Published online 28 July 2012 *Nucleic Acids Research, 2012, Vol. 40, No. 20* ***e160***
doi:10.1093/nar/gks697

Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data

Cem Sievers[1], Tommy Schlumpf[1], Ritwick Sawarkar[1], Federico Comoglio[1] and Renato Paro[1,2,*]

[1]Department of Biosystems Science and Engineering, Swiss Federal Institute of Technology Zurich, Mattenstrasse 26, 4058 Basel and [2]Faculty of Science, University of Basel, Klingelbergstrasse 50, 4056 Basel, Switzerland

Received April 23, 2012; Revised June 18, 2012; Accepted June 26, 2012

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

ABSTRACT

The Photo-Activatable Ribonucleoside-enhanced CrossLinking and ImmunoPrecipitation (PAR-CLIP) method was recently developed for global identification of RNAs interacting with proteins. The strength of this versatile method results from induction of specific T to C transitions at sites of interaction. However, current analytical tools do not distinguish between non-experimentally and experimentally induced transitions. Furthermore, geometric properties at potential binding sites are not taken into account. To surmount these shortcomings, we developed a two-step algorithm consisting of a non-parametric two-component mixture model and a wavelet-based peak calling procedure. Our algorithm can reduce the number of false positives up to 24% thereby identifying high confidence interaction sites. We successfully employed this approach in conjunction with a modified PAR-CLIP protocol to study the functional role of nuclear Moloney leukemia virus 10, a putative RNA helicase interacting with Argonaute2 and Polycomb. Our method, available as the `R` package `wavClusteR`, is generally applicable to any substitution-based inference problem in genomics.

INTRODUCTION

The interaction of RNA and RNA binding proteins (RBPs) occurs in a large variety of cellular processes and functional contexts. Corresponding processes can be as essential and diverse as mRNA splicing, microRNA (miRNA) mediated post-transcriptional regulation or translation. Hence, faithful measurement of direct RNA–RBP interactions constitutes a problem of great importance to various fundamental fields of biology (1–4). Probably the most recent, major implication was recognized in the field of epigenetics, where several non-coding RNAs (ncRNAs) have been shown to affect chromatin landscape through interaction with epigenetic regulators, and thereby control transcriptional activity of entire chromatin domains (5). To study direct RNA–RBP interactions globally and at high resolution, the Photo-Activatable Ribonucleoside-enhanced CrossLinking and Immuno Precipitation (PAR-CLIP) method has been developed (6). In this method, cells are cultured with 4-thiouridine (4SU) or 6-thioguanosine (6SG) ribonucleoside analogues, which are incorporated into nascent RNA molecules. Subsequent to incorporation, *in vivo* UV crosslinking introduces a covalent bond between the base analogue and a proximal amino acid residue of the interacting protein. The covalently linked RNA–RBP complex is isolated and the protein is removed, rendering RNA templates, which are employed for reverse transcription-mediated complementary DNA (cDNA) library generation. The cDNA library sequence content is determined using next-generation sequencing (NGS) technologies, resulting in a number of short reads. A crucial feature of PAR-CLIP is the UV-dependent induction of specific transitions observable in the short reads. The type of transition depends on the base analogue provided: 4SU and 6SG cause T to C or G to A transitions, respectively. RNA nucleotide positions engaging in the covalent bond with the nearby amino acid residue of interacting proteins exhibit transitions with increased probability, likely to be caused by incorrect reverse transcription (6). The consideration of transitions enables the detection of high confidence interaction sites. However, observed transitions may be caused by a variety

*To whom correspondence should be addressed. Tel: +41 61 387 31 20; Fax: +41 61 387 39 96; Email: renato.paro@bsse.ethz.ch

© The Author(s) 2012. 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.

of reasons besides the result of a crosslink. These include: (i) Sequencing errors intrinsic to any currently available NGS platform (7). (ii) Contamination with external RNA possibly introduced by the usage of recombinantly produced enzymes during the experimental procedure. This problem arises, if corresponding reads are similar to any subsequence of the reference genome such that alignments are still valid, while mismatches appear as substitutions. Existing tools account for this problem by removing all reads which can be aligned to a selected set of genomic sequences including known bacterial genomes (8). Another approach performs corrections based on the assumption that binding sites occur in sense orientation of annotated regions only (9). (iii) Cell line specific pre-existing genetic variation, such as single nucleotide polymorphisms (SNPs). Short reads originating from corresponding sites exhibit systematic differences with respect to the reference genome, bearing the risk of misinterpretation. The genetic background varies among cell lines and is not known a priori. This problem is not considered by existing tools.

Independent of the source, non-experimentally induced transitions increase the risk of false positives at worst leading to wrong conclusions and time-consuming, unsuccessful validation attempts. In order to account for these problems, without making prior assumptions, we developed a non-parametric two-component mixture model that distinguishes between experimentally and non-experimentally induced transitions and identifies transition frequencies most affected by PAR-CLIP.

Another challenge remains the accurate resolution of clusters. Clusters represent genomic regions encoding for the protein binding site within the corresponding transcript and were previously defined as contiguous regions of non-zero coverage (8). The resolution of clusters can be difficult especially when multiple binding sites localize in close proximity on the same RNA molecule, where they spuriously appear as single site. Highly resolved binding sites can lead to a better characterization of the RNA–protein interaction (e.g. by improving results of motif search), or might be important for the deduction of protein complex structure, whenever binding information of complex components is integrated. For this reason, our algorithm exploits geometric properties of the coverage function, which can be defined as the number of aligned reads as a function of the genomic position. Binding sites of known RBPs, as detected by this method, resemble sharply peaking rectangle functions (6). This information was taken into account using the continuous wavelet transform (CWT), which provides an efficient way to compute local signal-to-noise ratios and thereby detects corresponding peaks.

We employed this method to study global RNA binding characteristics of the protein Moloney leukemia virus 10 (MOV10), a putative RNA helicase known to be involved in the miRNA pathway through interaction with RNA-induced silencing complex (10). More recently, MOV10 was recognized to be involved in Polycomb-mediated regulation of the *INK4a/ARF/INK4b* tumor suppressor locus, relevant in various cancer (11). MOV10 presumably facilitates direct interaction between ncRNA *ANRIL* and the Polycomb protein CBX7 (12) during recruitment. To investigate global involvement of MOV10 in RNA-dependent chromatin regulation, a modified PAR-CLIP method was applied to the nuclear fraction of HEK293 cells. Our method identifies high confidence interaction sites providing a faithful representation of the MOV10 binding profile and thereby reflecting binding preferences.

MATERIALS AND METHODS

Modified PAR-CLIP method

Using the Invitrogen Flp-In T-REx system, HA-Streptavidin tagged MOV10 was expressed in HEK293 cells. To validate expression levels and functionality, tagged and endogenous MOV10 were compared (Supplementary Figure S1a). The PAR-CLIP protocol by (6) was modified to allow for a nuclear isolation step prior to the immunoprecipitation as well as the use of the Streptavidin tag.

