SOFTWARE TOOL ARTICLE

HDCytoData: Collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats [version 1; peer review: 2 approved with reservations]

Lukas M. Weber[1,2], Charlotte Soneson[3,4]

[1]SIB Swiss Institute of Bioinformatics, Zurich, 8057, Switzerland
[2]Institute of Molecular Life Sciences, University of Zurich, Zurich, 8057, Switzerland
[3]SIB Swiss Institute of Bioinformatics, Basel, 4058, Switzerland
[4]Friedrich Miescher Institute for Biomedical Research, Basel, 4058, Switzerland

v1 **First published:** 19 Aug 2019, **8**:1459
https://doi.org/10.12688/f1000research.20210.1

Latest published: 04 Dec 2019, **8**:1459
https://doi.org/10.12688/f1000research.20210.2

Abstract

Benchmarking is a crucial step during computational analysis and method development. Recently, a number of new methods have been developed for analyzing high-dimensional cytometry data. However, it can be difficult for analysts and developers to find and access well-characterized benchmark datasets. Here, we present HDCytoData, a Bioconductor package providing streamlined access to several publicly available high-dimensional cytometry benchmark datasets. The package is designed to be extensible, allowing new datasets to be contributed by ourselves or other researchers in the future. Currently, the package includes a set of experimental and semi-simulated datasets, which have been used in our previous work to evaluate methods for clustering and differential analyses. Datasets are formatted into standard SummarizedExperiment and flowSet Bioconductor object formats, which include complete metadata within the objects. Access is provided through Bioconductor's ExperimentHub interface. The package is freely available from http://bioconductor.org/packages/HDCytoData.

Keywords

benchmarking, high-dimensional cytometry, Bioconductor, ExperimentHub, clustering, differential analyses

This article is included in the Bioconductor gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 04 Dec 2019	✓ view	✓ view
	↑	↑
version 1 19 Aug 2019	? view	? view

1. **Shila Ghazanfar**, Cancer Research UK Cambridge Institute (CRUK CI), Cambridge, UK

2. **Laurent Gatto**, University of Louvain (UCLouvain), Brussels, Belgium

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

This article is included in the RPackage gateway.

Corresponding authors: Lukas M. Weber (lukas.weber@imls.uzh.ch), Charlotte Soneson (charlottesoneson@gmail.com)

Author roles: Weber LM: Conceptualization, Data Curation, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Soneson C**: Conceptualization, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: LMW was supported by a Forschungskredit (Candoc) grant from the University of Zurich [FK-17-100].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Weber LM and Soneson C. **HDCytoData: Collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats [version 1; peer review: 2 approved with reservations]** F1000Research 2019, **8**:1459 https://doi.org/10.12688/f1000research.20210.1

First published: 19 Aug 2019, **8**:1459 https://doi.org/10.12688/f1000research.20210.1

Introduction

Benchmarking analyses are frequently used to evaluate and compare the performance of computational methods, for example by users interested in selecting a suitable method, or by developers to demonstrate performance improvements of a newly developed method. A critical part of any benchmark is the selection of appropriate benchmark datasets[1,2]. In some cases, suitable publicly available datasets may be found in the literature. Alternatively, new experimental or simulated datasets containing a known ground truth may be created by the authors of the benchmark[1,2].

High-dimensional cytometry refers to a set of recently developed technologies that enable measurement of expression levels of up to dozens of proteins in hundreds to thousands of cells per second, using targeted antibodies labeled with various types of reporter tags. This includes multi-color flow cytometry, mass cytometry (or CyTOF), and sequence-based cytometry (or genomic cytometry). Due to the large size and high dimensionality of the resulting data, numerous computational methods have been developed for analyzing these datasets[3]. Many of these methods are based on the fundamental concept of analyzing cells in terms of cell populations, for example using clustering to define cell populations, or detecting differential cell populations between conditions.

