Bioinformatics, 32(7), 2016, 976–983
doi: 10.1093/bioinformatics/btv520
Advance Access Publication Date: 5 September 2015
Original Paper

OXFORD

Genome analysis

BMix: probabilistic modeling of occurring substitutions in PAR-CLIP data

Monica Golumbeanu[1,2,†], Pejman Mohammadi[1,2,†] and Niko Beerenwinkel[1,2,*]

[1]Department of Biosystems Science and Engineering and [2]SIB Swiss Institute of Bioinformatics, CH-4058 Basel, Switzerland

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Benjamin Raphael

Received on March 19, 2015; revised on July 24, 2015; accepted on August 18, 2015

Abstract

Motivation: Photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) is an experimental method based on next-generation sequencing for identifying the RNA interaction sites of a given protein. The method deliberately inserts T-to-C substitutions at the RNA-protein interaction sites, which provides a second layer of evidence compared with other CLIP methods. However, the experiment includes several sources of noise which cause both low-frequency errors and spurious high-frequency alterations. Therefore, rigorous statistical analysis is required in order to separate true T-to-C base changes, following cross-linking, from noise. So far, most of the existing PAR-CLIP data analysis methods focus on discarding the low-frequency errors and rely on high-frequency substitutions to report binding sites, not taking into account the possibility of high-frequency false positive substitutions.

Results: Here, we introduce *BMix*, a new probabilistic method which explicitly accounts for the sources of noise in PAR-CLIP data and distinguishes cross-link induced T-to-C substitutions from low and high-frequency erroneous alterations. We demonstrate the superior speed and accuracy of our method compared with existing approaches on both simulated and real, publicly available human datasets.

Availability and implementation: The model is freely accessible within the BMix toolbox at www.cbg.bsse.ethz.ch/software/BMix, available for Matlab and R.

Supplementary information: Supplementary data is available at *Bioinformatics* online.

Contact: niko.beerenwinkel@bsse.ethz.ch

1 Introduction

RNA molecules interact with proteins and form ribonucleoprotein complexes actively involved in a plethora of essential biological processes such as translational regulation, alternative splicing or RNA transport (Lunde *et al.*, 2007; Muller-McNicoll *et al.*, 2013). A well-known example of RNA-binding proteins (RBPs) consists of members of the Argonaute family, components of the RNA-induced silencing complex (RISC), which bind to diverse small RNA molecules and regulate gene silencing (Meister, 2013). Gerstberger *et al.* (2014) report 1542 RBPs in humans, many of these having been found dysregulated in diseases including cancer (Kechavarzi *et al.*, 2014). Therefore, characterizing the interactions between RNA and RBPs represents an important step towards understanding RNA function.

High-throughput sequencing technology allows querying the binding sites of a specific RBP in a transcriptome-wide fashion (Blencowe *et al.*, 2009; Kloetgen *et al.*, 2014). One of the recently developed high-throughput sequencing-based experimental

protocols aiming to identify the binding sites of RBPs throughout the transcriptome is Photo-Activatable Ribonucleoside-enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) (Hafner *et al.*, 2010). According to this method (Fig. 1A), a synthetic photoactivatable ribonucleoside such as ^{4S}U (4-thiouridine) or, less commonly, ^{6S}G (6-thioguanosine) is integrated into the RNA of cultured cells. Upon exposure to ultraviolet (UV) light, cross-linking of RBPs to RNA occurs. The cross-linked RNA-RBP pairs are subsequently isolated using immunoprecipitation with an antibody targeting the protein of interest, and the RNA fragment is retrieved upon protein digestion. A complementary DNA (cDNA) sequencing library is generated by reverse complementing the RNA fragments. Due to the incorporated nucleoside, systematic T-to-C (for ^{4S}U) or G-to-A (for ^{6S}G) substitutions appear in the cDNA library at the interaction sites (Hafner *et al.*, 2010). Therefore, PAR-CLIP brings the advantage of having an additional layer of evidence by introducing specific base changes at the binding sites.

PAR-CLIP data are characterized by prevalent T-to-C substitutions observed at different mismatch frequencies (Hafner *et al.*, 2010). However, compared with RNA-Seq, PAR-CLIP has been observed to also introduce a large number of other substitutions, different from T-to-C, notably at low and high mismatch frequency (see Section 3). This indicates the presence of noise and contamination in the PAR-CLIP procedure which can as well introduce erroneous T-to-C substitutions, with high potential to be mistaken for true cross-link alterations. Therefore, of great importance in PAR-CLIP data analysis is discarding the low-frequency sequencing errors, as well as high-frequency spurious substitutions.

Currently, there is a handful of methods available for analyzing PAR-CLIP data, trying to identify the RNA-RBP-binding sites using various techniques, such as kernel density estimation (Corcoran *et al.*, 2011), non-parametric mixture models (Sievers *et al.*, 2012; Comoglio *et al.*, 2015), Bayesian hidden Markov models (Yun *et al.*, 2014) and binomial tests (Chen *et al.*, 2014). Some of the available methods focus on analyzing data from one specific type of RBP, such as, e.g. AGO2 PAR-CLIP data (Erhard *et al.*, 2013). However, the non-cross-link, high-frequency substitutions are usually reported within high-confidence-binding sites by most of these methods. To the best of our knowledge, only one of the currently existing PAR-CLIP analysis methods, namely WavClusteR (Sievers *et al.*, 2012; Comoglio *et al.*, 2015), accounts for these substitutions. Sievers *et al.* (2012) show that cross-link loci reside in moderately altered sites and, within the WavClusteR package, use a mixture model based on the relative substitution frequencies to exclude sites with too low or high nucleotide substitution rates. However, the method is not widely used due to complex implementation and long execution time. Additionally, the approach underlying the WavClusteR package does not explicitly model read counts within the genome and eventually uses a fixed cutoff of the substitution frequencies to select high-confidence T-to-C alterations, which leads to a higher error rate, especially in low-coverage regions. Here, we present a novel probabilistic approach, based on a constrained three-component binomial mixture, to explicitly describe substitution counts observed in PAR-CLIP data. The method, coined BMix, uses a maximum likelihood approach to estimate the substitution rates induced by PAR-CLIP and separates both low- and high-frequency erroneous T-to-C alterations from the true cross-link substitutions (Fig. 1B). We perform an exhaustive comparison of BMix to existing approaches and show its increased performance in terms of speed, accuracy and consistency on synthetic and real data.

2 Methods

2.1 Data preprocessing

PAR-CLIP and RNA-Seq reads were clipped from their adapter using the `fastx clipper` tool (http://hannonlab.cshl.edu) and only reads larger than 13 nucleotides were kept for further analysis. PCR duplicates were included in the analysis by all the tested methods. The reads were aligned to the human reference genome assembly *hg19* with the `bowtie` alignment tool version 0.12.9 (Langmead *et al.*, 2009). The same parameters employed by PARalyzer and WavClusteR were used, `-n 1 --best -m 100 -k 1 -l 50` (1 allowed mismatch, only one best alignment is reported, at most 100 allowed alignments per read, and seed length of 50, respectively).

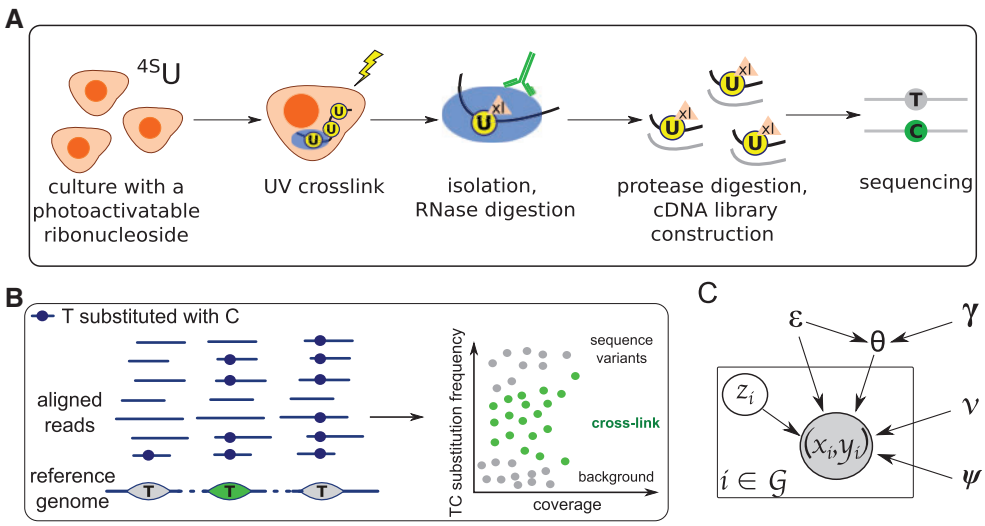

Fig. 1. **(A)** The main steps of the PAR-CLIP protocol including cell culture with 4-thiouridine (^{4S}U), UV cross-link, isolation of the RNA-RBP complexes and sequencing of the bound RNA fragments where systematic T-to-C substitutions occur. **(B)** Schematic representation of the BMix rationale. The method takes as input the aligned sequencing reads and uses a three-component mixture based on the observed substitution counts to infer whether the observed T-to-C substitutions are most likely caused by sequencing errors, sequence variants or cross-link. **(C)** Graphical representation of the statistical model. For each evaluated locus i on the genome $\mathcal{G}$, the coverage x_i and the mismatch count y_i are observed. The latent variable z_i used to indicate the different components of the mixture, and the parameters ϵ, γ, θ, ν and ψ define the model (cf. Section 2)

2.2 Probabilistic modeling of substitution counts

The data consist of PAR-CLIP cDNA sequencing reads. After alignment to the reference genome, at each position in an aligned read, the observed nucleotide can either match or differ from the reference. When the nucleotide differs from the reference, several causes are possible. First, the observed nucleotide could be a sequence variant, caused by single-nucleotide polymorphism (SNP), or foreign, non-cross-linked RNA fragments mapped to a reference location highly similar to their sequence (contamination). Second, if the reference is T and the corresponding read nucleotide is C, then a cross-link-induced substitution could have occurred. Third, a mismatch could have also happened due to sequencing error. These three events are not mutually exclusive.

In order to detect RNA-protein cross-link-induced T-to-C substitutions, we model, for each position i in the genome, where the reference, r_i, is different from C, the probability of the observed T-to-C, A-to-C or G-to-C substitution. We define x_i as the sequencing coverage at position i, and y_i as the number of times the reference nucleotide is substituted with C in all the reads covering position i. We assume that the observed T-to-C alterations are due to (i) sequencing error, (ii) SNPs or contamination or (iii) PAR-CLIP cross-link-based substitution, whereas the observed A-to-C and G-to-C substitutions are assumed to originate only from (i) sequencing error, or (ii) SNPs or contamination. We ignore the cases of photo-activated sequence variants (i.e. where the reference is A, or G, while the sequence variant is T, and is substituted to C due to photo-activation by PAR-CLIP). With this simplification, we introduce the latent random variable $z_i \in \{1,2,3\}$ corresponding to the three possible reasons (i) to (iii) that can explain the observed nucleotide at locus i on the genome. Specifically, for reference T positions, $z_i = 1$ refers to background, $z_i = 2$ corresponds to a sequence variant and $z_i = 3$ refers to an RNA-RBP cross-link. For reference A or G positions, only $z_i = 1$ and $z_i = 2$ are possible.

We define ϵ as the probability of inducing a substitution due to sequencing noise. This probability accounts for all the modeled nucleotide substitutions (i.e. T-to-C, A-to-C, G-to-C, C-to-T, C-to-A and C-to-G), and is expected to be low. Consequently, at background positions ($z_i = 1$), the probability of occurrence of a specific substitution is ϵ. In the case of sequence variant loci ($z_i = 2$), where one can assume that the aligned reads originate from a genomic sequence which differs from the reference genome at the sequence variant locus (Supplementary Fig. S1), we expect the aligned reads to contain either the nucleotide from the sequence of origin (which is different from the reference), or transitions due to sequencing error. There are three possible transitions from the nucleotide in the sequence of origin, each with probability ϵ. Thus, the probability of substitution at sequence variant loci becomes $1 - 3\epsilon$. Finally, at cross-link loci ($z_i = 3$), which can, at the same time, be affected by sequencing errors, T-to-C substitutions occur with probability

$$\theta = (1-\gamma)\epsilon + (1-3\epsilon)\gamma, \tag{1}$$

where γ corresponds to the probability of a T nucleotide to be mutated to C following photo-activation and cross-link during PAR-CLIP, i.e. to the efficiency of the protocol to induce T-to-C substitutions at cross-link loci. We assume that the probability θ is bounded between ϵ and $1 - 3\epsilon$, which results in the constraint $\epsilon \leq 0.25$.

We denote by $\nu = P(z_i = 2)$ the probability of a locus on the genome to be a sequence variant, and, for the remaining cases which are not sequence variants, by ψ the probability that a genomic locus is a cross-link site. The observed data at each T reference position on the genome is then modeled by a constrained mixture of three binomial distributions, and the probability of an observed data point is

$$\begin{aligned} P((x_i,y_i)|r_i = T) &= \sum_{z_i=1}^{3} P((x_i,y_i)|z_i, r_i = T)P(z_i|r_i = T) \\ &= \underbrace{(1-\psi)(1-\nu)Bin(y_i;x_i,\epsilon)}_{\text{background}} \\ &\quad + \underbrace{\nu Bin(y_i;x_i,1-3\epsilon)}_{\text{sequence variant}} + \underbrace{\psi(1-\nu)Bin(y_i;x_i,\theta)}_{\text{cross-link}} \end{aligned} \tag{2}$$

where $\epsilon, \theta, \gamma, \nu$ and ψ are the parameters of the model, and the notation $Bin(k;n,p)$ corresponds to the probability mass function of the Binomial distribution, precisely the probability of having k successes within n trials with success probability p.

In absence of a control PAR-CLIP experiment, our model readily incorporates information from A-to-C and G-to-C alterations for a better estimation of the sequencing error ϵ and the probability ν. The probability of observed coverage and mismatch count for observed A-to-C and G-to-C alterations is

$$P((x_i,y_i)|r_i \in \{A,G\}) = \underbrace{(1-\nu)Bin(y_i;x_i,\epsilon)}_{\text{background}} + \underbrace{\nu Bin(y_i;x_i,1-3\epsilon)}_{\text{sequence variant}}. \tag{3}$$

The model (Fig. 1C) is thus fully defined by Equation (2) for reference genome T positions, and Equation (3) for A and G positions in the reference genome. Using these equations, we can derive the likelihood for the entire set of observations $\mathcal{D} = \{(x_i,y_i)\}_{i\in\mathcal{G}}$ throughout the whole genome $\mathcal{G}$ as follows:

$$L(\epsilon,\theta,\gamma,\nu,\psi) = P(\mathcal{D}|\epsilon,\theta,\gamma,\nu,\psi) = \prod_{i\in\mathcal{G}} P((x_i,y_i)|r_i) \tag{4}$$

We infer all the parameters of our model by maximizing the above defined likelihood with a gradient-based nonlinear constrained optimization (Powell, 1978). To classify each T locus on the genome as either background, sequence variant or cross-link, we choose the maximum of the posterior probabilities of the latent variable z_i:

$$\begin{aligned} P(z_i = 1|(x_i,y_i), r_i = T) &\propto (1-\psi)(1-\nu)Bin(y_i;x_i,\epsilon) \\ P(z_i = 2|(x_i,y_i), r_i = T) &\propto \nu Bin(y_i;x_i,1-3\epsilon) \\ P(z_i = 3|(x_i,y_i), r_i = T) &\propto \psi(1-\nu)Bin(y_i;x_i,\theta) \end{aligned} \tag{5}$$

where the $\propto$ symbol is used to represent proportionality between the posterior probability and the product between the likelihood and the prior, thus the normalization constant (same in all three cases) being omitted.

2.3 Construction of RNA-RBP-binding sites

By *RNA-RBP-binding site* we denote the region on the transcriptome where the protein of study attaches in order to fulfill a specific function. Once high-confidence cross-link T loci were identified as described in the previous section using a posterior cutoff of 95% for classification, BMix reports candidate-binding sites by using all the aligned sequencing reads that span these loci. Precisely, in order to construct the binding sites, all the reads spanning cross-link loci are grouped into clusters, and the cluster boundaries with coverage of 1 are trimmed out. Overlapping clusters (by at least 1 nucleotide) are grouped into contigs and reported as candidate RNA-RBP-binding sites (Supplementary Fig. S2). The user of BMix has the liberty to choose a different posterior cutoff, and a different trimming coverage.

2.4 Generation of the simulated data

For the generation of realistic synthetic data, we used the AGO2 PAR-CLIP data published in Kishore *et al.* (2011) for chromosme 1,

and, after read alignment, we introduced systematic A-to-C substitutions. We could not use the T-to-C substitutions, since these were already affected by the real PAR-CLIP protocol. These substitutions were thus ignored by all the applied methods on the simulated data, and did not affect the analysis outcome. Instead, A-to-C substitutions played the role of cross-link transitions in the simulated data. Precisely, after read alignment, we randomly chose 2000 A loci on chromosome 1. In each of these loci, we centered a window of size w. With probability μ, we inserted A-to-C conversions in the aligned reads at the A loci within each window (Supplementary Fig. S3). By altering A loci within a window w, we built regions where the incorporated A-to-C substitutions were more dense, simulating binding sites. The probability μ has the interpretation of how likely a photoactivated nucleotide within a binding site is efficiently substituted to C in the PAR-CLIP protocol. We simulated data both considering μ uniform across the different genomic positions, as well as non-uniform, drawn from a Beta distribution (see Section 3). For the latter case, the parameters of the Beta distribution were shared between all the altered genomic loci. The produced simulated data contain a realistic amount of sequencing errors and contamination, and is based on the alteration of a reference base different from T. We assessed the performance of BMix, PARalyzer and WavClusteR on these data.

2.5 Comparison with other methods

We compared BMix to PARalyzer v1.1 (Corcoran *et al.*, 2011) and WavClusteR v2.0.0 (Comoglio *et al.*, 2015) on the produced synthetic data, as well as on publicly available PAR-CLIP datasets published in Kishore *et al.* (2011), Sievers *et al.* (2012) and Hafner *et al.* (2010). On simulated data, the three methods were compared in terms of accuracy. The accuracy is defined as the ratio of true positive and true negative loci over the entire set of observed loci. The true positives correspond to the A loci where A-to-C substitutions were introduced in the synthetic datasets (cf. Section 2.4), and also reported as cross-link loci by the tested method. The true negatives correspond to the loci which were not altered in the synthetic datasets, and which were not reported as cross-link loci by the tested method. On real data, the methods were evaluated according to specific characteristics of the studied proteins, such as their affinity for microRNA (miRNA), 3′-untranslated regions (3′UTRs), and introns annotated in the RefSeq database (http://www.ncbi.nlm.nih.gov/refseq/), enrichment of protein-specific RNA recognition elements (RREs), as well as execution time. The reported execution time corresponds to the amount of time spent running by each method on one core of a Linux machine with a clock rate of 2.3 GHz.

3 Results

We validated our model on synthetic and real human PAR-CLIP data and assessed its performance compared with PARalyzer and WavClusteR methods (see Section 2).

3.1 Performance on simulated PAR-CLIP data

In the absence of a ground truth dataset for protein-binding sites, we generated a set of synthetic datasets in order to evaluate our model and compare it to the existing methods WavClusteR and PARalyzer. We started from real world PAR-CLIP data and mimicked *in silico* the PAR-CLIP protocol for a different substitution than T-to-C, precisely A-to-C. In this way, we kept the intrinsic noise and contamination levels specific to PAR-CLIP data and, at the same time, introduced validation cross-link loci. Furthermore, the simulated data was built independently from our model, providing an unbiased test set for all the methods. Knowing thus the genomic loci where A-to-C substitutions were introduced, we could test how accurate BMix as well as other methods were in detecting these loci.

We first performed a simulation study assuming μ is uniform and tested different values for μ, between 0.1 and 0.9, each time with three different window sizes: 15, 21 and 31 nucleotides long. We applied BMix, PARalyzer and WavClusteR on 100 simulated datasets (cf. Section 2), for each combination of μ and window size w. In all the different simulation scenarios, BMix had a significantly higher accuracy (Wilcoxon test $P < 10^{-4}$) than the other two methods (Fig. 2 and Supplementary Fig. S4). For example, for $w = 21$, BMix had on average an accuracy of 96%, compared with 88.6% for WavClusteR, and 76% for PARalyzer. All methods reached their lowest accuracy at $\mu = 0.1$. However, while WavClusteR and PARalyzer had similar accuracy for this case (75 and 71%, respectively), BMix outperformed them with an accuracy of 88%. By looking at the outcome of the three methods on a randomly chosen synthetic dataset with $\mu = 0.2$ and $w = 21$ (Supplementary Fig. S5), the main characteristics of each method were exposed. Precisely, after learning non-linear classification boundaries, BMix reached a true positive rate of 93%, and a true negative rate of 98% (Supplementary Fig. S5B). WavClusteR detected only 68% of the true validation loci, having difficulty in detecting the cross-link loci located in the low-coverage regions (Supplementary Fig. S5C). PARalyzer was affected by the low substitution probability, $\mu = 0.2$, and detected only 53.6% of the validation cross-link loci. Specifically, the method preferentially reported sites which had high-frequency A-to-C substitutions, and ignored validation sites with low-frequency transitions (Supplementary Fig. S5D).

Next, we performed simulations where the conversion probability μ was drawn from a Beta distribution whose parameters were the same across the simulated cross-link loci. For the previously defined window sizes, we tested different simulation scenarios, where we varied the mean of the Beta distribution between 0.1 and 0.9, and the variance was 0.001 or 0.005 (density of distributions showed in Supplementary Figs. S6 and S7). On all the newly simulated datasets, BMix had significantly superior accuracy. For example, for $w = 21$ and 0.001 variance of the Beta distribution, BMix had an average accuracy of 90%, while WavClusteR and PARalyzer had an average accuracy of 79.4 and 55.7%, respectively (Supplementary Fig. S8). The same behavior of the methods was observed as before, namely with

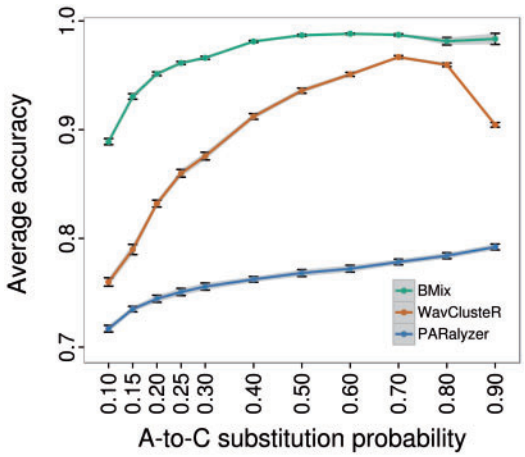

Fig. 2. Accuracy of BMix, WavClusteR, and PARalyzer on 100 simulated datasets generated with different uniform substitution probabilities, and window size $w = 21$ bases. The error bars and grey shade correspond to one SD of the accuracy over 100 synthetic datasets

PARalyzer reporting all the loci with high-frequency substitutions as cross-link sites, and WavClusteR having difficulties in identifying the cross-link loci within low coverage regions (Supplementary Fig. S9).

We further compared the three methods in terms of reporting the correct binding site boundaries on three synthetic datasets generated with conversion windows w of 31, 21 and 15 bases, and uniform probability $\mu = 0.2$. To assess the performance, for each identified validation locus, we computed the Jaccard Coefficient (Levandowsky and Winter, 1971) between the generated synthetic-binding site (window w centered in the locus) and the binding site reported for that locus with each method. For a window $w = 31$, BMix had a superior average Jaccard coefficient of 0.72 compared with 0.685 obtained with WavClusteR, and 0.643 with PARalyzer (Supplementary Fig. S10). For shorter lengths of the synthetic-binding sites, all the three methods decreased their performance in reporting the correct boundaries, having comparable average Jaccard coefficients (Supplementary Figs. S11 and S12).

3.2 Application to real PAR-CLIP datasets

We ultimately applied our method on three published human PAR-CLIP datasets corresponding to proteins AGO2, HUR (Kishore *et al.*, 2011) and MOV10 (Sievers *et al.*, 2012), and inferred the model parameters for each dataset. With BMix, one can choose to learn the parameters either separately for the forward and reverse strands, or combine the strands. Depending on the studied protein and experimental setup (e.g. strand-specific libraries), one of these procedures might be more appropriate. In our analysis, in order to avoid any influence from strand bias, we have learned the parameters of the model independently for both strands. However, for the three tested datasets, the results are equivalent when both strands are combined (Supplementary Table S1). We applied WavClusteR and PARalyzer on the same data and we compared the three different methods by evaluating their results according to specific characteristics of the proteins such as miRNA, 3′UTR, and intron affinity, as well as enrichment of protein-specific RREs.

By comparing the PAR-CLIP data for the AGO2 protein to matched RNA-Seq data from the same sample, published in Kishore *et al.* (2011), we observed the expected prevalence of T-to-C substitutions (Supplementary Fig. S13), but also a significantly larger amount of A-to-C and G-to-C alterations (one-tailed Wilcoxon test $P < 10^{-3}$) in the AGO2 PAR-CLIP dataset (Fig. 3 and Supplementary Fig. S14), indicating a high level of contamination.

3.2.1 Identification of AGO2-binding sites

Two replicate PAR-CLIP datasets were tested for the AGO2 protein. Because AGO2 is one of the proteins involved in RISC, its affinity to miRNAs and 3′UTRs was expected to be elevated (Hendrickson *et al.*, 2008).

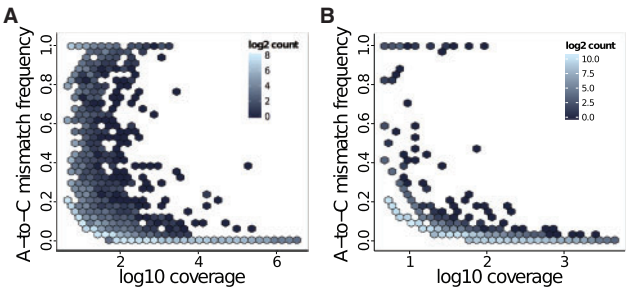

Fig. 3. Observed genome-wide A-to-C observed mismatch frequency as a function of log10 of coverage in a PAR-CLIP dataset **(A)**, as well as in the matched control RNA-Seq dataset **(B)**. The color intensity corresponds to log density of the data points

BMix identified 15 317 binding sites for replicate A, and 9615 binding sites for replicate B of the AGO2 dataset. We annotated the three classes of loci reported by BMix (background, sequence variant and cross-link) according to the Ensembl gene types retrieved using the UCSC Table Browser (Karolchik *et al.*, 2004). A higher proportion of background and sequence variant loci overlapped with ribosomal RNA (rRNA) and unannotated regions than the cross-link loci which mostly covered protein-coding regions (Fig. 4A). Both PARalyzer and WavClusteR reported around 4000 more binding sites than BMix (Supplementary Table S2). Annotation of the binding sites according to the same Ensembl types showed that a large proportion of these additional sites covered significantly more rRNA and unannotated regions and less protein-coding regions compared with the common sites (Fig. 4B). To assess the reproducibility of the three methods, each method was applied on each AGO2 replicate dataset independently, and the number of common miRNAs found within the binding sites between replicates was reported. All the three methods yielded a similar high percentage of reproducible miRNAs: 88.6% for BMix, 88.98% for PARalyzer and 89.1% for WavClusteR (Fig. 5A and Supplementary Table S2). All the three tested methods identified $\approx 30\%$ less binding sites for replicate B than for replicate A. We investigated the difference between replicates and we observed that, unlike replicate A, replicate B had a lower percentage of aligned reads (42.5% as opposed to 60.6%), a lower coverage, and a lower prevalence of T-to-C substitutions (Supplementary Fig. S15), explaining the lower amount of identified binding sites. Furthermore, Kishore *et al.* (2011) report a correlation of only 61% between these two replicates.

For both replicates, over 95% of the binding sites identified with BMix overlapped with the sites found by the two other methods. BMix reported on average 4% more of its binding sites within 3′UTRs than the other two methods (Fig. 5B and Supplementary Table S2). Over 70% of the binding sites reported by BMix were <30 nucleotides long (Supplementary Fig. S16A), in concordance with the expected small length of miRNA targets. Furthermore, given that it has been previously reported that some targeted 3′UTRs may have high miRNA target-site abundance (Garcia *et al.*, 2011), we investigated the number of identified binding sites in each 3′UTR. We found that over 85% of the covered 3′UTRs contained at most two binding sites (64% had only one binding site), while a small proportion had more than two binding sites (Supplementary Fig. S20A). A similar percentage was obtained when only binding sites in mRNAs were considered, precisely 86% 3′UTRs had at most two mRNA-binding sites (64.4% had only one binding site, Supplementary Fig. S21A).

We performed *de novo* motif discovery using the Software MEME (Bailey and Elkan, 1994) for the AGO2-binding sites found by each method in 3′UTRs. The motifs were ranked following their E-value, and only motifs with an E-value < 0.05 were reported. We observe that the identified motifs are analogous between the three methods, with the sequence 'TGCTGCT' found to be the most significant by all methods. Furthermore, these motifs were also found prevalent within sequences of various known miRNAs (Supplementary Tables S3–S5). In terms of execution time, the MEME motif search on BMix-binding sites was considerably faster, running for 11 h, as compared with 26 h for PARalyzer, and 2 days for WavClusteR.

3.2.2 Identification of MOV10-binding sites

Next, we applied the three methods on a dataset for the MOV10 protein, also expected to preferentially bind in 3′UTRs (Kenny *et al.*, 2014; Sievers *et al.*, 2012). In absence of replicates, we could not assess

