An introduction to the nuCpos package

Hiroaki Kato*, Takeshi Urano

January 19, 2023

1 About nuCpos

nuCpos, a derivative of *NuPoP*, is an R package for predicting ***nuc***leosome ***pos***itions. In *nuCpos*, a duration hidden Markov model is trained with a ***C***hemical map of nucleosomes either from budding yeast *Saccharomyces cerevisiae* (Brogaard et al. (2012)), fission yeast *Schizosaccharomyces pombe* (Moyle-Heyrman et al. (2012)), or embryonic stem cells of house mouse *Mus musculus* (Voong et al. (2016)). *nuCpos* outputs the Viterbi (most probable) path of nucleosome-linker states, predicted nucleosome occupancy scores and histone binding affinity (HBA) scores as *NuPoP* does. *nuCpos* can also calculate local and whole nucleosomal HBA scores for a given 147-bp sequence.

The parental package *NuPoP*, licensed under GPL-2, was developed by Ji-Ping Wang and Liqun Xi. Please refer to Xi et al. (2010) and Wang et al. (2008) for technical details of *NuPoP*. Their excellent codes were adapted in *nuCpos* to demonstrate the usefulness of chemical maps in prediction. In dHMM-based predicton, users of *nuCpos* can choose whether HBA scores will be smoothed or not in the output. Note that when *nuCpos* was released, *NuPoP* only used an MNase-seq-based map of budding yeast nucleosomes to train a duration hidden Markov model. However, as *NuPoP* now provides chemical map-based prediction, users are encouraged to use *NuPoP* functions to conduct dHMM-based prediction in their original way.

2 nuCpos functions

nuCpos has three functions: `predNuCpos`, `HBA`, and `localHBA`. The `predNuCpos` function provides dHMM-based prediction of nucleosome occupancy. This function is built to demonstrate the usefulness of chemical maps in prediction. Users can obtain unsmoothed HBA scores with this function. Note: *NuPoP* now provides chemical map-based prediction.

The functions `HBA` and `localHBA` receive a sequence of 147-bp DNA and calculate whole nucleosomal and local HBA scores. These functions invoke core Fortran codes for HBA calculation that were adapted from the excellent dHMM code of *NuPoP*.

nuCpos requires the *Biostrings* package, especially when DNA sequences are given as DNAString objects to the functions `HBA`, and `localHBA`. These functions can also receive DNA sequences as simple character string objects without loading the *Biostrings* package. Note: *nuCpos* requires the *NuPoP* package to perform some example runs.

Load the *nuCpos* package as follows:

```
> library(nuCpos)
```

*hkato@med.shimane-u.ac.jp

3 Performing predictions with predNuCpos

The `predNuCpos` function acts like the `predNuPoP` function of *NuPoP*. When the *ActLikePredNuPoP* argument is set as TRUE, `predNuCpos` reads a DNA sequence file in FASTA format and invokes a Fortran subroutine to perform predictions. The prediction results will be saved in the working directory. *TRP1ARS1x1.fasta*, the DNA sequence of *TRP1ARS1* circular minichromosome (1,465 bp) (Fuse et al. (2017)), in `extdata` can be used for an example run. Call the `predNuCpos` function as follows:

```
> predNuCpos(file = system.file("extdata", "TRP1ARS1x1.fasta",
+     package = "nuCpos"), species = "sc", smoothHBA = FALSE,
+     ActLikePredNuPoP = TRUE)
```

The argument *file* is the path to the fasta file. The argument *species* can be specified as follows: mm = *M. musculus*; sc = *S. cerevisiae*; sp = *S. pombe*. Re-scaling of the nucleosome and linker models for the prediction of other species' nucleosomes are not supported. *nuCpos* uses 4th order Markov chain models for the prediction.

The name of the output file will be like *TRP1ARS1x1.fasta_Prediction4.txt*. As in the output file produced by the parental *NuPoP* package, it will contain five columns:

1. `Position`: position in the input DNA sequence.
2. `P-start`: probability that a nucleosome starts at.
3. `Occup`: nucleosome occupancy score.
4. `N/L`: Viterbi path (1 and 0 for the nucleosome and linker states, respectively).
5. `Affinity`: histone binding affinity score.

