What is Bioconductor?

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Principally a collaborative software development project

But it is also:

- a software repository
- a bioinformatics support site
- data repository
- publisher for supplementary materials
- source for tutorials and instructional documentation

Managed and maintained by a core team of 6 people, with contributions coming from all over the world

What is Bioconductor?

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Principally a collaborative software development project

But it i

- a so
- a bio
- data
- publ
- sour

Manag
with

What is Bioconductor ?

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

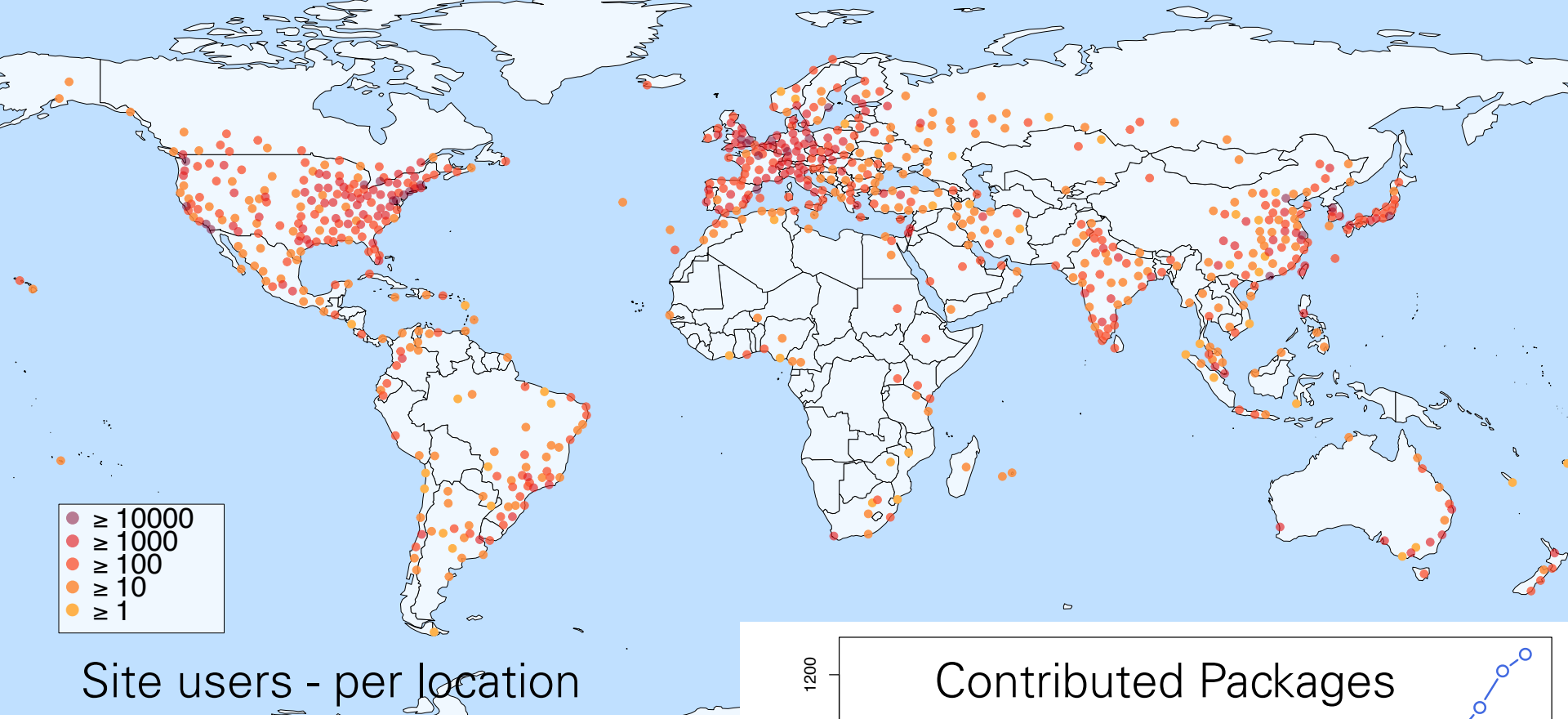
Started 2002 as a platform for analysis & understanding of microarray data