In our previous work, we have collected a number of benchmark datasets to evaluate methods for clustering[4] and differential analyses[5] in high-dimensional cytometry data. This includes publicly available datasets previously published by other groups or our experimental collaborators, as well as new semi-simulated datasets that we generated. In these previous publications, we recorded links to original data sources and made all data available via FlowRepository[6]. FlowRepository is a widely used resource in the cytometry community, which has also been used by other authors to distribute benchmark datasets (e.g., 7,8). However, downloading and loading the data from these sources for further analysis in R requires customized code and matching of metadata (e.g., sample information), which can hinder accessibility and reproducibility.

Here, we introduce the `HDCytoData` package, which provides a resource for re-distributing high-dimensional cytometry benchmark datasets through Bioconductor's *ExperimentHub*[9], in order to improve accessibility. *ExperimentHub* provides a flexible platform for hosting datasets in the form of R/Bioconductor objects, which can be directly loaded within an R session. `HDCytoData` provides datasets in the form of standard `SummarizedExperiment` and `flowSet` Bioconductor object formats[10–12], which include all required metadata within the objects and facilitate interoperability with R/Bioconductor-based workflows. We envisage that these datasets will be useful for future benchmarking studies, as well as other activities such as teaching, examples, and tutorials. The package is extensible, allowing new datasets to be contributed by ourselves or other researchers in the future. The package is freely available from http://bioconductor.org/packages/HDCytoData.

Methods

Implementation

The benchmark datasets currently included in the `HDCytoData` package consist of experimental and semi-simulated data, and can be grouped into datasets useful for benchmarking algorithms for (i) clustering and (ii) differential analyses. Table 1 and Table 2 provide an overview of the datasets.

The raw datasets were collected from various sources (Table 1 and Table 2), and have been extensively reformatted and documented for inclusion in the `HDCytoData` package. Each dataset is stored in both `SummarizedExperiment` and `flowSet` formats, since these are the most commonly used R/Bioconductor data structures for high-dimensional cytometry data. The objects each contain one or more tables of expression values, as well as all required metadata. Following standard conventions used for cytometry data[13], rows contain cells, and columns contain protein markers. Row metadata includes sample IDs, group IDs, patient IDs, reference cell population labels (where available), and labels identifying 'spiked in' cells (where available). Column metadata includes channel names, protein marker names, and protein marker classes (cell type or cell state). Note that raw expression values should be transformed prior to performing any downstream analyses. Standard transformations include the inverse hyperbolic sine (`asinh`) with `cofactor` parameter equal to 5 for mass cytometry or 150 for flow cytometry data (14, Supplementary Figure S2); several other alternatives also exist[15].

Most of these datasets include a known ground truth, enabling the calculation of statistical performance metrics. The ground truth information consists of reference cell population labels for the clustering datasets, and labels identifying computationally 'spiked in' cells for the differential analysis datasets. The datasets without a ground truth instead consist of experimental datasets that contain a known biological signal, which can be used to evaluate methods in qualitative terms; i.e., whether methods can reproduce the known biological result.

Table 1. Summary of benchmark datasets for evaluating clustering algorithms. For more details on these datasets, see Table 2 in [4], or the `HDCytoData` help files.

Dataset	ExperimentHub ID	Number of cells	Number of dimensions	Number of reference cell populations	Type of ground truth	FlowRepository ID	Original reference
Levine_32dim	EH2240 – EH2241	265,627	32	14	Manual gating	FR-FCM-ZZPH	16
Levine_13dim	EH2242 – EH2243	167,044	13	24	Manual gating	FR-FCM-ZZPH	16
Samusik_01	EH2244 – EH2245	86,864	39	24	Manual gating	FR-FCM-ZZPH	17
Samusik_all	EH2246 – EH2247	841,644	39	24	Manual gating	FR-FCM-ZZPH	17
Nilsson_rare	EH2248 – EH2249	44,140	13	1 (rare population)	Manual gating	FR-FCM-ZZPH	18
Mosmann_rare	EH2250 – EH2251	396,460	14	1 (rare population)	Manual gating	FR-FCM-ZZPH	19

Table 2. Summary of benchmark datasets for evaluating methods for differential analyses. For more details on these datasets, see Supplementary Note 1 in [5], or the `HDCytoData` help files.

