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](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...

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
Bioconductor documentation exists at several levels:

- [http://www.bioconductor.org/help](http://www.bioconductor.org/help)
  - Workflows, mailing lists, newsletters, courses, blogs, books

  - *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:**
  - 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

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

**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

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

<table>
<thead>
<tr>
<th>Use case</th>
<th>Packages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motif matching</td>
<td><em>Biostrings</em>: <code>matchPWM</code> and <em>MotIV</em>: <code>motifMatch</code></td>
</tr>
<tr>
<td>Motif discovery</td>
<td><code>motifRG</code>, <em>rGADEM</em></td>
</tr>
<tr>
<td>Find palindromic regions</td>
<td><em>Biostrings</em>: <code>findPalindromes</code></td>
</tr>
<tr>
<td>Find intramolecular triplexes (H-DNA) in DNA sequences</td>
<td><em>triplex</em></td>
</tr>
<tr>
<td>Map probe sequences to a reference genome</td>
<td><em>altcdfenvs</em>: <code>matchAffyProbes</code>, <em>waveTiling</em>: <code>filterOverlap</code></td>
</tr>
<tr>
<td>Find probe positions in a set of gene sequences</td>
<td><em>GeneRegionScan</em>: <code>findProbePositions</code></td>
</tr>
<tr>
<td>Specialized matching/alignment tools</td>
<td><em>DECIPHER</em> (AlignSeqs, AlignProfiles, and FindChimeras)</td>
</tr>
<tr>
<td>Design of hybridization probes</td>
<td><em>DECIPHER</em></td>
</tr>
<tr>
<td>Import and analysis of Roche’s 454 sequencing data</td>
<td><em>R453Plus1Toolbox</em> and <em>rSFFreader</em></td>
</tr>
<tr>
<td>Operation type</td>
<td>Functions</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>Arithmetic</td>
<td>shift, resize, restrict, flank</td>
</tr>
<tr>
<td>Set</td>
<td>intersect, union, setdiff, gaps</td>
</tr>
<tr>
<td>Summary</td>
<td>coverage, reduce, disjoin</td>
</tr>
<tr>
<td>Comparison</td>
<td>find0overlaps, nearest, order</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Domain</th>
<th>Packages</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Epi-)Genomic Data</td>
<td>\textit{Gviz} and \textit{epivisr} (genome browsers), \textit{rtracklayer} (UCSC)</td>
</tr>
<tr>
<td>Networks</td>
<td>\textit{Rgraphviz}, \textit{RCytoscape}</td>
</tr>
<tr>
<td>Chemical Structure</td>
<td>\textit{ChemmineR}</td>
</tr>
<tr>
<td>Flow Cytometry</td>
<td>\textit{flowViz}, \textit{flowPlots}, \textit{spade}</td>
</tr>
<tr>
<td>Big Data</td>
<td>\textit{supraHex}</td>
</tr>
</tbody>
</table>

**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
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


- The *Bioconductor* community