Bioc Technical Advisory Board Minutes

2 February 2023

Attending: Vince Carey, Lori Kern, Laurent Gatto, Charlotte Soneson, Marcel Ramos, Michael Lawrence, Jayaram Kancherla, Robert Shear, Alexandru Mahmoud, Hervé Pagès, Wolfgang Huber, Levi Waldron, Sean Davis, Rafael Irizarry, Jennifer Wokaty, Aedin Culhane, Robert Gentleman, Kasper Hansen
Regrets: Stephanie Hicks, Davide Risso, Shila Ghazanfar, Mike Love

:03 - :04 January minutes approved.

:04 - :05 Good progress on changing the default branch name from master to devel.

:05 - :25 Jayaram Kancherla: BiocPy

- Enabling Bioconductor workflows in python.
- Motivation: increasing use of python in bioinformatics.
  - AnnData, scanpy, scverse.
  - napari and other efforts.
  - AI/ML - keras, PyTorch.
  - backend systems.
- Previous isolated efforts to bring Bioc representations to python - fragmented user/developer experience.
  - pyRanges (GRanges), AnnData (SE), MuData (MAE).
- Goals:
  - Develop Bioconductor representations in python.
    - Interval-based operations.
    - Rely on numpy, scipy, pandas etc.
    - Don't aim to be exhaustive, implement as needed.
  - Usability/easy transition for polyglot analysts/scientists.
    - Similar user interface as in R.
  - Standardize package management in python.
  - Foster community efforts to build analysis methods, visualization packages etc.
- What's in BiocPy today:
  - Core packages: GenomicRanges, SummarizedExperiment, SingleCellExperiment, MultiAssayExperiment, BiocFrame, SpatialExperiment (WIP).
  - Utility packages: rds2py, mopsy (interface similar to base R matrix methods/MatrixStats), pyBiocFileCache.
- Get started: `pip install biocpy` (wrapper), or install packages separately.
- https://biocpy.github.io/GenomicRanges/tutorial.html
- Properties provide an interface to instance attributes (consistent in all the biocpy classes). Getting and setting attributes are in-place operations.

- Most methods work on shallow copies, not in-place operations (unless specified).
- Assays - use numpy or scipy sparse representations for matrices.
- colData - Pandas DataFrame.
- https://biocpy.github.io/mopsy
- Use downstream python packages to visualize data.
- Challenges with python development process:
    - No standard way of writing packages (Pyscaffold).
    - Doesn't enforce testing, documentation.
    - Build/tests are done locally, artifacts uploaded to PyPI.
    - Unlike vignettes, documentation snippets are not executed (solution: all snippets used in docs are also part of tests).
    - Tricky if you need to interface with underlying C/C++ libraries -> rds2py.
- Future work:
    - Import from common genomic file formats (like rtracklayer).
    - Visualization interfaces (genome tracks, ComplexHeatmap-like).
    - Language-agnostic serialization, for interoperability with ArtifactDB (h5-based, protobufs?).
    - Delayed operations/lazy evaluation.
    - Interfaces to frequently used analysis methods.
    - Outreach and community feedback.
- https://github.com/BiocPy/
- https://github.com/LTLA/bioconductor.js/ (JavaScript version)
- Q: Lots of things in IRanges are implemented in C (e.g. findOverlaps) - can you reuse that code in the python package, or did you reimplement it? A: Writing interfaces to underlying C/C++ code is the easy part, the tricky thing is the build process (different architectures etc). Reimplemented the interval operations in GenomicRanges. In principle, C code from the R implementation could be reused.
- Q: How many users currently/are they happy? A: Hasn't been advertised much yet, but is used internally (e.g. ArtifactDB).
- Q: How to collaborate/interoperate with Bioconductor? A: Would be good with an active collaboration.

:25 - :35 Laurent Gatto: MsDataHub

- ExperimentHub package for sharing MS data.
- In the vignette, there are code chunks showing what you would type to get the data, but they are not executed (would download large data files - slow and requires lots of disk space). What should we advise developers? Systematically download all the data (which may be large)? Would lead to large caches on the build system, but makes it easier to find bugs.
- In general, "executing everything" improves robustness, but it may not be sustainable.
- Limited example of download functionality to be shown in the vignette. Do as much as is necessary to show that you have a rational approach to getting the data.
- Put data on Zenodo, use ExperimentHub to define metadata.

:35 - :40 Working group invitation for cloud service concepts and GitHub Actions.

- Several people are working actively on GitHub Actions in Bioc environments - have been invited to form a working group on cloud service concepts and GHA (Leo, Mike, Brian, Erdal, Jen, Alex).

:40 - :50 Machine Learning: how to distribute large fitted model objects.

- E.g. keras models can be saved in hdf5 format and read back into R with keras::load_model_hdf5(). The same model should in principle work also in python.
- At the moment, Bioconductor doesn't have a suitable DispatchClass for distributing such models via ExperimentHub.
- Resources like kipoi (http://kipoi.org/), Sfaira (https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02452-6) and huggingface (https://huggingface.co/models) provide means for developers to distribute models.
- kipoi is accessible from R using reticulate (http://kipoi.org/docs/using/R/), submissions are highly structured (see e.g. http://kipoi.org/docs/contributing/02_Writing_model.yaml/). Last addition to https://github.com/kipoi/models was 6 months ago. Model weights are stored on Zenodo or FigShare. Most models so far seem to take a sequence as input.
- huggingface is more flexible, but may be difficult to provide a general interface to access models. Not directly biology-related (computer vision, NLP, audio, etc).
- Strategy for being more active in machine learning in genomics.
- Container image from rocker - rocker/ml. In progress: bioc/ml image that contains some tools for linking Bioc to the ML ecosystem, connect to GPUs etc. Ready for testing.
- Important problem, but don't overengineer before we have more experience.
- Create a section for ML models to make them more exposed to people working in the area?
- Working group, Slack channel.

:50 - :60 Other business

- TAB/CAB joint meeting in March.
- Bioconductor mastodon account created.
- CAB elections next year aligned with the BioC conference.
- Submitted proposal for ELIXIR All Hands meeting workshop in June.
- Bioconductor was again selected by Outreachy for internships - search more broadly for mentors.
- Tutorial was accepted at ISMB (Davide).
- EuroBioC2023 will take place in Ghent in September.