Chapter 9

HAT: A Novel Statistical Approach to Discover Functional Regions in the Genome

Erdogan Taskesen, Bas Wouters, and Ruud Delwel

Abstract

Tiling arrays are useful for exploring local functions of regions of the genome in an unbiased fashion. The exact determination of those genomic regions based on tiling-array data, e.g., generated by means of hybridization with immunopreciptated DNA-fragments to the arrays is a challenge. Many different statistical methodologies have been developed to find biological relevant regions-of-interest (ROI) by using the quantitative signal intensity of each probe. We previously developed a method called Hypergeometric Analysis of Tiling arrays (HAT) for the analysis of tiling-array data, but it is developed such that it can also be used to study data derived by genome-wide deep sequencing approaches. Here we applied HAT to analyze two publicly available tiling-array data sets. After the detection of statistically significant ROI, these are often used in additional analysis for hypothesis testing. We therefore discuss, by using the results of the tiling-array experiment, pathway and motif analyses.

Key words HAT, Tiling array, Peak calling, Motif analysis, CEBPA

1 Introduction

Tiling arrays are a subtype of microarrays which are designed with probes that cover contiguous regions of a genome. The locations of probes do not necessarily cover genomic regions that are known to be functional, as is the case for gene expression or promoter arrays. Therefore tiling arrays differ from these microarrays as they are not by definition designed to cover known or predicted genes in the genome. Moreover, the coverage of probes in unknown genomic regions has been useful for exploring the genome in an unbiased fashion. Examples of applications for tiling arrays are (1) protein–DNA interaction by conducting chromatin immunoprecipitation (ChIP-on-chip) experiments [1], (2) epigenetic modifications by Methyl-DNA immunoprecipitation [2] (MeDIP-on-chip), or (3) identification of DNAse hypersensitive sites, which can be used to predict regulatory elements such as promoter regions, enhancers, and silencers [3]. Although tiling arrays are

Tin-Lap Lee and Alfred Chun Shui Luk (eds.), *Tiling Arrays: Methods and Protocols*, Methods in Molecular Biology, vol. 1067,
DOI 10.1007/978-1-62703-607-8_9,

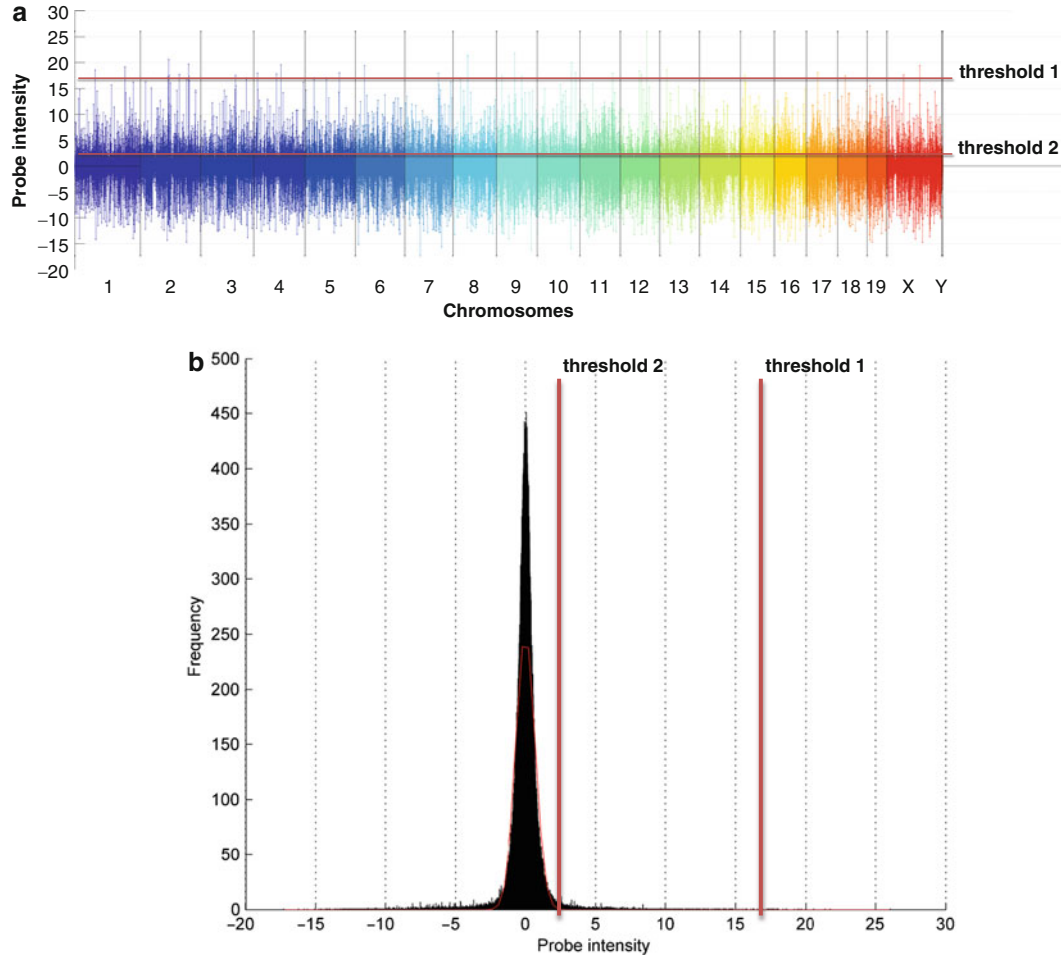

Fig. 1 Graphical representation of probe intensities in a ChIP-on-chip tiling-array experiment. (**a**) Normalized probe intensity of 4.6 million probes among 22 chromosomes. Colors illustrate the different chromosomes whereas the length of a lollipop represents the probe intensity. (**b**) Distribution of the probe-intensity values. The probe-intensity values are normalized against a reference file. Threshold 1 indicates a high threshold cutoff, whereas threshold 2 indicates a low intensity cutoff. HAT uses many different threshold cutoffs to determine significantly enriched ROI

useful for genome-wide studies, the coverage of the genome on the arrays depends on the species that is being studied. As an example, probes can cover the majority of a small genome such as for Arabidopsis [4], whereas probes will cover only contigs in a large genome, such as for human. Thus for larger genomes, as is the case for mouse or humans, the choice of the content depends on the questions one wishes to address using a particular tiling array.

