Some Basic Analysis of ChIP-Seq Data

July 23, 2010

Our goal is to describe the use of Bioconductor software to perform some basic tasks in the analysis of ChIP-Seq data. We will use several functions in the as-yet-unreleased chipseq package, which provides convenient interfaces to other powerful packages such as ShortRead and IRanges. We will also use the lattice and rtracklayer packages for visualization.

```r
> library(chipseq)
> library(GenomicFeatures)
> library(lattice)
```

Example data

The cstest data set is included in the chipseq package to help demonstrate its capabilities. The dataset contains data for three chromosomes from Solexa lanes, one from a CTCF mouse ChIP-Seq, and one from a GFP mouse ChIP-Seq. The raw reads were aligned to the reference genome (mouse in this case) using an external program (MAQ), and the results read in using the the readAligned function in the ShortRead, in conjunction with a filter produced by the chipseqFilter function. This step filtered the reads to remove duplicates, to restrict mappings to the canonical, autosomal chromosomes and ensure that only a single read maps to a given position. A quality score cutoff was also applied. The remaining data were reduced to a set of aligned intervals (including orientation). This saves a great deal of memory, as the sequences, which are unnecessary, are discarded. Finally, we subset the data for chr10 to chr12, simply for convenience in this vignette.

We outline this process with this unevaluated code block:

```r
> qa_list <- lapply(sampleFiles, qa)
> report(do.call(rbind, qa_list))
> ## spend some time evaluating the QA report, then proceed
> filter <- compose(chipseqFilter(), alignQualityFilter(15))
> cstest <- GenomicRangesList(lapply(sampleFiles, function(file) {
+ as(readAligned(file, filter), "GRanges")
+ }));
> cstest <- cstest[seqnames(cstest) %in% c("chr10", "chr11", "chr12")]
```

The above step has been performed in advance, and the output has been included as a dataset in this package. We load it now:

```r
> data(cstest)
> cstest
```
GRangesList object of length 2:
$cctcf
GRanges object with 450096 ranges and 0 metadata columns:

```
seqnames ranges strand
<Rle> <IRanges> <Rle>

[1] chr10  3012936-3012959 +
[2] chr10  3012941-3012964 +
[3] chr10  3012944-3012967 +
[4] chr10  3012955-3012978 +
[5] chr10  3012963-3012986 +
...
[450092] chr12  121239376-121239399 -
[450093] chr12  121245849-121245872 -
[450094] chr12  121245895-121245918 -
[450095] chr12  121246344-121246367 -
[450096] chr12  121253499-121253522 -
```

<1 more element>
-------
seqinfo: 35 sequences from an unspecified genome

cstest is an object of class GRangesList, and has a list-like structure, each component representing the alignments from one lane, as a GRanges object of stranded intervals.

```
> cstest$cctcf
GRanges object with 450096 ranges and 0 metadata columns:

```
seqnames ranges strand
<Rle> <IRanges> <Rle>

[1] chr10  3012936-3012959 +
[2] chr10  3012941-3012964 +
[3] chr10  3012944-3012967 +
[4] chr10  3012955-3012978 +
[5] chr10  3012963-3012986 +
...
[450092] chr12  121239376-121239399 -
[450093] chr12  121245849-121245872 -
[450094] chr12  121245895-121245918 -
[450095] chr12  121246344-121246367 -
[450096] chr12  121253499-121253522 -
```
-------
seqinfo: 35 sequences from an unspecified genome

Extending reads

Solexa gives us the first few (24 in this example) bases of each fragment it sequences, but the actual fragment is longer. By design, the sites of interest (transcription factor binding sites) should be somewhere in the fragment, but not necessarily in its initial part. Although the actual lengths of
fragments vary, extending the alignment of the short read by a fixed amount in the appropriate direction, depending on whether the alignment was to the positive or negative strand, makes it more likely that we cover the actual site of interest.

It is possible to estimate the fragment length, through a variety of methods. There are several implemented by the `estimate.mean.fraglen` function. Generally, this only needs to be done for one sample from each experimental protocol. Here, we use SSISR, the default method:

```r
> fraglen <- estimate.mean.fraglen(cstest$ctcf, method="correlation")
> fraglen[!is.na(fraglen)]
```

```r
chr10 chr11 chr12
340 340 340
```

Given the suggestion of 190 nucleotides, we extend all reads to be 200 bases long. This is done using the `resize` function, which considers the strand to determine the direction of extension:

```r
> ctcf.ext <- resize(cstest$ctcf, width = 200)
> ctcf.ext
```

```r
GRanges object with 450096 ranges and 0 metadata columns:
  seqnames ranges strand
  <Rle> <IRanges> <Rle>
  [1] chr10 3012936-3013135 +
  [2] chr10 3012941-3013140 +
  [3] chr10 3012944-3013143 +
  [4] chr10 3012955-3013154 +
  [5] chr10 3012963-3013162 +
  ... ... ... ...
  [450092] chr12 121239200-121239399 -
  [450093] chr12 121245673-121245872 -
  [450094] chr12 121245719-121245918 -
  [450095] chr12 121246168-121246367 -
  [450096] chr12 121253323-121253522 -
```

seqinfo: 35 sequences from an unspecified genome

We now have intervals for the CTCF lane that represent the original fragments that were precipitated.

**Coverage, islands, and depth**

A useful summary of this information is the `coverage`, that is, how many times each base in the genome was covered by one of these intervals.

```r
> cov.ctcf <- coverage(ctcf.ext)
> cov.ctcf
```

```r
RleList of length 35
$chr1
integer-Rle of length 197195432 with 1 run
  Lengths: 197195432
```
For efficiency, the result is stored in a run-length encoded form.
The regions of interest are contiguous segments of non-zero coverage, also known as *islands*.

```r
> islands <- slice(cov.ctcf, lower = 1)
> islands
RleViewsList of length 35
$chr1
Views on a 197195432-length Rle subject
views: NONE

$chr2
Views on a 181748087-length Rle subject
views: NONE

$chr3
Views on a 159599783-length Rle subject
views: NONE

...  
<32 more elements>

For each island, we can compute the number of reads in the island, and the maximum coverage depth within that island.

```r
> viewSums(islands)

IntegerList of length 35
[["chr1"]]
integer(0)
[["chr2"]]
integer(0)
[["chr3"]]
integer(0)
[["chr4"]]
integer(0)
[["chr5"]]
integer(0)
[["chr6"]]
integer(0)
[["chr7"]]
integer(0)
[["chr8"]]
integer(0)
[["chr9"]]
integer(0)
[["chr10"]]
2400 200 200 200 200 200 200 600 ... 200 200 400 200 200 200 200 200

...  
<25 more elements>

> viewMaxs(islands)

IntegerList of length 35
[["chr1"]]
integer(0)
[["chr2"]]
integer(0)
[["chr3"]]
integer(0)
[["chr4"]]
integer(0)
```r
> nread.tab <- table(viewSums(islands) / 200)
> depth.tab <- table(viewMaxs(islands))
```
> depth.tab[,1:10]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr10</td>
<td>68149</td>
<td>14748</td>
<td>2386</td>
<td>547</td>
<td>256</td>
<td>180</td>
<td>150</td>
<td>129</td>
<td>120</td>
<td>101</td>
</tr>
<tr>
<td>chr11</td>
<td>71677</td>
<td>17945</td>
<td>3527</td>
<td>862</td>
<td>362</td>
<td>268</td>
<td>205</td>
<td>179</td>
<td>181</td>
<td>130</td>
</tr>
<tr>
<td>chr12</td>
<td>59181</td>
<td>12441</td>
<td>2078</td>
<td>482</td>
<td>191</td>
<td>131</td>
<td>131</td>
<td>108</td>
<td>95</td>
<td>77</td>
</tr>
<tr>
<td>chr13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chrX</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chrY</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chrM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr1_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr3_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr4_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr5_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr7_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr8_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr9_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr13_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr16_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chr17_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chrX_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chrY_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>chrUn_random</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Processing multiple lanes

Although data from one lane is often a natural analytical unit, we typically want to apply any procedure to all lanes. Here is a simple summary function that computes the frequency distribution of the number of reads.

