Genomics Data 4 (2015) 17–21

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Analysis of changes to mRNA levels and CTCF occupancy upon TFII-I knockdown

Maud Marques, Rodrigo Peña Hernández, Michael Witcher *

The Lady Davis Institute of the Jewish General Hospital, Departments of Oncology and Experimental Medicine, McGill University, Montreal, Canada

ARTICLE INFO

Article history:
Received 26 September 2014
Accepted 29 September 2014
Available online 18 October 2014

Keywords:
CTCF
TFII-I
Microarray
ChIP-seq

ABSTRACT

CTCF is a key regulator of nuclear chromatin structure, chromatin organization and gene regulation. The impact of CTCF on transcriptional output is quite varied, ranging from repression, to transcriptional pausing and transactivation. The multifunctional nature of CTCF is mediated, in part, through differential association with protein partners having unique properties. We identified the general transcription factor TFII-I as an interacting partner of CTCF. To gain an understanding of the function of TFII-I in regulating gene expression and CTCF binding genome wide, we conducted microarray experiments following TFII-I knockdown and chromatin immunoprecipitation of CTCF followed by next generation sequencing (ChIP-seq) from the same TFII-I depleted cells. Here, we described the experimental design and the quality control and analysis that were performed on the dataset. The data is publicly available through the GEO database with accession number GSE60918. The interpretation and description of these data are included in a manuscript in revision (1).

Specifications	
Organism/cell line/tissue	*Mus Musculus, Wehi-231, B lymphocyte immature*
Strain	*(BLAB/c x NZB) F1*
Sequencer or array type	*Illumina HiSeq 2000 and Illumina BeadChips Mouse WG-6*
Data format	*ChIP-seq: Raw (Fastq) and processed (bed file and bedgraph file)* *Microarray: excel spreadsheet before and after normalization.*
Experimental factors	*Wehi231-CT vs Wehi231-TFII-I knockdown*
Experimental features	Microarray gene expression profiling to identify genes that are regulated by TFII-I. ChIP-seq purpose was to map CTCF binding sites affected by TFII-I depletion.
Consent	*NA*
Sample source location	*NA*

Direct link to deposited data

Deposited data are available here: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60918.

Experimental design, materials and methods

Cell line

The mouse B lymphocyte cell lines Wehi-231 expressing shRNA construct Control (Wehi-CT) or a shRNA construct directed against the transcription factor TFII-I (Wehi-TKII-I-KD) were used to investigate the effect of TFII-I depletion on global gene expression and CTCF binding.

Microarray and quality control

To identify genes regulated by TFII-I, we extracted total RNA from Wehi-CT and Wehi-TFII-I-KD from three independent samples. The quantity and the quality of the RNA samples were assessed by a Nanodrop spectrophotometer and Agilent Bioanalyser. Illumina BeadChIPs MouseWG-6 was used to perform expression analysis. Data preprocessing was carried out with Bioconductor package "lumi", and we used log2 transformation followed by quantile normalization [2,3]. Quality controls were performed before (Fig. 1A) and after (Fig. 1B) microarray data preprocessing. Reproducibility between biological replicates was evaluated by calculating the correlation coefficient R^2 (see example of the scatter plot Fig. 1C and D). Clustering of the microarray was performed to ensure correct segregation between Control and TFII-I knockdown samples (Fig. 1E). Identification of differentially expressed genes between Wehi-CT and Wehi-TFII-I-KD was made

* Corresponding author.
E-mail address: michael.witcher@mcgill.ca (M. Witcher).

http://dx.doi.org/10.1016/j.gdata.2014.09.012

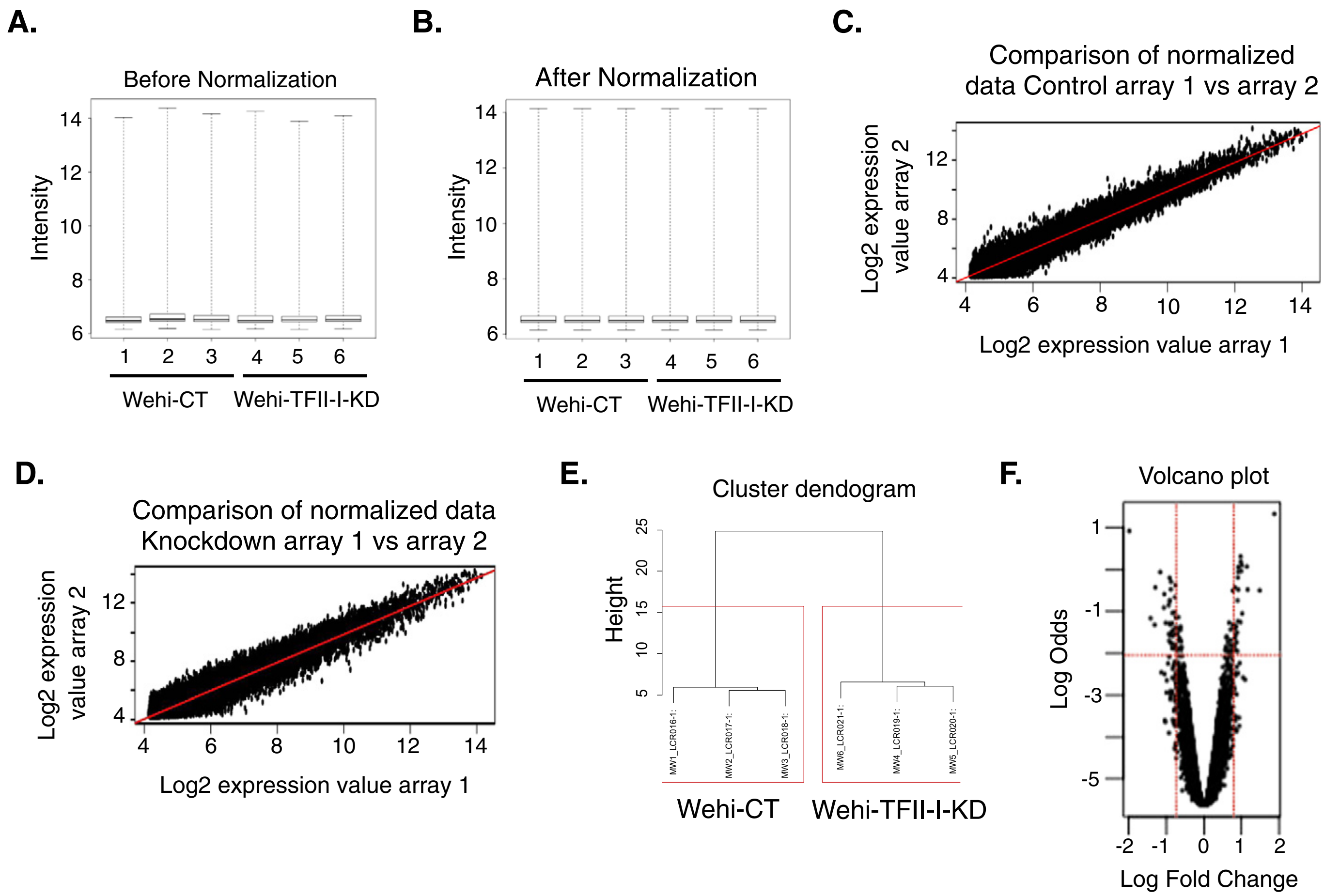

Fig. 1. Effect of normalization on microarray signal intensity. Before (A) and after (B) normalization distribution of signal intensity by array. (C) and (D) are scatter plots showing the comparison between two biological replicates of the log2 expression value. $R^2 = 0.95$ and $R^2 = 0.94$. (E) Cluster dendogram of the arrays in function of change in gene expression. (F) Volcano plots contrast significance as the negative logarithm of the *p*-value against log fold change between control cells and TFII-I knockdown cells.

Table 1
List of genes differentially regulated.

List of differentially expressed genes ($p < 0.05$) with a fold change >2 identified by microarray							
Up-regulated genes (55)				Down-regulated genes (62)			
ALDH3B1	WDR6	ATP6AP2	STARD13	GTF2I	ZFP219	MIB1	LIAS
CNR2	LMCD1	SFRS11	RAB8B	EGFL7	CYTIP	ZBTB17	RILPL2
CNR2	LSM14A	DEK	POLR3G	CYTH4	TBC1D10C	SHC1	STC2
LMCD1	ZFYVE26	MSH6	AATF	SLMO2	GSTT1	PFN1	FTL1
LRRC33	ANKRD49	HPRT1	NPM3	IL12A	NANS	D10ERTD610E	2310033F14RIK
BLK	AGPS	PLSCR1	POLE3	3300001G02RIK	2310008H09RIK	1600002K03RIK	GSTO1
AURKA	AF067061	RNF145	FAM178A	KHK	6330442E10RIK	TRUB2	1810026J23RIK
DDX24	TCIRG1	HAAO	VEGFB	CLEC2D	EBPL	ACTB	BST2
CREG1	BLVRB	GNAS	YBX3	1600012P17RIK	EIF2S2	RPN2	LOC629364
POLR2A	RBBP7	VPREB3	C730026J16	CALM3	PICK1	TMEM11	GUSB
ARPP19	MLLT4	CHFR	PLEKHA2	SERPINF1	MARCKS	HIST1H2BJ	AP3D1
PREI4	PANK4	GPR107	UBE2G1	PSMD8	CBR3	SEC63	RBM47
CEP120	DCPS	MT1	CKM	CDR2	SYNCRIP	VARS	LOC100044172
TWSG1	PDZD11	CDC5L		LCE1M	FCRL5	GPHN	DYNC1LI1
				KEAP1	JAGN1	FCGR2B	RRM2
				WDR68	EHD1		

Table 2
Reads count and numbers of peaks.

			Number of reads in millions			
Sample names	Antibody	Cell lines	Raw	No duplicate	MAPQ ≥ 20	Peak number
Ctl1	CTCF	Wehi-CT	43.58	36.1	28.6	24467
Ctl2	CTCF	Wehi-CT	36.1	32.9	26.1	23873
KD1	CTCF	Wehi-TFII-I-KD	36.2	32.3	25	19076
KD2	CTCF	Wehi-TFII-I-KD	36.46	23.7	16.9	15309

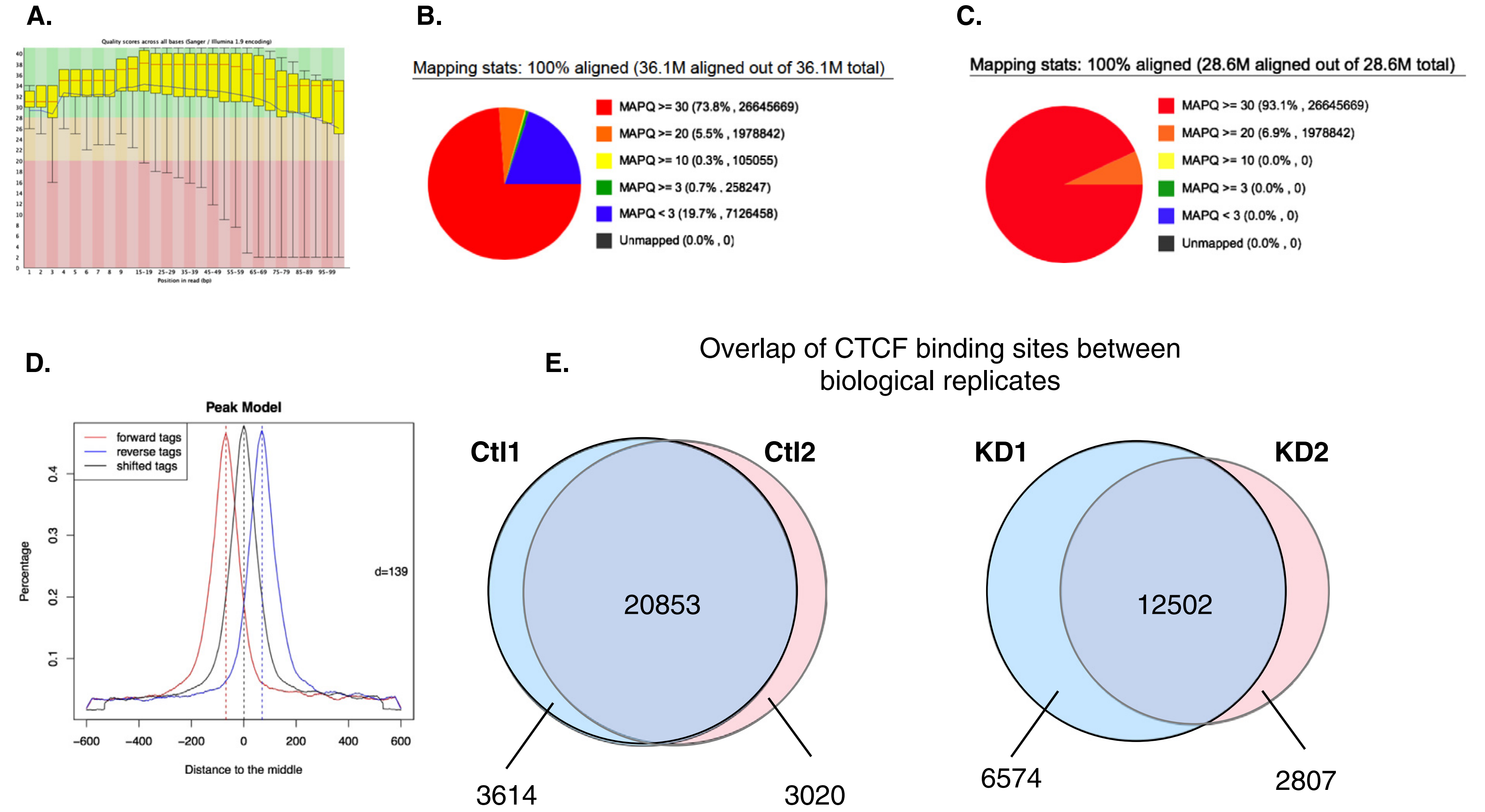

Fig. 2. Quality control for ChIP-seq raw data and alignment file. (A) Graph representing the per base quality using the Phred score. Pie chart obtained with SAMstat describing the distribution of the sequence alignment quality score before (B) or after (C) filtering. (D) Peak model produce by MACS. (E) Venn diagram representing the overlap of CTCF binding sites between biological replicates.

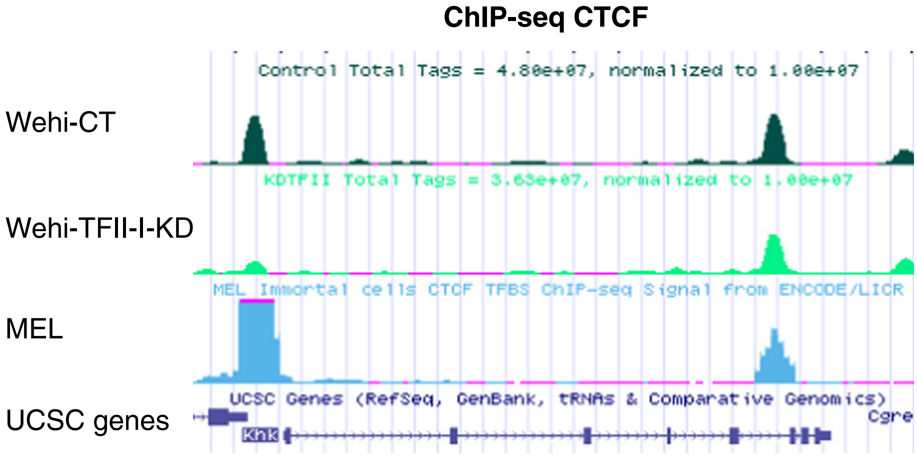
ChIP-seq CTCF
Wehi-CT
Control Total Tags = 4.80e+07, normalized to 1.00e+07
Wehi-TFII-I-KD
KDTFII Total Tags = 3.63e+07, normalized to 1.00e+07
MEL
MEL Immortal cells CTCF TFBS ChIP-seq Signal from ENCODE/LICR
UCSC genes
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Cgre
Khk

Fig. 3. Visualization of CTCF ChIP-seq data in the UCSC genome browser. Screenshot of UCSC genome browser showing CTCF ChIP-seq results in the Control and TFII-I knockdown samples. Previously published dataset for CTCF ChIP-seq in another hematopoietic cell line is also shown.

using Bioconductor package "limma" as shown with a volcano plot in Fig. 1F [4]. We identified 117 genes differentially regulated with a fold chance ≥2 and p-value ≤ 0.05 listed in Table 1. As a confirmation of the knockdown efficiency, we found *Gtf2i,* the gene coding for TFII-I, being the gene the most down regulated in our data.

ChIP-seq

To identify the CTCF binding sites that were affected by TFII-I depletion, we carried two independent ChIP-seq assays CTCF in Wehi-CT and Wehi-TFII-I-KD cells with CTCF antibody. Briefly, cells were collected and crosslinked with 1% folmaldehyde in PBS for 10 min at room temperature. Crosslinking reaction was stooped with Glycerine 125 mM and cells were washed with PBS and stored at −80 °C until assay was carried out. Cells were lysed and DNA sheered by sonication with cell lysis/ChIP buffer (0.25% NP-40, 0.25% Trinton-X, 0.25% Sodium deoxycholate, 0.1% SDS, 50 mM Tris pH 8.0, 50 mM NaCl, 5 mM EDTA) for 15 s, 15 times. Lysed cells were centrifuged for 15 min at 14,000 rpm at 4 °C and supernantant was collected. 1 mg of protein was precleared for 2 h with Protein G agarose beads (50% slurry blocked with salmon sperm) at 4 °C. Immunoprecipitation was carried out by adding 2 μg of antibody and 30 μl of agarose G beads and nutated overnight at 4 °C. After immunoprecipitation, beads were pelleted by centrifugation and were washed 4 times to remove unspecific binding using buffers with varying concentrations of salt. Buffers 1 to 3 contained 0.1% SDS, 1% Triton-X, 2 mM EDTA, 20 mM Tris pH 8.0 and 150 mM NaCl, 300 mM Nacl, 500 mM NaCl respectively. Buffer 4 contained 0.25 M LiCl, 1% NP-40, 1% Sodium deoxycholate, 1 mM EDTA and 10 mM Tris pH8.0. Two additional washes with TE were done to remove any residual buffer from the beads. Complexes bound to the beads were eluted with 500 μl of elution buffer (1% SDS, 1 mM EDTA, 50 mM Tris pH 8.0) at 65 °C for 25 min with occasional vortexing. Beads were pelleted by centrifugation and supernatant was collected. Crosslink reversal was achieved by adding 0.2 mM NaCl at 65 °C overnight. Next proteins (including DNA bound factors and antibodies) were degraded by a treatment with Proteinase K, carried at 45 °C for 1 h and a second incubation of 15 min at 65 °C. PCR purification kit (Qiagen) was used to retrieve the DNA following manufactured instruction and store at −20 °C. DNA was sent to the IRIC (Institut de Recherche en Immunologie et Cancérologie, Montreal, Canada) sequencing facility where both the library construction and sequencing (100bases, paired-end, HiSeq2000, Illumina) were carried out (Table 2).

ChIP-seq quality control and analysis

Quality of the sequencing was assessed using FastQC software, an example is presented in Fig. 2A (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Using FastX tool kit (http://hannonlab.cshl.edu/fastx_toolkit/), DNA sequences obtained were trimmed to 45 bases, filtered for high quality scores (>30), and duplicates were removed before being aligned to the mouse genome (U.S. National Center for Biotechnology Information (NCBI) Build 37, July 2007, mm9) using the BWA algorithm [5]. Quality of the alignment was assessed using SAMStat and only the sequences with MAPQ score ≥30 were kept for further analysis (Fig. 2B and C) [6]. The model based analysis of ChIP-Seq peak-finding algorithm was used to identify peaks in Wehi-CT and Wehi-TFII-I-KD conditions using the default settings and an example of peak model obtain with MACS is presented in Fig. 2D [7]. Overlap for CTCF binding sites between biological replicates was assessed using the intersect function of bedtools [8], the results are shown with Venn diagram (Fig. 2E). HOMER was used to annotate CTCF peaks, determine their genomic distribution and generated the bedgraph files to visualize the results in UCSC Genome Browser (homer.salk.edu/). We used previously published CTCF ChIP-seq data available in the UCSC genome browser as controls for our dataset (Fig. 3).

Discussion

Here, we described a dataset containing gene expression profiling using Illumina BeadChips (microarray) and ChIP-seq analysis of CTCF binding in mouse B cell lymphocyte cell lines expressing a shRNA construct against TFII-I, a general transcription factor. These data were generated to analyze the influence of TFII-I on the genomic targeting of the epigenetic regulatory protein CTCF, and understand how these two factors co-regulate gene transcription. With this dataset, we were able to show that TFII-I is important for targeting CTCF to a cohort of promoter regions where they co-operate to activate transcription. This finding sheds new light on how CTCF targeting at specific genomic regions can occur.

Conflict of interest

The authors have no conflicts of interest.

Acknowledgments

The authors would like to thank all the contributors to the original paper [1]. This work was made possible in part by grants from the Fonds de recherche Québec–Santé (# 003157) and Canadian Institut of Health Research (#201203MOP-273025-CBT-CFAF-109483) to Michael Witcher. Maud Marques is supported by a CIHR postdoctoral training grant (FRN53888).

References

[1] Maud Marques, Rodrigo Peña Hernández, Khalid Hilmi, Teijun Zhao, Sonia Victoria del Rincon, Todd Ashworth, Ananda Roy, Beverly Marie Emerson, Michael Witcher, Genome wide targeting of the epigenetic regulatory protein CTCF to gene promoters by the transcription factor TFII-I. 2014.

[2] P. Du, W.A. Kibbe, S.M. Lin, lumi: a pipeline for processing Illumina microarray. Bioinformatics 24 (2008) 1547–1548.

[3] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, W. Huber, BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21 (2005) 3439–3440.

[4] G. Smyth, Limma: linear models for microarray data. in: R. Gentleman, C. V., S. Dudoit, R. Irizarry, W. Huber (Eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Springer, New York, 2005, pp. 397–420.

[5] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25 (2009) 1754–1760.

[6] T. Lassmann, Y. Hayashizaki, C.O. Daub, SAMStat: monitoring biases in next generation sequencing data. Bioinformatics 27 (2011) 130–131.

[7] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li, X.S. Liu, Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9 (2008) R137.

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