seqPattern: an R package for visualisation of oligonucleotide sequence patterns and motif occurrences

Vanja Haberle *

February 2, 2015

Contents

1 Introduction **2**

2 Getting started **2**

3 Example data **2**
3.1 Zebrafish promoter sequences 3
3.2 Zebrafish promoter coordinates and associated features 3

4 Visualising oligonucleotide and consensus sequence densities **4**
4.1 Preparing input sequences 4
4.2 Plotting dinucleotide density maps 5
4.3 Plotting consensus sequence density map 9

5 Visualising motif occurrences **10**
5.1 Plotting density of motif occurrences 10
5.2 Plotting motif scanning scores 11

6 Getting occurrence of sequence patterns and motifs without visualisation **12**

7 Session Info **13**

*vanja.haberle@gmail.com

1 Introduction

This document briefly describes how to use the package *seqPattern*. *seqPattern* is a Bioconductor-compliant R package designed to visualise sequence patterns across a large set of sequences (*e.g.* promoters, enhancers, ChIPseq peaks, *etc.*) centred at a common reference position (*e.g.* TSS, peak position) and ordered by some feature. The visualisations includes plotting the density of occurrences of di-, tri-, and in general any oligo-nucleotides, consensus sequences specified using IUPAC nucleotide ambiguity codes and motifs specified by a position weight matrix (PWM). Such visualisations are useful in discovering sequence patterns positionally constrained with respect to the reference point and their correlation with the specified feature [1].
Here is a list of functionalities provided in this package:

- Obtaining positions of occurrences of oligonucleotides or consensus sequences in an ordered set of sequences centred at a common reference point in a matrix-like representation.
- Visualising density of oligonucleotide or consensus sequence occurrences in an ordered set of sequences centred at a common reference point. Multiple oligonucleotides can be analysed simultaneously and their density plotted on the same colour scale allowing visual comparison of enrichments/depletions of various oligonucleotides.
- Obtaining positions of occurrences of motifs defined by a PWM in an ordered set of sequences centred at a common reference point in a matrix-like representation. A custom threshold can be applied to report only motif matches with score above specified percentage of the maximal PWM score.
- Visualising density of motif occurrences in an ordered set of sequences centred at a common reference point. Only motif matches with score above specified percentage of the maximal PWM score are visualised.
- Visualising motif scanning scores for an ordered set of sequences centred at a common reference point in the form of a heatmap.

2 Getting started

To load the *seqPattern* package into your R envirnoment type:

```
> library(seqPattern)
```

3 Example data

The package contains two data sets provided for illustrating the functionality of the package.

3.1 Zebrafish promoter sequences

The first dataset is `zebrafishPromoters` and can be loaded by typing:

```
> data(zebrafishPromoters)
> zebrafishPromoters
```

```
  A DNAStringSet instance of length 1000
       width seq
   [1]  1000 ATCACTGGGTGACAAGCTGTTATAACACGCTG...ATGGGTAATGTAAAAAATAATATGAAAAATGC
   [2]  1000 TGACAGAAATGTGGATGATGTGTGGATAATTG...TCCAGACAAGTAAGAGACCACCCCCTCAGAAA
   [3]  1000 GTAAATCTAGTTTTGGCTGTTCTCAGGGGTGT...TGAACAATAGAGCATATGAACTGAAAGATTTT
   [4]  1000 TTGTCACTATACACTGCCATTATATGAATCAC...TGTTACTTTTAGCCAGTGTCTGAGTAAGTTTA
   [5]  1000 CACACTTTATATAATGAACAAATAAATATATT...CGTTAGCATTAATGCTAGCTTATTTTCGGGGC
   ...   ... ...
 [996]  1000 AACTGATTGTATTTTAATGACATTACAACATC...AAAGTTTAAAGCCTTGAGGCTCTAAGGCCTTT
 [997]  1000 CGCAATGCTCAGTAAACTCTCTGAAACAGACA...TTAACTGTTTTAAACCTCAGAAGGAGTTTTCT
 [998]  1000 TTGAAAATAATAGGGGTGAATGTATGACATTT...TCAGTCTGATAGATGACGTCAGTGCTCTTCTT
 [999]  1000 GATGTTGTTTTTTGGCTCAAGAACTGAAGTAT...ACTGTATGCAATATTTAATGTGATGTATTTAC
[1000]  1000 TACACACTCTTGCACAAACTCGCATACACTAC...TGACCAGCCTGGTTTAAGCTGGGCTCCCAGCC
```

```
> head(zebrafishPromoters@elementMetadata)
```

```
DataFrame with 6 rows and 2 columns
        tpm interquantileWidth
  <numeric>          <numeric>
1  21.64224                 58
2 166.72616                 59
3  34.23795                 29
4  33.93793                191
5  34.17674                 52
6  76.30356                 10
```

It is a `DNAStringSet` object that contains sequence of 1000 randomly selected promoters active in zebrafish (*Danio rerio*) embryos at 24 hours past fertilisation (hpf). The data is taken from Nepal *et al.* [2], and represents regions flanking 400 bp upstream and 600 bp downstream of the dominant TSS detected by Cap analysis of gene expression (CAGE). In addition to genomic sequence, the object contains metadata providing CAGE tag per million values and interquantile width for each promoter. This small example dataset is intended to be used as input for running examples from *seqPattern* package help pages.

3.2 Zebrafish promoter coordinates and associated features

The second dataset is `zebrafishPromoters24h`, which can be loaded by typing:

```
> data(zebrafishPromoters24h)
> head(zebrafishPromoters24h)
```

```
   chr  start    end strand dominantTSS      tpm interquantileWidth
1 chr1  53498  53920      +       53734 58.10913                 31
2 chr1 101877 102026      +      101934 43.38573                 14
3 chr1 134648 134774      +      134725 80.04638                 34
4 chr1 293528 293621      +      293616 54.59192                 48
5 chr1 368940 369095      +      369041 55.50777                 27
6 chr1 388651 388827      +      388696 10.14479                 13
```

This is a `data.frame` object that contains genomic coordinates of all (12078 in total) promoters active in zebrafish *Danio rerio* embryos at 24 hpf [2]. For each promoter additional information is provided, including position of the dominant (most frequently used) TSS position, number of CAGE tags per million supporting that promoter and the interquantile width of the promoter (width of the central region containing >= 80% of the CAGE tags). All examples in this vignette use this dataset to demonstrate how to use various functions provided in the package and illustrate the resulting visualisation.

4 Visualising oligonucleotide and consensus sequence densities

In this part of the tutorial we will be using data from zebrafish (*Danio rerio*) that was mapped to the danRer7 assembly of the genome. Therefore, the corresponding genome package *BSgenome.Drerio.UCSC.danRer7* has to be installed and available to load by typing:

```
> library(BSgenome.Drerio.UCSC.danRer7)
```

4.1 Preparing input sequences

As input we will use a full set of zebrafish promoters active in 24 hpf embryos that were precisely mapped using CAGE [2]. To load the zebrafish promoters data type:

```
> data(zebrafishPromoters24h)
> nrow(zebrafishPromoters24h)
```

```
[1] 12078
```

```
> head(zebrafishPromoters24h)
```

```
   chr  start    end strand dominantTSS      tpm interquantileWidth
1 chr1  53498  53920      +       53734 58.10913                 31
```

```
2 chr1 101877 102026      +      101934 43.38573                  14
3 chr1 134648 134774      +      134725 80.04638                  34
4 chr1 293528 293621      +      293616 54.59192                  48
5 chr1 368940 369095      +      369041 55.50777                  27
6 chr1 388651 388827      +      388696 10.14479                  13
```

The loaded `data.frame` contains genomic coordinates, position of the dominant (most frequently used) TSS position, number of CAGE tags per million and the interquantile width (width of the central region containing $>=$ 80% of the CAGE tags) for each promoter.

Next, we need to obtain the genomic sequence of the promoter region for which the oligonucleotide density will be visualised, for instance the region flanking 400 bp upstream and 800 bp downstream of the dominant TSS. Thus, in this case the dominant TSS will be the reference point to which all promoter sequences will be aligned. We also want to keep the information about promoter interquantile width, since this feature will be used to order the promoters in the density map.

```
> zebrafishPromotersTSS <- GRanges(seqnames = zebrafishPromoters24h$chr,
+             ranges = IRanges(start = zebrafishPromoters24h$dominantTSS,
+             end = zebrafishPromoters24h$dominantTSS),
+             strand = zebrafishPromoters24h$strand,
+             interquantileWidth = zebrafishPromoters24h$interquantileWidth,
+             seqlengths = seqlengths(Drerio))
> zebrafishPromotersTSSflank <- promoters(zebrafishPromotersTSS, upstream = 400,
+                                         downstream = 800)
> zebrafishPromotersTSSflankSeq <- getSeq(Drerio, zebrafishPromotersTSSflank)
```

Note that all regions need to have the same width, and in cases when flanking regions fall outside of chromosome boundaries they need to be removed prior to plotting the oligonucleotide density map.

