The *Bioconductor* Project: Current Status

Martin Morgan

Roswell Park Cancer Institute
Buffalo, NY, USA
martin.morgan@roswellpark.org

17 November, 2017

Bioconductor

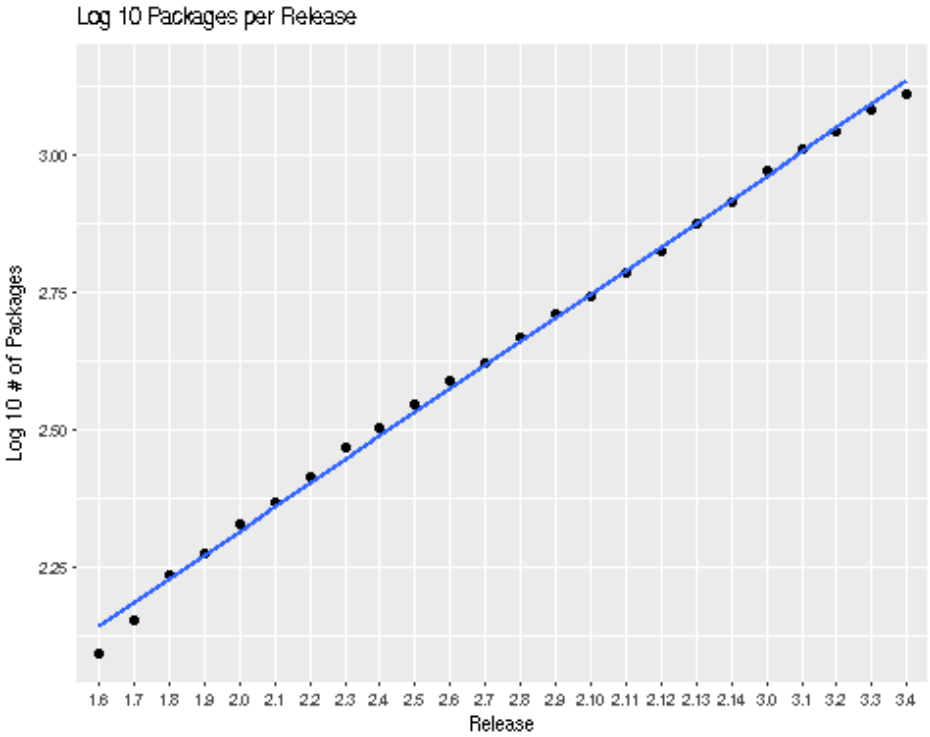
Analysis and comprehension of high-throughput genomic data.

- Started 2002
- 14736 *R* packages – developed by 'us' and user-contributed.

Well-used and respected.

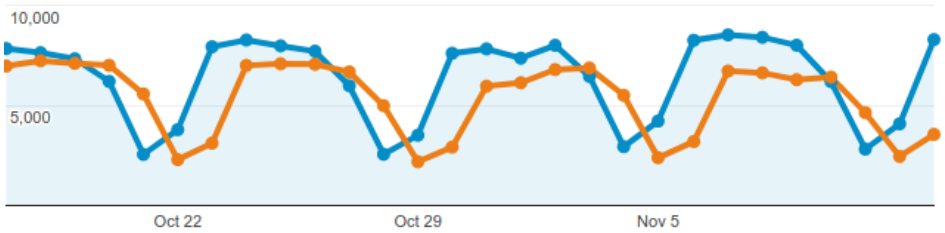
- 53k unique IP downloads / month.
- 21,700 PubMedCentral citations.

State of the project



- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

State of the project



1.	United States	**58,384** (32.78%)
2.	China	**20,910** (11.74%)
3.	United Kingdom	**12,265** (6.89%)
4.	Germany	**10,024** (5.63%)
5.	France	**5,536** (3.11%)
6.	Canada	**4,999** (2.81%)
7.	Spain	**4,864** (2.73%)
8.	Japan	**4,539** (2.55%)
9.	India	**4,397** (2.47%)
10.	Australia	**4,043** (2.27%)