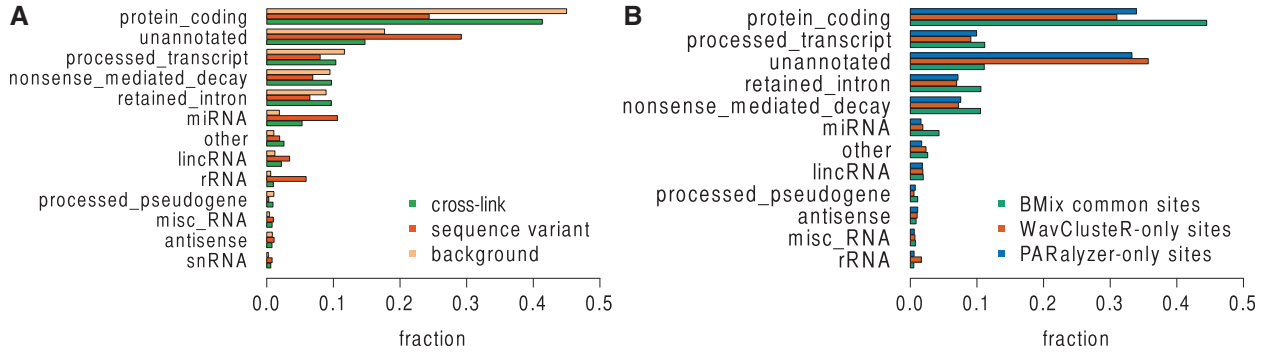

Fig. 4. (A) Proportion of BMix-classified loci within each Ensembl gene type retrieved using the UCSC table browser (Karolchik *et al.*, 2004) for replicate A of the AGO2 PAR-CLIP dataset (Kishore *et al.*, 2011). **(B)** Annotation according to the Ensembl gene types for binding sites commonly identified by BMix and the other three methods, as well as for the additional sites reported by PARalyzer and WavClusteR. The proportion of binding sites within each gene type is displayed. All the Ensembl types which contained <0.1% sites were grouped under the name '*other*' and all the sites which did not fall within any annotation were marked as '*unannotated*'

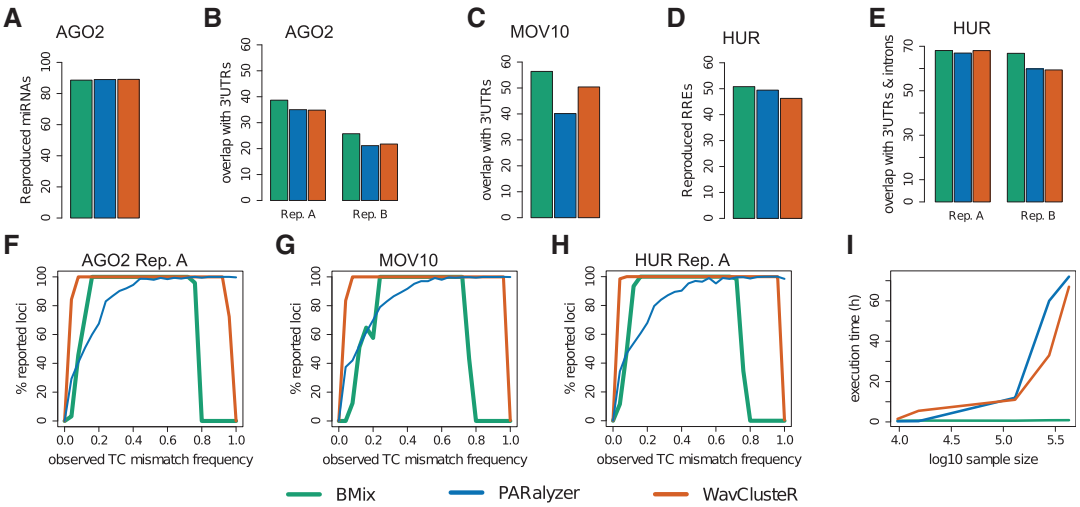

Fig. 5. Summary of results from applying BMix, PARalyzer and WavClusteR on human datasets for proteins AGO2, MOV10 and HUR. **(A)** Reproducibility of the methods in terms of percentage of commonly reported miRNAs between two AGO2 PAR-CLIP replicates. **(B)** Percentage of binding sites reported in 3′UTRs for the AGO2 datasets. **(C)** Percentage of binding sites reported in 3′UTRs for the MOV10 dataset. **(D)** Reproducibility of the methods in terms of commonly reported RREs between two HUR PAR-CLIP replicates. **(E)** Percentage of binding sites reported in 3′UTRs and introns for the HUR datasets. **(F–H)** Fraction of reported loci with a particular T-to-C substitution frequency out of the total number of observed loci with that substitution frequency. **(I)** Execution time for the three methods as function of log10 of the number of reported binding sites

the reproducibility of the methods on this dataset. However, as for the previous dataset, BMix found a higher percentage of its binding sites (56.39%) in 3′UTRs than PARalyzer and WavClusteR which obtained an overlap of 40.14 and 50.4%, respectively (Fig. 5C and Supplementary Table S6). Furthermore, >96% of the binding sites reported by BMix overlapped with the ones found by the two other methods.

3.2.3 Identification of HUR-binding sites

Finally, we applied a similar evaluation scheme on two replicate PAR-CLIP datasets for the HUR protein, a well-characterized RBP involved in maintaining mRNA stability and regulating gene expression (Peng *et al.*, 1998). It has been shown that this protein preferentially binds AU-rich regions in 3′UTRs of messenger RNAs, as well as intronic regions (Lebedeva *et al.*, 2011). Therefore, we quantified the amount of binding sites found by each method within these genomic features for each replicate independently, as well as their enrichment for HUR-specific RREs described in Ma *et al.* (1996).

For both replicates, the binding sites found by BMix overlapped over 90% with the binding sites reported by the other methods (Supplementary Table S7). To assess the reproducibility of the methods, we evaluated the percentage of RRE-enriched sites common between the two replicates. Even though it reported less binding sites than the other methods, BMix reached a higher reproducibility of 50.8% compared with 49.45% obtained with PARalyzer and 46.27% reported by WavClusteR (Fig. 5D and Supplementary Table S7). For both replicates, BMix had also a superior percentage of RRE-enriched sites than the other two methods (Supplementary Table S7).

BMix, PARalyzer and WavClusteR were applied on each replicate HUR dataset independently. For both replicates, BMix reported more binding sites within 3′UTRs and intronic regions compared with the other two methods (Fig. 5E). Precisely, 68.1% of the sites reported by BMix for replicate A overlapped with 3′UTRs and intronic regions, while PARalyzer and WavClusteR attained 66.95 and 68.06%, respectively. For replicate B, despite the lower number of reported sites, BMix had a superior overlap with 3′UTRs, namely 66.83%, compared with 59.83 and 59.37% obtained by PARalyzer and WavClusteR, respectively (Supplementary Table S7).

