R / *Bioconductor* for Sequence Analysis

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

June 27-July 1, 2011

[1] mtmorgan@fhcrc.org

Bioconductor

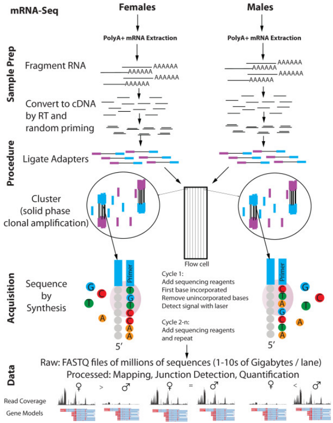
Goal	Help biologists understand their data
Focus	▶ Sequence analysis ▶ Expression and other microarray ▶ Imaging, flow cytometry, . . .
Themes	▶ Based on the *R* programming language – statistics, visualization, interoperability ▶ Reproducible – scripts, *vignettes*, packages ▶ Open source / open development ▶ Contributions from ‘core’ members and (primarily academic) user community
Status	> 460 packages; very active web site and mailing list; annual conferences; courses; . . .

Sequence analysis

Overall work flow

1. Experimental design
2. Sample preparation
3. Sequencing – fastq files
4. Alignment – bam files
5. *Quality assessment* (before & after alignment)
6. ‘Domain-specific’ analysis – RNAseq, ChIPseq, . . .

Italic: role for *Bioconductor*



Malone and Oliver (2011)

Sequence analysis

Overall work flow

1. Experimental design
2. Sample preparation
3. Sequencing – fastq files
4. Alignment – bam files
5. *Quality assessment* (before & after alignment)
6. 'Domain-specific' analysis – **RNAseq**, ChIPseq, . . .

Italic: role for *Bioconductor*

RNAseq: gene abundance

- Estimate or *count reads overlapping genes*
- *Machine learning*
- *Between-group comparison*
- *Gene set enrichment*
- *Annotation*

Sequence analysis

Overall work flow

1. Experimental design
2. Sample preparation
3. Sequencing – fastq files
4. Alignment – bam files
5. *Quality assessment* (before & after alignment)
6. 'Domain-specific' analysis – **RNAseq**, ChIPseq, . . .

Italic: role for *Bioconductor*

RNAseq: transcript abundance

- Alignment to known gene models, or to whole genome
- *Count reads overlapping transcripts or exons*
- *Machine learning*
- *Between-group comparison*
- *Gene set enrichment*
- *Annotation*

Example work flow in *passila* experiment data package vignette

Sequence analysis

Overall work flow

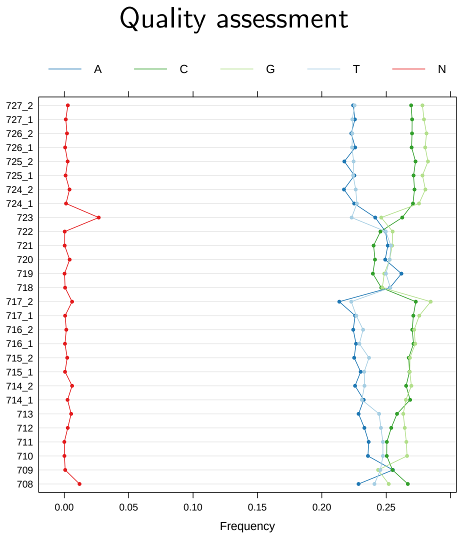
1. Experimental design
2. Sample preparation
3. Sequencing – fastq files
4. Alignment – bam files
5. *Quality assessment* (before & after alignment)
6. 'Domain-specific' analysis – RNAseq, **ChIPseq**, . . .

Italic: role for *Bioconductor*

ChIPseq

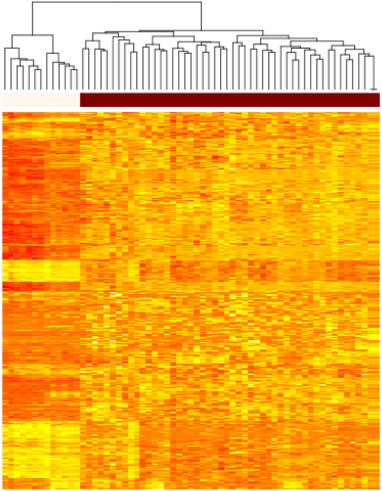
- Find peaks, e.g., MACS, *chipseq*, 59 others. . .
- *Annotation*
- *Designed experiments?*

A Package Tour



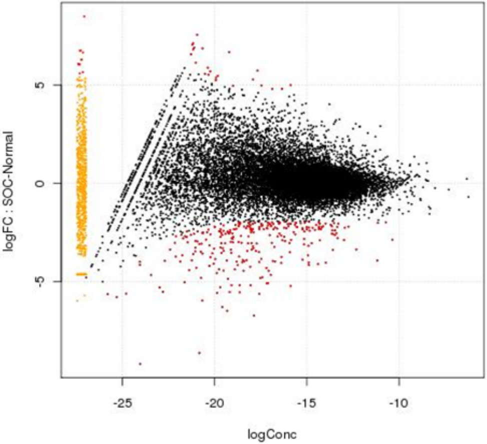
A Package Tour

50 ovarian cancer, 13 benign / normal RNAseq samples



A Package Tour

Differential representation in ovarian cancer vs. control



A Package Tour

KEGG terms under-represented in ovarian cancers

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

⇒ Investigate intron abundances

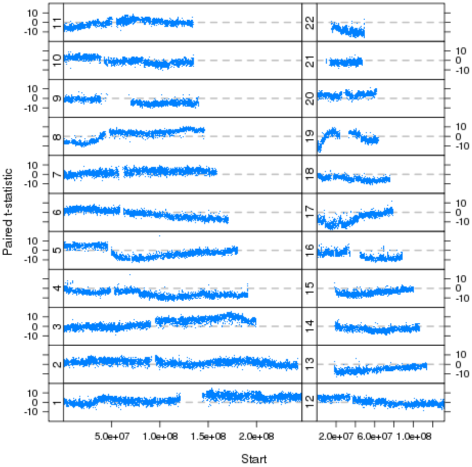
A Package Tour

Annotation and data integration

- Retrieve gene models (coordinates)
- Identify human genes in ‘spliceosome’, ‘ribosome’, and ‘cell cycle’ KEGG pathways.
- Discover and retrieve GEO expression arrays related to ovarian carcinomas.
- Query 1000 genomes BAM files for regions of interest, e.g., ‘spliceosome’ genes.

A Package Tour

Integrate 86 Paired HMS HG-CGH-244A TCGA samples



Common work flows

Input / output

- Fasta, fastq – *ShortRead*
- SAM / BAM – *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., *rGADEM*, *ChIPpeakAnno*

Annotation / variants

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

Useful data structures

DNAString, *DNAStringSet*

- Sequences and character-encoded quality scores
- *Biostrings*, *BSgenome*, *ShortRead*

GappedAlignments

- Sequence alignment coordinates
- CIGAR, e.g., a read aligning with 25 matches or mismatches, then an insertion relative to reference of 5 nucleotides, and then 7 more matches or mismatches is `25M 5I 7M`
- *GenomicRanges*, *Rsamtools*

GRanges / *GRangesList*

- Ranges of genomic coordinates
- E.g., simple genes (*GRanges*), exons within transcripts (*GRangesList*)
- *GenomicFeatures*, *GenomicRanges*, *IRanges*

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

Citations

J. H. Malone and B. Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.*, 9:34, 2011.