After 365 nm crosslinking, the cells were harvested, washed with cold PBS and pelleted at 500 rcf for 5 min. The supernatant was removed and every 6×10^6 cells were resuspended in a cold mixture of 3 ml of PBS, 9 ml of mQ water and 3 ml of Nuclear Isolation Buffer (1.28 M Sucrose, 40 mM Tris-HCl pH 7.5, 20 mM MgCl2, 4 % Triton X-100) with Aprotinin 3.3 µg/ml, Leupeptin 10 µg/ml, Pepstatin 4 µg/ml added. The resulting suspension was rolled for 10 min at 4°C and subsequently spun at 2500 rcf for 10 min. After aspirating the supernatant, nuclei were resuspended in 1 ml of Lysis Buffer (50 mM HEPES pH 7.5, 150 mM KCl, 1 mM NaF, 1% NP40, Roche Complete Protease Inhibitors) for every 6×10^6 cells. The resuspended nuclei were sonicated in a diagenode Bioruptor for 15 cycles of 30 s on and 30 s off. Except for minor changes, all subsequent steps followed the protocol in (6).

Nuclear RNA-sequencing

Nuclei were extracted from cells treated with 4SU overnight in the same way as the protocol mentioned above. Nuclear RNA was extracted using TRIzol (Invitrogen) and reverse transcribed using random hexamer, and sequenced using Illumina Hi-Seq.

Bioinformatic processing of PAR-CLIP data

The PAR-CLIP experiment rendered a total of 59.5 million reads, using the Solexa Illumina Genome Analyzer sequencing platform, of which 23.2 million passed the Illumina quality filter. Adapter sequences were removed, reads of length <15 were discarded. Remaining reads were aligned to the human genome assembly 'hg19' using the Bowtie aligner (13). The following parameters varied from default values: –best –chunkmbs 512 -n 1 -S -M 100. A total of 9.1 million reads were aligned to the reference genome. Further processing of short reads was mainly based on samtools (14) and the Python interface pysam (see: http://www.cgat.org/~andreas/documentation/pysam/contents.html

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

(15 July 2012, date last accessed)). MOV10 wavClusters (protein binding sites) exhibit a mean length of 35.7 bases, wavClusters shorter than 12 bases were expanded in either direction to the minimum length of 12 bases (15). The MEME motif search (16) within the binding sites was done using the following parameters: -mod anr -maxsize 1000000 -nmotifs 30 -minw 12 -maxw 20 -dna. The wavelet-based peak calling was done using the `R` package `wmtsa` [see Constantine,W. and Percival,P. (2011) `wmtsa`: Wavelet Methods for Time Series Analysis. R package version 1.1-1]. The following parameter settings were used in the indicated functions: wavelet = 'gaussian2' (`wavCWT`), n.octave.min = 2 (`wavCWTTree`), noise.span = 0 and snr.min = 3 (`wavCWTPeaks`).

Implementation of the R package wavClusteR

We implemented the algorithm as a package for the statistical environment `R` (17) (details discussed below). The package requires a sorted BAM file as input and allows users with little experience in `R` to perform the analysis. The obtained wavClusters can be exported for visualization. `wavClusteR` supports multicore parallel computing if available and has been implemented accounting for modularity. Hence, advanced users can integrate `wavClusteR` functions in their own pipelines. The package will be made available along with the relevant documentation at: http://www.bsse.ethz.ch/egg/software/index (16 July 2012, date last accessed).

Co-immunoprecipitation

Nuclei of HEK293 cells were prepared as per the PAR-CLIP protocol and lysed in Co-immunoprecipitation (CoIP) buffer (50mM Tris-HCl pH 8, 150 mM NaCl, 1mM EDTA, 1% Triton X-100) for 2 h at 4°. The samples were then centrifuged for 15 min at 20 000*g* in cold to get rid of debris. Supernatant was incubated with 5µg of MOV10 antibody (ab80613, Abcam) overnight with or without 2 units of RNase (Roche) per 500 µl of suspension. To pull down the complexes, the solution was incubated with Protein A dynabeads (Invitrogen) for 3 h in cold. Immunoprecipitates were washed with CoIP buffer three times before eluting the complexes in SDS-gel loading buffer. The immunoblots were probed with anti-Argonaute 2 (AGO2) antibody (18).

RESULTS

PAR-CLIP induced transitions preferentially occur at specific frequencies

To study the function of MOV10, we generated MOV10 PAR-CLIP data using nuclei derived from HEK293 cells. Purity of the nuclear fraction taken for PAR-CLIP was confirmed by fractionation experiments (Supplementary Figure S1). The cells were cultured with 4SU (Supplementary Materials and Methods), important information regarding protein binding sites is therefore contained in T to C transitions. To identify PAR-CLIP specific transition frequencies, we calculated relative substitution frequencies (RSFs) for each substitution, i.e. the sum of observed substitutions within aligned reads divided by the total number of aligned reads (coverage) at a particular genomic position. This analysis revealed that a number of genomic sites exhibit relatively high RSF values. Figure 1a shows the total number of genomic positions exhibiting RSFs falling into the interval [0.82, 1], i.e. 82–100% of all reads aligned to these genomic sites showed the specified substitutions.

It can be seen that the number of genomic positions having substitutions within the specified RSF interval is of similar prevalence for all substitutions (Figure 1a), implying that this RSF interval is unaffected by the experimental procedure. Similar results were obtained for RNA-Seq control experiments (Supplementary Figure S2a, Supplementary Materials and Methods), rendering these genomic T to C transition sites highly questionable. Henceforth, genomic positions of relative T to C transitions frequency in [0.82,1] (Figure 1a) will be referred to as high-TC sites. The consideration of the RSF interval [0.1, 0.82) changes proportions substantially (Figure 1b). In this case, the majority is clearly constituted by genomic positions exhibiting T to C transitions. A comparable increase in T to C transitions is missing in the control (Supplementary Figure S2b), confirming PAR-CLIP induced transitions occur at specific frequencies and UV-crosslinking is necessary. This specific partitioning of the RSF intervals maximized the relative difference between the substitution profiles shown in Figure 1a and b. These results indicate the existence of specific RSF subsets for which non-experimental causes dominate, as all substitutions are observed at similar levels across the genome (Figure 1a).

One simple approach for identification of protein binding sites can be taken by ranking all clusters according to absolute number of observed T to C transitions occurring within the cluster in decreasing order (8). Cluster were previously defined as contiguous genomic regions of non-zero coverage, representing possible binding sites (8). To examine the approach, we analyzed how many high-TC sites (Figure 1a) localize within the top 1000 ranked clusters. Figure 1c represents the fraction of top ranked clusters containing a high-TC site normalized to number of clusters. About 24% of top 100 ranked MOV10 clusters included a high-TC site. Closer examination suggested two potent sources: (i) RNA contamination and (ii) cell line-specific SNPs. Examples corresponding to contaminations are shown in Figure 1d and Supplementary Figure S2c, as no reads were obtained in nuclear RNA-Seq experiments, indicating the transcriptional inactivity of the locus. High-TC sites resulting from cell-type specific SNPs are shown in Figure 1e and Supplementary Figure S2d. Corresponding reads reflect genomic alterations, as judged by direct sequencing of genomic DNA or nuclear RNA-Seq experiments. The large fraction of reads exhibiting transitions is consistently observed, independent of the experimental source. Here, we sought a way to distinguish experimentally and non-experimentally induced T to C transitions without prior assumptions about the source. For this purpose, we developed a non-parametric two-component mixture model exploiting the RSF differences shown in Figure 1a and b.

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

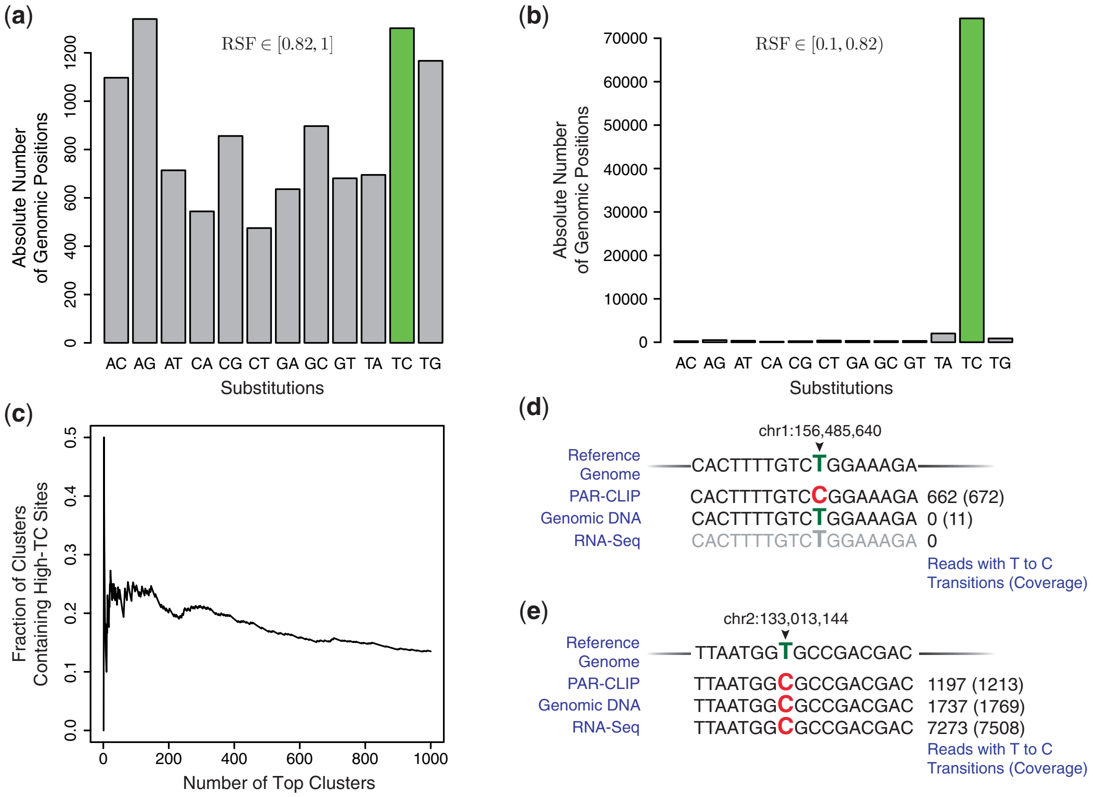

Figure 1. PAR-CLIP induces transitions at specific frequencies. MOV10 PAR-CLIP data analysis: (**a**, **b**) Absolute number of genomic sites exhibiting specified substitutions within the RSF intervals specified in the figure. To obtain reasonable estimates on RSFs, only genomic positions of coverage ≥20 were considered. (**c**) All clusters were ranked according to the absolute number of T to C transitions. Vertical axis represents the number of clusters containing high-TC sites (a) normalized to the total number of clusters being considered (horizontal axis). (**d**) Example of a high-TC site likely to be the result of external RNA contamination. Genomic position is indicated on top. Total number of all aligned reads (in brackets) and observed T to C transitions are shown below. Experimental sources are indicated on the left. PAR-CLIP: reads obtained from MOV10 PAR-CLIP experiments. Genomic DNA: reads obtained from pooling multiple RNA-Seq and ChIP-Seq experiments performed in the same cell line (unpublished data). Since no substitutions are induced experimentally, majority of reads correspond to the actual genomic sequence and can be used to determine SNPs. RNA-Seq: reads obtained from nuclear RNA-Seq control experiments (Supplementary Materials and Methods). (**e**) Example of a high-TC site likely to be the result of a HEK293-specific SNP. Annotations are the same as in Figure 1d.

Defining the mixture model

Let $\mathcal{A} = \{A, C, G, T\}$ be the nucleotide alphabet and $\mathcal{S} = \{(g, r)|g, r \in \mathcal{A} \wedge g \neq r\}$ be the set of substitutions of any base g in the reference genome to any other base r contained in the mapped read. The RSF can be quantified as:

$$\hat{x}_{s,i} = \frac{y_{s,i}}{z_i},\ s \in \mathcal{S}$$

where $y_{s,i}$ indicates the total number of observed substitutions s at position i and z_i represents the total coverage at position i. Since the largest proportion of genomic positions with at least one substitution shows only one particular kind of substitution (Supplementary Figure S3), it was assumed at any position the number of any substitution can be regarded as independent binomially distributed random variable $y_{s,i} \sim \text{Bin}(z_i, x_{s,i})$ parametrized by sample size z_i and probability $x_{s,i}$. Consequently, $\hat{x}_{s,i}$ represents the maximum likelihood estimate (MLE) of $x_{s,i}$. Since the variance of the MLE is a function of the coverage it was required that $z_i \geq c$, where $c = 20$ was chosen in this study, as regions of low coverage will give rise to MLE with high variance. The choice of this value might be adjusted to the sequencing depth of the data set and hence represents a tradeoff between variance and recall. Considering all genomic positions exhibiting a particular substitution, the parameter x_s will be distributed according to some probability density function (PDF) p_s, $x_s \sim \mathrm{p}_s$. It can be assumed that all observed substitutions may either result from PAR-CLIP specific experimental induction or any other non-experimental causes (e.g. sequencing errors, contamination, SNPs). Therefore, p_s can be expressed as mixture of two components

$$\mathrm{p}_s(x) = \lambda_{s,1}\mathrm{p}_{s,1}(x) + \lambda_{s,2}\mathrm{p}_{s,2}(x) \qquad (1)$$

subject to $\lambda_{s,k} \geq 0$ and normalization $\sum_k \lambda_{s,k} = 1$, where k indicates the component (Supplementary Text S1 provides a brief introduction to mixture models). Here, the first component accounts for non-experimentally induced substitutions, whereas the second component models experimentally induced substitutions. A reasonable assumption is that all non-experimentally induced substitutions have approximately the same distribution, therefore,

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

$p_{s,1}(x) = p_1(x)$. PAR-CLIP induces T to C transition only. Hence, $\lambda_{n,2} = 0$, $\forall n \in \mathcal{N}$ where $\mathcal{N} = \mathcal{S}\backslash(T, C)$ is the set of all non-T to C substitutions. Equation (1) simplifies for those instances to:

$$p_n(x) = p_1(x). \tag{2}$$

Consequently, the second component only exists within the following expression:

$$p_{TC}(x) = \lambda_1 p_1(x) + \lambda_2 p_2(x). \tag{3}$$

The problem of identifying a subset within the support of $p_{TC}(x)$ dominated by $\lambda_2 p_2(x)$, and therefore likely result from experimental induction, can be treated as binary density classification problem. For this reason, either the posterior class probability:

$$\mathbb{P}(K = 2|X = x) = \frac{\lambda_2 p_2(x)}{\lambda_1 p_1(x) + \lambda_2 p_2(x)} \tag{4}$$

or the log-odds ratio:

$$\log \frac{\mathbb{P}(K = 2|X = x)}{\mathbb{P}(K = 1|X = x)} = \log \frac{\lambda_2 p_2(x)}{\lambda_1 p_1(x)} \tag{5}$$

can be considered. Thus, one is left with the problem of estimating λ_1, λ_2, $p_1(x)$ and $p_2(x)$.

Estimation of the mixing coefficients

To estimate the mixing coefficients, a count function f is introduced:

$$f : \mathcal{S} \to \mathbb{N}_0, f(s) = \sum_{j=1}^{G} \mathbb{1}(s, j)$$

with

$$\mathbb{1}(s, j) = \begin{cases} 1 & \text{if } s \text{ is observed at least once at position } j \\ 0 & \text{otherwise} \end{cases}$$

where G equals the size of the genome. Therefore, $f(s)$ indicates the number of genomic positions exhibiting at least one substitutions s. Figure 2a shows $f(s)$ for the MOV10 data set. Assuming experimentally and non-experimentally induced T to C transitions are additive, as implied by the model, and the number of non-experimentally induced substitutions are approximately equal for all substitutions, λ_2 can be estimated using the following conservative estimator:

$$\hat{\lambda}_2 = \frac{f(TC) - \tilde{f}}{f(TC)}, \tilde{f} = \arg\max_{n \in \mathcal{N}} f(n)$$

From the normalization it follows:

$$\hat{\lambda}_1 = 1 - \hat{\lambda}_2$$

Estimating PDFs using Bayesian inference

Since observed substitution are modeled as part of a binomial process, a Bayesian framework can be employed to estimated the overall density using available observations (Supplementary Text S1 provides a brief introduction to Bayesian inference). Due to its conjugacy property with respect to the binomial likelihood function, a beta prior of the form:

$$\text{Beta}(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

with $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$ was considered. Requiring the prior to be uniform, the hyperparameters were set to α, $\beta = 1$. The resulting posterior PDF takes the form:

$$g(x_{s,i}|y_{s,i}, z_i) = \frac{\Gamma(z_i + 2)}{\Gamma(y_{s,i} + 1)\Gamma(z_i - y_{s,i} + 1)} x_{s,i}^{y_{s,i}}(1 - x_{s,i})^{z_i - y_{s,i}}. \tag{6}$$

Equation (2) states that $\forall n \in \mathcal{N}$ samples were obtained from $p_1(x)$. Defining $\mathcal{D}_s = \{(y_{s,i}, z_i)|1 \leq i \leq f(s)\}$ to be the set of all observations of a particular substitution s and employ the union

$$\mathcal{D}_{\mathcal{N}} = \bigcup_n \mathcal{D}_n, n \in \mathcal{N}$$

a non-parametric estimate of $p_1(x)$ can be obtained using:

$$\hat{p}_1(x) = \frac{1}{N}\sum_{i=1}^{N} g(x_{s,i}|y_{s,i}, z_i), N = |\mathcal{D}_{\mathcal{N}}| \tag{7}$$

This estimator does not assign point mass of one as in case of the MLE. It naturally accounts for different variances arising as a consequence of variable sample sizes. Regions of high coverage will enter the estimation in Equation (7) by means of sharply peaked posterior PDFs of the form of Equation (6), reflecting higher confidence in the parameter estimate. Similarly; $p_{TC}(x)$ in Equation (3) is estimated using $\mathcal{D}_{TC}$. Inserting all estimates, Equation (3) can be solved for:

$$\hat{p}_2(x) = \frac{\hat{p}_{TC}(x) - \hat{\lambda}_1 \hat{p}_1(x)}{\hat{\lambda}_2}$$

Using all results, estimates on Equations (4,5) can be obtained. Figure 2b shows the densities estimates of Equation (3) and log-odds ratio [Equation (5)]. The posterior probability [Equation (4)] of an observation being generated by the experimental component is shown in Figure 2c, large probabilities indicate RSF intervals most affected by the experimental procedure. In principle, the entire domain can be classified according to the Bayes classifier, which assigns any x to the class that maximizes the posterior probability. The number of transitions within true interaction sites is thought to reflect the strength of interaction. To obtain a preferably clear MOV10 binding profile, it was decided to prioritize precision over recall and focus analysis on T to C transitions having RSF values within [0.2, 0.7], as they are likely to represent strong and high confidence interaction sites.

Resolving binding sites using wavelet transforms

Binding sites identified by PAR-CLIP appear as narrow regions exhibiting jump discontinuities and often localize