For all the datasets, the additional cross-link loci reported by PARalyzer and WavClusteR compared with BMix had either a low substitution frequency and low coverage, or a high substitution frequency. To illustrate this, we calculated, for each method, the fraction of reported loci with a particular T-to-C substitution frequency out of the total number of observed loci with that substitution frequency. PARalyzer reported all the high-frequency altered loci with no exception, while WavClusteR learned a more lenient threshold than BMix for the low and high substitution frequency regions (Fig. 5F–H and Supplementary Fig. S23). Furthermore, the additional binding sites reported by the two other methods were more prevalent in rRNA and unannotated regions compared with the BMix sites (Fig. 4B and Supplementary Fig. S25). A large proportion of the binding sites identified with BMix were short, with a median length of 30 bases. Paralyzer and WavclusteR yielded slightly shorter binding sites, with median lengths of 17 and 21 bases, respectively (Supplementary Figs. S16–S18). One can decrease the median length of binding sites reported by BMix by progressively increasing the trimming coverage threshold used to refine the clusters before merging them (Supplementary Fig. S19).

We analyzed the impact of the alignment strategy on the results of our method by testing BMix on Bowtie-aligned AGO2 PAR-CLIP data (Kishore *et al.*, 2011), allowing for one to three mismatches. The number of BMix-reported binding sites increased to over 40%, from 15 317 sites with one mismatch to 21 607 and 25 289 sites with two and three mismatches, respectively. Over 50% of the identified binding sites with two or three mismatches were also detected with one mismatch (Supplementary Table S9). More than 85% of the binding sites identified with BMix using one mismatch were also found by using two or three mismatches. Nevertheless, the configuration of the three types of loci classified with BMix changed as the number of mismatches was increased. Precisely, the fraction of reported cross-linked loci decreased, whereas the other two classes gained in size as the number of allowed mismatches increased (Supplementary Fig. S26A). On the other hand, PARalyzer doubled the number of reported binding sites when more than one mismatch was allowed, reporting 19 248 sites for one mismatch, 40 522 sites for two mismatches and 51 029 sites for three mismatches. WavClusteR functions specifically for alignments with one allowed mismatch, therefore testing for more mismatches was not possible. We performed the same analysis with BMix by choosing TopHat (Trapnell *et al.*, 2009) as aligner (Supplementary Fig. S26B) on the same dataset, although PAR-CLIP reads are generally too short to be efficiently used by a splice-aware aligner. Similar results were obtained with TopHat as with Bowtie, especially when one or two mismatches were allowed (Supplementary Tables S9 and S10, and Supplementary Fig. S26).

In terms of execution time, on one core of a Linux machine with a clock rate of 2.3 GHz, BMix proved to be considerably faster than the other two methods, running on average in <40 min on all the datasets. On the contrary, the execution time of both PARalyzer and WavClusteR on the same machine increased with the sample size from several hours to multiple days for the HUR datasets (Fig. 5I).

3.2.4 Comparison to RNA Bind-n-Seq for IGF2BP1

We also performed an analysis on a PAR-CLIP dataset for protein IGF2BP1 (Hafner *et al.*, 2010), where an RNA Bind-n-Seq dataset was also available [Lambert *et al.* (2014), project ENCSR928XOW on ENCODE]. RNA Bind-n-Seq (RBNS) is a recent method that queries *in vitro* the binding predilections of RBPs against an elaborate, randomized pool of RNA short fragments at different protein concentrations. It has been shown that RBNS provides an extensive, biologically relevant landscape of RBP-binding motifs, covering various affinities and has been proposed as an enhancement and validation step for CLIP-based methods (Lambert *et al.*, 2014). We used the 7mers found by RNBS at concentration 320 nM, which had an enrichment value (K) two SDs larger than the overall mean [as proposed by Lambert *et al.* (2014), Supplementary Fig. S22] to validate the PAR-CLIP-binding sites identified with the three methods. Surprisingly, over 76% of the binding sites reported by BMix also contained at least one of the significant RBNS motifs, while only 46.06% of the Paralyzer sites, and 65.24% of the WavClusteR sites presented RBNS motifs within their sequences. Furthermore, the canonical motifs 'CAUU' and 'CUUU'were contained within 30% of the RBNS significant 7mers, and were also found more prevalent within the BMix-binding sites as compared with the other methods (Supplementary Table S8).

4 Discussion

In this work, we have proposed BMix, a new probabilistic model for identifying high-confidence RNA-protein interaction sites from PAR-CLIP data. BMix uses a constrained three-component binomial mixture model to describe the T-to-C substitutions observed at genomic loci in three categories: low-frequency errors due to sequencing noise, true cross-link sites or high-frequency sequence variants caused by SNPs or contamination. Therefore, our model brings the novelty of accounting for both low and high-frequency erroneous T-to-C alterations in PAR-CLIP data. We validated and demonstrated the superior performance of BMix compared with the methods WavClusteR v2.0.0 and PARalyzer v1.1 both on synthetic and real data.

Most of the current PAR-CLIP analysis methods focus on filtering low-frequency altered loci and consider the high-frequency substitutions as reliable indicators of cross-linking. We have observed this behavior also within our study on synthetic and real data, where PARalyzer has selected all the high-frequency altered loci within its reported binding sites (Figs. 5F–H). However, methods like PARalyzer ignore the possibility that high-frequency alterations could have been caused by other factors, such as contamination, or single nucleotide variants. By comparing PAR-CLIP and matched control RNA-Seq data from the same experiment, the prevalence of highly altered loci was clearly observed also for non T-to-C substitutions in published data (Fig. 3), which motivates the need of identifying and discarding spurious highly altered loci from the analysis. BMix had on average an accuracy at least 20% larger than PARalyzer on simulated data (cf. Section 3.1 of Results). So far, only WavClusteR accounts for high-frequency substitutions. However, because it uses relative substitution frequency values instead of actual read counts, the method looses performance especially in low and high substitution regions, and where the coverage is weak, as presented in real and synthetic data applications, having an accuracy on average inferior by 7% to BMix. In an extensive simulation study, by varying the probability μ, we have observed that the accuracy of both BMix and WavClusteR attains a global maximum at $\mu = 0.6$ and 0.7, respectively, and then decreases. In other words, a perfect 100% PAR-CLIP T-to-C substitution efficiency would not improve the result; on the contrary, it would make difficult to differentiate between induced substitutions and high-frequency errors.