Dataset	ExperimentHub ID	Type of data	Number of cells	Number of dimensions	Type of ground truth	Type of differential analysis	FlowRepository ID	Original reference
Krieg_Anti_PD_1	EH2252 – EH2253	Experimental	85,715	24 (cell type)	Qualitative	Differential abundance	FR-FCM-ZYL8	20
Bodenmiller_BCR_XL	EH2254 – EH2255	Experimental	172,791	24 (10 cell type; 14 cell state)	Qualitative	Differential states	FR-FCM-ZYL8	21
Weber_AML_sim	EH3025 – EH3046	Semi-simulated (multiple simulation scenarios)	157,593 (excluding spike-in)	16 (cell type)	Spike-in cell labels	Differential abundance	FR-FCM-ZYL8	5
Weber_BCR_XL_sim	EH3047 – EH3064	Semi-simulated (multiple simulation scenarios)	85,331 (main simulation; excluding spike-in)	24 (10 cell type; 14 cell state)	Spike-in cell labels	Differential states	FR-FCM-ZYL8	5

Extensive documentation is available via the help files for each dataset — including descriptions of the datasets, details on accessor functions required to access the expression tables and metadata, and links to original sources. In addition, reproducible R scripts demonstrating how the formatted `SummarizedExperiment` and `flowSet` objects were generated from the original raw data files are included within the source code of the package. New datasets may be contributed by ourselves or other authors by providing (i) formatted `SummarizedExperiment` and `flowSet` objects containing the data as well as all necessary metadata, (ii) reproducible R scripts showing how the formatted objects were generated from the original raw data files, and (iii) comprehensive documentation.

Operation

The `HDCytoData` package can be installed by following standard Bioconductor package installation procedures. All datasets listed in Table 1 and Table 2 are available in Bioconductor version 3.10 and above. Minimum system requirements include a recent version of R (3.6 or later; this paper was prepared using R version 3.6.1), on a Mac, Windows, or Linux system. Example installation code is shown below.

```
# install BiocManager
install.packages("BiocManager")
```

```
# install HDCytoData package
BiocManager::install("HDCytoData")
```

Once the HDCytoData package is installed, the datasets can be downloaded from *ExperimentHub* and loaded directly into an R session using only a few lines of R code. This can be done by either (i) referring to named functions for each dataset, or (ii) creating an *ExperimentHub* instance and referring to the dataset IDs. Example code for each option for one of the datasets is shown below. Note that each dataset is available in both SummarizedExperiment and flowSet formats. After an object has been downloaded, the *ExperimentHub* client stores it in a local cache for faster retrieval. For more details on accessing *ExperimentHub* resources, refer to the *ExperimentHub* vignette available from Bioconductor.

```
# load HDCytoData package
library(HDCytoData)

# option 1: load datasets using named functions
d_SE <- Bodenmiller_BCR_XL_SE()
d_flowSet <- Bodenmiller_BCR_XL_flowSet()

# option 2: load datasets by creating ExperimentHub instance
ehub <- ExperimentHub()
query(ehub, "HDCytoData")
d_SE <- ehub[["EH2254"]]
d_flowSet <- ehub[["EH2255"]]
```

Once the datasets have been downloaded and loaded, they are available to the user as R objects within the R session. They can then be inspected and manipulated using standard accessor and subsetting functions (for either the SummarizedExperiment or flowSet object class). Example code to inspect a SummarizedExperiment is displayed below. For more details on how to load and inspect datasets, including the expected output from each function shown here, refer to the HDCytoData vignette available from Bioconductor.

```
# inspect SummarizedExperiment object
d_SE
assays(d_SE)
rowData(d_SE)
colData(d_SE)
metadata(d_SE)
```

Documentation describing each dataset is available in the help files for the objects, which can be accessed using the standard R help interface, as shown below.

```
# display documentation (help files)
?Bodenmiller_BCR_XL
help(Bodenmiller_BCR_XL)
```

Use cases

The datasets currently included in the HDCytoData package (Table 1 and Table 2) can be used to benchmark methods for either (i) clustering or (ii) differential analyses. In addition, these datasets may be useful for other activities such as teaching, examples, and tutorials (e.g., demonstrating how to use a new computational tool).

