Bioc Technical Advisory Board Minutes

1 February 2024

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

:03 - :05 Welcome and topics from previous meetings

- Previous meeting minutes approved.
- bibliometrics – any interest in a working group? (Kasper is interested).

:05 - :19 Conference

- Nominations for Bioc Awards
    - Can't have diversity without enough nominations!
    - Be mindful that diversity is not only important for the awards, but for the project as a whole.
    - Awards can be a good entry point to encourage contributors to become more deeply involved.
    - How can we improve our understanding of the composition of the community as a whole?
    - Consider how to increase the reach to other communities (e.g. via the conferences).
    - https://bioconductor.org/about/awards/
- Long workshops at BioC2024 will be "invite-only" (not solicited via the regular submission system). Does the TAB have any insight/opinions on what topics would be good to cover?
    - Intro-type workshops, annotation, using public data, GRanges ('Fundamentals of Bioconductor').
    - Spatial and single-cell.
    - Working across packages in a domain.

:19 - :20 Upcoming release, contact R core for release date of R 4.4.0.

:20 - :21 Voting on governance: outcome

- 12 votes: 8 Yes, 2 No, 2 Abstain -> new governance document accepted and will be uploaded to the website.

:21 - :31 Workflows vs books (Davide)

- Does Bioconductor still accept workflow submissions (https://www.bioconductor.org/packages/release/BiocViews.html#___Workflow)? Yes! Normally receive 1-2 per year.

- Workflows are featured prominently under the "Learn" tab of the new website.
- Journals in general consider publication as a single, static artifact.
- More recently, books have become increasingly common (e.g. OSCA).
- Workflows are (compared to books) less monolithic, easier to maintain and to discover for users interested in a specific task.
- Consider the success of Seurat's "frequently asked vignettes" - could we leverage the website redesign to provide more prominent workflows?
- As an example, Davide could think of a “best practices” vignette on working with very large data on Bioconductor. Part of these best practices are covered in Chapter 14 of the advanced OSCA book, but it would be nice to have it as a standalone resource.
- A similar workflow could be a version of what Marcel, Ludwig, and Davide taught at ISMB last year (https://bioconductor.github.io/ISMB.OSCA). Another one is Pete Hickey's excellent DelayedArray tutorial (https://petehaitch.github.io/BioC2021_DelayedArray_workshop/).
- We need more curated datasets, especially for teaching and for new domains (it's a hurdle to put together the curated dataset for a workflow) - could some of this be centrally coordinated?
- Need a range of different types of data sets - some simple, some more advanced. It would be helpful to also mention what to look out for in the data set (e.g., 'this is a data set with a lot of small differences between the groups').
- Seurat workflows are task-oriented (and has the advantage that everything is organized within a single framework).
- We (as a project) should not be picking 'winners'.
- Could we have a template workflow package available?
- Several curated data sets exist for single-cell and spatial domains.
- In which fields are we missing good curated data sets?
- Examples of useful data packages: SingleCellMultiModal, cBioPortalData, curatedTCGAData, curatedMetagenomicData, airway, fission, parathyroidSE, oct4, macrophage.
- Small data sets from GTEx and new ENCODE releases would be useful.
- Finding an appropriate data set can be quite difficult in our current search (experiment data packages are very diverse).

:31 - :39 Misc, quick topics

- New website: metrics on use/visit patterns – google analytics?
- GPU computing – we have a host at DFCI. What do we need to do for developers/users in this area? A100. IMHO (VJC) we should test packages that use tensorflow, keras, torch, and provide guidance on how to support validation, how to verify cuda infrastructure, etc.
- Long reads
    - Biostrings buffer size increased to 200k. https://github.com/GoekeLab/sg-nex-data has lots of tutorial content, the NanoporeRNASeq package vignette is more terse.

        - slack channel (#longread) and biocViews available: https://www.bioconductor.org/packages/release/BiocViews.html#___LongRead
- OSCA maintenance
    - Currently all but one part is building.
    - Are there elements that are substantively obsolete?
- quarto
    - Vignette engine moving forward.
    - quarto evolution
    - BiocStyle needs changes to be used with quarto.
- basilisk maintenance
    - Repo has been transferred from LTLA to the Bioconductor organization.
    - Steering committee concept.
    - How to make sure all platforms are viable?
        - Windows is challenging.
    - How to take advantage of new developments like mamba (now default solver in conda).
    - Guidelines on best practices for python-dependent packages in Bioc.
    - Testing is excessive in devel and causes timeouts, tests will be modified.
    - Clients are a useful extra testing resource.

:39 - :46 wasm for R/Bioc in the browser

- Beyond proof of concept - what will users actually do?
    - Interactive example(), quick-access demos in teaching
- shinylive example by Charlotte: https://github.com/csoneson/shinylive_test
- Quarto example from Wolfgang: https://www.huber.embl.de/group/posts/horseshoes-embeddings.html
- Prospects for all Bioc packages in r-wasm?
    - Matt McCormack
    - Alex will try to repurpose the existing dispatcher (used to build container binaries) to take advantage of the GitHub Action provided to build wasm binaries.
    - Building binaries: https://docs.r-wasm.org/webr/latest/building.html.
- HB: E.g. system() calls, sockets not supported (which affects parallelization).
- webR binary package repository: https://repo.r-wasm.org/
- r-universe builds wasm binaries for hosted packages: https://ropensci.org/blog/2023/11/17/runiverse-wasm/
- Packages are built, but currently not checked.
- #wasm slack channel created for interested parties.

:47 - :60 Bioconductor build system

- Recent work by Sean in harvesting longitudinal record of build outcomes.
    - https://seandavi.github.io/BiocBuildDB/ - based on tgz files from the build system. Build report (increment daily). Log files are included in the data frames that are created as well.

    - Are the report files (for a package) small enough to be version controlled? tgz files are between 5 and 60MB. 6-7 million lines in complete build report history over the captured time span. Size is not a limitation for keeping the historical data going forward.
- Discussion of cloud working group
    - Ephemeral infrastructure goal: in case on-premises compute for BBS becomes infeasible (scale, security, administrative constraints), we may need to be cloud-native.
- gitolite replacement concept is relevant.
- Recent work on GitHub Actions to run checks and builds is promising.
- Build report for M1 mac is provided separately from the others - should be brought together.
- There is a fair amount of packages for which mac arm binaries are not available - what can we do to get more robust availability of these binaries? M1 build machine performs a bit differently from the Intel machine in some aspects - configuration is ongoing.