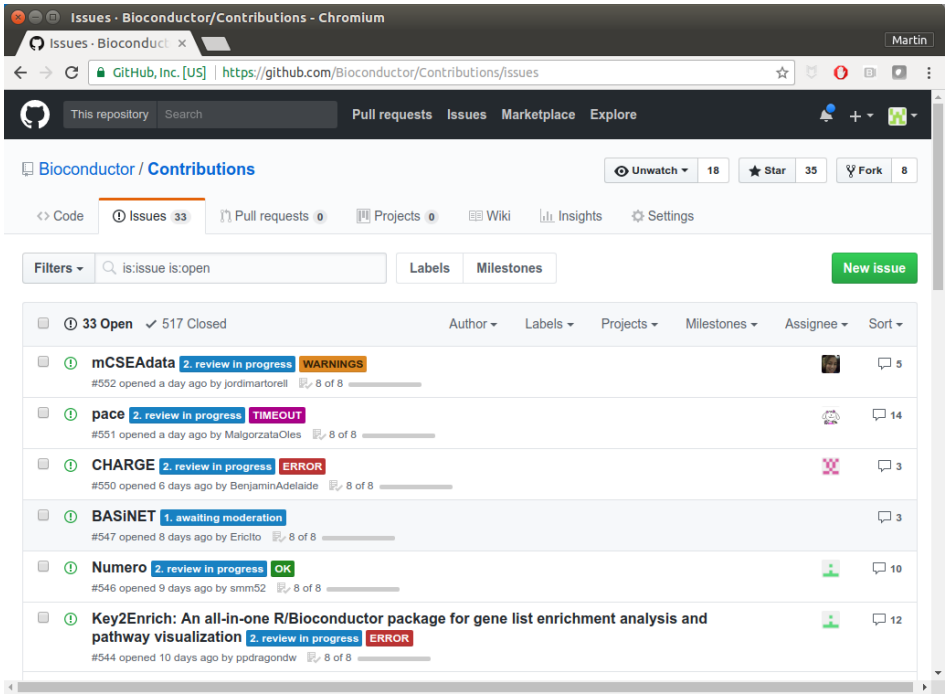
- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

State of the project

https://bioconductor.org
https://support.bioconductor.org

- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

State of the project



- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

State of the project

- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

Recent developments

- Git!

```
git clone https://git.bioconductor.org/packages/limma
git clone git@git.bioconductor.org:packages/DESeq2
```

- Large Single Cell
  - *SingleCellExperiment*
  - *HDF5Array*
- Lessons from 100's of package reviews

Large single-cell data

```
> sce = TENxBrainData::TENxBrainData()
snapshotDate(): 2017-10-30
> sce
class: SingleCellExperiment
dim: 27998 1306127
metadata(0):
assays(1): counts
rownames: NULL
rowData names(2): Ensembl Symbol
colnames(1306127): AAACCTGAGATAGGAG-1 AAACCTGAGCGGCTTC-1 ...
  TTTGTCAGTTAAAGTG-133 TTTGTCATCTGAAAGA-133
colData names(4): Barcode Sequence Library Mouse
reducedDimNames(0):
spikeNames(0):
```

Large single-cell data

- Chunk-wise iteration (often transparent to the user / developer).
- Marginal summaries in `rowData`, `colData`.
- Supporting infrastructure: *ExperimentHub*, *rhdf5*, *HDF5Array*, *DelayedMatrixStats*, *beachmat*.

Why use or contribute to *Bioconductor*?

- Recognition.
- Access & Permanance.
- Interoperability.
- Documentation.
- Support.
- Tested.

Why use or contribute to *Bioconductor*?

- Recognition.
- Access & Permanance.
- Interoperability.
- Documentation.
- Support.
- Tested.

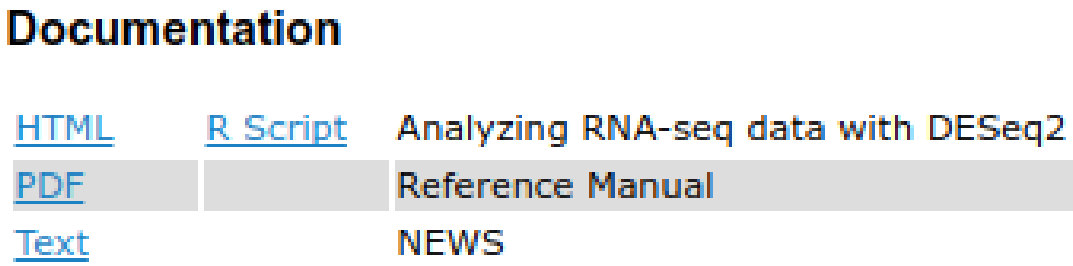
Why use or contribute to *Bioconductor*?