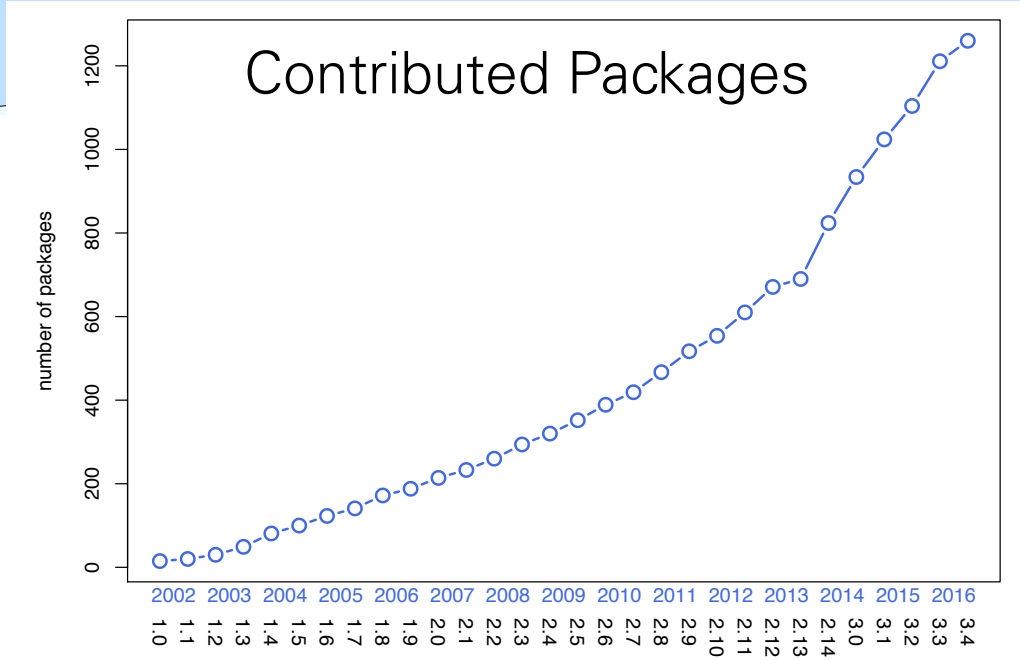
More than 1,300 packages. Domains of expertise:

- Sequencing (RNASeq, ChIPSeq, single-cell, called variants, ...)
- Microarrays (methylation, expression, copy number, ...)
- Flow cytometry
- Proteomics
- Multi-Omics data integration

Important themes

- Reproducible research
- Interoperability between packages & workflows
  … even from different authors
- Usability





World largest bioinformatics project

10,000s users

>18,000 papers in PubmedCentral

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Data import,
preprocessing
Integration of data types
Based on R

