Bioc Technical Advisory Board Minutes

5 December 2024

Present: Vince Carey, Hervé Pagès, Stephanie Hicks, Helena Crowell, Lori Kern, Henrik Bengtsson, Charlotte Soneson, Ludwig Geistlinger, Rafael Irizarry, Wolfgang Huber, Davide Risso, Kasper Hansen, Alexandru Mahmoud, Michael Lawrence, Levi Waldron, Marcel Ramos, Jacques Serizay
Apologies: Laurent Gatto, Erdal Cosgun

:03 - :04 Previous minutes approved.

:04 - :09 Discussion of working group proposals from Tim Triche

- Pangenome working group proposal
    - The Bioconductor project already possesses infrastructure for operating on pangenomes and genome graphs, which are rapidly proving to be the only viable means of reducing reference bias for structural polymorphisms across diverse ancestries. The PanVizGenerator package (somewhat unfortunately unrelated to the released PanViz package) was implemented by Thomas Petersen to facilitate genome and pangenome graph visualization using the panViz javascript library. Tools such as sequence tube maps (cf. https://cpang.netlify.app/post/testing-ivg/) are excellent visual aids but more relevant in the context of (for example) igv.js, which is spawned for Bioc users by tools such as Paul Shannon's igvR. There is an urgent need for interested parties to revive and improve infrastructure for exploratory data analysis on graphs (rather than e.g. linear assemblies with alt contigs), especially when mapping regions such as the polymorphic MHC and KIR regions, but also for structurally variable regions on the X chromosome across populations. This need will only increase as the IGVF and SMAHT consortia generate functional and somatic maps of variation (resp). At least 4 active Bioc developers from 3 separate institutions have expressed interest in this specific direction and we expect more.
- WASM working group proposal
    - The Bioconductor project and its associated Carpentries educational material suffers from mild code rot. The succession of foundational Bioc S4 classes (GenomicRanges, SummarizedExperiment) and inherited versions thereof provides an idiosyncratic view of even relatively stock tasks such as bulk RNAseq analysis. There is little incentive for researchers to modify or modernize these code bases other than to pass build checks. An interesting opportunity to both expand the Bioc community and modernize lightweight data structures, suitable for data interchange with (e.g.) python, is migration of documentation to statically compiled WebAssembly (WASM) code which users can execute, modify, and evaluate in their browser. Alexandru Mahmoud's efforts to create a Bioconductor .wasm repository have provided an enormous head start relative to

(for example) pyodide package creation, and it is in the Bioconductor community's own best interest to explore it further. While a 4GB-8GB memory limit per browser process enforces certain limitations, Carpentries instructors and students have reacted positively to the instant interaction and learning-by-doing aspect of webR-enabled documentation. Further value is provided by the static compilation of specific software and package versions to .wasm binaries, allowing a degree of reproducible analysis difficult to achieve otherwise (yet relatively straightforward in this context). Paradoxically, the memory limitations of this venue encourage lightweight packages that minimize dependencies, an additional net positive for the Bioc project. The participation of a core Bioconductor/Galaxy member in webR support for Bioconductor, and interest from Carpentries lesson maintainers in a low-setup direction for the (incubator-stage) Bioconductor Carpentries curriculum, as well as CZI support for both, speaks in favor.

- The TAB is supportive of the formation of these working groups.
- Working group chairs will be identified and will be invited to TAB calls for the lifetime of the working group.

:09 - :11 Ad hoc comments on possible industry consortium and grant support.

- CZI - no EOSS call next year.
- Multi-PI grants based in Europe, with some subcontracts to other parts of the world.

:11 - :15 Updates from CAB.

- Kevin and Johannes stepped down as co-chair and secretary. Stevie Pederson and Lori Kern to replace (ran unopposed).
- Updates from the Workflow/Publication working group (rebranding of workflows).
  - Suggestion to make current workflows (and other new basic content like Docker usage, class usage etc) available at workshop.bioconductor.org and/or in AnVIL.
- Plans to resolve broken F1000R pipeline.
  - Looking for replacements and suggestions. JOSE was proposed but indexing in Scopus/Web of Science is important.
- Suggestions to switch from 3 separate conferences (BioC202x, EuroBioC, BiocAsia) to a single larger conference that rotates hosting. Pro: better community building. Con: organization and financing might be difficult. Re-establish and have more communication between the three organizing conferences as currently they meet and plan separately. We attempted an all conference meeting but no follow up or continuation.
  - If we keep separate conferences it is also suggested to rebrand BioC202X to make it more “equal”. US BioC/BioC North America or something of the like, do not advertise as the "main" conference. Some slack discussion here: https://community-bioc.slack.com/archives/G01KMRQP3C7/p1702485210812569
  - Related topic: broadening selected and keynote speakers outside the regions of these conferences. Naturally that also brings a higher cost, virtual speakers have had problems in the past and are not as engaging for live attendees.