- Recognition.
- Access & Permanance.
- Interoperability.
- Documentation.
- Support.
- Tested.

```
git$ grep -l SummarizedExperiment \
  */DESCRIPTION | wc -l
165
```

Why use or contribute to *Bioconductor*?

- Recognition.
- Access & Permanance.
- Interoperability.
- Documentation.
- Support.
- Tested.

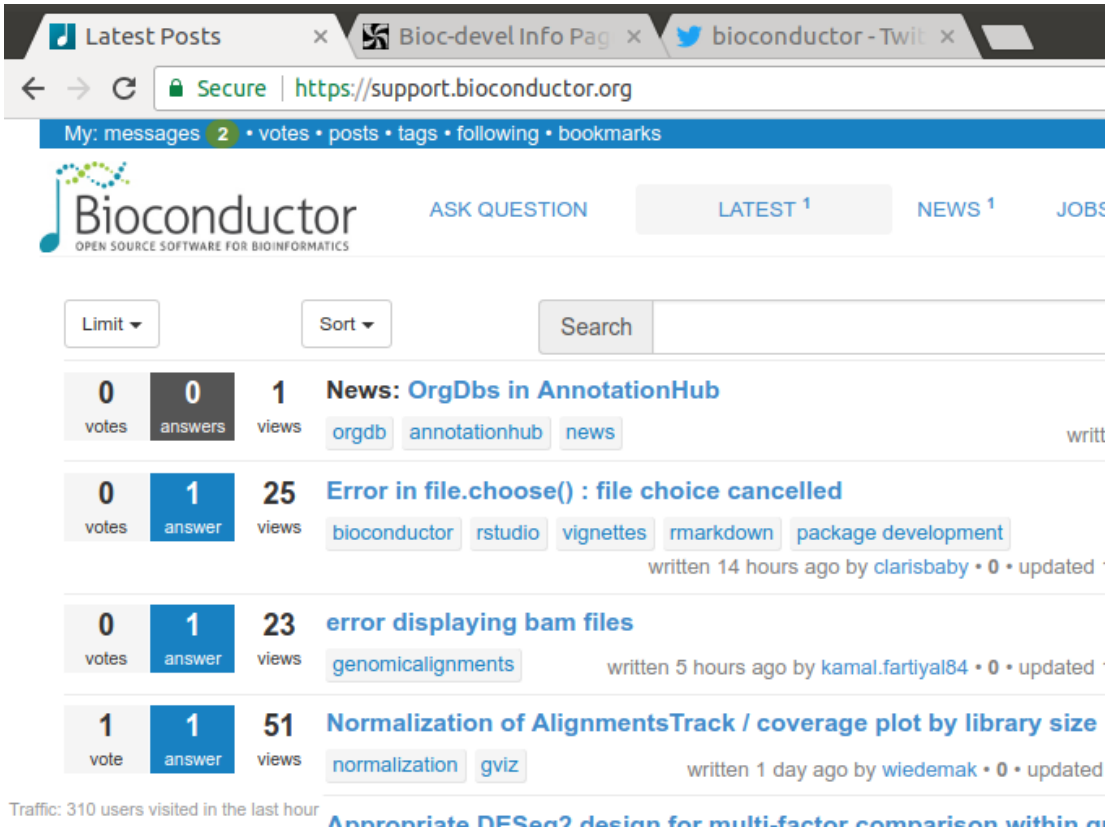


Documentation

HTML	R Script	Analyzing RNA-seq data with DESeq2
PDF		Reference Manual
Text		NEWS

