Annotation and High Throughput Sequencing

Martin Morgan
Fred Hutchinson Cancer Research Center

19-21 January, 2011

Annotation Resources – Genes and Genomes

AnnotationDbi

- Chip, 'org', GO, KEGG, homology
- Curated from NCBI, GO, other sources for each *Bioconductor* release.
- SQL 'under the hood'

biomaRt

- Large online annotation collection
- Curated by OICR / EMBL-EBI

BSgenome

- Genome sequences – try `available.genomes`

Demo

AnnotationDbi, *biomaRt*

Work Flow: Sequence Analysis

Prior to analysis

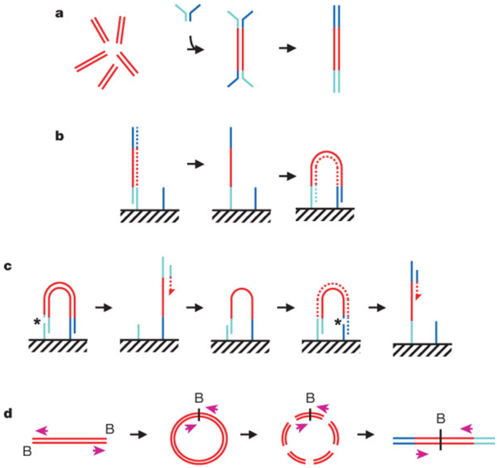
- Biological experimental design – treatments, replication, etc.
- Sequencing preparation – library preparation, manufacturer protocol, etc.

Analysis

1. Pre-processing (sequencing, alignment, quality assessment)
2. Count, e.g., reads per transcript – ChIP-seq; RNA-seq; novel transcript identification; microbiome; . . .
3. Differential representation / ChIP-seq / SNP / . . .
4. Annotation
5. . . .

`http://bioconductor.org/workflows` for common analyses.

Bridge PCR



Bentley et al., 2008, Nature 456: 53-9

Bioconductor entry points

- Quality assessment.
- Preliminary read processing, e.g., demultiplexing, remediation
- Specialized alignment, e.g., `matchPDict` in *Biostrings*.
- 'Upstream' domain-specific work flows, e.g., ChIP-seq peak calling (*chipseq*), RNA-seq reads per transcript (*GenomicRanges* / *IRanges* / . . .)
- Statistical analysis of designed experiments, e.g., *edgeR*, *DESeq*
- Specialized analysis, e.g., microbiome sequence processing and ecological analysis (*vegan*, *ape*, . . .)

Sequence I/O

Packages

Biostrings	DNA sequence, pattern matching
Rsamtools	BAM manipulation
ShortRead	‘traditional’ aligned reads; quality assessment
rtracklayer	GFF and other formats; browser interaction
GenomicRanges	Regions of interest / aligned reads as collections of ranges on genomes

Functions

- `readFasta`, `readFastq`, `writeFasta`, `writeFastq`
- `scanBam` (also sort, index, filter BAM files; BCF, indexed fasta)
- `import` / `export` (for GFF & friends)
- `readAligned`, `readGappedAlignments`

Representing Sequence Information

DNAStringSet

- Collections of DNA sequences, e.g., microarry probes, Illumina reads
- Quality scores

GRanges

- Genome coordinates – reference sequence name, start and end coordinates, strand; e.g., aligned reads
- *GRangesList* – hierarchical structure, e.g., exons within transcripts

Additional classes: *AlignedRead*, *GappedAlignment*, . . .

Sequence Annotations

- Existing infrastructure for gene-level annotation

GenomicFeatures

- Idea: retrieve annotations from common sources, e.g., UCSC genome browser 'known genes' track; save as a local data base.
- Query for regions of interest, e.g., exons per transcript

Demo

DNAStringSet, *GRanges*, *AlignedRead* and *GappedAlignment*, *GenomicFeatures*

Lab activity

Goal: Explore sequences and their annotation

1. Data input and exploration
2. Gapped alignments
3. Transcript annotations
4. Counting reads aligned to regions
5. (Differential representation)
6. Annotation to biological function

Example Data

Nagalakshmi et al., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science* 320: 1344–1349 [**?**].

- Original ‘RNA-seq’ experiment
- Two different primers to generate DNA from poly(A) RNA:

RH	Random hexamer
dT	oligo(dT)

- Biological and technical replicates
- Illumina GAI – relatively small number (<5 million / lane) of short (33bp) reads; poor trailing base quality.