Bioc Technical Advisory Board Minutes

3 December 2020

Attending: Vince Carey, Levi Waldron, Charlotte Soneson, Michael Love, Wolfgang Huber, Martin Morgan, Aaron Lun, Stephanie Hicks, Rafael Irizarry, Kasper Hansen, Aedin Culhane, Lori Shepherd (guest), Laurent Gatto, Hector Corrada Bravo, Robert Gentleman (joined at :30)
Regrets: Shila Ghazanfar

:03-:05 - 2020-11-05 minutes approved.

- Governance process progress. Suggestion to hold quarterly joint CAB/TAB meetings.
- How to request a budget for a TAB project? Request for proposals, relation to BiocChallenges (see below). Bottom line: TAB members are invited to consider how to propose and how to manage a granting process for Bioc-oriented work.

:05-:20 - Technical issues

- Getting ready for Apple M1 chips -- apparently some issues with inferior numeric support (see also R-core, below). Seems likely that R-related problems will be resolved in the relatively short term.
- Mikhail's question to slack on liftOver of paired bed data
- Hubs: https://community-bioc.slack.com/archives/CDSG30G66/p1606036502028400 -- are there technical problems with data service? Timeouts, locked directories -- transient. ExperimentHub (but not AnnotationHub) occasionally fails to find the MySQL server. Seems unrelated to multiple simultaneous connections, unlikely to be related to resource consumption. Still a problem, under active investigation, but currently occurs relatively rarely.
- Acceptability of .github/workflows in main branch of git repo for a Bioconductor package.
    - Initial report of issues on slack.
    - Reference to submission policy, excluding GHA workflows.
    - BiocCheck events -- part of reviewers' rubric, submission instructions
    - Is it necessary to have a main branch, a devel branch for syncing, and a third branch for development?
    - Can the workflow be lodged in a non-main branch?
    - Systemic solutions seem possible
    - What are the downsides of just allowing these files in the default branch? Is there a case for denying presence of non-package-related files? Security vs. convenience -- definiteness of review policies.
    - Need for a canonical statement, term of validity, conditions for revision, of submission/review policies.
    - Comments: If a non-package file is included in .Rbuildignore, that's a sign that the developer has actively/consciously put the file there.

    - Explicitly allow/disallow specific types of files. .Rdata and .Rhistory, for example, should not be included (e.g., since they can contain sensitive information).
    - Challenge: the 'main' branch from the developer point of view is not necessarily "the same" as the 'main' branch from Bioconductor's point of view. Make the transition between the two as easy as possible.
    - There are already many git-related questions on the mailing list. Need to think about the effects of adding more requirements. Predominantly experienced developers who want to add workflows/websites/etc?
- Requirement that "ownership" of package name passes to Bioconductor so that orphaned packages can be maintained by new maintainers if desired (to be discussed at next meeting). CRAN policy: "Package maintainers give the right to use that package name to CRAN when they submit, so the CRAN team may orphan a package and allow another maintainer to take it over." (https://cran.r-project.org/web/packages/policies.html)
- https://jorainer.github.io/SpectraTutorials/ with docker. Suggestion: invite Johannes and Laurent for TAB tech talk on proteomics infrastructure/analysis tools
- https://kevinrue.github.io/BiocChallenges/ - collection of 'challenges' related to Bioconductor, compiled for EuroBioc2020. Could turn into a useful platform for collaborative work.
    - TAB should revisit on a regular basis
    - Scope and engagement of general core activities
    - Each challenge has a leader identified who can answer questions but need not do technical work on the solution
    - Motivate discussion of challenges in the TAB
    - Vince has prepared a new challenge on a fork -- https://github.com/vjcitn/BiocChallenges/tree/vince-1
- dropbox endpoints for downloading resources -- BiocCheck bitbucket … needs strategic alignment of core and developer interest -- Rirods, globus -- transferring large data
- R-core activities. Preparing R for 'Apple silicon' (changing macOS compiler & underlying hardware). Already some package changes (requested by Prof. Ripley); likely future challenge is a build system machine based on Apple silicon required to produce binary packages compatible with both old and new systems. Must also be compatible with CRAN binaries, which usually dictates that we follow Simon Urbanek's lead -- buy an old machine, or purchase VM software emulating an old OS on new hardware.
- Do we need to update the devel support pages to reflect main rather than master in git commands? Git is using main by default.

:20-:30 - CAB (last meeting was Nov 12th 2020).

