Trends in Genomic Data Analysis with R / Bioconductor

Levi Waldron
CUNY School of Public Health, Hunter College

Martin T. Morgan
Fred Hutchinson Cancer Research Center

Michael Love
Dana-Farber Cancer Center

Vincent J. Carey
Harvard Medical School

16 July, 2014

Introductions

- Levi Waldron
  - Specializations: data curation and meta-analysis, gene expression, predictive modeling
- Martin T. Morgan: *Genomic data and annotation through AnnotationHub*
  - *Bioconductor* project leader
  - Specializations: sequence data analysis, genomic annotation
- Vincent J. Carey *Scalable integrative bioinformatics with Bioconductor*
  - *Bioconductor* founding member
  - Specializations: eQTL, integrative genomic data analysis, performant computing
- Michael Love: *RNA-Seq workflows in Bioconductor*
  - Specializations: RNA-Seq

Introduction: *Bioconductor*

Analysis and comprehension of high-throughput genomic data

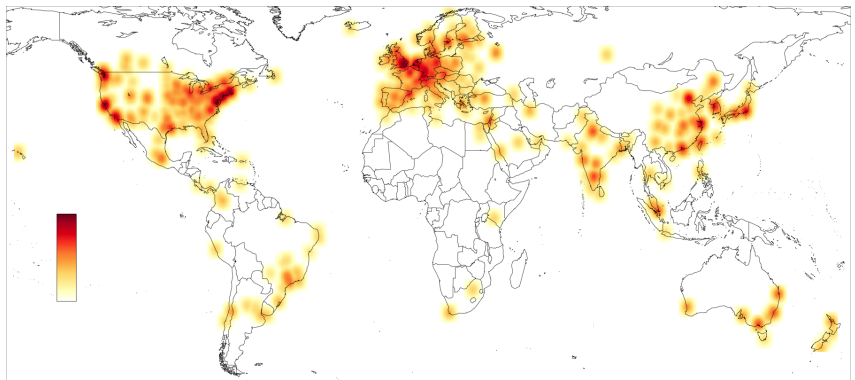
- `http://bioconductor.org`
- $>$ 11 years old, 824 packages

Themes:

- Rigorous statistics
- Reproducible work flows
- Integrative analysis
- distributed development

Introduction: *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; ...



Bioconductor Conference, July 30 - Aug 1, Boston, USA

Bioc2014: July 30 - Aug 1, 2014 (Boston)

- July 30: Developers Day (current and prospective)
- Morning scientific talks
- afternoon practicals (2h hands-on sessions)
  - Introduction, Variant Calling, Intro Sequence Analysis, RNA-seq differential expression, ChIP-seq, 450K methylation data analysis, genomic annotation resources, meta-analysis, parallel computing...

https://register.bioconductor.org/BioC2014

Introduction: Application areas of *Bioconductor*

- Microarray analysis: expression, copy number, SNPs, methylation, ...
- Sequencing: RNA-seq, ChIP-seq, called variants, ...
  - Especially *after* assembly / alignment
- Annotation: genes, pathways, gene models (exons, transcripts, etc.), ...
- Epigenetics
- Gene set enrichment analysis
- Network analysis
- Flow cytometry
- Proteomics and metabolomics
- Cheminformatics
- Images and high-content screens

Levels of documentation

Bioconductor documentation exists at several levels:

- http://www.bioconductor.org/help
  - Workflows, mailing lists, newsletters, courses, blogs, books
- **Workflows:** Common tasks spanning multiple packages, http://www.bioconductor.org/help/workflows/
  - *e.g.*: Sequence Analysis, RNAseq differential expression, oligonucleotide arrays, variants, accessing annotation data, annotating ranges...
- **Package Vignettes:** Working "literate code" demonstrating use of a package
  - Some vignettes of mature packages are extensive introductions, *e.g.* limma
- **Function man pages** and Reference Manuals

Additional Sources of Documentation

- **Courses and Workshops:**
  - http://www.bioconductor.org/help/course-materials/
  - Notes from dozens of courses and workshops, including today's.
- **BiocViews** hierarchical controlled vocabulary
  - Software (824)
  - AnnotationData (867)
  - ExperimentData (202)
- **Classic textbooks:**
  - Bioinformatics and Computational Biology Solutions Using R and Bioconductor
  - Bioconductor Case Studies
  - R Programming for Bioinformatics
- **Bioconductor mailing list**

Key Data Structures

Container (package)	Data type
`ExpressionSet` (*Biobase*)	Matrix-like dataset plus experiment/sample/feature metadata
`SummarizedExperiment` (*GenomicAlignments*)	Analogous to `ExpressionSet`, but features defined in genomic coordinates.
`GRanges` (*GenomicRanges*)	Genomic coordinates and associated qualitative and quantitative information, e. g., gene symbol, coverage, *p*-value.

Table 1 : Key common data structures in *Bioconductor*. `SummarizedExperiment` and `GRanges` are standard for genome-linked data; `ExpressionSet` is standard for most other experimental data.

Microarray Analysis

- 300 packages with microarray biocViews term
  - Classic packages: *affy* (RMA preprocessing), *limma* (linear modeling)
  - Newer packages: *oligo* (tools for modern microarrays), *pdInfoBuilder* (for building annotation packages)
- All kinds of arrays supported
  - See Arrays workflow
  - Excellent Vignettes, *e.g.* of *limma* and *affy*

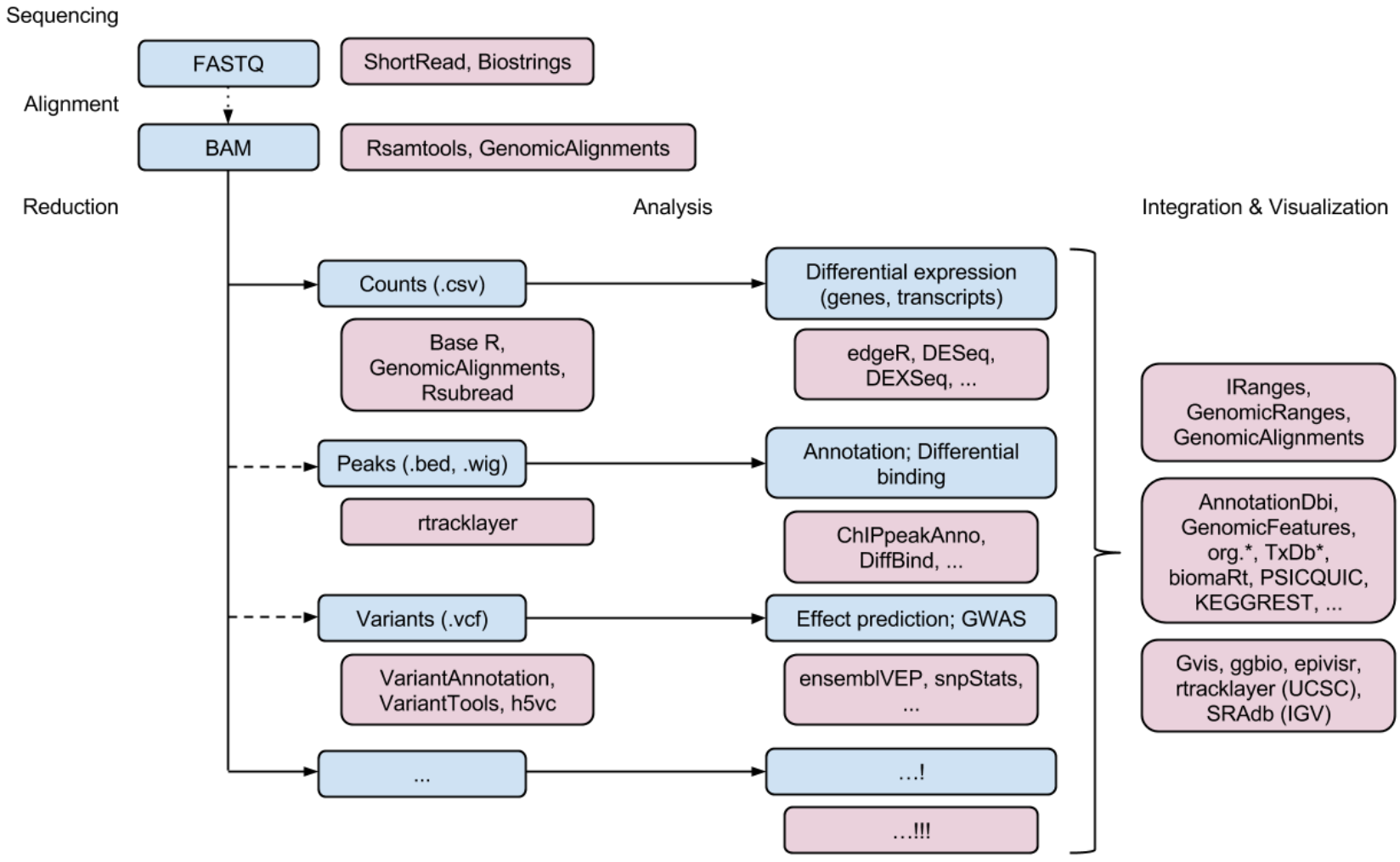
RNA-seq differential expression analysis

