R / *Bioconductor* for Sequence Analysis

Martin Morgan[1]

June 20-23, 2011

[1] mtmorgan@fhcrc.org

Bioconductor

Goal Help biologists understand their data

Focus
- Expression and other microarray
- Sequence analysis
- Imaging, flow cytometry, ...

Themes
- Based on the *R* programming language – statistics, visualization, interoperability
- Reproducible – scripts, *vignettes*, packages
- Open source / open development
- Contributions from 'core' members and (primarily academic) user community

Status $>$ 460 packages; very active web site and mailing list; annual conferences; courses; ...

Using *R* / *Bioconductor*

- **Programming language**

```
> library(GEOquery)
> eset = getGEO('...')
```

- Scripts, vignettes, packages
- Appeal

Flexibility

Leveraging resources, e.g., SQL, XML, third party libraries (e.g., *samtools*)

R statistical methods and visualization

Using *R* / *Bioconductor*

- Programming language

```
> library(GEOquery)
> eset = getGEO('...')
```

- **Scripts, vignettes, packages**
- Appeal

1. Reproducibility
2. Communication
3. Enabling

Using *R* / *Bioconductor*

- Programming language

```
> library(GEOquery)
> eset = getGEO('...')
```

- Scripts, vignettes, packages
- **Appeal**

Statisticians

Bioinformaticists

. . . but not everyone!

A Package Tour

Bioconductor

- **Expression and other microarrays**
- Sequence analysis
- Annotation and archive resources
- Additional

All of CRAN

Pre-processing

Quality assessment

Differential expression (e.g., *limma*)

Gene set enrichment

Many features for free, e.g., machine learning, visualization

A Package Tour

Bioconductor

- **Expression and other microarrays**
- Sequence analysis
- Annotation and archive resources
- Additional

All of CRAN

Array CGH (e.g., *DNAcopy*)

Methylation, epigenetics, miRNA

Genotyping (e.g., *snpStats*)

A Package Tour

Bioconductor

- Expression and other microarrays
- **Sequence analysis**
- Annotation and archive resources
- Additional

All of CRAN

I/O, QA, manipulation

RNAseq differential representation (e.g., *DESeq*)

Gene set analysis (e.g., *goseq*)

ChIPseq

Metabiome

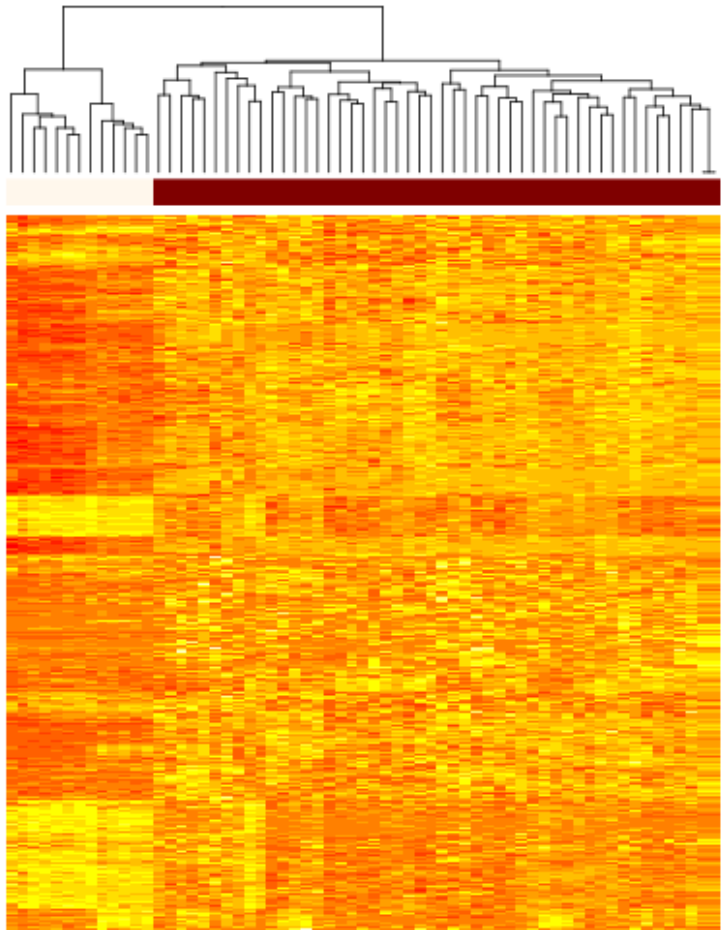
A Package Tour

Bioconductor

- Expression and other microarrays
- **Sequence analysis**
- Annotation and archive resources
- Additional