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

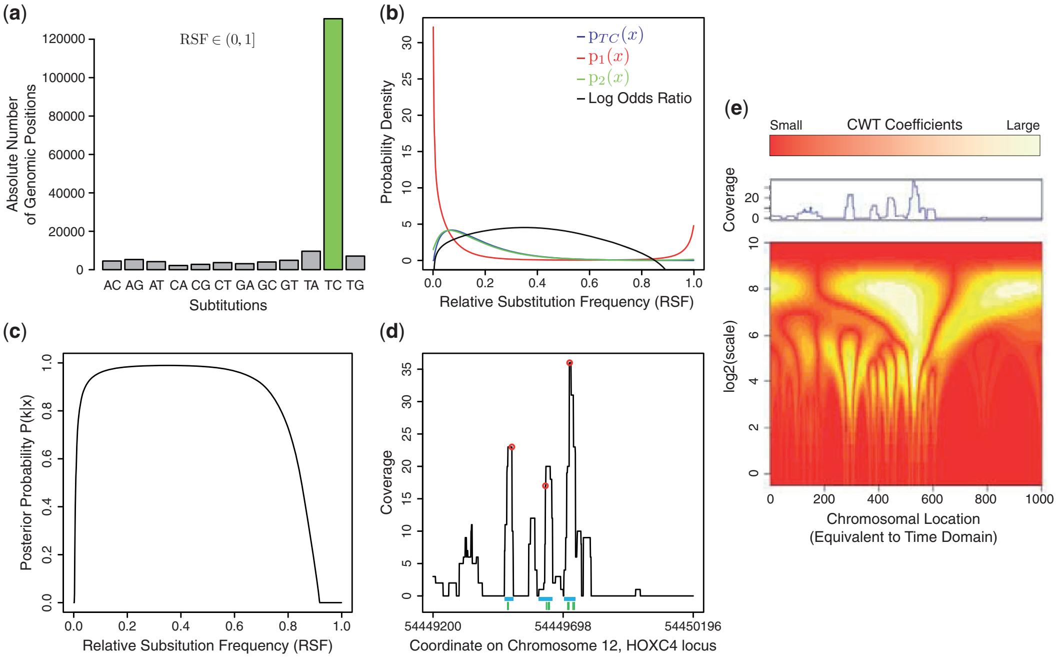

Figure 2. Model fit and wavelet peak calling on the MOV10 PAR-CLIP data. (**a**) The count function $f(s)$ represents the absolute number of genomic positions exhibiting at least one substitution. A minimum coverage $c = 20$ was required. (**b**) The densities estimates [Equation (3)] as well as the log-odds ratio [Equation (5)] which were estimated from the data. Vertical axis represents density as well as log-odds ratio, respectively. (**c**) Estimated posterior probability [Equation (4)] of a given observation belonging to class 2 (experimentally induced transition) computed from the estimated densities and mixing coefficients. (**d**) Coverage function observed within the 3′UTR of the *HOXC4* gene. Ticks (green) below indicate high confidence $T->C$ transitions as determined by the mixture model. Red circles indicate wavelet peaks, horizontal lines (blue) below represent the clusters. (**e**) Continuous wavelet transform of the coverage function shown in Figure 2d, color coding and the corresponding coverage at each position is indicated above.

within broader regions of non-zero coverage. Geometric properties of the coverage function at binding sites can be used for two purposes: (i) proximal binding sites can be resolved and regions exhibiting low signal-to-noise ratio can be excluded, referred to as peak calling and (ii) the coverage function at high confidence interaction sites can be utilized to determine cluster boundaries. Peak calling is performed within the time-scale domain using the CWT of the coverage function. The CWT of a function z is defined as:

$$\mathcal{T}^{wav}z(a,b)=\frac{1}{|\sqrt{a}|}\int z(t)\Psi\left(\frac{t-b}{a}\right)dt$$

with $a > 0$ and $b \in \mathbb{R}$ (19). The CWT represents the inner product of z with a family of wavelets indexed by shift (b) and scale (a) parameter (20) (Supplementary Text S2 provides a brief introduction to wavelet analysis). In this approach, the symmetric mexican hat wavelet defined as:

$$\Psi(t) = (1 - t^2)\exp(-t^2/2)$$

was used, as it was successfully applied in peak detection before (21). To illustrate the two-step algorithm, an exemplary signal is shown in Figure 2d. The corresponding CWT is illustrated in Figure 2e. Prior to peak calling, ridges are identified as local maxima of CWT coefficients connected across scale dimension, localizing within bright, vertical areas shown in Figure 2e. The set of all ridges constitutes the branches of a tree, employed for peak detection, i.e. branches are pruned starting from small scales until a specified signal-to-noise ratio is exceeded. The time coordinate corresponding to the scale coordinate closest to zero is returned as peak location. Red circles in Figure 2d indicate peaks detected by this approach (for more details see documentation of the R package, 'wmtsa'). In order to determine cluster boundaries (clusters correspond to the protein binding site), the difference quotient Δz_i (with $\Delta i = 1$) of the coverage function is considered. The left boundary is obtained as follows: starting from the wavelet peak position, the algorithm determines the left boundary as position of the first positive difference quotient followed by a negative difference quotient, possibly separated by regions of constant coverage. Movements towards larger values in coverage are permissible only if they correspond to first overall change in height, allowing the algorithm to surmount coverage peaks whenever starting sites do not exactly correspond to maxima. Similarly, the right boundary is identified as first positions having a negative difference quotient followed by a positive difference quotient. Horizontal lines (blue) in Figure 2d represent clusters obtained by this rule. To recapitulate, according to our definition, a cluster has to meet the following conditions:

Beginning from a wavelet peak, cluster boundaries, determined as described above, are required to encompass a high confidence RNA–RBP interaction site as determined by the mixture model. Hence, valid clusters

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

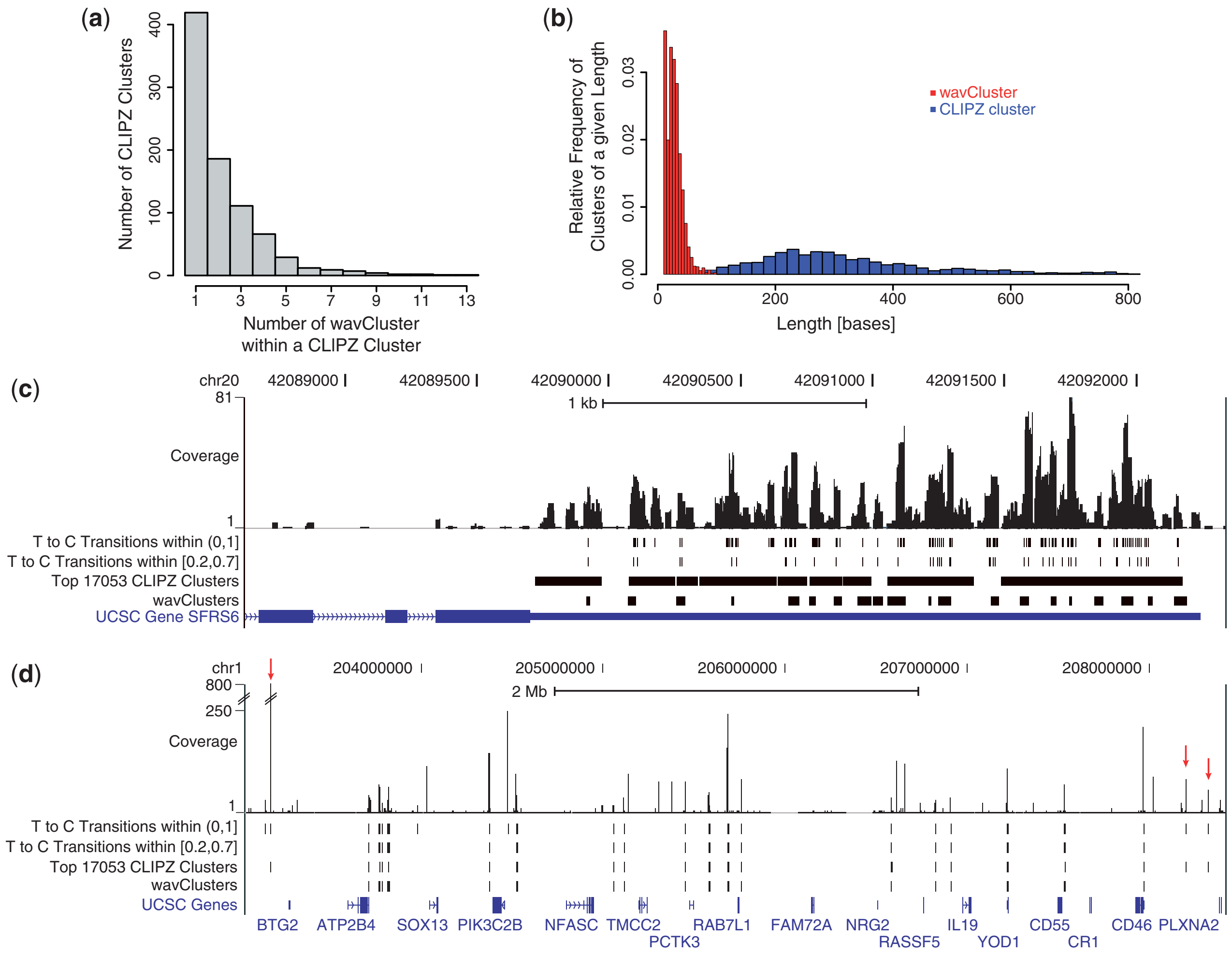

Figure 3. Comparison of CLIPZ cluster and wavClusters. (**a**) Number of wavClusters overlapping with each 841 CLIPZ cluster. (**b**) Length distribution of the 841 out of the top 1000 CLIPZ clusters and the overlapping 2455 wavCluster. (**c**) Exemplary genome browser view corresponding to the genomic location of the 3′UTR of a the SFRS6 gene. MOV10 coverage function, obtained from the PAR-CLIP data, is shown as 'Coverage'. Tracks below indicate positions exhibiting at least one T to C transition, high confidence interaction sites determined by mixture model, the top 17053 CLIPZ clusters, wavClusters and the genomic annotation. (**d**) Broader genome browser view. Labeling is the same as in Figure 3c. Red arrows indicate false positives called by CLIPZ, as they localize to untranscribed regions.

contain at least one high confidence interaction site and a wavelet peak. To distinguish different cluster definitions, clusters obtained by this two-step algorithm will be referred to as wavClusters following the nomenclature introduced by the R package 'wmtsa'.

In total, 17053 MOV10 wavClusters were identified across the transcriptome. To contrast wavClusters with conventional cluster definitions, the same data set was processed using CLIPZ (8). Comparison reveals substantial overlap between the results. 841 out of the top 1000 CLIPZ clusters (sorted according to absolute T to C transitions in decreasing order) overlapped with 2455 wavClusters, indicating that binding sites are more resolved in the latter. Hence, CLIPZ clusters contain more than one wavCluster on average (Figure 3a), which have shorter mean length (Figure 3b). The example illustrated in Figure 3c shows that CLIPZ clusters cover broader regions, whereas wavClusters usually span relatively narrow stretches corresponding to peaks in the coverage function. Figure 3d shows a broader genomic locus to emphasize the overall correspondence of the results. The high coverage peaks close to the boundaries (red arrows), called by CLIPZ, are likely to be false positives, as these regions are silent according to the nuclear RNA-Seq control experiments (Supplementary Materials and Methods).

Analysis of MOV10 binding sites

Analysis of all 17053 wavClusters revealed that 88.9% localized to annotated genomic regions. The genomic annotation is shown in Figure 4a. The largest fraction of wavClusters maps into 3′UTRs in sense orientation. A substantially smaller number falls into coding sequences and only few wavCluster localize within 5′UTRs. wavClusters found in antisense strand orientation of any feature are less abundant. The small mean length of MOV10 wavClusters (Figure 4b) led us to perform sequence analysis, which is usually more efficient on shorter sequences. We analyzed whether bound RNAs