For benchmarks using the clustering datasets (Table 1), performance can be evaluated by calculating metrics such as the mean F1 score or adjusted Rand index, which measure the similarity between two sets of cell labels (i.e., the cluster labels and the ground truth reference cell population labels)[1]. For examples (including reproducible R code), see the evaluations in our previous study[4]. An additional visual example is displayed in Figure 1, which compares the performance of three different dimensionality reduction algorithms (principal component analysis [PCA], t-distributed stochastic neighbor embedding [tSNE][22,23], and uniform manifold approximation and projection [UMAP][24,25]) in visually separating the known cell populations in the Levine_32dim dataset (see Table 1). R code to reproduce Figure 1 using data downloaded from the HDCytoData package is available at http://github.com/lmweber/HDCytoData-example.

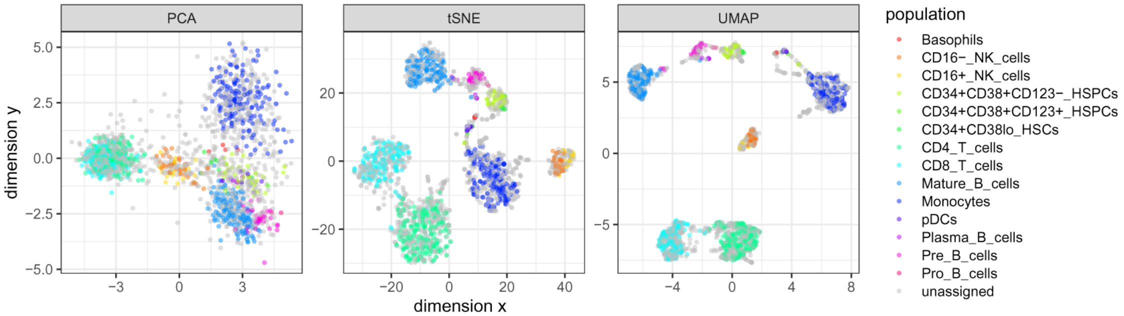

Figure 1. Example of use case for benchmark datasets in the HDCytoData package. This example compares performance (in visual terms) of three dimensionality reduction algorithms — principal component analysis (PCA), t-distributed stochastic neighbor embedding (tSNE), and uniform manifold approximation and projection (UMAP) — for representing known cell populations in the `Levine_32dim` dataset (Table 1).

For benchmarks using the differential analysis datasets (Table 2), methods can be evaluated by their ability to recover the known differential signals, either in quantitative terms using the ground truth spike-in cell labels (for the semi-simulated datasets), or in qualitative terms (for the experimental datasets). The differential signals consist of either differential abundance of cell populations, or differential states within cell populations (i.e., differential expression of additional functional markers within cell populations), providing conceptually distinct differential analysis tasks. For examples (including reproducible R code), see the evaluations in our previous study[5].

Summary

The `HDCytoData` package is an extensible resource providing streamlined access to a number of publicly available benchmark datasets used in our previous work on high-dimensional cytometry data analysis. Datasets are provided in standard Bioconductor object formats, and are hosted on Bioconductor's *ExperimentHub* platform. By facilitating access to these datasets, we hope they will be useful for other researchers interested in designing rigorous benchmarks for method development or other computational analyses, as well as other activities such as teaching, examples, and tutorials.

Data availability

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

Software available from: http://bioconductor.org/packages/HDCytoData

Source code available from: https://github.com/lmweber/HDCytoData

Archived source code at time of publication: https://doi.org/10.5281/zenodo.3362847[26]

Licence: MIT License

Grant information

LMW was supported by a Forschungskredit (Candoc) grant from the University of Zurich [FK-17-100].

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

Acknowledgments

The authors thank Mark D. Robinson (University of Zurich, Switzerland) for supervising the projects where these datasets were previously used for benchmarking, and feedback on the manuscript; and Lori Shepherd (Bioconductor Core Team and Roswell Park Cancer Institute, Buffalo, NY, USA) for assistance in making the datasets available through Bioconductor's *ExperimentHub*.