```r
> islandReadSummary <- function(x)
+ {
+   g <- resize(x, 200)
+   s <- slice(coverage(g), lower = 1)
+   tab <- table(viewSums(s) / 200)
+   df <- DataFrame(tab)
+   colnames(df) <- c("chromosome", "nread", "count")
+   df$nread <- as.integer(df$nread)
+   df
+ }
```

Applying it to our test-case, we get

```r
> head(islandReadSummary(cstest$ctcf))
```

<table>
<thead>
<tr>
<th>chromosome</th>
<th>nread</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>chr2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>chr3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>chr4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>chr5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>chr6</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

We can now use it to summarize the full dataset, flattening the returned `DataFrameList` with the `stack` function.

```r
> nread.islands <- DataFrameList(lapply(cstest, islandReadSummary))
> nread.islands <- stack(nread.islands, "sample")
> nread.islands
```

<table>
<thead>
<tr>
<th>sample</th>
<th>chromosome</th>
<th>nread</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ctcf</td>
<td>chr1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ctcf</td>
<td>chr2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ctcf</td>
<td>chr3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ctcf</td>
<td>chr4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ctcf</td>
<td>chr5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>gfp</td>
<td>chr16_random</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>gfp</td>
<td>chr17_random</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>gfp</td>
<td>chrX_random</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>gfp</td>
<td>chrY_random</td>
<td>34</td>
<td>0</td>
</tr>
<tr>
<td>gfp</td>
<td>chrUn_random</td>
<td>34</td>
<td>0</td>
</tr>
</tbody>
</table>
```
> xyplot(log(count) ~ nread | sample, as.data.frame(nread.islands),
+     subset = (chromosome == "chr10" & nread <= 40),
+     layout = c(1, 2), pch = 16, type = c("p", "g"))
```
If reads were sampled randomly from the genome, then the null distribution number of reads per island would have a geometric distribution; that is,

\[ P(X = k) = p^{k-1}(1 - p) \]

In other words, \( \log P(X = k) \) is linear in \( k \). Although our samples are not random, the islands with just one or two reads may be representative of the null distribution.

```r
> xyplot(log(count) ~ nread | sample, as.data.frame(nread.islands),
+   subset = (chromosome == "chr10" & nread <= 40),
+   layout = c(1, 2), pch = 16, type = c("p", "g"),
+   panel = function(x, y, ...) {
+     panel.lmline(x[1:2], y[1:2], col = "black")
+     panel.xyplot(x, y, ...)
+   })
```
We can create a similar plot of the distribution of depths.

```r
> islandDepthSummary <- function(x) 
+ { 
+   g <- resize(x, 200) 
+   s <- slice(coverage(g), lower = 1) 
+   tab <- table(viewMaxs(s) / 200) 
+   df <- DataFrame(tab) 
+   colnames(df) <- c("chromosome", "depth", "count") 
+   df$depth <- as.integer(df$depth) 
+   df 
+ } 
> depth.islands <- DataFrameList(lapply(cstest, islandDepthSummary)) 
> depth.islands <- stack(depth.islands, "sample") 
> xyplot(log(count) ~ depth | sample, as.data.frame(depth.islands), 
+   subset = (chromosome == "chr10" & depth <= 20), 
+   layout = c(1, 2), pch = 16, type = c("p", "g"), 
+   panel = function(x, y, ...) { 
+     lambda <- 2 * exp(y[2]) / exp(y[1]) 
+     null.est <- function(xx) { 
+       xx * log(lambda) - lambda - lgamma(xx + 1) 
+     } 
+     log.N.hat <- null.est(1) - y[1] 
+     panel.lines(1:10, -log.N.hat + null.est(1:10), col = "black") 
+     panel.xyplot(x, y, ...) 
+   }) 
> 
> ## depth.islands <- summarizeReads(cstest, summary.fun = islandDepthSummary) 
>
```
The above plot is very useful for detecting peaks, discussed in the next section. As a convenience, it can be created for the coverage over all chromosomes for a single sample by calling the `islandDepthPlot` function:

```r
> islandDepthPlot(cov.ctcf)
```
Peaks

To obtain a set of putative binding sites, i.e., peaks, we need to find those regions that are significantly above the noise level. Using the same Poisson-based approach for estimating the noise distribution as in the plot above, the `peakCutoff` function returns a cutoff value for a specific FDR:

```r
> peakCutoff(cov.ctcf, fdr = 0.0001)
[1] 6.959837
```

Considering the above calculation of 7 at an FDR of 0.0001, and looking at the above plot, we might choose 8 as a conservative peak cutoff:

```r
> peaks.ctcf <- slice(cov.ctcf, lower = 8)
> peaks.ctcf
```

RleViewsList of length 35
$chr1
Views on a 197195432-length Rle subject
views: NONE

$chr2
Views on a 181748087-length Rle subject
views: NONE

$chr3
Views on a 159599783-length Rle subject
views: NONE

...
<32 more elements>

To summarize the peaks for exploratory analysis, we call the `peakSummary` function:

```r
> peaks <- peakSummary(peaks.ctcf)
```

The result is a `GRanges` object with two columns: the view maxs and the view sums. Beyond that, this object is often useful as a scaffold for adding additional statistics.

It is meaningful to ask about the contribution of each strand to each peak, as the sequenced region of pull-down fragments would be on opposite sides of a binding site depending on which strand it matched. We can compute strand-specific coverage, and look at the individual coverages under the combined peaks as follows:

```r
> peak.depths <- viewMaxs(peaks.ctcf)
> cov.pos <- coverage(ctcf.ext[strand(ctcf.ext) == "+"])  
> cov.neg <- coverage(ctcf.ext[strand(ctcf.ext) == "-"])
> peaks.pos <- Views(cov.pos, ranges(peaks.ctcf))
> peaks.neg <- Views(cov.neg, ranges(peaks.ctcf))
> wpeaks <- tail(order(peak.depths$chr10), 4)
> wpeaks
```
[1] 971 989 1079 922

> 

Below, we plot the four highest peaks on chromosome 10.
\texttt{coverageplot(peaks.pos$chr10[wpeaks[1]], peaks.neg$chr10[wpeaks[1]])}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig1}
\caption{Position vs. Coverage for Chr10 at positions 79388100 to 79388400.}
\end{figure}

\texttt{coverageplot(peaks.pos$chr10[wpeaks[2]], peaks.neg$chr10[wpeaks[2]])}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig2}
\caption{Position vs. Coverage for Chr10 at positions 79627800 to 79628100.}
\end{figure}
> coverageplot(peaks.pos$chr10[wpeaks[3]], peaks.neg$chr10[wpeaks[3]])

> coverageplot(peaks.pos$chr10[wpeaks[4]], peaks.neg$chr10[wpeaks[4]])
Differential peaks

One common question is: which peaks are different in two samples? One simple strategy is the following: combine the two peak sets, and compare the two samples by calculating summary statistics for the combined peaks on top of each coverage vector.

```
> ## find peaks for GFP control
> cov.gfp <- coverage(resize(cstest$gfp, 200))
> peaks.gfp <- slice(cov.gfp, lower = 8)
> peakSummary <- diffPeakSummary(peaks.gfp, peaks.ctcf)
> xyplot(asinh(sums2) ~ asinh(sums1) | seqnames,
+       data = as.data.frame(peakSummary),
+       panel = function(x, y, ...) {
+         panel.xyplot(x, y, ...)
+         panel.abline(median(y - x), 1)
+       },
+       type = c("p", "g"), alpha = 0.5, aspect = "iso")
```

We use a simple cutoff to flag peaks that are different.

```
> mcols(peakSummary) <-
+   within(mcols(peakSummary),
+     {
+       diffs <- asinh(sums2) - asinh(sums1)
+       resids <- (diffs - median(diffs)) / mad(diffs)
+       up <- resids > 2
+       down <- resids < -2
+       change <- ifelse(up, "up", ifelse(down, "down", "flat"))
+     })
```
Placing peaks in genomic context

Locations of individual peaks may be of interest. Alternatively, a global summary might look at classifying the peaks of interest in the context of genomic features such as promoters, upstream regions, etc. The GenomicFeatures package facilitates obtaining gene annotations from different data sources. We could download the UCSC gene predictions for the mouse genome and generate a GRanges object with the transcript regions (from the first to last exon, contiguous) using makeTxDbFromUCSC; here we use a library containing a recent snapshot.

