

R / Bioconductor Packages for Short Read Analysis

Martin Morgan (mtmorgan@fhcrc.org)

Fred Hutchinson Cancer Research Center

7-9 June, 2010

Announcement / Acknowledgments

- ▶ Annual conference in Seattle, 29-30 July ('Developer Day' 28 July) <https://secure.bioconductor.org/BioC2010>
- ▶ Two positions available – software and web development <http://www.fhcrc.org/about/jobs/index.html> and search for positions 23129, 23133.

Bioconductor team

- ▶ Patrick Aboyoun, Marc Carlson, Nishant Gopalakrishnan, Hervé Pagès, Chao-Jen Wong
- ▶ Wolfgang Huber, Vince Carey, Rafael Irizarry, Robert Gentleman.

Outline

Work flow

Experiment

Technology

Pre-processing

Analysis

Annotation and Integration

Examples (Psuedo-Code)

Quality Assessment

454 Microbiome Pre-Processing

Digital Gene Expression

Differential Expression

Resources

Experiments

Sequence-based analysis

- ▶ ChIP
- ▶ Differential expression
- ▶ RNA-seq (alternate splicing)
- ▶ Metagenomic
- ▶ ...

Important issues

- ▶ Experimental design
- ▶ Replication
- ▶ Sample preparation artifacts

Technology

Platforms

- ▶ Illumina / Genome Analyzer
- ▶ Roche / 454
- ▶ AB / SOLiD
- ▶ Complete Genomics
- ▶ Third-generation: PacBio, Ion Torrent, Oxford Nanopore

Important issues

- ▶ Experimental design (blocking)
- ▶ Technology artifacts

Pre-processing

Vendor and third-party

- ▶ Image processing, base calling
- ▶ Machine quality assessment
- ▶ Alignment

Bioconductor

- ▶ Quality assessment and representation: *ShortRead*, *GenomicRanges*
- ▶ Read remediation, trimming, primer removal, specialized manipulation: *IRanges*, *ShortRead*, *Biostrings*
- ▶ Specialized alignment tasks: *Biostrings*, *BSgenome*

Analysis

Domain-specific, e.g.,

- ▶ ChIP-seq: *chipseq*, *ChIPseqR*, *CSAR*, *BayesPeak*
- ▶ Differential expression: *DESeq*, *edgeR*, *baySeq*
- ▶ RNA-seq: *Genominator*

Examples

- ▶ *EatonEtAlChIPseq*, *leeBamViews*

Annotation and Integration

Annotation

- ▶ Genome coordinate / gene (and other) relationships, *GenomicFeatures*, *ChIPpeakAnno*

Integration

- ▶ Digital and microarray differential expression
- ▶ RNAseq and gene ontology / pathway, *goseq*
- ▶ HapMap, 1000 genomes, UCSC, Sequence Read Archive, GEO, ArrayExpress, *rtracklayer*, *biomaRt*, *Rsamtools*, *GEOquery*, *SRAdb*

Outline

Work flow

Experiment

Technology

Pre-processing

Analysis

Annotation and Integration

Examples (Psuedo-Code)

Quality Assessment

454 Microbiome Pre-Processing

Digital Gene Expression

Differential Expression

Resources

Quality Assessment

```
> library(ShortRead)
> library(multicore)      # Use all cpu cores for qa()
> dir <-                  # Input
+   "/mnt/fred/solexa/xxx/100524_HWI-EAS88_0005"
> sp <- SolexaPath(dir)  # Many other formats
> qa <- qa(sp)           # Collate statistics -- slow
> rpt <- report(qa)      # Create report
> browseURL(rpt)         # View in browser
```

454 Microbiome Pre-Processing

```
> library(ShortRead)
> dir <- "/not/public"
> bar <- read454(dir)           # Input
> code <- narrow(sread(bar), 1, 8) # Extract bar code
> aBar <- bar[code == "AAGCGCTT"] # Subset one bar code
> noBar <-                      # Remove bar code
+   narrow(aBar, 11, width(aBar))
> pcrPrimer <- "GGACTACCVGGGTATCTAAT"
> trimmed <-                    # Remove primer
+   trimLRPatterns(pcrPrimer, noBar, Lfixed=FALSE)
> writeFastq(trimmed,           # Output
+   file.path(dir, "trimmed.fastq"))
```

Digital Gene Expression

```
> library(GenomicFeatures)
> bamFile <- "/path/to/file.bam"
> aligns <- readGappedAlignments(bamFile)
> ## ... txdb: transcripts from UCSC 'knownGenes'
> exonRanges <- exonsBy(txdb, "tx")
> ## ... housekeeping
> counts <- countOverlaps(exonRanges, aligns)
> ## ... normalization --> 'highScores' variable
> txs <- transcripts(txdb,
+                   vals=list(tx_id=names(highScores)),
+                   columns=c("tx_id", "gene_id"))
> systematicNames <- elementMetadata(txs)[["gene_id"]]
```

Differential Expression

```
> library(DESeq)
> tsvFile <- # Input, or previous work flow
+   system.file("extra", "TagSeqExample.tab",
+             package="DESeq")
> counts <- read.delim(tsvFile, header=TRUE,
+                   stringsAsFactors=TRUE, row.names="gene")
> condition <- factor(c("T", "T", "T", "Tb", "N", "N"))
> cds <- newCountDataSet(counts, condition)
> cds <- # Effective library size
+   estimateSizeFactors(cds)
> cds <- # Variance, estimated from mean
+   estimateVarianceFunctions(cds)
> res <- # Negative binomial test
+   nbinomTest(cds, "T", "N")
```

Outline

Work flow

Experiment

Technology

Pre-processing

Analysis

Annotation and Integration

Examples (Psuedo-Code)

Quality Assessment

454 Microbiome Pre-Processing

Digital Gene Expression

Differential Expression

Resources

Resources

Bioconductor Web site

- ▶ `http://bioconductor.org`
- ▶ 'Installation', 'Software', and 'Mailing lists' links.

Help in *R*

- ▶ `help.start()` to view a help browser.
- ▶ `help(package = "Biostrings")`
- ▶ `?readAligned`
- ▶ `browseVignettes("GenomicRanges")`