BioC 2014 Developer Day Public Notes

Twitter: #BioC2014

Project Updates

Martin Morgan

- Why Bioconductor? S4 (reusability, interoperability, robustness); release / devel branch (user stability, devel flexibility); central repository (coordinate packages into 'Biocondcutor'; definitive repository)
- [KDH]: some interesting stats on usage. (1) One stat had a year-by-year growth: For grants I think this is useful so I think we should collect some (the same) statistics every year so we can easily calculate growth. Answer: The Annual Report contains multi-year data; some statistics have not always been available. (2) Should we to a greater extent encourage the citation of Bioconductor (again, for grants).

Valerie Obenchain

- BiocParallel
- GenomicFiles -- high-level reduceBy functions, e.g., reduceByFile(X, MAPPER, REDUCER)

Herve Pages

- Improvements to sequence infrastructure, especially package organization
- Challenge: mastering relevant classes. Key classes: GRanges, DataFrame, GenomicAlignments, Rle, *List (e.g., IntegerList)

Sonali Arora

- Videos! link in Description: file, hosted on YouTube channel - https://www.youtube.com/user/bioconductor
- biocViews more useful
- GenomeInfoDb for managing sequence level styles

Marc Carlson

- New support site
- Plan: beta in august, 'live' (replacing mailing list) in September
- Concerns: markdown for posting, rather than wysiwyg; changes required versus using a mailing list
- question: will the site also replace bioc-devel? Answer: the current plan is to maintain the Bioc-devel mailing list
- question: will the messages be indexed by google (that would be great)? Answer: yes
- question: can I follow tags using an rss feed? Answer: yes

Dan Tenebaum

- BiocCheck for consistency
- AMI now has StarCluster for provisioning clusters [I forgot to mention that the AMIs are now automatically generated with Vagrant and Chef and can be re-used to provision virtual or physical machines, scripts here: https://github.com/Bioconductor/setup-starcluster-image]

Nate Hayden

- Rsamtools::pileup: https://www.youtube.com/watch?v=Rfon-DQYbWA
    - Question: Is there any consensus on whether (or in which scenarios) to use pileup as opposed to ReadGAlignments/SummarizeOverlaps (or similar workflow)? Answer: pileups are really for nucleotide-level summaries (e.g., variant calling), whereas GAlignments / SummarizeOverlaps is about read-level summaries (e.g., RNA-seq differential expression, ChIP-seq)
- Mavericks C++ compiler / C++11
    - C++/Mavericks Best Practices guide: http://bioconductor.org/developers/how-to/mavericks-howto/
    - [KDH]: the main issue with Mavericks is that BioC is now being compiled by more than 1 compiler (in past we just used different versions of GCC), which always exposes errors (experience from past reported issues in packages I maintain under eg. the Intel compiler). From my POV, this is really different from the C++11 issue.
    - Issue raised that many corporate clusters run other Linux distros such as RHEL5, which have very outdated compilers. Bottom line: shy away from C++11 features unless you want to limit your user base.

Group Activity

Flashlight Talks (5 minutes)

- Di Wu. Rank-based Rotation Gene Set Enrichment Analysis (ROMER), accounting for genewise correlation and dealing with complex designs.
- Ying Shen. ASSIGN: Context-specific and Integrative Genomic Profiling of Heterogeneous Biological Pathways.
- Adam Mark. mygene.R: An R package to access MyGene.info gene query and annotation services. (https://bitbucket.org/sulab/mygene.r)
    - Important to track ‘metadata’, e.g., database version
- Michael Love. Shiny clickable plots.
    - How to make this executable to a non-R collaborator, short of making a *shiny* server? (Vagrant or Docker?)
        - Link to site discussing stand-alone shiny apps: http://blog.analytixware.com/2014/03/packaging-your-shiny-app-as-windows.html (this is only for windows, but should be possible to expand to other platforms)
- Stephanie Hicks. When to use quantile normalization?
    - What to do in the case when the groups are not well-defined (e.g. tumor purity?)
- Leonardo Rodrigues. LAT, streamlining lipid data analysis in R.
    - Question about reusing, e.g., xcms, vs. writing new code; value to re-using as much as possible

- Robert Castelo. Integrated variant annotation and filtering using R/Bioconductor and the VariantFiltering package.
- Jianhong Ou. motifStack for graphic representation of multiple motifs in one canvas.
- Laurent Gatto. Makefile for R packages: maker.
- Houtan Noushmehr. Official Inauguration of Bioconductor in Latin America: LAb.Foundation [presentation]

[KDH]: Didn't we used to limit the number of slides in a flash talk: I think that would be a good idea.

