Bioc Technical Advisory Board Minutes

4 March 2021

Attending: Vince Carey, Laurent Gatto, Kasper Hansen, Levi Waldron, Charlotte Soneson, Michael Love, Wolfgang Huber, Shila Ghazanfar, Stephanie Hicks, Martin Morgan, Robert Gentleman, Aedin Culhane, Hector Corrada Bravo, Rafael Irizarry, Aaron Lun
Apologies:

:01 - :03) 2021-02-04 minutes approved

:03 - :08) Funding pursuits:

- Vince/Aedin: NIH R03 for Common Fund clients and tutorials (e.g., 4D nucleome, GTEx, IDG Pharos)
- Vince: AWS cloud credits (due 31 March but email says decisions delayed).
- CAB intend to apply for CZI EOSS. LOI March 30, 2021. Full application May 19, 2021. 2 year $50k - $200k total costs/year (inclusive 15% indirect/overhead costs). https://chanzuckerberg.com/rfa/essential-open-source-software-for-science/

:08 - :10) Open comments on joint TAB/CAB meeting

:10 - :15) Discussion of options for the mission statement - board members are encouraged to provide feedback/preferences among the options.

:15 - :25) CAB liaison/teaching committee

- CAB: nominations for new members. 19 applications received. Up to 5 positions available. Currently reviewing applications. Recycling positions - need to define the number to appoint.
- Planning Bioc awards 2021.
- Conferences: H3 Africa conference (April 26-30) contribution, useR! (July 5-9)
- Package Review Working Group -- reviewed current practices; comparison with rOpenSci / JOSS. Starting toward development of reviewer training material.
- New developer group mentorship program (draft seeking suggestions/ideas). Coordinate with package review working group. There may be scope for engaging the monthly developer forum in providing more introductory topics.
- Teaching committee: monthly calls, three lessons in development (intro, Bioc project specifics, RNA-seq). Several suggestions (and volunteers) for more advanced/specialized lessons to build on top of the introductory ones. Carpentries membership being discussed.

:25 - :35) How should we identify project core API/core packages? They should be enumerated at www.bioconductor.org, with named teams identified as developers. Identify milestones for

testing, documentation, new feature planning, any breaking API changes for the core packages should not be committed until discussion at a TAB meeting.

Comments:

- Identify strategies to distribute the knowledge (within the core team), heavy dependence on single individuals with critical expertise
- Identify critical packages - identify (and fix) issues early
- Core team should take a forward-looking position, "ahead of the curve" (e.g. in terms of infrastructure)
- Phase out/deprecate old packages
- Possible analogy: https://elixir-europe.org/platforms/data/core-data-resources
- Once core resources have been identified, it will be important to - on a per package basis - identify "threaths" to the package's short and long term viability

As an example, using as a measure of "core-ness" the number of package that (directly or indirectly) depend on or import a package (could also be weighted by the number of downloads) gives the following top list (code):
BiocGenerics - 1486
Biobase - 1249
S4Vectors - 1231
zlibbioc - 1197
IRanges - 1173
XVector - 1134
GenomeInfoDb - 862
Biostrings - 861
GenomicRanges - 861
MatrixGenerics - 762
DelayedArray - 760
SummarizedExperiment - 745
BiocParallel - 696
KEGGREST - 607
AnnotationDbi - 600
limma - 430
Rsamtools - 430
GenomicAlignments - 394
BiocFileCache - 360
BiocIO - 359
rtracklayer - 354
biomaRt - 316
graph - 275
annotate - 255
GenomicFeatures - 252

:35 - :40) The subgroup on AnnotationHub "2.0" has met.

- Hector, Sean Davis, Mikhail, Vince, Nathan Sheffield.
- Nathan's bedbase.org has bed and bigbed files and can support targeted queries; a postgres database manages metadata.
- The group will experiment with client design in Bioc framework - e.g., a RESTful GenomicFiles used to define and execute queries to bedbase API. Core developer effort would be useful.
- Aim: Enable richer queries in AnnotationHub:
    - being able to search and query based on metadata on annotations
    - subsetting/filtering on annotations based on genomic regions
    - finding annotations based on properties (e.g. GC content)
- Think about findability for *Hub resources.
- Big challenge: indexing and organizing *Hub content, documenting origin/provenance.
- Figure out what the most common tasks are and make them easy to do (e.g. via wrappers)
- Governance of the hubs perhaps more important than the technical discussion. Who does curation, who gets credit. What about retrospective curation of metadata.
- Guidelines from the "curated*" data packages? More mandatory fields in the metadata?

:40 - :50) Quick review of Aaron's recent contributions to SE, SCE, S4Vectors

- Ability to combine different objects (e.g., DataFrame, SummarizedExperiment) by row or column (equivalence to rbind(), cbind() that would be more relaxed/flexible, similar to dplyr::bind_rows()) -> combineRows() [for DataFrame, SE], combineCols() [for DataFrame]
- ConstantMatrix (to be added to DelayedArray) - avoid the need to store all NAs explicitly

:50 - :55) Items related to 3.13 release preparations

- Will use R 4.1 (including, e.g., the native pipe)
- CRAN has not yet announced the release date for R 4.1
- Non-maintained packages are more actively being deprecated - affects also packages depending on these

:55 - :60) Open discussion

Appendix: working group notes

Governance:
*Hub: call noted above
spatial/multiomic:
matrix/array processing: significant innovation with HDF5/(AWS) S3
build/check containerization:
package review process analysis: