Genomics Data 5 (2015) 368–370

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Description of an optimized ChIP-seq analysis pipeline dedicated to genome wide identification of E4F1 binding sites in primary and transformed MEFs☆

Thibault Houlès, Geneviève Rodier, Laurent Le Cam, Claude Sardet *, Olivier Kirsh **

Inserm U1194, Université Montpellier, Institut Régional du Cancer de Montpellier (ICM), 208 rue des Apothicaires, 34298 Montpellier, France

ARTICLE INFO

Article history:
Received 26 June 2015
Accepted 12 July 2015
Available online 14 July 2015

ABSTRACT

This Data in Brief report describes the experimental and bioinformatic procedures that we used to analyze and interpret E4F1 ChIP-seq experiments published in Rodier et al. (2015) [10]. Raw and processed data are available at the GEO DataSet repository under the subseries # GSE57228.

E4F1 is a ubiquitously expressed zinc-finger protein of the GLI-Kruppel family that was first identified in the late eighties as a cellular transcription factor targeted by the adenoviral oncoprotein E1A13S (Ad type V) and required for the transcription of adenoviral genes (Raychaudhuri et al., 1987) [8]. It is a multifunctional factor that also acts as an atypical E3 ubiquitin ligase for p53 (Le Cam et al., 2006) [2]. Using KO mouse models we then demonstrated that E4F1 is essential for early embryonic development (Le Cam et al., 2004), for proliferation of mouse embryonic cell (Rodier et al., 2015), for the maintenance of epidermal stem cells (Lacroix et al., 2010) [6], and strikingly, for the survival of cancer cells (Hatchi et al., 2007) [4]; (Rodier et al., 2015) [10]. The latter survival phenotype was p53-independent and suggested that E4F1 was controlling a transcriptional program driving essential functions in cancer cells.

To identify this program, we performed E4F1 ChIP-seq analyses in primary Mouse Embryonic Fibroblasts (MEF) and in $p53^{-/-}$, H-Ras^{V12}-transformed MEFs. The program directly controlled by E4F1 was obtained by intersecting the lists of E4F1 genomic targets with the lists of genes differentially expressed in E4F1 KO and E4F1 WT cells (Rodier et al., 2015). We describe hereby how we improved our ChIP-seq analyses workflow by applying prefilters on raw data and by using a combination of two publicly available programs, Cisgenome and QESEQ.

Specifications	
Organism/cell line/tissue	*Mus musculus*, Mouse Embryonic Fibroblasts (MEF). Primary MEFS and p53 KO, H-Ras^{V12} MEFs
Sex	
Sequencer or array type	Illumina HiSeq 2000
Data format	ChIP-seq Raw: FASTQ. Processed: BED, COD, BAR, TXT files
Experimental factors	
Experimental features	ChIP-seq experiments to identify E4F1 direct target genes.
Consent	NA
Sample source location	NA

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57228

2. Experimental design, materials and methods

2.1. Cells

$E4F1^{-/-}$ and $E4F1^{-/flox}$ mouse models have been previously described [1,6]. Primary MEFs were isolated from $E4F1^{+/flox}$ and $E4F1^{-/flox}$-E13.5 embryos obtained by crossing $E4F1^{+/-}$ and $E4F1^{flox/flox}$ animals [10]. Embryos were sliced and incubated for 1 h in trypsin at 37 °C, 5% CO_2. Isolated MEFs were then harvested and plated in new dishes for amplification.

To generate $p53^{-/-}$, H-Ras^{V12}-transformed cells we infected p53 KO, $E4F1^{+/flox}$ (or $E4F1^{-/flox}$) MEFs with retroviruses encoding the activated mutant of H-RAS (H-RAS^{v12}). E4F1 KO MEFs (primary and

☆ Equipe Labellisée Ligue Contre le Cancer (2012–2014).

* Corresponding author.
** Correspondence to: O. Kirsh, Epigenetics and Cell Fate, University Paris Diderot, Sorbonne Paris Cite, UMR7216 CNRS, 75013 Paris, France.
E-mail addresses: claude.sardet@igmm.cnrs.fr (C. Sardet), olivier.kirsh@univ-paris-diderot.fr (O. Kirsh).

http://dx.doi.org/10.1016/j.gdata.2015.07.004

transformed) were obtained by infecting cells with a retrovirus encoding the Cre recombinase and were used for ChIP QPCR validation and microarray analysis.

2.2. Chromatin immunoprecipitation (ChIP)

Two subconfluent 150 cm^2 dishes of MEFs ($\pm 3 \times 10^7$ cells) of primary or transformed p53 KO, $E4F1^{+/flox}$, were trypsinized and counted before crosslink for 8 min with 1% formaldehyde (SIGMA) directly added in medium. Fixation was stopped with addition of Glycine at a final concentration of 125 mM for 5 min. Cells were rinsed twice with cold PBS. Cell nuclei were isolated by incubation for 5 min on ice, with 20 mM Hepes at pH 7.8, 10 mM KCl, 0.25% TritonX100, 1 mM EDTA, 0.5 mM EGTA and proteases inhibitors. After centrifugation, nuclei were resuspended in 10 mM Tris at pH 8, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, and protease inhibitors and incubated for 10 min on ice. Nuclei were lysed and the chromatin was extracted with a lysis buffer (10 mM Tris at pH 8, 140 mM NaCl, 0.1% SDS, 0.5% TritonX100, 0.05% NaDoc, 1 mM EDTA, 0.5 mM EGTA and protease inhibitors) at a final concentration of 10^7 cells per ml. Chromatin was sheared with a Vibra-Cell™ (bioblock) sonicator in 2 ml tubes floating in melting ice. A complete fragmentation of genomic DNA (fragments bellow 600 base pair) was obtained after 5 series of 3 min pulses (5 s ON, 1 s OFF). An aliquot of sheared chromatin was decrosslinked and deproteinized for quality control before immunoprecipitation. E4F1 and IgG (use for QPCR controls) ChIP are carried out within 3 tubes for each antibody containing 2 μl of an affinity-purified rabbit anti-E4F1 polyclonal antibody [3] incubated in the presence of 2 ml of MEF chromatin and 20 μl of Dynabeads protein G. After over-night incubation, immunoprecipitates were successively washed out with 1.5 ml of the 5 following buffers (W1: Tris at pH 8 10 mM, KCl 150 mM, NP40 0.5%, EDTA 1 mM, W2: Tris at pH 8 10 mM, NaCl 100 mM, NaDoc 0.1%, TritonX100 0.5%, W3a: Tris pH 8 10 mM, NaCl 400 mM, NaDoc 0.1%, TritonX100 0.5%, W3b: Tris at pH 8 10 mM, NaCl 500 mM, NaDoc 0.1%, TritonX100 0.5%, W4: Tris pH 8 10 mM, LiCl 250 mM, NaDoc 0.5%, NP40 0.5%, EDTA 1 mM, W5: Tris at pH 8 10 mM, EDTA 1 mM). IPed DNAs were eluted from beads with 100 μl TE + 1% SDS. 50 μl of input chromatin is diluted in 50 μl of TE + 2% SDS. Samples were decrosslinked over-night at 65 °C. They were diluted with 100 μl TE, 50 mM NaCl and 4 μg of RNAseA and incubated for 45 min at 37 °C and deproteinized with proteinase K (4 μg, 55 °C, 45 min). Proteins are removed with phenol–chloroform–isoamylic-alcohol and DNAs are recovered by chromatography with nucleospin extract II columns (Macherey-Nagel). DNA concentration was then determined with Qubit.

2.3. ChIP library, sequencing and base calling

Library preparation and deep-sequencing were performed according to the Illumina protocols at the MGX—Montpellier GenomiX facility (www.mgx.cnrs.fr/).

Libraries of 200 bp ± 25 DNA fragments were prepared with a "ChIP-seq sample prep" kit from Illumina and quality checked on Agilent DNA1000 chips. Flow cells were loaded following the Illumina instructions.

Sequencing runs were done on a HiSeq 2000 sequencer. We generated 50 base pair reads for input and E4F1 ChIP from primary MEF DNA and 36 base pairs for input and E4F1 ChIP from transformed MEFs. HiSeq Control Software and RTA Illumina software were respectively used for image acquisition and base calling.

2.4. Mapping

ChIP-seq raw and processed files are fully available at GEO DataSet repository under the subseries # GSE57228. FASTQ files can be recovered from the NCBI SRA database under the reference #SRP041609. FASTQ files were aligned over mouse genome (mm9) with a maximum of two mismatches in the first 32 nucleotides with CASAVA 1.8 and fulfil the quality control standards performed by the MGX facility. Multiple mapping reads were discarded from further analyses. Raw files were delivered as BED files.

2.5. ChIP-seq data analysis

Read density exploration with IGB confirms E4F1 enrichment in ChIP-seq runs versus inputs. However the first's peak calling lists obtained with MACS [11] or Cisgenome [5] were not fully relevant for further functional analyses or motif discovery. With MACS, transformed MEF E4F1 ChIP-seq gives 293 peaks but at least 50% are false positive peaks. MACS also failed to detect any significant peaks in primary MEF E4F1 ChIP-seq. In order to bypass these problems, we decided to develop our own pipeline to strengthen ChIP peak list accuracy. The different steps are described below.