References

1. Weber LM, Saelens W, Cannoodt R, *et al.*: **Essential guidelines for computational method benchmarking.** *Genome Biol.* 2019; **20**(1): 125. PubMed Abstract | Publisher Full Text | Free Full Text

2. Mangul S, Martin LS, Hill BL, *et al.*: **Systematic benchmarking of omics computational tools.** *Nat Commun.* 2019; **10**(1): 1393. PubMed Abstract | Publisher Full Text | Free Full Text

3. Saeys Y, Van Gassen S, Lambrecht BN: **Computational flow cytometry: helping to make sense of high-dimensional immunology data.** *Nat Rev Immunol.* 2016; **16**(7): 449–462. PubMed Abstract | Publisher Full Text

4. Weber LM, Robinson MD: **Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data.** *Cytometry A.* 2016; **89**(12): 1084–1096. PubMed Abstract | Publisher Full Text

5. Weber LM, Nowicka M, Soneson C, *et al.*: **diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering.** *Commun Biol.* 2019; **2**: 183. PubMed Abstract | Publisher Full Text | Free Full Text

6. Spidlen J, Breuer K, Rosenberg C, *et al.*: **FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications.** *Cytometry A.* 2012; **81**(9): 727–731. PubMed Abstract | Publisher Full Text

7. Aghaeepour N, Finak G, The FlowCAP Consortium, *et al.*: **Critical assessment of automated flow cytometry data analysis techniques.** *Nat Methods.* 2013; **10**(3): 228–238. PubMed Abstract | Publisher Full Text | Free Full Text

8. Aghaeepour N, Chattopadhyay P, Chikina M, *et al.*: **A benchmark for evaluation of algorithms for identification of cellular correlates of clinical outcomes.** *Cytometry A.* 2016; **89**(1): 16–21. PubMed Abstract | Publisher Full Text | Free Full Text

9. Bioconductor Package Maintainer: **ExperimentHub: Client to access ExperimentHub resources.** R package, version 1.10.0. 2019. Publisher Full Text

10. 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 | Free Full Text

11. Morgan M, Obenchain V, Hester J, *et al.*: **SummarizedExperiment: SummarizedExperiment container**. R package, version 1.14.0, 2019.

12. Ellis B, Haaland P, Hahne F, *et al.*: **flowCore: Basic structures for flow cytometry data.** R package, version 1.50.0, 2019. Reference Source

13. Spidlen J, Moore W, Parks D, *et al.*: **Data File Standard for Flow Cytometry, version FCS 3.1.** *Cytometry A.* 2010; **77**(1): 97–100. PubMed Abstract | Publisher Full Text | Free Full Text

14. Bendall SC, Simonds EF, Qiu P, *et al.*: **Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum.** *Science.* 2011; **332**(6030): 687–696. PubMed Abstract | Publisher Full Text | Free Full Text

15. Finak G, Perez JM, Weng A, *et al.*: **Optimizing transformations for automated, high throughput analysis of flow cytometry data.** *BMC Bioinformatics.* 2010; **11**: 546. PubMed Abstract | Publisher Full Text | Free Full Text

16. Levine JH, Simonds EF, Bendall SC, *et al.*: **Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis.** *Cell.* 2015; **162**(1): 184–197. PubMed Abstract | Publisher Full Text | Free Full Text

17. Samusik N, Good Z, Spitzer MH, *et al.*: **Automated mapping of phenotype space with single-cell data.** *Nat Methods.* 2016; **13**(6): 493–496. PubMed Abstract | Publisher Full Text | Free Full Text

18. Rundberg Nilsson A, Bryder D, Pronk CJ: **Frequency determination of rare populations by flow cytometry: a hematopoietic stem cell perspective.** *Cytometry A.* 2013; **83**(8): 721–727. PubMed Abstract | Publisher Full Text

19. Mosmann TR, Naim I, Rebhahn J, *et al.*: **SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 2: biological evaluation.** *Cytometry A.* 2014; **85**(5): 422–433. PubMed Abstract | Publisher Full Text | Free Full Text