All of CRAN

50 ovarian cancer, 13 benign / normal RNAseq samples



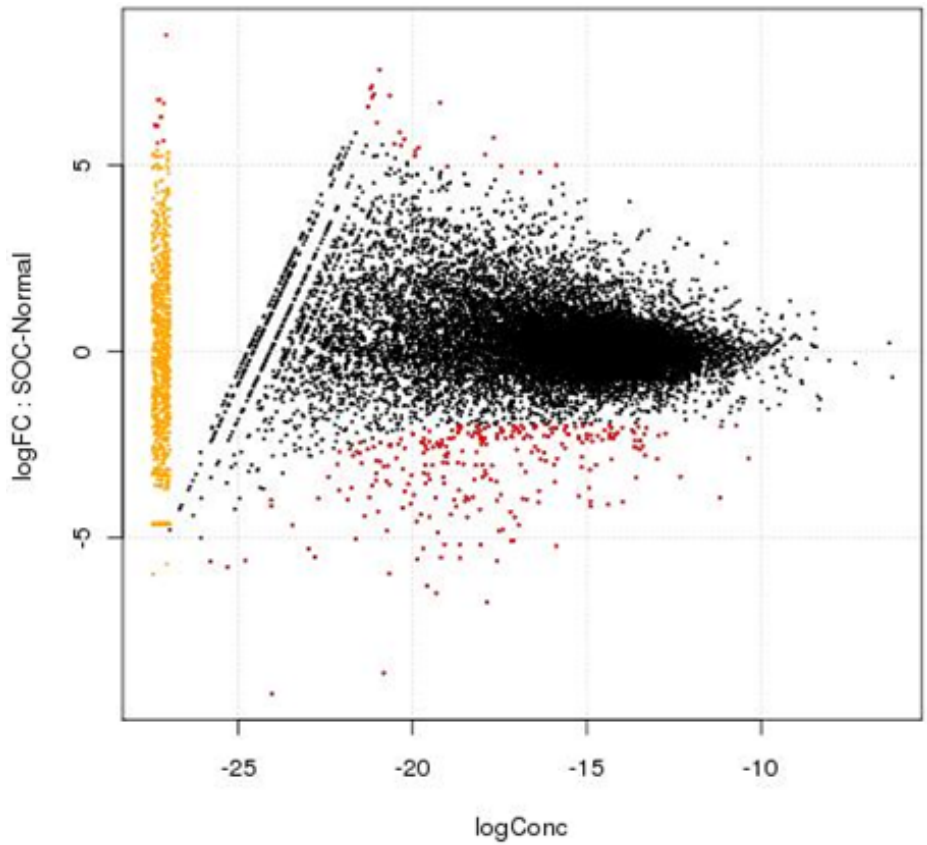
A Package Tour

Bioconductor

- Expression and other microarrays
- **Sequence analysis**
- Annotation and archive resources
- Additional

All of CRAN

Differential representation in SOC vs. Control



A Package Tour

Bioconductor

- Expression and other microarrays
- **Sequence analysis**
- Annotation and archive resources
- Additional

All of CRAN

KEGG terms under-represented in SOC

	Description	P Value
1	Spliceosome	0.0017
3	Ribosome	0.0073
5	Cell cycle	0.0123
...		

Investigate intron abundances

A Package Tour

Bioconductor

- Expression and other microarrays
- Sequence analysis
- **Annotation and archive resources**
- Additional

All of CRAN

Curated, versioned (semi-annual)

- Chip
- Organism
- Pathway
- Homology
- miRNA

biomaRt, UCSC

GEO, *ArrayExpress*, SRA

A Package Tour

Bioconductor

- Expression and other microarrays
- Sequence analysis
- **Annotation and archive resources**
- Additional

All of CRAN

Examples:

Identify human genes in 'spliceosome', 'ribosome', and 'cell cycle' KEGG pathways.

Discover and retrieve GEO expression arrays related to ovarian carcinomas.

Remotely query 1000 genomes BAM files for regions of interest, e.g., 'spliceosome' genes.

Input TCGA ovarian cancer copy number and clinical data.