Collaborative
and distributed development
Open source

Interoperable
components
Rapid development
Code re-use

Lower barrier of entry
Training
Turn users into
developers

Publication of software
Computational reproducibility

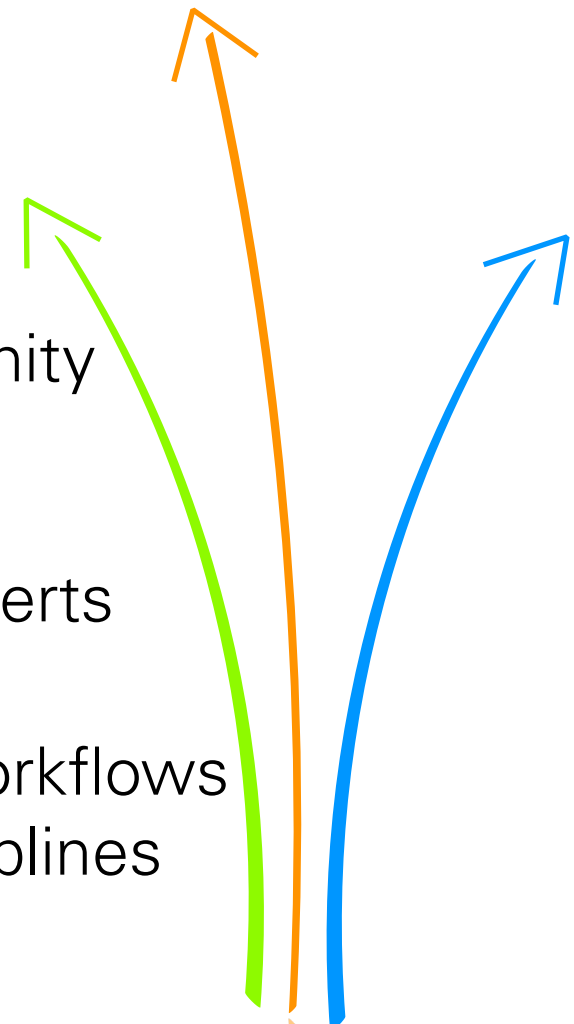
Motivating principles

→ Provide a compelling user experience: documentation, demos, tutorials

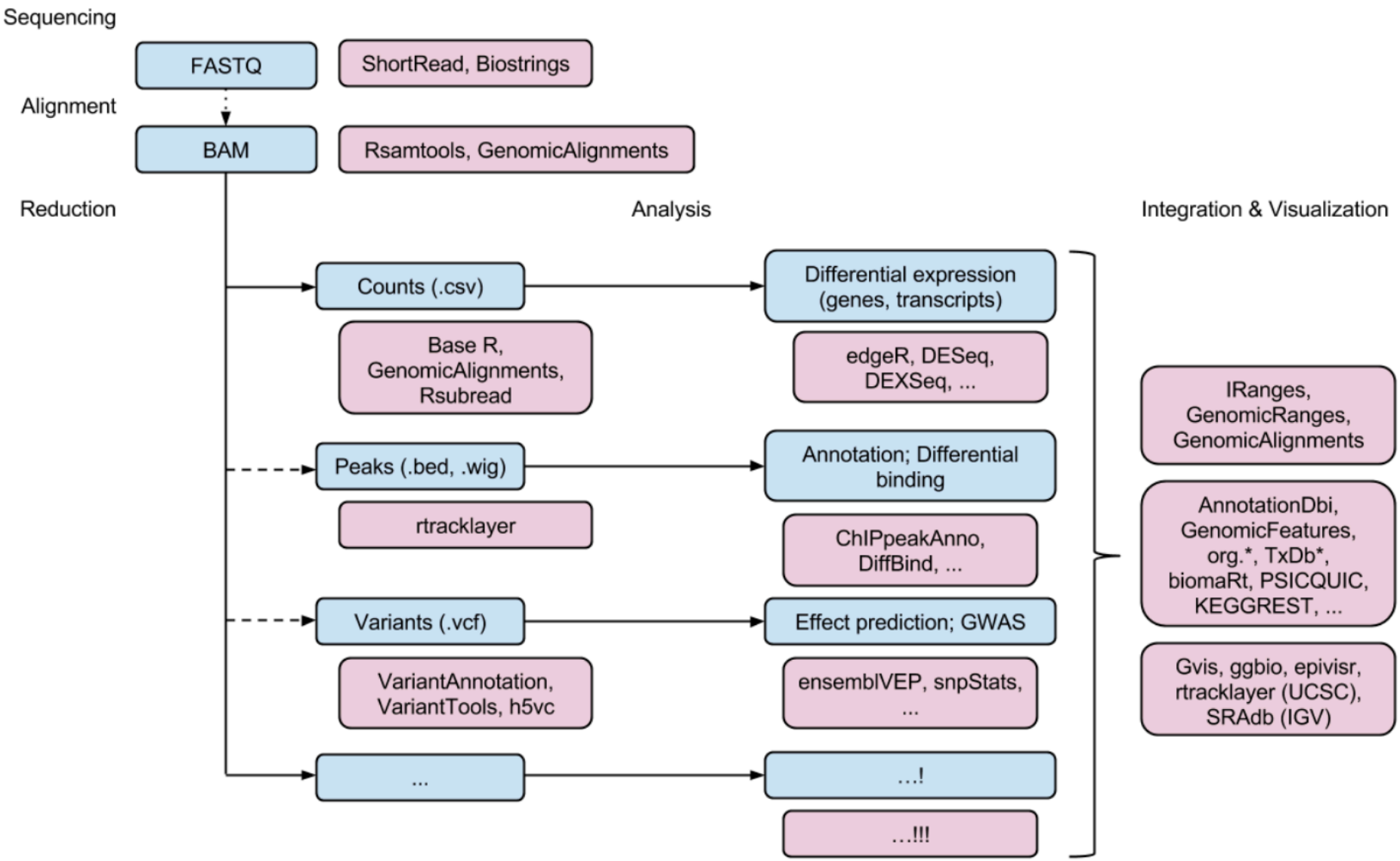
- workflows
- package vignettes
- function manual pages