20. Krieg C, Nowicka M, Guglietta S, *et al.*: **High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy.** *Nat Med.* 2018; **24**(2): 144–153. PubMed Abstract | Publisher Full Text

21. Bodenmiller B, Zunder ER, Finck R, *et al.*: **Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators.** *Nat Biotechnol.* 2012; **30**(9): 858–867. PubMed Abstract | Publisher Full Text | Free Full Text

22. van der Maaten L, Hinton G: **Visualizing data using t-SNE.** *J Mach Learn Res.* 2008; **9**: 2579–2605. Reference Source

23. van der Maaten L: **Accelerating t-SNE using tree-based algorithms.** *J Mach Learn Res.* 2014; **15**: 3221–3245. Reference Source

24. McInnes L, Healy J, Melville J: **UMAP: Uniform manifold approximation and projection for dimension reduction.** *arXiv, 1802.03426 (v2).* 2018. Reference Source

25. Becht E, McInnes L, Healy J, *et al.*: **Dimensionality reduction for visualizing single-cell data using UMAP.** *Nat Biotechnol.* 2019; **37**(1): 38–44. PubMed Abstract | Publisher Full Text

26. Weber LM, Soneson C: **lmweber/HDCytoData: Version from paper (Weber and Soneson, 2019) (Version v1.5.12).** *Zenodo.* 2019. http://www.doi.org/10.5281/zenodo.3362847

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 02 September 2019

https://doi.org/10.5256/f1000research.22200.r52679

?

Laurent Gatto
De Duve Institute, University of Louvain (UCLouvain), Brussels, Belgium

Weber and Soneson present HDCytoData, a Bioconductor data package providing pre-formatted high-dimensional cytometry data. The preparation of the datasets as SummarizedExperiment and flowSet objects makes these amendable for benchmarking, a crucial step when developing new methods.

My main comment centres around the contribution of new data. While the curated/formatted data in the package have already been useful to the authors in their previous work, the ambition is to make it possible for others to benefit from them and, to enable this in the longer term, to expand the package with additional data. These contributions are anticipated to come from the original authors and, ideally, also by new contributors.

The contribution procedure, while crucial, (1) isn't described very clearly and, at least in its current form, (2) only applies to seasoned R users/programmers. These two points constitute a serious barrier to external contributions.

Indeed, the only information that is provided are a list of three required artefacts (objects, scripts and documentation), without details as to how to produce these, nor how to provide them. I would suggest to add a 'How to contribute' vignette to the package, describing all these aspects, including an example for one of the existing data. I would also suggest to include a contribution code of conduct, given that external contributions are explicitly advertised.

I would suggest asking new contributors to send a pull request (PR) on Github, with possible alternative methods for those that aren't familiar with GitHub. The use a PR provides traceability (as opposed to an email, for instance) and publicly recognises the external contribution, as PRs are publicly recorded on GitHub. I would also suggest to explicitly define how external contributions are to be acknowledged in the contribution guide (for example addition as a 'contributor' in the DESCRIPTION file).

These additions will clarify what is expected for a contribution to be considered, how it will be managed by the authors, and how it will be acknowledged, thus hopefully facilitating the process.

Minor suggestions:

- How can a potential user find out if/when new data have been added to the package? While `?HDCytoData` gives a list of dataset, a function returning a vector or dataframe with dataset names and possibly some annotation would be useful for programmatic access (given here that `data(package = "HDCytoData")` doesn't work for data on ExperimentHub).

- It could be useful to expand the 'Use cases' section with (1) example calculations of the F1 scores and Rand indices for the clustering example and (2) adding a similar short example for the differential analysis use case.

- I am curious as to why the content of the lmweber/HDCytoData-example isn't included as a vignette in the HDCytoData package (and thus lacking the usual control and documentation that comes with R packages).

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?

Partly

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, method development, research software engineering.

I confirm that I have read this submission and 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.

Author Response 26 Nov 2019

Lukas Weber

Thank you for your comments and suggestions. As suggested, we have provided significant additional material on the procedure for contributing new datasets. We have expanded the section in the text on external contributions, and added an additional Bioconductor vignette titled “Contribution guidelines”. This vignette describes the required files (data objects, scripts, documentation, metadata), as well as the submission procedure. We have requested that contributions be submitted via GitHub issues and pull requests, clarified the acknowledgment procedure, and added a code of conduct.

Regarding the minor suggestions, we have also (i) updated the main vignette and package help file to show how to programmatically retrieve a data frame of all available datasets, and (ii) added a new vignette titled “Examples and use cases”, which includes the example from the previous repository (https://github.com/lmweber/HDCytoData-example), as well as new examples showing how to use the datasets in the HDCytoData package to evaluate clustering performance (e.g. adjusted Rand index) and perform differential analyses.

Competing Interests: No competing interests were disclosed.

Reviewer Report 28 August 2019

https://doi.org/10.5256/f1000research.22200.r52681

?

Shila Ghazanfar

Cancer Research UK Cambridge Institute (CRUK CI), Cambridge, UK

Weber and Soneson have written a software article presenting HDCytoData (currently version 1.4.0), a Bioconductor package aimed at making multiple high-dimensional cytometry (HDC) datasets available in a consistent R friendly format as either SummarizedExperiment or flowSet objects. The authors have framed this as an aid for facilitating benchmarking studies and for use for examples or tutorials in future. HDCytoData provides links to these datasets through ExperimentHub and includes some helpful commands for downloading such data. Currently eight datasets are included in the package which are accessed with easy-to-use function calls in the workspace.

With this in mind, I have some comments & questions below that could improve the manuscript and useability of the HDCytoData package.

- This approach effectively duplicates the data from flowRepository and into the Bioconductor ExperimentHub ecosystem, is it more worthwhile to provide the functions to extract and process the data from the original flowRepository source? This manuscript could contain more motivation for hosting the processed data versus providing functions to download + process the data from flowRepository.

- Is it extensible to other additional characteristics? e.g. data arising from imaging mass cytometry with further measured features, or similar. The authors could discuss the breadth of experimental data types they imagine HDCytoData to encompass or accept from contributors.

- The authors should discuss the continued curation of the data within the HDCytoData package and mention how it behaves in case of changes or updates to the 'original' data in flowRepository. Describe further how one can contribute their dataset(s)?

- I'm confused as to the framing of this package as principally for benchmarking studies. Whilst it's an important aspect of understanding and improving methodology, users of this package may be more interested in a convenient and consistent way of loading the flow cytometry data altogether, especially so for some integrative analysis of multiple HDC datasets.

- It's unclear how large the data files are that are being downloaded into the local cache, ideally the user would want to know this information before going ahead and downloading it.

- It's not clear in this manuscript how one would remove the data once it's no longer needed, or how to clear the cache. It appears that it's assumed users are also fairly familiar with the ExperimentHub interface. The authors should make it more clear what level of experience they imagine package users should have, i.e. who are they aiming the software towards?

- It would be useful to have a bulk download to cache, or possibly a bulk load to workspace option for these datasets.

- Is there a functionality to switch from SummarizedExperiment object to flowSort format? If this exists in another package it should be pointed to.

- For the Bodenmiller data, I was surprised to find that the help file for Bodenmiller_BCR_XL_SE() says there are measurements for 24 proteins but the exprs assay has 35 columns, looking at the colData, features are classed into "type" and "state" with the remainder having a class of "none". Some of these columns do not appear to measure specifically protein abundance but rather cell-specific (as opposed to sample-specific) features, for instance "Cell_length". Have the authors anticipated this type of extra information and how it would fit into the SummarizedExperiment or flowSort object? The slot name "exprs" suggests that the data within this slot should be some molecular quantities, should these other features go into the rowData() slot instead? Or furthermore, whether ideally for downstream analysis (such as differential expression) these extra columns should be discarded? (Note these extra columns than what is listed in the function help descriptions appears for multiple datasets.)