4.2 Plotting dinucleotide density maps

Once a `DNAStringSet` object is obtained, it can be used to plot the density of oligonucleotides of interest. In the following example, we will plot the density of TA, AA, GC and CG dinucleotides for the obtained set of sequences sorted by the promoter interquantile width (Figure 1):

```
> plotPatternDensityMap(regionsSeq = zebrafishPromotersTSSflankSeq,
+             patterns = c("TA", "AA", "GC", "CG"),
+             seqOrder = order(zebrafishPromotersTSSflank$interquantileWidth),
+             flankUp = 400, flankDown = 800)
```

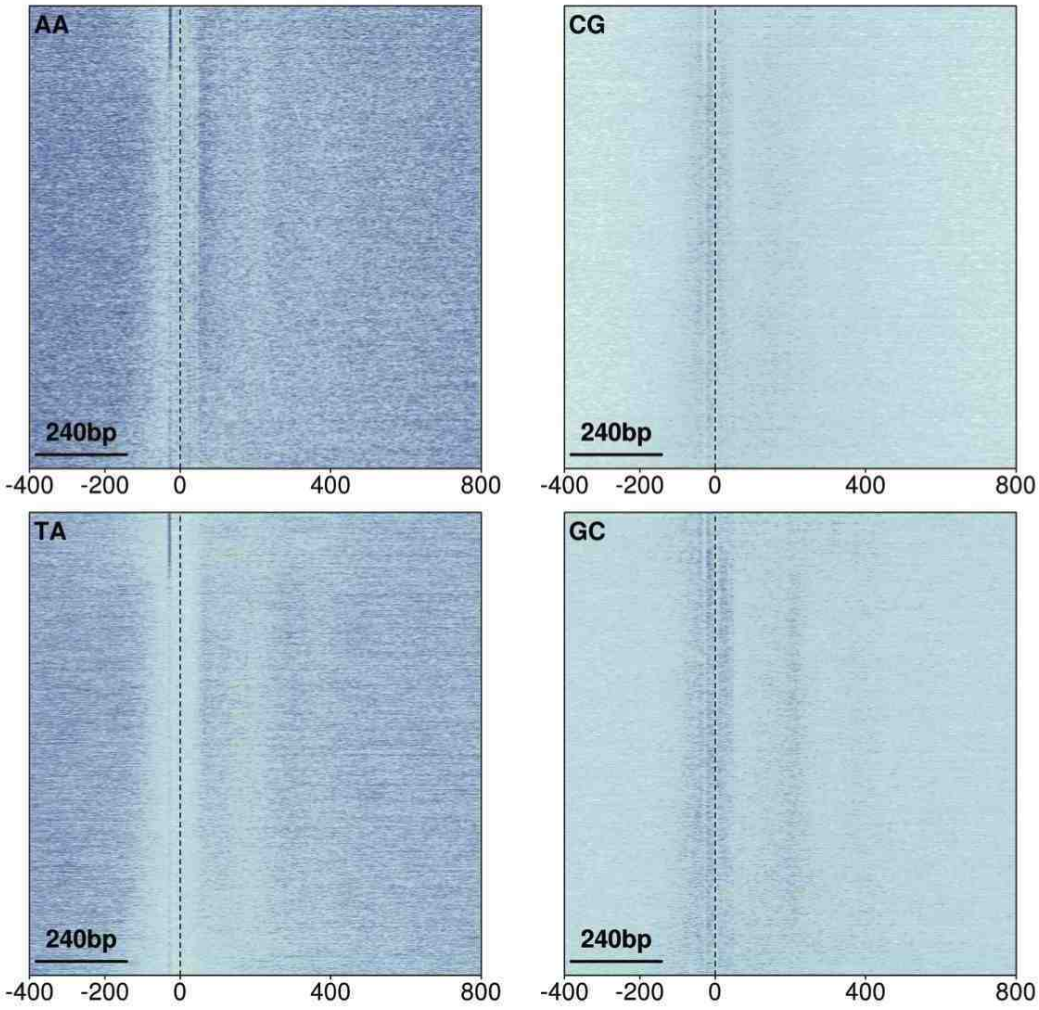

Figure 1: Density of TA, AA, GC and CG dinucleotides in regions flanking dominant TSS in zebrafish 24h embryo promoters ordered by promoter width

