**DATA ACCESS**

Sequence data generated for this paper can be found at: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=qpobwiyuhjqppsr&acc=GSE93662>.

Coordinate files used to generate the figures for this paper can be found at: <https://github.com/CEGRcode/2018-Rossi_GenomeResearch>

ChIP-exo datasets were deduplicated because they were generated on the NextSeq (dual read)

All other datasets were not deduplicated because they were generated on the HiSeq (single read)

Analysis was performed on the GUI **ScriptManager v.010**, which is available for download at:

<https://github.com/CEGRcode/scriptmanager>

**Heatmaps and Composite Plots of ChIP-exo, WhIP-exo, and PB-exo data**

**Heatmaps of ChIP-seq and Native PB-seq**

Figure 1A-C; 2A,B; 3D; 4B,E; 5A; S2C; S3A; S4C; S5E; S6A,C; S7A,B; S8A,B; S10A,B; S11B

* Heatmaps and composite plots were generated using **Tag Pileup** from ScriptManager v0.10.
  + Read 1; Separate Strand, 0 bp tag shift, 1 bp bin size, set tags to be equal; sliding window 3
  + For ChIP-exo, WhIP-exo, and PB-exo; sorts were performed on the central 60 bp around the motif using **Sort BED by CDT** from ScriptManager v0.10.
  + For ChIP-seq and Native-PB-seq; sorts were performed on the central 400 bp around the motif using **Sort BED by CDT** from ScriptManager v0.10.
  + CDT files visualized in Java Treeview
  + Composites were constructed in Prism 7

**Composite Plots of H3 MNase ChIP-seq**

Figure 1D, F; 5C; S1B-F

* Composite plots were generated using **Tag Pileup** from ScriptManager v0.10.
  + H3 MNase data was from Batta et al (SourceID: 01101)
  + Read 1; Separate Strand, 75 bp tag shift, 3 bp bin size, set tags to be equal; sliding window 21
  + Composites were constructed in Prism 7

**Defining candidate motif occurrences via MEME analysis**

**Defining bound sites via Peak Pairs**

Figure 1E; 2C; 4A,D; S2A; S4A; S5E; S6B; S7C,D; S8C,D; S10A,B

* TAB files were generated using **BAM to scIDX** from ScriptManager v0.10.
  + Read 1
* Peak pair files (.gff) and meme.txt (position weight matrix of motif) were generated from the TAB files using the shell script **standard\_XO\_pipeline.sh**
  + Genetrack.py
    - For ChIP-exo, WhIP-exo, and PB-exo: Smooth=5, exclusion=10, filter=1
    - For ChIP-seq and Native PB-seq: Smooth=20, exclusion=40, filter=1
    - Blacklist filtered: rDNA locus
  + Cwpair\_gff.py
    - For ChIP-exo, WhIP-exo, and PB-exo: Up=0, down=80
    - For ChIP-seq and Native PB-seq: Up=0, down=400
  + meme -dna -minw 10 -maxw 20 -nmotifs 3 -time 600 -mod zoops -revcomp -oc <output folder name> <input fasta file>
    - expansion=80, topsite=500
  + fimo --oc <output folder name> --verbosity 1 --thresh 1.0E-4 <input MEME file> <input fasta file>
* When a motif p-value was used to generate lists, a new meme.txt was generated manually and **fimo** was run from the command line
  + fimo --oc fimo\_output --verbosity 1 --thresh 1.4E-4 candidate\_motif.txt <sacCer3.fa>
* When a specific sequence was used to generate lists, Will’s script was used. This script has been incorporated into ScriptManager and will be available in the next release.
* Bed files were expanded to 60 bp using **Expand BED File** from ScriptManager v0.10. Bedtools intersect command was used to find the overlap between the expanded BED files and peak pair files (.gff)
  + Bedtools intersect –u –a <coordinate file.bed> -b <peakpair file.gff>
* Lists were compared and venn diagrams generated using the following websites:
  + http://www.pangloss.com/seidel/Protocols/venn.cgi
  + http://jura.wi.mit.edu/bioc/tools/venn3way/

**Defining the location of motif occurrences**

Figure 1E, 2A, 5A; S2B; S3B; S7C; S8C; S10C

* Location of motifs was determined by taking the FIMO output and running the following scripts:
  + **Filter\_BED\_by\_proximity.pl**
    - Cutoff distance=100bp, clustered motifs removed
  + **Parse\_Coord\_by\_sacCer3\_feasures.sh**
    - Telomere: sacCer3\_Telomere.bed
    - ORF: sacCer3\_VerfifiedORF.bed
    - Promoter: sacCer3\_VerifiedORF\_Promoter\_500bp.bed

**Four Color (or Two Color) Plots**

Figure 2A,D; 4B,E; 5A; S3A; S5E; S10E

* Four color plots were generated by expanding the bed file to 30 bp using **Expand BED File** from ScriptManager v0.10.
* The sequence was retrieved using **FASTA from BED** from ScriptManager v0.10.
  + Genome fasta: sg11\_all.fa; force strandedness
* The plot was generated using **4Color Sequence Plot** from ScriptManager v0.10.
  + For 2 color plot: A/T=red, G/C=green

**DNA shape analysis**

Figure 3A,B,E; 4C,F; 5B; S5B,C,F; S6D,E

* For DNA shape analysis, sorted bed files were expanded to 30 (or 50) bp **Expand BED File** and as input for **DNA Shape from BED** from ScriptManager v0.10.
* The DNA shape values for the “Top 100” and “Bottom 100” motifs were rank ordered in Excel (=RANK.AVG(cell, range)), then a Mann-Whitney *U* test was performed on the ranks of the two groups.
  + U1 = sum of ranks of the “Top 100” sites
  + U2 = sum of ranks of the “Bottom 100” sites
  + U = U1 +U2
  + n1 = 100, n2 = 100
  + mean of ranks = U/2
  + Standard deviation =SQRT(n1\*n2\*(n1+n2+1)/12)
  + Z-score =(U1-(mean of ranks))/(Standard Deviation)
* If |Z| > 2, then the position was considered to have a significant DNA shape difference between the two groups of sites.
  + The Z threshold varied between figures depending on the number of sites that were analyzed.

**Correlation analysis**

Figure S4B

* Pearson Correlation heatmaps and dendrogram showing Euclidean distance average linkage between replicates was calculated using **correlation\_coeff\_v2.py**

**Structure visualization**

Figure S5A; S9A; S10C

* Crystal structures visualized with Pymol