Community Needs

Topics for Discussion

Publications

- [KDH] It might be useful to have a discussion of how and where to publish software papers, with people's experience.
- [SH] Some data: JSS can take up to 2 years to handle papers. Plos ONE has very good visibility.

Integrative analysis

- [DGC] Data integration and Bioconductor: considering the integration of different omics, such as metabolomics and RNA-seq, which classes are we to select use when creating such packages? Two comments: (1) In mRNA analysis there is a lot of development and interoperability, but in other omics is less, so good to think about it. (2) maybe in other omics it is possible to reuse mRNA classes.
    - Don't reinvent classes
    - Interoperability
    - What about data types that are not well-represented
    - Integrative analysis
- Missing domains of analysis -- importance of integration
- One class for all, or separate classes for each data type
- github -- biocMultiAssay class -- Levi / Vince
    - e.g., like TCGA assays, some linked to genome, some not
    - Want container to coordinate across analysis
    - Shared subsetting methods, e.g., gene coordinates
    - In-memory solutions, because at this level data is not that voluminous

user outreach

- User training -- interactive lessons via *swirl* http://swirlstats.com/
- Industry doesn't like working w/ R. Use by non-biologists. Needs to be 'easy' but managing 'power'. GUIs for Bioconductor / other libraries
- Community competitions to spur tool / method development?
- What users are we targeting?
    - Importance of power users
    - Not a real choice? Catering to different audiences. Lots of people developing for Bioc, so room for many approaches

    - Next generation of biologists will *have* to have computing skills
- Re-use existing tools -- *RGalaxy*
- Value of workflows and courses
    - Course material hard to navigate
    - Training for critical application

Visualization

- 'Biologists' want a way to play with the (analyzed) data; presenting complex multi-dimensional data to users
    - Trade-off between 'universal' / comprehensive solutions, vs. standard shiny interactions to GRanges, for e.g. (interactiveDisplay)
- Theme-garden of shiny apps for Bioconductor; widgets

performance

- Profiling and other code development tools, especially for non-coders
- 50 bp windows x 140 samples → v. large matrix that "can't be done" w/ R's memory limit
    - Solution: bedtools via BiocParallel
- Other approaches to 'packaging', e.g., cloud, virtual machine; commercial (AMI), XSEDE

Functionality

- Emerging research areas -- single cell sequencing

Documentation

* data.table, data.frame, DataFrame
* Navigating large object hierarchies

Developer Workshops

2:15- 3:05

- Creating packages (Laurent / Marc) +
- EfficientR (Martin) (From this lab)
- Running R in Batch on Large Clusters the pbdR way (George)
- Using shiny + (JJ)

3:00 - 3:30 Break

3:30 - 4:30

- New package submission process (Marc); Social pkg development with a svn/github bridge (Dan) Yawkey 306
- Writing Rcpp code (Michael I. Love) Yawkey 307 +
  here's some demo code:
  https://gist.github.com/mikelove/a78c8d3ccc0f2b41bd2a

  http://gallery.rcpp.org/
  JJ shows me up: even easier sourcing with sourceCpp():
  http://gallery.rcpp.org/articles/subsetting/
  http://dirk.eddelbuettel.com/code/rcpp/Rcpp-attributes.pdf
  Rcpp::checkUserInterrupt inside loops to allow users to escape safely
  Rcpp parallelization:

https://github.com/RcppCore/RcppParallel
External pointers Rcpp:::XPtr described here:
http://cran.r-project.org/web/packages/Rcpp/vignettes/Rcpp-modules.pdf

- Debugging R, solutions, session transcript (Laurent) ++
- Strategies for Scalable Computing in R (Michael F Lawrence) + in Room: Smith 308
- S4 classes (+); overall classes; NAMESPACE import/export and DESCRIPTION enhances/suggests/depends recommendations for Bioconductor packages (Herve, Val) Yawkey 308

4:30 - 5:00
Suggestions (please add!):

- C++/STL and Mavericks
- Debugging C (Martin)
- Writing C code (Martin)
- Modular GUI development (Thomas)

Please email your slides to Dan (dtenenba@fhcrc.org) so that we can add them to BioC2014 page here:
http://www.bioconductor.org/help/course-materials/2014/BioC2014/
(He sent out an email last week- just reply to that)

Developer Day Address

RStudio
REVOLUTION ANALYTICS
Genentech
A Member of the Roche Group
NOVARTIS
PHARMACEUTICALS
FRED HUTCHINSON CANCER RESEARCH CENTER
A LIFE OF SCIENCE