The information about the number of base pairs upstream and downstream of the reference point (`flankUp` and `flankDown`) needs to be specified in cases where these are asymmetric. If not specified, it is assumed that the reference point is located in the middle of the provided sequences (*i.e.* half of the total length in bp). There is also a number of graphical parameters that can be adjusted, such as width and hight of the plot, axis labels, scale bars, *etc.*, and they are explained in detail in the help page of the `plotPatternDensityMap` function. The two main parameters that define how the density of the pattern will be calculated and plotted are `nBin` and `bandWidth`. The `nBin` parameter specifies the number of bins in which the plot will be divided across x (horizontal, corresponding to the sequence length) and y axes (vertical, corresponding to the number of sequences). The default value is to calculate

the density at each position in every sequence, *i.e.* the number of bins in the horizontal direction is set to the sequence length and in the vertical direction to the total number of sequences. The `bandWidth` parameter specifies the standard deviation of the bivariate Gaussian kernel along the x (*i.e.* number of nucleotides) and y (*i.e.* number of sequences) axis that is used to compute the density for each bin. The schematic in Figure 2 illustrates how the density of three different dinucleotides is calculated for a set of 10 sequences each 10 bp long, using a 2D Gaussian kernel with standard deviation of 5 in both directions. The calculations for larger sets of sequences and for any specified sequence pattern are done analogously. Note that *seqPattern* package supports parallel processing using multiple cores on Unix-like platforms, which significantly reduces the computational time when visualising density of multiple patterns. For instance, the above example that calculates the density of 4 dinucleotides can be run on 4 cores by setting the `useMulticore` and `nrCores` parameters:

```
> plotPatternDensityMap(regionsSeq = zebrafishPromotersTSSflankSeq,
+             patterns = c("TA", "AA", "GC", "CG"),
+             seqOrder = order(zebrafishPromotersTSSflank$interquantileWidth),
+             flankUp = 400, flankDown = 800, useMulticore = TRUE, nrCores = 4)
```

Calling the `plotPatternDensityMap` function will create one .png file per specified pattern in the working directory. The resulting dinucleotide density plots reveal complex pattern of dinucleotide enrichments/depletions in zebrafish promoters (Figure 1). The density of all dinucleotides is plotted on the same colour scale, which allows direct comparison. For instance, it is clear that CG dinucleotides are generally abundant than other dinucleotides. The region immediately upstream and downstream of TSS is enriched for CG and GC dinucleotides and depleted for TA and AA dinucleotides. Within this region, there is narrow band of TA and AA enrichment ~30 bp upstream of the TSS visible only in sharp promoters (top). Region downstream of TSS is characterised by alternating bands of enrichments and depletions visible for all four dinucleotides and this pattern is more prominent in broad promoters (bottom).Thus, visualising sequence in such way reveals differences in underlying sequence composition and its relation to the reference point, as well as how this changes with the provided feature.

