R / *Bioconductor* for Everyone

Martin Morgan[1]
Fred Hutchinson Cancer Research Center
Seattle, WA

18 November 2013

[1] mtmorgan@fhcrc.org

Outline

R and *Bioconductor*

R

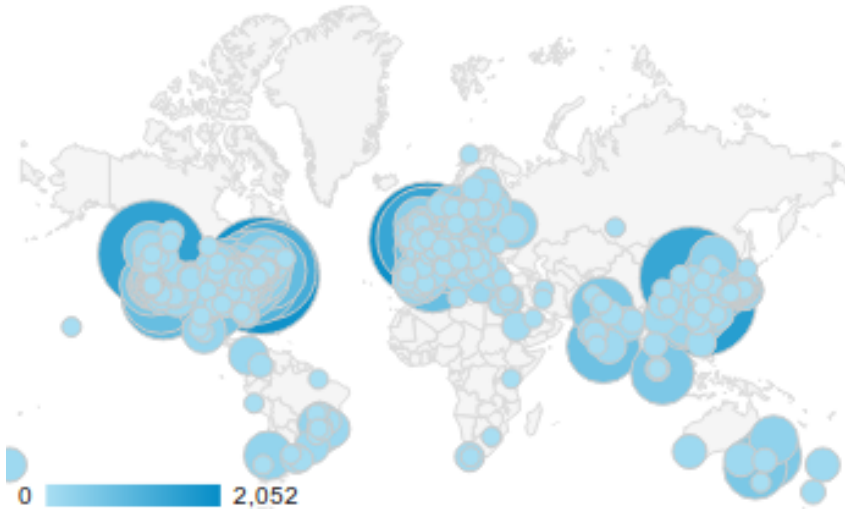
- http://r-project.org
- Open-source, statistical programming language; widely used in academia, finance, pharma, . . .
- Core language, 'base' and $>$ 5000 contributed packages
- Interactive sessions, scripts, packages

Bioconductor

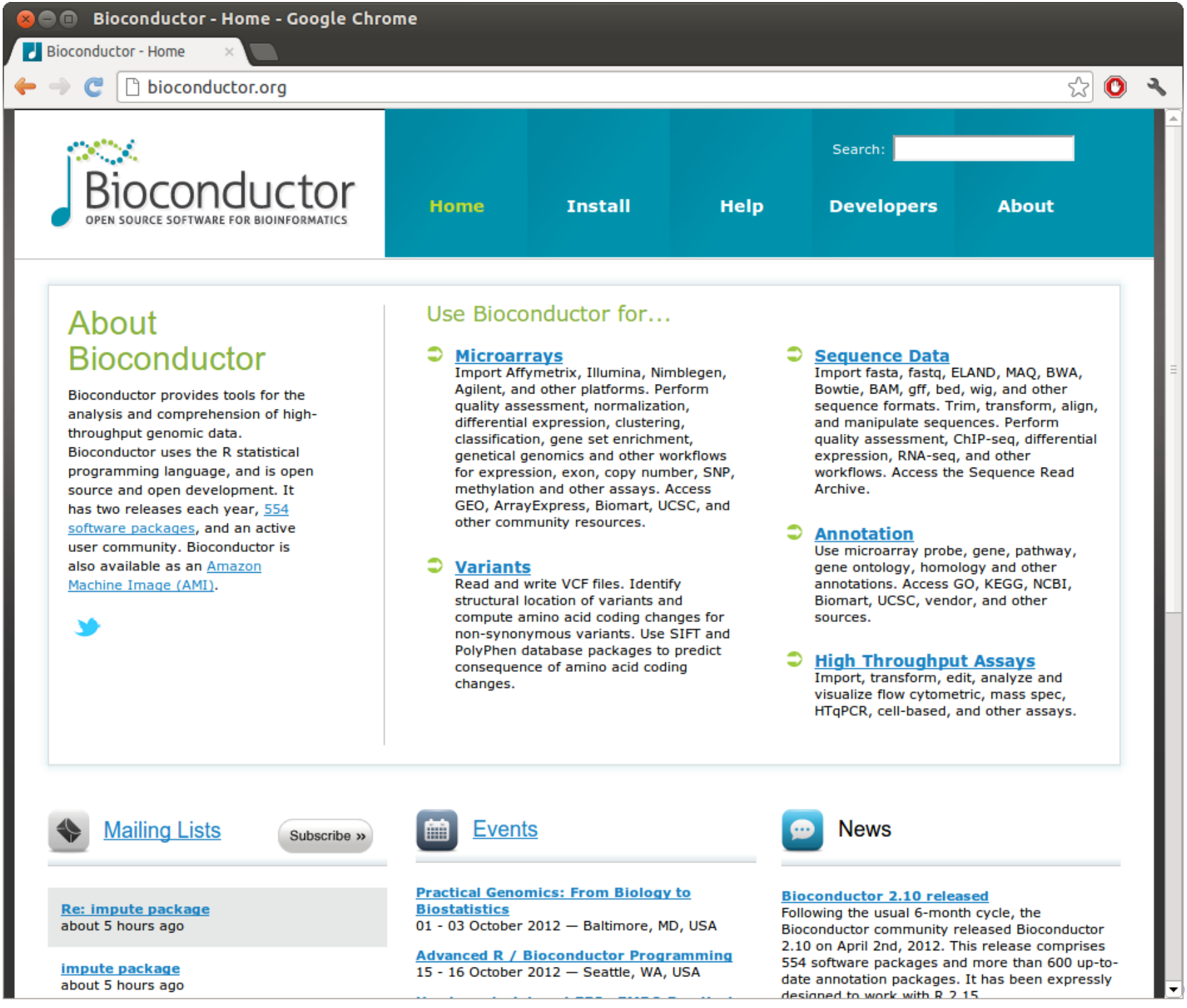
- http://bioconductor.org
- Analysis and comprehension of high-throughput genomic data
- Themes: rigorous statistical analysis; reproducible work flows; integrative analysis
- $>$ 11 years old, 749 packages