→ Support active & open developer community

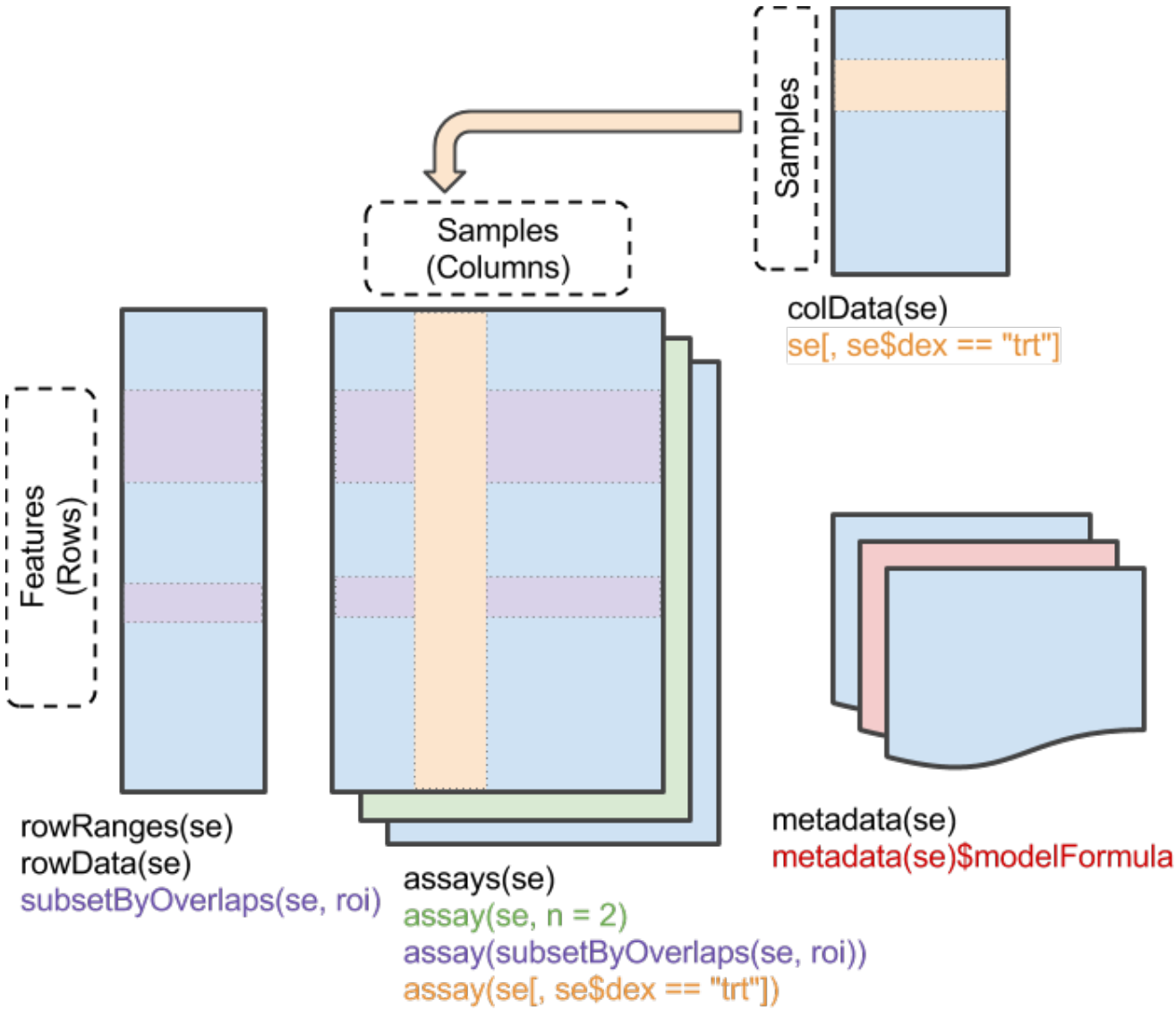
- training on software development & programming techniques
- distributed development by domain experts (→interoperability)
- common data structures that enable workflows integrating multiple data types and disciplines