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

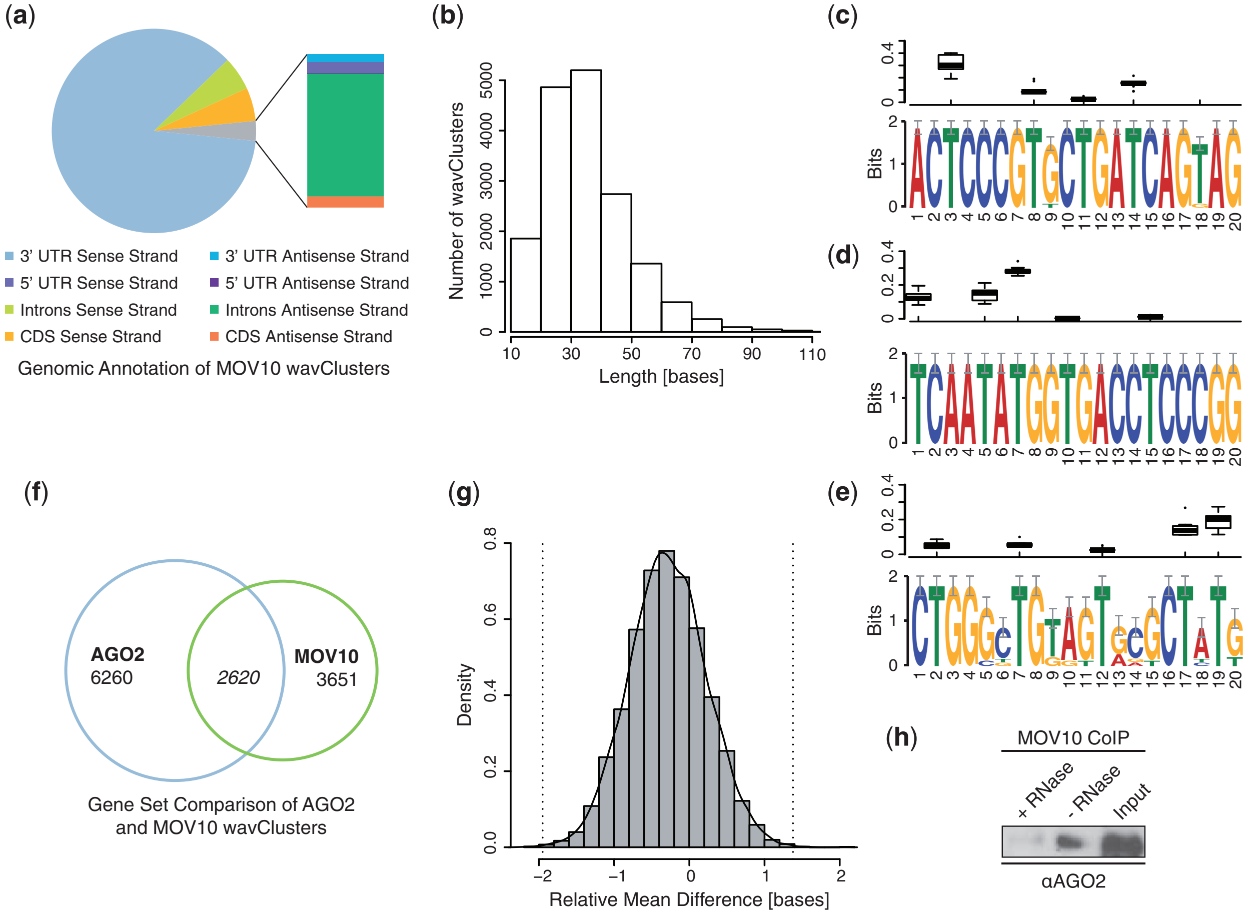

Figure 4. Analysis of MOV10 and AGO2 wavClusters. (**a**) Genomic annotation of MOV10 wavClusters. Only wavCluster, which unambiguously localize to one genomic feature only were considered. (**b**) Length distribution of MOV10 wavClusters, mean length = 35.7 bases. (**c–e**) MEME results: logos of the top three motifs ranked increasingly by E-values (1.1e − 74, 6.3e − 47 and 1.2e − 25, respectively). Vertical axis represents information content in bits. The distribution of T to C RSF values for a given position over all motif occurrences is shown above each motif logo. (**f**) Venn diagram indicating the number of genes bound by AGO2, MOV10 or both proteins. Overlap is significant according to hypergeometric testing (*P*-value < 2.2e − 16). (**g**) Bootstrap distribution of mean wavCluster center difference of overlapping AGO2 and MOV10 wavCluster. 10 000 bootstrap samples were computed. Dashed vertical lines (−1.95, 1.38) indicate bootstrap confidence intervals considering a significance level of 0.001. Solid line represents the kernel density estimate. (**h**) Co-immunoprecipitation in nuclear extract of HEK293 cells (Input). MOV10 antibody was used for the immunoprecipitation. Western blotting was done using AGO2 antibody. Samples were prepared with and without RNase treatment as indicated in the figure.

showed primary sequences similarities possibly reflecting general MOV10 recruitment mechanisms. Motif search (16) performed on all wavCluster sequences resulted in 30 significant motifs reported by the algorithm. Figure 4c–e shows the three motifs having lowest E-values. Motif occurrences are highly similar between wavCluster sequences as indicated by the high information content. However, despite this conservation and the high significance assigned by the algorithm, reported motifs occur only in small subsets of all sequences (number of occurrences of motifs in Figure 4c–e: 14, 10, 10, respectively) indicating primary sequence features are not crucial regarding general recruitment, possibly involving secondary or tertiary structures instead. In order to identify the interacting positions in the motifs, the T to C RSF distribution over all motif occurrences was considered (Figure 4c–e). Certain positions within the motif are more likely to engage in RNA–protein interaction, as the strength of interaction is reflected in the magnitude of the RSF values.

The large number of MOV10 clusters within 3′UTRs appears to reflect previously described functional involvement in the miRNA pathway. To investigate this further, published AGO2 PAR-CLIP data (22) integrated into the analysis and processed in the same way (Supplementary Figure S4a,b). A significant overlap among MOV10 and AGO2 target genes (Figure 4f) was observed substantiating this assumption. Functional GO term enrichment analysis of the gene set bound by AGO2 and MOV10 indicates a general involvement in regulation of transcription (Supplementary Table S1). We further looked for systematic differences between adjacent AGO2 and MOV10 binding sites, potentially conveying information regarding RNA interactions or complex structures. The mean cluster center (mean of start and end position rounded to the closest integer) difference of overlapping wavClusters was determined to be −0.3 bases (MOV10 center position subtracted from AGO2 center position). The distribution of this statistic was estimated using a non-parametric bootstrap (23) (Figure 4g).

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

The results do not suggest that there are systematic differences within adjacent binding sites. The distribution of the mean start site difference and mean end site difference (Supplementary Figure S4e and d) suggests that on average, MOV10 wavClusters are enclosed within AGO2 wavClusters, which are fractionally longer in mean length (36.6 bases). It appears as if MOV10 and AGO2 bind to the same positions in the sequence. In addition, a HEK293 miRNA expression profile, providing quantitative information regarding miRNAs expression within this cell line, was obtained (24). Supplementary Table S2 lists all expressed miRNAs and indicates MOV10 and AGO2 binding as judged by presence of wavClusters. It can be seen that a large number (30 out of 33) of miRNAs exhibits AGO2 wavClusters, whereas the number of MOV10 wavClusters is considerably smaller (4 out of 33). This suggests that wavClusters accurately detect known miRNA–AGO2 interactions. To test for physical interaction of MOV10 and AGO2 in the nucleus, CoIP using nuclear extract was performed (Figure 4h). RNase treatment of the sample diminishes the interaction implying that the complex contains an RNA component required for interaction.

DISCUSSION

PAR-CLIP offers great potential to faithfully study and characterize RNA–RBP interactions. The strength of this method, the specific induction of transitions, is accompanied by challenges regarding the discrimination of signal and noise. In this work we present a two-step algorithm, employing different sources of information available in PAR-CLIP data, to identify high confidence interaction sites at high resolution. The mixture model identifies most affected substitution frequencies and thereby discriminates high confidence interaction sites from non-experimentally induced transitions. One major problem in this respect appears to be contamination with external RNA. This problem can be approached by exhaustively aligning short reads to selected genomes prior to analysis (8). One major disadvantage of this approach is the requirement of complete knowledge regarding genomic sequences of possible contaminants. Lack of information translates into leakiness of this filter and consequently increases the risk of false positives. In addition, pre-existing genetic variation is not correctable in this way. All spurious interaction sites discussed in Figure 1d and e; Supplementary Figure S2c and d localize within the top 1000 CLIPZ clusters underlining the occurence of this problem. Another proposed way for PAR-CLIP data analysis defines cutoffs, based on the assumption that clusters occur in known transcripts (9). Making such strong assumptions can be problematic whenever proteins interact with RNAs, which are mainly unknown. Since MOV10 binding preferences were not known in advance, this approach was not applicable in this setting. The PARalyzer software (15) does not explicitly account for non-experimental transitions and read filtering is not performed in advance. However, PARalyzer implicitly controls this error in part by hard-thresholding i.e. only considering 'read groups' (equivalent to cluster) with at least two transition sites, in this way decreasing the probability of false positive detection.

The mixture model developed here is not restricted to PAR-CLIP data. In principle, it can be applied to other substitution-inducing NGS-based methods, since it exploits intrinsic information in terms of unaffected substitutions to perform background correction. Recently, a base-pair-resolved genome-wide cytosine methylation map (methylome) was generated using whole-genome bisulphite sequencing (BisSeq) (25). Bisulphite treatment of DNA causes specific transitions of cytosine to uracil, whereas methylated cytosines remain unaffected. Consequently, transitions, detectable using DNA sequencing, comprise information regarding methylation status. BisSeq data analysis presents similar challenges, such as discriminating SNPs from signal, which was accounted for by complete reconstruction of the cell-type specific genotype using NGS (25). The disadvantages of genome-sequencing based background correction comprise large experimental expenses as well as the restricted applicability to the particular cell culture. BisSeq data analysis constitutes a potential application of the mixture model described here and similar NGS-based experimental procedures are likely to be developed in the near future.

