Cell, Volume *165*

Supplemental Information

Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape

Ronan C. O'Malley, Shao-shan Carol Huang, Liang Song, Mathew G. Lewsey, Anna Bartlett, Joseph R. Nery, Mary Galli, Andrea Gallavotti, and Joseph R. Ecker

SUPPLEMENTAL INFORMATION

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

DAP-seq Genomic DNA Library Preparation

The DAP genomic DNA (gDNA) library was prepared as a standard high-throughput gDNA sequencing library for the Illumina platform. First, gDNA (5-10µg in 130µL elution buffer (EB): 10mM Tris-Cl, pH 8.5) was fragmented to an average of 200bp using a Covaris S2 and manufacturer recommended settings. The fragmented gDNA was then purified using Sera-Mag beads (Thermo) at a 1:2 DNA to beads ratio (130µL DNA, 260µL beads). The beads were incubated with the gDNA for 10 minutes, placed on a magnet to immobilize the beads, and the supernatant was removed. The beads were then washed twice with 500µL 80% ethanol and allowed to dry. Once dry they were resuspended in 100µL EB, incubated for 10 minutes, placed on the magnet, and the DNA-containing supernatant was transferred to a new tube. 5µg (in 34µL) of DNA was end repaired in a 50µL reaction using the End-It DNA End-Repair Kit (Epicentre), and incubated at 22°C for 45 minutes. The reaction was precipitated with isopropanol, resuspended in 32µL elution buffer, and used in a 50µL A-tailing reaction using dATP and Klenow Fragment 3'->5' exo- (NEB) incubated at 37°C for 30 minutes. A second isopropanol precipitation was performed and DNA was resuspended in 10µL EB. The DNA was ligated in a 50µL ligation reaction using T4 DNA Fast Ligase (Promega)) at 22°C for 50 minutes followed by heat inactivation at 70°C for 10 minutes. Following ligation, a Sera-Mag bead purification was performed at a 1:1 ratio and the resulting DAP-seq library was resuspended in 30µL EB. The adapters sequences are truncated Illumina TruSeq adapters; the TruSeq Universal and Index adapters correspond to the DAP-seq Adapter A: CACGACGCTCTTCCGATCT and Adapter B: GATCGGAAGAGCACACGTCTG. In the PCR after the DNA affinity-precipitation step, the full-length Illumina adapter sequences including the flow cell attachment and index information are introduced in the PCR primers.

DNA Affinity Purification Sequencing (DAP-seq)

A collection of 1,812 Gateway compatible full-length Arabidopsis TF-ORFs were assembled from Pruneda-Paz et al. and AtORFeome2.0 (Arabidopsis Interactome Mapping Consortium, 2011). 27 families present in Pruneda-Paz et al. that were not annotated as DNA-binding proteins were excluded from our collection. Clones were recombined using LR clonase (Life Technologies) into the pIX *in vitro* expression vector (Arabidopsis Interactome Mapping

Consortium, 2011)) modified to contain an N-terminal HALO-Tag (Promega). pIX-HALO-ORF plasmid DNA was extracted using the Qiaprep 96-Turbo DNA extraction kit, quantified with the Quantifluor dsDNA system (Promega), and normalized prior to expression. HALO fusion proteins were expressed using the TNT SP6 Coupled Wheat Germ Extract System (Promega) following the manufacturer's specifications for expression in a 50μL reaction containing on average ~800ng DNA with a 2hr incubation at 30°C. Typical reactions yields ranged from 50-500ng of protein as measured relative to purified Halo-GST protein standards detected with anti-Halo antibody (Promega) by western blot. Expressed proteins were directly captured using Magne HaloTag Beads (Promega; 10μL per expression reaction). Proteins were incubated with the beads on a rotator for 1hr at RT. The beads were then washed three times with 125μL of wash/bind buffer (PBS with 0.005% NP-40).