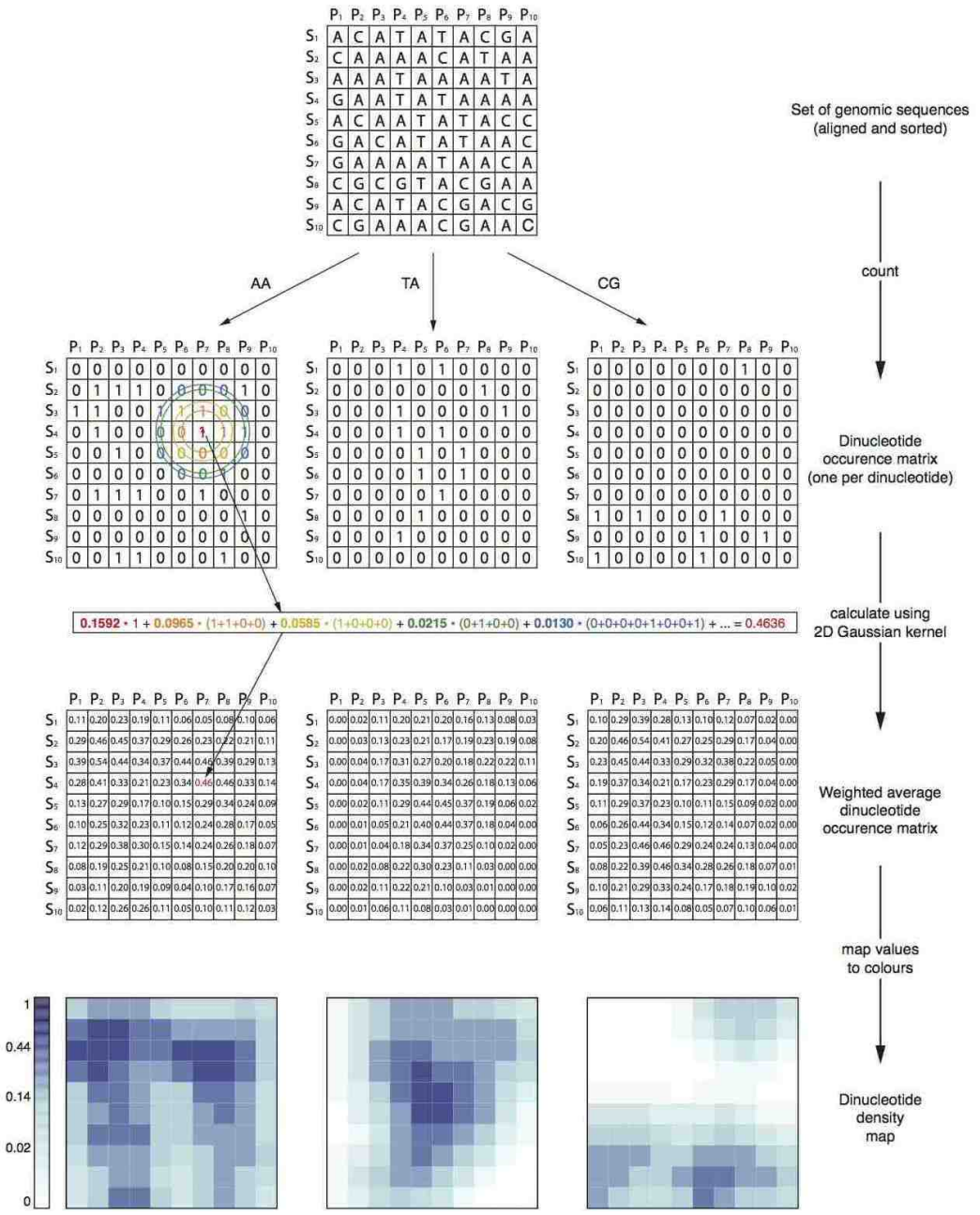

Figure 2: Schematics illustrating steps in pattern density calculation and visualisation. Genomic sequences (of the same length) are sorted and aligned into a matrix-like representation. Marking the presence of selected dinucleotide by 1 and the absence by 0 creates an occurrence matrix. Next, a weighted average is calculated at each position by placing a 2D Gaussian kernel at that position and assigning weights to surrounding positions. An example of calculating the value at position S4,P7 is shown. Surrounding positions are coloured on the basis of the weights assigned to them by the Gaussian kernel (bandwidth=5 in both dimensions, and covariance=0 between the two dimensions). Averaged values are mapped to different shades of blue to visualise the dinucleotide density.

In addition to plotting density map of individual dinucleotides, a "metadinucleotide" can be specified using IUPAC nucleotide ambiguity codes. For instance, for the same set of promoter regions we can plot the density of all WW and all SS dinucleotides in the same plot (Figure 3):

```
> plotPatternDensityMap(regionsSeq = zebrafishPromotersTSSflankSeq,
+             patterns = c("WW", "SS"),
+             seqOrder = order(zebrafishPromotersTSSflank$interquantileWidth),
+             flankUp = 400, flankDown = 800, labelCol = "white")
```