Each tiling-array produces quantitative signal intensity for each probe by the hybridization of labeled DNA. Normalized probe intensities are illustrated by the different peaks in Fig. 1, where the colors indicate the probe signals at different chromosomes.

Although single probe hybridization with high signal intensity suggests strong hybridization, it is not necessarily the result of specific hybridization of labeled DNA (illustrated by the probes above threshold 1 in Fig. 1a, b). Multiple contiguous probes that show increased signal intensity upon hybridization across a particular genomic region are more likely to be the result of true hybridization in a biological experiment. These genomic regions are denoted as a putative region-of-interest (ROI). In order to find such ROI, a low threshold must be employed which may compromise the results by introducing false-positives ROI (Fig. 1a, b, threshold 2). To detect biological relevant ROI, probe-intensity signals should be discriminated from nonspecific signals. A challenge in the analysis of tiling-array data is the detection of true ROI, and to minimize the number of false positives. A straightforward approach is to choose a fixed number of consecutive probes above a certain threshold and indicate it as an ROI. Nevertheless, this definition of ROI may be inadequate because of the required number of consecutive probes and the optimal threshold may be difficult to establish. In addition, the probe resolution varies across the genome, and across different tiling-array platforms.

Multiple methods have been developed to analyze tiling-array data which all serve one goal, i.e., the detection of true ROI and thereby discriminating positive-probe intensity from the background. The developed methods differ in their statistical approaches: methods incorporate the hypergeometric distribution [5], hidden Markov models [6–8], correlation structures [9], heuristics [10], mixture models [11], Bayesian modeling [12, 13], wavelets [14], or by using other methodologies [15–22]. All methods have shown to be useful in filtering large data sets for candidate gene discovery. It is of importance to note that biological experiments are always a necessity to validate particular findings.

Here we discuss the previously developed method, Hypergeometric Analysis of Tiling arrays (HAT) [5], that uses the hypergeometric distribution to assess the probability of a consecutive number of probes in a particular genomic region while controlling multiple testing (Family Wise Error: FWER). Furthermore, HAT uses multiple threshold cutoffs, it does not necessarily require experimental replicates, and can be normalized against reference files. It furthermore employs a single user defined parameter: the significance level alpha. Note that alpha is not used to determine the threshold cutoff using the data distribution (Fig. 1b), instead it computes the probability to observe a specific number of probes for a particular genomic region (window) over multiple threshold cutoffs. Furthermore, specifying parameters such as fragment-size may improve the detection of ROI, whereas parameters for gene-mapping and sequence-of-interest are required for additional analysis (Fig. 2). HAT is generically built and therefore independent of probe-intensity distribution, probe-sets coverage

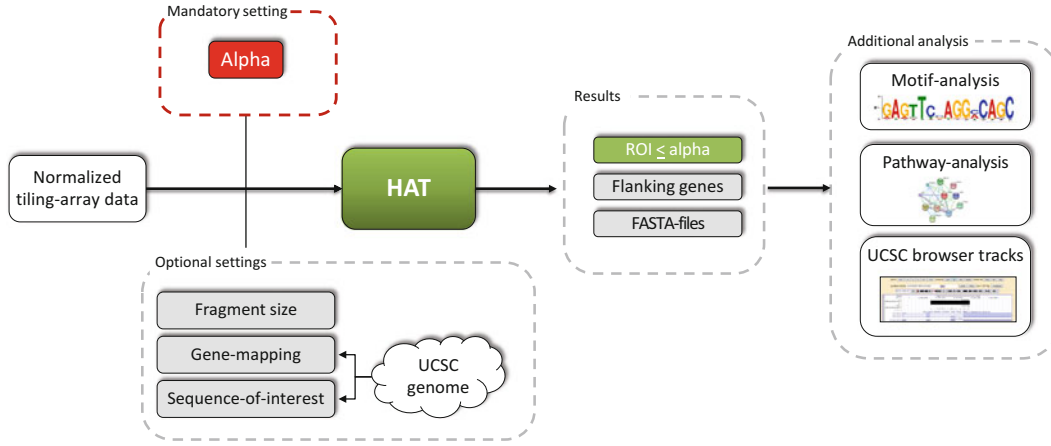

Fig. 2 Schematic overview of tilling-array data analysis. Stepwise illustration of normalized tiling-array data towards the detection of significantly enriched ROI, the flanking genes, sequences files (FASTA), motif analysis, pathway analysis, and the UCSC-browser

and probe-sets resolution across the genome and tiling-array platform. It is successfully applied in multiple types of biological research questions, i.e., the detection of protein–DNA interactions (ChIP-on-chip [5]), identification of genomic locations that are involved in viral integration and potentially harbor tumor suppressor genes (MeDIP-on-chip) [2], the identification of regions enriched for histone modifications such as trimethylation of histone 3 at lysine 4 or lysine 27 (H3K4 me3, H3K27 me3) [5], and for the identification of anthocyanin-specific genes that flank enriched genomic DNA in black rice using 3′-TILLING 135 K *Oryza sativa* microarray [23]. Many detected ROI among these different studies were confirmed by quantitative polymerase chain reaction (qPCR).

Although tiling arrays have been applied successfully for genome-wide applications, high-throughput sequencing of for instance chromatin immunoprecipitated DNA-fragments (ChIP-Seq) show genome-wide associations in higher resolutions and will therefore be superior to chip technology. Even though ChIP-Seq is becoming the standard for genome-wide applications, numerous high-quality tiling-array data sets are publicly available at the gene expression omnibus website (GEO: http://www.ncbi.nlm.nih.gov/geo/). These can be of value to address particular research questions raised by investigators and to which HAT may be very useful. Furthermore, although HAT was initially developed for the analysis of tiling-array data, the application is not limited to the studies discussed in this chapter, but can be applied for the analysis of ChIP-Seq data as well.

Here we stepwise discuss how to apply HAT to analyze tiling-array data. As case examples we used two publicly available ChIP-on-chip data sets. In addition we discuss two types of analyses that frequently follow upon the detection of ROI, namely motif and pathway analyses.