Subsequent to the mixture model, the coverage function at potential interaction sites is transformed using the CWT. This representation is suitable to perform peak calling and identify sharp peaks within broader regions of non-zero coverage (Figure 3c), detecting binding sites at higher resolution. Optimal parameter settings for a particular data set are in general difficult to assess in the absence of a positive training set. Wavelet peak calling was therefore performed using less stringent conditions, as stringency was imposed by the mixture model beforehand, where this problem did not apply. Although wavelets have been successfully employed for detection of mass spectrometry peaks (21), this algorithm constitutes, to our knowledge, the first application of wavelets to any type of NGS data. The short length (Figure 4b) of detected wavClusters increases resolution at which protein binding sites are detected. This additional information can be advantageous for an improved characterization of the RNA–RBP interactions or for the inference of protein complex structure. For example, examining the binding sites of the snRNPs, which constitute the splicosome, on the pre-mRNA, might provide a better insight into the process of splicing, wheras a coarser resolution of binding sites could impede such an analysis. PARalyzer detects clusters of similar size (15), but uses transitions to determine the exact cluster location and not the coverage function. This could be misleading whenever interacting nucleotides (Ts or Gs) are not centrally located within binding sites. In those cases, the coverage might be more appropriate. Determining cluster boundaries using the difference quotient can cause problems in regions where the coverage function fluctuates. This can cause the algorithm to stop untimely and return boundaries located within the wavCluster. Smoothing the coverage function using

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

wavelet or other filtering methods might be a possible solution. However, such procedures presents different challenges like determining reasonable degrees of smoothing, while preventing the signal from abolishing. Hence, it was decided to omit a smoothing step.

Analysis of the MOV10 wavCluster revealed a preferential binding to 3′UTRs of annotated genes, probably reflecting the previously reported functional involvement in the miRNA pathway (10). This is further substantiated by significant co-localization with AGO2 at ~2600 genes, where MOV10 possibly facilitates target recognition by affecting accessibility. As the major proportion of MOV10 localizes to 3′UTRs, association with non-coding or unannotated regions, possibly representing components of the epigenetic regulatory system, is less frequent. The proposed involvement of MOV10 in CBX7 and *ANRIL* mediated regulation of the *INK4a/ARF/INK4b* tumor suppressor locus (12) was not detected, probably explained by use of a different cell systems.

The modified PAR-CLIP method presented in this work represents a general approach to study nuclear or chromatin-associated RNA–RBP interactions. A promising extension might be achieved by applying the modified PAR-CLIP procedure in conjunction with ChIP-Seq experiments and thereby measuring RNA interaction and protein binding across the genome. Integration of such data sets could reveal precise information regarding *cis*-interaction at chromatin. Hence, the modified PAR-CLIP method developed here can be used to study interactions in the nucleus with high precision and confidence. In addition, we present an algorithm of general design, which is applicable to existing or future substitution-inducing NGS-based data and therefore of potential interest to scientists of various fields.

ACCESSION NUMBERS

GSE37524.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–2, Supplementary Figures 1–4, Supplementary Materials and Methods, Supplementary Texts 1–2 and Supplementary References [19,25–27].

ACKNOWLEDGEMENTS

We thank Lukasz Jaskiewicz, Shivendra Kishore and Mihaela Zavolan for all their help and support with PAR-CLIP. We also thank Pejman Mohammadi for helpful discussion and ideas, Christian Beisel and Makiko Seimiya for the cell lines and Ina Nissen for sequencing, which was performed at the Quantitative Genomics Facility at the Department of Biosystems Science and Engineering.

FUNDING

Funding for open access charge: Swiss Federal Institute of Technology Zurich.

Conflict of interest statement. None declared.

REFERENCES

1. Jackson,R.J., Hellen,C.U.T. and Pestova,T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Bio.*, **11**, 113–127.
2. Matlin,A.J., Clark,F. and Smith,C.W.J. (2005) Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Bio.*, **6**, 386–398.
3. Ghildiyal,M. and Zamore,P.D. (2009) Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.*, **10**, 94–108.
4. Konig,J., Zarnack,K., Luscombe,N.M. and Ule,J. (2012) Protein-RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.*, **13**, 77–83.
5. Wang,K.C. and Chang,H.Y. (2011) Molecular mechanisms of long noncoding RNAs. *Cell*, **43**, 904–914.
6. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M., Jungkamp,A., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
7. Zagordi,O., Klein,R., DŁumer,M. and Beerenwinkel,N. (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res.*, **38**, 7400–7409.
8. Khorshid,M., Rodak,C. and Zavolan,M. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **39**, D245–D252.
9. Lebedeva,S., Jens,M., Theil,K., Schwanhäusser,B., Selbach,M., Landthaler,M. and Rajewsky,N. (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell*, **43**, 340–352.
10. Chendrimada,T.P., Finn,K.J., Ji,X., Baillat,D., Gregory,R.I., Liebhaber,S.A., Pasquinelli,A.E. and Shiekhattar,R. (2007) MicroRNA silencing through RISC recruitment of eIF6. *Nature*, **447**, 823–828.
11. Kim,W.Y. and Sharpless,N.E. (2006) The regulation of INK4/ARF in cancer and aging. *Cell*, **127**, 265–272.
12. Yap,K.L., Li,S., Muoz-Cabello,A.M., Raguz,S., Zeng,L., Mujtaba,S., Gil,J., Walsh,M.J. and Zhou,M. (2010) Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by Polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell*, **38**, 662–674.
13. 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.
14. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079, 1000 Genome Project Data Processing Subgroup.
15. Corcoran,D.L., Georgiev,S., Mukherjee,N., Gottwein,E., Skalsky,R.L., Keene,J.D. and Ohler,U. (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, **12**, R79.
16. Bailey,T.C. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Inte. Conf. Intel. Syst. Mol. Biol*, **2**, 28–36.
17. R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
18. Ruedel,S., Flatley,A., Weinmann,L., Kremmer,E. and Meister,G. (2008) A multifunctional human Argonaute2-specific monoclonal antibody. *RNA*, **14**, 1244–1253.
19. Daubechies,I. (1992) *Ten Lectures on Wavelets*. SIAM, PA, USA.

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020

20. Li,P. (2003) Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, **19**, 2–9.
21. Du,P., Kibbe,W.A. and Lin,S.M. (2011) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, **22**, 2059–2065.
22. Kishore,S., Jaskiewicz,L., Burger,L., Hausser,J., Khorshid,M. and Zavolan,M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods*, **8**, 559–564.
23. Efron,B. and Tibshirani,R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, **1**, 54–75.
24. Hausser,J., Berninger,P., Rodak,C., Jantscher,Y., Wirth,S. and Zavolan,M. (2009) MirZ: an integrated microRNA expression atlas and target prediction resource. *Nucleic Acids Res.*, **37**, W266–W272.
25. Stadler,M.B., Murr,R., Burger,L., Ivanek,R., Lienert,F., Schler,A., Wirbelauer,C., Oakeley,E.J., Gaidatzis,D., Tiwari,V.K. *et al.* (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490–495.
26. Bishop,C. (2007) *Pattern Recognition and Machine Learning*. Springer, NY, USA.
27. Hastie,T., Tibshirani,R. and Friedman,J. (2007) *The Elements of Statistical Learning*. Springer, NY, USA.
28. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.

Downloaded from https://academic.oup.com/nar/article-abstract/40/20/e160/2414476 by MPI Immunobiology user on 29 June 2020