Bioc Technical Advisory Board Minutes

7 November 2024

Present: Vince Carey, Stephanie Hicks, Helena Crowell, Rafael Irizarry, Jacques Serizay, Charlotte Soneson, Wolfgang Huber, Alexandru Mahmoud, Marcel Ramos, Laurent Gatto, Kasper Hansen, Erdal Cosgun, Michael Lawrence, Henrik Bengtsson, Lori Kern, Ludwig Geistlinger, Levi Waldron, Jen Wokaty, Tim Triche
Apologies: Davide Risso

:03 - :04 Previous minutes approved.

:05 - :18 Review of topics of interest to TAB members, proposed briefly at October TAB meeting or in interim.

- Spatial omics data - call occurred 30 Oct 2024.
- bluesky social media platform [events on #social-media on slack] - account launched.
- Single-cell data - nf/core – NextFlow core … scrnaseq pipeline, should we make a PR so that it outputs SCE, or uses more of our infrastructure/methods?
- Data repositories / ExperimentHub - concerns about transparency of what a user gets with eh[[x]], concern with potential for malice. eh[[will run new() on class definition specified in user-supplied metadata.
- Build system move to GitHub Actions - active, later in meeting.
- (Generative) AI - time available for discussion later in call.
- Performance improvement of packages - Rcollectl as a measurement tool.
- Interactions with other communities (CRAN, scverse) - engaged with scverse, Kurt Hornik and Tomas Kalibera at CRAN.
- Multi-modal data.
- Infrastructure building.
- Making sure that Bioconductor stays competitive for users - what's the metric? Requirement for more advertising and demonstration of value (pair with justifying developers submitting to Bioc below).
    - Front page of website could be more informative - more tailored, more specific images/example plots. Something like CRAN task view for specific disciplines/applications (bulk, single-cell, spatial).
    - Suggestion: collaborative "bioconductor-science" website/book (pkgdown?) that has content that can inspire new developers. Dynamic, easy to contribute. Note that pkgdown doesn't support all kinds of vignettes.
    - The devel part of the website is not as easy to navigate (harder to find the path from the landing page). Developers tab is more informative (different from clicking "For developers" on landing page).
    - Gap between the landing page and the software pages.
    - Suggestion: add a switch to the website that would show tailored information to groups of users (change the content of the displayed page).

  - Would be nice to have automated tracker/feed of newly added package on the website. There is https://bioconductor.org/dashboard/ (but it's not well advertised - perhaps could have some form of carousel of new packages). Slack has a channel with new submissions as well. Include description as well, to avoid having to click through to each package landing page.
  - The website should have a button to easily toggle to devel version of packages (instead of release).
  - Development of tutorials should be given more attention. Don't want to oversimplify/'monopolize' on methods - tutorials/easy-to-run workflows that are open to contributions, perhaps a voting/ranking scheme would be worth thinking about. Having to consolidate information from dozens of packages may lead to users giving up on using Bioconductor before learning about all the benefits. Credit/control/methodology stay with the individual developers.
- Cloud infrastructure/on-disk representations of mass spec data – note phantasus package uses HSDS, parquet and TileDb in cloud also relevant.
- Perturb-seq.
- Long read genomics - Matt Ritchie group?
- Long term survival of the project - how long?
- How to justify to developers submitting to Bioconductor - logistical advantages, immediate access to annotation, mentoring.
- Teaching, outreach, conferences - website could highlight much better
- Workflows spanning multiple packages - solving practical problems - Mike Love spearheading a working group.
- GPU computing - Davide/Levi project.

:18 - :35 Tim Triche: pangenome data and methods concepts

- PanVizGenerator: https://bioconductor.org/packages/3.13/bioc/vignettes/PanVizGenerator/inst/doc/panviz_howto.html
- sequenceTubeMap: https://github.com/vgteam/sequenceTubeMap
- Kasper worked with genomes from multiple mouse strains - similar in many ways (e.g. converting coordinates, lots of potential to make this easier with the foundation provided by GenomicRanges).
- One of the reasons for using genome graphs in the alignment phase is to reduce reference bias.
- Previous work by Tim's group driven by a locus of interest with local structural variants.
- Map to genome reference graph.
- Extract the "tangle" or "snarl" (characterizing the genetic variation seen across human populations at a locus).
- At least 6 different paths in a 800 bp region across ~30 patients.
- Possible wasm/webR working group
  - Start from example that users can tweak slightly, e.g. https://trichelab.github.io/webR/mixtures/, https://trichelab.github.io/webR/germline/

    - Expand "Intro to Bioconductor" workshop (from Martin/Lori) to include more classes, integrate wasm.
    - Carpentries curriculum.