To import the output into R, the `readNuPoP` function of *NuPoP* can be used:

```
> library(NuPoP)
> results.TRP1ARS1x1 <- readNuPoP("TRP1ARS1x1.fasta_Prediction4.txt",
+     startPos = 1, endPos = 1465)
> results.TRP1ARS1x1[1:5,]
```

```
  Position P.start Occup N/L Affinity
1        1   0.000 0.000   0       NA
2        2   0.000 0.000   0       NA
3        3   0.000 0.000   0       NA
4        4   0.001 0.001   0       NA
5        5   0.005 0.006   0       NA
```

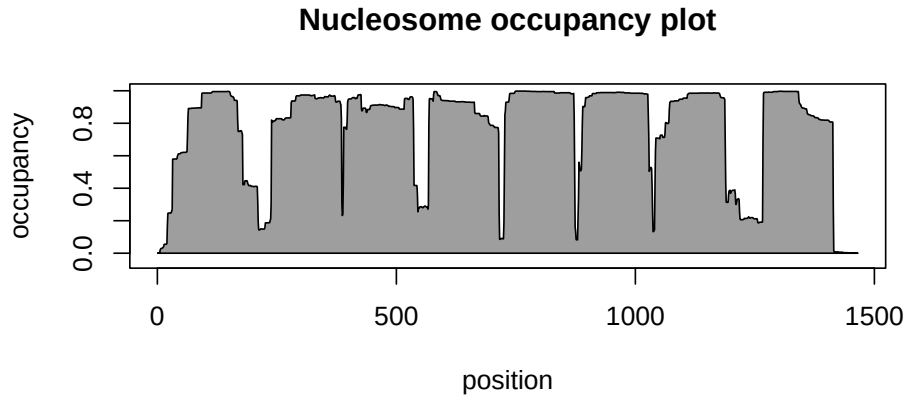
The arguments *startPos* and *endPos* are used to import a part of the prediction results. In this example, the prediction results for the whole tested sequence is imported. First and last 73-bp regions do not have HBA scores (`Affinity`) as they cannot be calculated. The HBA scores start from the 74th position:

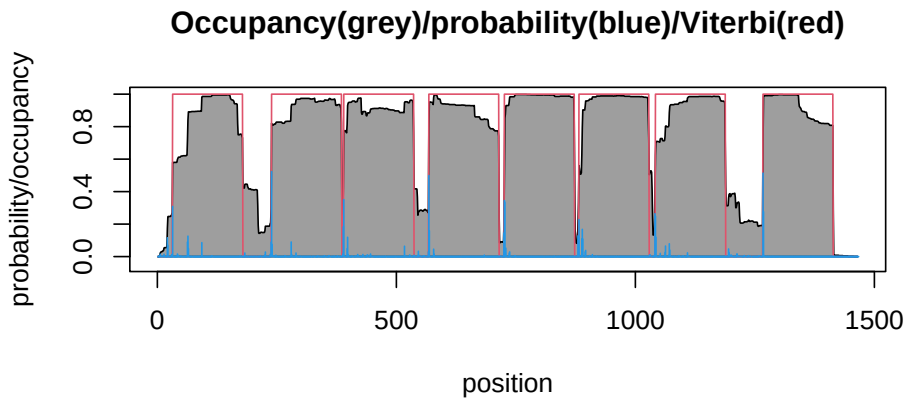
```
> results.TRP1ARS1x1[72:76,]
```

```
   Position P.start Occup N/L Affinity
72       72   0.001 0.893   1       NA
73       73   0.000 0.893   1       NA
74       74   0.000 0.893   1    0.346
75       75   0.000 0.893   1   -4.435
76       76   0.000 0.893   1   -3.429
```

For visualization of the prediction results, the `plotNuPoP` function of *NuPoP* can be used. This function draws two plots in the graphical window. The top one shows predicted nucleosome occupancy. In the bottom one, probability of a nucleosome to start at the given position (blue vertical lines) and the Viterbi path (red lines) are shown as well as the nucleosome occupancy (gray).

```
> plotNuPoP(results.TRP1ARS1x1)
```





