

RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets

Morgane Thomas-Chollier, Carl Herrmann, Matthieu Defrance, Olivier Sand, Denis Thieffry, Jacques Van Helden

▶ **To cite this version:**

Morgane Thomas-Chollier, Carl Herrmann, Matthieu Defrance, Olivier Sand, Denis Thieffry, et al.. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Research, 2012, 40 (4), pp.e31-e31. 10.1093/nar/gkr1104 . hal-01624284

HAL Id: hal-01624284
https://amu.hal.science/hal-01624284

Submitted on 9 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Published online 8 December 2011

Nucleic Acids Research, 2012, Vol. 40, No. 4 ***e31***
doi:10.1093/nar/gkr1104

RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets

Morgane Thomas-Chollier[1], Carl Herrmann[2], Matthieu Defrance[3], Olivier Sand[4], Denis Thieffry[2,5] and Jacques van Helden[2,6,*]

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, Berlin 14195, Germany, [2]Technological Advances for Genomics and Clinics (TAGC), INSERM U928 & Université de la Méditerranée, Campus de Luminy, Marseille F-13288, France, [3]Centro de Ciencias Genomicas, Universidad Nacional Autónoma de México, Avenida Universidad, Cuernavaca, Morelos 62210, Mexico, [4]CNRS-UMR8199 Institut de Biologie de Lille, Génomique et maladies métaboliques. 1, rue du Pr Calmette, Lille 59000, [5]Institut de Biologie de l'Ecole Normale Supérieure – UMR ENS & CNRS 8197 & INSERM 1024, 46 rue d'Ulm, Paris 75005, France and [6]Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Université Libre de Bruxelles, Campus Plaine, CP 263. Bld du Triomphe, Bruxelles B-1050, Belgium

Received September 3, 2011; Revised November 3, 2011; Accepted November 5, 2011

Downloaded from https://academic.oup.com/nar/article/40/4/e31/2411061 by guest on 09 November 2021

ABSTRACT

ChIP-seq is increasingly used to characterize transcription factor binding and chromatin marks at a genomic scale. Various tools are now available to extract binding motifs from peak data sets. However, most approaches are only available as command-line programs, or via a website but with size restrictions. We present peak-motifs, a computational pipeline that discovers motifs in peak sequences, compares them with databases, exports putative binding sites for visualization in the UCSC genome browser and generates an extensive report suited for both naive and expert users. It relies on time- and memory-efficient algorithms enabling the treatment of several thousand peaks within minutes. Regarding time efficiency, peak-motifs outperforms all comparable tools by several orders of magnitude. We demonstrate its accuracy by analyzing data sets ranging from 4000 to 1 28 000 peaks for 12 embryonic stem cell-specific transcription factors. In all cases, the program finds the expected motifs and returns additional motifs potentially bound by cofactors. We further apply peak-motifs to discover tissue-specific motifs in peak collections for the p300 transcriptional co-activator. To our knowledge, peak-motifs is the only tool that performs a complete motif analysis and offers a user-friendly web interface without any restriction on sequence size or number of peaks.

INTRODUCTION

ChIP-seq (1,2) has recently become a method of choice to study the binding preferences of transcription factors, as well as the localization of epigenetic regulatory marks at a genomic scale. The first steps of the computational analysis (read mapping and peak calling) typically result in several thousands of peak regions ranging between 200 and 10 000 bp. Motif analysis is required to extract the relevant information from these regions: discover binding motifs that capture the binding specificity of the pulled-down factor and their possible co-regulators; compare discovered motifs to databases to predict associated transcription factors; predict the exact positions of the binding sites (usually much shorter than the peak regions); study the binding specificity of transcription factors in various contexts (cell types, mutant strains and transcription factor isoforms).

Specialized software tools have recently been developed for the analysis of ChIP-seq peaks, supporting different combinations of motif-related tasks (Table 1). An important bottleneck for most existing tools is that the underlying algorithms were originally developed to discover binding motifs from a small set of co-regulated promoters, and can hardly treat the thousands of peaks produced by ChIP-seq experiments. This limitation is typically circumvented by restricting motif discovery to a few hundreds peak regions and by truncating the peaks to a maximal width (e.g. 100 bp) to further reduce the total size of the sequence set (3–5). However, given the power of the genome-wide experimental approach, one would like to be able to analyze the full data set. Some alternative algorithms support the analysis of large-scale data sets but

*To whom correspondence should be addressed. Tel: + 32 (0) 2 650 20 76; Fax: + 32 (0) 2 650 54 25; Email: Jacques.van.Helden@ulb.ac.be

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1. Features of software tools used for analyzing motifs in ChIP-seq peak seqm

Program	Peak-motifs	ChipMunk	CompleteMotifs	MEME-ChIP	MICSA	GimmeMotifs
Web interface	Yes	Yes	Yes	Yes	No	No
Size limitation	Unrestricted (website tested with 22 Mb)	100 kb (website)	500 kb (web site)	Unrestricted, but analysis limited to 600 peaks clipped to 100 bp	Motif discovery restricted to a few hundred base pairs	–
Stand-alone version	Yes	Yes	No	Yes	Yes	Yes
Tasks						
Peak finding	No	No	No	No	Yes	No
Annotation of peak-flanking genes	No	No	Yes	No		No
Sequence composition (mono- and di-nucleotides)	Yes	No	No	No		No
Motif discovery	Yes	Yes	Yes	Yes	Yes	Yes
Enrichment in motifs from databases	No	No	Yes	Yes		No
Enrichment in discovered motifs	Yes	No	No	No		No
Peak scoring	No	No	No	Yes	Yes	No
Motif clustering	No	No	No	No		Yes
Comparison discovered motifs/motif DB	Yes	No	No	Yes		Yes
Sequence scanning for site prediction	Yes	No	No	Yes		No
Positional distribution of sites inside peaks	Yes	No	Yes	No		Yes
Visualization in genome browsers	Yes	No	Yes	No		No
Motif discovery algorithms	RSAT oligo-analysis RSAT dyad-analysis RSAT local-word-analysis MEME ChIPMunk	ChipMunk	ChipMunk MEME Weeder	MEME DREME	MEME	MEME Weeder MotifSampler BioProspector Gadem Improbizer MDmodule Trawler MoAn
Pattern matching algorithms	RSAT matrix-scan-quick	No	patser	MAST + AME (enrichment)		No
Motif comparison algorithm	RSAT compare-motifs	No	STAMP	TOMTOM		STAMP
Motif clustering algorithm						STAMP
Comparison between discovered motifs	Yes	No	Yes	No		Yes
Motif database comparisons	JASPAR UNIPROBE DMMPMM RegulonDB upload your own database	No	JASPAR TRANSFAC	JASPAR TRANSFAC UNIPROBE FLYREG DPINTERACT SCPD DMMPMM and many others		No
Motif sizes	Variable (multiple word assembly)	User-specified	≤25 for MEME ≤12 for Weeder ≤13 for ChipMunk			Predefined ranges (small, medium, large, extra-large)
Multiple motifs	Yes		Yes	Yes		Yes
Ref (PMID)	This article	20736340	21183585	21486936	20375099	21081511

The table summarizes the tasks, algorithms and usability properties to compare the different software options for the users. Most programs offer a web interface, but apply restrictive limitations on the size of the data sets to process. Although all programs support motif discovery, the other tasks are quite diverse and not all covered by a single program.

Downloaded from https://academic.oup.com/nar/article/40/4/e31/2411061 by guest on 09 November 2021

are only available via a Unix shell interface (6–8), or as MATLAB functions (9), and are thus of poor usability for life-science researchers.

We have developed a computational pipeline called 'peak-motifs', motivated by the pressing need for a statistically reliable, time-efficient and user-friendly framework to analyze full data sets of ChIP-seq peaks or similar data (ChIP-PET, ChIP-on-chip, CLIP-seq). This comprehensive pipeline takes as input a set of peak sequences, discovers exceptional motifs, compares them with motif databases, predicts binding site positions and returns a structured HTML report with direct links to visualization in the UCSC genome browser (Figure 1). This tool can also be used for differential analyses, where two datasets are given as input (e.g. test versus control, or peaks from two experimental conditions), to discover motifs specific to one of the datasets.

We first show that this motif discovery approach is significantly faster than other available alternatives, thereby allowing processing of comprehensive ChIP-seq data sets, even from the web server. We then demonstrate the biological relevance of the motifs discovered by our pipeline with two study cases, highlighting the benefit of analyzing complete datasets and using complementary approaches for motif discovery.

MATERIALS AND METHODS

The motif discovery step relies on a combination of tried-and-tested algorithms integrated in the software suite regulatory sequence analysis tools (RSAT, http://rsat.ulb.ac.be/rsat/) (10–12), which use complementary criteria to detect exceptional words (oligonucleotides and spaced motifs): global over-representation of oligonucleotides (*oligo-analysis*) or spaced pairs (*dyad-analysis*), heterogeneous positional distribution (*position-analysis*) and local over-representation (*local-word-analysis*) (12–15).

The motif comparison step is performed by *compare-matrices* (12), which supports a wide range of scoring metrics and displays the results as multiple alignments of logos, enabling to grasp the similarities between a discovered motif and several known motifs. This feature is particularly valuable to reveal adjacent fragments of the discovered motif showing similarities with two distinct known motifs, suggesting a bipartite motif for two factors (see the SOCT motif in Figure 4 and below).

As the individual components of the workflow have been described previously (12), we briefly explain here the choice of parameters for the different steps of peak-motifs analyses. The full list of commands and parameters are automatically reported at the end of each peak-motifs report. The parameters used for the case studies are available in the peak-motifs reports on the supporting website (http://rsat.bigre.ulb.ac.be/~rsat/supp_material_peak-motifs/).

Motif discovery

Word-based analysis is performed with hexanucleotides ($k = 6$) and heptanucleotides ($k = 7$). The significance tests underlying pattern detection ensure a control of the rate of false positives, with suitable multi-testing corrections. The motif discovery algorithms support higher order background models, which are of particular importance for modeling genomic sequences of vertebrates. For *oligo-analysis*, expected word frequencies were estimated with a Markov model of order $m = k - 2$, trained in the peak sequences. The website also allows to select lower order Markov models, which are less stringent but achieve a higher sensitivity with small data sets. For differential analysis, the expected frequency of each k-mer is estimated by taking the observed frequency of the same k-mer in the control set. Significant words are assembled using '*pattern-assembly*' and converted to position-specific scoring matrices with *matrix-from-patterns*.

Motif comparison

Discovered motifs are compared (using *compare-matrices*) to one or several databases of known transcription factor binding motifs. The website directly supports comparisons with JASPAR (16), UNIPROBE (17), REGULONDB (18) and *Drosophila*-specific collections (19), thus providing a vast choice of known motifs, for a wide range of organisms. Personal or license-protected motif collections can also be uploaded. Several metrics are computed to measure the similarity between each matrix pair (Pearson correlation, width normalized correlation, logo dot product, correlation of information content, normalized Sandelin–Wasserman, sum of squared distances and normalized Euclidian similarity). As these metrics span over very different ranges, we convert them to ranks and compute a mean rank in order to obtain a robust comparison metrics.

Matrix scanning

Peak sequences are scanned to predict binding sites with the program *matrix-scan*, using as background model a Markov chain of order 1 trained on the peak sequences themselves. Noteworthy, a Markov order $m \geq 1$ is required to account for the CpG avoidance observed in vertebrate genomes, and for other types of context-dependent residue probabilities. Predicted binding sites are mapped onto the genome (*convert-features*) and exported as BED files to be automatically loaded as custom tracks on the UCSC genome browser.

RESULTS

Peak-motifs processes full-sized ChIP-seq data sets in a few minutes

We assessed the time efficiency of peak-motifs by analyzing data sets of increasing sizes (from 100 to 1 000 000 peaks of 100 bp each), with total sequence sizes ranging from 10 kb to 100 Mb The computing time of the motif discovery algorithms integrated in peak-motifs increases linearly with sequence size and outperforms all the other existing motif discovery tools used in this comparison (Figure 2, Supplementary File S1). Data sets of several tens of megabytes are

Downloaded from https://academic.oup.com/nar/article/40/4/e31/2411061 by guest on 09 November 2021

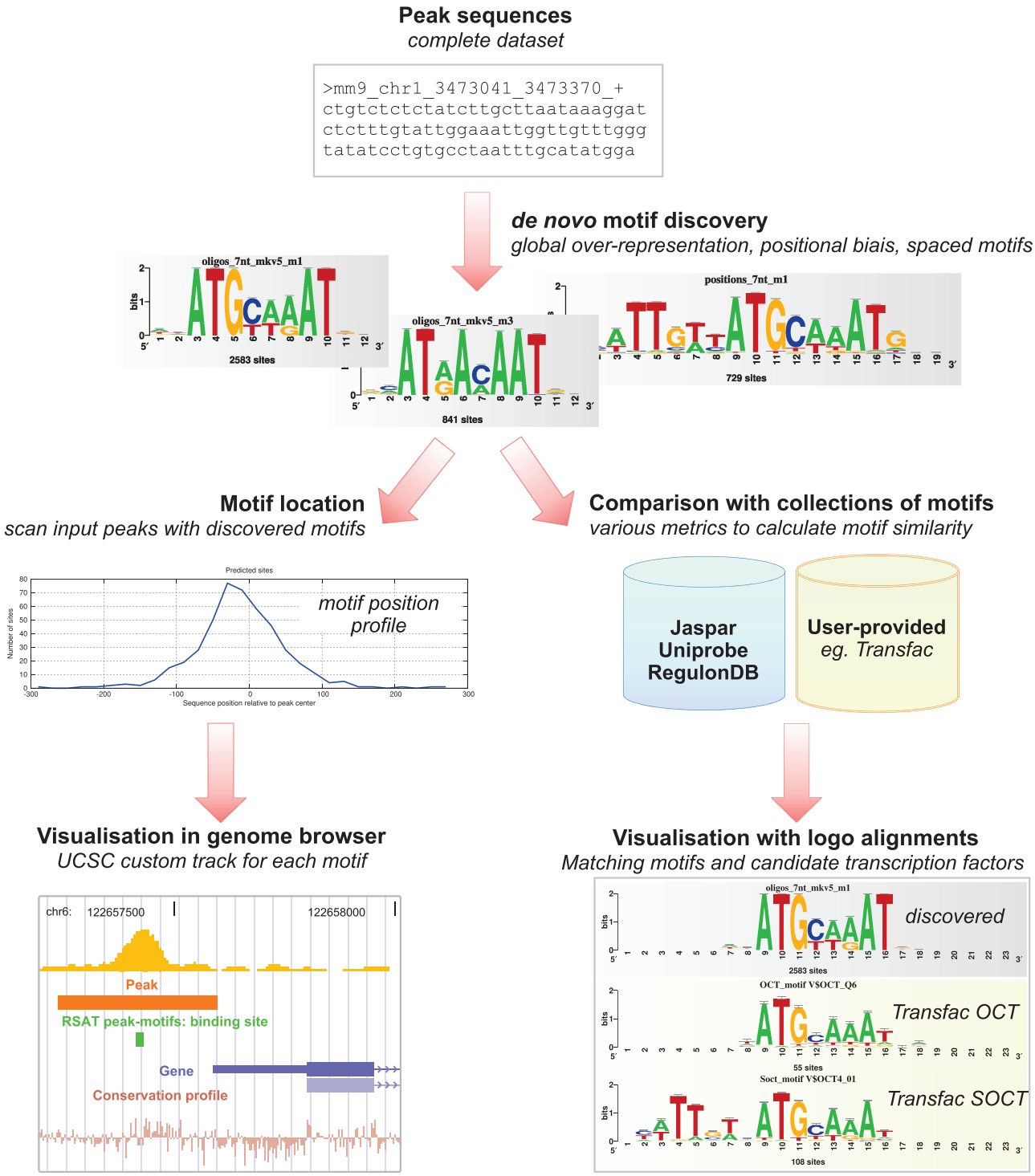


Figure 1. Schematic flow chart of the peak-motifs pipeline. For sake of clarity, only the main analysis steps are depicted. The pipeline takes as input a set of peak sequences, and runs several *de novo* motif discovery algorithms based on different detection criteria: over-representation, differential representation (test versus control), global position bias or local over-representation along the centered peaks. Transcription factors are predicted by matching discovered motifs against several public motif databases and/or against user-uploaded motif collections. Peak sequences are scanned with the discovered motifs to predict precise binding positions. These positions are then automatically exported as an annotation track for UCSC genome browser, thus enabling a flexible visualization in their genomic context.

processed in a few minutes on a personal computer (the most efficient tool, *oligo-analysis*, treats 100 Mb in 3 min). This linear time response enables peak-motifs to scale up efficiently with sequence size, and allows us to provide an easy access via a web interface, without any data size restriction. This moreover gives us the possibility to run four distinct algorithms in order to detect motifs of various types (oligonucleotides, spaced pairs) based on complementary criteria (over-representation, positional heterogeneity).

Downloaded from https://academic.oup.com/nar/article/40/4/e31/2411061 by guest on 09 November 2021

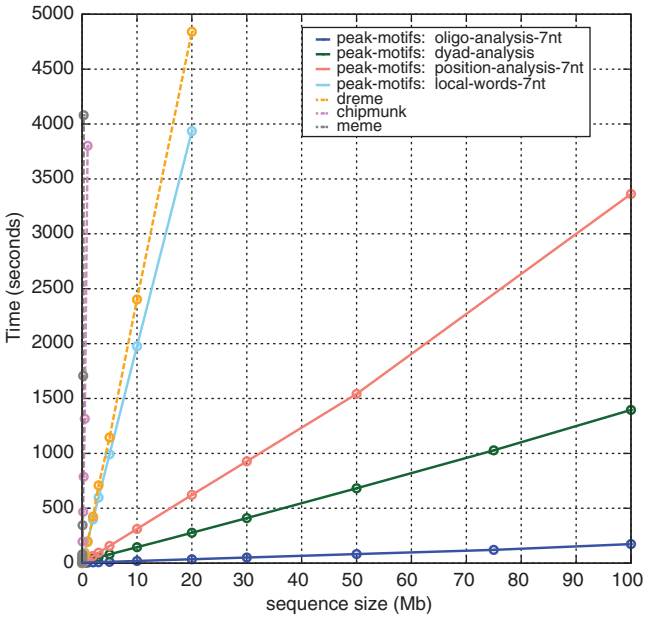


Figure 2. Time efficiency of motif discovery algorithms integrated in peak-motifs (plain lines) compared to alternative algorithms (dotted lines). The abscissa indicates sequence sizes, the ordinate processing times. The programs oligo-, dyad-, position-analysis and DREME show a linear time complexity (the power is ~1), ChIPMunk has a quasi-linear complexity (power 1.27) and MEME a more than quadratic complexity (power 2.21). See Supplementary File S1 for the detailed analysis.

Analysis of the ChIP-seq peak sets for 12 DNA-binding transcription factors involved in mouse ES cell pluripotency and self-renewal

To evaluate the accuracy of the predicted motifs, we analyzed the ChIP-seq peak sets for 12 DNA-binding transcription factors involved in mouse embryonic stem cell pluripotency and self-renewal (20). The read sequences were downloaded from the Gene Expression Omnibus website and mapped with Bowtie (21) on the mouse mm9 assembly. Peak regions were extracted from reads using MACS (22) with a false discovery rate threshold of 0.2, and processed with PeakSplitter (23) to obtain actual peaks. For the Smad1 data set, MACS did not return a single peak with the selected parameters. We, therefore, used the peaks from the initial data set GSM288348, which contains 1084 ChIP-seq peaks for the Smad1 factor. The other data sets comprise between 4249 peaks for Stat3 (totaling 1.4 Mb) and 1 28 469 peaks for Esrrb (36.6 Mb).

For each of the 12 tested factors, peak-motifs discovered the correct motif (Figure 3). The relevant motifs were generally detected independently by several of the four algorithms, indicating that they are not only over-represented (*oligo-analysis*, *dyad-analysis*) but also positionally biased around peak centers (*local-word-analysis*, *position-analysis*). For several peak sets, recent studies (5,24) using novel motif-finding programs returned more accurate motifs than the original study, which was restricted to the 500 top-scoring peaks. Our comprehensive analysis also returned more accurate motifs than the original study, and performed as well or better than other recent motif-finding programs, as detailed below.

In the Sox2 and Oct4 peak sets, peak-motifs found not only the composite 'SOCT' motif bound by the Sox2/Oct4 complex (reported by Chen and co-workers), but also the distinct motifs recognized by Sox2 (CTCTTTGTT) and Oct4 (ATGyAAAt), respectively (Figure 4, top). Interestingly, in the Oct4 data set, unknown motifs were returned with a high significance, (i.e. motifs with no significant similarity with the consensus encompassed by the common databases). Such motifs may reveal alternative consensus, as in the case of the motif crTATGCGCA TAyg, which actually corresponds to an alternative Oct4 motif, also detected in other recent studies (5,24).

As discussed by Chen and co-workers, Nanog and Smad1 frequently bind the same regions as Sox2/Oct4, which raises a particular difficulty for motif discovery. Indeed, their analysis of the Nanog peak set returned a Sox2-like motif instead of the Nanog binding motif. Subsequently, this Sox2-like motif was erroneously annotated as Nanog binding in the TRANSFAC database (matrix V$NANOG_02), although its consensus (CYWTTGTTNT) clearly differs from the previously annotated Nanog consensus (GGGNCCATTKCC, TRANSFAC matrix V$NANOG_01). The prevalent motif discovered by peak-motifs in Nanog peaks corresponds to the SOCT binding motif, while the canonical Nanog motif is not found. However, peak-motifs reports a motif (sCGCmaTCAbg) that is not similar to any motif found in the databases (Figure 4, middle). A similar motif with a ccAT(C/T)A core was also reported by Bailey (5), and actually corresponds to an experimentally validated alternative Nanog motif (25).

For the Smad1 factor, the peak size distribution of the original data set seems to be biased toward very small peaks (smallest peak is 1bp, mean size id 30 bp); nevertheless, peak-motifs was able to discover a motif agAAACA AAGCmar that matches the canonical Smad1 motif (V$SMAD1_01 adAAACAAAGcm). In addition, several other discovered motifs match a Sox-like motif wGAACAATAga, confirming the frequent co-binding of Smad1 and Sox.

In the E2f1 peak set, peak-motifs discovered several motifs matching the generic E2F consensus (GGCGsg, matrix V$E2F_Q2) but distinct from the E2f1-specific consensus (TTTsGCGG, in Transfac matrix V$E2F1_Q4; TTTsGCGC in JASPAR matrix MA0024.1) (Figure 4, bottom). Whereas no E2f1 motif was detected in the original study by Chen and co-workers (20), an E2f-like motif similar to ours was reported in ref. (5).

In summary, our analysis of the 12 peak sets from Chen and co-workers significantly improved motifs as compared to the original study, highlighting the value of applying motif discovery to full-size data sets. Remarkably, in addition to the motifs corresponding to the transcription factors targeted by the experiments, peak-motifs also returned several motifs corresponding to transcription factors presumably involved in the same regulatory pathways.

Downloaded from https://academic.oup.com/nar/article/40/4/e31/2411061 by guest on 09 November 2021

Data set	Nb peaks	Total seq size (Mb)	Best-matched ref motif ID	Best-matched reference motif	Best-matching discovered motif	Cor	Cov	oligo-analysis	dyad-analysis	position-analysis	local-word-analysis	Algorithms finding the motif	Reference Motifs found	Other motifs found
Smad1*	1.084	0,03	M01216	adAAACAAAGcm	agAAACAAAGCmar	0,81	0,70		1			1	V$SMAD1_01	Not a single peak selected by MACS with FDR<=0.2 ; we used the peaks from the GSM288348 dataset instead
Stat3	4.249	1,36	MA0144.1	TTCCaGGAAr	syTTCCwGGAAGtsm	0,98	0,67		1	**1**		2	V$STAT_Q6 V$STAT_01	oligo-analysis returns SP, AP, and ER motifs
Nanog	7.699	3,03	V$NANOG_01	ggGvyCATTkcC	WMATTWSCATTW	0,81	0,50	1				1	V$NANOG_01	All programs match the Sox2 motifs Note: the TRANSFAC motif called "V$NANOG_02" is actually a Sox2 motif. It is found by all programs. Additional motifs found with oligo-anlaysis, local-words, and positional motifs, in particular YY variants.
Sox2	8.014	2,78	MA0143.1	CCwTTGTyaTGcaaA	YWTTGTYATKY	0,93	0,73	**1**	1	1	1	4	V$SOX2_Q6 V$SOX_Q6	Sox but also Oct motifs are found by the different algorithms
Oct4	9.198	2,92	V$OCT4_01	tdATTTgCATw	HATTWRCATWW	0,96	0,79	1	1	**1**	1	4	V$OCT_Q6 V$OCT4_01	The Oct4 motif is clearly found, but some of the discovered instances also match the compositite Sox/Oct motifs.
c-Myc	13.742	4,62	V$MYC_Q2	CACGTGb	rcCACGTGgy	0,99	0,70	**1**	1	1	1	4	V$MYCMAX_03 V$MYCMAX_02 V$CMYC_01	LBP1, SP (oligo-analysis) MEF2 (position-anlaysis)
n-Myc	21.513	6,09	MA0104.1	CACGTG	ryCACGTGry	1,00	0,60	**1**		1	1	3	MA0104.1 V$NMYC_01 V$EBOX_Q6_01	AP, SP (oligo-analysis) c-Myc, MEF2 (position-analysis) + other motifs (dyad-analysis)
Klf4	30.577	7,72	V$GKLF_02	rCCmCrCCCwkc	rrCCmCrCCCTyy	0,99	0,86	1	1	**1**	1	4	V$GKLF_02	E2F, FXR, ER, LBP1 (oligos_6nt)
CTCF	44.519	14,48	MA0139.1	ygrCCAsyAGrkGGCrsyr	grCCACyAGrkG	0,95	0,63	1		1	**1**	3	V$CTCF_02 V$CTCF_02	MEF2, SP (oligo-analysis) other motifs (dyad-analysis)
Zfx	50.017	14,39	MA0146.1	ssscAGGCCkcrscss	srGscAGGCCywGss	0,87	0,88	1	1	1	**1**	4	V$ZFX_01	AP, SP (oligo-analysis) FOXO1 (positions-analysis) + other motifs (dyad-analysis
Tcfcp21	57.975	19,47	MA0145.1	CCrGyyyaadCCrG	hrrArCCAGyyTgr	0,91	0,47			**1**		1	MA0145.1	Tcfcp21 found only with position-analysis other motifs returned by other programs
E2F1	91.490	24,58	V$E2F_Q2	ttTTTCsCGsc	SSCGGSRGCGSS	0,90	0,50	**1**			1	2	V$E2F_Q2	match only covers the right side of the motif
Esrrb	128.469	36,59	V$ERR2_01	yCAAGGTCAcss	gtTCAAGGTCAkc	0,94	0,79	**1**		1	1	3	V$ERR2_01 Jaspar MA0141.1	AP4 (oligo-analysis), ER (oligo-analysis)

