Bioc Technical Advisory Board Minutes

4 May 2023

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

:00 - :05 Welcome

- Prior minutes approved.
- Thanks to the dev team for a successful 3.17 release!

:05 - :15 Various technical topics

- Do we need a policy on chatGPT answers on support site, code contributions, etc.? The discussion of the StackOverflow ban is of some interest.
    - Concerns about copyright.
    - In general, people should be able to use tools like GitHub Copilot etc. to generate code.
    - One goal of our code review is to improve coding style.
    - Seems similar to some human agent copy-pasting some code from their scripts into an answer box, despite perhaps not really knowing the provenance or how it works.
- Efforts to interoperate with python/scverse – recall BiocPy, now scverse using AnnHub, see https://github.com/scverse/genomic-features.
    - Twitter thread about the recent scverse hackathon: https://twitter.com/scverse_team/status/1653418406450798595
    - Note the distinction between core and ecosystem in scverse.
    - Rubric: Use the interpreted language as a tool, but do not let its 'native' data representation or documentation methods 'get in the way' … HDF5, SOMA?
    - Our BBS can handle quarto vignettes with R and python mixed (e.g., BiocHail). Should be mentioned in the developer instructions (issue has been opened).
- r2u/bioc2u for rapid 0-config endowment of "empty" Linux machines, good for users and scalability? Compare to the situation with conda which seems to require more at the user level to get consistent environments and avoid bloat.
    - Having binary packages on Linux allowing us to build containers rapidly is important.
    - Can r2u be used with Docker/Singularity? Yes, as long as the containers are debian-based, it's the same whether a new machine or a container (our RStudio containers are currently ubuntu 22, and jupyter ubuntu 20, so they would work in both).
    - Should we use r2u for our containers? In those controlled environments it's actually faster to get the binaries directly. .deb packages are the binaries plus

    system deps, but system deps are already pre-installed in the containers. Balance between container config and deb config. Packages don't always do a good job of defining their system requirements.
    - Maybe we could solicit a support site "tutorial" from Dirk on how to use r2u to quickly install Bioc packages on deb/ubuntu.
- How to assign DOIs to book content? We can do it but what is the right approach?
- Distinction between "build system failure" and "developer's code frailty" needs to be clearer. Containerization cannot solve peculiarities like network problems, contention events, filled disk. Recent example.

:15 - :30 Hervé on SparseArray

- https://docs.google.com/presentation/d/1_ZKZdAUCKV3sGceU-nspF21oAkjLh0TodKSuilXFXjU/edit?usp=sharing
- Q: How do you read a huge csv file? Do you read it directly into a SparseArray, or first read and then convert to SparseArray? A: The package has a function to read directly from a csv file into a SparseArray object. A function to load from hdf5 is in preparation.

:30 - :37: Bob Shear on AWS footprint and cloud usage renovations

- https://docs.google.com/presentation/d/11QVpeD09TD1V5FvDPO1Co25_DfyrOtK_/edit?usp=sharing&ouid=105707534490078575160&rtpof=true&sd=true
- Deeper look at the traffic egress from bioconductor.org, specifically the packages. Presentation from FigShare (as part of data sharing consortium) - GREI (where NIH-funded researchers can place data to comply with requirements). They budget for egress costs - (how) can we leverage this?
- Fraction of requests that go to mirror (not served by bioconductor.org directly) ~ 35%.
- ~50% of traffic (in terms of size) was made up of a small number of requests for very large packages (dbSNP, reference genomes). Note that most of the content of these packages are provided free of charge from other sources already.
- CloudFront usage is cost-effective currently.
- Not much to do in the short term.

:37 - :42 Lori on the future of *Hubs

- Verification of content accessibility, object evaluation testing, retirement of resources in progress.
- Can some AH/EH resources be offloaded to zenodo or other permanent distribution resources?
- Deeper look at hub pipeline that is run at release time - see if pipelines need revamping to be made more efficient. Generate 1500 resources at release time, while only a subset of them are actually widely used. Takes a long time (weeks at the moment).
- "The hubs never forget" - no process for removing resources, may need to be re-evaluated.

:42 - :56 Alex on containers, binaries, workshop.bioconductor.org

- https://docs.google.com/presentation/d/1C0vvsQuasnHkJsHyHDMPGd9UIz0dXwcALYBUD6qPKaM/edit?usp=sharing
- The build for each package starts from the empty container - intermediate bucket in Jetstream that keeps dependencies to avoid pulling from bioconductor excessively.

:56 - :60 Other business

- 4 terms are expiring - nominations needed.
- Nomination form: https://forms.gle/2vjhUvykn6zECsnt9. Aim to post on May 6, nomination deadline May 31. Discussion in June meeting, followed by voting. Suggested announcement:
  Do you want to join the Bioconductor Technical Advisory Board, or do you know someone who would be a great fit? Then fill this short form (if you're nominating someone else, please first confirm that they are interested) before May 31 (at midnight in a time zone of your choice). If you are unable to access Google Forms, please see this pdf of questions, and email your nomination to charlottesoneson@gmail.com. For more information about the current board and the election process, see https://bioconductor.org/about/technical-advisory-board/, and if you have questions, feel free to post them in the #tech-advisory-board channel. Please help spread the word!
- BioC2023: Proposal from the conference committee that the 'Meet the TAB/CAB/core' sessions are done in the form of a (very short) overview followed by a panel discussion, with pre-formulated questions as backup.
- The August TAB meeting will take place during BioC2023 - do we move it?
- Kozo questioning and wanting feedback on Bioconductor packages in Python.
  - Is the Bioconductor community planning to support BiocPy?
  - If we re-implement the Bioconductor packages in Python, does BiocPy plan to collect contributions from those packages?
  - slack channel #biocpython created.
- Franck Richard’s posts about slowness of indexing and alignment operations on basic machines. Any interest in incorporating some Cuda into packages?