2 Materials

We previously reported the successful usage of HAT on two novel data sets [5]. Here we demonstrate HAT on previously reported STAT4-chromatin immunoprecipitation (ChIP-on-chip) experiments ($n=2$), compared to controls ($n=2$). Secondly, we use HAT to analyze the DNA-binding capacity of a C-terminal mutant *CEBPA* ($n=2$), compared to controls (ER) ($n=2$). Both data sets are available on the gene expression omnibus (GEO), GSE19321 and GSE16845, respectively. Data were generated using the Affymetrix GeneChip Mouse Promoter 1.0 Array. This chip generates 4.6 million perfect match probes over 28,000 mouse promoter regions. Promoter regions cover 6 kb upstream to 2.5 kb downstream of 5′ transcription start sites. Each probe has a size of 25 nucleotides (nt). RAW probe-intensity values are normalized by utilizing Model-based analysis of tiling arrays for ChIP-chip (MAT) [22, 24].

3 Analyzing Tiling-Array Data Sets

In this paragraph we demonstrate the usage of HAT for the identification of significant ROI and define the parameters for ChIP-on-chip experiments. Before starting the peak-detection algorithm (HAT), pre-knowledge about the experimental setup is highly recommended. The experimental protocol requires shearing of the DNA by using a sonication process which results in DNA-fragments of approximately 600 base pairs (bp). Subsequently, chromatin fragments are *immunoprecipitated* using antibodies directed to the protein of interest, known to interact with DNA. The consecutive probes can, therefore, cover up to 600 bp after the hybridization process per fragment. This information can be used in the model for the detection of ROI. Note that significant ROI can be detected that are larger or smaller in width than 600 bp. In addition, we set the significance level on 0.05.

The first ChIP-on-chip data set to which we applied HAT is a study that was previously reported and in which STAT4-mediated transcriptional regulatory networks in Th1 cell development were investigated [1]. STAT4 is a critical component in the development of inflammatory adaptive immune responses. Although STAT4 was subject in various other studies [25, 26], it was claimed that the genetic program, activated by STAT4 that results in an inflammatory cell type, is not well characterized. A ChIP-on-chip experiment was therefore conducted as previously reported [1]. Here, we analyzed both experimental replicates by choosing a fragment-size of 600 nt and α: 0.05, and detected $n=2{,}903$ and $n=3{,}106$ ROI. Moreover, 84 % ($n=2{,}499$) overlapped in both replicates compared to the controls (sized between 215 and

Table 1
Motif enrichment analysis on the detected regions-of-interest in the STAT4 experiment

Transcription factor	Recognized factors	Fold-increase	*P*-value
V$STAT1_01	STAT1, STAT1alpha,STAT1beta	8,0588	5,5308E–29
V$STAT5B_01	STAT5A, STAT5B	4,0922	1,052E–26
V$STAT1_05	STAT1	5,6114	2,1506E–26
V$STAT_01	STAT1, STAT1alpha, STAT1beta,STAT2, STAT3, STAT3-isoform1,STAT4, STAT5A, STAT5B, STAT6	3,5424	5,4465E–22
V$STAT3_01	STAT3, STAT3-isoform1	5,9629	1,3443E–19
V$STAT1STAT1_Q3	CBF3, STAT1:STAT1, ehf	4,0458	1,1127E–18
V$IRF_Q6	IRF-10, IRF-2, IRF-3, IRF-4, IRF-5, IRF-6, IRF-7, IRF-7A, IRF-7B, IRF-7H, IRF-8, IRF4-1, irf1	3,603	1,0817E–11
V$AP1 _Q6_01	AP-1, FOSB, FosB, Fra-1, Fra-2, JunB, JunB:Fra-1, JunB:Fra-2, JunD, JunD:Fra-2, JunD:deltaFosB, c-Fos, c-Jun, c-Jun:FosB, c-Jun:JunD, c-Jun:c-Fos, deltaFosB	2,7031	2,0413E–11
V$STAT5A_01	STAT5A	4,0795	3,8123E–11
V$BACH1_01	Bach1, Bach1t	3,1209	3,3523E–09

The top ten enriched TFBS among the detected binding regions using HAT for the STAT4 study (ChIP-on-chip). A TFBS is called when the position weight matrices (PWM) is enriched at $P \leq 0.001$. Recognized factors: the transcription factors that are recognized by the TFBS. Fold-increase: the frequency that a TFBS is detected among the binding regions compared to the reference set (5,000 randomly chosen genes)

4543 nt, median: 1002 nt). It was previously demonstrated that the analysis method, GenPathway, identified 4,669 genes that were seen in both replicates [1]. This list is subsequently filtered for genes with binding intensity >4 and thereby resulted in 1,540 genes. This indicates that using the unfiltered list, GenPathway detects almost twice the number of ROI when compared to HAT. To investigate the validity of the ROI that were detected by HAT, a motif enrichment analysis was conducted on the 2,499 common ROI by using F-MATCH [27, 28]. We hypothesize that the detected ROI should contain a STAT-binding site. We detected a total of 38 transcription factor binding sites (TFBS) of which the STAT-motifs were highly enriched ($P < 0.001$). Moreover, 7 STAT-motifs were detected in the top 10 after ranking the TFBS on significance (Table 1). This suggests high specificity of the detected ROI. Note that the STAT-motif is also highly enriched in the genes detected by GenPathway [1]. Although both methods detected high enrichment for the STAT-motifs, the overlap of genes between both methods was 897 genes. In other words, 1,211 genes were

Table 2
Motif enrichment analysis on the detected regions-of-interest that are exclusively detected using HAT in the STAT4 experiment

Transcription factor	Recognized factors	Fold-increase	*P*-value
V$STAT1_01	STAT1, STAT1alpha, STAT1beta, STAT2, STAT3, STAT3-isoform1, STAT4, STAT5A, STAT5B, STAT6	7,235	4,96E–16
V$STAT3_01	STAT3, STAT3-isoform1	5,862	7,36E–12
V$STAT5B_01	STAT5A, STAT5B	3,6333	2,16E–11
V$STAT1_05	STAT1	4,6577	2,45E–09
V$STAT_01	STAT1, STAT1alpha, STAT1beta	3,4047	4,05E–09
V$GADP_01	GABP	4,4162	1,88E–08
V$SAP1A_01	SAP-1a	4,1057	5,32E–08
V$STAT1STAT1_Q3	CBF3, STAT1:STAT1, ehf	3,5072	5,26E–07
V$ELK1_02	Elk-1, Elk1-isoform1	4,0252	7,68E–07
V$CETS1P54_01	Ets-1, Ets-1 deltaVII, c-Ets-1, c-Ets-1 54, c-Ets-1A, c-Ets-1B	3,9813	8,78E–07