Figure 3. Most significant motifs discovered with the different algorithms encompassed by peak-motifs for ChIP-seq peak collections pulled down with 12 transcription factors involved in ES cell pluripotency (20). The first three columns indicate the studied transcription factor and the size of the data set (in number of peaks and in Mb). The fourth and fifth columns display the ID and consensus of the chosen reference motif. The sixth column shows the best motif found by peak-motifs, followed by two estimations of the correlation between the discovered and the matched motifs (Cor and Cov). The following columns detail which algorithm(s) detected this motif, and which motifs from the Jaspar and Tranfac databases were similar to the found motif.

Downloaded from https://academic.oup.com/nar/article/40/4/e31/2411061 by guest on 09 November 2021

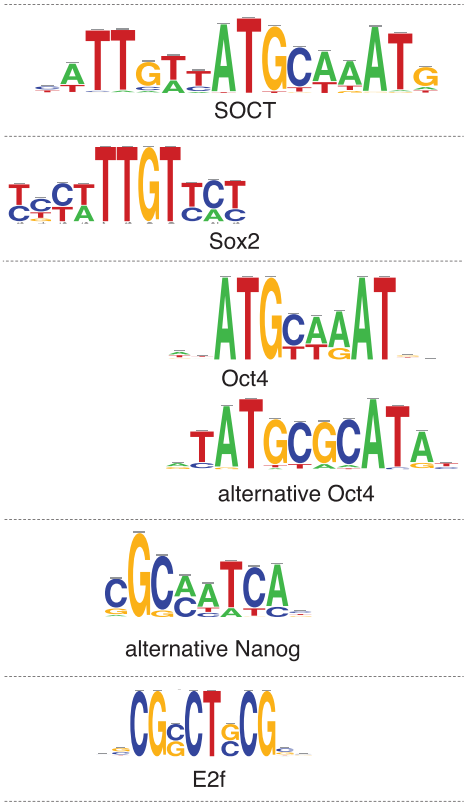


Figure 4. Logos of the motifs discovered by peak-motifs for the factors Oct4, Sox2, Nanog and E2f1 adapted from the ChIP-seq data set by Chen *et al.* (20).

Analysis of the ChIP-seq peak sets for p300 in four different mouse embryonic tissues

Beyond the analysis of motif-specific DNA-binding transcription factors, the ChIP-seq approach can be used to characterize binding profiles of epigenetic regulators, chromatin marks and generic cofactors. In contrast with the transcription factors analyzed above, such cofactors do not recognize specific DNA motifs, but interact with various specific DNA-binding transcription factors and facilitate the activation of their target genes by modifying DNA structure. Genome-wide location analyzes of the generic cofactor p300 have been performed to reveal regions transcriptionally active in different tissues during embryonic development (26,27). Since the DNA regions identified by this approach likely contain binding sites for the transcription factors specifically active in the analyzed tissues and developmental stages, we wondered if peak-motifs would be able to detect the corresponding motifs. In this respect, we used peak-motifs to detect motifs from the ChIP-seq peaks of the generic enhancer-associated p300 cofactor. In the two aforementioned studies, binding profiles of this cofactor were characterized in several embryonic mouse tissues (heart, midbrain, forebrain and limb) and some binding regions were validated as tissue-specific enhancers. However, the transcription factors bound to those enhancers remain unknown.

We retrieved the peak locations for all four tissues. By running peak-motifs in the p300 peak sets in each of these four tissues, we were able to identify motifs potentially bound by tissue-specific regulators, as well as some motifs common to all four data sets, probably corresponding to ubiquitous activators (Supplementary File S2). Peak-motifs compared these discovered motifs to motifs of known factors stored in databases, including Transfac, JASPAR and UniProbe. Tissue-specific motifs include a motif found in the limb data set alone, which matches the consensus of Hox9, known to be involved in limb development (28). We also identify a GATA motif specific to the heart data set, which presumably points to a key factor of the cardiac gene regulatory network. As a validation of these predictions, we verified that the predicted transcription factors are indeed expressed in the corresponding tissues, using expression data from the MGI database (29) (Supplementary File S2).

For further validation, we analyzed data generated by ChIP-seq experiments targeting various heart-specific transcription factors (Mef2, SRF, GATA4, Nkx2.5) in the mouse HL1 cardiomyocyte cell line. Predicted motifs for these data sets strengthen our findings (Figure 5): the predicted GATA motif from the p300 heart data set clusters with similar motifs obtained from the GATA4 data set. Similarly, several motifs obtained in HL1 data sets cluster with the set of motifs from the p300 data sets matching the Mef2 consensus, giving insight into the highly combinatorial nature of cardiogenesis. We also found two 'ubiquitous' motifs significantly over-represented in all four data sets. The first is a C-rich motif, which matches the binding motif of Sp1, consistent with the fact that Sp1 functionally interacts with the acetylase domain of p300 (30,31). The second motif matches the Mef2 consensus (ATTTTTA). Interestingly, Mef2 is known to be involved not only in muscle formation, explaining its presence in the heart and limb data set, but also in CNS development (in particular neuron differentiation).

The relevance of the discovered motifs opens the exciting prospect of predicting which transcription factors and enhancers are active in a given tissue and/or at a given developmental stage, by discovering specific TF motifs in the peaks pulled down by generic cofactors such as p300.