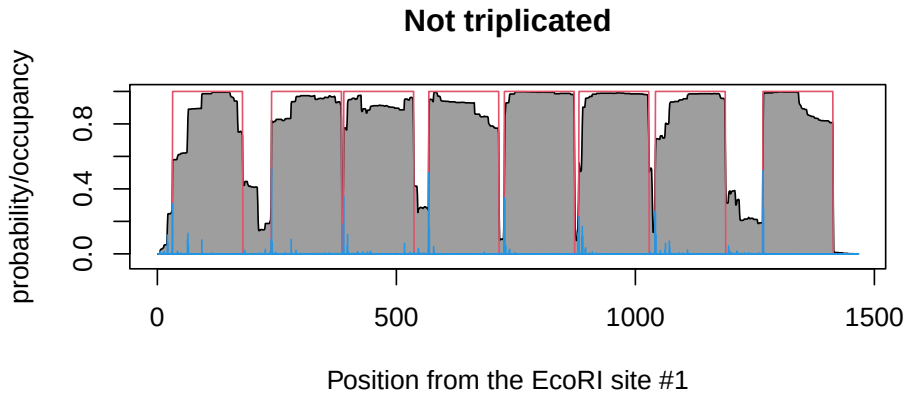
For prediction of nucleosome positioning in short circular DNA, one can use a triplicated sequence for prediction and read only the central copy for the evaluation. By triplicating the DNA, inaccurate prediction near the DNA ends, which are joined to each other in the circular form, can be avoided.

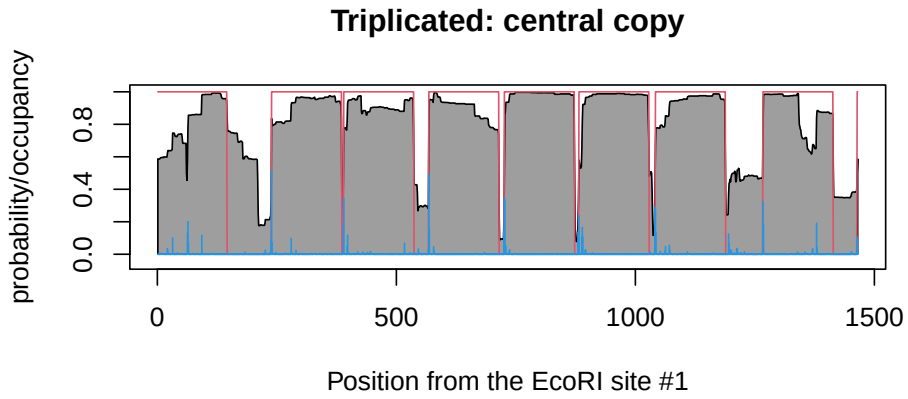
```
> predNuCpos(file = system.file("extdata", "TRP1ARS1x3.fasta",
+     package = "nuCpos"), species = "sc", smoothHBA = FALSE,
```

```
+     ActLikePredNuPoP = TRUE)
> results.TRP1ARS1 <- readNuPoP("TRP1ARS1x3.fasta_Prediction4.txt",
+     startPos = 1466, endPos = 2930)
```

Here, `TRP1ARS1x3.fasta` in `extdata` is a triplicated sequence (4,395 bp) of the *TRP1ARS1* minichromosome (1,465 bp). The central part (from the coordinate 1,466 to 2,930) of the prediction results is read by `readNuPoP`. They are apparently different from the previous results near the terminal regions.

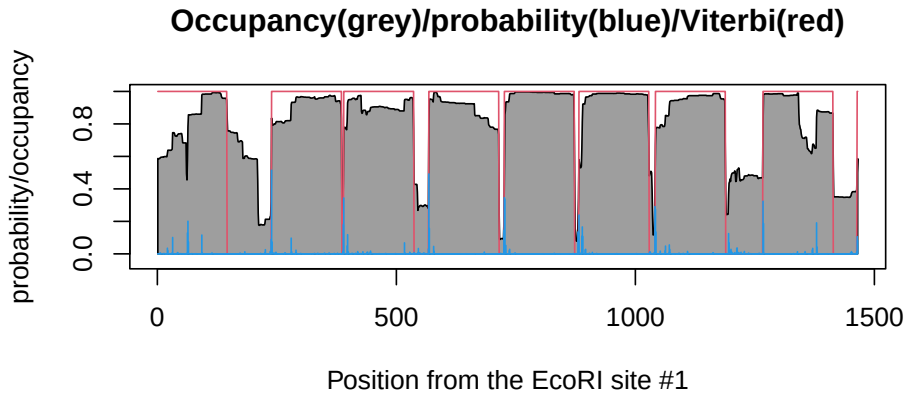
```
> par(mfrow = c(2, 1))
> plot(x = 1:1465, y = results.TRP1ARS1x1[,3], type = "n",
+     ylim = c(-0.05, 1), xlab = "Position from the EcoRI site #1",
+     ylab = "probability/occupancy")
> title("Not triplicated")
> polygon(c(1, 1:1465, 1465), c(0, results.TRP1ARS1x1[,3], 0), col = 8)
> points(x = 1:1465, y = results.TRP1ARS1x1[,4], type = "l", col = 2)
> points(x = 1:1465, y = results.TRP1ARS1x1[, 2], type = "h", col = 4)
> plot(x = 1:1465, y = results.TRP1ARS1[,3], type = "n",
+     ylim = c(-0.05, 1), xlab = "Position from the EcoRI site #1",
+     ylab = "probability/occupancy")
> title("Triplicated: central copy")
> polygon(c(1, 1:1465, 1465), c(0, results.TRP1ARS1[,3], 0), col = 8)
> points(x = 1:1465, y = results.TRP1ARS1[,4], type = "l", col = 2)
> points(x = 1:1465, y = results.TRP1ARS1[, 2], type = "h", col = 4)
```