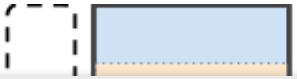
Workflows for HT Sequencing



Bioconductor Classes: e.g., Summarized Experiment



Bioconductor Classes: e.g., Summarized Experiment



- Validity checking: enforces contracts with the user
- Synchronized handling of multiple tables / matrices (subsetting, dangling pointers)
- Encapsulation: separation of interface from implementation
- Specialized highly efficient methods for manipulation (e.g. GRanges class)

```
subsetByOverlaps(se, roi)
assays(se)
assay(se, n = 2)
assay(subsetByOverlaps(se, roi))
assay(se[, se$dex == "trt"])
```

Annotation & Datasets

Annotation of genes, transcripts, proteins, pathways, metabolites; GO, Reactome, Pubmed, …

Sequences

You don't have to download text files from NCBI / EBI and parse them into R - use ready made packages with nice interfaces.

ExperimentHub: published datasets already curated into efficient R objects, with documentation

Modes of documentation

- Manual Pages (for each function)
- Vignettes: Narrative overviews on what you can do with a package
- Workflows: end-to-end descriptions of a scientific question
- F1000Research papers, Bioinformatics application notes: peer-reviewed, citable

Support Forum - http://support.bioconductor.org

Latest Posts

Secure | https://support.bioconductor.org

My: messages • votes • posts • tags • following • bookmarks

Mike Smith • 1.7k | log out • about • faq • rss

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ASK QUESTION | LATEST | NEWS | JOBS | TUTORIALS | TAGS | USERS

Limit | Sort | Search

votes	answers	views		
0	1 answer	51	Testing for DE genes	limma, differential gene expression, affymetrix microarrays — written 6 hours ago by Lennart.Vermander • 0 • updated 6 hours ago by Aaron Lun • 14k
0	0 answers	21	read.maimages(targets, source="genepix") Error in `[.data.frame`(obj, , columns[[a]]) : undefined columns selected	limma, bioconductor, agilent microarrays — written 7 hours ago by sankhwar.madhu • 0
2	1 answer	45	Feature request: syntax for dropping unused seqlevels in GRanges objects	seqlevels, granges, feature request — written 1 day ago by Charles Plessy • 0 • updated 14 hours ago by Hervé Pagès ♦♦ 12k
0	0 answers	28	Job: Research Fellow or Junior Research Fellow, Statistical Genetics and Genomics, Australia	genetics, job, biostatistics — written 14 hours ago by hong.lee • 0
0	0 answers	33	Bioconductor for predicting transcription factors from a gene list?	transcription, transcriptdb — written 18 hours ago by hkarakurt • 0
0	0 answers	30	DESeq2 DESeqDataSetFromTximport design formula	deseq2 — written 18 hours ago by stephensmith • 0
0	0 answers	26	check splice isoforms	deseq2, tximport — written 22 hours ago by ry82722 • 0
5	2 answers	63	Experiment design contrasts any time edgeR	rnaseq, edger, rna, dge — written 1 day ago by delfino.pietro • 10 • updated 1 day ago by londonflavido • 0
0	0	37 views	binnedaverage - is it possible to add an na.rm option?	genomicranges — written 1 day ago by Janet Young • 730

Traffic: 123 users visited in the last hour

Recent...

Replies
- C: Testing for DE genes by Aaron Lun • 14k
- C: Testing for DE genes by Lennart.Vermander • 0
- A: Testing for DE genes by Aaron Lun • 14k
- A: Feature request: syntax ... by Hervé Pagès ♦♦ 12k
- C: Probeset ID's to entrez ... by akhattri • 0

Votes
- recount counts in the examp...
- [GenomicRanges] subsetByOve...
- normalization with mas5
- Preprocessing of Human Gene...
- Preprocessing of Human Gene...

Awards • All »
- Scholar to Hervé Pagès ♦♦ 12k
- Scholar to Aaron Lun • 14k
- Teacher to Aaron Lun • 14k
- Teacher to James W. MacDonald ♦ 43k
- Popular Question to Rory Stark • 1.7k
- Popular Question to Michael Love ♦ 11k

Locations • All »
- Switzerland/Lausanne/EPFL, 3 minutes ago
- Germany, 40 minutes ago

