and semantically meaningful programmatic idioms, while abstracting the remote interface from end users and developers.

Methods: Data structures and remote back ends

The `SummarizedExperiment` class and related methods

Let Q denote a matrix of quantifications arising from a genome scale assay with G assay features measured on N experimental samples. The elements of Q are the numbers q_{ij}, $i = 1, \ldots, G$, $j = 1, \ldots, N$. Bioconductor's SummarizedExperiment structure manages feature quantifications with associated metadata about assay features and samples.

In the 10x mouse neuron dataset, $G = 27998$ and $N = 1.3$ million. Each of the G features is a gene, and it is useful to have handy a number of feature annotations like gene name, location, functional role; suppose each gene has F such features recorded. When these quantifications and associated annotations are managed in a Bioconductor `SummarizedExperiment` `X`, the matrix Q is programmatically bound to a $G \times F$ table of feature-level metadata accessible by the `rowData` method, and to an $N \times R$ table of sample-level metadata accessible by `colData`, where R denotes the number of sample-level metadata features recorded (Huber *et al.*[1]). See Figure 1.

In the context of R programming, let `K` denote a vector of feature identifiers, `S` denote a vector of sample identifiers. The standard subsetting idiom `X[K,S]` expresses filtering of the all the information in Q and the associated metadata to features `K` and samples `S`. A `GRanges` instance (Lawrence *et al.*[2]) defining genomic coordinates for features may be bound to `X`, facilitating queries defined by genomic location (using, for example,

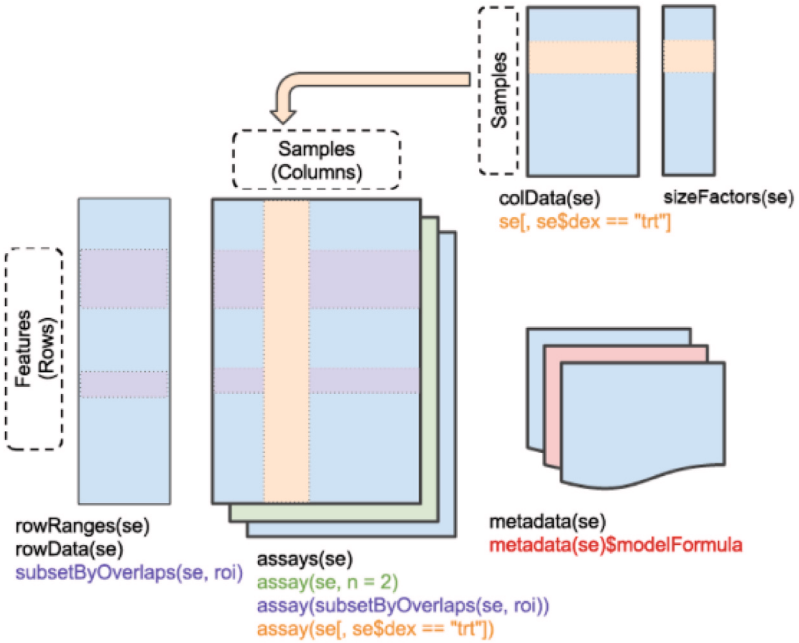

Figure 1. Schematic of SummarizedExperiment class structure. Colored regions of panels within the schematic are linked with command examples in colored text beneath the panels. For example, the purple command `subsetByOverlaps(se, roi)` would produce a restricted `RangedSummarizedExperiment` instance with features limited to those colored purple. The `sizeFactors` component is specific to a subclass for single cell data.

`subsetByOverlaps`) to isolate features coincident with or near the elements of a set of query genomic ranges (eg., binding peaks). This outline of genomic data representation and analysis is characteristic of Bioconductor.

Examples of remote back ends

Google BigQuery. The Institute for Systems Biology Cancer Genomics Cloud project (ISB-CGC) (ISB[3]) uses Google BigQuery to provide access to various public cancer genomics resources including TCGA and the PanCancer Atlas (Hoadley *et al.*[4]). The `pancan_SE` function of *restfulSE* constructs queries that derive `SummarizedExperiment` instances using quantifications and annotations for PanCancer atlas experiments managed in BigQuery tables.

HDF Scalable Data Service (HSDS). An AWS S3-based distributed data object model for HDF5 datasets, including a RESTful API to structure, populate, and query HDF5 archives, has been implemented by the HDF Group. A number of datasets of interest in bioinformatics are served through HDF Kita Lab in the `/shared/bioconductor` folder.