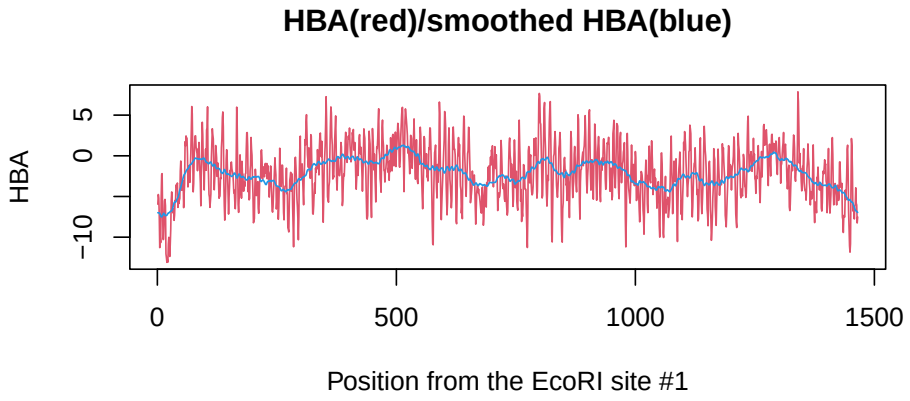




By specifying the argument *smoothHBA* as TRUE, HBA scores can be smoothed in a 55-bp window as being done by the `predNuPoP` function of *NuPoP*.

```
> predNuCpos(file = system.file("extdata", "TRP1ARS1x3.fasta",
+     package = "nuCpos"), species = "sc", smoothHBA = TRUE,
+     ActLikePredNuPoP = TRUE)
> results.TRP1ARS1.smooth <- readNuPoP("TRP1ARS1x3.fasta_Prediction4.txt",
+     startPos = 1466, endPos = 2930)
> par(mfrow = c(2, 1))
> plot(x = 1:1465, y = results.TRP1ARS1[,3], type = "n",
+     ylim = c(-0.05, 1), xlab = "Position from the EcoRI site #1",
+     ylab = "probability/occupancy")
> title("Occupancy(grey)/probability(blue)/Viterbi(red)")
> polygon(c(1, 1:1465, 1465), c(0, results.TRP1ARS1[,3], 0), col = 8)
> points(x = 1:1465, y = results.TRP1ARS1[,4], type = "l", col = 2)
> points(x = 1:1465, y = results.TRP1ARS1[, 2], type = "h", col = 4)
> plot(x = 1:1465, y = results.TRP1ARS1[,5], type = "n",
+     xlab = "Position from the EcoRI site #1",
+     ylab = "HBA", main = "HBA(red)/smoothed HBA(blue)")
> points(x = 1:1465, y = results.TRP1ARS1[,5], type = "l", col = 2)
> points(x = 1:1465, y = results.TRP1ARS1.smooth[,5], type = "l", col = 4)
```





As shown as a red line in the bottom one of the above plots, non-smoothed HBA scores in eukaryotic sequences exhibit about 10-bp periodicity. The dyads of predicted nucleosomes usually locate at the coordinates with high HBA scores. HBA scores in the output of `predNuCpos` can be standardized as being done by the `predNuPoP` function of *NuPoP* by specifying the argument *std* of `predNuCpos` as TRUE. The default setting for *std* is FALSE.

When the argument *ActLikePredNuPoP* is set as FALSE, which is the default setting, `predNuCpos` receives a character string or DNAString object as *inseq*. In this case, prediction results will be returned to the R environment, and no file will be generated in the working directory. The input sequence (*inseq*) must not contain characters other than A/C/G/T.