Bioconductor

- 1341 PubMed full-text citations in trailing 12 months
- 28,000 web visits / month; 75,000 unique IP downloads / year
- Annual conferences; courses; active mailing list; ...



European Bioconductor Conference, December 9-10, Cambridge, UK


Bioconductor - Home - Google Chrome
Bioconductor - Home
bioconductor.org
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Search:
Home
Install
Help
Developers
About
About Bioconductor
Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 554 software packages, and an active user community. Bioconductor is also available as an Amazon Machine Image (AMI).
Use Bioconductor for...
Microarrays
Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.
Variants
Read and write VCF files. Identify structural location of variants and compute amino acid coding changes for non-synonymous variants. Use SIFT and PolyPhen database packages to predict consequence of amino acid coding changes.
Sequence Data
Import fasta, fastq, ELAND, MAQ, BWA, Bowtie, BAM, gff, bed, wig, and other sequence formats. Trim, transform, align, and manipulate sequences. Perform quality assessment, ChIP-seq, differential expression, RNA-seq, and other workflows. Access the Sequence Read Archive.
Annotation
Use microarray probe, gene, pathway, gene ontology, homology and other annotations. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.
High Throughput Assays
Import, transform, edit, analyze and visualize flow cytometric, mass spec, HTqPCR, cell-based, and other assays.
Mailing Lists
Subscribe »
Re: impute package
about 5 hours ago
impute package
about 5 hours ago
Events
Practical Genomics: From Biology to Biostatistics
01 - 03 October 2012 — Baltimore, MD, USA
Advanced R / Bioconductor Programming
15 - 16 October 2012 — Seattle, WA, USA
News
Bioconductor 2.10 released
Following the usual 6-month cycle, the Bioconductor community released Bioconductor 2.10 on April 2nd, 2012. This release comprises 554 software packages and more than 600 up-to-date annotation packages. It has been expressly designed to work with R 2.15

Why use *R* / *Bioconductor*?