The top ten enriched TFBS among the exclusively detected binding regions of HAT for the STAT4 study (ChIP-on-chip). A TFBS is called when the position weight matrices (PWM) is enriched at $P \leq 0.001$. Recognized factors: the transcription factors that are recognized by the TFBS. Fold-increase: the frequency that a TFBS is detected among the binding regions compared to the reference set (5,000 randomly chosen genes)

solely detected by HAT and not by GenPathway. To assess the validity of these ROI, we conducted a motif analysis for only those 1,211 ROI and detected again high enrichment for the STAT-motifs, i.e., 6 STAT-TFBS are detected in the top ten ranked list (Table 2). We hypothesize that these 1,211 genes may be present in the initial 4,669 genes detected by GenPathway, but are excluded from the list as these did not comply the above mentioned criteria. This is supported by the notion that significantly lower probe-intensity levels are observed ($P < 0.0001$) in the 1,211 ROI compared to the 897 ROI. Note that the probe-intensity levels, of all the detected ROI, are significantly higher compared to the background. Unfortunately, we were not able to analyze the motifs among the exclusively detected genes by GenPathway, as the exact genomic positions of the ROI were not specified. These differences may occur due to alternatively defined gene-mapping procedures (Fig. 3) and the differences in statistical methodologies. In conclusion, we identified another set of genes that were highly enriched for the STAT-motif.

The second ChIP-on-chip data set is used to study the DNA-binding capacity of a variant of CCAAT enhancer binding protein

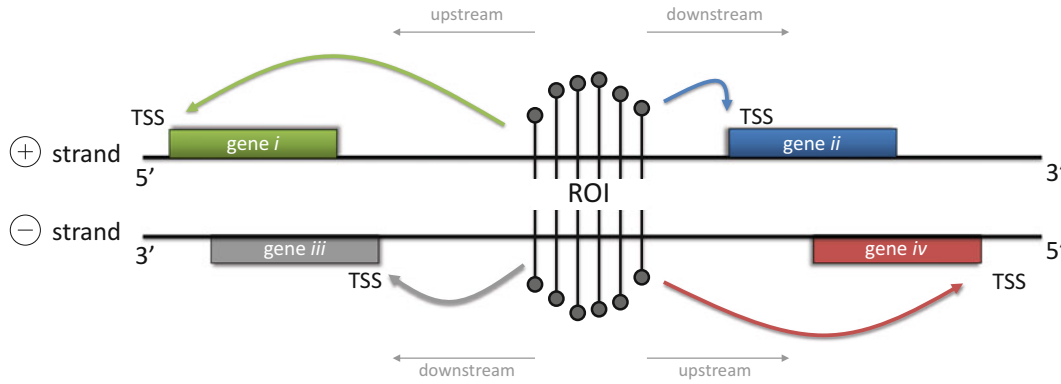

Fig. 3 Mapping of detected regions-of-interest to genes located in close vicinity. A single ROI is illustrated with four neighboring genes: two on the positive strand (upstream and downstream) and two on the negative strand (upstream and downstream). Mapping of ROI to genes is crucial for additional analysis (e.g., pathway analysis). ROI: region-of-interest, TSS: transcriptional start site, 3′: three prime UTR, 5′: five prime UTR

alpha (*CEBPA*) that carries a C-terminal mutation. *CEBPA* is a transcription factor and master regulator of myeloid differentiation [29, 30]. It is frequently mutated in patients with acute myeloid leukemia (AML) (5–14 %) [31]. Abnormalities in *CEBPA* may contribute to a block in differentiation of progenitor cells of granulocytes, which can result in leukemogenesis. Mutations in *CEBPA* are associated with a particular prognosis of patients with AML [31]. In AML patients, two types of *CEBPA* mutations are known to exist: mutations in the N-terminus and the C-terminus. C-terminal mutations are found in the DNA-binding domain. Since the mutant protein can still interact with other proteins that may interact with DNA, we propose that mutant *CEBPA* may indirectly interact with DNA. We wondered to which loci mutant *CEBPA* might interact in an indirect manner. We created a similar C-terminal mutation as found in one particular human AML patient [32], with an insertion of six amino acids in the C-terminal bZIP domain. We used it in the ChIP-on-chip experiment to identify genes that may play a role in leukemogenesis. Promoter array hybridizations were conducted from a myeloid cell line model (32D) that expresses either beta-estradiol inducible C-terminal mutant *CEBPA* (2 clones) or control-ER (2 clones). The question that we wished to address is whether mutated *CEBPA* can bind to the DNA, thereby identifying the associated genes. Using a fragment-size of 600 bp and an alpha of 0.05, we detected in total $n = 89$ and $n = 109$ significant binding regions in the two clones with C-terminal mutant *CEBPA* that was not seen in the controls (Fig. 4). The ROI are sized between 154 and 2,481 nucleotides (median 717 nt) and forty-eight were commonly detected in both clones.