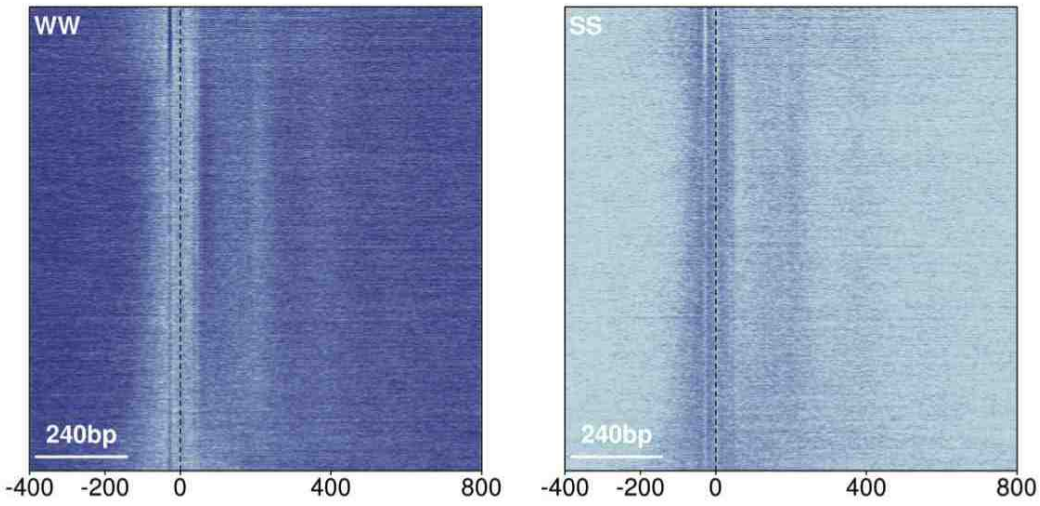

Figure 3: Density of WW and SS dinucleotides in regions flanking dominant TSS in zebrafish 24h embryo promoters ordered by promoter width

4.3 Plotting consensus sequence density map

Analogously to plotting dinucleotide density as described above, the `plotPatternDensityMap` function can be used to visualise the density of longer consensus sequences specified using IUPAC nucleotide ambiguity codes. For instance, one can use a consensus sequence for binding of a certain transcription factor to visualise density of these sites with respect to some reference point. Here we show an example of plotting density of the TATA-box consensus sequence (TATAWARA) across zebrafish promoters aligned to dominant TSS and sorted by promoter width:

```
> plotPatternDensityMap(regionsSeq = zebrafishPromotersTSSflankSeq,
+             patterns = c("TATAWAWR"),
+             seqOrder = order(zebrafishPromotersTSSflank$interquantileWidth),
+             flankUp = 400, flankDown = 800, nBin = c(1200, 3000),
+             bandWidth = c(2,6), addPatternLabel = FALSE)
```

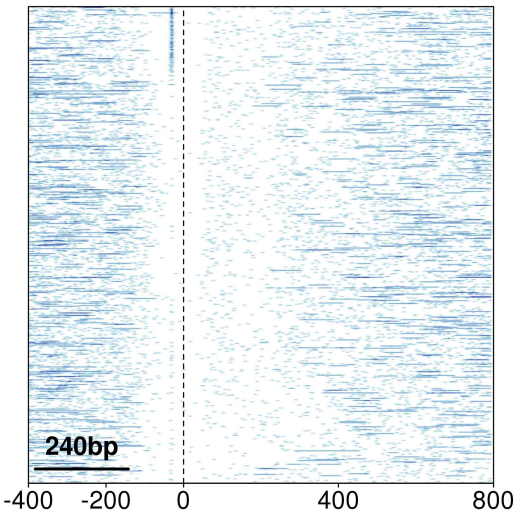

Figure 4: Density of the TATA-box consensus sequence (TATAWAWR) in regions flanking dominant TSS in zebrafish 24h embryo promoters ordered by promoter width

The resulting density plot (Figure 4) shows that the TATA-box consensus sequence is positioned very precisely at ~30 bp upstream of the TSS and that it is present only in very sharp promoters (top), but not in the broad promoters (bottom).

