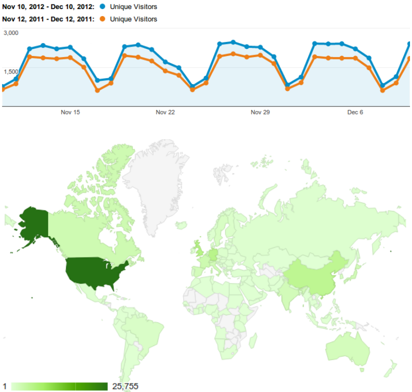
Bioconductor: State of the Project

Martin Morgan (mtmorgan@fhcrc.org)
Fred Hutchinson Cancer Research Center

13-14 December 2012

Project Status



Google analytics, 1-month access, 10 December 2012

Doing great!

- Packages
    - 109 packages added, 10 removed in last 12 months
    - October: downloads to 19,591 distinct addresses[1]
- Web traffic
    - 35k visitors / month, 20% more than last year
    - International participation
    - 'Hot-spots'
- Citations
    - 72 PubMed citations for 'Bioconductor' in 2012

---

[1] `http://bioconductor.org/packages/stats/`

Project Status – Seattle Staff

Marc Carlson	New packages, annotation, *GenomicFeatures*, *uniprot.ws*
Valerie Obenchain	*VariantAnnotation*, *ensemblVEP*, *GenomicRanges* and friends
Hervé Pagès	*IRanges*, *GenomicRanges*, *Biostrings*, etc.
Paul Shannon	Biological networks, *MotifDb*, *RCytoscape*. Mentored projects
Dan Tenenbaum	Build system, AMI, *RGalaxy*, . . .

Birth, Life and Death of a Package

Birth

- A scientifically sound idea
- Packaged according to *Bioconductor* guidelines[2]
  - Class use and re-use
  - Vignettes and examples
  - Increasingly: unit tests[3]
- Submitted to *Bioconductor* for technical review[4]
  - Automatic cross-platform build & check

http://www.dec.ny.gov/pubs/43776.html

---

[2] http://bioconductor.org/developers/package-guidelines/

[3] http://bioconductor.org/developers/unitTesting-guidelines/

[4] http://bioconductor.org/developers/package-submission/

Birth, Life and Death of a Package

Early life

- Added to our subversion repository[5] and to the 'devel' branch, as version 0.99.z
- `R CMD build`, `R CMD check` run nightly on Windows, Mac, and Linux
  - Using devel version of *R* and other packages
  - Build report available[6]
- Package 'landing page' available[7]
- New versions (changes in 'z' part of version number) made available via `biocLite` to 'devel' users

---

[5] http://bioconductor.org/developers/source-control/

[6] http://bioconductor.org/checkResults/devel/bioc-LATEST/

[7] http://bioconductor.org/packages/devel/

Birth, Life and Death of a Package

Later life

- At next release (∼ every 6 months), 'devel' branched to becomes release – version 1.0.0!
- Bug fixes to release branch
- New development in devel version 1.1.0
- Monitor mailing list[8] to support users

---

[8] `http://bioconductor.org/help/mailing-list/`

Birth, Life and Death of a Package

Death of a package

- Original author no longer able to maintain package, no one from community willing to step forward. E.g., *cosmo*
- Ideally, 'deprecated' for one release cycle before removal

R is for Rhoda consumed by a fire

The Gashlycrumb Tinies, Edward Gorey

Highlights from the Core Team

1. *BiocInstaller*: installation, updates, & upgrades
2. *BiocGenerics*: avoid name conflicts for S4 generics
3. Easier annotations: `select`; *Homo.sapiens* and friends; *AnnotationHub*
4. *MotifDb* & gene regulation
5. Variants: *VariantAnnotation*, *ensemblVEP*
6. *GenomicRanges* and friends: many changes. Metadata column accessors `mcols`, `$`; *GappedAlignmentPairs*, `encodeOverlaps`; *Hits* to represent overlaps; discouraging use of *RangedData*
7. ‘Cloud’ facilities: Amazon machine instance[9]; *RGalaxy*
8. And... mentored projects; *Rgraphviz*; ...

---

[9] `http://bioconductor.org/help/bioconductor-cloud-ami/`

BiocGenerics

Rationale

- Avoid creating multiple generics for a single *R* function

Use `duplicated,GeomicRanges-method`

- DESCRIPTION file: `Imports: GenomicRanges`
- NAMESPACE file:

  ```
  importMethodsFrom(GenomicRanges, duplicated)
  ```

Define a new method `duplicated,MyClass-method`

- DESCRIPTION file: `Imports: BiocGenerics`
- NAMESPACE file:

  ```
  importFrom(BiocGenerics, duplicated)
  exportMethods(duplicated)
  ```

- And in a file in the `R` directory:

  ```
  setMethod(duplicated, "MyClass",
      function(x, incomparables=FALSE, ...) {})
  ```

Annotation

```
library(Homo.sapiens)

## discovery
cols(Homo.sapiens)              # 52 columns of data
keytypes(Homo.sapiens)          # 27 ways to query

## query
keys <- c("NOVA1", "NOVA2")     # 'SYMBOL' keys
cols <- c("TXNAME", "GENENAME")
anno <- select(Homo.sapiens, keys, cols, "SYMBOL")

## to come: GRanges / GRangesList
exByTx <- exonsBy(Homo.sapiens, "tx", keys, "SYMBOL")
```

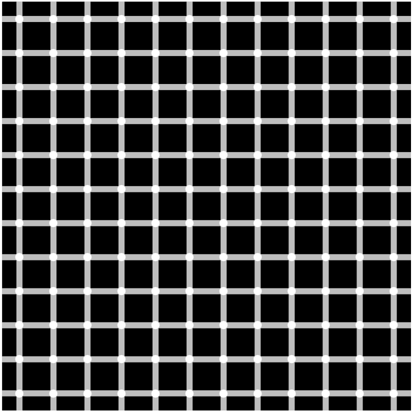
AnnotationHub (work in progress)

```
library(AnnotationHub)
hub <- AnnotationHub() # web query
names(hub)             # available resources; many
df <- metadata(hub); names(df)
hub$pub<tab>

gr <- hub$pub.release.69.gtf.danio_rerio.
    Danio_rerio.Zv9.69.gtf.RData

gr <- hub$goldenpath.hg19.encodeDCC.wgEncodeRikenCage.
    wgEncodeRikenCageCd20CellPapTssHmm.
    bedRnaElements.RData
```

Opportunities



http://rpubs.com/kohske/R-de-illusion

- Flexible vignettes (e.g., markdown; *knitr*) and documentation (*roxygen2*)
- Use of *git* and github[10] to augment subversion; use of social media, e.g. twitter[11]
- Accessible documentation
- *BiocParallel* for more transparent, standardized access to parallel evaluation

---

[10] https://github.com/Bioconductor

[11] https://twitter.com/bioconductor