We next searched for binding-motifs among the detected ROI of the C-terminal mutant *CEBPA*. Although it is known that the

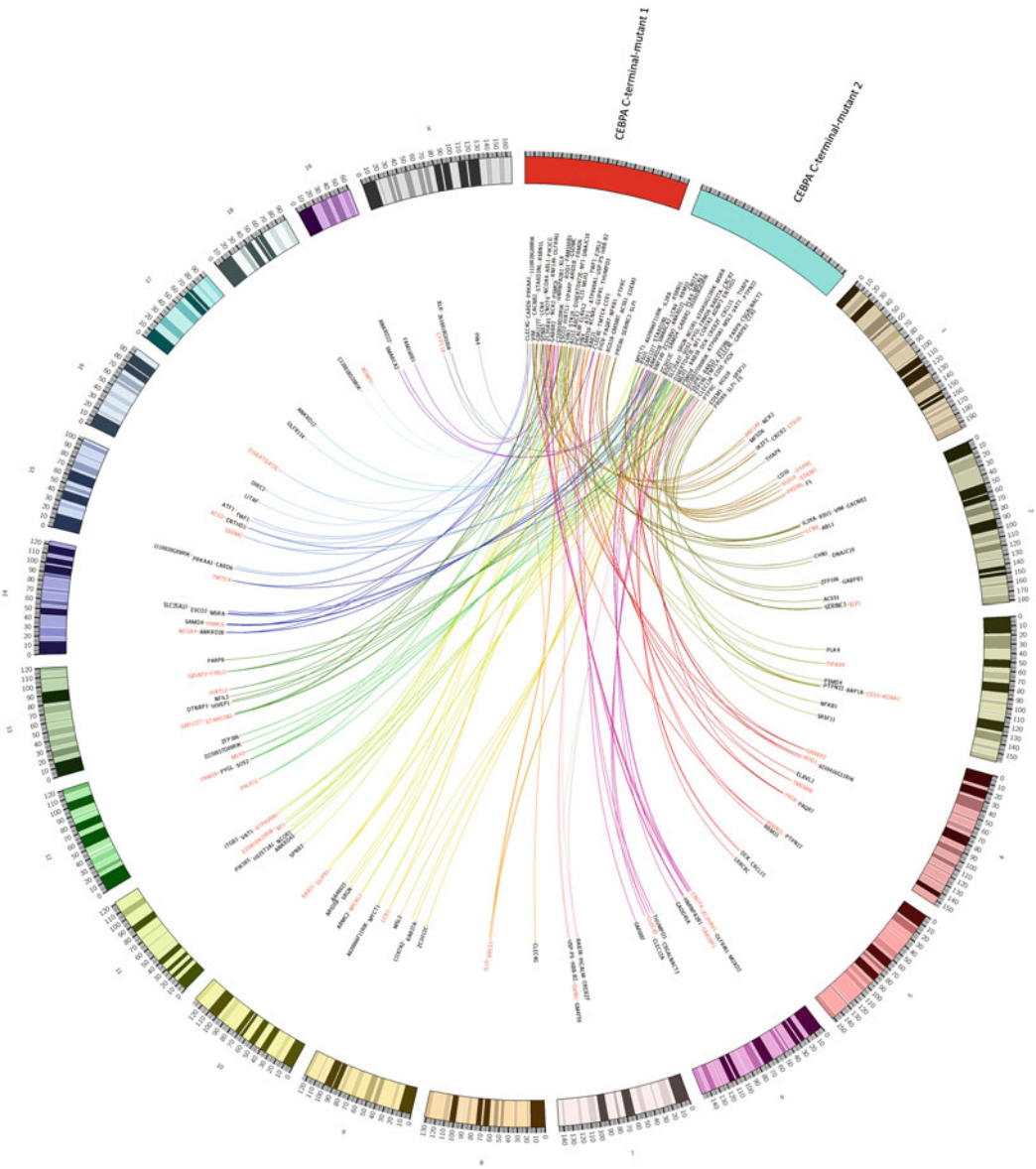

Fig. 4 Graphical representation of the genes that are bound by the C-terminal mutant *CEBPA*. One hundred and forty mapped genes from the detected ROI of the C-terminal mutant *CEBPA* experiments are illustrated. Candidate genes in experiment 1 are indicated by the *red box*, whereas the candidate genes from experiment 2 are indicated by the *blue box*. Forty-six genes (mapped from 48 ROI) that overlap between experiment 1 and 2 are indicated with a *red text-color*. *Line-colors* are colored similar as the chromosomes which are numbered from 1 to 19 and X, and show the relative location of the genes using mouse genome-build 8 (mm8)

C-terminal mutant *CEBPA* lacks binding capacity, we identified three enriched motifs namely, core-binding factor (CBF), ETS, and ESE-1 (P-value < 0.001). Core-binding factors have been shown to fulfil an important role in hematopoiesis [33] and ETS family members, such as ESE-1, fulfil an important role in several

Table 3
Motif enrichment analysis on the 2 kb upstream regions-of-interest of the C-terminal mutant *CEBPA* genes

Transcription factor	Recognized factors	Fold-increase	*P*-value
V$POU3F2_02	POU3F2, POU3F2 (N-Oct-5a), POU3F2 (N-Oct-5b)	1,9479	6,8068E–12
V$CDP_02	CDP, CDP-isoform1, CDP2	2,0723	9,8643E–09
V$FOXP3_Q4	FOXP3	4,753	5,9801E–08
V$OCT1_01	Oct-1, POU2F1, POU2F1a	1,7131	6,2367E–08
V$IPF1_Q6	PDX1, ipf1	1,6051	1,3651E–07
V$CLOX_01	Cutl	1,7556	8,1928E–07
V$SATB1_01	CBF-C	1,5865	8,4804E–07
V$OTX_Q1	Otx1, Otx2	1,9029	9,0873E–07
V$HMGIY_Q3	HMGI-C, HMGIY, HMGIY-isoform1, HMGIY-isoform2	1,546	1,7164E–06
V$FOXO1_01	FOXO1A	1,7142	1,9312E–06
V$DMRT4_01	DMRT4	1,581	2,2881E–06
V$NFAT_Q6	NF-AT, NF-AT1, NF-AT1C, NF-AT2, NF-AT3, NF-AT4, NFAT1, NFAT1-is&$$$;	3,009	2,7411E–06
V$TEF_Q6	TEF-xbb1, Thyrotroph embryonic factor, Thyrotroph embryonic factor&$$$;	1,9549	5,1394E–06
V$SRF_C	SRF, SRF-I, SRF-L, SRF-M, SRF-S	2,1108	5,9055E–06
V$CEBPGAMMA_Q6	C/EBPgamma	1,8284	6,5027E–06