The Ensembl annotation of the three classes of loci reported by BMix showed that our model captures more rRNA and unannotated RNA within its background and sequence variant mixture components, while the reported cross-link loci mainly cover protein-coding regions and miRNAs. The PARalyzer approach, based on selecting high-frequency mutations as high-confidence-binding sites is therefore at risk of reporting a large amount of spurious alterations as

cross-link loci. WavClusteR is exposed to the same risk by using relative frequencies instead of substitutions counts, thus disregarding uncertainty in the low coverage regions. BMix identified less binding sites than the other two methods, at the same time keeping high the proportion of reproducible binding sites overlapping with features of interest. The other methods detected many additional sites that typically overlapped with rRNA and unannotated regions. This suggests that BMix output is more reliable and contains less false positives. This facilitates downstream analyses, such as motif search, or understanding RNA regulation. The binding sites identified with BMix had a higher prevalence of canonical-binding motifs previously reported in literature, as well as RNA Bind-n-Seq motifs, compared with the other methods.

BMix, as well as the other PAR-CLIP analysis methods, relies on the alignment of sequencing reads to a reference genome. In this work, we have chosen the standard alignment strategy employed by PARalyzer and WavClusteR, aligning the sequencing reads with Bowtie and allowing for one mismatch. However, a strict alignment strategy would potentially discard reads with viable cross-link T-to-C alterations. Due to explicit modeling of noise and contamination, our model is less sensitive to the choice of alignment and is able to control the false positive rate even for more lenient alignment parameters. As a result, the user of BMix can pick the alignment procedure which best suits the data without having to maintain a too strict control on the alignment parameters. This procedure depends on multiple aspects such as, e.g. the quality of the sequencing, the length of the reads or the binding protein (König *et al.*, 2012). Ultimately, a systematic comparison of the results obtained from different alignment strategies can be utilized for a better quantification of the binding sites.

A limitation of our method consists in the difficulty of sorting out T-to-C substitutions of moderate frequency which have not been introduced by PAR-CLIP. These can occur following diverse molecular processes such as, e.g. RNA editing by ADAR enzyme (Samuel, 2011), following contamination, or as heterozygous SNPs. In this case, any analysis relying entirely on coverage and substitution counts does not have sufficient statistical power to discard these false loci, and our method can report false positives. Nevertheless, these substitutions would also appear during a control experiment. Therefore, a potential solution to this limitation would be to use a control PAR-CLIP, or RNA-seq experiment. These would reveal spurious substitutions to be subtracted from the cross-link T-to-C loci identified with BMix.

Due to the reduction of sequencing costs, a high increase in sequencing-based experiments such as PAR-CLIP is expected in the near future. BMix provides a rigorous probabilistic method which is significantly faster and more accurate than the current state-of-the art methods for detecting RNA-protein interaction sites in PAR-CLIP data.

Acknowledgements

The authors thank Federico Comoglio and Cem Sievers for valuable feedback and discussions as well as their support with WavClusteR.

Funding

Monica Golumbeanu was financially supported by the Swiss National Science Foundation (project "Dynamics of HIV latency and reactivation at population and single-cell level", grant number 31003A_146579/1).

Conflict of Interest: none declared.

References

Bailey,T. and Elkan,C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. 28–36.

Blencowe,B.J. *et al.* (2009). Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev.*, **23**, 1379–1386.

Chen,B. *et al.* (2014). PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biol.*, **15**, R18.

Comoglio,F. *et al.* (2015). Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. *BMC Bioinformatics*, **16**, 32+.

Corcoran,D. *et al.* (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, **12**, R79.

Erhard,F. *et al.* (2013). PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome Biol.*, **14**, R79.

Garcia,D.M. *et al.* (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 1139–1146.

Gerstberger,S. *et al.* (2014). A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.

Hafner,M. *et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.

Hendrickson,D.G. *et al.* (2008). Systematic identification of mRNAs recruited to Argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. *PloS One*, **3**, e2126.

Karolchik,D. *et al.* (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.

Kechavarzi,B. *et al.* (2014). Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol.*, **15**, R14.

Kenny,P.J. *et al.* (2014). MOV10 and FMRP regulate AGO2 association with microRNA recognition elements. *Cell Rep.*, **9**, 1729–1741.

Kishore,S. *et al.* (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, **8**, 559–564.

Kloetgen,A. *et al.* (2014). Biochemical and bioinformatic methods for elucidating the role of RNA-protein interactions in posttranscriptional regulation. *Brief. Funct. Genomics*, **14**, 101–114.

König,J. *et al.* (2012). Protein-RNA interactions: new genomic technologies and perspectives. *Nature*, **13**, 77–83.

Lambert,N. *et al.* (2014). RNA bind-n-seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol. Cell*, **54**, 887–900.

Langmead,B. *et al.* (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lebedeva,S. *et al.* (2011). Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR. *Mol. Cell*, **43**, 340–352.

Levandowsky,M. and Winter,D. (1971). Distance between sets. *Nature*, **234**, 34–35.

Lunde,B. *et al.* (2007). RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.

Ma,W.-J. *et al.* (1996). Cloning and characterization of HuR, a ubiquitously expressed elav-like protein. *J. Biol. Chem.*, **271**, 8144–8151.

Meister,G. (2013). Argonaute proteins: functional insights and emerging roles. *Nat. Rev. Genet.*, **14**, 447–459.

Muller-McNicoll,M. *et al.* (2013). How cells get the message: dynamic assembly and function of mRNA - protein complexes. *Nat. Rev. Genet.*, **14**, 275–287.

Peng,S.S.-Y. *et al.* (1998). RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein. *EMBO J.*, **17**, 3461–3470.

Powell,M. (1978). A fast algorithm for nonlinearly constrained optimization calculations. In: Watson,G.A. (ed.) *Lecture Notes in Mathematics*. **Vol 630**. Springer, Berlin Heidelberg.

Samuel,C.E. (2011). Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. *Virology*, **411**, 180–193. Special Reviews Issue 2011.

Sievers,C. *et al.* (2012). Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res.*, **40**, e160.

Trapnell,C. *et al.* (2009). Tophat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Yun,J. *et al.* (2014). Bayesian hidden Markov models to identify RNA - protein interaction sites in PAR-CLIP. *Biometrics*, **70**, 430–440.