

Bioc Technical Advisory Board Minutes

4 June 2020

Attending: Martin Morgan, Laurent Gatto, Charlotte Sonesson, Levi Waldron, Stephanie Hicks, Kasper Hansen, Robert Gentleman, Aaron Lun, Sean Davis, Vince Carey, Wolfgang Huber, Rafael Irizarry, Aedin Culhane

Regrets: Matt Ritchie

Schedule

:00 - :05 Minutes

- 2020-05-07 [minutes](#) approved

:05 - :15 Material for review

- Highlights
 -
- Funding
 - [U41](#): (NHGRI; core) Renewal submitted (as U24) January 24. Panel date 2020-06-09
 - [U24](#): Monthly coordination meetings; private slack channel
 - [AnVIL](#): <https://anvilproject.org>: bi-weekly bioc meetings, public [slack](#); weekly tech calls. Bioconductor-specific [progress](#).
 - [CZI seed network](#): monthly meetings, private slack. Annual report June 1
 - CZI EOSS (Vince)
 - Other opportunities?
 - CZI EOSS call "The third distinct cycle will open mid-June 2020."
 - NSF Conferences -- [Computational illiteracy](#)
 - Burroughs-Welcome trust -- conference sponsorship through Rafa for next year, have an interest in Bioconductor
- Events
 - BiocAsia: [2020](#): Virtual meeting, Oct 17-18.
 - BiocEurope: Padova, Italy - Davide Risso as local host. Format (spatially localized or online) not confirmed yet. In either case, planned in the week of Dec 7-11 2020, usual 2+1 days format if physical, more drawn out 5x(1/2) days if online(?). We're trying to learn from the BioC2020 conference in the summer.
 - [Bioc2020](#): July 27-31 2020 virtual conference. Public [slack](#).
 - CSAMA June 21-26 2020: Postponed to 2021: 2021-06-27 to 2021-07-02
 - Other
- Community Advisory Board
 - [CAB page](#)
 - [Minutes](#) from previous meeting (May 14)

- CZI EOSS grant submission (fostering Bioc community activities) planned
- Committees
 - Code of Conduct Committee (Levi, with Stephanie): nothing to report
- Working groups
 - [Carpentries training](#) (Laurent, Charlotte):
 - instructor training took place online May 19-20
 - 5 participants from Bioc (Charlotte Soneson, Robert Castelo, Marcel Ramos, Nitesh Turaga, Kayla Interdonato)
 - currently going through the certification process
 - lots of useful ideas on how to structure teaching material and courses, giving and collecting feedback
 - next steps: finish certification, initialize discussion on development of lesson material. Larger group of people interested (including some from the Carpentries). Provide guidance to BioC2020 workshop instructors.
 - Developer forum (Michael, Martin): ongoing activities
 - Fluent Genomics (Michael)
 - Build system / continuous integration (Vince)
- Other:
 - European Bioconductor Foundation (Laurent / Wolfgang / Simone B.): no substantial news here, has been stalled due to lockdowns. Likely we'll be going for charitable registered association (gemeinnütziger eingetragener Verein) [https://en.wikipedia.org/wiki/Registered_association_\(Germany\)](https://en.wikipedia.org/wiki/Registered_association_(Germany)). This would be a legal person under German law, but could operate internationally and in particular in the EU. (A similar legal entity under EU law does not yet exist, but appears to be in discussion (for decades...) and may come eventually.)

:15 - :20 BioC 2020

- Remember: APPROVED: BioC2020 a virtual conference in the same week in July.
- Workshop group is meeting weekly to develop programme and platform
- Will be spread over 5 days instead of usual 3 (July 27-31)
- New registration fee is \$50. Strong registration. Will extend deadline for "travel scholarships" (now just registration waivers)

:20 - :45 Working with large data

- Inputs from Kasper, Sean, Aaron, Stephanie
 - Aaron's [working doc](#) on matrix multiplication / DelayedArray
 - Sean Davis's [big-picture notes](#) and [rendered](#)
 - Missing capability around producing e.g. documentation that lives outside packages
 - Goal to have not just a computable document but a reproducible computing environment associated with it