The protein-bound beads were then incubated with 50ng of adapter-ligated gDNA fragments on a rotator for 1hr at RT in 50μL wash/bind buffer. Beads were washed again three times using the same wash buffer to remove unbound DNA fragments. The HaloTag beads were then resuspended in 30uL EB and heated to 98°C for 10 minutes to denature the protein and release the bound DNA fragments into solution. The supernatant was transferred to a new well. For HaloTag protein reactions, 25μL were used in a 50μL PCR reaction using Phusion polymerase and the same cycling conditions as described above for ampDAP-seq for 20 cycles. PCR primers consisted of the full-length Illumina TruSeq Universal primer (5'AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT) and an Illumina TruSeq Index primer (5'GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-NNNNNN-ATCTCGTATGCCGTCTTCT GCTTG) where NNNNNN represents the 6 base pair sequence index used for sample identification. The PCR product was precipitated with 95% ethanol, washed with 70% ethanol, and resuspended in 21μL EB. DNA concentrations were determined using a Qubit (Life Technologies). Multiple sets of indexed DAP-seq sequencing libraries could be combined in a single flow cell lane. Most datasets described here were sequenced as pools of 48 DAP-seq libraries per flow cell lane (Illumina 2500). We found that ~2-3 million reads per experiment produced a high signal-to-noise DAP-seq datasets for Arabidopsis. The correct DAP-seq library concentration to achieve a specific read count can be calculated based on library fragment size. Negative control mock DAP-seq libraries were prepared as described above without the addition of protein to the beads.

Expression and binding-efficiency of Halo-fusion proteins were confirmed by western

blot for a subset of TFs from multiple families. All samples tested showed protein expression at the expected size (Figure S1E). We observed no noticeable difference in expression or Halo-Tag binding efficiency in samples that generated high quality DAP-seq results and those that failed to pass our quality filters, indicating that the failure of certain TFs was due to factors other than lack of expression of full length protein in the wheat germ extract.

We retested a set of 201 TFs to determine the reproducibility rate of DAP-seq using TFs that either succeeded or failed in the initial experiments (Table S2). A set of 32 TFs that were successful when initially tested, reproduced at a rate of 88% indicating a technical failure rate estimate of ~12%. We believe this technical failure is likely due to operational errors associated with the scale of the experiment (201 TFs tested) and may be lower when assaying a smaller number of TFs. However as our goal was to establish a technical failure rate of the high throughput approach, we believed it important to perform our quality control in a large-scale experiment. We also retested 168 factors that failed in the first experiment to determine the rate of rescue of failed TFs. We examined several families that showed either a high (bZIP, NAC) or low (MADS, GRF, C2H2) initial success rate to determine if family success rates influence the rate of recovery in retests. Repeated attempts resulted in successful DAP-seq datasets for only ~6% of TFs tested, compared to an 88% reproducibility rate for initially successful TFs (Table S2). Moreover, families with high initial success rates (bZIP, NAC) showed higher rescue, 18% and 13% respectively, than those with lower initial success rates. This is illustrated by the failure to rescue any of the 87 MADS family members in a retest (Table S2) suggesting that some TF families may require alternative conditions for optimal DNA-binding.

DAP-seq datasets for ARF5, and ZmARF5 were performed using *E. coli* expressed GST-fusions generated during the initial phases of DAP-seq development. TFs were recombined into pDEST15 and expressed as follows: GST-HAT2 fusion protein was expressed in a 10mL culture of One Shot BL21 Star (DE3) cells using 1mM IPTG to induce protein expression for 5hr at room temperature. After expression, the cells were pelleted, frozen overnight, and lysed using 1mL B-PER II Bacterial Protein Extraction Reagent (Thermo Scientific) mixed with 15mg lysozyme (Thermo Scientific) and 5µL protease inhibitors (Sigma). The GST-ARF5 and GST-ZmARF5 fusion proteins were expressed in 500ml of BL21 DE3 codon plus cells (Stratagene) with 0.4mM IPTG at 23°C for 4 hours. Cells were lysed similar to that described above with an additional sonication step. GST-fusion proteins were column purified using Glutathione Sepharose 4B (GE Healthcare) and ~2ug of purified protein was used in the affinity capture reactions as described for Halo-fusion proteins with the following

modifications: GST-fusion proteins were captured using Glutathione-Superflow Resin (Clontech; 300μL per pellet) and washed with 1mL of GST wash buffer (25mM Tris pH 7.5, 150mM NaCl, 1mM EDTA). The protein-bound resin was incubated with 800ng (or 5ug for maize) of adapter-ligated gDNA fragments, washed, and resuspended in 200μL EB prior to DNA elution. 100μL were used in a 200μL PCR reaction split across two wells. PCR primers and conditions were identical to those described above for Halo-fusions. The full-length maize ARF5 co-ortholog ZmARF29 (GRMZM2G086949; (Xing et al., 2011)) was PCR amplified from B73 cDNA (mixed stage ear).

Generation of Transgenic Lines

Recombineered gene tagging of ABI5 (AT2G36270), HB5 (AT5G65310) and ANAC055 (AT3G15500) were carried out as described previously (Alonso and Stepanova, 2014). An YPET-6xHis-3xFLAG tag was fused in frame right before the stop codon of the *ABI5*, *HB5* and *ANAC055* genes in the transformation-competent bacterial artificial chromosome (TAC) clone JAtY64K17, JAtY58F22 and JAtY59D10, respectively. The resulting constructs were introduced into wild-type *A. thaliana* (accession Col-0) by the floral dip transformation method (Bent and Clough, 1998). Single-insertion transgenic lines were selected by the Chi-square test on plates containing Linsmaier and Skoog medium (Caisson labs, USA) with 0.8% agar and 15 μg/ml glufosinate ammonium (Sigma-Aldrich, USA) from plants in the T2 generation. The expression of the tagged transcription factors was confirmed by western blotting. Homozygous transgenic lines were selected from the subsequent generation and used for bulking seeds. The ABI5-YPET line using in our ChIP-seq experiments exhibited mild over-expression of ABI5, but the plant responded to ABA treatment phenotypically and transcriptionally.

ChIP-seq

To ChIP ABI5, seedlings were germinated and grown on nylon mesh in hydroponics for 36 hours under long day light conditions, followed by 5 μM ABA treatment for 4 hours. To ChIP HB5, seedlings were grown in dark for 3 days. ChIP-Seq was carried out as described previously with minor modifications (Chang et al., 2013). Briefly, harvested seedlings were cross-linked by 1% formaldehyde solution (Sigma-Aldrich, USA) under vacuum for 20 minutes. After nuclei isolation, chromatin was sonicated to 100-400bp fragments. Native ABI5 in wild-type Col-0 plants was immunoprecipitated by a rabbit polyclonal ABI5 antibody (cat # ab98831, Abcam, USA). Tagged ABI5 or HB5 proteins in the transgenic lines were immunoprecipitated by

a rabbit polyclonal GFP antibody (cat # A11122, Thermo Fisher Scientific, USA). Col-0 was immunoprecipitated with either rabbit IgG or the GFP antibody as mock IP for the native ABI5 and tagged transcription factors, respectively. After elution and reverse crosslinking, ChIP DNA was used to generate sequencing libraries according to the Illumina ChIP-Seq instructions. The libraries were sequenced on an Illumina HiSeq 2500 with 100bp SR according to the manufacturer's instructions (Illumina, USA).

To ChIP ANAC055, etiolated seedlings expressing recombineered ANAC055 were grown in the dark on agar plates, as described previously (Chang et al., 2013). Plates were sprayed with a control solution of 0.5x Linsmaier and Skoog medium containing 0.05% (w/v) ethanol and seedlings harvested two hours later. Cross-linking, nuclei extraction and immunoprecipitation were conducted as for tagged ABI5, but using polyclonal anti-GFP antibodies supplied by David Drechsel (Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany). Control ChIP experiments were conducted using wild-type *A. thaliana* treated in the same manner. Immunoprecipitation was conducted using rabbit whole IgG (Covance) for these experiments. ChIP-seq libraries were generated and sequenced as for ABI5.

DAP-seq Data Processing

Reads were mapped to the TAIR9/10 genome sequence using bowtie2 (Langmead and Salzberg, 2012) version 2.0-beta7 with default parameters and post-processing to filter out reads mapped to multiple locations. Peak calling was done using GEM peak caller version 2.5 (Guo et al., 2012) with the TAIR9/10 genome sequence and following parameters "--f SAM --k_min 6 --kmax 20 --k_seqs 600 --k_neg_dinu_shuffle" limited to nuclear chromosomes only. For factors with technical replicates, GEM was called with the replicate mode. Quality control metrics were computed by the R package ChIPQC (Carroll et al., 2014). Read coverage across the genome was computed in deepTools (Ramírez et al., 2014) by extending the reads to 150bp (DNA fragment size of the assay) and normalized by the number of reads mapped to nuclear chromosomes to base pair FPKM values. Association of DAP-seq peaks with gene features and repeat regions was computed by the Genome Association Tester (Heger et al., 2013). The peak calling and analysis pipeline was created in the snakemake framework (Köster and Rahmann, 2012).

Correlation between Factors and Factor Replicates

A consensus peak set for each TF family was generated by merging peak summits from all the TFs in the family that were within 100bp of one another. Average FPKM values of DAP-seq read counts were calculated for 100bp up- and downstream of each region in the consensus and used in computing the Pearson correlation coefficients between replicate libraries of the same factor or between (merged) libraries of different factors.

Motif Discovery, Clustering and Scoring

Since the kmer search approach used by GEM tends to discover short motifs, we used the meme-chip tool (Machanick and Bailey, 2011) in MEME suite 4.10.1 to allow identification of longer motifs. For each set of GEM-called TF binding events, we retrieved 200bp sequences surrounding the top 600 events and ran meme-chip with default parameters. The top motif from each factor was selected and evaluated for central enrichment. Motif PWM models were clustered by functionalities provided in the MotIV BioConductor package (Mahony and Benos, 2007; Mahony et al., 2007; Mercier et al., 2011) and visualized by the motifStack package. The PWM were used to score for motif matches in the TAIR10 genome sequence by the motif search tool FIMO (Grant et al., 2011), with a zero-order background model of TAIR10 nucleotide composition. To compare the number of motif matches from DAP-seq and PBM PWM models, a threshold of 65% maximum score for each motif was used for all the factors.

Comparison of DAP-seq and ChIP-seq

ChIP-Seq reads were mapped to the TAIR9/10 genome sequence using bowtie2 (Langmead and Salzberg, 2012) version 2.0-beta7 with default parameters. Since the GEM peak caller reports individual point source binding events but the ChIP-seq peaks have a broad appearance, to identify ChIP-seq peaks we used MACS2 peak callers. For ABI5 ChIP-seq (Ab etiolated, Ab light, YPET light) we used MACS version 2.0.10 with the parameters "--gsize 1.19e8 --nomodel –shiftsize 150 –keep-dup auto –call-summits --bdg" and a mock IP sample as control. For consistency MACS2 was also used to call peaks for DAP-seq of these factors, with parameters "--gsize 1.19e8–keep-dup auto –call-summits –bdg --SPMR". The peaks reported (default minimum q-value cutoff 0.05) were further filtered by a minimum fold enrichment threshold of 3 for DAP-seq, 2.5 for ABI5 YPET, and 2 for the other ChIP-seq experiments. For HB5 and ANAC055 ChIP-seq, each with two replicates, we used the IDR pipeline with the MACS peak caller (Li et al., 2011). Peak comparison calculations were done using BEDTools 2.19.1 (Quinlan and Hall, 2010), BEDOPS 2.4.1 (Neph et al., 2012) and UCSC genome browser

utilities (Kuhn et al., 2013). ChIP peaks were divided into equal-sized quartiles by the minimum p-value of the motifs in each peak and percentage of peak overlap with DAP peaks was computed for each quartile. To calculate overlapping peak categories for the empirical cumulative distribution curves, a union peak set was first created by merging peak regions from the two assays. Each merged region was designated as “DAP-ChIP” if it contained peak regions from both assays; “DAP-only” or “ChIP-only” if it contained regions only from the respective assay. A peak was considered to be in a DHS if it overlapped with a DHS region by more than 50%. For comparisons of DAP and motif-based predictions, motifs discovered from the top 600 DAP-seq peaks and motifs from PBM or SELEX assay were used to scan the TAIR10 genome sequence with p-value threshold of 1E-4. A motif match was considered bound in DAP-seq or ChIP-seq if it completely fell within peak regions of either assay, and was in DHS if it fell inside a DHS region. DAP signals at motif matches were computed by averaging the normalized DAP-seq read depth (in FPKM) over regions in -10bp to +10bp flanking regions of the motif match, with non-covered bases counted as zero. Precision-recall curves and AUC were computed by the R PRROC package (Keilwagen et al., 2014).

The random forest classifiers including motif and environment features for predicting in vivo binding sites of ABI5 and ANAC055 from ChIP-seq can be represented as:

$$Y \sim M + E + X + S$$

where Y is a binary variable indicating whether a motif is bound in ChIP-seq, M is motif feature (described below), E is the level of methylation inside the motif, X is the number of motifs of the same factor found within 100bp and S is a set of four first order DNA shape features for 17-bp immediately flanking both sides of each motif (30 features for minor groove width, 30 features for propeller twist, 32 features for roll and 32 features for helix twist), computed by the DNAshapeR package (Chiu et al., 2015). To compare motif scores and DAP-seq signal, two random forest classifiers were trained for each ChIP-seq dataset, consisting of the same motif environment features (E, X, S) but differed by the motif feature M: in the first one M was the motif score for motifs with p-value threshold of 1.5E-4 and in the second one M was the DAP-seq read depth computed as above on the same set of motifs as in the first model. Training and testing of the classifiers were performed by functionalities provided by the R caret package (Kuhn, 2008) calling the randomForest package (Liaw and Wiener, 2002). For training, 50% of the data were used in 5-fold cross-validation repeated 5 times, using the one standard error rule for selecting tuning parameters and ROC as the performance metric. Precision-recall metrics

were computed using the remaining 50% of the data. Scaled variable importance were retrieved by the caret package and plotted as heatmap.

GO Enrichment and Gene Feature Overlap Analysis

Peak summits called by GEM were associated with the closest TAIR10 gene model using the BioConductor package ChIPpeakAnno (Zhu et al., 2010). Enriched GO terms were identified by the g:Profiler web service accessed via the R API (Reimand et al., 2011) limited to the Biological Process ontology, maximum size of functional category of 1500, no hierarchical filtering and corrected for multiple hypothesis testing by the default method g:SCS. The two most significantly enriched terms were plotted in the heatmap by the aheatmap function in the R package NMF (Gaujoux and Seoighe, 2010). Enrichment/depletion of DAP-seq peaks with gene features and repeat/non-repeat regions were computed by the Genome Association Tester (Heger et al., 2013).

Identification of mCG-only and mC-all Sites and Comparison of mC to DAP-derived Motifs

For the methylation analysis we used methylcytosine calls from *Arabidopsis thaliana* Col-0 leaf from the GEO accession number GSM1085222 (Schmitz et al., 2013). The same seed stock used to create this leaf methylome map was used for all the DAP-seq leaf gDNA libraries. A 5'-methylcytosine site was called at any cytosine with read coverage of three or more and a ratio of mC over total read depth (methylation site-frequency) greater than 15%. Using this threshold, 11% of all cytosines with sufficient coverage were found to be methylated in the Col-0 leaf tissue. To identify the mC-all regions, we examined 100bp on either side of each methylcytosine. If the site was a non-mCG then the site was considered mC-all. If the site was mCG but an additional non-mCG site was found in the 30bp window the region was also designated as an mC-all. All mCG sites that did not neighbor mC-all regions were designated as the mCG-only set. A motif match was considered methylated if the total methylation level in the motif match region (mC/C) exceeded 0.66. This high threshold ensured that most gDNA fragments associated with the motif matches will contain a 5'-methylcytosine to allow us to measure the impact of the chemical mark on TF binding. Unmethylated motif matches (mC/C of 0.0) are used to establish background control levels.

Quantification of Cytosine Content and Classification of Cytosine Context for Motifs

Normalized cytosine content of a motif was calculated by dividing the sum of cytosine information content (IC) across all positions in the motif by the total IC of the motif. To assign motifs to CG, CHG and CHH contexts, we first computed distributions of IC for each nucleotide A, C, G, and T in all positions in all motifs. Nucleotides at each position were considered *strong* if their IC at this position was within the top 15% of the IC distribution for this nucleotide. A motif was considered to contain CG if it contained neighboring strong C and strong G. Similarly, a motif contained CHG if it contained neighboring strong C, strong A/C/T and strong G, while CHH-containing motifs were classified as containing neighboring strong C, strong A/C/T and strong A/C/T. Both the motif and its reverse complement were scored. All IC calculations were computed by functionalities in the R package motifStack with TAIR10 nucleotide frequencies as background.

Microarray Analysis to Identify ARF5/MP Targets

Series matrix file of accession GSE13881 was downloaded from NCBI Gene Expression Omnibus and analyzed as described in (Schlereth et al., 2010). Briefly, the normalized intensity values were log2 transformed and probe sets with low variance across samples (the lowest 0.5 percentile in terms of interquartile range) were removed. A linear model was fitted for each gene, and moderated t-statistics were calculated by an empirical Bayes method (Ritchie et al., 2015) for two comparisons 1) between *mp* and GR-*bdl* –DEX and 2) between GR-*bdl* +DEX and –DEX. A probe set was differentially expressed if the adjusted p-value was lower than 0.001 and fold change was at least two. Sixty-nine genes that were differentially expressed in both comparisons were considered ARF5/MP targets. A set of 62 genes that had p-values greater than 0.6 in both comparisons were considered background genes.

SUPPLEMENTAL REFERENCES

Alonso, J.M., and Stepanova, A.N. (2014). A Recombineering-Based Gene Tagging System for Arabidopsis. In Methods in Molecular Biology, (New York, NY: Springer New York), pp. 233–243.

Bent, A.F., and Clough, S.J. (1998). Agrobacterium Germ-Line Transformation: Transformation of Arabidopsis without Tissue Culture. In Plant Molecular Biology Manual, (Dordrecht: Springer Netherlands), pp. 17–30.

Carroll, T.S., Liang, Z., Salama, R., Stark, R., and Santiago, I. de (2014). Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Frontiers in Genetics *5*, 1–11.

Chang, K.N., Zhong, S., Weirauch, M.T., Hon, G., Pelizzola, M., Li, H., Huang, S.-S.C., Schmitz, R.J., Urich, M.A., Kuo, D., et al. (2013). Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. eLife Sciences *2*, e00675.

Chiu, T.-P., Comoglio, F., Zhou, T., Yang, L., Paro, R., and Rohs, R. (2015). DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics 1–3.

Gaujoux, R., and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinformatics *11*, 367.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Guo, Y., Mahony, S., and Gifford, D.K. (2012). High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. PLoS Comput. Biol. *8*, e1002638.

Heger, A., Webber, C., Goodson, M., Ponting, C.P., and Lunter, G. (2013). GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics *29*, 2046–2048.

Keilwagen, J., Grosse, I., and Grau, J. (2014). Area under Precision-Recall Curves for Weighted and Unweighted Data. PLoS ONE *9*, e92209.

Köster, J., and Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1–26.

Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Briefings in Bioinformatics *14*, 144–161.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357–359.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics *5*, 1752–1779.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R News *2*, 18–22.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696–1697.

Mahony, S., and Benos, P.V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. *35*, W253–W258.

Mahony, S., Auron, P.E., and Benos, P.V. (2007). Inferring protein DNA dependencies using motif alignments and mutual information. Bioinformatics *23*, i297–i304.

Mercier, E., Droit, A., Li, L., Robertson, G., Zhang, X., and Gottardo, R. (2011). An Integrated Pipeline for the Genome-Wide Analysis of Transcription Factor Binding Sites from ChIP-Seq.

PLoS ONE *6*, e16432.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.

Reimand, J., Arak, T., and Vilo, J. (2011). g:Profiler--a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. *39*, W307–W315.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47–e47.

Schlereth, A., Möller, B., Liu, W., Kientz, M., Flipse, J., Rademacher, E.H., Schmid, M., Jürgens, G., and Weijers, D. (2010). MONOPTEROS controls embryonic root initiation by regulating a mobile transcription factor. Nature *464*, 913–916.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., et al. (2013). Patterns of population epigenomic diversity. Nature *495*, 193–198.

Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R., Zhao, Y. (2014). NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. 42, D98-D103.

Zhu, L.J., Gazin, C., Lawson, N.D., Pagès, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics *11*, 237.