R / *Bioconductor* Packages for Short Read Analysis

Martin Morgan (mtmorgan@fhcrc.org)

Fred Hutchinson Cancer Research Center

14-18 June, 2010

Announcements / Acknowledgments

Announcements

- Annual conference in Seattle, 28-30 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.

Resources

- Bioconductor web site `http://bioconductor.org`

Outline

- Work flow
  - Experiments and Technologies
  - Pre-processing
  - Analysis
  - Annotation and Integration
- Examples (Psuedo-Code)
  - Quality Assessment
  - 454 Microbiome Pre-Processing
  - Differential Expression
- ShortRead
  - Input and exploration
  - Manipulation
- Resources

Experiments and Technologies

Sequence-based experiments

- ChIP, Differential expression, RNA-seq, Metagenomic, . . .

Technology

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

Relevant issues in analysis

- Experimental design, replication, sample preparation artifacts

Pre-processing

Vendor and third-party

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

Bioconductor packages

- 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 (‘experiment data’ packages)

- *EatonEtAlChIPseq*, *leeBamViews*

Annotation and Integration

Annotation

- Gene-centric: *AnnotationDbi*, *org.*.db*, *KEGG.db*, *GO.db*, *Category*, *GOstats*
- Genome coordinate: *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

- Experiments and Technologies
- Pre-processing
- Analysis
- Annotation and Integration

Examples (Psuedo-Code)

- Quality Assessment
- 454 Microbiome Pre-Processing
- Differential Expression

ShortRead

- Input and exploration
- Manipulation

Resources

Quality Assessment

```
> library(ShortRead)
> 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"))
```

Differential Expression

```
> library(DESeq)
> tsvFile <-            # Input
+     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)
> cds1 <-               # Effective library size
+     estimateSizeFactors(cds0)
> cds2 <-               # Variance, estimated from mean
+     estimateVarianceFunctions(cds2)
> res <-                # Negative binomial test
+     nbinomTest(cds2, "T", "N")
```

Outline

- Work flow
  - Experiments and Technologies
  - Pre-processing
  - Analysis
  - Annotation and Integration
- Examples (Psuedo-Code)
  - Quality Assessment
  - 454 Microbiome Pre-Processing
  - Differential Expression
- ShortRead
  - Input and exploration
  - Manipulation
- Resources

ShortRead data input

```
> library(EatonEtAlChIPseq)
> fl <- system.file("extdata",
+   "GSM424494_wt_G2_orc_chip_rep1_S288C_14.mapview.txt.gz"
+   package="EatonEtAlChIPseq")
> aln <- readAligned(fl, type = "MAQMapview")
```

The *AlignedRead* class

```
> aln

class: AlignedRead
length: 478774 reads; width: 39 cycles
chromosome: S288C_14 S288C_14 ... S288C_14 S288C_14
position: 2 4 ... 784295 784295
strand: + - ... + +
alignQuality: IntegerQuality
alignData varLabels: nMismatchBestHit mismatchQuality nExac

> table(strand(aln), useNA="always")

     +      -      *   <NA>
 64170 414604      0      0
```

Accessing reads, base quality, and other data

```
> head(sread(aln), 3)

  A DNAStringSet instance of length 3
    width seq
[1]    39 CGGCTTTCTGACCG...AAAAATGAAAATG
[2]    39 GATTTATGAAAGAA...AAATGAAAATGAA
[3]    39 CTTTCTGACCGAAA...AATGAAAATGAAA
```

Alphabet by cycle

Expectation: nucleotide use independent of cycle

```
> alnp <- aln[strand(aln) == "+"]
> abc <- alphabetByCycle(sread(alnp))
> class(abc)
```

```
[1] "matrix"
```

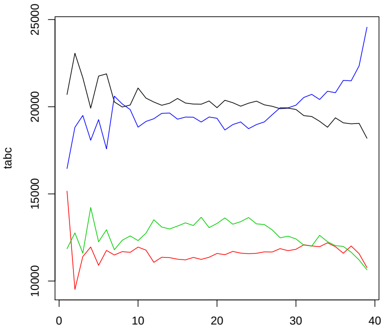
```
> abc[1:6,1:4]
```

```
        cycle
alphabet  [,1]  [,2]  [,3]  [,4]
       A 20701 23067 21668 19920
       C 15159  9523 11402 11952
       G 11856 12762 11599 14220
       T 16454 18818 19501 18078
       M     0     0     0     0
       R     0     0     0     0
```

Alphabet by cycle

`matplot` takes a matrix and plots each column as a set of points

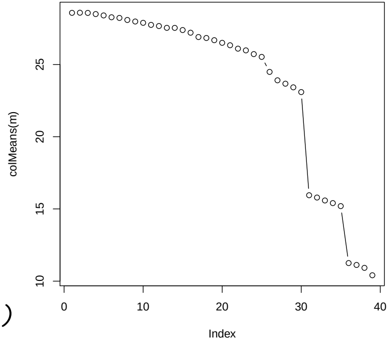
```
> tabc <- t(abc[1:4,])
> matplot(tabc, type="l",
+           lty=rep(1, 4))
```



Quality by cycle

Encoded quality scores can be decoded to their numerical values and represented as a matrix. Calculating the average of the column means creates a vector of average quality scores across cycle.

```
> m <- as(quality(alnp),
+           "matrix")
> plot(colMeans(m), type="b")
```



Recoding and updating

1. Access the chromosome information
2. Extract the chromosome number from the factor level
3. Recode the chromosome number to roman (!), create new levels, and update the chromosome
4. Update the *AlignedRead*

```
> chrom <- chromosome(alnp)
> i <- sub("S288C_([[:digit:]]+)", "\\1", levels(chrom))
> levels(chrom) <- paste("chr", as.roman(i), sep="")
> alnp <- renew(alnp, chromosome=chrom)
```

Outline

- Work flow
    - Experiments and Technologies
    - Pre-processing
    - Analysis
    - Annotation and Integration
- Examples (Psuedo-Code)
    - Quality Assessment
    - 454 Microbiome Pre-Processing
    - Differential Expression
- ShortRead
    - Input and exploration
    - Manipulation
- 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")`