High throughput sequence I/O, manipulation, and quality assessment

Martin Morgan (mtmorgan@fhcrc.org)

Fred Hutchinson Cancer Research Center

January 29, 2010

Work flow

Prior to analysis

- Biological preparation, e.g., ChIP.
- 'Sequencing': library preparation, cluster generation, imaging, ...

Analysis

1. Pre-processing, quality assessment, exploratory analysis
2. Domain-specific analysis
   - ChIP-seq
   - Digital gene expression
   - RNA-seq
   - Microbial / community structure
   - ...
3. Annotation & integration

GA II read characteristics and throughput

Read characteristics

- 30-100bp.
- Single-end: one end of the amplified fragment.
- Paired-end: both ends of the amplified fragment, $\approx$ 200bp apart.
- Mate pair: larger genomic sequence, circularized, fragmented to span circularized location, paired end sequencing.

Throughput (ours, November, 2009)

- 80bp sequences, 20 million reads per lane, 8 lanes per cell.

Other technologies

- Roche / 454: 300-500bp reads, 1 million reads.
- ABI SOLiD: 60 gigabase, 1 billion reads / run. High-accuracy reads from 'color-space' model (no *Bioconductor* support for color space).
- Also: Helicos (single-molecule); PacBio; ...

Bioconductor tools

- *IRanges*: range-based calculations, infrastructure, ...
- *Biostrings*: string manipulation, pattern matching, ...
- *BSgenome*: genome-scale data representations
- *ShortRead*: I/O, quality assessment, ...
- *GenomicFeatures*: transcript-level annotation (in development)

Today: Bloom et al., 2009

Background

- Bloom et al. (2009)
- 'Digital gene expression'
- Two yeast strains under two environmental conditions
- First-generation (GA I) Illumina technology
- Comparable array data available: Smith and Kruglyak (2008)

Data

- Authors provide *fastq* files containing reads and base qualities
- 'We' aligned reads to reference genomes using the *Bowtie* aligment program. Details avialable

```
> file.show(system.file("extdata", "README.TXT",
+                       package="day3"))
```

- *Bowtie* produces a file containing reads, base qualities, and the locations where the reads best align.

ShortRead

Functions to explore in *ShortRead*

Input	`readAligned`
Accessors	`sread`, `quality`, `strand`, `chromosome`
Coercion, update	`as`, `initialize`
Summary	`alphabetByCycle`, `qa`, `report`

Other

String and factor manipulation	`levels`, `sub`
Coercion	`as.vector`, `as.romanb`
Plotting	`plot`, `matplot`
Summary	`table`, `head`, `tail`, matrix `rowMeans`, `colMeans`, `row`, `col`

ShortRead data input

```
> library(ShortRead)

> bowtieFile <-
+     system.file("extdata", "BYe9.head.map",
+                 package="day3")
> aln <- readAligned(bowtieFile, type = "Bowtie")
> aln

class: AlignedRead
length: 1000000 reads; width: 32 cycles
chromosome: chrmt_S288C chrmt_S288C ... chr12_S288C chr12_S
position: 7021 12161 ... 446999 461957
strand: - - ... + -
alignQuality: NumericQuality
alignData varLabels: similar mismatch
```

The *AlignedRead* class

```
> aln

class: AlignedRead
length: 1000000 reads; width: 32 cycles
chromosome: chrmt_S288C chrmt_S288C ... chr12_S288C chr12_S
position: 7021 12161 ... 446999 461957
strand: - - ... + -
alignQuality: NumericQuality
alignData varLabels: similar mismatch

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

     -      +      *   <NA>
508233 491767      0      0
```

Accessing reads, base quality, and other data

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

  A DNAStringSet instance of length 3
    width seq
[1]    32 GATTTTATTTTTAA...TTATATATATATA
[2]    32 TATGCCAAATACCA...TTAATTAATTAAA
[3]    32 GTATTTCGTTGATA...GTGGCTATATAGT

> tail(quality(aln), 3)

class: FastqQuality
quality:
  A BStringSet instance of length 3
    width seq
[1]    32 hhhhhhhhhhhhhh...hhhhhhhhhhhhh
[2]    32 hhhhhhhShhVhhh...h[hdFhFhNS[^?
[3]    32 hhhhhhhhhhhhhh...hhhhhhhhahhhh
```

Alphabet by cycle

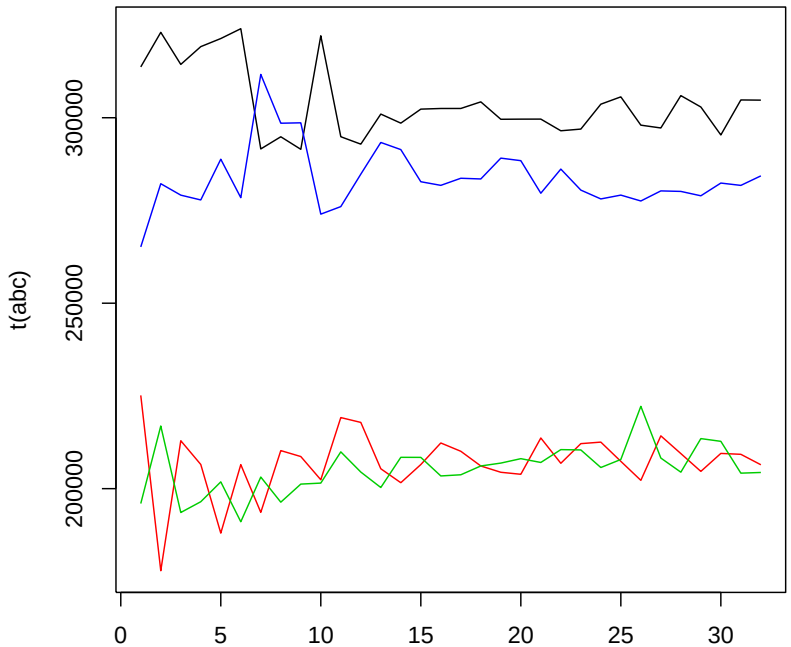
Expectation: nucleotide independent of cycle

