

An overview of yeastRNASeq

James Bullard and Kasper D. Hansen

Modified: 15 January 2010. Compiled: January 15, 2010

```
> require(yeastRNASeq)
```

This package contains data from

```
> x <- citation(package = "yeastRNASeq")[[1]]
> utils:::print.citation(x, bibtex = FALSE)
```

Albert Lee, Kasper D. Hansen, James Bullard, Sandrine Dudoit, Gavin Sherlock (2008). Novel low abundance and transient RNAs in yeast revealed by tiling microarrays and ultra high-throughput sequencing are not conserved across closely related yeast species.
PLoS Genet 4(12): e1000299.
doi:10.1371/journal.pgen.1000299

which describes some experiments in *S. cerevisiae* comparing various mutant strains to a wild-type strain. A full BibTex entry can be obtained by

```
> citation("yeastRNASeq")
```

The subset of the data which this package contains is more specifically data from a wild-type and a single mutant yeast. For each condition (mutant, wild-type) there is two lanes worth of data, each lane containing a sample of 500,000 raw (unaligned) reads from each of 2 lanes each. Each of the four lanes have been aligned against the yeast genome using Bowtie.

The raw reads are contained in 4 FASTQ files and the Bowtie alignment are contained in 4 Bowtie output files. There are 500,000 reads in each of the FASTQ files and fewer reads in each of the Bowtie files. The filenames are

```

> list.files(file.path(system.file(package = "yeastRNASeq"),
+   "reads"))

[1] "mut_1_f.bowtie.gz" "mut_1_f.fastq.gz" "mut_2_f.bowtie.gz"
[4] "mut_2_f.fastq.gz"  "wt_1_f.bowtie.gz"  "wt_1_f.fastq.gz"
[7] "wt_2_f.bowtie.gz"  "wt_2_f.fastq.gz"

```

The reads were aligned to the yeast genome obtained from <ftp://ftpmips.gsf.de/yeast/sequences> (which was the basis for the Bowtie index available at the Bowtie website at the time of alignment).

These files are ready to be parsed by the tools in the ShortRead package. As an example we read the alignment files by

```

> require(ShortRead)
> files <- list.files(file.path(system.file(package = "yeastRNASeq"),
+   "reads"), pattern = "bowtie", full.names = TRUE)
> names(files) <- gsub("\\.bowtie.*", "", basename(files))
> names(files)

[1] "mut_1_f" "mut_2_f" "wt_1_f"  "wt_2_f"

```

```

> aligned <- lapply(files, readAligned, type = "Bowtie")

```

The constructed object `aligned` is a list with 4 elements. Each element correspond to a lane and is an object of class `AlignedRead`.

The output from this operation has already been stored as an R object and is accessible by

```

> data(yeastAligned)
> yeastAligned[["mut_1_f"]]

class: AlignedRead
length: 423318 reads; width: 26 cycles
chromosome: Scchr05 Scchr15 ... Scchr08 Scchr13
position: 541317 885627 ... 488228 667296
strand: - + ... - +
alignQuality: NumericQuality
alignData varLabels: similar mismatch

```

The percent of aligned reads is

```
> round(sapply(aligned, length)/5e+05, 2)
```

```
mut_1_f mut_2_f wt_1_f wt_2_f
  0.85    0.84    0.82    0.86
```

There are two additional objects available in the package, purely for illustrative purposes (do not use them for analysis). The object `yeastAnno` is annotation obtained from Ensembl using `biomaRt` and is a `data.frame` of annotation:

```
> data(yeastAnno)
```

```
> dim(yeastAnno)
```

```
[1] 7124    6
```

```
> head(yeastAnno, n = 2)
```

```
ensembl_gene_id chromosome_name start_position end_position strand
1      YHR055C          VIII      214535      214720      -1
2      YPR161C          XVI      864445      866418      -1
  gene_biotype
1 protein_coding
2 protein_coding
```

```
> table(yeastAnno$gene_biotype)
```

ncRNA	protein_coding	pseudogene	rRNA
9	6698	21	14
snoRNA	snRNA	tRNA	
77	6	299	

The other object is called `geneLevelData` and is a `matrix` of counts per gene.

```
> data(geneLevelData)
```

```
> head(geneLevelData, n = 2)
```

```
      mut_1 mut_2 wt_1 wt_2
YHR055C    0    0    0    0
YPR161C   38   39   35   34
```

Such a matrix may be constructed from `yeastAligned` and `yeastAnno` using either the functionality in the `IRanges` and `ShortRead` packages or by using the functionality of the `Genominator` package (which also contains a vignette describing a simplified differential analysis of this dataset).

Note that the data does not contain any biological replicates. In the original publication this was addressed by also analyzing a set of tiling microarrays.