The results will contain five columns:

1. `pos`: position in the input DNA sequence.
2. `pstart`: probability that a nucleosome starts at.
3. `nucoccup`: nucleosome occupancy score.
4. `viterbi`: Viterbi path (1 and 0 for the nucleosome and linker states, respectively).
5. `affinity`: histone binding affinity score.

```
> TRP1ARS1 <- paste(scan(file =
+     system.file("extdata", "TRP1ARS1x1.fasta", package = "nuCpos"),
+     what = character(), skip = 1), sep = "", collapse = "")
> results.TRP1ARS1.internal <-
+     predNuCpos(inseq = TRP1ARS1, species = "sc", smoothHBA = FALSE,
+     ActLikePredNuPoP = FALSE)
> results.TRP1ARS1.internal[72:76,]

   pos       pstart  nucoccup viterbi   affinity
72  72 9.802925e-04 0.8926019       1         NA
73  73 3.182192e-04 0.8929201       1         NA
74  74 5.025299e-05 0.8929704       1  0.3456824
75  75 5.160730e-06 0.8929756       1 -4.4353083
76  76 1.397652e-06 0.8929769       1 -3.4287470
```

4 Histone binding affinity score calculation with HBA

HBA score can be calculated for a given 147-bp sequence with the `HBA` function. In the examples bellow, a character string object `inseq` and a DNAString object `INSEQ` with the same 147-bp DNA sequences are given to `HBA`. Note: the *Biostrings* package is required for the latter case.

```
> load(system.file("extdata", "inseq.RData", package = "nuCpos"))
> HBA(inseq = inseq, species = "sc")

      HBA
-2.460025

> for(i in 1:3) cat(substr(inseq, start = (i-1)*60+1,
+     stop = (i-1)*60+60), "\n")

ATCGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAA
ACGCACGTACGCGCTGTCCCCCGCGTTTTAACCGCCAAGGGGATTACTCCCTAGTCTCCA
GGCACGTGTCAGATATATACATCCGAT

> load(system.file("extdata", "INSEQ_DNAString.RData",
+     package = "nuCpos"))
> INSEQ

147-letter DNAString object
seq: ATCGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTC...TAGTCTCCAGGCACGTGTCAGATATATACATCCGAT

> HBA(inseq = INSEQ, species = "sc")

      HBA
-2.460025
```

The argument *inseq* is the character string object to be given. Alternatively, a DNAString object can be used here. The length of DNA must be 147 bp. The argument *species* can be specified as follows: mm = *M. musculus*; sc = *S. cerevisiae*; sp = *S. pombe*.

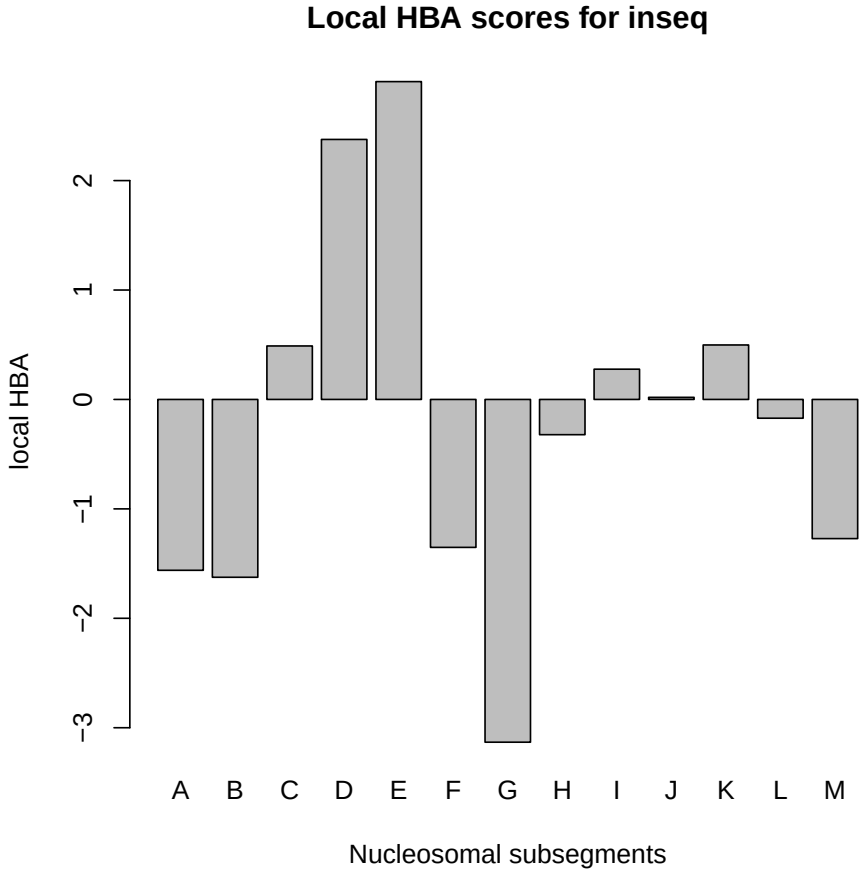
5 Local histone binding affinity score calculation with localHBA