The top 15 enriched TFBS among the 2 kb upstream genes of the C-terminal mutant ROI. A TFBS is called when the position weight matrices (PWM) is enriched at $P \leq 0.001$ and with fold-increase >1.5. Recognized factors: the transcription factors that are recognized by the TFBS. Fold-increase: the frequency that a TFBS is detected among the binding regions compared to the reference set (5,000 randomly chosen genes)

signal transduction pathways [34–36]. As expected, we did not find the consensus binding motif CEBP as we showed previously for wild-type *CEBPA* using the same model system [5]. The detection of these three enriched motifs and the absence of the CEBP motif suggest that DNA-binding by mutant CEBPA had occurred indirectly. We hypothesized that other factors may influence the DNA-binding capacity and therefore analyzed the 2 kb upstream regions, from the transcriptional start site (TSS) of the detected genes (Fig. 4). This resulted in the detection of 71 enriched TFBS with $P \leq 0.001$ and 1.5 times more frequently observed than in the reference set (fold-increase ≥ 1.5). As a reference set we selected 2 kb upstream sequences (starting from the transcription start site) of 5,000 randomly selected genes. The 2 kb upstream sequences are gathered using the UCSC database (http://hgdownload.cse.ucsc.edu). The top 15 TFBS are depicted in Table 3.

3.1 Detected Regions-of-Interest Can Be Mapped to Genes that Are Located in Close Vicinity

Although the goal is to detect ROI by using ChIP-on-chip tiling arrays, it often requires additional analysis, such as pathway analysis, to test a particular hypothesis. This requires the mapping of ROI to genes. Each ROI can, theoretically, be mapped to four genes that are located on: (1) the positive strand and upstream, (2) the positive strand and downstream, (3) the negative strand and upstream, and (4) the negative strand and downstream (Fig. 3). From these four genes, only one gene may be targeted (or two genes in a bi-directional promoter region). For promoter tiling arrays, where only the promoter regions are present on chip, it is straightforward to map the detected ROI to the nearest located TSS of a gene. To prevent incorrect gene-mapping, due to differences in genomic locations of TSS between species and/or genomic-build (hg18, hg19 for human and mm8, mm9 for mus musculus), it is highly recommended to use the same species and genomic-build for both the gene-mapping file as the one used in the normalization process. These gene-mapping files can be downloaded from the UCSC: http://hgdownload.cse.ucsc.edu.

Manually curating each detected ROI to a particular gene is possible using the UCSC-genome-browser track (generated using HAT, Fig. 2) but can be time-consuming. Alternatively, by specifying the species and genome-build in HAT, each ROI can automatically be mapped to the TSS of a gene in closest vicinity. We specified in both ChIP-on-chip experiments "mm8" because the experimental samples were derived from mus musculus and normalized with genomic-build 8. Because both analyzed data sets have been generated using promoter tiling arrays, it allowed the mapping of the ROI to genes in close vicinity. For the STAT study, the 2,499 detected ROI were mapped to 2,108 unique genes. For the *CEBPA* study, it resulted in the detection of 140 unique genes. These are graphically illustrated using a circos-plot [37] (Fig. 4). Such graphical representation indicates the chromosomal location of the genes, and whether genes are commonly detected in the independent experiments using different clones.

3.2 Motif and Pathway Analyses on the Detected Regions-of-Interest and Their Flanking Genes

Analysis on the detected ROI or the genes that are located in close vicinity of the ROI is an important next step for hypothesis testing. Both motif and pathway analyses are therefore useful in tiling-array studies (Fig. 2).

Motif analysis detects specific sequences involved directly in protein–DNA binding interactions, or alternatively whether the promoter regions of the flanking genes include overrepresented sequences of transcription factors. These so-called TFBS may suggest that the protein-of-interest interacts synergistically with other proteins or is involved in the formation of protein complexes. In general, two types of motif analysis exist: by using known TFBS that are derived from published collections (e.g., JASPAR or TRANSFAC databases). These databases should be used when

seeking specific factors or structural classes. Secondly, dé-novo motif analysis can be used to analyze similarities among the sequences to produce a description for each pattern it discovers. F-MATCH [27, 28] and MEME [38] are two algorithms which can be used for the detection of known TFBS and/or dé-novo motifs. These methods are online accessible and require FASTA-files as an input, which contain sequences of the ROI (generated by HAT).

Besides motif analysis, it can be useful to analyze the detected genes for enriched pathways. Pathway analysis is the process of identifying interactions and associated annotations [39]. For the detected flanking genes it may provide insight how genes are regulated and which processes, functions, or networks were involved. Both commercial and noncommercial entities provide pathway analysis. A commercial tool is Ingenuity Pathway Analysis (Ingenuity® Systems, http://www.ingenuity.com, IPA 8.8). Networks in IPA are created using literature-based records that are maintained in the Ingenuity Pathway Knowledge Base. It computes a network-score for the overlap of the focus genes with a global molecular network. Alternatively, Gene Set Enrichment Analysis (GSEA) [40] provides both software and a collection of annotated gene sets (MSigDB: Molecular Signature Database) that can be used for the detection of pathways and/or gene sets (noncommercial). Depending on the research question, different gene sets can be used: (1) BioCarta pathways, describing the molecular relationships derived from active research areas, (2) KEGG pathways, describing the molecular interactions and reaction networks, (3) Reactome pathways, manually curated and peer-reviewed pathways, (4) GO biological processes, gene sets describing the biological process ontology, (5) Transcription factor targets (TFT), gene sets contain genes that share a TFBS, and (6) MicroRNA targets, Gene sets that contain genes that share a 3′ UTR microRNA binding motif.

4 Notes

Different Methodologies Come to Different Results, What Is the Correct One to Choose?

All previously described methods have been reported to validate some of the detected ROI as described in Subheading 1. Nevertheless, different statistical methodologies lead to differences in the detected ROI. We hypothesize that various methodologies may results in similar detected ROI which are most likely the genomic regions that contain a contiguous number of probes with high probe-intensity levels (the results of two methods are shown in Subheading 3). In addition, the differences between detected ROI among various methodologies are likely the genomic regions with subtle changes in probe-intensity levels. Note that some developed methodologies are designed for the analysis of one type of tiling-array application. Others may require various parameters to set before starting the analysis, e.g., by defining the ROI using

the maximum and/or minimum number of probes in a genomic region, maximum gap size between two probes and threshold. Changing one of the parameters will affect the final results. It is therefore always recommended to perform additional analysis after the detection of ROI to ensure confidence about the gained results. We demonstrated this in Subheading 3, where we detected 1,211 ROI that were exclusively found for HAT. A motif analysis showed significant enrichment for the STAT-consensus binding site. Such findings may help deciding which method to use. It is important to note that in the end laboratory experiments are indispensable to demonstrate the biological significance of particular that ROI, identified by means of tiling-array analysis.