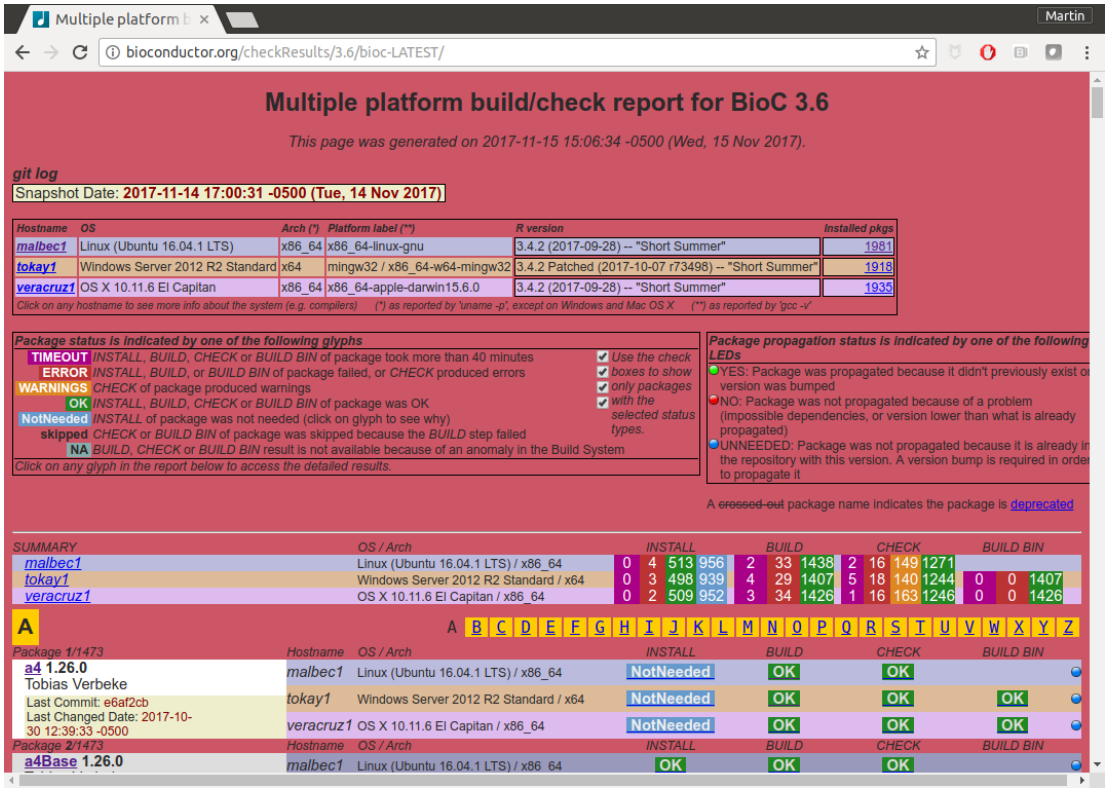
Why use or contribute to *Bioconductor*?

- Recognition.
- Access & Permanance.
- Interoperability.
- Documentation.
- Support.
- Tested.



Why use or contribute to *Bioconductor*?

- Recognition.
- Access & Permanance.
- Interoperability.
- Documentation.
- Support.
- Tested.



Lessons learned from package reviews I

1. Interoperability
   - Use feature $\times$ sample `SummarizedExperiment`, not sample $\times$ feature matrix
   - Use paradigms familiar to *Bioconductor* users
2. Reuse
   - Use `rtracklayer::import.bed()`, not custom parser
3. Robust code
   - Edge cases: `seq_len()` / `seq_along()`, not `1:n`
   - Code complexity: `vapply()`, not `sapply()`
4. Performant code
   - *Vectorize* rather than *iterate* (`for`, `lapply()`, `apply()` are all iterative).
   - Reuse (e.g., *matrixStats* before C / C++ implementation.

Lessons learned from package reviews II

5. Tested code
   - Essential: evaluated example and vignette code chunks.
   - Desireable: unit tests, e.g., *testthat*.
6. Time and space limits.
   - Excessive computation may represent inefficient code.
   - Challenging to identify rich but modest data for illustration.
   - Experiment data packages, work flows, F1000 papers as venues for more expensive / comprehensive reproducible analysis.
7. Ambition
   - Implement essential features well
   - Avoid dependencies on packages for marginal value

Future challenges

- Large data.
- Cloud. Possible visions:
  - As now, but 'in the cloud' – `https://rstudio.cloud`.
  - Integrated with 'third party' compute efforts, e.g., NCI, NIH in the United States.
  - Pay-as-you-play – use existing *Bioconductor* AMIs or docker containers.

Acknowledgments

Core team: Qian Liu, Valerie Obenchain, Hervé Pagès, Marcel Ramos, Lori Shepherd, Nitesh Turaga, Daniel van Twisk.

Technical advisory board: Vincent Carey, Kasper Hansen, Wolfgang Huber, Robert Gentleman, Rafael Irizzary, Levi Waldron, Michael Lawrence, Sean Davis, Aedin Culhane

Scientific advisory board: Vincent Carey (Brigham & Women's), Wolfgang Huber (EBI), Rafael Irizzary (Dana Farber), Jan Vitek (Northeastern University), Robert Gentleman (23andMe).

Research reported in this presentation was supported by the National Human Genome Research Institute and the National Cancer Institute of the National Institutes of Health under award numbers U41HG004059 and U24CA180996. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.