Assessment: Count table creation

We will first examine the read counts in genes generated with `summarizeOverlaps()` using:

- `genes()`, which gives a range for each gene, from the start position to the end position,
- `exonsBy()`, which produces a *GRangesList*, with a *GRanges* of the exons for each gene.

Load the package with the BAM files and the transcript database:

```
library(pasillaBamSubset)
library(TxDb.Dmelanogaster.UCSC.dm3.ensGene)
txdb <- TxDb.Dmelanogaster.UCSC.dm3.ensGene
```

Create the genes object and subset to chromosome 4:

```
g <- genes(txdb)
g <- g[seqnames(g) == "chr4"]
```

Create the *GRangesList* of exons by gene, and subset to the same genes as in g:

Generating Speech Output

```
grl <- exonsBy(txdb, by="gene")
grl <- grl[names(g)]
```

Test for the same names:

```
all.equal(names(g), names(grl))
```

Determine the path to the BAM file of single-end RNA-seq data that we have been using:

```
library(Rsamtools)
bf <- BamFile(untreated1_chr4())
```

Question 1

1/1 point (graded)

Load the **GenomicAlignments** package:

```
library(GenomicAlignments)
```

Now call the `summarizeOverlaps()` function on the *BamFile* `bf` once using the *GRanges* object `g`, and once using the *GRangesList* object `grl` with exons per gene. Since the experiment was not strand-specific, the strand information in the BAM file is not biologically meaningful. Use the `ignore.strand=TRUE` setting to when finding overlaps. You do not need to specify any additional arguments.

If desired, you can plot the reads overlapping `g` and `grl` in a scatterplot, on the log scale as in the previous video. Add an `abline` with `a=0`, and `b=1`.

Generating Speech Output

o of the counts in `grl` and `g`, after removing genes where `g` had zero counts?

0.8929273 ✔

0.8929273

Submit You have used 2 of 10 attempts

In the exercise above, the *GRangesList* feature set gives a lower count value because this method of counting does not include reads which fall in introns. If we include only the exons as features (as in the *GRangesList* object above), the reads must fall in exons to be counted. There are many options for counting modes which can be specified to the `summarizeOverlaps` function, including fully custom counting rules provider by the user as a function. See `?summarizeOverlaps` for more details.

One reason for preferring the results based on the `grl` *GRangesList* above is that reads which fall in introns may not be from the fully mature mRNA. Instead, these might be contamination of precursor mRNA or possibly from some other source, leading to noisier and less reliable estimates of gene expression.

Question 2

1/1 point (graded)

The number of reads which align to a gene depend upon, among other factors, the sequencing depth (the total number of sequenced reads), and the size of the gene.

Let's consider the single column count table obtained using the `grl` *GRangesList* above. Divide the counts in this table by [redacted] to obtain the proportion of reads aligning to each read. Now multiply these proportions by 1 million. This operation scales each column of the count table such that we get the number of reads expected if the

Generating Speech Output

sample were sequenced to have 1 million reads mapping to genes on chromosome 4.

In this dataset, we are only looking at the subset of reads mapping to chromosome 4. Suppose, however, that this were the complete set of mapped reads for this sample. Then, the computed quantities would be the number of reads mapping to each gene for every million mapping reads. This quantity is commonly referred to as reads per million (RPM). The term *fragments* per million, FPM, is also often used. ("Fragments" are used more generally, because in a paired-end experiment two *reads* represent one observed DNA or RNA *fragment* and we are almost always more interested in the fragment counts rather than the read counts.)

Supposing that all reads mapped to chromosome 4, what is the FPM for the first gene, FBgn0002521?

4275.607 ✔

4275.607

Submit You have used 1 of 5 attempts

Question 3

1/1 point (graded)

We said that the count depends on the number of mapped reads, and also on the size of the gene. If you think about reads randomly arising from fragmented RNA molecules, the number of reads (or fragments) should be proportional to the number of basepairs in the exons (roughly), if all other variables are held equal (gene expression and sequencing depth).

Generating Speech Output

Using the `reduce()` and `width()` functions, compute the total width of the exonic regions of each gene in `grl`. The `reduce()` call is necessary to prevent overcounting positions in overlapping exons. Check the `summary()` of these values. The mean should be `4383`.

Divide the FPM values by the number of basepairs in the exons of each gene, and then finally multiply by 1000. This is called the FPKM, the number of fragments per kilobase of exonic basepairs, per million mapped reads (again we suppose that we've observed all the genes and reads, rather than just the ones on chromosome 4).

What is the FPKM for the first gene, FBgn0002521?

1782.245648 ✔

1782.245648

Submit You have used 1 of 5 attempts

Note that there are more robust ways of estimating FPM and FPKM. For example, using a more robust estimate of the sequencing depth, and taking into account multiple transcript isoforms, which might be expressed at different levels. Here, we have demonstrated the simplest kind of sequencing depth and gene length normalization introduced for RNA-seq analysis.

Discussion

Hide Discussion

Generating Speech Output NGS data

Add a Post

Show all posts ▼ by recent activity ▼

Reads difference in two objects: genes vs. exon?

For question 1, I have checked the respective reads of the genes in "g" and "grl". Something I found a bit weird to me is that in some case,... 4

Generating Speech Output