R / Bioconductor packages for high-throughput sequence analysis: work flows and data management

Martin Morgan
Bioconductor / Fred Hutchinson Cancer Research Center
Seattle, WA, USA

8-10 June 2009

Bioconductor and high-throughput sequencing

Bioconductor

- Open source, open development, based on the R statistical programming language
- $>$ 300 contributed and internally developed packages for microarray, flow cytometry, high-throughput sequence, . . . analysis

High-throughput sequencing

- Focus especially on down-stream (e.g., after alignment) analysis
- Quality assessment, data manipulation, ChIP-seq (and other) peak calling, annotation, visualziation

Packages for high-throughput analysis

Currently 'releasesd'

- ShortRead: I/O and quality assessment
- Biostrings: Sequence manipulation, pattern matching
- BSgenome: Whole-genome representations and manipulation
- IRanges, genomeIntervals: Range-based calculations
- rtracklayer: 'Genome browser' input and output
- HilbertViz: advanced visualization

In development

- chipseq: ChIP-seq specific tools
- Contributed packages for base calling, ...

Additional R / Bioconductor packages, e.g., AnnotationDbi (gene-centric annotation), lattice (graphics), edgeR (regression analysis of count data), ...

Work flows

1. Biological sample preparation
2. Sequencing: base calls, quality scores
3. Alignment: manufacturer or third-party tools; also Biostrings (especially for specialized tasks)
4. Input, quality assessment, remediation – ShortRead, Biostrings
5. Appilcation-specific processing – chipseq, BSgenome, IRanges, third-party tools
6. Annotation, visualization, genome browser manipulation – rtracklayer, HilbertViz, ...

Input

- Manufacturer files, e.g., Solexa aligned reads and qualities
- Third party software
  - MAQ, Bowtie, SOAP, ...
  - fastq (sequence and base quality scores), fasta
- Other data sources: Delimited text (e.g., `readXStringColumns`), data base (RSQLite, RMySQL, ...), NetCDF (ncdf), ...

```
> library(ShortRead)
> dir <- "~/proj/a/bioC/Courses/EMBL2009/extdata"
> aln <- readAligned(dir, ".*1985.map",
+     "MAQMapShort")
```

Exploration & quality assessment

- Query object, e.g., aligned chromosome, strand, and position; alignment score, file-specific information, . . .
- Explore, e.g., nucleotide frequency, position-specific quality, duplicated reads, . . .
- Manipulate, e.g., trim leading / trailing nucleotides, filter uninformative information
- Perform quality assessment

```
> levels(chromosome(aln))
> sum(quality(alignQuality(aln)) == 0)
> filt <- compose(chromosomeFilter("_random",
+     fixed = TRUE, exclude = TRUE), alignQualityFilter(1L)
> faln <- aln[filt(aln)]
> xtabs(~chromosome(faln) + strand(faln))
```

Transform for relevant biological question

- Coverage, e.g., for ChIP-seq; read depth at each aligned position
- Position-specific consensus matrix, e.g., for SNP analyses
- ...

```
> library(BSgenome.Mmusculus.UCSC.mm9)
> Mmusculus
> seqlen <- seqlengths(Mmusculus)
> seqlen <- seqlen[names(seqlen) %in% levels(chromosome(fal
> cvg <- coverage(faln, start=1L, end=seqlen)
```

Manipulate and export

- Coverage and other objects
  - Small and easily manipulated
  - Easily leverage existing R infrastructure, e.g., for linear models
- Visualize using R packages, particularly lattice
- Export to genome browsers

```
> library(rtracklayer)
> chr1 <- cvg[["chr1"]]
> chr1 >= 10
> chr1peak <- chr1 * (chr1 >= 10)
> chr1peaka <- slice(chr1, lower = 10)
> rng <- as(chr1, "RangedData")
> export(rng, "/tmp/chr1.wig")
```

Summary

- Tools for diverse high-throughput analyses
- Performant – reasonable memory management, fast operations
- Flexible, both established work flows and connection with R functionality
- Open-ended and interactive – ability to develop creative, customized analyses addressing truly novel challenges

Resources

- Bioconductor web site `http://bioconductor.org`
- `bioconductor` and `bioc-sig-seq` mailing lists