Lazy data retrieval via DelayedArray

The *restfulSE* package provides interfaces to BigQuery and HSDS so that the numerical content housed in these services satisfies the API of the Bioconductor *DelayedArray* (Pagès and Hickey[5]). Any `DelayedArray` instance can serve as the `assay` component of a `SummarizedExperiment` instance. Thus the capacities of `SummarizedExperiment` to bind semantically rich metadata to genome-scale assays are extended implicitly to data resources for which no standards exist for associating substantive metadata.

In conjunction with the *rhdf5client* and *bigrquery* packages, *restfulSE* functions translate filtering and selection operations which are readily defined using `rowData`, `rowRanges`, `colData` into formal queries resolvable by the HDF5 and BigQuery services. Numerical results are transmitted from server to client only when needed.

Results

The RESTful `SummarizedExperiment` representation allows complicated research queries to be obtained in a concise, fast, convenient and robust fashion, as illustrated by the following examples.

Hybrid data/annotation strategy for integrative analysis

The following code chunk, which generates Figure 2, illustrates the use of the *restfulSE* protocol with the ISB-CGC BigQuery back end.

```
library(SummarizedExperiment)
library(BiocOncoTK)       # uses restfulSE for cancer bioinformatics
bq = pancan_BQ()          # need CGC_BILLING to authenticate
seCOAD = buildPancanSE(bq, acronym="COAD", assay="RNASeqv2")
seCOAD = bindMSI(seCOAD)  # update to include MSIsensor scores
par(mfrow=c(1,2))         # figure layout
amap = c("29126"="PD-L1", "925"="CD8A") # entrez:symbol mapping
bxs <- lapply( c("29126", "925"),       # for genes of interest
  function(x) boxplot(split(log2(as.numeric(assay( seCOAD[x,]))+1),
      seCOAD$msiTest >= 4), names = c("<4", ">=4"), ylab=amap[x],
      xlab="MSIsensor score")
  )
```

Our interest is in replicating part of Figure 5C of Bailey *et al.*[6]. In that paper, it is shown that microsatellite instability (MSI) is associated with different expression signatures of immune cell infiltration for adenocarcinomas of colon (COAD) and stomach (STAD), and uterine corpus endometrial carcinoma (UCEC). The MSI scores developed using MSIsensor are found in Table S5 of Ding *et al.*[7]. These scores are not available in BigQuery, but can be combined with the assay data using standard R programming, leading to a hybrid data/annotation strategy.

Functions in the *BiocOncoTK* package (Carey[8]) build on *restfulSE* functionality to a) authenticate the user to the BigQuery platform, b) select a tumor type (COAD) and assay for *SummarizedExperiment* construction, c) bind Ding *et al.*'s MSI values as sample-level data variable `msiTest`, d) acquire and transform the PD-L1 and CD8A (Entrez IDs 29126 and 925) expression values, and e) form the stratified boxplot. The basic findings of Bailey *et al.* are replicated. Enhancement of the code to produce a display covering more genes and tumor types is demonstrated

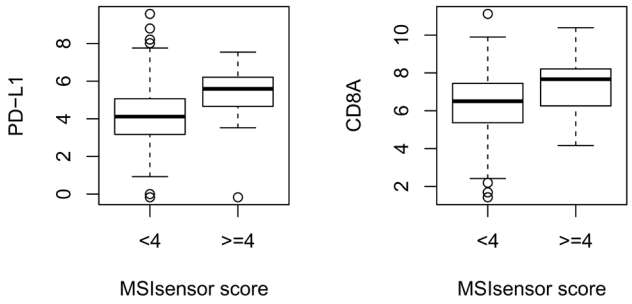

Figure 2. Association of MSI sensor scores with distributions of PDL-1 and CD8A in TCGA colorectal adenocarcinoma samples (COAD).

in the BiocOncoTK package vignette. Note that in this example, expression values are only downloaded for the genes requested, without altering the end user programming paradigm of working with a SummarizedExperiment instance.