Local HBA scores are defined as HBA scores for 13 overlapping subnucleosomal segments named A to M. They can be calculated for a given 147-bp sequence with the `localHBA` function. Like `HBA`, this function can receive either a character string object or a DNAString object. The segment G corresponds to the central 21 bp region, in which the dyad axis passes through the 11th base position. This means that the local HBA score for the G segment implies the relationship between DNA and histone proteins at around superhelical locations -0.5 and +0.5. The neighboring F segment, which is 20 bp in length, is for SHLs -1.5 and -0.5. The result of example run shown below suggests that subsequence of `inseq` around SHL -3.5 and -2.5 is suitable for nucleosome formation.

```
> localHBA(inseq = inseq, species = "sc")
```

```
     lHBA_A      lHBA_B      lHBA_C      lHBA_D      lHBA_E      lHBA_F
-1.56140949 -1.62502354  0.48885990  2.37615568  2.90458625 -1.35195919
     lHBA_G      lHBA_H      lHBA_I      lHBA_J      lHBA_K      lHBA_L
-3.13228907 -0.32208031  0.27650871  0.01922002  0.49787625 -0.17151500
     lHBA_M
-1.27186158
```

```
> barplot(localHBA(inseq = inseq, species = "sc"),
+     names.arg = LETTERS[1:13], xlab = "Nucleosomal subsegments",
+     ylab = "local HBA", main = "Local HBA scores for inseq")
```



6 Acknowledgements

We would like to thank Drs. Shimizu, Fuse and Ichikawa for sharing DNA sequences and *in vivo* data, and giving fruitful comments. We would like to thank Dr. Ji-Ping Wang and his colleagues for distributing NuPoP under the GPL-2 license. In this package, their excellent code for dHMM-based prediction was adapted for chemical map-based prediction to demonstrate the usefulness of chemical maps in prediction. As we noticed that canceling of HBA smoothing helps predicting rotational settings, predNuCpos provides this option. However, for those who want to predict nucleosome occupancy in the original way with chemical maps, we encourage users to use NuPoP functions as it now provides chemical map-based predictions. In our functions HBA and localHBA, their excellent code was also adapted to calculate the scores of given 147-bp sequences independently of the genomic context.

References

Wang JP, Fondufe-Mittendorf Y, Xi L, Tsai GF, Segal E and Widom J (2008). Preferentially quantized linker DNA lengths in *Saccharomyces cerevisiae*. *PLoS Computational Biology*, 4(9):e1000175.

Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J and Wang JP (2010). Predicting nucleosome positioning using a duration hidden markov model. *BMC Bioinformatics*, 11:346.

Brogaard K, Xi L, and Widom J (2012). A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486(7404):496-501.

Moyle-Heyrman G, Zaichuk T, Xi L, Zhang Q, Uhlenbeck OC, Holmgren R, Widom J and Wang JP (2013). Chemical map of *Schizosaccharomyces pombe* reveals species-specific features in nucleosome positioning. *Proc. Natl. Acad. Sci. U. S. A.*, 110(50):20158-63.

Ichikawa Y, Morohoshi K, Nishimura Y, Kurumizaka H and Shimizu M (2014). Telomeric repeats act as nucleosome-disfavouring sequences in vivo. *Nucleic Acids Res.*, 42(3):1541-1552.

Voong LN, Xi L, Sebeson AC, Xiong B, Wang JP and Wang X (2016). Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell*, 167(6):1555-1570.

Fuse T, Katsumata K, Morohoshi K, Mukai Y, Ichikawa Y, Kurumizaka H, Yanagida A, Urano T, Kato H, and Shimizu M (2017). Parallel mapping with site-directed hydroxyl radicals and micrococcal nuclease reveals structural features of positioned nucleosomes in vivo. *Plos One*, 12(10):e0186974.