Bioconductor for Sequence Analysis

Alex Sánchez[1]

June 7, 2014

[1] Adapted from MArtin Morgan's slides

Introduction: What is *Bioconductor* good for?

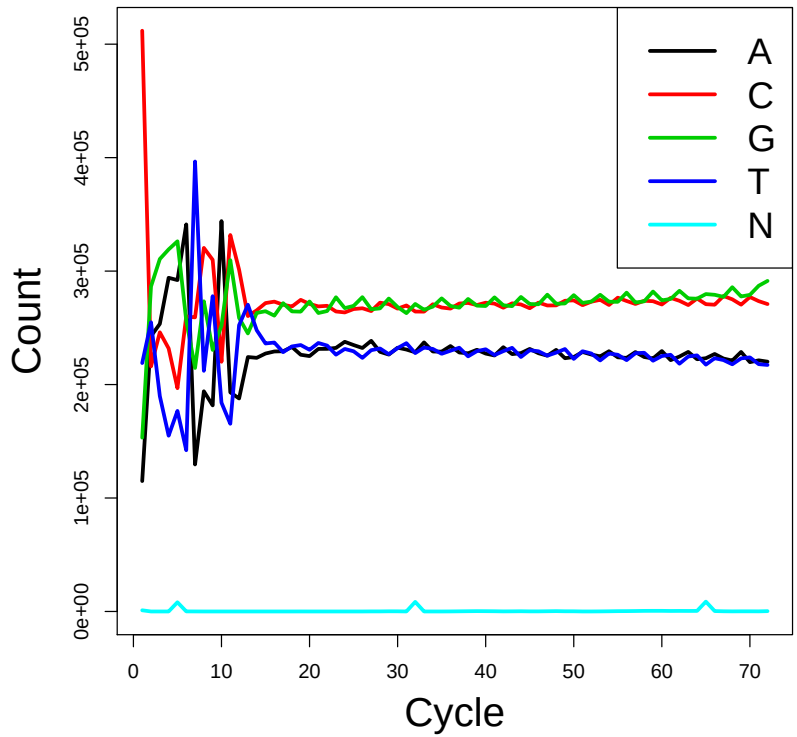
- Sequencing: RNA-seq, ChIP-seq, called variants, ...
  - Especially *after* assembly / alignment
- Annotation: genes, pathways, gene models (exons, transcripts, etc.), ...
- Microarrays: expression, copy number, SNPs, methylation, ...
- Flow cytometry, proteomics, image analysis, high-throughput screens, ...

Sequencing: The *ShortRead* package

```
## Use the 'ShortRead' package
library(ShortRead)
## Create an object to represent a sample from a file
sampler <- FastqSampler("ERR127302_1.fastq.gz")
## Apply a method to yield a random sample
fq <- yield(sampler)
## Access sequences of sampled reads using `sread()`
## Summarize nucleotide use by cycle
## 'abc' is a nucleotide x cycle matrix of counts
abc <- alphabetByCycle(sread(fq))
## Subset of interesting nucleotides
abc <- abc[c("A", "C", "G", "T", "N"),]
```

Sequencing: The *ShortRead* package

```
## Create a plot from a
## matrix
matplot(t(abc), type="l",
  lty=1, lwd=3,
  xlab="Cycle",
  ylab="Count",
  cex.lab=2)
## Add a legend
legend("topright",
  legend=rownames(abc),
  lty=1, lwd=3, col=1:5,
  cex=1.8)
```



Sequencing: Essential packages and classes

- *Biostrings* and *DNAStringSet*
- *GenomicAlignments* and *GAlignments*
- *GenomicRanges* and *GRanges*
- *GenomicFeatures* and *TranscriptDb*
- *VariantAnnotation* and *VCF*
- Input and output: *rtracklayer* (WIG, BED, etc.), *Rsamtools* (BAM), *ShortRead* (FASTQ) file input

Reads

Data	Short reads and their qualities
Tasks	Input, quality assessment, summary, trimming, ...
Packages	*ShortRead*, *Biostrings*
Functions	▶ `readFastq`, `FastqSampler`, `FasqtStreamer`. ▶ `qa`, `report`. ▶ `alphabetFrequency`, `alphabetByCycle`, `consensusMatrix`. ▶ `trimTails`, `trimLRPatterns`, `matchPDict`, ...