```
> abc <- alphabetByCycle(sread(aln))
> class(abc)

[1] "matrix"

> abc[1:6,1:4]

        cycle
alphabet   [,1]   [,2]   [,3]   [,4]
       A 313699 323017 314371 319165
       C 225143 177874 212925 206521
       G 195988 216896 193575 196479
       T 265170 282213 279129 277835
       M      0      0      0      0
       R      0      0      0      0

> abc <- abc[1:4,]
```

Alphabet by cycle

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

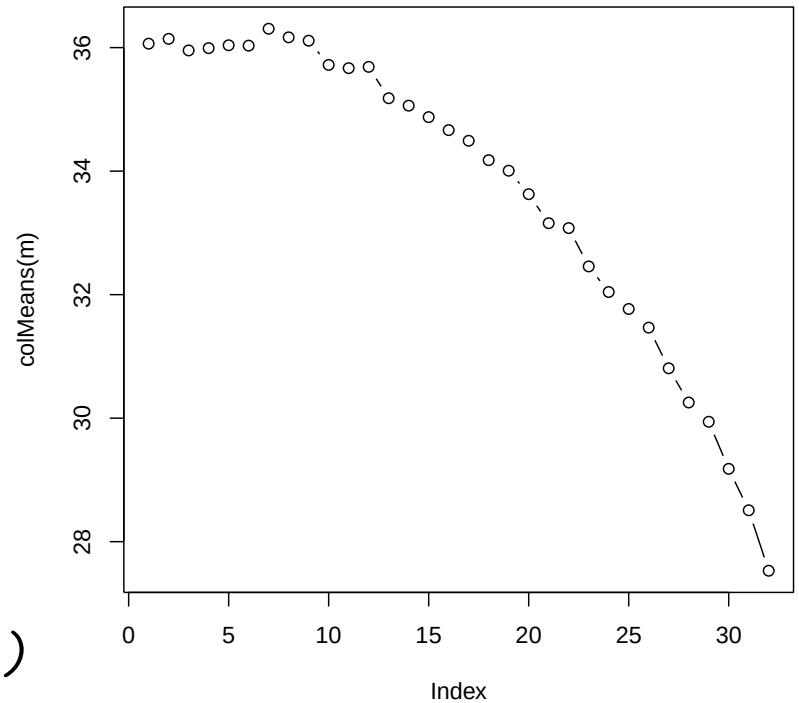
```
> matplot(t(abc), type="l",
+         lty=rep(1, 4))
```



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(aln),
+         "matrix") - 33
> plot(colMeans(m), type="b")
```



Subsetting *AlignedRead*

Two issues:

- Mitochondrial genome a 'special case'
- Chromosomes labels different from other resources

Creating a subset of aligned reads

```
> tail(table(chromosome(aln)))
```

```
chr12_S288C chr13_S288C chr14_S288C
     562222       26811       24655
chr15_S288C chr16_S288C chrmt_S288C
      27330       26763      128230
```

```
> aln <- aln[chromosome(aln) != "chrmt_S288C"]
```

Recoding and updating

1. Access the chromosome
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(aln)
> i <- sub("chr(.+)_S288C", "\\1", levels(chrom))
> levels(chrom) <- paste("chr", as.roman(i), sep="")
> aln <- initialize(aln, chromosome=chrom)
```

Quality assessment

Two-step process

1. qa: visit each input file and collate statistics. Long and computationally intensive; can be done in parallel.
2. report: summarize collected statisitcs into an HTML-based report

```
> ## bowtieDir <- "/path/to/alignments"
> ## qa <- qa(bowtieDir, ".*map$", type="Bowtie")
> data("qa_caudy_28_jan_2009")
> rpt <- report(qa)
> browseURL(rpt)
```

References

Joshua Bloom, Zia Khan, Leonid Kruglyak, Mona Singh, and Amy Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, 10(1):221, 2009. ISSN 1471-2164. doi: 10.1186/1471-2164-10-221. URL `http://www.biomedcentral.com/1471-2164/10/221`.

Erin N Smith and Leonid Kruglyak. Gene-environment interaction in yeast gene expression. *PLoS Biol*, 6(4):e83, 04 2008. doi: 10.1371/journal.pbio.0060083. URL `http://dx.doi.org/10.1371%2Fjournal.pbio.0060083`.