- Comments from TAB:
    - Developer Day (alternative - integrate more speakers that are developer-focused into the regular conference schedule).
    - Workshops - used to be more 'established' workflows rather than focused on newer packages.
    - Similar discussions taking place for useR (more technical/developer focus).

:21 - :30 Collaboration with r-universe to evolve the Bioconductor Build System processes to GitHub Actions-based framework.

- Background:
    - The BBS codebase is large (but has been working pretty well for a long time).
    - Currently managing 5 platforms (Windows, 2 x macOS, 2 x Linux).
    - Management of all these platforms is a significant undertaking.
    - Annotation pipeline needs an update, to reduce investment needed at each release.
- gitdevtest setup: developer sets up 3 remotes:
    - origin (developer's GitHub)
    - bioc (Bioconductor git)
    - gitdevtest (temporary GitHub remote, for testing - eventually will become git.bioconductor.org; gitdevtest is a proxy to a GitHub repository)
- See for example https://bioc.r-universe.dev/builds which is based on a GitHub organization produced by Gabor Csardi.
- TLDR: Hosting source code on GitHub instead of hosted git server. R-universe actions will build “dev” binaries/tarballs only available manually, and our GitHub Action will propagate to repository for BiocManager after validity checks.
- The tentative model is that Bioconductor will have a GitHub organization that holds the source code of all active packages, and will define a common action workflow for them. Developers will make changes to their repository via personal forks/clones and push to the Bioconductor remote when they are ready to build/test their changes. The R-Universe GHA workflows build/check across platforms and provide BBS-like feedback. If everything builds successfully a Bioconductor action will evaluate package validity (e.g., that version numbering is consistent with the history and current state, that desired checks pass) and propagate new versions to distribution as devel branch packages when validity is confirmed. If errors are detected, the package is only available from dev location and not via BiocManager. Handling of changes to "release branch" packages will be similar, targeting a different branch, and having additional rules for the Bioconductor action validity checks.
- Implications in terms of large-scale code ingestion by crawlers? The use of public and licensed material without proper credit is an issue that needs further consideration and discussion.
- R Consortium is now funding r-universe as a top-level project.
- r-multiverse (similar to Bioc in terms of having devel/release cycles).
- Important to not lock into GitHub specifically, portability important.

:30 - :43 Spatial omics workshop review.

- SpatialData workshop in Basel, 12-14 Nov, organized by the scverse team.
- 9th hackathon so far.
- Set up themes in advance.
- Results shared on BioHackrXiv.
- R interoperability participants: Helena Crowell, Louise Deconinck, Estella Dong, Dario Righelli, Artür Manukyan, Vince Carey.
- Aim: interoperability in terms of interacting with disk-based file formats for different types of spatial data.
- images & labels use Rarr -> ZarrArray (delayed), points & shapes use (geo)arrow -> Table/query (delayed).
- Some public data sets are available via Bioconductor.
- Elements can live in different coordinate spaces - represent coordinate system relationships as a graph (for alignment between elements).
- Zarr arrays have multiple scales/resolutions (pyramid levels).
- https://github.com/HelenaLC/SpatialData
- https://github.com/HelenaLC/SpatialData.plot

:43 - :60 Computo and workflow publication.

- Indexing is important for academic credit.
- Document by Workflows working group.
- Other options:
    - R Journal
    - Bioinformatics Application Notes
    - Bioinformatics Advances
    - JOSE

Addendum: Some notes on possible alternative funding channels for OSS (not aiming to be comprehensive)

- EC - consortium (somewhat like SOUND, RADIANT; ELIXIR?).
- https://www.sovereign.tech/ The Sovereign Tech Agency [German Gov't] supports the development, improvement, and maintenance of open digital infrastructure. No specific Biomed angle though, ideally should be more generic.