Peak-motifs is accessible through a user-friendly web interface

The simplest way to use peak-motifs is via its user-friendly web interface, where all parameters (background models, word lengths, etc.) are pre-selected according to the optimal conditions found from our study cases. The only required input is the set of peak sequences. A second set of peak sequences can also be provided to serve as background for differential analyses (treatment versus control). Although peak-motifs is designed to process full data sets, the interface offers the possibility to easily reduce the analysis to a subset of top sequences, or yet to clip peaks at a maximal size from their centers, thereby reducing the need for data manipulation on the user

Downloaded from https://academic.oup.com/nar/article/40/4/e31/2411061 by guest on 09 November 2021

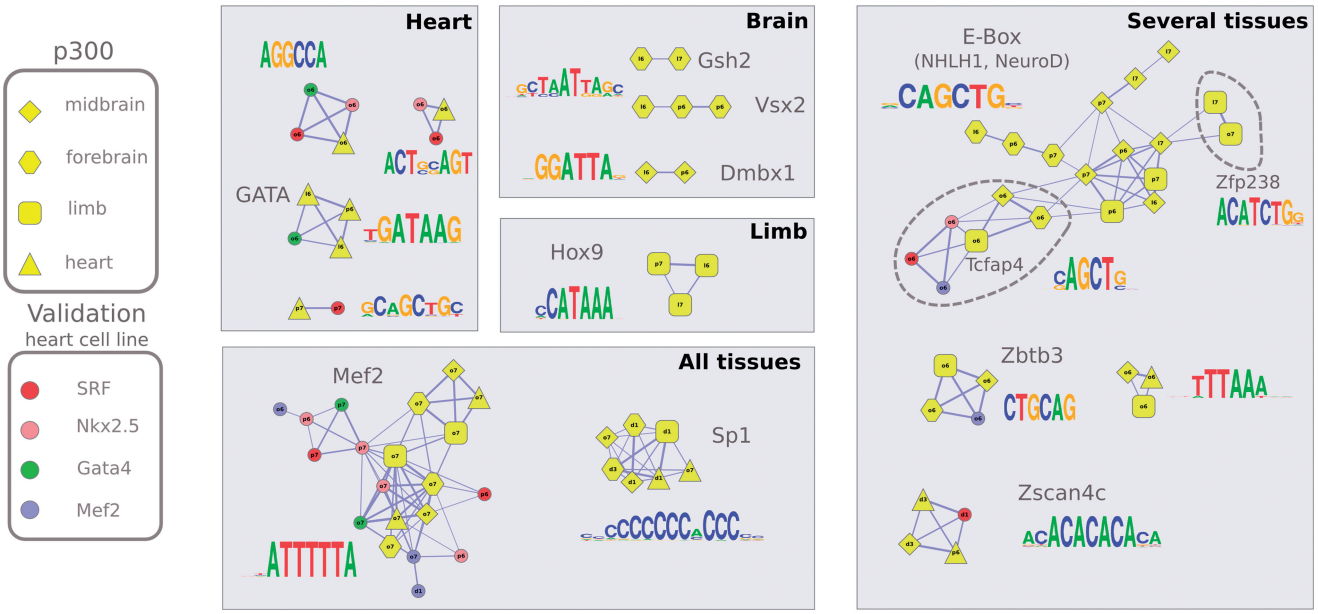


Figure 5. Network of motifs discovered in the p300 data set. Each node represents a motif; the shape and color of the node denote the tissue (for the p300 datasets) and the ChIPed-factor (for the HL1 cell-line datasets, used as a validation), respectively. Two motifs are joined by a line if their normalized correlation is above 0.75; the width of the line denotes the degree of correlation. Node labels refer to the algorithm used to discover the motif: L (local-words), P (position-analysis), O (oligo-analysis), D (dyad-analysis) as well as the considered word length (6 or 7). The names of the transcription factor(s) likely associated with the motif clusters are also indicated, together with a representative logo.

side. The web page is documented with a manual providing detailed information about each option. A 'demo' button fills up the form with a typical test set. A tutorial further guides new users through choices of parameters and explains how to interpret the results. In addition to its website access, peak-motifs can be used as a stand-alone application (Unix shell), as well as SOAP/WSDL web services (thereby enabling bioinformaticians to automate its use, without installing it on their machine).

A particular effort has been made to generate a clear and easily interpretable output for less-advanced users, while providing links to the raw results for the expert users. All result files are presented in standard formats and are downloadable as an archive along with the summary web page, to allow further analysis with third-party software. To our knowledge, peak-motifs is the only ChIP-seq pipeline offering direct visualization of the predicted binding sites as custom tracks in the UCSC genome browser. This feature is of prime importance to interpret the results in light of the genomic annotation, in order to plan experiments for further validation of the results.

DISCUSSION

Peak-motifs is a comprehensive pipeline to efficiently discover motifs and identify putative transcription factors in ChIP-seq and related data sets. We demonstrated its biological validity by recovering the correct motifs from 12 ChIP-seq sets corresponding to known transcription factors (20). We also performed an original analysis of the binding profiles of the generic cofactor p300 (26), which led us to predict specific motifs and transcription factors that are active in specific tissues at specific developmental stages. Our benchmarks showed that for large data sets peak-motifs outperforms its most serious competitors by a factor of at least 100, allowing us to analyze full data sets in a matter of minutes. This time efficiency enables an interactive web access for comprehensive data sets, thereby constituting a convenient tool for ChIP-seq data analyses even for naive users. This tool will be of broad interest to the increasing community of experimentalists and bioinformaticians who are confronted to the challenging issue of extracting interpretable information from the massive amounts of data resulting from next generation sequencing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Files 1 and 2.

