Bioc Technical Advisory Board Minutes

5 October 2023

Present: Vince Carey, Charlotte Soneson, Lori Kern, Levi Waldron, Helena Crowell, Marcel Ramos, Alexandru Mahmoud, Ludwig Geistlinger, Robert Shear, Nikhil Mane, Hervé Pagès, Jen Wokaty, Stephanie Hicks, Laurent Gatto, Kasper Hansen, Aedin Culhane, Maria Doyle, Sean Davis, Brian Schilder
Apologies: Wolfgang Huber, Michael Love, Rafael Irizarry, Davide Risso, Henrik Bengtsson

Welcome

- Previous minutes approved

Some older business reviewed

- Packages with no accessible maintainer.
    - Bioc Orphaned Package Policy
- Mentoring process.
    - Engage with the mentorship program - there were resources created by the mentors/mentees of the first round.
    - Discussion of S4 classes might fit best as a chapter in some kind of online living book / resource where "best practices for Bioconductor" are discussed.
- Books.
    - We are going to have to provide some guidance on submitting books to Bioconductor. This could be heavy and involve a working group and specs, or light with a single books repository that primarily links to content that authors stand up and stand behind. Suggestion that books should be static entities released by editions fixed to Bioc release versions.
- TAB Meetings: breakout rooms?
- Rsamtools in limbo (CRAM, Rhtslib feature match to htslib?, CSI indexing…)
- Hacktoberfest? Articulate a project on Rsamtools?

Funding

- A supplement was awarded. M1 Mac Pro ordered to assist with building of mac arm binaries.

EOSS planning

- Coordination among people planning to submit LOIs.
- scverse, interaction with python community (leverage other NumFocus projects).
- Carpentries, website.
- Conferences (NumFocus).
- Community manager.
- Strategic planning for Bioconductor - Bioconductor is entitled to a certain amount of support from CZI (as a previous grantee).

- Build on the developer mentorship program.

Submissions

- We have some outstanding submissions for 3.18 but it is not clear how to fairly highlight them. Any members interested in reviewing a portfolio of vignettes? Maybe the chatbot could summarize usefully?
- Important that such highlights come from a named source, and with a motivation.

Security profile concerns

- Related to libwebp.

New website progress

- Important design/evolution question: how to introduce modern search facilities? AI?
- New and old websites currently running in parallel (there are links from one to the other).
- Comments and feedback welcome (#biocwebsite slack channel).

Need spearheading of developer forum

Conference abstracts

- CSHL Genome Informatics: Towards a cloud-agnostic scalable ecosystem for open genomic data science with Bioconductor and Galaxy (Alex lead), Genome representation in Bioconductor (Hervé lead).
- Let's get into the habit of preparing and submitting to useful conferences.
- CSHL Biological Data Sciences may be good to consider (only every 2 years).

Developer status reporting tools

- https://github.com/Bioconductor/BiocPkgDash creates a dashboard of packages for a given maintainer email.
- Install and open *app.R* in base directory and click on *Run App* at the top left in RStudio. If running R from the terminal, call shiny::runApp().

Build system

- What do we require? (Jen).
- Input from other projects: Levi's encounters at NumFocus
- The container at ghcr.io/almahmoud/bioconductor_salt:jammy-r-4.3.1 duplicates the Linux BBS runtime environment.
- Slides
- Discussion
    - Any caching from day to day? Dependencies are not all installed from scratch, packages are obtained anew from git.bioconductor.org
    - Is the DAG aware of changes in dependencies? Moving to a more distributed system requires awareness of what has been changed and what has not. No automation involved in installation of system dependencies at the moment. All Bioc packages are built and checked every night, regardless of whether they

were changed or not. Could adapt the code to allow a more 'incremental' mode of running the build system. The dependency graph is built every day, the build system figures out what is required. CRAN package dependencies are not reinstalled if they are not changed.
    - Dagster useful for orchestrating large DAGs. Offers quite a bit of observability, and 'code locations' that can be hosted on their own hardware environments. Captures logs, manages dependencies between tasks, etc.
    - https://github.com/Bioconductor/BBS has information about dependencies in the form of a document for each OS in the Doc path. In the Ubuntu-files path, apt and pip dependencies are listed.