- How do you ensure that there is enough information available here to be able to accurately normalise/standardise the data, especially so for flow cytometry data, given the particular combinations of fluorescent markers associated with the proteins, and potential overlap of the fluorescence for these markers?

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

Partly

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: statistics, high throughput genomics, transcriptomics, R software, high-dimensional data analysis

I confirm that I have read this submission and 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.

Author Response 26 Nov 2019

Lukas Weber

Thank you for your comments and suggestions. We have updated the text, vignettes, and help files to clarify each of the issues raised above. Below are also responses to the specific questions:

(1) Code to process the raw .fcs files from FlowRepository into the SummarizedExperiment and flowSet formats is provided in the ‘make-data’ scripts saved in the ‘inst/scripts’ directory in the source code of the HDCytoData package. Here we have followed the standard setup for ExperimentHub packages – i.e. processed data objects that are ready to load into R, together with reproducible scripts saved in ‘inst/scripts’ – as described in the ExperimentHub vignettes. We believe this is a useful setup for these datasets. FlowRepository is primarily intended as a permanent public repository for .fcs files associated with peer-reviewed publications, which cannot be updated. FlowRepository is also primarily accessed via the web interface, so it would be much less user-friendly to only provide scripts that re-format the .fcs files after downloading. Providing these datasets as pre-formatted SummarizedExperiment and flowSet objects makes them much more easily accessible for users.

(2) In principle, any data types that can be formatted into SummarizedExperiment and flowSet formats could be added to the package. However, we believe it makes sense to keep the scope of the package relatively limited, to facilitate modularity and maintainability. For now, we plan to include only the current set of technologies, although in the future it may make sense to develop similar packages for other data types (e.g. imaging mass cytometry) (see Summary).

(3) According to the policies of FlowRepository, original .fcs files stored in FlowRepository cannot be updated after publication of the associated peer-reviewed paper. Similarly, data objects stored in ExperimentHub can only be updated manually by contacting the ExperimentHub maintainers. Therefore, we do not expect any major updates to the datasets currently stored in the HDCytoData package (except possibly minor bug fixes). We have included some additional text explaining this. We have also included a new vignette on “Contribution guidelines” (see comments for Reviewer 2).

(4) While users could indeed use the HDCytoData package to load these datasets in a consistent way for other purposes, we believe the main use cases for these particular datasets are for benchmarking and teaching / examples / tutorials. These datasets are well-characterized and have been studied in a number of previous publications, making them ideal for benchmarking. Formatting the datasets into consistent SummarizedExperiment and flowSet formats requires significant effort, so we expect this will mainly be worthwhile for datasets that can be re-used a number of times, e.g. for benchmarking.

(5-7) We have updated the text, main vignette, and help files to mention the size of the data files. The datasets range in size from 2.4 MB to 194.5 MB. We have also explained how to clear the local download cache, and updated the text to mention the expected level of experience with Bioconductor. We are not aware of a bulk download option in the ExperimentHub interface, so we have not included this. (If this functionality were added in the future, we believe it would better belong in the ExperimentHub package than in HDCytoData.)

(8) There is no simple way to convert between the SummarizedExperiment and flowSet formats. This is one of the major contributions of this package – we have pre-processed the datasets into both of these formats (with reproducible code saved in the ‘inst/scripts’ directory), so that users do not need to do this manually. We have included additional text to mention this.

(9) The additional columns of raw data (which are labeled as “none” in the “marker_class” column) contain additional information from the raw .fcs files from the mass cytometry machine, including barcodes for sample deconvolution, and event length and DNA content to identify live single cells. These columns are usually stored in the expression matrices in the original raw .fcs files, so we have also left them in the objects, e.g. for users who wish to check the pre-processing steps. We labeled these columns as “marker_class = none” to make them easier to identify, especially for users who are not already familiar with mass cytometry data. We have updated the help files to clarify that these columns are not needed for downstream analyses.

(10) Compensation for fluorescence spillover has already been performed by the original authors of the flow cytometry datasets, so users of these datasets do not need to perform this. However, users still need to apply a transformation (e.g. arcsinh), which we have described in the vignettes and help files.

Competing Interests: No competing interests were disclosed.