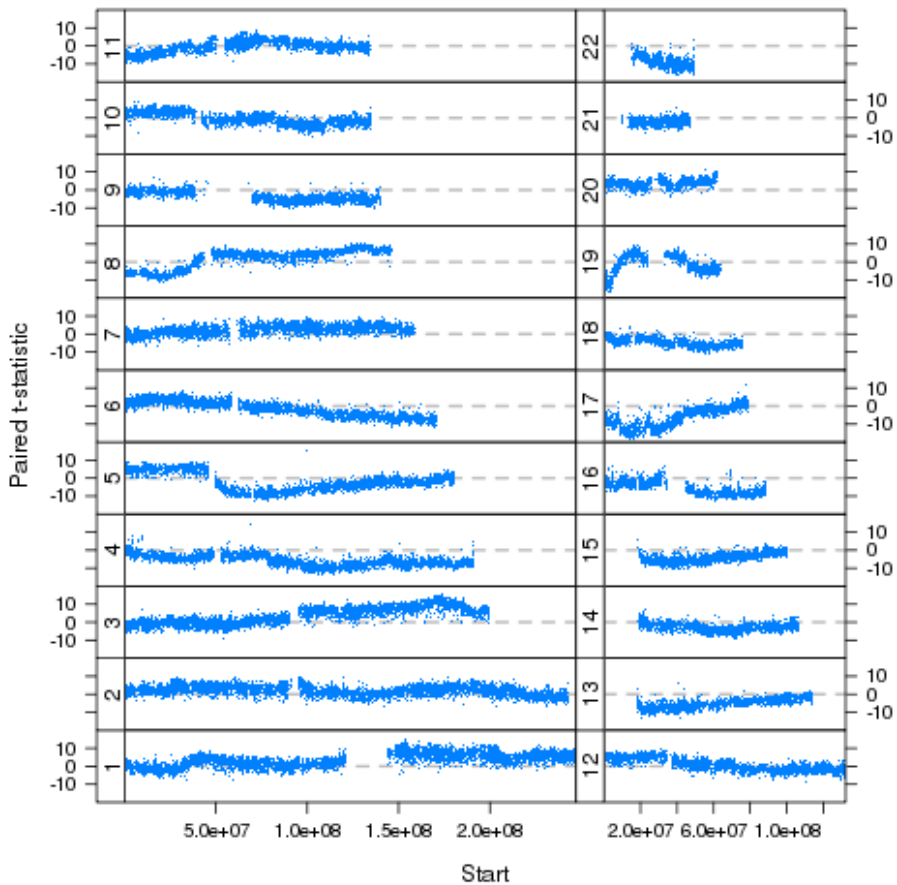
A Package Tour

Bioconductor

- Expression and other microarrays
- Sequence analysis
- **Annotation and archive resources**
- Additional

All of CRAN

86 Paired HMS HG-CGH-244A TCGA samples



A Package Tour

Bioconductor

- Expression and other microarrays
- Sequence analysis
- Annotation and archive resources
- **Additional**

All of CRAN

Pathways and networks

Flow cytometry

High-throughput qPCR

Image processing (*e.g.*, *EBImage*)

A Package Tour

Bioconductor

- Expression and other microarrays
- Sequence analysis
- Annotation and archive resources
- Additional

All of CRAN

3000+ packages

Novel approaches, e.g., *cghFLasso*

Advanced statistical analyses, e.g., Bayesian network models

Common work flows

Input / output

- Fasta, fastq – *ShortRead*
- SAM / BAM, tabix, indexed fasta – *Rsamtools*
- Genome tracks & related formats – *rtracklayer*

Pre-processing / manipulation / count & measure

- String manipulation, pattern matching *Biostrings*
- Quality assessment *ShortRead*
- finding / counting overlaps *GenomicRanges*

Analysis domains

- RNAseq, e.g., *DESeq*, *edgeR*, *goseq*
- ChIPseq, e.g., *ChIPpeakAnno*

Annotation / variants

- *AnnotationDbi* / *org.**, *GenomicFeatures*, *BSgenome*, *biomaRt*

Useful data structures

Biostrings, *BSgenome*

- *XString*, *XStringSet*

GenomicRanges

- *GappedAlignments* – CIGAR
- *GRanges* / *GRangesList* – sequence, strand

IRanges

- *IRanges* / *IRangesList* / *RangedData*– ranges
- *Rle* – run length encoding
- *Views*

Effective compulational software

Effective computational biology software

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

Bioconductor

Who

- FHCRC: Hervé Pagès, Marc Carlson, Nishant Gopalakrishnan, Valerie Obenchain, Dan Tenenbaum, Chao-Jen Wong
- Robert Gentleman (Genentech), Vince Carey (Harvard / Brigham & Women's), Rafael Irizzary (Johns Hopkins), Wolfgang Huber (EBI, Hiedelberg)
- A large number of contributors, world-wide

Resources

- http://bioconductor.org: installation, packages, work flows, courses, events
- Mailing list: friendly prompt help
- Conference: Morning talks, afternoon workshops, evening social. 28-29 July, Seattle, WA. Developer Day July 27