How to Continue If No Significant Regions-of-Interest Are Detected?

The analysis of tiling-array data (Subheading 3) can result in the absence of significantly enriched ROI. This indicates that probe-intensity values, by the hybridization of DNA-fragments on chip, showed no significant differences compared to the background data-file. In case the hybridization process on chip is successfully performed (i.e., DNA-fragments are immunoprecipitated) and the background data-file is correctly provided into the model, it still may result in the absence of significantly enriched ROI. Note that analyzing experimental data-files without the usage or incorrect usage of a background data-file can lead to the absence of significantly enriched ROI or the detection of false-positive ROI. If no significantly enriched ROI are detected, it should be considered that no DNA-binding did take place and therefore no ROI were detected. Alternatively, one could decide to increase the significance level alpha and rerun the analysis. Note that the false-positive rate increases by using alpha >0.05. It is therefore highly recommended to validate the ROI by qPCR. As an example, it is demonstrated that a MeDIP-on-chip experiment resulted in the detection of 15 ROI [2]. These are detected without using a background [2]. Although there was supporting evidence that all 15 ROI may be valid (additional analysis showed that all ROI contained a nearby restriction site), only eight viral integration sites could be validated by directed PCR followed by Sanger sequencing [2]. The remaining seven ROI may therefore be the result of technical variation which may have been prevented by using a background file.

How to Determine the Best Flanking Gene for a Detected Regions-of-Interest?

The usage of tiling-array data does not provide information regarding the strand (positive or negative) or genes affected by the putative promoter. It only indicates the probe-intensity values and their genomic positions. If a particular genomic region is marked as a potential ROI, the responsible immunoprecipitated DNA-fragment is suggested to show binding, e.g., via an immunoprecipitated transcription factor, that could bind to the DNA strand. The ROI is then linked to the gene in close vicinity (Fig. 3). The use of an UCSC-browser track may help manually curating the ROI to

a gene. Alternatively, it requires biological experiments to validate whether the binding had an effect on the regulation of a gene. Note that promoter tiling arrays (as described in Subheading 2) only contain probes of which the genomic locations are in the promoter regions of genes and therefore simplifies the gene-mapping procedure.

How to Run HAT with RAW Cell Files?

HAT is build generically to analyze different applications and platforms of tiling-array data (as described in Subheading 1). On the contrary, normalization may differ between different applications and platforms of tiling arrays, e.g., one-color arrays of Affymetrix versus two-color arrays of Nimblegen. Including a normalization step into the model would therefore limit the model to one type of tiling array. RAW cell files need to be normalized based on the type of tiling array [24], and then used as an input into the model (Fig. 2).

How to Prevent "Out-of-Memory" Problems When Analyzing Tiling-Array Data?

When using HAT, it is recommended to use at least 4GB of RAM memory and Windows-64bits version or UNIX-based system. The methodology is tested on tiling-array data containing 4.6 million perfect match probes, and developed in such a way that it is analyzed per chromosome which reduces high memory loads. Nevertheless, when memory problems occur, it is recommended to kill unused running processes when running HAT. In a Windows environment this can be done in the "task manager"; find the "Run" window in the start-menu en type "taskmgr" and then press "Ok" or press the <ENTER>-key.

How to Install HAT in Windows or an UNIX Environment?

The installation of HAT requires an x86-64 Windows or UNIX-based system and 4GB memory or more is highly recommended. Both platforms require the installation of MATLAB or the MATLAB Compiler Runtime (MCR) which is a standalone set of shared libraries that enable full functioning of HAT. The MCR installer and HAT can be downloaded from http://www.erasmusmc.nl/hematologie/.

Windows x86-64 Platform

1. Download HAT and the MCRinstaller setup-file.
2. Run the setup-file and check the option: "MCR-package for win64."
3. After the installation, the MCR directory (<mcr_root>\<version>) will automatically be included in the PATH. Alternatively the environment variable can be set using the command-prompt (**3.1**) or without the command-prompt (**3.2**).
   3.1. Open a command-prompt and issue the command:

        set PATH=<mcr_root>\<version>\runtime\win64;%PATH%

   3.2. For Windows, add the PATH environment variable as follows.

- Select the My Computer icon on your desktop or in the configuration window.
- Right click the icon and select Properties from the menu.
- Select the Advanced tab.
- Click Environment Variables.
- Click "Edit" for variable "PATH"
- Add the "<mcr_root>\<version>\runtime\win64" to the "PATH" (delimited by semicolons (;)).
- Run HAT by executing "HAT.exe".

UNIX x86-64 Platform

1. Download the HAT and MCRinstaller file.
2. Install the MATLAB Compiler Runtime with the following command.

   ```
   ./MCRInstaller.bin -is:extract
   ./MCRInstaller.bin -console
   ```

   For a noninteractive and non-GUI installation.

   ```
   ./MCRInstaller.bin –P bean421.installLocatio
      n="desiredInstallPath" –silent
   ```

3. After the installation, add the MCR directory (<mcr_root>/<version>) to the environment variable.

   ```
   setenv LD_LIBRARY_PATH
   <mcr_root>/<version>/runtime/glnxa64:
   <mcr_root>/<version>/bin/glnxa64:
   <mcr_root>/<version>/sys/os/glnxa64:
   <mcr_root>/<version>/sys/java/jre/glnxa64/
      jre/lib/amd64/native_threads:
   <mcr_root>/<version>/sys/java/jre/glnxa64/
      jre/lib/amd64/server:
   <mcr_root>/<version>/sys/java/jre/glnxa64/
      jre/lib/amd64:
   setenv XAPPLRESDIR<mcr_root>/<version>/X11/
      app-defaults
   ```

4. Run HAT by issuing the command: "./HAT".

Acknowledgment

We wish to thank Claudia Erpelinck-Verschueren for technical assistance in the preparation of the *CEBPA* C-terminal mutant samples.