Support Forum -

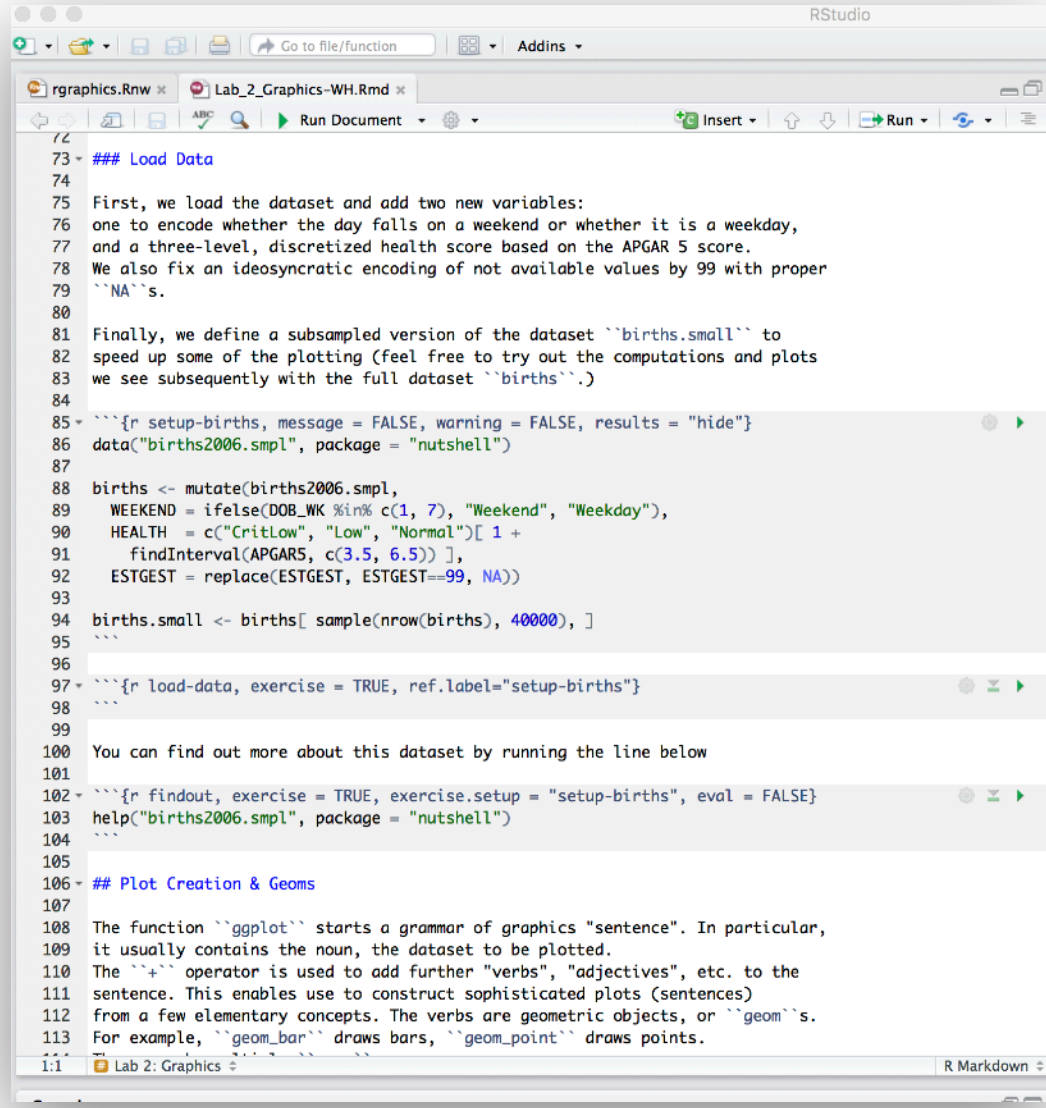
Etiquette (Posting Guide)

- Make sure you use most recent versions
- Read the documentation
- Use Google to see if a similar question has already been asked
- Prepare a minimal working example and post its code
- Remember your manners when reporting "bugs" or "missing features"
- Use descriptive subject line and precise language
- Post 'devtools::session_info()

https://www.bioconductor.org/help/support/posting-guide/

Scientific software should be assessed by similar criteria as a scientific publication

- Reproducible
- Peer-reviewed
- Easy to access by other researchers & society
- Builds on the work of others
- Others will build their work on top of it



Code re-use

Writing good software is hard

Existing, well-used and maintained software contains fewer bugs

Common problems are already solved

Avoid re-implementation — produce interfaces

Focus on new things

→ Lots of package interdependencies (>1000 packages, 100s developers)

Don’t reinvent the wheel

Shared code base, maintained by core team

Bioconductor already has code to:

- Read common file formats
- Represent common data types e.g. Genomic Ranges, Summarized Experiments
- Load genomes and annotation
- etc.

Let users become developers

What are the benefits from using the Bioconductor development environment?

Standardised and powerful data structures for representing datasets incl. metadata

Many tools for data I/O and preprocessing. Access to databases of primary data and annotation (ExperimentHub)

Support for writing good documentation

Support for supporting your users

What are the benefits from using the Bioconductor development environment?

Free code review

Package system

Daily checks - continuous integration

Version control system

Release & devel branches

Six monthly release cycle

Stable version for most users, but easy to make new features public

Why R?

- high-level, interpreted programming language
- rapid prototyping, creativity, flexibility and reproducibility
- scientific and statistical computing capabilities
- publication quality graphics system ('grammar of graphics')
- convenient data I/O & wrangling
- mature package management system
- inter-language interfaces (C, C++, Java, JavaScript)
- lots of momentum with recent language innovations (RStudio, tidyverse, Jupyter, commercial adaptations, …)

LISP/Scheme inside