:36 - :48 LLM/chatbot training/general AI - open discussion

- How to build assistants that allow people (who may not have the required coding experience) to perform analyses (with R/Bioc) - let the LLM create/iteratively refine an Rmd document that performs a certain analysis. Work in progress at Genentech. Also exploring chunking options - make it easier to search through documentation.
- RAG to generate embeddings.
- Discussions also ongoing in the cloud working group.
- https://techcommunity.microsoft.com/blog/healthcareandlifesciencesblog/genomics--llms-a-case-study-on-adding-variant-annotations-to-llms-through-rag-an/4255228
- https://shiny.posit.co/blog/posts/shiny-assistant/ (knows a lot of Bioconductor as well).
- https://elmer.tidyverse.org/
- Bringing in extra information may "level the playing field" - provide more information about not only the most downloaded/used packages.
- Claude, GitHub CoPilot.
- Direct bots to prefer core Bioc packages when possible, to avoid suboptimal/unnecessarily complex/niche solutions.

:48 - :55 Update on 3.20 production, GitHub-based system, Discourse (support site transition; should be a smooth migration without loss of information) – now that 3.20 is out there is time to work on this.

- Significant diversion the past few days for BiocAsia, with Alex tending to a number of last minute workshop wrinkles. Conference leaders should identify a point person who can help workshop authors meet deadlines and have successful testing ahead of the workshop. The workshop instance is stable, and last minute revisions of containers should not be allowed.
- Form to volunteer packages for GitHub Dev Hosting: https://forms.gle/jSDBdKPE4Bvf73Vv5

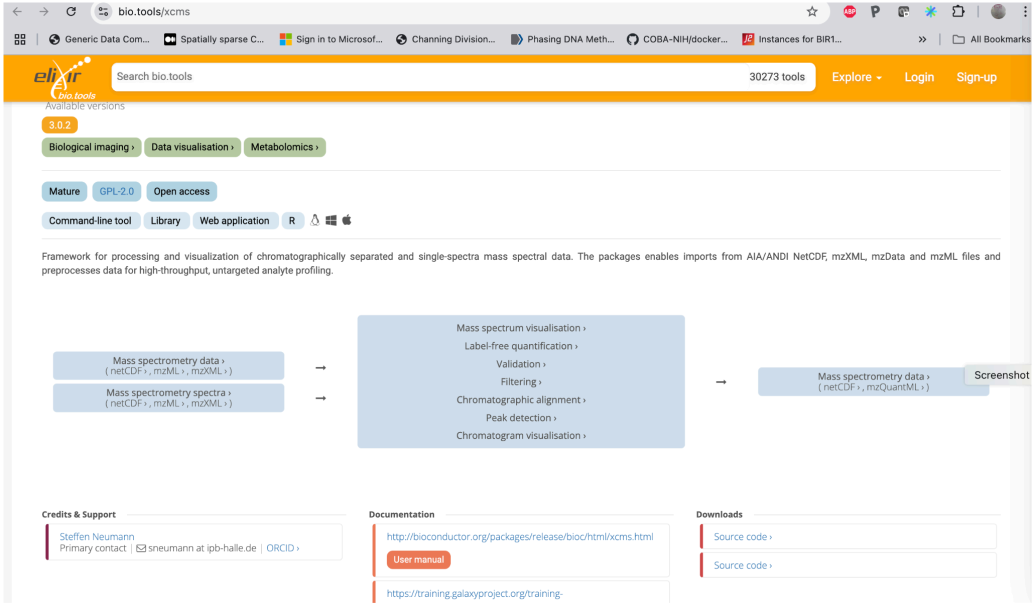
:55 - :60 Open floor

- S7 (https://github.com/RConsortium/S7) - becoming more mature, new release just out. Ready for users to start playing around with. S7 generics can be used with S4 objects.

Other topics (not covered)

- Improving discoverability and schematization of metadata about Bioc packages (Steffen Neumann)
    - At the ongoing ELIXIR hackathon in Barcelona, there has been a lot of investigation of EDAM ontology for Bioconductor packages, with respect to properly annotating ELIXIR's bio.tools repository. Steffen has suggested obtaining metadata that would allow production of a flow graph like the one in the

middle here. Not clear how automatic that is at present.



Framework for processing and visualization of chromatographically separated and single-spectra mass spectral data. The packages enables imports from AIA/ANDI NetCDF, mzXML, mzData and mzML files and preprocesses data for high-throughput, untargeted analyte profiling.

Steffen Neumann
Primary contact | sneumann at ipb-halle.de | ORCID ›

http://bioconductor.org/packages/release/bioc/html/xcms.html

https://training.galaxyproject.org/training-

- Queries:
  - Do we want to pursue a description-custom-field for Bioconductor (e.g., EDAM/topic, EDAM/operation, EDAM/data)?
  - Should we consider a new file that captures detailed operations, inputs, and outputs—similar to the graph in the center of bio.tools xcms?
  - If so, creating-a-new-roclet could be a good approach, rendering all operations, inputs, and outputs to a mypackage/biotools.json file.

- New paradigms for programming statistical algorithms
  - JAX: https://statmodeling.stat.columbia.edu/2024/10/08/defining-statistical-models-in-jax/