Hallmarks of effective computational software

1. Extensive: data, annotation, integration
2. Statistical: volume, technology, experimental design
3. Reproducible: long-term, multi-participant science
4. Leading edge: novel, technology-driven
5. Accessible: affordable, transparent, usable

Outline

Basic data types

- Vectors of *logical*, *integer*, *numeric*, *complex*, *character*, or *raw* types
- Statistical concepts such as *factor*, `NA`
- More complicated data structures: *data.frame*, *matrix*, *list*
- Object-oriented classes – 'S3' and 'S4' systems

```
> df <- data.frame(
+           age = c(27, 32, 19),
+           sex = factor(c("Male", "Female", NA)))
> df

  age    sex
1  27   Male
2  32 Female
3  19   <NA>
```

Functions

- Act on *vectors*
- Required and / or optional arguments
- Matching by name, then position

```
> y <- 5:1     # vector: 5, 4, 3, 2, 1
> log(y)       # log of each element, 'vectorized'

[1] 1.6094379 1.3862944 1.0986123 0.6931472 0.0000000

> args(log)    # discovery; argument 'base' has default

function (x, base = exp(1))
NULL

> log(base=2, y) # match by name, then position

[1] 2.321928 2.000000 1.584963 1.000000 0.000000
```

Classes and methods

- Coordinate complicated data
- *methods* specialize functions; *accessors*

```
> x <- rnorm(1000, sd=1); y <- x + rnorm(1000, sd=.5)
> fit <- lm(y ~ x); class(fit)
[1] "lm"
> head(methods(class=class(fit)), 3)
[1] "add1.lm"  "alias.lm" "anova.lm"
> anova(fit)
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x           1 1053.44 1053.44  3876.3 < 2.2e-16 ***
Residuals 998  271.22    0.27
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.0
```

S4 classes and methods

- S4 a more formal class system, used extensively in *Bioconductor*

```
> library(Biostrings)
> dna <- DNAStringSet(c("AACA", "ATTA"))
> ## showMethods(class=class(dna),
> ##     where=search())
> alphabetFrequency(dna, baseOnly=TRUE)
```

```
     A C G T other
[1,] 3 1 0 0     0
[2,] 2 0 0 2     0
```

Packages

- Core and contributed; many
- Technical standards imposed, e.g., *man* page for each exposed function, *Bioconductor* vignettes, examples
- Considerable room for author personality, quality variation
- `biocLite` to install a new package, once only
- `library` to attach an installed package

Installation – once only

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("ShortRead")  # install 'ShortRead' package
> biocLite()             # update all installed packages
> library(ShortRead)     # attach to current session
```

Help

