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. Weber1,2, Charlotte Soneson3,4

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

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.

This article is included in the RPackage gateway.
Corresponding authors: Lukas M. Weber (lukas.weber@imls.uzh.ch), Charlotte Soneson (charlotte.soneson@fmi.ch)

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.

Copyright: © 2019 Weber LM and Soneson C. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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\(^1\). 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\(^2\).

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\(^1\), 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.
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.

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

---

### 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.

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

### 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ExperimentHub ID</th>
<th>Type of data</th>
<th>Number of cells</th>
<th>Number of dimensions</th>
<th>Type of ground truth</th>
<th>Type of differential analysis</th>
<th>FlowRepository ID</th>
<th>Original reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krieg_Anti_PD_1</td>
<td>EH2252 – EH2253</td>
<td>Experimental</td>
<td>85,715</td>
<td>24 (cell type)</td>
<td>Qualitative</td>
<td>Differential abundance</td>
<td>FR-FCM-ZYL8</td>
<td>20</td>
</tr>
<tr>
<td>Bodenmiller_BCR_XL</td>
<td>EH2254 – EH2255</td>
<td>Experimental</td>
<td>172,791</td>
<td>24 (10 cell type; 14 cell state)</td>
<td>Qualitative</td>
<td>Differential states</td>
<td>FR-FCM-ZYL8</td>
<td>21</td>
</tr>
<tr>
<td>Weber_AML_sim</td>
<td>EH3025 – EH3046</td>
<td>Semi-simulated (multiple simulation scenarios)</td>
<td>157,593 (excluding spike-in)</td>
<td>16 (cell type)</td>
<td>Spike-in cell labels</td>
<td>Differential abundance</td>
<td>FR-FCM-ZYL8</td>
<td>5</td>
</tr>
<tr>
<td>Weber_BCR_XL_sim</td>
<td>EH3047 – EH3064</td>
<td>Semi-simulated (multiple simulation scenarios)</td>
<td>85,331 (main simulation; excluding spike-in)</td>
<td>24 (10 cell type; 14 cell state)</td>
<td>Spike-in cell labels</td>
<td>Differential states</td>
<td>FR-FCM-ZYL8</td>
<td>5</td>
</tr>
</tbody>
</table>
# 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). For examples (including reproducible R code), see the evaluations in our previous study. 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], and uniform manifold approximation and projection [UMAP]) 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.
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. Data sets 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
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


Open Peer Review

Current Peer Review Status:  ?  ?

Version 1

Reviewer Report 02 September 2019

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

© 2019 Gatto L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 Imweber/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.

Reviewer Report 28 August 2019

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