Bioc Technical Advisory Board Minutes

4 August 2022

Attending: Wolfgang Huber, Sean Davis, Lori Shepherd, Davide Risso, Vincent Carey, Charlotte Soneson, Aedin Culhane, Michael Love, Alexandru Mahmoud, Rafael Irizarry, Stephanie Hicks, Jennifer Wokaty, Nitesh Turaga, Marcel Ramos
Regrets: Kasper Hansen, Robert Gentleman, Levi Waldron, Laurent Gatto, Shila Ghazanfar

:03 - :04 July minutes approved.

:04 - :05 Welcome to new members: Sean Davis and Davide Risso. Thank you to departing members: Aaron Lun and Hector Corrada Bravo.

:05 - :06 Slides for Meet the TAB session at BioC2022.

:05 - :13 Working group concepts
- Aim to clarify the tasks/purpose of the various working groups.
- Do we have a procedure for evaluating how the interactions between the working groups and the CAB/TAB will be managed? It would be good to have a policy and a high-level process to evaluate (e.g. annually) how things are going. There will be bumps along the road and getting feedback early if the process is working well for all parties is important.
- CAB has done a few things. Initially all working groups presented at each meeting, however that took too long. Then moved working group reports to a document and only asked working groups to present if they had actions. Now moved to 2 working groups presenting each CAB meeting, on a rotating schedule.
- Add the information in the document to http://workinggroups.bioconductor.org/ that was advertised at the conference for members of the community to get involved.
- All working groups should have a document describing the group and could also have a dedicated slack channel.

:13 - :30 Presentation by Wolfgang. Upcoming presentations (update the schedule to include new members)
- Work in progress: "Linear models for multi-condition single-cell data without clustering of 'cell types', using Procrustes regression on SO(n) in dimension reduced (PCA) space". With Constantin Ahlmann-Eltze.
- Questions:
  - What would be the failure modes of the rotation? When does it not work? Don't know yet, need some model fit diagnostics.
  - Is it a problem if the cell type composition differs between samples? Intended for fairly subtle differences, for other situations other methods may be more suitable.
  - Nice review comparing procrustes to CCA: http://pbil.univ-lyon1.fr/JTHome/Biblio/Drayetal2003.pdf

- ○ Using standard tools for differential expression? Mixed effect model? Still the issue of correlation among cells coming from the same individual. Right now a standard linear model - can use the full language of linear models. A bit like the GLM where the link function maps into $SO(n)$. Can include external effects as well.
  - ○ One could do procrustes (or other joint decomposition) between conditions and look at distance between feature sets in latent space. Then test these feature distances for association with covariates.

:30 - :36 Archive distribution methods.
- BiocStorage package with dashboard has been on core agenda for a relatively short time.
- Redundancies - Bioc-funded AWS EC2/EBS, S3 buckets; Azure cloud storage and distribution; NSF Open Storage Network, AWS OpenData.
- Currently working on figuring out how well the Open Storage performs for worldwide access.
- Diversity of vehicles: Annotation packages, Experiment packages, AHub, EHub, APIs.
- Estimated time to achieve a plan to unify and simplify "data services" methodology - one year. Not clear how to prioritize this.
- Could there be a TAB working group on Data services?
- Would be good to have an active catalog of what's where, how often it's accessed etc.
- Also important to consider data integrity - some platforms provide guarantees, others don't.
- Want to avoid paying for shipping things out more than once. Minimize egress costs for most of the ecosystem.

:36 - :41 Outcomes/interactions from BioC2022
- Nitesh Mishra from Scripps Institute: HIV virology, B cell receptor VDJ recombination modeling – connected to Kelly Street's VDJdive submission?
- Stefano Mangiola: HCA as ExperimentHub resource? Or use the hca package. HCA data resources are at very diverse levels of maturity and a curatedHCA concept would make sense and add value – and the SOMA/TileDb work has to be examined to see how we want to work on this.
  - ○ EMBL/EBI centrally involved; Wolfgang will be involved in oversight.
  - ○ Hard to understand licensing of data in the HCA Portal. Which datasets are already published (and can be used/redistributed), which are not yet?
  - ○ hca may be sufficient for the moment, the annotation required for ExperimentHub would be a big effort.
- Nathan Sheffield: AnnotationHub, bedbase

:41 - :46 Adding "BiocType" as a DESCRIPTION file field was in discussion, to avoid heuristic inference on whether a package is Annotation or Experiment or workflow or book … or docker (?) from the biocViews field.
- biocViews and EDAM ontology - biocViews modernization as a TAB working group.

- Why the need to make a distinction? Build frequency, time constraints.
- We do already use BiocType for Workflows and Books to adjust time limits (at the moment that is the only use). The thought was that it makes some of the core team tasks easier and more transparent, and since we already require it for certain package types, why not standardize it across all.
- Suggestion: add the bioc type to the biocViews rather than as a new field - this would need to be enforced.
- Is the plan to do this retroactively? Yes.
- If BiocType is related to the build system, would it make sense to call it BiocBuildType? Core team could add during the submission process, as we may want to modify it in the future as the build system develops.
- This feels a bit like an implementation detail on the part of the build system - do we need to expose it to package developers, or is there a possibility of just keeping a mapping of packages to "build environment"? We have the manifests, the idea here was to attempt to make it more direct rather than having to do a match/mapping for builds/hooks.

:46 - :50 Licensing issues (see recent [discussion](#) on bioc-devel).
- Add note to contributor guidelines to clarify responsibility of the author relative to Bioconductor.
- Package reviewers can ask for clarification for derivative packages.

:51 - :60
- The Bioconductor Foundation contributed $1000 to the Mexican bioinformatics program.
- [Quarto](#) publishing system announced, 'next generation of RMarkdown', implications for vignettes in Bioconductor?
  - Should largely be backwards compatible.
- Discussion from Core over Annotation packages. Currently, they are treated differently.
  - Do not bump and branch.
    - Justification: annotations do not often change and a version bump implies a change.
    - Justification: not all Annotation packages are on git. Git version of annotation packages is a relatively new concept. Legacy packages were uploaded as tar.gz and checked manually before addition. For annotation packages with large data objects, git does not necessarily make sense. Is there a possibility of a hybrid AnnotationHub/ExperimentHub package?
    - Git annotation packages now have at least the concept of a release/master (soon to be devel) branch for consistency.
  - Need a different hook to allow for git pushes because of different versioning schema as the hook will prevent rogue version bumps for all other types of packages. But odd/even version for release/devel does not exist and versioning of certain Bioconductor packages are quite different (e.g using Bioconductor release version as version of package).
    - Two versions to consider: the underlying data resource, and the package that needs to match up with the AnnotationHub API etc.

- Slack pro to be purchased by foundation for a year at $2K. CZI grant?
- Thoughts on interfacing with python
  - Aaron would like someone to take over support for basilisk.
    - How much should Bioconductor do (centrally) in terms of supporting python interaction? Anything that brings the worlds more together is useful, especially in single-cell. Could we do more outreach to the python community?
    - Reach out to posit? basilisk solves a problem that reticulate doesn't.
  - Can we support python users to use our annotation resources for example? pyBiocManager? pyBiocAHub?

Other notes
- Governance update/discussion about the TAB structure and mission - postpone until next month.