Alignments

Data BAM files of aligned reads

Tasks Input, BAM file manipulation, pileups

Packages *GenomicAlignments*, *Rsamtools* (also: *GenomicRanges*)

Functions

- `readGAlignments`
- `BamFile`, `BamFileList`
- `scanBam`, `ScanBamParam` (select a subset of the BAM file)
- `asBam`, `sortBam`, `indexBam`, `mergeBam`, `filterBam`
- `BamSampler`, `applyPileups`

Ranges

Data Genomic coordinates to represent data (e.g., aligned reads) or annotation (e.g., gene models).

Tasks Input, counting, coverage, manipulation, ...

Packages *GenomicRanges*, *IRanges*

Functions

- `readGAlignments`, `readGAlignmentsList`
- Many intra-, inter-, and between-range manipulating, e.g., `narrow`, `flank`, `shift`, `intersect`, `findOverlaps`, `countOverlaps`

Variants

Data	VCF (Variant Call Format) file
Tasks	Calling, input, summary, coding consequences
Packages	*VariantTools* (linux only), *VariantAnnotation*, *ensemblVEP*
Functions	▶ `tallyVariants` ▶ `readVcf`, `locateVariants`, `predictCoding` ▶ Also: SIFT, PolyPhen data bases

Annotations

Data	Gene symbols or other identifiers
Tasks	Discover annotations associated with genes or symbols
Packages	*AnnotationDbi* (*org.**, *GO.db*, ...), *biomaRt*
Functions	▶ Discovery: `columns`, `keytype`, `keys` ▶ `select`, `merge` ▶ *biomaRt*: `listMarts`, `listDatasets`, `listAttributes`, `listFilters`, `getBM`

Features

Data	Genomic coordinates
Tasks	Group exons by transcript or gene; discover transcript / gene identifier mappings
Packages	*GenomicFeatures* and *TxDb.** packages (also: *rtracklayer*)
Functions	▶ `exonsBy`, `cdsBy`, `transcriptsBy` ▶ `select` (see Annotations, below) ▶ `makeTranscriptDb*`

Genome annotations

Data	FASTA, GTF, VCF, . . . from internet resources
Tasks	Define regions of interests; incorporate known features (e.g., ENCODE marks, dbSNP variants) in work flows
Packages	*AnnotationHub*
Functions	▶ `AnnotationHub`, `filters` ▶ `metadata`, `hub$<tab>`

Sequences

Data Whole-genome sequences

Tasks View sequences, match position weight matricies, match patterns

Packages *Biostrings*, *BSgenome*

Functions

- `available.genomes`
- `Hsapiens[["chr3"]]`, `getSeq`, `mask`
- `matchPWM`, `vcountPattern`, `...`
- `forgeBSgenomeDataPkg`

Import / export

Data	Common text-based formats, `gff`, `wig`, `bed`; UCSC tracks
Tasks	Import and export
Packages	*rtracklayer*
Functions	▶ `import`, `export` ▶ `browserSession`, `genome`

And...

Data representation: *IRanges*, *GenomicRanges*, *GenomicFeatures*, *Biostrings*, *BSgenome*, *girafe*. Input / output: *ShortRead* (fastq), *Rsamtools* (bam), *rtracklayer* (gff, wig, bed), *VariantAnnotation* (vcf), *R453Plus1Toolbox* (454). Annotation: *GenomicFeatures*, *ChIPpeakAnno*, *VariantAnnotation*. Alignment: *Rsubread*, *Biostrings*. Visualization: *ggbio*, *Gviz*. Quality assessment: *qrqc*, *seqbias*, *ReQON*, *htSeqTools*, *TEQC*, *Rolexa*, *ShortRead*. RNA-seq: *BitSeq*, *cqn*, *cummeRbund*, *DESeq*, *DEXSeq*, *EDASeq*, *edgeR*, *gage*, *goseq*, *iASeq*, *tweeDEseq*. ChIP-seq, etc.: *BayesPeak*, *baySeq*, *ChIPpeakAnno*, *chipseq*, *ChIPseqR*, *ChIPsim*, *CSAR*, *DiffBind*, *MEDIPS*, *mosaics*, *NarrowPeaks*, *nucleR*, *PICS*, *PING*, *REDseq*, *Repitools*, *TSSi*. Motifs: *BCRANK*, *cosmo*, *cosmoGUI*, *MotIV*, *seqLogo*, *rGADEM*. 3C, etc.: *HiTC*, *r3Cseq*. Copy number: *cn.mops*, *CNAnorm*, *exomeCopy*, *seqmentSeq*. Microbiome: *phyloseq*, *DirichletMultinomial*, *clstutils*, *manta*, *mcaGUI*. Work flows: *ArrayExpressHTS*, *Genominator*, *easyRNASeq*, *oneChannelGUI*, *rnaSeqMap*. Database: *SRAdb*. ...