- Two new committees established
    - **Package submission and review** will form guidelines to facilitate package submission and assist the core team in package review. Members of sub committee: Lori, Kayla, Johannes,Yagoub.
        - Regarding core packages where a user/developer complained that core Bioc packages are not accessible/available on GitHub (see issue here).

Decided it might be indeed helpful to have (all) core Bioc packages on GitHub. A simple first solution/fix could be to ensure that the BugReports field in the package's DESCRIPTION points to the github repo, and that the GitHub repo says in the README that it is the official package repo.

    - **Bioconductor Education and Training Committee.** Proposed to CAB by Charlotte Soneson and Laurent Gatto. Concrete outcomes so far have been newly trained Carpentries instructors from the Bioconductor community, development of Bioconductor-oriented lessons as part of the Carpentry lesson incubator (still early stage), monthly meetings, and considering joining the Carpentry organisation membership program. Will announce at EuroBioc2020. Leaders Laurent, Charlotte, CAB members: Saskia, Susan.
- **CoC.** Bioconductor-wide CoC documents almost ready, awaiting final review/approval before submission to TAB and full CAB. 2 CoC incidents (on slack).
- **Slack**:
    - CoC will help hopefully with slack communication but CAB considering how to make slack more user friendly and welcoming
    - A "buddy system" whereby those new to Slack workspace have a trusted friend in the workspace
    - Navigating a growing number of channels. The "Browse channels" filter is tedious to scroll through when searching for "the right channel".
    - CAB approved/agreed with the suggestion to disable creation of new private channels on slack without an admin. Admin unable to monitor private channels and DM messages or even know the number of private channels created.
- **Events**. Discussed need for document that summarises minimum requirements of a “Bioc” event
    - **Bioc2021**. New website, logo. Discussion about virtual/hybrid. Currently planning for virtual with some in-person events/watch party in Seattle. Virtual conferences are resulting in larger registration numbers
    - **BioC Europe** (Dec 2020)
        - 8 confirmed invited speakers, 28 contributions (talks, workshops, posters) and 103 people registered as of Nov 12. Registration is free and open until Dec 7.
        - Kevin Rue-Albrecht starting some “BiocChallenges”. https://kevinrue.github.io/BiocChallenges/ and asks for contributing challenges.
        - European Bioconductor Society to be founded during this conference, primarily as a vehicle to buffer money between events (conference, courses). Not aiming to create a distinct identity, rather a low-profile substructure of the global BioC project.
    - **Asia** BioC Asia 2020 (October 15-18. half days, virtual format)
        - 440 people registered, peak attendance of ~ 150. Workshops in English and Mandarin. Recorded all sessions. Need to put on Bioconductor YouTube

        - Potential to host 2021 event in Japan (Kozo to lead)
    - **Japan** Bio"Pack"athon Nov 11 (Japan Standard Time), virtual format Next event Dec 9
        - 8 participants. Ongoing teaching videos, package development
        - 3 Japanese teaching videos (about Bioconductor) https://togotv.dbcls.jp/en
        - Trying Twitter Polls to know the Japanese Bioconductor needs https://twitter.com/biopackathon/status/1326802411730530305
    - **Mexico** Leo awarded CS&S event funding to support Mexican conference/training
    - **H3Africa**
        - Add name to **potential speakers for H3ABioNet webinar series**
        - Potential collaborations (in education/ training) with H3Africa (Aedin sent a LOS of CAB support in their application for NIH funding proposal).
        - Yagoub has contacted Rolanda Julius about collaborating with H3Africa, and providing training sessions for H3Africa, also about the possibility of organizing BioC Africa.
    - **Bioconda / Bioconductor hackathon** https://github.com/bioconda/bioconda-recipes/issues/25225. Hackathon to fix ~ 75 Bioc packages that are not compiling for bioconda. Most are "Can't resolve environment", "Can't compile" or "fails on OSX". Currently, all issues appear to be solved.
- *Mission statement* revision? Work in progress. A proposal started by Levi:

  "The Bioconductor Project (https://bioconductor.org/) mission is to promote the statistical analysis and comprehension of current and emerging high-throughput biological assays. The Bioconductor Project is committed to open source, collaborative, distributed software development, and literate, reproducible research.

  The project aims to build a robust, diverse, inclusive, supportive, and skilled community of developers and users. Bioconductor is not an "insiders" club but a transparent organization with leadership opportunities open to all with shared interests in open-source Bioinformatics software and an interest in furthering the goals of the organization."