```r
> library(TxDb.Mmuseulus.UCSC.mm9.knownGene)
> gregions <- transcripts(TxDb.Mmuseulus.UCSC.mm9.knownGene)
> gregions

GRanges object with 55419 ranges and 2 metadata columns:

<table>
<thead>
<tr>
<th>seqnames</th>
<th>ranges</th>
<th>strand</th>
<th>tx_id</th>
<th>tx_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>4797974-4832908</td>
<td>+</td>
<td>1</td>
<td>uc007afg.1</td>
</tr>
<tr>
<td>chr1</td>
<td>4797974-4836816</td>
<td>+</td>
<td>2</td>
<td>uc007afh.1</td>
</tr>
<tr>
<td>chr1</td>
<td>4847775-4887990</td>
<td>+</td>
<td>3</td>
<td>uc007afi.2</td>
</tr>
<tr>
<td>chr1</td>
<td>4847775-4887990</td>
<td>+</td>
<td>4</td>
<td>uc011wht.1</td>
</tr>
<tr>
<td>chr1</td>
<td>4847775-4887990</td>
<td>+</td>
<td>5</td>
<td>uc011whu.1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>chrUn_random</td>
<td>2204169-2216886</td>
<td>-</td>
<td>55415</td>
<td>uc009sjw.2</td>
</tr>
<tr>
<td>chrUn_random</td>
<td>2674945-2678407</td>
<td>-</td>
<td>55416</td>
<td>uc009skb.2</td>
</tr>
<tr>
<td>chrUn_random</td>
<td>2889607-2891056</td>
<td>-</td>
<td>55417</td>
<td>uc009ske.1</td>
</tr>
<tr>
<td>chrUn_random</td>
<td>3830796-3837247</td>
<td>-</td>
<td>55418</td>
<td>uc009skg.1</td>
</tr>
<tr>
<td>chrUn_random</td>
<td>4677114-4677187</td>
<td>-</td>
<td>55419</td>
<td>uc009skh.1</td>
</tr>
</tbody>
</table>

seqinfo: 35 sequences (1 circular) from mm9 genome

We can now estimate the promoter for each transcript:

```r
> promoters <- flank(gregions, 1000, both = TRUE)
```

And count the peaks that fall into a promoter:

```r
> peakSummary$inPromoter <- peakSummary %over% promoters
> xtabs(~ inPromoter + change, peakSummary)

   change  inPromoter
          down flat
FALSE     21  5158
 TRUE      2   625
```

Or somewhere upstream or in a gene:

```r
> peakSummary$inUpstream <- peakSummary %over% flank(gregions, 20000)
> peakSummary$inGene <- peakSummary %over% gregions
```

```r
> sumtab <-
>  as.data.frame(rbind(total = xtabs(~ change, peakSummary),
```

18
Visualizing peaks in genomic context

While it is generally informative to calculate statistics incorporating the genomic context, eventually one wants a picture. The traditional genome browser view is an effective method of visually integrating multiple annotations with experimental data along the genome.

Using the rtracklayer package, we can upload our coverage and peaks for both samples to the UCSC Genome Browser:

```r
library(rtracklayer)
session <- browserSession()
genome(session) <- "mm9"
session$gfpCov <- cov.gfp
session$gfpPeaks <- peaks.gfp
session$ctcfCov <- cov.ctcf
session$ctcfPeaks <- peaks.ctcf
```

Once the tracks are uploaded, we can choose a region to view, such as the tallest peak on chr10 in the CTCF data:

```r
peak.ord <- order(unlist(peak.depths), decreasing=TRUE)
peak.sort <- as(peaks.ctcf, "GRanges")[peak.ord]
view <- browserView(session, peak.sort[1], full = c("gfpCov", "ctcfCov"))
```

We coerce to GRanges so that we can sort the ranges across chromosomes. By passing the full parameter to browserView we instruct UCSC to display the coverage tracks as a bar chart. Next, we might programmatically display a view for the top 5 tallest peaks:

```r
views <- browserView(session, head(peak.sort, 5), full = c("gfpCov", "ctcfCov"))
```

Version information

```r
sessionInfo()
```

R version 3.6.0 (2019-04-26)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 18.04.2 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.9-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.9-bioc/R/lib/libRlapack.so
locale: