Genomics Data 2 (2014) 230–236

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

ChIP-seq profiling of the active chromatin marker H3K4me3 and PPARγ, CEBPα and LXR target genes in human SGBS adipocytes

Mafalda Galhardo [a], Lasse Sinkkonen [a], Philipp Berninger [b], Jake Lin [c], Thomas Sauter [a,*], Merja Heinäniemi [d,*]

[a] Life Sciences Research Unit, University of Luxembourg, 162a Avenue de la Faïencerie, L-1511 Luxembourg, Luxembourg
[b] Biozentrum, Universität Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland
[c] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, House of Biomedicine, 7 Avenue des Hauts-Fourneaux, L-4362 Esch/Alzette, Luxembourg
[d] Institute of Biomedicine, School of Medicine, University of Eastern Finland, FI-70120 Kuopio, Finland

ARTICLE INFO

Article history:
Received 29 June 2014
Accepted 10 July 2014
Available online 6 August 2014

Keywords:
Adipocyte
ChIP-seq
Transcription factor

ABSTRACT

Transcription factors (TFs) represent key factors to establish a cellular phenotype. It is known that several TFs could play a role in disease, yet less is known so far how their targets overlap. We focused here on identifying the most highly induced TFs and their putative targets during human adipogenesis. Applying chromatin immunoprecipitation coupled with deep sequencing (ChIP-Seq) in the human SGBS pre-adipocyte cell line, we identified genes with binding sites in their vicinity for the three TFs studied, PPARγ, CEBPα and LXR. Here we describe the experimental design and quality controls in detail for the deep sequencing data and related results published by Galhardo et al. in Nucleic Acids Research 2014 [1] associated with the data uploaded to NCBI Gene Expression Omnibus (GSE41578).

Introduction

Specifications [*standardized info for the reader*] where applicable, please follow the Ontology for Biomedical Investigations: http://obi-ontology.org/page/Main_Page

Organism/cell line/tissue	Human/SGBS preadipocyte/adipose tissue
Sex	Male
Sequencer or array type	Illumina Genome Analyzer II
Data format	Raw and analyzed data
Experimental factors	ChIP-antibody used
Experimental features	Genome-wide binding or chromatin marker level (6 samples, including input control). SGBS preadipocyte cells originate from a patient with SGB syndrome. See Wabitsch M. et al. Int J Obes Relat Metab Disord. 2001 [2] for more details on differentiation protocol and origin of cells
Consent	See Wabitsch M. et al. Int J Obes Relat Metab Disord. 2001 [2]
Sample source location	See Wabitsch M. et al. Int J Obes Relat Metab Disord. 2001 [2]

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41578
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41629
http://www.ncbi.nlm.nih.gov/sra?term=SRP016497

Experimental design, materials and methods

Cell differentiation and experimental design

Chromatin was collected at day 0 and day 10 of adipogenesis for ChIP. SGBS cells differentiate within 10–12 days as determined by microscopic analysis (Oil Red O Staining). At this time point the cells are filled with small sized lipid droplets and are most responsive, whereas at later time points (20 days) the lipid droplets fuse and cells are less active (personal communication, Dr. Martin Wabitsch).

Specifically, SGBS cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM)/Nutrient Mix F12 (Gibco) containing 8 mg/l biotin, 4 mg/l pantothenate, 0.1 mg/mg streptomycin and 100 U/ml penicillin (OF medium) supplemented with 10% FBS in a humidified 95% air/5% CO_2 incubator. The cells were seeded into 10 cm plates, which were coated with a solution of 10 μl/ml fibronectin and 0.05% gelatine in phosphate-buffered saline. Confluent cells were cultured in serum-free OF medium for 2 days followed by stimulation to differentiate with OF media supplemented with 0.01 mg/ml human transferrin, 200 nM T3, 100 nM cortisol, 20 nM insulin, 500 μM IBMX and 100 nM rosiglitazone (Cayman Chemicals). After day 4, the differentiating cells

* Corresponding authors.
E-mail addresses: thomas.sauter@uni.lu (T. Sauter), merja.heinaniemi@uef.fi (M. Heinäniemi).

http://dx.doi.org/10.1016/j.gdata.2014.07.002

were kept in OF media supplemented with 0.01 mg/ml human transferrin, 100 nM cortisol and 20 nM insulin.

Chromatin immunoprecipitation

Nuclear proteins were cross-linked to DNA by adding formaldehyde directly to the medium to a final concentration of 1% for 8 min at room temperature. Cross-linking was stopped by adding glycine to a final concentration of 0.125 M and incubating for 5 min at room temperature on a rocking platform. The medium was removed and the cells were washed twice with ice-cold PBS. The cells were then collected in lysis buffer (1% SDS, 10 mM EDTA, protease inhibitors, 50 mM Tris–HCl, pH 8.1) and the lysates were sonicated by a Bioruptor UCD-200 (Diagenode, Liege, Belgium) to result in DNA fragments of 200 to 500 bp in length. Cellular debris was removed by centrifugation and the lysates were diluted 1:10 in ChIP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 167 mM NaCl, protease inhibitors, 16.7 mM Tris–HCl, pH 8.1). Chromatin solutions were incubated overnight at 4 °C with rotation with antibodies against H3K4me3 (4 μl per IP of 17–614, Millipore, Billerica, MA, USA), PPARγ (mixture of 0.5 μl per IP of sc-7196x, Santa Cruz Biotechnologies, Santa Cruz, CA, USA and 5 μl per IP of 101700, Cayman, Ann Arbor, MI USA), CEBPα (5 μl per IP of sc-61, Santa Cruz Biotechnologies), and LXRα (5 μl per IP, kind gift from Eckardt Treuter, Karolinska Institute, Stockholm, Sweden). The LXR antibody recognizes also LXRβ that maintains a constant low level of expression during differentiation. The immuno-complexes were collected with 20 μl of MagnaChIP protein A beads (Millipore) for 1 h at 4 °C with rotation. Non-specific background was removed by incubating the MagnaChIP protein A beads overnight at 4 °C with rotation in the presence of BSA (250 μg/ml). The beads were washed sequentially for 3 min by rotation with 1 ml of the following buffers: low salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris–HCl, pH 8.1), high salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 500 mM NaCl, 20 mM Tris–HCl, pH 8.1) and LiCl wash buffer (0.25 M LiCl, 1% Nonidet P-40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris–HCl, pH 8.1). Finally, the beads were washed twice with 1 ml TE buffer (1 mM EDTA, 10 mM Tris–HCl, pH 8.1). The immuno-complexes were then eluted by adding 500 μl of elution buffer (25 mM Tris–HCl, pH 7.5, 10 mM EDTA, 0.5% SDS) and incubating for 30 min on rotation. The cross-linking was reversed and the remaining proteins were digested by adding 2.5 μl of proteinase K (Fermentas) to a final concentration of 80 μg/ml and incubating overnight at 65 °C. The DNA was recovered by phenol/chloroform/isoamyl alcohol (25:24:1) extractions and precipitated with 0.1 volume of 3 M sodium acetate, pH 5.2, and 2 volumes of ethanol using glycogen as carrier. Immunoprecipitated chromatin DNA was then used as a template for real-time quantitative PCR or for library preparation and sequencing (performed at EMBL core facility).

Data processing and alignment

Sequencing reads were quality controlled using the FASTQC software v.0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The quality scores were consistently high along the read length and the samples had overall good quality based on multiple metrics (see Supplementary data file 1, Figs. 1–5). Possible clonality in the PCR step of library preparation was evaluated by counting reads mapping per

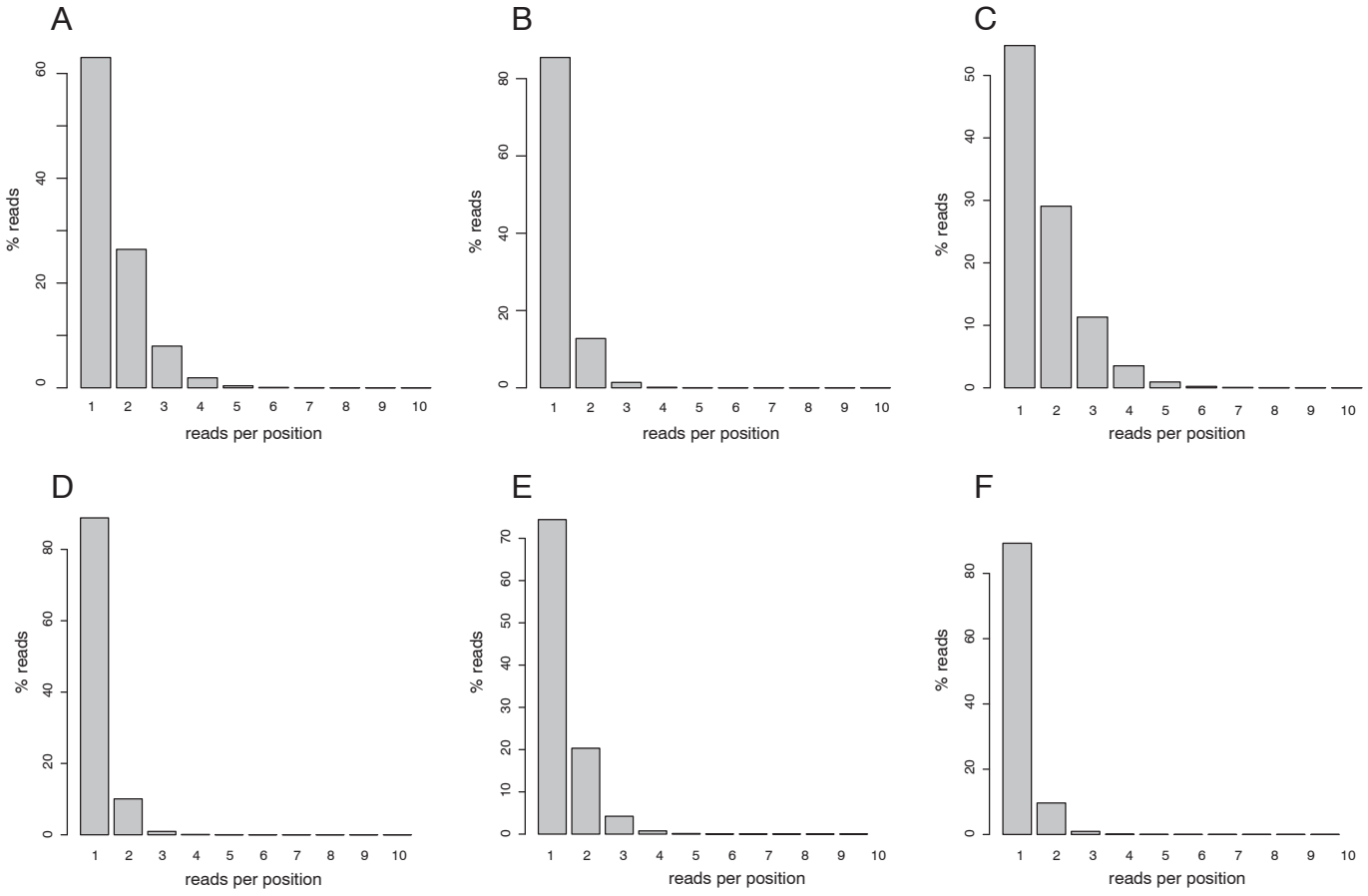

Fig. 1. Analysis of sample clonality. Histograms of clonal read depth are shown for the ChIP-seq samples. The bars indicate the number of reads per unique position. In an ideal ChIP-seq experiment there is a high fraction of single reads per position. Panels A–E show data of differentiated SGBS cells, and panel F shows data of preadipocytes. A. PPARg, B. CEBPa, C. LXR, D. H3K4me3, E. Input, F. H3K4me3 preadipocyte.

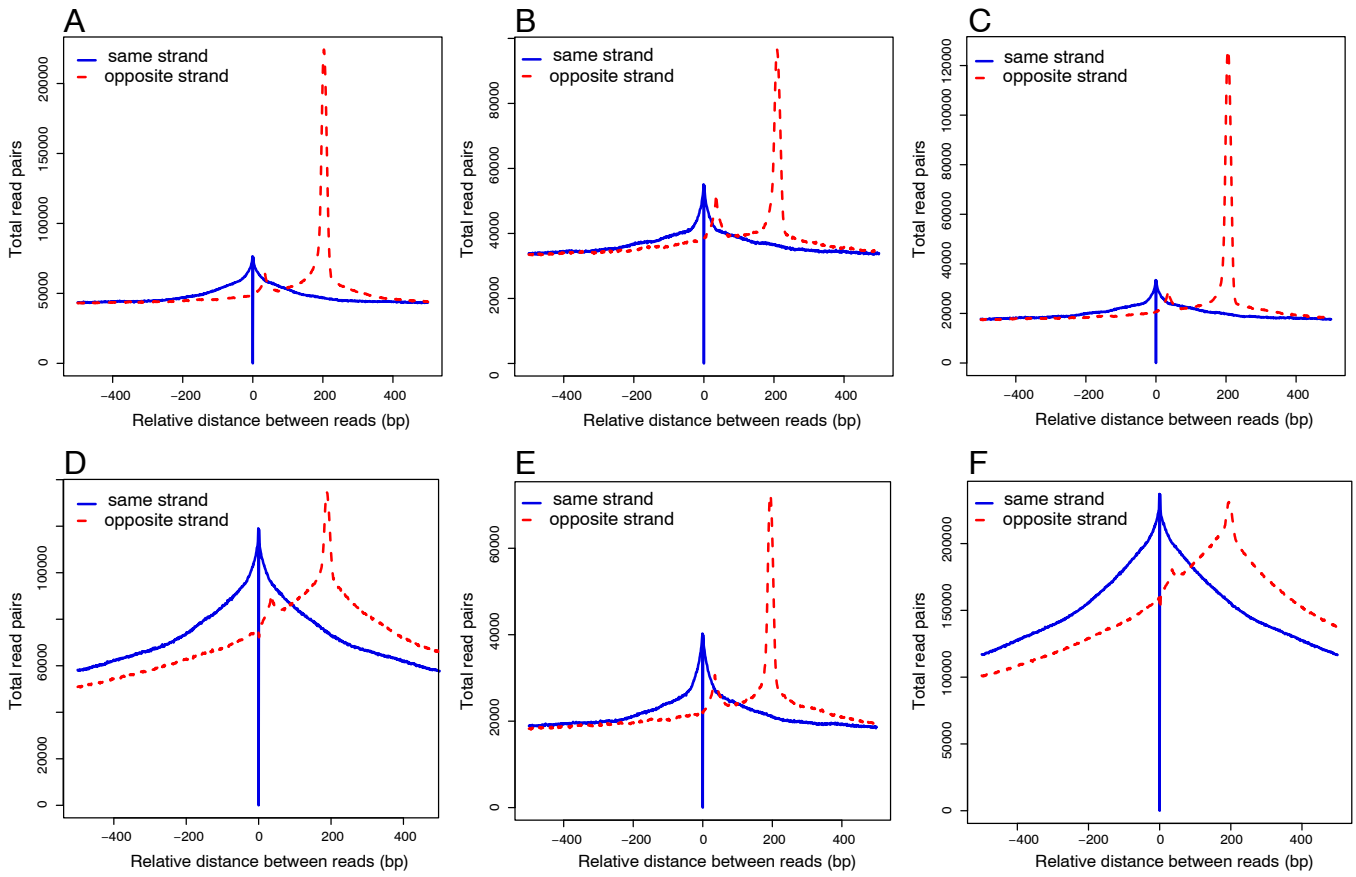

Fig. 2. Analysis of fragment length. The relative distance of reads mapping to ChIP-seq signal maximal from the two strands (positive and negative) is shown for the ChIP-seq samples. In a typical ChIP-seq experiment the peaks from opposite strands are 100–300 bp separated. Panels A–E show data of differentiated SGBS cells, and panel F shows data of preadipocytes. A. PPARg, B. CEBPa, C. LXR, D. H3K4me3, E. Input, F. H3K4me3 preadipocyte.

genomic position (Fig. 1). This revealed some degree of clonal amplification in the samples for PPARγ and LXR. Therefore, we chose to include a stack collapsing step to the preprocessing. The fragment length was estimated based on distance between reads mapping to positive and negative strands at peak locations (Fig. 2) and agreed well in each sample with the expected size of approximately 200 bp. When examining the read base pair content (Figs. 3 and 4), a deviation from the expected GC-content was observed in the input sample of SGBS cells and this sample was replaced in the downstream analysis by a new input obtained from similarly differentiated cells. The slightly higher GC-content in H3K4me3 peaks is expected as these reads derive from promoter proximal regions that have typically higher GC-content than the rest of the genome.

Specifically, the FASTX software v.0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html) was used to remove read artifacts and those reads that had low quality base pair calling (minimum quality score of phred 10 across the read length was required) and to collapse read stacks. Subsequently, reads were aligned to the hg19 human genome using the Bowtie software v0.1.25 [3] with the following settings: one mismatch was allowed, maximum three locations in the genome were allowed, and the highest quality match was reported. This resulted in 9608582 mapped reads for PPARγ, 19889853 for CEBPα, 15375177 for LXR, 12253403 for adipocyte H3K4me3, 12550706 for preadipocyte H3K4me3 and 18109349 for input. A script that downloads the deposited reads from the NCBI SRA database and produces aligned reads with these settings is provided (see Additional Data File 2).

Analysis of ChIP-seq signal

The H3K4me3 histone mark is often found at active transcription start sites (TSS). We were interested to assign each gene to H3K4me3-positive vs -negative categories. For this purpose, the mixture modeling approach implemented in the EpiChIP software v.0.9.7 [4] was applied. The results have been presented elsewhere [1] and a thorough user-guide is available from the tool website (http://epichip.sourceforge.net/tutorial.html). As instructed in the user guide, differently sized windows around the TSS regions were quantified to choose a proper window size. The region −750 to +1250 centered at Refseq TSS coordinates had the highest amount of signal and was therefore chosen for signal quantification. Two distributions were clearly visible: the higher values corresponding to actual chromatin marker signal distribution separated from the background distribution. The software then assigns a probability for each TSS region that indicates whether it corresponds to the signal distribution. We employed the default settings and used the noise, unclassified and signal results to compare the preadipocyte and the adipocyte TSS activity.

TF peak detection was performed using the Quest software v.2.4 [5]. To allow configuring all settings, we turned on the advanced mode. Parameters were generated using the command QuEST_2.4/generate_QuEST_parameters.pl -bowtie_align_ChIP sample.bowtie -bowtie_align_RX_noIP input.bowtie -gt genome_table_hg19 -ap sampleFolder -ChIP_name sampleChIP –advanced. Default settings were otherwise applied except for the mappable genome fraction

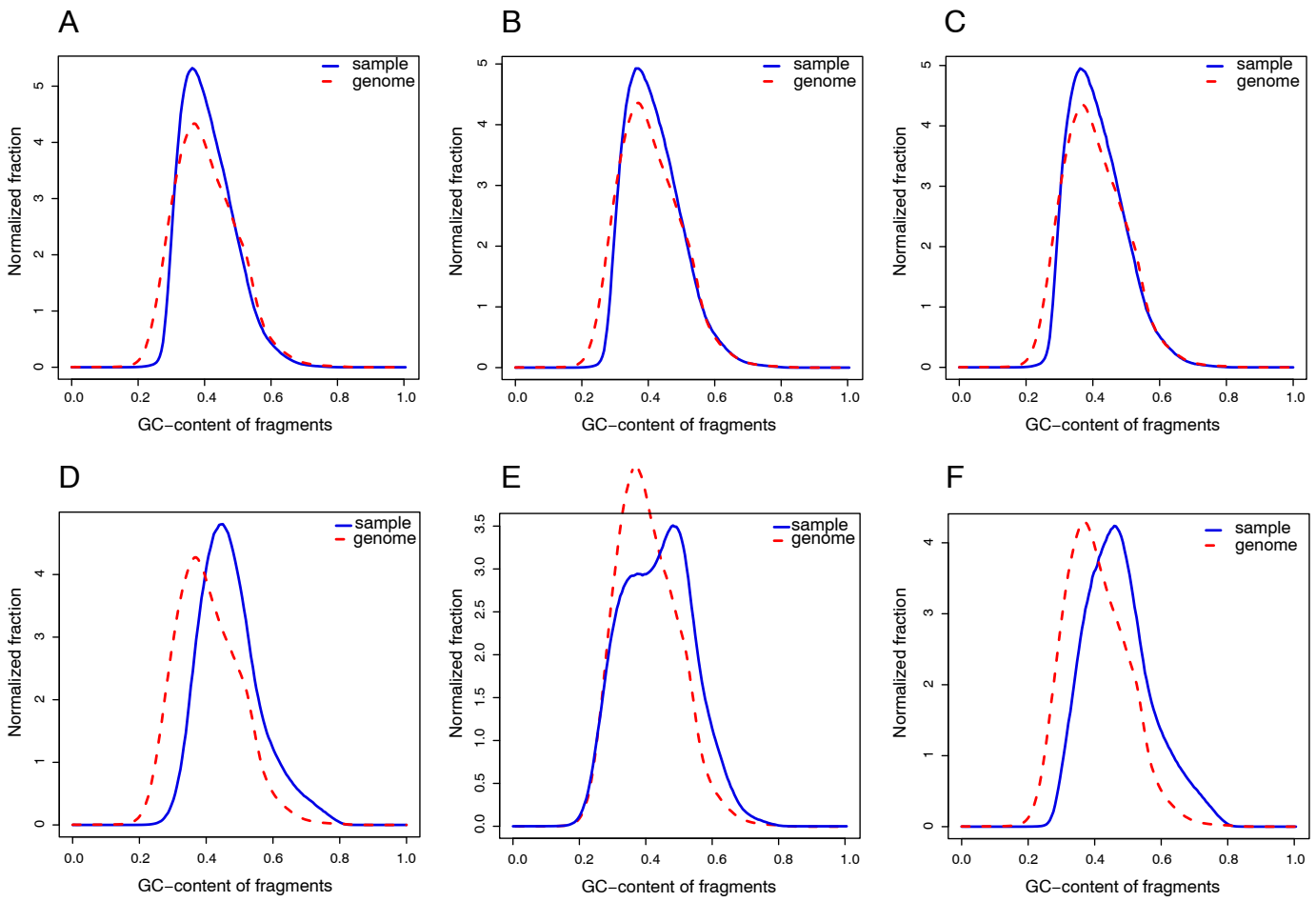

Fig. 3. GC-content of reads. The GC-content of the reads compared to that of the genome (hg19) is shown for the ChIP-seq samples. Typical ChIP-seq experiments with TF antibodies show a distribution that closely resembles that of random fragments from the genome. Gene proximal areas typically have higher GC-content and this is reflected in the H3K4me3 samples that have peaks nearby transcription start sites. Panels A–E show data of differentiated SGBS cells, and panel F shows data of preadipocytes. A. PPARg, B. CEBPa, C. LXR, D. H3K4me3, E. Input, F. H3K4me3 preadipocyte.

(set to 0.88) and enrichment (ChIP enrichment set to 15 and ChIP to background enrichment to 2.5). The final peak lists were filtered to remove peaks with q-value (fdr) above 0.001 (−logQvalue > 3). Based on examining the signal wiggle files, cut-offs for low-occupancy (enrichment > 15) and high occupancy (enrichment > 30) binding sites were defined. Finally, using the UCSC Table Browser, we obtained a file (group: Repeats, track: RepeatMaster) corresponding to satellite repeats (#filter: rmsk.repClass = 'satellite') and removed peaks overlapping these regions.

TF motif detection by the MEME-ChIP software [6] was performed using the high occupancy peaks. We used the setting -nmotifs 10 -minw 6 -maxw 30 and matched the motifs found to the JASPAR 2009 CORE database. MEME analysis using 600 randomly chosen trimmed (central 100 bp) input sequences revealed the respective TF motif as top motif present in the sample in each case (Fig. 6). The canonical binding sites matched to the TF analyzed were: MA0065.2 PPARG::RXRA and MA0065.1 PPARG::RXRA for PPARγ peaks; CEBPα: MA0102.2 (CEBPA), MA0102.1 (Cebpa) for CEBPα peaks; while a close match to motif reported by Feldman et al. [7] was found for LXR. This indicates that the antibody collection and downstream analysis were successfully performed. A script to run the motif analysis is provided (see Additional data file 2).

Discussion

Here we describe deep sequencing data obtained from human SGBS preadipocyte differentiation. This dataset is composed of data derived using the Illumina Genome Analyzer II. We demonstrated genome-wide binding pattern of key adipogenic TFs that were shown to co-occupy several loci. Further, this dataset is part of a GEO Superseries (GSE41578) and we have used it in combination with gene expression data to associate putative target genes in [1]. To further analyze the data in an integrative manner, we introduced the web portal IDARE (Integrated Data Nodes or Regulation) in [1] for interactive data exploration of the results within the metabolic network context, available at http://systemsbiology.uni.lu/idare.html, including a detailed user guide. Direct links to our ChIP-seq track hub which can be used to visualize the signal at any genomic loci are available from this website. We also provide results on motif analysis presented in this paper that can be used for further analysis

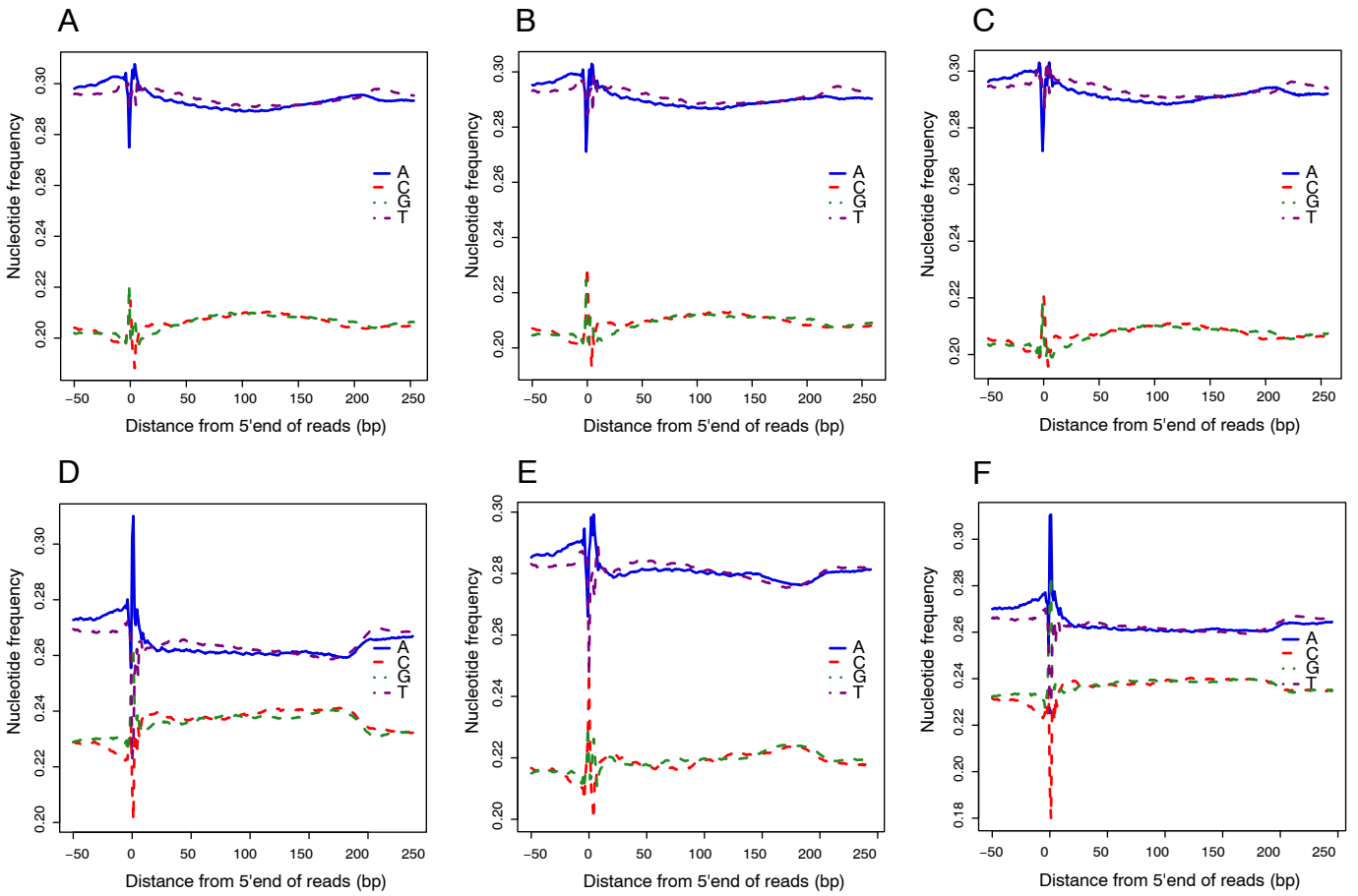

Fig. 4. Nucleotide frequencies along the read length. The frequency of each nucleotide along the read is plotted for each ChIP-seq sample. The frequencies of A and T are typically very similar to those of G and C. Gene proximal areas typically have higher GC-content and this is reflected in the H3K4me3 samples that have peaks nearby transcription start sites. Panels A–E show data of differentiated SGBS cells, and panel F shows data of preadipocytes. A. PPARg, B. CEBPa, C. LXR, D. H3K4me3, E. Input, F. H3K4me3 preadipocyte.

of combinatorial TF binding. Results from the data have increased our understanding of the TF-mediated control of human adipogenesis.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.07.002.

Acknowledgments

We would like to acknowledge the following funding sources: Fonds National de la Recherche Luxembourg [AFR, AM2c and C08/BM/01]; University of Eastern Finland and Fondation du Pélican de Mie et Pierre Hippert-Faber under the aegis of Fondation de Luxembourg.

References

[1] M. Galhardo, L. Sinkkonen, P. Berninger, J. Lin, T. Sauter, M. Heinäniemi, Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network. Nucleic Acids Res. 42 (2014) 1474–1496.

[2] M. Wabitsch, R.E. Brenner, I. Melzner, M. Braun, P. Möller, E. Heinze, K.M. Debatin, H. Hauner, Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. Int. J. Obes. Relat. Metab. Disord. 25 (2001) 8–15.

[3] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10 (2009) R25.

[4] D. Hebenstreit, M. Gu, S. Haider, D.J. Turner DJ, P. Liò, S.A. Teichmann, EpiChIP: gene-by-gene quantification of epigenetic modification levels. Nucleic Acids Res. 39 (2011) e27.

[5] A. Valouev, D.S. Johnson DS, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R.M. Myers, A. Sidow, Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat. Methods 5 (2008) 829–834.

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

[7] R. Feldmann, C. Fischer, V. Kodelja, S. Behrens, S. Haas, M. Vingron, B. Timmermann, A. Geikowski, S. Sauer, Genome-wide analysis of LXRα activation reveals new transcriptional networks in human atherosclerotic foam cells. Nucleic Acids Res. 41 (2013) 3518–3531.

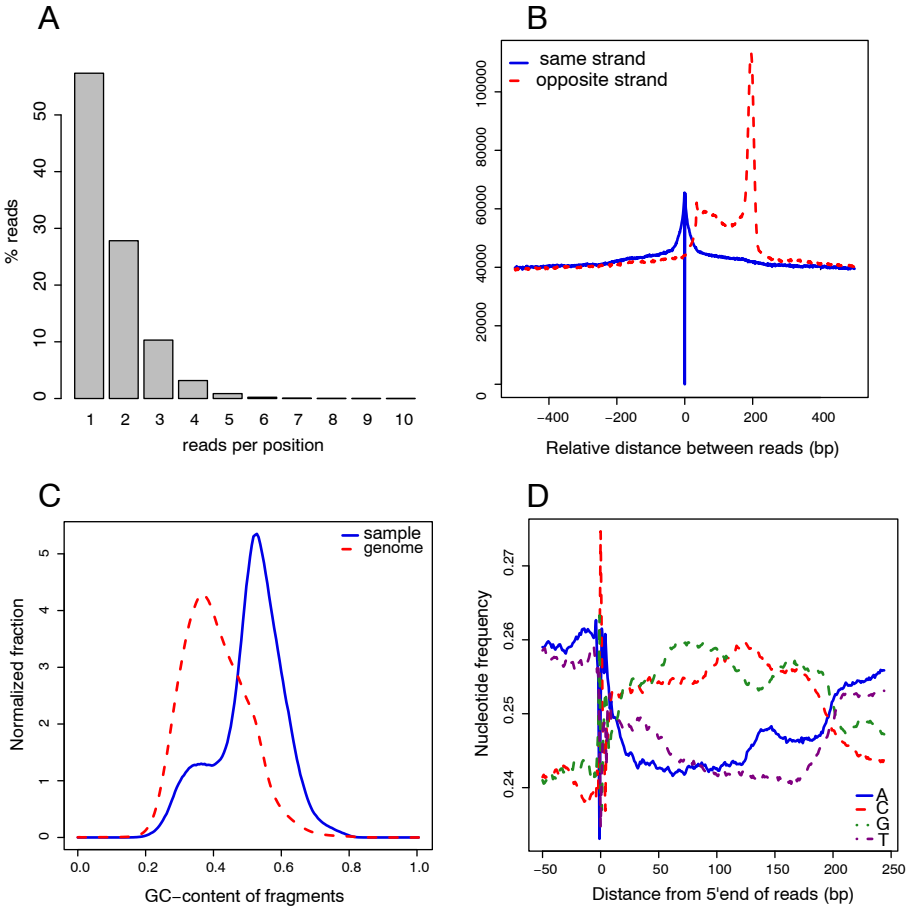

Fig. 5. Quality control data for discarded sample. One input sample did not match the other samples in terms of the quality results. As in Figs. 1–4, analysis of sample clonality is shown in A, analysis of fragment length in B, GC-content in C and nucleotide frequencies along the read length in D.

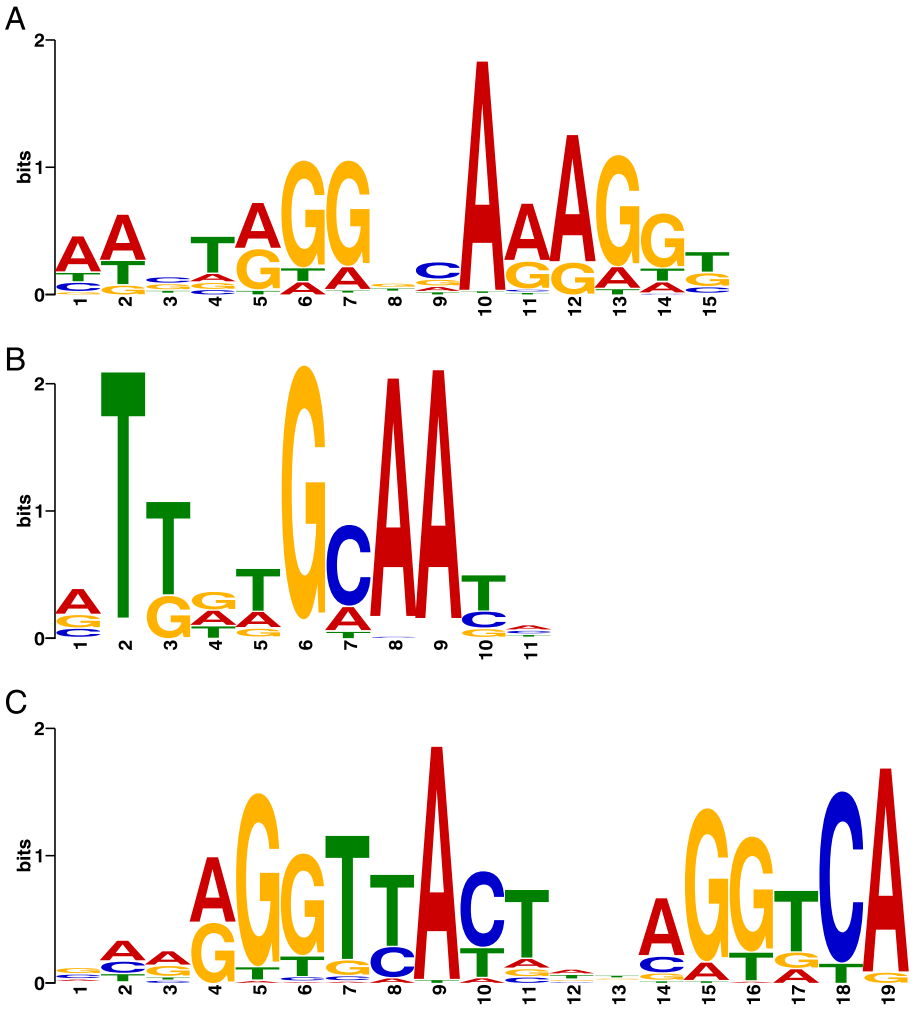

Fig. 6. TF *de novo* motif analysis. The top motifs detected in PPARg (in A), CEBPa (in B) and LXR (in C) peaks are shown. Letter height indicates information content in bits. Those positions with high information content are typically well conserved between binding sites and correspond to protein–DNA contacts.