```
> help.start()
> ?data.frame
> ?anova
> ?anova.lm     # anova generic, method for class lm
> class ? DNAStringSet
> method ? "alphabetFrequency,DNAStringSet"
> vignette("GenomicRangesIntroduction", "GenomicRanges")
> help(package="Biostrings")
> RShowDoc("R-intro")
```

Useful functions

`dir`, `read.table`, `scan` List files; input data.

`c`, `factor`, `data.frame`, `matrix` Create vectors, etc.

`summary`, `table`, `xtabs` Summarize or cross-tabulate data.

`t.test`, `lm`, `anova` Compare two or several groups.

`dist`, `hclust` Cluster data.

`plot` Plot data.

`ls`, `library` List objects; attach packages.

`lapply`, `sapply`, `mapply` Apply function to elements of lists.

`match`, `%in%` find elements of one vector in another.

`split`, `cut` Split or cut vectors.

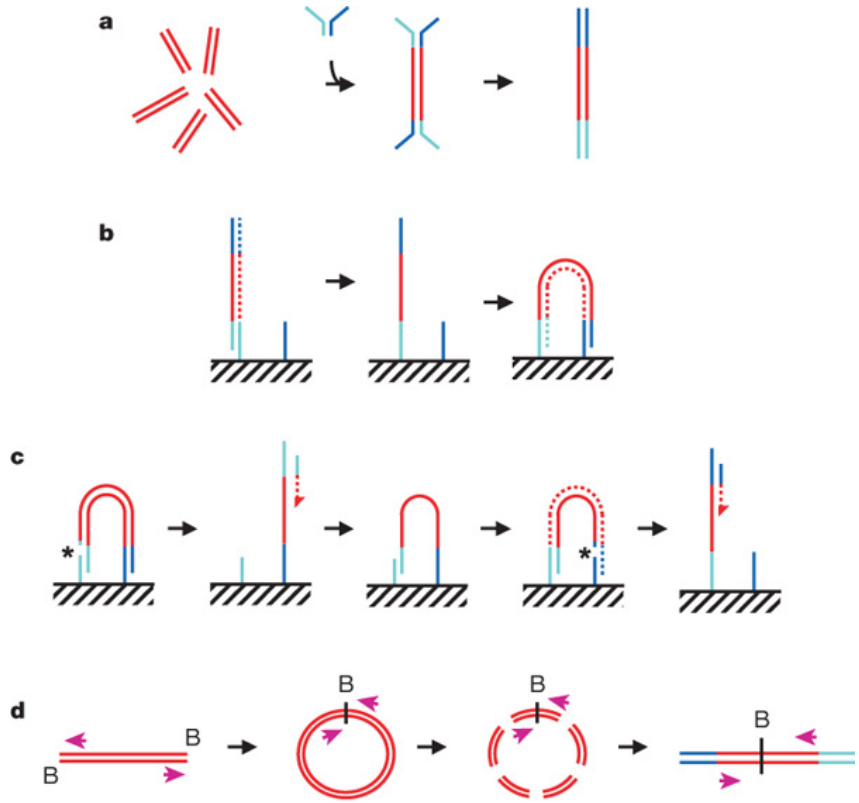
`strsplit`, `grep`, `sub` Operate on character vectors.

`biocLite` Install a package from an on-line repository.

`traceback`, `debug`, `browser` Help debug errors.

Outline

1. Wet lab: 10's of millions of 'random' DNA fragments, 300nt
2. Sequence: single or paired end, FASTQ files
3. Align: BAM files
4. Summarize: counts (RNASeq), coverage peaks (ChIPSeq; BED files), variants (DNASeq; VCF files), ...



Illumina bridge PCR; Bentley et al., 2008, Nature 456: 53-59.

Reads

Data	Short reads and their qualities
Tasks	Input, quality assessment, summary, trimming, ...
Packages	*ShortRead*, *Biostrings*
Functions	▶ `readFastq`, `FastqSampler`, `FasqtStreamer`. ▶ `qa`, `report`. ▶ `alphabetFrequency`, `alphabetByCycle`, `consensusMatrix`. ▶ `trimTails`, `trimLRPatterns`, `matchPDict`, ...

Alignments

Data	BAM files of aligned reads
Tasks	Input, BAM file manipulation, pileups
Packages	*Rsamtools* (also: *GenomicRanges*)
Functions	▶ `BamFile`, `BamFileList` ▶ `scanBam`, `ScanBamParam` (select a subset of the BAM file) ▶ `asBam`, `sortBam`, `indexBam`, `mergeBam`, `filterBam` ▶ `BamSampler`, `applyPileups`

Variants

Data	VCF (Variant Call Format) file
Tasks	Calling, input, summary, coding consequences
Packages	*VariantTools* (linux only), *VariantAnnotation*, *ensemblVEP*
Functions	▶ `tallyVariants` ▶ `readVcf`, `locateVariants`, `predictCoding` ▶ Also: SIFT, PolyPhen data bases

Ranges

Data Genomic coordinates to represent data (e.g., aligned reads) or annotation (e.g., gene models).

Tasks Input, counting, coverage, manipulation, ...

Packages *GenomicRanges*, *IRanges*

Functions

- `readGAlignments`, `readGAlignmentsList`
- Many intra-, inter-, and between-range manipulating, e.g., `narrow`, `flank`, `shift`, `intersect`, `findOverlaps`, `countOverlaps`

Annotations

Data	Gene symbols or other identifiers
Tasks	Discover annotations associated with genes or symbols
Packages	*AnnotationDbi* (*org.**, *GO.db*, ...), *biomaRt*
Functions	▶ Discovery: `columns`, `keytype`, `keys` ▶ `select`, `merge` ▶ *biomaRt*: `listMarts`, `listDatasets`, `listAttributes`, `listFilters`, `getBM`

Features