- 55 packages with RNASeq biocViews term
  - *edgeR*, *DESeq2* for differential abundance analysis
  - *Rsubread* for read alignment, quantification and mutation discovery
  - *QuasR* provides an integrated work flow using *Rbowtie* for alignment and *GenomicRanges* for read counts.
  - *cummeRbund* for post-processing of *cufflinks* isoform assemblies

Epigenetics

- 53 packages with Epigenetics-related biocViews term
  - 450K methylation arrays: **minfi, methylumi, lumi**, *methyAnalysis*, *wateRmelon*, *ChAMP*
  - Whole-genome bisulfite sequencing: *bsseq*, *MethylSeekR*, *BiSeq*, *QuasR*
  - affinity or restriction enzyme based assays such as ME-dip or MBD-seq: *Repitools*, *MEDIPS*
  - ChIP-seq: *DiffBind*, *DBChIP*, *ChIPpeakAnno*

Bioconductor ecosystem of sequencing tools



Credit: Martin Morgan

String-related data structures and tools

Use case	Packages
Basic operations on `DNAString` and `DNAStringSet` objects	*Biostrings*
Extract sequences of an arbitrary set of regions	*BSgenome*::`getSeq`
Extract transcript, CDS, or promoter sequences from a reference genome and gene model	*GenomicFeatures*
Import sequences from BAM file	*Rsamtools*, *GenomicAlignments*
Pileup functions	*GenomicAlignments* (`pileLettersAt` and `stackStringsFromBam`), *Rsamtools*`::applyPileups`, *VariantTools*`::tallyVariants`
Representation of ref/alt alleles	*VariantAnnotation*::`VCF` and `VRanges` classes)
Predict amino acid coding	*Biostrings*::`translate`, *VariantAnnotation*`::predictCoding`
Short read quality assessment	*ShortRead*::`qa`
Assess technical bias in NGS data	*seqbias*
Identify low-complexity sequences	*ShortRead*::`dustyScore`
Measure CpG enrichment	*MEDIPS*::`MEDIPS.CpGenrich`

String-related data structures and tools (cont'd)

Use case	Packages
Motif matching	*Biostrings*::`matchPWM` and *MotIV*::`motifMatch`
Motif discovery	*motifRG*, *rGADEM*
Find palindromic regions	*Biostrings*::`findPalindromes`
Find intramolecular triplexes (H-DNA) in DNA sequences	*triplex*
Map probe sequences to a reference genome	*altcdfenvs*::`matchAffyProbes`, *waveTiling*::`filterOverlap`
Find probe positions in a set of gene sequences	*GeneRegionScan*::`findProbePositions`
Specialized matching/alignment tools	*DECIPHER* (`AlignSeqs`, `AlignProfiles`, and `FindChimeras`)
Design of hybridization probes	*DECIPHER*
Import and analysis of Roche's 454 sequencing data	*R453Plus1Toolbox* and *rSFFreader*

Operation type	Functions
Arithmetic	`shift`, `resize`, `restrict`, `flank`
Set	`intersect`, `union`, `setdiff`, `gaps`
Summary	`coverage`, `reduce`, `disjoin`
Comparison	`findOverlaps`, `nearest`, `order`

Table 2 : Some of the important functions in the `ranges` algebra. They are flexible and fast.

Visualization

Domain	Packages
(Epi-)Genomic Data	*Gviz* and *epivisr* (genome browsers), *rtracklayer* (UCSC)
Networks	*Rgraphviz*, *RCytoscape*
Chemical Structure	*ChemmineR*
Flow Cytometry	*flowViz*, *flowPlots*, *spade*
Big Data	*supraHex*

Table 3 : 134 Bioconductor packages are currently tagged with the 'Visualization' keyword.

Annotation resources

Pre-built packages	
*org.**	Identifier mapping (*AnnotationDbi*)
*TxDb.**	Gene models (*GenomicFeatures*)
*BSgenome.**	Whole-genome sequences (*BSgenome*)
Web access (examples)	
biomaRt	Ensembl (and other) biomart
rtracklayer	UCSC genome browser tracks
ensemblVEP	Ensembl Variant Effect Predictor
PSICQUIC	Molecular interactions data bases
AnnotationHub (Bioc-hosted transparent-access databases)	
	UCSC, ENCODE, Ensembl, dbSNP

Table 4 : Annotation resources in *Bioconductor*.