HDF Scalable Data Service

Figure 3 demonstrates use of a RESTful `SummarizedExperiment`, with assay data provided in the object `/shared/bioconductor/darmgcls.h5` at `hsdshdflab.hdfgroup.org`. Briefly, as a prelude to single-cell RNA-sequencing of glioblastoma (GBM) tumors from four patients, Darmanis *et al.*[9] used immuno-panning to increase the proportion of non-neoplastic cells that constitute the "migrating front" of progression of glioblastoma. Antibody to CD45 was used to capture microglial cells. Figure 3 provides code to compare the distribution of CD45 expression among the classes of cells as labeled in the metadata of GSE84465, the NCBI GEO archive from which the quantifications were derived. In this example, data on one gene from all cells is retrieved when the statement defining vector `vals` is executed. The display can be recapitulated for other genes by substituting different symbols in the statement computing `ind`. The `DelayedArray` framework leveraged here enables basic computations of this kind without loading the entire matrix into memory.

```
library(rhdf5client)
library(SummarizedExperiment)
library(ggplot2)
cdar = BiocOncoTK::darmGBMcls
ind = match("PTPRC", rowData(cdar)$symbol)
var = gsub("selection: ", "",
       cdar$characteristics_ch1.8)
vals = log10(assay(cdar[ind,])+1)
ddd = data.frame(log10norm=vals, pan=var)
ggplot(ddd, aes(x=log10norm, colour=pan)) +
  geom_density() + ylim(0,1) +
  xlab("log10 CD45+1")
```

Performance

We focus on pursuit of reliability, expressivity, and scalability using *restfulSE*.

Reliability: The *restfulSE*, *rhdf5client* and *BiocOncoTK* packages are accompanied by detailed unit tests that compare retrievals to known values. In the case of BigQuery table queries, the test suite composes random queries in both BigQuery SQL and in the `SummarizedExperiment` idiom. Results are checked for elementwise equality.

Expressivity: The code segments for Figure 2 and Figure 3 are complex but easy to break down. The joining and reshaping of pancan-atlas tables in BigQuery corresponding to the code in Figure 2 can be checked through the query history in the BigQuery interface. The acquisition of expression values employed five nested SELECT statements; the query for assay quantifications was 6000 characters in length. The R code is less than 500 characters including comments.

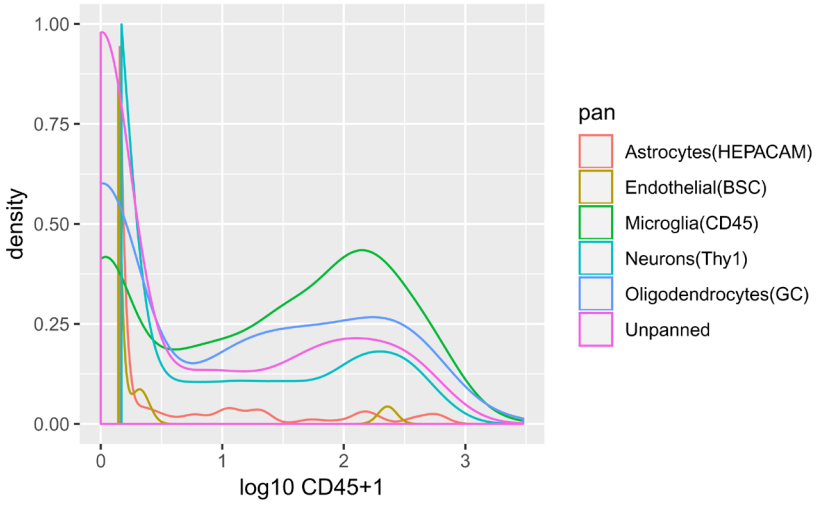

Figure 3. Density estimates for log10 CD45 expression in single-cell RNA-seq studies of glioblastoma.

Scalability. BigQuery is intrinsically auto-scaling, but charges accrue with the amount of data scanned, so query design can have effects on throughput and cost. We rely on the *bigrquery* (Wickham[10]) and *dbplyr* (Wickham and Ruiz[11]) packages for efficient translation of R-oriented data manipulations to BigQuery SQL. Throughput with the HDF Scalable Data Service is dependent upon the configuration of the object server, the relationship of numerical data layout to prevalent access patterns, and the degree to which queries capitalize on API efficiencies like chunk-based retrieval. For both back ends, proper design and deployment of the querying client can lead to throughput that scale with client-side resources.

Conclusions

Cloud-scale storage and retrieval strategies are of significant interest for genome science. The `SummarizedExperiment` class unifies assay data with substantive sample- and experiment-level metadata, and its API for managing and interrogating genome-scale experiment archives is used in numerous analytic packages. The *restfulSE* package exposes high-performance cloud-resident data stores to users and algorithms as `SummarizedExperiments`. Continued improvements in efficiency of representation and query resolution for assay data and metadata will help to achieve the potential of a federated data ecosystem for enhanced discovery in biology through interactive genome-scale analysis.

Software availability

restfulSE package available from: https://bioconductor.org/packages/3.9/restfulSE Source code available from: https://github.com/shwetagopaul92/restfulSE Archived source code as at time of publication: DOI: 10.18129/B9.bioc.restfulSE[12] License: Artistic-2.0

Grant information

Support for the development of this software was provided by NIH grants NCI U01 CA214846 (Carey, PI), NCI U24 CA180996 (Morgan, PI), and NHGRI 1U24HG010263-01 (J Taylor, PI), and Chan Zuckerberg Initiative DAF 2018-183436 (Carey, PI).