3. Preprocessing filters

A global overview with IGB of the raw read distributions allows us to define the shape of potent real and fake peaks. Snapshots are displayed as Supplementary Fig. S2 in [10]. Despite completing the MGX platform QC controls, our samples contain a large fraction of non-unique reads (pile-up) for a given position on the genome that could be responsible for false positive peak detection by peak caller algorithms. The first filter can easily be achieved in Galaxy and consist of reducing the occurrence of each tag to 1. The Statistics tool "count occurrence of each record" is run on BED files containing raw reads. Read occurrence is given by the first column of "count" function output. BED are recovered by removing this first column with the "cut" function in the Text Manipulation tools directory. This can also be achieved in R [9] with the command line #1 (sup data).

Our raw files undergo a second filtering by the removal of Chromosomes M and Y mapped reads which display unusual density that could bias peak calling. A manual survey of read density over these two chromosomes did not reveal any significant enrichment that could have been missed because of this filter. Removal is done with the function "filter" (filter and sort tools directory) and "not c1==‘chrY’ and not c1==‘chrM’" arguments. This could also be done in batch under a Linux console with the command lines #2 (sup data).

The third step of filtering is applied on our raw data. Read tag density background is not random, neither equally distributed over mouse genome and samples. Nevertheless these fluctuations are found both on input and ChIP samples on different chromatins that we analyzed, and nearly at the same position on the genome indicating that they are not relevant for our experiments. Because these fluctuations do not present the same amplitude on each sample, the peak calling algorithms that we tested, like MACS [11], Cisgenome [5] or QESEQ [7] can call these regions as significantly enriched regions and so contaminate the final lists with false positives. Removal of these potentially false positive regions on input and ChIP-seq samples before peak calling greatly reduces false positive peak calling and greatly increases functional interpretation of the ChIP-seq experiments.

We first run "hts_windowsummaryv2" with a 100 bp exploration window to get the metrics of our samples and defined a cutoff for calling these artifactual peaks. We performed a peak calling on inputs alone with the "one sample analysis" module of Cisgenome (hts_peakdetectorv2 -i C:\...\XX_.bar -d C:\....\OSXX -o C:\...\XX__1002510 -w 100 -s 25 -c 10) in order to identify these regions with high density reads. We generated a mask constituted of genomic regions called in primary and transformed MEF inputs. These regions were extended by 1000 base pairs in 5′ and 3′ in order to cover the edges of these artifactual regions. We used this 'mask' to remove these regions in input and in ChIP-seq files with Galaxy's "substract" function in operate on genomic interval tool directory. Finally, we did a random sampling of 10 million tags on each sample before peak

calling. Even if most of the peak callers perform a normalization between the IP and control samples to handle imbalanced read numbers between samples, our preliminary ChIP-seq data analyses reveals a higher rate of high confidence peaks when we analyzed 10 million reads per samples. This step can be done in Galaxy with "select random lines" in text manipulation tools or in R with the command lines #3 (sup data).

4. Peak calling

Once data had been pre-filtered (Chromosomes Y and M, pile-up, artifactual region removal, random sampling) we applied a combined peak calling strategy by processing E4F1–ChIP and control input in parallel with two independent programs, Cisgenome and QESEQ. For Cisgenome we used the Seqpeak tool with the following arguments (Seqpeak -b 50 -w 1 -e 150 -ts 1 -c 3 -maxgap 50 -minlen 100 -br 1 -bar 1 -dat 1 - bw 5 -lpois 1 -lpwin 10000 -lpcut 1e-5). Cisgenome generates BAR files for input and ChIP samples which are binary files useful for visualization of read densities in a genome browser like IGB. QESEQ is run as follows QESEQ -s 150 -v 1 -p 0.01 -c 20. Both algorithm outputs consist of a list of enriched regions with genomic coordinates and associated statics. These tables are provided as COD files for Cisgenome output and TXT files for QESEQ output. These two lists are transformed as BED files and are intersected to retain only high confidence peaks. The intersection files have been generated with Galaxy (intersect) and are supplied as "_intersection.bed" files. All processed files (BAR, BED, COD and TXT files) are fully available in GEO DataSet (GSM1377538 and GSM1377540) and as supplementary data in [10]. Tag densities and significantly enriched regions are visualized with IGB to ensure that our filters and cut-off give a good positive signal/background ratio. These lists have been then used for further functional annotation. ChIP–QPCR validations had been performed on a set of representative peaks from the top to the bottom of the list to confirm or not the accuracy of our analysis pipeline.

4.1. Functional annotation

BED files containing E4F1 bound regions were annotated for the closest gene transcription start site (TSS) nearby a peak with four independent tools (Cisgenome, ChIPPeakanno (R), Nebula (Galaxy server) and Seqminer).

4.2. Sequence homology analysis

Peak DNA sequences have been fetched from BED files containing E4F1 bound region coordinates and analyzed with MEME logo suite (http://meme.nbcr.net/meme/).