- Approaches in the larger community, e.g., Apache Arrow
- Deep indexing of AnnotationHub / ExperimentHub
- knowledge graph is a flexible big-data approach not leveraged much in Bioc
- People doing deep learning / machine learning have the impression that python is required for performance issues.
 - interactive analysis tends to lead toward R
- altrep in R 4.0 is another framework for big data that may be more natural than DelayedArray (Robert)
 - but may be dangerous exposed to user (Aaron)
- DelayedArray-like activities
 - Biggest issues
 - Documentation and use limited, so we are not seeing adoption & use cases
 - Addressing big issues
 - Document current capabilities and software
 - Proposal to create a set of Rmarkdown documents associated with 1) data, 2) single or set of highly related use cases, 3) possible solutions using DelayedArray
 - Recognize and document gaps associated with new use cases
 - Capture advancements in approaches and “backfilling” documentation of capabilities
 - Relate concepts and practices to each other across packages and time
 - Expose experts to problems--socialize the problems
 - Implementation challenges
 - Stages of grief
 - It is possible: all of Aaron's packages work 'efficiently' with DelayedArray
 - Key is to ensure 'reasonably efficient' row access and column access
 - Specific implementation challenges
 - Minimizing independent iterations
 - Minimizing 'read' operations
 - Coordinating out-of-memory outputs, e.g., keeping external files in sync
 - Is parallelization basically broken / not useful
 - Action items (DelayedArray-centric)
 - Use cases document illustrating transformation of package 'X' to support DelayedArray
 - Use cases outlined by Kasper, Stephanie, tied to data
 - Technical contributions by, e.g., Aaron
 - Vetted by Hervé
 - Revised by Stephanie, Kasper
 - List of packages for inspiration
 - Surfacing and solving technical limitations (gaps)

- e.g., 'native' sparse matrix operations, e.g., matrix multiplication (is this not the domain of DelayedMatrixStats?)
- Alternative back-ends, e.g., TileDB with native sparse representation
- How are these best incorporated into main-stream best-practices?
- Code review of DelayedArray reverse dependencies (backfill) (Hervé)
 - How is it being used?
 - What types of limitations do authors encounter and work around?
 - How can these problems be solved intrinsically
- Socialize via developer forum, ...
- Approaches in the larger community
- Client-server model
- Often client and / or server in the cloud
- Bioconductor services to exploit the cloud, e.g., provide a very large machine to perform quantile normalization on 100,000 EPIC arrays
- Snaptron queries of very large RNAseq collections at base-pair resolution; a better path forward than trying to make ExperimentHub / AnnotationHub more granular?
- Enabling exploratory analysis at scale
 - Rafa's experience -- smart subsetting for exploratory analysis
 - "The best way to explore big data is to make it small" R. Gentleman, 2020.
- Use cases for graph-based (for e.g.,) analyses
- Deep learning / modern machine learning & movement to python. Scripting versus interactivity.
- ALTREP -- more transparent representation
- Other challenges (separate topic(s))
 - Milestones & overall coordination of progress
 - Community around production of new software
 - Venue for large-scale products like osca, ...
 - Notebooks (in addition to markdown) as a communication / engagement vehicle

:45 - :50 Technical Advisory Board executive nominees

- Martin Morgan is completing his period as Chair. Martin made following nominations:
 - Chair: Vince (currently Vice-Chair)
 - Vice-Chair: Levi (currently Secretary)
 - Secretary: Charlotte
 - Motion by Robert, seconded by Stephanie. Will vote next month.

:50 - :60 Technical Advisory Board nominees

- Procedures

- [Governance document](#): (two-thirds) supermajority (of cast votes) approval by current board members
- Today: review and discuss nominees
- Next week
 - Levi will send out ballot
 - Each TAB member votes to approve or disapprove each nominee
 - Nominees ranked, all with supermajority elected to board; if more than five nominees with supermajority, discussion amongst TAB executive to decide path forward; communication and approval by the TAB.
- Members standing for re-appointment should leave the call