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

References

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

2. Lawrence M, Huber W, Pagès 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

3. ISB: **ISB Cancer Genomics Cloud 1.0.0 Documentation**. 2018; Accessed: 2018-08-17.
Reference Source

4. Hoadley KA, Yau C, Hinoue T, *et al.*: **Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer.** *Cell.* 2018; **173**(2): 291–304.e6.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Pagès H, Hickey P: **DelayedArray: Delayed operations on array-like objects**. R package version 0.7.28. 2018.

6. Bailey MH, Tokheim C, Porta-Pardo EP, *et al.*: **Comprehensive Characterization of Cancer Driver Genes and Mutations.** *Cell.* 2018; **173**(2): 371–385.e18.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Ding L, Bailey MH, Porta-Pardo E, *et al.*: **Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics.** *Cell.* 2018; **173**(2): 305–320.e10.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Carey V: **BiocOncoTK: Bioconductor components for general cancer genomics**. R package version 1.1.16. 2018.
Reference Source

9. Darmanis S, Sloan SA, Croote D, *et al.*: **Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma.** *Cell Rep.* 2017; **21**(5): 1399–1410.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Wickham H: **bigrquery: An Interface to Google's 'BigQuery' 'API'**. R package version 1.0.0. 2018.
Reference Source

11. Wickham H, Ruiz E: **dbplyr: A 'dplyr' Back End for Databases**. R package version 1.2.1. 2018.
Reference Source

12. Carey V, Gopaulakrishnan S: **restfulSE: Access matrix-like HDF5 server content or BigQuery content through a SummarizedExperiment interface**. R package version 1.4.0. 2018.
Publisher Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 25 February 2019

https://doi.org/10.5256/f1000research.19158.r42652

✔ **Sheila Reynolds**
Institute for Systems Biology, Seattle, WA, USA

The restfulSE interface described in this article by Gopaulakrishnan et al. is a very useful extension to the SummarizedExperiment class which provides a convenient approach to storing and manipulating rectangular matrices of experimental results, along with associated meta-data. This new extension allows users to query remote data, eliminating the common “download” step that still precedes many large-scale analyses.

As these datasets grow, and are more commonly made available in cloud-hosted technologies such as Google or AWS object stores or data warehouses such as Google BigQuery, tools that allow users to easily access and query these datasets become critical. The restfulSE interface permits targeted queries of such remote datasets.

As background information, the article includes a nice summary of the SummarizedExperiment class and related methods, for researchers (such as this reviewer) who had not come across this package before. The authors go on to describe two separate remote back ends: one which accesses PanCancer Atlas TCGA, hosted in Google BigQuery by the ISB-CGC; and the other which access HDF5 data hosted in AWS S3. Both of these backends further make use of the DelayedArray package, which implements delayed or block-processing operations to facilitate working with large datasets that cannot be stored in-memory. This enables “lazy” data retrieval, with numerical results transmitted from server to client only when needed.

The authors provide two concrete examples, illustrating the usage of both remote back ends. This reviewer ran into some issues trying to run these examples and reached out to the authors who provided additional information in video and Jupyter notebook form. Making additional tutorial resources available with this article will render this information useful and usable by a wider audience and is strongly encouraged.

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

Is the description of the software tool technically sound?
Yes

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

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?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology, cloud-computing, integrative analyses of heterogeneous and large-scale cancer data sets

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 04 February 2019

https://doi.org/10.5256/f1000research.19158.r43093

Dennis J. Hazelett
The Center for Bioinformatics and Functional Genomics, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA

The restfulSE software package for bioconductor purports to extend a very useful data structure, the SummarizedExperiment to handle very large datasets wherein dynamic download of the full dataset is neither necessary nor practical. Therefore, Gopaulakrishnan et al. have created restfulSE to make this data structure interactive with remote databases on an as-needed basis.

This is a very useful idea from the Bioconductor core team, and likely to be impactful as datasets grow larger, cheaper to produce, and it becomes increasingly necessary for bioinformaticians to leverage available data against local experiments.

The tool is technically sound, built on Google BigQuery and HDF5, and the paper is well written and clear. The manuscript includes code examples making it simple to get a quick start and see how the software works.

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

Is the description of the software tool technically sound?
Yes

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?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, regulatory genomics, cancer genomics and epigenomics

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.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research