  Comments: Is the main purpose software or statistical analysis? Is it clear what is meant by 'literate'? Avoid definition by negation. Is the last sentence self-referential?

- We need a document that describes relationships between TAB, CAB, SAB, core team managers, and probably have to draft it ourselves. This document would include processes of policy review, approval, implementation with dates of inception and expiration, and enforcement.

- As funding from Foundation is offered, accountability process is needed. Invoice / reimbursement concepts.

:30-:50 - Mike Love, Hector Corrada Bravo, Aaron Lun, Lori Shepherd - Bioc Hub directions

- Recent ExperimentHub usage statistics are summarized in an appendix to this document
- Comments:
    - AWS Open Data project (https://registry.opendata.aws/) - glue database. bed files etc could go into such a structure. (Sean Davis is in conversation with AWS opendata group)
    - Re. organization/shared caches: perhaps, a local mirror of the hubs may be what one actually wants.
    - Wish to increase the flexibility of ExperimentHub (e.g., recount3 ~75TB). Not everything fits in the current structure. Could be more like an 'index' of useful resources instead of host of data itself.
    - Working group on EHub needed
    - If 'recipes' are submitted (e.g. as for AnnotationHubData), who is responsible if they break (submitter, core)?
    - What about having a 'form' for hub submission instead of the current submission approach?
    - Make it easier to submit data, perhaps with less long-term accountability requirements?

:50-:60 - Open discussion

Appendix: ExperimentHub statistics

Some quick stats on ExperimentHub, over the last 9ish days

```
# A tibble: 1 x 4
  days          n_total n_builder n_per_day
  <drtn>          <int>     <int>     <dbl>
1 9.281262 days  437350    282621    47122.
```

- 430k accesses, about ⅔ from our build machines

For the builders:

```
# A tibble: 2 x 2
  op         n
  <chr>  <int>
1 GET    48424
2 HEAD  234197
```

- ExperimentHub uses BiocFileCache, which uses HEAD to see if a file needs downloading; this saves a lot of traffic
- Each HEAD means potentially that the actual data is not downloaded, saving network traffic and significant 'egress' charges from AWS.

These are the files accessed by the builders:

```
# A tibble: 1,716 x 3
   path                              GET  HEAD
   <chr>                           <int> <int>
 1 /metadata/experimenthub.sqlite3   290 42558
 2 /fetch/1035                        63   315
 3 /fetch/912                         63   306
 4 /fetch/913                         63   306
 5 /fetch/919                         63   306
 6 /fetch/2061                        90   264
 7 /fetch/3348                         8   298
 8 /fetch/1957                        63   189
 9 /fetch/1958                        63   189
10 /fetch/2143                        63   135
# … with 1,706 more rows
```

- The sqlite file might be downloaded (GET) if a resource is added, or if a package sets ExperimentHub to use a temporary location
- Only about 1/2 the resources are referenced in evaluated code
- 'fetch' ids (e.g., 1035 in `/fetch/1035`) are not the same as 'EH' ids (`"EH123"`) entered by the user, so some additional work needs to be done to map these to actual resources...

For non-builder access -- each row represents a distinct IP address, arranged in descending order of GET + HEAD requests

```
# A tibble: 1,696 x 2
     GET  HEAD
   <int> <int>
 1  4173 18899
 2     3 10261
 3  1986  4486
 4     3  4035
 5  1003  2420
 6   213  2365
 7     4  1928
```

```
 8     2  1684
 9   785   899
10   485  1185
# … with 1,686 more rows
```

- About 1700 distinct IP addresses over 9 days
- A few very heavy consumers -- are these CI platforms or robots or rogue scripts or ...?
- HEAD is important in reducing traffic

What's being accessed?

```
# A tibble: 1,607 x 3
   path                              GET  HEAD
   <chr>                           <int> <int>
 1 /metadata/experimenthub.sqlite3  4079 32137
 2 /                               16777   226
 3 /fetch/3148                        72  1847
 4 /fetch/3106                       320  1541
 5 /fetch/3107                       270  1507
 6 /fetch/3508                       259  1296
 7 /fetch/3509                       194  1244
 8 /fetch/3314                       172  1088
 9 /fetch/2573                       182   994
10 /fetch/2577                       180   992
# … with 1,597 more rows
```

- Not really sure when `/` is a reasonable access point -- robots? (robots.txt tells well-behaved bots not to visit…)
- Would be interesting to know the actual resources being accessed