Experimental data packages

- 202 packages with `ExperimentData biocViews`
- `Relatively static data for:`
  - `Package testing (e.g. ALL)`
  - `Reproducible analysis for published papers (e.g. Hiiragi2013)`
  - `Meta-analysis of curated cancer datasets (e.g. curatedOvarianData, curatedCRCData, curatedBladderData)`

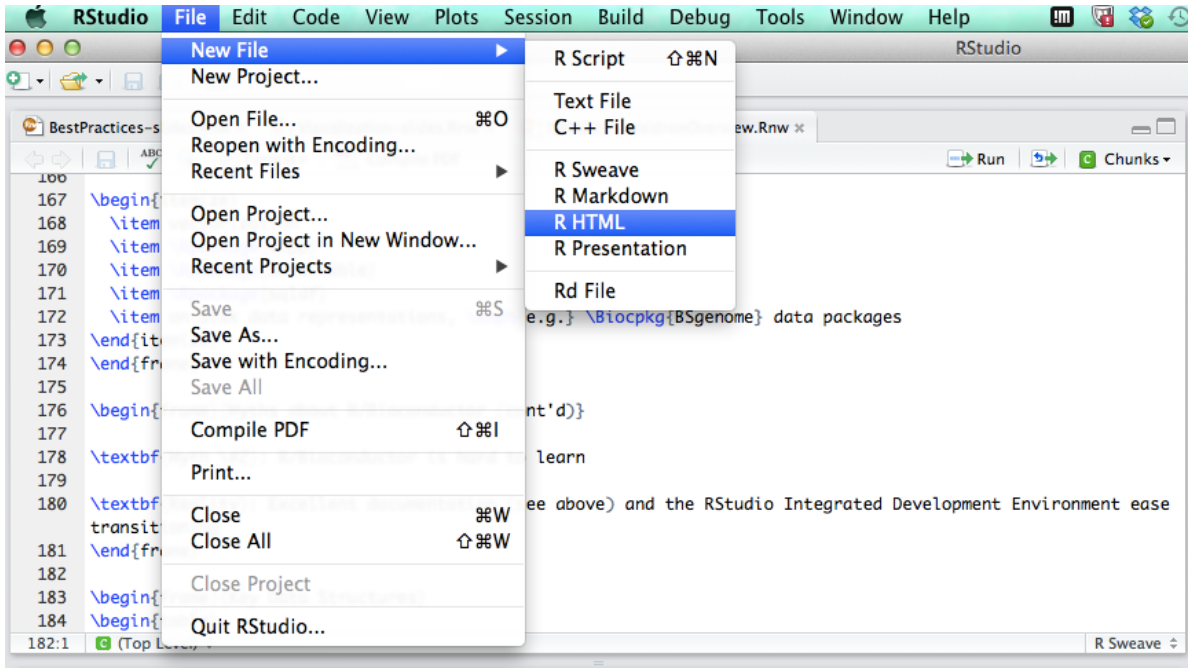
Acquiring experimental data from online databases

- *GEOquery*: Import data from NCBI Gene Expression Omnibus (GEO)
- *GeoMetaDB*: SQLite database of all GEO metadata
- *SRAdb*: SQLite database of NCBI Short Read Archive + download / send tracks to IGV
- *ArrayExpress*: Import ArrayExpress data
- *CGDS-R*: cBioPortal TCGA import
- Synapse R Client for TCGA

Myths about R/Bioconductor (cont'd)

Myth #1: R/Bioconductor is hard to learn

Reality: Multi-level documentation (see above), RStudio Integrated Development Environment, online courses ease transitioning



Summary - Myths about R/Bioconductor

Myth #2: R/Bioconductor is slow / uses too much memory

Reality: R/Bioconductor *can* slow or memory intensive, depending on how it's used:

- vectorization
- *Rcpp*, traditional **C** and **Fortran** function interfaces
- library(*data.table*)
- library(*sqldf*)
- on-disk data representations, *e.g.* *BSgenome* data packages
- *knitr* provides caching with dependency tracking
- *parallel*, *BiocParallel* for parallelization

Acknowledgements

- slide contributions: Wolfgang Huber, Vincent Carey, Robert Gentleman,Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper Hansen, Rafael A. Irizarry, Michael Lawrence, Michael I. Love, James MacDonald, Valerie Obenchain, Andrzej K. Oleś, Hervé Pagès, Paul Shannon, Gordon Smyth, Dan Tenenbaum, Martin Morgan
- The *Bioconductor* community