Debian binary packages for Bioconductor software

- Ubuntu binaries for Bioconductor packages.
- Eliminates need for containers when using Ubuntu (focal or jammy).
- https://github.com/bioconductor/bioc2u
- Use cases: binder (can't use the container binaries because of missing dependencies), can start from a smaller container than the 'big' Bioc container that has all the system dependencies.
- Similarities to the approach used by conda. In principle, could we have a conda channel for Bioc (repackaging the binaries)? Leverage what's available e.g. at conda-forge build system.
- Alex did some benchmarking to showcase the utility of bioc2u. Picked `cbpManager` (i.e. `system.time(BiocManager::install("cbpManager", update=TRUE, ask=FALSE))`) somewhat randomly, just looking through packages for one with close to 100 dependencies, and compared time of using r2u/bioc2u vs our containers and binary repo vs source.
- Logs in this gist: https://gist.github.com/almahmoud/df31672452448114918 2ae5ebaf04837
- Summary:
    - Source, using BBS salt container: `docker run --rm ghcr.io/almahmoud/bioconductor_salt:jammy-r-4.3.1 -e 'system.time(BiocManager::install("cbpManager", update=TRUE, ask=FALSE))'`

```
   user  system elapsed
174.478  12.474 191.327
```

    - Container binaries `docker run --rm -it bioconductor/bioconductor:3.17 Rscript -e 'system.time(BiocManager::install("cbpManager", update=TRUE, ask=FALSE))'`

```
   user  system elapsed
 17.008   6.321  31.459
```

    - New Bioc2u user container which is plain ubuntu plus apt repository, R, and BiocManager `docker run --rm -it ghcr.io/bioconductor/bioc2u-user:jammy-r-4.3.1 Rscript -e 'system.time(BiocManager::install("cbpManager", update=TRUE, ask=FALSE))'`

```
   user  system elapsed
```

```
          11.442   2.432  18.637
```

- Note that it's running on a Jetstream VM and bioc2u is hosted on Jetstream, so faster to pull from JS2 than GCP for container binaries.

- If I am not mistaken the awareness of this approach arose from a developer forum meeting with Dirk E. presenting. We need that forum to proceed regularly. Maybe include an honorarium component of an EOSS submission?

Python interoperability

- Aedin - interactions with Isaac Virshup from scverse. Bioconda Bioconductor packages are not always up to date. Discuss potential collaborations.
- Genentech: ArtifactDb, CollaboratorDb (early self-deployable demo). From Aaron:

```
import collaboratordb as cdb
obj = cdb.fetch_object("scRNAseq:ZeiselBrainData@2023-08-08")
## Class SummarizedExperiment with 20006 features and 3005 samples
##   assays: ['counts']
##   row_data: ['featureType']
##   col_data: ['tissue', 'group #', 'total mRNA mol', 'well', 'sex', 'age',
'diameter', 'cell_id', 'level1class', 'level2class']
```

- Long-range game plan: When desired, Bioconductor contributors apply alabaster converters to artifacts, generating "language agnostic" (e.g., FASTA, HDF5, JSON) components and metadata that can go into a performant, indexed, versioned store. This "hub" can be queried from any language to get genomic data artifacts that can be as self-describing as they are in Bioc.

Support for scientific innovation

- Is infrastructure at Bioconductor too cumbersome? Adhering to the requirements may feel like a burden rather than an opportunity (may be a matter of documentation).
- Could be a good opportunity for supplemental/additional funding (for developer support).
- Publication explaining the BBS, why to submit etc would be useful. Also explain the review process.
- Should development start from the general, or from the specific? Define classes early, or do more specific development first and see what is shared/general?

Other

- Sean has been experimenting with building LLM agents for specific purposes. A Bioconductor chatbot could do things such as 1) recommend a package, 2) show how to do something with Bioconductor, 3) find data resources that Bioconductor maintains, 4) find community resources, 5) find people to collaborate with.
- Vince reached out to CRAN to ask when Quarto will be supported for vignettes - not clear yet. The quarto package does not support Quarto vignettes, but that's easy to add. Indeed, there is a PR (https://github.com/quarto-dev/quarto-r/pull/57); I just verified it with a standalone package

(https://github.com/quarto-dev/quarto-r/pull/57#issuecomment-1751093594). There's also a discussion at https://github.com/quarto-dev/quarto-cli/discussions/2307.