Bioc Technical Advisory Board Minutes

1 September 2022

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

:03 - :08
- Previous [minutes](#) approved.
- M1 binaries - good progress.
- [r-universe](#) building and distributing Bioconductor packages.
    - They have an app to add to your own GitHub repos, that will add them to r-universe.
    - Lots of redundancy with what Bioconductor does - building/testing, vignettes.
    - There are ways to [prevent](#) a package from being added.
    - Suggestion: write up a short blog post (on the bioc blog) - if someone finds r-universe and has installation issues, googling "difference between r-universe and Bioconductor" they would land on this post where they learn about potential problems for installing packages. Could also include differences between bioconda and Bioconductor. Collect a bit more data on possible inconsistencies. BiocManager functions are helpful in cases one hits binary incompatibilities somehow, when some packages have been installed by other means.
- Consider also a blog post about Bioconductor support for quarto - any issues with BiocStyle?
- basilisk updated to python 3.8, working on arm64 (M1).
- Likely 3.16 release 9 November 2022.
- Conference follow-up - upcoming CZI meeting, [EuroBioC](#), BiocAsia.

:08 - :10 Should the October meeting be a joint CAB/TAB meeting? Lori to send out suggestions for meeting times.

:10 - :15 Mike Love ad hoc talk
- Working on trying to detect allelic imbalance (AI) at sub-gene resolution ([slides](#)).
- Measure within an individual (or cell) the effective regulatory SNPs.
- An individual must be heterozygous in regulatory region, and in exon, to have/measure AI.
- Most methods are SNP-based; look at coverage per SNP.
- Lots of exciting work ongoing in AI:

- ○ Combs and Fraser (spatial AI dataset)
  https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007631.
  Also I've added this to Bioconductor as a data package: {spatialDmelxsim}.
  - ○ Findley et al (AI across few donors, many treatments)
    https://elifesciences.org/articles/67077.
  - ○ spASE (spatial AI) from Irizarry lab
    https://www.biorxiv.org/content/10.1101/2021.12.01.470861v1.
  - ○ scDALI (single cell AI, Gaussian Processes) from Stegle et al
    https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02593-8.
  - ○ Airpart (single cell AI, fused lasso) https://bioconductor.org/packages/airpart.
- Isoforms may be under different control.
- Preprint: https://www.biorxiv.org/content/10.1101/2022.08.12.503785v1.
- Vignette:
  https://bioconductor.org/packages/release/bioc/vignettes/fishpond/inst/doc/allelic.html.

:15 - :20 Sean Davis ad hoc talk
- GEOquery updates - working with NCBI to provide homogeneously processed human RNA-seq data.
- GEOmetadb and SRAdb broken (harder to get metadata from GEO/SRA and keep it up to date) - working on next gen replacements.
- RefineBio R package (https://refine.bio client, Alex Ho (medical student)) - SCAN for affy processing, Salmon for RNA-seq processing.
- GenomicSuperSignature extensions (Sehyun and Levi).
- Orchestra extensions - self-service addition, new tagging/organization.

:20 - :25 Scientific Advisory Board members invited, annual report needs to be revised, meeting will occur October 28. Current membership:
- Robert Gentleman (chair)
- Rafael Irizarry
- Vincent Carey
- Christopher Wellington
- Wolfgang Huber
- Martin Morgan
- Benjamin Neale
- Michael Schatz
- Aviv Regev
- Sandrine Dudoit
- John Marioni
- Daniela M Witten
- Barbara Engelhardt
- Susan Holmes
- Benilton Carvalho

Develop a charter for the SAB as a guide - think about what kind of advice is useful to get, current pain points. Suggestions: How they use Bioconductor, what they need, current/future directions in technology, what they see as shortcomings of Bioconductor.
Other sources of potential members: https://h3africa.org/index.php/contacts/,

:25 - :30 G-DADS genomic data and analysis development services: storage/compute/GPU/global distribution
- 50TB storage, virtual machines, GPU allocation.
- NSF estimates the value of the allocation at $132,000 per year. To get it renewed we need to demonstrate usage.
- Task of actually testing GPU-oriented packages in the context of BBS needs work.
- GDSCN for diversification.
- IAM.

:30 - :33 Dependency heaviness paper in arXiv
- Identify core packages, vulnerabilities.

:34 - :60 Governance discussion
Motivations/other examples:
- What does the TAB do, what could it do, what is it not doing?
- Could we do more on providing leadership to the community, helping recruit technical support for the project, giving technical advice to the core team?
- The CAB works along a somewhat different model.
- In Galaxy, the TAB is made up of working group representatives.
Ideas:
- Switch all/part of the TAB to consist of representatives of working groups organized around technical themes.
- Expand the board to make rooms for representatives of working groups.
- Increase connection between working groups and core/boards.

Comments:
- Need to think about what we get out of changing the system.
- How do we measure how well a specific system works?
- The TAB should be a mix of people who are broadly interested in the project and people who are involved in specific working groups.
- One important goal of Bioconductor is to provide statisticians/analysts a way to analyze genomics data rigorously, and promote good statistical practices through packages.
- Are we attracting cutting-edge tools/methods? Working groups can be leveraged to develop cutting-edge infrastructure/tools, rather than individuals.
- Open development has provided Bioconductor with lots of good packages. Is openness a potential issue if we get 'too many' packages? Puts a lot of pressure on the core team.
  - Concept of initial triage of submitted packages, look at the 'history' of the package (e.g. engagement on GitHub)?
- Hard to predict in advance what is or will be valuable (esp for a niche community).

- We also want multiple packages doing the same thing, to encourage 'competition'.
- In terms of lessons learned from single-cell (and why folks may look at tools outside of Bioconductor), recommend creating a working group for this new biological data type and reach out to companies producing new technologies for creating workflows for this new data type. This is how other projects are establishing a strong hold when new technologies come out. The complication becomes NDAs, which is less clear with a working group (vs a single PI lab).
- Are there specific reasons why some good R packages that are not in Bioconductor, were not submitted to Bioconductor? Can we gather such data?
- Think about a formalization of working groups - charter, reporting.
- NIH example of SIGs/working groups: https://oir.nih.gov/sigs.
- We have two roles: (i) provide the best tool(s) for doing certain tasks, (ii) journal-like distribution platform 'for the scientific record'. Getting (i) is not done on command/by committee. We need to be open to and compelling for whoever the creator of the next limma/edgeR/DESeq2 etc. is.