Data	Genomic coordinates
Tasks	Group exons by transcript or gene; discover transcript / gene identifier mappings
Packages	*GenomicFeatures* and *TxDb.** packages (also: *rtracklayer*)
Functions	▶ `exonsBy`, `cdsBy`, `transcriptsBy` ▶ `select` (see Annotations, below) ▶ `makeTranscriptDb*`

Genome annotations

Data	FASTA, GTF, VCF, . . . from internet resources
Tasks	Define regions of interests; incorporate known features (e.g., ENCODE marks, dbSNP variants) in work flows
Packages	*AnnotationHub*
Functions	▶ `AnnotationHub`, `filters` ▶ `metadata`, `hub$<tab>`

Sequences

Data Whole-genome sequences

Tasks View sequences, match position weight matricies, match patterns

Packages *Biostrings*, *BSgenome*

Functions

- `available.genomes`
- `Hsapiens[["chr3"]]`, `getSeq`, `mask`
- `matchPWM`, `vcountPattern`, `...`
- `forgeBSgenomeDataPkg`

Import / export

Data	Common text-based formats, `gff`, `wig`, `bed`; UCSC tracks
Tasks	Import and export
Packages	*rtracklayer*
Functions	▶ `import`, `export` ▶ `browserSession`, `genome`

And...

Data representation: *IRanges*, *GenomicRanges*, *GenomicFeatures*, *Biostrings*, *BSgenome*, *girafe*. Input / output: *ShortRead* (fastq), *Rsamtools* (bam), *rtracklayer* (gff, wig, bed), *VariantAnnotation* (vcf), *R453Plus1Toolbox* (454). Annotation: *GenomicFeatures*, *ChIPpeakAnno*, *VariantAnnotation*. Alignment: *Rsubread*, *Biostrings*. Visualization: *ggbio*, *Gviz*. Quality assessment: *qrqc*, *seqbias*, *ReQON*, *htSeqTools*, *TEQC*, *Rolexa*, *ShortRead*. RNA-seq: *BitSeq*, *cqn*, *cummeRbund*, *DESeq*, *DEXSeq*, *EDASeq*, *edgeR*, *gage*, *goseq*, *iASeq*, *tweeDEseq*. ChIP-seq, etc.: *BayesPeak*, *baySeq*, *ChIPpeakAnno*, *chipseq*, *ChIPseqR*, *ChIPsim*, *CSAR*, *DiffBind*, *MEDIPS*, *mosaics*, *NarrowPeaks*, *nucleR*, *PICS*, *PING*, *REDseq*, *Repitools*, *TSSi*. Motifs: *BCRANK*, *cosmo*, *cosmoGUI*, *MotIV*, *seqLogo*, *rGADEM*. 3C, etc.: *HiTC*, *r3Cseq*. Copy number: *cn.mops*, *CNAnorm*, *exomeCopy*, *seqmentSeq*. Microbiome: *phyloseq*, *DirichletMultinomial*, *clstutils*, *manta*, *mcaGUI*. Work flows: *ArrayExpressHTS*, *Genominator*, *easyRNASeq*, *oneChannelGUI*, *rnaSeqMap*. Database: *SRAdb*. ...

Outline

Resources

- Packages and their vignettes:
  http://bioconductor.org/packages/release
- Course and conference material:
  http://bioconductor.org/help/course-materials
- Introduction to *R* – `RShowDoc('R-intro')`
- Mailing list
  http://bioconductor.org/help/mailing-list for
  support

Acknowledgements

- *Bioconductor* team: Marc Carlson, Valerie Obenchain, Hervé Pagès, Paul Shannon, Dan Tenenbaum
- Technial advisory council: Vincent Carey, Wolfgang Huber, Robert Gentleman, Rafael Irizzary, Sean Davis, Kasper Hansen
- Scientific advisory board: Simon Tavaré, Vivian Bonazzi, Vincent Carey, Wolfgang Huber, Robert Gentleman, Rafael Irizzary, Paul Flicek, Simon Urbanek.
- NIH / NHGRI U41HG0004059
- . . . and the *Bioconductor* community!