References

1. Good SR, Thieu VT, Mathur AN, Yu Q, Stritesky GL, Yeh N et al (2009) Temporal induction pattern of STAT4 target genes defines potential for Th1 lineage-specific programming. J Immunol 183:3839–3847
2. Beekman R, Valkhof M, Erkeland SJ, Taskesen E, Rockova V, Peeters JK et al (2011) Retroviral integration mutagenesis in mice and comparative analysis in human AML identify reduced PTP4A3 expression as a prognostic indicator. PLoS One 6:e26537
3. Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR et al (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods 3:503–509
4. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ et al (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science 302:842–846
5. Taskesen E, Beekman R, de Ridder J, Wouters BJ, Peeters JK, Touw IP et al (2010) HAT: hypergeometric analysis of tiling-arrays with application to promoter-GeneChip data. BMC Bioinformatics 11:275
6. Munch K, Gardner PP, Arctander P, Krogh A (2006) A hidden Markov model approach for determining expression from genomic tiling micro arrays. BMC Bioinformatics 7:239
7. Yu WH, Hovik H, Chen T (2010) A hidden Markov support vector machine framework incorporating profile geometry learning for identifying microbial RNA in tiling array data. Bioinformatics 26:1423–1430
8. Knott SR, Viggiani CJ, Aparicio OM, Tavare S (2009) Strategies for analyzing highly enriched IP-chip datasets. BMC Bioinformatics 10:305
9. Kuan PF, Chun H, Keles S (2008) CMARRT: a tool for the analysis of ChIP-chip data from tiling arrays by incorporating the correlation structure. Pac Symp Biocomput 515–526
10. Zhang Y (2008) Poisson approximation for significance in genome-wide ChIP-chip tiling arrays. Bioinformatics 24:2825–2831
11. Wu H, Ji H (2012) JAMIE: a software tool for jointly analyzing multiple ChIP-chip experiments. Methods Mol Biol 802:363–375
12. Mo Q, Liang F (2010) A hidden Ising model for ChIP-chip data analysis. Bioinformatics 26: 777–783
13. Mo Q, Liang F (2010) Bayesian modeling of ChIP-chip data through a high-order Ising model. Biometrics 66:1284–1294
14. Karpikov A, Rozowsky J, Gerstein M (2011) Tiling array data analysis: a multiscale approach using wavelets. BMC Bioinformatics 12:57
15. Otto C, Reiche K, Hackermuller J (2012) Detection of differentially expressed segments in tiling array data. Bioinformatics 28:1471–1479
16. Lan X, Bonneville R, Apostolos J, Wu W, Jin VX (2011) W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. Bioinformatics 27:428–430
17. Kechris KJ, Biehs B, Kornberg TB (2010) Generalizing moving averages for tiling arrays using combined p-value statistics. Stat Appl Genet Mol Biol 9:Article29
18. Zacher B, Kuan PF, Tresch A (2010) Starr: simple tiling ARRay analysis of Affymetrix ChIP-chip data. BMC Bioinformatics 11:194
19. Droit A, Cheung C, Gottardo R (2010) rMAT–an R/Bioconductor package for analyzing ChIP-chip experiments. Bioinformatics 26:678–679
20. Judy JT, Ji H (2009) TileProbe: modeling tiling array probe effects using publicly available data. Bioinformatics 25:2369–2375
21. Cesaroni M, Cittaro D, Brozzi A, Pelicci PG, Luzi L (2008) CARPET: a web-based package for the analysis of ChIP-chip and expression tiling data. Bioinformatics 24:2918–2920
22. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS (2006) Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci USA 103:12457–12462
23. Kim CK, Kikuchi S, Hahn JH, Park SC, Kim YH, Lee BW (2010) Computational identification of anthocyanin-specific transcription factors using a rice microarray and maximum boundary range algorithm. Evol Bioinform Online 6:133–141
24. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193
25. Lieberman LA, Banica M, Reiner SL, Hunter CA (2004) STAT1 plays a critical role in the regulation of antimicrobial effector mechanisms, but not in the development of Th1-type responses during toxoplasmosis. J Immunol 172:457–463
26. Nguyen KB, Watford WT, Salomon R, Hofmann SR, Pien GC, Morinobu A (2002) Critical role for STAT4 activation by type 1

interferons in the interferon-gamma response to viral infection. Science 297:2063–2066

27. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31:3576–3579
28. Kel A, Voss N, Jauregui R, Kel-Margoulis O, Wingender E (2006) Beyond microarrays: find key transcription factors controlling signal transduction pathways. BMC Bioinformatics 7(Suppl 2):S13
29. Rosenbauer F, Tenen DG (2007) Transcription factors in myeloid development: balancing differentiation with transformation. Nat Rev Immunol 7:105–117
30. Friedman AD (2007) Transcriptional control of granulocyte and monocyte development. Oncogene 26:6816–6828
31. Taskesen E, Bullinger L, Corbacioglu A, Sanders MA, Erpelinck CA, Wouters BJ et al (2011) Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. Blood 117:2469–2475
32. van Waalwijk B, van Doorn-Khosrovani S, Erpelinck C, Meijer J, van Oosterhoud S, van Putten WL, Valk PJ et al (2003) Biallelic mutations in the CEBPA gene and low CEBPA expression levels as prognostic markers in intermediate-risk AML. Hematol J 4:31–40
33. de Bruijn MF, Speck NA (2004) Core-binding factors in hematopoiesis and immune function. Oncogene 23:4238–4248
34. Rabault B, Ghysdael J (1994) Calcium-induced phosphorylation of ETS1 inhibits its specific DNA binding activity. J Biol Chem 269: 28143–28151
35. Tenen DG, Hromas R, Licht JD, Zhang DE (1997) Transcription factors, normal myeloid development, and leukemia. Blood 90:489–519
36. Grall FT, Prall WC, Wei W, Gu X, Cho JY, Choy BK et al (2005) The Ets transcription factor ESE-1 mediates induction of the COX-2 gene by LPS in monocytes. FEBS J 272: 1676–1687
37. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D et al (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645
38. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28–36
39. Viswanathan GA, Seto J, Patil S, Nudelman G, Sealfon SC (2008) Getting started in biological pathway construction and analysis. PLoS Comput Biol 4:e16
40. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102:15545–15550