5 Visualising motif occurrences

5.1 Plotting density of motif occurrences

In addition to using consensus sequence, the binding motif of a certain transcription factor can be described by a position-weight matrix (PWM), which gives the probability of occurrence of each of the four nucleotides at a given position in the motif. More specifically, the values in the PWM are derived from the position-specific frequency matrix and represent log-ratio between nucleotide probabilities derived from observed frequency and expected background probability for the corresponding nucleotide [3]. An example of a PWM describing the binding motif for the TATA-box binding transcription factor (TBP) is provided in the package and can be loaded:

```
> data(TBPpwm)
> TBPpwm
```

```
       [,1]      [,2]      [,3]      [,4]       [,5]      [,6]       [,7]
A -2.588668  1.853661 -4.928518  1.861830  1.4605751  1.886064  1.1891247
C -1.076769 -8.928518 -3.256093 -8.928518 -8.9285184 -4.928518 -5.4690868
```

```
G -2.420724 -5.469087 -5.469087 -4.228079 -8.9285184 -2.270307 -1.1406158
T  1.665806 -1.469087  1.941075 -1.690114  0.3146556 -3.974322  0.3146556
        [,8]
A  0.6713945
C -1.1406158
G  0.6898671
T -1.5534789
```

The `plotMotifDensityMap` takes a PWM as an input, scans all sequences for the occurrence of the motif above the specified match threshold (*e.g.* 90%) and visualises the density of the motif (Figure 5, left):

```
> plotMotifDensityMap(regionsSeq = zebrafishPromotersTSSflankSeq,
+             motifPWM = TBPpwm, minScore = "90%",
+             seqOrder = order(zebrafishPromotersTSSflank$interquantileWidth),
+             flankUp = 400, flankDown = 800)
```

5.2 Plotting motif scanning scores

On the other hand, using the `plotMotifScanScores` function it is possible to visualise the PWM scanning scores along entire sequences in a form of a heatmap (Figure 5, right):

```
> plotMotifScanScores(regionsSeq = zebrafishPromotersTSSflankSeq,
+             motifPWM = TBPpwm,
+             seqOrder = order(zebrafishPromotersTSSflank$interquantileWidth),
+             flankUp = 400, flankDown = 800)
```

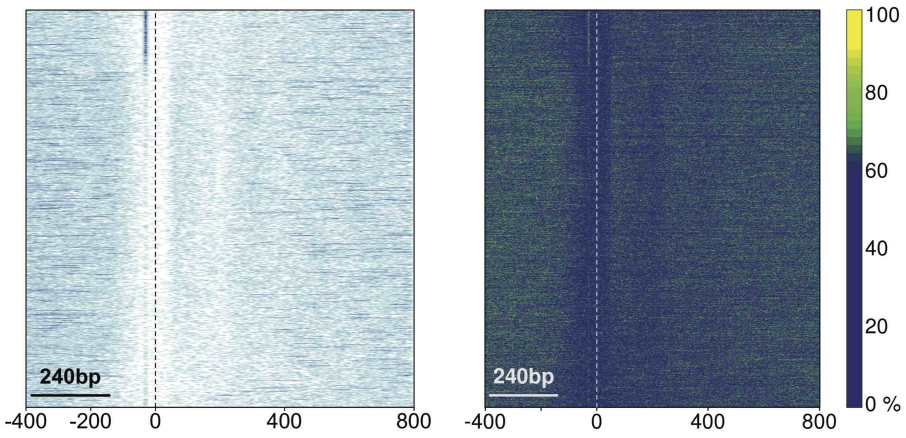

Figure 5: Density of TATA-box motif occurrence above 90% of PWM score (left) and heatmap of PWM scanning scores along all sequences (right) in regions flanking dominant TSS in zebrafish 24h embryo promoters ordered by promoter width

In addition to showing positioning and enrichment of strong motif occurrences with high match to the provided PWM, this form of visualisation can also reveal positional constraints for occurrences of motifs with varying strength. For instance weaker motifs might have different positional preference with respect to the reference point and might occur only in a subset of sequences correlating with some feature, which will be visible when the sequences are sorted according to that feature.

6 Getting occurrence of sequence patterns and motifs without visualisation

The above described functions find the occurrence of specified sequence patterns or motifs in an ordered set of sequences, calculate their density and visualise the result as a density map. The *seqPattern* package also provides functions for finding only the occurrence of patterns or motifs without calculating the density and visualising it in a plot. These are `getPatternOccurrenceList` and `motifScanHits` for finding occurrence of patterns/consensus sequences and motifs specified by a PWM, respectively.

```
> motifOccurrence <- motifScanHits(regionsSeq =
+             zebrafishPromotersTSSflankSeq[1:50],
+             motifPWM = TBPpwm, minScore = "90%", seqOrder =
+             order(zebrafishPromotersTSSflank$interquantileWidth[1:50]))
> head(motifOccurrence)
```

```
  sequence position value
1        1      269     1
2        1      370     1
3        1      506     1
4        1      686     1
5        1     1011     1
6        1     1015     1
```