ACKNOWLEDGEMENTS

The collaboration between BiGRe and ENS has been stimulated by a 2-months invitation of JvH as visiting professor at Ecole Normale Supérieure.

FUNDING

Alexander von Humboldt foundation (to M.T.C.); Agence Nationale de la Recherche (ANR) partner of the

Downloaded from https://academic.oup.com/nar/article/40/4/e31/2411061 by guest on 09 November 2021

ERASysBio+ initiative supported under the EU ERA-NET Plus scheme in FP7; ANR Young Researchers Grant 'CardiHox' (to C.H.); the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office [project P6/25 (BioMaGNet)]; EU-funded COST action [BM1006 'Next Generation Sequencing Data Analysis Network']; FP7 MICROME Collaborative Project ('Microbial genomics and bio-informatics', contract number 222886-2). Funding for open access charge: Belgian Program on Interuniversity Attraction Poles [project P6/25 (BioMaGNet)].

Conflict of interest statement. None declared.

REFERENCES

1. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.
2. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein–DNA interactions. *Science*, **316**, 1497–1502.
3. Boeva,V., Surdez,D., Guillon,N., Tirode,F., Fejes,A.P., Delattre,O. and Barillot,E. (2010) *De novo* motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.*, **38**, e126.
4. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.
5. Bailey,T.L. (2011) DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
6. Hu,M., Yu,J., Taylor,J.M., Chinnaiyan,A.M. and Qin,Z.S. (2010) On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Res.*, **38**, 2154–2167.
7. Kulakovskiy,I.V., Boeva,V.A., Favorov,A.V. and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.
8. van Heeringen,S.J. and Veenstra,G.J. (2011) GimmeMotifs: a *de novo* motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics*, **27**, 270–271.
9. Agius,P., Arvey,A., Chang,W., Noble,W.S. and Leslie,C. (2010) High resolution models of transcription factor–DNA affinities improve *in vitro* and *in vivo* binding predictions. *PLoS Comput. Biol.*, **6**, e1000916.
10. van Helden,J., Andre,B. and Collado-Vides,J. (2000) A web site for the computational analysis of yeast regulatory sequences. *Yeast*, **16**, 177–187.
11. Thomas-Chollier,M., Sand,O., Turatsinze,J.V., Janky,R., Defrance,M., Vervisch,E., Brohee,S. and van Helden,J. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–W127.
12. Thomas-Chollier,M., Defrance,M., Medina-Rivera,A., Sand,O., Herrmann,C., Thieffry,D. and van Helden,J. (2011) RSAT 2011: regulatory Sequence Analysis Tools. *Nucleic Acids Res.*, **29**, W86–W91.
13. van Helden,J., Andre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
14. van Helden,J., Rios,A.F. and Collado-Vides,J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.
15. van Helden,J., del Olmo,M. and Perez-Ortin,J.E. (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res.*, **28**, 1000–1010.
16. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
17. Robasky,K. and Bulyk,M.L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Res.*, **39**, D124–D128.
18. Gama-Castro,S., Salgado,H., Peralta-Gil,M., Santos-Zavaleta,A., Muniz-Rascado,L., Solano-Lira,H., Jimenez-Jacinto,V., Weiss,V., Garcia-Sotelo,J.S., Lopez-Fuentes,A. *et al.* (2011) RegulonDB version 7.0: Transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.
19. Kulakovskiy,I.V., Favorov,A.V. and Makeev,V.J. (2009) Motif discovery and motif finding from genome-mapped DNase footprint data. *Bioinformatics*, **25**, 2318–2325.
20. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
21. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
22. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
23. Salmon-Divon,M., Dvinge,H., Tammoja,K. and Bertone,P. (2010) PeakAnalyzer: Genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics*, **11**, 415.
24. Mason,M.J., Plath,K. and Zhou,Q. (2010) Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics*, **26**, 2826–2832.
25. He,X., Chen,C.-C., Hong,F., Fang,F., Sinha,S., Ng,H.-H. and Zhong,S. (2009) A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS ONE*, **4**, e8155.
26. Blow,M.J., McCulley,D.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.*, **42**, 806–810.
27. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
28. Duboule,D. (1993) The function of Hox genes in the morphogenesis of the vertebrate limb. *Ann. Genet.*, **36**, 24–29.
29. Zhu,Y., King,B.L., Parvizi,B., Brunk,B.P., Stoeckert,C.J. Jr, Quackenbush,J., Richardson,J. and Bult,C.J. (2003) Integrating computationally assembled mouse transcript sequences with the Mouse Genome Informatics (MGI) database. *Genome Biol.*, **4**, R16.
30. Suzuki,T., Kimura,A., Nagai,R. and Horikoshi,M. (2000) Regulation of interaction of the acetyltransferase region of p300 and the DNA-binding domain of Sp1 on and through DNA binding. *Genes Cells*, **5**, 29–41.
31. Billon,N., Carlisi,D., Datto,M.B., van Grunsven,L.A., Watt,A., Wang,X.F. and Rudkin,B.B. (1999) Cooperation of Sp1 and p300 in the induction of the CDK inhibitor p21WAF1/CIP1 during NGF-mediated neuronal differentiation. *Oncogene*, **18**, 2872–2882.

Downloaded from https://academic.oup.com/nar/article/40/4/e31/2411061 by guest on 09 November 2021