5. Discussion

We described here an original approach for the analysis of the first E4F1 ChIP-seq. Our first attempt to analyze our ChIP-seq data with MACS or Cisgenome alone failed to generate a high confidence list of genomic regions where E4F1 interacts with DNA. However, since peaks were clearly and easily detectable when exploring read density over the mouse genome, we decided to develop our own analysis pipeline that allowed the detection of these peaks with a low rate of false positive.

In brief, our workflow was as follows: 1/ removal of the read pile-up and Chromosomes M and Y; 2/ we refined our raw data by removing genomic regions of high artifactual read density found by analyzing inputs alone. Finally, using a combined approach of two peak callers (Cisgenome and QESEQ) on 10 million reads per samples, we defined high confidence lists of genomic regions where E4F1 actually binds DNA. This binding was confirmed by ChIP QPCR validation experiments and by the discovery of a hitherto undescribed DNA motif in most peaks, the latter showing a high affinity for recombinant E4F1 protein when tested *in vitro*. Moreover this motif is conserved over species as we were able to ChIP the orthologous regions in human cells based on the motif conservation.

Functional annotations and intersection of these ChIP-seq lists with the lists of genes that are differentially expressed between wild type vs E4F1$^{-/-}$ cells (not described in this article), reveal that E4F1 directly controls a limited set of genes important for cell cycle progression, genomic integrity, mitochondrial homeostasis and energy metabolism.

Acknowledgments

This work was supported by grants from the French Ligue Nationale Contre le Cancer (LNCC, C.S. Equipe Labellisée 2011), from the Agence Nationale pour la Recherche (ANR SVSE2-YinE4F1Yang2 and MetaboCycle 2 to C.S. and L.L.C.), and from the Association pour la Recherche contre le Cancer (to G.R.). Institutional support was provided by the Institut National de la Santé et de la Recherche Médicale (L.L.C.) and the Centre National de la Recherche Scientifique (C.S.), and technical support by the Montpellier Genomics (MGX) and animal facilities. T.H. was supported by fellowships from the LNCC, and O.K. by fellowships financed on the ANR grants JCJC-0014-01 and SVSE2-YinE4F1Yang2.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.07.004.

References

[1] L. Le Cam, M. Lacroix, M.A. Ciemerych, C. Sardet, P. Sicinski, The E4F protein is required for mitotic progression during embryonic cell cycles. Mol. Cell. Biol. 24 (2004) 6467–6475.
[2] L. Le Cam, L.K. Linares, C. Paul, E. Julien, M. Lacroix, E. Hatchi, R. Triboulet, G. Bossis, A. Shmueli, M.S. Rodriguez, et al., E4F1 is an atypical ubiquitin ligase that modulates p53 effector functions independently of degradation. Cell 127 (2006) 775–788.
[3] L. Fajas, C. Paul, O. Zugasti, L. Le Cam, J. Polanowska, E. Fabbrizio, R. Medema, M.L. Vignais, C. Sardet, pRB binds to and modulates the transrepressing activity of the E1A-regulated transcription factor p120E4F. Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 7738–7743.
[4] E. Hatchi, G. Rodier, M. Lacroix, J. Caramel, O. Kirsh, C. Jacquet, E. Schrepfer, S. Lagarrigue, L.K. Linares, G. Lledo, et al., E4F1 deficiency results in oxidative stress-mediated cell death of leukemic cells. J. Exp. Med. 208 (2011) 1403–1417.
[5] H. Ji, H. Jiang, W. Ma, D.S. Johnson, R.M. Myers, W.H. Wong, An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat. Biotechnol. 26 (2008) 1293–1300.
[6] M. Lacroix, J. Caramel, P. Goguet-Rubio, L.K. Linares, S. Estrach, E. Hatchi, G. Rodier, G. Lledo, C. de Bettignies, A. Thépot, et al., Transcription factor E4F1 is essential for epidermal stem cell maintenance and skin homeostasis. Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 21076–21081.
[7] M. Micsinai, F. Parisi, F. Strino, P. Asp, B.D. Dynlacht, Y. Kluger, Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. Nucleic Acids Res. 40 (2012) e70.
[8] P. Raychaudhuri, R. Rooney, J.R. Nevins, Identification of an E1A-inducible cellular factor that interacts with regulatory sequences within the adenovirus E4 promoter. EMBO J. 6 (1987) 4073–4081.
[9] R Development Core Team, R: a language and environment for statistical computing. http://www.R-Project.org 2008.
[10] G. Rodier, O. Kirsh, M. Baraibar, T. Houlès, M. Lacroix, H. Delpech, E. Hatchi, S. Arnould, D. Severac, E. Dubois, et al., The transcription factor E4F1 coordinates CHK1-dependent checkpoint and mitochondrial functions. Cell Rep. 11 (2015) 220–233.
[11] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li, et al., Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9 (2008) R137.