The occurrences are returned as coordinates in a matrix-like representation as follows: Input sequences of the same length are sorted according to the index in `seqOrder` argument creating an `n x m` matrix where `n` is the number of sequences and `m` is the length of the sequence. For each pattern match the coordinates within such matrix are reported, *i.e.* the ordinal number of the sequence within the ordered set of sequences (`sequence` column) and the start position of the pattern within that sequence (`position` column) are returned in the resulting `data.frame`.

Similarly, the matrix of PWM scanning scores along all sequences can be obtained using `motifScanScores` function:

```
> scanScores <- motifScanScores(regionsSeq = zebrafishPromotersTSSflankSeq[1:50],
+              motifPWM = TBPpwm, seqOrder =
+              order(zebrafishPromotersTSSflank$interquantileWidth[1:50]))
> dim(scanScores)
```

```
[1]   50 1193
```

```
> scanScores[1:6,1:6]
```

```
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
[1,] 30.33369 57.86230 30.33369 57.89343 29.49237 55.09883
[2,] 69.22622 49.77369 62.22976 79.66015 66.11636 77.34020
[3,] 59.39872 58.38168 67.79532 45.71215 52.94822 58.83006
[4,] 42.98148 79.86212 63.36395 64.24627 73.58978 83.06478
[5,] 44.02943 41.17408 68.54642 50.96918 74.85985 49.42530
[6,] 54.02857 51.50187 58.33996 54.29635 61.60616 66.96074
```

By default, the values corresponding to the percentage of the maximal possible PWM score are returned.

7 Session Info

```
> sessionInfo()
```

```
R version 3.2.0 alpha (2015-03-20 r68043)
Platform: x86_64-unknown-linux-gnu (64-bit)
Running under: Ubuntu 14.04.2 LTS

locale:
 [1] LC_CTYPE=C                 LC_NUMERIC=C
 [3] LC_TIME=C                  LC_COLLATE=C
 [5] LC_MONETARY=C              LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
 [1] BSgenome.Drerio.UCSC.danRer7_1.4.0 BSgenome_1.35.19
 [3] rtracklayer_1.27.9                 Biostrings_2.35.11
```

```
 [5] XVector_0.7.4                   GenomicRanges_1.19.47
 [7] GenomeInfoDb_1.3.15             IRanges_2.1.43
 [9] S4Vectors_0.5.22                BiocGenerics_0.13.9
[11] seqPattern_0.99.3

loaded via a namespace (and not attached):
 [1] XML_3.98-1.1             Rsamtools_1.19.47        GenomicAlignments_1.3.32
 [4] bitops_1.0-6             futile.options_1.0.0     KernSmooth_2.23-14
 [7] zlibbioc_1.13.3          limma_3.23.11            futile.logger_1.4
[10] BiocStyle_1.5.3          lambda.r_1.1.7           BiocParallel_1.1.21
[13] tools_3.2.0              RCurl_1.95-4.5           plotrix_3.5-11
[16] marray_1.45.0
```

References

[1] Vanja Haberle, Nan Li, Yavor Hadzhiev, Charles Plessy, Christopher Previti, Chirag Nepal, Jochen Gehrig, Xianjun Dong, Altuna Akalin, Ana Maria Suzuki, Wilfred F J van IJcken, Olivier Armant, Marco Ferg, Uwe Strähle, Piero Carninci, Ferenc Müller, and Boris Lenhard. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, 507(7492):381–385, 2014.

[2] Chirag Nepal, Yavor Hadzhiev, Christopher Previti, Vanja Haberle, Nan Li, Hazuki Takahashi, Ana Maria S. Suzuki, Ying Sheng, Rehab F. Abdelhamid, Santosh Anand, Jochen Gehrig, Altuna Akalin, Christel E.M. Kockx, Antoine A.J. van der Sloot, Wilfred F.J. van IJcken, Olivier Armant, Sepand Rastegar, Craig Watson, Uwe Strähle, Elia Stupka, Piero Carninci, Boris Lenhard, and Ferenc Müller. Dynamic regulation of coding and non-coding transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Research*, 23(11):1938–1950, 2013.

[3] Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 2004.