Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic

Annelien Verfaillie[1], Dmitry Svetlichnyy[1], Hana Imrichova, Kristofer Davie[1], Mark Fiers[2], Zeynep Kalender Atak[1], Gert Hulselmans[1], Hana Imrichova[1], Valerie Christiaens[1], and Stein Aerts[1#]

[1]Laboratory of Computational Biology, Center for Human Genetics, University of Leuven, Leuven, Belgium

[2]VIB Center for the Biology of Disease, Leuven, Belgium

correspondence to: stein.aerts@med.kuleuven.be

Supplemental Materials and Methods

CHEQ-seq plasmid

The pGL4.23 plasmid (Promega cat No E8411) was altered between the KpnI and XbaI restriction site to contain a sequence consisting of a super core promoter, a synthetic intron (Arnold et al. 2013) and a venus reporter gene (Roure et al. 2007). To incorporate the barcodes into the modified pGL4.23 backbone a PCR was performed using the following primers:

GTCGGCGCGCCGATCNNNNNNNNNNNNNNNNNGCTTCGAGCAGACATGATAAGATAC

and TATGGCGCGCCTTACTTGTACAGCTCGTCCATGC. The product was restricted with AscI and ligated overnight at 16°C. The final plasmid was purified using an ethanol precipitation before electroporating them into competent bacteria and grown overnight. All ligated product was electroporated to maintain complexity, with no more than 120 ng of plasmid being transfected per 20 µl of electrocompetent cells (Invitrogen, cat No C6400-03). The final CHEQ-seq plasmid was extracted using a Giga prep (Qiagen no. 12191).

Bait design

ChIP-seq peaks for TP53 were called against input (GSE47043) as described before (Janky et al. 2014) and were filtered for centromere and telomere regions and ranked based on their peak score. The top 1700 regions were selected for bait design. Peaks with length <500 were extended to 500 bp. Larger peaks were binned in designed regions going up by 50 bp. The designed regions were checked for overlap, with overlapping regions being removed. As control, 100 promoters of housekeeping genes were selected (Eisenberg and Levanon 2013) and regions spanning the TSS 400 bp upstream and 100 bp downstream were designed. Likewise, 100 random regions unlikely to display any binding in the genome were selected and 500 bp regions were designed (Yip et al. 2012). After several assessment rounds for production quality, 1526 regions spanning ChIP-peaks, 94 positive controls and 66 negative controls were used as final input for the bait design.

Generating the CHEQ-seq input library

Genomic DNA was extracted from MCF7 cells (TP53 wild type) using the Qiagen blood and tissue kit (cat no. 69581). DNA fragments were sheared to 500-1000 bp using the Covaris, size selected on a gel and purified using Qiagen PCR purification columns. Next, adapters were ligated to around 3-4 µg input DNA following the NEBnext DNA library prep protocol using the following adapter sequences: CCATCTCATCCCTGCGTGTCTCCGACTCAG*T and CTGAGTCGGACACGCAACAGGGGATAG. The adapter-ligated DNA was amplified for 10-15 cycles using the KAPA-HIFI hot start ready mix (annealing temp 60°C; extension time 30 sec) using following primers: ATCTGTGTGTTGGTTCCATCTCATCCCTGCGTGTC, GTACCGGCCAGTTAGCTATCCCCTGTTGCGTGTC. DNA was purified using the AMPure XP Beads (Beckman Coulter, cat no A63880) after which it was used as input for capturing the targeted regions following the MYbaits protocol (Custom bait libraries, MYcroarray). At least three captures were performed and pooled after purification. The CHEQ-seq plasmid containing the barcode pool was linearized via PCR with the Phusion High-fidelity PCR master mix (cat no M0532S) (annealing temp 63°C; extension 100 sec) and using the following primers: CTAACTGGCCGGTACCTGAG, AACCAACACACAGATGTAATGAAAA. It was then treated with DpnI and purified. The restricted CHEQ-seq plasmid and ~250 ng input DNA were combined in a total of 4 infusion reactions (Clontech). The recombined library was precipitated overnight and transformed at 100 ng per 20 µl electrocompetent cells. Transformed cells were grown overnight (max 12h) and the final library harvested using a giga prep (Qiagen no. 12191).

Generating the STARR-seq input library

The same procedure was followed as described for CHEQ-seq with some minor changes. The captured fragments were ligated and PCR-ed respectively with the following adapters and primers: GATCGGAAGAGCACACGTCT and ACACTCTTTCCCTACACGACGCTCTTCCGATC*T; TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT and GGCCGAATTCGTCGAGTGACTGGAGTTCAGACGTGTGCTTCCGATCT.

The STARR-seq plasmid was linearized using restriction as described in (Arnold et al. 2013) .

Cell culture and transfection

MCF7 cells (TP53 wild type or TP53 knock down) were cultured at 37°C with 5% CO_2 in RPMI medium (+L-glutamate, Gibco) supplemented with 10% fetal bovine serum (Invitrogen), 0.4 mM sodium pyruvate (Gibco), 100 µm/ml penicillin/streptomycin (Invitrogen), 1x non-essential aminoacids (Gibco) and 10 µg/ml Insulin (Sigma). Cells were plated at 10 million cells per 15 cm plate at a confluency of ~70%. The day after, around 50 µg of DNA was transfected using PEI (Sigma, cat no. 764892-1G) at a 1.2:1 ratio. The next day, cells were stimulated with 5 µM Nutlin-3a for 24h before extracting RNA. A constitutively active GFP construct was transfected alongside as a control for transfection. Two conditions were obtained: "p53-high" (TP53 wild type cells stimulated with Nutlin-3a) and "p53-low" (TP53 knock down cells).

DNA and RNA extraction and cDNA preparation

CHEQ-seq and STARR-seq libraries were treated similarly. Cells were collected 48h after transfection. 1/6 of cells were used to extract the plasmid DNA using the qiagen spin miniprep kit (cat no 27104). The rest of the cells were used to extract RNA using the qiagen RNeasy mini kit (cat no 74104). mRNA was isolated using Dynabeads mRNA purification kit (thermoFisher) and followed by cDNA preparation using the Goscript reverse transcription system (cat no. A5000).

Library preparation for Illumina quantitative barcode sequencing

cDNA or (extracted) plasmid DNA was amplified with two rounds of PCR using the Phusion High-fidelity PCR master mix (cat no M0532S). The first round of PCR (annealing temp 56°C; extension time 8 sec.) used the following primers: CCTACACGACGCTCTTCCGATCTAGCTGTACAAGTAAGGCG and CAGACGTGTGCTCTTCCGATCTTGTATCTTATCATGTCTGCTCGAA. After purification, the second round

of PCR was done (annealing temp 64°C; extension time 10 sec) using the Illumina Truseq adapters. The library was purified using the AMPure XP Beads (Beckman Coulter, cat no A63880) and checked using the bioanalyzer (Agilent) before being sequenced on the Illumina high-seq platform. For STARR-seq the DNA and cDNA libraries were created as described before (Arnold et al. 2013)

Library preparation for PacBio long-read enhancer identification

The input library was amplified (annealing temp 65°C; extension time 65sec) using the Phusion High-fidelity PCR master mix (cat no M0532S) with the following primers: CCTACACGACGCTCTTCCGATCTTGTGTTGGTTCCATCTCATCC; CAGACGTGTGCTCTTCCGATCTTCATCAATGTATCTTATCATGTCTGC. The linear fragments of around 2-3 kb were purified using AMPure XP Beads (Beckman Coulter, cat no A63880). Libraries were subsequently sent for SMRT PacBio sequencing (Pacific Biosciences).

Illumina short-read quantitative data analysis

Reads were cleaned using cutadapt (-k 0, -x 0) to remove any residual Illumina adapters and the 5' and 3' plasmid backbone sequences. For CHEQ-seq remaining barcodes were filtered on length, keeping only those that match exactly 17 bp. A counts matrix with all barcodes was created from the cDNA samples and subjected to library size normalization. This was followed by normalization for the distribution of each barcode by taking the sum of the barcode count measured in the input and plasmids DNA samples. For STARR-seq reads were cleaned using cutadapt and mapped using STAR (Dobin et al. 2012). Reads were extended to 400 bp and reads mapping to the same chromosomal location were collapsed into unique regions. A count matrix for all obtained regions was generated from the cDNA samples and normalized for library size and library coverage using the plasmid DNA region counts.

PacBio long-read data analysis and barcode association

Reads representing the reverse complement were altered to match the correct template. Cutadapt was used to extract the enhancer-regions and barcode sequences using the intermediate venus sequence as an alternative adapter recognition sequence. Barcodes were filtered on size and trimmed to contain only the variable 17 nucleotides. Enhancer-regions were mapped to the genome using bowtie2 (Kim et al. 2013). Barcodes were coupled back to their corresponding enhancer-region using their PacBio ID. Only barcodes uniquely linked to one enhancer-region were kept. Next, enhancer-regions were collapsed if mapping to the same location in the genome.

Generating a final list of enhancer-regions

For CHEQ-seq a stringent grep approach was used to retrieve the PacBio identified barcodes from the barcode expression matrix. Next, expression data for both the p53-high and p53-low condition were retrieved and filtered for barcodes having zero counts in one or more samples. Next, the data was normalized in DESeq2 (Love et al. 2014, 2) after which differential expression was calculated. For STARR-seq the count matrix was normalized using DESeq2 and differential expression calculated. For both methods enhancer-regions were overlapped with the designed regions as well as with previously called ChIP-seq peaks (Janky et al. 2014) using intersectBed. Regions not overlapping were considered random regions in the genome. Also see Supplemental Table S1 and S3 for the final list of CHEQ-seq and STARR-seq regions and accompanying information.

Determining subsets and comparison of the methods

Regions overlapping for 60% with the design were considered good representatives of a ChIP-peak. This assumption was based on the fact that ROIs were designed to center around and cover the targeted ChIP peak. If we assume that within most called peaks from a ChIP-peak experiment the presumed binding site of the factor is located in the center of the peak, then a minimum overlap of 50% of a captured region with the design would be needed for the region to have had potential binding and thus show activity. Taken that most designed regions

were slightly larger than the peaks they were modeled on, a 60% overlap seemed prudent to ensure a good representation. In addition, a more stringent overlap would result in a larger loss of fragments (**Supplemental Figure S7**). Functional enhancers (positives) were determined as ChIP-seq peaks having a representative region overlapping at least 60% and significantly up-regulated (p-val <0.05 and log2FC >= 1.5). Down-regulated enhancers (down) were determined as ChIP-seq peaks having a representative region overlapping at least 60% and significantly down-regulated (p-val < 0.05 and log2FC =< 1.5). Negative enhancers were determined as ChIP-seq peaks represented by a region (min 60% overlap), not classified to significantly up or down-regulated peaks and not having a region that has a log2FC > 0 and p-val < 0.1 or log2FC > 0.5 and p-val < 0.2. Remaining ChIP-seq peaks (differentially expressed but not significant) were termed greyzone. ChIP peaks classified as both significantly up and down-regulated were termed ambiguous. ChIP peaks are not allowed to be classified into more than one category. Also see Supplemental Tables S2 and S4. Comparison of CHEQ-seq and STARR-seq was based on the determined subsets and the represented ChIP-seq peaks. Only peaks represented by both methods were considered in these analyses. For the 2-fold cross validation, peaks were randomly sampled from the positives and negatives resulting in a training group (164 positives and 172 negatives) and a validation group (173 positives and 178 negatives.

Motif discovery and motif scoring

HOMER (Heinz et al. 2010) and RSAT peak motifs (Thomas-Chollier et al. 2011) were run on the positive set with the negative set as background, both on the entire sets and on the randomly selected training sets. i-cisTarget (Imrichová et al. 2015) was run on the positive set. For HOMER, length of the motif was set at either length 19 or 20 (-len) to allow for discovery of the consensus TP53 motif in the *de novo* option. Motifs were selected based on their overall performance and low occurrence in the negative set. 3 motifs were selected from i-cisTarget (ranked first, second and sixth). The top motif of each subcategory for HOMER was selected: known motif and *de novo* motif of length 19 or 20. One of the top motifs for

each RSAT category was selected (6 nucleotides, 7 nucleotides and dyads) in addition to the best representative half site. The matrices for the selected TP53 motifs were retrieved from each analysis and images were generated using Enologos (Workman et al. 2005). For the 2-fold cross validation the following motifs were selected: the HOMER known motif and de novo motif length 19, the first, second and sixth iRegulon motifs and the top 6nt and 7nt RSAT motifs. For each obtained motif or the combination of all 10 motifs Cluster-Buster (Frith et al. 2003) was used to score the positive and negative sets using –c 0 and –m 0. The highest motif score (or CRM score) for each region was obtained and used to determine the predictive value of each motif to classify regions into positives or negatives. In short, the sensitivity and specificity for each motif was calculated and the area under the receiver operating characteristic curve (AUC) determined.

ChIP-seq against H3K27ac

ChIP-seq against H3K27ac was performed as described before (Verfaillie et al. 2015) using the anti-histineH3 acetyl K27 antibody (ab4729, Abcam). Data processing was performed as described in (Janky et al. 2014). in short, reads were cleaned and mapped to the human reference genome hg19 using bowtie (v2.0.0-beta3). Reads with a mapping quality below 4 were removed. Peak calling was performed using MACS (version 1.4.2, (Zhang et al. 2008)) with default p-value settings.

RNA-seq, differential expression and Gene assignment

RNA-seq for MCF7 TP53 knock down was extracted and performed as described previously (Janky et al. 2014). RNA-seq reads were mapped to the genome (Gencode v18) using TopHat2 2.0.9 with Bowtie2 2.1.0 (Kim et al. 2013). Read counts per gene were obtained from the aligned reads using *htseq-count* command from the HTSeq framework. The Bioconductor/R packages DESeq2 (Love et al. 2014) was used for normalization and differential gene expression analysis. For each ChIP-seq peaks, the closest gene was determined using GREAT (McLean et al. 2010) with the following association rule settings:

'single nearest gene' within 20kb. The following gene ontology terms were collected: GO Biological Processes, Panther Pathways, MsigDB Pathways and MSigDB Perturbations. Only TP53 related function were recovered and reported.

Training and cross-validation of a Random Forest model

The Random Forest was trained using 9 TP53 PWMs (eliminating the halfsite) on the CHEQ-seq TP53-responsive peaks (positives). As control either the CHEQ-seq indirect (negatives) were used or genomic sequences were randomly selected from the genome (20x selection), matching in length and GC content to the positives. Both trained models were subsequently applied genome-wide to predict TP53 binding sites. The overlap these predictions yielded ~21000 predicted binding loci, which clustered into strong, weak and unbound groups. A third model was trained on the 21000 predicted binding sites using the strongly bound sites as positives and the unbound as negatives. In addition to the 9 TP53 motifs several additional features were used: distance to CpG island or TATA promoters (or both); DNA shape features like propeller twist, helical twist and GC content; GC content; number of coding genes or lncRNAs in a 10kb, 50kb or 100kb space around the site.

As Random Forest implementation we used the scikit-learn Python package. Each classifier uses ensemble of 151 decision trees. The parameter max_features (responsible for number of features to consider when looking for the best split) was set to sqrt (number of features). To calculate the feature importance we used the Gini impurity criterion averaged across trees, using the whole training data. The quality of each model was estimated in 5-fold cross-validations.

Feature-vector representation of the DNA sequence

For each PWM the motif score was calculated employing a Hidden Markov Model as implemented in Cluster-Buster (Frith et al. 2003). Number of coding genes and lncRNAs was calculated using bedTools and a custom bash script. The file with TSSs of genes was downloaded from the UCSC Genome Browser. High confident subsets of lncRNAs were

downloaded from LNCipedia (Volders et al. 2015). Files with the positions of promoters with TATA-box and/or GpC islands have been downloaded from the FANTOM5 resource (Lizio et al. 2015, 5).

Processing, analysis and ranking of public data

The collected public ChIP-seq data against TP53 are summarized in Supplemental Table S6. Fastq data was downloaded and mapped using bowtie2 2.1.0 (Kim et al. 2013) and peaks were called using macs2 with q-value set at 0.05. All peaks were called against the same input data (GSE47043) as not all datasets provided their own input and to keep uniformity across the analysis. DHS-seq data for Estradiol treated MCF7 was downloaded from GSE29692. GRO-seq data was downloaded from GSE53964 and GSE53966 and mapped in the same way as the ChIP-seq data.

To generate the ranking of all predicted TP53 binding sites, OrderStatistics (Aerts et al. 2006; Stuart et al. 2003) was applied on scores representing counts of alignments from all TP53 ChIP-seq data, resulting in aggregation of the scores into a final ranking.

seqMINER and clustering of genome-wide TP53 binding sites

BAM files of all public data (15 samples) and our in-house data (TP53 ChIP-seq data in Nultin-3a stimulated MCF7 cells, TP53 ChIP-seq data in non-stimulated MCF7 cells and input (see also GSE47043) were loaded into seqMINER (Ye et al. 2011). Additionally, a bed file with all 21659 predicted TP53 binding sites was loaded. Alternatively in-house TP53 ChIP-seq data, H3K27ac in MCF7 cells or DHS-seq BAM files were loaded and compared across the positive and negative CHEQ-seq peaks. The flanking area was set at 2000 bp around the binding site. Heatmaps show the raw tag count coverage from each BAM file for each input site or peak. To determine the best number of clusters representing differentially bound sites, different values for k were tested in a Kmeans ranked clustering method and mean density determined per cluster. K larger than three did not add to the initial distinction

of three clusters with regards to the initially detected three distinct levels of mean density of binding. Therefore a k = 3 was selected (**Supplemental Figure S26**).

DNA shape calculations

DNA shape data indicating Helix Twist (HelT) and Propellor Twist (ProT) for HG19 were downloaded from ftp://rohslab.usc.edu/hg19/ in bigWig format (Chiu et al. 2015). Regions of 1000 bp flanking the predicted TP53 binding sites were extracted using bwtool (Pohl and Beato 2014). Fraction GC was derived from DNA sequence data extracted from the UCSC geneome browser (ref) using the same TP53 10000bp flanking regions. A mean score per nucleotide for HelT, proT and GC content were calculated across all regions, and each of the three subsets (strong, weak and unbound). Subsequently the mean was calculated over a rolling window and plotted.

Prediction of TP53 binding using Deep Learning

The deep convolutional network represents a multilayer structure applying series of functional mappings where the input for each layer comes from the previous one. The first layer receives initial data and the last layer outputs class label. The convolutional layer performs a one-dimensional convolution with a number of kernels (filters) where weights are learned from data. Next, the pooling layer computes maximal value in a window for each kernel output coming from the convolution layer. Finally, transformed information is integrated by two fully connected layers after which sigmoid activation function is applied. The network was trained using the RMSprop algorithm for Stochastic Gradient Descent with 100 training samples in each mini-batch and binary cross-entropy loss function for minimization. For implementation we used the Keras 0.2.0 library (https://github.com/fchollet/keras) with the Theano 0.7.1 backend. Calculations have been performed with NVIDIA K40c accelerator. The regularization parameters are: dropout proportion (fraction of outputs randomly set to 0) for layer 2: 10%; layer 3: 10%; layer 6: 50%; all other layers: 0%. The details of the CNN model architecture are as follows:

1. Convolution layer (120 kernels. Window size: 26. Step size: 1; activation=rectified linear unit)

2. Pooling layer (Window size: 13. Step size: 13)

3. Convolution layer (120 kernels. Window size: 20. Step size: 1; activation=rectified linear unit)

4. Convolution layer (120 kernels. Window size: 16. Step size: 1; activation=rectified linear unit)

5. Pooling layer (Window size: 8. Step size: 8)

6. Fully connected layer (output dimension=925; activation=rectified linear unit)

7. Fully connected (output dimension=2; activation=sigmoid output)

References

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent L-C, De Moor B, Marynen P, Hassan B, et al. 2006. Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**: 537–544.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**: 1074–1077.

Chiu T-P, Yang L, Zhou T, Main BJ, Parker SCJ, Nuzhdin SV, Tullius TD, Rohs R. 2015. GBshape: a genome browser database for DNA shape annotations. *Nucleic Acids Res* **43**: D103–D109.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2012. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* bts635.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends Genet* **29**: 569–574.

Frith MC, Li MC, Weng Z. 2003. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* **31**: 3666–3668.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**: 576–589.

Imrichová H, Hulselmans G, Kalender Atak Z, Potier D, Aerts S. 2015. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res* gkv395.

Janky R 's, Verfaillie A, Imrichová H, Van de Sande B, Standaert L, Christiaens V, Hulselmans G, Herten K, Naval Sanchez M, Potier D, et al. 2014. iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLoS Comput Biol* **10**: e1003731.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Koeppel M, van Heeringen SJ, Kramer D, Smeenk L, Janssen-Megens E, Hartmann M, Stunnenberg HG, Lohrum M. 2011. Crosstalk between c-Jun and TAp73alpha/beta contributes to the apoptosis-survival balance. *Nucleic Acids Res* **39**: 6069–6085.

Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S, et al. 2015. Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310165/ (Accessed December 22, 2015).
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*. http://biorxiv.org/content/early/2014/05/27/002832 (Accessed July 22, 2014).
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**: 495–501.
Pohl A, Beato M. 2014. bwtool: a tool for bigWig files. *Bioinforma Oxf Engl* **30**: 1618–1619.
Roure A, Rothbächer U, Robin F, Kalmar E, Ferone G, Lamy C, Missero C, Mueller F, Lemaire P. 2007. A Multicassette Gateway Vector Set for High Throughput and Comparative Analyses in Ciona and Vertebrate Embryos ed. J.-N. Volff. *PLoS ONE* **2**: e916.
Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**: 249–255.
Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, Helden J van. 2011. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res* gkr1104.
Verfaillie A, Imrichova H, Atak ZK, Dewaele M, Rambow F, Hulselmans G, Christiaens V, Svetlichnyy D, Luciani F, Van den Mooter L, et al. 2015. Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat Commun* **6**. http://www.nature.com/ncomms/2015/150409/ncomms7683/full/ncomms7683.html (Accessed April 25, 2015).
Volders P-J, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, Mestdagh P. 2015. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res* **43**: D174–180.
Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV. 2005. enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res* **33**: W389–392.
Ye T, Krebs AR, Choukrallah M-A, Keime C, Plewniak F, Davidson I, Tora L. 2011. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res* **39**: e35.
Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, et al. 2012. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**: R48.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.